Vol.2010-SLP-84 No.13
2010/12/21

# 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

西田昌史[†1]　細川光政[†1]　山本誠一[†1]

従来のアンカーモデルではアンカーモデルを無作為に選択しており，多数のモデルを必要としていた．それに対して，本研究ではアンカーモデルを最適化するために，GMM 間の KL 距離をベースとして非負値行列因子分解に基づいたアンカーモデルのクラスタリング手法を提案した．本手法により，UBM を初期モデルとして MAP 推定で学習したアンカーモデルに対して，音響的に類似した話者をクラスタリングすることで効率的な認識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ならびに GMM 間の KL 距離による階層的クラスタリング手法との話者認識実験を行った結果，提案手法は認識精度を低下させることなく，従来手法に比べて大幅にモデル数を削減することができた．

# Speaker Recognition Based on Speaker Clustering of Anchor Model Using Non-negative Matrix Factorization

Masafumi Nishida,[†1] Mitsumasa Hosokawa[†1] and Seiichi Yamamoto[†1]

In conventional methods based on anchor model, it was needed many models and selected anchor models at random. We proposed a clustering method based on non-negative matrix factorization for anchor models trained by UBM-MAP. The proposed method can perform the speaker clustering based on KL divergence between GMMs. We conducted speaker recognition experiments using a clustering method based on BIC, agglomerative clustering method using KL divergence between GMMs, and proposed method. As a result, the proposed method was able to reduce the number of models without decreasing the recognition accuracy compared with the conventional methods.

## 1. はじめに

近年，セキュリティのための生体認証としての話者認識や，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た話者分類に関する研究がさかんに行われている．

従来の話者認識の手法としては，登録話者の音声データから抽出した特徴を統計的にモデル化する GMM（Gaussian Mixture Model）がよく用いられてきた．この GMM による手法では多くの学習データを用いれば高い認識精度が得られるが，学習データ量が少ない場合には認識精度が劣化してしまう．それに対して，登録話者のモデルを仮定せずに登録話者以外の多くの話者モデル (GMM) を用いることで，少量の学習データで認識を行うアンカーモデルという手法が提案されている．このアンカーモデルに基づいた手法は，会議や討論などの音声データベースを対象とした話者インデキシング[1)2)] や話者照合[3)] に用いられている．また，アンカーモデルによる話者空間を判別分析などを用いて構成する手法[4)] や，各話者ごとに学習した音素モデルをアンカーモデルとする話者識別手法が提案されている[5)]．

従来のアンカーモデルによる手法では，多くの話者モデルを用意することで高い認識精度を実現しているが，アンカーモデルを選択する基準については検討されておらず無作為に選択している．そこで，これまで我々は，認識対象の話者を識別するのに有効なアンカーモデルを構成する手法として，GMM 間の KL 距離[6)7)] に基づいて階層的にクラスタリングする手法を提案した[8)]．本手法は，アンカーモデルを直接 GMM にてモデル化し，GMM 間の KL 距離をもとに音響的に類似した話者をマージし，識別に有効なアンカーモデルを効果的に生成することができる．しかし，ボトムアップな階層的クラスタリングでは，モデル数が多いほどモデル間の距離の算出やクラスタリングの処理コストがかかってしまう．

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[10)] 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，より効果的なアンカーモデルによる話者認識を実現する．また，これまで取り組んできたアンカーモデルの学習では直接 GMM を学習していたが，UBM(Universal Background Model) を初期モデルとした MAP 推定[11)] によりアンカーモデルを学習する手法を適用した．

†1 同志社大学
Doshisha University

Vol.2010-SLP-84 No.13
2010/12/21

これを踏まえて本研究では，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[12] ならびに，GMM 間の KL 距離による階層的クラスタリング手法との比較実験を行う．

以降，2 章にて通常のアンカーモデルによる認識，3 章にて BIC に基づくクラスタリング，4 章にて GMM 間の KL 距離に基づく階層的クラスタリング，5 章にて提案手法である NMF に基づくクラスタリング，6 章にて評価実験により得られた結果，7 章にてまとめと今後の課題について述べる．

## 2. アンカーモデルによる話者認識

### 2.1 Universal Background Model を用いたアンカーモデルの学習

本研究では，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，各アンカーモデルの学習データにより MAP 推定を行うことで話者モデルを学習する．

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

$$n_i = \sum_{t=1}^{T} Pr(i|x_t) \tag{2}$$

$$E_i(x) = \frac{1}{n_i} \sum_{t=1}^{T} Pr(i|x_t) x_t \tag{3}$$

ここで，$x_t$ は各アンカーモデルの学習データ，$T$ は各アンカーモデルの学習データの総フレーム数，$M$ が UBM の混合分布数，$w_i$ は UBM の各混合分布の重みを表している．

以上で求めた結果をもとに，UBM の各混合分布の重み $w$，平均 $\mu$，分散 $\sigma^2$ を以下の式により適応する．

$$\hat{w}_i = [\alpha_i n_i/T + (1-\alpha_i) w_i]\gamma \tag{4}$$

$$\hat{\mu}_i = \alpha_i E_i(x) + (1-\alpha_i)\mu \tag{5}$$

$$\hat{\sigma_i}^2 = \alpha_i E_i(x^2) + (1-\alpha_i)(\sigma_i{}^2 + \mu_i{}^2) - \hat{\mu_i}^2 \tag{6}$$

ここで，$\gamma$ は混合分布の重みの総和を制御する係数，適応データの割合を制御する係数は

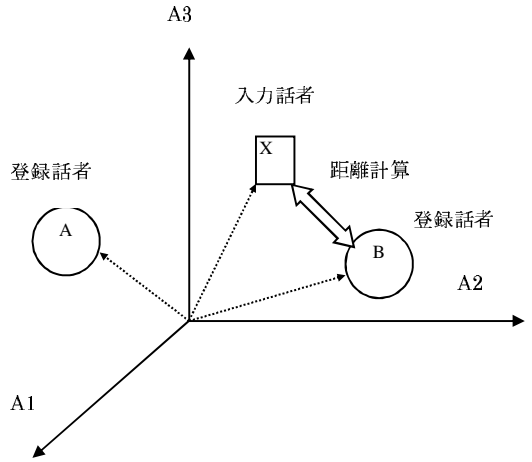


**図 1** アンカーモデルによる認識

$\alpha_i = n_i/(n_i + r)$ により求める．

### 2.2 アンカーモデルによる認識

アンカーモデルによる話者認識では，認識対象以外の多くの話者の音声データを集め，話者ごとに GMM を学習する．本研究では，UBM を初期モデルとした MAP 推定により話者モデルを学習する．そして，入力された発話と認識対象以外の話者ごとの尤度を求め，この尤度を話者ベクトルの要素とし登録話者のベクトルと入力話者のベクトル間のユークリッド距離にて認識を行う手法である．また，話者ベクトルは発話間のスコア変動を抑えるために平均 0，分散 1 に正規化される．本手法では，入力音声から話者ベクトルを生成し，識別対象話者の登録音声の話者ベクトルとのユークリッド距離を求め，距離が最短となる話者ベクトルをもつ話者が入力音声の話者であると識別する．

図 1 に 3 次元での話者ベクトル空間の概念図を示す．それぞれの軸は，認識対象以外の話者であるアンカーモデルを表している．

GMM に基づく従来の話者認識手法では，識別対象話者の話者モデルを作成する必要があり，学習用の発話が複数文必要であった．それに対してアンカーモデルによる認識手法では，識別対象話者のためにモデルを学習する必要がなく，話者ベクトルの生成には 1 発話程

度あればよい．

しかしながら，認識対象以外の不特定多数の話者の音声データからアンカーモデルを作成する必要があり，モデル数が多いほど処理時間がかかってしまうという問題がある．また，従来アンカーモデルは実験的に選択されており，登録話者を識別するにあたりどのような話者をアンカーモデルとして用意すべきかが重要である．

## 3. BIC に基づくアンカーモデルのクラスタリング

BIC(Bayesian Information Criterion) に基づくアンカーモデルのクラスタリング手法について述べる．BIC は，ベイズ推定に基づいてモデル選択を行う基準として用いられている．各話者のデータ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．この手法では，2 つの話者が似た特徴を持つと仮定した場合と，異なる特徴を持つと仮定した場合の BIC 値との差分に基づいて判定する．

2 つの話者をマージしたときの共分散行列を $\Sigma_0$，1 人目の話者の共分散行列を $\Sigma_1$，2 人目の話者の共分散行列を $\Sigma_2$，各話者のフレーム数を $N_i$，特徴ベクトルの次元数を $d$ とすると BIC 値の差分は式 (7) により求まる．係数 $\alpha$ は，BIC を用いた最適化で導入される重み係数であり，実験的に決める必要がある．

$$\begin{aligned}&\Delta BIC\\&= \frac{N_1+N_2}{2}\log|\Sigma_0| - \frac{N_1}{2}\log|\Sigma_1| - \frac{N_2}{2}\log|\Sigma_2|\\&-\alpha\frac{1}{2}(d+\frac{d(d+1)}{2})\log(N_1+N_2)\end{aligned} \tag{7}$$

BIC に基づくクラスタリングでは，式 (7) の ΔBIC 値が負であれば 2 つの話者をマージする．その際，BIC 値が最も大きい話者間から順次マージしていく．全ての発話間で BIC 値が正になれば，どの発話もマージすべきでないとしてクラスタリングの処理を終了する．

以上の処理により得られたクラスタごとに，UBM を初期モデルとした MAP 推定により GMM を学習してアンカーモデルとする．

## 4. KL 距離に基づくアンカーモデルの階層的クラスタリング

本章では，GMM 間の KL 距離に基づくアンカーモデルの階層的クラスタリング手法について述べる．本手法では，アンカーモデルをクラスタリングするにあたり，GMM 間の KL 距離を用いた．なお，GMM は UBM を初期モデルとした MAP 推定により学習した．一般的に，KL 距離は単一ガウス分布間の距離尺度であるので，本研究では式 (8) のように混合分布間の距離尺度に拡張して用いた．

$$d(t,s) = \sum_{p=1}^{m} w_p \max_q KL(p,q) \tag{8}$$

$$\begin{aligned}KL(p,q) = \sum_{i=1}^{d}\{&\frac{{\sigma^2}_{pi} - {\sigma^2}_{qi} + (\mu_{qi}-\mu_{pi})^2}{{\sigma^2}_{qi}}\\&+\frac{{\sigma^2}_{qi} - {\sigma^2}_{pi} + (\mu_{qi}-\mu_{pi})^2}{{\sigma^2}_{pi}}\}\end{aligned}$$

ここで，p は話者 t のモデルの分布番号，q は話者 s のモデルの分布番号，m は話者モデルの混合分布数，$w_p$ は混合分布の重み，n は特徴ベクトルの次元数を示している．また，$\mu$，$\sigma$ は混合分布の平均ベクトル，共分散行列の要素を表している．

次に，GMM 間の KL 距離に基づいたアンカーモデルの階層的クラスタリング手法について述べる．本研究では，すべてのアンカーモデル同士の GMM 間の KL 距離を求め，最小距離が閾値よりも小さければそれらの話者同士をマージする．その後，どれにもマージされなかった話者とクラスタとの KL 距離を比較し，最小距離が閾値よりも小さければその話者をクラスタに加える．単独の話者のクラスタリングが終わってから，クラスタ同士の KL 距離を比較し，最小距離が閾値よりも小さければそれらのクラスタをマージする．このようにすることで，特定のクラスタに話者が集中してクラスタリングされないように対応した．一連のクラスタリング処理が終わって得られたクラスタごとに GMM を UBM-MAP により再学習して，これらをアンカーモデルとして認識を行う．

クラスタリングの処理の流れを以下に示す．

(1) アンカーモデルの GMM 間の KL 距離を全てのモデル間で計算
(2) KL 距離が最小となるモデル同士を新たなクラスタとする
(3) 2 でマージしたモデル以外で KL 距離が最小となる話者を選出．全てのモデル同士の KL 距離が閾値より大きくなるまで 2，3 を繰り返す．
(4) 3 までの処理で得られたクラスタと単独モデルの KL 距離が最小となるクラスタを探す．ここで，クラスタと単独モデルとの距離は，クラスタ内の各 GMM との KL 距離の平均距離により求める．この距離が閾値より大きくなるまで処理を繰り返す．

(5) クラスタ同士の KL 距離を比較し，距離が最小となるクラスタ同士をマージする．ここで，クラスタ間の距離はクラスタ内の各 GMM 間の KL 距離の平均距離により求める．この距離が閾値より大きくなるまで処理を繰り返す

(6) 以上より得られたクラスタごとに UBM-MAP により GMM を再学習し，これらをアンカーモデルとする．

## 5. NMF に基づくアンカーモデルのクラスタリング

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．

$$V \approx WH \tag{9}$$

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．

$$D(V||WH) = \Sigma_{ij}(V_{ij} log \frac{V_{ij}}{(WH)_{ij}} - V_{ij} + (WH)_{ij}) \tag{10}$$

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．

$$W_{ia} \leftarrow W_{ia} \frac{\Sigma_\mu H_{a\mu} V_{i\mu}/(WH)_{i\mu}}{\Sigma_\nu H_{a\nu}} \tag{11}$$

$$H_{a\mu} \leftarrow H_{a\mu} \frac{\Sigma_i W_{ia} V_{i\mu}/(WH)_{i\mu}}{\Sigma_k W_{ka}} \tag{12}$$

本研究では，全アンカーモデルの GMM 間の KL 距離を求めてそれらを要素にもつ行列を $V$ とする．ここで，アンカーモデルの GMM は UBM を初期モデルとした MAP 推定により学習を行っている．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 $H$ の列は各アンカーモデルを表しているので，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アンカーモデルは同じ性質をもつと考え，クラスタリングを行う．これにより得られたクラスタごとに UBM-MAP により GMM を再学習して，これらを最終的なアンカーモデルとする．

このように，アンカーモデル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ことで，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．

## 6. 評価実験

本章では，通常のアンカーモデルによる話者認識，アンカーモデルのクラスタリングにおける比較手法として BIC，KL 距離による階層的クラスタリングならびに提案手法である NMF に基づくクラスタリングによる話者認識実験を行う．

### 6.1 実験条件

本研究では，NTT の話者認識用データベースを用いて話者認識実験を行った．話者 30 名（男性 21 名・女性 9 名）が約 1 年間の 7 時期（1990 年 8 月・9 月・12 月，1991 年 3 月・6 月・9 月，1992 年 3 月）に発声した各時期 10 文章のデータで，各文章における 3 種類の発声速度 (普通，遅い，速い) の計 30 発話である．

UBM ならびにアンカーモデルの学習データには，認識対象のデータと異なる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た．1 人あたり無音区間を除いたおよそ 60 秒の発話で，218 名の話者のデータを UBM の学習，それとは異なる 500 名の話者をアンカーモデルの学習に用いた．なお，300ms 以上の無音区間を基準に発話を分割した．また，UBM の混合分布数は 256 とした．

アンカーモデルによる認識手法では，学習データとして最初の時期 90 年 8 月の普通の速さ 1 発話を用いて行い，認識では全 7 時期の学習とは異なる 5 文の 3 速度の 15 文章で，話者ごとに合計 105 発話を用いた．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．

### 6.2 実験結果と考察

通常のアンカーモデルによる認識において，アンカーモデル数を 300,400,500 に変化させて認識を行った結果を図 2 に示す．

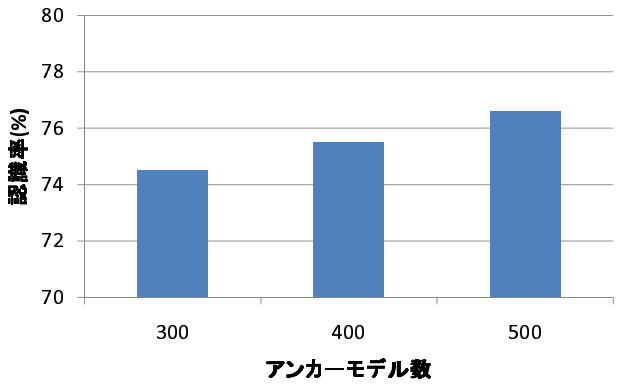


**図 2** 従来のアンカーモデルによる認識結果

アンカーモデル数が 300 のとき 74.5%，400 のとき 75.5%，500 のとき 76.6% という認識率が得られ，アンカーモデル数が 500 のときに最も高い認識精度が得られた．そこで，以下のアンカーモデルのクラスタリング手法においては，初期のアンカーモデル数を 500 として実験を行った．

図 3 に各クラスタリング手法におけるアンカーモデルによる認識結果を示す．図中の $BIC$ は従来手法である BIC に基づくクラスタリング，$KL$ は従来手法である GMM 間の KL 距離による階層的クラスタリング，$NMF$ は提案手法である NMF に基づくクラスタリングによる結果を示している．

BIC に基づくクラスタリングでは 73.8%，GMM 間の KL 距離による階層的クラスタリングでは 75.5%，提案手法である NMF に基づくクラスタリングでは 75.0%の認識率が得られた．

図 4 に各クラスタリング手法により得られたアンカーモデル数を示す．なお，図 3 における各手法の認識率は，図 4 のアンカーモデル数のときの結果を示している．

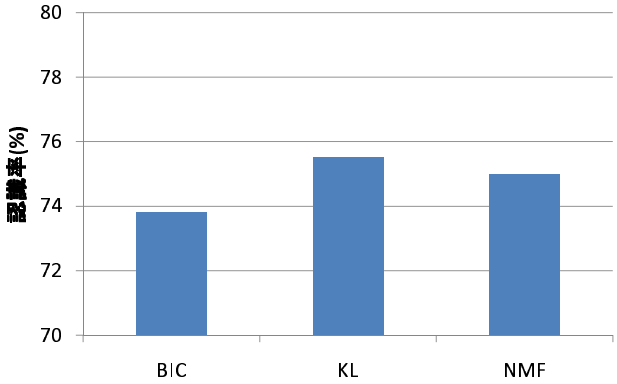


**図 3** 各クラスタリング手法による認識結果

初期のアンカーモデル数に比べて，BIC に基づくクラスタリングでは 52%，GMM 間の KL 距離による階層的クラスタリングでは 45%，提案手法である NMF に基づくクラスタリングでは 76%のアンカーモデル数を削減することができた。

以上の結果から，提案手法は無作為に選択された 400 個のモデル数での通常のアンカーモデルによる認識精度と同じ結果が得られていることから，400 個を初期モデルと考えた場合 70%の割合でアンカーモデル数を削減することができ，提案手法が認識精度を劣化させることなく最もモデル数を減らすことができた．また，提案手法は全アンカーモデル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となり，従来法に比べて大幅な計算コストを抑制することができ，提案手法の有効性を示すことができた．

## 7. おわりに

本研究では，非負値行列因子分解に基づくアンカーモデルのクラスタリングによる話者認識手法を提案した．また，アンカーモデルの学習には UBM を初期モデルとした MAP 推定により実現した．

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，GMM 間の KL 距離に基づく階層的クラスタリングとの比較実験を行った．その結果，提案手法はアンカーモデル数が 400 のときと同等の認識精度が得られ，初期モデルを 400 としたときに

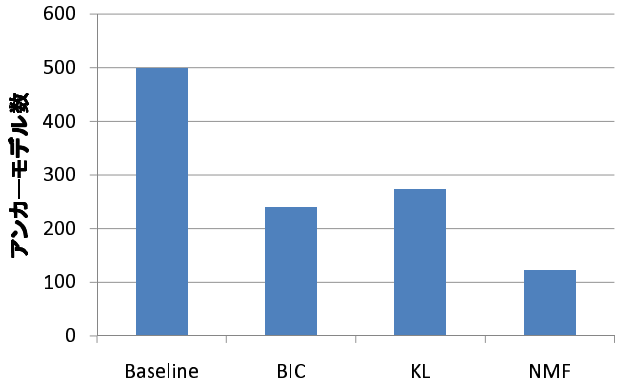


**図 4** 各クラスタリング手法により得られたアンカーモデル数

比べて 70%の割合でアンカーモデル数を削減することができ，従来手法に比べてより多くモデル数を削減することができた．本手法は，モデル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，大幅な計算コストを削減することができ，アンカーモデルのクラスタリングにおいて有効であることがわかった．

今後は，より多くの認識対象話者での評価を行い，さらに有効なアンカーモデルの構成法について検討する．また，本手法を話者クラスタリングによる話者適応や，討論などの複数話者の音声データに対する話者分類への適用などについて検討する予定である．

## 参 考 文 献

1) D. Sturim, D. Reynolds, E. Singer, and J. Campbell: Speaker indexing in large audio databases using anchor models, Proc. ICASSP, Vol.1, pp.429-432, (2001).
2) 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).
3) Y. Yang, M. Yang, and Z. Wu: A Rank based Metric of Anchor Models for Speaker Verification, Proc. ICME, pp.1097-1100 (2006).
4) Yassine Mami, Delphine Charlet: Speaker recognition by location in the space of reference speakers, Speech Communication 48, pp.127-141 (2006).
5) 小坂哲夫，赤津達也，加藤正治，好田正紀: 音素モデルを用いた話者ベクトルに基づく話者識別, 電子情報通信学会論文誌，Vol.J90-D No.12，pp.3201-3209 (2007).
6) 西田昌史，堀内靖雄，市川熹，河原達也: 統計的モデル選択に基づくクラスタリングを用いた話者適応，日本音響学会講演論文集，2-11-5，pp.109-110 (2004).
7) J.E.Rougui, M.Rziza, D.Aboutajdine, M.Gelgon, and J. Martinez: Fast Incremental Clustering of Gaussian Mixture Speaker Models for Scaling Up Retrieval in On-line Broadcast, Proc. ICASSP, vol.5, pp.521-524 (2006).
8) 細川 光政，西田 昌史，山本 誠一: KL 情報量による Anchor model の階層的クラスタリングに基づく話者認識, 情報処理学会研究報告, 2010-SLP-82, No.14, pp.1-6 (2010).
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，山本 誠一: 非負値行列因子分解に基づく Anchor Model のクラスタリングによる話者認識, 日本音響学会講演論文集, 2-9-14, pp.75-76 (2010).
11) 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).
12) 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).