メディア教育研究　第９巻　第１号
Journal of Multimedia Education Research 2012, Vol.9, No.1, S19－S28

特集（招待論文）

音声認識技術の発音学習への応用

新田　恒雄[1) 2)]・入部　百合絵[2)]

近年，外国語学習者を対象とするCALL（Computer Assisted Language Learning）教材の開発が盛んに行われている。しかし，学習者の調音動作の誤りを正確に指摘できる教材はまだ開発されていない。本論文では，著者らが開発中の「教師と学習者の調音の違いを簡単に理解でき，さらに正しい調音への矯正方法を直感的に読み取ることのできる，英語発音学習システム」の調音編を中心に紹介する。このシステムは，学習者の発音を①調音部位や調音様式の違いを軸とした平面上に，表示する「発音マップシステム」と，②発声器官の動作をMRI画像から学習して教師と利用者の調音動作の違いをアニメーションで示す「調音アニメ生成システム」を提供する。①，②とも，学習者の音声から多層ニューラルネットワーク（Multi-Layer Neural network; MLN）を用いて調音特徴を抽出し，①では発音マップ上へ座標変換することにより，また②では対応する発音器官（声道）上の座標に変換することで，調音動作をリアルタイムに表示し発音矯正を支援する。本文ではこれまでに行った評価実験結果についても紹介する。

キーワード
音声認識，発音学習，調音特徴抽出，発音マップ，調音アニメ生成

1. はじめに

コンピュータ性能とネット速度の向上に伴い，ICTを導入して語学教育を支援するシステムの研究開発が盛んである。またユーザインタフェース（以下UI）についても，タッチ入力，音声入出力，さらに種々のセンサーを組込んだ端末が市販され，次世代UIとしてのマルチモーダル対話（以下MMI（Multi-Modal Interaction）と呼ぶ）への移行が始まっている（新田・桂田，2012）。こうした進展の中で，本特集号のテーマである音声認識を利用し，発音訓練・発音矯正を行う，CAPT（Computer-Assisted Pronunciation Training）の研究開発が盛んになっている（Eskenazi，2009）。また，2012年６月にはStockholmに於いて “Automatic Detection of Errors in Pronunciation Training” に関する国際ワークショップが開催され，この分野の研究者が集い活発な討議が行われた（IS ADEPT，2012）。

音声認識は，連続する波形から分節化された記号（音素・単語・単語列（文））への変換を目的としているが，一連の処理過程の中では，波形から特徴系列を抽出し，続いて音素単位に分類する。ディクテーション（口述筆記）ソフトは，音素分類と同時に単語の連鎖確率（豊橋に接続する語は，市・駅・技科大・圏・格助詞などに限られる）を利用して単語列を確定している（鹿野・河原・山本・伊藤・武田，2001）。このうち特徴系列は，スペクトル時系列に由来するものを用いることが多いが，最近は調音や音素の素性をベクトル表現して特徴系列とするものが提案されている（Nurul・Kawashima・Nitta，2009; Nitta・Onoda・Kimura・Iribe・Katsurada，2010）。特に調音特徴系列は，発音器官への動作指令を表現しているため，調音特徴抽出結果から発音エラーを呈示したり，教師と生徒の調音の差異を示すなど，発音学習支援への応用が期待される（Iribe・Manosavanh・Katsurada・Hayashi・Zyu・Nitta，2012a; Iribe・Mori・Katsurada・Nitta，2012b）。

以下では，２．で発音学習システムの現状と課題を述べた後，３．で調音特徴，調音特徴抽出器，調音運動のワンモデルに基づく音声認識・合成について説明する。続いて４．では調音特徴抽出に基づく発音マップシステム，および調音アニメ生成システムについて述べる。また，評価実験結果についても紹介する。

2. 発音学習システムの現状と課題

近年，外国語習得を目的に様々なCALL教材が導入され，また学習者が音声認識技術を利用して発音を習得するCAPTシステムも開発されている。一方，英会話教室

[1)] 早稲田大学
[2)] 豊橋技術科学大学

や教育機関では，教師が模範となる音声を学習者に聴かせるとともに，舌や口唇など調音器官の動作（以下，調音動作と呼ぶ）をface-to-faceで伝えながら，正確に発音できるよう学習者を指導してきた。第二言語（L2）の習得は開始年齢によって大きな差があると言われているが，導入時期に正しい発音を身に着けることができるか否かは，聴取能力向上を含めてその後の言語習得に大きな影響を持つ。

これまでに学習者と正しい発音に対する調音を視覚的に評価することを目指すCAPTシステムが幾つか開発されている（Jo・Kawahara・Doshita・Dantsuji，2000；菊池・島崎・境，2010）。Sonic Printは，学習者音声のフォルマント周波数をリアルタイムに分析してF1-F2平面上に表示し，自身の発音が正しい母音領域に収まるよう，繰り返し練習することを推奨している（㈱アルカディア，2007）。ここでF1，F2とは，振幅スペクトル上の二つの主要なピークを与える周波数（下から第１，第２フォルマント周波数）を指し，これらを横軸と縦軸とした図をF1-F2平面と呼ぶ。

F1-F2平面上の各母音領域は口腔断面図に描かれる調音と対応するが，具体的な調音情報（舌の位置，口の開閉）は表示されたプロット位置から読み取る必要がある。このため音響音声学の知識を持たない学習者には，どのように調音動作を改善すればよいか分かりづらい。また，教師と学習者の音声波形を比較表示する教材が多いが，一般の学習者にとって波形の違いは分かっても，具体的にどのように調音を矯正すれば良いのか理解するのは困難である。

一方，音声認識を利用するCAPTシステムでは，予め用意した教材（単語・単語列）に対する音素認識結果をもとに，発音誤りを示すことが多い（坪田・壇辻・河原，2000）。即ち，正解音素と誤り音素との調音動作の違いを定型の単語や文を音声記号で示すのみであるため，どの調音がどのように誤り，具体的にどのように矯正すればよいのか分かりづらい。正しい発音を身につけるには，少人数の会話教室で教師が実施しているように，学習者の調音誤りをリアルタイムに，的確に指示して矯正することが重要である。

3. 調音特徴の抽出

調音特徴を音声認識へ利用する研究は，まず発声器官とその動作をモデル化する研究が行われた（白井，2010）。ここでは近年盛んな，音声学上の知見を基に調音動作の諸属性を抽出する技術を中心に，著者らが進めている調音特徴抽出器と調音運動のワンモデルに基づく音声認識・合成について説明する。

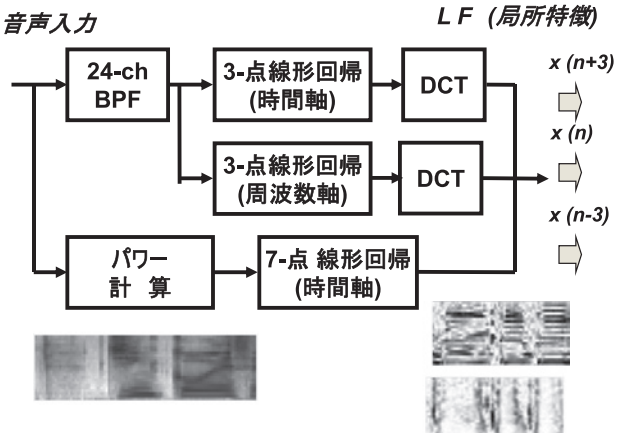

図１　局所特徴抽出器

3.1 調音特徴

調音特徴（Articulatory Feature; AF）は，単音（phone）分類に用いられる調音様式（破裂音，摩擦音，破擦音，鼻音，半母音など）と調音部位（口唇，歯茎，口蓋，咽喉などの位置（子音の場合）や，舌の最も盛り上る位置と口の開閉度（母音））の諸属性から構成される特徴量である。従って，調音特徴が得られると学習者の調音動作を推定することができ，IPA（International Phonetic Alphabet）（IPA，2012）の母音チャートや子音チャートへプロットすることが可能になる。

現在使用している調音特徴セットは，IPAから英語と日本語に関する部分（次元数：28）を取り出したもので，英語音素数46（/sil/を含む），日本語音素数13（英語と重複分を除く）を対象にしている。

3.2 局所特徴抽出器

学習者端末から入力された音声は，16kHzでサンプリング後，25msのハミング窓で10ms毎に高速フーリエ変換（FFT）処理される。この結果はパワースペクトルの形で積分され，中心周波数を聴感覚に近似したメル尺度間隔の24チャンネルBPF（Band Pass Filter）出力にまとめられる。なおメル尺度は，音の高さに対する人間の知覚特性から得た尺度で，ほぼ対数周波数軸に近いが高域と低域では非線形性を示す。ここまでが分析処理である（図１参照）。

続いて局所特徴（Local Feature; LF）の抽出が行われる。パワースペクトルの時系列が構成する曲面は，多様体として見ると時間と周波数方向の局所的な微分要素で表現できる（微分多様体）。そこで，BPF出力を３×３の局所特徴とし，主要な二つの軸，時間軸と周波数軸に沿って３点の線形回帰（Linear Regression; LR）演算を施し，微分特徴としてのLFに変換する（新田・井上・正井・松浦，2000）。二つのLFは各24次元であるが（図１参照），次に，離散余弦変換（Discrete Cosine Transform; DCT）によって，半分の12次元に圧縮され

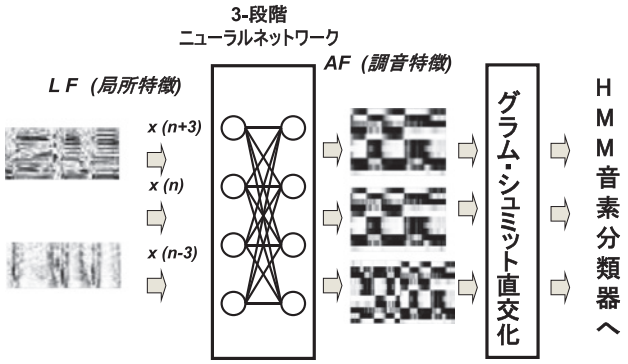

図２　調音特徴抽出部の構成

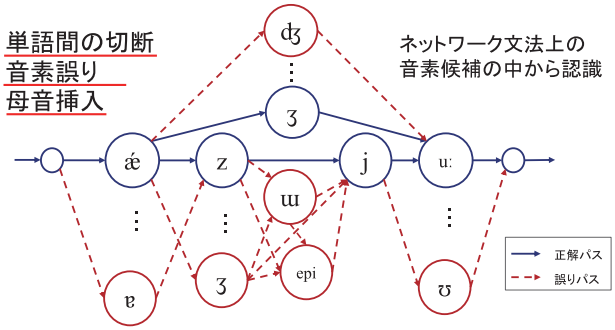

図３　日本人の発音誤りを含むネットワーク文法
（例：“as you”）

る。これに対数パワー成分の微分を加えた25次元の特徴を以後は局所特徴LFと呼ぶ。

3.3　調音特徴抽出器（Fukuda・Nitta，2004）

局所特徴LFから大域的特徴である調音特徴（Articulatory Feature; AF）を抽出する。AFの抽出は，２段のMLN（Multi-Layer Neural network）と直交化処理により行われる（図２参照）。１段目は10msec間隔で７フレームのLF（このうちt-3，t，t+3の３フレームを使用）をMLNへ入力して，AFの候補列を抽出する。続いて，音素境界で生じる分類誤りを補正するため，AF系列に加えてその速度成分（ΔAF），および加速度成分（ΔΔAF）を調音運動への制約として，２段目のMLNに加える。AF系列は，図２の濃淡グラフ出力に見られるように，離散的情報（segmental information）の性質を持ち，音素列と一対一に対応することから，原初的言語情報と見做すことができる。音声に対応するAF系列を利用するだけの場合は，この系列データを使用して発音マップや調音アニメを表示できる。AF系列信号からさらに，単語や文中の発音誤りを検出したい場合は，単語・文を音素単位でグラフ表現したネットワークに通して誤りを検出する（図３参照）。各音素は調音特徴系列を学習データに，隠れマルコフモデル（Hidden Markov Model;

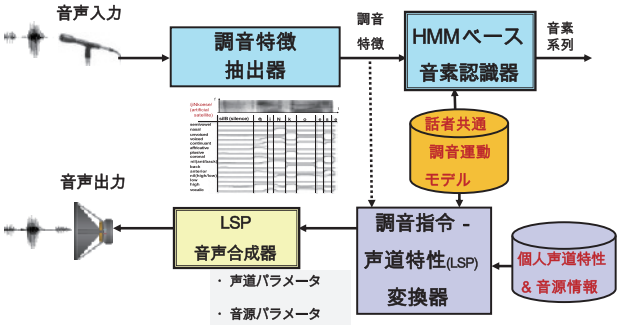

図４　調音運動のワンモデル音声認識・合成

HMM）を使用して調音運動モデルを形成している。HMMは，入力データ（時間的に前後の複数系列から成る）内での無相関性を要請するため，特徴系列間の独立性（直交性）を与えるGram-Schmidtの直交化処理を施したものをAF入力系列とする。

3.4　調音運動のワンモデル音声認識・合成（Nitta et al，2010）

調音特徴（AF）を学習データとして構成したHMMは，音素ごとの調音運動の振舞いを確率的に表現する。図４の上半分に示す音声認識エンジンでは，調音特徴系列がHMMに入力され，ここで話者に共通の調音運動モデルを参照しながら入力系列を処理する。また，図の下半分の音声合成エンジンでは，音声認識エンジンと共通の調音運動モデルからなるHMMを音素単位に結合しつつ，HMMの各状態から読み出されたAF系列を，話者固有のLSP（Line Spectral Pair；もしくはLine Spectral Frequencies；LSFと呼ばれる）で表現した声道パラメータ系列に変換する。合成音声は，LSPデジタルフィルタで構成される合成器に，LSP系列と音源信号を入力して生成される。音源信号は，HMMから音源符号を読みだし，PSOLA（Pitch Synchronous Overlap and Add）方式（Hamon・Moulines・Charpentier，1989）を用いて，ピッチの音調曲線（pitch contour：現在は音声から抽出したものを使用）に沿った制御を行う。この方式は，共通のモデルを使用することから，「調音運動のワンモデル音声認識・合成」方式と呼んでいる。

4.　調音特徴に基づく発音学習システム

この節では，調音特徴抽出に基づく発音学習システムについて説明した後，発音マップシステムと調音アニメ生成システムについて述べる。また，各システムの評価実験結果についても紹介する。

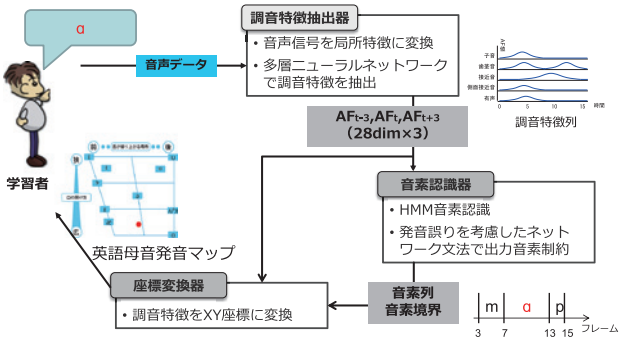

図５　発音マップシステム全体図

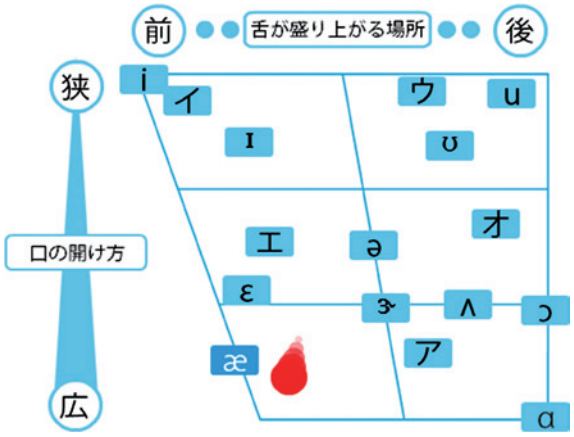

図６　母音発音マップ（/æ/の学習）

4.1　発音マップシステムの概要（森・入部・桂田・新田，2012）

図５に発音マップシステムの全体図を示す。システムは学習者の発声を検知すると，調音特徴抽出部で10ms毎に28次元の調音特徴を抽出する。母音発音マップでは，抽出された調音特徴から，母音に関する10次元の特徴ベクトルを元に，座標変換器で２次元平面上のX，Y座標に変換する。子音発音マップでも同様に，子音に関係する14次元の特徴ベクトルが座標に変換される。この際，HMMから得た音素継続時間を用いて，調音特徴毎の平均値を算出しプロットしている。

(1)　母音発音マップ

母音発音マップの画面例を図６に示す。発音マップでは，IPA母音チャートを模した梯形図に発音記号が配置され，口唇の開き具合を示すスケール（縦軸），舌の盛り上がる位置を示すスケール（横軸）をもとに，ユーザの調音状態を赤い光点で示している。また，調音の軌跡を薄い赤円で表現することで，矯正結果を学習者にフィードバックできるようにしている。なお，調音特徴から発音マップへプロットする際には，母音発音マップが台形状のため，図のようにX-Y座標を台形座標にスケール変換する。

(2)　子音発音マップ

子音発音マップの例を図７，図８に示す。子音発音マップでは，等分割した矩形図に発音記号が配置され，調音部位を示すスケール（横軸）と調音様式を示すスケール（縦軸）を用いて，ユーザの発音を赤い光点でプロットする。マップ上には，日本人の発音で置換誤りが多く見られる英語子音の組み合せを表示している（坪田他，2000）。即ち，日本人が誤り易い音素対に焦点を当て，発音マップ上でそれらの調音動作の違いを明確にしつつ，矯正できることを目指している。横軸と縦軸には，マップ上の音素間の差異を特徴付ける素性を配している。図７に/b/と/v/の発音マップの例を示した。図に示すように，音素間で調音部位と調音様式が異なる場合はそれらを縦軸と横軸にとる。一方，図８に示す/r//l//ɾ/の発音マップの例のように，音素間の差異が調音部位か調音様式のいずれかに限定される場合は，一次元でのマップ表示となる。

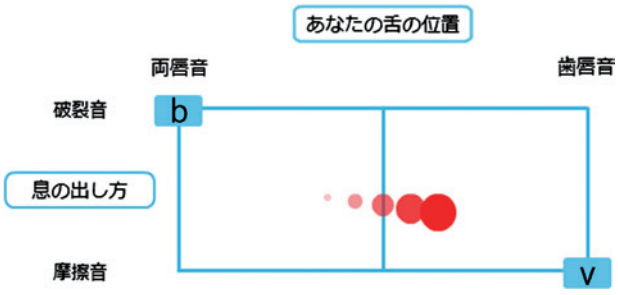

図７　子音発音マップ(横軸：調音部位，縦軸：調音様式)

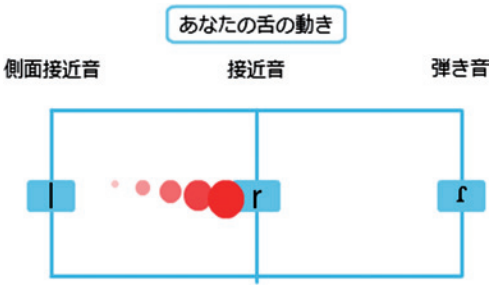

図８　子音発音マップ（横軸：調音様式）

発音マップを用いた矯正は，調音部位と調音様式を独立に評価することができるため，例えば「舌の位置は正しいが摩擦音がうまく発音できていない」というように，似通った音素群の調音上の差異を，より具体的かつ詳細に示すことができる。

(3)　評価実験と結果

(a)　調音特徴の抽出精度

プロット精度に大きな影響を及ぼす調音特徴抽出精度を評価する。後述の学習セットで学習したMLNを用いて，評価セットの音声から抽出した調音特徴28次元を対象に，次式のAF抽出精度（AF-Correct Rate;

AFCR）を求めた。

$$AFCR = \frac{\text{正しく抽出できた特徴数}}{\text{フレーム数} \times 28} \times 100[\%] \quad (1)$$

ここで正しく抽出できた特徴とは，MLNの教師信号（AF値）を“+”（実際には1.0の実数値）と与えたAFに対して0.5以上を出力した場合を，また“−”（実際には0.0の実数値）と与えたAFに対しては0.5未満の値を出力した場合を指す。

図９と図10に，各々母音と子音の一部に対する音素別調音特徴抽出精度を示す。このうち，英語全体では平均93%，日本語全体では平均96%の抽出精度が得られている。

(b)　母音発音マップに対するプロット精度

英語母音発音マップは，学習者の発声をIPA母音チャート上にプロットし，マップ上の発音記号との相対的な位置から発音動作の差異を視覚的に教示することを目標にしている。従って，ネイティブ英語発音に近い発音がなされた場合は，対応する発音記号近傍にプロットされることが理想である。また，日本語と英語の調音の違いを理解するため，誤って（もしくは故意に）日本語の母音を発声した場合は，日本語母音の発音記号の近くにプロットされなければならない。

そこで，ネイティブ英語話者の母音がプロットされた座標と，各母音の正解座標（IPAチャート上の座標）について次式に示す距離を算出し，プロット精度を評価した。

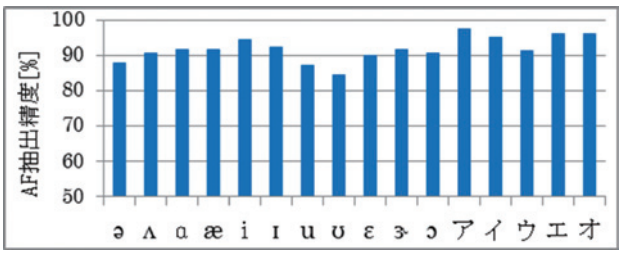

図９　母音に対する調音特徴抽出精度

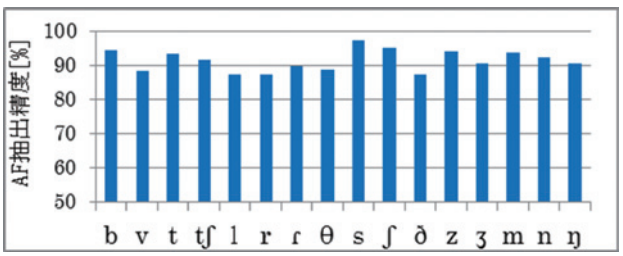

図10　子音（一部）に対する調音特徴抽出精度

$$Dist_p = \frac{1}{n}\sum_{i=1}^{n}\sqrt{(x_i - x_{corr})^2 + (y_i - y_{corr})^2} \quad (2)$$

ここで，x_i，y_iは音素pの音素継続長毎のX，Y座標を指す。また，x_{corr}，y_{corr}は音素pの正解座標のX，Y座標で，MLNの教師信号を座標変換器に入力して得た座標である。nは評価データ中に母音pが出現する回数を示す。なお(3)の評価実験で用いた音声試料は，英語母音の調音特徴を抽出するMLN学習にTIMIT（Texas Instruments Massachusetts Institute of Technology）コーパス（Garofolo・Lamel・Fisher・Fiscus・Pallett・Dahlgren・Zue, 1993）の2,600文（男性話者325名）を，日本語母音の調音特徴を抽出するMLNの学習には日本語話し言葉コーパス（Corpus of Spontaneous Japanese; CSJ）（人間文化研究機構国立国語研究所，2010）の22時間分（男性話者92名）を使用した。プロット精度の評価には，学習に用いないTIMIT896文（男性話者112名）とCSJコーパス2.5時間分（男性話者10名）を使用している。

図11に母音のプロット精度を示す。値が小さいほど正確にプロットできている。図９で抽出精度の高かった/i/，/ɪ/，/イ/は正解座標とほぼ同じ座標にプロットされており，抽出精度の低かった/u/，/ʊ/，/ウ/は正解座標からの距離が大きいことから，調音特徴の抽出精度がプロット精度に大きく影響することが確認できる。

さらに話者によるプロット精度のばらつきを確認するため，TIMITの各英語話者を出身地情報別に８つのグループに分け，またCSJの日本語10話者に対してもプロットを行った。話者グループ毎の平均座標を台形のマップに適応させる変換処理を行った後，プロットした結果を図12に示す。図中の破線領域はIPAの母音図に相当する領域である。幾つかの母音は正解座標から離れた位置にプロットされたが，話者グループ内のばらつきは小さい。さらに，調音の異なる英語と日本語では，発音マップ上のプロット位置がほぼ正確に分離できていることが確認できた。

(c)　子音発音マップに対するプロット正解率

子音発音マップは，音素グループ毎に誤り易い調音

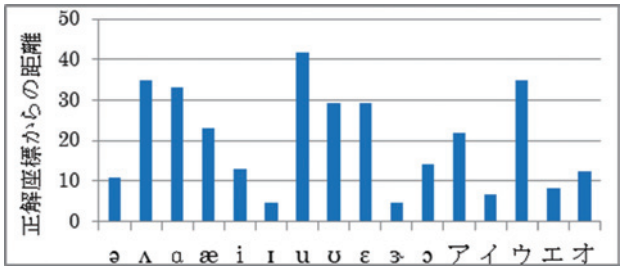

図11　母音に対するプロット精度

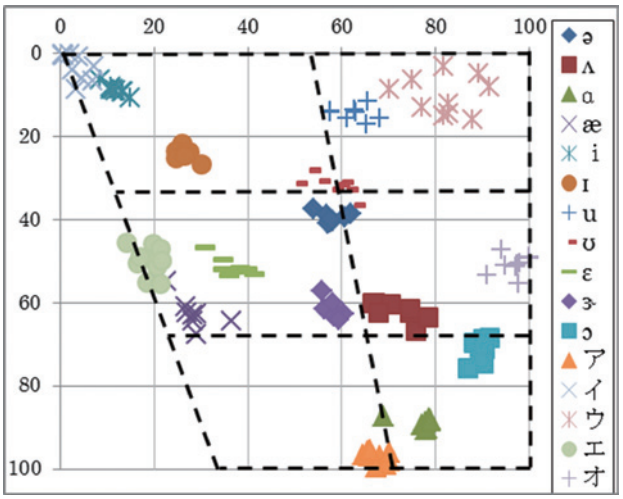

図12　母音に対する話者グループ毎のプロット例

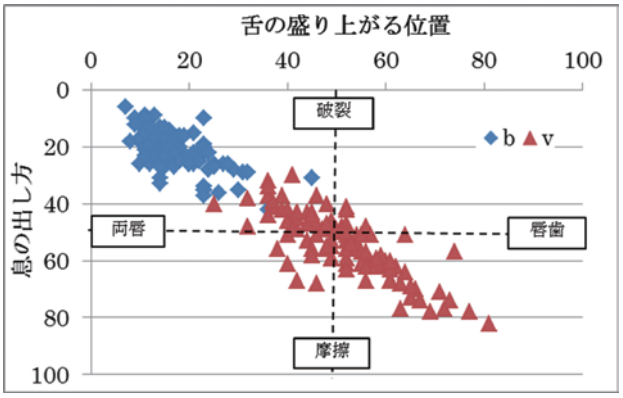

図13　/b/と/v/に対するプロット例

部位と調音様式を組み合わせた平面上に，学習者の音声をプロットし，マップ上の発音記号との相対的な位置から発音動作の差異を視覚的に教示する。従って，ネイティブ話者の英語発音に近い場合は，母音マップ同様，発音記号と同じ座標領域にプロットされることが理想である。そこでネイティブ話者の子音発音が正しくプロットされた場合を正解として，プロット正解率を評価した。ただし，母音発音マップと異なり，図７，図８に示したように，各音素として許容できる領域がマップ上の線分で明確に区切られるため，正解音素と同じ領域にプロットされた場合を正解とした。例えば図13では，/b/のラベルが付与された区間に対応する座標が，両唇の領域（左上の領域）なら正解とし，正解数を評価データ中の/b/全ての発話区間数で除した値をプロット正解率とした。

子音発音マップにプロットした際の正解率を図14に示す。/v/と/θ/を除いて，50%以上のプロット正解率が得られた。また，/θ/と/ð/はそれぞれ/s/と/ʃ/，/z/と/ʒ/に比べてプロット正解率が低い。図10からこれらの音素は，調音特徴抽出精度も低いことが分かる。

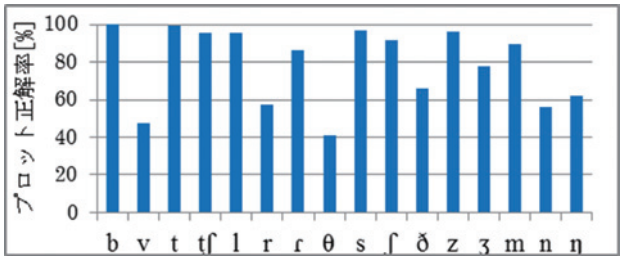

図14　子音プロット正解率

調音特徴抽出器の歯音に関する特徴抽出性能が低いことが，プロット正解率の低さの原因になっていると考えられる。有声・無声共に歯摩擦音は歯茎摩擦音に比べて弱いことが知られており（竹林・斉藤，2008），歯音特徴を正確に抽出できるように，特徴抽出器を改良する必要がある。さらに，/v/のプロット精度は/b/に比べて低いため，摩擦音の調音部位に対する抽出性能が改善されると，プロット正解率が向上すると考えられる。

4.2　調音アニメ生成システム（Nitta・Manosavanh・Iribe・Katsudara・Hayashi・Zyu，2012）

音声から調音動作を抽出しCGアニメーション（以降，調音アニメと呼ぶ）を表示できるなら，学習者は自身の調音誤りを視覚的に知ることができる。さらに，教師の音声から抽出した調音アニメと比較すれば，調音上の差異が分かり，発音器官の何処をどのように動作させると発音矯正できるかを指示することもできる。

高精度な調音アニメを実現するには，3. で述べた音声からのAF抽出のほか，AF系列からの調音アニメ生成を精確に行う必要がある。ここでは，調音アニメ生成システムの概要を説明すると共に，MRI動画像をMLNの教師データとしたAF系列からの調音アニメ生成と評価実験について述べる。

(1)　システムの構成

調音アニメ生成システムの構成を図15に示す。音声はAF系列に変換された後MLNに入力され，発音器官中の調音に関わる座標値に逐次変換される。MLNは，MRI動画像から次に説明する座標値を教師データとし，対応する入力データに動画中の音声をAF系列に変換したものを使用して学習した。学習に利用可能なMRI動画像は未だ少ないことから，画像フレームから抽出する座標値（特徴点）の数は，(2)で述べるように６点に制限した。

各特徴点のデータ系列は，続いて調音アニメのグラフィックス生成部に送られ，音声に対応する発音器官の調音動作が自動的に生成される。

(2)　MRI動画像を教師データとした調音の特徴点抽出

MRI動画像の初期フレームを使用して，調音に関わる部位（舌，口蓋，口唇，下顎など）の輪郭に沿って特徴

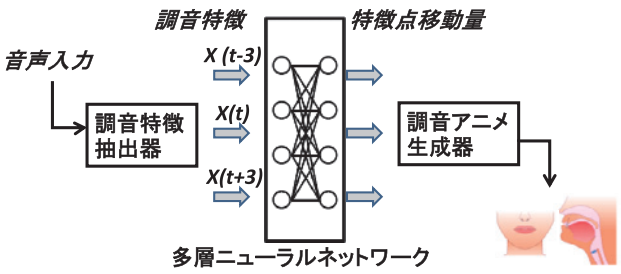

図15　調音アニメ生成システムの構成

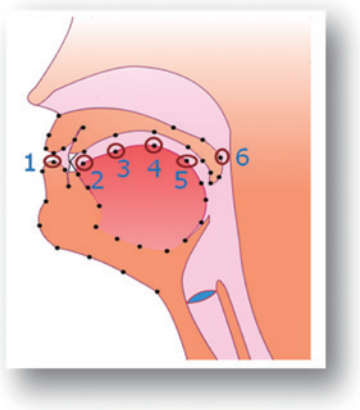

図16　調音に関わる特徴点

点を与えた（図16参照）。このうち，図中○で囲んだ６部位の特徴点に対して座標値を学習する。座標値は，オプティカルフローを計算することによって取得し，フレーム毎の特徴点の移動量を計算して，結果をMLNの教師信号として与える。

選定した６個の特徴点は，移動量が比較的大きく，かつ調音動作を教える上で重要な部位である。MRIデータ収録には多大な費用がかかるため（明瞭な画像を得るには，同一音声を数十から100を越える回数収録し，これらを同期加算する必要がある）（Takemoto・Honda・Masaki・Shimada・Fujimoto，2006），少量のMRIデータでMLNを学習する必要があるが，それには学習空間を制約することが有効になる。

さらに，調音動作アニメの可視化精度を向上させるため，個々の音素に対して調音上必須と考えられるアンカーポイントを設定すると共に，ポイントに対応する図16の特徴点座標を修正して，MLNの教師信号とした。音素毎のアンカーポイント対応を表１に示す。

導入したアンカーポイントは，発話の際に両唇や舌により声道途中に閉鎖や狭めを作るなど，調音上重要な部位に相当する。例えば，/p/のアンカーポイントは両唇になるため，特徴点１（下唇）が上唇と一致するように特徴点１の座標を修正する。なお，アンカーポイント以外の特徴点に関しては，オプティカルフローから得た移動量を教師信号に含める。MRI動画像から調音動作を観

表１　音素とアンカーポイントの対応表

音　素	アンカーポイント	特徴点
p, b, m	両唇	1
f, v	唇歯	
θ, ð, t_J, d_J	歯	2
t, d, z, ɾ, l, ɹ, dz_J, ts_J, s	歯茎	
ʃ, ʃ_J, tʃ, dʒ, ʒ, ʒ_J	後部歯茎	3
k^j, g^j	硬口蓋	
k, g	軟口蓋	4
ŋ	軟口蓋（後方）	5
m, n, ŋ以外	口蓋垂	6

表２　実験データ

学習単語	MRIデータ： －英語ネイティブ話者：男女２名（102単語） －日本人話者：男女２名（75単語）
評価単語	英語ネイティブ話者：102単語 日本人話者：75単語 （評価はLeave-One-Out Cross-Validation）

表３　特徴点別に見た相関係数

特徴点	評価データ			
	ネイティブ話者		日本人話者	
	アンカーポイント修正なし	アンカーポイント修正あり	アンカーポイント修正なし	アンカーポイント修正あり
1	0.68	0.80	0.67	0.81
2	0.66	0.80	0.68	0.79
3	0.73	0.85	0.71	0.81
4	0.71	0.82	0.72	0.80
5	0.67	0.77	0.68	0.75
6	0.72	0.83	0.70	0.81
平均	0.70	0.81（+16%）	0.69	0.79（+14%）

察する場合，多数の画像を同期加算する結果，調音部位が明確でないことが少なくない。このため，重要な調音部位にアンカーポイントを設定し，MLNを学習する手法は，音声に対応する調音動作アニメをより明瞭化でき，発音訓練に有用と考えられる。

(3)　調音アニメの生成

調音アニメはActionscript3.0で実装した。プログラム

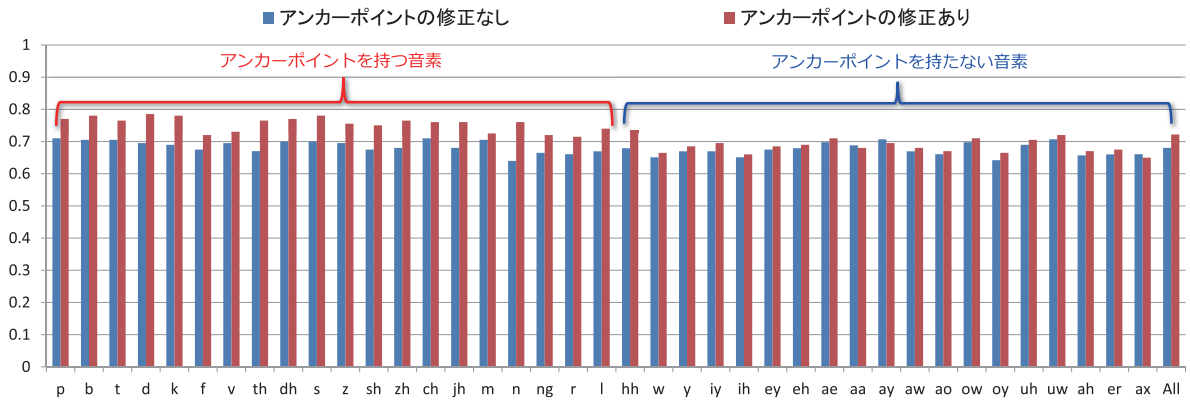

図17　音素毎に算出した調音アニメとMRIとの相関係数（英語ネイティブ話者）

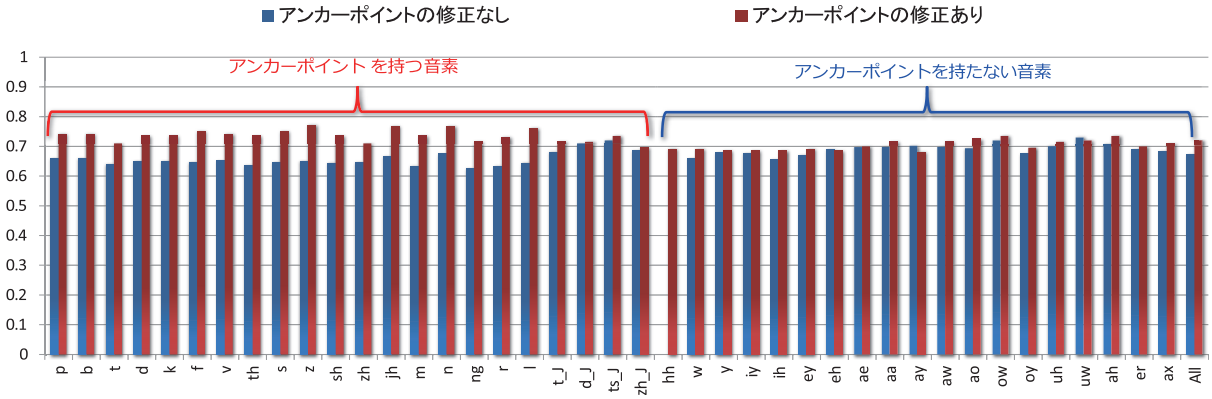

図18　音素毎に算出した調音アニメとMRIとの相関係数（日本人話者）

はFlash Player10上，もしくはFlash Playerプラグインが有効なウェブブラウザ上で動作する。なお，発音器官の特徴点座標から滑らかなアニメを生成するために，中間値フィルタとスプライン曲線補間法を用いている。

(4)　評価実験と結果

生成された調音アニメの精度を評価するため，試作システムが音声から生成した調音アニメとMRI動画像の各特徴点における相関係数を算出した。表２に実験に用いたMRI動画像データを示す（林・中村・エリクソン・朱・定延，2010)。表３は英語母語話者の音声から生成した場合と，日本語母語話者の英語音声について，生成された調音アニメとMRI動画像との相関係数（平均値）を示したものである。アンカーポイントを導入してMLNを学習したことにより，ネイティブ話者，日本人話者音声とも，相関係数が各々16%と14%向上した。図17，図18は，表３の内容を音素毎の相関係数から示したものである。図から，アンカーポイントを付与した全ての音素について，相関係数が改善されていることが分かる。以上の実験結果から，MLNの教師信号にアンカーポイント修正を経た特徴点を使用することにより，調音アニメの可視化精度を向上できることが明らかになった。

5.　まとめ

音声信号から調音特徴を自動抽出する技術を用いて，英語の発音訓練を支援するシステムについて紹介した。この中で，(a)学習者の調音動作をIPAチャート上にリアルタイムで表示する「英語発音マップ」と，(b)学習者の発音器官の動きを教師のそれと比較し差異を示す「調音アニメ」の二つの機能を説明した。また，英語母語話者と日本語母語話者の発話を用いた評価実験により，母音発音マップのプロット精度の確認と，子音発音マップのプロット正解率の現状を紹介した。調音アニメの評価では，MRI動画像との比較結果を示し，アンカーポイントの導入によって比較的高い相関結果が得られることを示した。今後，発音マップのプロット精度と調音アニメの描画精度を一層向上するとともに，調音編以外の機能（韻律編，短文編など）についても順次整備したい。また，英語発音教育に携わる教師の方たちと，効果的な教示方法など発音学習システムの改善を進めたいと考えている。紹介したシステムは，音声―調音変換をベースにしており，日本語や他の言語にも比較的容易に応用できる。音声研究者やL2言語（第二言語）の研究者の方たちと，この面でも徐々に適用範囲を拡張していきたい。

謝辞

MRIデータをご提供頂いた神戸大学の朱春躍教授および林良子准教授に深く感謝いたします。

引用文献

Christian Hamon, Eric Moulines, and Francis Charpentier（1989）. A diphone synthesis system based on time-domein prosodic modifications of speech Proc. ICASSP 1989, 238–241

Chul-Ho Jo, Tatsuya Kawahara, Shuji Doshita, Masataka Dantshuji（2000）. Japanese Pronunciation Instruction System Using Speech Recognition Methods IEICE transactions on information and systems, **E83–D**(11), 1960–1968.

大学共同利用機関法人　人間文化研究機構国立国語研究所（2010）. 日本語話し言葉コーパス〈http://www.ninjal.ac.jp/csj/〉（2012年10月15日）

林良子，中村淳子，ドナエリクソン，朱春躍，定延利之（2011）. MRI動画による英語音声の調音動態の観察―日本人英語学習者との比較―　第25回日本音声学会全国大会予稿集，91–96.

Hironori Takemoto, Kiyoshi Honda, Shinobu Masaki, Yasuhiro Shimada, and Ichiro Fujimoto（2006）. Measurement of temporal changes in vocal tract area function from 3D cine-MRI data Journal of the Acoustical Society of America, **119**(2), 1037–1049.

Huda Mohammad Nurul, Hiroaki Kawashima, Tsuneo Nitta（2009）. Distinctive Phonetic Feature（DPF）Extraction Based on MLNs and Inhibition/Enhancement Network IEICE transactions on information and systems, **E92–D**(4), 671–680.

IPA（International Phonetic Association）（2012）. IPA vowels〈http://www.langsci.ucl.ac.uk/ipa/vowels.html〉（2012年10月15日）

IS ADEPT（International Symposium on Automatic Detection on Errors in Pronunciation Training）（2012）. ISADEPT-proceedings.pdf〈http://www.speech.kth.se/isadept/ISADEPT-proceedings.pdf〉（2012年10月18日）

John S. Garofolo, Lori F. Lamel, William M. Fisher, Jonathan G. Fiscus, David S. Pallett, Nancy L. Dahlgren, and Victor Zue（1993）. TIMIT Acoustic Phonetic Continuous Speech Corpus, Linguistic Data Consortium（Pennsylvania, USA）

株式会社アルカディア（2007）. 株式会社アルカディア Sonic Print〈http://www.arcadia.co.jp/SP/index.html〉（2012年10月15日）

菊地歌子，島崎のぞみ，境一三（2010）. 日本人フランス語学習者のための発音学習教材　電子情報通信学会技術研究報告SP，110（**452**），25–29.

Maxine Eskenazi（2009）. An overview of spoken language technology for education Speech Communication, **51**(10), 832–844.

森拓郎，入部百合絵，桂田浩一，新田恒雄（2012）. 調音特徴抽出に基づくIPAチャートへの英語発音リアルタイム表示　電子情報通信学会技術研究報告，SP2011–169，77–82.

新田恒雄，井上雄，正井康之，松浦博（2000）. 複合音響特徴平面に基づく音声認識のための局所特徴抽出法　電子情報通信学会論文誌，**J83–D–II**(11)，2341–234.

新田恒雄，桂田浩一（2012）. マルチモーダル対話システム基盤技術とその応用　電子情報通信学会誌，**95**(5)，446–451.

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

白井克彦編著（2010），音声言語処理の潮流　コロナ社

Takashi Fukuda, Tsuneo Nitta（2004）. Orthogonalized Distinctive Phonetic Feature Extraction for Noise-robust Automatic Speech Recognition IEICE transactions on information and systems, **E87–D**(5), 1110–1118.

竹林滋，斉藤弘子（2008）. 英語音声学入門　大修館書店

坪田康，壇辻正剛，河原達也（2000）. 日本人の誤りパターンの対判別を利用した英語発音教示システム，電子情報通信学会技術研究報告SP，125, 25–32.

Tsuneo Nitta, Silasak Manosavan, Yurie Iribe, Kouichi Katsurada, Ryoko Hayashi and Chunyue Zhu（2012）. Pronunciation Training by Extracting Articulatory Movement from Speech Proc. of IS ADEPT（International Symposium on Automatic Detection of Errors in Pronunciation Training），211–216.

Tsuneo Nitta, Takashi Onoda, Masashi Kimura, Yurie Iribe, Kouichi Katsurada（2010）. One-model speech recognition and synthesis based on articulatory movement HMMs Proc. Interspeech 2010, 2970–2973.

Yurie Iribe, Silasak Manosavanh, Kouichi Katsurada, Ryoko Hayashi, Chunyue Zhu and Tsuneo Nitta（2012a）. Improvement of Animated Articulatory Gesture Extracted from Speech for Pronunciation Training Proc. of ICASSP（IEEE International Conference on Acoustics, Speech, and Signal Processing）2012, 5133–5136.

Yurie Iribe, Takuro Mori, Kouichi Katsurada, Tsuneo Nitta（2012b）. Real-time Visualization of English Pronunciation on an IPA Chart Based on Articulatory Feature Extraction Proc of InterSpeech 2012, 1023–1026.

メディア教育研究　第９巻　第１号
Journal of Multimedia Education Research 2012, Vol.9, No.1, S19－S28

新田（にった）　恒雄（つねお）
1969年東北大学工学部電気工学科卒業。㈱東芝を経て1998年豊橋技術科学大学大学院工学研究科教授。現在，同大名誉教授および同大・早稲田大学客員教授。工学博士。情報処理学会フェロー。音声認識・合成，マルチモーダル対話，概念獲得の研究に従事。IEEE，電子情報通信学会，情報処理学会，人工知能学会，日本音響学会各会員。

入部（いりべ）　百合絵（ゆりえ）
2001年名古屋大学大学院人間情報学研究科修士課程修了。2004年名古屋大学大学院人間情報学研究科博士課程満期退学。博士（学術）。現在，豊橋技術科学大学情報メディア基盤センター助教。教育支援，ユーザインタフェースに関する研究に従事。情報処理学会，人工知能学会，教育システム情報学会各会員。

Applying Speech Recognition Technology to Pronunciation Training

Tsuneo Nitta[1) 2)] ・ Yurie Iribe[2)]

Typical CAPT systems evaluate pronunciation using speech recognition, however, they cannot indicate how the learner can correct his/her articulation. This paper describes a novel CAPT system based on articulatory feature extraction from learner's utterance. The proposed system has two novel functionalities of: (A) plotting a phone uttered by the learner on a pronunciation map with the two axes of articulation manner and articulation place in real time, and (B) visualizing animated articulatory gesture by highlighting the phone-specific part of an articulatory organ on a screen and comparing with the correctly pronounced gesture. In the system, a multi-layer neural network (MLN) is applied to convert learner's speech into (A) the IPA chart and/or (B) the coordinate of a vocal tract using MRI data.

Keywords
Pronunciation Training, Articulatory Feature Extraction, Articulatory Map, Animated Articulatory gesture

[1)] Waseda University
[2)] Toyohashi University of Technology