情報処理学会論文誌　Vol. 52　No. 12　3671–3681 (Dec. 2011)

# 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

峯松 信明[†1]　鎌田 圭[†1,*1]　朝川 智[†1,*2]
鈴木 雅之[†1]　牧野 武彦[†2]
西村 多寿子[†1]　広瀬 啓吉[†1]

外国語を習得する場合，母語干渉に起因する外国語訛りが頻繁に観測される．本研究では外国語発音の自動分析およびそれに基づく自動評定技術の構築を目指している．外国語発音を音響分析する場合，年齢・性別・体格に起因する音響バイアスを捨象したうえで分析することが望まれるが，本研究では音声の構造的表象を用いてこれを検討している．すでに先行研究で，構造的表象を用いた学習者分類および発音矯正度推定についての有効性が示されている．本研究では前者に対しては，音声学者による学習者分類との比較を通してその妥当性検証を，後者に対しては，より高精度に矯正度推定を行う手法を提案する．実験の結果，音声学者による分類とほぼ同等の分類結果を得ることができ，また，矯正度に関しても 7.4 ポイント精度向上を果たすことができた．

# Verification of Clustering Language Learners and Improvement of Estimating which Vowel to Correct Based on the Structural Representation of Speech

Nobuaki Minematsu,[†1] Kei Kamata,[†1,*1]
Satoshi Asakawa,[†1,*2] Masayuki Suzuki,[†1]
Takehiko Makino,[†2] Tazuko Nishimura[†1]
and Keikichi Hirose[†1]

When students learn a foreign language, due to language transfer, foreign accented pronunciations are often found. In this study, we aim at developing methods of automatic analysis of foreign accented pronunciations and automatic assessment of them. Acoustic analysis of the pronunciations should follow the process of removing acoustic feature biases from observations, which are extra-linguistic and caused by differences in age, gender and physical figure of students. This process is realized in this study by using the structural representation of speech. In our previous study, the effectiveness of using this representation to cluster students and estimate the need of correction for vowels was shown. In the current study, the validity of the clustering result is verified by comparing it to a phonetician's clustering of the same students and a new method is proposed to estimate the need of correction. Experimental results show that the automatic clustering performs reasonably similarly to the manual clustering and the new method outperforms the previous method by 7.4 points.

## 1. はじめに

近年，国際語である英語を用いたコミュニケーション能力のきわめて高い重要性が，社会的に広く認知されるようになった．たとえば，日系企業が国際的なビジネス展開を図るために，英語を社内公用語とする例がある[1)]．また，多くの国立大学において学部授業の一部英語化が計画されている．さらに，文部科学省は 2002 年度に「英語が使える日本人育成に向けた戦略構想」を策定し，2011 年度より外国語活動が小学 5・6 年生において必修化されている[2)]．小学校での試みでは「話し言葉」としての英語活動が行われることとなっており，「話す/聞く」能力の養成が求められている．外国語発音を習得する場合，母語干渉に起因する訛りが頻繁に観測される．これらを自動分析し，各学習者の発音習熟度を推定したり，発音のどの部分を矯正すべきなのかを提示したりすることは，外国語発音の効率的な習得につながる[3)]．本研究は音声情報処理技術を用いてこれらの高精度自動化を目指している．

外国語発音を自動分析する場合の問題点の 1 つは，同種の発音が音としては異なる音響事象として観測される点にある．同じような発音誤りを有する 2 学習者を考えた場合，彼らが異なる性別，年齢，体格を有する話者であれば，当然観測される音響事象は異なってくる．音声の音響的特徴は，発音能力だけでなく，上記の非言語的要因に容易に影響を受ける．そ

†1 東京大学
The University of Tokyo
†2 中央大学
Chuo University
*1 現在，JR 東日本旅客鉄道株式会社
Presently with East Japan Railway Company
*2 現在，ソニー株式会社
Presently with SONY Corporation

の結果，システム構築時の学習データと利用時の評価データとの間に音響的ミスマッチが生じてしまう．小学生にまで拡大する英語学習者の現状を考えると，性別，年齢，体格などに起因する声質の違いに頑健に動作する技術が必要であると考えられる．

従来，ミスマッチ問題に対する解決法として，多数話者を用いた母語話者音声の統計モデル（隠れマルコフモデル，HMM）を構築したり，学習者音声を常時参照し，音響モデルに話者適応をかける方法が検討されている[4),5)]．しかしこの方法では，適応用データに発音誤りが含まれるため，学習者・環境への適応だけでなく，発音習熟度に対しても適応がかかり，適応後の音響モデルを用いると，誤った発音を正しいと判断するという問題が生じる[6)]．さらに，適応用データがごく少量である場合，適応モデルによる発音評価精度の劣化も報告されている[7)]．そもそもミスマッチ問題は，非言語的要因による音響変動（たとえば話者の違い）も，音韻の発音の違いによる変動（たとえば/r/と/l/の違い）も，音響的にはスペクトル包絡の違いとして観測されるが，これらを分離して扱う技術が構築されていないことに起因する[8)]．

小学生の英語音声を母語話者（たとえば男性教師）モデルと比較する場合，母語話者モデルを子供化する（話者適応）解決方法は，次のような非技術的な問題を生じる．年齢や性別に起因する話者性を含めて他者の声を模倣する行為は声帯模写といわれるが，上記の 2 種類の変動が分離できない発声比較技術は，発音の良し悪しではなく，声帯模写の良し悪しを推定する枠組みと解釈すべきである．これを発音習熟度推定に応用する場合には，前処理として「声質合わせ」が必要となる．そもそも，英語教師の発声を模倣する行為は声帯模写の一形態なのだろうか？　それとも，英語教師の発声から話者性に起因する音響的側面を無視して（分離して）模倣しているのだろうか？　筆者らは以下の理由により，後者であると考える．英語教師や他者の発声を模倣する際に，声帯模写的になる（話者性の分離が難しい）例は自閉症者に見られるが[9)–11)]，この場合，音声コミュニケーションそのものが困難となることが多い（たとえば，母親の声しか理解困難になる[12)] など）．また，幼児の言語獲得は親の発声に対する模倣行為が基盤となるが，これは物理的な模倣ではない．相手の声の話者性を無視した模倣をする．このような事実に鑑みると⋆1，英語教師が行う発音評定の自動化を目的とした技術構築においても，非言語的な要因に起因する音響バイアスを分離して発音の様態をモデル化・表象し，これを学習者，教師間で比較する枠組みが必要であると考える[15),16)]．

⋆1 なお，動物の音声模倣は一般的に物理的模倣となる[13)]．類似した例として，動物は移調前後のメロディの同一性知覚が困難である（相対音感を有していない）[14)]．異なる 2 話者は通常，異なる音色バイアスおよび音高バイアスを有する．メロディの移調は通常，音高バイアスのみが異なる．動物はこれを無視することが難しい．

先行研究において筆者らは，非言語的要因に起因する静的な音響バイアスを除去したうえで外国語発声をモデル化し（音声の構造的表象），これを 2 話者間で比較する技術を提案した[8),15)]．そこでは，話者性に非依存な学習者発音分類や，英語母音を対象とした矯正度推定が検討されている．本研究では，前者に対して，音声学者による学習者分類との比較を通して提案技術の妥当性を示し，後者に対して，より高精度な矯正度推定手法を提案する．

## 2. 音声の構造的表象とそれを用いた外国語発音分析

### 2.1 音声に不可避的に混入する静的な音響的バイアス

音声はある話者によって生成され，ある音響系を通して収録され，ある伝送系を通して伝搬されるが，いずれの過程においても音響的なバイアスが混入する．話者によって異なる声道形状や声道長に起因するバイアス，マイク特性の差異に起因するバイアス，発声環境（部屋）や伝送経路の差異に起因するバイアス，などである．学習者音声を収録するたびにこれらの音響バイアスは変化するが，各収録においては，いずれも静的なバイアスとして考えられる．これらは音声のスペクトル包絡特性を変形させるが，特徴パラメータ空間でこの変形を考えれば，それは空間写像ととらえることができる（**図 1** 参照）．特徴パラメータとしてケプストラム $c$ を考え，より具体的な写像を考える．マイクや伝送経路の差異（および声道形状差異の一部）は伝達関数をかける演算となるため，ケプストラムに対する加算 $c' = c + b$ としてモデル化される．一方，声道長の差異は共振周波数を高低させる変形となるため，ケプストラムに対して行列を乗じる演算 $c' = Ac$ としてモデル化される（**図 2** 参照）[17)]．なお，文献 17) で提案されている行列 $A$ は，回転性の非常に強い行列となる[18)]．

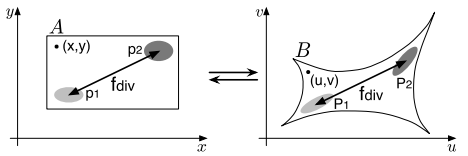


**図 1**　空間写像とそれに不変な $f$-divergence
Fig. 1　Feature transformation and transform-invariant $f$-divergence.

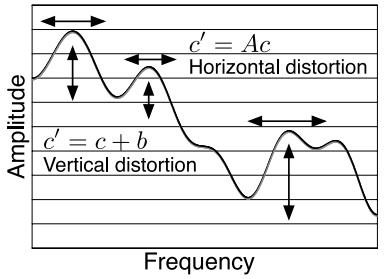


**図 2**　ベクトル $b$ および行列 $A$ によるスペクトル変形
Fig. 2　Two kinds of spectral distortions caused by vector $b$ and matrix $A$.

### 2.2　音響バイアス不変な音声の構造的表象

筆者らは，$f$-divergence [19] が，任意の連続かつ可逆な写像に対して不変性を持つ計量であること，また，不変となる計量は $f$-div. のみであることを証明している（図 1 参照）[20]．$f$-div. は $t > 0$ において凸な関数 $g(t)$ に対して，式 (1) で定義される．ここで，$p_i(\boldsymbol{x})$ は $i$ 番目の事象である．事象は点ではなく，確率密度分布として表現される．

$$f_{div}(p_1, p_2) = \oint p_2(\boldsymbol{x}) g\left(\frac{p_1(\boldsymbol{x})}{p_2(\boldsymbol{x})}\right) d\boldsymbol{x} \tag{1}$$

$g(t)$ を換えることで様々な $f$-div. が定義可能であるが，$g(t) = t\log(t)$ とすれば，$f$-div. は KL-div. となり，$g(t) = \sqrt{t}$ とすれば[⋆1]，$-\log(f\text{-div.})$ はバタチャリヤ距離になる．つまり，これらの分布間距離尺度は変換不変量である．

事象を単一ガウス分布でモデル化することを考えると，これは，非線形写像によって非ガウス分布へと変換されてしまう．しかし図 2 に示したように，対象とする写像に線形性が仮定できる場合，変換後もガウス分布となる．複数の事象群（分布群）が与えられたとき，任意の 2 事象間の距離を $f$-div. で計測し，これら事象群を距離行列で表象すると，任意の線形変換に対してこれは不変となる（**図 3** 参照）．これを（音声の）構造的表象と呼んでいる[8),15)]．

本研究では，事象をガウス分布でモデル化し，事象間距離をバタチャリヤ距離（以下，BD と略す）の平方根で定義する．2 つのガウス分布 $\mathcal{N}(\mu_1, \Sigma_1)$，$\mathcal{N}(\mu_2, \Sigma_2)$ に対して BD は式 (2) で定義される．$T$ は行列の転置を表す．

$$\mathrm{BD}(\mu_1, \Sigma_1, \mu_2, \Sigma_2) = \frac{1}{8}\mu_{12}^T \left(\frac{\Sigma_1 + \Sigma_2}{2}\right)^{-1} \mu_{12} + \frac{1}{2}\ln\frac{|(\Sigma_1 + \Sigma_2)/2|}{|\Sigma_1|^{\frac{1}{2}}|\Sigma_2|^{\frac{1}{2}}} \tag{2}$$

### 2.3　構造的表象を用いた外国語発音分析

先行研究において筆者らは，日本人学習者によって発声された英語母音群を構造的に表現し，これを，各学習者の英語母音発音表象として利用する手法を提案した[8)]．母音群を構造的に表現した場合，これは，個々の母音の絶対的な音響特性を捨象し，個々の母音が他の母音群とどのような関係性で結ばれているのか，その相対的な配置特性のみに着眼することになる．このように母音群の配置（母音体系）の様子をとらえることは，言語学的には言語の方言性をとらえることに相当する．日本語では方言の違いはアクセントやイントネーションなどの韻律的特徴に表出されるが，より一般的には，方言の違いは母音の音色を変化させる[21)–23)]．たとえば文献 21) は，北米各地より集めた約 400 名の音声を分析し，声道長を正規化したうえで，各話者の母音図における体系的差異を方言性として示している．**図 4** は米語方言における $F_1/F_2$ 図である．方言差による母音体系の変形は，一般に線形変換では表現困難であり，2.1 節で示した変形とは異なる変形である．文献 8) では，母音体系を $\sqrt{\mathrm{BD}}$ による距離行列で表現し，母音体系間距離（構造間距離）を式 (3) で定義している．

$$D_1(P, Q) = \sqrt{\frac{1}{M}\sum_{i<j}(P_{ij} - Q_{ij})^2} \tag{3}$$

$M$ は母音数，$P$，$Q$ は 2 話者の距離行列であり，$P_{ij}$，$Q_{ij}$ はその要素である．$D_1(P, Q)$ は幾何学的には，2 つの母音体系を回転およびシフトにより重ね合わせた後の，対応する 2 母音間距離の差の総和におよそ比例することが実験的に示されている（**図 5** 参照）[24)]．これは，明示的な適応処理，正規化処理を行うことなく，マイクなどの収録機器によるバイア

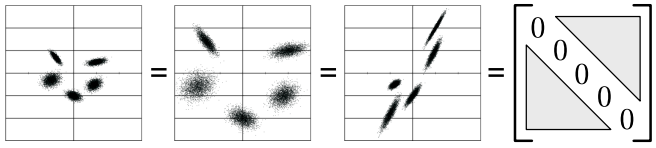


**図 3**　$f$-div. の距離行列として表現される内在する不変構造
Fig. 3　The invariant underlying structure represented as $f$-div.-based distance matrix.

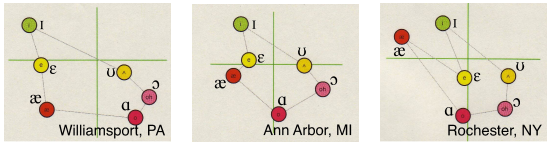


**図 4**　米国方言における母音配置（ただし一部）の差異[21)]
Fig. 4　Vowel arrangement (in part) of several American English dialects [21)].

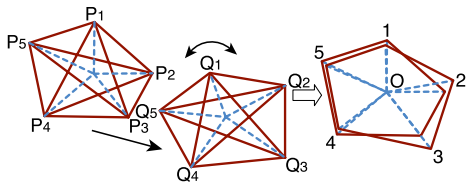


**図 5**　回転およびシフトによる構造の重ね合わせ
Fig. 5　Overlapping two structures through rotation and shift.

⋆1 文献 19) の定義式では $g(1) \equiv 0$ であるが，$g(t) = \sqrt{t}$ はこの条件を満たさない．しかし，$f$-div. の不変性は $g(1) \equiv 0$ を要求しないため，ここでは $g(1) \neq 0$ である $g(t)$ を用いた場合でも $f$-div. と呼んでいる．

**表 1**　母音置換によって模擬された 8 種類の日本語訛り
Table 1　Eight accented pronunciations of English vowels.

| | ɑ | æ | ʌ | ə | ɚ | ɪ | i | ʊ | u | ɛ | ɔ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| P1 | J | J | J | J | J | J | J | J | J | J | J |
| P2 | A | A | A | A | A | J | J | J | J | J | J |
| P3 | J | J | J | J | J | A | A | A | A | A | A |
| P4 | A | A | J | J | J | A | A | J | J | A | A |
| P5 | J | J | A | A | A | J | J | A | A | J | J |
| P6 | A | J | A | J | A | J | J | J | J | A | A |
| P7 | J | A | J | A | J | A | A | A | A | J | J |
| P8 | A | A | A | A | A | A | A | A | A | A | A |

A：米語母音を使用，J：日本語母音で置換

**表 2**　米語母音と日本語母音の置換表
Table 2　Vowel substitution table.

| 米語母音 | → 日本語母音 |
|---|---|
| /æ/, /ʌ/, /ɑ/, /ɚ/, /ə/ | /あ/ |
| /i/, /ɪ/ | /い/ |
| /ʊ/, /u/ | /う/ |
| /ɛ/ | /え/ |
| /ɔ/ | /お/ |

スや声道長差異によるバイアスをキャンセルした後の母音体系間差異を推定することに相当する[15]．なお式 (3) は，対角行列である距離行列の上三角部分をベクトルとして解釈し，2 つのベクトル間のユークリッド距離を計測しているわけだが，このベクトルのノルムを正規化する前処理を導入している．このノルム正規化は図 5 にある構造のサイズ正規化に相当する処理であり，調音音声学的には，調音努力の正規化に相当する演算である[25),⋆1]．

### 2.4　構造的表象を用いた学習者分類

さらに文献 8) では，模擬学習者発音を用いた学習者分類を試みている．12 名の帰国子女に米語 11 単母音を含む 11 単語[⋆2]を 1 回ずつ，および，日本語 5 母音を含む 5 単語[⋆3]を 5 回ずつ発声してもらい，母音区間を自動切り出しし，米語母音を 1 サンプルずつ，日本語母音を 5 サンプルずつ取得する．これに対して**表 1** に示すように，米語母音の一部を日本語母音と置換し，日本語訛りを模擬した．ここで置換対象の母音は**表 2** に従った．なお，異なる米語母音を同一種類の日本語母音と置換する場合は，同一種類・異発声の日本語母音を使用した[⋆4]．

12 名（A〜L）× 8 種類の訛り方（P1〜P8）として得られた 96 種類の日本語英語に対し，ケプストラム空間において各母音区間を自動検出し，HMM によりモデル化する（話者別，母音別に HMM を学習する）．1 サンプルから HMM のパラメータ推定を行うため，話者・母音種類に非依存な母音 HMM をまず構築し，上記母音区間を使って MAP 適応（話者・母音種類適応）を行った．音響分析条件を**表 3** に示す．また，2 母音間の距離は対応する状態間距離の平均として定義した．次に，任意の 2 母音体系間の距離を式 (3) で計算し，96 × 96 の母音体系間距離行列を得た．この距離行列に対して Ward 法[26] を適用してボトムアップクラスタリングを行った結果が**図 6** である．一方，2 母音体系間距離を式 (4) で定義して，クラスタリングを行った結果が**図 7** である．ここで $v_i^P$ とは，母音体系 $P$ における母音 $i$ のガウス分布である．$D_2(P,Q)$ は，2 母音体系を音響的に直接比較して，その差異（音としての違い）を定量化したものである．

⋆1 たとえば，怠けた発声になれば，母音は弱母音（schwa 化，/ə/化，図 8 参照）し，母音図中心に集まることになる．この場合，構造のサイズは小さくなることが実験的に知られている[8),25)]．
⋆2 beat, bit, bet, bat, but, put, boot, pot, bought, bird, about の 11 単語．
⋆3 /bVto/の V の箇所に日本語の母音を入れた 5 単語．
⋆4 そのため，日本語母音は 5 サンプルずつ取得している．

**表 3**　音響分析条件
Table 3　Acoustic analysis condition.

| | |
|---|---|
| サンプリング | 16 bit/16 kHz |
| 窓 | 窓長 25 ms，シフト長 1 ms |
| パラメータ | FFT ケプストラム（1〜10 次元） |
| HMM | 1 混合 monophone（全角分散行列） |
| トポロジー | 3 状態 left-to-right モデル |
| 音素間距離 | 対応するガウス分布間の $\sqrt{BD}$ の平均値 |

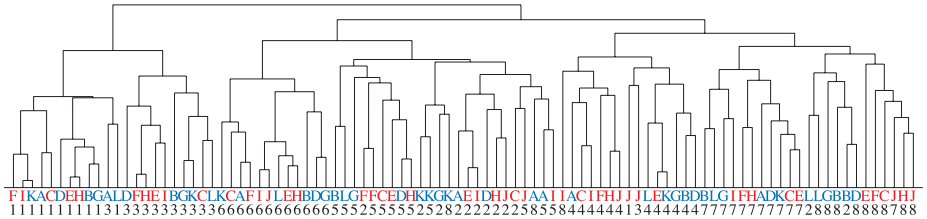


**図 6**　構造的表象に基づく 96 母音体系（[12 話者 A〜L] × [8 体系 1〜8]）の分類
Fig. 6　Clustering of 96 vowel systems using the structural representation.

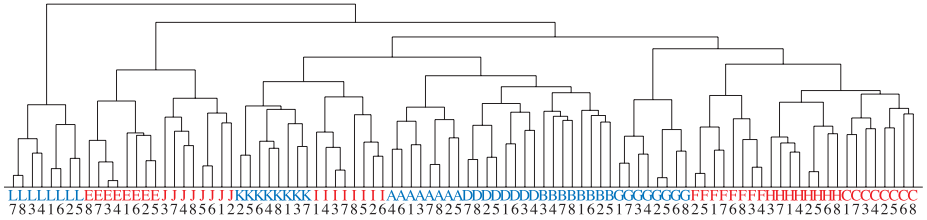


**図 7**　母音実体間の差に基づく 96 母音体系（[12 話者 A〜L] × [8 体系 1〜8]）の分類
Fig. 7　Clustering of 96 vowel systems using the vowel substances.

$$D_2(P,Q)=\sqrt{\frac{1}{M}\sum_i BD(v_i^P,v_i^Q)} \tag{4}$$

図 6 はおよそ母音体系に従って分類されている（左から 1，3，6，5，2，4，7，8 となっている）．一方図 7 は，完全に話者に従って分類されている（左から L，E，J，K，I，A，D，B，G，F，H，C となっている）．いい換えれば，前者は方言性（発音）の違いに基づく話者分類が行われ，後者は各話者の声道形状の違いに基づく話者分類となっている．

### 2.5　構造的表象を用いた母音矯正度の自動推定

式 (3) は，2 話者 $P$，$Q$ から得られた英語母音体系（構造的表象）に対して，両者の差異を定量化しているが，文献 8) では，学習者と教師間で比較を行い，その差異だけでなく，どの母音から矯正すべきか，母音矯正度の自動推定も検討している．文献 8) では，$P$，$Q$ 間の母音 $v$ に対する構造歪み（矯正度）を次式で定義している．すなわち，$d_1(P,Q,v)$ の大きな母音ほど矯正が必要とされる母音とし，この定義式の有効性も実験的に示している[⋆1]．

$$d_1(P,Q,v)=\sum_{j=1}^{M}|P_{vj}-Q_{vj}| \tag{5}$$

以上，先行研究について紹介した．以降本研究では，1) 学習者分類の音声学的妥当性を実験的に検討し，2) 矯正すべき母音種類の推定に関する技術的精度向上を狙う．

## 3.　日本語訛りに基づく学習者分類結果の妥当性検証

文献 8) で得られた図 6 は，方言性（母音体系）に基づいた分類が行われている様子は示されているが，クラスタ構成の様子（まず {13}，{52}，{78} が結合され，次に後者 2 つが {652}，{478} となるなど）の妥当性までは検証できていなかった．本章では，この分類の妥当性について検証する．英語音声学を専門とし，日本語訛りについて十分な知識と指導経験を持つ音声学者（第 5 著者．文献 27)，28) の著者）に，文献 8) で使用した 96 組の 11 母音群を聴取させ，これを母音図化させた．母音図は舌の（口腔内相対）位置で各母音を表現する方法であり，調音音声学では広く用いられている（**図 8** 参照）．得られた 96 枚の母音図を，各々，母音距離行列化し，文献 8) と同様の手続きで樹型図化し，分類の様子を図 6 と比較した．

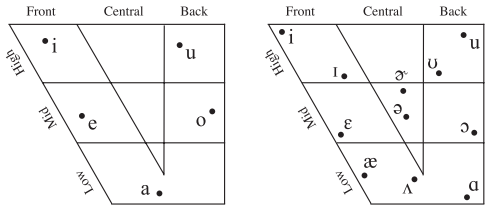


**図 8**　日本語母音と米語母音の母音図
Fig. 8　Vowel charts of Japanese and American English.

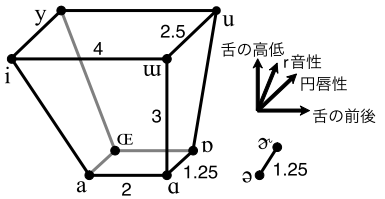


**図 9**　4 次元母音図
Fig. 9　Four-dimensional vowel chart.

### 3.1　英語音声学者による日本語訛りを有する米語母音発音の母音図化

通常，母音図は，1) 舌の前後位置，2) 舌の高低位置，を 2 次元平面上に配置する形で生成される．しかし，IPA（International Phonetic Association）による母音の定義では，上記 2 要因以外に，3) 円唇性，4) r 音性，なども考慮されている．その結果，円唇性の高い母音のみの 2 次元母音図と，低い母音のみの 2 次元母音図を構成している．本研究では様々な日本語訛りを扱うため，音声学者との協議の結果，舌の前後・高低位置だけでなく，円唇性，r 音性も同時に考慮して母音図を作成できる枠組みを採択した．すなわち，4 次元空間に母音体系をプロットすることになる．このような 3 次元以上の母音図は文献 29) でも検討されている．

文献 29) を参考にし，また，音声学者との協議の結果，本研究では，円唇性は無円唇，弱円唇，強円唇の 3 段階で評価し，r 音性は r 音性あり（+），なし（−）の 2 段階で評価することとした．旧来から使われている母音図（図 8 参照）の台形の枠組みは，母音間の知覚的距離の等価性を考慮して設計されている．そこで 4 次元母音図も，この 4 : 3 : 2 の台形に対して円唇性，r 音声の軸を付与する形で設計した．**図 9** に本研究で使用した母音図（の枠組み）を示す．予備的検討の結果，4 : 3 : 2 の 2 次元母音図に対して，円唇性の軸の重みは上舌母音，下舌母音に対して 2.5，1.25 とした．また，r 音性の軸の重みは 1.25 とした[⋆2]．なお，母音図プロット作業の効率を考え，Web 上でプロットするシステムを構築し，これを用いた．

### 3.2　手動作成された母音図に基づく学習者分類

作成された 96 種類の 4 次元母音図の各々に対して，母音間距離を求め，96 種類の母音

⋆1 式 (3) との類似性でいえば，$d_2(P,Q,v)=\sqrt{\sum_{j=1}^{M}(P_{vj}-Q_{vj})^2}$ が，より妥当な定義と思われるが，実データを用いた検証の結果，文献 8) では式 (5) を採択している．なお，$d_1$ と $d_2$ の比較については 4.3 節を参照．

⋆2 米語を対象とした場合，r 音性の有無が生じるのは/ə/と/ɚ/の対だけである．

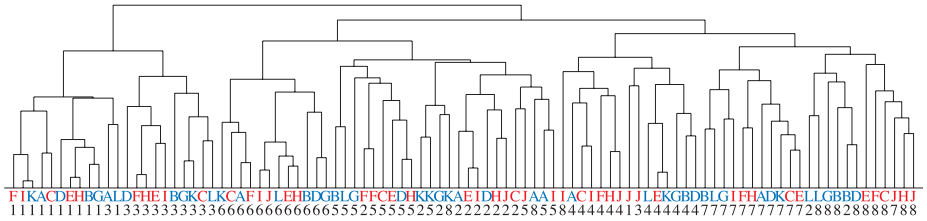


**図 6**（再掲）　構造的表象に基づく 96 母音体系（[12 話者 A〜L] × [8 体系 1〜8]）の分類
Fig. 6　Clustering of 96 vowel systems using the structural representation.

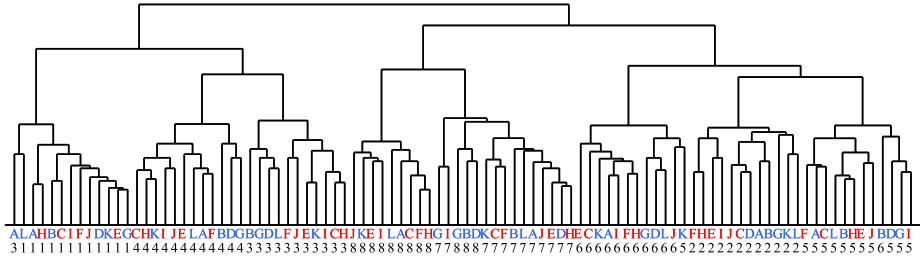


**図 10**　手動母音図に基づく 96 母音体系（[12 話者 A〜L] × [8 体系 1〜8]）の分類
Fig. 10　Clustering of 96 vowel systems using the manually drawn vowel charts.

距離行列を算出した．その後，式 (3) を用いて母音体系間距離を求め，96 × 96 の学習者間距離行列より，学習者分類を行った．結果を**図 10** に示す．比較のため，図 6 を再掲する．図 6 同様，図 10 も発音状態に基づいたクラスタリングとなっている．クラスタ結合であるが，結合順序が一部異なるのみで，非常に類似した結果となった．図 6 では，{1,3}，{{2,5},6}，{4,{7,8}} の 3 つの中クラスタが作られ，その後，後者 2 つが結合する．図 10 では，{1,{3,4}}，{{2,5},6}，{7,8} の 3 つの中クラスタが構成され，その後，後者 2 つが結合する．両者の違いは，状態 4 の発音の結合のみである．

式 (3) は，発音構造の各エッジ（音素間 BD）における教師学習者間差異を足し合わせる演算である．この演算に対して文献 7) では発音習熟度推定をタスクとして，各エッジに適切な重みを乗じることで，精度向上を実現している．本研究が対象とする学習者分類の場合も，図 6 を図 10 に近づけるために，エッジ重みを導入することは可能である．しかしこの場合，文献 7) にならえば，重み推定は教師あり学習となり，たとえば日本人英語分類に最適化された重みが推定されることとなる．筆者らは発音分類の対象を，非日本語を母語とする学習者（アジア圏英語や世界英語）に広げつつある[⋆1]．各種母語干渉の多様性を考えると，必ずしも教師あり学習は得策ではない．本章では，教師なしの条件下で（各エッジを平等に扱う）構造表象に基づく学習者分類時のクラスタ結合の様子が，音声学者による母音図から得られる樹型図とほぼ等しいクラスタ結合を有していることを示すことができた．

## 4. 母音矯正度推定の高精度化

### 4.1 提案する母音矯正度推定の高精度化手法

式 (5) は母音 $v$ の正しさを，$v$ と $v$ 以外との関係性（具体的には距離）を参照することで推定している．これは $v$ 以外の母音がおよそ正しく発音されていることを暗に仮定している．たとえば P1（表 1 参照）のように，すべての英語母音が日本語母音と置換されている場合，この仮定の妥当性は低減すると予想される．また式 (5) は，他母音との距離の 2 話者（教師・学習者）間差異を足し合わせて母音矯正度としており，各他母音を平等に扱っている．他母音の発音の正しさに応じて（学習者の母国語とは独立），他母音との距離の 2 話者間差異に重みを乗じる（式 (6)）ことで，より高精度な矯正度推定が可能になると考えられる．

$$d'_1(P,Q,v) = \sum_{j=1}^{M} w_j \, |P_{vj} - Q_{vj}| \tag{6}$$

問題は，各母音の発音の正しさが未知である状態でいかにして重み $w_j$ を推定するのか，である．ここでは，$d_1$ を初期矯正度（初期構造歪み）と考え，式 (7) で $w_j$ を定義する．

$$w_j = c \times \frac{1}{d_1(P,Q,j)} \tag{7}$$

$c$ は，$\sum_{j=1}^{M} w_j = M$ とするための正規化係数である．そして，この $w_j$ を用いて矯正度を再推定する（式 (6)）．当然 $d'_1$ を用いて各母音の重みを再推定し，$w'_j$（$\propto \frac{1}{d'_1(P,Q,j)}$）を算出できる．これを繰り返すことで $w_j$ をより精度良く推定することが可能になると考えられる．

### 4.2 模擬音声を用いた提案手法の実験的検証

3.1 節で得られた 96 枚の母音図を用いて提案手法の有効性を検証する．2 話者 $P$，$Q$ の

⋆1 世界英語を学習者単位で分類すれば，ある学習者の発音を「最も理解しやすい（intelligible な）発音」と評価する別の学習者（方言的に最も類似している別の学習者）を，世界中から探すことが可能になると考えられる．

母音図が与えられた場合，その枠組み（図 8 の台形）を重ね合わせることで，個々の母音を直接的に 2 話者間で比較できる．構造表象（距離行列）のみに基づく発音比較は，この枠組みが与えられていない状態で両者を自動的に重ね合わせて（図 5）比較することに相当する．本節では前者を「絶対的な枠組み（基準）が与えられている」という意味で「絶対的比較」と呼び，後者を「相対的比較」と呼ぶことにする．絶対的比較を行い，対応する 2 母音間の距離を求めれば，その距離の大小そのものが各母音の「絶対的母音矯正度」となる．本節ではこの絶対的母音矯正度と，式 (5) や式 (6) で計算される「相対的母音矯正度」との相関分析を行うことで，式 (6) の式 (5) に対する優位性を実験的に検証する．

まず，ある話者（帰国子女）の P8（米語教師相当）母音図と，同一話者の P1（極端な日本語訛りに相当）母音図を比較し，枠組みを用いることで絶対的母音矯正度を母音ごとに求める．次に，相対的母音矯正度（式 (5)）と重み付き相対的母音矯正度（式 (6)）を各母音に対して求める．話者数は 12 であるため，12 人 × 11 母音に対して求まった 132 対の絶対的母音矯正度と（重み付き）相対的母音矯正度の間で相関分析を行う．同様に，P8 と同一話者の P2〜P7 を比較し，相関分析を行う．最後に，96 枚ある母音図から任意の 2 枚を取り出し，これを教師，学習者と仮定して各母音の絶対的母音矯正度と（重み付き）相対的母音矯正度を算出して相関分析を行う（${}_{96}C_2$ 教師・学習者対 × 11 母音．結果として，50,160 対の絶対的矯正度と相対的矯正度が得られる）．得られた相関係数を**図 11** にまとめる．

相対的矯正度（式 (5)）と絶対的矯正度の相関について考察する．P2〜P7 および 96 × 96 は 0.750〜0.885 と比較的高い相関を示しているのに対し，P1 のみが 0.677 とより弱い相関となった．これは 4.1 節における考察の妥当性を示している．多くの母音が不適切な発音をしている場合，式 (5) の信頼性は低くならざるをえない．この相対的矯正度に対して，提案手法を用いることで（重み付き相対的母音矯正度，式 (6)），P1 を含め，すべての場合において相関は向上した．96 × 96 の場合，3.5 ポイントの向上であった．図 11 には再推定を 10 回繰り返した結果も示しており，いずれも相関は向上する．96 × 96 の場合，10 回の重み再推定でさらに 2.0 ポイント向上し，P1 においても相関は 0.755 となった．**図 12** は，提案手法による相関の向上の様子を示している．これらの実験結果より提案手法の有効性が示された．

なお図 11 は，同一話者内での P8 と P$i$（$i \neq 8$）の比較，同一および非同一話者内での任意の 2 母音図間の比較を通して得られた結果であるが，**表 4** に，各話者の P1〜P8 に対して，異なる話者の任意の母音構造を教師と見立てて相関係数を算出した結果（すなわち，異なる話者間での相関係数）を，話者（学習者）ごとに示す．重み推定回数は 10 回である．

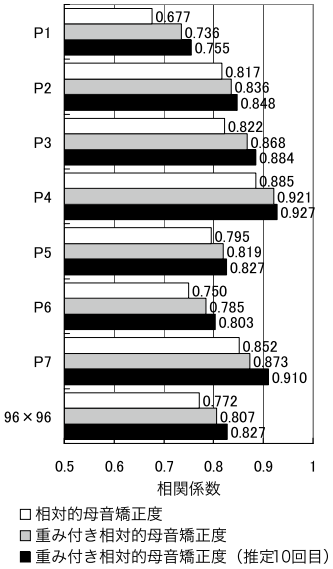


**図 11**　絶対的母音矯正度と（重み付き）相対的母音矯正度
Fig. 11　Absolute priority and relative priority of vowel correction.

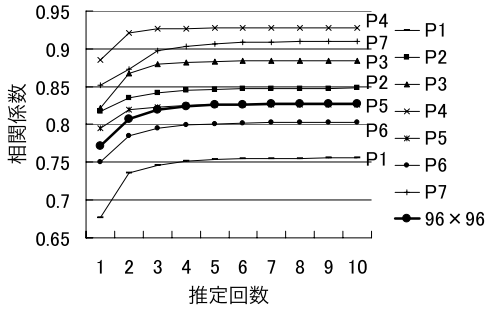


**図 12**　重み再推定による相関係数の変化
Fig. 12　Increase of correlations through iterated estimation.

**表 4**　各話者ごとの相関係数
Table 4　Correlation analysis per speaker.

| A | B | C | D | E | F | G | H | I | J | K | L | avg. | std. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.81 | 0.79 | 0.83 | 0.82 | 0.80 | 0.82 | 0.81 | 0.82 | 0.81 | 0.82 | 0.82 | 0.82 | 0.81 | 0.011 |

話者（学習者）間での性能差が非常に小さいことが分かる．

### 4.3　平方二乗和を用いた母音矯正度推定

式 (5) に対して，式 (8) で定義される平方二乗和を用いた $d_2$ に対する重み導入を考える．2 話者間の母音体系距離を式 (3) で定義するならば，4.1 節で述べたように，式 (5) より式 (8) の方が，幾何学的には自然な構造歪みの導出であると考えられる．

$$d_2(P,Q,v)=\sqrt{\sum_{j=1}^{M}(P_{vj}-Q_{vj})^2} \tag{8}$$

このように母音矯正度を定義すると，式 (3) との間に

$$D_1(P,Q)=\sqrt{\frac{1}{M}\sum_{i<j}(P_{ij}-Q_{ij})^2}=\sqrt{\frac{1}{2M}\sum_{i=1}^{M}d_2(P,Q,i)^2} \tag{9}$$

が成り立ち，構造間距離と各要素の構造歪みとの関係が明確となる．$d_2$ を用いた場合でも 4.1 節同様，重み $w_j$（$\propto \frac{1}{d_2(P,Q,j)}$）を導入することが可能である．この場合の重み付き母音矯正度は式 (10) となり，これを用いて再推定を繰り返す．

$$d_2'(P,Q,v)=\sqrt{\sum_{j=1}^{M}\{w_j\,(P_{vj}-Q_{vj})\}^2} \tag{10}$$

前節と同様の相関分析を行った結果（推定回数は 10 回）を**図 13** と**図 14** に示す．図 13 より，いずれの場合も平方二乗和の方が絶対値和より高い相関を示している．特に P1 の場合，相関は 0.8 を超えるようになり，効果が大きい．図 14 は 96 × 96 に対して求めた，再推定による相関係数の向上の様子である．重みを導入しない場合，絶対値和の方が性能が良く，文献 8) と一致する．しかし，他母音の発音の正しさに応じて重みを乗じ，再推定を繰り返すことでより適切な母音矯正度を算出する場合，$d_2$ を使うべきであることが分かる．96 × 96 の場合，文献 8) の重みなし・絶対値和と比較すると，最終的に 7.4 ポイント上昇している．

**表 5** は，表 4 同様，異なる 2 話者を教師・学習者と見立てて算出した，各学習者の相関係数である．ここでも，学習者によらず一定の性能が得られていることが分かる．

### 4.4　実データを用いた提案手法の実験的検証

前節では，文献 8) で使われた模擬学習者音声を用いて提案手法の有効性を示したが，本節では，実際の学習者音声を用いた検証を行った．日本人中学生 36 名（平均 13.5 歳）と 5 名の英語教師（平均 45.2 歳）に 11 単語セットを読み上げさせ，それを，音声学者（第 5 著者）に母音図化させたものを用い，4.2 節，4.3 節と同一の分析を行った．すなわち，実際の英語学習者と英語教師間で，彼らの母音図を用いた分析を行った．絶対的評価と相対的評価間の相関係数を，重み再推定回数の関数として**図 15** に示す．図には，絶対値和（$d_1'$）と平方二乗和の（$d_2'$）の両方を示している．この結果より，実際の学習者音声を用いた場合で

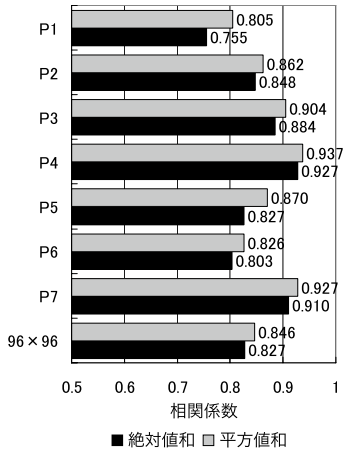


**図 13**　2 種類の重み付き相対母音矯正度
Fig. 13　Two kinds of weighted relative priority of vowel correction.

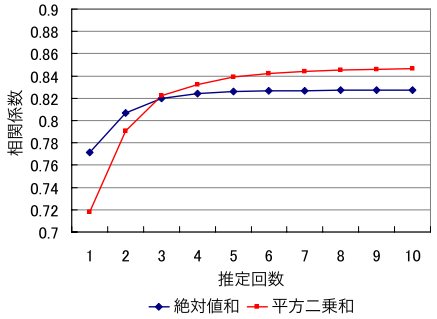


**図 14**　重み再推定による相関係数の変化
Fig. 14　Increase of correlations through iterated estimation.

**表 5**　各話者ごとの相関係数
Table 5　Correlation analysis per speaker.

| A | B | C | D | E | F | G | H | I | J | K | L | avg. | std. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.83 | 0.80 | 0.84 | 0.84 | 0.82 | 0.84 | 0.82 | 0.83 | 0.83 | 0.83 | 0.83 | 0.83 | 0.83 | 0.011 |

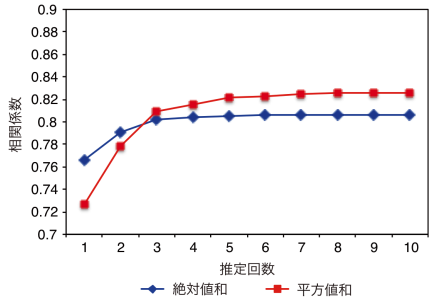


**図 15**　重み再推定による相関係数の変化
Fig. 15　Increase of correlations through iterated estimation.

もエッジに適切な重みをかける本手法が有効であること，および，図 14 同様，適切な重みの導入により平方二乗和の方がより高い相関値を示すようになることが分かる．

## 5. おわりに

本研究では，音声に不可避的に混入される非言語的かつ静的な音響的バイアスを捨象した音声の構造的表象を，外国語発音表象として利用することを想定し，その音声学的な妥当性と，技術的な精度向上について検討した．前者では，先行研究で示された発音分類と等価な処理を音声学者に依頼し，自動発音分類と手動発音分類とを樹型図として比較し，両者の良好な類似性を確認することができた．一部クラスタ構成に差異が見られたが，これは教師あり学習を導入し，エッジ重みを導入することで容易に解決できる問題である．本研究では，多様な母語干渉を考慮し，エッジを平等に扱う教師なしの条件下での検討にとどめた．後者においては，各母音の発音の良し悪しを考慮して，エッジに適宜重みを乗じ，かつ，重みの再推定を繰り返す手法を提案し，その妥当性を実験的に示すことができた．特に，完全に日本語発音となっている場合において，相関を 0.677 から 0.805 まで向上させることができた．さらに，実際の英語学習者と英語教師の発声データを用いた実験も行い，提案手法の有効性を示すことができた．なお本研究では，構造表象を用いた発音評価と従来の音響モデル適応に基づく発音評価とを明示的には比較しなかった．両者の比較と統合，および前者の技術的有効性については文献 7) を参照していただきたい．たとえば本研究で扱ったような，少量の音声データを用いた発音評価を音響モデルの話者適応を通して実装すると，不適切なモデルパラメータの修正が行われ，評価精度の低下を招く様子が報告されている．構造的な特徴量を用いた場合は，そのような現象は生じていない．その一方，無声子音など非言語的要因による音響変動が比較的小さい音素に関しては，構造的な特徴量のみでは十分な評価精度が得られていない．文献 7) では，絶対的な評価手法と相対的な評価手法の適切な融合を検討している．

## 参考文献


1) 現代ビジネス：経済の死角「社内公用語が英語って，なんか違うんじゃない？　楽天，ユニクロ，日産」，講談社（オンライン），入手先⟨http://gendai.ismedia.jp/articles/-/960⟩（参照 2011-03-21）．
2) 文部科学省：小学校外国語教育サイト，文部科学省（オンライン），入手先⟨http://www.mext.go.jp/a_menu/shotou/gaikokugo/index.htm⟩（参照 2011-03-21）．
3) 壇辻正剛：共通教育における ICT 支援の外国語教育と発音指導，電子情報通信学会技術研究報告，SP2010-115, pp.1–6 (2011).
4) Ohkawa, Y., Suzuki, M., Ogasawara, H., Ito, A. and Makino, S.: A speaker adaptation method for non-native speech using learners' native utterances for computer-assissted language learning system, *Speech Communication*, Vol.51, pp.875–881 (2009).
5) Hamada, H., Miki, S. and Nakatsu, R.: Automatic evaluation of English pronunciation based on speech recognition techniques, *IEICE Trans. Inf. & Syst.*, Vol.E76-D, No.3, pp.352–359 (1993).
6) Luo, D., Qiao, Y., Minematsu, N. and Hirose, K.: Regularized maximum likelihood linear regression adaptation for computer-assisted language learning systems, *IEICE Trans. Inf. & Syst.*, Vol.E94-D, No.2, pp.308–316 (2011).
7) 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，情報処理学会論文誌，Vol.52, No.5, pp.1899–1909 (2011).
8) 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，電子情報通信学会論文誌，Vol.J90-D, No.5, pp.1249–1262 (2007).
9) Willey, L.：アスペルガー的人生，東京書籍 (2002).
10) 綾屋紗月，熊谷晋一郎：発達障害当事者研究（ただし，筆者らの対談を含む），医学書院 (2008).
11) 深見　憲：ひろしくんの本（V），中川書店 (2006).
12) 東田直樹，東田美紀：この地球（ほし）にすんでいる僕の仲間たちへ，エスコアール (2005).
13) 岡ノ谷一夫：小鳥の歌と言語：共通する進化メカニズム（ただし，質疑応答含む），音響学会春季講演論文集，1-7-15, pp.1555–1556 (2008).
14) Hauser, M.D. and McDermott, J.: The evolution of the music faculty: A comparative perspective, *Nature neurosciences*, Vol.6, No.7, pp.663–668 (2003).
15) 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，電子情報通信学会論文誌，Vol.J94-D, No.1, pp.12–26 (2011).
16) 峯松信明：グローバル時代における英語発音とその科学的な分析方法，大学英語教育学会関東支部学会誌，No.7, pp.5–14 (2011).
17) Pitz, M. and Ney, H.: Vocal tract normalization equals linear transformation in cepstral space, *IEEE Trans. SAP*, Vol.13, No.5, pp.930–944 (2005).
18) Saito, D., Minematsu, N. and Hirose, K.: Rotational properties of vocal tract length difference in cepstral space, *Journal of Research Institute of Signal Processing*, Vol.15, No.5, pp.363–374 (2011).
19) Csiszár, I.: Information-type measures of difference of probability distributions

and indirect observations, *Studia Scientiarum Mathematicarum Hungarica*, Vol.2, pp.299–318 (1967).
20) Qiao, Y. and Minematsu, N.: A study on invariance of $f$-divergence and its application to speech recognition, *IEEE Trans. Signal Processing*, Vol.58, No.7, pp.3884–3890 (2010).
21) Labov, W., Ash, S. and Boberg, C.: *Atlas of North American English*, Mouton and Gruyter (2005).
22) Wells, J.: *Accents of English*, Cambridge University Press (1982).
23) Schneider, E.W., Burridge, K., Kortmann, B., Mesthrie, R. and Upton, C.: *A Handbook of Varieties of English: A Multi-Media Reference Tool*, Mouton de Gruyter (2004).
24) 峯松信明，志甫　淳，村上隆夫，丸山和孝，広瀬啓吉：音声の構造的表象とその距離尺度，電子情報通信学会技術研究報告，SP2005-13, pp.9–12 (2005).
25) Minematsu, N., Asakawa, S. and Hirose, K.: Para-linguistic information represented as distortion of the acoustic universal structure in speech, *Proc. ICASSP*, pp.261–265 (2006).
26) 宮本定明：クラスター分析入門，森北出版 (1999).
27) 英語音声学研究会：大人の英語発音講座，NHK 出版 (2003).
28) 牧野武彦：日本人のための英語音声学レッスン，大修館書店 (2005).
29) Ladefoged, P.: *A Course in Phonetics, 4th Edition*, Heinle & Heinle (2001).

(平成 23 年 4 月 11 日受付)
(平成 23 年 9 月 12 日採録)

**峯松　信明**（正会員）

1995 年東京大学大学院工学系研究科博士課程修了．博士（工学）．現在，同大学院情報理工学系研究科准教授．2002〜2003 年在外研究員（KTH，スウェーデン）．科学から工学に至るまで，音声コミュニケーションに関する研究に従事．IEEE，ISCA，SLaTE，IPA，CALICO，音響学会，電子情報通信学会，人工知能学会，音声学会，音声言語医学会各会員．

**鎌田　　圭**

2008 年東京大学大学院新領域創成科学研究科修士課程修了．修士（科学）．現在，JR 東日本勤務．音声信号処理，特に外国語発音分析や CALL システムに関する研究に従事．

**朝川　　智**

2008 年東京大学大学院新領域創成科学研究科博士課程修了．博士（科学）．2006〜2008 年日本学術振興会特別研究員 DC1．現在，ソニー株式会社勤務．音声信号処理，パターン認識に関する研究に従事．電子情報通信学会会員．

**鈴木　雅之**

2010 年東京大学大学院工学系研究科修士課程修了．修士（工学）．現在，同大学院工学系研究科博士課程に所属．音声認識，音声分析，音声強調に関する研究に従事．IEEE，ISCA，電子情報通信学会，日本音響学会各会員．

**牧野　武彦**

1991 年東京外国語大学大学院外国語学研究科ゲルマン系言語専攻修士課程修了．文学修士．1987〜1998 年カンザス大学に留学．現在，中央大学経済学部准教授．英語（特に発音）の教育法，方言性や母語干渉をともなう英語音声の分析，英語辞書編集に関する研究に従事．IPA，ADS，ISCA，音声学会，音韻論学会，英語音声学会，言語学会，英語学会各会員．

**西村多寿子**

1997 年東京大学大学院医学系研究科国際保健学専攻修士課程修了．現在，同研究科公共健康医学専攻客員研究員．看護師，保健師の実務経験後，医療翻訳者として独立．現在，医学サイトの論文紹介記事等を執筆．

**広瀬　啓吉**（正会員）

1977 年東京大学大学院工学系研究科博士課程修了．工学博士．現在，同大学院情報理工学系研究科教授．1987 年客員研究員（米国 MIT）．音声言語情報処理分野一般，特に韻律に着目した研究に従事．IEEE，米国音響学会，ISCA（Board メンバ），電子情報通信学会（フェロー），日本音響学会，人工知能学会，言語処理学会，信号処理学会各会員．