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*

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

1 はじめに

唯一の世界共通語である英語は，各話者の母語や英語の学習環境などの影響を受け，各個人の特性に応じて異なる発音訛りを伴って話されている．英語によるコミュニケーションでは，各自が英語母語話者に近い発音を有していることよりも，双方の持つ発音訛りが近いことの方が，相手の英語に対する了解度の高さに大きく関わることが知られている [1]．近年英語教師に広まりつつある “World Englishes” (WE, 世界諸英語) [2, 3] の立場で考えても，母語話者の発音が絶対的規範とはならない．重要なのは，各自が WE の多様さを認識することであるが，本研究では，WE を発音分類し，各話者を分類結果（地図）に位置づけることを検討している．本研究の究極目標は，15 億人と言われる英語話者の発音世界地図を作ることにある．

話者の自動分類には，話者間の発音の違いを表す発音距離を，音声学的に妥当な形で自動計測する必要がある．先行研究 [4] では，Speech Accent Archive (SAA) [5] の IPA 書き起しを用いて，二話者間の参照発音距離を定義し，音声情報処理技術を用いて，これを予測することを試みた．しかしながら，[4] で用いた参照距離は本タスクに最適化された参照距離とは言えず，より適切な参照距離の定義のためには，種々の実験的検討が必要である．本稿では，幾つか異なる距離定義を行ない，これらの比較を行なった．

2 参照距離の算出

2.1 先行研究における手法

先行研究 [4] では，発音の参照距離を， Dynamic Time Warping (DTW) によって求まる二話者の IPA 書き起しの整合コストとして定義した．DTW は時系列の整合，類似度計算で使われているアルゴリズムである．

DTW を用いて計算するためには，局所スコアとして全ての二単音間距離を用意する必要がある．そこで [4] は，ある男性の音声学者（以下， P01 と称する）の各単音の発声を用いて 3 状態 1 混合の単音 HMM を構築し，2 つの対応する状態間で特徴量分布の Bhattacharyya distance (BD) をとりその平均を単音間距離とした．

SAA パラグラフの各単語毎に，二話者の書き起しで対応する単音系列間で DTW を行い，正規化単語コストの総和（69 単語）を話者間参照距離と定義した．本研究においては，フィラーなどの余分に挿入された単語は手作業で削除し，脱落した単語はコストを 0 とした．

[4] では音響モデルで求めた単音間距離と DTW により参照距離を定義したが，これ以外にも定義の仕方は考えられる．[6, 7, 8] は各弁別素性に値を与え，弁別素性の差分で単音間距離として定義した．しかし，この弁別素性に基づく局所距離定義は主観によるもので，議論の必要がある．[9] は自己相互情報量 (pointwise mutual information, PMI) で整合スコアを表し，レーベンシュタイン距離として話者間距離を定義した．[10] は naïve discriminative learning (NDL) を用いて単音系列と単語の繋がりを定量化し，会話の了解度を推定した．

2.2 先行研究の問題点

先行研究での参照距離計算手法については，幾つか検討の余地が残されている．

- 分布間距離は BD が最適か
- 単音間距離に特定話者依存性があることで問題が起こらないか
- 前後の文脈を考慮しない monophone による単音のモデル化は適切か
- 単音 HMM 間距離で状態アライメントは必要か

本研究では， SAA 話者 395 人を対象として分析条件を変えながら音声セグメント間距離計算，及び話者間距離計算を行い，どの分析条件が最適なのかを検討する．最終的に求めるべきは，より適切な発音分類を可能とする参照距離である．これは例えば各条件における分類結果を専門家の分類結果と比較することで可能となるが，実験環境の構築が困難である．そこで，先行研究 [9, 10] で行なわれた，“母語話者らしさ” の主観実験結果と比較することで，より適した分析条件を客観的に求めることを検討する．

3 分布間距離の定義についての検証

[4] では単音モデル間の距離に BD を用いていた．しかし，単音間距離を求める際の分布間距離としては BD の他に，Hellinger distance (HD), Kullback-Leibler divergence (KL), Mahalanobis distance (MD) といったものを採択することも可能であり，どの定義が最適であるかは用途により異なる．本節では， P01 の単音音響モデルを用いて， 4 つの距離定義により単音間距離を求め，相関を調べる．

---

*A study on the calculation method of reference distance between speakers toward automatic clustering of accents of World Englishes. by T. Shi (Tsinghua University, The University of Tokyo), S. Kasahara, N. Minematsu, D. Saito, K. Hirose (The University of Tokyo)

Table 1　Corr. of phone-to-phone distances between distance metrics using P01's HMMs.

	BD	HD	KL	MD
BD	1.00			
HD	0.56	1.00		
KL	0.84	0.54	1.00	
MD	0.71	0.41	0.60	1.00

Table 2　Corr. of speaker-to-speaker distances between distance metrics using P01's HMMs.

	BD	HD	KL	MD
BD	1.00			
HD	0.97	1.00		
KL	0.99	0.98	1.00	
MD	0.96	0.89	0.94	1.00

正規分布における解析式が以下の式で表される [11].

$$D_{BD}(p,q) = \frac{1}{8}(\mu_p - \mu_q)^T \left(\frac{\Sigma_p + \Sigma_q}{2}\right)^{-1} (\mu_p - \mu_q) + \frac{1}{2}\ln\left(\frac{|(\Sigma_p + \Sigma_q)/2|}{\sqrt{|\Sigma_p||\Sigma_q|}}\right)$$

$$D_{KL}(p,q) = \frac{1}{2}(\mu_q - \mu_p)^T(\Sigma_p^{-1} + \Sigma_q^{-1})(\mu_q - \mu_p) + \frac{1}{2}tr(\Sigma_p^{-1}\Sigma_q + \Sigma_q^{-1}\Sigma_p - 2I)$$

$$D_{HD}^2(p,q) = 1 - e^{-D_{BD}(p,q)}$$

$$D_{MD}(p,q) = (\mu_p - \mu_q)^T(\Sigma_p\Sigma_q)^{-1}(\mu_p - \mu_q)$$

p, q は 2 つの正規分布を，μ, Σ は正規分布の平均ベクトルと共分散行列を表す．非対称性を持つ KL と MD については，平均化処理をしている．

表 1 に 4 つの距離定義間での単音間距離の相関を，表 2 に話者間距離の相関を示す．局所的な単音間距離については定義による違いがあるものの，累積した話者間距離ではその差が見られなくなっている．

表 2 の結果から，話者間距離算出において BD, HD, KL には相関が 1 に近く類似した性質があるように見られるが，MD については他の距離定義との相関が少し下がっている．

4　単音 HMM における話者依存性についての検証

[4] の参照距離の計算は全て一人の音声学者による発話音声に基づいている．そのため単音 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．ここでは，新たにもう一人の男性の音声学者 (以下，P02 と称する) の単音発声を収録し，P01, P02 それぞれの音声を用いた 2 つの話者間発音距離がどのくらいの相関を持つのかを調べ，話者依存性の影響を検証した．

収録では [4] と同様の 153 種類の単音について，各 20 回ずつ発声させ，この録音データから P01, P02 それぞれで単音 HMM を構築した．単音間距離としては，各単音モデルの MFCC12 次元と ΔMFFC 12 次元の計 24 次元を用いて分布間距離を求めた．これと DTW により，2 つの話者間発音距離を計算した．

Table 3　Corr. of phone-to-phone distances between P01 and P02, separately shown for each distance metric.

	BD	HD	KL	MD
P01-P02	0.66	0.86	0.56	0.29

Table 4　Corr. of speaker-to-speaker distances between P01 and P02, separately shown for each distance metric.

	BD	HD	KL	MD
P01-P02	0.98	1.00	0.99	0.90

表 3 に 4 つの距離定義それぞれで求めた P01, P02 の単音間距離の相関を，表 4 に話者間距離の相関を示す．単音間距離の相関は距離定義によっては低くなっているが，話者間距離については，どの場合も高い相関が得られている．単音 HMM の話者依存性の発音距離計算に対する影響は十分に小さいと言える．

表 4 の結果によると，話者間距離算出において，BD, HD, KL についてはほぼ同じくらいの相関が出ているのに対し，MD のみ相関が 0.1 ほど低いものとなっていて，若干の傾向の違いが観察できる．

5　Triphone を利用した距離の有効性

単音 HMM の構築には monophone を用いているが，これは単音間距離（DTW における局所スコア）が前後の単音の並びに関わらず計算されることになる．しかし，IPA 書き起しの中では，無声音の直後の単音に無声化を表す装飾記号が付けられやすいなど，修飾記号の付与は前後の文脈に依存している．この場合，単音は，biphone や triphone など単音の並び方を考慮した HMM を用いる方が厳密なモデル化が可能であると考えられる．

単音は本研究で扱うだけでも 153 種類あり，triphone を作成しようとすると，あらゆる単音の並び方を考慮して収録する必要があり，その作業量は膨大なものとなってしまう．ここでは代わりに，米語音素 triphone を構築し，米語音素による各話者の発音書き起しを用いて話者間発音距離の計算を行う．米語音素の種類数は CMU 発音辞書 [12] を参照すると 39 種類である．書き起しの米語音素系列への変換は，IPA から米語音素への変換表を用いて行う．

triphone の構築では，音素決定木を用いた状態クラスタリングにより状態の共有を行う． DTW の局所スコアとして用いる音素間距離行列は，状態共有後の全物理モデル間で計算する．米語音素 triphone の性能と比較するため，米語音素 monophone も構築し，別に音素間距離を求めた．

米語音素 HMM は English Read by Japanese (ERJ) コーパス [13] の男性母語話者 M08 の発話より構築した． HMM 構築における音響分析条件を表 5 に示す．状態共有によって，物理モデル数が

Table 5　Acoustic analysis conditions

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

Table 6　Corr. of speaker-to-speaker distances between P01's phone HMM, M08's monophone and triphone HMMs.

	P01	Tri4	Tri2	Mono
P01	1.00			
Tri4	0.88	1.00		
Tri2	0.88	0.98	1.00	
Mono	0.87	0.98	1.00	1.00

約 100 (Tri2) と約 10,000 (Tri4) の 2 種類の triphone を作成した．

米語音素 monophone HMM (Mono), Tri2, Tri4 を用いて算出した話者間発音距離と，P01 単音音響モデルを用いた場合の話者間発音距離との相関を表 6 に示す．分布間距離としては BD を採用している．

monophone 及び triphone 同士では話者間距離にほとんど差はなく， triphone にして音素列の前後関係を考慮することや，物理モデルを細かくすることによる違いはほとんど見られなかった．米語音素情報を利用した話者間の発音距離計算を行う場合は，最も計算コストの低い monophone を用いた音素間距離を利用すれば十分だということになる．

P01 による話者間距離と monophone による話者間距離の相関は， P01 と P02 の相関よりも有意に低くなっている．どちらの距離が参照用の距離としてより妥当であるか，第 7 節にてさらに検証を行っている．

6　状態間アライメントの必要性の検証

[4] での単音 HMM 間の距離は，対応する 3 状態それぞれの BD の平均としている．しかし，これは全ての単音間で整合のとれた比較であるとは言いがたい．3 状態のうち初めと終りの状態は前後の音声セグメントとの境界が曖昧であるという理由から，中心状態のみがより正確な特徴量分布とみなし，その BD のみをモデル間距離とすることもできる．またこの他にも，[14] で試みられているように，モンテカルロ法 (Monte Carlo, MC) を用いて状態間のアライメントを取ることも考えられる．ある 2 つの単音モデル間の距離を計算する場合，シミュレーション毎に HMM の遷移確率に従って 2 つの状態系列をランダムに生成し，状態系列間で DP マッチングをとり最小距離を求める (図 1)．シミュレーションを繰り返し，得られた距離の平均をモデル間距離として採用する．

P01　音声のモデルを用いた単音間距離，または M08 音声の monophone モデルを用いた音素間距離を使い，それぞれ分布間距離を BD, MD とした場合で，状態間アライメントを考慮した場合の寄与を

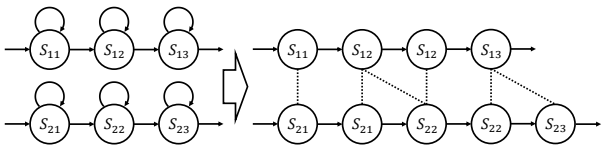

Fig. 1　An example of state alignment using Monte-Carlo.

Table 7　Corr. of segment-to-segment distances between different state alignment techniques, using two speakers and two distance metrics. (B:BD, M:MD)

	M08 B	M08 M	P01 B	P01 M
Center-MC	0.93	0.77	0.75	0.50
Mean-MC	0.97	0.93	0.93	0.72
Center-Mean	0.98	0.90	0.83	0.94

Table 8　Corr. of speaker-to-speaker distances between different state alignment techniques, using two speakers and two distance metrics. (B:BD, M:MD)

	M08 B	M08 M	P01 B	P01 M
Center-MC	0.996	0.975	0.995	0.949
Mean-MC	0.998	0.992	0.997	0.967
Center-Mean	0.999	0.988	0.997	0.993

調べた．モデル間距離を， 3 状態の平均 (Mean) とした場合，中心状態間の距離のみ (Center) とした場合， MC により計算 (MC) した場合について，音声セグメント間距離と話者間距離を算出した．なお，ここでの MC のシミュレーション計算の終了条件は，100 回シミュレーションする毎に平均値を比べ差が 0.1% 以下となった時，としている．

表 7 に各状態間アライメント手法における音声セグメント間距離の相関を，表 8 に話者間距離の相関を示す．音声セグメント間距離の相関の傾向は話者（単音または音素），距離定義の組み合わせにより異なっている．話者間距離に関しては，組み合わせや状態アライメントの手法による相違は見られない．

7　主観実験との比較及び可視化

前節までは， [4] で行われている話者間発音距離の計算方法について，分析条件を変え違いを観察し，各条件の発音距離計算における影響を検証してきた．最後に本節にて，提案された手法により求めた発音距離の妥当性を検証する．

発音距離の妥当性は，例えば専門家によって定量的に定義された二話者間の発音距離との整合性によって評価されるが，そのような実験環境を整えることは難しい．そこで本研究では，[9, 10] で行なわれているように，任意の話者 X の発音と母語話者発音との発音距離を，話者 X の発音に対する “母語話者らしさ” の主観的評定値と比較する．なお，母語話者らしさが正しく自動推定できることは，適切な発音距離推定の必要条件でしかない．

ある距離定義における SAA 話者の母語話者らしさは，その話者と米語母語話者 115 人との発音距離の平均として数値化する．主観実験では，各音声を米語母語話者 1,143 名に聴かせ，それぞれの音声が

Table 9 Corr. between automatically predicted scores and subject native-likeness scores.

	P01	M08
HD	**-0.81**	-0.77
MD	-0.72	-0.72

	Baseline
PMI	-0.77
NDL	-0.75

どれだけ母語話者に近い発音をしているかをスコア付けさせている [9]．主観実験で使用された SAA 話者 286 人を対象に，母語話者らしさを計算した．

先行研究では， PMI を用いた場合の相関は −0.77[9], NDL を用いた場合の相関は −0.75[10] となっている．これらの値を比較のためのベースラインとする．P01 音声のモデルを用いた単音間距離，または M08 音声の monophone モデルを用いた音素間距離を使い，分布間距離を HD, MD とした場合で，発音距離を算出した．この発音距離から各話者の母語話者らしさを求め，主観実験との相関を計算した．

表 9 に結果をまとめる．主観実験との相関は，MD よりも HD を用いた場合の方が高い．第 3 節，及び第 4 節にて BD, HD, KL と MD とでは話者間発音距離算出において若干の性質の違いが見られたが，ここでの結果から，発音距離の問題については BD, HD, KL の定義の方が適していると言える．また M08 による音素間距離を用いるよりも P01 による単音間距離を用いた場合の方が高くなった．単音から音素への変換は抽象化であり，音素での書き起しは音声学的情報の一部は消失されることになる．その一方で， Tri2, Tri4 では文脈依存性という形でその情報を明示的に組み込んでいる．実験結果としては，P01 の方が，僅かではあるがより適切な距離定義を提供できることが示された．

また P01 HD の相関は，先行研究での PMI, NDL を用いた結果よりも少し高いものとなっている．PMI, NDL は知覚的な根拠のある距離定義である．本研究で提案した音響モデルに基づく発音距離で，それらの距離と同等かそれ以上の評価となる結果が得られ，本研究での距離が参照距離としてある程度の妥当性があることが示された．

最後に， P01 HD で求めた話者間の発音距離を MDS により可視化した結果を図 2 に示す．各点は英語話者に相当し，それぞれを母語話者と非母語話者に二値化して，表している．母語話者と非母語話者が二分されているのが容易に見て取れる．

8 おわりに

本稿では， [4] における，音響モデルを利用した話者間の発音距離算出手法について，分析条件を変え，様々な別の距離定義を提案し，比較検討した．この結果，どの定義においても，発音距離に関してはほとんど同じ値が算出された．また，英語話者の発音の母語話者らしさのスコア付けにおいて，音響モデルによる手法は人の主観による評価と高い相関があること

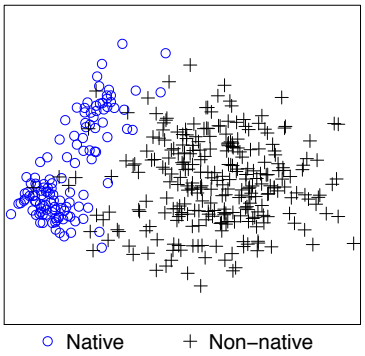

Fig. 2 An example of MDS-based visualization of WE pronunciation diversity.

が明らかになった．これらの結果より，先行研究で提案された発音距離が参照用の距離としてある程度妥当なものであることが示された．

本研究での話者間発音距離算出において，分布間距離などの定義の仕方により若干の特性の違いが見られた．それにも関わらず，発音距離では定義毎に大きな差が見られなかった．今後はこれらのことについて，詳細な検討を試みる．

参考文献

[1] J. Flowerdew, “Research of relevance to second language lecture comprehension: An overview”, Academic listening: Research perspectives, pp.7-29, 1994.

[2] B. Kachru, *et al.*, “The handbook of world Englishes”, Wiley-Blackwell, 2009.

[3] J. Jenkins, “World Englishes: A resource book for students”, Routledge, 2003.

[4] H. Shen, *et al.*, “Automatic pronunciation clustering using a World English archive and pronunciation structure analysis”, IEEE Workshop on Automatic Speech Recognition and Understanding, pp.222-227, 2013.

[5] S. Weinberger, Speech Accent Archive, `http://accent.gmu.edu`.

[6] J. Nerbonne, *et al.*, “Measuring dialect distance phonetically”, Proceedings of the Third Meeting of the ACL Special Interest Group in Computational Phonology (SIGPHON), 1997.

[7] H. L. Somers, “Similarity metrics for aligning children’s articulation data” , Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, pp.1227-1232, 1998.

[8] G. Kondrak, “A new algorithm for the alignment of phonetic sequences”, Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference, pp.288-295, 2000.

[9] M. Wieling, *et al.*, “Measuring foreign accent strength in English. Validating Levenshtein Distance as a Measure”, The Mind Research Repository, 2014.

[10] M. Wieling, *et al.*, “A cognitively grounded measure of pronunciation distance”, Public Library of Science (PLoS) one: e75734, 2014.

[11] J. Sooful, *et al.*, “Comparison of acoustic distance measures for automatic cross-language phoneme mapping”, The 7th International Conference on Spoken Language Processing (ICSLP), pp.521-524, 2002.

[12] The CMU pronunciation dictionary, `http://www.speech.cs.cmu.edu/cgi-bin/cmudict`.

[13] N. Minematsu, *et al.*, “Development of english speech database read by japanese to support CALL research”, The 18th International Congress on Acoustics, pp.557-560. 2004.

[14] H. You *et al.*, “A statistical acoustic confusability metric between hidden markov models”, International Conference on Acoustics, Speech and Signal Processing(ICASSP), pp.745-748, 2007.