The L^{2}F Query-by-Example Spoken Term Detection system for the ALBAYZIN 2016 evaluation

Anna Pompili and Alberto Abad

L^{2}F - Spoken Language Systems Lab, INESC-ID Lisboa
IST - Instituto Superior Técnico, University of Lisbon
{anna.pompili,alberto.abad}@inesc-id.pt
http://www.l2f.inesc-id.pt

Abstract. Query-by-Example Spoken Term Detection (QbE-STD) is the task of finding occurrences of a spoken query in a repository of audio documents. In the last years, this task has become particularly appealing, mostly due to its flexibility that allows, for instance, to deal with low-resourced languages for which no Automatic Speech Recognition (ASR) system can be built. This paper reports experimental results of the L^{2}F system built for the QbE-STD Albayzin 2016 evaluation. The system exploits frame level phone posteriors followed by a Dynamic Time Warping (DTW) search procedure. In order to ease the search process, audio documents are first partitioned into smaller segments using an audio segmentation module. Then, given a query and an audio document, the normalized distance matrix is computed between their phone posterior representations and a segmental DTW matching procedure is performed between the query and each segment of the audio document. Phone decoders of different languages have been exploited and no assumption has been made about the language of the audio collection and the queries. In the end, different sub-systems are combined together based on a discriminative calibration and fusion approach.

Keywords: Spoken Term Detection, Phone Posteriorgrams, Dynamic Time Warping, Score Calibration and Fusion

1 Introduction

The task of QbE-STD aims to find occurrences of a spoken query in a set of audio documents. In the last years, QbE-STD has gained the interest of the research community for its versatility in settings where untranscribed, multilingual and acoustically unconstrained spoken resources must be searched, or when searching spoken resources in low-resource languages. The query-by-example task can be considered as a sort of generalization of the problem of speech search based on text queries, wherein, usually, the search space involves a single language for which there is plenty of resources to build ASR systems. Under these conditions, a simple approach to the task would first perform speech-to-text conversion

of the queries and then apply any of the methods used in text-based speech search. However, when the spoken language is unknown (or, equivalently, when multiple languages may appear) or when there are not enough resources to build robust ASR systems, alternative approaches that do not rely on well-trained acoustic models are needed. In the case of QbE-STD, some of the most recent approaches are based on template matching methods, such as different flavours of DTW of posterior derived features [1, 2]. Other systems use acoustic keyword spotting (AKWS) [3, 4], exploiting multilingual acoustic models in several ways. A common trend in current QbE-STD systems is the combination of several (probably weak) detectors, each providing complementary information, which usually leads to improved detection performance [4–8].

In this paper, we describe the QbE-STD system developed by the INESC-ID's Spoken Language Systems Laboratory (L^{2}F) for the Albayzin 2016 evaluation. Detailed information about the task and the data used for this evaluation can be found in the evaluation plan [9]. The L^{2}F proposed system is formed by the fusion of four individual sub-systems, and exploits an integrated approach composed of different modules. At the first stage, different frame-level phone posteriors are extracted from both queries and documents collection. Phone posteriors are obtained using two different phone decoder engines: the AUDIMUS in-house decoders [10] and the Brno University of Technology (BUT) [11] decoders. In total, seven different sub-systems based on 7 different language-dependent decoders have been built. Additionally, an audio segmentation module is used to segment the documents collection and then apply a DTW algorithm between each query and each sub-segment [12, 13]. This process results in a query detection candidate for each sub-segment of the collection, so that no further iterative DTW procedures are performed. Finally, the best results from each sub-system are retained and fused together following a discriminative approach [8].

This paper is organized as follows. First, the databases used in the QbE-STD task are briefly introduced in Section 2. Then, Section 3 describes the approaches followed in this work. Section 4 presents and discusses the performance of the baseline sub-systems and the fused ones. Finally, conclusions are given in Section 5.

2 Train and development data

Two different data sets have been provided for system evaluation: MAVIR and EPIC databases. However, in this work, only the MAVIR database is used for evaluation. The MAVIR database consists of a set of talks extracted from the MAVIR workshops [14] held in 2006, 2007, and 2008. This corpus amounts to about 7 hours of speech that are further divided for the purpose of the evaluation into training (4 hours), development (1 hour) and test sets (2 hours). In this work, the training data partition has not been used. Further details about the data used for this evaluation can be found in the evaluation plan [9].

3 Overview of the L^{2}F QbE-STD system

The system submitted for the Albayzin evaluation is composed of four main modules: feature extraction, speech segmentation, DTW-based query matching, and score calibration and fusion. Two different phone decoder engines were used to extract frame level phone posteriors, overall seven different acoustic models were used, leading to seven sub-systems. Then, the documents collection has been partitioned into small segments of speech, using the audio segmentation module of the in-house speech recognizer AUDIMUS [12, 13]. The search for a match is then performed with the segmental DTW algorithm applied between each segment and a query. This approach provides the benefit of improving the performance of the search using a reduced and parallelizable search space. Finally, the best results from each sub-system are retained, calibrated and fused together.

3.1 Feature Extraction

AUDIMUS decoders are based on hybrid connectionist methods [16]. Four phonetic decoders have been used exploiting four different language-dependent acoustic models trained for European Portuguese (PT), Brazilian Portuguese (BR), European Spanish (ES) and American English (EN). The acoustic models from each system are in fact multi-layer perceptron (MLP) networks that are part of L^{2}F in-house hybrid connectionist ASR system named AUDIMUS [10, 13]. AUDIMUS combines four MLP outputs trained with Perceptual Linear Prediction features (PLP, 13 static + first derivative), PLP with log-RelAtive SpecTrAl speech processing features (PLP-RASTA, 13 static + first derivative), Modulation SpectroGram features (MSG, 28 static) and Advanced Font-End from ETSI features (ETSI, 13 static + first and second derivatives). The language-dependent MLP networks were trained using different amounts of annotated data. Each MLP network is characterized by the input frame context size (13 for PLP, PLP- RASTA and ETSI; 15 for MSG), the number of units of the two hidden layers (500), and the size of the output layer. In this case, only monophone units are modeled, which results in posterior vectors of the following dimensionality: EN (41), PT (39), BR (40) and ES (30). Finally, frames for which the non-speech posterior is the highest unit are considered silence frames and they are discarded.

BUT decoders are based on Temporal Patterns Neural Network (TRAPs/NN) phone decoders [11]. The open software developed by the Brno University of Technology (BUT) provides acoustic models for Czech (CZ), Hungarian (HU) and Russian (RU), that have been exploited in this work to obtain frame-level phone posterior probabilities. The original phone state-level outputs and multiple non-speech units have been reduced to single-state phone outputs and a unique silence output unit, which results in feature vectors of 43, 59 and 50 log-likelihoods for the systems based on the CZ, HU and RU decoders, respectively.

Like in the case of the AUDIMUS decoders, frames with the non-speech class as the most likely one are removed.

3.2 Speech segmentation

The audio documents collection has been pre-processed using our in-house audio segmentation module [12] that was mostly developed for automatic segmentation and transcription of broadcast news. This module performs speech/non-speech classification, speaker segmentation, speaker clustering, gender and background conditions classification. The speech/non-speech segmentation is implemented using an artificial neural network of the multi-layer perceptron (MLP) type, based on perceptual linear prediction (PLP) features, followed by a finite state machine. This finite state machine smooths the input probabilities provided by the MLP network, using a median filter over a small window. The smoothed signal is then thresholded and analyzed using a time window ($t_{\min}$). The finite state machine consists of four possible states ("probable non-speech", "non-speech", "probable speech", and "speech"). If the input audio signal has a probability of "speech" above a given threshold, the finite state machine is placed into the "probable speech" state. If, after a given time interval ($t_{\min}$), the average speech probability is above a given confidence value, the machine changes to the "speech" state. Otherwise, it transitions to the "non-speech" state. The finite state machine generates segment boundaries for "non-speech" segments larger than the resolution of the median window. Additionally, "non-speech" segments larger than $t_{\min}$ are discarded. The value of $t_{\min}$ has been optimized to maximize non-speech detection.
With the speech segmentation module, we obtain for each document a partition into smaller segments. The resulting "speech" segments are further processed for query searching, while "non-speech" ones are discarded. This strategy offers two computational advantages. First, since the same query may occur multiple times in an audio document, a DTW-based search should proceed sequentially or iteratively, over all the audio document, storing the candidate matches found along the execution and initiating a new process with the remaining audio until a certain stopping criteria is met. By partitioning the audio document into smaller segments, the search could be parallelized, allowing for different searches of the same query at the same time. Second, since segments classified as not containing speech are discarded, the performance of the DTW benefits from the overall reduction of the search space. On the other hand, this strategy conveys at least two drawbacks that may affect the query matching ability of the proposed system. First, the errors of the audio segmentation module can result in missing speech segments that may eventually contain query terms that are lost. Second, we assume in this work that only a single match per query can occur in a sub-segment, which can eventually introduce miss detection errors.

3.3 DTW-based query matching

Given two sequences of feature vectors, corresponding to a spoken query q and to an audio document x, the cosine distance is computed between each pair of vectors $(q[i], x[j])$ as shown in Eq. 1. This information is used to build a distance matrix.

$$d(q[i], x[j]) = -\log \frac{q[i] \cdot x[j]}{|q[i]| \cdot |x[j]|} \tag{1}$$

The distance matrix is then normalized with regard to the audio document, such that matrix values are all comprised between 0 and 1 [17]. The normalization is performed as follows:

$$d_{norm}(q[i], x[j]) = \frac{d(q[i], x[j]) - d_{min}(i)}{d_{max}(i) - d_{min}(i)} \tag{2}$$

where:

$$d_{min}(i) = \min_{j=1,\ldots,n} d(q[i], x[j])$$

$$d_{max}(i) = \max_{j=1,\ldots,n} d(q[i], x[j])$$

In this way a perfect match would produce a quasi-diagonal sequence of zeroes. This normalization was found highly important for achieving good performance in the Mediaeval Spoken Web Search (SWS) 2013 [17].
Since the normalization needs to be computed on the whole audio document, this is done once for each audio of the collection. Then, the DTW procedure looks for the best alignment of the query under evaluation and a partition of the normalized distance matrix corresponding to a "speech" segment. The algorithm uses three additional matrices to store the accumulated distance of the optimal partial warping path found (AD), the length of the path (L), and the path itself.

The best alignment of a query in an audio document is defined as that minimizing the average distance in a warping path of the normalized distance matrix. A warping path may start at any given frame of x, k_1, then traverses a region of x which is optimally aligned to q, and ends at frame k_2. The average distance in this warping path is computed as:

$$d_{avg}(q, x) = AD[i, j]/L[i, j].$$

The detection score is computed as $1 - d_{avg}(q, x)$, thus ranging from 0 to 1, where 1 represents a perfect match. The starting time and the duration of each detection are obtained by retrieving the time offsets corresponding to frames k_1 and k_2 in the filtered audio document. This approach finds an alignment, and consequently a match candidate, for each query-segment pair. Subsequently, the detection results are filtered out to reduce the number of matches per query to a fixed amount of hypothesis. Different values, ranging from 50 to 500, were experimented in order to empirically determine the right threshold. It was found that the best results were achieved with a threshold equal to 100 query detection candidates for each hour of the data collection.

3.4 Systems fusion

The problems that should be addressed in the combination of STD systems are twofold: on the one hand there is the need to define a common set of candidates for all the systems, on the other hand multiple system scores have to be combined in order to produce a single score per candidate detection. In this work, the scores that are obtained by the different systems are transformed according to the approach described in [8]. In this approach, system scores are first normalized to have per-query zero-mean and unit-variance (*q-norm)*, thus allowing scores to be in the same range. Then, scores are filtered according to an heuristic scheme known as Majority Voting (MV). Under this approach, only candidate detections given by at least half of the systems are kept for further processing. Detections produced by different systems are aligned by considering their initial and final time stamps, i.e. if they partially overlap in time. Missing scores are hypothesized using a *per-query minimum* strategy, i.e. the minimum score produced by the system for that query. Then, the resulting list of scores of each system are used to estimate, through linear regression, the combination weights that result in well calibrated fused scores. Given that this procedure is expected to produce well-calibrated scores, the theoretical optimum Bayes threshold can then be used for making hard decisions.

4 Experimental evaluation

Seven basic QbE-STD systems were developed as described in Section 3, using the phone posterior features provided by the AUDIMUS decoders for European Portuguese (PT), Brazilian Portuguese (BR), European Spanish (ES) and American English (EN); and by the BUT decoders for Czech (CZ), Hungarian (HU) and Russian (RU). Table 1 reports the Actual/Maximum Term Weighted Value (ATWV/MTWV) achieved by these systems, on the development data set of the Albayzin 2016 QbE-STD task. Calibration and fusion parameters have been estimated on the development set. The decision threshold is theoretically determined and set to ~ 6.9[8].

Table 1. MTWV/ATWV performance for single QbE-STD systems. ATWV is shown for the optimal heuristic threshold in development set.

System	development	
	MTWV	ATWV
BR	0.092	0.063
CZ	0.022	0.000
EN	0.067	0.027
ES	0.102	0.064
HU	0.025	0.015
PT	0.077	0.042
RU	0.031	0.008

From Table 1 one can observe that system scores are not as well calibrated as expected, as revealed by the ATWV being far from the MTWV. We hypothesize that the linear regression estimation failed to provide a more accurate calibration configuration due to the small size of development data set. The lack of data for system calibration is known to be particularly critical when the task operation point is placed in a very low false-alarm region, as it is in the case of the Albayzin 2016 task.

As shown in Table 1, among the sub-systems obtained with the AUDIMUS decoder, the EN sub-system is the one that achieved the poorest performance. Further experiments that included the combination of the EN sub-system have always produced weaker results than when using the same combination but without this sub-system. For this reason, the EN sub-system was no longer included in subsequent experiments. Regarding the sub-systems obtained with the BUT decoder, one can observe that the best result was achieved by the RU sub-system, followed by the HU and CZ sub-systems. However, further experiments that included the combination of each of these sub-systems with the three best sub-systems obtained with the AUDIMUS decoder, have shown a different trend as reported in Table 2. In fact, the best result is achieved with the fusion of the BR, ES, PT, and CZ sub-systems. Moreover, this combination seems to provide the best calibration configuration. Consequently, the system resulting from this fusion of four sub-systems was selected as the L^{2}F primary submission to the Albayzin 2016 QbE-STD evaluation.

Table 2. MTWV/ATWV performance for the fusion of four QbE-STD sub-systems. ATWV is shown for the optimal heuristic threshold in development set.

System	development	
	MTWV	ATWV
BR, ES, PT, CZ	0.178	0.172
BR, ES, PT, HU	0.163	0.154
BR, ES, PT, RU	0.175	0.149

From Table 2, it seems that the scores resulting from the fusion of the four sub-systems are better calibrated than the ones obtained with the single sub-systems. Also, comparing Table 1 and Table 2, it is clear that the fusion of the four single sub-systems yields to remarkable MTWV improvements. From 0.102 obtained with the best individual sub-system (ES) to 0.178 (BR, ES, PT, CZ), which corresponds to more than 70% relative improvement. Overall, however, we believe that the homogeneity of the sub-systems, and also the presence of one sub-system more adequate for the task (same decoder language as the data evaluation language), limits the potential benefits of the fusion scheme.
Finally, the Detection Error Trade-off (DET) curve of the L^{2}F submitted system is shown in Figure 1. As anticipated previously, the actual system operation point is located in a very low false-alarm region (False Alarm around 0.002% and probability of Miss of 79.7%). In these operation conditions, very few query

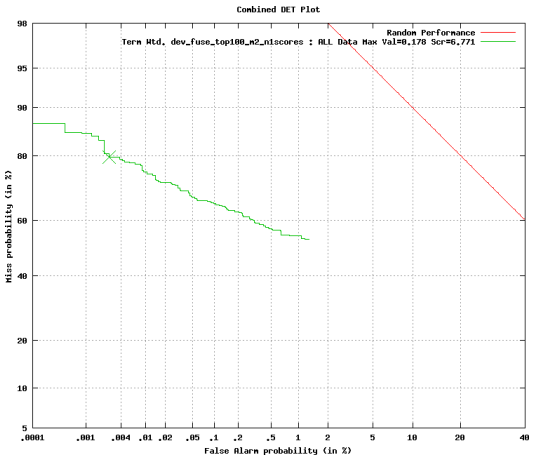

Fig. 1. DET curve for the fused system on the development set.

matches are hypothesized by the system as it can be noticed by the large steps of the DET curve, which generally results in poorer fusion and calibration configurations.

5 Conclusions

In this work, the L^{2}F QbE-STD Albayzin 2016 system formed by the combination of posterior-based DTW query matching sub-systems has been described. In particular, two different phone decoder engines –AUDIMUS and BUT– have been used to extract frame level phone posteriors with different acoustic models. In order to ease the search process, the audio document collection was partitioned into smaller segments, which allowed for a computationally optimized search with a reduced search space. Then, DTW was applied to search for each query in every segment of the audio documents. In the last step, the scores produced by the different sub-systems were normalized, filtered, and combined together with other sub-systems with the aim of obtaining well-calibrated scores. The different possible sub-system combinations have been exhaustively explored and a summary of the most remarkable results have been reported in this document. In accordance with these results, the L^{2}F submitted system was finally composed by the fusion of 4 sub-systems.
Overall, we acknowledge that the results achieved by the submitted system are below the current state of the art for this task. Thus, as future work, we plan to incorporate heterogeneous sub-systems based on AKWS, which will very likely provide performance improvements, as we have previously observed in similar tasks.

6 Acknowledgements

This work was supported by national funds through – Fundação para a Ciência e a Tecnologia (FCT), under Grants SFRH/BD/97187/2013 and with reference UID/CEC/50021/2013.

References

1. X.Anguera, "Telefonica system for the spoken web search task at MediaEval 2011," in Proc. MediaEval Workshop, 2011.
2. A. Muscariello, G. Gravier, and F. Bimbot, "A zero-resource system for audio-only spoken term detection using a combination of pattern matching techniques," in Proc. Interspeech, 2011.
3. I. Szöke, J. Tejedor, M. Fapso, and J. Colás, "BUT-HCTLab approaches for spoken web search," in Proc. MediaEval Workshop, 2011.
4. A. Abad and R. F. Astudillo, "The L2F Spoken Web Search system for MediaEval 2012," in Proc. MediaEval 2012 Workshop, 2012.
5. N. Brummer, L. Burget, J. Cernocky, O. Glembek, F. Grezl, M. Karafiat, D. van Leeuwen, P. Matejka, P. Schwarz, and A. Strasheim, "Fusion of Heterogeneous Speaker Recognition Systems in the STBU Submission for the NIST Speaker Recognition Evaluation 2006," IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, no. 7, pp. 2072-2084, 2007.
6. L. Rodríguez, M. Peñagarikano, A. Varona, M. Díez, G. Bordel, D. Martínez, J. Villalba, A. Miguel, A. Ortega, A. Lleida, A. Abad, O. Koller, I. Trancoso, P. Lopez-Otero, L. Fernández, C. García-Mateo, R. Saeidi, M. Soufifar, T. Kinnunen, T. Svendsen, and P. Fränti, "Multi-site heterogeneous system fusions for the Albayzin 2010 Language Recognition Evaluation," in Proc. ASRU, 2011.
7. H. Wang, T. Lee, C.-C. Leung, B. Ma, and H. Li, "Using parallel tokenizers with DTW matrix combination for low-resource Spoken Term Detection," in Proc. ICASSP, 2013.
8. A. Abad, Luis J. Rodríguez Fuentes, M. Peñagarikano, A. Varona, M. Díez, and G. Bordel, "On the calibration and fusion of heterogeneous spoken term detection systems," in Interspeech, Lyon, France, August 25-29, 2013.
9. Javier Tejedor, and Doroteo T. Toledano, "The ALBAYZIN 2016 Search on Speech Evaluation Plan", in Proc. IberSPEECH, 2016.
10. H. Meinedo, A. Abad, T. Pellegrini, I. Trancoso, and J. Neto, "The L2F Broadcast News Speech Recognition System," in Proc. Fala, 2010.
11. P. Schwarz, "Phoneme recognition based on long temporal context," Ph.D. dissertation, Faculty of Information Technology, Brno University of Technology, http://www.fit.vutbr.cz/, Brno, Czech Republic, 2008.
12. Hugo Meinedo, João Neto, "A Stream-based Audio Segmentation, Classification and Clustering Pre-processing System for Broadcast News using ANN Models", in Proc. Interspeech 2005.
13. A. Abad, J. Luque, and I. Trancoso, "Parallel Transformation Network features for Speaker Recognition", In Proc. ICASSP, 2011.
14. MAVIR corpus. http://www.lllf.uam.es/ESP/CorpusMavir.html
15. SoX - Sound eXchange. http://sox.sourceforge.net/
16. N. Morgan and H. Bourlad, "An introduction to hybrid HMM/connectionist continuous speech recognition," IEEE Signal Processing Magazine, vol. 12, no. 3, pp. 25-42, 1995.

17. L. J. Rodríguez-Fuentes, A. Varona, M. Peñagarikano, G. Bordel and M. Díez, "High-performance Query-by-Example Spoken Term Detection on the SWS 2013 evaluation," 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, 2014, pp. 7819-7823.