The CMTECH Spoken Web Search System for MediaEval 2013

Ciro Gracia
University Pompeu fabra
Barcelona, Spain
ciro.gracia@upf.edu

Xavier Anguera
Telefonica Research
Barcelona, Spain
xanguera@tid.es

Xavier Binefa
University Pompeu fabra
Barcelona, Spain
xavier.binefa@upf.edu

ABSTRACT

We present a system for query by example on zero-resources languages. The system compares speech patterns by fusing the contributions of two acoustic models to cover both their spectral characteristics and their temporal evolution. The spectral model uses standard Gaussian mixtures to model classical MFCC features. We introduce phonetic priors in order to bias the unsupervised training of the model. In addition, we extend the standard similarity metric used comparing vector posteriors by incorporating inter cluster distances. To model temporal evolution patterns we use long temporal context models. We combine the information obtained by both models when computing the similarity matrix to allow subsequence-DTW algorithm to find optimal subsequece alignment paths between query and reference data. Resulting alignment paths are locally filtered and globally normalized. Our experiments on Mediaeval data shows that this approach provides state of the art results and significantly improves the single model and the standard metric baseline.

1. INTRODUCTION

The task of searching for speech queries within a speech corpus without a priori knowledge of the language or acoustic conditions of the data is gaining interest in the scientific community. Within the Spoken Web Search task (SWS) in the Mediaeval evaluation campaign for 2013 [3] systems are given a set of acoustic queries that have to be searched for within a corpus of audio composed of several languages and different recording conditions. No information about the transcription of the queries or speech corpus, nor the language spoken is given.

To tackle this task we propose a system using a zero resources approach by extending some ideas from the state of the art.

We adopt posteriorgram features[9, 5] in order to improve comparison between speech features. Posteriorgram features are obtained from an acoustic model and allow to consistenly compare acoustic vectors by removing factors of feature variance. The difficulty at this point relies into how to obtain meaningful acoustic models in an unsupervised manner and how to properly compare posterior features. The difficulty at this point rely into how to obtain meaningful acoustic models unsupervisedly and how to properly compare posterior features. In order to obtain meaninfull acoustic models with unsupervised data we introduce linguistic prior information to the unsupervised training by using an specific pre-trained model as initialization. In addition, instead of use standard dot product to compare normalized posteriorgram vectors, we extend this approach by incorporating to the comparison a specially crafted matrix defining an inter-cluster similarity.

Previous approaches [8] to Mediaeval data have shown that using different acoustic models to fuse different sources of knowledge provides a significant improvement on evaluation. Despite of that, it is important to determine which types of information can complement each other in order to guarantee a gain for the extra computational cost. Our appoach to fusion is to combine temporal and spectral information. As stated above, one of the models is focused into spectral configuration of the acoustic vectors while the complementary model is focused into model temporal evolution of the feature dimensions.

For sequences matching we use the subsequence-dynamic time warping algorithm (s-DTW) [7]. With it we obtain the alignment paths and the scores of all the potential matches of the query inside the utterance. The major difficulty relies in how to decide which ones of the provided alignments are acceptable as potential query instances and how to deal with intra-inter query results overlap. In our system we used lowpass filtering to reduce the number of spurious detections and keept only the highest score of the intra query overlapping paths. Inter-query overlap is complex and remains for future work. Finally, We explore two different approaches to global score normalization: the standard Z-norm approach and score mapping based on continuous density function.

2. THE CMTECH SYSTEM DESCRIPTION

The system is based on standard MFCC39 features computed by means of HTK at (25ms windows , 10 ms shift time).

2.1 Spectral Acoustic Model

The first acoustic model based on a gaussian mixture model (GMM). We originally trained this model using TIMIT phonetic ground truth. We trained a 4 gaussians GMM for each of the 39 Lee and Hon [6] phonetic classes and then combined all of them into a single GMM. This GMM is used as initialization for an unsupervised training of the final 156 components GMM using SWS2013 utterances.

Using this model we build an inter-cluster distance matrix D (156x156) using Kullback Leibler divergence:

MediaEval 2013 Workshop, October 18-19, 2013, Barcelona, Spain

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

When comparing posterior features $\vec{x}, \vec{y}$ we use:

$$d_s(\vec{x}, \vec{y}) = \vec{x}e^{-D}\vec{y}^\top \quad (2)$$

We found this extended comparison providing above 0.05 absolute MTWV points gain in mediaeval 2012 data.

2.2 Temporal Acoustic Model

The objective of this temporal model is to extend the context information and to effectively complement the frame based acoustic model. The temporal model is based on long temporal context approach [1] trained on Mediaeval 2012 data. We process each of the MFCC39 dimensions independently. We first segmented Mediaeval 2012 data using an unsupervised phonetic segmentation approach[4] and extraced a 150 ms context from the center of each of the segments forming a collection of $\mathbb{R}^{31}$ vector. Each context vector is standarized to zero mean and unity variance, windowed using a Hanning window, decorrelated using discrete cosinus transform and only the 15 first coefficients become the final $\mathbb{R}^{15}$ vector. The modeling is performed by hierarchical k-medioid together with a final covariance matrices estimation. The resulting model is composed of a Gaussian Mixture model of 128 components for each of the original 39 dimensions.

The comparison between two input vector is done in each band b indepently by means of its model posterior $\vec{x}_b, \vec{y}_b$, and then we fuse them using the median operator:

$$d_t(\vec{x}, \vec{y}, b) = \frac{\vec{x}_b\vec{y}_b^\top}{\|\vec{x}_b\| \, \|\vec{y}_b\|}$$
$$d_t(\vec{x}, \vec{y}) = median(d_t(\vec{x}, \vec{y}, b)); \quad (3)$$

Inside Mediaeval 2012 data, the incorporation of this acoustic model boosted our system MTWV results from 0.47 to 0.53 points.

2.3 Query Search

For each pair of Query q and utterance u patterns we build a distance matrix M of size ($|q|x|u|$) using:

$$M(q,u) = -log(d_t(q,u)d_s(q,u)) \quad (4)$$

We use S-DTW to obtain the score of alignment paths for each possible ending position in u. In order to select relevant local maxima scores, we first lowpass filter the results by using a 25 frames gaussian window. Depite that the resulting selected alignment paths retain their original score values.

2.4 Global normalization

When all utterances have been processed for a given query, we perform a normalization step. The first system presented (primary) uses a standard Z-normalization excluding the first 500 results from the parametter estimation. Similarly to contrast enhacing performed by histogram equalization in image processing[2], our mapping approach replaces resulting query scores with their corresponing value at the query probability continuous density function (cdf). This effectively maps the scores distribution into a uniform distribution and their cdf as a linear function. Our second system (contrastive) replaces global Z-normalization by the cdf equalization aproach.

Table 1: System results: MTWV/ATWV

Normalization	Dev-Dev	Dev-Eval
CDF equalization	0.2685-0.2683	0.2623-0.2619
Z-normalization	0.2642-0.2638	0.2575-0.2552

3. RESULTS

Table 1 shows the results obtained by our systems. We can see how CDF equalization system obtains slightly better results than the Z-normalization system. The Runtime Factor is 0.0056 and the average memory usage is 11,5GB. Many of the difficulties in the results come from a set of noisy and reververant examples. We feel that denoising algorithms like spectral substraction would be useful to improve models training and performance on these samples.

4. CONCLUSIONS

Our future work will be related to explore the relationship between system performance and voice activity detection. Face the inter query overlap problem its inherent open set classification problem. We are interested into distiguish which are the key elements that garantee the suitability of an acoustic model for the task, Specially interesting is explore rigid and elastic distribution matching methods like maximum likelihood linear transforms in order to be able to adapt pre-trained models to new data unsupervisedly.

5. REFERENCES

[1] P. Ace, P. Schwarz, and V. P. Ace. Phoneme recognition based on long temporal context.

[2] T. Acharya and A. K. Ray. *Image processing: principles and applications*. Wiley. com, 2005.

[3] X. Anguera, F. Metze, A. Buzo, I. Szoke, and L. J. Rodriguez-Fuentes. The spoken web search task. In *MediaEval 2013 Workshop*, Barcelona, Spain, October 18-19 2013.

[4] C. Gracia and X. Binefa. On hierarchical clustering for speech phonetic segmentation. 2011.

[5] T. J. Hazen, W. Shen, and C. White. Query-by-example spoken term detection using phonetic posteriorgram templates. In *Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on*, pages 421–426. IEEE, 2009.

[6] C. Lopes and F. Perdigão. Broad phonetic class definition driven by phone confusions. *EURASIP Journal on Advances in Signal Processing*, 2012(1):1–12, 2012.

[7] M. Müller. Dynamic time warping. *Information Retrieval for Music and Motion*, pages 69–84, 2007.

[8] H. Wang and T. Lee. Cuhk system for the spoken web search task at mediaeval 2012. In *MediaEval*, 2012.

[9] Y. Zhang and J. R. Glass. Unsupervised spoken keyword spotting via segmental dtw on gaussian posteriorgrams. In *Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on*, pages 398–403. IEEE, 2009.