# BAYESIAN MODELS FOR UNIT DISCOVERY ON A VERY LOW RESOURCE LANGUAGE

*Lucas Ondel*[1]*, Pierre Godard*[2]*, Laurent Besacier*[3]*, Elin Larsen*[6]*, Mark Hasegawa- Johnson*[4]
*Odette Scharenborg*[5]*, Emmanuel Dupoux*[6]*, Lukas Burget*[1]*, François Yvon*[2]*, Sanjeev Khudanpur*[7]

1. Brno University of Technology, Brno, Czech Republic, 2. LIMSI, CNRS, Université Paris Saclay
3. LIG, CNRS, Université Grenoble Alpes, 4. University of Illinois, Urbana, IL, USA
5. Centre for Language Studies, Radboud University, Nijmegen, Netherlands
6. CoML, ENS/EHESS/PSL Research University/CNRS/INRIA, Paris, France
7. Johns Hopkins University, Baltimore, MD USA

## ABSTRACT

Developing speech technologies for low-resource languages has become a very active research field over the last decade. Among others, Bayesian models have shown some promising results on artificial examples but still lack of *in situ* experiments. Our work applies state-of-the-art Bayesian models to unsupervised Acoustic Unit Discovery (AUD) in a real low-resource language scenario. We also show that Bayesian models can naturally integrate information from other resourceful languages by means of *informative prior* leading to more consistent discovered units. Finally, discovered acoustic units are used, either as the 1-best sequence or as a lattice, to perform word segmentation. Word segmentation results show that this Bayesian approach clearly outperforms a Segmental-DTW baseline on the same corpus.

***Index Terms***— Acoustic Unit Discovery, Low-Resource ASR, Bayesian Model, Informative Prior.

## 1. INTRODUCTION

Out of nearly 7000 languages spoken worldwide, current speech (ASR, TTS, voice search, etc.) technologies barely address 200 of them. Broadening ASR technologies to ideally all possible languages is a challenge with very high stakes in many areas and is at the heart of several fundamental research problems ranging from psycholinguistic (how humans learn to recognize speech) to pure machine learning (how to extract knowledge from unlabeled data). The present work focuses on the narrow but important problem of unsupervised Acoustic Unit Discovery (AUD). It takes place as the continuation of an ongoing effort to develop a Bayesian model suitable for this task, which stems from the seminal work of [1] later refined and made scalable in [2]. This model, while rather crude, has shown that it can provide a clustering accurate enough to be used in topic identification of spoken document in unknown languages [3]. It was also shown that this model can be further improved by incorporating a Bayesian "phonotactic" language model learned jointly with the acoustic units [4]. Finally, following the work in [5] it has been combined successfully with variational auto-encoders leading to a model combining the potential of both deep neural networks and Bayesian models [6]. The contribution of this work is threefold:

- we compare two Bayesian models ([2] and [6]) for acoustic unit discovery (AUD) on a very low resource language speech corpus,
- we investigate the use of "informative prior" to improve the performance of Bayesian models by using information from resourceful languages,
- as an extrinsic evaluation of AUD quality, we cascade AUD with sequence/lattice based word discovery [7].

## 2. MODELS

The AUD model described in [1, 2] is a non-parametric Bayesian Hidden Markov Model (HMM). This model is topologically equivalent to a phone-loop model with two major differences:

- since it is trained in an unsupervised fashion the elements of the loop cannot directly be interpreted as the actual phones of the target language but rather as some acoustic units (defined as 3-states left-to-right sub-HMM) whose time scale approximately corresponds the phonetic time scale.
- to cope with the unknown number of acoustic units needed to properly describe speech, the model assumes a theoretically infinite number of potential acoustic units. However, during inference, the prior over the weight of the acoustic units (a Dirichlet Process [8])

O. Scharenborg was supported by a Vidi-grant from NWO (grant number: 276-89-003)

will act as a sparsity regularizer leading to a model which explains the data with a relatively small number of units.

In this work, we have used two variants of this original model. The first one (called HMM model in the remainder of this paper), following the analysis led in [9], approximates the Dirichlet Process prior by a mere symmetric Dirichlet prior. This approximation, while retaining the sparsity constraint, avoids the complication of dealing with the variational treatment of the stick breaking process frequent in Bayesian non-parametric models. The second variant, which we shall denote Structured Variational AutoEncoder (SVAE) AUD, is based upon the work of [5] and embeds the HMM model into the Variational AutoEncoder framework [10]. A very similar version of the SVAE for AUD was developed independently and presented in [6]. The main noteworthy difference between [6] and our model is that we consider a fully Bayesian version of the HMM embedded in the VAE; and the posterior distribution and the VAE parameters are trained jointly using the Stochastic Variational Bayes [5, 11]. For both variants, the prior over the HMM parameters were set to the conjugate of the likelihood density: Normal-Gamma prior for the mean and variance of the Gaussian components, symmetric Dirichlet prior over the HMM's state mixture's weights and symmetric Dirichlet prior over the acoustic units' weights. For the case of the uninformative prior, the prior was set to be vague prior with one pseudo-observation [12] [1].

## 3. INFORMATIVE PRIOR

Bayesian Inference differs from other machine learning techniques by introducing a distribution $p(\boldsymbol{\xi})$ over the parameters of the model. A major concern in Bayesian Inference is usually to define a prior that makes as little assumption as possible. Such a prior is usually known as uninformative prior. Having a completely uninformative prior has the practical advantage that the prior distribution will have a minimal impact on the outcome of the inference leading to a model which bases its prediction purely and solely on the data. In the present work, we aim at the opposite behavior, we wish our AUD model to learn phone-like units from the unlabeled speech data of a target language given the knowledge that was previously accumulated from another resourceful language. More formally, the original AUD model training consists in estimate the *a posteriori* distribution of the parameters given the unlabeled speech data of a target language $\mathbf{X}_t$:

$$p(\boldsymbol{\xi}|\mathbf{X}_t) = \frac{p(\mathbf{X}_t|\boldsymbol{\xi})p(\boldsymbol{\xi})}{p(\mathbf{X}_t)} \tag{1}$$

[1] Because of lack of space, we have only given a rudimentary description of the models. Note that the HMM model was described at length in [2] whereas the full description of the SVAE model is yet to be published. However, the implementation of both models is available at https://github.com/amdtkdev/amdtk

The parameters are divided into two subgroups $\boldsymbol{\xi} = \{\boldsymbol{\eta}, \mathbf{u}_t\}$ where $\boldsymbol{\eta}$ are the global parameters of the model, and $\mathbf{u}_t$ are the latent variables which, in our case, correspond to the sequences of acoustic units. The global parameters are separated into two independent subsets : $\boldsymbol{\eta} = \{\boldsymbol{\eta}_A, \boldsymbol{\eta}_L\}$, corresponding to the acoustic parameters ($\boldsymbol{\eta}_A$) and the "phonotactic" language model parameters ($\boldsymbol{\eta}_L$). Replacing $\boldsymbol{\eta}$ and following the conditional independence of the variable induced by the model (see [2] for details) leads to:

$$p(\mathbf{u}_t, \boldsymbol{\eta}|\mathbf{X}_t) \propto p(\mathbf{X}_t|\mathbf{u}_t, \boldsymbol{\eta}_A)p(\mathbf{u}_t|\boldsymbol{\eta}_L)p(\boldsymbol{\eta}_L)p(\boldsymbol{\eta}_A) \tag{2}$$

If we further assume that we have at our disposal speech data in a different language than the target one, denoted $\mathbf{X_p}$, along with its phonetic transcription $\mathbf{u}_p$, it is then straightforward to show that:

$$p(\boldsymbol{\eta}, \mathbf{u}_t|\mathbf{X}_t, \mathbf{X}_p, \mathbf{u}_p) \propto \quad p(\mathbf{X}_t|\mathbf{u}_t, \boldsymbol{\eta}_A)p(\mathbf{u}_t|\boldsymbol{\eta}_L)p(\boldsymbol{\eta}_A|\mathbf{X}_p, \mathbf{u}_p) \tag{3}$$

which is the same as Eq. 2 but for the distribution of the acoustic parameters which is based on the data of the resourceful language. In contrast of the term uninformative prior we denote $p(\boldsymbol{\eta}_A|\mathbf{X}_p, \mathbf{u}_p)$ as an informative prior. As illustrated by Eq. 3, a characteristic of Bayesian inference is that it naturally leads to a sequential inference. Therefore, model training can be summarized as:

- given some prior data $\mathbf{X}_p$ from a resourceful language, estimate a posterior distribution over the acoustic parameters $p(\boldsymbol{\eta}_A|\mathbf{X}_p)$
- for a new unlabeled speech corpus, estimate the posterior distribution but considering the learned posterior distribution $p(\boldsymbol{\eta}_A|\mathbf{X}_p)$ as a "prior".

Practically, the computation of the informative prior as well as the final posterior distribution is intractable and we seek for an approximation by means of the well known Variational Bayes Inference [13]. The approximate informative prior $q_1(\boldsymbol{\eta}_A)$ is estimated by optimizing the variational lower bound of the evidence of the prior data $\mathbf{X}_p$:

$$\begin{aligned} q_1^* = \arg\max_{q_1} \; & \mathrm{E}_{q_1(\boldsymbol{\eta}_A)}\left[\ln p(\mathbf{X}_p, \boldsymbol{\eta}_A|\mathbf{u}_p, )\right] \\ & - \mathrm{D}_{\mathrm{KL}}(q_1(\boldsymbol{\eta}_A)||p(\boldsymbol{\eta}_A)) \end{aligned} \tag{4}$$

where $\mathrm{D}_{\mathrm{KL}}$ is the Kullback-Leibler divergence. Then, the posterior distribution of the parameters given the target data $q_2(\mathbf{u}_t, \boldsymbol{\eta}_A, \boldsymbol{\eta}_L)$ can be estimated by optimizing the evidence of the target data $\mathbf{X}_t$:

$$\begin{aligned} q_2^* = \arg\max_{q_2} \; & \mathrm{E}_{q_2(\mathbf{u}_t, \boldsymbol{\eta}_A, \boldsymbol{\eta}_L)}\left[\ln p(\mathbf{X}_t, \mathbf{u}_t, \boldsymbol{\eta}_A, \boldsymbol{\eta}_L)\right] \\ & - \mathrm{D}_{\mathrm{KL}}(q_2(\boldsymbol{\eta}_A)||q_1(\boldsymbol{\eta}_A)) \\ & - \mathrm{D}_{\mathrm{KL}}(q_2(\mathbf{u}_t, \boldsymbol{\eta}_L)||p(\mathbf{u}_t, \boldsymbol{\eta}_L)) \end{aligned} \tag{5}$$

Note that when the model is trained with an uninformative prior the loss function is the as in Eq. 5 but with $p(\boldsymbol{\eta}_A)$ instead of the $q_1(\boldsymbol{\eta}_A)$. For the case of the uninformative prior, the Variational Bayes Inference was initialized as described in [2]. In the informative prior case, we initialized the algorithm by setting $q_2(\boldsymbol{\eta}_A) = q_1(\boldsymbol{\eta}_A)$.

## 4. EXPERIMENTAL SETUP

### 4.1. Corpora and acoustic features

We used the Mboshi5K corpus [14] as a test set for all the experiments reported here. Mboshi (Bantu C25) is a typical Bantu language spoken in Congo-Brazzaville. It is one of the languages documented by the BULB (Breaking the Unwritten Language Barrier) project [15]. This speech dataset was collected following a real language documentation scenario, using Lig_Aikuma[2], a mobile app specifically dedicated to fieldwork language documentation, which works both on android powered smartphones and tablets [16]. The corpus is multilingual (5130 Mboshi speech utterances aligned to French text) and contains linguists' transcriptions in Mboshi (in the form of a non-standard graphemic form close to the language phonology). It is also enriched with automatic forced-alignment between speech and transcriptions. The dataset is made available to the research community[3]. More details on this corpus can be found in [14].

TIMIT is also used as an extra speech corpus to train the informative prior. We used two different set of features: the mean normalized MFCC + $\Delta$ + $\Delta\Delta$ generated by HTK and the Multilingual BottleNeck (MBN) features [17] trained on the Czech, German, Portuguese, Russian, Spanish, Turkish and Vietnamese data of the Global Phone database.

### 4.2. Acoustic unit discovery (AUD) evaluation

To evaluate our work we measured how the discovered units compared to the forced aligned phones in term of segmentation and information. The accuracy of the segmentation was measured in term of Precision, Recall and F-score. If a unit boundary occurs at the same time (+/- 10ms) of an actual phone boundary it is considered as a true positive, otherwise it is considered to be a false positive. If no match is found with a true phone boundary, this is considered to be a false negative. The consistency of the units was evaluated in term of normalized mutual information (NMI - see [2, 4, 6] for details) which measures the statistical dependency between the units and the forced aligned phones. A NMI of 0 % means that the units are completely independent of the phones whereas a NMI of 100 % indicates that the actual phones could be retrieved without error given the sequence of discovered units.

[2] `http://lig-aikuma.imag.fr`

[3] It will be made available for free from ELRA, but its current version is online on: `https://github.com/besacier/mboshi-french-parallel-corpus`

| Features | Precision | Recall | F-score | NMI |
|---|---|---|---|---|
| MFCC | 28.40 | **54.36** | 37.36 | 17.92 |
| MBN | 24.60 | 41.71 | 30.95 | 14.81 |

**Table 1**: AUD results of the baseline (HMM model with uninformative prior) - Mboshi5k corpus

### 4.3. Extension to word discovery

In order to provide an extrinsic metric to evaluate the quality of the acoustic units discovered by our different methods, we performed an unsupervised word segmentation task on the acoustic units sequences, and evaluated the accuracy of the discovered word boundaries. We also wanted to experiment using lattices as an input for the word segmentation task, instead of using single sequences of units, so as to better mitigate the uncertainty of the AUD task and provide a companion metric that would be more robust to noise. A model capable of performing word segmentation both on lattices and text sequences was introduced by [7]. Building on the work of [18, 19] they combine a nested hierarchical Pitman-Yor language model with a Weighted Finite State Transducer approach. Both for lattices and acoustic units sequences, we use the implementation of the authors with a bigram language model and a unigram character model[4]. Word discovery is evaluated using the *Boundary* metric from the *Zero Resource Challenge 2017* [21] and [22]. This metric measures the quality of a word segmentation and the discovered boundaries with respect to a gold corpus (Precision, Recall and F-score are computed).

## 5. RESULTS AND DISCUSSION

First, we evaluated the standard HMM model with an uninformative prior (this will be our baseline) for the two different input features: MFCC (and derivatives) and MBN. Results are shown in Table 1. Surprisingly, the MBN features perform relatively poorly compared to the standard MFCC. These results are contradictory to those reported in [4]. Two factors may explain this discrepancy: the Mboshi5k data being different from the training data of the MBN neural network, the neural network may not generalize well. Another possibility may be that the initialization scheme of the model is not suitable for this type of features. Indeed, Variational Bayesian Inference algorithm converges only to a local optimum of the objective function and is therefore dependent of the initialization. We believe the second explanation is the more likely since, as we shall see shortly, the best results in term of word segmentation and NMI are eventually obtained with the MBN features when the inference is done with the informative prior.

[4] It would be more natural to use a 4-gram or an even higher order spelling model for word discovery, but we wanted to be able to validate our metric by matching it with the model of [20] (*dpseg*) which implements a bigram language model based on a unigram model of characters (see details in Table 2).

| Features | Model | Inf. Prior | Precision | Recall | F-score | Precision | Recall | F-score |
|---|---|---|---|---|---|---|---|---|
| | | | 1-best Word Seg. | | | Lattices Word Seg. | | |
| MFCC | HMM | no | 28.8 | 74.5 | 41.5 | 29 | 75.9 | 41.9 |
| MFCC | HMM | yes | 28.5 | **79.1** | 41.9 | 29.3 | **78.1** | 42.6 |
| MFCC | SVAE | no | 28.7 | 77.2 | 41.9 | 30 | 74.4 | 42.8 |
| MFCC | SVAE | yes | 29.3 | 73.1 | 41.8 | 30.4 | 69.6 | 42.3 |
| MBN | HMM | no | **30.3** | 69 | 42.1 | 30.8 | 66.1 | 42 |
| MBN | HMM | yes | 29.2 | 67.5 | 40.8 | 29.9 | 67.8 | 41.5 |
| MBN | SVAE | no | 29.2 | 68.3 | 41 | 29.6 | 68.1 | 41.3 |
| MBN | SVAE | yes | 29.8 | 73.4 | **42.4** | **30.9** | 72.2 | **43.3** |

**Table 2**: Precision, Recall and F-measure on word boundaries, using different AUD methods. Segmental DTW baseline [23] gave F-score of 19.3% on the exact same corpus; $dpseg$ [20] was also used as a word segmentation baseline and gave similar (slightly lower) F-scores to *1-best* (best config with $dpseg$ gave 42.5%) - Mboshi5k corpus

| Features | Model | Average Unit duration (s) |
|---|---|---|
| phones | | 0.091 |
| MFCC | HMM | 0.082 |
| MFCC | SVAE | 0.096 |
| MBN | HMM | 0.093 |
| MBN | SVAE | 0.102 |

**Table 3**: Average duration of the (AUD) units (AUD) for the HMM and SVAE models trained with an uninformative prior. "phones" refers to the forced aligned phone reference.

| Features | Model | Inf. Prior | Precision | Recall | F-score | NMI |
|---|---|---|---|---|---|---|
| MFCC | HMM | no | 28.4 | **54.36** | **37.6** | 17.92 |
| MFCC | HMM | yes | 29.88 | 47.34 | 36.64 | 20.42 |
| MFCC | SVAE | no | 30.1 | 49.29 | 37.38 | 21.03 |
| MFCC | SVAE | yes | **35.85** | 25.59 | 29.87 | 21.67 |
| MBN | HMM | no | 24.6 | 41.71 | 30.95 | 14.81 |
| MBN | HMM | yes | 27.8 | 36.58 | 31.56 | 20.34 |
| MBN | SVAE | no | 26.8 | 41.51 | 32.57 | 18.33 |
| MBN | SVAE | yes | 30.75 | 37.94 | **33.97** | **23.49** |

**Table 4**: Effect of the informative prior on AUD (phone boundary detection) - Mboshi5k corpus

Next, we compared the HMM and the SVAE models when trained with an uninformative prior (lines with "Inf. Prior" set to "no" in Table 4). The SVAE significantly improves the NMI and the precision showing that it extracts more consistent units than the HMM model. However, it also degrades the segmentation in terms of recall. We further investigated this behavior by looking at the duration of the units found by both models compared to the true phones (Table 3). We observe that the SVAE model favors longer units than the HMM model hence leading to fewer boundaries and consequently smaller recall.

We then evaluated the effect of the informative prior on the acoustic unit discovery (Table 4). On all 4 combinations (2 features sets $\times$ 2 models) we observe an improvement in terms of precision and NMI but a degradation of the recall. This result is encouraging since the informative prior was trained on English data (TIMIT) which is very different from Mboshi. Indeed, this suggests that even speech from an unrelated language can be of some help in the design of an ASR for a very low resource language. Finally, similarly to the SVAE/HMM case described above, we found that the degradation of the recall is due to longer units discovered for models with an informative prior (numbers omitted due to lack of space).

Word discovery results are given in Table 2 for the *Boundary* metric [21, 22]. We observe that i) the best word boundary detection (F-score) is obtained with MBN features, an informative prior and the SVAE model; this confirms the results of table 4 and shows that better AUD leads to better word segmentation ii) word segmentation from AUD graph *Lattices* is slightly better than from flat sequences of AUD symbols (*1-best*); iii) our results outperform a pure speech based baseline based on segmental DTW [23] (F-score of 19.3% on the exact same corpus).

## 6. CONCLUSION

We have conducted an analysis of the state-of-the-art Bayesian approach for acoustic unit discovery on a real case of low-resource language. This analysis was focused on the quality of the discovered units compared to the gold standard phone alignments. Outcomes of the analysis are i) the combination of neural network and Bayesian model (SVAE) yields a significant improvement in the AUD in term of consistency ii) Bayesian models can naturally embed information from a resourceful language and consequently improve the consistency of the discovered units. Finally, we hope this work can serve as a baseline for future research on unsupervised acoustic unit discovery in very low resource scenarios.

## 7. ACKNOWLEDGEMENTS

This work was started at JSALT 2017 in CMU, Pittsburgh, and was supported by JHU and CMU (via grants from Google, Microsoft, Amazon, Facebook, Apple), by the Czech Ministry of Education, Youth and Sports from the National Programme of Sustainability (NPU II) project "IT4Innovations excellence in science - LQ1602" and by the French ANR and the German DFG under grant ANR-14-CE35-0002 (BULB project). This work used the Extreme Science and Engineering Discovery Environment (NSF grant number OCI-1053575 and NSF award number ACI-1445606).

## 8. REFERENCES


[1] C. Lee et al., "A nonparametric bayesian approach to acoustic model discovery," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, Stroudsburg, PA, USA, 2012, ACL '12, pp. 40–49, Association for Computational Linguistics.

[2] L. Ondel et al., "Variational inference for acoustic unit discovery," in *Procedia Computer Science*. 2016, vol. 2016, pp. 80–86, Elsevier Science.

[3] S. Kesiraju et al., "Topic identification of spoken documents using unsupervised acoustic unit discovery," in *Proceedings of ICASSP 2017*. 2017, pp. 5745–5749, IEEE Signal Processing Society.

[4] L. Ondel et al., "Bayesian phonotactic language model for acoustic unit discovery," in *Proceedings of ICASSP 2017*. 2017, pp. 5750–5754, IEEE Signal Processing Society.

[5] M. Johnson et al., "Composing graphical models with neural networks for structured representations and fast inference," in *Advances in Neural Information Processing Systems 29*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, Eds., pp. 2946–2954. Curran Associates, Inc., 2016.

[6] Janek Ebbers et al., "Hidden markov model variational autoencoder for acoustic unit discovery," in *Proceedings of INTERSPEECH 2017*, 2017.

[7] J. Heymann et al., "Unsupervised word segmentation from noisy input," in *2013 IEEE Workshop on Automatic Speech Recognition and Understanding, Olomouc, Czech Republic, December 8-12, 2013*, 2013, pp. 458–463.

[8] Y. W. Teh et al., "Hierarchical Bayesian nonparametric models with applications," in *Bayesian Nonparametrics: Principles and Practice*, N. Hjort, C. Holmes, P. Müller, and S. Walker, Eds. Cambridge University Press, 2010.

[9] K. Kurihara et al., "Collapsed variational Dirichlet process mixture models," in *IJCAI'07: Proceedings of the 20th international joint conference on Artifical intelligence*, San Francisco, CA, USA, 2007, pp. 2796–2801, Morgan Kaufmann Publishers Inc.

[10] D. Kingma et al., "Auto-encoding variational bayes," *CoRR*, vol. abs/1312.6114, 2013.

[11] D. Matthew et al., "Stochastic variational inference," *Journal of Machine Learning Research*, vol. 14, pp. 1303–1347, 2013.

[12] C. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*, Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[13] M. Jordan et al., "An introduction to variational methods for graphical models," *Mach. Learn.*, vol. 37, no. 2, pp. 183–233, Nov. 1999.

[14] P. Godard et al., "A very low resource language speech corpus for computational language documentation experiments," 2017.

[15] G. Adda et al., "Breaking the unwritten language barrier: The Bulb project," in *Proceedings of SLTU (Spoken Language Technologies for Under-Resourced Languages)*, Yogyakarta, Indonesia, 2016.

[16] D. Blachon et al., "Parallel speech collection for under-resourced language studies using the LIG-Aikuma mobile device app," in *Proceedings of SLTU (Spoken Language Technologies for Under-Resourced Languages)*, Yogyakarta, Indonesia, May 2016.

[17] František Grézl et al., "Adapting multilingual neural network hierarchy to a new language," in *Proceedings of the 4th International Workshop on Spoken Language Technologies for Under- resourced Languages SLTU-2014. St. Petersburg, Russia, 2014*. 2014, pp. 39–45, International Speech Communication Association.

[18] D. Mochihashi, "Bayesian unsupervised word segmentation with nested pitman-yor language modeling," 2009, ACL '09.

[19] G. Neubig, "Bayesian learning of a language model from continuous speech," *IEICE Transactions on Information and Systems (¡A href="http://search.ieice.org/"¿link¡/a¿)*, vol. E95-D, no. 2, pp. 614–625, February 2012.

[20] S. Goldwater et al., "A Bayesian framework for word segmentation: Exploring the effects of context," *Cognition*, vol. 112, no. 1, pp. 21–54, 2009.

[21] B. Ludusan et a., "Bridging the gap between speech technology and natural language processing: an evaluation toolbox for term discovery systems," in *Proceedings of LREC*, 2014.

[22] E. Dunbar et al, "The zero resource speech challenge 2017," in *Automatic Speech Recognition and Understanding (ASRU), 2017 IEEE Workshop on*. IEEE, 2017.

[23] A. Jansen et al., "Efficient spoken term discovery using randomized algorithms," in *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*. IEEE, 2011, pp. 401–406.