英語習熟度を考慮した発音辞書と音響モデル逐次適応による非母語音声認識*

☆辻岡 聡, サクティ サクリアニ, 吉野 幸一郎, ニュービッグ グラム, 中村 哲 (奈良先端大)

1 はじめに

急速な国際化に伴い，国際会議などでは非英語母語話者が英語で意思疎通を図る場面が増加している．そのため，非母語話者の音声を認識して会議録を作成するなどの応用技術を考えた場合，非母語音声認識を高精度で認識する必要がある．これまでに著者らは，非母語音声認識において，非母語話者の音素認識結果を発音変換知識として用いた発音辞書学習による適応を提案している [1]．しかし，話者の英語習熟度に応じた適応は考慮されてこなかった．

そこで本研究では，英語習熟度別に行った音素認識結果を G2P (Grapheme-to-Phoneme) ツールの学習データに用いたデータ駆動型逐次発音学習を行うとともに，話者適応学習 (Speaker Adaptive Training: SAT) を組み合わせた手法を提案する．その結果，日本語母語話者の英語音声認識において提案手法が有効であることが確認できた．

2 確率的発音モデル

2.1 確率的発音モデルの定式化

確率的発音モデルでは，各単語の音響的特徴を直接モデル化するのではなく，各単語の発音をモデル化する発音辞書を用意し，この発音に対して音響モデルを定義する．よって観測された音響特徴量を $\mathbf{X}$，認識結果の単語列を $\hat{\mathbf{W}}$ とした時，従来の音声認識の定式から以下の式 (1) のように書き換えられる．

$$\hat{\mathbf{W}} = \underset{\mathbf{W}}{\operatorname{argmax}} P(\mathbf{W}) \sum_{B \in \Psi_{\mathbf{W}}} P(\mathbf{X}|\mathbf{B}) P(\mathbf{B}|\mathbf{W}) \quad (1)$$

ここで $\mathbf{B} = \{\mathbf{b}_1, ..., \mathbf{b}_n\}$ は単語列 $\mathbf{W} = \{\mathbf{w}_1, ..., \mathbf{w}_n\}$ の発音系列候補を表しており，単語列に対する各発音系列候補の確率を $P(\mathbf{B}|\mathbf{W})$ で表している．$\mathbf{b}_i$ は単語 $\mathbf{w}_i$ の発音である．$\Psi_{\mathbf{W}}$ は単語列 $\mathbf{W}$ の考えうる全ての発音系列候補の集合を表す．今回は各単語の発音は当該単語のみに依存すると仮定し，各単語の発音確率を以下の式のように表す．

$$P(\mathbf{B}|\mathbf{W}) = P(\mathbf{b}_1|\mathbf{w}_1) \cdots P(\mathbf{b}_n|\mathbf{w}_n) \quad (2)$$

各単語に複数の発音系列候補がある場合，それぞれに対して発音確率を付与する．

$$P(\mathbf{b}_i = \mathbf{p}_j|\mathbf{w}_i) = \theta_{ij}, \quad j = 1, \ldots, J_i \quad (3)$$

$$\sum_{j=1}^{J_i} \theta_{ij} = 1 \quad (4)$$

J_i は単語 $\mathbf{w}_i$ の考えうる発音系列候補の数であり，$\mathbf{p}_j$ とは発音確率 θ_{ij} を持つ発音系列候補を指す．

2.2 発音確率の更新

発音系列候補は通常 G2P で推定されるが，G2P の発音推定誤りは多々あり，誤った発音系列候補は音声認識に悪影響を及ぼす．この問題を解決するために，データ駆動型の逐次発音学習を用いることで，実音声から正しい発音系列候補を推定する手法が提案されている [2]．このデータ駆動型の逐次発音学習の過程を以下に説明する．

1. 初期学習発音辞書を G2P ツールの学習データとして使用し，G2P モデルの学習を行う．
2. 学習された G2P モデルから，各単語ごとに複数の発音系列候補を生成するとともに，全ての発音確率に同等の確率を付与する．
3. 等確率の発音確率が付与された発音辞書を用いて，音響モデルの学習を行う．
4. 学習データの実音声に対して認識を行い，認識結果の音素ラティスを取得する．
5. 認識された各単語における発音系列の音素ラティス出現回数を計算し，その単語の出現回数で割ることにより，発音確率を更新する．この際，更新された発音確率が閾値を下回った発音系列を削除する．
6. 更新された発音確率が付与された発音辞書を使用して，再度認識を行うとともに，音響モデルの再学習を行う．
7. 5 の過程を経て更新された発音辞書を，G2P ツールの学習データとして使用し，G2P モデルの再学習を行う．

これらの過程から発音確率が更新され，実音声に対して尤もらしい発音系列候補を選択する事が可能となる．この結果，話者の実際の発音に合わせた発音辞書を作成することができる．

3 英語習熟度を考慮した発音辞書生成法

本研究では，G2P ツールの学習データとなる初期学習発音辞書に着目する．我々の先行研究 [1] では非母語話者の音素認識結果を G2P ツールの学習データとして用いていた．しかし，Wang らにより，日本語母語話者における英語発音生成において，英語習熟度によって発音生成が異なることを明らかにしている [3]．[1] では，非母語話者の英語習熟度別に適応した発音学習を行っておらず，習熟度によって変化する発音に対応できない．

本手法では，一部の評価データの非母語音声の音素認識を英語習熟度別に行う．そして，各習熟度別の音素認識結果から単語とのアライメントを取ったものを G2P の学習データとして使用し，前節で述べた逐次発音学習を用いた発音辞書生成法を提案する．本手法では，人手による発音変換知識が無くても G2P を用いて非英語母語話者の英語習熟度を考慮した潜在的な発音候補を生成することが可能である．また，逐次発音学習の音響モデル学習の過程で，音響モデルでの代表的な適応手法である特徴量空間最尤線形回帰 (feature-space Maximum Likelihood Linear Regression: fMLLR) による SAT を用いた音響モデル学習を組み合わせた逐次適応手法を提案する．

---

*Non-native Automatic Speech Recognition Utilizing Acoustic Data-driven Pronunciation Learning and Acoustic Model Adaptation, by TSUJIOKA, Satoshi , SAKTI, Sakriani, YOSHINO, Koichiro, NEUBIG, Graham, NAKAMURA, Satoshi (NAIST)

Table 1　実験データ

学習データ		人数	時間	単語数 (千)
WSJ		282	82.9	370
ERJ	LOW	6	1.0	5.8
	MID	93	14.3	72.4
	HIGH	26	4.2	18.0
評価データ		**人数**	**時間**	**単語数 (千)**
ERJ	LOW	5	0.8	4.3
	MID	40	6.6	33.8
	HIGH	20	3.3	16.6

4　実験的評価

4.1　実験条件

本研究では Minematsu[4] による ERJ (English Read by Japanese) データベースの一部を学習・評価に用いる．このデータベースでは，日本人学生が読み上げた英語音声に対して英語母語話者の英語教師 5 名が (1) 音素生成 (2) リズム生成 (3) イントネーション生成の三つの観点から，1.0〜5.0 の範囲でスコアリングしている．我々はこの三つのスコアリングの加算平均を行い，発話者を三つの英語習熟度別に分割し，1.0〜2.5 を LOW（初級者），2.5〜3.5 を MID（中級者），3.5〜5.0 を HIGH（上級者）とした．

音声認識器は Kaldi tool kit[5] を使用し，音響モデルの特徴量は 39 次元の MFCC+Δ+ΔΔ を用いている．また，線形判別分析 (Linear Discrinative Analysis: LDA) と最尤線形変換 (Maxium Likelihood Linear Transform: MLLT) を用いた特徴量変換にもとづく次元数圧縮を行っており，対象フレームの前後 3 フレーム (計 7 フレーム) を考慮した音響モデル学習を行っている．学習データは音響モデル・言語モデルともに，WSJ (Wall Street Journal) と ERJ の一部を使用した．評価データには学習データに含まれていない ERJ の一部を使用した．また，音素認識においても同様の実験条件で行っている．これらのデータの詳細を Table 1 に示す．評価基準は単語誤り率 (Word Error Rate: WER) を用いる．英語母語話者の発音辞書には CMU 発音辞書を使用している．G2P ツールには，SequiturG2P を使用した [6].

4.2　実験結果

LOW，MID，HIGH それぞれの実験結果を Fig.1 に示す．縦軸は WER を表しており，横軸は逐次学習のそれぞれの状態を表している．Initial は各単語の発音候補に対して等確率の発音確率が付与された状態での認識精度，Step-5 は更新された発音確率を用いて認識を行った際の認識精度，Step-6 は音響モデルの再学習を行った際の認識精度を表している．

従来の発音辞書を拡張しない場合の WER は，LOW では 52.8%，MID では 36.5%，HIGH では 23.8%となった (Baseline)．これらを基準として，提案手法の評価と分析を行う．また，比較実験として，音響モデルの代表的な適応手法である fMLLR による SAT を用いた認識評価実験を行った場合も示す．この WER は，LOW では 45.6%，MID では 33.0%，HIGH では 21.3%となった (SAT)．また，我々の先行研究 [1] である英語習熟度を考慮しない逐次発音学習手法を No-KnowledgeG2P としている．

提案手法である英語習熟度を考慮した逐次発音学習手法をそれぞれ，LOWG2P，MIDG2P，HIGHG2P とし，さらに話者適応学習を組み合わせた手法を LOWG2P + SAT，MIDG2P + SAT，HIGHG2P + SAT とする．

その結果，提案手法である英語習熟度を考慮した発音辞書と SAT を組み合わせた手法の WER が，LOW では 32.7%，MID では 27.7%，HIGH では 17.4%となり，各英語習熟度において認識精度が改善していることが確認できた．また，SAT のみを用いた手法と比較しても，提案手法の有用性が確認できた．

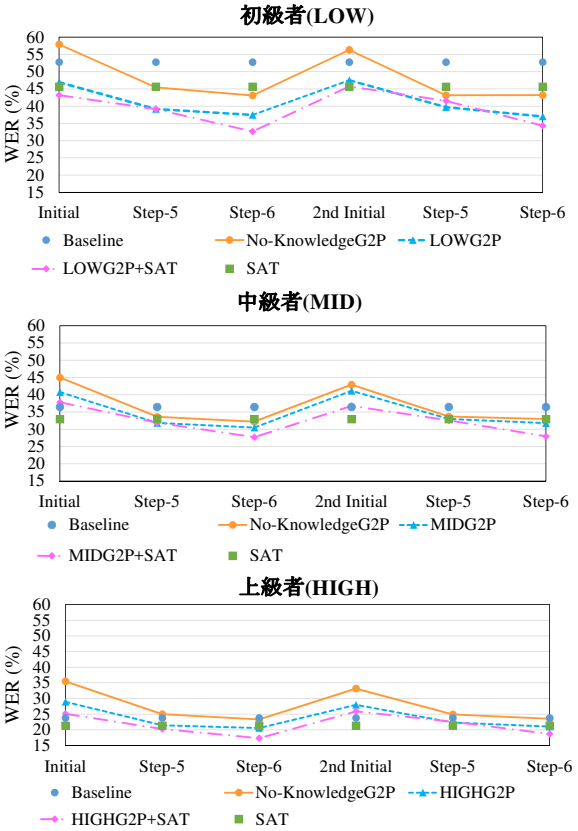

Fig. 1　LOW, MID, HIGH における評価実験結果

5　まとめ

本稿では，非母語音声の認識精度改善のために，非英語母語話者の英語習熟度を考慮した発音辞書生成法と音響モデル適応を組み合わせた手法を提案した．実験的評価から，提案手法が先行研究 [1] と比較して各習熟度において認識精度を向上させることを確認できた．また，従来の SAT のみを用いた場合と比べて，提案手法が有効であることが確認できた．

今後は，非母語話者の英語習熟度の自動判別を行うとともに，本稿の提案手法を用いて日本人以外の複数の非英語母語音声の認識に適用する手法を検討する．

謝辞　本研究の一部は，(独) 情報通信研究機構の委託研究「知識・言語グリッドに基づくアジア医療交流支援システムの研究開発」および JSPS 科研費 24240032 および 26870371 の助成を受け実施した．

参考文献

[1] 辻岡聡 他. 情報処理学会研究報告, Vol. 2015–SLP–109, No.14, pp. 1–6, Dec. 2015.
[2] L.Lu *et al.*, in *Proc.ASRU*, pp. 374–379, 2013.
[3] X.Wang *et al.*, *Interspeech* 2015, pp. 1265–1269, 2015
[4] N.Minematsu *et al.*, in *Proc.LREC* 2002, pp. 896–903, 2002
[5] D.Povey *et al.*, in *Proc.ASRU*, 2011.
[6] M. Bisani *et al.*, Speech Communication, vol. 50, No.5, pp. 434–451, 2008.