修士論文

MFCC-based GMMを用いたNon-negative Matrix Factorizationの高精度化とその雑音環境下音声認識への応用

2015 年 2 月 5 日

指導教員　峯松 信明 教授

電気系工学専攻

37-136484　藤垣 健太郎

内容梗概

現在，音声認識システムは新たなユーザーインタフェースとして利用が広がっている．例えば，手を使わず操作する手段としてカーナビゲーションシステムやスマートフォンで利用されている．また，コールセンターでの電話自動対応システムや議事録の自動作成システムといったビジネス分野での応用例もある．

これらのアプリケーションを実環境で利用する場合，入力音声以外の環境雑音の混入が不可避であり，これが認識率の低下の大きな原因となっている．そこで，雑音に頑健な音声認識のための技術がこれまで検討されてきた．

現在の音声認識システムでは，入力音声の特徴量と事前に学習された音響モデルとのマッチングによって認識を行っている．雑音に対する頑健性を高めるためには，雑音の影響により特徴量と音響モデルとの間に生じる環境のミスマッチを解消する必要がある．その方法として大きく 2 つのアプローチがある．雑音の乗った音声を雑音のないクリーンな音声に近づける音声強調と音響モデルを雑音の乗った音声に適応させるモデル適応である．本論文では音声強調を扱い，その中でも特に，スペクトル領域での音声強調である Nonnegative Matrix Factorizastion (NMF) をパワースペクトルの時系列に適用する手法に注目した．

NMF は非負の行列を非負の行列の積に分解する（$\boldsymbol{A} \rightarrow \boldsymbol{BC}$）アルゴリズムであり，コスト関数を最小化する基準のもと繰り返しパラメータ更新を行うことで行列の分解を行う．音声強調では，雑音の乗った音声を音声と雑音それぞれの構成要素である基底行列とその重み行列の積で表現することで，音声と雑音の分離を実現する．従来の NMF では入力音声に依らずすべての基底を同等に扱っていた．しかし，入力音声に対して適応的な基底選択を行うことで，音声強調の精度向上が期待できる．そこで，NMF による音声強調に対して基底選択を導入することで，音声強調の精度向上を目指した．

GMM を用いたモデルベースの音声強調手法をふまえて，MFCC 領域における GMM (MFCC-based GMM) 分類による事後確率を用いた NMF を提案し，雑音環境下音声認識実験によってその有効性を検証した．事後確率の利用方法として，事後確率最大となる分布をクラスとするハードな分類とコスト関数に事後確率を導入したソフトな利用の 2 つを行った．具体的には，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，事後確率を導入したコスト関数による学習が有効であった．また，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．これは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．結果としてまだ課題点は残るものの，MFCC-based GMM の事後確率を用いた NMF には一定の有効性を確認できた．

目次

図目次

表目次

第1章

序論

1.1　背景

現在，音声認識システムは新たなユーザーインタフェースとして利用が広がっている．例えば，手を使わずテキスト入力やコマンド操作をする手段の一つとして，カーナビゲーションシステムやスマートフォンの入力システムで利用されている．近年は対話型音声案内デジタルサイネージ [1] のように音声対話システムの実用化も広まっており，音声認識技術の需要は高まっている．また，このような一般ユーザー向けのシステムにとどまらず，コールセンターでの電話自動対応システムや議事録の自動作成システム [2] といったビジネス分野での応用例もある．

音声認識システムは，音声波形から発話内容をテキストに書き起こすシステムである．音声波形は，同一の発話内容であっても話者の特性，周囲の雑音，チャネル特性など，さまざまな要因で変化する．したがって，発話内容以外の変化に頑健な音声認識システムが必要となる．特に，実環境で利用するにあたっては周囲の雑音が大きな問題となる [3]．実環境では入力音声以外の環境雑音の混入が不可避であり，認識率の低下の大きな原因となっている．そこで，雑音に頑健な音声認識のための技術がこれまで検討されてきた [4]．

現在の音声認識システムでは，まず音声波形からパワースペクトルを求め，そこから Mel-Frequency Cepstrum Coefficients (MFCC) などの特徴量を抽出する．そして，その特徴量と Hidden Markov Model (HMM) による音響モデルとのマッチングによって認識を行う．ここで問題となるのが音響モデルを学習する際の環境と実際の音声を入力する際の環境のミスマッチであり，その一つの要因が環境雑音である．この雑音によるミスマッチ低減する手法は大きく次の 2 つに分けられる．雑音の乗った音声を雑音のないクリーンな音声に近づける手法と音響モデルを雑音の乗った音声に適応させる手法である．前者の手法は音声強調と呼ばれている．

本論文では，音声強調の中でも NMF による事例ベースの手法に注目する．音声強調は Gaussian Mixture Model (GMM) を用いたモデルベースの手法の研究例が多かったが，近年では事例ベースの手法の研究も進められている．事例ベースの手法の例として，スペクトル領域での音声強調である Nonnegative Matrix Factorizastion (NMF) [5] をパワースペクトルの時系列に適用する手法が挙げられる．NMF は非負の行列を非負の行列の積に分解する（$\boldsymbol{A} \rightarrow \boldsymbol{BC}$）アルゴリズムである．音声強調では，雑音の乗った音声を音声と雑音それぞれの構成要素である基底行列とその重み行列の積で表現することで，音声と雑音の分離を実現する．NMF は音声強調だけでなく声質変換や音源分離にも利用されており，音声強調と声質変換を同時に行った例 [6] や雑音として音楽を扱った例 [7] もある．声質変換においては，入力音声に応じた基底の選択によって精度向上が確認されている [8, 9]．そこで本論文では，NMF による音声強調に対して基底選択を導入することで，音声強調の精度向上を目指す．GMM を用いたモデルベースの音声強調手法をふまえて，MFCC 領域における GMM を併用した基底選択の手法を提案する．

1.2　本論文の構成

本論文は全 6 章から構成される．まず第 1 章では，本論文の背景と目的について述べる．第 2 章では，音声認識システムの枠組みとそれを実現する要素技術について解説する．第 3 章では，雑音によるミスマッチを低減するアプローチとして音声強調に注目し，GMM を用いたモデルベースの音声強調と NMF による事例ベースの音声強調の 2 つについて代表的な手法を紹介する．第 4 章では，音声強調手法の中でも NMF に注目し，その精度向上を図る．NMF は入力音声に対して適応的に基底を選択して利用することで精度の向上が示されており，雑音環境下音声認識においてこれを実現する手法として MFCC-based GMM の事後確率を用いた NMF を提案する．第 5

章では，雑音環境下音声認識実験によって提案手法の有効性を検証する．最後に第 6 章で本論文をまとめ，今後の展望について述べる．

第2章

音声認識システムの枠組み

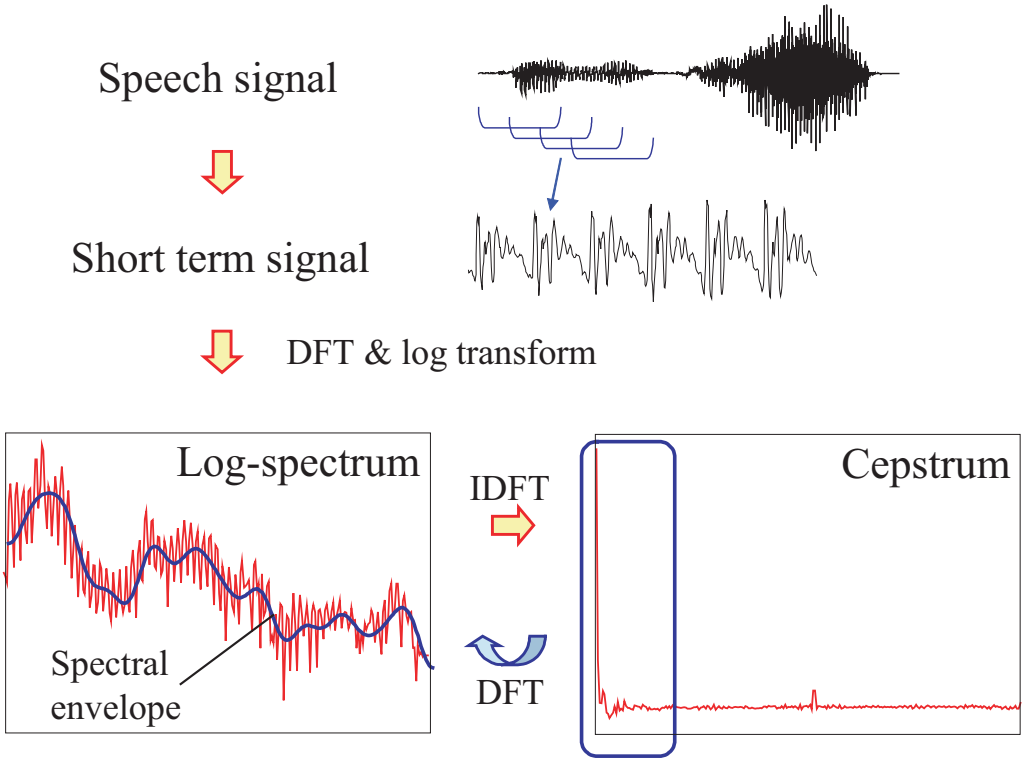

図 2.1: 音声信号からのケプストラム抽出

2.1　はじめに

本章では音声認識システムの枠組みとして，音声認識で用いられる特徴量と HMM による音響モデルについて説明する．また，雑音環境下での音声認識について述べる．

2.2　音響的特徴

2.2.1　ケプストラム

音声信号は声帯で生み出された音源波と声道フィルタの畳み込みで表される (ソースフィルタモデル)．主に基本周波数やパワーの情報が音源成分に含まれ，主に発話内容といった言語情報が声道成分に含まれることが多い．時間領域で畳み込まれた音声信号を分析するには，それらを分割して扱う必要がある．そのための特徴量として現在広く用いられているのが，音韻的特徴のひとつの技術的定義としてのケプストラムである．

音声信号の波形からケプストラムを抽出する過程を図 2.1 に示す．まず離散フーリエ変換 (Discrete Fourier Transform; DFT) を施し，その対数をとることで対数パワースペクトルを抽出する．対数パワースペクトルに逆離散フーリエ変換 (Inverse DFT; IDFT) を施したものがケプストラムである．ケプストラムの低次の成分は図 2.1 のスペクトルの概形，つまりスペクトル包絡の特性を示し，それが主に声道成分の特性を表す．高次の成分は図 2.1 のスペクトルの細かな櫛状になっている高周波成分の特性を示し，それが音源の特性を表す．このように，それぞれの成分を独立して扱うことができる．

音声信号の波形は時間変化に伴い刻々と変化するが，非常に短い時間フレームで切り取って見

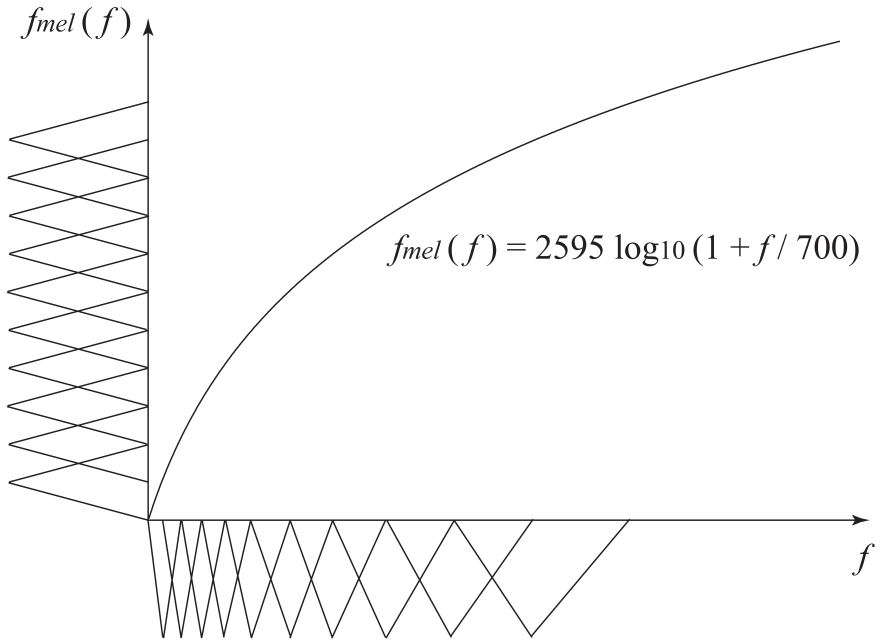

図 2.2: メル周波数軸上に等間隔で配置された三角窓

るとそのフレーム内では定常状態とみなすことができるため，その区間ごとにケプストラムを求めていく．フレームを一定のシフト長ごとにずらして切り出し, それぞれのフレームからケプストラムを抽出していくことで，その音声全体のケプストラム系列を得ることができる．ケプストラムは次元間の相関がほとんどないことが知られており，ケプストラムに対する様々な計算を簡便に行うことができる．

2.2.2　聴覚的特性に基づくケプストラム

人間の音の高さに対する知覚特性は，低周波ほど分解能が細かく，高周波ほど分解能が粗くなっており，周波数分解能が周波数に対してほぼ対数に近い特性となっている．この特性を表す尺度としてメル尺度があり，これに合わせて音声分析を行うことで，より人間の感覚に合った特徴量を抽出できる．そこで，メル尺度をケプストラムに反映させた特徴量が提案されている．その一つとして，Mel-Frequency Cepstrum Coefficient (MFCC) がある．図 2.2 のようにメル周波数 (メル尺度化された周波数) 軸上に等間隔で配置された三角窓を用意し，フィルタバンク分析を行うことでメル周波数が求められる．なお，メル周波数 f_{mel} は周波数 f[Hz] に対して，

$$f_{mel}(f) = 2595 \log_{10}(1 + \frac{f}{700}) \tag{2.1}$$

という周波数ウォーピングを施すことで得られる．このメルフィルタバンクの出力を求めることでメルスペクトルが得られ，さらに離散コサイン変換を施すことで，MFCC が求められる．

2.3　音響モデル

人間の音声活動において，音素列 P を発話する上でどのような音響特徴量系列 $\boldsymbol{O}$ が出力されるかを表現するモデルが音響モデルである．このモデルは，音響的特徴に関するあるパラメータ

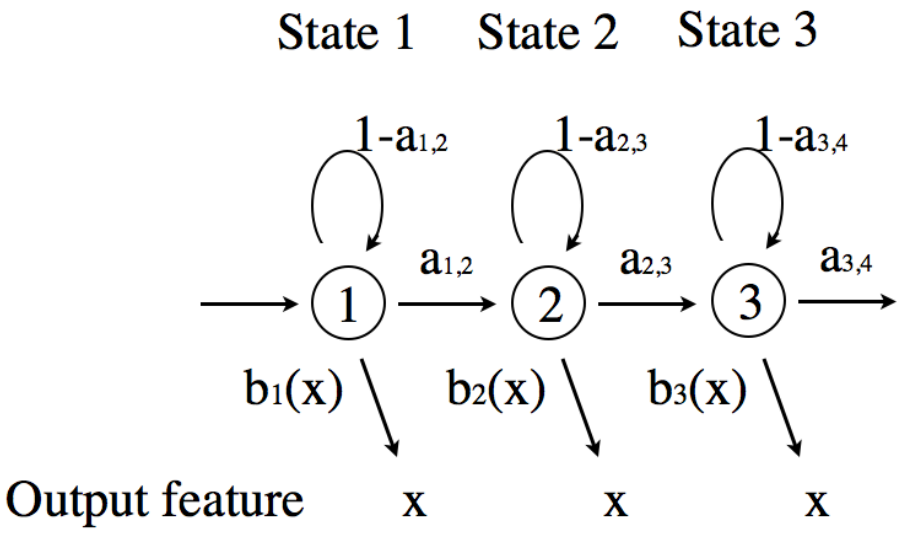

図 2.3: 隠れマルコフモデル (HMM)

θ で制御されるモデルとして考えることができる．音響モデルは，$p(\boldsymbol{O}|P;\theta_1)$ を最大化する θ_1 を学習することで得られる生成モデルや，$p(P|\boldsymbol{O};\theta_2)$ を最大化する θ_2 を学習することで得られる識別モデルに分類される．

2.3.1　隠れマルコフモデル (HMM)

音声は発声によって時間的に長さ，速さが変化し，また，現在どのような音響事象にあるかを音響特徴量から直接観測することができないという特性を持つ．そこで一般的には，生成モデルとして状態を隠れ変数とする隠れマルコフモデル (Hidden Markov Model; HMM) を利用することが多い [10]．HMM の概念図を図 2.3 に示す．S_i は i 番目の状態，a_{ij} は状態 S_i から状態 S_j への状態遷移確率，$b_i(\boldsymbol{o})$ は状態 S_i から音声特徴量 $\boldsymbol{o}$ が出力される出力確率である．出力確率 $b_i(\boldsymbol{o})$ の分布形としては，ガウス分布を複数用意し，その重み付け和で $b_i(\boldsymbol{o})$ を表現する混合ガウス分布 (Gaussian Mixture Model; GMM) がよく用いられる．HMM のパラメータ a_{ij}，$b_i(\boldsymbol{o})$ を θ とおけば，ある音素 P_1 に対する HMM を利用することで $p(\boldsymbol{O}|P_1;\theta)$ をモデル化できる．

一つの音素につき一つの HMM を用いることが一般的であり，これを音素 HMM と呼ぶ．音声は時間に伴って変化する巻き戻し不可能なものであるため，HMM のトポロジーとしては図 2.3 のような 3 状態程度の Left-to-Right 型が利用されている．音素の特徴は調音結合により前後の音素に依存して変化するため，音素 HMM には調音結合を考慮していないものとそれを考慮するものとがある．前者は monophon，後者は triphone と呼ばれる．また，音素 HMM を連結した音素列 HMM を用いることで，音素列 P について $p(\boldsymbol{O}|P;\theta)$ をモデル化できる．音素列として単語を単位として作成したモデルは，単語 HMM と呼ばれる．

2.3.2　HMM の学習

音素列 P に対応する $\boldsymbol{O}$ が得られるモデルを考えるとき，HMM の学習すべきパラメータは $\theta = \{a_{ij}^p, b_i^p(\boldsymbol{o})\}$ であり，これを最尤 (Maximum Likelihood; ML) 推定する．それは

$$\underset{\theta}{\operatorname{argmax}}\, p(\boldsymbol{O}|P;\theta) \tag{2.2}$$

を求めることで行われる．しかし，HMM ではどのような状態遷移を通して $\boldsymbol{O}$ が出力されたのかは未知であり，その状態が隠れ変数となっている．つまり，ある出力系列 $\boldsymbol{O}$ が観測されたときに各々がどの状態から生じたものなのか観測することはできないため，式 (2.2) を解析的に解くことは不可能である．しかし，ある時刻 t における出力がある状態から出力されたものである確率を推定することは可能である．したがって，隠れ変数が存在する統計モデルの ML 推定値を一般的に見出すことができる Expectation-Maximization (EM) アルゴリズムを用いてその局所最適解を得る．ここで，EM アルゴリズムの収束結果は各パラメータの初期値に依存するため，その初期値設定は非常に重要である．初期値設定法は様々であり，全データのグローバルな平均と分散をすべての HMM の初期値とするフラットスタートや，triphone の初期値に monophone のパラメータを用いる手法などがある．EM アルゴリズムの実装では，HMM の EM 学習に特化して効率を高めたアルゴリズムである Baum-Welch アルゴリズムを利用できる．

2.3.3　HMM による音素認識

観測された音響特徴量の時系列をもとに音声認識することは，その音響特徴量の時系列 $\boldsymbol{O}$ を出力する確率が最も高くなるような音素列 P の音素列 HMM を求める事である．つまり，音素列 P の音素列 HMM から音響特徴量の時系列 $\boldsymbol{O}$ が出力される事後確率 $p(P|\boldsymbol{O};\theta_i)$ が最大になる P を求めれば良い．ベイズの定理により

$$\begin{aligned}
\underset{P}{\operatorname{argmax}}\, p(P|\boldsymbol{O};\theta) &= \underset{P}{\operatorname{argmax}}\, \frac{p(P,\boldsymbol{O};\theta)}{p(\boldsymbol{O};\theta)} \\
&= \underset{P}{\operatorname{argmax}}\, p(P,\boldsymbol{O};\theta) \\
&= \underset{P}{\operatorname{argmax}}\, p(\boldsymbol{O}|P;\theta)p(P;\theta)
\end{aligned} \tag{2.3}$$

となり，これを求めればよい．ここで $p(\boldsymbol{O}|P;\theta)$ は HMM でモデル化されており，$p(P;\theta)$ は，言語的な制約条件によってモデル化されている．

実際の計算では，隠れ変数である HMM の状態がどのように遷移したかを考慮する必要がある．例として，音響特徴量の時系列データ $\boldsymbol{O} = \{\boldsymbol{o}(1), \boldsymbol{o}(2), \cdots, \boldsymbol{o}(7)\}$ が，ある音素 P に対応する音素 HMM から出力される場合において，とりうる状態遷移の経路を図 2.4 に示す．ある 1 つの経路を通って $\boldsymbol{O}$ が出力される確率は，その経路の状態遷移確率 a_{ij} と経路上の各状態での出力確率 $b_i(\boldsymbol{o})$ の積によって計算できる．図 2.4 に示された経路全てに対してこの確率を求めて和をとることで，音素 P の HMM から音響特徴量の時系列 $\boldsymbol{O}$ か出力される確率，すなわち $p(\boldsymbol{O}|P;\theta)$ を求めることができる．しかし，全ての経路からの出力確率の和をとると計算量が増大してしまうため，実際には最も出力確率の大きな経路のみを計算し，その確率値で $p(\boldsymbol{O}|P;\theta)$ を近似する Viterbi アルゴリズムが用いられる．このような近似は Viterbi 近似と呼ばれる．

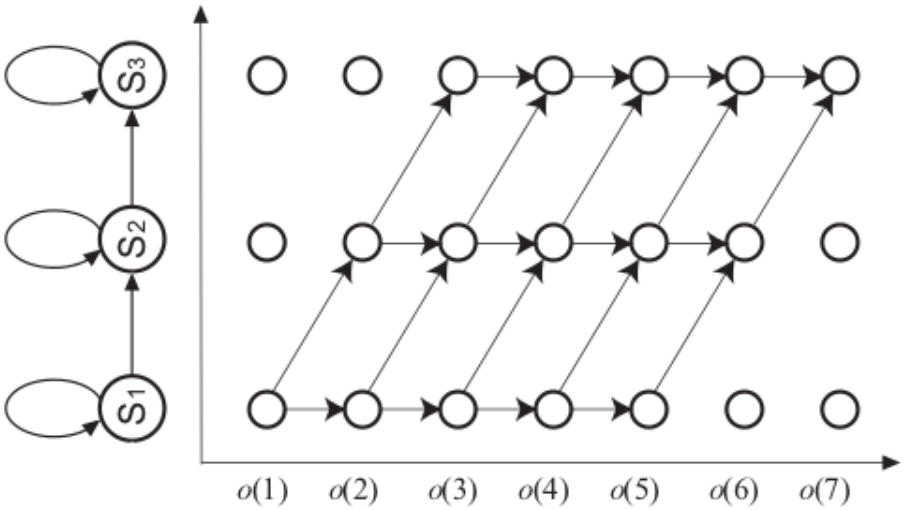

図 2.4: HMM の状態遷移の経路

2.4　雑音環境下での音声認識

2.4.1　雑音の性質

雑音は大きく加算性雑音と乗算性雑音の 2 つにわけられる．加算性雑音の例としては，生活音や自動車の走行音といった背景雑音が挙げられる．乗算性雑音の例としては，電話などのチャネル歪みが挙げられる．本論文では，主に加算性雑音の乗った音声を想定する．パワースペクトルでは，音声と雑音は足し合わせで表現されるため，音声と雑音を明示的に分離することができる．一方，MFCC では，雑音は特徴量の非線形な歪みとして表れるため，明示的に分離することが困難である．そこで，クリーンな音声と雑音の乗った音声の関係をモデル化し，統計的に扱う必要がある．

2.4.2　雑音に頑健な音声認識のためのアプローチ

2.3.3 節で述べた HMM による認識は，観測された音響特徴量と音響モデルとのマッチングと考えることができる．音響モデルを学習する際の環境と実際の音声を入力する際の環境のミスマッチが起こると，認識率が低下する．そのミスマッチの大きな要因が雑音である．雑音によるミスマッチ低減するための手法は大きく次の 2 つに分けられる．雑音の乗った音声を雑音のないクリーンな音声に近づける手法と音響モデルを雑音の乗った音声に適応させる手法である．前者の手法は音声強調，後者はモデル適応と呼ばれる．本論文では，音声強調に注目し，第 3 章でその具体的な手法を解説する．

第3章

雑音環境下音声認識のための音声強調手法

3.1　はじめに

雑音環境下音声認識のためのアプローチとして，雑音重畳音声を音響モデルに近づける音声強調の手法に注目する．音声強調はモデルベースの手法と事例ベースの手法の大きく 2 つがある．モデルベースの手法では，学習データを統計的にモデル化し，そのモデルに基づいてクリーン音声の推定を行う．事例ベースの手法では，学習データから得られた多数のサンプルを用いてクリーン音声の推定を行う．本章では，それぞれについて具体的な手法を紹介する．

3.2　GMM を用いたモデルベースの音声強調

モデルベースの音声強調の例として，Gaussian Mixture Model (GMM) を用いた 4 つの手法を紹介する．

まず，GMM の概要を述べる．GMM は任意の確率分布をガウス分布の重み付きの線形重ね合わせで表現したモデルである [11]．この時，重ね合わせるガウス分布の数を混合数と呼ぶ．ある特徴量 $\boldsymbol{x}$ に対する混合数 M の GMM は以下のように表される．

$$\boldsymbol{p}(\boldsymbol{x}) = \sum_{m=1}^{M} \pi_m \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m) \tag{3.1}$$

$$\text{ただし}\ \ 0 \leqq \pi \leqq 1,\ \ \sum_{m=1}^{M} \pi_m = 1$$

ここで π_m，$\boldsymbol{\mu}_m$，$\boldsymbol{\Sigma}_m$ はそれぞれインデックス m のガウス分布の重み，平均，分散である．また，$\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$ はガウス分布を表し，ベクトル $\boldsymbol{x}$ を Ω 次元とすると

$$\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m) = \frac{1}{(2\pi)^{\Omega/2}} \frac{1}{|\Sigma|^{1/2}} \exp\left\{ -\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu}_m)^\top \boldsymbol{\Sigma}_m^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_m) \right\} \tag{3.2}$$

と書ける．また確率分布が与えられた場合の各ガウス分布の重み π_m，平均 $\boldsymbol{\mu}_m$，分散 $\boldsymbol{\Sigma}_m$ は EM アルゴリズムによって求められる．

3.2.1　VTS 強調 [12]

Vector Taulor Series (VTS) は音声強調と音響モデルの雑音環境適応のどちらにも用いられており [12]，音声強調は雑音重畳音声音声の特徴量 $\boldsymbol{y}$ をクリーンな音声の特徴量 $\boldsymbol{x}$ と雑音の特徴量 $\boldsymbol{n}$ から VTS 近似することで実現される．ここでは，簡単のため特徴量として FBANK[1] を用いた場合を紹介する．

まず，以下のクリーン音声特徴量 $\boldsymbol{x}$ の GMM と雑音特徴量 $\boldsymbol{n}$ のガウス分布を学習する．

$$p(\boldsymbol{x}) = \sum_m p(\boldsymbol{x}, m) = \sum_m p(\boldsymbol{x}|m) p(m) \tag{3.3}$$

$$p(\boldsymbol{n}) = \mathcal{N}(\boldsymbol{n}; \boldsymbol{\mu}^n, \boldsymbol{\Sigma}^n) \tag{3.4}$$

ただし，

$$p(\boldsymbol{x}|m) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_m^x, \boldsymbol{\Sigma}_m^x) \tag{3.5}$$

[1] 音声波形の対数パワースペクトルにメルフィルタバンクをかけたもの

ここで，m，$\boldsymbol{\mu}_m^x$，$\boldsymbol{\Sigma}_m^x$ は $\boldsymbol{x}$ についての GMM の各ガウス分布のインデックス，分布 m の平均と分散である．また，$\boldsymbol{\mu}^n$，$\boldsymbol{\Sigma}^n$ は $\boldsymbol{n}$ についてのガウス分布の平均と分散である．

雑音重畳音声の特徴量 $\boldsymbol{y}$ は以下のように近似できる．

$$\boldsymbol{y} \approx g(\boldsymbol{x}, \boldsymbol{n}) = \log(\exp \boldsymbol{x} + \exp \boldsymbol{n}) \tag{3.6}$$

ここで，$\boldsymbol{y} \approx g(\boldsymbol{x}, \boldsymbol{n}, m)$ に対し，$\boldsymbol{x}$，$\boldsymbol{n}$ をそれぞれ $\boldsymbol{\mu}_m^{\boldsymbol{x}}$，$\boldsymbol{\mu}^{\boldsymbol{n}}$ を中心とした 1 次の VTS 近似を行うと以下のようになる．

$$\begin{aligned} \boldsymbol{y} &\approx g(\boldsymbol{x}, \boldsymbol{n}, m) = \log(\exp \boldsymbol{x}_t + \exp \boldsymbol{n}) && (3.7) \\ &\approx g(\boldsymbol{\mu}_m^{\boldsymbol{x}}, \boldsymbol{\mu}^{\boldsymbol{n}}) + \frac{\boldsymbol{x} - \boldsymbol{\mu}_m^{\boldsymbol{x}}}{\mathbf{1} + \exp(\boldsymbol{\mu}^{\boldsymbol{n}} - \boldsymbol{\mu}_m^{\boldsymbol{x}})} + \frac{\boldsymbol{n} - \boldsymbol{\mu}^{\boldsymbol{n}}}{\exp(\boldsymbol{\mu}_m^{\boldsymbol{x}} - \boldsymbol{\mu}^{\boldsymbol{n}}) + \mathbf{1}} && (3.8) \end{aligned}$$

log，exp，分数計算はそれぞれベクトルの要素ごとの演算，$\mathbf{1}$ は全ての要素が 1 のベクトルとする．このとき，

$$\begin{aligned} p(\boldsymbol{y}|m) &= \mathcal{N}(\boldsymbol{y}; \boldsymbol{\mu}_m^{\boldsymbol{y}}, \boldsymbol{\Sigma}_m^{\boldsymbol{y}}) && (3.9) \\ \boldsymbol{\mu}_m^{\boldsymbol{y}} &= g(\boldsymbol{\mu}_m^{\boldsymbol{x}}, \boldsymbol{\mu}^{\boldsymbol{n}}) && (3.10) \\ \boldsymbol{\Sigma}_m^{\boldsymbol{y}} &= \frac{\boldsymbol{\Sigma}_m}{(\mathbf{1} + \exp(\boldsymbol{\mu}^{\boldsymbol{n}} - \boldsymbol{\mu}_m^{\boldsymbol{x}}))^2} + \frac{\boldsymbol{\Sigma}^n}{(\exp(\boldsymbol{\mu}_m^{\boldsymbol{x}} - \boldsymbol{\mu}^{\boldsymbol{n}}) + \mathbf{1})^2} && (3.11) \end{aligned}$$

として $\boldsymbol{y}$ の確率分布を求める事ができる．したがって，以下のように $p(m|\boldsymbol{y})$ も求められる．

$$\begin{aligned} p(m|\boldsymbol{y}) &= \frac{p(m, \boldsymbol{y})}{p(\boldsymbol{y})} \\ &= \frac{p(\boldsymbol{y}|m)p(m)}{\sum_{m=1}^{M} p(\boldsymbol{y}|m)p(m)} \end{aligned} \tag{3.12}$$

これを用いて，以下のようにクリーン音声の特徴量の推定値を求める．

$$\begin{aligned} \hat{\boldsymbol{x}} &= \operatorname*{argmax}_{\boldsymbol{x}} p(\boldsymbol{x}|\boldsymbol{y}) \\ &= \operatorname*{argmax}_{\boldsymbol{x}} \sum_m p(\boldsymbol{x}, m|\boldsymbol{y}) \\ &= \operatorname*{argmax}_{\boldsymbol{x}} \sum_m p(m|\boldsymbol{y}) p(\boldsymbol{x}|\boldsymbol{y}, m) && (3.13) \\ &\approx \sum_m p(m|\boldsymbol{y}) g(\boldsymbol{\mu}_m^{\boldsymbol{x}}, \boldsymbol{\mu}^{\boldsymbol{n}}) && (3.14) \end{aligned}$$

入力の雑音に応じて式 (3.11) を計算するため，精度はよいが計算コストが大きくなってしまう．特徴量として FBANK を用いた場合は式 (3.11) が対角になるため計算コストはある程度抑えられるが，MFCC を用いた場合はそれが全角になるため，計算コストが非常に大きくなってしまうという欠点がある．

3.2.2　SPLICE[13]

MFCC 領域において，雑音による音声の歪みは特徴量の非線形な変換として表される．Stereo Piecewise LInear Compensation for Environments (SPLICE) [13] では，その非線形な変換が微

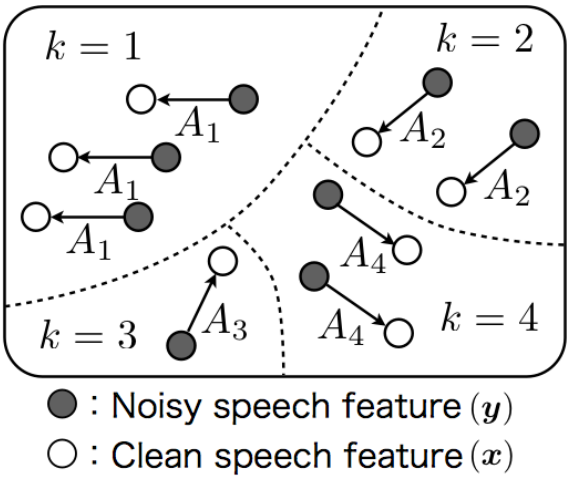

図 3.1: 区分的線形変換

小区間においては線形変換で近似されるという仮定に基づいて MFCC 領域における音声強調を行う．図 3.1 のような区分的線形変換により雑音重畳音声の特徴量 $\boldsymbol{y}$ からクリーン音声の特徴量の推定値 $\hat{\boldsymbol{x}}$ を求める．

学習データとしては，時系列の対応がとれたクリーン音声の特徴量系列 $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T]$ と雑音重畳音声の特徴量系列 $\boldsymbol{Y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_T]$ のステレオデータを用いる．特徴量の次元数を N，フレーム数を T とする．まず，$\boldsymbol{Y}$ を用いて $\boldsymbol{y}$ を出力する混合数 M の GMM を学習する．これにより，GMM の m 番目の分布の重み $p(m)$ とその分布 $p(\boldsymbol{y}|m)$ の平均，分散のパラメータが推定できる．

これらを用いることで，以下のようにクリーン音声の推定値 $\hat{\boldsymbol{x}}$ が得られる．

$$\hat{\boldsymbol{x}} = \sum_{m=1}^{M} p(m|\boldsymbol{y}) \boldsymbol{A}_m \begin{bmatrix} 1 \\ \boldsymbol{y} \end{bmatrix} \tag{3.15}$$

$$\begin{aligned} p(m|\boldsymbol{y}) &= \frac{p(m, \boldsymbol{y})}{p(\boldsymbol{y})} \\ &= \frac{p(\boldsymbol{y}|m)p(m)}{\sum_{m=1}^{M} p(\boldsymbol{y}|m)p(m)} \end{aligned} \tag{3.16}$$

つまり，図 3.1 のように GMM の各分布ごとに線形変換 $\boldsymbol{A}_m$ をかければよい．ここで，$\boldsymbol{A}_m$ は $N \times (N+1)$ 行列であり，重み付き最小二乗誤差基準で以下のように推定する．

$$\boldsymbol{A}_m = \underset{\boldsymbol{A}_m}{\operatorname{argmin}} \sum_{t=1}^{T} p(m|\boldsymbol{y}_t) |\boldsymbol{x}_t - \boldsymbol{A}_m \begin{bmatrix} 1 \\ \boldsymbol{y}_t \end{bmatrix}|^2 \tag{3.17}$$

これは解析的に以下のように解ける．

$$\boldsymbol{A}_m = \boldsymbol{X} \boldsymbol{P}_m \overline{\boldsymbol{Y}} (\overline{\boldsymbol{Y}} \boldsymbol{P}_M \overline{\boldsymbol{Y}})^{(-1)} \tag{3.18}$$

ここで，$\overline{\boldsymbol{Y}}$ は $\begin{bmatrix} 1 \\ \boldsymbol{y}_t \end{bmatrix}$ をフレーム数分並べた行列であり，$\boldsymbol{P}_m$ は $p(m|\boldsymbol{y}_t)$ をフレーム数分並べたものを対角成分にもつ対角行列である．

SPLICE は事前に GMM と変換が学習されており，音声強調を行う際の計算コストが非常に低い．しかし，学習データの雑音環境に基づいて学習しているため，学習に用いた既知の雑音には非常に強いが未知雑音や非定常雑音に対しては弱いという特徴がある．

ステレオデータを用いる手法に対しては，雑音重畳音声の学習データのみからクリーン音声を推定し，それをステレオデータとして利用する手法も検討されており，ステレオデータがなくとも適用できうる [14, 15].

3.2.3　SSM[16]

Stereo-based Stochasstic Mapping (SSM) [16] も SPLICE と同様に，MFCC 領域において，雑音による非線形な変換を近似した GMM をステレオデータ $\boldsymbol{X}$，$\boldsymbol{Y}$ を用いて学習し，それに基づいた区分的線形変換によりクリーン音声特徴量を推定する手法である．

まず，ステレオデータを用いて事前に GMM の学習を行う．SPLICE と異なる点は，対応する $\boldsymbol{x}$ と $\boldsymbol{y}$ の結合ベクトル $\boldsymbol{z} = [\boldsymbol{x}^\mathsf{T}, \boldsymbol{y}^\mathsf{T}]^\mathsf{T}$ を特徴量とした以下の GMM を学習しておく点である．

$$p(\boldsymbol{z}) = \sum_{m=1}^{M} c_m \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}_{z,m}, \boldsymbol{\Sigma}_{zz,m}) \tag{3.19}$$

$$\boldsymbol{\mu}_{z,m} = \begin{pmatrix} \boldsymbol{\mu}_{x,m} \\ \boldsymbol{\mu}_{y,m} \end{pmatrix} \tag{3.20}$$

$$\boldsymbol{\Sigma}_{zz,m} = \begin{pmatrix} \boldsymbol{\Sigma}_{xx,m} & \boldsymbol{\Sigma}_{xy,m} \\ \boldsymbol{\Sigma}_{yx,m} & \boldsymbol{\Sigma}_{yy,m} \end{pmatrix} \tag{3.21}$$

これを用いて，以下のように Maximum a Posteriori (MAP) 基準でクリーン音声特徴量の推定値 $\hat{\boldsymbol{x}}$ を求めることができる．

$$\begin{aligned} \hat{\boldsymbol{x}} &= \operatorname*{argmax}_{\boldsymbol{x}} p(\boldsymbol{x}|\boldsymbol{y}) \\ &= \operatorname*{argmax}_{\boldsymbol{x}} \sum_m p(\boldsymbol{x}, m|\boldsymbol{y}) \\ &\equiv \operatorname*{argmax}_{\boldsymbol{x}} \log \sum_m p(\boldsymbol{x}, m|\boldsymbol{y}) \end{aligned} \tag{3.22}$$

ここで，推定の際の入力は $\boldsymbol{y}$ のみであり，m についての知識は事前の GMM 学習で得られる $p(m|\boldsymbol{x}, \boldsymbol{y})$ のみである．そこで，$p(m|\bar{\boldsymbol{x}}, \boldsymbol{y})$ を用いて $\log p(\boldsymbol{x}, m|\boldsymbol{y})$ の期待値を求める．$\bar{\boldsymbol{x}}$ はあら

かじめ得られた推定値である．

$$\hat{\boldsymbol{x}} = \underset{\boldsymbol{x}}{\operatorname{argmax}} F(\boldsymbol{x}, \boldsymbol{y}) \tag{3.23}$$

$$\begin{aligned} F(\boldsymbol{x}, \boldsymbol{y}) &= \sum_m p(m|\bar{\boldsymbol{x}}, \boldsymbol{y}) \log p(\boldsymbol{x}, m|\boldsymbol{y}) \\ &= \sum_m p(m|\bar{\boldsymbol{x}}, \boldsymbol{y}) \log p(m|\boldsymbol{y}) p(\boldsymbol{x}|m, \boldsymbol{y}) \\ &= \sum_m p(m|\bar{\boldsymbol{x}}, \boldsymbol{y}) \{\log p(m|\boldsymbol{y}) + \log p(\boldsymbol{x}|m, \boldsymbol{y})\} \\ &= \sum_m p(m|\bar{\boldsymbol{x}}, \boldsymbol{y}) \{\log p(\boldsymbol{x}|m, \boldsymbol{y}) - \frac{1}{2}[\log |\boldsymbol{\Sigma}_{x|y,m}| + (\boldsymbol{x} - \boldsymbol{\mu}_{x|y,m})^T \boldsymbol{\Sigma}_{x|y,m}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_{x|y,m})]\} \end{aligned}$$

$$\begin{aligned} \hat{\boldsymbol{x}} &= \underset{\boldsymbol{x}}{\operatorname{argmax}} F(\boldsymbol{x}, \boldsymbol{y}) \\ &= \underset{\boldsymbol{x}}{\operatorname{argmax}} -\frac{1}{2} \sum_m p(m|\bar{\boldsymbol{x}}, \boldsymbol{y}) (\boldsymbol{x} - \boldsymbol{\mu}_{x|y,m})^T \boldsymbol{\Sigma}_{x|y,m}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_{x|y,m}) \end{aligned} \tag{3.24}$$

これを繰り返すことにより，クリーン音声特徴量の推定値 $\hat{\boldsymbol{x}}$ を求める．

ここで，式 (3.24) を $\boldsymbol{x}$ について微分して得られた導関数が 0 になるような $\boldsymbol{x}$ が $\hat{\boldsymbol{x}}$ であることから，

$$\begin{aligned} \sum_m p(m|\bar{\boldsymbol{x}}, \boldsymbol{y}) \boldsymbol{\Sigma}_{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{m}}^{-\mathbf{1}} \hat{\boldsymbol{x}} &= \sum_m p(m|\bar{\boldsymbol{x}}, \boldsymbol{y}) \boldsymbol{\Sigma}_{x|y,m}^{-1} \boldsymbol{\mu}_{x|y,m} \\ \hat{\boldsymbol{x}} = (\sum_m p(m|\bar{\boldsymbol{x}}, \boldsymbol{y}) \boldsymbol{\Sigma}_{x|y,m}^{-1})^{-1} &\sum_m p(m|\bar{\boldsymbol{x}}, \boldsymbol{y}) \boldsymbol{\Sigma}_{x|y,m}^{-1} \boldsymbol{\mu}_{x|y,m} \end{aligned} \tag{3.25}$$

ただし，

$$\boldsymbol{\mu}_{x|y,m} = \boldsymbol{\mu}_{x,m} + \boldsymbol{\Sigma}_{xy,m} \boldsymbol{\Sigma}_{yy,m}^{-1} (\boldsymbol{y} - \boldsymbol{\mu}_{y,m}) \tag{3.26}$$

$$\boldsymbol{\Sigma}_{x|y,m} = \boldsymbol{\Sigma}_{xx,m} - \boldsymbol{\Sigma}_{xy,m} \boldsymbol{\Sigma}_{xy,m}^{-1} \boldsymbol{\Sigma}_{yx,m} \tag{3.27}$$

である．さらに，式 (3.25) を変形すると，

$$\begin{aligned} \hat{\boldsymbol{x}} &= \sum_m p(m|\bar{\boldsymbol{x}}, \boldsymbol{y}) (\sum_m p(m|\bar{\boldsymbol{x}}, \boldsymbol{y}) \boldsymbol{\Sigma}_{x|y,m}^{-1})^{-1} \boldsymbol{\Sigma}_{x|y,m}^{-1} \boldsymbol{\mu}_{x|y,m} \\ &= \sum_m p(m|\bar{\boldsymbol{x}}, \boldsymbol{y}) (\boldsymbol{A}_m \boldsymbol{y} + \boldsymbol{b}_m) \end{aligned} \tag{3.28}$$

ただし，

$$\boldsymbol{A}_m = \boldsymbol{C}\boldsymbol{D}_m, \boldsymbol{b}_m = \boldsymbol{C}\boldsymbol{e}_m$$

$$\boldsymbol{C} = (\sum_m p(m|\bar{\boldsymbol{x}}, \boldsymbol{y}) \boldsymbol{\Sigma}_{x|y,m}^{-1})^{-1} \tag{3.29}$$

$$\boldsymbol{D}_m = \boldsymbol{\Sigma}_{x|y,m}^{-1} \boldsymbol{\Sigma}_{yy,m}^{-1} \boldsymbol{\Sigma}_{xy,m} \tag{3.30}$$

$$\boldsymbol{e}_m = \boldsymbol{\Sigma}_{x|y,m}^{-1} (\boldsymbol{\mu}_{x,m}) - \boldsymbol{\Sigma}_{yy,m}^{-1} \boldsymbol{\Sigma}_{xy,m} \boldsymbol{\mu}_{y,m} \tag{3.31}$$

となり，SPLICE と同様の区分的線形変換となる．SPLICE が最小二乗誤差基準であるのに対し SSM では MAP 基準と変換の基準は異なっているものの，数式上は同様の区分的線形変換によっ

て音声強調が実現される．そのため，最小二乗誤差基準での SSM の実装も存在する．SPLICE との違いは，結合ベクトル $\boldsymbol{z}$ の GMM を用いている点，一度推定されたクリーン音声特徴量 $\bar{\boldsymbol{x}}$ のフィードバックを用いて繰り返し推定を行う点の 2 点であり，これによってより入力音声に適した変換ができると期待される．Afify らの実験によると [16]，SPLICE よりも SSM の方が認識率がよいという結果が出ている．

SSM についても雑音重畳音声の学習データのみから推定されたクリーン音声をステレオデータとしてして利用する手法も検討されている [17].

3.2.4　REDIAL[18, 19]

VTS 強調は，MFCC で行う場合は全角の共分散をもつ GMM の事後確率計算があるため非常に大きな計算コストがかかる．その一方で，事前学習のみで行う SPLICE と比べて未知雑音や非定常雑音に強い．そこで，事後確率計算を直接行わず，識別モデルを利用して推定する REgularized piecewise linear mapping with DIscriminative region weighting And Longspan features (REDIAL) が提案されている [18, 19].

学習データとして，SPLICE と同様にパラレルなクリーン音声特徴量 $\{\boldsymbol{x}_t\}_{t=1,\ldots,T}$ とノイジー音声特徴量 $\{\boldsymbol{y}_t\}_{t=1,\ldots,T}$ を用いる．

まず，T フレームのクリーン音声特徴量 $\{\boldsymbol{x}_t\}_{t=1,\ldots,T}$ の GMM（クリーン GMM）を以下のように学習する．

$$p(\boldsymbol{x}_t) = \sum_{m=1}^{M} \pi_m^{\boldsymbol{x}} \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{\mu}_m^{\boldsymbol{x}}, \boldsymbol{\Sigma}_m^{\boldsymbol{x}}) \tag{3.32}$$

$\pi_m^{\boldsymbol{x}}$，$\boldsymbol{\mu}_m^{\boldsymbol{x}}$，$\boldsymbol{\Sigma}_m^{\boldsymbol{x}}$ は，それぞれ m 番目のインデックスに対応する GMM の重み，平均ベクトル，分散ベクトルである．ここで，事後確率 $\{p(m|\boldsymbol{x}_t)\}_{m=1,\ldots,M}$ を，$\boldsymbol{x}_t$ を観測することなく推定することを考える．ノイジー音声特徴量 $\boldsymbol{y}_{tt=1,\ldots,T}$ を前後数フレーム連結した特徴量 $\boldsymbol{d}_t$ から $\{p(m|\boldsymbol{x}_t)\}_{m=1,\ldots,M}$ を識別モデルによって推定する．識別モデルとしては，事後確率に基づくソフトな LDA を用いる．これにより，LDA の教師ラベルとして $\{p(m|\boldsymbol{x}_t)\}_{m=1,\ldots,M}$ を用い，出力として事後確率 $\{p(s|\boldsymbol{d}_t)\}_{s=1,\ldots,S}$ が得られる．

ソフトな LDA による事後確率の推定は次のように行う．まず，学習データ $\{\{p(m|\boldsymbol{x}_t)\}_{m=1,\ldots,M}, \boldsymbol{d}_t\}_{t=1,\ldots,T}$ を用いて $\boldsymbol{L}\boldsymbol{d}_t$ のように次元圧縮を行う行列 $\boldsymbol{L}$ を求める．

$$\boldsymbol{L} = \underset{\boldsymbol{W}}{\operatorname{argmin}} \frac{\boldsymbol{W}^\top \boldsymbol{\Sigma}^w \boldsymbol{W}}{\boldsymbol{W}^\top \boldsymbol{\Sigma}^b \boldsymbol{W}} \tag{3.33}$$

$$\boldsymbol{\Sigma}^w = \sum_{m=1}^{M} \sum_{t=1}^{T} p(m|\boldsymbol{x}_t) \left(\boldsymbol{d}_t - \boldsymbol{\mu}_m^w\right) \left(\boldsymbol{d}_t - \boldsymbol{\mu}_m^w\right)^\top \tag{3.34}$$

$$\boldsymbol{\Sigma}^b = \sum_{m=1}^{M} \left(\textstyle\sum_{t=1}^{T} p(m|\boldsymbol{x}_t)\right) \left(\boldsymbol{\mu}_m^w - \frac{\sum_{t=1}^{T} \boldsymbol{d}_t}{T}\right) \left(\boldsymbol{\mu}_m^w - \frac{\sum_{t=1}^{T} \boldsymbol{d}_t}{T}\right)^\top \tag{3.35}$$

$$\boldsymbol{\mu}_m^w = \frac{1}{\sum_{t=1}^{T} p(m|\boldsymbol{x}_t)} \sum_{t=1}^{T} p(m|\boldsymbol{x}_t) \boldsymbol{d}_t \tag{3.36}$$

次に，次元圧縮された空間 $\boldsymbol{v}_t = \boldsymbol{L}\boldsymbol{d}_t$ において，混合数 S の GMM を以下のように学習する．

$$p(\boldsymbol{v}_t) = \sum_{s=1}^{S} \pi_s^{\boldsymbol{v}} \mathcal{N}(\boldsymbol{v}_t; \boldsymbol{\mu}_s^{\boldsymbol{v}}, \boldsymbol{\sigma}_s^{\boldsymbol{v}}) \tag{3.37}$$

$\pi_s^{\boldsymbol{v}}$，$\boldsymbol{\mu}_s^{\boldsymbol{v}}$，$\boldsymbol{\Sigma}_s^{\boldsymbol{v}}$ は，それぞれ s 番目のインデックスに対応する分布の重み，平均ベクトル，分散ベクトルである．

ここで，$\boldsymbol{v}_t$ は $\{p(m|\boldsymbol{x}_t)\}_{m=1,\ldots,M}$ の情報を保存するように次元圧縮された空間であるので，その空間内で学習された GMM のインデックスの事後確率 $\{p(s|\boldsymbol{v}_t)\}_{s=1,\ldots,S}$ は $\{p(m|\boldsymbol{x}_t)\}_{m=1,\ldots,M}$ の m とは直接的関係はないものの，似たような情報を持っていると考えられる．音声強調においては，クリーン状態が異なる場合に異なるインデックスの事後確率が高くなればよいので，$\{p(s|\boldsymbol{v}_t)\}_{s=1\ldots,S}$ を直接音声強調に利用できる．したがって，以下のように $\boldsymbol{d}_t$ からインデックス s の事後確率を計算する．

$$\begin{aligned} p(s|\boldsymbol{d}_t) &= p(m|\boldsymbol{v}_t) \\ &= \frac{\pi_s^{\boldsymbol{v}}\mathcal{N}(\boldsymbol{v}_t;\boldsymbol{\mu}_s^{\boldsymbol{v}},\boldsymbol{\Sigma}_s^{\boldsymbol{v}})}{\sum_{s=1}^{S}\pi_s^{\boldsymbol{v}}\mathcal{N}(\boldsymbol{v}_t;\boldsymbol{\mu}_s^{\boldsymbol{v}},\boldsymbol{\Sigma}_s^{\boldsymbol{v}})} \end{aligned} \tag{3.38}$$

認識時には，クリーン音声特徴量の推定値 $\hat{\boldsymbol{x}}_t$ は以下のような区分的線形変換で求められる．

$$\hat{\boldsymbol{x}}_t = \sum_{s=1}^{S} p(s|\boldsymbol{d}_t)\boldsymbol{A}_s \begin{bmatrix} 1 \\ \boldsymbol{d}_t \end{bmatrix} \tag{3.39}$$

SPLICE と同様に雑音重畳音声から得られる特徴量 $\boldsymbol{d}_t$ にバイアス項 1 を加えている．

$\boldsymbol{A}_S$ は最小二乗誤差基準で SPLICE では $\boldsymbol{y}_t$ を用いているのに対し，$\boldsymbol{d}_t$ であり，次元数が大きくなっているため，過学習が起こりうる．そこで，最小二乗誤差基準での $\boldsymbol{A}_S$ の各要素が小さくなるように正則化を行う．二次の正則化項を導入する場合の $\boldsymbol{A}_s$ の学習は，以下のように行う．

$$\boldsymbol{A}_s = \operatorname*{argmin}_{\boldsymbol{A}_s} \sum_{t=1}^{T} p(s|\boldsymbol{d}_t)\,|\boldsymbol{x}_t - \boldsymbol{A}_s\boldsymbol{e}_t|^2 - \lambda R_s \tag{3.40}$$

$$R_s = \frac{\mathbf{1}^\top {\boldsymbol{A}_s'}^\top \mathrm{diag}(\boldsymbol{E}\boldsymbol{P}_s\boldsymbol{E}^\top)\boldsymbol{A}_s'\mathbf{1}}{M} \tag{3.41}$$

$$\boldsymbol{E} = [\boldsymbol{e}_1, \cdots, \boldsymbol{e}_T] \tag{3.42}$$

$$\boldsymbol{P}_s = \begin{bmatrix} p(s|\boldsymbol{d}_1) & 0 & \cdots & 0 \\ 0 & p(s|\boldsymbol{d}_2) & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & p(s|\boldsymbol{d}_T) \end{bmatrix} \tag{3.43}$$

ここで，$\boldsymbol{A}_s'$ は $\boldsymbol{A}_s$ のバイアス項に相当する部分を取り除いた行列である．解析的には以下のように解ける．

$$\boldsymbol{A}_s = \boldsymbol{X}\boldsymbol{P}_s\boldsymbol{E}^{\mathrm{T}}(\boldsymbol{E}\boldsymbol{P}_s\boldsymbol{E}^{\mathrm{T}} + \lambda\boldsymbol{I}'\mathrm{diag}(\boldsymbol{E}\boldsymbol{P}_s\boldsymbol{E}^\top))^{-1} \tag{3.44}$$

ここで，$\boldsymbol{I}'$ は $I'_{1,1} = 0$，それ以外の対角要素は 1，それ以外はすべて 0 となるような正方行列である．

3.3　NMF による事例ベースの音声強調

事例ベースの音声強調として，Nonnegative Matrix Factorizastion (NMF) [5] を用いた手法を紹介する．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（$\boldsymbol{A} \to \boldsymbol{B}\boldsymbol{C}$）

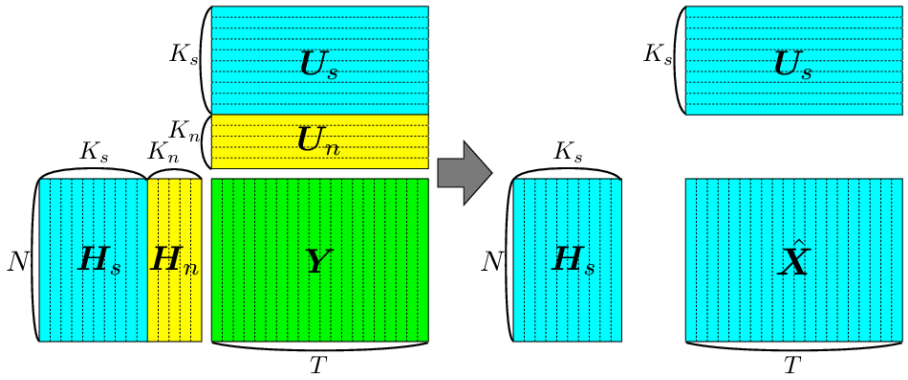

図 3.2: NMF による音声強調

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．

ここで，音声のパワースペクトルを考える．パワースペクトルは非負であり，加算モデルで表現できる．また，加算性雑音の乗った音声を考えた場合，雑音のパワースペクトルも音声のパワースペクトルに対して加算モデルとして近似できる．したがって，ある時間 t におけるスペクトルスライス $\boldsymbol{y}_t$ を考えたとき，K 個の構成要素（基底）$\boldsymbol{H}$ とその重みベクトル U_t による重み付け和で表現できる．Ω を $\boldsymbol{y}_t$ の次元数とする．

$$
\begin{aligned}
\boldsymbol{y}_t &\approx \boldsymbol{H}\boldsymbol{U}_t \\
&= \begin{bmatrix} H_{1,1} & H_{1,2} & \cdots & H_{1,K} \\ H_{2,1} & H_{2,2} & & \\ \vdots & & \ddots & \vdots \\ H_{\Omega,1} & H_{\Omega,2} & \cdots & H_{\Omega,K} \end{bmatrix} \begin{bmatrix} U_{1,t} \\ U_{2,t} \\ \vdots \\ U_{K,t} \end{bmatrix}
\end{aligned} \tag{3.45}
$$

ここで，NMF をパワースペクトル時系列の行列に対して適用することで，重みは行列として得られる．したがって，雑音重畳音声を音声と雑音の基底とその重み行列の積に分解することができる．それぞれの行列に非負の制約があるため，音声は少ない基底によってスパースに表現される．ここで，音声基底とその重み行列のみを抽出することで，クリーン音声を構成することができる．

3.3.1　NMF による音声強調 [20]

まず，NMF を用いたベーシックな手法を紹介する [20]．NMF による手法の概要を図 3.2 に示す．

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．

$$\boldsymbol{Y} = \boldsymbol{HU} = [\boldsymbol{H}_s\ \boldsymbol{H}_n] \begin{bmatrix} \boldsymbol{U}_s \\ \boldsymbol{U}_n \end{bmatrix} \tag{3.46}$$

$$\hat{\boldsymbol{X}} = \boldsymbol{H}_s\boldsymbol{U}_s \tag{3.47}$$

パラメータの更新は式 (3.48) のコスト関数 $D(\boldsymbol{Y}||\boldsymbol{HU})$ を最小化するように行う．

$$D(\boldsymbol{Y}||\boldsymbol{HU}) = d(\boldsymbol{Y}, \boldsymbol{HU}) + ||\boldsymbol{\lambda}. * \boldsymbol{U}||_p \tag{3.48}$$

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 (3.49) 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．

$$d(\boldsymbol{Y}, \boldsymbol{HU}) = \sum_{\omega,t}^{\Omega,T} \Big(Y_{\omega,t} \log \frac{Y_{\omega,t}}{\sum_k^K H_{\omega,k}U_{k,t}} - Y_{\omega,t} + \sum_k^K H_{\omega,k}U_{k,t} \Big) \tag{3.49}$$

コスト関数 (3.48) を最小化するアクティベーションの更新式として，式 (3.50) が得られる．

$$\boldsymbol{U} \longleftarrow \boldsymbol{U}. * (\boldsymbol{H}^\top(\boldsymbol{Y}./(\boldsymbol{HU})))./(\boldsymbol{H}^\top\mathbf{1}_Y + \boldsymbol{\lambda}) \tag{3.50}$$

$\boldsymbol{U}$ 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，$\mathbf{1}_Y \in \mathbb{R}^{\Omega\times T}$ はすべての要素が 1 の行列である．

3.3.2　Noise-transductive NMF による音声強調 [21]

通常の NMF による音声強調 [20] 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 [21]．

[21] 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 (3.48) 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 (3.51) としている．

$$D(\boldsymbol{Y}||\boldsymbol{HU}) = d(\boldsymbol{Y}, \boldsymbol{HU}) + ||\boldsymbol{\lambda}. * \boldsymbol{U}||_p + \boldsymbol{\eta}. * d(\boldsymbol{N}, \boldsymbol{H}_n) \tag{3.51}$$

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 (3.51) 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 (3.52) ， (3.53) のようになる．

$$\boldsymbol{U} \longleftarrow \boldsymbol{U}.*(\boldsymbol{H}^{\top}(\boldsymbol{Y}./(\boldsymbol{H}\boldsymbol{U})))./(\boldsymbol{H}^{\top}\mathbf{1}_Y+\boldsymbol{\lambda}) \tag{3.52}$$

$$\boldsymbol{H}_n \longleftarrow (\boldsymbol{H}_n.*(\boldsymbol{Y}./(\boldsymbol{H}\boldsymbol{U})\boldsymbol{U}_n^{\top})+\boldsymbol{\eta}.*\boldsymbol{N})./(\mathbf{1}_Y\boldsymbol{U}_n^{\top}+\boldsymbol{\eta}) \tag{3.53}$$

3.4　モデルベースと事例ベースの組み合せ

GMM を用いたモデルベースの手法では，特徴量が GMM によって拡張された高次の空間でマッピングされており，音声強調はその高次の空間から強調後の特徴量空間への射影と見なす事ができる．したがって，GMMM によらずとも，雑音の乗った音声の特徴量をある空間にマッピングすることができれば，その空間から強調後の特徴量空間への射影によって音声強調が実現できる．最近では，Dictionary Learning (DL) による手法が検討されている [22]．この手法では，NMF による事例ベースの手法を取り入れ，雑音の乗った音声の特徴量 $\boldsymbol{Y}$ を Dictionary 行列とその重み行列に分解して表している．ここで，Dictionary 行列の空間における重みの各要素を GMM における各分布の重み $p(m|\boldsymbol{y})$ に対応させる事で，区分的線形変換を行う事ができる．つまり，GMM の混合分割を Dictionary 行列の要素ごとの分解として考える事ができる．事例ベースとの対応づけのため，ここでは線形変換はバイアス項 $\boldsymbol{b}_m$ のみを考える．[2]

3.4.1　Dictionary Learning による手法 [22]

Dictionary Learning による手法の具体的な実現方法を紹介する．まず，Ω 次元，フレーム数 T の $\boldsymbol{X}$ の要素 $\boldsymbol{x}_t$ に対する区分的線形変換は，$\boldsymbol{X}$ 全体に対しては以下のようになる．

$$\begin{aligned}\boldsymbol{x}_t &= \sum_m p(m|\boldsymbol{y}_t)(\boldsymbol{y}_t+\boldsymbol{b}_m)\\ &= \boldsymbol{y}_t+\sum_m p(m|\boldsymbol{y}_t)\boldsymbol{b}_m \qquad (3.54)\\ \boldsymbol{X} &= \boldsymbol{Y}+\boldsymbol{B}\boldsymbol{\Gamma} \qquad (3.55)\\ &= \begin{bmatrix}\boldsymbol{I}_\Omega & \boldsymbol{B}\end{bmatrix}\begin{bmatrix}\boldsymbol{Y}\\ \boldsymbol{\Gamma}\end{bmatrix} \qquad (3.56)\end{aligned}$$

$\boldsymbol{I}_\Omega$ は $\Omega\times\Omega$ の単位行列である．Γ は $M\times T$ 次元であり，各要素は $\{\{p(m|\boldsymbol{y}_t)\}_{m=1}^{M}\}_{t=1}^{T}$ となっている．$\boldsymbol{B}$ は $\Omega\times M$ 次元であり，$\boldsymbol{B}=[\boldsymbol{b}_{m=1},\ldots,\boldsymbol{b}_{m=M}]$ となっている．ここで，バイアス項 $\boldsymbol{B}$ は最小二乗誤差基準で以下のように推定でき，解析的に式 (3.58) として求める事ができる．

$$\underset{\boldsymbol{B}}{\operatorname{argmin}}\,||\boldsymbol{X}-\boldsymbol{Y}-\boldsymbol{B}\boldsymbol{\Gamma}||_2^2 \tag{3.57}$$

$$\hat{\boldsymbol{B}}=(\boldsymbol{X}-\boldsymbol{Y})\boldsymbol{\Gamma}^{\mathrm{T}}(\boldsymbol{\Gamma}\boldsymbol{\Gamma}^{\mathrm{T}})^{-1} \tag{3.58}$$

この $\boldsymbol{B}$ を，Γ ではなく Dictionary Learning によって推定していく．

[2] これらの手法のゲインはバイアス項によるところが大きく，線形変換 $\boldsymbol{A}_m$ をかけることが必ずしも大きく性能向上に貢献するとは限らないため，実用上はバイアス項 $\boldsymbol{b}_m$ を考慮するだけでも十分である．

i)　Dictionary Learning

Dictionary Learning においては，以下のように $\boldsymbol{y}_t \approx \boldsymbol{D}\boldsymbol{w}_t$ の分解を行う．

$$\underset{\boldsymbol{D},\boldsymbol{w}_t}{\operatorname{argmin}} ||\boldsymbol{y}_t - \boldsymbol{D}\boldsymbol{w}_t||_2^2 + \Lambda(\boldsymbol{w}_t) \ \forall t \tag{3.59}$$

$\boldsymbol{D}$ は $D \times M$ の Dictionary 行列，$\boldsymbol{w}_t$ はフレーム t における D の各要素の重みである．事例ベースの手法と同様に，w_t はスパースとなるように学習する必要があるため，$\Lambda(\boldsymbol{w}_t)$ の正則化を行っている．

ここで，$\boldsymbol{D}$ と $\boldsymbol{W} = [\boldsymbol{w}_{t=1}, \ldots, \boldsymbol{w}_{t=T}]$ から式 (3.57) の Γ に対応するような Ψ を用意することで，式 (3.58) のように $\hat{\boldsymbol{B}}$ が得られる．

$$\begin{aligned} &\underset{\boldsymbol{B}}{\operatorname{argmin}} ||\boldsymbol{X} - \boldsymbol{Y} - \boldsymbol{B}\boldsymbol{\Psi}||_2^2 \\ &\boldsymbol{\Psi} = [\boldsymbol{\psi}_{t=1}, \ldots, \boldsymbol{\psi}_{t=T}] \end{aligned} \tag{3.60}$$

まず，スパースな $\boldsymbol{w}_t$ を得るための手法として，Compressive sensing を利用する [23]．例えば，式 (3.61) のような Orthogonal Matching Pursuit (OMP) [24] や，L1 正則化を用いた Lasso と呼ばれる式 (3.62) のような方法がある．

$$\underset{\boldsymbol{w}_t}{\operatorname{argmin}} ||\boldsymbol{w}_t||_0 \ \text{s.t.} \ ||\boldsymbol{y}_t - \boldsymbol{D}\boldsymbol{w}_t||_2^2 \leq \varepsilon \tag{3.61}$$

$$\underset{\boldsymbol{w}_t}{\operatorname{argmin}} ||\boldsymbol{y}_t - \boldsymbol{D}\boldsymbol{w}_t||_2^2 + \lambda||\boldsymbol{w}_t||_1 \tag{3.62}$$

上のような手法は多くの場合に適用可能な原理的な手法であるが，その他にもスパースな $\boldsymbol{w}_t$ を得るための手法がある [25, 26]．

$\boldsymbol{w}_t$ が学習できたとき，式 (3.57) の Γ に対応する Ψ として，以下のように $\boldsymbol{w}_t$ から事後確率 $p(m|\boldsymbol{y}_t)$ を求めて用いる．

$$\begin{aligned} p(m|\boldsymbol{y}_t) &= \frac{p(\boldsymbol{y}_t|m)}{\sum_{m=1}^{M} p(\boldsymbol{y}_t|m)} \\ &\propto \exp\Big(-\frac{||\boldsymbol{y}_t - w_{m,t}\boldsymbol{d}_m||_2^2}{2\sigma^2}\Big) \end{aligned} \tag{3.63}$$

ここで，$0 \leq p(m|\boldsymbol{y}_t) \leq 1$ であるため，σ による正規化を行っている．しかし，Dictionary Learning においては重みのレンジも重要であるため，重みそのものも Ψ として用い，2 通りを検討している．

- Weight: $\psi \triangleq w_{m,t}$
- Posterior: $\psi \triangleq p(m|\boldsymbol{y}_t)$

次に，得られた $\boldsymbol{W}$ から Dictionary Learning を行う．その一例として，Method of Optimal Direction (OMD) があり [27]，以下のように $\hat{\boldsymbol{D}}$ が推定できる．

$$\hat{\boldsymbol{D}} = f_{nc}(\boldsymbol{Y}\boldsymbol{W}^{\mathsf{T}}(\boldsymbol{W}\boldsymbol{W}^{\mathsf{T}})^{-1}) \tag{3.64}$$

ここで，$f_{nc}(\cdot)$ は $\hat{\boldsymbol{D}}$ を構成する縦ベクトル $\hat{\boldsymbol{d}}_m$ に対し，$\hat{\boldsymbol{d}}_m \rightarrow \hat{\boldsymbol{d}}_m/|\hat{\boldsymbol{d}}_m|$ と変換するものである．$\hat{\boldsymbol{D}}$ を求める手法はこの他にもいくつか提案されている [25, 28]．この $\hat{\boldsymbol{D}}$ と $\boldsymbol{W}$ の推定を繰り返して得られた $\boldsymbol{W}$ から $\boldsymbol{\Psi}$ を求める．

ii)　変換行列の推定

式 (3.55) は Γ を Ψ で置き換えることで，以下のようになる．

$$\boldsymbol{X} = \boldsymbol{Y} + \boldsymbol{B}\boldsymbol{\Psi} \tag{3.65}$$

したがって，式 (3.58) と同様にして，以下のように $\hat{\boldsymbol{B}}$ を求める事ができる．

$$\hat{\boldsymbol{B}} = (\boldsymbol{X} - \boldsymbol{Y})\boldsymbol{\Psi}^{\mathsf{T}}(\boldsymbol{\Psi}\boldsymbol{\Psi}^{\mathsf{T}})^{-1} \tag{3.66}$$

ここで，この変換行列の推定は繰り返し行う事ができる．式 (3.67) に基づくと，以下のように，$\boldsymbol{X}^n$ の推定値からさらに推定値 $\boldsymbol{X}^{(n+1)}$ を求める事ができる．n は繰り返し回数であり，$\boldsymbol{X}^1 \triangleq \boldsymbol{Y}$ とする．

$$\boldsymbol{X}^{(n+1)} = \boldsymbol{X}^{(n)} + \boldsymbol{B}^{(n)}\boldsymbol{\Psi}^{(n)} \tag{3.67}$$

$$\underset{\boldsymbol{D}^{(n)},\boldsymbol{w}_t^{(n)}}{\operatorname{argmin}} \ ||\boldsymbol{x}_t^{(n)} - \boldsymbol{D}^{(n)}\boldsymbol{w}_t^{(n)}||_2^2 + \Lambda(\boldsymbol{w}_t^{(n)}) \ \forall t \tag{3.68}$$

$$\hat{\boldsymbol{B}}^{(n)} = (\boldsymbol{X} - \boldsymbol{X}^{(n)})(\boldsymbol{\Psi}^{(n)})^{\mathsf{T}}(\boldsymbol{\Psi}^{(n)}(\boldsymbol{\Psi}^{(n)})^{\mathsf{T}})^{-1} \tag{3.69}$$

この繰り返しの 1 ステップにおいて，$\boldsymbol{X}^{(n)}$ の時系列全体を利用して $\hat{\boldsymbol{B}}^{(n)}$ を推定している．したがって，次の $\boldsymbol{X}^{(n+1)}$ は前後の時間関係による長時間特徴も考慮されていると見る事ができる．GMM を用いた手法では長時間特徴を考慮するため前後フレームを連結して特徴量として用いる場合があるが，特徴量の次元が増大し，GMM の混合数が多いと大きな計算コストがかかってしまう．しかし，この繰り返しによって計算コストを抑えつつ長時間特徴を取り入れることができる．

第4章

MFCC-based GMMの事後確率を用いたNMFとその雑音環境下音声認識への応用

4.1　はじめに

本章では，前章で紹介した NMF による音声強調に注目し，より高精度な NMF を検討する．NMF における改善点として，入力音声に応じた基底の選択が挙げられる．そこで，MFCC 領域における GMM の事後確率を用いて，適応的に基底を選択する NMF を提案する．

4.2　NMF における基底の選択

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して利用することができれば，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．

NMF は音声強調だけでなく声質変換にも用いられており，ここでも基底の選択が検討されている [8]．[8] においては，まず音素のカテゴリに基づいたクラスで基底の分類を行っている．さらに，アクティベーションが最大となるクラスの基底のみを選択して用いることで，変換精度の向上が確認されている．このようなアクティベーションを基準とした選択の場合，あくまでもスペクトル領域でのコスト関数に基づく基準で選択をしていることになる．

音声強調においても，基底の選択方法の一つとして音素情報を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．したがって，音素に依らない基底の選択方法を検討する必要がある．また，スペクトル領域でのコスト関数とは異なる基準を用いることでより精度のよい選択が可能になると期待できる．そこで，MFCC 領域における GMM の事後確率に基づく基底選択を用いた NMF を提案する．

4.3　NMF と GMM の組み合せ

入力音声に対する適応的な基底の選択は，どの基底にマッチしているかの識別という入力音声の状態識別と考えることができる．そこで，3.2 節で用いられている GMM による状態識別（GMM 分類）を組み合わせて基底の選択を行う．

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おり，音声強調にも用いられている [12, 13, 16] そこで，MFCC 領域での GMM 分類を導入する．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が期待できる．

4.4　GMM 分類による事後確率を用いた基底の選択

本節では，MFCC 領域での GMM 分類によって得られた事後確率を用いた NMF を提案する．GMM はクリーン音声の特徴量 $\boldsymbol{x}_t (t = 1, \ldots, T)$ から構成された以下の GMM（クリーン GMM）を用い，混合数を M とする．

$$p(\boldsymbol{x}_t) = \sum_{m}^{M} p(m) \pi_m^{\boldsymbol{x}} \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{\mu}_m^{\boldsymbol{x}}, \boldsymbol{\Sigma}_m^{\boldsymbol{x}}) \tag{4.1}$$

$\pi_m^{\boldsymbol{x}}$，$\boldsymbol{\mu}_m^{\boldsymbol{x}}$，$\boldsymbol{\Sigma}_m^{\boldsymbol{x}}$ はインデックスが $m = 1, \ldots, M$ の分布の重み，平均値，分散である．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

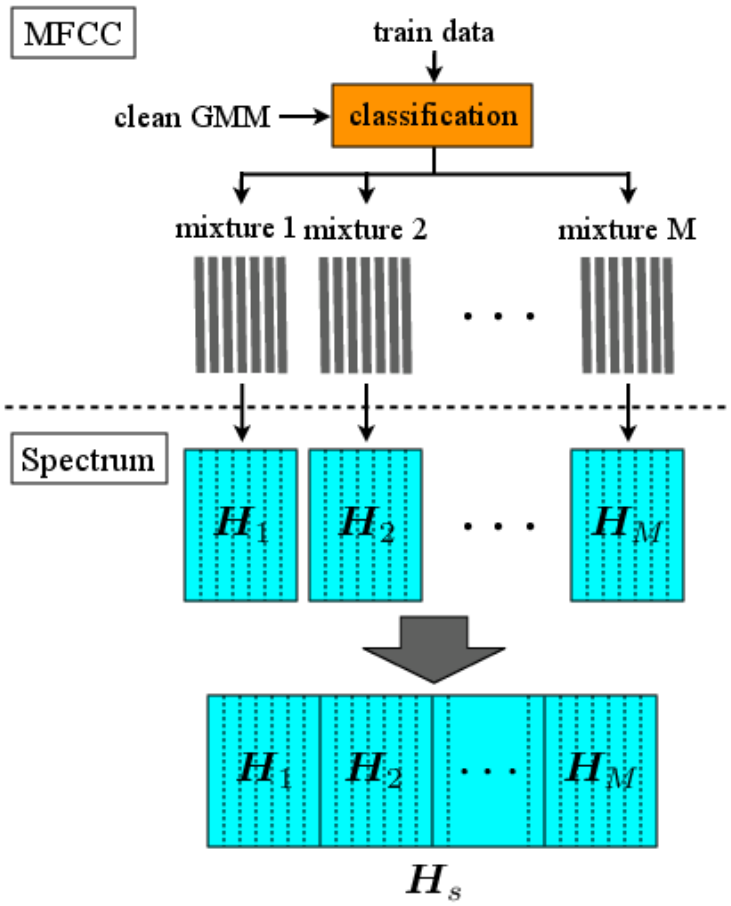

図 4.1: クラスごとの基底ベクトルの学習

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．GMM 分類による事後確率の利用方法として，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と，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の 2 つを提案する．

4.4.1　ハードな分類による手法

i)　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 GMM 分類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 4.1 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分類する．GMM の事後確率が最大となる分布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．

ii)　評価データに対する音声基底の選択

認識時の入力音声に対しても GMM 分類を行い，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分布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，事後確率を得るためにはクリーン音声の特徴量を用いることが望ましい．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 $\hat{\boldsymbol{X}}_{hard}$ を推定する．次に，$\hat{\boldsymbol{X}}_{hard}$ に対してクリーン GMM で分類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 $\hat{\hat{\boldsymbol{X}}}_{hard}$ を推定する．

iii)　REDIAL を用いたクラス推定

前節ではクラスの推定に一度推定したクリーン音声の特徴量を利用したが，あくまでも推定結果であるためクリーン GMM とのミスマッチが起こりうる．そこで，雑音重畳音声からのクリーン音声のクラス推定手法の一つとして，3.2.4 節で紹介した REDIAL を用いたクラス推定を提案する．

REDIAL と同様に，t 番目のフレームの特徴量と前後数フレームを結合した特徴量 $\boldsymbol{d}_t$ を LDA によって次元圧縮した特徴量空間 $\boldsymbol{v}_t = \boldsymbol{D}\boldsymbol{d}_t$ で GMM 分類を行うことで事後確率を求め，クラスを決定する．クリーン音声のクラスを推定するためには，$\boldsymbol{v}_t$ の GMM の各分布がクリーン GMM の分布と対応している必要がある．そこで，式 (4.1) のクリーン GMM を LDA の変換行列 $\boldsymbol{L}$ によって変換した以下の GMM (LDA-GMM) を用いる．

$$p(\boldsymbol{v}_t) = \sum_{m}^{M} p(m)\pi_m^{\boldsymbol{v}}\mathcal{N}(\boldsymbol{v}_t;\boldsymbol{\mu}_m^{\boldsymbol{v}},\boldsymbol{\Sigma}_m^{\boldsymbol{v}}) \tag{4.2}$$

$$\begin{aligned} \pi_m^{\boldsymbol{v}} &= \pi_m^{\boldsymbol{x}} \\ \boldsymbol{\mu}_m^{\boldsymbol{v}} &= \boldsymbol{L}\boldsymbol{\mu}_m^{\boldsymbol{d}} \\ \boldsymbol{\Sigma}_m^{\boldsymbol{v}} &= \boldsymbol{L}\boldsymbol{\Sigma}_m^{\boldsymbol{d}}\boldsymbol{L}^{Top} \end{aligned} \tag{4.3}$$

$\pi_m^{\boldsymbol{v}}$，$\boldsymbol{\mu}_m^{\boldsymbol{v}}$，$\boldsymbol{\Sigma}_m^{\boldsymbol{v}}$ はインデックスが $m = 1,\ldots,M$ の分布の重み，平均値，分散である．また，$\boldsymbol{\mu}_m^{\boldsymbol{d}}$，$\boldsymbol{\Sigma}_m^{\boldsymbol{d}}$ はそれぞれ，$\boldsymbol{d}_t$ で結合したフレーム数分だけ $\boldsymbol{\mu}_m^{\boldsymbol{x}}$，$\boldsymbol{\Sigma}_m^{\boldsymbol{x}}$ を結合したものである．LDA-GMM を用いた分類による事後確率 $p(m|\boldsymbol{v}_t)$ が最大となる m をクラスとすることで，雑音重畳音声からクリーン音声のクラスが推定できる．

4.4.2　事後確率をソフトに利用した手法

i)　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 $\boldsymbol{Y}$ と NMF により再構成された特徴量 $\boldsymbol{H}\boldsymbol{U}$ の距離として式 (4.4) を定義する．クリーン GMM から得られたクラス $m = 1,\ldots,M$ の事後確率を $\boldsymbol{\gamma}_m \in \mathbb{R}^{1\times T}$，その各要素を $\gamma_{m,t}$ とする．また，クラス m の音声基底 $\boldsymbol{H}_m$，雑音基底 $\boldsymbol{H}_n$ それぞれの基底数を 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．

$$d_{soft}(\boldsymbol{Y}, \boldsymbol{HU}) = \sum_{m}^{M} \sum_{\omega,t}^{\Omega,T} \gamma_{m,t} \Big(Y_{\omega,t} \log \frac{Y_{\omega,t}}{HU_{\omega,t,m}} - Y_{\omega,t} + HU_{\omega,t,m} \Big) \tag{4.4}$$

ただし，

$$HU_{\omega,t,m} = \sum_{k_m}^{K_m} H_{\omega,k_m} U_{k_m,t} + \sum_{k_n}^{K_n} H_{\omega,k_n} U_{k_n,t} \tag{4.5}$$

とする．

この定義において，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 (3.48)， (3.51) における第一項を式 (4.4) 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 4.4.1 節のハードな分類に対応している．

ii)　音声基底の学習

クリーン音声 $\boldsymbol{X}$ 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 $\boldsymbol{H}_n$ は除くため，基底は音声基底 $\boldsymbol{H}_m (m = 1, \ldots, M)$ のみとなり，全体の基底数 $K = \sum_m^M K_m$ 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 $\boldsymbol{U}_m (m = 1, \ldots, M)$ のみとなる．したがって，コスト関数は以下のようになる．

$$D(\boldsymbol{X}||\boldsymbol{HU}) = \sum_{m}^{M} \sum_{\omega,t}^{\Omega,T} \gamma_{m,t} \Big(X_{\omega,t} \log \frac{X_{\omega,t}}{\sum_{k_m}^{K_m} H_{\omega,k_m} U_{k_m,t}} - X_{\omega,t} + \sum_{k_m}^{K_m} H_{\omega,k_m} U_{k_m,t} \Big) + \sum_{k,t}^{K,T} \lambda_k U_{k,t} \tag{4.6}$$

式 (4.6) を最小化する音声基底の更新式は以下のようになる．

$$\boldsymbol{H}_m \longleftarrow \boldsymbol{H}_m .* (((\boldsymbol{1}_{\boldsymbol{Y}} .* \boldsymbol{X})./(\boldsymbol{H}_m\boldsymbol{U}_m))\boldsymbol{U}_m^{\top})./\boldsymbol{1}_{\Omega}\boldsymbol{\gamma}_m\boldsymbol{U}_m \tag{4.7}$$

$\boldsymbol{1}_{\Omega} \in \mathbb{R}^{\Omega \times 1}$ はすべての要素が 1 の行列である．式 (4.7) の導出は後述する．

iii)　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 (4.8) のようになる．

$$D(\boldsymbol{Y}||\boldsymbol{HU}) = d_{soft}(\boldsymbol{Y}, \boldsymbol{HU}) + ||\boldsymbol{\lambda} .* \boldsymbol{U}||_p + \boldsymbol{\eta} .* d(\boldsymbol{N}, \boldsymbol{H}_n) \tag{4.8}$$

コスト関数 (4.8) 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 (4.9)， (4.10) が得られる．

$$\boldsymbol{U}_m \longleftarrow \boldsymbol{U}_m .* (\boldsymbol{1}_{K_m}\boldsymbol{\gamma}_m .* (\boldsymbol{H}_m{}^{\top}(\boldsymbol{Y}./(\boldsymbol{HU})_{mn})))./(\boldsymbol{\gamma}_m .* \boldsymbol{H}_m{}^{\top}\boldsymbol{1}_Y + \boldsymbol{\lambda}_m) \tag{4.9}$$

$$\boldsymbol{U}_n \longleftarrow \boldsymbol{U}_n .* (\sum_{m}^{M} \boldsymbol{1}_{K_n}\boldsymbol{\gamma}_m .* (\boldsymbol{H}_m^{\top}(\boldsymbol{Y}./(\boldsymbol{HU})_{mn})))./(\boldsymbol{H}_n^{\top}\boldsymbol{1}_Y + \boldsymbol{\lambda}_n) \tag{4.10}$$

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，$\mathbf{1}_{K_m} \in \mathbb{R}^{K_m \times 1}$，$\mathbf{1}_{K_n} \in \mathbb{R}^{K_n \times 1}$ はすべての要素が 1 の行列である．また，雑音基底の更新式としては式 (4.11) が得られる．

$$\boldsymbol{H}_n \longleftarrow (\sum_m^M \boldsymbol{H}_n . * (((\mathbf{1}_Y \boldsymbol{\gamma}_m . * \boldsymbol{Y}) ./ (\boldsymbol{HU})_{mn}) \boldsymbol{U}_n^\top) + \boldsymbol{\eta} . * \boldsymbol{N}) ./ (\sum_m^M \mathbf{1}_\Omega \boldsymbol{\gamma}_m \boldsymbol{U}_n^\top + \boldsymbol{\eta}) \quad (4.11)$$

式 (4.9)，(4.10)，(4.11) の導出は後述する．

4.4.1 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 (3.51) に基づいて $\boldsymbol{H}_s$ 全体を用いた NMF でクリーン音声の特徴量 $\hat{\boldsymbol{X}}_{soft}$ を推定する．次に，$\hat{\boldsymbol{X}}_{soft}$ から得られた事後確率を用いてコスト関数 (4.8) 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 $\hat{\hat{\boldsymbol{X}}}_{soft}$ を得る．

iv)　更新式の導出

まず，更新式 (4.9)，(4.10)，(4.11) を導出する．コスト関数を最小化するパラメータを求めればよいので，通常の NMF[20]，Noise-transductive NMF[21] と同様にコスト関数の偏微分により導出できる．コスト関数 (4.8) における距離関数 $d_{soft}(\boldsymbol{Y}, \boldsymbol{HU})$ を以下のように展開する．

$$\begin{aligned}
& d_{soft}(\boldsymbol{Y}, \boldsymbol{HU}) \\
= & \sum_m^M \sum_{\omega,t}^{\Omega,T} \gamma_{m,t} \Big(Y_{\omega,t} \log \frac{Y_{\omega,t}}{HU_{\omega,t,m}} - Y_{\omega,t} + HU_{\omega,t,m} \Big) \\
= & \sum_m^M \sum_{\omega,t}^{\Omega,T} \gamma_{m,t} \Big\{ Y_{\omega,t} \log Y_{\omega,t} \underline{- Y_{\omega,t} \log \Big(\sum_{k_m}^{K_m} H_{\omega,k_m} U_{k_m,t} + \sum_{k_n}^{K_n} H_{\omega,k_n} U_{k_n,t} \Big)} - Y_{\omega,t} + HU_{\omega,t,m} \Big\} \\
\leq & \sum_m^M \sum_{\omega,t}^{\Omega,T} \gamma_{m,t} \Big\{ Y_{\omega,t} \log Y_{\omega,t} \\
& \underline{- Y_{\omega,t} \Big(\sum_{k_m}^{K_m} \delta_{k_m,K_n,\omega,t} \log \frac{H_{\omega,k_m} U_{k_m,t}}{\delta_{k_m,K_n,\omega,t}} + \sum_{k_n}^{K_n} \delta_{K_m,k_n,\omega,t} \log \frac{H_{\omega,k_n} U_{k_n,t}}{\delta_{K_m,k_n,\omega,t}} \Big)} \\
& - Y_{\omega,t} + HU_{\omega,t,m} \Big\} \\
= & Q_{soft}(\boldsymbol{Y}, \boldsymbol{HU}, \boldsymbol{\delta})
\end{aligned} \quad (4.12)$$

$d_{soft}(\boldsymbol{Y}, \boldsymbol{HU})$ の偏微分を考えたとき，下線部は微分が困難である．そこで，下線部に対して Jensen の不等式を導入することで上限をとっている．等号成立条件は

$$\sum_{k_m}^{K_m} \delta_{k_m,K_n,\omega,t} + \sum_{k_n}^{K_n} \delta_{K_m,k_n,\omega,t} = 1 \quad (4.13)$$

$$\frac{H_{\omega,k_m} U_{k_m,t}}{\delta_{k_m,K_n,\omega,t}} = \frac{H_{\omega,k_n} U_{k_n,t}}{\delta_{K_m,k_n,\omega,t}} = const \quad (4.14)$$

となる．$\delta_{k_m,K_n,\omega,t}$，$\delta_{K_m,k_n,\omega,t}$ は $\boldsymbol{\delta}$ の各要素である．したがって，等号成立時の $\delta_{k_m,K_n,\omega,t}$，$\delta_{K_m,k_n,\omega,t}$ を $\hat{\delta}_{k_m,K_n,\omega,t}$，$\hat{\delta}_{K_m,k_n,\omega,t}$ とすると式 (4.15)，(4.16) となる．

$$\hat{\delta}_{k_m,K_n,\omega,t} = \frac{H_{\omega,k_m} U_{k_m,t}}{HU_{\omega,t,m}} \quad (4.15)$$

$$\hat{\delta}_{K_m,k_n,\omega,t} = \frac{H_{\omega,k_n} U_{k_n,t}}{HU_{\omega,t,m}} \quad (4.16)$$

したがって，以下の式 (4.17) を最小化すれば良い．

$$Q_{soft}(\boldsymbol{Y}, \boldsymbol{HU}, \hat{\boldsymbol{\delta}}) + ||\lambda. * \boldsymbol{U}||_p + \boldsymbol{\eta}. * d(\boldsymbol{N}, \boldsymbol{H}_n) \tag{4.17}$$
$$\text{ただし}\ \ \delta_{k_m,K_n,\omega,t} = \hat{\delta}_{k_m,K_n,\omega,t}\ ,\ \ \delta_{K_m,k_n,\omega,t} = \hat{\delta}_{K_m,k_n,\omega,t}$$

式 (4.17) 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．

$$U_{k_m,t} \longleftarrow U_{k_m,t} \frac{\gamma_{m,t} \sum_{\omega}^{\Omega} X_{\omega,t} \frac{H_{\omega,k_m}}{HU_{\omega,t,m}}}{\gamma_{m,t} \sum_{\omega}^{\Omega} H_{\omega,k_m} + \lambda_{k_m}} \tag{4.18}$$

$$U_{k_n,t} \longleftarrow U_{k_n,t} \frac{\sum_{m}^{M} \gamma_{m,t} \sum_{\omega}^{\Omega} X_{\omega,t} \frac{H_{\omega,k_n}}{HU_{\omega,t,m}}}{\sum_{\omega}^{\Omega} H_{\omega,k_n} + \lambda_{k_n}} \tag{4.19}$$

$$H_{\omega,k_n} \longleftarrow \frac{\sum_{t}^{T} \sum_{m}^{M} \gamma_{m,t} Y_{\omega,t} \frac{H_{\omega,k_n} U_{k_n,t}}{HU_{\omega,t,m}} + \eta N_{\omega,k_n}}{\sum_{t}^{T} \sum_{m}^{M} \gamma_{m,t} U_{k_n,t} + \eta} \tag{4.20}$$

これらを行列の形で表現することで式 (4.9) , (4.10) , (4.11) を得る．

次に，更新式 (4.7) を導出する．コスト関数 (4.6) についても同様に Jensen の不等式で上限をとる．

$$\begin{aligned}
& D(\boldsymbol{X}, \boldsymbol{HU}) \\
= & \sum_{m}^{M} \sum_{\omega,t}^{\Omega,T} \gamma_{m,t} \Big(X_{\omega,t} \log X_{\omega,t} - X_{\omega,t} \log \sum_{k_m}^{K_m} H_{\omega,k_m} U_{k_m,t} - X_{\omega,t} + \sum_{k_m}^{K_m} H_{\omega,k_m} U_{k_m,t} \Big) + \sum_{k,t}^{K,T} \lambda_k U_{k,t} \\
\leq & \sum_{m}^{M} \sum_{\omega,t}^{\Omega,T} \gamma_{m,t} \Big(X_{\omega,t} \log X_{\omega,t} - X_{\omega,t} \sum_{k_m}^{K_m} \delta_{k_m,\omega,t} \log \frac{H_{\omega,k_m} U_{k_m,t}}{\delta_{k_m,\omega,t}} - X_{\omega,t} + \sum_{k_m}^{K_m} H_{\omega,k_m} U_{k_m,t} \Big) + \sum_{k,t}^{K,T} \lambda_k U_{k,t} \\
= & R_{soft}(\boldsymbol{X}, \boldsymbol{HU}, \boldsymbol{\delta})
\end{aligned} \tag{4.21}$$

等号成立条件は

$$\sum_{k_m}^{K_m} \delta_{k_m,\omega,t} = 1 \tag{4.22}$$

$$\frac{H_{\omega,k_m} U_{k_m,t}}{\delta_{k_m,\omega,t}} = const \tag{4.23}$$

となる．これを満たす $\delta_{k_m,K_n.\omega,t}$ を $\hat{\delta}_{k_m.\omega,t}$ とすると式 (4.24) となる．

$$\hat{\delta}_{k_m,\omega,t} = \frac{H_{\omega,k_m} U_{k_m,t}}{\sum_{k_m}^{K_m} H_{\omega,k_m} U_{k_m,t}} \tag{4.24}$$

したがって，以下の式を最小化すれば良い．

$$R_{soft}(\boldsymbol{Y}, \boldsymbol{HU}, \boldsymbol{\delta}) \quad \text{ただし } \delta_{k_m,\omega,t} = \hat{\delta}_{k_m,\omega,t} \tag{4.25}$$

式 (4.25) の $\boldsymbol{H}_m$ についての導関数から，以下の更新式を得る．

$$H_{\omega,k_m} \longleftarrow H_{\omega,k_m} \frac{\sum_t^T \gamma_{m,t} X_{\omega,t} \frac{U_{k_m,t}}{\sum_{k_m}^{K_m} H_{\omega,k_m} U_{k_m,t}}}{\sum_t^T \gamma_{m,t} U_{k_m,t}} \tag{4.26}$$

これを行列の形で表現することで更新式 (4.7) を得る．

第5章

実験

表 5.1: 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
クリーン GMM 混合数	4
LDA-GMM 混合数	4
連結フレーム数（LDA）	前後各 4 フレーム
NMF	Noise-transductive NMF
連結フレーム数（NMF）	前後各 9 フレーム
音響モデル	クリーン条件

5.1　はじめに

本章では，前章で提案した MFCC 領域における GMM 分類による事後確率を用いた NMF の有効性を検証するため，雑音環境下音声認識実験を行う．

5.2　データベース

認識実験用の音声データベースとして，雑音環境下音声認識のベースラインとしてよく用いられる AURORA2 データベース [29] を用いた．これは雑音環境下における英語連続数字発話音声認識の評価を行うデータベースである．

学習用データとして全 8440 発話のクリーンな連続数字読み上げ音声 (clean データ) とそれぞれに加法性雑音が重畳された音声 (multi データ) が与えられる．雑音は Subway，Babble，Car，Exhibit の 4 種類であり，それぞれについて SNR5，10，15，20，∞[dB] の 5 パターンがある．

評価用データには A セット，B セット，C セットの 3 つがあるが，今回は A セット，B セットを用いた．音声データとしては全 4004 発話があるが，4 分割された 1001 発話が基本単位となる．A セットは雑音環境クローズドの評価セットである．学習用の multi データと同様の 4 種類の雑音があり，各雑音がそれぞれ 1001 発話に重畳されている．SNR はそれぞれの発話についてに-5，0，5，10，15，20，∞[dB] の 7 パターン用意されている．B セットは雑音環境オープンの評価セットである．学習用の multi データとは異なる Restaurant，Street，Airport，Station の 4 種類の雑音が重畳されており，SNR は A セットと同様の 7 パターンである．最終的な実験結果の比較は各セットで SNR0，5，10，15，20 を見ることが多いため，今回もそれに沿った．

5.3　実験条件

実験条件を表 5.1 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 3.3.2 項の Noise-transductive NMF を用いた．式 (3.51) 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 5.2: 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 学習	コスト関数 テスト	テスト時の更新回数	事後確率を得る特徴量と GMM
NMF (4000)	4000	全発話	式 (3.48)	式 (3.51)	200 回	-
NMF (4000, 1 of 4div)	4000	ランダム 4 分割 1 セット	式 (3.48)	式 (3.51)	200 回	-
NMF (16000)	16000	全発話	式 (3.48)	式 (3.51)	200 回	-
NMF (16000, all of 4div)	4000 × 4	ランダム 4 分割 4 セット	式 (3.48)	式 (3.51)	200 回	-
NMF+GMM (hard1)	4000 × 4	クラスごとに分割	式 (3.48)	式 (3.51)	200 回	-
NMF+GMM (hard2)	4000	クラスごとに分割	式 (3.48)	式 (3.51)	50 回	NMF+GMM (hard1) $\hat{\boldsymbol{X}}_{hard}$ クリーン GMM
NMF+LDA-GMM (hard)	4000	クラスごとに分割	式 (3.48)	式 (3.51)	50 回	$\boldsymbol{V} = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_T]$ LDA-GMM
NMF+GMM (soft1)	4000 × 4	全発話＋事後確率	式 (4.6)	式 (3.51)	200 回	-
NMF+GMM (soft2, oracle)	4000 × 4	全発話＋事後確率	式 (4.6)	式 (4.8)	200 回	NMF+GMM (soft1) $\hat{\boldsymbol{X}}_{soft}$ クリーン GMM
NMF+GMM (hard2, oracle)	4000	クラスごとに分割	式 (3.48)	式 (3.51)	50 回	クリーン音声 $\boldsymbol{X}$ クリーン GMM
NMF+GMM (soft2, oracle)	4000 × 4	全発話＋事後確率	式 (4.6)	式 (4.8)	200 回	クリーン音声 $\boldsymbol{X}$ クリーン GMM

NMF の各条件における音声基底数，音声基底の学習データ，基底の学習とテスト時の更新に用いたコスト関数，テスト時の更新回数，事後確率を得るために用いた特徴量と GMM を表 5.2 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．したがって，全クラスを合わせた音声基底 $\boldsymbol{H}_s^h$，$\boldsymbol{H}_s^s$ の基底数は 4000 × 4=16000 である．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 (3.48) を，テスト時のクリーン音声の推定では式 (3.51) 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 $\boldsymbol{H}_s$ を用いてクリーン音声の特徴量 $\hat{\boldsymbol{X}}_{hard}$ を推定した．この $\hat{\boldsymbol{X}}_{hard}$ 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，$\hat{\boldsymbol{X}}_{hard}$ 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 $\boldsymbol{H}_m$ を選択した場合を NMF+GMM (hard2) とする．また，LDA-GMM を用いて雑音重畳音声から直接事後確率を求め，ハードな分類によって音声基底を選択した場合を GMM+LDA-GMM (hard) 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 (4.6) に基づいて音声基底 $\boldsymbol{H}_s = [\boldsymbol{H}_1 \ \cdots \ \boldsymbol{H}_M]$ を学習した．この音声基底 $\boldsymbol{H}_s$ を用いて式 (3.51) に基づいてクリーン音声の特徴量 $\hat{\boldsymbol{X}}_{soft}$ を推定した．この $\hat{\boldsymbol{X}}_{soft}$ 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，$\hat{\boldsymbol{X}}_{soft}$ から得られた事後確率を用いて式 (4.8) 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 $\boldsymbol{X}$ 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

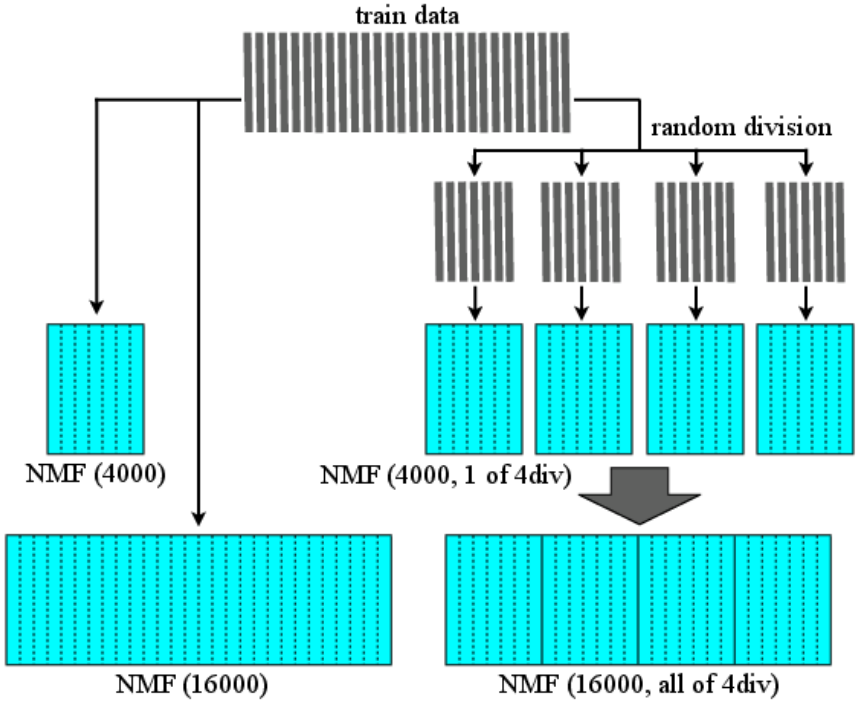

図 5.1: ベースラインにおける音声基底

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 5.1 にまとめる．

5.4　実験結果

各実験条件における set A，set B の認識率を表 5.3，5.4 に示す．

まず，NMF+GMM (hard1) ，NMF+GMM (soft1) 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，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．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 分類の事後確率を利用したことによる効果であることが確認できる．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．しかし，NMF+LDA-GMM (hard) で最も高い認識率が得られている．NMF+GMM (hard2) ，NMF+GMM (soft2) が 2 回の NMF を要するのに対し，NMF+LDA-GMM (hard) は NMF+GMM (hard1) ，NMF+GMM (soft1) と同様に 1 回の NMF でよいため，計算コストの面でも優位である．ハードな分類については，

表 5.3: 認識率 (set A) [%]

	NMF (4000)	NMF (4000, 1 of 4div)	NMF (16000)	NMF (16000, all of 4div)
SNR20	84.18	83.05	83.58	85.45
SNR15	80.76	79.74	80.60	82.53
SNR10	73.28	72.09	73.73	75.60
SNR5	59.54	58.17	60.16	61.79
SNR0	38.82	37.73	39.19	40.32
Average	67.31	66.15	67.45	69.14

	NMF+GMM (hard1)	NMF+GMM (hard2)	NMF+LDA-GMM (hard)	NMF+GMM (hard2, oracle)
SNR20	92.89	87.78	95.01	93.89
SNR15	89.33	84.56	91.51	92.58
SNR10	81.73	78.36	86.05	89.09
SNR5	67.83	65.68	74.31	80.72
SNR0	45.63	43.94	49.62	60.35
Average	75.48	72.06	79.30	83.32

	NMF+GMM (soft1)	NMF+GMM (soft2)	NMF+GMM (soft2, oracle)
SNR20	94.09	85.45	89.28
SNR15	90.60	82.11	86.22
SNR10	83.53	74.92	79.38
SNR5	69.49	60.05	64.35
SNR0	46.53	33.54	38.17
Average	76.85	67.22	71.48

NMF+GMM (hard2, oracle) で最も高い認識率が得られているため，クラス推定の精度が良ければハードな分類を利用した音声基底の選択が有効であることがわかる．実際の認識で利用する上では，NMF が 1 回かつ認識率が高い NMF+LDA-GMM (hard) を利用するのが望ましい．

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 (4.6) 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 (3.51) によって更新している．つまり，雑音の要素の更新を事後確率を用いて行ったことが NMF+GMM (soft1) より悪化した原因であると考えられる．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．特に，今回のコスト関数である式 (4.8) はハードな分類を参考に定義したものであるため，MFCC 領域での事後確率がスペクトル領域でどう振る舞うのかを考慮して検討する必要がある．

表 5.4: 認識率 (set B) [%]

	NMF (4000)	NMF (4000, 1 of 4div)	NMF (16000)	NMF (16000, all of 4div)
SNR20	81.17	79.31	79.61	81.59
SNR15	75.91	74.56	75.87	77.30
SNR10	67.67	66.74	68.43	69.63
SNR5	54.17	52.86	54.13	55.96
SNR0	35.02	33.63	34.58	36.26
Average	62.79	61.42	62.52	64.15

	NMF+GMM (hard1)	NMF+GMM (hard2)	NMF+LDA-GMM (hard)	NMF+GMM (hard2, oracle)
SNR20	90.57	82.78	95.67	92.53
SNR15	86.21	77.96	92.88	90.30
SNR10	77.76	70.95	87.87	86.50
SNR5	62.95	57.69	75.00	76.64
SNR0	41.78	37.39	48.75	57.32
Average	71.85	65.35	80.03	80.66

	NMF+GMM (soft1)	NMF+GMM (soft2)	NMF+GMM (soft2, oracle)
SNR20	92.22	73.32	82.67
SNR15	87.96	69.26	79.30
SNR10	79.53	61.97	71.89
SNR5	64.40	47.58	57.86
SNR0	43.03	27.19	36.23
Average	73.43	55.86	65.59

第6章

結論

6.1 本論文のまとめ

本論文では，雑音環境下音声認識におけるアプローチの中でも NMF による音声強調に注目し，その精度向上を目指した．従来の NMF では入力音声に依らず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た．ここで，入力音声に対して適応的な音声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が期待できる．そこで，MFCC-based GMM の事後確率を用いた NMF を提案し，雑音環境下音声認識における有効性を検証した．事後確率の利用方法として，ハードな分類とコスト関数に事後確率を導入したソフトな利用の 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，事後確率をソフトに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．

6.2 今後の展望

提案手法は，事後確率推定と NMF をそれぞれ独立したステップとして適用していた．NMF は繰り返し更新するアルゴリズムであるため，その更新に事後確率推定の枠組みを組み込むことが考えられる．そこで，NMF と事後確率の更新を同時に行えるような，NMF と GMM を統合したモデルで扱うことが最終的な目標となるだろう．また，NMF による音声強調は計算コストが大きい手法であったが，基底の選択によって利用する基底のサイズが小さくなることで，計算コストの低減が見込める．そこで，計算コストについても比較・検討することで，精度向上にとどまらない本手法の効果を見いだせると期待できる．

謝辞

まず本研究を進めるにあたり，2 年間指導教員として多大なご指導をして頂いた峯松信明教授，広瀬啓吉教授に深く感謝いたします．特に峯松信明教授には，常日頃からお忙しい中，多くのご助言をいただきました．また，日頃の研究活動を支えて下さった高橋登技術専門員，池上恵事務補佐員，折茂結実子事務補佐員にも感謝致します．

さらに，齋藤大輔助教、博士課程の柏木陽佑氏にも大変お世話になりました．深く感謝いたします．理論の構築，研究方針をはじめとした非常に多くの相談にのっていただき，未熟者の私を導いてくださいました．両者の助力なしには本研究を成し遂げることはできなかったでしょう．

また，研究に遊びにと非常に楽しい時間を過ごさせていただいた，柏木陽佑氏，橋本浩弥氏をはじめとした研究室の方々にも感謝いたします．特に，卒論からの 3 年間苦楽を共にした同期である笠原駿氏，橋本哲弥氏，水上智之氏，岡安貴大氏，中村新芽氏，槇佑馬氏には深く感謝しております．彼らなくして修士課程を無事に過ごすことはできなかったことでしょう．有意義で楽しい学生生活になったのも彼らのおかげです．そして最後に，これまで私を支えてくださったすべての方々に深く感謝いたします．本当にありがとうございました．

2015 年 2 月 5 日
藤垣 健太郎

参考文献

[1] 大浦 圭一郎, 山本 大介, 内匠 逸, 李 晃伸, 徳田 恵一, “キャンパスの公共空間におけるユーザ参加型双方向音声案内デジタルサイネージシステム”, 人工知能学会誌, Vol,28, No.1, pp.60–67, 2013.

[2] 河原達也, “議会の会議録作成のための音声認識－衆議院のシステムの概要－”, 情報処理学会研究報告, SLP-93-5, 2012.

[3] 中村 哲, “実音響環境に頑健な音声認識を目指して”, 電子情報通信学会技術研究報告 EA, 応用音響, Vol.102, No.33, pp.31–36, 2002.

[4] J. Droppo, A. Acero, “Environmental Robustness,” Springer Handbook of Speech Processing, 99.653–679, 2008.

[5] D. D. Lee, H.S. Seung, “Algorithms for non-negative matrix factorization,” Proc. Neural Information Processing Sistems, pp.556–562, 2001.

[6] 藤井 貴生, 相原 龍, 滝口 哲也, 有木 康雄, “話者適応を用いた NMF による雑音環境下の声質変換”, 日本音響学会秋季講演論文集, 2-Q-36, pp.345–348, 2014.

[7] 仲野 翔一, 山本 一公, 中川 聖一, “NMF と VQ 手法による音楽重畳音声の音声認識”, 電子情報通信学会技術研究報告, Vol.111, No.97, pp.23–28, 2011.

[8] 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, pp.71–76, 2013.

[9] 相原 龍, 滝口 哲也, 有木 康雄, “スパース辞書学習による構音障害者の話者性を維持した声質変換”, 電子情報通信学会技術報告, Vol.114, No.91, pp.39–44, 2014.

[10] 鹿野清宏, 伊藤克亘, 河原達也, 武田一哉, 山本幹雄, “IT Text 音声認識システム”, 情報処理学会編集, オーム社, 2001.

[11] C. M. ビショップ, “パターン認識と機械学習 上,” 株式会社シナノ, 2007.

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

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

[14] J. Du, Y. Hu, L.-R. Dai, R.-H. Wang, “HMM-based pseudo-clean speech synthesis for SPLICE algorithm,” Proc. ICASSP, pp4570–4573, 2010.

[15] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, T. Kitamura, “Speech parameter generation algorithms for HMM-based speech synthesis,” IEEE Transactrion on Audio, Speech and Language Processing, Vol.3, pp.1315–1318, 2000.

[16] M. Afify, X. Cui, Y. Gao, “Stereo-based stocastic mapping for robust speech recognition,” IEEE Transactrion on Audio, Speech and Language Processing, Vol.17, No.7, pp.1325–1334, 2009.

[17] J. Du, Q. Huo, “Synthesized stereo-based stochastic mapping with data selection for robust speech recognition,” ICASSP, pp.122–125, 2012.

[18] 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.

[19] 鈴木 雅之, “背景雑音と話者の違いに頑健な音声認識”, 東京大学大学院工学系研究科電気系工学専攻, 博士論文，2013.

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

[21] 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.

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

[23] D. L. Donoho, “Compressed sensing,” IEEE Transactions on Information theory, Vol.52, No.4, pp.1289–1306, 2006.

[24] Y. C. Pati, R. Rezaiifar, P. S. Krishnaprasad, “Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition,” Proc. ASILOMAR’93, pp.40–44, 1993.

[25] M. Aharon, M. Elad, A. Bruckstein, “K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation,” IEEE Transactions on Signal Processing, Vol.54, No.11, pp.4311–4322, 2006.

[26] M. Fornasier, H. Rauhut, “Compressive sensing,” Handbook of Mathmatical Methods in Imaging, Vol.1, pp.187–229, 2011.

[27] K. Engan, S. O. Aase, J. H. Husϕy, “Frame based signal compression using method of optimal directions,” the 1999 IEEE International Symposium on Circuits and Systems, Vol.4, pp.1–4, 1999.

[28] J. Mairal, F. Bach, J. Ponce, G. Sapiro, “Online learning for matrix factorization and sparse coding,” The Journal of Machine Learning Research, Vol.11, pp.19–60, 2010.

[29] H.G. Hirsc, D. Pearce,“The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions,” Proc. ISCAITRW ASR, 2000.

発表文献

国内研究会・全国大会

[1] 藤垣 健太郎, 柏木 陽佑, 齋藤 大輔, 峯松 信明, 広瀬 啓吉, “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, Vol.114, No.365, pp.69–74, 2014.

[2] 藤垣 健太郎, 柏木 陽佑, 齋藤 大輔, 峯松 信明, 広瀬 啓吉, “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”，日本音響学会春季講演論文集, 2-1-1, 2014.（発表予定）

学位論文

[3] 藤垣 健太郎, “時間構造が大きく異なる同一内容音声からの構造的特徴の抽出に関する実験的検討”，東京大学工学部電子工学科卒業論文, 2013.