MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

Semi-Supervised Learning of a Pronunciation Dictionary from Disjoint Phonemic Transcripts and Text

Shinozaki, T.; Watanabe, S.; Mochihashi, D.; Neubig, G.

TR2017-133 August 2017

Abstract

While the performance of automatic speech recognition systems has recently approached human levels in some tasks, the application is still limited to specific domains. This is because system development relies on extensive supervised training and expert tuning in the target domain. To solve this problem, systems must become more self-sufficient, having the ability to learn directly from speech and adapt to new tasks. One open question in this area is how to learn a pronunciation dictionary containing the appropriate vocabulary. Humans can recognize words, even ones they have never heard before, by reading text and understanding the context in which a word is used. However, this ability is missing in current speech recognition systems. In this work, we propose a new framework that automatically expands an initial pronunciation dictionary using independently sampled acoustic and textual data. While the task is very challenging and in its initial stage, we demonstrate that a model based on Bayesian learning of Dirichlet processes can acquire word pronunciations from phone transcripts and text of the WSJ data set.

Interspeech

Semi-Supervised Learning of a Pronunciation Dictionary from Disjoint Phonemic Transcripts and Text

Takahiro Shinozaki[1]*, Shinji Watanabe*[2]*, Daichi Mochihashi*[3]*, Graham Neubig*[4]

[1]Tokyo Institute of Technology, Japan
[2]Mitsubishi Electric Research Laboratories, USA
[3]The Institute of Statistical Mathematics, Japan
[2]Carnegie Mellon University, USA

shinot@ict.e.titech.ac.jp, watanabe@merl.com, daichi@ism.ac.jp, gneubig@cs.cmu.edu

Abstract

While the performance of automatic speech recognition systems has recently approached human levels in some tasks, the application is still limited to specific domains. This is because system development relies on extensive supervised training and expert tuning in the target domain. To solve this problem, systems must become more self-sufficient, having the ability to learn directly from speech and adapt to new tasks. One open question in this area is how to learn a pronunciation dictionary containing the appropriate vocabulary. Humans can recognize words, even ones they have never heard before, by reading text and understanding the context in which a word is used. However, this ability is missing in current speech recognition systems. In this work, we propose a new framework that automatically expands an initial pronunciation dictionary using independently sampled acoustic and textual data. While the task is very challenging and in its initial stage, we demonstrate that a model based on Bayesian learning of Dirichlet processes can acquire word pronunciations from phone transcripts and text of the WSJ data set.

1. Introduction

Even in highly tuned automatic speech recognition (ASR) systems that achieve human-level accuracy, their performance is heavily dependent on supervised learning. To support a new task domain or new words, labeled speech data and pronunciations of new words must be prepared. This often limits the usability of the system to the initially prepared domain. In contrast, humans can constantly learn from both speech and text data, even if they are not paired, recognizing new words in speech with unknown pronunciations by understanding the context in which the word is used. If the same ability to expand this *pronunciation dictionary* could be achieved in speech recognition systems without requiring labeled data, it would contribute not only to reducing maintenance cost but also allow for more natural communication between machines and humans when applied to interactive systems such as home robots.

In general, a word pronunciation dictionary plays two roles in a speech recognition system. One is to define word units, and the other is to define a mapping from pronunciation to spelling. Some recent end-to-end ASR frameworks avoid the necessity for both of these functions by performing character recognition [1]. However, many applications including information retrieval and spoken dialog systems still require word units, and we therefore focus on approaches that define these units. We propose a Bayesian semi-supervised framework for learning pronunciation dictionaries that can learn word pronunciations from disjoint phonemic transcripts and text of a language. The assumption is that a partial pronunciation dictionary is available; pronunciations are given for some words but missing for others. This makes it possible to learn additional dictionary entries that are compatible with the manually prepared word units, phone set, and spelling.

2. Related work

Methods to automatically obtain word pronunciations are classified into two groups according to whether a mapping from pronunciation to spelling is provided or not: without-spell-mapping methods and with-spell-mapping methods.

In the without-spell-mapping methods, usually a word is directly represented by a phone sequence. Representative methods in this paradigm are based on out of vocabulary (OOV) detection [2, 3] and phone recognition, where speech input is first decoded by a phone recognizer and a speech segment detected as an OOV is labeled by the decoded phone sequence. By combining this with a word decoder, it is expected that a word is output if it is included in the vocabulary of the decoder and otherwise a phone sequence is output [4]. To improve the performance, several extensions have been proposed such as including frequent phone sequences in the phone recognizer as word fragments [5]. Another approach is based on segmenting a phone sequences in a completely unsupervised manner by making a hierarchical Bayesian model [6, 7, 8, 9], which originates from the unsupervised word segmentation in text [10, 11]. The training is performed only using a phone sequence or a phone lattice, without using text, making it possible to learn spoken languages without a writing system.

While representing a new word by its pronunciation is useful for some applications, obtaining a corresponding spelling is often important. To address the problem, Parada et al. proposed a method that assigns a spelling to an OOV word based on heuristics on context information and web search [12]. Another approach is *grapheme to phoneme (G2P) conversion* [13, 14, 15, 16], where a G2P converter is applied to a new word to estimate its pronunciation. While G2P is mathematically well-formulated and is easy to use, a limitation is that it is not accurate for words for which the pronunciation is hard to infer from the spelling. This is the case for English acronyms, or for ideograms such as Chinese or Japanese characters, which have a weak correlation between characters and pronunciations. To handle the mapping at a word level in a statistical framework, learning methods using pronunciation mixture models have been proposed [17, 18, 19]. These methods model a word pronunciation by a finite categorical distribution of possible pronunciations. The parameters of the distributions

Table 1: *Description of the nodes in the Bayesian model shown in Figure 1.*

Node	Description
Pronunciation dictionary (Δ)	Probabilistic pronunciation dictionary
Language model (Θ)	Hierarchical Bayesian language model
Word sequence ($\mathbf{w}$)	Word sequence of an utterance
Segmented phone sequence ($\boldsymbol{\psi}$)	Phone sequence of an utterance with word boundaries
Phone sequence ($\boldsymbol{\phi}$)	Phone sequence of an utterance without a word boundary

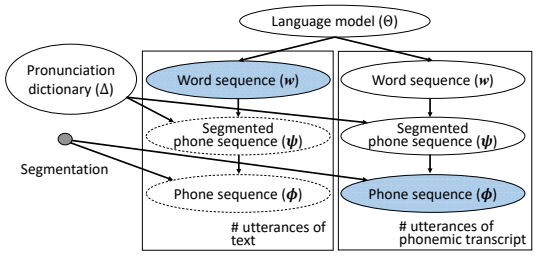

Figure 1: *Proposed Bayesian model to train a pronunciation dictionary from disjoint phonemic transcripts and text.*

are estimated by the EM algorithm [20], or a posterior distribution is inferred using Bayesian methods. However, a disadvantage is that these are supervised methods that require parallel speech and text data, which is hard to come by.

3. Proposed method

To perform semi-supervised learning of word pronunciations from unaligned phone and word utterances utilizing a partial pronunciation dictionary, phonetic and linguistic knowledge must be combined. We do so by creating a single Bayesian model integrating a pronunciation dictionary and a language model. Compared to the hierarchical Bayesian model of the unsupervised without-spell-mapping approach, a pronunciation dictionary is introduced to support the mapping from pronunciation to spelling. Likewise, compared to the supervised methods using the pronunciation mixture model, a language model is integrated to use distributional information of words.

3.1. Integrated Bayesian model

Figure 1 shows the structure of the proposed Bayesian model for semi-supervised pronunciation dictionary learning. Within the Bayesian model, each node has an internal structure. The definitions of the nodes are summarized in Table 1.

The three nodes "Word sequence ($\mathbf{w}$)", "Segmented phone sequence ($\boldsymbol{\psi}$)", and "Phone sequence ($\boldsymbol{\phi}$)" represent an utterance given as a word sequence, a word segmented phone sequence, and a phone sequence with no word segmentation, respectively. For example, "the sale of the hotels" is a word sequence, "DH AH </w> S EY L </w> AH V </w> DH AH </w> HH OW T EH L Z </w>" is a word segmented phone sequence where </w> represents a word boundary, and "DH AH S EY L AH V DH AH HH OW T EH L Z" is a phone sequence.

The node "Language model (Θ)" is a hierarchical Bayesian language model [21], and the node "Pronunciation dictionary (Δ)" is a pronunciation dictionary. A word sequence is generated from the language model, and a word segmented phone sequence is generated from the word sequence and the pronunciation dictionary. The node represented by a filled small circle represents a simple deterministic rule to convert the word segmented phone sequence to the phone sequence. Each utterance is assumed to be independent given the language model and the pronunciation dictionary. Unlike the unsupervised word and LM learning [6, 7, 8, 9], the hierarchical language model is based on words represented by a character sequence rather than a phone sequence, which allows for the use of text data for the training.

3.2. Probabilistic model of the pronunciation dictionary

A pronunciation dictionary is a set of pairs of a word w (e.g. "hello") and its pronunciation ρ (e.g. "HH AH L OW") for words in a vocabulary. Sometimes, a pronunciation dictionary is designed so that a word can have multiple pronunciations with optional probability weights that represent their relative frequency [22]. From probabilistic modeling point of view, this is equivalent to considering a finite mixture model of pronunciations for each word where a mixture component is the pronunciation and a mixture weight is its probability. In [17], the mixture weights are trainable parameters, and are estimated by the EM algorithm as a categorical distribution. For the Bayesian approach, a Dirichlet distribution is used in [19] to give a prior probability for the mixture weights.

To potentially allow any pronunciation for a word, we extend the finite pronunciation mixture model to an infinite mixture model. To give a prior probability to the infinite distribution $p_w(\rho)$ of the word pronunciations, we use the Dirichlet process [23, 24]

$$\begin{aligned} G_w &\sim \mathrm{DP}\left(\alpha, G_0\right) \quad (i.i.d.\ w \in V)\,, \\ p_w(\rho) &= G_w, \end{aligned} \tag{1}$$

where G_0 is a base distribution, and $\alpha\,(>0)$ is a concentration parameter. A draw from the base distribution G_0 is a pronunciation ρ. The pronunciation dictionary Δ is defined as a set of word pronunciation models G_w as shown in Equation (2), and the prior probability of Δ is the joint probability of the infinite mixture models G_w.

$$\Delta = \{G_w\}_{\{w \in V\}}\,. \tag{2}$$

The predictive distribution of the pronunciations based on the pronunciation dictionary is obtained by applying the Chinese restaurant process (CRP) [25] at each word. Let us assume that a set of utterances U are observed in which a word w has appeared $c(w)$ times. Let's also assume that a pronunciation ρ of the word w has appeared $c_w(\rho)$ times where $\sum_\rho c_w(\rho) = c(w)$. Then, the predictive distribution of the pronunciation ρ for the word w is given by Equation (3).

$$p(\rho|w, U) = \frac{c_w(\rho)}{\alpha + c(w)} + \frac{\alpha}{\alpha + c(w)}\, G_0(\rho). \tag{3}$$

The fact that a word usually has only a few (often only one) pronunciations is represented by choosing α close to 0, assigning a large prior probability mass $\frac{1}{1+\alpha}$ to the first pronunciation.

3.3. Gibbs sampling for learning and evaluation

For inference on the Bayesian model, we use Gibbs sampling [26, 27] starting with an initial assignment, then repeatedly randomly picking an utterance and updating the values of its hidden variables by drawing a sample from their joint posteriors given values of the rest of the variables.

Let $\boldsymbol{\phi}_n, \boldsymbol{\psi}_n, \mathbf{w}_n$ be the phone sequence, word segmented phone sequence, and word sequence of the selected utterance respectively. Similarly, let $\Phi^{\backslash n}, \Psi^{\backslash n}, \mathbf{W}^{\backslash n}$ be the sets of phone sequences, word segmented phone sequences, and word sequences of the remaining utterances. As shown in the left plate of Figure 1, given a word-segmented text the word sequence $\mathbf{w}_n$ is an observed variable and the word segmented phone sequence $\boldsymbol{\psi}_n$ and the phone sequence $\boldsymbol{\phi}_n$ are hidden. Similarly, when an utterance is given as a phone level transcript, the phone sequence $\boldsymbol{\phi}_n$ is observed and the other two variables $\boldsymbol{\psi}_n$ and $\mathbf{w}_n$ are hidden.

The joint posterior of any combination of the hidden nodes of a selected utterance is obtained from a joint posterior of the three variables:

$$\begin{aligned} & p(\boldsymbol{\phi}_n, \boldsymbol{\psi}_n, \mathbf{w}_n | \Phi^{\backslash n}, \Psi^{\backslash n}, \mathbf{W}^{\backslash n}) \\ &= \int p(\boldsymbol{\phi}_n, \boldsymbol{\psi}_n, \mathbf{w}_n, \Theta, \Delta | \Phi^{\backslash n}, \Psi^{\backslash n}, \mathbf{W}^{\backslash n}) d\Theta d\Delta \\ &= p(\boldsymbol{\phi}_n | \boldsymbol{\psi}_n) p(\mathbf{w}_n | \mathbf{W}^{\backslash n}) p(\boldsymbol{\psi}_n | \mathbf{w}_n, \mathbf{W}^{\backslash n}, \Psi^{\backslash n}). \quad (4) \end{aligned}$$

The derivation of Equation (4) is based on the chain rule, conditional independencies that are read from the Bayesian model by d-separation [28], and marginalization of the language model Δ and the pronunciation dictionary Θ.

In the equation, predictive distributions $p(\mathbf{w}_n | \mathbf{W}^{\backslash n})$ and $p\left(\boldsymbol{\psi}_n | \mathbf{w}_n, \mathbf{W}^{\backslash n}, \Psi^{\backslash n}\right)$ are obtained by CRP, and are used through collapsed Gibbs sampling [27]. The predictive distribution $p(\mathbf{w}_n | \mathbf{W}^{\backslash n})$ is easily evaluated by the CRP because $\mathbf{W}^{\backslash n}$ is in the conditional, which means it is treated as if it is observed, and it works as a language model for the selected utterance.[1] Similarly, the predictive distribution $p\left(\boldsymbol{\psi}_n | \mathbf{w}_n, \mathbf{W}^{\backslash n}, \Psi^{\backslash n}\right)$ is easily evaluated by CRP because both $\mathbf{W}^{\backslash n}$ and $\Psi^{\backslash n}$ are in the conditional part, which means they are treated as if their alignments were known. This works as a pronunciation dictionary for the selected utterance. $p(\boldsymbol{\phi}_n | \boldsymbol{\psi}_n)$ corresponds to the segmentation rule, which takes a value of 1 only when $\boldsymbol{\phi}_n$ is obtained by removing the word boundaries in $\boldsymbol{\psi}_n$, and otherwise it is 0. In a word-segmented text, the segmented phone sequence and the phone sequence may be marginalized out instead of sampling their values because of the Bayesian model's structure.

3.4. WFST-based implementation

Sampling from the joint posterior distribution of the hidden variables of the selected utterance is not a simple task due to its complex internal structure. To implement Gibbs sampling, we make use of WFSTs, extending the implementation of the unsupervised word and LM learning [6, 7, 8] to introduce the pronunciation dictionary. In order to perform sampling from a joint probability of $p(\boldsymbol{\psi}_n, \mathbf{w}_n | \boldsymbol{\phi}_n, \Phi^{\backslash n}, \Psi^{\backslash n}, \mathbf{W}^{\backslash n})$ given an input phone sequence $\boldsymbol{\phi}_n$ of a selected utterance, first the phone sequence $\boldsymbol{\phi}_n$ and each component of Equation (4) are represented by WFSTs and they are composed to form a single WFST. The composed WFST expresses an unnormalized distribution of the posterior. Then a sample is obtained by applying the forward filtering backward sampling algorithm, which uses dynamic programming to effectively sample from the WFST.

There is a problem, however, when composing a WFST for our proposed framework. Specifically, the intermediate symbols are removed if we use normal composition operations. This means the necessary information about the segmented phone sequence $\boldsymbol{\psi}_n$ is marginalized out. To address the problem, we modify the composition operation so that intermediate symbols are accumulated in the input label. When an arc having "a" and "b" as the input and output labels and an arc having "b" and "c" are composed by the modified composition, the composed arc has "a_b" as the input label and "c" as the output label instead of "a" and "c".

4. Experimental setup

Experiments were performed using the WSJ corpus [31, 32] and the CMU dictionary. As the phone transcript, true phone labels were used. The number of phones was 39 and lexical stress was not used. The word entries of the pronunciation dictionary were made from a word level transcript, in which pronunciations were initially given to a subset of the words. The task was to find pronunciations for the remaining words using unaligned word and phone level transcripts. As the base distribution for the pronunciation, a phone 0-gram model was used. The concentration parameter α for the pronunciation dictionary was set to 0.1. Gibbs sampling was initialized by performing the Viterbi assignment to the hidden variables at the first epoch. For the initialization, all the utterances were processed before updating the statistics. After the first epoch, the distributions were updated utterance by utterance. During the sampling, the vocabulary was fixed so that no new word was generated with unknown spelling. The language model was first initialized using the word-segmented text, where the segmented phone sequences and the phone sequences were marginalized out. The software was implemented by modifying LatticeWordSegmentation [8, 33, 6],[2] which implements a hierarchical Pitman-Yor language model. As the baseline, G2P was evaluated where Sequitur G2P [13][3] was used for the implementation.

5. Results

To investigate the basic properties of the proposed method, we first run experiments using a small data set, and then scale it up to a larger setting. The first experiments were performed using 100 utterances with phone level transcript and 100 utterances with word level text. In this experiment, the word and phone level transcripts were obtained from the same 100 utterances in the corpus. However, no utterance level alignment information was given to the system. The vocabulary size was 849 in which pronunciations were given to 70%, which was 594 words, with the remaining 30% not given pronunciations. The language model was a word 2-gram, and the perplexity was 30.8. When a third order G2P model was trained using the 594 words and applied to the remaining 255 words to compensate for the missing pronunciations, 60.4% of the words were assigned a wrong pronunciation. Looking at the whole dictionary, 18.1% of the words had a wrong pronunciation.

[1] Particularly, it is an extended version of the Kneser-Ney N-gram [29] when a hierarchical Pitman-Yor language model is used [30].

[2] https://github.com/fgnt/LatticeWordSegmentation

[3] https://www-i6.informatik.rwth-aachen.de/web/Software/g2p.html

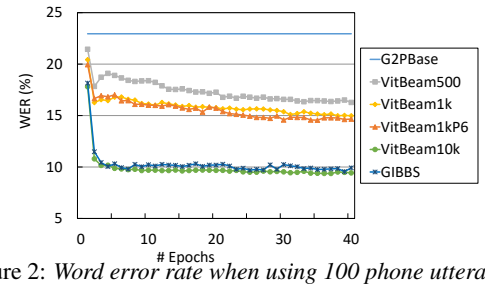

Figure 2: *Word error rate when using 100 phone utterances.*

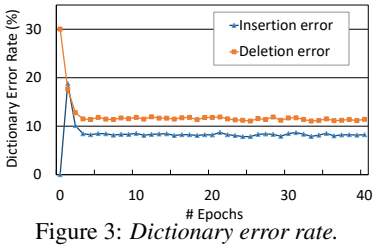

Figure 3: *Dictionary error rate.*

Figure 2 shows WER results evaluated for output word labels of WFST paths obtained by the sampling for the phone input data. The plotted WER is an average of three independent runs. In the figure, "G2PBase" is the baseline result when a Viterbi assignment was performed using the G2P-augmented dictionary pinning the pronunciations by choosing α almost 0 (1E-9) and prohibiting model update. "GIBBS" indicates the result of the proposed method using the initial dictionary with the missing pronunciations. "VitBeam$N$" is a result when Viterbi beam approximation with beam width N was used in the Gibbs sampling. "VitBeam1kP6" is similar to VitBeam1k but six utterances were processed in parallel before updating the statistics. By proceeding the sampling epochs, smaller WERs were obtained for all the conditions compared to 23.0% of the baseline WER showing the effect of the proposed method. VitBeam1kP6 gave similar WER as VitBeam1k, which indicates parallel processing for faster computation does not harm the WER. The minimum WERs by GIBBS, VitBeam500, VitBeam1k, VitBeam1kP6, and VitBeam10k after 40 epochs were 9.6%, 16.3%, 15.0%, 14.6% and 9.4%, respectively. When a Xeon X5650 CPU was used, their processing times were 37 min, 2.1 min, 4 min, 1.3 min, and 99 min per epoch. This result shows the Viterbi beam approximation is useful in reducing the computational time, while the improvement in WER is reduced if the beam width is narrow.

To analyze the learned dictionary obtained by GIBBS, Figure 3 depicts the number of insertion and deletion dictionary errors normalized by the vocabulary size. Here, an insertion error means an extra wrong pronunciation appeared in one of the sampled pronunciations for the word, and a deletion means a correct pronunciation is missing. For the G2P-augmented dictionary, these errors are both 18.1% since one wrong pronunciation results in one insertion and one deletion error. In the figure, the zeroth epoch is the initial condition and only deletion errors existed. At the first epoch, insertion errors increased since a new pronunciation was generated for the words having no pronunciation. After that, both insertion and deletion errors mostly monotonically decreased.

Table 2 shows an example of a part of sampled word sequences obtained for a phone input when GIBBS was used. It can be seen that the correct sentence was obtained at the third epoch, which was the result of successful pronunciation assignment.

We next perform a larger experiment using all the training data in the WSJ corpus after removing duplicated utterances. The first 8000 utterances were used as the word transcript and the remaining 2796 utterances were used as the phone transcript. There was no overlap between them. The vocabulary size was 12.5k in which 15% and 30% words were not given pronunciations initially. A word 3-gram was used as the language model. The perplexity was 200.8 and the OOV rate was 2.4%. When G2P was used to compensate for the 15% and the 30% of the missing pronunciations, the error rates were 38.7% and 39.7%, respectively. For fast computation and suppressed memory usage, VitBeam1kP6 was used (denoted as VitBeam). We additionally tested an extension of the proposed method where the G2P-augmented dictionary was used as an initial dictionary (VitBeam+G2P). Table 3 shows the results. The WERs of the G2P baseline were 16.7% and 21.1% for the 15% and 30% initial conditions, respectively. VitBeam gave an improvement as epochs progressed in this condition as well, but the result obtained at fifth epoch was worse than the G2P baseline when the 30% of pronunciations were initially missing. However, by combining the G2P with the proposed method, VitBeam+G2P successfully gave a significantly smaller WER of 8.9% and 13.5%, respectively.

Table 2: *Example of a part of sampled sentences. Pronunciations of "but", and "times" were initially unknown.*

Reference	you're a friend in bad times as well as good
Epoch1	you're a friend in , as well as good
Epoch2	you're a friend in bad time close as well as good
Epoch3	you're a friend in bad times as well as good

Table 3: *WERs when non-overlapping larger data set was used.*

Init condition	15% missing			30% missing		
Epoch	1	2	5	1	2	5
G2PBase	16.7	-	-	21.1	-	-
VitBeam	20.3	15.0	14.5	33.8	25.1	24.1
VitBeam+G2P	11.1	8.8	8.9	18.4	13.8	13.5

6. Conclusion and future work

We have proposed a Bayesian semi-supervised pronunciation dictionary learning method using a disjoint phone and word data. Experiments using WSJ corpus have demonstrated its effectiveness in obtaining reductions in WER. Future work includes improving the base distribution for improved performance. In fact, purely context information was utilized in this paper based on the phone 0-gram base distribution to find or correct pronunciations of words, but utilizing a more informative base distribution could potentially be useful. Investigating other sampling strategies such as beam sampling [34] would improve the computational efficiency. Another important extension is to use automatically recognized phone transcripts. For this, a WFST encoding a phone lattice could be used instead of the one best hypothesis, or a layer of HMMs could be appended to the framework to form a full Bayesian model of a speech recognition system.

7. Acknowledgements

Part of this work 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 MERL.

8. References

[1] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *International Conference on Machine Learning (ICML)*, 2014, pp. 1764–1772.

[2] T. J. Hazen and I. Bazzi, "A comparison and combination of methods for OOV word detection and word confidence scoring," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2001, pp. 397–400.

[3] C. Parada, M. Dredze, D. Filimonov, and F. Jelinek, "Contextual information improves OOV detection in speech," in *North American Chapter of the Association for Computational Linguistics (NAACL)*, 2010, pp. 216–224.

[4] I. Bazzi, "Modelling out-of-vocabulary words for robust speech recognition," Ph.D. dissertation, Massachusetts Institute of Technology, 2002.

[5] A. Rastrow, A. Sethy, and B. Ramabhadran, "A new method for OOV detection using hybrid word/fragment system," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2009, pp. 3953–3956.

[6] G. Neubig, M. Mimura, S. Mori, and T. Kawahara, "Learning a language model from continuous speech." in *INTERSPEECH*, 2010, pp. 1053–1056.

[7] G. Neubig, "Unsupervised learning of lexical information for language processing systems," Ph.D. dissertation, Kyoto University, 2012.

[8] J. Heymann, O. Walter, R. Haeb-Umbach, and B. Raj, "Iterative Bayesian word segmentation for unsupervised vocabulary discovery from phoneme lattices," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 4057–4061.

[9] H. Kamper, A. Jansen, and S. Goldwater, "Unsupervised word segmentation and lexicon discovery using acoustic word embeddings," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 24, no. 4, pp. 669–679, 2016.

[10] S. Goldwater and T. Griffiths, "A fully Bayesian approach to unsupervised part-of-speech tagging," in *Association for Computational Linguistics (ACL)*, 2007, pp. 744–751.

[11] D. Mochihashi, T. Yamada, and N. Ueda, "Bayesian unsupervised word segmentation with nested Pitman-Yor language modeling," in *Proc. ACL-IJCNLP*, 2009, pp. 100–108.

[12] C. Parada, A. Sethy, M. Dredze, and F. Jelinek, "A spoken term detection framework for recovering out-of-vocabulary words using the web," in *INTERSPEECH*, 2010.

[13] M. Bisani and H. Ney, "Joint-sequence models for grapheme-to-phoneme conversion," *Speech Communication*, vol. 50, no. 5, pp. 434–451, 2008.

[14] S. F. Chen, "Conditional and Joint Models for Grapheme-to-Phoneme Conversion," in *INTERSPEECH*, 2003.

[15] P. Taylor, "Hidden Markov models for grapheme to phoneme conversion." in *INTERSPEECH*, 2005, pp. 1973–1976.

[16] J. R. Novak, N. Minematsu, and K. Hirose, "WFST-based grapheme-to-phoneme conversion: open source tools for alignment, model-building and decoding," in *10th International Workshop on Finite State Methods and Natural Language Processing*, 2012, p. 45.

[17] 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.

[18] L. Lu, A. Ghoshal, and S. Renals, "Acoustic data-driven pronunciation lexicon for large vocabulary speech recognition," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2013, pp. 374–379.

[19] C.-y. 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.

[20] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of Royal Statistical Society B*, vol. 39, pp. 1–38, 1976.

[21] Y. W. Teh, "A hierarchical Bayesian language model based on Pitman-Yor processes," in *ACL/COLING*, 2006, pp. 985–992.

[22] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey *et al.*, "The HTK book (for HTK version 3.4)," *Cambridge University Engineering Department*, 2006.

[23] T. S. Ferguson, "A Bayesian analysis of some nonparametric problems," *The Annals of Statistics*, vol. 1, no. 2, pp. 209–230, 1973.

[24] Y. W. Teh, "Dirichlet processes," in *Encyclopedia of Machine Learning*. Springer, 2010.

[25] J. Pitman, "Exchangeable and partially exchangeable random partitions," *Probability theory and related fields*, vol. 102, no. 2, pp. 145–158, 1995.

[26] A. E. Gelfand and A. F. Smith, "Sampling-based approaches to calculating marginal densities," *Journal of the American statistical association*, vol. 85, no. 410, pp. 398–409, 1990.

[27] J. S. Liu, "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.

[28] D. Geiger, T. Verma, and J. Pearl, "Identifying independence in Bayesian networks," *Networks*, vol. 20, no. 5, pp. 507–534, 1990.

[29] R. Kneser and H. Ney, "Improved backing-off for m-gram language modeling," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, 1995, pp. 181–184.

[30] Y. W. Teh, "A Bayesian interpretation of interpolated Kneser-Ney," Technical Report TRA2/06, School of Computing, NUS, Tech. Rep., 2006.

[31] J. Garofolo, D. Graff, D. Paul, and D. Pallett, "CSR-I (WSJ0) Complete LDC93S6A. DVD. Philadelphia: Linguistic Data Consortium," 1993.

[32] "CSR-II (WSJ1) Sennheiser LDC94S13B. DVD. Philadelphia: Linguistic Data Consortium," 1994.

[33] J. Heymann, O. Walter, R. Haeb-Umbach, and B. Raj, "Unsupervised word segmentation from noisy input," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2013, pp. 458–463.

[34] J. Van Gael, Y. Saatci, Y. W. Teh, and Z. Ghahramani, "Beam sampling for the infinite hidden Markov model," in *Proceedings of the 25th International Conference on Machine Learning*, ser. ICML '08. ACM, 2008, pp. 1088–1095.