Multilingual Query by Example Spoken Term Detection for Under-Resourced Languages

Andi Buzo, Horia Cucu, Mihai Safta, Corneliu Burileanu
Speech and Dialogue (SpeeD) Research Laboratory
University Politehnica of Bucharest
Bucharest, Romania
{andi.buzo, horia.cucu, mihai.safta, corneliu.burileanu}@upb.ro

***Abstract* — We propose a query-by-example approach to multilingual Spoken Term Detection for under-resourced languages based on Automatic Speech Recognition. The approach overcomes the main difficulties met under these conditions, i.e., providing a new method for building multilingual acoustic models with few annotated data and searching in approximate Automatic Speech Recognition transcriptions providing high scalability. The acoustic models are obtained by adapting well trained phonemes to the ones from the envisaged languages. The mapping is made according to 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. Experimental validation was conducted on a standard data set consisting of 3 hours of mixed African languages. The recorded speech has telephonic quality and it is a mix of read and spontaneous speech.**

Keywords — spoken term detection; multilingual acoustic model; under-resourced languages;

I. INTRODUCTION

The increase of the available spoken data has made Spoken Term Detection (STD) an important tool for retrieving precise and relevant information. STD for high-resourced languages is usually approached based on Automatic Speech Recognition (ASR), relying on large amounts of training data, including recordings (for acoustic modeling) and text data (for language modeling) in the target languages. The good accuracy of the ASR for such languages and the improved indexing and searching techniques have assured high quality STD. 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 useless [1, 2, 3].

The recent advances of the technology have increased user's expectations about the usability of the speech applications. One of the current main challenges is the need to support multiple language input, especially for applications dedicated to linguistically diverse communities.

The query-by-example (QbyE) STD has proved to be a suitable framework for the case where pronunciation is unavailable as in the case of OOV words [4, 5]. This is also the case for under-resourced languages where the phonetic dictionary is often not provided. Not only is QbyE a compulsory approach for multilingual under-resourced STD (some languages and dialects don't even have the written form) but it is also an attractive solution that leads to a full speech system.

In this paper, we propose an approach to multilingual STD for under-resourced languages in the case of spoken query. This means that we are searching for spoken term within a spoken content by using a spoken query. The system uses the STD architecture proposed at NIST 2006 campaign [6], which separates the indexing and the searching modules rather than searching the corpus directly for each query term. The evaluation of the approach is made in the context of the MediaEval 2012 Spoken Web Search campaign [7]. The Lwazi database of African languages contains recorded speech over the telephonic channel. Language labels and phonetic dictionaries are not provided.

The remainder of the paper is organized as follows: related work is presented in Section 2, Section 3 describes the proposed approach for the multilingual acoustic model and the searching method, results and discussions are given in Section 4 followed by the conclusions.

II. RELATED WORK

Multilingual speech systems have to deal with multiple languages input. If the language is known a priori, then the multilingual STD system could choose the corresponding acoustic and language models. Otherwise, a Language Identification (LID) module is needed in order to pick the correct STD system. There are a few previous studies on multilingual STD for example [8] and [9]. 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 [10-12].

Our study is closely related to works that provide multilingual acoustic models, even though they are mainly used in ASR. One way of obtaining such universal models is by training with multiple languages speech data and by using IPA phoneme classification [11]. The authors use 8 languages to cover all the phonemes from the target languages. Eventually, the acoustic models can be adapted to a specific language by using a small amount of data. In addition to knowledge based techniques, data driven techniques can be used if less languages are available. In [10], the authors use Discriminative Combination Models starting from 4 languages,

while in [12] the authors use a confusion matrix obtained by recognizing a small amount of speech from the target language (Vietnamese) with acoustic models trained in the source language (French).

Another approach is by building new acoustic models. For example, Imseng et al. [13] proposed the Kullback-Leibler divergence based acoustic modeling method, which is able to exploit multilingual information in a principled way.

We mainly relate our work to QbyE STD studies on OOV words because this case is closer to the under-resourced languages condition. QbyE STD based on phone recognition and Dynamic Time Warping (DTW) as searching method is previously approached in [4, 14, 5]. The methods described in [4] are based on confusion networks and treat the search as a string-distance comparison problem. The transcriptions are obtained by a neural network based phone recognizer trained with 10h of speech (from Switchboard2, phase 4 data). The approach is suitable for under-resourced languages, but it uses only single language input. The same authors show that QbyE STD can also be approached by phonetic posteriorgrams [14]. They are obtained by a time-vs.-phoneme-class matrix representation in which each element represents the posterior probability of a phoneme at the respective time frame. They use the same recognizer as in [4], but the evaluation database is different, hence it is not possible to directly compare the approaches. In [5], the authors use a weighted finite state transducer based indexing system in order to search directly with the phoneme lattices from the ASR output. Even though the experiments are made on OOV words, the acoustic model is trained with 300 hours of speech, which are not available for under-resourced languages. QbyE STD is also approached by pattern recognition methods as in [15], where the search is made directly on the speech features. Despite several optimizations made on the computation speed, such systems are not scalable because they don't make any indexing, leaving the entire computation load to the search process.

We propose a novel multilingual acoustic modeling method and an improved scalable DTW searching method. In our approach, the proposed indexer is a multilingual phoneme recognizer. We start from acoustic models of a single language (Romanian) and we use the IPA classification for mapping the phonemes from the under-resourced languages to the Romanian phonemes. Because the Romanian phonemes do not cover all the phonemes of the envisaged language we propose a novel approach for mapping the rest of the phones based on phoneme similitude (according to IPA) and on a confusion matrix. It is similar to the approach proposed in [12], but in addition we introduce an intermediate step for mapping phonemes that are similar according to the full IPA chart [16]. This solution is motivated by the high number of phonemes (77 in the envisaged African languages vs. 28 Romanian phonemes). The advantage of this approach is that it does not require multiple languages speech data but only little data for acoustic model adaptation.

We have also improved the common DTW string comparison method by including in the distance formula the effect of query length and DTW match spread. Their effect is weighted in order to find an optimal configuration.

III. PREPARE YOUR PAPER BEFORE STYLING

Our system is built based on the architecture proposed at NIST 2006 campaign [6]. This means that it has an indexing stage which is run offline and a searching stage which is run online. This architecture makes the search faster provided that the indexing method simplifies the search. We use ASR for indexing, thus all the speech contents are transformed in strings of phonemes. During the retrieval of the target term, the spoken query is transcribed by the ASR system and a search is performed over the speech contents to find a string of phonemes that exceeds a threshold of similitude with the given query.

A. The indexing component

The proposed STD system is based on an ASR system previously developed in our lab for the Romanian language [17]. The acoustic model is trained with 64 hours of speech from 30 speakers. It is based on Hidden Markov Models with 3 states per phoneme, 4000 senones, 8 Gaussian components per state and MFCC + Δ + $\Delta\Delta$ used as speech features. The language model is N-gram based and it is built with 175 million words. The overall system performance is 18% Word Error Rate (WER) on clean continuous speech. Since the envisaged Lwazi database consists in telephonic recordings, the first step was to resample (8 kHz) our Romanian recordings and re-train the acoustic model.

In the case of under-resourced languages, the lack of a lexicon and of a phonetic dictionary makes phone recognition a compulsory choice. Otherwise, the great number of out-of-vocabulary words that may appear in the query terms will increase the number of errors. Thus we have tuned the ASR system in order to minimize the Phone Error Rate (PhER) which is calculated similarly to WER. The tuned parameters are relative beam width, word insertion probability, language weight, etc. By tuning the relative beam width related parameters we have reduced the PhER from 36.8% to 31.4%. By tuning the language related parameters, such as language model weight and word insertion penalty, we have obtained a further reduction up to 25.3%.

The envisaged under-resourced languages are the ones provided in the Lwazi database. The development data set consists in 3 hours of annotated speech at sentence level from several African languages (isiNdebele, Siswati, Tshivenda and Xitsonga). By applying a Viterbi forced alignment, we have obtained annotation at phone level too. There are 77 distinct phonemes and the IPA classification is known for each of them [18]. At this point, the main challenge is to come up with a multilingual acoustic model with few annotated data.

The multilingual acoustic model is built by mapping all the 77 African phonemes to the 28 Romanian phonemes. After the mapping, a maximum a priori (MAP) adaptation is run for the Romanian phones by using the Lwazi development data set. The proposed mapping algorithm is the following:

For each African phoneme:

1. If the IPA classification is identical to the one of the Romanian phoneme then the phonemes are mapped directly.

2. Else, it is mapped to the closest Romanian phoneme according to the full IPA chart [16].

3. If none of the above is applicable, then it is mapped to the Romanian phoneme that it is confused most often with, according to a confusion matrix.

The confusion matrix is built by recognizing the whole development dataset with the Romanian acoustic models. It consists in the counts of the confusion pairs obtained by aligning the recognition output and the transcription of the recording.

With this mapping, the system has a PhER of 61.2%, which is rather high. Therefore, the Romanian acoustic models are adapted to the development data set by using MAP. Using the phone-level annotation, obtained by force alignment, proved to be more efficient than the use of phrase-level annotation. In the first case we obtained a PhER of 48.1% while in the second case a PhER of 50.3%.

B. The searching component

If the accuracy of the ASR used for indexing is 100%, then the searching problem is reduced to a mere string search. In the multilingual context for under-resourced languages the PhER is rather high (48%) which means that it must be compensated with a robust searching algorithm. Besides accuracy, the computational cost and the scalability are also important criteria in evaluating the searching component.

The proposed Dynamic Time Warping String Search (DTWSS) uses DTW for aligning the query within one content, where both query and contents are string of phonemes. 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. The size of the window must be greater than the query size because, otherwise the insertions that may occur in the content will make some matched phones fall outside the window. The window cannot be too long either, because the phonemes from the query will eventually be found in the window yielding in a false match. For each window, the alignment is given a score *s* which is complementary to the cost (calculated the same as *PhER*):

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

where *s* is a score of similitude while the detection is based on a threshold which is determined empirically. The windowing process and the recognition timestamps make the term localization possible. In the evaluation process we will use this method as the baseline.

We have improved this common searching method by introducing a penalization for short queries and large DTW match spread (how compact the aligned phonemes are).

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

where *LQ* is the length of the query, *LQM* and *LQm* are the maximum and minimum values respectively for *LQ, LW* is the window length, *LS* is the spread of the DTW match, while α and β are tuning parameters that control the amount of penalization. The query length normalization is motivated by the fact that for imperfect recognition conditions the probability to confuse short terms is higher than for long terms. The weighting of the DTW match spread is explained by the fact that the higher the compactness of the aligned phonemes, the higher the probability that these phonemes belong to the same term. The query-length normalization is used in previous works as in [2], but no solution is provided for controlling and optimizing its effect. The introduction of the tuning parameters (α and β) is a novelty of the proposed approach. The optimal value for these parameters is determined empirically.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

For the evaluation of the method we have used the evaluation Lwazi database from the MediaEval 2012 campaign [7]. It has common characteristics with the development database described in Section 3 but it is not used for refining the proposed methods. The evaluation metric used for STD performance is the Actual Term Weighted Value (ATWV) proposed in [6] which also takes into account the localization of the term.

A. The refinment of the method

The first experiments are run on the development database (which is split 90% for adaptation and 10% for testing) in order to determine the optimal values for α and β from the DTWSS method. We consider as the baseline method the one based on DTW score defined by equation (1) (α=0 and β=0). The results are presented in Table I. While in the description of the method we motivated the effectiveness of advantaging long queries and short DTW match spread, the experimental results show that α and β cannot be too high either. A high value for α means that a short term will not obtain a sufficient score even though it is perfectly recognized and perfectly matched (only because it is penalized for being short). Similarly, a high value for β means that compact DTW matches will be given a higher score even though their pure DTW score (1-*PhER*) is low. Hence, there exist optimal values for α and β. For comparison to other methods we have chosen the following (α,β) combination: (0.8,0.4), (0.6,0.6) and (0.1,0.4).

$$\alpha + \beta = \chi. \tag{1}$$

$$\alpha + \beta = \chi. \tag{1}$$

TABLE I. THE DTWSS RESULTS

ATWV	α=0	α=0.1	α=0.2	α=0.4	α=0.6	α=0.8	α=1
β=0	0.21	0.21	0.22	0.22	0.23	0.22	0.22
β=0.2	0.29	0.29	0.31	0.30	0.32	0.28	0.25
β=0.4	0.31	**0.33**	0.33	0.33	0.33	**0.34**	0.30
β=0.6	0.31	0.32	0.32	0.32	**0.33**	0.31	0.31
β=0.8	0.28	0.31	0.30	0.32	0.30	0.28	0.26

$$\alpha + \beta = \chi. \tag{1}$$

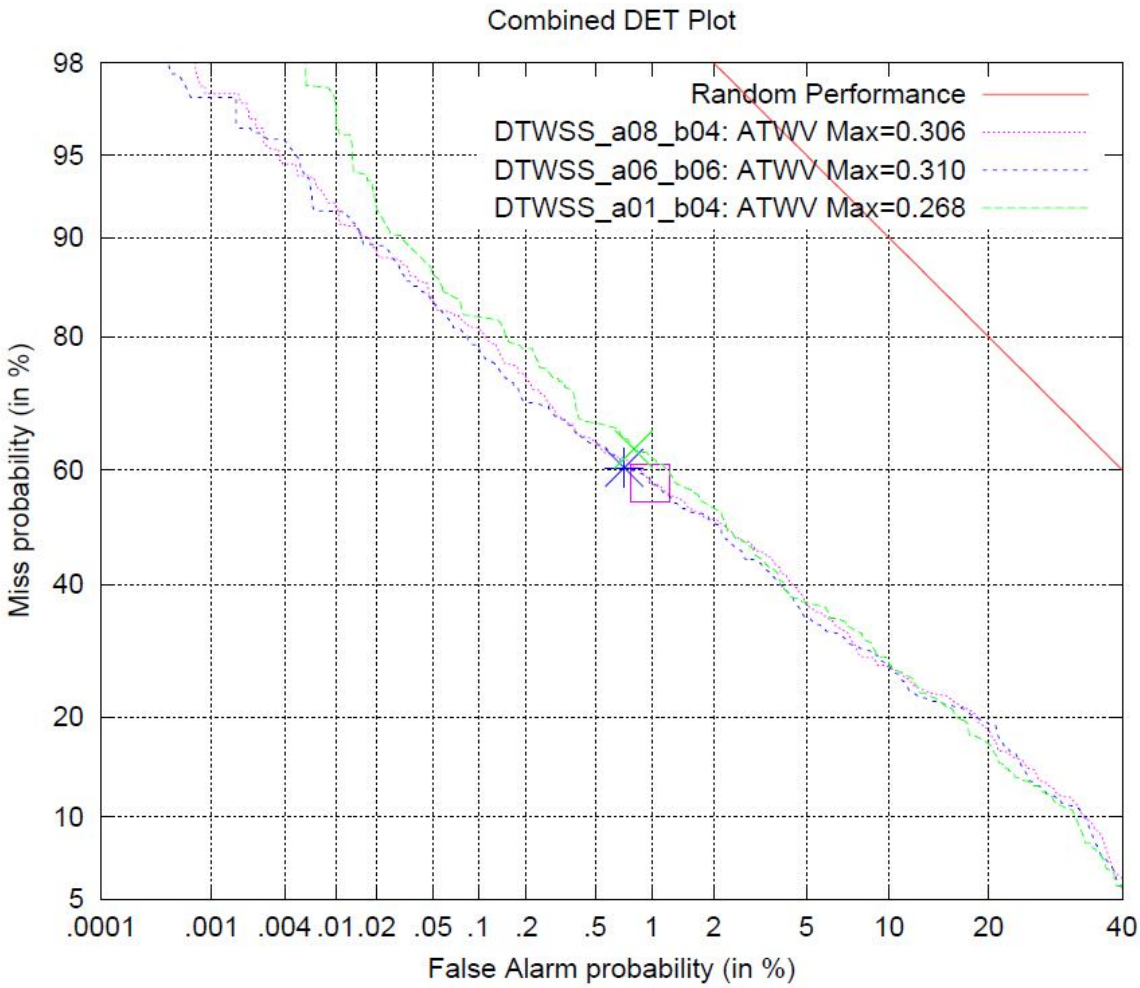

Fig. 1. The DET curves for the proposed methods

Figure 1 shows the Detection Error Trade-off (DET) curves for the proposed methods obtained for the evaluation database, while Table II shows how well the methods perform on seen/unseen data by testing with all combinations of data sets (dev/eval queries, dev/eval contents). All the methods suffer a performance degradation by 0.13-0.18 ATWV when moving from seen data to unseen data. The results show that the DTWSS outperforms the other two methods.

TABLE II. RESULTS ON EVALUATION/DEVELOPMENT DATA SET

ATWV	*evalQ-evalC*	*evalQ-devC*	*devQ-evalC*	*devQ-devC*
DTWSS (α=0.8 β=0.4)	0.31	0.47	0.33	0.49
DTWSS (α=0.6 β=0.6)	0.31	0.48	0.33	0.47
DTWSS (α=0.1 β=0.4)	0.27	0.44	0.32	0.47

B. Comparison to other methods

Table III compares our DTWSS method with other methods from other research teams that have used the Lwazi database for STD evaluation. The system names are the same the authors have used in the papers reference in the table. Most of the proposed methods (RAILS, IRDTW, CDTW, DTWSVM) treat STD as a pattern recognition problem by aligning speech features. The aligning algorithms are mainly based on DTW with significant variations. Some of these methods obtain good results, but they are not highly scalable. Only feature extraction can be done offline, while the rest of the operations (including scoring and searching, which are quite sophisticated) are done online.

The AKWS method uses an out-of-grammar branch which consists in a phone loop. It is expected that speech segments which do not contain the query term will score better in this branch, thus reducing the false alarms which are the main problem in our solution. The AKWS method assumes that a recognition process is run for the whole content for each query search. The computation cost of the recognition process becomes an important barrier for the scalability. This method, as the presented pattern recognition methods, are efficient if few searches are made within a content, but for multiple searches it is best if the greatest part of the computation is made offline.

Our DTWSS is highly scalable. The recognition for the content is made offline while the search is reduced to a string comparison. The proposed method has also a competitive

accuracy and the high score obtained when testing the development database shows that there is margin for improvement if the recognition accuracy is increased or if more data are available.

TABLE III. COMPARISON WITH OTHER METHODS BY ATVW

Team-Method	*evalQ-evalC*	*devQ-devC*
BUT-AKWS [19]	0.49	0.49
JHU_HLTCOE-RAILS [20]	0.36	0.38
TID-IRDTW [21]	0.33	0.38
DTWSS (α=0.8 β=0.4)	**0.31**	**0.49**
TUM-CDTW [22]	0.29	0.26
GTTS-Phone_Lattice [23]	0.08	0.09
TUKE-DTWSVM [24]	0	0

V. CONCLUSIONS

The proposed method approaches multilingual STD by cross-language transfer of acoustic models. The acoustic model adaptation after the hierarchical phoneme classification reduced the PhER from 61% to 48%. The experimental results show that there exists optimal weights for the query normalization and DTW match spread effects. The query length normalization is motivated by the fact that for imperfect recognition conditions the probability to confuse short terms is higher than for long terms. The weighting of the DTW match spread is explained by the fact that the higher the compactness of the aligned phonemes, the higher the probability that these phonemes belong to the same term. The proper weighting of these effects improved ATWV by 0.13. Compared to other methods that have used the same evaluation database, our approach have the same accuracy as the DTW-based pattern recognition methods (ATWV of 0.31 vs. [0.29, 0.36]). Despite our system is not trained with large data from the Lwazi dataset, the trained phones represent good classes for the classification of speech. Only the keyword spotting method performed significantly better (ATWV of 0.49). However, our approach is more scalable to large database than the other proposed methods, because the search is reduced to string comparison. Moreover, the high ATWV (0.49) obtained on seen data shows that better acoustic models can increase the accuracy.

ACKNOWLEDGMENT

We thank the organizers of the Media Eval 2012 Spoken Web Search task for making available the Lwazi database [7].

REFERENCES

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

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

[3] D. Wang, S. King, J. Frankel and P. Bell, "Stochastic pronunciation modeling and soft match for out-of-vocabulary spoken term detection," ICASSP: Proceedings of the Acoustics, Speech, and Signal Processing, USA, pp. 5294-5297, 2010.

[4] W. Shen, C. White and T. Hazen, "A comparison of query-by-example methods for spoken term detection," Proceedings of INTERSPEECH, UK, pp. 2143-2146, 2009.

[5] C. Parada, A. Sethy and B. Ramabhadran, "Query-by-Example Spoken Term Detection for OOV Terms," IEEE Workshop on Automatic Speech Recognition & Understanding (ASR), Italy, pp. 421-426, 2009.

[6] NIST Spoken Term Detection (STD) 2006 Evaluation Plan, http://www.itl.nist.gov/iad/mig//tests/std/2006/docs/std06-evalplan-v10.pdf

[7] F. Metze et al., "The spoken web search task," Working Notes Proceedings of the MediaEval 2012 Workshop, Pisa, Italy, October 4-5, 2012, CEUR-WS.org, ISSN 1613-0073, http://ceur-ws.org/Vol-927/mediaeval2012_submission_4.pdf.

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

[9] H. Lee, Y. Tang, H. Tang and L. Lee, "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.

[10] W. Byrne, et al., "Towards language independent acoustic modeling," ICASSP: Proceedings of the Acoustics, Speech, and Signal Processing, Turkey, pp. II1029 - II1032 vol.2, 2000.

[11] 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.

[12] V-B. Le and L. Besacier. "First steps in fast acoustic modeling for a new target language. Application to Vietnamese," ICASSP: Proceedings of the Acoustics, Speech, and Signal Processing. USA, pp. 821 - 824, 2005.

[13] D. Imseng, H. Bourlard and P.N. Garner, "Using KL-Divergence and Multilingual Information to Improve ASR for Under-Resourced Languages," ICASSP: Proceedings of the Acoustics, Speech, and Signal Processing, Japan, pp. 4869-4872, 2012.

[14] T. Hazen, W. Shen and C. M. White, "Query-by-Example Spoken Term Detection using Phonetic Posteriorgram Templates," IEEE Workshop on Automatic Speech Recognition & Understanding (ASR), Italy, pp. 421-426, 2009.

[15] Y. Zhang, K. Adl and J. R. Glass, "Fast spoken query detection using lower-bound Dynamic Time Warping on Graphical Processing Units," ICASSP: Proceedings of the Acoustics, Speech, and Signal Processing, Japan, pp. 5173-5176, 2012.

[16] http://www.langsci.ucl.ac.uk/ipa/fullchart.html.

[17] H. Cucu, L. Besacier, C. Burileanu and A. Buzo, "Investigating the Role of Machine Translated Text in ASR Domain Adaptation: Unsupervised and Semi-supervised Methods," in the Proceedings of the 2011 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), USA, pp. 260-265, 2011, ISBN: 978-1-4673-0366-8.

[18] http://hlt.mirror.ac.za/Phoneset/Lwazi.Phoneset.1.2.pdf.

[19] I. Szöke, M. Fapšo and K. Veselý, "But2012 Approaches for Spoken Web Search - MediaEval 2012," Working Notes Proceedings of the MediaEval 2012 Workshop, Pisa, Italy, October 4-5, 2012, CEUR-WS.org, ISSN 1613-0073, http://ceur-ws.org/Vol-927/mediaeval2012_submission_29.pdf.

[20] A. Jansen, B. Van Durme and P. Clark, "The JHU-HLTCOE Spoken Web Search System for MediaEval 2012," Working Notes Proceedings of the MediaEval 2012 Workshop, Pisa, Italy, October 4-5, 2012, CEUR-WS.org, ISSN 1613-0073, http://ceur-ws.org/Vol-927/mediaeval2012_submission_30.pdf.

[21] X. Anguera, "Telefonica Research System for the Spoken Web Search Task at Mediaeval 2012," Working Notes Proceedings of the MediaEval 2012 Workshop, Pisa, Italy, October 4-5, 2012, CEUR-WS.org, ISSN 1613-0073, http://ceur-ws.org/Vol-927/mediaeval2012_submission_34.pdf.

[22] C. Joder, F. Weninger, M. Wollmer and B. Schuller, "The TUM Cumulative DTW Approach for the Mediaeval 2012 Spoken Web

Search Task," Working Notes Proceedings of the MediaEval 2012 Workshop, Pisa, Italy, October 4-5, 2012, CEUR-WS.org, ISSN 1613-0073, http://ceur-ws.org/Vol-927/mediaeval2012_submission_25.pdf.

[23] A. Varona, M. Penagarikano, L.J. Rodriguez-Fuentes, G. Bordel and M. Diez, "GTTS System for the Spoken Web Search Task at MediaEval 2012," Working Notes Proceedings of the MediaEval 2012 Workshop, Pisa, Italy, October 4-5, 2012, CEUR-WS.org, ISSN 1613-0073, http://ceur-ws.org/Vol-927/mediaeval2012_submission_24.pdf.

[24] J. Vavrek, M. Pleva and J. Juhar, "TUKE MediaEval 2012: Spoken Web Search using DTW and Unsupervised SVM," Working Notes Proceedings of the MediaEval 2012 Workshop, Pisa, Italy, October 4-5, 2012, CEUR-WS.org, ISSN 1613-0073, http://ceur-ws.org/Vol-927/mediaeval2012_submission_23.pdf.