修士論文

音声の構造的表象を用いた発音分析の高精度化

2010 年 2 月 9 日

指導教員　峯松 信明 准教授

電気系工学専攻融合情報学コース

37-086488　鈴木 雅之

内容梗概

近年コンピュータを用いた語学学習（Computer Aided/Assisted Language Learning; CALL）システムが広く用いられるようになった．現在広く用いられている CALL システムは，主に文法能力と聞き取り能力を訓練するシステムである．これらの他にも，学習者の発音を評価・分析して発音学習を行うシステムがあり，コミュニケーションに主眼をおいた語学教育が多く行われるようになった現在，そのニーズは非常に大きいといえる．しかし，このような発音評価・分析システムの構築には大きな問題がある．音声は，様々な声道形状を持つ話者に，様々な環境下で発声され，様々な伝送経路を通り様々な音響機器により収録される．これらのプロセスの中で，音声の物理的実体は，たとえ同じ発音情報を持っていたとしても，非言語的特徴によって変形されてしまう．そのため，あるネイティブ話者の発音と学習者の発音を比較してどこに違いがあるか分析しようとしても，前述した非言語的特徴によるミスマッチにより，発音の間違いのみを正しく見つけることは難しい．この非言語的特徴の違いによる「ミスマッチ問題」が解決され，十分に高い精度をもつ発音分析・評価システムが構築できれば，非常に多くのユーザに有効活用されると予想される．

従来の発音評価技術は，多くの話者の多様な環境による音声データを集め，統計的手法を駆使することによってあらかじめ音響モデルを作成し，それを生徒専用に話者適応させることでミスマッチ問題を解決しようとしてきた．しかし，音響モデルを生徒専用に適応させるときに，生徒の発音誤りにすら適応がおこってしまうという過適応の問題が発生してしまうという問題をかかえている．この問題は，非言語的特徴の違いと，発音の情報の違いをまったく区別していないために起こる問題である．

近年，上記の非言語的特徴に理論的には不変性をもつ，発音の違いのみに影響を受ける音声の構造的表象が提案された．これは，音声の物理的・絶対的実体を捨て，それらの相対関係のみをとらえることによって得られるものである．音声の構造的表象を用いることで，非言語的特徴に頑健な音声アプリケーションを実現することが可能となる．既に，発音分析・評価，そして孤立単語音声認識への応用が研究され，その非言語的特徴に対する頑健性が実験的に示されている．

しかしながら，音声の構造的表象を用いた音声分析にはいくつかの欠点が残されている．1 つは精度の問題である．音声の構造的表象を用いて音声分析を行うと，非言語的特徴のミスマッチには非常に頑健なシステムを構築することができるが，非言語的特徴のミスマッチが比較的小さい場合には，既存の手法と比較して精度が十分でないという問題がある．また，音声の構造的表象を用いた音声認識において，母音に関しては非常に高い精度が得られるものの，子音を含む場合には精度が低下することが報告されている．もう 1 つは，

柔軟性の問題である．音声の構造的表象を抽出するためには，それが音声の相対関係を利用している都合上，複数の文や単語の読み上げが必要になる．しかし現実的な問題として，多くの文や単語を読み上げるのは労力が必要であるため，より少ない文から発音分析を行えた方が都合がよい．

本論文では，音声の構造的表象を用いた母音・子音を含むすべての発音の分析・評価を主な対象として，その精度を大幅に向上させる手法を提案する．具体的には，構造の辺（エッジ）の長さの分散が考慮されていなかった問題を解決するエッジ長正規化と，エッジの数が多すぎることによる次元の呪いの問題を解決する部分構造化と二段階重回帰分析，さらに特徴量を増やすことで精度を改善させるマルチストリーム構造化と三段階重回帰分析を提案する．提案手法を用いて日本人大学生が読み上げた米語音声の発音評価実験を行ったところ，ミスマッチが大きくない条件下においても，提案手法と従来の音声の物理的・絶対的実体を用いた発音評価法は同程度の精度が実現できることが示された．

さらに，音声の構造的表象を用いた発音分析法と，従来の音声の物理的・絶対的実体を用いた発音分析法は，互いに対立しあう手法ではないという観点から，それらを組み合わせることによるさらなる精度の向上も試みた．実験の結果，両者を組み合わせを用いることで，さらなる精度の向上を実現することができた．

さらに，少ない読み上げ文章から音声の相対関係を利用した音声分析を行うことを目標として，局所的なアフィン変換不変量（Localized Affine Invariant Feature; LAIF）を提案する．LAIF を用いることで，従来の音声の物理的・絶対的実体を用いた手法と同様の枠組みを用いながら，相対関係を考慮した音声分析を行うことが可能になる．今回は LAIF を発音分析に用いる初期検討として，孤立単語音声認識実験において LAIF の有効性を検討した．実験の結果，特にミスマッチがある場合に，LAIF を用いることで音声認識率を向上させられることがわかった．

目次

図目次

表目次

第1章

序論

1.1　研究の背景

現代社会は，国際化社会である．図 1.1 に，日本への出入国者数の推移を示しているが，年間の出入国者の総計はこの 20 年およそ線形に増加し続けていることがわかる [1]．また，図 1.2 に，海外在留邦人数の推移を示しているが，海外在留邦人数の年による推移もこの 20 年およそ線形に増加し続けていることがわかる [2]．既に現在，一年間に日本の総人口の半数程度が日本の国境を行き来し，また日本人のおよそ 100 人に 1 人が日本国外で生活している．世界は高度に国際化されており，今後国際化はさらに進んでいくものと予想される．

国際化の進展に伴ない，語学教育の重要性が盛んに叫ばれるようになった．そのため，語学教育をとりまく環境は近年大きく変化している．例えば，2011 年度から，毎年約 120 万人の子供たちが新たに語学学習を受ける予定になっている．これは，文部科学省が 2002 年度に定めた「英語が使える日本人」育成に向けた戦略構想において小学校の英会話支援が構想され，2006 年度に中央教育審議会外国語専門部会により 2011 年度からの外国語活動必修化が決定されたためである [3]．また，小学生のみならず，語学教室に通う社会人の数も増えている．これは，ビジネスにおいて英語を使用する機会が増加したためであり，現在の語学教室市場の規模は 3,000 億円を超えるまでなった [4]．このように，日本人が語学教育を受ける機会は着実に増加している．このような語学教育を受ける機会の増加に加え，語学教育の方法も変化している．例えば，文部科学省が新しく定めた小学校での外国語活動の指導要領によると，話し言葉としての外国語が重要視され，外国語をたくさん聞く・話すことが活動の目標となっている [5]．これは，従来の中学校以上での英語教育において，読み書きが重要視されていたことと大きく異なる．また，ほとんどの社会人向けの語学教室でも，コミュニケーションを重視した教育が行われている．これは，ビジネス英語において，英語を読み書きする能力よりも，実際に外国人と英語を使って会話を行い，意思伝達を行う能力が必要とされるためである．このように，近年の語学教育の多くでは，聞く・話すことを中心としたコミュニケーション能力の向上に主眼をおいた教育がなされている．

語学教育の機会が増え，しかもその教育がコミュニケーションに重点をおいていることから，現在，十分な教育能力を持つ語学教師数の不足が問題になっている．例えば，2011 年度からの小学校における外国語教育では，優秀な外国人指導助手（Assistant Language Teacher; ALT）の不足から，語学を教えた経験がほとんどなく，かつ外国語によるコミュニケーションが必ずしも堪能ではないクラス担任たちが外国語を教える予定になっている [6]．また，社会人向けの語学教室においては，専任講師の数に対する非常勤講師の割合が年々高まってきており，講師の指導力低下が業界全体としての問題になっている．

そのため近年，語学教師の不足をコンピュータを用いて補う，Computer Assisted/Aided Language Learning（CALL）システムが非常に多く開発され，利用されるようになった．特に，ニンテンドー DS で動作する CALL システムは，非常に高い人気を集めている．なかでも，英語のリスニングを訓練する CALL システムである Plato 社の「えいご漬け」シリーズは，ニンテンドー DS ソフトウェア全体の中でも，非常に高い売り上げ本数を記録している [7]．ニンテンドー DS の他にも，近年ユーザ数が非常に増えているソーシャルネッ

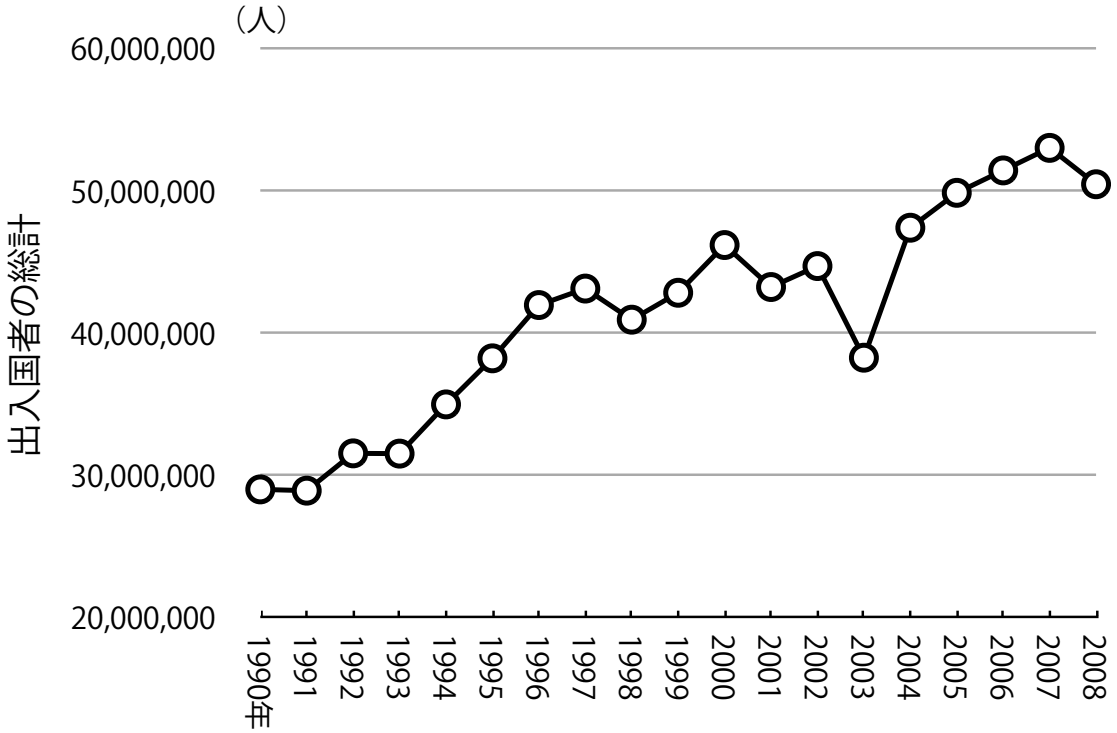

図 1.1: 日本の出入国者の推移

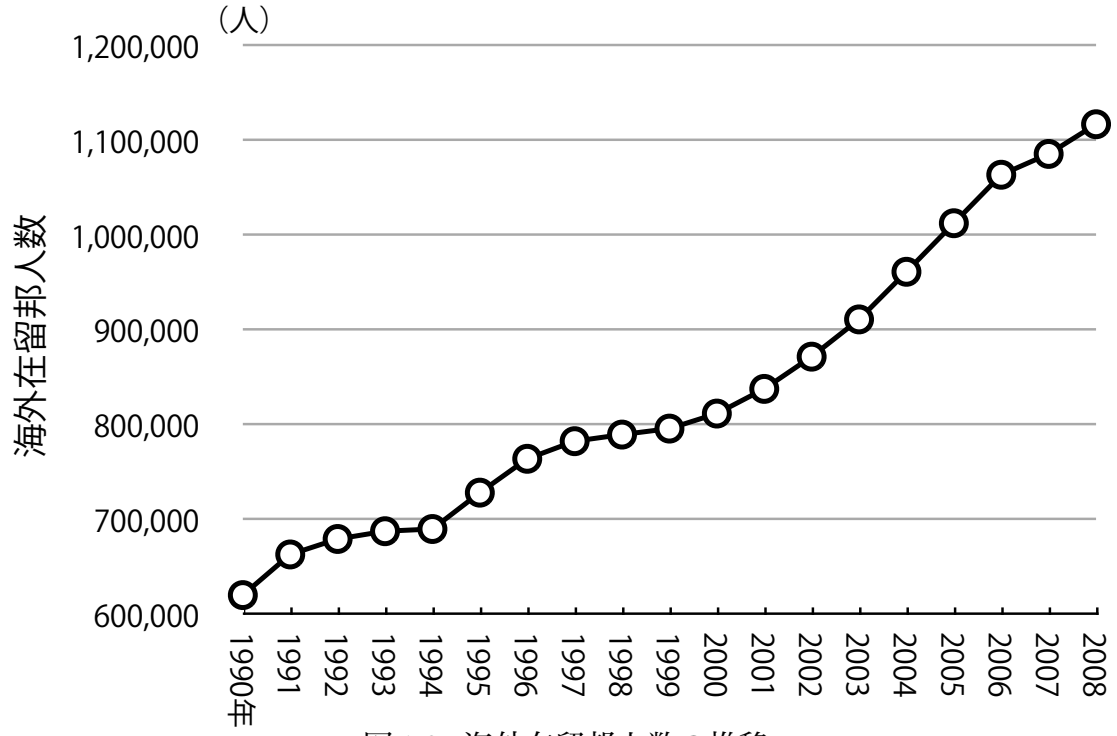

図 1.2: 海外在留邦人数の推移

トワーキングサービス（Social Networking Service; SNS）の中で動作する CALL システムが多くのユーザに利用されている．例えば，日本人の英語学習専用の SNS サイトである「smart.fm」には，50 万人以上が会員登録をし，英語学習を行っている [8]．他にも，数多くの CALL システムが実用化されている．CALL システムは，いまや広く一般に受け入れられた技術といえよう．

CALL システムの一つとして，学習者の発音を分析・評価する発音学習システムがある．発音学習は，相手に言いたいことを伝える能力を向上させること以外にも，聞き取り能力を向上させる効果があるなど，コミュニケーション能力の養成に非常に高い効果がある [9, 10]．そのため，発音学習を行う CALL システムは，コミュニケーション能力に主眼をおいた近年の語学教育の状況において非常に有効であると考えられる．

しかし，発音学習システムの構築には大きな問題がある．ある音素が正しく発声できているか否かと話者の声道形状の違いは，共に，音声のスペクトル包絡を変形させる．そのため，スペクトル包絡を表現する特徴量を直接利用して教師の発音と生徒の発音を比較すると，両者における声道形状のミスマッチに依存して，比較結果が変わってしまう．つまり，生徒や教師の性別・年齢などに依存して，評価の結果が変化してしまう問題がおこる．この「ミスマッチ問題」のため，発音学習は外国語によるコミュニケーション能力を訓練する学習法として非常に有効であるにも関わらず，発音学習を行う CALL システムはまだ広く一般に受け入れられているとは言い難い．

ミスマッチ問題を解決するため，従来様々な方法が提案されてきている．このような方法のうち，もっとも単純なのは，データを大量に集め，学習者と似た声質をもつ教師の発音を用意する方法である．このようなデータさえ集めることができれば，精度の高い発音分析が可能になる．例えば「ベネッセ小学生向け英語学習プログラム BE-GO」では，小学生の英語ネイティブ話者の発声を集めることで，学習者が小学生であった場合のみ，高い精度を実現している [11]．しかしながら，一般的にデータベースの構築には膨大な費用がかかり，かつ利用できるデータの量には限りがある．音声工学の分野では，よく “There is no data like more data.” と言われる．データをたくさん集めることによってミスマッチを解決する方法は，本質的な困難が付きまとう．

そこで，ミスマッチのあるデータからさまざまな手法を用いてミスマッチを取り除く処理を行うことで，ミスマッチのない音声比較を行う手法が広く研究されている．なかでも，ミスマッチのある音声から学習した音響モデルから，少量の入力話者の音声を利用して，入力話者とミスマッチのない音響モデルを作成する音響モデルの適応が現在標準的に利用されている．この技術は，もともと音声認識の精度向上のために提案された技術であり [12]，これが発音評価にも導入されている [13]．しかし，発音分析用音響モデルの適応には，音声認識の適応とは異なる独自の問題を抱えている．発音分析タスクでは，話者の声道形状の他，母語や発音習熟度によっても音声が変動するため，本来評価対象である発音習熟度に対し適応がかかってしまい，下手な発音を正しいと評価してしまう問題が発生する [14]．また，小学生の英語活動必修化などを受け，発音分析システムのユーザとして小さな子供が増えることが予想されるが，子供は特に声道形状に大きな個人差があるため，子供の音声の適応の精度は大人の音声の適応の精度より低いという問題がある [15]．

近年，発音分析におけるミスマッチ問題の解決手法として，解くべき課題の多い従来手法とまったく異なる，静的な変動に不変な音声の相対関係の利用が提案された [16]．これは，音声の静的な変動に不変な特徴量空間を用いることでをミスマッチを低減させる手法であり，音響モデルの適応とは同じ目的で異なる視点からのアプローチとなる．この音声の相対関係から得られる情報表象は，音声の構造的表象と呼ばれ，既に発音分析における有効性が示されている [17, 18, 19, 20, 21, 22]．音声の構造的表象を用いた手法の大きな利点としては，この表象は話者の違いに近似的に不変であるため，学習者の発音構造と教師の発音構造を一対一で比較できることである [23]．例えば，学習者は発音の目標としたい教師を選択し，その発音に近づくためにはどの音素を優先的に直していくべきかを知ることができる [20]．一口に発音学習と言っても，アメリカ英語の発音を目指す場合，イギリス英語の発音を目指す場合，さらには「多少日本なまりがあってもちゃんと通じる英語」を目指す場合では，学習方法は異なるはずである．構造を用いることで，従来手法ではデータ数の問題で実現することが困難であった，学習者にあわせた柔軟な学習目標を設定することができる．

しかしながら，構造を用いた発音分析にも，少なからず問題点がある．まず，ミスマッチがもともとないデータを用いた従来手法との比較実験があまり報告されておらず，実際比較実験を行うと精度の面で大きく劣っている．また，構造を用いた発音分析で主に扱われているのは母音であり，子音に関する詳細な分析は行われていない．また，音声から構造的表象を抽出するためには，音素バランスのとれた単語もしくは文章を，ある程度の量読み上げなければ発音評価ができない．具体的には，母音のみの構造を抽出するためには孤立単語 11 発声など，子音を含む構造を抽出するためには読み上げ文章 60 文程度などが必要になる．

1.2　本研究の目的

本論文の目的は，音声の構造的表象に基づく，母音と子音をすべて含む発音分析の精度を向上させることである．さらに，構造を用いた発音分析の適用範囲を広げるため，少ない読み上げ音声から構造に似た特徴を抽出できる手法の初期検討も行う．

1.3　本論文の構成

本論文は，全 8 章から構成される．第 1 章（本章）では，本論文の背景，目的について述べた．第 2 章では，現在広く用いられている外国語発音評価法の枠組みを概説する．その中でも，特に音響特徴量抽出と音響モデルについて詳しく説明を行う．第 3 章では，音声の物理的実体を変形する非言語的特徴について述べ，非言語的特徴によるミスマッチを解決するための従来手法について述べる．さらに，発音評価用音響モデルを適応する際に，どのような問題がおき，その問題に対しこれまでどのような対処がなされてきたのかについても述べる．そして第 4 章で，本研究で用いる音声の構造的表象について述べる．音声の構造的表象を用いた音声分析に用いられる基礎技術の他，具体的な応用として行われて

いる音声認識，そして本研究で扱う発音分析についても述べる．

第 5 章から，本論文の提案手法を述べていく．まず第 5 章では，音声の構造的表象を用いた発音分析の精度を向上させるさまざまな手法を提案し，その米語発音評価実験においてその有効性を検証する．具体的には，エッジ長正規化，部分構造化，二段階重回帰分析，マルチストリーム構造化と三段階重回帰分析を提案する．これらの手法を用いることにより，従来広く用いられている発音評価手法と同程度の精度が実現できることを示す．第 6 章では，従来手法と音声の構造的表象を用いた手法を組み合わせることでより高い精度を持つ発音分析手法を提案する．また，提案手法の非言語的特徴に対する頑健性を実験的に検証するため，多様な話者性を含む音声データを用意して実験を行い，詳細な比較実験を行う．第 7 章では，連続発声に対する音声の構造的表象を用いた分析の初期検討として，構造を短時間特徴量として利用する手法を提案し，音声認識実験においてその有効性を検証する．最後に，第 8 章で本論文をまとめ，今後の課題と展望について述べる．

第2章

従来の発音評価法

2.1　はじめに

本章では，現在最も広く用いられている発音評価法として，短時間音響特徴量を出力する音響モデルの事後確率を用いた，Goodness Of Pronunciation（GOP）スコアとその応用を紹介する [24]．はじめに，GOP スコアを算出するための要素技術として，短時間音響特徴量として利用されるメルケプストラム（Mel Frequency Cepstrum Coefficients; MFCC）と，音響モデルとして利用される隠れマルコフモデル（Hidden Markov Model; HMM）ついて詳細な説明を行う．その後，具体的に GOP スコアを算出する方法について説明し，GOP スコアを用いた発音評価法について述べる．

2.2　短時間音響特徴量時系列の抽出

音声信号のうち，言語的特徴の多くは，音声信号のスペクトル包絡に含まれている．そのため，音声信号を周波数変換しスペクトル包絡を抽出することで，発音評価に用いる特徴量を得ることができる．しかし，一般的に周波数変換は信号が時間的に変動しないことを前提にしてスペクトルを算出するが，実際の音声は時間的に変化する信号である．そのため，音声信号処理の特徴量抽出においては，ある程度の幅を持つ窓関数を，少しづつずらしながら音声信号にかけ，それらの出力を周波数変換しスペクトル包絡を算出する短時間フーリエ変換を利用し，短時間音響特徴量時系列を特徴量として用いる．

2.2.1　窓関数

窓関数とは，ある有限区間以外で 0 となる関数である．窓関数を単に窓ともいい，データに窓関数を掛け合わせることを窓を掛けるという．音声の音響的特徴がおよそ定常とみなせる時間長の窓を時間をずらしながら掛けて周波数分析することで，音響特徴量の時系列が得られる．発音評価のための音声分析としては，窓の幅を 25msec 程度，窓のずらし幅を 10msec 程度にすることが多い．

窓関数の具体例としては，方形窓，ハミング窓，ブラックマン窓などがある．t を時間として，$0 \leq t \leq 1$ の区間に窓をかけるとすると，方形窓 W_R，ハミング窓 W_H，ブラックマン窓 W_B はそれぞれ以下のような関数となる．

$$W_R(t) = 1 \tag{2.1}$$

$$W_H(t) = 0.54 - 0.46 \cos 2\pi t \tag{2.2}$$

$$W_B(t) = 0.42 - 0.5 \cos 2\pi t + 0.08 \cos 4\pi t \tag{2.3}$$

ただし，$t \leq 0, 1 \leq t$ の区間では $W_R(t) = W_H(t) = W_B(t) = 0$ である．それぞれの窓は，周波数分解能やダイナミックレンジが異なっており，分析によって最適な窓は異なる．本研究では，ハミング窓を利用している．

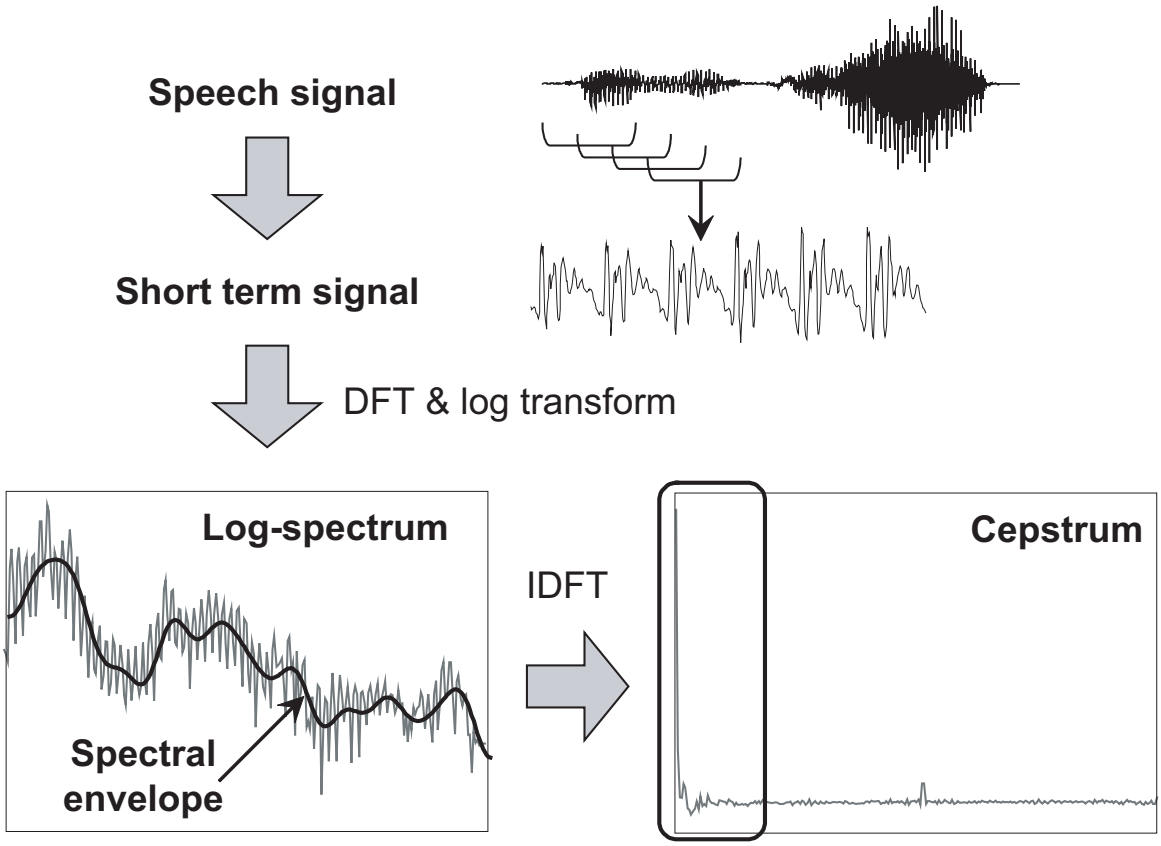

図 2.1: 音声信号からケプストラムの抽出

2.2.2　ケプストラム

音の大きさに対するヒトの感覚は，パワーに対する対数軸におよそ比例しているため，発音評価の音声分析では，音声信号の対数パワースペクトル包絡を短時間特徴量として用いることが有効である．この対数パワースペクトル包絡を効率的に低次元の特徴量で表現する方法として，現在最も広く用いられているのがケプストラムである．

音声波形に窓をかけ，そこからケプストラムを抽出するまでの様子を図 2.1 に示す．まず音声波形から，窓掛けにより数十ミリ秒程度のフレームを切り出し，その区間に対して離散フーリエ変換（Discrete Fourier Transform; DFT）を施し，その対数パワー成分を抽出する．ここで，特徴量として利用したいのは，対数パワースペクトルの包絡成分である．そこで，いったん対数パワースペクトルに対して逆離散フーリエ変換（Inverse DFT; IDFT）を施す．これがケプストラムと呼ばれる特徴量である．このケプストラムのうちの低次項数十次元のみを残し，高次項を 0 にして DFT してスペクトル領域に変換することにより，スペクトル領域における低周波成分，すなわちスペクトル包絡が得られる．そのため，ケプストラムの低次元はスペクトル包絡の情報のみを小さな次元数で表現した特徴量といえる．そこで，このケプストラムの低次元 10〜20 次元分を，短時間音響特徴量として利用する．このケプストラムの低次元のみをぬきだす操作は，リフタリングと呼ばれている．な

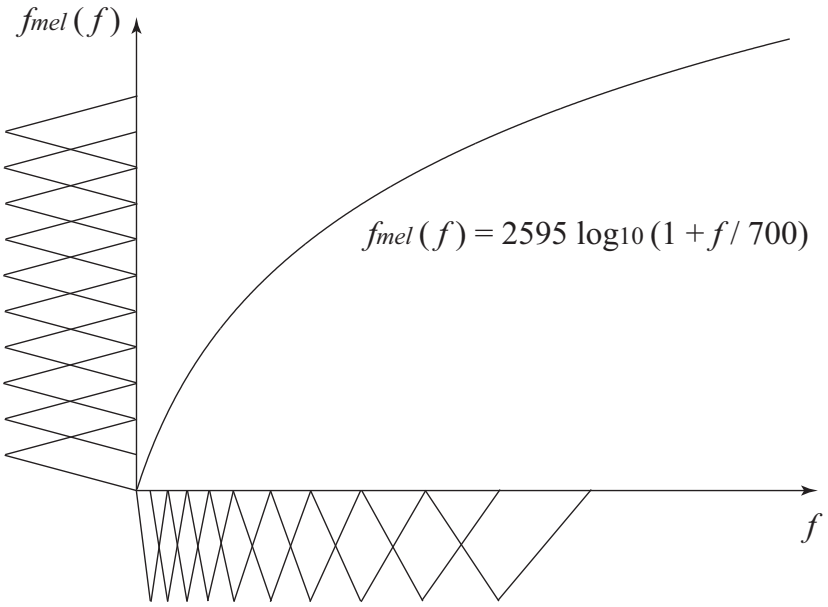

図 2.2: メル周波数軸上に等間隔で配置された三角窓

お，ケプストラムの 0 次元は，対数スペクトル領域でいうオフセット成分に対応しており，これは音声のパワー成分に相当する．パワーはマイクと口との距離などでも変動し，全体的にパワーが変化しても音素は基本的に変化しないため，ケプストラムの 0 次元は音響特徴量から排除されることが多い．ただし，パワーの時間的な変動の様子は情報として有用であるため，ケプストラム 0 次元の変動成分は音響特徴量として用いられることがある．

2.2.3　ヒトの聴覚特性を考慮したケプストラム

音の高さに対するヒトの知覚特性は，低域ほど分解能が高く，広域ほど分解能が低い．具体的には，周波数分解能は周波数に対する対数関数で近似できる．そこで，ヒトの知覚特性に合わせて周波数分解能を変化させて音声分析を行うことで，よりヒトの感覚にあった特徴量が抽出できる．

ヒトの知覚特性を反映した尺度であるメル尺度を利用したケプストラムは，数多く提案されている．現在の短時間音響特徴量のデファクトスタンダードとなっている MFCC（Mel-Frequency Cepstrum Coefficient）も，その一つである．まず，図 2.2 に示すようにメル周波数（メル尺度化された周波数）軸上に等間隔で配置された三角窓を用意し，フィルタバンク分析を行なう．なお，メル周波数 f_{mel} は周波数 f [Hz] に対して，

$$f_{mel}(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \tag{2.4}$$

という周波数ウォーピングを施すことで得られる．このメルフィルタバンクの出力を求めることにより，メルスペクトルが得られる．これに離散コサイン変換を施し，リフタリン

グを行った物が MFCC である [25]. なお，発音評価においては，メルフィルタバンクの数は 24 程度，MFCC の次元数は 12 程度とされることが多い.

2.2.4　ケプストラムの動的特徴量

ケプストラム係数は，数 10msec 程度の音声区間（フレーム）を定常とみなした上で得られる静的な特徴であるが，音素の音響的な特徴は周辺の音素に影響を受けて変化する調音結合がおこり，スペクトルは時間とともに連続的に変化している．そこで，フレーム分析によって得られる静的な特徴に加え，時間とともに変化する方向成分を動的特徴量として加えることで精度が大きく向上することが知られている [26].

動的特徴量として最もよく用いられるのは，ケプストラムをある時間幅において重み付き最小二乗法で直線近似した傾きとして定義される，Δ ケプストラムである．フレーム番号 n における Δ ケプストラム $\Delta c(n)$ は，その前後 T フレームのケプストラムと，各フレームに対する重み係数 w_t により以下のように算出される.

$$\Delta c(n) = \frac{\sum_{t=-T}^{T} t w_t c(n+t)}{\sum_{t=-T}^{T} t^2 w_t} \tag{2.5}$$

T としては 2 を利用することが多い.

さらに，Δ ケプストラムの動的特徴量である ΔΔ ケプストラムも用いられることがある.

2.3　音響モデル

ヒトの音声活動を，パラメータ θ で制御され，音素列 P から音響特徴量時系列 $\boldsymbol{O}$ を出力するモデルとして考える．このようなモデルは，音響特徴量を $\boldsymbol{O}$，音素などの音響モデルの単位を P として，$p(\boldsymbol{O}, P|\theta)$ を最大化する θ を学習することで得られる生成モデルや，$p(P, \theta|\boldsymbol{O})$ を最大化する θ を学習することで得られる識別モデルに分類される.

ヒトの音声活動のモデルのうち，音素列 P を話そうとした上で，どのような音響特徴量時系列 $\boldsymbol{O}$ が出力されるかを表現するモデルを，音響モデルと呼ぶ．発音分析は，音素列 P が既知のタスクであるため，この音響モデルが特に重要になる.

発音分析に用いる音響モデルには，隠れマルコフモデル（Hidden Markov Model; HMM）を生成モデルとして利用することが多い [27]．以下，HMM とはどういったモデルであるかと，P の手動書き起こし付きの $\boldsymbol{O}$ が得られたときに $p(\boldsymbol{O}, P|\theta)$ を最大化する HMM のモデルパラメータ θ を学習する方法と，θ を学習した後，短時間音響特徴量時系列 $\boldsymbol{O}$ が得られたときにそれが P から生成されたとする事後確率 $p(P|\boldsymbol{O}, \theta)$ を計算する方法について述べる.

2.3.1　隠れマルコフモデル (HMM)

HMM の概念図を図 2.3 に示す．S_i は i 番目の状態，a_i は状態 S_i から状態 S_{i+1} への状態遷移確率，$b_i(\boldsymbol{o})$ は状態 S_i から短時間音響特徴量 $\boldsymbol{o}$ が出力される出力確率である．出力確

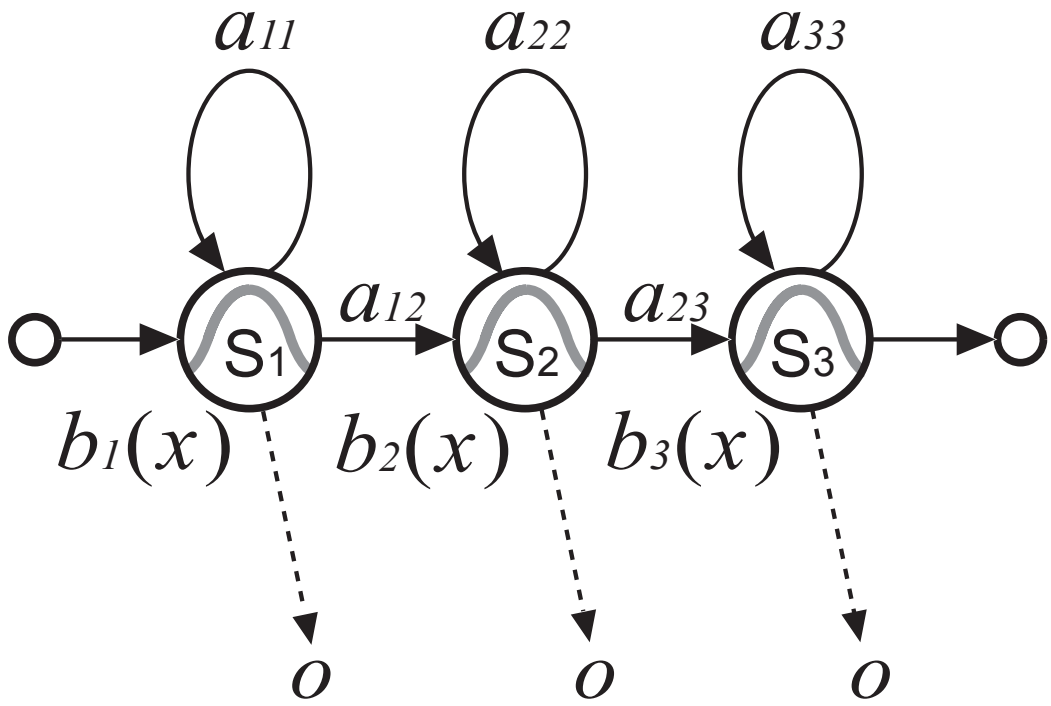

図 2.3: 隠れマルコフモデル（HMM）

率 $b_i(\boldsymbol{o})$ の分布形としては，ガウス分布を複数用意し，その重み付け和で $b_i(\boldsymbol{o})$ を表現する混合ガウス分布（Gaussian Mixture Model; GMM）がよく用いられる．HMM のパラメータを θ とすると，HMM を用いることで，$p(\boldsymbol{O}|\theta)$ がモデルされることになる．

多くの場合，一つの音素につき一つの HMM を用いる．これは音素 HMM と呼ばれ，HMM のトポロジーとしては図 2.3 で示しているような 3 状態程度の left-to-right 型が多く利用される．音素 HMM により，音素系列を P として，$p(\boldsymbol{O}|P,\theta)$ がモデル化される．

音素 HMM のうち，前後の音素環境を考慮しない音素 HMM を monophone，前後の音素環境を考慮する音素 HMM を triphone と呼ぶ．実際の音声は，調音結合と呼ばれる現象により，音素の音響的特徴が前後の音素環境に依存して大きく変化する．そのため triphone を用いることで，音響モデリングを精緻にできる利点がある．ただし，triphone には HMM の数が膨大に増える欠点も存在する．そのため，音声認識タスクにおいては triphone が主に用いられているが，発音評価タスクでは monophone を用いることが多い．

2.3.2　HMM の学習

音素系列 P の書き起こし付きの $\boldsymbol{O}$ が得られたときに，HMM の学習すべきパラメータは $\theta = \{a^p_{ij}, b^p_i(\boldsymbol{o})\}$ であり，これを生成モデルとして最尤（Maximum Likelihood; ML）推定

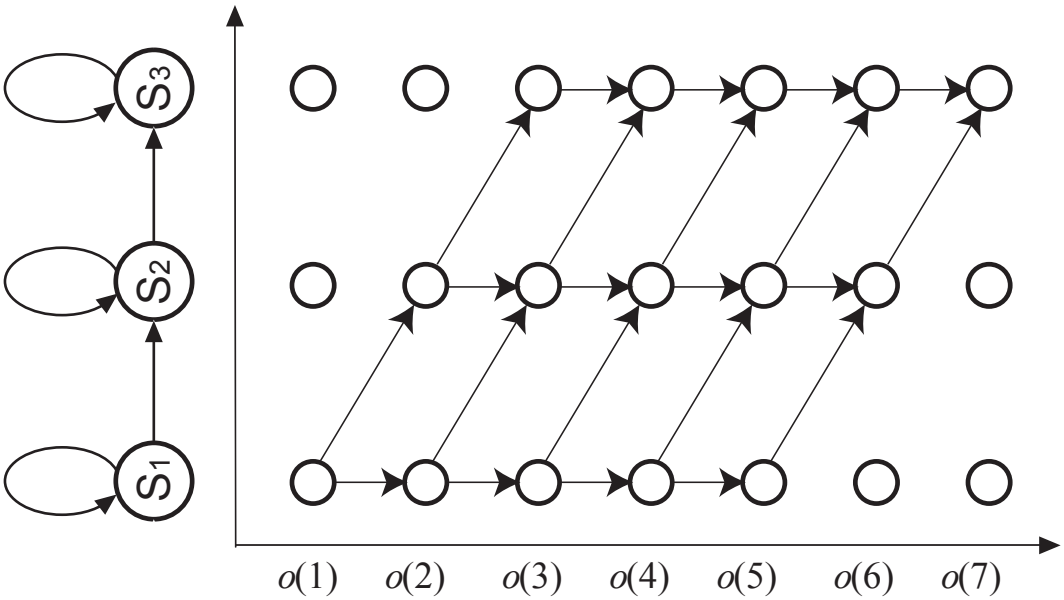

図 2.4: HMM の状態遷移の経路

することを考える．これは

$$\begin{aligned} \underset{\theta}{\operatorname{argmax}}\, p(\boldsymbol{O}, P|\theta) &= \underset{\theta}{\operatorname{argmax}}\, \frac{p(\boldsymbol{O}|P,\theta)}{p(P|\theta)} \\ &= \underset{\theta}{\operatorname{argmax}}\, p(\boldsymbol{O}|P,\theta) \end{aligned} \tag{2.6}$$

を解くことで行われる．しかし，HMM の現在の状態 z は隠れ変数であり，(2.6) を解析的に解くのは不可能である．そのため，隠れ変数が存在する統計モデルの ML 推定値を一般的に見出すことができる Expectation-Maximization（EM）アルゴリズムを用いて (2.6) の局所最適解を得る．ここで EM アルゴリズムの収束結果は初期値に依存するため，パラメータの初期値設定法は重要である．初期値設定法としては，全音素を同一として HMM を学習し，そのパラメータをすべての HMM の初期値として利用するフラットスタートや，triphone の初期値に monophone のパラメータを用いる手法などが利用される．なお，EM アルゴリズムの実装には，HMM の EM 学習に特化して効率を高めたアルゴリズムである Baum-Welch アルゴリズムを利用することができる．

2.3.3　HMM を用いた音素認識

すべての音素 HMM が学習され，そのパラメータを θ とする．その上で，短時間音響特徴量の時系列 $\boldsymbol{O}$ が観測されたときに，それが音素系列 P の音素 HMM から生成された事後確率 $p(P|\boldsymbol{O},\theta)$ を計算することを考える．これを計算し，尤度の最も高い音素 P を出力

することは，音声認識をすることにほかならない．ベイズの定理より

$$\begin{aligned}\underset{P}{\operatorname{argmax}}\, p(P|\boldsymbol{O},\theta) &= \underset{P}{\operatorname{argmax}}\, \frac{p(P,\boldsymbol{O}|\theta)}{p(\boldsymbol{O}|\theta)} && (2.7)\\ &= \underset{P}{\operatorname{argmax}}\, p(P,\boldsymbol{O}|\theta) && (2.8)\\ &= \underset{P}{\operatorname{argmax}}\, p(\boldsymbol{O}|P,\theta)p(P|\theta) && (2.9)\end{aligned}$$

となる．ここで $p(P|\theta)$ は，どのような音素が出現しやすいかを表すものであり，言語的な制約条件によってモデル化されるべきもので，音響モデルではモデル化されていない．音声認識問題ではなく，音素認識問題を解く場合は，$p(P|\theta) = const.$ として，以下の最大化問題を解けばよい．

$$\underset{P}{\operatorname{argmax}}\, p(\boldsymbol{O}|P,\theta) \tag{2.10}$$

$p(\boldsymbol{O}|P,\theta)$ は HMM でモデル化されているので，これで音素認識問題を解くことができる．

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

2.4　GOP スコア

発音評定値，すなわち生徒の発声から音響特徴量時系列 $\boldsymbol{O}$ が観測され，それが本来発声されるべき音素系列 P としてどれだけ良く発声されたかを示すスコアとして Goodness of Pronunciation（GOP）スコアが広く用いられている [24]．GOP の計算に用いる HMM のパラメータを θ として，GOP スコアは以下の式で定義される．

$$\mathrm{GOP}(\boldsymbol{O},P,\theta) = -\log p(P|\boldsymbol{O},\theta) \tag{2.11}$$

すなわち，音響特徴量時系列 $\boldsymbol{O}$ が観測されたときに，それが P として正しく発声された事後確率を，GOP スコアと定義する．GOP スコアは，以下のように書き下せる．

$$\begin{aligned}-\log p(P|\boldsymbol{O},\theta) &= -\log \frac{p(\boldsymbol{O}|P,\theta)p(P|\theta)}{p(\boldsymbol{O}|\theta)} && (2.12)\\ &= -\sum_{i=1}^{N} \log \frac{p(\boldsymbol{o}^{p_i}|p_i,\theta)}{\sum_{q\in Q} p(\boldsymbol{o}^{p_i}|q,\theta)} && (2.13)\end{aligned}$$

ここで，N は音素数，p_i は各音素，Q は考慮している全音素である．また，$\boldsymbol{o}^{p_i}$ は音素 p_i の HMM から出力される特徴量時系列である．なお，(2.12) の計算には，発声された音素が既知であるため $p(P) = 1$ が成立することを用いている．(2.13) を見ると，分子のスコアが分母のスコアで正規化される形になっている．すなわち，HMM の学習データと学習者の音声で多少のミスマッチがあったとしても，およそキャンセルできるスコア手法となっている．

2.4.1　GOP スコアの計算

GOP スコアを計算するためには (2.13) を計算する必要があるが，これを直接計算するのは難しい．そこで，以下のような近似を導入する．

$$-\sum_{i=1}^{N} \log \frac{p(\boldsymbol{o}^{p_i}|p_i,\theta)}{\sum_{q\in Q} p(\boldsymbol{o}^{p_i}|q,\theta)} \quad \approx \quad -\sum_{i=1}^{N} \log \frac{p(\hat{\boldsymbol{o}}^{p_i}|p_i,\theta)}{\max_{q\in Q} p\left(\hat{\boldsymbol{o}}^{p_i}|q,\theta\right)} \tag{2.14}$$

ただし，$\hat{\boldsymbol{o}}^{p_i}$ は，$\boldsymbol{O}$ を P に対して強制アライメントしたときに音素 p_i に対応する部分の $\boldsymbol{O}$ である．分子は，$\hat{\boldsymbol{o}}^{p_i}$ を計算するために強制アライメントを計算したときに算出される $\hat{\boldsymbol{o}}^{p_i}$ の事後確率である．これは (2.13) の分子に対するビタビ近似であり，簡単に計算することができる．分母は，連続音素認識を行った場合の，分子の計算で得られる $\boldsymbol{o}^{p_i}$ に対応する部分の事後確率である．これは (2.13) の分母に対するビタビ近似であり，簡単に計算することができる．以上により，GOP スコアを近似的に計算することが可能になる．

2.4.2　GOP スコアを用いた発音評価

発音評価タスクは，生徒の発声から，その生徒全体を通してのスコアを算出するタスクと，どの音素の発声が誤っているのかを診断するタスクの二つに大きく分類できる．GOP スコアは，各強制アライメント区間ごと（音素ごと）にスコアを算出することが可能であり，これら二つのタスクどちらにも利用することができる．

生徒のスコアを算出する場合には，生徒が発声した文章から (2.14) を計算すればよい．また，音素ごとに発音誤りを見つける場合には，各強制アライメント区間ごとに GOP スコアを算出し，しきい値を設けて GOP スコアがそれより大きい場合に誤り，小さい場合に正解とすればよい [28].

2.5　まとめ

本章では，従来の発音評価の枠組みを説明するために，その構成要素である音響特徴量抽出と音響モデルに関して説明した．音響特徴量として MFCC，音響モデルとして HMM を利用することは，これまで長く研究されてきており，デファクトスタンダードとなっている．そして，これらを用いた発音評価スコアとして GOP スコアを紹介し，その算出方法を示した．

第3章

音声に含まれる非言語的特徴

3.1　はじめに

従来の発音評価手法の問題点として，音声に混入する非言語的特徴に起因するミスマッチ問題がある．音声には言語的な情報以外にも非言語的な情報が含まれており，この非言語的特徴により音声の物理的実体（MFCC などで表現されるスペクトル包絡）は様々に変化する．一方で発音評価は，音声が言語的な情報のみを，正しく音として表現できているかを評価するタスクである．そのため，学習データに用いた音声と，評価対象である入力音声との間で非言語的特徴にミスマッチが存在する場合，発音評価の性能が劣化してしまう．例えば GOP スコアを計算する場合には，強制アライメント結果が本来あるべきアライメントからずれるなどといった問題が発生する．

ミスマッチ問題は，発音評価に本質的に発生する不可避な問題である．音声は必ずある特定の話者によって発声されねばならないし，背景雑音の存在する外気を伝わらなければならないし，周波数特性を持つマイクロフォンを用いて収録しなければならない．つまり，非言語的特徴は音声に不可避的に混入する．ミスマッチ問題を起こさないためには，教師の発音と生徒の発音が，似た声質で，同じ録音条件で録音される必要があるが，このような音声を集めるのは現実的に不可能である．そのため，発音評価の精度を向上させるためには，ミスマッチ問題が発生するとした上で，それにどう対処するのかが重要になる．

本章では，まず音声のその物理的特徴を変化させる要因を 3 つに分けて整理した上で，ミスマッチ問題の原因となる非言語的特徴による音声の変動を数学的にモデル化する．次に，これらの非言語的特徴による変動に対する従来手法を紹介する．

3.2　音声を変化させる要因

音声の物理的実体を変動させる要因は，大きく以下の 3 つに分けることができる．

言語的特徴　発話内容，語彙など

パラ言語的特徴　発話スタイル，意図，感情など

非言語的特徴　話者の性別，年齢，体格，健康状態，音響機器の特性，背景雑音，残響など

言語的特徴は，文字言語としても表現可能である情報であり，発話内容・発話テキストや語彙の情報を表す．発音評価はこの情報を正しく音として生成できているかを評価するタスクである．パラ言語的特徴は，人間が音声を意図的に制御することによって伝達される情報であるが，文字では表現することはできない情報のことである．例えば, 発話スタイル（速く話す・ゆっくり話すなど），発話者の意図や感情[1]などである．発音評価タスクにおいても，例えばある意図を伝えるために日本語にはない発話スタイルが要求されることもあるため，パラ言語的特徴に着目した CALL システムも可能ではあるが，本論文では対象にしない．非言語的特徴は，話者の性別の違いや年齢の違いといった話者の身体性の違い

[1]ただし，感情は，発話者が制御できるものではないという考え方から，非言語的特徴に含まれることもある．

に起因する情報や，音響機器や周囲の環境など話者以外の要因によって変動する情報であり，一般に話者が制御不可能な情報である．

音声の物理的実体を変動させる 3 つの要因の中で，話者によって制御不可能なのは非言語的特徴のみである．この非言語的特徴の違いに起因するミスマッチこそが，発音評価の精度を劣化させる原因となる．

3.3　非言語的特徴のモデル化

本節では，非言語的特徴を，それが音声をどう変動させるのかといった観点から分類し，それぞれを数学的にモデル化する．音声分析に用いる短時間音響特徴量は，MFCC などの対数パワースペクトルを周波数変換した空間の特徴量なので，非言語的特徴がこの空間に対しどのような変動をもたらすかを見る．

3.3.1　話者性の違い

非言語的特徴の中でも，最も違いが大きくかつ本質的に不可避なものが，話者性の違いである．話者性の違いは，話者の性別や年齢などによる，声道形状の違いによって発生する．

話者の違いは，ケプストラム空間に対する時間変動しない可逆な変換で近似することができる．例えば，声道長の差異を近似する対数スペクトル領域における周波数ウォーピングは，ケプストラムに対する線形変換 $\boldsymbol{c}' = \boldsymbol{A}\boldsymbol{c} + \boldsymbol{b}$ で表されることが示されている [29]．また，音声から言語情報を保持したまま話者性を変換する技術である話者変換の研究では，話者の違いをケプストラム空間における可逆な非線形変換 $\boldsymbol{c}' = f(\boldsymbol{c})$ と仮定し，その変換関数 f を混合正規分布（Gausiann Mixture Model; GMM）を利用して学習することで精度のよい話者変換を実現させている [30]．また，後述する MLLR 適応では，話者の違いを線形変換と仮定した上で，話者の違いを表す線形変換のパラメータを推定することで，音響モデルの適応を実現している．以上のことから，数学的にも経験的にも，話者の違いはケプストラム空間に対する可逆な変換 $\boldsymbol{c}' = f(\boldsymbol{c})$ で近似できることがわかる．

3.3.2　音響デバイスの周波数特性の違い

話者性の違いと並んで音声に不可避的に含まれる非言語的特徴として，音響機器や伝送経路の周波数特性の違いがある．音声分析を行うためには，コンピュータに電気的信号に変換された音声信号を送らなければならないため，なんらかの音響デバイスを利用して音声を収録しなければならない．しかしながら，音響デバイスは周波数特性を持っており，この周波数特性が異なれば収録される音声は異なるものになる．

音響デバイスなどの周波数特性の違いは，スペクトルに対する時間変動しない乗算で表現されるため，ケプストラム空間では時間変動しない足し算で表現される．足し算は，当然可逆な変換であるので，話者性の違いと音響デバイスの違いは，両方含めてケプストラム空間に対する時間変動しない可逆な変換 $\boldsymbol{c}' = f(\boldsymbol{c})$ で近似されることになる．

3.3.3　背景雑音の違い

不可避的ではないが多くの場合に含まれてしまう非言語的特徴として，背景雑音の違いがある．背景雑音の違いに関しては，静かな部屋に移動したり，指向性の高いマイクを用いることなどで，ある程度ミスマッチを低減することができるが，現実的にはある程度の背景雑音は含まれてしまう．

背景雑音の違いは，スペクトルに対する足し算で表現されるため，ケプストラム空間では可逆な非線形変換になる．背景雑音が定常であった場合，この非線形変換は，話者の違いや音響デバイスの違いと同様，ケプストラム空間に対する時間変動しない可逆な変換で近似されることになる．しかし，背景雑音が非定常であった場合，これはケプストラム空間における時間変動する可逆な非線形変換で表現されることになる．

3.4　ミスマッチ問題に対する従来手法

非言語的変動に起因するミスマッチ問題に対処するために，これまで様々な手法が検討されている．最も単純な解決手法は，データをたくさん集めることである．何百，何千，何万もの話者による発声を用いて音響モデルを学習すると，不特定話者音響モデルを得ることができる．不特定話者音響モデルは，認識対象となる入力音声における話者性の違いに対処するために，様々な特性をもつ話者を集めてどのような入力に対しても広くカバーできるように意図されて構築される．しかし，不特定話者音響モデルを用いても，認識性能が高くならない話者がどうしても存在する．これは，不特定話者音響モデルでは話者性の違いを完全にはカバーしきれないことが原因であり，集めることによる対処法は根本的な解決にはならないといえる．

そこで，データを集めることに頼らないミスマッチ問題の解決のために，入力音声の正規化や音響モデルの適応技術が広く研究されてきた．入力音声の正規化とは，音声に含まれる変動をできるだけ取り除くように入力音声を標準話者の音声に変換する手法であり，一方，適応手法は入力音声と音響モデルとのミスマッチを低減させるようにモデルのパラメータを変更する手法である．正規化・適応手法はこれまで様々な手法が研究され，現在でもその改良が続けられている．本節では，これらのうち最も基本的な，ケプストラム平均正規化（Cepstrum Mean Normalization; CMN），スペクトルサブトラクション（Spectral Subtraction; SS），声道長正規化（Vocal Tract Length Normalization; VTLN），最尤線形回帰（Maximum Likelihood Linear Regression; MLLR）適応の 4 つの手法について説明する．

3.4.1　ケプストラム平均正規化

ケプストラム平均正規化（Cepstrum Mean Normalization; CMN）は，簡便でかつ効果的なケプストラムの足し算による変動に対する正規化手法である [31]．CMN では，認識対象である入力音声のケプストラム時系列 $\boldsymbol{c}_t(t = 0, \cdots, T)$ に対して，各フレームのケプス

トラムベクトルから発声全体のケプストラム系列の平均ベクトル

$$\hat{\boldsymbol{c}} = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{c}_t \tag{3.1}$$

を差し引くことで，入力の音響特徴量の正規化を行う．

これにより，音響デバイスの周波数特性の違いと，背景雑音や話者性の一部を消去することができる．

3.4.2　スペクトルサブトラクション

スペクトルサブトラクション（Spectral Subtraction; SS）は，簡便でかつ効果的な定常な背景雑音による変動を抑制する手法である [32]．SS では，あらかじめ音声の含まれない無音区間から，背景雑音のスペクトル成分を得ておき，それを，音声区間のスペクトルから引き算することにより定常は背景雑音による変動を抑制する．

3.4.3　声道長正規化

声道長正規化（Vocal Tract Length Normalization; VTLN）は，話者の声道長の違いに起因するミスマッチを周波数ウォーピングによって標準的な声道長に変換することでミスマッチを低減する手法である [33]．VTLN では，入力発話に対して，それを標準的な声道長に変換する周波数ウォーピングのパラメータを推定することで，入力発話の声道長正規化を行っている．

3.4.4　最尤線形回帰によるモデル適応

最尤線形回帰 (Maximum Likelihood Linear Regression; MLLR) による音響モデルの適応は，ケプストラム $\boldsymbol{c}$ に対する線形変換 $\boldsymbol{A}\boldsymbol{c}+\boldsymbol{b}$ で近似される非言語的特徴による変動を，$\boldsymbol{A}$ と $\boldsymbol{b}$ を推定し，音響モデルのパラメータを変換することによってミスマッチを低減する [34]．この手法は，扱いやすさと性能の良さから，音声認識における音響モデル適応手法として広く用いられている．

もっとも単純な MLLR は，HMM の各ガウス分布の平均ベクトル $\boldsymbol{\mu}$ にアフィン変換 $\boldsymbol{A}\boldsymbol{\mu}+\boldsymbol{b}$ を施すことで行なわれる．$\boldsymbol{A}$ と $\boldsymbol{b}$ は， ML 推定の枠組みで推定される．全ガウス分布で一つの $\boldsymbol{A},\boldsymbol{b}$ を用いる場合は，入力データ量が少ないときに良い性能を示すが，入力データ量が増加するとモデル過小問題が生じる．これは大局的なアフィン変換 $\boldsymbol{A}\boldsymbol{c}+\boldsymbol{b}$ のみでは話者性を表現する能力が限られているためである．この問題を解決するために，HMM のガウス分布をボトムアップクラスタリングなどを用い複数のグループに分け，各同一グループにつき一つの $\boldsymbol{A}$ と $\boldsymbol{b}$ を利用し，話者性を表現する非線形変換を区分的な線形変換の連続で近似することにより話者性を表現する能力を高める手法も広く利用されている．

しかしながら，複数のグループに分けてアフィン変換を行う MLLR 適応を発音評価タスクに用いるには，問題がある．発音評価タスクの評価データには，ある音素はうまく発声

されているが，ある音素は間違って発声されているようなデータが含まれる．そのため，このような MLLR 適応を用いると，非言語的特徴に対する適応だけでなく，間違った音素を正しく発声したかのように変換する適応まで行われてしまうことがある．[14] では，このような MLLR 適応ではなく，アフィン変換のパラメータをひとつだけ用いる MLLR を用いる場合に GOP スコア算出の精度が最も高くなることを示している．

3.5　まとめ

本章では，まず音声の物理的特性を変化させる要因を 3 つに分けて説明し，その中で人間が音声を用いてコミュニケーションする上で不可避的に混入する非言語的特徴がミスマッチ問題の原因となることを述べた．また，非言語的特徴のうち，話者の違い，音響デバイスの違いは，ケプストラム空間に対する時間変動しない可逆な変換でモデル化できることを述べた．さらに，ミスマッチ問題を解決するために従来行われている手法のうち，最も基本的な 4 つの手法について説明を行った．これらの手法を用いることで，ミスマッチ問題はいくらか低減される．今回説明した 4 つの手法以外にも，これらを拡張・発展させたさまざまな改良手法が，現在でも研究されている．

第4章

音声の構造的表象

4.1　はじめに

音声は環境により大きく変動する．そのため，音響モデルを構築する際どんなに多くのデータを集めても，モデルと入力のミスマッチ問題を避けることができない．ミスマッチ問題の解決法として，音響モデルの適応が広く行われているが，“There is no data like more data.” の言葉通りデータは本質的に不足しており，汎化能力の高いなんらかの手法による解決が求められる．例えば Furui は音声認識のレビュー論文において，HMM を識別学習・適応する際，汎化能力を高めるさまざまな工夫をした手法を紹介すると共に，ゴールデン・スタンダードと呼べる汎化手法はまだ存在しないと述べている [35].

外国語発音評価システム用音響モデルの学習・適応では，汎化能力を高めることはより難しい問題になる．例えば，母語や発音習熟度によっても音声が変動するため，過適応の問題が発生しやすくなる [14]．また，2011 年度からの小学校 5・6 年生の英語活動必修化を受けてユーザに小さな子供が増えることが予想されるが，子供の音声は大人の音声よりも変動が大きく，汎化はより難しい [15].

近年，解くべき課題の多い HMM の適応とはまったく異なるミスマッチ問題の解決法として，静的な変動に不変な音声の相対関係の利用が提案された [16]．これは，音声の変動に頑健な特徴量を用いることでをミスマッチ低減させる手法であり，HMM の適応とは同じ目的で異なる視点からのアプローチとなる．この音声の相対関係から得られる情報は音声の構造的表象と呼ばれ，外国語発音評価における有効性が示されている [17, 18, 19, 20, 21, 22]. また，構造表象を利用した音声認識 [36, 37, 38, 39, 40, 41, 42] も検討され，構造表象による音声分析技術を高度化されている．

本章では，まず音声の構造的表象と，その抽出方法について述べる．次に，先行研究として既に行われている構造表象に基づく孤立単語音声認識，外国語発音分析に関する実験的検討について述べる．

4.2　音声の構造的表象

話者の違いやマイクなどの伝送特性の違いは，ケプストラム空間における可逆な空間写像で近似できる．例えば MLLR 適応では，この空間写像を線形変換と仮定し，適切な変換パラメータを推定することによって実現される．また，GMM を用いた話者変換では，話者変換に対応する非線形写像を GMM で学習することによって実現されている．

二つの空間が可逆な空間写像で結びつけられ，それぞれの空間において対応する複数の分布が存在する場合，それぞれの空間における分布間の f-divergence は，常に不変となる [43]．f-divergence とは，分布間距離尺度の一種であり，二つの分布 p_i, p_j 間の f-divergence は以下の汎関数で表される．

$$f\text{-div.}\,(p_i, p_j) = \int p_j(\boldsymbol{x})\, g\left(\frac{p_i(\boldsymbol{x})}{p_j(\boldsymbol{x})}\right) d\boldsymbol{x} \tag{4.1}$$

ただし，$g(t)$ は $t > 0$ で定義する凸関数であり，$g(1) = 0$ を満たすとする．

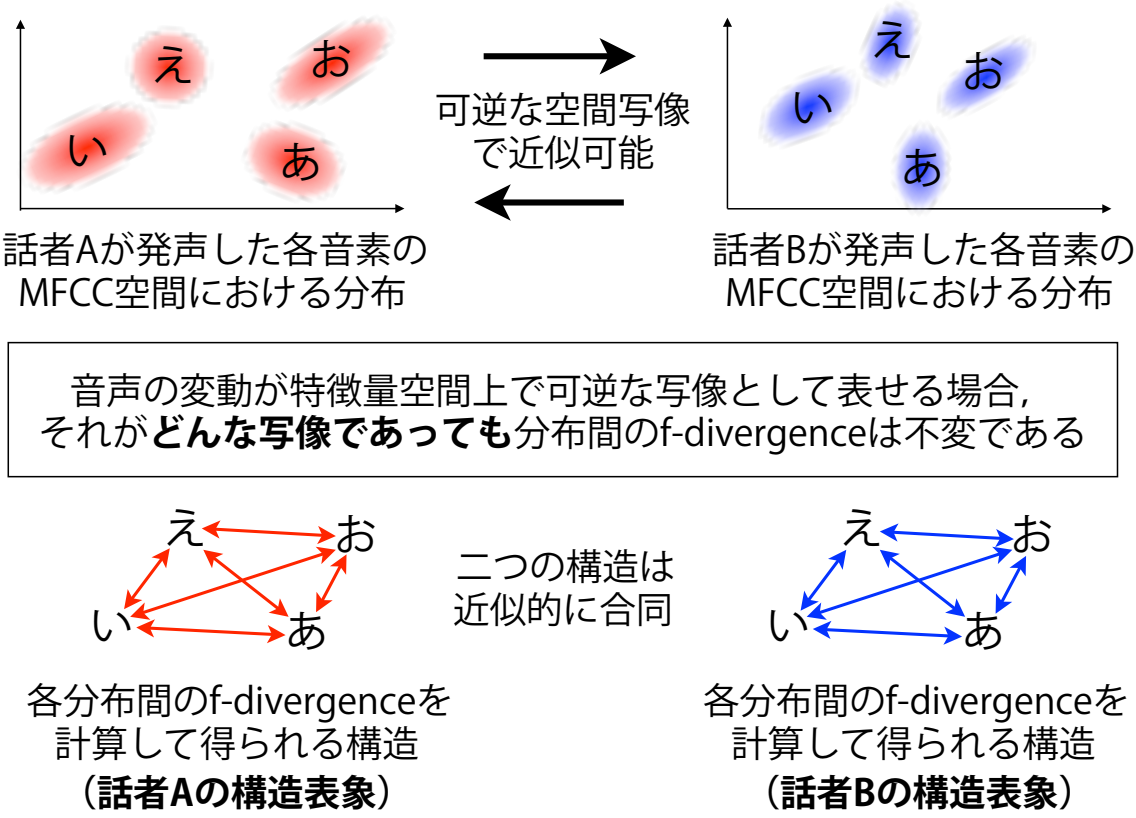

図 4.1: 静的な変動に不変な音声の構造的表象

図 4.1 に，MFCC 空間における対応する音素分布間の f-divergence が不変になる様子を示す．ただし，ここで音素が MFCC 空間において 1 つの分布から生成されると仮定している．この仮定の下，ケプストラム空間上で分布として表現された音素分布間の f-divergence 及びその関数は，時間変動しない可逆な変換で近似できる話者の違い，音響デバイスの違い，定常雑音に近似的に不変になる．我々は，すべての音響イベント分布間の f-divergence またはその関数を計算することによって得られる行列表象を，音声の構造的表象と呼んでいる．以降，音声の構造的表象のことを省略して「構造」と呼ぶ．

4.2.1　f-divergence の不変性の証明

可逆な変換 $\boldsymbol{y} = h(\boldsymbol{x})$ により，$p_i(\boldsymbol{x}), p_j(\boldsymbol{x})$ がそれぞれ $q_i(\boldsymbol{y}), q_j(\boldsymbol{y})$ に変換される場合の f-divergence の不変性は，$J(\boldsymbol{y})$ を $h^{-1}(\boldsymbol{y})$ のヤコビアンの行列式の絶対値として，以下のよ

うに証明できる．

$$
\begin{aligned}
f\text{-div.}\,(p_i, p_j) &= \int p_j(\boldsymbol{x})\, g\left(\frac{p_i(\boldsymbol{x})}{p_j(\boldsymbol{x})}\right) d\boldsymbol{x} && (4.2)\\
&= \int p_j(h^{-1}(\boldsymbol{y})) g\left(\frac{p_i(h^{-1}(\boldsymbol{y}))J(\boldsymbol{y})}{p_j(h^{-1}(\boldsymbol{y}))J(\boldsymbol{y})}\right) J(\boldsymbol{y}) d\boldsymbol{y} && (4.3)\\
&= \int q_j(\boldsymbol{y}) g\left(\frac{q_i(\boldsymbol{y})}{q_j(\boldsymbol{y})}\right) d\boldsymbol{y} && (4.4)\\
&= f\text{-div.}\,(q_i(\boldsymbol{y}), q_j(\boldsymbol{y})) \quad \square && (4.5)
\end{aligned}
$$

以上により，f-divergence が可逆な変換に不変になる十分性が証明された．なお，証明は省くが，可逆な変換 $\boldsymbol{y} = h(\boldsymbol{x})$ に不変な分布間距離は f-divergence でなければならないという必要性も証明することができる [43].

4.2.2　f-divergence の実装

$g(t)$ を変化させた f-divergence の関数は，具体的には Bhattacharyya distance，KL-divergence，Hellinger distance などになる．本研究では，構造表象を計算する実装において，Bhattacharyya Distance（BD）の平方根を利用する．BD は，$g(t) = \sqrt{t}$ とした場合の f-divergence の $-\ln$ をとったものであり，二つの分布 $p_i(\boldsymbol{x}), p_j(\boldsymbol{x})$ 間の BD は下記で定義される．

$$
\mathrm{BD}\,(p_i, p_j) = -\ln \int \sqrt{p_i(\boldsymbol{x}) p_j(\boldsymbol{x})} d\boldsymbol{x} \tag{4.6}
$$

また，2 つの分布がガウス分布 $\mathcal{N}_a(\boldsymbol{\mu_a}, \boldsymbol{\Sigma_a}), \mathcal{N}_b(\boldsymbol{\mu_b}, \boldsymbol{\Sigma_b})$ だった場合，BD はガウス分布の平均と分散共分散行列の閉形式で記述できる．

$$
\mathrm{BD}\,(\mathcal{N}_a, \mathcal{N}_b) = \frac{1}{8}\left(\boldsymbol{\mu_a} - \boldsymbol{\mu_b}\right)^T \left(\frac{\boldsymbol{\Sigma_a} + \boldsymbol{\Sigma_b}}{2}\right)^{-1} \left(\boldsymbol{\mu_a} - \boldsymbol{\mu_b}\right) + \frac{1}{2}\log \frac{\left|\left(\boldsymbol{\Sigma_a} + \boldsymbol{\Sigma_b}\right)/2\right|}{\left|\boldsymbol{\Sigma_a}\right|^{\frac{1}{2}} \left|\boldsymbol{\Sigma_b}\right|^{\frac{1}{2}}} \tag{4.7}
$$

4.2.3　音声の構造的表象の抽出

f-divergence を計算して構造抽出するためには，音響イベントごとに，その音響イベントを生成する分布を推定する必要がある．音響イベントごとに分布化されていれば，それらすべての分布間 f-divergence 距離行列を計算することで，構造が抽出できる[1]．しかし，分布を推定する際に真の分布を得ることは不可能であり，適当な仮定や近似をしなければならない．また，そもそも音響イベントが 1 つの分布から生成されるというものは，あくまで仮定であり，これによる近似誤差は当然生じている．

このような問題は，音響モデルのデファクトスタンダードである HMM にも発生している．HMM では，音響イベントを各音素につき約 3 つ割り当て，それらの分布をガウス分布を複数組み合わせた混合ガウス分布（Gaussian Mixture Model; GMM）を用いてモデル化している．このような近似を行っている HMM は，現実的なタスクにおいて大きな成功

[1]距離行列が与えられれば，構造の形は一意に定まるため．

をおさめているため，構造を用いた音声分析においても，HMM と似た仮定をおいて分布を推定している．

以下，構造を抽出するために提案されている音声を分布化手法を述べる．具体的には，1) 孤立単語発声から母音などの特定の音響イベントを切り出してガウス分布化する方法，2) 特定話者音素 HMM を学習する方法，3) 一発声から一つの HMM を学習する方法，4) 一発声をボトムアップクラスタリングすることによりガウス分布系列化する方法を説明する．

孤立単語発声から音響イベントを切り出して分布化する方法では，孤立単語発声を強制アライメント等をして音響イベントを切り出し，所望の区間がガウス分布から生成されたと仮定して平均値と分散共分散行列を計算することで分布化する．例えば，[20] では，V を母音として/bVt/，もしくは/pVt/となる米語単語（pot，bat，but，bought，bet，bird，bit，beat，put，boot）を発声させ，それらの中心母音を切り出して分布化を行っている．

特定話者音素 HMM を学習する方法では，ある話者が発声した複数の文から特定話者音素 HMM を学習し，各音素 HMM の状態が出力する確率分布をひとつの音響イベント分布として利用する．例えば，[17] では，TIMIT の音素バランス文の一部である 60 文程度の音声を用いて特定話者音素 HMM を学習し，構造を抽出している．

一発声から HMM を学習する方法では，文の読み上げ音声が 1 つの HMM から生成されたと仮定し，それを学習することで，音響イベント分布を得る．例えば，[38] では，20 状態の HMM 学習することで分布化している．一発声から分布をつくる別の手法として，ボトムアップクラスタリングを用いて分布化する手法もある．例えば，[44] では，情報量規準に基づき適当な分布数になるまでボトムアップクラスタリングを行うことで分布化を行っている．なお，一発声から分布を得る方法では，分布を推定するためのデータ数が少ないという問題が不可避であるので，ML 推定で分布を推定するのではなく，事前知識を利用して分布 MAP 推定することが有効である．

4.3　構造的表象を用いた音声分析

構造表象を用いて音声分析を行う際には，二つの構造間差異を定義する必要がある．ここで，二つの構造間差異を，適切に構造を回転・シフトさせて構造を合わせてもなお残る構造の各頂点の差の和と定義する．構造間差異の概念図を図 4.2 に示す．このとき，構造間差異は，構造を表す f-divergence 距離行列の上三角成分を並び替えて作ったベクトルのユークリッド距離と非常に相関が高くなる[2]ことが知られている [45]．そのため，構造の回転やシフトに対応する演算を明示的に行わなくても，構造間差異を計算することができる．f-divergence 距離行列の上三角成分を並び替えて作るベクトルのことを，構造ベクトルと呼ぶ．そのため，構造ベクトル空間でさまざまな処理を行うことで，構造的表象を用いた音声分析が可能になる．

構造表象は，あらゆる可逆な変換に対して不変である．そのため構造表象を用いた音声分析は，理論的には，CMN，最適なスペクトル包絡が得られた SS，最適な声道長パラメータが得られた VTLN，最適な変換行列が得られた MLLR 適応などを内包する効果を，それ

[2]相関係数は 0.99 以上になる．

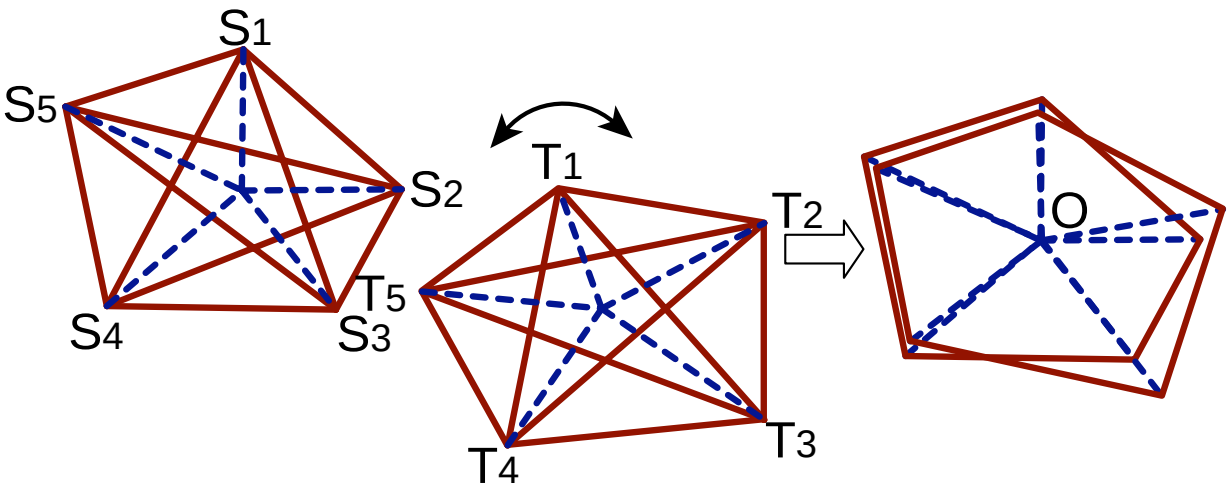

図 4.2: 二つの構造間差異の定義

らを行うことなしで得ることができる．ただし，音響イベントの分布推定精度の問題に起因する歪みを不可避に有しており，さらには後述するように不変性が強すぎるという問題もあり，これが構造表象を用いた音声分析における問題となっている．

4.3.1　構造的表象を用いた孤立単語音声認識

構造的表象を用いて孤立単語音声認識を行うには，入力単語音声を構造ベクトル化した後，構造ベクトル空間において単純なパターン認識問題を解けばよい．最も簡単な方法としては，構造を単語単位に統計モデル化し，入力の構造と最も近い構造統計モデルに対応する単語を認識結果とする方法がある [37]．一発声から HMM を用いて構造化し，構造統計モデルを学習して音声認識を行う枠組みを，図 4.3 に示す．

しかし構造を用いた音声認識には，二つの問題点がある．一つは，f-divergence の不変性が強すぎて，異なる単語でさえも同じ単語であると判定されてしまう問題．もう一つは構造ベクトルの次元数が高すぎて次元の呪いが発生してしまう問題である．

まず f-divergence の不変性が強すぎる問題について述べる．f-divergence の不変性は，線形・非線形を問わずあらゆる変換に対して成り立つ．この非常に強い不変性により，効率的に話者性を取り除くことが可能となるが，それと同時に，非言語的特徴を表す変換としてあり得ないような変換にも不変となってしまい，言語的には全く異なる単語同士が同一と見なされてしまうことが起こりうる．単語音声認識タスクにおいては，異なる単語同士を比較することが必要となるため，この強すぎる不変性は認識性能の低下を引き起こすと考えられる．

この問題を解決する方法として，f-divergence 不変性に制限をかけるマルチストリーム構造化が提案されている [38]．マルチストリーム構造化は，声道長の違い，音響デバイスの違い，定常雑音による音声の変換が，近い次元間での変換のみで近似できることに基づく手法である．例えば [46] では，声道長の違いが以下のような帯行列による線形変換で近

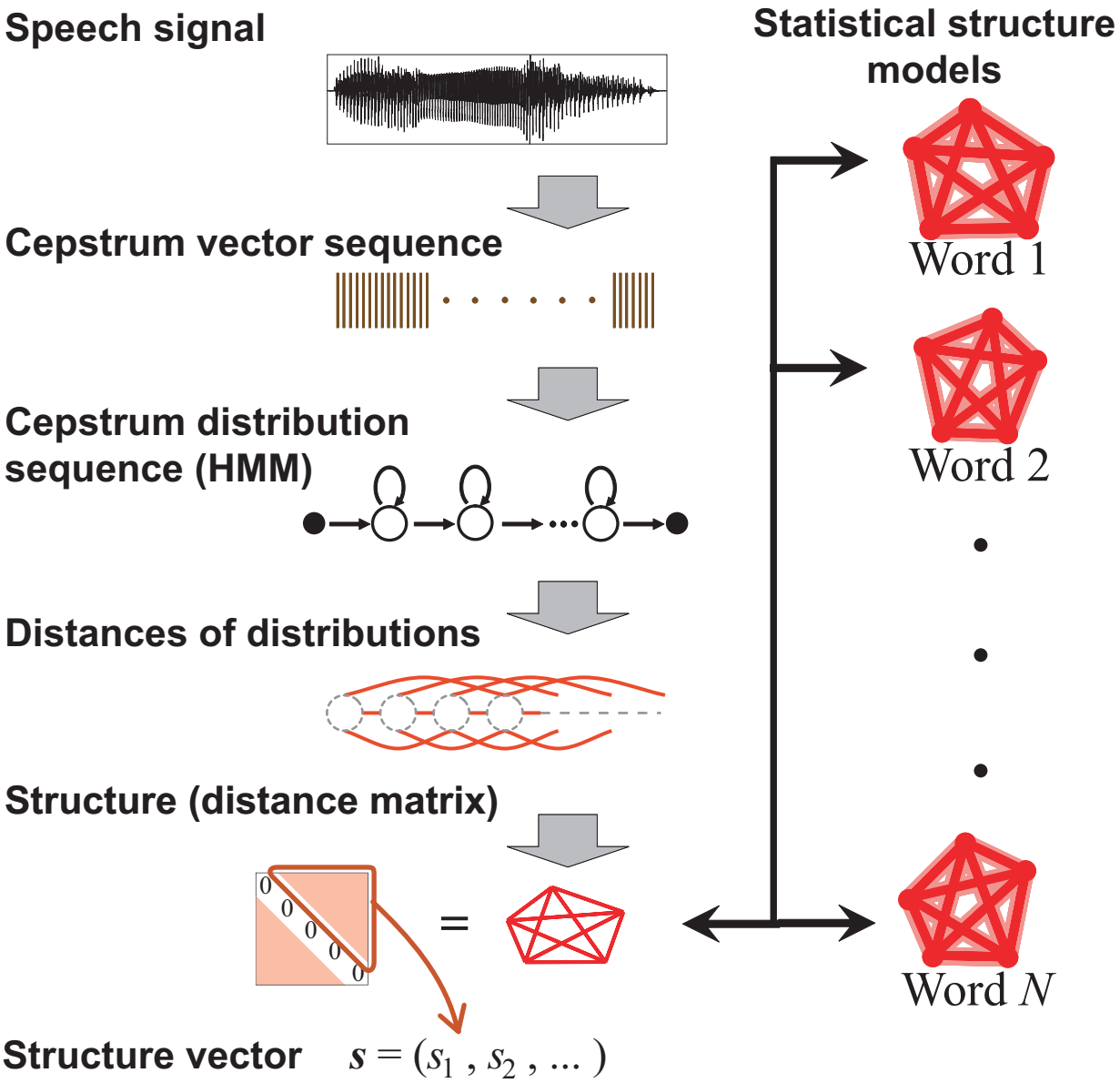

図 4.3: 構造統計モデルを用いた孤立単語音声認識

似できることを示している．

$$\boldsymbol{A} = \begin{pmatrix} 1 & \alpha & \alpha^2 & \alpha^3 & \cdots \\ 0 & 1-\alpha^2 & 2\alpha - 2\alpha^3 & \cdots & \cdots \\ 0 & -\alpha + \alpha^3 & 1 - 4\alpha^2 + 3\alpha^4 & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix} \tag{4.8}$$

ここで α は声道長の違いの度合いを表す周波数ウォーピングパラメータであり，$|\alpha| < 1$ で，$\alpha = 0$ のときに変換なし（$\boldsymbol{A} = \boldsymbol{I}$），$\alpha > 0$ のときに声道長を短くする変換，$\alpha < 0$ のとき

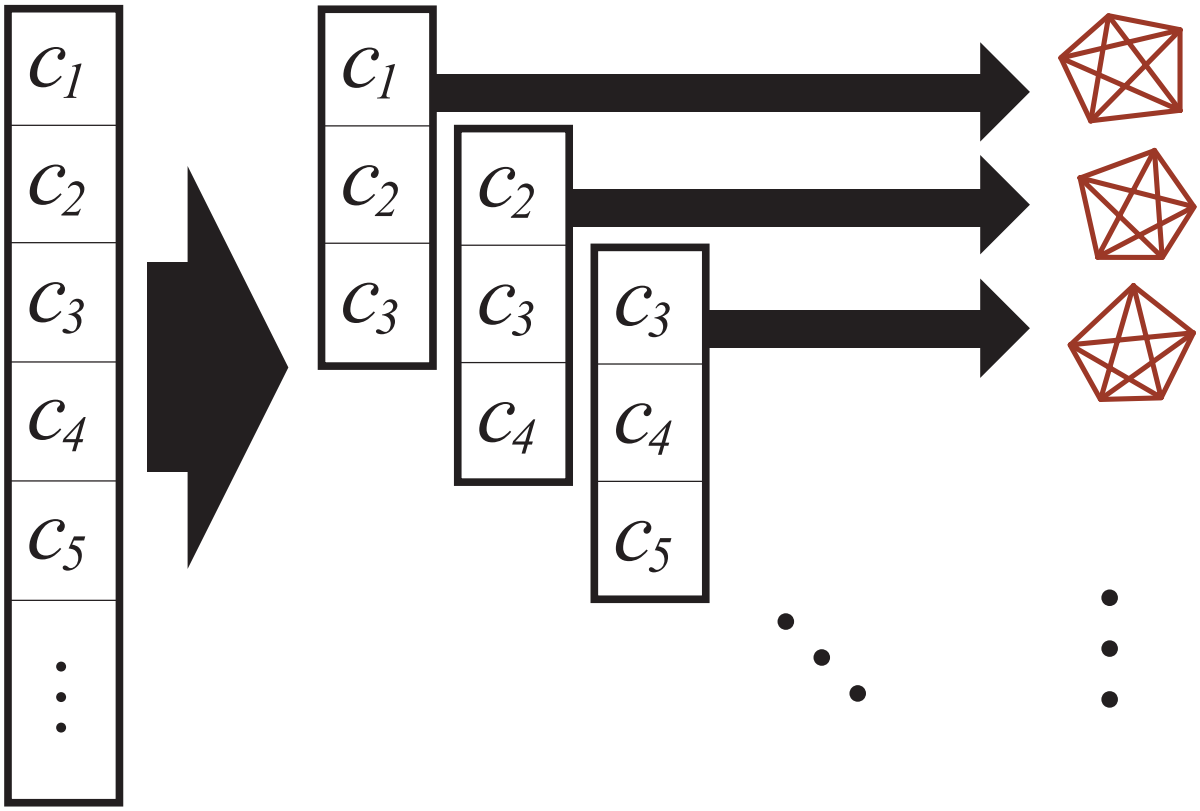

図 4.4: マルチストリーム構造化

に声道長を長くする変換に対応している．なお，$\alpha = \pm 0.4$ のときに声道長をおよそ 1/2・2 倍に変換することに対応する．このような帯行列の線形変換は，次元に番号をふった場合に，番号が近い次元間では依存関係があるが，番号が遠い次元間には依存関係が無いような変換であるといえる．また，音響デバイスの違いや，定常雑音の違いも同様に，異なる番号の次元には依存関係が無いような非線形変換になっている．

マルチストリーム構造化は，次元番号が近いもの同士においてのみ不変性を導入し，次元番号が遠いものに関しては制約をかけないための手法である．マルチストリーム構造化の概念図を，図 4.4 に示す．図 4.4 のようにケプストラムをブロックごとに分割することによりマルチストリーム化させてそれぞれ構造を抽出すると，それぞれの構造は，そのストリーム内の変換のみに不変となり，ケプストラム全体を使った構造化と比較して，不変性に制約をかけることができる．図 4.4 では，各ストリームの特徴量ブロックを 3 次元ずつにしているが，この次元数 s は不変性の強さを制限するパラメータとなる．すなわち，s がもとの特徴量の次元数と等しいのときに不変性はもっとも強く，通常の構造と同じになり，逆に $s = 1$ のときに不変性はもっとも弱く，定数倍の変換のみに不変になる．

マルチストリーム構造化は適切に s を選択することで，認識率を向上させることができるが，同時に，次元数を大幅に増やしてしまうため，次元の呪いが発生してしまう．もともと構造ベクトルは，音響イベントの数を n として ${}_nC_2$ 次元と高く，それがマルチストリーム構造化によりストリーム数倍されてしまうことになる．そこで，適当な次元圧縮を用い

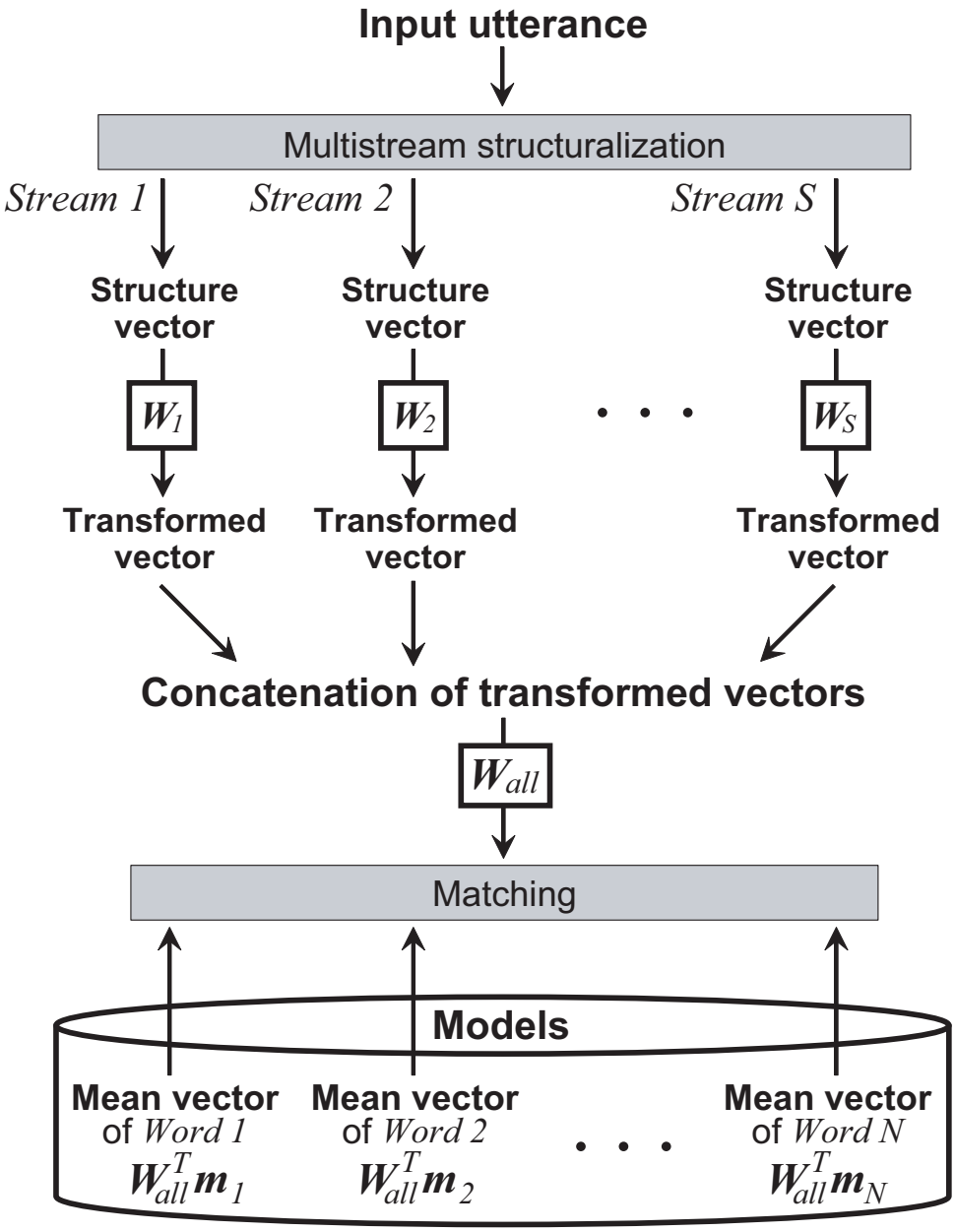

図 4.5: 2 段階 LDA を用いた音声認識

ることで次元の呪いを解決する手法が提案されている．ただし，教師あり次元圧縮を単純に用いると，あまりに高すぎる次元により，過学習がおき汎化性能が低下する可能性がある．そのため，適当な制約条件をかけた次元圧縮法が提案されている．例えば，ランダムに構造のエッジを選択したのちに判別分析を行う手法 [39]，教師なし次元圧縮法である PCA をかけてから LDA を用いる手法 [41]，2 段階で LDA をかける手法 [40]，パラメータ共有によって次元数を下げる手法 [42] が提案されている．図 4.5 に，[40] で提案されている 2 段階 LDA を用いて音声認識を行う枠組みを示す．まず，マルチストリーム構造化を用いて，S

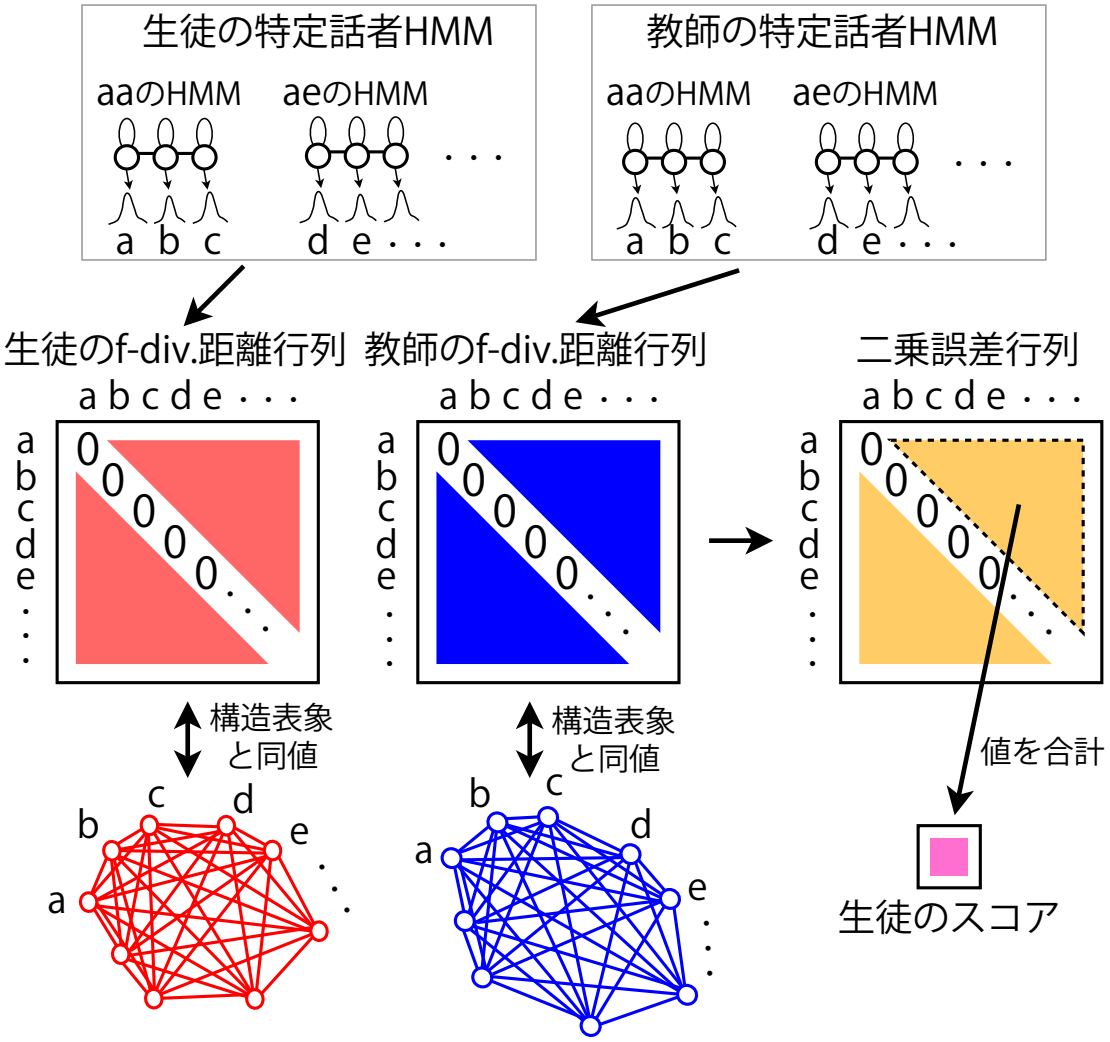

図 4.6: 構造表象を用いた外国語発音評価

個の構造ベクトルを図 4.3 の要領で計算する．その後，各ストリームにおいて別々に，LDA を用いて次元圧縮を行う．次に，各ストリームでそれぞれ次元圧縮されたベクトルを結合し，それに対して LDA を用いて認識を行う．

4.3.2　構造的表象を用いた外国語発音分析

構造的表象を用いて外国語発音分析を行うには，構造ベクトル空間において，教師の構造ベクトルと生徒の構造ベクトルを比較し，どの部分が大きくずれているかを調べればよい．最も単純な外国語発音評定としては，f-divergence の距離行列を教師と生徒で二乗誤差をとり二乗誤差行列を計算し，対称行列である二乗誤差行列の上三角成分の和，すなわち構造ベクトルのユークリッド距離をとることで生徒のスコアを算出する方法がある [17].

特定話者音素 HMM を学習して構造を抽出し，二乗誤差行列の上三角成分の和をとることで外国語発音評定を行う枠組みを，図 4.6 に示す．

二乗誤差行列は，生徒の全体を通してのスコアの情報意外にも，どの音響イベントの発音がどう違うのかといった情報も含んでいる．これを用いることで，どの音素から発音を直して行くべきかなどといった発音分析を行うことができる．例えば [20] では，二乗誤差行列の各行の和を，生徒の発音をどの音響イベントから直していくべきかの推定値として利用できることを示している．

生徒のスコアや各音響イベントのスコアの推定の他に，構造の可視化や，発音に基づくクラスタリングも行われている．例えば [21] では，樹形図や多次元尺度構成法（MDS）を用いて構造の可視化を行っている．ある話者の米語母音と日本語母音の発音をこれらの方法で視覚化したものを図 4.7，図 4.8 に示す．また [22] では，構造ベクトル空間において話者分類を行うことにより，発音のみに基づく話者分類行い，それを可視化している．MDS をもちいて発音に基づく話者分類を視覚化したものを図 4.9 に示す．

4.4　まとめ

本章では，音声の構造的表象について述べた．まず，音声を適当な音響イベント単位で分布化し，変換不変性を持つ分布間距離尺度 f-divergence を用いることにより，あらゆる変換に不変な音声の構造的表象を抽出することを述べた．さらに，二つの構造間の比較には，f-divergence 距離行列の上三角成分を並べてベクトルにした構造ベクトルのユークリッド距離を利用することを述べた．さらに，構造の持つこれらの性質を利用した，音声認識や発音分析手法についても述べた．

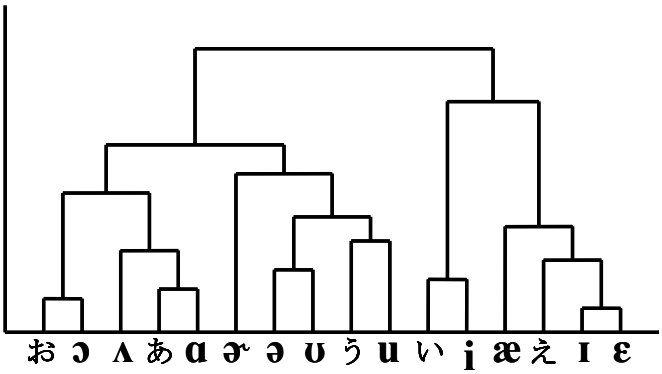

図 4.7: 発音構造の樹形図による視覚化

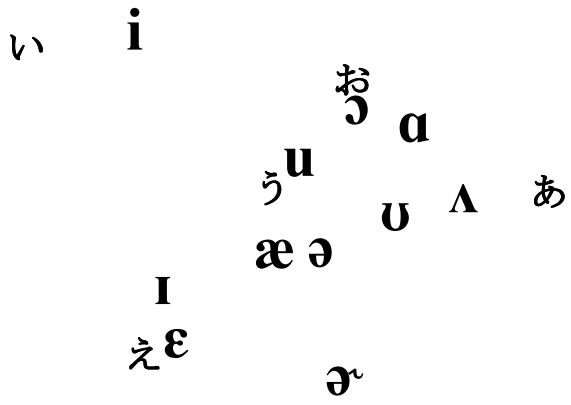

図 4.8: 発音構造の MDS による視覚化

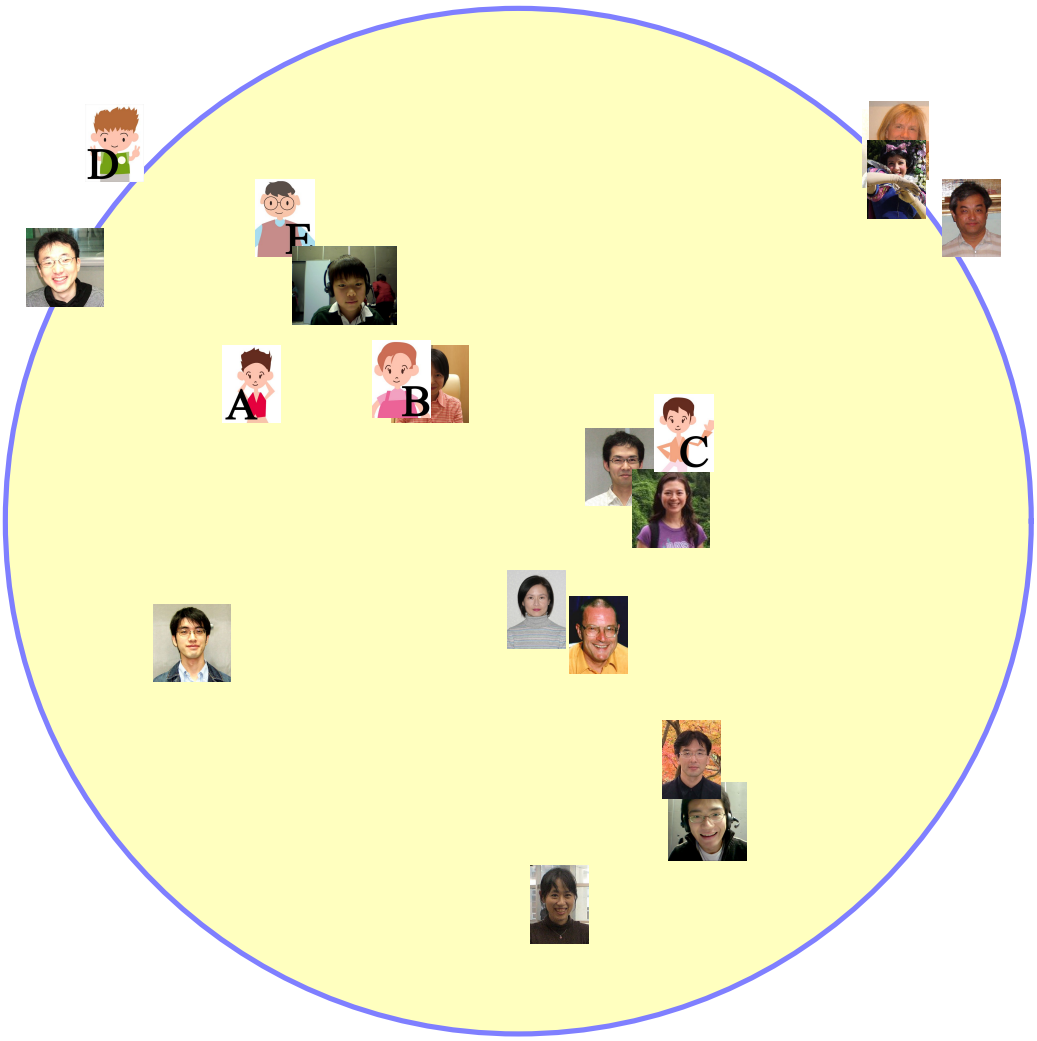

図 4.9: 発音に基づく話者の分類の MDS による視覚化

第5章

構造を用いた発音分析の高精度化

5.1　はじめに

第 4 章では，音声の構造的表象と，それを用いた音声分析法について述べた．外国語の発音分析においても，構造を利用することにより，話者の違いなどに非常に頑健な処理が可能になる．しかし，構造を用いた外国語発音分析の実装は，音声認識タスクと比較して非常に単純な実装が利用されており，様々な観点から高精度化させることが可能であると考えられる．

そこで本章では，構造を用いた発音分析に対する改善手法を提案する．具体的には，構造を比較する際における正規化，タスクに合わせた適切な次元圧縮，マルチストリーム化による精度向上法を提案し，それぞれの効果を実験的に検証する．

5.2　エッジ長正規化

従来，構造間の差異尺度として，f-divergence 距離行列の上三角部分の要素の二乗誤差の和

$$D_1(\boldsymbol{S}, \boldsymbol{T}) = \sqrt{\frac{1}{M} \sum_{i<j} (S_{ij} - T_{ij})^2} \tag{5.1}$$

を用いていた．これは，二乗誤差行列の上三角成分を並び替えたベクトルにしたもののユークリッド距離である．ここで，構造のエッジには，話者平均的に長いもの，短いものが混在している．そのため，(5.1) を使う場合，平均的に長いエッジの差異が結果に大きく反映され，短いエッジの差が無視される危険性がある．この問題を避けるため，(5.1) の代わりに

$$D_2(\boldsymbol{S}, \boldsymbol{T}) = \sqrt{\frac{1}{M} \sum_{i<j} \left(\frac{S_{ij} - T_{ij}}{S_{ij} + T_{ij}} \right)^2} \tag{5.2}$$

を用いることを考える．(5.2) を使うことで，各エッジの差の二乗の値域が 0 から 1 に正規化され，エッジ長の相対的な差によって構造間差異を計算することができる．このような操作を，今後エッジ長正規化と呼ぶ．また，エッジ長正規化を用いて計算した二乗誤差行列を，正規化二乗誤差行列と呼ぶ．

なお，構造統計モデルを用いた音声認識では，このようなエッジ長正規化の効果は低いと考えられる．構造統計モデルでは，それぞれのエッジ長をガウス分布でモデル化し，それを音響モデルとして利用している．そのため，モデルがエッジ長の分散を正規化しており，エッジ長正規化を行わなくてもよいと考えられる．一方，発音評価タスクでは，教師の構造と生徒の構造を一対一で比較している．このような場合には，エッジ長正規化の効果は高いと考えられる．

5.2.1　実験的検証

エッジ長正規化を導入した場合としていない場合の外国語発音評価の精度を比較する実験を行う．精度比較には，手動評価値との相関を利用し，相関が高いものをより精度の高

表 5.1: 構造を抽出するための音響分析条件

サンプリング	16bit / 16kHz
窓	25 msec 幅，10 msec シフト
学習データ	一名につき 60 文（set 8 を読み上げた話者のみ 40 文）
HMM 学習時の特徴量	MFCC(12)+Energy(1)+ΔMFCC(12)+ΔEnergy(1)
構造抽出時の特徴量	MFCC(12)
HMM の種類	monophone，特定話者音素 HMM
出力確率分布	対角共分散行列を持つガウス分布
トポロジー	3 状態の left to right 型
音素の種類	aa,ae,ah,ao,aw,ax,axr,ay,b,ch,d,dh,eh,er,ey,f,g,hh,ih, iy,jh,k,l,m,n,ng,ow,oy,p,r,s,sh,t,th,uh,uw,v,w,y,z,zh,sil 合計 42 種類

い評価法とする．

発音評価用データベースには，ERJ データベースを用いる [47]．ERJ データベースは，200 名の日本人大学生が読み上げた米語音声，同じ文章を 20 名の米語ネイティブスピーカーが読み上げた音声が収録されている．読み上げられた文章セットには，8 つの種類があり，各セットには，TIMIT に含まれる 60 文などが含まれている．200 名の学生それぞれは，この文セットのうち，1 セットだけを読み上げている．そのため，1 セットにつき約 25 名の学生が読み上げている計算になる．また，20 名の米語ネイティブスピーカーそれぞれは，文セットのうち 4 セットを読み上げている．ただし，20 名中 2 名（男性 M08 と女性 F12）は，全 8 セットを読み上げている．さらに，学生の発声の一部に対する手動評価値も収録されている．各学生が読み上げた文章のうち，1 人につき 10 文の発声に対して手動評価が行われている．評価者は日本人学習者の癖をよく理解している，米語ネイティブスピーカーである音声学者 5 名である．評価は，音素的に正しく発声されたか，リズムが正しく発声されたか，イントネーションが正しく発声されたか，の 3 つの尺度が用いられ，それぞれ 5 段階で評価されている．

構造を抽出するための音響分析条件を表 5.1 に示す．まず，200 名の学生から，各々 42 個ずつの特定話者音素 HMM を作成し，それらの間の $\sqrt{\mathrm{BD}}$ 距離行列を計算することで構造を抽出する．HMM のトポロジーとしては，3 状態の left-to-right 型 HMM を用いた．そのため，音響イベントの単位としては音素を 3 分割したものとなる．ただし，評価を音素ごとに行うために，3 状態音素 HMM のそれぞれの状態が出力する分布間の 3 つの $\sqrt{\mathrm{BD}}$ の和を 2 つの音素間距離として構造を作成する[1]．これにより，${}_{42}C_2 = 861$ 本のエッジからならなる構造が得られる．次に，同様の読み上げ文セットを学習データ，同様の条件で教師の構造も抽出する．ここで教師としては，男性教師 M08 1 名のみを利用した．

このように抽出した学生の構造と教師の構造の構造間差異をエッジ長正規化あり／なし

[1]音素を 3 分割したものを構造のノードとして利用することも可能である．ただし，ノードを音素にした方が結果がわかりやすいことや，予備実験の結果構造のノードとして音素を用いた方が良かったため，音素をノードとした．

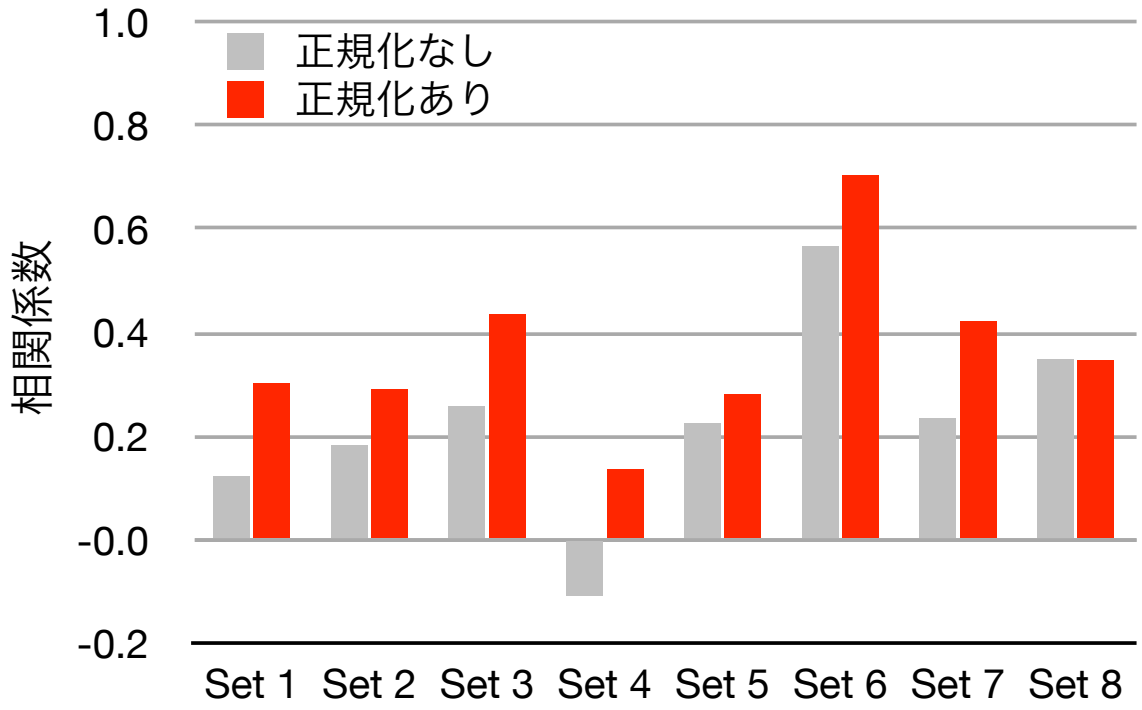

図 5.1: エッジ長正規化による手動評価値との相関の変化

計算し，それをセットごとに話者ごとに手動評価値との相関値を算出した．ここで手動評価値には，ERJ の各話者につけられた 10 文 ×5 名の音素的正しさに対する評価値の平均値を，話者の手動評価値として利用した．

5.2.2　実験結果

実験結果を図 5.1 に示す．結果，ほぼすべてのセットで，エッジ長正規化により精度が向上していることが分かる．8 セットそれぞれの相関の平均は，正規化なしで 0.23，正規化ありで 0.36 となった．構造間差異計算時における正規化は，単純な手法ではあるものの高い効果が得られることがわかる．ただし，正規化を用いた場合でも，十分な相関が得られているとは言い難く，なんらかの他の手法を用いて精度をより改善させる必要がある．

5.3　部分構造化

構造を表現する特徴量の次元数，すなわち構造のエッジの数は，音響イベント数 M に対して $M(M-1)/2$ となり，M^2 のオーダーである．そのため，子音も評価に含めるなどして M を大きくした場合，構造の次元数は非常に大きくなってしまう．本章で検討している発音評価タスクでは，具体的に $M = 42$，$M(M-1)/2 = 861$ となっている．次元数が高

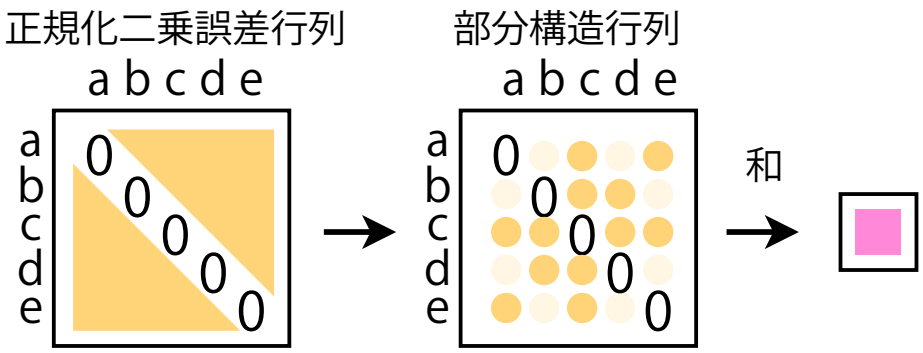

図 5.2: 正規化二乗誤差行列からの部分構造抽出と比較

くなると，識別に無関係な次元が増え，球面集中現象と呼ばれる現象により，一般的に二つの特徴ベクトル間の距離がほとんど一定になってしまい，パターン認識問題の精度が極端に低下してしまう「次元の呪い」の問題が発生することが知られている．

このような場合，構造を用いた音声認識と同様，なんらかの次元圧縮を用いることが有効であると考えられる．しかし，構造を用いた音声認識で行われているような，構造ベクトルに対する PCA や LDA などの線形変換を用いた次元圧縮を発音分析に用いるには問題がある．これは，構造を用いた外国語発音評価では，二乗誤差行列の各行の和など，二乗誤差行列の配置そのものが有効な情報持っており，構造ベクトルの線形変換はその情報を消してしまうためである．

そこで，二乗誤差行列のかたちをまったく崩さない次元の呪いの解決法として，部分構造化を提案する．構造の特徴量は各エッジの長さであるので，特徴量選択は構造の部分構造を抽出する操作に対応する．そのため，特徴量選択による構造の次元圧縮を部分構造化と呼ぶ．部分構造化を用いることで，二乗誤差行列のかたちを崩すことなく，次元の呪いの問題を緩和させることが可能になる．部分構造化の概念図を図 5.2 に示す．部分構造化は，エッジ長（$\sqrt{\mathrm{BD}}$）を特徴量として，特徴量選択を行うことで実現される．なお部分構造化は，評価が不十分ではあるものの，[17] で既に一部検討されている．ここでは，それをより厳密かつ詳細に検討する．

5.3.1　特徴量選択

部分構造化の性能は，いかにして部分構造を選択するか，すなわち特徴量選択をどう行うかに依存する．特徴量選択の方法にはさまざまなものが考えられるが，ここでは手動評価値との相関値を規準としてフィルター法で特徴選択を実行する方法と，ルールベースで決定する方法の，二通りの手法を検討する．

i)　フィルター法による特徴量選択

手動評価値との相関値を規準としてフィルター法で特徴選択する方法では，あらかじめ学習データとして用意された手動評価値を用い，その評価値と部分構造を用いた評価値との相関が最大されるような部分構造を選択する．選択アルゴリズムとしては，単純な全探索は計算量が高すぎ，また過学習になる可能性が高いため，forward stepwise selection や backward stepwise selection などを利用する．forward stepwise selection では，相関を上げる効果が一番高いエッジ一つずつ追加していくことで部分構造を得る．backward stepwise selection では，forward stepwise selection とは逆向きに，相関を下げる効果が一番高いエッジを一つずつ捨てていくことで部分構造を得る．また，forward stepwise selection と backward stepwise selection を両方行い，選択されたエッジ数が同じところで二つの方法を比較して良い方を選ぶアルゴリズムもある．なお，これらのアルゴリズムを用いた場合，最終的にいくつの特徴量を用いるのかは，別の手法を用いて決定する必要がある．

ii)　ルールベースによる特徴量選択

ルールベースで決定する方法は，あらかじめ重要だと考えられるエッジをトップダウンに決定してしまう方法である．例えば日本人の米語発音を評価する場合，日本人が混同しやすい発音はある程度傾向がある．例えば，/r/と/l/であったり，/s/と/sh/といった音素を混同しやすい．そこで，混同しやすいとわかっているエッジを手動で選び出し，そのエッジのみを使った部分構造を用いる．この方法はフィルター法と異なり，学習データとして手動評価値を用意する必要がなく，対象者の発音に対する知識さえあれば部分構造を選択することができる．

選択するエッジの選び方としては，混同しやすい音素間のエッジを選ぶという方法の他，不変性が強いと考えられるエッジを選択することも有効である．f-divergence そのものは，あらゆる可逆な変換に常に不変になるが，f-divergence を計算するために分布を推定する時点において近似が含まれている．分布の推定の実装上，我々は分布をガウス分布と仮定しているため，分布のかたちは正しく推定されているとは限らず，特に，実際の観測点から遠い部分の分布の精度は低い．そのため，特徴量空間で大きく離れた音素，例えば/aa/と/s/のようなエッジはあまりつかわず，より近い音素，母音なら他の母音とのエッジを選んだ方が，f-divergence の不変性がより強いという意味でよいと考えられる．

5.3.2　実験的検証

部分構造を用いることによる外国語発音評価の精度の変化を見る実験を行う．実験は，先のエッジ長正規化の実験において，エッジ長正規化を導入した場合の条件と同様の実験条件で行う．

部分構造化における特徴量選択のアルゴリズムには，手動評価値との相関を規準として forward stepwise selection を用いたフィルター法による特徴量選択と，日本人が混同しやすい音素ペアのみを選んだルールベースによる特徴量選択の両方を試した．forward stepwise selection による特徴量選択の学習データには， 8 つの文セットのうち，1 つを除いた 7

表 5.2: ルールベースの特徴量選択で選択した音素組

選択した米語音素組	対応するカタカナ発音
aa, ae, ah, ao, er	ア
ih, iy, y	イ
uh, uw, w	ウ
s, sh, th	サシスセソ
z, jh, dh	ザジズゼゾ
m, n, ng	ナニヌネノ／マミムメモ
f, hh	ハヒフヘホ
b, v	バビブベボ
l, r	ラリルレロ

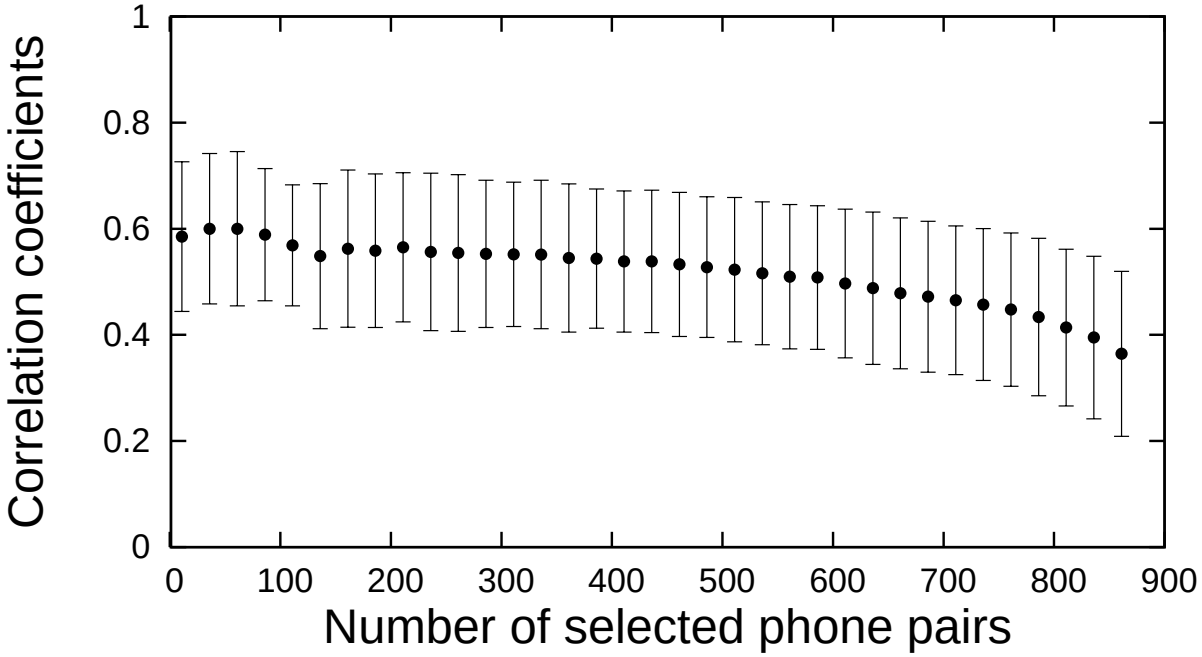

図 5.3: 部分構造分析したときの相関値の平均と標準偏差

セットを利用し，残った 1 セットを評価データとした（1set-leave-out）．ルールベースによる特徴量選択では，米語発音をカタカナ表記したときに同じカタカナで表現されることの多い音素ペアのみを選択した．具体的に選んだ音素ペアを表 5.2 に示す．これにより，${}_5C_2 + {}_3C_2 \times 5 + {}_2C_1 \times 3 = 28$ 本のエッジが選択されることになる．

5.3.3　実験結果

Forward stepwise selection により選択した部分構造による発音評価の結果を図 5.3 に示す．縦軸は手動評価値との相関を，横軸は選択されたエッジ数を，各プロットは 8 セット

表 5.3: 選ばれた 71 本のエッジ（音素の組み合わせ）

1 ∼ 5	L–R	S–AA	OY–TH	AX–IH	L–EY	
6 ∼ 10	AW–NG	D–L	P–Z	JH–NG	P–JH	
11 ∼ 15	V–AW	IH–AXR	HH–AXR	Z–sil	AH–AX	
16 ∼ 20	N–UW	G–HH	F–CH	IY–JH	R–IH	
21 ∼ 25	Y–EH	AE–AX	AH–IH	ER–OW	DH–HH	
26 ∼ 30	R–HH	AE–IY	HH–IY	T–sil	K–L	
31 ∼ 35	AO–AXR	N–AE	G–UH	M–OW	AO–AX	
36 ∼ 40	V–JH	P–EY	T–DH	F–Z	R–Z	
41 ∼ 45	CH–TH	AO–UH	Z–AW	AY–OW	AA–EY	
46 ∼ 50	D–JH	M–AA	R–UH	R–ER	F–AXR	
51 ∼ 55	AE–AY	T–CH	D–EY	D–TH	F–HH	
56 ∼ 60	AE–OY	P–AY	AY–ER	M–CH	F–DH	
61 ∼ 65	S–UW	F–M	AE–AH	AO–ER	AO–NG	
66 ∼ 71	P–S	D–P	V–W	S–Z	AO–AY	P–OY

の実験の平均値を，エラーバーは標準偏差を表す．なお，図 5.3 のうち，一番右側のプロット，すなわちすべてのエッジを利用したものが，部分構造化を用いずに構造を用いた発音評価を行った場合の結果である．

結果，全体的に左上がりの傾向がある．すなわち部分構造化によって相関が向上していることが分かる．条件によるが，今回の場合はエッジを 71 本選んだときに相関の平均 0.63 で最も高い精度が得られた．部分構造化を用いていない場合には相関の平均 0.36 であったので，大幅に精度が向上していることがわかる．

参考に，セット 1 を評価データとして学習した時に，71 番目までに選ばれた音素ペアを表 5.3 に示す．一番最初に L と R のエッジが選ばれるなど，おおよそ直感とあったエッジが選択されていることがわかる．なお，ここで部分構造として選ばれた 71 本のエッジからなる部分構造のノードには，42 音素中，セットによって出現しない場合がある音素 /ZH/ を除く 41 音素すべてが含まれている．

ルールベースで選んだ部分構造の結果は，相関の平均 0.56，標準偏差 0.14 となった．これは，最適なエッジ数が選択されたの forward stepwise selection の結果（平均 0.63，標準偏差 0.12）には若干劣るものの，ほぼ同等の精度になっている．すなわち，学習データを全く用いなくても，学習者の発音の癖に対する知識さえあれば，有効な部分構造が選択できるといえる．

5.4　2 段階重回帰

部分構造化による次元圧縮は，正規化二乗誤差行列の形を崩さないよう，各エッジに対して使う／使わないの 2 値のラベルをふることによって実現されている．これは，次元圧縮手法としては自由度が非常に低いものである．本来は，非常に重要なエッジ・重要なエッ

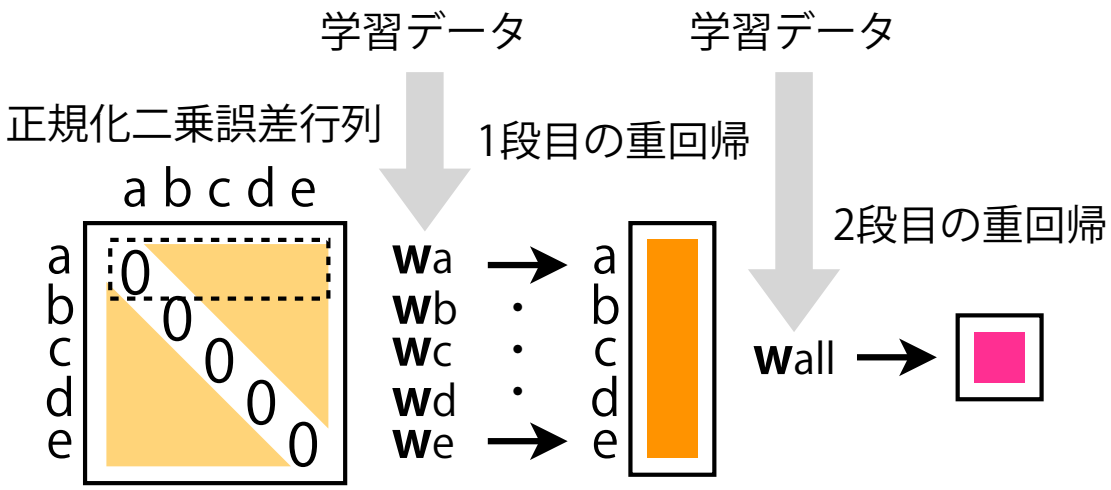

図 5.4: 2 段階の重回帰を用いた外国語発音評価

ジ・少し重要なエッジ・必要ないエッジといったように，段階的に重みをつけられる方が，次元圧縮法としてより望ましい．

そこで，部分構造化より自由度が高く，かつ二乗誤差行列の形も大きく崩さず意味解釈の行いやすい次元圧縮手法として，2 段階重回帰を用いた手法を提案する．2 段階重回帰を用いた外国語発音評価の枠組みを，図 5.4 に示す．なお，$\boldsymbol{w}_i$ は重回帰分析により推定された重みベクトルである．2 段階重回帰では，二乗誤差行列の上三角ベクトルを並べ替えた構造ベクトルに対して重回帰分析を行うのではなく，2 段階にわけて二乗誤差行列に対し重回帰分析を行う．構造ベクトルに対して重回帰分析を用いると，正規化二乗誤差行列の形態の情報が失われてしまう欠点の他，重回帰分析で推定すべき重みパラメータの数が多くなりすぎて，学習データの不足および過学習の問題が発生してしまう問題が発生する．2 段階で重回帰分析を行うことにより，学習するパラメータを大幅に減らして過学習を防ぎ，かつ二乗誤差行列の形態を残しながら次元圧縮を行うことができる．

1 段目の重回帰では，二乗誤差行列の行ごとに重回帰を行う．重回帰分析の目的変数としては，例えば各音素に対する手動評価値が利用できる．重回帰の重みパラメータ学習時には，ある音響イベントの発音を評価する際にどの音響イベントとの相対関係を重要視するかが学習される．例えば，日本人が発音する米語の/s/を評価する際には，/s/と混同しやすい/sh/や/th/などとの相対関係が重要視されると予想される．1 段目の重回帰の結果は，各音響イベントごとの評価値として利用できる．

2 段目の重回帰では，1 段目の重回帰で計算した各音響イベントの評価値に対して重回帰を行う．重回帰分析の目的変数としては，例えば生徒に対する手動評価値が利用できる．重回帰の重みパラメータの学習時には，生徒の発音全体のスコアを算出する際に，どの音響イベントを重要視するかが学習される．例えば，日本人の米語発音を評価する場合には，日本人が一般的に苦手にしやすい/r/や/er/などが重要視されると予想される．2 段目の重回帰の結果，生徒のスコアが得られる．

5.4.1　リッジ回帰

2 段階重回帰を用いることで，部分構造化より次元圧縮の自由度が高くなる．しかし，それと同時に過学習により汎化性能の低下し性能が低下するリスクも高くなる．そこで，単純な二乗誤差最小規準による重回帰分析ではなく，二次の正則化項を導入したリッジ回帰を行うことで汎化性能の向上を図る．

通常の最小二乗法による線形回帰分析では，推定すべき回帰係数（重みパラメータ）$\boldsymbol{w}$ の誤差関数 $E(\boldsymbol{w})$ は，N 個ある学習データセットのうち，目的変数に対応する n 番目のデータを t_n，説明変数に対応する n 番目のデータを $\boldsymbol{x}_n$ として，

$$E(\boldsymbol{w}) = \sum_{n=1}^{N} (t_n - \boldsymbol{w}^T \boldsymbol{x}_n)^2 \tag{5.3}$$

となり，誤差関数を最小化する $\boldsymbol{w}$ は，${\boldsymbol{x}_n}^T$ を縦に並べた行列を $\boldsymbol{X}$，t_n を縦に並べたベクトルを $\boldsymbol{t}$ として

$$\underset{\boldsymbol{w}}{\operatorname{argmin}} E(\boldsymbol{w}) = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{t} \tag{5.4}$$

となる．

これに，二次の正則化項を導入したリッジ線形回帰の誤差関数 $\tilde{E}(\boldsymbol{w})$ は，正則化パラメータを λ として

$$\tilde{E}(\boldsymbol{w}) = \sum_{n=1}^{N} (t_n - \boldsymbol{w}^T \boldsymbol{x}_n)^2 + \lambda \boldsymbol{w}^T \boldsymbol{w} \tag{5.5}$$

であり，誤差関数を最小化する $\boldsymbol{w}$ は，

$$\underset{\boldsymbol{w}}{\operatorname{argmin}} \tilde{E}(\boldsymbol{w}) = (\lambda \boldsymbol{I} + \boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{t} \tag{5.6}$$

となる．なお，$\lambda = 0$ のとき最小二乗法による線形回帰と同値になる．

5.4.2　実験的検証

二段階重回帰を用いることによる精度の変化を見る実験を行う．実験条件には，先の部分構造化と同じ条件を用いる．

2 段階重回帰の目的変数には，1 段目・2 段目の重回帰共にこの話者の手動評価値を用いた．本来，1 段目の重回帰では，各音素に対する手動評価値を利用することが望ましいが，ERJ にはそのようなデータは含まれていないため，1 段目においても話者全体を通しての手動評価値を目的変数とした．また，リッジ回帰の正則化パラメータ λ は，1 段目の重回帰に用いる λ_1 と 2 段目の重回帰に用いる λ_2 を二つ設定し，それらを変化させながら実験を行った．

表 5.4: 2 段階重回帰分析と手動評価値との相関の平均

	$\lambda_2 = 0$	$\lambda_2 = 0.01$	$\lambda_2 = 0.1$	$\lambda_2 = 1$	$\lambda_2 = \infty$
$\lambda_1 = 0$	0.39	0.56	0.61	0.62	0.62
$\lambda_1 = 0.01$	0.46	0.60	0.67	0.67	0.67
$\lambda_1 = 0.1$	0.58	0.62	0.68	0.69	0.68
$\lambda_1 = 1$	0.61	0.63	0.67	0.69	0.68
$\lambda_1 = \infty$	0.61	0.63	0.67	0.69	0.68

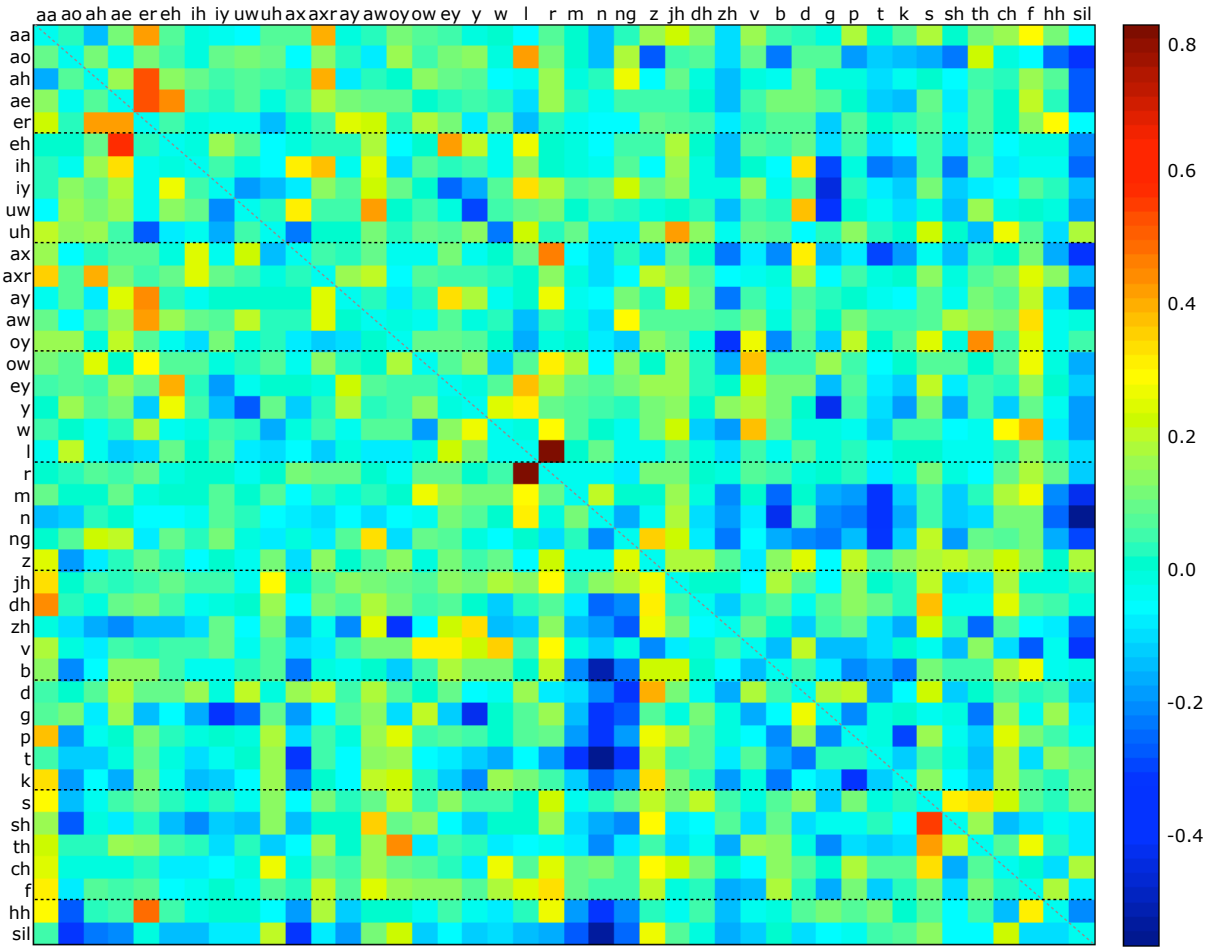

図 5.5: 1 段目の重回帰分析による回帰係数

5.4.3　実験結果

実験結果を表 5.4 に示す．結果，適切な正則化パラメータを選ぶことで，部分構造より大幅に精度が向上していることが分かる．具体的には，$\lambda_1 = 1, \lambda_2 = 1$ のときに相関の平均 0.69 で最大となった．部分構造化において最適な特徴量の数を選択した場合でも，相関の平均の最大は 0.63 であったので，二段階重回帰は部分構造化より高い精度が得られることが分かる．

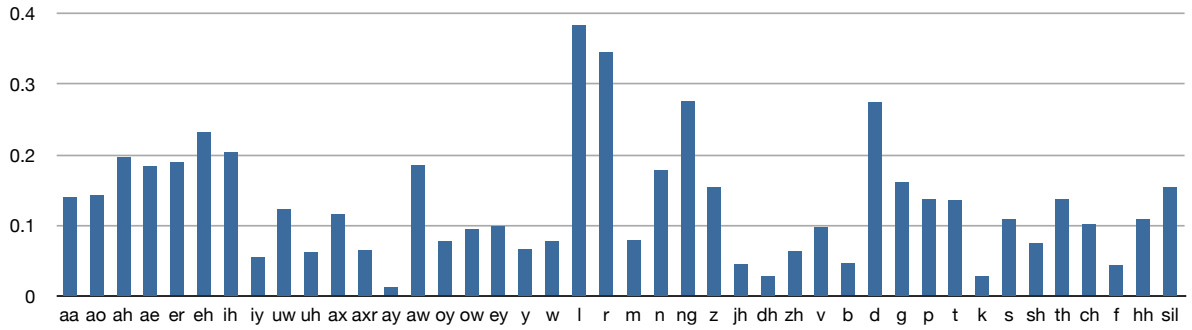

図 5.6: 2 段目の重回帰分析による回帰係数

なお，正則化パラメータは，$\lambda_1 = \infty, \lambda_2 = \infty$ でも相関の平均 0.68 とほぼ最高性能が得られる．このとき，リッジ回帰の解 $\hat{\boldsymbol{w}}$ は，

$$\hat{\boldsymbol{w}} = \boldsymbol{X}^T \boldsymbol{t} \tag{5.7}$$

となり，$\boldsymbol{X}$ の分散共分散行列を単位行列と仮定して $\boldsymbol{X}$ と $\boldsymbol{t}$ の単純な類似度（内積）を $\boldsymbol{w}$ として推定していることになる．

参考に，セット 1 を評価データ，$\lambda_1 = \lambda_2 = 1$ として学習した時に，1 段目の重回帰分析で得られる重みを図 5.5 に，2 段目の重回帰分析で得られる重みを図 5.6 に示す．1 段目の重回帰分析の重みは，行列の各行が，一回の重回帰分析で得られる重み $\boldsymbol{w}$ に相当する．例えば，*AA* を評価するときに日本語の/あ/に似た母音との重みが大きくなるなど，おおよそ直感とあった重みが学習されていることがわかる．2 段目の重回帰分析の重みも，日本人の苦手とする/R/や/L/といった音素の重みが大きくなるなど，おおよそ直感とあった重みが学習されている．

5.5　マルチストリーム化と 3 段階重回帰

次に，マルチストリーム構造化と，2 段階重回帰を拡張した 3 段階重回帰分析を用いたさらなる精度向上法について述べる．

5.5.1　マルチストリーム化

構造表象を用いた音声認識では，構造の不変性に制約を加えるために，ケプストラムを次元の近いブロックごとに分割し，特徴量マルチストリーム化させてマルチストリーム構造を用いていた．構造表象を用いた外国語発音評定では，構造の不変性が強すぎる問題は大きな問題となっていないが，同様の処理は発音評定タスクにも利用することができる．

また，マルチストリーム構造化は，短時間音響特徴量のブロック化に基づくマルチストリーム構造化を行わなくても実現できる．構造表象は，例えば MFCC 空間における分布間 $\sqrt{\mathrm{BD}}$ を計算することにより算出されるが，例えば ΔMFCC 空間における分布間 $\sqrt{\mathrm{BD}}$ を計

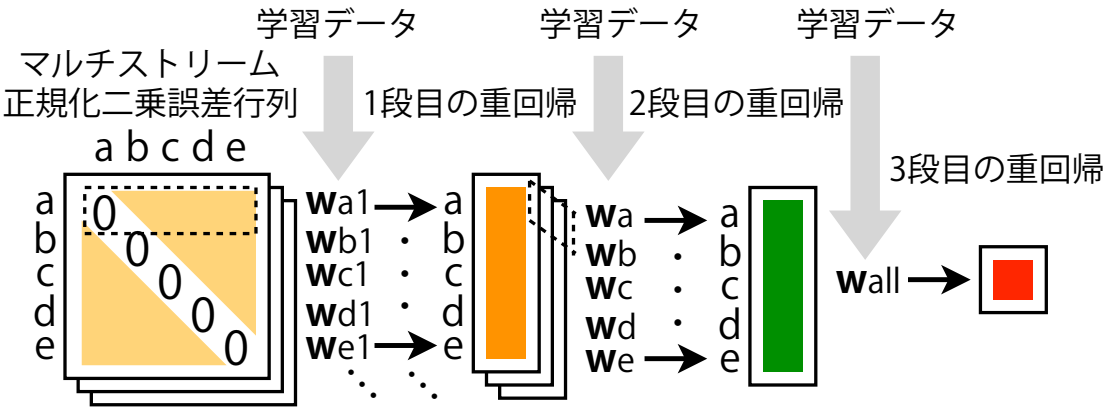

図 5.7: 3 段階の重回帰を用いた外国語発音評価

算することによっても算出することができる．つまり，異なる音響特徴量を用いれば，異なる構造表象を抽出することができる．短時間音響特徴量を変化させる他にも，教師を変えて正規化二乗誤差行列を計算すれば，これもストリームを追加できる．このように，ブロック化により不変性に制約条件を加えるわけではなく，単純に情報量を増やす意味でもマルチストリーム構造化が利用できる．

5.5.2　三段階重回帰

マルチストリーム構造化を用いると，マルチストリーム正規化二乗誤差行列が計算できる．これにより，特徴量の数がストリーム数倍されることになり，当然次元の呪いの問題が発生する．そのため，適切な次元圧縮を利用する必要がある．そこで，マルチストリーム正規化二乗誤差行列に対し，3 段階の重回帰分析を用いて外国語発音評価を行う手法を提案する．3 段階重回帰分析の枠組みを図 5.7 に示す．

1 段目の重回帰では，各正規化二乗誤差行列の行ごとに重回帰を行う．これは 2 段階重回帰の 1 段目の重回帰と同様の処理である．

2 段目の重回帰では，各ストリームから計算された音響イベントのスコアに対して重回帰分析を行う．重回帰分析の目的変数としては，例えば 1 段目と同じ生徒に対する手動評価値が利用できる．重回帰のパラメータの学習時には，ある音響イベントの発音を評価する際に，どの音響特徴量空間の構造を重要視するかが学習される．例えば，16kHz サンプリング音声の MFCC と 6kHz サンプリング音声の MFCC を使ってマルチストリーム構造化した場合，低周波数領域に音素の特徴が多く含まれる母音は 6kHz サンプリングの構造が重要視され，高周波数領域にも音素の特徴が多く含まれる子音は 16kHz サンプリングの構造が重要視されると予想される．2 段目の重回帰の結果，各音響イベントごとの評価値が得られる．

3 段目の重回帰では，2 段目の重回帰で計算した各音響イベントの評価値に対して重回帰

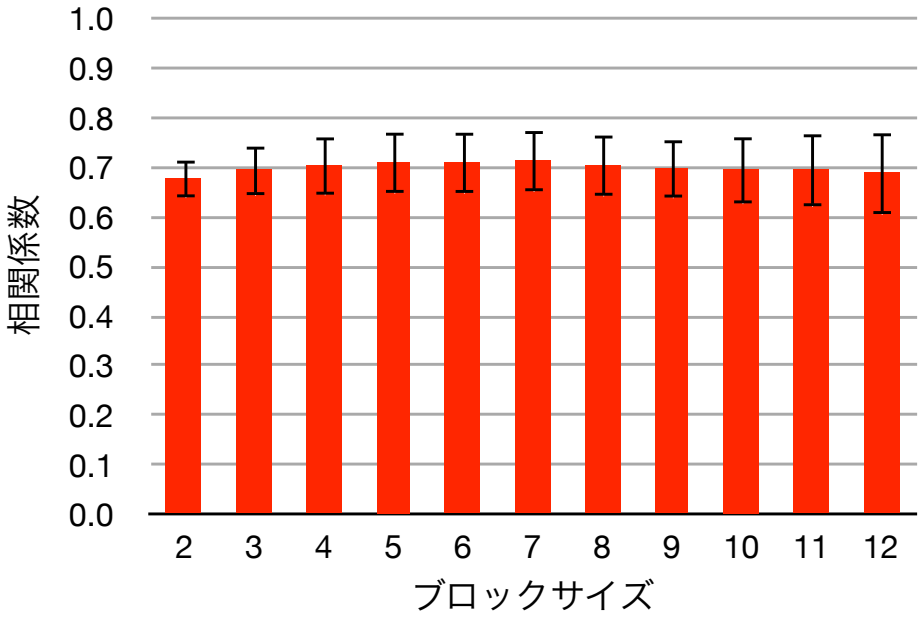

図 5.8: ブロックサイズを変更したときの相関値の変化

分析を行う．これは 2 段階重回帰の 2 段目の重回帰と同様の処理である．

2 段階の重回帰と 3 段階の重回帰を比較すると，マルチストリーム化を行い，2 段目の重回帰によってストリームごとに重みを付けているところが差分であり，この処理により，精度がさらに向上すると考えられる．

5.5.3　実験的検証

マルチストリーム化と三段階重回帰を用いることによる精度の変化を見る実験を行う．実験条件は，先の 2 段階重回帰分析の実験と同じ条件で行った．3 段階重回帰の目的変数には，1 段目・2 段目・3 段目の重回帰共にこの話者の手動評価値を用いた．また，リッジ回帰の正則化パラメータ λ は，1 段目の重回帰に用いる λ_1 と 2 段目の重回帰に用いる λ_2 と 3 段目の重回帰に用いる λ_3 の三種類のパラメータを動かしながら実験を行った．

マルチストリーム化に用いる特徴量としては，まず，構造を用いた音声認識でも利用されている MFCC のブロック化によるマルチストリーム化を試した．次に，母音を特徴付けるフォルマント周波数はおよそ 3 kHz 以下に存在するという音声学的知見から，3 kHz のローパスフィルタをかけた音声の MFCC や，教師として M08 以外に F12 を利用するなど，さまざまな種類の特徴量を用いた実験を行った．

表 5.5: 手動評価値との相関の平均（結果の良い順にソート）

教師:短時間音響特徴量	平均	標準偏差
M08:MFCC+M08:MFCC$_{\mathrm{lowpass}}$+F12:MFCC	0.72	0.08
M08:MFCC+M08:MFCC$_{\mathrm{lowpass}}$+F12:MFCC+F12:MFCC$_{\mathrm{lowpass}}$	0.71	0.08
M08:MFCC+M08:MFCC$_{\mathrm{lowpass}}$	0.71	0.09
M08:MFCC	0.69	0.08
M08:MFCC+M08:ΔMFCC	0.68	0.08
M08:MFCC+M08:Energy	0.62	0.09

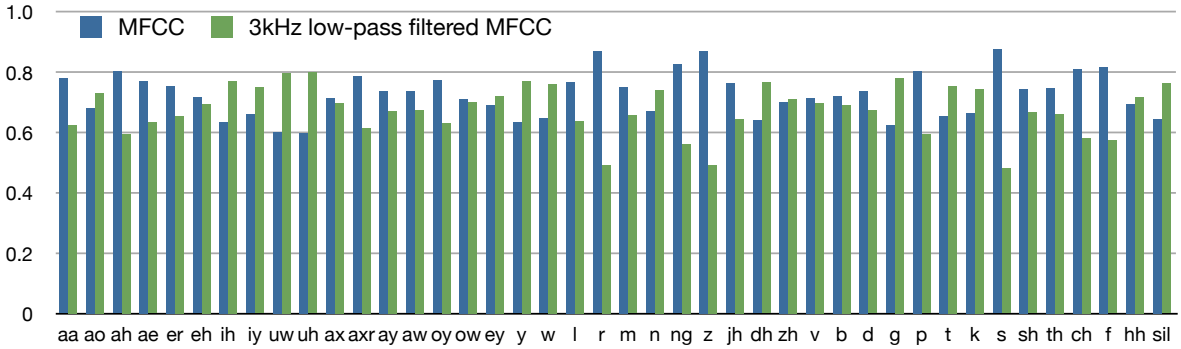

図 5.9: 3 段階重回帰分析の 2 段目の重回帰分析による回帰係数

5.5.4　実験結果

まず，MFCC をブロック化し，ブロックサイズをかえながらマルチストリーム化したときの実験結果を図 5.8 に示す．なおこの実験では，正規化パラメータは $\lambda_1 = \lambda_2 = \lambda_3 = 1$ と設定した．結果，ブロックサイズが 7 のときに，相関の平均は最大で 0.71 となった．なお，ブロックサイズが 12 のときが 2 段階重回帰と同じ条件の実験となっており，その相関値の平均 0.69 と比較して精度は微増している．

次に，3 kHz のローパスフィルタを通した音声の MFCC など異なる短時間音響特徴量を用いたり，教師を変更したりしてマルチストリーム構造化を行い，それを 3 段階重回帰分析を行った結果を示す．いくつかの組み合わせによる相関値を抜粋して表 5.5 に示す．今回の実験タスクでは，MFCC と 3kHz のローパスフィルタを通した MFCC を M08 と比較したものと，MFCC を F12 と比較した 3 ストリームを用いて 3 段階重回帰分析を行った時，相関の平均は最大で 0.72 となった．なお，このとき $\lambda_1 = \lambda_2 = \lambda_3 = 1$ とした．$\lambda_1 = \lambda_2 = \lambda_3 = \infty$ と設定した場合には，相関の平均は 0.70 となり，2 段階重回帰と同様に，3 段階重回帰でも，正則化パラメータを ∞ にしてもほぼ最良の結果が得られている．

参考に，MFCC と 3kHz のローパスフィルタを通した MFCC を M08 と比較した 2 ストリームで 3 段階重回帰を行った時，各ストリームの重みが音素によってどう割り振られているかを図 5.9 に示す．/ih/のような母音など，低域に重要な特徴が含まれている音素はローパスフィルタを通した MFCC の方が重みがつよく，逆に/s/のような無声子音など広

表 5.6: GOP 算出に用いる HMM を学習するための音響分析条件

サンプリング	16bit / 16kHz
窓	25 msec 幅，10 msec シフト
学習データ	ERJ に含まれる 20 名のネイティブスピーカーの音声すべて
話者適応用データ	60 文（set 8 を読み上げた話者のみ 40 文）
特徴量	CMN をかけた MFCC(12) + Δ MFCC(12) + Δ Energy(1)
HMM の種類	monophone，不特定話者音素 HMM
出力確率分布	対角共分散行列を持つ 4 混合 GMM
トポロジー	3 状態の left to right 型
音素の種類	aa,ae,ah,ao,aw,ax,axr,ay,b,ch,d,dh,eh,er,ey,f,g,hh,ih, iy,jh,k,l,m,n,ng,ow,oy,p,r,s,sh,t,th,uh,uw,v,w,y,z,zh 合計 41 種類

域にも重要な特徴が含まれている音素はローパスフィルタをかけていない通常の MFCC の方が重みが強くなっていることがわかる．

5.6　GOP を用いた発音評価との比較

本節では，ここまでに提案した構造表象を用いた発音評価手法と，従来から広く用いられている GOP スコアを用いた発音評価手法の比較実験を行う．

5.6.1　GOP スコアと重回帰分析

GOP スコアは，各音素ごとにスコアが算出される．それを今回の実験タスクで扱っている話者ごとのスコアに変換するためには，構造と同じく，重回帰分析を利用することで，話者のスコアをより精度よく算出できると考えられる．すなわち，2 段階重回帰分析や 3 段階重回帰分析の，最終段階の重回帰分析と同様の重回帰分析を行うことで，音素ごとの GOP スコアから話者のスコアを算出する．

5.6.2　実験条件

HMM を学習するときの音響分析条件を表 5.6 に示す．HMM の学習話者を ERJ に含まれるすべての米語ネイティブ話者 20 名にしていること，特徴量を CMN をかけた MFCC_E_D_Z_N としていることが主な先の実験との差分である．また，HMM は，各学習話者の読み上げ音声すべて（セット 8 以外は 60 文，セット 8 は 40 文）を用いて MLLR 適応による話者適応を施した．

表 5.7: GOP スコアを重回帰分析したときの手動評価値との相関

正則化パラメータ λ	平均	標準偏差
0	0.68	0.10
1	0.68	0.10
100	0.70	0.08
10000	0.71	0.06
∞	0.71	0.07

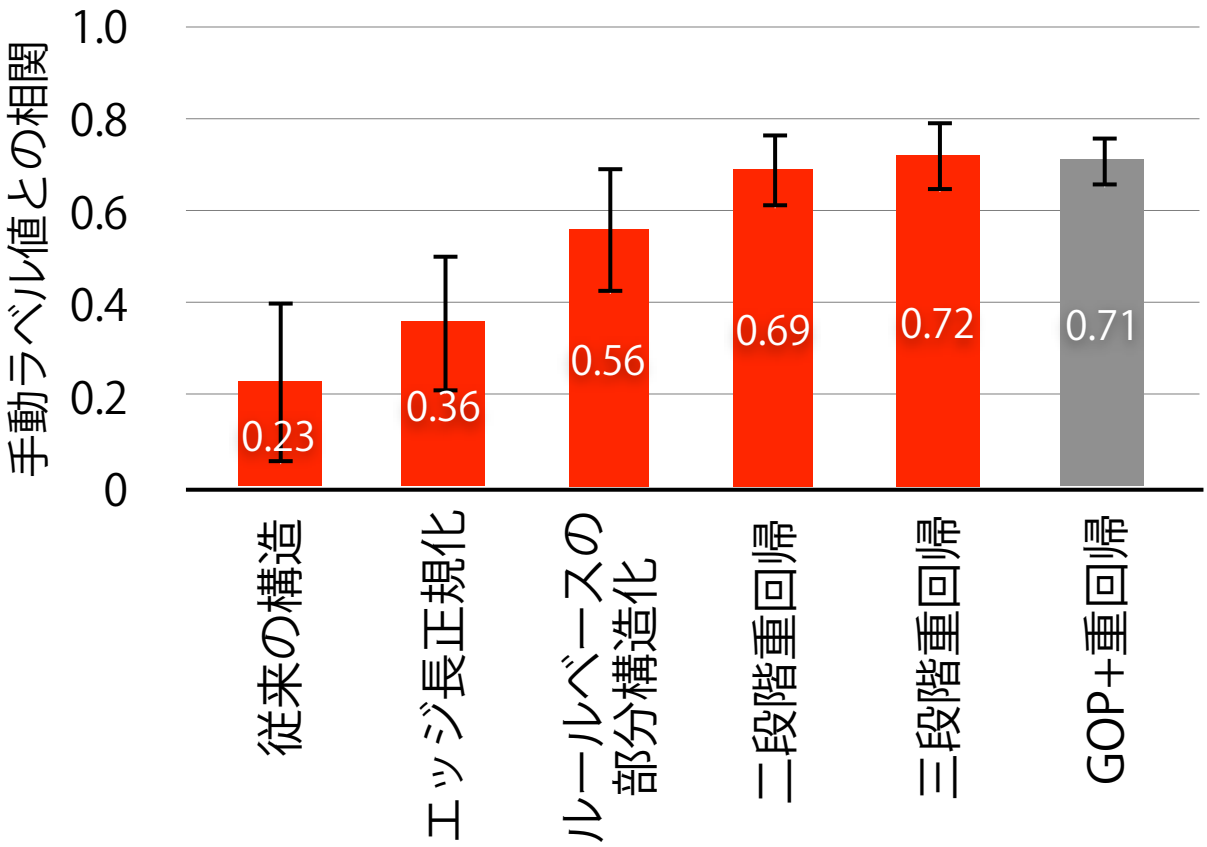

図 5.10: 米語発音自動評価値と手動ラベル値との相関

5.6.3　実験結果

重回帰分析の正則化パラメータ λ を変化させながら相関をだした結果を表5.7 に示す．GOP スコアと手動評価値の相関の平均の最大は，正則化パラメータ $\lambda = 10000$ としたとき，0.71 となった．

本章で行った結果をすべてまとめて図5.10 に示す．本章で提案したエッジ長正規化，マルチストリーム化，三段階重回帰分析を利用して発音評価を行った結果と，GOP スコアと

重回帰分析を用いて発音評価を行った結果は，ほぼ同等の精度を示していることがわかる．

5.7　まとめ

本節では，構造表象を用いた外国語発音分析において，構造間差異計算時における正規化，適切な次元圧縮，マルチストリーム化を用いることにより，大幅に精度を向上させる手法について述べた．話者間差異の正規化は，非常に単純な手法ながら 2 話者を直接比較する際には精度を大幅に向上させる効果がある．部分構造化は，多段階重回帰と比べると精度は低いものの，学習データが得られない状況でも，発音に対する知識を利用することで適切な次元圧縮を実現できる．2 段階重回帰は，学習データが得られたときに適切な自由度と汎化性能で次元圧縮を実現することができる．さらにマルチストリーム化と 3 段階重回帰で，その精度をさらに向上させることができる．

最終的には，ERJ に含まれる話者ごとの子音・母音すべてを含む発音のスコアを評価する実験において，話者間差異の正規化を用い，MFCC と 3kHz のローパスフィルタを通した MFCC を M08 と比較したものと，MFCC を F12 と比較した 3 ストリームをマルチストリーム構造としてを選び，適切な正則化パラメータを設定した三段階重回帰分析を行うことで，手動評価値との相関を 0.23 から 0.72 まで上昇させることができた．この相関値は，教師の音声として 1 名ないしは 2 名の音声しか用いていないにも関わらず，教師の音声として 20 名分を用いている，従来広く用いられている GOP を用いた場合の相関（0.71）とほぼ同等の相関を示している．

第6章

相対量と絶対量を用いた発音分析

6.1　はじめに

第 5 章では，構造表象を用いた発音分析の改善手法について述べ，ERJ を用いて話者ごとの子音・母音すべて含む発音の評価値を推定する実験においては，GOP スコアを用いた手法とほぼ同等の精度を実現した．

ここで，構造表象と GOP スコアは音声の異なる特徴を捉えているものであり，それらは対立するものではない．構造表象は，音と音との相対関係を記述している．一方 GOP スコアは，HMM により，音そのものの絶対的特徴を捉えている．そのため例えば，母音のような話者の影響を強く受ける音素は，構造表象を用いて話者不変な相対的な特徴を見た方が，頑健な発音分析が可能になると考えられる．一方，無声子音のような話者によって大きな変化がない音素では，GOP スコアを用いて音そのものの絶対的な特徴を利用した方がより頑健な発音分析が可能になると考えられる．

そこで，本章では，構造表象を用いた相対量に基づく発音分析と，GOP スコアを用いた絶対量に基づく発音分析を組み合わせた発音分析法について検討する．さらに，相対的特徴を用いた発音分析と絶対的特徴を用いた発音分析の利点と欠点を調べるために，意図的に話者のミスマッチを発生させた場合の発音分析実験も行う．

6.2　3 段階重回帰と GOP スコアの組み合わせ

3 段階重回帰分析を用いれば，構造スコアと GOP スコアを簡単に組み合わせることができる．例えば，GOP スコアは音素ごとに算出されるため，2 段目の重回帰の説明変数として各音素の GOP スコアを追加する方法が考えられる．この方法で 3 段階重回帰に GOP スコアを組み込む概念図を図 6.1 に示す[1]．

今回，3 段階重回帰の 2 段目の重回帰の説明変数に GOP スコアを入れる手法を提案したが，GOP スコアの他にも，さまざまなスコアを重回帰の説明変数に導入してもよい．例えば，MFCC 以外の別の音響特徴量から算出した GOP スコアを利用することなどが考えられる．なお，音素ごとにスコアが得られる場合には 2 段目の重回帰の説明変数に加えればよいし，話者ごとのスコアが得られる場合には 3 段目の重回帰の説明変数に加えればよい．

なお，$\sqrt{\mathrm{BD}}$ と GOP スコアのように，単位の異なる説明変数を用いて正則化ありの重回帰分析する場合は，オーダーの違いに注意する必要がある．通常の重回帰分析では，オーダーの違いは分析結果にまったく影響を与えないが，リッジ回帰のように正則化項を導入する場合には，オーダーの大きい変数には正則化が弱く，オーダーの小さい変数には正則化が強く働くようになってしまう．この問題を解決するためには，オーダーをそろえるために，定数でスコアを割り算することで分散を 1 に正規化する標準化を行えばよい．

[1] 1 段目の重回帰分析と 2 段階目の重回帰分析の順番を入れ替えるなど，今回提案している手法以外も考えられる．ただし，予備実験の結果，今回提案している組み合わせ法がもっとも精度が高くなったため，この手法を用いている

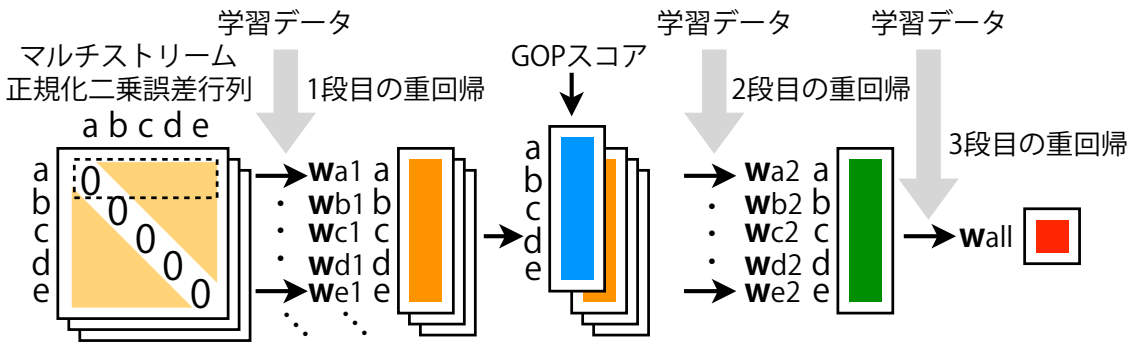

図 6.1: 構造と GOP スコアを用いた 3 段階の重回帰

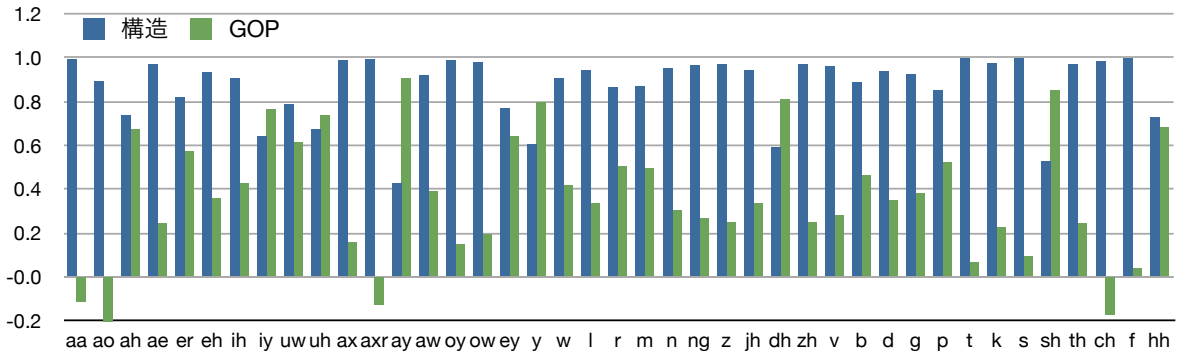

図 6.2: 音素ごとのストリームに対する重み付け

6.2.1　実験的検討

GOP と構造表象と多段階重回帰を用いた外国語発音評価の効果を実験により検証する．実験条件は，第 5 章の実験条件とまったく同じである．

また，構造スコアと GOP スコアのオーダーをそろえるために，1 段目の重回帰が終わった後の構造スコアすべての，学習データの標準偏差と GOP スコアすべての学習データの標準偏差を 1 に正規化した．

6.2.2　実験結果

マルチストリーム構造化と三段階重回帰分析でもっとも精度が高くなった M08 を教師にした MFCC と 3 kHz ローパスをかけた MFCC の構造，F12 を教師にした MFCC の構造を用いたものと GOP スコアをつかって 3 段階重回帰を行ったところ，結果は相関の平均 0.75，標準偏差 0.08 となり，構造のみの 3 段階重回帰や，GOP のみの重回帰の結果よりも高い相関が得られた．なおこのとき，正則化パラメータは $\lambda_1 = 1, \lambda_2 = 0, \lambda_3 = 1$ とした．さらに，GOP スコアを算出するときに，MLLR 適応を行っていない HMM を使って GOP

表 6.1: 評価者間の手動ラベル値の相関

	評価者 1	評価者 2	評価者 3	評価者 4	評価者 5
評価者 1	1	0.78	0.85	0.77	0.76
評価者 2	0.78	1	0.79	0.83	0.71
評価者 3	0.85	0.79	1	0.71	0.73
評価者 4	0.77	0.83	0.71	1	0.80
評価者 5	0.76	0.71	0.73	0.80	1

スコアを算出し，同様に構造と GOP を三段階重回帰で組み合わせたところ，結果は相関の平均 0.76，標準偏差 0.08 となり，さらに高い相関が得られた．

GOP と構造を組み合わせただけで相関が上昇するという結果は，HMM でモデル化される絶対的特徴と構造でモデル化される相対的特徴とが，発音の異なる特徴を捉えており，音素によってそれらを使い分けることで精度が向上したものと考えられる．

参考に，M08 を教師とした MFCC の構造と，MLLR 適応なしの GOP スコアを用い，セット 1 以外の音声を学習データとして 3 段階重回帰分析したときの，音素ごとに構造と GOP の重み付けがどうなるかを図 6.2 に示す．なお，この条件では，相関の平均 0.75，標準偏差 0.08 となる．/aa/や/ao/母音などは構造の重みが，/sh/などの無声子音などには GOP スコアの重みが大きくなっていることがわかる．

ここで，手動評定を行なった 5 名の音声学者の評定値に対し，それぞれの話者間の相関係数を計算すると，表 6.1 のようになる．なお，この評価は 1 名につき 10 文章を用いて行われている．これと今回の結果（0.76）を比較すると，提案手法は上限に近い相関値が得られていることがわかる．すなわち，今回の提案手法により，60 文分の米語文章を読み上げると，10 文読み上げて手動評価した場合と同程度の精度で発音分析が可能になったということができる．

6.3　多様な話者性に対する頑健性の分析

ここまでの実験では，教師の音声としてに成人の米語ネイティブ話者の読み上げ音声を用い，生徒の音声として日本人大学生の音声を利用していた．このようなタスクでは，非言語的特徴のミスマッチは比較的小さく，従来手法でも高い認識性能を示す．一方，構造を用いた発音分析法は，非言語的特徴の違いに理論的に非常に頑健な手法であり，これまでの実験では提案手法の有効性を完全には検証できていない．

そこで本節では，意図的にミスマッチを大きくした実験データを用いた発音評価実験を行い，話者性のミスマッチが大きい場合に精度がどの程度変化するのかについて定量的な実験を行う．

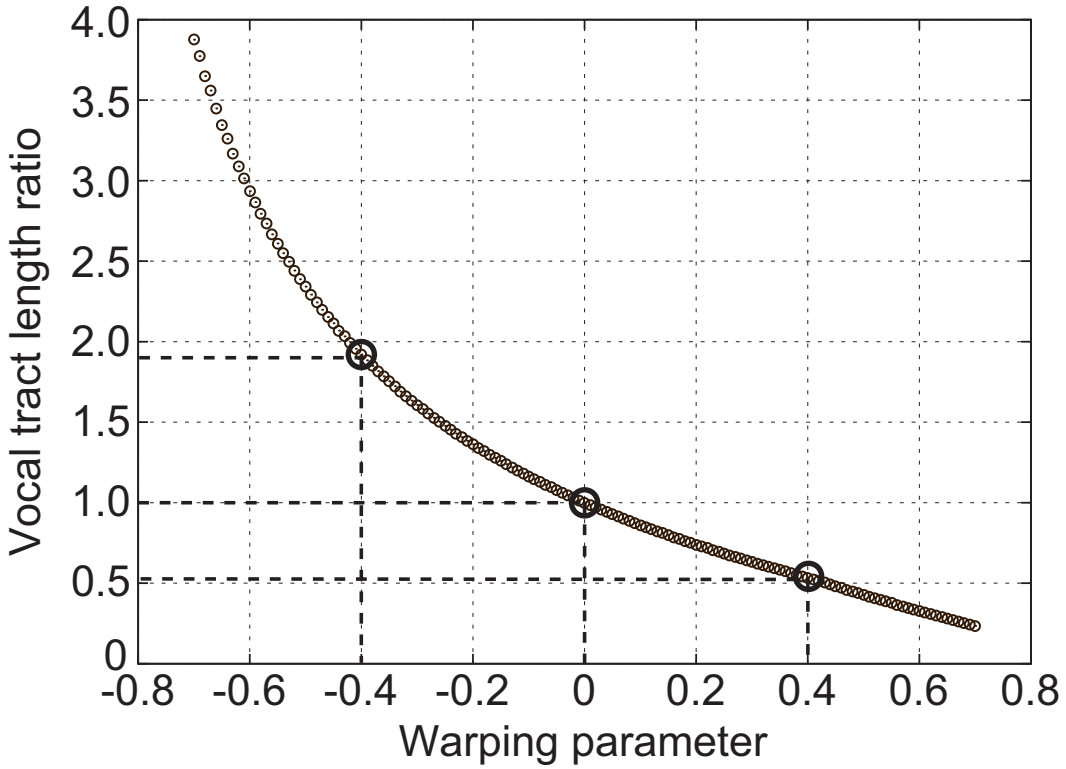

図 6.3: 音素ごとのストリームに対する重み付け

6.3.1　ミスマッチデータの作成

ミスマッチのあるデータの作成のために，ERJ に含まれる生徒の音声を，音声分析変換合成法である STRAIGHT[48] を利用して人工的に声道長を変化させた音声を作成した．声道長を変化させる周波数ウォーピング関数は，[46] で用いられている，周波数ウォーピングを 1 次の全域通過フィルタ

$$\hat{z}^{-1} = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}} \tag{6.1}$$

とした近似を利用した．このとき声道長の変換度合いは，ウォーピングパラメータ α によって調整される．このとき，$|\alpha| < 1$ であり，$\alpha = 0$ のときに変換なし，$\alpha = \pm 0.40$ の時に，声道長が約半分／倍になることに対応している．図 6.3 に，ウォーピングパラメータ α で声道長変換をかけることにより，どの程度声道長変化するかを示す．

6.3.2　発音評価実験

図 6.4 に，人工的に声道長変換をかけた音声をこれまで実験してきた方法で発音評定した場合の相関の平均値の変化を示す．構造+GOP，構造，GOP は，それぞれにおける実験で最も高い性能を示した条件，すなわち構造+GOP については M08 を教師にした MFCC と 3kHz ローパスフィルタをかけた MFCC，F12 を教師にした MFCC のマルチストリーム

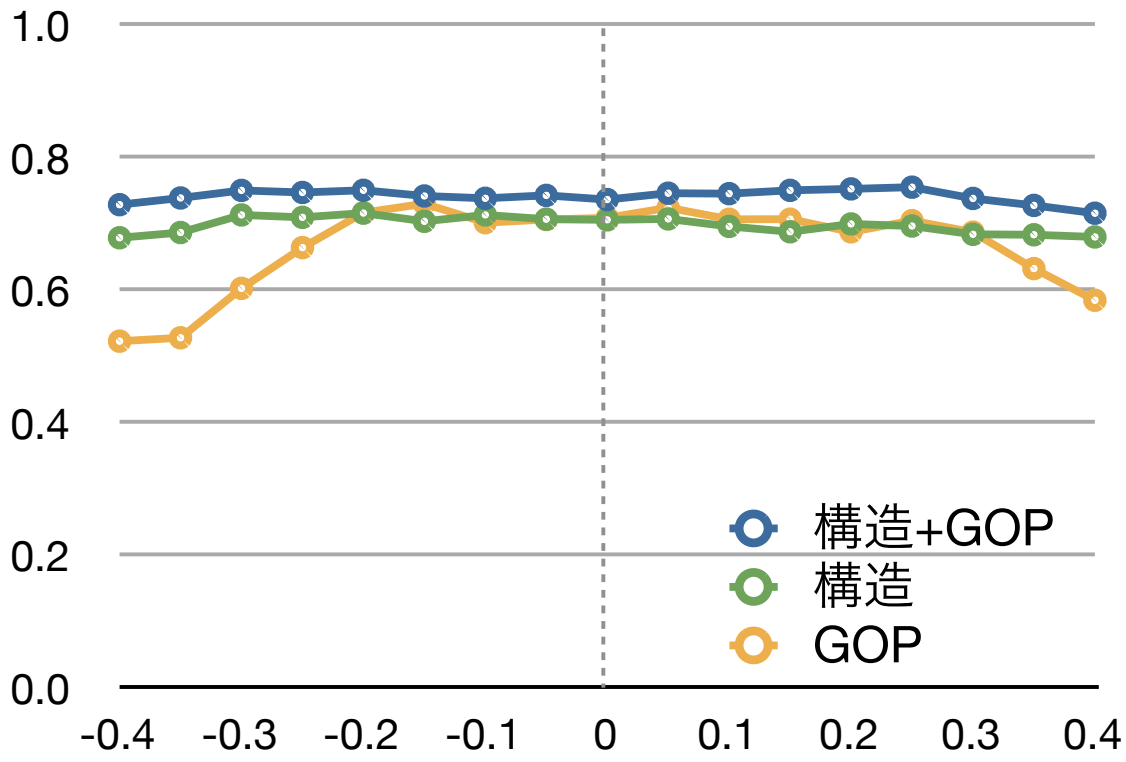

図 6.4: ウォーピングされた音声を利用した場合の相関値

構造と GOP スコアをつかった 3 段階重回帰で $\lambda_1 = 1, \lambda_2 = 0, \lambda_3 = 1$ としたもの，構造については先と同じマルチストリーム構造を用いた 3 段階重回帰で $\lambda_1 = \lambda_2 = \lambda_3 = 1$ としたもの，GOP については重回帰分析で $\lambda = 10000$ としたもの）で実験を行った．

結果，どの手法も，今回用意した人工的なミスマッチに対してはある程度頑健性があることがわかる．構造を用いた結果や構造+GOP の結果は，大きなミスマッチ条件でもほとんど精度は低下しておらず，非常に頑健性の強い分析手法になっているということができる．一方，GOP スコアの結果も，MLLR 適応により頑健性はあるが，ミスマッチが非常に大きい時には精度が低下している．MLLR のミスマッチキャンセル能力は，構造のそれに比べると低いことがわかる．

6.3.3　話者分類実験

次に，$\alpha = 0$ の声道長変換をかけていない音声と，$\alpha = +0.15$ の，およそ大学生の平均身長から小学 5 年生の平均身長へ変換する声道長変換を行った音声を利用して，話者分類を行う実験を行った．ただし，データが多くなりすぎるのを防ぐため，set 1 に含まれる音声と，米語ネイティブスピーカである男性 M08 と女性 F12 のみで実験を行った．

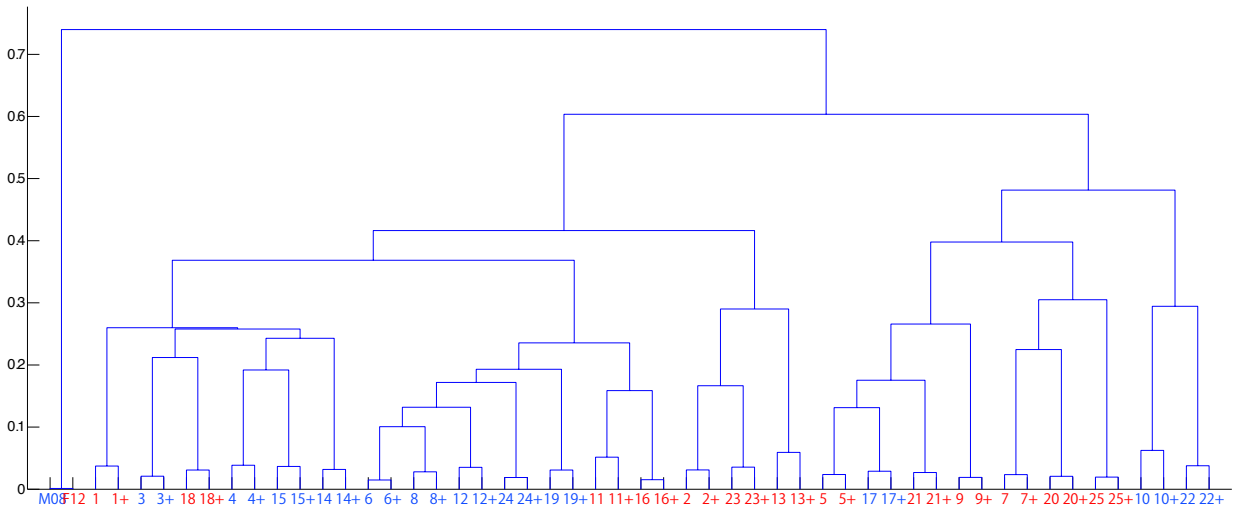

図 6.5: 構造間差異を用いた樹形図による話者分類の可視化

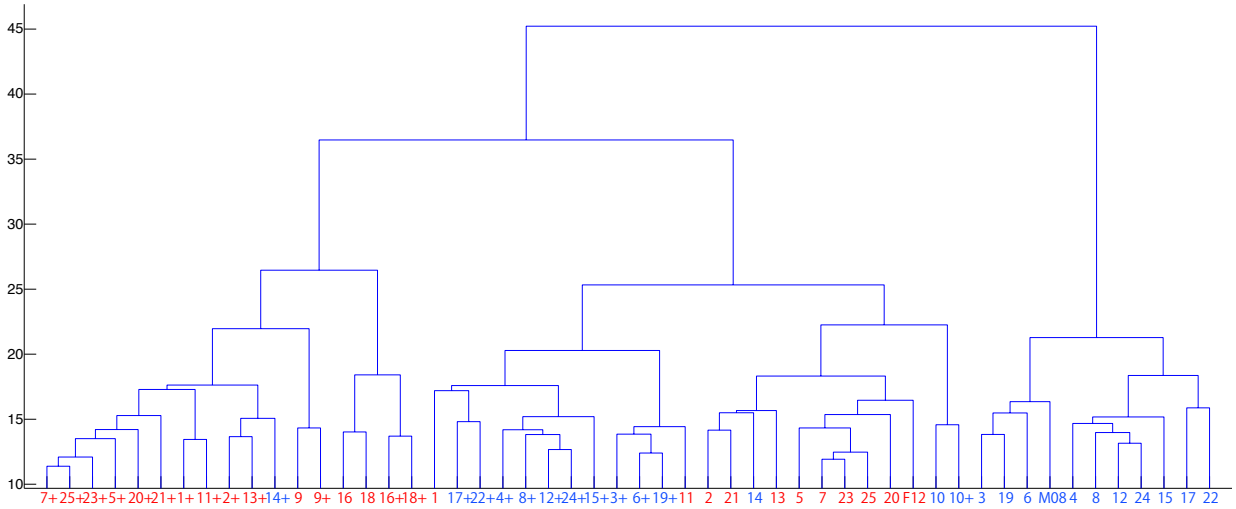

図 6.6: 絶対量の差異を用いた樹形図による話者分類の可視化

話者分類には，2 段階構造間によって構造間差異を計算し，それを元に ward 法によるボトムクラスタリングを行い樹形図によりクラスタを可視化する方法と，Multi Dimensional Scaling（MDS）によって 2 次元平面にプロットする方法の二通りを行った．また，比較実験のために，構造間差異ではなく，対応する 2 話者の音響イベント分布間の $\sqrt{\mathrm{BD}}$ を，GOP の重回帰分析と同じ重みベクトルを用いて重み付けした絶対量に基づく話者間差異を定義し，それによる同様の話者分類実験を行った．

樹形図による可視化の結果を図 6.7 に，MDS による二次元平面への可視化を図 6.8 に示す．なお，数字が話者の ID であり，数字のあとに + がついているものが声道長変換により小学校 5 年生程度の声道長に変換した音声である．また，青が男性，赤が女性を表す．

結果，構造間差異を用いた話者分類では，米語ネイティブである M08 と F12 と，それ以外の日本人話者が大きく分類された結果となっている．また，同じ話者（数字が同じ）で声道長変換をかけるかかけないかの違いにはほとんど影響をうけず，数字が同じものが同

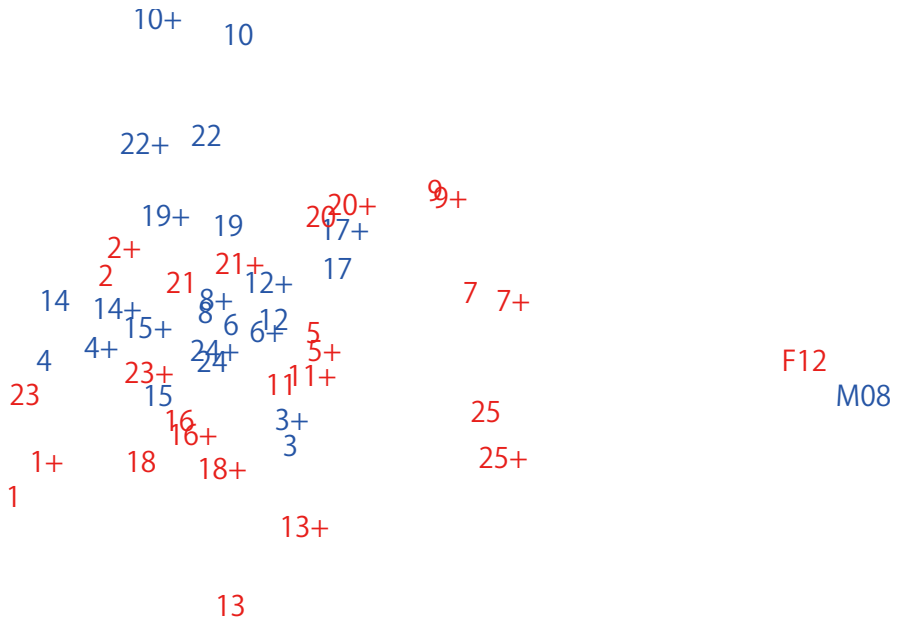

図 6.7: 構造間差異を用いた MDS による話者分類の可視化

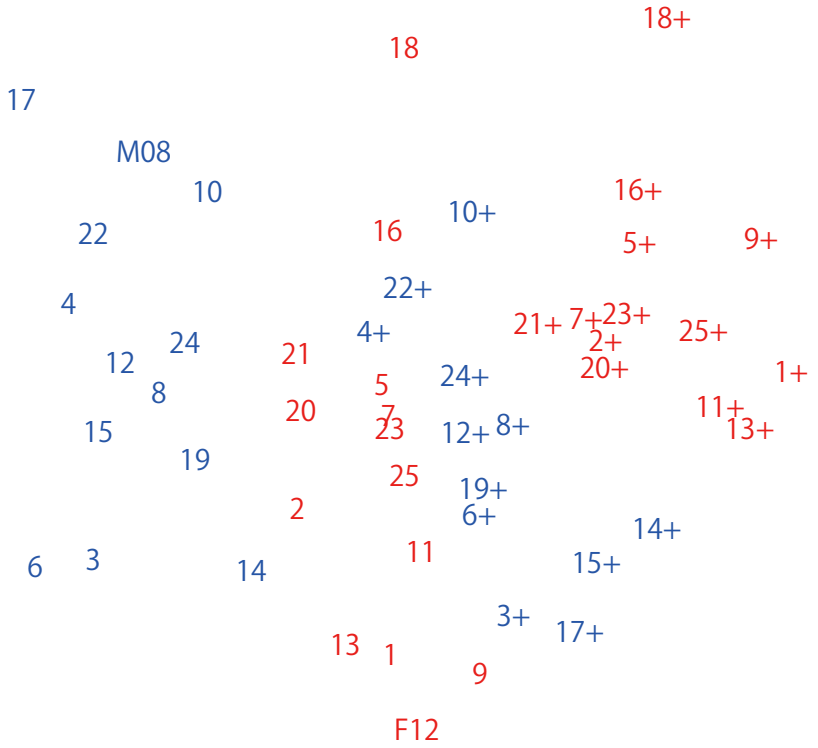

図 6.8: 絶対量の差異を用いた MDS による話者分類の可視化

じクラスタに分類されていることがわかる．すなわち，話者の声道長などの声質ではない，発音の違いなどの要素によってクラスタリングが行われているといえる．

一方，絶対量の差異を用いた話者分類では，男性，女性，声道長変換をかけた男性，声道長変換をかけた女性と，大きく 4 つのクラスタに分類されていることがわかる．また，M08 や F12 は，他の日本人の発音と混ざっている．すなわち，話者の声道長など，声質によってクラスタリングが行われてしまっているといえる．

6.4　まとめ

本章では，構造による相対量と，GOP スコアによる絶対量を組み合わせることで，より精度の高い分析を行えることを示した．また，ミスマッチ条件下の実験の結果，相対量を用いる提案手法は，ミスマッチに非常に頑健な処理が可能であることを示した．

第7章

局所的なアフィン変換不変量

7.1　はじめに

これまで，音声の構造的表象を用いた発音評価を扱ってきた．前章までで，構造を用いた発音評価は単独で GOP スコアと同程度の精度を持ち，さらに，それらを組み合わせることでさらに高い精度を実現できることを実験的に示した．

しかしながら，構造を用いた音声分析には 1 つの大きな欠点がある．構造表象を抽出するためには，音声を分布化せねばならず，そのためには少なくない読み上げ音声データを必要とする．これまで扱ってきたタスクでは，約 60 文程度の読み上げ文章を用いて発音評価を行っていた．

そこで本章では，複数の文章や単語の読み上げを必要としない，音声の相対量に着目した音声分析手法を提案する．具体的には，短時間音響特徴量として，f-divergence と同じような性質をもつアフィン変換不変量を利用する手法を提案する．今回は，これを音声分析に用いるための初期検討として，孤立単語音声認識タスクにおいてその有効性を実験的に確かめる．

7.2　アフィン変換不変性を有する局所特徴量

短時間音響特徴量として，f-divergence と同じような変換不変性をもつ特徴量として局所的アフィン変換不変量（Localized Affine Invariant Feature; LAIF）を提案する．f-divergence は，音声を音響イベント単位で分布化しなければ計算することができないが，LAIF は短時間音響特徴量として，Δ パラメータなどと同じように扱えることができる．

7.2.1　LAIF

まず，ケプストラムの時系列データから LAIF を抽出する計算方法の説明と，f-divergence との関係について示す．

$\boldsymbol{X} = [\boldsymbol{x_1}, \boldsymbol{x_2}, \cdots, \boldsymbol{x_T}]$ を，d 次元のケプストラムベクトル $\boldsymbol{x_t}$ の時系列データとおく．ここで，以下のような $\boldsymbol{x_t}$ に対するアフィン変換について考える．

$$\hat{\boldsymbol{x}}_{\boldsymbol{t}} = \boldsymbol{A}\boldsymbol{x_t} + \boldsymbol{b} \tag{7.1}$$

ここで $\boldsymbol{A}$ は $d \times d$ の正則行列であり，$\boldsymbol{b}$ は d 次元の定ベクトルである．また，アフィン変換後の $\boldsymbol{X}$ を，$\hat{\boldsymbol{X}} = [\hat{\boldsymbol{x}}_1, \hat{\boldsymbol{x}}_2, \cdots, \hat{\boldsymbol{x}}_T]$ で表すことにする．このようなケプストラムに対するアフィン変換は，適切な A や b をおくことにより，話者の違いや録音機器の違いなどを近似することができる．

LAIF は，$\boldsymbol{X}$ のある時間フレーム t 付近における部分ベクトル列 $\boldsymbol{X}_{\boldsymbol{t-k_1:t+k_2}} = [\boldsymbol{x}_{\boldsymbol{t-k_1}}, \cdots, \boldsymbol{x}_{\boldsymbol{t+k_2}}]$ と，それにアフィン変換をかけたもの $\hat{\boldsymbol{X}}_{\boldsymbol{t-k_1:t+k_2}}$ において共通の値を持つ関数，すなわち

$$F(\boldsymbol{X}_{\boldsymbol{t-k_1:t+k_2}}, t) = F(\hat{\boldsymbol{X}}_{\boldsymbol{t-k_1:t+k_2}}, t) \tag{7.2}$$

が成り立つような F のことである．

$F(\boldsymbol{X_{t-k_1:t+k_2}})$ の計算には，t より k_1 だけ前のフレームから k_2 だけ後のフレームまでのケプストラムデータを用いる．データの部分ベクトル列から新たな特徴量を計算するという考え方は，デルタ特徴量抽出の考え方と同じである．なお，デルタ特徴量 $\boldsymbol{\Delta}(\boldsymbol{X_{t-k_1:t+k_2}})$ は，$k = k_1 = k_2$ として以下の式で抽出される特徴量である．

$$\boldsymbol{\Delta}(\boldsymbol{X_{t-k:t+k}}, t) = \frac{\sum_{\tau=1}^{k} \tau \left(\boldsymbol{x_{t+\tau}} - \boldsymbol{x_{t-\tau}}\right)}{2\sum_{\tau=1}^{k} \tau^2} \tag{7.3}$$

デルタ特徴量はケプストラムの回帰係数に相当する特徴量であり，音声の動的な特性を表現するのに有用である．しかし，通常 $\boldsymbol{\Delta}(\boldsymbol{X_{t-k:t+k}}) \neq \boldsymbol{\Delta}(\hat{\boldsymbol{X}}_{\boldsymbol{t-k:t+k}})$ となるため，デルタ特徴量は LAIF ではない．

ここで，以下に示す $F(\boldsymbol{X_{t-k_1:t+k_2}})$ はアフィン変換に不変な LAIF である．

$$F(\boldsymbol{X_{t-k_1:t+k_2}}, t) \;\; = \sqrt{(\boldsymbol{\mu_a} - \boldsymbol{\mu_b})^T (\boldsymbol{\Sigma_a} + \boldsymbol{\Sigma_b})^{-1} (\boldsymbol{\mu_b} - \boldsymbol{\mu_a})} \tag{7.4}$$

ただし，$\boldsymbol{\mu}$ は平均ベクトルを，$\boldsymbol{\Sigma}$ は分散共分散行列を表す．添字 a はフレーム t より前の部分ベクトル列 $[t-k_1, \cdots, t-1]$ を，添字 b はフレーム t 以後の部分ベクトル列 $[t, \cdots, t+k_2]$ を表す．すなわち，$\boldsymbol{\mu_a}$ および $\boldsymbol{\Sigma_a}$ は以下のように推定できる．

$$\boldsymbol{\mu_a} \;\; = \;\; \frac{1}{k_1} \sum_{\tau=t-k_1}^{t-1} \boldsymbol{x_\tau} \tag{7.5}$$

$$\boldsymbol{\Sigma_b} \;\; = \;\; \frac{1}{k_2+1} \sum_{\tau=t}^{t+k_2} (\boldsymbol{x_\tau} - \boldsymbol{\mu_b})(\boldsymbol{x_\tau} - \boldsymbol{\mu_b})^T \tag{7.6}$$

この式は平均や共分散行列の ML 推定値を与える．$\boldsymbol{\mu_b}$ および $\boldsymbol{\Sigma_b}$ も同様に推定できる．

ここで，LAIF と f-divergence の関係性について述べる．2 つのガウス分布 $\mathcal{N}_a(\boldsymbol{\mu_a}, \boldsymbol{\Sigma_a}), \mathcal{N}_b(\boldsymbol{\mu_b}, \boldsymbol{\Sigma_b})$ 間の，BD はガウス分布の平均と分散共分散行列を用いて以下のようにかける．

$$\mathrm{BD}\left(\mathcal{N}_a, \mathcal{N}_b\right) = \frac{1}{8} \left(\boldsymbol{\mu_a} - \boldsymbol{\mu_b}\right)^T \left(\frac{\boldsymbol{\Sigma_a} + \boldsymbol{\Sigma_b}}{2}\right)^{-1} \left(\boldsymbol{\mu_a} - \boldsymbol{\mu_b}\right) + \frac{1}{2} \log \frac{\left|\left(\boldsymbol{\Sigma_a} + \boldsymbol{\Sigma_b}\right)/2\right|}{\left|\boldsymbol{\Sigma_a}\right|^{\frac{1}{2}} \left|\boldsymbol{\Sigma_b}\right|^{\frac{1}{2}}} \tag{7.7}$$

先に導入した LAIF は，ガウス分布間の BD の閉形式における第一項を定数倍してルートをとったものとなっている．なお，BD の第二項も，LAIF となる．

BD の閉形式の第二項のように，(7.4) 以外にも，LAIF は数多く存在する．例えば，以

下はすべてアフィン変換に不変な LAIF である [49].

$$
\begin{aligned}
F_{T1}(\boldsymbol{X},t) &= (\boldsymbol{\mu_t^b}-\boldsymbol{\mu_t^a})^T(\boldsymbol{\Sigma_t^b})^{-1}(\boldsymbol{\mu_t^b}-\boldsymbol{\mu_t^a}) && (7.8)\\
F_{T2}(\boldsymbol{X},t) &= (\boldsymbol{\mu_t^b}-\boldsymbol{\mu_t^a})^T(\boldsymbol{\Sigma_t^a})^{-1}(\boldsymbol{\mu_t^b}-\boldsymbol{\mu_t^a}) && (7.9)\\
F_{T3}(\boldsymbol{X},t) &= (\boldsymbol{\mu_t^b}-\boldsymbol{\mu_t^a})^T(\boldsymbol{\Sigma_t^b}+\boldsymbol{\Sigma_t^a})^{-1}(\boldsymbol{\mu_t^b}-\boldsymbol{\mu_t^a}) && (7.10)\\
F_{T4}(\boldsymbol{X},t) &= \mathrm{Trace}\left((\boldsymbol{\Sigma_t^a})^{-1}\boldsymbol{\Sigma_t^b}\right) && (7.11)\\
F_{T5}(\boldsymbol{X},t) &= \mathrm{Trace}\left((\boldsymbol{\Sigma_t^b})^{-\frac{1}{2}}\boldsymbol{\Sigma_t^a}(\boldsymbol{\Sigma_t^b})^{-\frac{1}{2}}\right) && (7.12)\\
F_{T6}(\boldsymbol{X},t) &= \frac{|\boldsymbol{\Sigma_t^a}|}{|\boldsymbol{\Sigma_t^b}|} && (7.13)\\
F_{T7}(\boldsymbol{X},t) &= \frac{|\boldsymbol{\Sigma_t^a}|}{|\boldsymbol{\Sigma_t^a}+\boldsymbol{\Sigma_t^b}|} && (7.14)
\end{aligned}
$$

ただし，Trace は，行列の対角要素の和をとる演算を，$|\cdot|$ は行列式をとる演算を表す．先に示した LAIF は，F_{T3} のルートをとったものである．

さらに，これらの，重み付けバージョンが存在する．$W^b=\{w_j\}_{j=i-k_1}^{i}$ を，$\boldsymbol{X^b}$ に対する非負の重み付け値と置く．同様に，$W^a=\{w_j\}_{j=i+1}^{i+k_2}$ を，$\boldsymbol{X^a}$ に対する非負の重み付け値と置く．ここで，W^b と W^a は異なっていてもよい．ただし，正規化のために $\sum_j w_j=1$ とする．ここで $\boldsymbol{W_b}$ を，対角要素が W^b で $(k_1+1)\times(k_1+1)$ の対角行列，$\boldsymbol{W_a}$ を，対角要素が W^a で $k_2\times k_2$ の対角行列と置くと，重み付け後の平均値 $\boldsymbol{\mu_i^{W_b}}$ と分散共分散行列 $\boldsymbol{\Sigma_i^{W_b}}$ は以下のようになる．

$$
\begin{aligned}
\boldsymbol{\mu_t^{W_b}} &= \sum_{j=i-k_1}^{i} w_j^b\boldsymbol{x_j} && (7.15)\\
\boldsymbol{\Sigma_t^{W_b}} &= \sum_{j=i-k_1}^{i} w_j^b(\boldsymbol{x_j}-\boldsymbol{\mu_i^{W_b}})(\boldsymbol{x_j}-\boldsymbol{\mu_i^{W_b}})^T && (7.16)
\end{aligned}
$$

$\boldsymbol{\mu_i^{W_a}}$ と $\boldsymbol{\Sigma_i^{W_a}}$ も同様である．これらを用いて，重み付けバージョンの LAIF は以下のように書ける．

$$
\begin{aligned}
F_{T1}^W(\boldsymbol{X},t) &= (\boldsymbol{\mu_t^{W_b}}-\boldsymbol{\mu_t^{W_a}})^T(\boldsymbol{\Sigma_t^{W_b}})^{-1}(\boldsymbol{\mu_t^{W_b}}-\boldsymbol{\mu_t^{W_a}}) && (7.17)\\
F_{T2}^W(\boldsymbol{X},t) &= (\boldsymbol{\mu_t^{W_b}}-\boldsymbol{\mu_t^{W_a}})^T(\boldsymbol{\Sigma_t^{W_a}})^{-1}(\boldsymbol{\mu_t^{W_b}}-\boldsymbol{\mu_t^{W_a}}) && (7.18)\\
F_{T3}^W(\boldsymbol{X},t) &= (\boldsymbol{\mu_t^{W_b}}-\boldsymbol{\mu_t^{W_a}})^T(\boldsymbol{\Sigma_t^{W_b}}+\boldsymbol{\Sigma_t^{W_a}})^{-1}(\boldsymbol{\mu_t^{W_b}}-\boldsymbol{\mu_t^{W_a}}) && (7.19)\\
F_{T4}^W(\boldsymbol{X},t) &= \mathrm{Trace}\left((\boldsymbol{\Sigma_t^{W_a}})^{-1}\boldsymbol{\Sigma_t^{W_b}}\right) && (7.20)\\
F_{T5}^W(\boldsymbol{X},t) &= \mathrm{Trace}\left((\boldsymbol{\Sigma_t^{W_b}})^{-\frac{1}{2}}\boldsymbol{\Sigma_t^{W_a}}(\boldsymbol{\Sigma_t^{W_b}})^{-\frac{1}{2}}\right) && (7.21)\\
F_{T6}^W(\boldsymbol{X},t) &= \frac{|\boldsymbol{\Sigma_t^{W_a}}|}{|\boldsymbol{\Sigma_t^{W_b}}|} && (7.22)\\
F_{T7}^W(\boldsymbol{X},t) &= \frac{|\boldsymbol{\Sigma_t^{W_a}}|}{|\boldsymbol{\Sigma_t^{W_a}}+\boldsymbol{\Sigma_t^{W_b}}|} && (7.23)
\end{aligned}
$$

7.2.2　LAIF のアフィン変換不変性の証明

ここまでに示した LAIF のアフィン変換不変性を証明する．重みありバージョンの LAIF のアフィン変換不変性を示せば，重みなしバージョンの LAIF のアフィン変換不変性も同時に示せるので，ここでは重みありバージョンの LAIF を扱う．アフィン変換であることを証明するには，$F_{Tk}(\boldsymbol{X},t) = M_{Tk}(\hat{\boldsymbol{X}},t)$ となることを示せばよい．

重み付き平均ベクトル $\boldsymbol{\mu}_t^{W_a}$ や重み付き分散共分散行列 $\boldsymbol{\Sigma}_t^{W_a}$ を用いて，アフィン変換後の平均ベクトル $\hat{\boldsymbol{\mu}}_{\boldsymbol{t}}^{\boldsymbol{W_a}}$ と分散共分散行列 $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{t}}^{\boldsymbol{W_a}}$ を表すと，

$$\begin{aligned} \hat{\boldsymbol{\mu}}_{\boldsymbol{t}}^{\boldsymbol{W_a}} &= \boldsymbol{A}\boldsymbol{\mu}_{\boldsymbol{t}}^{\boldsymbol{W_a}} + \boldsymbol{b} && (7.24)\\ \hat{\boldsymbol{\Sigma}}_{\boldsymbol{t}}^{\boldsymbol{W_a}} &= \boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{t}}^{\boldsymbol{W_a}}\boldsymbol{A}^T && (7.25) \end{aligned}$$

となる．これを利用して，LAIF のアフィン変換不変性を証明する．

まず，F_{T1} の場合について証明する．

$$\begin{aligned} F_{T1}(\hat{\boldsymbol{X}},t) &= (\hat{\boldsymbol{\mu}}_{\boldsymbol{t}}^{\boldsymbol{W_b}} - \hat{\boldsymbol{\mu}}_{\boldsymbol{t}}^{\boldsymbol{W_a}})^T(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{t}}^{\boldsymbol{W_b}})^{-1}(\hat{\boldsymbol{\mu}}_{\boldsymbol{t}}^{\boldsymbol{W_b}} - \hat{\boldsymbol{\mu}}_{\boldsymbol{t}}^{\boldsymbol{W_a}})\\ &= (\boldsymbol{A}\boldsymbol{\mu}_{\boldsymbol{i}}^{\boldsymbol{b}} - \boldsymbol{A}\boldsymbol{\mu}_{\boldsymbol{i}}^{\boldsymbol{a}})^T(\boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{i}}^{\boldsymbol{b}}\boldsymbol{A}^T)^{-1}(\boldsymbol{A}\boldsymbol{\mu}_{\boldsymbol{i}}^{\boldsymbol{b}} - \boldsymbol{A}\boldsymbol{\mu}_{\boldsymbol{i}}^{\boldsymbol{a}})\\ &= (\boldsymbol{\mu}_{\boldsymbol{i}}^{\boldsymbol{b}} - \boldsymbol{\mu}_{\boldsymbol{i}}^{\boldsymbol{a}})^T\boldsymbol{A}^T(\boldsymbol{A}^T)^{-1}(\boldsymbol{\Sigma}_{\boldsymbol{i}}^{\boldsymbol{b}})^{-1}\boldsymbol{A}^{-1}\boldsymbol{A}(\boldsymbol{\mu}_{\boldsymbol{i}}^{\boldsymbol{b}} - \boldsymbol{\mu}_{\boldsymbol{i}}^{\boldsymbol{a}})\\ &= F_{T1}(\boldsymbol{X},i). \quad \square && (7.26) \end{aligned}$$

同様に F_{T2}, F_{T3} の場合も証明できる．

$$\begin{aligned} F_{T2}(\hat{\boldsymbol{X}},t) &= (\hat{\boldsymbol{\mu}}_{\boldsymbol{t}}^{\boldsymbol{W_b}} - \hat{\boldsymbol{\mu}}_{\boldsymbol{t}}^{\boldsymbol{W_a}})^T(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{t}}^{\boldsymbol{W_a}})^{-1}(\hat{\boldsymbol{\mu}}_{\boldsymbol{t}}^{\boldsymbol{W_b}} - \hat{\boldsymbol{\mu}}_{\boldsymbol{t}}^{\boldsymbol{W_a}})\\ &= (\boldsymbol{A}\boldsymbol{\mu}_{\boldsymbol{t}}^{\boldsymbol{W_b}} - \boldsymbol{A}\boldsymbol{\mu}_{\boldsymbol{t}}^{\boldsymbol{W_a}})^T(\boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{t}}^{\boldsymbol{W_a}}\boldsymbol{A}^T)^{-1}(\boldsymbol{A}\boldsymbol{\mu}_{\boldsymbol{t}}^{\boldsymbol{W_b}} - \boldsymbol{A}\boldsymbol{\mu}_{\boldsymbol{t}}^{\boldsymbol{W_a}})\\ &= (\boldsymbol{\mu}_{\boldsymbol{t}}^{\boldsymbol{W_b}} - \boldsymbol{\mu}_{\boldsymbol{t}}^{\boldsymbol{W_a}})^T\boldsymbol{A}^T(\boldsymbol{A}^T)^{-1}(\boldsymbol{\Sigma}_{\boldsymbol{t}}^{\boldsymbol{W_a}})^{-1}\boldsymbol{A}^{-1}\boldsymbol{A}(\boldsymbol{\mu}_{\boldsymbol{t}}^{\boldsymbol{W_b}} - \boldsymbol{\mu}_{\boldsymbol{t}}^{\boldsymbol{W_a}})\\ &= F_{T2}(\boldsymbol{X},t) \quad \square && (7.27) \end{aligned}$$

$$\begin{aligned} F_{T3}(\hat{\boldsymbol{X}},t) &= (\hat{\boldsymbol{\mu}}_{\boldsymbol{t}}^{\boldsymbol{W_b}} - \hat{\boldsymbol{\mu}}_{\boldsymbol{t}}^{\boldsymbol{W_a}})^T(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{t}}^{\boldsymbol{W_b}} + \hat{\boldsymbol{\Sigma}}_{\boldsymbol{t}}^{\boldsymbol{W_a}})^{-1}(\hat{\boldsymbol{\mu}}_{\boldsymbol{t}}^{\boldsymbol{W_b}} - \hat{\boldsymbol{\mu}}_{\boldsymbol{t}}^{\boldsymbol{W_a}})\\ &= (\boldsymbol{A}\boldsymbol{\mu}_{\boldsymbol{t}}^{\boldsymbol{W_b}} - \boldsymbol{A}\boldsymbol{\mu}_{\boldsymbol{t}}^{\boldsymbol{W_a}})^T(\boldsymbol{A}(\boldsymbol{\Sigma}_{\boldsymbol{t}}^{\boldsymbol{W_b}} + \boldsymbol{\Sigma}_{\boldsymbol{t}}^{\boldsymbol{W_a}})\boldsymbol{A}^T)^{-1}(\boldsymbol{A}\boldsymbol{\mu}_{\boldsymbol{t}}^{\boldsymbol{W_b}} - \boldsymbol{A}\boldsymbol{\mu}_{\boldsymbol{t}}^{\boldsymbol{W_a}})\\ &= (\boldsymbol{\mu}_{\boldsymbol{t}}^{\boldsymbol{W_b}} - \boldsymbol{\mu}_{\boldsymbol{t}}^{\boldsymbol{W_a}})^T\boldsymbol{A}^T(\boldsymbol{A}^T)^{-1}(\boldsymbol{\Sigma}_{\boldsymbol{t}}^{\boldsymbol{W_b}} + \boldsymbol{\Sigma}_{\boldsymbol{t}}^{\boldsymbol{W_a}})^{-1}\boldsymbol{A}^{-1}\boldsymbol{A}(\boldsymbol{\mu}_{\boldsymbol{t}}^{\boldsymbol{W_b}} - \boldsymbol{\mu}_{\boldsymbol{t}}^{\boldsymbol{W_a}})\\ &= F_{T3}(\boldsymbol{X},t) \quad \square && (7.28) \end{aligned}$$

次に，F_{T4} の場合を証明する．$\mathrm{Trace}\,(\boldsymbol{X}^{-1}\boldsymbol{A}\boldsymbol{X}) = \mathrm{Trace}(\boldsymbol{X})$ という性質に注目すると，

$$\begin{aligned} F_{T4}(\hat{\boldsymbol{X}},t) &= \mathrm{Trace}\left((\hat{\boldsymbol{\Sigma}}_{\boldsymbol{t}}^{\boldsymbol{W_a}})^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{t}}^{\boldsymbol{W_b}}\right)\\ &= \mathrm{Trace}\left((\boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{t}}^{\boldsymbol{W_a}}\boldsymbol{A}^T)^{-1}\boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{t}}^{\boldsymbol{W_b}}\boldsymbol{A}^T\right)\\ &= \mathrm{Trace}\left((\boldsymbol{A}^T)^{-1}(\boldsymbol{\Sigma}_{\boldsymbol{t}}^{\boldsymbol{W_a}})^{-1}\boldsymbol{A}^{-1}\boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{t}}^{\boldsymbol{W_b}}\boldsymbol{A}^T\right)\\ &= F_{T4}(\boldsymbol{X},t). \quad \square && (7.29) \end{aligned}$$

次に，F_{T5} の場合を証明する．Trace $(\boldsymbol{Y}^{-1}\boldsymbol{A}\boldsymbol{Z}^{-1}\boldsymbol{B}\boldsymbol{Y}\boldsymbol{A}\boldsymbol{Z}) = \mathrm{Trace}(\boldsymbol{A}\boldsymbol{B}\boldsymbol{A})$ という性質に注目すると，

$$\begin{aligned} F_{T5}(\hat{\boldsymbol{X}}, t) &= \mathrm{Trace}\left((\hat{\boldsymbol{\Sigma}}_{\boldsymbol{t}}^{\boldsymbol{W_b}})^{-\frac{1}{2}}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{t}}^{\boldsymbol{W_a}}(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{t}}^{\boldsymbol{W_b}})^{-\frac{1}{2}}\right) \\ &= \mathrm{Trace}\left((\boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{t}}^{\boldsymbol{W_b}}\boldsymbol{A}^T)^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{t}}^{\boldsymbol{W_a}}\boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{t}}^{\boldsymbol{W_b}}\boldsymbol{A}^T)^{-\frac{1}{2}}\right) \\ &= \mathrm{Trace}\left((\boldsymbol{A}^T)^{-\frac{1}{2}}(\boldsymbol{\Sigma}_{\boldsymbol{t}}^{\boldsymbol{W_b}})^{-\frac{1}{2}}\boldsymbol{A}^{\frac{1}{2}}\boldsymbol{\Sigma}_{\boldsymbol{t}}^{\boldsymbol{W_a}}(\boldsymbol{A}^T)^{\frac{1}{2}}(\boldsymbol{\Sigma}_{\boldsymbol{t}}^{\boldsymbol{W_b}})^{-\frac{1}{2}}\boldsymbol{A}^{-\frac{1}{2}}\right) \\ &= F_{T5}(\boldsymbol{X}, t). \quad \square \end{aligned} \tag{7.30}$$

次に，F_{T6} の場合について証明する．

$$\begin{aligned} F_{T6}(\boldsymbol{X'}, i) &= \frac{|\hat{\boldsymbol{\Sigma}}_{\boldsymbol{t}}^{\boldsymbol{W_a}}|}{|\hat{\boldsymbol{\Sigma}}_{\boldsymbol{t}}^{\boldsymbol{W_b}}|} \\ &= \frac{|\boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{t}}^{\boldsymbol{W_a}}\boldsymbol{A}^T|}{|\boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{t}}^{\boldsymbol{W_b}}\boldsymbol{A}^T|} \\ &= F_{T6}(\boldsymbol{X}, t) \quad \square \end{aligned} \tag{7.31}$$

同様に，F_{T7} の場合についても証明できる．

$$\begin{aligned} F_{T7}(\hat{\boldsymbol{X}}, t) &= \frac{|\hat{\boldsymbol{\Sigma}}_{\boldsymbol{t}}^{\boldsymbol{W_a}}|}{|\hat{\boldsymbol{\Sigma}}_{\boldsymbol{t}}^{\boldsymbol{W_a}} + \hat{\boldsymbol{\Sigma}}_{\boldsymbol{t}}^{\boldsymbol{W_b}}|} \\ &= \frac{|\boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{t}}^{\boldsymbol{W_a}}\boldsymbol{A}^T|}{|\boldsymbol{A}(\boldsymbol{\Sigma}_{\boldsymbol{t}}^{\boldsymbol{W_a}} + \boldsymbol{\Sigma}_{\boldsymbol{t}}^{\boldsymbol{W_b}})\boldsymbol{A}^T|} \\ &= F_{T7}(\boldsymbol{X}, t) \quad \square \end{aligned} \tag{7.32}$$

以上により，今回提案したすべての LAIF のアフィン変換不変性が証明された．

時系列データとして，t をひとつずつずらしながら $F(\boldsymbol{X}_{\boldsymbol{t-k_1:t+k_2}}, t)$ を抽出していくと，$t = n$ の場合，F は，$\boldsymbol{X}_{\boldsymbol{n-k_1:n+k_2}}$ に対するアフィン変換に不変となる．また，$t = m$ の場合は，F は，$\boldsymbol{X}_{\boldsymbol{m-k_1:m+k_2}}$ に対するアフィン変換に不変となる．ここで，$\boldsymbol{X}_{\boldsymbol{n-k_1:n+k_2}}$ に対するアフィン変換と，$\boldsymbol{X}_{\boldsymbol{m-k_1:m+k_2}}$ に対するアフィン変換は，同一のアフィン変換でなくてもよい．すなわち LAIF は，ケプストラムの時系列データすべてに対する単一のアフィン変換のみに不変なのではなく，局所的な部分におけるアフィン変換に不変な特徴量となる．

7.2.3　特徴量マルチストリーム化

LAIF はアフィン変換に不変である．ここでアフィン変換は，話者の違いなどを近似する変換であるが，同時に，言語情報の一部も表現してしまう．あらゆるアフィン変換に不変になることは，音声認識タスクにおいては不変性が強すぎる [38]．つまり，LAIF は話者の違いに頑健であると同時に，単語弁別能力まで低くなってしまう．そこで，話者の違いなどのみに不変で，言語情報には不変にならないように，マルチストリーム化を導入することで適切な制約条件をかける．

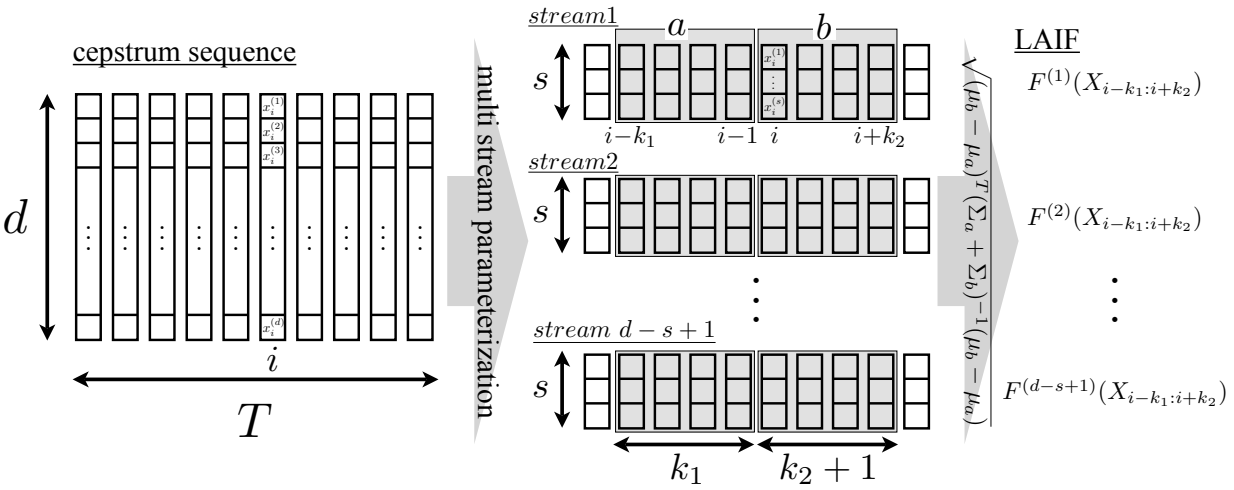

図 7.1: 特徴量次元分割を導入した LAIF の抽出

特徴量をブロック化するマルチストリーム化を用いることで，多次元の短時間特徴量として LAIF が計算できることになる．図 7.1 に，マルチストリーム化を導入した LAIF の抽出法を示す．このように LAIF を抽出することで，帯行列 $\boldsymbol{A}$ を用いたアフィン変換 $\boldsymbol{Ax}+\boldsymbol{b}$ のみに不変な短時間特徴量を計算することができる．

7.2.4　LAIF と他の短時間特徴量との関係

マルチストリーム化において，ブロックサイズ $s=1$ の場合，LAIF は $\boldsymbol{A}$ を対角行列とした場合の $\boldsymbol{Ax}+\boldsymbol{c}$ のみに不変になる．このとき，一つのストリームから式 (7.4) で抽出する LAIF は以下のように計算することができる．

$$\begin{aligned} f_t(X_{i-k_1:i+k_2}) &= \sqrt{(\mu_a-\mu_b)^T\left(\sigma_a^2+\sigma_b^2\right)^{-1}(\mu_a-\mu_b)} \\ &= \frac{|\mu_a-\mu_b|}{\sqrt{\sigma_a^2+\sigma_b^2}} \end{aligned} \tag{7.33}$$

ここで，ブロックサイズ s は 1 であるので，μ や σ はスカラーである．

さらに，$|\mu_a-\mu_b|$ は，$k=k_1=k_2$ とするとさらに以下のように書き下せる．

$$|\mu_a-\mu_b| = \left|\sum_{\tau=1}^{k} w_\tau\left(x_{i+\tau-1}-x_{i-\tau}\right)\right| \tag{7.34}$$

ただし w_τ は τ によらず $1/k$ である．ここで，w_τ の値を変えることは，各ケプストラムベクトルに重みをかけることに相当する．そのため，w_τ に任意の定数を与えても先に示したように式 (7.4) は LAIF となる．そこで，$w_\tau=\tau/(2\sum_{\tau=1}^{k}\tau^2)$ とおく．すると，式 (7.34) はデルタ特徴量の計算式 (7.3) とインデックスが一つずれていることと絶対値をとっていること除き同じものになる．

表 7.1: 使用する特徴量

Features(# of dimension)
MFCC(12)
MFCC(12) + LAIF$_{s=1}$(12)
MFCC(12) + LAIF$_{s=2}$(11)
MFCC(12) + ΔMFCC(12)
MFCC(12) + ΔMFCC(12) + LAIF$_{s=1}$(12)
MFCC(12) + ΔMFCC(12) + LAIF$_{s=2}$(11)

これをふまえ式 (7.33) に戻る．すると，ブロックサイズ $s = 1$ のときの LAIF は，おおまかにいって，分散を正規化して絶対値をとったデルタ特徴量に等しいことがわかる．

次に，ブロックサイズ s が 2 以上の場合を考える．今回我々は，ケプストラムから LAIF を抽出しているが，ケプストラムとスペクトルは直交変換の関係にあるので，スペクトル領域でも同様に LAIF を抽出することができる [50]．そのため，LAIF は一種の時間-周波数領域特徴量 (Spectro-Temporal Features; STF) ということができる．STF は，近年盛んに研究が行われている．例えば Muroi らは，パワースペクトルにある時間幅と帯域幅をもった 2 次元の窓をかけ，得られたパターンを識別学習して用いることにより話者性に頑健な音声認識を実現している [51]．LAIF も，ある時間幅と帯域幅をもった 2 次元の窓の出力から特徴量を計算している点は Muroi らの手法と同じである．しかし，そこにアフィン変換不変性という数学的基盤を持っていることが，大きく異なる点となっている．話者性を表すアフィン変換 $\boldsymbol{Ax} + \boldsymbol{b}$ の $\boldsymbol{A}$ は，幾何学的には回転性が強いことが知られており [52]，回転に影響を受けない LAIF を用いることは理にかなっている．

7.3　実験

この節では，不特定話者音声認識に対する効果を実験的に検証する．それに先立ち，LAIF を抽出する際の各種パラメータの設定値を予備実験により決定した．

LAIF を抽出する際の窓の幅を決めるパラメータである k_1, k_2 としては，サンプリング周波数が 16kHz の場合 $k_1 = k_2 + 1 = 16$ を用いる．$k_1 + 1$ や k_2 が 16 のときに最も成績がよいというのは，音声に含まれる変調周波数成分のうち，4Hz 付近に最も言語情報が含まれているという Kanedera らの結果におおよそ対応がとれる [53]．特徴量マルチストリーム化におけるブロックサイズ s は，1 または 2 を用いる．

音声認識実験においては，LAIF は MFCC のようなスペクトル特徴量とは異なる音声の特徴を捉えていると考えられるため，MFCC と LAIF の結合ベクトルを用いることにする．また，ΔMFCC の抽出における窓の幅 k は 2 を用いているため，LAIF$_{s=1}$ と ΔMFCC も，それぞれ異なる音声の特徴を捉えているものと考えられる．結局，従来のスペクトル特徴量のみのものも含めて表 7.1 に示した 6 種類それぞれについて認識実験を行うことにした．

表 7.2: 認識実験の結果

Method	M	M+$L_{s=1}$	M+$L_{s=2}$	M+Δ	M+Δ+$L_{s=1}$	M+Δ+$L_{s=2}$
通常条件	98.35%	99.24%	98.88%	99.47%	99.51%	99.39%
男学習 – 女評価	72.71%	83.22%	83.83%	82.79%	88.35%	89.27%
女学習 – 男評価	70.59%	83.25%	83.21%	85.34%	89.88%	90.70%

7.3.1　データベース

実験には，東北大松下音声単語データベースを用いる [54]．このデータベースは，男声話者 30 名と女声話者 30 名による，日本語 212 単語の孤立発声音が収録されている．また，音声は 12bit/16kHz でサンプリングされている．実験では，これを 16bit/16kHz に再サンプリングした音声ファイルを用いた．

7.3.2　孤立単語音声認識

HMM を用いた孤立単語音声認識実験を行ない，LAIF の不特定話者音声認識に対する有効性を評価する．HMM は単語単位で作成し，1 単語につき 25 状態の left-to-right 型 HMM，出力分布としては対角共分散の単一正規分布を用いた．通常の不特定話者音声認識タスク（Matched condition）として，学習話者を男声 15 名／女声 15 名，評価話者を男声 15 名／女声 15 名としたタスクと，学習データと評価データのミスマッチを大きくした条件の実験として，学習話者を男声 30 名，評価話者を女声 30 名とする実験，学習話者を女声 30 名，評価話者を男声 30 名とする実験も行なった．

実験結果を表 7.2 に示す．結果，MFCC や MFCC+ΔMFCC 単独で認識を行うよりも，LAIF を付け加えた方がより不特定話者に対する頑健性が高くなることがわかる．特に，学習データと評価データに性別のミスマッチがある場合，MFCC に対し，$\mathrm{LAIF}_{s=2}$ を結合することによるエラー削減率は 41%，MFCC+ΔMFCC に対し $\mathrm{LAIF}_{s=2}$ を結合することによるエラー削減率は 37%となった．また，ブロックサイズ s が 1 のときと 2 のときを比べると，ミスマッチがない条件では単語弁別能力がより高い $s = 1$ の方が認識率が良く，ミスマッチがある条件では話者不変性がより高い $s = 2$ の方がよいという結果になっている．

7.4　まとめ

本章では，アフィン変換不変性を持つ局所的特徴量 LAIF を提案し，その有効性を音声認識実験によって確認した．LAIF は，MFCC などといった既存のスペクトル特徴量に対し特徴量マルチストリーム化と簡単な計算を行うことにより容易に抽出できるものである．ケプストラムベクトルへのアフィン変換は話者の違いを近似するため，LAIF は話者の違いに対しておよそ不変となり，話者の違いに頑健な音声分析に効果があると考えられる．LAIF を使った音声認識実験を行った結果，LAIF+MFCC+ΔMFCC を用いることで，ミ

スマッチ条件下において 37%の誤り削減率を得た.

LAIF は，話者不変性を持ち，しかも話者正規化や話者適応などと異なり非常に簡単に計算できるという性質から，音声認識以外にも発音分析応用などにも有効であると考えられる.

第8章

結論

8.1　まとめ

本論文では，静的な変換で表現される非言語的特徴に不変となる音声の構造的表象を用いた，外国語発音分析・評価法の精度を大幅に改善させる手法を提案した．さらに，少ない読み上げ文章から音の差に基づく音声分析を行うために，局所的なアフィン変換不変量（LAIF）の利用を提案し，孤立単語音声認識実験により LAIF を発音分析に用いるための初期検討をおこなった．

音声の構造的表象を用いた発音分析・評価法の改善では，5 つの手法を提案した．一つ目は構造のエッジ長の分散の違いに影響を受けないためのエッジ長の正規化であり，非常に簡単なアイデアにも関わらず精度を大きく向上させることが実験的に示された．二つ目は構造の高すぎる次元数に対応するための部分構造化であり，特にあらかじめ手動で評価したデータが得られないような状況においても，発話者の発音の癖に対する知識を持つ人が部分構造をルールベースで選択することにより，精度を大きく向上させられることが分かった．三つ目は部分構造化よりさらに精度の高い次元圧縮法である二段階重回帰分析であり，適切な正則化項を導入したリッジ回帰を行うことで精度を大きく向上させられることが分かった．四つ目は，特徴量を増やすことで精度を向上させるマルチストリーム構造化とそれにより次元が増えることに対応する三段階重回帰分析であり，二段階重回帰分析を用いた手法からさらに精度を向上させられることがわかった．これにより，従来広く用いられている GOP スコアを用いた発音評価法と同等の精度で発音評価が行えることが実験的に示された．五つ目は，提案手法である構造を用いた手法と GOP スコアを，三段階重回帰分析の枠組みを用いて統合し組み合わせる手法であり，これにより構造を用いた手法や GOP スコアを用いた発音評価法よりさらに高い精度で発音分析が実現できることが実験的に示された．

少ない読み上げ文章から音の差に基づく音声分析を行うために，従来の HMM などを利用する発音分析法に，短時間音響特徴量のレベルで音の差に基いてアフィン変換不変性を有する LAIF を導入することで精度の向上をはかった．このように音の相対量を導入することで，既存の HMM などを利用する発音評価の枠組みとまったく同じ方法を利用することが可能になり，従来から広く研究されてきたさまざまな手法と組み合わせて利用することが可能になる．今回は，孤立単語音声認識実験において，特に学習データと評価データにミスマッチがある場合に，精度を向上させることができることが示された．

以上に述べたように，本論文では音声の構造的表象を用いた発音分析・評価を，研究室レベルでの実験から，現実的なアプリケーションとして有効活用できるレベルに改良を行った．すべてのタスクで実用上問題ない精度が実現できたわけではないものの，構造を用いることで，一部のタスクにおいては state-of-the-art の精度を実現することができた．

音声の相対量に基づく発音分析は，従来あまり行われてこなかった．今回行った研究により，音声の相対量を用いることが，さらに音声分析の精度向上を実現する可能性を秘めていることが確認され，音声情報処理研究全体に対してもひとつの貢献となることができたと考えている．

8.2　課題と今後の展望

本論文で解決できなかった大きな課題として，少ない文章のみを用いた発音分析法の実現がある．LAIF を用いることで，その初期検討は行ったが，ミスマッチのないときに精度が逆に低下しているなど，十分な精度を得られたとは言い難い．LAIF が十分な精度を実現できなかったことに対し，考えられる問題点は二つある．一つ目は，LAIF はある時間フレーム t の前後固定フレーム（本論文では 16 フレーム）を常に利用しているという点である．本来音声は時間的に伸縮する信号であるので，このフレーム数は状況に応じて動的に変更させることが必要であると考えられる．二つ目は，ある時間フレームの前後フレームしか見ていないという点である．部分構造化や多段階重回帰分析の実験でも示されているが，どの音素ペアの相対関係情報が有効であるかは，その音素ペアに強く依存している．LAIF では，常に時間的に隣り合う音響イベント間の相対関係しか利用していないため，その相対関係だけでは情報が足らず，もっと他の音響イベントとの相対関係をとる必要性があるものと考えられる．

十分な量の読み上げ文章が得られた場合のタスクにおける構造発音分析の精度の向上についても，まだまだ精度を向上させる余地は残されている．今回提案した多段階重回帰分析においては，重回帰分析としてリッジ回帰を利用している．その結果，得られる重みベクトルはマイナスの要素を含んでいる．しかし，重みベクトルがマイナスになることは物理的な意味から考えておかしなことであり，本来は非負の制約をつけた重回帰分析を利用するべきである．非負の制約をつけることのほか，リッジ回帰分析のような二次正則化ではなく，一次正則化を利用すると，必要ない音素ペアの重みが 0 になり，重みベクトルをスパースにする効果が高くなるため，部分構造化との関連も議論しやすくなる．このように，重回帰分析の方法はまだ検討する余地が残されている．

謝辞

本研究を進めるにあたり，常日頃からご指導，ご鞭撻を賜りました指導教員の峯松信明准教授に深く感謝致します．卒論，修士課程の 3 年間で，日頃の研究の進め方から，論文執筆，学会発表など，様々な形で熱心にご指導頂きました．広瀬啓吉教授には，本研究に対しての鋭いご意見やご助言を頂きました．ここに感謝の意を申し上げます．

研究環境の設備など，研究活動を様々な面で支えてくださった高橋登技官，秘書の池上恵さん，楠本由香里さんに深く感謝します．

広瀬・峯松研究室の皆様には，日頃より様々なご協力を頂きました．特に，理論的・数学的側面から多くの深い洞察力のある意見を頂いた研究員の喬宇氏に深く感謝します．また，博士課程の齋藤大輔氏には，日頃から白熱した議論をさせて頂き，沢山の有意義なアドバイスを頂きました．また，残念ながらすべての方の挙げることはできませんが，研究室の先輩，同期，後輩，OB の皆様のおかげでとても有意義な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．

最後に，今日まで自分を支えてくれた友人と家族に感謝いたします．

2010 年 2 月 9 日
鈴木 雅之

参考文献

[1] 法務省. 出入国管理統計統計表.
http://www.moj.go.jp/TOUKEI/ichiran/nyukan.html, 2008.

[2] 外務省領事局政策課. 海外在留邦人数統計（平成 21 年速報版）.
http://www.mofa.go.jp/toko/tokei/hojin/09/pdfs/1.pdf, 2008.

[3] 中央教育審議会外国語専門部会. 小学校における英語教育について（外国語専門部会における審議の状況）. http://www.mext.go.jp/b_menu/shingi/chukyo*ash*chukyo3/015/siryo/06032708/003.pdf, 2006.

[4] 中小企業基板設備機構. サービス産業業種別実態調査（対個人サービス業）報告書 2. 語学教室業. http://www.smrj.go.jp/keiei/dbps_data/_material_/chushou/b_keiei/service/pdf/H20houkokushogogaku.pdf, 2008.

[5] 文部科学省. 新しい学習指導要領.
http://www.mext.go.jp/a_menu/shotou/new-cs/youryou/syo/syo.pdf, 2008.

[6] 鳥飼玖美子. 危うし！小学校英語. 文藝春秋, 2006.

[7] えいご漬け. http://www.nintendo.co.jp/ds/angj/.

[8] smart.fm. http://smart.fm/.

[9] L.Bachman. Fundamental considerations in language testing. Cambridge University Press, 1990.

[10] J. Bernstein, M.Lipson, G.Halleck, and J. Martinez. Comparison of oral interviews and automatic tests of spoken language. In *Language Tesgin Research Colloquium (LTRC'1999)*, 1999.

[11] ベネッセ小学生向け英語学習プログラム be-go. http://be-go.benesse.ne.jp/be-go/.

[12] 篠田浩一. 確率モデルによる音声認識のための話者適応化技術. 電子情報通信学会論文誌, Vol. J87-D-II, No. 2, pp. 371–386, 2004.

[13] Y. Ohkawa, M. Suzuki, H. Ogasawara, A. Ito, and S. Makino. A speaker adaptation method for non-native speech using learners' native utterances for computer-assisted language learning system. *Speech Communication*, Vol. 51, No. 10, pp. 875–882, 2009.

[14] D. Luo, Y. Qiao, N. Minematsu, Y. Yamauchi, and K. Hirose. Analysis and utilization of mllr speaker adaptation technique for learners' pronunciation evaluation. In *Proc. INTERSPEECH'2009*, pp. 608–611, 2009.

[15] M. Russell and S. D'Arcy. Challenges for computer recognition of children's speech. In *Proc. ISCA Tutorial and Research Workshop on Speech and Language Technology in Education (SLaTE'2007)*, 2007.

[16] N. Minematsu. Yet another acoustic representation of speech sounds. In *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP'2004)*, pp. 585–588, 2004.

[17] 峯松信明. 音声の音響的普遍構造の歪みに着眼した外国語発音の自動評定. 電子情報通信学会技術研究報告, SP2003-180, pp.31–36, 2004.

[18] 朝川智, 峯松信明, 広瀬啓吉. 音声の構造的表象に基づく英語学習話者発音の音響的分析. 電子情報通信学会論文誌, Vol. J90-D, No. 5, pp. 1249–1262, 2007.

[19] N. Minematsu, K. Kamata, S. Asakawa, T. Makino, T. Nishimura, and K. Hirose. Structural assessment of language learners' pronunciation. In *Proc. INETERSPEECH*, pp. 210–213, 2007.

[20] 鎌田圭, 朝川智, 峯松信明, 牧野武彦, 広瀬啓吉. 音声の構造的表象に基づく発音矯正必要度の計算手法の検討. 電子情報通信学会技術研究報告, SP2007-36, pp.73–78, 2007.

[21] 鎌田圭, 高澤真章, 朝川智, 峯松信明, 牧野武彦, 広瀬啓吉. 音声の構造的表象に基づく英語発音分析結果の視覚化に対する一考察. 日本音響学会春季講演論文集, 3-Q-23, pp.425–426, 2008.

[22] 鎌田圭, 高澤真章, 竹内京子, 朝川智, 峯松信明, 牧野武彦, 広瀬啓吉. 大規模英語学習者を対象とした音声の構造的表象に基づく発音分類とその応用. 情報処理学会全国大会講演集, 2ZL-6, pp.275–276, 2008.

[23] 高澤真章, 鎌田圭, 竹内京子, 朝川智, 峯松信明, 牧野武彦, 広瀬啓吉. 大規模英語学習者を対象とした音声の構造的表象に基づく発音評価とその応用. 日本音響学会春季講演論文集, 3-10-12, pp.489–492, 2008.

[24] S.M. Witt and S. Young. Phone-level pronunciation scoring and assessment for interactive language learning. *Speech Communication*, Vol. 30, pp. 95–108, 2000.

[25] S. Young, et al. *The HTK Book*.

[26] S. Furui. Speaker independent isolated word recognition using dynamic features of speech spectrum. *IEEE Trans. Acoustics, Speech and Signal Processing*, Vol. 34, No. 1, pp. 52–59, 1986.

[27] 鹿野清宏, 伊藤克亘, 河原達也, 武田一哉, 山本幹雄. IT Text 音声認識システム. オーム社, 2001.

[28] S. Kanters, C. Cucchiarini, and H. Strik. The goodness of pronunciation algorithm: a detailed performance study. In *Proc. ISCA Tutorial and Research Workshop on Speech and Language Technology in Education (SLaTE'2009)*, 2009.

[29] M. Pitz and H. Ney. Vocal tract normalization equals linear transformation in cepstral space. *IEEE Trans. Speech and Audio Processing*, Vol. 13, No. 5, pp. 930–944, 2005.

[30] 戸田智基, 陸金林, 猿渡洋, 鹿野清宏. 周波数軸伸縮を用いた混合正規分布モデルに基づく声質変換法. 電子情報通信学会論文誌, Vol. J84-D-II, No. 10, pp. 2181–2189, 2001.

[31] B. Atal. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *Journal of the Acoustical Society of America*, Vol. 55, No. 6, pp. 1304–1312, 1974.

[32] S. Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. Acoustics, Speech and Signal Processing*, Vol. 27, No. 2, pp. 113–120, 1979.

[33] E. Eide and H. Gish. A parametric approach to vocal tract length normalization. In *Proc. Int. Conf Acoustics, Speech, and Signal Processing (ICASSP'1996)*, pp. 346–348, 1996.

[34] C. J. Leggetter and P. C. Woodland. Maximum likelihood speaker adaptation of continuous density hidden markov models. *Computer Speech and Language*, Vol. 9, pp. 171–185, 1995.

[35] S. Furui. Generalization problem in asr acoustic model training and adaptation. In *Proc. Int. Workshop on Automatic Speech Recognition and Understanding (ASRU'2009)*, pp. 1–10, 2009.

[36] 村上隆夫, 峯松信明, 広瀬啓吉. 音声の構造的表象に基づく日本語孤立母音系列を対象とした音声認識. 電子情報通信学会論文誌, Vol. J91-A, No. 2, pp. 181–191, 2008.

[37] 朝川智, 村上隆夫, 峯松信明, 広瀬啓吉. 音声の構造的表象に基づく日本語母音系列連続発声の認識. 電子情報通信学会技術研究報告, SP2006-105, pp.119–124, 2006.

[38] S. Asakawa, N. Minematsu, and K. Hirose. Multi-stream parameterization for structural speech recognition. In *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP'2008)*, pp. 4097–4100, 2008.

[39] Y. Qiao, S. Asakawa, and N. Minematsu. Random discriminant structure analysis for continous japanese vowel recognition. In *Proc. Int. Workshop on Automatic Speech Recognition and Understanding (ASRU'2007)*, pp. 576–581, 2007.

[40] 朝川智, 喬宇, 広瀬啓吉. 音声の構造的表象と判別分析を用いた単語音声認識. 電子情報通信学会技術研究報告, SP2008-113, pp.203–208, 2008.

[41] Y. Qiao, S. Asakawa, N. Minematsu, and K. Hirose. On invariant structural representation for speech recognition: theoretical validation and experimental improvement. In *Proc. INTERSPEECH*, pp. 3055–3058, 2009.

[42] 斎藤大輔, 松浦良, 峯松信明, 広瀬啓吉. 話者不変な相対関係特徴を音響単位とする音響モデリングに関する実験的検討. 電子情報通信学会技術研究報告, SP2009-77, pp.7–12, 2009.

[43] Y. Qiao and N. Minematsu. f-divergence is a generalized invariant measure between distributions. In *Proc. INTERSPEECH*, pp. 1349–1352, 2008.

[44] Y. Qiao and N. Minematsu. Metric learning for unsupervised phoneme segmentation. In *Proc. INTERSPEECH*, pp. 1060–1063, 2008.

[45] 峯松信明, 志甫淳, 村上隆夫, 丸山和孝, 広瀬啓吉. 音声の構造的表象とその距離尺度. 電子情報通信学会技術研究報告, SP2005-13, pp.9–12, 2005.

[46] 江森正, 篠田浩一. 音声認識のための高速最ゆう推定を用いた声道長正規化. 電子情報通信学会論文誌, Vol. J83-D-II, No. 11, pp. 2108–2117, 2000.

[47] 峯松信明, 富山義弘, 吉本啓, 清水克正, 中川聖一, 壇辻正剛, 牧野正三. 英語 call 構築を目的とした日本人及び米国人による読み上げ英語音声データベースの構築. 日本教育工学会論文誌, Vol. 27, No. 3, pp. 259–272, 2004.

[48] H. Kawahara. Straight, exploitation of the other aspect of vocoder: Perceptually isomorphic decomposition of speech sounds. *Acoust. Sci. & Tech.*, Vol. 27, No. 6, pp. 349–353, 2006.

[49] Y. Qiao, M. Suzuki, and N. Minematsu. Affine invariant features and its application to speech recognition. In *Proc. Int. Conf. Acoustics, Speech, & Signal Processing (ICASSP'2009)*, pp. 4629–4632, 2009.

[50] 鈴木雅之, 朝川智, 喬宇, 峯松信明, 広瀬啓吉. スペクトル特徴量を用いた音声の構造的表象に関する実験的検討. 電子情報通信学会技術研究報告, SP2008-32, pp.7–12, 2008.

[51] T. Muroi, T. Takiguchi, and Y. Ariki. Speaker independent phoneme recognition based on fisher weight map. *International Journal of Hybrid Information Technology*, Vol. 1, No. 3, 2009.

[52] 斎藤大輔, 松浦良, 朝川智, 峯松信明, 広瀬啓吉. ケプストラムの声道長依存性に関する幾何学的考察. 電子情報通信学会技術研究報告, SP2007-128, pp.189–194, 2007.

[53] N. Kanedera, T. Arai, H. Hermansky, and M. Pavel. On the relative importance of various components of the modulation spectrum for automatic speech recognition. *Speech Communication*, Vol. 28, No. 1, pp. 43–55, 1999.

[54] 牧野正三, 二矢田勝行, 真船裕雄, 城戸健一. 東北大-松下単語音声データベース. 日本音響学会誌, Vol. 48, No. 12, pp. 899–905, 1992.

発表文献

学術論文

[1] N. Minematsu, S. Asakawa, M. Suzuki, “Speech structure and its application to robust speech processing,” Journal of New Generation Computing (2009, submitted)

国際会議論文

[2] M. Suzuki, N. Minematsu, D. Luo, and K. Hirose, “Sub-structure-based estimation of pronunciation proficiency and classification of learners,” Proc. Int. Workshop on Automatic Speech Recognition and Understanding (ASRU’2009), pp.574-579 (2009-12)

[3] Y. Qiao, M. Suzuki, and N. Minematsu, “A study of Hidden Structure Model and its application of labeling sequences,” Proc. Int. Workshop on Automatic Speech Recognition and Understanding (ASRU’2009), pp.118-123 (2009-12)

[4] M. Suzuki, L. Dean, N. Minematsu, K. Hirose, “Improved structure-based automatic estimation of pronunciation proficiency,” Proc. ISCA Tutorial and Research Workshop on Speech and Language Technology in Education (SLaTE), CD-ROM (2009-9)

[5] N. Minematsu and M. Suzuki, “Structure-based pronunciation assessment,” Proc. ISCA Tutorial and Research Workshop on Speech and Language Technology in Education (SLaTE), Demo Session (2009-9)

[6] H. Hirano, M. Suzuki, K. Innami, N. Minematsu, and K. Hirose, “Development of an on-line word accent dictionary of Japanese,” Proc. Int. Conf. on Japanese Language Education (ICJLE’2009) (2009-7)

[7] Y. Qiao, M. Suzuki, and N. Minematsu, “Affine invariant features and its application to speech recognition,” Proc. Int. Conf. Acoustics, Speech, & Signal Processing (ICASSP’2009), pp.4629-4632 (2009-4)

国内研究会論文

[8] 鈴木雅之，羅徳安，峯松信明，広瀬啓吉，“音声の構造的表象を用いた自動発音評定法の改善”，情報処理学会音声言語情報処理研究会，2009-SLP-77-17，pp.1-6 (2009-7)

[9] Y. Qiao, M. Suzuki, and N. Minematsu, “An Investigation of Hiden Structure Model,” 情報処理学会音声言語情報処理研究会, 2009-SLP-77-5, pp.1-6 (2009-7)

[10] 鈴木雅之，喬宇，峯松信明，広瀬啓吉 “アフィン変換不変性を有する局所的特徴量を用いた音声認識”，電子情報通信学会音声研究会，SP2008-114，pp.209-214 (2008-12)

[11] 國越晶，喬宇，鈴木雅之，峯松信明，広瀬啓吉，“空間写像に基づく手の動きを入力とした音声生成系の構築”，電子情報通信学会音声研究会，SP2008-78，pp.45-50 (2008-11)

[12] 鈴木雅之，朝川智，喬宇，峯松信明，広瀬啓吉，“スペクトル特徴量を用いた音声の構造的表象に関する実験的検討”，電子情報通信学会音声研究会，SP2008-32，pp.73-78 (2008-6)

国内研究会論文

[13] 鈴木雅之，喬宇，峯松信明，広瀬啓吉，牧野武彦，“構造表象と多段階の重回帰を用いた外国語発音評価”，日本音響学会春季講演論文集 (2010-3，発表予定)

[14] 中村綾乃，鈴木雅之，峯松信明，広瀬啓吉，牧野武彦，“音声の構造的表象を用いた英語二重母音の発音評価に関する検討”，日本音響学会春季講演論文集 (2010-3，発表予定)

[15] 千々岩圭吾，鈴木雅之，齋藤大輔，峯松信明，広瀬啓吉，“非周期性に注目した基本周期パターン生成過程モデルのパラメータ自動抽出の高精度化” 日本音響学会春季講演論文集 (2010-3，発表予定)

[16] 清水信哉，齋藤大輔，鈴木雅之，峯松信明，広瀬啓吉，“類似単語の置換による言語モデルの平滑化”，日本音響学会春季講演論文集 (2010-3，発表予定)

[17] 鈴木雅之，喬宇，峯松信明，広瀬啓吉，“音声の構造的表象と多段階の重回帰を用いた外国語発音分析”，情報処理学会全国大会講演集，1U-9 (2010-3，発表予定)

[18] 鈴木雅之，羅徳安，峯松信明，広瀬啓吉，“発音構造を用いた話者の違いに頑健な発音評定と学習者分類”，日本音響学会秋季講演論文集，1-2-5，pp.243-246 (2009-9)

[19] 喬宇，鈴木雅之，峯松信明，“Proposal of Hidden Structure Model,” 日本音響学会秋季講演論文集，1-1-3，pp.7-10 (2009-9)

[20] 鈴木雅之，喬宇，峯松信明，広瀬啓吉，“アフィン変換不変性を有する局所特徴量を用いた音声認識”，日本音響学会春季講演論文集，1-5-4，pp.11-14 (2009-3)

[21] 國越晶，喬宇，鈴木雅之，峯松信明，広瀬啓吉，坂野秀樹，“ジェスチャー空間と音響空間の写像に基づくリアルタイム音声生成系”，日本音響学会春季講演論文集，2-P-2，pp.445-448 (2009-3)

[22] 鈴木雅之，朝川智，喬宇，峯松信明，広瀬啓吉，“スペクトル領域特徴量を用いた音声の構造的表象による音声認識”，日本音響学会秋季講演論文集，1-R-26，pp.473-476 (2008-9)

[23] 桜庭京子，峯松信明，廣瀬啓吉，鈴木雅之，田山二朗，今泉敏，山内俊雄，“MtF のボイスセラピーにおける成功症例の型の分類”，性同一性障害学会研究大会 (2008-3)

学位論文

[24] 鈴木雅之，“周波数領域の特徴量を用いた音声の構造的表象による雑音環境下音声認識”，東京大学工学部電子情報工学科卒業論文 (2008)

付録A

米語発音のアルファベット表記法

付録 A　米語発音のアルファベット表記法

本論文で用いた ERJ データベースで用いられ，本論文でも利用した米語発音のアルファベット表記とその発音に対応する IPA 記号，実際にその発音が使われる単語のサンプルの対応を表 A.1 に示す.

表 A.1: 米語発音のアルファベット表記と IPA 記号の対応

アルファベット表記	IPA 記号表記	単語の例 (発音)
AA	ɑ	POT (P AA T)
AE	æ	BAT (B AE T)
AH	ʌ	BUT (B AH T)
AO	ɔ	BOUGHT (B AO T)
AW	au	MOUNT (M AW N T)
AX	ə	ABOUT (AX B AW T)
AXR	ɚ (unstressed)	BUTTER (B AH T AXR)
AY	ai	BITE (B AY T)
B	b	BAY (B EY)
CH	tʃ	CHOKE (CH OW K)
D	d	DAY (D EY)
DH	ð	THEN (DH EH N)
EH	ɛ	BET (B EH T)
ER	ɚ (stressed)	BIRD (B ER D)
EY	ei	BAIT (B EY T)
F	f	FIN (F IH N)
G	g	GAY (G EY)
HH	h	HAY (HH EY)
IH	ɪ	BIT (B IH T)
IY	i	BEET (B IY T)
JH	dʒ	JOKE (JH OW K)
K	k	KEY (K IY)
L	l	LAY (L EY)
M	m	MOM (M AA M)
N	n	NOON (N UW N)
NG	ŋ	SING (S IH NG)
OW	ou	BOAT (B OW T)
OY	ɔi	BOY (B OY)
P	p	PAY (P EY)
R	r	RAY (R EY)

アルファベット表記	IPA 記号表記	単語の例 (発音)
S	s	SEA (SH IY)
SH	ʃ	SHE (S IY)
T	t	TEE (T IY)
TH	θ	THIN (TH IH N)
UH	ʊ	BOOK (B UH K)
UW	u	BOOM (B UW M)
V	v	VAN (V AE N)
W	w	WAY (W EY)
Y	j	YOU (Y UW)
Z	z	ZONE (Z OW N)
ZH	ʒ	MEASURE (M EH ZH ZXR)