修士論文

主成分分析に基づく特徴量強調を用いた雑音環境下における音声認識フロントエンドの高精度化

2012年 2月 8日

指導教員　峯松 信明 准教授

情報理工学系研究科
電子情報学専攻

48-106423　千々岩　圭吾

内容概要

スマートフォンなどの普及により，車内やレストランといった雑音のある環境でも広く音声認識が用いられるようになってきた．しかし雑音環境下における音声認識では，静かな環境で収録した音声で事前に学習した音響モデルと，雑音で歪んだ音声との間のミスマッチにより，認識精度が著しく低下するという問題がある．

この問題に対応するため，これまで様々な手法が提案されてきた．例えば，雑音に頑健な特徴量を用いる手法や，音響モデルを雑音環境に適応させる手法などがある．その中でも本研究では特に，比較的計算量が少なく，非常に大きな効果を発揮している特徴量強調に着目した．特徴量強調は，音声特徴量の統計的情報などを利用して，雑音付加音声特徴量からクリーン音声特徴量に変換する手法である．具体的には，SPLICE と呼ばれる既存手法の主成分分析に基づいた高精度化に取り組んだ．

従来の SPLICE は変換関数を学習した環境と似た雑音に対しては非常に有効であるが，未知の雑音環境下においては十分な性能を発揮することが保証されていない．そこで本研究では，未知の雑音環境にも適応できる Eigen-SPLICE を提案した．基本的な枠組みとしては，計算量を抑え，少量の適応データで変換関数を適応できるように，主成分分析を用いて適応すべきパラメータを削減した．この際，入力発声の雑音のみの区間を学習データ中のクリーン音声に重畳することで，適応に必要な擬似パラレルデータを作成した．その他にも，固有声に基づいた声質変換手法を，雑音環境下における特徴量強調の枠組みに導入した Eigen Joint GMM 法なども提案した．

これら２つの提案手法は，基本的な枠組みは似ているが，幾つかの違いがある．Eigen SPLICE は適応の対象が変換関数であるために，適応データとして擬似パラレルデータが必要があるが，重みベクトルの推定は解析的に計算できる．一方，Eigen Joint GMM 法は適応の対象が GMM のパラメータであるため，擬似パラレルデータは必要としないが，重みベクトルの推定には繰り返し計算によって局所最適解を求めることしかできない．

本研究では雑音環境下における音声認識データベース AURORA-2 を用いて実験を行い，従来手法との性能比較を行い，提案手法の有効性が示された．特に，Eigen-SPLICE の性能改善は大きく，未知雑音環境下においても十分な性能を発揮した．

提案手法の今後の課題としては，乗法性雑音への対応がある．擬似パラレルデータを作るときにクリーン音声に雑音を重畳する必要がり，現在の手法では原理的に乗法性雑音には対応できない．また，今後の展望としては，より頑健な特徴量の導入や Uncertainty Decoding の導入が考えられる．

目次

図目次

表目次

第1章

序論

1.1 研究の背景

近年の統計的な音声認識システムの発達により，身近な場面でも音声認識が使われるようになってきた．例えば，会議の議事録制作支援のために音声認識が用いられたり [1]，テレビの字幕をリアルタイムで制作するのを支援するのに音声認識が用いられたりしている [2]．また，最近ではカーナビゲーションシステムやスマートフォンの普及により [3]，スタジオのような静かな環境だけでなく，車内やレストランなど雑音の多い場面でも音声認識が用いられるようになってきた．雑音環境で収録された歪んだ音声は，クリーン音声を用いて学習した認識用の音響モデルと間にミスマッチを生じ，認識精度を著しく低下させるという問題がある [4].

この問題に対応するために，これまで様々な手法が提案されてきた．これらの耐雑音性に関する研究は，大きく分けて以下の５つのアプローチがある [5].

- 雑音に頑健な特徴量
- 音声強調
- モデル適応
- ミッシングフィーチャー理論
- モデルのマルチ条件学習

まず，雑音に頑健な特徴量を用いる手法である．つまり，音声波形が雑音によって歪められてしまっても，その特徴量は変化しにくいものを音声認識に用いるのである．例えば，係数間の相関が低く，低次元で音声信号の情報を表現できる MFCC(Mel-Frequency Cepstrum Coefficient)[6] やそれらに NMN(Noise Mean Normalization) や HEQ (Histogram Equalization) などの正規化を施した特徴量が挙げられる [7].

次に音声強調を用いた手法がある．音声強調は，雑音によって歪められてしまった音声特徴量を，元のクリーンな音声特徴量に変換する手法である．単純なものとしては，雑音のスペクトルの平均成分を雑音付加音声のスペクトルから差し引く，スペクトルサブトラクションなどがある．また，部分的線形変換によってより複雑な特徴量変換を実現す

る SPLICE(Stereo Piecewise LInear Compensation for Environments)[8] や VTS((Vector Taylor Series:ベクトルテーラー展開) を用いた手法 [9] なども高い性能を発揮している．

更に，モデル適応などの手法もある．これは，クリーン音声で学習した音響モデルを雑音付加音声の音響モデルへ適応する手法である．例えば，音響モデル HMM(Hidden Markov Model:隠れマルコフモデル) のトポロージやその他のパラメータを変更してモデル適応する手法 [10] やトポロジーは変更せずに計算量を比較的抑えた手法 [11] などがある．しかし，これらのモデル適応の手法は一般的に，非常に計算量を多く要する．

ミッシングフィーチャー理論に基づく手法は，雑音によって歪んだ特徴量を信頼できる部分と信頼できない (ミッシングフィーチャー) 部分に分けて，音声認識する．信頼できない部分を除いて認識する手法 [12] と信頼できない部分を再構成する手法 [13] がある．

最後のマルチ条件での音響モデル学習は，音響モデルを雑音付加音声によって学習するものである．しかし，これには様々な種類の雑音付加音声が必要になり，また計算量も非常に多く必要とする．

1.2　研究の目的

耐雑音性を得るために様々な手法が提案されている．その中でも計算量が低く，多くの場面において適用可能で高い効果を発揮する音声強調の技術が着目されている．その中でも，雑音付加音声特徴量の確率分布を GMM によってモデル化し，その特徴量空間を確率的に分割して，区分的線形変換を実現する SPLICE[8] や，VTS 近似を用いてクリーン音声特徴量の GMM と雑音の推定特徴量から雑音付加音声の GMM を作成し，区分的線形変換を実現する手法 [9] などが特に高い性能を発揮している．

しかし，これらの手法には幾つかの問題がある．まず，従来の SPLICE は変換関数を事前に学習した環境と，入力の雑音環境が似ているということを暗に仮定しており，未知の雑音環境下において十分な性能を発揮するということが保証されていない．また，VTS を用いた手法は雑音付加音声の GMM を作成する際に，雑音による影響をモデル化した歪関数を用いる．MFCC などの雑音に頑健な特徴量を用いた場合，その歪関数が複雑になりすぎて，膨大な計算量を必要とするという問題がある．

そこで本研究の目的は，比較的少ない計算量で未知の雑音環境にも対応できる，音声強調手法の提案である．

1.3　本論文の構成

本論文の構成は以下のとおりである．既に，序論で研究の背景と目的について述べた．第２章においては，音声認識システムの概要と，雑音がそれにどのように影響するのかを説明する．さらに第３章では，本研究のベースとなった従来手法，SPLICE，VTS を用いた手法，Joint GMM を用いた手法について詳しく説明していく．第４章では，従来手法の問題点を解決する Eigen-SPLICE や Eigen Joint GMM について説明する．第５章では，提案

手法の有効性を確認するために行った，雑音環境下における音声認識実験について説明する．最後に第６章の結論では，本論文のまとめと今後の課題や展望について述べていく．

第2章

音声認識システム
雑音による影響

2.1 はじめに

様々な場面において，音声信号は雑音によって歪められる．雑音の影響は，主に加法性雑音と乗法性雑音に分けられる．加法性雑音は所望の音源以外から発生される雑音で，例えばレストランなどでの他の人の会話や食器の音，電車の中でのモーター音や車輪が擦れる音などである．乗法性雑音は使用するマイクや回路の違いによって生じる雑音で，例えば学習データとは異なるマイクを用いた場合や，電話回線を通することによって発声するチャンネル歪みである．

これらの雑音によって歪められた音声は，音声認識においてクリーン条件で学習された音響モデルとのミスマッチを生じ，音声認識の精度を著しく低下させる．これは，近年広く用いられている統計的な音声認識システムが，入力の音声と学習に用いた音声がどれだけの確率的に似ているかを基準にして，認識しているためである．

この章では，まず音声特徴量や音声認識システムの概要を説明する [14]．その後，加法性雑音や情報性雑音がどのように音声波形やその特徴量に影響を与えるのか，そのモデル化について説明する．

2.2 音響特徴量

2.2.1 スペクトル

音声認識を行う際には一般的に，収録された音声波形をそのまま用いることは殆どない．一般的に，音声認識に必要な特徴は特にその周波数特性であるスペクトルによって表現される．スペクトルは，まず音声波形を短時間毎に切り出し，適切な窓関数をかける．それにより，離散フーリエ変換 (Discrete Fourier Transform: DFT) が可能になり，それぞれの周波数がどの程度の振幅や位相を持っているか表すスペクトル (Spectrum) を得ることができる．

2.2.2　ケプストラム

音声分析によって抽出される特徴量のうち，上記のスペクトル情報を表現しかつ最も扱いやすい特徴量として広く用いられているのがケプストラム（Cepstrum）である．まずスペクトルを求めた後，その対数パワースペクトルを求め，それに対して逆離散フーリエ変換（Inverse DFT: IDFT）を施して得られるのがケプストラムである．

ケプストラムは現在の音声情報処理の基礎である線形分離等価回路モデル（ソースフィルタモデル）に基づいている．ソースフィルタモデルでは，人間の音声生成の過程に基づき，人間の音声が声帯の振動による音源特性 $G(\omega)$ に対して，人間の声道における調音の特性 $H(\omega)$ を伝達関数として我々の耳に届いていると考える．すなわち周波数領域において生成される音声 $S(\omega)$ を以下の式で表す．

$$S(\omega) = G(\omega)H(\omega) \tag{2.1}$$

式 (2.1) の絶対値の対数をとりこれを逆フーリエ変換する．対数をとる処理により積関係を和の形で分離できる．

$$\begin{aligned} c(\tau) &= \mathcal{F}^{-1} \log |S(\omega)| \\ &= \mathcal{F}^{-1} \log |G(\omega)| + \mathcal{F}^{-1} \log |H(\omega)| \end{aligned} \tag{2.2}$$

このとき式 (2.2) における $c(\tau)$ が連続量としてのケプストラムである．IDFT により得られたベクトルの低次項のみを残し，高次項を 0 とした後に，DFT することでスペクトル包絡が得られる．これは式 (2.2) における $\mathcal{F}^{-1} \log |H(\omega)|$，すなわち声道の調音特性に対応する．スペクトル包絡の山の部分は声道の共鳴周波数に対応しフォルマント周波数と呼ばれる．音声の音韻的特徴はこのフォルマント周波数によく表れる．つまりケプストラムはスペクトル情報を効率的に表現するベクトル特徴量となる．

以上のような流れで音声波形からケプストラムを抽出する過程を図 2.1 に示す．

2.2.3　聴覚特性を反映したケプストラム

人間の音の高さに対する周波数分解能は低い周波数ほど細かく，高い周波数ほど粗い事が知られている．このような聴覚特性をケプストラム特徴量に反映させたものが幾つか存在する．

MFCC はメル周波数と呼ばれる，人間の聴覚特性を反映した周波数軸上において等間隔に配置された三角窓を用意し，フィルタバンク分析を行う事で求められる．各窓毎に対応する周波数帯域のパワーを求め，窓の大きさの重みをつけて和をとることでメル化したスペクトルの離散情報が得られ，これに離散コサイン変換を施すことで MFCC が求められる．なおメル周波数は以下の周波数ウォーピングで求める．

$$f_{mel}(f) = 2595 \log_{10} \left(1 + \frac{f}{700} \right) \tag{2.3}$$

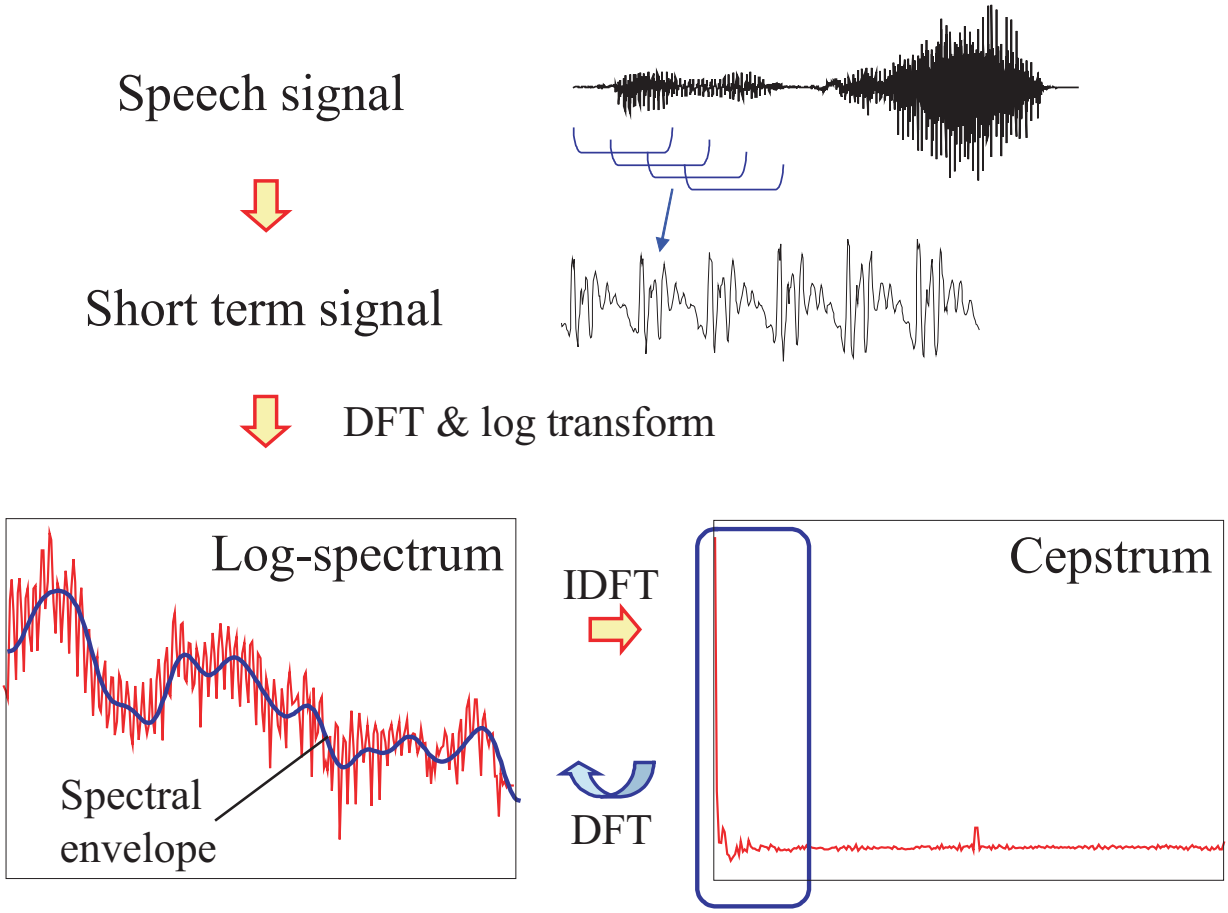

図 2.1: Extraction of cepstrum

2.2.4　動的特徴を考慮したケプストラム

スペクトルの時間軸に対する動的な特徴をとらえるため，ケプストラムベクトルの時間変化量である Δ ケプストラムがある．Δ ケプストラムは差分に基づく特徴量であるため，マイクの伝達特性の変化等に頑健で，時間変化量として動的特徴を表現するのに適していると考えられており，広く用いられている．また，更にその動的な特徴量を捉えるために，それを微分した ΔΔ ケプストラムも音声認識の精度を向上させるために用いられている．

2.3　HMM による音響特徴のモデル化

2.3.1　隠れマルコフモデル（HMM）

隠れマルコフモデル (HMM) は信号源間の状態遷移確率と信号源からの出力ベクトルの確率分布をパラメータとして持ち，状態遷移とベクトルの出力を繰り返す生成モデルである．図 2.2 に HMM の構造を示す．図 2.2 において S_i は i 番目の状態を，a_{ij} が S_i から S_j への遷移確率を表している．各状態 S_i はベクトル $\boldsymbol{x}$ を出力する確率 $b_i(\boldsymbol{x})$ をパラメータと

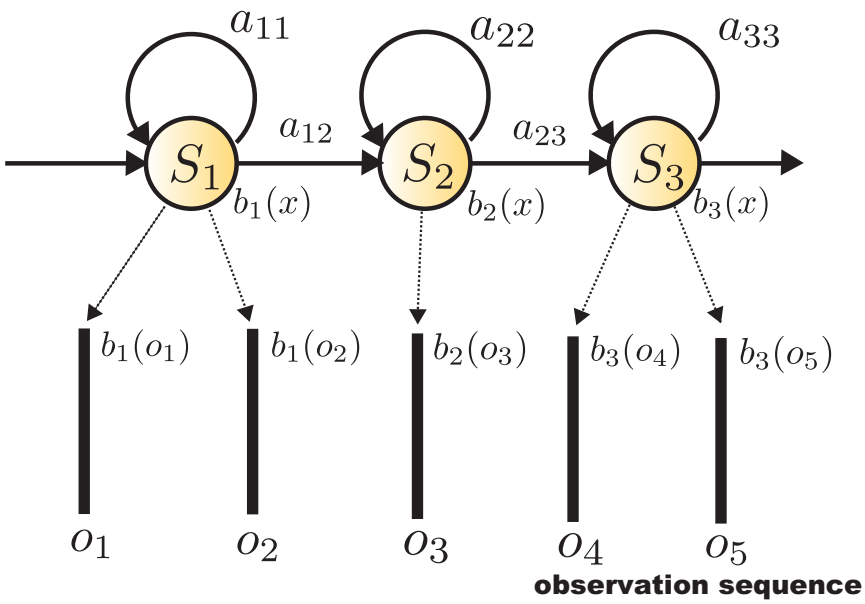

図 2.2: Structer of HMM

して持つ．このとき $b_i(\boldsymbol{x})$ の分布系として混合ガウス分布に基づくものが広く用いられている．

2.3.2　HMM の学習

HMM において学習すべきパラメータは $\boldsymbol{\theta} = \{a_i, b_i(\boldsymbol{x})\}$ であるが，これは最尤（Maximum Likelihood; ML）推定に基づいて行なわれる．即ち，学習データから音声特徴量の時系列データ $\boldsymbol{X}$ が観測されたとき，その尤度を最大化する $\boldsymbol{\theta}$ を求める問題に帰着され，

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, P(\boldsymbol{X}|\boldsymbol{\theta}) \tag{2.4}$$

が求めるパラメータとなる．しかし，HMM の場合は隠れ変数 (外部から直接観測することができない変数．HMM の枠組みでは，データ系列 $\boldsymbol{X}$ が観測されたとき，その各々がどの状態から生じたものなのかまでを観測することはできない．) が存在し，式 (2.4) を解析的に解くのは困難である．このため，実際には式 (2.4) の局所最適解を求める Baum-Welch アルゴリズムが用いられる．

Baum-Welch アルゴリズムでは前向き変数 $\alpha_i(t)$，後向き変数 $\beta_i(t)$ と呼ばれる変数が登場する．これらは，時刻 t における状態が i であれば 1，そうでなければ 0 をとる隠れ変数を z_{ti} とすると，

$$\alpha_i(t) = P(z_{ti} = 1, \boldsymbol{x}_1, \cdots, \boldsymbol{x}_t|\boldsymbol{\theta}) \tag{2.5}$$
$$\beta_i(t) = P(\boldsymbol{x}_{t+1}, \cdots, \boldsymbol{x}_T|z_{ti} = 1, \boldsymbol{\theta}) \tag{2.6}$$

と表すことができるものである．この前向き変数 $\alpha_i(t)$ 及び後向き変数 $\beta_i(t)$ を用いて，時

刻 t における状態が i である確率 $\overline{z_{ti}}$ を，

$$\overline{z_{ti}} = P(z_{ti}=1|\boldsymbol{X},\boldsymbol{\theta}) \tag{2.7}$$

$$= \frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^{N}\alpha_t(j)\beta_t(j)} \tag{2.8}$$

のようにして求めることができる．上記は，パラメータ $\boldsymbol{\theta}$ と学習データからの音声特徴量の時系列データ $\boldsymbol{X}$ を用いれば，そのデータ系列の各々が特定の状態から生じた確率を求めることができることを意味する．式 (2.5) から式 (2.8) を用いれば，新しいパラメータを最尤推定によって求めることができる．例えば，出力確率 $b_i(\boldsymbol{x})$ の分布形として単一ガウス分布 $\mathcal{N}(\boldsymbol{x};\boldsymbol{\mu}_i,\boldsymbol{\sigma}_i^2)$ を用いる場合，パラメータ $\boldsymbol{\theta}=\{a_i,\boldsymbol{\mu}_i,\boldsymbol{\sigma}_i^2\}$ に対して，新しいパラメータ $\hat{\boldsymbol{\theta}}=\{\hat{a}_i,\hat{\boldsymbol{\mu}}_i,\hat{\boldsymbol{\sigma}}_i^2\}$ を，

$$\hat{a}_i = \frac{\sum_{t=1}^{T-1}\alpha_i(t)a_ib_i(\boldsymbol{x}_{t+1})\beta_{i+1}(t+1)}{\sum_{t=1}^{T-1}\alpha_i(t)\beta_i(t)} \tag{2.9}$$

$$\hat{\boldsymbol{\mu}}_i = \frac{\sum_{t=1}^{T}\overline{z_{t,i}}\boldsymbol{x}_t}{\sum_{t=1}^{T}\overline{z_{t,i}}} \tag{2.10}$$

$$\hat{\boldsymbol{\sigma}}_i^2 = \frac{\sum_{t=1}^{T}\overline{z_{t,i}}(\boldsymbol{x}_t-\boldsymbol{\mu}_i)^2}{\sum_{t=1}^{T}\overline{z_{t,i}}} \tag{2.11}$$

のようにして求めることができる．式 (2.9) から式 (2.11) の算出には，式 (2.5) から式 (2.8) を求めておく必要があり，一方，式 (2.5) から式 (2.8) の算出には，パラメータを式 (2.9) から式 (2.11) によって求めておく必要があるので，両者は互いに依存関係にある．しかしながら，このようにして得たパラメータ $\hat{\boldsymbol{\theta}}$ は，パラメータ $\boldsymbol{\theta}$ に対して常に，

$$P(\boldsymbol{X}|\boldsymbol{\theta}) \leq P(\boldsymbol{X}|\hat{\boldsymbol{\theta}}) \tag{2.12}$$

が成立するので，式 (2.5) から式 (2.8) の算出と，式 (2.9) から式 (2.11) の算出を繰り返す反復アルゴリズムによって，パラメータは局所最適解に収束する．

2.4　音声認識システム

次に，抽出された特徴量や学習した音響モデルを用いて，どのように音声認識を行うのかについて説明していく [15]．音声認識システムでは，収録された音声波形から特徴量を抽出し，それを入力とする．その後，音響モデルとのマッチングを行うことで音声を認識する．また大語彙連続音声認識などの場合は，それに加えて言語としての単語の生起しやすさをモデル化した言語モデルも用いて，より精度の高い音声認識を可能にする．

2.4.1　定式化

このような音声認識は，入力音声の特徴量を $\boldsymbol{x}$ としたとき，事後確率が最大となる単語列 $\boldsymbol{W}$ を見つける問題として以下のように定式化できる．

$$\hat{\boldsymbol{W}} = \underset{\boldsymbol{W}}{\operatorname{argmax}}\, p(\boldsymbol{W}|\boldsymbol{x}) \tag{2.13}$$

これをベイズの定理で展開すると

$$p(\boldsymbol{W}|\boldsymbol{x}) = \frac{p(\boldsymbol{W})p(\boldsymbol{x}|\boldsymbol{W})}{p(\boldsymbol{x})} \tag{2.14}$$

となり分子が，どのような単語が出現しやすいかという言語モデル $p(\boldsymbol{W})$ の部分と，ある単語を発音するときにどのような特徴量になりやすいかという音響モデル $p(\boldsymbol{x}|\boldsymbol{W})$ の部分に分解される．ここで，分母の $p(\boldsymbol{x}) = \sum_{\boldsymbol{W}} p(\boldsymbol{W})p(\boldsymbol{x}|\boldsymbol{W})$ は，$\hat{\boldsymbol{W}}$ の決定に寄与しないために，認識では無視される (認識の信頼度を計算する場合には必要)．こうすることで，音声認識は音響モデル的にも言語モデル的にも生起確率が最も高い単語列を探す問題に帰着する．

2.4.2　音響モデル

音響モデルとしては，前節で述べた HMM が広く用いられる．一般的な音声認識に用いられる HMM の各状態は確率分布を混合正規分布 (Gaussian Mixture Model: GMM) で表す．一般的に音響モデルは，学習に用いた音声の性質 (話者性など) に大きく依存し，高い精度を得るためには多くの学習データが必要である．そのため，学習には静かで雑音のない環境で収録された複数の話者による大量の音声データを用いる．

2.4.3　言語モデル

言語モデルは，認識結果が言語として妥当なものにする制約条件である．一般的には，単語 n-gram モデルが用いられることが多い．単語 n-gram は与えられた単語列 $w_1, w_2 \ldots w_n$ に対して，その出現確率を与える統計モデルである．例えば，“私は発表” という単語列の後に，“する” という単語が来る確率が，“です” という単語が来る確率よりも高いという様なことを記述するモデルである．これによって，音響モデルと入力特徴量とのマッチングによって推定された認識単語の候補の中から，言語的に最も尤もらしいものを選び出すことが可能になる．

この n-gram モデルの学習は，大量の文章に出現する単語列を集計することによって学習する．しかし，単語の組み合わせは非常に多くあり，全部の単語の組み合わせに対して十分な量の学習データを得ることは現実的には難しい．そこで，この学習データのスパース性の問題を解決するために，様々な手法で言語モデルの平滑化が行われる．平滑化とは，学習データの中に登場していない単語列であっても，その生起確率を 0 にせず，ある小さな値を与えることである．もし平滑化を施さない言語モデルであった場合，学習データに

は登場しなかったが言語的には正しい単語列を認識することができなくなってしまう．このような言語モデルに対する平滑化処理は数多く提案されている．

2.4.4　デコーディング

音響モデルや言語モデルなど学習された後，入力された音声特徴量に対して適切な単語列を見つけ出すのは，デコーダ (認識エンジン) の役割である．デコーダは，様々な単語列 $\boldsymbol{W}$ に対して仮説を生成し，その生起確率を計算し，それが最大になるものを選択する．

$$\begin{aligned} \hat{\boldsymbol{W}} &= \operatorname{argmax} p(\boldsymbol{W}|\boldsymbol{x}) && (2.15)\\ &= \operatorname{argmax} p(\boldsymbol{W})p(\boldsymbol{x}|\boldsymbol{W}) && (2.16)\\ &= \operatorname{argmax} \{\log p(\boldsymbol{W}) + \log p(\boldsymbol{x}|\boldsymbol{W})\} && (2.17) \end{aligned}$$

この仮説を検証するなかで，すべての仮説に対して確率を計算してしまうと計算量が膨大になりすぎるために，尤度の高い仮説のみ検証していく枝刈りなどの処理が適宜施されていく．

以上のように，モデルの学習，特徴量の抽出，デコーディングなどの流れで統計的な音声認識は行われている．この概要を図 2.3 に示す．

2.5　雑音による影響

レストランや会議室，駅や車の中といった環境では，様々な種類と大きさの雑音によって音声信号は歪む．その歪によって入力音声と音響モデルとのミスマッチを生じさせている．以下，音声波形領域，スペクトル領域，ケプストラム領域で，どのように雑音による歪みがモデル化されるのかを説明する [5].

2.5.1　音声波形領域

歪みを生じさせる雑音は大きく分けて，乗法性雑音と加法性雑音の二つに分けられる．乗法性雑音は，残響やマイク特性といった物であり，スペクトル領域においては音声スペクトルへの乗算によって表現できる．一方，加法性雑音は，他の音源から発生する背景雑音などがあり，スペクトル領域において加算によって表現される．これらをモデル化すると，図 2.4 および (2.18) の様になる．

$$\boldsymbol{y} = (\boldsymbol{x} \otimes \boldsymbol{h}) + \boldsymbol{n} \tag{2.18}$$

ただし，ここで $\boldsymbol{y}$ は雑音によって歪んだ音声を表す．また，$\boldsymbol{x}$ は雑音によって歪む前の音声を表し，$\boldsymbol{n}$ は他の音源などの加法性雑音を表し，$\boldsymbol{h}$ はチャンネル歪みなどの乗法性雑音を表す．またこのとき，$\otimes$ は畳み込み積分を表す．

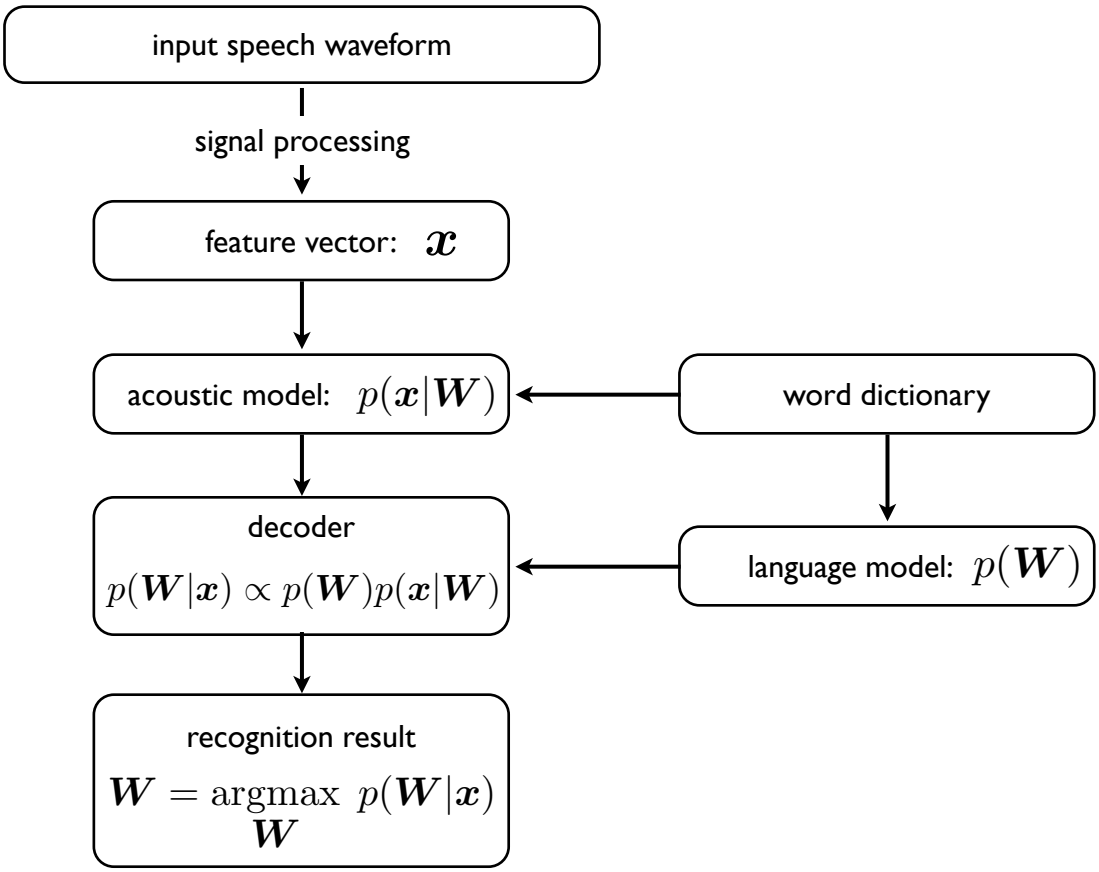

図 2.3: Overview of speech recognition

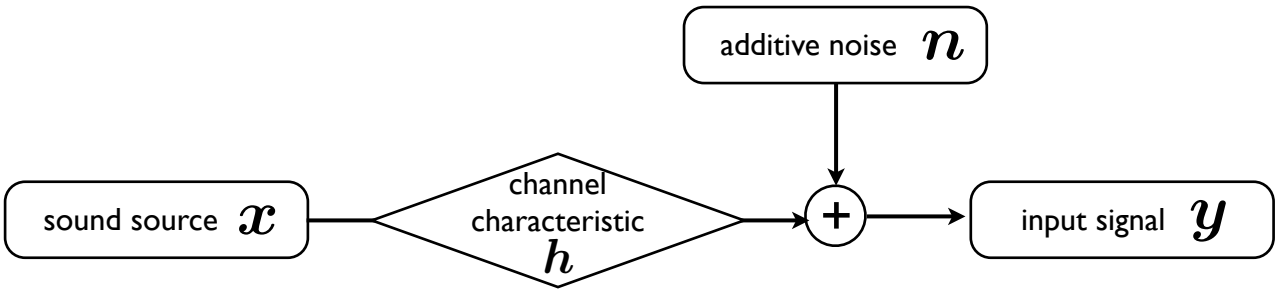

図 2.4: Model of distortion by noise

2.5.2　スペクトル領域

これらに対して，短時間フーリエ変換を施すと，以下のようになる．

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{H} + \boldsymbol{N} \tag{2.19}$$

ここで，$\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{H}, \boldsymbol{N}$ はそれぞれ雑音付加音声，クリーン音声，乗法性歪みと加法性歪みを短時間フーリエ変換したものである．

さらに，パワースペクトルをとると，

$$\begin{aligned} |\boldsymbol{Y}|^2 &= (\boldsymbol{X}\boldsymbol{H} + \boldsymbol{N})^* (\boldsymbol{X}\boldsymbol{H} + \boldsymbol{N}), \\ &= |\boldsymbol{X}|^2|\boldsymbol{H}|^2 + |\boldsymbol{N}|^2 + 2\mathrm{Re}(\boldsymbol{X}\boldsymbol{H}\boldsymbol{N}^*), \\ &= |\boldsymbol{X}|^2|\boldsymbol{H}|^2 + |\boldsymbol{N}|^2 + 2\cos(\Phi)|\boldsymbol{X}\boldsymbol{H}||\boldsymbol{N}|. \end{aligned} \tag{2.20}$$

ここで，$*$ は複素共役を，Re[] は複素数の実部を求める演算子，Φ は加法性雑音とクリーン音声の位相の差を表す．しかし，最後の式の第三項は位相差 Φ は，$[-\pi, \pi]$ の範囲に分布し，$\cos(\Phi)$ の期待値が 0 になる．そのため多くの場合，第三項は省略され

$$|\boldsymbol{Y}|^2 = |\boldsymbol{X}|^2|\boldsymbol{H}|^2 + |\boldsymbol{N}|^2. \tag{2.21}$$

となる．

2.5.3　ケプストラム領域

次に，メル周波数フィルタバンク処理を施すと，ケプストラム領域においては常用対数と離散コサイン変換で以下のように表される [16]．

$$\mathbf{y} \simeq f(\mathbf{x}, \mathbf{h}, \mathbf{n}), \tag{2.22}$$

$$= \mathbf{x} + \mathbf{h}\mathbf{C}\log(1 + \exp(\mathbf{C}^+(\mathbf{n} - \mathbf{h} - \mathbf{x}))) \tag{2.23}$$

ここで，$\mathbf{C}$ は離散コサイン変換行列を表し，$\mathbf{C}^+$ はそのムーア・ペンローズ擬似逆行列を表す．

以上のようにして，各領域における雑音による歪みはモデル化されている．

2.6　まとめ

この章では，まず音声特徴量の抽出と言った音声情報処理の基本的な枠組みについて述べた．その後，音声認識システムの概要について述べ，最後に雑音による歪みが各領域において，どのようにモデル化されるのかについて述べた．このように，様々な雑音によって歪められた音声特徴量は，音声認識システムにおいて，音響モデルとのミスマッチを発生させ，認識精度を著しく低下させる．

第3章

従来手法

3.1 はじめに

この章では，雑音環境下における特徴量強調の先行研究について述べる．まず，SPLICE と Joint GMM を用いた手法について説明する．これらの手法は，雑音付加音声とクリーン音声の対であるパラレルデータを用いて，入力された雑音付加音声特徴量を出力のクリーン音声特徴量へと変換する手法である．その後，クリーン音声特徴量の確率分布と推定された雑音の特徴量から雑音付加音声特徴量の GMM を VTS によって近似し，それを用いて特徴量変換を実現する手法を説明する．

3.2 SPLICE

この節では，まず SPLICE[8] について詳しく説明する．SPLICE は，雑音付加音声の特徴量の統計的な性質を GMM でモデル化し，その GMM の各コンポーネント毎に特徴量を線形変換することで，雑音付加音声からクリーン音声への変換を実現する．これは，雑音付加音声の特徴量からクリーン音声の特徴量への非線形変換を，区分的線形変換によって近似的に実現している．以下，従来手法の各段階を詳しく説明する．

3.2.1 仮定

SPLICE は，まず雑音環境下の音声の特徴量ベクトルの分布が GMM で表現出来ると仮定している．その分布を EM(Expectation-Maximization) アルゴリズムによって学習する．

$$\begin{aligned} p(\boldsymbol{y}) &= \sum_s p(\boldsymbol{y}, s) = \sum_s p(\boldsymbol{y}|s)p(s), \\ \text{where}, p(\boldsymbol{y}|s) &= \mathcal{N}(\boldsymbol{y}; \boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s) \end{aligned} \tag{3.1}$$

ここで $\boldsymbol{y}, s, \boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s$ はそれぞれ，雑音付加音声の特徴量ベクトル，GMM の各コンポーネントのインデックス，平均，分散を表す．

ここで，以下の仮定を設ける．雑音付加音声の特徴量ベクトルとそれが所属するコンポーネントが与えられたとき，その変換後のクリーン音声の確率分布も正規分布によって表され，その正規分布の平均は雑音付加音声のアフィン変換によって表されると仮定する．つまり次式で表される仮定を設ける．

$$p(\boldsymbol{x}|\boldsymbol{y}, s) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{A}_s\boldsymbol{y} + \boldsymbol{r}_s, \boldsymbol{\Gamma}_s) \tag{3.2}$$

ここで，$\boldsymbol{x}, \boldsymbol{r}_s, \boldsymbol{A}_s, \boldsymbol{\Gamma}_{\boldsymbol{s}}$ はそれぞれ，クリーン音声の特徴量ベクトル，コンポーネント s における平均ベクトルの補正ベクトル，線形変換を表す行列，分散共分散行列を表す．ただし，分散共分散 $\boldsymbol{\Gamma}_{\boldsymbol{s}}$ は，以後の変換手法には関与しない．

3.2.2　変換手法

以上のような仮定を設けることによって，ある入力の雑音付加音声の特徴量ベクトルが与えられたときの，出力のクリーン音声の特徴量は，下記の期待値として推定することができる．

$$\hat{\boldsymbol{x}} = E_x[\boldsymbol{x}|\boldsymbol{y}] = \sum_s p(s|\boldsymbol{y}) E_x[\boldsymbol{x}|\boldsymbol{y}, s] \tag{3.3}$$

ここで，(3.2) を用いると，

$$E_x[\boldsymbol{x}|\boldsymbol{y}, s] = \boldsymbol{A}_s\boldsymbol{y} + \boldsymbol{r}_s \tag{3.4}$$

となる．よって，雑音付加音声の特徴量から推定したクリーン音声の特徴量は，以下のようになる．

$$\hat{\boldsymbol{x}} = \sum_s p(s|\boldsymbol{y}) \left(\boldsymbol{A}_s\boldsymbol{y} + \boldsymbol{r}_s\right) \tag{3.5}$$

この変換は，入力が GMM のどのコンポーネントに所属しているかという事後確率を求め，その事後確率を重みとして各コンポーネントでの変換結果の重み付け和と解釈することが出来る．

3.2.3　学習

以上の特徴量の変換に必要なコンポーネント s における $\boldsymbol{r}_s, \boldsymbol{A}_s$ は，クリーン音声とそれに雑音が付加された雑音付加音声の対を用い，最小誤差基準によって以下のように学習する．

まず，以下のように $\boldsymbol{r}_s$ のように計算する．

$$\begin{aligned} \boldsymbol{r}_s &= \frac{\sum_t p(s|\boldsymbol{y}_t)(\boldsymbol{x}_t - \boldsymbol{y}_t)}{\sum_t p(s|\boldsymbol{y}_t)}, \\ \text{where}, p(s|\boldsymbol{y}_t) &= \frac{p(\boldsymbol{y}_t|s)p(s)}{\sum_s p(\boldsymbol{y}_t|s)p(s)} \end{aligned} \tag{3.6}$$

その後，以下のように $\boldsymbol{A}_s$ を計算する．

$$\boldsymbol{A}_s \quad = \quad \frac{\sum_t p(s|\boldsymbol{y}_t)\,(\boldsymbol{x}_t - \boldsymbol{r}_s)\,\boldsymbol{y}_t^{\mathrm{T}}}{\sum_t p(s|\boldsymbol{y}_t)\boldsymbol{y}_t\boldsymbol{y}_t^{\mathrm{T}}}. \tag{3.7}$$

このとき t は特徴量系列のインデックスである．あるコンポーネントの補正ベクトルは $\boldsymbol{r}_s$ は，雑音付加音声の特徴量がそのコンポーネントに所属する確率を重みとして，雑音付加音声とクリーン音声の差分の重み付け平均とすることが出来る．また，そのときの所属する事後確率は，ベイズの定理を用いて求めることが出来る．このときのクリーン音声と雑音付加音声の対，パラレルデータは実環境では口元のマイクで録音した歪が殆ど無い音声と離れたところで録音した歪んだ音声というようにして得ることが可能である．今回用いた AURORA–2 データベースではシミュレーションによって，クリーン音声に雑音を重畳している．

この SPLICE は用いる GMM の混合数を上げていくと，つまり特徴量空間をより細かく分割していくと，行列 $\boldsymbol{A}_s$ を単位行列にしても十分な効果が得られることが知られている [8]．しかし，本報告では $\boldsymbol{A}_s$ も非単位行列として推定した．

3.2.4　SPLICE の問題点

以上の SPLICE は変換を学習した雑音環境と入力の雑音環境が似ているということを暗に仮定しており，未知の雑音環境においては十分な性能を発揮することが保証されていない．そこで， 環境毎に GMM および区分的線形変換の方法を切り替える EMS(Environmental Model Selection) が SPLICE の改善手法として提案されている [17]．これは， 各々の学習環境依存の GMM を学習し，その GMM を用いて学習環境依存の区分的線形変換を学習する．そして変換の際には， 入力系列 $\boldsymbol{y}_t$ がどの環境 e に依存しているかを推定し，もっとも尤度の高い環境の GMM とその区分的線形変換関数を用いて変換する．

$$\hat{e} = \underset{e}{\mathrm{argmax}}\, p(\boldsymbol{y}_t|e) \tag{3.8}$$

本来は事後確率 $p(e|\boldsymbol{y}_t)$ を最大化する e を求めるべきであるが，事後確率を求めるためには下式のように，環境の確率 $p(e)$ が必要になり，計算できないので，$p(\boldsymbol{y}_t|e)$ の最高化を基準とした．

$$p(e|\boldsymbol{y}_t) = \frac{p(\boldsymbol{y}_t|e)p(e)}{\sum_e p(\boldsymbol{y}_t|e)p(e)} \tag{3.9}$$

これによってある程度の性能改善が見込まれるが，この EMS も依然として，事前に学習した変換方法でしか変換出来ないため，未知の雑音環境下において高い性能を発揮することが保証されていない．

3.3　Joint GMM を用いた変換手法

この節では，Joint GMM を用いた特徴量変換手法 (以下 J-GMM 法と記述)[18] について説明する．この手法は，まず学習データ中の入力特徴量と出力特徴量を連結することで

Joint Vector を得る．次に，得られた Joint Vector の確率分布を GMM を用いてモデル化する．これを Joint GMM と呼び，その平均や分散共分散行列などのパラメータを用いて，与えられた入力特徴量を所望の特徴量に変換する．この細かい手順について雑音環境下における音声特徴量強調の枠組みとして，以下述べていく．

3.3.1　パラレル学習

まず，Joint GMM の学習について述べる．ここでは，t フレーム目 (最大 T まで) の入力の雑音付加音声の特徴量を $\boldsymbol{y}_t$，出力のクリーン音声特徴量を $\boldsymbol{x}_t$ と記述する．まず，クリーン音声特徴量とそれに雑音が付加された雑音付加音声の特徴量を対応するフレーム毎に連結し，Joint Vector $\left[\boldsymbol{x}_t^{\mathsf{T}}, \boldsymbol{y}_t^{\mathsf{T}}\right]^{\mathsf{T}}$ を作成する．但し，$^{\mathsf{T}}$ は転置を表す．次に得られた Joint Vector を用いて，結合確率密度 $p\left(\boldsymbol{x}_t, \boldsymbol{y}_t|\boldsymbol{\lambda}\right)$ を GMM として学習する．

$$\hat{\boldsymbol{\lambda}} = \underset{\boldsymbol{\lambda}}{\operatorname{argmax}} \prod_{t=1}^{T} p\left(\boldsymbol{x}_t, \boldsymbol{y}_t|\boldsymbol{\lambda}\right). \tag{3.10}$$

$$\begin{aligned} p\left(\boldsymbol{x}_t, \boldsymbol{y}_t|\boldsymbol{\lambda}\right) &= \sum_{s=1}^{S} \alpha_s \mathcal{N}\left(\boldsymbol{x}_t, \boldsymbol{y}_t; \boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s\right), \\ \boldsymbol{\mu}_s = \begin{bmatrix} \boldsymbol{\mu}_s^x \\ \boldsymbol{\mu}_s^y \end{bmatrix}&, \ \boldsymbol{\Sigma}_s = \begin{bmatrix} \boldsymbol{\Sigma}_s^{xx} & \boldsymbol{\Sigma}_s^{xy} \\ \boldsymbol{\Sigma}_s^{yx} & \boldsymbol{\Sigma}_s^{yy} \end{bmatrix}. \end{aligned} \tag{3.11}$$

$\boldsymbol{\lambda}$ は GMM のパラメータを表し，学習には EM アルゴリズムを用いる．ここで $\mathcal{N}\left(\boldsymbol{x}_t, \boldsymbol{y}_t; \boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s\right)$ は，平均・分散共分散がそれぞれ $\boldsymbol{\mu}_s$・$\boldsymbol{\Sigma}_s$ となる正規分布であり，s は正規分布のインデックスを表している．また，α_s は s 番目の正規分布の重みを表す．

3.3.2　最尤変換

J–GMM 法では，下記の尤度関数の最大化基準に基づき，変換を施す．

$$p\left(\boldsymbol{x}_t|\boldsymbol{y}_t, \boldsymbol{\lambda}\right) = \sum_{s=1}^{S} p(s|\boldsymbol{y}_t, \boldsymbol{\lambda}) p(\boldsymbol{x}_t|\boldsymbol{y}_t, s, \boldsymbol{\lambda}), \tag{3.12}$$

t フレーム目における $p(s|\boldsymbol{y}_t, \boldsymbol{\lambda})$ および $p(\boldsymbol{x}_t|\boldsymbol{y}_t, s, \boldsymbol{\lambda})$ は以下のようになる．

$$p(s|\boldsymbol{y}_t, \boldsymbol{\lambda}) = \frac{\alpha_s \mathcal{N}\left(\boldsymbol{y}_t; \boldsymbol{\mu}_s^y, \boldsymbol{\Sigma}_s^{yy}\right)}{\sum_{i=1}^{S} \alpha_i \mathcal{N}\left(\boldsymbol{y}_t; \boldsymbol{\mu}_i^y, \boldsymbol{\Sigma}_i^{yy}\right)}, \tag{3.13}$$

$$p(\boldsymbol{x}_t|\boldsymbol{y}_t, s, \boldsymbol{\lambda}) = \mathcal{N}\left(\boldsymbol{x}_t; \boldsymbol{E}_t(s), \boldsymbol{D}(s)\right), \tag{3.14}$$

ただし，

$$\boldsymbol{E}_t(s) = \boldsymbol{\mu}_s^x + \boldsymbol{\Sigma}_s^{xy} \boldsymbol{\Sigma}_s^{yy^{-1}} \left(\boldsymbol{y}_t - \boldsymbol{\mu}_s^y\right), \tag{3.15}$$

$$\boldsymbol{D}(s) = \boldsymbol{\Sigma}_s^{xx} - \boldsymbol{\Sigma}_s^{xy} \boldsymbol{\Sigma}_s^{yy^{-1}} \boldsymbol{\Sigma}_s^{yx}. \tag{3.16}$$

所望のクリーン音声の特徴量は以下のように求められる．

$$\hat{\boldsymbol{x}}_t = \underset{\boldsymbol{x}}{\operatorname{argmax}}\, p(\boldsymbol{x}_t|\boldsymbol{y}_t, \boldsymbol{\lambda}), \tag{3.17}$$

$$= \underset{\boldsymbol{x}}{\operatorname{argmax}} \sum_{s=1}^{S} p(s|\boldsymbol{y}_t, \boldsymbol{\lambda}) p(\boldsymbol{x}_t|\boldsymbol{y}_t, s, \boldsymbol{\lambda}). \tag{3.18}$$

これから出力される特徴量は，

$$\hat{\boldsymbol{x}}_t = \sum_{s=1}^{S} p(s|\boldsymbol{y}_t, \boldsymbol{\lambda}) \left\{ \boldsymbol{\mu}_s^x + \boldsymbol{\Sigma}_s^{xy} \boldsymbol{\Sigma}_s^{yy^{-1}} (\boldsymbol{y}_t - \boldsymbol{\mu}_s^y) \right\} \tag{3.19}$$

のように推定される．

3.3.3　Joint GMM を用いた手法の問題点

この J-GMM 法も SPLICE と同様に，Joint GMM を学習した雑音環境と入力音声の雑音環境が似ているということを暗に仮定しており，未知の雑音環境下において十分な性能を発揮する h ことが保証されていない．一方，J-GMM は後述する主成分分析による改良を施した際に，SPLICE が変換関数の適応に擬似パラレルデータを必要とするのに対して，それを必要としないという利点が有る．

3.4　VTS を用いた手法

この節では，ベクトル・テーラー展開 (VTS) を用いた特徴量変換について説明する．この手法は，雑音付加音声の特徴量の確率分布を，歪関数を用いてクリーン音声特徴量と雑音特徴量から VTS 近似することで，特徴量変換を実現する．

3.4.1　歪関数

ここで雑音付加音声の特徴量を $\boldsymbol{y}$，クリーン音声の特徴量を $\boldsymbol{x}$，加法性雑音の特徴量を $\boldsymbol{n}$ とする．特徴量が対数メルフィルタバンク出力 (FBANK) で，乗法性雑音と位相を無視した場合に，クリーン音声と雑音付加音声の差分を表す歪関数 $g(\boldsymbol{x}, \boldsymbol{n})$ は以下のように表される．

$$\boldsymbol{y} = \boldsymbol{x} + g(\boldsymbol{n}, \boldsymbol{x}) \tag{3.20}$$

$$g(\boldsymbol{n}, \boldsymbol{x}) = \log(1 + \exp(\boldsymbol{n} - \boldsymbol{x})). \tag{3.21}$$

また，特徴量が MFCC で同様に乗法性雑音を無視した場合は，歪関数は以下のようになる．

$$g(\boldsymbol{n}, \boldsymbol{x}) = \mathbf{C} \log(1 + \exp(\mathbf{C}^{+}(\boldsymbol{n} - \boldsymbol{x}))). \tag{3.22}$$

ここで，$\mathbf{C}, \mathbf{C}^{+}$ はそれぞれ離散コサイン変換行列，それの擬似逆行列である．

3.4.2　確率分布の学習

VTS 近似を用いた手法では，まずクリーン音声特徴量を GMM によって以下のように学習する．

$$p(\boldsymbol{x}) = \sum_s p(\boldsymbol{x}, s) = \sum_s p(\boldsymbol{x}|s)p(s), \tag{3.23}$$

$$\text{where}, p(\boldsymbol{x}|s) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_s^x, \boldsymbol{\Sigma}_s^x) \tag{3.24}$$

但し，$s, \boldsymbol{\mu}_s^x, \boldsymbol{\Sigma}_s^x$ はそれぞれ，GMM のコンポーネントのインデックス，コンポーネント s の平均と分散共分散行列である．

次に，雑音の特徴量の確率分布を正規分布として以下のように学習する．

$$p(\boldsymbol{n}) = \mathcal{N}(\boldsymbol{n}; \boldsymbol{\mu}^n, \boldsymbol{\Sigma}^n) \tag{3.25}$$

このとき，$\boldsymbol{\mu}^n, \boldsymbol{\Sigma}^n$ はそれぞれ正規分布の平均と分散共分散行列である．

3.4.3　VTS 近似

次に，クリーン音声と雑音の特徴量が既知となったとき，雑音付加音声の特徴量の確率分布を GMM として近似するために，歪関数を VTS によって以下のように近似する．

$$g(\boldsymbol{n}, \boldsymbol{x}) \simeq \frac{\partial}{\partial \boldsymbol{n}} g(\boldsymbol{n}_0, \boldsymbol{x}_0)\boldsymbol{n} + \frac{\partial}{\partial \boldsymbol{x}} g(\boldsymbol{n}_0, \boldsymbol{x}_0)\boldsymbol{x} + c(\boldsymbol{n}_0, \boldsymbol{x}_0) \tag{3.26}$$

ここで，$\boldsymbol{n}_0, \boldsymbol{x}_0$ は展開の中心であり，c は $\boldsymbol{n}, \boldsymbol{x}$ によらない定数である．これを利用して，雑音付加音声特徴量の確率分布を GMM として以下のように近似する．

$$p(\boldsymbol{y}) = \sum_s p(s)\mathcal{N}(\boldsymbol{y}; \boldsymbol{\mu}_s^y, \boldsymbol{\Sigma}_s^y), \tag{3.27}$$

$$\text{where}, \boldsymbol{\mu}_s^y \simeq \boldsymbol{\mu}_s^x + g(\boldsymbol{\mu}^n, \boldsymbol{\mu}_s^x) \tag{3.28}$$

$$\boldsymbol{\Sigma}_s^y \simeq \frac{\partial}{\partial \boldsymbol{n}} g(\boldsymbol{n}_0, \boldsymbol{x}_0)\boldsymbol{\Sigma}^n \frac{\partial}{\partial \boldsymbol{n}} g(\boldsymbol{n}_0, \boldsymbol{x}_0)^\mathsf{T} + \frac{\partial}{\partial \boldsymbol{x}} g(\boldsymbol{n}_0, \boldsymbol{x}_0)\boldsymbol{\Sigma}_s^x \frac{\partial}{\partial \boldsymbol{x}} g(\boldsymbol{n}_0, \boldsymbol{x}_0)^\mathsf{T} \tag{3.29}$$

3.4.4　特徴量変換

VTS 近似を用いた手法では，学習した雑音付加音声特徴量の GMM を用いて，以下のようにクリーン音声特徴量へと変換する．

$$\hat{\boldsymbol{x}} = \sum_s p(s|\boldsymbol{y}) \left(\boldsymbol{y} - g(\boldsymbol{\mu}^n, \boldsymbol{\mu}_s^x)\right), \tag{3.30}$$

$$\text{where}, p(s|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|s)p(s)}{\sum_i p(\boldsymbol{y}|i)p(i)} \tag{3.31}$$

このとき SPLICE と同様に，一般的に GMM のある一つのコンポーネントの事後確率がほぼ 1 となることから，計算量の削減のため以下のような近似を施しても十分な性能を発揮

する．

$$\hat{\boldsymbol{x}} = \boldsymbol{y} - g(\boldsymbol{\mu}^n, \boldsymbol{\mu}^x_{s^*}), \tag{3.32}$$

$$\text{where}, s^* = \underset{s}{\text{argmax}}\, p(s|\boldsymbol{y}) \tag{3.33}$$

3.4.5　VTS 近似を用いた手法の問題点

VTS 近似を用いて，雑音環境下における特徴量強調を行った場合，その効果は高い．しかし，計算量が膨大になるという問題がある．それは，雑音の推定特徴量が変わる毎に，Σ^y_s を計算する必要があるからである．特徴量が FBANK である場合には，歪関数 (式 3.21) は単純で Σ^y_s は対角行列となり，さほど大きな計算量は必要としない．しかし，音声認識で広く用いられている MFCC を特徴量とした場合には，歪関数が式 (3.22) のように複雑になり，Σ^y_s は全角行列となり，逆行列を計算するには非常に大きな計算量が必要となる．

3.5　まとめ

表 3.1: Comparison among conventional methods.

	SPLICE	J-GMM 法	VTS 近似を用いた手法
GMM でモデル化する音声	雑音付加	雑音付加・クリーン	クリーン
未知雑音環境への適応	不可	不可	可
計算量	小	小	特徴量が MFCC だと膨大

この章では，雑音環境下における特徴量変換について，従来手法を説明してきた．まず，SPLICE について述べた．この手法は雑音付加音声特徴量の確率分布を GMM でモデル化し，雑音付加音声とそのクリーン音声の対であるパラレルデータから GMM の各コンポーネントでの変換関数を学習する．そのことで，雑音付加音声特徴量からクリーン音声への非線形変換を，区分的線形変換により実現する．次に，J-GMM 法について述べた．この手法は雑音付加音とそのクリーン音声のパラレルデータの特徴量の Joint Vector から，Joint GMM を学習し，それを用いて特徴量変換を実現にする．

この２つの手法は，基本的にどのような特徴量を用いたとしても，計算量は変わらないという利点がある．しかし，GMM や変換関数を事前に学習し，それを入力の雑音環境において変化させることがないため，未知の雑音環境下において十分な性能を発揮することが期待されていないという問題がある．

更にこの章では，VTS 近似を用いた手法についても述べた．この手法は，クリーン音声特徴量の確率分布を表す GMM と推定された雑音の特徴量から，VTS 近似により雑音付加音声特徴量の GMM を推定し，それを用いて特徴量変換を施す．この手法は，入力の雑音環境下に応じて適切な変換関数を用いることができるが，雑音特徴量の推定値が変化する

たびに膨大な計算量を必要とし，特に MFCC などの複雑な歪み関数が必要な特徴量では多くの計算量が必要となるという問題がある．

これらの手法の違いをまとめると表 3.1 のようになる．

第4章

主成分分析を用いた高精度化

4.1 はじめに

この章では，前章で述べたSPLICEとJoint GMMを用いた手法の改善方法について述べる．提案手法は，計算量が比較的少なく，どのような特徴量にも導入可能という利点を残した上で，未知の雑音環境にも対応できる手法を目指す．具体的には，変換関数を入力環境に適応できるようにし，また適応に必要な計算量を抑えるために主成分分析 (Principal Component Analysis:PCA) によって適応すべきパラメータ数を削減する．それぞれの手法を今後，Eigen-SPLICE[2]，Eigen JointGMM法と呼ぶこととする．

Eigen-SPLICEは，変換関数のパラメータを連結してSuper Vectorを作成し，それらに対してPCAを施す．そこで得られた主成分とバイアスベクトルを用いて，変換関数を低次元の重みベクトルで適応できるようにする．一方Eigen Joint GMM法は，Joint GMMの雑音付加音声特徴量のパラメータを連結することでSuper Vectorを作成し，それらに対してPCAを施す．そこで得られた主成分行列とバイアスベクトルを用いて，低次元の重みベクトルを推定することによって未知雑音環境下用のJoint GMMを作成する．以下，それぞれの手順について詳しく説明していく．

4.2 Eigen-SPLICE

Eigen–SPLICEは，区分的線形変換のパラメータを入力環境に対して適応することで，未知雑音環境下でも性能を発揮することを目指す．また，少数のデータでも適応出来るように，変換を表すベクトルに主成分分析を施すことで，推定すべきパラメータを削減する．

今回は簡単のため各コンポーネント毎の変換関数のうち，補正ベクトル$\boldsymbol{r}_s$の項のみについて適応した．主成分分析によって推定すべきパラメータを削減する手法は，Eigen–MLLR[20]や固有声に基づいた声質変換 [21] などで広く用いられている．しかし，提案手法は確率分布のパラメータではなく変換関数のパラメータを推定するので，入力音声だけでなく雑音付加音声とクリーン音声のパラレルデータを必要とする点でこれらの手法と異なる．

実際の場面では，未知環境のパラレルデータを得ることは難しい．そこで，提案手法では，雑音付加音声から雑音のみの区間を抽出し，それを手持ちの学習データのクリーン音声に重畳することで擬似的にパラレルデータを作成する．

4.2.1　主成分の学習

まず，従来の SPLICE を用いて学習データの中の全ての環境に共通な変換関数の行列項として $\boldsymbol{A}_s^0$ を学習する．次に，学習データの中の特定の種類・SNR の雑音環境での変換関数の補正ベクトル項 $\boldsymbol{r}_s^i$ を次式のように求める．ただし，このとき添字 i は特定の種類・SNR の雑音環境を表すインデックスである．

$$\hat{\boldsymbol{r}_s^i} = \underset{\boldsymbol{r}_s^i}{\operatorname{argmin}} \sum_t \sum_s p(s|\boldsymbol{y}_t^i) \left\{ \boldsymbol{x}_t^i - (\boldsymbol{A}_s^0 \boldsymbol{y}_t^i + \boldsymbol{r}_s^i) \right\}^2 \tag{4.1}$$

こうすることで学習データ中の特定の環境のための変換パレメータの $\boldsymbol{r}_s^i$ を得る．ただし，行列 $\boldsymbol{A}_s^0$ に関しては全ての環境で共通である．

次に，GMM の各コンポーネントの補正関数を全て連結することによって，特定の環境の変換関数を表すスーパーベクトル $\boldsymbol{SV}^i$ を得る．ただし，S は GMM の混合数である．また，添字 $^\mathsf{T}$ は行列の転置を表す．

$$\boldsymbol{SV}^i = \{\hat{\boldsymbol{r}_1^i}^\mathsf{T}, \cdots, \hat{\boldsymbol{r}_s^i}^\mathsf{T}, \cdots, \hat{\boldsymbol{r}_S^i}^\mathsf{T}\} \tag{4.2}$$

このスーパーベクトルは，学習データの中の全ての環境に関して各々学習する．そして，得られた複数のスーパーベクトルに対して主成分分析を施す．学習データの中の全ての環境の変換関数のスーパーベクトルの平均を表すバイアスベクトル $\boldsymbol{BV}$ と，その主成分を表すベクトル $\boldsymbol{PC}^m$ を得る．ただし，m は主成分のインデックスである．

$$\begin{aligned}
\boldsymbol{BV} &= \{\boldsymbol{b}_1, \cdots, \boldsymbol{b}_s, \cdots, \boldsymbol{b}_S\} && (4.3)\\
\boldsymbol{PC}^1 &= \{\boldsymbol{c}_1^1, \cdots, \boldsymbol{c}_s^1, \cdots, \boldsymbol{c}_S^1\} \\
&\vdots \\
\boldsymbol{PC}^m &= \{\boldsymbol{c}_1^m, \cdots, \boldsymbol{c}_s^m, \cdots, \boldsymbol{c}_S^m\} && (4.4)\\
&\vdots \\
\boldsymbol{PC}^M &= \{\boldsymbol{c}_1^M, \cdots, \boldsymbol{c}_s^M, \cdots, \boldsymbol{c}_S^M\} && (4.5)
\end{aligned}$$

これらのバイアスベクトルと主成分を用いて，ある環境での変換関数は以下のように表せる．

$$\begin{aligned}
\hat{\boldsymbol{x}}_t &= \sum_s p(s|\boldsymbol{y}_t) \left(\boldsymbol{A}_s^0 \boldsymbol{y}_t + \boldsymbol{B}_s \boldsymbol{w} + \boldsymbol{b}_s \right), \\
\text{where, } \boldsymbol{B}_s &= \{\boldsymbol{c}_s^1, \cdots, \boldsymbol{c}_s^M\} && (4.6)
\end{aligned}$$

ただし，ここで添字 $\boldsymbol{w}$ は主成分の重み付けを表す．

4.2.2　重みの推定

以上の操作によって，新たな雑音環境が現れたときに，適応すべき変換関数のパラメータが高次元の補正ベクトルから，低次元の重みベクトルになった．次にこの重み $\boldsymbol{w}$ の推定について述べる．この重みベクトルは，少数の未知環境下における雑音付加音声とクリーン音声のパラレルデータを用いて，最小誤差基準で下記のように推定される．

$$\hat{\boldsymbol{w}} = \underset{w}{\operatorname{argmin}} \sum_t \{\boldsymbol{x}_t - \sum_s p(s|\boldsymbol{y}_t)\,(\boldsymbol{A}_s \boldsymbol{y}_t + \boldsymbol{B}_s \boldsymbol{w} + \boldsymbol{b}_s)\}^2 \tag{4.7}$$

これを解くと，以下のような重み付けになる．

$$\begin{aligned} \hat{\boldsymbol{w}} &= \left(\sum_t \boldsymbol{M}_t^{\mathsf{T}} \boldsymbol{M}_t\right)^{-1} \left(\sum_t \boldsymbol{M}_t^{\mathsf{T}} \boldsymbol{E}_t\right), \\ \text{where, } \boldsymbol{M}_t &= \sum_s p(s|\boldsymbol{y}_t) \boldsymbol{B}_s \\ \boldsymbol{E}_t &= \boldsymbol{x}_t - \sum_s p(s|\boldsymbol{y}_t)\,(\boldsymbol{A}_s \boldsymbol{y}_t + \boldsymbol{b}_s,) \end{aligned} \tag{4.8}$$

以上の手順によって，従来の SPLICE を少数の適応データを用いて，重みを推定することで未知の雑音環境下における変換関数を得られるようになった．通常の SPLICE との比較を図 4.2.2 に示す．

4.2.3　擬似パラレルデータ

通常の雑音除去において，未知の雑音環境下において雑音付加音声とクリーン音声のパラレルデータを得ることは難しい．しかし未知の雑音付加音声は，学習用のクリーン音声に未知雑音を付加することで擬似的に作成可能である．具体的には，未知の雑音付加音声が得られたときに，雑音のみの区間を取り出して，それを学習データ中のクリーン音声に付加することで，擬似パラレルデータを得る．そして，この擬似パラレルデータを用いて未知環境における主成分の重み付けを推定する．

4.3　Eigen Joint GMM 法

前章で述べた既存手法の J-GMM 法を，主成分分析を用いて変換方法を適応できるように改善する手法について説明する．これは多対一の声質変換を可能にする固有声に基づいた声質変換手法 (Eigenvoice Conversion: EVC)[21] を雑音環境下の音声特徴量強調に導入したものである．

この手法は，まず全学習データを使って特定の雑音環境に依存しない汎用的な Joint GMM を学習し，次に特定の雑音環境のデータのみを用いて，その環境に適するような Joint GMM へとパラメータを再学習する．その後，再学習したパラメータを連結し，雑音の種類数と同じだけの Super Vector を作成する．得られた Super Vector に主成分分析を施し，高次元

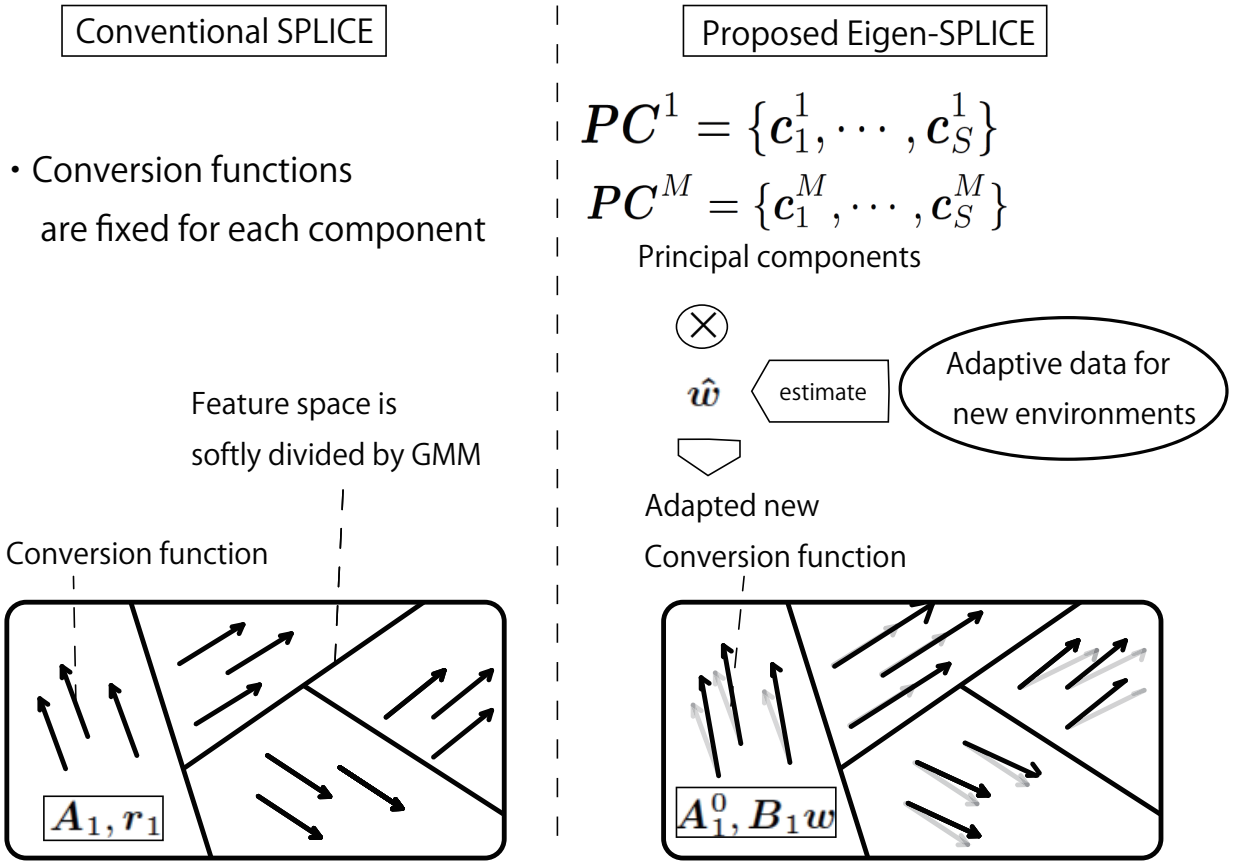

図 4.1: Overview of Eigen-SPLICE

の Joint GMM のパラメータを低次元の重みベクトルで表現できるようにする．その結果，未知の雑音環境下であったとしても，低次元の重みベクトルを推定するだけで Joint GMM を適応することが可能になる．この Eigen Joint GMM を用いた特徴量変換手法を，以下 E-J-GMM 法と記述し，その手順について詳しく説明していく．

4.3.1　学習

まず，学習データ中の全種類の雑音環境のパラレルデータを用いて，環境非依存の Joint GMM，$\boldsymbol{\lambda}^0$ のパラメータを学習する．その後，雑音のタイプや SNR の異なる雑音環境毎の学習データを用いて，$\boldsymbol{\lambda}^0$ を環境 e に適した $\boldsymbol{\lambda}^e$ へ再学習する．この際，平均ベクトル $\boldsymbol{\mu}^y$ のパラメータのみを再学習する．次に，環境依存 $\boldsymbol{\lambda}^e$ の平均ベクトル $\boldsymbol{\mu}_s^{y^e}$ を連結し，以下のように Super Vector $\boldsymbol{SV}^e$ を作成する．

$$\boldsymbol{SV}^e = \left\{ \boldsymbol{\mu}_1^{y^e}, \cdots, \boldsymbol{\mu}_s^{y^e}, \cdots, \boldsymbol{\mu}_S^{y^e} \right\}. \tag{4.9}$$

得られた複数の Super Vectors に対して主成分分析を施すことで，Super Vector の主成分行列 $\boldsymbol{PC}^m$ とバイアスベクトル $\boldsymbol{BV}$ を得る．

$$\begin{aligned} \boldsymbol{BV} &= \{\boldsymbol{b}_1, \cdots, \boldsymbol{b}_s, \cdots, \boldsymbol{b}_S\}, & (4.10) \\ \boldsymbol{PC}^1 &= \{\boldsymbol{c}_1^1, \cdots, \boldsymbol{c}_s^1, \cdots, \boldsymbol{c}_S^1\}, \\ &\vdots \\ \boldsymbol{PC}^M &= \{\boldsymbol{c}_1^M, \cdots, \boldsymbol{c}_s^M, \cdots, \boldsymbol{c}_S^M\}. & (4.11) \end{aligned}$$

ただし，m は主成分のインデックスであり，M は用いる主成分の数である．

4.3.2　変換

これら $\boldsymbol{BV}, \boldsymbol{PC}$ を用いて，ある環境 e に適する Joint GMM の平均ベクトルは，重みベクトル $\boldsymbol{w}$ によって下記のように記述できる．

$$\begin{aligned} \boldsymbol{\mu}_s^e &= \boldsymbol{B}_s \boldsymbol{w} + \boldsymbol{b}_s, \\ \text{where}, \boldsymbol{B}_s &= \{\boldsymbol{c}_s^1, \cdots, \boldsymbol{c}_s^M\} \end{aligned} \tag{4.12}$$

つまり，この Joint GMM を用いた変換は下記のようになる．

$$\hat{\boldsymbol{x}}_t = \sum_s p(s|\boldsymbol{y}_t, \boldsymbol{\lambda}^e) \left\{ \boldsymbol{\mu}_s^x + \boldsymbol{\Sigma}_s^{xy} \boldsymbol{\Sigma}_s^{yy^{-1}} \left(\boldsymbol{y}_t - \boldsymbol{B}_s \boldsymbol{w} - \boldsymbol{b}_s \right) \right\}. \tag{4.13}$$

4.3.3　入力環境への適応

次に，入力雑音環境に適する重みベクトル $\boldsymbol{w}$ を決定する手順について説明する．推定では，以下の最尤推定を行う．

$$\hat{\boldsymbol{w}} = \underset{\boldsymbol{w}}{\operatorname{argmax}} \int p(\boldsymbol{x}, \boldsymbol{y}^{(\mathrm{IN})}|\boldsymbol{\lambda}^e)d\boldsymbol{x}, \tag{4.14}$$

$$= \underset{\boldsymbol{w}}{\operatorname{argmax}}\, p(\boldsymbol{y}^{(\mathrm{IN})}|\boldsymbol{\lambda}^e). \tag{4.15}$$

ただし，$\boldsymbol{y}^{(\mathrm{IN})}$ は入力未知環境の雑音付加音声特徴量の系列を表す．このとき以下の関数を繰り返し最大化することによって，教師なしで重みベクトルを推定することができる．

$$Q(\boldsymbol{w}, \boldsymbol{w}^{\mathrm{new}}) = \sum_{t=1}^{T}\sum_{s=1}^{S} p(s|\boldsymbol{y}_t^{(\mathrm{IN})}, \boldsymbol{\lambda}^e) \log p(s|\boldsymbol{y}_t^{(\mathrm{IN})}, \boldsymbol{\lambda}^{e\,\mathrm{new}}) \tag{4.16}$$

更新式は以下のようになる．

$$\boldsymbol{w}^{\mathrm{new}} = \left\{ \sum_{s=1}^{S} \overline{\gamma_s^{(\mathrm{IN})}} \boldsymbol{B}_s^{\mathsf{T}} \boldsymbol{\Sigma}_s^{yy^{-1}} \boldsymbol{B}_s \right\}^{-1} \sum_{s=1}^{S} \boldsymbol{B}_s^{\mathsf{T}} \boldsymbol{\Sigma}_s^{yy^{-1}} \overline{\boldsymbol{y}_s^{(\mathrm{IN})}} \tag{4.17}$$

ただし，

$$\overline{\gamma_s^{(\mathrm{IN})}} = \sum_{t=1}^{T} p(m_s|\boldsymbol{y}_t^{(\mathrm{IN})}, \boldsymbol{\lambda}^e) \tag{4.18}$$

$$\overline{\boldsymbol{y}_s^{(\mathrm{IN})}} = \sum_{t=1}^{T} p(m_s|\boldsymbol{y}_t^{(\mathrm{IN})}, \boldsymbol{\lambda}^e) \left(\boldsymbol{y}_t^{(\mathrm{IN})} - \boldsymbol{b}_s \right) \tag{4.19}$$

なお，最初の事後確率を求める際には，環境非依存の $\boldsymbol{\lambda}^0$ のパラメータを用いる．E–J–GMM 法の概要を図 4.3.3 に示す．

4.4　まとめ

以上のように，この章では既存手法の主成分分析を用いた改善について 2 つの手法を説明してきた．Eigen-SPLICE は，変換関数を表す Supere Vectors に対して主成分分析を施

表 4.1: Comparison among proposed methods.

	Eigen-SPLICE	Eigen J-GMM 法
GMM の学習に用いる特徴量	雑音付加音声	クリーン音声と雑音付加音声
適応対象	変換関数	Joint GMM
適応に必要なデータ	擬似パラレルデータ	入力音声のみ
重みベクトルの計算	解析的に計算	繰り返し計算

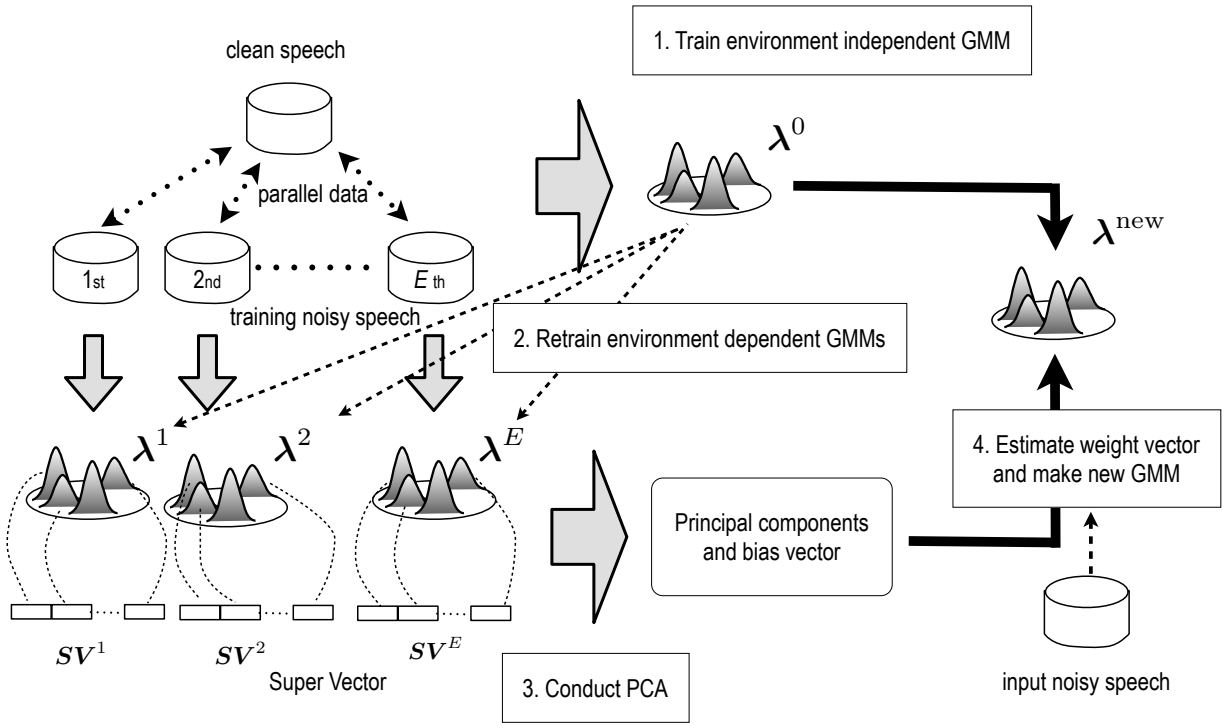

図 4.2: Overview of Eigen Joint GMM method

すことで，少量の適応データで未知の雑音環境下においても変換関数を適応できる手法である．一方 Eigen Joint GMM 法は，声質変換の分野で既に提案されている固有声に基づいた手法を，雑音環境下における音声特徴量強調に導入したものである．E-J-GMM 法も PCA を用いることで，少量の適応データで未知の雑音環境に Joint GMM のパラメータを適応できるようになった．

この２つの手法は，非常によく似ているが幾つかの点で異なる (表 4.1)．まず GMM でモデル化する特徴量が，Eigen-SPLICE が雑音付加音声の特徴量であるのに対して，E-J-GMM 法は雑音付加音声とクリーン音声の特徴量の結合ベクトルである．そのため，同じ学習データを用いて，同じ混合数の GMM を学習したとしても，その GMM による確率的な特徴量空間の分割の様子も変わってくる．

また次に，適応するものが Eigen-SPLICE は変換関数の補正ベクトルであるのに対して，E-J-GMM 法は Joint GMM の雑音付加音声特徴量の部分の平均ベクトルであるという点も異なる．そのため，Eigen-SPLICE が適応の際にパラレルデータを必要とし，擬似パラレルデータを作成するのに対して，E-J-GMM 法が入力の雑音付加音声の特徴量のみだけで計算できるという違いを生む．同様に，適応の際の重み計算も，Eigen-SPLICE が適応データに対して解析解が求まるのに対して，E-J-GMM 法は繰り返し計算によって求める必要があるという違いも生む．

第5章

実験的検証

5.1 はじめに

この章では，提案した手法の有効性を実験的に検証する．まずはじめに，Eigen-SPLICE や Eigen Joint GMM 法の幾つかのパラメータを決定するために，予備実験を行った．その後，従来手法との性能を比較するために，雑音環境下における音声認識精度で特徴量変換の性能比較を行った．

5.1.1 雑音環境下音声認識データベース AURORA-2

この一連の実験的検証は実験では，雑音環境下における音声認識用データベース AURORA-2[22] を用いて行った．このデータベースについて，簡単に説明していく．

AURORA-2 データベースは，雑音環境下における連続数字音声認識タスクであり，大きく学習セットと評価セットに分かれている．通常，学習セットを学習データとして，認識用の音響モデルや変換方法を学習する．学習セットには，成人男性 55 名，成人女性 55 名による合計 8440 発声のクリーン音声が収録されている．またそれらを 20 環境，4 タイプ (Subw., Babble, Car, Exhibit) × 5SNR(5, 10, 15, 20, ∞ [dB])，それぞれ 422 発声に分割し，加法性雑音を重畳されたものがある．ただし，∞ [dB] はつまりクリーン音声である．

一方，評価セットは学習したものの検証のために用いられ，A セット，B セット，C セットの 3 つに分かれている．ABC の各セットには，成人男性 52 人，成人女性 52 人による合計 4004 発声文の音声があり，それを 4 つに分割した 1001 発声が基本単位となっている．その 4 つの単位に対して異なる種類の雑音を 7SNR(-5, 0, 5, 10, 15, 20, ∞ [dB]) で重畳されている．ただし，今回の一連の評価実験においては極端な SNR のものを除いて，5SNR(0, 5, 10, 20[dB]) の平均で評価している．

A セットには，学習データと同じ 4 種類の雑音環境が重畳されている．つまり，4 タイプ (Subw., Babble, Car, Exhibit) × 7SNR(-5, 0, 5, 10, 15, 20, ∞ [dB]) 計 28 タイプの雑音環境がある．したがって，A セットでは基本的に雑音の種類に対してクローズドな実験が行えるようになっている．

表 5.1: AURORA2 data sets.

	学習	評価		
		A set	B set	C set
雑音	加法性	加法性	加法性	加法性・乗法性
SNR 数	5	7	7	7
備考		既知環境	未知環境	

B セットには，学習データとは異なる 4 種類の雑音環境が重畳されている．その 4 種類は (Rest., Street, Airport, Sta.) である．これが，A セットと同様に 28 タイプの雑音環境でそれぞれ 1001 発声分ある．したがって，B セットは学習データに対して未知雑音下における実験が行える．

A,B セットが加法性雑音を重畳したものであるのに対して，C セットは加法性雑音を重畳したものに，更に電話通信を想定したフィルタを通して乗法性雑音を加えたものである．

5.2　予備実験：Eigen-SPLICE

この節では，まず Eigen-SPLICE において用いる主成分の数と擬似パラレルデータの量を決定するために行った予備実験について説明する．

5.2.1　実験条件

まず，学習セット 4 タイプ (Subw., Babble, Car, Exhibit) × 4SNR(5,10,15,20[dB]) 計 16 雑音環境を使って，雑音環境非依存の変換関数 $\boldsymbol{A}_s^0$ を学習する．次に $\boldsymbol{A}_s^0$ を用いて，16 雑音環境毎の $\boldsymbol{r}_s^i$ を学習する．その後，$\boldsymbol{r}_s$ スーパーベクトルを構成し，主成分分析を施し，主成分とバイアスベクトルを得る．

今回は擬似パラレルデータを作る際に，入力雑音付加音声の始端と末端 250[ms] は雑音のみの区間という粗い仮定をして，雑音区間を切り出した．この切り出した雑音区間を，学習データの中からランダムに選び出したクリーン音声に繰り返し重畳することで，入力発声毎に変換関数を適応する．特徴量としては MFCC13 次元に Δ と $\Delta\,\Delta$ を加えた 39 次元 (HTK[23] での MFCC_D_A_0)，雑音付加音声の特徴量の分布を表現する GMM の混合数は 64 混合とした．認識用の HMM は学習セットのクリーン音声のみから学習し，1 単語あたり 18 状態，1 状態あたり 20 混合の GMM を持つ単語 HMM を用いた．

5.2.2　実験結果

以上の条件下で，用いる主成分の数を 4, 6, 8 と変化させ，また適応に用いる擬似パラレルデータの量を 1, 2, 4, 8, 16, 32 と変化させ，そのとき未知雑音環境下である評価 B セットの雑音のタイプ Rest. SNR 5[dB](N1 SNR5 セット) での認識精度の変化を見た．その様子が図 5.1 のようである．

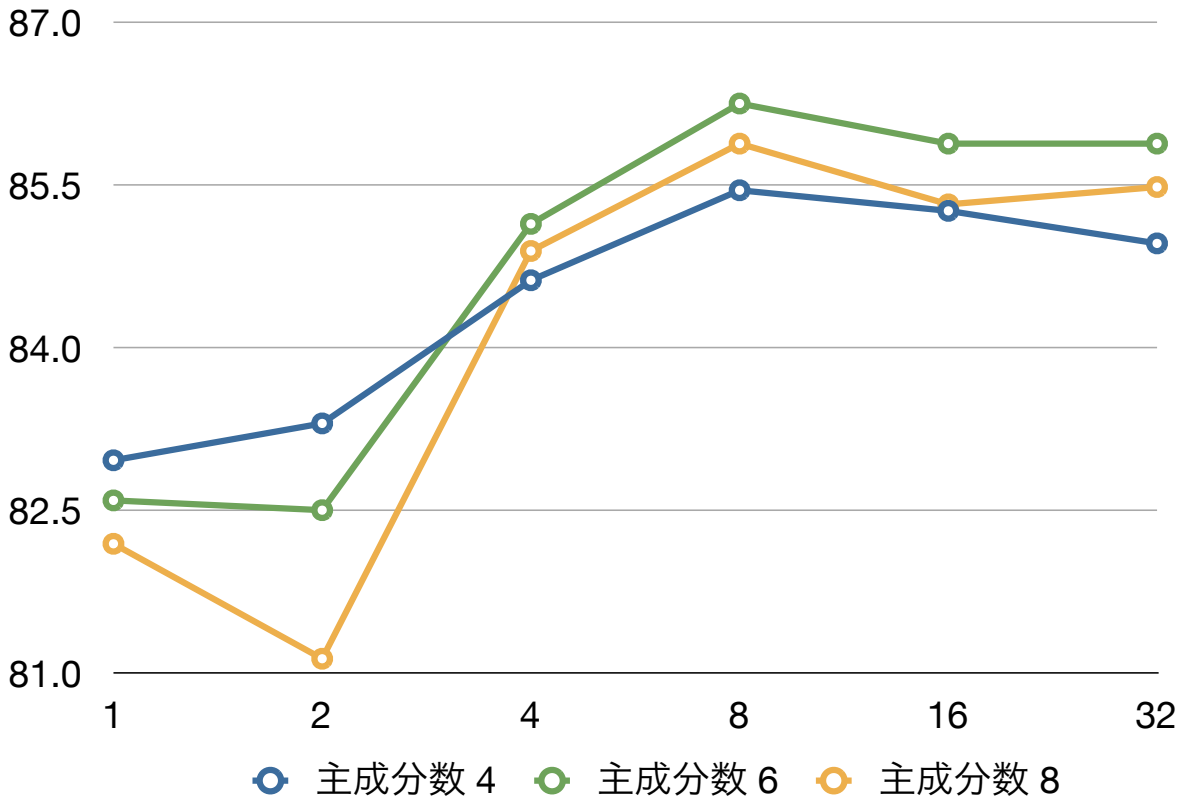

図 5.1: Word accuracy in test B N1 SNR 5 set with Eigen SPLICE.

結果を見てみると，主成分数は 6 のとき最も良い性能を発揮しており，また用いる適応データ量は 8 発声分の時に高い性能を発揮し，それ以上は安定する．これにより以下，Eigen-SPLICE では用いる主成分の数は 6，適応に用いるデータは 8 発声程度とする．主成分数が 6 で最も高い性能を発揮するのは，用いる主成分数が少な過ぎると，それらの重み付け足し合わせだけでは十分に未知の雑音環境に対応出来ないが，逆に用いる主成分の数が多すぎても，未知の雑音環境に対応するのに必要ない成分まで混じってきてしまうためだと考えられる．

5.3　予備実験：Eigen Joint GMM 法

次にこの節では，Eigen Joint GMM 法において用いる主成分数を決めるために行った予備実験について説明する．

5.3.1　実験条件

実験手順は基本的に Eigen-SPLICE の予備実験を行ったときとほぼ同じであるが，計算機の都合で用いる特徴量や GMM の混合数などが異なる．まず AURORA–2 の学習セット

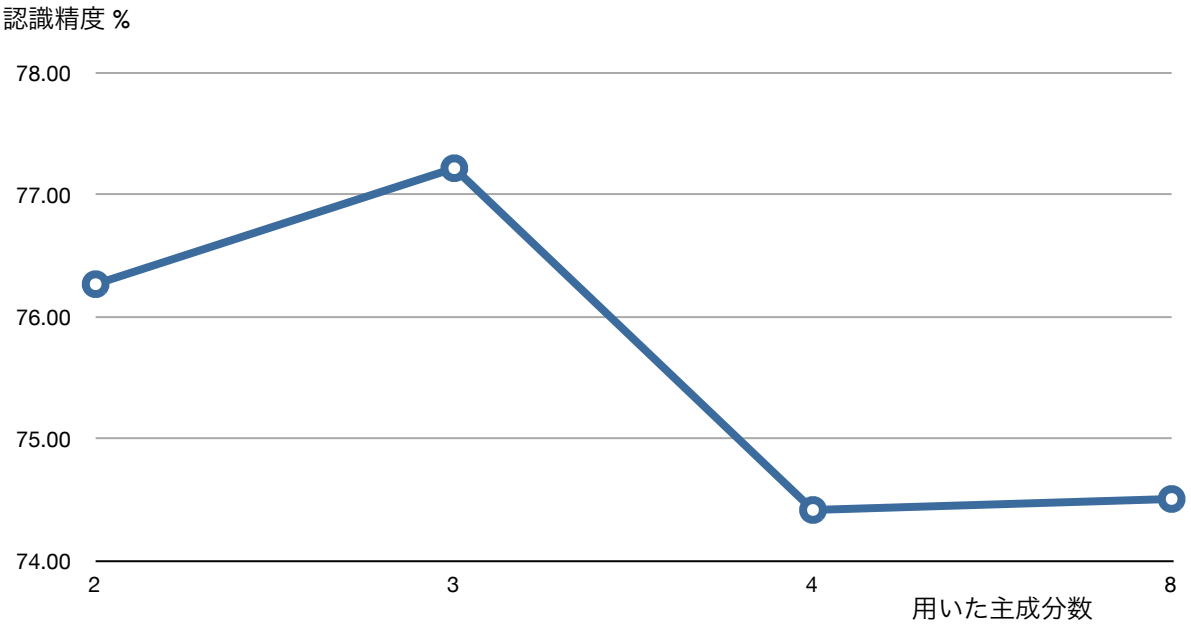

図 5.2: Word accuracy in test B N1 SNR 5 set with Eigen Joint GMM.

の 4 タイプ (Subw., Babble, Car, Exhibit) × 4SNR(5, 10, 15, 20[dB]) 計 16 雑音環境を使って，雑音環境非依存の Joint GMM$\boldsymbol{\lambda}^0$ を学習する．次に，それぞれの雑音環境のデータを用いて再学習を行い，雑音環境毎の $\boldsymbol{\lambda}^e$ を得る．その後，$\boldsymbol{\mu}_s^y$ のスーパーベクトルを構成し，主成分分析を施し，主成分とバイアスベクトルを得る．

今回は，入力一発声毎に 5 回更新を繰り返すことで重みを推定した．計算機の都合上，まず MFCC13 次元に対して変換を行い，変換後の特徴量から Δ と ΔΔ を計算し，MFCC+Δ+ΔΔ の 39 次元として認識した．認識用の音響モデルは前回同様，クリーン音声のみから学習し，1 単語あたり 18 状態，1 状態あたり 20 混合の GMM を持つ単語 HMM を用いた．また，GMM の混合数は 256 混合とした．

5.3.2　実験結果

以上の条件下で，用いる主成分の数を 2,3,4,8 と変化させ，そのとき未知雑音環境下である評価 B セットの雑音のタイプ Rest. SNR 5[dB] での認識精度の変化を見た．その様子が図 5.2 のようである．

結果を見てみると，E-J-GMM 法の場合は主成分数が 3 のときに最も良い性能を発揮することが分かる．これにより，以下 E-J-GMM 法では 3 つの主成分を用いることにする．

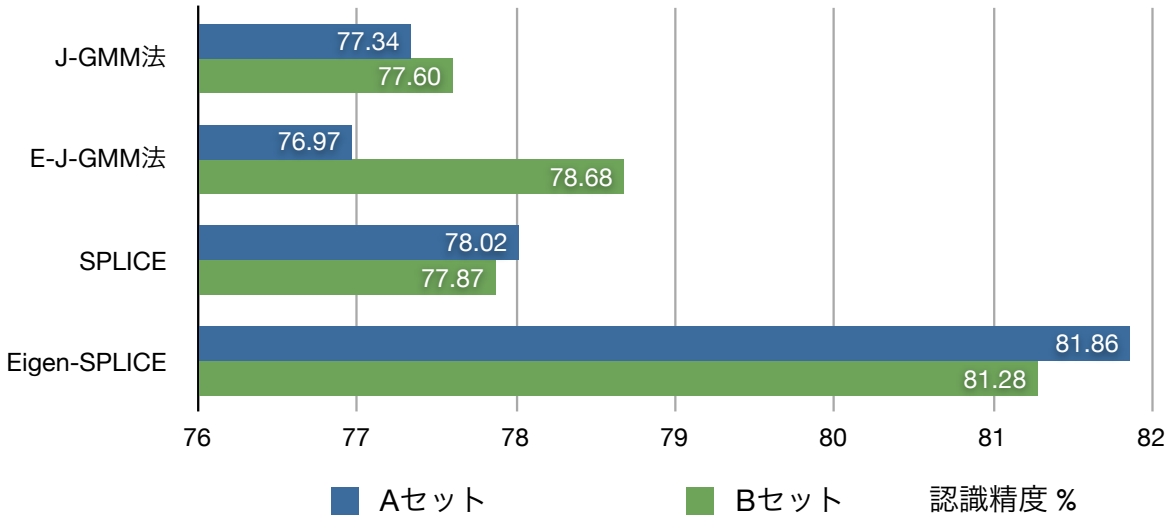

図 5.3: Word accuracy with proposed methods and conventional methods.

5.4　Eigen-SPLICE と Eigen Joint GMM 法の比較

この節では，Eigen-SPLICE と Eigen Joint GMM 法の有効性の確認及び性能比較のため行った認識実験について説明する．

5.4.1　実験条件

まず，Eigen-SPLICE と Eigen Joint GMM 法も予備実験と同じ手順で学習を行う．また，従来手法とも性能を検討するために，SPLICE や Joint GMM 法も同様の学習データを用いて，学習する．ただし，今回は手法の評価のために，GMM の混合数，用いる特徴量の次元数は計算機による制限のある JointGMM 法に合わせて，256 混合，MFCC13 次元で統一した．また予備実験と同様にまず MFCC13 次元に対して変換を行い，変換後の特徴量から Δ と $\Delta\Delta$ を計算し，MFCC+Δ+$\Delta\Delta$ の 39 次元としてクリーン条件で学習した HMM で認識した．認識には，評価 A セット，B セットを用いて SNR が 0[dB] から 20[dB] の認識精度の平均を用いて評価した．

5.4.2　結果と考察

認識結果を図 5.3 に示す．ただしこのとき，特徴量変換を施さないもの場合の Baseline の結果は，A セットで 51.45%，B セットで 44.86%である．また詳細の結果についても表 5.2 に示す．

結果を見ると，単純な J–GMM 法であったとしても評価 B セットで 77.60%の精度を示しており，やや劣るものの従来の SPLICE と同程度の性能を発揮している．つまり，声質

変換の分野で用いられていた Joint GMM の手法が雑音環境下における音声特徴量強調にも導入できるということが言える．また，E–J–GMM 法は評価 A セットにおいて J–GMM 法と比較して 1.6%の誤り増加が見られたものの，未知の雑音環境である評価 B セットにおいては相対的に 4.8%の誤り削減が見られた．しかし，Eigen–SPLICE が評価 B セットにおいて SPLICE と比較して 15.4%の誤りの削減をしていることを考慮すると，改善の効果は限定的である．

Eigen–SPLICE と違い E–J–GMM 法では雑音付加音声とクリーン音声とが GMM のインデックスを共有しているため，事後確率 $p(s|\boldsymbol{y}_t, \boldsymbol{\lambda}^e)$ を求める際に必要な GMM が，雑音の特性に基づいて学習されていないことが原因であると考えている．

以上より，混合数が同じであるような状況下では，Eigen Joint GMM 法よりも Eigen-SPLICE の方が高い性能を示すということが分かった．

5.5　Eigen-SPLICE と従来手法との性能比較

次に，Eigen-SPLICE の混合数を増やした条件下で，SPLICE やその既存の改善法である Environment Model Selction SPLICE(EMS SPLICE) との性能比較を行った．

5.5.1　実験条件

基本的に，SPLICE と Eigen-SPLICE に関しては，以前の実験と同じ手順で GMM や変換関数，および主成分などを求めた．ただし，今回は雑音付加音声の特徴量の確率分布をモデル化する GMM の混合数を 512 としている．また，用いる特徴量に関しては MFCC+Δ+ΔΔ の 39 次元を用いた．つまり，以前の実験のように MFCC13 次元を特徴量変換した後の特徴量から Δ+ΔΔ を計算するのではなく，39 次元の特徴量をそのまま特徴量変換して，それを認識に用いた．

EMS SPLICE も同様に，MFCC+Δ+ΔΔ 39 次元で特徴量変換をした後に，それを用いて認識する．ただし，EMS SPLICE の各環境毎の GMM の混合数は予備実験で求めた 128 であり，学習セットの中の 4 タイプ (Subw., Babble, Car, Exhibit) × 4SNR(5, 10, 15, 20[dB]) 計 16 環境分の GMM およびそれに対応する変換関数を用意した．

5.5.2　実験結果

実験結果は図 5.4 のようである．また，詳細の結果についても表 5.3 に示す．

結果を見てみると，学習データと同じ雑音環境を含む A セットに関しては，事前に学習した各環境毎の変換関数と GMM を切り替えながら変換する EMS SPLICE が最も高い性能を示している．しかし，EMS SPLICE は未知の雑音環境下である B セットにおいては，逆に最も性能が低くなってしまっている．

一方，Eigen-SPLICE は，変換関数を発声毎に適応していくので，未知の雑音環境下の B セットにおいても，A セットと同様に高い性能を発揮している．A セットにおける性能は，

表5.2: Word recognition accuracies [%] (a) without enhancement, baseline, (b) Joint GMM, (c) Eigen Joint GMM (d) SPLICE (e) Eigen-SLICE.

			A set					B set		
(a)	Subw.	Babble	Car	Exhibit	Avg.	Rest.	Street	Airport	Station	Avg.
20dB	97.06	85.12	94.26	97.12	93.39	90.16	94.63	86.88	88.18	89.96
15dB	88.91	64.41	78.44	90.79	80.64	71.31	84.59	66.39	69.01	72.83
10dB	67.31	33.90	49.61	71.81	55.66	44.77	59.90	40.42	42.77	46.97
5dB	31.05	1.58	21.45	39.62	23.42	10.51	31.08	11.08	15.82	17.12
0dB	12.62	-19.8	9.86	13.84	4.13	-15.2	11.24	-6.26	0.00	-2.55
Avg.	59.39	33.04	50.72	62.64	51.45	40.31	56.29	39.70	43.16	44.86
(b)	Subw.	Babble	Car	Exhibit	Avg.	Rest.	Street	Airport	Station	Avg.
20dB	96.81	98.07	97.91	97.13	97.48	98.07	96.70	98.33	98.09	97.80
15dB	93.40	96.28	95.17	92.59	94.36	96.53	93.26	95.94	95.90	95.41
10dB	86.80	91.11	87.35	83.92	87.30	89.75	83.10	90.49	89.39	88.18
5dB	72.92	73.25	65.40	64.76	69.08	75.04	59.22	71.37	72.69	69.58
0dB	47.25	37.06	32.66	37.06	38.51	42.46	28.93	37.76	39.06	37.05
Avg.	79.44	79.15	75.70	75.09	77.34	80.37	72.24	78.78	79.03	77.60
(c)	Subw.	Babble	Car	Exhibit	Avg.	Rest.	Street	Airport	Station	Avg.
20dB	97.91	98.55	97.97	98.77	98.30	98.56	97.40	97.55	98.18	97.92
15dB	96.13	96.80	95.82	96.30	96.26	97.67	95.92	95.38	96.39	96.34
10dB	90.97	90.51	84.73	91.33	89.39	91.19	89.33	90.01	89.17	89.92
5dB	73.13	69.53	47.09	79.08	67.21	77.22	66.87	72.83	67.42	71.09
0dB	35.25	36.85	16.94	45.79	33.71	46.70	32.32	41.46	31.97	38.11
Avg.	78.68	78.45	68.51	82.25	76.97	82.27	76.37	79.45	76.63	78.68
(d)	Subw.	Babble	Car	Exhibit	Avg.	Rest.	Street	Airport	Station	Avg.
20dB	97.21	98.31	97.91	97.59	97.75	98.43	96.49	98.30	98.18	97.85
15dB	94.75	96.67	96.24	94.42	95.52	96.81	94.80	97.23	96.39	96.31
10dB	87.75	92.38	90.16	86.70	89.25	91.13	86.52	92.25	90.44	90.08
5dB	73.35	74.85	69.67	65.57	70.86	74.85	63.75	71.40	70.50	70.12
0dB	45.81	32.77	34.83	33.45	36.72	37.73	32.56	34.72	34.90	34.98
Avg.	79.77	79.00	77.76	75.55	78.02	79.79	74.82	78.78	78.08	77.87
(e)	Subw.	Babble	Car	Exhibit	Avg.	Rest.	Street	Airport	Station	Avg.
20dB	98.16	98.82	98.69	98.43	98.52	98.83	97.31	98.99	98.77	98.47
15dB	95.98	97.19	96.96	96.02	96.54	97.85	95.71	98.12	96.64	97.08
10dB	91.71	92.38	91.47	91.14	91.67	92.72	89.00	94.60	90.87	91.80
5dB	79.31	77.33	74.53	76.80	76.99	79.06	71.64	79.57	74.45	76.18
0dB	50.20	40.96	42.65	48.53	45.59	45.62	39.21	47.06	39.52	42.85
Avg.	83.07	81.34	80.86	82.18	81.86	82.82	78.57	83.67	80.05	81.28

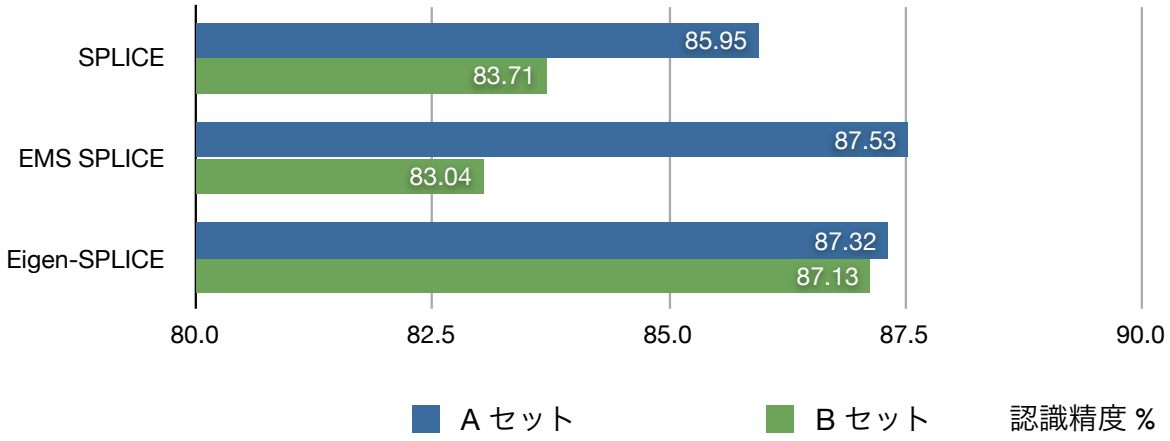

図 5.4: Word accuracy with proposed SPLICE and conventional SPLICE.

やや EMS SPLICE に劣るものの，A セット，B セット両者を考えた場合，Eigen-SPLICE は非常に有効な手法であると言える．

5.6　まとめ

この章では，提案手法の有効性を確認するために，雑音環境下における音声認識データベース AURORA-2 を用いて音声認識実験を行った．

まず，AURORA-2 の概要について説明した．AURORA-2 は，学習セットとテストセットに分かれており，テストセットは更に A セット, B セット, C セットに別れており，それぞれ既知加法性雑音環境，未知加法性雑音環境，既知および未知加法性乗法性雑音環境の音声を含む．Eigen-SPLICE は雑音を擬似的にクリーン音声に重畳する作業を必要とするため，原理的に乗法性雑音については対応できない，そのため認識実験では主に A セットおよび B セットを用いた．

次に，Eigen-SPLICE や Eigen Joint GMM 法で必要な主成分数や適応データ数を決定するために，予備実験を行った．この際，未知雑音環境下における改善を見るために，未知の雑音環境の B セットのサブセットを用いて，パラメータを求めた．その結果，Eigen-SPLICE で用いる主成分の数は 6，適応データの量は 8 発声分とした．また，Eigen Joint GMM 法に関しては，用いる主成分の数を 3 とした．

また，提案手法の Eigen-SPLICE 及び Eigen Joint GMM 法の性能を比較するために，計算機によって制限された GMM 混合数 256 および特徴量 MFCC13 次元の条件下で，A セット及び B セットを用いて実験を行った．その結果，この条件下では Eigen-SPLICE の方がより高性能を発揮するということが分かった．この性能差は，事後確率を求める際に必要な GMM が，Eigen-SPLICE の方が適切に特徴量空間を確率的に分割しているためだと考える．

表 5.3: Word recognition accuracies [%] (a) without enhancement, baseline, (b) SPLICE (mix #516), (c) EMS Joint GMM (mix #128)(d)Eigen-SPLICE (mix #516)

			A set					B set		
(a)	Subw.	Babble	Car	Exhibit	Avg.	Rest.	Street	Airport	Station	Avg.
20dB	97.06	85.12	94.26	97.12	93.39	90.16	94.63	86.88	88.18	89.96
15dB	88.91	64.41	78.44	90.79	80.64	71.31	84.59	66.39	69.01	72.83
10dB	67.31	33.90	49.61	71.81	55.66	44.77	59.90	40.42	42.77	46.97
5dB	31.05	1.58	21.45	39.62	23.42	10.51	31.08	11.08	15.82	17.12
0dB	12.62	-19.8	9.86	13.84	4.13	-15.2	11.24	-6.26	0.00	-2.55
Avg.	59.39	33.04	50.72	62.64	51.45	40.31	56.29	39.70	43.16	44.86
(b)	Subw.	Babble	Car	Exhibit	Avg.	Rest.	Street	Airport	Station	Avg.
20dB	98.68	98.97	99.11	98.80	98.89	99.20	98.55	99.02	99.04	98.95
15dB	97.64	98.43	98.21	97.69	97.99	98.68	97.70	98.63	97.90	98.23
10dB	94.87	96.46	96.06	94.14	95.38	95.76	93.53	95.50	94.26	94.76
5dB	87.23	85.19	81.90	82.57	84.22	85.45	76.00	83.72	77.72	80.72
0dB	61.77	51.15	47.54	52.64	53.28	56.16	40.51	50.70	36.07	45.86
Avg.	88.04	86.04	84.56	85.17	85.95	87.05	81.26	85.51	81.00	83.71
(c)	Subw.	Babble	Car	Exhibit	Avg.	Rest.	Street	Airport	Station	Avg.
20dB	98.83	99.06	99.16	97.87	98.73	99.29	98.40	99.08	99.04	98.95
15dB	97.85	98.34	98.36	96.39	97.73	98.43	97.67	98.36	97.72	98.05
10dB	95.09	96.13	96.00	92.72	94.99	95.52	92.62	94.69	94.01	94.21
5dB	89.07	86.06	85.80	83.59	86.13	83.70	75.73	81.36	77.41	79.55
0dB	69.36	51.72	57.53	61.59	60.05	52.01	40.75	47.24	37.70	44.42
Avg.	90.04	86.26	87.37	86.43	87.53	85.79	81.03	84.15	81.18	83.04
(d)	Subw.	Babble	Car	Exhibit	Avg.	Rest.	Street	Airport	Station	Avg.
20dB	98.86	99.09	99.14	98.83	98.98	99.26	98.58	99.02	99.11	98.99
15dB	97.88	98.52	98.54	97.78	98.18	98.80	97.82	99.08	98.33	98.51
10dB	95.43	96.34	96.45	94.72	95.74	96.56	94.23	97.14	95.34	95.82
5dB	87.96	86.00	84.64	86.02	86.15	87.93	82.86	88.16	84.60	85.89
0dB	62.63	55.65	52.43	59.55	57.56	61.93	51.60	62.24	49.98	56.44
Avg.	88.55	87.12	86.24	87.38	87.32	88.90	85.02	89.13	85.47	87.13

最後に，Eigen-SPLICE と既存の SPLICE およびその既存の改善手法 EMS SPLICE の性能を比較した．Eigen-SPLICE および SPLICE の GMM の混合数は 512 とし，EMS SPLICE の GMM は混合数 128 のものを 16 個用意して，それに対応する変換関数も 16 組用意した．A セット，B セットを用いて認識実験を行ったところ，A セットにおいては EMS SPLICE が最も高い性能を発揮したものの，Eigen-SPLICE も A セットでも EMS SPLICE と同程度の性能を発揮し，また未知環境の B セットにおいては Eigen-SPLICE がもっとも高い性能を発揮した．

以上の実験によって，提案手法である Eigen-SPLICE は変換関数を少数のパラメータを適応することによって，未知の雑音環境下においても高い精度で特徴量強調ができるということが示せた．

第6章

結論

6.1 本論文のまとめ

第 1 章では，研究背景および研究目的について述べた．近年の統計的音声認識技術の発展やスマートフォンなどの普及によって雑音環境下における音声認識が広く一般的に用いられるようになってきた．しかし，雑音によって歪められた音声特徴量は，事前にクリーン音声を用いて学習した音響モデルとのミスマッチを生じさせる．これを解決するために様々な手法が提案されているが，本研究では GMM によって雑音付加音声特徴量の分布をモデル化し，それを用いて特徴量変換する SPLICE をベースに行われた．本研究はその SPLICE を PCA を用いて改善することによって，高精度化を目指した．

第 2 章では雑音による影響について，統計的音声認識システムの仕組みと併せて説明した．近年用いられている統計的音声認識システムでは，音声波形を直接認識に用いずに，音声の音韻的な性質をよく表すスペクトルやケプストラムといった特徴量がよく用いられており，その抽出手順について説明した．その中でも特によく用いられているのが MFCC であり，今回の実験でも用いた．

統計的音声認識システムでは，一般的に音響モデルを HMM でモデル化する．この音響モデルは，大量のスクリプト付きの音声データによって学習する必要があり，事前に用意したクリーン音声によって学習される．そのため，背景雑音などの加法性雑音やチャンネル歪などによる乗法性雑音によって歪んだ雑音付加音声特徴量と音響モデルとの間にはミスマッチが生じ，認識精度を著しく低下させる．その雑音による歪みに関しても，モデル化の仕方やその定式化について述べた．

第 3 章では，既存の GMM を用いた特徴量変換手法について述べた．SPLICE は雑音付加音声特徴量の確率分布を GMM モデル化し，それを用いて GMM の各コンポーネント毎の変換関数を学習する．変換時には，雑音付加音声の GMM で事後確率を計算し，その事後確率の重み付きの変換結果の総和をとることによって，所望のクリーン音声特徴量へと変換する．

Joint GMM 法は，雑音付加音声特徴量とクリーン音声の特徴量を連結した Joint Vector を GMM によってモデル化する．変換時には，Joint GMM の雑音付加音声部分のみを用いて事後確率を計算し，Joint GMM の平均や分散共分散を用いて特徴量変換する．

ベクトルテーラー展開 (VTS) を用いた手法は，クリーン音声の特徴量の確率分布を GMM でモデル化し，雑音の推定された特徴量を用いて，雑音付加音声特徴量の GMM を VTS 近似する．その後，近似された雑音付加音声の GMM のパラメータを用いて，特徴量変換する．

従来手法は高い性能を発揮しているが，SPLICE や Joint GMM 法には未知の雑音環境下においては十分な性能を発揮することが保証されていないという問題が，VTS を用いた手法には歪み関数が複雑な MFCC などの特徴量では計算量が膨大になってしまうという問題がある．

第 4 章では，前章で述べた従来手法の問題点を解決する，主成分を用いた高精度化を提案した．基本的な枠組みは，本来高次元である変換関数などのパラメータを Super Vectors 化して，それを主成分分析にかけ低次元化することで，未知の環境に対して少量の適応データで適応可能にするというものである．この枠組みで，Eigen-SPLICE を提案し，固有声に基づいた声質変換を Eigen Joint GMM 法として雑音環境下における特徴量変換に導入した．

これらの提案手法は，基本的な枠組みは同じであるが，幾つかの違いがある．Eigen SPLICE は適応の対象が変換関数であるために，適応データとして擬似パラレルデータを作成する必要があり，その代わり重みベクトルの推定は適応データが与えられると解析的に計算できる．一方，Eigen Joint GMM 法は適応の対象が GMM のパラメータであるため，擬似パラレルデータは必要としないが，重みベクトルの推定には繰り返し計算によって局所最適解を求めることしかできない．

第 5 章では，提案手法の有効性を確認するために，雑音環境下における音声認識データベース AURORA-2 を用いて，従来手法との性能比較を行った．その結果，Eigen-SPLICE と Eigen Joint GMM 法に必要な主成分数などのパラメータが求まり，提案手法の有効性が示された．特に，Eigen-SPLICE の性能改善は大きく，未知雑音環境下においても十分な性能を発揮した．

6.2　今後の課題

6.2.1　Eigen-SPLICE

本研究で有効性が示された Eigen-SPLICE には，幾つかの課題も残されている．まず，擬似パラレルデータ作成の際に雑音をクリーン音声に重畳するが，これは加法性雑音には対応できるが，乗法性雑音の場合単純に雑音を重畳することは難しい．乗法性雑音に関しては，その歪を推定し，その歪みフィルタをクリーン音声に施すなどの対策が必要でであると考える．

また，現在の手法では擬似パラレルデータを作成する際に，入力発声の始端と終端が雑音であるという仮定を設定して，雑音のみの区間を抽出している．しかし，この仮定が崩れた場合には，適応データが上手く作成できない可能性がある．そこで今後は，雑音環境下の音声区間検出 (Voice Activity Detection: VAD)[24] などの手法を用いて，音声区間を

適切に検出する必要がある．また，適切な VAD が可能になれば，発声単位ではなく単語単位などより細かい区間で雑音のみの区間が検出できるようになり，より細かく変換関数を適応できるようになる．これらの対策により，さらに Eigen-SPLICE の導入可能な範囲が広がり，精度が向上していくことが期待される．

6.2.2　Eigen Joint GMM 法

また，Eigen Joint GMM 法に関しては，計算機の制約で GMM の混合数と特徴量が低くなってしまっているので，計算アルゴリズムの改善，及び計算資源の増強によって，より大きな混合数および特徴量の次元数で試してみる必要がある．混合数を上げることにより，ある程度の精度向上が見込まれるが，やはり Joint Vector で GMM を学習する限り，GMM による確率的な特徴量空間の分割が適切に行われづらい可能性がある．

6.3　今後の展望

6.3.1　GMM 学習の改善

SPLICE などの GMM に基づいた特徴量変換手法では，GMM の学習が非常に重要になってくる．GMM によって雑音付加音声の特徴量空間をその後の変換に適するように，確率的に分割できれば，より高い性能が得られることが予想される．例えば GMM の学習に，雑音付加音声ではなく，それを NMN したものを用いた手法が，通常の手法よりも高い性能を示していたり [8]，GMM の学習を線形判別分析を用いて学習した手法なども高い性能をしめしたりしている [25]．

これらの改善法は，GMM の学習方法を変えるだけなので，今回の提案手法である Eigen-SPLICE に組み込むことが可能であると考える．雑音付加音声特徴量の GMM の各コンポーネントが出来るだけ，対象のクリーン音声の音韻的なラベルなどを基準として，似たもので固まるようにすれば，そのコンポーネントの変換関数も無理なく学習することが出来る．

6.3.2　効果的な特徴量

今回提案した，Eigen-SPLICE などの手法は基本的に，どんな特徴量であったとしても雑音付加音声とクリーン音声のペアがあれば導入可能である．そのため，この特徴量をより雑音に強いものにすると，性能の改善が見込まれる．例えば，従来の SPLICE に HEQ を施した特徴量を用いることで高精度化する手法が提案も提案している [7]．また，近年注目されている multi layer perceptron[26] によって抽出された特徴量などを用いることも考えれる．

特徴量を正規化することで，耐雑音性が向上するということ以外にも，前述の雑音付加音声特徴量の取りうる値の分散を小さくすることにより，より適切な GMM が学習されるという効果も期待できると考える．

6.3.3 Uncertainty Decoding の導入

以上のような改善方法の他に，より大きな精度改善が見込める Uncertainty Decoding(UD) の枠組みの導入がある．

雑音環境下における音声認識では，一般的にモデル適応のアプローチが最も高い性能を示す．しかし，モデル適応では非常に多くの計算量を必要とするという問題点がある．そこで，SPLICE などの特徴量変換手法ではなく，UD を組み合わせた手法が特徴量変換程度の計算量でモデル適応に近い性能を発揮する枠組みとして提案されてきた．以下，その枠組みを簡単に説明して行く．

特徴量変換では，推定されたクリーン音声の特徴量のみをデコーダに渡し，推定された特徴量をクリーン音声の音響モデルでデコードする．一方，UD ではデコーダに条件付きの確率分布 $p(\boldsymbol{y}|\boldsymbol{x})$ を渡す．デコーダでは，$p(\boldsymbol{y}|\boldsymbol{x})$ を用いて，クリーン音声の音響モデル $p(\boldsymbol{x}|\theta)$ から近似的に雑音付加音声の音響モデル $p(\boldsymbol{y}|\theta)$ を求め，でデコーディングしていく．但しここで，θ は音響モデルのある正規分布を表すインデックスとする．

クリーン音声の音響モデル $p(\boldsymbol{x}|\boldsymbol{\theta})$ と条件付き確率分布 $p(\boldsymbol{y}|\boldsymbol{x})$ 雑音付加音声の音響モデルは，下記の式の様に求められる．

$$p(\boldsymbol{x}|\theta) = \int p(\boldsymbol{y},\boldsymbol{x}|\theta)d\boldsymbol{x} \tag{6.1}$$

$$= \int p(\boldsymbol{y}|\boldsymbol{x},\theta)p(\boldsymbol{x}|\theta)d\boldsymbol{x} \tag{6.2}$$

ここで，$p(\boldsymbol{y}|\boldsymbol{x})$ はモデルの状態に依存しないという仮定を置き，下の式のような近似をする．

$$p(\boldsymbol{y}|\boldsymbol{x},\theta) \simeq p(\boldsymbol{y}|\boldsymbol{x}). \tag{6.3}$$

このとき雑音付加音声の音響モデルのインデックス θ の確率分布が正規分布となるために，$p(\boldsymbol{y}|\boldsymbol{x})$ も正規分布であることが必要となる．

SPLICE などの特徴量変換手法と UD の枠組みの違いの概要を図 6.1 に示す．

UD with SPLICE

この条件付き確率を求める際に，UD with SPLICE [27] では，以下のように式を展開してく．

$$p(\boldsymbol{y}|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\boldsymbol{y})p(\boldsymbol{y})}{p(\boldsymbol{x})}. \tag{6.4}$$

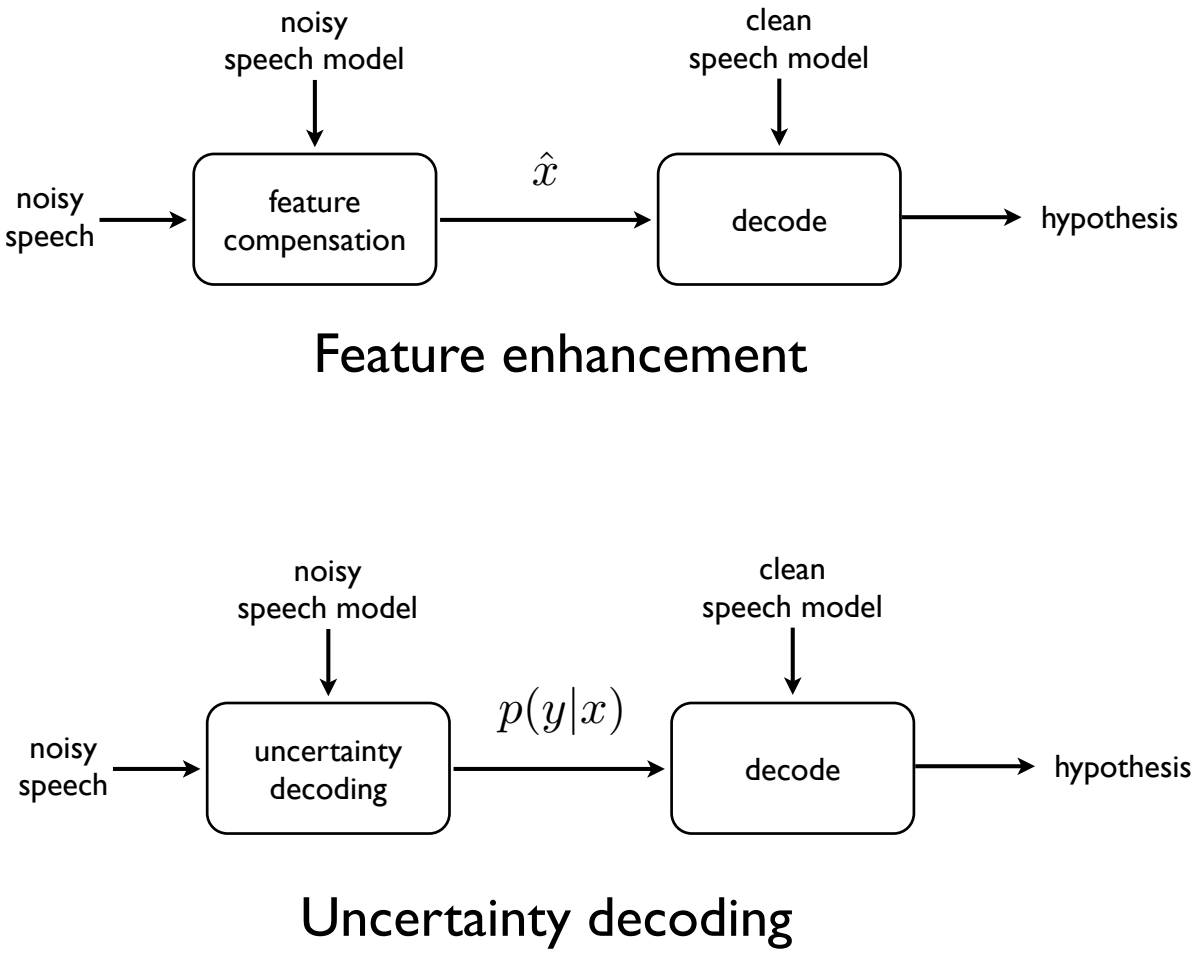

図 6.1: Difference between Feature enhancement and Uncertainty decoding

ここで $\boldsymbol{x}$ は，$\boldsymbol{y}$ によって周辺化できるものと仮定し，$\boldsymbol{y}$ の GMM によって以下のように近似できる．

$$p(\boldsymbol{x}) \simeq \sum_k p(\boldsymbol{x}|k)p(k) \tag{6.5}$$

$$= \sum_k \int p(\boldsymbol{x}, \boldsymbol{y}|k)p(k)d\boldsymbol{y} \tag{6.6}$$

$$= \sum_k \int p(\boldsymbol{x}|\boldsymbol{y}, k)p(\boldsymbol{y}|k)p(k)d\boldsymbol{y}. \tag{6.7}$$

但し，このとき k は $\boldsymbol{y}$ を表す GMM のインデックスである．最後に，求められた $p(\boldsymbol{x})$ が正規分布である必要があるため，この GMM のそれぞれのコンポーネントの平均や分散から，一つの正規分布として近似する．こうすることで，SPLICE の枠組みで $p(\boldsymbol{y}|\boldsymbol{x})$ が正規分布として求められる．

Joint UD

一方，Joint UD[28] では，GMM ベースの声質変換 (GMM based VC) と同様の枠組みで $p(\boldsymbol{y}|\boldsymbol{x})$ を求める．

$$p(\boldsymbol{y}|\boldsymbol{x}) = \sum_k p(\boldsymbol{y}|\boldsymbol{x}, k)p(k|\boldsymbol{x}). \tag{6.8}$$

但し，このときの k は $\boldsymbol{x}, \boldsymbol{y}$ のジョイントベクトルの GMM のインデックスである．ここで，$\boldsymbol{x}$ の所属コンポーネントの事後確率は，$\boldsymbol{y}$ の所属コンポーネントの事後確率によって求められるという近似を置く．

$$p(k|\boldsymbol{x}) \simeq p(k|\boldsymbol{y}). \tag{6.9}$$

すると，式 (6.8) は

$$p(\boldsymbol{y}|\boldsymbol{x}) \simeq \sum_k p(\boldsymbol{y}|\boldsymbol{x}, k)p(k|\boldsymbol{y}), \tag{6.10}$$

となる．ここで $p(\boldsymbol{y}|\boldsymbol{x})$ を正規分布として求めるために，もっとも尤度 $p(k|\boldsymbol{y})$ の高い k^* をもとめる．

$$p(\boldsymbol{y}|\boldsymbol{x}) \simeq p(\boldsymbol{y}|\boldsymbol{x}, k^*)p(k^*|\boldsymbol{y}) \tag{6.11}$$

$$k^* = \underset{k}{\operatorname{argmax}}\, p(k|\boldsymbol{y}) \tag{6.12}$$

以上の枠組みによって，Joint UD で条件付き確率 $p(\boldsymbol{y}|\boldsymbol{x})$ を正規分布として求められた．

これらの Uncertainty Decoding の手法を提案手法に組み込むことによって大きな性能改善が見込まれる．今後，この手法が雑音環境下における音声認識システムの発展に貢献することを大いに期待する．

謝辞

本論文を執筆するにあたり，常日頃のご指導，ご鞭撻を頂きました指導教官の峯松信明先生に深く感謝致します．また，打合せ等で深い洞察とご指摘をいただきました広瀬啓吉先生にも感謝いたします．博士課程でいらっしゃるときから，理論的および技術的な指導をいただきました齋藤大輔先生にも深く感謝いたします．

また，研究室の機材の整備など，本研究を様々な面で支援してくださった当研究室技官の高橋登氏，出張や物品の購入の際に事務手続きなどの面から支えてくださった秘書の池上恵氏に深く感謝致します．

本論文のアイディアおよび研究方針についてさまざまな助言をくださり，論文添削など丁寧なご指導をいただきました博士課程の鈴木雅之氏に深く感謝いたします．また，研究室でお世話になった先輩方，いつでも議論に応じてくれた同期の皆や後輩にも，感謝いたします．最後に，公私にわたり私を支えてくれた家族と友人たちに感謝します．

2012 年 2 月 8 日
千々岩圭吾

参考文献

[1] 秋田祐哉, 三村正人, 河原達也. 会議録作成支援のための国会審議の音声認識システム (画像符号化・映像メディア処理レター特集). 電子情報通信学会論文誌. D, 情報・システム, Vol. 93, No. 9, pp. 1736–1744, 2010.

[2] 安藤彰男, 今井亨, 小林彰夫, 本間真一, 後藤淳, 清山信正, 三島剛, 小早川健, 佐藤庄衛, 尾上和穂ほか. 音声認識を利用した放送用ニュース字幕制作システム. 信学論, D, Vol. 84, pp. 877–887, 2001.

[3] comSCORE http://www.comscore.com/. 日本のスマホ利用動向最新情報.

[4] 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.

[5] V. Stouten. Robust Automatic Speech Recognition in Time-varying Environments. *PhD thesis*, 2006.

[6] S. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, Vol. 28, No. 4, pp. 357–366, 1980.

[7] Tsunenobu Kai, Masayuki Suzuki, Keigo Chijiiwa, Nobuaki Minematsu, and Keikichi Hirose. Combination of splice and feature normalization for noise robust speech recognition. *Proceeding of International Workshop on Nonlinear Circuits Communications and Signal Processing*, (2012. to appear).

[8] J. Droppo, L. Deng, and A. Acero. Evaluation of the SPLICE algorithm on the Aurora2 database. *Proc. Eurospeech*, Vol. 1, pp. 217–220, 2001.

[9] J.C. Segura, A. De La Torre, M.C. Benitez, and A.M. Peinado. Model-based compensation of the additive noise for continuous speech recognition. Experiments using the AURORA II database and tasks. *Proc. Eurospeech*, Vol. 1, pp. 221–224, 2001.

[10] AP Varga and RK Moore. Hidden markov model decomposition of speech and noise. In *Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on*, pp. 845–848. IEEE, 1990.

[11] M.J.F. Gales. Model-ased techniques for noise ro ust speech recognition. *PhD thesis*, 1995.

[12] M. Cooke, P. Green, L. Josifovski, and A. Vizinho. Robust automatic speech recognition with missing and unreliable acoustic data. *Speech communication*, Vol. 34, No. 3, pp. 267–285, 2001.

[13] B.R. Ramakrishnan. *Reconstruction of incomplete spectrograms for robust speech recognition*. PhD thesis, Carnegie Mellon University, 2000.

[14] 齋藤大輔. 音声の構造的表象に基づく音声合成技術に関する基礎的研究. 東京大学大学院 院新領域創成科学研究科 修士論文, 2008.

[15] 鹿野清宏, 伊藤克亘, 河原達也, 武田一哉, 山本幹雄. 音声認識システム. オーム社, 2001. ISBN: 4-274-13228-5.

[16] P. Alexandre and P. Lockwood. Root cepstral analysis: A unified view. application to speech processing in car noise environments. *Speech Communication*, Vol. 12, No. 3, pp. 277–288, 1993.

[17] J. Droppo, A. Acero, and L. Deng. Efficient on-line acoustic environment estimation for fcdcn in a continuous speech recognition system. Vol. 1, pp. 209–212, 2001.

[18] A. Kain and M.W. Macon. Spectral voice conversion for text-to-speech synthesis. *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, Vol. 1, pp. 285–288, 1998.

[19] K. Chijiiwa, M. Suzuki, N. Minematsu, and K. Hirose. Evaluation of speech recognition in noisy environments using eigen-splice. *Proc. ASJ, 2011, Autumn, 2–Q–19, (in Japanese)*, 2011.

[20] K. Chen, W. Liau, H. Wang, and L. Lee. Fast speaker adaptation using eigenspace-based maximum likelihood linear regression. *Proc. ICSLP*, Vol. 3, pp. 742–745, 2000.

[21] T. Toda, Y. Ohtani, and K. Shikano. Eigenvoice conversion based on gaussian mixture model. *Proc. ICSLP*, pp. 2446–2449, 2006.

[22] D. Pearce, H.G. Hirsch, et al. The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. *Proc. ICSLP*, Vol. 4, pp. 29–32, 2000.

[23] http://htk.eng.cam.ac.uk/. *The Hidden Markov Model Toolkit*.

[24] K. Ishizuka, T. Nakatani, M. Fujimoto, and N. Miyazaki. Noise robust voice activity detection based on periodic to aperiodic component ratio. *Speech Communication*, Vol. 52, No. 1, pp. 41–60, 2010.

[25] 鈴木雅之, 吉岡拓也, 峯松信明, 広瀬啓吉. 非定常雑音環境における線形判別分析を用いた静的・動的 mfcc の強調. 日本音響学会講演論文集 秋, Vol. 1-10-6, pp. 15–18, 2011.

[26] Z. Tüske, C. Plahl, and R. Schlüter. A study on speaker normalized MLP features in LVCSR. *Proc. INTERSPEECH*, pp. 1089–1092, 2011.

[27] J. Droppo, A. Acero, and L. Deng. Uncertainty decoding with splice for noise robust speech recognition. *Proc. ICASSP*, Vol. 1, pp. I–57, 2002.

[28] H. Liao and M.J.F. Gales. Issues with uncertainty decoding for noise robust automatic speech recognition. *Speech Communication*, Vol. 50, No. 4, pp. 265 – 277, 2008.

発表文献

[1] 千々岩圭吾: “言語的制約に非依存な基本周波数パターン生成過程モデルパラメータ自動抽出の高精度化,” 東京大学工学部電子情報工学科卒業論文, 2010.

[2] 千々岩圭吾, 鈴木雅之, 齋藤大輔, 峯松信明, 広瀬啓吉: “基本周波数パターン生成過程モデルにおけるフレーズ指令パラメータ抽出の高精度化,” 日本音響学会春季講演論文集, pp. 495–498, 2010.

[3] 千々岩圭吾, 鈴木雅之, 齋藤大輔, 峯松信明, 広瀬啓吉: “Eigen-SPLICE を用いた雑音環境下における音声認識,” 情報処理学会研究報告, SLP, 音声言語情報処理 2011–SLP–87 No.15, 2011.

[4] 千々岩圭吾, 鈴木雅之, 峯松信明, 広瀬啓吉: “Eigen-SPLICE を用いた雑音環境下における音声認識の実験的検討,” 日本音響学会秋季講演論文集, pp. 113–116, 2011. (**学生優秀発表賞**)

[5] 千々岩圭吾, 鈴木雅之, 峯松信明, 広瀬啓吉: “主成分分析を用いた GMM に基づくた耐雑音音声認識フロントエンドの高精度化,” 日本音響学会春季講演論文集, (2012. 発表予定)

[6] Keigo Chijiiwa, Masayuki Suzuki, Nobuaki Minematsu, Keikichi Hirose: “Unseen Noise Robust Speech Recognition using Adaptive Piecewise Linear Transformation,” *Proceeding of IEEE International Conference on Acoustics, Speech, and Signal Processing,* (2012. to appear)

[7] Tsunenobu Kai, Masayuki Suzuki, Keigo Chijiiwa, Nobuaki Minematsu, Keikichi Hirose: “Combination of SPLICE and Feature Normalization for Noise Robust Speech Recognition,” *Proceeding of International Workshop on Nonlinear Circuits, Communications and Signal Processing,* (2012. to appear)