I. INTRODUCTION

Current approaches to spoken term discovery rely on different languages, some of those languages even considering the linguistic constraints. With the increasing availability of spoken documents in language resources, there is a growing need for unsupervised methods of information extraction. An important topic, there is a need for unsupervised methods of information extraction. An important topic, there is a need for unsupervised methods of information extraction. Among such algorithms, spoken term discovery allows the extraction of word-like units (terms) directly from the continuous speech signal, in an unsupervised manner and without any knowledge of the language at hand. Since the performance of any downstream application depends on the goodness of the terms found, it is relevant to try to obtain higher quality automatic terms. In this paper we investigate whether the use input features derived from of multi-language resources helps the process of term discovery. For this, we employ an open-source phone recognizer to extract posterior probabilities and phone segment decisions, for several languages. We examine the features obtained from a single language and from combinations of languages based on the spoken term discovery results attained on two different datasets of English and Xitsonga. Furthermore, a comparison to the results obtained with standard spectral features is performed and the implications of the work discussed.

Keywords: spoken term discovery; posteriorgrams; multi-resource methods; phone recognition; automatic speech segmentation

II. RELATED WORK

The use of Gaussian posteriorgram representations for unsupervised speech pattern discovery was introduced in [1, 2, 3, 4]. Applications employing these algorithms have appeared, ranging from topic segmentation [5] to document classification [6] or spoken document summarization [7].

Among such algorithms, spoken term discovery allows the extraction of word-like units (terms) directly from the continuous speech signal, in an unsupervised manner and without any knowledge of the language at hand. Since the performance of any downstream application depends on the goodness of the terms found, it is relevant to try to obtain higher quality automatic terms. In this paper we investigate whether the use input features derived from of multi-language resources helps the process of term discovery. For this, we employ an open-source phone recognizer to extract posterior probabilities and phone segment decisions, for several languages. We examine the features obtained from a single language and from combinations of languages based on the spoken term discovery results attained on two different datasets of English and Xitsonga. Furthermore, a comparison to the results obtained with standard spectral features is performed and the implications of the work discussed.
The current work employs an open source spoken term discovery system, called MODIS [13], based on the discrete cosine transform (DCT) features and Split Temporal Context (STC) algorithm, together with Temporal Pattern Recognition (TPR) features and Split Temporal Context (STC) algorithm, as described in [4]. The functioning of the system is based on a combination of classical spectral features (MFCC) and a 1 sec long temporal vector of critical band energies [17], extracted using the HTK toolkit [15], for all the evaluation datasets in Section 4 and obtained results in Section 5.

The remainder of this paper is structured as follows: Section 2 presents the system used for spoken term discovery. Section 3 describes the features employed in the discovery system, called MODIS [13]. The functioning of the system is based on a combination of classical spectral features (MFCC) and a 1 sec long temporal vector of critical band energies [17], extracted using the HTK toolkit [15], for all the evaluation datasets in Section 4 and obtained results in Section 5.

The paper concludes with some final remarks and future work directions.

II. DISCUSSION

For this task, we used a state vector (SV) representation. Each input speech frame is converted to a binary value, 1 indicating that at that time instance a particular phone was found by the recognizer and 0 signaling the opposite case.

A Cyclic Rhythm of Words (CRW) is defined as a sequence of binary values, 1 indicating that at that time instance a particular phone was found by the recognizer and 0 signaling the opposite case.

5. The paper concludes with some final remarks and future work directions.
A. Experimental settings

For the spoken term discovery experiments we varied the sequence similarity score threshold. As optimization metric we used also the best performance attained by the systems: "int" (intermittent noise), "Spk" (speaker noise) and "pau" (silent pause) [23]. We use in this paper the datasets released with the official evaluations of the NCHLT Speech Recognition Challenge. As features for term discovery we use both MFCCs and phoneme vectors. As mentioned in Section 3, we employed phoneme vectors to discard noise and other nonterm speech events from speech files. The three types of non-term speech files, except for the EN and CZ systems: "int" (intermittent noise), "Spk" (speaker noise) and "pau" (silent pause) [23]. We use in this paper the datasets released with the official evaluations of the NCHLT Speech Recognition Challenge. As features for term discovery we use both MFCCs and phoneme vectors. As mentioned in Section 3, we employed phoneme vectors to discard noise and other nonterm speech events from speech files. The three types of non-term speech files, except for the EN and CZ systems: "int" (intermittent noise), "Spk" (speaker noise) and "pau" (silent pause) [23].

We use in this paper the datasets released with the official evaluations of the NCHLT Speech Recognition Challenge.

As features for term discovery we use both MFCCs and phoneme vectors. As mentioned in Section 3, we employed phoneme vectors to discard noise and other nonterm speech events from speech files. The three types of non-term speech files, except for the EN and CZ systems: "int" (intermittent noise), "Spk" (speaker noise) and "pau" (silent pause) [23].

For the spoken term discovery experiments we varied the sequence similarity score threshold. As optimization metric we used also the best performance attained by the systems: "int" (intermittent noise), "Spk" (speaker noise) and "pau" (silent pause) [23]. We use in this paper the datasets released with the official evaluations of the NCHLT Speech Recognition Challenge. As features for term discovery we use both MFCCs and phoneme vectors. As mentioned in Section 3, we employed phoneme vectors to discard noise and other nonterm speech events from speech files. The three types of non-term speech files, except for the EN and CZ systems: "int" (intermittent noise), "Spk" (speaker noise) and "pau" (silent pause) [23].
of the matching process, and it rewards systems having both a high precision and a high recall.

Fig. 1. Feature extraction module

B. Evaluation

Besides the two datasets, the challenge offers on its website (www.zerospeech.com) also an evaluation software based on the measures introduced in [25], toolkit which was used in this paper. The only difference between the implemented measures and the ones proposed in the previous paper is that all metrics are computed on the entire corpus (as opposed to the discovered set), in order to be able to compare systems that cover different parts of the evaluation set.

Several measures are implemented in the evaluation package, ranging from metrics on the quality of the matching process, to those characterizing the clustering stage and some which compute natural language processing metrics, like token and type F-scores. We focus here on the matching metrics, as they indicate the performance of the DTW search. We have chosen this measure because all the other measures are directly affected by the matching quality and we expect that a good first matching stage would also translate into better performance downstream.

Precision, recall and F-score are computed from the set of discovered motif pairs, with respect to all matching substrings in the dataset. Precision is defined as being the proportion of discovered substrings pairs that belong to the list of gold pairs, weighted by the type frequency. Similarly, recall is computed as the proportion of gold motif pairs discovered by the algorithm. Matching F-score is defined as the harmonic mean between precision and recall. For a formal definition of these measures, the reader is invited to consult [25].

C. Results

The results obtained are presented in terms of matching F-score, computed over all speakers in the respective datasets. Since the optimal value of the DTW threshold was set on the sample set, part of the English dataset, we are particularly interested in the performance obtained by the system on a different language. We expect that good results on another language, not seen by the system, will further validate the generalizability of the approach. We report results for the matching precision, recall and F-score and for all the features/combinations of features we tested. By doing so, we expect to have a better insight into the role that each feature plays in the term discovery process.

The matching F-score results on the two tested languages are illustrated in Figure 2. It shows the performance of our baseline (MFCC) on the first column and that of the systems using either posteriorgrams or phoneme vectors, computed with the different single language acoustic models or combinations of them, as input features.

Fig. 2. Matching F-score obtained using the posteriorgrams and the phoneme vectors features (individual and combination of languages), on the English and Xitsonga datasets.
I. Malioutov, A. Park, R. Barzilay, and J. Glass, “Making sense of...”

Y. Zhang and J. Glass, “Towards multi...”

A. Muscariello, G. Gravier, and F. Bimbot, “Unsupervised motif...”

Xitsonga dataset, although the DTW thr... we can see an important increase in performance also on the... When comparing the performance of the different features,... contrary to its usefulness in spoken term detection... As future research directions we plan to extend the current...


