修士論文

学習者の音声で
矯正フィードバックする
英語発音教育システム

2008 年 02 月 03 日

東京大学大学院
情報理工学系研究科
電子情報学専攻　66443
三輪 周作

内容梗概

近年の国際化社会の中で, 英語という言語の重要性が叫ばれている. しかし, 我々日本人の英語発音能力は世界的に見て低いレベルにあると言われている. その原因のして, 言語の持つリズム構造の違い, 音節構造が挙げられる. 英語には「強勢が置かれる音節が等時的に現れる」という特徴がある. また, 英語は子音で音節が終わることを許す言語である. 正しい英語の発音を身に付ける為には, この英語特有の特徴を身に付ける必要がある.

一方で近年, 計算機の急速な発展により, 語学発音学習を支援するソフトウェアが多数発表されている. これまで様々なソフトウェアが提案され, 音声認識技術を用いたものもあるが, 学習者が何をどう直したらいいのがを視覚的に表示し, 教師音声を流すだけのものがほとんどであった. しかし, それだけでは発音学習において学習者は自分の発音の誤りを理解することが困難であり, より効果的なフィードバックというものが求められる.

そこで本研究では, 英語文強勢のモデル化とその検出に関する技術と, 音声修正技術を用いて, 学習者の音声で矯正フィードバックする英語リズム発音教育システムを作成した.

目次

図目次

表目次

第1章

序論

1.1　はじめに

1.1.1　英語発音教育の現状

近年の計算機能力の向上, 及び音声認識・合成などに代表される音声情報処理技術や言語情報処理技術の発展により, それらの技術を応用した語学学習支援システムに関する研究が盛んに行なわれており,PC 上のソフトとして市場に流通したり, 教育機関等で積極的に導入されている例もある. しかしながら, 市販のソフトは単にマルチメディア技術を用いて学習者の学習意欲向上を目的としたものも多い. また音声情報処理技術を用いたものでも音声認識技術を直接的に利用した形態が殆んどであり, その結果分節的特徴に基づく処理系となっている. この場合, リズムやイントネーションなどを伝搬する韻律的特徴は無視された処理形態となっており, 語学学習支援システムとしては不十分なものとなる [2].

また, これまでのほとんどの学習者支援システムでは単に結果を表示したりするものが多く, 学習者が実際にどのように発音をなおしたらいいか, という点について考慮したものはあまりないといえる.

現在, 教育機関で行なわれている語学教育現場では,[l] と [r] に代表される単音などの音韻の発音指導に比べ, リズムやイントネーションといった韻律的特徴は, 指導の困難さから後回しにされがちである [3]. しかしながら, 米語母国語話者が最も重要に感じる発音情報が強勢であるとの報告 [4] があるように, 韻律的特徴の習得が母国語話者とのコミュニケーションには重要であり, 実際, 外国語教育法の中には, 対象とする言語の音韻の学習に先立ち, リズムやイントネーションといった韻律の学習を重要視するものもある (例えば, Verbo-tonal method[5] など).

そこで本研究では, 強・弱勢やイントネーションなどの韻律に関して学習者の音声で矯正フィードバックする発音教育システムを提案する.

1.1.2　本研究の目的

先行研究において,Fred らは, 日本語アクセントについて学習者の音声で矯正フィードバックするシステムを提案している [6]. そこでは, 学習者の音声のピッチ・duration・パワーなどを教師音声のそれからマッピングすることで学習者の音声で正しいアクセントの音声を得ており, それを聞かせて学習させたところ, 教師音声のみを聞いて学習した場合よりもより少ない回数で目標を達成できたことが報告されている. この手法では, 学習者が教師音声の単なる声真似に陥ることなく学習することができ, 自分の誤った音声と修正した音声を聞き比べることでより顕著に誤った部分を認識できると考えられる.

また, 先行研究において, 峯松らは, 孤立発声された英語の単語に対する強勢音節の自動検出技術を構築している [7], そこでは, 音節を強/弱勢のみならず, 母音の種類や連接する子音からなる音節構造, 単語内位置, 前後の音節環境 (強/弱勢) などを考慮することで, いくつかの音節カテゴリを定義し,HMM を用いて音節単位でモデル化を行なっている. この際, 日本語のアクセントが声の高さで表現されるのに対し, 英語では強/弱勢の差が音の強さ, 高さ, 長さ, そして母音の音質にまで現れるとの知見 [8] から, 上記の韻律的特徴に加え, 粗い分節

的特徴としての低次のケプストラムも特徴量に導入している.

これらを踏まえ, 本研究では, 英語のリズム, イントネーションに着目した, 学習者の声で矯正フィードバックする英語発音教育システムについて提案する.

1.2　本論文の構成

本稿は以下のように構成されている.

- 第 2 章:本研究に関する英語音声学に関する基礎知識や, 日本語と英語の違い, それらを考慮した音響モデルについて述べる.
- 第 3 章:発音教育システムについて概観し, システムで重要となるユーザ入力の評価技術やフィードバック, システムの評価方法について述べる.
- 第 4 章:本論文で構築したシステムで用いている TD-PSOLA というピッチ・duration・パワーの修正技術について述べる.
- 第 5 章:本論文で提案する, 学習者の音声で矯正フィードバックする英語発音教育システムについて述べる.
- 第 6 章:最後に本論文のまとめと今後の課題を述べる.

第2章

日本人にとっての英語発音

2.1　はじめに

本章では, 本研究の背景としての英語音声学の基礎知識について述べ, 既存の英語発音教育システムを紹介し, いくつかの関連研究例についてまとめる.

2.2　英語音声学に関する基礎知識

英語のリズムを議論する際に必要となる英語音声学に基づく基礎知識について述べておく.

2.2.1　用語の定義

音の高さ (ピッチ)(pitch)　人間が聴覚的に感じる音の高さは音響的にその音の**基本周波数**[1] (fundamental frequency, $F0$) に対応する. 会話における基本周波数は男性では,50∼250Hz, 女性では 120∼480Hz と言われている. 音声データから F_0 を抽出するツールがいくつか開発されている. 例えば,TEMPO[10] などがその一例である.

リズム (rhythm)　ある決まった音声的な型 (pattern) がほぼ等間隔で繰り返し現れる現象をいう. これをリズムにおける**等時性** (isochrony) と呼ぶ. どのような音声的/音響的イベントが繰り返されるかは言語に依存する.

イントネーション (intonation)　音の高さ (ピッチ) の上がり下がり, つまりピッチの時間的パターンが句またはそれ以上の長い範囲に及ぶときにはイントネーションという.

2.2.2　日本語と英語の言語学的差異

著者は, 日本人による英語の発音教育システムの構築を目的とした研究を目的としている. 以下に日本語と英語の韻律に注目した言語学的差異について述べる.

2.2.2.1　音節

後述するが (第 2.3.0.2 節), 日本語と英語におけるリズム構造に関する基本的韻律単位には差異が存在する. 英語は音節, あるいは**シラブル** (syllable) と呼ばれるものを基本的単位とする言語であり, 日本語は音節より小さな単位である**モーラ** (mora) を基本的単位とする言語である [11]. 表 2.1 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.[7]

表 2.1 でわかる通り, 日本語には子音で終わる規則がない. そのため, 日本人が英語の子音で終わる英単語を発話した場合最後の子音の後に母音を挿入してしまう現象がよくみられ

[1] 声帯の振動周期のことを基本周期 (fundamental period) といい, その逆数を基本周波数と呼ぶ.[9]

表 2.1: 日本語のモーラと英語のシラブルの構造的差異

モーラ	母音 (V), 子音+母音 (CV), 撥音 (/N/), 促音 (/Q/)
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.

る. 英語においては母音が増えると音節がひとつ増えたと知覚されるので, 母音を挿入してしまうことは発話の伝わりやすさに大きく影響すると言われている.

表 2.2 に主な日本人の音素挿入の傾向を挙げる. [12]

表 2.2: 日本人の主な音素挿入

最後の子音	挿入母音の種類
t,d	お
c,b,g,f,k,l, p,s	う
te, de, ce, be, fe, ke, le, pe （上記に e がついたもの）	ト, ド, ク, ブ, フ, ク, ル, プ

2.3　アクセント

日本語において, 音の高さの変化を指し, **高さアクセント**と呼ばれ, 音声情報処理分野では, ピッチのみを用いて記述される. 一方の英語においては, **強さアクセント**と呼ばれる. 音声情報処理の分野ではパワーのみでは十分に記述できず, ピッチ, 短時間パワー, 短時間パワーの時間累積値, 持続時間, 母音の音質などがアクセントの記述に用いられている. 英語のアクセントを以降, 特に「強勢」という言葉を用いる.

強勢 (stress)　ある音節を発音するに当たって音源である呼気が強くなったりその量が多くなると喉頭や調音器官が緊張して調音のエネルギーが強くなり, 聞き手が感じる**音の大きさ** (loudness) が増大する現象をいう. 強勢を受けた音節はピッチが高まり, 音が長めになる. 強勢は**強強勢** (strong stress) と**弱強勢** (weak stress) とに 2 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.

単語強勢 (word stress)　語中にある音節に置かれた強い強勢

文強勢 (sentence stress)　文中の特定の音節が持つ強勢およびその強弱の差. 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.

2.3.0.2　リズム

リズムにおける等時性に関しても日本語と英語は異なる. 日本語は, フランス語やイタリア語と同様に, 一つ一つの音節が等間隔で発音され, **音節拍のリズム** (syllable-timed rhythm [2]) を持つといわれる. 一方の英語の方は, ドイツ語, ロシア語と同様に強い強勢が等間隔に繰り返され, **強勢拍のリズム** (stress-timed rhythm) を持つといわれる. 特に, 英語のリズムを形成するものとして, ある一つの強勢音節から次の強勢音節までを**脚** (foot) という.

2.4　日本人英語のリズムの分析

2.4.1　音節単位の持続時間による分析

Tarui[13] は, 日本語と英語の異なる音声体系のために, 英語のリズムの習得は日本人学習者にとって大きな壁の一つとなっていることを踏まえ, 英語の習得度に応じて脚の持続時間と無強勢音節[3] の測定を行ない日本人学習者の発声の英語のリズムタイミングの要因の比較を行なっている.

被験者を J(英語学習経験無し),B(初心者),M(中級者),A(上級学習者),N(母国語話者) の 5 つのグループに分けて, 流暢に読めるまで練習を行なった英文を朗読したものを録音し, 測定を行なった.

図 2.1 に各話者グループごとに含まれるシラブル数に対する脚の持続時間の関係を示す. これによると,A と N,B と M の間に類似関係が見られる. 全話者グループに対して含まれる音節数が増えると脚の持続時間が長くなるという英語が持つ等時性に反した結果となっている.

図 2.2 に強勢音節に対する無強勢音節の持続時間の比を示す. これによると, 日本人学習者は母国語学習者と比較して, 強勢音節と無強勢音節の差が少ない. この傾向は,A,M,B の順で増え, 弱形の適切な使用に支えられる無強勢音節の発音の実現は適切な英語のタイミングの習得に関して重要であることが示される.

強勢音節と無強勢音節を持続時間に以外にも, ピッチやパワーなどの強形[4] となる時の変化が予想されるパラメータについても比較を行なうことが考えられる.

2.4.2　母音単位の持続時間による分析

また,[15] ではリズム測定法 PVI(Pairwise Variability Index) を用いて, 日本語と英語のリズムを客観的に比較し, 日本人学習者による英語リズムの問題点を指摘している. PVI は以下の式 (2.1) で表される. 母音長のみを考慮し, 発話速度は標準化されている. 強勢拍のリズムは高い PVI, 音節拍のリズムは低い PVI となる傾向がある.

[2]日本語の場合, 音節構造がモーラを単位としているため,mora-timed rhythm と呼ぶべきであろう.

[3]弱母音を持つ音節. 弱音節 (weak syllable).

[4]単音節語の機能語の多くは弱い文強勢の結果, 弱母音を含む弱い形となることが多い. これを弱形 (weak form) と呼ぶ. 強母音を含む本来の場合を強形 (strong form) という.[14]

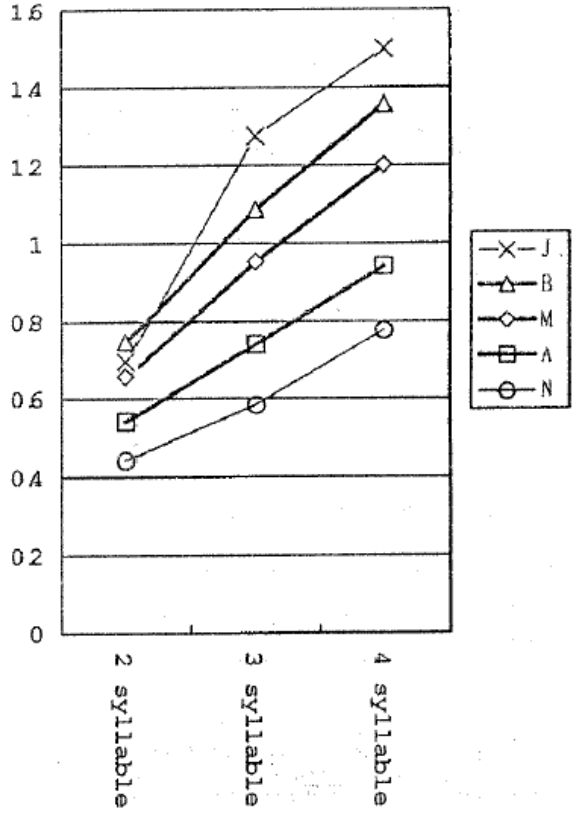

図 2.1: 平均脚持続時間 (秒)

$$PVI = 100 \times \left[\sum_{k=1}^{m-1} \left| \frac{d_k - d_{k-1}}{(d_k + d_{k+1})/2} \right| /(m-1) \right] \tag{2.1}$$

$$\begin{aligned} \text{ただし}, m &= \text{発話中の母音の数} \\ d_k &= k\text{ 番目の母音の持続時間} \end{aligned}$$

この PVI を用いて, 英語母国語話者と日本人話者による, 日本語, 日本人英語, イギリス英語の比較を行なっている.

- 被験者
  - 英語母国語話者 (英語文のみ録音)
  - 日本語母国語話者 (英語文, 日本語とも録音)
- 発声文
  - 英語 8 文
    - ∗ S set 4 文 · · · 母音は全て強母音
    - ∗ SW set 4 文 · · · 母音は全て強母音と弱母音の交互

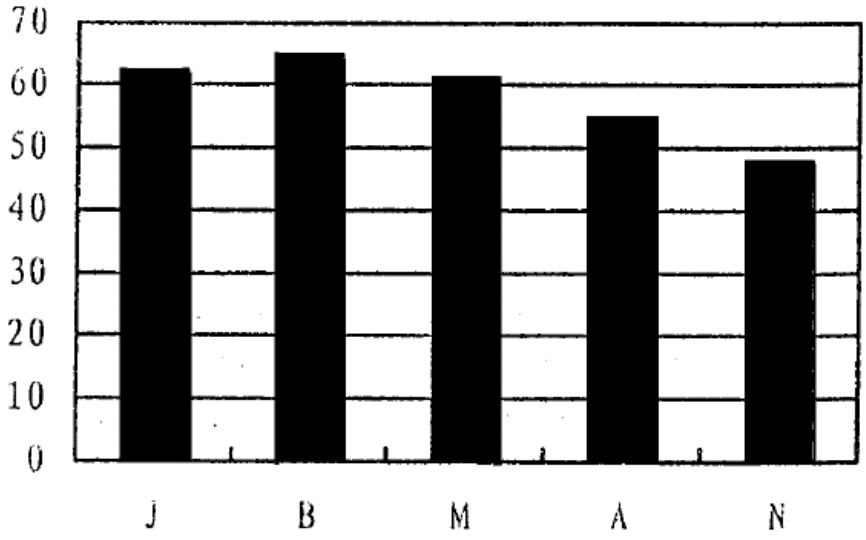

図 2.2: 無強勢音節対強勢音節の持続時間の比

- 日本語 8 文

● 分析
- スペクトログラムを用いて母音長測定
- 全話者の全文について同一の分節化の基準
- 各話者の各文単位で PVI を測定

● 結果
- 日本語, 日本人英語, イギリス英語の比較
  - ∗ 日本語の PVI > イギリス英語の PVI
  - ∗ 日本人英語の PVI ≃ 日本語の PVI

  PVI 比較では日本人英語はイギリス英語よりも日本語に近い.
- 日本人, 英語母国語話者による S set と S-W set の PVI 比較
  - ∗ イギリス英語の PVI: S-W set ≫ S set
  - ∗ 日本人英語の PVI:S-W set ≥ S set
  - ∗ S set の PVI:イギリス英語 ≃ 日本人英語
  - ∗ S-W set の PVI:イギリス英語 > 日本人英語

  日本人学習者の英語リズムの問題点の一つは, 強母音に比べ, 弱母音を短く発音できない

[16] でも同様に, 日本人の英語の問題点が弱勢を弱勢らしく発声できないことにあると言われている.

2.5　日本人の英語を考慮した forced alignment

2.5.1　音響モデル

2.5.1.1　音素単位の音響モデル

音響モデルとは, ある音素 W が発声された時にどのような特徴パラメータ列 X が生成されるのかを確率的に記述するモデルである. 単語単位の音響モデルは, 音素単位の音響モデルを作成し, これと発音辞書とを組み合わせることで構築する. その理由は単語単位で音響モデルを構築すると, 単語の種類は非常に多いためモデル数が膨大になり, モデル学習が効果的に行なえなくなってしまうからである. 音素単位でモデルを構成する際の問題点は, 同じ音素でも調音結合の影響により前後の音節により音響的特徴が大きく影響を受けることにある. この問題を解決する為に同一音素でも前後の音素により別のモデルを用意する方法がある. 前または後の音素によってモデルを用意するものを biphone, 前後の音素を共に考慮するモデルを triphone といい, 前後の音素を全く考慮しないものを monophone という.

本研究では, 日本人学習者の英語文音声に対して, 英語の音韻モデルを用いて時間情報を抽出することになる. [17] と同様の議論により, 日本人の音声中には必ずしも英語の文脈に依存した音素系列での発声とはならないと思われる. そこで本研究では文脈自由の音響モデル, すなわち monophone を用いることとする.

2.5.1.2　隠れマルコフモデル (HMM)

ある音素からある特徴パラメータが生成される確率を与えるモデルとして隠れマルコフモデル (Hidden Marcov Model) が用いられる. 音素単位の音声認識には図 2.3 のような 3 状態の left-to-right 型 HMM が用いられ, 一つの HMM が一つの音素モデルに対応する.

HMM は, 遷移確率 a_{ij} で状態遷移を行ないながら, 各状態で定義された出力確率分布 $b_i(X)$ に従って特徴パラメータ X を出力する, といったモデルである. 出現確率分布 $b_i(X)$ としては, 最も単純な場合は正規分布が用いられるが, 実際にはもう少し複雑な出力確率を表現する為に混合正規分布が用いられることが多い.

本研究では, 音素の代りに, 強勢音節 HMM/弱勢音節 HMM によるモデル化を行なっている.

2.5.1.3　HMM からの時系列出力確率の計算

特徴パラメータの時系列 X が観測されたとき, この系列が与えられた HMM から出力される確率の計算について示す. 図 2.4 は, 特徴パラメータの時系列 $X(1), X(2), \cdots, X(7)$ が 3 状態の HMM から出力される場合に可能な状態遷移を示している. ある状態遷移によって時系列 X が観測される確率は, 状態遷移確率 a_{ij} と各状態での特徴パラメータの出力確率 $b_i(X)$ の積によって計算できる. その確率を全ての経路について和を取ることにより, この HMM から特徴パラメータの時系列 X が出力される確率を求めることができる.

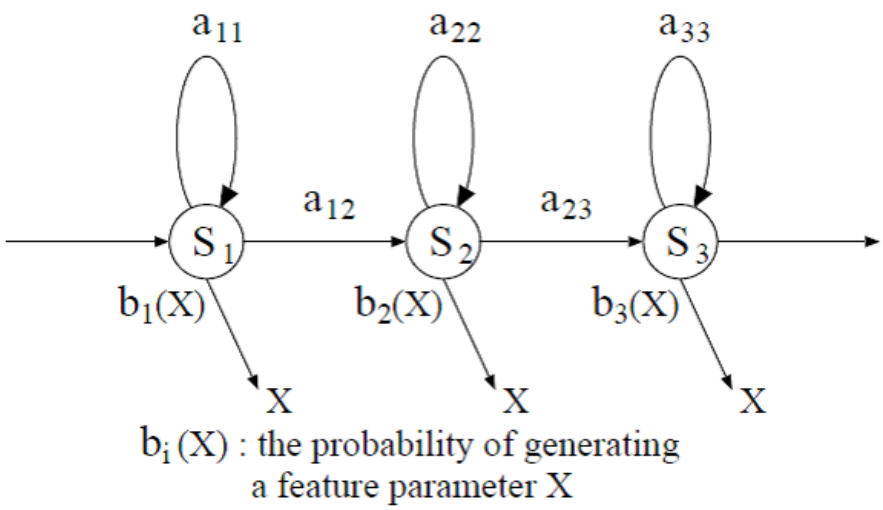

図 2.3: 隠れマルコフモデル

しかし, 全経路からの出力確率の和をとると計算量が膨大になってしまうため, 実際には出力確率が最大の経路を選択していくという近似により計算量を削減している. このアルゴリズムを Viterbi アルゴリズムという.

2.5.1.4　HMM のパラメータ学習

HMM のパラメータ λ は, その HMM が表している音素を発声したときに観測される特徴パラメータ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 $P(X|\lambda)$ を最大にする λ を求めれば良い. 次の式で表すと以下のようになる.

$$\hat{\lambda} = \underset{\lambda}{\operatorname{argmax}}\, P(X|\lambda) \tag{2.2}$$

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なうことになる.

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, そのような手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. この方法は初期パラメータ λ をもとに, 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, この $\hat{\lambda}$ を初期パラメータとして同様の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 に対応する音素 (音節) の学習音声データを用いて Baum-Welch 法を適用することで,HMM のパラメータ学習が行なわれる. これには大量の学習データが必要となり, 近年の音声認識の世界では音声データベースの整備は非常に重要になっている.

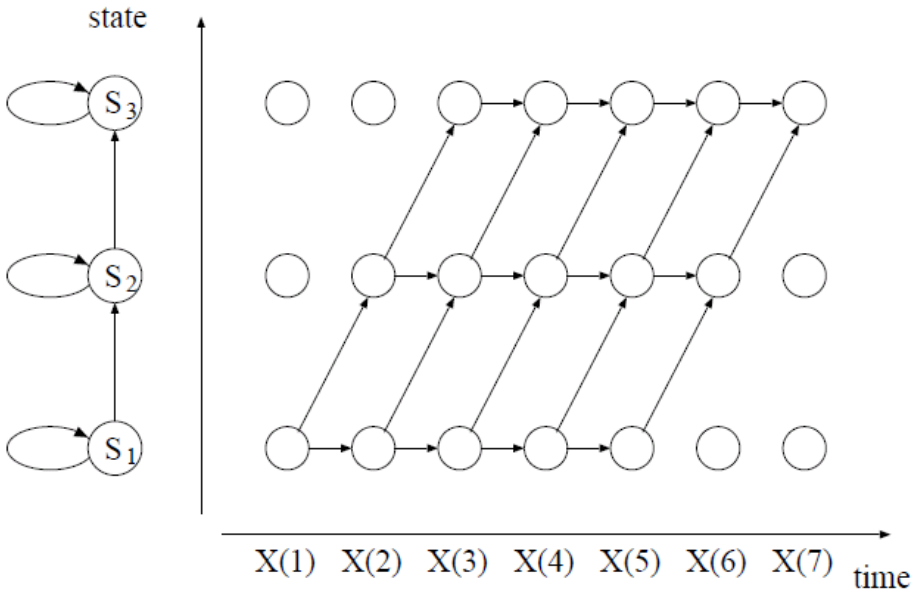

図 2.4: 状態遷移

2.5.1.5　forced alignment の実際

以上で構築された英語音韻モデルを用いて, 音声から各音韻の時間情報を抽出する forced alignment を行なう.

まず, どのような単語の並びが許容されるのかと記述する必要がある. これは, 発声する文通りに記述すれば良い. ただし, 文頭/文末に存在するポーズ (silB/silE) を忘れてはならない.

次に, 各単語に対してどのような音素系列が許されるのかを書いた発音辞書が必要となる. 発音辞書の例を以下の表 2.3 に示す. また, 発音辞書は PRONLEX 0.2(COMLEX English pronouncing lexicon from the LDC[5]) の short label を元に作成した.

表 2.3 に示す通り, 発音記号が複数考えられる場合は, 全てのエントリーを記述しておく. その上で,forced alignment をかけると最も可能性の高い音素系列で音韻認識を行なう. 各単語の後にショートポーズ (sp) の挿入を許した発音辞書となっている.

母国語話者の音響モデルが存在する場合に日本人話者の英語音声の認識精度を向上させる手段として, 以下のようなものが挙げられる.

1. 音素モデル間の類似度に従い, 音素モデルの置換を行なう. 例えば, 日本人英語の音素/v/のモデルを母国語話者の音素/b/のモデルを用いる
2. 日本人英語における発音様態を表す音素系列と, 正規の母国語話者の音素系列のマルチモデルを用いる

[5]Linguistic Data Consortium http://www.ldc.upenn.edu/Catalog/

表 2.3: 発音辞書の例

単語	発音記号列
amused	x m y u z d
amused	x m y u z d sp
by	b Y
by	b Y sp
i'm	Y m
i'm	Y m sp
man	m n
man	m n sp
the	D A
the	D A sp
the	D i
the	D i sp
the	D x
the	D x sp
the	D I
the	D I sp
silB	silB
silE	silE

3. 日本人英語音声データにおける複数の単語音声データから平均的な発話様態を表す単語モデルを求め, 母国語話者用の辞書による単語モデルと併用する
4. 発話様態に着眼したクラスタリングを行なって, 単語モデルを作成する

後述した本研究で構築したシステム (第 5 章) では日本人学習者による母音挿入を考慮して, 各子音の後に日本語の 5 母音が挿入することを許す文法を書いて,forced alignment を行なった.

第3章

発音教育システム

3.1　はじめに

近年, その必要性から多くの英語教育に関するソフトが市場に流通しており, 英語発音教育システムに関する研究も盛んに行なわれている. 多くの発音教育システムは, 課題の提示, 学習者の入力, 分析, 評価, システムによるフィードバック, のような順で進行する. 本章では, これらに関して関連する先行研究についてまとめた.

3.2　発音教育システムの概要

本章では, 発音教育システムにおける概要について述べる.

CAI(Computer Assisted/Aided Instruction) **システム**

指導者が学習者を指導する時の主要な機能をコンピュータに代行させ, あらかじめ用意された教材 (コースウェア）に従って学習者個々に応じた最適な学習指導を行うシステム[1]

CALL(Computer Aided Language Learning) **システム**

コンピュータが生身の先生を取って代わるのではなく, コンピュータを使って言語学習の促進を図ることが主な目的[2]

現在, 多くの英語 CAI システムが商用のソフトとして市場に流通している. 以下にその英語 CAI ソフトの例を示す.

3.2.1　マルチメディア技術に基づく英語 CAI システム

一般的に商用の英語 CAI ソフトには音声情報処理技術を用いたものよりも, ビデオ映像などのマルチメディア技術を伴ったものが多い. その一例である Microsoft ENCARTA[18] を図 3.1 に示す.

これらの多くが, 音声処理技術よりもビデオ映像を見た後にそれについての設問に答える形式になっている. また, 千葉大学の協力でメディア開発センターが英語の CALL 教材 [19] を発表しているが, これもリスニング中心でリスニング後に設問に答える教材である.

3.2.2　音声情報処理技術に基づく英語 CAI システム

音声情報処理技術を用いた英語 CAI ソフトの一例である English Now![20] を図 3.2 に示す. この例のように音声波形を出力して比較をするものも多い.

この教材には「リズム」に関する教師音声との比較が導入されているが, この教材における「リズム」というのは「音の長さ」のみを考慮した結果である. また, 分析結果によるフィードバックの意味するところが不明確で分かりにくく効果は薄いように考えられる.

[1] http://sociolab.tamacc.chuo-u.ac.jp/ saizeHP/2000zemi/komine-p/txt1.html
[2] http://www15.freeweb.ne.jp/diary/yujii/english/qa/tesol.htm#CALL

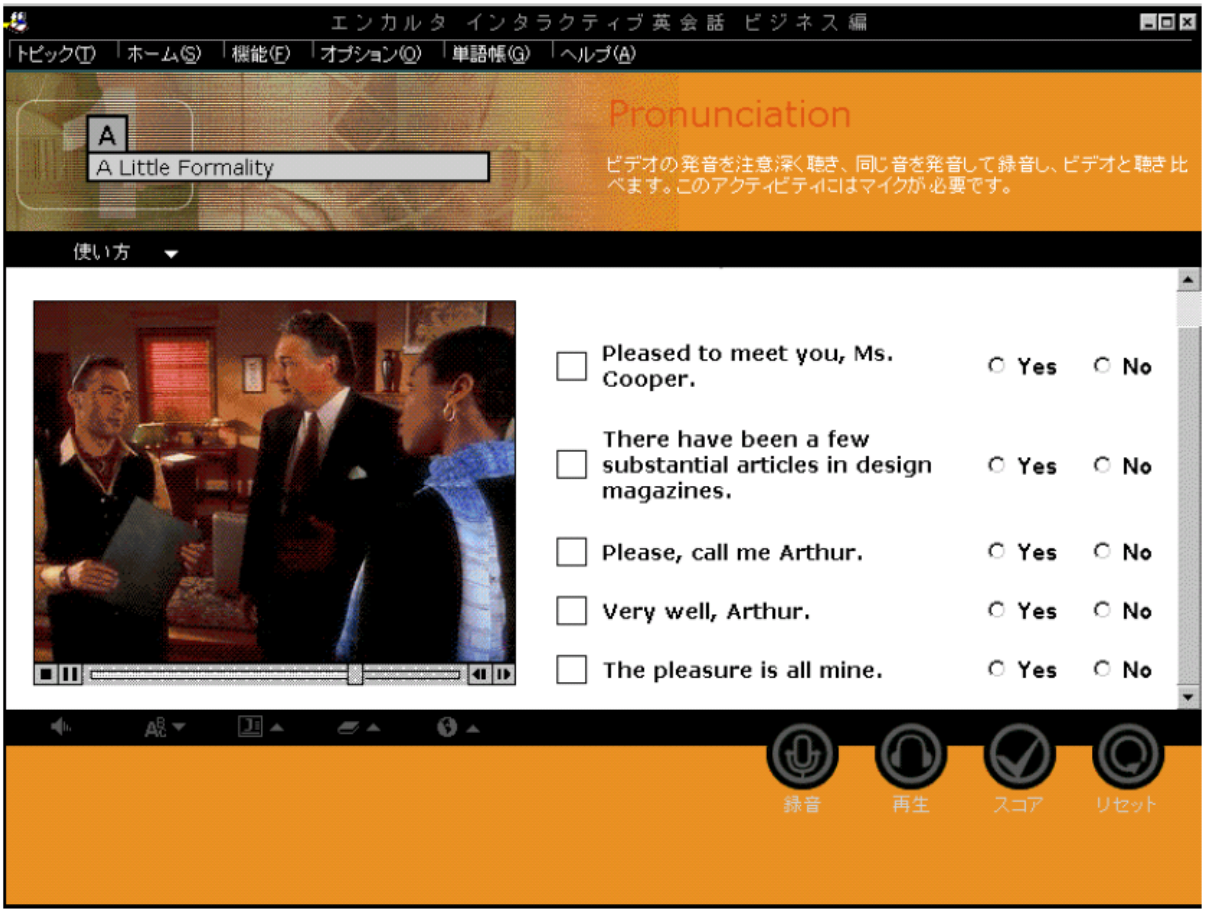

図 3.1: マルチメディア技術に基づく英語 CAI システム:Microsoft ENCARTA

また音声認識技術を用いた英語 CALL システムとして [21] では, インターネット上で公開実験を行なっているが, 韻律的特徴よりも音韻学習が中心の内容になっている.

3.3　ユーザ入力に対する評価

3.3.1　時間的特徴量による発音評点

Strik ら [22] は, 連続音声認識を用いた発音自動評点手法の研究の大部分が朗読音声に対して行なわれているため, 自然発話音声への適用可能性に疑問を持ち, 朗読音声と自然発話音声の時間的測定を行なっている.

連続音声認識システムを用いて, オランダ語の朗読音声と自然発話音声に対して以下の測定項目の時間的測定の自動評定を行なっている.

1. ros(音声の割合)=音声の数/無音抜き全持続時間
2. ptr(音声/時間比)=100% * 無音抜き全持続時間/全持続時間

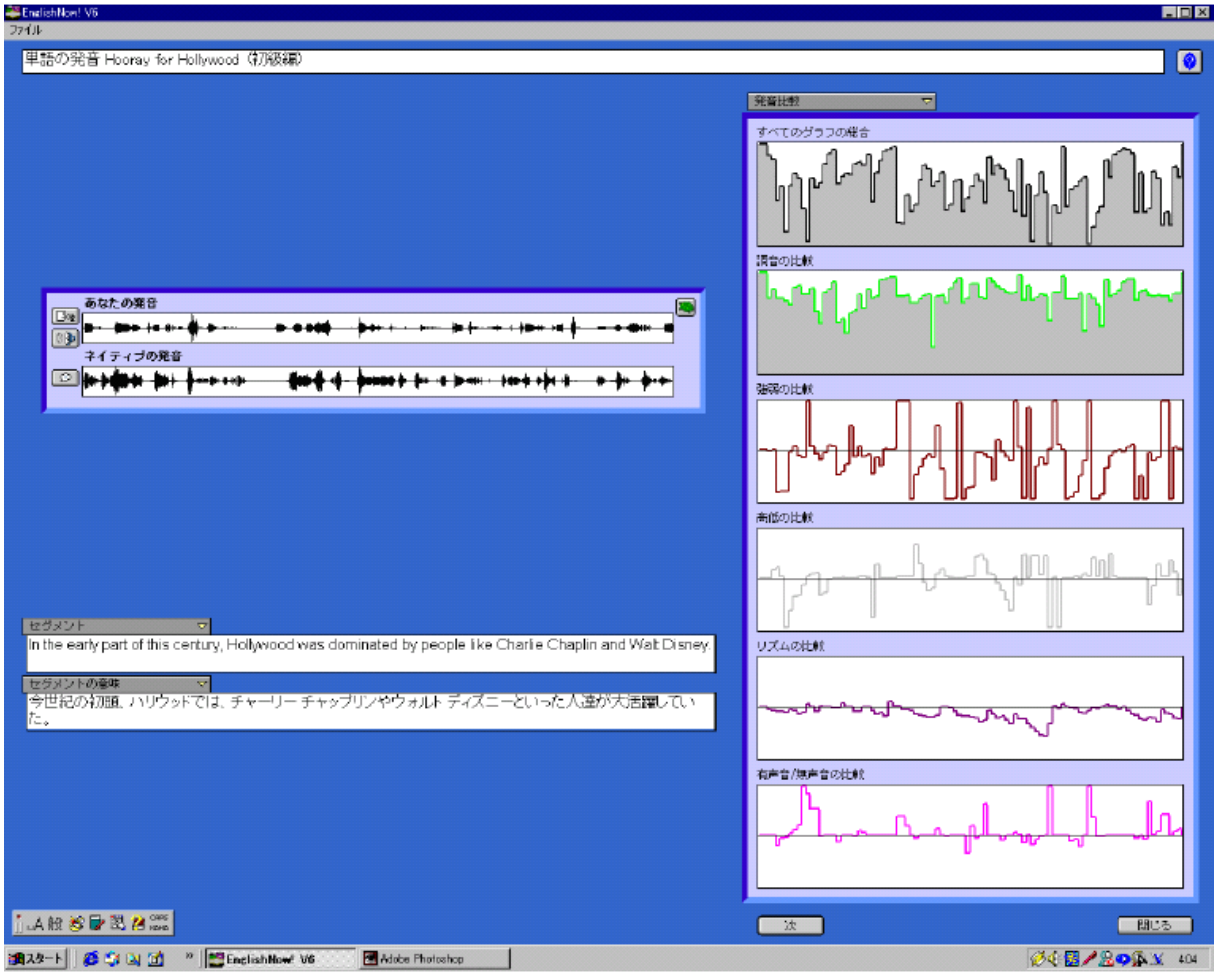

図 3.2: 音声情報処理技術に基づく英語 CAI システム:English Now!

3. art(発音率) =音声の数/全持続時間
4. #ps(単位時間当たりのポーズ[3]数) =ポーズの数/全持続時間
5. mlp(ポーズ平均長) = 全ポーズの長さの平均
6. mlr(連続発声の平均長) = ポーズ間の平均音声数

これらの測定値と専門家のによる評定基準である, 全体の発音品質 (OP), 文節的品質 (SQ), 流暢さ (FL), 話速 (SR) との相関を調べている. その結果, 朗読音声に関してはFL と SR のような時間と関係する専門家の測定値が自動測定と ros,ptr,art,#ps,mlr の評価項目と強く相関している. それに対して, 自然発話に関しては, ポーズの頻度の情報を含む ros,ptr,#ps,mlr と相関しており, 特にポーズの分布よりも頻度のみを考慮している mlr が強い相関を示している.

一方, この研究にいていくつかの問題点が指摘される. 一つは自然発話の測定方法である. 自然発話であるため, 与えられるタスクや被験者の応答が異なるため, 直接比較することが

[3]ポーズは 2 秒以上の無音としている

できないと考えられる．また, 発話を促す質問者も被験者の言語習得度に応じて質問内容も変え, また評点にも差が生じる可能性が高いことも挙げられる．自然発話の場合, 評定者の主観が介入する可能性があり, 人による公平な測定が困難であると言える．もう一つの問題点として, 専門家の測定基準として, 流暢さ (FL) や話速 (SR) という項目が実際の発音能力を直接反映しているかどうかという疑問が生じる．これらは, 発音の時間的特徴を人が評価した結果であり, 同じく時間的特徴の計算機による自動評価の結果と相関するというのは当然と言えるかもしれない.

ただし, 発音の品質の評価に関して, 時間的測定が重要であることが言える. 特に文発声の発音評価の場合, ポーズの取り扱いも考慮しなくてはならない.

3.3.2　単語強勢検出

峯松ら [23] は日本人による発声の英単語音声に対する韻律的評定の自動化を目的として, シラブル HMM を用いた強制位置の検出を行なっている.

強勢音節検出過程を図 3.3 に示す. 学習者に発声させる英単語を提示し, それを見て学習者発声する形態のシステムを想定して, 入力単語のスペル (発音記号列), 音節数, 強勢音節数は既知とする. 単語強勢であるため, 強勢音節数は 1 としている. 入力単語音声に対してシステムが音響的パラメータを抽出し, それを用いて単語を音節に分解する．音節分節は強勢パターン候群と照合される．強勢パターンの候補数は, 単語内の強勢音節数を 1 としているため, 単語内の音節数通りとなる．強勢パターン中の強勢音節モデルと弱勢音節モデルは, 単語内位置属性 (頭/尾/ 他), 音節構造属性などを参照し, 複数の音節カテゴリのうち該当するものの中で照合が行なわれ, 照合尤度が最大になったものが選択され, 選択された強勢パターンのうちの強勢音節位置が検出結果として出力される.

さらに, 峯松らは強勢位置検出に用いられる特徴パラメータのうち, 母音の持続時間に対する重みを固定して, 残りの母音の品質, パワー, ピッチについての重みを変化させ検出率を比較している. 各重みパターンと検出率を視覚化し, 日本人話者, 英語母国語話者における最高検出率に対応する重みパターンの分布を図 3.4 に示す．この図は高さアクセントというアクセント体系を持つ日本語を母国語とする日本人話者は声の高さの制御によって, 本来強さアクセント体系を持つ英語のアクセントを生成してしまっていることを反映している.

このような三角表示によって, 学習者の発音の傾向が視覚化され, これを提示することにより英単語の発音の学習ができる. しかしながら, 英単語を孤立発声した場合と, 実際のコミュニケーションに使われる文で発声した場合に強勢が変わる可能性もあるため, 文強勢への拡張が課題となる.

3.3.3　文強勢検出 (1)

井本ら [24][25] は, 英語文強勢の教育システムの構築を目指し, 文強勢知覚のモデル化と自動検出方法の検討を行なっている．基本周波数, パワー, 母音の持続時間の特徴パラメータそれぞれに対して文強勢の検出を行ない, また各特徴パラメータを重み付け線形結合した線形識別関数を定義し, それに基づき文強勢のモデル化を行なっている．各パラメータの離

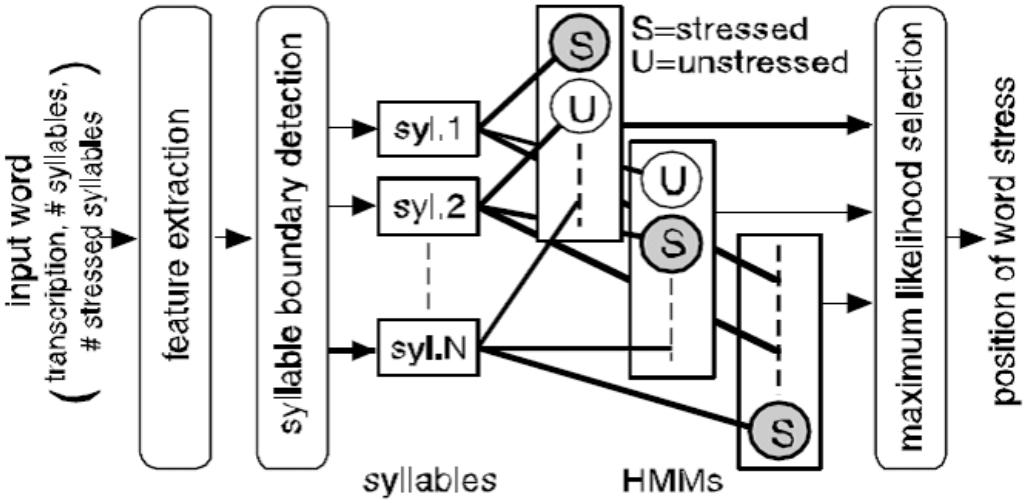

図 3.3: 強勢音節自動検出器

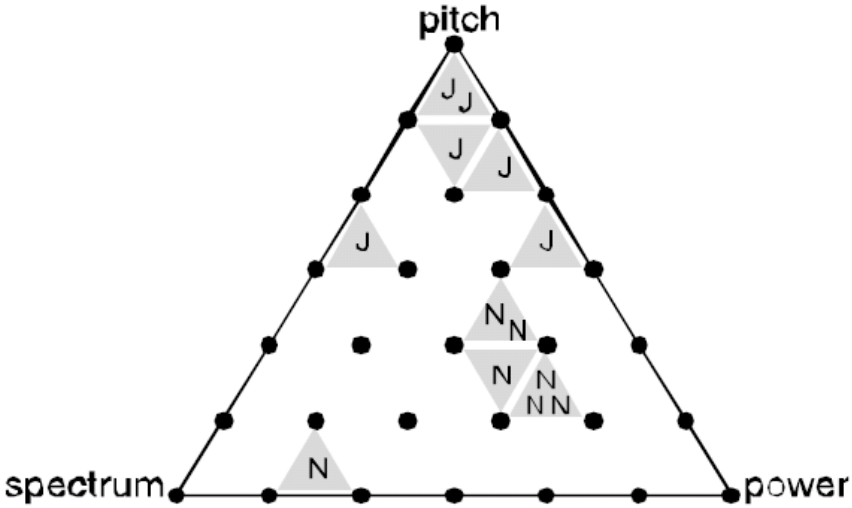

図 3.4: 日本人話者, 母国語話者における最高検出率に対する最小三角形の分布

散化について, 音節単位での基本周波数降下度, 平均化したパワーの音節内の最大値, 音節内の母音の短母音/長母音/二重母音/曖昧母音の各カテゴリ毎に標準化した母音持続時間, 曖昧母音からの距離を用いている. 母国語話者の音声に対する文強勢検出率 (約 95%) に対して, 日本人話者に対する検出率はそれと比較してやや低い結果 (約 84%) となっている. ここでは, 強勢を第一強勢/第二強勢とし, 弱勢を無強勢とし,3 段階の強勢レベルの検出を行なっている.

3.3.4　文強勢検出 (2)

小橋川ら [26] は, 英語文リズム学習支援を目的として, 文強/弱勢音節のモデル化, 及びその検出に関する技術を構築した. この際, 強勢と弱勢の差異が, 音の強さのみならず, 高さ, 長さ, そして母音の音質まで現れるとの知見から, 種々の韻律的特徴に加え, 分節的特徴である低次のケプストラムまで特徴量として導入している.

まず, 英語母国語話者により孤立発声された英単語から構築された既存の単語強/弱勢音節モデルを用いて, 米語母国語話者の英語文音声に対して, 文強勢音節検出実験を行なったその結果, 孤立単語に対する同定率の約 9 割程度の性能を示す結果が得られた. そこでは, 基本周波数生成過程モデルにおけるフレーズ成分に着目し, 文強勢を生成する F_0 に対するイントネーションの影響を削除することを実験的に検討し, 効果を示した. また, 音節の持つ構造 (中心母音の前後に子音が連接するか否か) に加え, 上記の各特徴量の重みの最適化により検出精度の向上が見られた.

次に, 英語文音声を収集し, 米語母国語話者による英語文音声から文強/弱勢音節モデルを構築した. その際, 当該音節の音節構造, 句中位置等の強勢/弱勢以外の音節の持つ属性を考慮することで, 数十の音節カテゴリを定義し,HMM を用いてモデル化を行なった (以下, 当該音節モデルとする). 音節切り出しに関しては, 英語音韻モデルによる forced alignment や syllabification プログラム[4]を用いている. 当該音節モデルを用いて, 強勢検出実験を行なったところ, 話者クローズ&テキストオープン実験において, 約 91%の検出性能が得られた. また, 他の話者に同手法を用いたところ, 米語母国語話者に対しては約 84%, 日本人話者に対しては約 73%の同定率が得られた. この時,「単語内強勢音節数は 1 である」といった強/弱勢ラベル手法の原則に基づく制約を加えている.

また, 音節構造の差における検出性能差を軽減, 及び実装の簡便性を考慮して, 母音区間のみを考慮したモデル化を行なった. 米語母国語話者に対しては, 数% 性能の劣化が観測されたが, 日本人話者に対して, 約 4%の性能向上が得られた.

そして, 強勢の本質が周囲の音節と比較してより,“目立つ” 存在にあると考え, 周囲の音節間の韻律に関する差分情報を用いたモデル化, 及び強勢検出を行なった (以下, 隣接音節正規化モデル). そこでは, 当該音節とその前後に隣接する音節における強さ, 高さに関する韻律的特徴の平均を基準値とした正規化により隣接音節との差分情報を表現している. 隣接音節正規化モデルと当該音節モデルの強/弱勢判定の尤度の線形和により, 両モデルを統合し, 強勢検出を行なった (以下, 統合モデル). その結果, 話者クローズ&テキストオープン実験において, 約 92%の検出性能が得られた. また, 米語母国語話者に対して約 85%, 日本人話者に対しては約 74%の性能が得られた. また, 当該音節/隣接音節正規化モデルの両結果が一致した場合のみの部分データに関しては, 話者クローズ&オープンの両実験においてともに約 90%の検出性能が得られた. さらに, この両話者による文音声がリズムに着目して発声された文であることから, 話者性の正規化のみならずリズムに特化したモデルであることが示唆された.

さらに, 検出精度向上の為, 隣接音節正規化モデルにおける正規化の基準値の選択手法に関して実験的検討を加えた. そこでは, 当該音節の頭/末時刻における音の強さ/高さに関す

[4]tsylb2, http://www.nist.gov/speech/tools/

る韻律的特徴を基準値に用いた正規化により検出精度の向上を計った. その結果, 話者オープン&テキストクローズ実験において, 約 94% の検出性能が得られた. また, 米語母国語話者に対して約 89%, 日本人話者に対して約 78%の同定率が得られた.

この研究ではまず,forced alignment の際, 英語母国語話者音声で学習した音韻モデルを用いている等, 母音挿入をはじめとする日本人英語特有の発音誤りに対する検討が少ない. 特に, 音節切り出しに関して, 母音挿入等による音節数の増加や音節構造の誤りは重大な問題となる.

3.4　フィードバックのレベル

発音教育システムにおける重要な要素としてシステムから学習者に与えるフィードバックが挙げられる. 以下に発音教育システムのフィードバックによるレベル分けについて整理した.[27]

3.4.1　レベル 1 のフィードバック:認識結果の表示

システムのユーザに対するフィードバックとしてユーザの音声から音声認識の結果をテキストとしてスクリーン上に表示する. 通常の音声認識システムの出力はテキストであるので, 発音学習への有効性は限られているが,CALL システムとして音声認識システムはレベル 1 のフィードバックとしてたびたび用いられる.

例 : rice[ráɪs](米) と lice[láɪs](虱) の発音

S:「Please speak,“I like rice”.」

U:「アイ ライク ライス」

S:「I think you said, “I like lice”」

S:システム　U:ユーザ

3.4.2　レベル 2 のフィードバック:スコアの表示

レベル 2 のフィードバックは, レベル 1 の単純な認識を増強して, 発音の音韻の正確さのスコアを表示する. 音声認識器は, 蓄積された母国語話者による正確な発音とユーザの発声の近さの尤度を用いているため, これらの数値を変換してスコアを与える.

レベル 2 のフィードバックは時に母国語話者に対し低いスコアを与えたり, 非音声や不正確な発音に高いスコアを与えることがある. マイクの高品質化や雑音の低減などにより改善されることもある.

音素レベル, 単語レベルなどのスコアは, 多くの比較が行なわれるため, 望ましいスコアが与えられる. 一方, 文や発声全体のレベルでの有効なスコアを与えるには, 多くの正確な文や発声全体のデータが必要となる.

3.4.3　レベル 3 のフィードバック:発音矯正のヒントの表示

スコアつきの認識に加え, 発音誤りに対する向上のヒントを表示する. スコアが低い場合正確な発音を発声するためのヒントが与えられる.

> 例 フランス語の/y/という音の発音に関して
>
> S: “ou” と発音するような唇の形で, “ee” の音を発声してみて下さい.

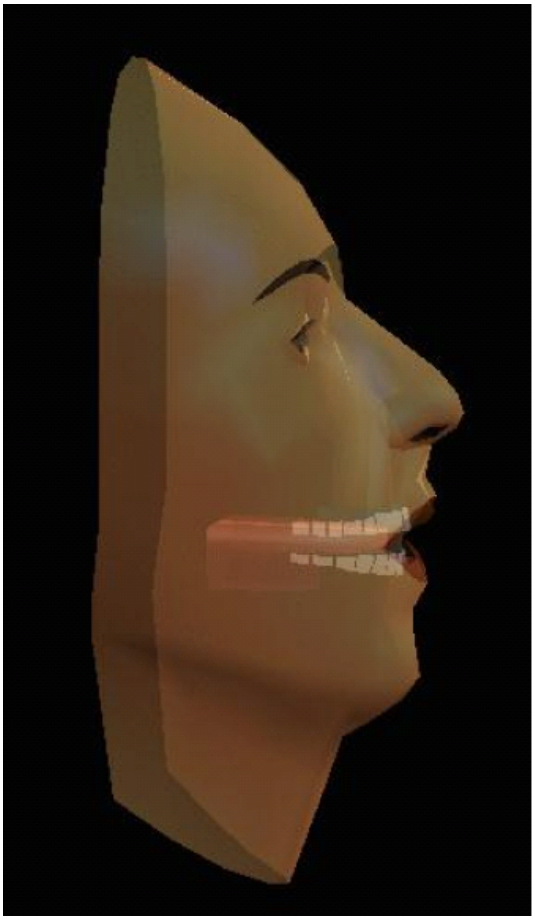

図 3.5: コンピュータアニメーショントーキングヘッド

レベル 3 のフィードバックのグラフィカルな例として, コンピュータアニメーショントーキングヘッド Baldi[28](図 3.5) の存在が挙げられる.

Massaro ら [29] は, 言語学習における視覚による音声知覚について述べている. 読唇術に代表されるように, 騒音, 帯域制限や聴覚障害のために聴覚による音声の品質が低下している場合, 視覚による音声知覚は特に効果的となる.

Baldi は,text-to-speech 分析によって制御され, 合成音声や自然音声を発声し, パラ言語的情報や感情も表現する.

心理学によると間隔をあけた長時間に渡る学習が, 短期集中の学習よりも良い学習になることが実証されており, いつでも利用できる Baldi は有効な学習の手段として用いることができる.

Baldi の皮膚は透明にすることもできる為, 正確に発音するための調音器官の位置などを観察することにより学習することができる.

3.4.4　その他：学習者の音声によるフィードバック

Fred ら [6] は学習者の音声を修正してフィードバックする日本語アクセントパターン CALL システムを提案している.

従来のフィードバックは母国語話者の音声とアクセント核（ピッチの落ちる場所）の表示をする視覚的なものであった. 提案手法では, 学習者自身の声によって正しい音声のフィードバックを行うことで, 学習者によりわかりやすく, 聞いた音声を再構成しやすくしようと試みている. この方法のもうひとつの特徴は元の音声と, 正しい自身の音声を比較することでより間違いの場所や違いがわかり易くなっているところである.

これらの案を, ここでは Time Domain Pitch Synchronous OverLap Add(TD-PSOLA) によって母国語話者の特徴量を抽出し, 学習者の誤った音声に対して適用している.TD-PSOLA については第 4 章で詳しく見ていく.

3.5　システムの評価手法

本章では発音教育システムに対する評価手法の例を述べる.

3.5.1　専門家のスコアとシステムのスコアの相関による評価

Bernstein ら [30] は不特定話者 HMM[5]音声認識システムから得られる日本人の英語文朗読音声のスペクトル距離尤度を評価スコアとして使用し, 専門家による発音品質の評価スコアの相関によりシステムを評価している.

Hamada ら [31] は, 複数の評価手法を組み合わせた評価基準を提案している.

音声スペクトルの静的特徴の評価指標 (E_{s1},E_{s2}) として非母国語話者と母国語話者の音声スペクトルの一致頻度を表す写像ベクトルにより計算されたものを用いる.

スペクトル列の動的特徴を,DTW マッチング法により計算される非母国語話者と母国語話者の単語発声におけるスペクトル距離を評価している (E_d).

韻律的特徴のパラメータとして基本周波数 (E_{f1},E_{f2}) と音声パワー (E_{p1},E_{p2}) を非母国語話者と母国語話者との間の発音パターンのフレームごとの差により計算される評価手法を用いる.

これらの複数の評価手法に各々対して, 専門家による評価との相関係数を計算することによって, 図 3.6 に示すように相関の高い指標が選択される. 選択された指標に対して多重回帰分析が行なわれ, これらの指標を組み合わせることにより新しい評価基準
$f(E_{s,ave}, E_{d,min}, E_{f2,ave}, E_{p1,ave})$ が求められる. この新しい評価基準によるスコアを縦軸に, 専門家によるスコアを横軸にプロットしたものを図 3.7 に示す.

3.5.2　システムの使用前後の学習者の向上度による評価

Akahane-Yamada ら [33][32] は学習の前後にテストを行い, 学習者の向上度の評価を行なっている.

[33] では, 日本人にとって知覚, 発声が難しいとされる英語の流音 (/r/と/l/) の知覚訓練の発声への効果を調査している. 河合ら [4] によると, 日本人の発音評価に関して, 流音の評

[5]隠れマルコフモデル:不確定な時系列のデータをモデル化するための有効な統計的手法

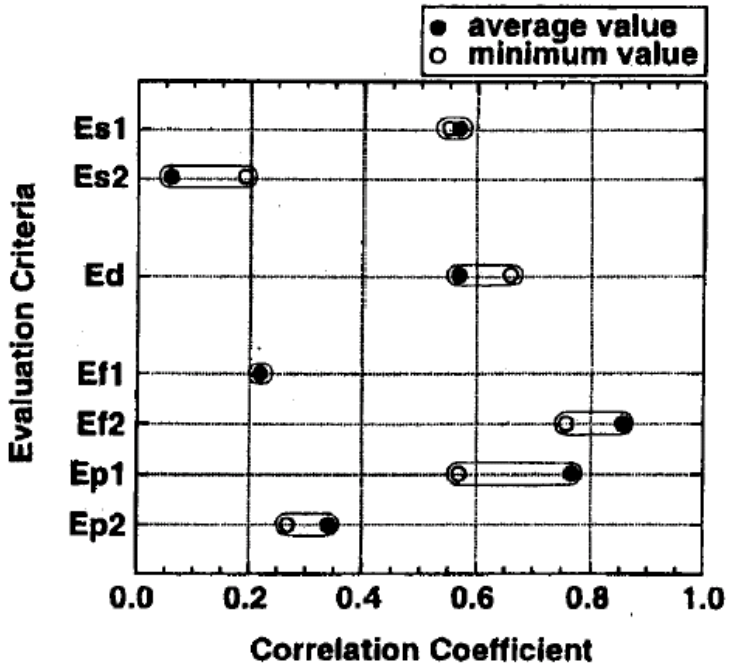

図 3.6: 評価基準と人による審査の相関分析結果

価と発話全体の評価に高い相関が見られ, 流音の測定が発音能力の向上に役立つ情報を得やすいことを示している.

日本人被験者は,/r/と/l/のみが異なる単語ペアのうち, ヘッドフォン越しに再生された一方の単語を聴いて選択する二者択一の知覚テストを受ける. その際, 正解/不正解などのフィードバックはない. この知覚テストは, 照合グループは連続した 2 日に渡って行なわれ, 被訓練者のグループは学習の前後に行なわれる.

日本人被験者のうち, 被訓練者のグループは知覚テストに用いられたものと同様のタスクが訓練段階に用いられる. しかしながら, フィードバックとしてチャイム (正解) やブザー (不正解) の応答や累積正解率の表示 (図 3.8 右上) が与えられ, さらに, 図 3.8 のようなグラフィックのコインが 3 回正解につき報酬として 1 枚与えられる.

日本人被験者はランダムに並んだ単語のリストを復唱するタスクを行ない, 録音される. 日本人被験者は単語リストの提示だけでなく, 発音の仕方のモデルとして米語母国語話者による発声も与えられる.

米語母国語話者が日本人被験者の発声の録音に対して分かりやすさを評価する. 別の米語母国語話者が日本人被験者の意図した単語を理解した上での評価も行なう.

発声能力の維持も評価するために,3ヶ月後と 6ヶ月後の追跡調査として日本人被験者の録音の評価が行なわれる.

その結果, 被訓練者のグループは知覚能力に関して改良が見られる. さらに, 被訓練者の発声能力にも改良が見られる. また, 被訓練者の発声能力は 3ヶ月後や 6ヶ月後においても, 訓練直後のテストのレベルで維持されたという結果が示されている.

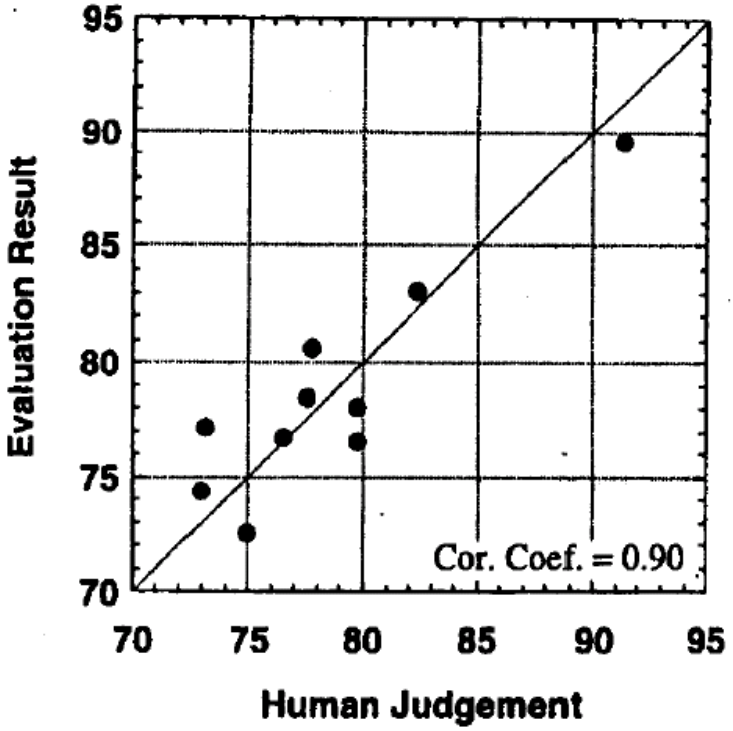

図 3.7: 評価結果と人による審査の関係

[32] では, 図 3.9 に示すようなフォルマント周波数を重ねたサウンドスペクトルグラム[6] のフィードバックと HMM による評価スコアのフィードバックが共に効果的であることを示している. この論文では追跡調査は行なっていないが, 訓練の前後で発声の分かりやすさに関してテストを行なっている.

[6]音声スペクトルの時間的な変化を, 濃淡図形によって目に見える形で表示したもの. 声紋 (voice print).[9]

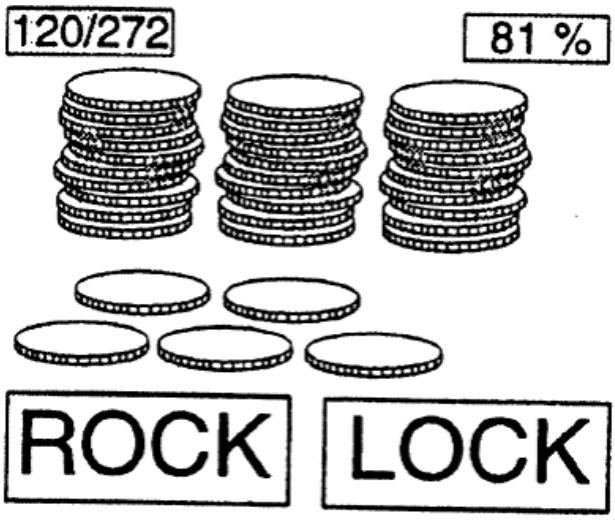

図 3.8: 訓練セッションにおける画面表示

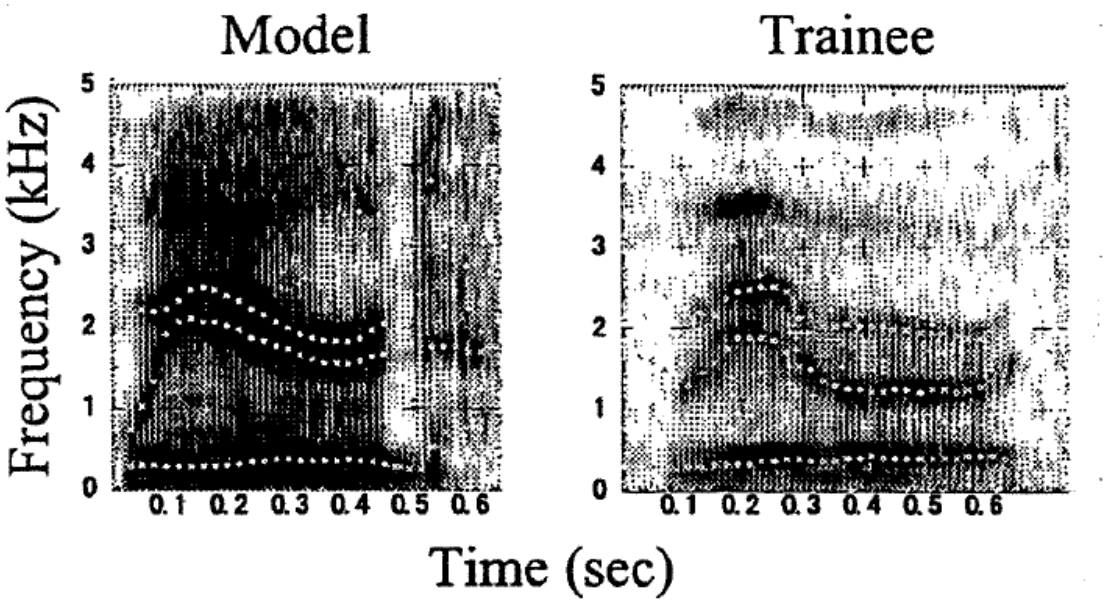

図 3.9: ピッチ追跡とサウンドスペクトログラム

第4章

TD-PSOLA法

4.1　はじめに

本研究でフィードバック部において用いた, ピッチ, 長さ, パワーを修正することができる TD-PSOLA 法について述べる.

4.2　TD-PSOLA 法

TD-PSLA 法は以下の 3 段階で構成される.

1. 元の音声の分析
2. 中間データのための修正
3. 中間データからの再合成

これをまとめたのが図 4.1 である.

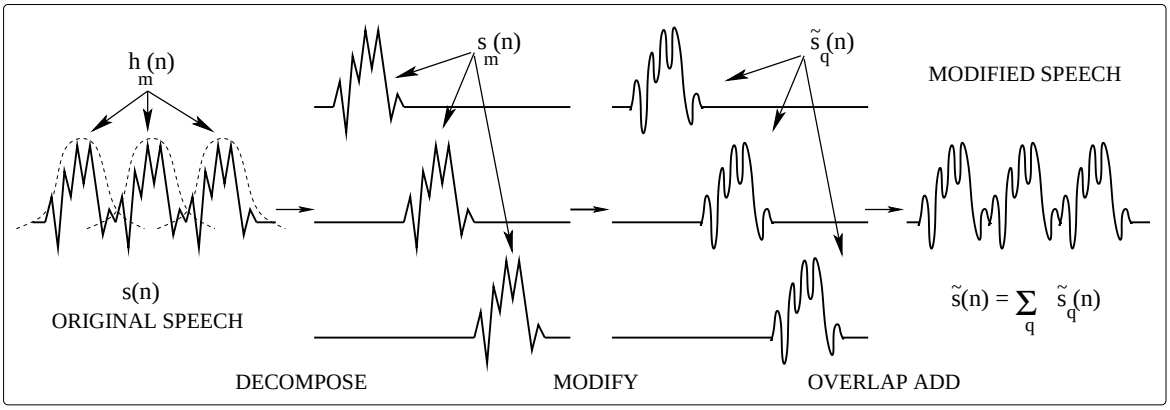

図 4.1: PSOLA の原理

4.2.1　ピッチ同期分析

音声信号を $s(n)$, 一連のピッチ同期窓を $h_m(n)$, 窓をかけた音声信号を $s_m(n)$ とし ST-Signal と名づける.

$$s_m(n) = h_m(t_m - n)s(n) \tag{4.1}$$

窓は hamming 窓などであり, ピッチマークと呼ばれるピッチに同期した時間 t_m を中心に常に 1 ピッチ周期より長くとる.[1]（通常ピッチ周期の 2 倍の長さ）したがって連続した ST-Signal 同士はオーバーラップすることになる.

[1]無声区間は仮のピッチを定めて一定周期でピッチマークをとる

4.2.2　ピッチ同期修正

修正後のピッチマーク $\tilde{t}_q$ にしたがって,$sm_(n)$ を合成信号 $\tilde{s}_q(n)$ に修正する. 修正は以下の 3 つによって行う. （図 4.2 参照）

1. ST-signal の増減
2. ST-signal の間隔の修正
3. ST-signal 自体の修正

窓長 $\mathcal{L}$ は $\mathcal{L} = min\{2(t_{m+1} - t_m), 2(\tilde{t}_{q+1} - \tilde{t}_q)\}$ で取る.

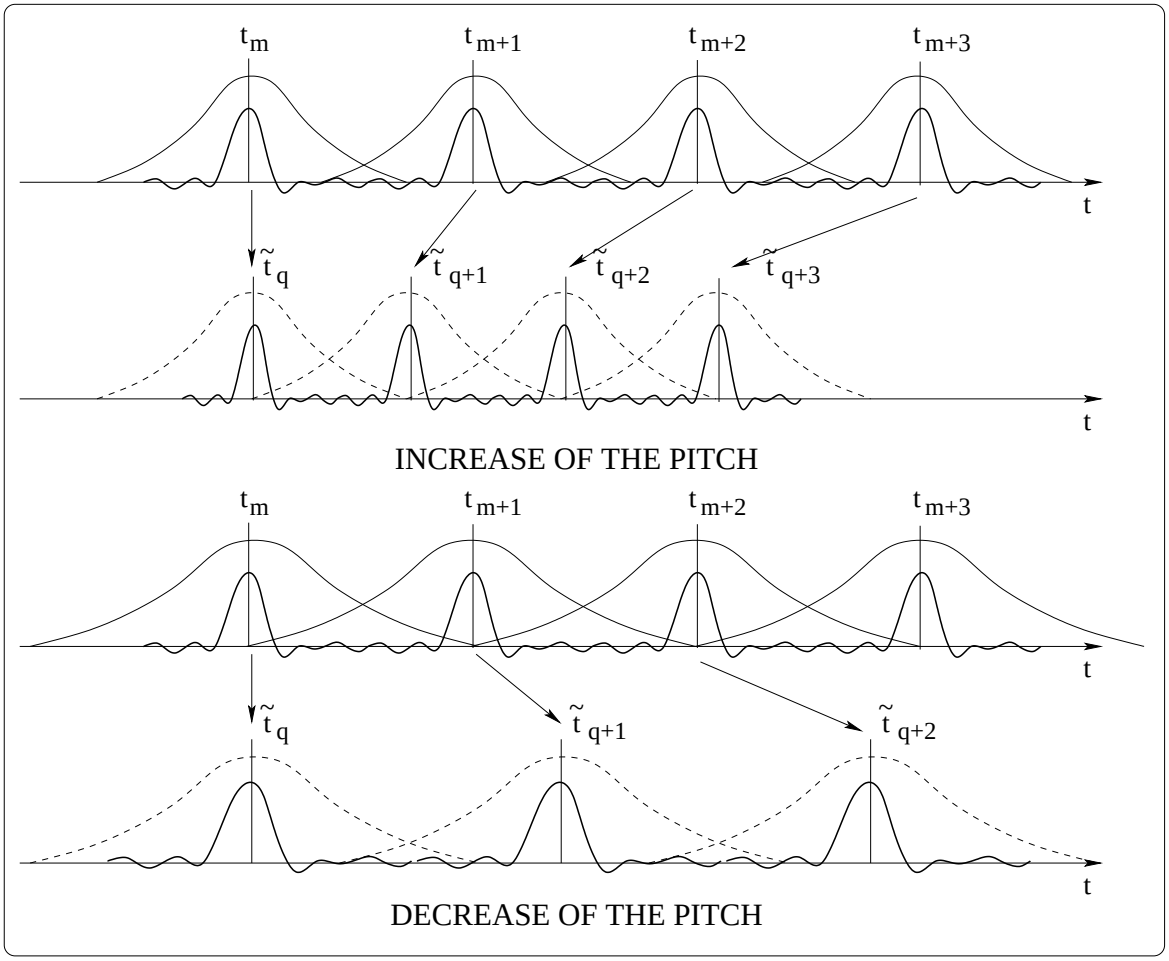

図 4.2: Pitch Mapping Process involved in PSOLA

4.2.3　ピッチ同期波形重畳

修正信号は式 4.2 で得られる.

$$\tilde{s}(n) = \sum_q \tilde{s}_q(n) \tag{4.2}$$

4.3 Automatic Pitch Marking

TD-PSOLA の効果はピッチマークの決定に依るところが大きい. しかし, 学習システムでは手作業によるラベル振りができないので, 自動的に抽出する必要がある.

自動ピッチマーク付与は

1. ピッチ推定
2. ピッチ推定に基づく粗いピッチマーク付与
3. DP によるより精細なピッチマーク付与

の順に行う.

4.3.1 ピッチ推定

最初のピッチ推定は難しい問題であるが, サードパーティー製の get_f0s(Galatea 付属のツール[2]) などである程度の精度で行うことができる.

4.3.2 粗いピッチマーク推定

次に粗いピッチマーク付与である. まず, ピッチ推定時に推定された有声音に対して, ピッチ ($\mathcal{P}_{estimation}[t(i)]$) に比例した間隔でピッチマークを置いていく. 付与したピッチマーク $\{p_r(i)\}$ に対して, 再帰的に次のピッチマークを決めてゆく.$t_r(i) = p_r(i-1) + \mathcal{P}_{estimation}[p_r(i-1)]$ で推定した $t_r(i)$ の周りの局所最大値をピッチマークとする. 無声部分は一定の仮のピッチと仮定する.

4.3.3 精細なピッチマーク推定

ピッチマーク $p_r(i)$ を中心として, 局所ピッチの 2 倍の幅の窓でフレームを抽出. その信号列を i 番目の列とする $N \times P$ 行列 $(m)_{n,p}$ を作る. ここで N は最大ピッチ周期の 2 倍の大きさである. すると $(m)_{n,p}$ は

$$m_{n,p} = h_{Hanning}\left\{s_p\left(p_r(n) - \frac{N}{2} + n\right)\right\} \tag{4.3}$$

とかける. ここで s_p は p 番目の ST-signal である. この行列に対して, 一定のビーム幅で DP 法を適用し最大コストのパス $p_a(i)$ を決定する. ここでコストは式 4.4 で与えられる.

$$C(\{p_a(i)\}) = \sum_{i=1}^{P} m_{p_a(i)-p_r(i)+\frac{N}{2},i} \tag{4.4}$$

制約条件は以下.

$$\forall i \quad (i \neq 1), \quad p_a(i-1) - B \leq p_a(i) \leq p_a(i-1) + B \tag{4.5}$$

[2]http://hil.t.u-tokyo.ac.jp/ galatea/index-jp.html

ここで B はビーム幅である. 以上によりピッチマークはより最大のピークを通るように推定できると考えられる.B を徐々に減らしつつこれを繰り返すことでより精細な値がえられる. 得られた結果を示したのが図 4.3 である.. 元のピッチマーク $p_r(i)$ が図の中心 0 に合わせてあり, 推定されたピッチマーク $p_a(i)$ が線で示してある.

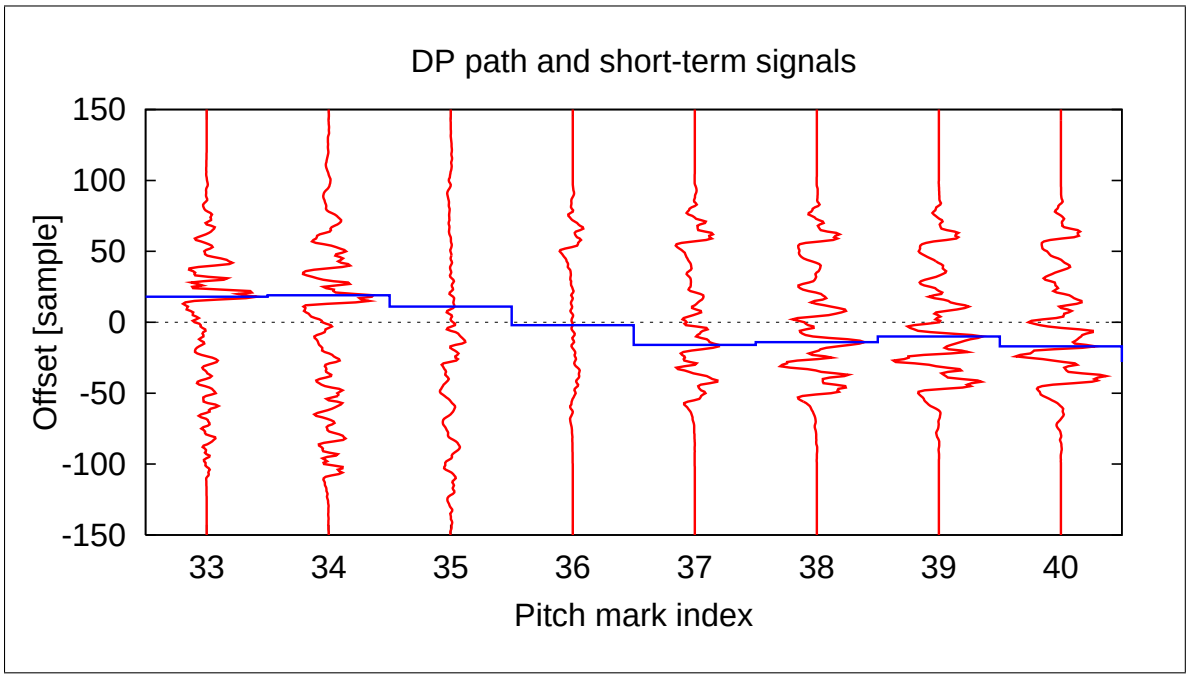

図 4.3: 振幅のピークについてのピッチマークアライメント

4.4　教師マッピング技術

4.5　ピッチ修正

ピッチ修正は, 学習者が発声した誤りであると判定された単語, 文などの音声に対し, 教師音声を適用することで行われる.

4.5.1　ピッチマッピング

教師と生徒で対応するピッチマークがわかれば, 教師のピッチマークに生徒のピッチマークをあわせればよい. ここでピッチマークは教師と生徒のピッチの比 P_{ratio} で補正する.

$$P_{ratio} = \frac{PmD_{ut.}(S)}{PmD_{ut.}(T)} \tag{4.6}$$

ここで,PmD はピッチマーク間隔の平均,$ut.$ は全発話に対して,S は学生,T は教師をあらわす.

4.6　長さ修正

学習者の音声で正しい発音をフィードバックするためには, フィードバック音声は学習者のピッチに近くなければならないが, 話速についても同様のことが言える.

4.6.1　話速

もっとも簡単な話速の定義は, 発話時間 ($L_{ut.}$) をモーラ数 (N_{mora}) で割ったもので定義できる.

$$SR_{Estimation} = \frac{L_{ut.}}{N_{mora}} \tag{4.7}$$

また話速の比は以下のように定義できる.

$$SR_{ratio} = \frac{L_{ut.}\{S\}}{L_{ut.}\{T\}} \tag{4.8}$$

ここで N_{mora} は 2 つの発話で同じとする.

4.6.2　長さ調整

これまでの定義を使って, ある音素に含まれるピッチマークの数を以下のように修正する. 学習者の音素の長さと教師の音素の長さは SR_{ratio} を適用したあとに同じにならなければならない.

$$L_{ph.}(T) \times SR_{ratio} = L_{ph.}^{new}(S) \tag{4.9}$$

ここで "$ph.$" は音素はあらわしている.

また,$L_{ph.}^{new}(S)$ と $L_{ph.}(T)$ はピッチマーク間隔の合計長が等しいことから, 音素について

$$L_{ph.}^{new}(S) = (N_{pm}(S) + X_{new\_pm}(S))\, PmD_{ph.}^{new}(S) \tag{4.10}$$

$$L_{ph.}(T) = N_{pm}(T) \times PmD_{ph.}(T) \tag{4.11}$$

がいえる. ここで N_{pm} は元の音声にあったピッチマークの数, $X_{new\_pm}(S)$ は加減するピッチマークの数を表す. $PmD_{pm}^{new}S$ はまた,

$$PmD_{ph.}^{new}(S) = PmD_{ph.}(T) \times P_{ratio} \tag{4.12}$$

とかけるので, 式 (4.10) (4.11) (4.12) (4.9) から

$$N_{pm}(S) + X_{new\_pm}(S) = N_{pm}(T)\frac{SR_{ratio}}{P_{ratio}} \tag{4.13}$$

がいえる. もうひとつの制約条件として

$$N_{pm}(T) + X_{new\_pm}(T) = N_{pm}(S) + X_{new\_pm}(S) \tag{4.14}$$

があるので, 式 (4.13) (4.14) から最終的に

$$X_{new\_pm}(T) = N_{pm}(T)\left(\frac{SR_{ratio}}{P_{ratio}} - 1\right) \tag{4.15}$$

が得られる.

4.6.3　ピッチマークの加減

表 4.1,4.2 のようにすることでそれぞれピッチマークの増減が実現される.

Original pm	
$Pmark(1)$	$Pitch_1$
$Pmark(2)$	$Pitch_2$
$Pmark(3)$	$Pitch_3$
$Pmark(4)$	$Pitch_4$
$Pmark(5)$	$Pitch_5$
$Pmark(6)$	$Pitch_6$
$Pmark(7)$	$Pitch_7$

$\Longrightarrow$

Modified pm	
$Pmark(1)$	$Pitch_1$
$Pmark(2)$	$Pitch_2$
$Pmark_{New}(1)$	$\frac{Pitch_3+Pitch_2}{2}$
$Pmark(3)$	$Pitch_3$
$Pmark_{New}(2)$	$\frac{Pitch_4+Pitch_3}{2}$
$Pmark(4)$	$Pitch_4$
$Pmark_{New}(3)$	$\frac{Pitch_5+Pitch_4}{2}$
$Pmark(5)$	$Pitch_5$
$Pmark_{New}(4)$	$\frac{Pitch_6+Pitch_5}{2}$
$Pmark(6)$	$Pitch_6$
$Pmark(7)$	$Pitch_7$

表 4.1: ピッチマークの追加

Original pm	
$Pmark(1)$	$Pitch_1$
$Pmark(2)$	$Pitch_2$
$Pmark(3)$	$Pitch_3$
$Pmark(4)$	$Pitch_4$
$Pmark(5)$	$Pitch_5$
$Pmark(6)$	$Pitch_6$
$Pmark(7)$	$Pitch_7$
$Pmark(8)$	$Pitch_8$
$Pmark(9)$	$Pitch_9$
$Pmark(10)$	$Pitch_{10}$

$\Longrightarrow$

Modified pm	
$Pmark(1)$	$Pitch_1$
$Pmark_{New}(1)$	$\frac{Pitch_2+Pitch_3}{2}$
$Pmark_{New}(2)$	$\frac{Pitch_4+Pitch_5}{2}$
$Pmark_{New}(3)$	$\frac{Pitch_6+Pitch_7}{2}$
$Pmark_{New}(4)$	$\frac{Pitch_8+Pitch_9}{2}$
$Pmark(10)$	$Pitch_{10}$

表 4.2: ピッチマークの削除

第5章

Corrective Feedback System

5.1　はじめに

本章では, 構築した, 日本人のための英語リズム学習システムについて述べる.

5.2　本システムの概要

既に提案された, 日本語学習者のためのアクセント型学習システム [6] では, 学習者の間違った日本語アクセント型の音声に対して, 教師の正しい F0 と duration とパワーを TD-PSOLA によってマッピングする. 学習者の声でフィードバックすることのメリットとして, 問題となる点を修正することによってどんな発声になるかを具体的に把握でき, 先生の声真似の練習に陥らない点があげられる. 本研究では, 同様の手法を利用し, 日本人による英語の第 2 章に示したような問題を直すための CALL システムを構築する. 学習者自身の声で, 母音挿入を修正し強勢弱勢が正しく配置された音声をフィードバックすることで学習者に発音のどこが悪いのかを提示する. 日本人による英語は母音挿入によってリズムが大幅に崩れるため, 挿入母音がある場合は修正時にその母音の削除を行う必要がある. 次に F0・duration を教師音声からマッピングした. 構築したシステムの概念図を図 5.1 に示した. 以下に詳細を示す.

5.3　発音誤りの検出

日本人による英語リズム学習に主眼を置いた本システムでは, 日本人が発話した際にリズムに特に影響を及ぼすと考えられる, 母音挿入誤りと強勢弱勢に関して検出を行っている. ここではそれぞれについて詳しく見ていく.

5.3.1　母音挿入

2.2.2.1 で述べたように, 日本人の母音挿入はある程度規則性がある. そこで,2.5.1 で述べた英語不特定話者音響モデルと, 母音挿入部分のみ日本語不特定話者音響モデルを用いための日米混合 monophone HMM 音響モデルを構築し, 表 2.2 を参考にしてネットワーク文法を用いて母音挿入を予測し検出した.

5.3.1.1　日本語 HMM の構築

音声認識システム [34] 付属の CD にある monophone モデル（男 or 女 or 性別非依存 (GID)）を使っている. データは音素バランス文＋新聞記事読み上げ（男女各 20k 文/132 名）である. この音響モデルで使われている音素を表 5.1 に示した.

表 5.1: 日本語音素

a i u e o a: i: u: e: o: N w y p py t k ky b by d dy g gy ts ch m my n ny h hy f s sh z j r ry q sp silB silE

5.3.1.2　英語 HMM の構築

英語の音響モデルは DARPA の
CSR-I(WSJ0) corpus 男女各 4 千文, 約 50 名）を使って学習したものを用いた. HMM の構築における各パラメタを表 5.2 に示す.

表 5.2: 音響分析条件

サンプリング周波数	16kHz
プリエンファシス	0.97
分析窓	Hamming 窓
分析窓長	25ms
窓間隔	10ms
特徴パラメタ	MFCC(12 次) + Δ MFCC + Δ Pow(計 25 次元)
周波数分析	等メル間隔フィルタバンク
フィルタバンク	24 チャネル

triphone の構築は, 日本語音素＋英語音素の音素環境のデータが得られないこと, 学習データ統合のためのトップダウンクラスタリングにおける Question Set に関して作成が困難であったので本システムでは行わなかった.

5.3.2　強勢弱勢検出器の構築

3.3.4 で触れた強勢検出技術を使って, 強勢・弱勢検出器を構築した.

5.3.2.1　強勢・弱勢データ

学習データとして, 英語教師により 54 文を用意した. なお 54 文は表 5.5 に記載した. 表において #は単語境界,

は音節境界を表す. 文を選定するに当たっては,1 文あたり 5 単語〜10 単語程度で, 母音挿入が起こりやすい子音で終わる単語が多く使われ, 意味が比較的わかりやすいものとした. これらの文を, カリフォルニア出身の米語男性話者によって各 2 回,108 発声, 932 シラブルを読んでもらい, それに対して, 英語教師 1 名が 3 段階のラベルを振ったものを用意した. 実

際のシステムでは, 強い, 弱いのメリハリを重視するために,3 段階の最も強い段階を強勢, 残り 2 つの段階を弱勢として 2 値のラベルにより学習を行った.

5.3.2.2　音響パラメータ

前述した不特定話者英語音響モデルを用いて, 音素セグメンテーションを行った後に, syllablification software を用いて音声を音節ごとに分け, それぞれの音節に対して, 表 5.3 のような分析条件で表のパラメタのベクトル系列を抽出した. なお,F0 は基本周波数生成モデル（付録 A）によって文全体でスムージングしたものを用いた.

表 5.3: 音響分析条件

サンプリング	12kHz/16bit
分析法	14 次の MelCep 分析
フレーム幅	21.3 msec (256 samples)
フレーム周期	8.0 msec (96 samples)
ピッチ抽出	遅れ時間幅に比例したフレーム幅を使用した自己相関法
抽出周期	8.0 msec (96 samples)

表 5.4: HMM 学習条件

特徴量	$F_0 + \Delta F_0$, ΔPOW, MCEP(1〜4)+ΔCEP (合計 11 次元)
topology	6 状態 4 分布, left-to-right
	継続時間長制御
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
学習用音声	英語母国語話者 (GA) 男性 1 名による 108 文発声

5.3.2.3　学習

上で得た音節ごとのベクトル系列に対して, 正解ラベルを強勢 (S), 弱勢 (W) として各音節に振り,HMM として Forward-Backward アルゴリズムで学習し,monophone 強勢弱勢音響モデルを得た.

先行研究 3.3.4 では CVC や CV などの音素環境や短母音, 長母音, 二重母音などのカテゴリに分けてさらによい検出率を得ていた. しかし, ここでは学習データが 10 文の 1 程度しかないので, シンプルなモデル化を行った. 後置音節が強勢 (S) で当該音声が強勢 (S) の SS, 後置音節が強勢 (S) で当該音節が弱勢 (W) の WS, 後置音節が弱勢 (W) で当該音節が強勢 (S) の SW, 後置音節が弱勢 (W) で当該が弱勢 (W) の WW の 4 つのラベルを定義し, それぞれの音節に正解データを振ったものを使って, 後置音節依存 bi-phone 強勢弱勢 HMM 音響モデルを作成した.

5.3.3　強勢弱勢検出器の予備評価実験

上述した HMM を用いて強勢, 弱勢音節を検出した結果, 話者クローズ, テキストクローズで,monophone 強勢弱勢 HMM 音響モデルで 68%, 最後端の音節を monophone 強勢弱勢 HMM で認識し, そこから前の音節を後置音節依存 bi-phone 強勢弱勢 HMM 音響モデルで認識した結果,73%の認識率を得た.

話者クローズ, テキストオープンでは 107 文:1 文の Cross Validation で bi-phone により,69%の結果を得た.

5.4　評価・判定部

本システムでは, 母音挿入があった場合と, 教師音声の強勢弱勢正解ラベルと学習者音声の強勢弱勢正解ラベルが違った場合に誤りありとし, 音声の修正を行うこととした.

5.5　フィードバック部

ここでは本システムのフィードバックに関して, 学習者音声の修正, 表示について述べる.

5.5.1　挿入母音, ショートポーズの削除

非母語英語話者は母語話者がいれないようなところにショートポーズ（sp）を入れがちである. また, 母音挿入は英語の音節構造を変化させてしまうため, 発音の理解度に重大な影響をもたらすと言われている.

そこで, 本システムでは音声認識結果と教師音声との対応付けで挿入されたと判断された sp（ショートポーズ）と, ネットワーク文法で認識した挿入母音をリズムに悪影響を及ぼすものと考え, 削除するようにした.

5.5.2　ピッチ・duration 修正

第 4 章で述べた PSOLA 法を用いて, 同一文を読んだ教師音声を手本として,TD-PSOLA によってピッチの変換,duration（持続時間）の変換を行った. あらかじめ行った HMM による音素セグメンテーションを用い, 学習者音声と教師音声の音素の対応をとり,durtaion については, 対象とする母音の全体の発話時間に対する長さの比が両者で同じになるように学習者の母音継続長を変換した. ピッチについては教師の平均ピッチからの対数変化を, 学習者の平均ピッチからの対数変化にマッピングすることでピッチを決定した.

5.5.3　表示

音声波形とピッチを表示し, 音素アライメント結果を音素とともに表示した. 音素のうち母音に関して, 強勢と判定された場合は大きめに, 弱勢と判定された場合は小さめに表示す

るようにした. 誤りを検出した部分に対しては, 赤い字で該当部分を指摘するようにして, 直感的でわかりやすくい提示を目指し, ユーザーに修正を促すようにした. 図 5.1 にシステムの表示例を示す.

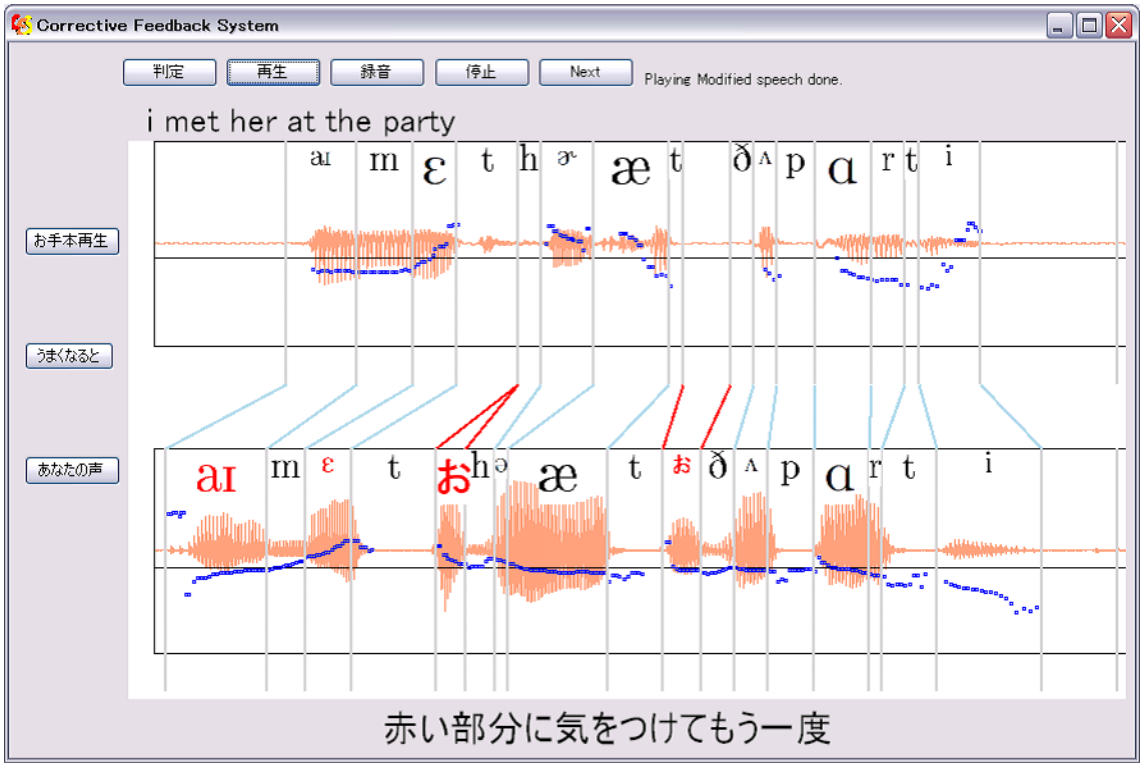

図 5.1: システム表示例

5.6　修正音声の英語らしさの評価

TIMIT460 文から選んだ 22 文について日本人が読んだものと, 母音挿入を削除し, F0 とパワーを教師音声のそれによって修正したものについて米語話者 (CA 出身) 男性 1 名にどちらが英語らしいかを 2 択で選択させた. その際, 音声の質については無視するように指定した. 結果,22 文中 21 文 (95 %) について元の音声より修正後の音声のほうが英語らしいと評価された.

表 5.5: 音節に分けた母音挿入しやすい英語文 54 文

(tell him to do it)
[t E l] # [h I m] # [t x] # [d u] # [I t]
(i met her at the party)
[Y] # [m E t] # [h R] # [@ t] # [D A] # [p a r] [t i]
(tom has oranges and apples)
[t a m] # [h @ z] # [o] [r I n] [J I z] # [@ n d] # [@] [p x l z]
(ask him to do it)
[@ s k] # [h I m] # [t x] # [d u] # [I t]
(she likes driving a car)
[S i] # [l Y k s] # [d r Y] [v I G] # [x] # [k a r]
(don't smoke in this area)
[d o n t] # [s m o k] # [I n] # [D I s] # [e] [r i] [x]
(he met a friend of mine)
[h i] # [m E t] # [x] # [f r E n d] # [A v] # [m Y n]
(he had a glass of milk)
[h i] # [h @ d] # [x] # [g l @ s] # [A v] # [m I l k]
(this story is very interesting)
[D I s] # [s t o] [r i] # [I z] # [v E] [r i] # [I n] [t r I] [s t I G]
(do you often go to philadelphia)
[d u] # [y u] # [c f] [t I n] # [g o] # [t u] # [f I] [l x] [d E l] [f i] [x]
(she can swim very fast)
[S i] # [k @ n] # [s w I m] # [v E] [r i] # [f @ s t]
(he plays the flute well)
[h i] # [p l e z] # [D I] # [f l u t] # [w E l]
(tony did it by himself)
[t o] [n i] # [d I d] # [I t] # [b Y] # [h I m] [s E l f]
(i have plenty of dresses)
[Y] # [h @ v] # [p l E n] [t i] # [A v] # [d r E] [s I z]

表 5.6: 音節に分けた母音挿入しやすい英語文 54 文 2

(there were hundreds of problems)
[D e r] # [w R] # [h A n] [d r I d z] # [A v] # [p r a] [b l x m z]
(we found them in the gymnasium)
[w i] # [f W n d] # [D E m] # [I n] # [D I] # [J I m] [n e] [z i] [x m]
(i'll go out next sunday)
[Y l] # [g o] # [W t] # [n E k s t] # [s A n] [d e]
(copy as a text file)
[k a] [p i] # [@ z] # [x] # [t E k s t] # [f Y l]
(look at that screen carefully)
[l U k] # [@ t] # [D @ t] # [s k r i n] # [k e r] [f x] [l i]
(i met him on the street yesterday)
[Y] # [m E t] # [h I m] # [c n] # [D I] # [s t r i t] # [y E] [s t R] [d e]
(we'll go to australia next spring)
[w I l] # [g o] # [t u] # [c] [s t r e l] [y x] # [n E k s t] # [s p r I G]
(these texts are difficult for us)
[D i z] # [t E k s t s] # [x r] # [d I] [f I] [k x l t] # [f o r] # [A s]
(please solve the problem promptly)
[p l i z] # [s c l v] # [D A] # [p r a] [b l x m] # [p r a m p t] [l i]
(dogs eat meat)
[d c g z] # [i t] # [m i t]
(the dogs eat the meat)
[D x] # [d c g z] # [i t] # [D x] # [m i t]
(the dogs were eating the rotten meat)
[D I] # [d c g z] # [w R] # [i] [t I G] # [D x] # [r a] [t x n] # [m i t]
(the dogs were eating the rotten meat at the back of macdonald's)
[D I] # [d c g z] # [w R] # [i] [t I G] # [D x] # [r a] [t x n] # [m i t] # [@ t] # [D x] # [b @ k] # [A v] # [m x k] [d a] [n x l d z]

表 5.7: 音節に分けた母音挿入しやすい英語文 54 文 3

(tell her to do it)
[t E l] # [h R] # [t x] # [d u] # [I t]
(i met him at the party)
[Y] # [m E t] # [h I m] # [@ t] # [D A] # [p a r] [t i]
(tom likes oranges and apples)
[t a m] # [l Y k s] # [o] [r I n] [J I z] # [@ n d] # [@] [p x l z]
(ask her to do it)
[@ s k] # [h R] # [t u] # [d u] # [I t]
(he likes driving a car)
[h i] # [l Y k s] # [d r Y] [v I G] # [x] # [k a r]
(don't smoke in this room)
[d o n t] # [s m o k] # [I n] # [D I s] # [r u m]
(i met a friend of mine)
[Y] # [m E t] # [x] # [f r E n d] # [A v] # [m Y n]
(i had a glass of milk)
[Y] # [h @ d] # [x] # [g l @ s] # [A v] # [m I l k]
(this book is very interesting)
[D I s] # [b U k] # [I z] # [v E] [r i] # [I n] [t r I] [s t I G]
(do you often go to macdonald's)
[d u] # [y u] # [c] [f x n] # [g o] # [t x] # [m x k] [d a] [n x l d z]
(he can swim very fast)
[h i] # [k @ n] # [s w I m] # [v E] [r i] # [f @ s t]
(she plays the flute well)
[S i] # [p l e z] # [D A] # [f l u t] # [w E l]
(mike did it by himself)
[m Y k] # [d I d] # [I t] # [b Y] # [h I m] [s E l f]

表 5.8: 音節に分けた母音挿入しやすい英語文 54 文 4

(she has plenty of dresses)
[S i] # [h @ z] # [p l E n] [t i] # [A v] # [d r E] [s I z]
(there are hundreds of problems)
[D e r] # [R] # [h A n] [d r I d z] # [A v] # [p r a] [b l x m z]
(i found them in the gymnasium)
[Y] # [f W n d] # [D E m] # [I n] # [D I] # [J I m] [n e] [z i] [x m]
(we'll go out next sunday)
[w I l] # [g o] # [W t] # [n E k s t] # [s A n] [d e]
(save as a text file)
[s e v] # [@ z] # [x] # [t E k s t] # [f Y l]
(look at this screen carefully)
[l U k] # [@ t] # [D I s] # [s k r i n] # [k e r] [f x] [l i]
(i met her on the street yesterday)
[Y] # [m E t] # [h R] # [c n] # [D x] # [s t r i t] # [y E] [s t R] [d e]
(she'll go to australia next spring)
[S I l] # [g o] # [t u] # [c] [s t r e l] [y x] # [n E k s t] # [s p r I G]
(these texts are easy for us)
[D i z] # [t E k s t s] # [R] # [i] [z i] # [f o r] # [A s]
(please answer the question promptly)
[p l i z] # [@ n] [s R] # [D x] # [k w E s] [C I n] # [p r a m p t] [l i]
(cats eat fish)
[k @ t s] # [i t] # [f I S]
(the cats eat the fish)
[D I] # [k @ t s] # [i t] # [D I] # [f I S]
(the cats were eating the rotten fish)
[D I] # [k @ t s] # [w R] # [i] [t I G] # [D x] # [r a] [t x n] # [f I S]
(the cats were eating the rotten fish at the back of macdonald's)
[D I] # [k @ t s] # [w R] # [i] [t I G] # [D x] # [r a] [t x n] # [f I S]
[@ t] # [D I] # [b @ k] # [A v] # [m x k] [d a] [n x l d z]

第6章

結論

6.1　はじめに

本論文では, 日本人のための英語リズム学習システムとして, 強勢・弱勢を検出し, 学習者の音声で矯正フィードバックするシステムを提案した. また, 矯正フィードバック音声に関してそのフィードバックがより英語としてよくなっているかどうかを評価した. 本章では, それらについてまとめるとともに, 今後の課題について述べる.

6.2　まとめ

本論文では, 第 2 章において, 基本単位となる英語の音節と日本語のモーラの違い, 日本語のピッチアクセントと英語の強勢アクセントの違い, 英語における強勢拍のリズムと日本語における音節（モーラ）拍のリズムの違い, など日本人が英語を発音する上で留意しなければならない日本語と英語の違いについて明らかにし, それらが原因となって, 音素挿入やリズムの崩れなどがおきてしまうことをまとめた. 次に, 第 3 章では発音教育システムの概要を述べ, システムの中でユーザ入力に対してどうのような評価手法やフィードバックがあるか, またシステムをどう評価すればよいかについてまとめた. 第 4 章では, 本システムで, ピッチの修正,duration の修正などの矯正フィードバックのために使用した TD-PSOLA 法について詳しく説明した. 最後に第 5 章で本論文で提案したシステム, すなわち、強勢・弱勢検出, 音素挿入検出を用いて学習者の音声を評価し, 学習者の音声を修正して矯正フィードバックするシステムについて詳しく述べた. 本システムでは, 挿入音素部分を削除し F0・dutation を教師音声に合わせて変更し, 得られた音声に対して, 元の音声より英語らしいという評価を得た.

6.3　今後の課題

本システムで考えられる課題について列挙した.

- 本システムでは現在, 強勢・弱勢の検出を誤り検出として行っており, 誤っていた場合に教師のピッチと duration をマッピングしているが, どこをどう直せば強勢が弱勢に, 弱勢が強勢になるかが明確になっていない点が問題であり, 実際にパラメータを変えた知覚実験などで, どこが知覚に効いているのかを調査する必要がある.
- 本システムにおいて行ったピッチ修正, 音素挿入削除, ショートポーズ削除は, HMM による音素セグメンテーションを利用して適応範囲を決定しているため, 音質などがセグメンテーションの結果に大きく依存しており, セグメンテーションが正確でない部分では音質が極端に悪くなってしまうという問題がある.
- 学習者音声の認識という問題に対しては, 日本人によるなまった英語に対しての音素セグメンテーションであるため, モデル化時の音素と実際に認識させる音素は大きく違う可能性があるが, それに適した音響モデルの構築は困難である. 便宜的に英語音響モデルと日本語音響モデルを混在させて用いたが, より適した音響モデルが求められる.

- TD-PSOLA ではピッチ修正幅が大きいと音質が大きく劣化してしまう問題があり, より品質のよいピッチ分析修正法を用いた音質の向上が必要である.
- 最後に, 実際の日本人の英語学習者が本システムを使用して学習者の音声で矯正フィードバックした音声を利用した場合に, 有効な学習効果があるかどうかの検証が必要である.

謝辞

本論文を執筆するにあたり, 指導教員である広瀬啓吉教授, また研究室の共同運営者である峯松信明准教授には, 日頃から研究や論文執筆等において, 様々なご指導, ご鞭撻を賜わりました．深く感謝の意を表します.

また, 英語の相談者, ラベラーとして本研究に多大な労力をかけてくださった山内豊東京国際大学教授, 実験データなどを提供し, 協力しててくださった Eric Grunow 君, 井川明彦様, 小橋川哲先輩, 研究の相談に乗ってくださった越智景子先輩には特に深く感謝致します.

また, 研究室環境の整備など, 本研究を様々な面で支援してくださった高橋登技官, 秘書の楠本由香里さん, 前秘書の武田祥子さん, 峯松研秘書の笠島恵さんに, 深く感謝いたします.

最後に,3 年間公私ともにお世話になった八木裕司先輩, 渡辺美知子先輩, 稲垣貴彦君, また修士から 2 年間お世話になった同輩の篠田知宏君, 石井英資君, ナジャンド・アリ君, アントニオ君, エルハン君, ホアン君には感謝の意を表してここに記します.

2008 年 02 月 04 日
三輪 周作

参考文献

[1] H. Fujisaki and S. Nagashima: “A model for synthesis of pitch contours of connected speech,” Annual Report of Engineering Research Institute, University of Tokyo, vol.28, pp.53-60, 1969.

[2] http://www.slp.tutics.tut.ac.jp/CALLsoft/.

[3] 斉藤明子, 長井克己. “VT 法を用いた日本語音声教育の実践報告” 第 1 回日本語音声教育方法研究会, 1999.

[4] 河合剛, 石田朗. “日本人の英語の発音評価の信頼性に関する実験的研究” 信学技報, Vol. ET95-44, pp. 89–96, 06 1995.

[5] http://www.kanto-gakuen.ac.jp/verbo/verbo.htm.

[6] Frederic Gendrin , 修士論文 “Accent Pattern CALL System using Speech Modification based Corrective Feedback” 広瀬峯松研究室

[7] 峯松信明, 藤澤友紀子, 中川聖一. “HMM を用いた英単語音声からの強勢音節の自動検出とそれに基づく発音能力の韻律的評定” 電子情報通信学会誌 D-II, Vol. J82-D-II, No. 11, pp. 1865–1876, 11 1999.

[8] A.C.Gimson 著, 竹林 滋訳. ギムスン “英語音声学入門” 金星堂, 1990.

[9] 古井貞熙 “ディジタル音声処理” 東海大学出版会, 1985.

[10] Hideki Kawahara, Alain de Cheveigné, and Roy D. Patterson. “An instantaneous-frequency-based pitch extraction method for high-quality speech transformation: revised TEMPO in the STRAIGHT suite” Proc. Int. Conf. Spoken Language Processing (ICSLP’98), Vol. 4, pp. 1367–1370, 1998.

[11] 窪園晴夫, 太田聡. 音韻構造とアクセント. 研究社出版, 1998.

[12] 宮地裕, 他編著 川原崎幹夫 “講座 日本語と英語教育 第 9 巻日本語の文字・表記（下）” 明治書院 1989

[13] Takeshi Tarui. “Rhythm Timing in Japanese English” Proc. Int. Conf. Spoken Language Processing (ICSLP2000), Vol. 4, pp. 592–595, 2000.

[14] 竹林滋 “英語音声学” 研究社, 1996

[15] 杉本淳子. “母音長コントロールによる日本人英語学習者による英語リズムの問題点” 日本音声学会 第303回研究例会 研究発表, 6 2001.

[16] 小川直樹. “理屈で分かる英語の発音 特有のイントネーションが身に着くステップ80.” NOVA, 2000.

[17] 河合剛, 石田朗, 廣瀬啓吉. “2言語の音響モデルを用いた音声認識による非母語発音誤りの検出と発音評価” 日本音響学会誌, Vol. 57, No. 9, pp. 659–580, 9 2001.

[18] Microsoft Corporation. “ENCARTA” 2000.

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

[20] Transparent Language. “English Now! Ver.6.0” スリー・エー・システムズ, 1998.

[21] 山田玲子. “英語リスニング/スピーキング” http://atrcall.isd.atr.co.jp/ja/exp/proc.html, 2000.

[22] Helmer Strik, Catia Cuccharini, and Diana Binnenpooete. “L2 PRONUNCIATION QUARITY IN READ AND SPONTANEOUS SPEECH” Proc. Int. Conf. Spoken Language Processing (ICSLP 98), Vol. 3, pp. 582–585, 2000.

[23] 峯松信明, 藤澤友紀子, 中川聖一. “英単語発音上の癖の自動推定・視覚化とそれに基づく発音能力の韻律的評定” 電子情報通信学会誌 D-II, Vol. J83-D-II, No. 2, pp. 486–499, 2 2000.

[24] 井本和範, 壇辻正剛, 河原達也. “CALLシステムのための英語文強勢知覚のモデル化” 信学技報 SP2000-1, pp. 1–8, 5 2000.

[25] 井本和範, 壇辻正剛, 河原達也. “日本人話者の英語文強勢誤りの自動検出” 日本音響学会講演論文集,3-7-2, pp. 343–344, 10 2001.

[26] 小橋川哲, “英語文強勢のモデル化とその発音教育への応用” 修士論文 広瀬峯松研究室 2002

[27] Kathleen B. Egan and Stephen A. LaRocca. “Speech Recognition in Learning: A Must” Proc. InSTIL2000, pp. 4–9, 2000.

[28] UCSC Perceptual Science Laboratory. http://mambo.ucsc.edu/psl.

[29] Dom Massaro and Ron Cole. “From ”Speech is Spectial” to Talking Heads in Language Learning” Proc. InSTIL2000, pp. 153–161, 2000.

[30] Jared Bernstein, Michael Cohen, Hy Murveit, Dimitry Rtischev, and Mitchel Weintraub. “Automatic Evaluation and Training in English Pronunciation” Proc. Int. Conf. Spoken Language Processing (ICSLP’90), Vol. 2, pp. 1185–1188, 1990.

[31] Hiroshi HAMADA, Satoshi MIKI, and Ryouhei NAKATSU. “Automatic Evaluation of English Pronunciation Based on Speech Recoginition Techniques” IEICE TRANS.INF.& SYST, Vol. E76-D, No. 3, pp. 352–359, MARCH 1993.

[32] Reiko Akahane-Yamada, Erick McDermott, Takahiro Adachi, Hideaki Kawahara, and John S. Pruitt. “COMPUTER-BASED SECOND LANGUAGE PRODUCTION TRAINING BY USING SECTROGRAPHIC REPRESENTATION AND HMM-BASED SPEECH RECOGNITION SCORES” Proc. Int. Conf. Spoken Language Processing (ICSLP’98), pp. 1747–1750, 1998.

[33] Reiko Akahane-Yamada, Yoh’ichi Tohkura, Ann R.Bradlow, and David B.Pisoni. “DOES TRAINING IN SPEECH PERCEPTION MODIFY SPEECH PRODUCTION ?” Proc. Int. Conf. Spoken Language Processing (ICSLP’96), pp. 606–609, 1996.

[34] 鹿野 他, “音声認識システム” 情報処理学会, 2006

[35] 深澤俊昭. “英語の発音パーフェクト学習辞典” アルク, 2000.

[36] H. Fujisaki: “From information to inotation,” Proc. 1993 International Symopsium on Spoken Dialogue, pp.7-18, 1993.

[37] H. Fujisaki and K. Hirose: “Analysis of voice fundamental frequency contours for declarative sentences of Japanese,” J. Acoust. Soc. Jpn(E), Vol.5, No.4, pp.233–242, 1984.

発表文献

[1] 三輪周作 広瀬啓吉，峯松信明，山内豊 “学習者の音声でフィードバックする英語発音教育システム” 日本音響学会秋季講演論文集，1-2-4，pp.491-492 (2007-9)

[2] 竹内京子，三輪周作，峯松信明，広瀬啓吉，“日本人学習者のフランス語鼻母音のモデル音声模倣時の生成特徴”

[3] Michiko Watanabe, Yasuharu Den, Keikichi Hirose, Shusaku Miwa, and Nobuaki Minematsu. “Features of pauses and conjunctions at syntactic and discourse boundaries in Japanese monologues. In Proceedings of the 10th European Conference on Speech Communication and Technology” Interspeech 2007, pp. 118-121. Antwerp, Belgium.

[4] Michiko Watanabe, Yasuharu Den, Keikichi Hirose, Shusaku Miwa, and Nobuaki Minematsu. “Factors affecting speakers’ choice of fillers in Japanese presentations. In Proceedings of the 9th International Conference on Spoken Language Processing” Interspeech 2006, pp. 1256-1259. Pittsburgh, PA.

[5] Michiko Watanabe, Keikichi Hirose, Yasuharu Den, Shusaku Miwa, and Nobuaki Minematsu. “Factors influencing ratios of filled pauses at clause boundaries in Japanese. In Proceedings of the ISCA tutorial and research workshop on Experimental linguistics” 2006, pp. 253-256. Athens, Greece.

付録A

基本周波数パターン生成過程モデル

A.1　基本周波数パターン生成過程モデル

音声の韻律的特徴を表現する客観的・物理的な量として，基本周波数（F_0）パターンは言語の構文や意味の伝達に重要な役割を果たしている．F_0 パターンは日本語を含む多くの言語の音声の抑揚を表し，一般に単語レベルのアクセントに対応する局所的で急激な起伏と，句・節・文レベルの，より広い範囲にわたる緩やかな起伏とから成るが，この F_0 パターンが生成される過程のモデルとそのモデルを用いる分析手法は，藤崎らによって初めて考案された [1]．このモデルでは，発話の言語学的情報と密接に関係のある少数のパラメータで，実際に観測される F_0 パターンを極めて良く近似できることが知られている [37].

A.2　F_0 パターンとその生成過程モデル

音声の F_0 パターン生成過程モデルは，対数軸上で表現した F_0 パターンが 2 種類の成分の和により表されるとしている．その 1 つは，句頭から句末に向かう緩やかな下降に対応するもので，これをフレーズ成分と呼ぶ．2 つめは，個々の単語または単語の連鎖に付属する局所的な起伏に対応するもので，これをアクセント成分と呼ぶ．実測される F_0 パターンを話者ごとのほぼ一定な基準値にこれらの 2 種類の成分が加えられたものと考えれば，単語及び文音声の F_0 パターンの特徴を統一的に把握することができる．

F_0 パターンを測定することにより，声の主観的な高低の型を，それと対応する客観的な物理量として表すことができる．

A.2.1　フレーズ成分

F_0 パターンを構成している成分の 1 つであるフレーズ成分は，およそ全ての発話に共通なもので，声帯振動の開始よりもおよそ 200〜400ms 以内から準備され始め，やや上昇しながら最大値に達した後，緩やかに下降してある一定の値に漸近し，発話の終端近くで急激に下降する成分である．

これは単独発話では 1 個であるが，文の発話では複数個存在し得る成分で，その形は質量とバネ定数とを持つ 2 次の力学系が瞬間的な外力（インパルス引力）を受けた場合の運動とよく似ている．これを数学的に表現するため，フレーズ成分を質量・バネ定数・摩擦抵抗を持った何らかの力学系のインパルス応答を用いて近似し，かつ仮想的な力学系は線形性を持ち，臨界制動系であると仮定すれば，フレーズ成分に相当する系のインパルス応答 $G_p(t)$ は次式で表される．

$$G_p(t) = \begin{cases} \alpha^2 t \exp(-\alpha t) & t \geq 0 \\ 0 & t < 0 \end{cases}$$

ここで，α はフレーズ指令に対する系の速さを定める係数であり，日本人の話者の平均的な値として，経験的に 3.0 を用いると良いことがわかっている [36].

A.2.2　アクセント成分

F_0 パターンを構成しているもう 1 つの成分であるアクセント成分は，個々の単語または連続した単語に付随するもので，主観的に高い拍の発音にやや先行して始まり，始めはゼロから緩やかに上昇し，途中はかなり急激に上昇し，その後また緩やかに一定のレベルに漸近するもので，高い拍が続けばそのレベルを保ち，高い拍から低い拍への移行に際しては，上記とは逆に低い拍の発音にやや先行して緩やかな下降を始め，途中は急激に下降し，その後またゼロとなる成分である．

これは，質量とバネ定数とを持つ 2 次の力学系が，ある時間持続する一定の外力（ステップ入力）を受けた場合の運動とよく似ている．これもフレーズ成分と同様に数学的に表現すると，アクセント成分に相当するステップ応答 $G_a(t)$ は次式で示される．

$$G_a(t) = \begin{cases} \min[1-(1+\beta t)\exp(-\beta t), \gamma] & t \geq 0 \\ 0 & t < 0 \end{cases}$$

ここで，β はアクセント成分の立上りの速さを定める係数であり，平均的な値として，20.0 が用いられる [36]．記号 $\min[x, y]$ は，x と y のうち小さい方を取ることを意味する．実際の F_0 パターンでは $G_a(t)$ が有限の時間内に上限値 γ に達し，以後その値を保持することに対応する．γ の値は，通常 0.9 が用いられる [36]．

A.2.3　基本周波数パターンの生成

これら 2 つの成分を用いて，F_0 パターンの特徴を非常に良く近似することができる．フレーズ成分を比較的時定数の長い線形 2 次系のインパルス応答，アクセント成分を比較的時定数の短い線形 2 次系のステップ応答で近似できるものとし，それらの和に非礼して対数 F_0 パターンが変形するものとしている．また，これらの表現を用いれば，F_0 パターンの特徴を良く近似できるだけでなく，発話の言語学的意図から F_0 パターンが生成される過程について，図 A.1 のようなモデルを考えることができる．

図 A.1 は，文音声の F_0 パターンを想定したもので，入力となる 2 種類の指令のうち，フレーズ成分の指令は正または負のインパルスとして，正のインパルスは文頭・文中のフレーズの先頭に，また，負のインパルスは文の終わりに生起してそれまでのフレーズ成分を下降させる役割を持っている．また，アクセント指令は正の方形波として，個々の単語または単語連鎖ごとに生起してアクセント成分を生ずる．最後にこの 2 種類の成分は相加され，声帯振動の基本周波数の対数値に変化を生ずる．ここで，時刻 t における基本周波数の値を $F_0(t)$ で表せば，その対数値は具体的には次式で表される．

$$\log_e F_0(t) = \log_e F_b + \sum_{i=1}^{I} A_{pi} G_p(t - T_{0i}) + \sum_{j=1}^{J} A_{aj}\{G_a(t - T_{1j}) - G_a(t - T_{2j})\}$$

ここで，F_b は F_0 パターンの基底値（基底周波数），I は文中のフレーズ指令の数，J は文中のアクセント指令の数，A_{pi} は i 番目のフレーズ指令の大きさ，A_{aj} は j 番目のアクセント指令の大きさ，T_{0i} は i 番目のフレーズ指令が生起する時点，T_{1j} は j 番目のアクセント指令の立上り時点，T_{2j} は j 番目のアクセント指令の立下り時点を表す．

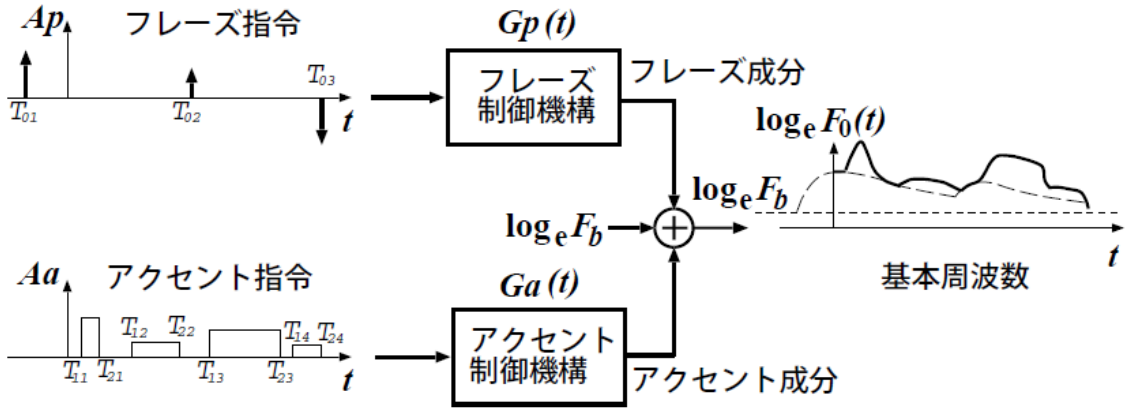

図 A.1: 基本周波数パターン生成過程のモデル [1]