Institute of Electronics, Information, and Communication Engineers

情報処理学会研究報告
IPSJ SIG Technical Report

Vol.2012-SLP-94 No.27
2012/12/21

クラスタ選定によるボトムアップ話者ダイアライゼーションの高精度化

陳 伯翰[1,†1,a)]　北岡 教英[1,†1]　武田 一哉[1,†1]

概要：従来, ボトムアップアプローチを用いて話者ダイアライゼーションをするとき, 殆どの場合話者内の音声類似度の変異が, 話者間より小さいことを仮定として扱う. もし実際のデータが仮定を満たさない場合, ボトムアップアプローチは機能しない. 本稿では, 従来のレベル内類似度評価以外に, クラスタ合併時の時間（レベル間）次元の変化を評価基準として加えるボトムアップクラスタリング法を提案する. 実験では, 従来のボトムアップクラスタリング法で停止条件を最適に設定した場合の最大可能精度を上回るの結果を得た.

キーワード：ボトムアップ　クラスタリング　クラスタ選定　高精度

1. はじめに

会議や放送音声において，事前情報を一切持たず「いつ，だれが話しているか」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システムの結果は音声データの分析や資料整理に利用することができる．特定の話者の音声を選択できるので，音声の再生も効率的になる．さらに話者ダイアライゼーションの結果を入力データとして，音声認識システムの性能を上げることも可能であると考えられる．

話者ダイアライゼーションは従来一種のクラスタリング技術として，ボトムアップクラスタリング，または階層的なクラスタリングアプローチ（Hierarchial Agglomerative Clustering ；HAC）を用いることが多い．ボトムアップクラスタリングでは，一回の反復で音声類似度が一番大きいと判断されたクラスタ対を合併する. レベル内のクラスタ間類似度がある閾値より小さい場合に反復を停止する．つまり, 停止条件に至るまで, 毎回合併されるクラスタ対が同じ話者に属さなければならない．しかしながら, 同じレベルにあるクラスタ間の音声類似度のみを合併基準として扱うには限界がある. 音声類似度空間で, 話者内の変異が話者間より大きい場合, 合併すべきクラスタ対を残したまま, 合併すべきでないクラスタ対を合併することがある. 我々はこの問題を解決するため, 他の視点からボトムアップの合併を評価する基準が必要だと考える.

本稿では, 従来のレベル内類似度評価以外に, レベル間の変異も合併の評価基準として用いることを考える. 従来のレベル内類似度評価は合併すべきクラスタ対を選定するのに対して, 提案法は合併すべきでないクラスタ対を選定するのであり, 一種の停止条件のための評価関数としても考えられる. 従来の評価関数と異なる, 提案法で評価する話者内変異は, 話者ごとにそれぞれ違うと考える. これは話者ごとに違う停止条件を設置することを意味する. これにより, 従来のようにクラスタリングを繰り返しあるレベルで停止するのではなく，各レベルでクラスタ毎に停止を判断する. 従来のレベル停止法と区別するため, 我々はこの方法をクラスタ停止法と呼ぶ. 評価の結果, 従来のボトムアップクラスタリングでは達成不可能の精度を実現できた.

2. ボトムアップクラスタリング

話者ダイアライゼーション問題に対して，ボトムアップクラスタリングアプローチを用いるのは主流である [2][3][4] [5].

ボトムアップクラスタリングアプローチの流れは以下のようになる（図 1）

(1) データを K 個のクラスタに分けて初期化
(2) クラスタ対ごとに類似度を計算
(3) 類似度が最も高いクラスタ対を合併
(4) クラスタのモデルを用いてリセグメント
(5) クラスタ数が 1 になるまで（あるいは停止条件を満たすまで）3-5 を繰り返す
(6) 閾値を用いて最適の結果を選択

ボトムアップクラスタリングでは，繰り返し毎にクラス

1　日本音響学会
†1　現在，名古屋大学
Presently with Nagoya Uniersity
a)　bohan.chen@g.sp.m.is.nagoya-u.ac.jp

© 2012 Information Processing Society of Japan

NII-Electronic Library Service

Institute of Electronics, Information, and Communication Engineers

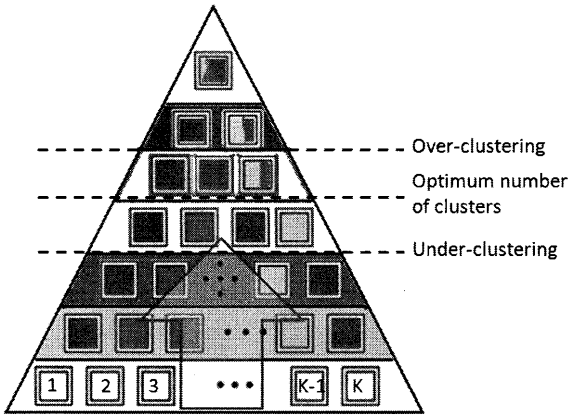

図 1 ボトムアップクラスタリングの概念図, 矢印は合併の方向, 小さい四角がクラスタ, 同じ色のクラスタは同じ話者に属することを示す. また, 色をつけたレベルは, このレベルで, 対応話者の理想的なクラスタができたことを意味する. この図では, クラスタ単位のオーバークラスタリングとアンダークラスタリングが同じレベルで発生した. 従来のボトムアップクラスタリングでは, このようなデータに対応できない

Fig. 1 Bottom-up clustering schema, small squares in the same colour belong to the same speaker. The speaker level ideal cluster appears in the iteration with the same color

タの数を一個減らし，クラスタ構成を更新する．新たなクラスタ構成を用いて，全てのフレームに対して再びラベル付けを行う必要がある．この再ラベル付けの過程をリセグメントと呼ぶ, 各クラスタを GMM でモデル化する場合には,Viterbi アライメントより実現することが多い. しかしながら, リセグメントというステップはあくまでも異なる話者の間の境界を最適化するためのステップである [6]. リセグメントの結果，同じ話者に属するクラスタの間で音声フレームの入れ替わりが発生するが, 各クラスタの話者構成は変わらないと仮定する. 実際, リセグメントを最後に一回しか使わない（あるいは全然使わない）ダイアライゼーションシステムも多く存在する. これが，この仮定がほぼ成り立つことを暗に示している.

したがって, ある初期化されたデータにおいて, 唯一クラスタの話者構成を大きく操作できるのは合併というステップである. ボトムアップクラスタリングでは, 毎回合併されるクラスタ対が類似度評価により決定される. 適切な結果を得るためには, クラスタリングを停止させるまでに合併されるクラスタ対は常に同じ話者に属さなければならない. これは類似度評価関数が

(1) 任意話者 A に属するクラスタ i に対する, 同じ話者 A に属するクラスタ j の類似度が,A と違う話者 B に属するクラスタ k の類似度より高い

(2) 任意話者 A に属するクラスタ i に対する, 同じ話者 A に属するクラスタ j の類似度が, 任意の違う話者 C と D との類似度より高い.

以上二つの条件を満たさなければならないことを意味する. 条件 1 は,

$$\sigma(C_{i\in S_A}, C_{j\in S_A}) > \sigma(C_{i\in S_A}, C_{k\in S_B}) \quad (S_A \neq S_B) \tag{1}$$

と書ける. ここで,$\sigma(,)$ は 2 つのデータ集合の音声類似度を評価する関数,S_A は実際話者 A のフレームの集合,C_i はボトムアップクラスタリング中のあるクラスタ i のフレームの集合を表す. この式は話者認識で求めたいとのと同様の類似度関数である. これによりボトムアップフレームワークで話者 A のクラスタは，他の話者のクラスタと混合されることなく 1 つのクラスタまで合併されることが保証される. 今回の議論では式 (1) のこと常に成り立つと仮定する.

一方条件 2 は

$$\sigma(C_i \in S_A, C_j \in S_A) > \theta > \sigma(S_C, S_D), \quad (S_A \neq S_C \neq S_D) \tag{2}$$

と書ける. この式は合併すべきクラスタ対が完全に合併される前に, 既に 1 つのクラスタとなった話者のクラスタが別の話者のクラスタと合併されないことを保証する. この条件を満たす類似度関数があるならば, あるレベルにおけるすべてのクラスタの間の類似度が閾値 θ より低い場合, これ以上合併する必要がないと判断できる. 今まで提案されている停止条件は多くにこの枠組みで停止時点を判断する [4][2][7].

しかしながら, 音声情報のみを話者ダイアライゼーションの入力情報として扱う時, 音声情報で提供できるの個人性情報には限界があり, 式 (2) が常に成り立つわけではない. 発話には話者の違いと音韻的な違いがあるため，式 (2) が成り立たないことはしばしば起こる．このように話者内の類似度の差より, 話者間の類似度の差の方が少ない場合, ボトムアップアプローチ自体がエラーを発生する（図 1, 色が話者を表す．上から 4 番目のレベルで，合併すべき赤の話者の 2 つのクラスタではなく，既に 1 つになった青と緑の話者クラスタが合併すべきと判断されている．この合併が行われると，以降で通常のやり方で適切なクラスタリングを実現するのは不可能である）.

一回の会話において, ある話者の音声が発話内容や話者個人の状態によるある程度の変化が起こると考えられる. 話者の個人性より，この変化はある範囲に限られるが，話者の個人性がそれぞれ違うので，音声の変化の可能な範囲もそれぞれ違う．そして，話者自身の音声の変化の範囲内であると判断される時のみ，合併が行われる．もし合併されるクラスタ内の音声の変化がその話者の音声の変化の範囲を超えるなら，類似度の評価では最適な合併であっても，

© 2012 Information Processing Society of Japan

NII-Electronic Library Service

Institute of Electronics, Information, and Communication Engineers

合併すべきでない．逆にいえば，例え合併より起こる音声の変化がやや大きいとしても，１人話者の変化として許容できる範囲であれば，合併を許可するべきと考える．したがって，従来のボトムアップクラスタリングのように単純に一番小さい音声類似度の変化を話者内の音声類似度の変化として扱うのは問題がある．

したがって，我々はボトムアップフレームワークで話者ダイアライゼーションをするとき，クラスタ間の音声類似度でなく，各クラスタのクラスタ内の変化の可能範囲も評価基準として考慮すべきと考える．従来のクラスタ間類似度評価とは異なり，クラスタ内の変化の範囲が個人の変化の範囲内であるか否かの評価により合併してもよいか否かが判定できる，もちろんこの評価基準自体も停止条件に用いることができる．

3. クラスタ停止法

実際には話者ダイアライゼーションの結果はいろいろな要因から影響を受ける．本節で議論するのはクラスタ内変化が可能な範囲内か否かを合併を不可とする基準として扱う問題であり．それ以外の要因はクラスタリングに影響しないと仮定する．これはつまり，クラスタ i の純度 $\pi(C_i)$ を

$$\pi(C_i) \equiv \max_K \frac{N(C_i \cap S_K)}{N(C_i)} = 1 - \epsilon_i \tag{3}$$

と定義すると，初期クラスタの平均純度 $\bar{\pi}_0$ は

$$\bar{\pi}_0 \equiv \frac{\sum_i N(C_i^{(0)})(1-\epsilon_i^{(0)})}{N_{all}} \simeq 1 \tag{4}$$

を満たすこと．ここで,$C_i^{(0)}$ は初期クラスタ i のフレームの集合を表す．また，$N_{all} = \sum_i C_i^{(0)}$ である．つまり，初期クラスタはそれぞれほぼ唯一の話者のフレームからなり，純度が高い．

また類似度評価について，今回は問題を簡単化するため式 (1) が常に成り立つことを仮定する．

ボトムアップのフレームワークでは，式 (1) の仮定に従えば，任意話者に対して，この話者に対応するクラスタ同士は他の話者クラスタとの合併より優先して合併されるため，話者のほぼすべてのフレームからなる「理想的な話者クラスタ」がクラスタリングのあるレベルに存在することになる．また，式 (4) の仮定より初期化時の初期クラスタはまだ同一の話者が別のクラスタに分けられているアンダークラスタリング状態である（例外の状況は後で議論する）．故に，いずれの話者も，その対応話者クラスタの状態をクラスタリング全体から見ると，必ず「アンダークラスタリング」，「理想的なクラスタ」，そして更に他の話者と合併されてしまう「オーバークラスタリング」の順番で発生する．

前節の議論で，クラスタリング中，唯一クラスタ構成を変更できるのは合併であることはすでに述べた．すなわち，提案法では「オーバークラスタリング」を生み出す合併を避けることを目的とする．逆に考えれば，「オーバークラスタリング」となる合併の直前までの合併が，すべて有意な合併であり，これらの合併の結果が理想的なクラスタリングに近い結果となると期待される．．

前節で述べたリセグメントは，データ上のクラスタ境界線を最適化させる操作である．フレームごとに，クラスタとの音声類似度に基づいてラベルが付けられる．ここで，同じ話者に属するクラスタは音声類似度が高いので，これらのクラスタ間ではある程度のフレームの交換が発生することが考えられる．故に，ある特定のレベルにおいて，ある話者 A に対して，そのレベル内の話者 A に属するあらゆるクラスタはある確率で，話者 A のすべてのフレームからなるデータプールからフレームを抽出して構成していると見ることができる．話者 A に属するフレーム $f \in S_A$ がクラスタ i に分類される確率は

$$p(C_i|f \in S_A) = \frac{p(C_i)p(f \in S_A|C_i)}{p(f \in S_A)}. \tag{5}$$

となる．ここで,$p(C_i) = N(C_i)/N_{all}$, $p(f \in S_A|C_i) = N(C_i \cap S_A)/N(C_i)$, $p(f \in S_A) = N(S_A)/N_{all}$ である.$N()$ はデータ集合内のフレーム数を表し，N_{all} は入力データのフレーム数 $N_{all} = \sum_i C_i$ を表す．式 (5) に代入すると

$$p(C_i|f \in S_A) = \frac{N(C_i \cap S_A)}{N(S_A)}, \tag{6}$$

となる．正しい分類と正しくない分類を区別するため，ここである話者 A に属するクラスタ i に対して以下の α を定義する

$$\alpha_A^{(i)} \equiv p(C_i \cap S_A|S_A) \tag{7}$$

$\alpha_A^{(i)}$ はクラスタのアンダークラスタリング程度を反映する量である．話者 A に属するクラスタ i に関して，ある一回の合併前後の $\alpha_A^{(i)}$ の変化量 $\Delta\alpha_A^{(i)}$ のことを考えると式 (6) を用いて

$$\begin{aligned}\Delta\alpha_A^{(i)} &= \alpha_A^{(i+j)} - \alpha_A^{(i)} = p(C_{i+j}|S_A) - p(C_i|S_A) \\ &= \frac{N(C_i) - \epsilon_i N(C_i) + N(C_j \cap S_A)}{N(S_A)} - \frac{N(C_i) - \epsilon_i N(C_i)}{N(S_A)} \\ &= \frac{N(C_j \cap S_A)}{S_A}\end{aligned} \tag{8}$$

で与える．ここで ϵ_i は式 (3) クラスタ i の純度を表す量である．クラスタ i とクラスタ j は互い ni 重なり合う部分がないため，$N(C_{i+j}) = N(C_i) + N(C_j)$ になる．正しい合併の場合,$N(C_j \cap S_A) = (1-\epsilon_j)N(C_j)$ である．式 (8) は

© 2012 Information Processing Society of Japan

NII-Electronic Library Service

Institute of Electronics, Information, and Communication Engineers

$$\Delta\alpha_A^{(i)} = \frac{N(C_j) - \epsilon_j N(C_j)}{N(S_A)} \leq \frac{N(C_j)}{N(S_A)} \tag{9}$$

となる．オーバークラスタリングが発生する場合，異なる話者 A と B のクラスタが合併され，$N(C_j \cap S_A) \leq \epsilon_j N(C_j)$，式 (8) より

$$\Delta\alpha_A^{(i)} \leq \frac{\epsilon_j N(C_j)}{S_A} \tag{10}$$

最初の仮定に従って $\epsilon_i \simeq 0, \epsilon_j \simeq 0$．したがって，正しい合併では $\frac{\Delta\alpha}{N(C_J)} \simeq \frac{1}{N(S_A)}$，アンダークラスタリングでは，$N(S_A)$ は常に定数であるため，合併より 1 フレーム増やすと $\Delta\alpha$ の値が一定である（これが α の最初の定義と一致する），オーバークラスタリングになる合併では，$\Delta\alpha$ が 0 に近い．もし一回目の合併が正しいと仮定すれば，次の合併での $\Delta\alpha/N(C_j)$ を計算し，前回と比較するにより，合併を評価することが可能になる．

しかしながら，話者 A のデータは未知であるので，推定する必要がある．最初の仮定に従って，初期クラスタの平均純度は $\bar{\pi}_0 \simeq 1$ とする．故に，特定の話者 A に属する純度の高い初期クラスタを集めることにより，話者 A のフレームを集めた集合を作ることができると考える．この近似を

$$\tilde{S}_A \equiv \sum_j C_i^{(0)} \quad (if \quad \epsilon_i \simeq 0 \quad then \quad C_i^{(0)} \in S_A) \tag{11}$$

のように行う．あるクラスタ i から，合併によりこのクラスタに至った初期クラスタまでさかのぼる．そのように選択した初期クラスタ $C_j^{(0)}$ に含まれるデータを用いて，話者データプールの近似 $\tilde{S}_i$ を得る

$$\tilde{S}_i \equiv \sum_j p(C_i|C_j^{(0)}) > \theta_1 \tag{12}$$

クラスタの純度問題を考慮して，選ばれる初期クラスタがクラスタ i とある程度のデータ共有が必要と考える．故に θ_1 を閾値として設置する，値は初期クラスタの平均純度と関連する．この近似話者データプールを用いて合併を評価したを例を図 3．示す．一回目の合併が常に正しいと考え，後ろの合併の $\frac{\Delta\alpha_A^{(i)}}{N(C_j)}$ が前の合併に近い場合は正しいと，逆に 0 に近い場合（三回目の合併）は正しいくないと判断する．この時，この合併をやめて，他のクラスタ対を合併する．

この提案法では，最初の合併を常に正しい合併（可能範囲内の変化）と考える．そして合併より発生する近似話者データプールの変化を用いて話者内の類似度変化を表現する．しかしながら，もし最初から合併する必要がないクラスタが存在すれば，提案法が適応できないと考える．こういう

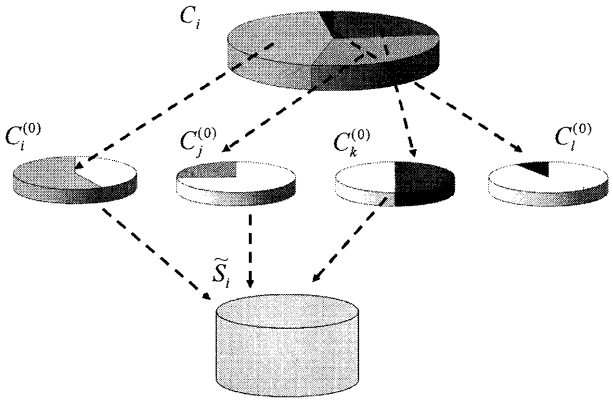

図 2 クラスタ i の対応話者データプールを得る方法．違うレベルにおける同じデータを同じ色で示す．黒の部分はデータが式 (12) を満たさないため，初期クラスタ l のデータを近似話者データに用いていない．

Fig. 2 Estimation of the approximated speaker data pool for cluster i. Same frames in different iteration have been set to the same colors. The frames in black do not satisfied Equ.(8). So we ignored initial cluster l from the approximated speaker data pool

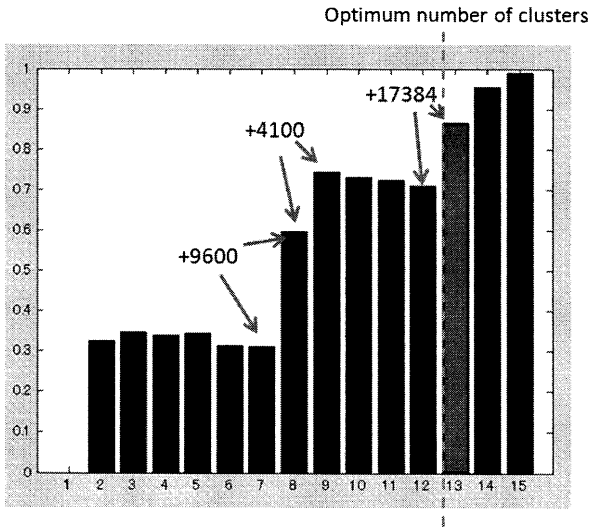

図 3 合併の評価の例，数字は合併により増えるフレームの数，3 回の合併で 100 フレームから得られる α の改善量が左から 3.0×10^{-3}，3.6×10^{-3}，8.9×10^{-4}（実際この合併を許可しない）で，赤い棒がクラスタ数が最適とするラベル

Fig. 3 Example of the proposed method. The number means how many numbers of frames changed by a combination. 100 frames can cause (from the left) 3.0×10^{-1}%, 3.6×10^{-1}%, 8.9×10^{-2}% change for the $\Delta\alpha$. Red bar indicates the level with optimum number of clusters

状況も対応するため，今回では最初（初期レベルを除いて最初なレベル）から

$$p(C_i|\tilde{S}_i) > \theta_2] \tag{13}$$

© 2012 Information Processing Society of Japan

NII-Electronic Library Service

Institute of Electronics, Information, and Communication Engineers

を満たすクラスタを合併する必要がないクラスタと判断する. ここでの θ_2 もまた初期平均純度と関わる閾値である.

まとめて, 今回提案するボトムアップクラスタリング法の流れは以下のようになる.

(1) データを K 個のクラスタに分けて初期化
(2) クラスタ対ごとに類似度を計算
(3) 類似度の高い順でクラスタごとに $\Delta\alpha$ を計算
(4) 合併を許可されるのクラスタ対の中, 類似度が一番高いクラスタ対を合併
(5) クラスタのモデルを用いてリセグメント
(6) 合併を許可されるクラスタ対が 0 になるまで 3-5 を繰り返す

4. 評価実験

AMI meeting corpus の会議録音データ [8] の中の 8 の会議を用いて評価実験を行った. 話者ダイアライゼーションの基本性能を評価するため, 無音区間とオーバーラップ区間（複数の話者が同時に話している）は人手より排除した. データ全体の長さは 3.6 時間で, 無音区間とオーバーラップを排除した後の長さは 2.26 時間である. データ毎に参加メンバーは異なる, 参加人数は全て 4 人である. 初期クラスタの数 K は経験的に $K = 16$ に設定した. 初期クラスタの分け方は, 最も単純なやり方, 入力データを時間順に等間隔に K 分割して, それぞれを 1 つの初期クラスタとして扱う [2]. 得られた初期クラスタの平均純度 π_0 の平均値は 12.4% である. 今回の実験では $\theta_1 = 20\%$,$\theta_2 = 85\%$ と設定した. 提案法はボトムアップクラスタリング向きの改良案とも考えられ, ボトムアップクラスタリングに適用できるの類似度関数であればよいと考える. 本研究では文献 [9] の Binary Key 法を類似度関数として用いる. 実験で比較するのは同じ類似度関数を用いて, 従来のボトムアップクラスタリングで得られるダイアライゼーションエラーレート（DER）の最小となるレベルを手動で選択して得るレベル選択で得られる最高性能の DER の値である. 実験の結果を図 4 に示す. データが式 (2) を満たす場合, 停止条件として提案法を用いる場合クラスタリングが最適なレベルで停止される. データが式 (2) を満たさない場合（8 個中の 2 個）, 通常のボトムアップクラスタリングより上回るの結果を得る.

5. まとめと今後の課題

本報告では, 従来の話者認識のように, なるべく話者内の音声類似度の変化を排除した特徴を用いて, クラスタリングすることを前提とするのではなく, 話者内の音声類似度変化がある程度起こるという現状を認め, そして変化の範囲の推定を用いてクラスタ毎に停止条件を適用する方法を提案した. 実験より提案法が従来のボトムアップクラスタリングで可能な最高の精度より上回る結果を得た.

提案法の判定条件を導き出すため, 類似度関数と初期クラスタ等についていくつかの前提条件を仮定した. 仮定と大きく異なる場合, 提案法は良好に動作しない. この点について改良する必要がある.

更に, 今後の課題として, 今最先端の類似度関数（例えば i-vector）と提案法を組み合わせて, ダイアライゼーションシステムを構築し, 提案法の有効性を示す予定である.

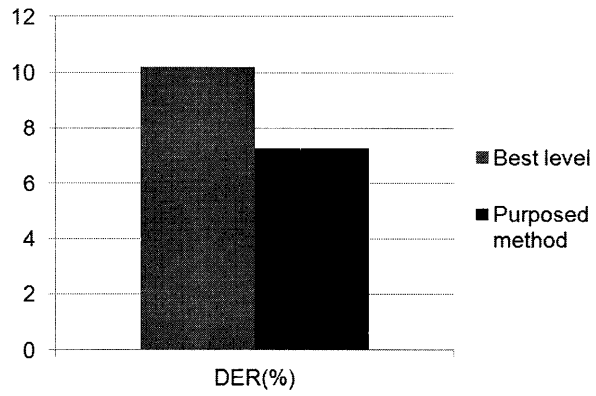

図 4 従来のボトムアップ法で得られる最大可能精度（青）と提案法で得る精度（黒）. 提案法が殆どの場合, ボトムアップクラスタリングを最大精度のレベルに停止させる. 更に式 (2) が満たさない場合, 最大可能精度より上回るの結果を得る

Fig. 4 DER from a common bottom-up clustering(blue) and from the purposed method(black). Purposed method can stop the clustering at the level which has the highest accuracy. When the data do not satisfied (2) the purposed method make the diarization accuracy even higher than the best one which a common method can.

参考文献

[1] Reynolds, D. A., and Torres-Carrasquillo, P., "Approaches and Applications of Audio Diarization," *Special Session on HLT, ICASSP 2005.*

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

[3] Friedland, G., Jamin, A., Imseng, D., Anguera, X., Gottlib, L., Huijbregts, M., Knox, M. T., and Vinyals, O.,, "The ICSI RT-09 Speaker Diarization System," *Audio, Speech, and Language Processing vol. 20, on. 2, pp. 371-381, Feb. 2012.*

[4] Zhu, X., Barras, C., Lamel, L., and Gauvain, J.-L., "Multi-stage speaker diarization for conference and lecture meetings," *Multimodal Technol. Perception of Humans: Int. Eval.Workshops CLEAR 2007 and RT 2007, Baltimore, MD, May 8-11, 2007, Revised Selected Papers, Berlin, Heidelberg: Springer-Verlag, 2008, pp. 533-542.*

[5] Reynolds, D. A., and Torres-Carrasquillo, P.,, "The MIT Lincoln Laboratory RT-04F Diarization Systems: Applications to Broadcast Audio and Telephone Conversations," *DARPA EARS RT-04F Workshop, White Plaines, NY, Nov. 8-11, 2004*

[6] Anguera, X., Bozonnet, S., Evans, N., Fredouille, C.,

© 2012 Information Processing Society of Japan

NII-Electronic Library Service

Institute of Electronics, Information, and Communication Engineers

and Friedland, G. "Speaker Diarization: A Review of Recent Research" *Audio, Speech, and Language Processing, vol. 20, no. 2, pp 356-370, Feb. 2012.*

[7] Nguyen, T. H., Chng. E. S., and Li, H. "T-test distance and clustering criterion for speaker diarization,' *Interspeech, Brisbane, Australia, 2008.*

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

[9] Anguera, X., and Bonastre, J.-F., "A novel speaker binary key derived from anchor models," *Interspeech, 2010*

© 2012 Information Processing Society of Japan

NII-Electronic Library Service