Vol.2015-SLP-105 No.11
2015/2/28

i-vectorを用いたスペクトラルクラスタリングによる雑音環境下話者クラスタリング

俵 直弘[1]　小川 哲司[1]　小林 哲則[1]

概要：i-vector による話者表現とスペクトラルクラスタリングを組み合わせることで，雑音に頑健な話者クラスタリングを実現する．まず，雑音を含む音声に対して話者クラスタリングを行う場合，高精度な話者特徴量として知られる i-vector を用いて発話間類似度を計算しても，話者の類似度を適切に推定できないことを実験的に明らかにする．また，この問題に対してスペクトラルクラスタリングを適用することの妥当性をグラフラプラシアンの固有ベクトルを分析することで確認する．最後に，スペクトラルクラスタリングの雑音に対する頑健性を実験的に確認するために，日本語話し言葉コーパスに様々な種類の雑音を重畳して得た音声を用いて話者クラスタリング実験を行い，クリーンな音声と同程度の精度で雑音を含む音声のクラスタリングが可能であることを明らかにする．

キーワード：スペクトラルクラスタリング，i-vector，雑音環境下話者クラスタリング

1. はじめに

話者クラスタリングは発話者が未知の発話音声に対し，どの発話が同じ話者によって発話されたものかを推定する問題である．話者クラスタリングには，大別してボトムアップとトップダウンの 2 種類のアプローチが存在する．ボトムアップアプローチでは，最も類似するクラスタのペアの推定を所望のクラスタ数になるまで繰り返すことでクラスタリングを行う [1], [2], [3]．一方，トップダウンアプローチでは全発話を所望のクラスタ数に分割した際に，クラスタ間の類似度を最も小さくし，かつクラスタ内の類似度を最も大きくする分割方法を直接推定することでクラスタリングを行う [4], [5], [6]．いずれのアプローチにおいても，高いクラスタリング精度を達成するためには，発話間で定義した類似度が話者の類似性を適切に表現していなければならない．音声に雑音が含まれる場合，雑音そのものの類似性が発話間の類似度に影響を与えるため，話者類似度の正確な算出は困難となる．したがって，雑音を含む音声に対しても頑健に話者クラスタリングを行うためには，雑音に頑健な発話表現手法が必要となる．

i-vector は話者認識において高い性能を達成している発話表現の一つで，発話ごとに算出した混合ガウス分布（GMM）に因子分析を適用することで得られる．このとき得られるベクトルは，Linear discriminant analysis (LDA) や Wihin class covariance normalization (WCCN) [7]，probabilistic LDA (PLDA) [8] といった統計処理を適用することで，セッションやチャネルの違い等，話者に依存しない変動の影響を低減できることが知られている．また，雑音環境においても，LDA や PLDA で用いる射影行列を雑音を含んだ音声で学習することで，高精度な話者照合が実現できることが明らかになっている [9], [10]．

i-vector は，話者照合のみならず話者クラスタリングにおいてもその有効性が確認されている [5], [6], [11]．i-vector を話者クラスタリングに適用する場合，発話ごとに算出した i-vector のコサイン類似度に基づく k-means 法が多く用いられている．k-means 法以外では，スペクトラルクラスタリング法の適用が検討されている [6]．スペクトラルクラスタリングでは，発話データを類似の発話が近傍となるような多様体上に射影した上でクラスタリングが行われるため，特にデータが元の空間で複雑な分布に従う場合，k-means 法よりも高い精度が得られる．実際に，i-vector 以外の発話表現を用いた先行研究では，スペクトラルクラスタリングが従来法に比べて高精度な話者クラスタリングを実現している [12], [13]．一方，[6] では，i-vector を発話表現として用いたとき各発話は超球上において概ね線形分離可能となり，単純なコサイン類似度基準の k-means 法で十分に高い精度でクラスタリングが可能であると結論づけている．しかし，当該研究では比較的雑音の少ない環境

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

を想定しており，発話の分布がより複雑になる雑音環境下での評価は行っていない．そこで本研究では，i-vector を用いて計算した発話の類似度を基準としたスペクトラルクラスタリングの雑音に対する頑健性を検証する．

以降，2 にて i-vector の概要について簡単に述べ，3 にて，雑音を含む音声に対しては，i-vector に基づく発話類似度が話者の類似性を表現するのに不十分である例を示す．4 にてスペクトラルクラスタリングの概要を述べ，この枠組により雑音によって影響を受ける発話の類似度を補正できることを示す．5 にて雑音を含む音声を用いた話者クラスタリング実験によりスペクトラルクラスタリングの有効性を示す．6 にてまとめと今後の展望について述べる．

2. i-vector に基づく発話類似度

i 番目の発話に依存する C 混合 GMM のスーパーベクトル $\boldsymbol{m}_i \in \mathbb{R}^{CF}$ を因子分析モデルとして以下で表す．

$$\boldsymbol{m}_i = \boldsymbol{m}_0 + \boldsymbol{T}\boldsymbol{x}_i. \tag{1}$$

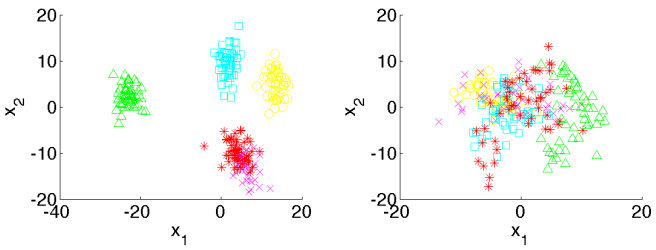
ただし，F は音響特徴ベクトルの次元数である．$\boldsymbol{m}_0 \in \mathbb{R}^{CF}$ は話者および発話非依存の C 混合 GMM のスーパーベクトルで不特定話者による複数の発話により学習した Universal background model (UBM) のスーパーベクトルを用いる．$\boldsymbol{T} \in \mathbb{R}^{CF \times D}$ は低ランクの矩形行列で各列は total variability (TV) 空間の基底ベクトルを表す．TV 空間へ射影されたベクトル $\boldsymbol{x}_i$ は i-vector と呼ばれ，i 番目の発話の特徴量として用いる．ここで得られる i-vector には話者情報に加えて発話内容やチャネル，背景雑音の影響等も含まれるが，これら話者認識に不要な情報は LDA や WCCN 等により抑圧することができる．また，その話者変動成分は超球上に分布することが実験的に示されている [7]．したがって i-vector 間の類似度は，以下のコサイン類似度として定義される．

$$S(\boldsymbol{x}_i, \boldsymbol{x}_j) = \frac{\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{x}_j}{||\boldsymbol{x}_i|| \cdot ||\boldsymbol{x}_j||}. \tag{2}$$

ただし，$\boldsymbol{x}_i$，$\boldsymbol{x}_j$ はそれぞれ i 番目と j 番目の発話から算出された i-vector である．

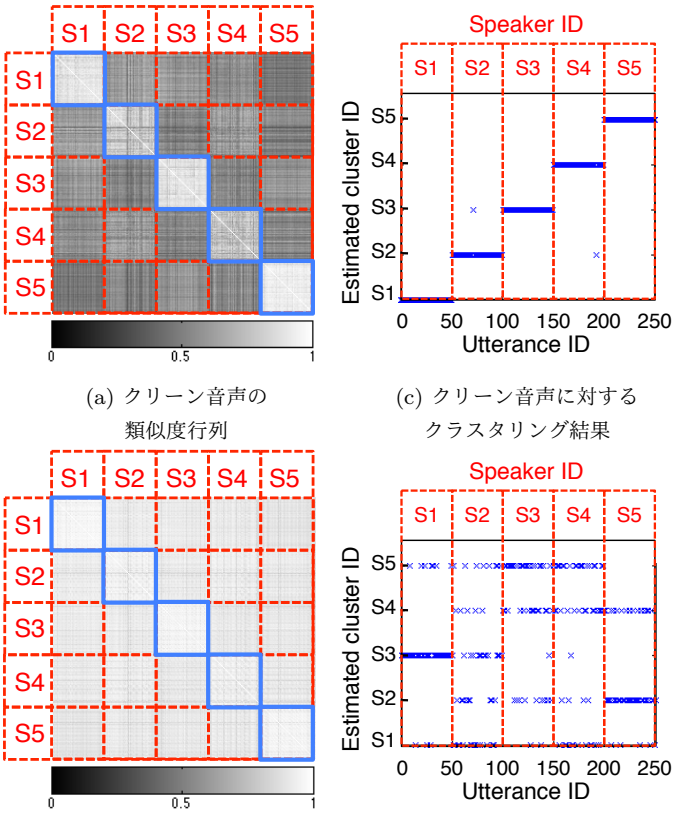
3. 雑音環境下における話者クラスタリング

背景雑音に頑健な話者クラスタリングを実現するため，式 (2) で定義される類似度に背景雑音が与える影響を調査する．図 1 (a) は，背景雑音がない環境 (クリーン環境) で録音した 5 名の発話に対する i-vector の分布を表す．ただし，可視化のため LDA / WCCN により 2 次元に射影した．図 1 (b) は (a) と同じ音声に加法性雑音を重畳して得た音声の i-vector の分布である．図 1 から，クリーン音声の i-vector は話者ごとに識別的に分布するのに対し，雑音下音声に対する話者分布は互いに重なり比較的広い分散を



(a) クリーン環境　(b) 雑音環境

図 1　5 話者による発話に対する i-vector の分布．雑音を含む音声で学習した LDA / WCCN 行列により 2 次元へ射影したものを示した．プロットの色は話者の違いを表す．



(a) クリーン音声の類似度行列　(c) クリーン音声に対するクラスタリング結果

(b) 雑音を含む音声の類似度行列　(d) 雑音を含む音声に対するクラスタリング結果

図 2　クリーン音声および雑音を含む音声に対する発話の類似度行列と k-means クラスタリングの結果．

持つことがわかる．また，i-vector のコサイン類似度を図 2 (a)，(b) に示す．ここでは 5 話者がそれぞれ 50 発話ずつ発話して得られた 250 発話に対する類似度行列を示す．図 2 (a) はクリーンな発話に対する i-vector の類似度であり，(b) は (a) と同一の発話に加法性雑音を重畳して得た音声に対する i-vector の類似度である．両図において，各話者の発話は点線に囲まれた $S1$ から $S5$ の領域で示され，同一話者の発話間の類似度は青い実線に囲まれた領域で示される．このとき発話間の類似度が最大のときに白，最小のときに黒となるように表示した．この図から，クリーン音声に対しては，同一話者による発話の類似度は異なる話者間の発話類似度と比較して高い値となり，雑音環境下で

は，異なる話者の発話であっても同一話者の発話間の類似度と同程度に高い値となることがわかる．式 (1) で表される TV 空間 $\boldsymbol{T}$ には，話者のみならず発話に依存した情報も含まれるため，この空間上に射影された i-vector には背景雑音に依存する情報も含まれる．このとき，i-vector を用いて計算した類似度には雑音の違いに起因する類似度が常に加算されるため，異なる話者間の発話であっても類似度が高い値となったと考えられる．このような類似度に基づき k-means 法を行った結果を図 2 (c)，(d) に示す．図 2 (c) はクリーン音声に対する結果，(d) は雑音下音声に対する結果である．この図から，クリーン音声に対しては，i-vector のコサイン類似度によるクラスタリングはほぼ全ての発話において成功しているのに対し，雑音環境下においてはほぼ全ての発話について失敗していることがわかる．

以上より i-vector のコサイン類似度は，クリーン音声に対しては話者性の類似度を表す適切な類似度となるのに対し，雑音環境下で取得した音声に対しては正しく機能しないことがわかる．次章では，この問題を解決するためにスペクトラルクラスタリングが有効であることを述べる．

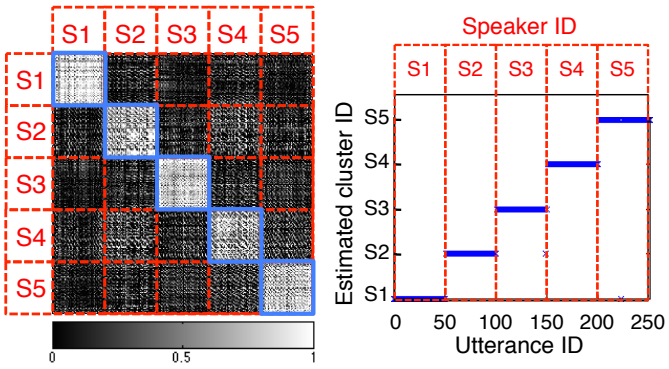
4. スペクトラルクラスタリング

前章で述べた通り，i-vector のコサイン類似度に基づく k-means クラスタリングは，雑音を含む音声に対して性能が劣化する可能性がある．この問題を，スペクトラルクラスタリングを導入することで解決することを試みる．本章では，スペクトラルクラスタリングの概要を述べ，雑音を含む音声に対してスペクトルクラスタリングが有効に働く理由について考察する．

目的関数を定義するため，クラスタに対するデータの割り当ての有無を示す変数である indicator vector $\boldsymbol{t}_i = [t_{i,1}, \cdots, t_{i,j}, \cdots, t_{i,n}]^{\mathrm{T}} \in \mathbb{R}^n$ を導入する．これはクラスタ毎に定義される n 次元ベクトルで，j 番目の発話が i 番目のクラスタに属する場合にのみ $t_{i,j} \neq 0$，それ以外は $t_{i,j} = 0$ となる変数である．このとき，類似度行列 $\boldsymbol{W}$ のグラフラプラシアン $\boldsymbol{L} = \boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{-1/2}$ を定義すると，任意の $\boldsymbol{t}_i$ に対して $\boldsymbol{L}$ は以下を満たす．

$$\boldsymbol{t}_i^{\mathrm{T}}\boldsymbol{L}\boldsymbol{t}_i = \frac{1}{2}\sum_{j=1}^{n}\sum_{j'=1}^{n} w_{jj'}(t_{i,j} - t_{i,j'})^2. \qquad (3)$$

ただし，$\boldsymbol{D}$ は $d_i = \sum_{j=1}^{n} w_{ij}$ を i 番目の対角要素に持つ対角行列，w_{ij} は i 番目と j 番目の i-vector 間の類似度，n は総発話数である．式 (3) からわかるように，類似した発話 (類似度 $w_{jj'}$ が大きい i-vector) に対して $t_{i,j}$ と $t_{i,j'}$ の差が小さくなるとき $\boldsymbol{t}_i^{\mathrm{T}}\boldsymbol{L}\boldsymbol{t}_i$ は小さくなる．したがって，式 (3) を最小化する indicator vector は類似度が大きい発話を同じクラスタに割り当てたクラスタリング結果に対応する．また，各データはいずれかのクラスタに排他的に割り当てられるため，indicator vector $\boldsymbol{t}_i$ は他の全てのクラスタの indicator vector $\boldsymbol{t}_j$ $(j \neq i)$ と直交しなければならない．以上より，最適なクラスタリングを行うためには，全ての indicator vector が互いに直交するという制約のもとで，$\mathcal{F} = \sum_i \boldsymbol{t}_i^{\mathrm{T}}\boldsymbol{L}\boldsymbol{t}_i$ を $\{\boldsymbol{t}_i\}_{i=1}^{K}$ について最小化すれば良いことがわかる．ただしこの最適化問題は NP 困難であり，indicator vector が離散値であるという条件を緩和すると，その解は結局，グラフラプラシアン $\boldsymbol{L}$ の固有値の上位 k 個に対する固有ベクトルと一致する (より詳しい導出は [14] を参照)．このような緩和のもとで得られた indicator vector はクラスタリング結果には対応しない．そこで，得られた k 個の indicator vector を並べて $n \times k$ 次元行列を作成し，各行を各発話データの特徴量とみなして一般的なクラスタリング手法を適用する．ここで得られる発話特徴量は，類似したデータであれば近い値になることが式 (3) により保証されるため，元の i-vector に比べより識別が容易な特徴量と言える．



(a) (b)

図 3 5 話者の発話から算出されたグラフラプラシアンの固有ベクトルに対する類似度行列 (a) とグラフラプラシアンの固有ベクトルに基づく発話特徴量に対し k-means 法を適用して得られるクラスタリング結果 (b)．

このことを確認するため，図 2 (b) で示した類似度行列から計算したグラフラプラシアンの固有値上位 20 個に対応した固有ベクトルを抽出し，それらを用いて類似度行列を計算した．この類似度行列を図 3 (a) に示す．また，図 3 (b) に上述のグラフラプラシアンの固有ベクトルとして得られた発話特徴に対して k-means クラスタリングを適用した結果を示す．これより，上記手順によって得られた発話特徴を用いることで，雑音を含む音声に対してもクラスタリングが可能であることがわかる．

Algorithm 1 に示した通り，スペクトラルクラスタリングでは，最終的に類似度行列のグラフラプラシアンの固有ベクトルをその固有値が大きい順に K 個並べたものを新たな特徴量として，一般的なクラスタリング法を適用する．そこで，雑音を含む発話音声に対して得られた固有ベクトルを観察することにより，スペクトラルクラスタリングが雑音下音声に対してどのように機能するかを考察する．グラフラプラシアン $\boldsymbol{L}$ とその固有ベクトル $\boldsymbol{t}_i$ には以下の関

Algorithm 1 スペクトラルクラスタリングのアルゴリズム [14].

1: Calculate cosine-based distance between all pair of i-vectors $S(\boldsymbol{x}_i, \boldsymbol{x}_j) = 1 - \cos(\boldsymbol{x}_i, \boldsymbol{x}_j), \forall i, j$.
2: Calculate adjacency matrix $\boldsymbol{W} \in \mathbb{R}_+^{n \times n}$, where $(\boldsymbol{W})_{ij} = \exp\{-S(\boldsymbol{x}_i, \boldsymbol{x}_j)\}$ for $i \neq j$ and zero otherwise.
3: Calculate the diagonal matrix $\boldsymbol{D}$ whose (i,i)-th components is sum of i-th row of $\boldsymbol{W}$ (i.e. $(\boldsymbol{D})_{ii} = \sum_{j=1}^{n} w_{ij}$), and construct the graph Laplacian $\boldsymbol{L} = \boldsymbol{D}^{-\frac{1}{2}} \boldsymbol{W} \boldsymbol{D}^{-\frac{1}{2}}$.
4: Select $\boldsymbol{t}_1, \cdots, \boldsymbol{t}_K$, K smallest eigenvectors of $\boldsymbol{L}$ and form $\boldsymbol{T} = [\boldsymbol{t}_1, \cdots, \boldsymbol{t}_K] \in \mathbb{R}^{n \times K}$.
5: Normalize each row of $\boldsymbol{T}$ to have unit length (i.e. $\{\hat{\boldsymbol{T}}\}_{ij} = \{\boldsymbol{T}\}_{ij} / (\sum_k t_{ik}^2)^{1/2}$)
6: Cluster row vectors of $\hat{\boldsymbol{T}}$ via cosine similarity-based k-means clustering.

係が成り立つ．

$$\boldsymbol{L} = \boldsymbol{T}\boldsymbol{\Lambda}\boldsymbol{T}^{\mathrm{T}} = \sum_{i=1}^{n} \lambda_i \boldsymbol{t}_i \boldsymbol{t}_i^{\mathrm{T}}. \tag{4}$$

ただし，$\boldsymbol{\Lambda} = [\lambda_1, \cdots, 0; 0, \cdots, 0; 0, \cdots, \lambda_n]$，$\boldsymbol{T} = [\boldsymbol{t}_1, \cdots, \boldsymbol{t}_n]$ はそれぞれ，固有値 λ_i を要素とする $n \times n$ 対角行列と，固有ベクトル $\boldsymbol{t}_i$ を列とする $n \times n$ 行列とする．雑音環境下で発話された音声の類似度行列は，話者の類似性を与える行列や雑音の類似性を与える行列に分解できると考えられる．このとき，同一話者の発話間の類似度は異なる話者の発話間の類似度よりも相対的に大きくなると考えられる．一方，雑音信号同士は一般に無相関で，その類似度に明白な関係はないと考えられる．また，グラフラプラシアンは半正定値行列であるためその固有値はすべて 0 以上であることを考慮すると，式 (4) の分解が成り立つためには，以下が満たされる必要がある．

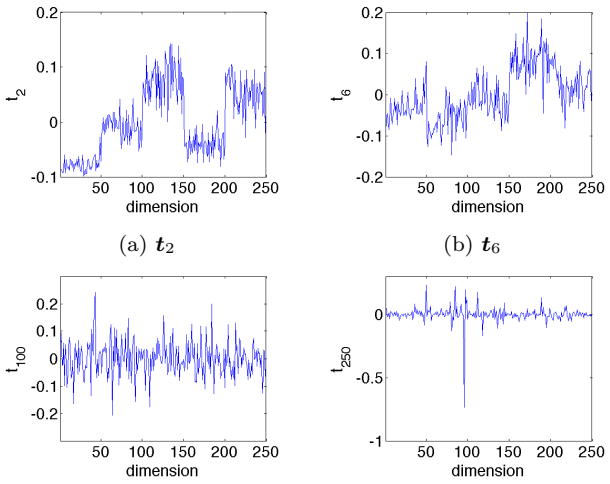
- 類似度行列に含まれる話者の類似度パターンを復元するために，話者の類似度パターンに対応する固有ベクトルは固有値が大きくなければならない．
- 雑音の類似度パターンに対応する固有ベクトルは固有値が小さくなければならない．

図 4 に 2 番目，6 番目，100 番目，250 番目に大きい固有値に対応した固有ベクトルを示す．この図から，固有値が大きい固有ベクトルは話者同士の類似度パターンを含み，固有値が小さい固有ベクトルは雑音間の類似度パターンを含んでいることが確認できる．以上より，固有値が大きい固有ベクトルを選択すれば話者の類似度に起因するパターンのみを特徴量として利用でき，より高い精度でのクラスタリングが可能となる．

5. 話者クラスタリング実験

スペクトラルクラスタリングの雑音に対する頑健性を評価するため，以下の 3 手法を雑音下話者クラスタリングに適用し比較した．

- **GMM-HAC**: 混合ガウス分布で表現された各クラスタについて，cross likelihood ratio [15] をクラスタ間類似度とした凝集的手法．
- **IV-KMEANS**: 発話毎に算出した i-vector のコサイン類似度を用いた k-means クラスタリング [5], [6].
- **IV-SC**: 各発話毎に算出した i-vector のコサイン類似度を用いたスペクトラルクラスタリング．



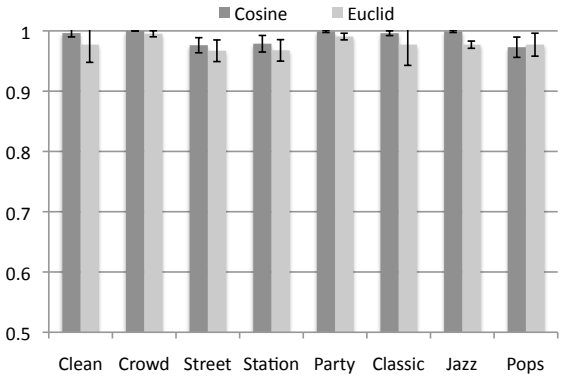
(a) $\boldsymbol{t}_2$ (b) $\boldsymbol{t}_6$ (c) $\boldsymbol{t}_{100}$ (d) $\boldsymbol{t}_{250}$

図 4 雑音を含む音声に対するグラフラプラシアンの上位 (a) 2 番目，(b) 6 番目，(c) 100 番目，(d) 250 番目の固有値に対する固有ベクトル．

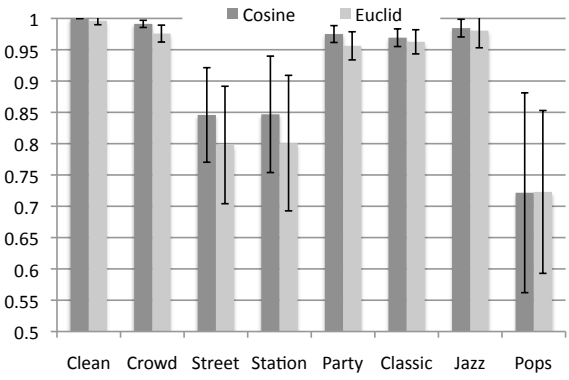
5.1 実験条件

日本語話し言葉コーパス (CSJ) に含まれる講演音声から 10 話者を無作為に選択し，以下の手順によりクリーン環境下評価セットを作成した．まず，抽出した講演データについて，300 ms 以上の無音区間により発話単位に分割した．得られた発話の中から発話長が 1 秒以上の発話を話者ごとに 50 発話選択した．以上の処理を話者と発話の組み合わせを変えて複数回実行し，計 4 つの評価データセットを作成した．異なる性質の雑音がクラスタリング精度に与える影響を調査するため，この 4 つ の評価セットそれぞれに対し，音楽と騒音を重畳することで，雑音環境下における発話データを疑似的に作成した．音楽については，RWC 研究用音楽データベースに含まれる 3 種類の楽曲 (クラシック音楽 (Classic), ジャズ音楽 (Jazz), ポピュラー音楽 (Pops)) をそれぞれ SNR 0 dB で重畳した．また，騒音データとして，電子協騒音データベース付属の展示会場 (通路) および (Party)，駅 (Station)，幹線道路・交差点 (Street)，人混み (Crowd) を，評価データにそれぞれ SNR 0 dB で重畳した．このとき，Party および Crowd はバブルノイズを多く含み，Station および Street は車道を走る車や構内に進入する列車等の非定常なノイズを多く含んでいる．

話者クラスタリング精度の評価尺度として，発話ごとに付与された話者ラベルと推定された話者ラベルから算出し



(a) 平均発話長: 20 秒



(b) 平均発話長: 10 秒

図 5 k-means クラスタリングに用いる距離尺度の比較

た平均クラスタ純度および平均話者純度の幾何平均値 (K 値) [16] を用いた．

スペクトラルクラスタリングでは，最終的に得られるベクトルに k-means 法を適用する．このときクラスタ重心の初期位置を変えた k-means 法を 200 回繰り返し実行し，各データとクラスタ重心との平均二乗距離が最も小さい結果を選択した．

5.2 前処理

作成した音声データに対し，フレーム長 25 ms のハミング窓，フレーム周期 10 ms で音響分析を行い，フレームごとに 12 次の MFCC とエネルギー，およびその Δ パラメータから成る 26 次元の特徴量を抽出した．日本語新聞記事読み上げコーパス（JNAS）と連続音声研究コーパス（ASJ-JIPDEC）に含まれる全ての音声を用いて対角共分散行列を持つ性別非依存の UBM を学習した．このとき発話長の違いによる変動を抑えるため，各発話の長さが 10 秒程度になるように連結した．TV，LDA，WCCN 行列の学習データとして，上記データベースに JEIDA 雑音データベースから抽出した 4 種類のノイズ（Air conditoner, car, factory, plant）を SNR 0 dB で重畳したものを用いた．学習された TV，LDA，WCCN 行列を用いて 150 次元の i-vector を算出し，LDA/WCCN により最終的に 100 次元のベクトルを得た．

5.3 実験結果と考察

5.3.1 スペクトラルクラスタリングにおける k-means 法の距離尺度

スペクトラルクラスタリングの後段の k-means 法で用いる距離尺度がクラスタリング性能に与える影響を調査する．図 5 に，ユークリッド距離とコサイン類似度を距離尺度として k-means 法を適用した結果を示す．これより，ほぼ全ての条件下でコサイン類似度の方がユークリッド距離よりも高い精度が得られた．したがって以降はコサイン類似度を用いた結果のみを示す．

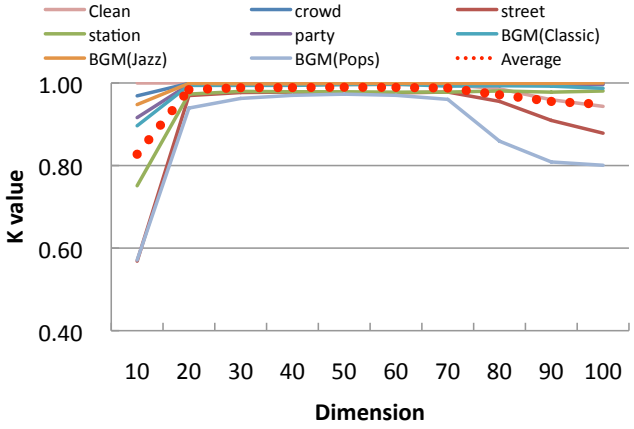


図 6 使用する固有ベクトル数とクラスタリング精度の関係

5.3.2 固有ベクトル数

理想的な条件下では，スペクトラルクラスタリングにより得られる基底ベクトルは indicator vector に一致する．このとき，スペクトラルクラスタリングで必要な固有ベクトルの数は真の話者数に一致する．しかし雑音環境下では，雑音の影響により異なる話者に属する発話間の類似度が不当に大きくなり，この仮定が成り立たない可能性が高い．図 6 に使用する固有ベクトル数とクラスタリング精度との関係を示す．この図から，クリーン環境下では実際の話者数 10 と同じ数の固有ベクトルを用いることで十分な性能が得られたのに対し，雑音環境下では，実際の話者数よりも多い固有ベクトルを使用した際に最も高いクラスタリング精度が得られたことがわかる．この傾向は非定常の雑音環境下において特に顕著であり，平均で 50 個程度の固有ベクトルを用いた場合に最も良い結果が得られた．

5.3.3 クラスタリング精度

表 1 および表 2 に，**GMM-HAC**，**IV-KMEANS** および **IV-SC** をクリーン音声と雑音下音声に対して適用したときの話者クラスタリング性能 (K 値) を示す．表 1 および表 2 は各々平均発話長が 20 秒，10 秒の発話音声に対する結果である．本実験においては，クリーン音声に対しては，発話の長さおよび手法に依らずほぼ完全にクラスタ

表 1　平均発話長 20 秒の雑音環境下話者クラスタリング実験結果（K 値）

Environment		GMM-HAC	IV-KMEANS	IV-SC
Clean		0.955	**1.000**	**1.000**
Stationary noise	Crowd	0.906	0.997	**1.000**
	Party	0.907	0.958	**0.999**
Non-Stationary noise	Street	0.425	0.540	**0.976**
	Station	0.591	0.591	**0.979**
BGM	Classic	0.769	0.930	**0.996**
	Jazz	0.821	0.989	**0.999**
	Pops	0.301	0.383	**0.973**

表 2　平均発話長 10 秒の雑音環境下話者クラスタリング実験結果（K 値）

		GMM-HAC	IV-KMEANS	IV-SC
Clean		0.900	**1.000**	**1.000**
Stationary noise	Crowd	0.672	0.809	**0.981**
	Party	0.727	0.752	**0.964**
Non-Stationary noise	Street	0.225	0.331	**0.876**
	Station	0.398	0.470	**0.820**
BGM	Classic	0.355	0.604	**0.964**
	Jazz	0.467	0.789	**0.983**
	Pops	0.193	0.263	**0.665**

リングできることがわかる．一方，雑音環境下においては，GMM-HAC および IV-KMEANS はクラスタリング精度が著しく劣化した．この傾向は特に Street と Station，および BGM (Pops) において顕著に見られた．Street および Station には非定常な雑音が多く含まれ，BGM (Pops) には歌手の音声が含まれることが，発話の類似度に特に大きな影響を与えたと考えられる．一方，スペクトラルクラスタリングを適用した場合，発話長が 20 秒程度あれば非定常な雑音環境も含めた全ての環境において，クリーン環境下とほぼ同程度の精度が得られた．また，平均発話長が 10 秒の場合，全ての雑音環境下においてスペクトラルクラスタリングの性能が劣化した．発話長が短く，各発話に含まれるフレーム数が少ない場合，発話ごとに推定される i-vector の分散が大きくなり，その類似度の信頼性が低下したと考えられる．しかしこのような条件下においても，スペクトラルクラスタリングは従来手法よりも依然として高い精度でクラスタリングを実現した．

6. 結論と今後の展望

i-vector のコサイン類似度により定義した類似度行列に対してスペクトラルクラスタリングを適用することで，様々な雑音に対して頑健な話者クラスタリングの実現を試みた．その結果，スペクトラルクラスタリングに基づく手法が従来の凝集的手法や k-means 法よりも高い精度でクラスタリング可能であることを示した．

本実験では適切な話者クラスタ数は既知とした．スペクトラルクラスタリングを用いた話者数の推定法としては，固有値の勾配情報を用いた手法がクリーン音声に対して適用されている．しかし本研究で明らかにしたように，雑音環境下では最適な固有ベクトル数と話者数は一致せず，このアプローチは適用できない．そのため今後の展望として，雑音環境の影響に対しより頑健なクラスタ数推定手法を導入することで雑音環境下においても頑健に話者数を推定できる手法の実現を検討している．

参考文献

[1] Chen, S. S. and Gopalakrishnan, P. S.: Clustering via the Bayesian information criterion with applications in speech recognition, *ICASSP*, pp. 645–648 (1998).
[2] Vijayasenan, D., Valente, F. and Bourlard, H.: Agglomerative information bottleneck for speaker diarization of meetings data, *IEEE Automatic Speech Recognition and Understanding Workshop* (2007).
[3] Khoury, E., El Shafey, L., Ferras, M. and Marcel, S.: Hierarchical speaker clustering methods for the NIST i-vector challenge, *Odyssey: The Speaker and Language Recognition Workshop* (2014).
[4] Zhang, X., Gao, J., Lu, P. and Yan, Y.: A novel speaker clustering algorithm via supervised affinity propagation, *ICASSP*, pp. 4369–4372 (2008).
[5] Shum, S., Dehak, N., Chuangsuwanich, E., Reynolds, D. A. and Glass, J. R.: Exploiting intra-conversation variability for speaker diarization, *Interspeech*, pp. 945–948 (2011).
[6] Shum, S., Dehak, N. and Glass, J.: On the use of spectral and iterative methods for speaker diarization, *Interspeech* (2012).
[7] Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P. and Ouellet, P.: Front-end factor analysis for speaker verification, *IEEE Trans. Speech Audio Process.*, Vol. 19, No. 4, pp. 788–798 (2011).
[8] Kenny, P.: Bayesian speaker verification with heavy-tailed priors, *Odyssey: The Speaker and Language Recognition Workshop* (2010).
[9] Garcia-Romero, D., Zhou, X. and Espy-Wilso, C. Y.: Multicondition training of Gaussian PLDA models in i-vector space for noise and reverberation robust speaker recognition, *ICASSP*, pp. 4257–4260 (2012).
[10] Lei, Y., Burget, L., Ferrer, L., Graciarena, M. and Scheffer, N.: Towards noise-robust speaker recognition using probabilistic linear discriminant analysis, *ICASSP* (2012).
[11] Zheng, R., Zhang, C., Zhang, S. and Xu, B.: Variational bayes based i-vector for speaker diarization of telephone conversations, *ICASSP*, pp. 91–95 (2014).
[12] Ning, H., Liu, M., Tang, H. and Huang, T.: A spectral clustering approach to speaker diarization, *ICSLP* (2006).
[13] Iso, K.: Speaker clustering using vector quantization and spectral clustering, *ICASSP*, pp. 4986–4989 (2010).
[14] Ng, A. Y., Jordan, M. I. and Weiss, Y.: On spectral clustering: Analysis and an algorithm, *NIPS*, pp. 849–856 (2001).
[15] Reynolds, D. A., Quatieri, T. F. and Dunn, R. B.: Speaker verification using adapted Gaussian mixture models, *Digital Signal Processing*, Vol. 10, No. 1–3, pp. 19–41 (2000).
[16] Solomonoff, A.: Clustering speakers by their voices, *ICASSP*, pp. 757–760 (1998).