自己視点からの世界諸英語分類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, 北原俊, 峯松信明, 笠原駿, 齋藤大輔, 広瀬啓吉 (東大)

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，上記のように多様な発音が存在する現代において国際的な現場での英会話相手は 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する必要がある。最近では TED [2] のように，様々な訛りの英語音声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，これらの英語スピーチ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。実際に先行研究 [3] では，ユーザー自身を中心とし，周りにアーカイブ内の話者を図 1 のように配置・地図化している。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。

想定するシステムは各自の環境での英語読み上げ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] の可視化を用い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 SPLICE [8] を導入することの効果を検討した。

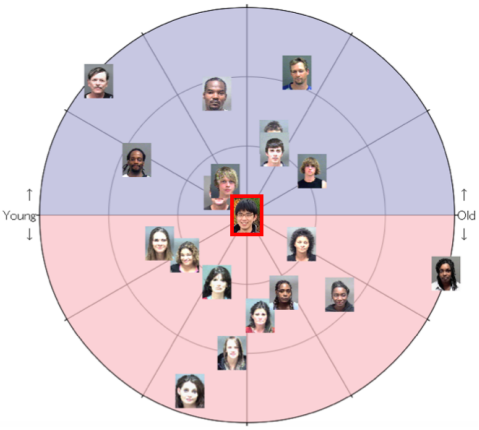

Fig. 1 ユーザーを中心にした世界諸英語話者の地図化; 自身が原点に配置され，自身と他者との距離が発音距離に相当する。上半円が同性，下半円が異性である。x 軸負方向から x 軸正方向への角度が年齢を示している。

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.

Fig. 2 SAA 読み上げ文章

[pliːz kəl ˈstelːʌ as hɐr tu brɪŋ diz θɪŋs wɪθ hɐr frʌm ðə stɑɪ sɪks spuːnz ʌɣ fɹɛʃ ˈsno piːz faɪɣ θɪk ˈslɛbz ʌv bluː ʧiːz æn meɪbiː eɪ snækˀ foɹ hɛɹ bɹʌðɜ bɑbˀ wĭ ɑlso nidˀ eɪ smɑlˠ plæstɪkˀ ˈsŋeɪk æn eɪ bɪg tʰɔɪ fɹɔgˀ fɔɹ ðə kɪdz ʃi kɐn ˈskuːbˀ ðiːz θɪŋs ɪntu θriː ɹɛdˀ bægs æn ə wɪl goː mitʰ hɛɹ wɛnzdeɪ ædˀ də tɹeɪn ˈstɛɪʃən]

Fig. 3 音声学者による IPA 書き起し例

2 発音距離予測

[5] では，英語パラグラフ読み上げ音声コーパス Speech Accent Archive (SAA) [6] 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。本研究における発音距離予測の枠組みは [5] 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慮して設計され，69 単語からなり，CMU 発音辞書 [7] を参照すると，221 米語音素系列に変換できる。

*Study on prediction of pronunciation distances for self-centered visualization of World Englishes diversity and its noise robustness. by Y. Sato, S. Kitahara, N. Minematsu, S. Kasahara, D. Saito, K. Hirose (The University of Tokyo)

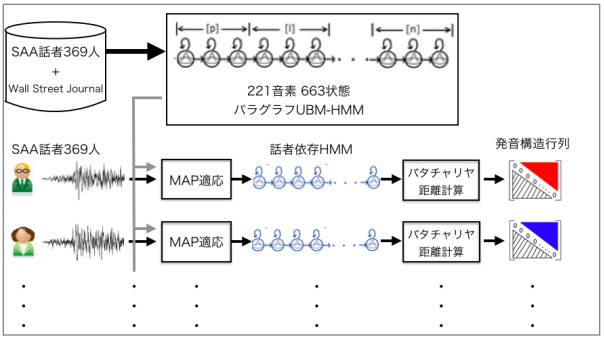

Fig. 4　発音構造算出手順

2.2　IPA 書き起しを用いた基準発音距離

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。[5] 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。

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」という高い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 [5]。なお SVR の入力には，音声の構造的表象 [9] 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 [11]。

2.4　発音構造

[5] での発音構造算出の概略図を図 4 に示す。

Wall Street Journal (WSJ) [10] 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 370 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と，この行列が発音構造に対応する。

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。[5] では話者対 open，話者 open という 2 通りの open 性での実験

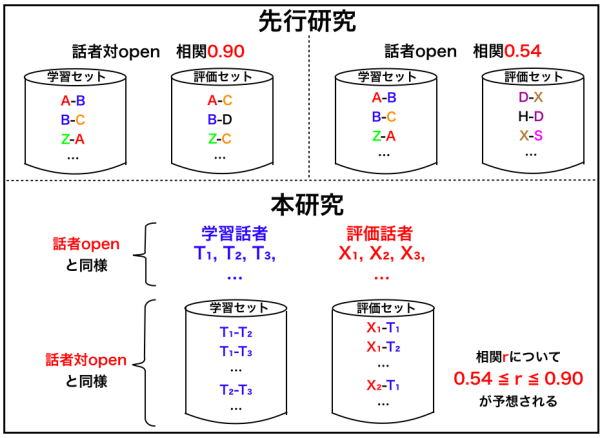

Fig. 5　先行研究と本研究の実験条件

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。

3.1.1　話者対 open

話者対 open 条件は，図 5 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，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 [5]。

3.1.2　話者 open

話者 open 条件は，図 5 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。

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 [5] 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 5 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。

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 SPLICE 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, 2.4 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未検討である。そこで，本実験ではより実用に即した想定のもと特徴量強調手法の一つである SPLICE [8] 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。

SPLICE は雑音重畳音声の特徴量 y からクリーン音声の特徴量の推定値 $\hat{x}$ を求める手法である。雑音による音声の歪みは特徴量ベクトルの非線形な変換として表される。SPLICE では，その非線形な変換が局所領域においては線形変換で近似されるという仮定に基づいて雑音抑制を行う。

4.1 原理

SPLICE は時間的対応のとれているステレオデータを学習データとして利用する。以下ではクリーン音声特徴量を x，雑音重畳音声特徴量を y とし $y' = [1, y^\top]^\top$ とする。

まず特徴量ベクトル y' の統計的性質を GMM を用いてモデル化する。

$$P(y') = \sum_{k=1}^{K} P(k, y') \quad (1)$$

$$= \sum_{k=1}^{K} P(k)P(y'|k) \quad (2)$$

$$= \sum_{k=1}^{K} \pi_k \mathcal{N}(y'; \mu_k, A_k) \quad (3)$$

この時，π_k, μ_k, A_k はそれぞれガウス分布の重み係数，平均，分散である。このモデルを仮定した上で，SPLICE では

$$\hat{x} = \sum_{k=1}^{K} P(k|y')A_k y' \quad (4)$$

のように推定特徴量 $\hat{x}$ を得る。ここで $P(k|y')$ は GMM から次のように求められる。

$$P(k|y') = \frac{P(k, y')}{P(y')} \quad (5)$$

$$= \frac{\pi_k \mathcal{N}(y'; \mu_k, A_k)}{\sum_{k=1}^{K} \pi_k \mathcal{N}(y'; \mu_k, A_k)} \quad (6)$$

A_k は GMM によって区分された領域 k に存在する入力 y' を変換するための行列であり，学習データから以下のように求められる。

まず，誤差の評価関数に以下の式を用いる。なお，t はパラレルデータの時間インデックスである。

$$\sum_{t=1}^{T}\sum_{k=1}^{K} P(k|y'_t)\|x_t - A_k y'_t\|^2 \quad (7)$$

その上で，A_k は次の式を解くことで求めることができる。

$$A_k = \arg\min_{A_k} \sum_{t=1}^{T}\sum_{k=1}^{K} P(k|y'_t)\|x_t - A_k y'_t\|^2 \quad (8)$$

最小値を求めるので，A_k で (8) 式の $\arg\min_{A_k}$ 内を微分することで

$$\sum_{t=1}^{T} P(k|y'_t)(x_t - A_k y'_t){y'_t}^\top = 0 \quad (9)$$

$$A_k = \left(\sum_{t=1}^{T} P(k|y'_t)x_t {y'_t}^\top\right)\left(\sum_{t=1}^{T} P(k|y'_t)y'_t {y'_t}^\top\right)^{-1} \quad (10)$$

と解析的に求められる。

4.2 SPLICE の特徴

SPLICE の利点として，一度パラメータを学習させれば (4) 式の重み付き線形変換の計算のみでクリーン特徴量を推定できるので演算が高速である点が挙げられる。そのため実時間処理に適しており本研究が目標とするシステムへの応用も十分考えられる。

一方，ノイジー特徴量空間で GMM を構成する為に，領域分割がパラレルデータに用いる雑音の種類に大きく依存してしまう欠点がある。学習データの雑音環境に過学習してしまい未知雑音に対して特徴量強調の精度が著しく低下してしまう恐れがある。

しかし本研究の目的とするシステムを運用することを考えた場合に，使用者は自分の英語発音の特性を知る目的でシステムを利用する。あえて多様な雑音が混入する収録条件を選ぶことはないと予想できるので，未知雑音に対する頑健性はそれほど重要ではない。

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 [5] と同じ話者セットの音声データ (clean data) を用いる。実験条件は 3.2 に即しており，369

Table 1　先行研究の条件と本実験の相関の比較

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

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。

実験結果を表 1 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。

5.2　SPLICE 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。SAA 内に散見される雑音レベルの高いサンプルを聴取し，似た雑音として電子協の騒音データベース [12] から，計算機 (中型)，空気室 (大型) の 2 種類の雑音 (noise) を使用した。clean data に対して noise を $SNR = 0, 5, 10, 15, 20$[dB] で重畳することで各 SNR のノイジー音声を得る。ノイジー音声から抽出した特徴量を評価データ noisy とし，noisy に SPLICE を施したものを評価データ splice とする。距離予測実験では，UBM として 5.2 で用いた clean なものを用い，これに noisy, splice それぞれを用いた MAP 適応により各話者モデルを作成する。

また，SPLICE の学習には clean data と重複がない SAA 話者 1016 人の音声 (dataA)[1] を使用し，雑音データは noise を使用した。dataA をクリーン音声として，これに noise 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，GMM の混合数は 512 とした上で SPLICE の学習を行なった。

実験結果を図 6 に示す。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。これは構造的表象が加算性の雑音に対して頑健でないことを表している。一方，splice の場合は極めて SNR が低い条件下でも比較的高い相関を実現できることが確認できた。特に $SNR = 10$ 以上程度の雑音環境下ではほぼ clean data を用いた場合の相関 (0.77) と同程度の距離予測を実現できると言える。

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い，加えてその耐雑音性についての検討を行なった。結果として「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 SPLICE を用いた特徴量強調によって，特に $SNR = 10$ 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。

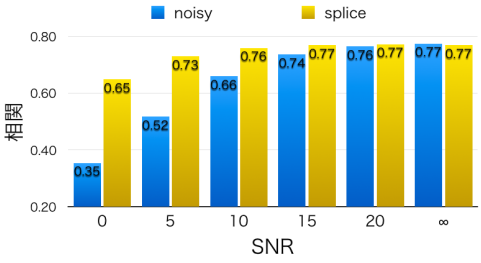

Fig. 6　SPLICE による特徴量強調の効果

今回の耐雑音性に関する実験においては未知雑音への頑健性を重視しなかったが，実際に未知雑音を混入した場合相関にどの程度影響が出るかは調べる必要がある。今後は DNN-SPLICE [13] のように未知雑音に対して強調効果を示す手法と比較を行なう。

[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。

参考文献

[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] 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.

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

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

[8] J. Droppo, *et al.*, “Evaluation of SPLICE on the Aurora 2 and 3 tasks,” in Proc. ICSLP, pp. 29–32, 2002.

[9] 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.

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

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

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

[13] 柏木他, “Deep Learning に基づくクリーン音声状態識別による雑音環境下音声認識,” 日本音響学会春季講演論文集, 9–12, 2013.