Exploring Spoken Term Detection with a Robust Multi-Language Phone Recognition System

Alexandru Caranica[1*],Horia Cucu, Andi Buzo, Corneliu Burileanu
Speech and Dialogue Research Laboratory, University Politehnica of Bucharest, Romania

Abstract: Information processing and retrieval has become a major interest topic in the speech community, with various applications in spoken document retrieval from speech data, like broadcast news, telephone conversations and roundtable meetings to audio query. The ultimate goal of Spoken Term Detection (STD) technology is to allow open vocabulary search over large collections of speech content, to find all of the occurrences of a specified term in a given corpus. For under-resourced languages, STD can only be addressed with Query-by-Example search where the queries are recordings of spoken words. This, coupled with growing web-accessible volumes of audio data, poses a difficult task consisting in finding repetitive patterns, while using as input only the speech signal. In this paper, we investigate whether the use of multi-language resources as input features helps the process of term discovery for under-resourced languages, by phone recognition with a multilingual acoustic model of three languages (Albanian, English and Romanian). The novel Power Normalised Cepstral Coefficients (PNCC) features are used for improved robustness to noise, along with the well know Mel-frequency cepstral coefficients (MFCC) features.

Key Words: spoken term detection, query-by-example search, multi-language resources, speech features, mfcc, pncc, STD evaluation metrics, under resourced languages.

1. INTRODUCTION

With the ever-increasing amounts of vast digital audio data being created and broadcasted daily from various sources, a pressing need exists for intelligent information extraction and retrieval methods, in the speech community. There are various applications for these methods, from document retrieval containing speech data like broadcast news, telephone conversations and roundtable meetings to audio query searches. Many of these spoken documents are in different languages, some of those even considered under resourced in the speech community, hencealso a growing need for an unsupervised method of information extraction and retrieval.

These problems are addressed by spoken term detection (STD) approaches, which identify all of the occurrences of a specified “term” in a given corpus of speech data. For the STD task, a term is considered a sequence of one or more words, and no terms will include more than five words (Patty, 2006). A particular feature that discriminates STD from other ASR-based tasks, such as speech transcription or keyword spotting, is that queries may contain words that are not limited to the system vocabulary. STD systems must cope with these so-called out-of-vocabulary (OOV) words.

Traditionally, most systems used large vocabulary continuous speech recognition tools to produce word transcripts (Mamou, 2007). These transcripts are further indexed and query terms are retrieved from the index. Most of the time, query terms that are not part of the recognizer’s trained vocabulary cannot be retrieved, decreasing the evaluation recall, with a significant drawback that such approaches return no results on queries containing out-of-vocabulary terms (Mamou, 2007). Thus, more advanced systems provide also phonetic transcripts, against which query terms can be matched phonetically. Such systems suffer from lower accuracy, but are a first step towards a language independent method of search.Some of the more advanced systems match phones from multiple language resources to improve the search. Current approaches to spoken term detection rely on variants of dynamic time warping (DTW) algorithm, to efficiently performa search within a given speech corpus anddetect the location of all query occurrences or terms (Park, 2008., Jansen, 2010., Flamary, 2011., Muscariello, 2012).Despite these different approaches, all spoken term discovery systems can be logically broken down and implemented in two phases: indexing and searching(Patty, 2006). In the indexing phase, the system must process the speech data without knowledge of the terms.The extraction of reliable features plays a very important role in this phase, for speech representation. In the searching phase, the system uses the terms and the speech parameterization (features), the index, and optionally the audio to detect term occurrences and their location. Theoretically, a perfect system, with the best methods, would detect the exact locations of all the query occurrences in the audio documents, and would yield no false detections.

In practice, however, acoustic, language and phonetic models are not available for all languages. Such languages are called under-resourced. In this case, it is not possible to process a query in a text form, because, due to the lack of phonetic models, mapping between the pronunciation and the written form of the words are not

*Corresponding author.
E-mail: alexandru.caranica@speed.pub.ro (Alexandru Caranica)

available. Therefore, STD is approached by query-by-example search. This means that queries are given in the form of recording of spoken terms and the task becomes searching for audio queries in audio contents.

In this paper, we attempt to resolve the Spoken Term Detection (STD) problem for under-resourced languages by phone recognition with a multilingual acoustic model. The Power Normalized Cepstral Coefficients (PNCC) features are used for improved robustness to noise.We investigate whether the use of multi-language resources as input features help the process of term detection. Our multilingual acoustic model (AM) is trained with three languages (Albanian, English and Romanian), and we evaluate our system on the ground truth and evaluation metrics proposed by the MediaEval 2014 Multimedia Benchmark Initiative (MediaEval, 2014).

The rest of the paper is organized as follows. Section 2 presents previous and different approaches in the field of information retrieval and audio motif discovery. Section 3 and 4 present the methodology for STD used in detail, along with information about the features and materials used in this paper. In Section 5, we present the experimental results of our tests with the proposed methodology and discuss the results, while Section 6 lists conclusions and future work directions.

2. RELATED WORK

A number of content-based retrieval methods have been explored, including topic detection and tracking, spoken term detection, spoken document retrieval, spoken term discovery and so forth. Research in these directions was supported by multiple evaluation campaigns. In 2006, the U.S. National Institute of Standards and Technology (NIST) created the STD (Spoken Term Detection) evaluation toolkit to facilitate research and development of technology for retrieving information from speech data (Fiscus, 2007). In recent years, numerous workshops hosted benchmarking initiatives to evaluate new algorithms for multimedia access and retrieval, such as MediaEval (MediaEval, 2011-2014), or as special sessions at relevant conferences in the field of speech communication (ZeroSpeech Challenge, InterSpeech 2015, OpenKWS).

In an ideal information retrieval scenario, the end user should be able to perform open vocabulary search and retrieval in any language, over a large collection of spoken documents, in a front-end application, with results being returned in a matter of seconds. For this reason, most of the systems employ some sort of pre-indexing of the speech corpus, prior to search, without the advanced knowledge of the query terms.Thus, a typical STD system is illustrated in Figure 1 and mainly consists of two components: in the pre-indexing phase, a speech recognition subsystem transcribes speech signals into intermediate representations, usually word or sub-word lattices, followed by a detection subsystem that searches for occurrences of the search terms, using a pattern matching or search algorithm (such as DTW). The later subsystem comprises (i) a term detector that searches the indexed content for all potential occurrences of a search term, and (ii) a decision making component that determines if a potential occurrence is reliable enough to be hypothesized as a term match. It is important to note that the recognition subsystem is run only once on the audio database and that the detection subsystem has access only to the decoded content (or lattices), hence the pre-indexing phase.

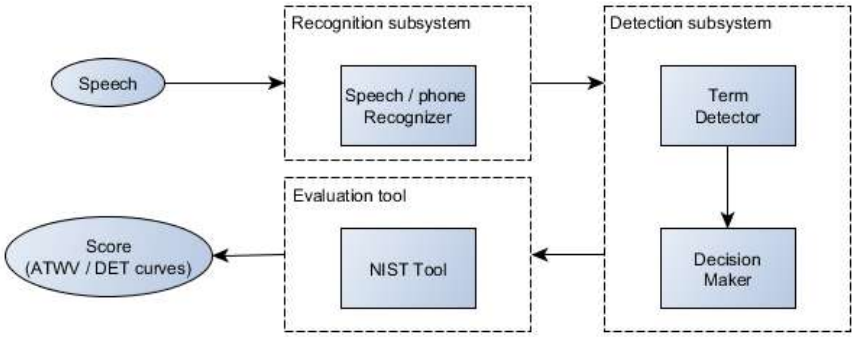

Figure 1 Illustration of a typical STD system, where the US NIST Tool is used to evaluate system performance. Adapted from (Dong, 2012)

Much of the prior work, done to date, focused on languages and domains where transcribed speech and phonetic lexicon resources are widely available.Thus, they relied on large amounts of training data, including recordings (for acoustic modeling) and text data (for language modeling) in the target languages. As such, the best current methods make heavy use of word-based speech recognition during the indexing process to build word lattices. The good accuracy of the Automatic Speech Recognition (ASR) systems for high-resourced languages has also assured a high quality STD.As such, these systems have some constraints, and assume well-trained recognizers

for the input language, with a search vocabulary to be well covered by the language models used during indexing (low Out-Of-Vocabulary for the query terms). Hence, the recent efforts are concentrated mainly on handling Out-Of-Vocabulary (OOV) words for which the pronunciation is unknown and the language model is unavailable (Parlak, 2008., Parada, 2010., Wang, 2010).In a recent paper (Wade, 2009), the authors made use of the classic lattice approach to build possible alternatives from both the query term and search indices, suggesting that lattice representations of search indices and queries can still improve STD performance.

To compensate for the languages where resources are scarce (low resourced languages), many state of the art systems also make use of phonetic search and data fusion techniques. Approaches based on subword units (phones), are widely used nowadays to solve the OOV issue. In this approach, subword representations of search terms are searched for within subword lattices that are generated by a subword-based ASR system. Authors in (Ng, 2000., Burget, 2006., Hori, 2007) made significant work with phonetic units for content based retrieval from speech, through a method of confusion networks applied to phones, outperforming lattice-based methods especially for OOV queries.

Regarding multilingual STD, there are a few previous studies(Lee, 2009., Motlicek, 2010). The first uses an out-of-language module based on confidence measures to detect only the English speech segments. The latter proposes a method for a switch between Chinese and English languages using code-switched lattice-based structures for word/subword units. An alternative solution is to build acoustic and language models that are shared across languages, like (Lin, 2009).

Less work has been done involving methods for speech search by example. In (Murao, 2005), the authors describe a method for example-based query generation for general search.(Buzo, 2013), proposed a query-by-example approach to multilingual Spoken Term Detection for under-resourced languages, based on ASR. The approach overcomes the main difficulties met under these conditions, providing a new method for building multilingual acoustic models with few annotated data. The acoustic models are obtained by adapting well trained phonemes to the ones from the envisaged languages. The mapping is made according to the International Phonetic Alphabet phoneme classification and a confusion matrix. The weighting of query length and alignment spread are incorporated in the Dynamic Time Warping technique to improve the searching method.

Using data fusion techniques to combine results from diverse ASR systems, one can improve robustness across a variety of talkers, channels, environments and target terms.Various Hybrid approaches which fuse word and subword approaches at the lattice level have also been proposed (Yu, 2004., Meng, 2008).

Our study is closely related to works that provide multilingual acoustic models, even though they are mainly used in ASR. All of the methods presented, in this section, show promising performance on the utterance retrieval task. STD remains a challenging task going forward. Unfortunately, state-of-the-art ASR systems are far from being reliable when it comes to transcribing unconstrained speech recorded in uncontrolled environments. Considering the heterogeneous nature of the large spoken databases, it is no surprise that speech retrieval research is mainly about compensating for ASR deficiencies (Can, 2011).

3. METHODOLOGY

This paper presents our approach for the Query-by-Example Spoken Term Detection (QbyE STD) task. We propose a multilingual acoustic modeling method with a scalable Dynamic Time Warping (DTW) search algorithm. To solve the QbyESTD problem for under-resourced languages, we use an indexer with a multilingual phoneme recognizer with acoustic models from three languages: Albanian, English and Romanian. Our final system implementation is based on the architecture proposed by NIST, illustrated in Figure 1.We separate the indexing and the searching modules, rather than searching the corpus directly for each query term, to make the search faster, provided that the indexing methodsimplifies the search (Buzo, 2013).

We use phone recognition for indexing, thus all the speech contents are transformed in strings of phonemes.Asstated in the introduction section, for under-resourced languages where resources are scarce, phone recognition makes an ideal choice.

Two speech feature types are used in this work, to parametrically represent speech: the common Mel Frequency Cepstral Coefficients (MFCC) and the Power Normalized Cepstral Coefficients (PNCC), the later offering improved robustness to noise.The MFCC featuresare widely used and well known, we used them as baseline features.MFCC`s are implemented in the Sphinx Toolkit used for development of the system. They are based on the known variation of the human ear's critical bandwidths with frequency. Filters spaced linearly at low

frequencies and logarithmically at high frequencies have been used to capture the phonetically important characteristics of speech.

The second set of features, Power Normalized Cepstral Coefficients (PNCC, figure 3), are relatively new and their development was motivated by a desire to obtain a set of features for speech recognition that are more robust to acoustical variability, hence they perform better in noisy environments. Their computational complexity is comparable to that of MFCCcoefficients. Major new features of PNCC processing include the use of a power-law non-linearity that replaces the traditional log non-linearity used in MFCC coefficients, a noise-suppression algorithm based on asymmetric filtering that suppress background excitation, and a module that accomplishes temporal masking (Kim, 2012). Experimental results demonstrate that PNCC processing provides substantial improvements in recognition accuracy compared to MFCC and PLP processing for speech, in the presence of various types of additive noise and in reverberant environments, with only slightly greater computational cost than conventional MFCC processing, and without degrading the recognition accuracy that is observed while training and testing using clean speech (Kim, 2012).

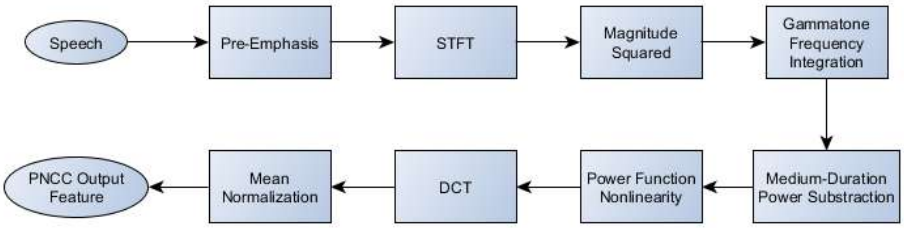

Figure 2 Block diagram of the PNCC feature extraction module

The triangular filter bank used by MFCC is replaced with a gamma tonefilter bank. The novel aspects of the algorithm are the use of a Power Function Nonlinearity (replacing MFCC's log non-linearity) and the use of Medium-Duration Power Bias Subtraction to suppress the effects of background excitation (Kelly, 2010), as shown in Figure 2.It can be seen as a variant on MFCC feature extraction with different stages of the conventional algorithm replaced with auditory motivated elements.

Regarding the Acoustic Model (AM), in our approach, we wanted to compare the effect of using multilingual resources against monolingual models, and in order to achieve this, we built six acoustic models described in Table 1.We start with acoustic models for each language and use the IPA classification for mapping common phones in language models (LM) trained with all data. Mapping common phones is motivated by the high number of phones obtained in AM4, wherephonemes from different languages are trained separately and they are seen as different entities. This allowed us to lower the number of phonemes to 98 for the multilingual model AM5. We used a moderate number of training data for individual languages, to have a balanced training data set among different languages. For comparison, we trained and additional acoustic model for Romanian (AM6), with a big amount of data (64h) and a relatively small set of phonemes (34).

Table 1 Training data

AM ID	Language	LM no. phonemes	Training data [h]
AM1	Romanian	34	8.7
AM2	Albanian	36	4.1
AM3	English	75	3.9
AM4	Multilingual separate phones	145	16.7
AM5	Multilingual common phones (IPA)	98	16.7
AM6	RomanianBig	34	64

The proposed search method uses a Dynamic Time Warping Algorithm (DTW) to align a string (a query) within a given content. Originally, DTW has been used to compare different speech patterns in automatic speech recognition. In fields such as data mining and information retrieval, DTW has been successfully applied to automatically cope with time deformations and different speeds associated with time-dependent data.

The search is not performed on the entire content, but only on a part of it by the means of a sliding window proportional to the length of the query, where both query and contents are string of phonemes. The term is considered detected if the DTW scores above a threshold.In addition to the classical DTW string algorithm, we include, in the distance formula, the effect of query length and DTW match spread. Their effect is weighted in order to find an optimal configuration. The accuracy of the ASR used for indexing plays an important role,

because the searching algorithm must compensate for the rather high Phone Error Rate (PhER), thus it must be robust.

Since the length of the content is usually greater than the length of the query, the comparison is made within a sliding window whose length is proportional to the query length. For each window, the alignment is given a score *s*:

$$s = (1 - PhER) \tag{1}$$

where *s* is a score of similitude.Detection is based on a threshold which is determined empirically.The detection method is refined by introducing a penalizationfor the short queries and the spread of the DTW match. Penalizations are motivated by the assumption that for two queries of different length that match their respective contents by the same phone error rate (PhER), the match of the longer query is more probable to be the right one. The formula for the score *s* is now given by equation (Buzo, 2013):

$$s = (1 - PhER)(1 + \alpha \frac{L_Q - L_{Qm}}{L_{QM} - L_{Qm}})(1 + \beta \frac{L_w - L_S}{L_Q}) \tag{2}$$

whereL_Q is the length of the query, L_{QM} and L_{Qm} are the maximum and minimum values respectively for L_Q, L_W is the window length, L_S is the spread of the DTW match, while α and β are tuning parameters that control the amount of penalization.

4. EVALUATION RESULTS

The development and evaluations datasets used are part of the 2014 QUESST task from Mediaeval 2014 evaluation campaign (http://www.multimediaeval.org/mediaeval2014/). Two separate sets of queries are provided, for development and evaluation, along with a single set of audio files. The set of development queries and the set of audio files are distributed including the ground truth and the scoring scripts. This allowed us to evaluate our results against the metrics proposed by the MediaEval 2014 Multimedia Benchmark Initiative.

The search corpus is composed of around 23 hours of audio in the following 6 languages: Albanian, Basque, Czech, non-native English, Romanian and Slovak, with different amounts of audio per language.TheQUESST 2014 dataset includes 560 development queries and555 evaluation queries, the number of queries per languagebeing more or less balanced with the amount of audio available in the search corpus. All audio files are PCM encoded at 8 kHz, 16 bits/sample, and stored in WAV format (Anguera, 2014).

The metrics used to evaluateQbyE STD performance obtained with different acoustic models on the development data set are shown in Figure 3. Comparison is made using the Maximum Term Weighted Value (MTWV), along with Detection Error Tradeoff (DET) curves.TWV is defined as a weighted combination of the miss and false alarm error rates, averaged over the set of queries, as follows (Fiscus, 2007):

$$TWV(\theta) = 1 - \frac{1}{|Q|}\sum_{\forall q \in Q}(P_{miss}(q,\theta) + \beta \cdot P_{fa}(q,\theta)) = 1 - (P_{miss}(\theta) + \beta \cdot P_{fa}(\theta)) \tag{3}$$

where the weight factor $\beta > 0$ is defined as:

$$\beta = \frac{C_{fa} \cdot (1 - P_{target})}{C_{miss} \cdot P_{target}}, \tag{4}$$

and $-\beta <$ TWV(θ) <1 with 1 for a perfect system.C_{miss}, $C_{fa} > 0$ are the costs of miss and false alarms, and $0 < P_{target} < 1$ is the prior probability of a target trial, assumed to be constant across queries. ATWV is also the reference metric in NIST Spoken Term Detection evaluations.In the STD 2006 evaluation campaign they used TWV(θ_{act}) for system hard decisions, known as Actual Term-Weighted Value. As usual, the Maximum Term Weighted Value (MTWV) is the highest value that can be attained by applying a single threshold to system scores.

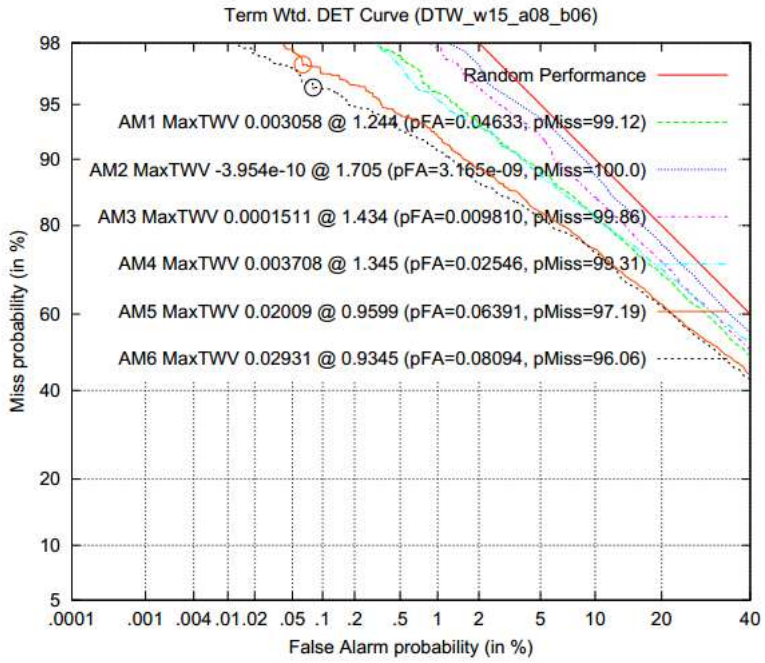

Figure 3 Results for the development data set

In Figure 3, a comparison of the results obtained with different acoustic models on the development data set are shown. We use a graphical performance assessment using a Detection Error Tradeoff (DET) curve that plots miss probability (pMiss) versus false alarm probability (pFA).Miss and false alarm probabilities are functions of the detection threshold, θ. This (θ) is applied to the system's detection scores, which are computed separately for each search term, then averaged to generate a DET line trace (Fiscus, 2007).

We can see that the Romanian acoustic model obtained the best results among individual languages, probably because it was trained with double the amount of data. AM4 multilingual performed slightly better than the monolingual acoustic models. The number of phonemes for this acoustic model is relatively high by combining data from all languages, thus it increases the uncertainty during recognition. Improvement is shown in multilingual model AM5, where using the IPA classification to merge phones helped and results show an improvement among the evaluation metrics. For comparison, the acoustic model AM6, trained with a bit amount of data, obtained the best results, even though it is trained with only one language (Romanian). It seems more data is needed to consolidate multilingual acoustic models, otherwise the larger phoneme set, even with merging, increase the detection uncertainty.

The results obtained on the development database with different speech features (PNCC and MFCC) are shown in table 2, and use a secondary metric, the normalized cross entropy cost (C_{nxe}). This metric has been used for several years in the language and speaker recognition fields to calibrate system scores, and correlate quite well with TWV metrics. C_{nxe} is based on system scores, in contrast to TWV, which evaluates system decisions. C_{nxe} measures the fraction of information, with regard to the ground truth, that is not provided by system scores, assuming that they can be interpreted as log-likelihood ratios. A perfect system would get $C_{nxe} \approx 0$ and a non-informative system would get $C_{nxe} = 1$ (Rodriguez-Fuentes, 2013). If we assume that the system under evaluation, *S*, submits a set of log-likelihood ratios llr_t for a set of trials *T(S)*, with the prior probability P_{tar}, then the empirical cross entropy, in information bits, is:

$$C_{xe} = \frac{1}{\log 2} \cdot \left(\frac{P_{tar}}{\left|T_{true}(S)\right|} \sum_{t \in T_{true}(S)} C_{\log}(llr_t) + \frac{1-P_{tar}}{\left|T_{false}(S)\right|} \sum_{t \in T_{false}(S)} C_{\log}(llr_t) \right), \tag{5}$$

where the logarithmic cost function is:

$$C_{\log}(llr_t) = \begin{cases} -\log(sigmoid(llr_t + \log it(P_{tar}))) & t \in T_{true}(S) \\ -\log(sigmoid(-(llr_t + \log it(P_{tar})))) & t \in T_{false}(S) \end{cases} \tag{6}$$

The empirical cross entropy of a system can be normalized ($llr_t = 0\ \forall t$) to obtain a trivial system, and we obtain the prior entropy, system that always gives non-informative scores:

$$C_{xe}^{prior} = \frac{1}{\log 2} \cdot \left(P_{tar} \cdot \log \frac{1}{P_{tar}} + (1 - P_{tar}) \cdot \log \frac{1}{1 - P_{tar}} \right) \quad (7)$$

Finally, the normalized empirical cross entropy is defined as (Rodriguez-Fuentes, 2013):

$$C_{nxe} = \frac{C_{xe}}{C_{xe}^{prior}} \quad (8)$$

Table 2 PNCC and MFCC performance comparison using actual and minimum C_{nxe}

AM ID	**PNCC**		**MFCC**	
	AC_{nxe}	$MinC_{nxe}$	AC_{nxe}	$MinC_{nxe}$
AM1	1.032	0.986	1.032	0.986
AM2	1.055	0.997	1.055	0.997
AM3	1.03	0.994	1.03	0.994
AM4	1.015	0.972	1.016	0.971
AM5	1.016	0.969	1.016	0.969

5. CONCLUSION AND FUTURE WORK

In this paper, we investigated whether the use of multi-language resources as input features help the process of term discovery for under-resourced languages, by phone recognition with a multilingual acoustic model of three languages (Albanian, English and Romanian). We compared single and multi-language resources used as a phoneme recogniser based on HMM for indexing the database, while an improved DTW based algorithm is used for searching a given query term in the content database.

Results show that we increase accuracy by training with multiple languages. Mixed language models increase the number of phonemes which leads to an increased uncertainty during the recognition phase. Using the IPA classification to merge phones helped and results show an improvement among the evaluation metrics. It seems that more data is needed in order to consolidate the multi-language acoustic models.

Regarding PNCC vs MFCC features, our results show no difference between the two types of parameters. In noisy environments, PNCCs is shown to obtain better accuracy, but it seems that the noise in MediaEval 2014 database is not significant enough to impact the results.

Further work will focus on improving the multi-language models by increasing the data used for training in the indexing stage, and also extend the study by usingan ASR phone recogniser based on state-of-the-art neural network classifiers and long temporal context.

ACKNOWLEDGEMENTS

The 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/134398and in part by the PN II Programme "Partnerships in priority areas" of MEN - UEFISCDI, through project no. 32/2014.

REFERENCES

Anguera X., Fuentes L.J.R, Szöke I., Buzo A., Metze F., "Query by Example Search on Speech at Mediaeval 2014", overview paper, Mediaeval Workshop (2014).

Burget L., et al., "Indexing and search methods for spoken documents," in ICTSD, (2006).

Buzo A., Cucu H., Safta M., Burileanu C., "Multilingual Query by Example Spoken Term Detection for Under-Resourced Languages", in Speech Technology and Human - Computer Dialogue (SpeD), (2013).

Can D., Saraçlar M., "Lattice Indexing for Spoken Term Detection", IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 19, NO. 8, (2011).

Dong W., Simon K., Joe F., Ravichander V., Nicholas E., and Raphaël T. "Direct posterior confidence for out-of-vocabulary spoken term detection", ACM Trans. Inf. Syst. 30, 3, Article 16 (2012).

Fiscus J.G., Ajot J., Garofolo J.S., Doddingtion G., "Results of the 2006 spoken term detection evaluation," in Proceedings of ACM SIGIR Workshop on Searching Spontaneous Conversational. Citeseer, pp. 51–55, (2007).

Flamary X., Anguera R. Oliver N., "Spoken wordcloud: Clustering recurrent patterns in speech," in Proc. Int. Workshop Content-Based Multimedia Index, pp. 133–138, (2011).

Hori T., Lee Hetherington I., Timothy J. Hazen, Glass J., "Open-vocabulary spoken utterance retrieval using confusion networks," in ICASSP, (2007).

Jansen A., Church K., and Hermansky H., "Towards spoken term discovery at scale with zero resources," in Proc. of INTERSPEECH-2010, pp. 1676–1679, (2010).

Jonathan M., Jia C., Xiaodong C., Mark J. F. G., Brian K., Kate K., Lidia M., David N., Michael P., Bhuvana R., Ralf S., Abhinav S., Philip C. W., "System Combination and Score Normalization for Spoken Term Detection", in Proceedings of Acoustics, Speech and Signal Processing (ICASSP), 8272 – 8276, (2013).

Kelly F., Harte N., "A COMPARISON OF AUDITORY FEATURES FOR ROBUST SPEECH RECOGNITION", in 18th European Signal Processing Conference (2010).

Kim C., Stern R. M., "Power-Normalized Cepstral Coefficients (PNCC) for robust speech recognition", in Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference, (2012).

Lee H., Tang Y., Tang H., Lee L., "Spoken term detection from bilingual spontaneous speech using code-switched lattice-based structures for words and subword units," Proceedings of ASRU, Italy, pp. 410-415, (2009).

H. Lin, L. Deng, D. Yu, Y. Gong, A. Acero and C.H. Lee, "A study on multilingual acoustic modeling for large vocabulary ASR," ICASSP: Proceedings of the Acoustics, Speech, and Signal Processing, Taipei, Taiwan, pp. 4333–4336, (2009).

Mamou J., Ramabhadran B., Siohan O., Vocabulary Independent Spoken Term Detection, SIGIR '07 Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, 615-622, (2007).

MENG, S., Y U, P., L IU, J., , SEIDE, F. "Fusing multiple systems into a compact lattice index for Chinese spoken term detection." In Proc. ICASSP'08. Las Vegas, Nevada, USA, 4345–4348, (2008).

Motlicek P., Valente F., "Application of out-of-language detection to spoken term detection," ICASSP: Proceedings of the Acoustics, Speech, and Signal Processing, USA, pp. 5098-5101, (2010).

Murao H., Kawaguchi N., Matsubara S., Inagaki Y., "Example-based query generation for spontaneousspeech," IEICE - Trans. Inf. Syst., (2005).

Muscariello A., Gravier G., and Bimbot F., "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).

Ng K., Zue V., "Subword-based approaches for spoken document retrieval," in PhD. Thesis, (2000).

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

Parlak S., Saraclar M., "Spoken Term Detection for Turkish Broadcast News," ICASSP: Proceedings of the Acoustics, Speech, and Signal Processing, USA, pp. 5244 - 5247, (2008).

Parada C., Sethy A., Ramabhadran B., "Balancing false alarms and hits in Spoken Term Detection," ICASSP: Proceedings of the Acoustics, Speech, and Signal Processing, USA, pp. 5286-5289, (2010).

Patty L., Spoken Term Detection Evaluation Plan, National Institute of Standards and Technology, (2006).

Rodriguez-Fuentes L. J., Penagarikano M., "MediaEval 2013 Spoken Web Search Task: System Performance Measures". Technical Report TR-2013-1, DEE, University of the Basque Country, (2013).

YU, P. SEIDE, F.. "A hybrid word / phoneme-based approach for improved vocabulary-independentsearch in spontaneous speech." In Proc. ICSLP'04. Jeju, Korea, 293–296, (2004).

Wade S., Christopher M. White, Timothy J. Hazen, "A Comparison of Query-by-Example Methods for Spoken Term Detection", MIT/Lincoln Laboratory, (2009).

Wang D., King S., Frankel J., Bell P., "Stochastic pronunciation modeling and soft match for out-of-vocabulary spoken term detection," ICASSP: Proceedings of the Acoustics, Speech, and Signal Processing, (2010).