解説論文

音声情報処理技術を用いた外国語学習支援

河原　達也[†]　　峯松　信明[††]

Computer-Assisted Language Learning (CALL) Based on Speech Technologies

Tatsuya KAWAHARA[†] and Nobuaki MINEMATSU[††]

あらまし　音声認識・合成に関する技術はこの十年ほどの間に大きな進歩を遂げており，言語学習支援 (CALL) システムに関しても，発音の自動評定や模擬会話訓練などの新たな可能性を広げている．音声分析・認識・合成技術を利用した CALL には，音韻的な観点と韻律的な観点がある．非母語話者の音声を正確に区分化・認識しながら，そこに含まれる誤りを検出するためのモデル化には多くの課題がある．本論文では，これらの課題と現在の方法論に関して概観を行う．まず，発音学習支援における音韻的な発音評定と誤り検出に関して，音声認識技術に基づく方法について述べる．統計的なアプローチの定式化を行い，非母語話者の音響・発音レベルのモデル化について述べる．次に，継続長や強勢・トーンなどの韻律的なモデル化と評価について述べる．更に，テキスト音声合成，分析合成，モーフィングなどの音声合成技術の利用についても述べる．最後に，実用化されている幾つかの外国語 CALL システム，及び非母語話者の音声データベースについて紹介する．

キーワード　音声情報処理，CALL，音声分析，音声認識，音声合成

1. まえがき

今世紀に入ってから我が国も国際化が加速し，事実上の国際標準語である英語の能力を身につけることが必要不可欠な状況になってきた．多くの大学で英語のみで卒業・修了できるコースが用意されたり，幾つかの企業で英語を標準語とする動きもある．この流れの中で，英語教育が小学校から実施されるようになった．英語以外の外国語学習の機会も多くなる一方，来日する外国人の増加に伴って，日本語を外国語として学習する需要も大きくなっている．

筆者らが学生だった数十年前は，実践的な語学学習 (LL) といえば，アナログのカセットテープを用いて行うのが一般的であった．1990 年代半ばになって，コンピュータを用いた語学学習 (CALL：Computer-Assisted Language Learning) が導入されるようになった [1]．昔の LL が音声のみのメディアでシーケンシャルアクセスしかできなかったのに対して，CALL では CD-ROM 等に，テキストや画像などが一式になったマルチメディア教材を用意でき，ランダムアクセスできるようになった点が第一の大きな変化であった．

第二の大きな変化として，発音訓練において，学習者の音声を単に録音・再生するだけでなく，様々な処理ができるようになったことが挙げられる．当初は，フォルマント分析や基本周波数分析などの基本的な音声信号処理を適用して，模範的な母語話者のパターンと比較・提示するレベルであった．ただし，このような単純な分析レベルの提示では，音声に関する知識あるいは語学教師の介在がないと，発音の何が問題でどう修正すればよいか分からないという問題があった．

ほぼ時期を同じくして，音声認識や音声合成の技術が大きな進歩を遂げた．これは主に，統計的なモデル化の洗練とデータベースの大規模化によるものである．それに応じて，これらの技術，特に音声認識技術を CALL システムに適用する試みが自然な流れとして行われるようになり，2000 年頃に研究分野として確立された [2], [3]．

これには，学習者の発音を評定することに加えて，学習を支援するという目的がある．発音評定に限定すると，米国では英語を母語としない人を対象とした PhonePass（現 Versant）[4] というシステムが早くから実用化されており，最近では TOEFL を主催してい

[†] 京都大学，京都市
Kyoto University, Kyoto-shi, 606–8501 Japan
[††] 東京大学，東京都
The University of Tokyo, Tokyo, 113–8656 Japan

る ETS が自動評定の開発・フィールド評価を進めている [5], [6]．また中国では，国内の方言話者を対象に普通語 (Putonghua) のレベルを評価するテスト PSC での導入も行われている [7]．

学習支援に関しては，大学等の CALL 教室で語学教師の介在を想定したものと，自学自習のものがあるが，徐々に後者が主流になっている．その理由として，学習者の都合のよい時間・場所で，他人を気にすることなく自己のペースで訓練に取り組めることが挙げられる．しかし，正しいフィードバックが行われないために誤った発音が固定化することを避ける方策に加えて，学習意欲を引き出すための “Edutainment” 的な工夫も必要となる．

CALL システムの対象は，非母語（方言を含む）話者に限らず，子供 [8] や聴覚・発声に障害のある方 [9] なども考えられ，実際に様々な取組みがなされているが，本論文では，非母語話者が外国語を学習する場合に焦点をおく．本論文の前半では，外国語学習のための音声情報処理技術について様々な側面から述べる．その後，実用化されている幾つかの CALL システム，及び非母語話者の音声データベースについて紹介する．

2. 外国語学習と音声情報処理

外国語学習には，リーディング，ライティング，リスニング，及びスピーキングの側面があるが，最初の二つについては基本的に音声メディアを扱わない．リスニングについては，例えば日本人が聞き取りにくい/l/と/r/の識別に着眼した研究開発 [10] 等があるが，本論文では，学習者の発声した音声を処理するスピーキング・発音の学習を主な対象として扱う．

これに関しても，自分で内容・文章を構成して話す場合（通常のスピーキング）と，与えられた文章を発声する場合（発音訓練）がある（表 1）．前者の場合，発音だけでなく，語彙・文法・語用論的な知識・運用能力も必要とするので，学習者にとっても，それを扱うシステムの開発も高度になるのは明らかである．したがって現状では，スピーキングの評定や，特定のシーン（例えばショッピングや商談）における会話に限定して，システム開発が行われている．

表 1　外国語を話す能力に必要な要素
Table 1　Factors in proficiency of foreign languages.

- スピーキング（語彙の選択や文の構成を含む）
  - 語彙
  - 文法
  - 語用論・社会的知識など
- （与えられた文の）発音
  - 音韻的要素
  - 韻律的要素... 継続長・強勢・トーンなど

ただし，スピーキングによるコミュニケーション全体において，発音能力が語彙力と並んで決定的に重要であることが指摘されている [11]．したがって，与えられた文章を発声する発音訓練がまず重視される．日本語と英語のように，母語と発音体系が大きく異なる外国語を学習する際には，特に重要である．母語話者レベルの発音が必ずしも要求されるわけではないが，円滑なコミュニケーションが成り立つレベルの発音は必要である．

発音が正しいかを確認したり，正しい発音を教示したりするためには，厳密には調音器官（舌や顎など）の動きを見る必要がある．しかし，口内の動きを捉えるのは簡易にできない．したがって，容易に収録・視覚化可能な音声信号を用いて処理を行う．音声と調音の関係に基づいて，調音器官の動きをアニメーションで視覚化する研究開発も行われている [12]．

音声には音韻的（分節的）な側面と韻律的（超分節的）な側面があり，両者が正しく構成されることにより，正しい発音が実現される．音声合成システムにおいては，両者のモデル化に関して長年の研究開発の蓄積がある．一方，音声認識システムにおいては，中国語の四声などを除くと，韻律的側面はほとんど扱われておらず，音韻的なモデル化に注力されている．これは，言語情報の伝達には音韻的側面だけでも十分であるが，自然なコミュニケーション（人間が聞き取る）には韻律的側面も重要であることを示唆している．あるいは，標準的な韻律パターンを合成することは可能であるが，不特定多数話者の韻律パターンの変動のモデル化は，音韻パターンに比べて困難であることを示唆している．

発音訓練においては，特定の音韻に着目した単語発声（例えば “right” と “light”）から，文やパラグラフ単位の発声までである．発声単位が長くなるほど，様々な韻律的な要因（イントネーションやプロミネンス）が加わることになる．また韻律は，ピッチ・パワー・継続長などの幾つかの特徴によって構成される．

音韻的な側面に関しては，音声認識研究において多くのモデル化が行われてきたので，その知見・モデルを利用するのが自然であると考えられる．ただし，音声認識が正しく発音された未知の文を特定するのに対

して，外国語学習支援では，既知の文が正しく発音されたか判定する点で定式化が大きく異なる．これについては **3.** で詳しく述べる．一方，韻律的な側面に関しては，音声分析・合成研究の知見が活用できる．ただしこの場合も，非母語話者を考慮したモデル化が必要となる．これについては **4.** で述べる．また，音声合成技術を活用して音韻・韻律両面から学習支援を行うことも考えられ，これについては **5.** で述べる．音声分析・合成に関しては，前述のように，単純な信号処理では模範的な母語話者と比較ができないので，比較が容易になるように特徴量を正規化したり（**3.1** 及び **3.8**），学習者の声質で理想的な母語音声を合成する (**5.**) などの処理を行う．

3. 音声認識技術を利用した音韻的な発音の評価

本章では，音韻的な側面に焦点をおいて発音の評価を行う方法について述べる．これには主に音声認識で用いられている定式化・モデル化を利用する．その典型的な処理の流れを図 1 に示す．図の括弧内の数字はおおむね以下の節番号に対応するが，図 1 のような構成をとらないシステムも考えられるため，以下の節の説明が必ずしも図 1 の各モジュールのものとは限らない．

3.1 音韻的側面に関する音声分析

フォルマントはスペクトル包絡においてピークとなる周波数で，特に母音に関しては開口度及び調音位置との対応がとれることが知られている．例えば，英語の “bat” と “but” の区別を判定・教示することができる．ただし，声道長に起因して個人差がある．例えば，男性と女性，日本人とアメリカ人ではかなりの違いがある．そこで，声道長正規化 (VTLN：Vocal Tract Length Normalization) を行った上で提示することが検討されている [13]．ただし，連続音声では前後の音素の影響を受けてフォルマント周波数も過渡的になり，専門家が目視で判定することはできても，頑健に高い精度で自動検出するのは容易でない．

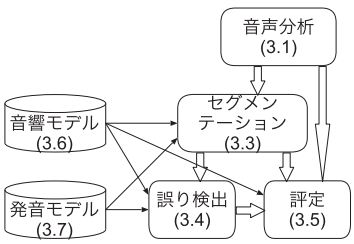

図 1 音韻的な発音評価の処理の流れ（括弧内の数字は説明している節を表す）

Fig. 1 Flowchart of pronunciation evaluation in the segmental viewpoint (Numbers in brackets correspond to Sections.)

調音位置のほかに，有声/無声，破裂性，鼻音性などの音素を記述する弁別素性を判別することで，単に発音を評価するだけでなく，調音の様子を学習者に視覚的にフィードバックできれば有用と考えられる [14]．例えば，英語の “thee” と “sea” と “she” の区別を判定・教示することができる．

しかしその一方で，現在の音声認識システムの音声分析においては，フォルマントや弁別素性などはほとんど用いられておらず [15]，スペクトル包絡の特徴を表現するメル周波数ケプストラム係数 (MFCC：Mel-frequency Cepstrum Coefficient) などが一般的な音響特徴量として用いられている [16]．ケプストラム平均正規化 (CMN：Cepstrum Mean Normalization) などは，話者正規化やチャネル正規化において効果的である．ただし，日本人の音声に対して，アメリカ人の音声データベースで学習した音響モデルを適用するような状況では，更なる正規化が必要であると考えられる．現在最も一般的な VTLN の方法は，音響モデルのゆう度が高くなるように周波数軸の区分線形変換係数を求めるものである [17]．

これに対して峯松らは，音韻カテゴリー間の f-divergence に基づく不変特徴量を抽出する方法を提案している（**3.8** 参照）．

3.2 音声認識と外国語発音学習支援

まず，通常の音声認識と外国語の発音学習支援の違いについて定式化を行う．音声認識は，未知の入力音声（正確にはその音響特徴量）X に対して，発話内容（音素列または単語列）W を推定する問題であり，事後確率 $p(W|X)$ を最大化する W を見つける問題として定式化される．これは，ベイズ則によって以下のように置き換えられ，

$$\arg\max_{W} p(W|X) = \arg\max_{W} p(W) * p(X|W) \quad (1)$$

W に含まれる個々の音素 w に対する音響モデルゆう度 $p(X|w)$ を乗算（対数スケールで加算）していき，言語モデルゆう度 $p(W)$ と組み合わせることで計算される．したがって，個々の音素 w の音響モデルはそれに対応する学習音声データを用いて，$p(X|w)$ を最大化するように学習（最ゆう推定）される．この前提・

定式化は，非母語話者の音声を自動認識するという設定でも基本的に同じである．

これに対して外国語の発音学習支援においては，発話内容 W は学習者に与えた（既知の）上で，必ずしも正しく発音されたとは限らない音声 X がシステムに入力されるという設定である．発音学習支援では，図 1 に示すように下記の 3 要素から構成される．

- セグメンテーション

既知の音素列 W に基づいて音声 X を強制アラインメントする．これはビタビアルゴリズムによって実現される．

- 誤り検出

$p(X|W') > p(X|W)$ となるような別の音素列 W' を見つける問題として定式化される．

- 評定

例えば，$p(X|W)$ を計算することにより実現できるようにも考えられるが，そのための音響モデルをどのように構築するかが大きな問題である．

以下の各節において，これらについて詳しく述べる．

3.3 セグメンテーション

入力音声の音素単位へのセグメンテーションは，誤り検出や評定を行うための重要な前処理である．一般に，入力音声 X の発話内容の音素列 W が既知の場合，W を表現する HMM を作成し，X に対してビタビアルゴリズムを適用することで，セグメンテーションは容易に実現される．

ただし，外国語の発音学習においては，音声 X が音素列 W の正しい発声になっているとは限らず，誤りが含まれる場合がある．特に，挿入誤りや脱落誤りが含まれていると，セグメンテーションに大きな影響を及ぼす．例えば，日本人が英語を発音する際に，連続する子音の間に母音が挿入される傾向がある．したがって，このような典型的な誤りを予測して検出する機構が必要になる．これは，後述する誤り検出や発音モデルとも密接に関連するので，各々の節で述べる．

3.4 誤り検出

音素列 W に対する発声 X に含まれる誤りの検出は，$p(X|W') > p(X|W)$ となるような別の音素列 W' を見つける問題として定式化される．これは単純には，セグメンテーションされた各音素 w の区間に対して，別の音素 w' に対するゆう度 $p(X|w')$ を計算することにより実現できる．しかし実際には，前述のように挿入誤りや削除誤りも考慮する必要があるので，それらの可能性を全て表現するネットワーク（図 2 参照）を用意し，そのネットワークを一種の制約（言語モデル）とみなして，音声認識を実行することになる．この誤り予測は，学習者の母語に関する先見的知識を考慮して発音モデルで記述する．

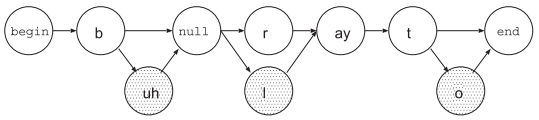

図 2 誤り予測を含む音素ネットワーク（英単語 “bright”）
Fig. 2 A phone network with error prediction for word “bright”.

上記のように可能な誤りパターンを用意して，ゆう度 $p(X|W')$ を計算するのは生成モデルに基づくアプローチといえる．これに対して，入力の音声区間が w よりも w' らしいかを直接的に識別するアプローチも考えられる．これは，誤り検出に特化して，通常の音声認識で扱う音響特徴量以外の様々な特徴・素性を導入できる利点がある．識別器としては，単純な線形判別のほかに，サポートベクトルマシンやロジスティック回帰モデルなどが考えられる．

誤り検出において留意すべき点として，数多くの誤りの中から重要な誤りを選別し，誤り検出自体の誤りの影響を抑えることが挙げられる．特に，正しく発音しているにもかかわらず誤りと判定されると，学習者が混乱したり意欲を失ったりすることにもつながるので，誤検出よりも検出漏れを許容するようにしきい値を設定するのが望ましい．更に，発音誤りに対してどのようにすれば修正できるか，適切なフィードバックを与えることも重要である．そのために，調音に則した教示が望ましい．

3.5 評　　定

外国語の発音を評定する際に二つの考え方がある．一つは「模範的な母語話者の発音にどのくらい近いか」という考え方で，この場合 “模範的な母語話者” のモデル λ_G を用意して，ゆう度 $p(X|W;\lambda_G)$ を計算すればよい．しかし，“模範的な母語話者” とは何かという教育哲学的な問題に加えて，音声分析 (**3.1**) で述べたように，話者正規化の問題もある．そもそも音声認識のゆう度 $p(X|W)$ は，話者変動や雑音等の影響も受けるので，絶対値を信頼度等に使うことは適切でない．

そこで，日本人が英語を学習する場合には，英語母語話者モデルによるゆう度と日本人話者モデルによるゆう度の比を求めることが検討され，効果が報告されている [18], [19]．すなわち，非母語話者のモデル λ_N

も用意して，下記のように，ゆう度の比（対数スケールではゆう度の差）の平均を計算する．

$$\frac{p(X|W;\lambda_G)}{p(X|W;\lambda_N)} = \prod_i \frac{p(X|w_i;\lambda_G)}{p(X|w_i;\lambda_N)} \quad (2)$$

$$= \prod_i \prod_t \frac{p(x_t|w_i;\lambda_G)}{p(x_t|w_i;\lambda_N)} \quad (3)$$

ここで，w_i は W の各音素であり，x_t はそれに対応する区間の各時間フレームである．つまりこの処理は，セグメンテーションの結果を前提としている．また上式で，$\prod$ と表記している部分は，実際には i や t に関して乗算平均（対数スケールでは加算平均）をとる．

発音評定のもう一つの考え方は，必ずしも “ネイティブらしさ” にこだわらず，コミュニケーション上の了解性を重視するという観点から，「他の音素とどの程度明確に区別できるか/まぎらわしくないか」というものである．これは，事後確率 $p(W|X)$ を計算することに相当し，前節の誤り検出が 2 値的な判定を行うのに対して，数値的な評価を行うことに対応する．これは，音声認識の信頼度計算と類似の考え方である．

これは通常下記の式で計算され，GOP (Goodness Of Pronunciation) スコア [20] と呼ばれる．

$$\frac{p(X|W)}{\sum_{W'} p(X|W')} = \prod_i \frac{p(X|w_i)}{\sum_{w'_i} p(X|w'_i)} \quad (4)$$

$$\simeq \prod_i \prod_t \frac{p(x_t|w_i)}{\max_{w'_t} p(x_t|w'_t)} \quad (5)$$

ここで w_i と x_t は前記と同じで，セグメンテーションの結果得られる．これに対して w'_t は，時刻フレーム t における最ゆうの音素であり，全ての音素（若しくは音節）の連鎖を許す制約（言語モデル）を用いた音声認識の結果，ビタビアルゴリズムにより得られる．上式は結果的に，これらの二つのゆう度の比を求めることを行っている．なお，上式で $\prod$ と表記している部分は，実際には t に関して乗算平均をとる．この式から，学習者の発音が正しい ($W = W'$) 場合には，GOP スコアが 1 に近くなり，逆にある音素 w_i の区間において GOP スコアが小さい場合は，そこに誤りがあることを示している．上式の w'_i に関して，全ての音素を考慮するのではなく，w_i と混同しやすい音素に絞り，更に音素ごとに重みを付ける方が効果的であるという知見もある [21]．また，日本人が英語を発音した場合に，英語のどの音素よりも日本語の音素に近いことも想定されるので，日本語の音素体系や音響モデルも考慮する必要がある．

上記で述べた評定スコアは正規化されているものの，人間（教師・評価者）の評定と必ずしも合致するとは限らない．そこで，線形回帰モデルなどを導入して，両者の写像を学習することも検討されている．またその際に，音韻的なモデルだけでなく，継続時間などの韻律的な要因も総合して評価関数を学習するのが望ましい．

GOP スコアに基づいて誤り検出を行うことも考えられるが，その際には，音素ごと若しくはそのクラスタごとにしきい値を設定する必要がある [19]．

3.6 音響モデル

外国語の発音学習を想定した音響モデルをどのように構築するかは，通常の音声認識の音響モデルの場合に比べて自明ではない．母語話者の音声データベースで学習したモデルは，“標準的な” 発音のモデルとしては妥当であっても，外国語学習者の発声には必ずしもマッチングしない．そこで，非母語話者（外国語学習者）の音声（例えば日本人が発声した英語）データベースを構築することが望まれる．しかし，そのようなデータを大規模に集めるのは容易でない．また，発音に誤りが含まれているので，その誤りを含めて忠実にラベル付与を行うのは，専門家による膨大な作業を要する．

したがって，母語話者の音響モデルを学習者・非母語話者の音声を用いて適応したり，IPA の体系で同一の単音とみなせる音素については，外国語学習者の母語音声（例えば日本人が発声した日本語）の音響モデルを利用するなどの解決策が考えられる．

坪田ら [22], [23] は，日本人の英語学習を対象として，音響モデル適応に関する詳細な検討を行った．ここでは，7 名の学習者が 850 単語を発声したデータベース（産総研で構築）を用いた．各発声には発音誤りを含めて人手でラベル付与が行われている．ベースラインの英語母語話者音響モデルは，TIMIT データベースを用いて学習したモノフォン（3 状態 16 混合）である．

まず，学習者ごとの話者適応の効果について調べた．各学習者について，100 単語発声を適応に用い，残り 750 単語発声で評価を行った（以下の表 2・表 3・表 4 で共通）．MLLR (Maximum Likelihood Linear Regression) 適応を行う際に，発音誤りを含む人手ラベルを用いる場合と，標準的な発音辞書に基づく場合を比較した．次節で述べる誤り予測に基づいて，発音

表 2　母語話者音響モデルの話者適応の効果
Table 2　Effect of speaker adaptation of native acoustic model.

話者適応	音素認識精度
適応なし	75.4%
人手ラベル	81.0%
辞書ラベル	80.6%

表 3　母語話者音響モデルと非母語話者音響モデルの比較
Table 3　Comparison of native acoustic model and non-native acoustic model.

	音素認識精度	
音響モデル	ベースライン	話者適応
英語母語話者モデル	75.4%	80.6%
日本人の英語モデル	78.0%	81.8%

表 4　学習者の母語音響モデルの併用の効果
Table 4　Effect of incorporating acoustic model of the learners' native language.

	音素認識精度	
音響モデル	ベースライン	話者適応
英語母語話者モデル＋日本語モデル	78.9%	81.3%
日本人の英語モデル＋日本語モデル	78.7%	81.5%

誤りを含めて音素が正しく認識された割合を表 2 に示す．話者適応により絶対値で約 5%の改善が得られたが，人手ラベルと辞書ラベルの差はほとんど見られなかった．すなわち，話者適応においては人手ラベルを用意する必要はないことが示された．これは，MLLR 適応がクラスタリングを介して行われるので，ある程度の誤りに対して頑健であるためと考えられる．

次に，非母語話者音響モデルを用いる効果について調べた．**7.2** で紹介する日本人の英語学習者の音声データベース (ERJ) を用いた．ERJ には発音誤りに関するラベル付与はされていないので，音響モデルを学習する際に，辞書ラベルを用いている．前述の話者適応の有無を含めて評価を行った結果を表 3 に示す．日本人話者の音声データベースで学習したモデルは，英語母語話者音響モデルに比べて高い認識精度が得られた．しかし，話者適応を行うとその優位性はほとんど見られなくなった．ただし，現実的な設定で必ずしも教師付きの話者適応を行えるわけではない．

更に，IPA の体系で同一の単音とみなせる音素について日本語の音響モデルを併用する効果を調べた．その結果を表 4 に示す．表 3 の結果と比較すると，英語母語話者音響モデルに日本語音響モデルを併用する効果が見られ，日本人の英語音声データベースで学習したモデルと同等の認識精度が得られるようになった．これは，多くの子音が日本語と英語で共有できるためと考えられる．

また，外国語の発音学習支援においては，トライフォンなどの音素文脈依存モデルよりも音素文脈独立モデル（モノフォン）の方が効果的であることが多い．これは，外国語の発音において必ずしも前後の音素文脈が信頼できないことに加えて，セグメンテーションを行う際には，音素間の境界が曖昧になる音素文脈依存モデルよりも，音素文脈独立モデルの方が精度が高くなるためである．

3.7　発音モデル

発音モデルは，通常の音声認識においては各単語の音素列を規定するものであるが，外国語の発音学習支援においては学習者が犯しやすい誤りを予測するものである．その際に，学習者の母語が特定されていれば，その知識を活用することができる．例えば，日本人が英語を発音する際に犯しやすい誤りに関しては，/l/と/r/, /v/と/b/のように多くの言語学的な知見がある．

京都大学の英語 CALL システム (**6.1**) においては，合計 79 種類の誤りパターンを用意している [22]．そのうち，37 種類は母音挿入に関するもので，35 種類が置換誤り，7 種類が脱落誤りに対応する．母音挿入については，特定の連続子音のパターン（/pl/や/tr/など）の間に/u/や/o/が挿入される場合と，単語末の特定の子音（/s/や/k/など）の後に/u/などが付加される場合を列挙している．置換誤りは，日本語の音節にないもの（/tu/→/tsu/など），日本語の音素にないもの（/v/→/b/など），母音の区別がつけられていないもの（/ou/→/o:/など）を列挙している．

これらの誤りパターンの規則を正しい発音の音素列に適用することによって，図 2 に示すような誤り予測を含むネットワークを構成することができる．このような誤りパターンの規則を記述するには，専門的な知識及び多くの知見を必要とする．規則を多くすれば，多くの現象をカバーできるものの，ネットワークが複雑になり，誤った検出が増えることにもなる．

したがって，人手で規則を記述するのではなく，発音誤りのラベルが付与されたデータから機械学習するアプローチが検討されている．例えば，Meng ら [24] は，大規模な中国人の英語音声データベースを構築し，誤りパターンを統計的に抽出している．また，Wang ら [25] は，日本語の発音学習を対象として，多くの誤りパターンの中から決定木学習を用いて有用なものを自動選択している．

3.8 音声の構造的表象に基づくアプローチ

音声認識で一般的に用いられている音響特徴量であるスペクトル包絡は，発音の善しあしだけでなく，話者の個人差（体格や性別）によっても変形する．したがって，学習者音声に対する音響モデルのゆう度は，その音声が模範的・平均的なモデルに音響的にどれだけ近いかを示す指標にすぎない．前節までに述べたように，音響モデルを学習者に事前に適応したり，特徴量正規化を施したり，あるいは，ゆう度比や GOP のようにゆう度の正規化を行うことにより，評定スコアと解釈できるよう対処することになる．

これに対して構造的表象とは，音声信号から（音韻情報とは無関係な）位相情報やピッチ情報を除去してスペクトル包絡が抽出されるように，この包絡特性から話者情報を除去して抽出される特徴量である [26]．話者（体格や性別）の違いは，二話者の音響空間の写像としてモデル化できるので，両空間で等しく観測される音響特徴量，すなわち写像不変の音響特徴量が定義できれば，それが話者の違いを超えた共通項，不変項となる．[27] では，二分布間距離の一定義である f-divergence[(注1)] が可逆かつ連続な任意写像に対する不変量であることの必要性及び十分性を証明している．

発声中の個々の音素の音響的特徴は話者性による影響を強く受けるため，構造表象では，音素（に相当する音声区間）と音素の関係性，すなわちこれらの差分量（コントラスト）のみに着眼する．ある発声に N 個の音素が観測された場合，これを N 個の確率分布として捉え，任意の 2 音素間の音響的距離を f-divergence で計測し，距離行列を構成する（図 3 参照）．この距離行列が不変項となる．個々の音素の音響的特徴をモデル化するのではなく，音素群を一つの体系として捉える方法論は，古くは構造音韻論 [28] で検討されている．

話者性の違いに頑健な発音評定 [29], [30] や誤り検出 [31]，学習者が選んだ特定の教師に近づけるために矯正すべき音韻の選定 [29], [32]，発音に基づく（体格や性別に依存しない）学習者分類 [29], [32] などの研究が行われている．

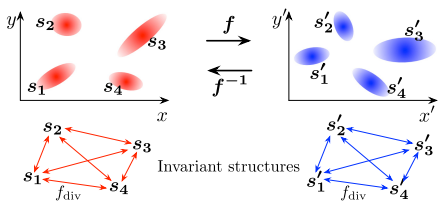

図 3 音群の体系として抽出される不変構造
Fig. 3 Invariant structure extracted as sound system.

4. 韻律的な発音の評価

学習者の発声には音韻的な誤り以外にも，不適切な高さ，長さ，強さの音の生成に起因する誤り，すなわち韻律的な誤りが多発する．韻律的な誤りの方が矯正し難く，学習後期になっても抜け切らないといわれており [33]，韻律的な矯正により母語話者に聞き取りやすい発声にするための教則本 [34] も発刊されている．

表 5 に基本的な韻律的特徴（心理量と物理量）と関連する言語的・非言語的現象についてまとめる．音の高さや大きさと異なり，音の長さについては心理量と物理量で異なる用語が用意されていない．音韻的側面の評価は，音声認識技術の適用という形で技術的構築がなされてきたが，音声認識では韻律的特徴を積極的に削除して（無視して）おり，韻律的側面の評価技術は（音韻的側面の評価技術と比較して）標準的な手段が確立されていない．本節では筆者らが調査した研究例を幾つかの着眼点から分類し，各々に対して解説する．

4.1 発音の流暢さと継続長に関する音響的特徴

どのような発声が流暢な発声といえるのだろうか．学習者の発声に対して教師が感じる流暢さ (fluency) と相関の高い音響的特徴に対する調査が古くから行われている [35], [36]．ここでは継続長に関する特徴量や，発声中の無音の数，言い淀みの数などが焦点となっている．例えば (1) rate of speech（[音素数]/[無音を含めた音声長]），(2) phonation ratio（[無音を省いた音声長]/[無音を含めた音声長]），(3) articulation ratio

表 5 韻律的特徴の種類と対応する言語的・非言語的現象
Table 5 Kinds of prosodic features and their corresponding linguistic or non-linguistic phenomena.

心理量	物理量	関連する現象
ピッチ	基本周波数（F_0）	イントネーション アクセント，個人性
ラウドネス	インテンシティー 音圧，パワー	アクセント
継続長	継続長	リズム アクセント
音色（声色）	スペクトル包絡	音素，個人性 アクセント

(注1)：$f_{div}(p_1, p_2) = \oint p_1(x) g\left(\frac{p_2(x)}{p_1(x)}\right) dx.$

([音素数]/[無音を省いた音声長])，(4) 文中に挿入されている無音長の総和，(5) 平均無音長，(6) 無音の数，(7) 無音と無音で挟まれた音声区間の平均音素数，(8) 言い淀み数などを計測し，相関分析を行っている．rate of speech だけでも約 0.9 ほどの高い相関が得られている．しかし（適度に）早口であることは流暢であることの必要条件であろうが，十分条件ではないだろう．

継続長と関連の深い言語特徴としてリズムがある．言語は (1) 強勢リズム（英語，独語など），(2) 音節リズム（仏語，伊語など），(3) モーラリズム（日本語など）のいずれかに分類される．母語話者の音声を対象に，その音響量からリズム識別を行う研究が行われている．[37], [38] では以下の音響特徴量が提案されている．

$$\mathrm{rPVI} = \frac{100}{m-1}\sum_{k=1}^{m-1}|d_k - d_{k+1}| \tag{6}$$

$$\mathrm{nPVI} = \frac{100}{m-1}\sum_{k=1}^{m-1}\frac{|d_k - d_{k+1}|}{(d_k + d_{k+1})/2} \tag{7}$$

d_k は vocalic interval（母音及びその連続で構成される区間）あるいは，consonantal interval（子音及びその連続で構成される区間）であり，上式は連続する二区間長の同一性を定量化している（後者は前者の正規化版）．これは等時性という考え方を基本にしている．例えば音節リズムは，音節と音節がおよそ等間隔に配置されていると知覚されることを意味している．[37], [38] では，(vocalic nPVI, consonantal nPVI) 平面を用いて，多言語の発声をリズム分類している．

一方 [39], [40] では，ΔV，ΔC，$\%V$ という尺度を用いてリズム分類を試みている．ΔV は vocalic interval の標準偏差，ΔC は consonantal interval の標準偏差であり，$\%V$ は発声に占める母音の割合である．PVI は連続する 2 区間で，ΔV，ΔC は発話全体で等時性を定量化しているといえる．[39], [40] では ($\%V$, ΔC) 平面でのリズム分類などが行われている．

以上紹介した音響特徴量は対象言語のリズム的な「その言語らしさ」を反映していると考えられ，学習者音声を用いた分析も行われている [41]．また，等時的リズムを直接扱ったものではないが，[42] では，時間制御に関する母語話者の知覚特性に基づいて学習者音声の客観的評価を試みている．

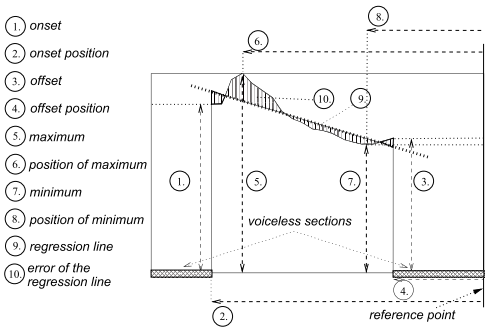

図 4　単語単位のF_0特徴量の定義
Fig. 4　Definition of word-based F_0 features.

4.2　様々な韻律的特徴量を用いた発音評定

様々な観点から韻律的特徴を定義し，それを用いて教師による「韻律的な発音習熟度」及び「総合的な発音習熟度」の自動推定が検討されている．この場合，教師スコアと相関値の高いスコアを自動推定する回帰モデルが構築される．線形回帰，リッジ回帰，サポートベクター回帰，ロジスティック回帰が使用されることが多い．

[43] では，言語非依存な韻律的特徴を用いて「韻律的な発音習熟度」の予測を試みている．図 4 に示すように，文中の各単語に対応するF_0パターンに対して，(1) 先頭のF_0値，(2) 終了時のF_0値，(3) 最大値，(4) 最小値，(5) 直線回帰の傾斜，(6) 回帰誤差，などを特徴量として採用している．これら単語単位での韻律的特徴以外にも文を単位とした韻律的特徴を定義し，最終的に 148 種類の韻律に関する特徴量を定義し，サポートベクター回帰を用いて「韻律的な発音習熟度」を予測している．

[44] では「総合的な発音習熟度」の推定をタスクとして，ゆう度比や GOP に相当する種々の音韻的スコアに幾つかの韻律的スコアを追加し，これらを線形回帰の枠組みで統合している．音韻的特徴と韻律的特徴は発声のメカニズムとしては独立であるが，実際の発声には相関が観測されており，最終的に，教師スコアとより相関の高い予測を行う説明変数の組合せが検討されている．各説明変数単独では，韻律的特徴の中では話速 (rate of speech) が最も高い相関を示したが，音韻的スコアとの組合せで有効なものはパワー値の分散であった．[43], [44] のいずれにおいても，0.88 ほどの相関が示されており，高い予測性能をもつ回帰モデルが構築されている．

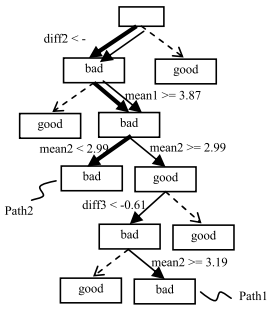

図 5　構築されたトーン正誤判定決定木の一例
Fig. 5　A decision tree built for tone judgment.

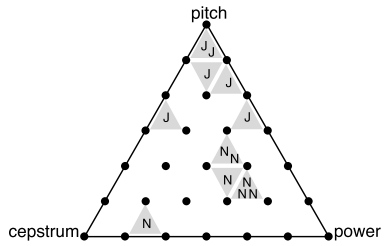

図 6　推定された強勢生成時の癖の様子
Fig. 6　Estimated habits observed in word stress generation.

4.3　種々の言語単位に着眼した韻律的評定

韻律的特徴は音素や節 (segment) を超えて存在する超分節的な特徴であるが，表 5 に示すように，音節，単語，句，文と，各単位に対して各々異なる言語的情報として存在している．以下では，中国語のトーン（音節レベル），英語の語アクセント（単語レベル），更には文発声時の韻律的特徴（句，文レベル）に対して行われている研究例を紹介する．

[45] では，中国語学習者のトーン評価を検討している．各音節で計測される F_0 パターンを 3 等分割し，各区間の平均値，及び，任意 2 区間の F_0 平均差を用いて特徴量を定義しており，一音節が六次元ベクトルとして表現される．トーンの正誤を自動判定するだけでなく，誤りであると判定された理由を学習者に示すことを目的としている．検定試験などに用いるような発音評定システムでは，スコア提示のみを実現すれば十分であるが，日々の学習では,「なぜ誤りと判定されたのか」「どのようにすればそれは改善されるのか」に関するフィードバックが重要である.[45] では，決定木に基づいて正誤判定及びフィードバック生成を試みている．構築された決定木の例を図 5 に示す．ルートノードから，各種 F_0 特徴量を参照し，質問に答えることで木をたどり，リーフノードに到達する．リーフノードには正誤のフラグがあり，この結果をユーザに提示するとともに，そこに至るまでの質問と特徴量からフィードバック生成を行う．

英単語アクセントに関する韻律的評定に関して,[46] では，孤立単語発声に対して強勢音節位置を同定する強勢検出器を構築し,[47] ではこれに基づいて，強勢が適切な韻律的特徴によって生成されているか，強勢生成時の発音癖の推定を検討している．

[46],[63] では F_0，パワー，継続長などの韻律的特徴と音韻的特徴（スペクトル特徴）の両方を用いた音節単位での HMM を構築し，強勢検出を行っている．音声認識での音素モデルの構築と同様に，様々なコンテキスト情報を使って音節モデルの精緻化が行われている．例えば，中心母音の種類（単母音，長母音，二重母音），音節の構造，単語内位置などである．母語話者発声の場合はより精緻なモデルが検出率を向上させるが，非母語話者の学習者音声の場合は（セグメンテーション同様），適度に粗いモデルが頑健で，結果的に精度も高いモデルとなる．

上記の強勢検出用音節 HMM のゆう度は，(1) F_0 ゆう度，(2) パワーゆう度，(3) 継続長ゆう度，(4) 音色ゆう度の 4 種類のゆう度の重み付け和として計算される．重みの最適化によって検出率を向上させることができるが，更に [47] では，最適重みを使って学習者の強勢生成癖の推定を検討している．これは日本人学習者は強勢生成を主に F_0 の上げ下げで実現する傾向があるからである [48]．図 6 に，母語話者の場合の最適重みと日本人学習者の最適重みの様子を示す．日本人の場合 F_0 に頼って強勢弱勢を生成する傾向があるが，母語話者の場合は各種の特徴をバランスよく用いている様子がうかがえる．

単語を超えた韻律パターンを扱う場合，例えば句や文を単位として扱うのではなく，単語単位で扱い，各単語でのスコアを統合する形で文レベルのスコアを求めることが多い．日本人による英文音声のF_0パターンは，(1) 単語ごとにポーズを置いて発声するためにF_0の山谷が多くなったり，(2) 逆に極端にフラットなカーブを描くこともある [49]．これを鑑みて [50] では，単語や句など複数の言語単位を用意して，F_0パターンやパワーパターンなど各種韻律的特徴を，学習者音声・モデル発声間で（一対一で）DP マッチング（動的計画法）により照合している．一方 DP を行わずに，文

中の各単語発声を 25 等分し，F_0パターンやパワーパターンを比較する試みもある [51]．韻律的特徴は発話スタイルなどの影響を強く受けて容易に変形されるため，異なる話者の同一単語数十発声を（等分割後に），数種類のテンプレートに分類し，一単語当り複数の正解韻律パターンを用意した上で，学習者パターンと複数のモデルパターンを比較する方式をとっている．

5. 音声合成技術の利用

テキスト音声合成 (Text-To-Speech：TTS) 技術は，入力テキストに対して適切な音声信号を生成する技術である．ここでは音声合成技術が外国語教育・学習支援にどのように貢献できるかについて解説する [52]．

5.1 TTS 技術の応用場面と要求される品質

近年の TTS 技術の進展により，母語話者の生活空間でその言語の TTS 出力を耳にすることが多くなった．母語話者に受け入れられるようになった TTS 品質は，当然ながら，その言語を学ぶ学習者に提供するモデル音声としての利用が検討されている．母語話者（教師）に発声を依頼せずとも読上げテキストを即座に音声化できる利便性は大きい．しかし，その音声に不適切な部分があったとしても学習者は気づくことが困難である．TTS に対する自然性評価試験は，母語話者評価と非母語話者評価は区別され，当然前者がより厳しい評価となる [53]. [54] ではこれらの点を鑑みて，2005 年当時の TTS システム出力が外国語学習においてどのような場面で利用可能か，以下の三つの場合について検討している．

（1） 電子化辞書など，初めてその単語と遭遇する学習者に呈示すべき音声（モデル音声）としての利用．

（2） Web テキストや学習者の読上げ原稿を読み上げさせたり，ディクテーションやシャドーイングの訓練で使われる音声としての利用．

（3） 対話形式で行われる CALL システムに登場する対話エージェントの音声としての利用．

当然前者の方ほど高精度，高自然性の音声が必要となる．上記論文では，(1) の場合は人間の音声が使われるべきであり，(3) の場合は利用可能であると述べている．英語を対象とした場合，会話相手が母語話者であることの方が少なく，その意味においても多少の不自然さを訛りと考えれば十分に許容できるのであろう．

上記 (2) に関して，最近国内において英語 TTS がディクテーションやシャドーイング練習に用いられるようになった．これは英語 TTS の高品質化と，英語授業での利用に特化した各種機能やインタフェースを実装した商用アプリの登場に起因するところが大きい(注2)．英語以外の言語でもディクテーション訓練にポルトガル語の TTS を利用している例がある [55]．TTS では話速を自由に制御できるので，学習者のレベルに合わせた話速設定などが可能になっている．

5.2 合成音声以外のシステム出力を用いた支援

TTS システムでは，与えられたテキストに対して，テキストには明示されていない種々の音韻的・韻律的情報を推定し，最終的な音声波形に反映させる必要がある．日本語の場合，(1) 母音はいつ無声化するのか，(2)「らりるれろ」はいつ/r/となり，いつ/l/となるのか，(3) アクセント句境界はどこにあるのか，(4) アクセント核はアクセント結合により移動するが，最終的にはどこに来るのか，などである．これらはいずれも日本語学習者が直面する問題でもある．TTS 出力を呈示すれば上記質問に「音声を使って」解答することになるが，視覚的に与えた方が分かりやすい．

学習者のテキスト読上げ支援の一貫として，TTS システムの内部モジュールの出力を学習者に明示的に示すことが検討されている. [56] では，入力テキストに対して，アクセント句境界推定，アクセント句内の（文発声としての）アクセント核位置推定を行い，核の位置を視覚的に示すシステムが検討されている．

5.3 分析合成技術を用いた学習者の知覚過程の分析

テキストを入力とする TTS 技術とは異なるが，分析合成技術を使えば，任意の音声を入力とし，その音声を変形する（モーフィング）ことが可能となる．高品質な分析合成システムである STRAIGHT [57] は，音声知覚実験用音声試料の音響的変形（定量的変形が頻繁に要求される）に広く使われており，非母語話者を対象とした知覚実験にも使われている．

音声のソースフィルタモデルに基づいて，入力音声を，(1) スペクトル包絡（パワー情報含む），(2) 基本周波数，(3) 有声度の時系列に分解し，これらを変形して再統合することで合成音声を得る. [58] では，同一話者の “right” と “light” の発声に対して，その中間の発声を数段階に分けて（内挿）構成し，母語話者及び日本人学習者を対象にして，/r/と/l/の同定実験が行われている．母語話者の場合，刺激の連続的な変化に伴い知覚がしきい値的に変化するが（カテゴリカ

(注2)：例えば，http://voicetext.jp/gv/pro_gve.html など．

ル知覚），日本人学習者ではそのような変化は見られない．

[59], [60] では，韻律的変形を用いた聴取実験が行われている．日本語・米語のバイリンガル話者による正しい日本語単語発声と，その単語の米語訛り発声とを内挿し，複数段階の訛った音声を作成した．その際に，ピッチのみ，パワーのみ，スペクトルのみ，それらの組合せの複数の方式に対して，複数段階の内挿による合成音声を作成し，知覚実験が行われた．タスクは米語訛りの度合いの回答（5 段階）である．日本語母語話者と日本語を学ぶ豪語話者を対象にして結果を比較すると，ピッチの変形に対して母語話者は敏感に反応するのに対し，豪語話者はピッチの変形による訛りには非常に鈍感であると報告されている [59]．

学習者の知覚過程分析とは異なるが，学習者音声に対して母語話者の韻律的特徴量を転写してモーフィングすることで正しい音声へと変形し，学習者にフィードバックする試みもある [61]．ただし，骨導音が考慮されていないため，本人が知覚する自分の声（自己聴取音）にはならないといった問題もある．

6. 外国語 CALL システムの紹介

本章では，これまでに開発されている代表的な外国語発音学習支援システムについて紹介する．

6.1 京都大学の英語 CALL システム Hugo

京都大学では 1998 年に総合情報メディアセンター（現学術情報メディアセンター）が設立されて以来，CALL 教室と教材の整備を進めてきた．CALL 教材は，英語のほかに中国語・フランス語・ドイツ語・ベトナム語などの言語に対して，担当教員が音声・映像の収録から自前で作成しており，一部について音声情報処理の技術を導入したシステムの研究開発も進めている．

英語 CALL システムは，日本人学生が日本の文化を外国人に紹介できるようになることを目標として設計・作成している．そのためのスキットを用意し，英語母語話者による会話を収録し，マルチメディア教材としている．学習者は，説明者役の文章を読み上げる訓練を行う（図 7 参照）．

Hugo では，その発音に対して自動評価と誤り検出を行う [23], [62]．スキットを一通り終えた後で，特に重要と思われる誤りパターンに関して，単語やフレーズの単位で重点的に訓練を行う．発音の評価は，音韻的な観点と韻律的な観点で行われる．音韻的な処理は

図 7　英語 CALL システムのスキット訓練画面
Fig. 7　Screen shot of role-play practice in English CALL system.

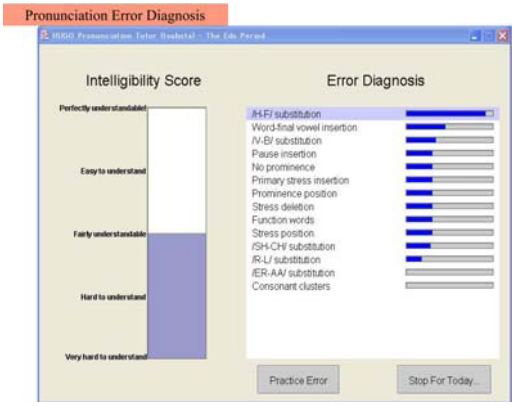

図 8　英語 CALL システムの発音評価画面
Fig. 8　Screen shot of pronunciation evaluation in English CALL system.

3. で述べたように，誤りパターンの予測を行う発音モデルと日本人の英語に対応した音響モデルを用いて，セグメンテーションと誤り検出を行う [22]．韻律的な処理は，**4.3** で説明された強勢のモデルに基づいて，強勢パターンの誤り検出を行う [63], [64]．その上で，日本人学習者が犯しやすい誤り 10 種類に関して，統計的な分析を行い，了解度を 5 段階で評定する．更に，了解度を改善するために最も効果的と考えられる誤りパターンを同定する（図 8 参照）[65]．その誤りパターンを含む単語やフレーズを提示して，模範的な母語話者のパターンと比較しながら集中的な訓練を行う．

本システムは，英語 CALL の授業の一部で試験的に使われた．若干の試行を経て，音声認識誤り等の誤動作はほとんどなくなり，学生からも好評であった [66]．

6.2 CMU の英語 CALL システム Native Accent

米国・カーネギーメロン大学 (CMU) の Eskenazi らは，早くから英語発音学習支援に関する研究を進めており，本システムはその成果を商用展開したものである [67]．音韻レベル並びに韻律レベルに関する誤り検出と調音器官の図示によるフィードバックを行う．全部で約 800 の演習課題と，日本語，ロシア語，フランス語など 28 の母語話者に対応した誤りパターン・フィードバックのモデルを用意しており，母語に応じて演習課題を設定できる．個々の学習者に応じた進度の表示や教師へのレポート機能なども備えている．

6.3 CUHK の英語 CALL システム

香港中文大学 (CUHK) の Meng らは，中国人を対象とした英語学習のための大規模な音声データベースの構築とシステムの研究開発を進めている [24]．音声データベースは，広東語話者 100 名と標準中国語話者 111 名が孤立単語や物語のパラグラフを読み上げたものである．発音誤りを予測するモデルは，人手で作成した規則とデータベースから統計的に抽出したものを比較・統合しているほか，GOP スコアに基づく評定など本論文で述べた技術が組み込まれている．学習者へのフィードバックでは，誤り部分を強調した音声の合成のほか，調音器官のアニメーションの生成も行っている．

前記の京都大学のシステムの試験評価 [66] でも報告されているが，CALL システムを実際に運用する際に大きな問題となるのは，録音レベルの問題やフィラーや言い直しなどにより，入力音声がシステムの想定外のものになることである．本システムでは，各課題の単語や文の各音素の継続長モデルを用いることにより，想定外の入力を棄却する機能を備えている．

6.4 シャドーイングによる訓練

峯松らは，模範的な母語話者の音声を聞きながら追随して発声を行うシャドーイングに基づく訓練システムの検討を行っている [68]．シャドーイングは聴覚呈示された母語話者の音声を即座に繰り返す訓練方法であり，リスニングとスピーキングを同時に訓練するため，認知負荷が高くなる．その結果，明確に調音されずに発声される傾向があるが，逆に，単純な読上げ音声よりも，学習者の英語能力をより適切に反映した音声資料となることが期待される．本システムでは音素単位の GOP スコアに基づいて，発話ごと及び話者ごとに評定を行っており，実験の結果，単純な読上げの場合に比べて GOP スコアと TOEIC スコアの相関が有意に高くなることが示されている．

6.5 ETS の Speech Rater

TOEFL を主催している ETS では，スピーキングの自動評価を行う研究開発を進めている [5], [6]．TOEFL のスピーキングでは，与えられた課題に対して数分のスピーチを行う．これは文章が与えられた前提の発音訓練と異なり，語彙や文法のスキルも必要となる．このような自由発声の評価に関しては，本論文の **3.** で述べた方法論を適用できず，通常の大語彙連続音声認識（ディクテーション）の方法論を用いることになる．音響モデルは非母語話者のデータ（ベースラインとして 30 時間）を用いて学習しており，言語モデルは非母語話者のデータに加えて，放送ニュースのテキストを用いて構築している．

この音声認識結果（単語とその信頼度），発話速度（秒当り音素・単語数），ポーズの頻度や長さなどの計 40 の特徴（=素性）を用いて，人間による評定への写像を求める回帰モデルを学習している．人間による評定との相関は 0.67 であり，一定レベルと考えられるが，異なる人間の評定間の相関 (0.94) に比べるとかなり低い．したがってまだ実用に供するレベルではないが，TOEFL iBT Practice Online (TPO) などで試験的に使われている．

6.6 POSTECH の対話型英語 CALL システム

韓国の浦項工科大学 (POSTECH) では，韓国人を対象とした英語の会話学習支援システムの研究開発を進めている [69]．これは，ショッピングや道案内などの限定された状況において会話の練習を行うものである．状況は限定されているが，文は与えられていないので，通常の音声対話システムの方法論を応用している．すなわち，タスクドメインに特化した言語モデルを用いた音声認識及び言語理解を行った上で，用例ベースの対話管理を行っている．近い用例に基づいて，学習者にフィードバックを行っている．ロボットを用いたシステムも構築しており，小学生に対して本格的なフィールドテストを実施している．

6.7 京都大学の日本語 CALL システム CALLJ

本システムは，日本語の基本的な文の生成・発音の学習支援を行うものであり，日本語検定 3 級及び 4 級に準拠して 30 レッスンから構成される．各レッスンごとに設定された典型的な文型に関して，学習者は図で示された状況を表す文をタイプ入力若しくは音声入力する（図 9 参照）．この課題は，設定された語彙と

図 9　日本語 CALL システムの画面
Fig. 9　Screen shot of Japanese CALL system.

文型をランダムに組み合わせることで生成される．それに応じて，学習者に与えるヒント，誤りパターンの予測を含む音声認識文法，誤りに応じたフィードバックが動的に生成される [25]．

7.　CALL 研究のためのデータベース

7.1　音声・言語コーパスに関する有用な情報源

音声認識や音声合成の研究と同様に，CALL に関連する研究においてもコーパスは非常に大きな役割を担う．表 6 に国内外の主要な音声・言語コーパス配布サイトを示す．これらの多くは母国語のコーパスであり，外国語のコーパス，更には外国語教育を目的としたコーパスとなると数が限られる．外国語教育や CALL 研究のためのコーパスに特化した情報源としては，[70] が当時の状況についてのレビュー論文となっており，その内容が Wikipedia に公開されている（表 6 参照）．

次節では，外国語学習者音声コーパス構成の典型例として，[70] でも取り上げられている「日本人学生による読上げ英語音声コーパス (English Read by Japanese, ERJ)」[71] について詳しく述べる．

7.2　ERJ の開発とその利用

本コーパスは，CALL システム開発や日本人の英語発音分析，更には英語教育での利用を目的として，技術的要請，教育的要請の両者を満たすように設計・構築された．話者数は成人男女 100 名ずつである．日本における英語教育の現状を考慮して，対象言語は米語 (General American English：GAE) とし，GAE を母語とする話者（20 名）の音声も収録されている．外国語の自由な発声の場合，言い淀みや言い直しなどの現象が多発し，技術的な扱いが困難となるため，読上げ音声のコーパスとなっている．また，読み間違いがなくなるまで繰り返し収録するよう指示を与えて，学習者が「正しい」と思った音声のみを収録している．様々な発声のヒントを読上げリストに付与しているが，多種多様な発音誤りが観測されている．

技術的及び教育的要請を考慮し，音韻的側面に着眼した単語・文セットと韻律的側面に着眼した単語・文セットを設計した（表 7，表 8 参照）．ここでは音素バランスや韻律のバリエーションが反映されており，合計約 800 文，1000 単語である．一人当りの読上げ量は約 120 文，約 220 単語である．母語話者には文セットの半分ずつを読ませた．読上げリストには音素記号，強勢記号，イントネーション記号などを必要に応じて付与し（付与しない版も用意），辞書を引かずとも読めるよう配慮した．

外国語音声コーパスを構築しても，個々の発声に対

表 6　音声・言語コーパスに関する有用 Web サイト
Table 6　Useful Web sites to search for speech and language corpora.

Linguistic Data Consortium (LDC, USA) http://www.ldc.upenn.edu/
European Language Resources Association (ELRA, EU) http://www.elra.info/
Speech Resource Consortium (SRC, Japan) http://research.nii.ac.jp/src/
Advanced LAnGuage INformation forum (ALAGIN, Japan) http://www.alagin.jp/
GSK (Gengo-Shigen-Kyokai, Japan) http://www.gsk.or.jp/index.html
Chinese Linguistic Data Consortium (C-LDC, China) http://www.chineseldc.org/
Non-native speech database [70] http://en.wikipedia.org/wiki/Non-native_speech_database

表 7　音韻的側面を考慮した単語・文セット
Table 7　Word and sentence sets designed for the segmental aspect of speech.

カテゴリ	サイズ
音素バランス単語	300
ミニマル単語対	600
音素バランス文	460
発音困難な音素列を含む文	32
音素学習に対する評価文	100

表 8　韻律的側面を考慮した単語・文セット
Table 8　Word and sentence sets designed for the prosodic aspect of speech.

カテゴリ	サイズ
種々の強勢パターンを含む単語	109
種々のイントネーションパターンを含む文	94
種々のリズムパターンを含む文	120

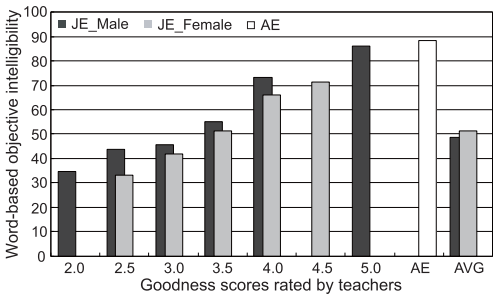

図 10　日本人英語及び母語話者英語の聞取り率
Fig. 10　Word-based intelligibility of Japanese English and American English.

する習熟度スコアや発音誤り箇所のラベリングが行われていなければ，CALL システム開発での利用は難しい．ERJ では，(1) 音素生成，(2) リズム生成，(3) イントネーション生成の三つの観点から，スコアリングを英語教師に依頼した．日本人学生に対する英語教育経験をもち，英語音声学に造詣の深い米語話者教師 5 名にスコア付けをしてもらった．話者別，観点別に数文ずつが選定され，発声単位でスコアリングが行われた（選定された文・単語は話者によって異なる）．なお，誤り箇所のラベリングは行われていない．

ERJ は CALL システム開発や日本人英語の音響分析など多角的に利用されているが（具体例は [71] 参照），ERJ を基に新たなコーパス構築も行われている．ERJ は日本人学習者が「読めた」と思った発声集であり，学習者にとっては正しい発声集である．これを日本人英語との接点の低い米語話者に呈示した場合，どのくらいの率で聞き取ってもらえるのだろうか. [72] では，ERJ の各文の単語数，単語 unigram，単語 bigram，及び各学習者の発音習熟度を基礎統計量として，ERJ のコンパクトセット（約 400 文）を定義し，これを用いて日本人英語発声（男女各約 400 発声ずつ）と，ランダム選択により抽出された GAE 話者発声（男女計 100 発声）を選定した．

上記音声を電話越しに米語話者に聴取させ，聞こえたとおりに口頭で繰り返させた．繰返し音声を，熟練した作業者によって書き起こさせた．最終的に ERJ コンパクトセットの各発声に対して約 20 名の米語話者による聞取り結果が得られた．図 10 は，各学習者の発音評定スコアを横軸に，その学習者の発声に対する単語単位での聞取り率（正解率）を縦軸にしてグラフ化したものである．今回の実験では，電話越しに前後の文脈なく文発声が呈示されることもあり，GAE 話者の音声であっても約 9 割の聞取り率であった．日本人英語音声は平均約 5 割の聞取り率となった．学習者が正しいと思った発声でも，文脈がない状況では，半分程度しか通じていないことを示している．

英語発音教育の現場では，「母語話者のような発音」を目指すべきか，「伝わる発音」を目指すべきかという議論がある [73], [74] が，本研究では後者に焦点を当て，聞誤り集を構築していることに相当する．World Englishes の重要性が叫ばれる今日では「伝わる発音」の方が実用的価値は高いと思われるが，この場合，「伝わりやすさ」は当然聞き手に依存する．上記コーパス開発は米語話者に対して行われたが，今後クラウドソーシング等を利用して，英語話者，豪語話者，加語話者などを対象にしたデータ収集が期待される．

8. むすび

以上，外国語学習支援のための音声情報処理技術，並びに CALL システムとデータベースに関する現状について概観を行った．

今世紀に入ってからの大きな変化は，国際化・グローバル化に加えて，ブロードバンドネットワークの普及に伴う Web ベースの多様なサービスの展開である．教育・学習の分野も例外ではない．例えば，講義のコンテンツやアーカイブが世界的に配信されるようになっている．更に，昨今のスマートフォンなどの普及に伴って，新たな様相を呈している．

外国語学習支援システムもスタンドアローンの計算機による CALL から，ネットワーク・Web を介した WELL (Web Enhanced Language Learning) に展開していくのは自然な流れであろう．これにより，教材やシステムを端末でなく，サーバ側で管理・更新できる．また，学習者のデータを集積することにより，大規模な評価やシステムの改善につなげていくこともできる．これは，音声認識システムをサーバベースで展開していくことにより，高精度化を実現できたのと同様の枠組みと捉えられる．このようなサイクルを加速するために，まずは大学の CALL 教室で試験運用を行っていくことが重要である．

更に，遠隔にいる教師や外国にいる母語話者と直接コミュニケーションを行うことも容易になってきており，これまで考えられなかったような学習支援の形態も模索できよう．音声認識・音声合成システムの高度化と併せて，今後ますます新たな可能性が広がることが期待される．

謝辞　京都大学の壇辻正剛教授には日ごろより数多くの知見を頂きました．坪田康助教には，本論文の実験データ (**3.6**) を提供頂くとともに，CALL システム (**6.**) の調査に協力頂きました．本論文で紹介した研究に従事された学生諸君にも感謝します．

文　献

[1] 壇辻正剛，“IT 化時代の語学環境としての CALL,” 情報処理，vol.42, no.10, pp.1001–1005, 2001.

[2] 中川聖一，牧野正三，壇辻正剛，“音声言語処理技術を用いた語学学習システム,” 日本音響学会誌，vol.59, no.6, pp.337–344, 2003.

[3] M. Eskenazi, “An overview of spoken language technology for education,” Speech Commun., vol.51, pp.832–844, 2009.

[4] J. Bernstein, “The theory and practice of speaking assessment,” TESOL Annual Convention, 2004.

[5] K. Zechner, D. Higgins, and X. Xi, “Speechrater.: A construct-driven approach to scoring spontaneous non-native speech,” Proc. SLaTE, 2007.

[6] L. Chen, “Audio quality issue for automatic speech assessment,” Proc. SLaTE, 2009.

[7] Q. Liu, S. Wei, Y. Hu, W. Guo, and R. Wang, “The application of phone weight in Putonghua pronunciation quality assessment,” Proc. ISCSLP, 2006.

[8] M. Russell, C. Brown, A. Skilling, R. Series, J. Wallace, B. Bonham, and P. Barker, “Applications of automatic speech recognition to speech and language development in young children,” Proc. ICSLP, pp.176–179, 1996.

[9] J. Bernstein, “Intelligibility and simulated deaf-like segmental and timing errors,” Proc. IEEE-ICASSP, pp.244–247, 1977.

[10] 山田玲子，ATR CALL 完全版英語リスニング科学的上達法，講談社，1999.

[11] J. Bernstein, “Objective measurement of intelligibility,” Proc. ICPhS, pp.1581–1584, 2003.

[12] P. Badin, Y. Tarabalka, F. Elisei, and G. Bailly, “Can you ’read’ tongue movements? evaluation of the contribution of tongue display to speech understanding,” Speech Commun., vol.52, pp.493–503, 2010.

[13] Y. Tsubota, M. Dantsuji, and T. Kawahara, “Computer-assisted English vowel learning system for Japanese speakers using cross language formant structures,” Proc. ICSLP, vol.3, pp.566–569, 2000.

[14] C.-H. Jo, T. Kawahara, S. Doshita, and M. Dantsuji, “Japanese pronunciation instruction system using speech recognition methods,” IEICE Trans. Commun., vol.E83-D, no.11, pp.1960–1968, Nov. 2000.

[15] C.-H. Lee and M. Siniscalchi, “An information-extraction approach to speech processing: Analysis, detection, verification and recognition,” Proc. IEEE, vol.101, no.5, pp.1089–1115, 2013.

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

[17] L. Lee and R.C. Rose, “Speaker normalization using efficient frequency warping procedures,” Proc. IEEE-ICASSP, pp.353–356, 1996.

[18] 中川聖一，“CALL と音声情報処理技術,” 音声研究，vol.9, no.2, pp.28–37, 2005.

[19] A. Ito, Y.L. Lim, M. Suzuki, and S. Makino, “Pronunciation error detection method based on error rule lustering using a decision tree,” Proc. INTERSPEECH, pp.173–176, 2005.

[20] L. Neumeyer, H. Franco, V. Digalakis, and M. Weintraub, “Automatic scoring of pronunciation quality,” Speech Commun., vol.30, pp.83–93, 2000.

[21] S. Wei, G. Hu, Y. Hu, and R. Wang, “A new method for mispronunciation detection using support vector machine based on pronunciation space models,” Speech Commun., vol.51, pp.896–965, 2009.

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

[23] Y. Tsubota, T. Kawahara, and M. Dantsuji, “An English pronunciation learning system for Japanese students based on diagnosis of critical pronunciation errors,” ReCALL J., vol.16, no.1, pp.173–188, 2004.

[24] H. Meng, W.-K. Lo, A.M. Harrison, P. Lee, K.-H. Wong, W.-K. Leung, and F. Meng, “Development of automatic speech recognition and synthesis technologies to support Chinese learners of English: The CUHK experience,” Proc. APSIPA ASC, pp.811–820, 2010.

[25] H. Wang, C.J. Waple, and T. Kawahara, “Computer assisted language learning system based on dynamic question generation and error prediction for automatic speech recognition,” Speech Commun., vol.51, no.10, pp.995–1005, 2009.

[26] 峯松信明，櫻庭京子，西村多寿子，喬　宇，朝川　智，鈴木雅之，齋藤大輔，“音声に含まれる言語的情報を非言語的情報から音響的に分離して抽出する手法の提案—人間らしい音声情報処理の実現に向けた一検討,” 信学論（D），vol.J94-D, no.1, pp.12–26, Jan. 2011.

[27] Y. Qiao and N. Minematsu, “A study on invariance of f-divergence and its application to speech recognition,” IEEE Trans. Signal Process., vol.58, no.7, pp.3884–3890, 2010.

[28] R. Jakobson, L. Waugh（著），松本克己（訳），言語音形論，岩波書店，1986.

[29] 朝川　智，峯松信明，広瀬啓吉，“音声の構造的表象に基づく英語学習者発音の音響的分析,” 信学論（D），vol.J90-D, no.5, pp.1249–1262, May 2007.

[30] 鈴木雅之，峯松信明，広瀬啓吉，“音声の構造的表象と多段階の重回帰を用いた外国語発音評価,” 情処学論，vol.52, no.5, pp.1899–1909, 2011.

[31] T. Zhao, A. Hoshino, M. Suzuki, N. Minematsu, and K. Hirose, “Automatic Chinese pronunciation error detection using SVM with structural features,” Proc.

IEEE-SLT, pp.473–478, 2012.
[32] 峯松信明，鎌田　圭，朝川　哲，牧野武彦，西村多寿子，広瀬啓吉，“音声の構造的表象に基づく学習者分類の検証と発音矯正度推定の高精度化，”情処学論，vol.52, no.12, pp.3671–3681, 2011.
[33] T. Shibata and R.R. Hurtig, “Prosody acquisition by Japanese learners,” in Understanding Second Language Process, ed. Z. Han, Multilingual Matters, 2007.
[34] 中川千恵子，中村則子，許　舜貞，さらに進んだスピーチ・プレゼンのための日本語発音練習帳，ひつじ書房，2009.
[35] C. Cucchiarini, H. Strik, and L. Boves, “Quantitative assessment of second language leaners’ fluency: An automatic approach,” Proc. ICSLP, 1998.
[36] C. Cucchiarini, H. Strik, and L. Boves, “Quantitative assessment of second language learners’ fluency: Comparisons between read and spontaneous speech,” J. Acoust. Soc. Am., vol.111, no.6, pp.2862–2873, 2002.
[37] E. Grabe, B. Post, and I. Watson, “The acquisition of rhythmic patterns in English and French,” Proc. ICPhS, pp.1201–1204, 1999.
[38] E. Grabe and E.L. Low, “Durational variability in speech and the rhythm class hypothesis,” in Laboratory Phonology 7, ed. C. Gussenhoven and N. Warner, pp.515–546, Berlin, New York (Mouton de Gruyter), 2002.
[39] F. Ramus, M. Nespor, and J. Mehler, “Correlates of linguistic rhythm in the speech signal,” Cognition, vol.73, no.3, pp.265–292, 1999.
[40] F. Ramus, “Autoamtic correlates of linguistic rhythm: Perspective,” Proc. Speech Prosody, 2002.
[41] 里井久輝，吉村満知子，薮内　智，“日本人英語学習者における発話のスピーチリズムと母音弱化の関係，”日本音声学会全国大会予稿集，pp.33–38, 2003.
[42] S. Nakamura, S. Matsuda, H. Kato, M. Tsuzaki, and Y. Sagisaka, “Objective evaluation of English learners’ timing control based on a measure reflecting perceptual characteristics,” Proc. IEEE-ICASSP, pp.4837–4840, 2009.
[43] A. Maier, F. Honig, V. Zeissler, A. Batliner, E. Korner, N. Yamanaka, P. Ackermann, and E. Noth, “A language-independent feature set for the automatic evaluation of prosody,” Proc. INTERSPEECH, pp.600–603, 2009.
[44] K. Hirabayashi and S. Nakagawa, “Automatic evaluation of English pronunciation by Japanese speakers using various acoustic features and pattern recognition techniques,” Proc. INTERSPEECH, pp.598–561, 2010.
[45] H.-C. Liao, J.-C. Chen, S.-C. Chang, Y.-H. Guan, and C.-H. Lee, “Decision tree based tone modeling with corrective feedbacks for automatic Mandarin tone assessment,” Proc. INTERSPEECH, pp.602–605, 2010.
[46] 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，”信学論（D-II），vol.J82-D-II, no.11, pp.1865–1876, Nov. 1999.
[47] 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，”信学論（D-II），vol.J83-D-II, no.2, pp.486–499, Feb. 2000.
[48] Y. Shibuya, “Differences between native and non-native speakers’ realization of stress-related durational patterns in American English,” J. Acoust. Soc. Am., vol.100, no.4, pt.2, p.2725, 1996.
[49] 清水克正，英語音声学—理論と学習，勁草書房，1995.
[50] Y. Yamashita, K. Kato, and K. Nozawa, “Automatic scoring for prosodic proficiency of English sentences spoken by Japanese based on utterance comparison,” IEICE Trans. Inf. & Syst., vol.E88-D, no.3, pp.496–501, March 2005.
[51] J. Cheng, “Automatic assessment of prosody in high-stakes English tests,” Proc. INTERSPEECH, pp.1589–1592, 2011.
[52] A. Black, “Speech synthesis for educational technology,” Proc. ISCA Workshop on Speech and Language Technology in Education (SLaTE), 2007.
[53] 徳田恵一，ブラック・アラン，“音声合成研究も協調と競争の時代に：The Blizzard Challenge，”音響誌，vol.62, no.6, pp.466–472, 2006.
[54] Z. Handley and M.-J. Hamel, “Establishing a methodology for benchmarking speech synthesis for computer-assisted language learning,” Language Learning & Technology, vol.9, no.3, pp.99–120, 2005.
[55] T. Pellegrini, A. Costa1, and I. Trancoso1, “Less errors with TTS? A dictation experiment with foreign language learners,” Proc. INTERSPEECH, USB-MEORY, 2012.
[56] 峯松信明，中村新芽，鈴木雅之，平野宏子，中川千恵子，中村則子，田川恭識，廣瀬啓吉，橋本浩弥，“日本語アクセント・イントネーションの教育・学習を支援するオンラインインフラストラクチャの構築とその評価，”信学論（D），vol.J96-D, no.10, 2013（掲載決定）.
[57] 河原英紀，“音声分析合成技術の動向，”音響誌，vol.67, no.1, pp.40–45, 2011.
[58] 久保理恵子，山田玲子，河原英紀，“STRAIGHT 分析合成音声を用いた/r/-/l/音聴取訓練，”音響春季講論集，1-8-22, pp.383–384, 1998.
[59] S. Kato, G. Short, N. Minematsu, C. Tsurutani, and K. Hirose, “Comparison of native and non-native evaluations of the naturalness of Japanese words with prosody modified through voice morphing,” Proc. SLaTE, 2011.
[60] C. Tsurutani and S. Ishihara, “Naturalness judgement of prosodic variation of Japanese utterances with prosody modified stimuli,” Proc. INTERSPEECH, 2012.
[61] K. Hirose, F. Gendrin, and N. Minematsu, “A pro-

nunciation training system for Japanese lexical accents with corrective feedback in learner's voice," Proc. EUROSPEECH, pp.3149–3152, 2003.
[62] T. Kawahara, H. Wang, Y. Tsubota, and M. Dantsuji, "English and Japanese CALL systems developed at Kyoto university," Proc. APSIPA ASC, pp.804–810, 2010.
[63] K. Imoto, Y. Tsubota, A. Raux, T. Kawahara, and M. Dantsuji, "Modeling and automatic detection of English sentence stress for computer-assisted English prosody learning system," Proc. ICSLP, pp.749–752, 2002.
[64] 井本和範，坪田　康，河原達也，壇辻正剛，"英語韻律発音学習支援システムのための英語文強勢のモデル化と自動検出，" 音響誌，vol.59, no.4, pp.183–191, 2003.
[65] A. Raux and T. Kawahara, "Automatic intelligibility assessment and diagnosis of critical pronunciation errors for computer-assisted pronunciation learning," Proc. ICSLP, pp.737–740, 2002.
[66] Y. Tsubota, M. Dantsuji, and T. Kawahara, "Practical use of English pronunciation system for Japanese students in the CALL classroom," Proc. INTERSPEECH, pp.1689–1692, 2004.
[67] M. Eskenazi, A. Kennedy, C. Ketchum, R. Olszewski, and G. Pelton, "The NativeAccent pronunciation tutor: measuring success in the real world," Proc. SLaTE, 2007.
[68] D. Luo, N. Minematsu, Y. Yamauchi, and K. Hirose, "Analysis and comparison of automatic language proficiency assessment between shadowed sentences and read sentences," Proc. SLaTE, 2009.
[69] S. Lee, H. Noh, J. Lee, K. Lee, and G. Lee, "POSTECH approaches for dialog-based english conversation tutoring," Proc. APSIPA ASC, pp.794–803, 2010.
[70] M. Raab, R. Gruhn, and E. Noeth, "Non-native speech databases," Proc. IEEE-ASRU, 2007.
[71] 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及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，vol.27, no.3, pp.259–272, 2004.
[72] N. Minematsu, K. Okabe, K. Ogaki, and K. Hirose, "Measurement of objective intelligibility of Japanese accented English using ERJ database," Proc. INTERSPEECH, pp.1481–1484, 2011.
[73] D. Crystal, English as a global language, Cambridge University Press, New York, 1995.
[74] J. Jenkins, The phonology of English as an international language, Oxford University Press, New York, 2000.

（平成 24 年 10 月 2 日受付，12 月 26 日再受付）

河原　達也　（正員）

1987 京大・工・情報工学卒．1989 同大大学院修士課程了．1990 同博士後期課程退学．同年京都大学工学部助手．1995 同助教授．1998 同大学情報学研究科助教授．2003 同大学学術情報メディアセンター教授，現在に至る．この間，1995～1996 米国・ベル研究所客員研究員．1998～2006 ATR 客員研究員．1999～2004 国立国語研究所非常勤研究員．2006～情報通信研究機構短時間研究員・招へい専門員．音声言語処理，特に音声認識及び対話システムに関する研究に従事．京大博士（工学）．1997 年度日本音響学会粟屋潔学術奨励賞，2000 年度情報処理学会坂井記念特別賞，2011 年度情報処理学会喜安記念業績賞，2012 年度科学技術分野の文部科学大臣表彰科学技術賞を受賞．IEEE SPS Speech TC 委員，IEEE ASRU 2007 General Chair，INTERSPEECH 2010 Tutorial Chair，IEEE ICASSP 2012 Local Arrangement Chair，言語処理学会理事，情報処理学会音声言語情報処理研究会主査を歴任．日本音響学会，情報処理学会各代議員．人工知能学会，言語処理学会，IEEE，ISCA，APSIPA 各会員．

峯松　信明　（正員）

1990 東大・工・電気卒．1995 同大大学院工学系研究科電子工学専攻博士課程了．博士（工学）．1995 豊橋技術科学大学情報工学系助手．2000 東京大学大学院工学系研究科情報工学専攻准教授．2012 東京大学大学院工学系研究科電気系工学専攻教授．この間，2002～2003 スウェーデン王立工科大学客員研究員．音声科学から音声工学に至るまで幅広い観点から音声コミュニケーションに関する研究に従事．2007 信号処理学会最優秀論文賞．2007 人工知能学会大会優秀賞．Speech Prosody 2004 Secretary，INTERSPEECH 2010 Secretary，L2WS 2010 Co-organizer，音声研究会幹事を歴任．IEEE，ISCA，IPA，SLaTE，音響学会，情報処理学会，人工知能学会，音声学会，音声言語医学会，発達心理学会，外国語教育メディア学会，日本語教育学会各会員．