

MFCC-based GMM による事後確率を用いた NMF と その雑音環境下音声認識への応用*

☆藤垣健太郎, 柏木陽佑, 齋藤大輔, 峯松信明, 広瀬啓吉 (東大)

1 はじめに

音声認識技術の利用が広まる中で、実環境で利用するためには雑音抑圧の技術が不可欠となっている。雑音抑圧手法の一つとして、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基底と雑音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例である [1]。

NMF 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われる。従来の NMF では、入力音声の特性に依らず、利用可能なすべての音声基底を同等に扱ってパラメータ更新を行っている。ここで、なんらかの情報に基づき、入力音声に応じて音声基底を選択的に利用できれば、より当該音声に適した更新が行えると期待できる。NMF のアルゴリズムは声質変換にも用いられており、基底の選択による精度向上が示されている。[2]。[2] においては、音素のクラスタリングに基づいて基底のクラスタリングを行い、入力音声の音素に対応する基底のみを利用している。

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、より高精度にクリーン音声を再構成できると期待できる。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、他の基準での基底の学習、選択の方法を検討する必要がある。そこで本稿では、MFCC 領域における GMM 分類によって得られた事後確率を基準として用いることを提案する。

事後確率の利用方法としては、2つの手法を検討している。既に [4] において、事後確率が最大となるクラスに対応する基底群を選択するハードな分類による手法を提案した。本稿では、クラスを一つに決めず、事後確率をソフトに利用した手法を提案する。

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 ($A \rightarrow BC$) である。元の行列 A と積型で表現した 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 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。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増定される [1]。

入力された雑音重畳音声特徴量 $Y \in \mathbb{R}^{\Omega \times T}$ を基底 $H \in \mathbb{R}^{\Omega \times K}$ とアクティベーション $U \in \mathbb{R}^{K \times T}$ 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 H を固定し、アクティベーション U のみを更新する。 H, U はそれぞれ、音声基底 H_s と雑音基底 H_n 、音声基底のアクティベーション U_s と雑音基底のアクティベーション U_n の連結により構成される。雑音重畳音声特徴量 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 $\hat{X}$ を得る。

$$Y = HU = [H_s H_n] \begin{bmatrix} U_s \\ U_n \end{bmatrix} \quad (1)$$

$$\hat{X} = H_s U_s \quad (2)$$

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

$$D(Y||HU) = d(Y, HU) + \|\lambda * U\|_p \quad (3)$$

第一項の $d(Y, HU)$ 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 λ 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 $d(Y, HU)$ 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 Y, H, U の各要素である。

$$d(Y, HU) = \sum_{\omega,t}^{\Omega,T} e(Y, HU)_{\omega,t} \quad (4)$$

$$e(Y, HU)_{\omega,t} = \log \frac{Y_{\omega,t}}{\sum_k^K H_{\omega,k} U_{k,t}} - Y_{\omega,t} + \sum_k^K H_{\omega,k} U_{k,t}$$

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。

$$U \leftarrow U * (H^\top (Y ./ (HU))) ./ (H^\top 1_Y + \lambda) \quad (5)$$

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、 $1_Y \in \mathbb{R}^{\Omega \times T}$ はすべての要素が 1 の行列である。

2.2 Noise-transductive NMF

通常の NMF による特徴量強調 [1] では、雑音基底が事前の学習により固定されているため、未知雑音

*Noise robust speech recognition using NMF integrated with GMM-based posteriors in the MFCC domain. by Kentaro FUJIGAKI, Yosuke KASHIWAGI, Daisuke SAITO, Nobuaki MINEMATSU, Keikichi HIROSE(The University of Tokyo)

への対応が課題となる。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 [3]。

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。[3] では、コスト関数を式 (6) としている。

$$D(Y||HU)=d(Y, HU)+\|\lambda * U\|_p+\eta * d(N, H_n) \quad (6)$$

第 3 項は現在の雑音基底 H_n と基準雑音基底 N の KL ダイバージェンスであり、 η はその重みである。 N は事前に雑音から用意された基底である。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。

$$\begin{aligned} U &\leftarrow U * (H^T(Y/(HU)))/(H^T 1_Y + \lambda) \quad (7) \\ H_n &\leftarrow (H_n * (Y/(HU)U_n^T) + \eta * N)/(1_Y U_n^T + \eta) \quad (8) \end{aligned}$$

入力音声に含まれる雑音に対し適応的に H_n を推定する本手法の効果は [3] を参照していただきたい。

3 GMM による事後確率を用いた NMF

従来の NMF では、入力音声の特性に依らず、利用可能なすべての基底を同等に扱ってパラメータ更新を行っている。ここで、入力音声に応じて適応的に基底を選択することで、より高精度にクリーン音声を再構成できると期待される。基底選択の方法の一つとして、音素情報をラベルとして利用することが考えられるが、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 領域における GMM 分類によって得られた事後確率を用いて基底の選択を行うことで、クリーン音声の構成精度の向上を図る。

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 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 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 領域での GMM 分類を導入する。まず、学習データに対して GMM 分類を行い、その事後確率に基づいて基底を学習する。また、評価データに対しても GMM 分類を行い、適応的に基底を選択して利用する。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が期待できる。

GMM 分類によって得られた事後確率の利用方法として、2つの手法を検討している。既に [4] において、事後確率が最大となる分布をクラスとするハードな分類による手法を提案した。本稿ではさらに、事後確率そのものをコスト関数に導入してソフトに利用した手法を提案する。

3.1 ハードな分類による手法 [4]

3.1.1 学習データの GMM 分類による音声基底の学習

MFCC 領域におけるハードな GMM 分類を利用して音声基底を学習することを考える。まず、混合数 M のクリーン GMM (クリーン音声特徴量の 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分類する。即ち、入力フレームに対して事後確率が最大となる分布を求め、そのインデックスを当該フレームのクラスとする。クラスごとに分類されたフレームのみを用いて音声基底を学習する。クラス m ($m = 1, \dots, M$) に対応する音声基底を H_m とする。

3.1.2 評価データに対する音声基底の選択

評価データに対しても同様に分類を行い、入力フレームのクラス m に対応した H_m のみを音声基底として用いる。ここで、クリーン GMM を用いるにはクリーン音声の特徴量が必要になるため、2段階の NMF を行う。まず全クラスの音声基底 $H_s = [H_1 \dots H_M]$ を用いた NMF によりクリーン音声の特徴量 $\hat{X}_{hard}$ を推定する。次に、 $\hat{X}_{hard}$ のクラス m を推定し、 H_m のみを用いた NMF により最終的なクリーン音声の特徴量 $\hat{X}_{hard}$ を得る。

3.2 事後確率をソフトに利用した手法

3.2.1 事後確率に基づく更新

ハードな分類による手法では、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 Y と NMF により再構成された特徴量 HU の距離として式 (9) を定義する。クラス $m = 1, \dots, M$ の事後確率を $\gamma_m \in \mathbb{R}^{1 \times T}$ 、その各要素を $\gamma_{m,t}$ とする。

$$d_{soft}(Y, HU) = \sum_m \sum_{\omega, t} \gamma_{m,t} e^{(Y, H_m U_m + H_n U_n)_{\omega, t}} \quad (9)$$

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 H_m と雑音基底 H_n から構成された特徴量 $H_m U_m + H_n U_n$ が事後確率 γ_m の割合だけ 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。ここで、事後確率を特定の分布のみを 1、それ以外の分布を 0 とした場合が 3.1 節のハードな分類に対応している。

3.2.2 音声基底の学習

クリーン音声 X から音声基底を学習することを考える。クリーン音声のみを学習データとし、雑音基底 H_n は除くため、基底は音声基底 H_m ($m = 1, \dots, M$) のみとなり、全体の基底数は $K = \sum_m^M K_m$ となる。アクティベーションも同様に音声基底に対応する U_m ($m = 1, \dots, M$) のみとなる。したがって、コスト関数は以下のようなになる。

$$D(\mathbf{X}||\mathbf{H}\mathbf{U}) = \sum_{\omega,t} \sum_m^{\Omega,T} \gamma_{m,t} e(\mathbf{Y}, \mathbf{H}_m \mathbf{U}_m)_{\omega,t} + \sum_{k,t}^{K,T} \lambda_k U_{k,t} \quad (10)$$

更新後の状態 m の音声基底を $\bar{\mathbf{H}}_m$ とすると、更新式は以下になる。

$$\bar{\mathbf{H}}_m = \mathbf{H}_m \cdot (((1_{\mathbf{Y}} * \mathbf{X}) ./ (\mathbf{H}_m \mathbf{U}_m)) \mathbf{U}_m^{\top}) ./ 1_{\Omega} \gamma_m \mathbf{U}_m \quad (11)$$

$1_{\Omega} \in \mathbb{R}^{\Omega \times 1}$ はすべての要素が 1 の行列である。

3.2.3 音声基底の選択的な利用

入力音声に対する事後確率を利用した更新を考える。Noise-transductive NMF の場合、コスト関数は式 (12) のようになる。

$$D(\mathbf{Y}||\mathbf{H}\mathbf{U}) = d_{soft}(\mathbf{Y}, \mathbf{H}\mathbf{U}) + \|\lambda * \mathbf{U}\|_p + \eta * d(\mathbf{N}, \mathbf{H}_n) \quad (12)$$

コスト関数 (12) 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 (13), (14) が得られる。

$$\mathbf{U}_m \leftarrow \mathbf{U}_m \cdot (\mathbf{1}_{K_m} \gamma_m \cdot (\mathbf{H}_m^{\top} (\mathbf{Y} ./ (\mathbf{H}\mathbf{U})_{mn}))) ./ (\gamma_m \cdot \mathbf{H}_m^{\top} \mathbf{1}_Y + \lambda_m) \quad (13)$$

$$\mathbf{U}_n \leftarrow \mathbf{U}_n \cdot (\sum_m^M \mathbf{1}_{K_n} \gamma_m \cdot (\mathbf{H}_m^{\top} (\mathbf{Y} ./ (\mathbf{H}\mathbf{U})_{mn}))) ./ (\mathbf{H}_n^{\top} \mathbf{1}_Y + \lambda_n) \quad (14)$$

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

λ_m, λ_n は λ の $\mathbf{U}_m, \mathbf{U}_n$ に対する要素, $\mathbf{1}_{K_m} \in \mathbb{R}^{K_m \times 1}$, $\mathbf{1}_{K_n} \in \mathbb{R}^{K_n \times 1}$ はすべての要素が 1 の行列である。 K_m, K_n はクラス m の音声基底 $\mathbf{H}_m$, 雑音基底 $\mathbf{H}_n$ それぞれの基底数である。また、雑音基底の更新式としては式 (15) が得られる。

$$\mathbf{H}_n \leftarrow (\sum_m^M \mathbf{H}_n \cdot (((1_Y \gamma_m * \mathbf{Y}) ./ (\mathbf{H}\mathbf{U})_{mn}) \mathbf{U}_n^{\top}) + \eta * \mathbf{N}) ./ (\sum_m^M \mathbf{1}_{\Omega} \gamma_m \mathbf{U}_n^{\top} + \eta) \quad (15)$$

式 (13), (14), (15) の導出は [4] を参照していただきたい。

3.1 節と同様に、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、2 段階の NMF を行う。まず、事後確率を用いない通常のコスト関数 (6) に基づいて $\mathbf{H}_s$ 全体を用いた NMF でクリーン音声の特徴量 $\hat{\mathbf{X}}_{soft}$ を推定する。次に、 $\hat{\mathbf{X}}_{soft}$ から得られた事後確率を用いてコスト関数 (12) 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 $\hat{\mathbf{X}}_{soft}$ を得る。

4 実験

4.1 実験条件

GMM 事後確率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。

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 $\Delta, \Delta\Delta$ の計 36 次元、ス

Table 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 + $\Delta\Delta$
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
音響モデル	クリーン条件

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 $\times$ 19 フレーム=456 次元の特徴量で NMF を行った。音声基底、雑音基底も同様に 456 次元である。NMF は 2.2 節の 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 $\lambda_m = 0.65$, $\lambda_n = 0.5$ とし、基準雑音基底 $\mathbf{N}$ は各要素をすべて 1 とした。

NMF の各条件における音声基底数、音声基底の学習データ、コスト関数、事後確率を得る特徴量を表 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。

ハードな分類に基づく手法においては、コスト関数は基底の学習では式 (3) を、テスト時のクリーン音声の推定では式 (6) を用いている。初めに学習データをハードに分割して学習した全クラスの音声基底 $\mathbf{H}_s$ を用いてクリーン音声の特徴量 $\hat{\mathbf{X}}_{hard}$ を推定した。この $\hat{\mathbf{X}}_{hard}$ を特徴量として認識した場合を NMF+GMM (hard1) とする。さらに 2 段階目として、 $\hat{\mathbf{X}}_{hard}$ から事後確率を得て、最大事後確率に基づくハードなクラタリングによって音声基底 $\mathbf{H}_m$ を選択した場合を NMF+GMM (hard2) とする。

事後確率をソフトに用いた手法では、まず初めに式 (10) に基づいて音声基底 $\mathbf{H}_s = [\mathbf{H}_1 \dots \mathbf{H}_M]$ を学習した。この音声基底 $\mathbf{H}_s$ を用いて式 (6) に基づいてクリーン音声の特徴量 $\hat{\mathbf{X}}_{soft}$ を推定した。この $\hat{\mathbf{X}}_{soft}$ を特徴量として認識した場合を NMF+GMM (soft1) とする。さらに 2 段階目として、 $\hat{\mathbf{X}}_{soft}$ から得られた事後確率を用いて式 (12) に基づいてクリーン音声を推定した場合を NMF+GMM (soft2) とする。また、参考として真のクリーン音声 $\mathbf{X}$ から得られた事後確率を用いた場合も行った。これを NMF+GMM (hard2, oracle), NMF+GMM (soft2, 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を行った。音声基底の基底数は 16000 とした。これを NMF (16000) とする。

4.2 実験結果

各実験条件における認識率を表 3 に示す。NMF+GMM (hard1), NMF+GMM (soft1) が高い認識率を得られているため、事後確率を用いた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。特に、事後確率をソフトに利用したコスト関数を用いた学習は期待通りハード

Table 2 NMF における条件

	音声基底数	音声基底の 学習データ	コスト関数		事後確率を得る特徴量
			学習	テスト	
NMF (16000)	16000	全発話	式 (3)	式 (6)	-
NMF+GMM (hard1)	4000 × 4	クラスごとに分割	式 (3)	式 (6)	-
NMF+GMM (hard2)	4000	クラスごとに分割	式 (3)	式 (6)	NMF+GMM (hard1) $\hat{\mathbf{X}}_{hard}$
NMF+GMM (soft1)	4000 × 4	全発話+事後確率	式 (10)	式 (6)	-
NMF+GMM (soft2)	4000 × 4	全発話+事後確率	式 (10)	式 (12)	NMF+GMM (soft1) $\hat{\mathbf{X}}_{soft}$
NMF+GMM (hard2, oracle)	4000	クラスごとに分割	式 (3)	式 (6)	クリーン音声 $\mathbf{X}$
NMF+GMM (soft2, oracle)	4000 × 4	全発話+事後確率	式 (10)	式 (12)	クリーン音声 $\mathbf{X}$

Table 3 認識率 [%]

	NMF (16000)	NMF+GMM (hard1)	NMF+GMM (hard2)	NMF+GMM (soft1)	NMF+GMM (soft2)	NMF+GMM (hard2, oracle)	NMF+GMM (soft2, oracle)
SNR20	83.58	92.89	87.78	94.09	85.45	93.89	89.28
SNR15	80.60	89.33	84.56	90.60	82.11	92.58	86.22
SNR10	73.73	81.73	78.36	83.53	74.92	89.09	79.38
SNR5	60.16	67.83	65.68	69.49	60.05	80.72	64.35
SNR0	39.19	45.63	43.94	46.53	33.54	60.35	38.17
Average	67.45	75.48	72.06	76.85	67.22	83.32	71.48

な分類よりも有効であった。NMF+GMM (hard2)、NMF+GMM (soft2) でさらなる認識率の向上を期待したが、NMF+GMM (hard1)、NMF+GMM (soft1) には及ばない結果となった。ハードな分類については、NMF+GMM (hard2, oracle) で NMF+GMM (hard1) より高い認識率が得られているため、ハードな分類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。

一方で、事後確率をソフトに利用した場合は、ハードな分類と異なり、事後確率が正しく得られている NMF+GMM (soft2, oracle) も NMF+GMM (soft1) に及ばない結果となった。有効性が確認できた NMF+GMM (soft1)、NMF+GMM (hard2, oracle) と異なる点は、事後確率を利用した更新が音声基底とそのアクティベーションに加えて雑音基底とそのアクティベーションに対しても行われている点である。音声基底の学習においては式 (10) を用いるためため雑音の要素は扱われておらず、NMF+GMM (soft1) でも入力音声に対しては事後確率を導入していない式 (6) によって更新している。したがって、雑音の要素の更新をふまえて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、入力音声に依らずすべての音声基底を利用して分解を行っている。したがって、入力音声に対して適応的に音声基底を選択して利用することで、クリーン音声の構成精度の向上が期待できる。そこで本稿では、MFCC 領域での GMM による事後確率を用いた NMF を提案した。事後確率の利用方法として、ハードな分類と事後確率をソフトに利用したコスト関数の導入の 2 つを行った。まず音声基底の学習において、MFCC 領域における GMM 分類によって得られた事後確率を利用し、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

同様に分類も行い、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、事後確率を利用した音声基底の学習の有効性が確認できた。特に、事後確率のソフトな利用したコスト関数による学習が有効であった。また、クリーン音声のハードな分類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。その一方で認識時に、事後確率をソフトに利用したコスト関数による音声基底の選択的な利用は効果が得られなかった。事後確率を導入した場合の雑音基底とそのアクティベーションの更新が課題と考えられるため、コスト関数を再検討する必要がある。

参考文献

- [1] J. F. Gemmeke *et al.*, “Exemplar-based sparse representations for noise robust automatic speech recognition,” *IEEE Transactions*, Vol.19, No.7, pp.2067–2080, 2011.
- [2] 相原 龍 他, “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.
- [3] Y. Luan *et al.*, “Semi-supervised noise dictionary adaptation for exemplar-based noise robust speech recognition,” *ICASSP*, pp.1764–1767, 2014.
- [4] 藤垣 健太郎 他, “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”, 電子情報通信学会技術報告, SP-2014-113, pp.1–6, 2014.