修士論文

音素重回帰と文強勢検出を用いたシャドーイング音声評価の高精度化

2013 年 2 月 6 日

指導教員　峯松 信明 教授

電子情報学専攻

48-116415　加藤 集平

内容梗概

本研究の目的は，従来から存在する Goodness of Pronunciation (GOP) を用いたシャドーイング評価手法を，重回帰および文強勢検出を用いることで高精度化することである．GOP とは，HMM 事後確率に基づく音素評価スコアで，発音評価に広く用いられている．従来手法では，シャドーイング音声中の各音素区間に対して GOP を計算し，それを文章全体にわたって平均した GOP_{all} を自動評価スコアとしている．そして，GOP_{all} と発話者の TOEIC スコアに良好な相関があることを見出している．つまり，GOP_{all} を説明変数とする単回帰で，発話者の TOEIC スコアを良好に推定できることになる．これに対して，本論文で述べる提案手法では，発話に対して GOP に基づいて計算した複数の自動評価スコア，および，文強勢検出に基づくスコアを説明変数とする重回帰を行うことによって，従来手法よりも高精度に TOEIC スコアを推定することを目指した．ただし，文強勢検出については，シャドーイング音声から直接文強勢検出を行うことは困難であるため，読み上げ音声を用いた文強勢検出の実験を行い，その技術を応用してシャドーイング音声の文強勢検出スコアを計算した．また，読み上げ音声を用いた文強勢検出の実験においては，識別モデルを用いて精度の高い文強勢検出を行うこと，音響特徴量以外の，テキストから得られる特徴量も積極的に使って精度を上げることを目指した．本論文では，重回帰を用いたシャドーイング評価の高精度化，読み上げ音声における文強勢検出，文強勢検出のシャドーイング評価への応用の順に述べる．

重回帰を用いたシャドーイング評価の高精度化については，結果として，説明変数に用いるスコアによっては，重回帰（最小二乗法あるいはリッジ回帰）によって単回帰の場合よりも高精度に TOEIC スコアを推定することができた．

読み上げ音声における文強勢検出については，日本人学生による英語読み上げデータベース（ERJ データベース）のリズム文の読み上げ音声を評価対象として，SVM を用いて高精度に文強勢の自動推定を行うことを目指し，音節ごとに強勢の度合いを自動推定する実験を行った．その過程で，評価対象音声に対して，手動判定により強勢の度合いを 9 段階でラベリングする作業を行ったが，3 段階に丸めて用いた場合と，9 段階のまま使用した場合の 2 通りについて実験を行った．特徴量セットに関して様々な検討を行い，評価者 closed の場合に最高で 81.8%（3 段階），47.0%（9 段階）の正解率を得ることができたが，手動判定の精度には及ばなかった．評価者 open の場合は手動判定に近い結果が得られた．

文強勢検出のシャドーイング評価への応用については，読み上げ音声に対する文強勢検出でも用いた SVM を，SVM を回帰に応用した SVR に変えることで強勢の度合いを連続値で推定し，それをもとに文強勢スコアを定義した．それを重回帰の説明変数に加えて，シャドーイング評価のさらなる高精度化を目指した．結果として，GOP をもとに重回帰を

行った場合よりも，同程度か上回る精度が得られた．

目次

図目次

表目次

第1章

序章

1.1　本研究の背景

現代社会は国際社会であり，日本も例外ではない．実際，日本の出入国者数は 1991 年の 2,885 万人から，2011 年の 4,853 万人へと，20 年間でおよそ 1.7 倍に増加している [1]．また，海外在留邦人数は 1991 年の 66 万人から，2011 年の 118 万人へと，20 年間でおよそ 1.8 倍に増加している [2]．現代の国際社会の状況を考えると国際化が後退するとは考えにくく，今後さらに進んでいくものと予想される．

このような状況の中で，国際的な意思疎通の手段としての英語を修得する必要性が叫ばれるようになり，そのための語学教育がより盛んに行われるようになってきている．例えば，2011 年度から，小学 5，6 年生での外国語活動が必修となった．これは，文部科学省が 2002 年度に定めた “「英語が使える日本人」の育成のための戦略構想” において小学校における英会話活動支援が構想され，2011 年度より実施された学習指導要領により先に述べた外国語活動が必修化されたことによるものである [3,4]．さらに，将来的には小学 4 年生以下でも外国語教育を必修化することが検討されている [5,6]．また，ビジネスにおいても英語の必要性は増しており，成人向け外国語教室の市場規模は 2011 年度で 1,951 億円に達しているほか，英語を社内公用語とする日本企業も現れている [7–9]．このように，日本人が語学教育を受けたり，英語を使用する機会は着実に増加している．

語学教育の方法も変化している．学習指導要領の規定によれば，先に述べた小学校での外国語活動では，外国語を用いてコミュニケーションを図る楽しさを体験する，積極的に外国語を聞いたり，話したりすることを指導することとされ，話し言葉としての外国語教育が重視されている [4]．また，多くの成人向け外国語教室でも，話し言葉によるコミュニケーションを重視した教育が行われている．これは，ビジネスの現場において，実際に外国人と英語を使ってコミュニケーションを行う能力が強く求められているからであろう．このように，近年の語学教育においては，聞く・話すことによるコミュニケーション能力の向上を目標とした教育が重視されている．

しかし，そのようなコミュニケーションを中心とした語学教育を十分に行える能力を持った教師が満足に確保されているかというと，そうではない．例えば，小学校の外国語活動の現状について日本英語検定協会が行った調査によると，外国語活動における問題・課題であると感じていることとして，“指導者（担当教員）の質・技術” が，教育委員会を対象とした調査では 15 項目中 1 番目に，公立小学校を対象とした調査では 16 項目中 3 番目にあげられている [10,11]．十分な教育能力を持った教師が不足している状況では，いくらコミュニケーション能力の向上を教育目標としたところで，実効性は限られたものになってしまうと考えられる．

そこで，教師不足をコンピュータによって補う，Computer Assisted/Aided Language Learning (CALL) システムが数多く開発されている．CALL には，学校の英語教室で導入されるもの [12–15]，PC やゲーム機上で動作するもの [16–19]，ウェブサイトで動作するもの [20] などがある．機能としては，単に発音練習を行うだけでなく，音声認識技術を応用して学習者の音声を自動評価して全体的なスコアを算出したり，発音分析を詳細に行ったりするものが登場している．教師不足が問題となっている以上，CALL の果たす役割は大

きく，その改良・発展は重要であると考えられる．

そこで，既存の CALL の機能をどのように改良・発展させることができるか考えることにする．まず，学習者の音声を自動評価して全体的なスコアを算出する機能について考えたい．市販のソフトウェアで自動評価を行う場合，通常，学習者がソフトウェアに指示された文章を読み上げた読み上げ音声に対して評価がなされることが多い．これに対して，シャドーイング音声に対して自動評価を行う研究がある [21,22]．シャドーイング (shadowing) とは，聴取した外国語音声を即座に繰り返して発声することで発音能力と聴取能力とを同時に鍛える外国語聴取・発音訓練法である．提示音声に対して影 (shadow) のように追従して発声することから，shadowing と呼ばれている．もともとは同時通訳者の訓練として広く行われていたが，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ている [23–25]．シャドーイングにおいては，学習者が提示音声をそのまま真似ることは難しく，学習者自身の話し方の癖や学習者の母語に関する言語知識が無意識のうちに使われることが知られている [26]．

シャドーイングでは提示音声の発話速度に追従する必要があるため，学習者のシャドーイング音声はかなり崩れた，不明瞭なものになることが多い．そのため，シャドーイング音声を評価しようとする場合，人手で評価するには膨大な時間を要する．そこで，発音評価技術を用いてこれを自動評価する手法が提案されている．[21] では，HMM 事後確率に基づいた発音評価スコアである Goodness of Pronunciation (GOP) を発話を/l/や/r/といった音素単位に区切った音声セグメントごとに計算し，それらを発話した文章全体にわたって平均した GOP_{all} を，発話に対する自動評価スコアとしている．そして，GOP_{all} と発話者の TOEIC スコアの間に高い相関があることを明らかにしている．また，[22] では，[21] と同様の手法による自動評価を，同一の発話者・テキストによるシャドーイング音声，テキスト付きシャドーイング音声，読み上げ音声に対して行い，発話者の TOEIC スコアとの相関を比較している．結果は，シャドーイング音声，テキスト付きシャドーイング音声，読み上げ音声の順に TOEIC スコアとの相関が高く，自動評価の対象としてシャドーイング音声を用いることの有効性が示されている．

ここで，[21] の手法を発展させることにする．[21] では，GOP を発話した文章全体にわたって平均した GOP_{all} を，自動評価スコアとして用いている．すなわち，発話に対して GOP_{all} という 1 つの自動評価スコアを計算し，その GOP_{all} を説明変数とした線形単回帰を行うことにより，発話者の TOEIC スコアを推定できることになる．本研究ではこれを発展させ，発話に対して GOP に基づいた複数の自動評価スコアを計算し，それらを説明変数とする線形重回帰を行うことにより，発話者の TOEIC スコアを単回帰の場合よりも高精度に推定することを検討した．また，文章によって各音素の出現回数は異なるので，GOP_{all} は文章への依存度が高いスコアと考えられる．そのため文章によって TOEIC スコアの推定精度に大きな差が出てしまう可能性がある．そこで，文章への依存度が低いと考えられるスコアを使用して，文章によらず TOEIC スコアを高精度に推定することを目指した．

次に，既存の CALL では実装されることの少ない，韻律の自動評価機能について考えたい．韻律とは強勢やリズム，イントネーションといったものの総称であり，これを正しく

身につけることは，/l/と/r/といった音素をそれぞれ正しく発音することと並んで，学習者が自然な英語の発音を身につけるためには重要なことである．ところが，市販のソフトウェアでは，発話中の各音素に対して自動評価を行ってその結果を表示するものは多い一方，韻律の自動評価を行うものは少ない．先に述べた重回帰によるシャドーイングの自動評価の高精度化の検討においても，あくまで GOP という音素に基づいたスコアの範疇で精度向上を図っているが，韻律評価に基づくスコアも使用することができれば，さらなる精度向上が期待できる．しかし，学習者のシャドーイング音声はかなり崩れた，不明瞭なものになることが多く，そのような音声に対して韻律評価を行うことは現状では困難である．そこで，本研究ではより簡単なタスクである読み上げ音声に対する文強勢の自動検出を行い，その技術を応用してシャドーイング評価に韻律評価の要素を導入することにした．韻律の中でも強勢，特に文強勢に注目したのは，文強勢は自然な英語リズムを形成するほか，強調などの話者の意図が表現されるため，母語話者と十分な意思疎通を行う上で重要な要素だからである [27]．また，読み上げ音声の音声試料としては,「日本人学生による読み上げ英語音声データベース」（ERJ (English Read by Japanese) データベース） [28] の「文強勢，文リズムに関する文」（以下リズム文とする）に対する読み上げ音声を用いることにした．この音声の特徴は，英語の強勢・リズムに着目して構成された文を，原稿に記載された強勢記号に従って学習者自身が読み上げたもので，学習者自身は正しく読めたと考えている音声となっている点である．読み上げ文が英語の強勢・リズムに着目して作られている点，学習者自身が正しく読めたと考えている音声である点で教育的な観点から評価対象としてふさわしいと考えた．実は，ERJ に収録されている音声を用いて文強勢検出を行う研究は既に存在する [29, 30]．ただし，[29, 30] では HMM を使って評価しているために，特徴量を増やしたりすることに限界がある．そこで本研究では，識別モデルを用いてより精度の高い文強勢の自動評価を目指す．具体的には，評価対象音声に対して音節ごとに強勢の度合いを高精度に自動評価することを目的とする．ただし，ERJ には音節ごとに強勢の度合いを評価したラベルは付与されていないため，手動評価という形でラベルを付与した．また，本研究では読み上げ音声を使用するため，評価対象音声の発話内容は既知である．そこで，音響特徴量以外の，テキストから得られる特徴量も積極的に使って評価精度を上げることを目指した．

1.2　本研究の目的

本研究では，[21] のシャドーイング評価手法を，重回帰を用いることで高精度化することを目的とする．また，特徴量として GOP に基づくスコアだけでなく，韻律評価，特に文強勢検出に基づくスコアも利用する．ただし，シャドーイング音声から直接文強勢検出を行うことは困難であるため，読み上げ音声を用いた文強勢検出の実験を行い，その技術を応用してシャドーイングに韻律評価の要素を導入することを目的とする．また，読み上げ音声を用いた文強勢検出の実験においては，識別モデルを用いて精度の高い文強勢の自動評価を行うこと，音響特徴量以外の，テキストから得られる特徴量も積極的に使って評価精度を上げることを目的とする．

1.3　本論文の構成

本論文は，全 8 章で構成される．**第 3 章**では，シャドーイング評価の先行研究およびそれに関連する知識について述べる．**第 4 章**では，文強勢検出の先行研究およびそれに関連する知識について述べる．**第 5 章**では，重回帰を用いたシャドーイングの高精度化について，**第 6 章**では，読み上げ音声における文強勢検出についてそれぞれ述べる．そして，**第 7 章**では，読み上げ音声における文強勢検出技術のシャドーイング評価に応用について述べる．最後に，**第 8 章**で本論文をまとめ，今後の展望について述べる．

第2章

本研究に関する基礎知識

2.1　はじめに

本章では，本研究に関連する基礎知識について述べる．まず，本研究で用いる音声学における用語，音声工学における用語を説明する．次に，日本語と英語の音声学的差異について，母音，子音，音節とモーラ，アクセント，リズムの観点から述べる．そして，日本語を母語とする英語学習者が犯しやすい誤りについて述べる．

2.2　用語の説明

2.2.1　音声学における用語

分節音　音声は一つ一つの音が連結してできるものであるが，この一つ一つの音を**分節音** (segment) または**単音**と呼ぶ [31]．分節音は国際音声記号 (International Phonetic Alphabet; IPA) を用いて表すことができ，[t] のように角括弧に挟んで表記する．

音素　ある言語で使われる単音のうち，意味の違いを起こさない単音の集合を**音素** (phoneme) と呼ぶ [32]．音素も IPA を用いて表すことができ，/t/のように斜線に挟んで表記する．

韻律　分節音を越えた範囲（音節，語，句，文など）で起こる現象を超分節的現象あるいは**韻律** (prosody) と呼ぶ．韻律には，アクセント，リズム，イントネーションなどが含まれる．

2.2.2　音声工学における用語

基本周波数　音声波形は時間とともに変化するが，短時間で見ると振幅がランダムに変化する雑音部分と，ほぼ一様な周期で繰り返す周期音部分とがある．後者の周期音部分の繰り返し周波数のうち，最も低い周波数を音声の**基本周波数** (fundamental frequency; F_0) と呼ぶ．単位はふつう Hz（ヘルツ）が用いられる．基本周波数は声帯振動の周期の逆数に相当するものであり，声帯振動を伴う有声音には存在し，声帯振動を伴わない無声音には存在しない．人間の知覚特性を考慮して，対数尺度に変換することもよく行われる．知覚的には，ピッチ（音の高さ）に対応する．

エネルギー　音声の**エネルギー** (energy) は，人間の知覚特性を考慮して，対数尺度である dB（デシベル）を用いた音圧レベルとして表されることが多い．ある音の音圧レベル L_p [dB] は，音圧を p [Pa]，基準となる音圧の実効値を p_0 [Pa] として以下の式で与えられる．

$$L_p = 20\log_{10}\frac{p}{p_0} \tag{2.1}$$

ここで，$p_0 = 20 \times 10^{-6}$ Pa である．知覚的には，音の強さに相当する．

音韻継続長　音韻継続長とは，音素，音節，句などの要素の物理的継続長のことをいう．知覚的には．音の長さに相当する．

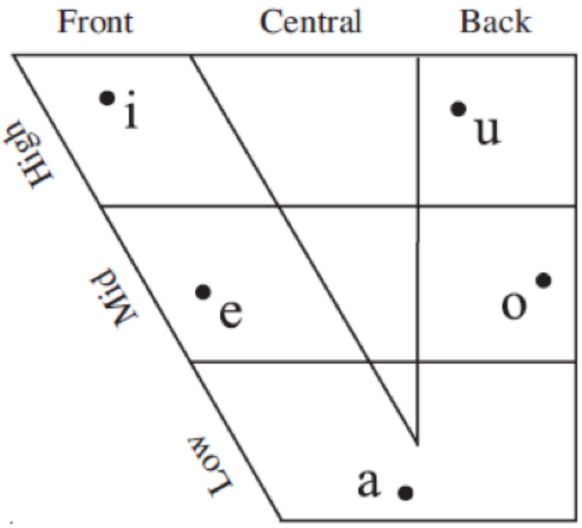

図 2.1　日本語の母音

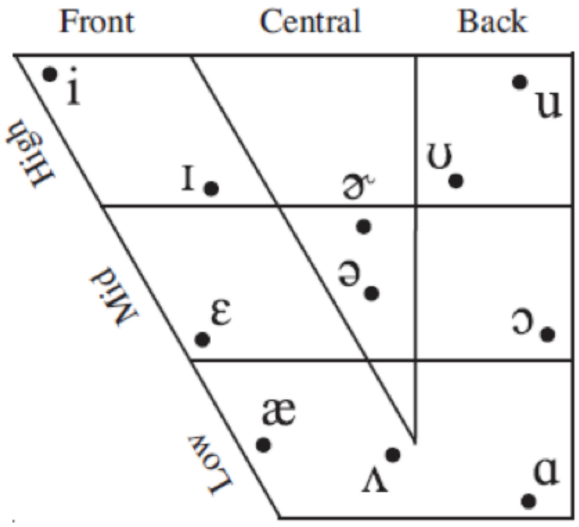

図 2.2　英語 (GA) の母音

2.3　日本語と英語の音声学的差異

2.3.1　母音

音声学において**母音**とは，声門を通ってきた気流が声道内を妨害されずに流れて生み出される音のことであり，すべての母音は声帯の振動を伴う有声音である [31,32]．日本語の母音を図 2.1 に示す [33].

表 2.1　日本語と英語の子音の比較

		調音位置									
調音方法		両唇	唇歯	歯	歯茎	後部歯茎	そり舌	硬口蓋	軟口蓋	口蓋垂	声門
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				ʦ/ʣ	ʧ/ʤ					
	英					ʧ/ʤ					
鼻音	日	m			n			ɲ	ŋ	ɴ	
	英	m			n				ŋ		
弾き音	日				ɾ						
	英				l ɹ		ɻ				
接近音	日	w						j	w		
	英	w						j	w		

図 2.1 中の黒丸は口腔内での舌の位置を示しており，横方向は舌の前後の位置（前，中，後），縦方向は舌の上下の位置（上，中，下）を意味している．日本語は図 2.1 のとおり，/a, i, u, e, o/の 5 つの母音を持っている．

次に，英語の数ある方言の中でもアメリカにおいて代表的なものの一つである，標準アメリカ英語 (General American; GA) の母音を図 2.2 に示す [33]．図 2.2 に書かれている母音のうち，継続長が比較的短い/æ, ʌ, ɛ, ɪ, ʊ/を**短母音**，比較的長い/ɑ, ɔ, i, u/を**長母音**，そして/ə/（およびこれが r 化した/ɚ/）を**弱母音**と呼ぶ（/ɚ/は後で述べる強勢が付与されることもある）．これらは途中で音色が変わらないので，**単母音**と総称される．一方，これらの単母音間を連続的に移動するようにして発音される母音は**二重母音**と呼ばれ，GA においては/aʊ, aɪ, eɪ, oʊ, ɔɪ/の 5 つが存在する．二重母音はあくまで第 1 要素が主で，第 2 要素は付属的に発音されるものであり，全体としてあくまで 1 つの母音である [32,34]．これは，日本語における母音連続（2 つの母音が続けて発音される）とはまったく異なる概念であることに注意されたい．以上のように，英語は日本語に比べて多くの母音を持っている．そのため，日本語を母語とする学習者が英語を発音した場合，英語の母音を日本語の 5 つの母音で置き換えてしまう誤りがよく見られる．

2.3.2　子音

子音は母音と対立する概念であり，声門を通ってきた気流が声道内において何らかの妨害を受けることにより生み出される音のことである [31,32]．子音は母音と異なり，声帯の振動を伴わない無声音と，声帯の振動を伴う有声音の両方が存在する．日本語と英語の子音を，表 2.1 に示す [32]．

ここでは個々の子音についての詳細な説明は省略する．日本語を母語とする学習者にとって発音が難しいものとしては，/l/ (light) や/ɹ/ (right) の区別が代表的であり，どちらも日本語のラ行子音 (/ɾ/) で置き換える誤りがよく見られる．このほか，日本語に存在しない/f/ (five)，/v/ (very)，/θ/ (thing)，/ð/ (that) について，それぞれ日本語のハ行，バ行，サ行，ザ行子音で置き換える誤りもよく見られる．

2.3.3　音節とモーラ

音節 (syllable) とは，単音（上で述べた母音や子音に相当）の集合で作られる，単音より大きな発音の基本単位のことである [32]．音節は 1 つの母音を中心としてその前後に子音が連なったものである．母音を V，子音を C で表したとき，日本語の音節は CV が基本の単位であり，1 音節内で CC と子音が連続することは基本的にない．一方，英語は CVC が基本の単位であり，1 音節内で CC と子音が連続することもある．最も長い音節としては，CCCVCCCC（例: strengths /stɹɛŋkθs/）が知られている．

モーラ (mora) とは，“音節の重さ” を規定する音の単位であるが，簡単に言えば，日本語でいうひらがな 1 文字分に相当する音の単位である．英語においてもモーラは存在し，たとえば短母音は 1 モーラ，長母音は 2 モーラであるとされる [32]．

後述するリズムの観点などから，日本語はモーラ，英語は音節を基調とした言語であるとされる．

2.3.4　アクセント

アクセントとは，単語や句の中で，他の部分よりも目立つ部分のことである [32]．アクセントの実現のしかたは言語によって異なり，音の高さで実現する**高さアクセント** (pitch accent)，強さで実現する**強さアクセント（強勢）** (stress accent) の 2 つに大きく分けられる．強勢は**強さ**アクセントであるが，実際には強さ，高さ，長さ，母音の音質などによって総合的に表現されることが分かっている．日本語は高さアクセント，英語は強さアクセントであると言われる．

強勢とは，ある音節を発声するにあたって，音源である呼気が強くなったりその量が多くなったりして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) が増大する現象のことをいう．強勢を受けた音節はピッチが高くなり，音が長めになる傾向にある．強勢は強勢 (stress) と弱勢 (weak stress) とに二分され，強勢アクセントでは全ての音節はいずれかを受ける．また，英語においては，特に長い単語中で強勢が 2 つ存在する場合があり，より強い方を第一強勢，もう一方を第二強勢という．音節数の少ない単語（3 音節以下）では，強勢は 1 つであることが知られている．

単語中で実現される強勢を**単語強勢**と言い，単語中のある音節に置かれた強勢のことを指す．一方，文中で実現される強勢を**文強勢**と言い，文中の特定の音節が持つ強勢およびその強弱の差のことを指す．語義を持つ内容語 (content word) は強い文強勢を受け，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．

2.3.5　リズム

リズム (rhythm) とは，あるパターンの規則的な繰り返しのことである [32]．リズムの実現のしかたは言語によって異なり，音節がほぼ等間隔で繰り返される**音節拍リズム** (syllable-timed rhythm)，モーラがほぼ等間隔で繰り返される**モーラ拍リズム** (mora-timed rhythm)，強勢がほぼ等間隔で繰り返される**強勢拍リズム** (stress-timed rhythm) がある．日本語はモーラ拍リズム，英語は強勢拍リズムの言語であるとされる．

2.3.6　イントネーション

イントネーション（音調，intonation）とは，アクセントより大きな単位でのピッチの変動のことである．イントネーションによってひとまとまりにされた発話の単位を**音調単位** (intonation unit) あるいは**音調句** (intonation phrase) といい，本論文では単に**フレーズ**と呼ぶことにする．で述べた文強勢については，フレーズ内でもっとも強く読まれるものを核強勢と呼び，英語においては基本的にフレーズ内の最後の内容語に置かれることが知られている [35]．

2.4　日本語を母語とする英語学習者が犯しやすい誤り

2.4.1　分節的誤り

日本語の母音・子音の種類が英語のそれよりも少ないことから，英語の母音・子音を日本語の少ないそれで置き換えてしまう誤りはすでに述べた．その他の誤りとしては，日本語の音節が CV を基本としていることの影響から，英語の音節構造を正しく捉えて発音できない誤りがある．たとえば，strike は/stɹaɪk/のように本来 1 音節で発音されるが，子音と子音の間にないはずの母音を挿入し，/su.to.ɾa.i.ku/のように 5 モーラ・5 音節で発音してしまう，という具合である．このような誤りは，日本語訛りの英語を聞き慣れていない英語話者とコミュニケーションする際の阻害要因となると考えられる．

2.4.2　韻律的誤り

日本語が高さアクセントの言語であることから，本来強さアクセント（強勢）によって実現されるべき英語のアクセントを，高さアクセントで実現する誤りがある．また，リズムについても，日本語がモーラ拍リズムの言語であることから，本来強勢拍リズムで発声すべき英文を，モーラ拍リズムで発声してしまう誤りがある．このような誤りも，日本語訛りの英語を聞き慣れていない英語話者とコミュニケーションする際の阻害要因となると考えられる [32]．

2.5　まとめ

本章では，本研究に関連する基礎知識として，本研究で用いる用語，日本語と英語の音声学的差異，日本語を母語とする英語学習者が犯しやすい誤りについて述べた．次章では，本研究の目的の一つであるシャドーイング評価の高精度化に関する先行研究について述べる．

第3章

シャドーイング評価の先行研究

3.1　はじめに

本章では，シャドーイング評価の先行研究として，羅らによる GOP を用いたシャドーイング自動評価について取り上げる [21]．GOP (Goodness of Pronunciation) は Witt らにより開発された発音評価スコアで，音声認識技術を用いた自動発音評価の分野で広く用いられているものである [36]．本章ではまず，GOP を算出するための要素技術として，短時間音響特徴量として利用されるメルケプストラム (Mel Frequency Cepstrum Coefficients; MFCC) と，音響モデルとして利用される隠れマルコフモデル (Hidden Markov Model; HMM) について詳細な説明を行う．その後，具体的に GOP を算出する方法について説明し，最後に，羅らによるシャドーイング自動評価について述べる．

3.2　短時間音響特徴量時系列の抽出

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

3.2.1　窓関数

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

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

$$W_R(t) = 1 \tag{3.1}$$

$$W_H(t) = 0.54 - 0.46 \cos 2\pi t \tag{3.2}$$

$$W_B(t) = 0.42 - 0.5 \cos 2\pi t + 0.08 \cos 4\pi t \tag{3.3}$$

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

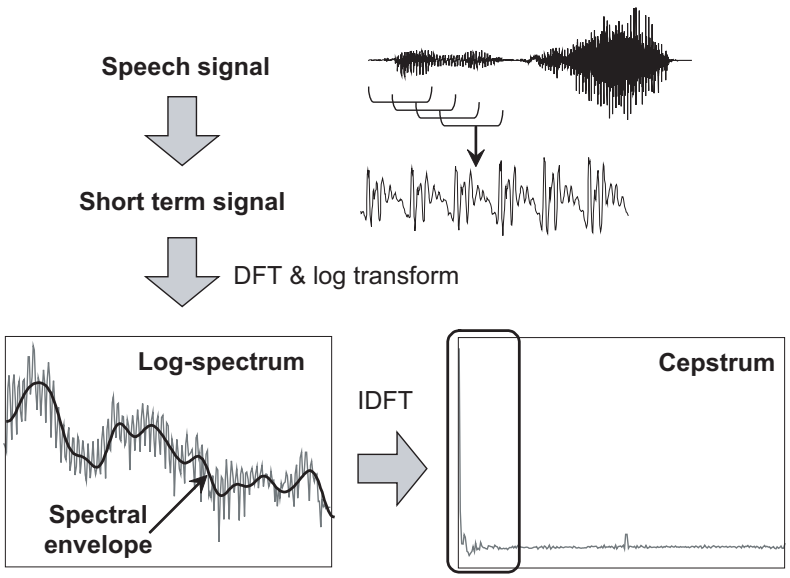

図 3.1　音声信号からのケプストラムの抽出

3.2.2　ケプストラム

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

音声波形に窓をかけ，そこからケプストラムを抽出するまでの様子を図 3.1 に示す．まず音声波形から，窓掛けにより数十 msec 程度のフレームを切り出し，その区間に対して離散フーリエ変換 (Discrete Fourier Transform; DFT) を施し，その対数パワー成分を抽出する．ここで，特徴量として利用したいのは，対数パワースペクトルの包絡成分である．そこで，いったん対数パワースペクトルに対して逆離散フーリエ変換 (Inverse DFT; IDFT) を施す．これがケプストラムと呼ばれる特徴量である．このケプストラムのうちの低次項数十次元のみを残し，高次項を 0 にして DFT してスペクトル領域に変換することにより，スペクトル領域における低周波成分，すなわちスペクトル包絡が得られる．そのため，ケプストラムの低次元はスペクトル包絡の情報のみを小さな次元数で表現した特徴量といえる．そこで，このケプストラムの低次元 10–20 次元分を，短時間音響特徴量として利用する．このケプストラムの低次元のみを抜き出す操作は，リフタリングと呼ばれている．なお，ケプストラムの 0 次元は，対数スペクトル領域でいうオフセット成分に対応しており，これは音声のパワー成分に相当する．パワーはマイクと口との距離などでも変動し，全体的にパワーが変化しても音素は基本的に変化しないため，ケプストラムの 0 次元は音響特徴量から排除されることが多い．ただし，パワーの時間的な変動の様子は情報として有用であるため，ケプストラム 0 次元の変動成分は音響特徴量として用いられることがある．

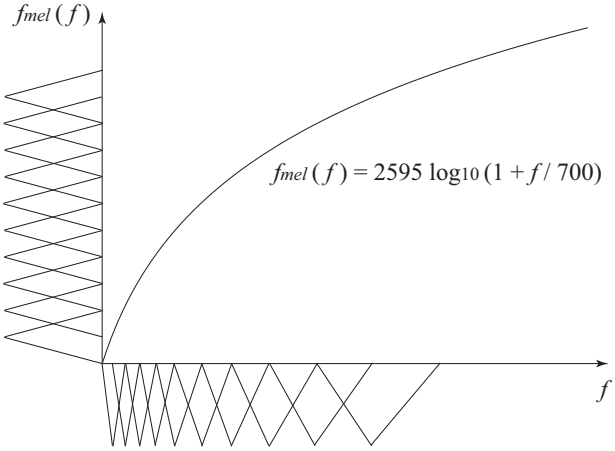

図 3.2　メル周波数軸上に等間隔で配置された三角窓

3.2.3　ヒトの聴覚特性を考慮したケプストラム

音の高さに対するヒトの知覚特性は，低域ほど分解能が高く，広域ほど分解能が低い．具体的には，周波数分解能は周波数に対する対数関数で近似できる．そこで，ヒトの知覚特性に合わせて周波数分解能を変化させて音声分析を行うことで，よりヒトの感覚にあった特徴量が抽出できる．ヒトの知覚特性を反映した尺度であるメル尺度を利用したケプストラムは，数多く提案されているが，現在の短時間音響特徴量のデファクトスタンダードとなっている MFCC(Mel-Frequency Cepstrum Coefficient) も，その一つである．まず，図 3.2 に示すようにメル周波数（メル尺度化された周波数）軸上に等間隔で配置された三角窓を用意し，フィルタバンク分析を行なう．なお，メル周波数 f_{mel} は周波数 f[Hz] に対して，

$$f_{mel}(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \tag{3.4}$$

という周波数ウォーピングを施すことで得られる．このメルフィルタバンクの出力を求めることにより，メルスペクトルが得られる．これに離散コサイン変換を施し，リフタリングを行った物が MFCC である [37]． なお，発音評価においては，メルフィルタバンクの数は 24 程度，MFCC の次元数は 12 程度とされることが多い．

3.2.4　ケプストラムの動的特徴量

ケプストラム係数は，数十 msec 程度の音声区間（フレーム）を定常とみなした上で得られる静的な特徴であるが，音素の音響的な特徴は周辺の音素に影響を受けて変化する調音結合が起こり，スペクトルは時間とともに連続的に変化している．そこで，フレーム分析によって得られる静的な特徴に加え，時間とともに変化する方向成分を動的特徴量として加えることで精度が大きく向上することが知られている [38]．動的特徴量として最もよ

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

$$\Delta c(n) = \frac{\sum_{t=-T}^{T} t w_t c(n+t)}{\sum_{t=-T}^{T} t^2 w_t} \tag{3.5}$$

T としては 2 を利用することが多い．さらに，Δ ケプストラムの動的特徴量である ΔΔ ケプストラムも用いられることがある．

3.3　音響モデル

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

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

発音分析に用いる音響モデルには，隠れマルコフモデル (Hidden Markov Model; HMM) を生成モデルとして利用することが多い [39]．以下，HMM とはどういったモデルであるかと，X の手動書き起こし付きの $\boldsymbol{O}$ が得られたときに $P(\boldsymbol{O}, X|\theta)$ を最大化する HMM のモデルパラメータ θ を学習する方法と，θ を学習した後，短時間音響特徴量時系列 $\boldsymbol{O}$ が得られたときにそれが X から生成されたとする事後確率 $P(X|\boldsymbol{O}, \theta)$ を計算する方法について述べる．

3.3.1　隠れマルコフモデル (HMM)

HMM の概念図を図 3.3 に示す．S_i は i 番目の状態，a_i は状態 S_i から状態 $S_i + 1$ への状態遷移確率，$b_i(\boldsymbol{o})$ は状態 S_i から短時間音響特徴量 $\boldsymbol{o}$ が出力される出力確率である．出力確率 $b_i(\boldsymbol{o})$ の分布形としては，ガウス分布を複数用意し，その重み付け和で $b_i(\boldsymbol{o})$ を表現する混合ガウス分布 (Gaussian Mixture Model; GMM) がよく用いられる．HMM のパラメータを θ とすると，HMM を用いることで，$P(\boldsymbol{O}|\theta)$ がモデルされることになる．

多くの場合，一つの音素につき一つの HMM を用いる．これは音素 HMM と呼ばれ，HMM のトポロジーとしては図 3.3 で示しているような 3 状態程度の left-to-right 型が多く利用される．音素 HMM により，音素を p として，$P(\boldsymbol{O}|p, \theta)$ がモデル化される．音素 HMM のうち，前後の音素環境を考慮しない音素 HMM を monophone，前後の音素環境を考慮する音素 HMM を triphone と呼ぶ．実際の音声は，調音結合と呼ばれる現象に より，音素の音響

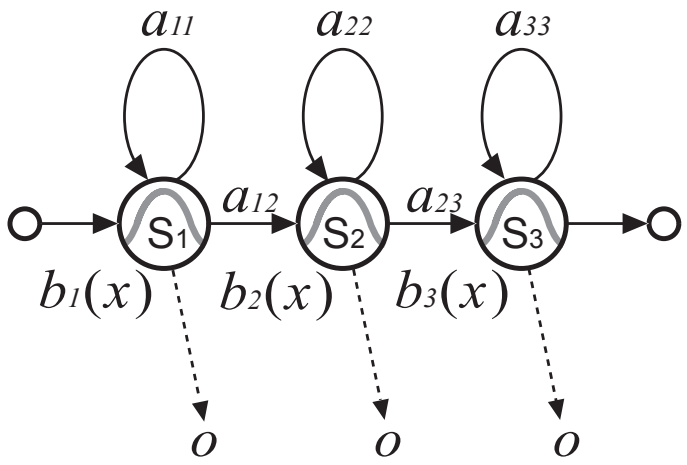

図 3.3　隠れマルコフモデル (HMM)

的特徴が前後の音素環境に依存して大きく変化する．そのため triphone を用いることで，音響モデリングを精緻にできる利点がある．ただし，triphone には HMM の数が膨大に増える欠点も存在する．そのため，音声認識タスクにおいては triphone が主に用いられているが，発音評価タスクでは monophone を用いることが多い．

3.3.2　HMM の学習

音素 p の書き起こし付きの $\boldsymbol{O}$ が得られたときに，HMM の学習すべきパラメータは $\theta = \{a_{ij}^P, b_i^P(\boldsymbol{o})\}$ であり，これを生成モデルとして最尤 (Maximum Likelihood; ML) 推定することを考える．これは

$$\underset{\theta}{\operatorname{argmax}}\, P(\boldsymbol{O}, p|\theta) = \underset{\theta}{\operatorname{argmax}}\, \frac{P(\boldsymbol{O}|p,\theta)}{P(p|\theta)} \tag{3.6}$$

$$= \underset{\theta}{\operatorname{argmax}}\, P(\boldsymbol{O}|p,\theta) \tag{3.7}$$

を解くことで行われる．しかし，HMM の現在の状態 z は隠れ変数であり，式 (3.7) を解析的に解くことは不可能である．そのため，隠れ変数が存在する統計モデルの ML 推定値を一般的に見出すことができる Expectation-Maximization (EM) アルゴリズムを用いて式 (3.7) の局所最適解を得る．ここで EM アルゴリズムの収束結果は初期値に依存するため，パラメータの初期値設定法は重要である．初期値設定法としては，全音素を同一として HMM を学習し，そのパラメータをすべての HMM の初期値として利用するフラットスタートや，triphone の初期値に monophone のパラメータを用いる手法などが利用される．なお，EM アルゴリズムの実装には，HMM の EM 学習に特化して効率を高めたアルゴリズムである Baum-Welch アルゴリズムを利用することができる．

3.3.3　HMM を用いた音素認識

すべての音素 HMM が学習され，そのパラメータを θ とする．その上で，短時間音響特徴量の時系列 $\boldsymbol{O}$ が観測されたときに，それが音素 p の音素 HMM から生成された事後確率 $P(p|\boldsymbol{O},\theta)$ を計算することを考える．これを計算し，尤度の最も高い音素 p を出力することは，音声認識をすることにほかならない．ベイズの定理より

$$\begin{aligned}\underset{p}{\operatorname{argmax}} P(P|\boldsymbol{O},\theta) &= \underset{p}{\operatorname{argmax}} \frac{P(p,\boldsymbol{O}|\theta)}{P(\boldsymbol{O}|\theta)} && (3.8)\\ &= \underset{p}{\operatorname{argmax}} P(p,\boldsymbol{O}|\theta) && (3.9)\\ &= \underset{p}{\operatorname{argmax}} P(\boldsymbol{O}|p,\theta)P(p|\theta) && (3.10)\end{aligned}$$

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

$$\underset{p}{\operatorname{argmax}}\, p(\boldsymbol{O}|p,\theta) \tag{3.11}$$

$P(\boldsymbol{O}|p,\theta)$ は HMM でモデル化されているので，これで音素認識問題を解くことができる． 実際に $P(\boldsymbol{O}|p,\theta)$ を計算するためには，HMM の隠れ変数である HMM の状態がどのように遷移したかを考慮する必要がある．例として，音声特徴量の時系列データ $\boldsymbol{O} = \{\boldsymbol{o}(1),\boldsymbol{o}(2),\ldots,\boldsymbol{o}(7)\}$ が音素 P に対応する音素 HMM から出力される場合の，可能な状態遷移の経路を図 3.4 に示す．ある 1 つの経路を通って $\boldsymbol{O}$ が出力される確率は，その経路の状態遷移確率 a_i と経路上の各状態での出力確率 $b_i(\boldsymbol{o})$ の積によって計算できる．図 3.4 に示された経路全てに対してこの確率を求めて和をとることで，音素 p の HMM から音響特徴量時系列 $\boldsymbol{O}$ が出力される確率，すなわち $P(\boldsymbol{O}|p,\theta)$ を求めることができる．しかし， 全ての経路からの出力確率の和をとると計算量が増大してしまうため，実際には最も出力確率の大きな経路のみを計算し，その確率値で $P(\boldsymbol{O}|p,\theta)$ を近似する Viterbi アルゴリズ ムが用いられる．このような近似は Viterbi 近似と呼ばれ，実験的に $P(\boldsymbol{O}|p,\theta)$ の非常によい近似となっていることが知られている．

3.4　GOP (Goodness of Pronunciation)

生徒の発声から音素 p を意図して発声された音響特徴量時系列 $\boldsymbol{O}^{(p)}$ が観測され，それが本来発声されるべき音素 p としてどれだけ良く発声されたかを示すスコアとして GOP (Goodness of Pronunciation) が広く用いられている [36]．GOP の計算に用いる HMM のパラメータを θ として，GOP は以下の式で定義される．

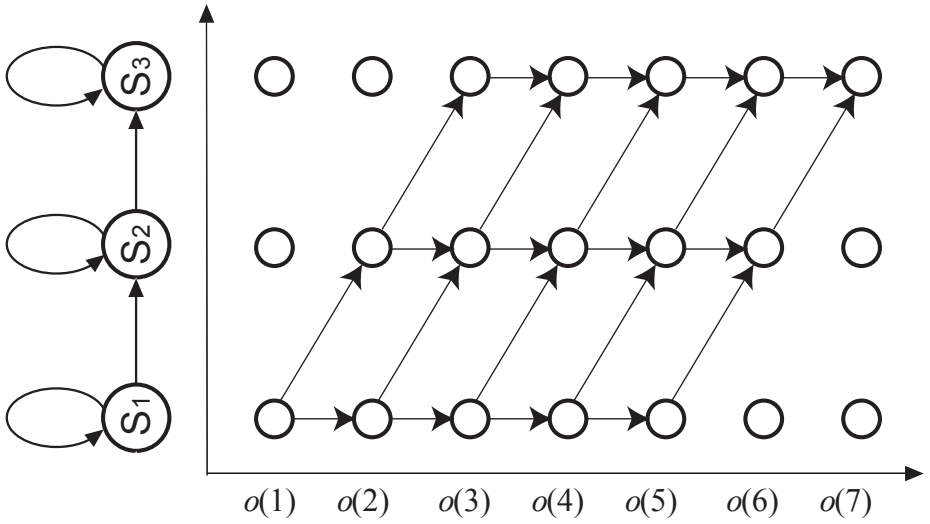

図 3.4　HMM の状態遷移の経路

$$GOP(\boldsymbol{O}^{(p)}, p, \theta) = \frac{1}{D_p} \log P(p|\boldsymbol{O}^{(p)}, \theta) \tag{3.12}$$

ここで，D_p は，$\boldsymbol{O}^{(p)}$ の継続長である．すなわち，GOP は，音響特徴量時系列 $\boldsymbol{O}$ が観測されたときに，それが p として正しく発声された事後確率を当該発声の継続長で正規化したものと定義される．GOP は，以下のように書き下せる．

$$\frac{1}{D_p} \log P(p|\boldsymbol{O}^{(p)}, \theta) = \frac{1}{D_p} \log \frac{P(\boldsymbol{O}^{(p)}|p, \theta)P(p)}{P(\boldsymbol{O}^{(p)}|\theta)} \tag{3.13}$$

$$= \frac{1}{D_p} \log \frac{P(\boldsymbol{O}^{(p)}|p, \theta)P(p)}{\sum_{q \in Q} P(\boldsymbol{O}^{(p)}|q)P(q)} \tag{3.14}$$

ここで，Q は考慮している全音素である．なお，式 (3.13) の計算には，発声された音素が既知であるため $P(p|\theta) = 1$ が成立することを用いている．式 (3.14) を見ると，分子のスコア が分母のスコアで正規化される形になっている．すなわち，HMM の学習データと学習者 の音声で多少のミスマッチがあったとしても，およそキャンセルできるスコア手法となっ ている．

3.5　GOP の計算

GOP を計算するためには式 (3.14) を計算する必要があるが，これを直接すると計算量が膨大になる．そこで，$P(p) = P(q) = const.$ という仮定（3.3.3 を参照）および，分母は最大値で近似されるという仮定を置いて，以下のような近似を行う．

$$\frac{1}{D_p}\log\frac{P(\boldsymbol{O}^{(p)}|p,\theta)P(p)}{\sum_{q\in Q}P(\boldsymbol{O}^{(p)}|q)P(q)} \approx \frac{1}{D_p}\log\frac{P(\boldsymbol{O}^{(p)}|p,\theta)}{\max_{q\in Q}P(\boldsymbol{O}^{(p)}|q)} \tag{3.15}$$

分子は HMM による強制 Viterbi アライメントにより算出する．分母は連続音素認識による尤度を用いることができる．以上のようにして，GOP を近似的に計算することが可能になる．

3.6　シャドーイング

シャドーイングは，聴取した外国語音声を即座に繰り返して発声することで発音能力と聴取能力とを同時に鍛える外国語聴取・発音訓練法である．もともとは同時通訳者の訓練として広く行われていたが，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ている [23–25]．シャドーイングにおいては，学習者が提示音声をそのまま真似ることは難しく，学習者自身の話し方の癖や学習者の母語に関する言語知識が無意識のうちに使われることが知られている [26]．

3.7　GOP を用いたシャドーイング評価

本節では，羅らによる GOP を用いたシャドーイング評価について述べる [21]．

[21] では，式 (3.15) で定義した GOP を，発話を音素単位に区切った音声セグメントごとに計算し，それらを発話した文章全体にわたって平均したもの（本論文中では GOP_{all} と呼ぶことにする）を，発話に対する自動評価スコアとしている．そして，GOP_{all} と発話者の TOEIC スコアとの相関を調べている．話者は，日本語を母語とする学習者，英語教師，米語を母語とする英語教師合わせて 27 名（TOEIC スコア 158 点–990 点），シャドーイングに用いた文章は 21 文からなる英文であり，発話者は聞こえてきた音声をできるだけ間を置かずに繰り返して発話するように指示されている．結果として，GOP_{all} と発話者の TOEIC スコアとの相関係数としては，0.82 という値を得ている．

3.8　まとめ

本章では，シャドーイング評価の先行研究として，羅らによる GOP を用いたシャドーイング自動評価について取り上げた [21]．本章ではまず，GOP を算出するための要素技術として，短時間音響特徴量として利用されるメルケプストラム (MFCC) と，音響モデルとして利用される隠れマルコフモデル (HMM) について詳細な説明を行った．その後，具体的に GOP を算出する方法について説明し，最後に，羅らによるシャドーイング自動評価について述べた．

次章では，文強勢検出の先行研究をいくつか紹介するとともに，各手法の長短について述べる．

第4章

文強勢検出の先行研究

4.1　はじめに

本章では，文強勢検出の先行研究をいくつか紹介するとともに，各手法の長短について述べる（文強勢検出の研究を見つけられなかった場合は，単語強勢検出となっている）．まずは，最も古く，単純な方法であるルールベースの手法を紹介する．次に，評価対象音声とモデル音声を時間軸上で伸縮・対応させた上で比較する，DTW による手法を紹介する．DTW はモデル音声と同じ内容の発声に対してしか評価を行えないため拡張性に乏しい．そこで，モデルを確率的に一般化して表現する手法として，HMM を用いた手法を紹介する．最後に，最新の手法の一つとして，識別モデルの一つである SVM (Support Vector Machine) を用いた手法を紹介する．

4.2　ルールベースによる強勢検出

ルールベースの手法では，入力音声から何らかの音響特徴量を抽出し，それらに対してルールを適用することにより，各音節の強勢判定を行う．ルールは音声学的知見や，試行錯誤により定義される．

4.2.1　Hieronymus の文強勢検出手法 [40]

Hieronymus はルールベースによる文強勢検出を行っている．[40] では，以下の手順で強勢判定を行う．

1. 基本周波数，エネルギー，継続長を音響特徴量として用いて，それぞれルールにより強勢判定を行う．
2. 1. の結果を統合して，最終的に強勢／弱勢を判定する．

基本周波数による強勢判定　まず，有声区間について基本周波数を測定し（無声部分には存在しない），スムージングを行う．次に，各有声区間ごとに以下を計算する．

- 最大基本周波数
- 平均基本周波数
- 区間の始点，終点における基本周波数
- 区間における基本周波数の傾き

これらの値を有限状態機械に入力し，出力として nothing accented，pitch accented，highest pitch accented のいずれかを得る．基本周波数パターンと対応する出力の例を図 4.1 に示す．

エネルギーによる強勢判定　まず，母音区間についてエネルギーを計算する．母音区間は手動でラベリングされている．次に，単語中の最大エネルギーから 11 db 以内の母音は stressed，20 db 以上小さい母音は，基本周波数および継続長による強勢判定によらず，全体として強勢とは判定されないものとする．

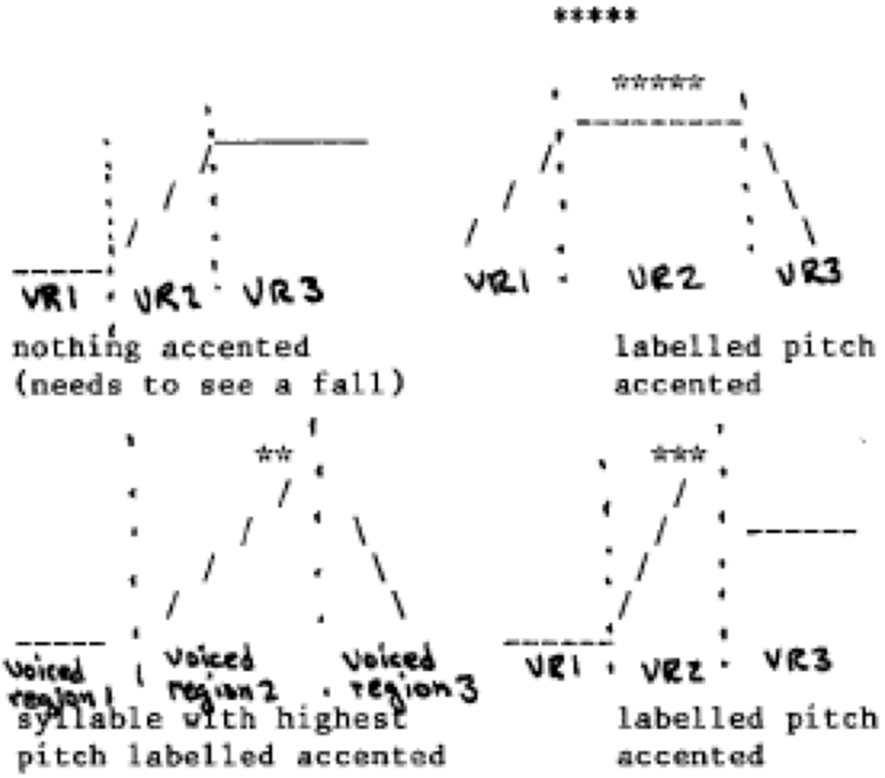

図 4.1　基本周波数のパターンと対応する強勢ラベルの例

継続長による強勢判定　継続長による判定は，単純に絶対値を比較する方法では，話速に左右されるので困難である（ゆっくり話せば，強勢の有無に関わらず継続長は長くなる）．そこで，継続長のヒストグラムに関して正規化処理を行った後，最大の継続長を持つものを選び，その継続長が 60 ms 以上のものを stressed と判定する．

最終的な強勢判定　基本周波数，エネルギー，継続長の合計 3 種類の強勢判定の結果のうち，2 種類以上で強勢と判定されたものを強勢と判定する．特に，3 種類全てで強勢と判定されたものは，最大強勢とする．結果として，67–75 %の正解率を得ている．

4.2.2　長短

長所としては，正解のパターンが不要であり，任意の発話に対して判定を行えることがあげられる．また，一般に計算量は小さいと考えられる．短所としては，適切なルールの定義が困難であることがあげられる．また，後述の手法に比べて正解率がやや低くなる傾向にある．

4.3　DTW による強勢検出

発音評価を行う場合，評価対象音声をモデル音声と照合し，モデル音声に近いほど高い評価を与えることが考えられる．ところが，両音声を単純に比較することはできない．音声の時系列パターンは，同じ話者が発声したものであっても発声のたびに継続長が変わるし，異なる話者が発声した場合はなおさらであるからである．しかもパターンは，時間的に一様に伸縮するわけではない．発話速度が変化しても，子音部，あるいは子音から母音への過渡部の継続長はある程度一定の長さを保っていて比較的変化が少ないが，母音部の継続長は大幅に伸縮する．したがって，時系列パターンの全区間では，非線形の時間伸縮が起こることになる．そこで，DTW（Dynamic Time Warping，動的計画法，DP マッチング）と呼ばれる，2 つの入力パターン間を伸縮を伴いながら照合し，パターン間の距離を求める方法を用い，両音声の時系列を揃えてやることが考えられる．DTW を用いた強勢判定では，同一単語についての入力パターン（評価対象音声）と参照パターン（モデル音声）のマッチングを行い，その結果を利用して強勢判定を行うことになる．

4.3.1　Arias らの単語強勢検出手法 [41]

Arias らは，DTW と相関を用いた単語強勢検出手法を提案している．この手法では，以下の手順で強勢検出を行う．

1. 音響特徴量として，基本周波数，エネルギー，MFCC を抽出する．基本周波数は正規化を行う．
2. DTW を用い，入力パターン（学習者の音声）と参照パターン（モデル音声）の時間方向のマッチングを行う
3. 入力パターンと参照パターンの間で，基本周波数，エネルギー時系列の相関を計算する
4. 閾値判定により強勢判定を行う

音響特徴量の抽出　音響特徴量として，基本周波数，エネルギー，MFCC をフレームごとに計算する．基本周波数とエネルギーは対数尺度に変換する．

DTW によるマッチング　入力パターン（学習者の音声）と参照パターン（モデル音声）は同一単語の発声であるが，両パターンの間には当然時間方向のずれが生じている．そこで，特徴量には 33 次元の MFCC を用い，両パターンの MFCC のマハラノビス距離が最小となるようにマッチングを行う．

相関の計算　マッチングの結果を元に，両パターン間で基本周波数時系列の相関 $TS[F0^R_{PP}(t), F0^S_{PP}(t)]$，エネルギー時系列の相関 $TS[E^R(t), E^S(t)]$ を計算する．

$$TS[F0^R_{PP}(t), F0^S_{PP}(t)] = \frac{\sum_{k=1}^{T}\{F0^R_{PP}[i_n(k)] - \overline{F0^R_{PP}}\}\{F0^S_{PP}[i_n(k)] - \overline{F0^S_{PP}}\}}{\sigma_{F0^R_{PP}} \cdot \sigma_{F0^S_{PP}}} \tag{4.1}$$

$$TS[E^R(t), E^S(t)] = \frac{\sum_{k=1}^{T}\{E^R[i_n(k)] - \overline{E^R}\}\{E^S[i_n(k)] - \overline{E^S}\}}{\sigma_{E^R} \cdot \sigma_{E^S}} \tag{4.2}$$

そして，両者を重み付けした上で足しあわせて，強勢判定のためのスコアとする．

$$TS(F0^R_{PP}, E^R, F0^S_{PP}, E^S) = \alpha TS(E^R, E^S) + (1-\alpha)TS(F0^R_{PP}, F0^S_{PP}) \tag{4.3}$$

この $TS(F0^R_{PP}, E^R, F0^S_{PP}, E^S)$ を閾値判定し，強勢判定を行う．

4.3.2　長短

長所としては，後述する HMM を用いた手法や SVM を用いた手法のようにモデルの構築に学習を必要としないので,（単語ごとに見れば）モデル音声の量が少なくて良いことがあげられる．逆に言えば，単語ごとにモデル音声を用意する必要があるため，他の単語へそのまま拡張することが難しいことが，短所と言える．また，単語ごとに 1 つのモデル音声しか用意しない場合,「音声学的に正しいけれども，モデル音声とは異なる」強勢の付け方をしていた場合に，正しく検出できない可能性がある．これに対しては，1 つの単語に対して複数のモデル音声を用意するなどの対処が考えられる．

4.4　HMM による強勢検出

上で述べた 2 つの手法に対して，HMM を用いた確率的な手法を用いて強勢を検出する手法がある．

4.4.1　小橋川らの文強勢検出手法 [29, 30]

小橋川らは，HMM を用いた文強勢検出手法を提案している．この手法では，以下の手順で文強勢検出を行う．

1. 学習データおよび評価データから音響特徴量として，基本周波数，エネルギー，MFCC，およびそれらの Δ を抽出する．基本周波数およびエネルギーは正規化を行う．
2. 1. で抽出した特徴量を用いて，強勢／弱勢を HMM でモデル化する．この際，音節のフレーズ内の位置などを考慮した複数の HMM を学習する．
3. 2. で学習した強勢／弱勢 HMM を用いて，評価対象音声の各音節に対して強勢／弱勢の判定を行う．

音響特徴量の抽出　学習データおよび評価データから音響特徴量として，基本周波数，エネルギー，MFCC（1〜4 次元），およびそれらの Δ をフレームごとに抽出する．基本周波数およびエネルギーは当該発声の平均値を引く形で正規化を行う．

HMM による強勢／弱勢のモデル化　学習データから抽出した音響特徴量を元に，強勢／弱勢 HMM の学習を行う．学習データは，ERJ データベース [28] のリズム文を，米語母語話者（女性）1 名が正しい強勢パターンで読み上げた音声である．このとき，音節を単に強勢／弱勢で分類するモデルのほか，フレーズ内の位置（頭／中／末）を考慮したモデル，音節構造を考慮したモデル，母音区間のみの情報を用いたモデルなどを構築する．

強勢／弱勢の判定　学習した HMM モデルを用いて，評価対象音声の各音節に対して強勢／弱勢の判定を行う．評価データは，学習用音声と同一話者・異なるテキストの音声のほか，学習用音声以外の母語話者や，比較的英語発音能力が高いと考えられる日本語母語話者による異なるテキストの音声が用いられている．結果としては，評価音声セットにより約 78%–94%の正解率を得ている．

4.4.2　長短

長所としては，モデルを確率的に一般化して表現できるため，単語の拡張が容易であることがあげられる．短所としては，モデル構築のために大量の発声および強勢ラベルを必要とすることがあげられる．母語話者ではなく学習者（正しく発音していない可能性が高い）の音声を用いる場合は，手動で強勢ラベルを付与する必要があり，特に問題である．また，過学習の問題から識別モデルにくらべて特徴量の数を増やしづらいことがあげられる．

4.5　SVM による強勢検出

前節で述べた手法は生成モデルの一つである HMM を用いた手法であったが，識別モデルを用いた強勢検出手法も提案されている．本節では識別モデルの一つである SVM (Support Vector Machine) を用いた Shi らによる単語強勢検出手法を紹介するが，まずは SVM について解説する．

4.5.1　SVM

SVM (Support Vector Mashine) は，識別モデルの一つである．入力ベクトル $\boldsymbol{x}$ とそれに対応するラベル t（1 または –1 の二値をとる）の組が学習データとして与えられたときに，未知の入力ベクトルに対する t を推定するタスクを考える．このとき，

$$y(\boldsymbol{x}) = \boldsymbol{w}^{\mathrm{T}}\phi(\boldsymbol{x}) + b \tag{4.4}$$

で定義される線形モデルを用いて二値分類問題を解くことにする．ここで $\boldsymbol{w}$ は入力データに対する重みベクトル，$\phi(\boldsymbol{x})$ は基底関数，b はバイアス項である．学習データは $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$

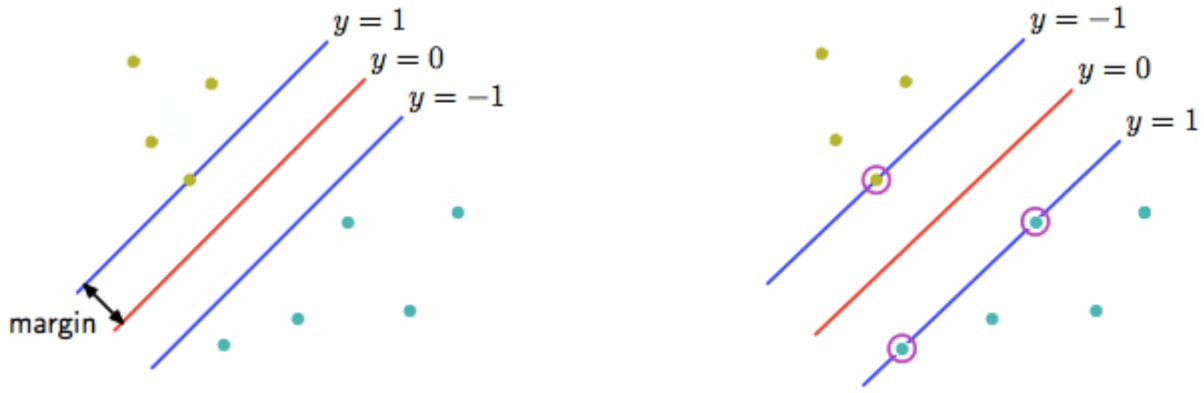

図 4.2　マージン (margin) の概念図．マージンは，左図のように分類境界 ($y = 0$) と最も近くのデータまでの距離のことである．マージン最大化という基準により，マージンは一意に定まる．またそのとき，分類境界の位置は右図で丸で囲んだような一部のデータ (support vectors) によって決まる．

の N 個の学習データと，それぞれに対応するラベル $t_1, \ldots, t_N (t_n \in \{-1, 1\})$ からなり，未知のデータ x は $y(\boldsymbol{x})$ の符号に応じて分類されるものとする．

このとき，学習データが特徴量空間で線形分離可能，すなわちある一組のデータ $\boldsymbol{w}$ と b に対して，$t_n = 1$ である点に対しては $y(\boldsymbol{x}_n) > 0$，$t_n = -1$ である点に対しては $y(\boldsymbol{x}_n) < 0$ が成り立つと仮定する．これらの条件は $t_n y(\boldsymbol{x}_n) > 0$ と表記できる．

一般に学習データが線形分離可能であれば，正確に分離できる解は多数存在するが，未知のデータに対してもできるだけ正しく分離するような解が望ましい．SVM では，このような解を**マージン** (margin) という概念を用いて求めようとする．マージンとは，図 **??**に示すような，分離境界と学習データの間の最短距離のことを指す．

SVM においては，マージンを最大化するような分類境界が選ばれる．詳細は省略するが，そのような解は，

$$\underset{\boldsymbol{w},b}{\operatorname{argmin}} \frac{1}{2}||\boldsymbol{w}||^2 \tag{4.5}$$

を，

$$t_n(\boldsymbol{w}^{\mathrm{T}}\phi(\boldsymbol{x}_n) + b) \geq 1 \tag{4.6}$$

の制約条件のもとで解くことで得られることが知られている．

ところで，上記の議論では学習データが特徴量空間で線形分離可能であることを仮定したが，現実には学習データが線形分離不可能である場合もある．そのような場合には，式 4.4 をカーネル関数というものにより書き換えることで対処可能となる．カーネル関数には，ガウスカーネル，多項式カーネルなどがある．

また，SVM は本来二値分類器であるが，複数の SVM の組み合わせにより多値分類を可能にした多クラス SVM や，SVM を回帰に応用した SVR (Support Vector Regression) といったものも開発されている．

4.5.2　Kim らの単語強勢検出手法 [42]

Kim らは，複数の手法による単語強勢検出の性能を比較する論文において，SVM を用いた単語強勢検出の実験を行っている．この手法では，以下の手順で単語強勢検出を行う．

1. 学習データおよび評価データから音響特徴量として，音節ごとに，母音の平均基本周波数，母音の平均エネルギー，音素の継続長を抽出する．これらを正規化したものも特徴量とする．
2. 1. で抽出した特徴量を用いて，SVM の学習を行う．
3. 2. で学習した SVM を用いて，単語強勢検出を行う．

音響特徴量の抽出　学習データおよび評価データから音響特徴量として，音節ごとに，母音の平均基本周波数，母音の平均エネルギー，音素の継続長を抽出する．これらを音素ごとに平均が 0，分散が 1 となるように正規化したものも特徴量とする．

SVM の学習　抽出した音響特徴量を用いて，SVM の学習を行う．学習・評価に用いているデータは英語母語話者（女性）1 名の単語発声であり，10 分割交差検定を行っている．単語は 3 音節または 4 音節からなっており，数としてはおよそ 14,000 単語にのぼる．強勢は，第一強勢，第二強勢，弱勢の 3 段階である．

単語強勢検出　学習した SVM を用いて，単語強勢検出を行う．結果としては，3 音節語で 81.6%，4 音節語で 85.3%の正解率を得ている．

4.5.3　長短

長所としては，HMM と同様，モデルを確率的に一般化して表現できるため，単語の拡張が容易であることがあげられる．また，SVM は一般に高い性能を示すことが知られていること，そして，識別モデルであることから，特徴量を増やしても過学習を起こしにくいこともあげられる．短所としては，HMM と同様，モデル構築のために大量の発声および強勢ラベルを必要とすることがあげられる．

4.6　まとめ

本章では，文強勢検出の先行研究として，ルールベースによる手法，DTW による手法，HMM による手法，SVM による手法を紹介し，それぞれの手法の長短を述べた．本研究では，読み上げ音声の文強勢検出実験を行うにあたり，音響特徴量以外の特徴量を積極的に利用することから，特徴量を増やしても過学習を起こしにくく，先行研究でも高い性能が示されている SVM を使用することにする．

次章では，重回帰を用いたシャドーイング評価の高精度化について検討を行う．

第5章

重回帰を用いたシャドーイング評価の高精度化

5.1　はじめに

本章では，羅らの GOP を用いたシャドーイング自動評価手法 [21] について，重回帰を用いた高精度化を検討する．まず，実験に使用したシャドーイング音声およびその発話者，GOP の計算にあたっての音響分析条件について述べる．次に，重回帰の説明変数として用いる，GOP を様々な観点で分類・集計したスコアについて説明する．そして，重回帰を用いて発話者の TOEIC スコアを推定する実験とその結果について述べ，結果を考察する．

5.2　音声収録

日本語母語話者 54 名および米語母語話者 2 名の計 56 名にシャドーイングを行わせた．シャドーイングに用いた英文は 2 種類で，それぞれ文章 A と文章 B とする．文章 A は 21 文，文章 B は 14 文からなっている．発話者のうち文章 A のシャドーイングを行ったのは 27 名，文章 B のシャドーイングを行ったのは 29 名であり，両方の文章をシャドーイングした発話者はいなかった．発話者の TOEIC スコアを表 5.1 に示す．

5.3　音響分析条件

GOP の計算には，WSJ および TIMIT データベースから学習した，英語の音素を単位とする monophone HMM を用いた [43]．HMM の音響分析条件を表 5.2 に示す．

5.4　GOP の分類・集計

シャドーイング音声に対する自動評価スコアは，シャドーイング音声を音素単位に区切った音声セグメントごとに式 (3.12) により計算した $GOP(p)$ を，以下に示す基準により分類・集計することによって用意した．なお，発話した文章には n 個（n 種類ではない）の音素が含まれ，先頭から順に $p_1 p_2 \ldots p_n$ のように並んでいるものとする．また，音素 sp および sil（どちらも無音を表す）に対する GOP は自動評価スコアの計算には用いない．

5.4.1　GOP_{all}

$GOP(p)$ を発話した文章全体にわたって平均したもの．以下の式で定義される．

$$GOP_{all} = \frac{1}{n} \sum_{k=1}^{n} GOP(p_k) \tag{5.1}$$

GOP_{all} では文章中での出現回数が多い音素の GOP ほど影響が大きくなる．文章によって各音素の出現回数は異なるので，文章への依存度が高いスコアと言える．そのため文章によって TOEIC スコアの推定精度に大きな差が出てしまう可能性がある．

表 5.1　発話者の TOEIC スコア

文章	TOEIC スコア
文章 A	990, 990, 968, 955, 940, 895, 825, 625, 601, 592, 581, 512, 436, 432, 427, 421, 395, 367, 308, 301, 289, 278, 275, 252, 202, 197, 158
文章 B	805, 792, 778, 722, 721, 721, 707, 693, 679, 677, 665, 636, 622, 594, 594, 594, 580, 566, 552, 481, 424, 410, 396, 368, 325, 311, 311, 255, 226

表 5.2　HMM の音響分析条件

標本化周波数	16 kHz
量子化ビット	16 bit
窓および窓長	ハミング窓／ 25 ms
シフト長	10 ms
音響特徴量	MFCC12 次元および対数パワー，それらの Δ，ΔΔ（計 39 次元）
音素の種類	ɑ, æ, ʌ, ɔ, aʊ, aɪ, b, ʧ, d, ð, ɛ, ɚ, eɪ, f, g, h, ɪ, i, ʤ, k, l, m, n, ŋ, oʊ, ɔɪ, p, r, s, ʃ, t, θ, ʊ, uː, v, w, j, z, ʒ, sil, sp（合計 41 種類）

5.4.2　GOP_{vow} および GOP_{cons}

GOP_{vow} は，母音に対する $GOP(p)$ を発話した文章全体にわたって平均したもの，GOP_{cons} は子音に対して同様の計算を行ったものである．文章中に母音が n_{vow} 個，子音が n_{cons} 個含まれるとしたとき，それぞれ以下の式で定義される．

$$GOP_{vow} = \frac{1}{n_{vow}} \sum_{p_k \in vowel} GOP(p_k) \tag{5.2}$$

$$GOP_{cons} = \frac{1}{n_{cons}} \sum_{p_k \in consonant} GOP(p_k) \tag{5.3}$$

文章によって各音素の出現回数は異なるので，GOP_{all} と同じく文章への依存度が高いスコアと言える．

5.4.3　GOP_{phone}

$GOP(p)$ を音素の種類ごとに平均したもの．例えば音素/ɑ/が文章中に $n_{\mathrm{ɑ}}$ 個含まれるとしたとき，$GOP_{\mathrm{ɑ}}$ は以下の式で定義される．

$$GOP_{\mathrm{ɑ}} = \frac{1}{n_{\mathrm{ɑ}}} \sum_{p_k = \mathrm{ɑ}} GOP(p_k) \tag{5.4}$$

文章中の当該音素の出現回数で正規化されるため，文章への依存度が低いスコアと考えられる．そのため GOP_{phone} を説明変数とした回帰式は，文章によらず適用できる可能性がある．

5.5　TOEIC スコアの推定

5.5.1　タスク

5.4 で定義した自動評価スコアをもとに線形回帰を行い，学習者の TOEIC スコアを推定した．推定タスクは以下の 4 種類を行った．

1. 文章 closed, speaker-open（文章 A）
2. 文章 closed, speaker-open（文章 B）
3. 文章 open, speaker-open（文章 A で学習した回帰式で文章 B の被験者の TOEIC スコアを推定）
4. 文章 open, speaker-open（文章 B で学習した回帰式で文章 A の被験者の TOEIC スコアを推定）

1. および 2. は，各文章の全サンプルから 1 話者のサンプルを除外したセットで学習した回帰式を用いて，除外した 1 話者の TOEIC スコアを推定する作業を，全話者について行うものである．3. および 4. については，一方の文章のサンプルで学習した回帰式を用いて，他方の文章の話者の TOEIC スコアを推定するものである．

推定精度の評価は，回帰式による TOEIC スコアの推定値と，実際の TOEIC スコアの相関係数を求めることにより行った．

5.5.2　回帰の種類および説明変数

回帰は最小二乗法による線形単回帰または線形重回帰を行った．説明変数（の組）としては以下の 4 つを用いた．

- GOP_{all}（従来手法）
- $\boldsymbol{GOP_{v+c}} = [GOP_{vow}, GOP_{cons}]^\top$
- $\boldsymbol{GOP_{each\_v}} = [GOP_{\mathrm{ɑ}}, GOP_{\mathrm{æ}}, \ldots, GOP_{\mathrm{u}}]^\top$（各母音種類に対する GOP）
- $\boldsymbol{GOP_{each\_c}} = [GOP_{\mathrm{b}}, GOP_{\mathrm{ʧ}}, \ldots, GOP_{\mathrm{ʒ}}]^\top$（各子音種類に対する GOP）

各音素種類（母音，子音とも）に対する GOP を説明変数の組として用いることは,「サンプル数 $\leq$ 音素の種類」となってしまうため行わなかった．

なお，$\boldsymbol{GOP_{each\_v}}$ および $\boldsymbol{GOP_{each\_c}}$ に対しては，二次の正則化項を導入したリッジ線形回帰も行った．リッジ回帰の正則化パラメータ λ は，予備実験により $\lambda = 50$ ($\boldsymbol{GOP_{each\_v}}$)，$\lambda = 20$ ($\boldsymbol{GOP_{each\_c}}$) に固定した．また，各説明変数は平均 0，分散 1 とする正規化を行い，バイアス項に対する重みについては正則化を行わなかった．

5.5.3　従来手法と提案手法の関係

GOP_{all} を説明変数とする単回帰で TOEIC スコアを推定する従来手法の回帰式は，各音素種類に対する GOP を用いて以下のように解釈できる．

表 5.3　TOEIC スコアの推定値と実際の TOEIC スコアの相関係数 (1)

タスク	GOP_{all}	$\boldsymbol{GOP_{v+c}}$	$\boldsymbol{GOP_{each\_v}}$	$\boldsymbol{GOP_{each\_v}}$（リッジ回帰）
文章 closed, speaker-open（文章 A）	0.789	0.772	0.772	0.810
文章 closed, speaker-open（文章 B）	0.694	0.716	0.679	0.758
文章 open, speaker-open（文章 A → B）	0.731	0.766	0.574	0.754
文章 open, speaker-open（文章 B → A）	0.818	0.826	0.582	0.793

表 5.4　TOEIC スコアの推定値と実際の TOEIC スコアの相関係数 (2)

タスク	$\boldsymbol{GOP_{each\_c}}$	$\boldsymbol{GOP_{each\_c}}$（リッジ回帰）
文章 closed, speaker-open（文章 A）	0.137	0.783
文章 closed, speaker-open（文章 B）	0.287	0.751
文章 open, speaker-open（文章 A → B）	0.116	0.629
文章 open, speaker-open（文章 B → A）	0.590	0.805

$$\hat{TOEIC} = wGOP_{all} + const. \tag{5.5}$$

$$= w\frac{n_{\text{ɑː}}}{n}GOP_{\text{ɑː}} + w\frac{n_{\text{æ}}}{n}GOP_{\text{æ}} + \ldots + w\frac{n_{\text{ʒ}}}{n}GOP_{\text{ʒ}} + const. \tag{5.6}$$

$$(n = n_{\text{ɑː}} + n_{\text{æ}} + \cdots + n_{\text{ʒ}})$$

式 (5.5) では，発話した文章全体から求められた GOP_{all} に対して重み w がかけられてているが，式 (5.6) の解釈では，各音素種類に対する GOP に対して，w に当該音素の文章中での出現割合を掛けたものが重みとしてかけられている．すなわち，GOP_{all} は，各音素種類に対する GOP に出現回数に比例した重みづけをして足しあわせたものであると解釈できる．

提案手法は，GOP_{all} が行っているように各音素種類に対する GOP に出現回数に比例した重みづけをするよりも，より高精度に TOEIC スコアを推定できるような重みづけを学習することを意図したものである．また，$\boldsymbol{GOP_{each\_v}}$ および $\boldsymbol{GOP_{each\_c}}$ は文章への依存度が低いと考えられるスコアであるため，文章によらず高精度に TOEIC スコアを推定できる可能性がある．

5.6　結果

実験結果を表 5.3 および表 5.4 に示す．$\boldsymbol{GOP_{v+c}}$ については，文章 closed, speaker-open（文章 A）において GOP_{all} と比べて相関が少し下がったものの，他のタスクにおいては相

関が上がった．$\boldsymbol{GOP_{each\_v}}$ については，最小二乗法による重回帰では文章 closed, speaker-open（文章 A）以外のタスクにおいて GOP_{all} と比べて相関が下がったものの，リッジ回帰では文章 open, speaker-open（文章 B → A）においてなお GOP_{all} と比べて相関がやや下回っているが，他のタスクにおいては相関が上がった．$\boldsymbol{GOP_{each\_c}}$ については，最小二乗法による重回帰ではすべてのタスクにおいて GOP_{all} と比べて相関が大きく下がったものの，リッジ回帰では文章 open, speaker-open（文章 A → B）においてなお GOP_{all} と比べて相関が下回っているが，他のタスクにおいては同程度かやや上回った．

5.7　考察

$\boldsymbol{GOP_{each\_v}}$ または $\boldsymbol{GOP_{each\_c}}$ を説明変数とした最小二乗法による重回帰を除いては，多くの場合で GOP_{all} と比べて相関が上がった．説明変数の数を増やすことで，よりよい重みづけができたものと考えられる．$\boldsymbol{GOP_{each\_v}}$ および $\boldsymbol{GOP_{each\_c}}$ が最小二乗法の場合に相関が下がったのは，話者数に対する説明変数の数が他に比べて相対的に多く，過学習が発生したからであろう．

なお，文章 open のタスクは，文章への依存度が高いと考えられる GOP_{all} や $\boldsymbol{GOP_{v+c}}$ を説明変数とした場合でも，使用した両文章で既に 0.7–0.8 程度と高い相関が見られた．そして両文章とも，文章への依存度が低いと考えられる $\boldsymbol{GOP_{each\_v}}$ や $\boldsymbol{GOP_{each\_c}}$ を説明変数としても大幅に相関係数が上がることはなかった．実験に用いた文章がたまたま両文章とも TOEIC スコアを高精度に推定できるような文章だったのか，どんな文章を用いても TOEIC スコアを高精度に推定できるのか，あるいはどのような文章なら TOEIC スコアを高精度に推定できるのかは今回の実験結果からは不明である．

5.8　まとめ

本章では，シャドーイング音声を自動評価し話者の TOEIC スコアを推定する手法の高精度化について検討した．従来の，発話から GOP_{all} という 1 つの自動評価スコアを説明変数とした線形単回帰を行う手法を発展させ，GOP に基づいた複数の自動評価スコアを計算し，それらを説明変数とする線形重回帰を行うことにより，話者の TOEIC スコアを単回帰の場合よりも高精度に推定することを検討した．また，文章への依存度が低いと考えられるスコアを用いて，文章によらず適用できる回帰式を推定することを検討した．結果としては，説明変数に用いるスコアによっては，重回帰（最小二乗法あるいはリッジ回帰）によって単回帰の場合よりも高精度に TOEIC スコアを推定することができた．一方．実験には 2 種類の文章を用いたが，文章への依存度が高いと考えられるスコアを説明変数とした場合でも，両文章間で TOEIC スコアを推定するタスクの精度に大差はなかった．そして，文章への依存度が低いと考えられるスコアを説明変数としても大幅に相関係数が上がることはなかった．

本章では GOP という音素に基づくスコアの範疇で重回帰を行ったが，第 7 章では，文強勢検出に基づくスコアを追加して評価のさらなる高精度化を目指す．しかし，学習者の

シャドーイング音声は非常に崩れたものになりがちで，直接文強勢を検出することは難しい．そのため，次章では読み上げ音声に対する文強勢検出実験を行い，その技術を第 7 章でシャドーイング評価に応用することにする．

第6章

読み上げ音声における文強勢検出

6.1　はじめに

本章では，読み上げ音声に対する文強勢検出実験を行う．評価対象音声には,「日本人学生による読み上げ英語音声データベース」（ERJ (English Read by Japanese) データベース）[28] の「文強勢，文リズムに関する文」（以下リズム文とする）に対する読み上げ音声を用いた．この音声の特徴は，英語の強勢・リズムに着目して構成された文を，原稿に記載された強勢記号に従って学習者自身が読み上げたもので，学習者自身は正しく読めたと考えている音声となっている点である．この音声を使用したのは，読み上げ文が英語の強勢・リズムに着目して作られている点，学習者自身が正しく読めたと考えている音声である点で教育的な観点から評価対象としてふさわしいと考えたからである．そして，4.5.1 で説明した SVM (Support Vector Machine) を用いてより精度の高い文強勢の自動検出を目指す．具体的には，評価対象音声に対して音節ごとに強勢の度合いを高精度に自動推定することを目的とする．ただし，ERJ には音節ごとに強勢の度合いを判定したラベルは付与されていないため，手動判定という形でラベルを付与した．また，本実験では読み上げ音声を使用するため，評価対象音声の発話内容は既知である．そこで，音響特徴量以外の，テキストから得られる特徴量も積極的に使って推定精度を上げることを目指した．

6.2　音声試料

評価対象の音声試料は，ERJ のリズム文である．ERJ は，日本語を母語とする大学生・大学院生相当の学生による読み上げ音声データベースである．リズム文の読み上げ原稿には，音節ごとに強勢の度合いを示す記号が 3 段階（弱勢，強勢，核強勢）で付けられている．ERJ の音声収録にあたっては，発声者は事前の発声練習が許されており，収録も発声者自身が「正しい」と判断できる発声が得られるまで繰り返されている．

ERJ のリズム文は全部で 120 文あるが，そのうち本研究では，共同研究者である英語教師が「発声者によって出来に差がつきやすい」と判断し選んだ 20 文を使用した．この 20 文について，すべての発声を評価対象とした．結果，評価対象となったのは 684 発声（7,414 音節），話者 202 名（男性 100 名，女性 102 名）である．

以下の実験においては，全 7,414 音節を話者および発声対象文が重複しないようにほぼ 3 等分に分割し，2 つを学習データ，残りの 1 つをテストデータとした．

6.3　手動判定による強勢ラベリング

6.3.1　評価者

評価者は 2 名である．英語音声学を専攻している大学院博士課程の学生（以下評価者 A）と英語教師（7.3 の英語教師とは別人．以下評価者 B）である．両者は事前に判定基準の統一をとることなく，それぞれ独立に判定を行った．

表 6.1　手動判定の結果（3 段階）

評価	0（発声されていない）	1–3（弱勢）	4–6（強勢）	7–9（核強勢）
評価者 A	80	4,128	2,776	431
評価者 B	86	5,248	1,986	1,096
（原稿の強勢記号）	0	4,693	1,342	1,379

表 6.2　手動判定の結果（9 段階）

評価	0	1	2	3	4	5	6	7	8	9
評価者 A	19	86	2,438	1,639	1,262	1,189	347	181	243	10
評価者 B	24	80	3,847	358	488	654	858	771	323	11

表 6.3　評価者内の判定の一致率および相関（3 段階）

評価者	一致率	相関
評価者 A	0.865	0.793
評価者 B	0.860	0.868

表 6.4　評価者内の判定の一致率および相関（9 段階）

評価者	一致率	相関
評価者 A	0.618	0.869
評価者 B	0.651	0.911

6.3.2　手動判定タスク

実験で用いる発声を聴取して，音節ごとに強勢の度合いを判定した．音声収録時の強勢記号は 3 段階（弱勢，強勢，核強勢）であるので，各々に対して，より弱い／適切／より強い，の 3 段階を考慮し，合計 9 段階とした（1 が最も弱く，9 が最も強い）．また，「発声されていない」ことを示す 0 を加えた．以下，9 段階を 3 段階に丸めて用いる（1 から 3 は「弱勢」としてひとくくりにする．以下同様）場合を「3 段階」，9 段階をそのまま用いる場合を「9 段階」と呼ぶことにする．

全 7,414 音節に対する手動判定の結果を表 6.1 および表 6.2 に示す．参考のため，収録時に読み上げ原稿に呈示された強勢記号の個数を表 6.1 に併記しておく．

また，手動判定の安定性を検証するために，評価対象である 684 発声のうち 64 発声を選び，数週間の時間を置いて再判定を行った．再判定対象発声の選定は，1 回目の判定順（684 回の判定作業で何番目に評価を行ったか）のバランスを考慮して，準ランダムに行った．判定条件は 1 回目と同様であるが，1 回目に自分が下した判定は参照させていない．1 回目の判定と再判定の一致率および相関を表 6.3 および表 6.4 に示す．

両評価者は事前に判定基準の打ち合わせ無しに判定を行ったが，評価者間の（1 回目の）

表 6.5　評価者間の判定の一致率および相関（3 段階）

一致率	相関
0.775	0.776

表 6.6　評価者間の判定の一致率および相関（9 段階）

一致率	相関
0.436	0.829

判定の一致率および相関を表 6.5 および表 6.6 に示す．SVM を用いた自動評価を試みるが，その精度評価は，表 6.3 から表 6.6 が比較対象となる．

6.4　HMM による手法を用いた実験

4.4.1 で，先行研究として HMM を用いた手法を紹介した [29,30]．本研究ではベースラインとして，[29,30] と同様の手法を実装し，実験を行った．なお，本節での実験は 3 段階についてのみ行った．

[29,30] では，音節の強勢／弱勢を HMM でモデル化し，入力文発声の各音節に対して HMM を用いて強勢／弱勢の判定を行っている．本研究でもこれにならい，弱勢／強勢／核強勢の 3 クラスのモデルを構築した（手動評価が 0 であったものについては，本実験では使用しなかった）．特徴量は SPTK を用いて抽出した [46]．音響分析条件を表 6.7，HMM の学習条件を表 6.8 に示す．

特徴量も [30] にならった．F_0 は，SPTK によって抽出した F_0 を母音区間だけ残して対数をとったものを線形補間する形で求めた．文頭（末）の無声区間の F_0 は，母音区間の F_0 の対数をとったものから平均 F_0 と標準偏差 σ を求め，母音区間から平均音節時間長前（後）に，平均 $F_0 - 3\sigma$ の値（基本周波数生成過程モデルにおける基底周波数 F_b に相当）にあるものとして線形補間を行った．さらに，F_0 とエネルギーに対しては，当該発声の平均値を引く形で正規化を行ったものを最終的な特徴量として用いた．

6.5　HMM による手法を用いた実験の結果

実験結果を表 6.9 に示す．評価者 A，評価者 B いずれのラベルを用いた場合でも，正解率はほぼ等しくなった．表 6.1 に示すように，強勢，弱勢の出現には数的偏りがある．先行研究ではこれらの事前分布を使わずに判定を行っており，表 6.9 の結果は，常に「弱勢である」と答えた場合よりも，正解率が劣る結果となった．

表 6.7　音響分析条件

標本化周波数	16 kHz
量子化ビット数	16 bit
抽出周期	8.0 ms (128 samples)

表 6.8　HMM の学習条件

特徴量	対数 $F_0 + \Delta$ 対数 F_0, エネルギー+Δ エネルギー, MFCC+ΔMFCC（各 1〜4 次元）
トポロジー	6 状態 4 分布，left-to-right
分散行列	3 種類の特徴量に対して個別に全共分散行列を算出

表 6.9　先行研究の手法を用いた実験の結果

学習データ	テストデータ	正解率
評価者 A	評価者 A	0.444
評価者 B	評価者 B	0.426

6.6　提案手法を用いた実験

提案手法では，SVM を用いて強勢の度合いを推定する．具体的には，R の kernlab パッケージに含まれる ksvm を用いた [47]．カーネルはガウスカーネル，ハイパーパラメータは ksvm のデフォルトのものを用いた．手動評価が 0 のデータはここでも使用しなかった．使用した特徴量セットを表 6.10 に示す．

特徴量セット a に含まれるのは音響特徴量で，音節の母音部の平均 F_0，継続長，平均エネルギー，平均 MFCC（12 次元），およびそれらの差または比を前後 2 音節に対して計算している．このような特徴量を用いたのは，1) 強勢が，高さ，長さ，強さ，音質の複合的な変化により実現されることと，2) 強勢とは，他の音節よりも際立って聞こえることを指すからである [34, 48]．

特徴量セット b に含まれるのは，音節が持つコンテキストに関するシンボリックな特徴量である．これらは，どのような音節に文強勢が置かれやすいかという観点から選んでいる [31, 34, 48]．

特徴量セット c に含まれるのは，ERJ の原稿に記載されていた強勢記号に関する特徴量である．発声者は原稿に記載されていた強勢記号のとおりに発声できていると考えているため，有用な情報であることが期待される．

特徴量セット b, c には，音響的な情報が含まれておらず，ある読み上げテキストのある音節に対しては，どの発声者に対しても同じ特徴量となることに注意されたい．なお，特徴量セット b, c に関しては，各特徴量を 0/1 の二値で表現した（「X である」という特徴量は，X である場合は 1，X でない場合は 0 をとる）．最終的に，a, b, c のうち 2 つを組み合わせた特徴量セット，すべてを組み合わせた特徴量セットを用意した．

実験は，手動評価の評価者 closed の場合と，評価者 open の場合の両方を行った．すなわ

表 6.10　提案手法で用いた特徴量セット

特徴量セット	特徴量
a（音響特徴量）	母音部平均対数 F_0 ／母音部継続長／母音部平均エネルギー／ 母音部平均 MFCC（12 次元）（いずれも発話ごとに正規化）， 母音部平均対数 F_0 ／平均エネルギーの差 （前後 2 音節まで）， 母音部継続長の比（前後 2 音節まで）
b（コンテキスト）	音節中の母音が単母音か否か，フレーズ頭の単語（最初の 1 単語）／ フレーズ中の単語／フレーズ末の単語（最後の 1 単語）中の音節である， 内容語中／機能語中の音節である，どの品詞中の音節か（名詞／ 代名詞／動詞／形容詞／副詞／前置詞／接続詞／間投詞），単語を単独 発声したときに第 1 強勢／第 2 強勢／弱勢で発声される
c（原稿の強勢記号）	収録時に参照した強勢記号（弱勢／強勢／核強勢）
a-b	（a と b を合わせたもの）
a-c	（a と c を合わせたもの）
b-c	（b と c を合わせたもの）
a-b-c	（a，b，c すべてを合わせたもの）

ち，評価者 closed の場合は評価対象の全 7,414 音節を発声者および発声対象文が重複しないように 3 分割したもののうち 2 つで学習，残り 1 つでテストするという作業を 3 回繰り返し平均をとった．評価者 open の場合は一方の評価者のラベルで学習し，他方の評価者のラベルでテストを行った．

6.7　提案手法を用いた実験の結果

提案手法を用いた実験の結果を表 6.11，表 6.12（3 段階）および表 6.13，表 6.14（9 段階）に示す．

まず，各特徴量セットに対する結果を比較する．3 段階に関しては，学習データとテストデータの組み合わせがいずれの場合でも，正解率および相関が最も低かったのは特徴量セット a であった．一方，最も高かった特徴量セットは，学習データとテストデータの組み合わせによって様々であった．4 つの組み合わせの結果を平均すると，正解率が最も高かったのは特徴量セット a-c（平均 0.786）であり，相関が最も高かったのは特徴量セット b-c および a-b-c（平均 0.737）であった．9 段階に関しては，学習データとテストデータの組み合わせがいずれの場合でも，正解率および相関が最も低かったのは特徴量セット a であり，3 段階と同じ結果となった．一方，学習データとテストデータの組み合わせがいずれの場合でも，正解率が最も高かったのは特徴量セット a-c であった．相関については学習データとテストデータの組み合わせによって様々であったが，4 つの組み合わせの結果を平均すると，最も高かったのは特徴量セット a-c（平均 0.761）であった．以上をまとめると，特徴量セット a-c を用いた場合がおおむね最も高精度であったと言える．また，3 段階に関しては，表 6.9 に示した先行研究の精度を大幅に上回る結果となった．

表 6.11　提案手法による実験の結果（正解率）(3 段階)

学習データ	テストデータ	a	b	c	a-b	a-c	b-c	a-b-c
評価者 A	評価者 A	0.695	0.807	0.810	0.769	0.807	0.818	0.799
評価者 B	評価者 B	0.659	0.753	0.759	0.734	0.796	0.765	0.769
評価者 A	評価者 B	0.694	0.768	0.763	0.774	0.767	0.770	0.776
評価者 B	評価者 A	0.728	0.774	0.764	0.785	0.775	0.768	0.782

表 6.12　提案手法による実験の結果（相関）(3 段階)

学習データ	テストデータ	a	b	c	a-b	a-c	b-c	a-b-c
評価者 A	評価者 A	0.477	0.688	0.693	0.610	0.686	0.712	0.675
評価者 B	評価者 B	0.519	0.707	0.753	0.664	0.773	0.748	0.737
評価者 A	評価者 B	0.634	0.788	0.778	0.800	0.781	0.797	0.805
評価者 B	評価者 A	0.614	0.697	0.690	0.733	0.705	0.691	0.732

表 6.13　提案手法による実験の結果（正解率）(9 段階)

学習データ	テストデータ	a	b	c	a-b	a-c	b-c	a-b-c
評価者 A	評価者 A	0.398	0.431	0.462	0.434	0.470	0.445	0.455
評価者 B	評価者 B	0.546	0.595	0.618	0.592	0.624	0.610	0.614
評価者 A	評価者 B	0.432	0.478	0.475	0.459	0.482	0.478	0.468
評価者 B	評価者 A	0.349	0.366	0.369	0.378	0.378	0.365	0.375

表 6.14　提案手法による実験の結果（相関）(9 段階)

学習データ	テストデータ	a	b	c	a-b	a-c	b-c	a-b-c
評価者 A	評価者 A	0.514	0.631	0.675	0.585	0.695	0.675	0.644
評価者 B	評価者 B	0.491	0.745	0.805	0.633	0.803	0.797	0.721
評価者 A	評価者 B	0.654	0.813	0.799	0.836	0.825	0.821	0.842
評価者 B	評価者 A	0.579	0.716	0.703	0.729	0.722	0.719	0.731

特徴量セット a, b, c をそれぞれ単独で用いた場合の結果については，特徴量セット a のみを用いた場合が正解率，相関とも最も低かった．特徴量セット b と c の結果の差は小さかった．2 つの特徴量セットを組み合わせた場合 (a-b, a-c, b-c) の結果については，特徴量セット a, b, c をそれぞれ単独で用いた場合の結果と比べて正解率，相関ともおおむね向上していた．

次に，すべての特徴量セットの中でおおむね最も高精度であった特徴量セット a-c を用いた場合の結果と，手動判定の結果を比較する．評価者 closed の結果については表 6.3 および表 6.4 と比較することになる．3 段階に関しては，どちらの評価者のラベルを用いた場合でも，正解率，相関とも手動判定の一致率，相関をそれぞれ少し下回る結果となった．9 段階に関してもやはり，どちらの評価者のラベルを用いた場合でも手動判定の精度を下

回った．しかし，評価者 B のほうが評価者 A よりも手動判定の精度に近い結果が得られた．評価者 open の結果については表 6.5 および表 6.6 と比較することになる．3 段階に関しては，正解率および相関を平均するとそれぞれ 0.771，0.743 で，手動判定に近い結果が得られた．9 段階に関しては，正解率および相関を平均するとそれぞれ 0.430，0.774 で， やはり手動判定に近い結果が得られた．

6.8　考察

特徴量セット a, b, c をそれぞれ単独で用いた場合は，特徴量セット a（音響特徴量）を用いた場合が最も精度が低かった．入力音声に関する音響的情報が含まれているのが特徴量セット a のみであることを考えると，(ERJ の発声者である)「日本語を母語とする大学生・大学院生相当の学生」という集団に対しては，発声者がどのような発声をしたかにかかわらず，音節の持つコンテキストあるいは原稿の強勢記号から推定した結果を提示したほうが，発声の情報のみから推定するよりも精度が高いということになる．すなわち，強勢記号が記載された原稿を渡されて，発声者が「正しい」と思って読み上げた場合，多くの発声者が似たようなパターンで読み上げることになることが予想される．これは逆に，非日本語を母語とする話者の読み上げ音声を対象にした場合は，ミスマッチが起こることも予想される．

2 つの特徴量セットを組み合わせた場合 (a-b, a-c, b-c) の結果については，a-c の組み合わせがおおむね最も高精度であった．音響的情報を加えることで，精度が向上したものと考えられる．一方，b-c の組み合わせについては，特徴量セット b, c をそれぞれ単独で用いた場合と比べて，さほど精度が向上しなかった．特徴量セット c で用いている収録時に参照した原稿に記載されていた強勢記号は，ある音声学者によって付与されたものであるが，そのラベリングの際に，特徴量セット b に含まれるようなコンテキスト情報が用いられたことが関係していると考えられる．すなわち，特徴量セット b から得られる情報と特徴量セット c から得られる情報は相関が高く，両者を組み合わせても，精度向上に寄与しなかったと考えられる．また，全ての特徴量を用いた場合 (a-b-c) は，特徴量セット a-c にくらべて若干精度が下回ったが，偶然なのかどうかは不明である．

おおむね最も高精度であったのは特徴量セット a-c を用いた場合であったが，評価者 closed の手動判定の精度には及ばなかった。これは各々の評価者が，今回検討した特徴量以外の独自の特徴量に着眼している可能性がある．なお，評価者 closed の結果と評価者 open の結果を比べると，一部，評価者 open のほうが closed の場合よりも高精度となっている．表 6.3・表 6.4 と表 6.5・表 6.6 からも，closed の結果が一般に高精度となるはずであるが，逆の傾向を示した結果に関する十分な考察はできていない．

6.9　まとめ

本章では ERJ データベースのリズム文の読み上げ音声を評価対象として，SVM を用いて高精度に文強勢の自動推定を行うことを目指し，音節ごとに強勢の度合いを自動推定す

る実験を行った．その過程で，評価対象音声に対して，手動判定による強勢の度合いを 9 段階でラベリングする作業を行ったが，3 段階に丸めて用いた場合と，9 段階のまま使用した場合の 2 通りについて実験を行った．特徴量セットに関して様々な検討を行い，評価者 closed の場合に最高で 81.8%（3 段階），47.0%（9 段階）の正解率を得ることができたが，手動判定の精度には及ばなかった．評価者 open の場合は手動判定に近い結果が得られた．

本実験に関する今後の課題としては，特徴量セットや識別器などの改良による精度向上を考えている．手動判定の判定基準をより明確なものとした上で評価者を増やし，より信頼のおける手動判定データを構築することを考えている．

次章では，本章の実験で用いた文強勢検出技術を，シャドーイング評価に応用した実験について述べる．

第7章

文強勢検出のシャドーイング評価への応用

7.1　はじめに

本章では，第 6 章の実験で用いた文強勢検出技術をシャドーイング評価に応用する．具体的には，評価対象のシャドーイング音声に対して，音節ごとに強勢の度合いを推定する．ただし，このとき第 6 章のように SVM によるクラス分類を行うのではなく，SVM を回帰に応用したものである SVR (Support Vector Regression) によって強勢の度合いを連続値で推定する．そして，本来発声すべき強勢の度合いとの差を，その発声者の強勢スコアとし，第 5 章で行った重回帰の説明変数に加えて，シャドーイング評価のさらなる高精度化を目指す．ただし，第 6 章の実験で用いた ERJ のリズム文と違ってシャドーイングは聴取した音声をそのまま繰り返すものであるから，参照すべき原稿の強勢記号は存在しない．そこで，第 6 章で強勢の手動評価を行った評価者に依頼して，聴取した音声に対して音節ごとに強勢の度合いを手動でラベリングする作業を行い，そのラベルを原稿の強勢記号に相当するものとして扱った．

本章ではまず，SVR についての説明を行う．次に，文強勢検出のシャドーイング評価への応用について述べ，実験結果を示し，考察を行う．

7.2　SVR (Support Vector Regression)

第 6 章でも用いた SVM (Support Vector Machine) は二値分類器であるが，これを回帰に応用したものが，**SVR** (Support Vector Regression) である．

7.3　音声試料

評価対象の音声試料は，第 5 章の実験で用いたものと同じ文章（文章 A，文章 B）および話者を使用した．結果，評価対象となったのは文章 A が 567 発声（12,663 音節），文章 B が 406 発声（5,104 音節）であった．

7.4　手動評価による強勢ラベリング

7.4.1　評価者

評価者は，第 6 章でも手動評価を担当した評価者 A（英語音声学専攻の大学院博士課程学生）であった．

7.4.2　手動評価タスク

文章 A，文章 B ともモデル音声を聴取して，音節ごとに強勢の度合いを評価した．ERJ の強勢記号（弱勢，強勢，核強勢）にならい，同様のラベルを付与した．また，ERJ の原稿でも示されている，フレーズの切れ目を表す記号を付与した．

表 7.1　TOEIC スコアの推定値と実際の TOEIC スコアの相関係数

タスク	$\boldsymbol{GOP_{each\_v}}$（リッジ回帰）	$\boldsymbol{GOP_{each\_v}}+SS$（リッジ回帰）
文章 closed, speaker-open（文章 A）	0.810	0.806
文章 closed, speaker-open（文章 B）	0.758	0.791
文章 open, speaker-open（文章 A → B）	0.754	0.766
文章 open, speaker-open（文章 B → A）	0.793	0.788

7.5　実験

7.5.1　シャドーイング音声に対する文強勢検出

第 6 章の実験の結果，おおよそ最も高精度に文強勢検出を行えたのは特徴量セット a-c であった．そこで，特徴量セット a-c を用いて，シャドーイング音声に対して文強勢検出を行った．ただし，第 6 章のように SVM を用いて強勢の度合いを分類するのではなく，SVR を用いて強勢の度合いを連続値で推定した．SVR の学習は，第 6 章で用いたすべての音声および，それに対する評価者 B の手動評価ラベル（9 段階: 1–9）を用いて行った．そして，学習した SVR を本実験の評価対象音声に適用し，強勢の度合いを推定した．

7.5.2　文強勢検出に基づくスコアリング

7.5.1 で推定した値をもとに，各話者に対する文強勢スコアを算出した．発声した文章には音節が $s_1 s_2 \ldots s_n$ のように並んでおり，s_i に対する 7.4.2 で付与した強勢ラベルを $w_{lab}(i)$ (第 6 章の「3 段階」にならい，弱勢を 2，強勢を 5，核強勢を 8 と数値化する)，7.5.1 で推定した強勢の度合いを $w_{mod}(i)$ として，文強勢スコア SS は以下の式で定義される．

$$SS = \sqrt{\frac{\sum_{i=1}^{n} (w_{est}(i) - w_{mod}(i))^2}{n}} \tag{7.1}$$

7.5.3　シャドーイング評価への応用

式 (7.1) で定義した文強勢スコア SS を第 5 章で行った重回帰の説明変数に加えて，5.5.1 と同じタスクで発話者の TOEIC スコアを推定する実験を行った．ベースラインは，第 5 章の実験で最も良い結果が得られた説明変数 $\boldsymbol{GOP_{each\_v}}$ のリッジ回帰 ($\lambda = 50$) を用いた場合とし，$\boldsymbol{GOP_{each\_v}}+SS$ の場合の λ も同じく $\lambda = 50$ とした．

7.6　結果

結果を表 7.1 に示す．

文章 closed, speaker-open（文章 B）の場合にベースラインより大きく相関が上がったほか，文章 open, speaker-open（文章 A → B）も少し相関が上がった．一方，他の 2 つのタスクについては相関がわずかに下がったものの，同程度であった．

7.7　考察

文章 closed, speaker-open（文章 B）で大きく相関が上がったのは，文章 A のモデル音声がが比較的淡々と話しているのに対し，文章 B のモデル音声は韻律を強調ぎみに話していることが影響している可能性がある．SVR の学習に使用した音声は ERJ のリズム文であり，リズム・強勢に注意して発声された音声である．モデル音声が韻律を強調ぎみに話している文章 B をシャドーイングしたほうがより韻律が強調された形のシャドーイング音声となり，ERJ で学習した SVR がうまく適用でき，結果として相関が上がったと推測することは可能である．

7.8　まとめ

本章では，第 5 章で議論した，音素に基づいたスコアである GOP をもとに重回帰を行ってシャドーイング評価を行う手法に，第 6 章の技術を応用した文強勢スコアという新たなスコアを定義し重回帰の説明変数に加えることで，シャドーイング評価のさらなる高精度化を目指した．結果として，GOP をもとに重回帰を行った場合よりも，同程度か上回る精度が得られた．

次章では，本論文をまとめ，今後の展望を述べる．

第8章

結論

8.1　まとめ

本研究では，[21] のシャドーイング評価手法を，重回帰を用いることで高精度化することを目的とした．また，特徴量として GOP に基づくスコアだけでなく，韻律評価，特に文強勢検出に基づくスコアも利用した．ただし，シャドーイング音声から直接文強勢検出を行うことは困難であるため，読み上げ音声を用いた文強勢検出の実験を行い，その技術を応用してシャドーイングに韻律評価の要素を導入した．また，読み上げ音声を用いた文強勢検出の実験においては，識別モデルを用いて精度の高い文強勢の自動検出を行うこと，音響特徴量以外の，テキストから得られる特徴量も積極的に使って精度を上げることを目指した．

[21] のシャドーイング評価手法を重回帰を用いることで高精度化することについては，第 5 章で議論した．結果としては，説明変数に用いるスコアによっては，重回帰（最小二乗法あるいはリッジ回帰）によって単回帰の場合よりも高精度に TOEIC スコアを推定することができた．

そして，第 6 章では，読み上げ音声を用いた文強勢検出の実験について議論した．ERJ データベースのリズム文の読み上げ音声を評価対象として，SVM を用いて高精度に文強勢の自動検出を行うことを目指し，音節ごとに強勢の度合いを自動推定する実験を行った．その過程で，評価対象音声に対して，手動判定により強勢の度合いを 9 段階でラベリングする作業を行ったが，3 段階に丸めて用いた場合と，9 段階のまま使用した場合の 2 通りについて実験を行った．特徴量セットに関して様々な検討を行い，評価者 closed の場合に最高で 81.8%（3 段階），47.0%（9 段階）の正解率を得ることができたが，手動判定の精度には及ばなかった．評価者 open の場合は手動判定に近い結果が得られた．

第 7 章では，第 5 章の重回帰によるシャドーイング評価に対して，第 6 章で議論した文強勢検出を応用した．第 6 章で用いた SVM を回帰に応用した SVR を用いて強勢の度合いを連続値で推定し，それをもとに文強勢スコアを定義した．それを重回帰の説明変数に加えることで，シャドーイング評価のさらなる高精度化を目指した．結果として，GOP をもとに重回帰を行った場合よりも，同程度か上回る精度が得られた．

8.2　今後の展望

今後の課題の第一は，シャドーイング評価および文強勢検出，個々の精度の向上である．特に，文強勢検出に関しては，評価者 closed の場合に手動判定の精度に及ばなかったので，改善の余地が残されている．方法としては，特徴量の工夫，識別器の改良などが考えられる．加えて，より信頼の置ける手動判定データの構築も重要である．

今後の課題の第二は，文強勢検出のシャドーイング評価への応用を，もっと深く検討することである．文強勢スコアの定義の良し悪しや，文強勢検出をどのようにしてシャドーイング評価への反映させるか，その方法の工夫はまだまだ可能であろう．

そして，今後の課題の第三は，シャドーイング音声からの文強勢検出である．本論文では，学習者のシャドーイングは非常に崩れたものであり，直接文強勢を検出することは難

しいとして，読み上げ音声からの文強勢検出を行うことにした．しかし，今回読み上げ音声に対して行ったのと同じように，シャドーイング音声から直接文強勢を検出する，すなわち音節ごとに強勢の度合いを評価することができれば，学習者にフィードバックできる情報が増え，学習者にとって有益だと考えられる．読み上げ音声を対象にした場合に比べると困難な技術ではあるが，今後の課題としたい．

謝辞

まず，学部時代から 3 年間に渡り研究指導を頂いた指導教員である峯松信明教授に深く感謝いたします．研究室のミーティングから学会発表まですべてにおいてお世話になりました．中国での国際学会で同室に宿泊した際に，朝から晩まで熱心な仕事ぶりを拝見できたことが印象に残っています．

峯松教授とともに的確なアドバイスをくださった広瀬啓吉教授に心より感謝いたします．また，研究活動を様々な面で支えてくださった高橋登技官，秘書の池上恵さんに心より感謝いたします．そして，この研究を進めていく上で気軽に相談に乗ってくださった，博士課程の鈴木雅之さんに改めて御礼を申し上げます．

また，第 5 章の先行研究は広瀬・峯松研究室のメンバーだった羅徳安さんによるもので，研究について詳しくご説明いただき，おかげで自らの研究をスムーズに始めることができました．ここに改めて感謝いたします．そして，本研究全体にわたって共同研究者として度々ご協力いただいた東京国際大学の山内豊教授，ならびに研究グループのメンバーで議論に参加していただいた，同じく東京国際大学の川村明美准教授，東海大学の西川惠准教授，HOYA サービス株式会社の藤田雅也氏に，それぞれ御礼申し上げます．特に，山内先生にはシャドーイング音声の収録ならびに第 6 章の評価対象文選定，手動評価の評価者募集でお世話になりました．西川先生にも，第 6 章の手動評価を行っていただき，大変助かりました．

一緒に研究室生活を過ごした広瀬・峯松研究室のメンバーにも感謝いたします．博士課程の Greg Short さんには，学部時代は直接研究のアドバイスをいただき，修士課程に進学してからも，英語の添削で度々お世話になりました．Short さんが日本語を身に付けられたことを見習い，もっと英語が上達するよう，努力します．同期の王程碩さん，Oraphan Krityakien さん，甲斐常伸君，柏木陽佑君，川口拓也君，橋本浩弥君，同じ時期に研究室に入った Teeraphon Pangkittiphan 君，一席挟んで隣の席だった尾崎洋輔君，休学中にその挟まれた席を尾崎君と折半して使わせてくださった毛利圭佑君，ティーチング・アシスタントを共に勤めた清水信哉さん，尾崎洋輔君（再掲），Nguyen Duc Duy 君，その他すべてのメンバーに改めて御礼申し上げます．

そして，毎日の生活を支えてくださった家族，友人に感謝いたします．

2013 年 2 月 6 日
加藤 集平

参考文献

[1] 法務省，出入力管理統計年報（2011 年）.

[2] 外務省，海外在留邦人数調査統計 平成 24 年速報版，2012.

[3] 文部科学省,「英語が使える日本人」の育成のための戦略構想，2002.

[4] 文部科学省，小学校学習指導要領，2008.

[5] “小 4 以下も英語必修、文科省検討 指導法を研究,” 日本経済新聞，2012 年 9 月 9 日，電子版.

[6] 文部科学省初等中等教育局，平成 25 年度概算要求主要事項【事項別表】，24，2008.

[7] 矢野経済研究所，語学ビジネス市場に関する調査結果 2012.

[8] 楽天，“楽天の歴史,” http://corp.rakuten.co.jp/about/history.html.

[9] 柳井正，“2012 年 8 月期 上期の振り返りと今後の展望,” ファーストリテイリング，2012.

[10] 日本英語検定協会 英語教育研究センター，小学校の外国語活動に関する現状調査 ⟨⟨ 教育委員会対象 ⟩⟩ 調査結果報告，2012.

[11] 日本英語検定協会 英語教育研究センター，小学校の外国語活動に関する現状調査 ⟨⟨ 小学校対象 ⟩⟩ 調査結果報告，2012.

[12] チエル，CALL システム CaLabo EX，2002.

[13] アドバンスト・メディア，AmiVoice CALL -pronunciation-，2005.

[14] HOYA サービス，GlobalvoiceCALL，2011.

[15] Brothers & Co., VSS Editor for CALL / VSS Player for CALL Ver. 2.6.1, 2012.

[16] CAI メディア共同開発，英語発音美人 Vol. 1–Vol. 5，2005–2006.

[17] プロンテスト，発音力，2007.

[18] アデュー，発音 PRO，2008.

[19] 学習研究社，えいご三昧 DS，2009.

[20] EnglishCentral, http://www.englishcentral.com/.

[21] Dean Luo, Naoya Shimomura, Nobuaki Minematsu, Yutaka Yamauchi and Keikichi Hirose, “Automatic pronunciation evaluation of language learners’ utterances generated through shadowing,” *Proc. INTERSPEECH*, 2807–2810, 2008.

[22] 羅徳安，喬宇，峯松信明，山内豊，広瀬啓吉，“シャドーイング・音読発音評価を目的とした話者適応の分析と応用,” 電子情報通信学会技術研究報告，SP，109 (99)，51–56，2009.

[23] 門田修平，“シャドーイングと音読の科学,” コスモピア，2007.

[24] 門田修平，“シャドーイング・音読と英語習得の科学,” コスモピア，2007.

[25] 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,” 風間書房，2005.

[26] Patrick W. Nye and Carol A. Fowler, “Shadowing latency and imitation: the effect of familiarity with the phonetic patterning of English,” *Journal of Phonetics*, 31, 63–69, 2003.

[27] 根間弘海，鈴木俊二，英語とリズムをマスターする英語音声学，英宝社，2000.

[28] 峯松信明，富山義弘，吉本啓，清水克正，中川聖一，壇辻正剛，牧野正三，“英語CALL構築を目的とした日本人および米国人による読み上げ英語音声データベースの構築,” 日本教育工学会論文誌，27 (3)，259–272，2003.

[29] 小橋川哲，峯松信明，広瀬啓吉，ドナ・エリクソン，“英語文リズム学習支援を目的とした文強勢音節のモデル化とその検出,” 電子情報通信学会技術研究報告，NLC，101 (520)，99–104，2001.

[30] Nobuaki Minematsu, Satoshi Kobashikawa, Keikichi Hirose and Donna Erickson, “Acoustic modeling of sentence stress using differential features between syllables for English rhythm learning system development,’ *Proc. ICSLP*, 745–748, 2002.

[31] 窪薗晴夫，音声学・音韻論，くろしお出版，1998.

[32] 川越いつえ，新装版 英語の音声を科学する，大修館書店，2007.

[33] 竹林滋，斎藤弘子，英語音声学入門，大修館書店，1998.

[34] 榎本正嗣，日英語 話し言葉の音声学，玉川大学出版部，2000.

[35] 根間弘海，英語の発音とリズム，開拓者，1996.

[36] S. M. Witt and S. J. Young, “Phone-level pronunciation scoring and assessment for interactive language learning,” *Speech Communication*, 30 (2-3), 95–108, 2000.

[37] Steve Young, Dan Kershaw, Julian Odell, Dave Ollason, Valtcho Valtchev, Phil Woodland, The HTK Book, 2000.

[38] Sadaoki Furui, “Speaker independent isolated word recognition using dynamic features of speech spectrum,” *IEEE Trans. Acoustics, Speech and Signal Processing*, 34 (1), 52–59, 1986.

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

[40] James L. Hieronymus, “Automatic sentential vowel stress labelling,” *Proc. EUROSPEECH*, 1226–1229, 1989.

[41] Juan Pablo Arias, Nestor Becerra Yoma, and Hiram Vivanco, “Word stress assessment for computer aided language learning,” *Proc. INTERSPEECH*, 1135–1138, 2009.

[42] Yeon-Jun Kim and Mark C. Beutnagel, “Automatic assessment of American English lexical stress using machine learning algorithms,” *Proc. SLaTE*, 2011.

[43] Keith Vertanen, HTK Wall Street Journal Training Recipe, http://www.keithv.com/software/htk/.

[44] Kazunori Imoto, Yasushi Tsubota, Antonie Raux, Tatsuya Kawahara and Masatake Dantsuji, “Modeling and automatic detection of English sentence stress for computer-assisted English prosody learning system,” *Proc. ICSLP*, 749–752, 2002.

[45] Chaolei Li, Jia Liu and Shanhong Xia, “English sentence stress detection system based on HMM framework,” *Applied Mathematics and Computation*, 185 (2), 759–768, 2007.

[46] Speech Signal Processing Toolkit (SPTK) Version 3.5, http://sp-tk.sourceforge.net/, 2011.

[47] Package ‘kernlab’ Version 0.9-15, http://cran.r-project.org/web/packages/kernlab/index.html, 2012.

[48] A. C. Gimson, *An introduction to the pronunciation of English*, 1962 （竹林滋訳，ギムスン 英語音声学入門，金星堂，1983）.

発表文献

国際会議論文

[1] Yi Luan, Masayuki Suzuki, Yutaka Yamauchi, Nobuaki Minematsu, Shuhei Kato and Keikichi Hirose, “Performance improvement of automatic pronunciation assessment in a noisy classroom,” *Proc. Spoken Language Technology (SLT)*, 2012-12.

[2] Shuhei Kato, Greg Short, Nobuaki Minematsu and Keikichi Hirose, “Effects of learners’ language transfer on native listeners’ evaluation of the prosodic naturalness of Japanese words,” *Proc. Speech Prosody*, 198–201, 2012-5.

[3] Shuhei Kato, Greg Short, Nobuaki Minematsu, Chiharu Tsurutani and Keikichi Hirose, “Comparison of native and non-native evaluations of the naturalness of Japanese words with prosody modified through voice morphing,” *Proc. SLaTE*, 2011-8.

国内研究会論文

[4] 加藤集平，鈴木雅之，峯松信明，広瀬啓吉，山内豊，西川惠，“識別モデルを用いた英文読み上げ音声からの強勢自動検出,” 電子情報通信学会技術研究報告，SP，2013-2（発表予定）.

[5] 山内豊，峯松信明，加藤集平，川村明美，西川惠，藤田雅也，“合成音声と自然音声による音声モデルの違いがシャドーイング・パフォーマンスに与える影響,” 外国語教育メディア学会関東支部第 128 回研究大会発表要項，10–11，2012-6.

[6] 加藤集平，ショート グレッグ，峯松信明，広瀬啓吉，“母語干渉が外国語発声の韻律的自然性に与える影響に関する知覚的検討,” 電子情報通信学会技術研究報告，SP，110 (452)，19-24，2011-3.

全国大会論文

[7] 加藤集平，鈴木雅之，峯松信明，広瀬啓吉，山内豊，西川惠，“識別モデルを用いた英文読み上げ音声中の弱・強勢自動評価,” 日本音響学会春季研究発表会講演論文集，2013-3（発表予定）.

[8] Yi Luan, Masayuki Suzuki, Yutaka Yamauchi, Nobuaki Minematsu, Shuhei Kato and Keikichi Hirose, “GOP performance improvement of automatic pronunciation assessment in a noisy environment,” 日本音響学会秋季研究発表会講演論文集，337–340，2012-9.

[9] 山内豊，峯松信明，加藤集平，川村明美，西川惠，“シャドーイングに必要な下位能力の研究: 単語認識能力とシャドーイングの関係,” 外国語教育メディア学会第 52 回全国研究大会発表論文集，76–77，2012-8.

[10] 加藤集平，ショート グレッグ，峯松信明，鶴谷千春，広瀬啓吉，“日本語単語の韻律モーフィング音声に対する母語話者と学習者による自然性評価の比較,” 日本音響学会秋季研究発表会講演論文集，579–582，2011-9.

[11] 加藤集平，鈴木雅之，峯松信明，広瀬啓吉，山内豊，“GOP と重回帰分析を用いたシャドーイング評価の高精度化,” 日本音響学会春季研究発表会講演論文集，417–420，2012-3.

学位論文

[12] 加藤集平，“母語干渉が外国語発声の韻律的自然性に及ぼす影響に関する知覚的検討,” 東京大学工学部電子情報工学科卒業論文，2011-3.

付録 A

シャドーイング評価で使用した文章一覧

A.1　文章 A

In nineteen ninety-six, three men in California were taken to a hospital
with strange symptoms. They felt dizzy, tired, and weak. They couldn't
speak, and they had trouble breathing. The hospital doctors thought the
men had been poisoned, but couldn't work out what was wrong with them.
Then they found out the three men were all chefs, and they had just shared
a dish of fugu. Fugu, the Japanese name for the puffer fish, is one of
the strangest fish in the ocean. The puffer fish gets its name from the
way the fish protects itself from enemies. Whenever it is attacked, the
fish puffs up (blows up) its body to over twice its normal size. The
reason the three men were taken to the hospital is because the puffer fish
is also very poisonous. As a rule, if you eat a whole puffer fish, you
will probably die. The three men had a close call, but they all survived.
The symptoms of fugu poisoning are a strange feeling around the mouth and
throat, and difficulty breathing. You can't breathe and your body can't
get any air. Your brain still works perfectly, however, so you know you
are dying, but you can't speak or do anything about it. Despite the
danger of fugu poisoning, this strange, ugly, and very poisonous fish is
actually a very expensive, and very popular, kind of food in Japan.

Customers pay up to two hundred dollars per person to eat a fugu meal.
+ - -　- - - + - - + - - @ - / - + - - - @ /
Because of the danger, fugu can only be prepared by chefs with a special
- + - - @ - / - - - + - - - @ - + / - - + -
license from the government. These chefs are trained to identify and
+ - - - - @ - - / + - - + - - + - - -
remove the poisonous parts of the fish. Most people who die from eating
- @ / - + - - + - - @ / + - - - + - @ -
fugu these days are people who have tried their hand at preparing the
- - + - - /- + - - - + - + - - + - -
fish themselves. Fugu is said to be so delicious that it has even started
- - @ / + - - + - - - - - @ - / - - - + - - -
to be imported into Hong Kong and the United States. Several tons of fugu
- - - + - - - - + + - - - + - @ / + - - + - - - -
are now exported from Japan every year.
- + - + - - - + + - @ /

A.2　文章 B

The MacDonald's house has been broken into. A policeman has come to
/ - - @ - + / - - + - @ -/ - - @ - - - + / -
check it out. He finds a boy standing nearby. The policeman is now
+ - @ / - @ - + / + - - @/ - - + - - -
talking to the boy. He wants to know how the door of the MacDonald's
+ - - - @ / - + - @ / - - + - - - + -
house was broken open. The boy said that it had already been broken
@ / - + - @ - / - @ + / - - - + + - - + -
before he and his friend went to the house. He said that they simply
- - + - - @ / - - - @ / @ + / - - + -
walked into the house. The police officer asked, Why were your
+ @ - - - / - - + + - - @ / + - -
fingerprints found all over the door? And why were your boots scratched?
@ - - + /+ + - - @ / - + - - @ + /
It was you who kicked the door open, wasn't it? Why did you steal the
- - @ / - + - - @ - / @ - / + - - + -
stereo and the CDs? Did you just want to have a bit of fun, or were you
@ --/- - +@ / - + + + - + - - - @ / - - -
trying to get some money? Now then, tell me the truth. I don't want to
+ - - - @ - / @ + / + + - @ / - + - -

hear any more of your lies.
 -   @ - +   -   -   + /

付録Ｂ

文強勢検出で使用した文章一覧

付録 B　文強勢検出で使用した文章一覧

文の前の文字列は，ERJ データベースにおける文番号である．

```
S_PR_R_1_004 Come to tea with John and Mary at ten.
          / +   - @ / -   +   -   +   -   @ /
S_PR_R_1_008 Thank you very much for everything that you did for us.
          /  +   -   + -  @  / -  @   -  -  /  -   -   @   -  - /
S_PR_R_1_025 Why won't you wait until Friday when he's back?
          /  -  -    -   @  /-  +   @ - /  -   -   @  /
S_PR_R_1_033 She's fallen in love with an artist again.
          /  -   + - -   @  / -   - + -   - @  /
S_PR_R_1_037 Tom looks as if he were tired.
          / +   @   /-  -   -  -   @   /
S_PR_R_1_038 Tom looks as if he were a little bit tired.
          / +   @   /-  -   -  -   -  +   -  -   @   /
S_PR_R_1_041 I think that he wants us to leave.
          /-   @  /  -   -  +   -   -  @   /
S_PR_R_1_042 I think that he wants Mary and Nancy to leave for school.
          /-   @  /  -   -  +   + -  -   @  -/ -  +    -    @  /
S_PR_R_1_046 I'll take him to the recital the day after tomorrow.
          /-     +   -   -   -  - @ - /  -  +  +  -   - @  - /
S_PR_R_1_056 I wanted him to be an engineer.
          /-  +  -   - / -  - -  -  - @  /
S_PR_R_1_068 Betty often cooks breakfast before seven for her familiy.
          / +  - +  -   +     @   -  / - +   @ - / -   -   @    /
S_PR_R_1_080 The girls have eaten some cookies.
          /  -  +    -   +  -   -   @  -  /
S_PR_R_1_081 The girls will have eaten some of the cookies.
          /  -  +    -   -   @  - / -   -   -  @    /
S_PR_R_1_085 The boys have sold some of the flowers.
          /  -  +   -   +   -   -   -   @ -  /
S_PR_R_1_086 The boys have been selling some of their flowers.
          /  -  @  / -   -   +  -   -   -   -    @ -  /
S_PR_R_1_102 The train arrived at the station on time.
          /  -   +   -  +   -   -   @ -  /-   @  /
S_PR_R_1_104 Could you give me a couple of examples?
          / -    -   +   - -  +   - -  - @   - /
S_PR_R_1_109 I will start work at nine and get off at six.
          /-  -    +   +   -   @  /-   -  +   -   @ /
S_PR_R_1_116 Who did you write the letter to?
          /  -  -   -   +    -  @  -   +/
```

付録 B　文強勢検出で使用した文章一覧

```
S_PR_R_1_119 Look it up in the dictionary, and tell me its meaning.
           / +   -  +  -    -  @  -  - -/ -    +    - -    @  -  /
```