Exploring multi-language resources for unsupervised spoken term discovery

Bogdan Ludusan[1], Alexandru Caranica[2], Horia Cucu[2], Andi Buzo[2], Corneliu Burileanu[2], Emmanuel Dupoux[1]

[1]Laboratoire de Sciences Cognitives et Psycholinguistique
EHESS/ENS/CNRS
Paris, France
bogdan.ludusan@ens.fr

[2]Speech and Dialogue (SpeeD) Research Laboratory
University Politehnica of Bucharest
Bucharest, Romania
alexandru.caranica@speed.pub.ro

***Abstract*—With information processing and retrieval of spoken documents becoming an important topic, there is a need of systems performing automatic segmentation of audio streams. 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-language resources;

I. Introduction

With the increasing availability of spoken documents in different languages, some of those languages even considered under-resourced in the speech community, there is a growing need for unsupervised methods of information extraction. An appropriate method for this task, spoken term discovery systems identify recurring speech fragments from the raw speech, without any knowledge of the language at hand [1].

Current approaches to spoken term discovery rely on variants of dynamic time warping (DTW) to efficiently perform a search within a speech corpus, with the aim of discovering occurrences of repeating speech (further called *terms* or *motifs*) [1, 2, 3, 4]. Applications employing automatically discovered terms have quickly appeared, having a wide focus, ranging from topic segmentation [5] to document classification [6] or spoken document summarization [7]. Besides the immediate applications it can have in languages with little or no resources, spoken term discovery can also have relevance to cognitive models of infant language acquisition [8].

In order for the obtained terms to be a viable source of information for any downstream application, they need to be of good quality and to sufficiently cover the target corpus. For this reason, the speech research community has worked towards improving the unsupervised term discovery process through different methods. Among the various approaches proposed, we mention the use of linguistic information in the input features [9, 4], the optimization of the search process [10], or the introduction of linguistic constraints during DTW search [11].

Since spoken term discovery works in an unsupervised manner, the extraction of informative features is an important aspect. Zhang and colleagues [9] were the first to explore the use of Gaussian posteriorgram representations for unsupervised discovery of speech patterns. They demonstrated the viability of using their approach, by showing that it provides significant improvement towards speaker independence. The investigation into the use of posteriorgrams for spoken term discovery was extended in [4], where the authors employed two types of posteriorgrams: both supervised and unsupervised ones, the former being trained either on the target language or on a different language. They showed that for one of their system settings, the posteriorgrams always outperformed the Mel Frequency Cepstral Coefficients (MFCC) features, while for the other setting only the target language supervised posteriors brought improvements over the MFCC baseline.

Taking advantage of the existence of open source recognition systems and the availability of acoustic models trained on different languages [12], we would like to build upon the study conducted by Muscariello and colleagues [4]. In this work we investigate a larger range of phone-based posteriorgrams, as well as combinations of posteriorgrams coming from different languages. Furthermore, we use the phoneme recognizer output to build an additional feature, a binary phoneme feature vector. The latter feature defines the presence (value 1) or absence (value 0) of a phoneme at a certain time instant, as returned by the speech recognizer. We compare the results given by a spoken term discovery system employing these linguistically-enhanced features with those obtained with an identical system using classical spectral features.

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 section, as well as the extraction process. We illustrate the evaluation datasets in Section 4 and obtained results in Section 5. The paper concludes with some final remarks and future work directions.

II. Spoken term discovery

The current work employs an open-source spoken term discovery system, called MODIS [13], based on the systems proposed in [4]. The functioning of the system follows the so called seed discovery principle [14], i.e. to search for matches of a short audio segment in a larger segment, with the search being performed by means of a segmental variant of the dynamic time warping (DTW) algorithm. In this framework the shorter segment is called seed, while the larger one is called buffer.

The algorithm inspects the acoustic sequences present in the buffer to assess whether it contains any repetition of the seed and a matching decision is taken by comparing the DTW score of the path with a DTW similarity threshold. If the computed score is lower than the threshold, the algorithm considers that a match was found. In that case, the seed will be extended and the process repeated using the longer seed, until the dis-similarity between the segments reaches the set threshold. When that happens, the term candidate is stored in the motif library, provided it has passed any length constraints imposed by the system. The algorithm continues parsing the speech looking for matches with respect to the motif library. If no match is found with respect to the motifs in the library, the DTW search process described previously is repeated. When further matches of the same term are found, the corresponding cluster model is updated accordingly. Once the corpus has been parsed in its entirety, found motifs are compared to each other in term of their overlap and overlapping elements are merged into one single term.

The algorithm has several important parameters that must be set: the *seed size*, the minimum stretch of speech matched against the buffer, the minimum term *size* the algorithm will find, the *buffer size* in which the seed is searched and the *similarity threshold* ϵ_{DTW}. The latter influences the level of similarity between the members of the same term class: the lower the threshold, the more similar the terms will be. The choice of this parameter has to be a compromise between a small number, but highly homogeneous terms, and a larger number of terms with a higher heterogeneity.

MODIS was designed to work with several types of input features. For this it implements two different distances: the Euclidean distance, generally used together with spectral representations, and the distance defined in (1), when posterior probabilities features are employed. The terms involved in (1), represent the two feature vectors for which the distance is calculated (*a* and *b*) and their length (*N*).

$$d(a,b) = -\log(\sum_{i=0}^{N}(a_i \cdot b_i)) \qquad (1)$$

III. Features

We decided to use for our baseline system MFCC features, a standard spectral representation in speech applications. The features were extracted using the HTK toolkit [15], for all the datasets used in this paper. The speech signal was analyzed using 25ms long frames with a 10ms frame rate and the original wider frequency band (16 kHz) was used for feature extraction. HTK was configured to compute 39 features (MFCC + Energy + Deltas + Accelerations) per frame. Critical bands' energy was obtained in conventional way.

Next, we considered phone-based posteriors as features. A phone posteriorgram is defined by a probability vector representing the posterior probabilities of a set of pre-defined phonetic classes for a speech frame, with entries summing up to one. By using a phonetic recognizer, each input speech frame is converted to its corresponding posteriorgram representation.

Finally, we use the phoneme recognizer output to build an additional feature, a binary phoneme vector. This vector represents the speech as a sequence of binary values, 1 indicating that at that time instant a particular phone was found by the recognizer and 0 signaling the opposite case.

For this task, we used a state-of-the-art recognition system based on long temporal context from BUT [16]. It uses a hybrid Hidden Markov Model - Artificial Neural Network (HMM/ANN) architecture, together with Temporal Pattern (TRAP) features and Split Temporal Context (STC) optimizations (Left-Right Context). TRAP features were introduced in order to address a particular drawback of the classical spectral features employed in automatic speech recognition systems, i.e. their rapid degradation in performance in realistic communication environments. In the case of TRAP features, a 1 sec long temporal vector of critical band logarithmic spectral energies from a single frequency band is used, to capture the temporal evolution of the band-limited spectral energy in a vicinity of the underlying phonetic class [17]. The latter characteristic of the recognition toolkit used, the STC optimization, aims at processing the two parts of the phoneme independently. The trajectory representing a phoneme feature can be decorrelated by splitting them into two parts, to limit the size of the model, in particular the number of weights in the neural-net (NN). The system uses two blocks of features, for left and right contexts (the blocks have one frame overlap). Before splitting, the speech signal is filtered by applying the Hamming window on the whole block, so that the original central frame is emphasized. The dimensions of vectors are then reduced by means of Discrete Cosine Transform (DCT) and results are sent to two neural networks. The optimal number of DCT coefficients was determined to be 11 [12]. The posteriors from both contexts are, in the final stage, merged, after the front-end neural networks are able to generate a three-state per phoneme posterior model [16].

The recognition system has several acoustic models, trained using data from one of the following four languages: English, with data from the TIMIT corpus [18], and three other languages from the SpeechDat-E corpus [19] (Czech, Russian and Hungarian). We summarize in Table 1 several statistics pertaining to the recognizers employed in this study: *phonemes* represents the number of phoneme classes modelled by the given recognizer, while *NN* is the number of neurons employed in the used neural networks. We also illustrate the system error rate *ERR* of each recognizer, as this measure might offer some insight into the performance obtained with the features derived for the different languages considered. It can be seen that the best performance is attained by the Czech and English systems, followed by the Hungarian recognizer and the Russian system.

TABLE I. BUT RECOGNIZER SYSTEMS USED

Rec. system	phonemes	ERR%	NN
CZ-8k	45	24.24	1500
HU-8k	61	33.32	1500
RU-8k	52	39.27	1500
EN-16k	39	24.24	500

IV. DATA SETS DESCRIPTION

We use in this paper the datasets released with the ZeroSpeech 2015 challenge [20]: an English dataset and one containing a surprise language. The English set contains recordings from the Buckeye corpus [21], while the surprise language, identified as being Xitsonga, one of the eleven official languages of South Africa, had its material drawn from the NCHLT speech corpus of the South African languages [22].

The Buckeye corpus contains spontaneous speech, recorded between an interviewer and an interviewee, discussing issues of local interest. It was completely transcribed orthographically and annotated at the phone- and word-level. While the corpus contains around 38 hours of recordings, coming from 40 speakers, for the challenge only a subset of the corpus was used. It was divided into two datasets: a sample set containing recordings from 2 speakers (one female, one male), totaling almost 2 hours, and an evaluation set more than 10 hours long, with data coming from 12 speakers (six females, six males).

A subset of the NCHLT Xitsonga Speech Corpus was used for the challenge. It contains more than 4 hours of recordings, coming from 24 speakers (12 females and 12 males). The corpus contains read speech, recorded through a smartphone-based interface and it was transcribed and annotated at the word and phone-level.

V. EXPERIMENTS

This section describes the experimental setup, the features involved in spoken term discovery and the evaluation metrics employed, while also illustrating the results obtained in the experiments.

A. Experimental settings

In a first step, the segments given in the output of the recognizer were processed to keep only speech zones, using an embedded voice activity detector module. Intervals for the Voice Activity Module (VAD) module were obtained either from the corpus annotation released with the challenge files, for the evaluation set, or calculated from the output of the phoneme recognizer, for the development set. These speech masks were applied on posteriorgrams, MFCCs and phoneme vectors to discard noise and other non-speech events from speech files. The three types of non-speech tokens in the BUT systems: “int” (intermittent noise), “Spk” (speaker noise) and “pau” (silent pause) [23] were all mapped to silence in our VAD module. The complete feature extraction procedure is illustrated in Fig. 1.

As features for term discovery we use both MFCCs and information coming from a speech recognizer: posteriorgrams and phoneme vectors. As mentioned in Section 3, we employed the BUT recognizer and acoustic models for the following languages: Czech (*CZ*), English (*EN*), Hungarian (*HU*) and Russian (*RU*). Down sampling to 8kHz was necessary for all speech files, except for the EN system, to match the recognizer acoustic models used for phoneme training. Besides extracting posteriorgrams and vectors of phonemes for each individual language, we created two combinations of such features, by concatenating the individual vectors into one, large, super-vector. The two combinations tested are the following: *All*, containing the vectors of all four languages, and *AllE*, obtained by concatenating the CZ, HU and RU models outputs. The latter was tested in order to see the effect of a combination of language posteriorgrams on the English data, without using knowledge from that language. The vector of phonemes feature represents the speech as a sequence of binary values, with 1 indicating that at that time instant, a particular phone was found by the recognizer and value 0 representing the opposite case.

For the spoken term discovery experiments we varied the similarity threshold, while keeping the rest of the parameters constant. The seed length was set to 0.25 s and the minimum term size considered was 0.5 s in order to able to find entire words. The buffer length to 600 s, as some of the materials used here had fewer repetitions of the same word. The model chosen to represent the term clusters was the median model, while self-similarity matrix checking [24] was employed for the matching between the model of the cluster and the seed. Regarding the distances used, the posteriorgram features employed the distance presented in (1), while the MFCCs and the phoneme vectors used the Euclidean distance.

Before proceeding to the experiments, good values for the term discovery similarity threshold had to be determined, for each of the different features employed. Besides the fact that the features used are quite diverse, the term discovery software uses also different distance functions for them. These differences made even more important the finding of a correct setting for the DTW threshold. For this reason, we employed the sample set released with the challenge as a development set, on which we searched for the optimum threshold value, given a certain evaluation metric. As optimization metric we chose the matching F-score [25], as it characterizes the quality of the matching process, and it rewards systems having both a high precision and a high recall.

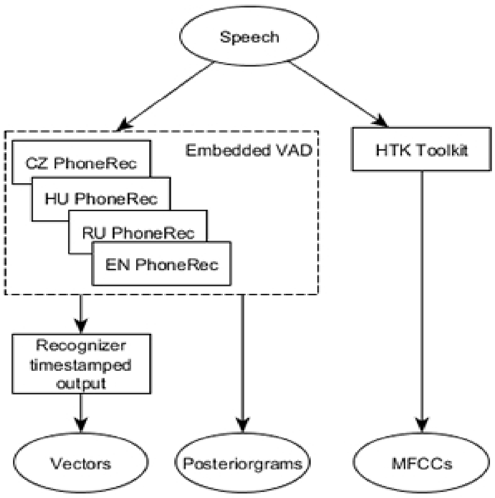

Fig. 1. Feature extraction module

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, weighed 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.

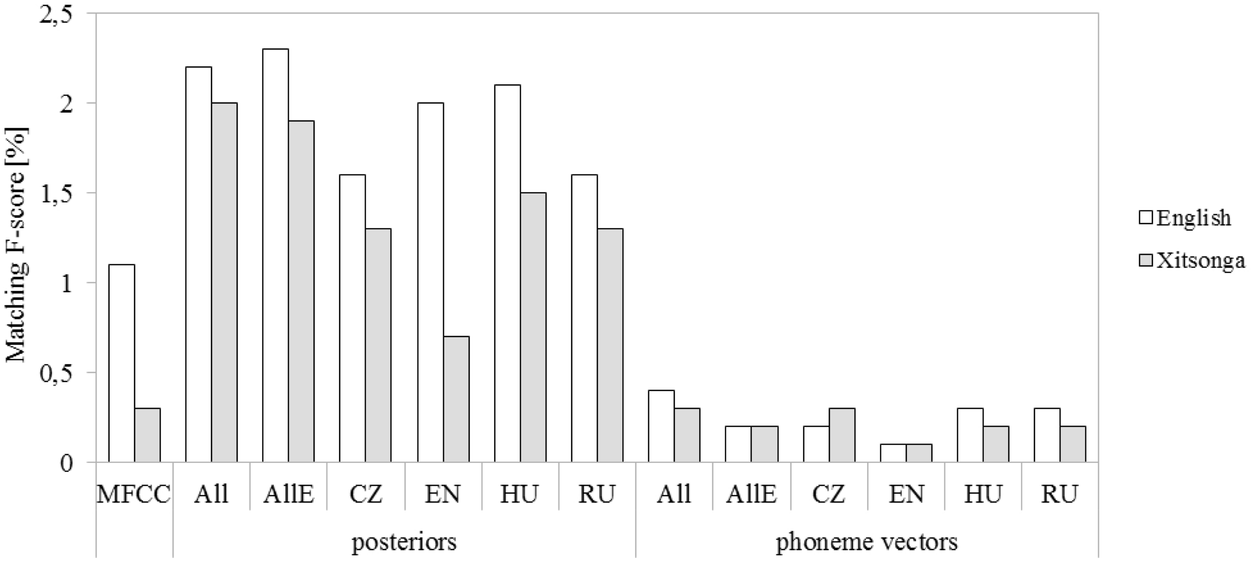

Fig. 2. Matching F-score obtained using the posteriorgrams and the phoneme vectors features (individual and combination of languages), on the English and Xitsonga datasets

When comparing the performance of the different features, we can see a clear advantage of posteriorgrams over MFCCs and phoneme vectors, for both languages. Furthermore, we observe an important increase in performance also on the Xitsonga dataset, although the DTW threshold was set on a totally different language (English). Phoneme vectors instead seem not to have enough discriminative power for spoken term discovery. It shows that the hard decision taken by recognizer introduces a significant amount of error, from which the system cannot recover even when multi-language resources are employed.

Regardless of the feature used (posteriorgrams or phoneme vectors), we can see the advantage of using combined features. These features give either the best metric values or they are close to the best one, being the most consistent ones, overall.

Next, we looked more in detail into the systems employing posteriorgrams as input features. Table II shows the precision, recall and F-score for each individual feature setting, on the two languages. It appears that the systems having in input posteriorgrams of combinations of languages behave better both in terms of precision and recall. Again, for both languages, we obtain either the best performance or close to it, in the case of multi-language posteriorgrams.

TABLE II. MATCHING PRECISION, RECALL AND F-SCORE OBTAINED ON ENGLISH AND XITSONGA WHEN MFCCS (BASELINE) AND POSTERIORGRAMS ARE USED AS INPUT FEATURES (BOLD REPRESENTS THE BEST OVERALL RESULT).

Setting	English			Xitsonga		
	P [%]	R [%]	F [%]	P [%]	R [%]	F [%]
MFCC	1.2	1.1	1.1	6.4	0.2	0.3
CZ	4.5	1.0	1.6	8.6	0.7	1.3
EN	6.1	1.2	2.0	5.7	0.4	0.7
HU	6.3	1.3	2.1	10.4	0.8	1.5
RU	3.6	1.1	1.6	6.9	0.7	1.3
AllE	**7.2**	**1.4**	**2.3**	**12.5**	1.0	1.9
All	4.5	1.5	2.2	8.7	**1.1**	**2.0**

VI. CONCLUSIONS

We have presented here an investigation into the use of multi-language resources for the task of spoken term discovery. Similarly to previous studies employing posteriorgrams as input features for term discovery, we have shown that they improve performance with respect to using MFCCs. We have also explored the use of combined features, by concatenating posteriorgrams coming from different languages and we observed that they generally improve over the single language features. Since individual language features might give good results for one metric and worse for other metrics, one can use the combination of posteriorgrams from different languages for more robust overall results, a desirable trait for a system working on an unknown language.

The second part of our study, exploring the use of phoneme identity information in spoken term discovery, showed that, contrary to its usefulness in spoken term detection [26], this type of information is not sufficient for the current task. Still, the use of combined features seems to give also in this case an overall better performance over single language features.

As future research directions we plan to extend the current study by investigating other language combinations that were not tested here. We also plan to explore the use of posteriorgrams coming from training an acoustic model with the phonemes of several languages, a sort of “universal” acoustic model.

ACKNOWLEDGMENT

Bogdan Ludusan and Emmanuel Dupoux's work was funded by the European Research Council (ERC-2011-AdG-295810 BOOTPHON). It was also supported by the Agence Nationale pour la Recherche (ANR-10-LABX-0087 IEC, ANR-10-IDEX-0001-02 PSL*), the Fondation de France, the École des Neurosciences de Paris, and the Région Île-de-France (DIM cerveau et pensée).

Alexandru Caranica and Horia Cucu`s work has been funded by the Sectoral Operational Programme Human Resources Development 2007-2013 of the Ministry of European Funds through the Financial Agreements POSDRU/159/1.5/S/132395, and POSDRU/159/1.5/S/134398.

Andi Buzo and Corneliu Burileanu’s work was funded by the PN II Programme "Partnerships in priority areas" of MEN - UEFISCDI, through project no. 25/2014.

REFERENCES

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

[2] A. Jansen, K. Church, and H. Hermansky, “Towards spoken term discovery at scale with zero resources,” in Proc. of INTERSPEECH, 2010, pp. 1676–1679.

[3] R. Flamary, X. Anguera, and N. Oliver, “Spoken WordCloud: Clustering recurrent patterns in speech,” in Proc. of International Workshop on Content-Based Multimedia Indexing, 2011, pp. 133–138.

[4] A. Muscariello, G. Gravier, and F. Bimbot, “Unsupervised motif acquisition in speech via seeded discovery and template matching combination,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, no. 7, pp. 2031–2044, 2012.

[5] I. Malioutov, A. Park, R. Barzilay, and J. Glass, “Making sense of sound: Unsupervised topic segmentation over acoustic input,” in Proc. of ACL, 2007, pp. 504–511.

[6] M. Dredze, A. Jansen, G. Coppersmith, and K. Church, “NLP on spoken documents without ASR,” in Proc. of EMNLP, 2010, pp. 460–470.

[7] D. Harwath, T. Hazen, and J. Glass, “Zero resource spoken audio corpus analysis,” in Proc. of IEEE ICASSP, 2013, pp. 8555–8559

[8] A. Jansen, E. Dupoux, S. Goldwater, M. Johnson, S. Khudan-pur, K. Church, N. Feldman, H. Hermansky, F. Metze, R. Rose, M. Seltzer, P. Clark, I. McGraw, B. Varadarajan, E. Bennett, B. Borschinger, J. Chiu, E. Dunbar, A. Fourtassi, D. Harwath, C. Lee, K. Levin, A. Norouzian, V. Peddinti, R. Richardson, T. Schatz, and S. Thomas, “A summary of the 2012 JHU CLSP workshop on zero resource speech technologies and models of early language acquisition,” in Proc. of ICASSP, 2013, pp. 8111–8115.

[9] Y. Zhang and J. Glass, “Towards multi-speaker unsupervised speech pattern discovery,” in Proc. of ICASSP, 2010, pp. 4366–4369

[10] A. Jansen and B. Van Durme, “Efficient spoken term discovery using randomized algorithms,” in Proc. of IEEE ASRU, 2011, pp. 401–406.

[11] B. Ludusan, G. Gravier, and E. Dupoux, “Incorporating prosodic boundaries in unsupervised term discovery,” in Proc. of Speech Prosody, 2014, pp. 939–943.

[12] P. Schwarz, P. Matejka, L. Burget, and O. Glembek, “Phoneme recognizer based on long temporal context,” Speech Processing Group, Faculty of Information Technology, Brno Univ. of Technology, Online Mar. 2015. Available: http://speech.fit.vutbr.cz/en/software/phoneme-recognizer-based-long-temporal-context

[13] L. Catanese, N. Souviraà-Labastie, B. Qu, S. Campion, G. Gravier, E. Vincent, and F. Bimbot, “MODIS: an audio motif discovery software,” in Proc. of INTERSPEECH, 2013.

[14] C. Herley, “ARGOS: Automatically extracting repeating objects from multimedia streams,” IEEE Transactions on Multimedia, vol. 8, no. 1, pp. 115–129, 2006.

[15] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, “The HTK book (Vol. 2)”, Cambridge: Entropic Cambridge Research Laboratory, 1997.

[16] P. Schwarz, “Phoneme recognition based on long temporal context,” PhD thesis, Brno University of Technology, 2009.

[17] H. Hermansky and S. Sharma, ”Temporal Patterns (TRAPS) in ASR of Noisy Speech”, in Proc. ICASSP’99, Phoenix, Arizona, USA, Mar, 1999

[18] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue “TIMIT: acoustic-phonetic continuous speech corpus,” Philadelphia: Linguistic Data Consortium, 1993.

[19] P. Pollák, J. Boudy, K. Choukri, H. van den Heuvel, K. Vicsi, A. Virag, R. Siemund, W. Majewski, P. Staroniewicz, H. Tropf, J. Kochanina, A. Ostroukhov, M. Rusko, and M. Trnka, “SpeechDat (E) - Eastern European telephone speech databases.” in Proc. of Workshop on Very Large Telephone Speech Databases, 2000.

[20] M. Versteegh, R. Thiolliere, T. Schatz, X.N. Cao, X. Anguera, A. Jansen, and E. Dupoux. "The Zero Resource Speech Challenge 2015." In Proc. of INTERSPEECH, 2015.

[21] M. Pitt, L. Dilley, K. Johnson, S. Kiesling, W. Raymond, E. Hume, and E. Fosler-Lussier, “Buckeye corpus of conversational speech (2nd release),” Columbus, OH: Department of Psychology, Ohio State University, 2007.

[22] N. de Vries, M. Davel, J. Badenhorst, W. Basson, F. de Wet, E. Barnard, and A. de Waal, “A smartphone-based ASR data collection tool for under-resourced languages,” Speech Communication, vol. 56, pp. 119–131, 2014.

[23] P. Matjka, P. Schwarz, and P. Chytil, “Phonotactic language identification using high quality phoneme recognition,” in Proc. of Eurospeech, 2005, pp. 2237–2240.

[24] A. Muscariello, G. Gravier, and F. Bimbot, “Zero-resource audio-only spoken term detection based on a combination of template matching techniques,” in Proc. of INTERSPEECH, 2011, pp. 921–924.

[25] 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 Proc. of LREC, 2014, pp.560–567.

[26] A. Buzo, H. Cucu, C. Burileanu, “SpeeD @ MediaEval 2014: Spoken Term Detection with Robust Multilingual Phone Recognition“, in Working Notes Proceedings of the MediaEval 2014 Workshop, Barcelona, Spain, 2014, ISSN: 1613-0073.