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.

自己視点からの世界諸英語クラスタリングを目的とした発音距離予測とその耐雑音性に関する検討

佐藤惟知† 柏木陽佑† 笠原駿† 峯松信明† 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

あらまし 近年，諸外国から日本を訪れる観光客が増加している．また，2020 年には東京オリンピックが行なわれる．彼らとのコミュニケーションは基本，英語となるが，当然様々に訛った英語を話す話者を相手にする必要がある．世界中の様々な英語発音（世界諸英語）に慣れ親しむことを目的として，世界諸英語の発音を話者を単位として自動分類し，可視化する技術を検討している．このためには，任意の二話者間の発音差異を定量的に予測する必要がある．本研究ではこれを，音声の構造的表象に基づく特徴抽出とサポートベクター回帰により実装している．本稿では，1) 自己視点からの可視化を想定した発音距離予測と，2) 発音距離予測における雑音抑制技術の有効性という 2 点に着目して実験的検討を行なった．その結果，音素書き起しを使った発音差異計算に相当する精度が得られ，また，10[dB] ほどの SN 比があれば，十分な雑音抑制が可能であることが示された．

キーワード 世界諸英語，発音分類，構造的表象，サポートベクター回帰，自己視点からの可視化，雑音抑制，DNN

Noise-robust Prediction of Pronunciation Distances Aiming at Clustering of World Englishes Using a Learner's Self-centered Viewpoint

Yuichi SATO†, Yosuke KASHIWAGI†, Shun KASAHARA†, Nobuaki MINEMATSU†,

Daisuke SAITO†, and Keikichi HIROSE†

† The University of Tokyo，7-3-1 Hongo，Bunkyo–ku，Tokyo，113-8654 Japan

Abstract In recent years，we have more and more international tourists and in 2020, we have Tokyo Olympic Games. For communicating with those tourists, the default language is English but they speak English with various accents. To realize smooth communication with these tourists, we are developing a technical infrastructure to accustom Japanese people to variously accented Englishes (World Englishes). The infrastructure aims at clustering a large diversity of English pronunciations on an individual basis and visualizing the diversity in an educationally effective way. For clustering, a technique is needed that can predict the accent gap between any speaker pair and we developed it by integrating pronunciation structure analysis and support vector regression. In this paper, the prediction performance is evaluated when the prediction technique is applied for visualization using a user's self–centered viewpoint and when it is applied with a noise suppression technique. Results show that the performance is comparable to that observed when we use phonemic, not phonetic, transcripts and that 10 [dB] is enough as SNR to guarantee the prediction performance realized in a clean condition.

Key words World Englishes, pronunciation clustering, structural representation，support vector regression, self-centered visualization, noise suppression, DNN

1. はじめに

英語は約 80 の国・地域で話されており，世界で最も広く使用されている言語である．英語を母国語としている話者はおよそ 3.5 億人おり，また公用語・外国語とする話者も含めればおよそ 15 億人にものぼる．このように広く用いられている英語はその普及の過程で文法，語用，綴り，発音など様々な面で変化してきた．発音に着目すればこの変化は多様な外国語訛りや

地方訛りという形で現れている．

一方，通常の学校教育における英語授業では米語 (特に General American，GA) や英語 (特に Received Pronunciation，RP) をモデル発音とすることが多いが，多様な発音が存在する現代において，実際の英会話相手は GA 話者，RP 話者ばかりではない．加えて近年では，Kachru らの提唱する「World Englishes (世界諸英語) 」[1] という概念を採択する英語教師が増えてきている．これは，GA や RP を「標準的な発音」としてみなすのではなく，これらも訛った英語の一種とみなし，発音が多様化した現状をそのまま受け入れる考え方である．

様々な訛りの英語話者とのコミュニケーション能力を向上させようとした場合に，世界にはどれほど多様な英語発音が存在するか，そして自分はその中でどのように位置しているのかを知ることが重要になると考えられる．このように世界諸英語を俯瞰する場合，様々な訛りの英語話者群アーカイブに対して自身の英語読み上げ音声を入力し，アーカイブ話者群と自身とを地図化できると良いだろう．最近では TED [2] のように，様々な訛りの英語音声資料に手軽にアクセスできる手段も増えてきており，地図化技術が確立すれば，これらの英語講演アーカイブを用いた，訛りに着眼した世界諸英語ブラウザが構築できる．実際に筆者らの先行研究 [3] では，ユーザー自身を中心とし，その周りにアーカイブ内の話者を図 1 のように配置・地図化する可視化手法が提案されている．図 1 は自身が原点に配置され，自身と他者との距離が発音距離に相当する．上半円が同性，下半円が異性である.x 軸負方向から x 軸正方向への角度が年齢を示している．2020 年の東京オリンピックでは，世界中から観光客が押し寄せることが予想される．彼らの話す様々な英語に対応できるよう，ホテルやレストランの従業員向けに，世界諸英語を教える教材開発なども行なわれており [4]，このようなケースにおいて図 1 のような世界諸英語ブラウザが有用であると考えられる．

しかし想定するシステムは，各自の環境で英語音声を収録する必要がある関係上，雑音に対して頑健であることが望ましい．従って本研究では [3] の可視化を想定した場合の発音距離予測精度を実験的に確認した上で，雑音への頑健性を向上させる目的で Deep Denoising Auto-Encoder(DDAE) [5] を導入することの効果についても，実験的に検討した．

2. 発音距離予測

筆者らの先行研究 [6] では，英語パラグラフ読み上げ音声コーパス Speech Accent Archive (SAA) [7] を用いて，任意の 2 話者間の発音距離を入力音声信号のみから予測することを試みている．本研究の発音距離予測の枠組みは [6] と同様である．

2.1 Speech Accent Archive

SAA は，図 2 に示す特定英語パラグラフの読み上げ音声と，各音声サンプルに対する国際音声記号（IPA）を用いた発音書き起しが提供されているコーパスである．図 3 に書き起し例を示す．書き起しは音声学を専門とする者らの手作業により，装飾記号 (diacritical mark) を用いて詳細になされている．パラグラフは米語音素及び米語音素対の被覆率が高くなることを考

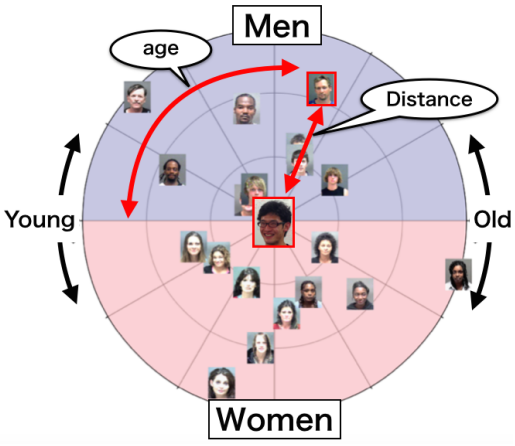

図 1　ユーザーを中心にした世界諸英語話者の地図化

Fig. 1　World Englishes Map Using a Learner's Self-centered Viewpoint

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.

図 2　SAA 読み上げ文章

Fig. 2　SAA paragraph

[pliːˌz kəl ˈstelːʌ as her tu brɪŋ diz θɪŋs wɪθ her frʌm ðə stɑɪ sɪks spuːnz ʌˌ fɪɛʃ ˈsno piːz farˌ θɪk ˈsleˌbs ʌv bluː ʧiːz æn meɪbiː eɪ snækˀ foɪ heɪ bɪʌðɜ bɑbˀ wɪˌ ɑlso nidˀ eɪ smalˠ plæstɪkˀ ˈsˌneɪk æn eɪ bɪg tʰɔɪ fɹɔgˀ fɔɹ ðə kɪdˌz ʃi kɛn ˈsku:bˀ ðiːz θɪŋs ɪntu θriː ɹɛdˀ bægs æn ə wɪl goː mitʰ heɪ wenzdeɪ ædˀ ˌdə tueɪn ˈstɛɪʃən]

図 3　音声学者による IPA 書き起し例

Fig. 3　An example of narrow IPA transcription

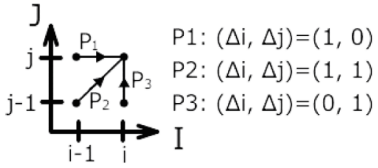

図 4　DTW において選択可能な経路

Fig. 4　Allowable paths of the DTW

慮して設計され，69 単語からなり，CMU 発音辞書 [8] を参照すると，221 米語音素系列に変換できる．

2.2 IPA 書き起しを用いた基準発音距離の定義

IPA 書き起しは，話者の年齢や性別などといった訛り以外の情報とは無関係に作られている．従って 2 話者の IPA 書き起し間の差異を定量的に計測できれば，それは 2 話者の発音の差異そのものといえる．[6] では IPA 単音間距離を局所コストとした Dynamic Time Warping (DTW) を用いて，単音の整合を取りながら求めたパラグラフ全体の累積コストを基準の発音距離としている．図 4 に採択した DTW のパスを示す．

なお，IPA 単音間距離の計算に際し，SAA の全使用話者 369 人の IPA 書き起こしに出現する最頻 95%に相当する 153 種類の単音を抽出し，音声学者にこれらを 20 回ずつ発音してもらった．この収録音声を用いて 3 状態 1 混合の話者依存単音 HMM

を学習し，各単音間の距離は，該当する 2 単音 HMM 間の対応する状態間バタチャリヤ距離 (BD) の平均で定義した．この時，残りの 5%の IPA は音響特性が近いと思われるものに置き換えて計算した．なお，IPA 書き起し間距離を定量的に計測し，これを発音差異とする手法は，上記以外にも幾つか提案されている [9], [10]. [11] では単音 HMM を使う我々の方法が，これらの従来手法よりも，母語話者の主観的発音差異により近い推定値となることを示している．

2.3 サポートベクター回帰を用いた発音距離予測

2.2 の方法で発音距離を求めるためには，「IPA を用いて両話者の発音を書き起す」という専門性を要する手作業行程を経る．しかし本研究で意図するシステムにおいて，入力音声に対して逐一手作業による IPA 書き起しを作るのは現実的ではない．

そこで先行研究では基準距離を予測対象としたサポートベクター回帰 (SVR) によって，書き起しの与えられていない話者間の距離を音声情報のみから予測する実験を行なっている [6]. なお SVR の入力には，2.4 で述べる音声の構造的表象 [12] の考えに基づく各話者の発音構造を使用している．発音構造特徴は，性別や年齢などの話者の声色の違いに対し変動が少なく，発音訛りの違いに対し多様に変化する特徴である [13].

2.4 音声の構造的表象

音声の音響的特徴は，話者の発音の差異の他にも，話者の声道形状や収録に用いる音響機器の差異，さらには収録時の背景雑音の違いによって，様々に変動する．収録環境を改善することで，この変動はある程度抑制することが可能であるが，声道形状や音響機器の伝達特性に由来する歪みは不可避的に混入する．音響分析に広く用いられるケプストラム特徴量は，これらの余計な情報 (非言語的情報) に対して頑健とは言えず，訛りの差異を抽出するのに適していない．そこで [12] では，非言語的特徴に頑健で発音訛りの違いに対しては多様に変化する特徴として音声の構造的表象を提案している．

音声に混入する非言語的特徴は主に，ケプストラムドメインでは，線形変換性歪みで近似されることが多い [14], [15]．マイクロフォンの伝送特性や声道形状の差異は，周波数ドメインでは乗算性歪みであり，ケプストラムドメインでは $\boldsymbol{c} \mapsto \boldsymbol{c} + \boldsymbol{b}$ という変換に相当する．話者間の声道長差異は周波数ドメインでは周波数軸変換となり，ケプストラムドメインでは $\boldsymbol{c} \mapsto \boldsymbol{A}\boldsymbol{c}$ という変換に相当する．以上，音声にはアフィン変換 $\boldsymbol{c} \mapsto \boldsymbol{A}\boldsymbol{c} + \boldsymbol{b}$ で近似される歪みが不可避的に混入すると言える．

音声の構造的表象は，ある話者の音声中に観測される音響事象に対してその絶対的な音響特性ではなく，相対的な配置特性のみでとらえるものである．2 つの空間が連続且つ可逆な空間写像で結びつけられ，それぞれの空間に対応する分布群が存在する場合，分布間の f-divergence(f-div) は空間に依らず不変となる [17]．f-div は 2 つの分布 p_1，p_2 の分布間距離尺度の 1 つで以下の式で表される．

$$f_{div}(p_1, p_2) = \int_{\chi} p_2(\boldsymbol{x}) g\left(\frac{p_1(\boldsymbol{x})}{p_2(\boldsymbol{x})}\right) d\boldsymbol{x} \qquad (1)$$

ただし，$g(x)$ は $x > 0$ で定義される凸関数であり，$g(1) = 0$ を満たす．$g(x)$ の種類によって f-div には，バタチャリヤ距離 (BD)，ヘリンジャー距離，KL ダイバージェンスなどの種類が存在する．

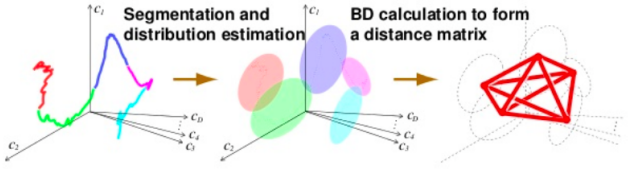

図 5　音声の構造的表象の概念図
Fig. 5　Structural representation of speech

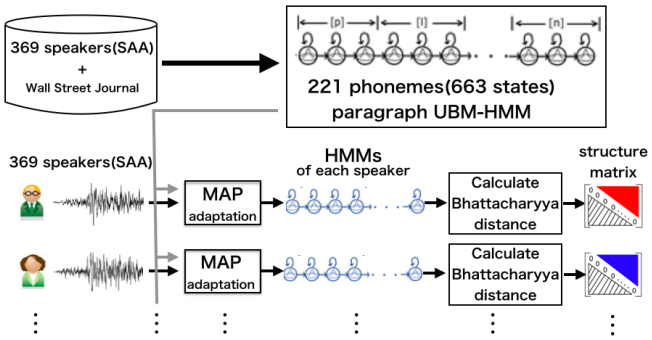

図 6　発音構造算出手順
Fig. 6　Procedure to calculate the pronunciation structure

図 5 に音声の構造的表象の概念図を示す．発声を音素単位などで分布化し，任意の 2 分布間の f-div を計算し，f-div の距離行列を用いて音声を構造的に表象する．

2.5 発音構造

[6] での発音構造算出の概略図を図 6 に示す．

Wall Street Journal (WSJ) [16] の英語音声データセットで学習した 3 状態音素 HMM を初期モデルとし，SAA の全使用話者 369 人で追加学習して，SAA パラグラフを単位とした 221 米語音素系列 Universal Background Model (UBM) を作成する．この UBM と各話者の音声に対する MAP 適応により，各話者の発音を表す話者依存パラグラフ HMM を作成する．同一話者内で音素モデル間の分布間距離（バタチャリヤ距離）を求めることにより各話者の発音を距離行列で表す (音声の構造的表象)．この行列を発音構造行列といい，本実験では発音構造行列の非対角成分を 1 列に並べたベクトルをサポートベクター回帰の説明変数としている．

3. 先行研究との実験条件の違い

3.1 先行研究における 2 つの実験条件

発音距離予測は話者対を対象に行う．筆者らの先行研究 [6] では，話者対 open，話者 open という 2 通りの open 性での実験が行なわれているが，この 2 条件は [3] での可視化において想定される条件とは合致しない．以下，[3] に沿った条件での発音距離予測について検討する．この条件は，話者対 open と話者 open の間に位置づけることができる．

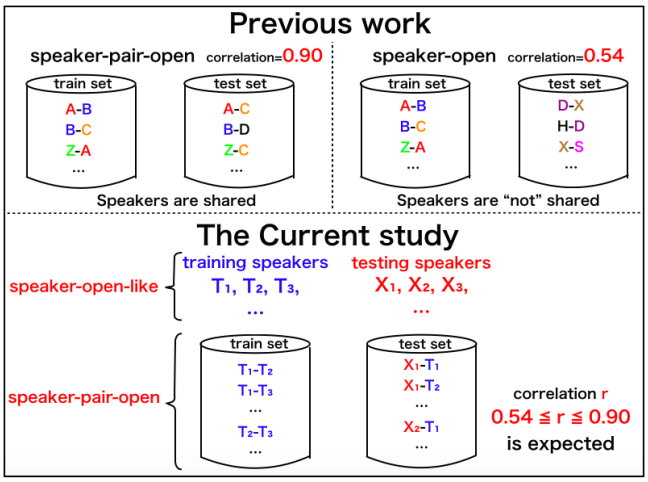

図 7　先行研究と本研究の実験条件

Fig. 7　The comparison of condition between previous work and the current study

3.1.1　話者対 open

話者対 open 条件は，図 7 上段左で示すように，あらかじめ全ての話者間で対を作ったのち，その話者対群を 2 つに分けて片方を SVR の学習セット，もう片方を評価セットとするものである．この条件では評価対 A-B が評価セットにある場合，学習セットには A-$\{x\}(x \neq B)$ または B-$\{y\}(y \neq A)$ が含まれることになる．したがって，学習セットと評価セットに同一話者対は存在しないが，話者を単位としてみると評価セットと学習セットに同一話者が存在しうる．この場合の予測距離と基準距離の相関は最大で 0.90 であった [6]．

3.1.2　話者 open

話者 open 条件は，図 7 上段右に示すように，あらかじめ全ての話者を SVR の学習話者と評価話者に分けたのち，学習話者間で対を作って学習セットとし，評価話者間で対を作り評価セットとする．この条件では評価対 A-B が評価セットにある場合，学習セットには A-$\{x\}$ も B-$\{y\}$ も一切含まれない．

3.1.1 の話者対 open と比較すると，学習セット内の話者と同じ話者が評価セットに含まれる設定上，話者対 open の方が易しい問題設定となっている．実際に [6] では，予測距離と基準距離の相関が話者対 open の場合の 0.90 に対し，話者 open では 0.54 と大きく下回っていることが確認できる．

3.2　本研究の想定に合った実験条件

本研究の想定するシステムに即した実験条件は，図 7 下段の実験条件になる．あらかじめ話者を学習話者と評価話者に分けて学習話者間で話者対を作り学習セットとする点は話者 open と同じだが，評価時には各評価話者と全学習話者との発音距離を予測するという点が話者 open とは異なる．話者を単位としてみると学習セットと評価セットに同一の話者が存在するが，話者対を単位としてみると同一のものが存在しない点は話者対 open となっている．

IPA 書き起しの情報が与えられている話者を既知話者，与えられていない話者を未知話者と呼ぶことにすると本研究の想定は「未知話者と既知話者との距離を予測する」問題に相当する．

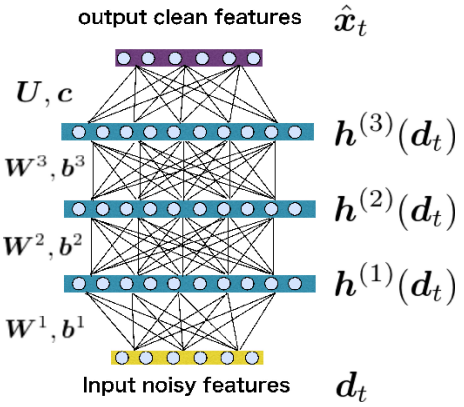

図 8　本稿における DDAE のトポロジー

Fig. 8　DDAE topology

本実験条件を先行研究の 2 つの実験条件と比較すると，3.1.1 の話者対 open は話者に関して closed になっているので本実験条件に対して限定されすぎている．一方，3.1.2 の話者 open は未知話者と未知話者の距離を予測する問題になっているが，本研究の未知話者と既知話者との距離予測に比較して難しすぎる設定になっている．以上のことから本実験で求められる予測距離と基準距離の相関はおよそ 0.54 と 0.90 の間の値を取ることが期待される．

4.　DDAE を用いた耐雑音性の向上

全世界の英語話者が利用できるシステムの構築を想定した場合，入力される音声は任意の録音環境で収録されたものとならざるを得ず，雑音の混入は容易に起こり得る．2.3，2.5 で述べたように，本実験では音声の構造的表象を用いているが，構造的表象が加算性の雑音に対して頑健であるかについては十分な検討がされていない．そこで，本実験ではより実用に即した想定の下，音声認識における雑音抑制技術の一つである Deep Denoising Auto-Encoder (DDAE) [5] を適応して，雑音下での発音距離予測精度の向上を検討する．

4.1　Deep Denoising Auto-Encoder

DDAE は Deep Neural Network (DNN) によって，観測ノイジー特徴量からクリーン音声特徴量を非線形かつ直接的に推定する手法である．図 8 に，使用した DDAE の形状を示す．

隠れ層は 3 層であり，各層のノード数は 1024 である．この DDAE では，クリーン特徴量を次のように推定する．

$$\hat{\boldsymbol{x}}_t = \boldsymbol{U}\boldsymbol{h}^{(3)}(\boldsymbol{d}_t) + \boldsymbol{c}, \tag{2}$$

$$\boldsymbol{h}^{(3)}(\boldsymbol{d}_t) = \boldsymbol{\sigma}(\boldsymbol{W}^{(3)}\boldsymbol{h}^{(2)}(\boldsymbol{d}_t) + \boldsymbol{b}^{(3)}), \tag{3}$$

$$\boldsymbol{h}^{(2)}(\boldsymbol{d}_t) = \boldsymbol{\sigma}(\boldsymbol{W}^{(2)}\boldsymbol{h}^{(1)}(\boldsymbol{d}_t) + \boldsymbol{b}^{(2)}), \tag{4}$$

$$\boldsymbol{h}^{(1)}(\boldsymbol{d}_t) = \boldsymbol{\sigma}(\boldsymbol{W}^{(1)}\boldsymbol{d}_t + \boldsymbol{b}^{(1)}), \tag{5}$$

ここで，$\boldsymbol{U}$ 及び $\boldsymbol{W}^{(n)}$ は重み行列で $\boldsymbol{c}$ 及び $\boldsymbol{b}^{(n)}$ はバイアスベクトルである．$\boldsymbol{d}_t$ は入力音声特徴量であり，MFCC+Δ+ΔΔ(39 次元) の特徴量を 7 フレーム連結した 273 次元の特徴量である．$\hat{\boldsymbol{x}}$ は出力される推定クリーン特徴量であり，39 次元のベクトルである．DDAE のプレトレーニングは Restricted Boltzmann Machine (RBM) として学習したパラメータを各層の初期モデ

表 1　音響分析条件
Table 1　Acoustic analysis conditions

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

表 2　先行研究の条件と本実験の相関の比較
Table 2　Correlation comparison between our previous works and the current study

話者対 open	話者 open	未知話者-既知話者
0.90	0.54	0.77

ルとして用い，最小二乗誤差基準のバックプロパゲーションを適用してファインチューニングを行なった．

5. 実　　験

5.1　システムを想定した条件での発音距離予測実験

本実験では [6] と同じ話者セットの音声データ (clean data) を用いた．実験条件は 3.2 に即しており，369 人の話者を 5 分割し，1 グループを評価用話者，残りを学習用話者として 5-fold の交差検定を行なった．UBM-HMM 学習時の音響分析条件を表 1 に示す．

実験結果を表 2 の未知話者-既知話者欄に示し，比較のために先行研究における 2 条件の相関も合わせて示す．予測距離と基準距離の相関の平均は 0.77 となった．これは 3.2 での「相関はおよそ話者 open の 0.54 と話者対 open の 0.90 の間になる」という予想が正しかったことを実験的に示せたと言える．

また，完全音素認識器を仮定して 2.2 の IPA 書き起こしを，規則によって音素書き起こしに変更したうえで，2.2 と同様の手順で計算した距離 (いわば音素ベース基準距離) と 2.2 で計算される距離 (IPA ベース基準距離) との相関は 0.76 になることが実験的に確かめられている．米語音素数は約 40 であるが，この結果は次のように解釈できる．

音素はしばしば普通の人が判別することができる言語学的に最も小さな単位だと言われており，その定義は聴取者の母国語に依存する．同様に音声記号 (IPA) は，音声学の専門家が聞き分けることのできる最小の単位であり，これは言語に非依存である．音素ベース基準距離は，SAA の IPA 書き起こしを米語音素書き起こしに変換したうえで DTW を用いて計算した．IPA 書き起こしが音声学の専門家への聞こえ方だとすると，米語音素書き起こしは普通のアメリカ人の聞こえ方を表していると言える．本研究の意図する発音距離の自動予測精度が，米語音素書き起こしを用いた場合の精度とほぼ同じであることから，提案手法は普通のアメリカ人が聴取して発音距離を予測した場合に匹敵する性能が得られたと言える．

さらに，SVR に用いる特徴量を改善することで距離予測精度の向上が検討されており [13]，これらの知見を応用することでさらに距離予測精度を向上させることが可能である．

5.2　DDAE を用いた耐雑音性向上実験

本実験では clean data に雑音を重畳してノイジー音声として使用した．また，より実用に即した環境での発音距離予測精度を確かめるためにノイズクローズド環境，ノイズオープン環境の 2 条件で実験を行なった．使用する雑音は電子協の騒音データベース [18] から，機械音ノイズと人混みノイズの 2 種類の雑音を用意した．

本実験ではノイズクローズド環境，ノイズオープン環境共に特徴量強調に使用する DDAE は同一のもの (同じノイズを学習させたもの) を使用しており，評価データに重畳するノイズが DDAE に学習させたものと同一かそうでないかでそれぞれの環境を実現している．

DDAE の学習には clean data と重複がない SAA 話者 1016 人の音声 (dataA)(注1) を使用し，雑音データは機械音ノイズを使用した．dataA をクリーン音声として，これに機械音ノイズを $SNR = 5, 10, 15, 20$[dB] で重畳したものをパラレルデータとして用意し，DDAE は図 8 のトポロジーを使用した．

clean data に対して機械音ノイズ及び人混みノイズをそれぞれ $SNR = 0, 5, 10, 15, 20, \infty$[dB] で重畳することで各 SNR のノイジー音声を得る．ノイズクローズド実験には機械音ノイズを重畳したノイジー音声から抽出した特徴量を使用し，ノイズオープン実験には人混みノイズを重畳したノイジー音声由来の特徴量を使用した．使用する DDAE は共通 (注2) である．距離予測実験では，UBM として 5.1 で用いた clean なものを用い，これをノイジー音声特徴量を使って MAP 適応する，あるいは，DDAE をかけて得られた音声特徴量を使って MAP 適応をすることで，二種類の各話者モデルを作成する．

実験結果を図 9，図 10 に示す．noisy はノイジー音声特徴量を用いた場合，ddae は noisy に DDAE をかけた推定クリーン音声特徴量を用いた場合である．ノイズクローズド，ノイズオープン両方で SNR が低くなるにつれて noisy の相関が著しく低下することが見て取れる．これにより，構造的表象は加算性の雑音に対して頑健でないことが確認できた．

フロントエンド処理として DDAE を適用した場合はノイズクローズド，ノイズオープン双方で距離予測精度は大きく向上していることがわかる．

ノイズクローズド環境では，DDAE をかけた場合 SNR が 0 付近の極めて劣悪なノイズ環境においても 0.7 程度の高い距離予測精度が実現できることがわかり，加えて SNR が 5 以上程度ではほぼクリーン環境と同等の距離予測が行えている．

一方ノイズオープン環境においても，DDAE をかけた場合は $SNR = 10$ 以上の条件であれば，クリーン状態の距離予測精度に匹敵する距離予測精度を実現できる．この実験では評価データに人混み雑音を使用しており，授業中に音声を収録している場合などを想定している．音声認識において比較的難しいとされる人ごみノイズ（人の声がノイズとなっている）を重畳した

(注1)：これらのデータは単語挿入・脱落など単語単位での誤りがあり，距離予測実験では用いていない SAA サンプルである．
(注2)：機械音ノイズを学習させたもの．

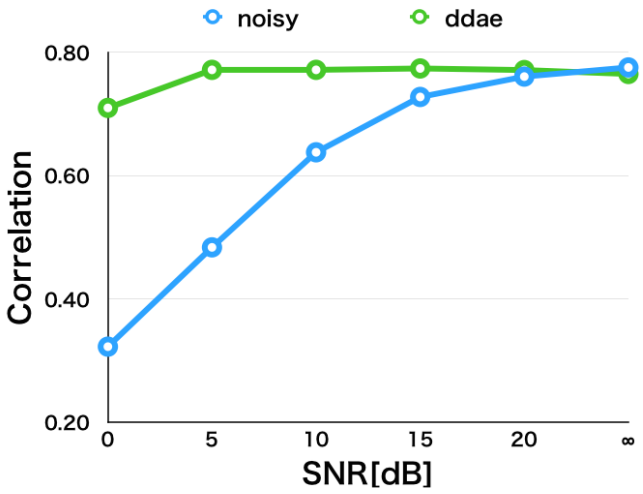

図 9　ノイズクローズド環境での実験結果

Fig. 9　Experimental results under noise-closed conditions

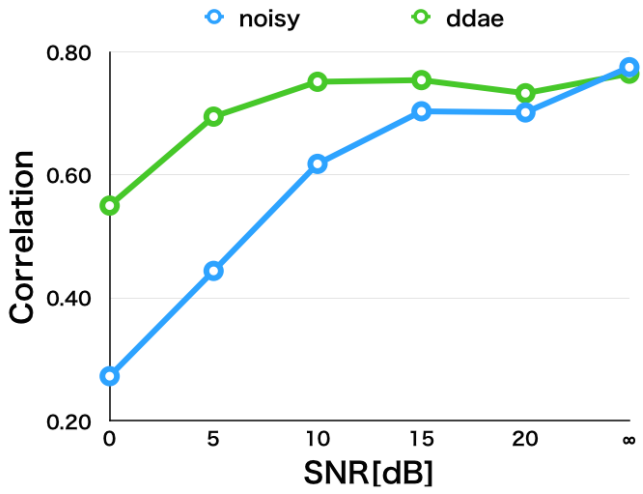

図 10　ノイズオープン環境での実験結果

Fig. 10　Experimental results under noise-open conditions

場合においても比較的高い精度の距離予測が行なえることが確認できたため，今回検討していない他種類のノイズに対しても効果的であることが期待される．

6. おわりに

本稿では世界諸英語の立場に基づき，英語の多様性を俯瞰できるより実用的な技術として「様々な訛りの英語話者群アーカイブに対して自身の英語発音の位置を提示するシステム」を想定した実験を行なった．加えて，意図するシステムは音声収録の行程で雑音が混入する可能性があることを考慮し，Deep Denoising Auto-Encoder を適用してシステムの耐雑音性向上についての検討も行なった．

結果としてシステムを想定した「未知話者-既知話者」の距離予測は 0.77 程度のある程度高い相関で行なえることを確認し，想定システムの実現可能性が高いことを示すことができた．また，耐雑音性に関する検討でも DDAE を用いた特徴量強調によって，ノイズを限定すれば SNR が 0 付近でも極めて高い距離予測が行なえることを確認し，ノイズが未知でも $SNR = 10$ 以上の音声が入力されれば，クリーン状態と同程度の距離予測精度を実現できることが確認できた．

文　献

[1] B. Kachru, *et al.*, *The handbook of World Englishes*, Wiley Blackwell, 2009.

[2] TED, https://www.ted.com/talks

[3] 川瀬他，“訛り・性別・年齢を考慮した自己視点からの世界諸英語発音の可視化，” 電子情報通信学会音声研究会資料，SP2014–12, pp.127–132 2014.

[4] 竹下他，“大学はグローバル人材をどう育てるのか：国際コミュニケーションマネジメント（ICM）のすすめ，” 外国語教育メディア学会全国大会講演集，p.160–161，2014.

[5] P. Vincent., *et al.*, “Extracting and composing robust features with denoising autoencoders,” Proc. Machine learning, pp. 1096–1103, 2008.

[6] 笠原他，“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，121–122，2014.

[7] S. Weinberger., Speech Accent Archive. George Mason University. http://accent.gmu.edu，2014

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

[9] M. Wieling., *et al.*, “Measuring foreign accent strength in english,” Language Dynamics and Change 4(2), 253269, 2014.

[10] M. Wieling., *et al.*, “Inducing a measure of phonetic similarity from pronunciation variation,” Journal of Phonetics 40(2), 307314, 2012.

[11] Tianze Shi., *et al.*, “A measure of phonetic similarity to quantify pronunciation variation by using ASR technology,” Proc. ICPhS, 2015 (to appear).

[12] 峯松他，“音声の構造的表象に基づく学習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，Vol. 52，No. 12, pp. 3671–3681，2011.

[13] N. Minematsu., *et al.*, “Speaker-basis accent clustering using invariant structure analysis and the speech accent archive,” Proc. Odyssay，pp. 158–165，2014.

[14] N. Minematsu., *et al.* “Implementation of robust speech recognition by simulating infants’ speech perception based on the in- variant sound shape embedded in utterances,” Proc. Speech and Computer，pp.35-40，2009.

[15] 峯松信明，“音声の音響的普遍構造の歪みに着眼した外国語発音の自動評定，” 電子情 報通信学会音声研究会，SP2003–180, pp.31-36，2004.

[16] HTK Wall Street Journal Training recipe http://www.keithv.com/software/htk/

[17] Y. Qiao., *et al.* “f-divergence is a generalized inbariant measure beween distributions,” In Proc. INTERSPEECH, pp.1349-1352，2008

[18] 電子協騒音データベース, http://www.sunrisemusic.co.jp/database/fl/noisedata01_fl.html