A Hierarchical Architecture with Feature Selection for Audio Segmentation in a Broadcast News Domain

Taras Butko and Climent Nadeu

TALP Research Center, Department of Signal Theory and Communications,
Universitat Politècnica de Catalunya, Barcelona, Spain
taras.butko@upc.edu, climent.nadeu@upc.edu

Abstract

This work presents a hierarchical HMM-based audio segmentation system with feature selection designed for the Albayzin 2010 Evaluations. We propose an architecture that combines the outputs of individual binary detectors which were trained with a specific class-dependent feature set adapted to the characteristics of each class. A fast one-pass-training wrapper-based technique was used to perform a feature selection and an improvement in average accuracy with respect to using the whole set of features is reported.
Index Terms: audio segmentation, broadcast news, international evaluations

1. Introduction

Audio segmentation is the task of segmenting a continuous audio stream in terms of acoustically homogenous regions. Recently, the audio segmentation has received increasing attention for its applications in automatic indexing, subtitling, content analysis and information retrieval. Many research works address the problem of audio segmentation in different scenarios. In [1] the authors propose a method for robust speech, music, environment noise and silence segmentation of the audio recorded in different conditions such as TV studio, telephone etc. In [2] the audio stream from broadcast news domain is segmented into 5 different types including speech, commercials, environmental sound, physical violence and silence. The content based retrieval using TV programs is considered in [3], where 7 similar classes are defined.

Besides, several speech technologies can benefit from audio segmentation done at early steps. A previous identification of speech segments facilitates the task of speech recognition or speaker diarization. Furthermore audio segmentation is widely used to make online adaptation of ASR models or generating a set of acoustic cues for speech recognition to improve overall system performance [4].

In the context of the Albayzín-2010 evaluation campaign, which is an internationally-open set of evaluations organized by the Spanish network of speech technologies, an audio segmentation task was proposed and organized by the authors. For this evaluation we propose a system that uses a hierarchical architecture with HMM-GMM-based binary detectors. Each detector, one per class, uses a specific feature set, which is designed by adapting a feature selection technique recently introduced by the authors. In the experimental part, the results obtained by using the training data of the evaluation for both training, development and testing are presented. When compared with a one-step multi-class system, our system shows a 25% average relative improvement.

2. Albayzin 2010 audio segmentation evaluation

2.1. The database

The database used for evaluations consists of a Catalan broadcast news database from the 3/24 TV channel that was recorded by the TALP Research Center from the UPC, and was manually annotated by Verbio Technologies. Its production took place in 2009 under the Tecnoparla research project. The database includes around 87 hours of annotated audio (24 files of approximately 4 hours long). According to this material five different audio classes were defined (Table 1). The distribution of classes within the database is the following: Clean speech: 37%; Music: 5%; Speech with music in background: 15%; Speech with noise in background: 40%; Other: 3%. The database was splitted into 2 parts: for training/development (2/3 of the total amount of data), and testing (the remaining 1/3). The audio signals are provided in pcm format, mono, 16 bit resolution, and sampling frequency 16 kHz.

The Corporació Catalana de Mitjans Audiovisuals, owner of the multimedia content, allows its use for technology research and development.

Table 1. *The 5 acoustic classes defined for evaluations.*

Class	*Description*
Speech [sp]	Clean speech in studio from a close microphone
Music [mu]	Music is understood in a general sense
Speech over music [sm]	Overlapping of speech and music classes or speech with noise in background and music classes
Speech over noise [sn]	Speech which is not recorded in studio conditions, or it is overlapped with some type of noise (applause, traffic noise, etc.), or includes several simultaneous voices (for instance, synchronous translation)
Other [ot]*	This class refers to any type of audio signal (including noises) that doesn't correspond to the other four classes

* Not evaluated in final tests

2.2. Metric

The metric is defined as a relative error averaged over all acoustic classes (ACs):

$$Error = \underset{i}{average}\left(\frac{dur(miss_i) + dur(fa_i)}{dur(ref_i)}\right) \quad (1)$$

where

dur(miss$_i$) – the total duration of all deletion errors (misses) for the i*th* AC.

dur(fa$_i$) – the total duration of all insertion errors (false alarms) for the i*th* AC.

dur(ref$_i$) – the total duration of all the i*th* AC instances according to the reference file.

The incorrectly classified audio segment (a substitution) is computed both as a deletion error for one AC and an insertion error for another. A forgiveness collar of 1 sec (both + and -) is not scored around each reference boundary. This accounts for both the inconsistent human annotation and the uncertainty about when an AC begins/ends.

3. The UPC audio segmentation system

3.1. Features

A set of audio spectro-temporal features, like those used in automatic speech recognition, is extracted to describe every audio frame. It consists of 16 frequency-filtered (FF) log filter-bank energies with their first time derivatives [5], which represent the spectral envelope of the audio waveform within a frame, as well as its temporal evolution. In total, a 32-dimensional feature vector is used. The FF feature extraction scheme consists in calculating a log filter-bank energy vector for each signal frame (in our experiments the frame length is 30 ms with 20 ms shift, Hamming window is applied) and then applying a FIR filter *h(k)* on this vector along the frequency axis. We use the *h(k)={1, 0, -1}* filter in our approach. The end-points are taken into account which represent the absolute energies of the first and the last filter banks.

3.2. The system architecture

Our hierarchical system architecture is a group of detectors (called modules), where each module is responsible for detection of one acoustic class of interest [6]. As input it uses the output of the preceding module and has 2 outputs: the first corresponds to audio segments detected as corresponding class of interest, and the other is the rest of the input stream. One of the most important decisions when using this kind of architecture is to put the modules in the best order in terms of information flow, since some modules may benefit greatly from the previous detection of certain classes. For instance, previous detection of the classes that show high confusion with subsequent classes potentially can improve the overall performance.

On the other hand, in this type of architecture, it is not necessary to have the same classifier, feature set and/or topology for different detectors. Tuning of parameters is done in each the system independently, and the two-class detection can be done in a fast and easy way. Given the modules, the detection accuracy can be computed individually and a priori. Those modules with best accuracies are then placed in the early stages to facilitate the subsequent detection of the classes with worst individual accuracies.

In our implementation, each binary detector consists of 2 HMMs: "Class" and "non-Class". Using the training approach known as one-against-all method [7], all the classes different from "Class" are used to train the "non-Class" model. Both HMMs have 3-states (with only 1 emitting state) and the observation distributions are Gaussian mixtures with continuous densities, and consist of 64 components with diagonal covariance matrices. The HTK[8] toolkit is used to perform training and the final segmentation.

The flow diagram of our hierarchical architecture is presented in Figure 1. The whole detection system consists of 5 binary detectors. Each binary detector (except silence) is trained using the features which were selected during the fast selection procedure (described in the next section).

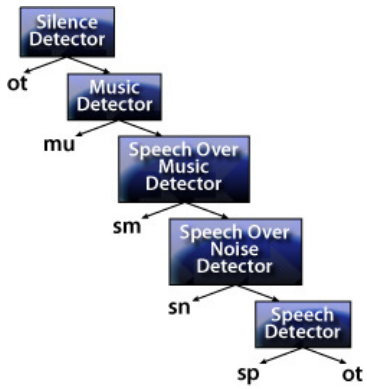

Figure 1: *Flow diagram of the hierarchical architecture.*

4. Feature selection

Actually, feature selection plays a central role in the tasks of classification and data mining, since redundant and irrelevant features often degrade the performance of classification algorithms [9]. In this paper, we use a fast one-pass-training feature selection technique [10] that avoids retraining of acoustic models during each evaluation of the candidate feature set.

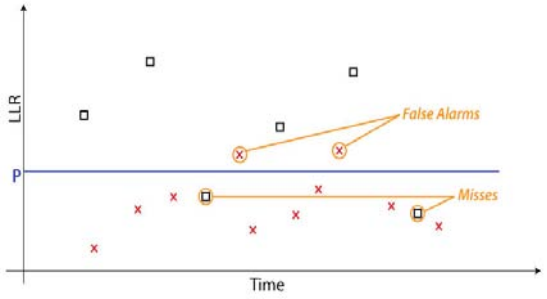

Figure 2: *Log-likelihood ratios Δi of the "Class" (squares) and the "non- Class" (crosses) segments.*

In order to perform feature selection, the database was divided into 2 parts: training and development. The fast one-pass-training feature selection technique when applied to the audio segmentation task consists in following steps:

1. Perform an initial training of "Class" and "non-Class" HMMs using the whole set of 32 FF features using the training part of the database.

2. Cut the development database into the short segments of 10 sec. Each such a segment belongs to either "Class" or "non-Class" (according to the ground truth labels).

3. Compute the log-likelihood ratios (LLRs) Δ_i of each such a segment given "Class" and "non-Class" models estimated on the step 1.

As an example, in Figure 2 we display the LLRs for all such segments in the labeled development database. Squares correspond to the "Class" instances while crosses correspond to "non-Class" instances. We consider the parameter P as a threshold (the horizontal line in Figure 2). We assume that the ith instance is detected as "Class" if its LLR ΔL_i is above P, otherwise it is detected as "non-Class". Thus, all "Class" instances (squares in Figure 2) below the P line are misses, and all "non-Class" instances (crosses) are false alarms. The parameter P is selected in such a way that the numbers of misses and false alarms are equal (equal error rate). The total number of errors (misses plus false alarms) is used as an objective function Ω for feature selection.

4. The LLR Δ_i of each segment is decomposed into a sum of "contributions" [10] coming from each feature j ($j \in 1..32$)

$$\Delta_i = \delta_{i,1} + \delta_{i,2} + \ldots + \delta_{i,32}, \qquad (2)$$

Using the sequential forward selection (SFS) approach, we iteratively select features (that correspond to the terms in (2)) that maximize the objective function Ω.

5. Experimental results

There are 16 sessions available for designing the audio segmentation system according to the evaluation plan. Half of the sessions we decided to use for training/development and the other half for testing.

First we select the appropriate number of Gaussians per HMM model for each binary detector. Actually, this number is a trade-off between the improvement in performance and the execution time needed to train the models with corresponding number of Gaussians. With 256 Gaussians we got the acceptable results. Fig. 3 demonstrates the mean error-rate obtained with increasing of the number of Gaussian mixtures per model.

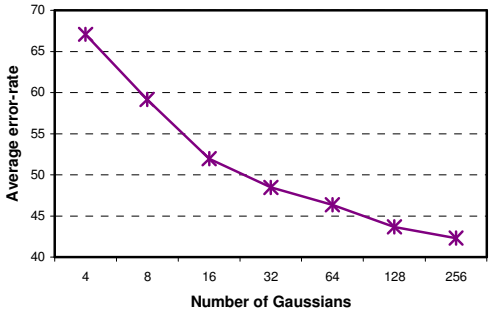

Figure 3: *Relation between mean error-rate and the number of mixtures per each GMM model.*

In Figure 4 we compare different system architectures. The "One-step multi-class" system corresponds to the HMM audio segmentation performed in one step. The "Hierarchical" architecture is described in sub-section 3.2. Finally, the system "Hierarchical + FS" is the same as previous but uses the feature selection.

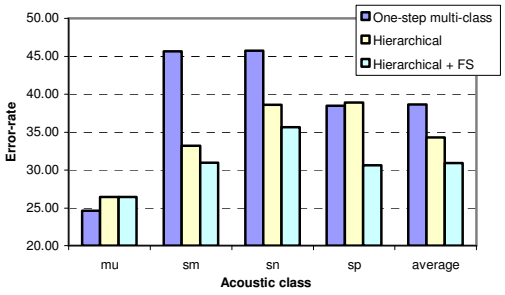

Figure 4: *Comparison of different detection systems.*

According to results from Figure 4, the hierarchical structure of audio segmentation outperforms the one-step multi-class detection system (about 8% of absolute error-rate reduction in average). Feature selection further improves the results for all classes (except music) and in average the absolute improvement is about 3%. Besides of that improvement we reduced the size of feature vector from 32 to 24 features in average.

The confusion matrix that corresponds to the "Hierarchical + FS" segmentation system is presented in Table 2, which shows the percentage of hypothesized AEs (rows) that are associated to the reference AEs (columns), so that all the numbers out of the main diagonal correspond to confusions.

Table 2: *The confusion matrix of acoustic classes.*

	mu	sp	sm	sn
mu	90	0	6	3
sp	0	83	1	16
sm	2	2	88	8
sn	0	10	7	83

As we see, the most common errors are confusions between "Music" and "Speech over music" classes and also among "Speech", "Speech over noise" and "Speech over music". Indeed, these classes have very similar acoustic content. Besides in many cases the ground truth labeling of audio is based on subjective reasoning of the annotator.

Table 3 we summarize the final results on testing database.

Table 3. *Final results on testing database.*

Database	**Error-rate**				
	mu	**sp**	**sm**	**sn**	**Average**
Result1	**26.40**	**44.20**	**33.88**	**41.52**	**36.50**
Result2	24.55	41.82	32.01	40.92	34.82

Where the *Result1* is obtained by the date of the presentation of final results. For that system we used the Gaussian models with only 64 mixtures and only 33% of training data were used to train

the models. The *Result2* was obtained using Gaussian models with only 256 mixtures and 100% of training data.

The CPU time employed to perform testing is described below:
Feature extraction: **546 sec;**
Viterbi segmentation: **3329 sec;**
Total: **3845 sec**.

This processes were executed on PC with Intel Core 2 CPU, 2.13 GHz, 1Gb of RAM.

6. Conclusions

In this work we proposed a hierarchical HMM-based system for broadcast news audio segmentation designed for Albayzin-2010 evaluation campaign. The main advantage of such a system is that each binary detector is placed in such order that previous detections improve the results of subsequent detector.

By using a fast one-pass-training feature selection approach we have selected the subset of features that shows the best detection rate for each acoustic class, observing an improvement in average accuracy with respect to using the whole set of features. The dimension of feature vector was reduced to 24 features (in average). Such a fast technique is a good alternative to the conventional SFS hill-climbing approach when the amount of data used for training the acoustic models is large.

When compared with a one-step multi-class system, our system shows a 25% average relative improvement.

7. Acknowledgements

The authors wish to thank our colleagues at GTTS, GTC-VIVOLAB, GSI, CEPHIS-UAB, ATVS-UAM, GTM, GTH-UPM for their participation in the evaluation. Also, the authors are very grateful to Henrik Schulz for managing the collection of the database and helping for its annotation. This work has been funded by the Spanish project SAPIRE (TEC2007-65470). The first author is partially supported by a grant from the Catalan autonomous government.

8. References

[1] L. Lie, J. Hao and Z. HongJiang, "A robust audio classification and segmentation method", Proc. *9th ACM conference on Multimedia*, p. 203-211, 2001

[2] T. L. Nwe H. Li, "Broadcast news segmentation by audio type analysis", in Proc. *ICASSP*, vol. 2, pp. 1065-1068, 2005

[3] D. Li, I.K. Sethi, N. Dimitrova, T. McGee, "Classification of general audio data for content-based retrieval", in *Pattern Recognition Letters*, v. 22, pp. 533-544, 2001

[4] H. Meinedo, J. Neto, "Audio Segmentation, Classification And Clustering in a Broadcast News Task", Proc. *ICASSP*, vol. 2, pp. 5-8, 2003

[5] C. Nadeu, J. Hernando, M. Gorricho, "On the decorrelation of filter-bank energies in speech recognition", in Proc. *European Speech Processing Conference*, pp. 1381–1384, 1995

[6] M. Aguilo, T. Butko, A. Temko, C. Nadeu, "A Hierarchical Architecture for Audio Segmentation in a Broadcast News Task", In Proc. *I Iberian SLTech - I Joint SIG-IL/Microsoft Workshop on Speech and Language Technologies for Iberian Languages*, 2009

[7] R. Rifkin, A. Klautau, "In defense of One-Vs-All Classification", *Journal of Machine learning Research,* vol. 5, pp.101-141, 2004

[8] S.J. Young et al., "The HTK Book (for HTK Version 3.2)", Cambridge University, 2002

[9] R. Kohavi, G. John, "Wrappers for feature subset selection", *Artificial Intelligence, Spec. Issue on Relevance*, vol. 97, pp. 273-324, 1997

[10] T. Butko, C. Nadeu, "A Fast One-Pass-Training Feature Selection Technique for GMM-based Acoustic Event Detection with Audio-Visual Data", in proc. *Interspeech*, 2010