Seshadri, Shreyas; Remes, Ulpu; Räsänen, Okko

Comparison of Non-parametric Bayesian Mixture Models for Syllable Clustering and Zero-Resource Speech Processing

Published in:
Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH

DOI:
10.21437/Interspeech.2017-339

Published: 01/08/2017

Document Version
Publisher's PDF, also known as Version of record

INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Comparison of Non-parametric Bayesian Mixture Models for Syllable Clustering and Zero-Resource Speech Processing

Shreyas Seshadri[1], *Ulpu Remes*[2], *Okko Räsänen*[1]

[1]Department of Signal Processing and Acoustics, Aalto University, Finland.
[2]Department of Mathematics and Statistics, University of Helsinki, Finland.
shreyas.sesahdri@aalto.fi, ulpu.remes@helsinki.fi, okko.rasanen@aalto.fi

Abstract

Zero-resource speech processing (ZS) systems aim to learn structural representations of speech without access to labeled data. A starting point for these systems is the extraction of syllable tokens utilizing the rhythmic structure of a speech signal. Several recent ZS systems have therefore focused on clustering such syllable tokens into linguistically meaningful units. These systems have so far used heuristically set number of clusters, which can, however, be highly dataset dependent and cannot be optimized in actual unsupervised settings. This paper focuses on improving the flexibility of ZS systems using Bayesian non-parametric (BNP) mixture models that are capable of simultaneously learning the cluster models as well as their number based on the properties of the dataset. We also compare different model design choices, namely priors over the weights and the cluster component models, as the impact of these choices is rarely reported in the previous studies. Experiments are conducted using conversational speech from several languages. The models are first evaluated in a separate syllable clustering task and then as a part of a full ZS system in order to examine the potential of BNP methods and illuminate the relative importance of different model design choices.

Index Terms: Non-parametric clustering, zero-resource processing, variational inference, Pitman-Yor process, von Mises-Fisher mixtures

1. Introduction

A recently emerged area of speech technology research is the so-called zero-resource speech processing (ZS) initiative where the aim is to create systems capable of learning structural representations of speech input in the absence of any data labeling [1–3], providing both scalability towards under-resourced domains and illuminating how human infants may learn spoken languages. A number of the existing ZS systems, including the best performing system at the word-level [1] in the Interspeech-2015 Zerospeech challenge and the state-of-the-art system in [2] are based on clustering and temporal grouping of syllable-like rhythmic units. The system in [1] first segments speech into syllable-like chunks, clusters the resulting tokens into categories using K-means, and decodes words as recurring n-grams over the syllabic clusters in the data. The work in [2] extends this method by creating a Bayesian segmental model that jointly optimizes word category identities (clustering, using a Bayesian GMM) and boundaries chosen from the syllable-like chunks (segmentation pruning). However, both systems used a heuristically set number of clusters for the data.

Since the overall goal in ZS is to work towards systems that can autonomously learn speech representations that are supported by the (statistical) properties of the available data, it would be highly beneficial if the number of clusters could be also inferred automatically from the input. In this context Bayesian non-parametric (BNP) models are potentially very powerful as they solve the model selection problem as an inherent part of their behaviour [3]. Dirichlet process mixture models (DPMMs) [4] are the most commonly used BNP models in various clustering problems in speech research. For instance, Lee and Glass [5] describe a DPMM for acoustic modelling where each mixture model is a hidden Markov model (HMM). Rosenberg [6] uses Gaussian DPMMs, called DPGMMs, to model prosodic sequences in speech. Kamper et al. [7] use DPGMMs to the task of lexical clustering whereas Chen et al. [8] use DPGMMs to learn acoustic models from speech. Nonparametric extensions to HMM frameworks have also been used in speaker diarisation [9, 10].

Despite their popularity, there are several open questions related to the use of BNP methods to ZS tasks. For instance, word counts in natural languages tend to follow a Zipfian distribution and hence syllable counts are also expected to be similarly distributed, especially for languages with a large proportion of monosyllabic words such as English [11]. Hence power-law producing priors such as the Pitman–Yor process (PYP) [12, 13] may perform better than Dirichlet process (DP) priors in syllable-based systems. In addition, the fixed-dimensional spectral representations such as those used in [1, 2] could also be modelled using some other parametric distribution than GMM. One candidate is the cosine distance -based von Mises–Fisher mixture model (here: VMM) [14] that is more suited for high-dimensional density estimation than GMMs [15] as long as the feature vectors can be unit normalised before clustering (see, e.g., [2]).

Given this background, the present paper focusses on investigating the feasibility of BNP methods in ZS settings, and especially on the impact of the prior and component model choice as such comparisons have not been reported in previous works. We first compare different BNP models with heuristic k-means and Bayesian GMM in a simplified syllable clustering task on conversational data from two different languages but using true syllable boundaries from the manual annotation. Then we compare the same methods on the 2015 ZS Challenge data [16] using the full system described in [1]. As a result, we can see 1) how the BNP methods compare against k-means with heuristically optimized number of clusters (”informed selection”), and 2) whether prior and component model choices have any practical impact in idealized and real ZS settings.

2. Bayesian Mixture Models

A Bayesian mixture model (BMM; Figure 1), G with K components, weights $\{\pi_k\}_{k=1}^{K}$ and mixture component model parameters $\{\boldsymbol{\theta}_k\}_{i=1}^{K}$ can be defined as

http://dx.doi.org/10.21437/Interspeech.2017-339

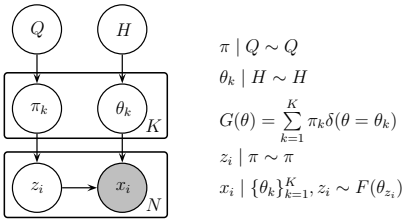

Figure 1: *Bayesian graph of a BMM*

$$G(\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \delta(\boldsymbol{\theta} = \boldsymbol{\theta}_k), \quad (1)$$

Here the weights and the mixture model component parameters are sampled from the prior distributions Q and H respectively. We define latent variables $\{z_i\}_{i=1}^{N}$, sampled from a multinomial distribution parameterized by $\{\pi_k\}_{k=1}^{K}$, as the clusters to which the N observations $\{\mathbf{x}_i\}_{i=1}^{N}$ are assigned. Finally, the observed variables $\mathbf{x}_i$ are then sampled from the model $F(\boldsymbol{\theta}_{z_i})$. We analyse the following models and priors on the weight distributions

2.1. Mixture Component Models

• ***Bayesian Gaussian Mixture Models*** **-** We consider the fixed spherical precision Gaussians for the mixture models as proposed in [2], i.e. $\boldsymbol{\theta}_k = \boldsymbol{\mu}_k$ representing the mean. The model is then $F(\boldsymbol{\theta}_{z_i}) = \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{z_i}}, \sigma\mathbb{I})$, where σ is the fixed spherical covariance. The prior is chosen as conjugate distribution, $H = \mathcal{N}(\boldsymbol{\mu}_0, \sigma_0\mathbb{I})$.

• ***Bayesian Von Mises–Fisher Mixture Models*** **-** Experiments are also conducted with a von Mises–Fisher mixture model [14]. VMMs model observation vectors as points on unit hypersphere and find clusters based on cosine distance between observations. The model parameters $\boldsymbol{\theta}_k$ include mean direction $\boldsymbol{\mu}_k$ and concentration parameter λ_k. The parameters are associated with a von Mises-Fisher (VMF)–Gamma prior as proposed in [17]: the mean direction is expected to have a VMF distribution with mean direction $\boldsymbol{\mu}_0$ and concentration parameter $\beta_0\lambda_k$ while λ_k is expected to have a gamma distribution with parameters a_0 and b_0.

2.2. Weight Distributions

• ***Dirichlet Distribution*** **-** Dirichlet distributions (DD) are simple parametric distributions that are commonly used as a prior distribution for the weights, as $\pi_k \sim Dir(\boldsymbol{\alpha})$ (e.g., [2]). The hyper-parameter α is a K dimensional vector of positive reals. If K overestimates the number of clusters represented in the data, unnecessary clusters may be emptied out in posterior estimation. In the current study, K is heuristically set to the expected number of clusters in the data, as proposed in [2].

• ***Dirichlet Process*** **-** A Dirichlet process (DP) [18, 4] is a non-parametric extension of the Dirichlet distribution. DPs model an infinite number of clusters i.e. $K = \infty$, so that the number of clusters that contributed in the observed dataset can be inferred automatically from the data. A DP is uniquely defined by the base distribution H of model parameter values on Θ and a positive scalar concentration parameter α, as $G \sim DP(\alpha, H)$. The weights, $\{\pi_k\}_{k=1}^{\infty}$, decrease exponentially, and can be sampled from the stick breaking process [19]

$$\pi_k = v_k \prod_{i=1}^{k-1}(1 - v_i) \quad (2)$$

where v_k are stick proportions, distributed as $v_k \sim Beta(1, \alpha)$.

• ***Pitman-Yor Process*** **-** The Pitman-Yor process (PYP) [12, 13] is another non-parametric model that generalises DP with an additional parameter, $0 \geq d > 1$. It is written as $G \sim PYP(d, \alpha, H)$. The weights, $\{\pi_k\}_{k=1}^{\infty}$ follow the power law, making PYP suitable for Zipf-distributed data [13]. We use a similar stick breaking process as with DPs to sample PYPs (see Eq.(2)), but the stick proportions are now distributed as $v_k \sim Beta(1 - d, \alpha + kd)$. When $d = 0$, PYP is equivalent to DP, i.e. $PYP(0, \alpha, H) = DP(\alpha, H)$.

2.3. Variational Inference

It is not possible to obtain a direct analytic solution for the BMM parameters. This paper focuses on variational inference methods [20, 21] that approximate the analytically intractable posterior with a tractable distribution called variational distribution. This is done by first making a number of independence assumptions that simplify the posterior distribution. Kullback–Leibler (KL) divergence to the true posterior is then minimised to find the variational distribution. In practice, the final update equations are similar to the expectation–maximisation (EM) algorithm that iterates between finding the probabilities of z_i (called responsibilities) based on the current model and updating model parameters based on the current responsibilities. The variational mixture model proposed in [21] is used to handle the non-parametric weight priors (DP and PYP), where K is truncated at a truncation limit T to deal with the infinite number of clusters.

3. Experiments

The first experiment was a syllable clustering task where the syllable tokens were extracted from the manual annotation, as we wanted to compare the methods in idealized settings without additional uncertainty caused by automatic syllabification. In the second experiment, the comparison was extended to an actual ZS pipeline where automatic syllable segmentation and clustering (compared here) was followed by decoding of words as recurring syllable n-grams of varying orders (see [1] for details). All tests were conducted on conversational speech.

3.1. Data and pre-processing

Switchboard [22] and Phonetic Corpus of Estonian Speech ([23]; studio section of the corpus) were used for the first experiment as they have manual syllable annotations available. Experiments were conducted in speaker independent settings for both Switchboard and Estonian (called "SB-I" and "Est-I" respectively) and speaker dependent settings for Estonian. The former was done using 1000 utterances (approx. 10000 syllable tokens) randomly chosen from each corpora. The latter using all data from each speaker in the Estonian corpus with the final results averaged across the speakers of the same gender. We call "Est-DM" for the 16 male and "Est-DF" for the 12 female talkers and containing 104074 syllables tokens in total.

As for the second experiment, a 10.5-h and 12-talker subset of the American English Buckeye corpus [24] and a Tsonga dataset [25] containing a total of 4.4 hours of speech from 24 talkers were used similarly to the ZS-2015 challenge (see [16] for details). Following [1], syllable clustering was done in a speaker dependent setting in the second experiment. Syllabification was carried out using a sonority envelope -based method described in [26], an improved version from the one described in [1].

In both experiments, standard 13-dimensional MFCC features (25-ms window and 10-ms step size) were extracted

for each syllable segment. Similarly to [1, 2], the syllable tokens were divided into 10 equal-length non-overlapping sub-segments, over which the MFCC features were averaged. The resulting vectors were then concatenated to create a 130-dimensional representation for each syllable. Finally, the vectors were normalised to unit vectors similarly to [2] after ensuring in preliminary experiments that the GMM performance is not affected by the process.

3.2. Evaluation

• ***Clustering*** **-** Performance was measured using purity of the resulting clusters, each phonetically annotated syllable type consisting of a separate class. This overall purity, Q_{tot}, was calculated as an extension to the commonly used standard cluster purity Q_{clust}, that is used in several zero-resource settings [2, 27, 28] defined as

$$Q_{clust} = \sum_z (\max_c p(c|z)) n_z / \sum_z n_z, \quad (3)$$

where $p(c|z)$ is the proportion of class c samples in cluster z and n_z is the number of samples. Concentration of class-specific samples into a specific cluster, Q_{class}, was computed similarly to (3) but using $p(z|c)$. Q_{tot} is then the harmonic mean of Q_{clust} and Q_{class}. Q_{tot} yields a value of one if one–to–one mapping between clusters and sample classes exists. It otherwise penalises clusters with data from multiple classes or having multiple parallel clusters for the same class.

• ***ZS evaluation*** **-** All evaluations were performed using the Zerospeech evaluation kit described in [29]; the reader is directed to the original paper for full technical details. The basic method in the kit is to represent each discovered pattern as a sequence of the underlying phonemes. The kit then measures several metrics, including the normalized edit distance (NED) between all phoneme sequences belonging to the same pattern class (cluster), the proportion of the corpus covered by the learned patterns (cov), and word level measures such as type and token selectivity of the clusters and word segmentation performance in terms of precision, recall and F-scores.

3.3. Compared methods

For the syllable clustering experiments, the baseline results were obtained using parametric methods with heuristically set number of clusters K, including K-means ("KM-heuristic"; as in [1]) and a BMM with the Dirichlet Distribution on the weights of Gaussians mixture components ("DDGMM"; as in [2]). These were compared to the four BNP models using either DP or PYP priors and VMF or Gaussian components. All models were also compared to K-means with K set to the actual number of syllable classes in the data ("KM-true"). For the ZS system the KM-heuristic was compared with the four BNP models on a speaker dependent basis.

Following [2], the number of clusters for the heuristic methods was set to 5% and 20% of the total number of syllable tokens available for the speaker independent and dependent cases, respectively. The hyperparameters of the spherical Gaussian were set as in [2]: μ_0 was set to mean of all the corresponding data, $\sigma_0 = 0.05$ and $\sigma = 10^{-3}$. The results generally depend on the fixed spherical covariance value, but were found to be consistent within the range $\sigma \in [10^{-2}, 10^{-4}]$. The hyper-parameters associated with the mean direction in VMF components were chosen to correspond to those of the Gaussian model: length-normalised dataset mean for $\boldsymbol{\mu}_0$ and $\beta_0 = 0.05$. The gamma distribution parameters a_0 and b_0 were chosen to favour unconcentrated solutions ($a_0 = 1$) but not to constrain the posterior too much ($b_0 = 0.01$). The hyper-parameters of the weight distribution were set as all ones vector for $\boldsymbol{\alpha}$ while using the DD (as in [2]), $\alpha = 1$ for the DP (similar to previous works in speech as in [5]) and $\alpha = 1$, $d = 0.5$ for the PYP.

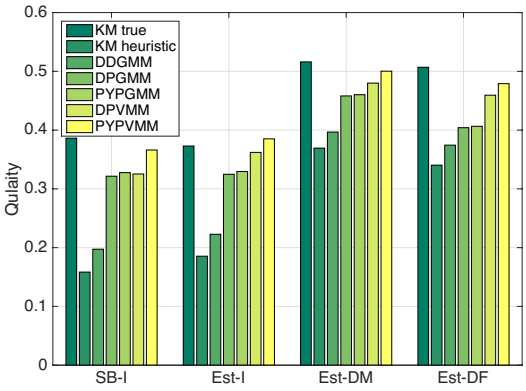

Figure 2: *Bar plots showing the total purities using different models with the 4 datasets.*

T for variational inference was set to 5000 for the speaker independent case and to 50% of the number of syllable tokens for the speaker dependent case. Since optimisation can converge to local maxima, variational distribution parameters were estimated 10 times with the means being reported. K-means was initialised randomly and its results are also averaged across 10 runs. The variational updates were continued until the difference between evidence lower bound in consecutive iterations did not exceed $10^{-4}\%$ or when 400 iterations were reached. MATLAB codes for the variational inference are available under an open source license[1].

4. Results

The results of the first experiment are shown in Figure 2. Since the variation in performance across the 10 runs of each method was very small, we only report means across the runs for each compared method. In general, the purities of the best methods are in the range of 35–50%, being similar to those reported by [2] for word clusters resulting from the full ZS system. This relatively low value indicates the difficult nature of the problem. As expected, speaker dependent clustering results are better than the speaker independent ones. The clustering methods performed at about the same level in the speaker independent experiments on the English and Estonian data, thus indicating language robustness of the approaches.

As expected, the oracle KM-true performs best among compared methods, except in the Est-I dataset where PYPVMM is marginally better. Between the parametric methods the DDGMM performed better than KM-heuristic. The BNP methods however performed better than both of these. This is much more pronounced in the speaker independent case, and shows that the 5% rule-of-thumb for the number of clusters [2] is not necessarily generalizable across data sets. Comparing the BNP methods, the performance of the VMM based methods is consistently higher than the GMMs. PYPVMM—the model with the best matching assumptions with respect to language data type and based on cosine distance—seems to achieve the best performance among the methods and almost approaching the level of KM-true. On the contrary, there is hardly any difference in the performance of the DPGMM and PYPGMMs.

We also analyzed the number of clusters found by different

[1] http://github.com/shreyas253/variational_NP_BMM/

Table 1: *ZS system performance on Buckeye and Tsonga datasets, comparing the baseline ZS system, KM-heuristic and the 4 BNP methods. NED, cov, PRC, RCL and F stand for normalized edit distance, coverage, precision, recall, and F-value (see [29] for details).*

	General		Phoneme grouping			Word token			Word type			Word boundary		
English	NED	cov	PRC	RCL	F	PRC	RCL	F	PRC	RCL	F	PRC	RCL	F
Baseline	**21.9**	16.3	**21.4**	**84.6**	**33.3**	5.5	0.4	0.8	6.2	1.9	2.9	44.1	4.7	8.6
KM-heur	73.4	**111.8**	9.7	7.7	8.5	15.4	**15.3**	**15.3**	**9.0**	25.3	**13.3**	56.7	**57.2**	**56.9**
DPVMM	79.5	111.0	5.8	9.6	7.1	**15.6**	14.3	14.9	**9.0**	**25.5**	**13.3**	**57.7**	54.7	56.1
PYPVMM	78.7	111.1	6.3	9.7	7.5	**15.6**	14.6	15.1	**9.0**	**25.5**	**13.3**	57.4	55.3	56.3
DPGMM	79.2	111.4	6.0	7.7	6.7	15.4	15.0	15.2	**9.0**	25.3	13.2	56.7	56.3	56.4
PYPGMM	79.2	111.3	6.0	7.5	6.7	15.5	15.0	15.2	**9.0**	25.3	13.2	56.7	56.3	56.4
Tsonga	NED	cov	PRC	RCL	F	PRC	RCL	F	PRC	RCL	F	PRC	RCL	F
Baseline	**12.0**	16.2	**52.1**	**77.4**	**62.2**	2.6	0.5	0.8	**3.2**	1.4	2.0	22.3	5.6	8.9
KM-heur	60.8	**100**	15.2	8.9	11.1	2.8	6.1	3.9	2.9	6.1	3.9	29.8	**54.6**	38.5
DPVMM	68.5	**100**	9.5	10.5	9.9	**3.1**	6.1	**4.1**	3.1	**6.4**	**4.2**	**30.5**	53.4	**38.8**
PYPVMM	67.9	**100**	10.2	10.4	10.2	3.0	6.1	4.0	3.1	**6.4**	**4.2**	30.2	53.9	38.7
DPGMM	67.7	**100**	10.3	7.4	8.6	2.8	6.1	3.8	2.8	5.9	3.8	29.1	55.0	38.1
PYPGMM	67.5	**100**	10.2	7.0	8.3	2.8	**6.2**	3.9	2.9	6.0	3.9	29.2	55.3	38.2

Table 2: *Number of syllable classes found by the 4 BNP methods on 4 datasets, shown as a percentage of the true number of syllable classes for each dataset.*

Datasets	**GMM**		**VMM**	
	DP	PYP	DP	PYP
SB-I	64.99	67.51	60.90	85.69
Est-I	66.93	69.73	87.56	109.56
Est-DM	74.31	75.59	73.71	87.88
Est-DF	70.56	72.41	74.01	89.39

BNP methods (Table 2). Since the PYP better approximates the Zipfian distribution typical to syllables, we would expect it to create more clusters than DP. This is also exactly what is observed in the results. In addition, the effect is again more pronounced with the VMMs than GMMs where the PYPVMMs find significantly more clusters than DPVMMs. Table 1 shows the results from the full ZS pipeline. Overall, the performance of all the BNP methods is similar to the heuristic k-means. As in the original pipeline [1], the performance of the BNP methods is greatly above the challenge baseline system [16, 30] in word discovery but much worse in terms of phoneme grouping. This is because the present system attempts to perform full parse of the corpus while the DTW-based baseline system outputs only a small number of well matching segments of speech, as reflected by its low coverage and word-level performance. The improvements of using VMM over GMM observed in the first experiment are still present, but much more marginal due to the increased complexity of the task due to less accurate syllabification.

5. Discussion and Conclusions

The present paper aimed at investigating feasibility of different BNP methods in clustering of syllabic units from speech. The comparison shows that they can all achieve relatively consistent performance across different subsets (and languages) of syllable data, having comparable or better performance than the earlier parametric methods used in ZS systems with heuristically set number of clusters [1, 2]. Still, K-means also leads to good performance in cases where the number of clusters K can be somehow defined in advance. However, setting the appropriate number of clusters is generally problematic in the zero-resource domain where no labeled data are assumed to exist, although some measures such as silhouette width [31] exist for that purpose. In contrast, the consistent performance of the BNPs across the four different languages implies that the BNP methods can be flexibly used in domains where validation of the obtained clustering solutions is not possible and without the need for external measures for quality of different clustering solutions. In addition, variational inference makes the BNP methods computationally tractable with even modest computational resources, which is in contrast to alternatives such as Gibb's sampling that is much slower and faces convergence issues.

As for the comparison between the BNP methods, it was observed that there was a difference in performance and in the number of clusters found between the PYP and DP weight distributions. This difference was more prominent with the VMM component model that also performed better than GMM in all cases. This confirms our initial assumptions that the PYP prior might be better to model Zipfian like data, while VMM is more suited for density estimation in a high-dimensional feature space with cosine distance. However, it should be noted that for the GMM components we had to use fixed covariance parameters in order to acquire well-performing solutions for the clustering tasks, a problem already encountered in [2]. While this may lead to suboptimal cluster shapes and sizes, it also indicates that the VMM is more suited for the present type of task where no such parameter tying/fixing is required for successful model inference with noisy high-dimensional data. These differences are less apparent when there is additional noise due to uncertainty in the syllable boundaries.

Overall, the present results show that non-parametric methods perform consistently across several data sets with the same set of prior parameters and therefore provide a potential alternative to more traditional parametric methods.

6. Acknowledgements

The study was supported by the Academy of Finland project no. 274479. Thanks to Pärtel Lippus for access to the Estonian corpus. The MATLAB scripts for variational inference the Bayesian methods can be found at `http://github.com/shreyas253/variational_NP_BMM/`

7. References

[1] O. Räsänen, G. Doyle, and M. C. Frank, "Unsupervised word discovery from speech using automatic segmentation into syllable-like units," in *Proc. Interspeech*, 2015, pp. 3204–3208.

[2] H. Kamper, A. Jansen, and S. Goldwater, "A segmental framework for fully-unsupervised large-vocabulary speech recognition," *CoRR*, vol. abs/1606.06950, 2016. [Online]. Available: http://arxiv.org/abs/1606.06950

[3] S. J. Gershman and D. M. Blei, "A tutorial on Bayesian nonparametric models," *Journal of Mathematical Psychology*, vol. 56, no. 1, pp. 1–12, 2012.

[4] C. E. Antoniak, "Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems," *The Annals of Statistics*, vol. 2, no. 6, pp. 1152–1174, 1974.

[5] C. Lee and J. Glass, "A nonparametric Bayesian approach to acoustic model discovery," in *Proc. Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2012, pp. 40–49.

[6] A. Rosenberg, "Modeling prosodic sequences with k-means and Dirichlet process GMMs," in *Proc. Interspeech*, 2013, pp. 520–524.

[7] H. Kamper, A. Jansen, S. King, and S. Goldwater, "Unsupervised lexical clustering of speech segments using fixed-dimensional acoustic embeddings," in *Proc. Spoken Language Technology Workshop*. IEEE, 2014, pp. 100–105.

[8] H. Chen, C.-C. Leung, L. Xie, B. Ma, and H. Li, "Parallel inference of Dirichlet process Gaussian mixture models for unsupervised acoustic modeling: A feasibility study," in *Proc. Interspeech*, 2015, pp. 3189–3193.

[9] E. B. Fox, E. B. Sudderth, M. I. Jordan, and A. S. Willsky, "An HDP-HMM for systems with state persistence," in *Proc. ICML*, 2008, pp. 312–319.

[10] S. H. Shum, N. Dehak, R. Dehak, and J. R. Glass, "Unsupervised methods for speaker diarization: An integrated and iterative approach," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2015–2028, 2013.

[11] S. Greenberg, "Speaking in shorthand–a syllable-centric perspective for understanding pronunciation variation," *Speech Communication*, vol. 29, no. 2, pp. 159–176, 1999.

[12] J. Pitman and M. Yor, "The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator," *The Annals of Probability*, vol. 25, no. 2, pp. 855–900, 1997.

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

[14] A. Banerjee, I. S. Dhillon, J. Ghosh, and S. Sra, "Clustering on the unit hypersphere using von Mises-Fisher distributions," *J. Machine Learning Research*, vol. 6, pp. 1345–1382, 2005.

[15] S. Seshadri, U. Remes, and O. Räsänen, "Dirichlet process mixture models for clustering i-vector data," in *ICASSP-2017*, 2017, pp. 5470–5474.

[16] M. Versteegh, R. Thiollière, T. Schatz, X. N. Cao, X. Anguera, A. Jansen, and E. Dupoux, "The Zero Resource Speech Challenge 2015," in *Proc. Interspeech*, 2015, pp. 3169–3173.

[17] J. Taghia, Z. Ma, and A. Leijon, "Bayesian estimation of the von-Mises Fisher mixture model with variational inference," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 36, no. 9, pp. 1701–1715, 2014.

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

[19] J. Sethuraman, "A constructive definition of Dirichlet priors," *Statistica Sinica*, vol. 4, no. 2, pp. 639–650, 1994.

[20] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[21] D. M. Blei and M. I. Jordan, "Variational inference for Dirichlet process mixtures," *Bayesian analysis*, vol. 1, no. 1, pp. 121–144, 2006.

[22] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, vol. 1. IEEE, 1992, pp. 517–520.

[23] Phonetic corpus of Estonian spontaneous speech. http://www.keel.ut.ee/en/languages-resourceslanguages-resources/phonetic-corpus-estonian-spontaneous-speech. Accessed: 2016-09-02.

[24] M. Pit, L. Dilley, K. Johnson, S. Kiesling, W. Raymond, E. Hume, and E. Fosler-Lussier, "Buckeye corpus of conversational speech (2nd release)," Columbus, OH: Department of Psychology, Ohio State University (Distributor)., Tech. Rep.

[25] N. J. De Vries, M. H. Davel, J. Badenhorst, W. D. Basson, F. De Wet, E. Barnard, and A. De Waal, "A smartphone-based ASR data collection tool for under-resourced languages," *Speech Communication*, vol. 56, pp. 119–131, 2014.

[26] O. Räsänen, G. Doyle, and M. C. Frank, "Pre-linguistic rhythmic segmentation of speech into syllabic units," *submitted for publication*.

[27] H. Kamper, A. Jansen, and S. Goldwater, "Unsupervised word segmentation and lexicon discovery using acoustic word embeddings," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 4, pp. 669–679, 2016.

[28] M. Sun *et al.*, "Joint training of non-negative Tucker decomposition and discrete density hidden markov models," *Computer Speech & Language*, vol. 27, no. 4, pp. 969–988, 2013.

[29] B. Ludusan, M. Versteegh, A. Jansen, G. Gravier, X.-N. Cao, M. Johnson, and E. Dupoux, "Bridging the gap between speech technology and natural language processing: an evaluation toolbox for term discovery systems," in *Language Resources and Evaluation Conference*, 2014.

[30] A. Jansen and B. Van Durme, "Efficient spoken term discovery using randomized algorithms," in *Proc. ASRU-2011*, 2011, pp. 401–406.

[31] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.