UNIZA System for the Spoken Web Search Task at MediaEval2013

Roman Jarina, Michal Kuba, Róbert Gubka, Michal Chmulik, Martin Paralič
Audiolab, Department of Telecommunications and Multimedia,
University of Žilina, Univerzitná 1, 010 26 Žilina, Slovakia
{roman.jarina, michal.kuba, robert.gubka, michal.chmulik, martin.paralic}@fel.uniza.sk

ABSTRACT

In this paper, we present an approach to detect spoken keywords according to a given query, as part of the MediaEval benchmark. The proposed approach is based on a concept of modelling the speech query as a concatenation of language-independent quasi-phoneme models, which are derived by unsupervised clustering on various audio data. Since only an initial version of the system is presented, issues concerning further system improvements are also discussed.

1. INTRODUCTION

Details about UNIZA submission for the Spoken Web Search (SWS) task within MediaEval 2013 benchmark iniciative are described below. SWS requires to develop a language–independent audio search system so that, given an audio query, it should be able to find the same speech phrases in audio content [1]. Our proposed method is motivated by the generally accepted approach of keyword spotting that relies on concatenation of probabilistic models (usualy Hidden Markov Models) of speech units [2]. Such approach is implicitly based on the fact that a language structure of the speech is a priori known and hence acoustical models of speech units can be developed in advance by using labeled training speech data. But this is not the case of the SWS task where neither language information nor any transcription is provided [1]. Hence the objective is to find a similar but more generalized language-independent and low-resource approach to acoustical modelling of speech.

The developed system generates stochastic models of "elementary sounds" (ES) derivated from the provided speech data in various languages. These ESs are used as building blocks for speech modelling instead of conventional phoneme based models. We have recently adopted this approach to the task of generic sound modelling and retrieval [3] with promising results. The approach in [3] is built on the assumption that, in general, many types of generic sounds can be modelled as a sequences of ES units, which are picked from a sufficiently large (much greater than the number of speech units), though finite inventory. Due to the diverse nature of generic sounds, it is infeasible to create an acoustical form of the elementary units; instead the ESs can be defined only by their stochastic models.

2. PROPOSED APPROACH

System layout is depicted in Figure 1. Each utterance (query) is modelled by a HMM-based statistical model. Due to lack of information about language and linguistic structure of the utterances, obviously, models of phonemes (or any other linguistic units) could not be built in advance. Instead, we have proposed to build models of language-independent quasi phoneme-like ES units with aid of unsupervised clustering. Then, the HMM for each query is built up by concatenation of such ES models (or fillers). Searching the query over the database is performed by Viterbi decoding [2]. The decoder generates output cumulative probabilities (confidence) and starting positions, from which the cumulative probabilities were computed. Candidates for search results are obtained by thresholding the confidence curve. We used only the provided SWS2013 development data during system development.

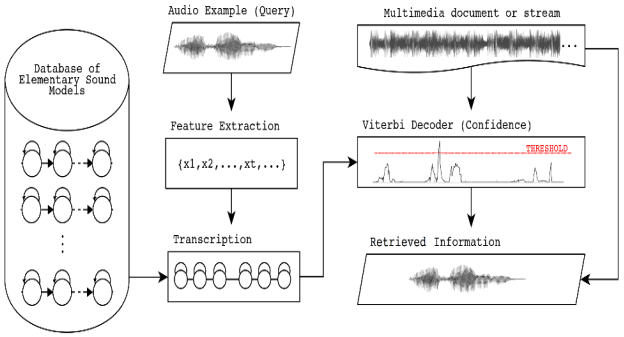

Figure 1. System layout

2.1 Feature extraction

We applied a 30 ms sliding window with 10 ms shift for computation of MFCCs (Mel Frequency Cepstral Coefficients), which may be considered as a standard in speech recognition. Each frame is represented by the 39–dimensional feature vector that is composed of 13 MFCCs along with their delta and delta–delta derivates (incl. 0th cepstral coefficient), followed by Cepstral Mean Normalisation. The feature vectors form observations for further acoustical modelling by HMM.

2.2 Concept of quasi phoneme-like models

The overall performance of our proposed approach for the SWS task relies on how precisely the developed models of ES units mimic real phoneme-based units. Thus they should preserve the linguistic information as much as possible while the non-linguistic features (related to speaker/gender variability, emotion, background noise, channel characteristics, etc) should be suppressed. This is a very challenging problem and can be solved only to a certain extent. We are aware of the fact that the developed ES units are only a rough approximation of linguistic units if no information about language structure is taken into account.

In these initial experiments, we developed the ES models by the following procedure:

1) All 20 hours of audio of the SWS2013 dataset were parameterized as a sequence of MFCC vectors.

2) Means and variances of Gaussian components of semi-continuous Gaussian Mixture PDF were estimated by unsupervised clustering of all MFCC vectors from the dataset.

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

The clustering was performed by the K–variable K–means procedure [4]. During the clustering, small clusters (containing less than 30 vectors) were discarded. This iterative procedure converged into 257 clusters, whose centroids defined the means of PDF's Gaussian components.

3) In the next step, the sequence of the MFCC vectors computed from the SWS2013 dataset was divided into segments of 20 vectors with 50% overlap, each spanning 0.1 seconds of audio, which is initially considered the average duration of ES units.. Then 1-state semi-continuous density HMM (SD-HMM) was estimated for each segment. Since all SD-HMMs utilize the PDFs with the same Gaussian components, only weights of the Gaussians in the mixtures were computed during ES model estimation. Thus the HMM states are defined by PDFs composed of Gaussian mixtures described by 257-dimensional weight vectors. Such approach is much less computational demanding than conventional HMM training. Since it results in the creation of about 1.5 million models, the amount of weight vectors was massively reduced, again by the K-variable K-means clustering (only 207 clusters with the highest amount of vectors were preserved). This procedure results in the creation of 207 models of ES units.

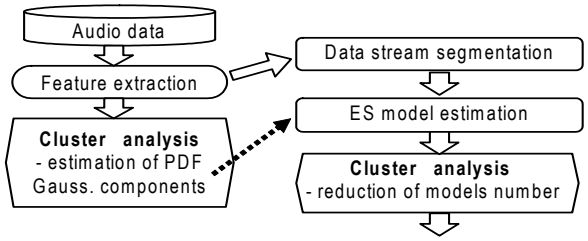

Figure 2. ES models development

The queries were represented by concatenation of these ES models as it is shown in Figure 1. For each query, the best sequence of the ES models was obtained by Viterbi decoding according to [2]. It should be remarked that during decoding, each 1-state HMM was replaced by the 10-equal-state HMM (obtained by concatenation of the same states), with an aim to avoid detection of very short segments. This procedure secured that the decoded sequence passed through the same state at least 10 times, i.e. the process retained in the same state at least 0.1 seconds.

3. RESULTS AND DISCUSSION

Maximal confidence obtained by Viterbi decoding was chosen as the score in submission. The segments with score about the given threshold (the threshold was tuned on the development queries) were labelled as YES decision. With an effort to decrease high false alarm rate, the segments with duration $D > 2Q$ or $D < 0.5Q$ (where Q is the duration of the query) were filtered out.

The performance of the submitted system was evaluated in terms of several measures, as defined by the SWS2013 task [1,5]: Actual/ Maximum Term-Weighted Values ATWV/ MTWV (weighted combination of miss and false alarm error rates); and a normalized cross-entropy metric Cnxe. The official results (late submission) are summarized in Table 1. In addition the amount of processing resources, namely Indexing/Searching Speed Factors (ISF/SSF) as defined in [5], are estimated.

We reckon that the following bottlenecks caused very low performance of this first version of the submitted system:

- Since the ES models are obtained by unsupervised clustering, a very important issue is that the features belonging to the same linguistic unit should be grouped together in the feature vector space. The problem with MFCC based features is that they are advisable for both speech recognition as well as speaker discrimination tasks, what is disadvantageous in our case. Some discriminant analysis transform on MFCC, to obtain the speaker–independent features, might be applied prior to clustering;
- We haven't investigated the impact of the size of the ES inventory on the search performance. Due to lack of time also only a simplified HMM training without re-estimation was applied for ES model development;
- After inspection of the results we have noticed that ES models of noise and silence were very often found in the decoding sequences, and overall confidence was affected by these "non linguistic" models. Prior proper speech activity detection on queries as well as audio content (to avoid modelling of non-speech events) would also help.

Table 1. The submission results

	ATWV	MTWV	Cnxe	Cnxe_min
Devel.	- 0.091	0	1.011	0.951
Eval.	- 0.027	0.001	1.011	0.945

The system was run on a workstation with 32-core CPU. All the programming was made in Matlab and it was not optimized for speed. In pre-processing (indexing) phase, only MFCC features were precomputed. It took about 1/60 of real-time, that means ISF = 1205/ (71839+696 sec.) = 0.017 for the evaluation data. Note that ES models computation is not considered in ISF. In the submitted version of the system, audio content-based adaptation of ES models is not considered, thus the development of ES models might be seen as an extra part. The processing time employed in searching (recomputed for single CPU) is approx. SSF = 1.08×10^6 / (71839 x 696 sec.) = 0.022. The peak memory usage in searching is very low because only the Viterbi algorithm is performed. Rough estimation is about 1 – 10 MBytes. Hence, the proposed system, if it is tuned (and compiled in other more suitable programming environment), might be computationally very efficient.

4. REFERENCES

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

[2] J. Nouza, J. Silovsky, "Fast keyword spotting in telephone speech". *Radioengineering*, 18(4), 2009, 665-670.

[3] R. Gubka, M. Kuba, "Elementary sound based audio pattern searching," *23rd Int. Conf. Radioelektronika 2013*, April 16-17, 2013, 325-328. DOI = http://dx.doi.org/10.1109/RadioElek.2013.6530940

[4] Reyes-Gomez, M.J.; Ellis, D.P.W., "Selection, parameter estimation, and discriminative training of hidden Markov models for general audio modeling," Int. Conf. on Multi-media and Expo, ICME '03, July 2003. I73-76

[5] L. J. Rodriguez–Fuentes, M. Penagarikano, "MediaEval 2013 Spoken Web Search Task: System Performance Measures," n. TR-2013-1, Dept. of Electricity and Electro-nics, Univ. of the Basque Country, 2013, http://gtts.ehu.es/gtts/NT/fulltext/rodriguezmediaeval13.pdf