COMBINING TEMPORAL AND SPECTRAL INFORMATION FOR QUERY-BY-EXAMPLE SPOKEN TERM DETECTION

Ciro Gracia[1], *Xavier Anguera*[2], *Xavier Binefa*[1]

[1]Universitat Pompeu Fabra, Dept. of Information and Communications Technologies, Barcelona, Spain
[2]Telefonica Research, Edificio Telefonica-Diagonal 00, 08019, Barcelona, Spain
{ciro.gracia, xavier.binefa}@upf.edu, xanguera@tid.es

ABSTRACT

We present a system for Query-by-Example Spoken Term Detection on zero-resource languages. The system compares speech patterns by representing the signal using two different acoustic models, a Spectral Acoustic (SA) model covering the spectral characteristics of the signal, and a Temporal Acoustic (TA) model covering the temporal evolution of the speech signal. Given a query and a utterance to be compared, first we compute their posterior probabilities according to each of the two models, compute similarity matrices for each model and combine these into a single enhanced matrix. Subsequence-Dynamic Time Warping (S-DTW) algorithm is used to find optimal subsequence alignment paths on this final matrix. Our experiments on data from the 2013 Spoken Web Search (SWS) task at Mediaeval benchmark evaluation show that this approach provides state of the art results and significantly improves both the single model strategies and the standard metric baselines.

Index Terms— Query by example, zero resources languages, unsupervised learning, long temporal context

1. INTRODUCTION

The objective of the Query-by-Example Spoken Term Detection (QbE-STD) task is to search for spoken audio within a speech corpus by using a speech query. This task is gaining interest in the scientific community in the last years. Within the SWS task in the 2013 Mediaeval evaluation campaign [1] evaluation systems were given a set of over 500 development and 500 evaluation queries and a corpus of audio composed of around 20 hours of audio and 9 different languages recorded in different acoustic conditions. No information about the transcription of the queries or the speech corpus, nor the language spoken in each utterance was given to participants. In addition, given that none of the languages in the dataset has additional extensive resources available to train full speech recognition systems, this can be considered as a zero-resource QbE-STD task.

To tackle this task, different approaches have been used in the literature. Many of them [2, 3] make use of posteriorgram features in order to improve the enhace comparison between speech patterns. Posteriorgram features are obtained as the posterior probabilities of a an acoustic model evaluated on the input speech features and allow to consistently compare acoustic patterns by removing factors of feature variability other than the content being spoken. Similarly, some approaches [4] take advantage of well trained phonetic recognizers available in some languages, and additional automatic speech recognition systems to produce posteriorgram representations. These systems are trained using quality annotated datasets that provide solid acoustic models. Despite of this, the performance of the models degrades when applied to different and mismatching data and some sort of adaptation must be always applied. The difficulty at this point relies on how to obtain meaningful acoustic models that provide adequate posteriorgram features and how to properly compare them to find matching results. Regardless of how the posteriorgrams have been obtained, query and reference features can be compared through a similarity matrix were the query is searched inside the reference by using the S-DTW [5] algorithm.

In order to improve the matching accuracy, some approaches [6] perform a fusion of the similarity matrices obtained from different feature posteriorgrams. Despite of that, it is important to determine which types of information can better complement each other in order to guarantee a performance gain for the extra computational cost. Many studies support that temporal and spectral information are complementary and crucial for speech processing by the human auditory system [7]. The exploitation of temporal information for supervised acoustic models has been widely studied in [8]. The temporal evolution is modeled for each band of the acoustic features by extracting temporal vectors on fixed time intervals. The resulting vectors are modeled with respect to phonetic classes using a supervised classifier. The resulting posteriors are then used to train a parallel grammar phonetic recognizer using hidden Markov models.

In this paper, we present a system based on pattern matching and the fusion of different knowledge sources. Instead of fusing information from different languages, we choose to combine the speech representations of the signals obtained from temporal and spectral models in a semi-unsupervised manner. In order to improve the acoustic modeling with unsupervised data, our approach is to drift the unsupervised training towards meaningful information by introducing linguistic priors. We obtain those priors from an annotated data set that, although it has a strong mismatch in both language and acoustic recording conditions with respect to the experimental corpus, it incorporates a valuable seed for the training of the acoustic models. We believe that zero-resource languages may take profit from the available well studied languages, since they probably share a certain amount of acoustic characteristics [9].

In addition, instead of using the standard cosine similarity to compare posteriorgram features, we extend this approach by incorporating to the comparison a specially crafted matrix defining an inter-cluster dissimilarity.

In order to find matching sequences we use a memory efficient subsequence-dynamic time warping algorithm (s-DTW) similar to that in [10]. With it we obtain the alignment paths and the scores of all the potential matches of the queries inside the reference utterances. Finally, we explore two different approaches to global score normalization: the standard Z-norm approach and score mapping based on a continuous density function.

2. SYSTEM DESCRIPTION

Figure 1 summarizes the proposed system. Initially, standard MFCC39 feature representation is compute, with 25ms windows size and 10 ms shift time. We apply cepstral mean and variance normalization to both queries and utterances at file level. We then use the spectral-acoustic model and the temporal-acoustic model to convert the input features into posterior probability vectors, which are then combined into a single similarity matrix to allow for search of the query into each utterance. We use the s-DTW algorithm to determine optimal alignment paths within this matrix. We then filter and normalize the results to determine the final hypothesized hits for each query. Each of these steps is further described below.

2.1. Spectral Acoustic Model

The SA model is based on a Gaussian mixture model (GMM). GMM models trained from acoustic data with no supervision have been reported as a successful way to model broad acoustic classes [2]. Despite of that, data preparation and model initialization are tricky steps that condition the model and therefore the performance of the entire system. There is a lack of methods to assess the quality of the obtained model with respect the representation of acoustic classes in the data space, especially the ones which are meaningful for the task. Alternatively, adaptation approaches can provide ways to apply well trained supervised models to new data. We would therefore like to transform a GMM model trained in a supervised manner from out-of-domain, out-of-language data to fit the target data specific acoustic conditions. Although different unsupervised adaptation approaches exist [11,12] we perform

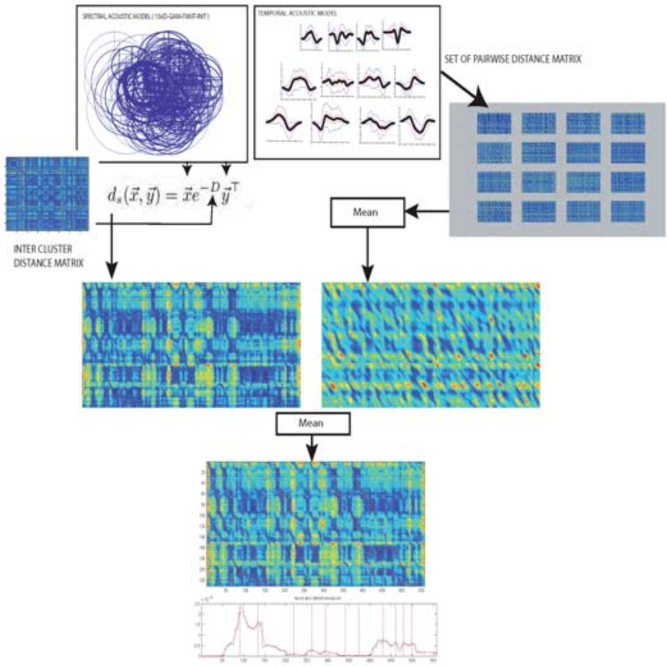

Fig. 1. Schematics of the system. Two acoustic models generate pairwise distances between query and reference. The matrices are fused into a single matrix where the alignment paths are searched and filtered.

here an unsupervised training of the target model by using a GMM model, trained in a supervised manner, as initialization.

We originally trained this supervised model using TIMIT phonetic ground truth. In particular, we trained a 4-Gaussian GMM for each of the 39 Lee and Hon [13] phonetic classes and then combined all of them into a single GMM model. This GMM is then used as initialization for an unsupervised training of the final 156-component model (39x4) using the Mediaeval 2013 database. The idea is to bias the unsupervised learning towards a phonetic-like structure and solve the problem of a proper initialization of the model. We assume that normalization on the data (CMVN) together with the dense GMM model structure will inhibit the unsupervised training from substantially modifying the original GMM structure.

2.1.1. *Comparison of posterior vectors*

Cosine similarity is generally used to compare posterior probability vectors [2]. Such measure has been shown to be slightly superior to other measures like the Kullback-Leibler (KL) divergence [14]. Assuming $s_x, s_y \in \Re^{156}$ being posterior probability vectors of the acoustic model for a given pair of acoustic vectors x, y, the cosine similarity is defined in Eq. 1. In addition to its geometric interpretation, it can be seen as the posterior probability of x and y to belong to the same cluster.

$$Cossim(s_x, s_y) = \frac{s_x s_y^\top}{\|s_x\| \, \|s_y\|} \tag{1}$$

In addition, we want our metric to take into account the similarity between posterior vectors and also penalize for the dissimilarities from the underlying acoustic classes. In consequence, we include a distance matrix into the similarity formulation. The distance matrix we use is defined as

$$Weightsim(s_x, s_y) = s_x e^{-D} s_y^{\top} \tag{2}$$

where $D \in M_{156x156}[\Re]$ is the KL divergence between each pair of Gaussian components in the acoustic model. Given a pair of Gaussian components (i, j), let μ_i, μ_j be the mean vectors and Σ_i, Σ_j the covariance matrices, then the KL divergence is : Eq. 3.

$$\begin{aligned} D(i,j) = \frac{1}{2}(log(\frac{|\Sigma_i|}{|\Sigma_j|}) + tr(\Sigma_i\Sigma_j + \Sigma_j\Sigma_i - 2I) \\ +(\mu_i - \mu_j)(\Sigma_i + \Sigma_j)(\mu_i - \mu_j)^{\top}) \end{aligned} \tag{3}$$

2.2. Temporal Acoustic Model

The objective of the temporal acoustic model is to use the information on the dynamics of the signal with a longer time span than the standard MFCC features can provide, therefore becoming a good complement to the spectral acoustic model. The temporal acoustic model is based on a long temporal context approach [8]. We build an independent temporal model for each of the 39 dimensions in the MFCC vector. The choice of using MFCC domain features for this model is motivated by the fact that the individual dimensions are mostly decorrelated and thus can be modeled independently. Initially, we segment the training data by using an unsupervised phonetic segmentation approach [15] and extract a 150 ms context from the center of each of the segments forming a collection of $\Re^{31}$ vectors per MFCC dimension. A postprocessing is performed on these vectors to avoid unnecessary overlap and to select only relevant context samples. Each context vector is then standardized to zero mean and unity variance, windowed using a Hanning window, and decorrelated using discrete cosine transform to choose the first 15 coefficients as the final $\Re^{15}$ vector. The modeling of each dimension is initialized by hierarchical k-medioids algorithm followed by an Expectation Maximization (EM) iteration to estimate the covariance matrices. The resulting model is composed of a Gaussian Mixture model of 128 components for each of the original 39 dimensions. We trained the temporal model using Mediaeval 2012 evaluation campaign database [16], which we used as development data and because the 2013 data was still not available at the time.

The comparison between two input vectors x, y is done in each of the b dimensions independently using the posteriors $p_x^b, p_y^b \in \Re^{128}$ obtained by the band temporal model. Then, the results from each band re fused by using Eq. 4.

$$\begin{aligned} d_t(x,y,b) &= \frac{p_x^b p_y^{b\top}}{\|p_x^b\| \, \|p_y^b\|} \\ d_t(x,y) &= \frac{1}{B}\sum_{b=1}^{B} -log(d_t(x,y,b)); \end{aligned} \tag{4}$$

2.3. Query Search

For each query $Q = \{q_1 \dots q_N\}$ and utterance $U = \{u_1 \dots u_K\}$ pair, we build a distance matrix $M \in M_{NxK}[\Re_{\geq 0}]$ by combining the similarity matrices from the SA and TA models as:

$$M(q_i, u_j) = -log(d_s(s_i, s_j)) + d_t(q_i, u_j); \tag{5}$$

We then use S-DTW to obtain the optimal alignment paths between every Q and U. S-DTW is a straightforward variation of the well-known DTW algorithm where no penalty is introduced by insertions in the utterances either at the beginning or at the end of the query, therefore allowing us to find the best matching subsequence in the utterance matching the entire query. When implementing the matching algorithm we incorporate a penalty term to each of the possible alignment steps in the S-DTW equation. We define the local constraints for S-DTW as shown in Eq. 6 where C is the resulting accumulated cost matrix and $P = \{P_1, P_2, P_3\}$ is a vector of positive penalties. We experimentally found $P = -log(\left[0.6, 0.6, 0.8\right])$ to be optimal. The penalties work together with the temporal model to avoid the presence of heavily warped paths.

$$C(i,j) = M(i,j) + min(\left\{ \begin{array}{c} C(i-1,j) + P_1 \\ C(i,j-1) + P_2 \\ C(i-1,j-1) + P_3 \end{array} \right\}) \tag{6}$$

The search alignment sequences between query and utterance returns a set of possible paths. The major difficulty at this point relies on how to decide which ones of the found alignments are acceptable as potential query-utterance hits and how to deal with intra-to-inter query results overlap. In order to select relevant local maxima scoring paths, we first lowpass filter the accumulated scores $M(q_N, u_i)|\forall i \epsilon 1 \dots K$ by using a 25 frames Gaussian window. Nonetheless, the resulting selected alignment paths retain their original score values. We solve intra-query overlap by selecting the best scoring path, but its difficult to solve the overlap between the detection of different queries at utterance level without priors about their score distributions. At this point we perform exclusively a global normalization and independent filtering of the query results, leaving the inter-query overlap problem for future work.

2.4. Global normalization

When all utterances have been processed for a given query, we perform a global normalization of all possible matches. This normalization step is critical when querying the database with multiple queries because we would otherwise have to set up a query-independent thresholds to separate between false alarm and true detections. The queries have different acoustic characteristics and the score distribution of their search results are also different. In order to align those distributions we initially used a standard Z-normalization approach. For this, we first excluded the top best 500 results from the parameter estimation to avoid true matches from biasing the normalization. Alternatively we have also explored a different approach for normalizing scores. Similarly to contrast enhancing performed by histogram equalization in image processing [17], our approach replaces each resulting query scores with their corresponding value within the query probability continuous density function (cdf) constructed from all scores for that query. This effectively maps the scores distribution into a uniform distribution and the cdf becomes a linear function.

3. EXPERIMENTS AND RESULTS

We have used three databases in our experimental setup. The phonetic model used to initialize unsupervised training of the SA model has been build using the 4620 utterances in the TIMIT training corpus [18]. The subsequent unsupervised training has been performed using the development set of the Mediaeval 2013 database. In particular, we used the search utterances and the development query set to represent the acoustic space. In addition, the TA model was trained using the database from Mediaeval 2012 [16], considered as development dataset in the 2013 evaluation. This corpus consists of 1580 utterances plus 100 queries collected from 4 different African languages. Finally, the evaluation of the system has been conducted on the development and test sets of the Mediaeval 2013 database.

The QqE-STD task requires the systems to perform language independent audio search. Given an audio query, systems should be able to locate the appropriate files and the location of the query term within the audio files. We evaluate the system performance using Term Weighted Value (TWV) as proposed by NIST in [19]. In this paper we show the maximum term weighted value (MTWV) and the actual term weighted value (ATWV), which are the primary metrics of the SWS-2013 evaluation.

Initially we evaluate the gain obtained by each of our processing steps on the development dataset. The Baseline system uses only the SA model and the standard cosine similarity measure. Alternatively, Baseline + DMatrix takes into account the inter-cluster matrix for computing the cosine similarity. Finally, the last system evaluated also takes into account the TA model. Table 1 shows how both additions help to increase the resulting performance, being remarkable the gain obtained by taking into account both acoustic models.

With respect to the normalization step, Table 2 shows the complete set of results obtained by our best system using different normalization functions. We can see how CDF equalization obtains better results than the Z-normalization. In addition, we include the results from other two campaing systems. In one hand, GTTS System [20] is the top scoring system of the evaluation but is based on fusion of supervised phonetizers. On the other hand, we include the top scoring purely unsupervised system: CUHK [21].

Finally, Figure 2 shows the DET curves for the systems shown in Tables 1 and 2. The curves are deliberately short because the system is defined to limit the number of hypothesized hits in order to reduce highly penalizing false alarms in the output. When comparing the figures it is important to note the big improvement of combining the TA model with the SA model. Also important is the change of tilt resulting from applying CDF normalization. While in the lower false-alarm region the Z-norm seems to perform better, when false-alarms increase, and in overall, the DCF normalization achieves best results.

System	CDF norm.	Z norm.
Baseline (SA model)	0.1699-0.1688	0.1606-0.1593
B.L. + DMatrix	0.1868-0.1865	0.1734-0.1734
B.L. + DMatrix + TA	0.2878-0.2863	0.2693-0.2690
CUHK System [21]	0.367-0.367	0.306-0.304
GTTS System [20]	0.4174-0.4078	0.3992-0.3806

Table 1. MTWV-ATWV scores on the development set for the different systems, using both proposed normalizations

Normalization	Dev. set	Eval. set
CDF equalization	0.2878-0.2863	0.2688-0.2673
Z-normalization	0.2693-0.2690	0.2561-0.2520

Table 2. Complete system results: MTWV-ATWV for the proposed system (SA model + DMatrix + TA Model) using both proposed normalization schemes

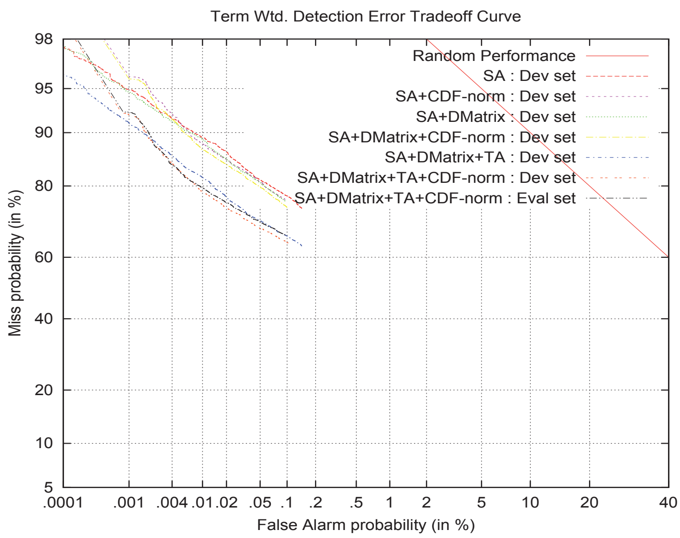

Fig. 2. DET plots for the presented systems

4. CONCLUSIONS

In this paper, we have presented a system for query-by-example spoken-term detection on zero-resource languages that uses information obtained from the spectral configuration of the signal, as well as information about the temporal evolution of the acoustic features. The fusion of both knowledge sources improves significantly the performance of the baseline system. In addition, we have extended the standard measure for comparing posterior features by incorporating a distance matrix into the dissimilarity computation, obtaining an additional extra performance boost of about 9% percent relative over the baseline approach. Finally, we have presented a different method for score normalization of the resulting hits for each query. Overall, the proposed system improves in 69% relative with regard to the considered baseline approach and achieved very competitive results within the Mediaeval SWS 2013 evaluation.

REFERENCES

[1] Xavier Anguera, Florian Metze, Andi Buzo, Igor Szoke, and Luis Javier Rodriguez-Fuentes, "The spoken web search task," in *MediaEval 2013 Workshop*, Barcelona, Spain, October 18-19 2013.

[2] Yaodong Zhang and James R Glass, "Unsupervised spoken keyword spotting via segmental dtw on gaussian posteriorgrams," in *Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on*. IEEE, 2009, pp. 398–403.

[3] Timothy J Hazen, Wade Shen, and Christopher White, "Query-by-example spoken term detection using phonetic posteriorgram templates," in *Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on*. IEEE, 2009, pp. 421–426.

[4] Amparo Varona, Mikel Peñagarikano, Luis Javier Rodríguez-Fuentes, Germán Bordel, and Mireia Díez, "Gtts system for the spoken web search task at mediaeval 2012.," in *MediaEval*, 2012.

[5] Meinard Müller, "Dynamic time warping," *Information Retrieval for Music and Motion*, pp. 69–84, 2007.

[6] Haipeng Wang, Tan Lee, Cheung-Chi Leung, Bin Ma, and Haizhou Li, "Using parallel tokenizers with dtw matrix combination for low-resource spoken term detection," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 8545–8549.

[7] Li Xu and Yunfang Zheng, "Spectral and temporal cues for phoneme recognition in noise," *The Journal of the Acoustical Society of America*, vol. 122, pp. 1758, 2007.

[8] Petr Schwarz, *Phoneme recognition based on long temporal context*, Ph.D. Thesis, 2008.

[9] Yu Qiao, Nobuaki Minematsu, and Keikichi Hirose, "On invariant structural representation for speech recognition: theoretical validation and experimental improvement.," in *INTERSPEECH*, 2009, pp. 3055–3058.

[10] Xavier Anguera and Miquel Ferrarons, "Memory efficient subsequence dtw for query-by-example spoken term detection," in *Multimedia and Expo (ICME), 2013 IEEE International Conference on*. IEEE, 2013.

[11] Peder A Olsen and Ramesh A Gopinath, "Extended mllt for gaussian mixture models," *Transactions in Speech and Audio Processing*, 2001.

[12] J-L Gauvain and Chin-Hui Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," *Speech and audio processing, ieee transactions on*, vol. 2, no. 2, pp. 291–298, 1994.

[13] Carla Lopes and Fernando Perdigão, "Broad phonetic class definition driven by phone confusions," *EURASIP Journal on Advances in Signal Processing*, vol. 2012, no. 1, pp. 1–12, 2012.

[14] Afsaneh Asaei, Benjamin Picart, and Hervé Bourlard, "Analysis of phone posterior feature space exploiting class-specific sparsity and mlp-based similarity measure," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE, 2010, pp. 4886–4889.

[15] Ciro Gracia and Xavier Binefa, "On hierarchical clustering for speech phonetic segmentation," 2011.

[16] Florian Metze, Etienne Barnard, Marelie Davel, Charl Van Heerden, Xavier Anguera, Guillaume Gravier, Nitendra Rajput, et al., "The spoken web search task," in *Working Notes Proceedings of the MediaEval 2012 Workshop*, 2012.

[17] Tinku Acharya and Ajoy K Ray, *Image processing: principles and applications*, Wiley. com, 2005.

[18] John S Garofolo, Lori F Lamel, William M Fisher, Jonathon G Fiscus, and David S Pallett, "Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1," *NASA STI/Recon Technical Report N*, vol. 93, pp. 27403, 1993.

[19] Jonathan G Fiscus, Jerome Ajot, John S Garofolo, and George Doddingtion, "Results of the 2006 spoken term detection evaluation," in *Proceedings of ACM SIGIR Workshop on Searching Spontaneous Conversational*. Citeseer, 2007, pp. 51–55.

[20] Luis J Rodriguez-Fuentes, Amparo Varona, Mikel Penagarikano, Germán Bordel, and Mireia Diez, "Gtts systems for the sws task at mediaeval 2013," in *MediaEval 2013 Workshop, Barcelona, Spain*, 2013.

[21] Haipeng Wang and Tan Lee, "The cuhk spoken web search system for mediaeval 2013," .