

話者間の英語発音距離予測に用いる音響的特徴の検討*

☆笠原駿（東大），△史天澤（清華大・中国，東大），峯松信明，
齋藤大輔，広瀬啓吉（東大）

1 はじめに

英語は唯一の世界共通語として受け入れられ、母国語として、公用語として、あるいは外国語として様々な国で話されている。英語が世界中に広まっていく中で、各国の地域性や民族性の影響を受けて、英語の統語、語用、綴り、発音など様々な側面が不可避的に変化してきた。発音に着目すれば、現在世界中には多くの訛り（外国語・地方訛り）英語が存在している。この訛りは話者の母国語や居住地のみだけでは決まらず、話者の学習方法や生活スタイル、家族や友人、英語教師といった周りの人間の有する訛りなどに影響されるもので、各自の訛りは個人により異なると言える。このような状況を鑑みて、近年 Kachru らにより、英語には基準となる発音は無いとする“World Englishes” (WE, 世界諸英語) [1] という考え方が提唱され、これを採択する教師が増えている。英語学習者が全員英語母国語話者の発音を理想とすることがなく、一人一人異なる英語発音を有する場合においては、学習者の英語が母語話者のそれと比べてどこがどのように違うのか、だけでなく一人一人の英語が WE の中でどのように位置づけられるか、を示すことも重要となる。本研究の最終的な目標は、発音訛りに焦点を置き、英語話者を単位として自動分類し、WE を一望できる発音地図を作成することである。話者を分類するためには、任意の話者間で発音がどれだけ違うかを定量的に表した“発音距離”が必要である。本研究では、この発音距離を定義するとともに、音声情報のみを用いて予測することを試みる。

[2] では、Speech Accent Archive [3] が提供する各話者の IPA 発音書き起しを用いて、特定話者依存の音響モデルに基づく単音間距離と Dynamic Time Warping により、基準の発音距離を算出していた。しかし、基準距離の算出には、この他にも複数の定義の仕方が考えられる。今回、別の話者による新たな単音音響モデルを導入し、二通りの発音距離に対する実験を行い、異なる定義の距離についても各音響特徴が有効であるかを確認した。

予測では、性別や年齢などによる音響特性の変動に対し頑健な特徴として、発音の相対特性のみを捉えた構造的表象 [4, 5, 6, 7] を用いているが、話者による違いが表れにくい無声音などについてはこの表

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.

[plɪz kəl ʔstɛləːs ɦər tu brɪŋ dɪz θɪŋz wɪθ ɦər frəm ðə stɔː sɪks spʊnz ʌv ˈfɪʃ ʃsno piːz ˈfɪʃ θɪk ʔsleɪbs ʌv blʊː ˈtʃiːz æn meɪbiː ɛɪ snæk ˈfɔː ɦər brʌðə bɒb ˈwɪ ɔːlsoʊ nɪd ˈɛɪ sməl ˈplæstɪk ˈʃsneɪk æn ɛɪ bɪɡ tɔɪ ˈfrɔːɡ ˈfɔː ðə kɪdz ʃɪ kən ʔskʊb ˈðɪz θɪŋz ɪntu θriː ˈæd ˈbæɡz æn ə wɪl goː mɪt ˈɦər wenzdeɪ ˈæd ˈðə tuem ʔstɛɪʃən]

Fig. 1 The elicitation paragraph used in the SAA and an example of IPA narrow transcription

象の仕方が適しているとは言えない。そこで本研究では、二話者間に対応する音素 HMM の差異を直接とる絶対的特徴と、さらに細かい比較として考えられる HMM スーパーベクトルを構造特徴と組み合わせることを検討した。

2 Speech Accent Archive

Speech Accent Archive (SAA) [3] は、英語パラグラフの読み上げ音声と、国際音声記号 (International Phonetic Alphabet, IPA) を用いたそれぞれの音声の書き起しがセットになって提供されているコーパスである。図 1 は SAA で指定されている読み上げ文と、IPA による書き起しの一例である。

各々の録音環境で録音された音声を収集しているため、話者によっては大きな背景雑音を含んでいる。発音の書き起しは音声学を専門とする者による手作業で為されている。これまで世界中の話者から約 2,000 人分の音声提供されており、このうち約 1,200 人の話者の発音が書き起されている。

書き起しには装飾記号 (diacritical mark) も用いられており、それらを考慮すると全書き起しで使われている異なりシンボル種類数は約 550 である。読み上げ文は 69 単語からなり、CMU 発音辞書 [8] を参照すると、221 米語音素系列に変換できる。

発音の書き起しにおいては、性別や年齢といった英語発音訛り以外の話者性の情報は捨象される。そのため、二話者の発音の距離として書き起し間の差異を定量的に評価できれば、これを発音訛りに関する正解距離として採択し、回帰モデルの学習や評価において使用できる。

本研究では、SAA の中から、背景雑音が少なく、単語の挿入・脱落のない話者 369 人のデータを選んで用いている。

*Use of various acoustic features for prediction of English pronunciation distances between speakers. by S. Kasahara (The University of Tokyo), T. Shi (Tsinghua University, The University of Tokyo), N. Mine-matsu, D. Saito, and K. Hirose (The University of Tokyo)

3 基準距離の複数の定義

各話者の IPA 発音書き起しを用いて，基準発音距離を計算する．計算は [2] と同様に，Dynamic Time Warping (DTW) を用いて，二話者の書き起し間で単音の整合をとりながら行う．

DTW の計算のために、局所距離として全単音間の距離を定義する必要がある。[2] では、使用話者の発音書き起しに出現する最頻 95% に相当する 153 種類の IPA を抽出し、男性の音声学者（本稿では P01 と称する）にそれらの IPA が示す単音を 20 回ずつ発声させた。この録音データから 3 状態 1 混合の単音 HMM を構築し、二つの単音 HMM の対応する状態間のバタチャリヤ距離 (BD) の平均を単音間距離とした。残りの 5% の IPA は、修飾記号なしの IPA など音響特性が近いと思われるものに置き換えた。

ここでの計算は全て一人の音声学者による発話音声を用いているため、単音 HMM 及び単音間距離は、この音声学者固有の声色や、発声の癖への依存性があることは否めない。今回、提案する発音距離の予測技術が、異なる基準距離に対しても高精度に予測できることを検証するために、新たにもう一人の男性の音声学者（以下、P02 と称する）に発声を依頼し、別の基準距離を算出して比較した。

二名の発話音声をそれぞれ用いて作成した単音間距離間の相関は 0.657 であった。各単音間距離を用いて算出した話者間の発音距離についても両者の相関をとると、0.913 となった。

本論文では複数の話者による単音間距離について扱っているが、基準距離の算出においてはこの他にも、分布間距離の定義の変更や、単音 HMM の代わりに triphone HMM の使用するなど、定義の仕方は様々に変えられる。定義を変えた時の基準距離間の相関や、それぞれの基準距離の発音距離としての妥当性などについては [9] を参照して頂きたい。

4 発音距離予測に用いる音響特徴量

4.1 発音の構造的表象

音声から英語発音の音響モデルを訛り毎に構築し、各モデルのパラメータの違いから発音距離の予測に有効な特徴を取得することを考える。発音距離予測は、年齢や性別といった訛り以外の声色に由来する音響変動に頑健であることが望ましい。話者固有の声色の影響を音響モデルに反映させたくない場合、通常、多数の話者の音声を用いることにより、声色による変動の平均化を図る。しかし本研究での発音距離予測においては、話者は一人一人異なる英語発音を有しているとみなしている。そのため、発音のモデル化は話者単位で行わなければならない一方で、体格などに

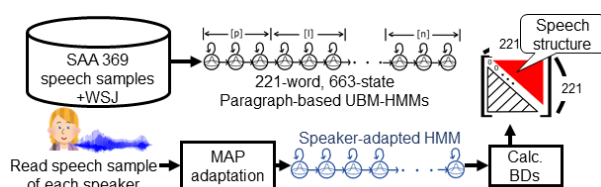


Fig. 2 Procedure to calculate the pronunciation structure

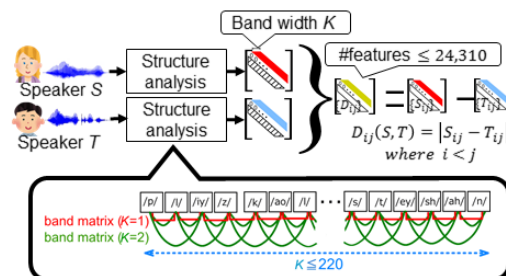


Fig. 3 Using a K-width band matrix as features of pronunciation difference between two speakers

起因する話者固有の声色を除去して行う必要がある。そのため、発音モデルからの特徴量抽出において、話者の声色の違いに頑健となるものを取得することを考える。本研究では、発音の構造特徴 [4, 5, 6, 7] を利用している。

話者の発音構造の算出過程を図 2 に示す. Wall Street Journal の音声データセットで学習した 3 状態米語音素モノフォン HMM [10] を初期モデルとして, SAA の全使用話者 369 人で追加学習し, SAA のパラグラフを単位とした HMM として 221 米語音素系列 Universal Background Model (UBM) を作成する. これと各話者音声から MAP 適応により各話者のモデル (パラグラフ HMM) を取得する. 同一話者の音素モデル間で分布間距離 f -divergence を求め, 発声を距離行列で表すと, これが話者の発音の構造特徴である. ある二つの分布 p_1, p_2 の f -divergence は次の式で表される.

$$f_{div}(p_1, p_2) = \int_{\mathcal{X}} p_2(x) g\left(\frac{p_1(x)}{p_2(x)}\right) dx \quad (1)$$

$g(t)$ は $t > 0$ の凸関数で, $g(t) = \sqrt{t}$ の時 $-\log(f_{div})$ は BD となる.

話者の違いはケプストラム空間上では回転や平行移動といったアフィン変換によって近似されるが、 f -divergence はそのような変換に対し不変となる。そのため、 f -divergence のみで構造的に表象すると、発話内容を揃えた上で発音が同じであれば、発音構造は話者性に対する非依存性が高くなる。逆に、話者間で発音構造に違いがあれば、その差異は発音の違い（訛りの違い）の特徴として利用できると考えられる。

二話者間の発音距離を予測する際の特徴として、二話者の発音構造の差行列 D_{ij} (図3) の各要素を利用する。実験では、発音構造の距離行列のうち幅 K ($1 \leq K \leq 220$) の帯行列部分のみを使用しているが、これは前後で K 個離れた音素モデルまでの分布間距離

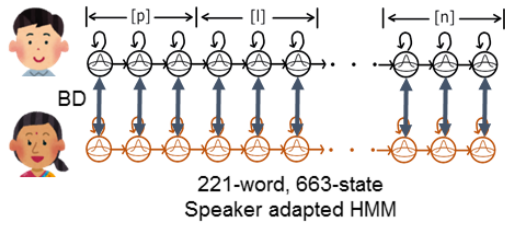


Fig. 4 Absolute features obtained by direct comparison

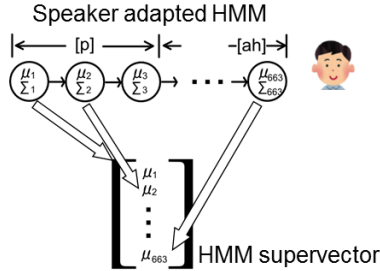


Fig. 5 HMM supervector

をとり構造化することに相当する． K を変えて距離予測実験を行い，時間的にどれだけ離れた発音間の差異までが予測に有効かを検証している．[2] では， K が最大になるまで予測性能が上がり続け，構造の距離行列の全要素が特徴として有効であることを示している．

構造特徴は，話者の音声セグメントを，空間におけるそれぞれの絶対的な位置ではなく，発声中の各セグメント間の相対的な配置のみで捉えることにより，話者の違いによる音響変動から発音の違いを切り離すものである．しかし音声には，摩擦音や破裂音といった無声のものもあり，これらは話者による影響が比較的少ない．これらの音声は，空間での位置そのものに話者の発音の特徴が表れていて，相対関係で記述してしまうことが最適とならない可能性が考えられる [11] ．

4.2 発音の絶対的特徴

二話者間に対応する音声セグメントモデルの差異を直接とり，これを話者間の発音の差異を表す絶対的特徴として導入し，構造的な特徴と組み合わせることを考える．絶対的特徴は，声色の違いが発音の違いと分けられず残るものの，話者の差異が表れにくい無声音などについては有効な特徴となり得る．

絶対的特徴は，図 4 のように，二話者間に対応する状態同士の BD をとり算出する．本研究では，同じ音素 HMM に属する 3 状態で BD の平均をとる音素を単位とした 221 次元の特徴と，状態毎に別の特徴としてそのまま扱う状態を単位とした ($221 \times 3 =$) 663 次元の特徴を使用する．

Table 1 Acoustic analysis conditions

Sampling frequency	16 bit / 16 kHz
Window size / Frame shift	25 ms length / 10 ms shift
Parameters	12 MFCC + 12 Δ MFCC
#mixtures / #states	1 / 3

4.3 HMM スーパーベクトル

各話者の 221 米語音素系列 HMM にある全 663 状態の平均ベクトルを連結し一つのベクトルとしたものを，話者の HMM スーパーベクトル (HMM-SV) として導入する (図 5) ．話者 X と話者 Y 間の発音距離予測に使う特徴量としては，次の HMM-SV の差ベクトル $\{d_k\}$ を用いる．

$$d_k(X, Y) = |SV_k^X - SV_k^Y|^2 \quad (2)$$

ただし， SV_k^X は話者 X の HMM-SV の k 次元目の要素を表す．話者 HMM 作成の際に用いた音響分析の特徴量数を m とすると，この時 $\{d_k\}$ の特徴量数は $663 \times m$ 次元となる．

HMM-SV は絶対的特徴と同様に，話者の声色の違いがそのまま影響する特徴である．また HMM-SV は，音素を単位とした絶対的特徴，状態を単位とした絶対的特徴のさらなる拡張で，状態が持つ音響分析特徴の各次元を単位としたさらに細かい比較が可能な特徴として考えることができる．なお，このような特徴量粒度を上げることは，構造的表象でも可能であるが，本稿では絶対的特徴においてより粒度の高い特徴量を使用することを検討する．

5 実験

P01, P02 二人の音声学者による単音 HMM を用いた二通りの基準距離に対し，音声情報のみを用いた予測実験を行った．単音 HMM, UBM, 各話者 HMM を構築する際の音響分析条件を表 1 に示す．この時，HMM-SV の次元数は ($663 \times 24 =$) 15,912 である．

本研究では，LIBSVM [12] の ϵ -SVR を利用して，サポートベクター回帰による回帰モデルの学習及び予測の評価を行う．カーネル関数として，放射基底関数 $K(x_1, x_2) = \exp(-\gamma|x_1 - x_2|^2)$ を適用している．

学習及び評価では，SAA の話者 369 人から得られる 67,896 の全話者対を基準距離が大きい順に並び替え奇数番，偶数番でデータを二分割し，2-fold 交差検定を行う．この実験は話者対 open, 話者 closed となっており，使用話者 369 人で学習した状態で，さらに発音書き起しの与えられていない未知話者が一人表れた場面の想定で，学習用話者と未知話者間の距離をどれだけ精度良く予測できるか，の評価として凡そ解釈できる [13] ．

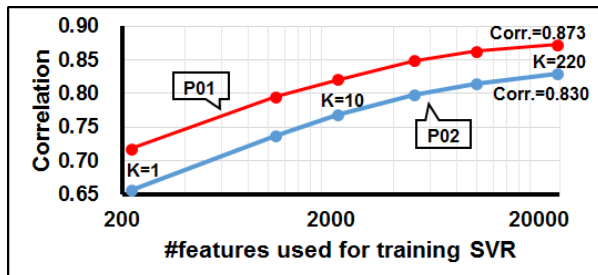


Fig. 6 Correlation between each of two reference distances and predicted distances using structural features

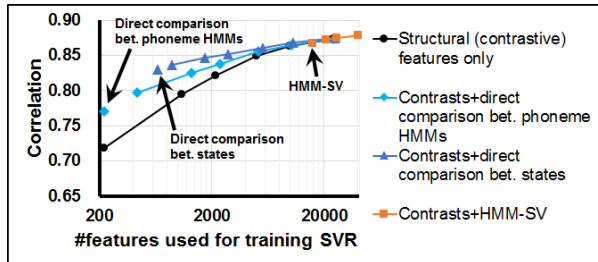


Fig. 7 Correlation between P01's reference distances and predicted distances using structural and absolute features

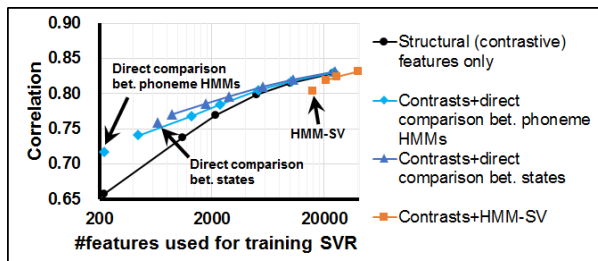


Fig. 8 Correlation between P02's reference distances and predicted distances using structural and absolute features

5.1 発音構造を用いた予測実験

最初に、構造特徴のみを用いて、二通りの基準距離を対象とする予測の結果を図6に示す。今回新たに導入した P02 の音声を用いた基準距離を予測した場合、P01 の音声を用いた場合と比べ少し性能が下がるものの、特徴量数の増加に応じて相関が最後まで上がり続けるという傾向は変わらず確認できた。

5.2 絶対的特徴を用いた予測実験

次に、構造特徴と絶対的特徴、または HMM-SV を組み合わせた場合の予測実験の結果を示す。図7は P01 の音声を用いた基準距離を、図8は P02 の音声を用いた基準距離を対象とした場合の結果である。

絶対的特徴、及び HMM-SV のみを用いた場合での相関を表2に示す。P01, P02 のどちらの基準距離に対しても、粒度の高い特徴を用いることによりより高い性能で予測ができています。

絶対的特徴、及び HMM-SV を、構造特徴に追加する形で用いた場合、特徴量数の合計が少ないところでは相関の向上がはっきりと見られたが、特徴量数が十分に多くなると、相関の差は僅かなものとなる。

Table 2 Corr. derived by using each abs. feature

	Comp. bet. phoneme HMMs	Comp. bet. states	HMM-SV (comp. bet. dimensions)
P01	0.771	0.830	0.866
P02	0.717	0.759	0.803

6 おわりに

本稿では、音声のみから話者間の英語発音距離を予測する研究において、まず学習及び評価に用いる基準距離の算出に別の定義を導入し、それぞれに対し発音構造特徴などを用いて予測実験を行った。その結果二つの距離定義に対し、ある程度高い予測性能が見られた。発音の基準距離の定義には、本稿で扱った二つ以外にも様々に考えられる。あらゆる定義の距離に対し実験を試み、予測が普遍的に高い精度で行えるか検証する必要がある。

また、予測の際に用いる特徴として、絶対的特徴と、さらにその拡張として考えられる HMM-SV を導入した。実験の結果、粒度が高い特徴ほど、距離予測に有効であることが示された。今後は HMM-SV の中から発音距離予測により有効な要素を抽出することで話者の声色の差異と切り分ける i-vector の考えを導入し、さらなる改善を検討していく。また、構造特徴についても、粒度を上げて利用することを考える。

参考文献

- [1] B. Kachru, *et al.*, "The handbook of World Englishes," Wiley-Blackwell, 2009.
- [2] 笠原他, "世界諸英語分類のための構造的表象を用いた発音距離推定の高精度化," 日本音響学会春季講演論文集, 121-124, 2014.
- [3] S. Weinberger, Speech Accent Archive. George Mason University. <http://accent.gmu.edu>, 2014.
- [4] N. Minematsu, *et al.*, "Speech structure and its application to robust speech processing," *Journal of New Generation Computing*, 28, 3, 299-319, 2010.
- [5] Y. Qiao, *et al.*, "A study on invariance of f-divergence and its application to speech recognition," *IEEE Trans. on Signal Process.*, 58, 7, 3884-3890, 2010.
- [6] 鈴木他, "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, 52, 5, 1899-1909, 2011.
- [7] 峯松他, "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," 情報処理学会論文誌, 52, 12, 3671-3681, 2011.
- [8] The CMU pronunciation dictionary, <http://www.speech.cs.cmu.edu/cgi-bin/cmudict>.
- [9] 史他, "世界諸英語発音分類のための話者間参照距離の算出手法に関する検討," 日本音響学会秋季講演論文集, 2014 (発表予定).
- [10] HTK Wall Street Journal Training Recipe, <http://www.keithv.com/software/htk/>.
- [11] 朝川他, "判別分析と構造表象を用いた話者の多様性に超頑健な音声認識," 日本音響学会秋季論文集, 113-116, 2008.
- [12] C.-C. Chang, *et al.*, LIBSVM, a library for support vector machine, 2001.
- [13] 笠原他, "構造的表象を用いた世界諸英語分類に関する分析的検討," 日本音響学会春季論文集, 125-126, 2014.