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

Unsupervised Speech Signal to Symbol Transformation for Zero Resource Speech Applications

Saurabhchand Bhati, Shekhar Nayak and K Sri Rama Murty

Department of Electrical Engineering, IIT Hyderabad, Hyderabad, India
ee12b1044@iith.ac.in, ee13p1008@iith.ac.in, ksrm@iith.ac.in

Abstract

Zero resource speech processing refers to a scenario where no or minimal transcribed data is available. In this paper, we propose a three-step unsupervised approach to zero resource speech processing, which does not require any other information/dataset. In the first step, we segment the speech signal into phoneme-like units, resulting in a large number of varying length segments. The second step involves clustering the varying-length segments into a finite number of clusters so that each segment can be labeled with a cluster index. The unsupervised transcriptions, thus obtained, can be thought of as a sequence of virtual phone labels. In the third step, a deep neural network classifier is trained to map the feature vectors extracted from the signal to its corresponding virtual phone label. The virtual phone posteriors extracted from the DNN are used as features in the zero resource speech processing. The effectiveness of the proposed approach is evaluated on both ABX and spoken term discovery tasks (STD) using spontaneous American English and Tsonga language datasets, provided as part of zero resource 2015 challenge. It is observed that the proposed system outperforms baselines, supplied along the datasets, in both the tasks without any task specific modifications.

Index Terms: Phonetic segmentation, Zero resource speech processing, Unsupervised learning, ABX, Spoken term discovery, Pitman-Yor language model, Deep neural network

1. Introduction

Speech signal conveys rich information about several sources including the message, speaker identity, language of communication, background environment etc. Generic features, like mel-frequency cepstral coefficients (MFCC), frequency domain linear prediction (FDLP) features [1], or modified group-delay (ModGD) features [2], contain information about all these sources. One of the important issues in speech processing is to extract features that highlight the characteristics of the desired source. For example, a keyword spotting system requires features that highlight the speech (phoneme) specific characteristics, while speaker recognition system requires features that highlight speaker-specific characteristics. In this paper, we present an unsupervised approach to extract features that highlight the speech specific characteristics, which can be used in applications like keyword spotting, spoken term detection, and speech recognition.

Speech signal can be considered as a sequence of basic sound units that constitute the intended message. One of the important tasks in speech signal processing is to segment the continuous speech signal into the basic sound units, with an objective to assign a symbolic representation of the sequence of acoustic segments [3]. It is desirable to have these symbols correspond to some meaningful linguistic units. The choice for the symbol could range from a phonetic unit at the lowest level to the word or even phrase at the highest level [4]. Depending on the availability of the data, signal to symbol transformation can be attempted in supervised or unsupervised manner. Supervised approaches require manual transcriptions of the acoustic speech signal, while the unsupervised approaches work with acoustic input alone. Over the past one decade, deep neural networks (DNNs) have been highly successful in supervised speech signal to symbol transformation. In this approach, the DNNs are used for acoustic modelling, which is aimed at mapping every frame of the speech signal to its manually transcribed phoneme label. The deep acoustic models are used to generate phoneme-posterior features, which offer better speaker-independence. Hence, the phoneme-posterior features achieved significant performance boost in several applications including speech recognition [5], keyword spotting [6], spoken term detection etc. Deep acoustic modelling is a supervised task and requires a large amount of manually transcribed speech data, and hence, it cannot be readily extended to situations where manually transcribed speech data is either limited or completely absent. Therefore, there is an increased interest in the speech community to develop acoustic models which depend little on linguistic knowledge and amount of transcribed data [7].

There have been attempts for unsupervised extraction of features that highlight the speech-specific characteristics while achieving speaker-independence. Gaussian posterior features extracted from Gaussian mixture model (GMM) improves the performance of STD task. GMM posteriors for different sound units are expected to occupy orthogonal subspaces, leading to a better inter-phone discrimination. Since the GMMs are trained on speech data collected from several speakers, the GMM posterior features exhibit better speaker independence compared to the raw spectral features. Artificial neural network models, especially autoencoders, have been used as an alternate to GMM models to capture the probability density of the raw spectral features [8]. The autoencoder networks are data-driven and are better suited for representation learning [9], since they relax the modelling assumptions that may be inaccurate. The bottleneck features derived from the autoencoders were shown to improve the performance of speech systems [10, 11]. A major disadvantage with GMM and autoencoder approaches is that they consider the data as a set of features, and ignore the sequence in which they evolve. As a result, there could be spurious symbol/feature switches even within a single sound unit segment. In order to overcome this issue, we attempt to model acoustic segments rather than individual frames.

In this paper, we propose an unsupervised approach for speech-specific feature extraction, which retains the advantages of supervised deep acoustic modelling. Our approach is similar to the acoustic segment modelling (ASM) approach [12, 13, 14, 15, 16], and consists of three important steps: speech segmentation, segment labelling and supervised modelling of labeled segments using DNNs. In the segmentation

http://dx.doi.org/10.21437/Interspeech.2017-1476

step, the continuous speech signal is divided into varying-length segments. We propose to use block diagonal structure of the affinity matrix in kernel space to identify the segments in the speech signal. In the segment labelling step, the varying-length segments are compared using a similarity measure and grouped into clusters. Since varying length segments do not admit fixed dimensional representations, clustering methods based on geometric proximity cannot be used. Hence, we employed graph clustering method, which exploits connectivity between the segments rather than geometric proximity. Dynamic time warping (DTW) is used to quantify the connectivity between two varying length segments. The proposed graph clustering approach assigns a unique label to each acoustic segment, which we refer to as virtual phoneme. An ideal labelling algorithm should consistently assign the same label to all the acoustically similar segments. In the segment modelling stage, the segment labels obtained from the previous stage are used to train DNN classifier on virtual phoneme labels. During the training process, the segment boundaries and their labels are iteratively refined to improve segmentation/labelling accuracy. The virtual phoneme posterior features extracted from the DNN, better characterize the speech specific information, and also exhibit better speaker independence. The effectiveness of the proposed approach is demonstrated on zero-resource - 2015 evaluation challenge, which involves ABX and STD tasks [17, 18, 19, 20].

The rest of the paper is organized as follows: An algorithm for segmentation of continuous speech signal into phoneme-like units is presented in Section 2. Section 3 proposes a graph clustering method for labelling the varying length acoustic segments. Unsupervised deep acoustic modelling of speech signals using virtual phoneme labels is discussed in Section 4. The effectiveness of virtual phoneme posteriors is demonstrated on zero-speech - 2015 challenge. Section 5 summarizes the contributions of this paper and highlights exciting future directions.

2. Speech segmentation

The first step in the proposed method is to segment the speech signal into acoustically similar regions. The core idea behind the proposed segmentation algorithm is that the frames belonging to the same segment exhibit higher degrees of similarity than those belonging to different segments. Consider an utterance represented by the sequence of feature vectors $X = (x_1, x_2, ..., x_N)$, where x_i is the d dimensional feature vector and N is the total number of frames. The kernel Gram matrix consisting of similarity between every pair of feature vectors, x_i and x_j, is computed in the Gaussian kernel space as

$$G(i,j) = \exp\left(-\frac{||\mathbf{x}_i - \mathbf{x}_j||^2}{h}\right), \quad 1 \leq i,j \leq N \quad (1)$$

where $||.||$ denotes the Euclidean norm of a vector and h is a free parameter which controls the width of the Gaussian kernel. In this work, we have used 39-dimensional Mel-frequency cepstral coefficients as features to compute the Gram matrix G. The feature vectors belonging to the same segment contribute square patches of higher similarity along the principal diagonal of the Gram matrix. The speech segmentation task can be viewed as identifying these square patches in the Gram matrix.

In order to identify the segment boundaries from the Gram matrix, we define a temporal neighbourhood criterion, similar to the neighbourhood in DBSCAN algorithm [21]. The ϵ-neighbourhood for the i^{th} frame $\mathbf{x}_i$ is defined as the set of all those frames in the utterance whose distance to $\mathbf{x}_i$ is less than a predefined threshold ϵ. Since all those frames in the segment containing $\mathbf{x}_i$ should be acoustically similar to $\mathbf{x}_i$, they will also be a part of ϵ-neighbourhood of $\mathbf{x}_i$. The set of consecutive frames that immediately follow $\mathbf{x}_i$ and falling in the ϵ-neighbourhood of $\mathbf{x}_i$ are referred to as temporally reachable frames from $\mathbf{x}_i$. The boundary of the segment containing $\mathbf{x}_i$ is located by identifying the first temporally unreachable frame from $\mathbf{x}_i$. Locating the boundary from just one point may lead to too many spurious boundaries. Hence, we use K-step unreachability criterion, i.e., K consecutive points being unreachable from $\mathbf{x}_i$, for detecting the boundaries. A smaller value of K leads to false alarms and a large value of K leads to missed detections. In this work, we have used a value of $K = 3$ for detecting the boundary locations. The search space for the boundary locations is restricted to minimum and maximum possible acoustic segment lengths of 20 ms and 500 ms, respectively.

Each frame predicts an end point and all the frames belonging to the same segment will predict the same or nearby frames as end points. For each frame, we count the number of frames that predicted it as the segment end point. The frames with a higher count than their adjacent frames are the eventual end points.

The proposed method depends critically on the choice of ϵ used to define the temporal neighbourhood. Typically voiced segments exhibit higher similarity than unvoiced segments. Hence we need an adaptive ϵ to define the neighbourhood, depending on the acoustic properties of the segment. In order to overcome this issue, we have chosen ϵ to be the running mean of the segment. Once a boundary is detected, the ϵ is reset based on the acoustic characteristics of the next segment.

3. Segment labelling using graph clustering

The speech data is segmented into a large number of varying length acoustic segments. The next step is to cluster the acoustic segments into a finite number of groups depending on their similarity. In this work, we have used a graph theoretic approach to cluster the acoustic segments. For this purpose, an unweighted graph is formed with the acoustic segments as the vertices of the graph, and similarity between pairs of segments as the edge weights connecting them. Since the acoustic segments are of varying length, dynamic time warping is used to define the similarity between them. In this approach, a pair of varying length segments is first aligned using DTW to arrive at fixed length representation. The similarity between the segments is computed as the sum of cosine similarities between corresponding frames in the aligned segments.

Once the graph is formed, we used an infinite range spin glass based community detection algorithm [22] to cluster it. One of the issues associated with this approach is that the size of the edge matrix grows quadratically with the number of acoustic segments. Clustering large graph is computationally very expensive and not feasible on moderate computational resources. In order to address this issue, we first cluster a graph of moderate size and grow it incrementally as and when new data is available. Given a clustered graph and a new acoustic segment, the average connectedness of the segment to every cluster is computed. Average connectedness between a cluster and a segment is defined as the mean similarity of the segment to the every segment in the cluster. The new segment is assigned to the cluster with the highest average connectedness. This approach reduces the computational cost to a great extent and allows clustering in an online manner. Once all the data is incrementally clustered using the proposed graph growing approach, each acoustic seg-

ment is labelled with the corresponding cluster index. The sequence of cluster indices, assigned to the acoustic segments in utterance, are treated as the sequence of virtual phoneme labels. The virtual phoneme labels are used for unsupervised acoustic modelling of the speech data.

4. Unsupervised acoustic modelling

We have used DNNs for unsupervised acoustic modeling from a large corpus of speech signals and their corresponding virtual phoneme transcriptions (cluster indices). In this work, we have used 50 virtual phone labels to transcribe the speech data. Each virtual phoneme is modelled as a 3-state continuous density hidden Markov model. The model parameters are estimated using Baum-Welch embedded reestimation from the virtual phone labels. The trained HMM models are used to force align the virtual phonemes to refine the boundaries obtained from the segmentation step. The state level alignments obtained from the HMM modelling are used as targets to train a DNN classifier. DNN, being a discriminative model, provides better estimates for emission probabilities of the HMM. We used a 6-layer DNN, with 1024 rectified linear units in each layer, to estimate posterior probabilities of the virtual phoneme states from the acoustic input. The input to the DNN is 39-dimensional (13 MFCCs + deltas + deltas deltas) MFCC features with 7-frame context window.

The trained DNN is capable of generating the virtual phoneme state posteriors, which can be used as a representative of speech specific information in the speech signal (ABX task). The state posterior features should, in principle, be speaker-invariant as the speaker-specific information is marginalized during the clustering stage. The DNN, in combination with the HMM, is used to decode the best possible virtual phoneme sequence for a given speech signal.

4.1. Experimental evaluation

The proposed unsupervised acoustic modelling can be used to obtain either a continuous representation of the speech signal in terms of (virtual phoneme) state posteriors or alternatively a discrete representation in terms of a sequence of virtual phonemes. Both these representations are useful in several applications, including speech recognition, spoken term detection, and speech summarization. The effectiveness of the continuous and discrete representations of the speech signal, obtained using the proposed method, is illustrated on the ABX and STD tasks, respectively. The performance of the proposed method on these two tasks is evaluated on the zero speech challenge 2015 dataset. This dataset consists of 10.5 hours of casual conversations in American English, and 5 hours of read speech in Xitsonga. Evaluation kit for both ABX and STD tasks were provided as part of the challenge.

4.1.1. Evaluation on ABX task

The objective of this task is to construct a representation of speech which is robust to within and across talker variation and supports word identification. The metric we will use is the ABX discriminability between phonemic minimal pairs [23, 18]. The ABX discriminability between the minimal pair ”beg” and ”bag” is defined as the probability that A and X are further apart than B and X, where A and X are tokens of ”beg”, and B a token of ”bag” (or vice versa), distance being defined as the DTW divergence of the representations of the tokens. We have used the virtual phoneme state posteriors as features to compute the DTW distance between the tokens. The average percentage error obtained on the ABX tasks for both the databases is reported in Table 1. The effects of within and across-talker variations are evaluated separately to assess the speaker-invariant nature of a feature representation. The performance of the proposed approach (DNN posteriors) is significantly better than the baseline system based on raw MFCC features and speaker adapted features, illustrating the speaker invariant nature of virtual phoneme state posteriors. The performance of the proposed features is better than the bottleneck features obtained from an autoencoder trained on MFCC features [10], illustrating the potential of supervised training of DNN. Notice that the labels for this supervised training are indeed obtained from unsupervised graph clustering. Though the performances of DPGMM [24] and ScatABNET [25] are superior to the proposed method, they cannot be used to obtain a discrete representation for speech signal, which is required in STD task.

Table 1: *Error rates on ABX tasks Zerospeech 2015 dataset. Topline performance is obtained with supervised posteriors.*

Model	English		Tsonga	
	within	across	within	across
Baseline (MFCC)	15.6	28.1	19.1	33.8
Proposed Approach	13.4	21.9	13.2	23.0
Autoencoder [10]	19.7	28.7	17.1	26.4
MFCC + VTLN [26]	14.6	24.0		
DPGMM [24]	10.8	16.3	9.6	17.2
Shallow ScatABnet [25]	11.0	17.0	12.0	15.8
Topline (Supervised)	12.1	16.0	3.5	4.5

4.1.2. Evaluation on STD task

The aim of the STD task is the unsupervised discovery of ”words” defined as recurring speech fragments. The systems should take raw speech as input and output a list of speech fragments (timestamps referring to the original audio file) together with a discrete label for category membership. The evaluation will use the suite of F-score metrics described in [19], which enables detailed assessment of the different components of a spoken term discovery pipeline (matching, clustering, segmentation, parsing) and so will support a direct comparison with the unsupervised word segmentation models. We propose to identify the repeating word like patterns from the sequence of virtual phonemes decoded from the speech signal. The choice of the length of the word plays an important role in searching for words from discrete symbols. Some of the existing methods search for fixed length patterns, like triphones [28] or pentaphones [16], as word candidates. However, words can be composed of varying length phoneme sequences, and the fixed length assumption does not suit for all possible words.

The task of word-level units from sequences of virtual phones in speech is analogous to word discovery from ”space removed” textual data. Pitman-Yor models have been successfully applied to discovering the words from continuous stream alphabets without any space[30]. The current situation is exactly similar to this except that the textual alphabets are replaced with virtual phonemes. Hence, we have used a hierarchical Chinese Restaurant Process [31] based nested Pitman Yor language models to automatically detect the varying length words from the sequence of virtual phonemes. The basic idea behind this method is that the word lengths, typically, follow a Poisson distribution [32]. The parameters of the Poisson distribution are

Table 2: *Results (in percentage) for STD task on Zerospeech 2015 datasets: English and Xitsonga (in brackets). The best scores for each evaluation metric are highlighted in bold. Topline performance is obtained with manual labels.*

System	NLP		type			token			boundary		
	NED	Cov	P	R	F	P	R	F	P	R	F
Baseline[27]	**21.9** (**12**)	16.3 (16.2)	6.2 (3.2)	1.9 (1.4)	2.9 (2.0)	5.5 (2.6)	0.4 (0.5)	0.8 (0.8)	44.1 (22.3)	4.7 (5.6)	8.6 (8.9)
Vseg[28]	89.6 (78.4)	40.6 (77.7)	13.5 (1.7)	11.3 (4.1)	12.3 (2.4)	21.6 (1.8)	4.8 (1.8)	7.9 (1.8)	76.1 (26.2)	28.5 (26.3)	41.4 (26.3)
EnvMin[28]	88 (61.2)	42.2 (95)	12.7 (1.1)	10.8 (3.3)	11.6 (1.7)	21.6 (0.8)	4.7 (1.3)	7.8 (1.0)	75.7 (16.3)	27.4 (24.4)	40.3 (19.5)
Osc[28]	70.8 (63.1)	42.4 (94.7)	**14.1** (2.2)	12.9 (6.2)	**13.5** (3.3)	**22.6** (2.3)	6.1 (3.4)	9.6 (2.7)	75.7 (29.2)	33.7 (39.4)	46.7 (33.5)
CC-PLP[29]	77.3 (36.1)	25.5 (30.2)	4.7 (3.0)	2.5 (2.7)	3.3 (2.8)	4.2 (2.0)	0.6 (0.9)	1.0 (1.2)	39.6 (19.4)	7.5 (11.2)	12.7 (14.2)
CC-FDPLS[29]	61.2 (43.2)	80.2 (89.4)	3.1 (**4.9**)	9.2 (**18.8**)	4.6 (**7.8**)	2.4 (2.2)	3.5 (**12.6**)	2.8 (3.8)	35.4 (18.8)	38.5 (64)	36.9 (29)
proposed	85.0 (66)	**100** (**95.8**)	5.4 (2.3)	**24.8** (8.0)	8.9 (3.6)	7.9 (**2.7**)	**13.9** (8.5)	**10.1** (**4.1**)	41.2 (22.5)	**71.1** (**74.8**)	**52.2** (**34.6**)
proposed 2	91.3 (80)	5.1 (4.9)							**81.4** (**61.2**)	15.7 (27.8)	26.2 (38.2)
Topline (supervised)	0 (0)	100 (100)	50.3 (15.1)	56.2 (18.1)	53.1 (16.5)	68.2 (34.1)	60.8 (49.7)	64.3 (40.4)	88.4 (66.6)	86.7 (91.9)	87.5 (77.2)

estimated from the continuous stream of virtual phoneme sequence. It also has a provision to adaptively build a word level language model from the discovered words that can be used to predict the following words from previous ones. Theoretical results show that it is capable of learning infinite order language model to achieve better segmentation performance.

The performance of the STD task is benchmarked against several well established unsupervised term discovery metrics [19]. Normalized edit distance (NED) measures the variability among the phoneme sequences of a word class, while coverage (Cov) measures the portion of the phoneme sequences covered in the discovered word units. Other evaluation metrics include token recall, type, and boundary. The token recall is the probability that a gold word (manual word transcription) token is found in obtained word classes. Token precision is the probability that an obtained word token would match a gold word token. A similar definition is used for calculation of type performance. Finally, the segmentation measures the accuracy of boundaries of discovered phoneme classes with respect to actual word boundaries.

4.1.3. STD Results

The performance of the proposed method on STD task is given in Table 2, for both English and Xitsonga. The most prominent finding is a full coverage segmentation algorithm with very high word segmentation accuracy on both languages. In English, found patterns cover the entirety of the speech data with 41.2% of the found boundaries matching a true boundary. It finds 71% of the existing boundaries in the data. Similar performance is observed for Tsonga, found patterns cover 96% of the data and locate 75% of the boundaries. The baseline system and other STD system achieve better precision. The high precision might be due to very selective nature of the systems (less coverage). So, we performed additional experiments to allow only high-quality patterns (proposed 2). We used a minimum similarity threshold while growing the graph. The precision of the new system increased as expected and is now almost twice the previous value. The precision reaches the topline precision. There is a decrease in recall performance. The overall boundary performance (F-score) is still better than the baseline and other STD algorithms. Precision and recall can be traded for each other depending on the application in hand. Higher type and token performance demonstrate the quality of obtained word units using the proposed algorithm. Overall, our algorithm achieves the best performance in the highest number of evaluation metrics on both the languages.

5. Conclusions and future work

The current results show that unsupervised labelling of speech data can be a very good start point for targeting zero speech applications. Both the tasks focus on specific information extraction from speech data and use targeted metrics for evaluating performance. Good performance on both the tasks implies that the proposed system preserves all round useful information. It makes the proposed algorithm an ideal candidate for pre-processing acoustic signals for zero resource speech applications. Combining more information can help in the development of speech applications. After labelling, task specific fine tuning can be used for improving performance. Results demonstrated that the neighbourhood density information can be used for segmenting the speech utterances in an unsupervised setting. Phonetic segmentation which in turn can give strong cues for word segmentation. A drawback with current strategy is that the performance of PY segmentation depends heavily on the labels obtained. Incorrect labels will lead to incorrect subwords/words which dampen the consistency of word labels.

ABnet and ScatAbnet can be trained on extracted segments using the proposed algorithm. Since our proposed system is on par with STD system used previously for training, it will improve the performance obtained on ABX task. Detailed analysis with different input features and with different clustering techniques as done in [29], can be used for finding the best performing system combination.

6. References

[1] M. Athineos and D. P. Ellis, "Frequency-domain linear prediction for temporal features," in *Automatic Speech Recognition and Understanding, 2003. ASRU'03. 2003 IEEE Workshop on*, pp. 261–266, IEEE, 2003.

[2] R. M. Hegde, H. A. Murthy, and V. R. R. Gadde, "Significance of the modified group delay feature in speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 190–202, 2007.

[3] Y. Qiao, N. Shimomura, and N. Minematsu, "Unsupervised optimal phoneme segmentation: Objectives, algorithm and comparisons," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pp. 3989–3992, IEEE, 2008.

[4] A. Tsiartas, P. K. Ghosh, P. Georgiou, and S. Narayanan, "Robust word boundary detection in spontaneous speech using acoustic and lexical cues," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pp. 4785–4788, IEEE, 2009.

[5] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[6] Y. Zhang and J. R. Glass, "Unsupervised spoken keyword spotting via segmental dtw on gaussian posteriorgrams," in *Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on*, pp. 398–403, IEEE, 2009.

[7] M. Versteegh, R. Thiolliere, T. Schatz, X.-N. Cao, X. Anguera, A. Jansen, and E. Dupoux, "The zero resource speech challenge 2015.," in *INTERSPEECH*, pp. 3169–3173, 2015.

[8] H. Bourlard and Y. Kamp, "Auto-association by multilayer perceptrons and singular value decomposition," *Biological cybernetics*, vol. 59, no. 4, pp. 291–294, 1988.

[9] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[10] L. Badino, A. Mereta, and L. Rosasco, "Discovering discrete subword units with binarized autoencoders and hidden-markov-model encoders.," in *INTERSPEECH*, pp. 3174–3178, 2015.

[11] D. Renshaw, H. Kamper, A. Jansen, and S. Goldwater, "A comparison of neural network methods for unsupervised representation learning on the zero resource speech challenge.," in *INTERSPEECH*, pp. 3199–3203, 2015.

[12] L. Badino, C. Canevari, L. Fadiga, and G. Metta, "An auto-encoder based approach to unsupervised learning of subword units," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pp. 7634–7638, IEEE, 2014.

[13] M. Huijbregts, M. McLaren, and D. Van Leeuwen, "Unsupervised acoustic sub-word unit detection for query-by-example spoken term detection," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pp. 4436–4439, IEEE, 2011.

[14] C.-y. Lee and J. Glass, "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*, pp. 40–49, Association for Computational Linguistics, 2012.

[15] B. Varadarajan, S. Khudanpur, and E. Dupoux, "Unsupervised learning of acoustic sub-word units," in *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, pp. 165–168, Association for Computational Linguistics, 2008.

[16] M.-h. Siu, H. Gish, A. Chan, W. Belfield, and S. Lowe, "Unsupervised training of an hmm-based self-organizing unit recognizer with applications to topic classification and keyword discovery," *Computer Speech & Language*, vol. 28, no. 1, pp. 210–223, 2014.

[17] T. Schatz, V. Peddinti, F. Bach, A. Jansen, H. Hermansky, and E. Dupoux, "Evaluating speech features with the minimal-pair abx task: Analysis of the classical mfc/plp pipeline," in *INTERSPEECH 2013: 14th Annual Conference of the International Speech Communication Association*, pp. 1–5, 2013.

[18] T. Schatz, V. Peddinti, X.-N. Cao, F. R. Bach, H. Hermansky, and E. Dupoux, "Evaluating speech features with the minimal-pair abx task (ii): resistance to noise.," in *INTERSPEECH*, pp. 915–919, 2014.

[19] 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.

[20] M. Versteegh, X. Anguera, A. Jansen, and E. Dupoux, "The zero resource speech challenge 2015: Proposed approaches and results," *Procedia Computer Science*, vol. 81, pp. 67–72, 2016.

[21] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise.," in *Kdd*, vol. 96, pp. 226–231, 1996.

[22] J. Reichardt and S. Bornholdt, "Statistical mechanics of community detection," *Physical Review E*, vol. 74, no. 1, p. 016110, 2006.

[23] Y. Zhang and J. R. Glass, "Towards multi-speaker unsupervised speech pattern discovery," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pp. 4366–4369, IEEE, 2010.

[24] 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 *INTERSPEECH*, pp. 3189–3193, 2015.

[25] N. Zeghidour, G. Synnaeve, M. Versteegh, and E. Dupoux, "A deep scattering spectrumdeep siamese network pipeline for unsupervised acoustic modeling," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pp. 4965–4969, IEEE, 2016.

[26] P. Baljekar, S. Sitaram, P. K. Muthukumar, and A. W. Black, "Using articulatory features and inferred phonological segments in zero resource speech processing.," in *INTERSPEECH*, pp. 3194–3198, 2015.

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

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

[29] V. Lyzinski, G. Sell, and A. Jansen, "An evaluation of graph clustering methods for unsupervised term discovery.," in *INTERSPEECH*, pp. 3209–3213, 2015.

[30] D. Mochihashi, T. Yamada, and N. Ueda, "Bayesian unsupervised word segmentation with nested pitman-yor language modeling," in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, pp. 100–108, Association for Computational Linguistics, 2009.

[31] D. J. Aldous, "Exchangeability and related topics," in *École d'Été de Probabilités de Saint-Flour XIII1983*, pp. 1–198, Springer, 1985.

[32] V. V. Kromer, "About word length distribution," in *Contributions to the Science of Text and Language*, pp. 199–210, Springer, 2007.