IEEE Transactions on Audio, Speech and Language Processing

Untranscribed Spoken Content Indexing and Retrieval through keyword discovery

Journal:	*Transactions on Audio, Speech and Language Processing*
Manuscript ID	T-ASL-10124-2023
Manuscript Type:	Regular Paper
Date Submitted by the Author:	28-Jul-2023
Complete List of Authors:	P, Sudhakar; Indian Institute of Technology Kharagpur, Advanced Technology Development Centre K, Sreenivasa; IIT Kharagpur, Computer Science and Engineering Mitra, Pabitra ; Indian Institute of Technology Kharagpur, Computer Science and Engineering
Subject Category Please select at least one subject category that best reflects the scope of your manuscript:	HUMAN LANGUAGE TECHNOLOGY
EDICS:	HLT-SDTM Spoken Document Retrieval and Text Mining < HUMAN LANGUAGE TECHNOLOGY

Author's Responses to Custom Submission Questions

Is this manuscript a resubmission of, or related to, a previously rejected manuscript?	No
If "Yes", specify the publication venue and manuscript ID of the previous submission and upload a supporting document detailing how the resubmission has addressed the concerns raised during the previous review. If this does not apply, type N/A.	N/A
Is this manuscript an extended version of a conference publication?	No
If "Yes", provide the full citation of the conference submission or publication. If this does not apply, type N/A.	N/A
Is this manuscript related to any other papers of the authors that are either published, accepted for publication, or currently under review, and that are not included among the references cited in the manuscript?	No
If "Yes", please list these papers below. Except for permitted preprints, explain why these papers are not included among the references cited in the manuscript and how they are different from the manuscript. Include any unpublished papers as "Supporting Documents". If this does not apply, type N/A.	N/A
What is the contribution of this paper, within the scope of Transactions on Audio, Speech and Language Processing?	The objective is to develop a language-independent spoken document indexing and retrieval system for the speech corpus that does not have annotations in an unsupervised way. Hence, this approach devised a pattern discovery method to accomplish the objective in the resource-free constraint.
Why is the contribution significant (What impact will it have)?	In the proposed approach, spoken content retrieval was achieved in the absolute absence of linguistic resources. In addition, the proposed method reduces false alarms by 26% and improves the accuracy by 7% in comparison with the other state-of-the-art systems in the same category.
What are the three papers in the published literature most closely related to this paper? Please provide full citation details, including DOI references where possible.	1. D. Ram, L. Miculicich, and H. Bourlard, â€œMultilingual bottle- neck features for query by example spoken term detection,â€□ in 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). IEEE, 2019, pp. 621â€"628

	2. N. San, M. Bartelds, M. Browne, L. Clifford, F. Gibson, J. Mansfield, D. Nash et al., â€œLeveraging pre-trained represen- tations to improve access to untranscribed speech from endan- gered languages,â€□ in 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). IEEE, 2021, pp. 1094â€" 1101. 3. O. RÃ¤sÃ¤nen and M. A. C. BlandÃ³n, â€œUnsupervised discovery of recurring speech patterns using probabilistic adaptive metrics,â€□ 2020.
What is distinctive/new about the current paper relative to these previously published works?	The aforesaid approaches aim to discover the similarity patterns based on the trained neural networks with existing resourceful languages and detect the matches according to the query. However, in the proposed approach, the similarity patterns are discovered from the speech corpus and indexed before the query occurs. This approach leads to two advantages. (i) The retrieval was achieved from the pre-determined similarity information; hence outliers are easily avoided. (ii) The search will be bounded to the limited indices discovered instead of the entire speech corpus.

Untranscribed Spoken Content Indexing and Retrieval through keyword discovery

Sudhakar P, *student member, IEEE*, Sreenivasa Rao K and Pabitra Mitra, *senior members, IEEE*

***Abstract*—In the recent growth of multimedia technologies, a large volume of spoken content generated without annotation becomes a challenge for the retrieval task. The pattern matching approaches aim to overcome the challenges by directly identifying the pattern similarities between the spoken query and the spoken documents. Despite feasibility, a real challenge is handling the variabilities that arise in natural speech due to speaker, language and environmental-specific changes. As a result, a lot of false alarm candidates are generated during the retrieval task and degrade the performance. In the proposed approach, we aim to overcome the challenges and achieve the retrieval task in four stages. At first, we aim to reduce the variability challenges at the acoustic feature level using RASTA-PLP and Mel-spectrogram representations. In the second stage, the similarity between spoken terms that exist in the corpus was discovered using our heuristic pattern match algorithm. In the third stage, the discovered spoken term similarities were grouped and indexed by the proposed keyword discovery approach. Finally, given a spoken query, relevant spoken terms were retrieved using the indices discovered. The proposed approach was evaluated using Microsoft Low-Resource Languages speech corpus in comparison with other state-of-the-art systems in the spoken term retrieval task. Based on the results, it is inferred that a 7% improvement in the hit ratio and a 26% reduction in the false alarm ratio was achieved.**

***Index Terms*—pattern similarity, spoken term discovery, keyword discovery, heuristic pattern match, spoken term detection, spoken content indexing, spoken term retrieval.**

I. Introduction

The Spoken Content Retrieval (SCR) task aims to retrieve the occurrence of similar spoken terms from the corpus given a spoken query. In the conventional approach, the SCR task was achieved by converting speech into text using Automatic Speech Recogniser (ASR) and text-based matching was carried out to retrieve similar spoken content. Such an approach demands a large volume of annotated spoken content to train the ASR. Further, it demands a language expert and time to prepare the annotations. The performance of the retrieval task relied on the ASR performance. In reality, a large volume of spoken content piled in the online repositories without annotation is unable to participate in the SCR task. Furthermore, a speech corpus belonging to the unwritten languages (zero-resource) or very little annotated content (low-resource) becomes challenging for the SCR task.

An unsupervised SCR task overcomes the resource constraint by discovering the pattern similarities directly from the speech signal without additional linguistic resources. Such a pattern-matching-based retrieval technique captures the similarity between the time-aligned representation of the speech signals. Dynamic Time Warping (DTW) is one of the well-known matching techniques [1] that captures the similarity between speech signals by identifying the optimal wrap path. Despite feasibility, a challenge in the DTW techniques is the global alignment problem [2]. The DTW technique aims to optimise the wrapping path globally by capturing the overall alignment between two spoken documents. However, in reality, the match may occur in a portion of the documents, creating a challenge for the DTW approach. The segmental DTW approach [3] overcomes the challenge by dividing the speech signal into a fixed duration of segments, and similarity detection was carried out in the fixed region. Though the global alignment problem was addressed in the segmental DTW approach, identifying the segment size introduces a new challenge. Moreover, the variabilities that arise due to the speaker, environment and language-specific changes further increase the complexity of the pattern-matching task.

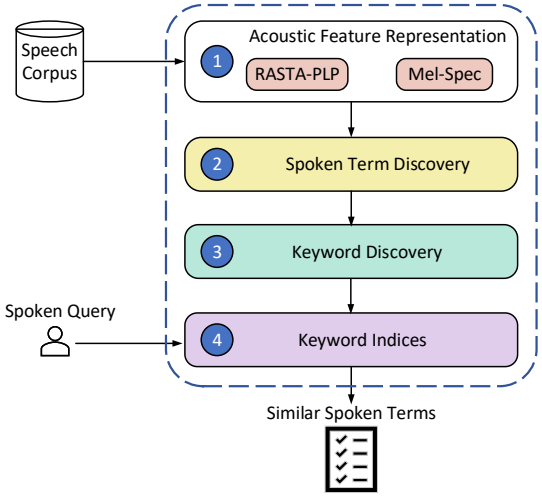

Figure 1. The schematic view of the overall approach.

In our approach depicted in Figure 1, we aim to

Sudhakar is with Advanced Technology Development Centre, Indian Institute of Technology, Kharagpur-721302, India.

Sreenivasa Rao and Pabitra Mitra are associated with Department of Computer Science and Engineering,Indian Institute of Technology, Kharagpur-721302, India.

overcome the aforementioned drawbacks and achieve the unsupervised SCR in four stages. At first, the spectral representation of the speech signal that emphasises the spoken content irrespective of the speaker variability was identified. In the second stage, the similarity between the two spoken contents was discovered. At this stage, the pattern match was computed using the Heuristic Pattern Match (HPM) algorithm that overcomes the variability challenges. In the next stage, the similarity discovered was grouped and indexed using our proposed keyword discovery algorithm. Finally, given a spoken query, the relevant spoken content was retrieved using the indices. In our approach, the SCR task was accomplished directly from the acoustic feature representation itself without additional information, and it is well-suitable for the zero-resource scenario. The contribution of this article is summarised as follows:

- The spoken term discovery was achieved by computing the pattern matches at the acoustic feature level using speaker-independent spectral representation. The HPM algorithm captures the appropriate matches and accomplishes the discovery task.
- The spoken term indexing was achieved using our keyword discovery approach that groups similar terms and identifies the appropriate index for each group.
- Given a spoken query, the SCR task was accomplished by matching the acoustic feature representation of the spoken query with the indices, and similar spoken terms were retrieved.

Further, the article was organised as follows. Section II outlines the related works carried out towards the SCR tasks. Section III discusses the acoustic feature representation that emphasises the speaker-independent spoken content representation. Section IV details the proposed spoken term indexing and retrieval technique to achieve the SCR in the zero-resource scenario. The experimental studies carried out towards the SCR task were presented in Section V. Section VI concludes the article with further scope for research.

II. Related works

The Spoken Content Retrieval task in the zero-resource or low-resource scenario was achieved through (i) the acoustic feature representation that emphasises the spoken content irrespective of the variabilities and (ii) the similarity matching task that captures the resemblance among the spoken terms. In this section, we summarise the related works towards the SCR task into two broader categories: (i) acoustic feature representation and (ii) similarity matching techniques.

In view of the acoustic feature representation, Mel-frequency cepstral coefficients (MFCCs) [4], [5], frequency domain linear prediction cepstral coefficients [6], and perceptually linear prediction cepstral coefficients [7] were used directly in the similarity matching task. Alternatively, the posterior representations of the MFCCs obtained from the Gaussian Mixture Model (GMM) were employed [8], [9] to overcome the speaker variabilities. In [10], the posterior representation of the perceptually linear prediction cepstral coefficients were used to achieve the discovery task. Instead of GMM, an Artificial Neural Network-based latent representation [11] obtained from the MFCCs was utilised to capture the spoken term matches. Similarly, in [12], [13], the Deep Neural Network (DNN) based posterior representations obtained from frequency domain linear prediction cepstral coefficients and perceptually linear prediction cepstral coefficients were employed to detect the spoken term matches. The aforesaid approaches do not require any additional information to obtain the acoustic feature representation. In contrast, the DNN models trained with resourceful languages were used to extract the acoustic feature representations of the low-resource languages. The phoneme-posteriorgram vectors [14] obtained from resourceful languages are used to accomplish the similarity-matching task in the resource-constrained scenario. Similarly, the bottle-neck features [5], [15], [16], [17] obtained from the DNNs using resourceful languages were also used as feature representations. In consideration of the resource-sharing approach, the amount of information shared across the languages become a constraint for the approaches.

In view of the similarity matching technique, two different approaches: (i) DTW-centric and (ii) Template matching techniques, were broadly used to accomplish the matching task. In the DTW-centric approaches, segmental-DTW was used [18], [19], [8] to capture the similarities within the limited duration of the speech signal. In segmental DTW, the speech signal was split into the fixed duration of multiple segments and similarity was computed within the segment. Alternatively, a probabilistic approach [20] with the DTW-matching technique was used to accomplish the SCR task in a zero-resource scenario. In contrast to the DTW-centric approaches, the template matching approach aims to capture the similarities from the acoustic feature representation itself. In [4], the spoken term similarities were captured between acoustic feature representations at the syllable level by mapping the variable length feature vectors into a fixed size. Further, the clustering technique was applied to the fixed-size vectors, and spoken terms were grouped based on their similarities. In the same way, in vector space model [21], variable length acoustic features are mapped to a fixed size using a Locality Sensitive Hashing technique and similarity was computed using the fixed dimension vectors. The graph clustering approach [13] computes the similarity by establishing the relationship between the segments of spoken content. In their approach, spoken content was split into segments, and each segment was represented as a node. Further, the same nodes were identified based on their affinity and a link was established among them. Similarly, in [22], variable length feature vectors are mapped into fixed dimensional space, and embedded segmental K-means clustering was used to compute the similarities. In contrast, an image processing-based angle histogram technique [9] was used to capture

the spoken term similarities in an unsupervised way. A diagonal similarity-based spoken term matching was achieved using the affinity kernel propagation approach in the zero-resource scenario [23].

In consideration of all the techniques specified, the main challenge is to handle the variabilities that occur at the acoustic feature representation as well as at the similarity matching task. The acoustic feature representation emphasises the spoken term similarities irrespective of the other artefacts, such as gender, language, and speaker-specific changes to be determined. Moreover, the similarity matching task should be capable of handling the variabilities and identifying the spoken term matches appropriately. The DTW-centric approach was very sensitive to the variabilities that occur in natural speech. Due to that, a lot of false alarms are produced during the matching task. The template matching approaches operate based on the fixed dimensional space to group the spoken terms, whereas the mapping space will not be the same for all languages. Further, a small change in the feature representation also creates an impact on the similarity-matching task. In the proposed approach, we aim to overcome the variability challenges in four stages. In the first stage, the speaker-independent acoustic feature representation was achieved using RASTA-PLP (Relative spectral-perceptual linear prediction) spectrogram and speaker-normalised Mel-spectrogram representations. In the second stage, the similarity between spoken documents was discovered using the HPM approach. The HPM approach handles the variabilities during the similarity matching task and captures the spoken term matches. In the next stage, the discovered similarities were grouped and indexed using our keyword discovery approach. Finally, the SCR task was achieved using the index discovered against a spoken query. Based on the evaluation, it was observed that the proposed approach reduces significant false alarms in the spoken term discovery task and improves the SCR performance.

III. Acoustic feature representation

The acoustic feature representation aims to emphasise the spoken content similarity by eliminating the speaker, environment and language-specific variabilities that exist in natural speech. In the proposed approach, the RASTA-PLP spectrogram and Mel-spectrogram based representations were employed to address the variability challenges.

The RASTA-PLP spectrogram [24] was computed based on analysing the speech signal using the RASTA filter specified in Eq. 1 in the frequency domain.

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

The RASTA-PLP spectrogram[1] representation is very less sensitive to speaker variabilities, and hence we used this for the similarity matching task. Figure 2 depicts the spoken content correlation obtained based on the RASTA-PLP spectrogram as feature representation computed from two different speakers' spoken content. Figure 2 (a) depicts the relevance between the same spoken content computed based on the RASTA-PLP spectrogram. A diagonal similarity in the principle diagonal region indicates a high correlation for the same content. Figure 2 (b) depicts the absence of correlation among different spoken content. Based on the figure, it is observed that the RASTA-PLP spectrogram captures the spoken content similarity across the speakers. Hence, it is considered one of the acoustic feature representations for the spoken term similarity detection task.

The Mel-Spectrogram based acoustic feature representation was obtained based on the Deep Convolutional Encoder-Decoder (DCE) [25] by disentangling the speaker-specific characteristics. An 80-dimensional speaker-independent Mel-Spectrogram, further denoted as Mel-Spec$_{norm}$, was obtained by separating the speaker-specific characteristic from the spoken content using the DCE network[2]. Figure 3 depicts the spoken content separation and speaker characteristic mapping from multiple sources to a target speaker. Based on the figure, it is inferred that the source speaker characteristics are mapped to a target speaker for the same spoken content. Similarly, it is viable to normalise the speaker-specific characteristics of the entire speech corpus; hence variability challenges are reduced. Finally, the speaker normalised Mel-spectrogram (Mel-Spec$_{norm}$), was obtained by exposing all acoustic features to the trained DCE network conditioning a specific target speaker.

IV. Unsupervised spoken document indexing and retrieval

The spoken document indexing task aims to identify the similar spoken terms that occur in the corpus and organise them appropriately. The spoken content retrieval task focus on retrieving similar terms given a spoken query. The spoken term indexing and retrieval task join together to achieve the SCR task in a zero-resource scenario. In the proposed approach, the SCR task was accomplished in three stages: (A) Spoken term discovery, (B) Keyword discovery and indexing and (C) Spoken content retrieval. At first, the spoken term similarity that exists in the corpus was discovered based on the acoustic feature representation using the HPM approach. In the second stage, the potential spoken terms (referred to as keywords) are identified based on the pattern similarity exhibited and grouped using the keyword discovery algorithm. For each group, an index was identified automatically to represent the group. Finally, given a spoken query, the match was computed based on the similarities between the query and the indices discovered.

[1]https://www.ee.columbia.edu/~dpwe/resources/matlab/rastamat/

[2]https://github.com/sudhakar-pandiarajan/heuristic

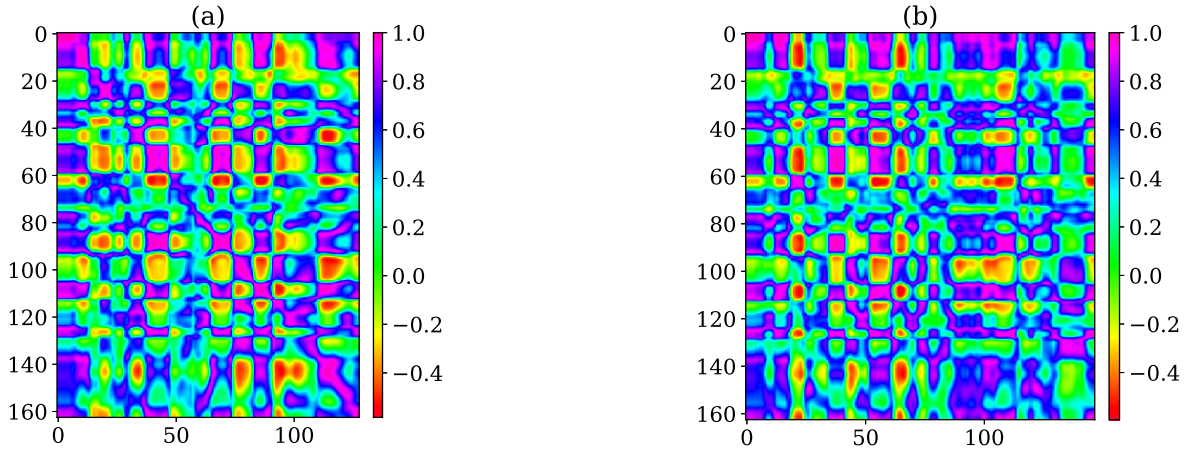

Figure 2. depicts the correlation between the RASTA-PLP spectrogram of (a) the same spoken content uttered by two different speakers and (b) different spoken content.

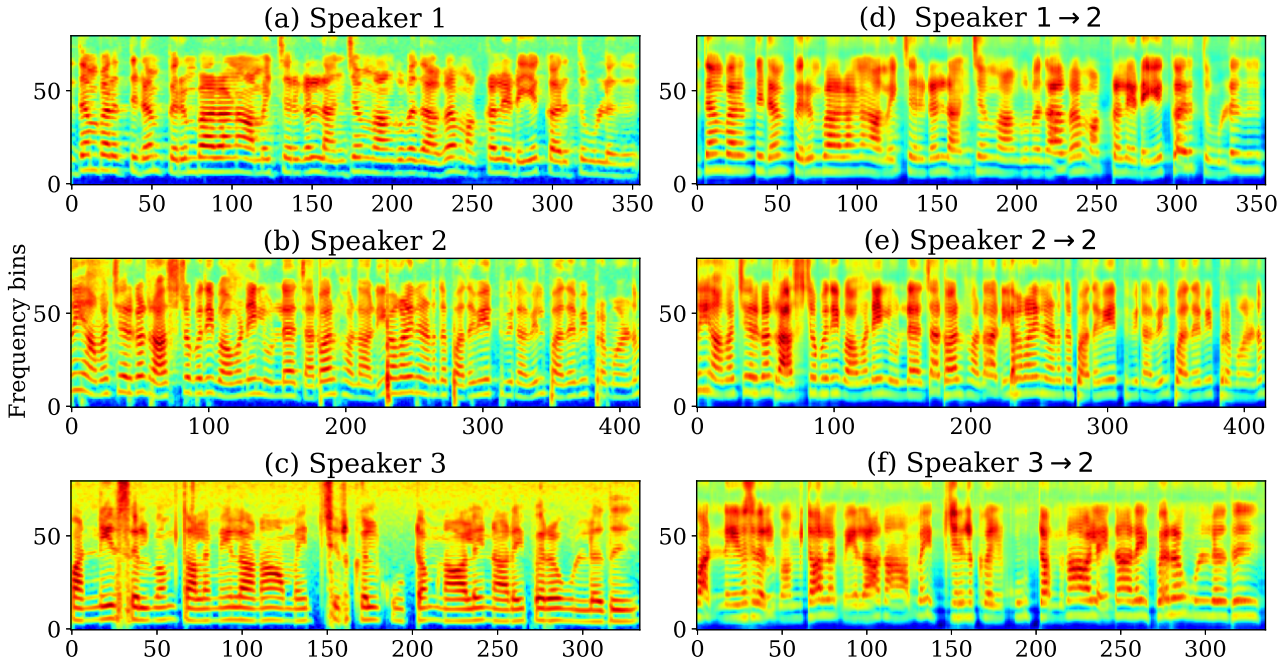

Figure 3. Speaker independent Mel-spectrogram representation achieved using the DeCoDE network. Figure (a), (b) and (c) depicts the Mel-spectrogram of the spoken content uttered by speakers 1, 2 and 3, respectively. (d) depicts the speaker's conversion from 1 to 2. Similarly, (e) and (f) depict the conversion from speaker 2 to 2 and 3 to 2, respectively.

A. Spoken Term Discovery

In view of the spoken term discovery task, the motivation is to capture the spoken term similarity by analysing the pattern matches at the acoustic feature level. The HPM approach [25] was used in the discovery task to identify similar terms. The motivation for the heuristic approach is to capture all similarities that exist between the spoken documents without constraining the length of the similarity propagation at the diagonal region. The match between two spoken terms was identified as a diagonal pattern that propagated for a sequence of time intervals. However, due to the variabilities in the natural speech, the diagonal pattern may occur in a non-contiguous manner (multiple segments of variable length), creating a challenge for the similarity detection task. The heuristic approach [25] addresses the challenges and captures the spoken term matches from the acoustic feature representation itself.

Figure 4 depicts the matches obtained between two spoken documents uttered in Tamil. The documents D^j and D^i contains the spoken content "இந்த போராட்டத்தில் அமைச்சர்கள் பங்கு" and "அமைச்சர்கள் அறிக்கை ஒன்றை வெளியிட்டனர்", respectively. From the figure, it is inferred that the diagonal pattern match occurs in the upper diagonal of the similarity matrix (Figure 4 (c)) for the spoken term match "அமைச்சர்கள்", transliterated to "Amaicargal, /a/ /m/ /ai/ /c/ /a/ /r/ /g/ /a/ /l/" in English. Figure 4 (c) highlights the matched region in the similarity matrix with rectangle boxes in red colour. The

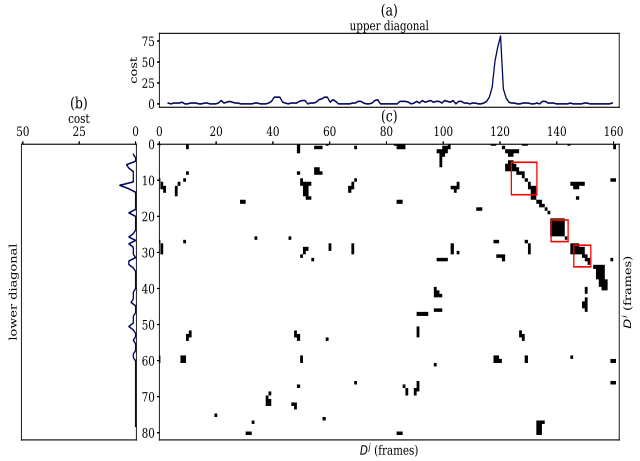

Figure 4. depicts the heuristic cost and diagonal similarity propagation. (a) and (b) depicts the heuristic cost obtained for the upper and lower diagonal, respectively. (c) depicts the diagonal similarity propagation. The rectangle region in red colour indicates the significant diagonal match above the threshold λ.

upper and lower diagonal cost obtained from the similarity matrix was plotted in Figure 4 (a) and (b). Figure 4 (b) highlights the significant peak as a spoken term match in the heuristic cost plot. Further, we also validate that the proposed HPM approach captures multiple matches in the diagonal region to qualify a spoken term match. Finally, the diagonal cost that is higher than the threshold λ is considered as a spoken term match and a link was established between spoken terms $D^i_{p,q}$ and $D^j_{r,s}$. The spoken term match pair ($D^i_{p,q}$, $D^j_{r,s}$) indicates the match between the document i, frames p to q and document j, frames r to s, respectively. Similarly, the HPM approach was applied to all document pairs, and a list of similarity pair information was computed for the keyword discovery task.

B. Keyword discovery and spoken document indexing

The keyword discovery process aims to group the spoken term matches and identifies the representative (denoted as index) for each group. During the discovery process, the matches identified in IV-A were clustered and indexed as key-value pairs. The key-value pair is represented as a function mapping $f_{key} : D^i_{p,q} \rightarrow M$, and $M \in \phi \cup \{D^j_{r,s}\}$ indicating that the key $D^i_{p,q}$ containing the spoken term match of document i from frames p to q, matches to a set M containing document j matched with frames r to s. The keyword discovery algorithm 1 groups the spoken terms and achieves indexing in multiple stages. At first, the discovered document pairs are assigned to a key-value pair by key as one of the document information and value as a document set containing another document. This phenomenon establishes a relationship between the document matches. For example, the spoken term match between the document pair ($D^1_{42,52}$,$D^2_{12,22}$) is denoted as $D^1_{42,52} \rightarrow \{D^2_{12,22}\}$. In the example, $D^1_{42,52}$ act as key and the set $\{D^2_{12,22}\}$ act as a associated document matches.

Algorithm 1 Keyword Discovery

Require: document pairs $\{(D^1_{p,q}, D^2_{r,s})...(D^i_{t,u}, D^j_{v,w})\}$
1: Map all document pairs as key-value pairs.
2: $K_{prime} \leftarrow \{\}$ ▷ Pseudo keywords dictionary
3: Compute key from the document pairs
4: **repeat**
5: for each key-value pair $D^i_{p,q} \rightarrow M_1$ do
6: **if** $D^i_{p,q}$ not in K_{prime} key list **then**
7: add $D^i_{p,q}$ to the K_{prime}.
8: **else** ▷ key exists
9: add documents M_1 with the key $D^i_{p,q}$ in K_{prime}.
10: **end if**
11: **until** all document pairs
12: Compute the key overlap in the same document
13: **for all** keys $D^i_{p,q}$,.., $D^i_{u,v} \in K_{prime}$ **do**
14: get document set $M_1 \leftarrow D^i_{p,q}$
15: get document set $M_2 \leftarrow D^i_{u,v}$
16: **if** $p, q \cap u, v$ **then** ▷ overlap
17: **if** $(u \geq p)$ and $(v \leq q)$ **then** ▷ middle
18: add documents M_2 with the M_1.
19: discard $D^i_{u,v}$
20: **else if** $(u < p)$ and $(v > q)$ **then** ▷ complete
21: add documents M_1 with the M_2
22: discard $D^i_{p,q}$
23: **else if** $(u < p)$ and $(v > p)$ **then** ▷ beginning
24: create new key $D^i_{u,q}$
25: $D^i_{u,q} \leftarrow M_1 \cup M_2$
26: discard $D^i_{p,q}$ and $D^i_{u,v}$
27: **else if** $(u < q)$ and $(v > q)$ **then** ▷ end
28: create new key $D^i_{p,v}$
29: $D^i_{p,v} \leftarrow M_1 \cup M_2$
30: discard $D^i_{p,q}$ and $D^i_{u,v}$
31: **end if**
32: **end if**
33: **end for**

In the next step, for all document pairs, if any one of the keys already exists, then the document set is merged with the existing document set M, indicating that the document has similarity. For example, the key-value pairs $D^1_{11,20} \rightarrow \{D^2_{15,25}\}$ and $D^1_{11,20} \rightarrow \{D^4_{82,92}\}$ contains the key similarity. Therefore, the merged key-value pair is established as $D^1_{11,20} \rightarrow \{D^2_{15,25}, D^4_{82,92}\}$. In the next stage, keyword discovery was achieved by eliminating the duplicate keys and merging the spoken terms together. At this stage, the spoken terms that are relevant to the key are grouped based on the locality information available with the key. There are four possible key overlaps (middle, complete, beginning and end) that occur during the keyword discovery that was merged according to the locality information. Figure. 5 depicts the possible overlaps that occur in the same document with different frame intervals. In all cases, the keyword discovery algorithm (Algorithm 1) addresses the overlaps and redefines the keys. After the merging task, the key-value pair represents a key and a set of documents associated with the key

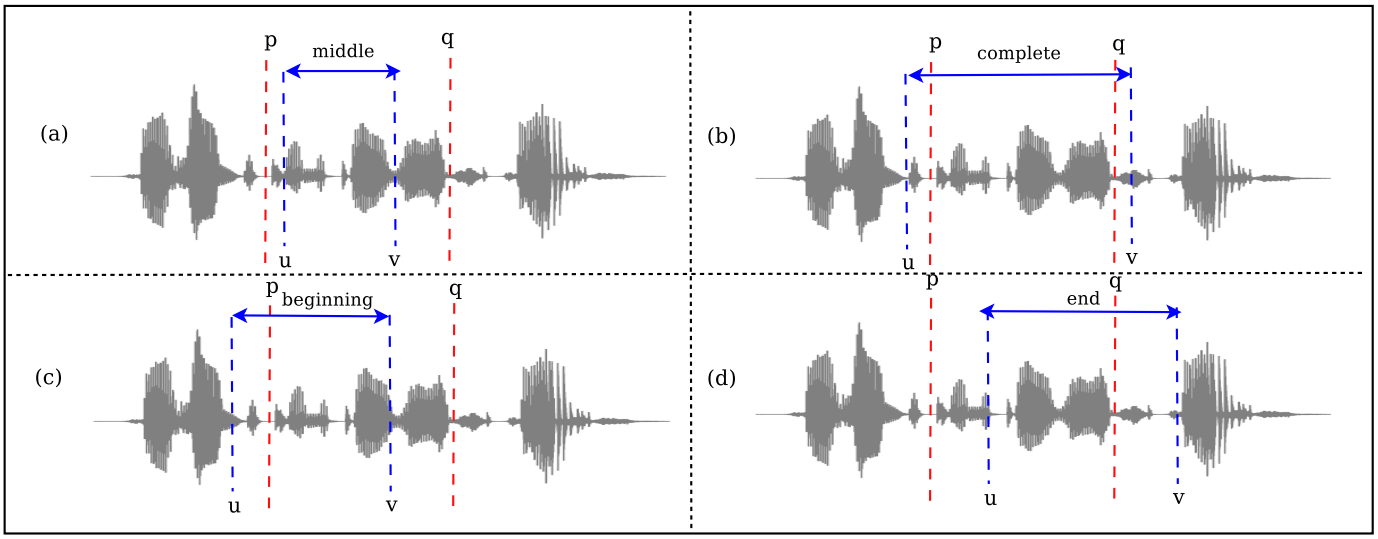

Figure 5. Possible key frame overlap during the keyword discovery task. (a) Overlap at the middle. Similarly, (b), (c) and (d) indicates the overlap in complete, beginning and end frame positions, respectively.

based on their similarity. Based on the evaluation, we infer that most of the key indices match with the ground truth keywords; hence, we label the proposed algorithm as keyword discovery and refer to the indices as keywords. Finally, during the retrieval task, the acoustic feature representation of the index was matched with the spoken query to retrieve similar terms. In contrast to the conventional pattern matching approaches that directly compute the match between the query and all spoken terms that exist in the corpus, the proposed approach removes the non-similarity pairs (outliers) in the keyword discovery process and retains only similar terms. As a result, it reduces the false alarms significantly and improves the retrieval results.

C. Query-by-example spoken term retrieval

The spoken term retrieval was achieved by searching similar spoken terms through the indices discovered. During the retrieval process, the acoustic feature representation of the spoken query $q \in Q$ and the spoken term index $D^i_{p,q} \in D$ were compared to capture the similarities. The HPM approach was used to compute the heuristic cost ($H_{cost}(.)$) between the query and spoken term. Further, the cost value obtained based on the HPM approach was normalised with the query length as specified in the Eq. (2). Finally, the match score $M = \{M|M > \beta\}$ between a query and spoken term has arrived. The threshold β act as a lower bound for the match detection and avoid insignificant matches.

$$M(q, D^i_{p,q}) = \frac{2 \times H_{cost}(q, D^i_{p,q})}{|q| \times |q+1|} \quad (2)$$

During the retrieval, the match score M was computed for all indices to find the similarity. If a significant match score ($M > \beta$) was obtained for an index, then the match score was computed for all members associated with the index to identify the spoken term matches. In the proposed approach, the entire retrieval process was achieved using the indices discovered by the keyword discovery algorithm.

V. Results discussion

The performance of the proposed approach was evaluated in three stages: (i) spoken term discovery, (ii) keyword discovery and (iii) spoken term retrieval. The hit, miss and false alarm ratios were used to measure the performance during the spoken term discovery stage. In view of the keyword discovery, the efficacy was measured based on (a) the keyword indices, (b) clustering quality and (c) the coverage of the corpus. The retrieval process was measured based on the spoken terms retrieved given the spoken queries. For all evaluations, the Microsoft Low-Resource Language (MSLRL) speech corpus was used.

A. Dataset

The performance of the proposed approach was evaluated using MSLRL speech corpus, released in the Interspeech-18 low-resource ASR challenge event [26]. The overall MSLRL speech corpus consists of ≈50 hours of spoken content from Gujarati, Telugu and Tamil languages. However, the corpus was curated for the keyword spotting task [27], and the annotation at the word level was released for a set of documents. Table I lists the statistics of the evaluation data considered in the experiments. The corpus comprises read and conversational modes of speech uttered by both male and female speakers. All speech files maintain a uniform sampling rate of 16 kHz with 16-bit resolution. Moreover, the occurrences of repeated words (referred to as keywords (KW_{act})) were analysed, and the frequency information was obtained. The length of the KW_{act} (in orthographic representation) spans between 3 to 20 for all languages, and the frequency of the KW_{act} occurrences span between 2 to 25. The dataset for evaluation was selected based on the availability of the ground

Table I
DATASET USED FOR THE EVALUATION TASK. # DOCS. SPECIFIES 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 DOCUMENTS.

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

truth information without any bias to create a real-time scenario.

B. Performance Metrics

The performance of the proposed approach was measured using three sets of performance indicators corresponding with (i) spoken term discovery, (ii) keyword discovery, and (iii) retrieval tasks. The spoken term discovery was measured using hit, miss and false alarm ratios [28]. The hit ratio in Eq. (3) indicates the fraction of spoken term (T) matches that are discovered by the proposed approach. Miss ratio in Eq. (4) indicates the fraction of matches that are not discovered. The false alarm ratio in Eq. (5) shows the fraction of non-similarity matches discovered by the algorithm. $|T_{NT\ trails}|$ indicates the number of non-target spoken terms exposed.

$$P_{hit} = \frac{|\ T_{act} \bigcap T_{detect}\ |}{|T_{act}|} \quad (3)$$

$$P_{miss} = 1 - P_{hit} \quad (4)$$

$$P_{fa} = \frac{|T_{fa}|}{|T_{NT\ trails}|} \quad (5)$$

In view of keyword discovery, keyword precision (KW_{prec}), NED (Normalised Edit Distance) and Coverage were used to measure the performance. The KW_{prec} in Eq. (6) measures the fraction of actual keywords that overlap with the keyword indices discovered (KW_{disc}). The NED score in Eq. (7) evaluates the clustering quality by computing the grapheme similarity between the keyword index and the spoken terms belonging to that cluster. In Eq. (7), the $Dist(.)$ function computes the *Levenshtein* distance between an index and spoken term based on the ground truth grapheme representation. The coverage mentioned in Eq. (8) measures the fraction of spoken terms that are discovered in comparison to the number of spoken terms that exist in the corpus.

$$KW_{prec} = \frac{|KW_{act} \bigcap KW_{disc}|}{|KW_{act}|} \quad (6)$$

$$NED = \sum_{\forall KW_{disc}} \frac{Dist(KW_{disc}, T)}{max(|KW_{disc}|, |T|)} \quad (7)$$

$$coverage = \frac{|T_{disc} \bigcap T_{all}|}{|T_{all}|} \quad (8)$$

The retrieval task was measured based on the quality of the retrieved documents against the spoken query. Mean Average Precision (MAP) mentioned in Eq. (9) was measured based on the average precision of the retrieved results against a spoken query for multiple trials.

$$\begin{aligned} precision(D^i) &= \frac{hit}{hit + false\ alarm} \\ avg.\ precision(query) &= \frac{\sum_{\forall k} precision(D^k)}{|D^k|} \\ MAP(Q) &= \frac{\sum_{\forall q \in Q} avg.\ precision(q)}{Q} \end{aligned} \quad (9)$$

The MAP score measures the outcome of the ranked results for different queries. Each metric measures the quality of the system performance based on spoken term discovery, keyword discovery and spoken term retrieval tasks.

C. Parameter selection

The performance of the similarity matching task relied on two parameters α and λ. The parameter α emphasises the cosine similarity between two acoustic feature vectors that are beyond the threshold for match consideration. The parameter λ determines the minimum temporal alignment (in frames) that exists between two sequences of feature vectors. By thresholding both α and λ, the system efficacy shall be verified. Figure 6 depicts the ROC curve [30] obtained for α and λ values in terms of hit and false alarm ratios. The objective is to identify the best parameter that maximises the hit ratio and minimises the false alarm ratio. In Figure 6, the plot region was divided into four quadrants originating at the centre, and each quadrant was named I, II, III and IV. The quadrant I represent the parameters that achieve the hit ratio in the range [0.5,1] and the false alarm ratio in the range [0,0.5]. Similarly, quadrants II, III and IV are defined. The parameters that are grouped within or nearest to the quadrant I achieve the best performance because it maximises the hit ratio and minimises the false alarms ratios. Accordingly, the parameters $\alpha = 0.99$ and $\lambda = 9$ obtained by the Mel-Spec$_{norm}$ achieve the higher hit ratio and lower false alarm ratio. Similarly, the parameters $\alpha = 0.98$ and $\lambda = 12$ obtained by the RASTA-PLP spectrogram maximise the hit ratio and minimises the false alarm ratio. Hence the identified threshold values are fixed for the respective features, and experiments were carried out further.

D. Experiments

The performance of the proposed approach was evaluated at different stages: (i) spoken term discovery, (ii) keyword discovery and (iii) spoken term retrieval stages.

In view of the spoken term discovery task, the proposed approach was expected to discover similar spoken terms in the corpus without any additional information. The Mel-Spec$_{norm}$ and RASTA-PLP spectrogram representations were used in the discovery task, and the results are reported in Table II. Based on the table, it is inferred that at least 50% hit rate was achieved across features

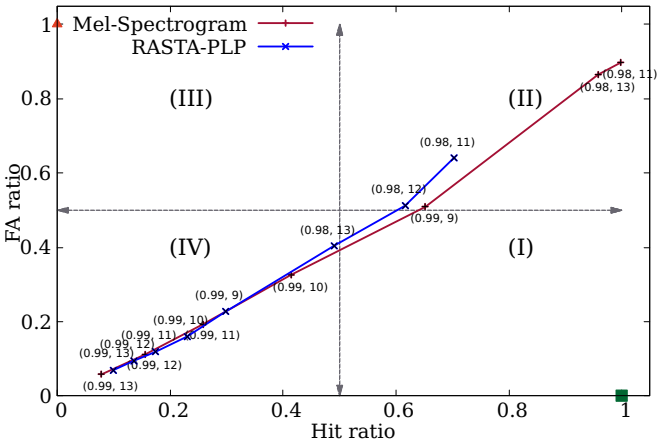

Figure 6. depicts the hit ratio and false alarm ratio produced by the proposed approach on various (α, λ) values.

Table II
EXPERIMENT RESULTS OF THE SPOKEN TERM DISCOVERY TASK. T_{act} SPECIFIES THE GROUND TRUTH SPOKEN TERM MATCHES. T_{detect} REPRESENTS THE DISCOVERED SPOKEN TERMS. FA INDICATES THE FALSE ALARM.

Language	T_{act}	T_{detect}	Hit	Miss	FA	P_{hit}	P_{miss}	P_{fa}
RASTA-PLP								
Gujarati	373796	792448	209700	164096	582748	0.561	0.439	0.0015
Telugu	16363	51098	9012	7351	42086	0.551	0.449	0.0007
Tamil	39339	170292	24620	14719	145672	0.626	0.374	0.0010
Mel-Spec$_{norm}$								
Gujarati	373796	734510	185777	188019	548733	0.497	0.503	0.0014
Telugu	16363	41477	8129	8234	33348	0.497	0.503	0.0006
Tamil	39339	154866	20575	18764	134291	0.523	0.477	0.0009

and languages. This phenomenon indicates that the proposed approach captures the spoken term similarities using both RASTA-PLP and Mel-Spec$_{norm}$. Further, we verified that the similarity information was captured across speakers and languages. Moreover, the highest hit rate was achieved in Tamil using RASTA-PLP as an acoustic feature representation. In comparison among features, RASTA-PLP has achieved a better hit rate than the Mel-Spec$_{norm}$. However, the Mel-Spec$_{norm}$ representation reduces the false alarms in Gujarati, Telugu and Tamil in comparison to RASTA-PLP. The reduction of false alarm indicates that the spoken term similarities captured were better using Mel-Spec$_{norm}$ than the RASTA-PLP.

In view of keyword discovery, first, we analysed the results of the keyword indices using KW_{prec}. Table III projects the results of the keyword indices performance with the actual keywords obtained from the corpus. From Table III it is inferred that the number of keys discovered in comparison with the actual keywords (KW_{act}) was high. This is viable in the unsupervised scenario because of the different types of matches that occur between the spoken terms. However, the keyword precision (KW_{prec}) score indicates that at least 62% of keywords are indexed by the proposed approach using RASTA-PLP. The maximum score of 77.5% was obtained for Tamil using Mel-Spec$_{norm}$ representation, indicating the retrieval was viable using the discovered indices.

Table III
STATISTICS OF THE KEY DISCOVERABILITY ACHIEVED BY BOTH RASTA-PLP AND MEL-SPEC$_{norm}$ ACROSS LANGUAGES.

Language	# KW_{disc}	# KW_{act}	KW_{prec}
RASTA-PLP			
Gujarati	15636	1995	0.620
Telugu	7399	944	0.695
Tamil	12548	734	0.745
Mel-Spec$_{norm}$			
Gujarati	12477	1995	0.698
Telugu	9667	944	0.748
Tamil	11819	734	0.775

In the second stage, we evaluated the cluster characteristics with respect to the indices discovered. During the evaluation, the grapheme representation of the key indices and the spoken terms associated with the key index was obtained from the ground truth information. The NED score was computed between the index and each term using the grapheme representation as specified Eq. (7). Figure 7 (a) and (b) depicts the NED score distribution of each index obtained based on RASTA-PLP and Mel-Spec$_{norm}$, respectively, for Tamil. From the figure, it is observed the NED score obtained from Mel-Spec$_{norm}$ representation (see Figure 7(b)) was dense towards the origin in comparison to the RASTA-PLP (7(a)). This scenario showcases that the NED score obtained by the Mel-Spec$_{norm}$ has less variability, i.e. High similarity was observed at the grapheme level in the NED score. Meanwhile, the RASTA-PLP spectrogram also shows the capability of grouping the spoken terms with the indices, whereas the cluster representation was better in the Mel-Spec$_{norm}$ based representation.

In the third stage, the coverage of the keyword discovery was evaluated. Based on the locality information of the spoken term associated with the indices, the ground truth term was obtained. The fraction of coverage was computed based on Eq. (8), and the results are presented in Table IV. Based on the table, it is inferred that at least

Table IV
KEYWORD DISCOVERABILITY BY THE PROPOSED APPROACH.

Language	T_{all}	T_{disc}	$Coverage$
RASTA-PLP			
Gujarati	11998	7532	0.627
Telugu	3924	1044	0.733
Tamil	5283	2756	0.521
Mel-Spec$_{norm}$			
Gujarati	11998	7492	0.624
Telugu	3924	1039	0.735
Tamil	5283	2828	0.535

52% of the spoken terms are discoverable by the RASTA-PLP spectrogram in Tamil. The maximum coverage was obtained in Telugu using Mel-Spec$_{norm}$. In both features,

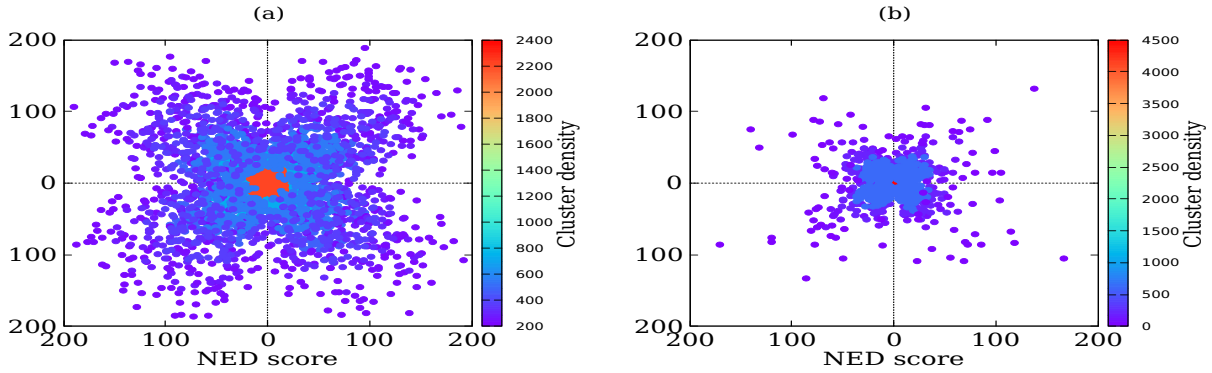

Figure 7. NED score of the key indices against associated spoken terms. The cluster characteristics of (a) RASTA-PLP spectrogram and (b) Mel-Spec$_{norm}$.

a similar performance was observed, indicating that both capture a similar amount of spoken terms. However, across the languages, a 16% and 17% improvement was observed in Telugu compared to Gujarati using RASTA-PLP and Mel-Spec$_{norm}$, respectively. In consideration of keyword discovery, clustering and coverage results, both the RASTA-PLP and Mel-Spec$_{norm}$ have contributed to grouping the spoken content in an unsupervised way using keyword discovery algorithm 1.

In view of the retrieval task, the acoustic feature representation of both spoken query q and indices were compared to identify the matches. For each query $q \in Q$, the match score M was computed with all indices. The match score M was computed based on the threshold β, where $M > \beta$. An empirical analysis was done to obtain an optimal β value based on the hit and false alarm ratios. Based on the performance, the $\beta \geq 1$ value for Mel-Spec$_{norm}$ and $\beta \geq 0.75$ for RASTA-PLP were identified as threshold and used further in the retrieval task. Table V showcases the performance of the retrieval task using the keyword discovery approach. From the table, it is observed that the maximum and minimum hit ratio of 52.4% and 31.6% was obtained using Mel-Spec$_{norm}$ and RASTA-PLP spectrogram as acoustic representation in Telugu. A similar trend was observed across the languages. However, the P_{fa} shows that the false alarm introduced by the Mel-Spec$_{norm}$ was slightly higher than the RASTA-PLP features. Further, it is inferred that when the hit ratio increases, it also increases the false alarms. Hence, deciding the optimal threshold value (β) have an impact on the system performance. In view of the miss ratio, the Mel-Spec$_{norm}$ outperforms the RASTA-PLP across languages. The quality of the retrieved spoken documents is evaluated based on the MAP score. The performance of the MAP score obtained by Mel-Spec$_{norm}$ gives a maximum confidence of 47.1% in Telugu. Similarly, a better trend was observed across the languages using Mel-Spec$_{norm}$ as feature representation in comparison with the RASTA-PLP features. However, RASTA-PLP still achieves 31.8% confidence in the Telugu language, indicating that the spoken term retrieval is viable by both Mel-Spec$_{norm}$ and RASTA-PLP across languages.

In addition, we evaluated the proposed approach with other state-of-the-art systems [32], [33] using MSLRL corpus. In the CNN-QBE approach [32], the acoustic feature representation was obtained from the pre-trained neural network model [34]. Further, the similarity between the query and the spoken term was measured based on the trained CNN network. The Feats-QBE approach [33] uses the pre-trained neural network representation to generate the acoustic features, and segmental DTW was used to capture the similarities between query and spoken term. In both approaches, the optimal parameters recommended by the authors are retained, and experiments were conducted against the MSLRL corpus. Table VI shows the results of the spoken content retrieval across languages. Based on the results, it is inferred that the proposed approach using RASTA-PLP spectrogram achieves the maximum P_{hit} in Tamil with 1.5% improvement in comparison to the Feats-QBE approach in Telugu. The minimum hit ratio was obtained in Gujarati using the CNN-QBE approach. In comparison with the proposed approach using RASTA-PLP, the Feats-QBE approach has slightly better performance in Gujarati and Telugu. This is due to the search space where the proposed approach aims to capture the similarities from the discovered spoken terms, whereas other approaches directly capture the similarity from the speech corpus. Meanwhile, the P_{fa} score obtained by RASTA-PLP discriminates the proposed approach by reducing the score significantly in comparison with other methods. A 24.6% reduction (average) in false alarms was observed in comparison with the CNN-QBE approach across languages. The reduction in false alarms was feasible due to the spoken term clustering strategy accomplished by the keyword discovery algorithm. During the discovery tasks, the proposed approach discards the outliers and retains only valid matches. Hence the false alarms were reduced during the retrieval. Similarly, the performance of the proposed approach was measured using Mel-Spec$_{norm}$ as an acoustic feature across the systems. The Mel-Spec$_{norm}$ achieves a 7%, 14.2% and 14.3% improvement in the hit ratio in comparison with the Feats-QBE approach. Meanwhile, a 26%, 30% and 36.1% reduction was observed in the false alarm ratio. The

Table V
Performance of spoken term retrieval over a set of query trails. # Queries indicates the number of spoken queries. $\#T_{occ}$ specifies the number of actual spoken term occurrences.

Language	Feats	# Queries	# T_{occ}	Hit	Miss	False alarm	$P_{hit}\uparrow$	$P_{miss}\downarrow$	$P_{fa}\downarrow$	MAP
Gujarati	RASTA-PLP	137	394	135	259	264574	0.343	0.657	0.124	0.323
	Mel-Spec$_{norm}$	137	394	170	224	297256	0.431	0.569	0.174	0.397
Telugu	RASTA-PLP	63	187	59	127	33693	0.316	0.684	0.072	0.318
	Mel-Spec$_{norm}$	63	187	98	89	55167	0.524	0.476	0.091	0.471
Tamil	RASTA-PLP	118	406	161	245	170916	0.397	0.603	0.116	0.387
	Mel-Spec$_{norm}$	118	406	198	208	168714	0.488	0.512	0.121	0.458

Table VI
Performance evaluation of the proposed approach with other state-of-the-art methods.

Approach	Feats	Language	$P_{hit}\uparrow$	$P_{miss}\downarrow$	$P_{fa}\downarrow$	MAP
CNN-QBE	BNF	Gujarati	0.295	0.705	0.35	0.062
		Tamil	0.329	0.671	0.33	0.031
		Telugu	0.374	0.626	0.37	0.039
Feats-QBE	BNF	Gujarati	0.361	0.639	0.44	0.33
		Tamil	0.345	0.655	0.421	0.375
		Telugu	0.382	0.618	0.452	0.342
Proposed	RASTA-PLP	Gujarati	0.343	0.657	0.124	0.323
		Tamil	0.397	0.603	0.116	0.387
		Telugu	0.316	0.684	0.072	0.318
	Mel-Spec$_{norm}$	Gujarati	0.431	0.569	0.174	0.397
		Tamil	0.488	0.512	0.121	0.458
		Telugu	0.524	0.476	0.091	0.471

improvement in hit ratio and reduction in the false alarm indicates that the proposed approach has better-spoken term detection capability using Mel-Spec$_{norm}$ as acoustic feature representation rather than the bottleneck features obtained from the Feats-QBE approach. Furthermore, the heuristic pattern match approach discovered the spoken terms appropriately; hence, the results are improved in the retrieval task. In comparison with the MAP score in both RASTA-PLP and Mel-Spec$_{norm}$ features, the Mel-Spec$_{norm}$ performed better by 7.4%, 7.1% and 15.3% improvement in Gujarati, Tamil and Telugu, respectively.

In consideration of all experiments, the following conclusions are arrived. (i) The proposed spoken term discovery task captures the appropriate spoken term matches. (ii) The keyword discovery task used the discovery information and grouped the spoken terms together based on their similarity. (iii) The indices obtained from the keyword discovery act as an index for the speech corpus without any additional resources. (iv) Spoken term detection was viable through the discovered indices. (v) Speaker-independent Mel-spectrogram representation has a better hit ratio than the RASTA-PLP spectrogram. (vi) The proposed approach directly uses the speech corpus without additional information and stands language independent.

VI. Summary and Conclusion

In this article, we demonstrated spoken term indexing and retrieval through spoken term discovery and keyword discovery techniques. During the pre-processing phase, two different acoustic feature representations: RASTA-PLP and Mel-spectrogram, were used to capture the speaker-independent spoken content information. Further, the spoken term discovery achieved by the heuristic pattern match algorithm captures the similarities in the acoustic feature representation and reduces the false matches. The keyword discovery technique groups the spoken term similarities discovered and generate the indices for the retrieval task. Finally, given a spoken query, the spoken term similarities were directly discovered from the indices. The performance evaluation of the proposed approach with other state-of-the-art systems indicates that a 7% gain in the hit ratio and a 26% reduction in the false alarm ratio was achieved. In future, we aim to reduce the false alarms by adopting a contextual learning approach to capture the similarities. Further, reducing the false alarms by capturing the latent representation of the acoustic feature introduces a new direction of research.

References

[1] *Dynamic Time Warping*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 69–84.

[2] C. Myers, L. Rabiner, and A. Rosenberg, "Performance tradeoffs in dynamic time warping algorithms for isolated word recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 6, pp. 623–635, 1980.

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

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

[5] 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.

[6] G. Mantena and K. Prahallad, "Use of articulatory bottle-neck features for query-by-example spoken term detection in low resource scenarios," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 7128–7132.

[7] A. Jansen, K. Church, and H. Hermansky, "Towards spoken term discovery at scale with zero resources," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[8] 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.

[9] K. K. R. and K. Sreenivasa Rao, "A novel approach to unsupervised pattern discovery in speech using convolutional neural network," *Computer Speech and Language*, vol. 71, p. 101259, 2022.

[10] Chun-an Chan and Lin-shan Lee, "Model-based unsupervised spoken term detection with spoken queries," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1330–1342, 2013.

[11] 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.

[12] J. Li, X. Wang, and B. Xu, "An empirical study of multilingual and low-resource spoken term detection using deep neural networks," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

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

[14] B. Ludusan, A. Caranica, H. Cucu, A. Buzo, C. Burileanu, and E. Dupoux, "Exploring multi-language resources for unsupervised spoken term discovery," in *2015 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*. IEEE, 2015, pp. 1–6.

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

[19] H. Tulsiani and P. Rao, "The iit-b query-by-example system for mediaeval 2015," in *MediaEval*, 2015.

[20] O. Räsänen and M. A. C. Blandón, "Unsupervised discovery of recurring speech patterns using probabilistic adaptive metrics," 2020.

[21] A. Jansen and V. D. Benjamin, "Efficient spoken term discovery using randomized algorithms," in *2011 IEEE Workshop on Automatic Speech Recognition Understanding*, 2011, pp. 401–406.

[22] 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.

[23] P. Sudhakar, K. S. Rao, and P. Mitra, "A novel zero-resource spoken term detection using affinity kernel propagation with acoustic feature map," *SN Computer Science*, vol. 4, no. 3, p. 310, 2023.

[24] H. Hermansky and N. Morgan, "Rasta processing of speech," *IEEE transactions on speech and audio processing*, vol. 2, no. 4, pp. 578–589, 1994.

[25] S. P, S. R. K, and P. Mitra, "Query-by-example spoken term detection for zero-resource languages using heuristic search," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 2023.

[26] B. M. L. Srivastava, S. Sitaram, R. K. Mehta, K. D. Mohan, P. Matani, S. Satpal, K. Bali, R. Srikanth, and N. Nayak, "Interspeech 2018 low resource automatic speech recognition challenge for indian languages." in *SLTU*, 2018, pp. 11–14.

[27] 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.

[28] L. J. Rodriguez-Fuentes and M. Penagarikano, "Mediaeval 2013 spoken web search task: system performance measures," *n. TR-2013-1, Department of Electricity and Electronics, University of the Basque Country*, 2013.

[29] 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 *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 2014, pp. 560–567.

[30] T. Fawcett, "An introduction to roc analysis," *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.

[31] L. Yujian and L. Bo, "A normalized levenshtein distance metric," *IEEE transactions on pattern analysis and machine intelligence*, vol. 29, no. 6, pp. 1091–1095, 2007.

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

[33] N. San, M. Bartelds, M. Browne, L. Clifford, F. Gibson, J. Mansfield, D. Nash *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.

[34] T. Schultz, N. T. Vu, and T. Schlippe, "Globalphone: A multilingual text & speech database in 20 languages," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 8126–8130.