修士論文

ボトムアップクラスタリングによる音声の教師無しセグメンテーションとその応用に関する研究

2008年1月29日

指導教員　峯松 信明 准教授

東京大学大学院新領域創成科学研究科
基盤情報学専攻　47-66327

下村 直也

内容梗概

音声処理技術の高度化を図る場合，音素境界ラベルが付与された音声コーパスが要求されることがある．通常，隠れマルコフモデル (Hidden Markov Model:HMM) などの音響モデルと読み上げテキスト (意図された音素列) を用いた強制切り出し（forced alignment）を行なうことが多い．この場合，学習用音声データを用いて構築された音響モデルの事前学習が必要であり，学習用音声データと同一発話スタイルのデータに対してのみ，高精度な処理が可能となる．例えば，非母国語者の音声は話者によって習熟度が異なるため，その話者の母国語，もしくは学習対象言語の音響モデルのみでは精度に限界があり，また，感情の入った音声や歌声等の特殊なデータに対しては読み上げ音声から構築された音響モデルでは精度が落ちる． HMM では，大量の音声データに対する数理統計的な手法によってその精度向上を実現してきたが，これは逆に，大量のデータを準備できない問題に対しては，その精度向上が困難とならざるを得ない．

一方，事前に用意された音響モデルを使用せず，音響分析のみによって音素境界検出 (セグメンテーション) を行なう手法が提案されている (教師なしセグメンテーション)．この場合，事前データを必要としないため，学習データ依存性は原理的に無い．音響分析のみに基づく方法論は，明示的に音素書き起こしを与えないため，挿入／削除誤りは，HMM を用いた音素アライメントに比較すれば格段に増加するが，検出された境界の時間的精度は比較的高い値を示す．また，上記したような学習データの依存性に対して高い頑健性が期待できる．

本研究では，学習データを必要としない，音響分析に基づく新しいセグメンテーション手法として，ボトムアップクラスタリングによる音素境界，及びイベント境界検出手法を提案する．単にクラスタリングを行なうのではなく，音声に不可避な時系列としての制約を反映させたクラスタリングを提案する．この手法は，クラスタ数を増減することによって対象とするイベントの「粒度」を制御可能である．この特徴を活かし，連続音声中に含まれる音素数を自動推定する手法についても検討し，更には，異なる言語単位に対する境界検出精度についても実験的に考察する．また，異言語音声に対する境界検出の頑健性についても報告する．

そして入力データに対する低い依存性を活かし，本提案手法をシャドーイング音声の自動評定に応用した．シャドーイングとは，聴取した外国語音声を即座に繰り返して発声する外国語聴取・発音訓練法である．学習負荷が非常に高く，シャドーイング音声は崩れた発声になることが多い．シャドーイング対象の文を適切に選定することで，自動評定スコアと教師による手動評点スコアとの相関は良好な値を示した．

目次

図目次

表目次

第1章

序論

1.1　コーパスベースの音声処理に根付く問題点

音声処理や自然言語処理の分野において，音声，言語コーパズが必要となることは言うまでもない．また，昨今の統計的手法の発展により，大規模コーパスがシステムの学習のために必要とされるようになってきた．そのような状況下，研究を進める上で，共通の学習データや評価データを構築し使用していくことは必然的に求められることである．

現在の web の進化により，書き言葉に関しての多くのテキストデータの収集が可能となってきた．特に，国立国語研究所では 2006 年から 5 年間かけ，インターネット上の文書 500 万語を加えた約 1000 万語のサンプルを有する現代日本語の書き言葉コーパスの構築を予定している．その完成により，言語学や日本語教育，自然言語処理などの分野で幅広い効果が見込まれている．

それでは，音声コーパスの場合どうだろうか．web 上に置かれている音声データは書き言葉データに対して極々僅かであることは容易に想像できる．また，音声コーパスには性別や年齢，地域，雑音などの情報が全て介入してくる．さらに言えば，同一話者が同一単語を発声した音声あったとしても，全く同一の音声であるということは基本的にあり得ない．このことからも，例えば日本語に絞ったところで，どれだけの時間的，経済的コストをかければ，統計的に十分な音声認識システムを構築できるのか，想像しがたい．

このような，永久的に続くかに思われる音声コーパスの欠乏を解決する根本的な方法論として，峯松が提案した音声の構造的表象がある [1]．この方法論は，非言語情報を除去することにより，非常に少数の学習話者のみで不特定話者の音声認識を実現している．

また，抜本的解決策とはいかないまでも，音声コーパスのラベリング簡略化をはかる技術に音声自動セグメンテーションの技術がある．音声を自動で切り分け，各音声処理システムに利用しやすいように加工することが行なわれている．それにより，欠点である時間的，経済的コストの大幅な軽減が可能となる．この音声自動セグメンテーションの技術は各研究の基幹であるため，高い精度はもちろんだが，どのような音声でも処理可能であること，つまりは高い頑健性が問われる．

1.2　本研究の目的

本研究では，新しい連続音声の教師無しセグメンテーション手法を提案し，主に 2 つの点について検討することを目的とする．1 つ目の目的は，他のセグメンテーション技術と比較検討を行ない，境界の検出精度や頑健性など，要素技術としての優位性を示すことである．

2 つ目の目的は，本手法が音声処理の基盤技術として位置づけられるため，今までに解決できなかった，もしくは手を出しづらかった音声処理を可能にすることを具体的に示すことにある．例えば [2] で紹介されている教師なしセグメンテーション手法は，その最終的目標を子供の言語獲得のシミュレーションにおいている．赤ん坊は連続した音声信号から，単語を取り出し，やがては音素単位といった細かい単位での切り出しをおこなっていく．その際一種の音声モデルを保持して切り分けているとすれば，鶏が先か卵が先か，と

いう議論と同様の問題が浮上してしまう．この 1 つの解として，聴覚特性を活かして (つまりはスペクトル分解を行ない) 音声を細かいイベントに分割し，カテゴリー化できるようになっていくのではないのかと提案し，そのシミュレーションに教師無しセグメンテーション手法を応用している．

本研究では，提案する教師無しセグメンテーション手法を外国語学習法であるシャドーイングの学習支援技術に応用している．シャドーイングは外国語音声を聴取しながら，即座にその通りに発声を行なう学習方法である．学習者にとって非常に負荷が高く，困難な作業であるため，シャドーイング音声は歪んだ音声になりやすい．よって，処理のしにくい対象であると言える．

1.3　本論文の構成

本論文では，まず第 2 章において，本研究に用いた音声研究における基礎的知識について述べ，第 3 章では，近年提案されている連続音声の自動セグメンテーション手法について先行研究を述べる．続いて第 4 章では，本研究が提案するボトムアップクラスタリングによる教師無しセグメンテーション手法を解説する．第 5 章では，本手法の特徴の洗い出しを行う．自動停止条件の設定による音素数推定や先行研究との比較，異なる言語単位による検出能力の検討，音素毎の検出精度の確認等を報告する．そして第 6 章では，提案手法を応用し，シャドーイング音声の自動評価手法について実験結果も含めて述べる．最後に第 7 章でまとめと今後の課題について述べる．

第2章

研究の背景

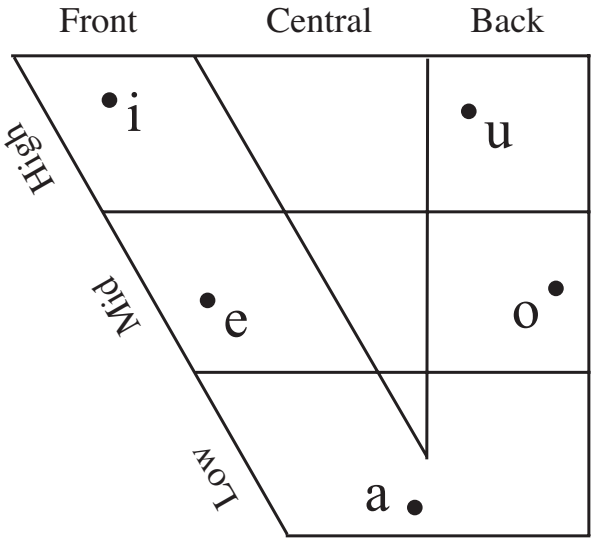


図 2.1: 日本語における母音図

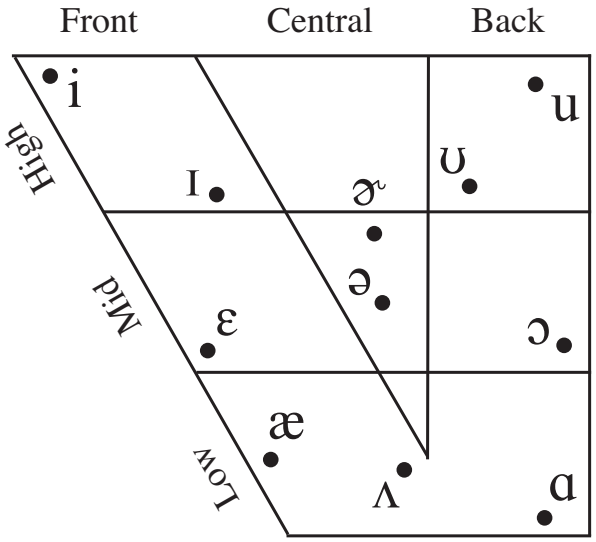


図 2.2: 英語における母音図

2.1　音声と言語

文を形作る基礎は**単語**である．単語は，音英語では音節あるいは**シラブル** (syllable) と呼ばれるものを発声の基本的単位により構成され，日本語は音節より小さな発声の単位である**モーラ** (mora) を基本的単位として構成される．音節は**音素**から成り立つ．音素には**母音**と**子音**がある．

2.1.1　母音

日本語における母音図を図 4.1 に示す [3]．母音図は発音するとき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），横軸は舌の部位（前舌か後舌か）を表している．日本語の母音は/a, i, u, e, o/の 5 母音である．一方，英語における母音図を図 4.2 に示す．図は二重母音に関しては省略してある．英語の母音は，**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 [4]．また，英語には弱母音 (schwa) と呼ばれる，最も脱力して発声できる (調音努力が少ない) 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 schwa の発音が特に苦手であると言われている．

2.1.2　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 [4]．

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 [ɹ]，イギリス英語は [ɹ]）は，日本語ではラ行の子音である [ɾ] に置き換えられることが多い．また，/f,v/や/θ, ð/は日本

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				
接近音 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.

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．

2.1.3　音節

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．

また，その他の言語でも 1 言語で用いられる音素の数が 50 を超えることは少ない．実際には，それぞれの音素が組み合わさって語になるときのつながり方について，いくつかの制限があり，全ての組み合わせが存在する訳ではない．このため，1 つの言語の中でも用いられる音節の数は，音素の組み合わせの数に比べればはるかに少ない．

2.2　ラベリングデータをもつ音声コーパス

2.2.1　音声コーパスに基づく統計的音声処理技術

音声研究において音声ラベリング情報が付与されたコーパスは非常に重要な役割を果たしてきた．音声認識，音声合成に限らず，各種音声システムは音声コーパスを用いた統計処理に基づくモジュール開発が行なわれることが多い．

現在の音声認識技術の基本となっているのが，音響モデルと言語モデルであり，音響モデルの中心的技術となっているのが**隠れマルコフモデル** (Hidden Markov Model:以下，**HMM**) である [6]．詳しくは 3.2 節で述べるが，HMM による音声認識は，コーパスを用いて事前学習を行ない，各音素毎に確率モデルを構築する．即ち，各音素からどのような音響量が視測されやすいのかをモデル化する．それにより，入力音声がどのような単語か，もしくはどのような音素であるか，認識可能なのである．当然，音声の音響特徴は話者によって変動するため，数多くの音声特徴量に関する知識を蓄積させていかなければ精度向上が見込めないのが一般的である．このように，音声コーパス環境を整えることは，音声工学技術底上げへとつながる．

2.2.2　大規模データベース

多くの音声研究で音声コーパスを用いるにもかかわらず，各研究毎に音声コーパスを構築していたのでは時間的，経済的コストがかかりすぎる．そこで，種々の研究目的を満たす音声コーパスの構築がなされてきた．ここでは，本研究で用いた大規模データベースである，**ATR データベース**と **TIMIT データベース**について述べる．両データベースとも，ラベリングは音声学の専門家がスペクトログラムの視察に基づき行なっている．

i)　ATR データベース

ATR データベースは，日本語読み上げ音声のコーパスである [5]．できるだけ少ない数の文で，実際の日本語に含まれてる音声現象を可能な限りカバーするように，発声現象の基本単位として音素をとり，基本単位の出現頻度のバランスが取られた 503 文を新聞，雑誌から選択している．

データセットは波形データとラベルデータから成り立ち，波形データはサンプリングレート 20KHz，16bit で量子化されている．ラベリング情報は用途に応じた利用が可能なように以下に示す階層的ラベル構成になっている．

- 音声記号層
  …発音を母音部と子音部に分割して記述しており，51 音素から構成されている．また，境界が決定できない場合は，その区間の分割を行なっていない．
- イベント層
  …音声記号層をさらに分割している．また，語頭，語尾に伴う過渡現象，及び何らかの原因でスペクトルパターンに乱れが生じている区間にセグメントを設ける．

- 異音化層
  …異音化が発声してる区間にセグメントを設ける．ただし，複数の音韻にわたって，異音化が発生している場合は，音声記号層の境界に関わらず，異音化の発生時点から終了時点までをスペクトログラムから読み取り，セグメントする．
- 融合化層
  …融合化が発生して，音声記号層，イベント層で予想される境界が明確でない場合，この層にセグメントを設ける．
- 母音中心層
  …スペクトログラムが母音の特徴を顕著に表現しているポイントをマークする．

2.2.3　TIMIT データベース

TIMIT データベースは，米語読み上げ音声のコーパスである．61 音素種から成り立つ．構成単語は，MIT が発行していた 1964 年度版 Merriam-Webster ポケット辞書に由来する常用単語 6229 単語である．波形データは，サンプリングレートは 16KHz であり，16bit で量子化されている．ラベリングデータは，ATR の音声記号層と比べ，境界が決定できない場合でも分割を行なっている分，詳細ではあるが，やや強引に分割が行なわれていると言える．

2.3　音声データ分析に用いられる音声特徴量

2.3.1　ケプストラム

音声分析で抽出される音声特徴量として現在最も広く用いられているのは**ケプストラム** (cepstrum) である．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 2.3 に示す．まず，音声波形から数十ミリ秒単位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, 求められる．

このケプストラムのうちの低次項 (通常は十数次元) 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．スペクトル包絡の局所的な最大値である山 (伝達関数の極に相当) のことをフォルマント周波数を呼ぶ．2 つ目までのフォルマントの位置がおよそ音韻を決定し，3 つ目以降のフォルマントに発声者の話者性等がよく表れる．つまりケプストラムは，音声の音韻的特徴や話者性などの情報を効率よく表すことのできるパラメータである．

2.3.2　メルケプストラム

人の聴覚は低い周波数では細かく、高い周波数では粗い周波数分解能を持つことが知られている．この特性に準ずるよう，音声波のスペクトルを人の聴覚に近い周波数間隔に切

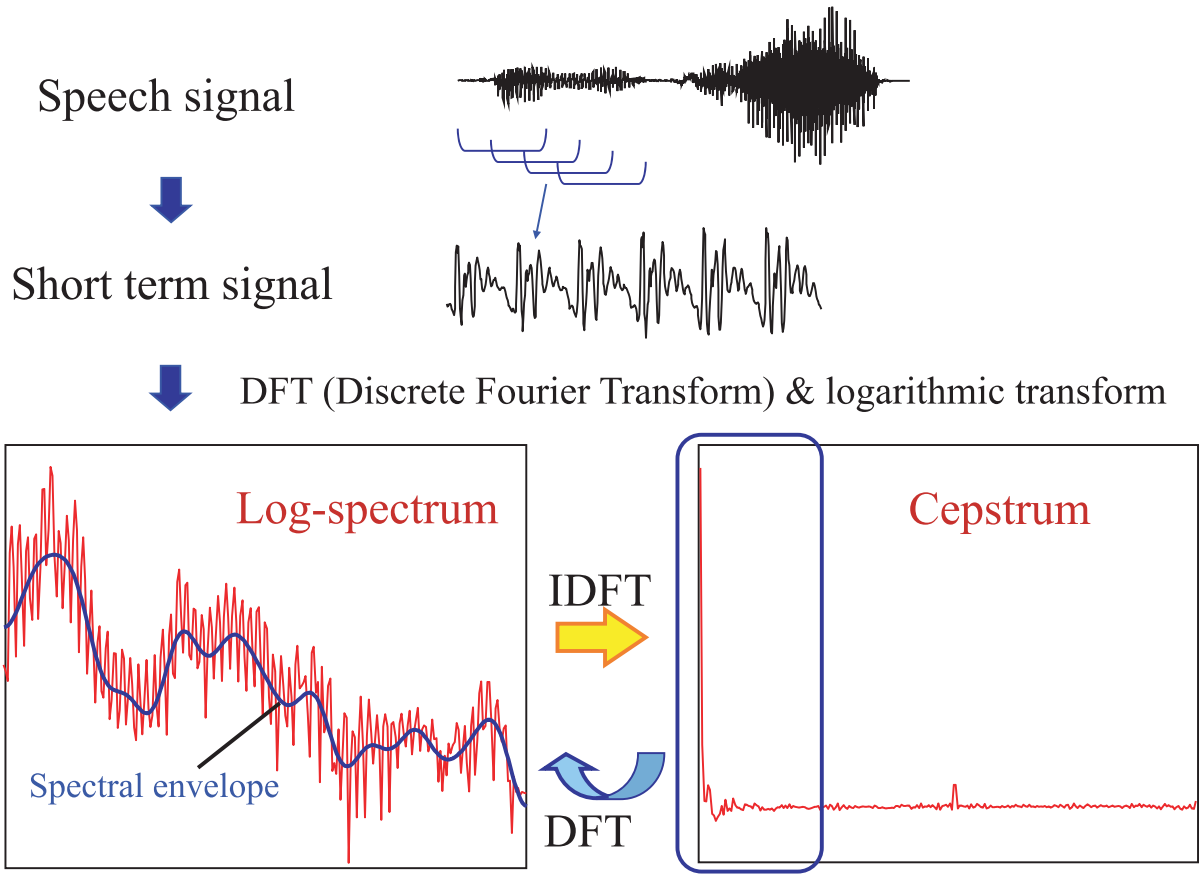


図 2.3: ケプストラムの抽出方法

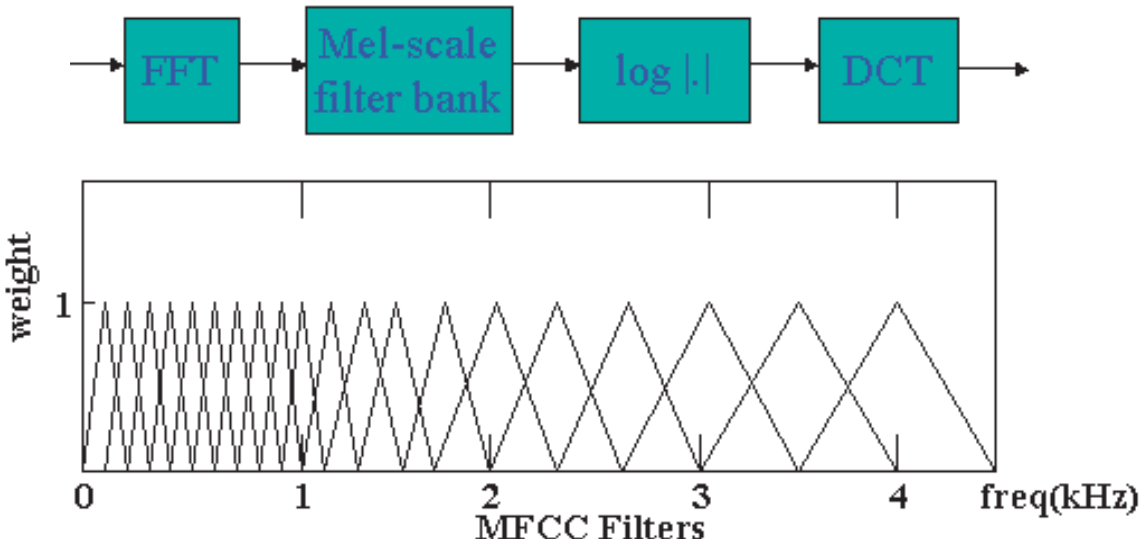


図 2.4: メルケプストラム導出過程及びメルスケール

り分けてケプストラム化したものを**メルケプストラム** (Mel-cepstrum) という．FFT によるスペクトルに対して**メルスケール**,

$$\mathrm{Mel}(f) = 2585 \log_{10}(1 + \frac{f}{700}) \tag{2.1}$$

を乗算し算出される．メルケプストラム導出過程及びメルスケールを図 2.4 に示す．

本研究において，メルケプストラムの算出は音声信号処理ツールキット，SPTK(Speech Signal Processing ToolKit[7]) を用いて算出した．

2.3.3　Δ ケプストラム

ケプストラムの変化量 (Δ=速度に相当)，Δ ケプストラム係数は次の一般式により計算される

$$\Delta c_t = \frac{\sum_{x=1}^{w} x(c_{t+x} - c_{t-x})}{2\sum_{x=1}^{w} x^2} \tag{2.2}$$

ここで c はケプストラムベクトル，c_t は時刻 t での Δ 項であり，c_{t-w} から c_{t+w} までの区間で算出される．

本研究ではケプストラムから Δ ケプストラムを抽出する際，HTK(Hidden Markov Model Toolkit - Speech Recognition toolkit) を用いている [8]．HTK ではデフォルトでは $w = 2$ と設定しており，本研究でも同様に設定した．つまり，時刻 t の Δc_t は前後 5 フレームにより計算する．

第3章

本研究に関連する先行研究

3.1　はじめに

2.2.2 節で紹介した音声データベースのセグメンテーション情報は，人の目視と聴取によって認識され，付与されたものである．そのような時間的，経済的コストが膨大にかかってくる音声データのセグメンテーション作業を自動化する手法がいくつか提案されてきた．本章では，本研究と競合しあう，音声の自動セグメンテーション手法を述べる．

3.2　HMM を用いた強制アライメント手法

HMM 音響モデルを用いた自動ラベリング手法は現在までに多く報告されている [9]〜[14]．ここでは，まず HMM による基本的ラベリング手法を述べた後，近年報告されてる技術について述べる．

3.2.1　音声認識の確率論的モデル化と音響モデル

入力音声の音響特徴量時系列 $X = (x(1), x(2), \cdots, x(t))$ について，発話されたある単語列 $W = (w(1), w(2), \cdots, w(t))$ から観測された音響特徴量の時系列 X を統計的に推定することが HMM による音声認識の目的である。つまり、観測された X が W である尤度

$$P(W|X) \tag{3.1}$$

を最大化するような単語列

$$\hat{W} = \arg\max_{W} P(W|X) \tag{3.2}$$

を求める問題といえる。W が未知なので式 3.2 の右辺を直接求めることは不可能であるが、ベイズの定理を用いて

$$P(W|X) = \frac{P(X|W)P(W)}{P(X)} \tag{3.3}$$

と変形することで、次のように定式化することができる。

$$\hat{W} = \arg\max_{W} P(X|W)P(W) \tag{3.4}$$

ここで、右辺第 1 項の $P(X|W)$ は、ある単語列 W が発声されたときに音響特徴量の時系列として X が観測される確率を意味する。このような W と X に関する統計モデルを**音響モデル**という。また、第 2 項の $P(W)$ は単語列 W が生成される確率を表しており、これは言語的な性質だけできまる統計モデルであり、これを**言語モデル**という。

式 3.4 が与える統計モデルが音響モデルであり、HMM によって実現する。これは、ある出力系列が与えられたときに、それを与える隠れ状態を仮定し、その隠れた状態の確率モデルのパラメータを推定することで得られる。推定された各状態のパラメータと、マルコ

フモデルの遷移確率によって与えられるモデルが音響モデルである。ひとつの音響モデルにつきひとつの音素モデルを対応させることで、音素 w を発声したときに得られる音響特徴量の系列 $\{x\}$ を得る確率（これが最終的に $P(X|W)$ を与える）を得ることができる。即ち、$\{x\}$ から w を推定することができる。ここでは、X を与える音響特徴量として、スペクトル包絡に基づいた音韻情報である。

音声認識で最もよく用いられる left-to-right 型の音素 HMM を図 3.1 に示す。HMM は遷移確率 a_{ij} で状態 i から状態 j へ遷移を行い、状態 i では確率分布 $b_i(X)$ に従って音響特徴量系列 X を出力する。HMM のひとつの状態は音声の定常的な部分信号を表し、状態遷移は信号の変化を表しているといえる。同じ音素でも状況や環境などによって音声信号は大きく変化するが、時間的な揺らぎは確率的な状態遷移によって、スペクトル的な変動は音響特徴量の出力確率分布によってそれぞれ吸収されるので音素の特徴をうまく認識することができる。そのため，HMM は音声の生成モデルとして非常に優れていると言える．

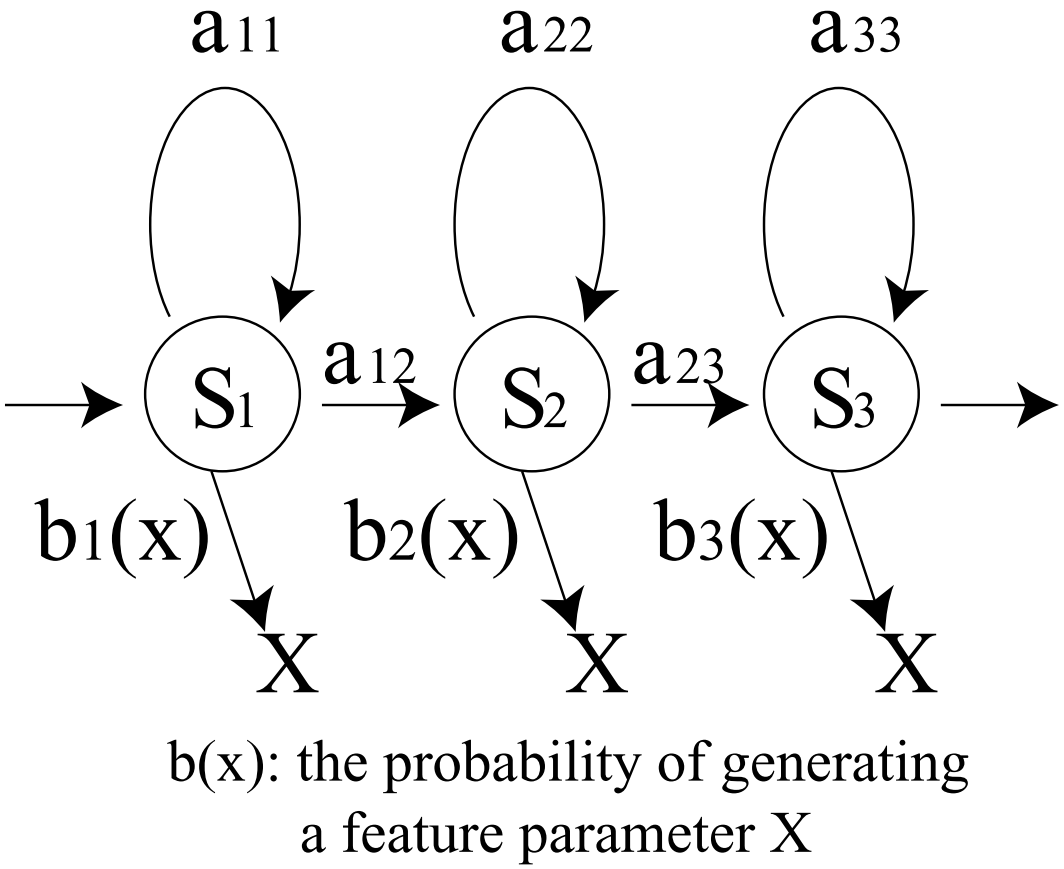


図 3.1: HMM における状態遷移の様子

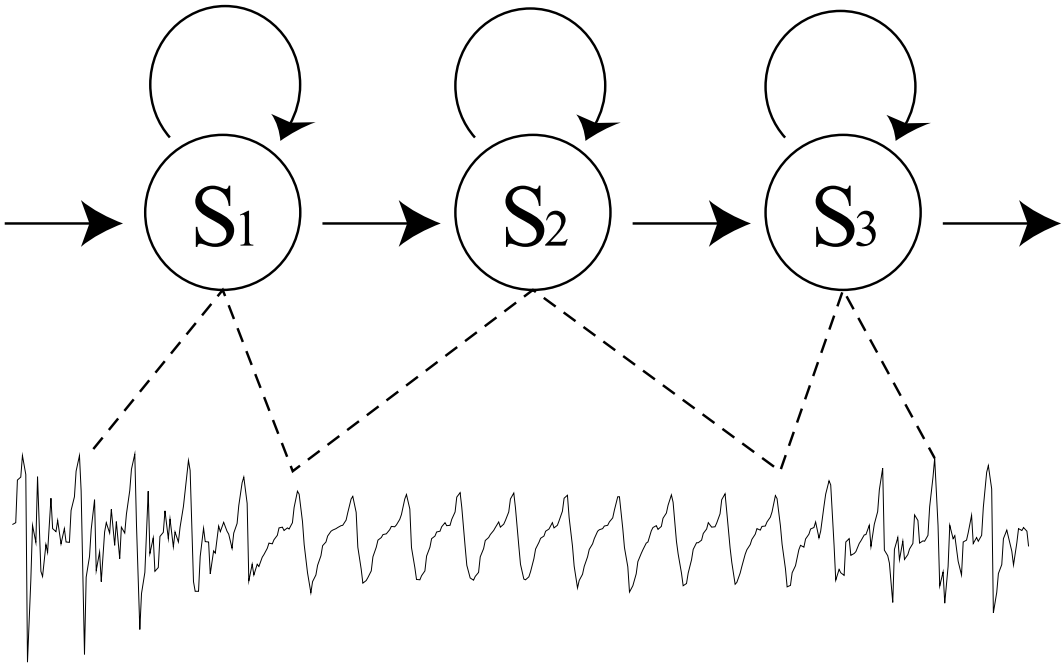


図 3.2: 音素 HMM と対応する音声波形

3.2.2　HMM からの音響特徴量時系列の出力確率の計算

前節で述べたように音響モデルは確率 $P(X|W)$ を与えるモデルである。よって音素 HMM を用いて，自身が表している音素 W から任意のある音響特徴量系列 X が生成される確率を求めることができる。この計算は次のように行うことができる。

図 3.3 は音響特徴量の時系列 $X(1), X(2), \cdots, X(7)$ が 3 状態の音素 HMM から出力される場合の可能な状態遷移の経路を表している。ある 1 つの経路を通って時系列 X が出力される確率は、その経路の状態遷移確率 a_{ij} と経路上の各状態での音響特徴量の出力確率 $b_i(X)$ の積によって計算できる。図 3.3 に示された経路全てに対してこの確率を求めて和をとることで、この音素 HMM から時系列 X が出力される確率を求めることができる。これはまさに $P(X|W)$ を求めたことに他ならない。

しかし、全ての経路からの出力確率の和をとると計算量が増大してしまうため、実際には最も出力確率の大きな経路のみを計算していく近似が行われる。この近似を行って出力確率を求める方法は**ビタビアルゴリズム**と呼ばれ、厳密に計算するのと比べて結果はそれほど変わらないが、大きく計算量を減らすことが可能である。

3.2.3　HMM の学習

HMM のパラメータ（これは a_{ij} と $b_i(X)$ であるが、以下ではまとめて λ とする）は、その HMM が表している音素を発声したときに観測されやすい音響特徴量系列を高い確率で生成するように学習されなければならない。ある音素の学習用音声データから、その音素

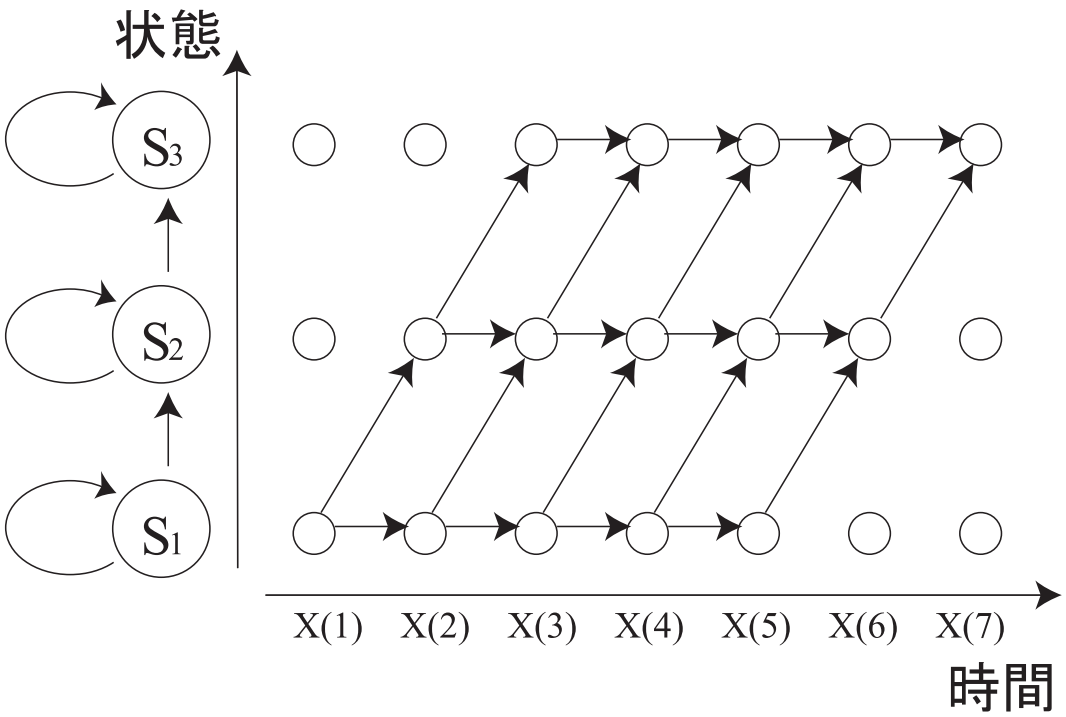


図 3.3: HMM の状態遷移

を表す HMM のパラメータ λ を学習するには、学習データから抽出された特徴パラメータ系列を X として、確率 $P(X|\lambda)$ を最大とする λ を求めればよいと考えられる。式で表すと次にようになる。

$$\hat{\lambda} = \arg\max_{\lambda} P(X|\lambda) \tag{3.5}$$

つまり、学習データの系列 X を最も高い確率で生成する HMM のパラメータ $\hat{\lambda}$ を求める**最尤推定**を行うことになる。

しかし最尤推定法で HMM のパラメータを解析的に決定するのは困難で、そのような方法は知られていない。そこで、$P(X|\lambda)$ を局所的に最大化するアルゴリズムとして **Baum-Welch** 法が考案された。Baum-Welch 法は初期パラメータ λ を元に、学習データ X を用いてパラメータを再推定し、$P(X|\hat{\lambda}) > P(X|\lambda)$ となる新しいパラメータ $\hat{\lambda}$ を求めることができる手法である。再推定されたパラメータ $\hat{\lambda}$ を再び初期パラメータとして計算を繰り返すことで $P(X|\lambda)$ を局所的に最大化するパラメータを得ることができる。

このように、1 つの HMM に対して対応する音素の学習用音声データを用いて Baum-Welch のアルゴリズムを適用することで、HMM のパラメータ学習が行われる。これには大量の学習データが必要であり、音声データベースの整備が非常に重要である。

近年、HMM 学習で変分ベイズ的な Baum-Welch アルゴリズムを用いることで過学習とデータ量の問題を解決し、さらに学習を高速に行う研究も提案されている [15]。

3.2.4　HMM と書きおこしテキストを用いた強制アライメントとその精度

HMM を用いて，セグメンテーションを行なう場合，入力音声の書きおこしデータを事前知識として与える．つまり，(3.4) 式において $P(W) = 1$ とし，音声認識を行なうことに相当する．単語列を与え，それに準じたモデルで強勢的に入力音声を切りわけることから，このことを**音声の強制アライメント**と呼ぶ．

HMM 技術が提案された当初，話者は数名に固定し，学習に用いたテキストデータと評価に用いたテキストデータを変化させた実験で，20msec 以内にセグメンテーションを確認できる率は，おおよそ 80 %程であった [9]．その後，学習に用いるデータベースの充実や，様々な提案技術により，精度の向上がはかられてきた．例えば，[10] では音響モデル作成の学習時の基準に，最尤法を用いるのではなく，アライメント時の境界誤りが最小となるような新しい尺度を提案し，精度向上を成し遂げている．また，[11] では，HMM により算出された音素境界の内，正解である境界と誤り境界を SVM により分割し，誤検出を低減することが提案されている．このように様々な面で最適化された近年の HMM の境界検出精度は，20msec 以内に約 90 %である．しかし，以上の結果は全て正しい書きおこしは与えた場合であり，書きおこしが入手困難な場面では使えない．

3.2.5　HMM によるアライメント技術の欠点

HMM ベースの手法は高い精度を示すものの，事前学習が必要であり，学習に用いた音声データと同一発話スタイルのデータに対してのみ，高精度な処理が可能となる．例えば，非母国語者の音声は，その話者の母国語，及び学習対象言語の音響モデルのみでは精度に限界があり [12]，また，感情の入った音声や歌声等の特殊なデータに対しては読み上げ音声から構築された音響モデルでは精度が落ちる．更には，音声データベースの構築が困難な（つまり，音響モデル構築が困難な）幼児音声のラベリングなど,「事前に大量の音声データが用意できない」状況に遭遇することは珍しいことではない．適応技術によりデータに対して頑健な処理を行なう研究が報告されているものの，根本的な解決には至っていない [13][14]．音声認識技術は大量の音声データに対する数理統計的な手法によってその精度向上を実現してきたが，これは逆に，大量のデータを準備できない問題に対しては，その精度向上が困難とならざるを得ない方法論であると言える．

3.3　スペクトルピーク検出によるセグメンテーション

異なる音素が発生させられた時，各音素間には必ず状態遷移が存在する．その遷移はスペクトルの変化として現れるため，スペクトルピークを抽出することにより音素境界を定めることができる．この手法を本稿では Maximum Spectral Transition Positions 法 (**MSTP 法**) と呼ぶ．

MSTP 法は音素セグメンテーションの基本的な手法として知られ，現在でも研究が続けられている [17, 18]．MSTP 法はスペクトルの遷移度（Δ ケプストラムの絶対値）がピー

クとなる時点を音素境界候補とし，後処理で候補を適宜削減することで精度向上を成し遂げている．

[17] では，まずスペクトル遷移の尺度 (Spectral transition measure(STM)) を，各フレームにおける 2 乗平均によって計算している．

$$STM(m) = (\sum_{i=1}^{D} a_i^2(m))/D \tag{3.6}$$

ここで D はスペクトル特徴量を表すベクトル (10 次元)，$a_i(m)$ はスペクトル特徴量，メルケプストラムの変化率を表し，

$$a_i(m) = (\sum_{n=-I}^{I} \mathrm{MelCepsturm_i(n+m)*n})/(\sum_{\mathrm{n=-I}}^{\mathrm{I}} \mathrm{n}^2) \tag{3.7}$$

により定義される．n はフレーム項を表し，I はフレーム数を表す．ここでは $I = 2$ と定める．フレームシフト長を 10ms とすると，STM は現在のフレームを中心に 40ms の間隔をとり，その変化率を求めることと一致する．I を大きくすると変化をより平均化してしまうので音素境界の見落としにつながり，I を小さくすると変化に敏感になり誤り音素境界を多く検出してしまう．

また，下記のような後処理を提言し，誤検出を削除している．

- あるピークと，時間的に隣り合っている遷移度との差分が，ある閾値以下の場合そのピークを除外する．この閾値は発話の最大ピーク値の 1%とする．
- あるピークに対して，その前後に存在する “谷”に着目する．ピークと谷との差分がピークの 10%以下なら，そのピークは排除する．

上記条件を満たす STM のピーク値を検出し音素境界と定める．

MSTP 法は変化を捉えるため，似ている音素については境界検出が困難になることがある．例えば/d/，/ow/，/m/等は非常にピークが小さく平坦になるので，音素境界結果として出力されにくい．また逆に 2 重母音では誤り検出されやすい．

正解音素境界数など一切与えていないため，40msec を許容誤差とした場合，挿入率 28.2%，脱落率 18.0% の精度を報告している．本稿では [17] の手法を，論文に沿って可能な限りそのアルゴリズムを忠実に実装した．MSTP 法の一例を図 3.4 に示す．

2007 年に発表された [18] では，8ms のフレームシフトで求めたスペクトルをスムージングした上で，0-500Hz，500-1420Hz，1420-2386Hz，2386-8000Hz の帯域に分けている．そして各帯域のいずれかにおいて，エネルギーの和がある閾値を超え，かつ，ピークとなっている時間を音素境界とすることを提案している．この手法では，20msec 誤差で 86 %の正解境界を検出し，8.4 %の誤検出を出力している．

MSTP 法は音響分析に基づく（事前学習を閾値設定以外では必要としない）音素セグメンテーション手法である．そのため，データの不一致問題は発生せず，言語間差異等には頑健な処理が可能であると言える．

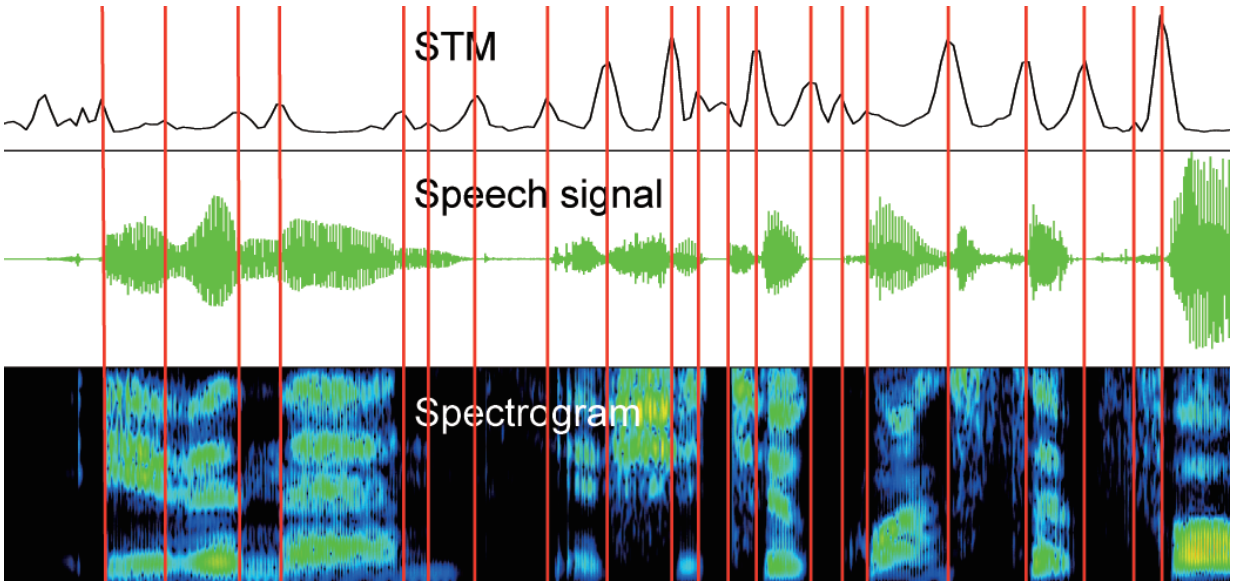


図 3.4: MSTP 法のセグメンテーション一例

3.4　その他のセグメンテーション手法

スペクトルの局所最大値をとる手法以外に，音素間のセグメンテーションにどのような変化が現れるのかを特別な特徴量やアルゴリズムを用いて捉え，セグメンテーションを行なう研究がなされている．ここでは，**SVM(サポートベクターマシン)** によるセグメンテーション手法を紹介する [2, 19].

[19] では，16kHZ の連続音声から 5ms フレームシフトさせながら，12 次元の MFCC とパワー及び，その Δ，ΔΔ，計 39 次元を算出し，ある時刻から時系列上で ±9 フレーム分の要素に対して，SVM を用いることで 2 種類のデータに分類を行なう．その時各フレームにおける結果が Fig. 3.5a に出力されている．基本的に ±0 フレーム付近で色が変わる (つまり，そのフレームを境に各フレームが SVM により分類されている) 場合，そこは境界として検出される．枠 B が一例を表している．

また SVM で分類された 2 クラスタの平均ベクトルのユークリッド距離を求めたものが，Fig. 3.5b である．Fig. 3.5b におけるピークは，2 分割されたフレーム群の距離が最も離れたところであり，そこに音素境界があると定めることは妥当である．

この手法では，Fig. 3.5a 及び Fig. 3.5b で検出された境界が，それぞれ k(=1,2,3) フレーム以内に存在するとき，音素境界として出力している．

この手法での検出精度は 20msec 以内誤差の検出精度が 76.0 % である．また，音素数は既知として扱っている．音響モデルに基づかない処理であるため，データ依存性は低いと言える．

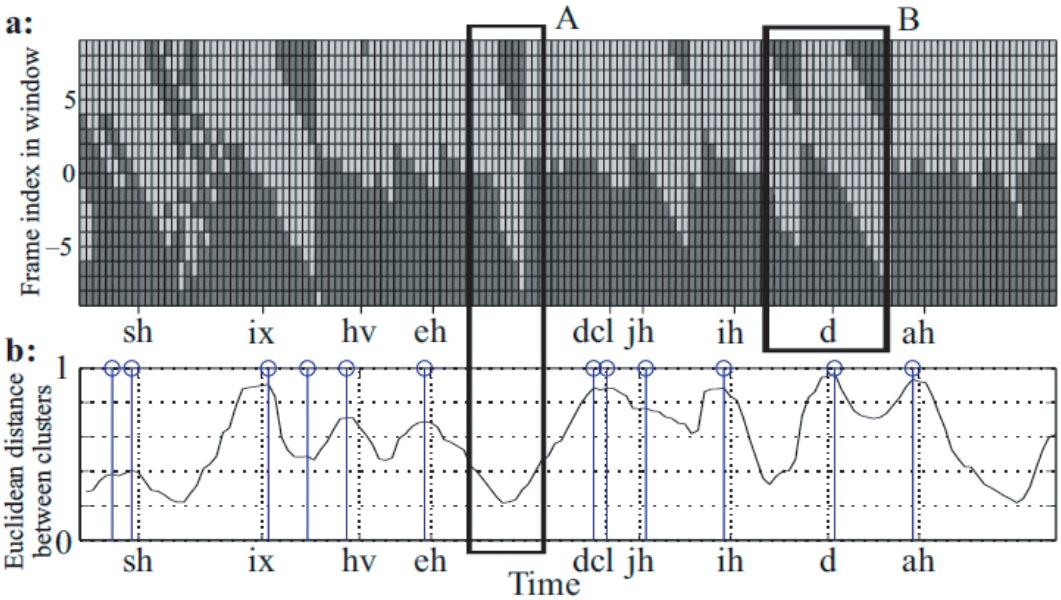


図 3.5: a:クラスタリング結果.b:クラスタ間の平均ベクトルのユークリッド距離

表 3.1: セグメンテーション手法の先行研究の比較

手法	誤差 20msec	音素数
HMM によるアライメント	92.4%	既知とする
MSTP 法	86.9%かつ 20%の挿入誤り	停止不可
SVM	76.0%	既知とする

3.5　まとめ

本章では，各先行研究のセグメンテーション手法について述べた．またその精度について述べた．表 3.1 に各手法の比較を簡単に示した．同条件で実験を行なったわけではないので，実験条件や評価用データベースは手法により多少の異なりがあることを明記しておく．検出精度では，HMM によるアライメント手法が群を抜いて良好な結果を示すが，大規模データにより音響モデルを学習し，処理データに適応させなければならないという，欠点が存在することを忘れてはならない．

第4章

時間制約を設けた
ボトムアップクラスタリング

4.1　ボトムアップクラスタリングによるセグメンテーション

異なる音素が連続的に発声された時，各音素間にはある音響事象から異なる音響事象への遷移が存在する．多くの場合，この遷移はスペクトル包絡の変化として観測される．しかし，連続音声の各音素内に注目すると，逆に，スペクトルの安定区間が存在する．MSTP 法はスペクトルの局所的変化を積極的に検出する方法であるが，本研究では，音響的に類似した連続するフレーム同士をマージする形で纏め上げ（**ボトムアップクラスタリング**），音声ストリームの大局的な階層構造を抽出することを考える．この構造抽出の結果，副産物として音声ストリームの区分分割，即ち，イベント境界位置が得られる．

本手法は MSTP 法では実現できない，言語単位に依存しない境界検出手法としても位置づけられる．MSTP の境界検出はスペクトルピークのみであるが，本手法は N 個のフレームに対して N 段階の粒度を設けることができる．そのため音素のみならず，シラブルやモーラといった境界検出可能性をもつ．入力データに対して非依存，出力とする言語体系に対しても非依存な非常にロバスト性の高い手法である．

4.2　コスト関数としての非類似度

ボトムアップクラスタリングは，クラスタ間に距離を定義し，距離が近いクラスタ同士をまとめていく作業である．本研究では 5 つの手法を検討する．

クラスタ (p) とクラスタ (q) を融合して新しくクラスタ (r) をつくることを考える．このとき，融合してつくられるクラスタ (r) と別の任意のクラスタ (s) との間の非類似度 d_{rs} の更新式がそれぞれの手法により異なる．それぞれのクラスタ内に含まれる構成サンプル数 (本研究ではサンプル数と呼ぶ) を n_p，n_q，n_r，n_s とおいたときの**非類似度** d_{rs} の各更新式を示す．この非類似度が最も低い (即ち距離が近い) クラスタ同士を結合させる [20].

4.2.1　最短距離法

最短距離法では非類似度 d_{rs} をクラスタ (r) に含まれる個体とクラスタ (s) に含まれる個体間との非類似度の最小値で定義する．

$$d_{rs} = \min(d_{ps}, d_{qs}) \tag{4.1}$$

4.2.2　最長距離法

最短距離法では非類似度 d_{rs} をクラスタ (r) に含まれる個体とクラスタ (s) に含まれる個体間との非類似度の最大値で定義する．

$$d_{rs} = \max(d_{ps}, d_{qs}) \tag{4.2}$$

4.2.3　群間平均法

群間平均法では，クラスタ (r) に含まれる個体とクラスタ (s) に含まれる個体の可能な全ての非類似度の平均により，両クラスタ間の非類似度 d_{rs} を定義する．

$$d_{rs} = \frac{n_p d_{ps} + n_q d_{qs}}{n_p + n_q} \tag{4.3}$$

4.2.4　重心法

重心法では，ユークリッド空間の距離による非類似度を前提にしている．その非類似度 d_{rs} はそれぞれのクラスタの重心間距離にもとづいて定義する．今，各クラスタについて m 次元の観測値 $(x_1,\cdots,x_m)$ が得られているとする．

クラスタ (p)，(q) の重心をそれぞれ $(\overline{x}_1^{(p)},\cdots,\overline{x}_m^{(p)})$，$(\overline{x}_1^{(q)},\cdots,\overline{x}_m^{(q)})$ とすれば，クラスタ (r) の重心 $(\overline{x}_1^{(r)},\cdots,\overline{x}_m^{(r)})$ は次のように表される．

$$\overline{x}_j^{(r)} = \frac{n_p \overline{x}_j^{(p)} + n_q \overline{x}_j^{(q)}}{n_p + n_q} \tag{4.4}$$

このとき，クラスタ (r) と (s) の重心間のユークリッド平方距離を非類似度 d_{rs} と定義すると

$$d_{rs} = \sum_{j=1}^{m} \{ \frac{n_p \overline{x}_j^{(p)} + n_q \overline{x}_j^{(q)}}{n_p + n_q} - \overline{x}_j^{(s)} \}^2 \tag{4.5}$$

4.2.5　Ward 法

Ward 法もユークリッド距離に基づくクラスタリングである．2 つのクラスタを結合した際の，群内平方和の増加量を非類似度と定義する．

クラスタ (p) に含まれている i 番目のサンプルを考え，その変量に関する観測値を x_{ij}^p と表せば，クラスタ (p) 内の偏差平方和の合計は

$$E(p) = \sum_{i=1}^{n_p} \sum_{j=1}^{m} (x_{ij}^{(p)} - \overline{x}_j^{(p)})^2 \tag{4.6}$$

となる．いま，クラスタ (p) とクラスタ (q) を融合してクラスタ (r) をつくる．このとき，クラスタ内の平方和の合計の増分を $\Delta E(p,q)$ とおけば

$$\begin{aligned} \Delta E(p,q) &= E(r) - \{E(p) + E(q)\} \\ &= E(p \cup q) - \{E(p) + E(q)\} \end{aligned} \tag{4.7}$$

となる．Ward 法では，クラスタ内平方和が，できるだけ小さいことが望ましいと考え，各段階でクラスタの融合による平方和の増分 $\Delta E(p,q)$ がもっとも小さい (p) と (q) を融合する．そのため，クラスタ (p) と (q) の非類似度 d_{pq} として $\Delta E(p,q)$ を用いる．2 つのクラス

タ (p)，(q) を融合してつくられたクラスタ (r) と，別のクラスタ (s) を融合するときの平方和の増分 $\Delta E(p,q)$ つまり非類似度 d_{rs} は

$$d_{rs} = \frac{n_r n_s}{n_r + n_s} \sum_{j=1}^{m} \{\overline{x}_j^{(r)} - \overline{x}_j^{(s)}\}^2 \tag{4.8}$$

となる．各段階でクラスタのマージによる偏差平方和の増分 $\Delta E(p,q)$ が最小となる p と q をマージする．

4.3　クラスタリングの時間制約

上記のクラスタリングを音声に適用する場合には注意すべきことがある．全フレームに対する距離行列を求め，通常のボトムアップクラスタリングを行なうと，時間的に離れたフレーム（クラスタ）がマージされることが頻繁に起こる．このようなマージ操作は，音声イベントセグメンテーションを求める本タスクにおいては意味を成さないため，時間的に連続する 2 クラスタのみをマージ対象とした**時間制約条件付きクラスタリング**を考える．

ここで非類似度として Ward 法を用いた場合の時間制約の実装方法を述べる．

ある段階でのクラスタを時間軸に沿って $\{p_0, p_1, \ldots, p_t, \ldots, p_T\}$ とおくと，(4.2.5) 式は下記となる．

$$\Delta E(p_t, p_{t+1}) = E(p_t \cup p_{t+1}) - E(p_t) - E(p_{t+1}) \tag{4.9}$$

即ち，連続する 2 クラスタに対して，偏差平方和の増分が最小となるクラスタを対象としてマージする．この時，より類似した連続 2 フレームをマージするのではなく，より類似した連続 2 クラスタをマージする点，即ち，フレームの部分系列をより大きな纏まりとして捉える手法である点に注意すべきである．

本手法は，初期フレーム系列長が N の場合，最終的に N 段の階層構造（樹形図）を生成する．この階層構造は N 通りの粒度におけるセグメンテーション結果を提供する．Ward 法では樹形図の縦軸は「偏差平方和の増分」の総和となり，これは，サンプル群を，与えられたコードブックサイズでベクトル量子化した際の量子化歪みに相当する．結局，許容する量子化歪みを与えれば，それに対応する粒度で階層構造を提供し，結果的に，対応する粒度のイベント境界位置を提供する．複数の粒度を考えた場合，それが，音素やモーラなど複数の言語単位の境界に相当するか否かについては，実験的に検討する．

クラスタリング手法の計算コストは初期フレーム系列長 N に対して $O(N^3)$ である．しかし，連続音声を無音区間ごとに区切れば N を小さくすることは十分可能である．

4.4　音素セグメンテーションの予備実験

本節では，教師なしセグメンテーション手法に関連する各分析条件を変化させ，連続音声に対して良好な音素境界検出結果を生む条件を実験的に検討する．

4.4.1　音声データ

使用したデータベースは米語読み上げ音声コーパス TIMIT データベース及び，日本語読み上げ音声コーパス ATR データベースである．音声データのサンプリング周波数は，TIMIT データベースが 16kHz であり，ATR データベースは 20kHz なので，ATR の音声データは全て 16kHz にダウンサンプリングを行なった上で処理する．

また，本章での実験では正解ラベルより音素境界数を算出し，その境界数をクラスタリングによる境界分割の最終境界数として与える．正解ラベルとして，TIMIT データベースでは男女各 2 名，各人 8 発話，計 32 発話 1203 音素を，ATR データベースでは男女各 2 名，各人 10 発話，計 40 発話，音声記号層の 2396 音素を用意した．

4.4.2　評価手法

正解ラベルとセグメンテーション結果との誤差が 30msec(480 サンプル点)，20msec(320 サンプル点)，10msec(160 サンプル点) となる割合により評価する．正解率 (Correct) は下記のように一定誤差以内の区間数 (Detected) と全区間境界数により算出する．

$$\text{Correct} = \frac{\text{Detected}}{\text{2396 or 1203}} \tag{4.10}$$

実際に人手で作成された境界を正解境界，本研究等で自動検出された境界を検出境界と呼ぶ．本章では，以下のアルゴリズムを用いて検出境界が正解であると定義した．

1. $x = 0$, X =4
2. 検出境界から見て，正解境界が (-x,+x) 以内のフレーム誤差までの中にあるか．あれば，それを対応位置としてその正解境界は以後使わない．但し，対応付けは時間の早いもの同士から行なわれる．
3. Step2 をすべての検出境界に対して行なう．
4. 対応位置の個数を x フレーム以内の誤差とし，出力．
5. 対応位置をリセット
6. x++; x<X ならば Step2 へ

ここでは誤差の小さい境界同士を優先していない．また，時間の逆行について考慮を入れていない．時間軸上で先に検出された境界を優先的に評価している．そのため図 4.1 の黒矢印で示したマッチングがあり得る．また，図 4.2 に示した時間の逆行による矛盾もあり得る粗い評価となっている．

4.4.3　変化させる分析条件

以下に列挙する条件を変化させ，最適となる分析条件を洗い出す．

- クラスタリング手法

すべての手法でマージする 2 クラスタは時間的に隣接しているという制約条件を設ける．

1. 最短距離法
2. 最長距離法
3. 群間平均法
4. 重心法
5. Ward 法

- 音響特徴量として用いる Cepstrum
  1. FFT-cepstrum(12 次元)
  2. MEL-cepstrum(12 次元)
- 音響特徴量に Δ - cepstrum を付与するか
  1. Δ 無し (cepstrum12 次元のみ)
  2. Δ 付与 (cepstrum12 次元+Δ12 次元 = 24 次元)
  3. ΔΔ 付与 (cepstrum12 次元+Δ12 次元+ΔΔ12 次元 = 36 次元)

  Δ 項はそれぞれ前後 2 フレーム分，計 5 フレーム間の変化量を用いて (2.2) 式により算出される．
- 音響特徴量にパワーを付与するか
  1. パワー付与
  2. パワーなし

なお，サンプリング周波数は 16kHz，フレーム長は 32msec(512 サンプル点) で固定とした．また，正解音素数を最終クラスタ数として与え，停止条件としている．

4.4.4　実験結果

i)　各クラスタリング手法による精度評価

図 4.3 は ATR データベースに対する，各クラスタリング手法別のセグメンテーション精度を表している．それぞれ，クラスタリング手法以外の実験条件は変化させ，その平均精度を出力している．この結果より，Ward 法が最もよい精度をあげていることがわかる．

他の手法と比較して，Ward 法が最も高いセグメンテーション結果をあげるため、ここからは Ward 法のみを用いて評価する．

ii)　ケプストラムの違いによる精度評価

図 4.4，図 4.5 は各ケプストラムにおける精度の違いを表す．ATR データベース，TIMIT データベースともにメルケプストラムの方が良い結果を生み出している．これから，人の知覚特性を考慮したメルケプストラムを音響特徴量として用いるべきであることがわかる．

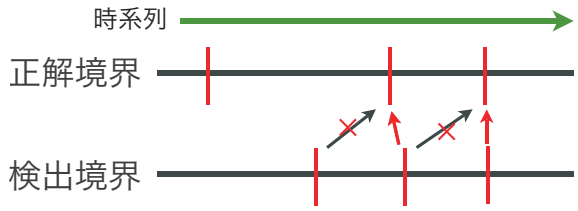


図 4.1: 誤差の小さい境界同士を優先する

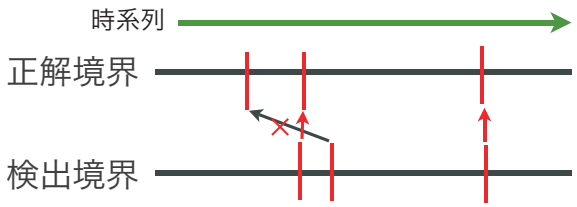


図 4.2: 時間の逆行を許さない

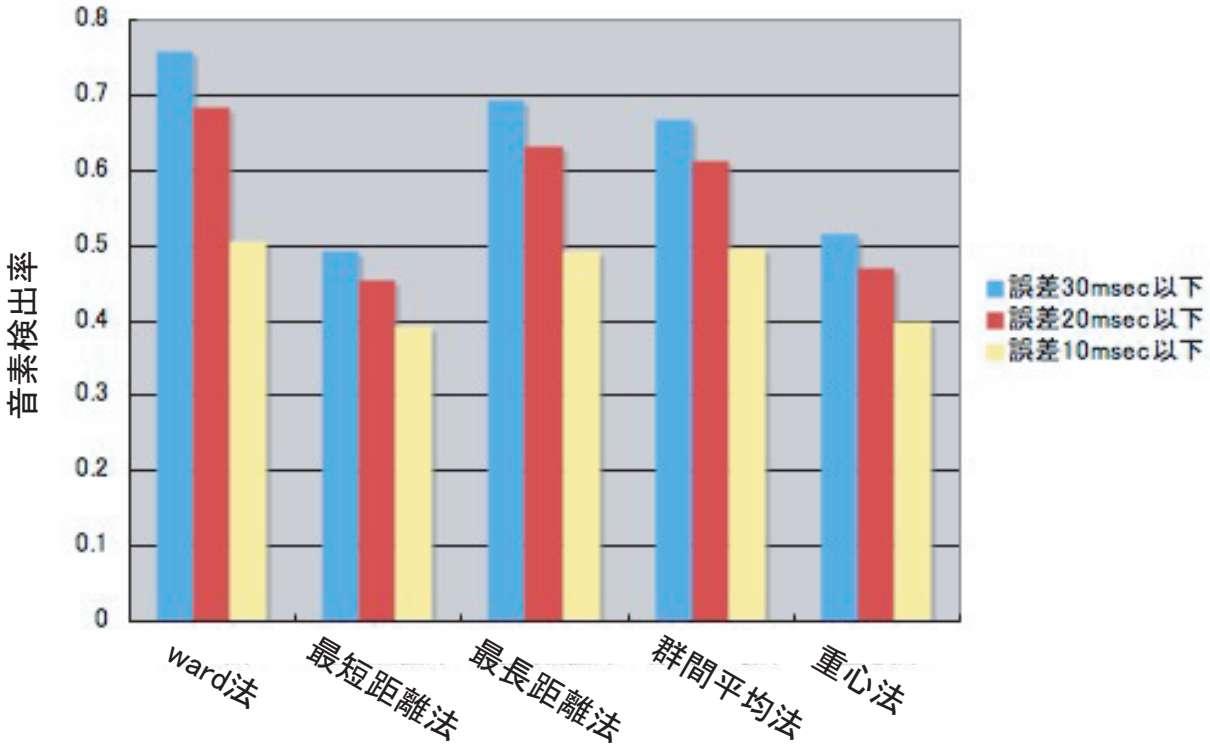


図 4.3: ATR データデースに対する各クラスタリング手法によるセグメンテーション結果

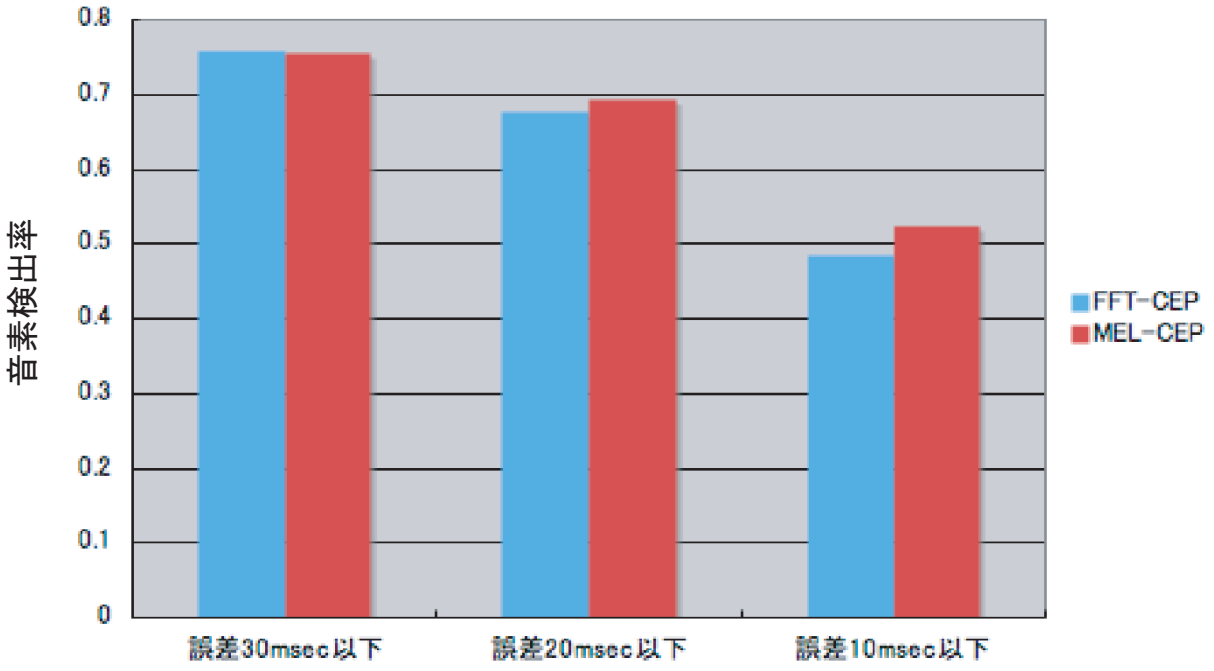


図 4.4: ATR データベースに対する各ケプストラムによるセグメンテーション結果

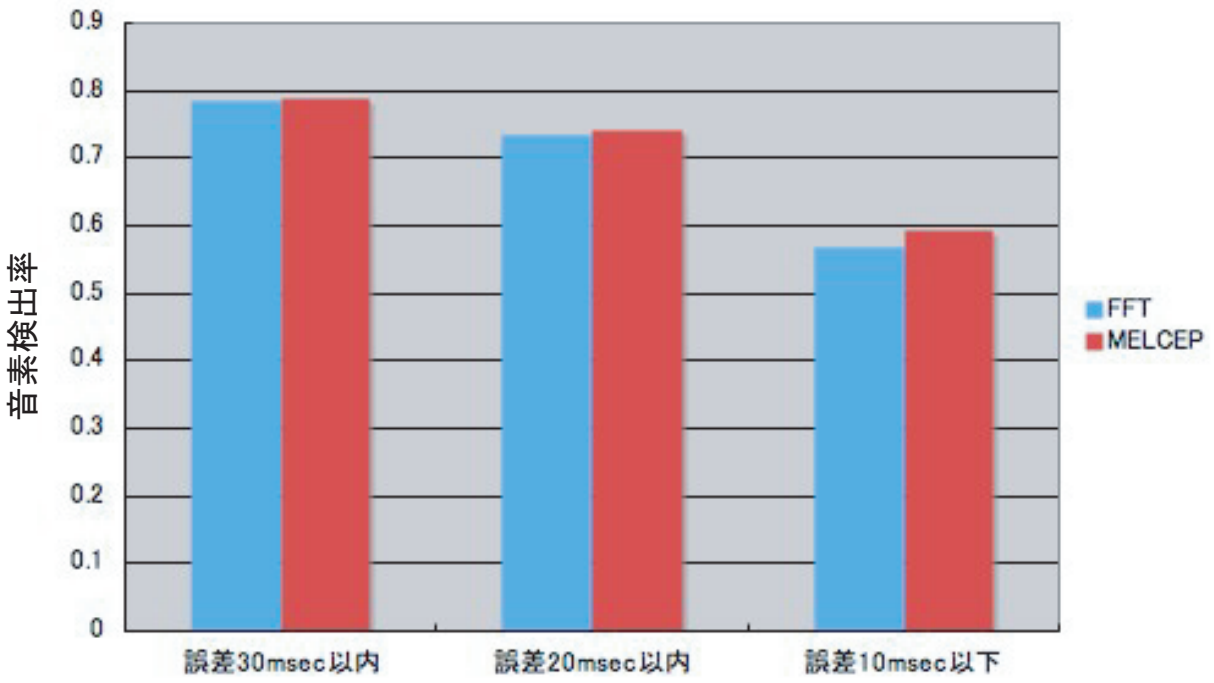


図 4.5: TIMIT データベースに対する各ケプストラムによるセグメンテーション結果

iii)　デルタ項の付与による精度評価

図 4.6，図 4.7 はデルタ項の付加による精度変化を表している．ATR に対してはデルタ項なしが，TIMIT に対してはデルタ項を付与した方が良い結果を出しているため，デルタ項の有無は処理音声に応じて変化させる必要があると考えられる．

iv)　パワーの有無による精度評価

図 4.8，図 4.9 はパワーの有無による結果である．図 4.8 の ATR に関しては，パワーの有無で結果に変化はないが，図 4.9 の TIMIT に対してはパワーを付与した方が高い精度を示している．これよりパワーの付与により多少の精度向上が見られることがわかった．英語には，強勢，弱勢が存在するため，音素間差異にパワー項を用いると有効な結果が得られると考えられる．しかし，日本語は強勢，弱勢が小さく，平坦な言語と言われる．そのため，パワー項を用いることが必ずしも精度向上に寄与するとは限らないと言える．

4.5　音素境界検出とその精度

前節で検討した分析条件を用いて，大規模データベースに対して音素境界検出を行なった結果を報告する．

4.5.1　使用した音声コーパス

2 つの異なる言語の音声コーパスを用意した．一方は米語読み上げ音声コーパス TIMIT の training データベースである[1]．462 人の各 10 発話，計 4,620 発話で構成され，172,460 個の音素境界を有している．他方は，日本語読み上げ音声コーパス ATR データベースの setA，連続発声の dsa パートの女性 3 名，男性 2 名，各 115 発話，計 575 発話を用いた．計 33,607 個の境界を含む音声記号層ラベルデータを使用した．前節同様，ATR の音声データは全て 16kHz にダウンサンプリングした上で処理した．

4.5.2　分析条件

分析条件を表 6.2 にまとめる．スペクトル変化を捉える音響特徴量として，聴覚特性を考慮したメルケプストラムを用いる．ここでは，前節の実験で言語依存性があった，パワー（MCEP の 0 次項）は用いなかった．

また，本節では各発話に含まれる音素数をクラスタ数として与えている．音素数を自動推定し，制約付きクラスタリングを自動的に中止する方法については次節で述べる．

[1]先行研究 [17] が，TIMIT の training データを評価データとしているため，本研究でも評価実験時にこれを用いた．

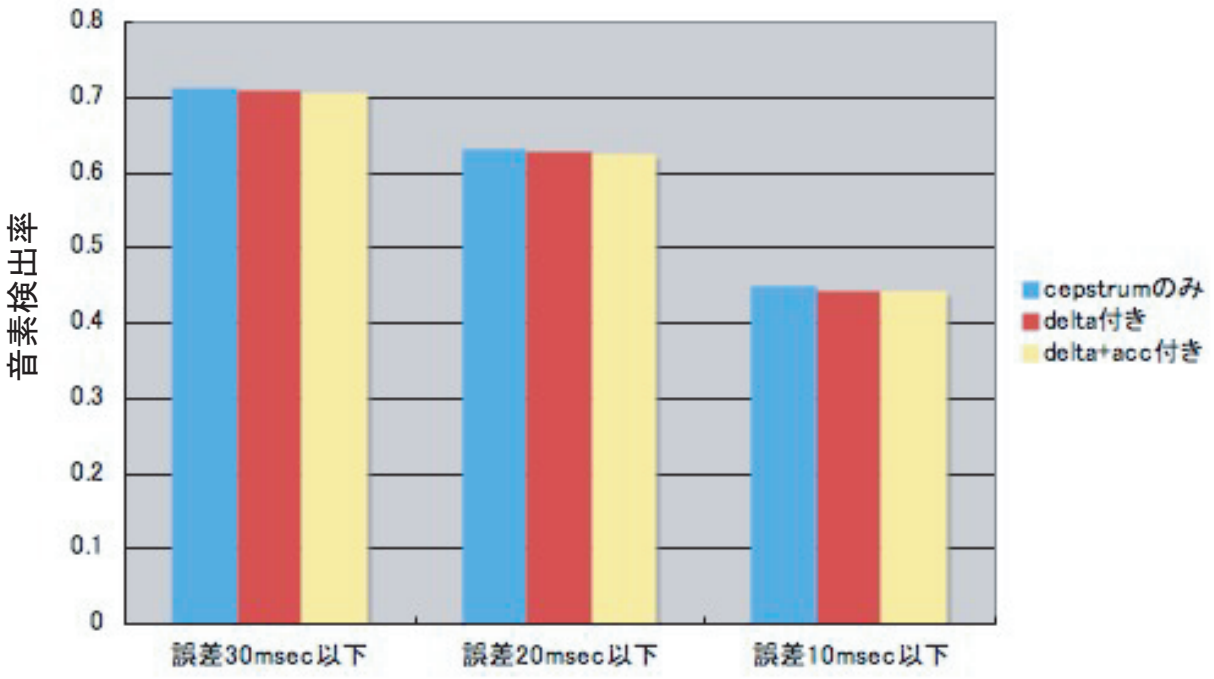


図 4.6: ATR データベースに対する各デルタケプストラムの付与によるセグメンテーション結果

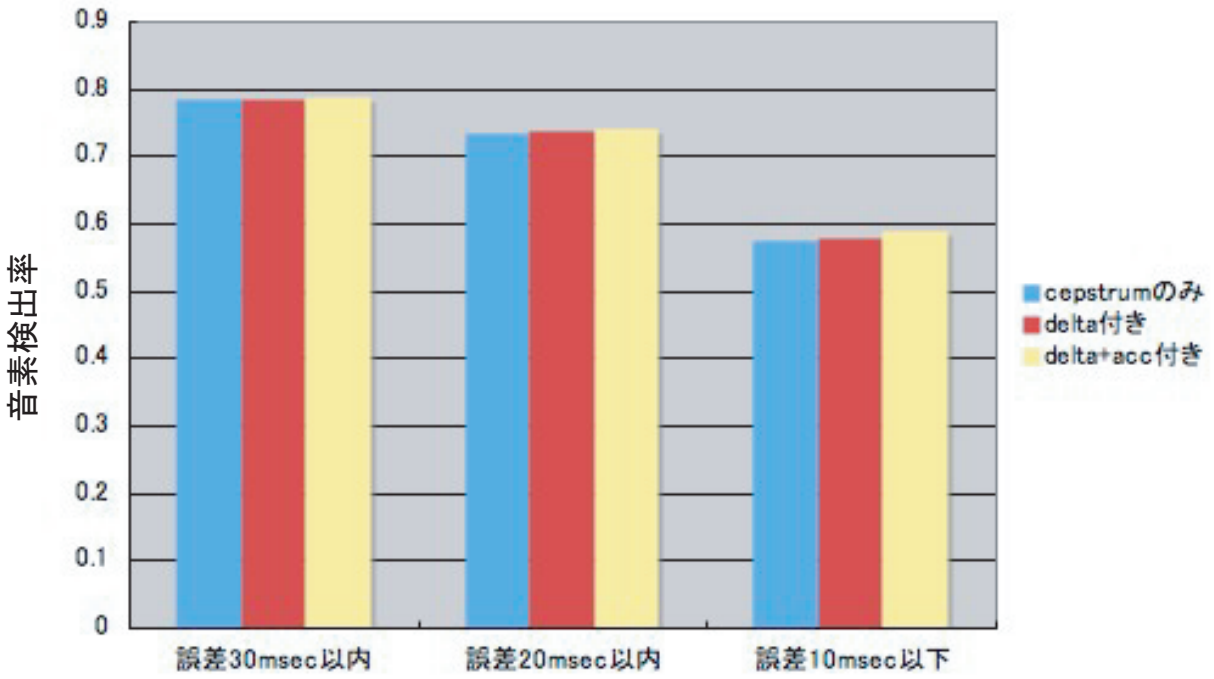


図 4.7: TIMIT データベースに対する各デルタケプストラムの付与によるセグメンテーション結果

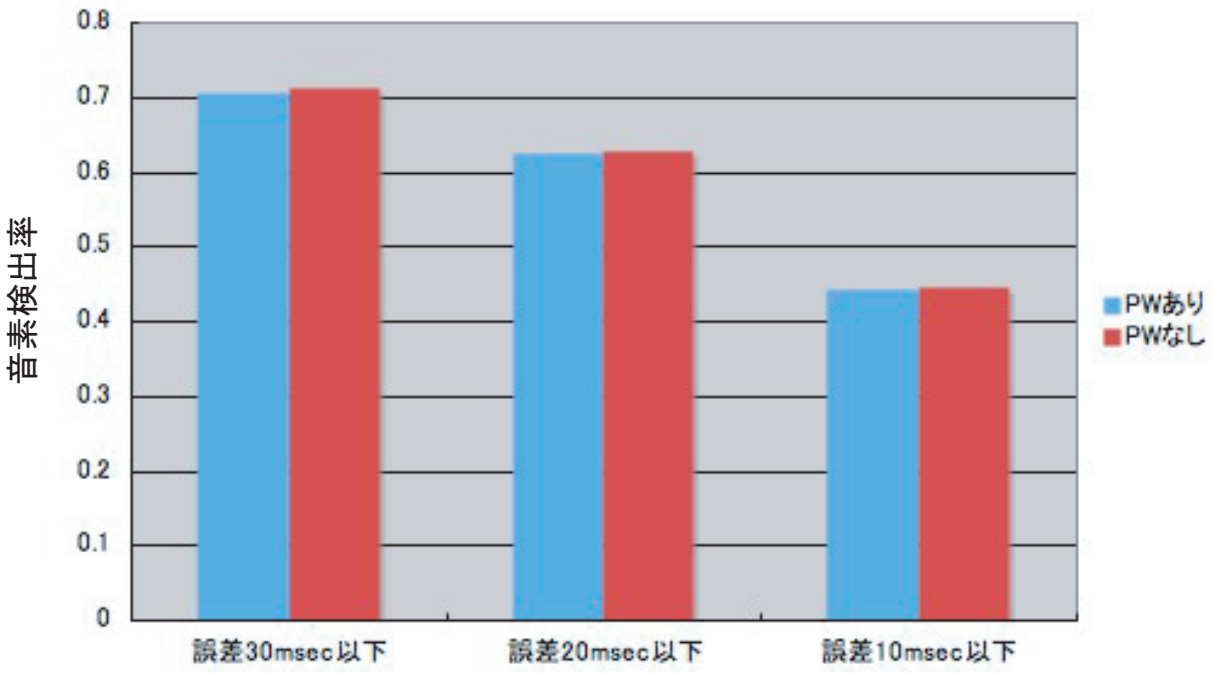


図 4.8: ATR データベースに対するパワーの付与によるセグメンテーション結果

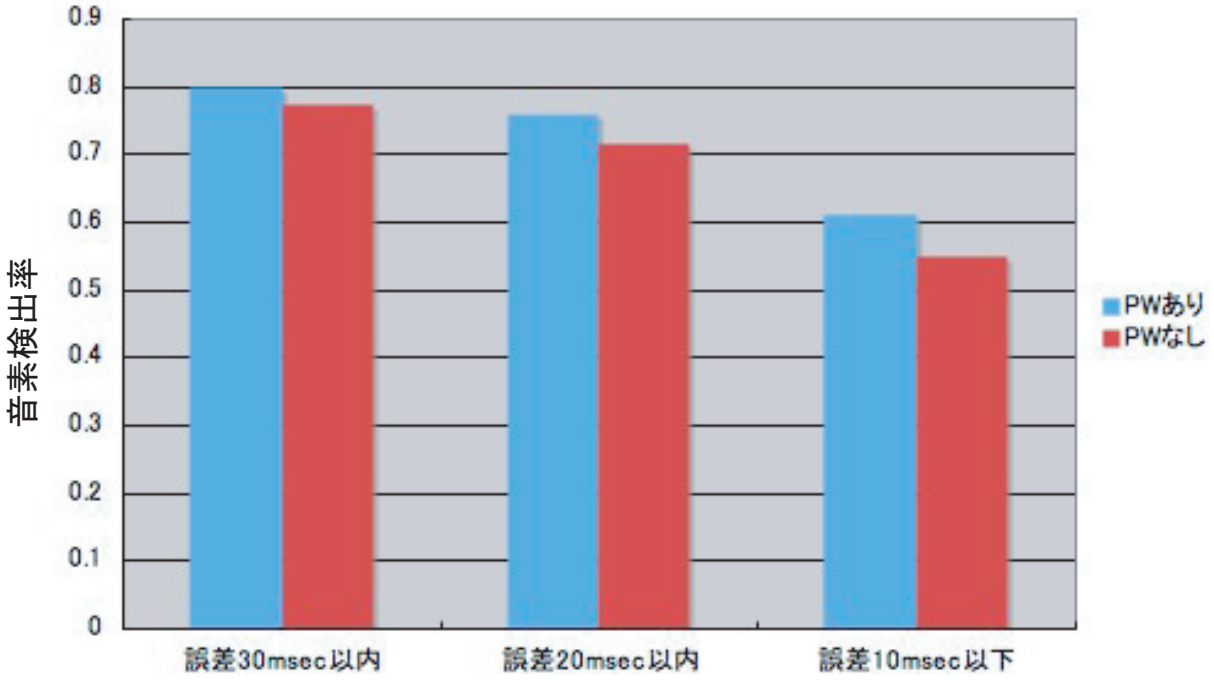


図 4.9: TIMIT データベースに対するパワーの付与によるセグメンテーション結果

表 4.1: 音響分析条件

サンプリング	16bit / 16kHz
フレーム幅	32msec
窓	ハミング窓 32msec
フレームシフト長	10msec
音響パラメータ	MCEP 1〜12 次元
帯域	fullband

4.5.3　評価方法

前節の評価では，誤差の小さい境界同士を優先していないこと，そして，時間の逆行について考慮に入れていないことから，正確な評価がなされていなかった可能性がある．本節からは，それら 2 つの問題を解決するように下記のアルゴリズムによって評価を行なう．

1. $x = 0$, X =4
2. 正解境界から見て，検出境界が (-x,+x) 以内のフレーム誤差までの中にあるか，かつ，時間の逆行を超えた対応づけではない．このような境界があれば，それを対応位置としてその正解境界は以後使わない．
3. Step2 をすべての検出境界に対して行なう．
4. 対応位置の個数を x フレーム以内の誤差とし，出力．
5. x++; x<X ならば Step2 へ

4.5.4　音素境界の自動検出とその精度

TIMIT データベースに対して，自動で音素境界を検出した結果の一例を図 4.10 に示す．原波形，スペクトログラム，自動検出結果及び手動でのラベリング結果を併記する．本手法の特徴として/k/や/t/等の破裂音に伴う休止区間/kcl/，/tcl/，またはポーズは検出されやすい．逆に母音が二つ続いた際はその 2 母音間の境界を検出しにくい[2]．

図 4.11 は TIMIT データベースと ATR データベースの全データに対して（音素境界数を与えて）音素検出処理を行なった結果である．自動で検出された境界（以後，検出境界と呼ぶ）とラベラーにより定められた境界（以後，正解境界と呼ぶ）との絶対時間誤差の程度に応じた検出率により評価を行なっている．共に 20msec 以内の誤差で 70%，30msec 以内の誤差で 75%程度の精度となっている．また，TIMIT データベースより ATR データベースの方が，検出精度は高い．ATR データベースのラベリングはスペクトログラム特徴に基づいて行なわれるが，境界を決定できない区間は無理に分割を行なっていない．ATR データベースの方が，纏まりを捉える本手法と，より近い処理をしていると考えられる．

[2]英語の二重母音（diphthong）を考えれば，この方がより正しいラベリングを行なっていると解釈することもできる．

4.6　まとめ

本章では，時間制約を設けたボトムアップクラスタリングによる自動セグメンテーション手法を提案した．さらに，本手法と音素検出精度という面で相性の良い分析条件を実験的に明らかにした．その結果，非類似度として Ward 法を，音響特徴量にはメルケプストラムを用い，英語に対するセグメンテーションの場合パワーを使用することで精度が高くなることが分かった．デルタケプストラムや日本語に対するパワーの使用により，精度に大きな変化は見られなかった．しかしこれは，音響特徴量として全て同じ比重で用いた結果である．主成分分析を行なうことにより，さらによい精度が得られる可能性はある．

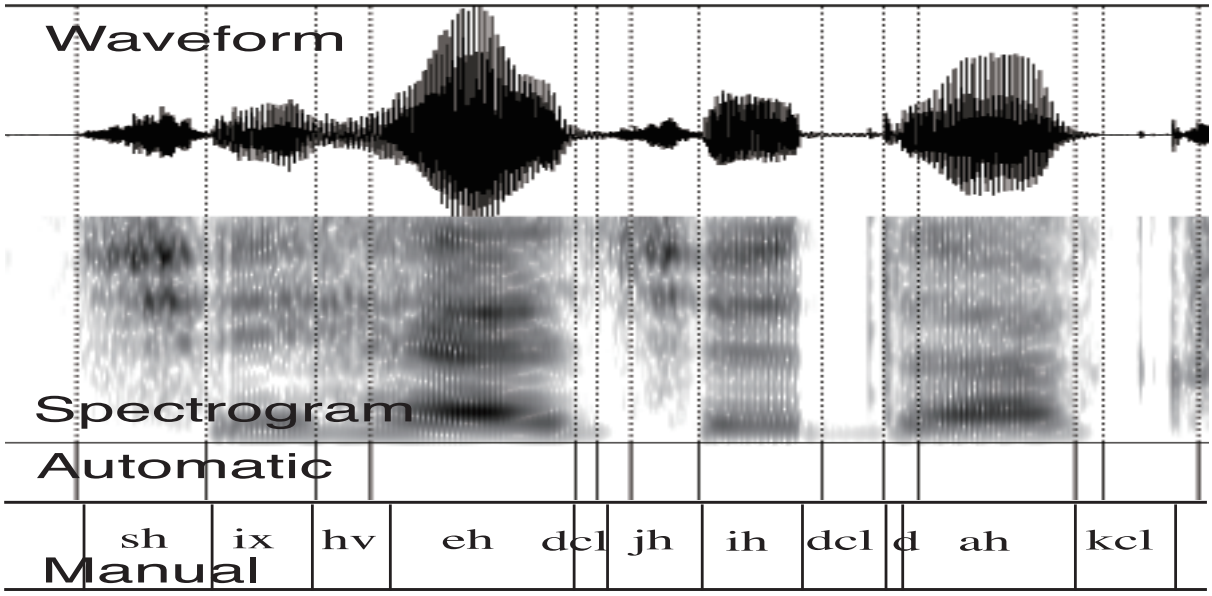


図 4.10: 自動セグメンテーション結果の一例

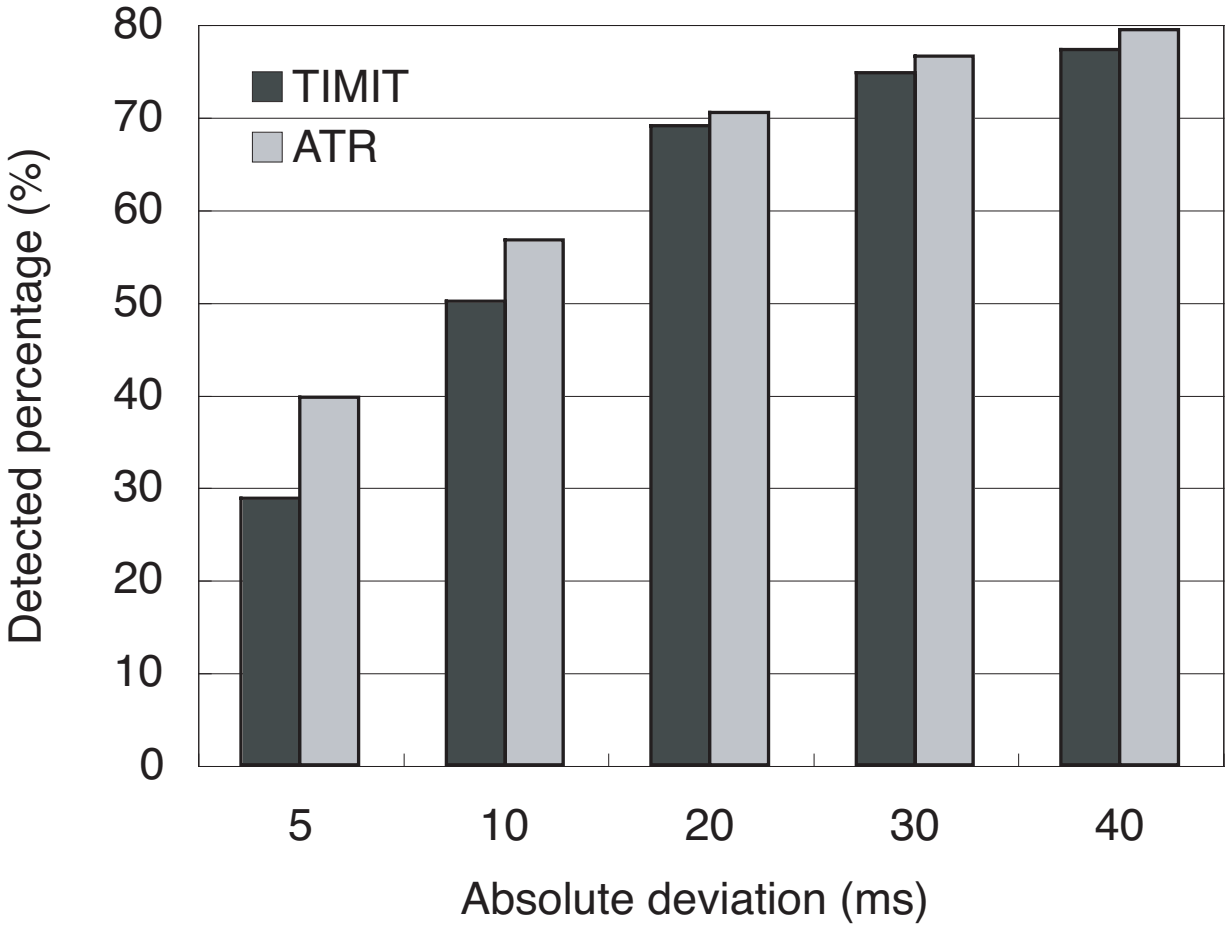


図 4.11: 自動セグメンテーション結果

第5章

イベント境界検出の様々な検討

5.1　閾値処理によるクラスタリングの自動停止

前章では，発声に含まれる音素数を既知として与えていた．しかし，HMM による音素セグメンテーションに代表される事前知識を用いる手法と頑健性で明確に優位に立つためには，音素数をも自動推定することが望まれる．また，事前学習を必要としない従来手法である MSTP 法は音素数を自動決定するので，精度比較実験を行なうためには何らかの方法論によりクラスタリングを停止した上で，境界検出を行なわなければならない．

そこで本節では，音素数を自動的に推定する手法について述べ，その自動推定結果の妥当性を実験的に示す．

5.1.1　コスト関数制限下でのクラスタリング

提案手法が基本的に階層的ボトムアップクラスタリングであることを考慮し，Ward 法の更新コスト（群内偏差平方和の増加量）に着眼することで実現する．以下の考察から，閾値処理により音素数を自動で決定すること，即ち，クラスタリングを自動停止することを検討する．

Ward 法では，クラスタ p，q に対して (4.2.5) 式の距離尺度を定義している．各段階で 2 クラスタのマージによる群内偏差平方和の増分 $\Delta E(p,q)$ が，最小となるクラスタをマージしようとする．ここで，ある段階において，各クラスタが凡そ各音素に対応している状態を考える．この場合次の操作で，異なる音素が強引にマージされることになる[1]．ある話者が生成する各音素の音響特徴量を考える．この場合，任意の 2 音素間距離（重心間距離）の最小値は，凡そ話者非依存であると仮定する．その結果，どの話者の発声した音声であっても，音素に対応した形でクラスタリングされた状態に対する次のマージ操作は，比較的大きな更新コスト（群内偏差平方和の増加量）を呈するはずである（$E(p \cup q) >> E(p), E(q)$）．ただし，以上の議論はパワー項を除いた，ケプストラム領域でのみ成り立つ．

図 5.1 は Ward 法の木構造の高さの増分を表す．図 5.1 の左図は正解境界 65 の男性音声をクラスタリングしたときの更新コストの増分であり，右図は正解境界数 36 の女性の音声データを同様に処理した結果である．正解境界数に赤線を引いている．これを見ると，共に増分が 0.2 程度でその正解境界数を向かえていることがわかる．

5.1.2　閾値の実験的検討

$\Delta E(p,q)$ に対応する閾値を実験的に定め，これを用いてクラスタリングの自動停止を検討する．このような手法はスペクトル遷移を局所的に走査する先行研究 [17, 18] では困難な方法論であると考えられる．局所的に定義されるスペクトル差は，閾値値を決定しにくいと考察されるためである．

[1]時間的に連続する 2 クラスタのみをマージ対象としていることに注意．時間的に離れた 2 クラスタを対象とすれば，非常に類似性の高い 2 クラスタ（例えば同一 2 音素）がマージ対象となる．

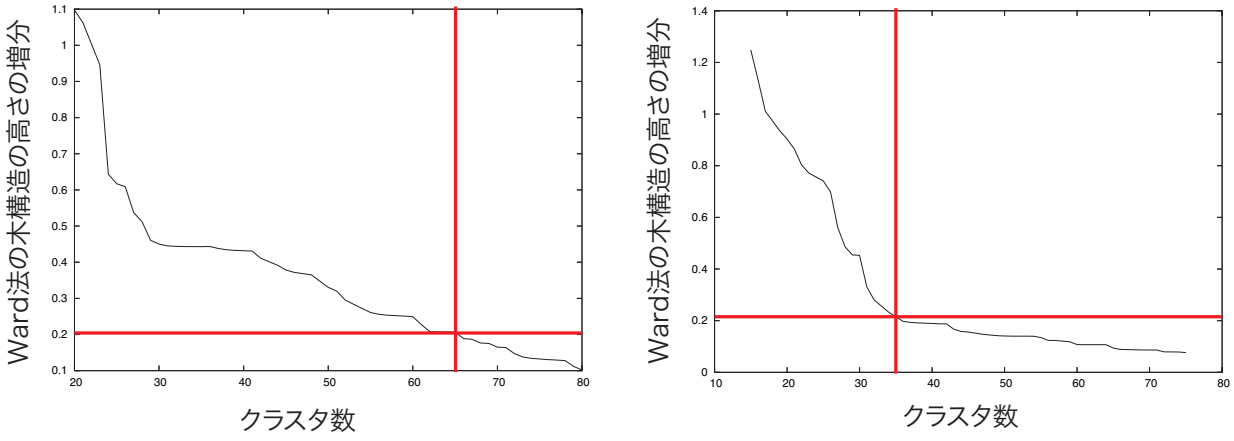


図 5.1: Ward 法の木構造の高さの増分

5.1.3　評価尺度

音素境界検出の評価尺度として F 値を用いる．F 値とは Precision(適合率) と Recall(再現率) により

$$\mathrm{Precision} = \frac{\mathrm{Detected}}{\mathrm{Detected} + \mathrm{Inserted}} \tag{5.1}$$

$$\mathrm{Recall} = \frac{\mathrm{Detected}}{\mathrm{Detected} + \mathrm{Missed}} \tag{5.2}$$

$$\mathrm{F-Measures} = \frac{\mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \times 2 \tag{5.3}$$

として表される尺度である．ただし，Detected とはシステムにより境界と判断され実際に境界があった総数，つまり正解数であり，Inserted とはシステムが境界と判断しながらも実際には境界がなかった誤検出数，Missed は正解境界が存在するがシステムにより境界として検出されなかった未検出数を表す．もし検出すべきクラスタ数を正解境界数と同等と設定すると，Detected + Missed = Detected + Inserted となる．よって全正解境界の内，正解を検出できた割合を x とおくと，

$$\mathrm{x} = \mathrm{Recall} = \mathrm{Precision} = \mathrm{F-Measures} \tag{5.4}$$

となる．Precision + Recall =(一定)　であれば Precision = Recall の時 F 値は最大値をとるが，クラスタ数を与えればその限りである．よって，クラスタ数を既知として与えた場合 (クラスタ数 = 正解境界数) の F 値と同等の F 値を示すクラスタ自動決定手法が存在すれば，それは非常に妥当性の高い手法だと言える．

5.1.4　閾値の決定

Fig. 5.2 は TIMIT の test データベースから無作為に 30 発話を選択し，$\Delta E(p,q)$ に対する閾値 K に対する絶対誤差 30msec 以内の F 値を求めた結果である．図より $K = 0.22$ の

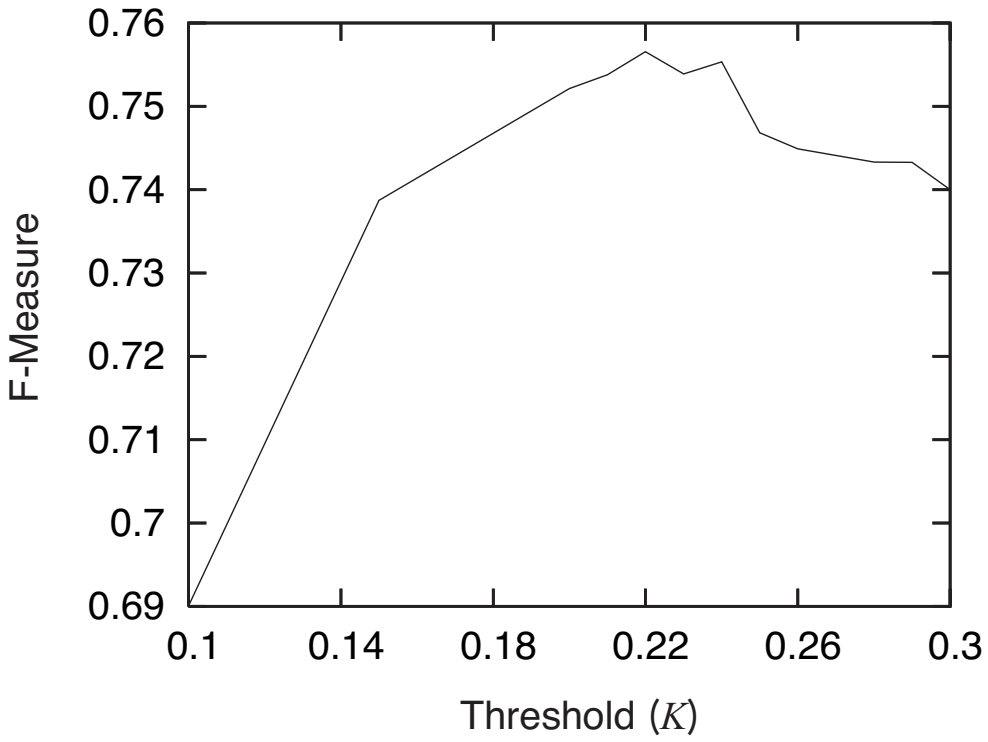


図 5.2: 閾値 K に対する F 値

時，最大値（0.757）となった．ATR データも含め，閾値 $K = 0.22$ を用いてボトムクラスタリングを自動停止させる．

5.1.5　提案する音素数自動推定手法の妥当性の検討

open データである TIMIT の training データの各発声に対して音素数推定実験を行なった．自動推定音素数と正解音素数の関係を図 5.3 に示す．自動推定音素数と正解音素数の相関係数は 0.84 となり，強い相関が確認できた．しかし，音素数推定という観点から見ると，正解音素境界数 50 に対して，検出境界数は 40 から 65 ほどの幅があり，検出漏れや過検出も多い．もし「乳児が泣き叫ぶ音声，笑い転げる音声」などを対象とした場合，明確な音素書き起こしが単に与えられた音声に対するラベラーの音韻化想像力を示すのみとなり，意味を持たないことも多い．これらの音声を対象とした音声区分化ツールを構築する場合，クラスタリングの更新コストに対する閾値をスライダー指定する GUI を用意し，ラベラーが検出境界の是非を判断しながらラベリング初期値を決定するような場面では，応用可能性がある．

図 5.4 は閾値処理によりクラスタリングを自動停止させた場合の，音素境界検出精度と，正解音素数を与えた場合の境界検出精度を F 値で比較した結果である．正解音素数を与えたとき，F 値は Precision(適合率) 及び Recall(再現率) と一致する．クラスタ数を自動推定しても音素数を与えた場合とほぼ同等の結果を出していることが分かる．これらの結果か

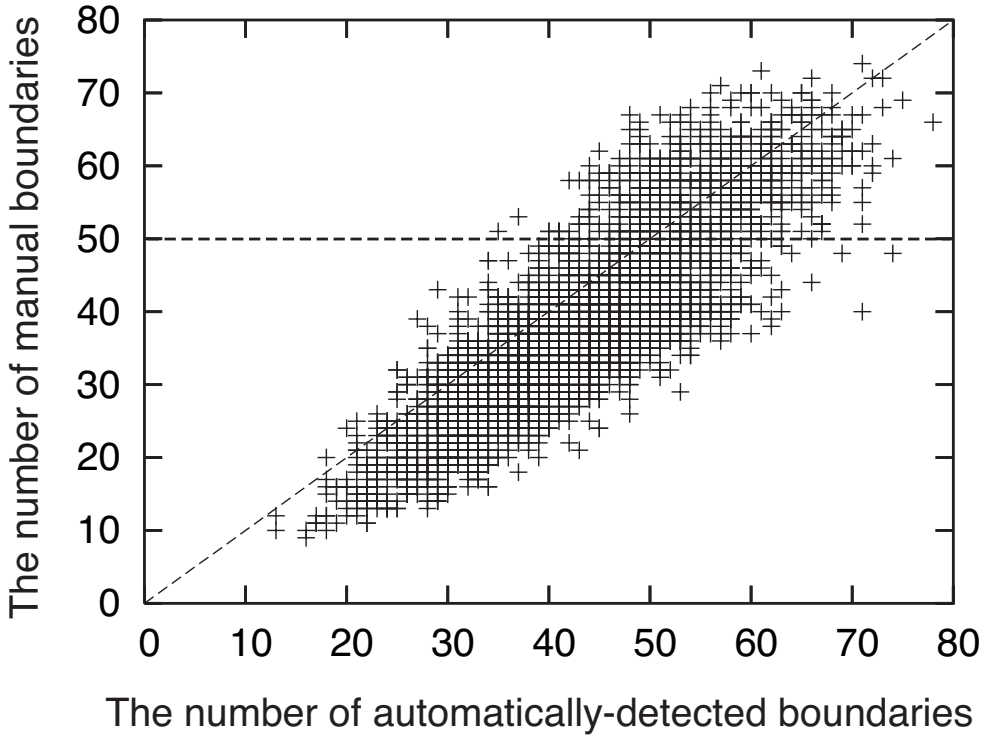


図 5.3: 自動推定音素数と正解音素数との関係

ら，閾値処理を用いることで，F 値ベースの評価においては事前知識を用いずとも音素数を与えた場合に近い出力が可能であり，提案手法は妥当性の高い手法であると言える．よって，この閾値をマニュアル操作しながら適切な音声区分化を行なう際の初期セグメンテーションを求めることは可能であると考えられる．

また，同一の閾値を異なる言語において用いたが，相応の結果であるため，定めた閾値は高い言語非依存性をもつことがいえる．

5.2　先行研究との比較実験及び考察

本節では，音響分析に基づく（事前学習を閾値設定以外では必要としない）先行研究である，MSTP 法と，同じデータベースに対して比較実験を行なうことで提案手法の有効性を確認する．

評価データベースは 4.5.1 節で用いたものと同様で，TIMIT データベース，ATR データベースである．比較実験結果を図 5.5 に示す．2 つのデータベース共に提案手法により精度が改善されている．特に ATR データベースに関しては，MSTP 法で精度が大きくて低下している．これには MSTP 法の後処理（特に閾値）が TIMIT データベースに依存していることを強く示唆する結果である．それとは対照的に，本手法で提案した閾値は TIMIT データを参照して求められたにも拘らず，ATR データデースに対して高い精度を呈してい

表 5.1: 各粒度のイベントラベルの境界数

粒度	境界数（個）
モーラ	20,621
音声記号層	33,607
TIMIT level	39,356
イベント層	46,212

る．本稿で提案する手法の言語依存性の低さを示していると考えている．

また，直接同じデータベースでの比較実験は行なっていないが，SVM による検出手法では音素検出率約 75%という結果が報告されている．しかしこの手法は音素数を自動で決定することが困難であるため，音声処理の前処理としての役割を考えた上で利用しづらい手法であると言える．

なお，音素数のみならず，個々の音素の種類，及び各音素の音響的特徴を，大規模音声データベースから計算される統計量としてシステムに与えた場合（即ち，HMM による強制アライメント），種々の後処理の導入により，20msec の誤差で，約 90%の精度が報告されている [10, 11, 16]．音響分析に基づく音素境界抽出は，先行研究を含め，通常の読み上げ音声を対象とすれば，HMM による強制切り出しに基づく方法論とは比較できないほど精度は悪い．音響モデルの学習データとその音響的特性が大きく外れた音声（例えば乳児が泣き叫ぶ音声，笑い転げる音声）などを本来は対象とすべき手法であろう．

5.3　様々な言語単位に対する境界検出

第 4 章で述べたように提案手法はボトムアップクラスタリングに基づいており，段階的な階層クラスタリングが結果として得られる．即ち，異なる粒度の音声区分化が得られる．本節では，ATR データベースに付属する複数の粒度の正解ラベルと，提案手法で得られる自動境界検出結果との比較を行なう．

5.3.1　様々な粒度の音声イベントラベル

ATR データベースに付属する複数粒度の正解ラベルを参照して，イベント層ラベル，TIMIT-level ラベル，音声記号層ラベル，モーララベルを定義した．各粒度に対して含まれる境界数を表 5.1 に表記する．音声記号層ラベル，イベント層ラベルは ATR に元から付属されているラベルである．前節までの実験では音声記号層ラベルのみを用いていた．モーララベルとは，音声記号層をモーラ単位に纏めて作成したラベルである．但し，音声記号層で区分化が不可能と判断されたラベルはそのままにしている[2]．TIMIT-level ラベルとは，イベント層を TIMIT のラベリングの粒度と同様になるよう，纏めたラベルである．{ > }，{ < }，{*> } 等の母音過渡区間を各母音に吸収させて生成した．

[2]{k,a,u} や {s,u} 等の表記部分であり，専門家がスペクトログラム上で分割できなかった区間である．

5.3.2　様々な粒度に対するイベント境界の検出実験

実験は ATR データベースの各イベントラベルに対して，正解境界数を既知として行なった．図 5.6 はラベル粒度を変化させた時の検出結果である．この結果より，音素単位もしくはそれよりも細かい粒度に対する境界推定は可能であるが，より大きいモーラ単位でのセグメンテーションでは精度が落ちる結果となった．その理由として，音素単位以上でのクラスタリングでは母音同士，または，母音+半母音+母音など，スペクトル形状の近い音素同士のマージが優先され，/k/や/t/等の破裂音に伴う休止区間等が最後まで残されてしまうことが理由の一つであると考察される．しかし，例えば英語音声学の中には/aʊɚ/（tower の t 以降の発声）を三重母音として認める考え方があるなど，自動ラベリング結果が必ずしも誤りであると言い切れない場合もある．そもそも日本人であれば英語の二重母音は，二つの異なる母音連鎖と感覚するのに対し，英語母語話者は一つの母音として感覚する．このような事実を考えれば，絶対的な正解ラベルを用意することの是非すら問われかねない．既存のデータベースに付与されている「ある知覚特性を持った人間が恣意的に付与したラベル」を正解とした評価よりも，本手法が呈する境界情報の利用価値を認める応用場面を模索した方が，意味のある評価となるのかもしれない．

しかし，本研究で検討した分析条件を変更することで，モーラ境界検出の精度を上げることも可能であろう．例えば，最終的に得られる境界に対して，境界間時間長の最小値を設ける，各クラスタの音声事象が有する全エネルギーをある範囲内に収める，など種々の実験的検討が可能である．或いは，音素レベルまでクラスタリングを行なった後，パワーの大小により母音を表すクラスタをまず特定する．日本語においてはモーラに含まれる子音は 1 つであるから，母音と推定された各クラスタの前のクラスタがもし母音と推定されればばマージせず，母音でないと推定された場合は対応づけるといった後処理によりモーラ境界を検出することも可能であろう．

5.4　各音素の境界検出精度

Fig. 5.7 は，ATR データベースに対する各音素間接続の検出実験結果である．モーラ構造をもつ日本語において，摩擦音や破擦音等の母音と大きく異なるスペクトル構造 (白色雑音性) をもつ音素は，その安定区間が母音と明確に異なるクラスタとして纏め上げられる傾向にあるため，境界が検出されやすいことがわかる．また，本実験ではパワーを特徴量として用いていないものの，破裂音に伴う休止区間 (/kcl/，/tcl/等) や，ポーズ等の無音区間も同様の理由から比較的検出されやすい．逆に，流音は顕著に検出精度が低くなっていることが実験的に検証された．

興味深い事に，流音と撥音以外の音素では，(母音 or ポーズ) → (子音) という音素接続により精度が劣化している．前節で音素単位以上に纏め上げを継続してもモーラ単位での境界検出は困難であると述べたが，このことからその理由の一端が説明されるに至った．

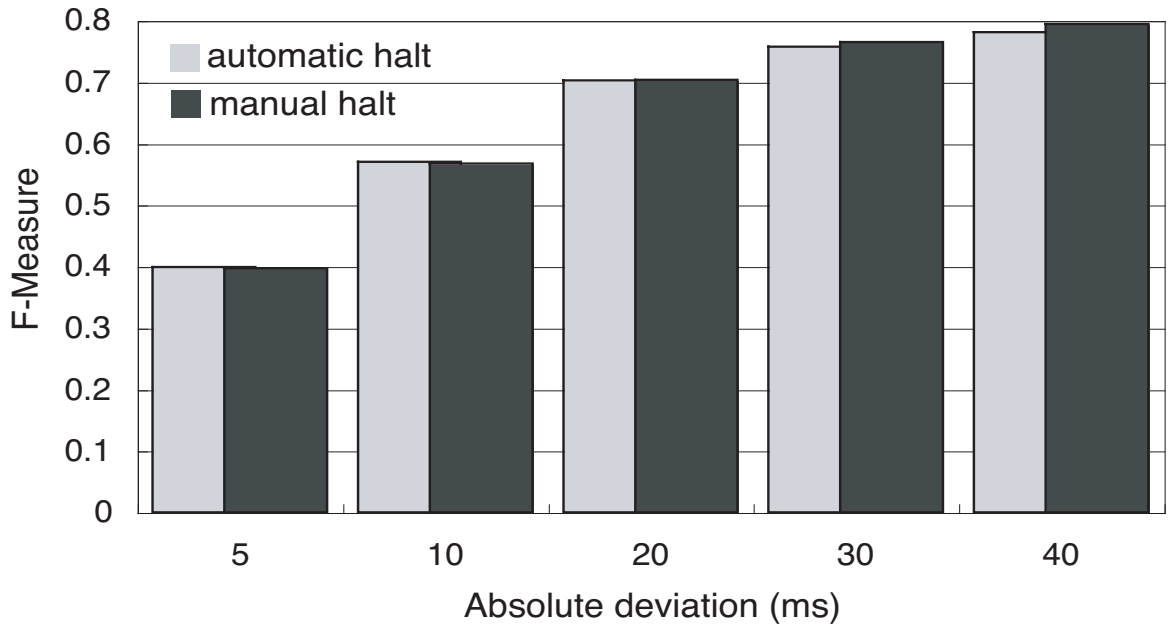
5.5　まとめ

本章では，ボトムアップクラスタリングによるイベント境界検出の性質を明らかにするために，様々な実験を行なった．

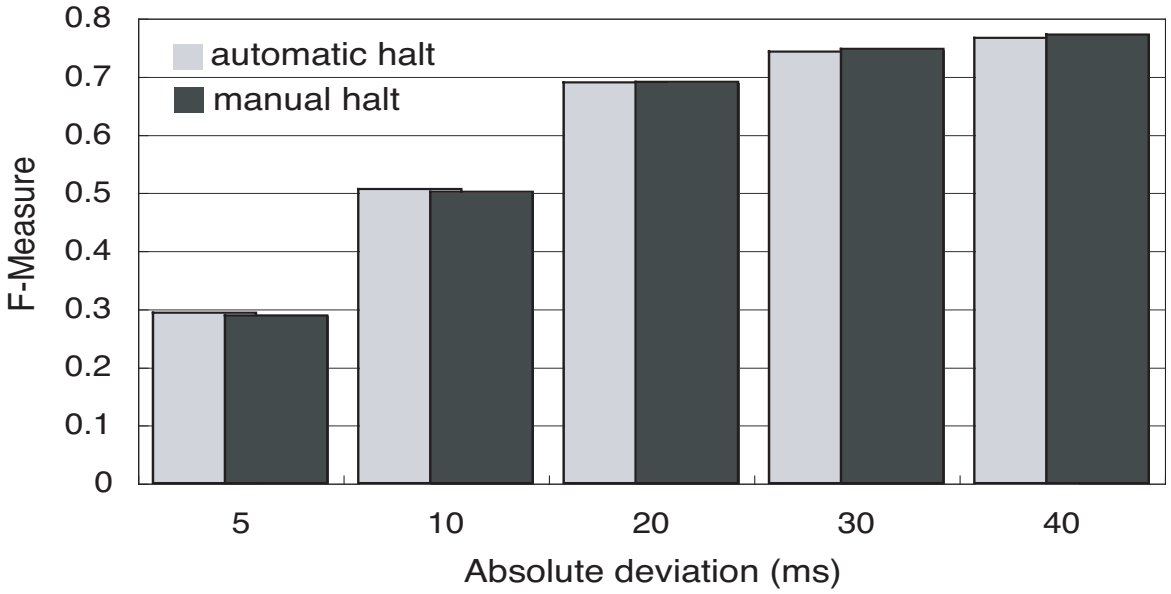
まずは，教師無し音素セグメンテーション手法として非常に重要な事柄である，音素数の自動推定手法について述べた．音素数を既知として与えずとも，適切な閾値処理により音素を自動で決定することで，精度がほとんど低下することなく音素セグメンテーションを実現することが可能となった．これは，先行研究の音素セグメンテーション手法では実現できなかったことである．

更には，自動音素推定手法を用いつつ，音響分析のみに基づく先行研究と比較実験を行なった．この結果より，MSTP 法より精度，頑健性の面で優位にたつことが示された．

そして，種々の言語単位でのセグメンテーション結果，各音素毎の検出精度について述べ，ボトムアップクラスタリングの利点，限界について言及した．人が境界数を手動で与えた場合にその数に応じて意味のある境界やクラスタが得られる点は大きな利点であると考えている．本研究では詳しく実験を行なっていないが，時間的制約条件を無くしクラスタリングを行なっていくと，最終クラスタ数を 3 クラスタに設定すると「有声音」＋「無声音」＋「無音」に，最終クラスタ数が 2 クラスタの時は「有声音」＋「無声音＋無音」というセグメンテーションができることが定性的に示されている．

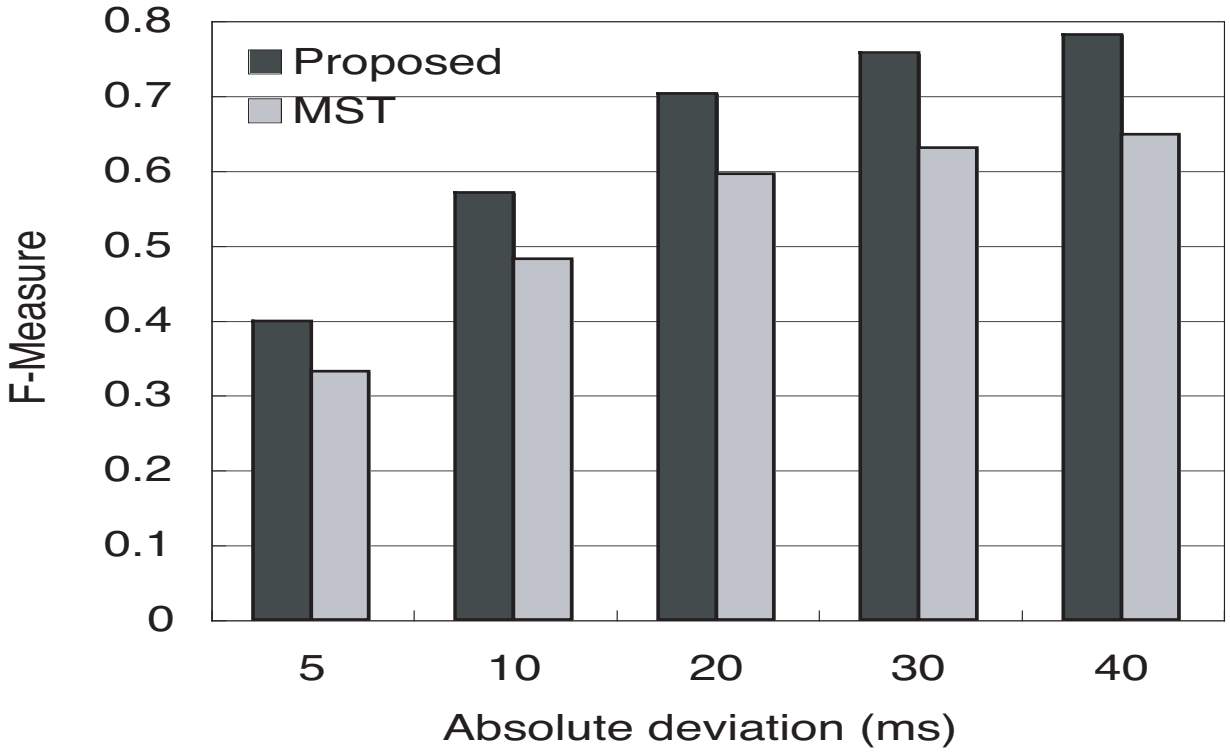


(a) ATR

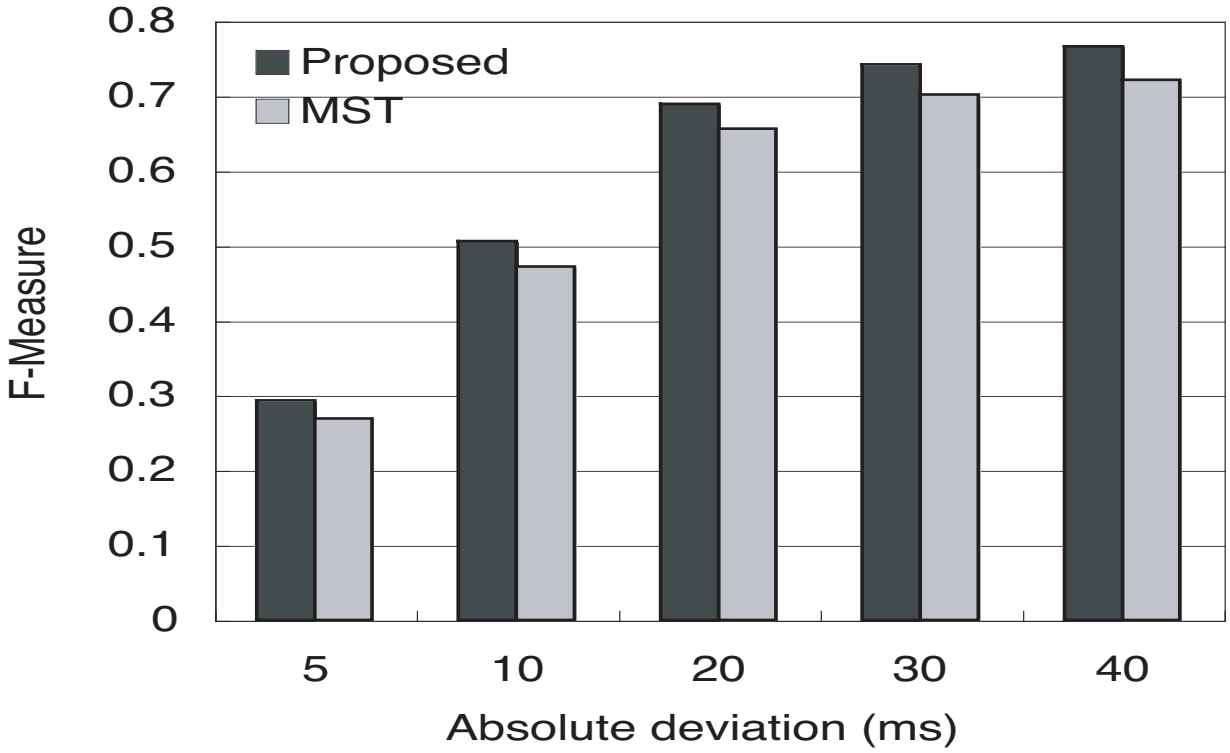


(b) TIMIT

図 5.4: 音素境界を自動停止した場合と，音素境界数を与えた場合の比較



(a) ATR



(b) TIMIT

図 5.5: MSTP 法と提案手法の比較実験

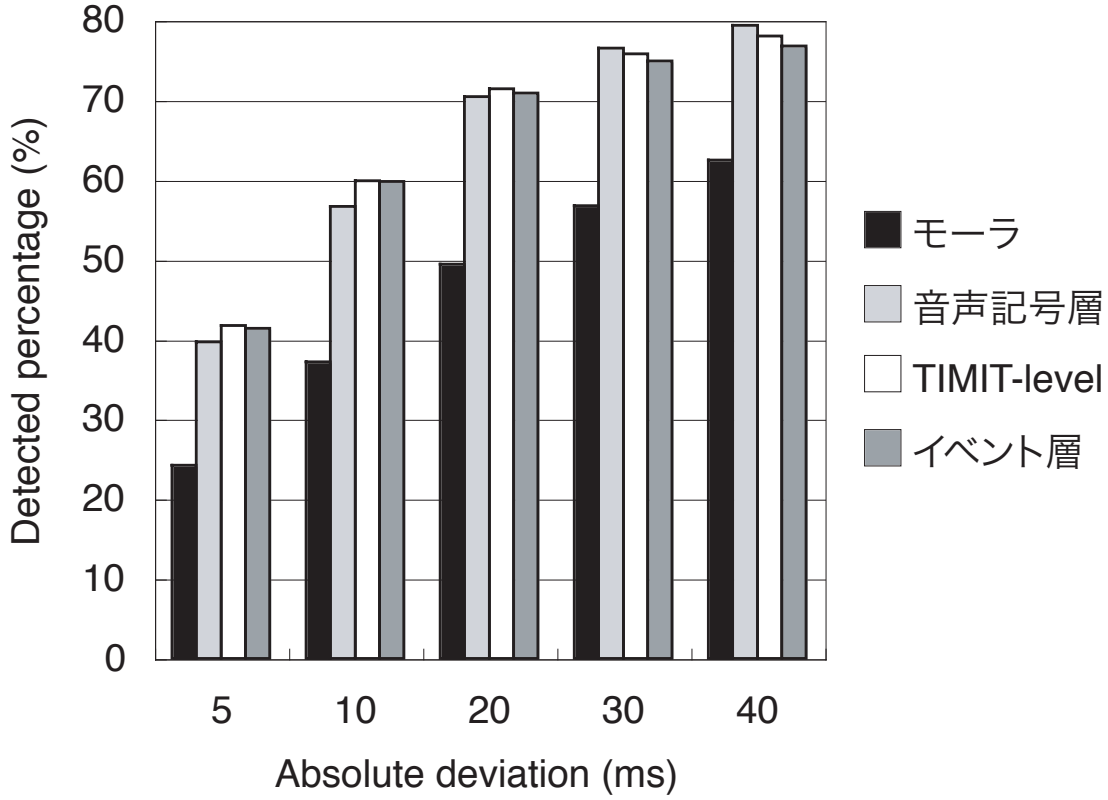


図 5.6: 各粒度のイベント境界に対する境界検出結果

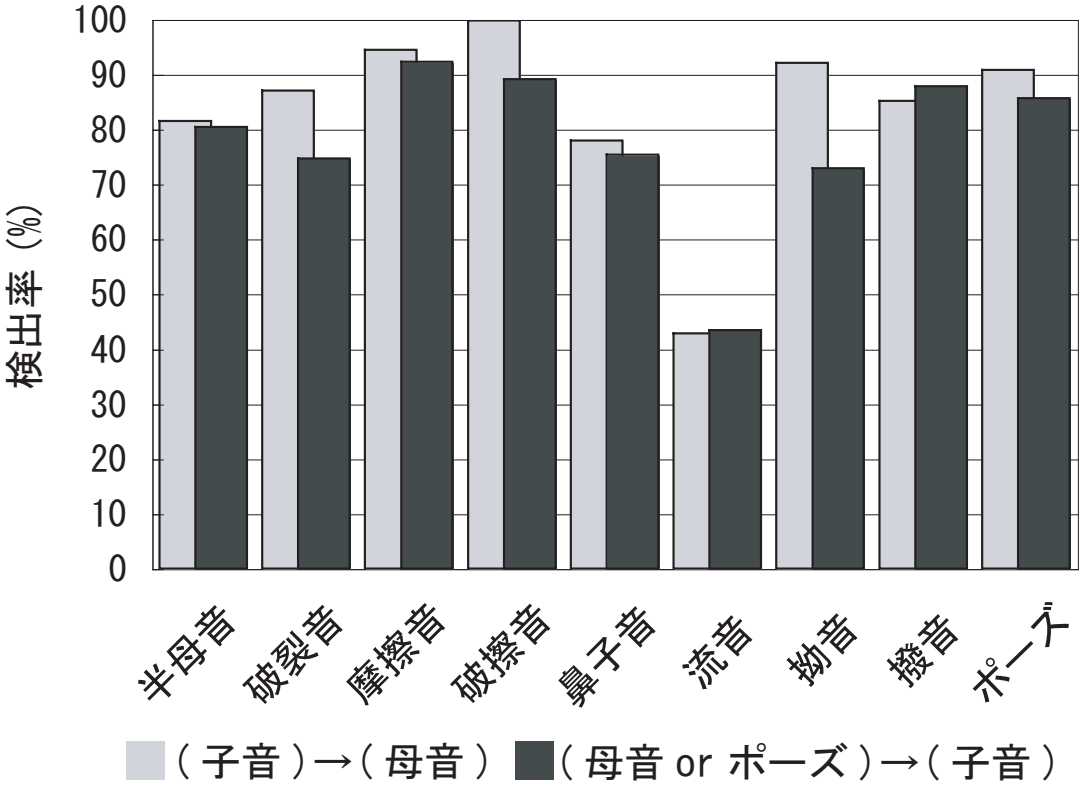


図 5.7: 日本語の各音素接続間の検出率

第6章

シャドーイング音声の自動評価に関する実験的検討

6.1　はじめに

前章までに検討してきた教師なしセグメンテーション手法は，非常に頑健な処理が可能であるが故に，崩れた音声の処理に適した手法である．そこで，一般に歪んだ音声となることが多いシャドーイング音声への応用を検討した．本章ではシャドーイングの自動評価方法について述べる．

6.2　シャドーイング学習とその特徴

近年，言語教育において**シャドーイング**が注目されている．シャドーイングとは，聴取した（母語話者により発声された）外国語音声を即座に繰り返して発声する外国語聴取・発音訓練法である．元来，同時通訳者の訓練として広く行なわれていたが（この場合，故意に delay を置いて通訳するなど，認知的により高いタスクを要求する），外国語学習においてもシャドーイング学習の効果が認められるようになった [21, 22].

学習初期段階の日本人が英語を発声すると，カタカナ英語と呼ばれる発音となることがある．認知心理学的には「英単語の発音が，日本語の音韻に変換された状態で，長期記憶中の心的辞書（メンタルレキシコン）に保持されていることに起因する」と考えられている．シャドーイングは，心的辞書から語彙情報を検索する時間を十分に与えずに発声を要求するため，入力音声の音的イメージをそのまま再生させることに繋がり，母国語の音韻体系に引きずられることなくスピーキング能力を向上させることができる [21, 22].

さらに，シャドーイングはリスニング能力の向上ももたらす．リスニングは「知覚」と「理解」から構成されているが，両段階において，認知資源を消費する．シャドーイングは，母語話者の発音を繰り返して聞くことで音声知覚過程を鍛え／自動化し，同時に，スピーキングを通して正確な発音（音的イメージ）を心的辞書に定着させることで，理解の段階により多くの認知資源を割り当てられるようになる．これらの結果，リスニング能力の向上が期待できる [21, 22].

このようにシャドーイングは，スピーキング／リスニング能力を同時に訓練できるため，コミュニケーション能力を重視する近年の外国語学習において広がりを見せている．学習意欲維持のためには学習者が自らの習熟度を把握し，また教師側は，学習者発声を短時間で評定し教示する必要がある．しかし，シャドーイングは非常に負荷の大きい訓練であり，シャドーイング音声は一般にかなり「崩れた」音声となる．人手でこれらを逐一評定することは膨大な時間を要するため，発音評定技術を用いた自動化が望まれるところである．しかしシャドーイング音声は，従来の評定技術が対象としてきた比較的「奇麗な」音声とはかなり異なる．筆者の知る限り，シャドーイング音声を対象とした自動評定手法は提案されていない．

6.3　従来のシャドーイング音声の評定方法

初期段階の英語学習者のシャドーイング音声は，非常に崩れた発声となる．耳から入る音声の知覚が十分に自動化されておらず，更には，英語の各音韻・音節を生成するための調音運動が十分に習得できていないため，時に言い黙り，言い淀みが生じる．逆に，熟練者のシャドーイング音声は調音努力が十分になされた流暢／明瞭な音声となる．このような発声の差異を捉える，手動によるシャドーイング音声評定方法が玉井によって 2 種類提案されている（音節法及びチェックポイント法 [21]）．また，これらの手法の問題点を考察し，3 つ目の手動評定方法として「全単語法」を考える．

6.3.1　音節法

音節法とは，英語における発話の最小単位と考えられている音節毎の正誤を判定する方法である．素材となる外国語テキストの書き起こしをもとに，1 音節語はそのままに，2 音節以上の単語は各音節毎に分け，評定を行なう．評定単位が単語より小さく，評定の信頼性が保たれる．しかし，採点者は音声を音節毎に区分化，評定する必要があり，時間的コストや体力的負担を被ることとなる．そのため，必ずしも実用性が高い方法ではないと筆者は考える．

6.3.2　チェックポイント法

音節法と違って，単語毎に評定する簡便法として，**チェックポイント法**が提案されている．英語テキストの全単語を n 単語毎に，その単語が正しく発声できているかどうかを判定する．[21] では，全単語が 350 語以上で，各文が 8 単語程度の長さで構成されている場合に n=5 を採用している．この場合，音節法との評定結果の相関として 0.89 が示されている [21].

6.3.3　全単語法

音節法は一見精度が高そうに見えるが，because を become とシャドーイングした場合，音節法では 50 %の正答を与えることになり，チェックポイント法では 0 %になる．because を become とシャドーイングするのは，明らかに単語を取り違えており，英語をコミュニケーション・ツールとして捉え，実践的コミュニケーション能力の養成を目的とする，近年の英語教育の方向性と乖離している．また，n の設定方法は十分に明らかとなっていない．これらを考慮し，本稿で行なう手動の評定は n=1，即ち全単語に対して「その単語が発声できているか」を判定し，手動の評定スコアとした (**全単語法**)．この場合，その単語として意図されたと思われる発声であれば，正解として判定している．

表 6.1: 被験者の TOEIC スコア

熟練度別学習者数	素点	平均点
中位者 3 名	432, 427, 421	427
下位者 3 名	301, 202, 197	233

6.4　音響事象群における事象間距離と調音努力

「個々の音が音響的に明瞭に区別できていない」場合，それは調音的にも区別できていないことを意味する．音響事象群に対して，全ての二事象間距離を求めて幾何学構造（距離行列）として事象群を表象する音声の構造的表象が提案されている．この場合，距離行列から構造のサイズ（構造の半径に相当する）が求まるが，この量が，凡そ調音努力に相当する定量的尺度になることが実験的に示されている [23]．調音努力とは，個々の音を区別して調音するために行なうべき調音運動量と解釈される量である．例えば母音構造を考えれば，その中心には弱母音，即ち，最も脱力した状態で発声される母音が位置しており，その他の母音は，その母音を発声すべく調音努力を払って声道形状を制御して生まれる音である (図 4.1，図 4.2 参照)．事象間距離に対する考察は，読み上げ音声／話し言葉音声の間でも行なわれており，当然話し言葉の方が「なまけ」などの理由で事象間距離が小さくなる [24]．これらを考慮すると，事象間距離の大小を通して，発声時に払われた調音努力の大小を推定することは十分妥当である．

結局，適切な固定閾値の下でクラスタリングを停止させ，その時のセグメント数の大小を議論することは，その発声において払われた調音努力の大小を推定することに相当する．言い換えれば，与えられた発声に対して，どの程度「滑舌の良い」「呂律の回った」発声であったのかを推定することになる．シャドーイング音声は，習熟度が低ければ「もごもごした」音声であり，高ければ「はきはきした」音声となることを考えると，筆者が提案する教師無しセグメンテーションはシャドーイング音声の評定に非常に相性の良い技術であると言える．以下，実験的に検証する[1]．

6.5　シャドーイング音声の自動評定実験

6.5.1　シャドーイング音声の収録と手動のよる評定

日本人英語学習者 6 名にシャドーイングを行なわせた．今回，特に低習熟度者の発声（より崩れた音声）に対する評定技術の構築を考えており，TOEIC テスト（990 点満点）における中位者 3 名，下位者 3 名を対象とした．彼らの TOEIC スコアを表 6.1 に示す．

シャドーイング用に提示した音声は，1 名の男性母語話者が読み上げた音声であり，全 21 文（335 単語）である．平均話速は 140 語/分であった．6 名による合計 126 発話のシャ

[1]なお，ボトムアップクラスタリングを行なわず，初期の $N \times N$ 距離行列（N= 総フレーム数）における構造サイズをもって，与えられた発声の調音努力は推定可能と考えられるが，本研究では固定閾値におけるセグメント数をもって調音努力と解釈した．

表 6.2: 音響分析条件

サンプリング	16bit / 16kHz
窓及び窓長	ハミング窓 / 16msec
シフト長	10msec
音響パラメータ	MCEP 1〜12 次元
クラスタリング停止条件	閾値 $K = 0.23$

ドーイング音声を実験に用いた．なお，収録に用いた教室では空調等の定常雑音が随時発生しており，提示音声の収録環境とは異なる．

手動による評定作業は，小・中・高校，及び，大学で英語授業を実践してきた英語教育の専門家（第三著者）によって全単語法で行なわれた．各発声に対して，その単語として意図された単語発声の個数を数え上げ，その文に含まれる語数で割った値（百分率）を，その発声のスコアとした．1 名分（約 4 分の音声データ）の評定に 2 時間程の時間を要した．

6.5.2　音響分析及びクラスタリングの諸条件

各種の分析条件を表 6.2 にまとめる．スペクトル変化を捉える音響特徴量として，聴覚特性を考慮したメルケプストラムを用いた．なお本実験では，収録されるシャドーイング音声によってはパワーが大きく異なることから，パワー項（MCEP の 0 次項）は用いていない．

クラスタリングの停止条件である閾値 K は，事前実験により，TIMIT データベース train パートの全 4620 発話に対して，正解音素数と自動推定音素数との相関が 0.83 と最も高かった $K = 0.23$ を用いた[2]．

6.5.3　各シャドーイング発声の自動評定結果

収録した 126 発声及び，提示した 21 発声を各々クラスタリングし，自動停止した時のセグメント数を算出した．そして「提示音声のセグメント数」に対する「発声者が生成したセグメント数」を百分率で算出し，これをその発声の自動評定スコアとした．自動評定／手動評定スコアの関係を図 6.1 に示す．TOEIC スコアの中位者／下位者を青／赤で示している．全体の相関は 0.57 となり，高い関係性は示されなかった．

しかし，この相関図を中位者／下位者別にプロットすると（図 6.2），両者の分布には大きな差があることが分かる．例えば，自動評定スコアが 50 未満の発声が散見されれば確実に下位者である．中位者にとっては今回のタスクの 21 文はどれも似たような難易度であり，その結果手動／自動両スコアとも固まって存在する．下位者にとっては提示された 21 文には容易な文から困難な文まで存在し，スコアは手動／自動ともに大きくばらついている．そこで 2 人以上の下位者が手動評定スコアで 40 点以下となった文（9 文）に限定して

[2]なお，5.1.4 節で述べた音素検出の F 値基準における最適閾値は $K = 0.22$ であった．ここでは，音素数推定の最適閾値を使用する．

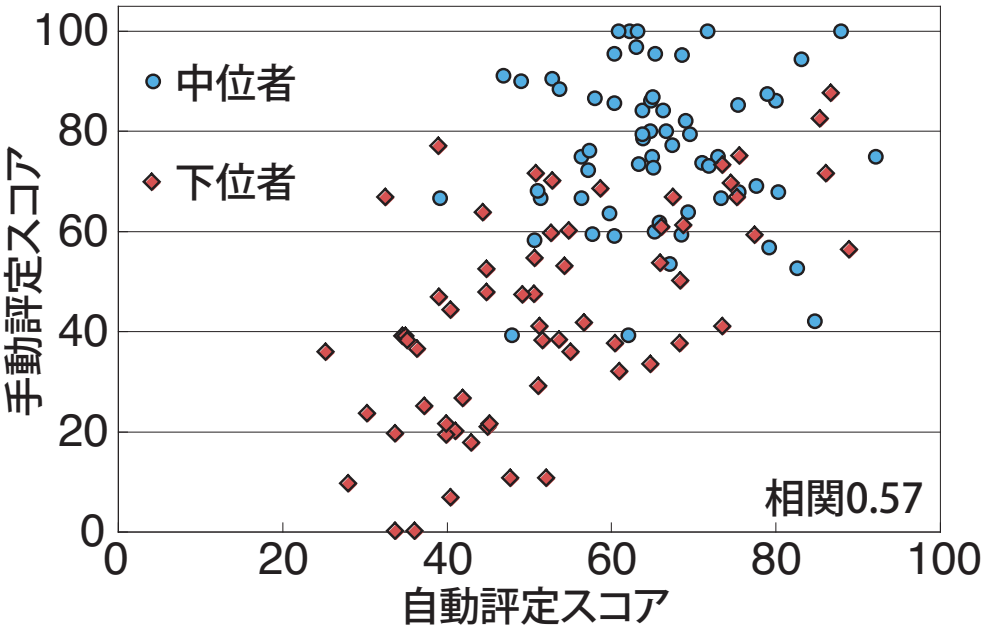


図 6.1: 自動評定スコアと手動評定スコア

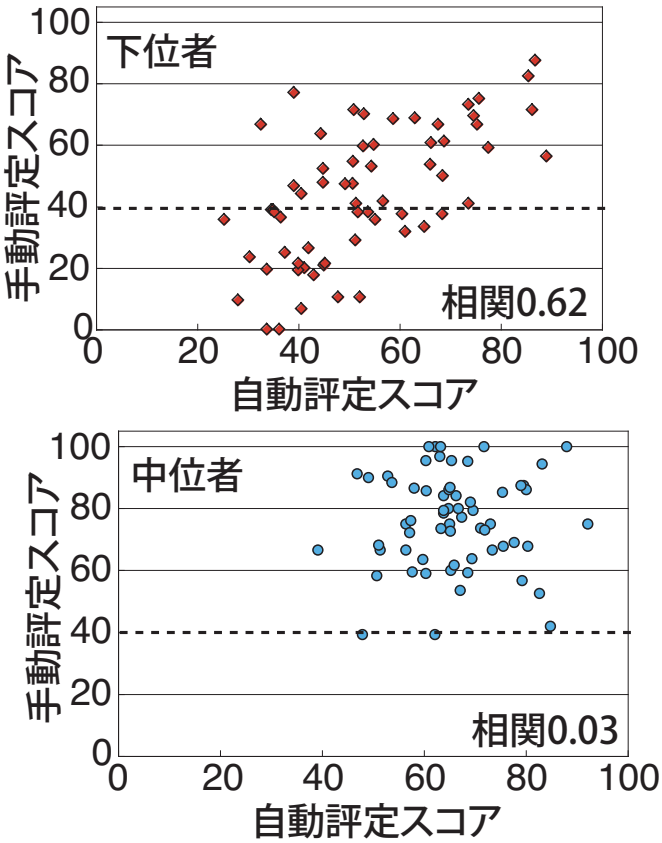


図 6.2: 各学習者グループに対する評定スコア

再度分析を行なった．結果が図 6.3 である．相関は 0.70 となった．なお，黄は TOEIC スコア 301 点の学習者である．図よりスコア 200 点群と 425 点群とは手動／自動スコア両者において比較的明確に分離され，300 点学習者がその両者に渡って存在する様子が分かる．習熟度を自動推定する場合，当然，習熟度の高低が適切に反映されるタスクを課す必要がある．タスク設定を適切に行なうことができれば，提案手法の実用性も向上すると考えられる．

6.5.4　DP マッチングによる自動評定結果

提案手法は，提示音声とシャドーイング音声間で一切音響的照合は行なっていない．シャドーイングは提示音声をそのまま真似る（再生する）という側面を有しており，比較対象として，提示音声と再生音声を音響的に照合することでスコアを算出する **DP マッチング**を行なった．この場合，両音声間の DP スコアが小さいほど「音響的に」類似している音声となる．その結果，DP スコアと手動評定スコアとは負の相関が見られるはずである．図 6.4 に図 6.3 で用いた 9 文に対する正規化 DP スコアと手動評定スコアの関係を示す．425 点群／ 200 点群間の手動評定スコアの差は明確に現れているが，正規化 DP スコアにはその差は全く現れておらず，無相関の結果となった．DP マッチングは，話者性の違いによってスコアが大きく変動する．不一致（ミスマッチ）問題が浮き彫りになった．

6.6　まとめ

筆者が提案している教師無しセグメンテーション手法が，調音努力（滑舌の良さ）に相当する定量的尺度を提供できることを鑑み，近年注目を集めているシャドーイングに着目し，その自動評定を試みた．その結果，シャドーイング対象となる文セットを適切に選択することができれば，手動の評価に沿った自動評定が可能であることを示した．その一方で，DP マッチングでは自動評定は極めて困難との結果を得た．

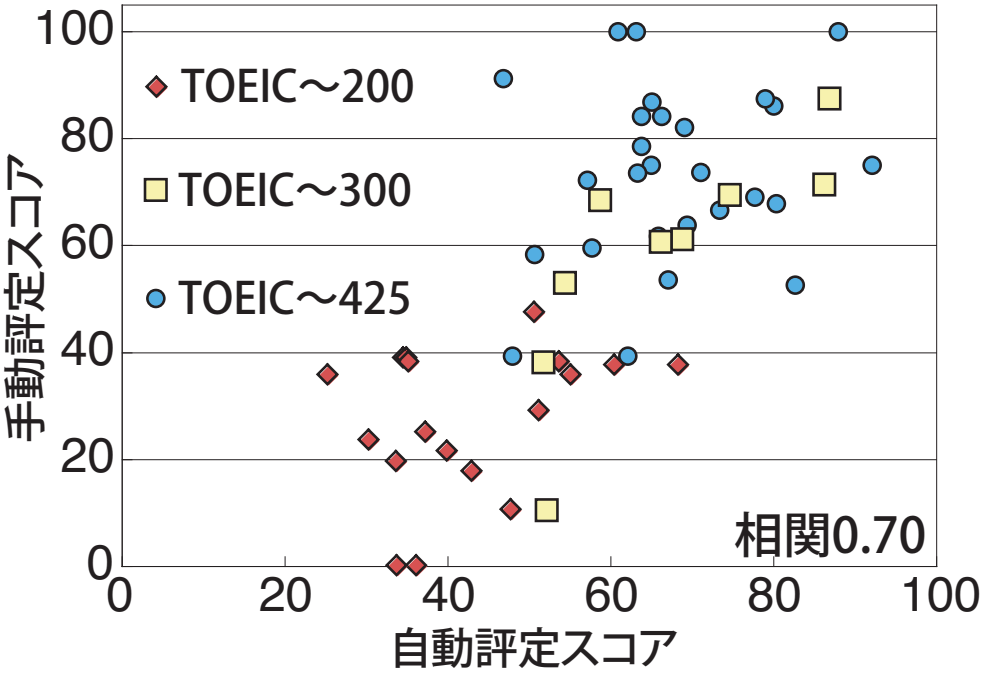


図 6.3: 低スコア文に対する手動／自動評定

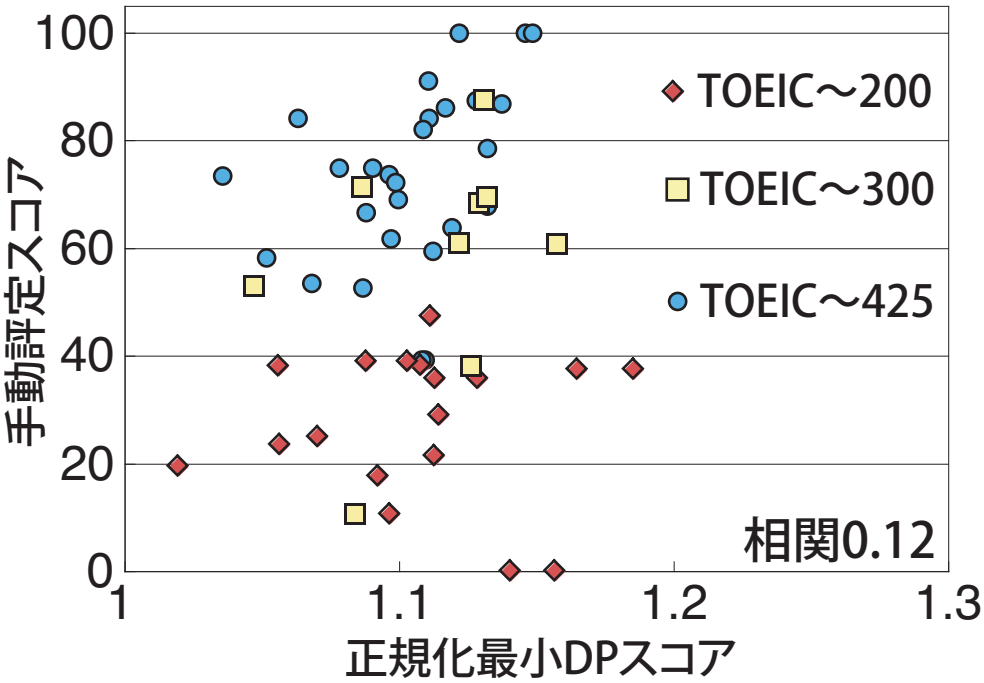


図 6.4: 正規化 DP スコアと正規化手動評定スコア

第7章

結論

7.1　本研究のまとめ

本研究では，自動音素境界検出というタスクを成すために，音声時系列上で隣接するクラスタのみをマージ対象とする制約条件付きボトムアップクラスタリングを提案した．クラスタリング手法を 5 つ紹介し，実験的に Ward 法が 20msec 以内の誤差での音素検出率が約 70% で，最も高いことを明らかにした．さらに，自動で音素数を推定する手法についても検討し，音素数を既知とした場合とほぼ同程度の F 値約 0.70 を得た．

また，音響モデルを用いない先行研究として，スペクトル特徴量に基づく検出手法 [17] においては F 値約 0.67，SVM による検出手法では音素検出率約 75%という結果が報告されている [19]．前者においては精度で，後者に対しては事前知識として音素数を与える必要がない点で，本手法は優位性をもつ．

しかしながら，読み上げ音声を対象とした場合，HMM による強制切り出しに基づく手法では，約 90%の音素を 20msec 以内の誤差に検出することが可能であり，精度では遠く及ばない．その意味で本手法は，音響分析を基本とするセグメンテーションが効果的に利用できる場面を具体的に検討することは必須である．例えば，感情の入った音声や歌声等の特殊な発声，非母国語話者の音声，幼児音声，動物や虫の鳴き声など，そもそもデータの入手が困難なデータのラベリングに利用することが考えられる．

そこで本研究ではセグメンテーションの応用として，崩れた音声となることが多いシャドーイング音声の評価に焦点を当て実験的検討を行なった．学習者レベルに応じた適切な難易度のテキストを選ぶ事により，シャドーイング音声の自動評定スコアと手動評定スコアの間に 0.70 という良好な相関を得ることができた．

本提案手法は言語非依存の技術である．仮に，新たに言語が発見された場合であっても，その直後に，その言語の発音・聴取能力の自動評定が可能となる技術である．また提案手法は不一致（ミスマッチ）問題とは無縁の技術である．距離行列計算で必要なのは，同一話者内での「音と音の距離計算」のみである．DP や HMM のように異なる話者間で「音と音の距離計算」を行なえば，不可避的に不一致問題が発生する．発音評定を行なうシステムを構築する場合，低いスコアを提示された時に，それが学習者の習熟度が低いからなのか，それとも学習者の声質がシステムの学習データに合致しないのか，不明であることも多い．教育応用には，安全かつ健全な技術構築が望まれると考える．

7.2　今後の検討課題

時間制約条件付きボトムアップクラスタリングによる音素セグメンテーションというタスクにおいて，現在さらなる検討がなされている．発表文献 [5, 6, 7] では，本稿で示した精度を上回る音素境界検出が達成している．そのため本研究の課題は，音素セグメンテーション技術よりも，シャドーイング自動評定技術に多く見受けられる．

まずは，シャドーイングデータ収集が課題として挙げられる．シャドーイング音声という特別な音声に一般的なデータベースはなく，自ら収録していく必要がある．本研究で用いたシャドーイング音声は，全 6 名，126 発話のみである．そのため，話者べ*ースの解析，

例えばTOEICと本手法によるシャドーイング音声評価との相関をとるなどの検討が不十分となっている．

次に，シャドーイング評価に用いる最適テキストデータ，聴取データの構築も求められる．非常にレベルの低いテキストを非常に低速で読んだ外国語音声をシャドーイングすることは，どのような学習者であっても容易い．学習上級者，中級者，下位者を線引きできるような難易度の高低があるテキストデータを作成する必要がある．

謝辞

本研究を進めるにあたって、懇切丁寧なご指導、ご意見を賜りました指導教官の峯松信明准教授，そして広瀬啓吉教授に感謝致します。峯松先生は，本研究の方針について様々な角度からヒントを与えてくださり、日々熱心な議論の機会を頂きました。さらには，研究者としての基本姿勢から，学会，輪講の原稿チェック，上手なプレゼンテーション方法等，多くのアドバイスを頂きました。広瀬先生には大域的な視点から研究の問題点，発展性等についてのご指導を頂きました。

本研究を行うにあたり必要になった，特殊音声であるシャドーイング学習音声の収録，採点，また，英語学習の専門家としての様々なご意見を頂きました東京国際大学の山内豊教授に深く感謝致します。

縁の下の力持ちとして研究の環境を整えていただいた高橋登技官、諸手続きを私に変わって行なっていただいた笠島恵美さんに感謝致します。おかげで私は研究活動に専念することができました。

入学当初から研究に関する密なご指導，ご鞭撻を頂きました博士課程の朝川智さん，研究員の喬宇さんに感謝申し上げます。朝川さんは研究の実験環境整備，具体的な実験方法，詳細かつ幅広い工学の知識について多くの解説を頂きました。喬さんは，研究の新たな視点と理論的部分の強化についてご指導頂きました。ありがとうございました。

また，同期の斎藤大輔君，鎌田圭君とは切磋琢磨し，様々な刺激を受けました。研究に関する深い議論を出来たことが頭の整理になったと感じています。そして，研究生活を日々盛り上げて頂いた広瀬・峯松研究室の先輩方、同期、後輩の全てのみなさんに感謝申し上げます。

多くの方々より音声データをご提供頂きました．音声収録等に協力してくださった全ての皆様にこの場を借りて深く感謝致します．

2008 年 1 月 29 日
下村直也

参考文献

[1] N. Minematsu, T. Nishimura, K. Nishinari, and K. Sakuraba, “Theorem of the invariant structure and its derivation of speech Gestalt”, Proc. Int. Workshop on Speech Recognition and Intrinsic Variations, pp.47-52 (2006)

[2] O. Scharenborg, M. Ernestus, V. Wan, “Segmentation of speech:Child’s play?”, Proc. EUROSPEECH, pp.1953-1956 (2007)

[3] International Phonetic Association, *Handbook of the International Phonetic Association*, Cambridge University Press (1999)

[4] 川越いつえ，英語の音声を科学する，大修館書店 (1996)

[5] 浦谷則好，竹沢寿幸，田代敏久，森元逞，匂坂芳典，“ATR の新音声言語データベース”，情報処理学会全国大会講演論文集，vol.48，pp.79-80

[6] 中川聖一, “音声認識の動向”, 電子情報通信学会論文誌, Vol.83-D2, No.2, pp.433-457 (2000)

[7] http://www.sp.nitech.ac.jp/˜tokuda/SPTK/index-j.html

[8] http://htk.eng.cam.ac.uk

[9] T. Svendsen, K. Kvale, “Automatic alignment of phonemic labels with continuous speech”, Proc. ICSLP, pp.997-1000 (1990)

[10] J. -W. Kuo and H. -M. Wang, “Minimum boundary error training for automatic phonetic segmentation”, Proc. ICSLP, pp.1217-1220 (2006)

[11] J. Keshet, S. Shalev-Shwartz, Y. Singer, and D. Chazan, “Phoneme alignment based on discriminative learning”, Proc. ICSLP, pp.2961-2964 (2006)

[12] Y. Tsubota, T. Kawahara, and M. Dantsuji, “Recognition and verification of English by Japanese students for computer-assisted language learning system”, Proc. ICSLP, pp.1205-1208 (2002)

[13] 菊池英明, 前川喜久雄, “自発音声に対する音素自動ラベリング精度の検証”, 春音講論, 2-5-12, pp.97-98 (2002)

[14] 米澤朋子, 水野秀之, 阿部匡伸, “HMM 音素モデルによる自動ラベリングのロバスト性の検討”, 信学技報, SP2002-74, pp.17-22 (2002)

[15] S. Watanabe, Y. Minami, A. Nakamura, and N. Ueda, “Variational bayesian estimation and clustering for speech recognition” IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING, Vol.12, No.4, pp.364-381 (2004)

[16] H. Lo, H.Wang, “Phonetic boundary refinement using support vector machine”, Proc. ICASSP, pp.933-936 (2007)

[17] S. Dusan, L., Rabiner, “On the Relation between Maximum Spectral Transition Positions and Phone Boundaries”, Proc. ICSLP, pp.645-648 (2006)

[18] L. Golipour, D. O’shaughnessy, “A new approach for phoneme segmentation of speech signals” Proc. EUROSPEECH, pp.1933-1936 (2007)

[19] Y. P. Esteva, V. Wan, O. Scharenborg, “Finding maximum margin segments in speech”, Proc. ICASSP, pp.937-940 (2006)

[20] 宮本定明, クラスター分析入門 ファジィクラスタリングの理論と応用, 森北出版 (1999)

[21] 玉井健，“リスニング指導法としてのシャドーイングの効果に関する研究”, 神戸大学大学院総合人間科学研究科博士学位論文（2001）

[22] 門田修平，“シャドーイングと音読の科学”, コスモピア株式会社（2007）

[23] N. Minematsu, S. Asakawa, and K. Hirose, “Para-linguistic information represented as distortion of the acoustic universal structure in speech”, Proc. ICASSP, vol.1, pp.261-264 (2006)

[24] M. Nakamura, S. Furui, and K. Iwano, “Acoustic and linguistic characterization of spontaneous speech”, Proc. Int. Workshop on Speech Recognition and Intrinsic Variations, pp.3-8 (2006)

発表文献

[1] 下村直也, 朝川智, 峯松信明, 広瀬啓吉, “制約条件つきクラスタリングによる連続音声からのイベント境界検出”，電子情報通信学会音声研究会, SP2007-6, pp.25-30 (2007)

[2] 下村直也, 朝川智, 峯松信明, 広瀬啓吉, “時間制約を持つクラスタリングによる連続音声の自動セグメンテーション”，日本音響学会秋季講演論文集, 3-4-4，pp.353-356 (2007)

[3] 下村直也, 峯松信明, 山内豊, 喬宇, 朝川智, 広瀬啓吉, “教師無しセグメンテーションを用いたシャドーイング音声の自動評定に関する実験的検討”，日本音響学会春季講演論文集 (2008, 発表予定)

[4] 下村直也, 峯松信明, 山内豊, 広瀬啓吉, “ボトムアップクラスタリングを用いたシャドーイング音声の自動採点”，電子情報通信学会音声研究会, SP2008-3 (2008, 発表予定)

[5] Y. Qiao, N. Shimomura, and N. Mimenatsu, “Toward optimal unsupervised phoneme segmentation – a theoretical and experimental investigation –” IEICE Technical Report, SP2007-12, pp.161-166 (2007)

[6] Y. Qiao, N. Shimomura, and N. Mimenatsu, “Unsupervised optimal phoneme segmentation: objectives, algorithm and comparisons” Proc. ICASSP (2008, accepted)

[7] Y. Qiao, N. Shimomura, and N. Mimenatsu, “Optimal phoneme segmentation using weighted cepstrum features” Proc. Spring Meeting of Acoust. Soc. Japan (2008 accepted)