Data-Centric Engineering (2025), 6: e28
doi:10.1017/dce.2024.48

CAMBRIDGE
UNIVERSITY PRESS

RESEARCH ARTICLE

Recurring spoken term discovery in the zero-resource constraint using diagonal patterns

Sudhakar Pandiarajan[1], Sreenivasa Rao K[2] and Pabitra Mitra[2]

[1]School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, Tamilnadu, India
[2]Department of Computer Science and Engineering, Indian Institute of Technology, Kharagpur, West Bengal, India
Corresponding author: Sudhakar Pandiarajan; Email: sudhakar.p@vit.ac.in

Received: 30 April 2024; **Revised:** 22 August 2024; **Accepted:** 03 October 2024

Keywords: diagonal pattern search; low-resource speech processing; pattern discovery; spoken term discovery; zero resource

Abstract

Spoken term discovery (STD) is challenging when a large volume of spoken content is generated without annotations. Unsupervised approaches resolve this challenge by directly computing pattern matches from the acoustic feature representation of the speech signal. However, this approach produces a lot of false alarms due to inherent speech variabilities, leading to performance degradation in the STD task. To overcome these challenges and improve performance, we propose a two-stage approach. First, we identify an acoustic feature representation that emphasizes spoken content irrespective of the variability challenge. Second, we employ the proposed diagonal pattern search to capture spoken term matches in an unsupervised way without any transcriptions. The proposed approach validated using Microsoft Speech Corpus for Low-Resource languages reveals that an 18% gain in hit ratio and 37% reduction in the false alarm ratio was achieved compared with the state-of-the-art methods.

Impact Statement

An unsupervised spoken term discovery task aims to capture the spoken term matches in the speech corpus directly from the acoustic feature representation without any transcriptions. A challenge to the matching task is to handle the variabilities that exist in natural speech. In the proposed approach, we aim to overcome the challenges in both the acoustic feature representation and pattern-matching strategy. In the acoustic feature representation, we analyzed the signal processing, deep learning, and transfer learning approaches toward feature representation that emphasize the spoken content. In view of matching, the diagonal pattern match technique was proposed to handle the variabilities. Based on the evaluation using Microsoft Speech Corpus for Low-Resource Indian Languages, it is inferred that the hit ratio was improved by employing both the signal processing and deep learning-based acoustic feature representation combined with the proposed matching technique.

1. Introduction

In the recent communication era, a lot of spoken content generated without transcription challenges the spoken content retrieval (SCR) task. In the conventional approach, the SCR task was achieved by converting the spoken query and spoken content to its equivalent text using the automatic speech

This research article was awarded an Open Data and Open Materials badge for transparent practices. See the Data Availability Statement for details.

https://doi.org/10.1017/dce.2024.48 Published online by Cambridge University Press

recognition (ASR) system, and text-based matching was applied to detect similar spoken content. The ASR-based system seeks a huge volume of annotated spoken content to train the system for optimal performance (Weintraub, 1993). The performance of the SCR task depends on the ASR system, and consequently, it expects a large volume of transcribed spoken content. Nowadays, a large volume of spoken content is piled up in online repositories without transcription, which challenges the SCR task. Pattern matching is an alternate approach that aims to capture the similarities between spoken terms directly from the acoustic feature representation of the speech signal. Such an approach does not seek annotations and is well-suitable for the SCR task in the zero-resource constraint (spoken content without any transcriptions).

Spoken term discovery (STD) without annotation is a subset of the SCR task that aims to discover similar spoken terms in the speech corpus in an unsupervised manner. Jansen and Van Durme (2011), Park and Glass (2005, 2008), Räsänen et al. (2015), Kamper et al. (2017) attempted an unsupervised STD task to resolve the resource constraints. However, a significant performance gap was observed in comparison to the ASR-based techniques. The existing approaches in the unsupervised STD tasks are broadly grouped into two categories: (i) dynamic time warping (DTW) centric and (ii) template matching centric. In the DTW approach, temporal alignment between two acoustic feature representations is obtained, and similarities were computed (Park and Glass, 2005; Zhang and Glass, 2010; Gupta et al., 2011; Karthik Pandia et al., 2016). The challenge in the DTW approach is global alignment. The segmental DTW approach (Park and Glass, 2008) overcomes the challenge and achieves the task at the segmental level. Similarly, the statistical word discovery model (Bosch and Cranen, 2007), the *n*-gram model (Aimetti, 2009) based on dynamic programming, Randomized Algorithm approach (Jansen and Van Durme, 2011), and the audio motif discovery (MODIS) approaches (Catanese et al., 2013) utilized the segmental DTW technique to accomplish the STD task in an unsupervised way. Despite its advantages, the segmental DTW approach completely relies on the segment size, and deciding on the segment size is another challenge.

The template matching approach aims to discover the pattern similarities between two acoustic feature representations and determines the spoken term match. The syllable boundary-based *n*-gram approach (Räsänen et al., 2015) maps similar spoken terms at the syllable level and identifies the similarities. Kamper et al. (2017) developed an embedded segmental K-means model to capture similar spoken terms in an unsupervised way. Alternatively, an image-processing approach (Birla et al., 2018) was developed to capture the spoken term matches. Ravi and Krothapalli (2022) attempted an unsupervised spoken content segmentation approach at the phoneme level and captured the spoken term matches using 3-NDFS traversal technique. Despite the feasibility, the aforementioned approaches introduce many false alarm matches during the STD task and degrade the performance.

One of the major concerns in the pattern discovery approach is eliminating the variabilities in the acoustic feature representation that arise due to the speakers, language, and environmental specific changes. As an effect, the pattern match propagation between the same spoken term varies in three ways: (i) a total match in sequence between the two acoustic feature representations of the spoken terms (referred Type-I), (ii) a partial match occurs at the prefix or the suffix portion or intermediate portion of the spoken term matched region (referred Type-II), and (iii) multiple noncontiguous partial matches in a spoken term matched region (referred Type-III). Therefore, the pattern discovery approach should be robust enough to capture all possible matches and determine the spoken term match appropriately. The proposed diagonal pattern search (DPS) overcomes the drawbacks by capturing all types of pattern similarity propagation (Types I, II, and III) in the matched region without constraint. Furthermore, the RASTA-PLP spectrogram (Hermansky and Morgan, 1994), Mel-Spec$_{norm}$ (Sudhakar et al., 2023), and Wav2vec embeddings (Conneau et al., 2020) were employed as acoustic feature representations and the STD task was analyzed in the zero-resource constraint. Based on the performance, it is inferred that the proposed DPS has achieved an 18% improvement in comparison with the segmental DTW approach using Mel-Spec$_{norm}$ representation.

Further, Section 2 briefly describes the acoustic feature extraction from the speech signal. Section 3 discusses in detail the proposed diagonal pattern matching technique to identify the spoken term matches

https://doi.org/10.1017/dce.2024.48 Published online by Cambridge University Press

in an unsupervised way. Section 4 outlines the corpus used to evaluate the performance along with performance metrics. The experimental results are discussed in detail in Section 5. Section 6 concludes the article with future research direction.

2. Acoustic feature representation

The acoustic feature vectors of the speech signal carry a comprehensive representation of the spoken content. Therefore, determining the spoken content matches from the speech signal is viable. However, a challenge to the spoken content match detection task is to handle the speech variabilities. Human speech carries spoken content along with additional information specific to the speaker, gender, environment, and so forth Disentangling the spoken content from the complex signal is a challenging task. In our approach, the acoustic feature representation obtained from the RASTA-PLP spectrogram, Mel-Spec$_{norm}$, and Wav2vec embeddings were analyzed toward the STD task.

2.1. RASTA-PLP spectrogram

The RASTA-PLP spectrogram representation (Hermansky and Morgan, 1994) was obtained by processing the speech signal in a sequence of stages. At first, the speech signal was divided into a sequence of frames of 20 ms duration with 50% overlap. In the second stage, the discrete fourier transform (DFT) was applied to the windowed speech signal at the frame level. In the third stage, the DFT values obtained were analyzed at the critical bands. In the fourth stage, the RASTA filter specified in Eq. (1) was applied.

$$H(z) = 0.1z^4 \times \frac{2 + z^{-1} - z^{-3} - 2z^{-4}}{1 - 0.98z^{-1}} \tag{1}$$

In the fifth stage, the filtered spectrum was uplifted with a factor for loudness and intensity amplification. Finally, the cepstral coefficients were computed at the frame level. Similarly, the coefficients were computed for all frames and stacked in sequence to represent the acoustic features of the speech signal.

2.2. Mel-Spec$_{\mathbf{norm}}$

The Mel-Spec$_{norm}$ representation (Sudhakar et al., 2023) of the speech signal was obtained by processing the speech signal in a sequence of steps. At first, the 80-dimensional Mel-spectrogram was obtained from the speech signal (Slaney, 1998). The Mel-spectrogram of the speech signal carries a spoken message along with additional information specific to the speaker and gender variabilities. In the second step, the spoken contents are disentangled from the Mel-spectrogram using Deep Convolutional Encoder-Decoder (DeCoED) architecture. The DeCoED network disentangles the spoken content from the Mel-Spectrogram representation directly without any additional information related to spoken content. Finally, the 80-dimensional speaker normalized Mel-spectrogram representation (Mel-Spec$_{norm}$) was used as an acoustic feature representation for the discovery task.

2.3. Wav2vec2 embedding

In the Wav2vec 2.0 framework (Baevski et al., 2020), self-supervised learning was exploited to understand the speech representation directly from a large volume of speech corpus in an unsupervised way. Wav2vec 2.0 model processes the raw audio through convolutional neural network (CNN) and Transformers. At first, the raw speech signal was fed to the multilayer CNNs to learn the latent representation. Further, the latent representation was fed to the transformers and trained by contrastive function[1]. Finally, the network was fine-tuned by adding an additional layer to the existing network to suit the speech recognition task. However, we obtained the embeddings from the pre-trained model with

[1] https://github.com/pytorch/fairseq

https://doi.org/10.1017/dce.2024.48 Published online by Cambridge University Press

960 hours of training data from the Libiri Speech corpus. The 512-dimensional latent representation of the speech was obtained from the Wav2vec 2.0 model and used as an acoustic feature representation.

3. STD

The STD task aims to capture the spoken term matches directly from the acoustic feature representation using the pattern matching approach. The spoken term match between the same spoken content exhibits a diagonal pattern match propagation in the similarity matrix. However, the match propagation strategies in the similarity matrix vary due to the temporal, speaker, gender, and environmental-specific conditions. Due to that, the diagonal pattern match propagation varies even though the same spoken content occurs at different time instances. In the proposed approach, we aim to overcome the variability challenge and capture the spoken term matches in the corpus directly from the acoustic feature representation.

In the proposed approach, the diagonal pattern match was computed between two acoustic feature representations in multiple stages. At first, the acoustic feature representation of two spoken documents $X = [x_1, x_2, \ldots, x_L], x \in X \in D^i$ and $Y = [y_1, y_2, \ldots, y_M], Y \in D^j$ was extracted from the speech signal. In the next stage, the cosine similarity between D^i and D^j was computed frame-wise using Eq. 2 for all frames $l, m \in L, M$.

$$sim[l,m] = \frac{x_l.y_m}{\sqrt{x_l^2}.\sqrt{y_m^2}} \tag{2}$$

In the next stage, the similarity matrix $sim[L,M]$ was binarised using Eq. (3) for all elements in u, v where $1 \leq u \leq L, 1 \leq v \leq M$.

$$sim_b[u,v] = \begin{cases} \mathbf{1}, & if\, sim[u,v] \geq \eta \\ \mathbf{0}, & otherwise. \end{cases} \tag{3}$$

In sim_b, **1** indicates that the frames correspond to the two documents having higher similarity and **0** refers to dissimilarity (mismatch). In the next stage, the sim_b was used to capture the pattern similarities between spoken terms. During the spoken term matching, the proposed diagonal pattern match algorithm captures the match(es) in three steps. At first, the sim_b was scanned diagonally to identify the segmental matches that occur in sequence in the diagonal. In the similarity matrix, $sim_b[u,v]$ represents the binarised values at u, v positions. The depth match $sim_d[u,v]$ at a specific diagonal region was obtained using Eq. (4).

$$sim_d[u,v] = \begin{cases} sim_b[u,v], if\, u = 1\, or\, v = 1 \\ sim_b[u,v] + sim_b[u-1, v-1], if\, u,v > 1 \end{cases} \tag{4}$$

Similarly, for all matrix elements, the depth match was computed in the diagonal, and the depth similarity matrix sim_d was arrived. The larger values in sim_d indicate the higher sequential similarity between the spoken terms. In the second stage, the diagonal match cost was computed for all diagonal entries $(L+M-1)$ in sim_d. The cost consideration in the diagonal path accounts for all types (I, II and III) of similarities and emphasizes the similarity level. Based on the analysis, it is observed that the affinity between two similar spoken content propagates diagonally in a contiguous as well as noncontiguous manner. This is due to the variability of natural speech. Hence, considering the diagonal matches of both contiguous and noncontiguous regions works for the variability issues. In the next step, the potential similarity region(s) were identified based on the diagonal match cost obtained using Eq. 5.

$$dia\_sum[k] = \sum_{\forall k=1}^{L+M-1} diag(k) \tag{5}$$

The $diag()$ function returns all cost values associated in the kth diagonal of the similarity matrix sim_d, and the dia_sum list maintains the diagonal match cost for each diagonal. Finally, potential similarity region(s) in sim_d were identified by analyzing the diagonal cost dia_sum with the threshold λ. The λ indicates the minimum depth of the similarity to be considered as a potential spoken term match. The computational

https://doi.org/10.1017/dce.2024.48 Published online by Cambridge University Press

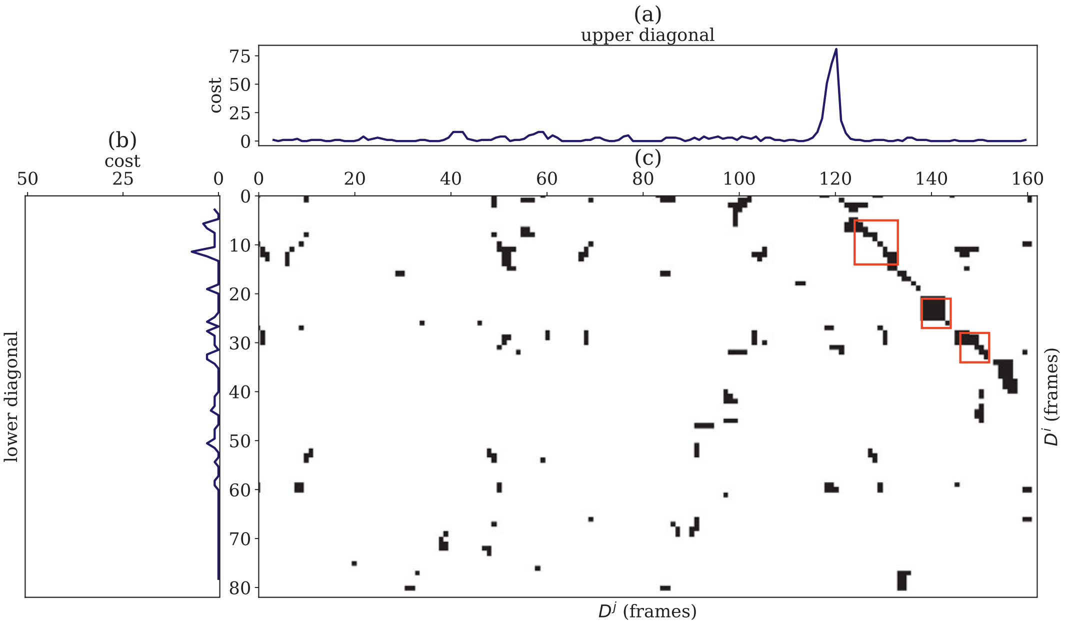

Figure 1. *depicts the spoken term match detected by the proposed approach. (a) and (b) indicate the upper and lower diagonal costs computed from the similarity matrix, respectively. (c) Highlights the matched regions in red color rectangle boxes.*

complexity of the proposed STD is computed as $O(L \times M)$, where L and M represents the number of frames in the documents D^i and D^j, respectively. The similarity matrix computation task performs $L \times M$ comparison for matrix computation and $O(k)$ for determining the match existence from k diagonals. Hence, the total computation involves $O(L \times M) + O(k)$.

Figure 1 depicts the matches obtained between two spoken documents uttered in Tamil. The documents D^j and D^i contains the spoken content “Inta pōrāṭṭattil amaiccarkaḷ paṅku” and “Amaiccarkaḷ ar̲brikkai on̲r̲ai vel̤̱iyiṭṭan̲ar”, respectively. From Figure 1(a), it is inferred that the diagonal cost for the spoken term similarity region is high (indicated by the line marked in blue colour) compared to other regions. By thresholding the diagonal cost with λ, the potential spoken term similarities were captured by the diagonal pattern match approach. The proposed method differs from the DTW-centric techniques (subsequence DTW and segmental DTW) by identifying the spoken term matches without constraining the segment size. Figure 2 shows the comparison between the proposed approach and DTW-centric approaches for the spoken term detection task. The spoken term “viḷaiyāṭṭu” in x-axis was search in the document containing “T.V kaḷin̲ varukaikku pin̲n̲arē kul̲antaikal̤̱in̲ viḷaiyāṭṭu mōkam kur̲aiya ārampittuviṭṭatu” in y-axis. Figure 2(a) and (b) indicates the cost matrix obtained from the proposed approach and its diagonal cost. Similarly, Figure 2(c)–(f) indicates the subsequence DTW and its cost matrix, segmental DTW, and its cost matrix, respectively. From the figure, it is clear that the proposed approach captures the similarity region and highlights the spoken term matches appropriately.

4. Performance Evaluation

The performance of the STD completely depends on both the acoustic feature representation and spoken term match detection. Therefore, the performance of the proposed approach was evaluated in two aspects: (i) acoustic feature representation and (ii) spoken term match detection. In view of acoustic feature representation, the RASTA-PLP spectrogram, Mel-Spec$_{norm}$ and Wav2vec embedding were evaluated toward the matching task. In view of spoken term match detection, the proposed diagonal pattern match

https://doi.org/10.1017/dce.2024.48 Published online by Cambridge University Press

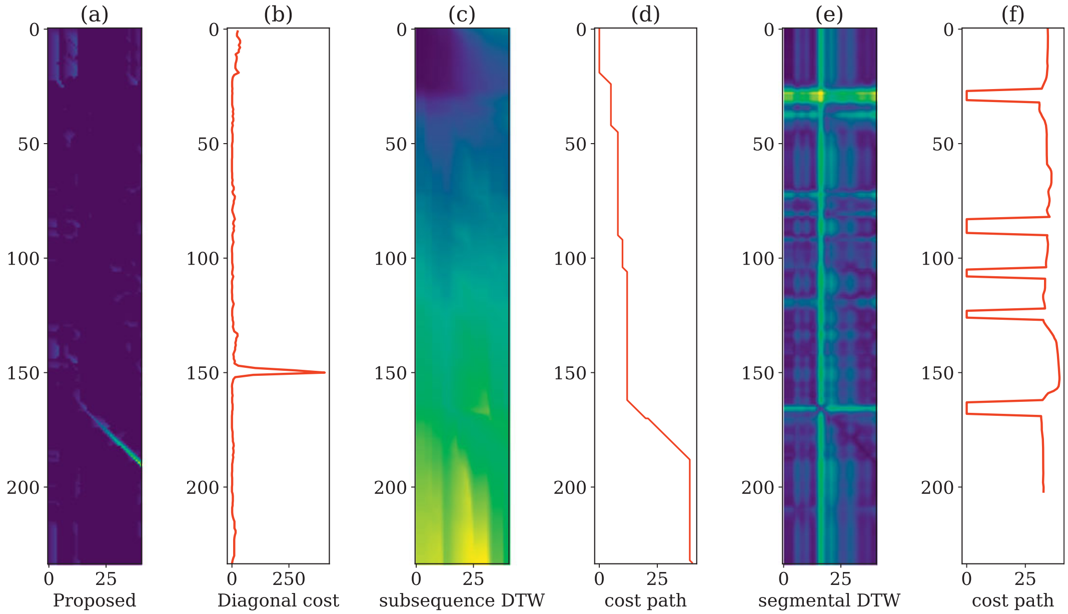

Figure 2. *(a) and (b) indicates the cost matrix and diagonal cost for the proposed approach. (c) and (d) indicates the cost matrix and cost path for the subsequence DTW. Similarly, (e) and (f) represent the segmental DTW approach.*

detection approach was compared with the other state-of-the-art system (San et al., 2021; Ram et al., 2019).

The MSLRL speech corpus defined for low-resource languages was used to analyze the performance of the discovery task in the zero-resource constraint.

The MSLRL speech corpus (Srivastava et al., 2018) was released in the Interspeech-2018 ASR challenge for low-resource Indian languages. It consists of ≈ 67 hours of training data without transcription and ≈ 5 hours of test data with transcriptions from Gujarathi, Tamil, and Telugu languages. The corpus consists of spoken utterances in both read and conversational modes to simulate the real-time scenario. All the speech files maintain the uniform sampling frequency of 16 kHz. Table 1 lists the number of spoken documents, duration, and speaker details of the training data considered. The training data in the MSLRL corpus contains only spoken documents and speaker information without any annotation, creating a real-time zero-resource scenario.

The standard performance measures, hit (H), miss (M), and false alarms (FA), were used to evaluate the discovery task. The hit is counted if the discovered spoken term belongs to the ground truth document pair. Otherwise, it is a false alarm. The miss reveals the absence of the spoken term pair(s) in the discovery task.

Table 1. *Details of the MicroSoft Low-Resource Language corpus. # Docs. represents the number of spoken documents. # Speakers indicate the number of speakers*

Language	# Docs.	Duration (hours)	# Speakers	# Docs.	Duration (hours)	# Speakers
	Training data			Test data		
Gujarathi	12,682	22.16	108	1155	2.27	89
Telugu	21,891	21.33	1210	603	1.02	53
Tamil	20,805	23.90	874	1032	1.55	84

https://doi.org/10.1017/dce.2024.48 Published online by Cambridge University Press

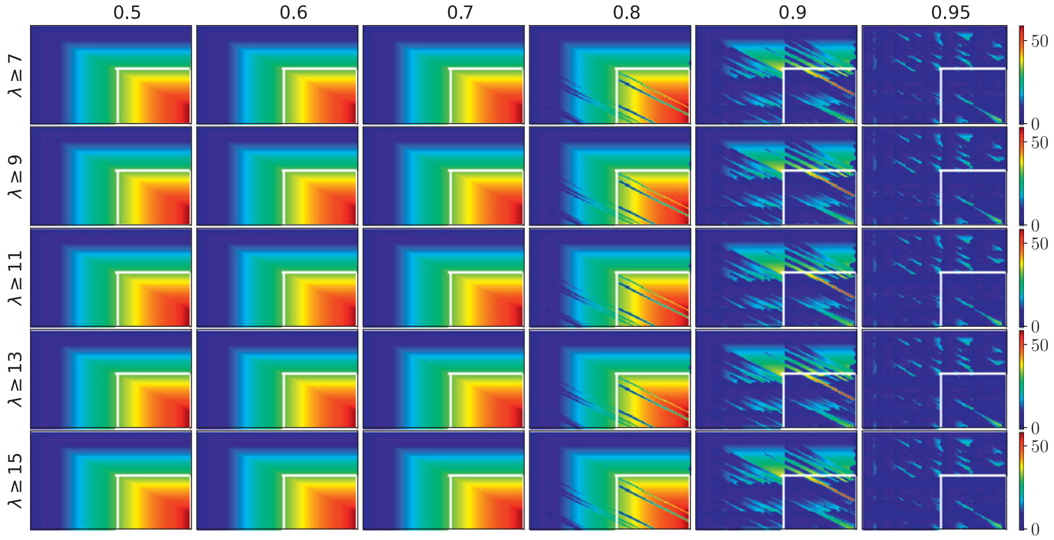

Figure 3. depicts the match propagation in the similarity matrix by varying the thresholds $0.5 \leq \eta \leq 1$ and $7 \leq \lambda \leq 15$.

True positive rate (hit rate) in Eq. (6) indicates the ratio of spoken term matches discovered that overlap with the ground truth match.

$$True\ positive\ rate\,(TPR) = \frac{Hit}{Hit + FA} \tag{6}$$

The false positive rate mentioned in Eq. (7) indicates the ratio of nonmatched spoken terms that are detected as a match by the system.

$$False\ positive\ rate\,(FPR) = \frac{FA}{Hit + FA} \tag{7}$$

An ideal system improves the TPR and reduces FPR in the spoken-term discovery task. In the proposed approach, the thresholding parameters η and λ were chosen empirically by maximizing the hit and minimizing the miss and false alarms. Figure 3 depicts the similarity matrix obtained between two spoken documents containing a spoken term match (“Labhistundi” in Telugu). The white-coloured rectangle indicates the ground truth region. From the figure, it is observed that the diagonal match propagation was distinct when $\eta > 0.95$ and $\lambda > 9$. Further increasing the η and λ values, discriminating the matched region clearly in the similarity matrix.

The optimal choices of parameters η and λ were computed for 100 documents chosen randomly that maximize the true positive rate and minimize the false positive rate. Figure 4 depicts the relationship between the TPR and FPR by varying the threshold parameters. Based on the result, it is inferred that the FPR reaches a minimum and TPR reaches a maximum when $\lambda = 9$ and $\eta = 0.97$. Further, varying the threshold either increases the FPR or reduces the TRP. Hence, the optimal choices ($\lambda = 9$ and $\eta = 0.97$) are retained for the Mel-Spec$_{norm}$ representation. Similarly, $\eta = 0.6$ and $\lambda = 9$ for Wav2Vec embeddings and $\eta = 0.8$ and $\lambda = 11$ for RASTA-PLP spectrogram were arrived, and the same was retained for all experiments.

5. Results and Discussion

The performance of the STD task was measured based on acoustic feature representation and pattern match detection. In view of acoustic feature representation, RASTA-PLP spectrogram, Mel-Spec$_{norm}$, and

https://doi.org/10.1017/dce.2024.48 Published online by Cambridge University Press

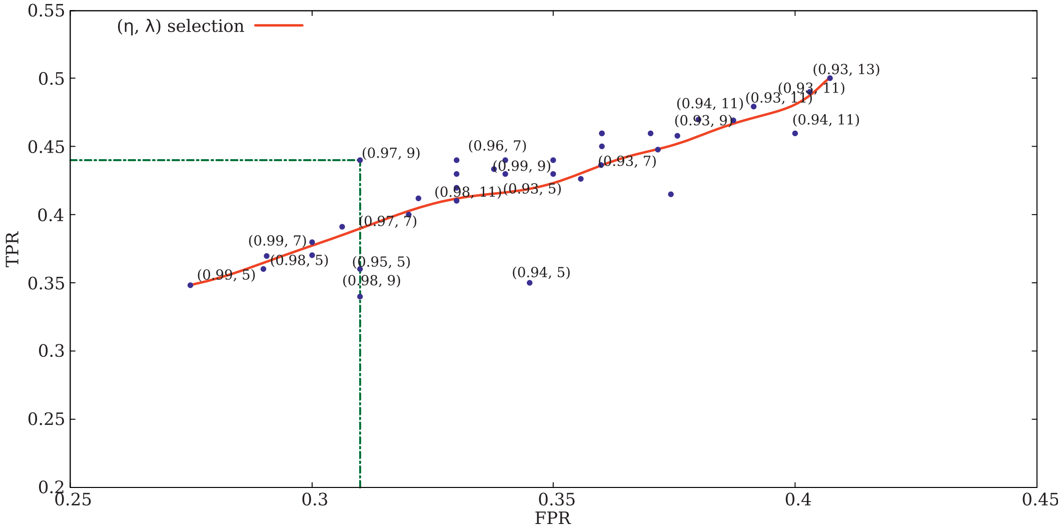

Figure 4. depicts the relationship between true positive rate (TPR) and false positive rate (FPR) obtained from the proposed approach by varying η and λ.

Wav2vec embedding were utilized to assess the system performance. Table 2 describes the performance of the discovery task employing the RASTA-PLP spectrogram. Based on the table, it is observed that a hit rate of 56.1%, 55.0%, and 62.5% was observed in Gujarati, Telugu, and Tamil, respectively. This phenomenon indicates that the RASTA-PLP representation carries the spoken content, which was detected by the proposed diagonal pattern match in an unsupervised way.

Table 3 represents the results of the discovery task using Wav2vec representation. Based on the results, it is inferred that the hit values decreased in the Wav2vec representation compared to the RASTA-PLP representation. A 30.9%, 27.9%, and 29.3% hit rate was achieved using Wav2vec representation. In comparison with the RASTA-PLP representation, a 25.1%, 27.0%, and 33.2% reduction in the hit ratio was observed in Gujarati, Telugu, and Tamil, respectively. This reduction indicates that the Wav2vec representation has language-specific information in the embeddings. Due to that, the performance of the

Table 2. *Performance of the STD task using RASTA-PLP representation. # Matches$_{act}$ represent the ground truth matches*

Language	# Matches$_{act}$	# Discovered	Hit	Miss	FA
Gujarati	373,796	792,448	209,700	164,096	582,748
Telugu	16,363	51,098	9012	7351	42,086
Tamil	39,339	170,292	24,620	14,719	145,672

Table 3. *Performance of the STD task using Wav2vec representation. # Matches$_{act}$ represent the ground truth matches*

Language	# Matches$_{act}$	# Discovered	Hit	Miss	FA
Gujarati	373,796	1,736,392	115,876	257,920	1,620,516
Telugu	16,363	41,858	4581	11,782	37,277
Tamil	39,339	148,947	11,534	27,805	137,413

https://doi.org/10.1017/dce.2024.48 Published online by Cambridge University Press

Table 4. *Performance of the STD task using Mel-Spec$_{norm}$ representation. # Matches$_{act}$ represent the ground truth matches*

Language	# Matches$_{act}$	# Discovered	Hit	Miss	FA
Gujarati	373,796	734,510	185,777	188,019	548,733
Telugu	16,363	41,477	8129	8234	33,348
Tamil	39,339	154,866	20,575	18,764	134,291

embeddings in unknown languages has reduced the hit rate. Meanwhile, the false alarms generated by the Wav2vec embeddings are also high in comparison with the RASTA-PLP representation.

The performance of the Mel-Spec$_{norm}$ representation in the STD task was presented in Table 4. From Table 4, it is observed that a 49.7%, 49.6%, and 52.3% hit rate was achieved in Gujarati, Telugu, and Tamil, respectively. In comparison with the RASTA-PLP representation, Mel-Spec$_{norm}$ has achieved a lesser hit ratio by 6%, 5%, and 10% in Gujarati, Telugu, and Tamil, respectively. However, it is noticed that the Mel-Spec$_{norm}$ has reduced the false alarm by 2.5%, 2.4%, and 1.0% in comparison with RASTA-PLP representation in Gujarati, Telugu, and Tamil, respectively. A minimum false alarm ratio of 9.1% was observed in Telugu.

Based on acoustic feature analysis in the discovery task, it is inferred the Mel-Spec$_{norm}$ representation has achieved an average of 50% hit rate, allowing 20% false alarm. The RASTA-PLP representation has an average of 57% hit rate, allowing 22.9% false alarms. The performance of both Mel-Spec$_{norm}$ and RASTA-PLP representation reveals that at least 50% of repeated spoken terms can be detected through the proposed diagonal pattern discovery approach. However, it is also noted that the Wav2vec embeddings are unable to perform well in the pattern discovery task.

Similarly, the performance of the proposed approach was evaluated using the other state-of-the-art methods (San et al., 2021; Ram et al., 2019) in the STD task. In the segmental DTW approach (San et al., 2021), a fixed segment of 10 frames (200 ms) was fixed for each segment, and the matching was computed using the DTW approach in the fixed region. The segment size was chosen empirically by considering ≈ 50% hit ratio. The CNN-based approach (Ram et al., 2019) uses the pre-trained CNN to detect the spoken term matches. In all approaches, the Mel-Spec$_{norm}$ was used as acoustic feature representation to avoid bias. Table 5 projects the performance of the STD task employing Mel-Spec$_{norm}$ representation. The highest hit ratio of 54.9% was observed in Telugu, allowing a false alarm ratio of 28.7% using the segmental DTW approach. Meanwhile, the proposed approach achieved the lowest false alarm ratio of 9.1% in Telugu, with a hit ratio of 49.6%. An average of 50.72% hit ratio, allowing a 20.97% false alarm ratio, was achieved by the proposed approach across languages. The segmental DTW approach achieved an average 51.26% hit ratio, allowing an average of 29.63% false alarms across languages. The CNN approach achieved an average of 33.26% hit ratio, allowing an average of 34.61% false alarms. This phenomenon indicates that the proposed approach captures ≈ 50% of spoken term matches and reduces the false alarm by 8% in comparison with the segmental DTW approach across languages. This is viable because of the diagonal pattern match strategy applied in the proposed approach, which captures the appropriate spoken term matches and minimizes the false alarms.

Table 5. *Performance of the proposed approach in comparison with the state-of-the-art systems*

	Gujarati	Telugu	Tamil
Language	Hit/Miss/FA	Hit/Miss/FA	Hit/Miss/FA
segmental DTW	191,384/182412/395539	8984/7379/104538	20,575/18764/324833
CNN based	110,270/263526/615166	6120/10243/94912	5383/10980/336982
Proposed	187,577/188019/548733	8129/8234/33348	20,575/18764/134291

https://doi.org/10.1017/dce.2024.48 Published online by Cambridge University Press

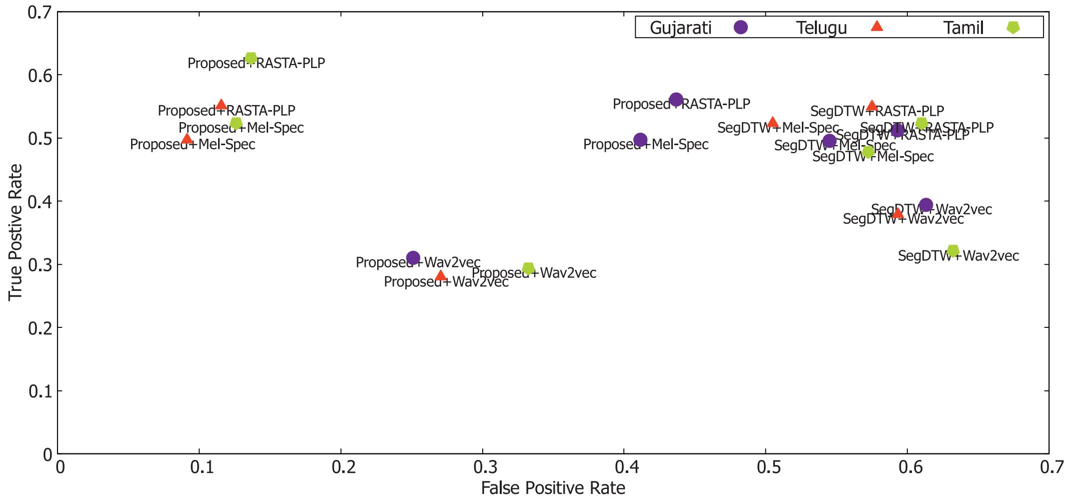

Figure 5. *Performance comparison across methods and features.*

Figure 5 depicts the performance of the proposed approach in comparison with the other state-of-the-art methods across acoustic features and languages. Based on the figure, it is observed that the proposed DPS approach with RASTA-PLP and Mel-Spec$_{norm}$ representation has reduced the false positive rate and improved the true positive rate for Tamil and Telugu. Meanwhile, it is also noted that the segmental DTW approach allows a lot of false alarms; hence, the false positive rate was high. This scenario confirms that the performance of the STD task depends on both the acoustic feature representation and pattern matching strategy.

In summary, the Mel-Spec$_{norm}$ representation achieved an average of $\approx 50\%$ hit ratio, allowing $\approx 20\%$ false alarm employing proposed diagonal pattern match representation. Similarly, the RASTA-PLP has achieved an average of 57% hit ratio with 23% false alarms. This scenario indicates that the proposed DPS search using Mel-Spec$_{norm}$ and RASTA-PLP representations identifies at least 50% of spoken term matches in the corpus. In addition, a performance gain of 18% in terms of hit ratio and a false alarm reduction of 37% was achieved in comparison with the segmental DTW approach employing Mel-Spec$_{norm}$ representation. Furthermore, the proposed approach has accomplished the STD task without any language-specific transcriptions and fits to the zero-resource scenario.

6. Conclusion

In this article, we demonstrated the STD task using the diagonal pattern match approach in the zero-resource constraint. The challenges arising from the speech variabilities are addressed in both the acoustic features and the pattern-matching approach. In view of acoustic feature representation, we infer that the language-specific features obtained from RASTA-PLP and Mel-Spec$_{norm}$ have better representation than the language-independent feature obtained from the self-supervised learning framework. Moreover, handling multiple types of pattern match propagation in the matched region improves the system's performance. As a result, an improvement of 18% hit ratio and a reduction of 37% false alarms was achieved in the proposed approach. The STD task is helpful in identifying the repeated spoken terms (keywords) in the corpus. Further, the repeated spoken terms can be organized with indices and retrieval can be achieved effectively given a spoken query. In future, we aim to augment the proposed approach pattern matches with deep learning techniques toward pattern-matching tasks, thereby improving the performance of the system.

Data availability statement. The Microsoft Low-Resource speech corpus available at: https://www.microsoft.com/en-us/download/details.aspx?id=105292. The annotated transcription for validation is available at: https://github.com/skesiraju/indic-kws. The code is available at: https://github.com/sudhakar-pandiarajan/unsupervised_spoken_term_discovery.

https://doi.org/10.1017/dce.2024.48 Published online by Cambridge University Press

Author contributions. S.P.: Conceptualization and Methodology formulation; S.R., P.M.: Methodology verification and validation; S.R., P.M.: Curation of the article; S.P.: Manuscript draft.

Funding statement. 'None'.

Competing interest. 'None'.

Ethical standard. The research meets all ethical guidelines, including adherence to the legal requirements of the study country.

References

Aimetti G (2009) Modelling early language acquisition skills: Towards a general statistical learning mechanism. In *Proceedings of the Student Research Workshop at EACL 2009*, pp. 1–9.

Baevski A, **Zhou H**, **Mohamed A and Auli M** (2020) wav2vec 2.0: a framework for self-supervised learning of speech representations. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20.

Birla L, *et al.* (2018) A robust unsupervised pattern discovery and clustering of speech signals. *Pattern Recognition Letters 116*, 254 –261.

Bosch L t and Cranen B (2007) A computational model for unsupervised word discovery. In *Eighth Annual Conference of the International Speech Communication Association*.

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

Conneau A, **Baevski A**, **Collobert R**, **Mohamed A and Auli M** (2020). Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979*.

Gupta V, **Ajmera J**, **Kumar A and Verma A** (2011) A language independent approach to audio search. In *Twelfth Annual Conference of the International Speech Communication Association*.

Hermansky H and Morgan N (1994) Rasta processing of speech. *IEEE Transactions on Speech and Audio Processing 2* (4), 578 –589.

Jansen A and Van Durme B (2011) Efficient spoken term discovery using randomized algorithms. In *2011 IEEE Workshop on Automatic Speech Recognition & Understanding*. IEEE, pp. 401–406.

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

Karthik Pandia DS, **MS Saranya, and Hema A Murthy** (2016) A fast query-by-example spoken term detection for zero resource languages. In *2016 International Conference on Signal Processing and Communications (SPCOM)*. IEEE, pp. 1–5.

Park A and Glass J (2008) Unsupervised pattern discovery in speech. *IEEE Transactions on Audio, Speech, and Language Processing 16* (1), 186 –197.

Park, A. and Glass JR (2005) Towards unsupervised pattern discovery in speech. In *IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 53–58.

Rao KS (2022) A novel approach to unsupervised pattern discovery in speech using convolutional neural network. *Computer Speech and Language 71*, 101259.

Ram D, **Miculicich L and Bourlard, H.** (2019) Multilingual bottleneck features for query by example spoken term detection. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, pp. 621–628.

Räsänen O, **Doyle G and Frank MC** (2015) Unsupervised word discovery from speech using automatic segmentation into syllable-like units. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Ravi KK and Krothapalli SR (2022) Phoneme segmentation-based unsupervised pattern discovery and clustering of speech signals. *Circuits, Systems, and Signal Processing 41* (4), 2088 –2117.

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

Slaney M (1998) Auditory toolbox: a matlab toolbox for auditory modeling. In *Work Technical Report, Interval Research Corporation*, pp. 29–32.

Srivastava BML, **Sitaram S**, **Bali K**, **Mehta RK**, **Mohan KD**, **Matani P**, **Satpal S**, **Bali K**, **Srikanth R and Nayak N** (2018) Interspeech 2018 low resource automatic speech recognition challenge for indian languages. In *SLTU*.

Sudhakar P, **Sreenivasa Rao K and Mitra P** (2023) Query-by-Example Spoken Term Detection for Zero-Resource Languages Using Heuristic Search. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

Weintraub M (1993) Keyword-spotting using sri's decipher large-vocabulary speech-recognition system. In *1993 IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE, Vol. 2, pp. 463–466.

Zhang Y and Glass JR (2010) Towards multi-speaker unsupervised speech pattern discovery. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4366–4369.

Cite this article: Pandiarajan S, K SR and Mitra P (2025). Recurring spoken term discovery in the zero-resource constraint using diagonal patterns. *Data-Centric Engineering*, 6, e28. doi:10.1017/dce.2024.48