ヒューマンインタフェース 97－2
音声言語情報処理 40－2
（2002. 2. 1）

ユーザー発話のセグメンテーションと発話評価機能をもつ英語学習支援システム

五十里 慎吾　佐野 輝希　緒方 淳　有木 康雄

龍谷大学 理工学部

〒 520-2194 大津市瀬田大江町横谷1-5
Tel: 077-543-7427
E-mail: ikari,teru.ogata@arikilab.elec.ryukoku.ac.jp, ariki@rins.st.ryukoku.ac.jp

あらまし 外国語によるコミュニケーションの学習においては，対象となる外国語発話の聞き取り能力，自己の発話能力，そして文作成能力を養う必要がある．それら3つの能力の習得を支援するComputer-Assisted Language Learning (以下CALL) システムの構築には，学習者の発話に対する評価と，適切な誤り部分の教示が重要な要素となる．本報告では，音声認識技術を利用したCALLシステムにおける学習機能として主に3つの機能 (セグメンテーション機能，フレージング機能，ディクテーション機能) を提案し，その原理と実装方法について述べる．実験として，実際に10人の学習者にシステムを利用してもらい，アンケートの結果から評価を行った．

キーワード : CALL, forced alignment, フレージング, ディクテーション

A CALL System with Segmentation and Evaluation Function of an User Utterance

Shingo Ikari　Teruki Sano　Jun Ogata　Yasuo Ariki

Faculty of Science and Technology, Ryukoku University

1-5 Yokotani, Oe-cho, Seta, Otsu-shi, 520-2194 Japan
Tel: +81-77-543-7427
E-mail: ikari, teru, ogata@arikilab.elec.ryukoku.ac.jp, ariki@rins.st.ryukoku.ac.jp

Abstract In communication learning of second language, three abilities have to be improved; listening, speaking and writing ability. In this sence, it is important to evaluate user's pronunciation ability and to detect mispronunciations in CALL (Computer-Assisted Language Learning) systems. In this paper, we propose three functions (segmentation, phrasing and dictation) in CALL system using speech recognition technology. As experiments, the system was evaluated from the result of a questionnaire to ten learners.

Key words : CALL, forced alignment, phrasing, dictation

1 はじめに

近年，音声情報処理技術が目覚ましく発展するなかで，計算機を利用した外国語学習支援システム（以下CALLシステム）が盛んに研究され，その注目を集め出している．こうした背景には，国際化により，英語を話すことが必要とされる場面が一般的に増え，国民のほとんどが国際語としての英語を学ぶことを希望するようになったことがあげられる．しかし，日本人の多くは，「読み」・「書き」はできても，「聞く」・「話す」ことは弱く，英語は必ずしも得意でない．

「聞く」・「話す」，つまり会話ができない要因は，根本的に日本人と英語母国語話者(以下，ネイティブ)との間で，発声の韻律操作が大きく異なるからである[1]．すなわち，英語にあって日本語にない発音や，また，日本語には無い子音の連続など，様々な言語的な違いが存在する．また，英語習得を困難にしているもう一つの要因として，定型文や慣用句を学習しても，日常的に英語を使用していないため，実際の場で使用できないことがあげられる．

これらの問題をふまえて，従来の「読む」・「聞く」と言った受信型教育に加え，「話す」・「書く」といった発信型教育を支援する形のCALLシステムが研究されている．[2][3]．また，強勢，リズム，イントネーションなどの韻律的特徴の発声に多くの誤りが含まれることが，ネイティブに指摘されており，この点から英語文強勢の教示システム構築が研究されている[4]．

このように，様々なCALLへのアプローチがなされているなかで，本研究では，自然な英会話の習得を研究目的として，映画を教材としたCALLシステムの構築を行う．映画を教材とすることによって，生きた英語，すなわち，リアリティを持って実際に使われている英語を学習することができ，又，学習者の興味を引き付け飽きさせないといった狙いもある．本報告では，音声認識技術を利用したCALLシステムにおける学習機能として次の3点を提案する．

- セグメンテーション機能:
  教師発声，ユーザー発声に対する単語毎の再生機能
- フレージング機能:
  教師発声，ユーザー発声に対する句毎の分節と再生機能
- ディクテーション機能:
  教師発声を聞き取り単語毎に書き取る機能

上記機能を実装し，システムの評価実験として，実際に10人の学習者にシステムを利用してもらい，アンケートの結果から評価を行った．

2 映画を教材としたCALLシステムの概要

2.1 映画を教材としたCALLシステム

CALLシステムでは，学習者に飽きさせないために，生きた英語，すなわち，リアリティを持って実際に使われている英語を教材として用いることが望ましい．生きた英語を用いることで，実際のコンテキスト（文脈）と感覚を体験することが出来るからである．このような生きた英語を豊富に含んでいる教材として映画がある．この点から，本研究では，映画を英語教育の教材としたCALLシステムの開発を行っている．

学習者が聞取能力や発声能力を高めていく一つの方法として，母国語話者の発声に続き，発話を真似て発声する方法が考えられる．しかし，学習者は母国語話者に続いて発声するため，一文単位の反復学習となり，自分自身で文を作成する能力を養うことができない．この問題を解決するためには，一文をフレーズという単位に分割し，フレーズの組み合わせで一文を組み立てていく方法が有効である．そこで，本研究では，映画のクローズドキャプションに含まれているフレーズに関して頻度分布を求め，そのフレーズリストに基づいて英語学習を行うシステムを開発している．まず日常会話が比較的多いと推測できる恋愛物，サスペンス，ホームドラマなど合計100本の映画を選択し，学習に用いるフレーズリストを選出した[5]．CALLシステムを構築する上では，初心者には上位100程度までのフレーズを使ったコース設定を，また，上級者には上位1000フレーズ程度のコースといったように，頻度分布を利用してコース設定することも可能である．

2.2 ユーザー発話のセグメンテーションと発話評価機能をもつCALLシステム

前節で述べた映画の教材やフレーズリストを用いて，本研究では，音声認識技術に基づくユーザー発話のセグメンテーションと発話評価機能を兼ね備えたCALLシステムについて検討する．音声処理として主に用いた技術は，映画のクローズトキャプションを基に，音声を単語単位に自動的に分割していくforced alignmentであり，そのセグメンテーション結果に基づいて以下のような機能を実現する．

- 単語単位の再生機能
- フレージング機能
- ディクテーション機能

実際に構築したシステムのユーザインターフェース画面を図1に示す．

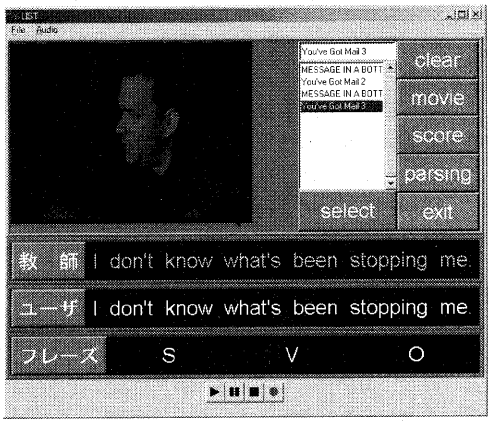

図 1: システムのユーザーインターフェース画面
教師ボタンをクリックすれば俳優の発声音声を，ユーザボタンをクリックすれば学習者の発話音声を聞くことができる．また，表示されている単語1つ1つがボタンになっており，そのボタンをクリックすると，発話を単語単位で聞くことができる．

3 発話音声のセグメンテーション

今回構築したCALLシステムでは，映画俳優の音声と学習者の音声を，単語単位で聞き比べる機能がある．その際，学習者の音声を単語単位でセグメンテーションする事が必要となる．以下にその手法を述べる．

3.1 セグメンテーションの手法

学習者が日本人であることから，学習者の英語発話には日本語らしさが入ることが考えられる．学習者の発音が未熟であれば，ネイティブの英語音素HMMを利用したセグメンテーションでは，発話音声を正確にセグメンテーションすることができない．そこで，学習者の発音が未熟なために発生するセグメンテーション誤りを防ぐために，日本語·英語，共に別々に学習したHMMを混在させ，英語46音素，日本語41音素，合計87音素で構成された音素HMMの集合により，forced alignmentを行い，学習者発話の単語区間の検出を行った．

また，単語レベルでforced alignmentを行うには単語辞書を作成する必要がある．今回，日本語·英語混在HMMを用いて forced alignment を行っているので，英語の音素表記に加え，日本語の音素表記，日本語·英語混在の音素表記を追加しなければならない．

表1に示すように，英語音素のみの音素表記，日本語音素のみの音素表記，日本語と英語音素混在の表記を，辞書内で許している．表中sh-jやi:-jは日本語音素であることを表わしている．

表 1: 音素表記の例(音素 -j は日本語音素)

	SHE	HAD
英語音素表記	sh iy	hh ae d
混在音素表記	sh i:-j	hh ae d-j
日本語音素表記	sh-j i:-j	h-j a-j d-j o-j

3.2 実験条件

音素認識には，音素環境独立なHMM（monophone HMM）を用いている．英語音響モデルの学習には，まずTIMIT連続音声データのすべての男性データを用いて初期モデルを作成し，次にWSJのうち，男性話者147名分，合計20614発話を用いて連結学習を行なった．日本語音響モデルの学習には，まずATR連続音声データベースa～jセットの6名分のデータの視察ラベルを用いて初期モデルを作成し，次に日本音響学会新聞記事読み上げコーパス(JNAS)のうち，男性話者137名分の21782発話を用いて連結学習を行なった．音響分析条件は日本語·英語HMMともに同じ条件で，39次元の特徴パラメータ(12次元のMFCCとパワー，およびそれぞれのΔ，ΔΔ係数)を用いた．

3.3 実験結果

学習者の発話音声に対して，音響モデルの種類ごとにセグメンテーションした実験結果を表2に示す．実験は，日本人4名に一人あたり3文，112単語を発声してもらい，英語音響モデル，日本語·英語混在音響モデルをそれぞれ用いてセグメンテーションし，単語毎に始端のずれ幅（単位ms)の平均値を求めた．ここでは，各単語の始端のずれが，3フレーム(30ms)以内であれば正解として正解率を求めた．

実験の結果，英語のみの音響モデルを用いたときよりも，日本語·英語混在音響モデルを用いた方がセグメンテーションの正解率が約10%高かった．以下では，システムの機能を実現するうえで，発話のセグメンテーションは日本語·英語混在音響モデルを用いて行った．

表 2: 音響モデルの違いによるセグメンテーションの結果

	英語音響モデル	混在音響モデル
ずれ幅の平均値	20.8 ms	12.5 ms
正解率	83.0 %	92.9 %

4 発話評価機能

ここでは，学習者が発声した文を，音声認識技術を用いて発音の良否を自動評価する機能について述べる．まず，実際に構築したシステムにおいて，学習者の発音を評価したスコアの表示と，誤りを教示するユーザーインターフェース画面を図2に示す．図の上部は1文全体のスコアと誤りの数，正しい音素列と学習者発話の認識された音素列を表しており，図の下部のうち左側が単語単位のスコア，右側が誤りの教示例を表している．学習者はこれらのスコアから自分の発話の中で，どの単語の発音が悪いかを知ることができる．

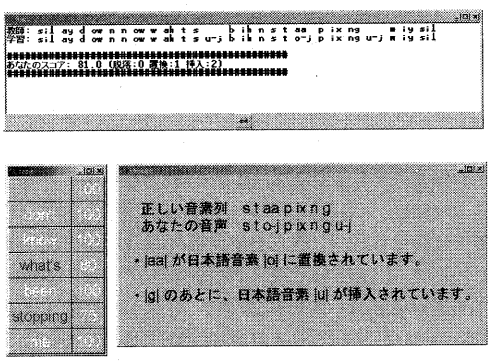

図 2: 発話のスコアと誤り教示のユーザーインターフェース画面

4.1 発音評価

発音評価の自動推定には，日本人の誤り音素パターンを考慮したforced alignmentを行う．具体的には，発音辞書中の単語の音素表記パターンを，通常の正しい英語音素表記，日本人の誤り音素パターンを考慮した音素表記というように複数登録する．実際のforced alignmentの際には，1つの単語に対して音素表記パターンの数だけ複数の候補が生成され，そのときの入力音声に対して尤も音響的にマッチした音素表記パターンが選択されることになる．これは河合ら[6]により提案された，日本人の音素誤りパターンを考慮した発音ラティスによる音声認識手法とほぼ同等であるといえる．ここで，日本人の誤りパターンとしては主に以下のものを用いた．

- 母音挿入誤り:
  英語の音節では，母音をはさまずに子音だけが連続して出現することがあるが，日本語では子音結合が存在しないため，日本人が英語を発話する際，子音結合でも母音を入れて発音しがちである．例えば，upの[p]のあとに[u]が挿入されるなどの誤りがある．
- 置換誤り:
  日本語では使用しない音を，日本語で使用する音で置き換えてしまう誤りが置換誤りである．例えば，英語の[r]のかわりに日本語の[r]を用いたり，[th]を[s]で代替し，thank youがsank youになったりする誤りがある．
- 脱落誤り:
  日本人が英語を発話する際，発音しなければいけない音をとばしてしまう誤りが脱落誤りである．例えば，音節末の[r]が落ちfarがfaになってしまうといった誤りがある．

4.2 誤りの教示

学習者の発声に対して，システムは学習者が発音したと判断した音素の並びを提示する(図2上)．学習者は発音の認識結果からスコアの低い単語のボタンをクリックすれば，その単語がどのように発音を誤ったかをテキスト形式で教示する(図2右下)．図2右下は，stoppingの教示例を示している．ここで，単語毎，文単位のスコアとしては，前節で述べた日本人の誤り音素パターンを考慮したforced alignmentの認識結果の音素列に対する，音素正解精度(置換誤り，挿入誤り，脱落誤りを含んだ正解率)を用いている．

5 フレージング機能

英語を学習する上で，英語のどのような形(構文)がどのような意味を表すかを正確にとらえていないと，英語を使う(聞く，話す，読む，書く)ことができない．そこで本研究では，文を意味的なまとまりでとらえて理解することができるように，俳優，ユーザーの発声をフレーズ単位で聞くことができるフレージング機能について検討する．このフレージング機能により，教師音声を意味的なまとまり毎に聞くことができ，より効果的なリスニングが可能であると考えられる．また，ユーザーの発声を意味的なまとまりとして再生することにより，ユーザー自身の発声を用いた正しい発声例を教示することができる．ここで，フレーズとは，意味的にまとまりのある単語の集合を表しており，本研究ではフレージング機能におけるフレーズの単位として，構文(S:主語，V:述語，O:目的語，C:補語，M:修飾語)を用いることにした．

図1のインターフェースには，S，V，Oといったフレーズが表されており，ユーザーはこのフレーズ1つ1つをクリックすることにより，フレーズ単位で音声を聞くことができる．したがって，聞き取りでは，まず「教師」

のボタンをクリックして1文全体を聞き，それで聞き取れない場合には，フレーズの「S」,「V」などのボタンをクリックすることでフレーズ単位で聞くことができる．それでも聞き取ることができない場合には，更に単語単位のボタンをクリックして単語毎に音声を聞くことができるようになっている．

6 ディクテーション機能

6.1 ディクテーション機能の概要

ここでのディクテーションとは，英文を聞き，それを正確に書き取ることである．英語学習法の1つとして，このディクテーションにより英語力を高める学習法がある．日本人は目から英語を覚えていることが多いので，読めばわかるのに聞くとわからない，ということが多い．そこで，ディクテーションを行うと次のような利点がある．

- リダクションやリエゾンの法則（音が消えたり，つながったりする法則）が体得できる．
- はっきり聞こえない部分は文法で補わないといけないので文法力の向上になる．
- 自分の苦手な音，聞き取れない音がわかり，リスニング力の向上になる．

6.2 実現方法

通常，英語学習におけるディクテーションは，教師音声としてネイティブ話者が発声した1文を注意深く聞き，それを正確に書き取っていく．教師音声1文の単語数などの情報は与えられず，また，連続音声特有のリダクションやリエゾンの現象もあるため，英語の習熟度の低い学習者にとっては非常に困難な学習方法といえる．

本研究では，上記のディクテーション学習を音声認識技術を用いることで拡張し，より効果的なリスニング能力向上を目指した機能について検討する．

まず，教師音声となる映画俳優の音声1文を，3.1節で述べた手法で単語ごとにセグメンテーションする．これにより，学習者は1文全体だけではなく，単語単位に教師音声を聞くことが可能である．また，一度聞いて聞き取れなかった単語のみを再生することもできる．

6.3 学習手順

1. 問題(1文単位)をリストから選択する．
2. 選択すると図1のような画面が出るので，startボタンを押してディクテーションする1文を聞く．この際には，正解のテキストは隠れており，単語数に応じた数のボタンのみが表示される．
3. 2を何度か繰り返しても聞き取ることができない場合はフレーズボタンを押してフレーズごとに区切られている1文を聞く．
4. 3を何度か繰り返しても聞き取ることができない場合は単語単位で聞く．ボタン押すことにより，聞き取れなかった単語を何度も聞く．
5. 白いところが答えを書き込む欄になっているので，そこに聞き取った英語を単語単位で書き込む．
6. 書き込んだ後にEnterキーを押すと，単語単位での正解(○×表示)が表示される．
7. 間違っていたところは，2，3，4，5，6を繰り返すことにより，正解を目指す．
8. どうしても聞き取れないときは，textボタンを押すと正解が表示される．

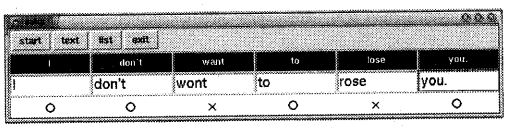

図 3: ディクテーション機能のユーザーインターフェース画面

7 システムの評価

システムの評価実験では，実際に被験者にシステムを利用してもらいアンケートの結果から評価を行うことにした．今回の被験者の人数は10人で，全員理系情報学科系の大学生である．被験者には実験の前に，システムの簡単な説明をした．システムを使用した後にアンケートに答えてもらった．アンケートの項目を以下に示す．

1. 英語は好きか
2. 年齢は
3. システムの操作は簡単にできたか
4. フレージング機能は役に立つか
5. ディクテーション機能は役に立つか
6. このシステムを使って楽しく学習できたか
7. 引き続きこのシステムを使いたいか
8. このシステムは実際の英語学習に役立つか

9. システムの良かった点は

10. システムの悪かった点は

11. このほかにどのような機能があれば良いか

12. システム全体は良いか

次にアンケートの結果を以下に示す．なお項目3～7，12では5段階(最大5点，最低1点)で評価してもらい，その平均値を算出した(表3)．

表 3: 評価スコアの平均値

項目	スコア
システムの操作は簡単にできたか	4.0
フレージング機能は役に立つか	3.0
ディクテーション機能は役に立つか	3.7
このシステムを使って楽しく学習できたか	3.9
引き続きこのシステムを使いたいか	4.2
このシステムは実際の英語学習に役立つか	4.2
システム全体は良いか	3.7

学習者が実際にシステムを使用して，特に良かったと述べた意見としては以下のようなものがあった．

- 映画の1シーンが学習に使用されている点．
- 学習者の発話と俳優の発話が比較できるので，間違いがわかりやすい点．
- 音声認識により発話が単語ごとに区切られる点．
- 発音評価に要する処理時間がそれほど長くない．

全体的に，特に良かったと回答した意見としては，リアルタイムで自分の発声を正しく単語ごとにセグメンテーションしてくれる点が多かった．また，ディクテーション機能は実用的であり，リスニング能力を向上させるのに非常に役立つという回答も多かった．

次に，学習者によって指摘された改良すべき点を以下に示す．

- 映画のシーンの数をもっと増やす．
- 俳優の台詞とクローズドキャプションが異なる場合がある．
- 俳優の発音が必ずしも正しいとは限らない．
- ホテルチェックインや入国審査など，海外旅行で直接役立つシーンが欲しい．

さらに学習者は，今後システムに以下のような機能があれば良いと答えている．

- 学習するフレーズの日本語訳を提示する機能．
- システムと簡単に対話できる機能．

8 おわりに

本報告では，音声認識技術を利用したCALLシステムにおける学習機能として，主に3つの機能(セグメンテーション機能，フレージング機能，ディクテーション機能)を提案し，その原理と実装方法について述べた．まず，3つの機能の基盤となる単語単位のforced alignmentの実験を行い，日本語・英語混在音響モデルを用いることにより，英語音響モデルのみを用いる場合と比べて，大幅にセグメンテーションの精度が向上することを確認した．

また，上記セグメンテーションに基づいた，発音誤り教示機能，フレージング機能，ディクテーション機能などを持つCALLシステムを実装した．実験として，実際に10人の学習者にシステムを利用してもらい，アンケートの結果から評価を行った．

今後の課題としては，映画のシーンを増やすことや，日本語訳の提示などのシステムの機能拡張などや，発音誤り教示の厳密な評価などがあげられる．

参考文献

[1] 柳ヶ瀬 裕則，山下 洋一: “言語教育支援のための発話比較方法の検討”，日本音響学会講演論文集，pp.83-84，2000.

[2] 早越 弘子，壇辻正剛，中村順一，河原達也: “京大総合情報メディアセンターにおけるCALLの試み”，音声言語情報処理，19-18，pp.115-122，1997.

[3] 中川聖一，Allan A. Reyes，鈴木英之，谷口泰広: “音声認識技術を利用した英会話CAIシステム”，情処学論，Vol.38，No.8.，1997.

[4] 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モデル化”，信学技報，SP2000-1，pp.1-8，2000.

[5] 緒方淳，阪口福太郎，五十里慎吾，佐野輝希，有木康雄: “映画を教材とした英語学習支援システム”，電子情報通信学会ソサエティ大会，SD-2-4，pp.303-304，2001-09.

[6] 河合剛，石田朗，広瀬啓吉: “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，57巻9号，pp.569-580，2001.