TUKE at MediaEval 2013 Spoken Web Search Task

Jozef Vavrek, Matúš Pleva, Martin Lojka, Peter Viszlay,
Eva Kiktová, Daniel Hládek, Jozef Juhár
Technical University of Kosice, Letna 9, 04200 Košice, Slovakia
{Jozef.Vavrek, Matus.Pleva, Martin.Lojka, Peter.Viszlay,
Eva.Kiktova, Daniel.Hladek, Jozef.Juhar}@tuke.sk

ABSTRACT

This paper provides a rough description of zero resource Query-by-Example retrieving system for the MediaEval 2013 spoken web search task. The proposed solution firstly implements the voice activity detection (VAD) utilizing variance of acceleration MFCC (VAMFCC) rule-based approach. A PCA-based segmentation, K-means clustering and GMM training are then used in order to built the posteriorgrams. Finally, two searching architectures based on posteriorgram matching (SDTW) and GMM modeling (GMM-FST) are evaluated. Results show that none of our systems is able to achieve the positive Actual Term Weighted Value, because of high number of insertions. We suppose that chosen clustering scheme caused generation of too many false alarms. Only provided data were used and no other resources were examined in any system component during the development.

1. MOTIVATION

The main purpose of our experiments was to check the proposed approaches for the language independent audio query detection and new speech feature analysis components. The mentioned approach is used in the MediaEval activity [1] and could be also applied in various speech [4] or non-speech [5] Query-by-Example applications.

2. SYSTEM OVERVIEW

Proposed solution for SWS task uses posteriorgram term matching and audio segment GMM modeling. The overall architecture of proposed system is depicted on Fig.1.

At first, a training phase is carried out using available development utterances. The VAMFCC-based silence detector performs the initial discrimination of silent parts in audio stream. The block of feature extraction is implemented after VAD utilizing 13 MFCCs. The phase of segmentation and clustering creates the audio segment units (ASU). ASU is represented as a small audio part (phoneme for example) with some spectral and temporal characteristics, different for each ASU. Then the training of acoustic models is performed using these ASU, where each ASU represents one class. Labels for these classes (ASUs) are assigned according to the number of GMM. Only the process of voice activity detection and feature extraction is then performed in preprocessing stage within the retrieving phase. Each utterance and query enters the block of model-based scorer, where scoring is performed by computing the Mahalanobis distance between MFCC vectors (frames with length 25 ms and 10 ms shift) and GMM acoustic model. The product of that are posteriorgrams. They are defined as probability vectors with the length of N, where N represents the number of GMMs. Segmental Dynamic Time Warping (SDTW) algorithm is then used for comparing these posteriorgrams and finding a possible occurrence of a query in the test utterance. The other solution GMM-FST (GMM-Finite State Transducers) implements a Viterbi algorithm to create a model for each query and state sequence network to find occurrences in test utterance by using this model.

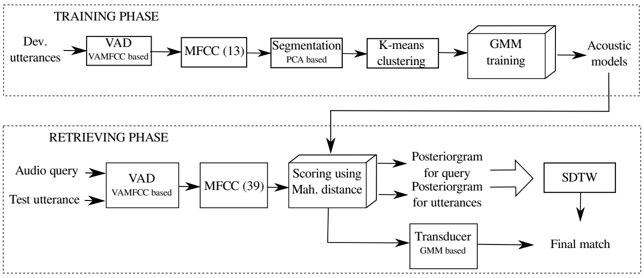

Figure 1: SWS framework architecture

2.1 Segmentation and Clustering

In order to identify and to distinguish the speech segments in the i-th utterance, PCA (Principal Component Analysis) was applied as follows. Each 13-dimensional MFCC vector $\mathbf{x}_j$ was reshaped to matrix X_j with row dimension n_r, where $j \in \langle 1; n_i \rangle$ is the number of vectors in i-th recording. In the next step, the covariance matrix C_1 was computed from the first matrix X_1 and its eigenvectors and eigenvalues were computed. The eigenvalue spectrum $\Lambda_1 = \{\lambda_{1j}\}_{j=1}^{n_r}$ was used to determine the significance $\Delta(\lambda_{1_{max}})$ of the dominant eigenvalue of C_1 as $\Delta(\lambda_{1_{max}}) = \frac{\lambda_{1_{max}}}{\sum_{j=1}^{n_r} \lambda_{1j}}$, where λ_{1j} are the eigenvalues of C_1. Then the matrix X_1 was spliced together with X_2 and the covariance matrix C_{12} and $\Delta(\lambda_{12_{max}})$ were computed again. If $\Delta(\lambda_{12_{max}})$ compared to $\Delta(\lambda_{1_{max}})$ changed significantly, a new speech segment was created and PCA started from the current frame. In the other way, if $\Delta(\lambda_{12_{max}})$ did not change significantly, the current matrix X_{12} was spliced together with X_3 and the process was repeated automatically until a new segment was indicated. The created segments corresponded to ASUs.

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

Table 1: Evaluation results of the tested algorithms

query	*system*	*ATWV*	C_{nxe}	C_{nxe}^{min}
dev	GMM-FST	-0.1371	0.980745	0.976363
eval	GMM-FST	-0.1372	0.98505	0.979883
dev	SDTW	-0.4176	0.998421	0.989649
eval	SDTW	-0.4252	0.998494	0.988404

In the next phase, the segments with similar acoustic and statistical properties were grouped together into several speech clusters using k-means clustering with $k = 50$ clusters and squared Euclidean distance metrics. As the input data for clustering the means of the segments were used. Each mean vector obtained an index (label) of the specific cluster. This label was assigned to the original feature vectors corresponding to the specific mean vector.

2.2 Searching techniques

2.2.1 GMM approach

A retrieving process uses Weighted Finite State Transducers (WFST) that allow us to find the most probable path (state sequence) in search network [2]. The search of a query consists of two steps. At first, query alone is recognized using search network, created from the trained acoustic model so that all GMM states are arranged in parallel. The result is a sequence of states that model the particular query. The process of recognition is done repeatedly with different insertion penalties in order to obtain multiple states sequences with different lengths. It helps to improve the model representation of retrieving query. The sequences are labeled and added to the previous search network in parallel. The second step involves the recognition of a test utterance using Viterbi algorithm. The final score for decision is computed as a difference between modeled likelihoods of query and utterance using *lambda* acoustic model where $score = (P(O_{occurence}|\lambda) - P(O_{query}|\lambda)) + \Theta$. Score is then shifted by predefined value Θ and then results with score below zero are removed.

2.2.2 SDTW detection

A simplified SDTW searching algorithm was utilized in our system, similar to that used in [3]. The adjustment window condition was set to $|(i_k - i_1) - (j_k - j_1)| \leq R$, where i_1 and j_1 are starting coordinates of warping path in each segment, i_k and j_k define the k-th coordinates and R represents the constraint parameter, set to $M/2$, where M is the length of query. The range of starting coordinates was conditioned by the constraint parameter and length of each utterance: $((2R+1)k+1, 1)$, where $0 \leq k \leq \left|\frac{N-1}{2R+1}\right|$. The process of finding the optimal local alignment between each utterance and query produces a set of local warp paths, equal to the number of diagonal regions. A score parameter was then set in the following form $score = \sqrt{\frac{2n}{N+M} \Big/ \frac{\sum_1^n warpDist}{n+1}}$,where n is the number of steps in local alignment, N is the length of utterance and M the length of query, $\sum_1^n warpDist$ represents a summation of components in each warping path, where components are computed from Bhattacharyya distance matrix.

Table 2: Processing resources measures

system	*ISF*	*SSF*	PMU_I	PMU_S	*PL*
GMM-FST	0.0054	0.0048	2GB	1.8GB	0.009
SDTW	0.0054	0.0046	2GB	2.2GB	0.01

3. EXPERIMENTAL RESULTS AND CONCLUSIONS

The official results for SWS task are listed in Tab. 1. Two metrics were used to asses the overall performance of GMM-FST and SDTW on *dev* and *eval* queries: the actual $ATWV$, normalized C_{nxe} and minimal cross-entropy C_{nxe}^{min}. The score normalization for both systems was performed only on development data. A minimum-cost alignment (MCA) for each segment was used as detection score at first level of search in case of SDTW. Final detection of retrieved query was then carried out utilizing score parameter defined in (2.2.2). A threshold for this parameter was set to 0.0819. A decision threshold for score parameter was set to 2.8 in case of GMM-FST based system, while $\Theta = 3$ (2.2.1). Both systems produced a huge amount of false alarms (FA) during the evaluation. Regarding the evaluation results, the GMM-FST system is more appropriate solution for SWS task, because of its lower tendency to detect spurious terms.

All the experiments were mainly done using 2x IBM System x3650 servers, 2x Intel® Xeon® QuadCore E5530 CPU @ 2.4 GHz Hyper-threading enabled (16 threads), 28 GB RAM, 1TB SAS HDD (RAID5), running Debian OS.

The Speed Factors (the ratio of the total time employed in searching{indexing} the set of queries in{and} the set of audio documents to the product{sum} of their total durations) and Peak Memory Usage during Searching{Indexing} tasks are presented in Tab. 2. The Performance Load equals $0.9 \cdot SSF \cdot PMU_S + 0.1 \cdot ISF \cdot PMU_I$ is derived from them.

In the future, an improved clustering and segmentation algorithm will be investigated in order to decrease the overlapping between individual ASUs. A minimal length of warping path algorithm will be integrated in SDTW approach, too.

4. ACKNOWLEDGMENTS

This research was supported by the ERDF funded projects ITMS-26220220155 (50%) & ITMS-26220220182 (50%).

5. REFERENCES

[1] X. Anguera, F. Metze, A. Buzo, I. Szoke, and L. J. Rodriguez-Fuentes. The Spoken Web Search Task. In *MediaEval 2013 Workshop*, page 4, Barcelona, Spain, 18-19 October 2013.

[2] M. Lojka and J. Juhár. Finite-state transducers and speech recognition in Slovak language. In *SPA 2009 Conference*, pages 149–153, art. no. 5941305, 2009.

[3] A. Park and J. Glass. Unsupervised pattern discovery in speech. *IEEE T Audio Speech*, 16(1):186–197, 2008.

[4] J. Vavrek, M. Pleva, and J. Juhár. TUKE MediaEval 2012: Spoken Web search using DTW and unsupervised SVM. In *MediaEval 2012 Workshop, Pisa - CEUR Workshop Proceedings*, volume 927, 2012.

[5] E. Vozáriková, M. Pleva, S. Ondáš, J. Vavrek, J. Juhár, and A. Čižmár. Detection and classification of audio events in noisy environment. *Journal of Computer Science and Control Systems*, 3(1):253–258, 2010.