

# 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

細川光政 西田昌史 山本誠一

同志社大学 大学院工学研究科

## 1. はじめに

近年、セキュリティのための生体認証や、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において話者認識に関する研究がさかんに行われている。

従来の話者認識の手法としては、GMM (Gaussian Mixture Model) がよく用いられてきた。GMM による手法では多くの学習データを用い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 (GMM, HMM) を用いることで、少量の学習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 [1][2]。

従来のアンカーモデルによる手法では、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が、アンカーモデルの明確な選択基準が無く無作為に選択している。そこで、UBM (Universal Background Model) を初期モデルとして MAP 推定により得られた GMM 間の KL 距離をもとに音響的に類似した話者をマージし、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を提案する。

## 2. Anchor model 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を学習する。

各パラメータは、UBM の各混合分布の重み  $w$ 、平均  $\mu$ 、分散  $\sigma^2$  をもとに以下の式により適応する。

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

$$\hat{\mu}_i = \alpha_i E_i(x) + (1 - \alpha_i) \mu \quad (2)$$

$$\hat{\sigma}_i^2 = \alpha_i E_i(x^2) + (1 - \alpha_i)(\sigma_i^2 + \mu_i^2) - \hat{\mu}_i^2 \quad (3)$$

ここで、 $\gamma$  は混合分布の重みの総和を制御する係数を、 $E_i(x)$  は学習データの各分布に対する事後確率を表し、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 Anchor model による認識

アンカーモデルによる認識では、入力された発話と認識対象以外の各話者の尤度を求め、この尤度を話者ベクトルの要素とする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。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。識別対象話者と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。

### 3. BIC に基づく Anchor model のクラスタリング

BIC (Bayesian Information Criterion) に基づく手法は、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  $\Sigma_0$ 、1 人目の話者の共分散行列を  $\Sigma_1$ 、2 人目の話者の共分散行列を  $\Sigma_2$ 、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 (4) により求まる。 $\alpha$  は、重み係数であり、実験的に決める必要がある。

$$\Delta BIC = \frac{N_1 + N_2}{2} \log |\Sigma_0| - \frac{N_1}{2} \log |\Sigma_1| - \frac{N_2}{2} \log |\Sigma_2| \quad (4) \\ - \alpha \frac{1}{2} \left( d + \frac{d(d+1)}{2} \right) \log(N_1 + N_2)$$

$\Delta BIC$  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、全ての発話間で BIC 値が正になれば、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。

#### 4. KL 距離に基づく Anchor model 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(5)のように混合分布間の距離尺度に拡張して用いた。

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

$$KL(p, q) = \sum_{i=1}^d \left\{ \frac{\sigma_{pi}^2 - \sigma_{qi}^2 + (\mu_{qi} - \mu_{pi})^2}{\sigma_{qi}^2} \right. \quad (5)$$

$$\left. + \frac{\sigma_{qi}^2 - \sigma_{pi}^2 + (\mu_{qi} - \mu_{pi})^2}{\sigma_{pi}^2} \right\}$$

GMM 間の KL 距離が閾値よりも小さい話者をマージし、再学習は行わずクラスタに属する GMM との平均距離に基づいてクラスタリングを行う。最終的に得られたクラスタ毎に 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。

#### 4. 評価実験

##### 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 時期に発声したデータである。UBM ならびにアンカーモデルの学習データには、「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 人あたり約 60 秒の発話で、600 名の話者データを UBM の学習、それと異なる 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手法では、学習データに最初の時期 90 年 8 月の普通の速さ 1 発話を用い、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。

##### 4.2 実験結果と考察

従来の UBM-MAP にて作られた GMM による手法では 81.0%、UBM-MAP にて作られたアンカーモデルによる手法では、アンカーモデル数 500 のときに 83.3%となった。学習データ量が少ないときは、アンカーモデルの方が有効であることが示せた。表 1 に BIC と KL 距離に基づくクラスタリングによって得られたクラスタ数を示す。ベースラインは、クラスタリングを行わないアンカ

ーモデルの数である。

表 1 各手法によるアンカーモデル数

| Baseline | BIC | KL  |
|----------|-----|-----|
| 500      | 250 | 265 |

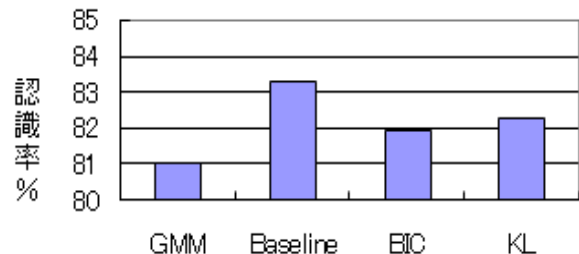


図 1 各手法による話者認識結果

図 1 に各クラスタリング手法におけるアンカーモデルによる認識結果を示す。BIC に基づくクラスタリングでは 81.9%、提案手法である GMM 間の KL 距離による階層的クラスタリングでは 82.3%の認識率が得られた。

以上の結果から、提案手法は無作為に選択された 500 個のモデル数での通常のアンカーモデルによる認識精度とほぼ同じ結果が得られていることから、500 個を初期モデルと考えた場合約 50%の割合でアンカーモデル数を削減でき、提案手法が認識精度を劣化させることなくモデル数を減らすことができた。

#### 5. おわりに

本研究では、UBMを初期モデルとしてMAP推定により学習した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。認識精度を劣化させずにモデル数を大幅に削減できた。

今後は、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きたいと考えている。

#### 参考文献

- [1]Yassine Mami , Delphine Charlet: Speaker recognition by location in the space of reference speakers, Speech Communication 48, pp.127-141 (2006).
- [2]小坂哲夫ら: 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).