修士論文

CALL応用のための
教師・学習者の発声間における
時間アライメントに関する研究

2010年2月9日

指導教員　峯松 信明 准教授

東京大学大学院 工学系研究科
電気系工学専攻 融合情報学コース
37-086491

高澤 真章

内容梗概

近年, 外国語教育の支援を目的とした CALL(Computer–Assisted Language Learning) システムの需要が高まり, 広く用いられるようになってきた．これらの CALL システムの利用する音声認識技術のほとんどが，スペクトルを用いて音声を表象することに基づいている．この場合音声は, 発音の良し悪しだけでなく，年齢や性別といった，発音とは無関係の要因によっても音声表象は変形してしまう．このため「不一致問題」と呼ばれる, 利用者の身体的特徴により発音評価の結果が不安定になる問題が起こる．現在の音声認識システムでは，年齢や性別などの話者の違いに起因する音声表象の変形に対する頑健性を確保すべく，多数話者の音声を収集し，発声を統計的にモデル化する技術（統計的音響モデル）や，必要に応じて音響モデルを利用者の声色に適応させる技術を用いて, この不一致問題の解決を図っているが，スペクトルを用いて音声を表象していることに変わりはなく，根本的な解決には至っていない．

この不一致問題に対する抜本的解決は，音声に含まれる非言語的な情報のみを除去することである．近年, この非言語的情報をそぎ落した上で発音を表現する「音声の構造的表象」が提案され, この表象を用いて教師と学習者の音声を「発音の良し悪し」のみで比較する CALL システムの構築が行われている．音声の構造的表象に基づくため，これらの CALL システムでは比較する音響イベントの対応は既知である必要がある．既に構築された CALL システムでは，発音の対応から音素同士の対応を明確にしたり，多量の発声を用いて話者の音響モデルを連結学習によって明確にしている．

既存のシステムでは，特定の音素や単語に依存した比較が行われていたが，著者が目標とするのは，構造的表象を用いてフレーズ発声などのより一般的な発音を比較するシステムであり，そのために二発声間の時間アライメントを精度よく求めることが必要である．本研究では，話者性の異なる教師・学習者それぞれの一発声から二発声間における時間アライメント手法に関して検討をおこなった．

目次

第 1 章　序論 **1**
1.1 外国語教育の技術支援 2
1.2 本研究の目的 2
1.3 本論文の構成 3

第 2 章　本研究の背景 **6**
2.1 外国語教育における発音教育の重要性 7
2.2 日本語発音と英語発音の差異 7
2.2.1 母音 7
2.2.2 子音 9
2.2.3 音節 9
2.2.4 アクセント 10
2.2.5 リズム 10
2.3 実用化されている英語教育システム 10
2.3.1 英語教育システムの区分 10
2.3.2 マルチメディア技術に基づく英語教育システム 11
2.3.3 音声情報処理技術に基づく英語教育システム 11
2.4 目標とする英語発音評価システム 12

第 3 章　従来の発音評価システムが抱える不一致問題 **14**
3.1 不一致問題の原因 15
3.2 連続 5 母音認識における不一致問題 15
3.2.1 認識実験に用いた学習用データと評価用データ 15
3.2.2 認識手法と認識結果 15
3.3 発音評価における不一致問題 16
3.3.1 発音評価に用いる英語音声 16
3.3.2 認識手法と認識結果 17

第 4 章　音声の構造的表象 **19**
4.1 音響特徴量ケプストラム 20
4.1.1 FFT ケプストラム 20
4.1.2 Mel ケプストラム 20
4.2 非言語的特徴による音声の歪み 20

4.3 空間写像不変量の一般解 22
4.4 音響的不変構造 23

第 5 章 音声の構造的表象に基づく英語発音評価システムに関する先行研究 24
5.1 音声の構造的表象に基づく英語発音状態記述 25
5.1.1 日本人英語学習者の模擬音声の収録 25
5.1.2 発音状態を示す発音構造の抽出 25
5.1.3 発音状態の記述 26
5.2 音声の構造的表象に基づく英語発音分類 28
5.2.1 日本人英語学習者の模擬音声の収録 28
5.2.2 構造間距離尺度 28
5.2.3 模擬英語学習者の発音構造の分類 30
5.3 音声の構造的表象に基づく英語発音の母音矯正度推定 32
5.3.1 日本人英語学習者の模擬音声の収録 32
5.3.2 母音構造間の要素差異に基づく母音矯正度の推定 32
5.3.3 構造的表象に基づく母音矯正度 32

第 6 章 動的計画法による話者性の異なる二発話の時間アライメント 34
6.1 実験に用いる音声の収録 35
6.2 絶対的物理量に基づく DP マッチングによるアライメント 35
6.3 話者変換を含む DP マッチングによるアライメント 37
6.4 本章のまとめ 39

第 7 章 制約付き HMM 学習による話者性の異なる二発話の時間アライメント 41
7.1 はじめに 42
7.2 特定話者音素 HMM 学習によるアライメント 42
7.3 2 混合 HMM 学習によるアライメント 44
7.4 状態を線状結合した HMM によるアライメント 47
7.5 本章のまとめ 50

第 8 章 結論 52
8.1 本研究のまとめ 53
8.2 今後の課題 54
8.2.1 英語発音音声への適用 54
8.2.2 発話者の倍加によるデータの増量 54

謝辞 55

参考文献 56

発表文献 59

図目次

1.1 複数の単語から抽出された母音群による母音構造の形成 4
1.2 回転とシフトによる二構造の重ね合わせ 4
1.3 一発声からの構造抽出の流れ 5

2.1 Bachman's Model 8
2.2 母音図（日本語母音） 8
2.3 母音図（米語母音） 8
2.4 マルチメディア技術に基づく英語 CAI システム:Rosetta Stone 11
2.5 音声情報処理技術に基づく英語 CAI システム:発音力 12

3.1 単語 HMM および構造的表象での認識率 16
3.2 $P(o|M)$ と $P(M|o)$ による発音評定 18
3.3 構造歪みスコアを用いた発音評定 18

4.1 音声波形からのケプストラム抽出 21
4.2 Mel 周波数とその軸上に等間隔で配置された三角窓 21
4.3 スペクトルに対する水平/垂直方向の音響歪み 23
4.4 音声に内在する音響的普遍構造 23

5.1 各状態における発音状態を表す樹形図 27
5.2 回転とシフトによる二構造の重ね合わせ (再掲) 29
5.3 構造的表象に基づく 96 発音構造 ([12 話者 A∼L] × [8 発音状態 1∼8]) の分類 31
5.4 母音実体間の差に基づく 96 発音構造 ([12 話者 A∼L] × [8 発音状態 1∼8]) の分類 31
5.5 自動推定された母音矯正順序 33

6.1 男性話者 A と女性話者 B の「こんにちわ」発声における DP パス 37
6.2 物理量を直接参照する DP マッチングでの認識率 38
6.3 話者変換を含む DP マッチングでの認識率 40

7.1 3 発声で連結学習した特定話者 HMM によるアライメント結果 (**7.2 節**，上:男性 A，下:女性 B) 43
7.2 特定話者音響モデルによる強制アライメントでの正解率 (**7.2 節**) 44
7.3 2 混合 HMM と 2 状態線状結合 HMM 45

7.4 二話者の 6 発声で連結学習した 2 混合の音響モデルによる「こんにちは」のアライメント結果 (**7.3 節**，上:男性 A，下:女性 B) 45
7.5 二話者の 6 発声で連結学習した 2 混合の音響モデルによるアライメント結果 (**7.3 節**，上:男性 A「こんにちは」，下:女性 B「こんにちわぃ」) 46
7.6 二話者の 6 発声で連結学習した 2 混合音響モデルによる強制アライメントでの正解率 (**7.3 節**) 47
7.7 線状結合 HMM を利用した二発話の共鳴音部における状態アライメント . . 47
7.8 二話者を線状結合した音響モデルによる「こんにちは」のアライメント結果 (**7.4 節**，上:男性 A，下:女性 B) 48
7.9 二話者を線状結合した音響モデルによるアライメント結果 (**7.4 節**，上:男性 A「こんにちは」，下:女性 B「こんにちわぃ」) 49
7.10 二話者を線状結合した音響モデルによる強制アライメントでの正解率 . . . 49

表目次

2.1 日本語・英語の主な子音の対照表 9
2.2 日本語のモーラと英語のシラブルの構造的差異 9

3.1 実験で使用した三種類の英語発音 17

5.1 収録に使用した単語と切り出される母音 26
5.2 日本語母音・米語母音の置換表 26
5.3 母音置換によって模擬された 6 つの日本人英語発音 28
5.4 音響分析条件（**5.1 節**） 28
5.5 母音置換によって模擬された 8 種類の発音状態 29
5.6 音響分析条件（**5.2 節**） 31

6.1 収録に使用した日本語単語 36
6.2 音響分析条件（**6.2 節**） 37
6.3 音響分析条件（**6.3 節**） 39

7.1 音響分析条件（**7.2 節**） 43
7.2 20msec まで許容した際の正解率 [%](**7.2 節**) 44
7.3 音響分析条件（**7.3 節**） 45
7.4 20msec まで許容した際の正解率 [%](**7.3 節**) 46
7.5 音響分析条件（**7.4 節**） 48
7.6 20msec まで許容した際の正解率 [%](**7.4 節**) 50

第1章

序論

1.1　外国語教育の技術支援

近年, 外国語教育の支援を目的とした CALL(Computer–Assisted Language Learning) システムの需要が高まり, 広く用いられるようになってきた．これまで CALL システムといえば視聴覚室や LL（Language Learning）教室などで用いられるものが多かったが，メディアの多様化やハードウェアの高性能化も相まって, 専用端末やパーソナルコンピュータを用いる高価なものだけでなく, iPhone や Nintendo DS などの携帯端末上で利用可能な廉価な外国語学習ソフトウェア [3, 4] も現れ始めており，気軽に利用できるなど利用者には裾野が広がっている．従来はディクテーションなどテキスト入力のソフトウェアが大半だったが, 最近では発音の良し悪しを評価するソフトウェアも多くなってきた．以後, 特に表記をしない限り, 外国語発音学習支援システムを「CALL システム」と呼ぶことにする．

多くの CALL システムは従来の音声認識技術を利用している．この場合音声は, そのスペクトルを用いて表象されるため，発音の良し悪しだけでなく，年齢や性別といった，発音とは無関係の要因によっても音声表象は変形してしまう．このため「不一致問題」と呼ばれる, 利用者の身体的特徴により評価結果が不安定になる問題が起こる．実現されている CALL システムのほとんどがこの不一致問題を抱えていることから, 教育現場への導入に懐疑的な意見も報告されていた [5]．現在の音声認識システムでは，後者に起因する音声表象の変形に対する頑健性を確保すべく，多数話者の音声を収集し，発声を統計的にモデル化する技術（統計的音響モデル）や，必要に応じて音響モデルを利用者の声色に適応させる技術を用いて, この不一致問題の解決を図っている．

根本的にこの不一致問題を解決するには, 音声に含まれる非言語的な情報を除去することが必須である．近年, この非言語的情報をそぎ落した上で発音を表現する「音声の構造的表象」が提案され, この表象を用いて教師と学習者の音声を「発音の良し悪し」のみで比較する CALL システムの構築が行われている．

1.2　本研究の目的

研究目的の明確化に先立ち, 先行研究である音声の構造的表象を利用した CALL システムについて簡単に整理する．

[7] では, **図 1.1** に示すように, 学習者に英語 11 単母音を含む 11 単語を発声させ，母音区間のケプストラム系列を分布として推定し，各母音間の距離のみによって表される 11×11 の母音構造を得る．これと教師の母音構造とを比較して発音評価を行う．

学習者，教師の母音行列を S，T とすると，次式で定義される構造間距離は，**図 5.2** に示す，学習者と教師の二構造を重ねたときの，対応する二点間距離の総和の最小値に比例することが実験的に示されている．

$$D(S,T) = \sqrt{\frac{1}{11}\sum_{i<j}(S_{ij} - T_{ij})^2}$$

また [8] では，上式を母音 i 毎に分解し，母音 i に対する構造歪みスコアを定義することで，発音矯正が最も必要な母音を特定する機能も実装されている．

[32] では，この手法を母音だけでなく子音にまで拡張している．約 70 文の発声から各音素の HMM を構築し，全音素間の距離行列を求めて構造間距離を計算し，発音評価を試みている．模擬的に作られた様々な体格の学習者の発声に対して，一人の教師から得られた構造だけで高精度な発音評定が可能であることを示している．

このように構造的表象は，年齢・性別の情報をそぎ落とした上で教師，学習者間の発音比較を可能にしている．しかし，構造間距離の定義 (上式参照) から分かるように，二構造間において頂点間の対応が自明である必要がある．この対応が不明だと，どのように二構造を重ねるべきかが不明となってしまう．母音構造の場合は，各単語発声において意図された母音は自明であり，この母音の検出も比較的容易に実行できる．全音素構造の場合も，発声した音素列は自明であり，また，発声数を増量し連結学習を通じて音素モデルを構築することで二構造の対応を明確化している．

CALL システムへの応用であることから，学習者の目線も考える必要がある．学習者にとって単母音のみの発音評価は毎回同じ単語を用いるため味気なく，逆に多量の英文の音読ではタスクが重過ぎ，ともに学習意欲を維持することは難しいと考えられる．また，特定の単語や文章を繰り返し利用することで，その単語や文章でのみ発音が改善され，他の単語や文章における発音では依然として日本語的発音に留まってしまう恐れもある．つまり，学習意欲を保ち，単語や文章に偏りなく自然に発音できるように支援するため，学習者に過度のストレスを与えることなく，同時に発話内容にバリエーションを持たせる必要がある．そのため，単語よりも自由度の高い一般的なフレーズ発声での発音評価をおこないたいというのが，研究の動機である．

教師の発声と，それを真似た学習者の発声を共に構造化して両者を比較することを考える．この場合，一発声からの構造化が必要となる．[10] では構造的表象に基づく音声認識において，状態数の多い HMM を一発声から学習することで一発声の構造化を実装している (**図 1.3** 参照)．しかしこの方法を直接 CALL に応用すると，(発音誤りがない条件下においてさえ) 教師発声における第 n 番目の分布と学習者発声における第 n 番目の分布が持つ言語的機能（音素の種類など）が頻繁に異なってしまう．つまり構造抽出をしても，二構造の各頂点の対応がとれないために比較ができない．

よって，構造表象に基づき教師・学習者間の一発声間比較を行うには，両者の時間アライメントを正確に抽出しながら構造化をおこなう必要がある．本論文では，時間アライメントの抽出を種々の方法を用いて検討する．

1.3　本論文の構成

まず，本章において，本研究の目的について述べた．**第 2 章**では，本研究の背景について述べる．**第 3 章**では，英語発音評価システムを構築する上で障害となる「不一致問題」について述べる．**第 4 章**では，英語発音評価システムの核を成す「音声の構造的表象」について説明する．**第 5 章**で，本表象に基づく英語発音評価システムの各要素について先行研究を紹介する．**第 6 章**では，動的計画法による話者性の異なる二発声の時間アライメントを行う方法を検討する．**第 7 章**では，一発声から構成した構造を話者性の異なる二話者

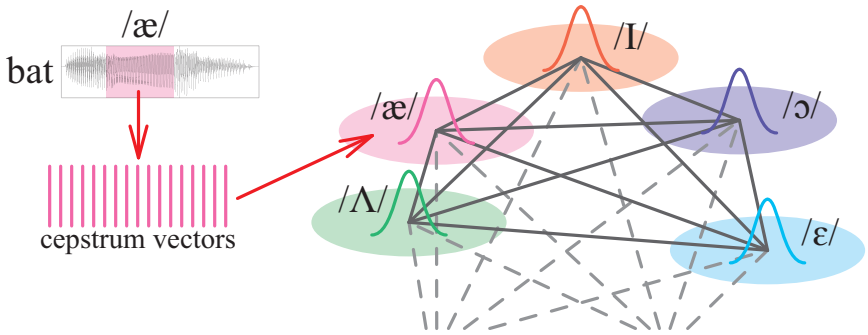


図 1.1: 複数の単語から抽出された母音群による母音構造の形成

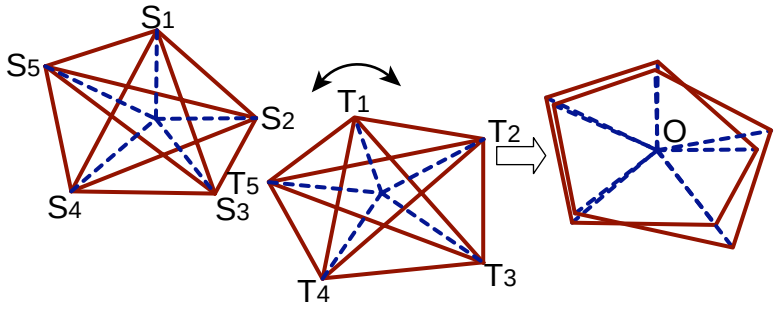


図 1.2: 回転とシフトによる二構造の重ね合わせ

間で比較することを目的として，時間アライメント及び構造の導出を考える．最後に，**第 8 章**にて，本研究のまとめと今後の課題について述べる．

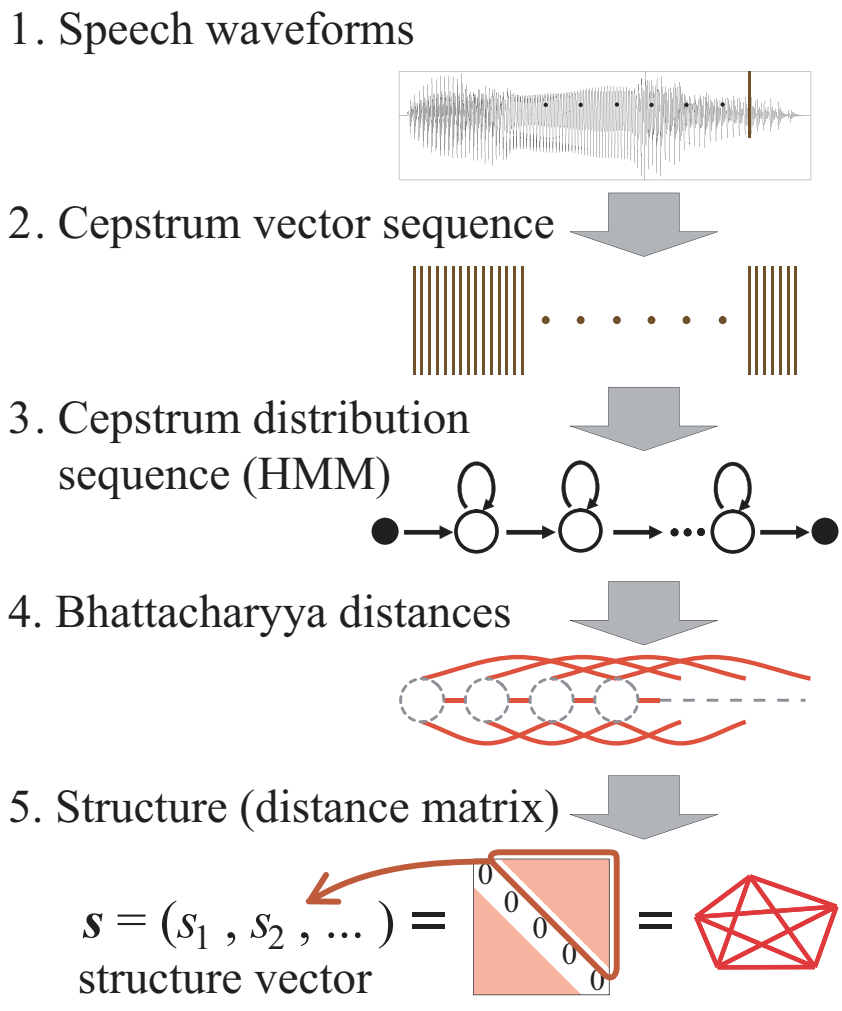


図 1.3: 一発声からの構造抽出の流れ

第2章

本研究の背景

2.1　外国語教育における発音教育の重要性

本論文の表題の「CALL 応用」は外国語教育の中でも発音教育を支援するものである．外国語教育の中で，この発音教育が重要な位置を占める理由を以下に記す．

外国語学習における学習目標として,「読む」・「書く」・「聞く」・「話す」の 4 技能の習得が挙げられる．近年ではこの中でも話す・聞く能力を重視したコミュニケーション能力の育成に主眼が置かれている．

外国語によるコミュニケーション能力を評価する場合には，上記のような 4 技能以外にも様々な観点から評価がなされるべきであり，例えば Bachman によって**図 2.1** のように外国語コミュニケーションにおける様々な要素の分類がなされている [11]．そして，これらの各要素の中で，発音（*P*ronunciation）はコミュニケーション能力を評価する上で非常に重要な要素となっており，下記のようなモデル式が存在している [12].

$$\text{comm.} \simeq \text{pron.} \bullet \text{lex.} \bullet (1 + \text{syn.} + \text{rhet.} + \text{illoc.} + \text{soc.}) \tag{2.1}$$

発音能力 (pron.) と語彙力 (lex.) の二つがコミュニケーション能力 (comm.) に大きく寄与していることが示されている．つまり，外国語学習をその言語によるコミュニケーション能力の習得として考えた場合，発音学習は非常に大きな位置を占めており，その技術支援には重要な意味を持つといえる．

2.2　日本語発音と英語発音の差異

本節では，外国語の一例として英語と，日本語の発音の差異について述べる．

2.2.1　母音

日本語における母音図を**図 2.2** に示す [13]．母音図は発音時の舌の位置を模式的に表しており，縦軸は舌の高さ (口の開き具合) を，横軸は舌の部位 (前舌か後舌か) を表現している．日本語の母音は/あ/,/い/,/う/,/え/,/お/[1][2]の 5 つである．一方，英語における母音図を**図 2.2** に示す[3]．図は二重母音に関しては省略していある．日本語母音はわずか 5 つであるのに対し，英語母音は，短母音 5 つ (/ɪ/ , /ʊ/ , /ɛ/ , /ʌ/ , /æ/) と長母音 4 つ (/i/ , /u/ , /ɔ/ , /ɑ/) の合計 9 つ，二重母音は/eɪ/ , /ɔɪ/ , /aɪ/ , /aʊ/ , /oʊ/の 5 つ，さらに弱母音 (/ə/) と，アメリカ英語独特の r 色の母音 (/ɚ/など) を入れると全部で 22 種類になる [14].

一般に，常用している言語の音素の範囲内の違いには鈍感であり，異なる音素にまたがる際には，それが仮に音響的に小さな差異であっても比較的敏感であるといわれている [15]. 例えば舌が下に下がる母音は，日本語では/あ/のみだが，英語では/ɑ/ , /æ/ , /ʌ/がある

[1]以下，音素に関しては/ /で示し，発音記号に関しては [ ] で示す.

[2]日本語母音は/a/，/i/，/u/，/e/，/o/と表記するのが慣習であるが，本論文では米語母音との混同を避けるため，/あ/,/い/,/う/,/え/,/お/と表記した.

[3]英語の母音は方言による違いが大きいが，以下ではアメリカ英語の一般的な方言 (米語，General American) を扱うものとする.

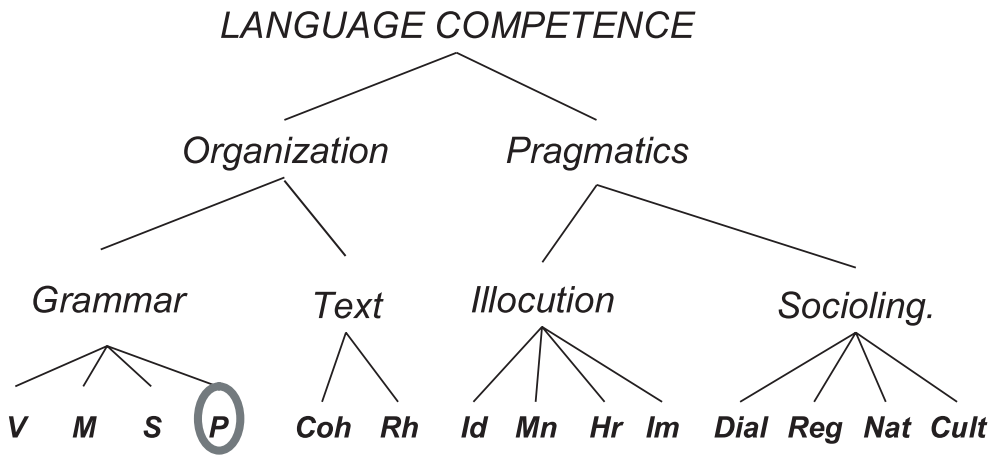


図 2.1: Bachman's Model

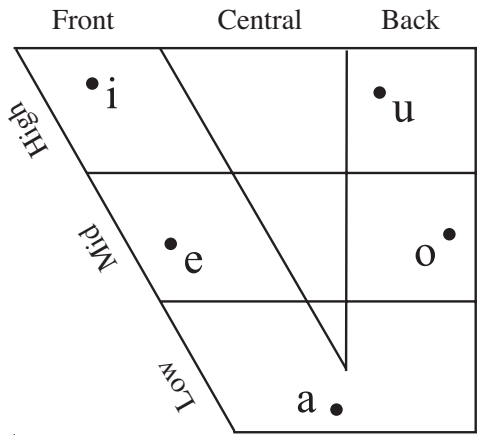


図 2.2: 母音図（日本語母音）

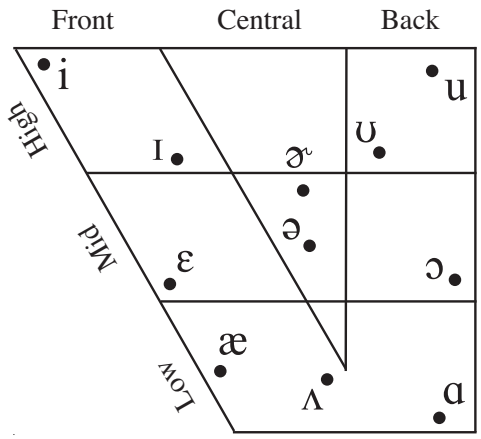


図 2.3: 母音図（米語母音）

ため，日本人学習者はこれら 3 つの母音の差異には鈍感である．/ɪ/ と /i/や/ʊ/ と /u/などの差異についても同様である．

日本語における長母音は音質的に見ると短母音が並んだものとしてとらえることができるが，英語の長母音は短母音を長く発音したものではなく，例えば beat の [i] と bit の [ɪ] は発音の仕方も音質も異なる別種の音である．二重母音も同様に，[aɪ] は [a]+[ɪ] ではなく，[aɪ] で 1 つの音である．そのため，長母音と二重母音は短母音とは別の音として取り扱われる．また，英語には弱母音 (schwa) と呼ばれる，弱く発音されるときのみの現れる母音が存在する．日本語では基本的にどの母音も同じ強さで発音されるため，日本人はこの弱母音の発音が特に苦手であると言われている．

表 2.1: 日本語・英語の主な子音の対照表

調音方法＼調音位置			唇音		歯	歯茎	後部歯茎	そり舌	硬口蓋	軟口蓋	口蓋垂	声門
			唇	唇歯								
閉鎖音		日	p / b			t / d				k / g		ʔ
		英	p / b			t / d				k / g		ʔ
摩擦音		日	ɸ			s / z			ç			h
		英		f / v	θ / ð	s / z	ʃ / ʒ					h
破擦音		日				ts / dz	tʃ / dʒ					
		英					tʃ / dʒ					
鼻音		日	m			n			ɲ	ŋ	ɴ	
		英	m			n				ŋ		
弾き音	流音	日				ɾ						
		英				l ɹ		ɻ				
閉鎖音	渡り音	日	w						j	w		
		英	w						j	w		

表 2.2: 日本語のモーラと英語のシラブルの構造的差異

モーラ	基本は 母音（V），子音+母音（CV） 他に 特殊拍（撥音（N），促音（Q），長音）が存在する
シラブル	母音を中心にその前後に 0 個以上の子音が連結した形をとる． 最長シラブルは CCCVCCCC.

2.2.2　子音

子音に関しては，調音位置と調音方法によって**表 2.1** のように分類される [14]．無声/有声の対で示しており，発音記号により表記してある．日本人が特に苦手であると言われる英語の/l/(発音記号では [l]) と/r/(アメリカ英語は [ɻ]，イギリス英語は [ɹ]) は，日本語ではラ行の子音である [ɾ] に置き換えられてしまうことが多い．また，/f/,/v/や/θ/,/ð/は日本語にない子音であるため，/v/と/b/，/f/と/h/，/θ/と/s/，/ð/と/z/などの混同が起こりやすい．

2.2.3　音節

英語は音節，あるいはシラブル (syllable) と呼ばれるものを発声の基本的単位とする言語であり，日本語は音節より小さな発声の単位であるモーラ (mora) を基本的単位とする言

語である [16]．**表 2.2** にモーラとシラブルの構造的差異を示す．日本語の母音数は 5 種類であるが，英語の母音の種類は約 20 種類となっており，シラブル/モーラの構造的差異から，モーラの種類数は約 100 であるが，シラブルは約 10,000 種類数を持つと言われる．

2.2.4　アクセント

日本語では，音の高さの変化によりアクセントを表す「高さアクセント」であり，音声情報処理においてはピッチのみを用いて記述される．一方，英語では「強さアクセント」と呼ばれる．ピッチ，パワー，持続時間，母音の音質などがアクセントによって変化する．以降，特に英語のアクセントを「強勢」(stress) と呼ぶ.「強勢」とは，ある音節を発音するに当たって音源である呼気が強くなったりその量が多くなると喉頭や調音器官が緊張して調音のエネルギーが強くなり，聞き手が感じる音の大きさ (loudness) が増大する現象をいう．強勢を受けた音節はピッチが高くなり，音が長めになる．強勢は強強勢 (strong stress) と弱強勢 (weak stress) とに二分され，強さアクセントでは全ての音節はいずれかを受ける．語中にある音節に置かれた強強勢は単語強勢（word stress）と呼ばれる．また，文中の特定の音節が持つ強勢は文強勢（sentence stress）と呼ばれる．語義を持つ内容語（content word）強い文強勢を受け，機能語（function word）と呼ばれる語義が希薄で主として内容語同士の文法的関係を示す働きをする語の文強勢は弱い．

2.2.5　リズム

リズムにおける等時性に関しても日本語と英語は異なる．日本語は一つ一つのモーラが等間隔で発音される,「モーラ拍のリズム」と呼ばれる特徴を持っている．一方の英語では，ドイツ語，ロシア語と同様に強い強勢が等間隔に繰り返される,「強勢拍のリズム」を持つといわれる．特に，英語のリズムを形成するものとして，ある一つの強勢音節から次の強勢音節までを「脚」(foot) という．

2.3　実用化されている英語教育システム

2.3.1　英語教育システムの区分

英語教育システムの区分として，ここでは 2 つ紹介する．一つは CAI(Computer Assisted/Aided Instruction) システムである．CAI システムとは，指導者が学習者を指導する時の主要な機能をコンピュータに代行させ，あらかじめ用意された教材 (コースウェア) に従って学習者個々に応じた最適な学習指導を行うシステムのことである．もう一つは CALL(Computer Aided Language Learning) システムである．CALL システムとは，コンピュータが生身の先生に取って代わるのではなく，コンピュータを使って言語学習の促進を図ることが主な目的のシステムのことである．

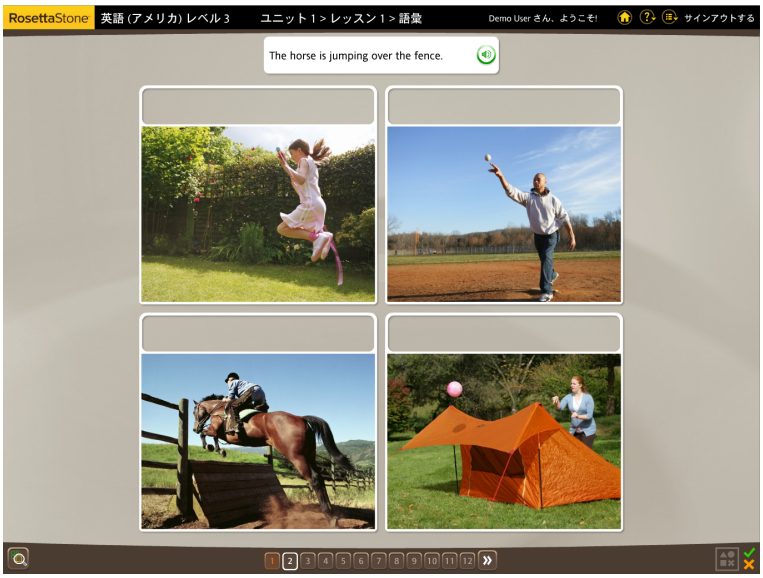


図 2.4: マルチメディア技術に基づく英語 CAI システム:Rosetta Stone

2.3.2　マルチメディア技術に基づく英語教育システム

一般的に商用の英語 CAI ソフトには音声情報処理技術を用いたものよりも，ビデオ映像などのマルチメディア技術を伴ったものが多い．その一例である Rosetta Stone[17] を**図 2.4** に示す．ネイティヴによる単語やフレーズ，文音声を聞き，それに合致するビデオ映像を指摘する形式が主な内容となっている．音声の再生中にテキストを字幕として表示することができるなど，マルチメディア技術を生かした構成になっている．本教材ではスピーキングのセクションも用意されているが，学習者の発音に対するフィードバックは音声波形を提示するのみでほとんど無く，音声処理は特に行われていないと思われる．また，千葉大学の協力でメディア開発センターが英語の CALL 教材 [18] を発表しているが，これもリスニング中心で，リスニング後に設問に答える形式の教材となっている．

2.3.3　音声情報処理技術に基づく英語教育システム

音声情報処理技術を用いた英語 CAI ソフトの一例である発音力 [19] を**図 2.5** に示す．この例のように，学習者が発声した音声を波形として表示し，教師音声の波形と比較して表示する形式のものが多い.「調音」「強弱」「高低」「リズム」等といった観点から比較が行われる．教師発音と異なる場合は，口内の断面図を図示し舌の動きなどを図示することで発音の仕方を教示する．他の音声認識技術を用いた英語 CALL システムである [20] などでは，韻律的特徴よりも音韻学習が学習内容の中心になっている．

図 2.5: 音声情報処理技術に基づく英語 CAI システム:発音力

2.4　目標とする英語発音評価システム

中学校や高校などにおいておこなわれることの多い，英語教師 1 人に対して数十人の生徒という「一対多」の英語の授業における発音教育は，それぞれの生徒の発音状態をその都度をチェックするということに関して，理に適っているとは言えない．発音学習において最も大切なことは，学習者が発声した発音の状態をその都度チェックし，結果を学習者にフィードバックすることであるが，一対多の授業では英語教師が全生徒の発音をチェックをするには時間が足りないからである．ここで情報処理によって効果的な支援をおこなうためには，学習者の発音状態を正確に記述し，学習者にわかりやすい教示をおこなえるような枠組みが必要である．

学習者の発音を波形やスペクトルで表示することは，音響音声学に加えパーソナルコンピュータを用いれば可能であり，これらを基にした発音学習アプリケーションは前節で紹介したように広く開発が行われている．しかし，波形やスペクトルには学習者の発音の善し悪し以外にも性別・年齢・体格・マイク・収録環境などの様々な情報が含まれる．発音学習という観点から見ればこれらは全て雑音であり，利用者環境とシステムとの間での相性問題（不一致問題）が常につきまとうことになる．このような技術を基にした発音学習

支援は結果として不安定なシステムとなることは明らかである．

また，音声学的な視点からの学習者の記述は，基本的には「単音」や「発話」を単位とすることとなる．つまり，それぞれの発話に対する発音の善し悪しに対しては記述可能であるが，学習者が今どのような状態にあるのかという「学習者」を単位とした記述に関しては，音声学的な観察眼のみでは不十分であるといえる．学習者の記述において求められるものは，単音や発話の記述ではなく，学習者を「英語音生成システム」としてシステムを規定するパラメータ群を考えたときの，そのパラメータ群を推定する形で学習者の発音状態を記述することである．つまりそれは音韻論的視点から学習者を記述することであり，これまでの音声学的観察眼から音韻論的観察眼への質的変化が必須であることを示している．

学習目標の設定についても検討する必要がある．従来の発音学習システムは母語話者との音響的照合を行うことで評定を行っているが，近年では native–sounding な発音ではなく，intelligible な発音が求められるようになってきている．つまり学習者が目指す目標が多様化しているため，それぞれの学習者が目指す目標を自分で設定することができ，かつ，現在の学習者の状態から目標に到達するまでの道のりが学習者に対して明瞭である必要がある．

発音学習支援システムに必要とされているものは以下の 3 点にまとめられる．

- 学習者の「現在」の発音状態を記述可能 · · · 学習者を「英語音生成システム」と捉え，学習者の現在の発音状態を正確に記述することが求められる．
- 実環境において安定に動作 · · · 低年齢者が学習者となった場合にも問題なく動作するなど，相性問題が原則的に起こらない技術でなければならない．
- 学習者が各自の学習目標を設定可能 · · · ビジネスマンが現場で通じる発音の習得を目指すなど，全ての学習者が母語話者のような発音を目指しているわけではなく，学習者が各人の学習意欲に応じた目標を適切に設定できなければならない．

このような発音学習の支援システムを構築するには，音声から性別・年齢・体格・マイク・収録環境などの雑音をそぎ落とし，発音状態を示す情報を抽出する必要がある．近年，音声からこれらの雑音となる非言語的特徴をそぎ落として発話を表象する手法である「音声の構造的表象」が提案されており [6]，この表象を利用して教師と学習者の発声を「発音の良し悪し」のみに基づいて比較するシステム構築がおこなわれている [7, 8]．

構造的表象に基づいて教師と学習者の発音を比較する場合，二つの発声に含まれる言語イベント (音素や音節) の時間アライメントを正確に求める必要がある．[7, 8] では，アライメントが比較的容易に求められる発声として，英語の単母音に焦点を絞っていた．本研究が目標とするのは，構造的表象を用いてフレーズ発声などのより一般的な発音を比較するシステムであり，そのために二発声間の時間アライメントを精度よく求めることが目的である．

第3章

従来の発音評価システムが抱える不一致問題

3.1　不一致問題の原因

従来の発音評価システムは音響特徴量としてケプストラムを利用し，その値を直接参照するため,「不一致問題」を抱えている．ケプストラムには話者の声道形状の特性，マイクロフォンなどの音響機器の特性，服の衣擦れや周囲の話し声などの背景雑音などの非言語的特徴が内包されており，発音の良し悪しの純粋な評価を妨げる要因となっている．

この不一致問題に対して，これまでは大量の音声を用いて個々の言語事象を統計的に統計的にモデル化することで対処してきた．しかし，モデル化に用いる特徴量には非言語的特徴が含まれていることに変わりなく，結局，話者によっては正しい認識がおこなわれないシステムとなっていた．多量のデータを用いたモデル化は，不一致問題の解決策にはならなかった．

現在，この不一致問題に対して話者適応など種々の適応技術が用いられている．しかし，これらの適応技術を発音の評価に利用すると，話者の違い (話者性) を吸収するばかりか間違った発音までも吸収してしまい，誤発音を許容する音響モデルになってしまうこともある [21]．話者の違いと発音の良し悪しとが，スペクトル包絡という同じ物理現象に基づくにも関わらず，この二つを切り分ける技術がないことが根本的な原因である．

3.2　連続 5 母音認識における不一致問題

本節では，従来手法 (ケプストラムを特徴量に用いた HMM 音声認識) による連続 5 母音音声認識実験を通して，不一致問題の一例を見る．また，後述する音声の構造的表象 (**第 4 章**) に基づく音声認識の結果を示し，不一致問題を解消していることを付記する．

3.2.1　認識実験に用いた学習用データと評価用データ

認識モデルを作るための学習用データには，日本人の成人 8 名 (男女各 4 名) の日本語 5 母音連続発声[1](120 単語) を用いた．評価用データには，学習用データとは異なる日本人成人 8 名 (男女各 4 名) の日本語 5 母音連続発声 (120 単語) を用いた．これに加え，不一致問題を引き起こす身長の低い人間の音声を用意するため，評価用データから音響分析による音声変換 [22] によって，身長が低い人間の音声を作成した．

3.2.2　認識手法と認識結果

学習用データを用いて単語 HMM を作成し，変換された評価用音声の認識を行った．結果は身長が低くなるに従って，認識率は低下した (**図 3.1**，従来手法参照)．これは従来手法が話者の違いに対応できず，不一致問題が起きていることを示している．なお，音声の構造的表象に基づく音声認識 [23] では，身長の変化による認識率の大きな変化は見られず，不一致問題に対応可能であることを示している (**図 3.1**，構造的表象参照)．従来手法では

[1]「あいうえお」「うおいえあ」など，${}_5P_5 = 120$ 通りの連続発声である．

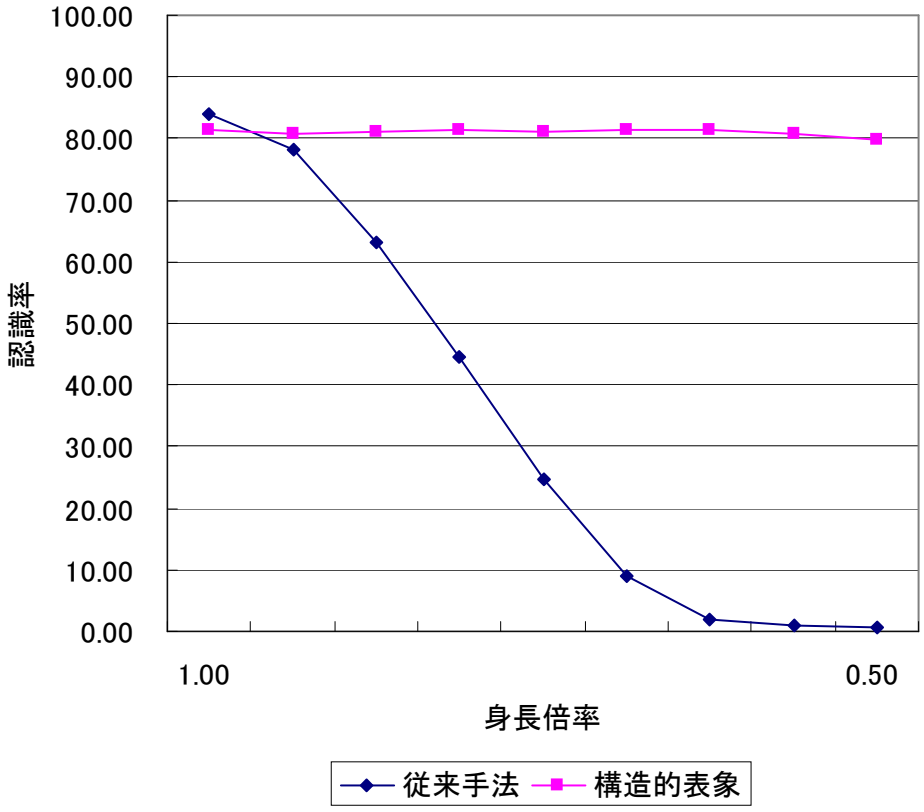


図 3.1: 単語 HMM および構造的表象での認識率

このような不一致問題に対応するために様々な話者適応技術が提案されており，音声認識率を改善することはできる．しかし，これらの話者適応技術が発音評価においては不自然な結果を残す場合がある．この例について次節で述べる．

3.3　発音評価における不一致問題

本節では，不一致問題解消のために広く用いられるパラメータによる発音評価を行い，発音分析の従来手法が不一致問題を抱えていることの一例を示す．また，音声の構造的表象に基づくことにより，不一致問題を解決可能である実験結果について述べる．

3.3.1　発音評価に用いる英語音声

3.2 節と同様，不一致問題を起こす英語音声を用意して実験する．まず，評価対象の音声として，日本人男性 (NM) の熟練した英語音声 (A) を用意した．NM は英語劇経験者であり，英語発音習熟度は極めて高い．次に，比較対象の音声として，この NM が故意に日本語訛り (カタカナ英語発音) で発音した習熟度の低い英語発音 (B) を用意した．さらに，もう一つの比較対象の音声として，英語教師である女性の米語母語話者 (USA/F12) の英語

表 3.1: 実験で使用した三種類の英語発音

話者	USA/F12(F)	NM(A)	NM(B)
性別	女	男	男
年齢	約 50	36	36
マイク	Sennheiser	特価品	特価品
録音室	防音室	リビング	リビング
AD	SONY DAT	PowerBook	PowerBook
習熟度	perfect	good	Japanized

音声 (F) を用意した．用意した音声を**表 3.1** にまとめる．また，結果の考察用に，男性の米語母語話者の英語音声 (M) も用意した．

3.3.2　認識手法と認識結果

熟練した男性英語音声 A を，習熟度の一致しない B，話者が異なる M，話者も性別も異なる F と比較する．比較対象の音声から作成した HMM によりスコアを求める．不一致問題を強く引き起こす比較が A と F である．スコアは，

- モデル (B, F, M) と A 間の尤度スコア $P(o|M)$
- モデル (B, F, M) と A 間の事後確率スコア $P(M|o)$

の二つを用いた．なお，後者の事後確率スコアは，モデルと入力話者間の相性や整合性を正規化する (不一致問題を解消する) 目的で発音評価では広く使われている．

$P(o|M)$ 及び $P(M|o)$ による評価結果を**図 3.2** に示す．それぞれのスコアを元に，評価対象 A の存在位置を内分点として示している．$P(o|M)$ では，A は限りなく B に近くなっており，発音の良し悪しよりも話者の一致・不一致に強く影響を受けていることがわかる．また $P(M|o)$ であるが，入力話者とモデルとの相性を正規化した後のスコアである [24] ので，母語話者の性別に関わらず，同じ値が得られるように意図したスコアであると言える．しかし，結果は**図 3.2** のとおり，異なるスコアとなった．これは，従来の不一致問題を事後確率によって解決する技術が不安定であることを示している．

一方，**5.2.2 節**にある音声の構造的表象に基づく構造間距離を利用したスコアによる結果が**図 3.3** である．こちらでは母語話者の性別に影響されることなく，ほぼ同じ値が得られている．**図 3.3** の下半分は，**3.3.1 節**の他に収録した米語話者，日本人の音声を評価した結果を同一軸上に示したものである．軸の上部は女性，下部は男性の評価結果を示している．米語話者の中に日本人が存在しているが，これは英語発音が堪能なバイリンガル話者である．図より分かるように,「バイリンガル以外のすべての日本人が話者 NM(A) 以上に，話者 NM(B) に近い」と判定されている．このような判定を下すことは，単純なスペクトル照合では不可能である．また，音声の構造的表象に基づく手法が不一致問題を回避し発音習熟度のみに注目していることを示す結果であると言える．

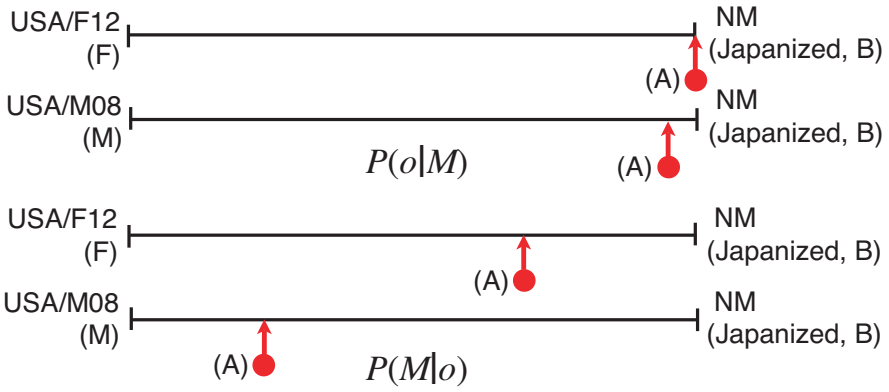


図 3.2: $P(o|M)$ と $P(M|o)$ による発音評定

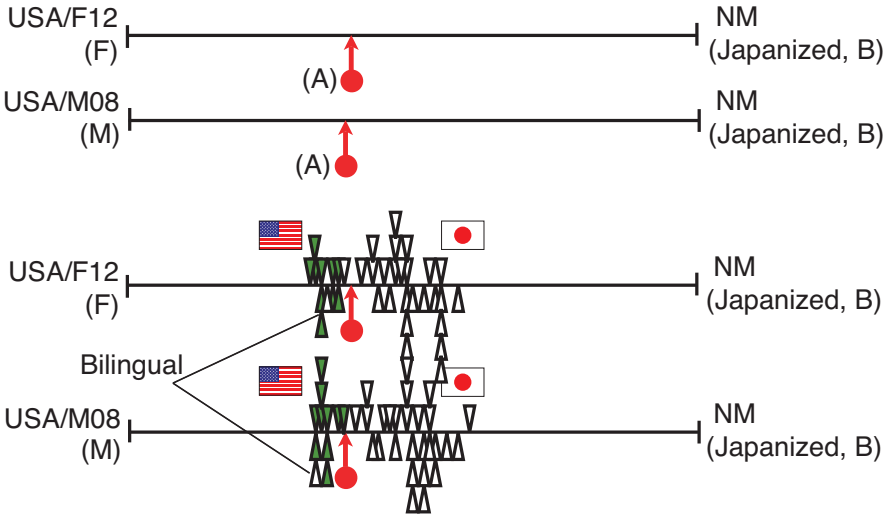


図 3.3: 構造歪みスコアを用いた発音評定

第4章

音声の構造的表象

4.1　音響特徴量ケプストラム

本節では，音響音声学で広く用いられる特徴量ケプストラム (Cepstrum) について述べる．音声の構造的表象においても，特徴量「音響的普遍構造」抽出の過程でケプストラムを利用する．

4.1.1　FFT ケプストラム

音声分析において，音声波形からケプストラムを抽出するまでの様子を**図 4.1** に示す．まず音声波形から，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，スペクトルを抽出する．その後，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) を施したものがケプストラムである．このケプストラムのうちの低次項のみを離散フーリエ変換すると，スペクトル包絡 (Spectrum Envelope) が得られる．声道管の共鳴によって強められた周波数をフォルマント周波数と呼び，これらはスペクトル包絡の山の部分におよそ相当するが，音声の音韻的特徴はこのフォルマント周波数によく表れる．つまりケプストラムは，音声の音韻的特徴を効率良く表すことのできるパラメータである．

4.1.2　Mel ケプストラム

人間の音の高さの感覚は Mel 尺度[1] と呼ばれるが，これは音の周波数に対してほぼ対数に近い特性を示し，人間の周波数分解能は低い周波数ほど細かく，高い周波数ほど粗いことが知られている．これをケプストラムに反映させた音声特徴量が提案されており，MFCC(Mel-Frequency Cepstrum Coefficient) はその一つである．MFCC は，**図 4.2** に示すように Mel 周波数 (Mel 尺度化された周波数) 軸上に等間隔で配置された三角窓を用意し，フィルタバンク分析を行うことで求められる．尚，Mel 周波数 f_{mel} は周波数 f[Hz] に対して，

$$f_{mel}(f) = 2595 \log_{10} \left(1 + \frac{f}{700}\right) \tag{4.1}$$

などの周波数ウォーピングを施すことで得られる．各窓毎に，対応する周波数帯域のパワースペクトルを求め，それに窓の大きさの重みを付けて和をとることで Mel スペクトルが得られる．これに離散コサイン変換を施すことで，MFCC が求められる．

これらの特徴量を直接参照することによって，不一致問題は引き起こされる．次節からは，特徴量「音響的普遍構造」の抽出過程を，不一致問題を回避する仕組みを交えて述べる．

4.2　非言語的特徴による音声の歪み

まず，不一致問題の要因となる，音声に混入し歪みを与える非言語的特徴について整理する．音声に混入する非言語的特徴は，加算性雑音・乗算性歪み・線形変換性歪みの 3 つ

[1] mel という名称は，“melody” に由来し，音高の比較に基づく尺度であることを示している．

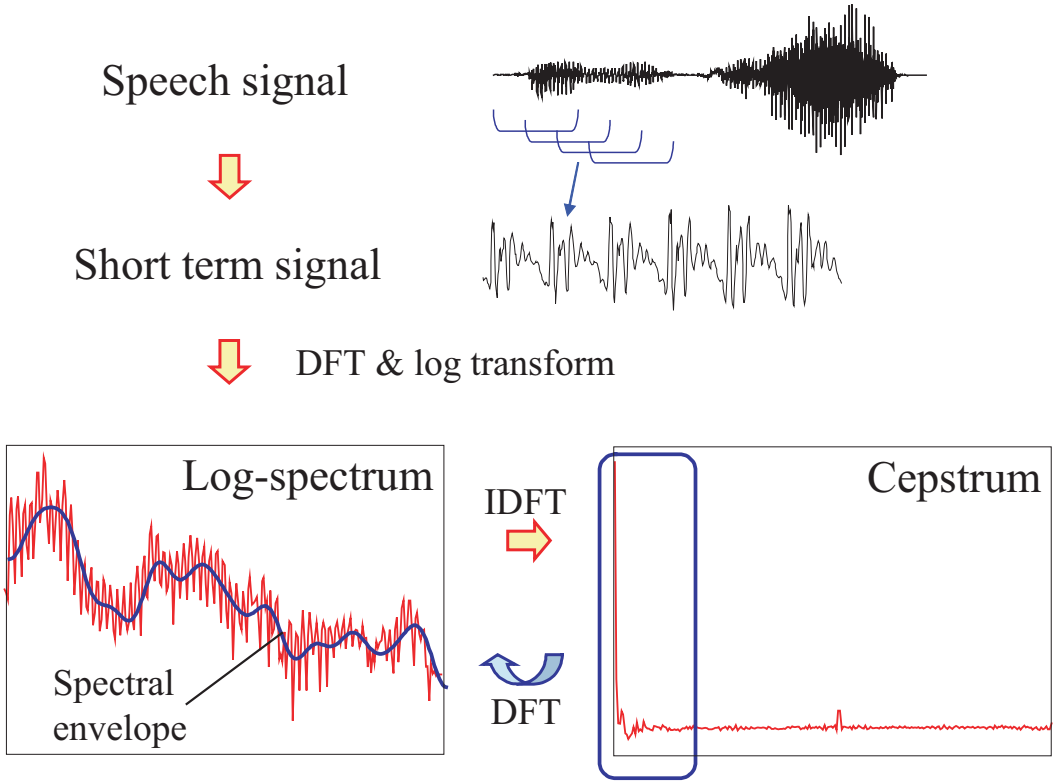


図 4.1: 音声波形からのケプストラム抽出

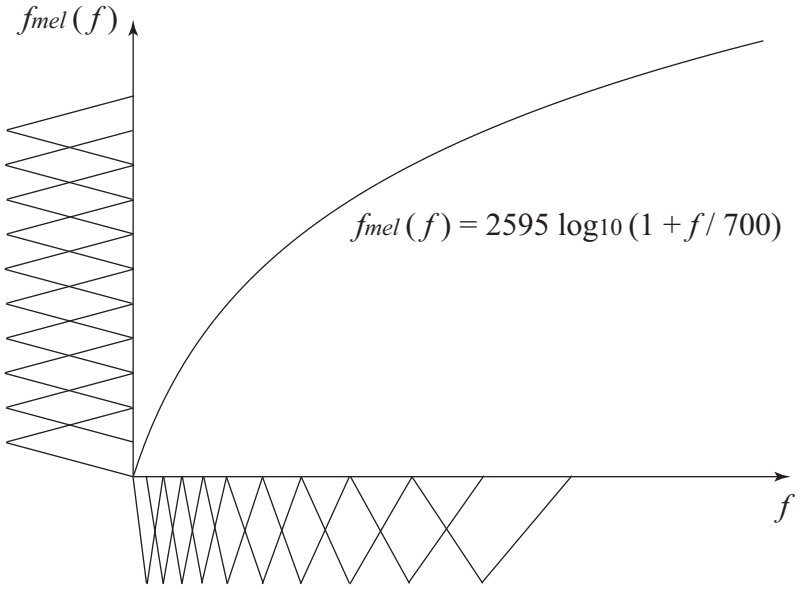


図 4.2: Mel 周波数とその軸上に等間隔で配置された三角窓

に大別される．

加算性雑音とは，時間軸上の加算で表現される雑音であり，例として服の衣擦れや周囲の話し声など背景雑音が挙げられる．これに関しては，発音評価システム利用時の音声収録環境を適切に整えることで回避することができるため，本研究では取り扱う対象から除く．

乗算性歪みは，スペクトルに対する乗算で表現される歪みであり，元のスペクトル特性に対して伝達関数を一つかけ合わせたもの (すなわちフィルタ) に相当する．例としてマイクロフォンなどの音響機器の伝送特性が挙げられる．また，話者の声道形状の違いの一部も乗算性歪みとして扱われる [25]．乗算性歪みを音響特徴量ケプストラムへの影響で表現すると，ケプストラムベクトル $\boldsymbol{c}$ に対するベクトル $\boldsymbol{b}$ の加算 $\boldsymbol{c}' = \boldsymbol{c} + \boldsymbol{b}$ となる．

線形変換性歪みは，$\boldsymbol{c}$ に対する行列 A の乗算 $\boldsymbol{c}' = A\boldsymbol{c}$ で表現される歪みである．話者の声道長の差異や聴取者の聴覚特性の差異を表すために，対数スペクトルに対して周波数ウォーピングが施されるが，単調増加かつ連続である周波数ウォーピングは，$\boldsymbol{c}$ に対する A の乗算で表すことができる [26][27]．すなわち，声道長の差異，聴覚特性の差異は近似的に線形変換性歪みとして扱うことができる．

音声を計算機で取り扱う場合，音声の発生源である話者と収録機器であるマイクロフォンを介することを考慮すれば，回避不可能な音響的歪みは乗算性歪み ($\boldsymbol{c}' = \boldsymbol{c} + \boldsymbol{b}$) と線形変換性歪み ($\boldsymbol{c}' = A\boldsymbol{c}$) である．よって，音声に不可避に混入する非言語的特徴の歪みは，$\boldsymbol{c}$ に対するアフィン変換 (一次変換) $\boldsymbol{c}' = A\boldsymbol{c} + \boldsymbol{b}$ で近似的に表現される．**図 4.3** は，アフィン変換が対数スペクトルに与える影響を示したものである．対数スペクトルの垂直変化が乗算性歪み，水平変化が線形変換性歪みである．また，**図 4.4** は非言語的特徴による歪みを空間写像として表したものである．線形変換や，更には非線形変換をも含めた空間写像に対して不変性をもつ特徴量を用いることで，年齢や性別などの身体的特徴や音響機器の特性に対して不変に，音声を表象することが可能となる．

4.3　空間写像不変量の一般解

連続かつ可逆な線形・非線形空間写像に対して不変性を有する尺度として，f-divergence がある [28]．これは二分布間の距離尺度であり，次式で表される．

$$f_{\mathrm{div}}(p_1, p_2) = \oint p_2(x) g\left(\frac{p_1(x)}{p_2(x)}\right) dx \tag{4.2}$$

$p_i(x)$ は確率密度関数であり，$f_{\mathrm{div}}(p_1, p_2)$ は任意の写像に対して不変である．更に，任意写像に不変となる (2 事象に関する) 尺度は，f-div. のみである [28]．この f-div. を用いて発声を表象すれば，それは話者不変な音声表象となる．筆者らは従来，$g(x) = \sqrt{x}$，$BD(p_1, p_2) = -\log(f_{\mathrm{div}}(p_1, p_2))$ で定義される，バタチャリヤ距離を用いて不変表象を検討してきた．二つの分布の確率密度分布関数をそれぞれ $p_1(x)$，$p_2(x)$ とすると，

$$BD(p_1(x), p_2(x)) = -\ln \int_{-\infty}^{\infty} \sqrt{p_1(x) p_2(x)} dx \tag{4.3}$$

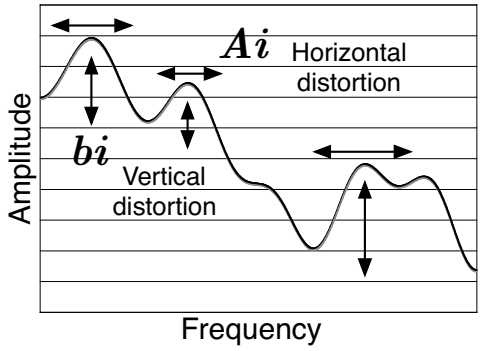


図 4.3: スペクトルに対する水平/垂直方向の音響歪み

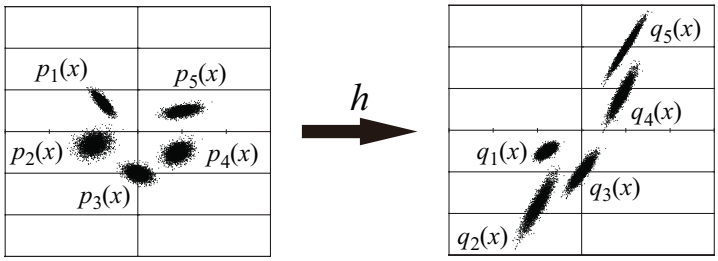


図 4.4: 音声に内在する音響的普遍構造

と表される．$0 \leq \int_{-\infty}^{\infty} \sqrt{p_1(x)p_2(x)}dx \leq 1$ を確率として解釈すれば，**式（4.3）** は自己情報量となり，単位は [bit] となる．また，二つの分布 $p_1(\mu_1, \sum_1)$，$p_2(\mu_2, \sum_2)$ がガウス分布とするとき，この二分布間の BD は

$$BD(p_1, p_2) = \frac{1}{8}(\mu_1 - \mu_2)^T \left(\frac{\sum_1 + \sum_2}{2}\right)^{-1} \mu_{12} + \frac{1}{2} ln \frac{|(\sum_1 + \sum_2)/2|}{|\sum_1|^{\frac{1}{2}}|\sum_2|^{\frac{1}{2}}} \tag{4.4}$$

と表わされる．

4.4　音響的不変構造

空間内に存在する N 点に対して，${}_NC_2$ 個のすべての二点間距離を求めると，その N 点で構成される構造は一意に規定される [6]．一般に距離行列は一つの幾何学的形態 (構造) に対応するため，全二点間距離の行列表象は構造的表象と考えることができる．

今，一つあるいは複数の発声に N 個の言語イベント (音素や音節など) が存在する場合を考える．個々のイベントを分布としてモデル化し，${}_NC_2$ 個全ての二分布間距離を BD で定義し距離行列を求めると，これは話者不変なイベント群の行列表象となる．

このように，言語イベントを分布として扱い各分布間距離を非言語的特徴に不変な距離尺度で距離行列化したものを,「音響的不変構造」と呼ぶ．この音響的不変構造を用いて英語発音評価システムを構築することにより，不一致問題を根本から解決したシステムを実現可能である．

第5章

音声の構造的表象に基づく英語発音評価システムに関する先行研究

本章では，英語発音評価システムの要素技術として 3 つの先行研究を紹介する．

5.1　音声の構造的表象に基づく英語発音状態記述

5.1.1　日本人英語学習者の模擬音声の収録

英語劇及び発音指導経験を持つ日本人男性に米語 11 単母音を/bVt/[1]の形で発声させた．ここで有意味語が存在しない場合は，/t/を/d/とするなどの処置を行い選定した 11 英単語を用い，11 単母音の収録を 1 回ずつ行った．/bVto/(日本語発音の「バト」,「ビト」，...) という形で日本語 5 母音を 5 回ずつ収録した．収録した 11 英単語と切り出される単母音を**表 5.1** に示す．これら $11 + 5 \times 5 = 36$ 個の音声資料を用いて，一部の米語母音を日本語母音と置換する形で，日本人英語発音を模擬した．**表 5.2** に用いた母音置換表を示す．母音置換の組み合わせは，日本人が混同しやすい組み合わせを考慮してあり，混同の様子は**図 2.2**，**図 2.3** の母音図を見ることにより窺うことができる．異なる米語母音が同一の日本語母音と置換される場合は，同一母音・異発声のサンプルを用いた[2]．これらの置換パターンを用いることで，

1. 日本語母音による発音 (いわゆる「カタカナ英語」発音)
2. 日本語母音/あ/に近い母音を矯正した発音
3. 日本語母音/あ/，/い/に近い母音を矯正した発音
4. 日本語母音/あ/，/い/，/う/に近い母音を矯正した発音
5. 日本語母音/あ/，/い/，/う/，/え/に近い母音を矯正した発音
6. 全ての母音を矯正した発音 (米語発音)

と 6 種類の発音状態 (S1〜S6) を規定した．S1 から S6 まで，徐々にカタカナ英語発音が米語発音に矯正されていくようにシミュレートしている．各置換パターンをまとめると**表 5.3** のようになる．

5.1.2　発音状態を示す発音構造の抽出

収録音声から目視により母音部分を切り出し，**表 5.4** に示す音響分析条件の下でケプストラムパラメータを求め，ガウス分布化した．一単語の一発声から母音部のみを切り出しておりデータが少量であるため，発音構造が不安定となることが考えられる．これの対処として，各ガウス分布には最大事後確率 (maximum a posteriori; MAP) 推定を施した [29]．MAP 推定後の全分布間のバタチャリヤ距離を求め，発音構造 (11×11 距離行列) を抽出した．

[1] /V/には/ʌ/や/æ/などの母音が相当する．
[2] 例えば，/ʌ/と/æ/を置き換える場合は，2 つの/あ/の発声を用いてそれぞれの発音を置き換える．

表 5.1: 収録に使用した単語と切り出される母音

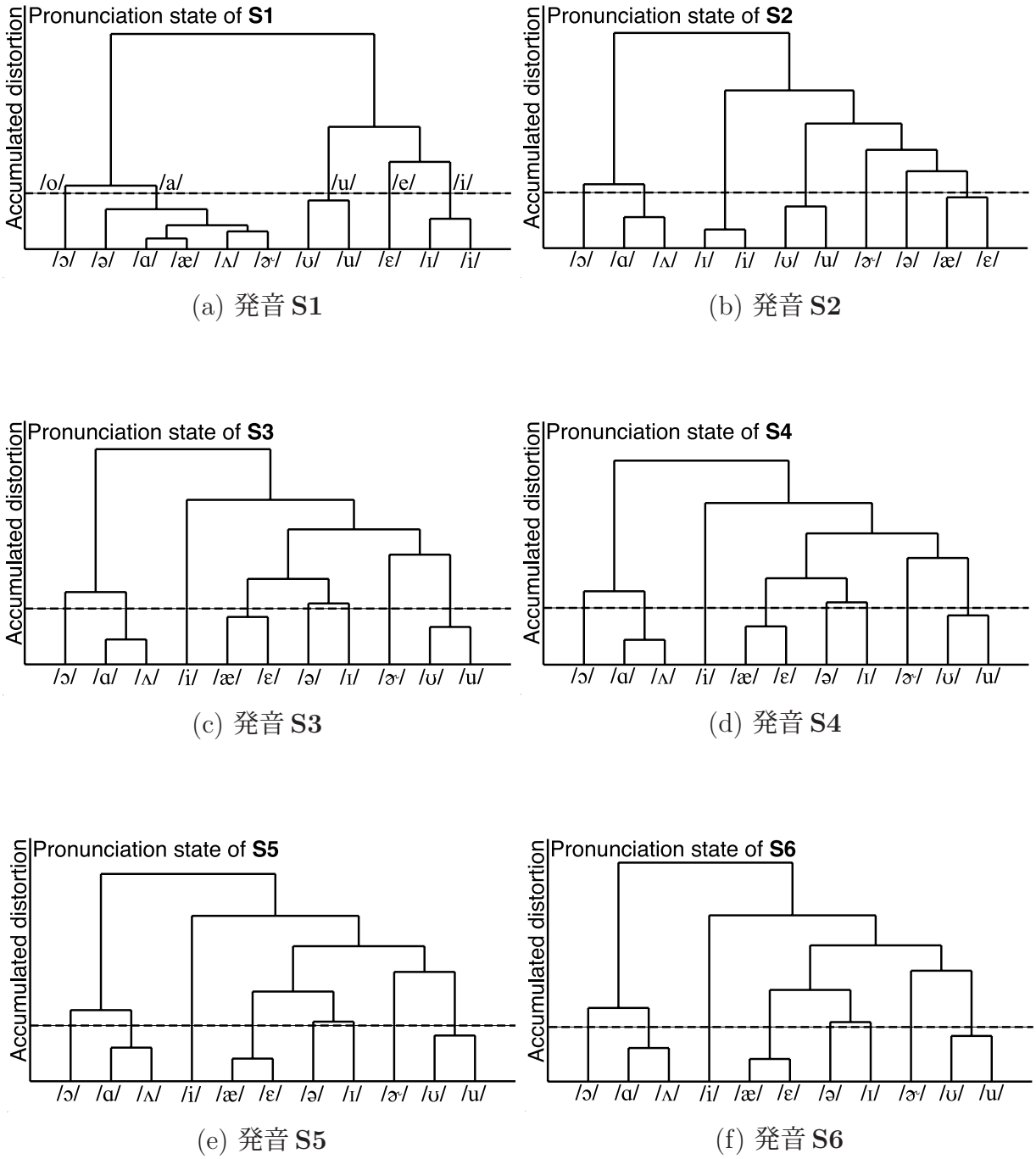
英語単語	単母音	英語単語	単母音	日本語単語	母音
beat	/i/	bought	/ɔ/	バト	/あ/
bit	/ɪ/	put	/ʊ/	ビト	/い/
bet	/ɛ/	boot	/u/	ブト	/う/
bat	/æ/	bird	/ɚ/	ベト	/え/
but	/ʌ/	about	/ə/	ボト	/お/
pot	/ɑ/				

表 5.2: 日本語母音・米語母音の置換表

日本語母音　↔	米語母音
/あ/	/æ/,/ʌ/,/ɑ/,/ɚ/,/ə/
/い/	/i/,/ɪ/
/う/	/u/,/ʊ/
/え/	/ɛ/
/お/	/ɔ/

5.1.3　発音状態の記述

11 × 11 の距離行列で表現される構造を，階層的クラスタリング手法の一つである，Ward 法に基づくボトムアップクラスタリングにより樹形図に表したものが**図 5.1** である．Ward 法とは，クラスター融合の際に発生するクラスター間距離の歪みが最小になるように距離行列を更新する手法である．分析対象のデータはあくまでも模擬音声であるが，完全なる日本語母音の使用によるカタカナ英語発音 (S1) から米語発音 (S6) に至るまでの一つの遷移の様子が示されている．



(a) 発音 **S1**　(b) 発音 **S2**

(c) 発音 **S3**　(d) 発音 **S4**

(e) 発音 **S5**　(f) 発音 **S6**

図 5.1: 各状態における発音状態を表す樹形図

表 5.3: 母音置換によって模擬された 6 つの日本人英語発音

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

A：米語母音を使用，J：日本語母音で置換

表 5.4: 音響分析条件（**5.1 節**）

サンプリング	16bit / 16kHz
窓	窓長 25ms，シフト長 4ms
パラメータ	FFT ケプストラム (1∼10 次元)
HMM	1 混合 monophones（全角分散行列）
トポロジー	3 状態 1 ガウス分布

5.2　音声の構造的表象に基づく英語発音分類

5.2.1　日本人英語学習者の模擬音声の収録

帰国子女，或いは英語劇経験者の日本人 12 名 (男性 6 名，女性 6 名; 各話者を A∼L とする) より，**5.1.1 節**と同様に，11 種類の米語母音と 5 種類の日本語母音を各々1 回，5 回ずつ収録した．これらの米語母音と日本語母音を同一話者内で置換することにより，様々な発音状態を規定する．本実験では，**表 5.5** に示す置換パターンを用い，話者ごとに 8 種類の発音状態 (P1∼P8) を用意した．なお，米語・日本語間の母音置換の対応は**表 5.2** の通りである．複数の英語母音を 1 種類の日本語母音で置き換える場合は，日本語母音はそれぞれ異なる発声を用いるものとした．このようにして模擬された 96 種類の発音状態に対して，下記で定義される距離尺度に基づいて距離行列化し，樹形図化することで分類を試みた．

5.2.2　構造間距離尺度

学習者が発声する 11 母音の母音間距離をすべて求めることにより，11×11 の母音間距離行列で表現される構造 (11 角形) を定義することが出来る．二話者の二つの 11 角形の構造間距離を規定することができれば，それは学習者と学習者の距離を，発音訛のみに着眼した形で (非言語的特徴を無視した形で) 定義することとなる．そこで，収録音声から目視により母音部分を切り出し，**表 5.6** に示す音響分析条件の下でケプストラムパラメータを求め，MAP 推定を用いて分布化した．そして，模擬された 96 種類 (12 話者 × 8 パターン)

表 5.5: 母音置換によって模擬された 8 種類の発音状態

	ɑ	æ	ʌ	ə	ɚ	ɪ	i	ʊ	u	ɛ	ɔ
P1	J	J	J	J	J	J	J	J	J	J	J
P2	A	A	A	A	A	J	J	J	J	J	J
P3	J	J	J	J	J	A	A	A	A	A	A
P4	A	A	J	J	J	A	A	J	J	A	A
P5	J	J	A	A	A	J	J	A	A	J	J
P6	A	J	A	J	A	J	J	J	J	A	A
P7	J	A	J	A	J	A	A	A	A	J	J
P8	A	A	A	A	A	A	A	A	A	A	A

A：米語母音を使用，J：日本語母音で置換

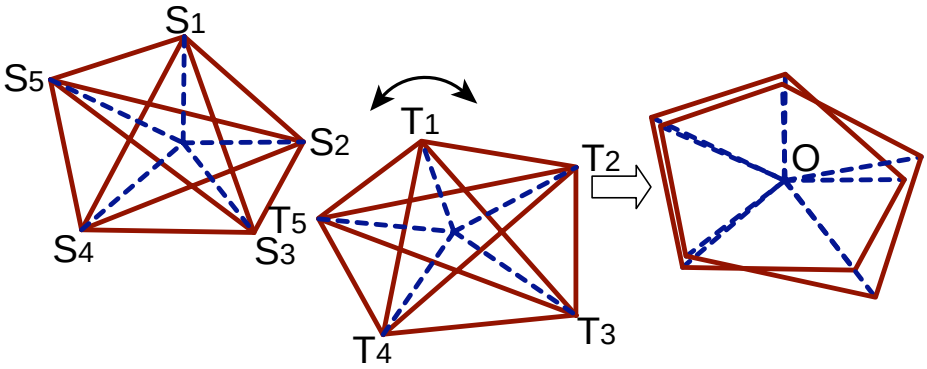


図 5.2: 回転とシフトによる二構造の重ね合わせ (再掲)

の発音状態に対して，それぞれ全ての音素分布間のバタチャリヤ距離を算出し，合計 96 種の学習者発音構造 (母音間距離行列) を得た．

構造間距離は**図 5.2** に示すように，二発声の非言語的特徴による歪みをキャンセルさせたとき，つまり一方の構造を回転及びシフトして他方にできるだけ近づけたときに観測される，対応する二点間距離の総和の最小値として規定できる[3]．この最小値は，二つの距離行列間のユークリッド距離として近似できることが示されており，この距離尺度を使用する [30]．構造 ($m \times m$ 距離行列)S，T の構造間距離 $D(S,T)$ は以下の式で表される．

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

ここで，m は母音数の 11 であり，S_{ij}，T_{ij} は構造 S，T の (i,j) 成分である．

[3]ケプストラムベクトル $\boldsymbol{c}$ に対する行列 A の乗算 $\boldsymbol{c}' = A\boldsymbol{c}$ はおよそケプストラム空間における回転として，$\boldsymbol{c}$ に対するベクトル $\boldsymbol{b}$ の加算はシフトとして解釈される．

5.2.3　模擬英語学習者の発音構造の分類

学習者間距離を，学習者の発音構造同士の構造間距離と定義することにより，96×96 の学習者間距離行列を算出することができる．この学習者間距離行列へ Ward 法によるボトムアップクラスタリングを適用することにより，96 人の学習者を分類することが可能である．

Ward 法によるボトムアップクラスタリングにより分類した結果を**図 5.3** に示す．樹形図のリーフノードにおける数字 (1∼8) が発音状態を意味し，A∼L が話者を意味する．樹形図を見ると，凡そ同じ発音状態が固まる形で 8 つのクラスタが形成されている．この結果より，構造的表象を用いることで，非言語的特徴に影響されることなく，発音の分類，即ち，言語的な分類が可能となることが実験的に示された．

一方，学習者間距離を (v_i^S を話者 S の母音 i の分布とする)

$$D'(S,T) = \sqrt{\frac{1}{m}\sum_i BhattacharyaDistance(v_i^S, v_i^T)} \tag{5.2}$$

のように，異なる話者間で母音を直接的に比較する形で定義すると，**図 5.4** に示すように完全なる話者分類となった [31]．これは**図 5.3** の言語的分類に対し，音響的分類であるといえる．言い換えれば，絶対的物理量の「差」を見ることで話者分類 (音響的分類) となり，「差の差」を見ることで発音分類 (言語的分類) となった訳である．

表 5.6: 音響分析条件（**5.2 節**）

サンプリング	16bit / 16kHz
窓	窓長 25ms，シフト長 1ms
パラメータ	FFT ケプストラム (1∼10 次元)
HMM	1 混合 monophones（全角分散行列）
トポロジー	3 状態 1 ガウス分布

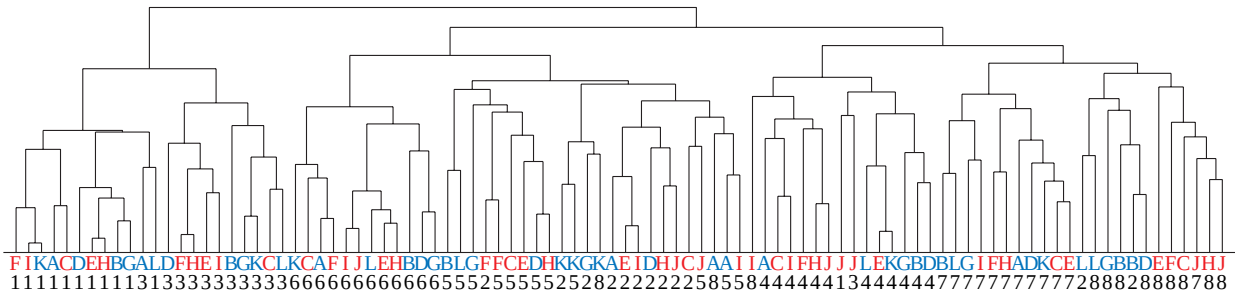


図 5.3: 構造的表象に基づく 96 発音構造 ([12 話者 A∼L] × [8 発音状態 1∼8]) の分類

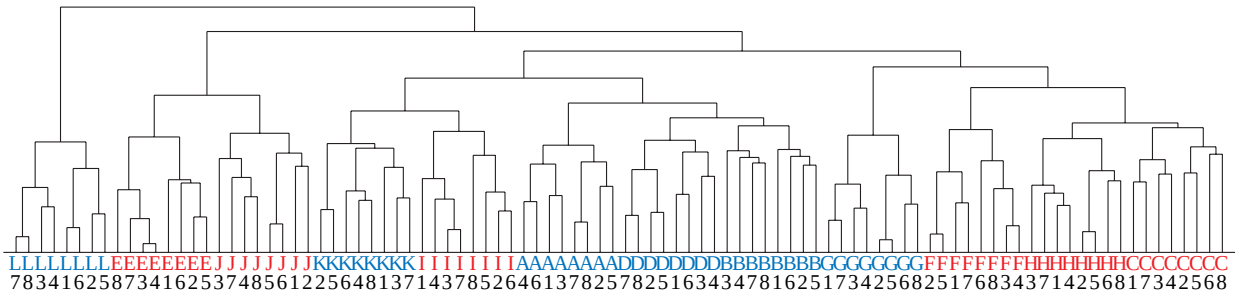


図 5.4: 母音実体間の差に基づく 96 発音構造 ([12 話者 A∼L] × [8 発音状態 1∼8]) の分類

5.3　音声の構造的表象に基づく英語発音の母音矯正度推定

5.3.1　日本人英語学習者の模擬音声の収録

使用した模擬音声は，**5.2.1 節**と同様の 96 種類の 11 発音である．本節では，**表 5.5**（母音置換表）の P1〜P7 はそれぞれ異なる日本人米語学習者の発音状態に相当し，P8 は米語教師の発音状態に相当している．

5.3.2　母音構造間の要素差異に基づく母音矯正度の推定

5.2.2 節と同様の手順で構造抽出をおこない，96 個の発音構造を得る．

異なる二つの構造を比較し，構造を構成する各要素 (構造の頂点) の構造歪みを求める．各要素が持つ構造歪みは，構造を歪ませることへの各要素の寄与度を表し，即ち，二つの構造を一致させるために必要な修正の労力を表す指標である．米語教師の発音構造と日本人米語学習者の発音構造を比較した場合，各要素の構造歪みは当該母音に関する母音矯正度と捉えることができる．

二つの構造 S と T の，要素 v の構造歪みを次式で定義する．

$$d(S,T,v)=\sum_{j=1}^{m}|s_{vj}-t_{vj}| \tag{5.3}$$

m は構造を構成する要素の数 (本節では 11) で，s_{ij} と t_{ij} は構造 (距離行列)S，T それぞれの 2 要素 i，j の要素間距離 (バタチャリヤ距離) を表す．

5.3.3　構造的表象に基づく母音矯正度

式（5.3）を **5.2.2 節**の模擬音声の発音構造に適用した結果が，**図 5.5** である．図中の P1〜P8 は，それぞれ話者 A〜L12 名の平均構造である．各図は米語教師相当の構造 P8 と日本人学習者状態の P1〜P7 を比較したときの各母音の構造歪みを表している．教師と学習者の比較であるので，各母音の構造歪みは各母音に対する母音矯正度と捉えることができる．図の縦軸は各母音の構造歪み (母音矯正度) を表し，色づけられているものは日本語母音に置換した母音 (カタカナ英語発音)，白抜きのものは米語発音の母音である．

半分の母音が置換された発音状態である P2〜P7 について見てみると，日本語母音の多くが，米語母音よりも高い母音矯正度を示している．また，すべての母音が置換された P1 に関しても，日本人が発音を苦手とする/ɔ˞/，/æ/，/ə/，/ɪ/が高い結果となった．このことより，構造的表象に基づいて各母音の母音矯正度を導出可能であることが実験的に示された．

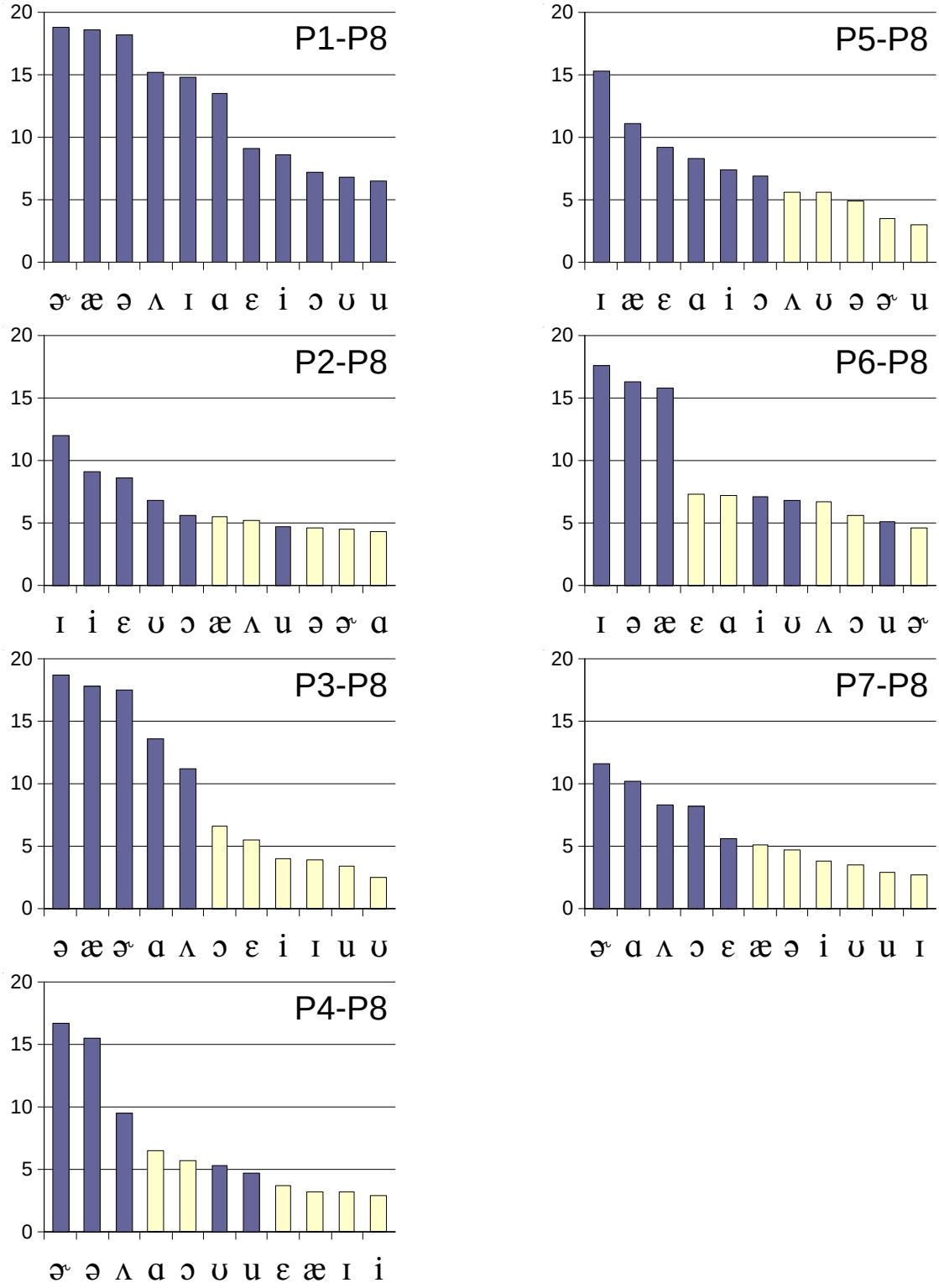


図 5.5: 自動推定された母音矯正順序

第6章

動的計画法による話者性の異なる二発話の時間アライメント

一般に，二発声間の時間アライメントは DP(Dynamic Programing; 動的計画法) マッチングによって求められることが多い．例えば GMM による話者変換では，変換元と変換先の二話者間でのパラレルデータが必要になるが，この場合，同一文セットの読み上げ音声に対して DP マッチングにより二話者間のアライメントが求められる．しかし，二話者の性別が異なるなど個人差が大きい場合，アライメントがずれ，変換音声の品質が劣化することになる [33].

6.1　実験に用いる音声の収録

日本人男性 1 名 (A) 女性 1 名 (B) の，**表 6.1** に示す「こんにちは」(/koNnichwa/) および同単語における音素の置換・挿入・脱落誤りを想定した無意味語，計 12 単語を各 3 発声ずつ収録した．音素の置換誤りは子音・単母音のそれぞれが置換された場合を，挿入誤りは母音が二重母音化し母音が挿入された場合を，脱落誤りは子音が脱落した場合を想定している．

2 単語の特徴ベクトルの系列を

$$A = \boldsymbol{X}(1), \boldsymbol{X}(2), ..., \boldsymbol{X}(I) \tag{6.1}$$

$$B = \boldsymbol{Y}(1), \boldsymbol{Y}(2), ..., \boldsymbol{Y}(J) \tag{6.2}$$

A と B のパターンの時間軸の対応付け (時間伸縮関数) を

$$F = c_1, c_2, ..., c_k, ..., c_K \tag{6.3}$$

$$c_k = (i_k, j_k) \tag{6.4}$$

とする．ここで

$$\min_{i,j} \sum_k ||\boldsymbol{X}(i_k) - \boldsymbol{Y}(j_k)||^2 \tag{6.5}$$

を基準とする，物理量を直接参照する一般的な DP マッチングをおこなった．分析条件を**表 6.2** に示す.

一例として，男性 A と女性 B の「こんにちわ」どうしの DP パスの様子を**図 6.1** の青線で示す．概ね斜め 45 度にパスが進むのは，話者性の影響が比較的少ない子音の部分で，逆に話者性に強く影響される共鳴音の部分は大きく歪み，直上，直右に進むパスが多く見られる．

6.2　絶対的物理量に基づく DP マッチングによるアライメント

DP マッチングによる時間アライメントのずれがどの程度なのかを測るため，収録した音声全てに対して付与した手動の音素アライメント[1]との比較をおこなった．具体的には，

[1]この手動アライメントは，単語 2~12 においても「こんにちは」と強制アライメントしたものである．そのため，音素の脱落が起こっている単語 2~4 においては，脱落した音素はその前後の音素境界の間の 1 フ

表 6.1: 収録に使用した日本語単語

Index	日本語単語	音素列
1	こんにちは	/koNnichiwa/
2	こにちは	/ko_nichiwa/
3	おんにちは	/_oNnichiwa/
4	こんにちあ	/koNnichi_a/
5	けんにちは	/keNnichiwa/
6	こんにちうぉ	/koNnichiwo/
7	そんにちは	/soNnichiwa/
8	こんみちは	/koNmichiwa/
9	こんにしは	/koNnisiwa/
10	こんないちは	/koNnaichiwa/
11	こうんにちは	/kouNnichiwa/
12	こんにちわい	/koNnichiwai/

1. 教師発声の音素境界に対応する教師発声のフレームを求める
2. DP マッチングの結果より，教師発声のフレームに対応する生徒発声のフレームを求める
3. 生徒発声のフレームと，生徒発声の音素境界に対応するフレームとの差を求める

という手順を追った．教師と生徒は別の話者とするため，教師:A・生徒:B の場合と教師:B・生徒:A の場合とし，それぞれ，教師音声 3 通り × 生徒単語 12 種 × 単語 3 通り × 音素境界 8ヶ所 =864ヶ所の差を求める．横軸に差の許容フレーム数を，縦軸に許容フレーム数以内に収まった音素境界の正解率を示したものが**図 6.2**(a)(b) である．参考のために，教師と生徒が同じ話者の場合の正解率 (教師音声 3 通り × 生徒単語 11 種 × 単語 3 通り × 音素境界 8ヶ所 =756ヶ所中の正解率) を**図 6.2**(c)(d) に示す．

日本語の場合，標準的な発声における 1 秒間あたりの音素数は 10〜15 であり，一音素当たりの平均持続時間は約 80msec 程度であると考えられる．ここではアライメントのずれを 20msec まで許容することにする[2]．話者が同じ場合，約 7 割の正解率となっている一方，話者が異なる場合は正解率が下落しそれぞれ約 35% ，約 55% となっている．絶対的な物理量を参照する一般的なの DP マッチングでは，話者の違いにより不一致問題が生じていることがわかる．

レームに対応する．置換が起きている単語 5〜9 では，置換前の音素が占めるフレームは，置換後の音素が占めるフレームに対応する．挿入が起き二重母音化している単語 10〜12 では，置換前の単母音が占めるフレームは置換後の二重母音の占めるフレームに対応する．

[2]一般的な音声認識の場合，特徴量抽出をおこなう際のフレームシフト長は 10msec であることが多く，今回の 20msec のずれはその 2 フレームに相当することになる．

表 6.2: 音響分析条件（**6.2 節**）

sampling	16bit / 16kHz
window	Blackman / 512 sample 25msec length / 1msec shift
特徴量	MCEP($C_{0\sim12}$, $\Delta C_{0\sim12}$)

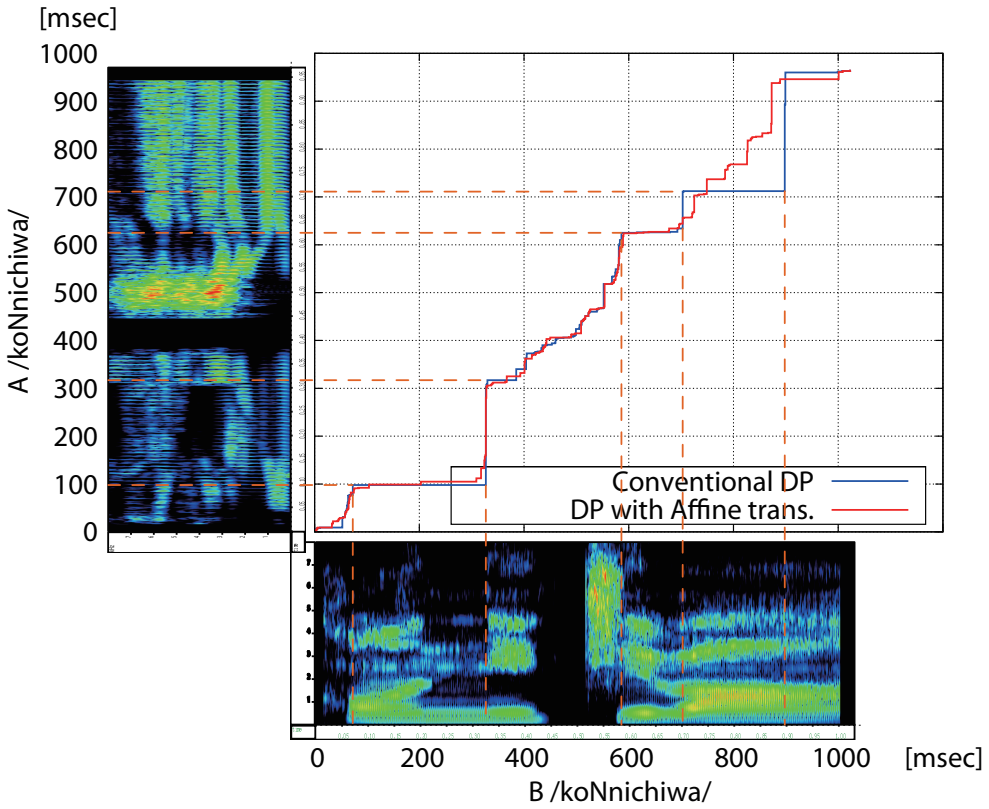


図 6.1: 男性話者 A と女性話者 B の「こんにちわ」発声における DP パス

6.3　話者変換を含む DP マッチングによるアライメント

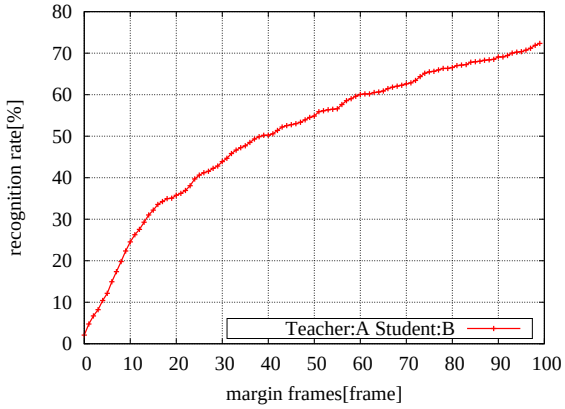
6.2 節と同様の条件において，**式（6.5）**の代わりに

$$\min_{i,j} \sum_{k} ||(\boldsymbol{W}\boldsymbol{X}(i_k) + \boldsymbol{b}) - \boldsymbol{Y}(j_k)||^2 \tag{6.6}$$

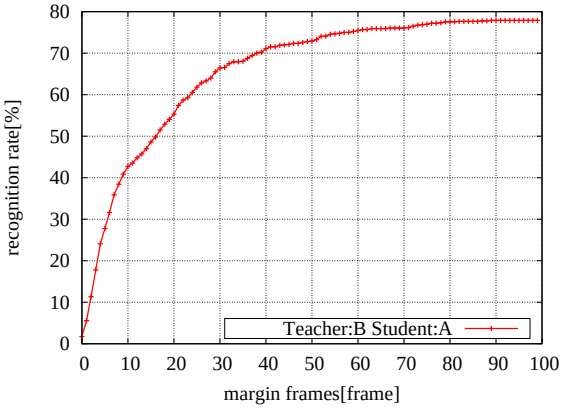
$$\boldsymbol{Y} = \boldsymbol{W}\boldsymbol{X} + \boldsymbol{b} \tag{6.7}$$

を基準とし，両発声から大局的な変換行列を GMM より構成し，アフィン変換で表わされるその話者変換をかけてから DP マッチングをおこなった．話者変換と DP マッチングを繰り返して行い，教師の発声を生徒に近付けていった．分析条件を**表 6.3** に示す．

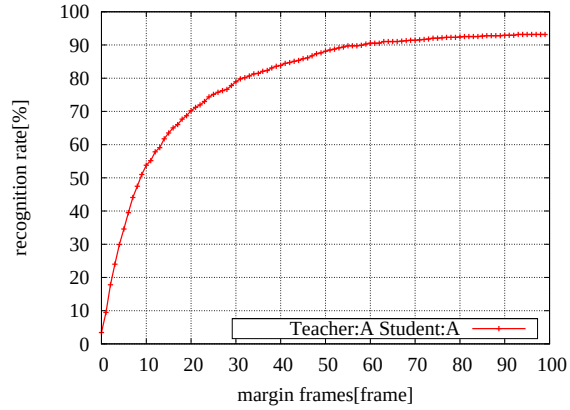
一例として，男性 A と女性 B の「こんにちわ」どうしの DP パスの様子を**図 6.1** の赤線で示す．話者変換が掛ることで，通常の DP パスでは大きく歪み直上，直右に進んだパス



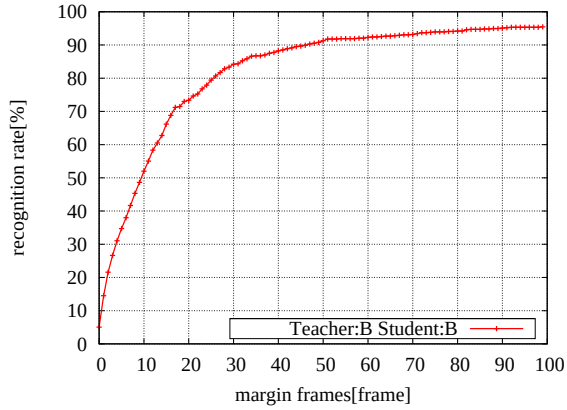
(a) 教師:A，生徒:B の場合



(b) 教師:B，生徒:A の場合



(c) 教師:A，生徒:A の場合



(d) 教師:B，生徒:B の場合

図 6.2: 物理量を直接参照する DP マッチングでの認識率

も一部斜め 45 度に進むようにパスが改善されているが，全てが改善されているわけでない．アフィン変換を考慮しても全ての対応が付かないのは，話者性が弱くアフィン変換をあまり必要としない子音部と話者性が強くアフィン変換が必要な共鳴音部との両方が存在するため，同じ変換行列をかけ合わせることで一度に両者を最適化することができないためだと考えられる．

また，**6.2 節**と同じ手順に従い，手動アライメントとの差を求めた．許容フレーム数に対する音素境界の正解率を示したものが**図 6.3** である．(a)(b) が話者性が異なる場合，(c)(d) が同じ話者の場合である．(a)(b) では，青線の通常 DP に対し赤線の話者変換をかけた DP では最大約 10% の正答率向上が見られ，話者変換により二発声の対応が改善されたことがわかる．それに対し，(c)(d) では話者変換の有無で正答率の変動はほとんどない．置換誤りは，その置換された音素のみに関して別の話者が発声したことと等価である．つまり「こんにちわ」の 9 音素のうち一つだけ話者性が異なることとなり，全体に対する話者の違

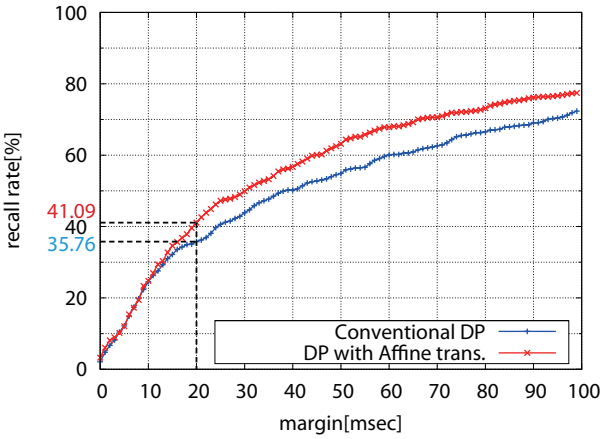
表 6.3: 音響分析条件（**6.3 節**）

sampling	16bit / 16kHz
window	Blackman / 512 sample 25msec length / 1msec shift
特徴量	MCEP($C_{0\sim12}$, $\Delta C_{0\sim12}$)
繰り返し回数	10 回

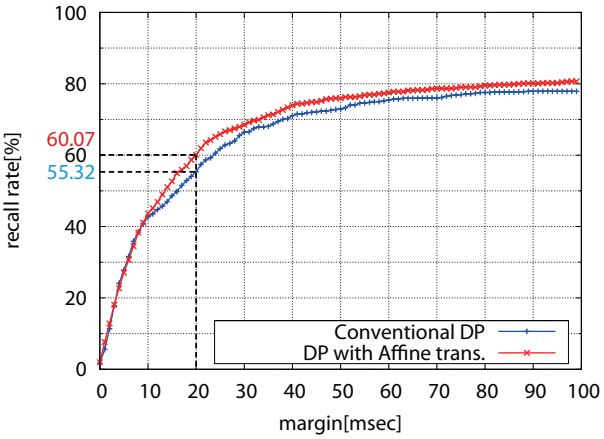
いの寄与が小さいため，ほとんど変動がないものと考えられる．これは，今回の場合「単母音が二重母音に置換されてしまった」と考えることができる挿入誤りにおいても同じことが言える．

6.4　本章のまとめ

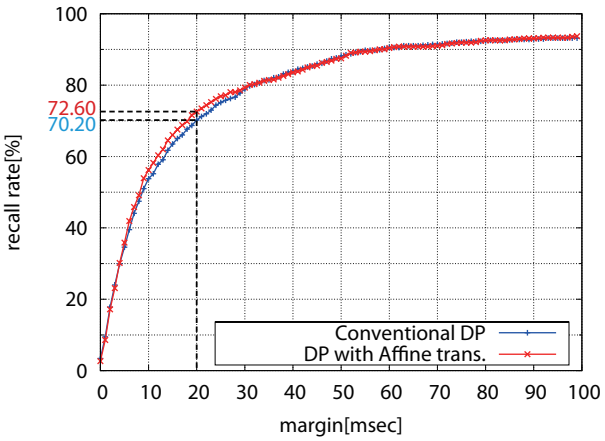
本章では，二発声間の時間アライメントを求めるため，動的計画法を用いる手法を検証した．一つは，音響特徴量そのものを参照しペナルティを計算する，最も単純な DP マッチングをおこなった．この場合，話者性の違いによって時間アライメントの精度が大きく低下することを確認した．もう一つは，二発声の話者性の違いをアフィン変換で表わされる話者変換によってキャンセルし DP マッチングをおこなった．上記の一般的な DP マッチングの結果に較べ，若干アライメントの精度が向上することを確認した．



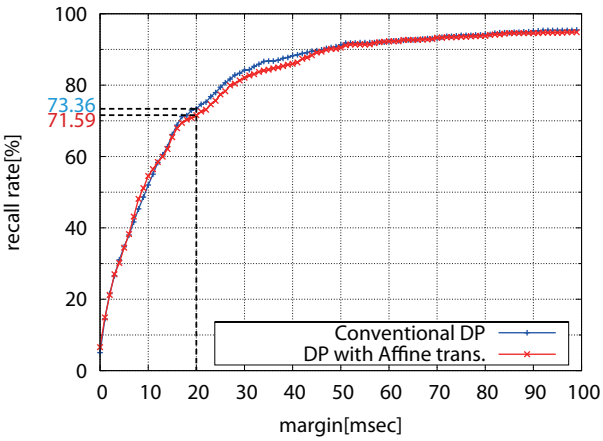
(a) 教師:A，生徒:B の場合



(b) 教師:B，生徒:A の場合



(c) 教師:A，生徒:A の場合



(d) 教師:B，生徒:B の場合

図 6.3: 話者変換を含む DP マッチングでの認識率

第7章

制約付きHMM学習による話者性の異なる二発話の時間アライメント

7.1　はじめに

CALL の場合，学習者が意図して発話した単語の種類は既知とすることが可能であり，不特定話者音響モデルを利用し強制アライメントを行い，音素単位で時間同期をとることができる．しかし，話者の年齢が児童から成人までを学習対象とするような場合では，あらゆる年代の話者の音響モデルを準備する必要があり，モデルと話者のミスマッチによるアライメント精度の劣化も十分起こりうる．

本章では，大規模話者から求められた不特定話者音響モデルを使用せず，一発声から構成した構造を教師・学習者間で比較することを目的として，時間アライメント (及び構造) の導出を考える．**1.2 節**で述べたように，各発声から独立に単語 HMM を学習し，これらから構造を求めると，同期のずれが頻繁に起こる．以下では，HMM 学習に対して各種の制約を導入することで，同期のとれた HMM の学習を試みる．

7.2　特定話者音素 HMM 学習によるアライメント

教師と学習者でそれぞれ一発声から単語 HMM を学習して構造を抽出すると同期のずれが生じるため，制約を設ける．一つは，発声回数を増やすことによるデータ量の倍加である．教師・学習者ともに 3 回ずつ単語を発声させ，データ量を倍加する．もう一つは，複数回出現する同一種類の音素のタイイングである．音素の結びの関係を考慮することで，推定するパラメータを減少させる (パラメータ当たりの学習データ量を増加させる)．

まず，教師・学習者ともに 3 回ずつ単語を発声させる[1]ことを考える．また音素のタイイングをし推定パラメータ数を減らして，連結学習を通し各話者毎に音素 HMM を構築する．最後に，得られた HMM を用いて各話者の発声の強制アライメントをとる．分析には，**6.1 節**で収録した音声を用いた．分析条件は**表 7.1** である．

一例に，得られた男性話者 A と女性話者 B のアライメント結果を**図 7.1** に示す (各話者，上段がスペクトログラム，中段が手動音素アライメント，下段がアライメント結果である．)．各発話を見ると，凡そ手動アライメントの境界付近に，いずれかの二状態の境界 (音素境界とは限らない) が存在しているものの，**図 7.1** で言えば状態番号 b3 や c3 のように二発声間では同期がとれていない．

特定話者音響モデルによる強制アライメントのずれがどの程度なのかを測るため，手動アライメントとの比較をおこなった．手順は以下に従う．

1. 教師発声の手動音素境界に最も近い，教師発声の強制アライメントによる状態境界を求める
2. 教師発声の状態境界と同じ生徒発声の状態境界と，生徒発声の手動音素境界との差を求める

ここでは教師と生徒は別の話者，つまり話者性の異なる 2 話者における発音のアライメントのみを考える．教師発話の手動音素境界とそれに最も近い状態境界との時間差と，生徒

[1]CALL 応用では，単語やフレーズを 3 回発声させることは，実用的には大きな問題にならない．

表 7.1: 音響分析条件（**7.2 節**）

sampling	16bit / 16kHz
window	Blackman / 512 sample 25msec length / 1msec shift
特徴量	MCEP($C_{0\sim12}$, $\Delta C_{0\sim12}$)
HMMs	特定話者 monophoneHMM
出力確率分布	対角共分散行列 / 単一ガウス分布
トポロジー	3 状態 left to right 型
音素	k, o, N, n, i, ch, w, a 計 8 種類

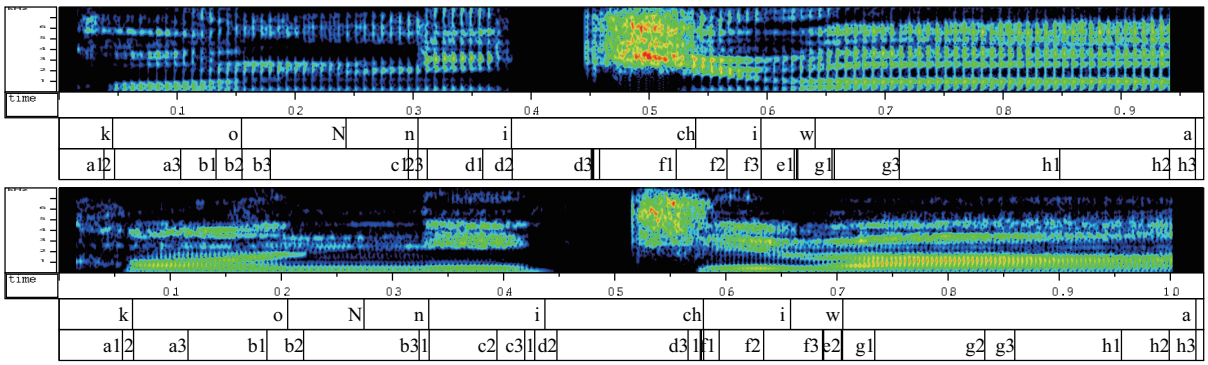
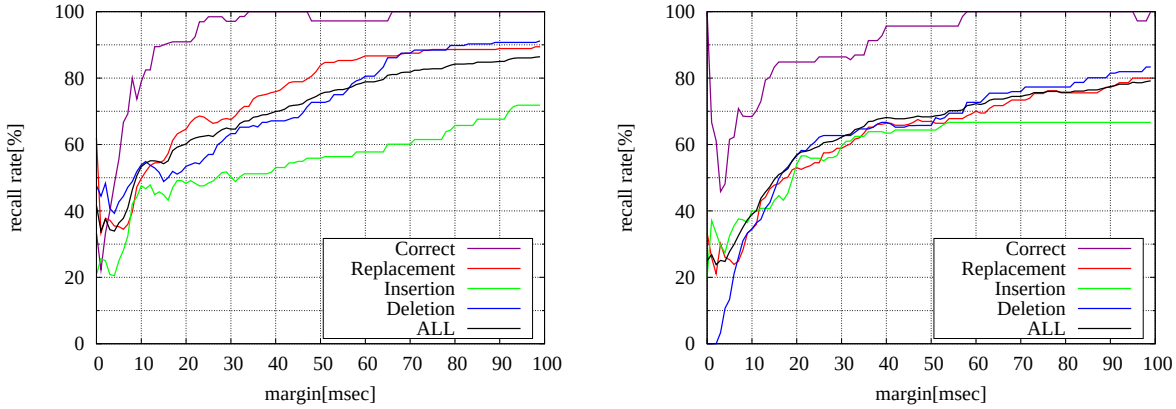


図 7.1: 3 発声で連結学習した特定話者 HMM によるアライメント結果 (**7.2 節**，上:男性 A，下:女性 B)

発話の手動音素境界と状態境界との時間差の両方を考慮する必要がある．そこで，教師発話での時間差が許容フレーム数以内のもののみを対象とし，横軸に差の許容フレーム数を，縦軸に生徒音声において許容数以内に収まった状態境界の正解率を示したものが**図 7.2** である．

手動で求めた音素境界と，強制アライメントによって求められた状態の境界のうち手動音素境界に最も近いものとの差が 20msec 以下のものは，男性話者 A では 88.89% ，女性話者 B では 83.68% であった．この時の正解率を**表 7.2** に示す．**??節**に比べ，全体的な精度は向上しているが，挿入・削除誤りにおいては不安定であるといえる．3 発声に学習データを増やし，タイイングによるパラメータ共有も考慮したが，十分な精度のアライメントではない．以降，二発声間で独立に HMM 学習するのではなく，二発声を網羅する一つの共通 HMM の学習を考える．



(a) 教師:A，生徒:B の場合　　(b) 教師:B，生徒:A の場合

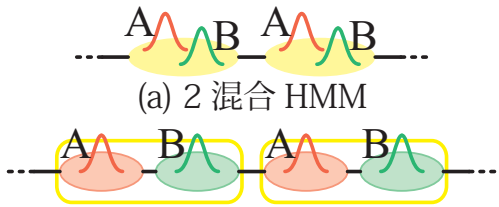
図 7.2: 特定話者音響モデルによる強制アライメントでの正解率 (**7.2 節**)

表 7.2: 20msec まで許容した際の正解率 [%](**7.2 節**)

	教師:A，生徒:B	教師:B，生徒:A
正解	90.91	84.85
置換	64.63	53.02
挿入	48.28	54.04
削除	53.44	56.61
平均	60.17	56.90

7.3　2 混合 HMM 学習によるアライメント

話者性が大きく異なる可能性があることを考えれば，**図 7.3**(a) のように，HMM の各状態の分布数を 2 として共通 HMM を学習することが望ましい．そこで **7.2 節**で得られた特定話者音素 HMM の平均ベクトル，分散を共通 HMM の各状態における二つのガウス分布として採用し，これを 2 混合の初期モデルとした．その後二話者の 6 発声を使って連結学習にて再推定を行った．また，**7.2 節**同様，結びによるパラメータ共有も行っている．分析条件は**表 7.3** である．



(a) 2 混合 HMM

(b) 2 状態を線状結合した HMM

図 7.3: 2 混合 HMM と 2 状態線状結合 HMM

表 7.3: 音響分析条件（**7.3 節**）

sampling	16bit / 16kHz
window	Blackman / 512 sample 25msec length / 1msec shift
特徴量	MCEP($C_{0\sim12}$, $\Delta C_{0\sim12}$)
HMMs	特定話者 monophoneHMM
出力確率分布	対角共分散行列 / 2 混合ガウス分布
トポロジー	3 状態 left to right 型
音素	k, o, N, n, i, ch, w, a 計 8 種類

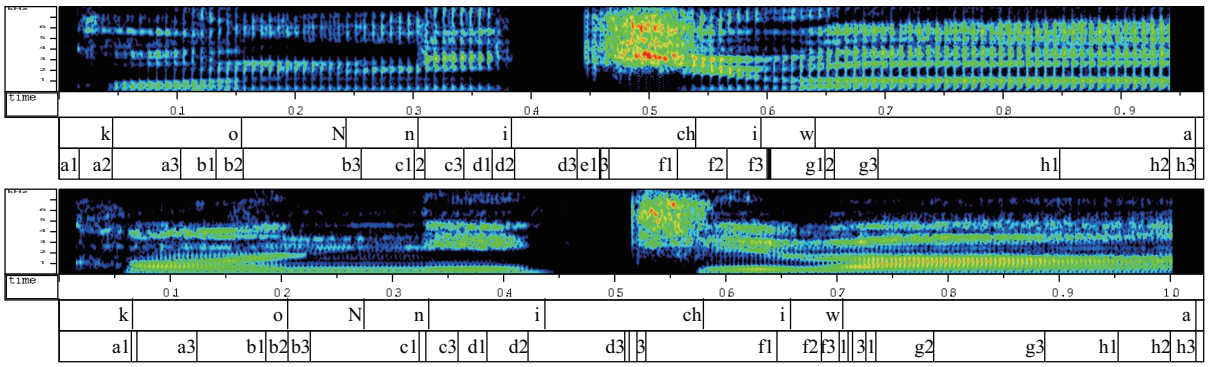


図 7.4: 二話者の 6 発声で連結学習した 2 混合の音響モデルによる「こんにちは」のアライメント結果 (**7.3 節**，上:男性 A，下:女性 B)

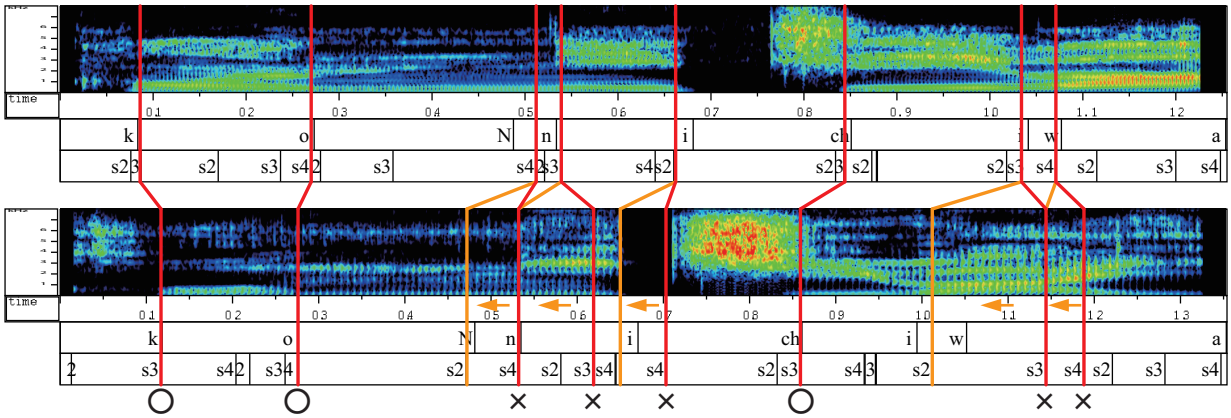


図 7.5: 二話者の 6 発声で連結学習した 2 混合の音響モデルによるアライメント結果 (**7.3 節**，上:男性 A「こんにちは」，下:女性 B「こんにちわぃ」)

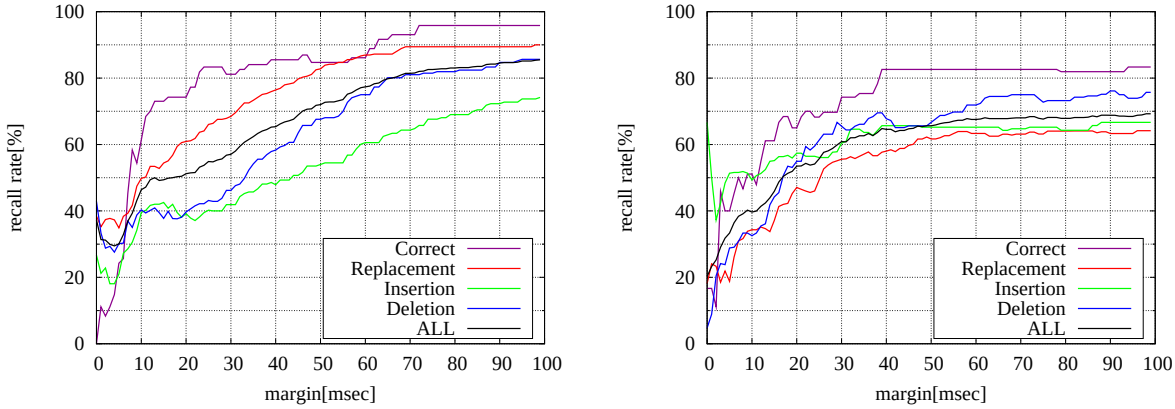
表 7.4: 20msec まで許容した際の正解率 [%](**7.3 節**)

	教師:A，生徒:B	教師:B，生徒:A
正解	74.24	65.00
置換	60.93	47.06
挿入	39.18	57.38
削除	39.68	54.94
平均	51.21	53.48

一例に，得られた男性話者 A と女性話者 B の「こんにちは」におけるアライメント結果を**図 7.4** に示す (各話者，上段がスペクトログラム，中段が手動音素アライメント，下段がアライメント結果である．)．**7.2 節**同様，各発話凡そ手動アライメントの境界付近にいずれかの二状態の境界がある．音素/o/と/N/との境界が状態 b2 と b3 の境界に対応したり音素/i/と/ch/の境界が状態 d2 と d3 の境界にほぼ合致するなど，手動アライメントの音素境界と得られた状態の境界は，第 7.2 節よりも改善している．一方で，**図 7.5** のように音素の挿入が行われた場合は，挿入された音素の分だけケプストラムベクトルの変化が増えるため，その部分に数状態が割り当てられている．このため，より変化の少ない部分が一つの状態へとまとまり，結果的に状態どうしの対応がずれることとなる．

7.2 節と同じように，二話者の 6 発声から得られた 2 混合の音響モデルによる強制アライメントのずれがどの程度なのかを測るため，手動アライメントとの比較をおこなった．**7.2 節**同様，教師発話の手動音素境界とそれに最も近い状態境界との時間差と，生徒発話の手動音素境界と状態境界との時間差の両方を考慮する必要があるため，教師発話での時間差が許容フレーム数以内のもののみを対象とした．横軸に差の許容フレーム数を，縦軸に生徒音声において許容数以内に収まった状態境界の正解率を示したものが**図 7.6** である．

手動で求めた音素境界と，強制アライメントによって求められた状態の境界のうち手動



(a) 教師:A，生徒:B の場合　　(b) 教師:B，生徒:A の場合

図 7.6: 二話者の 6 発声で連結学習した 2 混合音響モデルによる強制アライメントでの正解率 (**7.3 節**)

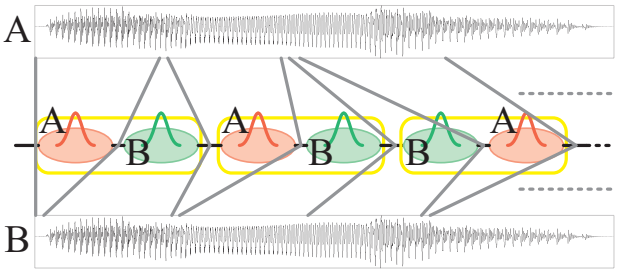


図 7.7: 線状結合 HMM を利用した二発話の共鳴音部における状態アライメント

音素境界に最も近いものとの差が 20msec 以下のものは，男性話者 A では 81.60% ，女性話者 B では 76.39% であった．この時の正解率を**表 7.4** に示す．**7.2 節**に較べ，挿入・削除誤りの正解率が極端に落ち込んでいるのがわかる．

7.4　状態を線状結合した HMM によるアライメント

本節では，挿入・削除誤りの存在を考慮して，2 話者を，線状結合された 2 状態で表現する（一方が教師，他方が学習者）ことで共通 HMM を学習する．つまり，HMM の状態数を時間方向に 2 倍し，連続する 2 状態が 2 話者に割り振られる共通 HMM を学習する (**図 7.3**(b) 参照)．**7.2 節**の HMM における 1 状態が，本節で構築する HMM の（連続する）2 状態に対応する．この共通 HMM を使ってアライメントをとる場合，当該話者ではない状態との照合が常に起きる．このような状態には（アライメント結果を見れば分かるように）通常 1 フレームが割り当てられることになる (**図 7.8** 参照)．これは 1 フレームが「挿入」されたことに相当する．

表 7.5: 音響分析条件（**7.4 節**）

sampling	16bit / 16kHz
window	Blackman / 512 sample 25msec length / 1msec shift
特徴量	MCEP($C_{0\sim12}$, $\Delta C_{0\sim12}$)
HMMs	特定話者 monophoneHMM
出力確率分布	対角共分散行列 / 単一ガウス分布
トポロジー	54 状態 left to right 型
単語	/こんにちは/ 1 種類

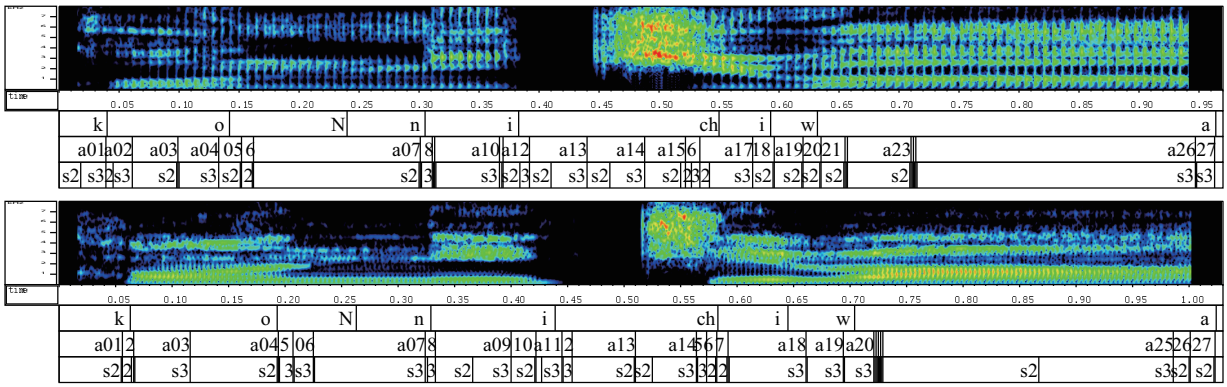


図 7.8: 二話者を線状結合した音響モデルによる「こんにちは」のアライメント結果 (**7.4 節**，上:男性 A，下:女性 B)

学習者の発話に挿入誤りが起こった際を仮定する[2]．学習者の発話には教師の発話よりも音響イベントが多く含まれることになり，音響イベント数が等しいときに 2 話者に割り振られる連続する 2 状態が，片方は学習者のみに，もう一方は学習者と教師の発話に割り振られることになる**図 7.7**．即ち，教師のみに割り当てられていた状態に学習者の挿入された音響イベントの影響も付加される．そのため，挿入・脱落誤りがあっても線状結合した状態単位では対応したアライメントとなると考えられる．

表 7.5 の分析条件で構築した共通 HMM を用いてアライメントをおこなった．なお，**7.2 節**で得られた特定話者音素 HMM に対して，状態単位で 2 話者間で線状に連結し，共通 HMM の初期モデルとした（「こんにちは」は音素数が 9 であるため，3 状態 ×9 音素 ×2=54 状態の HMM となる）．

一例に，得られた男性話者 A と女性話者 B の「こんにちは」におけるアライメント結果を**図 7.8** に示す (各話者，上段からスペクトログラム，手動アライメント，連続 2 状態を線状結合した状態単位のアライメント結果，状態単位のアライメント結果である)．発声ごとに見ると，**7.2 節**同様，手動アライメントの境界付近に線状結合した状態の境界が存在し

[2]音素の脱落は，もう一方の話者において挿入が起こったことと等価である．

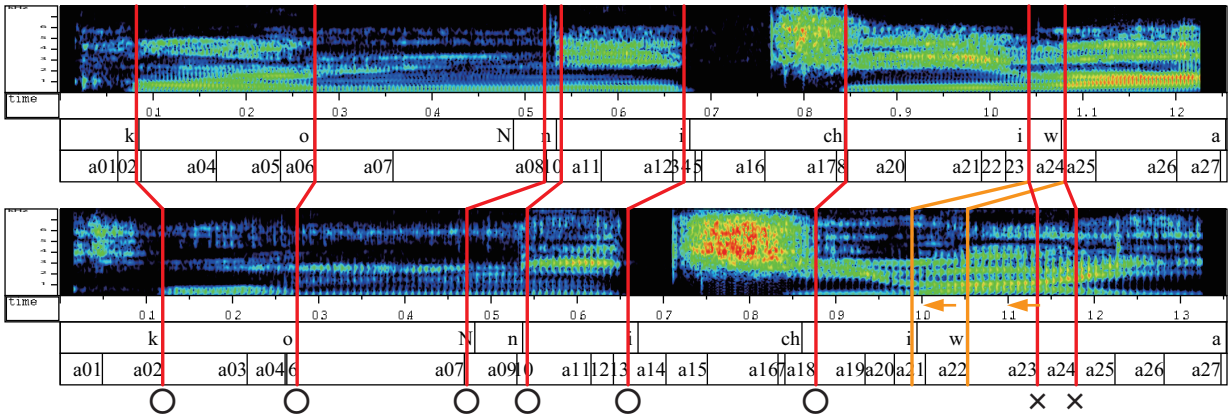
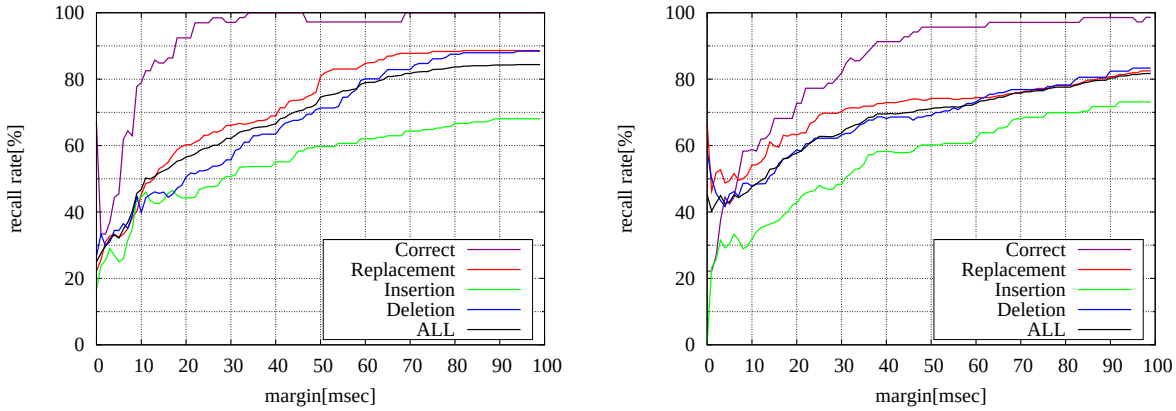


図 7.9: 二話者を線状結合した音響モデルによるアライメント結果 (**7.4 節**，上:男性 A「こんにちは」，下:女性 B「こんにちわぃ」)



(a) 教師:A，生徒:B の場合　　(b) 教師:B，生徒:A の場合

図 7.10: 二話者を線状結合した音響モデルによる強制アライメントでの正解率

ている．アライメントに注目すると，共鳴音においては，話者ごとに線状結合した 2 状態のうちのいずれかが多数のフレームに対応し，もう一方は 1 フレームのみに対応するようアライメントされている．対して，子音では二話者で線状結合した連続 2 状態を共有するようにアライメントされている．これは子音が非言語性歪みの影響が少ないため話者間でのばらつきが少なく，対して共鳴音は話者の違いに大きく影響を受けるため異なる物理表象を表していることに合致する．この実験から，話者性が違っていても線形結合した状態単位で凡そアライメントがとれることがわかる．

また，音素が挿入されたときの一例を**図 7.9** に示す．音素の挿入による状態のずれは起こっているものの，**図 7.5** の場合と異なり全体にその影響が波及してはいない．各話者の連続二状態が一つずつ 2 人の話者に割り当てられるため，2 混合のものに較べ学習時に大きなずれが生まれにくい．

表 7.6: 20msec まで許容した際の正解率 [%](**7.4 節**)

	教師:A，生徒:B	教師:B，生徒:A
正解	92.42	72.73
置換	60.24	63.17
挿入	44.27	42.71
削除	50.51	58.73
平均	56.58	57.74

7.2 節と同じように，二話者の 6 発声から得られた 2 混合の音響モデルによる強制アライメントのずれがどの程度なのかを測るため，手動アライメントとの比較をおこなった．教師と生徒は別の話者であるとし，横軸に差の許容フレーム数を，縦軸に許容数以内に収まった状態境界の正解率を示したものが**図 7.10** である．前節同様，教師発話での時間差が許容フレーム数以内のもののみを対象としている．手動で求めた音素境界と，強制アライメントによって求められた状態の境界のうち手動音素境界に最も近いものとの差が 20msec 以下のものは，男性話者 A では 90.62% ，女性話者 B では 88.19% であった．これは **7.2 節**, **7.3 節**に較べ最も良く，聴覚的アライメントとの親和性が高いことを示している．この時の正解率を**表 7.6** に示す．**表 7.4** に較べ，全体的に認識率の向上が見られた．

二話者の各状態を互い違いに線状結合し，他の話者の状態を強制的にアライメントすることで，挿入や削除誤りが起きた際にも増えた音響イベントを吸収し，2 混合音響モデルに較べて精度の高いアライメントが行えることがわかった．従来，二発話間のアライメントは話者の違いにより歪められてしまっていた (**6.2 節**参照) が，この手法では逆に話者性が異なることで，対応するイベントをより精度よくアライメントすることができると言える．

7.5　本章のまとめ

本章では，HMM 学習を用いて二発声間の時間アライメントを求めるため，各種の制約を検証した．一つは，発声回数の倍加によりデータ量を増やし，同一音素のタイイングを導入した学習をおこなって得られた，特定話者音響モデルによるアライメントをおこなった．この場合，**??節**に較べれば精度のよいアライメントが得られたが，挿入・削除誤りが起きた発声との対応付けではアライメントの精度が不安定であった．次に，HMM の各状態の分布数を 2 として二話者に共通の HMM を，両者の発声で連結学習し得られた 2 混合の音響モデルによるアライメントをおこなった．2 混合化したことで置換誤りに対しては特定話者音響モデルと同等の精度が得られたが，挿入・脱落誤りに対してはエラーを吸収することができず精度が低かった．最後に，線状結合した連続する二状態で二話者を表現する共通 HMM によるアライメントをおこなった．各話者は自分自身を表現するノードと他話者を表現するノードに対応するよう強制的にアライメントされるため，他話者に関するものは最小単位にアライメントされることとなる．挿入・脱落誤りが存在した場合は，その他話者を表現しているノードがアライメントされることが多く，結果的に挿入・脱落誤

りの区間は音響モデルに吸収されることとなる．置換誤りに関しては特定話者音響モデルと同等の精度が得られ，また，挿入・脱落誤りでは 2 混合の二話者共通音響モデルよりも精度よくアライメントが行われた．

第8章

結論

8.1　本研究のまとめ

本研究では，音声の構造的表象に基づく発音評価システムのための，話者性の異なる二話者の発声間における時間アライメント手法に関して検討をおこなった．

まず，**第 1 章**において，本研究の目的について述べた．

第 2 章では英語発音評価システムが担う発音教育の重要性について述べ，日本語発音と英語発音の差異について説明した．また，現在実用化されている英語教育システムについて触れ，本研究が目指す英語発音評価システムについて述べた．

第 3 章では，英語発音評価システムを構築する上で障害となる「不一致問題」について述べた．不一致問題を引き起こす場合について実験を行い，従来手法による対応では不十分であることを示した．

第 4 章では，話者の違いに頑健な発音評価システムの核をなす音声の構造的表象について説明した．不可避な要因（発話者や収録機器）によってもたらされる音声の歪みについて整理し，この歪みを除去した特徴量である「音響的不変構造」について説明した．

第 5 章では，音声の構造的表象に基づく英語発音評価システムの各要素について先行研究を紹介した．**5.1 節**では，英語学習者の音声から話者の身体的特徴の差異（年齢や性別など）を取り除き，英語の発音状態のみを評価し，記述可能であることを紹介した．**5.2 節**では，多数の英語学習者を話者性の違いに影響を受けずに，発音状態に基づいて分類できることを紹介した．**5.3 節**では，英語教師と英語学習者の発音を一対一に直接比較し，母音矯正必要度を推定する手法について紹介した．

第 6 章では，二話者の発声を動的計画法を用いた時間アライメントについて検討した．**6.2 節**では，絶対的物理量に基づき DP マッチングをおこなうことで話者性の違いにより時間アライメントが歪むことを示した．これに対し **6.3 節**では，話者性の違いがケプストラムベクトルに対する乗算と加算で表現できることから，話者の違いをなくすよう話者変換を行いながら両者の発声を DP マッチングした．しかし話者性の影響が強い母音部と影響の弱い子音部に対し同一の話者変換を掛けることになるため，発声全体としては最適な変換がかけられず，絶対的物理量に基づく DP マッチングからの精度の向上は必ずしも大きいとは言えなかった．

第 7 章では，大規模話者から求められた不特定話者音響モデルを使用せず，教師・学習者の発声のみから構成した構造を両者の間で比較することを目的として，HMM 学習に制約を設けることで時間アライメント (及び構造) の導出を検討した．**7.2 節**では発声回数を増やすことで学習データを増加させ，また同一音素のタイイングも行うことでパラメータ数を減らして学習をおこなった．**第 6 章**に較べ精度は上がったものの，挿入・削除誤りに対しては不安定であった．**7.3 節**では話者の違いを 2 つの分布の足し合わせで表現し，教師・学習者共通の HMM での連結学習をおこなった．**7.2 節**に較べ，音素境界と最も近い状態の境界はより似通ったが，特に挿入・置換誤りにおいて二発声での状態どうしの対応が芳しくなかった．この各状態における二分布の足し合わせによる共通 HMM の学習に対し，**7.4 節**では二話者を表す分布をそれぞれ互い違いに線状結合した共通 HMM の学習をおこなった．自分自身を表現する分布だけでなく他話者の分布も含まれるため，予期せぬ

音素の挿入や置換を吸収し二混合 HMM に較べ挿入・削除誤りに対する精度が向上した.

8.2　今後の課題

8.2.1　英語発音音声への適用

6.1 節において収録し，検討に用いてきた音声は日本語単語「こんにちは」およびその置換・挿入・脱落誤りをシミュレートした無意味語であった．外国語，特に英語発声においてなされる可能性のあるエラーを，日本語においてシミュレートするため，音素の置換誤りは子音・単母音のそれぞれが置換された場合を，挿入誤りは母音が二重母音化し母音が挿入された場合を，脱落誤りは子音が脱落した場合を今回想定した.

実際の英語の場合，子音が数音素連続することが頻繁に見られるのに対し，日本語ではそのような例は稀であるが，日本人学習者が英語発声を行った場合，それらの子音列の中に母音が含まれるエラーなどが考えられる．今回の日本語によるエラーのシミュレートでは挙げられていないエラーも多々想定されるため，実際の英語発声においてそれらのエラーをシミュレートし時間アライメントの精度を考える必要がある.

8.2.2　発話者の倍加によるデータの増量

今回は男性 1 名と女性 1 名の 2 話者間で時間アライメントについて検討してきたが，話者の違いは性別だけでなく体格の違いなども挙げられる．様々な話者どうしでの比較とアライメントの精度の確認が必要があると考えられる.

特にフレーズ音声においてであるが，ここで注意しなければならないのは発声のスタイルを統一しなければならないことである．収録の際には基準となる話者の発声を聴取してから発声するなどの手間が必要だと考えられる.

謝辞

二年間に渡る本研究を遂行するにあたり，多大なる御指導そして御協力を戴きました，指導教員である峯松信明准教授並びに広瀬啓吉教授に心より感謝いたします．特に，峯松信明准教授の熱意ある的確な指導は，研究を進める上での大きな推進力となりました．深く感謝いたします．また，快適な研究環境の維持に努め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いたします．博士課程の齋藤大輔氏には，ご自身の研究で忙しいにも関わらず，研究方針に関する助言など，本研究を全面に渡りサポートして戴きました．ここに深く感謝いたします．峯松研究室で研究生活を共にした鈴木雅之氏，高橋琢己氏を始めとする広瀬・峯松研究室の皆様に深く感謝いたします．修士課程の二年間が充実した日々となったのも，皆様のおかげであります．重ね重ね御礼申し上げます．

2010 年 2 月 9 日
高澤　真章

参考文献

[1] http://www.mext.go.jp/b_menu/shingi/chousa/shotou/020/sesaku/020702.htm

[2] http://www.mext.go.jp/a_menu/shotou/new-cs/youryou/syo/gai.htm

[3] http://www.rakuraku-inc.com/ihatsuon/

[4] http://www.gakken.jp/dc/ds/english/index.html

[5] A. Neri, C. Cucchiarini, and H. Strik, "Automatic speech recognition for second language learning: How and why it actually works," Proc. Int. Congress of Phonetic Sciences, pp.1157–1160, 2003.

[6] 峯松信明, 西村多寿子, 西成活裕, 櫻庭京子, "構造不変の定理とそれに基づく音声ゲシュタルトの導出," 電子情報通信学会音声研究会, SP2005–12, pp.1–8, May 2005.

[7] 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, Mar 2008.

[8] 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, vol.J90–D, no.5, pp.1249–126, May 2007.

[9] M. Suzuki, N. Minematsu, D. Luo, and K. Hirose, "Sub-structure-based estimation of pronunciation proficiency and classification of learners," Proc. Int. Workshop on ASRU'2009, pp.574–579, Dec 2009.

[10] 朝川智, 喬宇, 峯松信明, 広瀬啓吉, "判別分析と構造構造表象を用いた話者の多様性に超頑健な音声認識," 日本音響学会春季講演論文集, 2–P–3, pp.113–116, Sep 2008.

[11] L. Bachman, "Fundamental considerations in language testing," Cambridge University Press (1990)

[12] J. Bernstein, M. Lipson, G. Halleck, and J. Martinez, "Comparison of oral interviews and automatic tests of spoken language," Language Testing Research Colloquium (LTRC'99) (1999)

[13] International Phonetic Association, "Handbook of the International Phonetic Association," Cambridge University Press (1999)

[14] 川越いつえ, “英語の音声を科学する”, 大修館書店 (1996)

[15] 竹蓋幸生, “日本人英語の科学”, 研究社 (1982)

[16] 窪園晴夫, 太田聡, “音韻構造とアクセント”, 研究社 (1998)

[17] ロゼッタストーン・ジャパン株式会社, “Rosetta Stone” (2007)

[18] メディア教育開発センター, “Listen to Me!,” NHK エデュケーショナル (2000)

[19] 株式会社プロンテスト, “発音力”(2007)

[20] 国際電気通信基礎技術研究所, “ATR CALL”(2000)

[21] 羅徳安 他, 電子情報通信学会音声研究会, SP2009–32, p.51–56, Jun 2009.

[22] 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.

[23] 朝川智, 峯松信明, 広瀬啓吉, “音声の構造的表象を用いた音声認識における特徴両空間分割とその効果”, 日本音響学会秋季講演論文集,3–Q–10,pp.229–232, Sep 2007.

[24] Silke Witt,Steve Young, “Computer–assisted pronunciation teaching based on automatic speech recognition,” Proceedings of ICLSP, 1998.

[25] D.Reynolds and L.P.Heck, “Speaker verification:from research to reality,” Proc. Int. Conf. Acoustics, Speech, and Signal Processing, tutorial session, May 2001.

[26] M. Pitz and H. Ney, “Vocal tract normalization equals linear transformation in cepstral space,” IEEE Trans. Speech and Audio Processing, vol.13, pp.930–944, Sep 2005.

[27] 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 (D–II), vol.J83–D–II, no.11, Nov 2000.

[28] Y. Qiao and N. Minematsu, “f-divergence is a generalized invariant measure between distributions,” Proc. INTERSPEECH, pp.1349–1452, Sep 2008.

[29] 朝川智, 峯松信明, 村上隆夫, 伊勢井敏子・ヤッコラ, 広瀬啓吉, “音声の構造的表象に基づく非母語話者の英語発音分析”, 電子情報通信学会音声研究会,SP2005–24,pp.25–30, Jun 2005.

[30] 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”, 電子情報通信学会音声研究会,SP2005–13,pp.9–12, May 2005.

[31] N. Minematsu, S. Asakawa, and K. Hirose, “Structural representation of the pronunciation and its use for CALL,” Int. Workshop on Spoken Language Technology (SLT’2006), pp.126–129, Dec 2006.

[32] 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, Dec 2009.

[33] Y. Stylianou, O. Cappe and E. Moulines, “Continuous Probabilistic Transform for Voice,” Conversion, IEEE Trans. of Speech and Audio Processing, vol.6, no.2, 1998.

発表文献

[1] N. Minematsu, K. Kamata, M. Takazawa, K. Takeuchi, S. Asakawa, T. Makino, Y. Yamauchi, T. Nishimura, K. Hirose, "Pronunciation clinic – which part of your pronunciation to correct at first to become like your model speaker? –," WorldCALL, Courseware Session, Aug 2008.

[2] X. Ma, M. Takazawa, N. Minematsu, and K. Hirose, "Chinese dialect classification using acoustic universal structure in speech," Proc. Autumn Meeting of Acoust. Soc. Jpn., 2–P–22, pp.405–408, Sep 2008.

[3] N. Minematsu, M. Takazawa, X. Ma, "Pronunciation clinic & dialect-based speaker classification," Proc. Int. Conf. Acoustics, Speech, & Signal Processing (ICASSP'2009, Show & Tell), Apr 2009.

[4] X. Ma, N. Minematsu, A. Nemoto, M. Takazawa, Y. Qiao, K. Hirose, "Structural analysis of Chinese dialect speakers and their automatic classification," Proc. National Conference on Man-Machine Speech Communication (NCMMSC'2009), pp.440-445, Aug 2009.

[5] 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 (2008–3, 発表予定)

[6] 峯松信明，鈴木雅之，高澤真章，馬学彬，中村綾乃，"発音クリニック　～音声の構造的表象を用いた外国語・方言発音分析～"，情報処理学会講演論文集，Mar 2010(発表予定).