IberSPEECH 2018
21-23 November 2018, Barcelona, Spain

AUDIAS-CEU: A Language-independent approach for the Query-by-Example Spoken Term Detection task of the Search on Speech ALBAYZIN 2018 evaluation

Maria Cabello[1], *Doroteo T. Toledano*[1], *Javier Tejedor*[2]

[1]AUDIAS, Universidad Autónoma de Madrid
[2]Universidad San Pablo-CEU, CEU Universities

m.cabello.a@gmail.com, doroteo.torre@uam.es, javier.tejedornoguerales@ceu.es

Abstract

Query-by-Example Spoken Term Detection is the task of detecting query occurrences within speech data (henceforth utterances). Our submission is based on a language-independent template matching approach. First, queries and utterances are represented as phonetic posteriorgrams computed for English language with the phoneme decoder developed by the Brno University of Technology. Next, the Subsequence Dynamic Time Warping algorithm with a modified Pearson correlation coefficient as cost measure is employed to hipothesize detections. Results on development data showed an ATWV=0.1774 with MAVIR data and an ATWV=0.0365 with RTVE data.
Index Terms: language-independent QbE STD, template matching, SDTW

1. Introduction

The large amount of heterogeneous speech data stored in audio and audiovisual repositories makes it necessary to develop efficient methods for speech information retrieval. There are different speech information retrieval tasks, including spoken document retrieval (SDR), keyword spotting (KWS), spoken term detection (STD), and query-by-example spoken term detection (QbE STD). The advance of the technology in these tasks have been evaluated through different international evaluations related to SDR, STD, and QbE STD [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11] for different languages (English, Arabic, Mandarin, Spanish, Japanese, and low-resource language such as Swahili, Tamil, and Vietnamese). Specifically, Spanish language has been employed with the STD/QbE STD ALBAYZIN evaluations held in 2012, 2014 and 2016 [7, 8, 9, 10, 11].

Query-by-example Spoken Term Detection aims to retrieve data from a speech repository (henceforth utterance) given an acoustic query containing the term of interest as input. QbE STD has been mainly addressed from three different approaches: methods based on the word/subword transcription of the query that typically employ a word/phone-based speech recognition system for query detection [12, 13], methods based on template matching of features that are typically based on posteriorgram-based units and DTW-like search for query detections [14, 15, 16, 17], and hybrid approaches that take advantage of both approaches [18, 19, 20, 21].

This paper presents the system submitted by the AUDIAS-CEU research team to the QbE STD task of the Search on Speech ALBAYZIN 2018 evaluation [22], which deals with the Spanish language. However, our submission does not employ the target language, and hence is based on a language-independent approach that can be used for any target language. First, phoneme posteriorgrams are computed for query/utterance representation. These phoneme posteriorgrams are computed from neural networks trained for the English language. Next, the Subsequence Dynamic Time Warping (S-DTW) algorithm generates the query occurrences. Our system is largely based on the winner system of the QbE STD task of the Search on Speech ALBAYZIN 2016 evaluation [23].

The rest of the paper is organized as follows: Section 2 presents the system submitted to the evaluation, Section 3 presents the experiments and the results obtained in the development data provided by the organizers, and Section 4 concludes the paper.

2. QbE STD system

The QbE STD system, whose architecture is presented in Figure 1, integrates two different stages: feature extraction and query detection, which are explained next.

Figure 1: *QbE STD system architecture.*

2.1. Feature extraction

The English phoneme recognizer development by the Brno University of Technology [24] has been employed to compute 3-state phoneme posteriorgrams that represent both the queries and the utterances. This phoneme recognizer contains 39 units, which correspond to the 39 phonemes in English plus a non-speech unit to represent some other phenomena in speech such as laugh, noise, short silences, etc. These phoneme posteriorgrams have been computed each 10ms of speech, which have been further processed to compute a single posterior probability per phoneme. This posterior probability is obtained by summing the posterior probabilities of the three states corresponding to the given phoneme.

2.2. Query detection

Query detection involves different stages, as presented in Figure 2.

First, a cost matrix that stores the *similarity* between every query/utterance pair is computed. The Pearson correlation coefficient [25] has been employed to build the cost matrix, as presented in Equation 1:

10.21437/IberSPEECH.2018-51

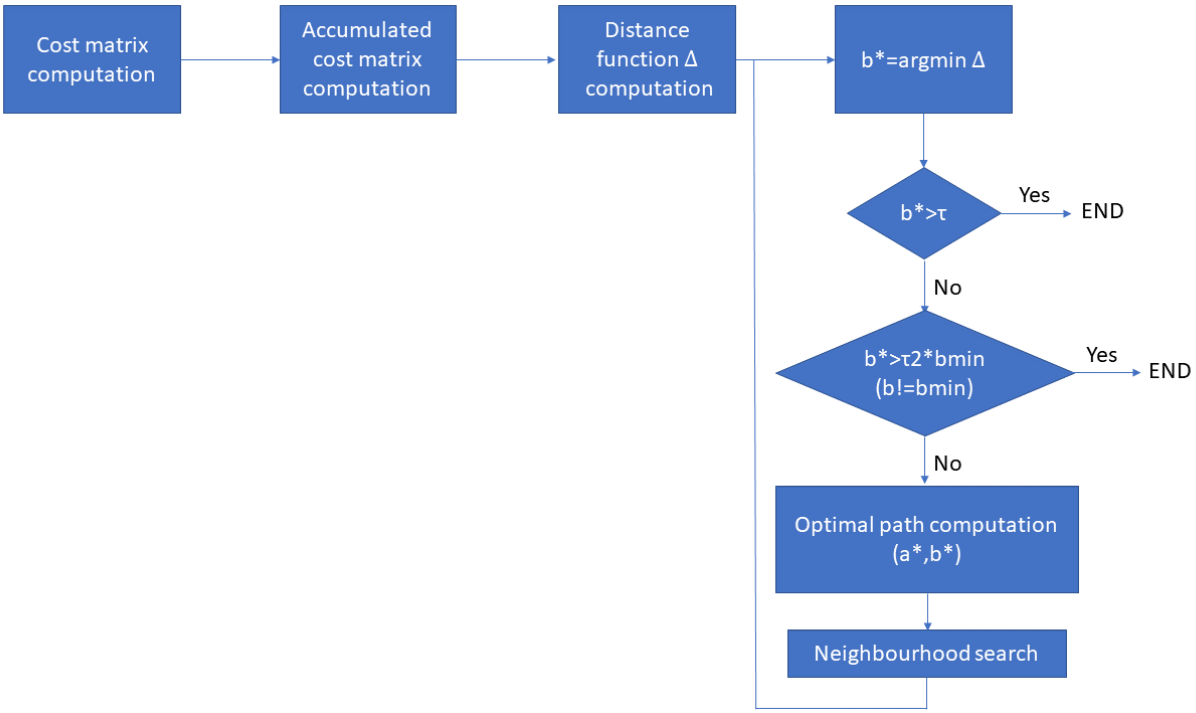

Figure 2: *Query detection stages.*

$$r(x_n, y_m) = \frac{U(x_n \cdot y_m) - ||x_m||||y_m||}{\sqrt{U||x_n^2|| - ||x_n||^2)(U||y_m^2|| - ||y_m||^2)}}, \quad (1)$$

where x_n represents the query phoneme posteriorgrams, y_m represents the utterance phoneme posteriorgrams, and U represents the number of phoneme units (40 in our case).

Then, the Pearson correlation coefficient is mapped into the interval [0 1], as given in Equation 2:

$$c(x_n, y_m) = \frac{1 - r(x_n, y_m)}{2}, \quad (2)$$

where $c(x_n, y_m)$ represents the cost matrix used during the S-DTW search.

Therefore, the cost $c(x_n, y_m)$ can take the values of 1 (when $r = -1$), 0.5 (when $r = 0$), or 0 (when $r = 1$). Figure 3 represents the cost matrix example with the *standard* Pearson correlation coefficient computation.

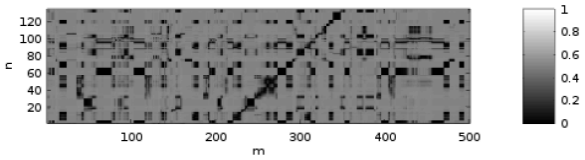

Figure 3: *Cost matrix example with the standard Pearson correlation coefficient.*

The final cost used during the search has been modified as follows: When $r <= 0$, r has been assigned the value of 0. Next, $c(x_n, y_m) = 1 - r(x_n, y_m)$. Therefore, for all the Pearson correlation coefficient values lower or equal to 0, the cost will be maximum, hence promoting the differences between aligned and non-aligned sequences in the next stage.

Figure 4 shows the cost matrix example with this modification of the Pearson correlation coefficient computation. This modification leads to more differences in the costs between query and utterance frames.

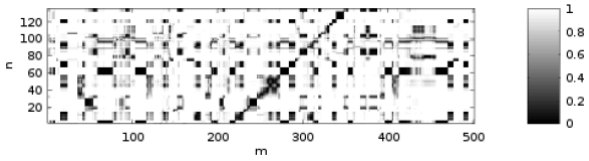

Figure 4: *Cost matrix example with the modified Pearson correlation coefficient.*

2.3. Subsequence Dynamic Time Warping-based search

The S-DTW algorithm [26] has been used to hipothesize query detections within the utterances. From the cost matrix $c(x_n, y_m)$, the accumulated cost matrix employed within the search is computed as given in Equation 3:

$$\mathrm{D}_{n,m} = \begin{cases} c(\mathrm{x}_n, \mathrm{y}_m) & \text{if} \quad n = 0 \\ c(\mathrm{x}_n, \mathrm{y}_m) + \mathrm{D}_{n-1,0} & \text{if} \quad n > 0,\ m = 0 \\ c(\mathrm{x}_n, \mathrm{y}_m) + \mathrm{D}^*(n, m) & \text{else,} \end{cases} \quad (3)$$

where

$$\mathrm{D}^*(n, m) = min\left(\mathrm{D}_{n-1,m}, \mathrm{D}_{n-1,m-1}, \mathrm{D}_{n,m-1}\right), \quad (4)$$

which implies that only horizontal, vertical, and diagonal path movements are allowed in the search.

Figure 5 shows the accumulated cost matrix from the cost matrix presented in Figure 3 (i.e., with the standard Pearson correlation coefficient computation), and Figure 6 shows that of the cost matrix presented in Figure 4 (i.e., with the modified Pearson correlation coefficient). The accumulated cost matrix from the modified Pearson correlation coefficient shows more cost in non-occurrence regions, which favors the final query detection.

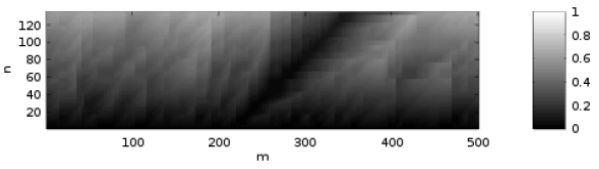

Figure 5: *Accumulated cost matrix from Figure 3 with the standard Pearson correlation coefficient.*

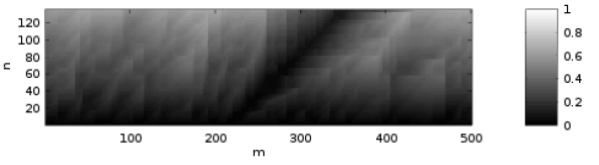

Figure 6: *Accumulated cost matrix from Figure 4 with the modified Pearson correlation coefficient.*

To hypothesize detections, a distance function Δ needs to be defined. This value corresponds to the last row/column of the accumulated cost matrix D, and is used as the initial value from which the minimum values that suggest possible paths are computed. Then, the S-DTW computes a global minimum b_{min} (also referred as $b*$ in Figure 2) in the accumulated cost matrix from which all the possible query detections are considered. This b_{min} has to be lower than a predefined threshold τ tuned on MAVIR development data. Next, the S-DTW computes all the local minima $b*$ that appear in the accumulated cost matrix. These local minima $b*$ need to be lower than a second threshold $\tau_2 = \tau * b_{min}$ to be considered as optimal paths where query detections reside. This second threshold τ_2 has been also tuned on MAVIR development data.

For each of values of $b*$ that meets the afore-mentioned conditions, the optimal paths ($a*$, $b*$) that represent the query detections are found as follows: Let $p = (p_1, ..., p_l)$ be a possible optimal path. Starting at $p_l = b*$, a reverse path that ends at $n = 1$ (i.e., the first frame of the query) is computed as presented in Equation 5:

$$p_{l-1} = argmin(D(n-1, m-1), D(n-1, m), D(n, m-1)). \quad (5)$$

Finally, a neighbourhood search is carried out so that all the paths (i.e., query detections) which overlap $500ms$ from a previously obtained optimal path are rejected in the final system output. An example of an optimal path found is presented in Figure 7.

3. Experiments and results

Experiments are carried out on the development data provided by the organizers for the QbE STD task. Two different databases were experimented with: MAVIR database, which

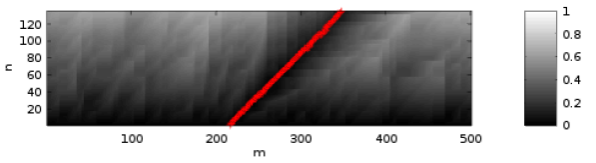

Figure 7: *Query detection example from the accumulated cost matrix in Figure 6.*

comprises a set of talks extracted from the Spanish MAVIR workshops[1] held in 2006, 2007, and 2008 (Corpus MAVIR 2006, 2007, and 2008) corresponding to Spanish language, and RTVE database, which comprises different Radio TeleVisión Española (RTVE) programs recorded from 2015 to 2018. For the MAVIR database, about 1 hour of speech material in total, extracted from 2 audio files was provided by the organizers, in which 102 queries extracted from the same development data were searched. For the RTVE database, about 15 hours of speech, extracted from 12 audio files, were provided by the organizers. In these RTVE data, 103 queries were searched. Organizers also provided with training data. However, since our submission is based on a previously trained phoneme recognizer for English language, the system does not employ any training data.

Results are shown in Table 1 for MAVIR and RTVE development data. These results show moderate performance for MAVIR data and a worse performance on RTVE data. This worse performance on RTVE data may be due to the optimal parameters found on MAVIR data were employed for the RTVE data, and no additional tuning on these data were carried out. Better performance should be obtained in case RTVE data were also fine-tuned.

Table 1: *System results for MAVIR and RTVE development data.*

Database	MTWV	ATWV	p(FA)	p(Miss)
MAVIR	0.1823	0.1774	0.00002	0.801
RTVE	0.0365	0.0365	0.00000	0.963

4. Conclusions

This paper presents the AUDIAS-CEU submission for the QbE STD task of the Search on Speech ALBAYZIN 2018 evaluation. The system relies on a language-independent approach for QbE STD, since no prior information of the Spanish language is employed for system building. Phoneme posteriorgrams and Subsequence Dynamic Time Warping with a modified Pearson correlation coefficient as cost measure were employed for system construction. System design was largely based on the winner system of the QbE STD task of the Search on Speech ALBAYZIN 2016 evaluation [23].

Future work will include some fusion techniques to get advantage of the query detections from different phoneme decoders, and feature selection techniques to retain the most meaningful phoneme units for each language.

[1] http://www.mavir.net

5. Acknowledgements

This work was partially supported by the project "DSSL: Redes Profundas y Modelos de Subespacios para Deteccin y Seguimiento de Locutor, Idioma y Enfermedades Degenerativas a partir de la Voz" (TEC2015-68172-C2-1-P, MINECO/FEDER).

6. References

[1] J. G. Fiscus, J. Ajot, J. S. Garofolo, and G. Doddingtion, "Results of the 2006 spoken term detection evaluation," in *Proc. of SSCS*, 2007, pp. 45–50.

[2] NIST, *The Ninth Text REtrieval Conference (TREC 9)*, Accesed: February, 2018. [Online]. Available: http://trec.nist.gov

[3] H. Joho and K. Kishida, "Overview of the NTCIR-11 SpokenQuery&Doc Task," in *Proc. of NTCIR-11*, 2014, pp. 1–7.

[4] X. Anguera, F. Metze, A. Buzo, I. Szöke, and L. J. Rodriguez-Fuentes, "The spoken web search task," in *Proc. of MediaEval*, 2013, pp. 921–922.

[5] X. Anguera, L. J. Rodriguez-Fuentes, I. Szöke, A. Buzo, and F. Metze, "Query by example search on speech at Mediaeval 2014," in *Proc. of MediaEval*, 2014, pp. 351–352.

[6] NIST, *Draft KWS14 Keyword Search Evaluation Plan*, National Institute of Standards and Technology (NIST), Gaithersburg, MD, USA, December 2013. [Online]. Available: https://www.nist.gov/sites/default/files/documents/itl/iad/mig/KWS14-evalplan-v11.pdf

[7] J. Tejedor, D. T. Toledano, P. Lopez-Otero, L. Docio-Fernandez, C. Garcia-Mateo, A. Cardenal, J. D. Echeverry-Correa, A. Coucheiro-Limeres, J. Olcoz, and A. Miguel, "Spoken term detection ALBAYZIN 2014 evaluation: overview, systems, results, and discussion," *EURASIP, Journal on Audio, Speech and Music Processing*, vol. 2015, no. 21, pp. 1–27, 2015.

[8] J. Tejedor, D. T. Toledano, X. Anguera, A. Varona, L. F. Hurtado, A. Miguel, and J. Colás, "Query-by-example spoken term detection ALBAYZIN 2012 evaluation: overview, systems, results, and discussion," *EURASIP, Journal on Audio, Speech, and Music Processing*, vol. 2013, no. 23, pp. 1–17, 2013.

[9] J. Tejedor, D. T. Toledano, P. Lopez-Otero, L. Docio-Fernandez, and C. Garcia-Mateo, "Comparison of ALBAYZIN query-by-example spoken term detection 2012 and 2014 evaluations," *EURASIP, Journal on Audio, Speech and Music Processing*, vol. 2016, no. 1, pp. 1–19, 2016.

[10] J. Tejedor, D. T. Toledano, P. Lopez-Otero, L. Docio-Fernandez, L. Serrano, I. Hernaez, A. Coucheiro-Limeres, J. Ferreiros, J. Olcoz, and J. Llombart, "Albayzin 2016 spoken term detection evaluation: an international open competitive evaluation in spanish," *EURASIP, Journal on Audio, Speech and Music Processing*, vol. 2017, no. 22, pp. 1–23, 2017.

[11] J. Tejedor, D. T. Toledano, P. Lopez-Otero, L. Docio-Fernandez, J. Proença, F. P. ao, F. García-Granada, E. Sanchis, A. Pompili, and A. Abad, "Albayzin query-by-example spoken term detection 2016 evaluation," *EURASIP, Journal on Audio, Speech, and Music Processing*, vol. 2018, no. 2, pp. 1–25, 2018.

[12] N. Sakamoto, K. Yamamoto, and S. Nakagawa, "Combination of syllable based N-gram search and word search for spoken term detection through spoken queries and IV/OOV classification," in *Proc. of ASRU*, 2015, pp. 200–206.

[13] R. Konno, K. Ouchi, M. Obara, Y. Shimizu, T. Chiba, T. Hirota, and Y. Itoh, "An STD system using multiple STD results and multiple rescoring method for NTCIR-12 SpokenQuery&Doc task," in *Proc. of NTCIR-12*, 2016, pp. 200–204.

[14] P. Lopez-Otero, L. Docio-Fernandez, and C. Garcia-Mateo, "Phonetic unit selection for cross-lingual Query-by-Example spoken term detection," in *Proc. of ASRU*, 2015, pp. 223–229.

[15] A. H. H. N. Torbati and J. Picone, "A nonparametric bayesian approach for spoken term detection by example query," in *Proc. of Interspeech*, 2016, pp. 928–932.

[16] H. Chen, C.-C. Leung, L. Xie, B. Ma, and H. Li, "Unsupervised bottleneck features for low-resource Query-by-Example spoken term detection," in *Proc. of Interspeech*, 2016, pp. 923–927.

[17] Y. Yuan, C.-C. Leung, L. Xie, H. Chen, B. Ma, and H. Li, "Pairwise learning using multi-lingual bottleneck features for low-resource Query-by-Example spoken term detection," in *Proc. of ICASSP*, 2017, pp. 5645–5649.

[18] S. Oishi, T. Matsuba, M. Makino, and A. Kai, "Combining state-level spotting and posterior-based acoustic match for improved query-by-example spoken term detection," in *Proc. of Interspeech*, 2016, pp. 740–744.

[19] M. Obara, K. Kojima, K. Tanaka, S. wook Lee, , and Y. Itoh, "Rescoring by combination of posteriorgram score and subword-matching score for use in Query-by-Example," in *Proc. of Interspeech*, 2016, pp. 1918–1922.

[20] C.-C. Leung, L. Wang, H. Xu, J. Hou, V. T. Pham, H. Lv, L. Xie, X. Xiao, C. Ni, B. Ma, E. S. Chng, and H. Li, "Toward high-performance language-independent Query-by-Example spoken term detection for MediaEval 2015: Post-Evaluation analysis," in *Proc. of Interspeech*, 2016, pp. 3703–3707.

[21] H. Xu, J. Hou, X. Xiao, V. T. Pham, C.-C. Leung, L. Wang, V. H. Do, H. Lv, L. Xie, B. Ma, E. S. Chng, and H. Li, "Approximate search of audio queries by using DTW with phone time boundary and data augmentation," in *Proc. of ICASSP*, 2016, pp. 6030–6034.

[22] J. Tejedor and D. T. Toledano, *The ALBAYZIN 2018 Search on Speech Evaluation Plan*, Universidad San Pablo CEU, Universidad Autónoma de Madrid, Madrid, Spain, June 2018. [Online]. Available: http://iberspeech2018.talp.cat/index.php/albayzin-evaluation-challenges/search-on-speech-evaluation/

[23] P. Lopez-Otero, L. Docio-Fernandez, and C. Garcia-Mateo, "GTM-UVigo systems for albayzin 2016 search on speech evaluation," in *Proc. of IberSPEECH*, 2016, pp. 306–314.

[24] P. Schwarz, "Phoneme recognition based on long temporal context," Ph.D. dissertation, FIT, BUT, Brno, Czech Republic, 2008.

[25] I. Szöke, M. Skacel, and L. Burget, "BUT QUESST 2014 system description," in *Proc. of MediaEval*, 2014, pp. 621–622.

[26] M. Muller, *Information Retrieval for Music and Motion*. New York: Springer-Verlag, 2007.