Vol.2017-SLP-119 No.11
2017/12/21

英語学習者の発声自動評価を目的とした DNN 音声認識システムの検討

加藤 拓[1]　篠崎 隆宏[1]

概要：日本人による英語音声は，英語母語話者による英語音声と比較すると，発音が不正確になることが多い．そのため英語母語話者の英語音声を用いて構築した音声認識システムによって，日本人の英語音声を認識した場合，認識精度が低くなることが予想される．しかしどの程度の認識率となるかは不明であり，実際に想定されるデータを用いた認識実験が必要である．そこで本研究では，WSJ コーパス及び SWBD コーパスを用いて構築した英語音声認識システムを用いて音声認識性能を評価する．さらに，日本人英語学習者の英語音声を用いて教師ありおよび教師なし適応を行い，その効果について検討を行う．WSJ や SWBD コーパスをそのまま用いた場合，日本人英語音声に対して非常に低い認識精度を示した．日本人英語音声のラベル付きデータを用いて適応することで，認識精度が大きく向上した．次にラベル付きデータ及びラベルなしデータを用いて適応を行った．1 つの出力層を持つ DNN を用いた場合は低い認識精度となったが，2 つの出力層を持つ DNN を用いた場合は認識精度が更に向上することを示した．

キーワード：自動評価，教師あり適応，教師なし適応，DNN，音声認識

1. はじめに

日本人英語学習者に対するスピーキングテストの評価は，人手によって行われることが多いため，多大なコストがかかる．そこで人手による評価の代わりに音声認識システムを用いた自動評価をすることが可能となれば，英語学習者はより手軽にスピーキングテストを受験でき，また採点結果をより素早く知ることができるようになる．

日本語と英語では音素の種類と数が異なるため，日本人の英語音声は，英語母語話者による英語音声と比較すると，発音が不正確になることが多い．そのため日本人の英語音声を，英語母語話者の音声によって構築された英語音声認識システムを用いて認識した場合，認識精度が非常に低くなることが予想される．しかし，実際にどの程度の認識率となるかは不明な点が多い．そこで本論文では，WSJ コーパス及び SWBD コーパスを用いて構築した DNN-HMM に基づく英語音声認識システムを用いて日本人話者の英語音声を認識し，基本的な認識性能を評価する．さらに，教師あり適応及び教師なし適応技術を用いて，非母語話者適応を試みる．

[1] 東京工業大学
Tokyo Institute of Technology, Kanagawa, Japan
www.ts.ip.titech.ac.jp

2. 日本人英語データ

本実験では日本人の英語音声として，国内の 4 校の高等学校において収録された，日本人高校生の英語音声を用いた．日本人英語音声のラベル付き学習データは 4.5 時間（137 人），ラベルなし学習データは 26 時間（379 人），評価セットは 1 時間（25 人）である．評価セットの話者はラベル付き学習データ及びラベルなし学習データには含まれない話者である．

また各話者に対して人手によりスコアが付けられており，発話内容や文法，発音の正確さなどが基準となっている．本実験では，話者毎のスコアと音声認識システムによる認識率との関係についても調査する．

3. 日本人英語音声認識システム

まず英語母語話者のデータを用いて，ベースラインシステムを構築する．次に日本人英語音声のラベル付きデータのみを用いて適応した後に，ラベルなしデータを併用して適応を行う．

3.1 ベースラインシステム

まず英語母語話者のラベル付きデータを用いて，ベースラインシステムの GMM-HMM を構築する．次に GMM-

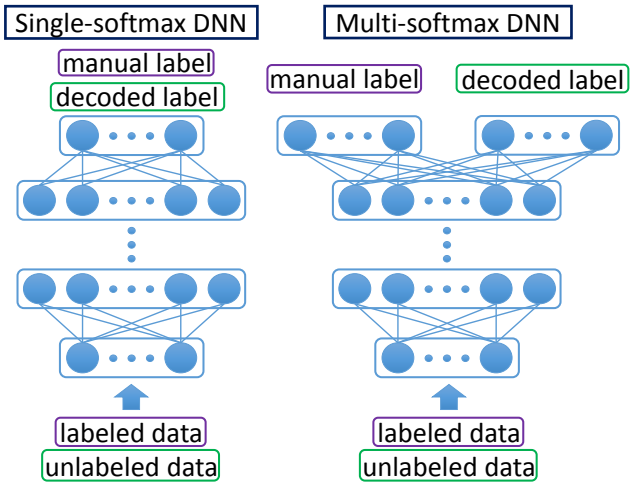

図 1　Single-softmax DNN / Multi-sotmax DNN.

HMM を用いて，学習データの各フレームに対応する HMM 状態（アライメント）を求める．最後に RBM による教師なし学習のプレトレーニングと，アライメントを用いた教師あり学習のファインチューニングにより DNN-HMM を構築する．

3.2　ラベル付きデータを用いた適応

ベースラインシステムの GMM-HMM を用いて，日本人英語音声のラベル付きデータに対するアライメントを求める．ラベル付き音声と得られたアライメントを用いて，ベースラインシステムの DNN-HMM を誤差逆伝播法により更新することで，日本人英語音声への適応を行う．

3.3　ラベル付きデータとラベルなしデータを併用した適応

まずラベル付きデータで適応されたシステムを用いて，ラベルなしデータを認識する．ラベル付きデータに対しては書き起こしテキストを，ラベルなしデータに対しては得られた認識仮説を正解文として用いる．ラベル付きデータとラベルなしデータをシャッフルして学習することで，ラベル付きデータで適応された DNN-HMM を更新する．

またラベルなしデータに対する認識仮説には誤りが含まれるため，この認識仮説を用いて DNN-HMM を更新した場合，DNN-HMM の認識精度が低下することが考えられる．この問題に対して，出力層を 1 つ持つ DNN（single-softmax DNN）の代わりに出力層を 2 つ持つ DNN（multi-softmax DNN）を用いた学習法が提案されている [1]．各 DNN の構造を図 1 に示す．Single-softmax DNN による学習ではラベル付きデータとラベルなしデータを同じ出力層を用いて学習しているのに対し，multi-softmax DNN では各データに対応する出力層を用いて学習する．ラベルなしデータを専用の出力層を用いて学習することで，認識仮説に含まれる誤りの影響を小さくすることが期待できる．

表 1　本実験で用いた各コーパスにおけるデータ量．”Non-native” は日本人による英語音声データを表す．

		Non-native	SWBD	WSJ
Training set	labeled	4.5h	319h	80h
	unlabeled	26h	-	-
Evaluation set		1h	-	-

4.　実験

4.1　実験条件

認識システムの学習および評価には Kaldi ツールキット[*1] を用いた．英語母語話者による音声データとしては Switchboard（SWBD）コーパス [2] 及び Wall Street Journal（WSJ）コーパス [3] を用いた実験を行った．SWBD コーパスのラベル付き学習データは 319 時間，WSJ コーパスのラベル付き学習データは 80 時間である．表 1 に本実験で用いた各コーパスにおけるデータ量を示す．語彙サイズはそれぞれ SWBD では 30k，WSJ では 146k である．

DNN における入力特徴量としては，40 次元の fMLLR（feature-space maximum likelihood linear regression）特徴量 [4], [5] を用いた．fMLLR 法における変換行列は，ラベル付きデータに対しては書き起こしテキストを用いた強制アライメントによって計算され，ラベルなしデータ及び評価セットに対してはデコード時に生成されるラティスから推定される．DNN の入力層の次元は 440 次元（splice : ±5）であり，隠れ層の数は 6 層，隠れ層の次元は 2048 である．出力層の次元は，SWBD では 8819，WSJ では 3382 である．DNN-HMM のファインチューニングにおける初期学習率は 0.008 である．

4.2　実験結果

SWBD コーパス及び WSJ コーパスをベースとしたシステムにおける実験結果を表 2 に示す．ベースラインシステムにおいて日本人の英語音声を認識した結果，SWBD ベースシステムにおいては 95.70%，WSJ においては 83.38%という高い WER を示した．日本人による英語音声の発音が，英語母語話者による発音と大きく異なっており，発話者が本来意図した音素とは異なる音素に識別されてしまったため，高い WER を示したと考えられる．

日本人英語音声のラベル付き学習データを用いてベースラインシステムを適応した結果，どちらのシステムでも WER が大幅に下がり，SWBD では WER が 50.88%，WSJ では 55.46%となった．これはシステムを日本人英語音声に適応することで，各音素における日本人の発音傾向をシステムが学習することができたためだと考えられる．またラベル付きデータで適応した後に，ラベル付きデータとラベルなしデータを同じ出力層を用いて学習した結果，

[*1] http://kaldi.sourceforge.net/index.html

表 2　各学習法による WER.

Corpus	System	WER[%]
SWBD	baseline	95.70
	adapted by labeled data	50.88
	adapted by labeled & unlabeled data (single-softmax DNN)	89.11
	adapted by labeled & unlabeled data (multi-softmax DNN)	**47.58**
WSJ	baseline	83.38
	adapted by labeled data	55.46
	adapted by labeled & unlabeled data (single-softmax DNN)	79.00
	adapted by labeled & unlabeled data (multi-softmax DNN)	**50.86**

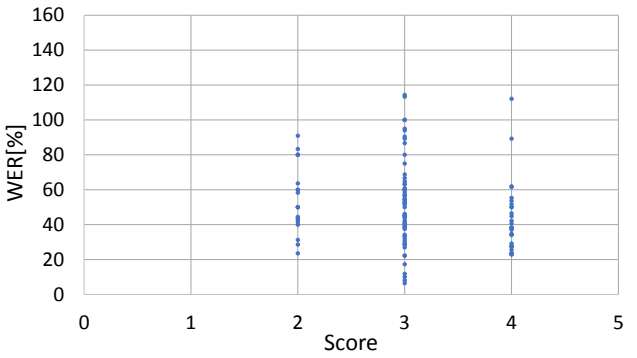

図 2　SWBD システムにおける，人手によるスコアと WER の散布図.

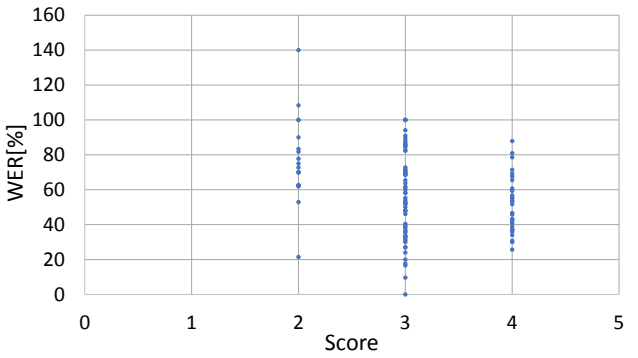

図 3　WSJ システムにおける，人手によるスコアと WER の散布図.

WER が増加した．誤りが多く含まれる認識仮説を，ラベルなしデータに対する正解文として用いて DNN-HMM を更新したため，高い WER を示したと考えられる．一方，異なる出力層を用いてラベル付きデータとラベルなしデータを学習することで，SWBD では WER が 3.3%，WSJ では 4.6%削減された．これは出力層におけるパラメタの更新において，ラベルなしデータの認識仮説に含まれる誤りの影響を小さくすることができたためだと考えられる．SWBD ベースシステムと WSJ ベースシステムを比較すると，適応前のシステムでは WSJ の方が WER が低いのに対し，適応後のシステムでは SWBD の方が低い WER を示している．これはおそらく日本人英語音声の発話内容が，新聞の読み上げ音声である WSJ コーパスよりも，会話音声である SWBD コーパスに近いことが原因だと考えられる．

次に日本人英語音声の評価セットに対する，人手によって付けられた話者のスコアと，multi-softmax DNN を用いてラベル付きデータ及びラベルなしデータで適応したシステムにおける WER の散布図を図 2 と図 3 に示す．図 2 ではベースのシステムとして SWBD を，図 3 では WSJ を用いている．話者のスコアは値が大きいほど優れていることを示す．話者のスコアと WER に対する Pearson の相関係数を求めたところ，SWBD システムでは −0.193，WSJ システムでは −0.319 となり，話者のスコアと WER には弱い相関関係があった．すなわち高いスピーキング能力を持つ話者は，低い WER を示す傾向があると言える．人手によるスコアリングは一般的に，発話内容や文法，発音の正確さなどが基準となる．高いスコアを持つ話者はより正確な発音で発話する傾向があるため，低い WER を示したと考えられる．しかし発話内容が出題に対して誤りの場合や，簡単な単語・文法のみが発話された場合であっても同じ基準で WER が計算されるために，スコアと WER は弱い相関関係に留まったと考えられる．よって英語学習者の発声についてのより正確な自動評価のためには，音声認識システムにおける WER 以外に，発声内容の簡易さや文法事項についても考慮する必要があると考えられる．

5. まとめ

英語母語話者データにより構築されたシステムを日本人英語音声に適応し，英語学習者の発声に対する認識率を用いた自動評価法について検討した．SWBD コーパス及び WSJ コーパスにより構築されたシステムにおいて，日本人英語音声を認識すると高い WER を示したが，ラベル付き日本人英語音声を用いて適応することによって WER が削減した．また 2 つの出力層を持つ DNN において，ラベル付き及びラベルなし日本人英語音声を用いて適応することにより，更に低い WER を示した．次に WER と話者の英語能力の関係について調査した．高い英語能力を持つ話者は，低い単語誤り率を示す傾向があることがわかったが，より正確な自動評価のためには，発話内容の簡易さや文法事項についても考慮する必要があると考えられる．今後の課題としては，言語モデルを用いた発話内容を考慮したス

コアリングの検討や，ラベルなしデータの適応における信頼度の利用などが挙げられる．

謝辞 本研究は JSPS 科研費 JP16H01935，JP26280055 の助成を受けたものです．

参考文献

[1] Su, H. and Xu, H.: Multi-softmax deep neural network for semi-supervised training, *Sixteenth Annual Conference of the International Speech Communication Association* (2015).
[2] Godfrey, J. J., Holliman, E. C. and McDaniel, J.: Switchboard:Telephone speech corpus for research and development, *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, Vol. 1, IEEE, pp. 517–520 (1992).
[3] Paul, D. B. and Baker, J. M.: The design for the Wall Street Journal-based CSR corpus, *Proceedings of the workshop on Speech and Natural Language*, Association for Computational Linguistics, pp. 357–362 (1992).
[4] Gales, M. J. F.: Maximum likelihood linear transformations for HMM-based speech recognition, *Computer Speech and Language*, Vol. 12, pp. 75–98 (1998).
[5] Povey, D. and Saon, G.: Feature and model space speaker adaptation with full covariance Gaussians, *Proc. Interspeech*, pp. 1145–1148 (2006).