Albayzin-2010 Audio Segmentation Evaluation: Evaluation Setup and Results

Taras Butko, Climent Nadeu and Henrik Schulz

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

Abstract

In this paper, we present the audio segmentation task from the Albayzín-2010 evaluation, and the results obtained by the eight participants from Spanish and Portuguese universities. The evaluation task consisted of the segmentation of audio files from the Catalan 3/24 TV channel into 5 acoustic classes: music, speech, speech over music, speech over noise and other. The final results from all participants show that the problem of segmenting broadcast news is still challenging. We also present an analysis of the segmentation errors of the submitted systems. Additionally, the evaluation setup, including the database and the segmentation metric, is also described.

Index Terms: audio segmentation, broadcast news, international evaluation, evaluation setup, database

1. Introduction

The Albayzín evaluation campaign is an internationally-open set of evaluations organized by the Spanish network of speech technologies every 2 years. In the context of Albayzín-2010, an audio segmentation task was proposed by the authors for the first time. It is motivated by the fast increase of audio data, which demands for efficient content-based automatic audio segmentation methods. Recently, researchers have put much effort on this problem due to its applications to tasks as audio indexing and retrieval [1], or automatic transcription of audio recordings [2]. Also, a previous identification of speech segments facilitates speech processing tasks as speech recognition or speaker diarization. Furthermore, audio segmentation is used to make online adaptation of ASR models, or to generate a set of acoustic cues for speech recognition in order to improve the overall system performance [1]. Additionally, multimedia surveillance and monitoring applications can benefit significantly from audio-based event detection [3].

Many research works address the problem of audio segmentation in different scenarios. In [4], the authors propose a method for robust speech, music, environment noise and silence segmentation of audio recorded in different conditions such as TV studio, telephone etc. In [5], the audio stream from broadcast news domain is segmented into 5 different classes, including speech, commercials, environmental sound, physical violence and silence. Content-based retrieval from TV programs is considered in [6], where 7 similar classes are defined.

The final results from 8 participants as well as the evaluation setup, including the database (which is freely available) and the segmentation metric, are described in this paper.

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).[1]

The manual annotation of the database was performed in 2 passes. A first annotation pass segmented the recordings with respect to background sounds (speech, music, noise or none), channel conditions (studio, telephone, outside and none), and speakers as well as speaking modes. A second annotation pass provided literal transcriptions and acoustic events of segments (such as throat, breath, voice, laugh, artic, pause, sound, rustle or noise). For the proposed evaluation we took into account only the first pass of annotation. According to this material, five different audio classes were defined (Table 1).

Table 1: *The five acoustic classes defined for evaluation.*

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

The distribution of the classes within the database is the following: Clean speech: 37%; Music: 5%; Speech over music: 15%; Speech over noise: 40%; Other: 3%. Although 3/24 TV is primarily a Catalan television channel, the recorded broadcasts contain a proportion of roughly 17% of Spanish speech segments. The gender conditioned distribution indicates a clear unbalance in favor of male speech data (63% versus 37%).

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

The database for evaluation 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.

2.2. Metric

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

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

where

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

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

dur(ref$_i$) is 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.

The proposed metric is slightly different from the conventional NIST metric for speaker diarization, where only the total error time is taken into account independently of the acoustic class. Since the distribution of the classes in the database is not uniform, the errors from different classes are weighed differently (depending on the total duration of the class in the database). This way we stimulate the participants to detect well not only the best-represented classes ("speech" and "speech over noise", 77% of total duration), but also the minor classes (like music, 5%).

2.3. Evaluation organization

Ten research groups registered for participation, but only eight submitted segmentation results: **GTTS** (Universidad del País Basco), **GTC-**VIVOLAB (Universidad de Zaragoza), **GSI** (Instituto de Telecomunicações, Universidade de Coimbra, Portugal), **TALP** (Universitat Politècnica de Catalunya) **CEPHIS** (Universitat Autònoma de Barcelona), **ATVS** (Universidad Autónoma de Madrid), **GTM** (Universidade de Vigo), **GTH** (Universidad Politécnica de Madrid / Universidad Carlos III de Madrid).

The database was splitted into 2 parts: for training/development (2/3 of the total amount of data, 16 sessions), and testing (the remaining 1/3, 8 sessions). The training/development audio data together with ground truth labels and evaluation tool were distributed among all the participants by the date of release.

About 3 months were given to all the participants to design their own audio segmentation system. After that period, the testing data was released and 2 weeks were given to perform testing.

Though the evaluation was carried out with the outputs from the primary system submitted by each participant, each site could also submit a contrast (alternative) system. Each evaluated system had to be applied to the whole test database. Each participant site was asked to provide also the total time required to run the set of tests for each submitted system (specifying the used computational resources). The evaluated systems could only use audio signals. Any publicly available data was allowed to be used together with the provided data to train the audio segmentation system. When additional training material is used, the participant was obliged to provide the reference regarding it. Indeed, listening to the test data, or any other human interaction with data, was not allowed before all test results had been submitted.

3. Final results

Table 2 presents the final evaluation scores from the eight participants.

Table 2. *Results of the audio segmentation evaluation.*

Participant	Error rate				
	mu	sp	sm	sn	**Average**
GTH	19.21	39.52	24.97	37.19	**30.22**
GTM	22.41	41.80	27.47	40.93	**33.15**
ATVS	31.01	40.42	33.39	39.80	**36.15**
TALP	26.40	44.20	33.88	41.52	**36.50**
CEPHIS	23.65	45.07	36.95	45.21	**37.72**
GSI	21.43	48.03	51.66	48.49	**42.40**
GTC-VIVOLAB	28.14	51.06	48.78	51.51	**44.87**
GTTS	26.94	52.76	47.75	52.93	**45.09**

Note that the winner of the evaluation (GTH), which obtained the highest average score, also got the highest scores individually for each class. According to the presented results, the "Music" class is the easiest for detection, while "speech" and "speech over noise" are the most difficult.

The distribution of miss and false alarm errors for all participants is presented in Figure 1. According to it, these two types of errors are balanced for "music" and "speech over noise" class, while for "speech" class the false alarm errors are more dominant, and for "speech over music" class the dominant errors are misses.

In Table 3 we present the confusion matrix, which shows the percentage of hypothesized AEs (rows) that are associated to the reference AEs (columns). Data represent averages across the eight audio segmentation systems.

Table 3. *The confusion matrix of acoustic classe*s.

	mu	sp	sm	sn
mu	89.4	0.1	8.0	2.5
sp	0.0	70.6	2.9	26.5
sm	1.8	1.2	87.0	10.0
sn	0.3	10.2	8.3	81.2

According to that confusion matrix, the most common errors are confusions between "Music" and "Speech over music", between "Speech over music" and "Speech over noise" and also between "Speech" and "Speech over noise" classes. Indeed, these classes have very similar acoustic content. Another interesting observation is the low proportion (almost 0%) of confusions between "Speech" and "Music" classes.

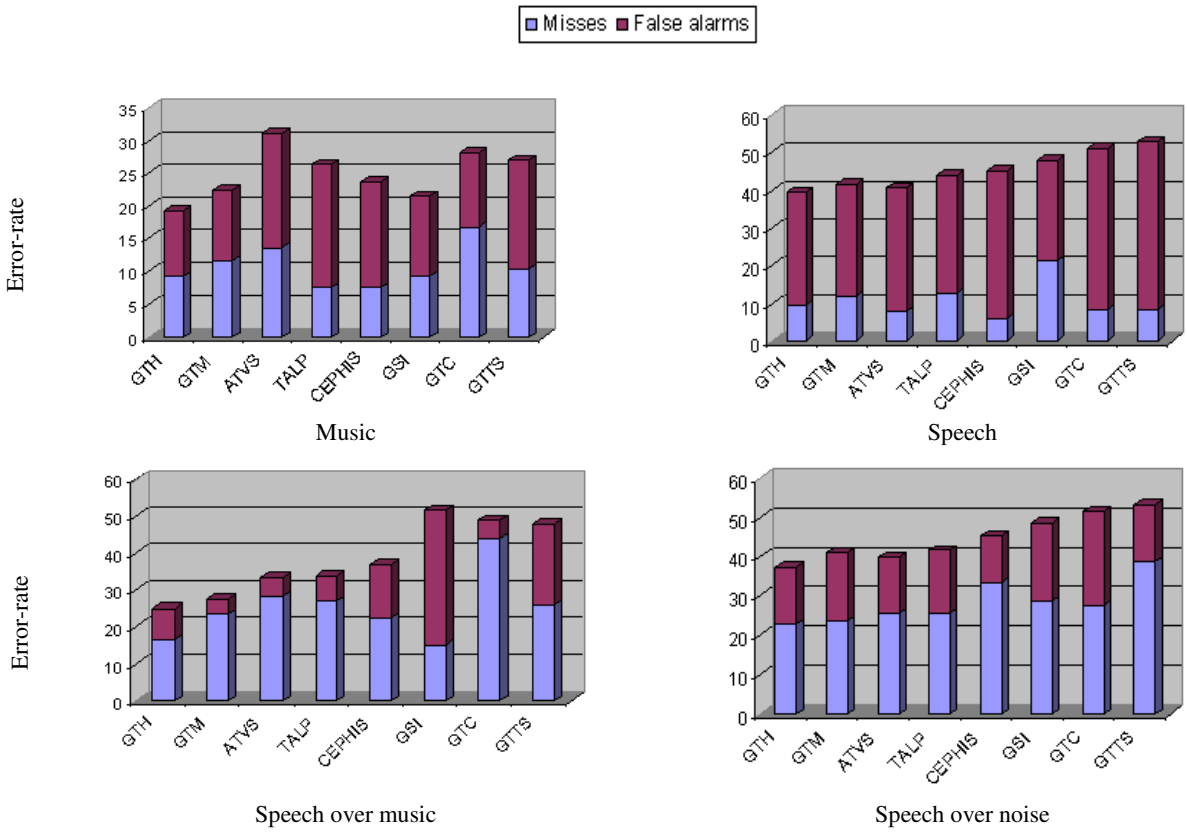

Figure 1: *Distribution of errors for the eight participants, and for each acoustic class.*

In Figure 2 we present the cumulative distribution of errors in terms of duration. Each point *(x, y)* of this plot shows the percentage *y* of total amount of errors with duration less than *x* seconds.

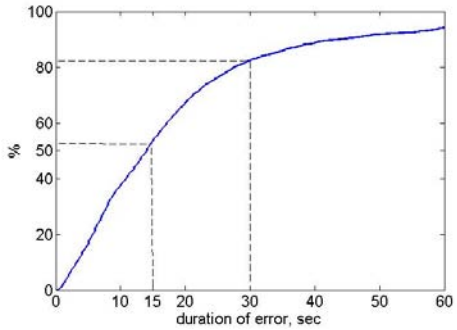

Figure 2: *The cumulative distribution of errors in terms of duration.*

According to that plot, more than half of the errors are shorter than 15 sec and more than 80% of the errors are shorter than 30 sec. In other words, almost each long segment is detected correctly.

In order to measure the difficulty of the proposed audio segmentation task, in Figure 3 we display the proportion of 3 different types of segments in the testing database: *very difficult*, *difficult* and *from winner*. *Very difficult* are those segments on which all 8 audio segmentation systems produced errors (misses or false alarms). *Difficult* segments are those where 7 out of 8 systems produced errors. Finally, *from winner* are those segments with errors produced by the winner system.

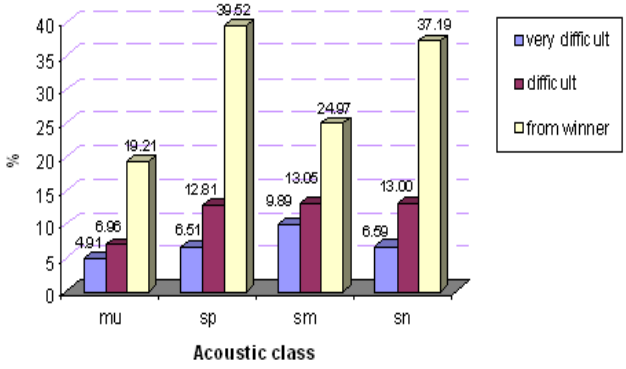

Figure 3: *Comparison of different types of segments in terms of segmentation difficulty.*

According to that plot, only 6.22% of segments are *very difficult*, while the rest of the segments were labeled correctly at least by one participant.

Table 4 and Figure 4 show a grouping of the errors which are common to all the 8 participant segmentation systems. The groups were defined after listening to all the error segments which are *very difficult* and longer than 5 seconds. Seven different semantic groups were distinguished, and the rest were included in *Other*.

Table 4. *Different semantic types of errors which are common to all eigth systems.*

Type of error	*Description*
Type 1	Low level of background sound
Type 2	Speech in background
Type 3	Annotation error
Type 4	The microphone is affected by the wind
Type 5	Singing in background
Type 6	Noise in background is more dominant than music for the [sm] class
Type 7	The quality of music in background is low
Type 8	Other

The percentages of distribution of the above mentioned errors are depicted in Figure 4.

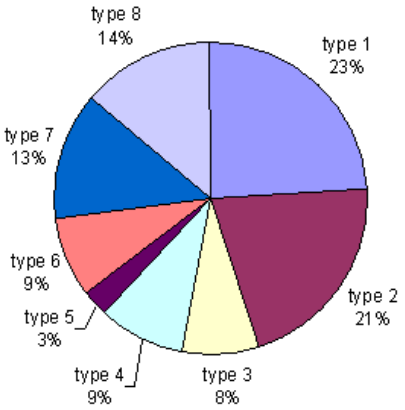

Figure 4: *Percentages of distribution of the different types of common errors.*

According to the plot, a large percentage of common errors are provoked by the presence of either a low level of sound in the background (23%) or overlapped speech (21%), while the annotators' mistakes cause only 8% of the total amount of common errors. Thus the audio segmentation task is still challenging.

4. Conclusions

In this paper, we have presented the submitted results of the Albayzín-2010 audio segmentation evaluation and the evaluation setup, including the database and the segmentation metric.

8 participants from different universities of Spain and Portugal submitted the evaluation results, and the winner got 30.22 % of error-rate in terms of proposed metric. By analyzing the submitted results from all participants we conclude that most of the errors are confusions between "Music" and "Speech over music", "Speech over music" and "Speech over noise", and also between "Speech" and "Speech over noise". Besides, more than half of the total amount of errors is shorter than 15 seconds.

By analyzing the semantic content of errors produced by all submitted systems we found that most of the errors are provoked by the presence of either a low level of sound in the background (23%) or overlapped speech (21%), while the annotators' mistakes cause only 8% of the total amount of common errors.

5. Acknowledgements

The authors wish to thank our colleagues at **GTTS**, **GTC-VIVOLAB**, **GSI**, **CEPHIS**, **ATVS**, **GTM**, **GTH** for their participation in the evaluation. The UPC's 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.

6. References

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

[2] P. C. Woodland, M. J. F. Gales, D. Pye, and S. J. Young, "The development of the 1996 HTK broadcast news transcription system," in DARPA Speech Recognition Workshop, 1997

[3] P. K. Atrey, N. C. Maddage and M. S. Kankanhalli, " Audio based event detection for multimedia surveillance", in Proc. ICASSP, vol. 5, pp. 813-816, 2006

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

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

[6] 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