論　文

発話セグメントクラスタの評価とそれに基づく改良ボトムアップクラスタリングによる話者ダイアライゼーションの高精度化

陳　伯翰[†]　　北岡　教英[†]　　武田　一哉[†]

Modified Bottom-up Clustering Based on Evaluation of Speech Segment Cluster for Improved Speaker Diarization

Bohan CHEN[†], Norihide KITAOKA[†], and Kazuya TAKEDA[†]

あらまし　本論文では，話者ダイアライゼーションにおけるクラスタリングエラーを減らし，高精度化する手法を提案する．クラスタがある 1 話者の発話セグメントのすべてを含み，かつ他のセグメントを含まないという条件を満たすか否かを判定できる評価があれば有用である．本論文では，評価するクラスタを表現する i-vector とデータ全体を分割したセグメントの i-vectors との間のコサイン類似度のヒストグラムに基づいて，それを実現する．さらに，この評価をボトムアップクラスタリングに基づくダイアライゼーションにおけるオーバーマージの防止に用いる．この方法により，ダイアライゼーションの精度を向上することができた．

キーワード　話者ダイアライゼーション，　ボトムアップクラスタリング，　クラスタ評価，　高精度化

1. まえがき

会議や放送音声などにおいて，事前情報を一切持たず「いつ，誰が話したか」を推定する問題を話者ダイアライゼーションと言う [1]．話者ダイアライゼーションは一般に，発話セグメントのクラスタリング問題とみなして実現する．話者ダイアライゼーションにおいてもっとも理想的な結果は，発話クラスタが各々単一話者のデータで構成され，かつ同一の話者のセグメントが別々のクラスタに分割されないことである．つまり発話セグメントは話者の個人性に基づいてクラスタリングされる．

一般的な手続きとしては，まず前処理として何らかの方法によって入力データから非音声とオーバーラップの部分を除去する．そして，残りのデータから適当な方法で区切ることにより初期セグメントを作る．最後にクラスタリングによって全てのセグメントを話者ごとに分類する [1]．セグメントの作り方については，適当なセグメントからはじめ，ダイアライゼーション中に再びセグメンテーション（リセグメンテーション）することを通じて各セグメント内のデータを調整する方法 [2] や，データを等間隔で区切ってセグメントとする方法も存在する [3]．このような様々な方法でセグメントを区切っても，結果に大きな違いがないことから，セグメントの作り方は最終結果に大きく影響を与えないと考えられる．したがって，その後の適切なクラスタリング方法を考案することを本研究の主なる目標とするのが妥当であろうと考える．

対話に関して事前情報を一切持たない点が話者ダイアライゼーションを困難にする一因である．これはつまり，対話の参加者の人数も各話者の音声の特性も推定しなければならないことを意味する．従来この推定を行うとき，全ての話者が平等に扱われる．つまりクラスタリングが停止して得られる，ダイアライゼーションの最後の結果である各クラスタが，それぞれ 1 名ずつの話者であるとする [2] [4]．しかし実際には，数名の話者の中には，他の話者と音声の特徴の違いがきわ立ってあり，識別しやすい話者が存在することが多い．例えば，手作業で話者を分類することを考えた時，このような“分りやすい”話者を最初に分離することにより，作業を効率的で正確にすることができることは，理解できるであろう．我々は計算機を用いて自動ダイアライゼーションをする場合にも，同じ方法を使

[†] 名古屋大学大学院情報科学研究科
Graduate School of Information Science, Nagoya University

うことで，より高い精度の結果を出せるのではないかと考えた．

本論文では，クラスタリングの途中の段階で，あるクラスタが１人の話者のほぼ全ての発話セグメントを含み，他の話者をほとんど含まないという，ダイアライゼーションにおけるクラスタの最終状態に達した状態か否かを判断するクラスタ評価法を提案する．さらに，この提案法を用いて改良したボトムアップクラスタリングを提案する．実験では，提案するクラスタ評価法が有効な判断を下せることを示す．そしてクラスタ評価法に基づいた改良ボトムアップクラスタリングにより，ダイアライゼーションの精度が向上することを示す．

本論文は以下のように構成される．２節は提案するクラスタ評価法について説明する，３節でクラスタ評価法を利用してボトムアップクラスタリングを改良する方法を説明する．４節では評価実験とその結果を述べ，５節で結論を述べる．

2. 話者空間に基づくクラスタ評価法

本節では，クラスタリングの過程においてクラスタが特定話者の全ての発話セグメントを含み（再現率 ≈ 1），かつ他の話者の発話セグメントを含まない（適合率 ≈ 1）という条件をほぼ満たしているか否かを判断する方法を提案する．話者の特徴を比較的よく表す特徴量を用いた場合には，その特徴量空間で，この理想的な条件に近いクラスタはある統計的な性質を示す．本論文ではこのことを利用してクラスタを評価し，判定する

2.1 話 者 空 間

音声信号には話者情報以外にも様々な情報が存在する．ダイアライゼーションにおいて，それらの情報を含んだままクラスタやセグメントを評価すれば，不要な情報の影響を受けて，精度が低下する．したがって，音響信号から話者の個人性のみ分離して評価に用いたい．これを実現するために次のベクトル空間（話者空間）が提案された [5] [8]．

本論文では，[5] で提案された話者空間を利用する．この方法は因子分析（Factor Analysis）を用いてデータ集合の話者性を表現する方法である [5] [6]．ある話者及びチャンネル（あるいはセッション）依存のスーパーベクトル M（通常はデータ集合から学習した GMM の平均ベクトルを連結したもの）は以下のように表現できる:

$$M = m + Tw, \tag{1}$$

ここで m は話者及びチャンネル独立のスーパーベクトルである．一般に大量のデータで学習した GMM（universal background model）の平均ベクトルを使う．この m が音響特徴量の中の話者及びチャンネル独立な成分を表現すると考えられる．そして T はいくつかの擬似話者を用いて構成する長方形行列であり，この行列を構成するベクトルによって張られる空間は Total variability space と呼ばれる．そして個々の話者を表す話者空間内のベクトル w は i-vector あるいは全因子ベクトル (Total factor vector) と呼ばれる．

従来 i-vector を使ってデータの話者特性を抽出する時，セッション変動の影響を抑えるため，LDA や NAP，WCCN 等の手段を用いて補正する必要がある [5]．しかし今回のダイアライゼーションタスクのデータは，１人の話者がごく短い時間内に，同じ場所同じマイクで収録されるため，話者内のセッション変動が存在しないと考えて差し支えないと考える．したがって，本論文では２つの i-vector 間の類似度を計る場合単純なコサイン類似度を使う．つまり，話者空間内のベクトル w_A と w_B を用いて表現されるデータ集合 A と B の類似度は

$$\sigma(w_A, w_B) = \frac{(w_A)^t (w'_B)}{||w_A|| \cdot ||w_B||} \tag{2}$$

を用いて計算できる．

2.2 クラスタ評価

2.1 節で説明した話者空間に基づいて，クラスタが特定話者のみの発話セグメントの全てを含むか否かを評価する方法を提案する．ここで話者空間内でクラスタ α を表現する i-vector のことを w_α と表記する．そして w_α を評価するため，全てのセグメントの i-vector と w_α との類似度の集合 $\mathbf{D}_\alpha$ を以下のように定義する．

$$\mathbf{D}_\alpha = \{\sigma(w_\alpha, \mathbf{s}_1), \sigma(w_\alpha, \mathbf{s}_2), ..., \sigma(w_\alpha, \mathbf{s}_N)\} \tag{3}$$

ここで $\mathbf{s}_i$ はセグメント s_i の i-vector を表す．N はセグメントの数である．セグメントの分割方法について，本論文では文献 [3] の方法を用いて，3 秒の窓幅と 1 秒のシフト幅でセグメントを作成する（オーバーラップあり）．そして $\sigma(.)$ は話者空間内の２つのベクトルの類似度を表す．

文献 [5] の仮定より，特定話者に属するデータより計算される i-vector　w はガウス分布に従う．つまり空間内の任意のベクトルに対して，同じ話者に属するベ

クトルとこの任意のベクトルとの類似度の値がある範囲内に集中する傾向を有すると考えられる. この場合話者 A の代表ベクトル（話者 A に属する全てのデータを用いて計算するもの）と話者 A に属する全てのセグメントの代表ベクトルとの類似度は高い値となる傾向を有する，また，それ以外の話者に属するセグメントの代表ベクトルとの類似度は低い値をとる傾向を有する. したがって，もしクラスタ α がほぼある特定の話者 A のセグメントからなる場合，全てのセグメントのうち，A に属するセグメントとクラスタ α の代表ベクトル w_α は類似度が大きく，それ以外のセグメントとは小さくなり，全セグメントとの類似度の分布は２峰性を示す．α に他の話者のデータが多く含まれば，w_α は A の代表ベクトルから遠くなり, 明確な２峰性ではなくなる. そこで, この２峰性分布を２混合ガウス分布で近似することとする．

式（3）の分布が２峰性を示すか否かを判断するため，評価関数が必要である. このようなモデル選択問題に対して従来 AIC や BIC を用いることが多い. しかし, 今回の問題に対して，AIC や BIC では「２峰性のモデル」と「その他のモデル」を用意する必要がある．その他のモデルは峰の数が不明のため, 多数用意する必要がある，したがって,「その他のモデル」の用意は困難と考える．そこで本論文では集合 $\mathbf{D}_\alpha$ 内の要素を用いて２混合 GMM を学習して，平均対数尤度を用いてクラスタ C_α を評価する方法を提案する．

$$\ln \hat{L} = \frac{1}{N}\sum_{i=1}^{N} \ln\{w_1 N(d_i;\mu_1,\sigma_1^2) + w_2 N(d_i;\mu_2,\sigma_2^2)\} \tag{4}$$

ここで，d_i は集合 $\mathbf{D}_\alpha$ 内の要素で，N はセグメントの数を表す．クラスタ C_α がほぼ１人の話者のデータからなり，その話者のデータをほぼすべて含むならば, その話者に対応するデータとそれ以外のデータの d の値がほぼ二峰性となり，モデルに合うために $\ln \hat{L}$ が大きい値となる．一方，その条件から大きくずれている場合には，$\ln \hat{L}$ は小さな値となる（図１参照）．

最後に，平均尤度が閾値を超えたら，

$$\omega_\alpha N(d_i;\mu_\alpha,\sigma_\alpha^2) > \omega_{\bar{\alpha}} N(d_i;\mu_{\bar{\alpha}},\sigma_{\bar{\alpha}}^2) \tag{5}$$

を満たすセグメント s_i をクラスタリングプロセスから除外し，それらを話者 A に属するセグメントとして扱い，１つのクラスタとする．ここで，μ_α,σ_α^2，ω_α と $\mu_{\bar{\alpha}}$,$\sigma_{\bar{\alpha}}^2$，$\omega_{\bar{\alpha}}$ は GMM 中の平均が大きい方のガウス分布の平均と分散と重み，平均が小さい方のガウス分布の平均と分散と重みである．

効果を比較するため，実験では式 (5) 以外にもう１つ判断基準を試す．これは，分散と重みを考慮せず

$$|\sigma(\mathbf{s}_i, w_\alpha) - \mu_\alpha| > |\sigma(\mathbf{s}_i, w_\alpha) - \mu_{\bar{\alpha}})| \tag{6}$$

を満たすセグメントをクラスタリングプロセスから除外して１つのクラスタとする．

3. 改良型ボトムアップクラスタリング

本節では，ボトムアップクラスタリングに基づいて，２節で提案する方法を用いて別の話者のクラスタと合併してしまうオーバーマージの発生を防止する改良案について述べる．ボトムアップクラスタリングはダイアライゼーション問題に対して最も利用される枠組みである [2] [4]．従来のボトムアップクラスタリングのアプローチは以下のようになる

（1） セグメントを K 個の初期クラスタに分類

（2） 類似度が最も高いクラスタ対を合併

（3） クラスタのモデルを更新し，そのモデルを用いてセグメンテーションをやり直す

（4） 最も高い類似度が閾値を下回るまで (2)-(4) を繰り返す

従来のボトムアップクラスタリングでは，繰り返し毎に，全ての入力データを用いてクラスタモデル（一般には GMM など）を更新する．そしてあらゆるクラスタ対の類似度を評価し，類似度が最も高いクラスタ対を合併する．あるレベルで類似度が最も高いクラスタ対の類似度が閾値より低くなったら，現在のクラスタはすべて異なる話者に属すると考えてクラスタリングを停止し，その時点のクラスタを結果として出力する．従来のボトムアップクラスタリングでは，クラスタに対するラベル付け（クラスタに含まれる全てのセグメントに，ある唯一の話者のラベルを付ける）を最後に行う．同一の話者のクラスタが合併される前に, 異なる話者のクラスタが合併されることもあり, その場合, 後の操作で正しいクラスタに戻すことはできなくなる. 従って，合併の制御が話者ダイアライゼーションで最も難しい問題である．各クラスタをオーバーマージさせないと同時に必要な合併を全て行うのが最も望ましい制御である．ダイアライゼーションの実際の応用を考えた場合, 合併されない話者が存在するよりもオーバーマージの方が悪影響がある場合が多い．したがっ

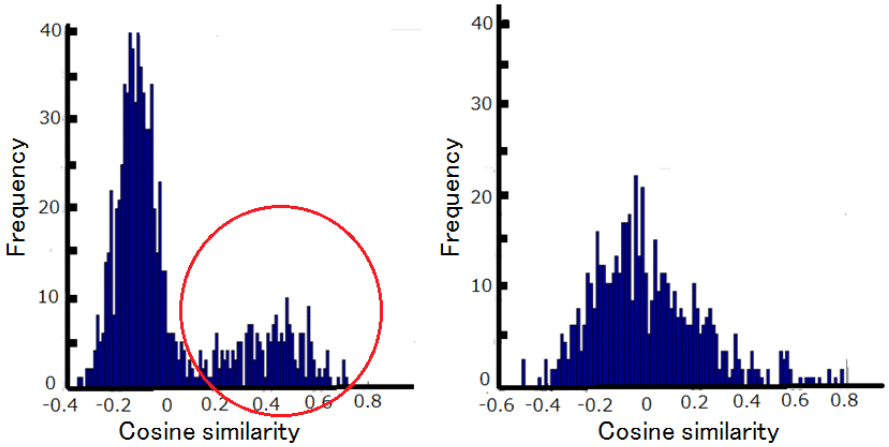

図 1　理想的なクラスタの代表ベクトル (w_α) の $\mathbf{D}_\alpha$ のヒストグラム（左）と理想でないクラスタの代表ベクトル (w_β) $\mathbf{D}_\beta$ のヒストグラム（右）．理想的にクラスタリングされていれば，赤い円の中のセグメントは 1 人の話者に対応する．

Fig. 1　The histograms of an ideal vector's(w_α) $\mathbf{D}_\alpha$(left) and an imperfect vector's(w_β) $\mathbf{D}_\beta$(right). Segments in the red circle all belong to one speaker when clustering doing the trick.

て比較的高い閾値を設定してオーバーマージを防止する研究もある [10]．我々は前節で提案するクラスタ評価を用いることにより，より早い段階で容易に区別できる話者のクラスタを確定でき，オーバーマージを防止できることを示す．従来のボトムアップと比較して，より低い類似度の閾値でアルゴリズムを実行でき，その結果クラスタリングの精度を上げることが可能であると考える．

我々は以下のようなクラスタ評価に基づく改良型ボトムアップクラスタリングを提案する

（1） セグメントを K 個の初期クラスタに分類

（2） 類似度が最も高いクラスタ対を合併

（3） 全てのクラスタを評価する．話者の全ての発話セグメントを含むクラスタが存在すれば，そのセグメントをクラスタリングプロセスから除外

（4） クラスタのモデルを更新，リセグメントする

（5） 最も高い類似度が閾値を下回るまで (2)-(4) を繰り返す

話者ダイアライゼーションの目標は話者ごとにクラスタを作ることであるため，既に話者の全ての発話セグメントを含むクラスタは，それ以上合併する必要がない．したがって，我々が提案する改良型ボトムアップクラスタリングでは，従来のボトムアップクラスタリングと異り，クラスタ評価法を用いてこのようなクラスタを検出し，クラスタリングプロセスから除外することによって，オーバーマージを防止する．そして残りのデータを従来のボトムアップクラスタリング法でクラスタリングする．

4. 実　　験

本節では，提案するクラスタ評価法とボトムアップクラスタリング法についての実験とその結果を述べる．実験では AMI English meeting corpus [11] の中の 8 つの会議録音（4.76 時間）を開発データとして用いる．そして開発データと異なる 8 つの会議データをテストデータ（5.43 時間）として用いる(注1)．AMI project はヨーロッパで開催されたマルチモーダル会議分析プロジェクトである．コーパスは総計 100 時間以上の会議のデータ（音声や画像，アノテーションなど）を提供する．会議のテーマは一定（商品の設計）であり，毎回の参加者は 3 ～ 5 人で，1 グループの参加者が 4 つの会議を行う．1 つの会議の長さはおよそ 30 分程度である．今回使用される会議は原則的にランダムで選んだ．ただし AMI コーパスの会議の話者数はほとん

(注1)：用いたデータは AMI コーパスの 10% 程度であるが，開発データ，テストデータ共に 5 時間程度あり，実験結果は信頼できる．ただし，他の AMI コーパスによる実験との比較はできない

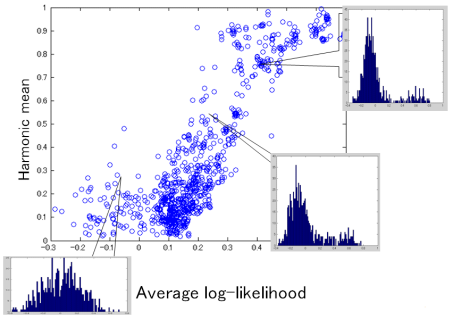

図 2　提案尺度とクラスタの再現率と適合率の調和平均との関係，縦軸がクラスタの再現率と適合率の調和平均，横軸が式 (4) を用いて計算した平均対数尤度

Fig. 2　Verification results of the proposed cluster evaluation method. The horizontal axis is the average log-likelihood calculated using4, and the vertical axis is the harmonic mean of the recall and the precision of the dominant person's segments in the cluster.

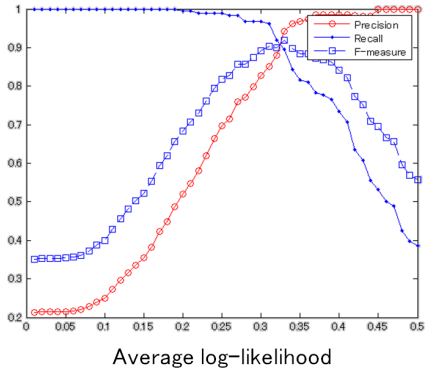

図 3　図 2 中の調和平均（縦軸）の閾値を 0.8 と設定する時，「理想的なクラスタ」の識別について提案の平均対数尤度の F 値

Fig. 3　F-measure of the proposed method when consider 0.8 as the threshold of the clusters' harmonic average(Fig.2)

どが 4 人であり，人数が固定であることの影響を避けるため，3 人で行う会議も意図的に選んだ（5 人で参加する会議は後の条件を満たさないため，選ばなかった）．今回は話者数既知の K-means 法を 1 つのベースラインとする．話者の発話時間の影響を避けるため，データを選ぶ時, 発話時間が 2 分以下の話者が含まれるデータは選択しなかった．録音環境はヘッドセットマイクで，会議はそれぞれ異なるメンバーで構成される．開発データでは 8 つ共に 4 人で，テストデータでは 3 人で行う会議は 3 つで，他はすべて 4 人である．ダイアライゼーションは会議ごとに行う．特徴量の抽出には ALIZE [12] を利用して，録音データからまず 30 ミリ秒の窓幅と 10 ミリ秒のシフト幅で 60 次元の LFCC（Linear Frequency Cepstral Coefficient）を抽出し，そして文献 [3] の方法を用いて，3 秒の窓幅と 1 秒のシフト幅でセグメントを作成し（オーバーラップあり），セグメント中に含まれる 300 個のフレームに基づいて i-vector を計算する．

今回は話者 1 人の発話データのみで構成するクラスタの識別及びこれをクラスタリングに用いる効果を評価することが目的である．そのため，コーパス中の非音声部分及び複数人が同時に発話しているオーバーラップ部分は，コーパスで提供されるアノテーションに基づいて事前に人手で除外した．

4.1　クラスタ評価法

本実験では，式（4）の平均対数尤度が「クラスタの理想状態」の評価尺度として妥当か否かを議論する．まずは「クラスタの理想状態」を表す尺度について，今回は各クラスタをクラスタ内のデータ量を最も占める話者に分類されるものとし，クラスタ内のデータにおける，この話者の再現率と適合率をフレーム単位で計算し，その調和平均をクラスタの理想さを表す尺度とする．開発セットのデータを 3 節で述べる従来法に従って，クラスタ間類似度が停止条件（同じデータを用いて算出）を満たすまで 1 回ボトムアップクラスタリングをする．クラスタリングの過程で現れるすべてのクラスタの理想さと提案の評価尺度との関係を調べた．その結果を図 2 に示す．クラスタが理想条件に近いか否かと平均尤度との間に，明らかな正の相関を持つ．そして調和平均 0.8 付近を境界線として，データが 2 組に分けられることが見て取れる．そこで，調和平均（縦軸）が 0.8 を超えるクラスタを便宜的に「理想的クラスタ」とし，提案の平均尤度を閾値として理想的クラスタか否かを判定した場合の再現率，精度および F 値を測った．その結果を図 3 に示す．平均尤度は 0.32 の付近で，0.9 程度の最大 F 値となった（参考に 0.7 と 0.9 を「理想クラスタ」境界とした場合の最大 F 値は．それぞれ尤度が 0.3 の時 0.95，尤度が 0.4 の時 0.8 であった）．したがって，以下は式 (4) の閾値を 0.3 と設定する．

4.2　改良型ボトムアップクラスタリング

今回の実験では初期クラスタの数 K を数名の会議を対象とすることを前提として十分と考えられる 16 に設定し，全データを等間隔に K 区間に区切って各

区間を 1 つずつ含むクラスタを初期クラスタとした [2]（K の値は 8 や 32 も試したが，結果はほぼ同じであった）．今回ボトムアップクラスタリング中の類似度評価においても i-vector を用いたが，実際には，クラスタリングにおける類似度評価は全く異なる方法でもかまわない．類似度に対する閾値を停止条件（開発セットで性能の良かった値）として用いる 3 節で紹介する従来のボトムアップクラスタリングをベースラインとする．i-vector に K-means 法 [7] を適用したものをもう 1 つのベースラインとする．つまり話者数を既知と仮定し，全てのセグメントの i-vector に対して K-means クラスタリングを用いて話者数のクラスタに分類する．K-means 法の分類結果はクラスタの初期重心に依存するため，今回の実験ではランダムな初期重心で K-means を 20 回行った結果で最も精度のよい結果を採用する．そして提案法は，4.1 節で得た最適平均対数尤度を用いて，3 節で提案したアルゴリズムで求めた最適な停止類似度（開発セットで性能の良かった値）を用いたボトムアップクラスタリングである．なお，開発セットで求めるのはこの 2 つの閾値のみである．ダイアライゼーションの結果は [9] で提案される DER(Diarization error rate) を用いて評価することがほとんどである．しかし DER はダイアライゼーションの全体の結果についての評価で，今回の実験では非音声とオーバーラップ部分のデータを人為的除外されるため，文献 [13] を参考に, 誤分類率，クラスタ純度，Rand Index の 3 つの指標を用いてクラスタリングの結果を評価する．

誤分類率 M はクラスタリングの結果と正解とを一対一マッピングして，

$$M = \frac{\sum_{j=1}^{N} e_j}{\sum_{i=1}^{C}\sum_{j=1}^{N} n_{ij}} \tag{7}$$

に従って計算する(注2)．ここで N は話者の数，C はクラスタの数，n_{ij} はクラスタ i に分類される話者 j に属するフレームの数を表す（ここでのフレームは LFCC を抽出する時に使うフレームである）．

クラスタ純度 P は

$$P = \frac{\sum_{i=1}^{C} f_j}{\sum_{i=1}^{C}\sum_{j=1}^{N} n_{ij}} \tag{8}$$

に従って計算する．クラスタ純度の計算にはマッピングは不要である．ここで f_i はクラスタ中データ量がもっとも大きい話者に属するフレームの数である．

最後に，二つの分類結果の一致性を表す Rand Index [14] の値 R は

$$R = \frac{a+b}{\binom{n}{2}} \tag{9}$$

を用いて計算する．ここで a は二つの分類結果において，共に同じクラスタに分類されるフレームペアの数，b は二つの分類結果において，共に異なるクラスタに分類されるフレームペアの数，n はデータ全体のフレーム数である．Rand Index はフレームペアの 2 つの分類が一致する数とフレームペアの組み合わせ数の比を用いて，分類結果の一致性を表す.

結果を表 1 に示す．ここで CB と KM がそれぞれ 3 節で紹介する従来のボトムアップ手法と文献 [7] で提案する手法に対応する．提案法はそれぞれ PM（式 (6)）と PM（式 (5)）に対応する．結果からわかるように，3 つの評価尺度に対して，提案法の方がこれらのボトムアップクラスタリングの改良法がベースラインより高い精度のダイアライゼーションを実現した．一方セグメントのラベル付けについては，式 (5) と式 (6) いずれも近い結果を得た．

そして，今度は提案法のオーバーマージ防止効果を確かめため，異なる停止閾値での提案法と従来のボトムアップ法とのクラスタリング結果を比較する. その結果を図 4 に示す. 図より，従来法と比べ提案法の方が停止閾値に対して頑健であることが分る. 特に停止閾値が低い時（すなわち，停止しにくいとき），従来法のクラスタリング結果の変化が大きい，これはオーバーマージによる影響である．

人間であれば, まずは簡単な話者のデータを処理するであろうというのが著者の最初の発想である．しかし，もともと全ての話者が区別されやすい場合 (ES2008b), 従来法でも十分な性能となり，提案法による改善がない．そして，クラスタ評価よってクラスタを検出できない場合 (EN2002c)（すなわち，人間の場合容易に区別できる話者が見つけられない場合に対応），従来法と全く同じとなる．さらに，クラスタ評価が誤検出して (今回のデータでは発生しなかった)，後続の処理で回復する術がなくため，誤り率が著しく増加する可能性も存在する．しかし全体の結果から見ると, 今回の提案法が従来のボトムアップクラスタリングより頑健である．これは大きな誤り（オーバーマージ）が発生

(注2)：オーバーラップと無音を除いた場合，ダイアライゼーションエラー（DER）はこの値と一致する．

表 1　Diariazation result

Data	Misclassification rate (%)				Cluster purity (%)				Rand Index			
	CB	KM	PM(式 (5))	PM(式 (6))	CB	KM	PM(式 (5))	PM(式 (6))	CB	KM	PM(式 (5))	PM(式 (6))
EN2002c	9.11	8.65	9.11	9.11	91.00	91.46	91.00	91.00	0.89	0.89	0.89	0.89
EN2009b	21.44	7.34	15.90	15.94	88.81	92.85	84.75	84.25	0.84	0.90	0.82	0.82
IN1001	9.23	22.71	9.55	9.89	90.86	83.82	90.54	90.20	0.89	0.80	0.86	0.86
ES2003b	16.15	12.24	5.19	5.86	83.97	87.88	95.40	95.11	0.86	0.89	0.95	0.94
ES2007b	33.65	17.63	11.75	12.56	78.38	83.00	88.49	87.68	0.80	0.84	0.88	0.87
ES2008b	5.08	6.06	8.62	8.54	91.39	94.07	92.43	91.60	0.94	0.94	0.92	0.92
ES2014b	23.71	8.19	7.76	8.18	89.47	91.99	92.41	91.99	0.85	0.91	0.92	0.91
ES2016c	15.18	4.21	4.91	5.59	84.87	95.96	96.17	95.86	0.92	0.96	0.96	0.96
Ave.	16.69	10.81	9.10	9.46	87.04	90.14	91.34	90.96	0.87	0.89	0.90	0.89

しにくいであり，平均的には従来の方法より高い精度が得られると考える．

5. 結　　論

本論文では，話者ダイアライゼーションにおけるクラスタリングでクラスタが１人の話者に属する全ての発話セグメントのみを含むという条件を満たすか否かを評価する方法を提案した．このクラスタ評価法に基づいて，ボトムアップクラスタリングの改良案も提案した．実験の結果，提案するクラスタ評価法が条件をほぼ満たすクラスタとそれ以外のクラスタを区別できることが分かった．さらに，ダイアライゼーションの実験では，提案法が２つのベースラインより高い精度を得た．

今回の提案法は対象クラスタと全てのセグメントとの間の統計的な性質を利用して実現する方法であるため，短いデータや発話区間が短い話者への対応が難しいと考えられる．今後の課題として，こうした発話への対処法を考える必要がある．また，今回提案するクラスタリング手法の精度はクラスタ評価の精度を依存する．もしクラスタの評価で誤りを起こったら，大きなかつ修正できないエラーが発生すると考える．今回のデータに対して提案のクラスタ評価手法の識別結果はかなり良い（0.9 程度の F 値）．故にクラスタ評価からなるエラーは存在しない．クラスタ評価の結果をもっとうまく利用し，評価が誤ってもそれより起こされるエラーを最小化するクラスタリング手法もまた今後の大として考える．さらに，今回利用した AMI meeting corpus は人数や対話のテーマが比較的狭い範囲に偏っている．提案法の一般性を検証するため，データ量を増してより多くの条件において実験する必要があると考えている．

文　　献

[1] S. Tranter, and D. A. Reynolds, “An overview of automatic speaker diarization systems”, *IEEE Trans. Audio, Speech, Lang. Process*, 14(5), Sep. 2006.

[2] C. Wooters, and M. Huijbregts, “The ICSI RT07s speaker diarization system, ” *NIST RT07 Meeting Recognition Evaluation Workshop*, 2007, pp. 509-519, Springer.

[3] X. Anguera, “Fast speaker diarization based on binary keys”, *IEEE ICASSP 2011* pp. 4428-4431, 2011.

[4] D. A. Reynolds, and P. Torres-Carrasquillo, “The MIT Lincoln Laboratory RT-04F Diarization Systems: Applications to Broadcast Audio and Telephone Conversations”, *DARPA EARS RT-04F Workshop*, Nov. 2004

[5] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Quellet, “Front-End Factor Analysis for Speaker Verification, ” *IEEE Trans. Audio, Speech, Lang. Process*, 19(4), May 2011.

[6] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, “Joint factor analysis versus eigenchannels in speaker recognition, ” *IEEE Trans. Audio, Speech, Lang. Process*, 15(4), May 2007.

[7] S. Shum, N. Dehak, D. Chuangsuwanich, D. Reynolds, and J. Glass, “Exploiting Intra-Conversation Variability for Speaker Diarization”, *Ineterspeech, 2011*

[8] W. Tsai, S. Cheng, and H. Wang, “Speaker Clustering of Speech Utterances Using A Voice Characteristic Reference Space” *ICSLP 2004*.

[9] http://www.nist.gov/itl/

[10] D. Vijayasenan, F. Valente, and H. Bourlard, “An information theoretic approach to speaker diarization of meeting data, ” *IEEE Trans. Audio, Speech, Lang. Process*, 17(7), Sep. 2009.

[11] https://corpus.amiproject.org/

[12] http://mistral.univ-avignon.fr/

[13] D. Liu and F. Kubala, “Online speaker clustering,” *ICASSP2004, SP-P2.6*

[14] W. M. Rand, “Objective criteria for the evaluation of clustering methods.” *Journal of the American Sta-*

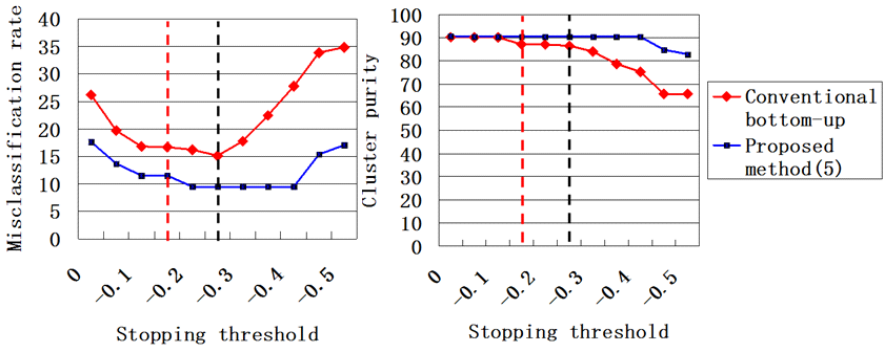

図 4　誤分類率と停止閾値との関係（左），クラスタ純度と停止閾値との関係（右）．低い停止閾数に対して，提案法の方が頑健性を持つ

Fig. 4　The relationship between misclassification rate and the stopping threshold(left).The relationship between cluster purity and the stopping threshold(right).The broken line is the stopping threshold computed from the develop data.

*tistical Association*66(336), Dec. 1971

（平成 xx 年 xx 月 xx 日受付）

陳　伯翰

2013 名古屋大学大学院情報科学研究科博士前期課程了，同年同大学院情報科学研究科博士後期課程入学，現在に至る．会議分析システムに関する研究を行っている．日本音響学会会員．修士（情報科学）．

北岡　教英

1994 京都大学大学院工学研究科修士課程了．同年（株）デンソー入社．1997 から 2000 豊橋技術大学大学院工学研究科博士後期課程在学．2001 同大情報工学系助手．2003 同講師．2006 名古屋大学大学院情報科学研究科助教授．2007 同准教授現在に至る．2009　Nanyang Technological University visiting associate professor．主として音声認識，音声対話，音声インタフェースに関する研究に従事．IEEE，ISCA，日本音響学会，情報処理学会，人工知能学会各会員．博士（工学）．

武田　一哉

1985 名古屋大学大学院工学研究科修士課程了．同年国際電信電話（株）入社．1986（株）ATR 国際電気通信基礎技術研究所．1990KDD 研究所復職（この間 1988 から 1989 マサチューセッツ工科大滞在研究員），1995 名古屋大学大学院工学研究科助教授，現在，同大学情報科学研究科教授．音声符号化，空間音響処理，音声情報処理など音声音響言語処理の研究に従事．日本音響学会理事，IEEE，情報処理学会等各会員．工博．

Abstract In this paper, we propose a method to reduce the diarization error and to improve its performance. It is helpful if we can evaluate whether a cluster includes all the speech segments belonging to a certain speaker. In this paper, histogram of the cosine similarity between cluster's i-vector and that of every segments are used to realize the evaluation. Modified bottom-up clustering algorithm based on this evaluation method is proposed to prevent the over-merge. Experimental results show that the proposed method can improve the diarization accuracy.

Key words Speaker diarization, Bottom-up clustering, Cluster evaluation, High accuracy