ROMANIAN JOURNAL OF INFORMATION SCIENCE AND TECHNOLOGY
Volume **20**, Number 3, 2017, 210–221

Survey on Multilingual Spoken Term Detection

Alexandru CARANICA , Horia CUCU , Andi BUZO , and Corneliu BURILEANU

Speech and Dialogue Research Laboratory, University Politehnica of Bucharest, Romania,
Email: alexandru.caranica@speed.pub.ro, {horia.cucu, andi.buzo, corneliu.burileanu}@upb.ro

Abstract. Multilinguality is a characteristic that involves the use of more than one natural language in automatic speech recognition applications. In today's world, it is a characteristic of a rapidly increasing class of features required by a "complete" digital assistant, through means of *Automatic Speech Recognition* (ASR) systems. In working environments, where more than one language is in use, the problem of storing and retrieving information acquires a multilingual dimension. Unfortunately, many languages from developing countries, or minorities, received very little attention so far. One way of improving this situation is to do more research on the portability of speech and language technologies for multilingual applications, especially for under-resourced languages. These problems, as well as that of processing spoken material in a multilingual low-resource environment, will be briefly covered in this overview paper.

1. Introduction

Automatic Speech Recognition (ASR) is a broad research area that absorbs many efforts from many digital signal processing groups, throughout the research community. Availability of very large on-line corpora has enabled statistical models of language at every level, from phonetics to discourse. Today, speech recognition systems powered by artificial intelligence (machine learning) and the latest hardware electronics, are improving dramatically every day and have become sophisticated enough to not only help commercially (handling phone inquiries at customer call centers), but have also become part of consumers daily lives: virtual voice assistants on mobile devices (Siri, Google Now) are used for hands-free text messaging, or even for making dinner reservations or schedule a meeting via uttered, spontaneous speech.

In today's world, multilinguality is a characteristic of a rapidly increasing class of features required by a "complete" digital assistant, through means of *Automatic Speech Recognition* (ASR) systems. This fact is most apparent in an increased need for on-the-fly translations and a consequent interest in alternatives to the traditional ways of producing them [1]. But multilinguality is more than just the automatic translation. Before any ASR system can begin processing raw audio, the language in which it is written must be identified. This so called Language Identification problem is therefore a pressing one, and one which current technology still hasn't fully

solved yet, at least not for languages where resources (to build phonetic, acoustic models, etc.) are scarce (low-resource scenario).

In working environments where more than one language is in use, the problem of storing and retrieving information from digital multimedia documents acquires a multilingual dimension. There are however more than 6900 languages in the world and only a small fraction offer the resources necessary for implementation of *Natural Language Processing Techniques* (NLP) [2]. Thus, current NLP means are mostly concerned with languages for which large resources are available or which have suddenly become of interest because of the economic or political scene. Unfortunately, most languages from developing countries or minorities received only little attention so far and one way of improving this situation is to do more research on the portability of speech and language technologies, for multilingual applications, especially for under-resourced languages. These problems, as well as that of processing spoken material in a multilingual low-resource environment, will be briefly overviewed in this paper.

Moreover, vast amounts of digital audio data, in many languages, are being created every day and broadcasted from various sources, hence a pressing need exists for intelligent information extraction and retrieval methods, independent of the language spoken in these documents. 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. In recent years, numerous workshops hosted benchmarking initiatives to evaluate new algorithms for multilingual multimedia access and retrieval, such as MediaEval (2011-2015), or as special sessions at relevant conferences in the field of speech communication (ZeroSpeech Challenge, InterSpeech 2015, OpenKWS). Most of these spoken documents are in different languages, some of those even considered under-resourced in the speech community, hence also a growing need for an unsupervised method of information extraction and retrieval. 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 audio retrieval systems employ some sort of pre-indexing of the speech corpus, prior to search, without the advanced knowledge of the query terms. They make use of unsupervised learning techniques to adapt to the low-resourced language. Information retrieval and extraction have direct applications in the field of Natural Language Processing: finding out where textual resources reside and then extracting pertinent facts from those textual resources.

A typical STD (Spoken Term Detection) 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 - *Dynamic Time Warping Algorithm* [39]). The detection subsystem comprises of (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.

The remainder of this paper will be devoted, in the second chapter, to a more in-depth discussion of current research directions regarding multilingual systems and ways of intelligently extracting information out of low-resourced, multilingual databases.

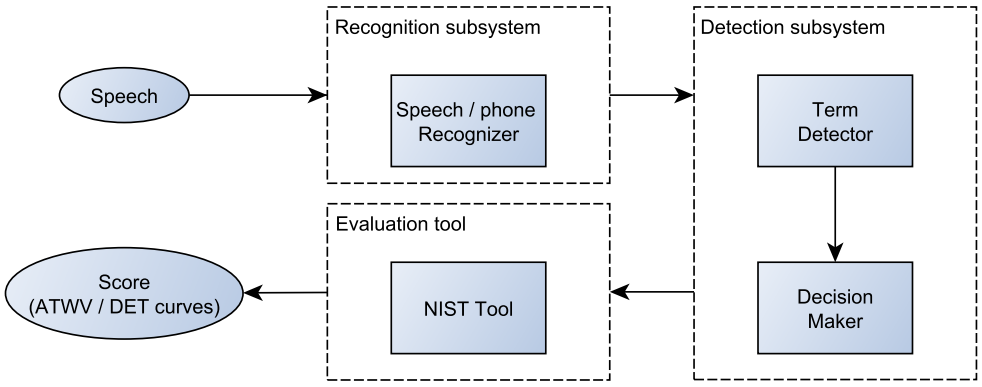

Fig. 1. Illustration of a typical STD (Spoken Term Detection) system, where the US NIST Tool is used to evaluate system performance. Adapted from [10].

2. Multilingual Spoken Content Recognition and Retrieval

State-of-the-art ASR systems typically use *Hidden Markov Models* (HMMs) or *Deep Neural Networks* (DNNs) and usually build on three components: acoustic-phonetic models, language model and a lexicon. If at least one of these components is multilingual, we refer to the whole system as a multilingual ASR system.

Multilingual language models are particularly useful when the speaker switches between languages (code-switching) or when the spoken language is unknown prior to decoding. Language models are normally trained on large amounts of text data. If text corpora from multiple languages are merged to estimate a multilingual language model, a language switch is in principle allowed at any time [3]. More restrictive approaches only allow language switches at common pause models [4]. Even though ASR systems with multilingual language models allow to implicitly identify the spoken language, if the spoken language is known apriori, usually the speech recognition performance is lower compared to ASR systems with monolingual language models, as shown in [5].

In a similar way, acoustic models can be trained on speech data from multiple languages. The main ndings of multilingual acoustic modeling studies such as [6, 7], can be generalized as follows [8]:

- If there is not enough training data (more than 100 hours), multilingual acoustic models perform worse than monolingual ones.
- The effect is more pronounced if data from more diverse languages are merged during training.
- Such systems have a high practical value, especially when little or no data exists in a particular language (low resourced scenario).

To model variability in the speech recordings, the acoustic models need to be trained on large amounts of acoustic data. Data collection involves large amounts of manual work, not only in time for the speakers to be recorded, but also for annotation of the subsequent recordings.

Therefore, developing ASR systems from scratch for a given language is expensive, and one of the main barriers in porting current systems to many languages is the large amount of data usually needed to train the models of current recognizers. On the other hand, large databases already exist for many languages and acoustic model training may in principle benet from data in languages other than the target language, assuming that all sounds produced by speakers across languages share a common acoustic space.

The problem of multilingual access to text databases can be seen as an extension of the general *information retrieval* (IR) problem. How does one retrieve documents containing expressions which do not exactly match the language or those found in the uttered query? Automatic characterization, in which the software attempts to duplicate the human process of "reading" is a very difficult problem. More specifically, "reading" involves attempting to extract information, both syntactic and semantic from the language, then using it to decide whether each document is relevant or not to a particular request. The difficulty is not only knowing how to extract the information, but also how to use it to decide relevance. Recent usage of "deep neural approaches" for Natural Language Processing brought "sentence meaning from the meanings of words and phrases" [9], and outperformed the classical statistical state-of-the-art systems on a variety of NLP tasks.

Many Natural Language Processing techniques have been used in Information Retrieval, such as stemming, part-of-speech tagging, compound recognition, de-compounding, chunking, word sense disambiguation, DTW etc [43]. It is interesting to see how these NLP techniques can be tailored to retrieve spoken documents from audio content, or discover word reoccurrences in a given audio corpus, as this might give a deeper level of understanding to an ASR system. Take, for example, Apple Siri or newly launched Microsoft Cortana. They are able to recognize speech, analyze it and then retrieve an answer to a question, and they are able to answer questions such as "how is the weather today?" or "when is my meeting scheduled for today?". This was not possible without document retrieval and processing techniques, and lately this has become a major interest topic in the speech community.

Regarding unsupervised multilingual processing, 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 [10]. In recent years, numerous workshops hosted benchmarking initiatives to evaluate new algorithms for multimedia access and retrieval, such as MediaEval (MediaEval, 2011-2015), or as special sessions at relevant conferences in the field of speech communication (ZeroSpeech Challenge, InterSpeech 2015, OpenKWS).

In the following section we look at two similar approaches for audio retrieval and discovery of speech related data (words reoccurrences, speech queries), in the context of under resourced languages. As opposed to the ASR task, there is no general way to evaluate such systems, and due to this fact, many researchers attended or submit their systems to popular evaluation campaigns, in the spoken document retrieval domain: The Zero Resource Speech Challenge [11], the first unified benchmark for zero resource speech technology and MediaEval 2015 QUESST - Query by Example Search on Speech Task [12]. The ZeroSpeech 2015 evaluation campaign targets the unsupervised discovery of linguistic units from raw speech in an unknown language. The idea behind this challenge is to push the envelope on the notion of flexibility in speech recognition

systems by setting up the rather extreme situation where a whole language has to be learned from scratch [11]. The Query by Example Search on Speech Task (QUESST) at MediaEval 2015 involves searching for audio speech content within audio content, using an audio content query, independent of the language. This task is particularly interesting for speech researchers in the area of spoken term detection or zero/low-resource speech processing, therefore requires researchers to build a language-independent audio-within-audio search system [12].

Sections to follow get into more detail about current approaches to solve each task, from multilingual systems to unsupervised processing of unknown speech.

2.1. Multilingual Speech Recognition

The process of determining the language of a speech utterance is called *Language Identication* (LID). LID describes the question of "which language is being spoken"? In a multilingual environment this information can then be exploited in various ways to recognize what was said. This task can be very challenging: it has to take into account various language-specic aspects, such as phonetic, vocabulary and grammar related. In multilingual speech recognition we try to nd the most likely word sequence that corresponds to an utterance where the language is not known apriori. This is a considerably harder task compared to monolingual speech recognition and it is common to use LID to estimate the current language. There are many general approaches to LID [14]. The rst approach uses "hierarchical multilayer perceptrons" to estimate language posterior probabilities given the acoustics in combination with hidden Markov models. The second approach "evaluates the output of a multilingual speech recognizer" to determine the spoken language. Different LID theoretical models have been proposed in literature:

- Spectral-similarity approaches: In these approaches several short-term spectra are extracted from the speech utterances. The spectra of the test utterances are then compared to those of the training utterances, using a Euclidean or another distance metric. The distant scores are accumulated and the language with the lowest distance score is selected [13].

- Prosody-based approaches: These approaches are based on pitch estimation and amplitude contours. They are then normalized to be "insensitive to overall amplitude, pitch and speaking rate" [13]. The accuracy of prosody-based approaches is highly language pair specific.

- Phone-recognition approaches: Phone-recognition approaches investigate the phone inventory of an utterance. Language characteristics are extracted based on the temporal order of the phones. Phono tactic constraints can be used in N-gram analysis to improve the result. These approaches require phonetically labelled corpora, but typically yield a higher performance [14].

- Artificial Neural Networks approaches: such methods are used to detect patterns that are not known in advance. This can be seen as a contrast to expert systems that rely on rules predened by the knowledge worker. [8].

Experiments show that, on a particular data set, LID can be used to significantly improve the performance of multilingual speech recognizers. Also, ASR dependent LID approaches yield the best performance due to higher-level cues and in general systems perform much worse on non-native data [14].

Other approaches can also be found in literature. In [15] the authors present a technique to vary the acoustic resolution of a phone decoder in LID by selecting the optimum set of phones. In [16] the authors describe how to build a Vietnamese and a Czech ASR system from scratch without any transcribed audio data. They use cross-language transfer from other languages, unsupervised training based on the "multilingual A-stabil" confidence score and bootstrapping. This approach is especially appropriate for under-resourced languages. Speech adaptation for non-native speech is a common way to improve ASR performance. Typical acoustic model adaptation techniques are *Maximum Likelihood Linear Regression* (MLLR) and *Maximum a Posteriori* (MAP) adaptation [44].

A possible approach to build a complete functional multilingual ASR system, for known languages, is to setup one system starting from those multiple languages. To recognize the language being spoken we require information about how to discriminate between languages, such as phonetic, phono tactic, vocabulary and grammar information, in order to bias towards the needed language. For this task, one can extract acoustic features and classify them by using hierarchical *Multilayer Perceptrons* (MLP). In a rst step, phone class posteriors are retrieved, then used to compute language posterior probabilities. The language posteriors from the MLPs are used as emission probabilities of a Hidden Markov Model that provides us with the correct timing. Different back-end metrics are presented and the systems are evaluated in terms of accuracy [14]. The LID results can then be used to choose from a set of monolingual speech recognizers, or to combine monolingual phone class posteriors for a multilingual speech recognizer. Furthermore, code-switches can effect the performance of a multilingual ASR system. A code-switch is a situation where one speaker changes the language during an utterance. It is a very common phenomenon in multilingual speaker communities. There has been a lot of research on a linguistic level on the nature of code-switches as well as the reason for speakers to use them [17], but relatively few studies have addressed the impact a code-switch situation has, on the overall speech recognition performance.

Multiple improvements and different approaches to this multilingual model were introduced, as it has been found to be very difficult to improve over separately trained systems. The usual approach has been to use some kind of "universal phone set" that covers multiple languages. In [18] authors report experiments on a different approach to multilingual speech recognition, in which the phone sets are entirely distinct but the model has parameters not tied to specific states that are shared across languages. A model called a "Subspace Gaussian Mixture Model" is used, where states distributions are Gaussian Mixture Models with a common structure, constrained to lie in a subspace of the total parameter space. The parameters that define this subspace can be shared across languages. They obtained substantial WER improvements with this approach, especially with very small amounts of training data, from an individual language.

In [19] the authors perform language identification experiments for four prominent South-African languages using a multilingual speech recognition system. Specifically, they show how successfully Afrikaans, English, Xhosa and Zulu may be identified using a single set of HMMs and a single recognition pass, further demonstrate the effect of language identification specific discriminative acoustic model training on both the per language recognition accuracy as well as the accuracy of the language identification process.

With the increasing availability of high amounts of processing power (the cloud model), Neural Network approaches are becoming more and more popular. In this context, it is of paramount importance to train accurate acoustic models for many languages within given resource constraints such as data, processing power, and time. Neural networks lend themselves naturally

to parameter sharing across languages, and distributed implementations have made it feasible to train large networks. In [20], the authors present experimental results for cross and multi-lingual network training of eleven languages, on over 10k hours of data in total. Cross-lingual training shares resources between languages when similarities can be found, for example, at acoustical or language level, where multi-lingual training treats each language individually. The average relative gains over the monolingual baselines are between 4% / 2% (data-scarce/data-rich languages) for cross and 7% / 2% for multi-lingual training. However, the additional gain from jointly training the languages on all data comes at an increased training time of roughly four weeks, compared to two weeks (monolingual) and one week (cross lingual). This large-scale experiment was enabled by a highly distributed software framework for deep neural networks, at Google, and seem to be the future for further advances in the multilingual department.

2.2. Multilingual Spoken Term Discovery

Multilingual Spoken term discovery systems identify recurring speech fragments from the raw speech, without any knowledge of the language at hand [21], to build classes of similar speech fragments. 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) [21–23]. Applications employing automatically discovered terms have quickly appeared, having a wide focus, ranging from topic segmentation to document classification [24] or spoken document summarization [25]. The 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 [23, 26], the optimization of the search process [27], or the introduction of linguistic constraints during DTW search [28].

Since spoken term discovery works in an unsupervised manner, the extraction of informative features is an important aspect. Zhang and colleagues [26] 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. They showed that for one of their system settings, the posteriorgrams always outperformed the *Mel Frequency Cepstral Coefficients* (MFCC) features. In [29] the authors approach the task by exploiting top-down information. They generate word-like pairs using an unsupervised term discovery system, then use the found matches as input to a neural network, in an effort to find a representation that brings the matches close together in the feature space. The results of this approach seem to consistently beat the baseline, in one case producing the best score in recent benchmark campaigns. In [30] the authors use features that are derived from a previously trained speech synthesis system for languages without a writing system. They compare features that are based on a cross-lingual phonetic system with features from segment-based inferred phones, using articulatory features derived directly from the acoustics. While this system uses side-information gleaned from a partially supervised system, it provides an intriguing insight into what is possible with articulatory features, which have been proven to be useful in supervised settings [31]. Finally, another interesting approach [32] proposes two auto-encoder variants (binary auto-encoders and hidden-markov-model encoders) to learn very compact representations of the input features. This results in representations that perform better than MFCCs with only six features.

2.3. Multilingual Spoken Term Detection

The Spoken Term Detection task is similar to the discovery one, only here, a query term is searched into a multilingual audio database. If Spoken Term Discovery can be compared with a library, where every book is classified in similar domains automatically, then Spoken Term Detection task is the search part, where the correct book is retrieved from this massive index.

As with most current web applications and search engines, end users expect good performance when it comes to front-end interaction. They should be able to instantly search and retrieve spoken documents, in any language, over a large collection of audio documents. It is why most systems that need to search over large quantities of data employ some sort of pre-indexing, prior to searching, and this rationale is also applied to audio based search systems. Thus, a typical STD system 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.

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 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. Hence, recent efforts are concentrated mainly on handling Out-Of-Vocabulary (OOV) words for which the pronunciation is unknown and the language model is unavailable [33, 34].

Regarding multilingual spoken term detection systems, there are a few previous studies [37, 38]. The first study uses an "out-of-language module based on confidence measures", to detect only the English speech segments. The latter study proposes a method for a "switch between Chinese and English languages using code-switched lattice-based structures for word/subword units". An alternative proposed solution is to build acoustic and language models that are shared across languages, like the study in [37] shows. Less work has been done involving methods for speech search by example. The authors in [37] proposed a "query-by-example approach to multilingual Spoken Term Detection" for under-resourced languages, based on ASR. This approach overcomes the main difficulties met under these conditions, providing "a new method for building multilingual acoustic models with few annotated data".

Many state of the art systems also make use of "phonetic search and data fusion techniques", to compensate for under-resourced language scenario. Widely used nowadays are approaches based on "subword units (phones)", to try and solve the out-of-vocabulary issue (OOV scenario). In this case, "subword representations of search terms are searched for within subword lattices", that are generated by a subword-based Automatic Speech Recognition system. Authors in [18,35, 36] made a significant breakthrough with a similar mechanism, using phonetic units for content based retrieval from speech, through a method of confusion networks applied to phones, finally outperforming lattice-based methods especially for out-of-vocabulary queries. Using data fusion techniques to combine results from diverse ASR systems, one can further improve robustness

across a variety of talkers, channels, environments and target terms. One good example is the system developed by [40]. The submitted system involves "dynamic time warping and symbolic search based approaches". The final submitted system was obtained by fusing 66 systems from 3 groups. Various other Hybrid approaches which fuse word and subword approaches at the lattice level have also been proposed in [41, 42].

3. Conclusion

This paper offered an overview of current multilingual approaches to spoken language recognition, with an extension to recent hot topics in the research community: Spoken Term Detection and Discovery, which helps extend the multilingual domain to under-resourced languages, where there are not enough resources to build reliable statistical models, in order to train reference models. Spoken Term Detection techniques are recommended for scenarios where information about the language is known and resources to build limited acoustic and language models are available. In other scenarios, where little or nothing is known about the language, Spoken Term Discovery might be the only way to extract some knowledge about the spoken documents (like similar speech fragments). Multilingual speech processing has been a topic of ongoing interest in the research community for many years, and the field is now receiving renewed interest. Recent developments, mentioned throughout the paper, confirm this trend. With this information at hand, further research directions are starting to focus on the development and addition of a deeper level of understanding, as the aim is to not only to recognize speech in any language, but also to extract the meaning and intent of what has been said, enabling multilingual voice driven systems as a whole to react in an intelligent way, appropriate to the user's needs.

4. 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

[1] R. COLE, J. MARIANI, H. USZKOREIT, G. BATTISTA VARILE, A. ZAENEN, A. ZAMPOLLI, *Survey of the State of the Art in Human Language Technology*, Cambridge University Press NY, 1998.

[2] BESACIER L., BARNARD E. B., KARPOV A. C., SCHULTZ T. D., *Automatic speech recognition for under-resourced languages: A survey*, in Speech Communication, **56**(1), 2014.

[3] T. WARD, S. ROUKOS, C. NETI, J. GROS, M. EPSTEIN, and S. DHARANIPRAGADA, *Towards speech understanding across multiple languages*, In The 5th International Conference on Spoken Language Processing, 1998.

[4] F. WENG, H. BRATT, L. NEUMEYER, and A. STOLCKE. *A study of multilingual speech recognition*, In Proceedings of Eurospeech, pp. 359–362, 1997.

[5] C. FUGEN, S. STUKER, H. SOLTAU, F. METZE, and T. SCHULTZ, *Efcient handling of multilingual language models*, In Proc. of ASRU Workshop, pp. 441–446, 2003.

[6] T. SCHULTZ and A. WAIBEL, Language independent and language adaptive acoustic modeling for speech recognition, Speech Communication, 35:3151, 2001.

[7] J. KÖHLER, *Multilingual phone models for vocabulary-independent speech recognition tasks*, Speech Communication, 35:2130, 2001.

[8] D. IMSENG, M. MAGIMAI.-DOSS, H. BOURLARD, *Hierarchical Multilayer Perceptron based Language Identication*, Proceedings of Interspeech, 2010.

[9] J. LI,, X. CHEN, E. HOVY, D. JURAFSKY, *Visualizing and Understanding Neural Models in NLP*, in Proc. of NAACL-HLT 2016, San Diego, California, June 12–17, 2016.

[10] 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.

[11] The Zero Resource Speech Challenge, accessed august 2015, http://www.lscp.net/persons/dupoux/bootphon/zerospeech2014/website/index.html

[12] MediaEval Benchmarking Initiative for Multimedia Evaluation, accessed august 2015, http://multimediaeval.org/mediaeval2015/quesst2015/.

[13] M. ZISSMAN, K. M. BERKLING, *Automatic Language Identication*, Information Systems Technology Group, Lincoln Laboratory, Massachusetts Institute of Technology, 2001.

[14] H. CAESAR, *Integrating Language Identication to improve Multilingual Speech Recognition*, Bachelor Thesis, IDIAP Research Institute - Martigny, Switzerland, 2012.

[15] P. KUMAR, H. LI, R. TONG, P. MATEJKA, L. BURGET, J. CERNOCKY, *Tuning Phone Decoders for Language Identication*, ICASSP, 2010.

[16] N. THANG VU, F. KRAUS, T. SCHULTZ, *Rapid building of an ASR system for Under-Resourced Languages based on Multilingual Unsupervised Training*, in Proc. of Interspeech, 2011.

[17] C. NILEP, Code Switching in Sociocultural Linguistics, University of Colorado, Bolder, 2006.

[18] L. BURGET, P. SCHWARZ, M. AGARWAL, P. AKYAZI, K. FENG, A. GHOSHAL, O. GLEMBEK, N. GOEL, M. KARA, D. POVEY, A. RASTROW, R. C. ROSE, S. THOMAS, *Multilingual Acoustic Modeling for Speech Recognition based on Subspace Gaussian Mixture Models*, in Proc. of ICASSP 2010.

[19] T. NIESLER, D. WILLETT, *Language identification and multilingual speech recognition using discriminatively trained acoustic models*, 2006.

[20] G. HEIGOLD, V. VANHOUCKE, A. SENIOR, P. NGUYEN, M. RANZATO, M. DEVIN, J. DEAN, *Multilingual acoustic models using distributed deep neural networks*, in Proc. of ICASSP, 2013.

[21] A. PARK and R. GLASS, *Unsupervised pattern discovery in speech*, IEEE Transactions on Audio, Speech, and Language Processing, **16**(1), pp. 186–197, 2008.

[22] A. JANSEN, K. CHURCH, and H. HERMANSKY, *Towards spoken term discovery at scale with zero resources*, in Proc. of INTERSPEECH, 2010, pp. 1676–1679.

[23] 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, **20**(7), pp. 2031–2044, 2012.

[24] M. DREDZE, A. JANSEN, G. COPPERSMITH, and K. CHURCH, *NLP on spoken documents without ASR*, in Proc. of EMNLP, 2010, pp. 460–470.

[25] D. HARWATH, T. HAZEN, and J. GLASS, *Zero resource spoken audio corpus analysis*, in Proc. of IEEE ICASSP, 2013, pp. 8555–8559.

[26] Y. ZHANG and J. GLASS, *Towards multi-speaker unsupervised speech pattern discovery*, In Proc. ICASSP, pp. 4366–4369, 2010.

[27] A. JANSEN and B. VAN DURME, *Efficient spoken term discovery using randomized algorithms*, in Proc. of IEEE ASRU, 2011, pp. 401–406.

[28] B. LUDUSAN, G. GRAVIER, and E. DUPOUX, *Incorporating prosodic boundaries in unsupervised term discovery*, in Proc. of Speech Prosody, 2014, pp. 939–943.

[29] THIOLLI'ERE R., DUNBAR E., SYNNAEVE G., VERSTEEGH M., DUPOUX E., *A hybrid dynamic time warping-deep neural network architecture for unsupervised acoustic modeling*, In: Proceedings of Interspeech, 2015.

[30] BALJEKAR P., SITARAM S., MUTHUKUMAR P.K., BLACK A., Using articulatory features and inferred phonological segments in zero resource speech processing, In Proc. of Interspeech. 2015.

[31] MITRA V., SIVARAMAN G., NAM H., ESPY-WILSON C., SALTZMAN E., *Articulatory features from deep neural networks and their role in speech recognition*, In: Proceedings of IEEE International Conference on Acoustic, Speech and Signal Processing. 2014, pp. 3041–3045.

[32] BADINO L., MERETA A., ROSASCO L. *Discovering discrete subword units with binarized autoencoders and hidden-markov-model encoders*, In: Proceedings of Interspeech, 2015.

[33] 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.

[34] 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.

[35] NG K., ZUE V., *Subword-based approaches for spoken document retrieval*, PhD. Thesis, 2000.

[36] HORI T., LEE Hetherington I., TIMOTHY J. Hazen, GLASS J., *Open-vocabulary spoken utterance retrieval using confusion networks*, in ICASSP, 2007.

[37] H. LIN, L. DENG, D. YU, Y. GONG, A. ACERO and C.H. LEE, *A study on multilingual acoustic modeling forlargevocabulary ASR*, ICASSP: Proceedings of the Acoustics, Speech, and Signal Processing, Taipei, Taiwan, pp. 4333–4336, 2009.

[38] MOTLICEK P., VALENTE F., *Application of out-of-language detection to spoken term detection*, ICASSP: Proc of the Acoustics, Speech, and Signal Processing, USA, pp. 5098–5101, 2010.

[39] BUZO A., CUCU H., SAFTA M., BURILEANU C., *Multilingual Query by Example Spoken Term Detection for Under-Resourced Languages*, in Proc of. SpeD 2013.

[40] J. HOU, V. T. PHAM, C. LEUNG, L. WANG, H. XU, H. L. XE, Z. FU, C. NI, X. XIAO, H. CHEN, S. ZHANG, S. SUN, Y. YUAN, P. LI, T. L. NWE, S. SIVADAS, B. MA, E. S. CHNG, H. LI, The *NNI Query-by-Example System for MediaEval 2015*, Working Notes Proceedings of the MediaEval 2015 Workshop, Wurzen, Germany, September 14–15, 2015.

[41] YU P., SEIDE F., *A hybrid word / phoneme-based approach for improved vocabulary-independent search in spontaneous speech*, In Proc. ICSLP04. Jeju, Korea, 293–296, 2004.

[42] MENG S., YU P., LIU J., SEIDE F., *Fusing multiple systems into a compact lattice index for Chinese spoken term detection*, In Proc. ICASSP08. Las Vegas, Nevada, USA, 2008.

[43] R. SUBHASHINIV, J. Senthil KUMAR, *A Framework for Efficient Information Retrieval Using NLP Techniques*, In Proc. of Communications in Computer and Information Science book series (CCIS, **142**).

[44] Y. H. SUNG, C. BOULIS, and D. JURAFSKY, *Maximum Conditional Likelihood Linear Regression and Maximum a Posteriori for Hidden Conditional Random Fields Speaker Adaptation*, in ICASSP, 2008.