Audio Segmentation System by Aholab for Albayzin 2012 Evaluation Campaign

David Tavarez, Eva Navas, Daniel Erro, and Ibon Saratxaga

Aholab - Dept. of Electronics and Telecommunications. Faculty of Engineering.
University of the Basque Country. Alda. Urquijo s/n 48013 Bilbao.
{david,eva,derro,ibon}@aholab.ehu.es

Abstract. *This paper describes the system developed by Aholab Signal Processing Laboratory for the Albayzin 2012 audio segmentation evaluation campaign. A model based audio segmentation system that considers seven models (clean speech, narrow band speech, speech+music, speech+noise, music, silence, silence+music) has been built. This main system has a 21% of segmentation error in the development recordings. A post-processing step that analyzes the speech segments has been added to the main system, with a 27% improvement in the results of the music class, reducing the final segmentation error to 20%.*

Keywords: Automatic Audio Segmentation, Albayzin Evaluation Campaigns, Broadcast Speech

1 Introduction

Automatic audio segmentation is the process of dividing the audio stream into homogeneous sections according to its content. Depending on the application, the goal of the automatic segmentation process may be very diverse: separating speech from noise and music [1], separating male voice from female voice [2], separating the segments corresponding to different speakers [3], etc. Automatic audio segmentation has many applications and usually is used as a first preprocessing step to improve the performance of other systems like automatic speech recognition [4], speaker identification [5], content-based audio indexing and information retrieval [6],...

Three segmentation techniques have been mainly applied in automatic audio segmentation:

- *Metric based segmentation*: in these methods an acoustic distance measure is used to evaluate the similarity between two adjacent windows shifted along the audio stream. The audio stream is segmented at maxima of the distances and different regions are labeled in the stream accordingly. The most common distance measures applied are Kullback-Leibler Distance [7], Hotlling"s T^2 Distance [8] and Generalized Likelihood Ratio [9].
- *Model based segmentation*: a set of statistical models are built for each of the acoustic classes to be labeled (speech, noise, music, speech+noise,...).

Segmentation boundaries are assumed where there is a change in the acoustic class. Gaussian mixture models (GMM) [10], Hidden Markov Models (HMM) [11] and Artificial Neural Networks (ANN) [12] have all been applied to the model based audio segmentation task.

– *Model selection based segmentation*: these techniques apply a criterion to select the most suitable way to model an audio segment between using a single model for all the segment or using two different models. Usually the Bayesian Information Criterion (BIC) is applied for the model selection [13] [14], although some other hybrid criteria have also been proposed [15].

The Spanish Thematic Network for Speech Technologies (Red Temática en Tecnologías del Habla[1]) organizes each two years evaluation campaigns to objectively assess the validity of the algorithms developed for different speech technology related systems. In these campaigns, different research groups test their algorithms with a shared database, which allows for performance comparison and helps identifying new trends. Evaluations of text to speech systems [16] [17], automatic machine translation, language identification, diarization systems [18] and automatic audio segmentation [19] have been organized in past editions. This year one of the campaigns organized deals again with automatic audio segmentation. This paper describes the system developed by Aholab Signal Processing Laboratory to take part in this evaluation campaign.

The rest of the paper is organized as follows: section 2 presents the description of the system developed and the database used for training. The results obtained by the system are detailed in section 3 and finally section 4 deals with the conclusions of the work.

2 Description of the Proposed System

Figure 1 shows a diagram of the proposed Audio Segmentation System. In the first step a Viterbi segmentation is performed to locate the boundaries of the different acoustic events (speech, music, speech+music, specch+noise...) in the audio stream. Then, the labels obtained in the first step are treated in order to refine the boundaries, discard short silences and correct other minor mistakes. Finally, the speech segments are post-processed individually in order to identify those with low level music in the background.

In the following sections, the database used will be presented and each step of the system will be described in detail.

2.1 Database

The Albayzin 2012 database includes broadcast speech from the archive of Corporación Aragonesa de Radio y Televisión (CARTV). It contains around four hours of audio for development and another sixteen hours for testing. The Albayzin 2012 organization also provided for training the Catalan broadcast news

[1] http://www.rthabla.es/.

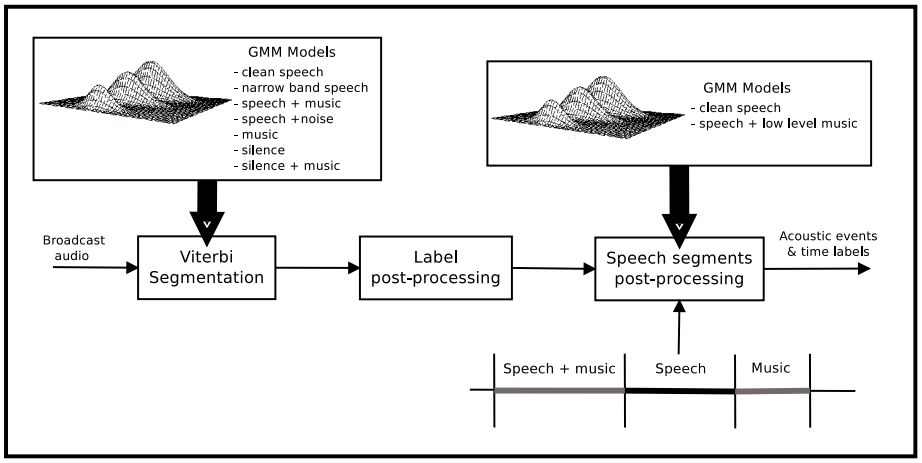

Fig. 1. Diagram of the proposed Audio Segmentation System

database from the 3/24 TV channel used in the 2010 Albayzin Audio Segmentation Evaluation [20]. The characteristics of this database are very different from the ones of CARTV. Due to the fact that the test database is also extracted from Aragón Radio database, only the CARTV database has been used in the development of the proposed audio segmentation system.

2.2 Viterbi Segmentation

A separate GMM model with 1024 mixtures was trained for clean speech, music, speech+noise, speech+music, narrow band speech, silence, and silence+music, using around four minutes of audio (for each model) of the development recordings of the database. The narrow band speech, silence and silence+music models have been included in order to improve the results of the system with these training recordings. The audio segments used to train the narrow band speech correspond to low pass filtered speech which is mostly labeled as speech+noise in the Albayzin 2012 reference labels. The silence and silence+music models have been trained with audio segments that correspond to pauses between sentences from the development recordings, taking into account the music in the background. These seven models are used in a Viterbi segmentation in order to detect the audio segments which contain the described acoustic events. Development experiments showed that the addition of derivatives of MFCC provided better segmentation results and that the use of 6 MFFC improved the location of the audio segments boundaries, therefore 6 MFCC with first and second derivatives were used for the classification. Finally, the segmentation labels obtained are mapped to the output classes that must be considered by the system (speech, music and noise) and are provided to the next module of the system, the label post-processing module.

2.3 Label post-processing

The aim of this step is to refine the boundaries, to discard short silences and to correct some minor errors made by the segmentation process. In the first place, silences shorter than 800 ms. are removed. These segments often correspond to pauses between words instead of boundaries between different audio segments, therefore they are eliminated and assimilated to the adjacent segments.

Development experiments also showed that short segments labeled as speech with noise in the segmentation process usually corresponded in fact to segments of speech with music in the background. This erroneous labeling occurred when the music was different from the one present in the training material of the model. In consequence, speech+noise segments shorter than 2.5 s. are relabeled in this post-processing step as speech+music. Similarly, segments labeled as silence+music longer than 800 ms. are relabeled as music. Finally, consecutive segments with the same label are unified in order to improve the performance of the post-processing step.

2.4 Speech segments post-processing

In this final step, the speech segments are processed in order to search for low level music in the background. When this low level music appears, the previous steps often mark the segments as clean speech, so a post-processing step as the one proposed in [21] is necessary to reduce the final error. First, a separate GMM G_{voice} with 1024 mixtures is trained using around fifteen minutes of real clean speech from the development recordings of the database. Similarly, a GMM G_{music} is trained using another fifteen minutes of speech with low level music. Finally, every segment labeled as clean speech is processed and if they are better modeled by G_{music} than by G_{voice} , the music label is added to the already existing speech label. Development experiments showed that, the use of 12 MFFC provided better results in this task, therefore, 12 MFCC with first and second derivatives were used for the classification in this step.

3 Results

This section presents the results obtained by the system in the development recordings of the Albayzin 2012 database. Table 1 shows the error of the proposed system after the segmentation and the label post-processing steps (main system) and after the speech segments post-processing module (post-processing). These results were obtained with the evaluation script provided by the Albayzin 2012 organization.

Table 2 presents the average CPU time required in order to process the development recordings. The time required for the post-processing step is also shown. These measures were made on a quad-core Intel Xeon 2.27 GHz computer with 32 GB memory.

As displayed in Table 1, there is a large variability in the obtained values, particularly for the main system, due principally to errors in the music and

Session	Main system	Post-processing	Session	Main system	Post-processing
1	25.67%	25.67%	17	1.61%	1.61%
2	39.68%	12.71%	18	5.02%	5.02%
3	43.99%	43.99%	19	28.36%	28.39%
4	20.74%	20.74%	20	24.71%	24.60%
5	28.29%	28.29%	21	28.36%	28.57%
6	35.82%	8.95%	22	2.24%	2.24%
7	37.09%	37.09%	23	8.92%	8.92%
8	3.25%	3.25%	24	11.85%	15.04%
9	36.25%	31.12%	25	31.73%	33.93%
10	34.44%	12.78%	26	8.75%	8.75%
11	18.65%	19.66%	27	30.55%	54.05%
12	37.84%	34.97%	28	13.51%	12.81%
13	28.54%	24.82%	29	31.32%	43.29%
14	6.74%	6.74%	30	9.68%	9.91%
15	19.60%	25.60%	31	16.30%	16.30%
16	8.61%	8.57%	32	40.09%	25.23%
ALL	21.45%	20.99%			

Table 1. Class Error Time of the system for the development sessions

Database	Main system	Post-processing
5 hours 16 minutes 34 seconds	2 hours 53 minutes 16 seconds	33 minutes 30 seconds

Table 2. CPU time required in order to process the development part of the database

noise labeling. By applying the speech segments post-processing the total error is reduced by 2.15% for the development part of the database, which barely proves the validity this step. However, if the acoustic events are evaluated individually, as it is displayed in tables 3 and 4, it becomes clearer how the post-processing step is decreasing the error. Table 3 shows the error of the proposed system after the segmentation step with the label post-processing and table 4 shows the error after the speech segments post-processing.

Error	Speech	Music	Noise
Missed Class Time	4.1%	38.7%	34.5%
False Alarm Class Time	1.1%	4.1%	9.9%
Total Class Error Time	5.19%	42.76%	44.33%

Table 3. Results of the main system for each speech, music and noise independently

As displayed in Table 3, the main source of the segmentation error resides on the music and noise detection (with 42.76% and 44.33% Class Error Time

respectively), while the speech labeling obtains good results (5.19%). In the post-processing step, low level music in the background of speech segments is located, and the result obtained can be seen in table 4.

Error	Speech	Music	Noise
Missed Class Time	4.1%	26.7%	34.5%
False Alarm Class Time	1.1%	4.3%	9.9%
Total Class Error Time	5.19%	31.02%	44.33%

Table 4. Results after post-processing for each speech, music and noise independently

As it can be seen in Table 4, the post-processing step decreases considerably the music Missed Class Time, reducing the obtained error by 27%, which proves the validity of this step for the enhancement of the labeling for music class.

4 Conclusions

This paper describes the system developed by Aholab Signal Processing Laboratory to take part in the Albayzin 2012 audio segmentation evaluation campaign. A model based audio segmentation system that considers seven models (clean speech, narrow band speech, speech+music, speech+noise, music, silence, silence+music) has been built. The error made by this system when working with the development recordings of the database from AragonRadio provided by Albayzin 2012 organization has been presented and analyzed. A main audio segmentation system has been built, with a 21.45% of Total Class Error Time in the development recordings. A post-processing step that analyzes the speech segments and tries to improve the labeling for music has been added to the main system, with a 27% improvement in the results for the music class, which translates in a 2.15% of Total Class Error Time Reduction in the whole system. With a 5.19% of speech class error, the proposed system presents as a good alternative to speech extraction in Broadcast Audio.

5 Acknowledgments

This work has been partially supported by UPV/EHU (Ayudas para la Formación de Personal Investigador), and the Spanish Ministry of Science and Innovation (Buceador Project, TEC2009-14094-C04-02).

References

1. Lu, L., Zhang, H., Jiang, H.: Content analysis for audio classification and segmentation. IEEE Transactions on Speech and Audio Processing, 10(7), pp. 504 - 516. (2002)
2. Ore, B.M., Slyh, R.E., Hansen, E.G.: Speaker Segmentation and Clustering using Gender Information. In: IEEE Odyssey 2006: The Speaker and Language Recognition Workshop, pp. 1 – 8. San Juan, Puerto Rico. (2006)
3. Moattar, M. H., Homayounpour, M. M.: A review on speaker diarization systems and approaches. Speech Communication, 54(10), pp. 1065 - 1103. (2012)
4. Rybach, D., Gollan, C.: Audio segmentation for speech recognition using segment features. In: IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP), pp. 4197 - 4200. Taipei, Taiwan (2009)
5. Reynolds, D. A., Torres-Carrasquillo, P.: Approaches and applications of audio diarization. In: IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP), pp. 953 - 956. Philadelphia, USA. (2005)
6. Meinedo, H., Neto, J.: Audio segmentation, classification and clustering in a broadcast news task. In: IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP), vol. 2, pp. 5 - 8. Hong-Kong, China. (2003)
7. Siegler, M., Jain, U., Raj, B., Stern, R.: Automatic segmentation, classification and clustering of broadcast news audio, In: DARPA Speech Recognition workshop, pp. 97 - 99. Cantilly, USA. (1997)
8. Zhou, B., Hansen, J.: Efficient audio stream segmentation via the combined T^2 statistic and Bayesian information criterion. IEEE Transactions on Speech and Audio Processing, 13(4), 467 - 474. (2005)
9. Meignier, S. Moraru, D. Fredouille, C., Bonastre, J. F., Besacier, L.: Step-by-step and integrated approaches in broadcast news speaker diarization, Computer Speech and Language, vol. 20, pp. 303 - 330. (2006)
10. Aronowitz, H.: Segmental modeling for audio segmentation. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 393 - 396. Honolulu, USA. (2007)
11. Ramabhadran, B., Huang, J., Chaudhari, U., Iyenfgar, G., Nock, H. J.: Impact of audio segmentation and segment clustering on automated transcription accuracy of large spoken archives. In Interspeech, pp. 2589 - 2592. Geneva, Switzerland. (2003)
12. Meinedo, H., Neto, J.: A stream-based audio segmentation, classification and clustering pre-processing system for broadcast news using ANN models. In: Interspeech, pp. 237 - 240. Lisbon, Portugal. (2005)
13. Chen, S., Gopalakrishnan, P.: Speaker, environment and channel change detection and clustering via the Bayesian Information Criterion. In: DARPA Broadcast News Transcription and Understanding Workshop, pp. 127 - 132. Lansdowne, USA. (1998)
14. Cettolo, M., Vescovi, M., Rizzi, R.: Evaluation of BIC-based algorithms for audio segmentation. Computer Speech and Language, 19(2), pp. 147 - 170. (2005).
15. Cheng, S., Wang, H.: METRIC-SEQDAC: A hybrid approach for audio segmentation. International Conference on Spoken Language Processing, pp. 1617 - 1620). Jeju Island, Korea. (2004)
16. Sainz, I., Navas, E., Hernez, I., Bonafonte, A., Campillo, F.: TTS Evaluation Campaign with a Common Spanish Database. In: Seventh International Language Resources and Evaluation (LREC'10), paper 454, pp. 2155-2160, Valletta, Malta. (2010)

17. Campillo-daz, F., Mendez, F., Arza, M., Docio, L., Bonafonte, A., Navas, E., Sainz, I.: Albayzin 2010 : a Spanish text to speech evaluation. In: 12th Annual Conference of the International Speech Communication Association (INTERSPEECH'11), pp. 2161 - 2164, Firenze, Italy. (2011)
18. Zelenák, M., Schulz, H., Hernando, J.: Albayzin 2010 Evaluation Campaign : Speaker Diarization. In: FALA 2010, pp. 301 - 304, Vigo (Spain). (2010)
19. Butko, T., Nadeu, C., Schulz, H.: Albayzin-2010 Audio Segmentation Evaluation Evaluation Setup and Results. In: FALA 2010, pp. 305 - 308, Vigo (Spain). (2010)
20. Butko, T., Nadeu, C.: Audio segmentation of broadcast news in the Albayzin-2010 evaluation: overview, results, and discussion. EURASIP Journal on Audio, Speech, and Music Processing, 2011(1), pp. 1 - 10. (2011)
21. Tavarez, D., Navas, E., Erro,D., Saratxaga,I.: Strategies to Improve a Speaker Diarisation Tool. In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), pp. 4117-4121, Istambul (2012)