INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Unsupervised Acoustic Segmentation and Clustering using Siamese Network Embeddings

Saurabhchand Bhati[†], *Shekhar Nayak*[‡], *K. Sri Rama Murty*[‡], *Najim Dehak*[†]

[†] Center for Language and Speech Processing, The Johns Hopkins University, USA
[‡]Department of Electrical Engineering, IIT Hyderabad, India

sbhati1@jhu.edu, ee13p1008@iith.ac.in, ksrm@iith.ac.in, ndehak3@jhu.edu

Abstract

Unsupervised discovery of acoustic units from the raw speech signal forms the core objective of zero-resource speech processing. It involves identifying the acoustic segment boundaries and consistently assigning unique labels to acoustically similar segments. In this work, the possible candidates for segment boundaries are identified in an unsupervised manner from the kernel Gram matrix computed from the Mel-frequency cepstral coefficients (MFCC). These segment boundary candidates are used to train a siamese network, that is intended to learn embeddings that minimize intrasegment distances and maximize the intersegment distances. The siamese embeddings capture phonetic information from longer contexts of the speech signal and enhance the intersegment discriminability. These properties make the siamese embeddings better suited for acoustic segmentation and clustering than the raw MFCC features. The Gram matrix computed from the siamese embeddings provides unambiguous evidence for boundary locations. The initial candidate boundaries are refined using this evidence, and siamese embeddings are extracted for the new acoustic segments. A graph growing approach is used to cluster the siamese embeddings, and a unique label is assigned to acoustically similar segments. The performance of the proposed method for acoustic segmentation and clustering is evaluated on Zero Resource 2017 database.

Index Terms: Zero resource speech processing, Representation learning, Spoken term discovery, Siamese network

1. Introduction

The advent of deep learning techniques, together with improved computational resources, has lead to significant improvements in the performance of automatic speech recognition (ASR) systems [1]. Most of the state-of-the-art ASR systems require thousands of hours of manually transcribed speech data [2, 3], lexicon and pronunciation dictionary [4]. The recent technological advances in ASR systems cannot be applied to under-resourced languages, for example, regional languages, for which transcribed speech data are not readily available. Hence, there is a pressing need to explore alternate methods for developing speech interfaces for under-resourced languages.

Such approaches commonly referred to as zero-resource speech processing [5,6] has several applications, which include, but not limited to, building speech interfaces in under-resourced languages, preserving endangered languages. At the core of the issues, in zero-resource speech processing, lies unsupervised discovery of linguistic units from the raw speech waveform [7–16]. Linguistic unit discovery, in turn, involves segmenting the speech waveform into acoustically homogeneous regions and consistently assigning unique labels to the segments with similar acoustic properties. Hence, the representation of the speech signal plays a vital role in the unsupervised discovery of linguistic units from the speech signal.

The acoustic speech waveform carries information about several sources, including linguistic units, speaker, emotion, etc. The traditional features, extracted from the magnitude spectral envelope of the speech signal, capture information about all these sources. For example, Mel-frequency cepstral coefficients (MFCC) have been widely used for speech recognition systems [17], language identification [18], and emotion recognition [19], as MFCC capture information about all these sources. In supervised ASR, a powerful classifier, like deep neural network (DNN), learns a nonlinear map between the MFCC features and the manual transcriptions [1]. Hence an ASR system can efficiently recognize the linguistic units from the MFCC features, even though they carry other information as well. However, in the zero-resource scenario, manual transcriptions are not available to guide the model in selecting only the relevant linguistic information. Hence, it is important to use improved features for zero resource applications as opposed to generic features like MFCC.

In this paper, we propose to extract features that highlight the speech-specific characteristics required for the application like unsupervised spoken term discovery. We propose full-coverage segmentation and clustering, where entire data is segmented and labeled. Full coverage system would allow development of speech indexing and query-by-example [20] search system in an unsupervised manner.

The extracted features from the Gaussian mixture model (GMM) or autoencoder are expected to occupy orthogonal subspaces for different speech units which helps in achieving better inter-phone discrimination [21, 22]. Both GMMs and autoencoders, however, do not model the time-sequence in which the feature vectors evolve. As a consequence, the new representation may vary even within a phoneme segment, which subsequently may lead to ambiguity at the clustering stage. Here, we learn unit-level features that are consistent with the segment properties as opposed to frame-level features.

Sequence-aware representation learning methods [23, 24] require initial segmentation and labeling information. Given preliminary segments and labels, correspondence autoencoder(CAE) [23], ABnet [24] learns feature representations that minimize distances among different instances of the same label. The learned features from CAE, ABnet outperform the MFCC features on both ABX and spoken term discovery (STD) tasks [23]. Both the CAE and the ABnet require segmentation and label information for extracting speaker-independent representation. Features learned after the clustering step combines the errors of both the segmentation and clustering steps. Usually, the segmentation step performs better than clustering step. So, we move the feature learning step before clustering and learn features directly from the segmentation information and

http://dx.doi.org/10.21437/Interspeech.2019-2981

use the improved features for clustering the acoustic segments. We propose a two-stage robust method for the segmentation of speech signal into acoustically homogeneous regions. In the proposed method, initial segmentation is obtained using a kernel Gram segmentation method [25]. It uses the block diagonal structure of the kernel Gram matrix to identify the segments in the speech signal. Next, we use a Siamese network [26] to learn representations that enhance the block diagonal structure. The frames from adjacent segments are used as mismatched pair examples. The frames within a segment are used as matched-pair examples. The learned embeddings are used for recomputing the Gram matrix and re-segmenting the speech signal. Since the Siamese network is trained on segments taken from various speakers, it can produce speaker-independent embeddings for the acoustic segments. We use the learned embeddings for clustering the discovered segments. The performance of the proposed for acoustic segmentation and clustering is evaluated on Zero Resource 2017 challenge dataset [6]. The Zero Resource 2017 dataset contains more than 100 hours of unlabelled data, from English, French, Mandarin and two surprise languages. Our proposed system outperforms existing unsupervised term discovery systems based on PLP and MFCC features.

2. Two-Stage Speech Segmentation

In this work, we propose a two-stage speech segmentation method to identify the boundaries of acoustically homogeneous regions. In the first step, kernel Gram segmentation method [25] is used to identify the candidate boundaries of the acoustic segments. This algorithm works on the assumption that the frames within a segment exhibit a higher degree of similarity than the frames across the segments. The similarity between two frames $\mathbf{x_i}$ and $\mathbf{x_j}$ is computed as

$$G(i,j) = \exp\left(\frac{-\|\mathbf{x_i} - \mathbf{x_j}\|}{h}\right), \quad 1 \leq i,j \leq N \qquad (1)$$

where $\|.\|$ denotes the Euclidean norm of a vector and h is the width of the Gaussian kernel. The frames within the same segment exhibit higher degree of similarity, and hence, they induce block diagonal structure in the kernel Gram matrix, as shown in Fig. 1(a). For the algorithm hypothesizes a frame as a candidate for the boundary if the next K consecutive frames have lower similarity than an adaptive threshold ϵ. A smaller value of K results in noisy segment boundaries (false alarms), and a larger K results in missed boundaries. The adaptive threshold ϵ is decided based on the running mean of the current segment so that it adapts itself based on the acoustic properties of the segment and it is reset after every segment boundary. We use a value of $K = 4$ for boundary detection. The minimum and maximum possible acoustic segment lengths are restricted to 20 ms and 500 ms, respectively [25].

2.1. Boundary refinement using siamese embeddings

The block diagonal structure of the kernel Gram matrix is the basis of the segmentation algorithm. We use a deep siamese network to enhance the block diagonal structure of the kernel Gram matrix which in turn improves the quality of the discovered segment boundaries. Siamese network can be trained to extract latent embeddings that improve the contrast between intrasegment to intersegment distances.

Siamese network training requires matched and mismatched pairs of data. The candidate boundaries obtained from the kernel-Gram segmentation were used to prepare the matched and mismatched pairs for training the siamese network. Pairs of frames from the same segment are considered as matched pairs, while pairs of frames from adjacent segments are considered as mismatched pairs. The number of training examples from matched and mismatched pairs are maintained in a similar range to avoid bias during training. The frames near the boundaries are not used in the training as they cause more confusion during the training.

Two identical copies of the same network project the input frames to an embedded space where a loss function is used to quantify the similarity between the pairs. Ideally, we want to use the cosine similarity as the objective. However, the cosine distance is maximum when the embeddings are collinear and it is minimal when they are anti-collinear. This forces the DNN to project the mismatched to be anti-collinear which is harder to achieve than orthogonality in high dimensional spaces [27]. The loss function is modified to be asymmetric for matched and mismatched data points. [24, 27]. The asymmetric loss between a pair of examples A and B is defined in terms of their siamese embeddings $\mathbf{Y}_A$ and $\mathbf{Y}_B$, as

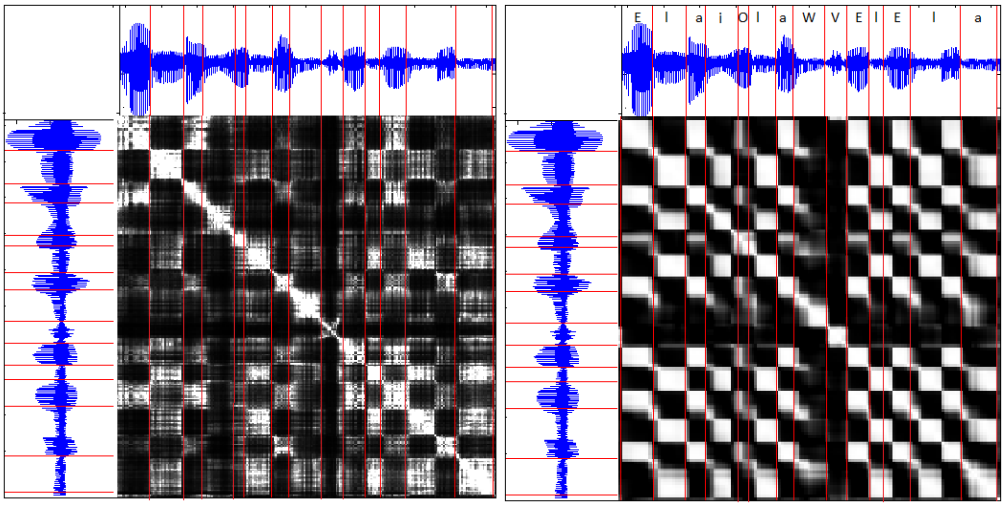
$$J(A,B) = \begin{cases} (1 - \cos(\mathbf{Y}_A, \mathbf{Y}_B))/2, & A\&B \in \text{Matched} \\ \cos^2(\mathbf{Y}_A, \mathbf{Y}_B), & A\&B \in \text{Mismatched} \end{cases}$$

where $\cos(\mathbf{x}, \mathbf{y})$ denotes the normalized inner product between the vectors $\mathbf{x}$ and $\mathbf{y}$. This loss function is minimized when the embeddings are collinear for matching frames and orthogonal for mismatching frames. The network is initialized using a glorot uniform initializer, and 70% of the data is used as training set, while the remaining 30% is used as the validation set. The Siamese network consists of two hidden layers with 500 hidden neurons each with a dropout of 0.5 and a bottleneck layer of with 20 neurons. A seven-frame context window, i.e., three frames on either side, is used to provide more temporal information to the network. The embeddings extracted from the penultimate layer of the trained siamese network are used as representative features for segmentation and clustering. In this work, the cosine angle between the siamese embeddings is used to quantify the similarity between the frames.

Fig. 1(b) shows the kernel-Gram matrix computed from the siamese embeddings of a speech signal. The square patches along the diagonal are well defined in the kernel Gram matrix computed from the siamese embeddings in Fig. 1(b), compared to the one computed from raw MFCC features in Fig. 1(a). This could be attributed to the fact that the MFCC kernel Gram matrix is computed from a predetermined kernel, whereas the siamese kernel Gram matrix is obtained from the data-driven kernel. Moreover, the kernel Gram matrix computed from MFCC features relies only on local information whereas one computed from the siamese embeddings utilizes contextual information from nearby frames. With a clearer transition between the segments, the boundary detection becomes easier, leading to improved segmentation performance.

3. Segment labeling using graph growing and word discovery

The segmentation step divides the speech data into a large number of varying length segments. The next step is to cluster the segments into acoustically similar clusters. Here, we use a graph-theoretic approach for clustering the acoustic segments. Graph clustering methods can capture the non-linear structure of the data and have achieved state-of-the-art performance on many clustering tasks [28, 29]. To achieve that, an undirected



(a) (b)

Figure 1: *(a) Kernel-Gram matrix computed from the MFCC features. (b) Self-similarity matrix computed from the siamese embeddings. The red lines indicate the manual phoneme boundaries.*

weighted graph is formed with acoustic segments as vertices and the similarity between the pairs of segments as the edge weights. There are two important issues that need to be addressed in the graph clustering approach:

i) Comparing varying length segments: To quantify the similarity (the edge weights) between two varying length segments, we need to extract fixed-dimensional representation from the varying length segments. In this work, we use a downsampling strategy to extract fixed dimensional representations from the varying length acoustic segments [30, 31]. Simply taking the average of the segments ignores the temporal information. So, each acoustic segment is divided into 10 uniform subsegments, and their means are appended to obtain a fixed dimensional representation for that segment. While this method yields fixed dimensional representations, they are typically high dimensional. For example, if 39-dimensional MFCC features are used to represent each frame of the speech signal, it results in a 390-dimensional feature vector for each segment. The siamese network embeddings can be used to arrive at a compact representation for the acoustic segments. Here, 20-dimensional siamese embeddings are used to represent each frame of the speech signal. As a result each acoustic segment, divided into 10 uniform subsegments, can now be represented using a 200-dimensional vector. Hence, the siamese embeddings help in reducing the computational complexity of graph formation.

ii) Size of the graph: The number of edges in the graph grow quadratically with the number of phone segments. For large datasets, the number of phone segments can reach up to several million. Clustering such large graphs is computationally very expensive. We address this issue by using a graph growing approach for clustering. In graph growing, a small graph of moderate size (typically order of ten thousand nodes) is clustered first and then it is grown incrementally. We use a spin glass based community detection method [32] for clustering the small graph. Given a small clustered graph and a new input segment, we compute the average connectedness between the segment and every cluster and assign the input segment to the cluster with the lowest score. The average connectedness between a segment and a cluster is defined as the mean similarity between the segments in the cluster and the segments. This procedure reduces the computational cost of the clustering step significantly. Since the labels are assigned in an incremental manner, a dataset of any size can be clustered.

Both these modifications, make the proposed algorithm scalable to large datasets which we demonstrate, here, by applying it to Zero Resource 2017 database (100+ hours of speech data). We refer to the resulting cluster indices of the graph as virtual phones, and any new speech signal can be represented as a discrete sequence of virtual phones.

3.1. Greedy bottom-up discovery of words

One of the challenging tasks in unsupervised speech processing is to discover the words "recurring speech fragments" from the untranscribed speech data. In our framework, words would correspond to repeating sequences of virtual phones. A bottom-up greedy mapping is used for discovering the words from sequences of virtual phones. First, we find all the longest recurring n-grams that occur at least twice in the data. Then, we locate the n-grams of the next highest order, and virtual phones that were part of the already discovered n-grams are excluded from this step. This process is repeated all the way down to unigrams, and all the remaining unigrams are included as words. The greedy nature of the algorithm makes it computationally efficient making word discovery feasible even for large datasets. We empirically use tri-grams as the possible words. Similar approach have successfully been used for word discovery [33].

4. Experimental Evaluation

We use Zero Resource speech 2017 challenge dataset [6] for evaluating the performance of the proposed approach. The challenge aims to measure the robustness of the unsupervised term discovery systems across speakers and languages. The 2017 challenge dataset consists of more than 100 hours of speech data distributed across 5 languages. The evaluation kit uses various well-established metrics [34] to quantify the system performance. All the metrics assume the availability of a time-aligned transcription of the speech data. Normalized edit distance (NED) measures the differences in the phoneme sequences of a word class, while the coverage (Cov) measures the fraction of

Table 1: *Performance of the baseline system and proposed siamese segmentation system on Zero Resource Speech 2017 Challenge*

Language	System	NLP		type			token			boundary		
		NED	Cov	P	R	F	P	R	F	P	R	F
English (45 hours)	Baseline [7]	**30.7**	2.9	4.5	0.1	0.2	4.0	0.1	0.1	37.5	0.9	1.8
	Syl-seg [15]	34.7	29.6	3.7	3.5	3.6	1.8	2.4	2.1	22.9	24.5	23.7
	ES-KMeans [16]	72.6	100	**8.3**	16.7	**11.1**	**13.0**	14.1	**13.5**	**51.0**	54.4	**52.7**
	Sia-seg	80.2	**100.9**	5.9	**20.5**	9.2	5.8	**18.6**	8.8	33.1	**76.8**	46.2
French (24 hours)	Baseline [7]	25.4	1.6	**6.9**	0.2	0.3	**5.2**	0.1	0.1	30.9	0.6	1.1
	Syl-seg [15]	**24.8**	28.8	4.7	4.0	4.3	2.9	3.5	3.2	24.8	25.6	25.2
	ES-KMeans [16]	67.3	97.2	3.1	6.3	4.2	3.5	3.9	3.7	**37.8**	41.6	39.6
	Sia-seg	78.9	**97.5**	5.2	**15.7**	**7.8**	4.7	**15.7**	**7.2**	31.7	**75.1**	**44.6**
Mandarin (2.5 hours)	Baseline [7]	**30.7**	2.9	4.5	0.1	0.2	4.0	0.1	0.1	**37.5**	0.9	1.8
	Syl-seg [15]	53.4	37.1	3.4	2.6	3.0	1.6	2.7	2.0	22.2	31.3	26.0
	ES-KMeans [16]	88.1	100	2.5	4.1	3.1	2.5	3.4	2.9	36.5	47.1	41.1
	Sia-seg	91.1	**117.5**	**5.6**	**22.0**	**8.9**	**10.4**	**22.0**	**14.1**	32.7	**87.0**	**47.5**
LANG 1 (25 hours)	Baseline [7]	30.5	3.0	5.5	0.3	0.6	4.0	0.1	0.2	28.2	1.2	2.3
	Syl-seg [15]	**29.7**	31.9	**58.5**	**36.0**	**44.4**	2.8	2.8	2.8	1.5	2.6	1.9
	ES-KMeans [16]	66.4	100.0	5.7	11.2	7.5	**10.3**	14.3	**12.0**	**42.6**	56.5	**48.6**
	Sia-Seg	71.5	**100.5**	4.9	15.2	7.4	4.5	**18.3**	7.2	27.9	**80.1**	41.4
LANG 2 (10 hours)	Baseline [7]	33.5	3.2	2.3	0.1	0.2	1.6	0.0	0.1	25.3	1.0	2.0
	Syl-seg [15]	**27.0**	17.4	2.9	1.9	2.3	1.4	1.1	1.2	18.9	14.2	16.2
	ES-KMeans [16]	72.2	100.0	4.6	10.0	6.3	**4.9**	5.2	5.0	**42.4**	44.3	43.3
	Sia-Seg	73.1	**100.7**	**6.3**	**20.1**	**9.6**	4.5	**15.9**	**7.0**	31.4	**76.4**	**44.5**

the data covered by the discovered word like units. The token recall is the probability that a gold word (manual word transcription) token is found in obtained word classes, while the token precision is the probability that a discovered word token would match a gold word token. A similar definition is used for the metric 'type'. It measures the correspondence between the discovered word and the true words in the data. The segmentation measures the quality of the boundaries of the identified word-like units with the manual word boundaries.

Following the guidelines of the challenge, the phonetic segmentation, clustering is done in a speaker and language independent manner, and the same system (same minimum phoneme length, the same number of clusters, etc.) is used for different languages without optimizing any language-specific parameters. As it is evident from the table, the proposed approach consistently achieves better type, token and boundary F-score on all the languages than the baseline system.

The baseline system [7] uses Locality Sensitive Hashing for mapping high dimensional PLP feature vectors into low dimensional bit signatures and then employes segmental DTW for locating matching patterns. The baseline system focuses on finding high-precision isolated segments. To achieve high precision, a lot of the discovered segments are discarded which results in a limited coverage of only 3%. As a result the baseline system performs better in terms of NED, which is computed only on the discovered patterns. On the other hand, the proposed system covers the entire data and therefore achieves better performance in word boundary, token, type and coverage. Although the acoustic segments clustered together by our approach have a poorer match to each other (high NED) as compared to the baseline, the discovered word-like units better match the true words (word type F-score). The proposed system consistently detects better word like units on all the languages, compared to the baseline approach.

The syllable segmentation based method [15] uses a syllable segmentor [33] to obtain syllable boundaries. These syllables are clustered and the cluster indices are used as labels for linear discriminant analysis (LDA) to learn compact representation from MFCC features. These new features are used for obtaining new clustering indices and new LDA based features. This process is repeated until there is no change in separability of the clusters. Uniformly down-sampled version of the final features are then used for detecting matching pairs of syllables. A dynamic time warping based analysis is carried out to refine the discovered words. This process improves the quality of the discovered words but reduces the coverage of the system. Our proposed method performs considerably better in majority of the evaluation metrics. The ES-KMeans algorithm [16] starts with an initial set of boundaries and then iteratively eliminates some of the boundaries to locate frequently occurring longer patterns. The ES-KMeans is a full coverage system and provides an important comparison metric for the unsupervised term discovery algorithms. We significantly outperform[1] the other existing algorithms [7, 15, 16]. Our system not only performs well on the train languages but also on the surprise languages which shows that our system generalizes well across languages and can be used on unseen languages.

5. Conclusions

This paper demonstrates that a good acoustic model can be learned by just using boundary information. The initial boundary candidates are obtained using a kernel Gram segmentation method. The embeddings obtained from the learned acoustic model enhance the block-diagonal structure of the self-similarity matrix and further refine the boundaries. The acoustic segments are represented as fixed dimensional vectors, and the Siamese embeddings capture the acoustic properties of the phonemes which is crucial for clustering the phonemes. The clustering is done in a speaker-independent manner. Our full coverage system finds words that are closer to actual words in the data as evident by word token, type, and boundary F-score.

[1]The complete output of the proposed system can be found here: https://zenodo.org/record/1473792

6. References

[1] 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.

[2] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, *et al.*, "Deep speech: Scaling up end-to-end speech recognition," *arXiv preprint arXiv:1412.5567*, 2014.

[3] 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, pp. 517–520, IEEE, 1992.

[4] A. Stolcke, J. Zheng, W. Wang, and V. Abrash, "Srilm at sixteen: Update and outlook," in *Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop*, vol. 5, 2011.

[5] 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.

[6] E. Dunbar, X. N. Cao, J. Benjumea, J. Karadayi, M. Bernard, L. Besacier, X. Anguera, and E. Dupoux, "The zero resource speech challenge 2017," in *Automatic Speech Recognition and Understanding Workshop (ASRU), 2017 IEEE*, pp. 323–330, IEEE, 2017.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] S. Bhati, S. Nayak, and K. S. R. Murty, "Unsupervised speech signal to symbol transformation for zero resource speech applications," *Proc. Interspeech 2017*, pp. 2133–2137, 2017.

[13] S. Bhati, H. Kamper, and K. S. R. Murty, "Phoneme based embedded segmental k-means for unsupervised term discovery," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5169–5173, April 2018.

[14] H. Kamper, A. Jansen, and S. Goldwater, "A segmental framework for fully-unsupervised large-vocabulary speech recognition," *Computer Speech & Language*, vol. 46, pp. 154–174, 2017.

[15] O. Räsänen and S. Seshadri, "Zs2017 aaltolag submission , https://doi.org/10.5281/zenodo.810808," june 2017.

[16] H. Kamper, K. Livescu, and S. Goldwater, "An embedded segmental k-means model for unsupervised segmentation and clustering of speech," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 719–726, IEEE, 2017.

[17] L. Muda, M. Begam, and I. Elamvazuthi, "Voice recognition algorithms using mel frequency cepstral coefficient (mfcc) and dynamic time warping (dtw) techniques," *arXiv preprint arXiv:1003.4083*, 2010.

[18] S. G. Koolagudi, D. Rastogi, and K. S. Rao, "Identification of language using mel-frequency cepstral coefficients (mfcc)," *Procedia Engineering*, vol. 38, pp. 3391–3398, 2012.

[19] C. S. Ooi, K. P. Seng, L.-M. Ang, and L. W. Chew, "A new approach of audio emotion recognition," *Expert systems with applications*, vol. 41, no. 13, pp. 5858–5869, 2014.

[20] 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.

[21] 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.

[22] 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.

[23] H. Kamper, M. Elsner, A. Jansen, and S. Goldwater, "Unsupervised neural network based feature extraction using weak top-down constraints," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pp. 5818–5822, IEEE, 2015.

[24] R. Thiolliere, E. Dunbar, G. Synnaeve, M. Versteegh, and E. Dupoux, "A hybrid dynamic time warping-deep neural network architecture for unsupervised acoustic modeling.," in *INTERSPEECH*, pp. 3179–3183, 2015.

[25] S. Bhati, S. Nayak, and K. Sri Rama Murty, "Unsupervised segmentation of speech signals using kernel-gram matrices," in *Computer Vision, Pattern Recognition, Image Processing, and Graphics: 6th National Conference, NCVPRIPG 2017, Mandi, India, December 16-19, 2017, Revised Selected Papers 6*, pp. 139–149, Springer, 2018.

[26] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature verification using a" siamese" time delay neural network," in *Advances in Neural Information Processing Systems*, pp. 737–744, 1994.

[27] G. Synnaeve, T. Schatz, and E. Dupoux, "Phonetics embedding learning with side information," in *Spoken Language Technology Workshop (SLT), 2014 IEEE*, pp. 106–111, IEEE, 2014.

[28] S. E. Schaeffer, "Graph clustering," *Computer science review*, vol. 1, no. 1, pp. 27–64, 2007.

[29] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Advances in neural information processing systems*, pp. 849–856, 2002.

[30] K. Levin, A. Jansen, and B. Van Durme, "Segmental acoustic indexing for zero resource keyword search," in *Proc. ICASSP*, 2015.

[31] H. Kamper, W. Wang, and K. Livescu, "Deep convolutional acoustic word embeddings using word-pair side information," in *Proc. ICASSP*, 2016.

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

[33] 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.

[34] 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.