博士論文

非言語情報の違いに頑健な特徴量表現に着目した
ニューラルネットワーク音声認識に関する研究

2015年12月01日

指導教員　峯松 信明 教授

東京大学大学院　工学系研究科
電気系工学専攻

37-137057　柏木 陽佑

あらまし

音声認識システムは，スマートフォンの普及や計算機パワーの発展などに伴い，非常に身近なものとなりつつある．単純な入力インタフェースとしての性能は実用に耐えうるものとなって来たが，その認識性能は人間のそれと比べた場合未だ充分とは言えない．今後，自動音声認識技術のさらなる応用を考えた場合，認識性能の向上は必要不可欠である．

我々が普段何気なく聞いている音声には様々な情報が内在する．これらは大きく言語情報，パラ言語情報，非言語情報の 3 つに分けることができ，音声認識とは，この内の言語情報を抽出するタスクであると考えることができる．なお，本研究において前提とする音声認識システムでは，パラ言語情報による影響は音声分析の過程で無くなるものと考えることができる．しかし，非言語情報は音響特徴量に影響を与えるためノイズであり，認識性能の低下の原因となる．そのため，これをいかに制御するかが音声認識において長年の課題であった．

さて，近年，統計的機械学習はニューラルネットワークの台頭という大きな転換点を迎えた．音声認識技術においてもこの影響は大きく，ニューラルネットワークベースの識別的な音声認識システムが高い性能を示すことも報告され，もはやニューラルネットワークベースのシステムが主流となったと言える．しかし，ニューラルネットワークは従来のガウス分布，ガウス混合分布に基づく生成的なモデルとは性質が大きく異なる．そのため，従来の非言語情報に関する要素技術をそのまま利用することが困難であり，現在のニューラルネットワークをベースとする音声認識における非言語情報の制御に関する研究の潮流は，手探りな状況であることが否めない．これを打開するためには，ニューラルネットワーク音声認識に適した理論的背景に重点を置いた非言語情報の制御技術の研究が重要である．

そこで，本論文では，非言語情報の違いに頑健なニューラルネットワーク音声認識システムの実現を目指す．非言語情報の制御として大きなウェイトを占める特徴量と音響モデルにおいて，従来のガウス混合分布に基づく非言語情報の制御に関する要素技術を基にした，ニューラルネットワークベースの非言語情報の制御手法を提案する．先に述べた通り，ニューラルネットワークとガウス混合分布は性質が大きく異なるため，単純な応用は困難である．そのため，ニューラルネットワークとガウス混合分布の融合により，両者の性質を組み合わせたアプローチを提案する．これにより，従来の生成的アプローチの持つパラメータの意味づけ，制御に関する要素技術をニューラルネットワーク音声認識に取り入れることが可能となり，認識性能の向上が可能となる．また，モデルの性質のみを考慮した

ニューラルネットワークとガウス混合分布の融合では，音声認識システムに対する非言語情報の扱いに対する本質的な解決策とはならない．そこで，本論文の後半では非言語情報の違いに頑健な特徴量表現である音声の構造的表象，そしてそれを構成する分布間距離の計算に対してニューラルネットワークを用いたアプローチを導入する．これにより，ニューラルネットワークの高い識別性能とそれを支える特徴量空間の表現能力を，従来のガウス分布をベースとする音声学的知見に基づいた手法との融合が可能となり，ニューラルネットワークを用いた新しい非言語情報の違いに頑健な特徴量表現が実現できる．

目次

図目次

図目次

表目次

第1章

序論

1.1　本論文の背景

自動音声認識は 20 世紀半ばから研究が始められ，統計的パターン認識技術の発達や大規模なコーパスの構築とともに精度が向上し，現在では高い認識率で大語彙の連続音声認識を実現出来るようになった．そのため，携帯電話やカーナビに代表される一般製品のインタフェースだけでなく，衆議院の議事録作成システム [1] などの幅広い分野に自動音声認識技術が利用されており，次世代のコンピュータインタフェースとして非常に注目されている．しかし，現在の自動音声認識システムの精度は，人間の認識能力には大きく劣り，未だ課題も多い．その原因の一つとして挙げられるのが非言語情報の制御の難しさである．

音声に内在する情報は言語情報・パラ言語情報・非言語情報の 3 種類に大別することができる．言語情報は文字言語により表現可能な情報であり，一般に現在の音声認識ではこの言語情報を抽出することを目的としている．パラ言語情報は，文字に表れないが，話者が意図を持って制御する情報であり，抑揚などの韻律的情報に代表される．なお，現在の音声認識で用いられる特徴量の一つであるメル周波数ケプストラム係数は，韻律情報を捨てており，このパラ言語情報の扱いに重きを置いていない．本研究でもパラ言語情報は取り扱わないため注意されたい．そして，非言語情報は，環境・話者の年齢・性別などに起因し，話者が通常制御しない情報である．

現在の自動音声認識システムは統計ベースの手法であるがゆえに，その学習には一般的に大規模なデータを要する．しかし，話者や収録環境等が整備された大規模なデータは収集すること自体が非常に高コストであるため，不特定話者・環境のデータで認識器を学習することが多い [2]．その結果，当然ながら，多様な非言語情報を持つデータを用いて認識器を学習することとなる．非言語情報はパラ言語情報と異なり，音声認識に用いる特徴量に影響を与えるため，認識器が多様な入力に対応できるようになるという利点はあるものの，特定の話者や環境のデータのみで学習した認識器と比較して認識性能が低下してしまう．その一例として，一般には単一話者の音声を認識する際，話者依存のモデルの方が話者非依存のモデルよりも高い認識率を示すことが知られている．そのため，さらなる認識性能の向上のためには，入力特徴量に内在する非言語情報をいかに制御するかが重要であると言える．この非言語情報の制御に関しては，特徴量ドメインの手法としては雑音抑圧や音声強調や特徴量正規化など [3,4]，音響モデルドメインの手法としてはモデル適応や正規化学習などが古くから研究されてきた [5–7]．これらは従来主流であった生成モデルである混合ガウスモデル（Gaussian Mixture Model ; GMM）をベースとした手法に対するアプローチであることが多い．

さて，近年，計算機の性能やディープラーニング技術の発達に伴い，ニューラルネットワークをベースとした識別モデルを用いる手法が多く提案されている．この背景として，ディープラーニングの登場による，深層ニューラルネットワーク (Deep Neural Network ; DNN) の安定した高い性能の実現の寄与するところが大きい [8]．自動音声認識における

音響モデルの主流は，GMM ベースからニューラルネットワークベースへと移り変わったと言っても過言ではない．DNN は従来のモデルが行っていた，認識に有効な特徴量の抽出と識別を内部において同時に行っていると解釈することができる．これは，従来 GMM をベースとした音声認識システムで用いられていた特徴量であるメル周波数ケプストラム係数の前段階である，メルフィルタバンク出力を直接特徴量として用いた方が認識性能が高いという報告からもわかる [8]．そのため，DNN 自体がそもそも GMM と比較して非言語情報に頑健なモデルであるとも考えられる．しかし，DNN も GMM と同様に話者依存モデルの方が話者非依存モデルより高い精度を示すことが知られている．また，例えば，音声認識における入力特徴量ドメインにおける話者正規化を行う Feature space Maximum Likelihood Linear Regression (fMLLR) は DNN ベースのモデルにおいても認識精度が向上することも報告されている [9]．したがって，DNN は非言語情報に対して十分に頑健であるわけではなく，GMM ベースのモデルと同様に DNN ベースのモデルに対しても非言語情報の制御は重要な課題であり，近年研究が盛んに行われている [10–18].

しかし，ここで DNN が識別モデルである点が大きな障害となる．従来の生成的なアプローチとはモデルの性質が大きく異なるため，非言語情報の制御という観点では，現状では未だ手探りな状況であることは否めない．DNN ベースのモデルは，その高い性能と引き換えに，各パラメータがどのような意味を持つのかを曖昧であるという特徴を持つ．これは機械学習の観点からは利点とも言えるが，それゆえ音声工学の観点からは今後のさらなる拡張が困難である．そのため，従来主流であった GMM ベースに対する手法が直接的に利用できないケースが多い．その一つとして話者適応技術が挙げられる．GMM は生成モデルであり，各混合の分布のパラメータが明示的にではないにしろ，音素などの音響的・言語的事象と比較的似た意味を持っていたため，話者性に対応するモデルパラメータの制御が容易であった．しかし，DNN は識別モデルであり，かつ非線形変換を多層に渡って積み上げるため，話者性とパラメータの対応が不明瞭である．その結果，生成モデルの様に話者性に相当するパラメータの制御が困難であるという欠点を持つ．

さて，音声の非言語情報に対する積み立ては，長い研究の積み立てと，生成モデルの性質から従来の GMM ベースに利があると言える．そのため，複雑な DNN における非言語情報の制御に対して，GMM ベースの従来の要素技術を道標とすることは，現状の手探りなニューラルネットワーク研究への一助となると考える．そこで，本研究では，従来の生成モデルをベースとした非言語情報の制御に関する要素技術の延長としてのニューラルネットワーク音声認識技術の提案と，それによる音声認識性能の向上を目指す．音声認識において非言語情報の制御に大きく寄与すると考えられる，特徴量ドメインと音響モデルドメインにおいて，DNN と GMM ベースのアプローチの融合を行い，その有効性を示す．また，単にモデルの性質を考慮した組み合わせではなく，音声学的知見に基づく非言語情報の違いに頑健な特徴量表現である音声の構造的表象 [19] にニューラルネットワークに基づくアプローチを導入することで，ニューラルネットワークに基づく新しい特徴量表現を提

案する．

1.2　本論文の目的

本研究の目的は，従来の GMM ベース音響モデルにおける非言語情報に関する要素技術を基にした，DNN ベース音響モデルの非言語情報に対する頑健性を向上と音声認識性能の向上である．本研究では，非言語情報の制御に大きく寄与するであろう特徴量ドメインのおける特徴量強調と音響モデルドメインにおけるモデル適応において，従来の GMM ベースの要素技術に由来する知見を基にしたニューラルネットワーク手法を提案する．また，さらに踏み込み，非言語情報の一つである話者の違いに頑健な特徴量である構造的表象とそれを構成する分布間距離へ，ニューラルネットワークによるアプローチを導入する．これにより非言語情報の違いに頑健な特徴量表現である構造的表象をニューラルネットワーク音声認識の領域へ導入することを検討する．

1.3　本論文の構成

本論文では，まず 第 2 章 で現在の自動音声認識技術の基礎とニューラルネットワークの音声認識への利用について述べ，現在の音声認識における重要な課題である，非言語情報とそれに伴うニューラルネットワーク音声認識の課題について紹介する．これに基づき，第 3 章では特徴量ドメインのアプローチとして，ニューラルネットワークと GMM を組み合わせることでニューラルネットワークの持つ高い識別性能と GMM の持つ汎化性能の両立を実現する．第 4 章では, 音響モデルドメインのアプローチとして，話者コードを用いた音響モデルの環境適応を提案する．これは，GMM 音響モデルにおいてモデル適応性能を向上する正規化学習法をニューラルネットワークにおいて実現したものである．また，第 5 章では話者の違いに頑健な特徴量表現である構造的表象を構成する分布間距離をニューラルネットワークにより計算する手法を提案する．さらに，これを特徴量ドメインにおける評価として言語識別タスクへ，音響モデルドメインにおける評価として音響モデルの話者適応の際の制約に導入することを提案する．最後に第 6 章で本論文をまとめる．

第2章

音声認識に関する基礎技術とニューラルネットワーク

2.1　はじめに

本研究は，非言語情報の制御による音声認識システムの認識性能の向上を目的とし，識別的アプローチへの従来の生成的アプローチによる知見の導入を行う．それに先立ち本章では，音声認識技術の基礎について解説し，ニューラルネットワークに関する要素技術の紹介を行う．その後，識別的なモデルであるニューラルネットワークが，従来の生成的なアプローチを基本とする音声認識にどのように利用されているのかを紹介する．これにより現在の音声認識システムの持つ，ある種歪な関係性について言及する．

2.2　音声認識の基礎技術

音声認識では，特徴量を抽出する特徴抽出部と，音響的特徴を記述する音響モデル，言語的特徴を記述する言語モデル，そしてこれらを統合し認識結果を出力するデコーダがある．なお，本研究では特徴量抽出と音響モデルに主眼を置いているが，音声認識全体におけるこれらの立ち位置を明確にするため，簡単ではあるが，言語モデルとデコーダについても紹介を行う．

2.2.1　定式化

音声認識は，まず観測された音声から音声分析により特徴量を抽出する．その後，得られた特徴量系列 $\boldsymbol{X}$ に対して事後確率 $P(\boldsymbol{W}|\boldsymbol{X})$ を最大とする単語系列 $\boldsymbol{W}$ を見つける問題として定式化することができる (図 2.1).

$$\hat{\boldsymbol{W}} = \underset{\boldsymbol{W}}{\operatorname{argmax}} \log P(\boldsymbol{W}|\boldsymbol{X}) \tag{2.1}$$

しかし，音声は系列情報であるため $P(\boldsymbol{W}|\boldsymbol{X})$ を直接モデル化することが困難である．そのため，一般にはベイズ則を適応することで,

$$\begin{aligned} \hat{\boldsymbol{W}} &= \underset{\boldsymbol{W}}{\operatorname{argmax}} \log \frac{P(\boldsymbol{X}|\boldsymbol{W})P(\boldsymbol{W})}{P(\boldsymbol{X})} && (2.2) \\ &= \underset{\boldsymbol{W}}{\operatorname{argmax}} \log P(\boldsymbol{X}|\boldsymbol{W})P(\boldsymbol{W}) && (2.3) \end{aligned}$$

として $P(\boldsymbol{X}|\boldsymbol{W})$ を音響モデル，$P(\boldsymbol{W})$ を言語モデルとしてモデル化する．

2.2.2　特徴量抽出部

音声認識に用いる特徴量は音声の生成メカニズムと密接な関係を持つ．人間の音声は，声帯の振動により発生する空気の振動が声道を通り，口からの放射特性を経て耳に届く．こ

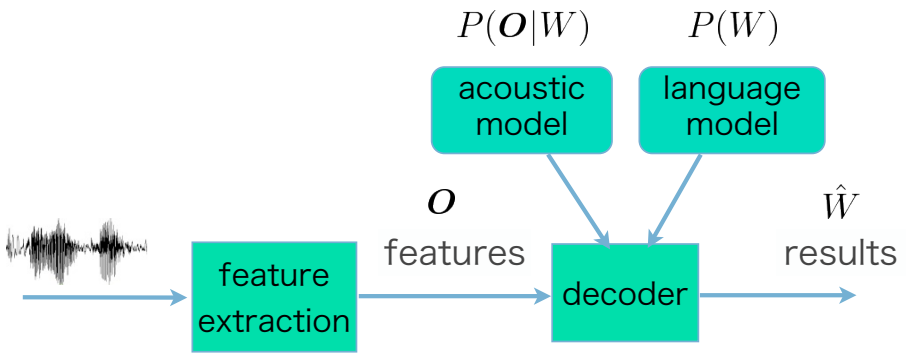


図 2.1: 自動音声認識システムの概形

の音声の生成メカニズムは声帯の振動や乱流などによって生じる音源と，声道による伝達特性を持つ調音フィルタによるソースフィルタモデルとして考えることができる．音声の最小単位である音素は，音源の種類（破裂音，声帯振動，乱流等）と調音フィルタの形状で決定されるが，特に音源の性質は声の高さや大きさ等の音声の韻律的な性質に影響している．そのため，音源の情報を除いた声道フィルタの伝達特性に相当する情報を抽出することで，音声認識に有効な音響特徴量が得られると考えられる．この考えに基づいて設計された特徴量がケプストラムである．本節では，まずケプストラム分析について述べ，その後音声認識において用いられることの多い特徴量であるメル周波数ケプストラム係数とメルフィルタバンク出力について紹介を行う．

i)　ケプストラム分析

音声は，声帯の振動や摩擦による乱流等の音源信号が調音フィルタの伝達特性によって音韻情報が付与された物であり，音素の音響的な特徴は，主に調音フィルタの振幅伝達特性によって担われている．そのため，音声認識に有効な特徴量を得るには，音声から音源成分を除去し，調音フィルタの伝達特性に相当する情報を抽出すれば良いことがわかる．ケプストラム（Cepstrum）は，声帯の振動成分と調音フィルタの伝達特性を比較した際に，調音フィルタの伝達特性が対数振幅スペクトルの包絡に相当することを利用して抽出する．図 2.2 にケプストラム分析の概形を示す．

まず，音声に対して窓関数との畳み込みを行った後，短時間フーリエ変換により，スペクトルを得る．音源信号のスペクトラムを $G(e^{j\omega})$，調音フィルタの伝達特性を $H(e^{j\omega})$ とすると，音声信号のスペクトル $S(e^{j\omega})$ は，

$$S(e^{j\omega}) = G(e^{j\omega}) \cdot H(e^{j\omega}) \tag{2.4}$$

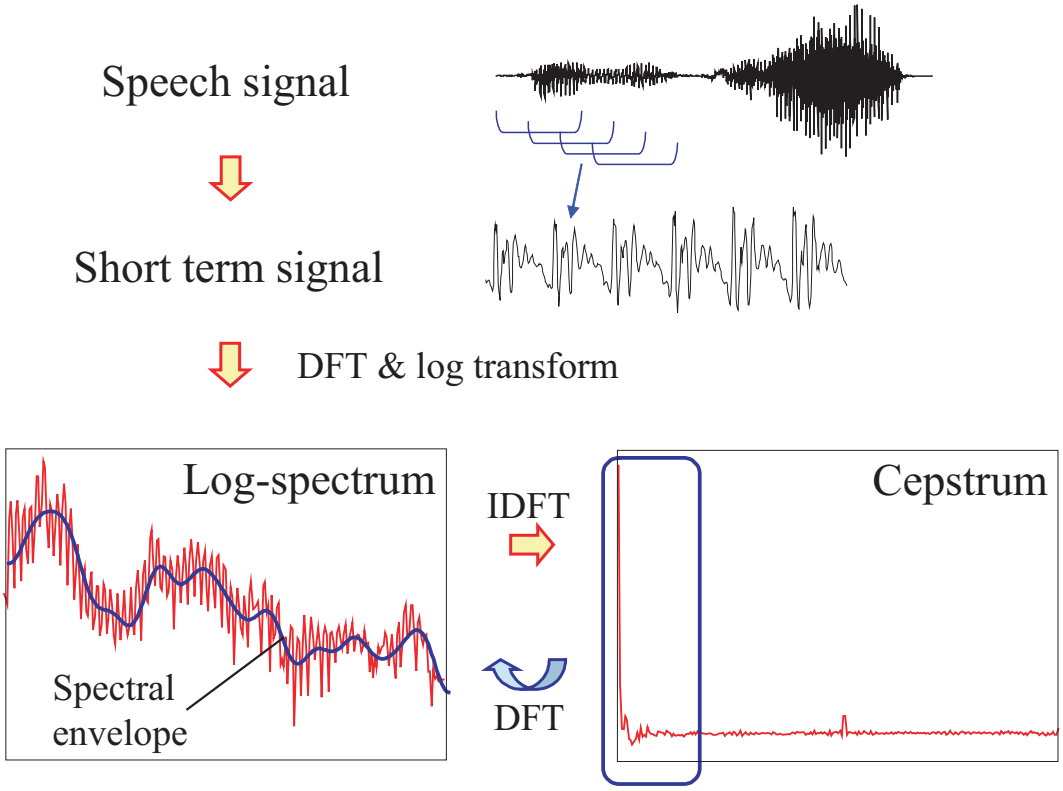


図 2.2: ケプストラム分析

として表すことができる．ここで，対数振幅スペクトルを考えると，

$$|S(e^{j\omega})| = |G(e^{j\omega})| \cdot |H(e^{j\omega})| \tag{2.5}$$

$$\log|S(e^{j\omega})| = \log|G(e^{j\omega})| + \log|H(e^{j\omega})| \tag{2.6}$$

となるため，音声の対数振幅スペクトルは音源と調音フィルタの対数振幅スペクトルの和となる．ここで，対数振幅スペクトルの微細構造は音源成分を，包絡は調音フィルタの成分が表れていることが知られている．そこで，これを時間信号と考えて逆フーリエ変換を行うことで，低い周波数帯域に調音フィルタに相当する情報が集中すると考えることができる．つまり，対数振幅スペクトルの逆フーリエ変換がケプストラム係数であり，ケプストラム係数の低次項からは韻律に相当する情報が取り除かれていると考えることができる．

ii)　メルフィルタバンク出力とメル周波数ケプストラム係数

一般に，音声認識では人間の聴覚特性にあわせて調整したメル周波数領域でケプストラム分析を行うことによって得られるメル周波数ケプストラム係数（Mel-Frequency Cepstrum Coefficient ; MFCC）を用いることが多い．メル周波数ケプストラム係数は，スペクトル空

間においてメルフィルタバンクを適用して得られるメルフィルタバンク出力を元に，逆フーリエ変換を行うことで得られる．音響分析に用いられるメル周波数領域は周波数領域を

$$Mel(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \tag{2.7}$$

により変換することで得られることできる．音声認識に用いる MFCC はこれと同様の処理を実現するため周波数軸上に L 個の三角窓を配置したメルフィルタバンクを用い，これを用いた分析により得ることが出来る．音声スペクトルを $S'(k)$ とするとし，各窓の中心点 $k_c(l)$ をメル周波数軸上で等間隔に配置すると，L 個の窓から得られるメルフィルタバンク出力は

$$m(l) = \sum_{k=k_{lower}}^{k_{upper}} W(k;l)|S'(k)| \quad (l = 1, \cdots, L) \tag{2.8}$$

$$W(k;l) = \begin{cases} \frac{k-k_{lower}(l)}{k_c(l)-k_{lo}(l)} & \{k_{lower}(l) \leq k \leq k_c(l)\} \\ \frac{k_{upper}(l)-k}{k_{upper}(l)-k_c(l)} & \{k_c(l) < k \leq k_{upper}(l)\} \end{cases} \tag{2.9}$$

となる．その後，得られた L 個のパワーを離散コサイン変換（DCT）することで，MFCC が求まる．

$$c_{MFCC} = \sqrt{\frac{2}{N}} \sum_{l=1}^{L} \log m(l) \cos\left\{\left(l - \frac{1}{2}\right)\frac{j\pi}{L}\right\} \tag{2.10}$$

また，近年メルフィルタバンク出力を直接ニューラルネットワーク音響モデルの入力特徴量として採用することがある．これは，(逆）フーリエ変換は線形変換と考えることができるため，ニューラルネット側にその機能を持たせることができるという考えに基づく．

2.2.3　音響モデル

音響モデルは音声特徴量の性質をモデル化するためのものである．音声特徴量は系列情報であり，時間伸縮するためモデル化の際にこれを吸収しなければならない．そこで，一般には隠れマルコフモデルを用いてモデル化される．隠れマルコフモデルの各状態の持つ特徴量分布として従来ガウス混合分布が用いられてきたが，近年ニューラルネットワークによりこれを擬似的に表現する手法が主流となってきている．ここでは，簡略のためガウス分布を各状態の特徴量分布として持つ隠れマルコフモデルについて説明する．

2.2.4　音響モデルの分類

ガウス混合分布（Gaussian Mixture Model; GMM）を各状態の特徴量分布として持つ隠れマルコフモデル（Hidden Markov Model ; HMM）を GMM/HMM と呼ぶ．この GMM/HMM は構築する音響モデルの単位によってワード型とサブワード型に分類することができる．サブワード型は音素などの音声の最小単位で HMM を学習するのに対して，ワード型は単語単位で HMM を構築する．これらはタスクによって向き不向きが変わるが，サブワード型は少量の HMM で多様な表現ができるため，大語彙音声認識は音素単位でのサブワード HMM が使われる．

また，サブワード型の HMM ではモノフォン，トライフォンモデルが一般に採用される．1 つの音素を 1 つの HMM でモデル化する枠組みがモノフォンモデルである．しかし音素はそれぞれで独立ではなく，前後の音素の影響を大きく受けることが知られている．これは，調音結合と呼ばれ，音声の認識を困難にしている要因の１つである．そのため，連結学習で単純に音素を当てはめるのでは，前後の音素の違うものも全て同一の音素であるとして学習が行われてしまう．これに対処するため，前後の音素を考慮した 3 つ組み音素単位として HMM のモデル化を行う枠組みがトライフォンモデルである．日本語の音素はおおよそ 40 種類と言われ，40 種類の音素に対応するモデルを学習すれば任意の単語の認識が可能となる．しかし，前後の音によって音素の音が影響をうけてしまう調音結合が起こるため，一般的には，直近の前後の音素を考慮するトライフォンを用い，このトライフォンモデルを学習する際の初期値としてモノフォンを用いることも多い．また，さらに細かなモデル化のため，4 つ組み以上も考慮する場合もある．逆に，学習データが少量の場合は，スパース性の問題からモノフォンで行われる．また，トライフォンは膨大なクラス数となる．このままではスパースなため，学習データ中に一度も出ない 3 つ組み音素もある．そこで，HMM を決定木などによりクラスタリングすることにより対処することが一般に行われる．

i)　隠れマルコフモデル

HMM は系列データのためのモデルであり，複数の状態と状態間の遷移確率，そして各状態からの出力の分布を持つ．図 2.3 に HMM の概形を示す．音声認識では一般的に，状態遷移が自身もしくは次の状態のみである left-to-right 型の HMM が用いられる．HMM は状態遷移を隠れ変数に持つモデルであり，学習はバウムウェルチアルゴリズムによって最尤推定により行うことができる．なお，これは EM アルゴリズム (Expectation Maximization algorithm) の HMM への拡張である．特徴量系列を $\boldsymbol{X}$，モデルパラメータを Θ とし，

$$\hat{\Theta} = \underset{\Theta}{\operatorname{argmax}} \sum_{l} \log P(\boldsymbol{X}_l|\Theta) \tag{2.11}$$

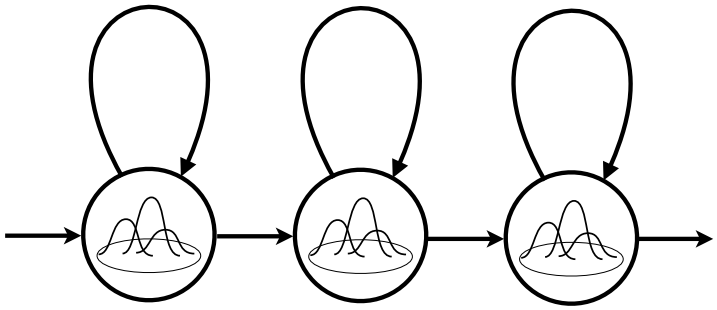


図 2.3: ガウス混合分布を各状態の特徴量分布に持つ隠れマルコフモデル

のように最尤基準でモデルパラメータを学習する．ここで，l はデータインデクスである．隠れ状態系列を $\boldsymbol{s}$ とすると，

$$\hat{\Theta} = \underset{\Theta}{\operatorname{argmax}} \sum_{l} \log \sum_{\boldsymbol{s}} P(\boldsymbol{X}_l, \boldsymbol{s}|\Theta) \tag{2.12}$$

$$= \underset{\Theta}{\operatorname{argmax}} \sum_{l} \log \sum_{\boldsymbol{s}} P(\boldsymbol{X}_l|\boldsymbol{s}, \Theta) P(\boldsymbol{s}|\Theta) \tag{2.13}$$

とすることができる．これを直接計算することは困難であるため，EM アルゴリズムでは下限を保証し反復計算を行うことで最大化を行う．条件付期待値である Q 関数は

$$Q(\Theta|\Theta^{(n)}) = \sum_{l} \sum_{\boldsymbol{s}} P(\boldsymbol{s}|\boldsymbol{X}_l, \Theta^{(n)}) \log P(\boldsymbol{X}_l|\boldsymbol{s}, \Theta) P(\boldsymbol{s}|\Theta) \tag{2.14}$$

となる．

さて，実際には状態系列は非常に膨大であるため，$P(\boldsymbol{s}|\boldsymbol{X}_l, \Theta^{(n)})$ を全ての状態系列について計算するのは困難である．そこで，バウムウェルチアルゴリズムでは，効率的に計算するために前向き・後ろ向きアルゴリズムを用いる．前向き確率 $\alpha_l(t, i)$ は，時刻 t において状態 i に至る確率であり，

$$\alpha_l(t, i) = P(s_t = i, \boldsymbol{x}_{l,1}, \boldsymbol{x}_{l,2}, \ldots, \boldsymbol{x}_{l,t}|\Theta) \tag{2.15}$$

となる．また，後ろ向き確率 $\beta_l(t, j)$ は，時刻 t において状態 j を出発して時刻 $T + 1$ に終状態 M に辿り着く確率であり，

$$\beta_l(t, j) = P(\boldsymbol{x}_{l,t+1}, \boldsymbol{x}_{l,t+2}, \ldots, \boldsymbol{x}_{l,T}|s_t = j, \Theta) \tag{2.16}$$

となる．これらを用いると，時刻 t において状態 m を通過する系列の出現確率は

$$P(s_t = m|\boldsymbol{X}_l, \Theta) = \frac{\alpha_l(t,m)\beta_l(t,m)}{\alpha_l(T+1,M)} = \psi_l(t,m) \tag{2.17}$$

として計算することが可能となる．

さて，各状態における特徴量分布をガウス分布と仮定すると，求めるべきパラメータは状態遷移確率 $a(m)$ と各状態における特徴量分布の平均 $\boldsymbol{\mu}_m$ と分散 $\boldsymbol{\sigma}_m^2$ である．最終的に，$b_m(\boldsymbol{x})$ を状態 m の特徴量分布から特徴量 $\boldsymbol{x}$ が出力される確率とすると，それぞれの更新式は

$$\hat{a}(m) = \frac{\sum_l \left[\frac{1}{\alpha_l(T+1,M)} \sum_t \alpha_l(t,m)a(m)b_m(\boldsymbol{x}_{lt})\beta_l(t,m)\right]}{\sum_l \sum_t \psi_l(t,m)} \tag{2.18}$$

$$\hat{\boldsymbol{\mu}}_m = \frac{\sum_l \sum_t \psi_l(t,m)\boldsymbol{x}_{lt}}{\sum_l \sum_t \psi_l(t,m)} \tag{2.19}$$

$$\hat{\boldsymbol{\sigma}}_m^2 = \frac{\sum_l \sum_t \psi_l(t,m)(\boldsymbol{x}_{lt} - \hat{\boldsymbol{\mu}}_m)(\boldsymbol{x}_{lt} - \hat{\boldsymbol{\mu}}_m)^\top}{\sum_l \sum_t \psi_l(t,m)} \tag{2.20}$$

となる．先に述べた通り音声認識の場合，一般的には特徴量分布を混合ガウス分布やニューラルネットワークによる識別モデルによって表現することが多い．以降では，混合ガウス分布を各状態の特徴量分布に持つ HMM を GMM/HMM と呼ぶこととする．混合ガウス分布であれば，最尤推定により同様にパラメータ推定することができるが，ニューラルネットワークのパラメータの推定は困難である．そのため，ニューラルネットワークを各状態に用いる場合は，ニューラルネットワークのみの学習を別途行う．そして，あらかじめ学習しておいた GMM/HMM の遷移確率を流用することとなる．詳細はニューラルネットワーク音響モデルにて述べる．

実際の学習データは，文章を読み上げた音声とそれに対応するテキストが与えられることが多い．しかし，音素 HMM を構築する場合，音素ラベルとそれに対応する音声が本来必要である．このミスマッチを解消するため，音素 HMM を連結し文 HMM を作成し，それを用いて HMM の学習を行う．これを連結学習と呼ぶ．図 2.4 のように音素 HMM のパラメータを連結した文 HMM と特徴量系列を用いてモデルパラメータを学習することにより，音素単位での学習データがなくとも学習することができる．

2.2.5　言語モデル

音声認識に用いられる代表的な言語モデルとしては，N-gram が挙げられる．ここでは，N-gram 言語モデルについて紹介する．N-gram 言語モデルは，次の単語の出現確率は，その直前の $N-1$ 単語にのみ依存するという仮定に基づきモデル化する．単語列 $w_1, w_2, \ldots, w_n$

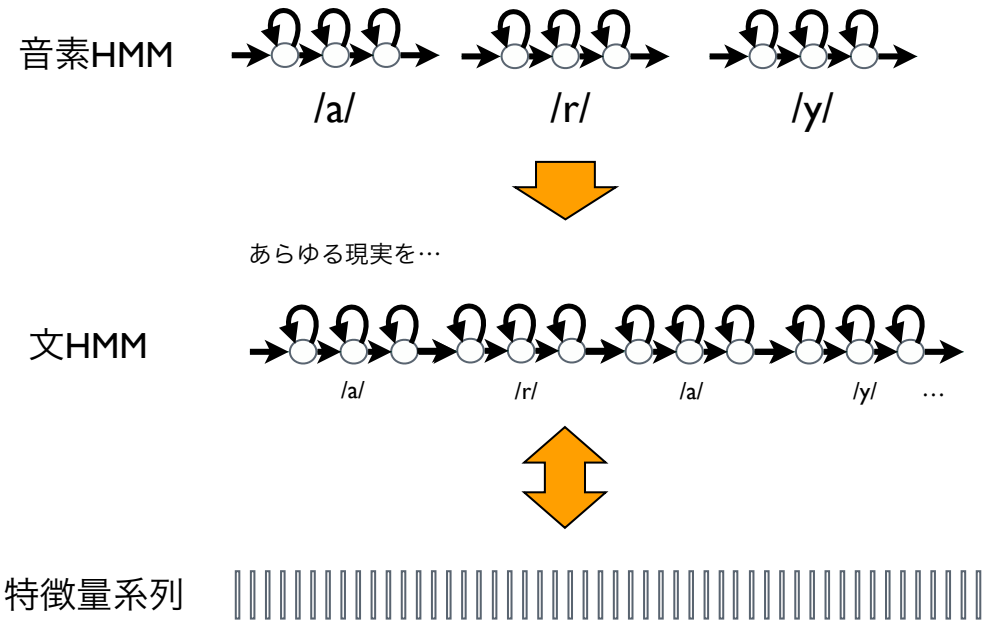


図 2.4: 連結学習

が与えられた時の，その出現確率 $P(w_1, w_2, \ldots, w_n)$ を

$$P(w_1, w_2, \ldots, w_n) = \prod_{i=1}^{n} P(w_i | w_{i-N+1}, \ldots, w_{i-1}) \tag{2.21}$$

と近似を行う．これは，i 番目の単語 w_i の生成確率が，直前の $N-1$ 単語 $w_{i-N+1}, \ldots, w_{i-1}$ にのみ依存すると仮定する．ここで，$N = 1$ の場合がユニグラム，$N = 2$ の場合がバイグラム，$N = 3$ の場合がトライグラムと呼ばれる．N-gram の学習は最尤推定によって行う．そのため，学習データに存在しないものについては確率が 0 となる．言語モデルは非常にスパースであるため，確率が 0 となるものが発生しやすく，これが認識誤りを引き起こす．そこで，確率の高いものを下げ，確率の小さなものを上げるスムージングを行うことによってこの問題に対処するすることができる．

2.2.6　デコーディング

デコーダでは，音響モデルと言語モデルから得られたスコアを用いて入力特徴量系列に対する最尤のパスを選択する．またデコードの際に音響モデルと言語モデルのスコアに対して重み付けを行うことで認識性能の向上が可能となる．言語モデルに対する重みを α と

すると，

$$\hat{\boldsymbol{W}} = \underset{\boldsymbol{W}}{\operatorname{argmax}}(1-\alpha)\log P(\boldsymbol{X}|\boldsymbol{W}) + \alpha \log P(\boldsymbol{W}) \tag{2.22}$$

として単語系列を計算する．なお，この重みはあらかじめ学習データ，評価データとは別に開発データを用意しておき，これを用いて決定することが多い．近年では，重み付き有限状態トランスデューサ (Weighted Finite-State Transducer ; WFST) によってネットワークを表現し，デコードする手法が主流となっている [20]．WFST デコーダは，複数のネットワークを合成することが可能であり，また，on-the-fly デコーディングにより部分的に合成することで省メモリ化も実現できる．

2.3　ニューラルネットワークに関する要素技術

本節では近年の深層ニューラルネットワーク (Deep Neural Network ; DNN) 技術の基礎と，その発展の重要な節目となった事前学習について述べる．

2.3.1　深層ニューラルネットワークの基礎

ニューラルネットワークは多くのノードとその間の連結からなり，各ノード間の連結が脳神経細胞のように情報が伝搬する連結がより強くなるように学習されるモデルである．音声認識で用いられるニューラルネットワークは，層ごとに一方向にのみ伝搬する順伝播型 (feedforward neural network) の人工ニューラルネットワークを基本とするものが多いため，ここでは順伝播型の深層ニューラルネットワークについて紹介する．

図 2.5 に順伝播型の深層ニューラルネットワークを示す．各層毎に多数のノードを持ち，入力から順に層を伝搬して出力層へと至る．中間の層は，中間層もしくは隠れ層と呼ぶが，本論文では隠れ層で統一する．隠れ層の各ノードは入力に対して活性化関数を作用させたものを出力する．学習はヘッブの法則に基づいて行い，誤差逆伝播法（Backpropagation ; BP）が用いられる．隠れ層が 1 層の入力 2 次元，隠れ層 2 次元，出力 1 次元の 3 層のニューラルネットワークの場合，入力層を $\boldsymbol{x} = \{x_1, x_2\}$，隠れ層を $\boldsymbol{h} = \{h_1, h_2\}$，出力層を O とすると，以下の様に書く事が出来る．

$$h_1 = f(w^1_{1,1}x_1 + w^1_{2,1}x_2 + w_{0,1}) \tag{2.23}$$

$$h_2 = f(w^1_{1,2}x_1 + w^1_{2,2}x_2 + w_{0,2}) \tag{2.24}$$

ここで w_{ij} は x_i から h_j への重みである．$w_{0,j}$ はバイアス成分を表している．簡略のため

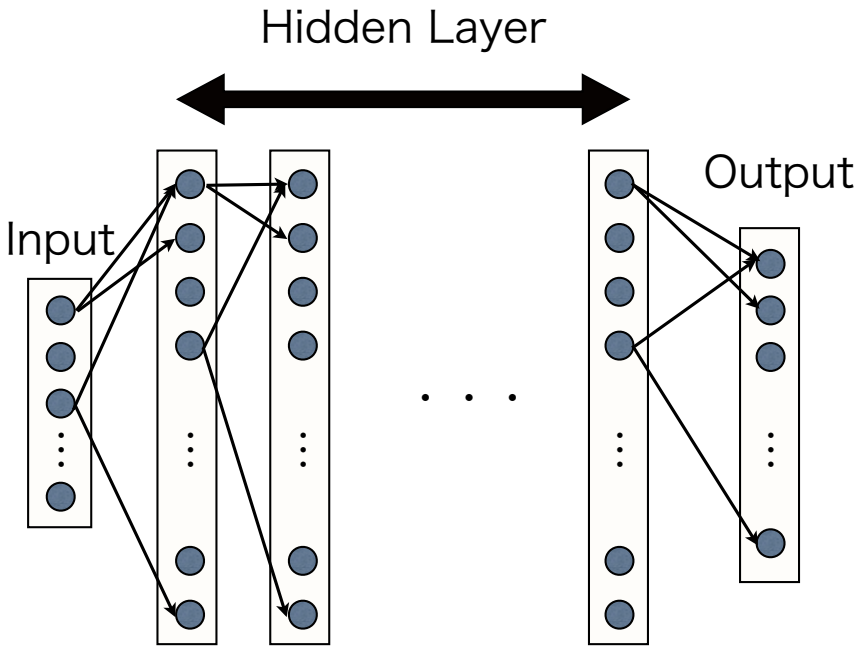


図 2.5: 深層ニューラルネットワーク

$x_0 = 1$ と置くと

$$h_1 = f(\sum_i x_i w_{i,1}) \tag{2.25}$$

$$h_2 = f(\sum_i x_i w_{i,2}) \tag{2.26}$$

と表す事が出来る．同様に h についても一般化を行うと

$$h_j = f(\sum_i x_i w^1_{i,j}) \tag{2.27}$$

となる．隠れ層から出力層も同様に一般化すると

$$O = f(\sum_i h_i w^2_{i,0}) \tag{2.28}$$

となる．活性化関数である f は tanh やシグモイド関数などの非線形関数を使うのが一般的である．シグモイド関数を導入すると

$$\sigma(x) = \frac{1}{1 + \exp(-x)} \tag{2.29}$$

となり，シグモイド関数の微分は

$$\frac{\partial}{\partial x}\sigma(x) = \sigma(x)(1-\sigma(X)) \tag{2.30}$$

となる．誤差逆伝播法は出力層側から再急降下法により順に誤差を修正していく．学習データを $\{\hat{\boldsymbol{x}}, \hat{O}\}$ とすると，

$$\begin{aligned} w'^{2}_{1,0} &= w'^{2}_{1,0} + k\frac{\partial O}{\partial w^{2}_{1,0}}(\hat{O}-O) && (2.31)\\ &= w^{2}_{1,0} + k^{2}_{1,0}h_1O(1-O)(\hat{O}-O) && (2.32)\\ & && (2.33) \end{aligned}$$

となる．ここで k はラーニングレートに相当し，これによって学習の速度や安定性が変化する．そして，最後に入力層と隠れ層の連結パラメータ部分を修正する．

$$\begin{aligned} w'^{1}_{1,1} &= w^{1}_{1,1} + k^{1}_{1,1}x_1h_1(1-h_1)(e_1w^{2}_{1,0}) && (2.34)\\ e_1 &= h_1O(1-O)(\hat{O}-O) && (2.35)\\ & && (2.36) \end{aligned}$$

以上を一般化して書くと,隠れ層 n と出力層の連結

$$\begin{aligned} w'^{n}_{i,j} &= w^{n}_{i,j} + k^{n}_{i,j}e^{n}_{i,j} && (2.37)\\ e^{n}_{i,j} &= h^{n}_{i}O_j(1-O_j)(\hat{O}_j-O_j) && (2.38)\\ & && (2.39) \end{aligned}$$

隠れ層 $n-1$ と隠れ層 n の連結

$$\begin{aligned} w'^{n-1}_{i,j} &= w^{n-1}_{i,j} + k^{n-1}_{i,j}e^{n-1}_{i,j} && (2.40)\\ e^{n-1}_{i,j} &= h^{n-1}_{i}h^{n}_{j}(1-h^{n}_{j})\{\sum_{k} e^{n}_{j,k}w^{n}_{j,k}\} && (2.41) \end{aligned}$$

入力層と隠れ層 1 の連結

$$\begin{aligned} w'^{0}_{i,j} &= w^{0}_{i,j} + k^{0}_{i,j}e^{0}_{i,j} && (2.42)\\ e^{0}_{i,j} &= h^{0}_{i}h^{1}_{j}(1-h^{1}_{j})\{\sum_{k} e^{1}_{j,k}w^{1}_{j,k}\} && (2.43) \end{aligned}$$

と更新を行うことによって学習が可能となる．学習の際はラーニングレートの他に，バッチサイズやエポック数等の学習に関するハイパーパラメータが数多く存在する．これらの制御に関しては多くの研究が提案されている [21].

2.3.2　事前学習

ニューラルネットワーク自体は古くからある技術であるが，多層のニューラルネットワークはその高い表現力から簡単に局所解に落ちてしまうため学習が困難であるという問題（勾配消失）があった．特に入力層に近くなる程，誤差逆伝搬法による誤差の伝搬効率が悪くなるため，パラメータの更新がほとんど起きないことが主な原因だと言われている．例えば，先に述べたシグモイド関数の微分は，

$$\frac{\partial}{\partial x}\sigma(x) = \sigma(x)(1-\sigma(X)) \leq 0.25 \tag{2.44}$$

となる．これは，各層毎に誤差の伝搬に対して 0.25 の係数が掛かることとに相当するため，入力層に近づくほど誤差が伝搬しないことは容易に理解できる．

これを解決する手法の一つが Hinton らにより提案された事前学習の導入である [22]．これによって，深層ニューラルネットワークのパラメータの初期値が比較的良いものとなるため，高い識別性能の実現が可能となった．近年のニューラルネットワークでは活性化関数の改良などによって事前学習なしに高い認識性能を得ることができるモデルが多く提案されているため，事前学習を行わないことも多くなっている．しかし，この事前学習は近年のニューラルネットワーク技術の根幹と大きく結びついており，重要な意味を持つ．ここでは，広く知れ渡った事前学習法である制約付きボルツマンマシンを用いたディープビリーフネットワークについて紹介する．

i)　ディープビリーフネットワークによる事前学習

ディープビリーフネットワークは制約付きボルツマンマシン (Restricted Boltzmann Machine ; RBM) を多層に積んだネットワークである．これを深層ニューラルネットワークのパラメータの初期値とすることて，効率的に深層ニューラルネットワークの学習が可能であることが Hinton らにより提案されたのが近年のニューラルネットワークの台頭の始まりである [22]．制約付きボルツマンマシンは可視素子の層と隠れ素子の層からなる無向２部グラフで各層内での結合はない特殊なボルツマンマシンである．可視素子はデータの入（出）力をおこなう素子であり，隠れ素子はモデルの内部自由度を増加させる．各素子は {0,1} の値を確率的に取るようになっている．RBM の定義は

$$P_{RBM}(\boldsymbol{v},\boldsymbol{h}|\boldsymbol{\Theta}) = \frac{1}{Z}\exp(-E(\boldsymbol{v},\boldsymbol{h})) \tag{2.45}$$

$$E(\boldsymbol{v},\boldsymbol{h}) = -\sum_{i\in V} b_i v_i - \sum_{j\in H} c_j h_j - \sum_{i\in V}\sum_{j\in H} v_i W_{ij} h_j \tag{2.46}$$

$$\tag{2.47}$$

である．ここで，b_i と c_i は可視素子と隠れ素子のバイアス項，W_{ij} はウェイトパラメータ．Z は分配関数であり，

$$Z = \sum_{\boldsymbol{v}} \sum_{\boldsymbol{h}} \exp(-E(\boldsymbol{v}, \boldsymbol{h})) \tag{2.48}$$

となる．なお，それぞれの層内でのノードの結合は想定しないため，条件付き確率は独立である．

$$P(\boldsymbol{v}|\boldsymbol{h}) = \prod_i p(v_i|\boldsymbol{h}) \tag{2.49}$$

$$P(\boldsymbol{h}|\boldsymbol{v}) = \prod_j p(h_j|\boldsymbol{v}) \tag{2.50}$$

したがって各ノードの条件付き確率は

$$P(v_i = 1|\boldsymbol{h}) = \sigma(b_i + \sum_j W_{ij} h_j) \tag{2.51}$$

$$P(h_j = 1|\boldsymbol{v}) = \sigma(c_j + \sum_i W_{ij} v_i) \tag{2.52}$$

のようにシグモイド関数で表現することができる．

この RBM のパラメータ $\boldsymbol{\Theta}$ は，尤度最大化基準（Kullback-Leibler Divergence 最小化基準）により求めることができる．

$$\begin{aligned}
\operatorname*{argmax}_{\boldsymbol{\Theta}} \sum_{\boldsymbol{v}} q(\boldsymbol{v}) \ln p(\boldsymbol{v}|\boldsymbol{\Theta}) &= \operatorname*{argmax}_{\boldsymbol{\Theta}} \sum_{\boldsymbol{v}} q(\boldsymbol{v}) \ln \sum_{\boldsymbol{h}} p_{RBM}(\boldsymbol{v}, \boldsymbol{h}|\boldsymbol{\Theta}) && (2.53)\\
&= \operatorname*{argmax}_{\boldsymbol{\Theta}} \left\langle \ln \sum_{\boldsymbol{h}} \exp^{-E(\boldsymbol{v},\boldsymbol{h}|\boldsymbol{\Theta})} \right\rangle_{q(\boldsymbol{v})} - \ln Z && (2.54)\\
q(\boldsymbol{v}) &= \frac{1}{N} \sum_k \delta(\boldsymbol{v} - \boldsymbol{v}^k) && (2.55)
\end{aligned}$$

ここで，$\delta(x)$ はディラックのデルタ関数であり，式 (2.55) は出現頻度をカウントして確率化している．RBM のパラメータ $\boldsymbol{\Theta}$ は勾配法を用いて推定する．式 (2.54) を微分すると次の式が得られる．

$$\begin{aligned}
\frac{\partial J}{\partial \Theta} &= -\left\langle \frac{1}{\sum_{\boldsymbol{h}} \exp(-E)} \sum_{\boldsymbol{h}} \frac{\partial E}{\partial \Theta} \exp(-E) \right\rangle_q && (2.56)\\
&\quad + \frac{1}{Z} \sum_{\boldsymbol{v}} \sum_{\boldsymbol{h}} \frac{\partial E}{\partial \Theta} \exp(-E) && (2.57)
\end{aligned}$$

これは

$$p(\boldsymbol{h}|\boldsymbol{v}) = \frac{p(\boldsymbol{v},\boldsymbol{h})}{p(\boldsymbol{v})} = \frac{\exp(-E(\boldsymbol{v},\boldsymbol{h}))}{\sum_{\boldsymbol{h}}\exp(-E(\boldsymbol{v},\boldsymbol{h}))} \tag{2.58}$$

$$\frac{\partial J}{\partial \Theta} = -\sum_{\boldsymbol{v}}\sum_{\boldsymbol{h}}\frac{\partial E}{\partial \Theta}p(\boldsymbol{h}|\boldsymbol{v})q(\boldsymbol{v}) + \sum_{\boldsymbol{v}}\sum_{\boldsymbol{h}}\frac{\partial E}{\partial \Theta}p(\boldsymbol{h},\boldsymbol{v}) \tag{2.59}$$

$$= -\left\langle\frac{\partial E}{\partial \Theta}\right\rangle_{data} + \left\langle\frac{\partial E}{\partial \Theta}\right\rangle_{model} \tag{2.60}$$

として整理することができる．ここで，$\left\langle\frac{\partial E}{\partial \Theta}\right\rangle_{data}$ は $q(\boldsymbol{v})$ が既知なので簡単に計算できる．しかし，$\left\langle\frac{\partial E}{\partial \Theta}\right\rangle_{model}$ はそのままでは求める事が困難である．そこで RBM の学習には Contrastive Divergence 法を導入する．Contrastive Divergence 法はマルコフ連鎖モンテカルロ（MCMC）の一種であり，多層ニューラルネットの事前学習には Contrastive Divergence 1 (CD1) と呼ばれる学習データ $\boldsymbol{v}$ を可視層と隠れ層の間を一往復だけ遷移させた結果 $\hat{\boldsymbol{v}}$ を利用する近似が用いられる．遷移は $p(\boldsymbol{h}|\boldsymbol{v})$，$p(\boldsymbol{v}|\boldsymbol{h})$ を用いてサンプリングにより行う．

$$p(\boldsymbol{v},\boldsymbol{h})_{model} \approx \frac{1}{N}\sum_{k}\delta(\boldsymbol{v}-\hat{\boldsymbol{v}}^k)p(\boldsymbol{h}|\hat{\boldsymbol{v}}^k) \tag{2.61}$$

CD1 により式 (2.60) が計算できるようになるため，パラメータ $\{W_{ij}, b_i, c_j\}$ ごとに微分する．最終的な更新式は α をラーニングレートとすると

W:

$$\frac{\partial E}{\partial W_{ij}} = -v_i h_j \tag{2.62}$$

$$\frac{\partial J}{\partial W_{ij}} = < v_i h_j >_{data} - < v_i h_j >_{model} \tag{2.63}$$

$$< v_i h_j >_{data} = \sum_{\boldsymbol{v}}\sum_{\boldsymbol{h}} v_i h_j q(\boldsymbol{v})p(\boldsymbol{h}|\boldsymbol{v}) \tag{2.64}$$

$$= \sum_{\boldsymbol{h}}\left\{\frac{1}{N}\sum_{k} v_i^k h_j p(\boldsymbol{h}|\boldsymbol{v}^k)\right\} \tag{2.65}$$

$$= \frac{1}{N}\sum_{k} v_i^k \sigma(c_j + \sum_{i'}\hat{v}_{i'}^k W_{i'j}) \tag{2.66}$$

$$< v_i h_j >_{model} = \frac{1}{N}\sum_{k}\hat{v}_i^k \sigma(c_j + \sum_{i'} v_{i'}^k W_{i'j}) \tag{2.67}$$

$$W'_{ij} = W_{ij} \quad + \quad \alpha\{\frac{1}{N}\sum_k v_i^k \sigma(c_j + \sum_{i'} v_{i'}^k W_{i'j}) \tag{2.68}$$

$$- \quad \frac{1}{N}\sum_k \hat{v}_i^k \sigma(c_j + \sum_{i'} \hat{v}_{i'}^k W_{i'j})\} \tag{2.69}$$

b:

$$b'_i = b_i + \alpha\{\frac{1}{N}\sum_k v_i^k - \frac{1}{N}\sum_k \hat{v}_i^k\} \tag{2.70}$$

c:

$$c'_j = c_j + \alpha\{\frac{1}{N}\sum_k \sigma(c_j + \sum_{i'} v_{i'}^k W_{i'j}) \tag{2.71}$$

$$-\frac{1}{N}\sum_k \sigma(c_j + \sum_{i'} \hat{v}_{i'}^k W_{i'j})\} \tag{2.72}$$

となる．なお，RBM の学習においても多くのハイパーパラメータが存在する．これに関しては [23] が詳しい．

2.4　ニューラルネットワークの音声認識への利用

本節では近年の音声認識システムにおいて，ニューラルネットワークがどのように利用されているのかを紹介する．

2.4.1　特徴量抽出器としての利用

i)　タンデムアプローチ

タンデムアプローチは教師あり特徴量抽出を用いた手法であり，Hermansky らによって提案された [24]．図 2.6 にタンデムアプローチのブロック図を示す．あらかじめ観測された音声の特徴量（MFCC 等）から音素ラベルや HMM の状態ラベルを識別する識別器を用いて有効な特徴量を抽出し，従来の混合ガウス分布を各状態の特徴量分布に持つ HMM 音響モデルを用いた認識をするアプローチである．この際に用いるラベルはあらかじめ学習した HMM により得る．そして，この識別器としてニューラルネットが用いられ，認識時は，音声特徴量を入力として得られる各ラベルに対する事後確率をそのまま特徴量としてもとの音声特徴量に連結して用いる．タンデムアプローチの本質は識別器の出力するラベルに対する事後確率を特徴量として用いることである．一般に音素事後確率は非常にスパースになることが経験的に知られており，これにより認識率が向上すると考えられる．

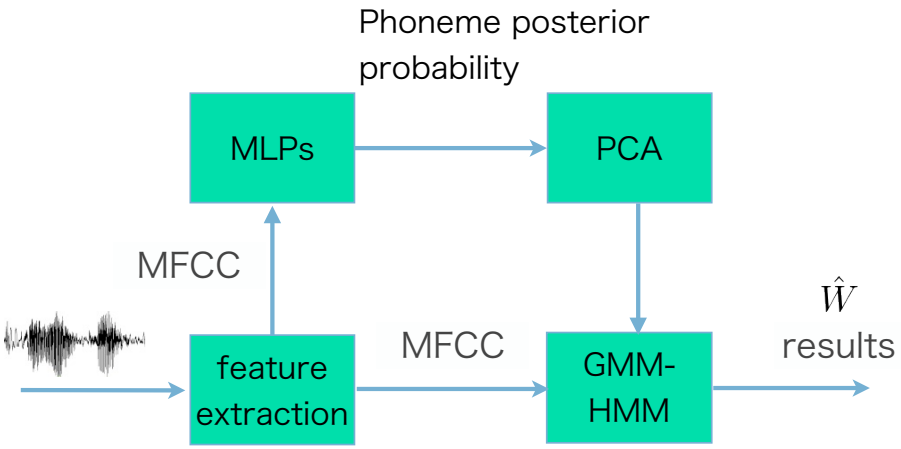


図 2.6: タンデムアプローチ

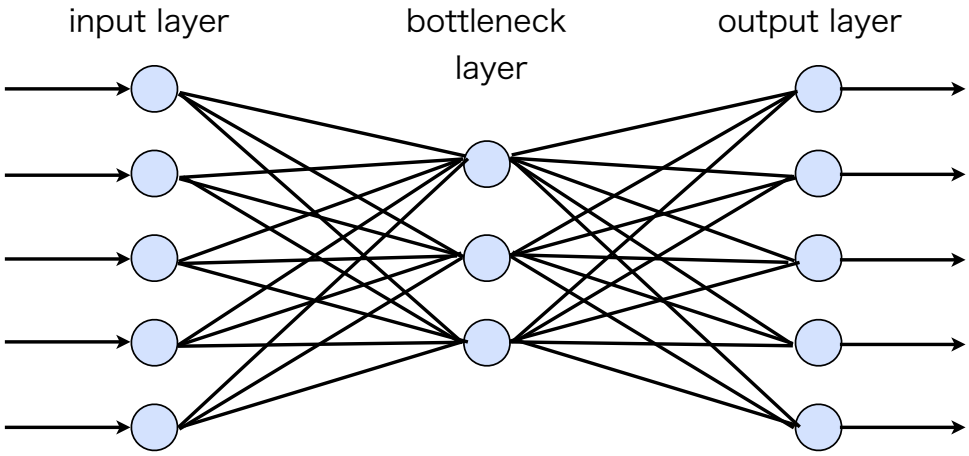


図 2.7: オートエンコーダー

ii)　オートエンコーダーによる特徴抽出

オートエンコーダーは教師なし特徴量抽出の一種であり，ニューラルネットの隠れ層が次元圧縮を行っていると考える．図 2.7 にオートエンコーダーの概略図を示す．まず，入力特徴量（MFCC 等）を入力とし，１層の隠れ層をもつ折り返したニューラルネットを構築し，誤差逆伝搬法によりパラメータを学習する．その後，出力層を取り除くと，隠れ層は特徴量抽出を行っていることに相当する．さらに層を追加して折り返し，誤差逆伝搬法により学習することを繰り返すことで，多層の特徴量抽出が可能となる．オートエンコーダーの特徴として，非線形素子により非線形の次元圧縮が可能である点が上げられる．なお，これはカーネル PCA の考え方にも似ており，一層のオートエンコーダーは活性化関数が線形の場合，主成分分析と等価となる．また，正則化を行ったり学習基準を変化させることで，よりデータのスパース性に強い特徴量抽出を行う研究も行われている [25].

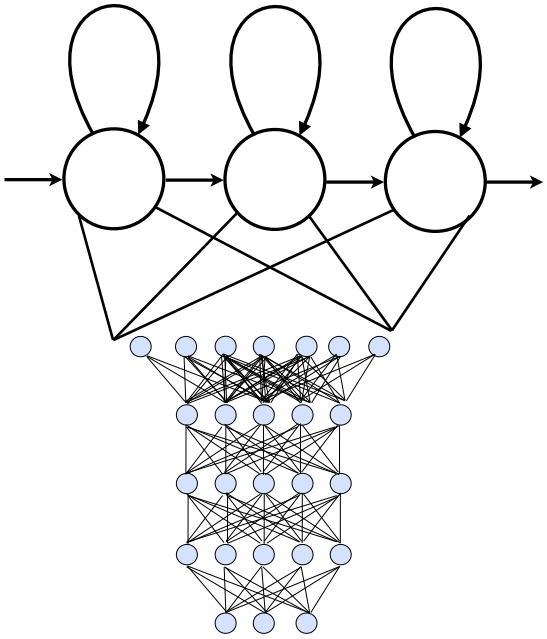


図 2.8: 各状態の特徴量分布をニューラルネットワークにより擬似的に表現した隠れマルコフモデル

2.4.2　音響モデルへの利用

i)　各状態の特徴量分布をニューラルネットワークにより擬似的に表現した隠れマルコフモデル

ニューラルネットワークを用いた音響モデルとしては GMM/HMM の GMM 部分を深層ニューラルネットワークで置き替えたモデルが現在広く用いられている．以後，このモデルのことを DNN/HMM と呼ぶこととする．しかし，GMM は生成モデルであるのに対してニューラルネットワークは識別モデルであるため，直接用いることはできない．そこで，DNN と HMM を組み合わせたモデルとして Hybrid approach が提案されている [26]．先に述べた通り，音声認識は式 (2.73) のように観測特徴量系列 $\boldsymbol{O}$ が得られた時，正解テキスト w の推定値 $\hat{W}$ を求めるタスクである．

$$\hat{W} = \underset{W}{\operatorname{argmax}} \log P(W|\boldsymbol{O}) \tag{2.73}$$

式 (2.73) を音響モデルから算出される確率 $P(\boldsymbol{O}|W)$ と言語モデルから算出される確率 $P(W)$ の積で表すと

$$\hat{W} = \underset{W}{\operatorname{argmax}} \log P(\boldsymbol{O}|W)P(W) \tag{2.74}$$

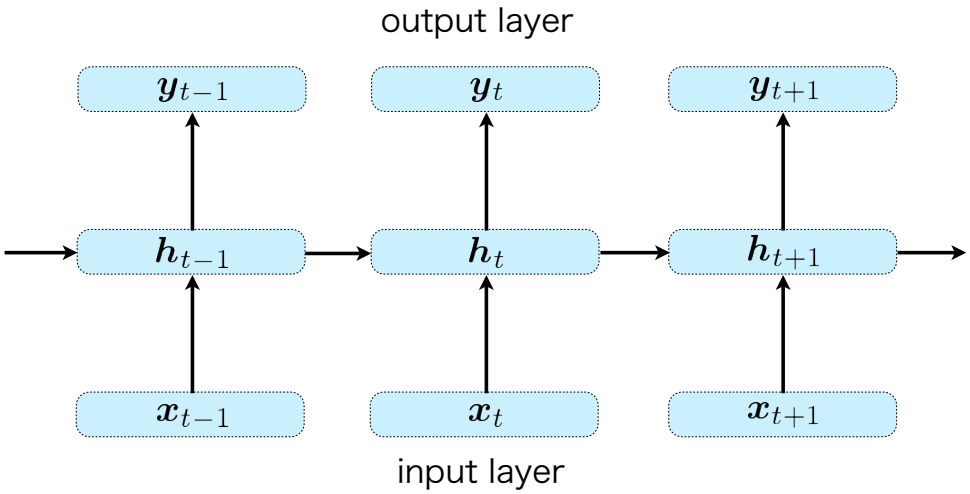


図 2.9: 再帰ニューラルネットワーク

となり，この音響モデル部分 $P(\boldsymbol{x}|l)$ をニューラルネットワークでモデル化することを考える．音声信号の時間方向の伸縮構造の表現には，従来通り HMM を用いることを考えて，音響モデル確率をベイズ則の導入と HMM 状態系列で周辺化した形で記述すると

$$P(\boldsymbol{O}|W) \stackrel{\text{def}}{=} \sum_{\boldsymbol{s}} \left\{ \prod_{t} Q(s_t|\boldsymbol{o}_t, \boldsymbol{\Lambda}) \right\} P(\boldsymbol{O}) P(\boldsymbol{s}|l) \tag{2.75}$$

$$Q(s_t|\boldsymbol{o}_t, \boldsymbol{\Lambda}) = \frac{p(s_t|\boldsymbol{o}_t, \boldsymbol{\Lambda})}{P(s_t)} \tag{2.76}$$

となる．ここで $\boldsymbol{\Lambda}$ はニューラルネットワークのパラメータ，$\boldsymbol{s}$ は HMM 状態系列である．$Q(s_t|\boldsymbol{o}_t, \boldsymbol{\Lambda})$ は $P(\boldsymbol{o}_t|s_t, \boldsymbol{\Lambda})$ に比例する項であり，$P(s_t)$ は一様分布を仮定する，あるいはコーパス中の出現回数をカウントし，最尤推定により求める．このようにして，識別モデルであるニューラルネットワークにより擬似的に HMM の各状態の特徴量分布を表現することで音響モデルとして利用することが可能となる．

ii)　再帰ニューラルネットワーク

HMM で時系列をモデル化を行うのは，DNN との親和性が悪い．また，波形全体を直接モデル化するのは，いかに DNN が過学習しづらい識別器であると言っても難しい．そこで，何らかの制約を導入した上で識別モデルで識別することが考えられる．再帰ニューラルネットワーク (Recurrent Neural Network ; RNN) は自然言語処理等で用いられていた系列情報を隠れ層を畳み込むことで考慮するモデルである．ニューラルネットの隠れ層に 1 つ前の隠れ層の状態を畳み込むことによって理論的には無限長の系列を考慮することがで

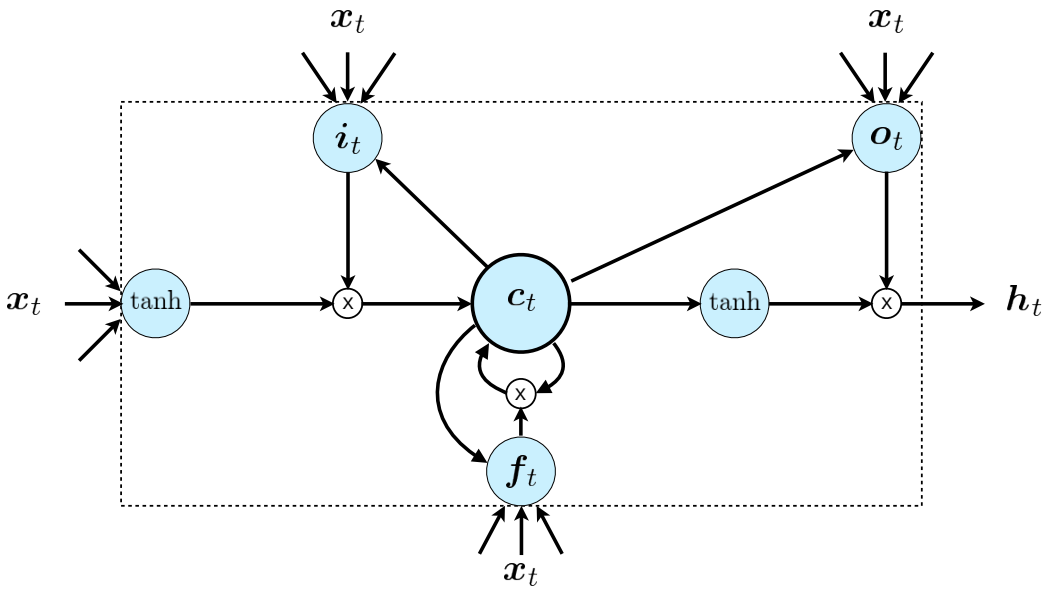


図 2.10: Long Short-term Memory Cell

きる．3 層の RNN を 2.9 に示す．ある時刻 t における入力特徴量を $\boldsymbol{x}_t$ とすると出力層 $\boldsymbol{y}_t$ は

$$\begin{aligned} \boldsymbol{h}_t &= \mathcal{H}(\boldsymbol{W}_{\boldsymbol{xh}}\boldsymbol{x}_t + \boldsymbol{W}_{\boldsymbol{hh}}\boldsymbol{h}_{t-1} + \boldsymbol{b}_{\boldsymbol{h}}) \\ \boldsymbol{y}_t &= \boldsymbol{W}_{\boldsymbol{hy}}\boldsymbol{h}_t + \boldsymbol{b}_{\boldsymbol{y}} \end{aligned} \tag{2.77}$$

として求める．ただし，$\mathcal{H}$ は隠れ層関数である．このように隠れ層を畳み込むことで，背後にある隠れ状態に相当するものが遷移する過程が暗にモデル化されることが期待される．また，隠れ層の関数 $\mathcal{H}$ は自由に設計することが可能であるため，より長時間の情報を伝搬させるための長時間と短時間の記憶が可能な Long Short Time Memory cell (LSTM) が提案されている [27]．図 2.10 に LSTM の概形を示す．LSTM は，複数の部位からなる複雑な構成を持ち，

$$i_t = \sigma(\boldsymbol{W}_{\boldsymbol{xi}}\boldsymbol{x}_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \tag{2.78}$$
$$f_t = \sigma(\boldsymbol{W}_{\boldsymbol{xf}}\boldsymbol{x}_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \tag{2.79}$$
$$c_t = f_t c_{t-1} + i_t \tanh(\boldsymbol{W}_{\boldsymbol{xc}}\boldsymbol{x}_t + W_{hc}h_{t-1} + b_c) \tag{2.80}$$
$$o_t = \sigma(\boldsymbol{W}_{\boldsymbol{xo}}\boldsymbol{x}_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \tag{2.81}$$
$$h_t = o_t \tanh(c_t) \tag{2.82}$$

となる．これにより，短時間の伝搬は従来の RNN の枠組みで考慮し，かつ長時間の情報を記憶することが可能となる．

また，音声認識では，後ろから伝搬させても何ら問題がないと考えられるため，単純に前

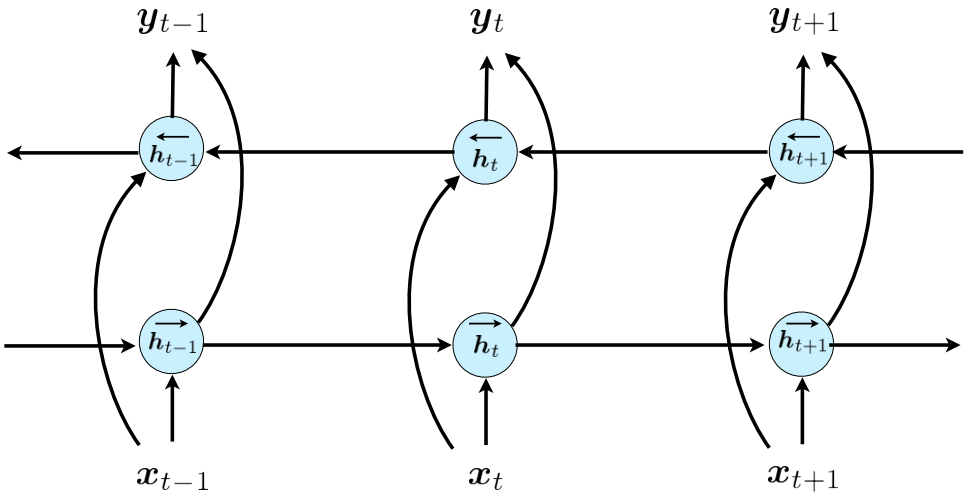


図 2.11: 双方向再帰ニューラルネットワーク

フレームの情報だけでなく，後ろのフレームの情報を逆向きに畳み込む双方向の再起ニューラルネットワークも提案されている [28]．これは，図 2.11 のように，前向きと後ろ向きの RNN を組み合わせることで実現できる．

$$\overrightarrow{\boldsymbol{h}}_t = f(\boldsymbol{W}_{\boldsymbol{x}\overrightarrow{\boldsymbol{h}}}\boldsymbol{x}_t + \boldsymbol{W}_{\overrightarrow{\boldsymbol{h}}\overrightarrow{\boldsymbol{h}}}\overrightarrow{\boldsymbol{h}}_{t-1} + \boldsymbol{b}_{\overrightarrow{\boldsymbol{h}}}) \tag{2.83}$$

$$\overleftarrow{\boldsymbol{h}}_t = f(\boldsymbol{W}_{\boldsymbol{x}\overleftarrow{\boldsymbol{h}}}\boldsymbol{x}_t + \boldsymbol{W}_{\overleftarrow{\boldsymbol{h}}\overleftarrow{\boldsymbol{h}}}\overleftarrow{\boldsymbol{h}}_{t-1} + \boldsymbol{b}_{\overleftarrow{\boldsymbol{h}}}) \tag{2.84}$$

$$\boldsymbol{y}_t = \boldsymbol{W}_{\overrightarrow{\boldsymbol{h}}\boldsymbol{y}}\overrightarrow{\boldsymbol{h}}_t + \boldsymbol{W}_{\overleftarrow{\boldsymbol{h}}\boldsymbol{y}}\overleftarrow{\boldsymbol{h}}_t + \boldsymbol{b}_{\boldsymbol{y}} \tag{2.85}$$

また，これらを組み合わせ深層ニューラルネットワークと同様に多層に拡張した deep recurrent neural network (DRNN) が提案されており，音声認識の音響モデルに利用した場合高い認識性能を示すことが報告されている [29].

2.4.3　言語モデルへの利用

近年，音響モデルにおいて述べた再帰ニューラルネットワークを用いた言語モデルも広く用いられるようになってきた．再帰ニューラルネットワークは，系列情報を考慮したニューラルネットワークモデルであり，理論上は離れた位置の情報を伝搬することができる．そのため，N-gram モデルと異なり離れた単語同士の関係性を記述することができると言われている．しかし，言語モデルにおいては再起ニューラルネットワーク単体の性能では N-gram モデルに劣る場合も多く，現在は N-gram モデルと併用するアプローチが主流となっている [30].

2.5　非言語情報の制御

音声に内在する情報は，大きく分けて 3 つに大別することができる．1 つ目は，発話内容，語彙などの言語的特徴であり，音声認識においてはこの言語情報を文字に表現することが目的である．2 つ目は，パラ言語情報であり，音の高さ，アクセント，韻律，抑揚などによって表現される発話スタイルや意図などである．これらは文字によっては表記することができない情報であり，一般に音声認識においては認識の対象とせず，特徴量抽出の段階で音の高さに関する情報を捨てることが多い．そして 3 つ目が非言語情報である．非言語情報は，背景雑音や話者や性別の違いなどにより混在する情報であり，音声認識性能の低下の一因となる．本節では，特徴量の正規化/適応，音響モデルの正規化学習，そして音響モデルのモデル適応の 3 つのドメインにおける非言語情報に関しての従来の制御技術について紹介する．

2.5.1　入力特徴量の正規化

非言語情報の違いに頑健な特徴量正規化としては，fMLLR や VTLN などのアプローチが代表的である．fMLLR は特徴量ドメインにおいて線形変換を行うことで，正規化を行う．その計算量の少なさから広く用いられている手法である．また，VTLN は話者の声道長に対応するパラメータを制御する [3]．これは周波数軸上でのウォーピングに対応し，これにより年齢の違いなどに起因する影響を取り除くことができる．

また，背景雑音に対するアプローチとして特徴量強調がある．背景雑音に対する頑健性の獲得は近年スマートフォンなどの携帯端末上での音声認識の機会の増加から，特に重要となっている．生成モデルを用いた雑音除去手法として代表的なものとして，Stereo-based Piecewise LInear Compensation for Environments (SPLICE) [31,32] や Vector Taylor Series (VTS) [33] などが挙げられる．また，ニューラルネットワークを用いた特徴量強調としては Denoising AutoEncoder (DAE) [34] がある．これらに関しては第 3 章で紹介する．

2.5.2　音響モデルの正規化学習/適応

音響モデルにおける非言語情報の制御に関するアプローチとしてモデル適応と，その適応性能の向上を目的とした正規化学習法が多く提案されている．

i)　正規化学習

先に述べたように，話者非依存の音響モデルは多種多様な話者，発話を集めて学習する．これは，例えば多種多様な話者の音素を GMM によりモデル化する場合，音素の違いではなく話者性の違いの影響によりクラスタリングされてしまう (図 2.12)．これは，モデル適応の初期モデルとしては不適切であり，具体的には共分散項が一話者のモデルに比べて広

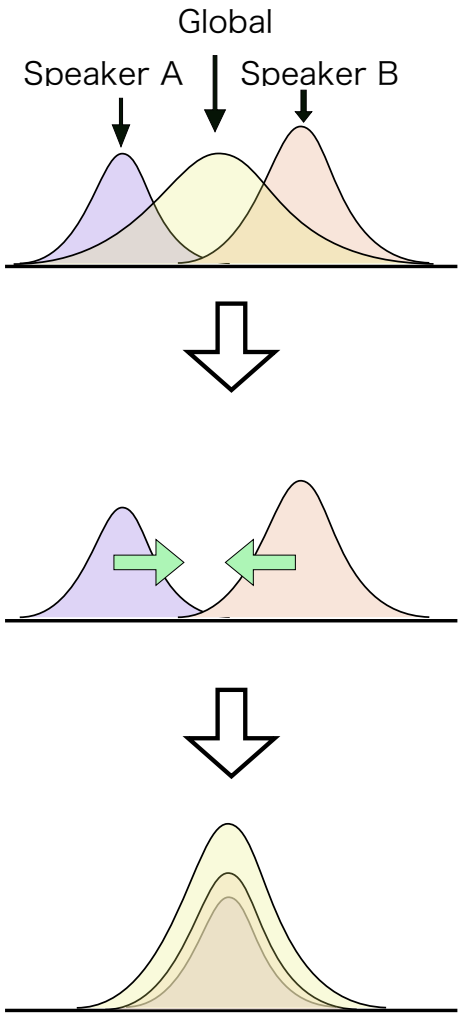


図 2.12: GMM ベースの音響モデルに対する話者正規化学習

がってしまうなどの影響がある．これを避けるため，話者の正規化とモデル化を繰り返し行い，話者性の違いではなく音素などの目的のラベルごとの話者非依存モデルを構築することが話者正規化学習 (Speaker adaptive training ; SAT) である，また，話者の違いだけではなく，環境の違いなどに起因するゆらぎを正規化することも可能である．

最も単純な SAT として，Constrained Maximum Likelihood Linear Regression (CMLLR) を音響モデルの学習に用いる特徴量に対して行うアプローチがある．CMLLR は GMM 音響モデルに対して平均と分散に数学的に対応した変換をかけることでモデルを適応すること手法である．これは fMLLR と同値であるため，特徴量に変換を行うことで GMM に限らず様々なモデルにおいても実現することができる．

DNN 音響モデルに対する正規化学習手法も提案されている [17]．GMM 音響モデルの場合と同様に，DNN の入力部分で fMLLR を行うことで話者依存性を予め打ち消すことで正

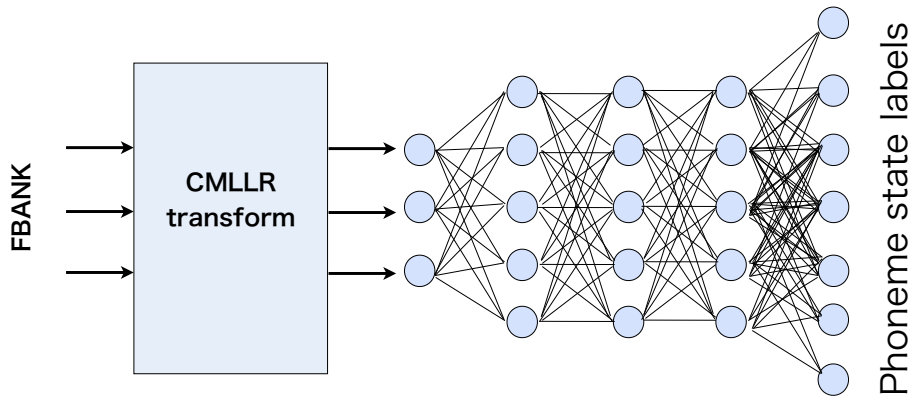


図 2.13: CMLLR を用いたニューラルネットワーク音響モデルの話者正規化学習

規化学習された DNN 音響モデルを実現することができる（図 2.13）．なお，この正規化学習を行ったモデルに対して適応する際は，この fMLLR パラメータを推定することで話者適応を行うことができる．

ii)　モデル適応

音響モデルのモデル適応は，特徴量ドメインにおける正規化と裏表の関係にあると言える．GMM 音響モデルに対する代表的なモデル適応手法として MLLR 適応 [6] が挙げられる．これは，入力特徴量に対して最尤となるように GMM 音響モデルのパラメータを更新することで，入力特徴量の認識に適したモデルに修正することができる．これは，GMM 音響モデルのパラメータと音響事象との対応が直感的であり，例えばケプストラム空間における線形変換のみでも話者性の制御が可能であることに由来する．

また，DNN 音響モデルに対するモデル適応手法も多く提案されている．例えば，話者性を表す特徴量である i-vector を用いたモデル適応手法が提案されている [12]．i-vector は話者認識などの分野において主流である特徴量であり，話者性をよく表現していると考えられる．この i-vector を入力特徴量として付与することにより，DNN に話者の違いに対する頑健性を獲得させることを期待している．(図 2.14）この手法は，入力特徴量から新たな特徴量を設計してそれを利用して識別するという非常に単純な手法ではあるが，効果として大きいのは，通常の DNN は入力として数フレームのセグメント特徴のみから識別するのに対し，i-vector は発話全体から話者に相当する特徴量を抽出するために，情報量が増えるという点であると考えられる．

また，GMM 音響モデルと異なり，DNN 音響モデルはモデルパラメータが非常に多いという特徴がある．これにより，適応データが少量の場合に容易に過学習が生じるという問

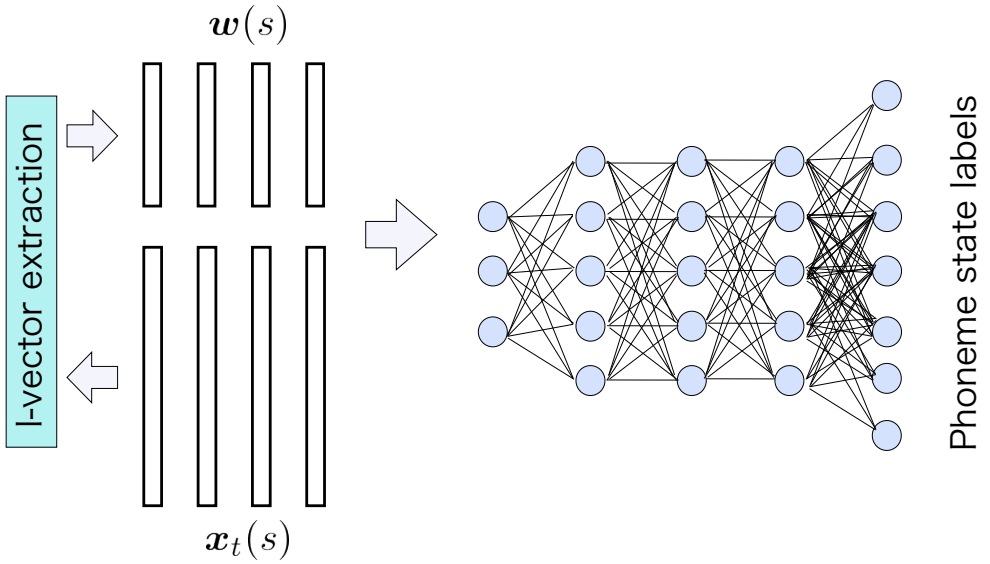


図 2.14: i-vector based

題がある．これを抑えるため，適応の際の制御するモデルパラメータを制限するという試みもある．特異値分解 (Singular value decomposition ; SVD) を利用する手法はその 1 つである [16]．パラメータ行列を $\boldsymbol{A}^{m\times n}$ とすると，これを SVD により

$$\boldsymbol{A}^{m\times n} = \boldsymbol{U}^{m\times n}\boldsymbol{\Sigma}^{n\times n}(\boldsymbol{V}^{n\times n})^{\top} \tag{2.86}$$

として分解することができる．ここで，$\boldsymbol{\Sigma}^{n\times n}$ は対角行列であり，対角成分に $\boldsymbol{A}^{m\times n}$ の特異値が並んでいる．また，$\boldsymbol{A}^{m\times n}$ が非常にスパースな行列な場合，特異値の数 k が n に比べて非常に小さかった場合は，

$$\boldsymbol{A}^{m\times n} \approx \boldsymbol{U}^{m\times k}\boldsymbol{N}^{k\times n} \tag{2.87}$$

として近似することが可能である．

これを用いて，学習された DNN のある層間の線形変換パラメータ $\boldsymbol{A}(l)^{m\times n}$ を SVD により分解することができる．これはつまり，ボトルネック層を持つ多層パーセプトロンに展開できることを示しており，SVD Bottleneck adaptation は，図 2.15 のように適応行列 $\boldsymbol{S}(l)^{k\times k}$ を

$$\begin{aligned}\boldsymbol{A}(l)^{m\times n} &\approx \boldsymbol{U}(l)^{m\times k}\boldsymbol{N}(l)^{k\times n} \\ &= \boldsymbol{U}(l)^{m\times k}\boldsymbol{S}(l)^{k\times k}\boldsymbol{N}(l)^{k\times n}\end{aligned} \tag{2.88}$$

として挿入することで，話者適応を行う．学習は，まず，通常の DNN の学習を行った後，SVD でパラメータを分解する．その後，$\boldsymbol{S}(l)^{k\times k}$ の層を挿入する．なお，初期値としては単位行列を用いる．適応時は back propagation による再学習により適応行列 $\boldsymbol{S}(l)^{k\times k}$ のパ

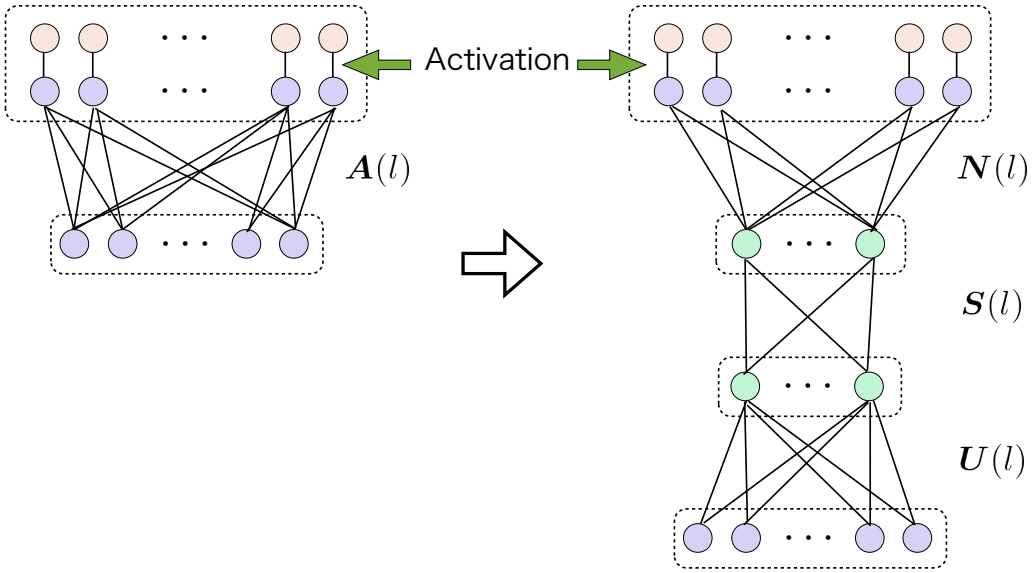


図 2.15: 特異値分解を用いた音響モデル適応

ラメータをアップデートする．

この手法の利点は，少ない適応データに対して通常のニューラルネットよりもより少ないパラメータを更新することで過学習を防ぐことができるだけでない．変換行列のランクが低い場合，モデルパラメータ自体が減少するため，開発コストの削減にも繋がるのである．

2.6　ニューラルネットワーク音声認識における非言語情報の制御に関する課題

現在の音声認識システムにおいて，ニューラルネットワークは単純な認識性能のみであれば高い精度を得ることができるものの，雑音除去やモデル適応などの非言語情報の制御による認識性能の向上では，生成モデルに対する技術程のゲインを得ることができていない．この理由として，GMM とニューラルネットワークの性質の違いにより，従来の非言語情報の制御に関する知見が活かし辛いということが挙げられる．一般的に GMM をベースとした手法はモデルパラメータの持つ意味が直感に沿っているという性質を持つ．これは，GMM/HMM でモデル化する単位が音素単位等であるため，その平均分散が，音素の平均的な特徴とばらつきを表す，などである．この性質はモデルパラメータを調整する際に非常に有益であり，例えば新たな入力話者に対するモデル適応の際に，この特徴量をア

フィン変換により適応することで比較的容易に話者適応が可能となる．なお，これは，話者の違いは GMM を構築する特徴量空間，一般にはケプストラム空間におけるアフィン変換で表されるという理論的な知見にも基づくものである．

一方，近年主流となっているニューラルネットワークは，非常に複雑な構造を持つ．複雑な構造によって多様な関数を近似することができることがニューラルネットワークの利点でもあるが，その複雑さから GMM のようなパラメータの意味づけが困難である．その結果，非言語情報の制御自体は音声認識において長年の課題であり，多くの研究が報告されてきたにも関わらず，ある種手探りな形を取らざるを得ず，効率的な非言語情報の制御が実現できているとは言えない．そこでこれを打開するため，本論文では，音響特徴量に内在する非言語情報の制御として重要となる特徴量ドメインと音響モデルドメインにおいて，従来の生成的アプローチとニューラルネットワークの識別的アプローチの融合を行う．これにより，従来の知見を活かしたアプローチにより認識性能の向上を図る．

2.6.1　特徴量ドメインにおける生成的アプローチと識別的アプローチの融合

特徴量ドメインにおける非言語情報の制御は，特徴量自体の設計と特徴量の話者適応，雑音適応，特徴量強調等が挙げられる．本論文では，特徴量強調において，GMM に基づく生成的アプローチとニューラルネットに基づく識別的アプローチを融合した手法を提案する．特徴量強調では，特徴量空間のクラスタリングと，その識別が重要となる．特徴量の分布のクラスタリングでは従来の GMM ベースのアプローチを用いることで，音声特徴量の分布に対する知見を用い，空間の識別では識別性能の高いニューラルネットを用いることで，各々の利点を組み合わせる．第 3 章において，詳細を述べる．

2.6.2　音響モデルドメインにおける生成的アプローチと識別的アプローチの融合

音響モデルドメインにおける非言語情報の制御に関する代表的なアプローチは，モデル適応とそれに付随する正規化学習である．本論文では，従来の GMM 音響モデルにおいてモデル適応性能の向上に寄与する正規化学習法を，ニューラルネットワーク音響モデルにおいて実現する．この際，話者・環境の情報を表す話者・環境コードを生成的に推定することで，効果的な正規化学習の実現とこれを用いたモデル適応性能の向上を図る．第 4 章において，詳細を述べる．

2.6.3　ニューラルネットワークを用いた非言語情報の違いに頑健な特徴量表現の実現とその利用

ニューラルネットワークと従来の音声学的知見の融合を考えた場合，従来のモデルの性質のみを考慮したアプローチのみでは不十分である．そこで，本論文では話者非依存な特徴量として提案されていた構造的表象に着目する．ケプストラム空間上において，話者の性別や年齢の違いはアフィン変換として表現することが可能であることが知られている [35]．音声の構造的表象は各音響イベント間の分布間距離を利用した特徴量であるが，f-divergence に代表される分布間距離はアフィン変換に対して不変である．そのため，音声の構造的表象は話者性の違いに対して頑健であると言える [36]．この性質を利用して，音声構造的表象を制約としてモデルのパラメータの更新を行う手法も提案されている [37,38]．本論文ではこの音声の構造的表象を構成する分布間距離をニューラルネットワークにより識別的に計算する手法を提案する．特徴量ドメインと音響モデルドメインにおいて比較を行い，提案法の有効性を検討する．第 5 章において，詳細を述べる．

2.7　まとめ

本章ではまず，音声認識の基礎技術とニューラルネットワークの基礎，そしてその音声認識への利用法について説明した．ニューラルネットワークの音声認識への利用は非常に多岐に渡るため，本章では，その基本となるアプローチのみを紹介し，関連する要素技術については該当する章にて関連研究として紹介する．また，非言語情報について説明し音声認識においてその制御の重要性を解説した．それを踏まえ，ニューラルネットワーク音声認識における非言語情報の制御に関する課題について述べ，従来の生成的アプローチ，音声学的知見を導入することの重要性を示した．

第3章

雑音環境下音声認識のための
ニューラルネットワークを用いた
識別的区分線形変換

3.1　はじめに

前章では，音声認識技術とニューラルネットワークの基礎について紹介し，現在のニューラルネットワーク音声認識において非言語情報の制御に関する課題について述べた．本章ではこれを踏まえ，特徴量ドメインにおける非言語情報の制御に関する要素技術の一つである特徴量強調において，GMM ベースと DNN ベースのアプローチを融合した手法の有効性について検討する．

現在の音声認識システムは雑音の影響が小さな静音環境においては高い認識性能を示す一方，雑音環境下では未だにその性能は十分とは言えない．雑音は話者が意図せずに音声に影響を与える非言語情報の一つであり，音声認識性能を低下させる大きな要因となる．そのため，ハンズフリーな音声認識システムの構築等を考えた場合，耐雑音性を高めることは非常に重要な課題と言える [39].

特徴量強調は耐雑音性を確保するためのフロントエンド処理の 1 つであり，Stereo-based Piecewise LInear Compensation for Environments (SPLICE) [31,32] や Vector Taylor Series (VTS) [33]，Stereo-based Stochastic Mapping (SSM) [40] などの区分的線形変換をベースとした手法や，ニューラルネットワークを用いた Denoising AutoEncoder (DAE) [34]，事例ベースの特徴量強調手法 [41] などが提案されている．これらは，波形ドメインではなく，音声特徴量のドメインにおいて，観測された雑音環境下における音声特徴量から対応する静音環境下の音声特徴量を再現する．

SPLICE に代表される区分的線形変換法は，モデルを 2 つの段階に分けて考えることができる．まず，雑音環境下における音声特徴量空間をガウス混合モデル (Gaussian Mixture Model; GMM) でモデル化し，観測された音声特徴量の各フレームに対する，GMM の各要素分布からの事後確率を計算する．次に，得られた事後確率を重みとして，観測された雑音環境下における音声特徴量からの線形変換の足し合わせにより静音環境下における音声特徴量を推定する．

GMM は特徴量空間をマハラノビス距離を用いて確率的にクラスタリングすることに相当する．従って，音声特徴量の GMM の各要素分布からの寄与率（事後確率）によって重み付けを行うことは，特徴量空間を確率的に領域分割することと等価である．SPLICE などの区分的線形変換は，この分割された各領域毎では入出力の対応が線形性を有すること（局所線形性）を仮定している．

局所線形性の仮定を考えた場合，理想的には静音環境下での特徴量空間の領域分割と雑音環境下での特徴量空間の領域分割が一致していることが望ましい．しかし，一般に雑音環境下においては，重畳されている雑音の影響で特徴量空間が縮退してしまう．そのため，例えば雑音が大きな環境では，音声特徴量が雑音によりマスクされてしまい入力特徴量空間を GMM でモデル化すると，各要素分布は雑音の種類や大きさに対応してモデル化される可能性が残る．雑音の種類による特徴量への影響は非線形であると考えられるため，

SPLICE のように雑音環境下における領域分割を基準にした場合，局所線形性の仮定が不適切となる場合が考えられる．

そこで，この問題解決へのアプローチとして，REgularized piecewise linear mapping with DIscriminative region weighting And Long-span features (REDIAL) が提案されている [42,43]．REDIAL は線形判別分析 (Linear Discriminant Analysis; LDA) により，領域分割が，静音環境下における音声特徴量の領域分割とより一致するような空間に特徴量を射影し，GMM によりモデル化する．しかし，LDA は線形変換であるため，雑音環境下における音声特徴量と静音環境における音声特徴量の複雑な関係を適切にモデル化できていない可能性がある．

一方，近年は深層ニューラルネットワーク（Deep Neural Network ; DNN) [22] の発展に伴い，ニューラルネットワークを特徴量強調に利用した手法も提案されている． それらのうちの 1 つである，DAE はニューラルネットワークを回帰モデルとして用い，観測された音声特徴量から対応する静音環境下の音声特徴量を非線形かつ直接的に推定する．さらに，DAE を多層にした Deep Denoising AutoEncoder (DDAE) は特に雑音環境が既知の条件において，高い精度で静音環境における音声特徴量を推定することが報告されている [34]．しかし，雑音環境が未知の場合では性能が低下することも分かっており，特定の環境に特化した強調となる傾向がある [44]．これは，ニューラルネットワークによる複雑な非線形変換によって，雑音環境下の音声特徴量空間を綿密にモデル化してしまうためと考えられる．

そこで，本提案手法はニューラルネットワークにより，静音環境における領域分割によって付与される領域ラベルを観測された雑音環境下の音声特徴量から識別する．その後，従来の区分的線形変換法と同様に，各領域毎の線形変換を事後確率による重みづけで足し合わせ，対応する静音環境下における音声特徴量を推定する．提案法では，ニューラルネットワークを回帰モデルとしてではなく，識別モデルとして扱うことで，ニューラルネットワークの高い識別性と雑音に対する汎化性の高い区分線形変換の両立をはかる．

3.2　関連研究

本節では，区分的線形変換手法である SPLICE とその拡張である REDIAL について説明を行う．また，ニューラルネットワークを用いた代表的な雑音抑圧手法である DAE についても言及する．

3.2.1　SPLICE

SPLICE は特徴量強調手法の 1 つであり，区分的線形変換によって，観測された時刻 $t \in T$ における雑音環境下における音声特徴量 $\boldsymbol{y}_t$ から，それに対応する静音環境下における音声

表 3.1: 雑音環境下における音声特徴量を用いて領域分割を行った場合の SPLICE と，正解の静音環境下における音声特徴量を用い領域分割を行った場合の SPLICE(oracle) の単語誤り率 (%)

	SPLICE	SPLICE (oracle)
clean	0.57	0.69
SNR 20	1.08	0.76
SNR 15	1.99	0.70
SNR 10	4.65	0.77
SNR 5	16.76	0.93
SNR 0	49.96	0.95
SNR -5	81.46	1.13
Average	14.89	0.82

特徴量 $\boldsymbol{x}_t$ を式 (3.1) で推定する．

$$\hat{\boldsymbol{x}}_t = \sum_{i=1}^{I} p(i|\boldsymbol{y}_t) \boldsymbol{A}_i \begin{bmatrix} 1 \\ \boldsymbol{y}_t \end{bmatrix} \tag{3.1}$$

ここで，$i \in I$ は混合のインデクスであり，$\boldsymbol{A}_i$ は各混合に対応する線形変換行列である．

SPLICE では，雑音環境下における音声特徴量の確率密度関数 $p(\boldsymbol{y})$ を GMM でモデル化する．混合 i における平均 $\boldsymbol{\mu}_i^{\boldsymbol{y}}$，分散 $\boldsymbol{\sigma}_i^{\boldsymbol{y}}$ のガウス分布を $\mathcal{N}(\boldsymbol{y}_t; \boldsymbol{\mu}_i^{\boldsymbol{y}}, \boldsymbol{\sigma}_i^{\boldsymbol{y}})$ とすると，GMM からの特徴量 $\boldsymbol{y}_t$ の出力確率は，各ガウス分布の重みを $\pi_i^{\boldsymbol{y}}$ とした場合

$$p(\boldsymbol{y}_t) = \sum_{i=1}^{I} \pi_i^{\boldsymbol{y}} \mathcal{N}(\boldsymbol{y}_t; \boldsymbol{\mu}_i^{\boldsymbol{y}}, \boldsymbol{\sigma}_i^{\boldsymbol{y}}) \tag{3.2}$$

となる．これを用いて，雑音環境下の音声特徴量の各フレームに対する GMM の要素分布の寄与率を事後確率 $p(i|\boldsymbol{y}_t)$ として計算する．

$$p(i|\boldsymbol{y}_t) = \frac{p(\boldsymbol{y}_t|i)p(i)}{\sum_{i'=1}^{I} p(\boldsymbol{y}_t|i')p(i')} \tag{3.3}$$

式 (3.2) のように，SPLICE では雑音環境下の音声特徴量のみを用いて領域分割に用いる GMM の学習を行う．しかし，雑音環境においては重畳されている雑音の影響により音声の特徴量空間が縮退する．そのため，雑音環境下の音声特徴量で学習した GMM による領域分割では，各要素分布が雑音の種類や大きさに対応してしまうことが考えられる．理想的には静音環境下での特徴量空間の領域分割と雑音環境下での特徴量空間の領域分割が一致していることが望ましいと予想される．

そこで，MFCC によって構築された音声特徴量空間において，静音環境下，もしくは雑

音環境下のそれぞれで，GMM によるモデル化と事後確率による重み付けを行った場合の比較を行った．表 3.1 は雑音環境下音声認識用のデータベースである Aurora-2 [45, 46] の既知雑音条件における評価セットであるセット A を用いた単語誤り率である．SPLICE (oracle) は静音環境下におけるモデル化と観測された雑音環境下音声に対応する静音環境下の音声を領域分割にのみ用いたものである．SPLICE (oracle) の場合，

$$\hat{\boldsymbol{x}}_t = \sum_{i^*=1}^{I^*} p(i^*|\boldsymbol{x}_t)\boldsymbol{A}_{i^*}\begin{bmatrix}1\\ \boldsymbol{y}_t\end{bmatrix} \tag{3.4}$$

$$p(i^*|\boldsymbol{x}_t) = \frac{p(\boldsymbol{x}_t|i^*)p(i^*)}{\sum_{i^{*\prime}=1}^{I^*} p(\boldsymbol{x}_t|i^{*\prime})p(i^{*\prime})} \tag{3.5}$$

として静音環境下における音声特徴量の要素分布 i^* に対する事後確率 $p(i^*|\boldsymbol{x}_t)$ を計算する．領域分割にのみ対応する静音環境下の音声特徴量を用いた SPLICE (oracle) の結果は，雑音の大きな環境でも頑健に対応する静音環境下の音声特徴量を推定できている．一方，これと比較した場合，通常の SPLICE の認識性能は特に雑音の大きな環境下において低下している．これは，雑音が大きくなった場合，特徴量空間の縮退により，要素分布が静音環境下における領域分割と異なる意味を持つためと考えられる．

3.2.2 REDIAL

静音環境下の音声特徴量をモデル化した GMM の混合インデクスを $k \in K$ とした時，学習データセットを $\{\{p(k|\boldsymbol{x}_t)\}_{k=1\ldots K}, \boldsymbol{d}_t\}_{t=1\ldots T}$ とする．ここで $\boldsymbol{d}_t$ は時刻 t における該当フレーム $\boldsymbol{y}_t$ とその前後 s フレームを連結した入力特徴量ベクトルであり，

$$\boldsymbol{d}_t = [\boldsymbol{y}_{t-s}^\top, \ldots, \boldsymbol{y}_{t-1}^\top, \boldsymbol{y}_t^\top, \boldsymbol{y}_{t+1}^\top, \ldots, \boldsymbol{y}_{t+s}^\top]^\top \tag{3.6}$$

である．学習段階では，静音環境下の特徴量空間で計算された事後確率と，観測特徴量から得られる事後確率が近くなるように，観測特徴量空間を次元圧縮する．次元圧縮行列 $\boldsymbol{L}$ は，ラベルを確率的に利用した LDA によって学習することができる．

$$\hat{\boldsymbol{L}} = \operatorname*{argmin}_{\boldsymbol{L}} \frac{\boldsymbol{L}^\top \boldsymbol{\Sigma}^w \boldsymbol{L}}{\boldsymbol{L}^\top \boldsymbol{\Sigma}^b \boldsymbol{L}} \tag{3.7}$$

$$\boldsymbol{\Sigma}^w = \sum_{k=1}^{K}\sum_{t=1}^{T} p(k|\boldsymbol{x}_t)\left(\boldsymbol{d}_t - \boldsymbol{\mu}_k^w\right)\left(\boldsymbol{d}_t - \boldsymbol{\mu}_k^w\right)^\top \tag{3.8}$$

$$\boldsymbol{\Sigma}^b = \sum_{k=1}^{K}\left(\sum_{t=1}^{T} p(k|\boldsymbol{x}_t)\right) \times \left(\boldsymbol{\mu}_k^w - \frac{\sum_{t=1}^{T}\boldsymbol{d}_t}{T}\right)\left(\boldsymbol{\mu}_k^w - \frac{\sum_{t=1}^{T}\boldsymbol{d}_t}{T}\right)^\top \tag{3.9}$$

$$\boldsymbol{\mu}_k^w = \frac{1}{\sum_{t=1}^{T} p(k|\boldsymbol{x}_t)} \sum_{t=1}^{T} p(k|\boldsymbol{x}_t)\boldsymbol{d}_t \tag{3.10}$$

次に，LDA により次元圧縮を行ったベクトル $\boldsymbol{v}_t = \hat{\boldsymbol{L}}\boldsymbol{d}_t$ を用いて K^* 混合の GMM を学習する．

$$p(\boldsymbol{v}_t) = \sum_{k^*=1}^{K^*} \pi_{k^*}^{\boldsymbol{v}} \mathcal{N}(\boldsymbol{v}_t; \boldsymbol{\mu}_{k^*}^{\boldsymbol{v}}, \boldsymbol{\sigma}_{k^*}^{\boldsymbol{v}}) \tag{3.11}$$

入力特徴量が得られた際の静音環境下における音声特徴量状態インデクス $k^* \in K^*$ に対する事後確率 $p(k^*|\boldsymbol{y}_t)$ を

$$p(k^*|\boldsymbol{y}_t) \simeq p(k^*|\boldsymbol{v}_t) = \frac{p(\boldsymbol{v}_t|k^*)p(k^*)}{\sum_{k^*=1}^{K^*} p(\boldsymbol{v}_t|k^{*\prime})p(k^{*\prime})} \tag{3.12}$$

として計算する．最終的に得られた事後確率を重みとして用いた区分的線形変換により静音環境下における音声特徴量を

$$\hat{\boldsymbol{x}}_t = \sum_{k^*=1}^{K^*} p(k^*|\boldsymbol{y}_t)\boldsymbol{A}_{k^*}\boldsymbol{e}_t \tag{3.13}$$

$$\boldsymbol{e}_t = [1, \boldsymbol{y}_{t-u}^\top, \ldots, \boldsymbol{y}_{t-1}^\top, \boldsymbol{y}_t^\top, \boldsymbol{y}_{t+1}^\top, \ldots, \boldsymbol{y}_{t+u}^\top]^\top \tag{3.14}$$

として推定する．ただし，$\boldsymbol{A}_{k^*}$ は，要素分布 k^* に対応する線形変換行列であり，$\boldsymbol{e}_t$ は当該フレームと，その前後 u フレームを連結した拡張行列である．

$$\boldsymbol{e}_t = [1, \boldsymbol{y}_{t-u}^\top, \ldots, \boldsymbol{y}_{t-1}^\top, \boldsymbol{y}_t^\top, \boldsymbol{y}_{t+1}^\top, \ldots, \boldsymbol{y}_{t+u}^\top]^\top \tag{3.15}$$

なお，$\boldsymbol{A}_{k^*}$ は重み付き最小二乗誤差基準で学習する．

$$\hat{\boldsymbol{A}}_{k^*} = \underset{\boldsymbol{A}_{k^*}}{\operatorname{argmin}} \sum_{t=1}^{T} p(k|\boldsymbol{y}_t)||\boldsymbol{x}_t - \boldsymbol{A}_{k^*}\boldsymbol{e}_t||^2 \tag{3.16}$$

これは解析解を得ることができ，$\boldsymbol{A}_{k^*}$ は，

$$\hat{\boldsymbol{A}}_{k^*} = \boldsymbol{X}\boldsymbol{P}\boldsymbol{E}^\top(\boldsymbol{E}\boldsymbol{P}\boldsymbol{E}^\top)^{-1} \tag{3.17}$$

として計算することができ，ここで，特徴量の次元数を D とすると，$\boldsymbol{X} \in \mathcal{R}^{D\times T}$ と $\boldsymbol{E} \in \mathcal{R}^{(D(2u+1)+1)\times T}$ はそれぞれ出力と入力特徴量の拡張ベクトルを並べたデータ行列，$\boldsymbol{P} \in \mathcal{R}^{T\times T}$ は $p(k^*|\boldsymbol{y}_t)$ を対角に並べた行列である．なお，$\boldsymbol{A}_{k^*}$ は非常に大きな行列になるため，学習の際に正則化を導入する．

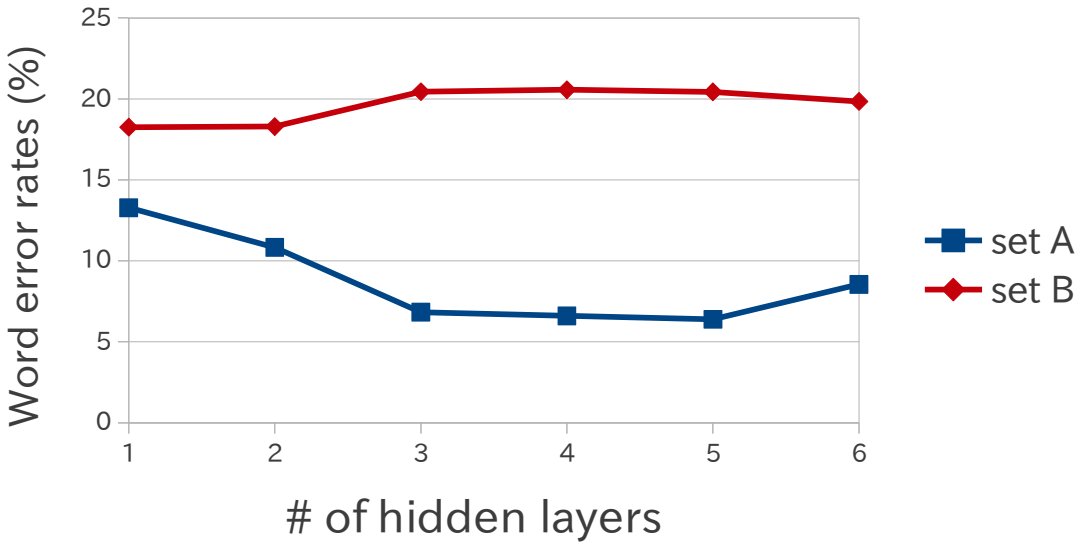


図 3.1: Aurora-2 データベースにおける，隠れ層の数を変化させた場合の DDAE の単語誤り率の変化

REDIAL は LDA によって雑音の影響を低減することで，雑音の大きな環境における観測特徴量の領域分割と，それに対応する静音特徴量の領域分割のミスマッチを減らし，静音特徴量の推定精度を向上させる．しかし，LDA はあくまで線形変換であるため，静音環境下における特徴量空間と雑音環境下における特徴量空間の非線形な対応を適切にモデル化できているとは言えない．

3.2.3　DAE

DAE はニューラルネットワークを用いて雑音環境下における音声特徴量から静音環境下における音声特徴量を直接的に推定する手法である．DDAE は DAE を多層にしたものであり，時刻 t の静音環境下における音声特徴量 $\boldsymbol{x}_t$ を

$$\hat{\boldsymbol{x}}_t = \boldsymbol{U}\boldsymbol{h}^{(n)}(\boldsymbol{d}_t) + \boldsymbol{c} \tag{3.18}$$

$$\boldsymbol{h}^{(n)}(\boldsymbol{d}_t) = \boldsymbol{\sigma}(\boldsymbol{W}^{(n)}\boldsymbol{h}^{(n-1)}(\boldsymbol{d}_t) + \boldsymbol{b}^{(n)}) \tag{3.19}$$

$$\boldsymbol{h}^{(1)}(\boldsymbol{d}_t) = \boldsymbol{\sigma}(\boldsymbol{W}^{(1)}\boldsymbol{d}_t + \boldsymbol{b}^{(1)}) \tag{3.20}$$

として推定する．ここで，$n \in N$ は中間層のインデクスであり，$\boldsymbol{U}, \boldsymbol{W}^{(n)}$ は重み行列，$\boldsymbol{c}, \boldsymbol{b}^{(n)}$ はバイアス項である．また，$\boldsymbol{h}^{(n)}$ は中間層の出力を表す．

図 3.1 は中間層の層数 N を変化した際の Aurora-2 における単語誤り率 (%) を示したものである．中間層の層数以外の詳細な実験条件は 4 章の実験と共通であるため，そちらを

参照して頂きたい．セット A とセット B はそれぞれ既知雑音条件と未知雑音条件のテストセットである．なお，ニューラルネットワークの各隠れ層のノード数は予備実験により 1024 で統一した．既知雑音条件では層の数を増やすと単語誤り率が減少するが，未知雑音環境では逆に単語誤り率が僅かながら増加し，既知雑音の場合との差が大きくなる．これは，ニューラルネットワークを回帰モデルとして利用した場合，その複雑な非線形変換によって，雑音環境下の音声特徴量空間を綿密にモデル化してしまうためと考えられる．

3.3　DNN に基づく領域分割を用いた区分線形変換

前章で示した通り，SPLICE に代表される区分的線形変換をベースとした特徴量強調手法は，局所線形性の仮定により，静音環境における特徴量空間の領域分割と雑音環境における特徴量空間の領域分割の対応がとれていることが望ましい．しかし，雑音の大きな環境では，雑音の影響により音声の特徴量空間が縮退するため，GMM による特徴量空間のクラスタリングでは，雑音の種類や大きさに各要素分布が対応することが想定される．従って，REDIAL のように観測特徴量空間をモデル化する際に識別的な基準を導入することが効果的だと考えられる．しかし，雑音の影響は非線形であると予想されるため，LDA のような線形変換をベースとしたモデルでは十分とは言えない．

一方，ニューラルネットワークを回帰モデルとして用い，観測特徴量から静音環境下の音声特徴量を推定するモデルは，高い認識性能を示すものの，未知雑音条件と既知雑音条件における単語誤り率には大きな差が存在する．これは，ニューラルネットワークのもつ非線形性により，雑音環境下の音声特徴量空間を綿密にモデル化してしまうことで，特定の環境に特化した傾向となるためだと考えられる．

そこで，提案手法では，観測特徴量から，それに対応する静音環境下における音声特徴量に対して最も高い寄与率を持つ要素分布をニューラルネットワークにより識別し，得られた事後確率を重みとした区分的線形変換により静音環境下の音声特徴量を推定する．これによって，ニューラルネットワークのもつ非線形性により，観測された雑音環境下における特徴量から，対応する静音環境下の特徴量の領域分割を高い精度で再現することが可能となる．また，ニューラルネットワーク自体は回帰モデルではなく領域の識別モデルとして機能するため，特徴量変換そのものは各要素分布である正規分布のもつ汎化性により，未知雑音環境下においても頑健に静音環境下の音声特徴量を推定することが可能となる．

今，時刻 t における D 次元の静音環境下における音声特徴量 $\boldsymbol{x}_t$ とそれに対応する雑音環境下における音声特徴量 $\boldsymbol{y}_t$ のパラレルデータ $\{(\boldsymbol{x}_t, \boldsymbol{y}_t)\}$ を考える．図 3.2 に学習段階の流れ

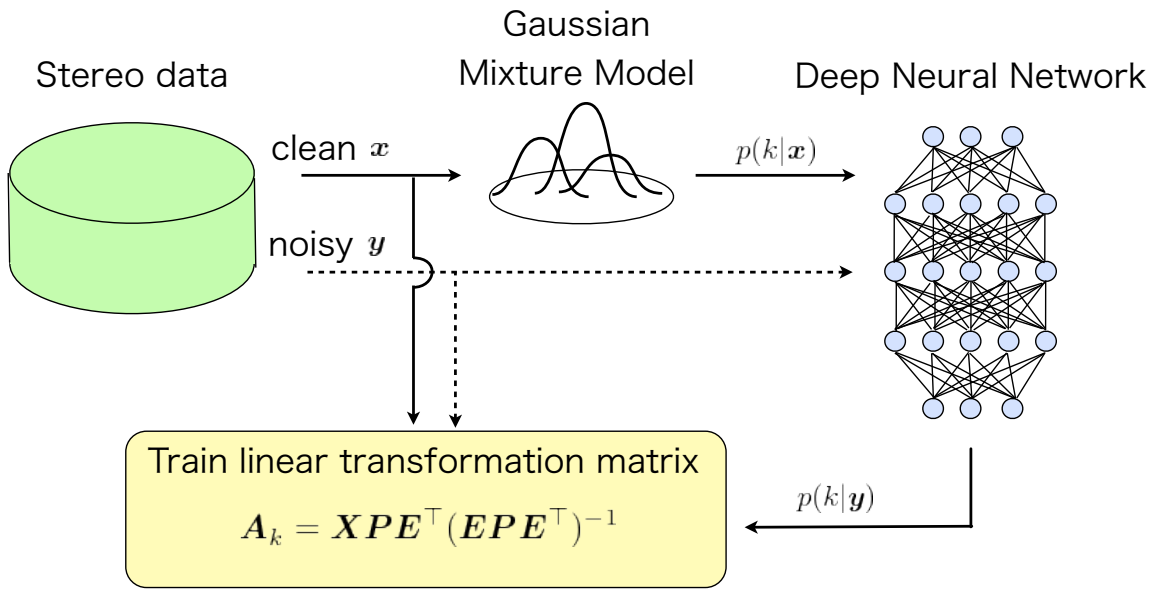


図 3.2: 提案手法の学習時の流れ

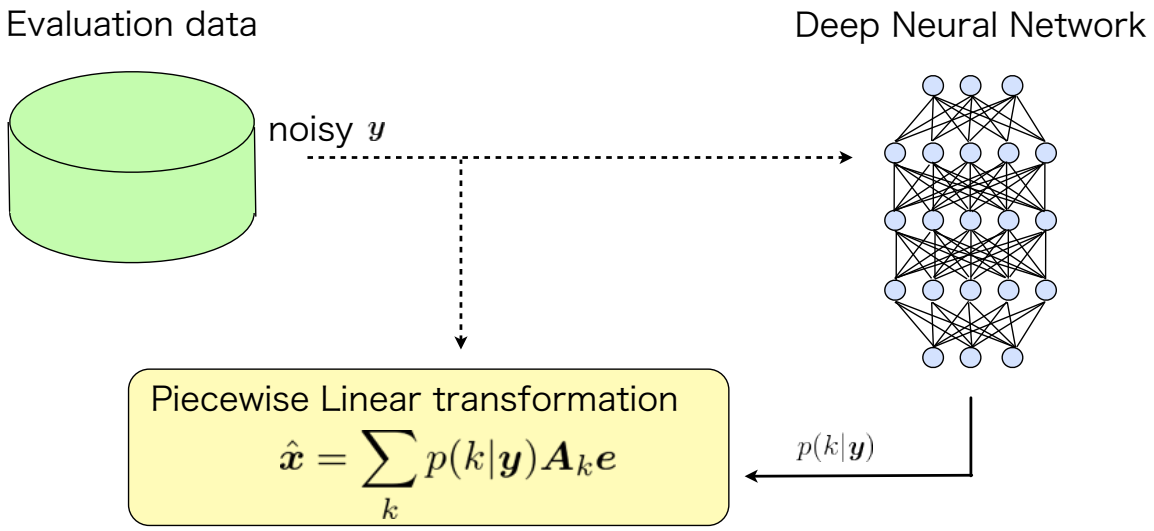


図 3.3: 提案手法の認識時の流れ

を示す．まず，静音環境下における音声特徴量の確率密度関数 $p(\boldsymbol{x})$ を GMM で学習する．

$$p(\boldsymbol{x}_t) = \sum_{k=1}^{K} p(k)p(\boldsymbol{x}_t|k) \tag{3.21}$$

$$p(\boldsymbol{x}_t|k) = \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{\mu}_k^{\boldsymbol{x}}, \boldsymbol{\sigma}_k^{\boldsymbol{x}}) \tag{3.22}$$

$$p(k) = \pi_k^{\boldsymbol{x}} \tag{3.23}$$

これを用いて，GMM の各要素分布 k に対する事後確率を，

$$p(k|\boldsymbol{x}_t) = \frac{p(\boldsymbol{x}_t|k)p(k)}{\sum_{k'=1}^{K} p(\boldsymbol{x}_t|k')p(k')} \tag{3.24}$$

と表すことができる．

一方，認識時には，静音環境下における音声特徴量 $\boldsymbol{x}_t$ は観測できないため，観測された雑音環境下における音声特徴量 $\boldsymbol{y}_t$ から，それに対応する静音環境下における音声特徴量の，各要素分布に対する事後確率 $p(k|\boldsymbol{y}_t)$ を推定する必要がある．そこで，静音環境下における音声特徴量に対して寄与率の最も大きい要素分布を観測された音声特徴量から識別するニューラルネットワークを学習する．

$$p(k|\boldsymbol{y}_t) \simeq p(k|\boldsymbol{d}_t) = \mathrm{softmax}_k(\boldsymbol{V}h^{(n)}(\boldsymbol{d}_t) + \boldsymbol{c}) \tag{3.25}$$

$$h^{(n)}(\boldsymbol{d}_t) = \boldsymbol{\sigma}(\boldsymbol{W}^{(n)}h^{(n-1)}(\boldsymbol{d}_t) + \boldsymbol{b}^{(n)}) \tag{3.26}$$

$$h^{(1)}(\boldsymbol{d}_t) = \boldsymbol{\sigma}(\boldsymbol{W}^{(1)}\boldsymbol{d}_t + \boldsymbol{b}^{(1)}) \tag{3.27}$$

ここで，$\boldsymbol{\sigma}$ はベクターシグモイド関数であり，$\boldsymbol{V}$，$\boldsymbol{W}^{(n)}$ と $\boldsymbol{c}$，$\boldsymbol{b}^{(n)}$ はニューラルネットワークの重みとバイアスのパラメータ，$\boldsymbol{h}^{(n)}(\boldsymbol{y})$ は n 番目の隠れ層の出力ベクトルである．なお，事前学習として各層の初期値を制約付きボルツマンマシン (Restricted Boltzmann Machine ; RBM) で学習した [22]．以上により，観測特徴量から静音環境下における音声特徴量に対する要素分布の事後確率 $p(k|\boldsymbol{y}_t)$ を推定することが可能となる．

評価段階では，ニューラルネットワークにより得られる事後確率 $p(k|\boldsymbol{y}_t)$ を重みとして，区分的線形変換によって静音環境下における音声特徴量 $\boldsymbol{x}_t$ を推定する (図 3.3)．

$$\hat{\boldsymbol{x}}_t = \sum_{k=1}^{K} p(k|\boldsymbol{y}_t)\boldsymbol{A}_k\boldsymbol{e}_t \tag{3.28}$$

なお，$\boldsymbol{e}_t$ は式 (3.15) にある通り，時刻 t を中心に数フレーム分の特徴量を連結したベクトルである．

3.4　実験

提案手法の有効性を Aurora-2 による連続数字読み上げ認識実験により評価した．音響モデルの学習と評価は complex backend と呼ばれる Aurora-2 の評価に標準的に用いられている設定を利用した [47]．データはいくつかの環境の雑音が重畳された雑音環境下における音声とそれに対応する静音環境下における音声が含まれている．学習データは 8,440 発話あり，本実験ではこれを全て音響モデルと雑音抑圧手法の学習に用いた．認識に用いる音響モデルは隠れマルコフモデル (Hidden Markov Model ; HMM) を利用した．各 HMM は

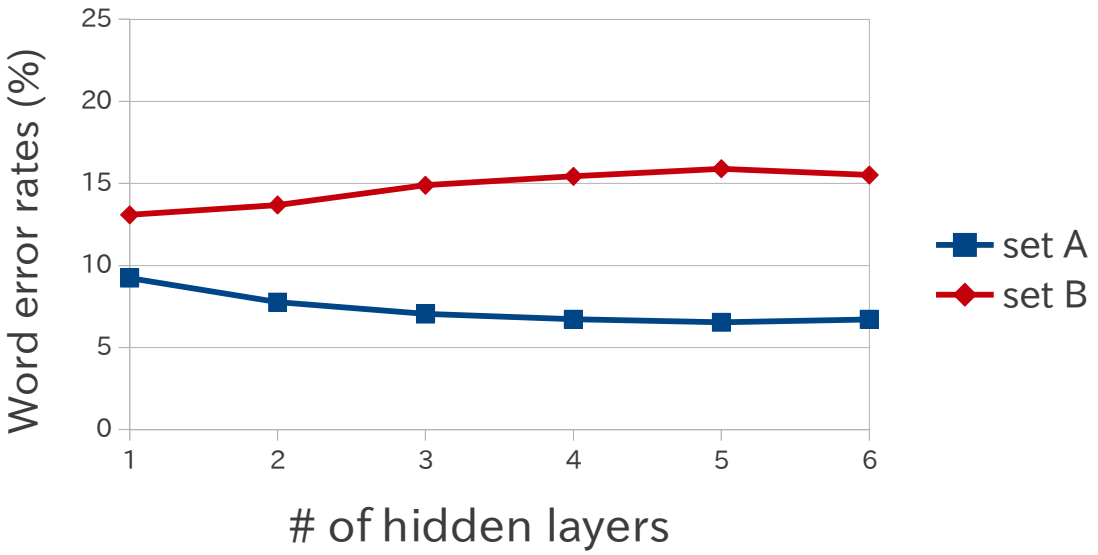


図 3.4: Aurora-2 データベースにおける，隠れ層の数を変化させた場合の提案手法の単語誤り率の変化

表 3.2: 提案手法と従来手法との単語誤り率 (%) の比較

	clean condition (WER. %)				multi condition (WER. %)			
	set A	set B	set C	Ave.	set A	set B	set C	Ave.
Baseline	48.93	55.80	39.23	47.98	10.57	**11.89**	14.33	12.27
SPLICE	14.89	19.31	21.59	18.60	9.20	14.50	15.22	12.97
REDIAL	16.70	20.59	21.14	19.48	8.98	13.26	**12.45**	11.56
DDAE	**6.39**	20.44	17.20	14.68	5.97	18.50	14.67	13.04
Proposed	7.04	**14.93**	**15.54**	**12.51**	**5.64**	15.20	13.29	**11.38**

単語モデルであり，単語は 0 から 9 の数字（0 は 2 通りの読み方がある）から成る．各単語は 18 状態の HMM であり各状態は 20 混合の GMM であり，無音区間は 4 状態 HMM で各状態が 36 混合の GMM で構成される．また，ネットワーク文法を用いて認識を行う．静音環境下における音声のみで学習したもの (clean condition) と雑音環境下における音声も含むデータで学習したもの (multi condition) の 2 通りで比較した．なお，雑音抑圧処理を行ったデータに対して認識を行う際は，multi condition のモデルを学習する際のデータにも同様の雑音抑圧処理を行っている．

評価データセットは 3 種類 (A,B,C) があり，セット A とセット B はそれぞれ 28,028 発話，セット C は 14,014 発話から構成されている．セット A は学習時と同じ雑音のみを含む既知雑音条件，セット B は学習時と異なる雑音を含む未知雑音条件，セット C は雑音は未知，既知共に含むが，伝達関数が異なる．特徴量として MFCC とパワー，その 1 次，2 次微分の計 39 次元を用いた．本実験では，異なる音声対雑音比（SNR）のデータを用いて雑

表 3.3: 音響モデルを静音環境下の音声特徴量で学習した場合の，既知雑音環境下における DDAE の雑音の種類，大きさ毎の単語誤り率 (%)

	closed-noise condition (Set A)				
	Subway	Babble	Car	Exhibition	Ave.
Clean	0.58	0.60	0.81	0.46	0.61
SNR 20	1.29	0.85	0.78	0.96	0.97
SNR 15	1.54	1.45	1.16	1.79	1.49
SNR 10	2.27	3.02	2.00	2.81	2.53
SNR 5	4.64	8.01	4.47	6.70	5.96
SNR 0	14.77	31.29	18.85	19.10	21.00
SNR -5	46.58	72.64	63.11	49.27	57.90
Average	4.90	8.92	5.45	6.27	6.39

表 3.4: 音響モデルを静音環境下の音声特徴量で学習した場合の，未知雑音環境下における DDAE の雑音の種類，大きさ毎の単語誤り率 (%)

	open-noise condition (Set B)				
	Restaurant	Street	Airport	Station	Ave.
Clean	0.58	0.60	0.81	0.46	0.61
SNR 20	1.11	2.21	1.82	1.14	1.57
SNR 15	2.21	6.32	3.91	3.46	3.98
SNR 10	6.05	18.02	10.11	8.67	10.71
SNR 5	18.33	40.39	25.95	24.38	27.26
SNR 0	51.30	69.35	58.75	55.29	58.67
SNR -5	97.91	90.93	96.42	86.52	92.95
Average	15.80	27.26	20.11	18.59	20.44

音抑圧を行うため，雑音の正規化に有効であると想定されるパワー項を用いている．また，ニューラルネットワークの学習は KALDI [48] を利用した．ニューラルネットワークを学習する際の誤差逆伝播法においては，学習データ 8,440 発話のうち 844 発話を開発セットとした．また，REDIAL と同様に線形変換の学習では正則化を導入した．

まず，ニューラルネットワークの層の数による単語誤り率の違いを図 3.4 に示す．入力に用いる特徴量 $\boldsymbol{d}_t$，$\boldsymbol{e}_t$ は当該フレームと，その前後 3 フレーム ($s = u = 3$) の計 7 フレームを用いた．あらかじめ予備実験により，ニューラルネットワークの各層のノード数は 1024 に設定し，静音環境下における音声特徴量空間をモデル化する際の GMM の混合数は $K = 1024$ に設定した．層の数を増やすと既知雑音条件では 5 層までは単語誤り率が単調に減少するが，未知雑音条件では DDAE と同様に効果が表れない．しかし，未知雑音条件と既知雑音条件で，層数 $N = 5$ の時に DDAE の場合は 14.05 ポイントの誤り率の差があったが，提案

表 3.5: 音響モデルを静音環境下の音声特徴量で学習した場合の，既知雑音環境下における提案手法の雑音の種類，大きさ毎の単語誤り率 (%)

	closed-noise condition (Set A)				
	Subway	Babble	Car	Exhibition	Ave.
Clean	0.49	0.57	0.60	0.65	0.58
SNR 20	1.01	0.91	0.66	0.86	0.86
SNR 15	1.72	1.21	1.22	1.57	1.43
SNR 10	2.36	2.60	2.12	2.81	2.47
SNR 5	5.25	7.62	6.80	7.41	6.77
SNR 0	16.40	31.59	26.04	20.77	23.70
SNR -5	49.86	72.13	69.31	54.46	61.44
Average	5.35	8.79	7.37	6.68	7.04

表 3.6: 音響モデルを静音環境下の音声特徴量で学習した場合の，未知雑音環境下における提案手法の雑音の種類，大きさ毎の単語誤り率 (%)

	open-noise condition (Set B)				
	Restaurant	Street	Airport	Station	Ave.
Clean	0.49	0.57	0.60	0.65	0.58
SNR 20	0.71	1.45	1.04	1.05	1.06
SNR 15	1.35	2.96	2.00	1.85	2.04
SNR 10	3.22	9.98	5.58	4.72	5.88
SNR 5	10.96	29.29	17.36	16.17	18.45
SNR 0	5.40	62.36	43.48	47.73	47.24
SNR -5	78.94	88.09	83.84	84.11	83.75
Average	10.33	21.21	13.89	14.30	14.93

手法では，8.91 ポイントまで減少することが確認できた．これは，ガウス分布の持つ雑音に対する汎化性能を効果的に利用できていると考えられる．

次に，SPLICE，REDIAL，DDAE を行った場合の単語誤り率 (word error rate; WER) の比較を表 3.2 に示す．提案手法のパラメータは図 4 と同様のものを用い，中間層の数は $N = 3$ とした．SPLICE の雑音環境下における音声特徴量をモデル化する際の GMM の混合数 I と，REDIAL の静音環境下と雑音環境下における音声特徴量をモデル化する際の GMM の混合数 K，K^* は予備実験により共に 1024 に設定した．また，REDIAL の LDA による次元圧縮後の特徴量 $\boldsymbol{v}_t$ の次元数は 64 としている．ニューラルネットワークのパラメータは，DDAE は中間層の数を $N = 5$ とした．

音響モデルを静音環境下における音声のみで学習した場合，提案手法が最も良い結果となった．特に，SPLICE と比較した場合，既知雑音条件では 53.72 %，未知雑音条件におい

ては 18.54 %の誤り削減率を得ることができている．また，音響モデルを雑音環境下における音声を含むデータで学習した場合[1]も，SPLICE と比較して未知雑音環境においては，4.82 %と誤り率が上昇してしまっているが，既知雑音条件では 38.70 %の誤り削減率を得た．なお，REDIAL が set B において高い性能を示している．これは，REDIAL が線形変換行列により次元圧縮を行った空間で対角共分散の GMM を再び構築するため，擬似的に全角共分散の GMM を構築することが可能であることが原因であると考えられる．

さらに，従来手法で最も単語誤り率が低かった DDAE と提案手法の既知雑音条件，未知雑音条件における単語誤り率を，雑音の種類と信号雑音比（SNR）別に表 3.3, 3.4 と表 3.5, 3.6 に示す．提案手法と DDAE のパラメータ設定は共に表 3.2 と同様のものを用い，音響モデルは静音環境下における音声のみで学習したものを用いた．既知雑音条件では，雑音が大きい場合，提案手法と比較した時 DDAE は 9.23 %誤り率が低い．しかし，未知雑音条件においては，雑音の大きな場合でも提案手法が有効であり，DDAE と比較して平均として 26.96 %の誤り削減率を得ることができた．これによって，静音環境下における音声特徴量空間をモデル化した際のガウス分布による汎化性能が効果的であることが確認できた．

3.5　まとめ

本章では，ニューラルネットワークによる，観測音声特徴量からそれに対応する静音環境下における音声特徴量の所属する要素分布の識別と，それにより得られる事後確率を重みとして用いた区分線形変換法を提案した．実験的にも未知雑音条件，既知雑音条件共に SPLICE と比較して雑音が既知の条件では 53.72 %単語誤り率を削減することができた．また，ニューラルネットワークを回帰モデルとして用いる DAE と比較した場合，ニューラルネットワークの持つ非線形性によって実現される複雑なモデル化と，静音環境下における音声特徴量のモデルとして用いたガウス分布の持つ汎化性能を組み合わせることで，既知雑音条件で低い誤り率を維持しつつ，雑音環境が未知な条件で 26.96 %の単語誤り率の削減が可能となった．

[1]Baseline が最も単語誤り率が低いが，これは Aurora-2 のデータセット特有の問題と考えられ，Droppo らの実験 [31] においても SPLICE により multiconditon における set B の性能が低下している．

第4章

話者コードによるパラメータ制御を用いたニューラルネット音響モデルの正規化学習

4.1　はじめに

前章では，特徴量ドメインにおいて非言語情報の一つである雑音の影響を取り除くため，ニューラルネットワークと GMM を組み合わせた特徴量強調手法を提案した．本章では，モデルドメインにおける非言語情報の制御に関する要素技術の一つである話者正規化学習をニューラルネットワークベースの音響モデルに対して適用する手法を提案する．

DNN を用いた音響モデルはその高い性能と引き換えに，各パラメータがどのような意味を持つのかが曖昧になった．GMM は生成モデルであり，各混合の分布のパラメータが明示的にではないにしろ，音素などの音響的事象と比較的似た意味を持っていたため，モデルパラメータの制御が容易であった．しかし，DNN は識別モデルであり，かつ非線形変換を多層に渡って積み上げるため，特徴量空間の分割が複雑でありパラメータの制御が困難である．そのため，GMM ベースの手法が利用できないケースが多い．その一つが音声に内在する非言語情報である話者の違いに起因する影響を低減する話者適応技術と，それに伴う話者正規化学習である．

DNN 音響モデルに対する話者適応技術の一つとして，Xue らの提案した話者コードを用いた話者適応がある [14]．これは，あらかじめ，通常の話者非依存の DNN を学習し，各層のバイアス項の制御を，サブネットワークを接続して話者依存の話者コードにより行なう．なお，認識時は，入力話者に対応する話者コードが不明なため，話者依存のパラメータを固定し，誤差逆伝搬法により話者コードを推定することで話者適応を実現する．しかし，この手法は話者非依存の DNN と話者依存のパラメータを別々に学習しており，これにより明確に話者依存／非依存の情報を分離できているとは言えず，話者適応に適した話者正規化が実現できているわけではない．

一方，DNN 音響モデルに対する話者依存／非依存のパラメータの同時推定法として，落合らの話者正規化学習法がある [18]．これは，層ごとに話者依存／非依存を決め，話者依存層を学習データの話者毎に切り替えて学習する．この枠組みを用いて，話者依存／非依存のパラメータの同時推定を行うことでそれぞれの情報を分離することが可能となる．こちらも，認識時は入力話者に対応する話者依存層のパラメータを誤差逆伝搬法により推定し話者適応を行う．しかし，この正規化法は誤差逆伝搬法の際に話者依存層のパラメータを切り替える必要があり，学習コストの削減のためミニバッチ学習を GPU 上で行うことの多い DNN の学習において，このコスト増大は致命的である．

本提案手法では，話者を表現する特徴量と音素状態識別を行う DNN の同時推定を行い，かつ，効率的な学習の実現を目的として，話者コードベースの話者正規化学習を提案する．

話者コードは少数のパラメータにより各層のバイアスを制御することが可能である．DNN が多層の構造により段階的な特徴量変換を行っていると考えると，話者コードにより異なる特徴量ドメインにおいて話者依存の成分を分離することが期待できる．また，話者コードはその性質上，学習時は入力特徴量と同様の扱いが可能となる．これにより，各層のパ

ラメータを学習中に切り替える必要がないため，誤差逆伝搬法中に特別な操作が必要無い．そのため，学習の際のコスト増加を抑えることが可能となる．

4.2　関連技術

4.2.1　話者コードを用いたモデル適応

Xue らは話者コード (speaker code) を用いた DNN の適応手法を提案している [14]．概要を図 4.1 に示す．この手法では，まずあらかじめ通常の DNN の学習を行い，話者非依存のモデルを構築する．その後，学習された話者非依存の DNN に対して，各層と話者コードと呼ばれる話者の特徴を表現するサブネットワークを連結する．これにより，層 l の出力 $\boldsymbol{O}^{(l)}$ は，$\boldsymbol{S}_c$ を話者 c の話者コードベクトルとすると，

$$\boldsymbol{O}^{(l)} = \boldsymbol{\sigma}(\boldsymbol{W}^{(l)}\boldsymbol{O}^{(l-1)} + \boldsymbol{b}^{(l)} + \boldsymbol{B}^{(l)}\boldsymbol{S}_c) \tag{4.1}$$

として，計算する．なお，ここで，$\boldsymbol{W}^{(l)}$ は l と $l-1$ 層間の線形変換パラメータであり，$\boldsymbol{B}^{(l)}$ は l 層への話者適応の重みパラメータである．また，$\boldsymbol{\sigma}$ はベクターシグモイド関数である．

サブネットワークの学習は，予め学習した DNN のパラメータ $\boldsymbol{W}^{(l)}$ を固定した誤差逆伝搬法によって行うが，話者コード $\boldsymbol{S}_c$ は話者毎に切り替えて学習を行う．これにより，サブネットワークのパラメータを通して各層のバイアスパラメータを話者コードにより制御するネットワークモデルを構築することが可能となる．なお，話者コードの誤差逆伝搬法はクロスエントロピー最小化などの目的関数を E とした場合，

$$\frac{\partial E}{\partial \boldsymbol{S}_{c,k}} = \frac{1}{L}\sum_{l}\sum_{j}\frac{\partial E}{\partial \boldsymbol{O}_j^{(l)}}(1-\boldsymbol{O}_j^{(l)})\boldsymbol{O}_j^{(l)}\boldsymbol{B}_{kj}^{(l)} \tag{4.2}$$

として計算することができる．これにより，話者依存の情報が話者コードに集約され，話者非依存の変換が線形変換行列 $\boldsymbol{B}^{(l)}$ に集約される．

モデル適応の際は，対応する話者の話者コードを入力することで効率的にモデルパラメータを制御することができる．なお，入力話者は未知話者であるため，教師あり適応を想定した場合，少量の適応データを用いて誤差逆伝搬法により入力話者の話者コードを推定する．

このモデルは，ベースの話者非依存 DNN のモデルパラメータ $\boldsymbol{W}, \boldsymbol{b}$，話者適応サブネットのモデルパラメータ $\boldsymbol{B}$，そして話者コード $\boldsymbol{S}_c$ の 3 種類のパラメータを学習する必要がある．しかし，話者コード $\boldsymbol{S}^{(c)}$ と話者適応サブネットのモデルパラメータ $\boldsymbol{B}^{(l)}$ は同時に学習するが，話者非依存の DNN のモデルパラメータはバイアスを除いて固定するため，各層間の線形変換行列の学習時に話者依存の情報の分離が不十分となる恐れがある．そのため，効率的に話者依存の情報を話者適応サブネットの学習に利用できているとは言えない．

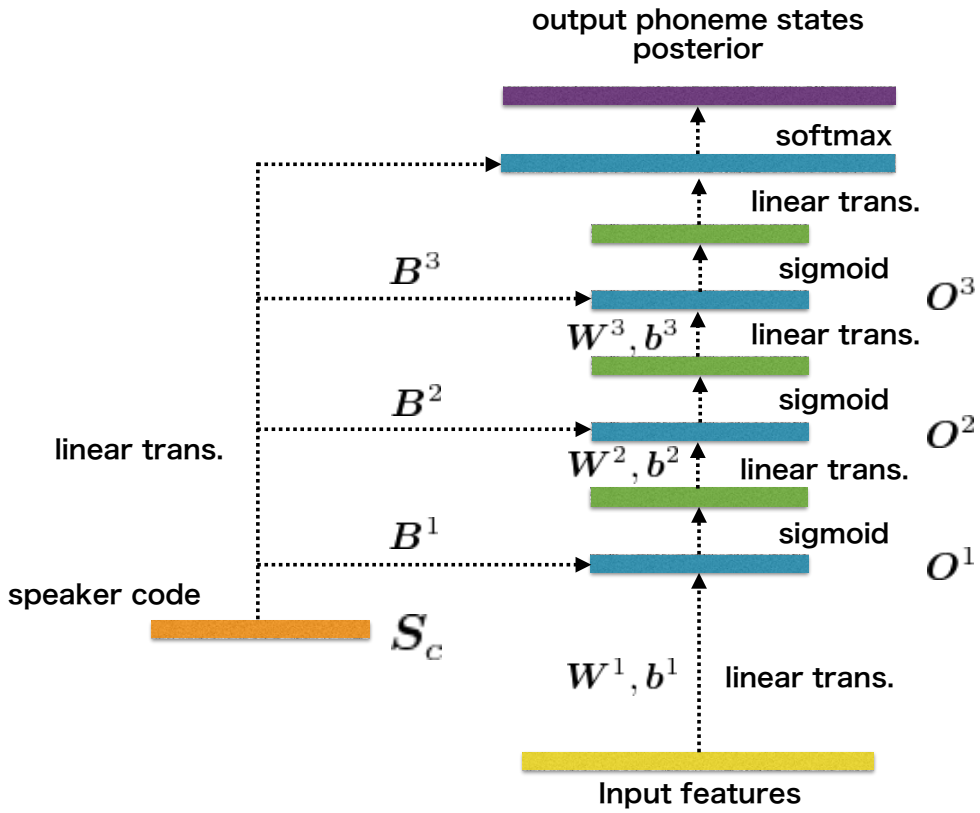


図 4.1: 話者コードを用いたニューラルネットワークの直接適応

4.2.2　話者依存層の切り替えによる話者正規化学習

落合らは，話者依存層を切り替えることによる話者正規化学習を提案している．図 4.2 にモデルの概要を示す．多層ニューラルネットにおいて話者依存／非依存層を設定し，学習時は話者毎に話者依存層のパラメータを切り替える．この枠組みを用いて誤差逆伝搬法により学習することで，話者依存／非依存のパラメータを同時に学習することが可能となる．認識時は，話者非依存層のパラメータを固定し，適応データを用いて誤差逆伝搬法により話者依存層のパラメータを再推定する．これにより話者正規化学習されたニューラルネットを用いた話者適応を実現することができる．なお，この際の話者依存層パラメータの初期値は，話者非依存層のパラメータを固定して全ての学習データで再学習を行ったものを用いている．

しかし，誤差逆伝搬法において話者依存層のパラメータを切り替えることは，非常にコストが大きい．DNN は学習コストの削減のため，学習時にミニバッチを用いて学習することが多い．そのため，各話者のデータ毎にミニバッチを構築することで，話者依存層の切り替えコストを削減することが可能となるが，学習データの偏りの問題が発生してしまう．また，話者依存層を多層に渡って設定した場合，過学習が発生してしまう恐れもある．

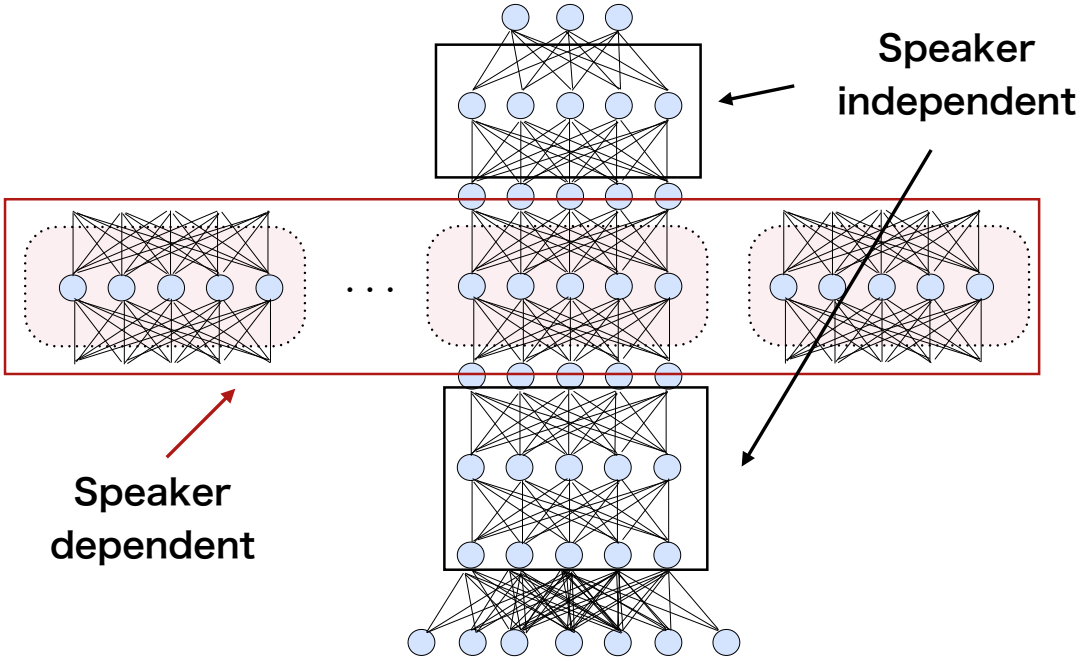


図 4.2: 層の切り替えを用いた話者正規化学習

4.3　話者コードを用いた話者正規化学習

4.3.1　話者依存/非依存パラメータの同時推定

話者コードを用いたモデル適応はバイアス成分に限定して話者の変動を表現することで，各層毎に話者適応を行うことができる．また，各層ごとに扱うパラメータ数も少ないため，多層に渡ってモデル適応が可能となる．しかし話者コードを用いた適応は，ベースとなるニューラルネットを予め学習するため，サブネットの学習時に話者に依存する情報が，効果的に伝搬しない．これは，あらかじめ学習する話者非依存のネットワーク自体がある程度の話者正規化能力を持つためである．

そこで，本提案手法は話者コード型の適応手法をベースとし，話者非依存のネットワークと話者依存ネットワークの同時推定学習を行う．これにより，話者情報を制御する機能をサブネットワークに集約することができ，より効果的な話者正規化学習を実現することが可能となる．

また，話者毎に特定層のパラメータを全て切り替えるタイプの正規化学習は，誤差逆伝搬法を行う際に必要となる，データの偏りを防ぐための入力セグメントのランダマイズが困難であるという問題があった．ランダムに並び替えた各セグメント毎にネットワーク構

造を変更しなければならず，これは計算コストが非常に高い．また，高速化のためミニバッチに同一話者のデータを集約する等の手段がとられるが，逐次更新を行う場合学習データの偏りは防ぐことができない．これは話者非依存部分のサブネットの学習にとって望ましいとは言えない．

それに対して，本提案手法は話者コードの枠組みにより正規化学習を行うことで，モデルパラメータの切り替えが不要となる．これは，話者コードを切り替えることがバイアスパラメータの切り替えに相当するためであり，ミニバッチ中でも各データに対応する話者インデクスさえ用意すれば，データのランダマイズは可能であり，通常の誤差逆伝搬法と同様の枠組みを利用できる．

中間層が 3 層の場合のネットワークの構造を図 4.3 に示す．学習データ中に存在する話者数が N 人の場合，各ビンを各話者に対応させた 1-of-N ベクトル $\boldsymbol{v} = [v_1, \ldots, v_n]^\top$ を考えると

$$\boldsymbol{S}_c \;=\; \boldsymbol{\sigma}(\boldsymbol{D}\boldsymbol{v}) \;\begin{cases} v_n = 1 \ (n = c) \\ v_n = 0 \ (n \neq c) \end{cases} \tag{4.3}$$

とする線形結合層 $\boldsymbol{D}$ の sigmoid 出力を話者コードとして再定義する．以下 $\boldsymbol{D}$ を辞書行列と呼ぶ．sigmoid 出力とすることで，誤差逆伝搬法の際の話者コードが閉区間 $[0, 1]$ の値のみを取るという制約を設けることに相当する．話者コードが全ての層との連結を持つことを考慮すると，それぞれの層が持つ意味が異なるため，スケーリングの効果は各層との連結行列である $\boldsymbol{B}^{(l)}$ に集約することを目的としている．

そして，各中間層の出力 $\boldsymbol{O}_{1,2,\ldots,l}$ を

$$\boldsymbol{O}^{(l)} = \sigma(\boldsymbol{W}^{(l)}\boldsymbol{O}^{(l-1)} + \boldsymbol{b}^{(l)} + \boldsymbol{B}^{(l)}\boldsymbol{S}_c) \tag{4.4}$$

とする．この枠組みを用いて，全てのパラメータを同時に学習することで，バイアス成分をグローバルな成分と話者に依存する成分に分離してモデル化することが可能となる．

4.3.2　ネットワーク構造

学習時は，入力特徴に加え各話者を 1-of-N 表現で表したベクトルを入力とし，誤差逆伝搬法により学習する．これは，学習話者毎に話者コードを切り替えて学習していることに相当する．ただし，辞書行列によりバイアスを制御するため，先行研究と異なり，明示的にモデルパラメータや話者コードを入れ替える必要がない．そのため，このモデル構造を通常の BP と同様のアルゴリズムで学習することにより，各セグメント単位でモデルパラメータの切り替えに相当する効果が得られるため，学習データの偏りを回避することが可能となる．

認識時は，認識する対象話者の話者コードを観測することができない．そこで，グロー

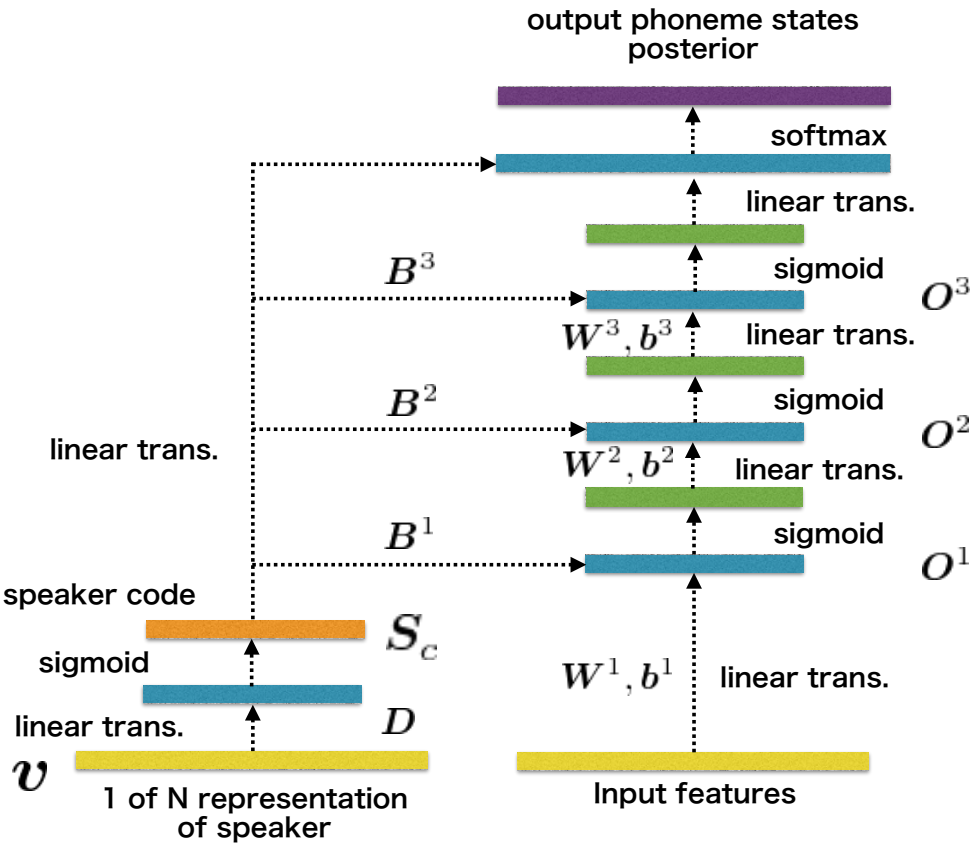


図 4.3: 制約付き話者コードを用いたニューラルネットワークの直接適応

バルな話者コードを擬似的に与えて認識に用いる．これは，

$$\boldsymbol{O}^{(l)} = \boldsymbol{\sigma}(\boldsymbol{W}^{(l)}\boldsymbol{O}^{(l-1)} + \boldsymbol{b}^{(l)} + \boldsymbol{B}^{(l)}\boldsymbol{S}_{global}) \tag{4.5}$$

として各隠れ層の出力を計算すれば良いことに相当する．なお，グローバルな話者コードの推定はニューラルネットのパラメータを固定したまま誤差逆伝搬法 により推定を行う．これは，図 4.4 のように話者コードをダミーのスカラーからの線形変換の出力を sigmoid 関数により $[0, 1]$ に正規化したモデルとして考えた場合，線形変換パラメータを推定することに相当する．そのため，誤差逆伝搬法の枠組みで学習することが可能である．なお，話者コードを与えた場合，当然ながら

$$\hat{\boldsymbol{b}}^{(l)} = \boldsymbol{b}^{(l)} + \boldsymbol{B}^{(l)}\boldsymbol{S}_{global} \tag{4.6}$$

$$\boldsymbol{O}^{(l)} = \boldsymbol{\sigma}(\boldsymbol{W}^{(l)}\boldsymbol{O}^{(l-1)} + \hat{\boldsymbol{b}}^{(l)}) \tag{4.7}$$

としてバイアス項を纏めることにより，通常のニューラルネットの形に変形することができる．したがって，認識時は既存の枠組みを流用することが可能となる．

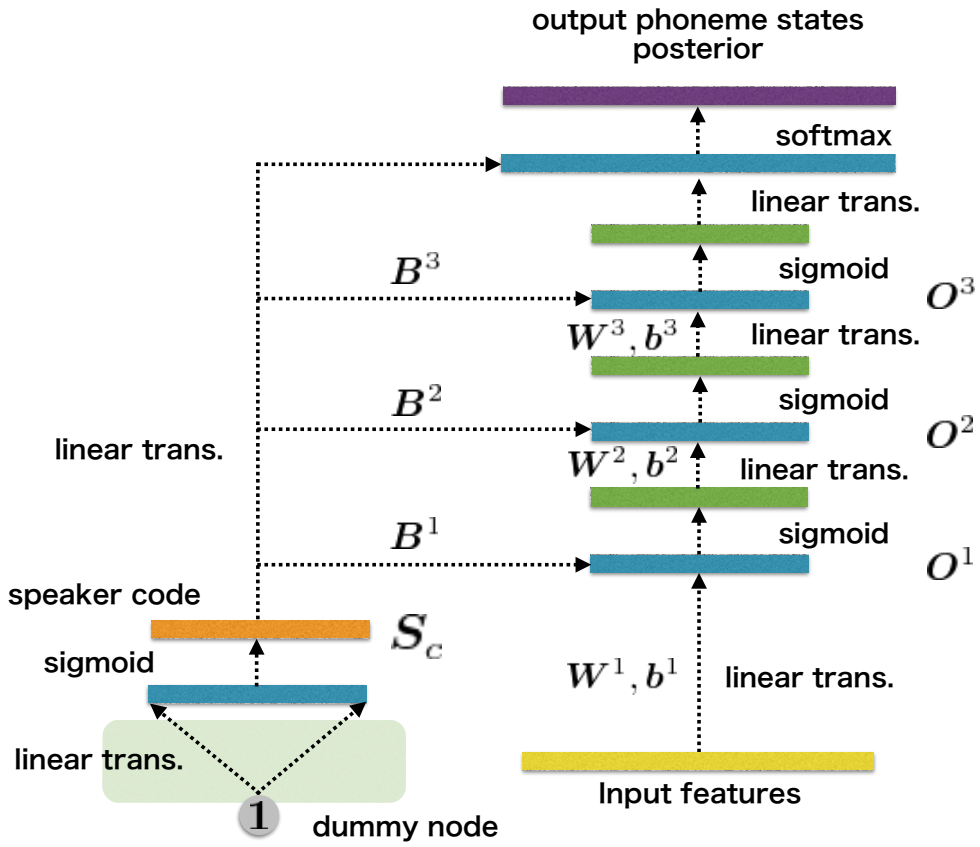


図 4.4: ダミーノードを用いた話者コードの推定

4.3.3　話者適応

認識時に用いるグローバルな話者コードは実際の話者の話者コードとは異なるため，ミスマッチが生じてしまう．そこで，図 4.4 のように新しく入力された話者に対する話者コードを話者コードに入力することによりモデルの適応を実現する．しかし，学習時と異なり，未知話者に対する話者コードは観測できないため，少量の教師あり適応データを用い，話者コードをモデルパラメータを固定した誤差逆伝搬法により推定する．この際，話者コードの初期値には正規化学習を行った際に推定したグローバルな話者コードを用いる．

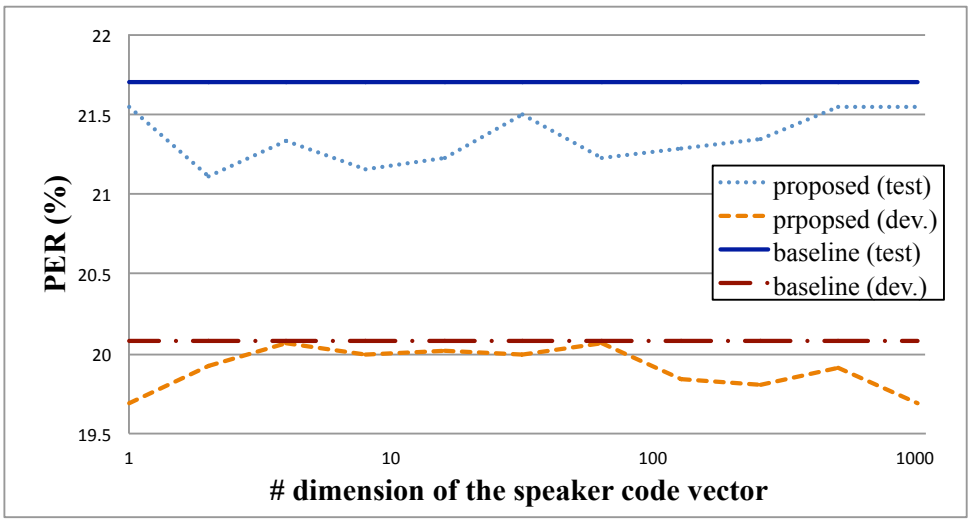


図 4.5: 制約付き話者コードを用いた話者正規化学習を行った DNN 音響モデルの TIMIT データベースにおける音素認識誤り率

4.4　実験

4.4.1　実験条件

提案手法により学習を行った DNN 音響モデルの評価を TIMIT データベース [49] を用いた連続音素認識実験により行った．音響モデルの学習には各話者 8 発話，全 462 話者の計 3696 発話を用い，評価セットは TIMIT のコアテストセットである 24 話者計 192 発話を用いた．また，デコード時の音響モデルの重みの調整として 50 話者 400 発話の開発セットを用いる．なお，学習セットと評価セット，開発セットで話者の重複はない．

DNN モデルの学習に用いる音素状態ラベルのアライメントには fMLLR を行ったトライフォンモデルを用い，音素状態は計 1951 状態である．DNN は隠れ層 6 層，各層 2048 ノード，活性化関数には sigmoid 関数を用いた．入力特徴量として MFCC に対して fMLLR を行ったものを用いている．なお，DNN の学習の際は開発セットと学習データで話者の重複が必要なため，学習データの内，フレーム単位で 5%を DNN 学習用の開発セットとして用いた．事前学習にはオートエンコーダーを用いており，また，dropout は行っていない．さらに，学習の際のハイパーパラメータのチューニングに起因する影響を軽減するため，ラーニングレート，エポック数等の設定は通常の DNN の学習の際の結果に対して最良値を採用した．このため，多少ではあるが，提案手法において不利な条件となっている．

表 4.1: TIMIT データベースにおける提案手法を用いて学習したモデルに対する適応性能（音素認識誤り率）

	PER (%)
baseline	21.50
+ BP adaptation	21.42
+ SC adaptation	21.21
SC-based normalized DNN	21.11
+ adaptation	**20.98**

4.4.2　話者正規化 DNN の性能評価

提案手法により話者正規化学習を行った DNN の性能を比較する．これは，提案手法においてグローバルな話者コード $\boldsymbol{S}_{global}$ を入力した場合に相当する．話者コードの次元数と音素認識誤り率の対応をプロットした実験結果を図 4.5 に示す．なお，baseline は，通常の DNN 音響モデルである．開発セットと評価セット共に特に話者コードの次元数が小さい場合，通常の DNN 音響モデルと比較して提案手法の方が良い結果が得られている．これにより，提案手法においては，話者に由来する各層のバイアスの変動が話者コード側のネットワークによって適切に吸収されて学習されていると考えられる．なお，話者コードの次元数が小さい場合に良い性能が得られていることにより，話者コードがボトルネック層として機能していることも期待されるが，学習データ話者の直感的なクラスタリングが実現できている様子はなかった．

4.4.3　話者正規化 DNN を用いた話者適応性能の評価

本提案手法により話者正規化を行った DNN をベースとした話者適応の結果を表 4.1 に示す．これは，提案手法において適応データにより推定した入力話者の話者コード $\boldsymbol{S}_c$ を入力した場合に相当する．適応データには各話者 2 発話を用いた．なお，この 2 発話は TIMIT 中の “sa” ラベルが振ってあるものであり，各話者について共通の文章となっている．適応の際のラーニングレート，バッチサイズ，エポック数等のハイパーパラメータ群は，各手法において最良値を採用した．baseline は通常の DNN での認識結果であり，BP adaptation は適応データを用いて誤差逆伝搬法により追加学習を行ったものである．なお，適応時は開発セットの利用が不可能であるため，開発セットを用いずにデコード時の重みを決定した．そのため，図 4.5 とベースラインの値が異なる．これは，適応時は開発セットの利用が不可能であるためである．また，SC adaptation は先行研究である Xue らの手法により適応を行ったものである．なお，先行研究，提案手法共に話者コードの次元数は 2 とした．

提案手法である SC-based normalized DNN は話者コードにグローバルな値を入れた場合でも既に他手法の適応後と同等の誤り率を得ることが出来ている．さらに，話者適応を行

表 4.2: 提案手法の TIMIT データベースにおける音素認識誤り率による他手法との比較

	PER (%)
monophone HMM	34.30
triphone HMM	30.42
triphone HMM (SAT)	25.47
Subsupeace GMM	23.06
Basic DNN/HMM	21.50
Proposed	21.11
Proposed + Subspace GMM	**20.64**

うことで，僅かではあるが誤り率を削減することが可能となる．なお，全体的にモデル適応の効果が低いが，これは入力特徴量に対して fMLLR を行っているため，あらかじめ話者によるばらつきが低減されているためと考えられる．

4.4.4　他手法との比較

提案手法により正規化学習を行った音響モデルと様々な音響モデルとの性能比較を表 4.2 に示す．monophone HMM と triphone HMM では MFCC を，それ以外では MFCC+fMLLR を特徴量として用いている．また Proposed + Subspace GMM は提案手法と subspace GMM を system combination したものであり，combination の際の重みは開発セットを用いて決定した．subspace GMM と提案手法の system combination により 20.64 %の音素認識誤り率を得ることができた．

4.5　まとめ

本章では，話者コードを用いた DNN 音響モデルの話者正規化学習を提案した．提案法はサブネットワークを接続し，1-of-N 表現の話者コードを入力としてネットワークの話者性を制御する．この構造を用いて話者依存／非依存のパラメータを同時に学習することで，話者依存／非依存の情報を分離することができ，正規化学習が実現が可能となる．また，話者コードをベースとすることにより，誤差逆伝搬法による学習を行う際に，明示的にモデルパラメータを切り替える必要がないため，学習コストの増加を抑えることができる．

第5章

識別的アプローチによる分布間距離計算とその利用

5.1　はじめに

前章では，音響モデルの話者・環境適応の性能向上を目的とした，話者コードを用いた正規化学習法を提案した．しかし，モデル適応の際に話者・環境コードを生成的に推定するとはいえ，ニューラルネットワークのパラメータの持つ厳密な意味づけは曖昧なままである．生成モデルとニューラルネットワークの効果的な融合を目指す上では，モデルの性質に着眼して各々の利点を組み合わせるだけではなく，従来の音声学的な理論的背景を考慮したアプローチが重要になると考えられる．そこで，本章では話者不変な特徴量である音声の構造的表象に着目する．

音声の構造的表象は分布間距離を利用した生成モデルに基づく特徴量である [19]．話者の違い，つまり声道のフィルタに起因する違いはケプストラム空間上でのアフィン変換によって表される [35]．特徴量の分布をガウス分布として仮定した場合の分布間距離はアフィン変換不変であるため，アフィン変換に対して普遍な分布間距離を用いた特徴量である構造的表象は話者の違いに対して普遍な特徴量となる．その結果，例えば音素の特徴量分布間の距離を想定した場合，アクセントや方言・言語の違いなどの情報を持つと考えられる．

分布間距離は，Kullback-Leibler 距離 (KL 距離) などが代表的であるが，情報量とも呼ばれパターン認識において重要な位置を占める．例えば，ガウス混合モデル (Gaussian Mixture Model ; GMM) に代表される隠れ変数を持つモデルの学習手法の一つである EM アルゴリズムは KL 距離を最小化する基準でパラメータのアップデートを行う．近年，音声認識の音響モデルにおいても主流となっているニューラルネットワークにおいても，KL 距離の変化を抑える基準を入れたモデル適応手法が効果的であるとの報告がなされている [13]．これは，適応の初期値として用いる複数話者のデータで学習された話者非依存のモデルと，適応後の話者依存のモデルとの KL 距離が大きく変わらないという制約により過学習を抑制していると考えられる．

しかし，そもそも観測特徴量の真の分布はガウス分布よりもさらに複雑な形状をしていると考えられる．そのため，単純なガウス分布ではなく，複雑な形状を仮定することで分布間距離の推定精度が向上することが考えられる．例えば，分布形状を GMM で仮定し，モデル適応を利用して分布間距離を推定する手法も提案されている [50]．しかし，変換前と後で対応するインデクスの分布形状はあくまで正規分布を仮定するため，変換によって分布形状が変化する場合妥当な近似とは言えない．

一方，先に述べた通り音声認識において近年，DNN に代表される識別モデルが，従来の GMM をベースとした生成モデルに対して高い精度を示すことが示されている [8, 22]．これは，DNN に代表される識別モデルは GMM などの生成モデルと比較して特徴量の分布に対する仮定が弱いため，柔軟なモデル化が可能であることが理由の一つと考えられる．

この識別モデルを利用した分布間距離の推定法として，Heigold らの対数線形モデルを用いた生成モデルを介する手法がある [51]．これは，対数線形モデルの性質を利用し，識

別モデルのパラメータからガウス分布のパラメータを推定することで，分布間距離を計算する．しかし，対数線形モデルのパラメータからでは，ガウス分布の平均と分散パラメータが一意に決定することができないため，分散パラメータを各ラベルに対して共有する等の仮定を置く必要がある．これを用いて DNN の最終層として用いられるソフトマックス関数から分布間距離を計算することで効率的に DNN を学習する手法も Li らにより提案されている [52]．だが，識別モデルの分布を明示的に仮定する必要のない性質を利用して分布間距離を推定する場合，生成モデルのパラメータを経由せずに直接推定することが理想的である．

そこで，本研究では，モデル適応に用いる分布間距離をニューラルネットワークにより計算する手法を提案する．これにより分布間距離の推定精度が向上するだけでなく，音響モデル適応に用いる際に問題となる計算コストの増加を抑えることが可能となる．

提案法により計算された分布間距離の有効性を特徴量ドメインと音響モデルドメインにおいて評価する．特徴量ドメインでは言語識別タスクにおける入力特徴量として利用を行い評価する．なお，言語識別システムを利用する理由は，音声認識システムと異なり現在の言語識別システムは識別器自体に時系列モデリングを行わないため特徴量自体の評価が明確となるためである．音響モデルドメインでは音響モデルの話者適応における制約として導入することで評価を行う．

5.2　関連研究

本節では，まず分布間距離を利用した非言語情報の違いに頑健な特徴量表現である音声の構造的表象について説明する．その後，GMM ベースの分布間距離推定法である Chang らの手法と，識別モデルを用いた分布間距離の推定法である Heigold らの手法を用いて DNN から分布間距離を推定する Li らの手法について紹介する．

5.2.1　音声の構造的表象

音声の構造的表象は，各音響イベント間の分布間距離を要素としてもつ特徴量表現である．音響イベントを N 個の音素状態とすると，音声の構造的表象は $N \times (N-1)/2$ の次元数を持つ [19]．音声の構造的表象を構成する分布間距離としては f-divergence の一種であるバタチャリヤ距離を用いることが一般的である．各音響イベントの特徴量分布をガウス

分布と仮定した場合のバタチャリヤ距離は,

$$BD(q_i, q_j) = -\ln \int \sqrt{p(\boldsymbol{x}|y=q_i)p(\boldsymbol{x}|y=q_j)}d\boldsymbol{x} \tag{5.1}$$

$$= \frac{1}{8}(\boldsymbol{\mu}^{(q_i)} - \boldsymbol{\mu}^{(q_j)})^\top \Sigma^{-1}(\boldsymbol{\mu}^{(q_i)} - \boldsymbol{\mu}^{(q_j)}) \tag{5.2}$$

$$+\frac{1}{2}\ln\left(\frac{\det\Sigma}{\sqrt{\det\Sigma^{(q_i)}\ \det\Sigma^{(q_j)}}}\right)$$

$$\Sigma = \frac{\Sigma^{(q_i)} + \Sigma^{(q_j)}}{2} \tag{5.3}$$

となる．ここで，$\boldsymbol{\mu}^{()}$，$\Sigma^{()}$ は各音響イベントのガウス分布の平均と分散である．

ケプストラム空間において，話者の性別や年齢の違いはアフィン変換によってよく表現できることが知られている [35]．バタチャリヤ距離はアフィン変換に対して不変であるため，各音響イベントをガウス分布として仮定した音声の構造的表象は，話者の性別や年齢の違いに対して頑健な特徴量であると言える [36].

5.2.2　特徴量分布に GMM を仮定した分布間距離推定

Chang らは，ガウス分布ではなく，GMM を仮定した場合の分布間距離推定法を提案している [50]．バタチャリヤ距離は次式によって定義される．

$$BD(a,b) = -\ln \int \sqrt{P(\boldsymbol{x}|y=a)P(\boldsymbol{x}|y=b)}d\boldsymbol{x} \tag{5.4}$$

ここで，a, b は音響イベントのラベルである．これは，音素や，音素状態ラベル等の特徴量との対応が取れるものであれば用いることができる．各ラベルに対応する特徴量の分布を正規分布 $P(\boldsymbol{x}|y) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_y, \Sigma_y)$ と仮定した場合，バタチャリヤ距離は

$$BD(a,b) = \frac{1}{8}(\boldsymbol{\mu}^{(a)} - \boldsymbol{\mu}^{(b)})^\top \Sigma^{-1}(\boldsymbol{\mu}^{(a)} - \boldsymbol{\mu}^{(b)})$$

$$+\frac{1}{2}\ln\left(\frac{\det\Sigma}{\sqrt{\det\Sigma^{(a)}\det\Sigma^{(b)}}}\right) \tag{5.5}$$

$$\Sigma = \frac{\Sigma^{(a)} + \Sigma^{(b)}}{2} \tag{5.6}$$

として正規分布のパラメータから推定することができる．従来はこれを用いることで観測特徴量により正規分布のパラメータを推定することを経由し，バタチャリヤ距離を推定することが一般的であった.

Li らは特徴量分布を GMM に置き換えた場合,

$$
\begin{aligned}
BD(a,b) &= \sum_{j}^{J} BD(a_j, b_j) && (5.7)\\
&= -\sum_{j}^{J} \ln \int \sqrt{w_j^{(a)} P(\boldsymbol{x}|y=a)} \sqrt{w_j^{(b)} P(\boldsymbol{x}|y=b)} d\boldsymbol{x} && (5.8)\\
&= \frac{1}{8} \sum_{j}^{J} (\boldsymbol{\mu}_j^a - \boldsymbol{\mu}_j^b)^\top \Sigma^{-1} (\boldsymbol{\mu}_j^{(a)} - \boldsymbol{\mu}_j^{(b)}) \\
&\quad + \frac{1}{2} \sum_{j}^{J} \ln \left(\frac{\det \Sigma}{\sqrt{\det \Sigma_j^{(a)} \det \Sigma_j^{(b)}}} \right) && (5.9)\\
&\quad - \frac{1}{2} \sum_{j}^{J} \ln (w_j^{(a)} w_j^{(b)}) \\
\Sigma &= \frac{\Sigma_j^{(a)} + \Sigma_j^{(b)}}{2} && (5.10)
\end{aligned}
$$

として計算している．各GMMはあらかじめ学習されたUniversal-Background Model (UBM)を適応したモデルを初期値としてGMM を学習している．そのため，GMM 同士のインデクスの対応が取れており，インデクスの対応する正規分布間のバタチャリヤ距離が重要であるという仮定に基づく．

5.2.3　出力分布基準に基づく分布間距離推定

識別モデルを用いた分布間距離の計算として，Heigold らの手法がある．これは，対数線形モデルの性質を利用して正規分布のパラメータを推定することで，分布間距離を推定することができる．対数線形モデルは，入力を $\boldsymbol{x}$ とすると

$$
P(y|\boldsymbol{x}) = \frac{1}{Z(\boldsymbol{x})} \exp\left(\sum_i \lambda_{yi} f_i(\boldsymbol{x})\right) \tag{5.11}
$$

として出力 y に対する事後確率を推定することができる．ここで，i は次数である．ニューラルネットワークなどで用いられる softmax 関数は,

$$
\begin{aligned}
f_1(\boldsymbol{x}) &= [\boldsymbol{x}^\top, 1]^\top && (5.12)\\
\lambda_{y1} &= [\boldsymbol{W}_y, b_y] && (5.13)\\
\lambda_{yi} &= 0 \quad (i \geq 2) && (5.14)
\end{aligned}
$$

とした場合と等価となる．各ラベルに対応する分布が正規分布であると仮定した場合，この対数線形モデルのパラメータから正規分布のパラメータへ対応づけを行うことができる．クラスに依存しない半正定値行列 Σ を与えた場合，各正規分布のパラメータは

$$\Sigma_y = \Sigma \tag{5.15}$$

$$\boldsymbol{\mu} = \Sigma^{-1}[\boldsymbol{W}_y, b_y] \tag{5.16}$$

となる．

これを用いて Li らは，DNN を学習する際に，大きなネットワークにより推定される正規分布と学習する DNN との KL 距離が，変化しない基準を導入して学習を行った．これによって認識性能の向上が実現されている．しかし，このアプローチは，分散パラメータの共有化は音声認識において効果的な仮定ではあるものの，分散と平均との間に相互依存が存在する．そのため，パラメータを一意に決定することができない．また，正規分布を仮定するため，実際の分布の形が正規分布と大きく異なっている場合，分布間距離の推定精度が大きく低下すると考えられる．

5.3　識別的アプローチによる分布間距離の推定と音声の構造的表象の構築

5.3.1　識別的アプローチによる分布間距離の推定

分布間距離の計算は，分布の形を仮定する必要があるが，近年の DNN に代表される識別モデルの台頭を鑑みた場合，分布の形状は複雑なものであると想定される．そのため，分布の形状を明示的に仮定せずに分布間距離を推定することが望まれる．本節では，構造的表象などに用いられる BD 距離を従来の生成モデルをベースとしたものではなく，識別モデルを用いて推定する手法を提案する．

さて，バタチャリヤ距離を明示的に分布の形状を仮定せず，識別モデルを用いて推定したい．ここで，我々は識別モデルにより事後確率 $P(y=a|\boldsymbol{x}), P(y=b|\boldsymbol{x})$ を直接計算することが可能である．そこで，式 (5.4) をベイズ則を用いることにより，

$$\begin{aligned} BD(a,b) = & -\ln \int P(\boldsymbol{x})\sqrt{P(y=a|\boldsymbol{x})P(y=b|\boldsymbol{x})}d\boldsymbol{x} \\ & +\frac{1}{2}\ln P(y=a) \\ & +\frac{1}{2}\ln P(y=b) \end{aligned} \tag{5.17}$$

とすることで，事後確率を用いて計算することが可能となる．さらに，事後確率を DNN な

どの分布の形状を明示的に仮定しない識別モデルにより計算する．これにより陽にそれぞれ音響イベントにおける特徴量の分布の形を決定することなく，バタチャリヤ距離の計算が可能となる．

データ集合を $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_L\}, \boldsymbol{Y} = \{y_1, y_2, \ldots, y_L\}$ とすると，

$$\begin{aligned} BD(a,b) &= -\ln \frac{1}{L} \sum_l \sqrt{P(y_l = a|\boldsymbol{x}_l, \theta) P(y_l = b|\boldsymbol{x}_l, \theta)} \\ &\quad + \frac{1}{2} \ln \frac{1}{L} \sum_l P(y_l = a) \\ &\quad + \frac{1}{2} \ln \frac{1}{L} \sum_l P(y_l = b) \end{aligned} \tag{5.18}$$

として計算することができる．なお，ここで θ は識別モデルのパラメータであり，例えば，以下のクロスエントロピー基準により学習することができる．

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} - \sum_l P(\hat{y}_l|\boldsymbol{x}_l, \theta) \ln q(y_l|\boldsymbol{x}_l, \theta) \tag{5.19}$$

5.3.2　音声の構造的表象の構築

観測発話から該当する音響イベント集合を出力ラベルとして持つニューラルネットワークを学習することで，観測発話から式 (5.25) により全音響イベント間のバタチャリヤ距離を計算することができる．構造的表象は，全イベント間の分布間距離を並べた特徴量である．従って，音響イベントを音素として考えた場合，DNN により推定する音素数を K とすると，最終的に観測発話から得られるバタチャリヤ距離は自身との距離を除けば，常に $K(K-1)/2$ 次元となり，それぞれの次元の持つ意味は必ず対応が取れている．ガウス分布により各音素の分布をモデル化した場合，構造的表象はアフィン変換に不変であり，話者性の変化に対して頑健である [36]．これは特徴量の分布の形状が，変換による前後で共にガウス分布であるという仮定に基づく．しかし，DNN により特徴量分布の形を陽に決定せずにバタチャリヤ距離を式 (5.25) を用いて計算した場合，DNN が精確に特徴量空間の対応を学習できたと仮定すると，あらゆる全単射である変換に対して不変となると考えられる．そのため，理想的には全ての発話に対してバタチャリヤ距離は等しくなる．

5.4　特徴量ドメインにおける評価：言語識別への利用

5.4.1　言語識別システムへの利用

提案法である識別的アプローチにより計算された分布間距離を特徴量として用いる場合の有効性を，後段の識別タスクに時系列処理が介在しない言語識別システムにおいて評価

する．識別的アプローチにより計算されたバタチャリヤ距離を用いた音声の構造的表象を特徴量として用いる際，観測発話の発話数によるバタチャリヤ距離の計算精度が問題となる．識別的アプローチによって計算されたバタチャリヤ距離の言語識別への応用を考えた場合，数発話のような少量のデータから推定する必要が往々にして生じる．しかし，データが少量の場合，その発話に対応する DNN を学習することは困難である．識別モデルの性質上，特徴量空間をそれなりの精度で埋める必要がある．そのため，本来有意に発生するであろうラベルが，その少量のデータ中に発生しなかった場合，式 (5.18) によるバタチャリヤ距離の計算は大きく誤ることが予想される．これは音響イベントを音素ラベル等で考えた場合，発話内容によっては発声されない音素があることは容易に想像がつく．また，仮に DNN が正しく学習された場合においても，観測特徴量は発話内容によって偏りが生じるため，事後分布の全積分自体が誤ると想定される．そこで，これらの問題を回避するために，DNN の適応とサンプリング法を導入する．あらかじめ大量のデータによりグローバルな DNN を学習しておき，これを入力発話によって適応することで少量のデータでも頑健に DNN パラメータの推定を行うことができる．また，サンプリングにより発話内容に依存せずに特徴量空間全体からデータ集合を生成することが可能となる．

DNN を用いて識別的に計算された音声の構造的表象を特徴量として用いる言語識別システムの概要を図 5.1 に示す．前提条件として多言語のデータには音素ラベルが付与されておらず，音素ラベルが付与されているデータは一部の言語のみである．まず，複数言語のデータを用いて学習された i-vector モデルにより，観測発話に対する i-vector を計算する．

観測される音響特徴量 $\boldsymbol{x}_t$ が K 混合の対角共分散行列を持つ GMM でモデル化される Universal Background Model (UBM) から生成されると仮定する．

$$\boldsymbol{x}_t \sim \sum_{k=1}^{K} c_k \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_k(0), \boldsymbol{\Sigma}_k) \tag{5.20}$$

ここで，c_k,$\boldsymbol{\mu}_k(0)$,$\boldsymbol{\Sigma}_k$ はそれぞれインデクス k のガウス分布の重み，平均，共分散行列である．UBM は多種多様な言語，話者の音声により学習した GMM であり，ある言語 s の発声特徴量の確率密度関数をこの UBM のパラメータを用いて

$$\boldsymbol{x}_t(s) \sim \sum_{k=1}^{K} c_k \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_k(s), \boldsymbol{\Sigma}_k) \tag{5.21}$$

$$\boldsymbol{\mu}(s) = \boldsymbol{\mu}(0) + \boldsymbol{T}\boldsymbol{w}(s) \tag{5.22}$$

として平均ベクトルの線形変換としてのモデル適応により表すことができると考える．ここで，$\boldsymbol{T}$ は主成分分析によって得られる $D \times M$ の基底ベクトルを並べた行列であり，$\boldsymbol{w}(s)$ は重みベクトルである．具体的なパラメータ学習や重み推定の導出は原論文を参照して頂きたいが，この重みベクトルは，ある発話が得られた時，UBM を初期値とした Maximum

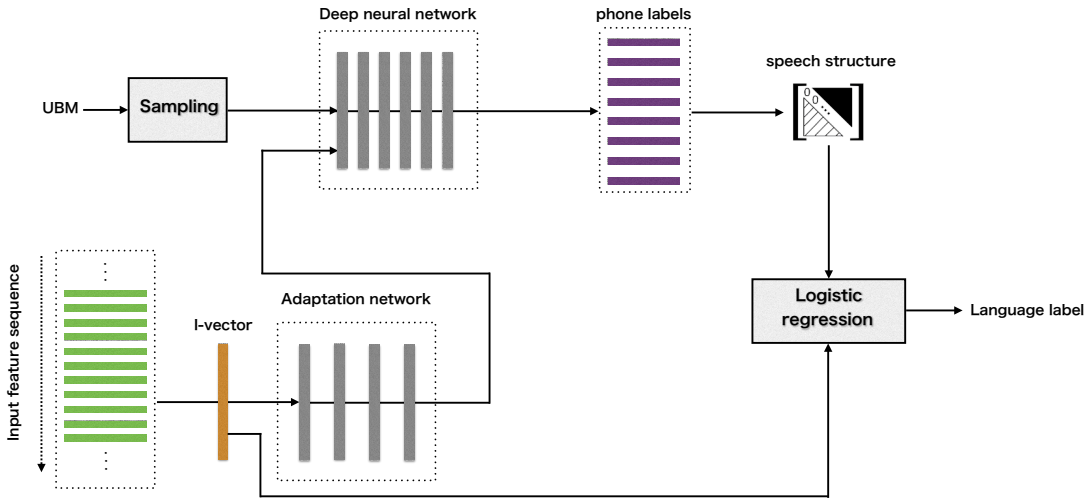


図 5.1: 想定する言語識別システムの概要

a posterior (MAP) 適応により得られた言語依存の GMM の平均ベクトルを用いて計算することができ，これを i-vector と呼ぶ．その後，推定された i-vector を入力特徴量系列に連結したものを入力として DNN により推定する．

その後，複数言語のデータを用いて学習された i-vector モデルにより，観測発話に対する i-vector を計算する．この i-vector を用いて，観測特徴量から音素を識別するニューラルネットワークの発話適応を行う．なお，適応に用いるベースとなるニューラルネットワークは，音素ラベルが存在する 1 つの言語のデータのみを用いて学習したものを用いる．その後，適応されたニューラルネットワークを用いてサンプリングにより音素間の分布間距離を推定し，構造的表象を構築する．これを特徴量として i-vector と連結しロジスティック回帰により言語識別を行う．

5.4.2　i-vector を用いた DNN の話者適応

言語識別に用いられる i-vector は言語による違いのだけでなく話者による違いも表現していると考えられる．そのため，DNN の発話適応は，Miao らの i-vector を用いた適応法である SAT-DNN を利用する [53]．この手法は，通常の DNN の入力に i-vector からのニューラルネット出力を足し合わせることによりモデル適応を行う．SAT-DNN のモデル構造を図 5.2 に示す．SAT-DNN は MFCC などの特徴量 $\boldsymbol{o}_t$ を入力として，ラベル $\boldsymbol{y}_t$ を

$$P(\boldsymbol{y}_t) = g(\boldsymbol{a}_t) \tag{5.23}$$

$$\boldsymbol{a}_t = \boldsymbol{o}_t + f(\boldsymbol{i}_s) \tag{5.24}$$

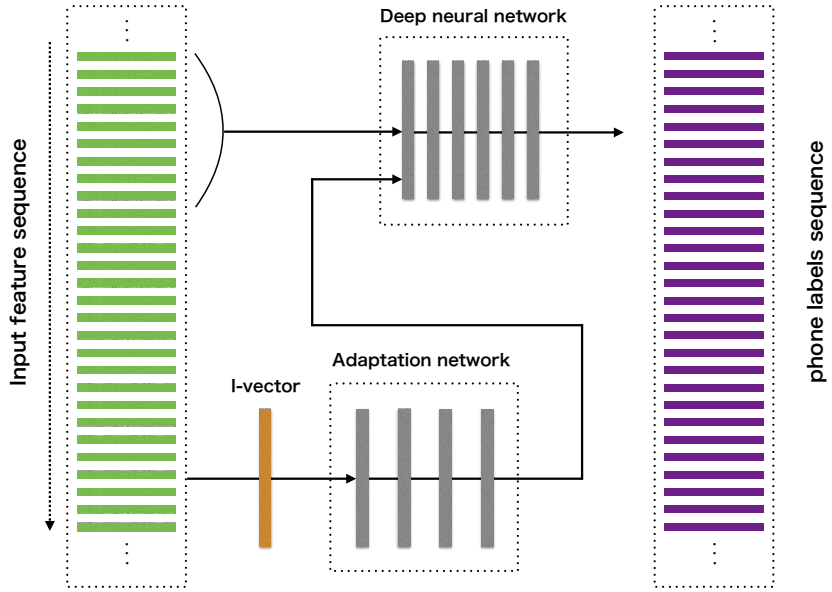


図 5.2: i-vector を用いた DNN のモデル適応

として識別する．ただし，$\boldsymbol{i}_s$ は入力話者を表現する i-vector である．また，$f(\cdot), g(\cdot)$ はそれぞれ，発話適応用の DNN，SAT モデルの DNN を表す．

学習の際は，まず，グローバルな DNN を学習する．その後，これを $g(\cdot)$ の暫定値とし，$g(\cdot)$ のパラメータを固定したまま誤差逆伝搬法により $f(\cdot)$ を学習する．その後，逆に $f(\cdot)$ のパラメータを固定して，$g(\cdot)$ のパラメータを更新することで，SAT モデルが学習される．この枠組みにより学習することにより，認識時は，各入力発話に対応する i-vector を入力することで入力層の対してバイアス適応の形で発話適応が可能となる．

なお，当然ながら DNN の学習には音素ラベル，もしくは音素状態ラベルが必要となる．不特定多数の言語に対して共通の音素ラベルを用意することは困難であるため，DNN は音素ラベルの存在する 1 言語の音声のみによって学習を行う．

5.4.3　発話適応 DNN を用いたバタチャリヤ距離の計算

式 (5.18) を発話適応により得られた DNN モデルパラメータ θ_c により計算することで，安定した事後確率の計算が可能となる．しかし，観測発話のデータ数が少ないという問題は依然として残る．そのため，データの偏りによって，特徴量空間全体に対する全積分を計算する必要のある分布間距離推定は計算することができない．そこで，話者非依存な分布，もしくはそれを基に適応を行った分布を用いて，特徴量空間をサンプリングすること

によりこれを回避する．式 (5.18) をベースとして，分布間距離を

$$\begin{aligned} BD(a,b) \;=\; & \ln \frac{1}{N} \sum_{l} \sqrt{P(y_n = a|\boldsymbol{x}_n, \theta_c) P(y_n = b|\boldsymbol{x}_n, \theta_c)} \\ & + \frac{1}{2} \ln \frac{1}{L} \sum_{l} P(y_l = a) \\ & + \frac{1}{2} \ln \frac{1}{L} \sum_{l} P(y_l = b) \end{aligned} \tag{5.25}$$

として推定する．ただし，$\{\boldsymbol{x}_n\}$ は DNN の学習に用いたデータと同じ言語のデータによって構築された UBM，もしくは UBM を基に発話適応を行った GMM より得られるサンプル集合である．

$$\boldsymbol{x}_n \sim \mathrm{GMM}(\phi_{global}) \tag{5.26}$$

$$or$$

$$\boldsymbol{x}_n \sim \mathrm{GMM}(\phi_c) \tag{5.27}$$

また，事前分布 $P(a), P(b)$ は，グローバルな DNN を学習した際の学習データ中の出現頻度により近似する．これは，話者により音素の発生頻度が大きく異なることはないという仮定に基づく．言語によって音素の発生頻度自体は変化すると考えらえるが，DNN は英語音声のみで学習しているため，アンカーモデルとして機能すると考えられるため，事前分布 $P(a), P(b)$ は全ての話者で英語の出現頻度を共通として用いた．なお，これは DNN を音響モデルとして利用する場合に用いられる近似と同じである．

$$\begin{aligned} P(a) &= \frac{1}{L} \sum_{l} P(y_l = a) \\ P(b) &= \frac{1}{L} \sum_{l} P(y_l = b) \end{aligned} \tag{5.28}$$

ただし，サンプリングにより計算する都合上，周辺分布と事前分布との誤差により推定されたバタチャリヤ距離が負の値となる場合があるため，注意が必要となる．

5.4.4　構造的表象と言語識別

提案手法により推定した分布間距離によって構築した構造的表象と i-vector を連結したものを特徴量とし，ロジスティック回帰により言語を識別する．音声の構造的表象は，分布間距離を特徴量として利用する手法の一つである．性別や年齢などの話者性の変換は，ケプストラム空間上におけるアフィン変換と対応している．構造的表象は音響イベントの分布間距離を用いて定義されるケプストラム空間上の構造体である．各音響イベントの分布

表 5.1: 学習／評価データ中における各言語の発話数

Language	Test set	Train set
Arabic	240	906
Bengali	240	200
Chinese	1194	3577
English	720	5739
Farsi	240	717
German	240	970
Hindustani	720	1135
Japanese	240	2050
Korean	240	1655
Russian	480	440
Spanish	720	3007
Tamil	480	1237
Thai	240	200
Vietnamese	480	707

をガウス分布によりモデル化することで分布間距離はアフィン変換不変となる．そのため，構造的表象を特徴量として用いることで，話者性などに起因するバラツキに対して頑健なシステムを構築することが可能となる．

さて，提案手法により識別的に推定された分布間距離を用いて構造的表象を構築する．前述のように i-vector を用いることによって DNN のモデル適応を行うため，i-vector を計算できさえすれば観測発話中に全音素が出現する必要なく，入力発話から式 (5.25) により全音素間のバタチャリヤ距離を計算することができる．しかし言語によって音素が異なるため，本予備実験では，DNN は音素ラベルの存在する言語のみにより学習する．この際，言語によって i-vector が学習に用いた言語のもつ i-vector と比較して大きく異なる場合，DNN が正しく適応できない．これによって適応後の DNN が想定する入力特徴量の分布が実際の分布から大きく崩れること想定される．この適応後のモデルを用いてサンプリングにより得た事後確率で構造的表象を構築した場合，本来観測特徴量に対して不変である構造が，逆に発話によっては異なる構造に変換される．つまり，学習に用いた言語をアンカーとした特徴量が生成されることが期待できる．

5.4.5　言語識別実験

言語識別用のデータベースである NIST LRE2007 データベースをベースとして言語識別実験により提案法の有効性を評価した．i-vector の学習は，NIST LRE2003，NIST LRE2005，NIST LRE2007 の学習セットを用いて行った．各言語の学習データ，評価データ中の発話

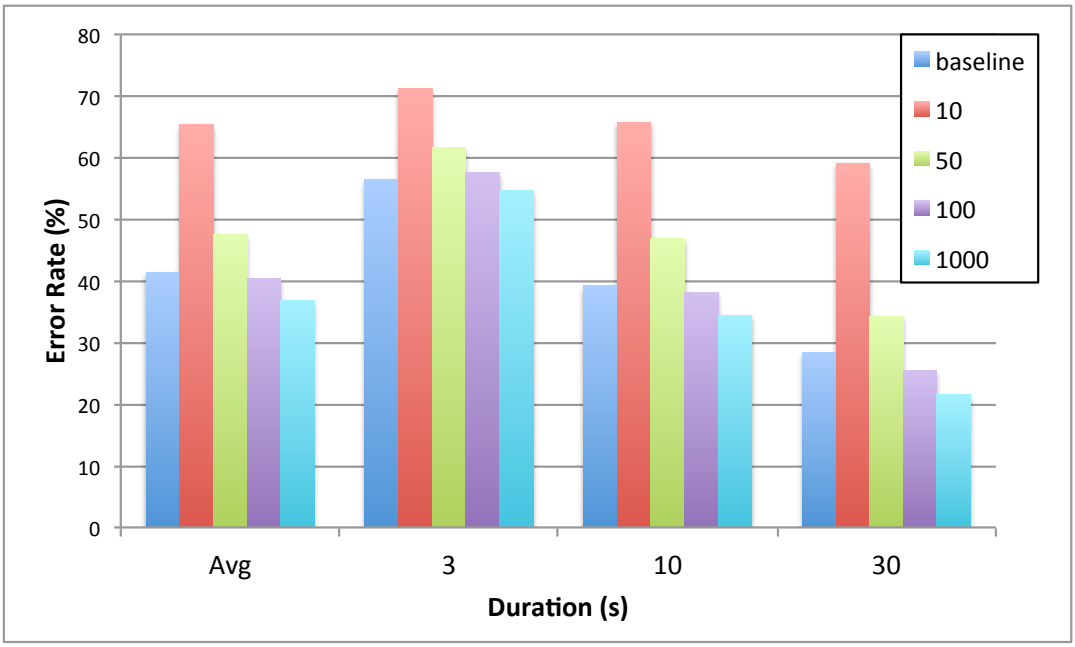


図 5.3: NIST LRE2007 データベースにおける言語識別誤り率

数を表 5.1 に示す．学習データ中の言語は英語が最も多いため，DNN の学習データは英語のデータである WSJ を用いて学習した．また，サンプリングに用いる UBM も WSJ を用いて学習した．評価は LRE07 のテストセットを用いており，識別対象は学習データ中の言語数と同じ 14 言語である．なお，言語クローズドセットであるため，評価セット中に未知の言語は存在しない．

DNN の入力特徴量は MFCC12 次元と C0 の計 13 次元を当該フレームとその前後 5 フレームの計 11 フレーム連結したものを用いた．DNN は各層の初期値を RBM により学習し，中間層の層数は 6 層，話者適応に用いるネットワークは中間層が 4 層に設定した．また各層のノード数は全て 1024 としている．サンプリングに用いる UBM の混合数は 1024 とし，サンプリングにより出力するデータは各発話毎に 10, 50, 100, 1000 サンプルの 4 通りを用意した．

ベースラインとサンプリング数毎の提案手法の識別実験結果を図 5.3，5.4 に示す．ベースラインは i-vector のみで識別したものであり，提案法はそれぞれのサンプリング数から計算された分布間距離により得られる構造的表象を i-vector と連結したものを用いたものである．なお，C_{avg} は LRE タスクで採用されている評価指標である．

$$C_{avg} = \frac{1}{N_L} \sum_{L_T} \left\{ \begin{array}{l} C_{Miss} P_{Target} P_{Miss}(L_T) \\ + \sum_{L_N} C_{FA} P_{Non-Target} \\ \qquad \times P_{FA}(L_T, L_N) \\ + C_{FA} P_{Out-of-Set} P_{FA}(L_T, L_O) \end{array} \right\} \tag{5.29}$$

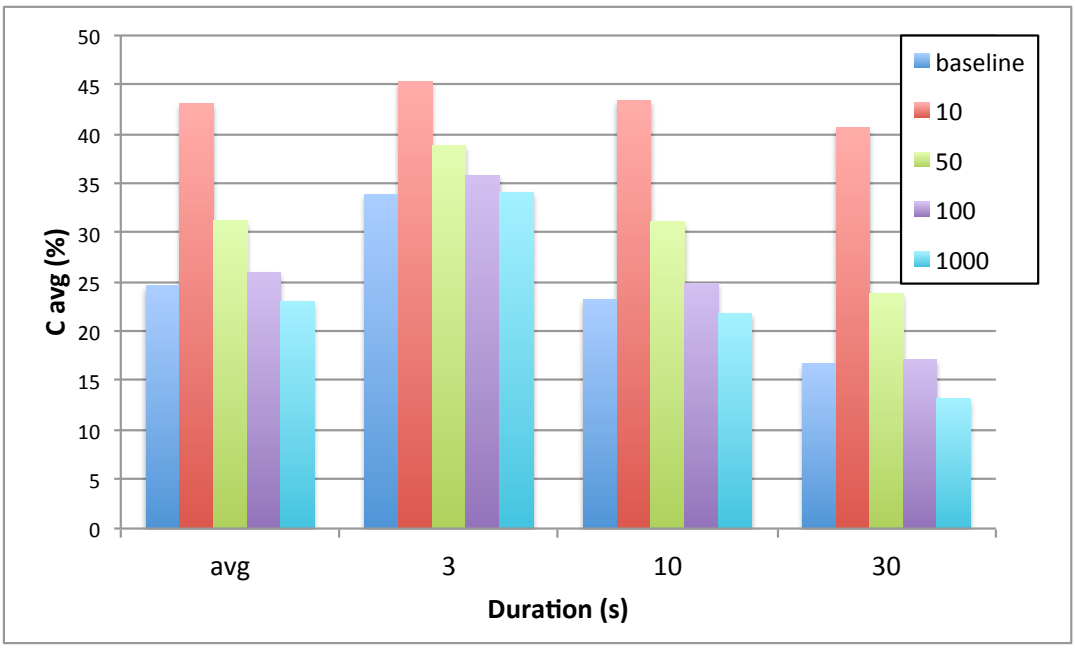


図 5.4: NIST LRE2007 データベースにおける言語識別における平均コスト（ C_{avg} ）

LRE07 の場合，$C_{Miss} = C_{FA} = 1.0$ であり，$P_{Target} = 0.5$ である．なお，言語クローズドセットを用いた評価であるため，$P_{Out-of-Set} = 0$ としている．また，$P_{Non-Target}$ は

$$P_{Non-Target} = (1 - P_{Target} - P_{Out-of-Set})/(N_L - 1) \tag{5.30}$$

として計算することができる．また，Duration は観測発話の秒数を表しており，秒数が短いほど難しいタスクとなる．

結果により，サンプリング数が少ない場合はベースラインと比較して認識誤りが増加してしまう．これは，サンプリング数が少ない場合，特徴量空間全体に対しサンプルの偏りが生じてしまうため，分布間距離の推定が誤るためと考えられる．しかし，サンプリング数が 100 程度あれば提案法によって認識誤りを低下することが可能となる．特に，サンプリング数が 1000 の場合，ベースラインのシステムと比較して 10.85%の認識誤り率の削減が可能となった．

5.5　音響モデルドメインにおける評価：話者適応への利用

5.5.1　再学習による DNN 音響モデルの話者適応

DNN は多層の複雑な構造を持つために，各パラメータの意味づけが直感的でない．そのため，ガウス混合モデルのように話者性に対応するパラメータの制御が困難である．そこで，あらかじめ大量の不特定話者のデータにより学習された不特定話者モデルのパラメータを初期値として，適応データを利用した追加の誤差逆伝搬法による再学習により全体のモデルパラメータを更新する手法が用いられることが多い．

新しい未知話者の適応データ $\{\boldsymbol{x}'_1, \boldsymbol{x}'_2, \ldots, \boldsymbol{x}'_{T'}\}$ が得られた際に，この適応データを用いて DNN を再学習することができる．まず，適応データを不特定話者モデルを用いてデコードすることにより，フレーム毎の音素状態事後確率ラベル $\{p_d(y'_1 = q|\boldsymbol{x}'_1), p_d(y'_2 = q|\boldsymbol{x}'_2), \ldots, p_d(y'_{T'} = q|\boldsymbol{x}'_{T'})\}$ を用意する．その後，不特定話者モデルのパラメータ Θ_g を初期値として，再学習によりパラメータを更新する．再学習時の目的関数 (5.32) は不特定話者モデルの学習時と同じであるが，誤差逆伝搬法により学習時のエポック数や学習係数等の制御を行うことで過適応を抑える．

$$\hat{\Theta}_{adapt} = \underset{\Theta}{\operatorname{argmin}} \frac{1}{T'} \sum_{t'=1}^{T'} D_{adapt}(\boldsymbol{x}_{t'}) \tag{5.31}$$

$$D_{adapt}(\boldsymbol{x}_{t'}) = \left\{ -\sum_{q}^{Q} p_d(y'_{t'} = q|\boldsymbol{x}'_{t'}, \Theta_g) \ln p(y_{t'} = q|\boldsymbol{x}'_{t'}, \Theta) \right\} \tag{5.32}$$

なお，適応データの音素状態事後確率ラベルを用意する際に，正解テキストを用いて強制アライメントを行うことで教師あり適応，正解テキストを用いずにデコード結果を用いることで教師なし適応となる．

5.5.2　分布間距離を制約として用いた話者適応

再学習法の学習時の目的関数を変更する手法として KL 距離を用いた正則化手法が提案されている [13]．再学習による DNN の話者適応は一般にエポック数や学習係数の制御によってモデルの過適応を抑えるが，適応データが少量の場合，もしくは教師なし適応のような適応データの音素状態事後確率ラベルが不安定である場合，学習係数等の制御では不十分である．これは DNN の持つ高い表現力により，適応データの偏りなどによって音素状態を識別するモデルの性質が簡単に失われてしまうためである．そこで，KL 距離を利用した正則化をかけることで，少量の適応データでも初期モデルからモデルパラメータが大きく外れないような制約をかけて適応を行う．

再学習は，あらかじめ学習された不特定話者モデルのパラメータ Θ_g を初期値として，目

的関数 (5.34) を用いてパラメータの更新を行う．

$$\hat{\Theta}_{adapt} = \underset{\Theta}{\operatorname{argmin}} \frac{1}{T'} \sum_{t'=1}^{T'} D_{kl}(\boldsymbol{x}_{t'}) \tag{5.33}$$

$$D_{kl}(\boldsymbol{x}_{t'}) = \left\{ -\sum_{i}^{Q} \tilde{p}(y_{t'} = q_i|\boldsymbol{x}_{t'}) \ln p(y_{t'} = q_i|\boldsymbol{x}_{t'}, \Theta) \right\} \tag{5.34}$$

$$\begin{aligned} \tilde{p}(y_{t'} = q_i|\boldsymbol{x}_{t'}) = \rho\, & p(y_{t'} = q_i|\boldsymbol{x}_{t'}, \Theta_g) \\ & +(1-\rho)\, p_d(y_{t'} = q_i|\boldsymbol{x}_{t'}, \Theta_g) \end{aligned} \tag{5.35}$$

ここで，$p(y_{t'} = q_i|\boldsymbol{x}_{t'}, \Theta_g)$ は話者非依存モデルから計算される事後確率であり，ρ はその重みである．つまり，話者非依存モデルパラメータを事前分布とした正則化をかけつつ再学習を行うことに相当する．

5.5.3　目的関数への音声の構造的表象の導入

音声の構造的表象の考えを基にした場合，音響イベント間の分布間距離は話者によらず不変であることが望ましい．しかし，過適応が生じる際は，適応データの偏り等に起因する場合が多いため，適応後のモデルの想定する特徴量空間における音声の構造的表象が変化してしまうことが想定される．そこで，本提案手法では，音声の構造的表象を基にした正則化を導入することで，再学習による DNN 音響モデルの話者適応の際に過適応を抑える．再学習時の目的関数を，あらかじめ学習された不特定話者モデルのパラメータ Θ_g を初期値として

$$\hat{\Theta}_{adapt} = \underset{\Theta}{\operatorname{argmin}}\, \rho D_{st} + (1-\rho)\frac{1}{T'} \sum_{t'=1}^{T'} D_{adapt}(\boldsymbol{x}_t) \tag{5.36}$$

$$D_{adapt}(\boldsymbol{x}_{t'}) = \left\{ -\sum_{i}^{Q} p_d(y_{t'} = q_i|\boldsymbol{x}_{t'}, \Theta_g) \ln p(y_t = q_i|\boldsymbol{x}_{t'}) \right\} \tag{5.37}$$

$$D_{st} = \frac{1}{\hat{Q} \times \hat{Q}} \left| \sum_{i}^{\hat{Q}} \sum_{j}^{\hat{Q}} BD^{SD}(q_i, q_j) - \sum_{i}^{\hat{Q}} \sum_{j}^{\hat{Q}} BD^{SI}(q_i, q_j) \right| \tag{5.38}$$

とする．ここで，$BD^{SD}(q_i, q_j)$ は適応データと適応後のモデルから得られるバタチャリヤ距離である．また，$BD^{SI}(q_i, q_j)$ は適応前の不特定話者モデルと，不特定話者のデータにより得られる，不特定話者モデルの持つ音響イベント間のバタチャリヤ距離である．$\hat{Q}$ は，音素状態をシェアリングした後の集合であり，これは DNN 音響モデルの出力であるトライフォン音素状態は次元数が大きいため，モノフォンなどの数まで削減することを視野に入れている．さらに，無音や子音などは話者によって不変である，もしくは大差がないこ

とが考えられるため，それらを除外したような集合等も考えられる．この目的関数を用いて誤差逆伝搬法による再学習によってパラメータの更新を行う．

この際のバタチャリヤ距離の計算には，各音響イベントに対する確率密度関数が必要となるため，ニューラルネットワークの学習時に各音響イベントの確率密度関数を計算することは非効率である．そこで，提案法によって分布間距離をニューラルネットワークを用いて識別的に計算する．これにより，誤差逆伝搬法のミニバッチ単位で容易に音声の構造的表象を計算することが可能となる．それぞれのバタチャリヤ距離はニューラルネットワークを用いて

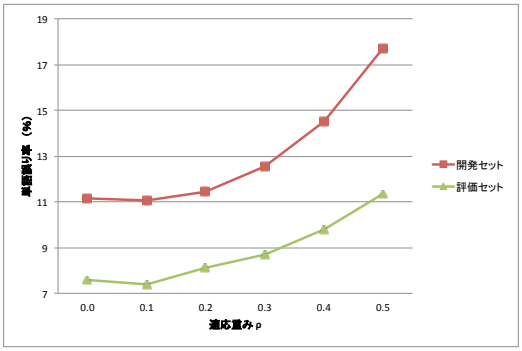
$$
\begin{aligned}
BD^{SD}(q_i, q_j) \approx & -\ln \frac{1}{T'} \sum_{t'} \sqrt{p(y_{t'} = q_i | \boldsymbol{x}_{t'}, \Theta) p(y_{t'} = q_j | \boldsymbol{x}_{t'}, \Theta)} && (5.39) \\
& + \frac{1}{2} \ln p(y = q_i) \\
& + \frac{1}{2} \ln p(y = q_j) \\
= & -\ln \frac{1}{T'} \sum_{t'} \sqrt{p(y_{t'} = q_i | \boldsymbol{x}_{t'}, \Theta) p(y_{t'} = q_j | \boldsymbol{x}_{t'}, \Theta)} && (5.40) \\
& + const.
\end{aligned}
$$

$$
\begin{aligned}
BD^{SI}(q_i, q_j) \approx & -\ln \frac{1}{T'} \sum_{t'} \sqrt{p(y_{t'} = q_i | \boldsymbol{x}_{t'}, \Theta_g) p(y_{t'} = q_j | \boldsymbol{x}_{t'}, \Theta_g)} && (5.41) \\
& + \frac{1}{2} \ln p(y = q_i) \\
& + \frac{1}{2} \ln p(y = q_j) \\
= & -\ln \frac{1}{T'} \sum_{t} \sqrt{p(y_{t'} = q_i | \boldsymbol{x}_{t'}, \Theta_g) p(y_{t'} = q_j | \boldsymbol{x}_{t'}, \Theta_g)} && (5.42) \\
& + const.
\end{aligned}
$$

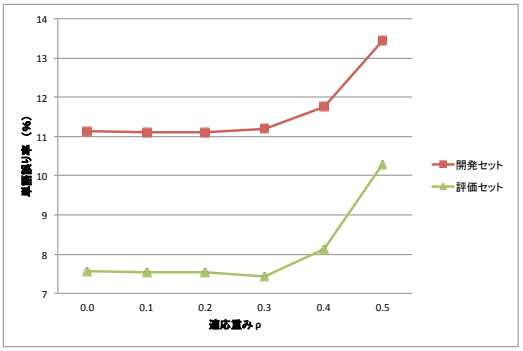
として計算する．なお，それぞれの音響イベントの事前確率に対応する $p(y = q_i)$ は，適応前と後で変化しないと仮定することで定数項として扱う．

5.5.4　話者適応実験

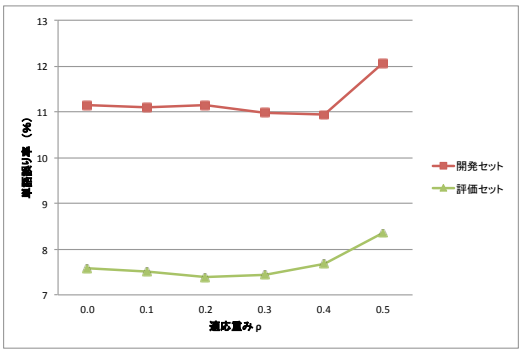
評価に用いるデータベースとして英語の大語彙音声認識用のデータベースである WSJ (01, 02) を用いた．学習データセットは 37,416 発話であり，評価データセットは 333 発話．また，デコード時の音響モデルスコアと言語モデルスコアの重み決定のため，開発データセット 503 発話が用意されている．また，評価セットは各話者約 40 発話，開発セットは各話者 50 発話程度ある．



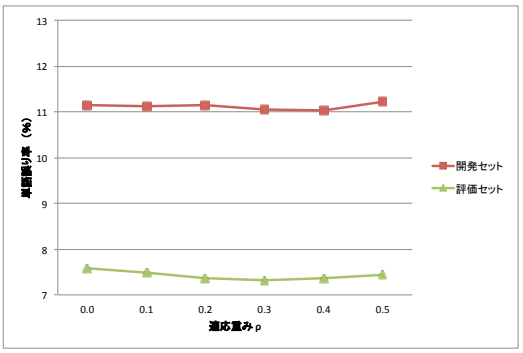
(a) シェアリングを行わず 1552 のトライフォン状態のまま計算した場合 (Prop. 1552)



(b) モノフォンラベルである 42 音素にシェアリングした場合 (Prop. 42)



(c) モノフォンラベルから無音ラベルに相当するものを除いた 39 のラベルにシェアリングした場合 (Prop. 39)



(d) 母音のみの 15 音素のモノフォンラベルにシェアリングした場合 (Prop. 15)

図 5.5: 適応重みを変化させた時の単語誤り率

認識システムは KALDI に付属の WSJ データベースの認識スクリプトをベースとして用いた [48]．入力特徴量はメルフィルタバンク出力の 40 次元であり，音響モデルは DNN/HMM である．ネットワークの入力は当該フレームとその前後 5 フレーム，計 11 フレームを入力とした．ネットワークの構造は中間層が 6 層であり，隠れ層は各層 1024 ノードである．また，活性化関数は maxout を採用している．出力は 1552 音素状態であり，学習データのラベルはあらかじめ学習した GMM/HMM のトライフォン音響モデルを用いた強制アライメントによってラベルを付与した．このトライフォン音響モデルは，特徴量として MFCC とその 1 次，2 次微分の 39 次元に平均分散正規化を行ったものを用いてモデル化した．

話者適応は教師なし適応を想定する．認識するデータに対して不特定話者モデルにより認識を行った結果を用いて擬似正解ラベルを作成する．これを適応データとして再学習によりモデル適応を行った．なお，適応時における音響モデルスコアと言語モデルスコアの重みも開発セットを用いて決定している．適応の際の音素状態数のシェアリングとして，1) シェアリングを行わず 1552 のトライフォン状態のまま計算した場合，2) モノフォンラベルである 42 音素にシェアリングした場合，3) モノフォンラベルから無音ラベルに相当

表 5.2: 認識実験結果：各状態シェアリングの条件，適応重み ρ における単語誤り率（開発セット/評価セット）

	適応なし	$\rho = 0.0$	$\rho = 0.1$	$\rho = 0.2$	$\rho = 0.3$	$\rho = 0.4$	$\rho = 0.5$
KL 正則化			11.20/7.51	11.20/7.53	11.25/7.55	11.32/7.57	11.34/7.57
Prop. 1552			11.06/7.39	11.42/8.12	12.55/8.67	14.50/9.80	17.69/11.34
Prop. 42	11.79/8.24	11.14/7.57	11.10/7.53	11.11/7.53	11.19/7.43	11.76/8.12	13.43/10.28
Prop. 39			11.09/7.50	11.15/7.39	10.98/7.43	10.94/7.69	12.06/8.36
Prop. 15			11.11/7.48	11.15/7.37	**11.05/7.32**	11.03/7.37	11.23/7.44

するものを除いた 39 のラベルにシェアリングした場合，4) 母音のみの 15 音素のモノフォンラベルにシェアリングした場合，の 4 パターンを検証した．それぞれの条件における単語誤り率を Fig. 5.5(a),5.5(b),5.5(c),5.5(d) に示す．また，それぞれの値の詳細は Table 5.2 に示す．$\rho = 0.0$ は正則化を行わずに再学習により適応した場合に相当し，Table 5.2 中の「適応なし」は適応を行っていない不特定話者モデルを用いて認識を行った結果である．また，Table 5.2 中の KL 正則化は KL 距離を制約として用いて再学習を行う先行研究 [13] である．

音素状態をシェアリングせずに用いた場合（Fig. 5.5(a), Prop. 1552）では，適応重みが $\rho = 0.1$ と小さい場合には単語誤り率が減少するが，それ以降は急速に誤り率が増加してしまう．また，モノフォンラベルである 42 音素にシェアリングした場合（Fig. 5.5(b), Prop. 42）では，適応重みの増加による誤り率の減少が抑えられており，音素状態をシェアリングすることの効果の有効性がわかる．さらに，モノフォンラベルから無音ラベルに相当するものを除いた 39 のラベルにシェアリングした場合（Fig. 5.5(c), Prop. 39）では，僅かではあるが誤り率が減少しており，話者によって共通であると考えられる無音に相当するラベルを取り除いたことによって構造特徴が安定したことが寄与したと考えられる．母音のみの 15 音素のモノフォンラベルにシェアリングした場合（Fig. 5.5(d), Prop. 15）では，適応重みを増加させた場合の単語誤り率の増加も抑えられており，適応重みが $\rho = 0.3$ の場合に最も良い単語誤り率が得られ，3.3%のエラー削減率が得られた．破擦音などの話者毎によって差が生じづらい子音の影響が取り除かれるため，さらに構造特徴が安定したと考えられる．

また，どの条件でも適応重みをより大きくした際に適応性能が程度の差はあれ低下することは共通して見られた．これは，構造的表象が時系列情報を吸収するため，適応重みを大きくした場合に逆にセグメント単位の音素状態識別性能が低下してしまうためだと考えられる．なお，先行研究である KL 距離を用いた正則化は本実験では適応性能の向上に寄与しなかった．これは，教師なし適応ではあるが比較的適応データの多い実験条件であるため，MAP 適応のように適応前と適応後のパラメータとの内挿を行う KL 距離の制約の効果が働かなかったためだと考えられる．

5.6　まとめ

本稿では，識別モデルを利用した分布間距離推定法と，これを用いた非言語情報の違いに対して頑健な特徴量表現である音声の構造的表象の利用を検討した．特徴量ドメインにおける評価として言語識別システムにおける特徴量としての利用を行なった．これにより分布間距離の実際の利用としてはデータ数が限られた場面での分布間距離の推定が必要となる．そのため，分布間距離であるバタチャリヤ距離を少量のデータから識別的に推定する手法について検討した．言語識別実験により，ロジスティック回帰の入力特徴量として i-vector にバタチャリヤ距離により得られた構造的表象を連結したもので認識を行ったところ，i-vector 単体と比較し 10.85%のエラー削減率が得られた．

また，音響モデルドメインにおける評価として，提案法により計算される分布間距離を制約として用いた音響モデルの話者適応を提案した．ケプストラム空間上における各音響イベントの分布をガウス分布として仮定した場合，音声の構造的表象は話者の違いに対して頑健な特徴量であると言える．ニューラルネットワークの学習時に音声の構造的表象を効率的に計算するために，本研究ではニューラルネットワークを用いて識別的に分布間距離を計算する手法を導入した．また，音声の構造的表象を構築する際の音響イベントの単位を，もともとのトライフォン音素状態からシェアリングすることで，無音や子音などの影響を取り除くことを行った．これにより，適応の際の正則化として音声の構造的表象が有効に機能することが実験によって明らかになった．最終的に，正則化を行わない再学習による話者適応と比較して，母音のみの 15 音素のモノフォンラベルにシェアリングした場合で 3.3%のエラー削減率が得られた．

全体として性能向上への寄与は小さいものの，音声学的知見とニューラルネットワークの融合が，期待された通りに有効に働くという実験結果は非常に有益であると言える．

第6章

結論

6.1　まとめ

本論文では，非言語情報の違いに頑健な特徴量表現に着目したニューラルネットワーク音声認識に関する研究を行った．現在の音声認識の精度はニューラルネットワークを用いた音声認識の出現により高いものとなってきたが，雑音や話者の違いなどに起因する非言語情報により認識性能が低下するという問題がある．そのため，非言語情報を適切に取り扱う枠組みがニューラルネットワーク音声認識においても非常に重要であると言える．従来の生成モデルに基づくアプローチは，生成モデルのパラメータの意味づけの容易さによって，非言語情報の操作に対する大きなバックグラウンドを持つ．これに対してニューラルネットワークは高い性能を持つが，非常に複雑なモデル構造を持つため，パラメータの意味づけと操作が直感的でなく困難であるという問題があった．

そこで，本論文では，従来の生成的アプローチをベースとした非言語情報に対する知見を基に，非言語情報の制御と密接な関係性を持つ，特徴量ドメイン，音響モデルドメインにおいて，生成モデルと識別モデルの融合によるニューラルネットワーク音声認識の精度向上を行った．まず，特徴量ドメインにおいて，非言語情報の一つである雑音に対して頑健な特徴量抽出を実現するため，従来の GMM によるアプローチとニューラルネットワークによるアプローチの融合を行った．音声特徴量のクラスタリングは従来の生成モデルが得意とするためガウス混合分布により行い，このクラスを識別性能の高いニューラルネットワークにより行った．これにより，それぞれのモデルの利点を組み合わせたモデルが実現できた．

また，音響モデルドメインにおいては，音響モデルの正規化学習の枠組みをニューラルネットワーク音声認識に導入した．音響モデルの正規化学習は環境・話者の違いに対する音響モデルの適応性能を向上するために用いられる．従来のガウス混合分布をベースとしたモデルにおいては，モデルパラメータの意味づけが直感的であるため，この正規化学習が容易であった．しかし，ニューラルネットワークはモデルパラメータが多く，かつ複雑な構成をしているため，正規化学習が困難である．そこで，ニューラルネットワークの学習時に生成的に話者情報を表す話者コードを同時に推定する手法を提案した．これにより適応性能の向上が実現することができた．

また，生成モデルと識別モデル，双方のモデルの性質のみに着眼してモデルの融合を行うだけでは，非言語情報の制御に対しての抜本的な解決策とならない．そこで，非言語情報，特に話者の違いに対する理論的バックグラウンドを持つ音声の構造的表象をニューラルネットワークにより計算する手法を提案した．これは，構造的表象の要素である分布間距離をニューラルネットワークによる識別的アプローチにより推定するものである．これにより，音声学的知見をニューラルネットワークへ導入することが可能となった．この手法の有効性を特徴量ドメイン，音響モデルドメインにおいて評価した．特徴量ドメインにおいては，言語識別の入力特徴量として用いることで識別性能の向上が実現できた．また，

音響モデルドメインにおいては，音響モデルの話者適応の際の制約として導入した．提案法はニューラルネットワークの学習時に同時に計算することが可能となるため，話者適応の際に効率的に音声の構造的表象の計算を行うことが可能となり，さらに適応性能の向上が可能となった．

本研究は単に従来の生成的アプローチと近年主流となっているニューラルネットワークに代表される識別的アプローチの融合に留まらず，音声学的知見に基づいたニューラルネットワークを用いた新しい特徴量表現の実現により，ニューラルネットワーク音声認識における非言語情報の制御に対して新たな道を切り拓くことができたと考える．

6.2　今後の展望

ニューラルネットワーク音声認識において，さらなる非言語情報の効果的な制御の実現を目指す上で重要となるのは，特徴量抽出と音響モデルの包括的なモデル化にあると考える．入力をより素な，例えば音声波形を直接用いる手法の研究は現在盛んに行われているが，これはニューラルネットワークの持つ高い識別性にのみ依存したモデルであることが多い．しかし，このような大規模なモデルを構築する際にこそ，従来の生成的アプローチにより培われた音声に対する知見が活きると考えられる．特に構造的表象に代表される従来の理論的背景を持つ手法は，ニューラルネットワークに導入する研究は大きな余地があると考えられる．

また，今後，音声認識システムとしては実際の応用場面を想定したものもさらに重要となると考えられる．その一つの例が，より砕けた話し言葉を対象としたものである．現在の音声認識は言語モデルの発展によりフィラーに対する対応力は比較的高くなったと言えるが，言い淀み，言い間違いの扱いは未だ困難である．これは言い淀みのモデル化が複雑であることが要因ではあるが，今後より自然な音声対話システム，音声インタフェースの実現を考えた上では避けて通ることができない問題であり，今後の研究が期待される．

謝辞

本研究ならびに本論文の執筆にあたり，多大なる御指導，御鞭撻を賜りました指導教員である峯松信明教授，並びに広瀬啓吉名誉教授に深く感謝いたします．峯松信明教授には修士課程からの 5 年間，研究の指導，論文執筆など熱心に指導頂きました．広瀬啓吉名誉教授には修士課程における研究の指導のみならず，博士課程における研究にも多くの的確な助言を頂きました．感謝の念が絶えません．また，研究活動を様々な面で支えて下さった高橋登技官，秘書の池上恵さん，折茂結実子さんに深く感謝します．

峯松研究室，広瀬研究室の皆様のおかげで 5 年間非常に多くのことを学ばせていただきました．論文の共著者として共に研究してくださった，皆様には深く感謝を申し上げます．また，研究だけではなく友人としても多くの研究室の方々にお世話になりました．齋藤大輔助教には，日頃から研究に関する多くのことについて相談に乗っていただきました．第 5 章の研究は，齋藤大輔助教のアイディアをベースとして私なりに試行錯誤を行い実現したものです．研究室の計算機環境等の整備も行って頂き，私の研究は助教のお力添えなくしては実現できなかったと言っても過言ではありません．感謝の念が絶えません第 3 章における研究については，博士課程の先輩である鈴木雅之氏の先行研究あってのものです．鈴木雅之氏には音声認識の研究を始めた当初から多くのことを学ばせていただきました．深く感謝いたします．残念ながら全ての方を挙げることはできませんが，研究室の先輩，同期，後輩の皆様には多くのことを学ばせて頂きました．大変実りの多い研究生活を送ることができました．この場を借りて感謝の意を申し上げます．

また，同志社大学の落合翼氏には学会で多くの意見交換をして頂きました．第 4 章の研究は，落合翼氏の先行研究あってのものです．深く感謝いたします．

本研究の一部は，日本学術振興会 科学研究費補助金（特別研究員奨励費 26・9167）の支援により行われました．ここに感謝の意を表します．

最後に，学生生活を支えて頂いた家族に感謝致します．

2015 年 12 月 1 日

柏木陽佑

参考文献

[1] 河原達也, 秋田祐哉, 三村正人, 政瀧浩和, 高橋敏. 衆議院会議録作成における音声認識システム全体の構成と評価. 2011.

[2] George Saon and Jen-Tzung Chien. Large-vocabulary continuous speech recognition systems: A look at some recent advances. *Signal Processing Magazine, IEEE*, Vol. 29, No. 6, pp. 18–33, 2012.

[3] Ellen Eide and Herbert Gish. A parametric approach to vocal tract length normalization. In *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, Vol. 1, pp. 346–348. IEEE, 1996.

[4] Olli Viikki and Kari Laurila. Cepstral domain segmental feature vector normalization for noise robust speech recognition. *Speech Communication*, Vol. 25, No. 1, pp. 133–147, 1998.

[5] Jean-Luc Gauvain and Chin-Hui Lee. Maximum a posteriori estimation for multivariate Gaussian mixture observations of markov chains. *Speech and audio processing, ieee transactions on*, Vol. 2, No. 2, pp. 291–298, 1994.

[6] Mark J.F. Gales and Philip C. Woodland. Mean and variance adaptation within the mllr framework. *Computer Speech & Language*, Vol. 10, No. 4, pp. 249–264, 1996.

[7] S.M. Ahadi and Philip C. Woodland. Combined Bayesian and predictive techniques for rapid speaker adaptation of continuous density hidden markov models. *Computer speech & language*, Vol. 11, No. 3, pp. 187–206, 1997.

[8] Abdel-rahman Mohamed, George E. Dahl, and Geoffrey Hinton. Acoustic modeling using deep belief networks. *Audio, Speech, and Language Processing, IEEE Transactions on*, Vol. 20, No. 1, pp. 14–22, 2012.

[9] Frank Seide, Gang Li, Xie Chen, and Dong Yu. Feature engineering in context-dependent deep neural networks for conversational speech transcription. In *Automatic*

Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on, pp. 24–29. IEEE, 2011.

[10] Kaisheng Yao, Dong Yu, Frank Seide, Hang Su, Li Deng, and Yifan Gong. Adaptation of context-dependent deep neural networks for automatic speech recognition. In *Spoken Language Technology Workshop (SLT), 2012 IEEE*, pp. 366–369. IEEE, 2012.

[11] Hank Liao. Speaker adaptation of context dependent deep neural networks. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pp. 7947–7951. IEEE, 2013.

[12] George Saon, Hagen Soltau, David Nahamoo, and Michael Picheny. Speaker adaptation of neural network acoustic models using i-vectors. In *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*, pp. 55–59. IEEE, 2013.

[13] Dong Yu, Kaisheng Yao, Hang Su, Gang Li, and Frank Seide. KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pp. 7893–7897. IEEE, 2013.

[14] Shaofei Xue, Ossama Abdel-Hamid, Hui Jiang, and Lirong Dai. Direct adaptation of hybrid DNN/HMM model for fast speaker adaptation in LVCSR based on speaker code. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pp. 6339–6343. IEEE, 2014.

[15] Ossama Abdel-Hamid and Hui Jiang. Rapid and effective speaker adaptation of convolutional neural network based models for speech recognition. In *INTERSPEECH*, pp. 1248–1252, 2013.

[16] Jian Xue, Jinyu Li, Dong Yu, Mike Seltzer, and Yifan Gong. Singular value decomposition based low-footprint speaker adaptation and personalization for deep neural network. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pp. 6359–6363. IEEE, 2014.

[17] Takuya Yoshioka, Anton Ragni, and Mark J.F. Gales. Investigation of unsupervised adaptation of DNN acoustic models with filter bank input. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pp. 6344–6348. IEEE, 2014.

[18] Tsubasa Ochiai, Shigeki Matsuda, Xugang Lu, Chiori Hori, and Shigeru Katagiri. Speaker adaptive training using deep neural networks. In *Acoustics, Speech and Signal*

Processing (ICASSP), 2014 IEEE International Conference on, pp. 6349–6353. IEEE, 2014.

[19] Nobuaki Minematsu. Yet another acoustic representation of speech sounds. In *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, Vol. 1, pp. I–585. IEEE, 2004.

[20] Mehryar Mohri, Fernando Pereira, and Michael Riley. Weighted finite-state transducers in speech recognition. *Computer Speech & Language*, Vol. 16, No. 1, pp. 69–88, 2002.

[21] Andrew Senior, Georg Heigold, Marc'Aurelio Ranzato, and Ke Yang. An empirical study of learning rates in deep neural networks for speech recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pp. 6724–6728. IEEE, 2013.

[22] Geoffrey E. Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural computation*, Vol. 18, No. 7, pp. 1527–1554, 2006.

[23] Geoffrey E. Hinton. A practical guide to training restricted Boltzmann machines. *Momentum*, Vol. 9, No. 1, p. 926, 2010.

[24] Hynek Hermansky, Daniel P.W. Ellis, and Sangita Sharma. Tandem connectionist feature extraction for conventional HMM systems. In *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, Vol. 3, pp. 1635–1638. IEEE, 2000.

[25] Li Deng, Mike Seltzer, Dong Yu, Alex Acero, Abdel-rahman Mohamed, and Geoffrey E. Hinton. Binary coding of speech spectrograms using a deep auto-encoder. In *Interspeech*, pp. 1692–1695. Citeseer, 2010.

[26] Hervé A. Bourlard and Nelson Morgan. *Connectionist speech recognition: a hybrid approach*, Vol. 247. Springer Science & Business Media, 2012.

[27] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, Vol. 9, No. 8, pp. 1735–1780, 1997.

[28] Mike Schuster and Kuldip K. Paliwal. Bidirectional recurrent neural networks. *Signal Processing, IEEE Transactions on*, Vol. 45, No. 11, pp. 2673–2681, 1997.

[29] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pp. 6645–6649. IEEE, 2013.

[30] Stefan Kombrink, Tomas Mikolov, Martin Karafiát, and Lukás Burget. Recurrent neural network based language modeling in meeting recognition. In *INTERSPEECH*, pp. 2877–2880, 2011.

[31] Jasha Droppo, Li Deng, and Alex Acero. Evaluation of the SPLICE algorithm on the Aurora2 database. In *INTERSPEECH*, Vol. 1, pp. 217–220, 2001.

[32] Jasha Droppo, Li Deng, and Alex Acero. Evaluation of SPLICE on the Aurora 2 and 3 tasks. In *INTERSPEECH*, 2002.

[33] Jinyu Li, Michael L. Seltzer, and Yifan Gong. Improvements to VTS feature enhancement. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pp. 4677–4680. IEEE, 2012.

[34] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pp. 1096–1103. ACM, 2008.

[35] Michael Pitz, Sirko Molau, Ralf Schlüter, and Hermann Ney. Vocal tract normalization equals linear transformation in cepstral space. In *INTERSPEECH*, pp. 2653–2656, 2001.

[36] Nobuaki Minematsu, Satoshi Asakawa, Masayuki Suzuki, and Yu Qiao. Speech structure and its application to robust speech processing. *New Generation Computing*, Vol. 28, No. 3, pp. 299–319, 2010.

[37] Yu Qiao, Masayuki Suzuki, Nobuaki Minematsu, and Keikichi Hirose. Structure-constrained distribution matching using quadratic programming and its application to pronunciation evaluation. In *Pattern Recognition (ACPR), 2011 First Asian Conference on*, pp. 350–354. IEEE, 2011.

[38] 内田秀継, 齋藤大輔, 峯松信明. 音声の構造的表象を用いた未観測調音運動の推定法の検討. 電子情報通信学会信学技報, 2016.

[39] Mark J.F. Gales. Model-based approaches to handling uncertainty. In *Robust Speech Recognition of Uncertain or Missing Data*, pp. 101–125. Springer, 2011.

[40] Mohamed Afify, Xiaodong Cui, and Yuqing Gao. Stereo-based stochastic mapping for robust speech recognition. *Audio, Speech, and Language Processing, IEEE Transactions on*, Vol. 17, No. 7, pp. 1325–1334, 2009.

[41] Jort F. Gemmeke, Tuomas Virtanen, and Antti Hurmalainen. Exemplar-based sparse representations for noise robust automatic speech recognition. *Audio, Speech, and Language Processing, IEEE Transactions on*, Vol. 19, No. 7, pp. 2067–2080, 2011.

[42] Masayuki Suzuki, Takuya Yoshioka, Shinji Watanabe, Nobuaki Minematsu, and Keikichi Hirose. MFCC enhancement using joint corrupted and noise feature space for highly non-stationary noise environments. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pp. 4109–4112. IEEE, 2012.

[43] Masayuki Suzuki, Takuya Yoshioka, Shinji Watanabe, Nobuaki Minematsu, and Keikichi Hirose. Feature enhancement with joint use of consecutive corrupted and noise feature vectors with discriminative region weighting. *Audio, Speech, and Language Processing, IEEE Transactions on*, Vol. 21, No. 10, pp. 2172–2181, 2013.

[44] Andrew L. Maas, Quoc V. Le, Tyler M. O'Neil, Oriol Vinyals, Patrick Nguyen, and Andrew Y. Ng. Recurrent neural networks for noise reduction in robust ASR. In *INTERSPEECH*. Citeseer, 2012.

[45] http://aurora.hsnr.de/aurora-2.html.

[46] Hans-Günter Hirsch and David Pearce. The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In *ASR2000-Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop (ITRW)*, 2000.

[47] D. Pierce and Ansela Gunawardana. Aurora 2.0 speech recognition in noise: Update 2. In *Proc. ICSLP Session on Noise Robust Rec., Colorado, USA*, 2002.

[48] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. The kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, December 2011. IEEE Catalog No.: CFP11SRW-USB.

[49] Victor Zue, Stephanie Seneff, and James Glass. Speech database development at MIT: TIMIT and beyond. *Speech Communication*, Vol. 9, No. 4, pp. 351–356, 1990.

[50] Chang Huai You, Kong Aik Lee, and Haizhou Li. GMM-SVM kernel with a Bhattacharyya-based distance for speaker recognition. *Audio, Speech, and Language Processing, IEEE Transactions on*, Vol. 18, No. 6, pp. 1300–1312, 2010.

[51] Georg Heigold, Hermann Ney, Patrick Lehnen, Tobias Gass, and Ralf Schlüter. Equivalence of generative and log-linear models. *Audio, Speech, and Language Processing, IEEE Transactions on*, Vol. 19, No. 5, pp. 1138–1148, 2011.

[52] Jinyu Li, Rui Zhao, Jui-Ting Huang, and Yifan Gong. Learning small-size DNN with output-distribution-based criteria. In *Proc. Interspeech*, 2014.

[53] Yajie Miao, Lu Jiang, Hao Zhang, and Florian Metze. Improvements to speaker adaptive training of deep neural networks. In *Spoken Language Technology Workshop (SLT), 2014 IEEE*, pp. 165–170. IEEE, 2014.

発表文献

学術論文

[1] 柏木陽佑, 齋藤大輔, 峯松信明, 雑音環境下音声認識のためのディープニューラルネットワークを用いた識別的区分線形変換, 電子情報通信学会学生論文特集（和文論文雑誌 D）, 2016. 公表予定

[2] 渡辺美知子, 柏木陽佑, 後続句の複雑さが文節境界におけるフィラーの出現率に与える影響, 音声研究, vol. 18, No. 1, pp. 45–56, 2014.

国際会議論文

[3] Y. Kashiwagi, D. Saito, N. Minematsu, “ DIVERGENCE ESTIMATION BASED ON DEEP NEURAL NETWORKS FOR LANGUAGE IDENTIFICATION, ”Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2016. 発表予定

[4] Y. Luan, D. Saito, Y. Kashiwagi, N. Minematsu, K. Hirose, “ Semi-supervised noise dictionary adaptation for exemplar-based noise robust speech recognition, ”Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2014.

[5] Y. Kashiwagi, D. Saito, N. Minematsu, K. Hirose, “ Discriminative piecewise linear transformation based on deep learning for noise robust automatic speech recognition, ”Proc. Automatic Speech Recognition and Understanding Workshop (ASRU), 2013.

[6] Y. Kashiwagi, M. Suzuki, N. Minematsu, K. Hirose,“ Audio-visual feature integration based on piecewise linear transformation for noise robust automatic speech recognition, ”Proc. Spoken Language Technology (SLT), 2012.

国内研究会・全国大会

[7] 柏木陽佑, 齋藤大輔, 峯松信明, 構造的表象を制約として用いたニューラルネットワーク音響モデルの話者適応の検討, 日本音響学会秋季講演論文集, 2015. 発表予定

[8] 柏木陽佑, 齋藤大輔, 峯松信明, 識別的アプローチによる分布間距離推定の検討とその言語識別への応用, 電子情報通信学会技術研究報告, vol. 115, no. 146, SP2015–38, pp. 13–18, 2015.

[9] 柏木陽佑, 齋藤大輔, 峯松信明, ニューラルネットワークを用いた識別的アプローチによる分布間距離推定の検討, 日本音響学会秋季講演論文集, 2015.

[10] 渡辺美知子, 柏木陽佑, 先行節の種類や後続節の長さが文境界・節境界のポーズ長・フィラー長に及ぼす影響, 日本音響学会春季講演論文集, 2014.

[11] 柏木陽佑, 齋藤大輔, 峯松信明, 広瀬啓吉, 制約付き話者コードの同時推定によるニューラルネット音響モデルの話者正規化学習, 日本音響学会春季講演論文集, 2014.

[12] 柏木陽佑, 齋藤大輔, 峯松信明, 広瀬啓吉, 話者コードに基づく話者正規化学習を利用したニューラルネット音響モデルの適応, 情報処理学会研究報告, no. 20, pp. 1–6, 2014.

[13] 橋本哲哉, 柏木陽佑, 齋藤大輔, 広瀬啓吉, 峯松信明, 話者依存サブネットワークを用いた深層学習による多対一声質変換, 日本音響学会春季講演論文集, 2014.

[14] バン・フクアンフイ, 齋藤大輔, 柏木陽佑, 峯松信明, 広瀬 啓吉, Noisy Channel Model に基づく音声特徴量強調に関する検討, 日本音響学会春季講演論文集, 2014.

[15] 杉田祐樹, 柏木陽佑, 齋藤大輔, 峯松信明, 広瀬啓吉, Deep Neural Network に基づく音素事後確率を用いた発音評価, 日本音響学会春季講演論文集, 2014.

[16] Y. Luan, D. Saito, Y. Kashiwagi, N. Minematsu, K. Hirose, Performance improvement of exemplar-based noise robust ASR, Autumn Meeting of Acoustic Society of Japan, 2013.

[17] 尾崎洋輔, 柏木陽佑, 齋藤大輔, 峯松信明, 広瀬啓吉, 音声雑音に頑健な主話者区間検出に関する検討, 日本音響学会秋季講演論文集, 2013.

[18] 柏木陽佑, 齋藤大輔, 峯松信明, 広瀬啓吉, Deep Learning に基づくクリーン音声状態識別による雑音環境下音声認識, 日本音響学会秋季講演論文集, 2013.

[19] 岡安貴大, 池島純, 柏木陽佑, 鈴木雅之, 峯松信明, 広瀬啓吉, 実環境下における GMM を用いた統計的声質変換の検討, 日本音響学会春季講演論文集, 2013.

[20] 柏木陽佑, 久保陽太郎, 中村篤, 峯松信明, 広瀬啓吉, Deep Neural Network 混合モデルを用いた環境・話者適応の検討, 日本音響学会春季講演論文集, 2013.

[21] 柏木陽佑, 鈴木雅之, 峯松信明, 広瀬啓吉, 区分的線形変換を用いた雑音環境下マルチモーダル音声認識, 日本音響学会秋季講演論文集, 2012.

[22] 柏木陽佑, 鈴木雅之, 峯松信明, 広瀬啓吉, SPLICE に基づく音声・口唇画像情報を用いた雑音環境下音声認識, 電子情報通信学会技術研究報告, 2012.