Vol.2010-SLP-82 No.14
2010/7/24

KL情報量によるAnchor modelの階層的クラスタリングに基づく話者認識

細川 光政[†]　西田 昌史[†]　山本 誠一[†]

従来のアンカーモデルではアンカーモデルを無作為に選択しており，多数のモデルを必要としていた．それに対して，本研究ではアンカーモデルを最適化するために，GMM 間の KL 距離に基づいてアンカーモデルを階層的にクラスタリングする手法を提案した．本手法により，無作為に選択されたアンカーモデルに対して音響的に類似した話者をクラスタリングすることで，効率的な認識を実現することができる．本手法の有効性を示すために，従来よく用いられている BIC に基づく話者クラスタリング手法との話者認識実験を行った結果，提案手法は従来手法に比べて認識精度の改善が得られた．

Speaker Recognition Based on Agglomerative Clustering Using KL Divergence for Anchor Model

Mitsumasa Hosokawa[†], Masafumi Nishida[†], and Seiichi Yamamoto[†]

In conventional methods, it was needed many models and selected Anchor models at random. We proposed a speaker recognition method based on agglomerative clustering using KL divergence for Anchor model. The proposed method can recognize efficiently by clustering similar speakers acoustically for Anchor models. We conducted speaker recognition experiments using clustering methods based on BIC and proposed method. As a result, the proposed method can improve the recognition accuracy compared with the conventional BIC method.

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や，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 GMM（Gaussian M ixture M odel）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を用いて構成する手法[7]なども提案されている．また，各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．そこで，本研究では，認識対象の話者を識別するのに有効なアンカーモデルを構成する手法として，話者クラスタリングによる手法について検討を行う．アンカーモデルに対して話者クラスタリングを行うことで，音響的に類似した話者をマージし，識別に有効なアンカーモデルを効果的に生成することができるのではないかと考える．

これを踏まえて本研究では，アンカーモデルを GMM により学習し，GMM 間の KL 距離[9]に基づいて階層的にクラスタリングする手法を提案する．本手法の有効性を示すために，従来よく用いられている BIC(Bayesian Information Criterion)に基づく話者クラスタリング手法[10]との比較実験を行う．

以降，2 章にて従来のアンカーモデルによる認識ならびに BIC に基づくクラスタリング，3 章にて提案手法である KL 距離に基づく階層的クラスタリング，4 章にて評価実験により得られた結果，5 章にてまとめと今後の課題について述べる．

[†] 同志社大学
Doshisha University

Vol.2010-SLP-82 No.14
2010/7/24

2. アンカーモデルによる話者認識

本章では，従来手法であるアンカーモデルによる話者認識と，BIC に基づくアンカーモデルのクラスタリングについて述べる．

2.1 アンカーモデルによる認識

GMM による話者認識では，認識対象話者の音声データで GMM を学習し，入力された発話と各 GMM との尤度により話者識別を行っていた．これに対して，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．そして，入力された発話と認識対象以外の話者ごとの尤度を求め，この尤度を話者ベクトルの要素とし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にて認識を行う手法である．

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(2.1)のように求められる．ここでx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．

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．本手法では，入力音声から話者ベクトルV_j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．

$$V_j = \begin{bmatrix} \dfrac{P(x_j \mid A_1) - \mu_j}{\sigma_j} \\ \dfrac{P(x_j \mid A_2) - \mu_j}{\sigma_j} \\ \vdots \\ \dfrac{P(x_j \mid A_U) - \mu_j}{\sigma_j} \end{bmatrix} \tag{2.1}$$

$$\mu_j = \frac{1}{U}\sum_{u=1}^{U} P(x_j \mid A_u) \tag{2.2}$$

$$\sigma_j = \sqrt{\frac{1}{U}\sum_{u=1}^{U}(P(x_j \mid A_u) - \mu_j)^2} \tag{2.3}$$

話者ベクトル間のユークリッド距離hは，式(2.4)により求められる．入力話者のu次元目のベクトル要素をv_u，登録話者のu次元目のベクトル要素をv_u'とする．

$$h = \sqrt{\sum_{u=1}^{U}(v_u - v'_u)^2} \tag{2.4}$$

図 1 に三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から求めたベクトルの要素を示す．このように，あらかじめ登録話者の音声から得られた話者ベクトルと，入力音声の話者ベクトルとの距離計算を行うことで，話者を識別する．

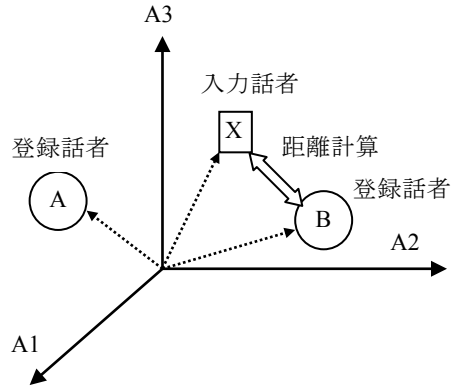

図 1　アンカーモデルによる話者ベクトル

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．

Vol.2010-SLP-82 No.14
2010/7/24

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るが，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．

2.2 BICに基づくアンカーモデルのクラスタリング

BIC(Bayesian Information Criterion)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の BIC 値と，異なる特徴を持つと仮定した場合の BIC 値と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(2.5)により求まる．係数αは，BICを用いた最適化で導入される重み係数である．重み係数αの値は実験的に決める必要がある．

$$\Delta BIC_{\mathrm{var}} = \frac{N_1+N_2}{2}\log|\Sigma_0| - \frac{N_1}{2}\log|\Sigma_1| - \frac{N_2}{2}\log|\Sigma_2| - \alpha\frac{1}{2}\left(d+\frac{d(d+1)}{2}\right)\log(N_1+N_2) \tag{2.5}$$

BICに基づくクラスタリングでは，式(2.5)の$\Delta \mathrm{BIC}_{\mathrm{var}}$値が正であれば 2 つの話者をマージする．その際，BIC値が最も大きい話者間から順次マージしていく．全ての発話間でBIC値が負になれば，どの発話もマージすべきでないとしてクラスタリングの処理を終了する．

以上の処理により得られたクラスタごとに GMM を学習してアンカーモデルとする．

3. KL距離に基づくアンカーモデルのクラスタリング

本章では，提案手法である 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手法について述べる．

3.1 GMM間のKL距離

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．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(3.1)のように混合分布間の距離尺度に拡張して用いた．ここで，b は話者 r のモデルの分布番号，c 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，n は特徴ベクトルの次元数を示している．また，μ，σ は各 GMM の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．

$$d(r,s) = \frac{1}{m}\sum_{b=1}^{m}\min_{c} KL(b,c) \tag{3.1}$$

$$KL(b,c) = \sum_{i=1}^{n}\left\{\frac{\sigma^2{}_{bi}-\sigma^2{}_{ci}+(\mu_{ci}-\mu_{bi})^2}{\sigma^2{}_{ci}} + \frac{\sigma^2{}_{ci}-\sigma^2{}_{bi}+(\mu_{ci}-\mu_{bi})^2}{\sigma^2{}_{bi}}\right\}$$

3.2 KL距離に基づく階層的クラスタリング

次に, GMM 間の 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について述べる．

本研究では，すべてのアンカーモデル同士の GMM 間の KL 距離を求め，最小距離が閾値よりも小さければそれらの話者同士をマージする．その後，どれにもマージされなかった話者とクラスタとの KL 距離を比較し，最小距離が閾値よりも小さければその話者をクラスタに加える．単独の話者のクラスタリングが終わってから，クラスタ同士の KL 距離を比較し，最小距離が閾値よりも小さければそれらのクラスタをマージする．このようにすることで，特定のクラスタに話者が集中してクラスタリングされないように対応した．

一連のクラスタリング処理が終わって得られたクラスタごとに GMM を再学習して，これらをアンカーモデルとして認識を行う．クラスタリングの処理の流れを図 2 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

2. KL 距離が最小となるモデル同士を新たなクラスタとする

Vol.2010-SLP-82 No.14
2010/7/24

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選出．

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 GMM を再学習し，これらをアンカーモデルとする．

図 2　クラスタリングの流れ

4.　評価実験

本章では，従来の GMM とアンカーモデルによる話者認識，従来の BIC ならびに提案手法によるアンカーモデルのクラスタリングによる話者認識実験を行う．

4.1　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 文章のデータで，各文章における 3 種類の発声速度(普通，遅い，速い)の計 30 文章である．

また，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国立国語研究所と通信総合研究所によって開発された「日本語話し言葉コーパス」（Corpus of Spontaneous J apanese：以下 CSJ と省略する）に含まれる講演音声を用いた．1 人あたり無音区間を除いたおよそ 30～90 秒の発話で，500 名の話者を用いた．なお，300ms 以上の無音区間を基準に発話を分割した．

従来の GMM による認識手法では，学習データとして最初の時期 90 年 8 月の普通の速さ 1 文章を用いて行い，認識では全 7 時期の学習とは異なる 5 文の 3 速度，計 15 文章を用いた．アンカーモデルの話者ベクトルの学習には，従来の GMM と同様に 90 年 8 月の 1 文を用いた．アンカーモデルによる認識でも従来の GMM と同じ 15 文を用いて行った．

本実験で用いた音声データは，音声データ（16kHz，16bit）に対しフレーム長 25ms のハミング窓，フレーム周期 10ms で音響分析を行っている．そして，フレーム毎に 12 次の MFCC の特徴量を求めている．

4.2　実験結果と考察

従来の GMM において混合分布数を変化させた際の認識結果を図 4.1 に示す．

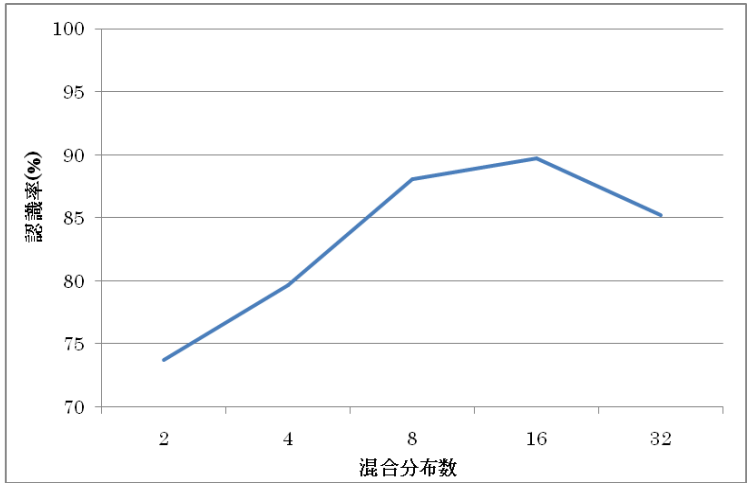

図 4.1　従来の GMM による認識結果

従来の GMM による認識を行った結果, 混合数が 2 のとき 73.8%, 4 のとき 79.6%, 8 のとき 88.1%，16 のとき 89.7%，32 のとき 85.2%となり，混合分布数が 16 のときに最も認識精度が高くなった．

次に従来のアンカーモデルにおいてアンカーモデル数が 100 のときに，GMM の混合分布数を変えた際の認識結果を表 4.1 に示す．

表 4.1　アンカーモデルにおける混合分布数の違いによる認識結果

混合数	32	64	128
認識率(%)	85.8	86.2	75.2

従来のアンカーモデルにおいてモデル数が 100 の際，混合分布数が 64 のときに認識精度が 86.2%と最も高くなった．したがって，以降のアンカーモデルによる実験では GMM の混合分布数を 64 に設定して行う．

アンカーモデル数を変化させたときの認識結果を図 4.2 に示す．

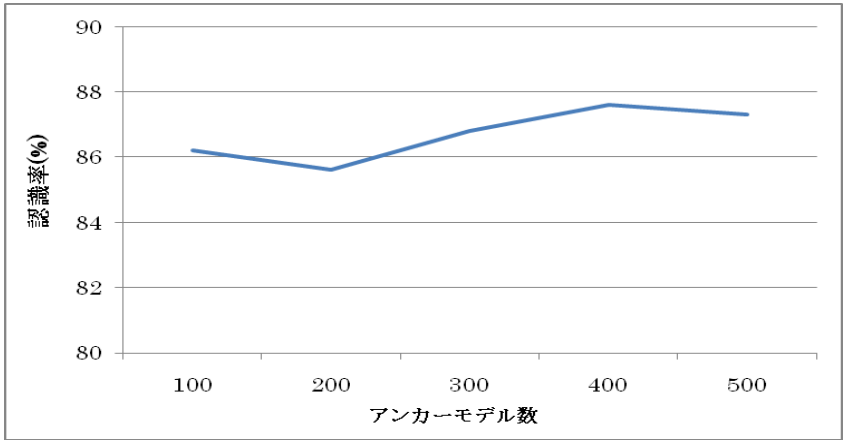

図 4.2　アンカーモデル数の違いによる認識結果

アンカーモデルによる認識を行った結果，モデル数が 100 のとき 86.2%，200 のとき 85.6%，300 のとき 86.8%，400 のとき 87.6%，500 のとき 87.3%という結果が得られ，モデル数によって認識精度が変動していることからどのようにアンカーモデルを選択するかが重要であることがわかる．今回の実験では，アンカーモデル数が 400 のときが最も高い認識精度になった．

次に，BIC ならびに提案手法に基づいてアンカーモデルをクラスタリングした際の認識結果を表 4.2 に示す．表中のアンカーモデルは従来のアンカーモデルでクラスタリングを行っていないときの結果，BIC は従来の BIC に基づいて重み係数を変化させてアンカーモデルをクラスタリングしたときの結果，提案手法は GMM 間の KL 距離に基づいて閾値を変化させてアンカーモデルをクラスタリングした際の結果である．また，各手法により得られたアンカーモデル数を表 4.3 に示す．

表 4.2　各手法における認識結果

アンカーモデル	BIC	提案手法
87.6%	85.4%	89.9%

表 4.3　各手法におけるアンカーモデル数

アンカーモデル	BIC	提案手法
400	200	249

従来のクラスタリングを行わないアンカーモデルによる手法では，アンカーモデル数が 400 のときに最も高い認識精度の 87.6%が得られた．従来の BIC によるクラスタリング手法では，アンカーモデル数が 200 のときに最も高い認識精度の 85.4%が得られ，従来のアンカーモデルに比べて半分にモデル数が削減された．それに対して，提案手法では，アンカーモデル数が 249 のときに最も高い認識精度の 89.9%が得られ，従来のアンカーモデルに比べて約 40%のモデル数を削減することができた．

以上の結果から，提案手法は無作為に選択された 400 個のアンカーモデルをクラス

タリングすることでモデル数を削減することができ，従来の BIC によるクラスタリングよりも高い認識精度が得られた．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認識を行っていたが，提案手法によりクラスタリングすることで，認識に有効なアンカーモデルを構成できることが明らかになった．

5. おわりに

本研究では，アンカーモデルを GMM で学習し，GMM 間の KL 距離に基づいてアンカーモデルを階層的にクラスタリングする手法を提案した．

本手法により，無作為に選択されたアンカーモデルをベースに，話者クラスタリングを行うことで約 40%のモデル数を削減することができた．また，従来よく用いられている BIC による話者クラスタリング手法との比較実験を行った結果，従来手法よりも高い認識精度が得られた．これらの結果から，提案手法により認識に有効なアンカーモデルを生成できることが明らかになった．

今後は，認識対象の話者を識別するの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きたいと考えている．

参考文献

1) A. Park and T.J. Hazen, “Asr dependent techniques for speaker identification”, Proc. ICSLP, pp.1337-1340，2002.

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 eith Speaker Adapted Syllable-Based HMM”IEICE TRANS.INF.&SYST，vol.E89-D,NO3,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) 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.

10) 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.