Subspace Regularized Dynamic Time Warping for Spoken Query Detection

Dhananjay Ram[†,‡], Afsaneh Asaei[†], Hervé Bourlard[†,‡]
[†]Idiap Research Institute, Martigny, Switzerland
[‡]École Polytechnique Fédérale de Lausanne (EPFL), Switzerland
{dhananjay.ram,afsaneh.asaei,herve.bourlard}@idiap.ch

Dynamic time warping (DTW) is an algorithm to find out the similarity between two temporal sequences of varying length. Previous works in this field can be traced back to as early as [1], for automatic speech recognition (ASR). Although this technique became obsolete for ASR with the advent of Hidden Markov Models (HMM) [2] and Deep Neural Network (DNN) based hybrid models [3], [4], DTW was found to be highly effective for spoken query detection, which refers to the task of searching a spoken query within an audio document. The key distinction is, unlike HMM and DNN solutions that require a large amount of annotated data to train the models, DTW can operate in low-resource conditions when training data is scarce. Therefore, DTW based systems are the state-of-the-art solutions for spoken query detection using one or a few examples of the query.

Traditional DTW algorithm performs an end-to-end comparison between two temporal sequences. This is not exactly applicable to spoken query search because, the query can occur anywhere in the test audio as a sub-sequence. Therefore, variants of DTW such as segmental DTW [5] and sub-sequence DTW [6] are developed to address this limitation. In order to use these methods, phone posterior features are extracted [8] from the speech data as shown in Fig. 1. Now, given a query and an audio document, a distance matrix is computed between their phone posterior representations where each element of the matrix represents a frame-level distance. It is followed by a dynamic programming technique to find an optimal alignment between the frames of a query and a test audio.

Although the methods discussed above are able to consider the sequential information present in a spoken query, they do not take into account the low-dimensional subspace structure of speech. Previously, we have proposed a novel sparse subspace modeling approach for query detection that exploits this property of speech [7], [8] where, we construct two dictionaries for sparse representation characterizing the subspace of the query and background speech independently. The sparse recovery reconstruction error is used as the score for query detection. To incorporate the sequential information, adjacent frames were concatenated to perform a frame-level detection. However, this approach lacks a proper framework to exploit the temporal information inherent to spoken queries.

We observe that the two kinds of systems discussed above use complementary information present in speech to perform the same task. In order to take advantage of both systems, we propose a new DTW technique considering the subspace structure in speech. This method relies on the notion that a spoken query lies in a low-dimensional subspace which can be represented as a sparse linear combination of corresponding training data. The training examples of the query are used to construct a dictionary for sparse representation which models the query subspace. These dictionaries can be used to obtain a sparse representation of test audio frames which can be further utilized to calculate reconstruction error for each frame [9]. The error for a test frame can be considered as the distance between the query subspace and the corresponding frame.

We propose to use the subspace based distance to regularize the distance matrix for DTW. Each column of the distance matrix corresponds to the frame-level distance between a test frame and all frames of the query. Whereas, we have only one number representing the distance from a test frame to the query subspace as a whole. Thus, to regularize the distance matrix, we consider a column of it corresponding to a test frame and take a weighted average of each element in this column with the subspace based distance obtained using the same test frame. Now, we perform dynamic programming on this regularized distance matrix to obtain the region of occurrence of the query and calculate the likelihood of its occurrence. A comprehensive block diagram for the proposed system is presented in Fig. 2.

The key idea behind the proposed method is, the frame-level distance provides local similarity and helps to capture the temporal information inherent to speech whereas, subspace based distance captures the similarity on subspace-level which considers all the frames present in the query for each test frame. A combination of these two distances provide better likelihoods for making a decision as can be seen through performance improvement. In principle, our approach can work with any variant of DTW by regularizing the corresponding distance matrix. However, in this work, we implement the system presented in [10] and perform the proposed regularization over the distance matrix followed by dynamic programming to obtain the region of occurrence along with likelihood score. The system in [10] is based on segmental DTW and is one of the best systems available for this task. Thus, we use this system as baseline for comparison purposes.

We have performed spoken query detection experiment on AMI meeting corpus [11] to show the potential of our approach. There are approximately 12k words in the training, out of which 200 frequent words are used as queries for our detection experiments. Then, these queries are divided into 2 sets of 100 queries each, to have development and evaluation queries. The feature vectors of each query serve as the dictionary for sparse coding as well as the reference template for DTW. Different parameters of the system are optimized using development queries. The results on evaluation queries are shown using detection error trade-off (DET) curve in Fig. 3. It is clear from the plots presented in Fig. 3 that, our proposed system significantly outperforms the baseline system. These results further show that, the low-dimensional subspace structure of speech can be very useful for spoken query detection.

I. Acknowledgments

The research leading to these results has received funding from the Swiss NSF project on "Parsimonious Hierarchical Automatic Speech Recognition and Query Detection (PHASER-QUAD)" grant agreement numbers 200021-153507, 200020-169398.

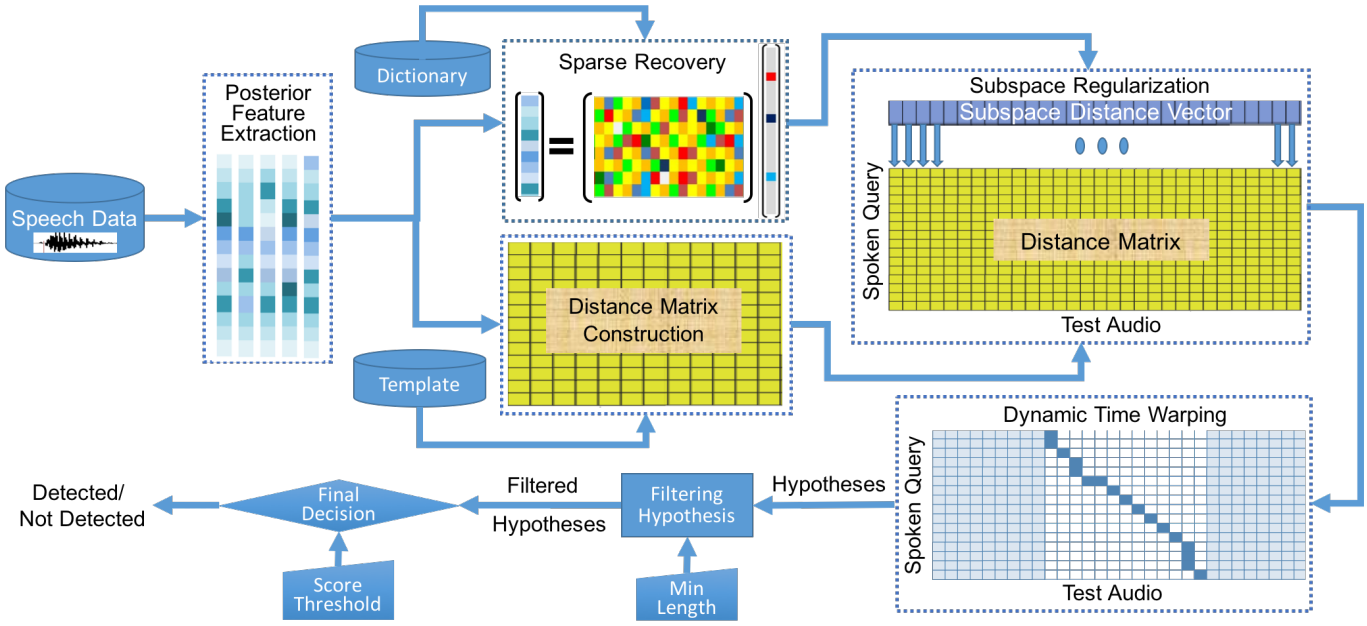

Fig. 2. Block diagram of the proposed system. The procedure is as follows: (1) Extract the posterior features from the test utterances, (2) Use the query dictionary (consisting of query posteriors) for sparse recovery and calculate the reconstruction error for each frame to generate the subspace based distance vector. (3) Calculate the distance matrix for DTW using the query posterior features as the template. (4) Regularize each column of the DTW distance matrix using the errors from the sparse reconstruction. (5) Apply DTW to detect the spoken query considering the hypotheses with more than half of the minimum length of the query to reduce false alarms. The detection score threshold is optimized on the cross-validation set.

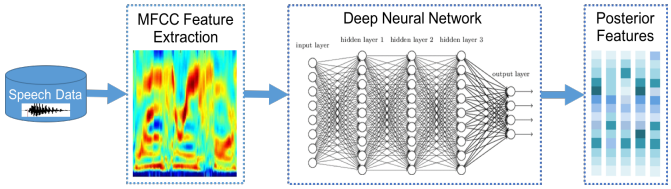

Fig. 1. Posterior feature extraction using a deep neural network: First, Mel Frequency Cepstral Coefficient (MFCC) based features are extracted over a sliding window of 25ms with a shift of 10ms. These features are then fed to a DNN to calculate phone conditional posterior probabilities.

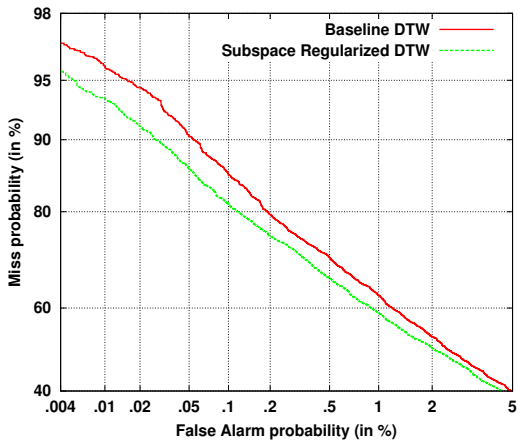

Fig. 3. DET curves for the proposed subspace regularized DTW and the baseline DTW system evaluated on the test set. Only a single example per query is used as the training data.

References

[1] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE transactions on acoustics, speech, and signal processing*, vol. 26, no. 1, pp. 43–49, 1978.

[2] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[3] H. Bourlard and N. Morgan, "Connectionist speech recognition: A hybrid approach," 1994.

[4] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 82–97, 2012.

[5] Y. Zhang and J. R. Glass, "Unsupervised spoken keyword spotting via segmental DTW on Gaussian posteriorgrams," in *IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)*, 2009, pp. 398–403.

[6] M. Müller, *Information retrieval for music and motion*. Springer, 2007, vol. 2.

[7] D. Ram, A. Asaei, P. Dighe, and H. Bourlard, "Sparse modeling of posterior exemplars for keyword detection," in *Sixteenth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2015.

[8] D. Ram, A. Asaei, and H. Bourlard, "Subspace detection of DNN posterior probabilities via sparse representation for query by example spoken term detection," in *Seventeenth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2016.

[9] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani *et al.*, "Least angle regression," *The Annals of statistics*, vol. 32, no. 2, pp. 407–499, 2004.

[10] 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 *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 7819–7823.

[11] I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos *et al.*, "The AMI meeting corpus," in *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*, vol. 88, 2005.