社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

音声情報処理技術を用いた計算機援用型外国語教育・学習

峯松　信明†

† 東京大学大学院工学系研究科，〒 113–8656 東京都文京区本郷 7–3–1
E-mail: †mine@gavo.t.u-tokyo.ac.jp

あらまし　音声分析・合成・認識の要素技術を用いて外国語の音声教育・学習を支援する研究は，古くは 1980 年代に遡る。1990 年代に数理統計的な枠組みにより音声認識・合成の技術は進展し，その恩恵を受ける形で CALL（Computer Assisted Language Learning，計算機援用型言語学習）に関する研究も進展し，様々な CALL ソフトが市場に出回るようになった。従来，音声認識技術を用いた発音の評価（スコアリング）や発音誤りを検出するものが多かったが，最近では，合成音声をモデル発話として用いたリーディング／シャドーイングが教室で行なわれるなど，音声合成利用も活発になってきた。一昔と比べ，実応用を意識した研究・開発も多くなってきたが，現在でも基礎研究は続けられており，音声認識・合成同様，機械学習を用いた方法論が最近の流行りである。本稿ではまず，これらの技術的枠組みを概観する。その後，計算機援用型の外国語教育・学習について筆者の私見を述べ，更に，筆者らの研究グループが行なっている幾つかの研究・システム開発例を紹介する。
キーワード　音声分析・合成・認識，外国語学習，音声教育，音韻的特徴，韻律的特徴，CALL，機械学習

Computer assisted foreign language teaching and learning by using speech technologies

Nobuaki MINEMATSU†

† Grad. School of Engineering, The University of Tokyo, 7–3–1, Hongo, Bunkyo-ku, Tokyo, 113–8656 Japan
E-mail: †mine@gavo.t.u-tokyo.ac.jp

Abstract　Early works aiming at technical support of teaching and learning pronunciation of foreign language can be found in the 1980s, where technologies for speech analysis, synthesis, and recognition were examined. In the 1990s, with the help of statistical frameworks, speech recognition and synthesis technologies were greatly improved and CALL (Computer Assisted Language Learning) studies benefited much from these improvements. Since then, various CALL systems can be found in the market. Many of them had functions of scoring pronunciation and/or detecting pronunciation errors using ASR technologies. These days, not a small number of teachers use synthesized speech as model utterances in reading and shadowing practices. Using these advanced technologies, many examples of developing CALL applications are found in technical papers but fundamental studies are also still being done. Similarly to studies of speech recognition and synthesis, researchers' attention is paid mainly to machine learning. In this paper, these CALL technologies are reviewed. Then, personal opinions on CALL are given and our recent research and development of CALL is introduced.
Key words　Speech analysis, synthesis and recognition, foreign language learning, pronunciation teaching, segmental features, prosodic features, CALL, machine learning

1. はじめに

外国語を学ぶ場合「読む，書く，聞く，話す」の四技能を培う必要がある。母語を獲得する場合は当然「聞く，話す」能力がまず獲得され，やがて文字に遭遇し「読む，書く」ようになる。音声言語と文字言語を比べると，情報伝達や知識獲得の効率・正確さは後者が高いため，外国語を学ぶ場合も，「読む，書く」教育が優先されることが多い。日本における英語教育，留学生向けの日本語教育，いずれも音声教育に割く時間は十分ではないが，これは日本特有の現象，という訳ではない。

このような場合，学習者の自律的な独習に期待することになる。しかし「読む，聞く」能力が，学習者一人で対象言語音声を聞き，話すことで身に付くことは稀であり，「教師の代わり・延長」としの計算機システム（CALL, Computer Assisted Language Learning）が期待されるようになった [1]。

音声分析・合成・認識の各種要素技術を使って模擬教師作成を試みることになるが，1990 年代には当時の音声認識技術を用いて，教師同様，機械に発音誤りを指摘させたり，発音のス

コア付けを行なわせる研究が始まる。モデル発声と学習者発声の Dynamic Time Warping (DTW) に基づく検討が行なわれ [2], [3]，その後，HMM（Hidden Markov Model）に代表される数理統計的な枠組みが音声認識に導入されるに至り，母語話者 HMM と学習者発声を照合することで計算される尤度スコアをどのように利用すべきか，という検討が始まる [4], [5]。

従来 CALL と言えば，音声認識技術の応用として捉えられることが多かったが，最近では音声合成技術の利用も盛んに行なわれるようになった。学習者に聞かせるモデル発話を，母語話者による自然発声から，合成音に置き換えようという動きである。現在市販されているテキスト音声合成の出力は，母語話者が聞くと不自然な箇所に気づくが，その言語を学ぶ学習者が聞くと，恐らく「自分より上手い」と思うに違いない。実際，リーディングやシャドーイング時の呈示音声として使う教師は少なくない [6], [7]。このような合成音の品質向上以外にも，母語話者に依頼して収録する時間的，経済的コストを考えると，合成音の手軽さも語学教師には魅力的に映っている。

STRAIGHT [8] に代表される高品質な音声分析・再合成技術も（音声合成の一モジュールとして使われるだけでなく），発音能力が十分上達した後の学習者音声の模擬 [9] や，学習者の知覚特性を研究する際の，聴取実験用刺激音声作成など [10], [11]，多様な目的で使われている。

音声認識，合成，分析技術を用いて CALL 用の技術やシステムを精緻化する場合，当然，学習者音声コーパスが必要となる。筆者が中心となって構築した English Read by Japanese (ERJ) database [12] や，Japanese Read by Foreigner (JRF) database [13] がその例である。ERJ に関しては，コンパクトセット（男女 800 発声）を定義し，それを日本人と会話したことがない米語母語話者に聴取させ，彼らの聞き取りの様子を調査するなど（日本人英語はどう聞き取られてしまっているのか），データベースの拡張が行なわれている [14]。

筆者は京都大学の河原と協力し，音声情報処理技術を用いた外国語学習支援に関して，国際会議で tutorial を行ない [15], [16]，また，電子情報通信学会論文誌に解説論文を発表している [17]。同様のサーベイ発表・論文としては [18] もある。[17] では，音声認識，合成，分析技術としての CALL 応用，更には外国語音声データベースについてまとめており，詳細はそちらをご覧戴きたい。本稿ではまず [17] の抜粋を示し，次に [17] には含めなかった CALL 研究に対する私見を述べる。更に，筆者の研究室で現在進行中のプロジェクト [19], [20] について紹介する。

なお，音声技術を用いた CALL 研究の多くは，海外の論文誌としては Speech Communication, Computer and Language, IEEE Trans. Audio, Speech and Language Processing, of Speech が代表的である。国際会議としては ISCA INTERSPPEECH, IEEE ICASSP，更には ISCA SIG である SLaTE (Speech and Language Technology in Education) workshop も情報源として有用である。

2. 発声の音韻的側面を評価する技術

外国語の発声を評価する場合，対象が音韻的側面なのか，韻律的側面なのかで，利用される音響特徴や技術が異なる。ここでは，音韻的特徴を評価する場合の CALL について述べる。なお誌面のスペースを考慮して，本節と次節では，[17] で引用した文献は本稿末尾には引用していない。[17] を参照して戴きたい。

発声の音韻的側面を評価する場合，基本的に音声認識技術を使うことになるが，当然相違点もある。音声認識は，未知の音声入力（正確にはその特徴量系列）X に対して，発話内容 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り，以下のように書き換えられる。

$$\arg\max_W p(W|X) = \arg\max_W p(W) * p(X|W) \qquad (1)$$

これに対して発音を評価する場合，発話内容 W は既知の上で，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。

- 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。発音誤りを考慮したセグメンテーションの場合は下記の誤り検出と同時に行なうことになる。

- 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，誤り検出と評定についてその概要を述べる。

2.1 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。音素挿入，置換，脱落など様々な誤りが含まれる可能性があるが，ここでは，その学習者の母語や学習歴を考慮して，これらの誤りが必要十分に含まれた音素ネットワーク文法を用意し，音声認識を実行することで，W' を探索する。上記のように可能な誤りパターンを用意して，尤度 $p(X|W')$ を計算するのは生成モデルに基づくアプローチと言える。これに対して，入力のある音声区間が音素 w よりも音素 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。

2.2 評　　定

外国語の発音を評定する際に 2 つの考え方がある。1 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，尤度 $p(X|W;\lambda_G)$ を計算すればよい。しかし，音声認識の尤度 $p(X|W)$ は，話者変動や雑音等の影響も受けるので，絶対値をそのまま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尤度と日本人話者モデルによる尤度の比を求めることが検討されている。

発音評定のもう 1 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する。発音評価では，この事後確率を GOP (Goodness Of Pronunciation) スコアと呼ぶ。

尤度比にしろ，GOP にしろ，人間の評定と必ずしも合致するとは限らない。最終的には，線形回帰や，SVR などを導入して，両者の写像が学習される。

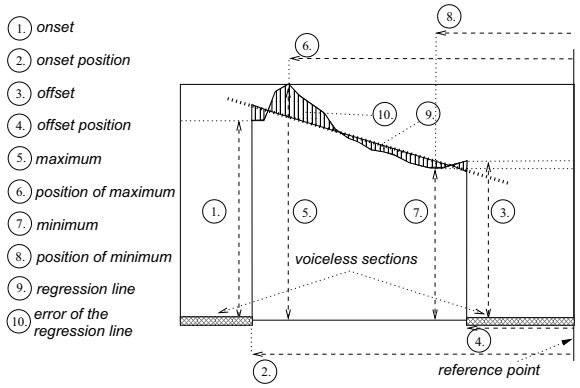

図 1　様々なF_0 素性 [21]

3.　発声の韻律的側面を評価する技術

韻律的側面を評価する場合，音韻的側面に対する音声認識技術のような，確立された技術体系が存在しないため，イントネーション，アクセント，リズムなど，韻律を構成する各側面毎に様々な技術が応用されている。基本的には，正解となる（統計）パターンと学習者の（入力）パターンとを照合し，スコア化する点においては音韻的側面と類似している。しかし，正解となるパターンを用意せずに評価する場合もある。これは，入力パターンと教師の評価スコアの組みが十分に与えられれば，前者から後者へ直接回帰する問題として捉えるアプローチである。例えば**図 1** に示すような様々な韻律素性を用いて SVR で教師スコアを予測している。

4.　発音学習に対して思うこと

4.1　Native-sounding or intelligible enough?

外国語学習，特に発音学習において native-sounding な発音と intelligible enough な発音のどちらを目標とすべきか，という議論が古くからある。筆者は大学時代，英語劇の舞台の上で発音を学び，発音を教えていた経験を持つ。Neil Simon の God's favorite の主役を任された時は，400 ものセリフを必死になって覚えた。当然のことながら，native-sounding な発音を目指した。ステージ・ボイスを出すために腹式呼吸をマスターし，より共鳴した（響く）英語らしい声を出すために喉を広げ（喉を落とし）(注1)，日本語にない調音制御（口型制御）が無意識的にできるよう，毎日口周りの筋肉訓練に勤しんだ(注2)。

CALL システムの発音スコアリングや発音誤り検出は，通常，母語話者の音声と比較して結果を出すこととなり，(特に母語話者モデルの尤度を用いた場合は）native-sounding な発音を目指す方々向けの支援ソフトと位置づけることもできる。当然，母語話者モデルから得られるスコアに対して閾値処理するなど，判定を甘くし「多少ずれていてもよい」との判定を出すことも可能であるが，母語話者との近さを基に議論していることは変わらない。しかし上記のような経験を持つ筆者は，それなりの地道な努力が必要であると考えており，果たして英語学習者の何割が，本気でそれを目指しているのかは懐疑的である。

Intelligible enough な発音を目指す発音教育は，native のような発音から少しずれた発音も許す教育と同値であるかどうかは議論の余地があるが，学習対象の言語が英語である場合，更に厄介な問題が存在する。それは，英語が国際語である，という事実からくる問題である。

4.2　国際語としての英語とその発音

学習対象が非国際語の場合，学習者がその言語を使って話しかける相手は，殆どがその言語を母語とする話者であろう。日本人以外を相手にして，日本語を話す機会はあまりない。しかし国際語である英語を話す場合，聞き手が英語の母語話者であることは少ない。これは，英語が国際語である所以である。

英語の利用者は世界に約 15 億人いると言われるが，母語として話す話者が約 1/4，公用語として英語を話す話者（インド，フィリピン，ナイジェリアなど）が約 1/4，そして，外国語として話す話者が約 1/2 である。母語として話す場合も，例えば英国訛り，米国訛りの英語であり，公用語の場合も然り，である。もちろん，外国語として話す場合に母語の影響が出るのは当然である。例えば，ある学習者が米国訛りを身に付けようと努力したところで，話し相手のインド人は 100 年後も，(筆者にとっては聞き取り難い）インド英語を話し続ける。それが彼らの公用語だからである。彼らに米国訛りを強制しても，それは受入れられない。訛りは彼らの identity だからである。同様に，英国人に米国訛りを身に付けるよう願っても，無駄である。

このような国際語としての英語を的確に捉えた用語として，World Englishes という言葉がある [22]。あるものが国際化される場合，それは各地の文化，風土によって多様化される。多様化を拒否すれば，国際化されることはない。英語も同様，発音・文法・語選択・綴り・談話戦略に至る様々な面での多様化は必然的結果と言える。国際社会における英語コミュニケーションでは，英語の多様性を積極的に許容する姿勢が必要となる [22]。この様子は，国連中継を見れば一目瞭然である。

さて,「みんな違ってそれでいい」であれば,「俺の英語」を直す必要があるのか？という哲学的な問いが浮上する。Native-sounding を目指すのでもなく，intelligible enough を目指すのでもなく「何もしなくてよいのか？」と。

筆者は [14] にて，日本人と会話したことがない米国人に，日本人大学生が読み上げた音声を電話越しで呈示し，彼らがどう聞き取るのかを調査した。一文毎に話者も内容も変わるため，前後の文脈が使えず，母語話者音声でも聞き取り率は約 90%であったが，日本人大学生の場合は約 50%であった。この読み上げ音声は，読んだ本人にとっては「正しく読めた」音声であり，本人はこれで伝わるものと思っている音声であるが，実際には 50%の単語しか聞き取ってもらえていない。この調査は自然な環境での日本人英語の聞き取り調査とはなっていないが,「何もしなければ」何が起こるのかを示す興味深い結果である。最低限，intelligible enough な発音が必要，ということになる。しかし，オバマ大統領と安倍総理の英語は，一般の日本人にとっては安倍総理の英語の方が聞きやすいだろう。つまり intelligible かどうかは，その発音が母語話者発音に近いかどうか，よりも，誰が聞いているのか，聞き手に大きく依存する [23]。

結局「国際語である英語，その発音を学ぶ場合，何を学ぶべきなのか」という振り出しに戻る。Native-sounding にする必要はないにしろ，intelligible enough な発音は，聞き手に依存し，一つの発音形式に特定できない。世界の誰もが intelligible であると認める発音形式があるのかどうか，筆者はそこまで知識がない。市販される発音教材の CD には通常，母語話者音声が入っている事実を考えれば，米国訛り・英国訛りとは異なる，

(注1)：最近では，英語喉と言われている。
(注2)：日本語は調音的には省エネ・節エネ言語であると筆者は思っている。

(あ はアクセント核)

図 2　フレージングとポージングに基づく韻律指導

地球人類がお墨付を与えた intelligible enough な発音形式は，まだ確立されていないのだろう(注3)。

4.3　非国際語学習の支援と国際語学習の支援

上記の問いに対する現在の筆者の私見を述べる。筆者は，非国際語を学習する場合の支援策と，国際語を学習する場合の支援策とを分けるべきだと考えている。

例えば非国際語の日本語を学ぶ学習者が目指すべき発音はどういうものか？それは，聞き手の多くを占める日本人が「聞き取りやすい」と感じる発音である。この場合，当然 the most intelligible な発音は，native-sounding な発音である。では，完全に native な発音でないと意味がないかと言われれば，母語話者でも方言による発音差があるように「native な発音から外れてもよいが，日本人にとって intelligible enough な発音」を目指すべきであり，そこに近づくコツの指導が必要になる。逆に言えば，どこは外れても，日本人は鈍感なのか，である。

では，国際語である英語の場合はどうであろうか？ある形式の発音を身につけても，それを intelligible enough と感じる聞き手もいれば，そうでない聞き手もいる。全人類が一つの発音形式を身に付けることは，国際化の特性を考えれば夢物語であり，結局，英語の多様性を大前提とした支援が必要となる。世界には約 15 億人の英語利用者がいる。英語の多様性として発音の多様性（訛り）を考える。通常，日本訛りや，ニューヨーク訛りなど，国，地方，都市を単位として訛りを考えることが多いが，本来訛りとは，その話者の言語背景に依存する。厳密には各個人の言語背景は皆異なっており，その意味において，約 15 億種類の英語発音が存在する，と言える。この多様性を大前提とした支援を考える必要がある。

上記のような戦略の元，日本語学習支援，英語学習支援として筆者の研究グループが行なっている研究例を以下，紹介する。

5.　日本語発音学習支援

5.1　発音学習のコツとそれに基づくシステム開発

[24] によれば，「日本人にとって聞き取り易い発音」にするコツとは，文の意味を考えて区切り（ポーズ）を置き，区切りから区切りまでに適切なイントネーションパターンを生成してフレーズとする（＝チャンキング）ことである（**図 2** 参照）。学習者はテキスト中，平仮名から漢字に移った時にポーズを置いてしまうことがあるが，不要なポーズが入ると聞きとり難くなる。実際に [24] の付属 CD にある，上記のコツを指導する前後の学習者発声を聞くと，変化の大きさに驚かされる。

図 2 では，句（フレーズ）を単位として「へ」の字型イントネーションパターンを付与しているだけであるが，アクセントによる局所的なピッチの上下まで考えると，このピッチパターンでは不十分である。一般にフレーズは複数のアクセント句から構成され，アクセント句には高々一つのアクセント核が存在する。その結果，一フレーズに複数のアクセント核が存在する。

日本語の音声教育では，アクセントを教えることは稀である。

(注3)：多数決でそれを求めれば，当然中国語訛りの英語発音となる。

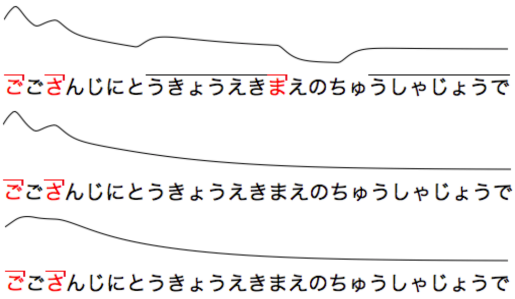

図 3　3 種類のピッチパターンと核位置表示

筆者が知る限りではこれは次の三点に起因する。1) そもそも日本語は方言性がアクセントに出ることが少なくなく，アクセントの間違いは方言性と捉えられ，コミュニケーションを大きく阻害することは少ない。2) 母語話者と言えども，任意の発声に対して各モーラを H/L で書き起こす（イントネーションによるピッチ変動を除去して，アクセントによる局所的なピッチ変動のみに着眼して H/L を振る）ことは困難であることが多い。日本語教師も例外ではなく，教えることが困難。3) 前後コンテキストによって頻繁に変わるアクセントを教えている時間がないし，そもそも，それを効率的に教える教材もない。その一方，学習者には声調言語を母語とする学習者も多く，例えば中国人学習者は教師の発声の各モーラが何声なのかを聞き分けながら聞いている。このように，教師よりも遥かに局所的なピッチ変動に敏感な耳をもつ学習者も多く，アクセントや，文脈によるその変形パターンを知りたがることも少なくない。

以上の現状を考慮し，「へ」の字のイントネーションパターンにアクセントの乗せ，両者が考慮されたピッチパターンを入力テキストの上に表示し，学習者が「日本人にとって聞き取りやすい」発声の習得を効率化するシステムを開発した。この時，フレーズ中の全てのアクセント核を付与して，それを発声させることは負担が大きい。より実践的な折衷案として [24] では，初級者向けに，「フレーズに最初に現れるアクセント核のみに注意を払い，その後の核は無視してよい(注4)」という指導戦略で臨んでいる。システム開発も，この戦略に従って行なった。

5.2　ピッチパターン描画

平仮名表記の直上に該当するピッチパターンを乗せる必要があるが，例えば，タノシム，という実際の発声のピッチパターンを乗せることは不適切である。各モーラの継続長は同一ではない。また，抽出誤りも避けられない。この場合，パターンを生成する数理モデルを用意し，そのモデル制約の下で，教師が示したいパターンの “イメージ” を描く必要がある。

本研究では，基本周波数パターン生成過程モデル [25] を用いた。基本周波数パターンをフレーズ成分（大局的な変化パターン）とアクセント成分（アクセントに伴う局所的な変化パターン）の足し合わせとして捉え，両成分を少数のパラメータで制御する。本モデルでは，アクセント成分に対応する制御パラメータが，アクセント核位置と直接対応がとれるため都合が良い。教師のイメージに沿ったパターンニングとなるよう，教師と協議しつつ各種パラメータの値を設定した。

全てのアクセント核を表示する上級者用と，フレーズ中の核を極力減らして表示する初級者用（2 種類）を用意している。実際の出力例を**図 3** に示す。

(注4)：誤った位置にアクセント核を付与するよりは，なだらかに下降するイントネーションとした方が誤りが目立ち難い。

“Please call Stella. Ask her to bring these things with her from the store: Six spoons of fresh snow peas, five thick slabs of blue cheese, and maybe a snack for her brother Bob. We also need a small plastic snake and a big toy frog for the kids. She can scoop these things into three red bags, and we will go meet her Wednesday at the train station.”

[pʰl̥iːz kʰɑːl stɛlʌ æsk hɚ ɾə bɹĩŋ ðiːz θĩŋz wɪθ hɛɹ fɪ̃ʌ̃m ðə stoːɹ sɪks spʊ̃ːnz əɣ fɪɛʃ snoʊ pʰiːʐ faːɣ θɪkˈ slæːbz ə blʊː ʃ͡fiːʐ ɛ̃n meɪbi ə snæ̰kˈ foɹ hɚ bɹʌðə bɑb wi ɑlso n̪ːd ə smɑːl pʰl̥æstɪk sneɪkˈ æ̃n ə bɪgˈ tʰɔɪ fɪog fʊ̃ɹ ðə kʰɪɪdʐ ʃĩ kʰĩn skụːp ðiːʐ θĩŋ̃ʐ ĩntʰụ θɪi ɹɛd bæːgz æ̃n wi wɪl miːt hɛɹ wɛ̃ntsdi æt ðə tɹeĩn steɪʃĩn]

図 4　読み上げ用パラグラフと IPA 書き起こし例

5.3　利 用 状 況

オンライン日本語アクセント辞書（Online Japanese Accent Dictionary）[19] の一部として本機能を提供している。基本的に音声合成器の裏側で動いている，アクセント核位置推定やアクセントパターン生成，イントネーションパターン生成を表舞台に出しているだけのシステムであるが，日本語教育史上初の「アクセント結合に対応した韻律指導教材」として広く利用されるに至っている。これまで読み上げ原稿のアクセント位置は，母語話者教師に付けてもらうしか方法がなかった学習者も少なくなく，既に海外 17 都市，国内 10 都市で講習会を開いている。

6.　英語発音学習支援

6.1　世界諸英語を前提とした支援

多様な発音（厳密な意味では約 15 億種類の発音）を前提にすると，学習者に提供すべき情報は「学習者の発音が母語話者とどう違うのか」ではなく，「学習者の発音が他者とどう違うのか，世界中の英語発音の中でどのように位置づけられるのか」に関する情報であると言える。現在，世界中の英語話者の音声データから，個人を単位とした世界英語発音分類（地図）の自動構築を検討している。このような地図ができれば，英語の現状を俯瞰することができ，また，自身の発音を客観的に捉えられ，更にインターネット上の英語音声コンテンツを地図とリンクすれば，世界英語ブラウザも構築できる。

6.2　Speech Accent Archive

Speech Accent Archive (SAA) [26] は，世界中の英語利用者に特定のパラグラフを朗読させ，その朗読音声とその IPA 書き起こし（補助記号を使った詳細な書き起こし）を提供しているコーパスである。パラグラフと書き起こし例を**図 4** に示す。現在世界中より収集した 1,700 名以上のデータを提供しており，これを用いて発音地図構築の技術的検討を行った。

ある集合の自動分類は，任意の二要素間の距離，即ち全要素に対する距離行列を計算することで可能となる。世界英語発音分類の場合，任意の二話者間の発音間距離を計測すれば良い。SAA は特定のパラグラフを読ませており，また，IPA 書き起こしも提供しているので，この目的には都合の良いコーパスである。二話者の IPA 書き起こし間の距離を計測すれば，距離行列は得られる。しかし，IPA 書き起こしは手間のかかる作業であり，特定のパラグラフを読ませた N 人の音声資料セットのみから，$N \times N$ の発音間距離行列を推定する技術が必要である。本研究では，IPA-based な二話者間距離を参照距離とし，当該二話者の音声データのみから，この参照距離を予測（回帰）する技術的枠組みを検討した。SAA は世界中からボランティアベースで音声を集めており，その録音環境は様々である（電話音声もある）。このような音声群から，発音だけに関する（年齢や性別などには依存しないように）二話者間距離を推定する。本研究では，構造的表象 [27] を応用する。

なお，SAA は発声中に言い直しがあっても訂正せず，そのまま IPA 書き起こしを行っている。本研究ではこれらの発声は手動で除外した。また，背景雑音レベルが非常に高いサンプル，語順を間違えて読み上げているサンプルも除外した。以下の検討は最終的に得られた 381 人の音声，${}_{381}C_2$=72,390 通りの発音間距離の推定を検討している。

6.3　IPA 書き起こしに基づく参照距離の算出

SAA に含まれる IPA 書き起こしは，補助記号も使われており，音声記号数は 153 種類であった。IPA 書き起こし間の距離は DTW により求めるが，この場合，153×153 の音声記号間距離行列が必要となる。熟練の音声学者 1 名に全音声記号を音声化してもらい（20 回／記号），これを使って HMM を構成した。次に，二つの HMM 間距離を状態間のバタチャリヤ距離の平均で定義し，音声記号距離行列を得た。これを用いて，任意の話者対（任意の IPA 書き起こし対）間の，IPA-based な発音間距離を DTW により求めた。

6.4　ベースラインシステム

構造的表象に基づく手法との比較のために，下記の参照距離予測を事前に行った。参照距離は，1)IPA を用いた手動書き起こし，2)DTW による自動距離計算により得られている。1) のプロセスを自動化できれば，参照距離の自動計算は可能である。入力音声を IPA 記号列へと変換するシステムは存在せず，ここでは，米語音素の音響モデル（HMM）を用いた連続音素認識器を代用した。これにより，全ての音声は米語音素系列に置き換わる。また 2) の DTW も，米語音素 HMM より計算した 43×43 の音素距離行列を用い，これを発音間距離とした。

音素認識率 100%の場合を想定して音素系列間距離を求めた。この場合，IPA 記号列を簡単な変換表により米語音素列へと変換した。IPA 記号は 153 種類もあり，これを 43 種類の音素へと変換すれば，様々が情報が消失する。音素列を用いた発音間距離と IPA を用いた参照発音間距離との相関は 0.88 であった。

次に実際の音素認識器を利用した。SAA より構築した monophone HMM を用い，入力音声以外の 380 発声から構成されたネットワーク文法(注5)を用いた。得られた音素正解率は 73.4%であった。各発声に対する音素系列と DTW を用いて，発音間距離を推定した。参照発音間距離との相関は僅か 0.31 であった。

6.5　構造表象と SVR による参照発音間距離予測

構造表象を使う場合，SAA パラグラフ読み上げ音声を分布系列，即ち HMM 化する必要がある。ここでは，パラグラフを 9 つに分割し（文や句），各々に対して音素数 ×3 だけの状態を有する HMM を構成した。まず，利用した音声資料全体から構築される不特定話者 HMM を 9 区間に対して各々構築した。これを UBM（Universal Background Model）として利用する。この UBM を当該発声で MLLR 適応して（クラス数 32），各話者の各発声を HMM 化した。

各話者に対して 9 個の HMM が構築され，各 HMM 毎に，3 状態を音素（相当）の単位と仮定して 3 状態単位でバタチャリ

(注5)：各単語の発音バリエーションを 380 音声から取得し，単語単位で構成したネットワーク文法。

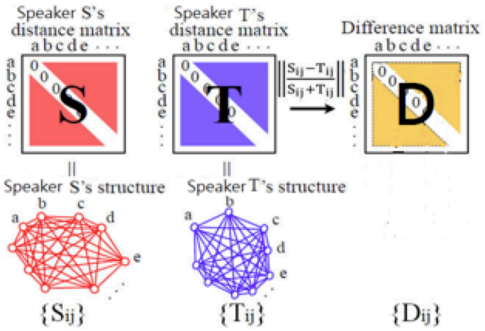

図 5　差行列 $\{D_{ij}\}$ の計算

ヤ距離の平均値 $d(i,j)$ を求めた。

$$d(i,j) = \sqrt{\frac{BD_1(i,j) + BD_2(i,j) + BD_3(i,j)}{3}}$$

i，j は音素インデックスであり，$BD_n(i,j)$ は音素 i，j 間の第 n 状態でのバタチャリヤ距離である。各話者に対して 9 種類の（音素単位の）距離行列が得られる。なお距離行列の上三角部分の要素数は全部で 2,804 であった。これがある話者の発音様態を表現する特徴ベクトル次元数である。

二話者 S，T の距離行列 $\{S_{ij}\}$，$\{T_{ij}\}$ に対して，その差を表現する差行列 $\{D_{ij}\}$ を以下のように求めた（**図 5**）。

$$D_{ij}(S,T) = \left| \frac{S_{ij} - T_{ij}}{S_{ij} + T_{ij}} \right|$$

二話者間の発音間差異も 2,804 次元の特徴量となる。

得られた発音間差異特徴量を使って，参照距離に対する回帰モデルを学習する。ここでは，識別的な回帰である SVR を用いた。実装系としては LIBSVM の ϵ-SVR を使用した。なお，カーネル関数は RBF である。

72,390 通りの話者対を二分し，2-fold な交差検定を行なった。参照発音間距離と，構造表象及び SVR によって予測された発音間距離との相関は 0.81 となった。100%の音素認識装置には及ばなかったが，実際の音素認識器（発音誤り検出器）を用いた実装よりも遥かに高い相関値を示した。現在，更なる精度向上を狙うと共に，1) パラグラフ音声収集の拡大化，2) 発音地図が構築された場合の教育利用，などについて検討を進めている。

7. ま と め

[17] に記述した，音声情報処理技術を用いた外国語学習支援の枠組みについて，その一部を述べると共に，[17] には示さなかった私見を述べ，筆者の研究グループの研究例を紹介した。本稿が，CALL 研究の更なる推進に繋がれば幸いである。

文　献

[1] *The history of computer assisted language learning web exhibition*, http://www.eurocall-languages.org/resources/history_of_call.pdf

[2] S. Hiller, *et al.*, "SPELL: An automated system for computer-aided pronunciation teaching," *Speech Communication*, 13, 463–473, 1993.

[3] H. Hamada *et al.*, "Automatic evaluation of English pronunciation based on speech recognition techniques," *IEICE Trans.*, E76-D, 3, 352–359, 1993.

[4] Y. Kim *et al.*, "Automatic pronunciation scoring of specific phone segments for language instruction," *Proc. EUROSPEECH*, 645–648, 1997.

[5] H. Franco *et al.*, "Automatic pronunciation scoring for language instruction," *Proc. ICASSP*, 1471–1474, 1997.

[6] 山内他，"合成音声と自然音声による音声モデルの違いがシャドーイング・パフォーマンスに与える影響"，外国語メディア学会関東支部研究大会発表要項，10–11, 2012.

[7] T. Pellegrini, *et al.* "Less errors with TTS? A dictation experiment with foreign language learners," Proc. INTERSPEECH, USB-MEORY, 2012.

[8] 河原，"音声分析合成技術の動向"，日本音響学会誌，67，1，40–45，2011.

[9] K. Hirose, *et al.*, "A pronunciation training system for Japanese lexical accents with corrective feedback in learner's voice," Proc. EUROSPEECH, 3149–3152, 2003.

[10] R. Kubo et al., "/r/-/l/ perception training using synthetic speech generated by STRAIGHT algorithm," Proc. Spring Meeting of Acoust. Soc. Japan, 1-8-22, pp.383–384, 1998 (in Japanese)

[11] C. Tsurutani *et al.*, "Naturalness Judgement of Prosodic Variation of Japanese Utterances with Prosody Modified Stimuli," Proc. INTERSPEECH, USB-MEMORY, 2012

[12] 峯松他，"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，27，3，259–272，2004.

[13] K. Nishina *et al.*, "Speech database construction for Japanese as second language learning," Proc. SNLP-Oriental COCOSDA, 187–191, 2002.

[14] N. Minematsu, *et al.*, "Measurement of objective intelligibility of Japanese accented English using ERJ database," Proc. INTERSPEECH, 1481–1484, 2011.

[15] T. Kawahara and N. Minematsu, "Tutorial on Computer-assissted language learning based on speech technologies," Tutorial session of APSIPA Conference, 2011.

[16] T. Kawahara and N. Minematsu, "Tutorial on Computer-assissted language learning (CALL) systems," Tutorial session of INTERSPEECH, 2012.

[17] 河原達也，峯松信明，"音声情報処理技術を用いた外国語学習支援"，電子情報通信学会論文誌，J96-D，7，1549–1565，2013.

[18] M. Eskenazi, "An overview of spoken language technology for education," Speech Communication, 51, 832–844, 2009.

[19] 峯松他，"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"，電子情報通信学会論文誌 D，J96-D，10，2496–2508，2013.

[20] H.-P. Shen *et al.* "Automatic pronunciation clustering using a world English archive and pronunciation structure analysis," Proc. ASRU, 2013 (to appear)

[21] A. Maier, *et al.* "A language-independent feature set for the automatic evaluation of prosody," Proc. INTERSPEECH, 600–603, 2009.

[22] B. Kachru, *et al.*, *The handbook of world Englishes*, Wiley-Blackwell, 2006.

[23] M. Pinet, *et al.*, "Second-language experience and speech-in-noise recognition: the role of L2 experience in the talker-listener accent interaction," Proc. SLaTE, CD-ROM, 2010.

[24] 中川他，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

[25] 藤崎他，"日本語単語アクセントの基本周波数パターンとその生成機構のモデル"，日本音響学会論文誌，27，9，445–453，1971.

[26] Speech Accent Archive, http://accent.gmu.edu

[27] 峯松他，"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～"，電子情報通信学会論文誌，J94-D，1，12–26 (2011)