Vol.2018-SLP-125 No.15
2018/12/10

スパース性に着眼したPosteriorgramのコンパクト化とDTW発話比較における効果

田中 公啓[1,a)]　椛島 優[1,b)]　齋藤 大輔[1,c)]　峯松 信明[1,d)]

概要：外国語の発音学習において，モデル音声を学習者が真似ることは日常的に行われている．最近では読上げ音声を聴取して再生する（真似る）だけでなく，呈示された音声に対してなるべく遅れずに再生するシャドーイングも広く行われている．学習者音声を呈示されたモデル音声を用いて自動評価する手法として，両音声を，DNN 音響モデルからの出力を正規化して得られる音素事後確率ベクトルの系列 (Posteriorgram) へ変換し，これらを DTW（Dynamic Time Warping）によって比較することが行われている．しかしながら，一般にフレームごとに出力される音素事後確率ベクトルの次元数は数千に及び，データ量が非常に大きくなってしまう．本研究では Posteriorgram の中で，確率の小さいクラスは値を 0 として切り捨てるコンパクト化を検討した．そして，コンパクト化によってシャドーイング音声の評価精度がどのように変化するのかについて実験的に検討した．提案手法の評価は教師による手動スコアとの相関値で行なうが，実験の結果，相関値は 0.79 から 0.88 となり，約 11% 改善した．

キーワード：発話比較，音素事後確率，Posteriorgram，DTW，スパース性，コンパクト化，正規化

1. はじめに

近年のグローバル化の進行，特にインターネットの普及に伴う文化の交流の活発化に伴い，外国語学習の需要はますます高まってきている．外国語学習における有用な手段の一つにシャドーイングがある．これは手本となるモデル音声を聞き，極力間をおかずに学習者がモデル音声を真似て同一の文を再生するというものである．シャドーイングは音声の聴取，理解，発声をほぼ同時に行なう認知負荷量の高いタスクであるが，これらの処理はほぼ自動的に，かつ並列に行えるようにならなければ，外国語の円滑な音声コミュニケーションは難しい．そのような意図の下，比較的負荷の高いシャドーイングが広く導入されている [1]．

シャドーイングは学習現場に広く導入されてはいるが，自分のシャドーイング音声がどれほどモデル音声，すなわち母語話者の発音に近いかについては現状では第三者に判定してもらうほかなく，完全な独学が難しいという問題がある．しかしながら，シャドーイング音声の自動評価に関して技術的検討が進められており，例えばテキスト（意図された音素列）と学習者音声との比較によって音声をスコア化する Goodness of Pronunciation (GOP) がある．一般に，音声特徴 o_t が与えられたときに，それが音素 x である確率 $P(x|o_t)$ を音素事後確率と呼ぶ．GOP は，意図された音素列を用いて音素事後確率を計算し，これにより学習者音声をスコア化する手法であり，隠れマルコフモデル (hidden Markov model; HMM)，またディープニューラルネットワーク (deep neural network; DNN) を用いた算出手法が検討されてきた [2][3]．

また，テキストと音声との比較ではなく，モデル音声と学習者音声とを特徴ベクトル系列として表し，これらを動的時間伸縮 (dynamic time warping; DTW) を通して比較することで学習者音声をスコア化する手法についても検討されてきており [3][4]，特に DNN 音響モデルの出力を正規化して得られる音素事後確率ベクトルを時系列として並べた Posteriorgram を対象として DTW を行う DNN-DTW は，発話評価に有効な手段である [3][5][6]．

DNN-DTW が音声の自動評価に有用な手法である一方，一般にフレーム毎の音素事後確率ベクトルは数千次元に及び，総フレーム分の値となるとデータ量は膨大になってしまう．本研究においては，この問題に対するアプローチとして，確率の小さい音素クラスは値を 0 として切り捨てるコンパクト化を実施し，値を残すクラス数を変えること

[1] 東京大学
The University of Tokyo
[a)] tanaka@gavo.t..u-tokyo.ac.jp
[b)] kabashima@gavo.t.u-tokyo.ac.jp
[c)] dsk_saito@gavo.t.u-tokyo.ac.jp
[d)] mine@gavo.t.u-tokyo.ac.jp

で，手動評価スコアとの相関 (評価精度) がどのように変化するのかを検討した．最終的な結果として，自動スコアと手動スコアの相関は，コンパクト化をかけない場合に比べ 11% の改善（0.79→0.88）を達成した．

2. 先行研究

2.1 DTW による発話評価

DTW は 2 つのフレーム系列が与えられたときに，時間軸を非線形に伸縮させて両系列を対応づけることで，累積距離を最小化する手法である．主に発話間の時間軸合わせ（時間アライメント）において利用され [8]，二系列がモデル音声，学習者音声であれば，最小化累積距離は学習者音声に対するスコアとなる．従来，スペクトル系列やケプストラム系列に対して DTW は適用されてきたが，この場合，話者や年齢の違いが特徴ベクトルを変化させるため，所望の比較結果が得られないことがある．DNN に基づく音声認識は，多数話者の音声データを用い，入力音声フレームに対して音素事後確率を出力するように DNN を学習する．音素事後確率は話者や年齢に非依存であるため，音声特徴を話者・年齢などの非言語情報を抑制した特徴に変換する変換器として DNN を考えることができる．したがって Posteriorgram を対象とした DTW では，スペクトル，ケプストラム系列に比べ話者性の違いに頑健な対応付けが可能となり [9]，検索語検出 (Spoken Term Detection; STD) においても利用されている [10]．また，Posteriorgram を用いた DTW では言語非依存性についても報告されている [4]．

2.2 DNN-DTW におけるデータ量の問題とアプローチ

前述の通り，フレーム毎の DNN 音響モデルの出力である事後確率ベクトルは数千次元に及び，これをフレーム数分保存するとデータ量が非常に大きくなるが，この問題に対する 2 つのアプローチが検討されてきた．

まず 1 つ目のアプローチとして決定木を用いたトライフォンの状態共有による音素クラス数削減がある．これに関しては，椛島らが音素事後確率のクラス数とシャドーイング音声評価精度との関係について検討しており，クラス数の大幅な削減を実現しているものの，依然として安定した評価精度を保つにはクラス数が 2500 以上必要であることが示されている [11]．この時，決定木に基づくクラス数削減では発話比較精度が向上する結果は得られていない．

2 つ目のアプローチが，Posteriorgram のコンパクト化である．DNN の学習では，正解ラベルとして 1 音素クラスのみが 1，残りの全クラスが 0 として与えられる．故に，音響特徴量の入力に対する出力を正規化し得られる事後確率ベクトルにおいては，大きな値を持つ値は限られており，多くのクラスにおいては値は非常に小さいものと推測され，これらのクラスの値を 0 として切り捨ててしまっても，DTW の際の影響は小さいと考えられる (コンパクト化)．このアプローチに関連する研究として，岩崎らが STD において事後確率ベクトル中の最大クラスのみ抜き出し最尤系列として発話比較を行っている [7]．しかしながら本研究で扱う音声は，1) 認知タスクの高いシャドーイング時の音声であること，2) 非母語話者による発声であることにより，しばしば調音的に不正確，曖昧な発声となる．その結果，最尤系列のみでの比較では不十分であると考えられるが，複数クラスを抜き出した学習者音声の発話比較に関する研究は未だになされていない．

3. 実験

3.1 実験環境

本研究の実験においては，以下すべて [11] の実験環境を踏襲する．モデル音声コーパスとして米語母語話者による 10 文の英文読み上げ音声を，またシャドーイング音声コーパスとして 124 名の日本人学生による延べ 1206 文のシャドーイング音声を使用する．全てのシャドーイング音声には，米語母語話者による 15 点満点での得点 (手動スコア) 付けがなされている．また，DNN 音響モデルの学習には WSJ の音声コーパスを用いており，フレーム毎の入力に対して出力音素クラスは 3500 クラスとした．

DTW 距離の計算にはバタチャリヤ距離を用いている．なお，参考としてクラス数 C の二つの事後確率ベクトル $\boldsymbol{p} = (p_1, \ldots, p_C)^\top \boldsymbol{q} = (q_1, \ldots, q_C)^\top$ のバタチャリヤ距離 $\mathrm{BD}(\boldsymbol{p}, \boldsymbol{q})$ の定義を式 (1) に示す．

$$\mathrm{BD}(\boldsymbol{p}, \boldsymbol{q}) = -\log\left(\sum_{i=1}^{C}\sqrt{p_i q_i}\right) \qquad (1)$$

3.2 予備実験

まず，予備実験として，モデル音声とシャドーイング音声の Posteriorgram がどれほど偏っているかを調べるため，モデル音声，シャドーイング音声それぞれについて，各フレーム毎の音素事後確率を降順に並べ替えて累積確率分布化し，その全フレームの平均をとった．**図 1** に結果を示す．なお，横軸は対数軸を取っている．

図 1 の累積分布によると，シャドーイング音声に比べ，モデル音声の方が音素事後確率の偏りが大きい，つまりスパース性が大きいことがわかる．これはモデル音声 (母語話者の発声) は曖昧さが少ないことが要因と考えられる．また，モデル音声，シャドーイング音声の両音声において，上位 10 位までの確率値の累積和が 0.9 以上に達しており，これ以降の確率値を 0 にしてしまっても十分 Posteriorgram 全体の分布について説明できると考えられる．

3.3 Posteriorgram のコンパクト化

ここでは，フレーム毎に音素事後確率値の上位 N_{top} クラス以外のクラスの確率を 0 とするコンパクト化を行い，

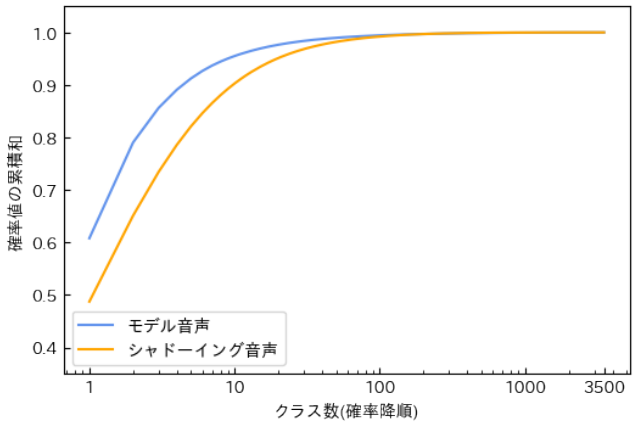

図 1 両音声における音素事後確率の累積分布.

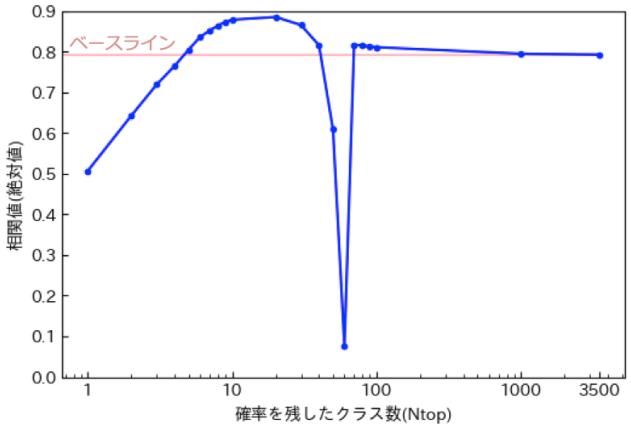

図 2 コンパクト化した Posteriorgram を対象にした DTW 距離と手動スコアとの相関.

Posteriorgram を得た (Posteriorgram のコンパクト化) . その後，モデル音声とシャドーイング音声との Posteriorgram の DTW 距離を算出し，評価精度，つまり手動スコアとの相関を計算した.

以下，コンパクト化の手順について説明する．DNN 最終層では活性化関数として softmax 関数を用いた事後確率化が行われる。以下，音素クラス x について，softmax 関数への入力を $A(x)$ とし，得られる事後確率を $P(x)$ とする。先行研究 [11] では，$P(x)$ は具体的には以下のように算出される。

$$e(x) = \exp\left\{A(x) - \max_x A(x)\right\}$$
$$P(x) = \frac{e(x)}{\sum_x e(x)} \quad (2)$$

このとき，$P(x)$ は全音素クラス x に対し正の値をとる．しかし，本研究においては，確率値の大きい上位 N_{top} クラス以外は確率を 0 にしたいため，式 (3) の活性化関数を使用することでコンパクト化を実現した.

$$e(x) = \begin{cases} \exp\{A(x) - \max_x A(x)\} & (\text{上位 } N_{top} \text{ のクラス}) \\ 0 & (\text{それ以外}) \end{cases}$$
$$P(x) = \frac{e(x)}{\sum_x e(x)} \quad (3)$$

この活性化関数を使用した場合，$P(x)$ が 0 となりえる．その結果，コンパクト化によって式 (1) の $\Sigma_i \sqrt{p_i q_i}$（バタチャリヤ係数）も 0 となりえるため，特に N_{top} が小さい場合にバタチャリヤ距離が無限大となってしまう．本実験では，$\Sigma_i \sqrt{p_i q_i}$ が 0 となった場合，十分小さな正定数 ϵ に置換してバタチャリヤ距離を計算した.

今回は，モデル，学習者の両音声に対してコンパクト化を適用し実験を行った.

3.3.1 コンパクト化：N_{top} が大きい場合

まず，予備実験の結果に基づき，N_{top} が大きい場合，つまり確率値の小さな音素クラスのみ切り捨てる場合についてコンパクト化を実施し，手動スコアとの相関を算出した．ここで，$N_{top} = 10, 100, 1000, 3500$ とし，また 10 以上 100 以下について，10 ずつ増分した場合も検討した．$N_{top} = 3500$ の場合は，コンパクト化をかけていない従来手法であるため，これをベースラインとする．結果は第 3.3.2 節の結果と併せ，**図 2** に示す．ただし，シャドーイング音声が母語話者の発声に近い，つまりモデル音声との類似度が高いほど DTW 距離は小さくなり，一方で手動スコアは高くなるため，相関値は負になる．故に，ここでは相関値の絶対値を評価尺度として扱う.

図 2 に見られるように，N_{top} が 70 以上の場合については，手動スコアとの相関がベースラインを上回った．これは，確率値を 0 にしたクラスが意図された発話のクラスとは無関係であり，ベースラインにおいてはかえってノイズとして作用し，精度が悪くなっていたと考えられる．換言すれば，コンパクト化によって学習者音声が，発音評価により適した形で表現されるようになったと解釈できる．また，$N_{top} = 60$ で極端な相関の落ち込みが見られる．N_{top} が小さくなると DTW の最適パスの中に，局所距離として ϵ を用いた距離計算が行なわれることが起こる．これは，モデル音声と学習者シャドーイング音声との距離を極めて大きく扱うことに相当し，この操作が観測されるのは N_{top} が 60 以下になった場合であった．$N_{top} = 60$ 付近の局所的な精度劣化は，この操作の副作用であると考えている．N_{top} が 60 を下回る場合，この操作が生起する頻度は上昇するが，N_{top} の数値によっては，不適切な学習者発声に対して，モデル音声との距離をより強調する効果を有しており，最終的には相関値の上昇に寄与することとなった．この相関値上昇は，ϵ の値にも依存しているが，例えば，零化しない場合の $\Sigma_i \sqrt{p_i q_i}$ の最小値を ϵ として設定すれば，このような副作用を防ぐことができるが，この場合でも相関値の上昇に寄与するのかどうか，実験的な検証が必要である.

表 1 $N_{top} = 10, 20, 30, 3500$ における DTW 距離と手動スコアとの相関.

N_{top}	10	20	30	3500
SN2	**0.88**	**0.88**	0.87	0.79 (B/L.)

3.3.2 コンパクト化：N_{top} が小さい場合

次に，N_{top} が小さい場合，つまり比較的確率値の大きなクラスにもコンパクト化を実施した．ここで，N_{top} を 1 以上 10 以下の全ての場合について実験を行った．前述の通り，結果は第 3.3.1 節の結果と併せ図 2 に示す．なお，N_{top} を 1 にすることは，フレーム毎に音素識別することに相当し（最尤クラスのみを採択），先行研究 [7] に凡そ相当する．

今回は，N_{top} を小さくする毎に，相関は低下した．これはさきに述べたとおり，シャドーイング音声はモデル音声に比べもともとのスパース性がモデル音声に比べ小さいため，極端なコンパクト化は，学習者音声を不適切に表現することになることが原因であろう．

3.3.3 コンパクト化実験のまとめ

表 1 には，特に相関が良かった $N_{top} = 10, 20, 30$，及びベースラインである $N_{top} = 3500$ について DTW 距離と手動スコアとの相関を示した．$N_{top} = 10, 20$ の場合，相関値は 0.88 に達しており，ベースラインスコア (表中に (B/L.) で示した) である 0.79 に比べ 11% もの改善を達成した．[6] では，ベースラインの DNN-DTW に対して，GOP スコアや単語認識率など幾つかの説明変数を用意し，回帰モデルを通して手動スコアの予測を行なっている．最高精度を示した回帰モデルの性能は相関値 0.90 であった．今回の検討は，今後の実験的検証が必要ではあるが，回帰モデルなどの学習や最適化を導入せずにほぼ同等の精度向上を実現している．

4. まとめ

本研究では先行研究を踏まえ，コンパクト化した Posteriorgram の発話比較，発話評価への応用を検討し，結果として手動スコアとの相関においては従来手法に比べ 11%（0.79→0.88）もの改善を示した．今後は，DTW を適用する対象を Posteriorgram ではなく，音声認識における有効な特徴量であることがわかっているネットワーク中のボトルネック特徴量 [12][13] にした場合に評価精度がどのように変化するのか検討していく予定である．

参考文献

[1] 玉井 健, “リスニング力向上におけるシャドーイングの効果について,” 日本通訳学会第 3 回年次大会 講演，2002.
[2] Dean Luo, Naoya Shimomura, Nobuaki Minematsu, Yutaka Yamauchi and Keikichi Hirose, “Automatic Pronunciation Evaluation of Language Learners’ Utterances Generated Through Shadowing,” in *Proc. INTERSPEECH* 2807–2810, 2008.
[3] Junwei Yue, Fumiya Shiozawa, Shohei Toyama, Yutaka Yamauchi, Kayoko Ito, Daisuke Saito and Nobuaki Minematsu, “Automatic Scoring of Shadowing Speech based on DNN Posteriors and their DTW,” in *Proc. INTERSPEECH* 1422–1426, 2017.
[4] Ann Lee and James Glass, “Pronunciation Assessment via a Comparison-based System,” in *Proc. SLaTE* 122–126, 2013.
[5] Yusuke Inoue, Suguru Kabashima, Daisuke Saito, Nobuaki Minematsu, Kumi Kanamura, Yutaka, Yamauchi, “A Study of Objective Measurement of Comprehensibility through Native Speakers’ Shadowing of Learners’ Utterances,” in *Proc. INTERSPEECH* 1651–1655, 2018.
[6] Suguru Kabashima, Yuusuke Inoue, Daisuke Saito, Nobuaki Minematsu, “DNN-based scoring of language learners’ proficiency using learners’ shadowings and native listeners’ responsive shadowings,” in *Proc. Spoken Language Technology*, 2018 (to appear) .
[7] 岩崎瑛太郎, 小原真人, 小嶋和徳, 李時旭, 伊藤慶明, “音声中の検索語検出における深層学習の事後確率を用いたクエリの最尤系列化方式,” 日本音響学会 秋季研究発表会, 2018.
[8] Müller and Meinard, “Dynamic time warping,” in *Information retrieval for music and motion* 69–84, 2007.
[9] Takuya Ozuru, Nobuaki Minematsu, Daisuke Saito, ”Prosodic Comparison of Utterances without Extracting Fundamental Frequencies based on Vocalized Subharmonic Summation,” in *Proc. Speech Prosody* 172–176, 2018.
[10] Timothy J. Hazen, Wade Shen and Christopher White, “Query-By-Example Spoken Term Detection Using Phonetic Posteriorgram Templates,” in *IEEE Workshop on Automatic Speech Recognition & Understanding*, 2009.
[11] 椛島優, 塩澤文野, 齋藤大輔, 峯松信明, 山内豊, 伊藤佳世子, “DNN-GOP と DNN-DTW に基づくシャドーイング音声自動評価の高精度化,” 日本音響学会 春季研究発表会, 2018.
[12] 向原康平, サクリアニ サクティ, グラム ニュービック, 戸田 智基, 中村 哲, “ボトルネック特徴量を用いた感情音声の認識,” 情報処理学会研究報告, 2015.
[13] Gautam Mantena and Kishore Prahallad, “IIIT-H SWS 2013: Gaussian Posteriorgrams of Bottle-Neck Features for Query-by-Example Spoken Term Detection,” in *MediaEval*, 2013.