Vol.2016-SLP-112 No.6
2016/7/29

世界諸英語発音分類を目的とした構造的特徴の不変性制御に関する検討

塩澤 文野[1,a)]　柏木 陽佑[1,†1,b)]　齋藤 大輔[2,c)]　峯松 信明[1,d)]

概要：唯一の国際共通語である英語は, 話者の言語的背景の違いにより, 様々な訛りと共に話されている (世界諸英語) . この発音多様性の可視化・地図化を目的として, 話者を単位とした発音自動クラスタリングが検討されている. ここでは, 二話者間の発音差異 (発音距離) を, 彼らの音声試料のみから定量的に推定する技術を構築している. 先行研究では, 年齢・性別などの非言語的情報を抑制することを目的として, 音声の構造的特徴を入力特徴量として用い, 回帰モデルを用いて距離を予測する方法が提案されている. しかし構造的特徴は理想環境下では, あらゆる変換に対して不変となるため, 発音の違いについても, これを無視する可能性がある. 本研究では, この不変性を適切に制御することで, 距離予測の精度向上を狙う. 先行研究において提案された, 異なるタスクにおいて検討された不変性の制御手法を本タスクに適用したところ, 次元を分割し, 特徴量を複数ストリーム化することによって, 予測精度が改善することが示された.

キーワード：世界諸英語, 発音分類, 構造的表象, f-divergence, DNN, 次元分割, サポートベクター回帰

A study of controlling the degree of invariant properties of structural features for World Englishes clustering

Shiozawa Fumiya[1,a)]　Kashiwagi Yosuke[1,†1,b)]　Saito Daisuke[2,c)]　Minematsu Nobuaki[1,d)]

1. はじめに

現在, 英語は唯一の世界共通語として認識され, 英語を母語としない多くの人も, 公用語・外国語として使用しており, それらを含めた英語の全話者数は世界人口の四分の一にものぼると言われている. 英語が世界中に広まる過程で, 話者によって文法・発音などは多様に変化している. 特に発音に着目すると, この多様性は外国語訛りや地方訛りとして現れている. 近年では英語には標準となるものを設けるべきではないという世界諸英語 (World Englishes, WE) [1] の考え方が提唱され, 英語の多様性は許容されている. WE の立場に立てば, 英語話者は自分の英語が標準からどれだけずれているのか, ではなく, 自分の英語が多様な英語の中でどのように位置づけられるのか, を考えるべきだと言える. このような背景から, 世界中の英語話者の発音訛りの多様性を, 話者単位で可視化した発音地図の作成を目的とした研究が行われてきた. その中で, 個人間の発音の違いをどのように定量化するのかは, 発音地図の妥当性を保証する意味で, 極めて重要であるといえる.

また, CALL(Computer Assisted Language Learning) や SLATE (Spoken Language Technology for Education) と呼ばれる分野では, 学習者音声を対象とした音声認識技術が研究されている [2]. 様々な学習者の存在を考えると, その英語発音には, 母語話者のものと比べより多くの異音 (allophone) が含まれることがあり, 非母語話者の発音から

[1] 東京大学 大学院工学系研究科
Graduate School of Engineering, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8656, Japan
[2] 東京大学 大学院情報理工学系研究科
Graduate School of Information Science and Technology, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
[†1] 現在, ソニー株式会社
[a)] shiozawa@gavo.t.u-toyko.ac.jp
[b)] Yosuke.Kashiwagi@jp.sony.com
[c)] dsk_saito@gavo.t.u-toyko.ac.jp
[d)] mine@gavo.t.u-toyko.ac.jp

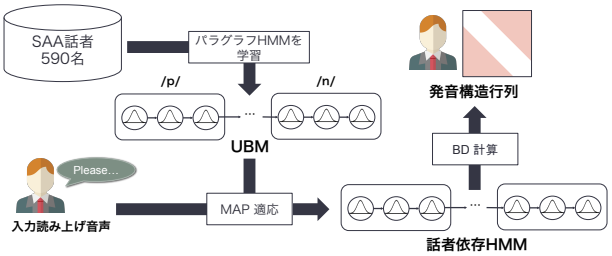

図 1　SAA 読み上げ音声を対象とした発音構造の抽出手法 [4]

抽出した音響特徴量は空間上により広く分布し, 音声認識の精度を低下させる原因の 1 つとなっている [3]. その際, これらの学習者群を発音に基づいて自動クラスタリングできれば, 事後的に形成されるクラスタ毎の音響モデルを構築することも可能となり, 認識精度の向上が見込める.

異なる話者間の音声には, 発音だけでなく話者や年齢の違いに由来する非言語的な音響差異が含まれるため, 二話者の音声試料のみから, 発音差異のみに基づく距離を推定する場合, これらの音響差異に対処する必要がある. 例えば, 全ての話者の英語発音を国際音声記号 (IPA) に基づく発音記号列に書き起こすことができれば, (書き起こし者が話者や年齢の情報を無視するため) 純粋に発音の違いのみを扱うことができる. しかし, 発音記号の書き起こしには極めて高い専門性が必要であり, 任意の話者の英語音声を即座に, かつ高精度に IPA 化することは困難である. そこで, [4] では, 音声の構造的表象と呼ばれる特徴量を利用することで, 非言語情報について不変な要素を音声から抽出し, 発音距離の推定を行っている.

構造的表象は, 音声に含まれる音響イベント群に対して, 各イベントを分布で表現し, 分布群に観測される相対的な位置関係のみを捉えたものである. しかし, 特徴量分布に, ある形状 (例えばガウス分布) を明示的に仮定すれば, その仮定が不変性に影響を及ぼす. 例えば不変性が強すぎる場合は, 本来着目したい発音差異を無視する (不変なものとして扱う) こととなり, 精度劣化を招く. 逆に, 不変性が弱すぎた場合には, 非言語情報の抑制が十分でなくなる可能性がある. 従来の研究では, 特徴量レベルで構造特徴の不変性を制御せず, (識別的な) 回帰モデルの学習プロセスの中で, これらの問題に間接的に対処していた. 本研究では特徴量抽出の段階で, 1) 識別モデルを利用して分布形状を陽に仮定せずに構造特徴量を抽出する手法と, 2) 音響空間を幾つかの部分空間に分割して特徴量抽出を行う手法を用いて, 構造特徴量の不変性を制御し, これが発音距離推定の精度に与える影響について検討する.

2. 発音距離予測に関する先行研究

[4] での発音距離予測では, 二話者に同一パラグラフを読ませ, それを IPA で書き起こし, この書き起こし間の発音距離を基準距離とし, これを音声特徴のみに基づいて予測する (回帰) . より具体的には, 音声特徴としては, 各話者から抽出された構造的特徴 (発音構造) の差行列を用い, 回帰モデルとしてはサポートベクター回帰 (Support Vector Regression, SVR) を用いている. **図 1** に, [4] で提案された発音構造の抽出手法を示す. 以下で, 図中に示す各処理について述べる.

2.1 Speech Accent Archive

[4] では, 距離予測実験に用いるデータベースとして Speech Accent Archive(SAA) を利用している [5]. これは, 69 単語から成る特定パラグラフの読み上げ音声と, 音声学者による国際音声記号 (International Phonetic Aphabet, IPA) の書き起こしが対になって提供されているコーパスである.

現在では約 2,000 人分の音声データが提供されており, そのうちおよそ 1,200 人分は専門家の手によって発音が書き起こされている. これらの音声は, 世界中のボランティア話者から提供されており, 様々な発音訛りを含む音声となっている. 一方で, 音声の収録環境は話者ごとに異なり, データによっては大きな背景雑音を含んでいたり, 読み上げ文章にない不要語を発音しているものがある. 本研究では, 比較的背景雑音レベルが低く, また, 69 単語をその語順で読み上げた話者 (590 人) を用いる.

2.2 読み上げ音声のモデル化

図 1 に, [4] で提案された, 構造特徴量を抽出する手法の概要を示す. [4] では, SAA の読み上げパラグラフを 221 の米語音素で表現し, 各話者の音声を 3 状態の音素 HMM が連結したパラグラフ HMM としてモデル化している (話者依存モデル) . 各状態の出力確率分布としては, 単一ガウス分布を仮定する. 話者依存モデルは, 全話者のデータを用いて学習した不特定話者モデル (UBM, Universal Background Model) を, 各話者に MAP 適応することで得られる.

2.3 f-divergence を利用した構造特徴量

得られた分布系列を用いて, 回帰学習の入力となる構造特徴量を算出する. この構造特徴量は, モデル化した 221 音素 HMM の各音素間距離を要素とする, 221×221 の距離行列として表現できる. この時音素 a, b 間の距離 $D_{phone}(a, b)$ は, 2 つの音素 HMM の対応する状態間の分布間距離 d の平均として, 以下の式で定義される.

$$D_{phone}(a, b) = \frac{1}{3} \sum_{i=1,2,3} d(p_{a_i}, p_{b_i}) \qquad (1)$$

ただし, p_{a_i}, p_{b_i} はそれぞれ a, b に対応した音素 HMM の i 番目の状態が持つ確率分布を表す.

ここでは, 分布間距離として f-divergence と呼ばれる距離尺度を利用する. 空間 A における確率分布 $p_1(x), p_2(x)$

について, 分布間の f-divergence は次式で定義される.

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

バタチャリヤ距離 BD は f-divergence の 1 つであり, 次式で表される.

$$BD(p_1,p_2)=-\ln\int\sqrt{p_1(\mathbf{x})p_2(\mathbf{x})}\mathrm{d}\mathbf{x} \qquad (3)$$

特に, 分布 p_1, p_2 が単一ガウス分布 $\mathcal{N}(\mu_1,\Sigma_1)$, $\mathcal{N}(\mu_2,\Sigma_2)$ であるとき, BD はパラメータ μ,Σ を用いて以下のように書き下せる. ただし, $\Sigma=\frac{\Sigma_1+\Sigma_2}{2}$ である.

$$\begin{aligned}BD(p_1,p_2)=&\frac{1}{8}(\mu_1-\mu_2)^T\Sigma^{-1}(\mu_1-\mu_2)\\&+\frac{1}{2}\ln\left(\frac{\det\Sigma}{\sqrt{\det\Sigma_1\det\Sigma_2}}\right)\end{aligned} \qquad (4)$$

さて, 空間 A を微分可能かつ可逆な変換によって射影した空間 B を考える. また, 空間 A 内の分布である p_1,p_2 に対応する, 空間 B の分布をそれぞれ P_1,P_2 とする. この時, 2 つの空間で f-divergence は等しくなる [6].

$$f_{div}(p_1,p_2)=f_{div}(P_1,P_2) \qquad (5)$$

話者の違いやマイクの歪みは, ケプストラム空間でのアフィン変換として表現できることが知られている. よって, f-divergence を要素とする構造的特徴は, これらの非言語特徴量に対して頑健であることが期待される.

回帰学習の入力には, 2 話者の発音の違いを表現する特徴量として, 各話者の距離行列の差行列を用いる.

2.4 IPA 書き起こしを用いた発音基準距離

教師データとして用いる二話者間の発音基準距離を, 読み上げ音声の IPA 書き起こし (単音記号系列) に対して, Dynamic Time Warping (DTW) により求まる最小整合コストとして定義している. DTW は各系列間の時間的な非線形な対応付けによって 2 発声の距離を求める手法であり, 次の漸化式に基づく動的計画法を用いて計算される.

$$D(i,j)=\min\begin{bmatrix}D(i,j-1)+d(i,j)\\D(i-1,j-1)+2d(i,j)\\D(i-1,j)+d(i,j)\end{bmatrix} \qquad (6)$$

i,j は着目している発音記号の各系列中でのインデックスであり, $D(i,j)$ は, 先頭の単音記号から, 対応する単音記号までの最小累積距離を表す. 局所コスト $d(i,j)$ として, ここでは単音音響モデル間の距離を用いた. 各単音に与えられる音響モデルは, 1 人の音声学者の単音読み上げによって学習された 3 状態 1 混合の HMM であり, 2 単音間の距離は対応する状態間のバタチャリヤ距離の平均として定義される.

3. 構造的表象の不変性制御

3.1 BD が示す不変性に対する考察

先行研究では, SAA の音声はパラグラフ HMM としてモデル化していた. 近年音声認識の分野では, 特徴量分布形状としてガウス分布を仮定した生成モデルである HMM から, 分布形状を仮定しない識別モデルである DNN を用いた音響モデルが主流となり, 大幅な精度向上を実現している. これは, 仮定した分布形状と真の分布形状の間にミスマッチが存在することが原因の一つと考えられている.

前節で検討した BD 計算において, 分布形状として単一ガウス分布を仮定し, 計算式を導出していた. 即ち, 先行研究では式 (4) を用いており, 個々のイベントが単一ガウス分布に常に従う, という強い仮定を置いている. 先行研究 [7] では, 識別モデルとしての DNN を効果的に利用することで, 分布形状を仮定せずに BD の推定を行ない, 言語識別性能の向上を実現している.

BD の定義式 (3) に従って, 音素 a,b 間の BD は,

$$BD(a,b)=-\ln\int\sqrt{p(\mathbf{x}|y=a)p(\mathbf{x}|y=b)}\mathrm{d}\mathbf{x} \qquad (7)$$

となる. これをベイズの定理により変形することで, 事後確率 $p(y=a|\mathbf{x}),p(y=b|\mathbf{x})$ を用いて表すことができる.

$$\begin{aligned}BD(a,b)=&-\ln\int p(\mathbf{x})\sqrt{p(y=a|\mathbf{x})p(y=b|\mathbf{x})}\mathrm{d}\mathbf{x}\\&+\frac{1}{2}\ln p(y_n=a)+\frac{1}{2}\ln p(y_n=b)\end{aligned} \qquad (8)$$

このとき特徴量系列 $\{\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_n\}$ の長さが十分であれば, 全積分を和の形で近似できる.

$$\begin{aligned}BD(a,b)=&-\ln\frac{1}{n}\sum_{i=1}^{n}\sqrt{p(y_i=a|\mathbf{x}_i)p(y_i=b|\mathbf{x}_i)}\\&+\frac{1}{2}\ln\frac{1}{n}\sum_{i=1}^{n}p(y_i=a)+\frac{1}{2}\ln\frac{1}{n}\sum_{i=1}^{n}p(y_i=b)\end{aligned} \qquad (9)$$

事後確率のラベルとなる音素状態数を n とすると, 上式に従って各音素状態間のバタチャリヤ距離を計算することで, $n\times n$ の発音構造行列を得る.

このように, 識別モデルを用いた場合, 特徴空間を十分に埋めるサンプルを用意できれば, 分布を陽に仮定する必要はなくなる. SAA パラグラフの読み上げ音声は, 各話者の十分な特徴量フレームを持つと考えられ, 本アプローチによる精度向上が期待される.

一方, 分布形状をガウス分布と仮定することのメリットもある. 仮に, 真の分布形状が常に得られれば, f-divergence は任意の微分可能な可逆変換に対して不変であるため, 構造特徴量は, 年齢, 性別, 訛りなど, あらゆる変換に対して不変になると考えられる. 本タスクでは年齢や性別の違いに対して不変で, 訛りの違いに対しては不変とならない特

徴量が必要である。即ち, 不変性を適度に制御することが必要である。ここで, 着目する音響イベントが全てガウス分布に従うことを仮定すれば, 変換前後でガウス分布に従う必要から, アフィン変換のみに不変となる. 例えばケプストラム空間では, 加算はチャネル歪みを表し, 乗算は声道長差異による歪みを表す [8] ため, これらに対して不変性を示すことになる. しかし, アフィン変換群全てに不変性を示す必要があるか否かは, チャネル歪み, 声道長歪みによる音響変化が, 全てのアフィン変換群を要求するか否かに関係する。当然, 必要となる不変性はアフィン変換の「一部」に対してのみ不変となる不変性である。

3.2 次元分割による不変性の制御

話者性の違いの要因の 1 つである声道長の違いは, ケプストラムベクトル c に変換行列 A をかける演算 Ac で近似できる [9]. A は次式のように表現できる.

$$A = \begin{pmatrix} 1 & \alpha & \alpha^2 & \cdots \\ 0 & 1-\alpha^2 & 2\alpha-3\alpha^2 & \cdots \\ 0 & -\alpha+\alpha^3 & 1-4\alpha^2+3\alpha^4 & \cdots \\ \vdots & \vdots & \vdots & \end{pmatrix} \quad (10)$$

α は $|\alpha| < 1$ なるウォーピングパラメータである. α が十分小さい場合であれば, A は対角成分とそれに隣接する要素のみが非零な行列 (帯行列) となる.

訛りの違いに対しては非不変で, 話者の違いに対しては不変性を示す特徴量は, このような帯行列を用いた変換に対してのみ不変であることが求められる. これは, もとの特徴空間を幾つかの次元毎に分割し, 各部分空間内で構造特徴量を抽出することで, 近似的に実現できることが報告されている [10]. 具体的には話者依存 HMM に対して, 低次から連続する n 個の次元で特徴量を構成し, これを用いて構造表象 (部分構造表象) を得る. 次の n 個の次元で構成される特徴量を使って, 次の部分構造表象を得る. このようにして, 特徴量ストリームを s 系列のストリームとした後に, 部分構造表象を s 種類, 得ることができる (すなわち, $s \times n$ が元の特徴量空間の次元数と一致することになる). ある変換前後で, 全体構造は不変であっても, s 個の部分構造は変化することとなり, 後者の方が不変性は弱くなる.

4. 発音距離予測の実験

4.1 予測精度の評価方法

本研究の目的である発音距離の予測について, 前節で述べた 2 つの手法を用いて特徴量を抽出し実験を行った. また比較のためのベースラインとして, 2 節で述べた手法についても実験を行った.

これらの異なる手法によって抽出した特徴量に対して, 以下に示す共通の枠組みで距離予測器の学習と評価を行う.

データには SAA の話者 590 人分の読み上げ音声と発音記号の書き起こしを使用した. 回帰はサポートベクター回帰によって行う. オープンソースライブラリである libsvm [11] の ϵ-SVR モードを使用し, カーネル関数としては放射基底関数 $K(u,v) = \exp(-\gamma|u-v|^2)$ を用いている.

590 人の SAA 話者を 4:1(=学習:評価) に分ける. まず, 学習話者 (472 人) で話者対 (111,156 話者対) を構成し回帰モデルを学習する. 評価時は, 任意の評価話者 1 人と任意の学習話者 1 人の間 (55,696 話者対) で発音距離を予測する. 以上の作業を 5 回繰り返し, 各試行で出力される予測距離と基準距離の相関によって予測の精度を評価する.

表 1 HTK の音響分析条件

サンプリング周波数	16kHz
窓長	25ms length
特徴量	MFCC 12 次元 + ΔMFCC 12 次元
混合数	1
状態数	3

4.2 ベースライン手法の実験条件

比較のためのベースラインとして, 2 節で述べた手法に従い回帰学習の入力特徴量を算出する. HMM の学習には, 公開ライブラリである HTK [12] を使用する. **表 1** にその際の音響分析条件を示す.

[4] では音素間の距離を要素とする構造特徴量を用いていたが, 今回は HMM の各状態を単位とした特徴量についても実験を行う. 2 つの音素間の距離は, 各音素 HMM が持つガウス分布間のバタチャリヤ距離の平均として定義されていた. 今回の設定では, SAA のパラグラフを 221 の音素で近似しているので, 音素間距離を用いた構造特徴量は 221×221 の距離行列になる. 一方で, 構築した HMM に含まれる 663 個の状態を単位として, 663×663 の距離行列としても表現できる. 先行研究で行われていたように, 音素間距離を求める際に状態間距離の平均を取ることは, モデルが持つ時間方向についての情報を一部無視することになる。逆に, 平均を取ること無く状態間距離をそのまま要素とした場合, 時間変化に伴う情報をより多く含んだ特徴量として (時間分解能を上げて) , 発音距離の予測を行うことに相当する. しかし, この場合得られる距離行列の次元数は, 音素を単位としたものと比べて 9 倍程度の大きさとなり, 回帰学習の際の計算コストは膨大なものになる. そこで, 計算の負荷を軽減するために, 全話者について共通な閾値を設定し, それ以下の値を持つ距離行列の差行列の要素を 0 に近似した. 閾値は計算が可能な範囲で最も小さいものを設定したが, この時平均して 8 割程度の要素が 0 に近似された.

4.3 分布を陽に仮定しない分布間距離

[7] で提案された手法に基づいて, 識別モデルを利用して音響特徴量の分布を陽に仮定せず構造特徴量を算出し, 回

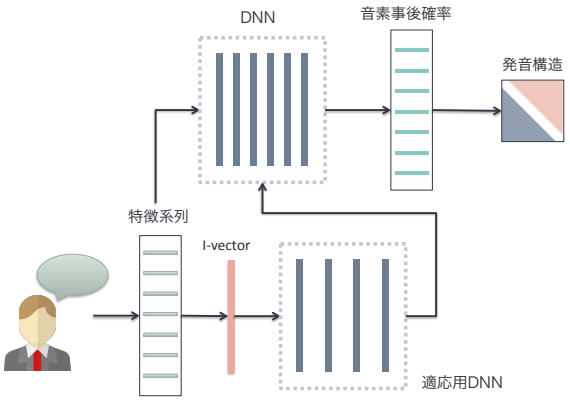

図 2　SAT-DNN を用いた発音構造の算出

表 2　DNN の構成と音響分析条件

入力特徴量	(MFCC 13 次元 $+\Delta+\Delta\Delta$) $\times$ 11 フレーム
サンプリング周波数	22.05kHz
窓長	20ms length
中間層	6 層 1024 ノード
出力状態ラベル	132

帰学習の入力特徴量として利用する.

図 2 に, 識別モデルを用いた構造特徴量の生成手法の概要を示す. 音素事後確率の推定は DNN によって行うが, その際 I-vector を用いた適応手法である SAT-DNN を利用する [13]. I-vector は言語識別においてよく用いられる特徴量であり, 英語発音の差異を扱う今回のタスクにおいても有効な情報を含むものと考えられる. SAT-DNN による適応は, 通常のニューラルネットワークの入力に, 適応用ネットワークの出力をバイアス項として足し合わせることで行う. 各ネットワークの学習は次の手順で行う. まず, メインとなる DNN を単体で学習し, 初期パラメータを決定する. 次に, このメイン DNN のパラメータを固定し, back-propagation により適応用 DNN を学習する. その後は逆に適応用 DNN のパラメータを固定し, メイン DNN のパラメータを更新する.

SAT-DNN の学習は音声認識用のツールキットである KALDI [14] を使用して行った. モデル適応に用いる I-vector は, 言語識別用データセットである, NIST LRE03, LRE05 , LRE07 によって学習した. また, メイン DNN の初期モデルの学習には, Wall Street Journal (WSJ) を使用した. **表 2** に DNN の構成とその入力の音響分析条件を示す.

学習した DNN に SAA 話者の音声を入力し, 出力として得られる音素ラベル毎の事後確率を用いて, 式 (9) に従い, 話者毎の発音距離行列を求める.

[7] では, 観測発話のデータ数が少ない場合, 特徴空間上での全積分が行えず, バタチャリヤ距離を計算出来ないという問題を挙げていた. そのため, 英語音声によって学習した UBM からサンプリングした音響特徴量を DNN の入力として用いていた. しかし, 本実験においては 1 話者の発話フレーム数は十分大きなものであり, また, 読み上げるパラグラフが共通であることから発話内容の偏りは小さくなると考えられる. そこで, 今回はサンプリングを行うこと無く, 各話者の発話から抽出した音響特徴量をそのまま DNN の入力として採用した.

表 3 に, 識別モデルによって分布を陽に仮定せずに発音構造を算出した場合の実験結果を示す. 事後確率から計算した構造特徴量のみを回帰に用いた場合, ベースラインと比較して大きく精度が下がった. このことから I-vector を用いたモデルの適応によってのみ発音訛りを表現することは難しいと言える. しかし, これを従来の構造特徴量に連結したものを入力特徴量とした場合, 僅かながら精度の改善が見られた.

4.4　次元分割

ここでは, 3.2 節での議論に基づき, 音響特徴量空間を分割することで, 複数ストリームの構造特徴量を回帰学習の入力として利用することを考える.

ベースラインと同様の方法で各話者の音声を HMM としてモデル化した後, HMM の各状態の持つガウス分布の平均ベクトル・分散共分散行列を s 個に分割する. 分割した部分空間上での発音構造は, ${}_{221}\mathrm{C}_2 = 24,310$ 個の音素間距離を要素として持つ距離行列で表現される. よって, 元の特徴量空間を s 個の部分空間に分割した場合, 各部分空間上で求めた発音構造特徴量を連結したものは $24,310 \times s$ 次元のベクトルとなる. f-divergence の変換不変性に対する制約は, 音響特徴空間を, より細かい部分空間に分割するほど強まると考えられる.

回帰学習の入力には, このベクトルの各成分の絶対値をとった後, しきい値以下の成分を 0 としたスパースなベクトルを用いる.

表 4 に, 次元分割を用いて, ストリーム数を 1,2,4 とした場合の実験結果を示す. 次元分割を行い, 特徴量のストリーム数を上げることで, スパース化による情報の損失を加味しても, 距離予測の精度が向上することが読み取れる.

5.　おわりに

本稿では, 英語発音間の距離予測の精度を向上させる目的で, 予測に用いる発音の構造特徴量の抽出手法について, 実験的な検討を行った.

今回は, 分布を陽に仮定せずに識別的なモデルを用いて構造特徴量を算出することで, 構造特徴量の変換不変性への制約を弱める手法と, 音響特徴量空間を部分空間に分割することで不変性に対する制約を強める手法との 2 つを用いて, 発音距離の予測の精度に与える影響について実験を行った. 結果として, 次元分割を行うことで距離予測の精度が改善することが示された.

表 3　識別モデルによって得られた特徴量での実験結果

特徴量	構造特徴量の次元数	set1	set2	set3	set4	set5	**Ave.**
ベースライン (音素間距離)	24,310	0.716	0.739	0.711	0.718	0.708	**0.718**
ベースライン (状態間距離)	219,453	0.712	0.736	0.709	0.721	0.740	**0.723**
DNN posterior	8,646	0.640	0.648	0.619	0.625	0.616	**0.630**
ベースライン (音素間距離) + DNN posterior	32,956	0.720	0.744	0.716	0.717	0.712	**0.722**

表 4　次元分割を行った特徴量での実験結果

ストリーム数	構造特徴量の次元数	set1	set2	set3	set4	set5	**Ave.**
1	24,310	0.716	0.739	0.711	0.718	0.708	**0.718**
2	48,620	0.717	0.733	0.710	0.734	0.730	**0.722**
4	97,240	0.720	0.746	0.705	0.736	0.730	**0.730**

今回の実験では, 次元分割に伴い特徴量の次元が増え, 計算のコストが増大する問題に対して, ある共通のしきい値以下の成分を 0 に近似することで対処した. この近似を行うことで, 距離予測の精度が劣化していることが考えられる. 今後は, 情報のロスの少ない次元圧縮手法について, より詳細な検討を行いたい. 具体的には, 発音構造を 1 つのグラフとみなすことで, グラフ理論に基づく手法を適用することを考えている.

謝辞　本研究は JSPS 科研費 JP26240022 および MEXT 科研費 JP26118002 の助成を受けた.

参考文献

[1] Kachru, B., Kachru, Y. and Nelson, C.: *The handbook of World Englishes*, Wiley Blackwell (2009).
[2] Eskenazi, M.: An overview of spoken language technology for education, *Speech Communication*, Vol. 51, No. 10, pp. 832–844 (2009).
[3] Tao, J., Ghaffarzadegan, S., Chen, L. and Zechner, K.: Exploring deep learning architectures for automatically grading non-native spontaneous speech, *Proc. of the IEEE ICASSP*, pp. 6140–6144 (2016).
[4] 笠原駿，峯松信明，　沈涵平，牧野武彦，齋藤大輔，広瀬啓吉：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，日本音響学会春季講演論文集，pp. 121–122 (2014).
[5] Weinberger, S. H. and Kunath, S. A.: The Speech Accent Archive: towards a typology of English accents, *Corpus-based Studies in Language Use, Language Learning, and Language Documentation*, Brill, pp. 265–281 (2011).
[6] Qiao, Y. and Minematsu, N.: A study on invariance of f-divergence and its application to speech recognition, *IEEE Transactions on Signal Processing*, Vol. 58, No. 7, pp. 3884–3890 (2010).
[7] 柏木陽佑，　張聡穎，齋藤大輔，峯松信明：識別的アプローチによる分布間距離推定の検討とその言語識別への応用，電子情報通信学会音声研究会資料，SP2015-38，pp. 77–82 (2015).
[8] Pitz, M. and Ney, H.: Vocal tract normalization equals linear transformation in cepstral space, *IEEE Transactions on Speech and Audio Processing*, Vol. 13, No. 5, pp. 930–944 (2005).
[9] 江森正，篠田浩一：音声認識のための高速最ゆう推定を用いた声道長正規化，電子情報通信学会論文誌 D，Vol. 83, No. 11, pp. 2108–2117 (2000).
[10] 朝川智，　喬宇，峯松信明，広瀬啓吉ほか：音声の構造的表象と判別分析を用いた単語音声認識，電子情報通信学会技術研究報告, SP2008-113，pp. 203–208 (2008).
[11] Chang, C.-C. and Lin, C.-J.: LIBSVM: A library for support vector machines, *ACM Transactions on Intelligent Systems and Technology*, Vol. 2, pp. 27:1–27:27 (2011).
[12] Young, S. J. and Young, S.: *The HTK hidden Markov model toolkit: Design and philosophy*, University of Cambridge, Department of Engineering (1993).
[13] Miao, Y., Jiang, L., Zhang, H. and Metze, F.: Improvements to speaker adaptive training of deep neural networks, *Spoken Language Technology Workshop (SLT), 2014 IEEE*, IEEE, pp. 165–170 (2014).
[14] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P. et al.: The Kaldi speech recognition toolkit, *IEEE 2011 workshop on automatic speech recognition and understanding*, No. EPFL-CONF-192584, IEEE Signal Processing Society (2011).