The ELiRF Query-by-Example STD systems for the Albayzin 2012 Search on Speech Evaluation

Emilio Sanchis, Lluís-F. Hurtado, Jon A. Gómez,
Marcos Calvo, and Raül Fabra

Departament de Sistemes Informàtics i Computació
Universitat Politècnica de València, València, Spain
{esanchis,lhurtado,jon,mcalvo}@dsic.upv.es
rafabbo@fiv.upv.es

Abstract. In this paper, we present two approaches to Query-by-Example Spoken Term Detection. In both systems a first phase that obtains posterior probabilities of phonemes is performed. One of the systems performs a Segmental DTW algorithm to obtain the best matchings between the query utterance and the speech data. The other system is based on a search algorithm that finds the best matchings between two graphs of phonemes that represent the query utterance and the speech data.

Keywords: Spoken Term Detection, Query-by-Example, Automatic Speech Recognition

1 Introduction

In recent years some approaches for mining in audio repositories have been developed. Classification, language identification, diarization, indexing or spoken term detection are some of these works. In particular, the Query-by-Example Spoken Term Detection is an interesting task, where it is not necessary to have any a-priori knowledge about the language corresponding to de audio. The search is based only in utterances of the words to search. Due to the problem of lacking the vocabulary of the task, many approaches are based on a phonetic representation of the speech or work directly on the parametrization of the signal [1–4].

In the work presented in this paper, we propose two approaches for the Query-by-Example Spoken Term Detection task. Both are based on a first phase where the posterior phonetic probabilities for each frame are obtained. This phonetic probabilities are computed from a previous process of acoustic clustering and classification in terms of acoustic classes. Once we have obtained the phonetic probabilities, in one of the systems we apply a Segmental Dynamic Time Warping algorithm to find the segments of the signal that best match the utterance. In the other system, we construct graphs of phonemes, with their corresponding probabilities, to represent the query utterance and the speech data. Then we have developed an algorithm to obtain the segments of speech that best match the graph representing the query utterance.

$$D(i,j) = \begin{cases} 0 & j < 1 \\ \min \begin{pmatrix} D(i-1,j-1) \\ D(i-2,j-1) \\ D(i-1,j-2) \end{pmatrix} + KL(A(i),B(j)) & j \geq 1 \end{cases}$$

where A(i) is the *a posteriori* probabilities of phonemes for the frame i of the speech data, B(j) is the *a posteriori* probabilities of phonemes for the frame j of the query, and KL is the Kullback–Leibler divergence described in the previous subsection.

4.3 Filtering the candidate detections

For each speech data A and each query B a SDTW matrix is obtained. The last row of this matrix ($D(i,|B|)$, $1 \leq i \leq |A|$) contains the scores of all the local alignment candidates. In order to filter this candidate list we have used the algorithm 1 as used in the ELiRF-PhGraph system, but considering the scores normalized by the length of their corresponding paths. The threshold used in algorithm 1 has been experimentally fixed to obtain the best AWVT score for the development set but ensuring at least 10% of coverage.

5 Conclusions

In this work, we have presented two approaches to Query-by-Example Spoken Term Detection. Both approaches are based on a search considering the *a posteriori* phonetic probabilities of both the query and the speech data. In one case the uncertainty is modeled by means of graphs of phonemes, while in the other a modified DTW is used to find the best matching.

Acknowledgments. This work is partially supported by the Spanish MICINN under contract TIN2011-28169-C05-01, by the Vic. d'Investigació of the UPV under contract PAID-06-10, and by the Spanish MICINN under FPU Grant AP2010-4193.

References

1. Anguera, X., Macrae, R., Oliver, N.: Partial sequence matching using an unbounded dynamic time warping algorithm. In: Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on, IEEE (2010) 3582–3585
2. Hazen, T., Shen, W., White, C.: Query-by-example spoken term detection using phonetic posteriorgram templates. In: Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on, IEEE (2009) 421–426
3. Zhang, Y., Glass, J.: Unsupervised spoken keyword spotting via segmental DTW on gaussian posteriorgrams. In: Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on, IEEE (2009) 398–403