Deep Neural Network に基づく音素事後確率を用いた発音評価*

☆杉田祐樹, 柏木陽佑, 齋藤大輔, 峯松信明, 広瀬啓吉 (東大)

1 はじめに

コンピュータを用いた語学学習支援 (Computer-Aided Language Learning, CALL) システムの構築において、学習者の発音評価をいかに精度良く行うかということは重要な問題である。CALL における発音評価指標の一つである Goodness of Pronunciation (GOP) は、発話者が意図した音素の事後確率を用いて、他の音素との弁別性によって発音を評価する指標である [1]。従来、音素事後確率は隠れマルコフモデル (Hidden Markov Model, HMM) を用いて計算されていたが、HMM の更なる改良には限界があった。一方、Hinton らによって Deep Neural Network (DNN) が考案され音声認識において著しい精度向上が実現すると [2][3]、DNN を GOP 計算に組み込むアプローチも試みられた。Hu らは DNN の出力として得られる各音素状態の時間フレーム毎の事後確率を用いて、HMM による強制アライメントで得られる音素状態の時間情報をもとに GOP を計算する手法 (以下 DNN-framewise) を提案した。この手法は HMM による従来手法 (以下 HMM-baseline) よりも高い評価精度を実現した [4]。しかしながら、HMM-baseline のように、音素状態間の遷移確率を考慮して状態の最尤遷移系列でもって当該音素区間の全フレームを単位とした事後確率を計算する手法は、音響的特徴としての時系列性をより明示的にモデル化していると考えられる。本研究では、TANDEM connectionist system[5] と呼ばれる、HMM による特徴量時系列構造のモデル化を DNN に組み込んだシステムによる GOP の計算手法を提案し、日本人英語の自動発音評価において HMM-baseline 及び DNN-framewise との比較検討を行った。

2 GOP

GOP は発話者が意図した音素 p の他の音素との弁別性に着目し、観測特徴量 $O^{(p)}$ の事後確率 $P(p|O^{(p)})$ によって発音を評価する手法である。音素 p の GOP は継続長 D_p を用いて次のように表される。

$$GOP_p = \frac{1}{D_p} \log P(p|O^{(p)}) \tag{1}$$

HMM-baseline では、事後確率を尤度を用いて近似的に変形した式 (2) によって計算する。分母・分子ともに HMM を用いて計算し、音素 p に該当するフレーム区間全体を単位とした事後確率を得る。

$$P(p|O^{(p)}) \approx \frac{P(O^{(p)}|p)}{\max_{q \in Q} P(O^{(p)}|q)} \tag{2}$$

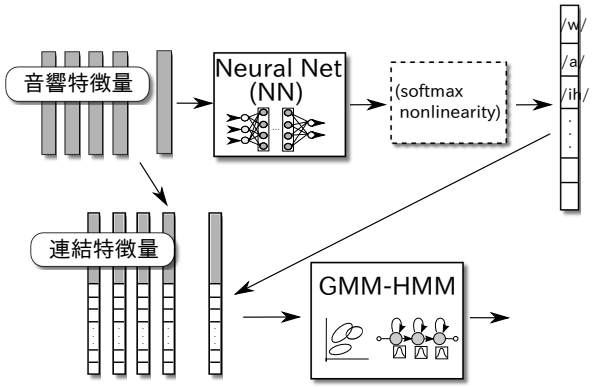

Fig. 1 TANDEM connectionist system

一方 DNN-framewise では、DNN の出力として得られる各音素状態の時間フレーム毎の事後確率ベクトルから、時間情報をもとに音素 p に該当するフレーム区間の p の事後確率を平均化する処理によって音素 p に該当する区間全体の事後確率を得る。

3 TANDEM connectionist system

TANDEM connectionist system (以下 TANDEM) は、Neural Network (NN) の出力を音響特徴量に連結した連結特徴量を HMM の入力とする手法である (Fig. 1)。この手法の利点は NN の音素識別機能と HMM による特徴量時系列構造のモデル化機能を融合できるところにある。また TANDEM では NN の出力をガウス分布に近似させることを目的に、NN の出力層の値の対数化、または出力層の softmax 非線形変換の除去が行われる。本研究では、NN として DNN を用いた TANDEM を GOP 計算に適用する手法を提案する。

3.1 TANDEM 発音評価器の作成

まず学習用音声から音響特徴量を抽出し、HMM と DNN を学習する。HMM は発話者の違いに適応させるため出力分布を混合ガウスモデル (Gaussian Mixture Model, GMM) とし、最尤 (ML) 基準によって学習する。DNN は Deep Belief Network (DBN) を用いて教師なし事前学習による初期化を行い、その後本学習する。本学習に使用するラベルは HMM による学習音声の強制アライメントによって取得する。次に学習用音響特徴量を DNN に入力し、得られた出力をもとの音響特徴量に連結して連結特徴量を作成する。識別性の向上を目的に、DNN の出力に対しては主成分分析 (PCA) 、連結特徴量に対しては音素クラスを

* Automatic pronunciation evaluation using phoneme posterior probabilities based on Deep Neural Network. presented by Y. Sugita, Y. Kashiwagi, D. Saito, N. Minematsu and K. Hirose, (The Univ. of Tokyo)

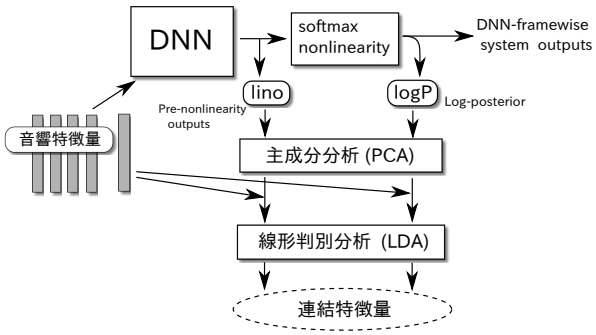

Fig. 2　連結特徴量作成の流れ

Table 1　実験条件

GMM-HMM	16-Gaussian 3 状態 41 音素
DNN ノード数	入力 273 中間 3 層各 1024 出力 64
音響特徴量	MFCC (0-12) + Δ + ΔΔ (計 39 次元)

ラベルとする線形判別分析 (LDA) による次元削減も試みた (Fig. 2)。最終的な発音評価器は作成した連結特徴量を用いて GMM-HMM の再学習を行うことで完成する。

4　実験

TANDEM を用いて GOP による発音評価器を作成する。作成した評価器をもとに日本人が読み上げた英語音声の発音評価を行い、従来手法としての HMM-baseline、DNN-framewise との比較検討を行った。

4.1　評価方法

GOP を説明変数として手動評価スコアとの相関を求める。手動評価スコアは各日本人発話音声に与えられた、英語教師 5 名による 5 段階の評価スコアの平均値を用いる。説明変数としては、加藤らの先行研究 [6] を参考に、各発話音声中の全音素の GOP を平均化したもの (GOP_{all}) と、母音音素および子音音素に分けてそれぞれ平均化したもの (GOP_v、GOP_c) を用意し、GOP_{all} による単回帰と GOP_v・GOP_c による重回帰を行った。

4.2　実験条件

評価用の日本人発話音声としては、日本人による英語読み上げ音声データベース (ERJ) [7] から音素バランスを考慮した文セット (460 文) の読み上げ音声 950 発声、学習用音声としては同じ文セットのアメリカ人による読み上げ音声 5054 発声を用いた。その他の実験条件については Table 1 に示す。GMM-HMM の学習には HTK[1], 1 音素につき 1 状態を出力とする DNN の学習には KALDI[2]を用いた。

4.3　実験結果と考察

実験結果を Table 2 に示す。提案手法では、すべての場合において HMM-baseline を上回る相関を得ることができた。また TANDEM lino+PCA の場合、単回帰では DNN-framewise を上回ることができたが重回帰ではほぼ差異がなく、提案手法の優位性を示すことができなかった。また LDA を用いても相関の向上は見られなかった。提案手法が重回帰において DNN-framewise に対して優位性を示せなかった原因としては、GOP_v による単相関を比較すると TANDEM が 0.561 に対して DNN-framewise が 0.548 と提案手法が上回っている一方で、GOP_c での単相関では両手法ともに 0.457 と差異がないことから、TANDEM の重回帰における説明変数同士の相関が高くなっていることが考えられる。

Table 2　評価実験の結果 “lino” は DNN から出力ノードの softmax 非線形化を除去したときの DNN の線形出力、“logP” は DNN 出力の対数を表す

実験手法	相関 GOP_{all}	相関 GOP_v・GOP_c
HMM-baseline	0.582	0.596
DNN-framewise	0.603	0.626
TANDEM lino	0.591	0.606
TANDEM lino+PCA	**0.612**	**0.624**
TANDEM lino+PCA+LDA	0.606	0.619
TANDEM logP	0.589	0.607
TANDEM logP+PCA	0.593	0.608
TANDEM logP+PCA+LDA	0.596	0.603

5　おわりに

CALL における発音評価指標の一つである GOP、その計算において DNN を活用するアプローチとして、DNN から出力される時間フレーム毎の音素事後確率を直接用いて GOP を計算する手法 (DNN-framewise) があるが、本研究ではさらに HMM を用いて特徴量の時系列性をよりよく表現できると考えられるアプローチ (TANDEM) の適用を提案した。単回帰では DNN-framewise を上回る相関を得ることができたが、重回帰では提案手法の優位性を示すことができなかった。今後の検討課題としては、説明変数間の相関除去による重回帰の改良と、学習データの一部を用いた GOP との相互情報量最大化 (MMI) 基準による HMM の識別学習を考えている。

参考文献

[1] S. Witt, *et al*, Speech Communication, vol. 30, pp. 95–108, 2000.
[2] G. Hinton, *et al*, Neural Computation, vol. 18, no. 7, pp. 1527–1554, 2006.
[3] A. Mohamed, *et al*, Proc. NIPS 22 workshop, 2009.
[4] W. Hu, *et al*, Proc. INTERSPEECH, pp. 1886–1890, 2013.
[5] H. Hermansky, *et al*, Proc. ICASSP, pp. 1635–1638, 2000.
[6] 加藤集平ら, 日本音響学会春季研究発表会講演論文集, pp. 417–420, 2012.
[7] 峯松信明ら, 日本教育工学会論文誌, vol. 27, no. 3, pp. 259–272, 2004.

[1]HTK, http://htk.eng.cam.ac.uk/
[2]KALDI, http://kaldi.sourceforge.net/index.html