Vol.2016-SLP-114 No.4
2016/12/20

第３回Frederick Jelinek記念サマーワークショップでの教師なし発音辞書学習の取り組み

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

概要：アメリカジョンズホプキンス大学で開催された第３回 Frederick Jelinek 記念サマーワークショップにおいて，我々は教師なし音声認識に関するプロジェクト「Building Speech Recognition System from Untranscribed Data」に参加し，教師なし発音辞書学習の実現に向けた取り組みを行った．単語ごとの発音を大きく偏った無限混合モデルとして表しその事前確率をディリクレ過程により与えることで，発音辞書全体を確率モデルとして扱う手法を提案した．予測分布は中華料理店過程を用いることで求めることができ，学習は言語モデルや学習データを含めてベイジアンネットワークによりモデル化しブロックギブスサンプリングを適用することで行える．また実験により，限定的ではあるものの本手法を用いることで同一の言語からサンプルされたアライメントのない音素列と単語列をもとに発音辞書の学習が可能であることを示した．

1. はじめに

近年深層学習の発展などにより音声認識性能が向上し，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，現在の音声認識技術は認識性能のタスクへの依存度が大きく，発話スタイルやトピックへの適応力に問題がある．この問題に対処るるためには音響的および言語的な様々な適応技術を発展させる必要がある．中でも特にチャレンジングな課題の一つとして，発音辞書の教師なし学習が挙げられる．

すなわち，システムを予め想定されたタスクとは異なるタスクに応用しようとする場合や，同じタスクでも時間の経過により新しい単語が増加した場合，未知語への対処が重要となる．人であれば，読みの分からない単語がテキスト中に出現した時とりあえず読みを不明としたままその単語が使用されるコンテキストを学習し，後日その単語が音声中に出現した時にそのコンテキストを元に表記と発音を結びつけて新しい単語の発音を獲得できる．そしてこのプロセスを繰り返すことで，必ずしも辞書を参照することなく，徐々に語彙を拡大していくことができる．このような機能は，自動的に様々なタスクあるいはタスクの変化に適応可能なシステムを実現する上で，決定的に重要となる．しかし，現在実用されているシステムには，そのような学習の仕組みが入っていない．そのため，そもそも新しい単語の追加登録は全くできないか，できる場合でも人手で発音を用意する必要がある．

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) が挙げられる [1], [2], [3], [4]．このプローチでは，対応関係のある単語と音素列を用いてG2P 変換器を予め教師あり学習する．そしてその G2P 変換器を発音が未知の単語表記に適用することで，その単語の発音を得る．このアプローチの問題点は，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できないことである．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 (e.g. [5])，パラレルデータを必要としている点で一般的な利用ができない制約がある．

もう一つのアプローチは，未知語検出 [6], [7] と音素認識の組み合せである．すなわち入力音声を音素認識機を用いて音素列に変換するのと並行して未知語検出を行う．そして未知語として検出された時間区間に対応する音素列をその未知語の発音とする．この方法では未知語の発音を音声から学習できるが，学習した未知語に文字表記を与えることはできない問題がある．

G2P アプローチではテキストを元に未知語の発音を推定し，未知語検出に基づいたアプローチでは音声を元に未知語の発音を推定している．また対応関係のあるテキストと音声から発音辞書を求めることは，アライメントを求めることで容易に行える．これに対して，対応関係の存在しな

[1] 東京工業大学
[2] MITSUBISHI ELECTRIC RESEARCH LABORATORIES (MERL)
[3] 統計数理研究所
[4] Carnegie Mellon University

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認識システムの語彙適応が可能となる．これはすなわち，同じ言語から独立してサンプルされた音声とテキストから発音辞書を推定する問題である．我々は Frederick Jelinek 記念サマーワークショップにおいてこの問題に取り組み，発音辞書を確率モデルとしてモデル化しベイズ推論 [8] により発音辞書を学習する教師なし発音辞書学習法を提案した．さらに対応のない単語テキストと音素書き起こし文を用いた実験を行い，提案法を用いた学習が実際に可能であることを実証した．本論文では，本年度実施されたサマーワークショップの概要と，我々が取り組んだ教師なし発音辞書学習法について報告する．

2. Frederick Jelinek 記念サマーワークショップについて

Frederick Jelinek 記念サマーワークショップ [*1]は，1995 年から続いたジョンズホプキンス大学サマーワークショップの後継ワークショップである．Frederick Jelinek 記念サマーワークショップとしては今年が 3 回目である．本ワークショップでは音声や自然言語処理を中心に画像処理など関連した広範な分野を対象として研究者や学生が集まり，チームを構成して共同研究を行う．研究テーマは公募制である．本年のワークショップに対するプロジェクトの公募は，昨年 10 月に行われた．その後書類選考と 2 日間に渡るピアレビューミーティングによりプロジェクトの絞り込みが行われた．ピアレビューミーティングでは，プロジェクトの提案者同士が議論を交わしながらプロジェクト案を改良しそれを外部からの審査委員も加わった評価発表会により評価するプロセスが繰り返された．そしてその過程においてプロジェクト案の統廃合が行われた．本年のワークショップにおいては最終的に得票数の多い 3 つのプロジェクトが採択され，ジョンズホプキンス大学において 6 月 27 日から 8 月 5 日の日程で開催されたワークショップにおいてそれらのテーマに対する取り組みが行われた．我々のサブチームによる教師なし発音辞書学習の研究は，その中の一つのプロジェクト「Building Speech Recognition System from Untranscribed Data」の取り組みの一環である．

3. ベイズ推論による教師なし発音辞書学習法の提案

3.1 発音辞書モデル

音声認識システムにおいて，発音辞書は認識語彙となる各単語についてその発音に対応した音素列を与えるものである．1 つの単語は，1 つあるいは複数の発音を持つ．通常の認識システムでは固定した発音辞書が用いられるが，ベイズ推論を用いた学習を行うためにはまず発音辞書を確率モデルとして定式化することが必要となる．単語の発音は短いものも長いものもあり，将来出現する可能性のある全ての単語まで考慮しようとすれば固定の最大長は存在しない．そのため，他の制約がなければ単語の発音となり得る音素列には無限の可能性がある．また，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であるが，こちらも決まった上限が存在するわけではない．

そこで，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ことを提案する．

$$\mathrm{P_w}\left(p\right) = \sum_{i=1}^{\infty} \theta_i \delta_{p_i}\left(p\right). \tag{1}$$

ここで，$\delta_{p_i}\left(p\right)$ は $p = p_i$ のとき 1 そうでない時 0 をとるデルタ関数である．また，θ_i は $0 \leq \theta_i, \sum_i \theta = 1$ を満たす分布のパラメタである．例として，「音声」の発音が「o N s e:」と「o N s e i」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 = “o n s e:”, p_2 = “o n s e i”, $\theta_1 = 0.8, \theta_2 = 0.2, \theta_{(2<)j} = 0.0$ 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 としたディリクレプロセス [9] が使用できる．

$$\mathrm{P}\left(\mathrm{P_w}\left(\right)\right) = DP\left[\mathrm{P}_w\left(\right)|\alpha, G_0\right]. \tag{2}$$

ここで，$(0<)\,\alpha$ は集中度パラメタであり，$\mathrm{P_w}\left(p\right)$ 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 0.0 に近づくほど分布は特定の発音に確率が偏ったものになる．

発音辞書は単語発音モデルの配列とみなすことができ，発音辞書の事前分布は単語事前分布の積として式 (3) により定義できる．

$$\mathrm{P}\left(PD\right) = \prod_{w \in V} DP\left[\mathrm{P}_w\left(\right)|\alpha, G_0\right], \tag{3}$$

ここで PD は発音辞書であり， V は語彙である．基底測度は全単語で共通としている．

発話集合 U を観測したとき，単語 w において発音 p の出現を予測する予測分布 $\mathrm{P}\left(p|w,U\right)$ は，式 (4) により与えられる．ここで Θ はパラメタの集合 $\{\theta_1, \theta_2, \theta_3, \cdots\}$ である．式 (4) における積分は，中華料理店過程 (CRP) [10] により求められ，式 (5) となる．ここで n_w は U の中での単語 w の出現回数，n_p は単語 w の発音として音素列 p が回出現した回数であり，$\sum n_p = n_w$ である．

$$\begin{aligned} &\mathrm{P}\left(p|w,U\right) \\ &= \int \mathrm{P}\left(p|w,\Theta\right)\mathrm{P}\left(\Theta|U\right)d\Theta \qquad (4) \\ &= \frac{n_p}{\alpha + n_w} + \frac{\alpha}{\alpha + n_w} G_0\left(p\right). \qquad (5) \end{aligned}$$

3.2 発音辞書の学習

対応のある単語テキストと単語区切りのある音素列が学

[*1] http://www.clsp.jhu.edu/workshops/16-workshop/

表 1 図 1 における確率変数のリスト.

確率変数	説明
Pronunciation dictionary (PD)	ディリクレプロセスの配列として実現された発音辞書
Language model (LM)	ピットマンヨー言語モデル
Word sequence (W)	1 発話に相当する単語列
Segmented phone sequence (sP)	単語区切りの与えられた 1 発話の音素列
Phone sequence (nP)	単語区切りの無い 1 発話の音素列

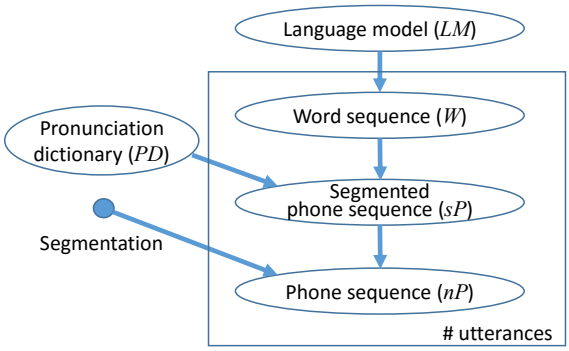

図 1 アライメントのない単語および音素データを用いた教師なし発音辞書学習のための全ベイズモデル.

習用パラレルデータとして与えられれば，単語発音の予測分布は直接式 (5) により容易に求められる．しかし，本研究が目的とする教師なし発音辞書学習を実現するためにはそのようなパラレルデータの使用は仮定できず，同じ言語から独立してサンプルされたアライメントのないテキストと，単語区切りのない音素書き起こしを用いて学習を行う必要がある．そのためには，テキストから得られる単語列と，音素書き起こしにおける音素列の出現パターンの両方の情報を双方向に活用する必要がある．直感的には，発音が分かっているある単語コンテキストの元で発音が未知のある単語の出現確率が高いことが分かっていれば，音素書き起こし中でその単語コンテキストに対応する音素列の前後の間の音素列がその単語の発音である可能性が高いと推測できる．ただし，単語パープレキシティが 1 でない限り，候補は一意には絞れない．しかし，同じ未知語が複数のコンテキストのもと複数回出現すれば，可能性を絞っていくことができると期待される．

単語列に関する言語的な知識と，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仕組みとして，図 1 に示す全ベイズモデルを提案する．また，表 1 にモデルで使用する確率変数の一覧を示す．ここで，LM は言語モデルである．データに応じて語彙を決める仕組みとするために，言語モデル自体も確率変数とする．具体的にはピットマンヨー言語モデル [11] を用いる．W は 1 発話を表す単語列であり，言語モデルより生成される．PD は，発音辞書であり，第 3.1 章において提案したモデル化法を用いる．sP は単語区切りの与えられた音素列であり，単語列と発音辞書から生成される．nP は単語区切りのない音素列であり，単語区切りの与えられた音素列から単語区切りを削除したものである．単語区切りの与えられた音素列と単語区切りのない音素列の関係は，セミマルコフモデル [12] として定式化される．これは，教師なし形態素解析における単語列と文字列との関係と同様である [13]．言語モデルと発音辞書は，全学習・評価用発話データに対して共通である．それに対して，単語列，単語区切りの与えられた音素列，および単語区切りのない音素列は発話毎の確率変数である．これらの発話を表す変数は，各発話において対応するデータが与えられた場合は観測変数となり，与えられない場合は隠れ変数となる．

このモデルでは，どの発話変数に観測値が与えられ，どの発話変数に観測値が与えられないかは発話ごとに任意である．本研究において目的としている教師なし発音辞書学習は，単語列のみが与えられた発話データと単語区切りのない音素列のみが与えられたデータを用いる条件での学習となる．もし同一の発話について単語列と音素列が与えられた条件で発音辞書を学習する場合は，教師あり学習に相当する．

提案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 [14], [15] 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．そしてどれか一つの発話を選択し，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事後確率分布からサンプルすることにより更新する．発話の選択とその隠れ変数の値の更新を繰り返すことで，学習データが与えられた条件での同時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おいて必要となる，他の全ての発話を固定した条件でのある発話におけるの発話変数の同時事後確率は，式 6 により得られる．

$$\begin{aligned}
&\mathrm{P}\left(nP_s, sP_s, W_s | nP_T, sP_T, W_T\right) \\
&= \int \mathrm{P}\left(nP_s, sP_s, W_s, LM, PD | nP_T, sP_T, W_T\right) dLM dPD \\
&= \mathrm{P}\left(W_s | W_T\right) \mathrm{P}\left(sP_s | W_s, W_T, sP_T\right) \mathrm{P}\left(nP_s | sP_s\right), \qquad (6)
\end{aligned}$$

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表し，サフィックス T は他の全ての発話に関する変数集合であることを表す．選択した発話において与え

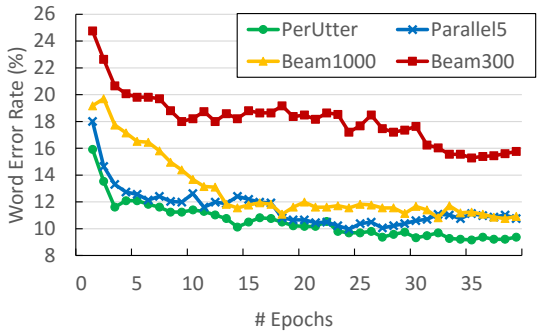

図 2 2-gram 言語モデルを用いた場合の単語誤り率.

られている学習データを反映させた事後確率は，式 6 の同時確率から求められる．データとして単語列が与えられた場合，ベイズモデルの構造から nP_s や sP_s は値をサンプルする代わりに周辺化により削除することも可能である．それにより，計算量を減らすことができる．

4. 実験条件

実験は WSJ コーパスおよび CMU 辞書を用いて行った．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．CMU 辞書では 1 音素はアルファベットの組み合わせで表現されているが，実験では始の 1 文字を残して他のアルファベットを一部実装上の都合から削除して縮約している．そのため，音素列から単語への変換における曖昧性が増加している．縮約した音素セットの音素数は 25 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%のみに発音を与えた．残りの 15%の単語の発音は初期値としては与えられず，それらを推定することがタスクとなる．発音の基底分布としては，音素 0-gram を用いた．発音長は 0 は含めず 1 以上とした．集中度パラメタ α は 0.001 に設定した．ブロック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，語彙は固定とした．これにより，本実験では音素書き起こしデータから表記の不明な単語が生成されることは無いこととし，表記は既知であるが発音が未知の単語の発音を推定することのみにフォーカスしている．実験に使用したソフトウエアは，LatticeWordSegmentation [16], [17], [18] をベースに用いて作成した．

5. 実験結果

図 2 に，ギブスサンプリングのエポック数と単語誤り率 (WER) の関係を示す．学習に用いたデータは，100 発話分の音素書き起こしと 100 発話の単語テキストである．これら音素書き起こしと単語テキストは，コーパス中の同じ発話セットに対応するものである．しかし，音素書き起こしと単語テキストを発話ごとに対応付ける情報はシステムに与えていない．語彙サイズは 849 で，初期条件として発音を与えたのはその内の 742 単語である．言語モデルは 2-gram であり，パープレキシティーは 30.8 である．単語書き起こしが与えられた条件においては他の発話に関する隠れ変数は周辺化により消去しており，1 エポックにおいてサンプリングを行っている対象は音素書き起こしが与えられた 100 発話である．単語誤り率は，サンプリングにより得られた単語列に対して計算している．

表 2 サンプルされた単語列（一部分）の例．単語 “fed” および “assets” は初期設定において未知である．

正解文	strategy to sell off assets and
epoch1	strategy to sell a fuss eight fed and
epoch2	strategy to sell officer bids and
epoch3	strategy to sell off assets and

ギブスサンプリングでは，いくつかのサンプリング戦略を試みた．図に示す “PerUtter” は，ブロックギブスサンプリングを 1 発話ごとに行っていることを示している．“Parallel5” では，5 発話を並列して処理し，変数の更新を 5 発話ごとに行っている．“Beam1000” および “Beam300” は，ビームサーチにより得られた仮説リストからのサンプリングにより近似を行った場合の結果である．1000 および 300 はビーム幅であり，探索中に保持する仮説数である．

図に示すように，一番低い単語誤り率は探索戦略として PerUtter を用いた場合に得られ，9.2%であった．Parallel5 はそれよりもやや高い単語誤り率となった．ビームサーチを行った場合，ビーム幅 1000 ではエポックを増やすことで Parallel5 と同程度まで単語誤り率が小さくなった．しかしビーム幅を 300 とした場合には，単語誤り率は初期値よりはよくなるものの，他の条件と比べて大きな値となった．計算時間は Xeon E5-2630v2 CPU を用いた場合で PerUtter, Beam1000, および Beam300 がそれぞれ 1 エポックあたり 161, 90, および 32 分であった．また Core i7 6800k CPU を用いた場合，PerUtter で 122 分，Parallel5 で 42 分であった．表 5 に，音素書き起こしを入力として得た単語列のサンプル例を示す．サンプリングは PerUtter により行っている．エポックを 3 回繰り返したところで，登場する単語について正しい発音が学習され，正しい単語列が得られている．

図 3 にエポック数と辞書誤り率の関係を示す．辞書誤り率は，サンプルした辞書中において誤った発音の割合である．挿入誤りは間違った発音があること，削除誤りは正しい発音が欠けていることを意味する．トータル誤り率はそれらのエラー率の和である．この実験においては，言語モデルには 3-gram を使用し，5 発話を並列処理した．図より，エポックを繰り返すことで，辞書エラー率が減少していくことがわかる．エポックの繰り返しを 30 回まで行ったとき，最小の辞書誤り率は 4.0%であった．

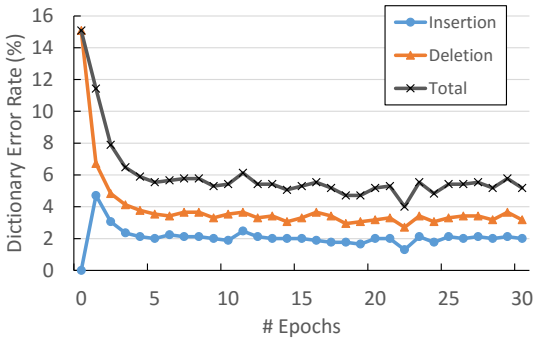

図 3 3-gram 言語モデルを用いた場合のエポック数と辞書誤り率の関係.

6. おわりに

対応の無い音素書き起こしと単語テキストを元に発音辞書を教師なし学習する新しい手法の提案を行った．言語モデルに 2-gram を用いパープレキシティーが 30.8 の条件で実験を行い，ブロックギブスサンプリングを用いた学習が進むにつれて発音未知の単語に正しい発音が割り当てられ，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していくことを実証した．また，言語モデルに 3-gram を用いた場合，辞書誤り率が初期状態の 15%から 4.0%まで減少した．今後の課題としては，サンプリングの効率を上げてより大規模なデータを用いた実験を行うことや，基底分布として G2P による予測を組み合わせて活用すること，認識誤りを含んだ音素列を入力に用いることなどが挙げられる．

謝辞 Part of the work reported here was carried out during the 2016 Jelinek Memorial Summer Workshop on Speech and Language Technologies. Takahiro Shinozaki, Daichi Mochihashi, and Graham Neubig were supported by Johns Hopkins University via DARPA LORELEI Contract No HR0011-15-2-0027, and gifts from Microsoft, Amazon, Google, and Facebook. Takahiro Shinozaki was also supported by JSPS KAKENHI Grant Number 26280055. Shinji Watanabe was supported by JSPS KAKENHI Grant Number 26280055 and MERL.

参考文献

[1] Bisani, M. and Ney, H.: Joint-sequence models for grapheme-to-phoneme conversion, *Speech Communication*, Vol. 50, No. 5, pp. 434–451 (2008).
[2] Chen, S. F. et al.: Conditional and joint models for grapheme-to-phoneme conversion., *INTERSPEECH* (2003).
[3] Taylor, P.: Hidden Markov models for grapheme to phoneme conversion., *INTERSPEECH*, pp. 1973–1976 (2005).
[4] 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).
[5] Lu, L., Ghoshal, A. and Renals, S.: Acoustic data-driven pronunciation lexicon for large vocabulary speech recognition, *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*, IEEE, pp. 374–379 (2013).
[6] Hazen, T. J. and Bazzi, I.: A comparison and combination of methods for OOV word detection and word confidence scoring, *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, Vol. 1, IEEE, pp. 397–400 (2001).
[7] Parada, C., Dredze, M., Filimonov, D. and Jelinek, F.: Contextual information improves OOV detection in speech, *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Association for Computational Linguistics, pp. 216–224 (2010).
[8] Watanabe, S. and Chien, J.-T.: *Bayesian Speech and Language Processing*, Cambridge University Press (2015).
[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] Teh, Y. W.: A hierarchical Bayesian language model based on Pitman-Yor processes, *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, pp. 985–992 (2006).
[12] Murphy, K. P.: Hidden semi-Markov models (hsmms), *unpublished notes*, Vol. 2 (2002).
[13] Mochihashi, D., Yamada, T. and Ueda, N.: Bayesian unsupervised word segmentation with nested Pitman-Yor language modeling, *Proc. ACL-IJCNLP*, pp. 100–108 (2009).
[14] 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).
[15] 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).
[16] Heymann, J., Walter, O., Haeb-Umbach, R. and Raj, B.: Iterative Bayesian word segmentation for unsupervised vocabulary discovery from phoneme lattices, *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 4057–4061 (2014).
[17] Heymann, J., Walter, O., Haeb-Umbach, R. and Raj, B.: Unsupervised word segmentation from noisy input, *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*, IEEE, pp. 458–463 (2013).
[18] Neubig, G., Mimura, M., Mori, S. and Kawahara, T.: Learning a language model from continuous speech., *INTERSPEECH*, Citeseer, pp. 1053–1056 (2010).