LIA @ MediaEval 2013 Spoken Web Search Task: An I-Vector based Approach

Mohamed Bouallegue, Grégory Senay,
Mohamed Morchid, Driss Matrouf, Georges
Linarès and Richard Dufour
LIA - University of Avignon (France)
{firstname.lastname}@univ-avignon.fr

ABSTRACT

In this paper, we describe the LIA system proposed for the MediaEval 2013 *Spoken Web Search* task. This multi-language task involves searching for an audio content query, in a database, with no training resources available. The participants must then find locations of each given query term within a large database of untranscribed audio files. For this task, we propose to build a language-independent audio search system using an i-vector based approach [2].

1. INTRODUCTION

The *Spoken Web Search* (SWS) task is characterized by two major difficulties. Firstly, the reference set is composed of audio files coming from different languages, accents and acoustic conditions. Secondly, no transcription or language resources are provided. Systems should then be built as generic as possible to succeed in finding queries appearing in these multiple condition sources.

In this work, a language-independent audio search system based on an i-vector approach is proposed. Inspired by the success of i-vectors in speaker recognition [2], we apply the same idea for this audio search task. To identify the locations of each query term within the audio files, our idea is to model each file and each query by a set of i-vectors and then align them.

2. PROPOSED APPROACH

Initially introduced for speaker recognition, i-vectors [2] have become very popular in the field of speech processing and recent publications show that they are also reliable for language recognition [5] and speaker diarization [3]. I-vectors are an elegant way of reducing the large-dimensional input data to a small-dimensional feature vector while retaining most of the relevant information. The technique was originally inspired by the Joint Factor Analysis framework [4]. Hence, i-vectors convey the speaker characteristics among other information such as transmission channel, acoustic environment or phonetic content of the speech segment.

2.1 I-vector extraction

The i-vector extraction could be seen as a probabilistic compression process that reduces the dimensionality of speech super-vectors according to a linear-Gaussian model. The speech (of given speech recording) super-vector m_s of concatenated Gaussian Mixture Model (GMM) means is projected in a low dimensionality space, named Total Variability space:

$$m_s = m + Tx_s \tag{1}$$

where m is the mean super-vector of *Universal Background Model* (UBM)[1]. T is a low rank matrix ($MD \times R$), where M is the number of Gaussians in the UBM and D is the cepstral feature size (39 in our case), which represents a basis of the reduced total variability space. T is named *Total Variability matrix*; the components of x_s are the total factors and they represent the coordinates of the speech recording in the reduced total variability space.

The proposed approach uses i-vectors to model speech segments. These short segments are considered as a language basic unit. Indeed, each file and each query are segmented in short segments of 20 frames. In our model, the segment super-vector $m_{(seg)}$ is modeled as follows:

$$m_{seg} = m + Tx_{(seg)} \tag{2}$$

2.2 System overview

In this section, the different steps used to build the proposed language-independent audio search system are detailed.

Step 0: Parametrization

In this first step, the MFCCs (39-dimensional feature vectors) are computed for all the database audio files. Each vector represents 30 ms of Hamming windowed speech signal (the window is shifted every 10 ms).

Step 1: Segmentation of database files

This step consists of segmenting the database files in short segments of 20 frames. In this step, a sliding window of 200 ms with an offset of 100 ms is used in order to avoid information lost. 713,315 short segments are obtained after the segmentation of the 10,762 audio files. The same procedure is applied on the sets of queries (development and evaluation). The 505 evaluation queries are segmented in

[1]The UBM is a GMM that represents all the possible observations. It is sometimes also called the world model.

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

6,602 short segments (nearly the same for the development queries).

Step 2: Estimation of the matrix T and the i-vectors of database files

The i-vectors of each of the 713,315 segments of the database files (see step 1) are estimated based on the equation 2. An i-vector x_{seg} of size 20 is then obtained for each segment. The UBM used (512 Gaussians) is estimated on all the database files. The *Total Variability matrix* T (dimension of $19,968 \times 20$) is estimated using all segments of audio files.

Step 3: Estimation of the i-vectors of queries

In this step, the i-vectors for the queries (development and evaluation) are estimated using the equation 2. We used the same UBM and the *Total Variability matrix* T obtained in the last step. Finally, 6,602 i-vectors (size 20) for the evaluation queries are obtained (nearly the same for the development queries).

Step 4: Alignment

In order to identify the locations of each query term within the database audio files, an alignment is performed between the i-vectors of each query and the i-vectors of all database audio files using an adapted *Dynamic Time Warping* (DTW) algorithm.

The dissimilarity between two i-vectors is computed with the Mahalanobis distance[2]. The matrix used in the Mahalanobis distance is the total covariance matrix estimated on all the i-vectors of the database audio files.

In order to find the query start, alignment can start at any point of the audio file but have to last at most two times the size of the query. The start and end times of a matching query are given by the start time of the first i-vector and the end time of the last i-vector of the alignment, respectively. The query score is the cost of the best path (more precisely, minus the cost). For each database file, the system searches the best costs of all queries although certain files do not contain any query. Only the n-best alignments for a document are kept (2, 3, 4, 6 or 8).

3. EXPERIMENTS

The proposed system is evaluated in the MediaEval 2013 SWS benchmark [1]. The number of queries is 500 and the number of audio files (dataset) is 10,000 (around 20 hours). The sets of audio files include many languages: non-native English, Albanian, Czech, Basque, Romanian and 4 African languages. The main metric evaluation used is the Actual Term Weighted Value (ATWV) [1]. Table 1 presents the results obtained on the development and the evaluation data in terms of ATWV and Cnxe scores. In the *Primary* system, the number of occurrences of each query in the audio files has been fixed to 2 (*i.e.* each query is detected twice in the database). In the four *contrastive* systems, we increased the number of occurrences to 3, 4, 6 and 8. Best results are obtained with the primary system whether it be on the development or the evaluation data. While the system applied on the development data reached a better performance than the baseline system provided by the organizers, these results are surprisingly low on the evaluation data.

[2]http://classifion.sicyon.com/References/M_distance.pdf

Table 1: Results obtained on the dev and the evaluation data in terms of ATWV and Cnxe scores.

	Dev		Evaluation	
	ATWV	**Cnxe**	**ATWV**	**Cnxe**
Primary (2)	0.0045	7.22683	-0.0013	93.2051
Contrastive 1 (3)	0.0040	7.23728	-0.0021	93.2154
Contrastive 2 (4)	0.0040	7.24818	-0.0029	93.2255
Contrastive 3 (6)	0.0029	7.26928	-0.0043	93.2461
Contrastive 4 (8)	0.0014	7.29037	-0.0055	93.2673

The indexing and the searching modules have been performed on a 48-core cluster (Intel Xeon processor 2.6 GHz). The memory peak reached 1.2 GB. The real-time ratio of the searching module have been computed:
Real-time ratio of the searching module: $((step\ 1+3+4)/(total\ duration\ of\ audio\ files * total\ duration\ of\ queries)) = 6,600/(71,839 * 696) =$ **0.000132**

4. CONCLUSIONS

In this paper, we proposed a language-independent audio search system based on an i-vector approach. Although the results on the evaluation queries are poor, the encouraging results obtained on the development data show that the i-vectors are an interesting and original unsupervised way to search audio content using an audio content query. In the future, we plan to investigate in details the mismatch between the development and the evaluation data performance. We will also explore the use of a *Voice Activity Detection* (VAD): it could help to discard silence sections that are contaminating audio queries.

5. ACKNOWLEDGMENTS

This work was funded by the ContNomina project supported by the French National Research Agency (ANR) under contract ANR-12-BS02-0009.

6. REFERENCES

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

[2] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet. Front-end factor analysis for speaker verification. In *IEEE TASLP*, 2009.

[3] J. Franco-Pedroso, I. Lopez-Moreno, and D. T. Toledano. Atvs-uam system description for the audio segmentation and speaker diarization albayzin 2010 evaluation. In *SLTech Workshop*, 2010.

[4] P. Kenny. Joint factor analysis versus eigenchannes in speaker recognition. In *IEEE Transactions on AShLP*, 2007.

[5] D. Martinez, O. Plchot, L. Burget, O. Glembek, and P. Matejka. Language recognition in i-vectors space. In *Interspeech*, 2011.