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*

☆笠原駿、峯松信明 (東大)、沈涵平 (成功大、台湾)、牧野武彦 (中央大)、
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、様々な国で話されている。各国に英語が広まっていく中で、英語は地域毎に変化して行き、現代では世界中に多くの訛り（外国語・地方訛り）英語が存在している。近年、カチュルらが提唱する「世界諸英語」 [1] という概念を採択する教師が増えている。これは、アメリカ英語やイギリス英語も訛った英語の一種としてみなし、全ての英語は等しく正しくて等しく間違っているとするものである。世界諸英語の考え方に基づき訛りを個性とみなす立場からすれば、自分の英語が母語話者のそれと比べどれだけ間違っているかよりも、自分の英語が世界中の英語に対しどのように位置付けされるのか、を知ることの方が重要である。

本研究では、訛りの中でも発音訛りに焦点を置き、任意の英語話者間の発音距離を音声分析のみで推定することを試みる。本研究の最終的な目標は、世界中の英語話者を個人単位で分類し、世界諸英語全体を一望できる発音地図を作成することである。地図により自分と訛りが近い話者が分かれば、英語学習者は英会話を容易に行える相手を探すことが可能になる。発音地図はまた、特定の地域の訛りに聞き慣れたいという場合、その訛りの話者の音声を Web 上で探すブラウジングシステムの構築にも役立てられる。

本研究では、Speech Accent Archive (SAA) [2] が提供する、世界中の話者による特定パラグラフ読み上げ英語音声を用いて、任意の二話者間の発音距離を推定することを試みる。訛り推定において問題となるのが、性別や年齢といった話者性によって音響特徴が変動することである。この変動は、ケプストラム領域におけるアフィン変換で凡そ近似される。音声の訛りのみから距離を推定するためには、これらの変動に頑健な特徴を用いるのが望ましい。本実験では、変換不変表象である、音声の構造的表象 [3, 4, 5, 6] を特徴として用いる。

2 Speech Accent Archive

Speech Accent Archive (SAA) は、英語の読み上げ音声と、各音声に対応する IPA 書き起こしから成るコーパスである。SAA の読み上げ文と書き起こしの例を Fig.1 に示す。音声は、世界中の 1800 人以上の話者が Fig.1 の文章を読み上げたものである。書き起しは修飾記号（diacritical mark）も使われており、

> 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.
>
> [pliːz̻ kəl ðstelːʌ as hɐr tu brɪŋ diz θɪŋs wɪθ hɐr frʌm ðə stɑɹ sɪks spuːnz ʌɣ̟ fɪeʃ ðsno piːz faɪɣ̟ θɪk ðsleb̪s ʌv bluː ʧiːz æn meɪbiː eɪ snæk˺ foɹ heɹ bɹʌðə bɑb˺ wɪ̈ ɑlso nid˺ eɪ smalˠ plæstɪk˺ ðs̻neɪk æn eɪ bɪg tʰɔɪ fɹɔg˺ fɔɹ ðə kɪdz̻ ʃi kɐn ðskuːb˺ ðiːz θɪŋs mtu θriː ɹɐd˺ bægs æn ə wɪl goː mitʰ heɹ wɐnzdeɪ æd˺ d̪ə tɹeɪn ðsteɪʃən]

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

これらを考慮すると全書き起しで使われた異なりシンボル種類数は数百に上る。読み上げ文は 69 単語から構成され、CMU 発音辞書 [7] を参照すると、221 個の米語音素系列となる。

IPA 書き起こしは音声学の専門家により年齢、性別などの話者性とは無関係に作られているので、書き起こし間の差異を定量的に定義できれば、これを発音に関する基準距離として採択し、サポートベクター回帰の学習や評価において使用できる。

本研究では、このコーパスの中から、背景雑音が少なく、単語の挿入・削除のない話者の音声データのみを選んで用いている。今回使用した話者は 370 人で、発音距離を推定する話者の組み合わせ数は 68,265(=370×369/2) である。

3 IPA 書き起しを用いた基準距離

発音距離推定の評価と、提案手法におけるサポートベクター回帰の学習に用いるため、基準となる発音距離を計算する。基準距離の計算は、各話者の IPA 書き起こしを比較して行う。今回書き起こし中の単語数は全て同じなので、二つの書き起こし間の単語の対応は容易にとれる。同じ単語同士の話者間比較は、単音単位の挿入・削除・置き換えに注意して整合をとりながら行う必要がある。本研究では、この整合に Dynamic Time Warping (DTW) [8] を用いている。

初めに、DTW の計算で用いる単音間距離行列を求める。まず、本研究で使用した 370 人の発音書き起こしに出現する全単音記号を抽出し、この内の 95%に当たる 153 種類の記号を、第四著者（音声学者）に 20 回ずつ発音してもらった。この録音データで 3 状態、1 混合 HMM を学習し、単音 HMM を得る。単音間距離は、対応する 2 単音 HMM 間の状態間バタチャリヤ距離の平均で定義した。残りの 5%の IPA は、全て

*Improved prediction of pronunciation distance based on structural representation for clustering World Englishes pronunciations. by S. Kasahara, N. Minematsu (Univ. of Tokyo), H. -P. Shen (National Cheng Kung Univ., Taiwan), T. Makino (Chuo Univ.), D. Saito, and K. Hirose (Univ. of Tokyo)

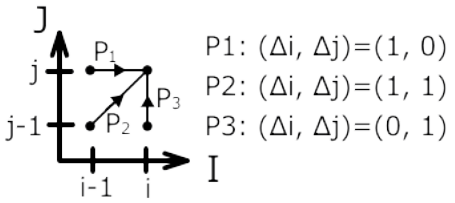

Fig. 2　Allowable path of the DTW

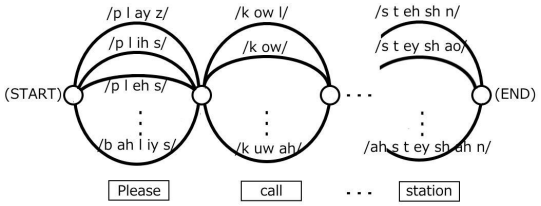

Fig. 3　word-based network grammar

修飾記号が付与されていたので、修飾記号なしの単音HMM で代用した。最終的に、153×153 の単音距離行列を用いて、DTW により単語間距離を計算する。

Fig. 2 は採択した DTW パスである。P_1, P_3 の経路は単音の挿入、削除に相当し、 P_2 は単音の一致もしくは置き換えに相当する。同単語の二つの単音系列を $a_1, ..., a_i, ..., a_I, b_1, ..., b_j, ..., b_J$ とすると、式 (1) より、(i, j) での最小累積距離 $DTW[i, j]$ が計算できる。

$$\begin{aligned} &DTW[i,j] = \\ &\min(DTW[i-1,j] + d(a_i, b_j), \\ &\quad DTW[i-1,j-1] + 2 * d(a_i, b_j), \\ &\quad DTW[i,j-1] + d(a_i, b_j)) \end{aligned} \tag{1}$$

$d(a_i, b_j)$ は a_i, b_j の単音間距離である。最終的に、$DTW[I, J]/(I + J - 1)$ が求める単語間距離である。

69 単語それぞれの単語間距離を求め、この合計を本研究における各話者間の基準距離とする。

4　ベースライン

本研究の提案手法との比較のため、ベースラインを二つ挙げる。一つは、第 3 節の発音距離算出を全て自動化したシステムで、もう一つは構造的表象を用いた距離算出の先行研究 [9] である。

4.1　自動発音書き起こしによる距離算出

第 3 節の発音距離の計算過程を再掲する。

1. 音声学の専門家による IPA 発音書き起こし
2. DTW による書き起こし比較と累積距離計算

このうち前者を、音素認識器を用いて自動化する[1]。

SAA の発音書き起こしは IPA が用いられているが、音素認識器を用いて自動書き起しを行なうため、仮に完全な音素認識の場合でも、得られるのは音素書き起しに過ぎない。単音から音素への変換は抽象化であり、この過程で音声学的情報はいくらか失われる。ここで、完全音素認識を仮定し、IPA 書き起しを規則により音素書き起しに変換し、後者を用いて発音距離算出を行うことで、完全音素認識器による距離推定の性能を評価した。DTW の計算で用いる音素間距離行列は、[10] のモノフォン HMM の音素モデル間

[1]著者らの知る限り、単音認識器は存在していない。そのため本研究においては、書き起こしの自動化はアメリカ音素認識器を代用して行っている。

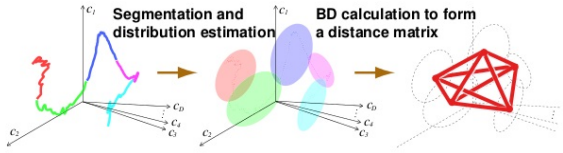

Fig. 4　Procedure of representing an utterance

のバタチャリヤ距離を使用した。完全音素認識器に基づく距離と IPA 基準距離との相関は 0.83 であった。

次に、実在の音素認識器を用いた場合の距離推定の性能を示す。認識の音響モデルとして、[10] のモノフォン HMM を初期モデルとし、370 人の全読み上げ音声を用いて追加学習したものを使用する。学習に用いるための音素ラベルファイルは、IPA 書き起しを音素化したものを用いている。得られた音素 HMM と、発音誤りを考慮した認識文法を用意することで、自動音素誤り検出が実現される。具体的な認識文法としては Fig.3 に示すように、370 人内で見られる各単語の音素系列で構築したネットワーク文法を使用する。実験の結果、得られた音素系列に対する音素正解率は 73.5%であった。また、IPA 基準距離に対する相関は 0.46 となった。発音誤り検出の精度は、発音距離推定に大きな影響を及ぼすことが分る。

4.2　構造的表象を用いた発音距離推定

発音距離推定の性能を下げる原因の一つは、話者性などの非言語的要因による音響変動であろう。精度向上には、より頑健な特徴が必要である。

性別や年齢などの話者性は、ケプストラム空間上ではアフィン変換で近似される。この変換に対し不変な特徴として、構造的表象が提唱されている [3, 4, 5, 6]。構造的表象は、ある一人の話者の音声を、絶対的な音響特性ではなく相対的な配置特性のみで捉えるもので、音声から話者性を分離して得られる特徴である。Fig.4 は発音構造を算出する過程である。入力発声（音響イベント列）中の各イベントを分布で表現し、任意の二分布 p_1, p_2 について、式 (2) の f-divergence を求める。この f-div. は任意の連続かつ可逆な写像（変換）に対して不変であることが証明されている [4]。

$$f_{div}(p_1, p_2) = \oint p_2(\mathbf{x}) g\left(\frac{p_1(\mathbf{x})}{p_2(\mathbf{x})}\right) d\mathbf{x} \tag{2}$$

$g(t)$ は $t > 0$ の凸関数で、$g(t) = \sqrt{t}$ の時 $-\log(f_{div})$ はバタチャリヤ距離（BD 距離）となる。全ての分布間の距離を求め、距離行列として発声を

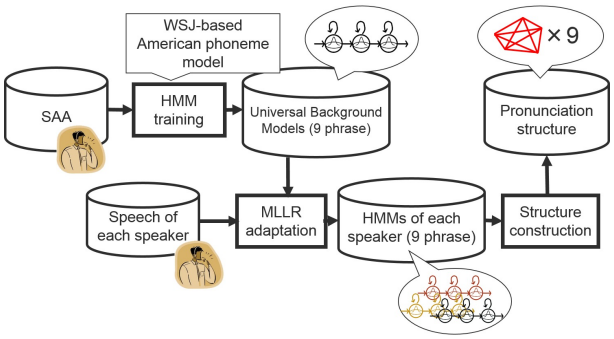

Fig. 5　Procedure to calculate the pronunciation structure [9]

表象する。この距離行列が、発音構造である。

全ての話者に同じ言語内容で発音してもらい、その音声から各話者の発音構造を算出すると、発音が同じであれば話者が違っていても構造は一致する。逆に、構造の違いがあれば、それは話者性とは切り離された訛りの違いであると考えることができる。

[9] では、381 人分の SAA の同文読み上げ音声を 9 つのフレーズに分割し、それぞれで構造的表象を算出し、9 つの発音構造を特徴として話者間の発音距離推定を試みた [9]。沈らの実験での話者構造算出までの概略図を Fig. 5 に示す。はじめに全音声を用いて Universal Background Model (UBM) となる HMM を用意する。次にクラス数 32 の MLLR 適応により 381 人の話者 HMM を作成する。これらの話者 HMM から、話者毎に、9 つの発音構造距離行列を算出する。任意話者 S と話者 T の発音構造の違いを示す特徴量として、以下の差行列 D を用いている。

$$D_{ij} = \left| \frac{S_{ij} - T_{ij}}{S_{ij} - T_{ij}} \right|, \quad \text{但し } i < j \tag{3}$$

但し、S_{ij} は話者 S における i 番目の音素と j 番目の音素の状態間 BD 距離の平均である。ここで i 番目の音素とは、HMM の先頭から 3 状態ずつを音素と仮定している。こうして得られる 9 つの差行列の全要素をサポートベクター回帰に入力し、発音距離を推定する。特徴量数は $2,804$ で、LIBSVM [11] の ϵ-SVR を用いている。カーネル関数としては放射基底関数 $K(x_1, x_2) = \exp(-\gamma|x_1 - x_2|^2)$ を適用している。

提案手法との比較のため、本研究で用いた 370 人の話者で [9] の距離推定の実験を行ったところ、基準距離との相関は 0.86 になった。

5　提案手法と実験

本研究では、沈らの行った実験で、サポートベクター回帰に入力する特徴量を変更することにより、距離推定の性能の改善を試みている。沈らは、SAA の読み上げ文を 9 つのフレーズに分けそれぞれで構造の算出を行っていたが、これでは文章全体からとり得る発音構造の距離行列のうち一部分のみしか利用していないことになる. 一つ目の実験ではこのフレーズの区切りをなくし、距離行列全体を求めてから特徴を選択する手法をとり、最終的に全発音構造を利用している。二つ目の実験では、相対的な特徴である発音構造に加えて、二話者間の対応する分布同士から距離をとった絶対的特徴を利用している。

Table 1　Acoustic analysis conditions

サンプリング	16 bit / 16 kHz
窓	25 ms length / 10 ms shift
特徴量	MFCC12 次元 + ΔMFCC
混合数	1
状態数	3

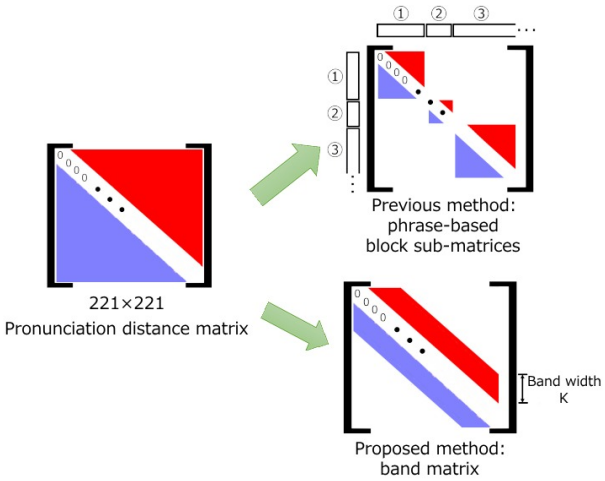

Fig. 6　Block sub-matrices and band matrix

本研究での HMM の音響分析条件を Table 1 に示す。本研究では UBM からのモデル適応は MAP 適応を用いた。サポートベクター回帰では、話者対を基準距離が大きい順に並び替え偶数番、奇数番でデータを分け、話者対を単位とした 2-fold 交差検定を行う。

[9] では差行列の計算に式 (3) を用いていた。これは線形回帰のための正規化として提案されている [5]。しかしここでの実験では距離推定にサポートベクター回帰を用いており、特徴量抽出の過程でも特徴量正規化は行われている。本研究では、正規化が二重になされることによる情報欠落が起こらないよう、$D_{ij} = |S_{ij} - T_{ij}|$ として発音距離推定を行っている。

5.1　文章を単位とした構造算出

本研究では、時間的に離れた分布間の構造も推定に有効であると考え、文章をフレーズに分割せず、全体から構造算出し距離推定を行う。文章単位で HMM を作成することで、全ての発音構造距離行列が得られ、これらを用いた差行列を入力として回帰を行う。沈らの実験は、この距離行列の内 9 つのブロック行列を用いることに相当する (Fig. 6)。本研究では代わりに、幅 K の帯行列を特徴として用いる ($K \leq 220$)。これは、SAA の文章を 221 個の音素ユニットの系列

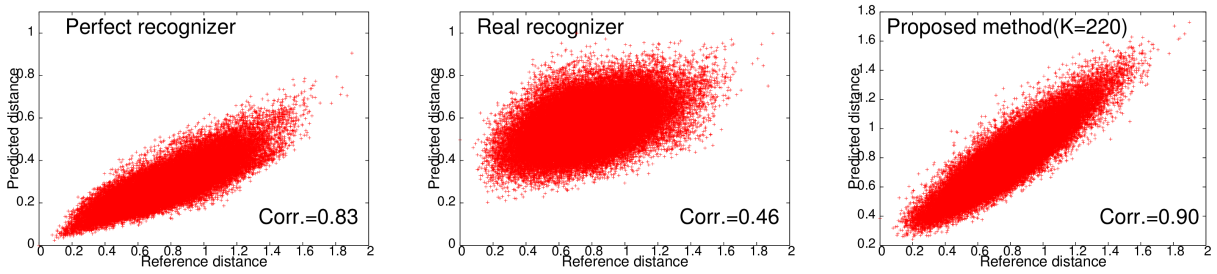

Fig. 7　Correlation between the reference distances and predicted distances

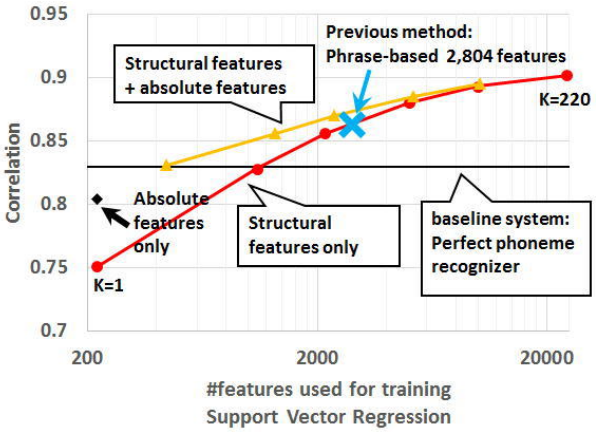

Fig. 8　Correlation improvement by increasing the band width

で考えた時に、前後 $K-1$ の幅の音素間で分布間距離をとり構造を算出することになる。

5.2　絶対的特徴の導入

発音構造の特徴に加えて、本研究ではさらに二話者間で対応する分布間の距離を直接計算することで得られる絶対的特徴を導入する。SAA の文章を 221 音素と仮定し、二話者間で対応する音素間距離を計測することで、221 次元の絶対的特徴を追加できる。

5.3　実験結果

Fig. 8 に結果を示す。幅 K を変え特徴量を増やしていくと、K が最大の 220 になるまで相関は上がり続けている。時間的に離れた発音間の相対特性も、全て訛りの推定に有効な特徴であることが示された、

絶対的特徴のみを用いた場合 (特徴量数 221) の相関は 0.80 で、発音構造でほぼ同じ特徴量数となる $K=1$(特徴量数 220) での相関よりも高いものとなった。また、発音構造の特徴と絶対的特徴を組み合わせた場合は、K が小さい時には相関の大きな向上が見られたが、K が十分大きくなるとその効果は見られなくなった。

相関は最終的に 0.90 となり、先行研究、及び理想的な完全音素認識器を超える距離推定性能を示した。

6　おわりに

本研究では、音声のみから任意の話者間の発音距離を推定する手法の改善を試みた。推定に用いる発音構造の特徴を従来手法より増やして全て用いることで、距離推定の性能は向上し、理想的な完全音素認識器を用いた場合よりも高い精度を示すことができた。

しかし今回の実験では、話者対を単位として学習・評価データを open データとして構成しているが、話者を単位として見れば、同一話者が学習・評価データに含まれることになる（話者を単位とした場合は closed データ)。距離推定の入力は話者対データであるので実験としては open 実験になっているが、入力データが本来もつ多様性を過小評価した実験的枠組みになっていると考察できる。この点については別途検討している。

参考文献

[1] B. Kachru, *et al.*, “The handbook of World Englishes”, Wiley-Blackwell, 2009.
[2] Weinberger and Steven, Speech Accent Archive. George Mason University. http://accent.gmu.edu, 2014.
[3] N. Minematsu, *et al.*, “Speech structure and its application to robust speech processing”, Journal of New Generation Computing, 28, 3, 299–319, 2010.
[4] Y. Qiao, *et al.*, “A study on invariance of f-divergence and its application to speech recognition”, IEEE Trans. on Signal Procession, 58, 7, 3884–3890, 2010.
[5] 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.
[6] 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.
[7] The CMU pronunciation dictionary, http://www.speech.cs.cmu.edu/cgi-bin/cmudict
[8] 迫江 博昭, 千葉 成美, “動的計画法を利用した音声の時間正規化に基づく連続単語認識”, 日本音響学会誌, 27, 483–490, 1971.
[9] H. -P. Shen, *et al.*, “Automatic pronunciation clustering using a world English arhcive and pronunciation structure analysis”, ASRU, 222–227, 2013.
[10] HTK Wall Street Journal Training Recipe, http://www.keithv.com/software/htk/
[11] C. -C. Chang, *et al.*, LIBSVM, a library for support vector machine, 2001.