Albayzin Evaluation: The PRHLT-UPV Audio Segmentation System

J. A. Silvestre-Cerdà, A. Giménez, J. Andrés-Ferrer, J. Civera, and A. Juan

Universitat Politècnica de València, Camí de Vera s/n,
46022 València, Spain,
jsilvestre@dsic.upv.es,
http://translectures.eu

Abstract. This paper describes the audio segmentation system developed by the PRHLT research group at the UPV for the Albayzin Audio Segmentation Evaluation 2012. The PRHLT-UPV audio segmentation system is based on a conventional GMM-HMM speech recognition approach in which the vocabulary set is defined by the power set of segment classes. MFCC features were extracted to represent the acoustic signal and the AK toolkit was used for both, training acoustic models and performing audio segmentation. Experimental results reveals that our system provides an excellent performance on speech detection, so it could be successfully employed to provide speech segments to a diarization or speech recognition system.

1 Introduction

Audio segmentation is a task with applications in subtitling, content indexing and analysis that has received notable attention due to the increasing application of automatic speech recognition (ASR) systems to multimedia repositories and broadcast news [6–9]. Formally, this task can be stated as the segmentation of a continuous audio stream into acoustically homogenous regions. Audio segmentation facilitates posterior speech processing steps such as speaker diarization or speech recognition.

Previous work on audio segmentation can be classified into those tackling this task at the feature extraction level [1, 10, 4, 3], and those approximations working at the classification level [2, 5]. This latter approximation is adopted in our audio segmentation system.

This paper describes the PRHLT-UPV audio segmentation system. First, the Albayzin Audio Segmentation Evaluation 2012 is presented in Section 2. Next, a complete system description is provided in Section 3. Experimental results are presented in Section 4. Finally, conclusions are drawn in Section 5.

2 Albayzin 2012 audio segmentation evaluation

2.1 Database description

The database used for the audio segmentation evaluation consists of a Catalan broadcast news database from the 3/24 TV channel, which comprises 87 hours of acoustic data for training purposes. In this dataset, speech can be found in a 92% of the segments, music is present a 20% of the time and noise in the background a 40%. Another class called others was defined which can be found a 3% of the time. Regarding the overlapped classes, 40% of the time speech can be found along with noise and 15% of the time speech along with music. Table 1 shows the audio time distribution over all overlapping acoustic classes as disjoint sets for the training set.

In addition, two sets, *dev1* and *dev2*, from the Aragón Radio database of the Corporación Aragonesa de Radio y Televisión (CARTV), are used for developing and internal testing purposes, respectively. Both sets sums up to 4 hours of acoustic data. All audio signals are provided in PCM format, mono, little endian 16 bit resolution, and sampling frequency of 16 kHz.

Table 1. Audio time distribution of all overlapping classes for the training set.

Class	Time (h)	Time (%)
sp	31.85	38.2
mu	4.94	5.9
no	0.91	1.1
sp+mu	12.58	15.1
sp+no	31.36	37.6
no+mu	0.06	0.1
sp+no+mu	1.65	2.0
Total	87	100

2.2 Evaluation metric

In order to assess the quality of our system, we used the Segmentation Error Rate metric (SER), defined as the fraction of class time that is not correctly attributed to that specific class (speech, noise or music):

$$SER = \frac{\sum_n T(n)\left[max(N_{ref}(n), N_{sys}(n),) - N_{correct}(n)\right]}{\sum_n T(n) N_{ref}(n)} \tag{1}$$

where $T(n)$ is the duration of segment n, $N_{ref}(n)$ is the number of reference classes that are present in segment n, $N_{sys}(n)$ is the number of system classes that are present in segment n, and $N_{correct}(n)$ is the number of reference classes in segment n correctly assigned by the segmentation system.

A forgiveness collar of one second, before and after each reference boundary, will be considered in order to take into account both inconsistent human annotations and the uncertainty about when a class begins or ends.

3 System description

Audio segmentation can be viewed as a simplified case of ASR, in which the system vocabulary is constituted by the power set of segment classes: Speech (*sp*), music (*mu*) and noise (*no*). For the Albayzin evaluation the silence (*si*) class is also included to denote that none of the three classes is present in a given time instant. Thus, the system vocabulary is defined as

$$\mathcal{C} = \{sp, mu, no, sp + mu, sp + no, mu + no, sp + mu + no, si\} \tag{2}$$

Provided an audio stream $\boldsymbol{x}$, the segmentation problem can be stated from a statistical point of view as the search of a sequence of class labels $\hat{\boldsymbol{c}}$ so that

$$\hat{\boldsymbol{c}} = \operatorname*{argmax}_{\boldsymbol{c} \in \mathcal{C}^*} p(\boldsymbol{x} \mid \boldsymbol{c})\, p(\boldsymbol{c}) \tag{3}$$

where, as in ASR, $p(\boldsymbol{x} \mid \boldsymbol{c})$ and $p(\boldsymbol{c})$ are modelled by acoustic and language models, respectively. In our case, it should be noted that each word is composed by a single phonem.

Acoustic models were trained on MFCC feature vectors computed from acoustic samples using the HTK HCopy tool. We used a 0.97 coefficient preemphasis filter and a 25 ms Hamming window that moves every 10 ms over the acoustic signal. From each 10ms frame, a feature vector of 12 MFCC coefficients is obtained using a 26 channel filter bank. Finally, the energy coefficient and the first and second time derivatives of the cepstrum coefficients are added to the feature vector.

Each segment class is represented by a single-state Hidden Markov Model (HMM) without loops, and its emission probability is modelled by a Gaussian Mixture Model (GMM). Acoustic HMM-GMM models were trained using the AK toolkit[1], which implements the conventional Baum-Welch algorithm. For each segment class, the number of mixture components per state was tuned on the development set. A 5-gram back-off language model with constant discount was trained on the sequence of class labels using the SRILM toolkit [11]. Constant discounts for each order were optimised on the development set. The segmentation process (search) was also carried out by the AK toolkit.

4 Experimental results

This section is devoted to the description of the experimental setup and results performed before submitting our final audio segmentation system. For these

[1] http://sourceforge.net/projects/aktoolkit

experiments, acoustic and language models were trained on the *training* set, while acoustic and language model parameters were tuned on the *dev1* set.

Table 2 shows SER figures computed on the *dev1* and *dev2* sets. In addition to the overall SER, SER values are provided for each acoustic class (speech, noise, music) in isolation. As observed in Table 2, our audio segmentation system offers an excellent performance in speech detection, so it could be successfully employed to provide speech segments to diarization or speech recognition systems.

Table 2. Segmentation error rate (SER) for the three acoustic classes (speech, music, noise) in isolation and overall SER, computed over the *dev1* and *dev2* sets.

Set	Speech	Music	Noise	Overall
dev1	1.2	25.3	71.4	24.9
dev2	2.2	20.2	71.2	26.4

However, the system provides low performance at detecting non-speech classes, specially the noise class. This fact can be explained by two reasons. First, we are using a feature representation of the acoustic signal that is focused on highlighting human voice characteristics, and conversely to abate acoustic features from music and noise. Secondly, few music and noise data samples appear in isolation (5% and 1%, respectively) to make feasible to robustly estimate acoustic models for these classes. For this reason, the global classifier suffers from a bias towards the isolated speech class. For instance, the posterior probability of an *sp+no* segment, given the isolated speech model parameter set is expected to be larger than that of the isolated noise model parameter set.

5 Conclusions

This paper has described the PRHLT-UPV audio segmentation system for the Albayzin evaluation. This system tackles the task of audio segmentation from the viewpoint of ASR system with a reduced vocabulary set. The vocabulary set comprises the speech, music and noise classes and combinations of those classes defined for this task. The experimental results show that this system provides excellent performance detecting speech segments, but its performance decays when dealing with music or noise segments.

Acknowledgments. The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 287755. Funding was also provided by the Spanish Government (iTrans2 project, TIN2009-14511; FPU scholarship AP2010-4349).

References

1. Ajmera, J., McCowan, I., Bourlard, H.: Speech/music segmentation using entropy and dynamism features in a hmm classification framework. Speech Communication 40(3), 351–363 (2003)
2. Bugatti, A., Flammini, A., Migliorati, P.: Audio classification in speech and music: A comparison between a statistical and a neural approach. EURASIP J. Audio Speech Music Process. pp. 372–378 (2002)
3. Gallardo-Antolín, A., Montero, J.M.: Histogram equalization-based features for speech, music, and song discrimination. IEEE Signal processing letters 17(7), 659–662 (2010)
4. Izumitani, T., Mukai, R., Kashino, K.: A background music detection method based on robust feature extraction. In: Proc. of ICASSP. pp. 13–16
5. Lavner, Y., Ruinskiy, D.: A decision-tree-based algorithm for speech/music classification and segmentation. EURASIP J. Audio Speech Music Process. 2009, 2:1–2:14 (2009)
6. Li, D., Sethi, I.K., Dimitrova, N., McGee, T.: Classification of general audio data for content-based retrieval. Pattern Recognition Letters 22(5), 533 – 544 (2001)
7. Lu, L., Jiang, H., Zhang, H.: A robust audio classification and segmentation method. In: Proc. of ACM International Conference on Multimedia. pp. 203–211 (2001)
8. Meinedo, H., Neto, J.: Audio segmentation, classification and clustering in a broadcast news task. In: Proc. of ICASSP. vol. 2, pp. 5–8 (2003)
9. Nwe, T., Li, H.: Broadcast news segmentation by audio type analysis. In: Proc. of ICASSP. vol. 2, pp. 1065–1068 (2005)
10. Panagiotakis, C., Tziritas, G.: A speech/music discriminator based on rms and zero-crossings. IEEE Transactions on Multimedia 7(1), 155–166 (2005)
11. Stolcke, A.: SRILM – an extensible language modeling toolkit. In: Proc. of ICSLP'02. pp. 901–904 (September 2002)