社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

MFCC 領域における GMM クラスタリングを併用した Non-negative Matrix Factorization による雑音環境下音声認識

藤垣健太郎† 柏木 陽佑† 齋藤 大輔† 峯松 信明† 広瀬 啓吉†

† 東京大学 〒 113–8656 東京都文京区本郷 7–3–1
E-mail: †{fujigaki,kashiwagi,dsk_saito,mine,hirose}@gavo.t.u-tokyo.ac.jp

あらまし 雑音環境下音声認識において，事例ベースの特徴量強調として Non-negative Matrix Factorization (NMF) を用いた手法が検討されている．スペクトル領域において雑音が加算性であることを利用し，雑音重畳音声のスペクトルを多数の音声基底，雑音基底とそのスパースな重み行列に分解することでクリーン音声を再構成する手法である．従来の NMF では，初期値を与え，これを繰り返し更新することで最終結果を得る．このとき，その音声の音素を教師として利用することができれば，基底や重み行列を音素依存で推定でき，より高精度に計算することが期待される．しかし，音声認識のタスクにおいて音素は認識すべき対象であり，事前には得られない．そこで本稿では，MFCC 領域での Gaussian Mixture Model (GMM) クラスタリングを併用した NMF を提案する．音素情報の代わりに，MFCC 領域における GMM クラスタリングによって得られたクラス情報を用いて基底を準備することで，従来の NMF に比べて認識率を向上できることを示す．

キーワード 雑音環境下音声認識，雑音抑圧, 特徴量強調，NMF，GMM クラスタリング

Noise robust speech recognition by non-negative matrix factorization using GMM clustering in MFCC domain

Kentaro FUJIGAKI†, Yosuke KASHIWAGI†, Daisuke SAITO†, Nobuaki MINEMATSU†, and Keikichi HIROSE†

† The University of Tokyo, 7–3–1, Hongo, Bunkyo-ku, Tokyo, 113–8656 Japan
E-mail: †{fujigaki,kashiwagi,dsk_saito,mine,hirose}@gavo.t.u-tokyo.ac.jp

Abstract Exemplar-based feature enhancement by non-negative matrix factorization (NMF) was proposed for noise-robust speech recognition. When we consider only additive noises, we can decompose a noisy speech spectrum into a linear but sparse combination of speech and noise bases. In the conventional NMF, decomposition is unsupervised. If we can give the phoneme sequence of an input utterance to the NMF processing, it is surely possible to realize much more precise decomposition. However, in the task of speech recognition, the phoneme sequence is unknown and unavailable. In this paper, therefore, we introduce unsupervised GMM clustering and classify each input frame by using GMM indexes. For NMF, speech bases are built separately for each GMM index. Experiments show that our proposed method of combining NMF with GMM clustering gives higher robustness of recognizing noisy speech than the original NMF.

Key words robust speech recognition, noise surpression, feature enhancement, NMF, GMM clustering

1. はじめに

現在，スマートフォンやカーナビゲーションをはじめ，音声認識技術が幅広く用いられるようになってきた．しかし，実環境で用いる場合，環境雑音の影響によって音響モデルと実際の入力音声との間に音響的なミスマッチが起こり，認識率が低下してしまう．環境雑音による影響を低減するために，現在までに様々な手法が提案されてきた [1]〜[6].

環境雑音を低減する手法の一つとして，Gaussian Mixture Model (GMM) クラスタリングを用いた統計モデルベースの特徴量強調が挙げられる．Vector Taylor Series (VTS) 近似によって雑音重畳音声を GMM でモデル化する VTS 強調 [1]

や，雑音重畳音声とクリーン音声の関係を GMM でモデル化する Stereo Piecewise LInear Compensation for Environments (SPLICE) [2] がその代表的な例である．一方で，モデルによらない事例ベースの特徴量強調が提案されている．スペクトル領域において，雑音重畳音声のスペクトルを多数の音声と雑音の基底の重み付け和で表現する Non-negative Matrix Factorization (NMF) による手法がその代表的な例である [3]．また，モデルベースの手法に事例ベースの手法を導入した例もある [4].

その中でも，本稿では NMF による事例ベースの特徴量強調に注目する．NMF による特徴量強調では，雑音重畳音声のスペクトルを多数の音声基底，雑音基底とこれらに対する重み行列の積に分解し，クリーン音声を再構成する．行列の積への分解はコスト関数を最小化するようなパラメータ更新を繰り返すことで行う．従来の NMF では，教師なしでパラメータ更新を行うため，クリーン音声を再構成する際にその音声に適さない基底が用いられていることが考えられる．NMF のアルゴリズムは声質変換にも用いられており，声質変換においても不適切な基底が用いられることが問題となっている [7]. [7] においては，音素のクラスタリングを用いた基底のクラスタリングと基底の選択によってその問題を解決し，実際に変換精度が向上することが確認されている．雑音環境下音声認識のための特徴量強調においても，その音声の音素情報を用いて基底を学習，選択することで，その音声に適した基底を用いたクリーン音声の構成が可能になると期待できる．しかし，音声認識のタスクにおいては音素は認識すべき対象であるため，基底の学習，選択の方法を検討する必要がある．そこで本稿では，MFCC 領域での GMM クラスタリングを併用した NMF による特徴量強調を提案する．音素情報の代わりとして MFCC 領域における GMM クラスタリングによって得られたクラス情報を利用することで，教師なしと比べて当該音声に適した基底を用いることが可能になる．まず，学習データに対して GMM でクラスタリングを行い，クラス依存で基底を学習する．また，テストデータに対しても GMM でクラスタリングを行い，基底を選択する．NMF はスペクトル領域での手法であるが，スペクトルは次元間の相関があり，特徴量として GMM クラスタリングには適していない．一方で MFCC は次元間の相関が低いため，GMM クラスタリングに適した特徴量であることが知られており，特徴量強調にも用いられてきた [1], [2]．そこで，MFCC 領域における GMM クラスタリングを導入する．スペクトル領域と MFCC 領域では特徴量の振る舞いが異なるため，2 つの領域で扱うことによる相乗効果も期待できる．

本稿では，まず第 2 節で NMF による特徴量強調を解説する．第 3 節で GMM クラスタリングを併用した NMF による特徴量強調を提案し，第 4 節で音声認識実験により提案手法の有効性を検証する．最後に，第 5 節でまとめる．

2. NMF による特徴量強調

2.1 NMF

NMF は，非負の行列を非負の行列の積に分解して表現するアルゴリズム（$\boldsymbol{A} \rightarrow \boldsymbol{BC}$）である．元の行列 $\boldsymbol{A}$ と積型で表現した $\boldsymbol{BC}$ の距離によるコスト関数を最小化するように繰り返しパラメータ更新を行うことで，分解された行列 $\boldsymbol{BC}$ を得る．加算性雑音の乗った音声を考えた場合，スペクトル領域においてその特徴量は非負であり，雑音は音声に対して加算性として近似できる．そこで，スペクトル領域において NMF を雑音重畳音声に適用することで，音声基底，雑音基底とその重み行列の積に分解することができる．したがって，音声基底，雑音基底それぞれの重み付け和で表現される．ここで，音声基底とその重み行列だけを抽出することでクリーン音声が推定される [3].

入力された雑音重畳音声特徴量 $\boldsymbol{Y} \in \mathbb{R}^{\Omega \times T}$ を基底 $\boldsymbol{H} \in \mathbb{R}^{\Omega \times K}$ とアクティベーション $\boldsymbol{U} \in \mathbb{R}^{K \times T}$ の積に分解することを考える．Ω は特徴量の次元数，T はフレーム数，K は基底数である．事前に学習された基底 $\boldsymbol{H}$ を固定し，アクティベーション $\boldsymbol{U}$ のみを更新する．$\boldsymbol{H}$，$\boldsymbol{U}$ はそれぞれ，音声基底 $\boldsymbol{H}_s$ と雑音基底 $\boldsymbol{H}_n$，音声基底のアクティベーション $\boldsymbol{U}_s$ と雑音基底のアクティベーション $\boldsymbol{U}_n$ の連結により構成される．雑音重畳音声特徴量 $\boldsymbol{Y}$ を以下のように分解し，再構成されたクリーン音声特徴量 $\hat{\boldsymbol{X}}$ を得る．

$$\boldsymbol{Y} = \boldsymbol{HU} = [\boldsymbol{H}_s \ \boldsymbol{H}_n] \begin{bmatrix} \boldsymbol{U}_s \\ \boldsymbol{U}_n \end{bmatrix} \tag{1}$$

$$\hat{\boldsymbol{X}} = \boldsymbol{H}_s \boldsymbol{U}_s \tag{2}$$

パラメータの更新は式 (3) のコスト関数 $D(\boldsymbol{Y}||\boldsymbol{HU})$ を最小化するように行う．

$$D(\boldsymbol{Y}||\boldsymbol{HU}) = d(\boldsymbol{Y}, \boldsymbol{HU}) + ||\boldsymbol{\lambda}. * \boldsymbol{U}||_p \tag{3}$$

第一項の $d(\boldsymbol{Y}, \boldsymbol{HU})$ は雑音重畳音声特徴量と NMF により再構成された特徴量の距離であり，第二項ではアクティベーションのスパース性をコントロールするためにゼロでないアクティベーションに対して $\boldsymbol{\lambda}$ でペナルティを設けている．これにより，アクティベーションがスパースになるように更新される．.*は行列の要素ごとの積を表す．今回は $d(\boldsymbol{Y}, \boldsymbol{HU})$ として式 (4) の Kullback-Liebler (KL) ダイバージェンスを用いた．$Y_{\omega,t}$，$H_{\omega,k}$，$U_{k,t}$ は $\boldsymbol{Y}$，$\boldsymbol{H}$，$\boldsymbol{U}$ の各要素である．

$$d(\boldsymbol{Y}, \boldsymbol{HU}) = \sum_{\omega,t}^{\Omega,T} \Big(Y_{\omega,t} \log \frac{Y_{\omega,t}}{\sum_k^K H_{\omega,k} U_{k,t}} - Y_{\omega,t} + \sum_k^K H_{\omega,k} U_{k,t} \Big) \tag{4}$$

コスト関数 (3) を最小化するアクティベーションの更新式として，式 (5) が得られる．

$$\boldsymbol{U} \longleftarrow \boldsymbol{U}. * (\boldsymbol{H}^\top (\boldsymbol{Y}./(\boldsymbol{HU}))))./(\boldsymbol{H}^\top \boldsymbol{1}_Y + \boldsymbol{\lambda}) \tag{5}$$

$\boldsymbol{U}$ の初期値はすべての要素を 1 として与える．ここで，.*と./は行列の要素ごとの積と商であり，$\boldsymbol{1}_Y \in \mathbb{R}^{\Omega \times T}$ はすべての要素が 1 の行列である．

2.2 Noise-transductive NMF

2.1 項のような通常の NMF による特徴量強調 [3] では，雑音基底が事前の学習により固定されているため，認識する入力音声に含まれる雑音がその雑音基底では表現できない未知の雑音であった場合には抽出できるクリーン音声の精度が劣化してしまう．あらゆる雑音に対応するためには非常に多くの雑音基底を用意する必要があり，計算量も非常に大きくなってしまう．そこで，入力音声から雑音基底を推定する NMF が提案されている [5].

[5] では，予め学習された雑音基底を用いるのではなく，アクティベーションの更新とともに雑音基底を更新することで，

入力音声に含まれる雑音の基底を推定する．これにより雑音基底を柔軟に扱うことができるため，自動的に多くの種類の雑音に対応することができる．しかし，コスト関数 (3) に基づいて雑音基底を更新すると，本来音声基底に割り当てられるべき要素も雑音基底に吸収されてしまう可能性がある．そこでコスト関数を式 (6) としている．

$$D(\boldsymbol{Y}||\boldsymbol{HU}) = d(\boldsymbol{Y}, \boldsymbol{HU}) + ||\boldsymbol{\lambda}.*\boldsymbol{U}||_p + \boldsymbol{\eta}.*d(\boldsymbol{N}, \boldsymbol{H}_n) \tag{6}$$

第 3 項は現在の雑音基底 $\boldsymbol{H}_n$ と基準雑音基底 $\boldsymbol{N}$ の KL ダイバージェンスであり，$\boldsymbol{\eta}$ はその重みである．$\boldsymbol{N}$ は事前に雑音から用意された基底であり，多くの種類の雑音に対応するため Ω 次元すべてにおいて 0 より大きい値であることが望ましい．第 3 項を導入することで，雑音以外の要素を雑音基底が吸収する効果を抑えている．コスト関数 (6) を最小化する更新式は，アクティベーション，雑音基底それぞれについて式 (7)，(8) のようになる．

$$\boldsymbol{U} \longleftarrow \boldsymbol{U}.*(\boldsymbol{H}^\top(\boldsymbol{Y}./(\boldsymbol{HU})))./(\boldsymbol{H}^\top\boldsymbol{1}_Y + \boldsymbol{\lambda}) \tag{7}$$

$$\boldsymbol{H}_n \longleftarrow (\boldsymbol{H}_n.*(\boldsymbol{Y}./(\boldsymbol{HU})\boldsymbol{U}_n^\top) + \boldsymbol{\eta}.*\boldsymbol{N})./(\boldsymbol{1}_Y\boldsymbol{U}_n^\top + \boldsymbol{\eta}) \tag{8}$$

入力音声に含まれる雑音に対し適応的に $\boldsymbol{H}_n$ を推定する本手法の効果は [5] を参照していただきたい．

3. GMM クラスタリングを併用した NMF

従来の NMF では，入力音声に対してラベルを用いず，教師なしでパラメータ更新を行っている．コスト関数 (3) において，アクティベーションがスパースになるよう第二項でペナルティを設けているが，教師なしの場合には値の小さなアクティベーションから成る多数の基底の重み付け和で表現されてしまうこともある．このとき，クリーン音声を再構成する際にその音声に適さない基底が用いられてしまう場合がある．ここでラベルとして音素情報を用いることで，その音声に適した基底を用いたクリーン音声の構成が可能になると期待できる．しかし，音声認識のタスクにおいては音素は認識すべき対象であり，事前には得られない．そこで本稿では，MFCC 領域での GMM クラスタリングを併用した NMF による特徴量強調を提案する．音素情報を用いる代わりに，識別的な情報として GMM クラスタリングによる結果を利用することで，クリーン音声の構成精度の向上を図る．スペクトルは次元間の相関がある一方で，MFCC は次元間の相関が低いため，GMM クラスタリングに適した特徴量であることが知られている．したがって，MFCC 領域における GMM クラスタリングを導入する．スペクトル領域と MFCC 領域では特徴量の振る舞いが異なるため，2 つの領域で扱うことによる相乗効果も期待できる．

3.1 音声基底の学習

MFCC 領域における GMM クラスタリングを利用して音声基底を学習することを考える．音声基底の学習の概要を図 1 に示す．まず，クリーン GMM を用いて学習データのクリーン音声をフレーム単位でクラスタリングする．クリーン GMM は学習データのクリーン音声から構成された GMM であり，混合数を M とする．GMM の事後確率が最大となる混合をそのフレームのクラスとする．次に，各クラスに分類されたフレームから音声基底を学習する．クラス m $(m = 1, \ldots, M)$ に対応する音声基底を $\boldsymbol{H}_m$ とする．また，$\boldsymbol{H}_m$ を結合し，全体の音声基底 $\boldsymbol{H}_s = [\boldsymbol{H}_1 \ \cdots \ \boldsymbol{H}_M]$ とする．

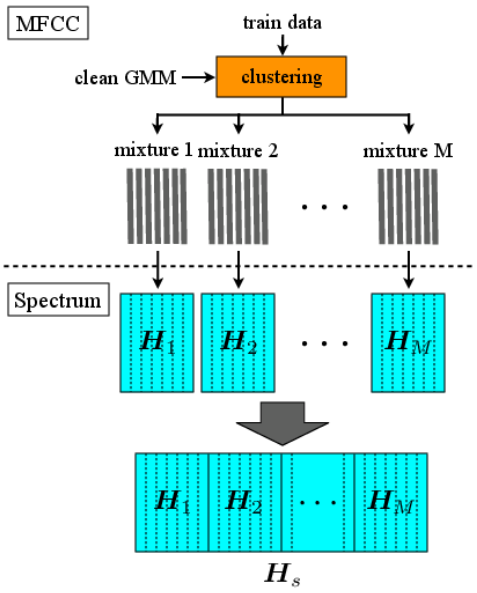

図 1　クラスごとの基底ベクトルの構成

3.2 音声基底の選択

認識時の入力音声に対しても GMM クラスタリングを行い，その結果に基づいて音声基底の選択を行う．具体的には，音声基底の学習時と同様に GMM の事後確率が最大となる混合をそのフレームのクラス m とし，音声基底として $\boldsymbol{H}_m$ のみを用いる．ここで，GMM はクリーン音声の特徴量の GMM を用いるため，クラスタリングのためにはクリーン音声の特徴量が必要になる．そこで今回は，2 段階の NMF を行う．まず全体の音声基底 $\boldsymbol{H}_s$ を用いて NMF によりクリーン音声の特徴量 $\hat{\boldsymbol{X}}$ を推定する．次に，$\hat{\boldsymbol{X}}$ に対してクリーン GMM でクラスタリングを行うことでクラス m を推定し，$\boldsymbol{H}_m$ のみを用いた NMF によりクリーン音声の特徴量 $\hat{\boldsymbol{X}}_{hard}$ を推定する．

3.3 事後確率に基づく更新

3.2 項では事後確率最大の混合としてハードにクラスを決定したが，事後確率に突出したピークがない場合は一つのクラスに絞ることは不適切である．そこでコスト関数に事後確率を導入し，各クラスの事後確率をソフトに利用することを考える．ここで，雑音重畳音声特徴量 $\boldsymbol{Y}$ と NMF により再構成された特徴量 $\boldsymbol{HU}$ の距離として式 (9) を定義する．クラス m の事後確率を $\boldsymbol{\gamma}_m \in \mathbb{R}^{1\times T}$，その各要素を $\gamma_{m,t}$，$\boldsymbol{H}_m$，$\boldsymbol{H}_n$ の基底数を K_m，K_n とする．H_{ω,k_m}，H_{ω,k_n}，$U_{k_m,t}$，$U_{k_n,t}$ は $\boldsymbol{H}_m$，$\boldsymbol{H}_n$，$\boldsymbol{U}_m$，$\boldsymbol{U}_n$ の各要素である．

$$\begin{aligned} &d_{soft}(\boldsymbol{Y}, \boldsymbol{HU}) \\ &= \sum_m^M \sum_{\omega,t}^{\Omega,T} \gamma_{m,t}\Big(Y_{\omega,t}\log\frac{Y_{\omega,t}}{\sum_{k_m}^{K_m} H_{\omega,k_m}U_{k_m,t} + \sum_{k_n}^{K_n} H_{\omega,k_n}U_{k_n,t}} \\ &\quad - Y_{\omega,t} + \sum_{k_m}^{K_m} H_{\omega,k_m}U_{k_m,t} + \sum_{k_n}^{K_n} H_{\omega,k_n}U_{k_n,t}\Big) \end{aligned} \tag{9}$$

MFCC 領域での事後確率をスペクトル領域における重みとして利用している．これはクラス m の音声基底 $\boldsymbol{H}_m$ と雑音基底 $\boldsymbol{H}_n$ から構成された特徴量 $\boldsymbol{H}_m\boldsymbol{U}_m + \boldsymbol{H}_n\boldsymbol{U}_n$ が事後確率 $\boldsymbol{\gamma}_m$ の割合だけ $\boldsymbol{Y}$ に占めることを意味する．この距離関数は通常の NMF，Noise transductive NMF のどちらにでも適用できる．式 (3)，(6) における第一項を式 (9) とすることで，事後確率をソフトに利用したコスト関数を得る．Noise-transductive NMF の場合，コスト関数は式 (10) のようになる．

$$D(\boldsymbol{Y}||\boldsymbol{HU}) = d_{soft}(\boldsymbol{Y}, \boldsymbol{HU}) + ||\boldsymbol{\lambda}.*\boldsymbol{U}||_p + \boldsymbol{\eta}.*d(\boldsymbol{N}, \boldsymbol{H}_n) \tag{10}$$

ここで，事後確率を特定の混合のみを 1，それ以外の混合を 0 とした場合が 3.2 節のハードな音声基底の選択に対応している．コスト関数 (10) を最小化する音声，雑音それぞれのアクティベーションの更新式として式 (11)，(12) が得られる．

$$\begin{aligned}\boldsymbol{U}_m \longleftarrow\ & \boldsymbol{U}_m.*(\boldsymbol{1}_{K_m}\boldsymbol{\gamma}_m.*(\boldsymbol{H}_m^\top(\boldsymbol{Y}./(\boldsymbol{HU})_{mn}))) \\ & ./(\boldsymbol{\gamma}_m.*\boldsymbol{H}_m^\top\boldsymbol{1}_Y + \boldsymbol{\lambda}_m)\end{aligned} \tag{11}$$

$$\begin{aligned}\boldsymbol{U}_n \longleftarrow\ & \boldsymbol{U}_n.*(\sum_m^M \boldsymbol{1}_{K_n}\boldsymbol{\gamma}_m.*(\boldsymbol{H}_m^\top(\boldsymbol{Y}./(\boldsymbol{HU})_{mn}))) \\ & ./(\boldsymbol{H}_n^\top\boldsymbol{1}_Y + \boldsymbol{\lambda}_n)\end{aligned} \tag{12}$$

ただし $(\boldsymbol{HU})_{mn} = \boldsymbol{H}_m\boldsymbol{U}_m + \boldsymbol{H}_n\boldsymbol{U}_n$

$\boldsymbol{\lambda}_m$，$\boldsymbol{\lambda}_n$ は $\boldsymbol{\lambda}$ の $\boldsymbol{U}_m$，$\boldsymbol{U}_n$ に対する要素，$\boldsymbol{1}_{K_m} \in \mathbb{R}^{K_m \times 1}$，$\boldsymbol{1}_{K_n} \in \mathbb{R}^{K_n \times 1}$ はすべての要素が 1 の行列である．また，雑音基底の更新式としては式 (13) が得られる．

$$\begin{aligned}\boldsymbol{H}_n \longleftarrow\ & (\sum_m^M \boldsymbol{H}_n.*((\boldsymbol{1}_Y\boldsymbol{\gamma}_m.*\boldsymbol{Y})./(\boldsymbol{HU})_{mn}\boldsymbol{U}_n^\top) + \boldsymbol{\eta}.*\boldsymbol{N}) \\ & ./(\sum_m^M \boldsymbol{1}_\Omega\boldsymbol{\gamma}_m\boldsymbol{U}_n^\top + \boldsymbol{\eta})\end{aligned} \tag{13}$$

$\boldsymbol{1}_\Omega \in \mathbb{R}^{\Omega \times 1}$ はすべての要素が 1 の行列である．式 (11)，(12)，(13) の導出は付録に記述する．

3.2 項と同様に，事後確率を得るためにはクリーン音声の特徴量が必要となるため，$\boldsymbol{H}_s$ 全体を用いた NMF で推定したクリーン音声の特徴量 $\hat{\boldsymbol{X}}$ を用いて事後確率を得る．その事後確率を用いてコスト関数 (10) による更新を行うことで，最終的なクリーン音声の特徴量 $\hat{\boldsymbol{X}}_{soft}$ を得る．

4. 実　　験

GMM クラスタリングの併用による効果を検証するため，AURORA2 データベースで認識実験を行った．

4.1 実 験 条 件

実験条件を表 1 に示す．特徴量として MFCC 領域では MFCC12 次元とその Δ，ΔΔ の計 36 次元，スペクトル領域では 23 次元のメルスペクトルとエネルギーの計 24 次元を用いた．スペクトル領域では長時間特徴を考慮するために前後各 9 フレームを結合し，24 次元× 19 フレーム=456 次元の特徴量で NMF を行った．NMF は 2.2 項の Noise-transductive NMF を用いた．式 (6) において，雑音基底の基底数 $K_n = 2$，スパース性に対するペナルティ係数 $\boldsymbol{\lambda}_m = 0.65$，$\boldsymbol{\lambda}_n = 0.5$ とし，基準雑音基底 $\boldsymbol{N}$ は各要素をすべて 1 とした．

表 1　実 験 条 件

特徴量（スペクトル）	メルスペクトル 23 次元+エネルギー
特徴量（MFCC）	MFCC 12 次元+Δ+ΔΔ
GMM 混合数	4
NMF	Noise-transductive NMF
連結フレーム数	前後各 9 フレーム
学習データ	AURORA2 のクリーン音声 8440 発話
テストデータ	AURORA2 セット A，セット B
音響モデル	クリーン条件

表 2　NMF における条件

	音声基底数	学習データ	更新回数
NMF (4000)	4000	全発話	200 回
NMF (4000, 1 of 4div)	4000	ランダム 4 分割 1 セット	200 回
NMF (16000)	16000	全発話	200 回
NMF (16000, all of 4div)	4000 × 4	ランダム 4 分割 4 セット	200 回
NMF+GMM (train)	4000 × 4	クラスごと	200 回
NMF+GMM (hard)	4000	クラスごと	50 回
NMF+GMM (soft)	4000 × 4	クラスごと	200 回
NMF+GMM (hard, oracle)	4000	クラスごと	50 回
NMF+GMM (soft, oracle)	4000 × 4	クラスごと	200 回

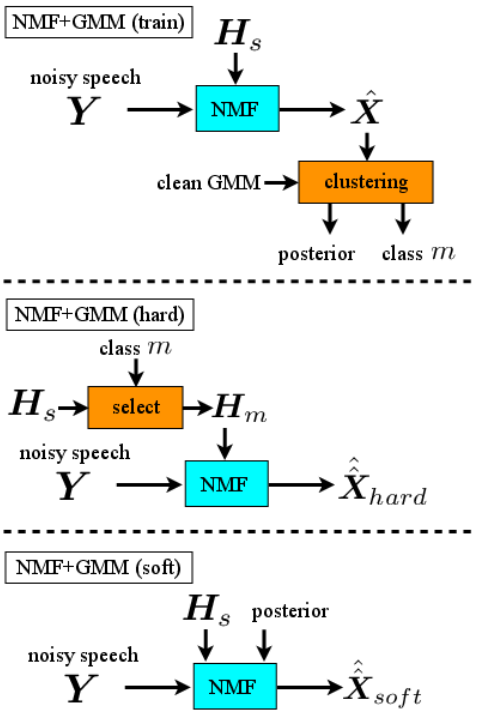

図 2　提 案 手 法

NMF の各条件における音声基底数，音声基底の学習データ，NMF の更新回数を表 2 にまとめる．クリーン GMM は 4 混合とし，1 クラスにつき基底数 4000 ($K_m = 4000(m = 1, 2, 3, 4)$) の基底ベクトルを学習した．学習は式 (3) のコスト関数に基づく通常の NMF により，200 回の更新で行った．したがって，全クラスを合わせた音声基底 $\boldsymbol{H}_s$ の基底数は 4000 × 4=16000 である．初めに $\boldsymbol{H}_s$ 全体を用いてクリーン音声の特徴量 $\hat{\boldsymbol{X}}$ を構

表 3　認識率（set A）[%]

	NMF (4000)	NMF (4000, 1 of 4div)	NMF (16000)	NMF (16000, all of 4div)	NMF+GMM (train)	NMF+GMM (hard)	NMF+GMM (soft)	NMF+GMM (hard, oracle)	NMF+GMM (soft, oracle)
SNR20	84.18	83.05	83.58	85.45	**92.89**	87.78	86.36	93.89	89.73
SNR15	80.76	79.74	80.60	82.53	**89.33**	84.56	82.51	92.58	86.93
SNR10	73.28	72.09	73.73	75.60	**81.73**	78.36	75.32	89.09	80.09
SNR5	59.54	58.17	60.16	61.79	**67.83**	65.68	59.64	80.72	64.88
SNR0	38.82	37.73	39.19	40.32	**45.63**	43.94	33.72	60.35	39.34
Average	67.31	66.15	67.45	69.14	**75.48**	72.06	67.51	83.32	72.19

表 4　認識率（set B）[%]

	NMF (4000)	NMF (4000, 1 of 4div)	NMF (16000)	NMF (16000, all of 4div)	NMF+GMM (train)	NMF+GMM (hard)	NMF+GMM (soft)	NMF+GMM (hard, oracle)	NMF+GMM (soft, oracle)
SNR20	81.17	79.31	79.61	81.59	**90.57**	82.78	74.11	92.53	85.13
SNR15	75.91	74.56	75.87	77.30	**86.21**	77.96	70.01	90.30	81.84
SNR10	67.67	66.74	68.43	69.63	**77.76**	70.95	62.53	86.50	74.28
SNR5	54.17	52.86	54.13	55.96	**62.95**	57.69	47.88	76.64	60.63
SNR0	35.02	33.63	34.58	36.26	**41.78**	37.39	27.62	57.32	38.09
Average	62.79	61.42	62.52	64.15	**71.85**	65.35	56.43	80.66	67.99

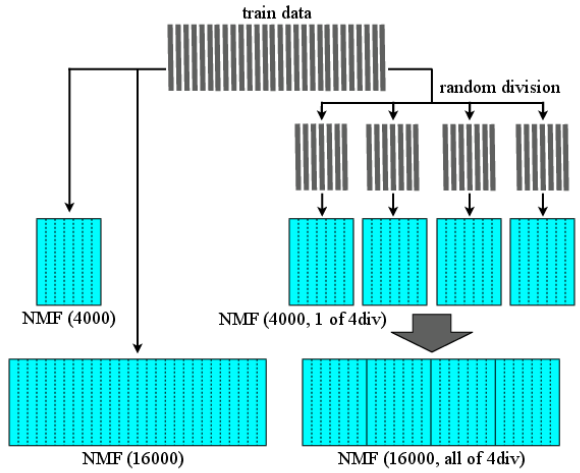

図 3　ベースラインにおける音声基底

成した場合を NMF+GMM (train) とし，$\hat{\boldsymbol{X}}$ から事後確率を得た．3.2 項のように事後確率を用いたハードなクラタリングによって音声基底を選択した場合を NMF+GMM (hard)，3.3 節のようにコスト関数に事後確率を導入した場合を NMF+GMM (soft) とする．それぞれの提案手法の概要を図 2 にまとめる．また，参考として真のクリーン音声 $\boldsymbol{X}$ から得られた事後確率を用いた場合も行った．これらをそれぞれ NMF+GMM (hard, oracle)，NMF+GMM (soft, oracle) とする．

ベースラインとして，従来の NMF のように全学習データから学習した音声基底でクリーン音声を構成した場合も行った．音声基底の基底数は提案手法に対応させるため，4000 と 16000 の 2 パターンである．これらを NMF (4000)，NMF (16000) とする．また，提案手法では学習データを分割しているため，学習データ数の変化による影響を考慮する必要がある．そこで，提案手法のように学習データを分割して音声基底の学習を行った．学習データをフレーム単位でデータ数が等しくなるようランダムに 4 セットに分割し，4 セットそれぞれから基底数 4000 の基底を学習した．そのうちの一つを音声基底として利用した場合を NMF (4000, 1 of 4div)，4 セット分すべての音声基底を利用した場合を NMF (16000, all of 4div) とする．これらのベースラインにおける音声基底の関係を図 3 にまとめる．

4.2　実験結果

各実験条件における set A，set B の認識率を表 3，4 に示す．NMF+GMM (train) を見ると各ベースラインよりも高い認識率を得られているため，クラスごとの音声基底の学習によってより適切な音声基底が構成できたことがわかる．NMF (16000, all of 4div) と比較すると，単なる学習データの分割による効果ではなく MFCC 領域における GMM クラスタリングの効果であることが確認できる．NMF+GMM (hard) でさらなる認識率の向上を期待したが，各ベースラインよりは高い認識率であるものの NMF+GMM (train) には及ばない結果となった．しかし，NMF+GMM (hard, oracle) では NMF+GMM (train) より高い認識率が得られた．このことから，ハードなクラスタリングを利用した音声基底の選択そのものは有効であることがわかる．実際の認識で利用するためには，雑音重畳音声から正確にクリーン音声のクラスを推定する手法が必要となる．

NMF+GMM (soft) では，ハードなクラスタリングで選択されず切り捨てられたクラスの情報が柔軟に利用できることから，NMF+GMM (hard) より高い認識率となることを期待した．しかし，実際には SNR によらず NMF+GMM (hard) に及ばない結果となった．音声基底の学習時はハードなクラスタリングを行い，テスト時は事後確率をソフトに用いているため，そのミスマッチが原因と考えられる．また，SNR20，15 では各ベースラインに勝るものの，SNR10，5，0 では各ベースラインより低い認識率となっている．低 SNR になるほどベースラインに比べて認識率が悪くなることから，式 (13) による雑音基底の更新が適切に行われていないかったことも原因と考えられる．NMF+GMM (soft, oracle) も同様の傾向になってい

るため，これらは事後確率の精度の問題ではない．したがって，事後確率を導入したコスト関数を再検討する必要がある．

5. おわりに

従来の NMF では入力音声に対する教師なしで分解を行うため，不適切な音声基底が用いられることが課題となっていた．そこで，本稿では，MFCC 領域における GMM クラスタリングを併用した NMF による特徴量強調を提案した．まず音声基底の学習において，MFCC 領域における GMM クラスタリングによって得られたクラスを利用し，クラス依存の音声基底を学習した．次に，入力音声に対しても同様にクラスタリングも行い，その結果に基づいた音声基底の選択による NMF を行った．認識実験により，クラスごとに音声基底を学習する手法の有効が確認できた．また，クリーン音声のハードなクラスタリング結果を用いた音声基底の選択によって，クリーン音声の構成精度が上がることも確認できた．しかし，クリーン音声の特徴量のクラス推定が課題となった．近年，MFCC 領域における GMM を用いた特徴量強調の一つとして，雑音重畳音声の特徴量からクリーン音声の特徴量のクラスを推定する手法を用いた特徴量強調が提案されている [6]．これをハードなクラスタリングによる音声基底の選択に導入したい．さらに，事後確率をソフトに利用したコスト関数を提案したが，ベースラインに及ばない結果となった．したがって，コスト関数を再検討する必要がある．また，事後確率を導入したコスト関数を用いることで，音声基底の学習においても事後確率をソフトに利用することも検討したい．

付　　録

事後確率を導入したコスト関数 (10) から更新式 (11)，(12)，(13) を導出する．補助関数を用いて，通常の NMF [3]，Noise-transductive NMF [5] と同様に導出できる．コスト関数 (10) における距離関数 $d_{soft}(\boldsymbol{Y}, \boldsymbol{HU})$ は以下のように上限をとることができる．

$$
\begin{aligned}
& d_{soft}(\boldsymbol{Y}, \boldsymbol{HU}) \\
=& \sum_{m}^{M}\sum_{\omega,t}^{\Omega,T} \lambda_{m,t}\Big(Y_{\omega,t}\log\frac{Y_{\omega,t}}{\sum_{k_m}^{K_m} H_{\omega,k_m}U_{k_m,t}+\sum_{k_n}^{K_n} H_{\omega,k_n}U_{k_n,t}} \\
& -Y_{\omega,t}+\sum_{k_n}^{K_n} H_{\omega,k_n}U_{k_n,t}+\sum_{k_n}^{K_n} H_{\omega,k_n}U_{k_n,t}\Big) \\
\leqq& \sum_{\omega,t}^{\Omega,T}\sum_{m}^{M}\gamma_{m,t}\Big\{Y_{\omega,t}\log Y_{\omega,t} \\
& \underline{-Y_{\omega,t}\Big(\sum_{k_m}^{K_m}\delta_{k_m,K_n,\omega,t}\log\frac{H_{\omega,k_m}U_{k_m,t}}{\delta_{k_m,K_n,\omega,t}}} \\
& \underline{+\sum_{k_n}^{K_n}\delta_{K_m,k_n,\omega,t}\log\frac{H_{\omega,k_n}U_{k_n,t}}{\delta_{K_m,k_n,\omega,t}}\Big)} \\
& -Y_{\omega,t}+HU_{\omega,t,m}\Big\} \qquad \text{(A·1)} \\
=& \; Q_{soft}(\boldsymbol{Y}, \boldsymbol{HU}, \boldsymbol{\delta}) \text{ s.t.} \sum_{k_m}^{K_m}\delta_{k_m,K_n\cdot\omega,t}+\sum_{k_n}^{K_n}\delta_{K_m,k_n\cdot\omega,t}=1
\end{aligned}
$$

ただし $HU_{\omega,t,m}=\sum_{k_m}^{K_m} H_{\omega,k_m}U_{k_m,t}+\sum_{k_n}^{K_n} H_{\omega,k_n}U_{k_n,t}$

下線部において Jensen の不等式を導入することで上限をとっている．$\delta_{k_m,K_n\cdot\omega,t}$，$\delta_{K_m,k_n\cdot\omega,t}$ は $\boldsymbol{\delta}$ の各要素である．等号成立条件は式 (A·2)，(A·3) となる．

$$\hat{\delta}_{k_m,K_n,\omega,t} = \frac{H_{\omega,k_m}U_{k_m,t}}{HU_{\omega,t,m}} \qquad \text{(A·2)}$$

$$\hat{\delta}_{K_m,k_n,\omega,t} = \frac{H_{\omega,k_n}U_{k_n,t}}{HU_{\omega,t,m}} \qquad \text{(A·3)}$$

したがって，以下の式 (A·4) を最小化すれば良い．

$$Q_{soft}(\boldsymbol{Y}, \boldsymbol{HU}, \hat{\boldsymbol{\delta}}) + ||\lambda .* \boldsymbol{U}||_p + \boldsymbol{\eta} .* d(\boldsymbol{N}, \boldsymbol{H}_n) \qquad \text{(A·4)}$$

式 (A·4) の $\boldsymbol{U}$，$\boldsymbol{H}_n$ についての導関数から，以下の更新式を得る．

$$U_{k_m,t} \leftarrow U_{k_m,t}\frac{\gamma_{m,t}\sum_{\omega}^{\Omega} X_{\omega,t}\frac{H_{\omega,k_m}}{HU_{\omega,t,m}}}{\gamma_{m,t}\sum_{\omega}^{\Omega} H_{\omega,k_m}+\lambda_{k_m}} \qquad \text{(A·5)}$$

$$U_{k_n,t} \leftarrow U_{k_n,t}\frac{\sum_{m}^{M}\gamma_{m,t}\sum_{\omega}^{\Omega} X_{\omega,t}\frac{H_{\omega,k_n}}{HU_{\omega,t,m}}}{\sum_{\omega}^{\Omega} H_{\omega,k_n}+\lambda_{k_n}} \qquad \text{(A·6)}$$

$$H_{\omega,k_n} \leftarrow \frac{\sum_{t}^{T}\sum_{m}^{M}\gamma_{m,t}Y_{\omega,t}\frac{H_{\omega,k_n}U_{k_n,t}}{HU_{\omega,t,m}}+\eta N_{\omega,k_n}}{\sum_{t}^{T}\sum_{m}^{M}\gamma_{m,t}U_{k_n,t}+\eta} \qquad \text{(A·7)}$$

文　　献

[1] P. J. Moreno and B. Raj R. M. Stern, “A vector taylor series approach for environment independent speech recognition,” ICASSP, pp.733–736, 1996.

[2] J. Droppo, L. Deng and A. Acero, “Ecaluation of the SPLICE algorithm on the AURORA2 database,” Proc. Eurospeech, Vol.1, pp.217–220, 2001.

[3] J. F. Gemmeke, T. Virtanen and A. Hurmalainen, “Exemplar-based sparse representations for noise robust automatic speech recognition,” IEEE Transactions, Vol.19, No7, pp.2067–2080, 2011.

[4] S. Watanabe and J. R. Hershey, “Stereo-based feature enhancement using dictionary learning,” ICASSP, pp.7073–7077, 2013.

[5] Y. Luan, D. Saito, Y. Kashiwagi, N. Minematsu and K. Hirose, “Semi-supervised noise dictionary adaptation for exemplar-based noise robust speech recognition,” ICASSP, pp.1764–1767, 2014.

[6] 鈴木 雅之, 吉岡 拓也, 渡辺 司, 峯松 信明, 広瀬 啓吉, “クリーン音声状態の識別に基づく特徴量強調”, 日本音響学会春季講演論文集, 1-7-10, pp.23–26, 2012.

[7] 相原 龍, 滝口 哲也, 有木 康雄, “辞書選択型非負値行列因子分解による構音障害者の声質変換”, 電子情報通信学会技術報告, Vol.113, No.366, SP-2013-87, pp.71–76, 2013.