Vol.2012-SLP-92 No.8
2012/7/20

i-vector に基づく発話類似度を用いた非負値行列分解と話者クラスタリングへの適用

福地 佑介[1]　俵 直弘[1]　小川 哲司[1]　小林 哲則[1]

概要：高精度な話者表現とクラスタリングアルゴリズムを統合した新たな話者クラスタリング手法を提案する．従来用いられる話者クラスタリング手法では，データ量が多くなると正確なクラスタリングが困難になるという問題があった．そのような条件下において正確な話者クラスタリングを実現するためには，音響変動に対して頑健なモデルにより話者を表現し，このモデルを用いて各発話を効率的にクラスタリングする手法が必要となる．そこで提案手法では，話者照合の分野で高い精度を達成している i-vector を話者の表現として用い，クラスタリング手法として非負値行列分解に基づいた効率的なクラスタリング手法を導入した．本手法の有効性を示すために，CSJ データを用いた話者クラスタリング実験を行い，従来手法と比較して，提案手法が発話データ量の変化に対し頑健に話者クラスタリングが行えることを確認した．

キーワード：i-vector，非負値行列分解，話者クラスタリング

Speaker clustering based on non-negative matrix factorization using i-vector-based speaker similarity

Fukuchi Yusuke[1]　Tawara Naohiro[1]　Ogawa Tetsuji[1]　Kobayashi Tetsunori[1]

Abstract: We have developed a novel speaker clustering method by integrating highly accurate speaker representation and a clustering algorithm. The conventional method caused significant degradation in clustering accuracy when the number of utterances increased. High-accuracy speaker representation and high-performance clustering method are required to realize robust speaker clustering system against such a condition. For this purpose, we used i-vectors for the speaker representation, which contributes to the realization of high-accuracy speaker verification systems, and efficient non-negative matrix factorization for the clustering algorithm. Experimental results show that the proposed method outperforms the conventional methods, irrespective of the amount of data.

Keywords: speaker clustering, i-vector, non-negative matrix factorization

1. はじめに

近年，Web を中心に大量の音声データが利用可能になった．話者クラスタリングはこれらの音声データを活用するための重要な要素技術である．例えば，音声データに対して付与された話者クラスタ情報は，特定話者の音声の検索や，音響モデルの話者適応学習への利用 [1] が期待される．このとき，話者クラスタリングを行うにあたり，話者ごとに発話数が大きく異なる，同一話者の音声であっても，発話スタイルや周辺の音環境が変動する，といった条件下においても，頑健な手法が求められる．

話者クラスタリングは，話者の類似性を記述するための話者の表現法と，クラスタリングアルゴリズムの観点から検討することができる．話者を表現するためのモデルとしては，ガウス分布や混合ガウス分布 (Gaussian mixture model; GMM) が広く用いられてきた [2], [3], [4], [5]．さらに，近年，チャネルの影響に頑健な話者表現として，因子分析に基づく手法 [6] が提案され，話者照合において有効性が

[1] 早稲田大学
Waseda University, Shinjuku, Tokyo 164–0042, Japan

© 2012 Information Processing Society of Japan

実証されている．一方，クラスタリングのアルゴリズムとしては，ベイズ情報量基準 (Bayesian information criteria; BIC) に基づく凝集的階層型クラスタリング (agglomerative hierarchical clustering; AHC) や k-means 法による非階層的なクラスタリング，行列分解を用いたクラスタリングなどが提案されている．

ガウス分布を用いて話者を表現する話者クラスタリング手法では，BIC に基づいて凝集的階層型クラスタリング (BIC-AHC) を行う手法 [4] が最も広く用いられている．しかし，この手法では，単峰のガウス分布を用いて話者を表現するため，各話者の発話データの数が増加するに従い分散が不当に広がり，クラスタリング精度が悪化するという問題があった．一方，行列分解に基づく手法として，非負値行列分解 (non-negative matrix factorization; NMF) に基づく話者クラスタリング手法 [5] が提案されている．この手法では，混合ガウス分布 (Gaussian mixture model: GMM) を話者モデルとして導入し，各発話間の類似度として，各発話の GMM から計算される cross likelihood ratio(CLR) を定義することにより，類似度行列の非負値行列分解による話者クラスタリングを実現している．この手法では，単峰のガウス分布に比べて表現能力の高い GMM を話者表現のモデルとして用いており，従来の BIC に基づく凝集的クラスタリング手法よりも高い精度でクラスタリング精度を達成している．しかし一方で，発話データが少ない場合，各発話の GMM を学習する際に過学習の問題が発生し，これにより，話者モデルの精度が低下する可能性がある．混合ガウス分布以外の話者表現を用いるクラスタリング手法として，発話ごとに算出した i-vector 間のコサイン類似度を尺度とした k-means 法に基づく話者クラスタリングが提案されており，会話中の音響変動に対して頑健な話者ダイアライゼーションが実現されている [6].

本研究では，以上の従来研究に着想を得て，環境変化に頑健な話者表現である i-vector と高精度な非負値行列分解によるクラスタリングを組み合わせた新たな話者クラスタリング手法を提案する．本手法の有効性を確認するために，比較手法として，BIC に基づく凝集的クラスタリング [4]，i-vector を話者表現として用いた k-means 法に基づく手法 [7]，GMM による話者表現を利用した NMF による話者クラスタリング [5] と比較を行い，提案手法が従来手法に対してデータ量の増加に対する頑健性の点で有効であることを示す．

以降，2 では本手法で用いる話者表現と発話類似度に焦点を当てて述べ，3 でクラスタリングのアルゴリズムに焦点を当てて述べる．4 では話者クラスタリングについて既存の手法と提案手法を比較する．5 では CSJ コーパスを用いた話者クラスタリング実験を通して提案手法の有効性を評価する．最後に 6 でまとめを述べる．

2. 話者のモデル化

本章では，発話データが与えられたときに，発話データの話者表現と，それを用いた話者類似度の計算手法について述べる．2.1 では，GMM に基づいて話者性を表現する方法と，その際に類似度として用いる Cross likelihood ratio (CLR) について述べる．2.2 では，i-vector に基づいて話者性を表現する方法と，その際に類似度として用いるコサイン類似度について述べる．

2.1 混合ガウス分布に基づく話者モデル

発話ごとに GMM を最尤学習により構築し，話者モデルとして用いる．与えられた発話を $\boldsymbol{X}_i$ とすると，各発話を表現する GMM は次式で表される．

$$p(\boldsymbol{X}_i|\lambda_i) = \sum_{k=1}^{K} m_{i,k}\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_{i,k}, \boldsymbol{\Sigma}_{i,k}), (\sum_{k=1}^{K} m_{i,k} = 1) \quad (1)$$

ここで，$\mathcal{N}(\cdot)$ は以下の式で定義されるガウス分布である．

$$\mathcal{N}(\boldsymbol{X}_i|\boldsymbol{\mu}_{i,k}, \boldsymbol{\Sigma}_{i,k}) = \frac{1}{(2\pi)^{\frac{D}{2}}\ |\boldsymbol{\Sigma}_{i,k}|^{\frac{1}{2}}} \cdot \exp\big(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_{i,k})^T\boldsymbol{\Sigma}_{i,k}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_{i,k})\big). \quad (2)$$

このとき，$\lambda_i = \{m_{i,k}, \boldsymbol{\mu}_{i,k}, \boldsymbol{\Sigma}_{i,k}\}_{k=1}^{K}$ は i 番目の発話 $\boldsymbol{X}_i$ を用いて推定した GMM のパラメータであり，$\boldsymbol{\mu}_{i,k}$，$\boldsymbol{\Sigma}_{i,k}$，$m_{i,k}$ はそれぞれ第 k 混合における平均ベクトル，共分散行列，重みである．ここでは，類似度として 2 つの GMM のクロスエントロピーに相当する CLR を利用した [5]．i 番目の発話と j 番目の発話の CLR は次式で定義される．

$$\mathcal{D}_{ij}^{\mathrm{CLR}} = \log\frac{p(\boldsymbol{X}_i|\lambda_i)}{p(\boldsymbol{X}_i|\lambda_j)} + \log\frac{p(\boldsymbol{X}_j|\lambda_j)}{p(\boldsymbol{X}_j|\lambda_i)} \quad (3)$$

$p(\boldsymbol{X}_i|\lambda_j)$ はパラメータとして，$\lambda_j = \{m_{j,k}, \boldsymbol{\mu}_{j,k}, \boldsymbol{\Sigma}_{j,k}\}_{k=1}^{K}$ を与えたときの GMM における発話 X_i の尤度であり，次式で表される．

$$p(\boldsymbol{X}_i|\lambda_j) = \prod_{t=1}^{T_i} p(\boldsymbol{x}_{i,t}|\lambda_j) \quad (4)$$

2.2 i-vector に基づく話者モデル

2.2.1 i-vector

発話 u について，話者とチャネル依存の GMM スーパーベクトル $\boldsymbol{M}_u$ は低次元の全変動 (Total variability; TV) 空間に写像する因子分析として以下のように定義される．

$$\boldsymbol{M}_u = \boldsymbol{m} + \mathbf{T}\cdot\boldsymbol{w}_u \quad (5)$$

m は universal background model (UBM) のような話者およびチャネル非依存の GMM スーパーベクトルである．$\mathbf{T}$ は話者とチャネル情報を同時にモデル化した低次元空間への射影行列であり，TV 空間を張る基底ベクトルから構成

© 2012 Information Processing Society of Japan

される．この $\mathbf{T}$ は全ての発話を別々の話者が発話したものとみなして，Eigen voice とほぼ同じ学習則により求めることが出来る [6]．このとき，$\boldsymbol{w}_u$ が発話 u の i-vector である．発話 u に対する i-vector は，TV 空間への射影行列 $\mathbf{T}$ と発話 u のデータから算出される統計データに基づいて以下により計算される．

$$\boldsymbol{w}_u = (\mathbf{I} + \mathbf{T}^t\boldsymbol{\Sigma}\mathbf{N}(u)\mathbf{T})^{-1}\mathbf{T}^t\boldsymbol{\Sigma}^{-1}\mathbf{F}(u) \tag{6}$$

$\mathbf{N}(u)$，$\mathbf{F}(u)$ は，次式で表される 0 次，1 次統計量を用いて表される．

$$N_c = \sum_{t=1}^{L} P(c|\boldsymbol{x}_t, \boldsymbol{\Omega}) \tag{7}$$

$$\boldsymbol{F}_c = \sum_{t=1}^{L} P(c|\boldsymbol{x}_t, \boldsymbol{\Omega}) \cdot (\boldsymbol{x}_c - \boldsymbol{m}_c) \tag{8}$$

ここで，混合数 C の UBM を $\boldsymbol{\Omega}$，特徴量の次元数を F で表す．$c = 1, \cdots, C$ は UBM の混合分布のインデックスであり，$\boldsymbol{m}_c$ は要素 c の平均ベクトルを表す．$\mathbf{N}(u)$ は対角ブロックが $N_c\mathbf{I}(\mathbf{I} \in \mathbb{R}^{F\times F})$ となる $CF \times CF$ 次元の対角行列である．$\mathbf{F}(u)$ は全ての $\boldsymbol{F}_c$ を結合することで得られる $CF \times 1$ 次元のスーパーベクトルである．$\boldsymbol{\Sigma}$ は TV 行列で表現できない残留変動成分をモデル化する $CF \times CF$ の対角行列で，因子分析により推定される [8]．以上を用いて推定される i-vector は話者の情報の他にチャネルの情報も含んでいるので，この影響を除去するために，フィッシャー判別分析 (Fisher discriminant analysis; FDA) およびクラス内分散正規化 (within-class covariance normalization; WCCN) を用いて話者の情報のみを抽出する．

2.2.2 チャネル情報の補正

i-vector に含まれる不要なチャネル情報を低減するために，FDA 及び WCCN を適用する [6], [10]．FDA は，同一クラスの分散を小さく，異なるクラスの分散を大きくするような座標変換 $\boldsymbol{w} \to \boldsymbol{A}^t\boldsymbol{w}$，$(\boldsymbol{w} \in \mathbb{R}^d, \boldsymbol{A} \in \mathbb{R}^{d\times d'}, d' < d)$ を行い，次元圧縮を行う手法である．これは教師あり次元削減として良く用いられる手法であり，話者を教師ラベルとし次元圧縮を行うと，チャネルの変動を除去する効果が得られる．変換行列 $\boldsymbol{A}$ は $\boldsymbol{S}_b$ をクラス間分散行列，$\boldsymbol{S}_w$ をクラス内分散行列としたときの固有値問題

$$\boldsymbol{S}_b\boldsymbol{v} = \lambda\boldsymbol{S}_w\boldsymbol{v} \tag{9}$$

の解として得られる固有値のうち，大きいものに対応する固有ベクトルから成る．

一方，WCCN では，クラス内の分散を正規化する座標変換 $\boldsymbol{w} \to \boldsymbol{B}^t\boldsymbol{w}$，$(\boldsymbol{w} \in \mathbb{R}^d, \boldsymbol{B}^{d\times d})$ を求める．ここで，クラスは話者であり，データは発話音声であるから，WCCN は各話者における音声データの分散を正規化する．これも，同一話者クラス内での音響変動 (チャネル変動) を補正する効果が期待できる．変換行列 $\boldsymbol{B}$ は話者 s の i 番目のデータを $\boldsymbol{w}_i^s$，話者 s のデータの平均を $\boldsymbol{w}_s$，話者数を S とすると，クラス内分散行列

$$W = \frac{1}{S}\sum_{s=1}^{S}\frac{1}{n_s}\sum_{i=1}^{n_s}(w_i^s - w_s)(w_i^s - w_s)^t \tag{10}$$

の逆行列 W^{-1} に対してコレスキー分解 $\boldsymbol{W}^{-1} = \boldsymbol{B}\boldsymbol{B}^t$ を行うことで得られる．

2.2.3 スコアリング

発話 i と発話 j 間の類似度として，各発話から求めた i-vector のコサイン類似度を計算する

$$\mathcal{D}_{i,j}^{\cos} = \frac{\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{z}_j}{\|\boldsymbol{z}_i\| \cdot \|\boldsymbol{z}_j\|} \tag{11}$$

ここで，$\boldsymbol{z}_u$ は i-vector $\boldsymbol{w}_u$ に対して FDA と WCCN を施し，チャネル情報を補正した u 番目の発話ベクトルである．

3. クラスタリングアルゴリズム

本章では，話者クラスタリングに適用するクラスタリングのアルゴリズムについて述べる．ここでは，BIC 基準に基づく凝集的階層型クラスタリングと NMF によるクラスタリングについて述べる．

3.1 BIC 基準に基づくクラスタリング

BIC に基づく話者クラスタリングでは，発話を単一ガウス分布の分散を用いて実現する．ここで，$BIC_{i,j}^0$ と $BIC_{i,j}$ を発話 $\boldsymbol{X}_i$ と発話 $\boldsymbol{X}_j$ が同一の話者による発話と仮定した場合と，異なる話者による発話と仮定した場合のモデル全体の BIC 値とすると，それぞれ以下のように書ける．

$$\begin{aligned} BIC_{i,j}^0 &= \frac{N_i + N_j}{2}\log|\boldsymbol{\Sigma}_0| \\ &\quad + \frac{\alpha}{2}(d + \frac{d(d+1)}{2})\log(N_i + N_j) \end{aligned} \tag{12}$$

$$\begin{aligned} BIC_{i,j} &= \frac{N_i}{2}\log|\boldsymbol{\Sigma}_i| + \frac{N_j}{2}\log|\boldsymbol{\Sigma}_j| \\ &\quad + \frac{\alpha}{2}(d + \frac{d(d+1)}{2})\log(N_i + N_j). \end{aligned} \tag{13}$$

$\boldsymbol{\Sigma}_0$ は 2 つの発話が同一話者によるものとしてマージされたときの分散を表し，$\boldsymbol{\Sigma}_i$ と $\boldsymbol{\Sigma}_j$ はそれぞれの発話における分散を表す．N_i と N_j はそれぞれの発話のフレーム数，d は特徴量の次元数を表す．α はモデルの複雑さにペナルティを与える重み係数で，実験により決定する．i 番目の発話と j 番目の発話が同一話者によるものかを以下の基準により決定する．

$$\Delta BIC_{i,j} = BIC_{i,j}^0 - BIC_{i,j} \tag{14}$$

全ての発話の組に対して $\Delta BIC_{i,j}$ を計算し，$\Delta BIC_{i,j}$ が正となった発話のペアの中から値が最も大きいものを同一話者による発話と判断しマージする．このとき，同一話者

© 2012 Information Processing Society of Japan

によるものであると判断された発話の組については，他の全ての発話の組に対して $\Delta BIC_{i,j}$ を計算し直す必要がある．以上の処理を，すべての発話の組に対して式 (14) が負になるまで繰り返す．

3.2 非負値行列分解によるクラスタリング

発話数 U の全ての発話の組み合わせに対してその類似度を行列 $\mathbf{V} \in \mathbb{R}^{U \times U}$ として定義する．非負値行列分解 (Non-negative Matrix Factorization; NMF) に基づく話者クラスタリングは，この類似度行列 $\mathbf{V}$ を次式のように基底の行列 $\mathbf{W} \in \mathbb{R}^{U \times S}$ と係数の行列 $\mathbf{H} \in \mathbb{R}^{S \times U}$ に分解することにより，クラスタリングを実現する．

$$\mathbf{V} \simeq \mathbf{WH} \tag{15}$$

ここで U は発話数，S は話者数を表す．このとき，基底行列 $\mathbf{W}$ の各列が話者の特徴に対応する．また，係数行列 $\mathbf{H}$ の u 列目のベクトルは，この発話に含まれる，話者 s の成分の大きさに対応する．そのため，u 番目の発話の話者は，次式を用いることにより推定することができる．

$$\hat{s}_u = \arg\max_s H_{s,u} \tag{16}$$

全ての発話 $u = 1, \cdots, U$ に対して式 (16) を評価することにより，各発話の話者 $\hat{s}_u$ を推定することができる．2.1 で述べたように，各発話の話者を GMM で表現した場合，その類似度行列は CLR から計算ことができ，i-vector で話者を表現した場合は，コサイン類似度を用いて計算することが出来る．$\mathbf{W}$ ， $\mathbf{H}$ 求めるために，真の分布を $\mathbf{V}$ として，以下の式でカルバックライブラー情報量 [11] を定義する．

$$\mathcal{D}(\mathbf{V}||\mathbf{WH}) = \sum_{i,j} \left(V_{i,j} \log \frac{V_{i,j}}{(\mathbf{WH})_{i,j}} - V_{ij} + (\mathbf{WH})_{i,j} \right) \tag{17}$$

$\mathbf{W}$ と $\mathbf{H}$ は $\mathcal{D}(\mathbf{V}||\mathbf{WH})$ を最小にするものを求める．これは，以下の更新式に基づいて推定することができる．

$$W_{i,a} \leftarrow W_{i,a} \frac{\sum_u H_{a,u} V_{i,u} / (\mathbf{WH})_{i,u}}{\sum_u H_{a,u}} \tag{18}$$

$$H_{a,j} \leftarrow H_{a,j} \frac{\sum_s H_{s,a} V_{s,j} / (\mathbf{WH})_{s,j}}{\sum_s H_{s,a}} \tag{19}$$

ここで， $V_{i,j}$，$W_{i,j}$，$H_{i,j}$ はそれぞれ行列 $\mathbf{W}$ ， $\mathbf{W}$ ， $\mathbf{H}$ の i 行 j 列目の要素を表す．行列分解による手法では，従来の凝集的な手法と異なり，類似度行列を一度計算すると，類似度の計算をし直すことなくクラスタリングを行うことができる．

4. 話者クラスタリング手法

本章では，本稿で比較する話者クラスタリング手法について述べる．表 1 に本稿で比較する話者クラスタリング手法を示す．BIC [4] では，話者表現に単一ガウス分布を利用し，クラスタリングに 3.1 を用いた．GMM-NMF は，話者表現として 2.1 で述べた GMM を用いて CLR により類似度行列を定義し，クラスタリングには 3.2 で述べた NMF を用いた．この手法は BIC に基づく話者クラスタリングよりも高精度であることが示されている [5]．このとき，CLR は小さいほど発話が似ていることを表す距離尺度なので，NMF で利用する際にはその逆数を用いて類似度行列として使用した．IV-kM は，i-vector 空間において k-means クラスタリングを適用することで k-means クラスタリングのように単純なアルゴリズムでも高い性能を達成する [7]．提案手法では高性能な話者モデル表現である i-vector と高精度なクラスタリングアルゴリズムである NMF を統合した．発話間の類似度行列は i-vector 間のコサイン類似度を計算することで求めた．ただし，コサイン類似度は非負値ではないため，式 (20) に示す変換関数により非負値に変換した．

表 1　話者クラスタリング手法．

Method	Speaker representation	Clustering algorithm
BIC [4]	single Gaussian	BIC-AHC
GMM-NMF [5]	GMM	NMF
IV-kM [7]	i-vector	k-means
IV-NMF (proposed)	i-vector	NMF

$$f(v) = \begin{cases} 0 & , \quad v < 0 \\ v^3 & , \quad \text{otherwise} \end{cases} \tag{20}$$

この変換により，類似度の大きい発話の影響は大きくなり，類似度の小さい発話の影響は小さくなるため，クラスタリングの精度が向上することが期待できる．この関数が良い性能を示すことを予備実験により確認した．

5. 話者クラスタリング実験

表 1 に示した各手法の話者クラスタリング精度を評価した．その際，発話数，話者数が異なる複数の評価セットを用意し，各手法の性能をそれぞれの条件下で評価した．

5.1 実験環境

5.1.1 音声データ

日本語話し言葉コーパス (CSJ) [12] の学会講演から男性 50 人，女性 50 人の合計 456 の講演データを無作為に選択し，以下の手順により異なる話者数と発話数を含む 4 種類の評価セットを作成した．まず，講演データを 500 ms 以上の無音区間で区切り，そのうち発話長が 5 s 以上 10 s 以下の発話を抽出した．抽出した発話の中から 5 話者または 10 話者をランダムに選び，各話者の発話の中からそれぞれ 10 あるいは 100 発話を選択した．全ての場合について，話者と発話が異なる 5 セットの評価セットを用意し，

© 2012 Information Processing Society of Japan

結果の平均を用いて評価を行った．

5.1.2 特徴抽出と話者モデル

予備実験の結果に基づき，BIC および GMM-NMF では音響特徴量として 12 次元の MFCC を用いた．GMM-NMF で用いた GMM の混合数は 8 で，発話ごとに 混合数 8 の GMM を作成した．CLR が 0 となる場合，発話間類似度の最大値を 1.5 倍した値を類似度として用いた．IV-kM と IV-NMF では， UBM，i-vector の射影行列 $\mathbf{T}$，FDA，WCCN における分散行列の学習データとして男性 183 人 が発話した新聞記事の読み上げ音声計 28,171 発話を利用した．ここで，UBM の作成には MFCC 12 次元と Δ MFCC 12 次元，$\Delta\Delta$ MFCC 12 次元の合計 36 次元の音響特徴量を用い，混合数は 512 とした．i-vector の次元数は 350 次元で，FDA によって 100 次元まで圧縮し，これに WCCN を施した．

5.1.3 クラスタリングパラメータ

本実験では，話者数は事前に与えられると仮定して推定を行わない．つまり，NMF の基底数と，k-means 法のクラスタ数は，正解の話者数 (5 人及び 10 人) を与えた．BIC に関しては，式 (12)，(13) における BIC のチューニングパラメータ α を各評価セットに対して最適化し，最も良い K 値が得られた値 (発話数が 10，100 のとき各々 $\alpha = 6.8$，$\alpha = 13.8$) に設定した．また，NMF に用いる行列 $\mathbf{H}$，$\mathbf{W}$ の初期値は乱数を用いて設定し，行列の更新値が閾値以下になるまで更新を繰り返した．このとき，閾値は予備実験により 0.0001 と決定した．

5.1.4 評価尺度

話者クラスタリング精度の評価には平均話者純度 (average speaker purity; ASP) と平均クラスタ純度 (average cluster purity; ACP)，これらの幾何平均である K 値を評価尺度として用いた [13]．クラスタ純度 p_i は，各クラスタに割り当てられた発話のうち同じ話者に起因するものの割合を表し，話者純度 q_j は，各話者が発話した発話のうち，同じクラスタに割り当てられた発話の割合を示し，それぞれ以下で表される．

$$p_i = \sum_{j=0}^{S} \frac{n_{ij}^2}{n_i^2} \tag{21}$$

$$q_j = \sum_{i=0}^{S} \frac{n_{ij}^2}{n_j^2} \tag{22}$$

ただし，S は話者数，U は総発話数，n_j は話者 j による発話数，n_i はクラスタ i に割り当てられた発話の数，n_ij は話者 j の発話のうちクラスタ i に割り当てられた発話数とする．平均クラスタ純度 V_{ASP}，平均話者純度 V_{ACP} はこれらを用いて以下で表される．

$$V_{\mathrm{ACP}} = \frac{1}{U} \sum_{i=0}^{S} p_i n_i \tag{23}$$

表 2　各発話数，話者数における K 値による話者クラスタリングの性能評価．各条件における最高性能を太字で記す．

	# speakers	# utterances	K value
BIC	5	10	**0.989**
		100	0.610
	10	10	**0.993**
		100	0.592
IV-kM	5	10	0.922
		100	0.914
	10	10	0.920
		100	0.896
GMM-NMF	5	10	0.977
		100	0.931
	10	10	0.940
		100	0.896
IV-NMF	5	10	0.984
		100	**0.987**
	10	10	0.966
		100	**0.944**

$$V_{\mathrm{ASP}} = \frac{1}{U} \sum_{j=0}^{S} q_j n_j \tag{24}$$

K 値はこれらの幾何平均として以下で表される．

$$K = \sqrt{V_{\mathrm{ACP}} \cdot V_{\mathrm{ASP}}} \tag{25}$$

K 値は値が大きいほど精度が高いことを示しており，正解ラベルとクラスタリングの結果が一致するときのみ 1 になる

5.2 実験結果

表 2 に，従来手法である BIC， IV-kM， GMM-NMF と提案手法 (IV-NMF) の話者クラスタリング精度を示す．この結果から，提案手法である IV-NMF は話者数と発話数によらず高い K 値が得られており，高精度に話者クラスタリングが実現できていることが分かる．このとき，IV-NMF は GMM-NMF および IV-kM に対しては条件によらず高い精度を与えた．一方，BIC-AHC に対しては，発話数が 10 のときには BIC-AHC が提案手法を若干上回ったものの，発話数が 100 に増加すると，提案手法の性能が BIC-AHC を上回った．

まず，条件による性能の変化を手法ごとにまとめる．

- **BIC**
  発話数が少ないときの性能は高いが，発話数の増加に伴い性能が著しく低下する．これは，発話数が多くなると単一ガウス分布では発話者の特徴を正確に捉えられないためと考えられる．
- **IV-kM**
  話者数，発話数の増加で性能が大きく低下することは無い．しかし，他の手法に比べて性能が伸びない．
- **GMM-NMF**

© 2012 Information Processing Society of Japan

発話数が少ないときの性能は比較的高く，発話数が増加したときの性能の低下も BIC ほど大きくない．しかし，IV-kM と比較すると性能の低下が大きく，発話数がさらに大きくなると，IV-kM よりも性能が悪化する可能性がある．

- **IV-NMF**
  発話数の変化に対して頑健に高い性能を与える．発話数が少ない場合には BIC の性能には及ばないものの，高い精度でクラスタリングが行えることがわかる．また，発話数が増えたときの性能の低下が，他の 3 手法と比較し小さいため，発話数がさらに増加した場合においても高い性能を維持することが期待できる．一方で，話者数が増加したときに性能が低下する可能性がある．

発話数の変化による影響について考えると IV-kM，IV-NMF のように話者表現として i-vector を用いた場合，BIC，GMM-NMF と比較して性能の低下がほとんど見られないことから，ガウス分布に基づく手法に比べて発話数の増加に頑健であると言える．これは，i-vector の抽出過程でチャネル変動の補償を行っているため，発話数が増えても話者の特徴を捉えることができたためと考えられる．

また，話者数の変化による影響について考えると，NMF に基づくクラスタリングを用いた場合 (GMM-NMF，IV-NMF)，他の 2 手法と比較して話者の増加による性能の低下が大きいと言える．これは，話者数が増えると，話者を表す基底数が増えることで，データの表現能力が十分でなくなったため，性能が低下したと考えられる．

6. まとめ

本稿では，i-vector を用いた話者性の表現と NMF に基づくクラスタリングを統合した新たな話者クラスタリング手法を提案した．話者数と発話数を変化させて話者クラスタリング実験を行い，提案手法の有効性を評価したところ，提案手法はデータ量の変化に頑健な話者クラスタリングを実現できることが分かった．大規模な実環境データに対応する話者クラスタリングに適した手法として期待できる．

参考文献

[1] T.J. Reynolds, *et al.*, “Clustering via the Baysian information criterion with applications in speech recognition,” Proc. ICASSP, vol.2, pp.645-648, May 1998.
[2] D.A. Reynolds, *et al.*, “Blind Clustering of Speech Utterances based on Speaker and Language Characteristics,” Proc. ISCLP, vol.7, pp.3193-3196, Nov. 1998.
[3] D.A. Reynolds, *et al.*, “Speaker Verification Using Adapted Gaussian Mixture Models,” Proc. Digital Signal Processing, vol.10, pp.19-41, Jan. 2000.
[4] S. Chen and P. Gopalakrishan, “Speaker, environment and channel change detection and clustering via the Bayesian information criterion,” Proc. DARPA Broadcast News Transcription and Understanding Workshop, pp.127-132, May 1998.
[5] M. Nishida *et al.*, “Speaker clustering based on non-negative matrix factorization,” Proc. Interspeech, pp.949-952, Aug. 2011.
[6] N. Dehak *et al.*, “Front-end factor analysis for speaker verification,” IEEE trans. Speech Audio Process., vol.19, no.4, pp.788-798, May 2011.
[7] S. Shum *et al.*, “Exploiting intra-conversation variability for speaker diarization,” Proc. Interspeech, pp.945-948, Aug. 2011.
[8] P.Kenny, *et al.*, “Eigenvice modeling with sparse training data,” IEEE trans. Speech Audio Process., vol.13, no. 3, pp.345-354 May 2005.
[9] P. Kenny *et al.*, “A study of interspeaker variability in speaker verification,” IEEE Trans. Audio, Speech, Lang. Process., vol. 16, no. 5, pp. 980-988, Jul. 2008.
[10] A. Hatch *et al.*, “Within-class covariance normalization for SVM-based speaker recognition,” Proc. ICSLP, pp.1471-1474, Sept. 2006.
[11] D. D. Lee and H. S. Seung, “Algorithms for non-negative matrix factorization,” Proc. NIPS, pp.556-562, Nov. 2000.
[12] K. Maekawa, “Corpus of spontaneous Japanese: its design and evaluation,” Proc. SSPR, pp.7-12, Apr. 2003.
[13] A. Solomonoff *et al.*, “Clustering speakers by their voices,” Proc. ICASSP, vol.2, pp.757-760, May 1998.

© 2012 Information Processing Society of Japan