Vol.2011-SLP-85 No.7
2011/2/4

非負値行列因子分解に基づく多人数会話における話者分類

西田昌史[†1]　石川勇樹[†1]　山本誠一[†1]

多人数会話における話者分類では，従来発話ごとにモデルを学習し，モデル間の距離に基づいて階層的なクラスタリングが行われてきた．しかし，発話数が多くなると距離の算出ならびにクラスタの探索に処理時間がかかってしまう．また，発話対のみの距離に基づいてクラスタリングが行われる．それに対して，本研究では非負値行列因子分解に基づく話者クラスタリング手法を提案した．本手法により，発話ごとに学習したモデル間の距離を要素にした行列を分解することで，全発話対の距離を総合的に捉えて高精度で高速な話者分類を実現することができる．本手法の有効性を示すために，従来手法として BIC に基づくクラスタリング，GMM 間の尤度比によるクラスタリング，UBM を初期モデルとした MAP 推定で学習した GMM に基づく手法との話者分類実験を行った．その結果，提案手法はすべての評価データにおいて従来手法に比べて最も高い話者分類精度を得ることができた．

Speaker Diarization Based on Non-negative Matrix Factorization in Multi-party Conversations

Masafumi Nishida,[†1] Yuuki Ishikawa[†1]
and Seiichi Yamamoto[†1]

In conventional speaker diarization, hierarchical clustering methods are used based on distances between models. However, the methods need many processes such as distance calculation and cluster search when there are many utterances in data. We proposed a novel clustering method based on non-negative matrix factorization. The proposed method can perform the fast and robust clustering by factorizing a matrix consisted of distances between models. We conducted speaker diarization experiments using a clustering method based on BIC, likelihood ratio between GMMs and GMMs trained by MAP adaptation from UBM, and proposed method. As a result, the proposed method was able to obtain the high diarization accuracy compared with the conventional methods.

1. はじめに

近年，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て複数話者が発話している音声からいつ誰が発話したかを自動的に抽出する話者分類に関する研究がさかんに行われている[1)]．

従来手法としては，発話間の VQ ひずみを用いて話者交替を識別し話者クラスタリングを行う手法[2)] や，発話ごとに直接モデルを学習するのではなく別途学習されたアンカーモデルを用いてそれらに対する尤度を要素とした話者ベクトルにより発話をクラスタリングする手法[3)]，発話の時間長のばらつきに対応するために学習データ量に応じて VQ と GMM を対象に最適な話者モデルを BIC に基づいて選択してクラスタリングする手法[4)] が提案されている．

また，事前知識として eigenvoice による話者空間を利用して特定話者の検出ならびに判別を行っているもの[5)] や，複数の特徴量を統合して KL 距離を用いて話者分類を行う手法[6)]，発話ごとに特定の音素の情報のみを用いて BIC に基づいてオンラインで話者識別を行う手法[7)] などが提案されている．

さらに，これまでの手法では主に音響情報のみを用いていたが，マイクロホンアレイと複数のカメラを用いて音響情報と映像情報を確率的に統合することでいつ誰が発話しているかを抽出する手法について検討されている[8)]．

これまでの従来研究では，複数話者の音声から発話区間を抽出し，発話間の類似度を計算してその値に応じて階層的にクラスタリングする手法が主に用いられてきた．しかし，本手法では全発話間の類似度を計算した後，階層ごとにクラスタ間の距離を再度計算してマージすべきクラスタを探索するため，発話数が多いほど処理時間がかかってしまう．また，単一発話対の類似度に基づいてクラスタリングが行われている．

そこで，本研究では非負値行列因子分解 (Non-negative Matrix Factorization, NMF)[9)] を用いた話者クラスタリング手法を提案し，モデル間の距離を行列で表現して処理することでクラスタリングのコストを削減し，全発話対の距離を総合的に捉えることでより高精度な話者分類を実現する．これまでに我々は，アンカーモデルによる話者認識において，ア

†1 同志社大学
Doshisha University

Vol.2011-SLP-85 No.7
2011/2/4

ンカーモデルのクラスタリングに NMF を適用した手法を提案し，その有効性を示している[10]．今回，本手法を教師なしの話者分類においても有効であることを示す．

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[11]，EM アルゴリズムにより学習した GMM 間の尤度比に基づく手法[12]，UBM(Universal Background Model) を初期モデルとした MAP 推定により学習した GMM[13] を用いた手法との比較実験を行う．

以降，2 章にて BIC に基づくクラスタリング，3 章にて EM アルゴリズムにて学習した GMM ならびに UBM をもとに MAP 推定で学習した GMM によるクラスタリング，4 章にて提案手法である NMF に基づくクラスタリング，5 章にて評価実験により得られた結果と考察，6 章にてまとめと今後の課題について述べる．

2. BIC に基づく話者クラスタリング

従来，BIC に基づく話者クラスタリング手法が主に用いられている．本手法は，各セグメント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方法である[11]．

この方法では，二つの発話が同一話者のものであるかどうかを，同一話者であるという仮説における BIC 値と，同一話者でないという仮説における BIC 値との差分に基づいて判定する．二つの発話を同一話者であるとしてマージした場合の BIC 値を式 (1)，同一話者でないとして二つの発話を分割した場合の BIC 値を式 (2) により求め，これらの BIC 値の差分を式 (3) により求める．

$$BIC^0 = \frac{N_1+N_2}{2}\log|\Sigma_0| + \frac{\alpha}{2}(d+\frac{d(d+1)}{2})\log(N_1+N_2) \quad (1)$$

$$BIC^{12} = \frac{N_1}{2}\log|\Sigma_1| + \frac{N_2}{2}\log|\Sigma_2| + \alpha(d+\frac{d(d+1)}{2})\log(N_1+N_2) \quad (2)$$

$$\begin{aligned}&\Delta BIC\\ &= BIC^0 - BIC^{12}\\ &= \frac{N_1+N_2}{2}\log|\Sigma_0| - \frac{N_1}{2}\log|\Sigma_1| - \frac{N_2}{2}\log|\Sigma_2|\\ &-\alpha\frac{1}{2}(d+\frac{d(d+1)}{2})\log(N_1+N_2)\end{aligned} \quad (3)$$

ここで，Σ_0 は二つの発話をマージしたときの共分散行列，Σ_1 は一つ目の発話の共分散行列，Σ_2 は二つ目の発話の共分散行列を表している．ここでは，共分散行列として対角成分以外の要素も用いている．また，N_i は各発話のデータサイズ (フレーム数)，d は特徴ベクトルの次元数を表している．重み係数 α は実験的に設定する必要がある．

式 (3) の ΔBIC_{var} の値が，正であれば二つの発話はマージされる．この過程を（時間的に連続しないセグメント間も含めて）繰り返すことにより，クラスタリングが行われる．もし，すべての発話間の ΔBIC_{var} の値が負になれば，どの発話もマージすべきでないとして，クラスタリング処理を終了する．

3. GMM に基づく話者クラスタリング

ここでは，GMM に基づく話者クラスタリングにおいて，発話ごとに EM アルゴリズムにより直接 GMM を学習する手法と，UBM をベースとして MAP 推定により発話ごとに GMM を学習する手法について述べる．

3.1 GMM 間の尤度比に基づくクラスタリング

本手法は，発話ごとに EM アルゴリズムにより直接 GMM を学習して，GMM 間の距離を Cross Likelihood Ratio (CLR)[12] に基づいて計算し，クラスタリングを行う．

以下に本手法による話者クラスタリングの処理の流れを示す．

(1) 初期学習：各発話に対して EM アルゴリズムにより直接 GMM を話者モデルとして学習する．このとき，各発話を一つのクラスタとする．

(2) 初期の距離計算：初期学習で得られた GMM 間の距離を全発話間を対象に次式の CLR に基づいて計算する．

Vol.2011-SLP-85 No.7
2011/2/4

$$d_{ij} = \log \frac{P(X_i|\lambda_i)}{P(X_i|\lambda_j)} + \log \frac{P(X_j|\lambda_j)}{P(X_j|\lambda_i)} \tag{4}$$

$$\log P(X_i|\lambda_j) = \frac{1}{T_i} \sum_{t=1}^{T_i} \log P(x_{it}|\lambda_j)$$

ここで，X_i は i 番目の発話，x_{it} は i 番目の発話の t フレーム目の特徴ベクトル，T_i は i 番目の発話の総フレーム数，λ_i は i 番目の発話から学習された GMM のパラメータを表す．

(3) クラスタリング：各クラスタ間に対して距離が最小になるクラスタを探索し，最小距離が閾値 θ より小さければ，対応するクラスタをマージする．ただし，マージされたときにモデルの再学習は行わない．

(4) クラスタ間の距離計算：すべてのクラスタ間において，各クラスタに属する発話間で式 (4) により CLR を計算してその平均値をクラスタ間の距離とみなす．

最小距離が閾値 θ より大きくなるまで，上記の処理 (3) と (4) を繰り返し実行する．

3.2 UBM を用いた MAP 推定により学習した GMM によるクラスタリング

本手法では，多数話者の音声データから学習した UBM(Universal Background Model) を初期モデルとして，各発話ごとに MAP 推定を行うことで GMM を学習する．

UBM の各混合分布の重み w，平均 μ，分散 σ^2 を各発話データをもとに以下の式により適応する．

$$\hat{w}_i = [\alpha_i n_i/T + (1-\alpha_i)w_i]\gamma \tag{5}$$

$$\hat{\mu}_i = \alpha_i E_i(x) + (1-\alpha_i)\mu \tag{6}$$

$$\hat{\sigma_i}^2 = \alpha_i E_i(x^2) + (1-\alpha_i)({\sigma_i}^2 + {\mu_i}^2) - \hat{\mu_i}^2 \tag{7}$$

$$Pr(i|x_t) = \frac{w_i p_i(x_t)}{\sum_{j=1}^{M} w_j p_j(x_t)} \tag{8}$$

$$n_i = \sum_{t=1}^{T} Pr(i|x_t) \tag{9}$$

$$E_i(x) = \frac{1}{n_i} \sum_{t=1}^{T} Pr(i|x_t) x_t \tag{10}$$

$$\alpha_i = \frac{n_i}{(n_i + r)} \tag{11}$$

ここで，x_t は各発話の特徴ベクトル，T は各発話の総フレーム数，M が UBM の混合分布数，γ は混合分布の重みの総和を制御する係数，α は適応データの割合を制御する係数である．

本研究では，各発話長が比較的短いため，重みと分散は適応せず平均のみを対象に MAP 推定により GMM の学習を行った．

これにより得られた GMM をもとに，3.1 節で述べたクラスタリング手法を適用する．UBM-MAP による手法においては，発話が比較的短いため，もとの UBM の分布から適応により平均が移動する分布の割合が少ない．したがって，すべての混合分布の出力確率を合計すると，モデル間の距離の差が小さくなってしまう．そこで，各混合分布の出力確率の大きい上位 5 個の分布に対しての尤度を求め，式 (4) の CLR を計算した．これによりモデル間の距離を強調したクラスタリングを行うことができる．

4. NMF に基づく話者クラスタリング

NMF は，n 行 m 列の非負値行列 V を，より要素数が少ない n 行 r 列の非負値行列 W と r 行 m 列の非負値行列に分解する手法であり，観測データに対して情報源がどれくらい混ぜ合わされたものであるかを推定することができる．非負値を対象としているため，確率や距離などの値を処理するのに適していると考えられる．

$$V \approx WH \tag{12}$$

ここで，行列 V は観測データ，行列 W は基底，行列 H は各基底における係数を表している．この行列 V から行列 W と H を求める際は，以下に示すカルバック・ライブラー情報量に基づいた手法を用いた．

$$D(V||WH) = \Sigma_{ij}(V_{ij} log \frac{V_{ij}}{(WH)_{ij}} - V_{ij} + (WH)_{ij}) \tag{13}$$

また，このカルバック・ライブラー情報量の目的関数に対する更新ルールは，以下のようになる．これらの式に基づいて，指定された回数分パラメータの更新を行う．

$$W_{ia} \leftarrow W_{ia} \frac{\Sigma_\mu H_{a\mu} V_{i\mu}/(WH)_{i\mu}}{\Sigma_\nu H_{a\nu}} \tag{14}$$

$$H_{a\mu} \leftarrow H_{a\mu} \frac{\Sigma_i W_{ia} V_{i\mu}/(WH)_{i\mu}}{\Sigma_k W_{ka}} \tag{15}$$

本研究では，3.1 節で述べた発話ごとに EM アルゴリズムにより直接 GMM を学習する手法をもとに，全発話間の CLR の値を要素にもつ行列を V とする．ここで，CLR は距離であるので値が小さいほうが特徴が近いと判断されるので，NMF によるクラスタリングに適用するために，CLR の値の逆数に要素を変換して処理を行った．

この行列 V を分解して得られた行列 H は，行が r 個の基底を表しており，列が各観測データに対してどれくらいの重みで結合されるかを表している．そこで，行列 H の列は各発話を表しているので，列ごとにどの基底に対する重みが最も大きいかを求め，最も大きい重みをもつ基底が同じ発話同士は同じ性質をもつと考え，クラスタリングを行う．

このように発話間の距離を一度計算しておけば，それらの値を要素とした行列を分解するだけでクラスタリングを行うことができるため，非常に高速に処理することが可能となる．また，全発話対の距離を総合的に捉えることができ，高精度な話者クラスタリングを実現することができる．

5. 評価実験

本章では，従来手法として BIC に基づくクラスタリング，GMM 間の CLR に基づくクラスタリング，UBM-MAP により学習した GMM に基づくクラスタリング，提案手法である NMF に基づくクラスタリングによる話者分類実験を行う．

5.1 実験条件

本研究では，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 Japanese：以下 CSJ と省略する）に含まれる講演音声を用いて，話者分類実験を行った．

各講演音声から発話区間を抽出し，得られた発話をランダムに並びかえることで，評価用データを作成した．評価データの構成を表 1 に示す．4 つのデータセットからなり，各データセットの話者数はすべて 6 名である．各データセットの発話は 10〜20 秒の音声データで，話者ごとに発話数は均一ではなく偏りがあり，話者ごとの発話数として最も少ない場合は 9，最も多い場合は 41 である．また，各データセットの総発話時間はおよそ 30 分である．

表 1 評価データの話者数と発話数

	A	B	C	D
話者数	6	6	6	6
総発話数	116	125	138	117
平均発話数	19	21	23	19
最少発話数	15	9	12	15
最多発話数	28	30	41	27

UBM の学習データにも CSJ を用いており，評価データと異なる 500 名の話者の講演音声から発話区間を抽出し，各話者ごとに約 60 秒の音声データを集めて 256 混合分布の GMM を作成した．

本実験で用いた音声データは，サンプリング周波数 16kHz でフレーム長 25ms のハミング窓，フレーム周期 10ms で音響分析を行い，フレーム毎に 12 次 MFCC の特徴量を求めている．

話者分類の性能は，SDER(Speaker Diarization Error Rate) により評価を行った．SDER は以下の式により算出した．

$$SDER = \frac{U_{miss} + U_{error}}{U_{ref}} \tag{16}$$

ここで，U_{miss} は得られたクラスタ数が正しい話者数よりも多い場合に，正しい話者に対応付けられなかった発話区間の時間長，U_{error} は誤った話者に対応付けられた発話区間の時間長，U_{ref} は正しい話者の総発話時間長を表している．SDER を算出する際は，クラスタリング結果と正しい話者ラベルの対応関係を SDER が最小になるように求めた．

5.2 実験結果と考察

各クラスタリング手法による話者分類結果を表 2 に示す．表中において，各データセットの上段の値は SDER，下段の値はクラスタリングにより得られた話者数を表している．*BIC*

は BIC に基づくクラスタリング結果，$UBM-MAP$ は UBM をもとに MAP 推定で学習した GMM に基づくクラスタリング結果，GMM は EM アルゴリズムにより学習した GMM で混合分布数が 4 のときの結果，NMF は提案手法である非負値行列因子分解に基づくクラスタリング結果を示している．なお，3 つの従来手法においては，全評価データセットに対して正しい話者数が得られるように共通に閾値を設定したときの結果を示している．また，提案手法においては，全評価データセットに対して正しい話者数が得られるように共通に次元圧縮を行ったときの結果を示している．

表 2 各評価データにおける話者分類結果

	BIC	UBM-MAP	GMM	NMF
A	24.3 (7)	32.9 (6)	18.9 (6)	6.9 (6)
B	34.8 (8)	42.2 (5)	18.3 (7)	17.6 (6)
C	34.3 (8)	74.6 (15)	19.9 (11)	9.2 (6)
D	23.0 (6)	46.5 (7)	42.6 (6)	16.7 (6)

従来手法である BIC に基づく手法では，どの評価データに対しても比較的安定した話者分類精度ならびに話者数を得ることができた．

UBM をもとに MAP 推定で学習した GMM に基づく手法では，全体的に話者分類精度が低く，クラスタリングにより得られた話者数の変動が大きかった．これは，本実験で用いた評価データに含まれる各発話は 10〜20 秒と比較的短かったため，MAP 推定による適応が十分に行えていないためではないかと考えられる．

EM アルゴリズムにより学習した GMM に基づく手法では，これまでの従来手法に比べて最も高い分類精度を得ることができ，UBM-MAP による手法と比べると話者数も正しく抽出することができている．今回，評価データに含まれる各発話が比較的短かったため，少ない混合分布数で GMM を学習することで，発話間の識別がより頑健に行えたのではないかと考えられる．また，BIC に基づく手法よりも GMM に基づく手法のほうが全体的に話者分類精度が高かったことから，単一分布よりも混合分布で発話をモデル化した効果が得られていると考えられる．

提案手法である NMF に基づく手法では，従来のどの手法と比べてもすべての評価データセットにおいて最も高い話者分類精度を得ることができた．

図 1 に各クラスタリング手法により得られた全評価データセットに対する平均 SDER の値を示す．

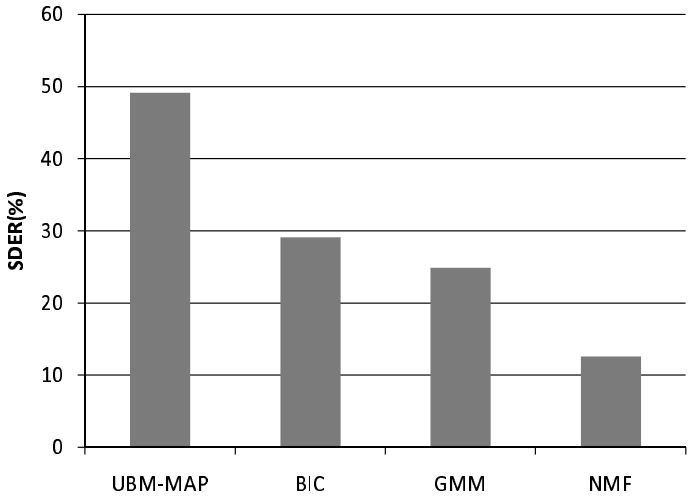

図 1 各クラスタリング手法による話者分類結果

この結果から，UBM をもとに MAP 推定で学習した GMM による手法では 49.1，BIC に基づく手法では 29.1，EM アルゴリズムにより学習した GMM による手法では 24.9，提案手法である NMF に基づく手法では 12.6 の平均 SDER が得られた．

従来の UBM-MAP，BIC，GMM に基づく手法では初期段階で全発話間の距離を求めて，その後距離が最も小さいものを探索してマージしていくという階層的なクラスタリングを行う必要があり，発話数が多いほど処理時間がかかってしまう．それに対して，提案手法である NMF に基づくクラスタリング手法では，初期段階で全発話間の距離を一度求めてしまえば，それらを要素とした行列を分解するだけでよいため，従来に比べて非常に高速なクラスタリングを実現することができる．また，従来手法では単一の発話対の距離に基づいてクラスタリングが行われるのに対して，提案手法では全発話対の距離を総合的に捉えることが

でき，高精度な話者分類を実現することができる．以上の結果から，NMF に基づくクラスタリングが話者分類において有効であることがわかった．

提案手法では，NMF による行列分解を行う際に次元圧縮を行うが，そのときに次元数を設定する必要がある．今回は，より正しい話者数が得られたときの話者分類性能を比較するために，次元数を話者数に設定したときの結果を求めた．今後は，最適な次元数を設定する手法について検討を行っていく必要がある．

6. おわりに

本研究では，教師なし話者分類を対象とした非負値行列因子分解に基づく話者クラスタリング手法を提案した．本手法の有効性を示すために，従来手法として BIC に基づくクラスタリング，EM アルゴリズムにより学習した GMM ならびに UBM をベースとした MAP 推定により学習した GMM 間の尤度比に基づくクラスタリングとの比較実験を行った．その結果，提案手法はすべての評価データに対して従来手法よりも高い話者分類精度を得ることができた．本手法は，発話間の距離を一度計算しておけば行列分解の処理のみでクラスタリングが可能であり，全発話対の距離を総合的に捉えることができ，高速でかつ高精度な話者分類を実現することができた．

今後は，より多くのデータでの評価やその他のデータベースやタスクでの評価について検討を行う．また，NMF における最適な次元数の設定などについて検討を行う予定である．

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) 中川 聖一，森 一将: 発話間の VQ ひずみを用いた話者交替識別と話者クラスタリング, 電子情報通信学会論文誌. D-II, 情報・システム, II-パターン処理 J85-D-II(11), 1645-1655 (2002).
3) 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- No.2，pp.495-503 (2004).
4) M. Nishida and T. Kawahara: Speaker Model Selection Based on the Bayesian Information Criterion Applied to Unsupervised Speaker Indexing, IEEE Transactions on Speech and Audio Processing, Vol.13, No.4, pp.583-592 (2005).
5) F. Castaldo, D. Colibro, E. Dalmasso, P. Laface, and C. Vair: Stream-Based Speaker Segmentation Using Speaker Factors and Eigenvoices, Proc. ICASSP, pp.4133-4136 (2008).
6) D. Vijayasenan, F. Valente, and H. Bourlard: KL Realignment for Speaker Diarization with Multiple Feature Streams, Proc. INTERSPEECH, pp.1059-1062 (2009).
7) 奥 貴裕, 佐藤 庄衛, 小林 彰夫, 本間 真一, 今井 亨: 音素情報を利用した BIC に基づくオンライン話者識別，情報処理学会研究報告，Vol.2010-SLP-80, No.9, pp.1-6 (2010).
8) 石塚 健太郎，荒木 章子，大塚 和弘，藤本 雅清，中谷 智広: 音響情報と映像情報の統合による多人数会話における話者決定技術，情報処理学会研究報告，Vol.2008-SLP-74, No.5, pp.25-30 (2008).
9) D. D. Lee and H. S. Seung: Algorithms for Non-negative Matrix Factorization, Proc. NIPS, pp.556-562 (2000).
10) 西田 昌史，細川 光政，山本 誠一: NMF に基づくクラスタリングを適用した Anchor Model による話者認識, 情報処理学会研究報告, Vol.2010-SLP-84, No.13, pp.1-6 (2010).
11) 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).
12) D.A. Reynolds, E. Singer, B.A. Carlson, G.C. O'Leary, J.J. McLaughlin, and M.A.Zissman: Blind Clustering of Speech Utterances based on Speaker and Language Characteristics, Proc.ICSLP, pp.3193-3196 (1998).
13) D.A.Reynolds, T.F.Quatieri, and R.B.Dunn: Speaker Verification Using Adapted Gaussian Mixture Models, Digital Signal Processing, Vol.10, pp.19-41 (2000).