On building and evaluating a broadcast-news audio segmentation system

Taras Butko and Climent Nadeu

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

Abstract

Audio segmentation is useful in diverse applications like audio indexing and retrieval, subtitling, monitoring of acoustic scenes, etc. Also, an initial audio segmentation stage may help to improve the robustness of speech technologies like automatic speech recognition and speaker diarization. In this paper, firstly, the Albayzín-2010 audio segmentation evaluation is reported, including some conclusions drawn from the analysis of the set of eight submitted systems and their results. Then an audio segmentation system build in agreement with those conclusions is described and tested. Finally, by using the gained experience, the initial design of both the acoustic classes and the detection scoring rules is refined aiming to obtain a more meaningful error rate measurement.

Index Terms: audio segmentation, broadcast news, evaluation

1. Introduction

The recent fast growth of available audio or audiovisual content strongly demands tools for analyzing, indexing, searching and retrieving the available documents. Given an audio document, the first processing step necessarily is audio segmentation (AS), which consists of partitioning the input audio stream into acoustically homogeneous regions, and label them according to a predefined broad set of classes like speech, music, noise, etc.

The research works on AS published so far address the problem in different contexts. The first prominent AS works are dated from 1996, the time when the speech recognition community moved from the newspaper (Wall Street Journal) era towards the broadcast news (BN) challenge [1]. In the BN domain the speech data exhibited considerable diversity, ranging from clean studio to really noisy speech interspersed with music, commercials, sports etc. This time the decision was made to disregard the challenge of transcribing speech in sports material and commercials. The work from [2] and then from [3] are the earliest works that tackled the problem of speech/music discrimination from radio stations. The authors found the first applications of AS in automatic program monitoring of FM stations, and in improvement of performance of ASR technologies, respectively. Both works showed relatively low segmentation error rates (around 2-5%).

Within the next years the research interest was oriented towards the recognition of a broader set of acoustic classes, like in [4] or [5] where, in addition to speech and music classes, the environment sounds were also taken into consideration. A wider diversity of music genres was considered in [6]. The authors in [7] tried to categorize the audio into mixed class types such as music with speech, speech with background noise, etc. The reported classification accuracy was over 80%. A similar problem was tackled by [8] and [9], dealing with the overlapped segments that naturally appear in the real-world multimedia domain and cause high error rates.

In the BN domain, where speech is typically interspersed with music, background noise, and other specific acoustic events, AS is primarily required for indexing, subtitling and retrieval. However, speech technologies that work on such type of data can also benefit from the acoustic segmentation output in terms of overall performance. In particular, the acoustic models used in automatic speech recognition or speaker diarization can be trained for specific acoustic conditions, such as clean studio vs. noisy outdoor speech, or high quality wide bandwidth studio vs. low quality narrow-band telephone speech. Also, AS may improve the efficiency of low bit-rate audio coders, as it allows that traditionally separated speech and music codec designs can be merged in a universal coding scheme which keeps the reproduction quality of both speech and music [10].

Taking into account the increasing interest in the problem of AS on the one hand, and the existence, on the other hand, of a rich variety of feature extraction approaches and classification methods, in 2010 we organized an international evaluation of BN audio segmentation in the context of the Albayzín-2010 campaign. The Albayzín evaluation campaign is an internationally-open set of evaluations organized by the Spanish Network of Speech Technologies (RTH) every 2 years.

NIST evaluation campaigns are a reference. In the NIST RT evaluation plan document [11] the diarization task is defined as "the process of annotating an input audio channel with information that attributes (possibly overlapping) temporal regions of signal energy to their specific sources". However, in NIST evaluations, only speaker diarization is addressed. For AS, several modifications to the speaker diarization evaluation plan have to be made, in particular, the definition and the way of evaluating the acoustic classes.

According to the results from the evaluation [12], the single main factor that accounts for most of the observed segmentation errors is the high degree of confusions between several acoustic classes due to the fact that they exhibit similar acoustic content. Those errors can indeed be partially removed using better acoustic models. An alternative approach may consist of redefining the acoustic classes in such a way that the overlap between different classes is minimal, while keeping the meaningfulness of the classes. We follow both of these approaches in our work presented here.

In the paper, we first describe the database and the evaluation protocol of the AS Albayzin-2010 evaluation as well as the main conclusions we have reached after analyzing both systems and results. Then we create a reference AS system that combines the best characteristics from the proposed systems. And, finally, we propose several changes to the definition and evaluation of the acoustic classes and see how the proposed changes affect the evaluation results.

2. Overview of the audio segmentation Albayzin-2010 evaluation

2.1. Experimental setup and metric

For the Albayzin-2010 evaluation, we used a BN audio database recorded from the 3/24 Catalan TV, and defined 5 acoustic classes (AC) as described in Table 1. The AC "Other" is not evaluated in the final tests.

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

Class	*Description*
Speech [sp]	Clean speech 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.
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
Other [ot]	This class refers to any type of audio signal (including noises) that doesn't correspond to the other four classes

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

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

where

$dur(miss_i)$ is the total duration of all deletion errors (misses) for the ith AC,

$dur(fa_i)$ is the total duration of all insertion errors (false alarms) for the ith AC, and

$dur(ref_i)$ is the total duration of all the ith AC instances according to the reference file.

An incorrectly classified audio segment (a substitution) is computed both as a deletion error for one AC and an insertion error for another.

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 AC. 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%).

2.2. Overview of the AS systems and results

Eight participants from Spanish and Portuguese universities competed in the evaluation [13]. By analyzing both the submitted AS systems and their corresponding segmentation results, several observations can be extracted which are outlined in the following.

1. *The conventional use of automatic speech recognition features for the AS task* (like MFCC, PLP or FF (frequency-filtered log-filter-bank energies)). Nevertheless, the best system additionally exploited chroma and spectral entropy features that showed to be useful. The systems that used other perceptual feature sets, like zero-crossing rate, spectral width etc, could not report significant improvement.

2. *The systems that used segment-based features outperformed the systems with frame-based features.*

The best two AS systems parameterized the audio signal using segment-based features. The best system used the mean and variance along a 1 sec segment; the second best system used a super-vector approach to parameterize even longer segments [13]. Presumably, this is the main reason for their superior detection rates. It may indicate that the models trained on frame-based features do not capture the structure of the acoustic classes sufficiently.

3. The majority of the AS systems used the HMM approach

The main advantage of the HMM approach is that it performs segmentation and classification jointly. Other alternatives like detection-and-classification or detection-by-classification require two independent steps to be carried out one after the other, so that the errors produced in the first step may propagate to the next one.

4. The hierarchical detection approach seems to be effective

Four research groups reported an improvement when using a hierarchical organization of the detection process. Those AS systems detect the easiest classes ([mu] and silence, which is included in [ot]) at the early steps, while a further discrimination among the rest of the classes is done on subsequent steps. In this type of architecture, it is not necessary to have the same classifier, feature set and/or topology for the various individual detectors.

5. Challenge of the AS task

Only 23% of errors produced by the best AS system were also produced by all the other AS systems. This indicates that there is still a large margin for improvement of the segmentation results. Taking into account that the main source of errors are confusions between [mu] and [sm], between [sm] and [sn], and also between [sp] and [sn], future research efforts should be devoted to improve detection of the background sounds.

3. Building a reference AS system

Inspired by the systems from the Albayzín-2010 evaluation and their results, we constructed a reference AS system that combines the best characteristics from the submitted systems.

We use the HMMs with 1 emitting state and 256 GMM mixture components to model each individual class. The one-against-all detection strategy is employed: 5 binary detectors are organized in a hierarchical way, as depicted in Figure 1.

One of the most important decisions when using this kind of architecture is to put the detectors in the best order in terms of information flow, since, for instance, a given detector may benefit greatly from the previous detection of the classes that show high confusion with the class under detection.

Each detector uses a different feature set, the one that showed the best accuracy for detection of the corresponding class using a cross-validation procedure. The following sets of features were considered:

- ASR features. 16 FF coefficients [14] with the first time derivatives computed in 20ms-long frames. Then mean and variance are computed over a 1 sec window.
- 12 chroma features [15] are computed in frames 50ms long. Then mean and variance are computed over a 1 sec window.
- 5 energy statistics features [16] are computed over a 10 sec window with 1 sec shift. Those features,

which are obtained from the amplitude histogram of the audio signal, were not used in the evaluation campaign by any of the participants. But, since we observed a high amount of confusions between the classes [sp] and [sn], we conjectured that they may improve the overall accuracy.

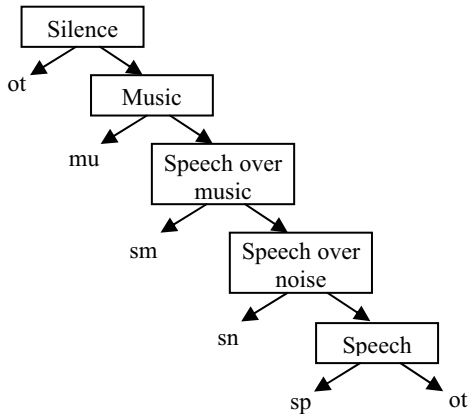

Figure 1: *Flow diagram of the hierarchical architecture.*

Using different combinations of the proposed features we found the best feature set for each detector separately. The trade-off between misses and false alarms is also optimized for each class independently. The detectors with best individual accuracies are placed in the early stages to facilitate the subsequent detection of the classes with worst individual accuracies. The segmentation results are presented in Table 2.

Table 2: *Error rates from different binary detectors using different feature sets.*

	FF	FF + Chroma	FF + Stat	FF + Stat + Chroma
Music	17.65	18.34	18.51	**17.52**
Speech	**39.67**	40.52	41.81	45.26
Speech over music	30.71	28.35	26.80	**23.00**
Speech over noise	41.93	**40.77**	43.14	44.00

Given the results from Table 2, we selected FF features for detecting [sp], FF + Chroma features for detecting [sn], and, finally, "FF + Statistical + Chroma" features for detecting both [mu] and [sm] classes.

4. Changes in the definition and scoring of the acoustic classes

The transcription of the database used for evaluations was performed according to TC-STAR European Parliament Plenary Session Transcription Guidelines [17]. Apart from speech transcription, the annotations include 3 different layers:
Speaker turn layer: None, Studio speaker, Outdoor speaker.
Background conditions: none, music, noise, speech, speech + music, speech + noise, noise + music, speech + noise + music.
Non-speech acoustic events: any short time non-speech sound like laugh, throat, knocking, etc.

Since the non-speech acoustic events affect just short portions of audio, the corresponding layer was discarded from the definition of the acoustic classes.

The distribution of the ACs in the database is presented in Table 3. The first number in each cell corresponds to the percentage of the corresponding AC in the whole database, the second one shows the percentage of the errors that the corresponding class provokes in the testing part of the database (the reference AS system is used to compute the error distribution). In the case when the first number is higher than the second one, the AC is considered easy for detection, in the opposite case it is considered difficult. From the table we observe that the classes *clean outdoor speech*, *studio speech in noise* and also *studio/outdoor speech with speech in background* are difficult for detection. In the table we also show how the classes were grouped for the Albayzin-2010 evaluation.

Table 3: *Distribution of the acoustic classes in the database.*

			Speaker turn			
			None	Studio	Outdoor	
Background conditions	None (clean conditions)	**OT**	3.85/ **-**	18.18/ **7.56**	18.55/ **26.34**	**SP**
	Music	**MU**	5.70/ **3.30**	11.21/ **9.99**	3.30/ **4.43**	**SM**
	Music+ Speech		0.68/ **0.00**	0.17/ **0.04**	0.30/ **0.62**	
	Music+ Noise		0.05/ **0.03**	1.07/ **0.72**	0.36/ **0.60**	
	Speech	**OT**	0.37/ **0.01**	1.27/ **1.99**	2.12/ **3.60**	**SN**
	Noise		0.01/ **0.02**	10.03/ **17.60**	21.90/ **22.93**	
	Noise+ speech		0.01/ **0.00**	0.20/ **0.01**	0.66/ **0.19**	

In the following we propose several alternatives to the initial design of the acoustic classes and the way they are evaluated, and report how these changes affect the segmentation results.

Refinement 1:

Include the *speech with speech in background* segments into the [sp] class. In fact, many times the overlapped speech appears when there is a synchronous translation. Although the amount of overlapped speech is not high (1.27 + 2.12 = 3.39% of total amount of data), the proposed refinement may partially remove confusion errors between [sp] and [sn].

Refinement 2:

We propose to make some refinement in the evaluation of the classes [sp] and [sn]. In fact, in Table 3 we see a clear unbalance between the two numbers in the *studio speech in the noise* cell, and in the *clean outdoor speech* cell. Indeed, *studio speech in the noise* is acoustically similar to the [sp] class, and conversely, *clean outdoor speech* is similar to the [sn] class. For the segments that are labeled as *studio speech in the noise* and *clean outdoor speech* we propose to assume that both hypothesis labels [sp] and [sn] are correct.

Refinement 3:

In the Albayzin-2010 evaluation we considered all confusion errors equally weighted in the metric. But, for instance, it seems reasonable to weight less the confusion between [mu] and [sm] than between [mu] and [sp]. In principle, the idea is to penalize confusion errors between ACs that have similar acoustic content less than confusion errors

between classes with very different acoustic content. The set of proposed weights is displayed in Table 4.

Table 4: *Weights for the different types of confusion errors.*

reference \ hypothesis	mu	sm	sn	sp
sp	1	0.5	0.5	0
sn	1	0.5	0	0.5
sm	0.5	0	0.5	0.5
mu	0	0.5	1	1

Refinement 4:

Although the use of single layer segments is practically convenient, we could also define the task in terms of a multiple layer segmentation. For instance, we could define the task of segmenting audio into 3 possibly overlapped ACs: "Speech", "Music" and "Noise". In that case the classes are acoustically different and could mutually overlap so there is no need to apply refinements 2 and 3.

Table 5: *Error rates with the proposed refinements.*

	mu	sp	sm	sn	Average
Baseline	17.70	37.42	22.80	38.93	**29.21**
Refinement 1	17.31	42.32	23.40	48.20	**32.82**
Refinement 2	17.31	34.27	23.40	31.84	**26.71**
Refinement 3	13.61	22.38	14.04	23.92	**18.49**

	Speech	Music	Noise	Average
Refinement 4	6.6	69.5	82.3	**52.8**

Table 5 shows the AS results with the proposed refinements. The main conclusions are:

1. Inclusion of *speech with speech in background* into the [sp] class increases the error rate of the reference AS system. Since in our database background speech is usually bubble noise, it is more appropriate to include the overlapped speech into [sn].

2. The two modifications related to the way of evaluating the ACs indeed decrease the error rate of the reference AS system. The main benefit from the proposed modifications is obtaining a more meaningful error rate measurement. For instance, while confusion errors were counted twice in the initial metric, one as deletion and the other as insertion, since the class "Other" [ot] is not evaluated in the final tests, the confusion errors with this class were counted just once for the remaining ACs. Therefore, there were confusion errors between semantically different ACs which were implicitly weighted less than other equally important errors. Conversely, with the proposed changes, we explicitly de-weight the confusion errors between semantically similar acoustic classes.

3. The definition of the three new ACs, "Speech", "Music" and "Noise", that can mutually overlap, to replace the five previously defined ACs, leads to very low recognition results. Presumably, the main reason for such behavior is the high proportion of audio segments that belong simultaneously to different ACs.

5. Conclusions

In this paper we have first described the database, the evaluation protocol of the AS Albayzin-2010 evaluation, and the main conclusions from the submitted systems and results. A reference AS system that combines the best characteristics from those systems is proposed and its results reported. Finally, several changes to the definition and evaluation of the acoustic classes are proposed, and their influence on the evaluation results is described. We conclude that the two class definition modifications yield worse results, while the changes in the evaluation metric offer a more meaningful error rate measurement.

6. Acknowledgements

This work has been funded by the Spanish project SARAI (TEC2010-21040-C02-01). The authors wish to thank our colleagues at *ATVS*, *CEPHIS, GSI*, *GTC-VIVOLAB*, *GTH*, *GTM*, and *GTTS* for their enthusiastic participation in the evaluation. Also, the authors are very grateful to Pedro Vizarreta for his contribution to the experimental part of the work, and to Henrik Schulz for managing the collection of the database and helping for its annotation. The first author is partially supported by a grant from the Catalan autonomous government.

7. References

[1] D. S. Pallet, "A look at NIST's benchmark ASR tests: Past, present, and future," Technical Report, National Institute of Standards and Technology (NIST), USA, 2003.

[2] J. Saunders, "Real-time discrimination of broadcast speech/music," in Proc. IEEE ICASSP, v. 2, pp. 993-996, 1996.

[3] E. Scheirer, M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," in Proc. IEEE ICASSP, 1997.

[4] T. Zhang, C.-C. Kuo, "Hierarchical classification of audio data for archiving and retrieving," in Proceedings IEEE ICASSP, vol. 6, pp. 3001–3004, 1999.

[5] L. Lu, H.-J. Zhang, H. Jiang, "Content analysis for audio classification and segmentation," in IEEE Transactions on Speech and Audio Processing, v. 10, no. 7, pp. 504–516, 2002.

[6] K. El-Maleh, M. Klein, G. Petrucci, P. Kabal, "Speech/music discrimination for multimedia applications," in Proc. IEEE ICASSP, v. 6, pp. 2445–2448, 2000.

[7] S. Srinivasan, D. Petkovic, D. Ponceleon, "Toward robust features for classifying audio in the cue video system," in Proc. 7th ACM Int. Conference on Multimedia, pp. 393–400, 1999.

[8] A. Bugatti, A. Flammini, P. Migliorati, "Audio classification in speech and music: a comparison between a statistical and a neural approach," EURASIP Journal on Applied Signal Processing, vol. 2002, no. 4, pp. 372–378, 2002.

[9] J. Ajmera, I. McCowan, H. Bourlard, "Speech/music segmentation using entropy and dynamism features in a HMM classification framework," Speech Communication, v. 40, no. 3, pp. 351–363, 2003.

[10] M. Exposito, G. Galan, R. Reyes, V. Candeas, "Audio coding improvement using evolutionary speech/music discrimination," in Proc. IEEE Conference on Fuzzy Systems, pp. 1-6, 2007.

[11] NIST. (2009) The NIST Rich Transcription evaluation project website. http://www.itl.nist.gov/iad/mig/tests/rt/

[12] T. Butko, C. Nadeu and H. Schulz, "Albayzin-2010 Audio Segmentation Evaluation: Evaluation Setup and Results", SLTech Workshop, Vigo, Spain, November, 2010

[13] http://fala2010.uvigo.es

[14] C. Nadeu, D. Macho, J. Hernando, "Frequency and time filtering of filter-bank energies for robust HMM speech recognition", Speech Communication, vol. 34, pp. 93-114, 2001.

[15] T. Fujishima, "Realtime chord recognition of musical sound: a system using common lisp music", Proc. of the Int. Computer Music Conference (ICMC), pp. 464–467, 1999.

[16] M. Buchler, S. Allegro, S. Launer, N. Dillier "Sound Classification in Hearing Aids Inspired by Auditory Scene Analysis", EURASIP J. on Appl. Sig. Proc, pp. 2991-3002, 2005.

[17] http://www.tc-star.org/.