# 世界諸英語を対象にした自己視点からの発音距離行列の可視化*

☆川瀬佑司, 峯松信明, 齋藤大輔, 広瀬啓吉（東大）, 沈涵平（成功大, 台湾）

## 1 はじめに

国際共用語である英語によるコミュニケーション能力が重要視されている。通常学校教育における英語授業では米語（特に General American, GA）や英語（特に Received Pronunciation, RP）をモデル発音として採択することが多いが、国際的な環境で GA 話者や RA 話者と会話する機会はあまりない。通常は母語が異なる者同士が英語で意志疎通を図り、母語話者も地方訛りの英語を話してくることが多い。このような状況を鑑み、国際語（英語）には標準発音などなく、GA や RP も一つの Accented English として捉え、世界中には多様な英語が存在すると考える教育者も多い（World Englishes、世界諸英語 [1, 2]）。

このような立場に立てば、英語を使った他者とのコミュニケーション能力向上を図る場合、世界にはどのような多様な英語発音が存在し、自らはその中のどこに位置しているのかを把握することは重要であると考えられる。筆者らの先行研究では [3]、世界中の英語利用者による特定パラグラフ読み上げ音声コーパス（Speech Accent Archive, SAA[*1]）を用いて、個人を単位とした世界諸英語発音の自動分類が検討されている。これは、任意の二話者間の発音距離（発音の違いのみに着眼した発話距離）を求め、対象話者群に対してその距離行列を推定するタスクである。

更に、話者を単位とした発音距離行列を樹形図や多次元尺度法（Multi-Dimensional Scaling, MDS）を用いて可視化し、対象話者群がどのように分類されるのかを学習者に呈示した研究もある [4, 5]。

本研究はこれらの先行研究を受け、発音距離行列に対する新しい可視化手法を提案する。樹形図や MDS では話者群全体を可視化するが、提案する手法では、可視化結果を提供する学習者を中心に据え、その学習者とそれ以外の学習者群との関係のみを可視化する。評価実験の結果、一定の評価を得ることができた。

## 2 従来の可視化手法の問題点

樹形図や MDS による二次元の可視化は距離行列全体を表現することが狙いである。しかし可視化結果を渡される特定の学習者（話者 $n$）にとってみれば、知りたい主情報は、自分とそれ以外の話者との関係性であり、他話者 $s$ と他話者 $t$ が近いのか遠いのかは重要ではない。即ち樹形図や MDS は無用な情報までも呈示している。発音距離行列 $\{D_{ij}\}$ において、話者 $n$ が欲しいのは $\{D_{nj}\}$ $(j \neq n)$ の可視化結果である。これは原点に話者 $n$ を置き、数直線上に話者 $j$ を、原点からの距離が $D_{nj}$ となるよう配置すればよい。

英語を外国語として学ぶ者が国際的な場で聞く英語は訛りによる多様性もあるが、当然年齢や性別などの要因によっても発音の音響特性は変わってくる。この場合、学習者に英語を教える教師と類似した声色の発声の方が聞き取り易い（聞きなれた声色の発声の方が聞き取り易い）などの報告もある [6]。そこで本研究では、$\{D_{nj}\}$ に加え、各話者の声色の違いも含めた可視化を考える。ここでは声色を特定する一要因である年齢に着眼し、$\{D_{nj}\}$ 及び話者 $j$ の年齢情報を用いた二次元の可視化を検討する。

なお本研究は、発音距離行列の可視化を検討するため、発音距離 $D_{ij}$ に関しては自動推定結果ではなく、SAA コーパスに付属する各発声の国際音声記号（International Phonetic Alphabet, IPA）書き起こしを利用し、話者 $i$・話者 $j$ 間の IPA 書き起こし間距離を $D_{ij}$ とする。また年齢情報については、話者から提供された実年齢情報を使うこともできるが、一般に任意の話者から実年齢を知ることは社会通念上、難しいこともある。そこで知覚的年齢推定技術を利用し、音声から「聞こえ」としての年齢情報を推定し [7]、それを用いることも検討する。

## 3 関連する先行研究

### 3.1 IPA 書き起こしに基づく二話者間発音距離推定

先行研究では、SAA コーパスに付属する IPA 書き起こしに対して動的時間伸縮法（Dynamic Time Warping, DTW）処理を行ない、任意の二話者間の発音のみに関する距離を算出している [8]。IPA 書き起こしは音声学者によって、話者の性別や年齢などに左右されずに行なわれるため、DTW 距離は純粋に発音の差異を定量化していると考えられる。

IPA 書き起こし間の DTW は単音間距離行列が必要であるが、先行研究では 153 種類の単音を各々、音声学者に 20 回発声させて単音 Hidden Markov Model（HMM）を構築し、単音 HMM 間のバタチャリヤ距離でもって単音間距離行列を取得している。これを用いて書き起こし間の DTW を行ない、得られた話者 $i$ と話者 $j$ との距離を $D_{ij}$ とする。

### 3.2 知覚的年齢推定

先行研究では 3 つの音声コーパス、JNAS（成人音声）[*2]、S-JNAS（高齢者音声）[*3]、CIAIR-VCV（子供音声）[*4]の各話者の発声に対して聴取実験により、「聞こえ」としての年齢である知覚的年齢をラベリングしており、それを用いた未知音声に対する知覚的年齢推定技術を提案している [7]。具体的には各話者

*Pronunciation distance matrix visualization from a specific speaker's viewpoint using a World Englishes corpus by Y. Kawase, N. Minematsu, D. Saito, K. Hirose (The University of Tokyo) and H.-P. Shen (National Cheng Kung University, Tainan)

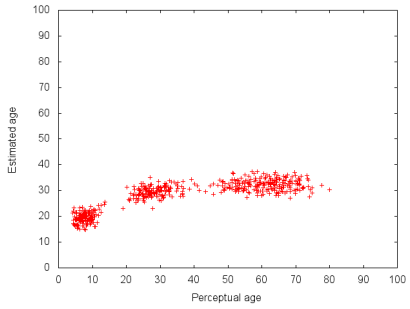


Fig. 1　GMM-SV+SVR による推定

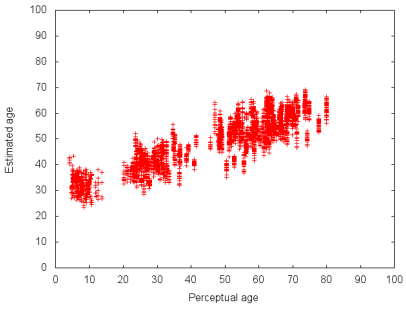


Fig. 2　60 秒分割による効果

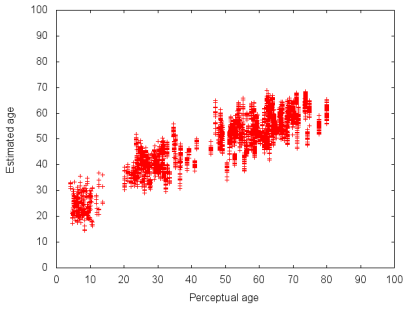


Fig. 3　F0 特徴による効果

毎に GMM を構築し、入力音声に対する GMM 尤度を計算し、各話者の知覚的年齢ラベル（分布）と尤度スコアの期待値操作によって知覚的年齢を計算する。知覚的年齢と推定年齢との相関は 0.88 である。

本研究では、先行研究 [7] と同様の検討を異なる技術的枠組みで行なったのでそれを報告し、次に、年齢情報と発音距離情報を用いた話者群の可視化について述べる。

## 4　知覚的年齢推定の再実験

### 4.1　知覚的年齢推定の再検討

先行研究 [7] は話者 GMM 尤度を用いているが、近年の話者認識研究では Supervector（SV）を特徴量とすることが多く、本実験では GMM-SV と Support Vector Regression（SVR）を使用し、再検討をした。なお、同様の検討は実年齢推定において行なわれている [9]。

### 4.2　実験条件

精度向上を狙い、複数の点に着目し実験を行った。尚、ここでは男性話者についてのみ報告する。

#### 4.2.1　使用した音声コーパスとデータ分割

CIAIR-VCV（6〜12 歳）の男性話者 145 名、JNAS（20〜60 歳）の男性話者 153 名、S-JNAS（60〜90 歳）の男性話者 202 名である。各話者の音声データを分割し、同一知覚年齢に対応するサンプル数を増加させた。具体的には各話者の音声データを 60 秒で分割し（知覚的年齢ラベルは同じ）、区分化データを使って、SVR を学習、評価した。

#### 4.2.2　使用した音声特徴量

先行研究では MFCC のみを用いて話者 GMM を構築した [7]。本研究でも MFCC を用いるが、構築した話者 GMM（混合数 64 の UBM-GMM から MAP 適応により推定）から SV を取得し、これを説明変数とする。

また F0 による SV も検討した。F0 抽出は Praat[*5] で行なった。MFCC 同様、混合数 64 で UBM を作成し、MAP 適応により話者 GMM を作成し、SV を抽出した。その後 2 つの SV を結合し、この結合 SV を説明変数として SVR により知覚的年齢を予測した。

なお SV は UBM からも構成できる。UBM-SV は、話者依存 GMM から得られる SV に含まれるバイアス項として解釈できる。そこで、GMM-SV から UBM-SV を差し引いた差ベクトルを用いた予測も検討した。

最終的な実験条件を、Table 1 に示す。

Table 1　実験条件

| | |
|---|---|
| 使用コーパス | CIAIR-CVC, JNAS, S-JNAS |
| UBM | 全発話 |
| 学習 | 偶数番号話者（60 秒分割）の UBM からの差分 GMM-SV |
| 評価 | 奇数番号話者（60 秒分割）の UBM からの差分 GMM-SV |
| 窓幅&シフト長 | 25ms length / 10ms shift |
| 特徴量 | 12MFCC + 12ΔMFCC + ΔEnergy + 対数 F0 |
| GMM 混合数 | 64 |

Table 2　年齢推定結果

| | 相関 |
|---|---|
| GMM-SV+SVR | 0.83 |
| + 60 秒分割 | 0.87 |
| + F0 | 0.89 |
| + 差分 SV | 0.89 |

### 4.3　実験手順と実験結果

以下の手順で実験を行った。

1. CIAIR-VCV、JNAS、S-JNAS 中の多数話者の音声から（60 秒分割せずに）UBM-GMM を、性別に依存させて学習した。
2. MAP 適応を用いて、GMM-UBM から各話者の GMM を得る。
3. 各話者 GMM の平均ベクトルを並べて連結し、SV を得る。MFCC が 25 次元であり、各々の GMM の混合数は 64 であるため、MFCC-SV の次元数は 1600 となる。
4. この SV を話者特徴量とし、知覚的年齢ラベルを用い、SVR の枠組みで知覚的年齢を予測した。

これにより推定された知覚的年齢を縦軸に、知覚的年齢ラベルを横軸にプロットすると Fig. 1 となり、

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

[*2] 新聞記事読み上げ音声コーパス, http://research.nii.ac.jp/src/JNAS.html

[*3] 高齢者話者データベース, http://research.nii.ac.jp/src/S-JNAS.html

[*4] 子供の声データベース, http://research.nii.ac.jp/src/CIAIR-VCV.html

[*5] Praat, http://www.fon.hum.uva.nl/praat/

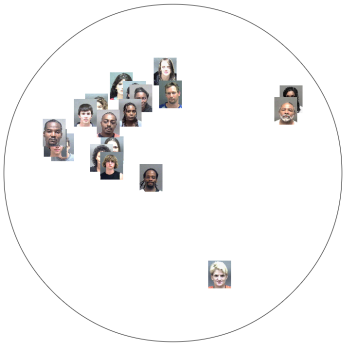

Fig. 4　俯瞰視点からの可視化

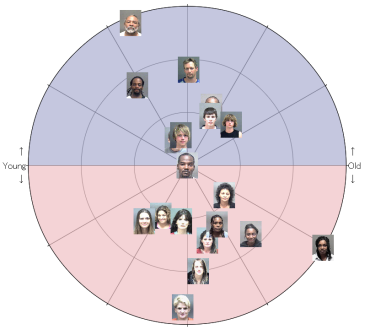


Fig. 5　自己視点からの可視化

相関は 0.83 となった。これに対し、音声データを 60 秒分割し、サンプル数を増加させたところ、Fig. 2 となり、相関は 0.87 へ増加した。

Fig. 2 によると、子供と大人とを十分に分類できていない。この点については、女性らしさを推定する実験において F0 を用いることで精度が向上している例もあり [10]、子供は大人よりも声のピッチが高いという点に着目し、F0 特徴量の追加を行った。MFCC と F0 に対して別個に UBM、GMM を作成し、次元数が 1664 の SV を作成した。その結果 Fig. 3 が得られ、相関 0.89 となった。子供と大人も分類できている。

最後に、UBM-SV を差し引く効果については、実験的には観測されなかった。以上の実験結果に対する相関値の上昇を Table 2 にまとめる。尚、先行研究 [7] に対して相関値の有意な上昇は観測されなかった。

## 5　可視化手法の検討

### 5.1　可視化手法

#### 5.1.1　俯瞰視点からの可視化

従来手法による、発音距離行列 $\{D_{ij}\}$ に MDS を用いた可視化結果を Fig. 4 に示す。ここでは、任意の二話者間の発音の近い・遠いが、二次元空間における距離の近い・遠いで表現されている。これは言語的特徴のみを用いた可視化である。

#### 5.1.2　自己視点からの可視化

従来手法の距離行列 $\{D_{ij}\}$ 全体の可視化に対し、本研究では特定話者（話者 $n$）に関する距離 $\{D_{nj}\}$ に加え、話者 $j$ の性別と知覚的年齢を用い可視化した。その結果を Fig. 5 に示す。ここでは、話者 $n$ を円の中心に据え、話者 $n$ と話者 $j$ の発音の近い・遠いを、円の中心からの距離の近い・遠いで表現した。また、Fig. 5 では上半円（青色部分）に男性、下半円（赤色部分）に女性を、さらに、円の左に寄る程、知覚的年齢が低く、円の右に寄る程、知覚的年齢が高くなるように配置した。これは言語的特徴と非言語的特徴を用いた可視化である。

尚、本研究では、各話者のプロット位置に点ではなく顔写真を配置することで、より直感的な可視化を検討した[*6]。

### 5.2　評価実験

これらの可視化手法について、アンケートによる主観的な評価実験を行った。尚、評価実験に際し、以下の 3 つの可視化手法を比較した。

**手法 M**　発音距離行列 $\{D_{ij}\}$ を用いた、俯瞰視点からの可視化

**手法 R**　発音距離 $\{D_{nj}\}$・性別・実年齢を用いた、話者 $n$ 視点からの可視化

**手法 P**　発音距離 $\{D_{nj}\}$・性別・知覚的年齢推定結果を用いた、話者 $n$ 視点からの可視化

各手法の比較に際して、同一の SAA 話者 20 名（男性 10 名、女性 10 名）を 2 グループ用意し提示した。一方のグループは実年齢が世代で均等になるように、もう一方のグループは知覚的年齢が世代で均等になるように考慮して選択したものである。

本評価実験は Web 上で行い[*7]、被験者が各手法を自由に選択し、手法 R・P に関しては中心話者を自由に選択できるよう提示した。また、顔写真をクリックすることで話者の音声が聞けるようにし、被験者が可視化に対する評価をより厳密に行えるよう配慮した。

アンケート項目に関して、10 段階評価で

- 各手法についての直感的なわかりやすさ
  （0:わかりにくい〜5:どちらともいえない〜10:わかりやすい）
- 手法 M と手法 R・P における発音の違いと可視化の位置関係の妥当性
  （0:妥当でない〜5:どちらともいえない〜10:妥当である）
- 手法 R と手法 P の年齢と可視化の位置関係の妥当性

を行い、

- 各可視化の長所短所
- 学習者目線ではない他の場面での活用
- 性別・年齢以外の要素を用いた可視化

に関して自由記述式で評価した。

[*6] SAA 話者自身の顔写真を使用しているわけではなく、SAA 話者情報から得られる性別・実年齢ラベルと、顔画像データベース [11] にある性別・実年齢ラベルを参照して、SAA の各話者に対し顔写真を対応させた。

[*7] 世界諸英語発音分類結果の可視化に関する評価アンケート，　http://www.gavo.t.u-tokyo.ac.jp/~kawase/speech_visualization/

### 5.3 実験結果・考察

評価実験に際して、28 名[*8]に 3 つの手法を見比べ・聞き比べてもらい、各手法の妥当性や分かりやすさを評価させた。

各手法について直感的なわかりやすさの評価結果は、M : 6.0、R : 6.4、P : 6.3 である。この結果より、手法 R・P のほうが若干の優位性を得た。しかし、被験者間における評価の差が大きく、手法 M か手法 R・P のどちらか一方が優れているという評価が多く見られた。この結果から示されるのは、手法 M と手法 R・P にはどちらも利点があるということである。それらを考慮して場面に合わせて的確に手法を使い分ける必要性があると考えられる。

また、被験者が発音の距離関係を把握できるか不明であるため、手法 M と手法 R・P における発音の違いと可視化の位置関係の妥当性を評価した。その結果は、M : 5.6、R・P : 5.9 であり、両手法の話者分布の様子は、被験者が感じる話者差異とある程度合致していることが考えられる。

また、手法 R と手法 P の年齢と可視化の位置関係の妥当性を評価した。その結果は、R : 6.0、P : 4.9 である。これは、まだ知覚年齢推定技術の精度が良くないことや、話者の顔写真が提示されたため、被験者判断が「見た目」の年齢に引きずられた可能性も考えられ、今後の課題として挙げられる。

自由記述結果によると、言語的特徴のみを用いた可視化手法よりも、言語的特徴に非言語的特徴を付加した可視化手法の方が分かりやすい・分かりやすくなる、という意見は多かった。可視化の際に、今回提案した性別や知覚的年齢だけでなく、出身地や所在地を用いることや、写真の大きさの大小による抑揚等の表現することはどうか、という意見も被験者から挙げられた。また、本実験では学習者目線での可視化ということで自己視点からの可視化を行ったが、教師目線からの可視化を行い、授業の前後で、中心話者の教師に対して、非中心話者の学習者がどの程度近づくかを見るという活用方法も挙げられ、これらの点においても更なる検討の余地がある。

## 6 まとめ

本研究では、知覚年齢推定手法の再検討を行い、発音距離行列を自己視点から可視化する手法を検討した。

知覚年齢推定手法の再検討では、先行研究 [7] に対して相関値の有意な上昇は観測されなかった。より高精度な推定を実現するためには、知覚年齢ラベルの改善や、韻律的特徴に基づく方法（話速やパワーの局所変動）を特徴量として加えることで、高齢者・非高齢者の識別率が向上している例もあり [12]、それらの検討の必要性が考えられる。

可視化手法の検討では、以下の 3 つの手法を比較し、アンケートによる主観的な評価実験を行った

- 従来手法である発音距離行列 $\{D_{ij}\}$ を用いた俯瞰視点からの可視化
- 発音距離 $\{D_{nj}\}$・性別・実年齢を用いた自己視点からの可視化
- 発音距離 $\{D_{nj}\}$・性別・知覚的年齢推定結果を用いた自己視点からの可視化

「実年齢」と「知覚的年齢」のどちらでの可視化がより理解に繋がるかという点は詳しくは測れなかったが、「俯瞰視点からの可視化」と「自己視点からの可視化」では、自己視点からの可視化の方が若干の優位性が示された。また、言語的特徴のみを含む可視化よりは、言語的特徴に加え非言語的特徴も含む可視化の方が分かりやすいという意見も多く見られた。しかしながら「俯瞰視点からの可視化」と「自己視点からの可視化」では、両手法とも利点があるため、場面に合わせて的確に手法使い分ける必要性がある。

これらを踏まえた上で、発音距離行列の可視化手法の更なる検討の必要があるのではないかと思われる。

## 参考文献


[1] B. Kachru *et al.*, *The handbook of World Englishes*, Wiley-Blackwell, 2009.

[2] J. Jenkins, *World Englishes: a resource book for students*, Routledge, 2009.

[3] H.-P. Shen *et al.*, “Automatic pronunciation clustering using a world English archive and pronunciation structure analysis,” *Proc. ASRU*, pp.222-227, 2013.

[4] N. Minematsu *et al.*, “Structural representation of the pronunciation and its use for classifying Japanese learners of English,” *Proc. SLaTE*, CD-ROM, 2007.

[5] 高澤他, “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”, 音講論, 3-10-12, pp.489-492, 2008.

[6] D. B. Pisoni, “Some thoughts on normalization in speech perception,”, in *Talker Variability in Speech Processing*, edited by K. Johnson and J. W. Mullennix, Academic Press, New York, pp.9–32, 1997.

[7] 山内他, “話者認識技術を応用した知覚的年齢分布の自動推定”, 信学技報, SP2002-186, pp.43–48, 2003.

[8] H.-P. Shen *et al.*, “Speaker-based Accented English Clustering Using a World English Archive,” *Proc. of SLaTE*, pp.184–188, 2013.

[9] 和田他, “年齢推定のための音声特徴量及び推定器の検討”, 信学技報, SP2010-27, pp.31–36, 2010.

[10] 王他, “スーパーベクトルと SVR に基づく MtF 話者のための女声度推定”, 信学技報, SP2012-120, pp.23–24, 2013.

[11] K. Ricanek and T. Tesafaye, “MORPH: A Longitudinal Image Database of Normal Adult Age-Progression,” *Proc. of Int. Conf. Automatic Face and Gesture Recognition*, pp.341–345, 2006.

[12] 峯松他, “話者認識技術を利用した主観的高齢話者の同定とそれに基づく主観的年代の推定,” 情報処理学会論文誌, vol.43, no.9, pp.2186–2196, 2002.


[*8] 28 名の内訳は学生 16 名・社会人 12 名である。被験者には日本語母語話者が多いが、それ以外にも英語・ポーランド語・ベトナム語・中国語・チェコ語・韓国語母語話者もいる。また語学に関する教育者や学者も含んでいる。