The SPL-IT-UC Query by Example Search on Speech system for MediaEval 2015

Jorge Proença[1,2], Luis Castela[2], Fernando Perdigão[1,2]
[1] Instituto de Telecomunicações, Coimbra, Portugal
[2] Electrical and Computer Eng. Department, University of Coimbra, Portugal
{jproenca, fp}@co.it.pt

ABSTRACT

This document describes the system built by the SPL-IT-UC team from the Signal Processing Lab of Instituto de Telecomunicações (pole of Coimbra) and University of Coimbra for the Query by Example Search on Speech Task (QUESST) of MediaEval 2015. The submitted system filters considerable background noise by applying spectral subtraction, uses five phonetic recognizers from which posterior probabilities are extracted as features, implements novel modifications of Dynamic Time Warping (DTW) that focus on complex queries, and uses linear calibration and fusion to optimize results. This year's task proved extra challenging in terms of acoustic conditions and match cases, though we observe the best results when merging all complex approaches.

1. INTRODUCTION

MediaEval's challenge for audio query search on audio, QUESST [1], keeps adding new relevant problems to tackle, and in depth details can be consulted in the referenced paper. Briefly, this year, the audio can have significant background or intermittent noise as well as reverberation, and there are queries that originate from spontaneous requests. These conditions further approach real case scenarios of a query search application, which is one of the underlying motivations of the challenge.

Systems for Query by Example (QbE) search keep improving with recent advances such as combining spectral acoustic and temporal acoustic models [2], combining a high number of subsystems using both Acoustic Keyword Spotting (AKWS) and Dynamic Time Warping (DTW) and using bottleneck features of neural networks as input [3], new distance normalization techniques [4] and several approaches to system fusion and calibration [5]. Some attempts have been made to address complex query types that were introduced in QUESST 2014, by segmenting the query in some way such as using a moving window [6] or splitting the query in the middle [7]. Our approach is based on modifying the DTW algorithm to allow paths to be created in ways that conform to the complex types, which has shown success in improving overall results [8, 9].

Using bottleneck features could be an important step to improve our systems, and although we did not consider them yet, we are certain that there are still improvements to be made not related to the feature extractor. Nevertheless, we tried to improve the system feature-wise by implementing additional phonetic recognizers. Compared to last year's submission, we reduce severe background noise in the audio by applying spectral subtraction, add two phonetic recognizers (five in total) from which to extract posteriorgrams, remove all silence and noise frames from queries, improve and add modifications to Dynamic Time Warping (DTW) for complex query search (filler inside a query being the novelty), and implement better fusion and calibration methods.

Copyright is held by the author/owner(s).
MediaEval 2015 Workshop, September 14-15, 2015, Wurzen, Germany

2. SYSTEM DESCRIPTION

2.1 Noise filtering

First, we apply a high pass filter to the audio signals to remove low frequency artefacts. Then, to tackle the existence of substantial stationary background noise in both queries and reference audio, we apply spectral subtraction (SS) to noisy signals (not performed for high SNR signals, which was worsening results). This implies a careful selection of samples of noise from an utterance. For this, we analyze the log mean Energy of the signal, consider only values above -60dB and calculate a threshold below which segments of more than 100ms are selected as "noise" samples, whose mean spectrum will be subtracted from the whole signal.

Using dithering (white noise) to counterbalance the musical noise effect due to SS didn't help. Nothing was specifically performed for reverberation or intermittent noise.

2.2 Phonetic recognizers

We continue to use an available external tool based on neural networks and long temporal context, the phoneme recognizers from Brno University of Technology (BUT) [10]. We used the three available systems for 8kHz audio, for three languages: Czech, Hungarian and Russian. Additionally, we trained two new systems with the same framework: English (using TIMIT and Resource Management databases) and European Portuguese (using annotated broadcast news data and a dataset of command words and sentences). Using different languages implies dealing with different sets of phonemes, and the fusion of the results will better describe the similarities between what is said in a query and the searched audio. This makes our system a low-resource one.

All de-noised queries and audio files were run through the 5 systems, extracting frame-wise state-level posterior probabilities (with 3 states per phoneme) to be analyzed separately.

2.3 Voice Activity Detection

Silence or noise segments are undesirable for a query search, and were cut on queries from all frames that had a high probability of corresponding to silence or noise, if the sum of the 3 state posteriors of silence or noise phones is greater than a 50% threshold for the average of the 5 languages. To account for queries that may still have significant noise, this threshold is incrementally raised if the previous cut is too severe (the obtained query having less than 500ms).

2.4 Modified Dynamic Time Warping

Every query must then be searched on every reference audio. We continue to use DTW as a way to find paths on the audio that may be similar to the query. As in [11], the local distance is based on the dot product of query and audio posterior probability vectors, with a back-off of 10^{-4} and minus log applied to the dot product, resulting in the local distance matrix of a query-audio pair, ready for DTW to be applied. Before dot product, removing the posterior probabilities of silence and noise altogether, and normalizing the probabilities of speech phones to 1 did not help.

In addition to separate searches on distance matrices from posteriorgrams of 5 languages, we add a 6th "language"/sub-system (called ML for multi-language) whose distance matrix is the average of the 5 matrices.

We employ several modifications to the DTW algorithm to allow intricate paths to be constructed that can correspond logically to the complex match scenarios of query and audio. The basic approach (named A1) outputs the best path (minimal normalized distance) by using the complete query and allows 3 movements in the distance matrix with equal weight: horizontal, vertical and diagonal. As in previous work [8, 9], 4 modifications are made that do not require repeating DTW with segmentations of the query:

(A2) Considering cuts at the end of query for lexical variations;

(A3) Considering cuts at the beginning of the query;

(A4) Allowing one 'jump' for extra content in the audio;

(A5) Allowing word-reordering, where an initial part of the query may be found ahead of the last.

Additionally, we designed an extra approach to address the case of type 3 queries where small fillers or hesitations in the query may exist:

(A6) Allowing one 'jump' along the query, of maximum 33% of query duration.

Each distance is normalized per the number of movements (mean distance of the path).

2.5 Fusion and Calibration

At this stage, we have distance values for each audio-query pair for 6 languages and 6 DTW strategies (36 vectors). First, modifications are performed on the distribution per query per strategy. While deciding on a maximum distance value to give to unsearched cases (such as an audio being too short for a long query), we found that drastically truncating large distances (lowering to the same value) improved results. Surprisingly, changing all upper distance values (larger than the mean) to the mean of the distribution was the overall best. We reason that since there are a lot of ground truth matches with very high distances (false negatives), lowering these values improves the Cnxe metric more than lowering the value of true negatives worsens it. The next step is to normalize per query by subtracting the new mean and dividing by the new standard deviation. Distances are transformed to figures-of-merit by taking the symmetrical value.

To fuse results of different strategies and languages we have two separate approaches/systems, both using weighted linear fusion and transformation trained with the Bosaris toolkit [12], calibrating for the Cnxe metric by taking into account the prior defined by the task:

- Fusion of all approaches and all languages (36 vectors).
- Fusion of the Harmonic mean (Hmean) of the 6 strategies for a given language (6 vectors, one per language). This is done to possibly counter overfitting to the training data from weighing 36 vectors, and only languages are weighed.

From a fusion, final result vectors with only one value per audio-query pair are obtained for development and test data.

Additionally, we provide side-info based on query and audio, added as extra vectors for all fusions. The 7 extra side-info vectors are: mean of distances per query before truncation and normalization from the best approach and language (the highest weighted from fusion of all); query size in frames and log of query size; 4 vectors of SNR values (original SNR of query and of audio, post spectral subtraction SNR of query and of audio).

2.6 Processing Speed

The hardware that processed our systems was the CRAY CX1 Cluster, running windows server 2008 HPC, and using 16 of 56 cores (7 nodes with double Intel Xeon 5520 2.27GHz quad-core and 24GB RAM per node).

Approximately, the Indexing Speed Factor was 2.14, Searching Speed Factor was 0.0034 per sec, and Peak Memory was 120MB.

3. SUBMISSIONS AND RESULTS

We submitted 4 systems for evaluation: fusion of all approaches and languages with and without side-info; fusion of harmonic mean with and without side-info. Table 1 summarizes the results of the Cnxe metric for the 4 systems.

Table 1. Results of Cnxe (and MinCnxe) for development and evaluation datasets.

Fusion Systems	Dev: Cnxe, MinCnxe	Eval: Cnxe, MinCnxe
All + side-info	**0.7782**, 0.7716	0.7866, 0.7809
Hmean + side-info	0.7862, 0.7800	**0.7842**, 0.7786
All, no side	0.7873, 0.7816	0.7930, 0.7875
Hmean, no side	0.7957, 0.7893	0.7914, 0.7865

It can be seen that the best result for the development set was our primary system that fused all languages and approaches plus some side-info. As suspected, the weighted combination applied to the Eval set may be too over fitted to the Dev set, as the best result on Eval was using the Hmean of approaches. The considered side-info always helped.

Analyzing the best result on Eval per query type (All-0.7842, T1-0.7107, T2-0.8147, T3-0.8115), the exact matches of type 1 are the easiest to detect compared to other types.

Using only each individual DTW approach and fusing, the results on the Dev set are: A1: 0.8041, A2: 0.7978, A3: 0.8335, A4: 0.8137, A5: 0.8184, A6: 0.8460. The overall best performing strategy is allowing cuts at the end of the query (A2), which may help in all cases due to co-articulation or intonation. The new strategy of allowing a jump in query (A6) performs badly and should be reviewed. Actually, a filler in a query may be an extension of an existing phone, which leads to a straight path and not a jump.

Below, we also report the improvements of some steps of our system on the Dev set (although the comparison may not be to the final approach). Using Spectral Subtraction resulted in 0.8130 Cnxe from 0.8368. Fusing 5 languages - 0.7971 Cnxe, using only the mean distance matrix ML - 0.8136, using 5 langs and ML - 0.7873. Using per query truncation to the mean – 0. 7873 Cnxe, without truncation – 0.7939.

4. CONCLUSIONS

Several steps were explored to improve the results of existing methods, and the main contributions came from: a careful Spectral Subtraction to diminish background noise which greatly influences the output of phonetic recognizers; using the average distance matrix of all languages as a 6th sub-system for fusion; including side-info of query and audio; and per-query truncation of large distances.

Including a DTW strategy that considers gaps in query did not prove very successful. This may be due to its target cases being too few in the dataset, and even some fillers in query being extensions and not unrelated hesitations.

5. REFERENCES

[1] I. Szöke, L.J. Rodriguez-Fuentes, A. Buzo, X. Anguera, F. Metze, J. Proença, M. Lojka, and X. Xiong, "Query by Example Search on Speech at Mediaeval 2015", in Working Notes Proceedings of the Mediaeval 2015 Workshop, Wurzen, Germany, September 14-15

[2] C. Gracia, X. Anguera, and X. Binefa, "Combining temporal and spectral information for Query-by-Example Spoken Term Detection," in Proc. European Signal Processing Conference (EUSIPCO), Lisbon, Portugal, 2014, pp. 1487-1491.

[3] I. Szöke, L. Burget, F. Grezl, J.H. Cernocky, and L. Ondel, "Calibration and fusion of query-by-example systems—BUT SWS 2013," in *Proc IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, 2014, pp. 7849-7853.

[4] L.J. Rodriguez-Fuentes, A. Varona, M. Penagarikano, G. Bordel, and M. Diez, "High-performance query-by-example spoken term detection on the SWS 2013 evaluation," in *Proc IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, 2014, pp. 7819-7823.

[5] A. Abad, L.J. Rodriguez Fuentes, M. Penagarikano, A. Varona, M. Diez, and G. Bordel, "On the calibration and fusion of heterogeneous spoken term detection systems," in *Proc. Interspeech 2013*, Lyon, France, 2013, pp. 20-24.

[6] P. Yang, et al., "The NNI Query-by-Example System for MediaEval 2014", in Working Notes Proceedings of the Mediaeval 2014 Workshop, Barcelona, Spain, October 16-17

[7] I. Szöke, M. Skácel and L. Burget, "BUT QUESST 2014 system description", in Working Notes Proceedings of the Mediaeval 2014 Workshop, Barcelona, Spain, October 16-17

[8] J. Proença, A. Veiga and F. Perdigão, "The SPL-IT Query by Example Search on Speech system for MediaEval 2014", in Working Notes Proceedings of the Mediaeval 2014 Workshop, Barcelona, Spain, October 16-17

[9] J. Proença, A. Veiga and F. Perdigão, "Query by Example Search with Segmented Dynamic Time Warping for Non-Exact Spoken Queries", Proc European Signal Processing Conf. - EUSIPCO, Nice, France, August, 2015.

[10] Phoneme recognizer based on long temporal context, Brno University of Technology, FIT, http://speech.fit.vutbr.cz/software/phoneme-recognizer-based-long-temporal-context

[11] T.J. Hazen, W.Shen and C.M. White, "Query-by-example spoken term detection using phonetic posteriorgram templates", In *ASRU 2009*: 421-426.

[12] N. Brummer, and E. de Villiers, "The BOSARIS Toolkit User Guide: Theory, Algorithms and Code for Binary Classifer Score Processing," Technical report, 2011. https://sites.google.com/site/bosaristoolkit/