未知話者に対する構造的発音距離推定に関する分析的検討*

☆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

各国で話される英語について、訛りは矯正されるべきものではなく話者の個性であるとする「世界諸英語」 [1] の考え方がある。この考えに基づき、世界中の多様な英語の中で自分の発音がどのように位置付けられるか一望できる発音地図の作成を目指して、先行研究にて、音声試料のみから話者間の英語発音距離を推定することを試みた [2]。音声の構造的表象 [3, 4, 5] とサポートベクター回帰を組み合わせることで、IPA 書き起こしに基づく基準発音距離との相関が 0.90 となり、高い精度で推定することができた。

しかし [2] で用いた学習・評価データは、話者対を単位とした open データセットとなっているが、話者を単位としてみると同一話者が学習・評価データに含まれ、closed となっている。[2] では二話者の発音間距離を、各話者の発音距離行列の差行列を用いて推定している。学習データ、評価データが話者対 open、話者 closed である場合、両データ間において、差行列が本来持つべき特徴量の多様性・ばらつきが過小評価されていることが危惧される。

発音距離推定の実用を想定すると、既存の書き起こしデータのみから学習したモデルで、未知の話者の音声データについても高い精度で距離推定が行える必要がある。本研究ではこの状況を想定し、サポートベクター回帰の学習と評価で話者を完全に分離し、発音構造を用いた距離推定の再評価を行った。

2 発音構造を用いた距離推定

先行研究 [2, 6] では、特定パラグラフの英語音声を世界中の英語利用者から集めている Speech Accent Archive (SAA) [7] の音声試料を用いて、話者間の発音距離を推定することを試みている。IPA 書き起こしを用いて基準となる発音距離を定義し、これを回帰モデル学習のラベルとして、また、評価時の正解ラベルとして用いている。回帰モデルとしてはサポートベクター回帰を使い、特徴量としては、性別や年齢などに対して不変量となる発音構造を利用している。

より具体的には、話者 370 人の読み上げ音声及び IPA 書き起こしを利用している。文字列 DTW に基づいて、任意の二話者間の書き起し距離を求め、これを基準発音距離とする。68,265 ($= 370 \times 369/2$) 通りある話者対を、その基準距離に基づいて昇順に並べ替え、偶数／奇数番の話者対を抽出することで、34,132 と 34,133 に二分し、学習データと評価データに分け 2-fold の交差検定を行っている。

推定距離と基準距離との相関は最大で 0.90 となり、十分な性能が得られたが、この実験では、話者対 open であるが、話者 closed な条件となっている。

2.1 サポートベクター回帰を用いた推定の問題点

N 人を対象とすると話者対数は N^2 となり、話者間の発音差異の多様性は $O(N^2)$ と考えられる。学習データが十分であれば、この多様性の元で、回帰モデルが獲得されることが期待できる。しかし以下に示すように、話者対 open、話者 closed の条件化では、発音差異の多様性が過小評価される可能性がある。

サポートベクター回帰では、入力（観測）特徴量をカーネル関数によって高次元特徴量空間に写像し、その空間で、入力特徴量と個々の学習サンプルの特徴量との内積をとる。内積値は類似度と解釈できるが、この類似度スコアを用いて回帰を行なう。評価実験において、話者 A と話者 B の距離 A-B を推定することを考える。この場合、学習データ内に話者 B と類似した話者 C が存在すると、A と C の差行列と距離 A-C が学習データに存在するため、A-B の推定はほとんど、AB の差行列と AC の差行列の内積を参照することで行なわれることになる。

学習データに A-$\{x_i\}$ ($x_i \neq B$) , B-$\{y_j\}$ ($y_j \neq A$) が含まれている時の A-B の距離推定は、$\{x_i\}$ の中に B に似た話者がいるか、$\{y_j\}$ の中に A に似た話者がいるかを探すだけの問題となる。言い換えれば距離推定の問題が、発音差異に対する $O(N)$ のだけの多様性で解けることになる。以上のように学習と評価で話者が混在している場合、学習データ数が十分ではなく発音差異の多様性（$O(N^2)$）に対して十分に汎化された回帰モデルが学習されていなくても、高い推定性能を示す可能性がある。

以上より本稿では、回帰モデルが発音差異の多様性にどれだけ対応できているのかを改めて検証する。即ち話者 open な発音距離推定の実験を行う。実験の手順は、[2] の発音距離推定実験に従う。

3 未知話者の音声に対する距離推定

3.1 発音構造を用いた未知話者音声の発音距離推定

370 人の話者を五等分し、74 人を評価用のデータ、残りの 296 人を学習用のデータとして、5-fold の交差検定を行う。学習データ中の話者対数は各々43,660 (=296×295/2) であり、[2] の場合よりも多い。

---

*An analytical study on structural estimation of pronunciation distances between unknown speakers. by S. Kasahara, N. Minematsu (The University of Tokyo), H. -P. Shen (National Cheng Kung University, Taiwan), T. Makino (Chuo University), D. Saito, and K. Hirose (The University of Tokyo)

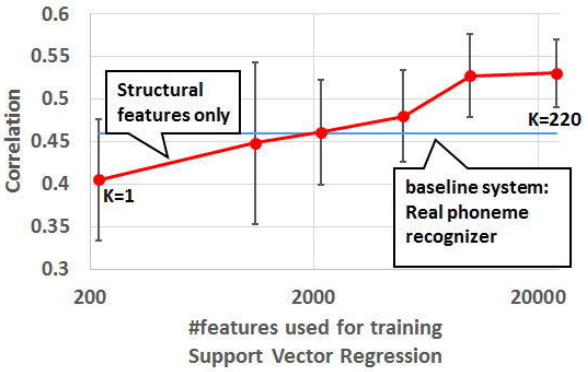

Fig. 1　Correlation improvement by increasing the band width

二話者の発音構造行列の差行列のうち、帯行列内の特徴を利用して回帰を行なった。各特徴量数における IPA 基準距離と推定距離間の相関を、Fig. 1 に示す。横軸が利用した総特徴量数であり，縦軸が相関である。未知話者に対する発音間距離推定値と基準距離との相関は、特徴量数が最大の時に最も高く 0.53 となった。相関の値は [2] の 0.90 と比べて大きく減少している。これは前節で検討した通り [2] の実験条件が、発音差異が本来もつ多様性を過小評価した条件となっていたと言える。しかし [2] でベースラインとして挙げられている、自動発音誤り検出器を通して得られる音素系列を用いた発音間距離と基準距離との相関 0.46 よりは高い値を示した。

更に、未知話者を評価に用いた場合でも特徴量数の増加に応じて、相関が単調に上がり続ける傾向は [2] 同様観測された。これは、最大数の特徴量を使った場合でも過学習は起きていないことを示唆する。差行列の特徴量を増やすことは、より長い時間離れた分布間の距離も説明変数として使用することを意味する。そして上記の結果は、時間的に最も離れた（220 音素離れている）分布間距離も、発音距離の推定に有効であることが [2] 同様，本実験でも示された。

3.2　未知話者に対する推定精度向上に向けた検討

以上の結果を受け，未知話者に対する精度向上に対する今後の検討事項について以下まとめる。

まず学習データを増加されることが考えられる。SAA は約 1800 名のデータがあり，この中から背景雑音が小さく，単語レベルでの置換，脱落が無い話者が 370 人である。SAA を使った先行研究 [8] では，この話者数を増加させるために IPA 書き起しを手動で修正し，有効データ数を増やしている。本タスクでも単純な不要語（“ah”など）挿入などは手動で削除するなどしてデータ量増加を検討する。

相対的特徴である構造的特徴以外の絶対的特徴を導入する。比較する二話者各々の話者適応 HMM を直接比較することで，絶対的な発話間距離が得られるが，これを導入する。この場合，性別や年齢の違いが悪影響すると考えられるので，話者適応学習の枠組みを導入し，仮想的な参照話者の話者性の下で二話者の発話を絶対的に比較する。

サポートベクター回帰学習におけるハイパーパラメータの最適化を行なう。サポートベクター回帰の学習は，(汎化能力を上げるために）回帰式の複雑さを抑えつつ，(学習データにおける）回帰誤差を最小化することで行なわれるが，両者はトレードオフの関係にある。前者により重きを置いた学習を行なうことも検討したい。また本稿では ϵ-SVR を用いているが，上記のトレードオフを制御した上で学習を行なう ν-SVR なども検討したい。

これら以外にもカーネル関数の設計・選択やマルチカーネル回帰の導入 [9] も検討事項であろう。

4　おわりに

本稿では [2] で採択した実験条件が、発音差異が本来持つ多様性を過小評価している可能性を指摘し、実験的にそれを示した。その結果、学習データに表れない話者に関しては、従来の自動音素誤り検出結果を利用する方法に比較すれば精度向上が見られたが，実用的には十分な精度が得られていないことが明らかとなった。しかしながら，時間的に離れた分布間距離を利用する利点については、本稿でも示すことができた。これらの結果を踏まえ，今後検討すべき事項についてまとめた。

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”, 音響学会, 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] H. -P. Shen, *et al.*, “Automatic pronunciation clustering using a world English arhcive and pronunciation structure analysis”, ASRU, 222–227, 2013.
[7] Weinberger, Steven, Speech Accent Archive. George Mason University. http://accent.gmu.edu, 2014.
[8] M. Wieling *et al.*, “A cognitively grounded measure of pronunciation distance,” *PLoS ONE*, 2013 (to appear).
[9] A. D. Dileep, *et al.*, “Representation and feature selection using multiple kernel learning”, Int. Joint Conf. Neural Networks, 717–722, 2009.