

大規模英語学習者を対象とした 音声の構造的表象に基づく発音評価とその応用*

◎高澤真章[†], 鎌田圭[‡], 竹内京子^{††}, 朝川智[‡], 峯松信明[‡], △牧野武彦^{*}, 広瀬啓吉^{††}
([†] 東大工学部, [‡] 東大新領域, ^{††} 東大総合文化, ^{*} 中央大学経済学部, ^{††} 東大情報理工)

1 はじめに

音声の音響的特徴は、話者や音響機器などの非言語的要因によって不可避免に変動し、同じ発話内容でも観測される音響事象は異なる。従来の音声システムでは、基本的に、大量の音声を用いて個々の言語事象を統計的にモデル化することで対処してきた。これは、不可避な音声歪みを内包した特徴量を利用していることに相当し、不一致問題の解決には至らず、話者を選ぶシステムとなっていた。音声認識の分野では、種々の適応技術（例えば話者適応）が用いられているが、本稿で検討する発音学習に適用すると、下手な発音に対して高いスコアを与える現象が免れない。これは、話者の違いと発音の上手下手とがスペクトル包絡という同一物理現象に基づくため、両者を切り分ける術を持っていなかったことが根本原因である [1]。

近年、非言語的要因に起因する静的な音声歪みを完全に排除して音声を表象する「音声の構造的表象」が提案された [2]。音声事象そのものではなく、事象間の相対的な関係のみを抽出／表象する。即ち、発話に含まれる有限個の事象から、全ての事象間差異（距離）を計算する。この際、一対一対応を満たす線形／非線形変換に対する変換不変量となる距離尺度 [3] を採択することで、非言語的変換不変の音声表象となる。この音声表象は構造音韻論の数学的／物理的実装に相当し、外国語学習者の発音状態（即ち、構造的歪み）を非言語的要因の影響を受けずに記述することが可能であり、種々の検討を行ってきた [1, 4, 5, 6, 7]。

本稿では、提案する構造表象を用いることで可能となる新しい発音学習応用（インタフェース）について検討する。具体的には、1) 学習者が選んだ一人の教師の発音と学習者の発音を比較することで矯正すべき発音部位を指摘するインタフェース、2) 身近な複数の教師及び複数の学習者の中に自分の発音を位置づけ、自らの発音と他者の発音の類似性／差異性を表示するインタフェース、そして、3) 大規模学習者群に対して発音分類を行うことで、例えば典型的な日本人英語を複数定義し、自身がどのタイプに分類されるのかを示すインタフェースなどである。

2 音声（発音）の構造的表象

静的な非言語的音響歪みは、乗算性歪みと線形変換性歪みに大別される。乗算性歪みは、マイクなどの伝送特性や話者の声道形状の違いの一部に相当し、

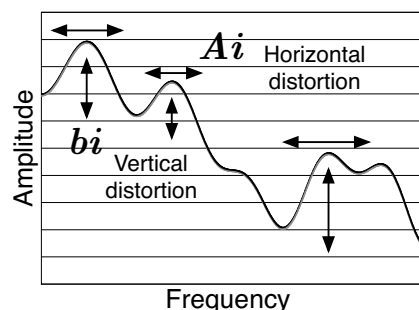


Fig. 1 非言語的要因によるスペクトル歪み

ケプストラム c に対する加算 $c' = c + b$ となる。線形変換性歪みは、声道長の差異や聴覚特性の差異が挙げられ、行列 A の乗算 $c' = Ac$ で精度良くモデル化される [8, 9]。以上より、不可避かつ静的な非言語的音響歪みは、線形（アフィン）変換でモデル化できる。図 1 はこの 2 種類の歪み（乗算性歪み＝垂直、線形変換性歪み＝水平）を示している。個々の音声事象を分布として捉え、全ての二事象間距離を計算すると（距離行列）、それは、事象群が成す幾何学構造を規定することになる。パタチャリヤ距離は線形／非線形変換に対して不変性を有するため [3]、これを用いることで、静的な非言語的音響歪みに一切不変な音声表象が得られる。この構造を音響の普遍構造と呼んでいる。

以下、日本人が英語（特に米語）を学習する環境を想定し、孤立単語発声に含まれる母音をガウス分布としてモデル化する。各学習者が発声した米語 11 単母音から構成される母音構造に対して検討する。

3 日本人英語学習者の模擬音声の作成

日本語・米語（英語）双方が話せる日本人話者 12 名（男性 6 名、女性 6 名）の米語 11 単母音と日本語 5 母音を収録した。収録では米語 /bVt/、日本語 /bVto/ を 5 回発声させ、母音部分のみを切り出して実験に用いた。米語 /bVt/ に有意義語が存在しない場合は、/b/ を /p/、/t/ を /d/ とするなどの処置を行った。

収録した米語母音と日本語母音を同一話者内で置換し、米語教師や日本人学習者の音声を模擬した。置換パターンには表 1 を用い、1 話者の音声から 7 人の日本人米語学習者（P1～P7）と 1 人の米語教師（P8）相当の発音を用意した。なお、米語と日本語の母音置換には、表 2 に示す日本人に頻出する置換母音を用いた。異なる米語母音が同一日本語母音と置換される場合は、同一母音・異発声のサンプルを用いた。

*Pronunciation assessment of a very large number of learners based on the structural representation of speech and its possible applications, by M.Takazawa, K.Kamata, K.Takeuchi, S.Asakawa, N.Minematsu, T.Makino, and K.Hirose

Table 1 母音置換によって模擬された8種類の発音

	ɑ	æ	Λ	ə	ɜ	ɪ	ɪ	ʊ	u	ε	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

Table 2 日本語母音・米語母音の置換表

日本語母音	↔	米語母音
/あ/		/æ/, /Λ/, /ɑ/, /ɜ/, /ə/
/い/		/i/, /ɪ/
/う/		/u/, /ʊ/
/え/		/ε/
/お/		/ɔ/

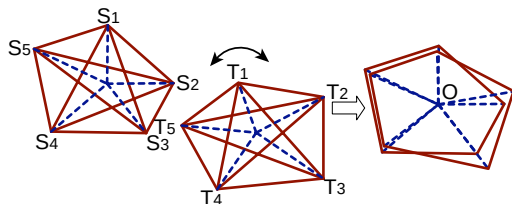


Fig. 2 回転とシフトによる構造の重ね合わせ

4 構造的表象に基づく母音矯正度推定

4.1 母音矯正度の自動推定

ある話者の母音距離行列 T と他話者の母音距離行列 S が与えられた時, 両者の距離を図2のようにして考える. 即ち一方の構造を回転・シフトして他方に合わせ, 対応する2事象間距離の最小和をもって両者の距離を定義する. この場合, 回転とは行列 A の掛算 (声道長差異の除去) に, シフトとはベクトル b の加算 (マイク差異の除去) に相当する.

ユークリッド空間に存在する二幾何学構造を対象にすれば, 上記の最小和は, 二つの行列定義のみから解析的に求めることができる [10]. 本研究ではバタチャリヤ距離を用いているため, 上記解析解を用いることが出来ず, タンパク質の構造解析でも用いられている下記の近似式を用いる. なお, この近似式の妥当性については [10] を参照して戴きたい.

$$D(S, T) = \sqrt{\frac{1}{m} \sum_{i < j} (s_{ij} - t_{ij})^2} \quad (1)$$

m は母音数であり, s_{ij} , t_{ij} は行列 S , T の要素である. D は二構造間の全体的な歪みを表現しており, 特定の母音 v に起因する歪みは以下で算出される.

$$d(S, T, v) = \sqrt{\frac{1}{m} \sum_{j=1}^m w_j (s_{vj} - t_{vj})^2} \quad (2)$$

d が大きい母音ほどより緊急に矯正が必要な母音として判定する. なお重み w_j は母音 j の「音声学的正しさ」を反映して求められる係数であり, その算出方法については [7] を参照して戴きたい.

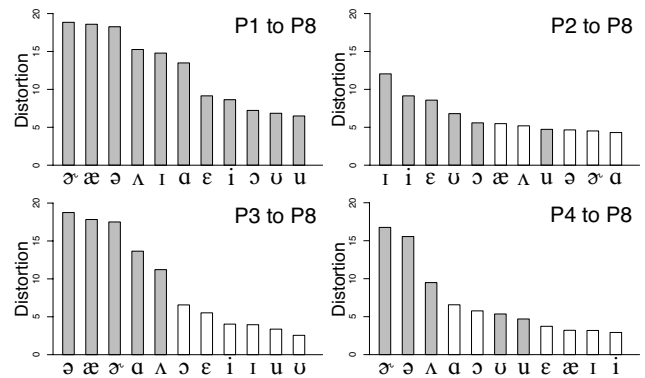


Fig. 3 P1 から P4 に対する母音矯正順序



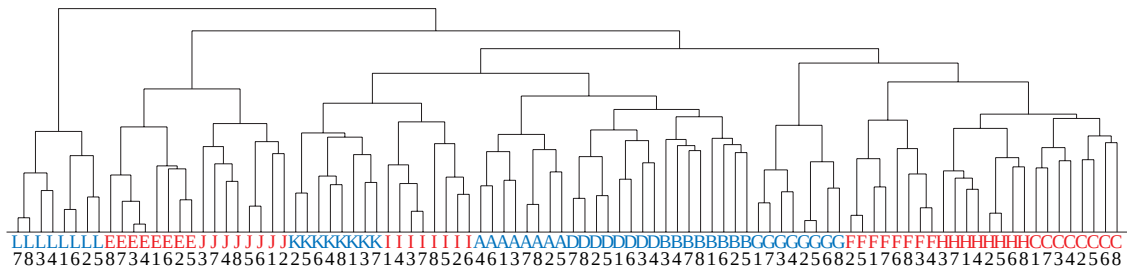
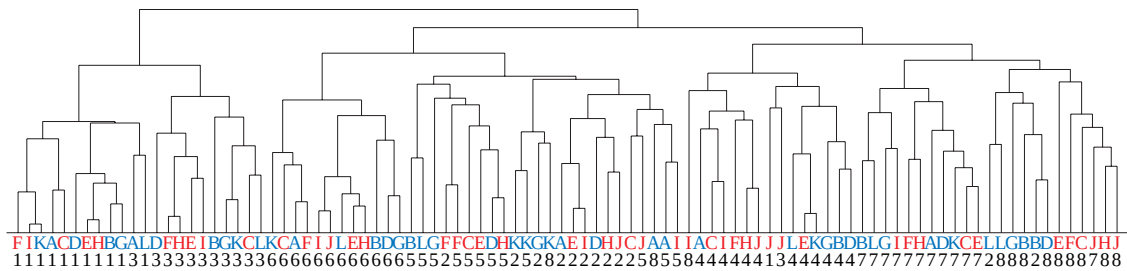
Fig. 4 Select your favorite teachers!

4.2 母音矯正度推定の妥当性の検証

日本人英語学習者の模擬音声を用いて, 式 (2) の妥当性を検証した. 詳細は [7] に譲るが, 例えば, 12 人の話者によって得られる平均的な $P_i (i < 8)$ 発音を, お手本となる P_8 と比較した場合に, どの母音から矯正すべきかを推定した結果を図3に示す (なお, $w_i \equiv 1.0$ の結果. 灰色が置換母音). この結果は「どの母音を日本語母音と置換したのか」という事前情報は一切用いていない. 二つの距離行列の比較のみを通して, 母音矯正順序を推定している. [7] では異なる話者間での構造歪み分析を行い, 異話者間でも妥当な矯正順序算出が可能であることを示している.

4.3 学習者が教師を選ぶインタフェース

学習者と「性別・年齢・声質の全く異なる話者」を適切に比較できるようになると, 複数教師の中から好みの教師を選ぶことで, 対象教師に近づく最も効率的な発音矯正法を得ることができる. 図4はオープンキャンパスデモの為に作成したインタフェースであるが, 憧れのメジャーリーガーや, 映画スターなどの著名人との比較も可能である. これらは, 継続学習に対するモチベーション向上につながる. 発音クリニックデモは年に一度行なわれるオープンキャンパスにて3年間実施してきた. 一昨年/昨年は晴天の下, 220 名, 207 名のデータを得ることが出来た. 本年度は雨天となり, キャンパス来訪者は約6割に減ったにも拘らず, 発音クリニックデモ参加者は225名であった. 本インタフェースの果たした役割は大きい.



第 5.1 節で検討した発音分類は、擬似的に作成された学習者発音を対象としていた。筆者らは、東京大学柏オープンキャンパスデモ及び種々の教育現場での収録を通して、三年間に渡り、これまで約 700 名の英語学習者（主に日本人）の収録を行ってきた。収録内容は日本語 5 単語（5 母音）と英語 11 単語（11 母音）を 1 回ずつである。収録した話者は 4 歳から 84 歳まで広い年齢層に渡る。同時に複数人が発声したり、咳・くしゃみなど背景ノイズも多少入ってしまうなど、録音環境が万全ではなかったため、事前に聴

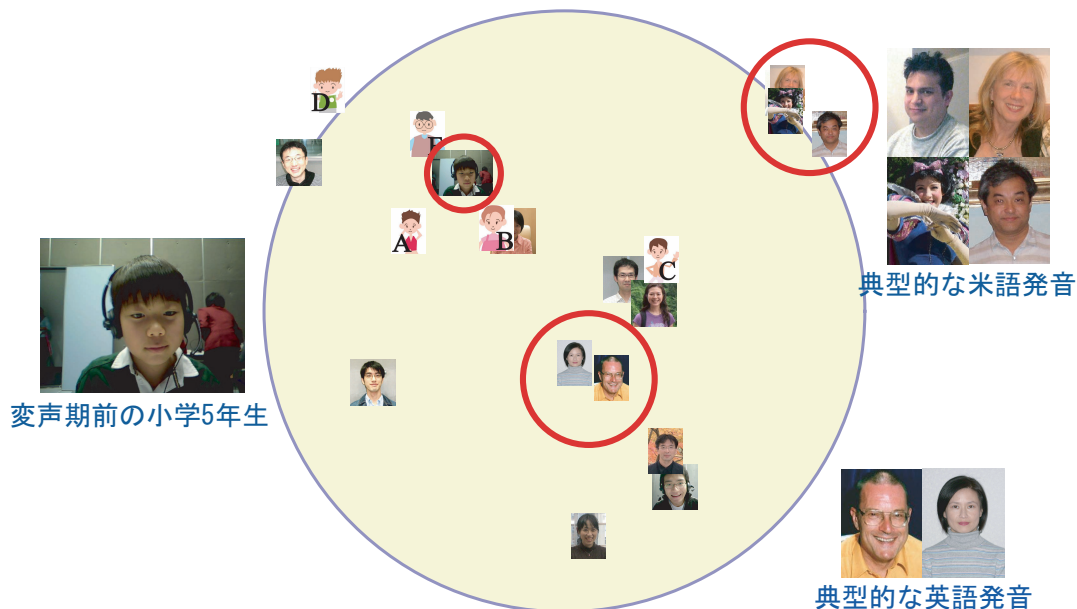


Fig. 7 身近な学習者／教師と自身との関係を表した発音分類マップ

Table 3 英語学習者クラスターの年齢・性別構成

No.	年齢				性別	
	0～19	20～39	40～59	60～	男	女
1	21	70	36	14	89	52
2	26	27	15	12	39	41
3	32	48	27	15	77	45
4	14	33	24	14	44	41
5	25	75	29	7	82	54

取を行ない、良好な音声として 564 名を選別した。

$N \times N$ の母音構造間距離行列 ($N=564$) に対し Ward 法によるクラスタリングを適用し、学習者分類を行った。5つのクラスターに分類したときの、各クラスターの年齢・性別構成を示したものが表3である。

年齢や性別といった話者の特徴に束縛されることがなく発音分類が行われていることがよく分かる。各クラスターに属する話者の音声を聴取したところ、「カタカナ発音（典型的な日本人英語）」や「/i/と/i/を混同している発音」といった分類が行われていた。参加者の多くは習熟度の低い学習者であったため、「アメリカ方言英語」や「イギリス方言英語」といったクラスターは得られず、「上手な英語発音」クラスターが1つ存在するのみであったが、5クラスター以上の分割を行うことで、そのクラスターに属する人数は少ないものの「アメリカ方言英語」や「イギリス方言英語」と呼べるクラスターを確認することができた。現在、本提案手法が算出する日本人英語分類の音声学／教育学的妥当性について音声学者の議論を行なっている。

6 まとめ

音声に混入する静的かつ非言語的な音響歪みをそぎ落とした形で音声を表象する構造表象の応用例として、筆者等は従来より、発音矯正アプリケーションを検討してきた。本稿では、構造表象を用いることで初めて可能となるインタフェースの実装を幾つか試みた。「複数話者の音声混ぜ合わせ」を基本とする従来法では、凡そ不可能な試みであると考えている。地球

上には約 20 億の英語学習者が存在するが、発音分類の究極の目標は、彼らを、個人を単位として分類し、地球規模の英語地図を作ることにある。言語学者らによる北米英語の地図は存在しているが [12], これは約 400 名の音声資料の音響分析に基づくものである。

参考文献

- [1] 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, vol.J90-D, no.5, 1249-1262, 2007.
- [2] 峯松信明, 西村多寿子, 西成活裕, 櫻庭京子, “構造不変の定理とそれに基づく音声ゲシュタルトの推定,” 信学技報, SP2005-12, 1-8, 2005.
- [3] 峯松信明, 朝川智, 広瀬啓吉, “線形・非線形変換不変の構造的情報表象とそれに基づく音声の音響モデリングに関する理論的考察,” 春音講論, 1-P-12, 147-148, 2007.
- [4] 朝川智, 峯松信明, 伊勢井敏子・ヤッコラ, 村上隆夫, 広瀬啓吉, “音声の構造的表象に基づく非母語話者の英語発音分析,” 信学技報, SP2005-24, pp.25-30, 2005.
- [5] 朝川智, 峯松信明, 広瀬啓吉, “発音の構造的表象に基づく学習者分類と発音教示生成,” 信学技報, SP2005-156, 37-42, 2006.
- [6] 鎌田圭, 朝川智, 峯松信明, 牧野武彦, 広瀬啓吉, “音声の構造的表象を用いた英語学習話者の分類に関する実験的検討,” 信学技報, SP2006-77, 7-12, 2006.
- [7] 鎌田圭, 朝川智, 峯松信明, 牧野武彦, 広瀬啓吉, “音声の構造的表象に基づく発音矯正必要度の計算手法の検討,” 信学技報, SP2007-36, 73-78, 2007.
- [8] M. Pitz and H. Ney, “Vocal tract normalization equals linear transformation in cepstral space,” IEEE Trans. Speech and Audio Processing, vol.13, 930-944, 2005.
- [9] 江森正, 篠田浩一, “音声認識のための高速最尤推定を用いた声道長正規化,” 信学論 (D-II), vol.J83-D-II, no.11, 2000.
- [10] 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,” 信学技報, SP2005-13, 9-12, 2005.
- [11] M. Russel, S. D’Arcy, “Challenges for computer recognition of children’s speech,” Proc. ISCA Tutorial and Research Workshop on Speech and Language Technology in Education, CD-ROM, 2007.
- [12] W. Labov, S. Ash, C. Boberg, “Atlas of North American English,” Mouton de Gruyter, 2005.