社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

世界諸英語分類のための構造的表象を用いた発音距離予測

笠原　駿†　峯松　信明†　沈　涵平††
牧野　武彦†††　齋藤　大輔†　広瀬　啓吉†

† 東京大学　〒 113–8654 東京都文京区本郷 7–3–1
†† 国立成功大学　〒 70101 台湾台南市大学路 1 号
††† 中央大学　〒 192–0393 東京都八王子市東中野 742–1
E-mail: †{kasahara,mine,dsk_saito,hirose}@gavo.t.u-tokyo.ac.jp、††† mackinaw@tamacc.chuo-u.ac.jp

あらまし　国際語として利用されている英語の様態を指し示す言葉として世界諸英語がある。発音の観点から世界諸英語を説明すれば、「英語には標準的な発音は存在せず、各国、地域、更には個人が各々異なった発音を有する」現状をそのまま受入れることを意味する。このような価値観に立てば、話者本人の英語発音が世界諸英語の中でどのように位置づけられるかを知ることは有益である。本研究では、個人を単位とした世界諸英語発音分類を念頭に置き、任意の二話者間の英語発音距離（英語訛りの違いの度合い）を、入力音声のみから自動で予測することを試みた。性別や年齢などによって音声特徴は変形するが、この変形に対し不変量となる構造的表象とサポートベクター回帰を用いて、発音距離を予測した。本稿では、回帰モデル学習を、話者対 open な学習・評価データセット、話者 open な学習・評価データセットを用いる二通りの条件下で行ない、各々の予測性能を検証した。実験の結果、話者対 open 条件では完全音素認識器を超える予測精度を持つことが示されたが、話者 open 条件では精度が低いことが示された。未知話者間の発音距離を予測するためには、さらなる改善が必要である。
キーワード　世界諸英語、発音分類、構造的表象、サポートベクター回帰、話者対 open，話者 open

Prediction of pronunciation distances based on structural representation for clustering World Englishes

Shun KASAHARA†, Nobuaki MINEMATSU†, Han-Ping SHEN††,
Takehiko MAKINO†††, Daisuke SAITO†, and Keikichi HIROSE†

† The university of Tokyo, 7–3–1 Hongo, Bunkyo-ku, Tokyo, 113–8654 Japan
†† National Cheng Kung University, 1 University Road, Tainan City, 70101 Taiwan
††† Chuo University, 742–1 Higashinakano, Hachioji-shi, Tokyo, 192–0393 Japan
E-mail: †{kasahara,mine,dsk_saito,hirose}@gavo.t.u-tokyo.ac.jp、††† mackinaw@tamacc.chuo-u.ac.jp

Abstract　The term of World Englishes is often used to indicate the current state of English as international language. It claims that English does not have the standard pronunciation and that every country, region, and even individual uses different pronunciations. From the viewpoint of World Englishes, it will be much more important to let each speaker know how his/her pronunciation is located in the diversity of World Englishes pronunciations, not how his/her pronunciation is incorrect compared to native pronunciations. This study tries to predict inter-speaker pronunciation distances only by speech analysis to examine the possibility of individual-basis pronunciation clustering of World Englishes. Speech features are often altered by non-linguistic factors such as age and gender differences. Considering this, the pronunciation structure, known as speaker-invariant feature, and support vector regression were applied for prediction. In the experiments, two conditions of a speaker-pair-open mode and a speaker-open mode were examined for training and testing the SVR. As a result, although a striking performance was obtained in the speaker-pair-open mode, only insufficient performances were found in the speaker-open mode. To predict pronunciation distances between unknown speakers, a further investigation is required.
Key words　World Englishes, pronunciation clustering, structural representation, support vector regression, speaker-pair-open, speaker-open

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、様々な国で母国語として（約 3.5 億人）、公用語として（約 4.0 億人）、あるいは外国語として（約 7.5 億人）話されている。各国に英語が広まっていく中で、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いる。発音に着眼すれば、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。このような状況を鑑み、近年、カチュルらが提唱する「世界諸英語」[1], [2] 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、英語には標準となる発音は無いと考えるものである。世界諸英語に基づき訛りを個性とみなせば、自分の英語が母語話者のそれと比べどこが間違っているかよりも、自分の英語が世界諸英語の中にどのように位置付けられるのか、を知ることの方が重要である。

本研究では、世界諸英語に見られる多様性の中でも発音訛りに焦点を置き、ある話者の英語発音と他の話者の発音との距離（発音距離）を、音声信号のみを用いて自動予測することを試みる。本研究の最終的かつ究極的な目標は、15 億人いると言われる世界中の英語話者を対象として話者間発音距離（行列）を算出し、それを用いて世界諸英語全体を分類し、その多様性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、英語学習者は会話し易い相手を探すことが可能になる。逆に訛りが大きく異なる話者を、敢えて探すことも可能になる。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) [3] 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予測することを試みる。距離予測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 [4]。英語発音の訛りのみに着眼して距離を予測するためには、これらの変動に頑健な特徴を用いるのが望ましい。本研究では変換不変表象である、音声の構造的表象 [5]〜[8] を特徴として用いる。

2. Speech Accent Archive

SAA は、英語の読み上げ音声と、各音声に対応する IPA 書き起しから成る。SAA の読み上げ文と書き起しの例を図 1 に示す。読み上げ文は、米語音素に対する被覆率を考慮して設計された文章である。69 単語から構成され、CMU 発音辞書 [9] を参照すると、221 個の米語音素系列となる。音声は、世界中の 1,800 人以上の話者が読み上げたものである。書き起しには修飾記号（diacritical mark）も使われており、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。IPA 書き起しは音声学の専門家により、話者の年齢、性別などとは無関係に作られており、書き起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、音声特徴量を用いた回帰処理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、発音距離を予測する話者の組み合わせ（話者対）数は ($_{370}C_2$ =)68,265 である。

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.

[pliːz̥ kəl ðstelːʌ as her tu brɪŋ diz θɪŋs wɪθ her frʌm ðə stɑɹ sɪks spuːnz ʌv̥ fɹɛʃ ðsno piːz faɪv̥ θɪk ðsleb̥s ʌv blu: ʧiːz æn meɪbiː eɪ snæk˺ foɹ heɹ bɹʌðɜ bɑb˺ wɨ ɑlso nid˺ eɪ smɑlˠ plæstɪk˺ ðs̥neɪk æn eɪ bɪg tʰɔɪ fɹɔg˺ fɔɹ ðə kɪdz̥ ʃi kɐn ðskuːb˺ ðiːz θɪŋs mtu θriː ɹɛd˺ bægs æn ə wi̥l goː mitʰ heɹ wɐnzdeɪ æd˺ d̥ə tɹeɪn ðsteɪʃən]

図 1　SAA の読み上げ音声と IPA 書き起しの例
Fig. 1　The elicitation paragraph used in the SAA and an example of narrow IPA transcription

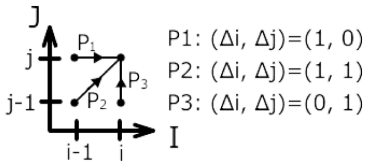

図 2　DTW において選択可能な経路
Fig. 2　Allowable paths of the DTW

3. IPA 書き起しを用いた基準距離

本研究では任意の二話者の IPA 書き起し間の差異（距離）を定量的に求め、これを基準距離として回帰処理（予測処理）の学習や評価に用いる。以下、基準距離の算出について説明する。書き起し中の単語数は話者間で同じなので、二つの書き起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（文字列を対象とした）Dynamic Time Warping (DTW) [10] を用いた。

この場合、任意の文字間距離（本研究のタスクで言うと、任意の IPA 記号距離）行列を求める必要がある。370 人の IPA 書き起しに出現する全単音記号を抽出し、この内の 95%の出現数に相当する 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。これを用いて 3 状態 1 混合の話者 closed な単音 HMM を学習した。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%の単音は全て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 の単音距離行列と DTW により、単語間距離を計算した。

図 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 $a_1, ..., a_I$, $b_1, ..., b_J$ とすると、(i, j) での最小累積距離 $DTW[i, j]$ が計算できる。

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

$d(a_i, b_j)$ 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した。

4. 音声特徴のみを用いた 2 種類の発音距離予測

本研究では音声特徴のみから、ある話者と別話者の IPA 基準距離を予測する（回帰する）。利用できる話者数が 370 人と限

定されており、以下の二種類の条件で検討することとした。回帰処理を行なう場合、学習データと評価データは open とすることが一般的であるが、話者対を単位とした open 実験と、話者を単位とした open 実験を行なった。

4.1 話者対 open 条件での発音距離予測

話者間発音距離を予測するというタスクの性質上、回帰処理に与える説明変数は主に、二話者（話者対）の発音差異に関する音声特徴となる。本条件では、学習データと評価データとの間に同一話者対が存在しないように両データを区分けする。例えば、${}_{370}C_2(= 68,265)$ だけ存在する話者対を IPA 基準距離でソートし、その偶奇でもって学習・評価データセットを用意することが考えられる。話者対 A–B が評価セットにある場合、学習セットには A–$\{x\}(x \neq B)$、B–$\{y\}(y \neq A)$ が含まれる。

本研究では回帰モデルとしてサポートベクター回帰を用いる。入力（観測）特徴量を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カーネル関数を用いてとる。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う。いま、評価セットの話者対 A–B の発音距離を予測する場合、学習セット中の話者対群 A–$\{x\}$、B–$\{y\}$ に対して、B に似た x、あるいは A に似た y が存在しているかどうかが回帰性能に影響する。

4.2 話者 open 条件での発音距離予測

学習データ、評価データに同一話者が全く含まれない、話者を単位とした openness を満たす条件下で予測実験を行なう。A–B が評価セットにある場合、学習セットには A–$\{x\}$、B–$\{y\}$ は一切含まれない。サポートベクター回帰を用いる場合、A–B の発音距離予測に対して、学習セットに話者対 A–B と似た話者対が存在しているかどうかが影響する。

いま、学習データに含まれる話者数を N とすると、話者対 open 実験では A–B に対して、A に似た話者あるいは B に似た話者が N の中に含まれるか否か、つまり、発音の多様性を $O(N)$ と評価できる枠組みでのタスク設定となっている。一方、話者 open 実験では、話者対 A–B に似た話者対が学習セットに含まれるのか否か、即ち、発音の多様性を $O(N^2)$ と評価するタスク設定となっている。話者 open 条件での回帰問題は、話者対 open 条件時よりもはるかに多くの学習データが必要となることが予想される。

4.3 両条件の実用的な意味合いについて

話者対 open 実験と話者 open 実験の実用的な価値について考察する。前者の場合、評価セットの話者対において、どちらか一方は必ず学習データに含まれることになる。このタスクは、複数話者の IPA 書き起しが用意された学習データに対して未知話者が与えられ、未知話者と学習データ中の各話者の距離を予測する問題に相当する(注1)。より具体的に考えれば、例えば、世界諸英語の分布状況を考慮してサンプリングされた各国、各地域の訛りを有する話者群の読み上げ音声と IPA 書き起しが（学習データとして）与えられた場合に、未知話者がどの話者とどのくらい離れているのかを予測する問題に相当する。

一方本研究の最終目標は、世界諸英語全体を一望できる発音地図を作成する（世界中の英語話者を個人単位で分類する）ことである。この状況を想定し、限られた話者数の学習データを用いることで、複数の未知話者間の発音距離予測がどの程度の精度で行なえるのかを実験的に検証する必要もある。

(注1)：最小距離を示す話者を結果として出力すれば、これは同定問題となる。

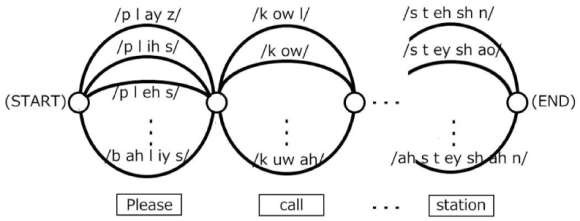

図 3　単語ネットワーク文法
Fig. 3　Word-based network grammar

以上の観点から、本研究では 2 通りの応用環境を想定し、話者対 open、話者 open の条件下で実験を行なう。IPA 書き起しを用いた基準距離を回帰の学習と評価に利用し、用意したベースラインシステムと比較して評価する。

5. ベースラインシステムとその性能

比較のため、第 3. 節の発音距離算出を全て自動化したシステムを構築し、これをベースラインとする。

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 書き起し

（2） DTW による IPA 書き起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(注2)。SAA の発音書き起しは IPA が用いられているが、音素認識器を用いると、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予測の性能を評価した。DTW の計算で用いる音素間距離行列は、[11] のモノフォン HMM の音素モデル間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発音距離と 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予測の性能を示す。音響モデルとして、[11] 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、自動音素誤り検出器が実現される。具体的な認識文法としては図 3 に示すように、370 人内で見られる各単語の発音のバリエーションをネットワーク化した文法を使用する。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。任意の二話者に対して得られる、自動音素書き起し間距離と、IPA 基準距離との相関は 0.46 となった。また、第 4. 3 節の考察に基づき、二話者の一方を完全音素認識結果とした場合（片方だけが自動音素書き起し結果）の相関を求めると、0.51 となった。発音誤り検出精度は、発音距離予測に大きな影響を及ぼすことが分る。

6. 構造的表象を用いた発音距離予測

発音距離予測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、頑健な特徴が必要である。以下、頑健な音声特徴の一つである音声の構

(注2)：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、書き起しの自動化は米語音素認識器を代用して行っている。

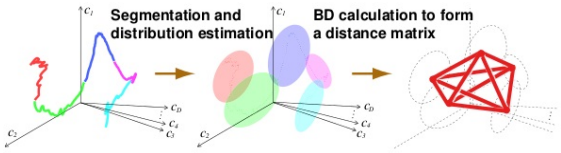

図 4　発音構造の抽出

Fig. 4　Extraction of the pronunciation structure

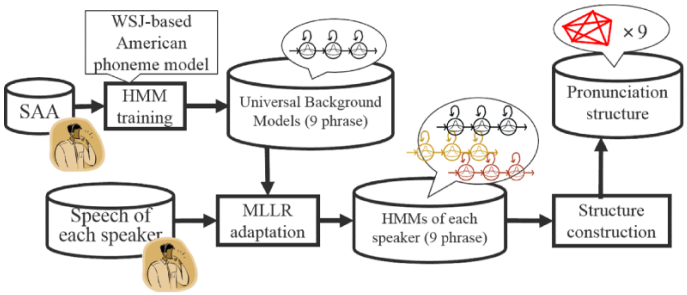

図 5　発音構造の算出手順

Fig. 5　Procedure to calculate the pronunciation structure [12]

造的表象を説明し、それを用いた先行研究 [12] について述べる。

6.1　構造的表象

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 [4]。この変換に不変な特徴として、構造的表象が提唱されている [5]〜[8]。構造的表象は、ある話者の音声中に観測される音響事象に対し、その絶対的な音響特性ではなく、事象群の相対的な配置特性のみで捉えるもので、音声から話者性（静的かつ非言語的な特性）と分離して得られる特徴である。図 4 は発音構造を算出する過程である。入力発声中の各事象を分布で表現し、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 [6]。

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

$g(t)$ 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）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。全ての話者に同一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、年齢や性別など静的な話者性として捉えることができる要因を削除した上で観測される、話者間の発音差異と考えられる [7], [8]。

6.2　発音構造を用いた話者間発音距離予測とその精度

発音構造により、ある話者の読み上げ音声は距離行列として表現される。[12] では、任意の二話者の発音構造（距離行列）に対して、その差異に相当する特徴量を定義し、これとサポートベクター回帰によって発音距離を予測している。より具体的には、SAA の読み上げ音声を 9 つのフレーズに分割し、フレーズを単位として構造的表象（距離行列）を算出し、各々の距離行列に対して、二話者間の差異に相当する特徴量を定義し、発音距離予測を試みた。[12] で行なわれた発音構造算出までの概略図を図 5 に示す。はじめに全音声を用いて Universal Background Model (UBM) となる HMM （UBM-HMM）を用意する。次にクラス数 32 の MLLR 適応により各話者の話者依存 HMM を作成する。これらの話者依存 HMM から、話者

表 1　音響分析条件

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/状態	1
状態数/音素	3

毎に、9 つの距離行列を算出する。話者 S と話者 T の発音構造差異を示す特徴量として、以下の差行列 D を用いている。

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

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 の平均である。なお、HMM はフレーズを単位として構成されるが、先頭から 3 状態ずつが音素に対応すると仮定している。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、発音距離を予測する。特徴量数は 2,804 で、LIBSVM [13] の ϵ-SVR を用いている。用いたカーネル関数は放射基底関数 $K(x_1, x_2) = \exp(-\gamma|x_1 - x_2|^2)$ である。

[12] では、話者対 open の条件での実験のみを行なっている。本研究で用いる 370 人の SAA 話者に対して、第 4.1 節で述べた方法で学習データ・評価データを用意し、2-fold の交差検定を行なった。実験の結果 0.86 の相関が得られ、完全音声認識器を用いた相関（0.83）や、話者対の片方を完全音素認識結果とした場合の相関（0.51）を越える精度が得られている。

7.　提案手法とその性能評価

本研究では、発音構造差異に相当する音声特徴量を変更することにより、距離予測の性能改善を試みる。また、話者対 open 条件のみならず、話者 open 条件での実験も行ない、その性能について実験的に検証する。

7.1　分析条件・発音差異特徴量の変更

本研究での UBM-HMM 学習時の音響分析条件を表 1 に示す。なお、MAP 適応を用いて話者依存 HMM を学習している。[12] では式 (3) を発音差異特徴としているが、この特徴は元来、発音差異を線形回帰の枠組みで予測するための特徴量として提案されている [7]。実験ではサポートベクター回帰を用いており、この場合、特徴量抽出の過程でも特徴量正規化は行われる。本研究では、二重正規化による情報欠落を避ける意味で、$D_{ij} = |S_{ij} - T_{ij}|$ を発音差異特徴量とした。

7.2　文章を単位とした構造算出

[12] では SAA の読み上げ文を 9 つのフレーズに分け、フレーズ単位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うち一部分のみしか利用できていない。本研究では、時間的に離れた分布間距離も予測に有効に働くと考え、フレーズに分割せず、文章全体から構造を算出する。

文章単位で HMM を作成することで、全体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。図 7 に、[12] で採択した算出方法と本研究のそれとの違いを模式的に示す。前者では、全体の距離行列を 9 つのブロック行列に分け、それのみを用いることに相当する。本研究では代わりに、幅 K の帯行列を特徴として用いる ($K \leqq 220$)。K が最大値を採った時、全体構造を利用することになる。なお SAA の文章を 221 個の音素系列と考えた場合、幅 K の帯行列は、最

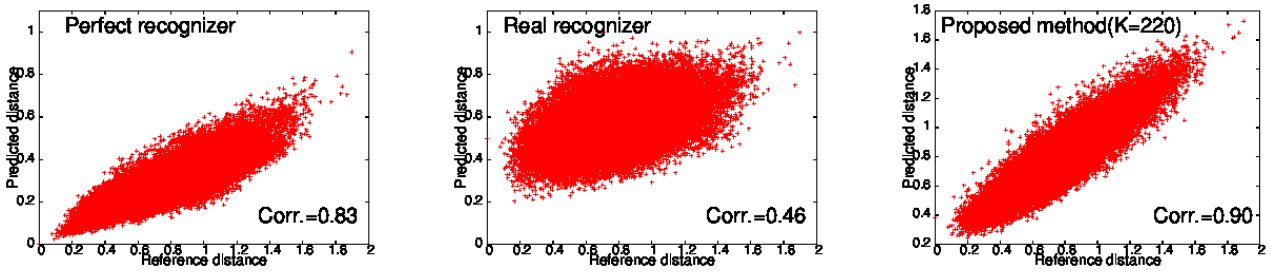

図 6　予測した発音距離と基準距離の相関図

Fig. 6　Correlations between the reference distances and predicted distances

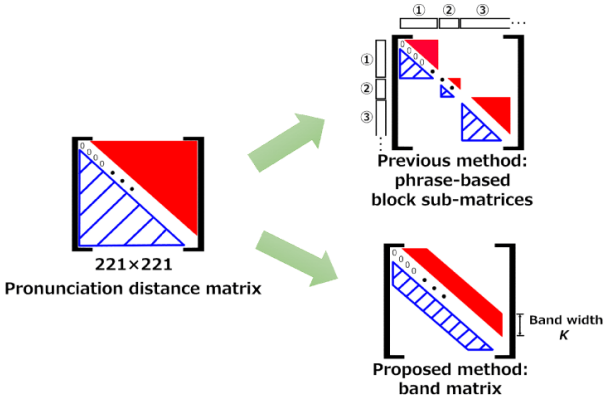

図 7　ブロック部分行列と帯行列の特徴利用

Fig. 7　Block sub-matrices and a band matrix

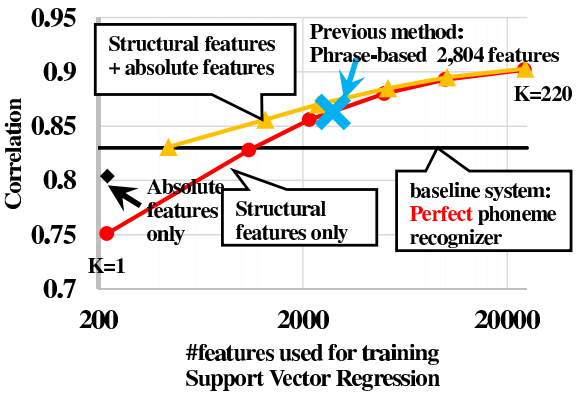

図 8　話者対 open 条件における帯幅の増加に対する相関の向上

Fig. 8　Correlation improvement by increasing the band width in a speaker-pair-open mode

長 $K-1$ 個分の音素だけ離れた分布間距離（音素間距離）を説明変数の一部として採択することを意味する。

7. 3　絶対的特徴に基づく差異特徴の追加

発音構造は音響事象間の距離のみを用いた相対的な特徴であり、[12] ではこれに基づく差異特徴のみを使って発音距離予測を行なっていた。本研究ではさらに二話者の文章 HMM に対して、対応する分布間距離をそのまま、発音差異を表現する説明変数として追加する。SAA の文章は 221 音素系列と考えられるため、対応する音素間距離（状態間距離の平均）を計測し、221 次元の（絶対的な）差異特徴を追加した。

7. 4　話者対 open 条件での発音距離予測実験結果

第 4. 1 節で述べた方法で学習データ・評価データを用意し、2-fold の交差検定を行なった。図 8 に結果を示す。幅 K 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。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。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、発音距離の予測に有効であることを示している。

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る効果は、K が小さい時には観測されたが、K が十分大きくなると見られなくなった。相関は最終的に 0.90 となり、先行研究 [12] を超える距離予測性能を示した。図 6 に、完全音素認識器を利用した相関、現実の音素誤り検出器を利用した相関、提案手法による相関の様子を示す。

7. 5　話者 open 条件での発音距離予測実験結果

学習データと評価データにおいて話者を完全に分け、話者間発音距離予測を試みた。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 の交差検定を行った。学習データ中の話者対数は各々43,660 ($=_{296}C_2$) であり、話者対 open 条件時（=34,132）よりも多い。

幅 K を変えることで利用する特徴量数を制御できるが、IPA 基準距離と予測距離間の相関を縦軸、利用した特徴量数を横軸として示したのが図 9 である。図には 5 回の実験を通して得られる相関の標準偏差も示している。相関値は話者対 open 時の実験同様、特徴量の増加に対して単調に増加する傾向にあり、特徴量数が最大の時に最大となった。時間的に離れた音素間距離の有効性は本研究でも示された。しかし相関値は 0.54 であり、話者対 open 条件の場合の 0.90 と比べて大きく減少している。但し、ベースラインシステムの相関値 0.46 よりは高い値を示した。

話者 open 条件においても、絶対的特徴を組み合わせる効果を検証した。絶対的特徴による差異特徴のみを用いた場合の相関は、0.44 となった。これに対し 5-fold 交差検定のうち、ある 1 つの学習・評価セットについて、絶対的特徴を追加した場合の効果を図 10 に示す。発音構造の特徴と絶対的特徴を組み合わせると相関の向上が見られ、話者対 open 条件の時と比較し、K が大きい時でも効果が見られた。話者正規化手法である話者適応学習 [14] を導入するなどして、絶対的特徴の定義を改良すれば、話者 open 条件における精度向上が見込める。

以上の結果より、提案手法は話者 open 条件では十分な精度を示すことは出来なかった。今後は、特徴の改善やサポートベクター回帰におけるハイパーパラメータの最適化などを行い、話者 open 条件での性能向上を検討したい。また、話者 open 条

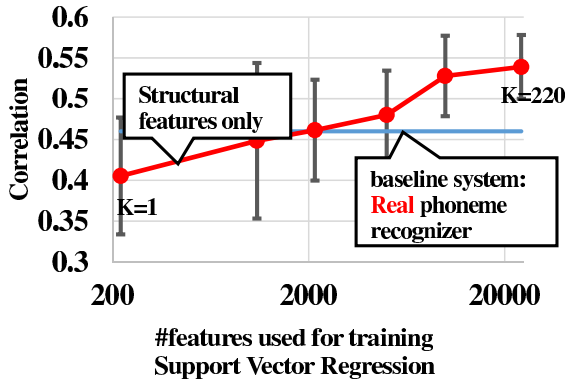

図 9　話者 open 条件における帯幅の増加に対する相関の向上

Fig. 9　Correlation improvement by increasing the band width in a speaker-open mode

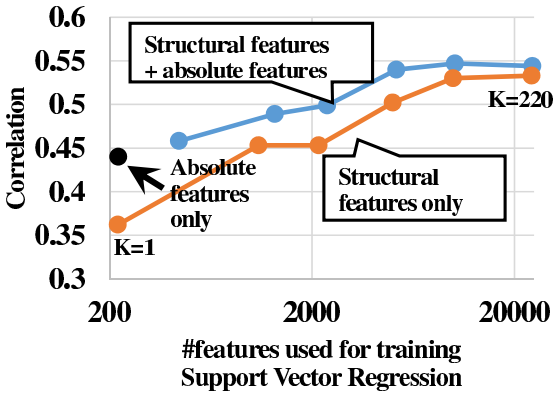

図 10　話者 open 条件における絶対的特徴の導入による相関の向上

Fig. 10　Correlation improvement by adding differential features based on absolute comparison in a speaker-open mode

件時の発音多様性に対応するためには、学習データの増加は不可欠であろう。SAA は約 1,800 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（即ち単語レベルでの言い誤りが無い）話者が 370 人である。SAA を使った先行研究 [15] 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、学習データ量を増やした上で再度検討したい。

8. 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予測する手法の改善を試みた。さらに、学習・評価データの整備に関して、話者対 open 条件、話者 open 条件の二種類の条件を考え、各々実験的に検討した。その結果、話者対 open の場合、予測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、距離予測精度は向上した。特徴量数を増やすことで性能が単調増加する様子は話者 open 時でも同様に見られた。しかし、話者 open 時の予測精度は話者対 open 時のそれを大きく下回り、学習データに現れない話者に関しては実用的な精度は得られなかった。本稿ではこれらの結果を受け、精度向上に対する今後の課題についてもまとめた。

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] S. Weinberger, *Speech Accent Archive*, 2014. http://accent.gmu.edu

[4] M. Pitz, H. Ney, “Vocal tract normalization equals linear transformation in cepstral space,” *IEEE Trans. Speech and Audio Processing*, 13, 5, 930–944, 2005.

[5] N. Minematsu, *et al.*, “Speech structure and its application to robust speech processing,” *Journal of New Generation Computing*, 28, 3, 299–319, 2010.

[6] Y. Qiao, *et al.*, “A study on invariance of f-divergence and its application to speech recognition,” *IEEE Trans. on Signal Processing*, 58, 7, 3884–3890, 2010.

[7] 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.

[8] 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.

[9] *The CMU pronouncing dictionary*, http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[10] 迫江他, “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, 9, 483–490, 1971.

[11] *HTK Wall Street Journal Training Recipe*, http://www.keithv.com/software/htk/

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

[13] C. -C. Chang, *et al.*, *LIBSVM, a library for support vector machines*, 2001.

[14] T. Anastasakos, *et al.*, “Speaker adaptive training: A maximum likelihood approach to speaker normalization,” *Proc. ICASSP*, 2, 1043–1046, 1997.

[15] M. Wieling, *et al.*,“A cognitively grounded measure of pronunciation distance,” *PLoS ONE*, 2013 (to appear).