A Hierarchical Architecture for Audio Segmentation in a Broadcast News Task

Mateu Aguilo, Taras Butko, Andrey Temko, Climent Nadeu

Department of Signal Theory and Communications,
TALP Research Center
Universitat Politècnica de Catalunya, Barcelona, Spain
{maguilo,butko,temko,climent}@gps.tsc.upc.edu

Abstract

In the broadcast news domain audio segmentation is an important pre-processing step for other speech technologies like speech recognition and speech diarization. In this work we propose an architecture that allows to integrate the individual detections of various acoustic classes. By implementing a different algorithm adapted to the characteristics of each class, we can obtain much better results than using a generic detector for all classes. Additionally, new features suited to detect telephone channel speech over wideband music that improve the accuracy are also introduced.

Index Terms: audio segmentation, acoustic event detection, music detection, telephone speech, software architecture

1. Introduction

The TECNOPARLA project aims to develop speech technologies in the Catalan language focusing on the broadcast news task. It involves language identification, automatic speech recognition (ASR), machine translation, speech synthesis and speaker diarization [1].

Audio segmentation is the task of segmenting a continuous audio stream in terms of acoustically homogenous regions. 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 systems. 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 [2]. In [3] audio classes are defined according to human perception which provide a clue towards detecting a particular event. The audio streams are segmented into five different types including speech, commercials, environmental sound, physical violence and silence. Similarly in [4] five audio classes are defined: silence, music, background sound, pure speech, and non-pure speech which includes speech over music and speech over noise. The definition of audio classes depends much on the data and application domain.

In this work the database consists of 43h and 25m of audio coming from audio-visual recordings of Àgora debate program of the Catalan TV (TV3). According to this material we define six different audio classes:

- "Speech". This is pure speech recorded in the studio without background such as music.
- "Speech over music". This category includes all studio speech with music in the background.
- "Telephone speech". Some sections of the program have telephonic interventions from the viewers. These interventions are mixed in the program's main audio stream as a wide band stream.
- "Telephone speech over music". The same as previous class but additionally there is music in the background.
- "Music". Pure music recorded in the studio without any speech on top of it.
- "Silence".

Since silences are not labeled, the evaluation of "silence" class is not included in our task. However it is detected to facilitate the detection of the other classes. Moreover, "telephone speech" class is poorly represented in the database (see Section 3), so this class is not evaluated either.

We propose a hierarchical architecture for detecting acoustic classes using a set of binary detection systems. For comparison we also show an alternative system with a one-step multiclass detector described in [5].

The rest of this paper is organized as follows: Section 2.1 describes the classical one-step multiclass segmentation approach. The hierachical structure for segmentation is presented in 2.2. Section 3 presents experimental results and Section 4 concludes the work.

2. Audio segmentation

2.1. One-step multiclass detection

2.1.1. Features

The audio signal (16 kHz sampling rate) is framed using 30 ms Hamming window and 10 ms shift. For each frame, a set of spectral parameters has been extracted. It consists of the concatenation of two types of parameters: 1) 16 Frequency-Filtered (FF) log filter-bank energies [10], along with the first and the second time derivatives; and 2) a set of the following parameters: zero-crossing rate, short time energy, 4 sub-band energies, spectral flux, calculated for each of the defined sub-bands, spectral centroid, and spectral bandwidth. In total, a vector of 60 components is built to represent each frame. The mean and the standard deviation parameters have been computed over all frames in a 0.5sec window with a 100ms shift, thus forming one vector of 120 elements.

2.1.2. Classifier

For each pair of classes an SVM classifier is trained. A dataset reduction algorithm based on PSVMs [6] to cope with the enormous amount of data available for training is applied . Those sets of feature vectors whose PSVM classifier accuracies in the middle (not the best classifiers nor the worst, in contrast with [6]) are finally used to train the final SVM classifier. Using a

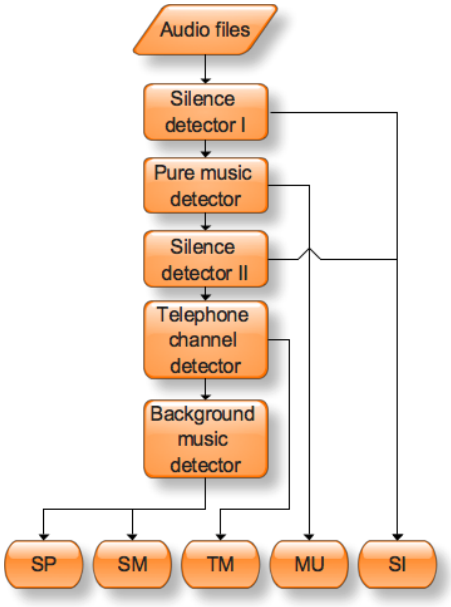

Figure 1: *Flow diagram of the hierarchical architecture.* **SP***: Pure speech.* **SM***: Speech over background music.* **TM***: Telephone speech over background music.* **MU***: Pure music.* **SI***: Silence.*

DAG architecture, as proposed in [9], each frame is classified in the final stage.

2.2. Hierarchical architecture

The hierarchical architecture (Figure 1) is a group of detectors (called modules), where each module is responsible for detection of one class of interest. 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.

2.2.1. Silence detection

The silence detector before the "music" detector is based on the derivative of the short time energy. This is done to avoid confusion with silences that have "musical" spectra. The algorithm can be described as follows:

- In the first stage the audio signal is low-band filtered at 1.5kHz. Although this filtering may cause problems with fricatives, that might become missdetected silences, this is dealt with in a post-process stage by using time constraints.
- The short time energy is convoluted with a 31 samples derivative filter, as proposed in [7] with the modifications in [8], to enhance the dynamics of the signal.
- Finally a threshold is tuned to separate the speech and non-speech frames. A final post process stage smooths the decisions and places time constraints (by using a finite state automat) to meet the evaluation requirements.

This detector has been tuned to give class "silence" at its output only when the confidence is high.

The second silence detector removes most of the silences to prepare the signal for the subsequent modules. Since there are no references for the silences it is trained in unsupervised manner. The algorithm can be described as:

- The short-time energy of the signal is then transformed to the logarithmic scale, and a GMM of N Gaussians is trained. The Gaussians with a lower weight than a fixed percent of the weight of the Gaussian with the highest weight are discarded (if any).
- The N_{sil} Gaussians with the lowest mean are selected for the silence class (as they represent the frames with low energy). The $N - N_{sil}$ other Gaussians are left for the non-silence class.
- With the Gaussians selected for silence, the whole show, is evaluated frame by frame. The same is done with the non-silence class. Comparing the silence and non-silence likelihoods, plus a penalty for silence, each frame is classified as silence or non-silence.
- Then the decisions are smoothed using a median filter. Finally, silences longer than the specified minimum duration are writen to the output file.

2.2.2. Music detection

Music segments usually appear at the beginning and the end of the show or when the topic of discussions changes. Music serves as introduction to show and it attracts attention of the audience towards its beginning. It is worth to mention that the melody in AGORA shows doesn't vary and only 2 or 3 different musical instruments could be distinguished: drums, saxophone and piano. To detect music segments, a one-against-all topology of detection process is selected [10] [11]. As disscussed in [10] the advantage of this topology is the possibility of using a specific kind of features for each particular classification task. The differences between the music and non-music class can be noticed in the spectral domain. The periodograms of 0.5 sec long music and speech segments are displayed in Figure 2 (we selected "speech" class as a representative of "non-music" metaclass).

As it seen from Figure 2, the spectral envelope is flatter for "music" class while for "speech" class the energy is concentrated in lower bands. Typical ASR features are used in this music detector (the FF coefficients with their first time derivatives. In total the feature vector has 32 components). Finally, mean normalization is applied. We model each of the two classes separately using Hidden Markov Models (HMMs) and apply Viterbi decoding for final segmentation. The "music" HMM model consists of 2 emitting states with 5 Gaussians per state, while "non-music" model has 3 emitting states and 9 Gaussians per state, as its observation distribution is more complex. Both of the models have left-to-right connected state transitions.

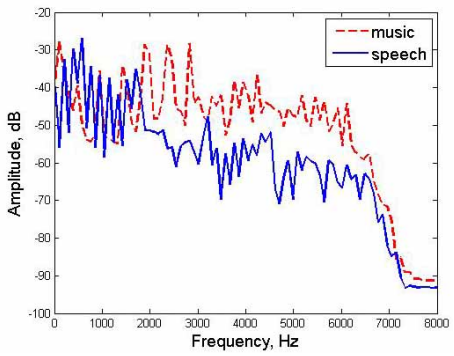

Figure 2: *Periodograms corresponding to "speech" and "music" classes. Sampling rate 16 kHz.*

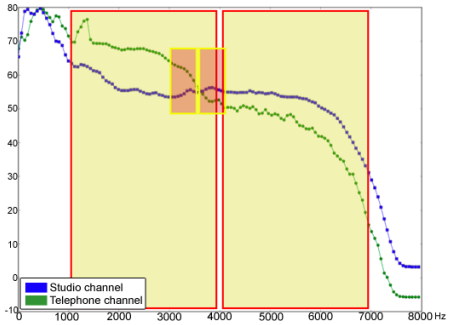

Figure 3: *Sub-band couples for the spectral slope superposed over periodograms corresponding to "speech" and "telephone speech over music" classes. Sampling rate 16 kHz.*

Using the proposed detection scheme the confusion between speech and music classes is minimal.

2.2.3. *Speech over music detection*

Often the discussions in Àgora shows start when music is still in the background. In this case we call it a "speech over music" segment. We use the same feature set as well as detection scheme as in the previously described music detector. Depending on the ratio between the energies of speech and background music, the spectrum will be more or less similar to the spectrum of the "speech" class and, in extreme case when the energy of music in background is very low, the differences between the corresponding spectra are negligible. In such cases the confussion between classes increases.

2.2.4. *Telephone speech over music detection*

To detect "telephone speech over music" class we use the two-class version of the system described in subsection 2.1. In our scenario "telephone speech over music" class is composed of the music that spans all the frequency range 0-8 Hz and telephone speech which is in low frequency range. New features, called spectral slopes, are concatenated to the existing ones to enhance the detection accuracy. To compute a spectral slope, two different couples of subbands are defined. These subbands have been chosen to discriminate between "telephone speech over music" and the rest of audio based on the slope of the spectrum in the region around 4000 Hz, the end of the band of telephone speech, beyond which only music frequency components exist. The first couple is made of the sub-bands $[1000-3000]Hz$ and $[3000-7000]Hz$ and the second is consists of the sub-bands $[3000-3500]Hz$ and $[3500-4000]Hz$ aims to parametrize the energy in the region where the energy drop should appear for the "telephone speech over music" class. A feature vector ss is computed for each couple as:

$$\boldsymbol{ss} = (S_1, S_2, \frac{S_1}{S_2}) \quad (1)$$

where S_1, S_2 are total energies of the first and second sub-band respectively.

Experimental results have shown that the dynamics of the spectral slope features are helpful for the detection of the "telephone speech over music" class. Thus the deltas and accelerations are added to the final feature vector. Finally we obtain a set of 18 values for each frame (with two sub-band couples), which is concatenated with all the features listed in subsection 2.1 leading to a feature vector of 78 components.

3. Experiments

3.1. Database description

As mentioned in Section 1, the database used to test the system consists of 43h and 25m of spontaneous speech in the context of a debate TV program. Each program has been cut in two parts to exclude the commercials, and each part has a duration of about 40 minutes. Àgora is a highly moderated program where around 7 different speakers discuss a wide variety of topics. The Àgora program has a fairly fixed structure, although no use of this information has been made in order to keep the system general. As can be observed in Table 1, the dominant class is "speech" appearing 83.21% of the time. The class "telephone speech" (without background music) has been discarded because of its extremely low appearance. The class "OV" (overlapped speech coming from two or more speakers) has been left out for future work.

Table 1: *Distributions of the events in the database.*

Acoustic class	Appearance (%)
Speech	83.21
Speech over music	9.78
Telephone speech	0.02
Telephone speech over music	2.48
Music	1.16
Overlapped speech	3.36

The Àgora database has been manually annotated. In order to evaluate the audio segmentation system the database has been divided in three sets: training, development and evaluation. The sets have been designed to have a similar distribution to the whole database (see Table 1). This leaves 8h of audio for development, 8h for evaluation, and 27h for training.

3.2. Results

In order to evaluate the improvements introduced by the hierarchical architecture two systems are compared:

- One-step (described in subsection 2.1).
- Hierarchical (described in subsection 2.2).

We use two metrics to compare both systems: the first metric is the ratio between the time when the hypothesis doesn't match the reference (error time) and the total time of the audio recordings. The second metric is the average ratio between the error time and the total time of audio per each class.

$$ERROR = \frac{t_{error}}{t_{total}} \quad (2)$$

$$MERROR = \frac{1}{N_{class}} \sum_{i=1}^{N_{class}} \frac{t_{error}(class_i)}{t_{total}(class_i)} \quad (3)$$

Table 2 shows that the use of hierarchical architecture improves

Table 2: *Segmentation results*

System	ERROR (%)	MERROR (%)
One-step	7.20	46.88
Hierarchical	**3.71**	**3.4**

Table 3: *Segmentation results per class*

Class	One-step (%)	Hierarchical (%)
Pure speech	6.5	4.8
Pure music	32.0	2.4
Speech over music	75.3	4.9
Telephone speech over music	73.7	1.5

both $ERROR$ and $MERROR$. As can be seen in Table 3 the large reduction of $MERROR$ can be explained by the poor results the *One-step* system achieves in the minority classes, while the *Hierarchical* system performs rather well for all classes.

The proposed spectral slope features yield a strong relative improvement in detection of the class "telephone speech over music", as displayed in Table 4.

Table 4: *"Telephone speech over music" detection results*

System	ERROR (%)
w/o Spectral Slope	3.5
w/ Spectral Slope	**1.5**

4. Conclusions

From the results in Table 2, it can be observed that the use of a more flexible architecture allows to develop a system that is more suited to a particular task. A large improvement can be obtained by using a set of detectors, which are properly combined and also tuned to the different target classes.

The one-step multiclass detection system tries to detect the most dominant class while doing worse in other classes; this is reflected in the large value of $MERROR$. On the other hand, an hierarchical system does not detect the most dominant class ("speech") explicitely, converse, it detects all other classes and "speech" is what is left.

Future work will be devoted to improve performance of the "speech over music" detector. For instance, the current system produces a large proportion of errors in the speech segments with very low level of music in the background. The forthcoming annotation of the "silence" class in the Àgora database will make it possible to tune the parameters of the silence detectors and get an improvementof the overall accuracy.

5. Acknowledgements

This work has been funded by the Generalitat de Catalunya in the framework of the TECNOPARLA project and also by the Spanish SAPIRE project (TEC2007-65470).

6. References

[1] H. Schulz, M. R. Costa-Juss, J. R. A. Fonollosa, "TECNOPARLA - Speech technologies for Catalanand its application to Speech-to-speech Translation", Procesamiento del Lenguaje Natural, vol. 41, pp. 319-320, 2008.

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

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

[4] L. Lu, S. Z. Li, H.-J. Zhang "Content-based Audio Segmentation Using Support Vector Machines", IEEE International Conference on Multimedia and Expo, pp. 956-959, 2001.

[5] A. Temko, C. Nadeu, J-I. Biel, "Acoustic Event Detection: SVM-based System and Evaluation Setup in CLEAR'07", in Multimodal Technologies for Perception of Humans, LNCS, vol.4625, pp.354-363, Springer, 2008.

[6] A. Temko, D. Macho, C. Nadeu, "Enhanced SVM Training for Robust Speech Activity Detection", IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP, 2007.

[7] A. T. Qi Li, J. Zheng and Q. Zhou, "Robust endpoint detection and energy normalization for real-time speech and speaker recognition", IEEE Transactions on Speech and Audio Processing, vol. 10, 2002.

[8] M. Aguilo, "Detección de actividad oral en un sistema de diarización", Final Degree Project, UPC, 2005.

[9] J. Platt, N. Cristianini, J. Shawe-Taylor, "Large Margin DAGs for Multiclass Classification", Proc. Advances in Neural Information Processing Systems, vol. 12, pp. 547-553, 2000

[10] T. Butko, C. Canton-Ferrer, C. Segura, X. Giró, C. Nadeu, J. Hernando, J.R. Casas, "Improving Detection of Acoustic Events Using Audiovisual Data and Feature Level Fusion", accepted to Interspeech, 2009.

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