Vol.2017-SLP-118 No.3
2017/10/13

ベイズ推論を用いた半教師あり発音辞書学習の日本語適用

池下 裕紀[1]　篠崎 隆宏[1]　渡部 晋治[2]　持橋 大地[3]　Neubig Graham[4]

概要：現在の音声認識システムにおいて発音辞書は人手による設定に基づいている．しかし日々出現する新しい単語のシステムへの取り込みや，ロボット等における音声対話を通した語彙の自動獲得を実現するためには，システム自身が言語的および音響的知識を結びつけ新しい単語を推定を交えながら獲得する能力を実現する必要がある．これまでに我々はノンパラメトリックベイズ法を応用し，同一の言語から独立してサンプルされた音素および単語テキストから発音辞書を学習する手法を提案し，英語データに対して有効性を示した．本研究では同手法を表音文字を用いる言語である英語に対して，表音文字と表意文字が混ぜて使用されるという点で大きく異なる日本語に適用する．実験には日本語話し言葉コーパス (CSJ) を用い，同手法の日本語における有効性について検討する．

1. はじめに

近年, 情報技術の飛躍的な進歩に伴い，一部のタスクにおいては人に匹敵する認識性能が得られるようになりつつある．しかしその一方で，それらの性能は教師あり学習への依存度が大きく，予め想定されたものとは異なるタスクへの応用や新たな単語への対応を行う場合，ラベリングの行われた音声データと新出単語の発音情報の準備が必要となる．これにより，当初に設定したドメインへのシステムの有用性が制限されるということがしばしば発生する．これに対して，人は発音が未知の単語に対しても音声情報と文章情報を元にその単語の発音を学習することが出来る．読みの分からない単語が文章中に出現した場合，とりあえず読みを不明としたままその単語が使用されているコンテキストを学習し，後にその単語が音声中に出現した際にそのコンテキストを元に綴りと発音を結びつけることで単語の発音を学習する．これは音声と文章が対になっていない場合においても，同様のコンテキストの元にその単語が出現すれば可能である．この能力を発音辞書に拡張することにより，ラベリングの行われたデータを必要としない音声認識システムを実現することが出来れば，システムのメンテナンスコストの削減だけに留まらず，機械と人間の間でのより自然なコミュニケーションを可能にすることができる．

一般的に，音声認識システムにおいて発音辞書は二つの役割を担っている．一つは単語の単位を定義するという働きであり，もう一つは単語の発音から綴りへのマッピングを行うという働きである．近年の end-to-end 音声認識では，文字単位での認識を行うことによってこれらの働きを必要としない枠組みも存在する [1]．しかし，依然として多くの音声認識アプリケーションが単語レベルでの認識を必要としており，様々なアプローチが行われている．これに対する取り組みの一つとして，教師なし発音辞書学習が挙げられる．教師なし発音辞書学習に関連する既存のアプローチとしては，書記素音素変換 (G2P)[2], [3], [4], [5] が挙げられる．このアプローチでは対応関係のある単語と音素列を用いて G2P 変換器の教師あり学習を行う．そして学習の完了した G2P 変換器を発音が未知の単語表記に適用することで，その単語の発音を獲得する．しかしこのアプローチの問題として，表記から直接発音を推定することが困難な単語には適用出来ないという点がある．音響データから発音を推定する手法も提案されているが，対応のあるパラレルデータを必要としている点で一般的な利用が難しい制約がある．これに対し，対応関係の存在しないテキストと音声から発音辞書を推定する手法が実現できればより柔軟で一般的な音声認識システムの語彙適応が可能となる．これについて，我々はこれまでに発音辞書を確率モデルとしてモデル化しベイズ推論により発音辞書を学習する教師なし発音辞書学習法 [6] を提案し，一定の成果を上げている．本論文ではこの枠組みを日本語に適用し，その有効性について検証を行う．

[1] 東京工業大学
[2] Johns Hopkins University
[3] 統計数理研究所
[4] Carnegie Mellon University

2. ベイズ推論による教師なし発音辞書学習

2.1 発音辞書の確率モデル化

音声認識システムにおける発音辞書は単語とその発音の対から構成される．発音辞書によって各単語についてその発音に対応した音素列が与えられる．しばしば，出現頻度に応じた確率重みを持たせることで同単語が複数の発音を持つことができるように発音辞書は設計される．確率モデルとして見ると，これは各単語において混合要素が発音，混合重みが確率として構成された有限混合モデルとして考えることができる．文献 [7] より混合重みは学習可能パラメタであり，EM アルゴリズムにより推定される．また，文献 [8] ではベイズ的アプローチがとられており，混合重みに事前確率を与えるためにディリクレ分布が用いられている．

1 単語あたりの発音の種類数は多くても数個程度であることが普通だが，上限が存在するわけではない．また，単語の発音を絞り込むための知識がない場合は，様々な発音の可能性を考慮する必要がある．そこで各単語にどのような発音でも割り当てられるように，有限混合モデルから無限混合モデルへと拡張する．具体的には，単語 w の発音を式 1 に示すように音素列を要素とする無限次元の離散分布としてモデル化する．

$$f(x) = \sum_{i=1}^{\infty} \theta_i \delta_{p_i}(p). \tag{1}$$

ここで，$\delta_{p_i}(p)$ は $p = p_i$ のとき 1，そうでないとき 0 をとるデルタ関数である．また，θ_i は分布のパラメタであり $0 \leqq \theta_i$，$\sum_{i=1} \theta = 1$ を満たす．例として，「情報」の発音が「j o: h o:」と「j o u h o u」の 2 通りだけでそれぞれの確率が 0.8 と 0.2 とすると，p_1 = 「j o: h o:」，p_2 = 「j o u h o u」，$\theta_1 = 0.8$，$\theta_2 = 0.2$，$\theta_{(2<)j} = 0.0$ となる．発音分布の事前分布としては，式 2 に示すように音素列を生成する分布を基底分布 G_0 としたディリクレ過程 [9] が使用できる．

$$\begin{aligned} G_w \sim DP(\alpha, G_0). \quad (i.i.d.w \in V), \\ p_w(p) = G_w, \end{aligned} \tag{2}$$

ここで，α は集中度パラメタであり，$P_w(p)$ が平均的にどれくらい G_0 と似ているかを制御する．α が 0 に近づくほど分布が特定の発音に確率が偏ったものになりやすくなる．

発音辞書は単語発音モデルの集合として，式 3 のように定義できる．

$$\Delta = \{G_w\}_{\{w \in V\}}. \tag{3}$$

ここで，Δ は発音辞書を表し，G_w は発音モデルを表す．

発音辞書に基づいて発音を予測する予測分布は中華料理店過程 (CRP)[10] により求められる．発話集合 U を観測した時，単語 w の発音 p の予測分布 $P(p|w, U)$ は，式 4 によって与えられる．ここで，n_w は U 内における単語 w の出現回数，n_p は単語 w の発音として音素列 p が出現した回数であり，$\sum n_p = n_w$ である．

$$P(p|w, U) = \frac{n_p}{\alpha + n_w} + \frac{\alpha}{\alpha + n_w} G_0(p) \tag{4}$$

2.2 発音辞書の学習

単語列に関する言語的情報，音素書き起こしにおける音素列のパターン情報，および単語発音に関する部分的な既存の知識をつなぎあわせて教師なし発音辞書学習を行うベイズモデルを図 (1) に示す．具体例としては，「実験 の 方法 は」という文章が単語列となり，「j i q k e N </w> n o </w> h o: h o: </w> w a」という文字列が単語区切りのある音素列となる．ここで，</w> は単語間の境界を示す．そして，「j i q k e N n o h o: h o: wa」という文字列が単語区切りのない音素列となる．

表 1 図 1 における各ノードの説明

ノード	説明
発音辞書	確率モデル化された発音辞書
言語モデル	階層ベイズ言語モデル
単語列	1 発話に相当する単語列
単語区切りのある音素列	単語区切りの与えられた 1 発話の音素列
単語区切りのない音素列	単語区切りのない 1 発話の音素列

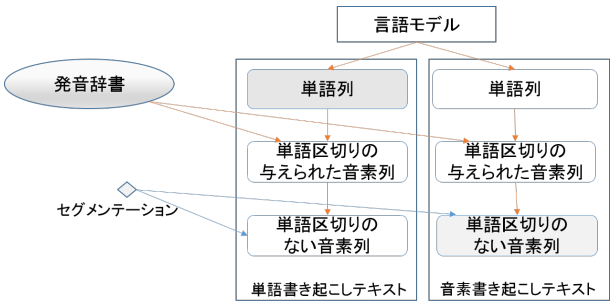

図 1 対応関係のない単語および音素データを用いた教師なし発音辞書学習のためのベイズモデル

ベイズモデルを用いた発音辞書の学習は，学習データが与えられた条件で隠れ変数の事後確率分布を求めることである．これは，発話を単位としてブロックギブスサンプリング [11], [12] を適用することで実現する．ブロックギブスサンプリングでは，まずベイズモデルの隠れ変数に対して適当な初期値を設定する．そしていずれか 1 つの発話を選択し，その発話の隠れ変数の値を他の発話の全ての変数を固定した同時事後確率分布からサンプルすることで更新する．これらを繰り返すことにより，学習データが与えられた条件での事後確率に従ったサンプル集合が得られる．

ブロックギブスサンプリングにおいて必要となる，他の

全発話を固定した条件下でのある発話における発話変数の同時事後確率は式 5 によって得られる．

$$\begin{aligned}&(nP_s, sP_s, W_s | nP_T, sP_T, W_T)\\&= \int P(nP_S, sP_s, W_s, LM, PD | nP_T, sP_T, W_T) dLM dPD\\&= P(W_s|W_T) P(sP_S|W_s, W_T, sP) P(nP_s|sP_s) \qquad (5)\end{aligned}$$

ここでサフィックス s は選択された発話に関する変数であることを示し，サフィックス T は他の全ての発話における変数集合であることを示す．選択した発話において与えられている学習データを反映させた事後確率は式 5 の同時確率から求められる．

3. 発音辞書学習の日本語適用

本論文では，表音文字を用いる英語において一定の成果をあげているベイズ推論による教師なし発音辞書学習の他言語に対する有効性の検証を行う．今回は，表音文字と表意文字が混同して用いられるという点で英語と大きく異なる日本語への適用を行う．日本語書き起こしデータから独立に単語列データと発音辞書，音素列データを生成し，発音辞書学習を行う．

4. 実験条件

実験は日本語話し言葉コーパス (CSJ) を用いて行った．音素列データとしては，認識誤りを含まない音素書き起こしテキストを用いた．音素列データに含まれる音素数は全部で 43 である．発音辞書の語彙は単語書き起こしテキストから作成し，その内の 85%，70%にのみ発音を与えた．残りの 15%の単語の発音は初期値としては与えられず，これらを推定することが本論文におけるタスクとなる．発音の基底分布としては音素 0-gram モデルを使用した．集中度パラメタ α は 0.1 に設定した．ギブスサンプリングでは，音素列が与えられた条件における事後分布からのサンプリングにおいて，語彙の固定を行った．これにより，本実験では音素列から表記が不明の単語が生成されることは無く，綴りのみが分かっており発音が未知である単語の推定にフォーカスしている．実験に用いたソフトウェアは，LatticeWordSegmentation をベースとして作成したものである [13], [14], [15]．また，比較実験として Sequitur G2P[16] を用いた G2P による手法での評価を行った．

5. 実験結果

CSJ 講演データのうち 4286 発話を利用して単語書き起こしテキストを生成し，1857 発話を利用して音素書き起こしテキストを生成した．これらの間に重複する文章は含まれず，完全にオープンなデータである．語彙サイズは 5452 であり，このうち 15%, 30%には初期値として発音を与えていない．言語モデルは 3-gram を用いた．パープレキシティーは 150.64 であり，OOV は 3.87%である．ここでは，OOV は単語列テキストに含まれず，音素列テキストにのみ含まれる単語のことを指す．G2P による実験では初期値として与えられていない 15%, 30%の単語について G2P を適用し，発音辞書を生成した．また，提案法の拡張として G2P で生成した発音辞書を初期値として与えた実験 (Bayes+G2P) を追加で行った．

表 2 オープンデータセットにおける単語誤り率

Init condition	15%missing			30%missing		
epoch	1	2	5	1	2	5
default	25.79			37.74		
G2P	22.58			29.16		
Bayes	25.72	23.75	23.29	37.49	34.47	33.76
Bayes+G2P	22.55	21.29	21.06	29.00	27.33	26.44

表 3 サンプリングされた文章の例

Reference	ここ で は 見せ ませ ん が 評価 話者 女性 の 二人 に
Epoch1	ここ で ． ま 線 が 評価 わ 者 女性 の 二人 に
Epoch2	ここ で は 見せ ませ ん が 評価 わ 者 女性 の 二人 に
Epoch3	ここ で は 見せ ませ ん が 評価 話者 女性 の 二人 に

表 (2) にそれぞれ 15%, 30%における結果を示す．初期条件の発音辞書での単語誤り率はそれぞれ 25.79%, 37.74%であった．ギブスサンプリングによる学習を行うことにより，単語エラー率が減少していくことがわかる．G2P と提案法を組み合わせることによってさらに単語誤り率が低下し，それぞれ 21.06%, 26.44%という結果が得られた．表 (3) にサンプリングの際に音素列から変換されて得られた単語列の例を示す．エポック 1 の時点では発音未知の単語に対して誤った単語が出力されているが，エポック 5 では発音の割り当てが成功し，その結果として正しく文章が出力されていることが分かる．

6. おわりに

対応の無い単語テキストと音素書き起こしデータを元に発音辞書の教師なし学習を行う手法を用い，日本語への適用を行った．日本語話し言葉コーパスを用いた実験を行い，音素列から単語並びへの変換において単語誤り率が減少することを実証した．また，初期辞書として G2P による発音の予測を組み合わせることにより，更に効果的に辞書学習を行えるということが確認できた．今後の課題としては英語データにおける実験と比較して性能の向上幅が小さいため，その原因の究明を行うことや，音素ラティスのエンコードを行う WFST を 1-best 仮説の代わりに用いるといった手法の拡張などが挙げられる．

7. 謝辞

This work was supported by JSPS KAKENHI Grant

Numbers 17K20001 and 26280055. This work was inspired by insights we gained from JSALT 2016, which was partially supported by Johns Hopkins University via DARPA LORELEI Contract No HR0011-15-2-0027, and gifts from Microsoft, Amazon, Google, and Facebook.

参考文献

[1] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks." in International Conference on Machine Learning (ICMK), 2014, pp. 1764-1772.
[2] Bisani, M. and Ney, H.: Joint-sequence models for grapheme-to-phoneme conversion, Speech Communication, Vol. 50, No. 5, pp. 434-451 (2008).
[3] Chen, S. F. et al.: Conditional and joint models for grapheme-to-phoneme conversion., INTERSPEECH (2003)
[4] Taylor, P.: Hidden Markov models for grapheme to phoneme conversion., INTERSPEECH, pp. 1973-1976 (2005).
[5] Novak, J. R., Minematsu, N and Hirose, K.: WFST-based grapheme-to-phoneme conversion: open source tools for alignment, model-building and decoding, 10th International Workshop on Finite State Methods and Natural Language Processing, p. 45 (2012).
[6] T. Shinozaki, S. Watanabe, D. Mochihashi, G. Neubig, "Semi-supervised learning of a pronunciation dictionary from disjoint phonemic transcripts and text" in INTERSPEECH, 2017, pp.2546-2550.
[7] I. McGraw, I. Badr, and J. R. Glass, "Learning lexicons from speech using a pronunciation mixture model," IEEE Transactions on Audio, Speech and Language Processing, vol. 21, no. 2, pp.357-366, Feb 2013.
[8] C. Lee, Y. Zhang, and J. R. Glass, "Joint Learning of Phonetic Units and word pronunciations for ASR." in Empirical Methods in Natural Language Processing (EMNLP), 2013, pp. 182-192.
[9] Teh, Y. W.: Dirichlet Processes, Encyclopedia of Machine Learning, Springer, (2010).
[10] Pitman, J.: Exchangeable and partially exchangeable random partitions, Probability theory and related fields, Vol. 102, No. 2, pp. 145-158 (1995).
[11] Gelfand, A. E. and Smith, A. F.: Sampling-based approaches to calculating marginal densities, Journal of the American statistical association, Vol. 85, No. 410, pp.398-409 (1990).
[12] Liu, J.: The Collapsed Gibbs Sampler in Bayesian Computations with Applications to a Gene Regulation Problem, Journal of the American Statistical Association, Vol. 89, No. 427 (1994).
[13] J. Heymann, O. Walter, R. Haeb-Umbach, and B. Raj, "Iterative Bayesian word segmentation for unsuprevised vocabulary discovery from phoneme lattices," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014, pp. 4057-4061.
[14] J. Heymann, O. Walter, R. Haeb-Umbach, and B. Raj, "Unsupervised word segmentatinon from noisy input," in IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), 2013, pp. 458-463.
[15] G. Neubig, M. Miura, S. Mori, and T. Kawahara, "Learning a language model from continuous speech." in INTERSPEECH, 2010, pp.1053-1056.
[16] M. Bisani and H. Ney, "Joint-sequence models for Graphemeto-phoneme conversion," Speech Communicatinon, vlo.50, no. 5, pp. 434-451, 2008,