修士論文

世界諸英語クラスタリングを目的とした発音距離予測の高精度化のための実験的検討 (An experimental study on improvement of prediction of pronunciation distance with the aim of World Englishes clustering)

2015 年 2 月 5 日

指導教員　峯松 信明 教授

電気系工学専攻 融合情報学コース

37-136444　笠原 駿

内容概要

国際語として世界中に広まっていく過程で英語は，各国各地域の地域性や民族性を反映して，様々な訛りを伴って話されることとなった．この英語に対し,「英語利用者が従うべき標準的な英語を設けるのではなく，英語が多様化した現状をそのまま受け入れる」世界諸英語という考えがある．国際的なビジネスの現場では，様々な地方訛り，外国語訛りの英語と遭遇することが多く，この多様性をどう対処するのかに着眼した英語教育を実践している教育者もいる．このように一人一人が異なる英語を話す場合，英語話者それぞれに英語の多様性をはっきりと認識させる技術もまた重要となるだろう．

本研究は発音に着目して英語話者群を個人を単位に分類し，世界諸英語の多様さを地図として提示すると共に，話者本人の英語発音が世界諸英語の中でどのように位置づけられるかを教示することを最終的な目標としている．発音分類のために本研究では，二話者間の英語発音の “距離” を定義し，またその距離を音声情報と回帰により予測することを試みている．

各国の話者による英語パラグラフ読み上げ音声のコーパスを対象として，音響特徴量を入力としたサポートベクター回帰により発音距離の予測実験を行なう．話者の声色の違いに頑健な発音構造特徴について，本稿ではその算出手法を変更することにより，距離予測の高精度化と分析を行なっている．提案する手法により発音距離予測の精度は先行研究から大幅に上げることができ，ベースラインシステムを超える結果を示すことができた．しかし同時に，より実用に即した実験条件での結果も新たに示し，本研究の最終目標に対してはまだ性能が不十分であるということも報告している．

本稿ではさらに，話者間の発音の差異を直接比較する絶対的特徴と，基準とした話者からの距離を利用した形状特徴を提案している．実験の結果，絶対的特徴については構造特徴に追加することでわずかながら精度改善に貢献することを示したが，これらの特徴のより効果的な利用方法については，今後の課題と位置づけることとなった．

目次

図目次

表目次

第1章

序論

1.1　国際共通語としての英語の現状

英語は唯一の世界共通語として受け入れられ，母国語として，公用語として，あるいは外国語として様々な国で話されている．英語が世界中に広まっていく中で，各国の地域性や民族性の影響を受けて，英語の統語，語用，綴り，発音など様々な側面が不可避的に変化してきた．発音に着眼すれば，現在世界中には多くの訛り（外国語・地方訛り）英語が存在している．この訛りは話者の母国語や居住地のみでは決まらず，話者の学習方法や生活スタイル，家族や友人・英語教師といった周りの人間の有する訛りなどに影響されるもので，より厳密に考えれば，各自の訛りは個人により異なると言っても良いであろう．英語が多様化した現状を鑑みて，近年 Kachru らにより，英語には標準となるものを設けないとする世界諸英語 (World Englishes, WE) [1, 2] の考え方が提唱され，これを採択する教師が増えている．国際交流のさらなる発展に伴い英語の実利性や必要性は今後も増していくが，多様性も同時に拡大し続けることになる．交流の現場に立つ人々は，多様な英語と接することを余儀なくされている．

1.2　本研究の目的

英語の多様性が許容され，一人一人が異なる発音を有する場合においては，世界にどういう英語が存在しているかを知り，またそれらの中で自身の英語がどこに位置づけられるのか客観的に捉えることも必要となるだろう．本研究の最終的な目標は，発音の訛りに焦点を置き，話者個人を単位として自動分類し， WE を一望できる発音地図を作成することである．発音地図により英語学習者は，自分と他者との英語訛りの近さを知ることができ，目的に沿った英会話相手を選択することが可能となる．また英語訛り音声を多数含む既存の Web アーカイブを利用すれば，アーカイブ上の話者を発音地図に配置することにより，特定の訛り英語の話者を探し出せる Web ブラウジングシステムを構築することができる．

話者を分類するためには，任意の話者間で発音がどれだけ違うかを定量的に表した “発音距離” が必要である．本研究では，この発音距離を定義するとともに，自動化のために音声の音響的特徴のみを用いて予測することを試みる．

本稿では [3] での距離予測実験の枠組みに従いながら，より目標とするシステムに近い実験条件を提案し，その条件下でも実験を行っている．距離予測に用いる音響特徴量の抽出手法の変更や新たな特徴量の提案により， [3] からの予測性能の改善と，予測に有効な特徴量の分析を試みている．

1.3　本稿の構成

本稿では，まず第 2 章にて，世界の英語の現状について述べ，この多様化した英語を収集し観察する動きについて述べる．第 3 章では，多様な英語を工学的技術により分類する関連研究，及び本稿の先行研究について説明する．第 4 章では，先行研究の手法を改善し時間的に離れた二音間の特徴も考慮した上で，2 種類の条件で発音距離予測実験を行う．先行研究の条件では予測精度を大幅に向上させたことと，より本研究の目標に即した実験条件では提案する手法でも不十分であるという結果が得られたことを報告している．第 5 章では，先行研究になかった新たな二種類の特徴量を導入した．二話者の発音の差異を直接みる特徴を用いることでわずかな精度改善が見られたが，どちらの特徴量にしても実用に対しては不十分な程度である．第 6 章ではまとめと残された課題について述べている．

第2章

研究の背景

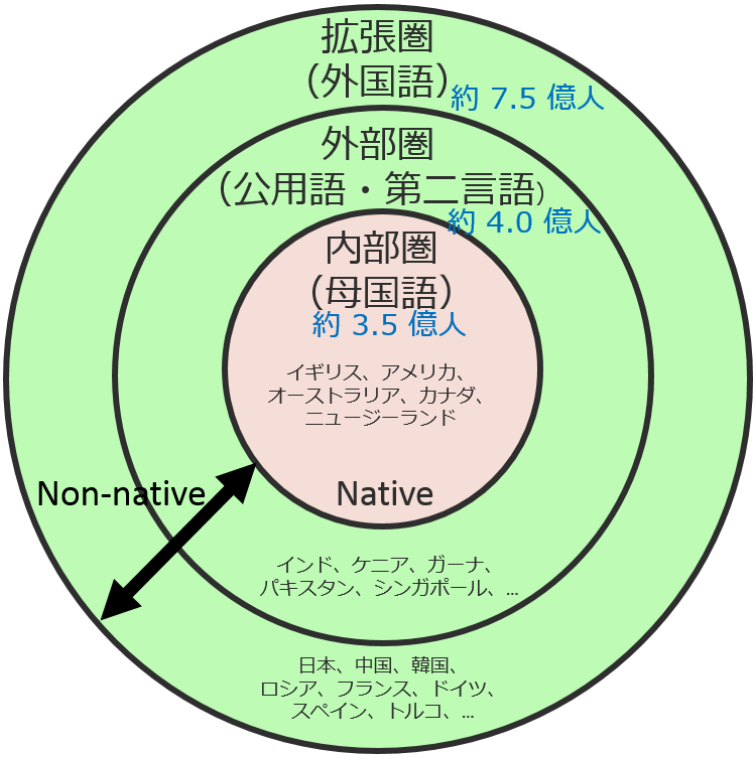

図 2.1: Kachru の同心円モデル（[4] を元に作成）

2.1　英語の国際化と多様化

英語には “国際化” と “多様化” という，他のどの言語にもない二つの大きな特徴がある．

図 2.1 は英語を母語とする国から第 2 言語または公用語とする国へ，さらに外国語として使用する国へと英語が広がっていく様子を同心円にして示したものである．現代では英語は 15 億人（世界人口の約 25%）に話されて，70ヵ国（約 36%）で通用する [5] 言語である．国際ビジネスや国際協力プロジェクトの現場では当然のように英語が用いられる．また，多くの企業や団体が，自身のウェブサイトを国語以外に英語でも作成している．国際交流は今後もさらに発展していき，英語の実利性は増す一方であろう．

世界中に広まり使用される過程で英語は，各地の地域性や民族性を反映して様々な訛りを伴った形で使用されている．以下にいくつかの訛りの具体例を示す．

- 統語訛り (syntax)：文法に置ける特徴を指す．
  例 マレーシアの英語
  -時世を区別せず，現在・過去・未来を全て現在形で表現する．過去完了を過去形で表現する．
  南アフリカの英語
  -否定疑問文に対する Yes/No の答えが問いの内容の成否に依存する．
  -(“Didn’t you find it?-Yes, I didn’t./No, I did.”)
- 語彙訛り (lexical choice)：元々英語にはなかった語彙が英語として使用されることを指す．
  例 土着語から英語になった言葉
  -penguin（ウェールズ語の「白い頭」）, pajama（ヒンディー語の「ズボン」）
  既存の英単語を組み合わせた言葉
  -American time/Filipino time （時間厳守/時間にルーズ，フィリピン）, bed town（日本）
- 発音訛り (pronunciation)：話者の母語に影響されて起こる発音の違いを指す．
  例 中国の英語
  -母音に続く r が発音されない．

日本の英語
-[r] と [l], [θ] と [s], [b] と [v] などの区別がない.

2.2　世界諸英語

英語が多様化した現状を鑑みて，近年 Kachru らにより，英語には標準となるものを設けるべきではないとする世界諸英語 (World Englishes; WE) [1, 2] の考え方が提唱され，これを採択する教師が増えている．ものごとの普及には必ず各地域への適応が伴う．言語においても同様であり，多様化することを許容しなければ，世界中で受け入れられ使用されることにはならない．英語はもはやアメリカやイギリスの言葉という範疇を越えていて，人々は英米文化を理解するために英語を学ぶのではなく，自国の文化を世界に向けて表現するための手段として英語を利用するのである．

2.3　多様な英語音声を取り扱ったデータベース

各話者の個性として許容されつつある多様な英語を理解するために，実際の訛り英語音声を聴取できる環境を構築することは重要である．近年では YouTube [1] や TED [2] に代表されるインターネット上の動画共有サイトにて，様々な訛り音声が多数共有されている．この他にも，多様な英語音声を収集することを主目的として作成された音声データコーパスが存在し，これらを利用すれば目的とする地域の訛り英語音声にさらに容易にアクセスすることが可能である．訛り英語音声コーパスとして，特定パラグラフの読み上げ音声を収録した Speech Accent Archive (SAA) [6] と International Dialects of English Archive (IDEA) [7] を紹介する．本研究では SAA のデータを対象として実験を行っている．訛り英語のコーパスはこれらの他にも，中立言語としての英語の使用状況を研究する目的で開発された Vienna-Oxford International Corpus of English (VOICE) [8], the Corpus of English as a Lingua Franca in Academic Setting (ELFA) [9, 10] といったものが挙げられる．VOICE, 及び ELFA では自然会話や独白などを収録しており，発話内容や発音についての書き起しはなされていない.

2.3.1　Speech Accent Archive

SAA は次に示す英語パラグラフの読み上げ音声と，各音声に対応する発音書き起しがセットになって提供されているコーパスである.

> Please call Stella. Ask her to bring these things with her from the store: Six spoons of fresh snow peas, five thick slabs of blue cheese, and maybe a snack for her brother Bob. We also need a small plastic snake and a big toy frog for the kids. She can scoop these things into three red bags, and we will go meet her Wednesday at the train station.

発音書き起しには国際音声記号 (International Phonetic Alphabet; IPA) が用いられている．IPA は音声学の専門的な記号で，各文字は発話音声の最小単位である単音に対応し，あらゆる言語の発音を書き下すために使用される．図 2.2 は IPA による書き起しの例である．IPA は無声化，鼻音化などを表す装飾記号 (diacritical mark) を用いることで同じ文字でも発音方法などにより細分化される．音素（特定の言語における音声の最小単位）の種類数が，例えば日本語の場合約 25 種類，米語の場合約 40 種類であるのに対して，IPA の異なりシンボル種類数は SAA にある全書き起しデータで使われているだけでも約 550 まである.

[1] `https://www.youtube.com/`, 2015.
[2] `http://www.ted.com/`, 2015.

"Please call Stella. Ask her to bring these things with her from the store."
English2: [pʰliːz kɒl stɛlə ʔɑsk hɜ tə bɹ̃ŋ ðiːz θ̃ŋz wɪθ hɜ fɹə̃m ðə stɔː]
Indonesian4: [plis kol stɛlɐ as hɜɪ ɾə brɪŋ θɪs θɪŋs wɪθ hɜɪ fɹəm ðə stɔːɹ]
Japanese4: [pəɾiz kɔl sɛʔə stɛɾɑ askə hɜ to bɹ̃ŋ dʱiʐ̪ə s̃ɪŋzˌ wɪʐ̪ə hɜ fɪʌm ʐ̪ə stɔɹ]

図 2.2: SAA の IPA 書き起し [6] の例

SAA のパラグラフは米語の音素（音素組み合わせ）に対する被覆率が高くなるように設計されている．パラグラフは 69 単語からなり，CMU 発音辞書 [11] を参照すると 221 個の米語音素系列に変換することができる．

これまでインターネットを通じて世界中のボランティアの話者約 2,000 人から音声が提供されており，このうち約 1,200 人分の発音が既に書き起されている．収録は各々の録音環境で行われており，音声によっては大きな背景雑音を含んでいる．

2.3.2 International Dialects of English Archive

IDEA は元々，役者が様々な地域訛りを覚えるのを支援するために作成されたインターネットサイトである．

IDEA では The Rainbow Passage [12] [3] または Comma gets a cure [4] というパラグラフの読み上げ音声が提供されている．The Rainbow Passage は 331 語からなり，話者の英語発音能力を測るための教材として IDEA の他でも頻繁に利用されている．Comma gets a cure は 375 語からなる，IDEA にて作成されたパラグラフである．Well が定義した標準語彙セット [13] を参照して単語が選択されており，発音の差がより見出しやすくなっている．

IDEA にはこれまで 1,000 人以上の音声が提供されているが，このうち IPA で発音が書き起されている話者は 30 人に満たない．

2.4　世界諸英語に関する日本語教材

世界諸英語を題材とした教材には日本語で書かれたものも数多く出版されている．[14] は，世界の英語と日本の英語の対比をしながら，国際共通語としての英語の特性を解説し，その中で各自が固有の英語を習得するためにどう考えるべきか，また日本語教育はどうあるべきかについて考察している．[15] は，英語母語国を含む各国・各地域について，見られる英語訛りの背景と特徴を具体例を交えながら解説している．[16] では，ダボス会議（世界経済フォーラムの年次総会）での非母語話者の発言を取り扱っていて，発言内容と背景の解説を読みながら，付属 CD で音声を聴取することができる．

[3] http://www.dialectsarchive.com/wp-content/uploads/2013/02/The-Rainbow-Passage.pdf, 2015.
[4] http://www.dialectsarchive.com/wp-content/uploads/2012/10/COMMA-GETS-A-CURE.pdf, 2015.

第3章

本研究に関連する先行研究

3.1　はじめに

本節ではまず 3.2, 3.3 にて，訛り・方言分析のための話者分類の関連研究について述べる．3.2 で説明するのは，問題毎に有限個の訛りクラスをあらかじめ設定し，各話者がどのクラスに属するか同定するトップダウンな分類問題である．これに対し 3.3 で説明するのは，話者間の発音の差異を定義し書き起しから直接求めることにより，各話者のパーソナルな情報を一切使わず発音の相対関係のみで話者をボトムアップに分類する（クラスタリングする）問題である．本研究は「英語の訛りは個人を単位とした多様性を持つ」という観点から，本研究の最終目標である世界諸英語を対象とした話者分類は 3.3 のボトムアップ・クラスタリングにより達成するべきだと考えている．世界中の英語話者を分類するための技術的検討として先行研究では，音声情報のみから発音差異を予測することを試みている．3.4 では，先行研究で利用している，話者の声色の違いに対し頑健な音声の構造的表象について説明する．

3.2　英語発音訛りの自動クラス分類

米語や中国語の方言，外国語訛りを対象として，各話者の発音がどの（方言・訛り）クラスに属するかを自動で推定する．英語音声を入力とし，その話者の母国語を推定するタスクもこの範疇である．各自であらかじめ設定した方言の分類しか行なえない限定的な問題となるが，分類の精度は高い．各話者に割り振られた訛りラベルを利用した学習が可能で，各訛りを HMM (Hidden Markov Model) や GMM (Gaussian Mixture Model) などを用いてモデル化し入力特徴量に対する最尤から訛りを推定する [17, 18, 19, 20]，SVM (Support Vector Machine) などを用いて訛りの識別器を構築する [21, 22] といった方策がとられる．

訛り分類器の実用について述べる， 分類器を前処理に用いることで入力音声の訛りを同定できれば，その訛りに適応した自動音声認識器を使用し認識性能を向上させることや，コールシステムで利用者と訛りの近いオペレータを自動的に選択することなどが可能になる．また音声情報からその発話者を特定したい場合に，音声の訛りを知ることができれば，発話者の国籍や居住地域，母国語を知る手がかりとすることができる．

訛り分類で用いられる特徴は，大きく分けて韻律的特徴と分節的特徴の二つでまとめられる．

韻律的特徴は訛り分類や言語分類において特に有効となる特徴である．イントネーションのパターン [17, 18] や各単語発話にかけられる時間，閉鎖音発声時に声門が閉じてから再び開放するまでにかかる時間 [17] などが利用される．

分節的特徴は一般的な音声認識・合成においても頻繁に用いられる特徴である．フォルマント周波数やメル周波数ケプストラム係数 [21] などが利用される．

3.2.1　基本周波数とフォルマント周波数

音声生成のプロセスは，音源の生成，調音，放射の 3 段階から成る．

人が声を出そうとする時，通常の呼吸時は開いている声帯が狭まり，その間を肺からの空気が通り抜けようとする．この時空気流と声帯の相互作用により，声帯が周期的に改変し，ほぼ規則的な空気の断続が生じる．この空気流の変化が音声の音源となる．この時の声帯の開閉周期（振動周期）を基本周期と呼び，その逆数を基本周波数と呼ぶ．基本周波数は声の高さ（ピッチ）に対応する．英語の訛り自動分類においては，発話中の基本周波数の勾配に着目し，米語母語話者と比べドイツ語を母語とする話者の方が勾配が大きい，マンダリンを母語とする話者は小さい [17] といった傾向が利用される．

音声を言語音として発するために，舌や口唇を動かし声道の形状を調整することを調音と呼ぶ．調音による音響的な（声道）フィルタを通過することで音源の伝達特性が変わり，共鳴作用で周波数に

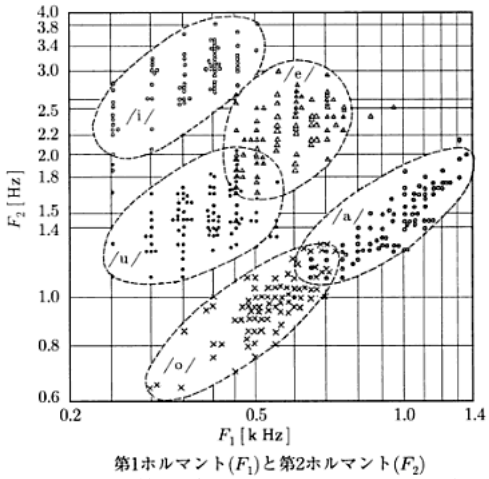

図 3.1: 日本語単母音のフォルマント周波数 [23]

よりエネルギーに強弱が生じる．そのため，音声の周波数スペクトルの包絡特性を抽出できれば，それは調音に対応するため音声の言語的特徴と考えることができる．

スペクトル包絡がピークとなる周波数スペクトル成分はフォルマントと呼ばれる．特に，周波数の低い方から第一フォルマント，第二フォルマント，...，それらの周波数をフォルマント周波数と呼ぶ．図 3.1 は， 7 歳から成人までの男女について，日本語単母音 (/a, i, u, e, o/) の第一，第二フォルマントを測定し記録したものである [23] ．フォルマント周波数は音韻だけでなく年齢，性別によっても変化する．全体的な傾向として，女性より男性の方が，また年齢が高い方がフォルマント周波数が低くなっており，成人男性の/a/と子供の/o/が，また同様に/e/と/u/が重なっている．話者の年齢，性別が分かっていれば，第二フォルマントまで利用することで母音をおおよそ区別することが可能である．訛り分類においてはしばしば第三フォルマントまで参照される [19, 20, 24].

3.2.2　メル周波数ケプストラム係数

信号波形の周波数スペクトルの対数をとり，さらに逆フーリエ変換をかけたものをケプストラムと呼ぶ．ケプストラム分析は，畳み込みで表されている複数の信号を，対数をとり和の形に直した上で変換することで，それぞれの信号に分離する処理である．

音声信号からケプストラムを抽出する過程を図 3.2 に示す．音声波形に窓をかけ数十ミリ秒程度のフレームを切り出し，その区間を離散フーリエ変換 (Discrete Fourier Transform; DFT) し，その対数成分を抽出する．これをさらに逆離散フーリエ変換 (Inverse DFT; IDFT) すれば，ケプストラムと呼ばれる特徴量が得られる．音声信号のスペクトルは音源信号に声道フィルタの伝達特性が畳み込まれたものとして考えることができる（放射特性は顕著な共振を持たないため，ここでは無視する）が，声道フィルタのスペクトル（音声のスペクトル包絡）は音源信号のスペクトルと比べ周波数に対し滑らかに変化する．このことから音声信号をケプストラム分析すれば，声道フィルタの対数振幅スペクトルの（逆）フーリエ変換は低い帯域に，音源信号の対数振幅スペクトルのフーリエ変換は高い帯域に集中することになる．よってこのケプストラムに低域通過フィルターを用いた後で再度フーリエ変換をすれば，声道特性に対応する対数スペクトル包絡を取り出すことができる．

人間の聴覚特性に合わせてスペクトルを周波数分解しケプストラム分析すれば，より人間の感覚に合った特徴量を抽出することができる．人間の周波数分解能は周波数に対する対数関数で近似でき，低い周波数ほど分解能が高く，高い周波数ほど低い．このような人間の感覚はメル尺度と呼ばれるが，ケプストラムにメル尺度を反映させた特徴の一つがメル周波数ケプストラム係数 (Mel-Frequency Cepstrum Coefficient; MFCC) である．MFCC の抽出にはまず図 3.3 のようにメル周波数 f_{mel} 軸上に等間隔で配置された三角窓を用意する．各窓に対応する周波数帯域のスペクトルを窓の大きさにより重み付けして和を計算し，それを IDFT することで MFCC が得られる．

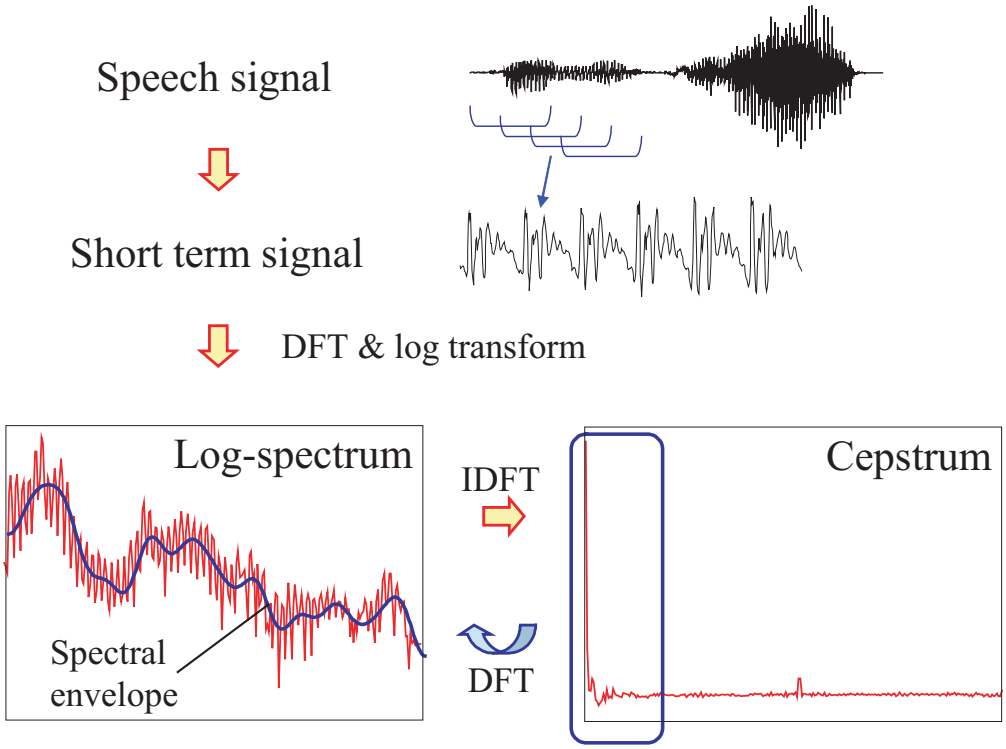

図 3.2: 音声信号からケプストラムを抽出する過程

3.3　話者の自動クラスタリングのための二話者間の英語発音距離の定義

3.2 では，母語や居住地の情報に基づいて話者にあらかじめ付けられた訛りラベルの情報を学習に利用し，訛りの分類を試みている．しかし，厳密には訛りは各自の英語学習方法や生活スタイル，家族や友人・英語教師といった周りの人間が有する訛りなどに影響される．各自の訛りは個人により異なるもので，母語・居住地ばかりで決定付けられるものではなく，また自身の訛りを自覚するのも困難であると言える．[21] では，ある米語方言データベースにおいて，話者が自称する方言と実際の方言が多くの話者で違っていることが指摘されている．

そこで本研究では，各話者の個人情報を基にしてトップダウンに発音訛り分類を行うのではなく，話者の発話の情報から個々の話者間で直接発音の類似度を量り，それに基づいて話者をボトムアップにクラスタリングすることを考える．機械学習の点から言えば，話者のラベル情報を利用しない教師なし学習による分類を行うことになる．具体的には，話者の発音の IPA 書き起しを比較することで，任意の二話者間について発音の違いの度合いを表す “発音距離” を算出し，距離に基づいて分類することを検討している．

発音の書き起しにおいては，性別や年齢といった英語発音訛り以外の話者性の情報は捨象されている．そのため，二話者の発音の距離として書き起し間の差異を定量的に評価できれば，これを発音訛りに関する正解の基準距離として採択し，回帰モデルの学習や評価に使用することができる．

発音距離を推定するタスクに関しては，計算機言語学 (computational linguistics) の研究者は，IPA の書き起し間の距離を推定する問題として捉え，音声工学の研究者は，書き起しを用いず，音声信号だけを用いて発音距離を推定する問題として捉え，種々の検討を行っている．以下，前者の研究例から紹介する．3.3.1 では，音声学的な知識に基づいて距離を定義している．3.3.2 及び 3.3.3 では，人間の感覚に根拠のある距離を多数の発音書き起しから機械学習により自動的に獲得することを試みている．3.3.4 では距離定義に単音音響モデルを用いている．本研究では，3.3.4 で算出される距離を基準距離としている．

また 3.3.5 では，発音距離としての妥当性を確認するための一つの検証として，各発音距離定義に

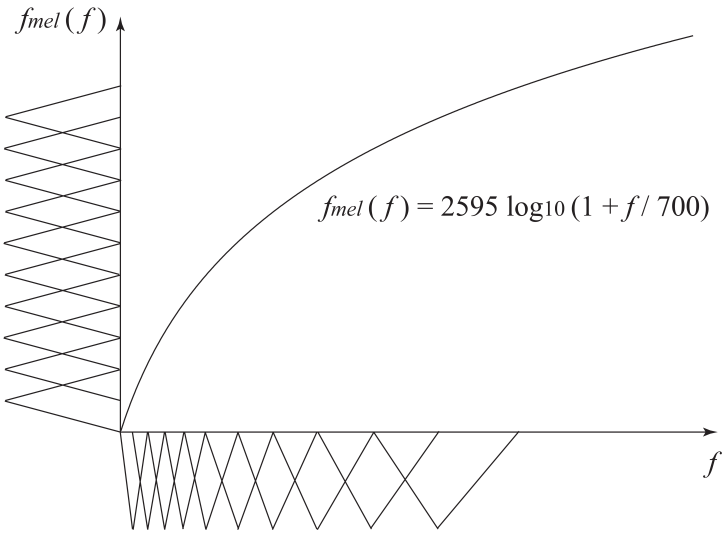

図 3.3: メル周波数とその軸上に等間隔で配置された三角窓

表 3.1: 弁別的素性のスコアの一例 [27]

音節性	5	高段性	5
有声性	10	鼻音性	10
側音性	10	有気性	5

より任意話者から米語母語話者への距離を求め，その値が母語話者が主観的に感覚する外国語訛りの大きさ [25, 26] とどれだけ相関を持つかについて調査している．

なお，韻律に関しては， IPA では書き起しの対象としていないため，以下の各距離定義においても考察対象とはしていない．

3.3.1　弁別素性に基づく距離定義

[27] では弁別的素性に表 3.1 に示すようにスコアを付け，この値を組み合わせて各単音をベクトル化する．このベクトルの差を単音間距離とし， DP マッチングにより単音を対応付け，累計により書き起し間距離を定義した．弁別的素性のスコアは [28] に基づいて付けられている．

3.3.2　自己相互情報量を用いたレーベンシュタイン距離

レーベンシュタイン距離は元々文章の類似度を表す数値であるが，[29] にてアイルランドの方言の比較のために用いられて以来，発音距離を測るために頻繁に適用されている．

ある二つの書き起し間の LD は，一方の文字系列からもう一方の対応する系列に変換するためにかかる最小の変換コストで表される．通常の LD では，文字の挿入，削除，置き換えという各操作のコストを全て 1 として算出される．変換を求める際に書き起し間で文字の対応付けが行われるが，基本的に母音は母音にのみ，子音は子音にのみ対応付けられる．

通常の LD では二系列で対応する単音が一致するかしないかのみが考慮されていて，置き換えが音が近い単音同士（例えば [i] と [ɪ] ）で起こっているか，または音が遠い単音同士（例えば [i] と [a]）で起こっているかで算出される発音距離を区別することができない．感覚に対しより根拠のある発音距離を求められるように，各変換操作のコストを自己相互情報量 (Pointwise Mutual Information; PMI) [30] に基づいて自動的に更新する手法が [31] で提案されている．以下が単音間変換コストの更新アルゴリズムである．

1. 全変換操作の初期コストを 1 として，LD の算出アルゴリズムにより文字の対応付けを行う．
2. 各文字 x について， x の挿入，削除，または x からの（ x への）置き換えの回数を全操作回数で割り，変換操作における x の生起確率 $P(x)$ を求める．
3. 各文字ペア (x, y) について， x または y の挿入，削除（これらの場合もう一方の文字は空白に相当），または x から y への（ y から x への）置き換えの回数を全操作回数で割り，変換操作における x と y の同時確率 $P(x, y)$ を求める．
4. 各文字ペア (x, y) について．式 (3.1) より PMI スコアを求める．

$$PMI(x, y) = \log_2 \left(\frac{P(x, y)}{P(x)P(y)} \right) \tag{3.1}$$

5. PMI スコアの符号を反転させ 0 から 1 の範囲で正規化することにより，各文字ペア (x, y) の PMI スコアに基づく文字間変換コストを求める．
6. 5 で求まった変換コストを用いて，LD の算出アルゴリズムにより文字の対応付けを行う．対応付けが更新されなくなるまで 2–5 を繰り返す．

x と y が真に相関を持ち共起しやすい ($PMI(x, y) \gg 0$) ということは，x と y の音声が近いことを表す．この場合文字間変換コストは 0 に近くなる．逆に音が共起しにくく音が遠い場合は，変換コストが 1 に近くなるため，LD を算出する際に二文字間の対応付けは起きにくくなる．

3.3.3　Naïve Discriminative Learning を用いた発音距離定義

ある一つの方言の話者らの感覚を基準とした，各発音を聴いた時の了解度を定量的に求め，二発音の了解度の差を発音の距離とする手法が [26] で提案されている．

ある単語を発話した時の発音（単音系列）の了解度は，その単語と各単音（単音組み合わせ）の関連度から推定できる．例えば， with という単語の発音書き起しが [wɪθ] である時，with と w+ɪ ，with と w-ɪ+θ ，with と ɪ-θ の関連度の合計が with の了解度となる．

単語と各単音組み合わせの関連度は Naïve Discriminative Learning (NDL) [32] により算出される．ある単語 O について，全発音書き起しから見られる単音組み合わせの種類数を n とすると，i 個目の単音組み合わせ C_i の関連度 V_i は式 (3.2) の Danks の平衡方程式 [33] から求められる．

$$P(O|C_i) - \sum_{j=0}^{n} P(C_j|C_i)V_j = 0 \tag{3.2}$$

NDL はこの方程式の解を効率良く推定するアルゴリズムである．NDL 計算時に基準とした方言に近い発音であれば，その方言にとって馴染みのある単音系列が入力されることになり，単語との関連度が高く了解度も高いことになる．逆に基準とした方言に遠い発音であれば，その方言ではあり得ない単音組み合わせが並ぶことになり，関連度が下がり了解度は低くなる．

NDL に基づいて発音距離を求める場合，どちらの話者の感覚を基準にするかで求まる距離は変わるため，この発音距離は非対称となる．

3.3.4　単音音響モデルを用いた距離定義

[3] では，二話者間の基準発音距離を，Dynamic Time Warping (DTW) [34] により求まる二話者の IPA 書き起しの単音系列整合コストとして定義した．単音系列の比較は単語毎に行われる． DTW の計算には局所コストが必要であるが，これには単音音響モデルから求まる単音間距離を用いている．本稿においても，この手法により算出される話者間距離を正解の距離として採用し，回帰の学習・評価に用いることにする．

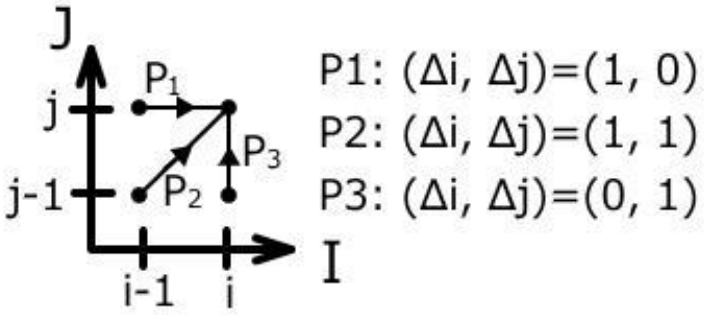

図 3.4: 選択可能な経路

図 3.4 は採択した DTW パスである．P_1, P_3 の経路は単音の挿入，削除に相当し， P_2 は単音の一致もしくは置き換えに相当する．同単語の二つの単音系列を $a_1, ..., a_i, ..., a_I$, $b_1, ..., b_j, ..., b_J$ とすると，式 (3.3) より，(i, j) での最小累積距離 $DTW[i, j]$ が計算できる．

$$DTW[i,j] = \min \begin{pmatrix} DTW[i-1,j]+d(a_i,b_j), \\ DTW[i-1,j-1]+2\times d(a_i,b_j), \\ DTW[i,j-1]+d(a_i,b_j) \end{pmatrix} \tag{3.3}$$

$d(a_i, b_j)$ は a_i, b_j の単音間距離である．最終的に，$DTW[I, J]/(I + J - 1)$ が求める単語間距離である．

LD とは異なり，挿入，削除，置き換えのコストをヒューリスティクスに基づいて設定する必要は無いが，局所コストとして全単音間の距離を定義する必要がある．[3] では実験に使用された SAA 話者の発音書き起しに出現する最頻 95% に相当する 153 種類の IPA（付録 A）を抽出し，男性の音声学者（本稿では P01 と称する）にそれらの IPA が示す単音を 20 回ずつ発声させた．この録音データから 3 状態 1 混合の単音 モノフォン HMM を構築し，二つの単音 HMM で対応する状態間のバタチャリヤ距離 (BD) の平均を単音間距離とした．残りの 5% の IPA は，修飾記号なしの IPA など音響特性が近いと思われるものに置き換えた．

前述の計算は全て一人の音声学者による発話音声を用いているため，単音 HMM 及び単音間距離は，この音声学者固有の声色や，発声の癖への依存性があることは否めない．今回，本研究で提案する発音距離の予測技術が，異なる基準距離に対しても高精度に予測できることを検証するために，新たにもう一人の男性の音声学者（以下， P02 と称する）に発声を依頼し，別の基準距離を算出して比較した．

二名の発話音声をそれぞれ用いて作成した単音間距離間の相関は 0.657 であった．各単音間距離を用いて算出した話者間の発音距離についても両者の相関をとると， 0.913 となった．

3.3.5　母語話者らしさの主観評定値を用いた発音距離の妥当性の検証

上記のように書き起し間の比較で計測された発音距離の妥当性は，例えば専門家の聴取によって定量的に定義された二話者間の発音距離との整合性によって評価されるが，そのような実験環境を整えることは難しい．そこで本稿では妥当性の一つの検証として， [25, 26] で行われているように，任意の話者 X の発音と母語話者発音との距離を，話者 X の発音に対する “母語話者らしさ” の主観的評定値と比較する．なお，母語話者らしさが正しく自動推定できることは，適切な発音距離予測の必要条件でしかないということには注意すべきである．

評価に用いる話者それぞれについて，複数の米語母語話者からの平均発音距離を求める．母語話者らしさの主観スコアが低いことはその話者の発音が母語話者から離れていることを意味するので，スコアが小さい話者ほどこの平均距離が大きくなる（スコアと平均距離の相関係数が -1 に近くなる）ような距離定義が妥当であると言える．主観実験では，各音声を米語を母語とする 1,143 人に聴かせ，それぞれの音声がどれだけ母語話者に近い発音かを 7 段階でスコア付けさせている [25]．スコアが 7 である時，最も母語話者発音に近い．[25] で利用された SAA 話者 286 人を対象に，各距離定義を用いて SAA 内の母語話者 115 人からの平均発音距離を算出し，主観スコアとの相関を調べた．

表 3.2: ある話者の母語話者発音からの平均距離とその話者の母語話者らしさの主観スコアとの相関

	P01
BD	-0.79

	Baseline
PMI	-0.77
NDL	-0.75

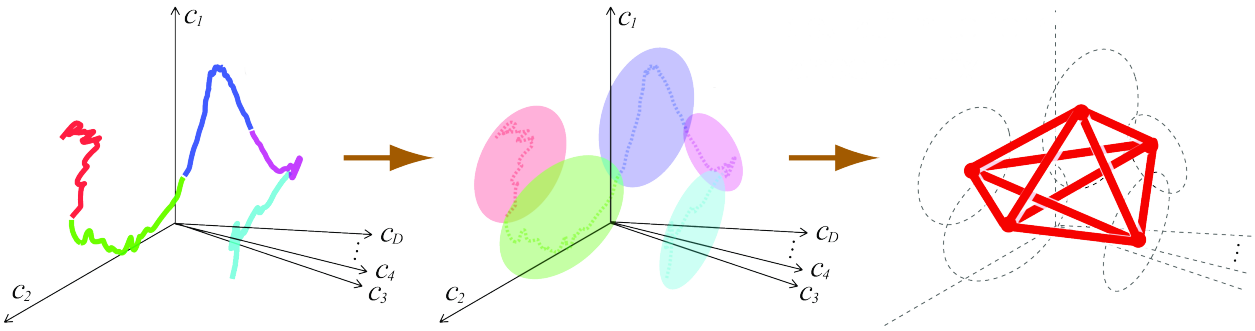

図 3.5: 発声を構造で表象する過程の概念図

他の研究では， PMI を用いた場合の相関は -0.77 [25], NDL を用いた場合の相関は -0.75 [26] となっている．これらの値を比較のためのベースラインとする．本稿では， P01 音声のモデルを用いた単音間距離を使った発音距離を評価した．表 3.2 に結果をまとめる．

P01 の相関は PMI, NDL を用いた結果よりも少し高いものとなっている．PMI, NDL は知覚的な根拠のある距離定義である．本稿でも利用する音響モデルに基づく発音距離が，それらの距離と同等かそれ以上の評価となる結果が得られ，基準距離としてある程度の妥当性があることが示された．

3.4　音声の構造的表象を用いた発音距離予測

3.3 で求まる話者間の発音距離により，人間の感覚に対しある程度妥当性のある話者クラスタリングを行なうことが可能である．しかし，この距離を算出するためには，対象とする話者の書き起しが必要である．世界 15 億人の英語話者を自動クラスタリングという本研究の目的に立ち返ると，世界中の全英語話者について，音声学者に発音書き起しを作成してもらうことは非現実的である．そこで [3] 及び本稿では，一部の発音書き起しから得られる基準距離を学習に利用したサポートベクター回帰により，発音書き起しが未知の話者対についても音声情報のみから距離を予測することを試みている．ここでは， [3] で回帰の入力として利用している音声の構造的表象について説明し，構造的表象を利用したその他の先行研究を紹介する．

3.4.1　音声の構造的表象

音声から英語発音の音響モデルを訛り毎に構築し，各モデルのパラメータの違いから発音距離の予測に有効な特徴を取得することを考える．発音距離予測は，年齢や性別といった訛り以外の声色に由来する音響変動に頑健であることが望ましい．話者固有の声色の影響を音響モデルに反映させたくない場合，通常，多数の話者の音声を用いることにより，声色以外の要因を隠れ変数として捉える．しかし本研究での発音距離予測においては，話者は一人一人異なる英語発音を有しているとみなしている．そのため，発音のモデル化は話者単位で行わなければならない一方で，体格などに起因する話者固有の声色を除去して行う必要がある．そこで，発音モデルには声色の影響を残し，モデルからの特徴量抽出において話者の声色の違いに頑健となるものを取得することを考える．[3] では，発音の構造特徴 [35, 36, 37, 38] を利用している．

ある一発声から発音構造を算出する時の概念図を図 3.5 に示す．発声（音響イベント列）中の各イ

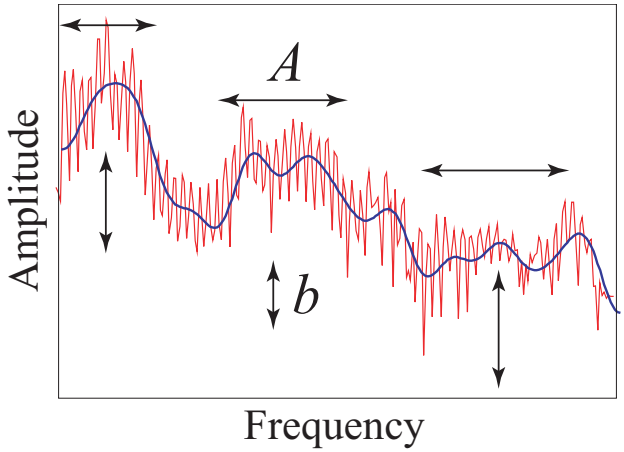

図 3.6: スペクトルに対する線形変換性歪み (A) と乗算性歪み (b)

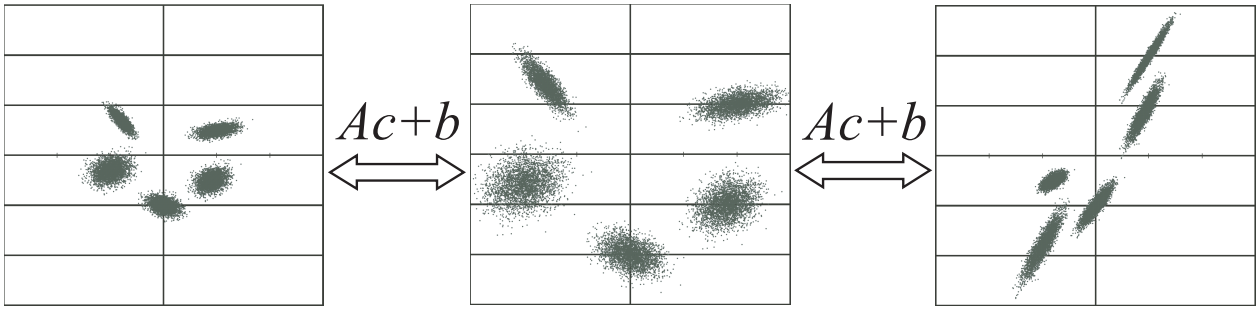

図 3.7: アフィン変換による分布群の変化（これらは全て同一の構造を持つ）

ベントを分布で表現し，任意の二分布 p_1, p_2 について，式 (3.4) により f-divergence を求める．

$$f_{div}(p_1, p_2) = \int_{\mathcal{X}} p_2(x) g\left(\frac{p_1(x)}{p_2(x)}\right) dx \tag{3.4}$$

$g(t)$ は $t > 0$ の凸関数で，$g(t) = \sqrt{t}$ の時 $-\ln(f_{div})$ は BD となる．

$$BD(p_1, p_2) = -\ln \int_{\infty}^{\infty} \sqrt{p_1(x) p_2(x)} dx \tag{3.5}$$

二分布 p_1, p_2 がそれぞれガウス分布である時，BD は式 (3.6) のように簡便な形に展開できる．

$$BD(p_1, p_2) = \frac{1}{8}(\mu_1 - \mu_2)^T \Sigma^{-1} (\mu_1 - \mu_2) + \frac{1}{2} \ln\left(\frac{\det \Sigma}{\sqrt{\det \mu_1 \det \mu_2}}\right) \tag{3.6}$$

但し，$p_1 = \mathcal{N}(\mu_1, \Sigma_1), p_2 = \mathcal{N}(\mu_2, \Sigma_2)$ とし,，$\Sigma = \frac{\Sigma_1 + \Sigma_2}{2}$ とする．

言語的内容が同じでも話者の性別や体格が異なれば個々の音の物理特性は変わる．しかし，各音の間の関係性のみを考え発声を構造的に捉えれば，この構造は話者の声色に対しほぼ不変となる．この事実を数学的に解釈し表現したものが，発音の構造的表象である．

音声に不可避的に混入する非言語的特徴を大別すると，主に加算性歪み，乗算性歪み，そして線形変換性歪みの 3 種類になる [38]．加算性歪みは背景雑音などに由来するもので，音声の周波数スペクトル領域での加算で表現される．この歪みは，雑音の少ない環境で録音するなどすれば取り除ける．乗算性歪み (b) はマイク特性などに由来するもので，スペクトル領域では図 3.6 のように加算に変換される．また線形変換性歪み (A) は声道長や声道形状の差異など話者の違いに由来するもの

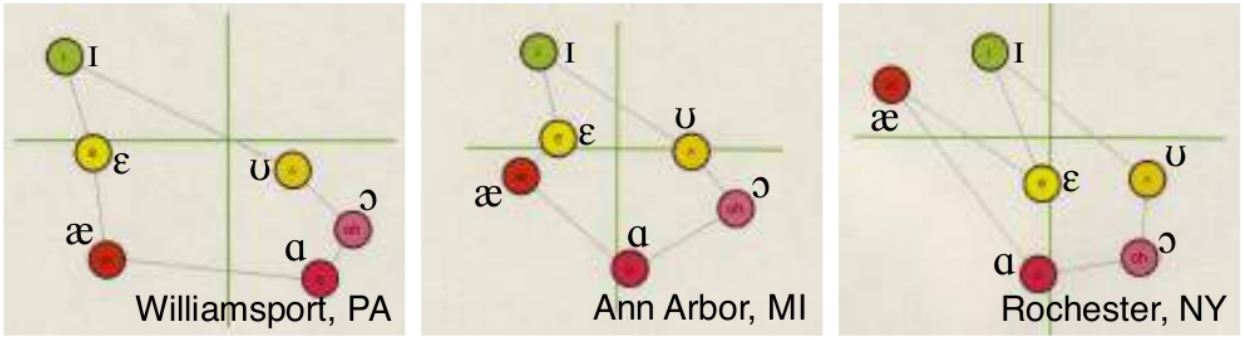

図 3.8: 米国方言における母音配置（ただし一部）の差異 [40]

で，図 3.6 のようにフォルマント周波数が周波数軸方向で移動することを表す [39]．これらをまとめると，非言語的特徴による歪みは，ケプストラム c に対するアフィン変換 $c' = Ac + b$ で近似できる．構造的表象では発声を f-divergence のみで表現したが，f-divergence はアフィン変換に対する不変性が証明されている [36]．図 3.7 に，アフィン変換による分布群の変化を示す．これらは一見全く異なる分布群に見えるが，その分布間距離は不変で，すなわち分布群の構造は不変となっている．

発音構造は声色に不変となるが，方言によっては多様に変化することが分かっている．図 3.8 はいくつかの米語方言について，声道長正規化をした後の単母音の平均点を第 1, 2 フォルマント周波数平面に配置したもの [40] で，音の配置が方言によって非線形的に変化することを示している．このことから，構造特徴が本研究の目的である発音分類に対し有効な特徴であることが予想される．発話内容を揃えた上で発音が同じであれば，発音構造は話者性に対する非依存性が高くなる．逆に，話者間で発音構造に違いがあれば，その差異は発音の違い（訛りの違い）の特徴として利用できると考えられる．

以下，音声の構造的表象を利用した先行研究をいくつか紹介する．

3.4.2　音声の構造的表象を用いた英語学習者の発音分類

[41] で朝川らは，構造的表象に基づいて，英語学習者の発音分類というタスクを設定し実験を行った．

朝川らはまず各被験者に， 11 種類の英語母音 (/ɑ, æ, ʌ, ə, ɚ, ɪ, i, ʊ, u, ɛ, ʉ/) と 5 種類の日本語母音 (/a, i, u, e, o/) を発声させ収録した．そして表 3.3 を用いて，11 種類の母音のうち一部の発音が日本語母音に置き換わった 8 種類の発音の状態（表 3.4 の S1 から S8）を定義し，各発音状態の母音組合せで発音構造を算出した．このように同一話者で英語と日本語の母音発声を交ぜ合わせた模擬的な音声を用いることで，発音構造における発音状態を既知とすることができる．これらの発音構造を対象に発音クラスタリングを行うことで，クラスタリング結果を発音状態をもとにして評価することを可能にしている．

帰国子女あるいは英語劇経験者の日本人 12 名（男性 6 名，女性 6 名）の母音音声を収録した．収録音声から表 3.5 に示す音響分析条件と MAP (Maximum A Posterior) 適応により発音をモデル化し，音素間距離を BD によって求め，表 3.4 の各状態に基づいて各話者毎に 8 種類の発音構造を算出した．これらの発音構造に対して，二構造間の距離を式 (3.7) により算出し，構造のボトムアップクラスタリングを行った．

$$D = \sqrt{\frac{1}{M}\sum_{i<j}(P_{ij} - Q_{ij})^2} \tag{3.7}$$

ここで， X_{ij} は話者 X の発音構造距離行列の (i, j) の要素を，M は音素数を表す．

比較のため，各話者の対応する発音モデル間の距離を式 (3.8) により直接求め同様のクラスタリン

表 3.3: 母音置換の組合せ [41]

Japanese vowels	English vowels
a	ɑ, æ, ʌ, ə, ɚ
i	ɪ, i
u	ʊ, u
e	ɛ
o	ɒ

表 3.4: 発音状態の定義 [41]

	ɑ	æ	ʌ	ə	ɚ	ɪ	i	ʊ	u	ɛ	ɔ
S1	J	J	J	J	J	J	J	J	J	J	J
S2	E	E	E	E	E	J	J	J	J	J	J
S3	J	J	J	J	J	E	E	E	E	E	E
S4	E	E	J	J	J	E	E	J	J	E	E
S5	J	J	E	E	E	J	J	E	E	J	J
S6	E	J	E	J	E	J	J	J	J	E	E
S7	J	E	J	E	J	E	E	E	E	J	J
S8	E	E	E	E	E	E	E	E	E	E	E

E:英語の母音発声を使用, J :日本語の母音発声で置換

表 3.5: [41] の実験における音響分析条件

サンプリング	16 bit / 16 kHz
窓	25 ms length / 10 ms shift
特徴量	FFT ケプストラム 10 次元
混合数	1
状態数	3

グも行った．

$$D' = \sqrt{\frac{1}{M}\sum_i BD(v_i^P, v_i^Q)} \tag{3.8}$$

ここで，v_i^X は話者 X における母音 i のモデルを示す．

2 種類の発音間距離に基づくクラスタリング結果を，それぞれ図 3.9，図 3.10 に示す．樹形図の葉ノードの数字は表 3.4 の各状態を表し，A〜L は話者を表している．直接比較した距離によるクラスタリングではほぼ話者により分類されている（図 3.9 ）のに対して，発音構造の距離によるクラスタリングでは話者と無関係にほぼ発音状態により分類されている（図 3.10 ）．これらの結果より，構造的表象を用いることで，確かに話者の性別や年齢といった声色に影響されることなく，発音状態（発音訛り）によって話者を分類することが可能となることが実験的に示された．

3.4.3　音声の構造的表象を用いた中国語の方言分類

[42, 43] で Ma らは，構造的表象を用いて中国語の方言で話者を分類することを検討している．実験では，七大方言の分類だけでなく，さらに区分を細かくした下位方言の分類まで行なっている．

Ma らの実験では，各話者に単一の漢字を複数発音してもらい，それぞれの音声毎に発音構造を算出する，分類時には，発音構造の差を定量的に求め，それを話者間距離としてクラスタリングを行なっている．二話者の発音構造の差は式 (3.7) で定義されている．なお，ここでの M には各漢字の音節の数を用いる．

クラスタリングの結果，性別と無関係にほぼ方言で話者が分類できたことが報告されている．

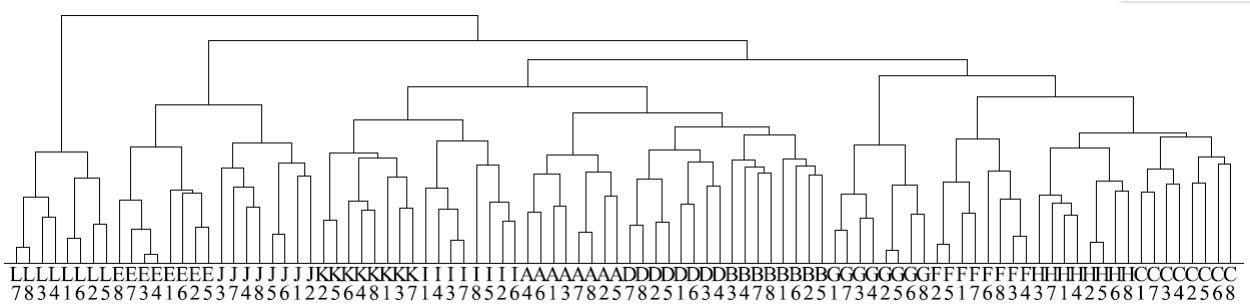

図 3.9: 音響的実体に基づく学習者発音構造のクラスタリング [41]

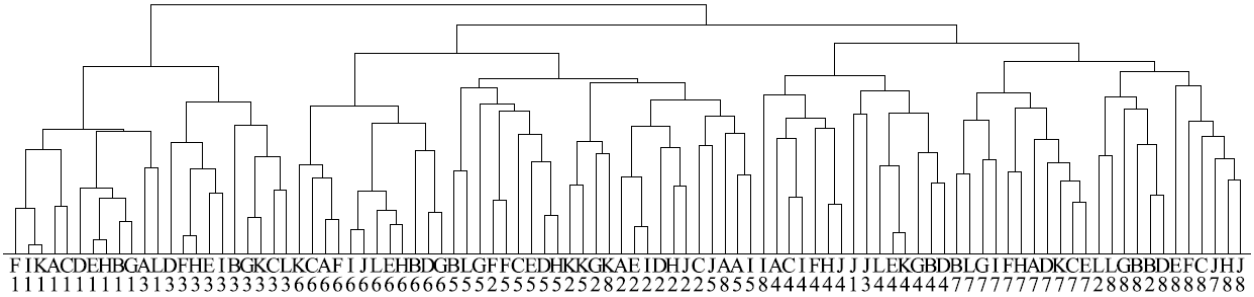

図 3.10: 構造に基づく学習者発音構造のクラスタリング [41]

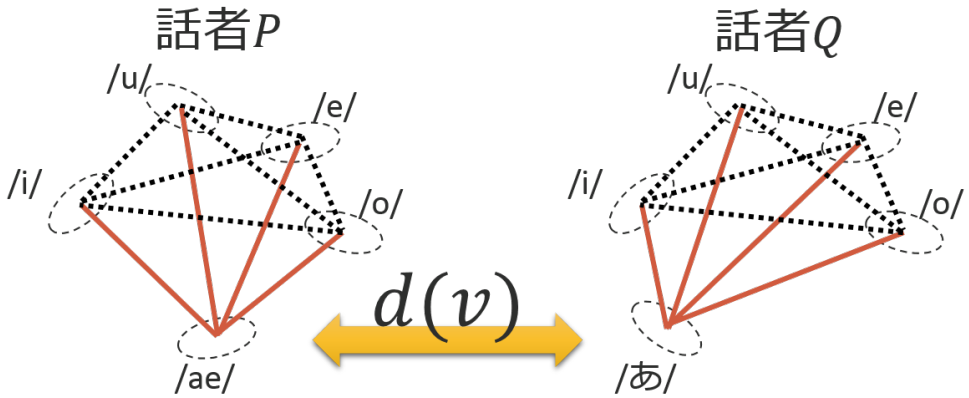

図 3.11: 二話者の発音構造間の構造歪み

3.4.4　発音構造間の構造歪みを用いた発音教示

[41] で朝川らはまた，英語学習者と教師の発音状況を比較し，学習者がどの音素の発音を矯正するのが教師の発音に到達するために効果的であるのかという発音教示について実験を行った．これは，構造的表象を利用し，学習者と教師の発音構造を比較し，どの音素が二話者の発音構造形状の差異（構造歪み）を特にもたらしているかを求めることにより可能となる．二名の話者の発音構造距離行列を P, Q とすると，朝川らは，式 (3.9) により，母音 v に起因する二構造間の構造歪み $d(v)$（図 3.11）を定義した．

$$d(v) = \sum_{i=1}^{M} |P_{vi} - Q_{vi}| \tag{3.9}$$

朝川らは模擬音声を用いた実験で，構造歪みを算出することで致命的な発音誤りを提示することが可能であることを示した．

3.5　おわりに

本節では発音に基づく話者分類に関する先行研究について説明した．3.3 では発音書き起しを用いた話者間発音距離の定義をいくつか紹介し，その中で本研究で採用する 3.3.4 の距離定義が知覚に対し根拠のある他の距離に劣らない妥当性を持つことを示した．音声を用いた発音距離予測について，3.4 で発音構造を用いた先行研究を示し，また同時に発音構造を用いた実験例をいくつか示した．次節では，[3] の距離予測手法の高精度化に向けた手法の改善と，いくつかの分析的検討について述べる．

第4章

発音構造を用いた英語発音距離予測の高精度化と分析

4.1　はじめに

本節では，[3] で行なわれた実験について，実験設定に関する考察と特徴量算出手法の変更を行なう．4.2 で述べるように問題のある話者のデータを実験対象から除外し，4.5 で述べる種々の手法の改善を施すことにより，[3] の実験の枠組みで距離予測精度が大幅に上がり，4.4 の理想的なベースラインシステムの性能を越えられる結果を示すことを実証する．ただしこの結果は 4.3 で考察するように，話者の発音データが本来持つべき多様性・ばらつきを過小評価した設定の上で成り立っている．本稿ではより実用状況に即した実験条件を提案し，この条件下では提案手法で実用に足る距離予測精度が出せていないことも報告する．4.5 では，特徴選択による分析的検討について報告している．

4.2　実験に使用するデータの選択

[3] では，SAA の話者のうち，発音書き起しがなされていて背景雑音が少なく，且つ発音書き起しの単語数がちょうど 69 単語であるものを実験の対象としていた．しかし SAA では，言い間違いやフィラーなどについても書き起しの対象としている．そのため書き起しが同じ 69 単語でも 他話者と単語の対応付けがとれない話者が存在することになる．[25] では，単語の言い間違いや言い直し，脱落を手動で修正した上で実験を行っている．本稿では，[3] の 381 人の話者のうち，69 単語を全て順番通り読み上げている話者 369 人のみを選択して使用する．369 人の話者は付録 B に列挙してある．

4.3　距離予測実験における open 性に関する考察

回帰処理を行う場合，学習データと評価データを分割し open とすることが一般的である．[3] では，72,390 ($= 381 \times 380/2$) 通りある話者対を，その基準距離に基づいて昇順に並べ替え，偶数／奇数番の話者対を抽出することで，36,195 ずつに二分し，学習データと評価データに分け 2-fold の交差検定を行っている．この実験では，話者対 open であるが，学習・評価データに話者が共通に入っており，話者 closed な条件となっている．

本研究では回帰モデルとしてサポートベクター回帰を用いる．入力（観測）特徴量を高次元特徴量空間に写像し，その空間で，入力特徴量と個々の学習サンプルの特徴量との内積を元空間のカーネル関数を用いて計算する．内積値は類似度と解釈できるが，この類似度スコアの線形結合により回帰を行なう．話者対 open 実験では，話者対 A–B が評価セットにある場合，学習セットには A–$\{x\}(x \neq B)$, B–$\{y\}(y \neq A)$ が含まれることになる．この時，話者対 A–B の発音距離を予測する場合，学習セット中の話者対群 A–$\{x\}$，B–$\{y\}$ に対して，B に似た x，あるいは A に似た y が存在しているかどうかが回帰性能に影響する．

一方，学習データ，評価データに同一話者が全く含まれない，話者を単位とした openness を満たす条件下でも実験を行うことができる．この時，A–B が評価セットにある場合，学習セットには A–$\{x\}$, B–$\{y\}$ は一切含まれない．サポートベクター回帰を用いる場合，A–B の発音距離予測に対して，学習セットに話者対 A–B と似た話者対が存在しているかどうかが影響する．

いま，学習データに含まれる話者数を N とすると，話者対 open 実験では A–B に対して，A に似た話者あるいは B に似た話者が N の中に含まれるか否か，つまり，発音の多様性を $O(N)$ と評価できる枠組みでのタスク設定となっている．一方，話者 open 実験では，話者対 A–B に似た話者対が学習セットに含まれるのか否か，即ち，発音の多様性を $O(N^2)$ と評価するタスク設定となっている．話者 open 条件での回帰問題は，話者対 open 条件時よりもはるかに多くの学習データが必要となることが予想される．

話者対 open 実験と話者 open 実験の実用的な価値について考察する．前者の場合，評価セットの話者対において，どちらか一方は必ず学習データに含まれることになる．このタスクは，発音書き起しが用意された学習用話者群に対して一名の書き起し未知話者が与えられ，未知話者と学習データ中

の各話者の距離を予測する問題に相当する [1]．より具体的に考えれば，例えば，世界諸英語の分布状況を考慮してサンプリングされた各国，各地域の訛りを有する話者群の読み上げ音声と IPA 書き起しが（学習データとして）与えられた場合に，未知話者がどの話者とどのくらい離れているのかを予測する問題に相当する．一方本研究の最終目標は，世界諸英語全体を一望できる発音地図を作成する（世界中の英語話者を個人単位で分類する）ことである．この状況を想定し，限られた話者数の学習データを用いることで，複数の未知話者間の発音距離予測がどの程度の精度で行なえるのかを実験的に検証する必要もある．

以上の観点から，本研究では 2 通りの応用環境を想定し，話者対 open，話者 open の条件下で実験を行なう．話者 open 条件では， 369 人の話者を五セットに分割し，一セットの話者からとれる話者対を評価用のデータ，残り四セットの全話者からとれる話者対を学習用のデータとして， 5-fold の交差検定を行なう．この時学習データ中の話者対数は 43,660 または 43,365 となる．発音書き起しを用いた基準発音距離を回帰の学習と評価に利用し，用意したベースラインシステムと比較して評価する．

4.4　ベースラインシステムの構築（米語音素誤り認識器を用いた発音書き起しの自動化）

[3] では比較のためにベースラインシステムとして，各話者の発話音声から書き起しを自動作成し，書き起し間距離を自動計算することで，距離計算を全自動化したものを挙げている．本稿においてもこのベースラインシステムを採用する．

3.3.4 の発音距離の計算過程を再掲する．

1. 音声学の専門家による IPA 発音書き起こし
2. DTW による書き起こし比較と累積距離計算

このうち前者を，米語音素の発音誤り検出器を用いて自動化する [2]．誤り検出器の出力の正解データとして，各話者発音の米語音素による書き起しが必要である．本研究ではこれを，付録 C の変換表を用いた各 IPA から米語音素文字への単純なマッピングにより SAA の書き起しを変換して作成している．

自動音素誤り検出器を用いた場合，仮に検出性能が完全であったとしても，得られるのは音素書き起しに過ぎない．単音から音素への変換は抽象化であり，この過程で音声学的情報はいくらか失われる．ここで，SAA の書き起しの単音音素変換で作成した正解データを直接用いて発音距離算出を行うことで，完全音素誤り検出器による距離予測の性能を評価した．DTW の計算で用いる音素間距離行列は，[44] のモノフォン HMM の音素モデル間のバタチャリヤ距離 (BD) を使用した．この時，SAA 369 人からなる全話者対について完全音素誤り検出器に基づく距離と (P01) IPA 基準発音距離との相関を求めたところ，0.756 であった．

次に，実在の音素誤り検出器を用いた場合の距離予測の性能を示す．認識の音響モデルとして，[44] のモノフォン HMM を初期モデルとし，369 人の全読み上げ音声を用いて追加学習したものを使用する．得られた音響モデルと，発音誤りを考慮した認識文法を用意することで，自動音素誤り検出が実現される．具体的な認識文法としては図 4.1 に示すように，369 人内で見られる各単語の音素系列で構築したネットワーク文法を使用する．実験の結果，得られた音素系列に対する音素正解率は 73.5%であった．また，IPA 基準距離に対する相関は 0.449 となった．音素誤り検出の精度は，発音距離推定に大きな影響を及ぼすことが分る．

[1] 最小距離を示す話者を結果として出力すれば，これは同定問題となる．

[2] 著者らの知る限り，単音の認識器は存在していない．そのため本研究においては，書き起こしの自動化は米語音素の自動発音誤り検出器を代用して行っている．

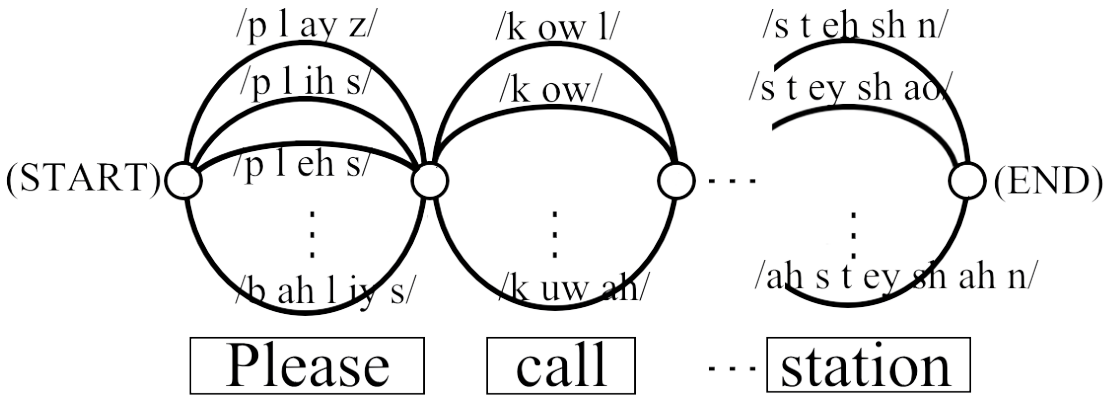

図 4.1: 単語ネットワーク文法

表 4.1: 音響分析条件

サンプリング	16 bit / 16 kHz
窓	25 ms length / 10 ms shift
特徴量	MFCC12 次元 + ΔMFCC
混合数	1
状態数	3

4.5　音声の構造的表象を用いた距離予測

4.5.1　先行研究における距離予測手法

[3] では，381 人分の SAA の同文読み上げ音声を次の 9 つのフレーズに分割し，それぞれで発音構造を算出し，9 つの発音構造を特徴として話者間の発音距離推定を試みた．

1. “Please call Stella.”
2. “Ask her to bring these things with her from the store:”
3. “Six spoons of fresh snow peas,”
4. “five thick slabs of blue cheese,”
5. “and maybe a snack for her brother Bob.”
6. “We also need a small plastic snake”
7. “and a big toy frog for the kids.”
8. “She can scoop these things into three red bags,”
9. “and we will go meet her Wednesday at the train station.”

[3] の実験での話者構造算出までの概略図を図 4.2 に示す．はじめに全音声を用いてフレーズを単位とした 9 つの Universal Background Model (UBM-HMM) を作成する．各音素セグメントは 3 状態でモデル化されている．音響分析条件を表 4.1 を示す．初期モデルは全音声の音響特徴量の平均から算出している（フラットスタート）．この UBM-HMM と MLLR (Maximum Likelihood Linear Regression) 適応により，各話者の発音を表す 381 人分の HMM を取得する．このように同一のモデル (UBM-HMM) をベースにして各話者発音モデルを作成することで，発音比較の際モデル間の時間的な対応付けがある程度正確なものとなるようにしている．MLLR 適応時に行われる状態クラスタリングでのクラス数は最大で 32 とする．同一話者の音素セグメントモデル間で距離を求め発声を距離行列で表すと，この行列が話者の発音構造に相当する．任意の話者 S の発音モデルにおける i 番目の音素セグメントと j 番目の音素セグメントとの距離 S_{ij} は，対応する状態間の分布間距離 BD の

平均の平方根とする．

任意の二話者 S と T の発音距離を予測する際の特徴として，式 (4.1) で求まる二話者の発音構造の差行列 $\{D_{ij}\}$ の各要素を利用する．

$$D_{ij} = \left| \frac{S_{ij} - T_{ij}}{S_{ij} + T_{ij}} \right|, \quad \text{但し } i < j \tag{4.1}$$

こうして得られる 9 つの差行列の全要素をサポートベクター回帰における入力特徴量とする．特徴量数は $2,804$ で，回帰は ϵ-SVR を用いている．カーネル関数としては放射基底関数 $K(x_1, x_2) = \exp(-\gamma|x_1 - x_2|^2)$ を適用している．

提案手法との比較のため，本稿で用いた 369 人の話者で [3] のフレーズ単位での構造特徴により発音距離予測実験を行なった．実験は話者対 open 条件で，学習・評価には P01 の基準距離を用いる．この時，予測距離と基準距離との相関は 0.781 となり，理想的な音素誤り認識器を用いたベースラインの性能をわずかに上回る結果となった．

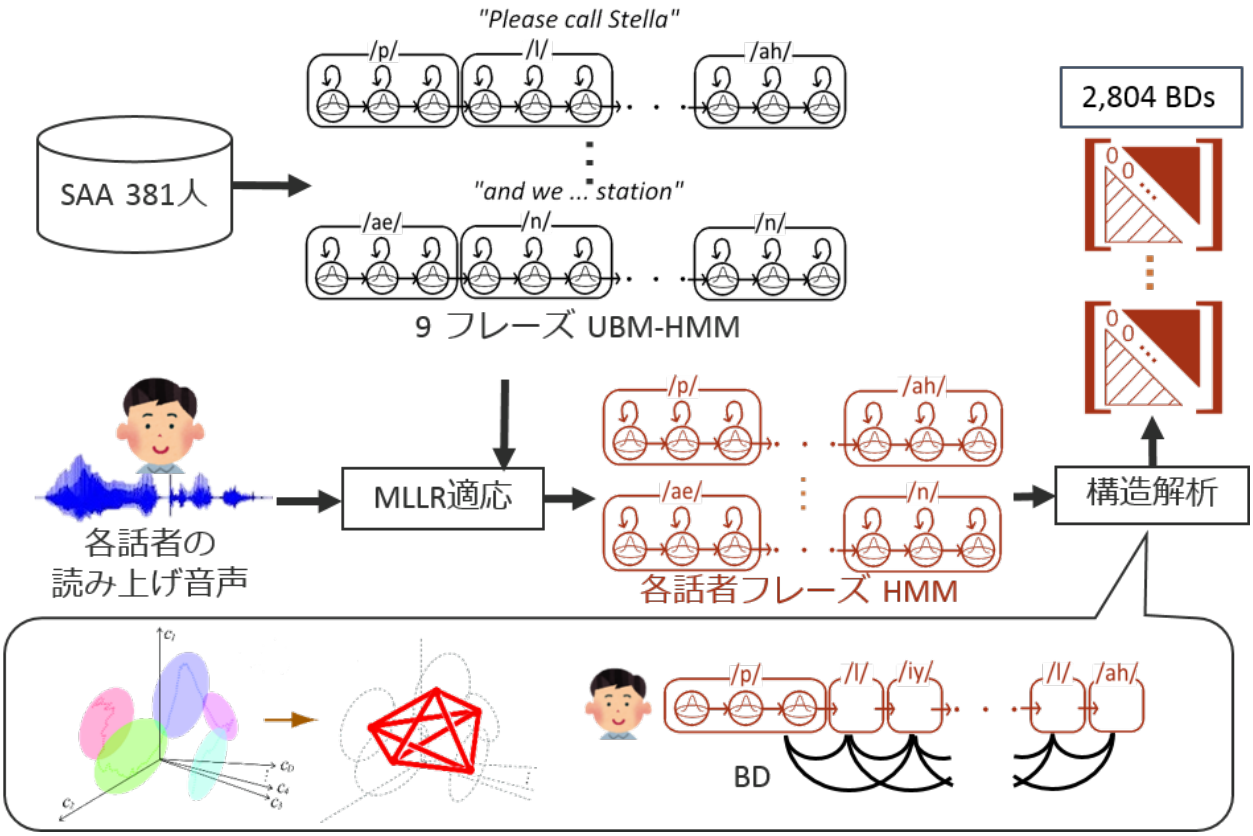

図 4.2: フレーズを単位とした構造算出 [3]

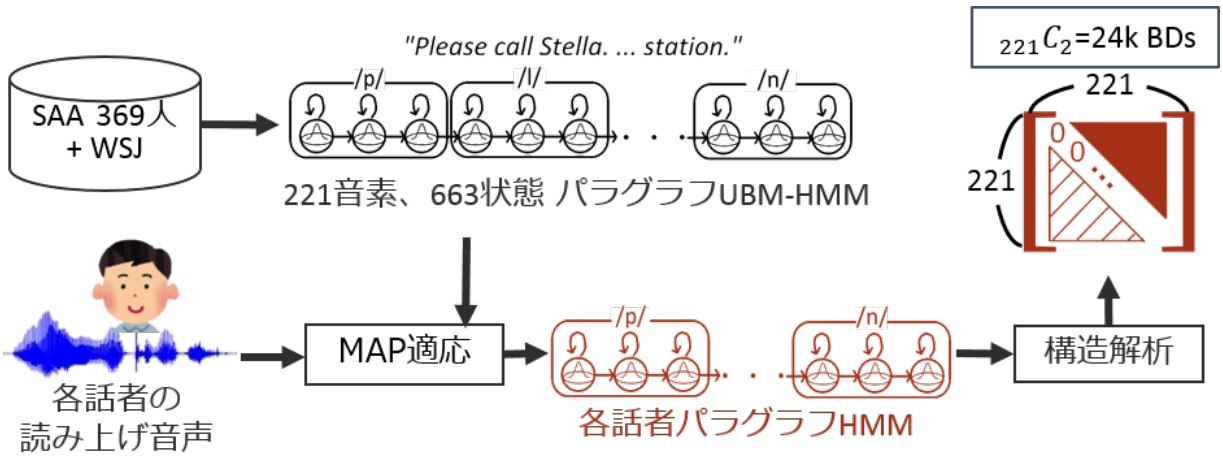

図 4.3: パラグラフを単位とした構造算出

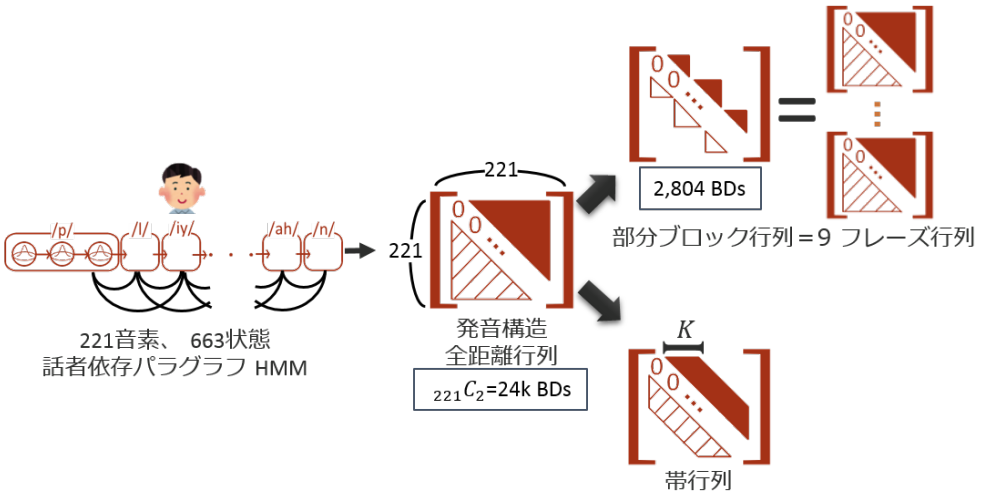

図 4.4: 部分ブロック行列（先行研究）と帯行列（提案）の利用

4.5.2　従来手法からの変更点

SAA では英語を母語としない話者も多数含まれており，それらの話者は言い詰まりが見られ単語間の無声区間が頻繁に発生している．本稿では，UBM-HMM の初期モデルとして Wall Street Journal を用いて学習した [44] のモノフォン HMM を使用することで，単語間の無声区間を sp (short pause) モデルで表現できるようにしている．

UBM-HMM から 各話者の発音モデルを作成する際の適応手法として， [3] では MLLR を使用していた． 適応に用いることのできるデータ（本研究では各話者につき一つ）が少量である場合，MLLR 適応を用いた方が MAP 適応を用いるよりも性能が良くなることが多い．しかし本研究の距離予測実験においては， MAP 適応の方が性能が良いことが実験により確認された．これより本稿では， MLLR 適応の代わりに MAP 適応を用いることにする．

[3] では差行列の計算に式 (4.1) を用いていた．これは線形回帰のための正規化として提案されている [37]．しかしここでの実験では距離推定にサポートベクター回帰を用いており，特徴量抽出の過程でも特徴量正規化は行われている．本研究では，正規化が二重になされることによる情報欠落が起こらないよう，$D_{ij} = |S_{ij} - T_{ij}|$ として発音距離推定を行っている．

以上の点を全て変更した上で，実験条件を同じくして SAA 話者 369 人を対象に話者対 open 実験を行ったところ，予測距離と基準発音距離との相関は 0.781 から 0.832 にまで上昇した．

4.5.3　パラグラフ全体を単位とした構造算出

本稿での HMM の音響分析条件を表 4.1 に示す．[3] では，SAA の読み上げ文を 9 つのフレーズに分けそれぞれで構造の算出を行っていたが，これでは文章全体からとり得る発音構造の距離行列のうち一部分のみしか利用していないことになる（図 4.4 ）．

本稿では，時間的に離れた分布間の構造も予測に有効であると考え，図 4.3 に示すように，文章をフレーズに分割せず，全体から構造を算出し距離予測を行う．文章単位で HMM を作成することで，全ての発音構造距離行列が得られ，これらを用いた差行列を入力として回帰を行う．実験では距離行列のうち幅 K $(1 \leq K \leq 220)$ の帯行列部分のみを使用した（図 4.4）．これは，SAA の文章を 221 個の米語音素系列で考えた時に，前後で K 個離れた音素モデルまでの分布間距離をとり構造を算出することに相当する． K を変えて距離予測を行い，時間的にどれだけ離れた発声間の差異までが予測に有効かを検証している．

話者対 open 条件で，帯行列の要素を入力特徴量とした場合の実験結果を図 4.5 に示す．図では，帯行列部分を利用する提案手法 (K-width contrasts) の他に，ベースラインシステムである完全米語

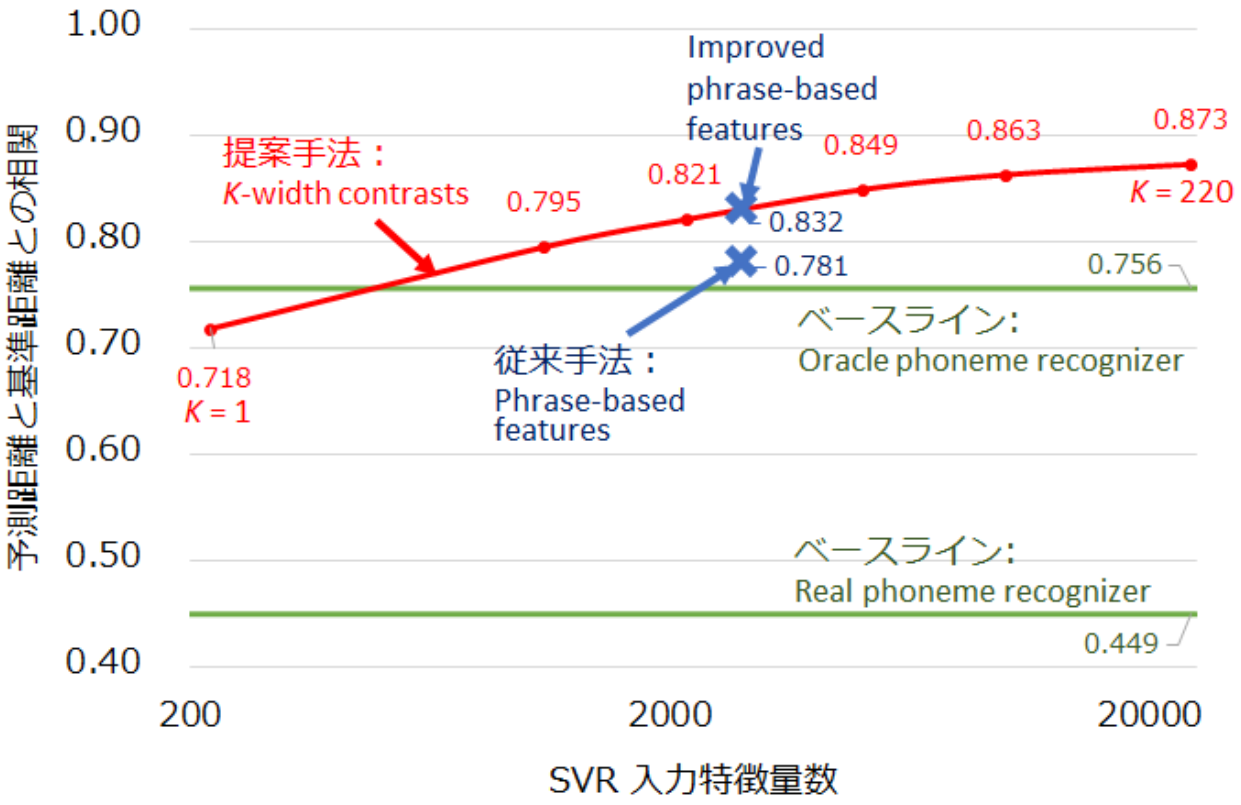

図 4.5: 話者対 open 条件における P01 距離予測でのベースライン，従来手法，提案手法の相関の比較

音素誤り検出器，及び実在の音素誤り検出器を用いた距離予測の結果 (oracle phoneme recognizer, real phoneme recognizer) と，従来手法のフレーズを単位として発音構造を算出し入力特徴とした場合の結果 (phrase-based features)，及び従来手法に 4.5.2 で挙げる変更を施した場合の結果 (improved phrase-based features) も同時に示している．

K-width contrasts について，帯幅 K を変え特徴量数を増やしていくと，K が最大の 220 になるまで相関は単調に上がり続けている．時間的に離れた発音間の相対特性も，全て訛りの特徴としてに有効であることが示された，

K-width contrasts と improved phrase-based features （提案手法と従来手法）を比較すると，ほぼ同じ特徴量数のところで相関に差はほとんど見られなかった．相関が特徴量数に応じて決まっており，発音構造距離行列のうち帯行列を利用する場合とフレーズに基づく部分を利用する場合とでは特徴量がほぼ同等の性質を持つことが分かった．

K-width contrasts で $K = 220$ の時，すなわち SAA パラグラフを単位とした発音構造特徴を全て用いた時相関は 0.873 となり，話者対 open 条件ではベースラインシステムである oracle phoneme recognizer を大きく超える距離推定性能を示した．

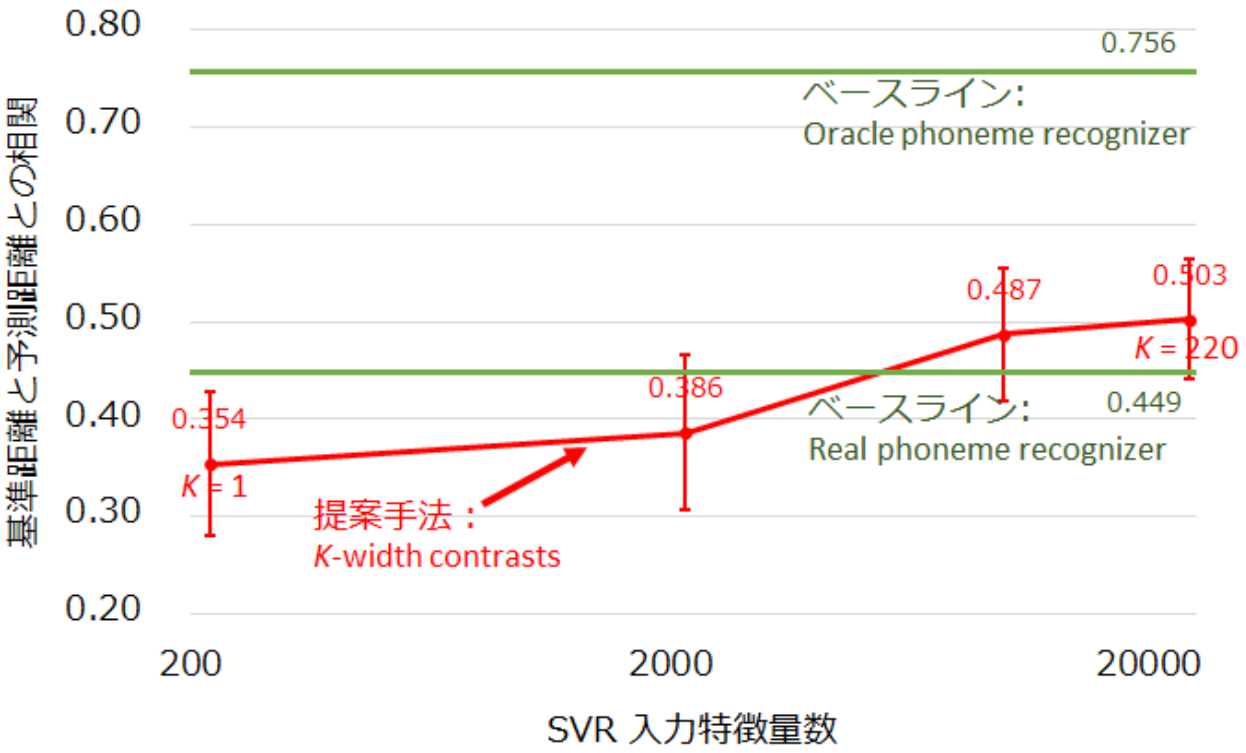

図 4.6: 話者 open 条件における P01 距離予測での ベースライン，提案手法の相関の比較

続けて，話者 open 条件で，帯行列の要素を入力特徴量とした場合の実験結果を図 4.6 に示す．

K-width contrasts の帯幅 K を変え特徴量を増やした時，話者対 open の結果と同様，K が最大 220 になるまで相関は上がり続ける．しかし，最終的な相関は 0.503 で留まり，ベースラインの real phoneme recognizer を超える予測性能ではあるものの，oracle phoneme recognizer に対しては大きく下回る結果となった．

4.5.4　異なる基準発音距離に対する距離予測

本研究で採用する基準発音距離は， 3.3.4 の手法で P01 一名の単音音声に基づき算出されている．ここでは， P01, P02 二名の音声から求まる二通りの基準距離それぞれを用いて実験を行い，どちらの定義による距離に対しても，発音構造差行列を入力とした SVR により同等の精度で予測が行えるかを検証する．

幅 K の帯行列の要素を入力特徴量とした場合の，話者対 open 条件での実験結果を図 4.7 に，話者 open 条件での実験結果を図 4.8 に示す．基準距離によって相関にわずかな差があるものの，どちらの条件においても，K の増加に応じて相関が単調に上がり続けるという傾向は変わらないことが確認できた．またベースラインとの比較についても両条件で同様のことが言えて，提案する距離予測手法の精度が，話者対 open 条件では oracle phoneme recognizer を上回り，話者 open 条件では oracle phoneme recognizer より低く real phoneme recognizer より高いという結果となった．

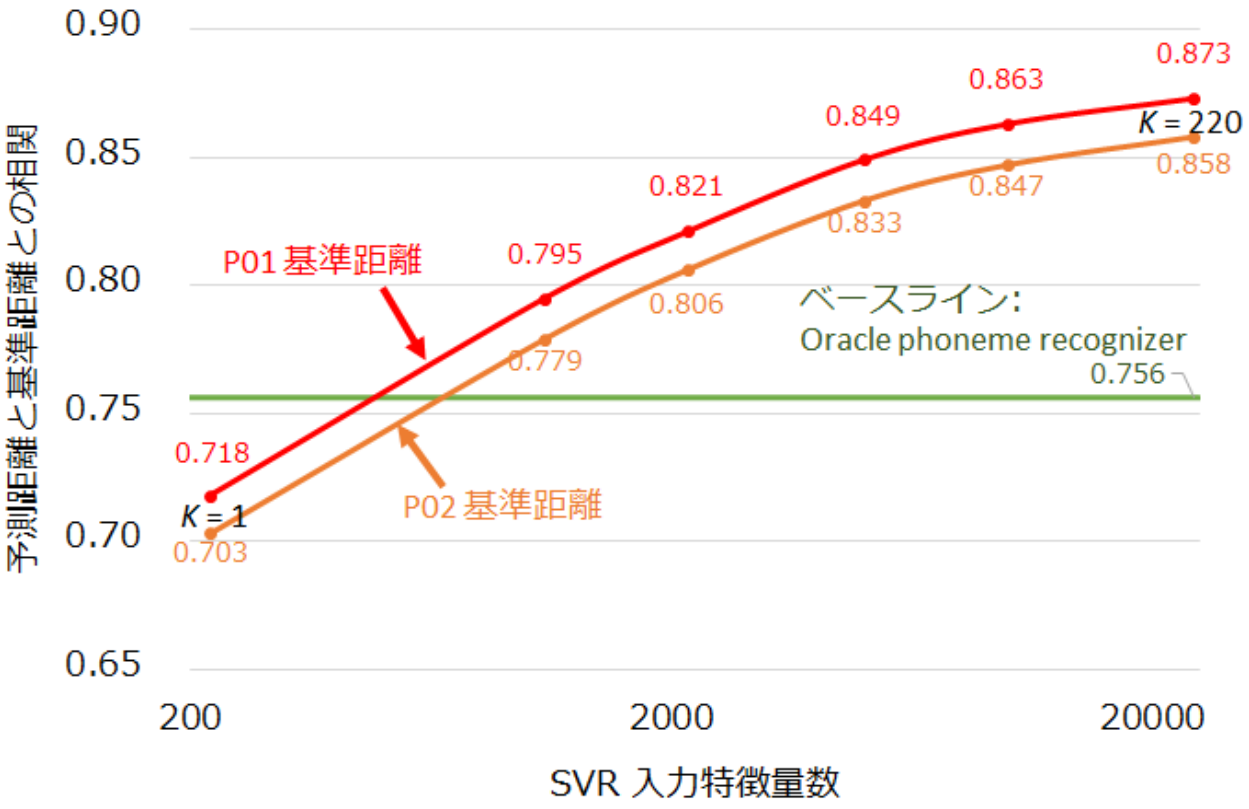

図 4.7: 話者対 open 条件における P01, P02 二通りの基準距離を対象とした予測での相関の比較

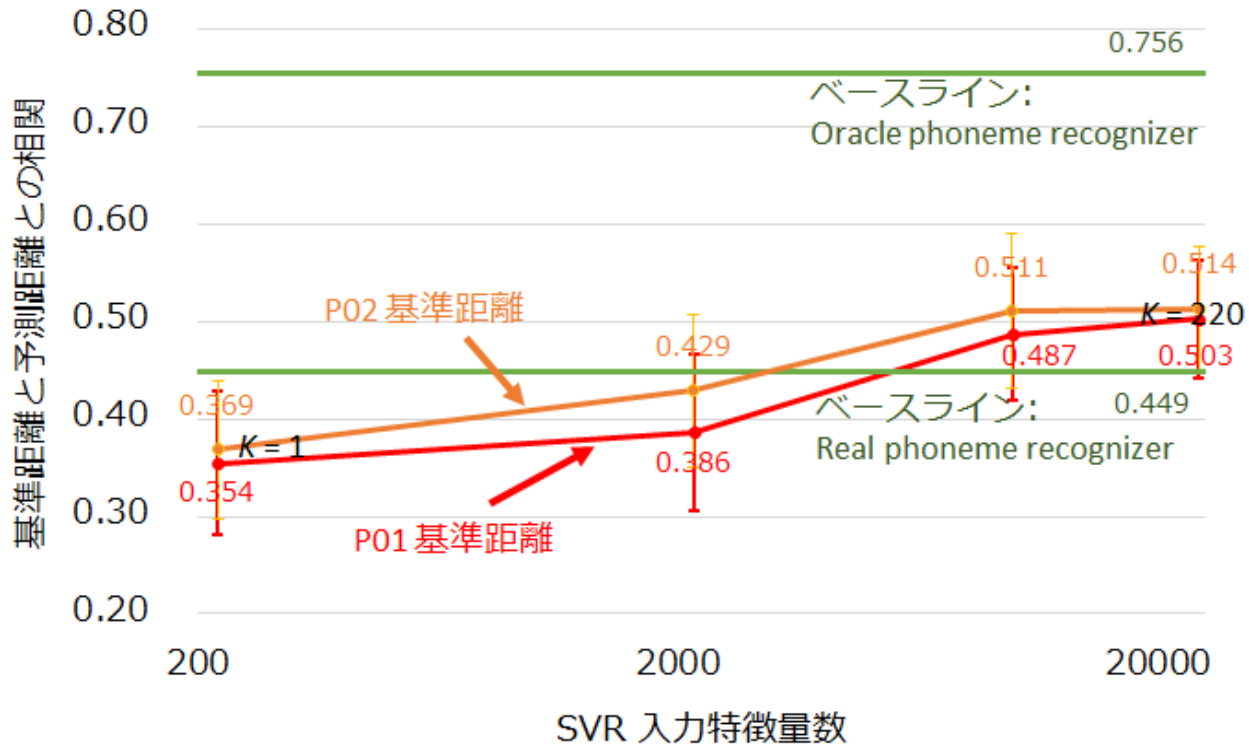

図 4.8: 話者 open 条件における P01, P02 二通りの基準距離を対象とした予測での相関の比較

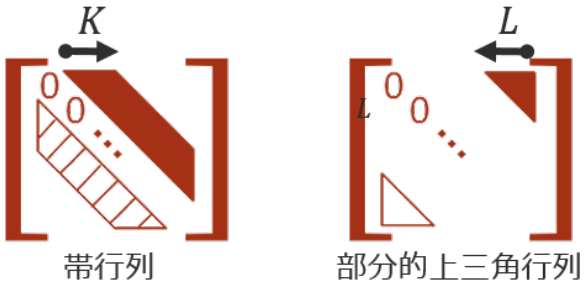

図 4.9: 局所性に着眼した特徴選択

4.6　発音構造の特徴選択による分析

4.5.3 で，SAA から算出できる発音構造特徴を全て入力することである程度高い精度で距離予測が行えることが確認された．本節では，入力する特徴量をある知識に基づいて選択することで，24,310 次元ある発音構造のうちどの部分の特徴がより発音距離予測に有効であるかを調査する．

4.6.1　音声セグメント間の時間的な遠近に着目した構造特徴の選択

4.5.3 では二話者の発音構造の差行列 D_{ij} の全要素を利用しているが，この中には時間間隔が狭い二音間の相対特性と，200 以上の音素分の発声を挟んだ二音間の相対特性が混在している．そこで，相対特性の局所性に着目した特徴選択を行い，時間間隔が狭い二音と広い二音の相対特性のうちどちらがより距離予測により有効であるかを調査した．

時間間隔の狭い二音間の相対特性は，構造距離行列のうち対角に近い帯行列部分に相当する．逆に，時間間隔の広い二音間の相対特性は，対角から遠い部分的上三角部分に相当する．ここでは，図 4.9 に示すように，帯行列部分で K を 220 まで増やしていった場合と，部分的上三角行列で L を 220 まで増やしていった場合とで，それぞれ距離予測精度がどう変化するか調べる．

4.6.2　音声学的な知識に基づく構造特徴の選択

構造特徴の要素である音声セグメント間の f-divergence の変換不変性は，どちらの音声に対してもかけられるアフィン変換が同一であるという仮定の元で成立する，しかし MLLR などの話者適応手法では，音声を類似度で分類し組毎に変換を別々にする場合がある．そこで本小節では，音声学的な分類によって話者性に基づく変換が異なるという予想のもと，知識に基づいて SAA パラグラフ中の 221 米語音素を分類し，各組に属する音素間のみで構造算出を行なう．

A. 母音
　/aa/, /ae/, /ah/, /ao/, /iy/, /uw/, /eh/, /ih/, /ay/, /ow/, /oy/, /ey/, /er/, /w/
B. 共鳴子音
　/m/, /n/, /ng/, /l/, /r/
C. その他の有声子音
　/b/, /v/, /g/, /d/, /z/, /dh/
D. 無声子音
　/p/, /t/, /k/, /ch/, /f/, /th/, /s/, /sh/, /hh/

ここでは，4 つの組（ A と B と C と D ）で別々に構造算出する場合 (4 groups) と，有声と無声だけを区別し（ A+B+C と D ）構造算出する場合 (2 groups) で実験を行う．この時特徴量数は前者の場合 6,593 ，後者の場合 14,752 となる．

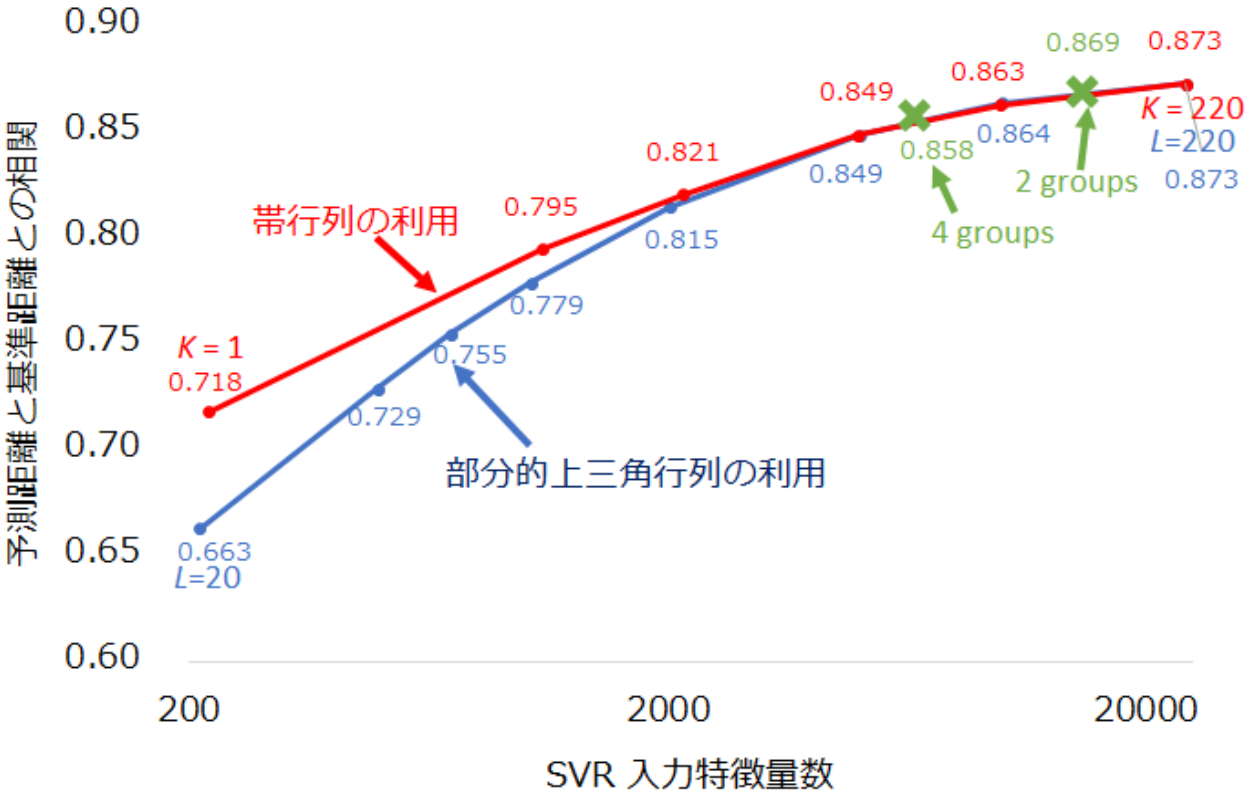

図 4.10: 話者対 open 条件における P01 距離予測での特徴選択による相関の比較

4.6.3　結果

P01 基準距離で，話者対 open 条件での 4.6.1, 4.6.2 の結果を図 4.10 に示す．

帯行列部分と部分的上三角部分の特徴を比べると，特徴量数の少ない所では帯行列部分の方が相関が少し高いものの，K, L を大きくするとすぐにその差が見られなくなる．また，部分的上三角行列部分の利用でも $L = 20$ の時相関は 0.663 で，既にベースラインの real phoneme recognizer を超える結果となっている． 200 程度の米語音素発音分の時間であれば，離れた音声セグメント間の相対特徴も十分距離予測に有効であると言える．

音声学的な知識に基づいた特徴選択では， 4 groups, 2groups で共に，特徴量数が同じとなる帯行列部分利用の場合と相関に差が見られなかった．音声学的な知識により変換が異なるという仮定の利用は有効ではなかった．

4.7　おわりに

本節では構造特徴を用いた発音距離予測実験において，予測精度の向上のための手法の改善と種々の特徴選択による分析的検討を行なった．SAA パラグラフから作成した発音モデルからとり得る構造特徴が全て発音距離予測に有効であることを示した．また提案する手法が，複数の基準距離に対してもほぼ同等の精度で予測ができることを示した．全構造特徴のうちどの要素がより予測性能に働いているかを解明するためには，さらなる検討を要する．

予測距離と基準距離との相関は，話者対 open 条件においては最大で 0.873 と十分な高いものとなったが，話者 open 条件では最大で 0.503 となり，実在のベースラインシステムを越えるものの不十分な結果となった．各 open 条件で想定される距離予測システムの実用の面で考えると，発音書き起しのある英語話者群で Web アーカイブを作成した場合に，書き起しのない利用者 1 名対各アーカイブ話者の発音距離は十分な精度で予測できるので，訛りの Web ブラウジングシステムといった応用については現在の技術で実現可能であると思われる．しかし，未知話者間の距離予測の精度が不十

分であるため，世界中の全英語話者を対象とした発音地図を作成し世界諸英語研究に貢献することはまだ叶わない．

第5章

英語発音距離予測に用いる新たな音響特徴量の検討

5.1　はじめに

前節では SVR の入力として，[3] と同様に構造特徴を利用することを基本にして，特徴の算出手段を変えることにより発音距離予測の精度を向上させることを検討していた．本節では，構造特徴とは別の特徴を導入することで精度の改善を試みている．発音構造は無声音など一部の音声セグメントに対し最適な記述となっていない可能性があるが，5.2 ではそれらの音声に特化して有効な特徴とするために，話者間の発音モデルの直接比較により話者間差異特徴を取得し予測に用いる．5.3 では全く新しい試みとして，複数の学習話者からの距離で未知話者を特徴付け，これを回帰の入力として発音距離を算出することを実験的に検討する．

5.2　音声の絶対的特徴を用いた距離予測

5.2.1　音声の絶対的特徴の導入

構造特徴は，発声中の各音声セグメントを相対的な配置のみで捉えることにより，話者の違いによる音響変動から発音の違いを切り離すものである．しかし音声には，摩擦音や破裂音といった無声のものもあり，これらは話者による影響が比較的少ない．これらの音声は，空間での位置そのものに話者の発音の特徴が表れていて，相対関係で記述してしまうことが最適とならない可能性が考えられる [45] ．そこで，二話者間で対応する音声セグメントモデルの差異を直接とり，これを話者間の発音の差異を表す絶対的特徴として導入し，構造特徴と組み合わせることを考える．絶対的特徴は．声色の違いが発音の違いと分けられず残るものの，話者の差異が表れにくい無声音などについては有効な特徴となり得る．

5.2.2　三段階の比較粒度での絶対的特徴

絶対的特徴は， 図 5.1 のように，二話者間で対応するセグメント同士のバタチャリヤ距離 (BD) をとり算出する．本研究では，同じ音素 HMM に属する 3 状態で BD の平均をとる音素を単位とした 221 次元の特徴 (phoneme-based direct comparison) と，状態毎に別の特徴としてそのまま扱う状態を単位とした (221×3=) 663 次元の特徴 (state-based direct comparison) を使用する．

また絶対的特徴のさらなる拡張として，各状態が持つ音響特徴量の各次元を単位として比較し特徴として導入することも考える (dimension-based direct comparison) ．HMM の各状態の音響特徴平均ベクトルと分散から次元毎に個別に BD を求める．音響分析条件が表 4.1 である時音響特徴量次元数は 24 であるので，dimension-based direct comparison の次元数は $663 \times 24 = 15,912$ となる．

このように特徴量粒度を上げることは，構造特徴でも可能である（図 5.2 ）が，本稿では絶対的特徴においてより粒度の高い特徴量を使用することを検討する．

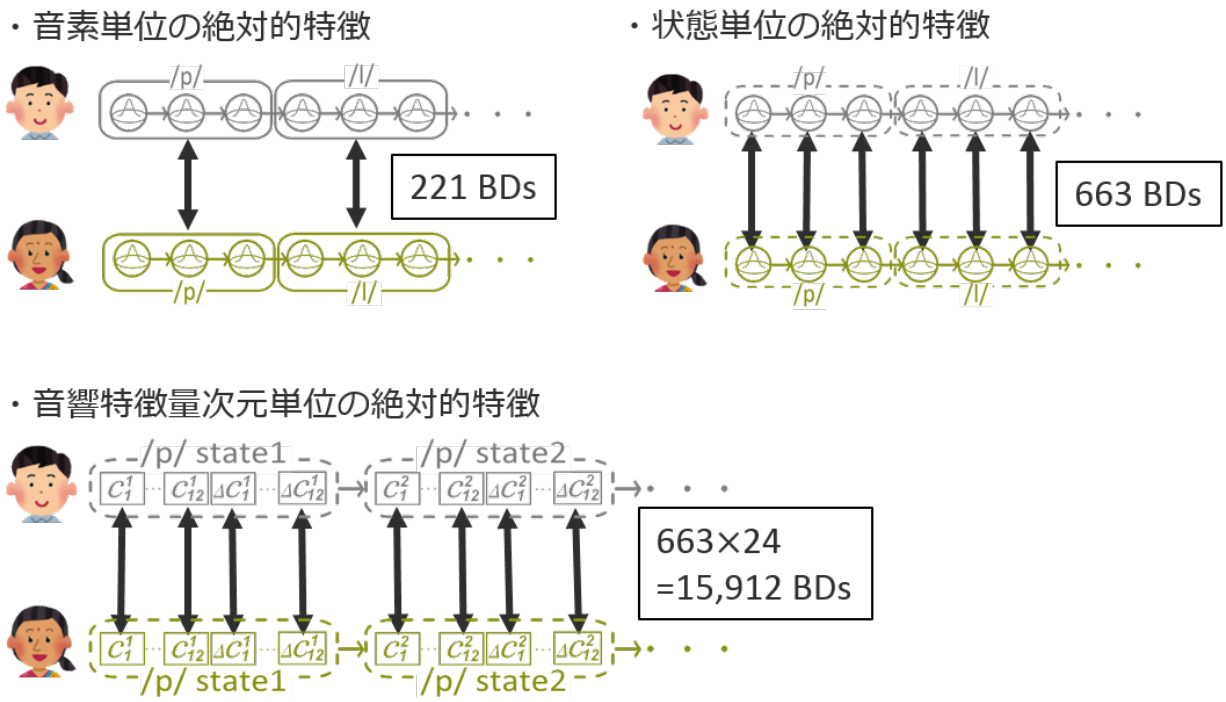

図 5.1: 絶対的特徴

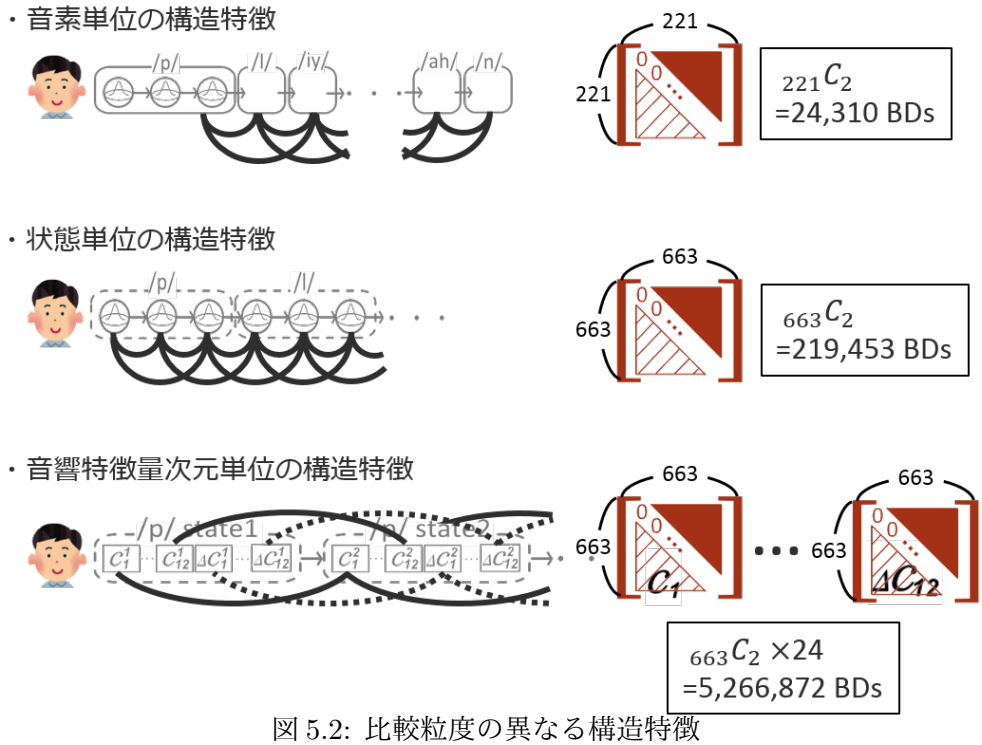

図 5.2: 比較粒度の異なる構造特徴

表 5.1: 話者対 open 条件における P01 距離予測での絶対的特徴のみを用いた時の相関

	phoneme-based direct comparison	state-based direct comparison	dimension-based direct comparison
P01 基準距離	0.771	0.830	0.867

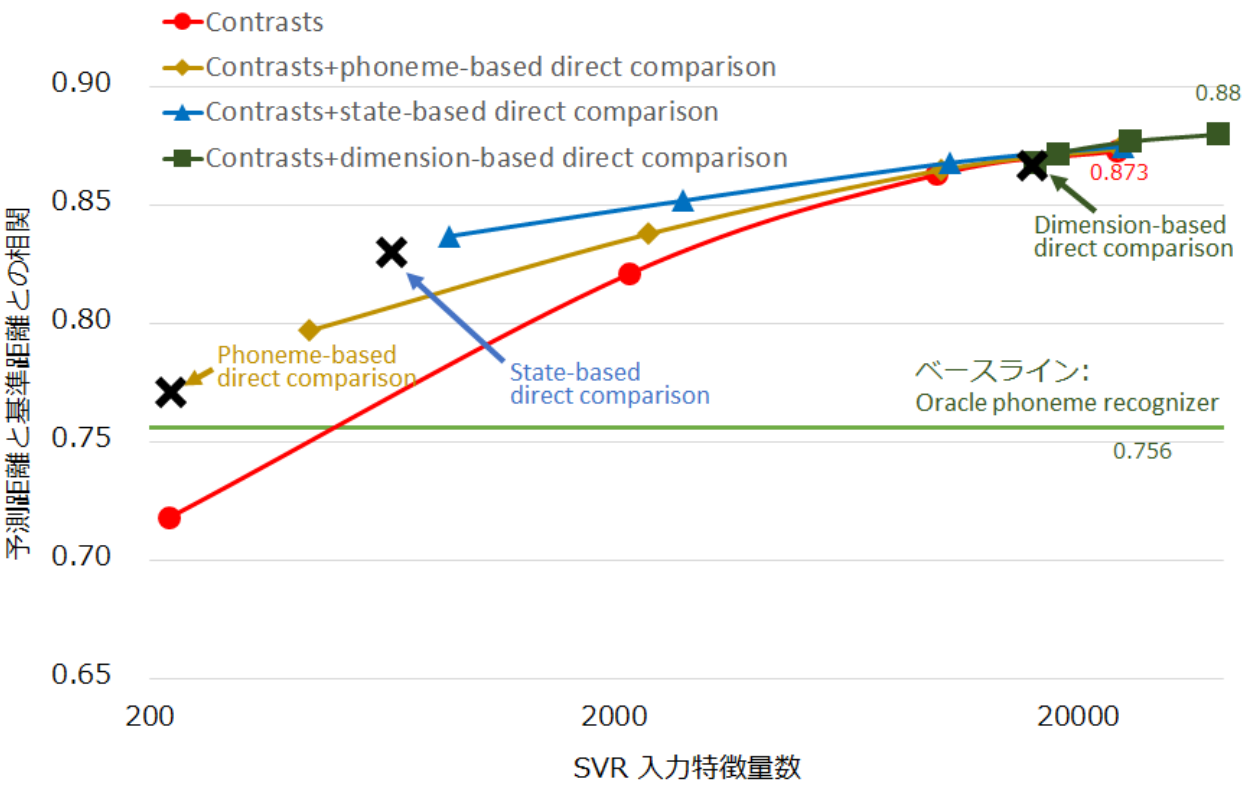

図 5.3: 話者対 open 条件における P01 距離予測での構造特徴・絶対的特徴を用いた時の相関

5.2.3　絶対的特徴を用いた距離予測の結果

話者対 open 条件で，各絶対的特徴のみを用いた場合の結果を，表 5.1 に示す．比較の粒度が高いほどよい良い性能で予測ができている．

話者対 open 条件で，絶対的特徴と構造特徴を比較した結果を図 5.3 に示す．構造特徴のみと絶対的特徴のみを用いた場合の結果を比べると，phoneme-based direct comparison 及び state-based direct comparison については，同じ特徴量数の構造特徴よりも高い相関を示した．絶対的特徴を構造特徴に追加する形で用いた場合，特徴量数の合計が少ないところでは相関の向上がはっきりと見られたが，特徴量数が十分に多くなると相関の差は僅かなものとなった．

次に，話者 open 実験での，絶対的特徴のみを用いた場合の結果を表 5.2 に示す．また，話者 open 条件で，絶対的特徴と構造特徴を比較した結果を図 5.4 に示す．

State-based direct comparison と dimension-based direct comparison に着目すると，これらの特徴単体で入力とした場合，同じ特徴量数の構造特徴を用いる場合と比べ相関が低くなっている．4.3 で考察したように，話者対 open 実験が学習データの中から類似した話者を探すような設定であるのに対し，話者 open 実験では類似した話者対を探すような設定となっている．絶対的特徴は訛り以外の声色による変動に影響を受けるものである．話者対 open 条件では声色の影響は無視できる程度だったのに対し，話者 open 条件では影響が二話者分の組み合わせで特徴量に働くこととなり，初めて結果を悪くさせることになったのだと思われる．

表 5.2: 話者 open 条件における P01 距離予測での絶対的特徴のみを用いた時の相関の比較

	phoneme-based direct comparison	state-based direct comparison	dimension-based direct comparison
P01 基準距離	0.390	0.346	0.463

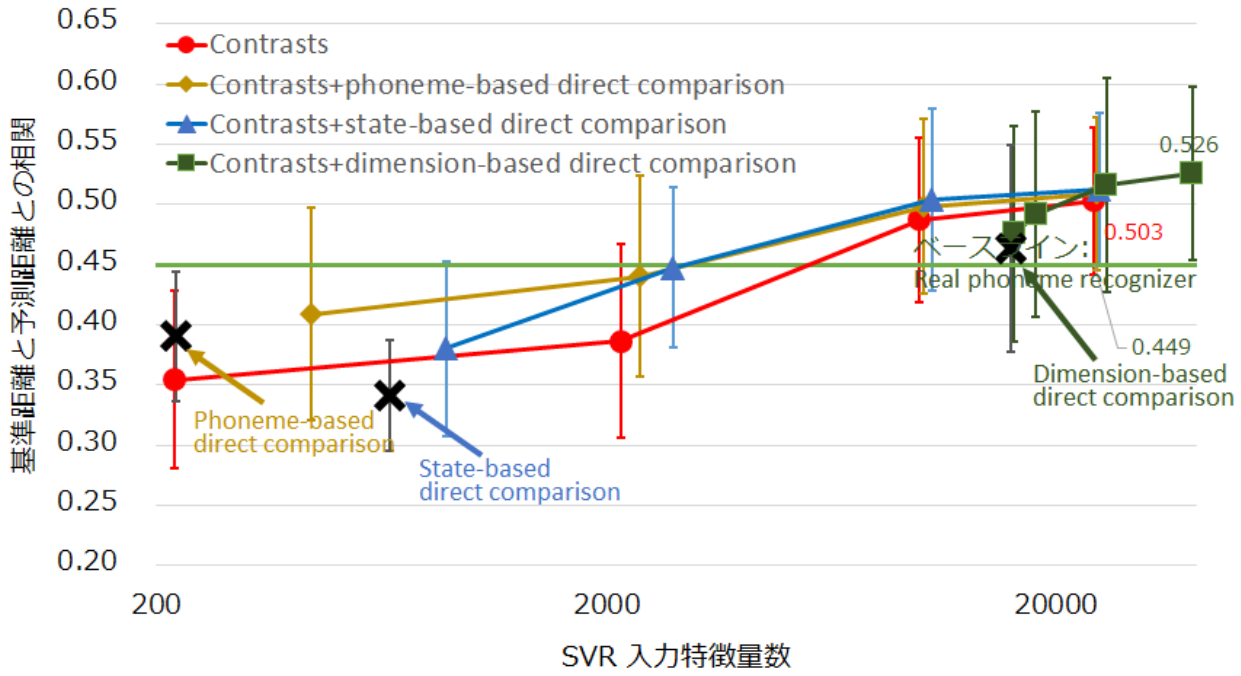

図 5.4: 話者 open 条件における P01 距離予測での構造特徴・絶対的特徴を用いた時の相関の比較

これらの特徴を構造特徴に追加する形で用いた場合は，構造の全特徴利用時の相関をわずかながらさらに上げることが可能となっている．絶対的特徴のみでは予測精度が低かったが，構造特徴と組み合わせた時には絶対的特徴の有効な部分のみが SVR によりうまく抽出されたのだと思われる．絶対的特徴は一部の音については相対特徴よりも良い記述となり得るという予想に沿う結果が見られた．

5.3　話者を頂点とした多角形の歪みを利用した発音距離回帰

ここでは，話者間がなす相対特性（発音距離）で各話者を特徴付けることにより，各人の発音を直接比較することなく話者間距離を予測することを考える．

3.4.4 では，各音素を他の音素からの距離という情報で特徴付けることで，対応する二音が発音構造に与える歪みを定量化できることを説明した．発音構造のノードは各母音音素に対応していたが，今この音素を世界の英語話者（発音）で置き換えることを考えると，ノード間の距離は発音距離によって置き換えられる．特に発音構造に歪みを与える二音を話者 open 実験での評価用話者（未知話者），その他の参照用の共通する音素を学習用話者（参照用話者）に置き換えた時に，構造歪みの考え方を拡張すれば，未知話者二名が話者群からなる形状に与える歪みは，それぞれの未知話者を全参照用話者からの距離で特徴付けることで表現できる，ということになる（図 5.5）．すなわち，未知話者間の差異（発音距離）はそれぞれの参照用話者からの距離と関連性が高いと考えれる．そこで，音響特徴量ではなく各参照用話者からの距離を特徴量として各話者を表現し，SVR の入力とすることを検討する．

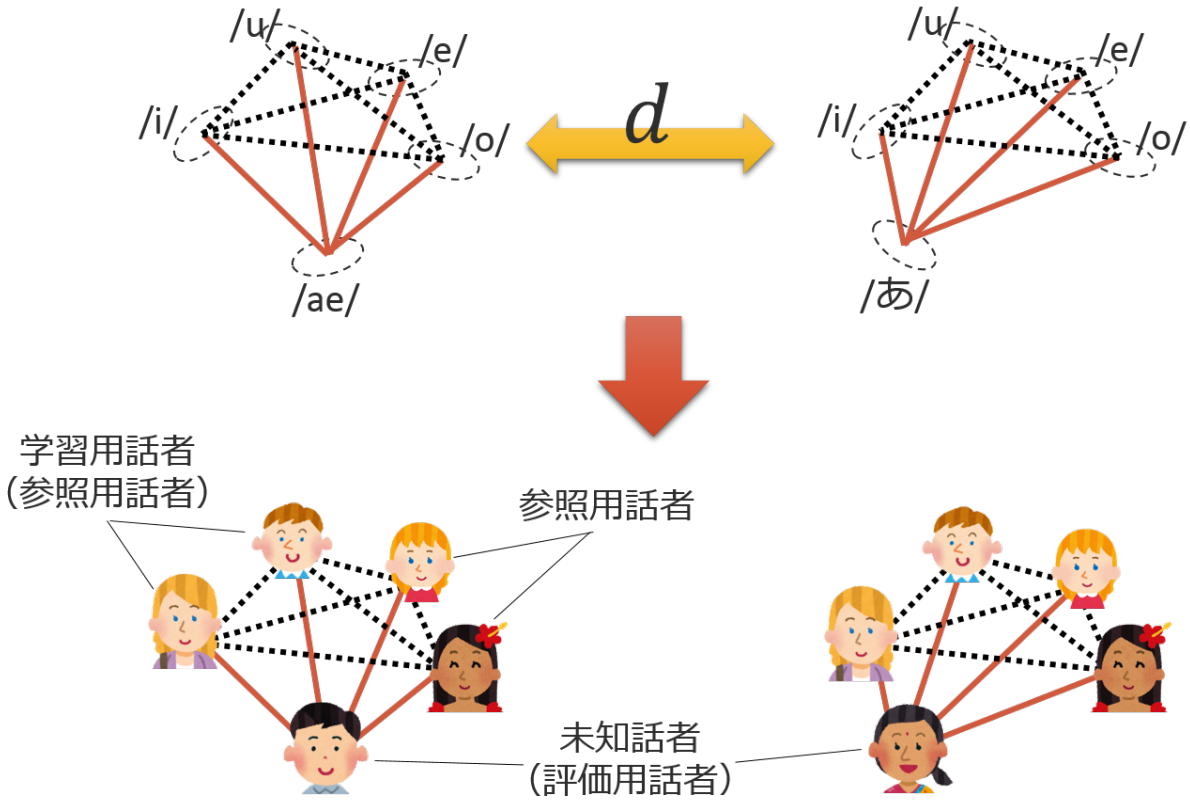

図 5.5: 構造歪みの発想を利用した未知話者間距離予測

5.3.1　話者群多角形の歪みを利用することを目的とした回帰の設計

本節では話者 open 実験において，各参照用話者から距離予測対象の二名の未知話者への距離を SVR の入力とし，未知話者間の発音距離を予測することを考える．この時，未知話者には発音書き起しの情報が与えられていないので，別に作成した SVR による距離予測器が出力する参照用-未知話者間の予測距離を入力特徴としなくてはいけない．ここでは，他の実験で学習用に用いていた話者を二分し，片方の 148 名を参照用話者とし，まず参照用話者のみを用いて距離予測器を作成する．音響分析条件，及び各話者の発音距離の算出方法は 4.5.3 と同様で，入力を二話者の発音構造差行列の全要素（ 24,310 次元）とする．次に，使用していないもう片方の学習用話者と未知話者それぞれについて，距離予測器により 148 名の参照用話者からの予測距離を求め，各話者の特徴とする（図 5.6）．最後に，参照用話者からの予測距離を入力とする SVR を，参照用話者以外の学習用話者からなる話者対で学習し，未知話者からなる話者対で評価する．二話者 A, B の特徴ベクトル（図 5.6）を $\{A_i\}, \{B_i\}$ とすると，SVR の入力には式 (5.1) で求まる差ベクトル $\{d_i\}_{i=1}^{148}$ を用いる．

$$d_i = |A_i - B_i| \tag{5.1}$$

5.3.2　話者群多角形の歪みを利用した回帰による未知話者間距離予測の結果

5.3.1 で述べた手順により，話者 open 条件の 5-fold 交差検定で未知話者間距離予測を行なう．基準距離として P01 の距離を用いる．

距離学習用話者 148 人からの予測距離で各話者の特徴量ベクトルを求める．この時の発音距離の予測距離と基準距離の相関は， 0.725 となった．学習データの数が揃っていないため厳密な比較ではないが，書き起しが与えられた話者対未知話者の距離予測が話者 open 条件と比べ高い精度で行なえることが実際に示されている．ここでの予測精度が最終的な未知話者間距離予測にどれだけ影響する

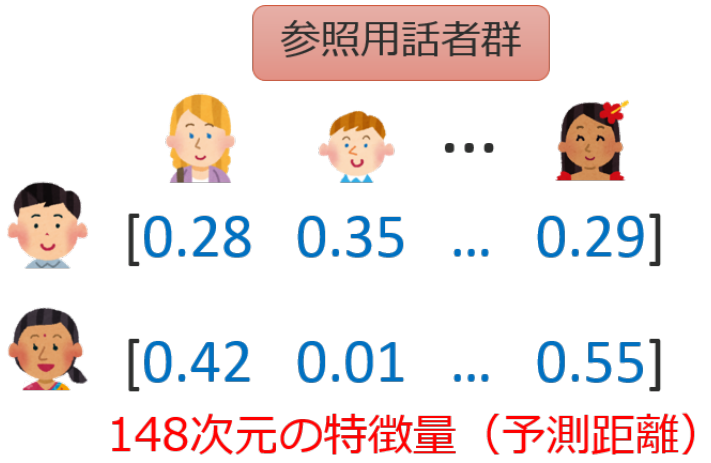

図 5.6: 距離学習用話者からの予測距離を特徴量として利用

表 5.3: 話者 open 条件における P01 距離予測での話者群多角形形状特徴を利用した場合の相関

P01 基準距離	structure	geometric (real)	geometric (oracle)	structure+geometric (real)
Set1	0.465	0.289	0.720	0.406
Set2	0.520	0.380	0.734	0.461
Set3	0.417	0.186	0.769	0.408
Set4	0.546	0.395	0.748	0.588
Set5	0.503	0.403	0.759	0.533
平均	0.503	0.330	0.746	0.479

かを検証するため，学習用話者及び未知話者の特徴として，実際の予測距離を用いた場合 (geometric (real)) ，及び全て P01 基準距離を用いた場合 (geometric (oracle)) の二通りで実験を行い比較する．

結果を表 5.3 に示す．比較として，SVR の入力を発音構造差行列の全特徴 24,310 次元とし，学習用話者を全て発音距離予測モデルの学習に用いて，未知話者間距離を直接予測した場合の結果 (structure, 図 4.6 に示す結果と同様) も示している．

geometric (real) と geometric (oracle) を比較すると，geometric (real) は相関が大幅に落ちている．話者の特徴である予測距離は基準距離と 0.725 の相関を持っていたが，予測精度が下がることが結果に大きく影響することが示された．geometric (real) はまたデータセットに対するばらつきが大きく，Set1 及び Set3 で特に相関が低くなっている．今回の実験では参照用話者 148 名がランダムに選択されているが，それが一部のセットで不適切な選択となったために，各話者の特徴が不十分なものとなり結果が悪くなったと考えられる．参照用話者としてどのような基準で選択を行なうことで距離予測精度を向上させることができるかについて，更なる検討が必要である．

相関の平均で比較すると，geometric (real) は structure より低く，geometric (oracle) は structure より高いという結果となっている．予測距離の精度をさらに上げることができれば，構造特徴を利用して直接予測するよりも高い精度となる可能性があると考えられる．

最後に，本節の特徴を構造特徴に追加する形で SVR に入力して実験を行なった（表 5.3 の structure+geometric (real) ）．この時の入力特徴量数は $(24,310 + 148 =)24,458$ である．相関の平均は structure から低くなっている．この結果からも，現状の距離予測精度では形状に基づく特徴を利用するには不十分であることが伺える．

5.4　おわりに

本節では SVR に入力する特徴として，絶対的特徴と，話者群からなる多角形の形状特徴を導入した．絶対的特徴を構造特徴に追加する形で用いることで，話者 open 条件でわずかながら精度が改善することができた．また，絶対的特徴は特徴算出時の比較粒度により性能が異なることが分かった．今後は構造特徴についても比較粒度を変えることを検討する必要がある．形状特徴については，oracle システムにおいて直接的な比較の回帰より高い精度が出たため今後有効な特徴となり得るが，直接的な発音距離予測の精度が低い現状では利用価値が低い．

第6章

結論

6.1　まとめ

英語は国際共通語として世界中に広まる中で多様化し，この多様化は今後も進んでいくと思われる．本研究では多様化した英語発音に基づいて英語話者をクラスタリングすることを目標に，そのために必要な技術として話者間の英語発音距離を音声情報から予測することを検討している．

本稿では先行研究と同じ実験条件において，先行研究で考慮されていなかった時間的に離れた音声セグメント間から得られる特徴も予測に使用することで，発音距離の予測精度をベースラインを超えるまで大幅に向上させられることを示した．しかし本稿で新たに設定した，さらに実用に即した実験設定においては，本稿の手法でも十分な予測精度は得られなかった．

次に本稿では回帰予測の入力として新たに二つの特徴量を提案し，実験を行なった．話者の違いが表れにくい音に対して有効と考えられる絶対的特徴の使用では，構造特徴に追加することで，実際に話者 open 実験でわずかながらさらに予測精度を向上させることができた．また話者群から成る多角形の形状特徴を利用し未知話者間の発音距離予測を行なうことを検討したが，結局この形状特徴を算出するために発音距離予測が高精度に行なえる必要であることが分かり，現状での利用価値は低いと思われる．

本稿の手法により，話者対 open 条件では十分と思われる結果が得られ，話者 open 条件では十分とならなかった．発音距離予測技術の実用を考えると現状では，発音に着眼した Web ブラウジングシステムには応用できる可能性が高いが，発音地図の作成に向けてはさらなる検討を要する．

6.2　残された課題

今回，構造特徴の性質を分析するために，音声学の知識に基づいて事前に特徴選択をした上で実験を行なったが，有意な結果は得られなかった．特徴量の有効な部分を抽出するための手法としては他に，自動学習により特徴選択を行なう手法や，特徴量ベクトルを低次元空間に写像し次元圧縮を行なう手法が数多く提案されている．本研究においてもさらなる精度改善のため，今後これらの手法を導入し，これまで使用された特徴量をより有効な形で表現することを検討すべきである．

今回の使用話者データの選択においては話者の母語や居住地といった情報は一切考慮されていないため，データの偏りが結果に悪影響を及ぼしていることは否定できない．また今回実験に使用した Speech Accent Archive は，世界諸英語の分析と異なる目的で作成されたコーパスであり，採用されている読み上げ文が英語の多様性を十分に網羅する設計になっていない．英語発音地図を作成するために，必要なデータ量の検討と，あらゆる英語発音の違いが見られるような読み上げ文の設計を行なわなくてはならない．

発音距離を実際に可視化して発音地図とする際に，どのような形式での提示が英語学習者にとって望ましいものとなるか，利用者のフィードバックを得ながら検討を進めていくべきである．また，本稿において，提案する発音距離が母語話者らしさに対し妥当性を持つことは検討しているものの，世界諸英語の発音地図における距離に対する妥当性の検証は行なわれていない．実際に作成された発音地図がどれだけ妥当性を持つのか，世界諸英語研究に対しどのような知見を提供することができるのか，言語学や分類学の知識と擦り合わせて探っていく必要がある．

謝辞

まず本研究を進めるにあたり，二年間指導教員として多大なるご指導をして頂いた峯松信明教授と広瀬啓吉教授に深く感謝致します．特に峯松信明教授には，ご多忙であるにも関わらず非常に多くの時間を割いて頂き，実験，執筆，発表などあらゆる面で数え切れないほどのご助言を賜りました．重ねて，感謝を申し上げます．また，日頃の研究活動を支えて下さいました高橋登技術専門員，池上恵事務補佐員にも感謝致します．

また，常日頃より親身になって熱心にご指導して頂いた齋藤大輔助教にも，深く感謝致します．幅広い知識と教養を有し，研究に関して的確なアドバイスをして下さり，また研究以外の様々なことについても相談に乗って下さいました．

元客員研究員の沈涵平氏は，私に本研究を引き継がれる際に研究を円滑に開始できるように尽力して下さいました．留学生の史天澤氏は，その高い技術力をもって本研究に積極的に取り組んで下さり，多くの知見を残して下さいました．研究室の OB である鈴木雅之氏は，本研究の実験条件に対する非常に鋭いご指摘をして下さり，そのことは本研究にとって重大な変換点となりました．彼らのご助力がなければ，本研究は今と全く違う所で停滞していたことと思います．深く感謝致します．

研究について議論し，多くの時間を過ごした研究室の方々に感謝致します．特に柏木陽佑氏には，日夜問わず様々な相談に乗って頂き，二年間で多くのことを学ばせて頂きました．また尾崎洋輔氏，内田秀継氏も，様々なことで面倒を見て下さいました．同期の橋本哲弥氏には，学部生の時から縁があり，多くのことを助けてもらいました．また藤垣健太郎氏，水上智之氏には，新しく研究室に入った私に多くのことを教えて下さいました．

NTT メディアインテリジェンス研究所の皆様に感謝致します．夏期学外実習を通して学ぶことのできた技術が，その後の研究活動で大いに役立ちました．

卒論研究でご指導して頂いた近山隆元教授，鶴岡慶雅准教授，近山・鶴岡研究室の皆様，そしてこれまで私を支えて下さった家族，友人，若菜根津一丁目店のスタッフの皆様，その他関わることのできた全ての方々に感謝致します．本当にありがとうございました．

2015 年 2 月 5 日
笠原 駿

参考文献

[1] B. Kachru, Y. Kachru and C. L. Neison, *The handbook of World Englishes*, Wiley-Blackwell, 2009.

[2] J. Jenkins, *World Englishes: a resource book for students*, Routledge, 2009.

[3] H. -P. Shen, N. Minematsu, T. Makino, S. H. Weinberger, T. Pongkittiphan and C. -H. Wu, "Automatic pronunciation clustering using a world English arhcive and pronunciation structure analysis," *ASRU*, 222–227, 2013.

[4] B. Kachru, "Standards, codification and sociolinguistic realism: The English language in the Outer Circle," *English in the World: Teaching and Learning the Language and Literatures*, R. Quirk, H. G. Widdowson and Y. Cantu (eds.), Cambridge University Press, 11–30, 1985.

[5] 本名信行, 世界の英語を歩く, 集英社新書, 2003.

[6] S. Weinberger, *Speech Accent Archive*, http://accent.gmu.edu, 2014.

[7] P. Meier, *International Dialects of English Archive*, http://www.dialectsarchive.com.

[8] B. Seidlhofer, "10.RESEARCH PERSPECTIVES ON TEACHING AS A LINGUA FRANCA," *Annual Review of Applied Linguistics*, 24, 209–239, 2004.

[9] A. Mauranen, "A rich domain of ELF-the ELFA corpus of academic discourse," *Nordic Journal of English Studies*, 5, 2, 145–159, 2006.

[10] A. Mauranen, E. Ranta, "English as an academic lingua franca-the ELFA project," *Nordic Journal of English Studies*, 7, 3, 199–202, 2008.

[11] *The CMU pronouncing dictionary*, http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

[12] G. Fairbanks, *Voice and Articulation Drillbook*, 124–139, Harper & Row, 1969.

[13] J. C. Wells, *Accents of English*, Cambridge University Press, 1982.

[14] 鳥飼玖美子, 国際共通語としての英語, 講談社現代新書, 2011.

[15] 田中春美, 田中幸子（編）, World Englishes-世界の英語への招待, 昭和堂, 2012

[16] 鶴田知佳子, 柴田真一, ダボス会議で聞く世界の英語, コスモピア, 2008.

[17] L. M. Arslan and J. H. L. Hansen, "A study of temporal features and frequency characteristics in American English foreign accent," *Acoustical Society of America*, 102, 28–40, 1997.

[18] A. Ljolje and F. Fallside. “Recognition of isolated prosodic patterns using hidden Markov models,” *Computer Speech & Language*, 2, 1, 27-33 1987.

[19] L. W. Kat and P. Fung, “Fast accent identification and accented speech recognition,” *ASSP*, 1, 221–224, 1999.

[20] S. Deshpande, S. Chikkerur and V. Govindaraju, “Accent classification in speech,” *AIAT*, 139–143, 2005.

[21] S. Zhang and Y. Qin, “Semi-supervised accent detection and modeling,” *ICASSP*, 7175–7179, 2013.

[22] M. H. Bahari, R. Saeidi, H. V. Hamme and D. V. Leeuwen. “Accent recognition using i-vector, gaussian mean supervector and gaussian posterior probability supervector for spontaneous telephone speech,” *ICASSP*, 7344–7348, 2013.

[23] 粕谷英樹, 鈴木久喜, 城戸健一, “年齢，性別による日本語 5 母音のピッチ周波数とホルマント周波数の変化,” 日本音響学会誌， 24, 6, 355–364, 1968.

[24] D. Stantic and J. Jo, “Accent identification by clustering and scoring formants,” *CCE*, 6, 232–237, 2012.

[25] M. Wieling, J. Bloem, K. Mignella, M. Timmermeister and J. Nerbonne, “Measuring foreign accent strength in English. Validating Levenshtein Distance as a measure,” *The Mind Research Repository*, 2014.

[26] M. Wieling, J. Nerbonne, J. Bloem, C. Gooskens, W. Heeringa and R. H. Baayen, “A cognitively grounded measure of pronunciation distance,” *PLoS one*: e75734, 2014.

[27] G. Kondrak, “A new algorithm for the alignment of phonetic sequences,” 1st North American chapter of the Association for Computational Linguistics conference. Association for Computational Linguistics, 2000.

[28] P. Ladefoged, *A Course in Phonetics*, Harcourt Brace Jovanovich, 1995.

[29] B. Kessler, “Computational dialectology in Irish Gaelic,” *the European ACL*, 60–67, 1995.

[30] K. W. Church and P. Hanks, “Word association norms, mutual information, and lexicography,” *Computational Linguistics*, 16, 1, 22–29, 1990.

[31] M. Wieling, J. Prokic and J. Nerbonne, “Evaluating the pairwise string alignment of pronunciation,” EACL 2009 workshop on language technology and resources for cultural heritage, social sciences, humanities, and education. Association for Computational Linguistics, 2009.

[32] R. H. Baayen, P. Milin, D. F. Durdevic, P. Hendrix and M. Marelli, “An amorphous model for morphological processing in visual comprehension based on naive discriminative learning,” *Psychological Review*, 118, 438–482, 2011.

[33] D. Danks, “Equilibria of the Rescorla-Wagner model,” *Mathematical Psychology*, 47, 109–121, 2003.

[34] 迫江博昭, 千葉成美, “動的計画法を利用した音声の時間正規化に基づく連続単語認識,” 日本音響学会誌, 27, 9, 483–490, 1971.

[35] N. Minematsu, S. Asakawa, M. Suzuki and Y. Qiao, “Speech structure and its application to robust speech processing,” *New Generation Computing*, 28, 3, 299–319, 2010.

[36] Y. Qiao and N. Minematsu, “A study on invariance of f-divergence and its application to speech recognition,” *Signal Processing*, 58, 7, 3884–3890, 2010.

[37] 鈴木雅之, 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評価,” 情報処理学会論文誌, 52, 5, 1899-1909, 2011.

[38] 峯松信明, 鎌田圭, 朝川智, 鈴木雅之, 牧野武彦, 西村多寿子, 広瀬啓吉, “音声の構造的表象に基づく学習者分類の検証と発音矯正度推定の高精度化,” 情報処理学会論文誌, 52, 12, 3671–3681, 2011.

[39] M. Pitz and N. Hermann, “Vocal tract normalization equals linear transformation in cepstral space,” *Speech and Audio Processing*, 13, 5, 930–944, 2005.

[40] W. Labov, S. Ash and C. Boberg, *The Atlas of North American English*, Mouton and Gruyter, 2005.

[41] 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者発音の音響的分析,” 電子情報通信学会論文誌, 90, 1249–1262, 2007.

[42] X. Ma, N. Minematsu, Y. Qiao, K. Hirose, A. Nemoto and F. Shi, “Dialect-based speaker classification of Chinese using structural representation of pronunciation,” *SPECOM*, 2009.

[43] X. Ma, A. Nemoto, N. Minematsu, Y. Qiao and K. Hirose, “Structural analysis of dialects, sub-dialects and sub-sub- dialects of Chinese,” *INTERSPEECH*, 2009.

[44] *HTK Wall Street Journal Training Recipe*, http://www.keithv.com/software/htk/.

[45] 朝川智, 喬宇, 峯松信明, 広瀬啓吉, “判別分析と構造表象を用いた話者の多様性に超頑健な音声認識, ” 日本音響学会秋季講演論文集, 113–116, 2008.

発表文献

国際会議論文

[1] S. Kasahara, S. Kitahara, N. Minematsu, H. -P. Shen, T. Makino, D. Saito and K. Hirose, "Improved and robust prediction of pronunciation distance for individual-basis clustering of World Englishes pronunciation," *ICASSP*, 3240–3244, 2014.

[2] S. Kasahara, N. Minematsu, H.-P. Shen, D. Saito and K. Hirose, "Structure-based prediction of English pronunciation distances and its analytical investigation," *ICIST*, 331–335, 2014.

[3] N. Minematsu, S. Kasahara, T. Makino, D. Saito and K. Hirose, "Speaker-basis accent clustering using invariant structure analysis and the speech accent archive," *Odyssay*, 158–165, 2014.

[4] T. Shi, S. Kasahara, T. Pongkittiphan, N. Minematsu and Hirose-Minematsu lab., "A measure of phonetic similarity to quantify pronunciation variation by using ASR technology," *ICPhS*, 2015 (to appear).

国内研究会・全国大会

[5] 笠原駿, 峯松信明, 沈涵平, 牧野武彦, 齋藤大輔, 広瀬啓吉, "世界諸英語分類のための構造的表象を用いた発音距離予測," 電子情報通信学会音声研究会資料, SP2013-109, 13–18, 2014.

[6] 笠原駿, 峯松信明, 沈涵平, 牧野武彦, 齋藤大輔, 広瀬啓吉, "世界諸英語分類のための構造的表象を用いた発音距離推定の高精度化," 日本音響学会春季講演論文集, 117–120, 2014.

[7] 笠原駿, 峯松信明, 沈涵平, 牧野武彦, 齋藤大輔, 広瀬啓吉, "世未知話者に対する構造的発音距離推定に関する分析的検討," 日本音響学会春季講演論文集, 121–122, 2014.

[8] 笠原駿, 史天澤, 峯松信明, 齋藤大輔, 広瀬啓吉, "話者間の英語発音距離予測に用いる音響的特徴の検討," 日本音響学会秋季講演論文集, 411–414, 2014.

[9] 史天澤, 笠原駿, 峯松信明, 齋藤大輔, 広瀬啓吉, "世界諸英語発音分類のための話者間参照距離の算出手法に関する検討," 日本音響学会秋季講演論文集, 415–418, 2014.

[10] T. Shi, S. Kasahara, N. Minematsu, D. Saito and K. Hirose, "Experimental investigation of the definition of reference accent distance between speakers toward automatic accent clustering of speakers of world Englishes," 日本音声学会全国大会予稿集, 147–152, 2014.

[11] 笠原駿, 史天澤, 峯松信明, 齋藤大輔, 広瀬啓吉, "発音クラスタリングを目的とした基準発音距離の定義と発音距離予測に用いる音響特徴量の実験的検討," 電子情報通信学会音声研究会資料, 47–52, 2014.

[12] 佐藤惟知, 北原俊, 峯松信明, 笠原駿, 齋藤大輔, 広瀬啓吉, “自己視点からの世界諸英語分類を目的とした発音距離予測とその耐雑音性に関する検討,” 日本音響学会春季講演論文集, 2015 （発表予定）.

学位論文

[13] 笠原駿, “コンピュータ大貧民における事前計算を用いたナッシュ均衡戦略の獲得に向けて”, 東京大学工学部電子工学科卒業論文, 2013.

付録 A

Speech Accent Archive 最頻の IPA 一覧

3.3.4 で抽出した，153 種類の IPA を表 A.1 に示す．これは，SAA 話者 369 人の発音書き起しに出現する IPA の最頻 95% に相当する．

表 A.1: 距離計算に使用された 153 種類の IPA

1	i	21	æː	41	ɐ	61	ɔ̈	81	m	101	s	121	ʒ	141	ɦ
2	ǐ	22	æ̃	42	ɐ̃	62	ɔ̃	82	m̥	102	s̪	122	ç	142	w
3	iː	23	a	43	ɯ	63	ɑ	83	m̝	103	sʲ	123	ɟ	143	ɥ
4	i̥	24	ã	44	ɯ̈	64	ɑː	84	n	104	z	124	j	144	pɸ
5	ï	25	ɨ	45	ɯ̃	65	ɑ̈	85	n̥	105	z̥	125	k	145	tθ
6	ĩ	26	ɨ̥	46	u	66	ɑ̃	86	n̪	106	ɹ	126	kʰ	146	dð
7	y	27	ɨ̃	47	ŭ	67	p	87	n̝	107	ɹ̥	127	k̟	147	ts
8	ɪ̞	28	ʉ	48	uː	68	pʰ	88	ɲ	108	ɹ̞	128	k’	148	dz
9	ɪ	29	ʉ̥	49	ü	69	p̚	89	ŋ	109	r	129	k̟ʰ	149	tɕ
10	ɪː	30	ʉ̃	50	ũ	70	b	90	ɴ	110	ɾ	130	k̚	150	dʑ
11	ɪ̥	31	ɵ	51	ũː	71	b̚	91	t	111	ɾ̝	131	ɡ	151	tʃ
12	ɪ̃	32	ɵ̃	52	ʊ	72	b̥	92	tʰ	112	l	132	ɡ̟	152	dʒ
13	e	33	ə	53	ɤ	73	ɸ	93	t̪	113	l̥	133	ɡ̚	153	kx
14	ë	34	ə̆	54	o	74	β	94	t̝	114	lˠ	134	ɡ̊		
15	ẽ	35	ə̥	55	ö	75	β̝	95	t’	115	θ	135	x		
16	ɛ	36	ə̃	56	õ	76	β̃	96	t̚	116	ð	136	ɣ		
17	ɛ̈	37	ə̰	57	ʌ	77	f	97	d	117	ɕ	137	ɣ̟		
18	ɛ̃	38	ɚ	58	ʌ̃	78	v	98	d̪	118	ʑ	138	ɰ		
19	æ̞	39	ɜ	59	ɔ	79	v̥	99	d̚	119	ʑ̥	139	ʔ		
20	æ	40	ɝ	60	ɔː	80	ʋ	100	d̥	120	ʃ	140	h		

付録B

距離予測実験で使用した話者一覧

第 4 章，及び第 5 章の実験で使用した話者 369 人の ID を列挙する．なお， SAA では話者 ID に話者の母語を使用している．

afrikaans2	dutch15	english148	english60	fulfuldeadamawa1
afrikaans3	dutch16	english149	english62	german13
afrikaans4	dutch18	english150	english65	german15
amazigh2	dutch19	english151	english67	german16
arabic16	dutch20	english153	english68	german17
arabic17	dutch21	english155	english69	german20
arabic18	dutch23	english157	english70	german21
arabic21	dutch24	english158	english71	german24
arabic22	dutch25	english159	english72	german6
arabic9	dutch27	english160	english74	german7
armenian4	dutch28	english162	english75	german8
armenian6	dutch29	english163	english76	german9
bambara1	dutch4	english165	english77	greek4
bambara4	dutch5	english166	english81	greek7
bavarian2	dutch9	english168	english82	gujarati2
belarusan1	english101	english169	english84	gujarati3
bengali11	english103	english170	english85	gujarati5
bengali6	english104	english171	english86	hainanese1
bengali9	english105	english172	english87	hausa2
bosnian3	english106	english173	english88	hebrew4
bosnian4	english107	english174	english90	hebrew7
bosnian6	english109	english177	english92	hebrew8
bosnian9	english117	english179	english95	hindi4
bulgarian2	english118	english18	english97	hindi5
bulgarian6	english119	english23	english98	hindi6
bulgarian7	english121	english28	estonian1	hindi7
bulgarian9	english122	english29	farsi13	hindi8
cantonese10	english123	english39	farsi7	hungarian1
cantonese11	english124	english40	farsi9	hungarian2
cantonese14	english125	english42	fijian1	hungarian3
cantonese17	english126	english43	finnish3	hungarian4
cantonese18	english128	english44	finnish4	hungarian5
cantonese6	english130	english46	finnish6	hungarian7
cantonese7	english131	english47	french12	icelandic1
cebuano1	english134	english48	french20	indonesian3
chichewa1	english135	english49	french21	indonesian4
croatian1	english136	english50	french23	indonesian5
croatian2	english137	english51	french24	italian11
croatian5	english138	english52	french25	italian14
czech2	english140	english53	french26	italian15
czech4	english142	english55	french28	italian17
dutch10	english146	english58	french33	italian18
dutch11	english147	english59	frisian1	italian19

italian21
italian22
italian23
italian24
italian4
italian6
italian8
italian9
japanese11
japanese13
japanese14
japanese7
japanese8
kabyle1
kambaata1
kazakh1
kazakh2
kiswahili4
kiswahili9
konkani1
korean11
korean20
korean24
korean25
korean7
kurdish7
lamaholot1
luxembourgeois1
macedonian1
malay1
malay2
malay3
malayalam2
maltese1
mandarin1
mandarin12
mandarin13
mandarin15
mandarin16
mandarin20
mandarin24
mandarin25
mandarin26
mandarin27
mandarin28
mandarin7
mandarin8
marathi3
mongolian3
nepali3
nepali4
norwegian2
norwegian3
norwegian4
norwegian6
oriya2
patois1
pohnpeian1
polish15
polish16
polish5
polish6
poonchi1
portuguese12
portuguese15
portuguese2
portuguese20
portuguese21
portuguese23
portuguese26
portuguese27
portuguese29
portuguese31
portuguese7
portuguese8
portuguese9
romanian14
romanian15
romanian4
romanian6
rotuman1
rotuman2
russian19
russian24
russian27
russian3
russian9
sa'a1
serbian11
serbian3
serbian4
serbian6
serer1
shona1
slovak2
slovak4
slovak5
slovenian1
somali2
somali5
spanish26
spanish27
spanish30
spanish34
spanish35
spanish36
spanish37
spanish38
spanish39
spanish41
spanish48
spanish50
spanish53
spanish56
spanish57
spanish60
spanish61
spanish64
spanish65
spanish66
spanish67
spanish79
susu1
swedish3
swedish4
swedish5
swedish9
swissgerman2
tagalog3
tagalog6
tagalog7
tamil3
thai9
tigrigna4
tswana2
turkish1
turkish13
turkish20
turkish21
turkish22
turkish23
turkish24
turkish6
turkish7
turkish9
urdu3
urdu4
vlaams1
wolof3
xiang2
yakut1
yiddish2
yoruba3
yupik1

付録C

IPA から米語音素への変換表

4.4 では，表 C.1 の変換表を用いて，各話者の IPA 発音書きから米語音素書き起しを作成した．なお，変換の際 IPA の装飾記号は全て無視している．

表 C.1: IPA から 米語音素への変換表

IPA	米語音素
ɜɯ, ɜu, ɜʉ, ɜʊ, aɯ, au, aʉ, aʊ, ɑɯ, ɑu, ɑʉ, ɑʊ, ɛɯ, ɛu, ɛʉ, ɛʊ	aw
ɜɪ, əi, əɪ, əʏ, ɵi, ɵɪ, ai, aɪ, ɑi, ɑɪ, ʌi, ʌɪ	ay
ei, eɪ, ɛi, ɛɪ	ey
ɵɯ, ɵu, ɵʉ, ɵʊ, əɯ, əu, əʉ, əʊ, ɘɯ, ɘu, ɘʉ, ɘʊ, oɯ, ou, oʉ, oʊ, ɔu, ɔʉ, ɔʊ, ɒu, ɒʉ, ɒʊ, ɵ, o	ow
oi, oɪ, ɔi, ɔɪ, ɒi, ɒɪ	oy
ʉɯ, ʉu, ʉʉ, ʉʊ, u, ʉ, ɯ	uw
æə, ɛə, ɛa, æ, œ, a	ae
ɑ	aa
ɨ, ʌ, ɜ, ɐ, ə	ah
ɞ, ɔ, ɒ	ao
e, ɛ	eh
ɜˤ, ɚ	er
ø, ɪ, ʏ	ih
i	iy
ɤ, ʊ	uh
j	y

IPA	米語音素
b	b
tʃ	ch
d	d
ð	dh
f, ɸ	f
g, ɣ, ʁ	g
h, x	hh
c	hh y
k	k
l, ʟ	l
m	m
n	n
ɲ	n y
ŋ	ng
p	p
r, ɹ, ʀ	r
s, ɬ	s
ʃ	sh
ɾ, t	t
θ	th
v, β	v
w	w
y	y uw
z	z
ʑ, ʒ	zh