INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

# Self-Paced Pattern Augmentation for Spoken Term Detection in Zero-Resource

*Sudhakar P*[1,3], *Sreenivasa Rao K*[2,3], *Pabitra Mitra*[2,3]

[1]Adavanced Technology Development Centre,
[2]Department of Computer Science and Engineering,
[3]Indian Institute of Technology, Kharagpur, India-721302

sudhakar.asp@iitkgp.ac.in

## Abstract

The spoken term detection task is challenging when a large volume of spoken content is generated without annotation. The pattern discovery approach aims to overcome the challenges by capturing the pattern similarities directly from the representation of the speech signal. A challenge to the pattern discovery task is handling the variabilities in natural speech. In the proposed approach, we aim to overcome the pattern variability challenges in the spoken term similarity region in three stages. At first, the pattern similarities between two spoken terms were captured using our heuristic search, and the pattern variabilities in the similarity region were observed. In the second stage, the observed pattern variabilities were augmented to the Siamese network to learn the relationship. Finally, the learned network is used to identify the matches between spoken query and document. Based on the experimental studies, it is observed that the proposed approach reduces the false alarms by 17.7% and improves the spoken term detection accuracy by 7.1% against the Microsoft Low-Resource Language corpus.

**Index Terms**: spoken term detection, pattern matching, pattern augmentation, Siamese network

## 1. Introduction

Speech signal carries a lot of variabilities that occur naturally due to the speaker, environment and language-specific changes. Due to that, capturing the perfect match between two similar spoken terms becomes challenging. The conventional approach uses Automatic Speech Recogniser (ASR) to convert speech into text. Further, text-based matching was carried out to accomplish the spoken term detection task. However, the ASR-centric approaches demand a large volume of annotated spoken content to accomplish spoken term detection. Moreover, preparing annotations for the spoken content demands time and language expertise. As an effect, spoken content retrieval (SCR) for speech corpora belonging to low-resource languages with minimal or no annotations remains challenging. Pattern discovery is one of the alternate approaches that aims to discover the similarities among spoken terms directly from the acoustic feature representation itself. Such an approach does not seek annotations and accomplishes the task across languages.

The spoken content retrieval task in the absence of annotation was achieved by capturing the similarity pattern between alike spoken terms in an unsupervised way. The studies carried out by [1, 2, 3, 4, 5] reveal that spoken term detection in the absence of resources is viable. The existing approaches for the spoken term detection task were broadly grouped into Dynamic Time Warping (DTW) based techniques and template matching centric techniques. In DTW-centric approaches [2, 6, 7, 8], the similarity between spoken terms was captured by computing the temporal alignment between the acoustic features. Despite feasibility, the challenge in the DTW approach is the global alignment problem. In the DTW approach, the optimal alignment path was computed globally, where the local alignments were discarded due to variability issues. The segmental DTW [3, 9] approach was introduced to overcome the global alignment problem by capturing the similarities at the segmental level. Furthermore, a statistical approach [10] to the spoken term detection task, dynamic programming approach [11] with a bag of grams, randomised algorithm approach [1] and audio motif discovery [12] approach for spoken term detection uses segmental DTW technique to accomplish the similarity matching task.

On the contrary, the template matching techniques achieve the spoken term matches by capturing the similarity pattern between two spoken terms. An n-gram approach [4] with syllable boundaries maps the variable length segments into fixed dimensions, and similarity was captured in the fixed dimensional space. In [5], the similarity patterns were discovered by grouping the fixed-size spoken segments with the k-means clustering approach. Alternatively, deep neural network (DNN) centric approaches [13, 14, 15, 16, 17] use the knowledge of the resourceful languages to capture the spoken term similarities of the low-resource languages. Despite viability, one of the major concerns for both DTW-centric and template-matching approaches is handling the variabilities that arise in natural speech. Due to that, the pattern similarities between similar spoken terms occur in multiple ways and capturing the variabilities becomes challenging. As a result, the pattern discovery task was severely affected by the false alarm candidates during the retrieval. In the proposed approach, we aim to capture the similarities through our heuristic search that accounts for all types of variabilities in the similarity region. Further, the discovered patterns were augmented to the Siamese network defined to learn the similarities from the acoustic feature representation itself. Based on the results, it is observed that the performance of the proposed approach was improved, and false alarm candidates were reduced compared to other approaches.

Further, the article is organised as follows: Section 2 discusses the pattern discovery approach. Section 3 details the spoken term detection network to capture the similarities from the acoustic feature representation. The experimental studies were discussed in Section 4. Section 5 concludes the article with further scope for research.

## 2. Pattern Discovery and Augmentation

In the proposed approach depicted in Fig. 1, the similarity patterns that occur between two spoken terms were discovered and augmented for the spoken term detection task. At the prepro-


10.21437/Interspeech.2023-1215

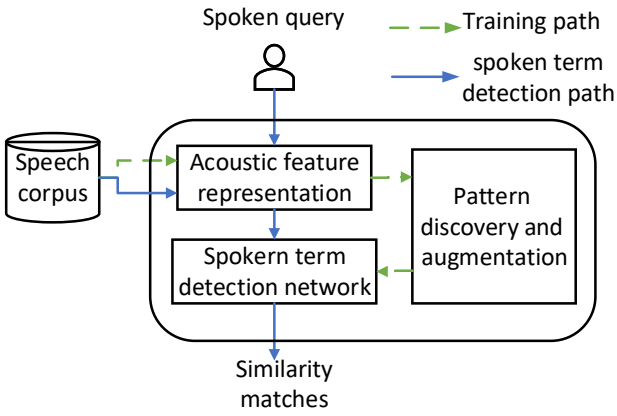


Figure 1: *depicts the overall schematic view of the proposed approach.*

cessing stage, the acoustic feature representation of the speech was obtained by computing the Mel-spectrogram of the speech signal over a 20 ms frame duration without overlap. The framed speech signal was windowed using the hamming window, and Discrete Fourier Transform was applied in the window region. Further, the Mel-filterbank was applied, and the log magnitude of the energy values was obtained in the window region. In our approach, 80 filter banks are used to capture the Mel-spectrogram. Similarly, for all frames, the Mel-spectrogram was computed and retained for the pattern discovery task. After preprocessing, the heuristic pattern match discovery was achieved in a sequence of steps. At first, the similarity between two spoken content was computed to identify the temporal alignment. The acoustic feature representation of the spoken document $D^i$, denoted as $X_L^i = \{x_1, x_2, ..., x_L\}, x \in \mathbb{R}^{80}$ was obtained from the speech signal. Similarly, the acoustic features for the document $D^j$ was computed and denoted as $X_M^j = \{x_1, x_2, ..., x_M\}, x \in \mathbb{R}^{80}$. Further, the cosine similarity between two feature vectors $X_l$ and $X_m$ was computed using Eq. (1)

$$S_c(X_l, X_m) = \frac{X_l . X_m}{||X_l||\,||X_m||} \tag{1}$$

Similarly, for all feature vectors $1 \leq l, m \leq L, M$, the cosine similarity was computed and the similarity matrix $sim[l, m]$ was obtained. In the next step, the matrix was binarised using Eq. (2) to discriminate the similar regions from dissimilar ones. The threshold $\eta$ was obtained empirically by maximising the spoken term match detection.

$$sim[l, m] = \begin{cases} 1, & sim[l, m] \geq \eta \\ 0, & otherwise \end{cases} \tag{2}$$

In the next step, the diagonal pattern similarity was determined based on capturing the diagonal matches that occur in the similarity matrix using Eq. (3). The spoken term similarity between two acoustic feature representations propagates diagonally. Hence, we capture the diagonal pattern similarity.

$$sim[l, m] = \begin{cases} 0, \; if \; l = 0 \; or \; m = 0 \\ sim[l, m] + sim[l-1, m-1] \end{cases} \tag{3}$$

At this point, our approach differs from the DTW-centric approaches. In DTW, it aims to obtain the best alignment between two temporal sequences by computing the optimal path cost. The challenge in cost computation is that it is easily affected by the variabilities that occur in the adjacent locations of the search space. Due to that, the DTW technique escapes from the local alignments. In the proposed approach, instead of finding the path cost, we aim to capture the alignment pattern that occurs in all diagonal regions and compute the similarities by considering the diagonal heuristic cost. Further, in the next step, the diagonal cost was computed for all diagonals $L + M - 1$ in the similarity matrix using Eq. (4)

$$cost_d^{ij}[k] = \sum_{k=1}^{L+M-1} diag(k) \tag{4}$$

where, the $diag(.)$ returns the elements associated in the diagonal. In the next step, the $cost_d^{ij}[k]$ was traced to capture the similar and dissimilar patterns that occur between the documents $D^i$ and $D^j$. The similarity patterns were obtained by Eq. (5) with the match threshold $\mu$ that aims to detect the significant matches that occur between $D^i$ and $D^j$.

$$T[i, j] = \arg\max(cost_d^{ij}[k] > \mu) \tag{5}$$

The match threshold assures the minimal segmental matches that occur between documents. In the same way, dissimilar patterns were captured using Eq. (6)

$$F[i, j] = \arg\min(cost_d^{ij}[k]) \tag{6}$$

Further, in the augmentation task, the acoustic feature representation ($X$) that occurs in the similar ($T$) and dissimilar ($F$) pattern regions were obtained and retained for the spoken term detection network training task. In this approach, the similarity and dissimilarity patterns were discovered with the help of the heuristic cost in a self-supervised fashion, discriminating our approach from supervised learning approaches.

## 3. Spoken Term Detection Network

The objective of the spoken term detection network is to capture the similarities between acoustic feature representations. To achieve this, we defined a Siamese network [18] depicted in Fig. 2 with convolution layers to learn the similarities from the feature representation itself. The network comprises a pair of identical networks consisting of a set of convolutional layers that learn together during the training phase. The learning strategy of the network aims to capture the similarity between two identical features by computing the minimal distance among them. Let $X_1$ and $X_2$ be a pair of acoustic feature representation and $Y \in 0, 1$ representing the match where $Y = 1$ indicates the identical pair and $Y = 0$ indicates the non-identical pair. The Siamese network learns the shared parameters ($\theta$) from the input pairs by finding the optimal value that reduces the distance between identical pairs and increases it for non-identical pairs. The learning parameters were controlled by the contrastive loss function defined in Eq. (7) that aims to obtain the appropriate distance by computing the optimal parameters in the search space.

$$\mathbb{L}(\theta, (Y, X_1, X_2)^i) = (1-Y) + \mathbb{L}_s(dist^i(\theta)) + Y(\mathbb{L}_d(dist^i(\theta))) \tag{7}$$

In Eq. (7), $(Y, X_1, X_2)^i)$ represents the $i^{th}$ data-pairs with label, $\theta$ indicates the learning parameter, $\mathbb{L}_s$ is the loss function

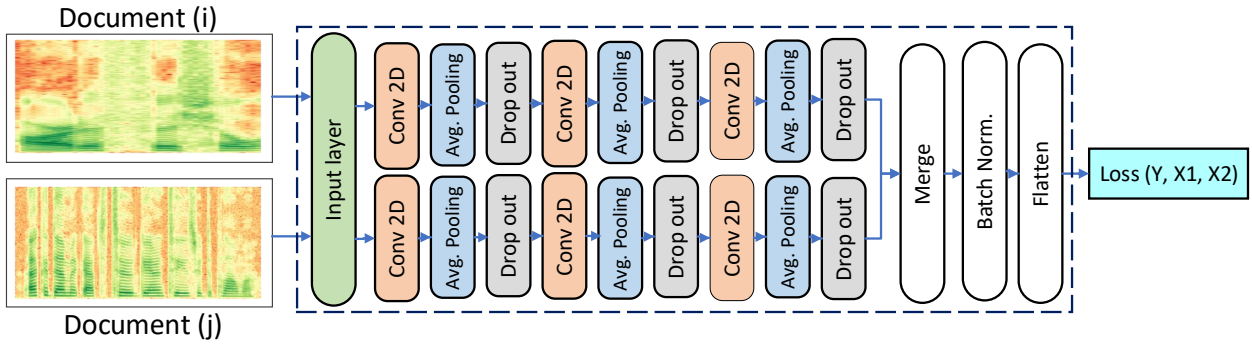


Figure 2: *depicts the Siamese network design for the spoken term match detection task.*

Table 1: *MSLRL speech corpus statistics. # Docs indicate the total number of spoken documents in the corpus. #Queries indicate the number of keyword queries used for evaluation. #Speaker represents the number of speakers who contributed to the specific language.*

| Language | #Docs | #Queries | | #Speakers | Duration (hrs) |
|---|---|---|---|---|---|
| | | Dev | Test | | |
| Gujarati | 1155 | 137 | 271 | 89 | 2.27 |
| Telugu | 603 | 63 | 131 | 53 | 1.02 |
| Tamil | 1032 | 118 | 225 | 84 | 1.55 |
| | | | | Total duration | 4.84 ($\approx$ 5) |

for the similar pair and $\mathbb{L}_d$ indicate the loss function for the dissimilar pair. By training the network with the pair of feature representations augmented from the pattern discovery (see Section 2) task, it is viable to learn the pattern similarities from the acoustic feature representation itself.

## 4. Experiments and Results Discussion

### 4.1. Corpus

The performance of the spoken term detection task was evaluated using the Mircosoft Low-resource Languages (MSLRL) speech corpus defined for INTERSPEECH-18 challenge [19]. The corpus consists of spoken content from Gujarati, Telugu and Tamil languages covering 50 hours of spoken content approximately for each language. Further, a subset of the speech corpus was studied [20] for the keyword spotting task, and a portion of the corpus was annotated at the word level for evaluation purposes. Table 1 lists the statistics of the speech corpus [20]. The corpus comprises two different query categories: dev and test, covering both seen and unseen spoken query terms for evaluation. In addition, the read and conversational modes of speech uttered by both genders create the real-time scenario. All speech files are uniformly sampled at 16kHz with 16bit resolution. The length of the spoken query (in orthographic representation) spans between 3 to 20 approximately for all languages, and the frequency of the spoken term occurrences in the corpus span between 2 to 25, indicating the availability of the spoken terms in both dense and sparse categories.

### 4.2. Performance Metrics

The performance of the proposed approach was evaluated using hit, miss and false alarms obtained for each query ($q$) trial. In addition, the retrieval performance was evaluated based on the TWV (Term Weight Value) scores. The TWV specified in Eq. (10) was obtained based on the number of hits ($P_{hit}$), miss ($P_{miss}$) and false alarm ($P_{fa}$) candidates that occur over multiple trials.

$$\begin{aligned} P_{miss}(q) &= 1 - P_{hit}(q) \\ &= 1 - \frac{N_{correct}(q)}{N_{act}(q)} \end{aligned} \tag{8}$$

$$P_{fa}(q) = \frac{N_{fa}(q)}{N_{NT}(q)} \tag{9}$$

$$TWV(\psi) = 1 - \frac{F}{\mid q \mid} \tag{10}$$

$$F = \sum_{\forall q} P_{miss}(q,\psi) + \beta . P_{fa}(q,\psi)$$

$$\beta = \frac{\omega_{miss} \times P_{target}}{\omega_{miss} \times P_{target} + \omega_{fa} \times (1 - P_{target})}$$

For MSLRL corpus, the $\beta$ was computed as 0.009 by applying $\omega_{fa} = 1$, $\omega_{miss} = 100$ and $P_{target} = 0.0001$ as specified in the corpus [20]. $N_{correct}$, $N_{act}$, $N_{fa}$ and $N_{NT}$ indicate the number of correct spoken terms detected, the actual number of spoken terms, the number of false alarms and a number of trials, respectively. The TWV was obtained based on a weighted average of the miss and false alarm candidates that occur for a set of queries. Further, the ATWV (Average Term Weight Value) was obtained by computing the average TWV for all query trials. The parameter $\psi$ indicates the threshold for detecting the similarity score as a valid match or not.

### 4.3. Parameters

In the pattern discovery stage, the acoustic feature representation obtained from all languages was used to discover the similarities among them. Each spoken document was compared with other documents to capture the pattern similarities using similarity threshold $\eta$ ($0.95 \leq \eta \leq 1$) and match length $\mu$ ($10 \leq \mu \leq 32$). During the experiments, the optimal values for $\eta = 0.99$ and $\mu \geq 10$ were obtained over multiple trials with similar spoken terms. Hence, both $\mu \geq 10$ and $\eta = 0.99$ were fixed throughout the experiments. During the spoken term match detection, the acoustic feature representations of the similar ($T$) and dissimilar ($F$) patterns were fed to the network for training. The $T$ pairs were obtained by capturing the highest heuristic cost obtained at the diagonal region of the similarity matrix ($sim[,]$) with the term threshold $\mu$. Simi-

larly, the $F$ pair is obtained by capturing the minimal heuristic cost. Both ($T$ and $F$) patterns are discovered in a self-paced manner, which distinguishes our approach from others. Further, the network was trained with similar and dissimilar pairs to learn the discrimination at the feature level. In our approach, the network was trained for 100 epochs with 318,000 samples (82,000+124,000+ 112,000) of segment size $10 \times 80$ from Gujarati, Telugu and Tamil languages. The training accuracy of the network achieved was 91.79%, and the validation accuracy was 90.21%.

### 4.4. Experiments

The performance of spoken term detection was measured using the ATWV score over multiple queries obtained from the MSLRL corpus. The spoken term match was considered when the distance between a pair of segments was minimal (i.e. $\psi \leq 0.1$); otherwise, it is discarded. The results of the proposed approach were compared with other state-of-the-art systems [13, 21] for the spoken term detection task. Table 2 projects the results obtained for the dev queries. From the table, it is observed that the proposed approach obtains a maximum hit ratio of 48.7% for Gujarati and a minimal false alarm ratio of 16.3% for Telugu. Further, the proposed approach was compared with CNN-QBE [13] and segmental DTW [21] approaches for the same task with MSLRL corpus, and the results are depicted in Table 3. Based on the results, it is inferred that the hit, miss and false alarm ratio of the proposed approach were better compared with other approaches. Especially the false alarm reduction confirms that the pattern augmented by the proposed approach was appropriate in capturing the similarities. Hence, the relationship between the similar and dissimilar spoken terms learnt by the network was helpful in identifying the matches from the feature itself.

Table 2: *Performance of the proposed approach for development queries.*

| Language | Gujarati | Tamil | Telugu |
|---|---|---|---|
| Occurances | 1,638 | 2,348 | 1,700 |
| Hit | 798 | 1071 | 774 |
| Miss | 1,077 | 1,556 | 1,071 |
| FA | 187,624 | 172,428 | 142,175 |
| $P_{hit}$ | 0.487 | 0.456 | 0.455 |
| $P_{miss}$ | 0.513 | 0.544 | 0.545 |
| $P_{fa}$ | 0.175 | 0.181 | 0.163 |

Table 3: *Performance evaluation of the proposed approach in comparison with other approaches in the dev category.*

| Language | Approach | $P_{hit}$ | $P_{miss}$ | $P_{fa}$ | ATWV |
|---|---|---|---|---|---|
| Gujarati | CNN-QBE | 0.295 | 0.709 | 0.35 | 0.292 |
| | Segmental DTW | 0.361 | 0.639 | 0.44 | 0.357 |
| | Proposed | 0.487 | 0.513 | 0.175 | 0.485 |
| Tamil | CNN-QBE | 0.329 | 0.671 | 0.33 | 0.326 |
| | Segmental DTW | 0.345 | 0.655 | 0.421 | 0.341 |
| | Proposed | 0.456 | 0.544 | 0.181 | 0.454 |
| Telugu | CNN-QBE | 0.374 | 0.626 | 0.37 | 0.371 |
| | Segmental DTW | 0.384 | 0.616 | 0.452 | 0.378 |
| | Proposed | 0.455 | 0.545 | 0.163 | 0.454 |

The proposed approach was evaluated with unseen queries obtained from the test category of the MSLRL corpus. Table 4 shows the statistics of the results obtained in comparison with other approaches. Based on the obtained results, once again, the false alarm reduction was observed for the proposed approach. The reason for such reduction is considering the appropriate diagonal matches that occur between similar spoken terms. In the CNN-QBE method, the spoken term similarity was obtained by computing the global alignment between query and spoken documents. Though the CNN-QBE approach performs well in the spoken term detection task, it also generates a lot of false alarm candidates. Due to that, the performance degradation in the ATWV was observed. The segmental DTW approach captures the optimal alignment between the segments of a size similar to query size. Though spoken term detection is viable, it also generates false alarms and degrades retrieval accuracy.

Table 4: *Performance evaluation of the proposed approach in comparison with other approaches in the test category.*

| Language | Approach | $P_{hit}$ | $P_{miss}$ | $P_{fa}$ | ATWV |
|---|---|---|---|---|---|
| Gujarathi | CNN-QBE | 0.31 | 0.69 | 0.43 | 0.306 |
| | Segmental DTW | 0.35 | 0.65 | 0.49 | 0.346 |
| | Proposed | 0.492 | 0.51 | 0.18 | 0.488 |
| Tamil | CNN-QBE | 0.292 | 0.71 | 0.46 | 0.288 |
| | Segmental DTW | 0.347 | 0.65 | 0.51 | 0.342 |
| | Proposed | 0.432 | 0.57 | 0.17 | 0.428 |
| Telugu | CNN-QBE | 0.34 | 0.66 | 0.41 | 0.336 |
| | Segmental DTW | 0.366 | 0.63 | 0.487 | 0.362 |
| | Proposed | 0.441 | 0.56 | 0.16 | 0.439 |

## 5. Summary

In the proposed approach, the pattern variabilities that occur between similar spoken terms were addressed via (i) a heuristic pattern discovery approach and (ii) a Siamese spoken term detection network augmented based on the pattern discovery. The heuristic pattern discovery aims to capture the diagonal pattern similarities that occur in a contiguous and non-contiguous manner due to speech variabilities and augment the acoustic feature. The network learns the relationship between the similar and dissimilar pairs from the augmented data and accomplishes the spoken term detection task across languages in zero-resource constraint. Based on the results, it is observed that the proposed approach reduces the false alarms by 17.7% on average and improves the spoken term detection accuracy by 7.1% at least. Hence the ATWV score was improved by 7.7%, at least for Telugu. Further exploring the approach with a deeper network and different acoustic feature representations opens a new direction of research.

## 6. References


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

[2] A. Park and J. R. Glass, "Towards unsupervised pattern discovery in speech," in *IEEE Workshop on Automatic Speech Recognition and Understanding, 2005.*, 2005, pp. 53–58.

[3] A. S. Park and J. R. Glass, "Unsupervised pattern discovery in speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 186–197, 2008.

[4] O. Räsänen, G. Doyle, and M. C. Frank, "Unsupervised word discovery from speech using automatic segmentation into syllable-like units," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[5] 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)*, 2017, pp. 719–726.

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

[7] V. Gupta, J. Ajmera, A. Kumar, and A. Verma, "A language independent approach to audio search," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[8] P. D. Karthik, M. Saranya, and H. A. Murthy, "A fast query-by-example spoken term detection for zero resource languages," in *2016 International Conference on Signal Processing and Communications (SPCOM)*. IEEE, 2016, pp. 1–5.

[9] S. H. Dumpala, K. R. Alluri, S. V. Gangashetty, and A. K. Vuppala, "Analysis of constraints on segmental dtw for the task of query-by-example spoken term detection," in *2015 annual IEEE India conference (INDICON)*. IEEE, 2015, pp. 1–6.

[10] L. ten Bosch and B. Cranen, "A computational model for unsupervised word discovery." Antwerp, Belgium: ISCA, 2007.

[11] G. Aimetti, "Modelling early language acquisition skills: Towards a general statistical learning mechanism," in *Proceedings of the Student Research Workshop at EACL 2009*, 2009, pp. 1–9.

[12] L. Catanese, N. Souviraa-Labastie, B. Qu, S. Campion, G. Gravier, E. Vincent, and F. Bimbot, "Modis: an audio motif discovery software," in *Show & Tell-Interspeech 2013*, 2013.

[13] D. Ram, L. Miculicich, and H. Bourlard, "Cnn based query by example spoken term detection." in *Interspeech*, 2018, pp. 92–96.

[14] 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*. Springer, 2018, pp. 139–149.

[15] J. Cui, B. Kingsbury, B. Ramabhadran, A. Sethy, K. Audhkhasi, X. Cui, E. Kislal, L. Mangu, M. Nussbaum-Thom, and M. Picheny, "Multilingual representations for low resource speech recognition and keyword search," in *2015 IEEE workshop on automatic speech recognition and understanding (ASRU)*. IEEE, 2015, pp. 259–266.

[16] K. Knill, M. Gales, A. Ragni, and S. P. Rath, "Language independent and unsupervised acoustic models for speech recognition and keyword spotting," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2014, pp. 16–20.

[17] Y. Yuan, L. Xie, C.-C. Leung, H. Chen, and B. Ma, "Fast query-by-example speech search using attention-based deep binary embeddings," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1988–2000, 2020.

[18] D. Chicco, "Siamese neural networks: An overview," *Artificial neural networks*, pp. 73–94, 2021.

[19] B. M. L. Srivastava, S. Sitaram, K. Bali, K. D. M. Rupesh Kumar Mehta and, P. Matani, S. Satpal, K. Bali, R. Srikanth, and N. Nayak, "Interspeech 2018 low resource automatic speech recognition challenge for indian languages," in *SLTU*, August 2018.

[20] V. L. V. Nadimpalli, S. Kesiraju, R. Banka, R. Kethireddy, and S. V. Gangashetty, "Resources and benchmarks for keyword search in spoken audio from low-resource indian languages," *IEEE Access*, vol. 10, pp. 34 789–34 799, 2022.

[21] N. San, M. Bartelds, M. Browne, L. Clifford, F. Gibson, J. Mansfield, D. Nash, J. Simpson, M. Turpin, M. Vollmer *et al.*, "Leveraging pre-trained representations to improve access to untranscribed speech from endangered languages," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2021, pp. 1094–1101.