EURASIP Journal on
Audio, Speech, and Music Processing
a SpringerOpen Journal

RESEARCH Open Access

Audio segmentation of broadcast news in the Albayzin-2010 evaluation: overview, results, and discussion

Taras Butko* and Climent Nadeu

Abstract

Recently, audio segmentation has attracted research interest because of its usefulness in several applications like audio indexing and retrieval, subtitling, monitoring of acoustic scenes, etc. Moreover, a previous audio segmentation stage may be useful to improve the robustness of speech technologies like automatic speech recognition and speaker diarization. In this article, we present the evaluation of broadcast news audio segmentation systems carried out in the context of the Albayzín-2010 evaluation campaign. That evaluation consisted of segmenting audio from the 3/24 Catalan TV channel into five acoustic classes: music, speech, speech over music, speech over noise, and the other. The evaluation results displayed the difficulty of this segmentation task. In this article, after presenting the database and metric, as well as the feature extraction methods and segmentation techniques used by the submitted systems, the experimental results are analyzed and compared, with the aim of gaining an insight into the proposed solutions, and looking for directions which are promising.

Keywords: Audio segmentation, Broadcast news, International evaluation

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 necessary, first processing step is audio segmentation, 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 studies on audio segmentation published so far have addressed the problem in different contexts. The first prominent audio segmentation studies began in 1996, the time when the speech recognition community moved from the newspaper (Wall Street Journal) era toward 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 was the time when the decision was made to disregard the challenge of transcribing speech in sports material and commercials. The earliest studies that tackled the problem of speech/music discrimination from radio stations are those of [2,3]. Those authors found the first applications of audio segmentation in automatic program monitoring of FM stations, and in the improvement of performance of ASR technologies, respectively. Both studies showed relatively low segmentation error rates (around 2-5%).

After those studies, the research interest was oriented toward the recognition of a broader set of acoustic classes (AC), such as in [4,5] wherein, 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]. Conventional approaches for speech/music discrimination can provide reasonable performance with regular music signals, but often fail to perform satisfactorily with singing segments. This challenging problem was considered in [7]. The authors in [8] 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 Bugatti et al. [9] and Ajmera et al. [10], dealing with the overlapped

* Correspondence: taras.butko@upc.edu
Department of Signal Theory and Communications, TALP Research Center, Universitat Politècnica de Catalunya, Barcelona, Spain

segments that naturally appear in the real-world multimedia domain and cause high error rates. The problem of audio segmentation was implicitly considered in the context of a meeting-room acoustic event detection task in two international evaluations: CLEAR 2006 and CLEAR 2007. The latter evaluation showed that the overlapping segments accounted for more than 70% of errors produced by every submitted system. Despite the interest shown in mixed sound detection in the recent years [11-13], it still remains a challenging problem.

In the BN domain, where speech is typically interspersed with music, background noise and other specific acoustic events, audio segmentation 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 (ASR) or speaker diarization can be trained for specific acoustic conditions, such as clean studio versus noisy outdoor speech, or high-quality wide-bandwidth studio versus low-quality narrow-bandwidth telephone speech. Also, audio segmentation may improve the efficiency of low bit-rate audio coders, as it allows for merging the traditionally separated speech and the music codec designs into a universal coding scheme, which keeps the reproduction quality of both speech and music [14].

Different techniques for audio segmentation are proposed in state-of-the-art literature. They mainly differ in either the feature extraction methods or the classification approaches. We can distinguish two main groups of features: frame-based and segment-based features. The frame-based features usually describe the spectrum of the signal within a short time period (10-30 ms), where the process is considered stationary. MFCCs and PLPs are examples of frame-based features routinely used in speech recognition [15], which represent the spectral envelope and also its temporal evolution. Some studies, such as [3], propose other types of features for audio segmentation: spectral roll-off point, spectral centroid, spectral flux, zeros-crossing rate, etc. Often, both types of features are also used in combination [16].

For segment-based feature extraction, usually a longer segment is taken into consideration. The length of the segment may be fixed (usually 0.5-5 s) or variable. Although fixing the segment size brings practical implementation advantages, the performance of a segmentation system may suffer from either the possibly high resolution required by the content or the lack of sufficient statistics needed to estimate the segment features because of the limited time span of the segment. According to [17], a solution with greater efficiency would be to extract global segments within which the content is kept stationary so that the classification method can achieve an optimum performance within the segment. The most usual segment-based features are the first- and second-order statistics of the frame-based features computed along the whole segment. Sometimes, high-order statistics are taken into consideration, like skewness and kurtosis, as well as more complex feature combinations that capture the dynamics of audio (e.g., the percentage of frames showing less-than-average energy), rhythm (e.g., periodicity from the onset detection curve), timbre, or harmonicity of the segment [18].

Audio segmentation can be performed in three different ways. The first one is based on detecting the sound boundaries and then classifying each end-pointed segment. Hereafter, we refer to it as the *detection-and-classification* approach. For example, in [19], an approach based upon exploration of relative silences has been proposed; a relative silence is considered as a pause between important foreground sounds. A different type of segmentation algorithm, which does not require any *a priori* information about the particular AC, is based on the BIC [20]. It assumes that the sequence of acoustic feature vectors is a Gaussian process, and measures the likelihood that two consecutive acoustic frames were generated by two processes rather than a single process.

The second approach consists of classifying consecutive fixed-length audio segments. We will refer to it as the *detection-by-classification* approach. A raw segmentation output is obtained in this case as a direct byproduct of the sequence of segment labels given by the classifier. However, to improve the segmentation (detection) accuracy, some kind of smoothing is required, under the assumption that a sudden or frequent change of sound types in an arbitrary way is unlikely. Many publications give preference to this second approach because of its natural simplicity. As an example, Saunders [2] used a multivariate Gaussian classifier to obtain a sequence of decisions, Lu et al. [5] applied a KNN-based classifier, and Bugatti et al. [9] used an MLP-based classifier in the experiments.

In the third approach, classification and segmentation are done jointly. For instance, in its decoding step, the HMM-based method attempts to find the state sequence (and, consequently, the AC sequence) with the highest likelihood given a sequence of observed feature vectors. The most common procedure for doing that is by Viterbi decoding, i.e., using a dynamic programming algorithm to find in a recursive a manner the most probable sequence of HMM states. The HMM-based audio segmentation approach borrowed from speech/speaker recognition applications has been successfully applied in [4,10,13] and many other studies.

Taking into account the increasing interest in the problem of audio segmentation, on the one hand, and the existence, on the other hand, of a rich variety of feature

extraction approaches and classification methods, 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. Actually, the quantitative comparison and evaluation of competing approaches is very important in nearly every research and engineering problem. The evaluation campaigns that independently compare systems from different research groups help us to determine which directions are promising and which are not [1].

For the proposed evaluation, we used a BN audio database recorded from the 3/24 Catalan TV, and defined five AC: "Music," "Speech," "Speech over music," "Speech over noise," and "Other." In the rest of the article after presenting the database and metric, we describe the different feature extraction methods, the segmentation techniques, and the organization ways of the segmentation process proposed by the eight groups that submitted their results to the evaluation. We also compare the various segmentation systems and results, to gain an insight into the proposed solutions. Section "Database and metric" gives an overview of the database and metrics used in evaluation. In Section "Participating groups and methods", a short description of the methods that were applied by the individual groups is given. The results of the evaluation are presented and discussed in Sections "Results" and "Discussion." Finally, this article concludes with conclusion section.

Database and metric

The database used for the evaluations consists of BN audio from the 3/24 Catalan TV channel, which 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 24 files of approximately 4-h duration each, and a total duration of approximately 87 h of annotated audio.[a] The manual annotation of the database was performed in two passes. The first annotation pass segmented the recordings with respect to background sounds (speech, music, noise, or none), channel conditions (studio, telephone, outside, and none), speakers, and speaking modes. The second annotation pass provided speech transcriptions and acoustic events (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, a set of five different audio classes was defined (Table 1), which includes overlapping of speech with either music or noise.

The distribution of the classes within the database is the following: "Speech": 37%; "Music": 5%; "Speech over music": 15%; "Speech over noise": 40%; and "Other": 3%.

Table 1 The five acoustic classes defined for evaluation

Class	Description
Speech [sp]	Clean speech from a close microphone without any kind of background sound
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 silence and noises) that does not correspond to the other four classes

The class "Other" is not evaluated in the final tests. Although 3/24 TV is primarily a Catalan-spoken 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 vs. 37%). The audio signals are provided in pcm format, mono, 16 bit resolution, and 16-kHz sampling frequency.

The metric is defined as a relative error averaged over all the AC:

$$\text{Error} = \underset{i}{\text{average}} \left(\frac{\text{dur}(\text{miss}_i) + \text{dur}(\text{fa}_i)}{\text{dur}(\text{ref}_i)} \right) \quad (1)$$

where $\text{dur}(\text{miss}_i)$ is the total duration of all deletion errors (misses) for the ith AC, $\text{dur}(\text{fa}_i)$ is the total duration of all insertion errors (false alarms) for the ith AC, and $\text{dur}(\text{ref}_i)$ 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. A forgiveness collar of 1 s (both + and -) is not scored around each reference boundary. This accounts for both the inconsistencies of 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 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%).

The database was split into two parts: 2/3 of the total amount of data, i.e., 16 sessions, for training/development, and the remaining 1/3, i.e., 8 sessions, for testing.

The training/development audio data together with the ground truth labels and the evaluation tool were distributed among all the participants by the date of release.

The evaluated systems should only use audio signals. Any publicly available data were allowed to be used together with the provided data to train the audio segmentation system. When additional training material was used, the participant was obliged to provide the reference regarding it. Listening to the test data, or any other human interaction with data, was not allowed before the test results were submitted by all the participants.

Participating groups and methods

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

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

In the following, the systems presented by the participant groups are briefly described. The systems are listed in the order in which they are ranked in the table of final results. The full description of the systems can be found in FALA 2010 conference proceedings [21].

System 1

Features: segment-based. First, 15 MFCCs, the frame energy, and their first and second derivatives (delta and delta-delta) are extracted. In addition, the spectral entropy and the CHROMA coefficients are calculated. Second, the mean and variance of these features are computed over 1-s interval.

Segmentation approach: HMM-based.

The acoustic modeling is performed using five HMMs with three emitting states and 256 Gaussians per state. Each HMM corresponds to one acoustic class. An hierarchical organization of binary HMM detectors is used. First, audio is segmented into "Music"/"non-Music" portions. Second, the "non-Music" portions are further segmented into "Speech over music"/"non-Speech over music" portions. Finally, the "non-Speech over music" portions are segmented into "Speech"/"Speech over noise."

System 2

Features: segment-based. First, 13 MFCCs including the zero (energy) coefficient and their first and second derivatives (delta and delta-delta) are extracted. Second, a background model based on GMM (GMM-UBM) of M mixture components is trained using data from all classes. Then, given an audio segment represented by N feature vectors of dimension D, the GMM-UBM is adapted to that audio segment using MAP adaptation. By stacking the resulting means, a supervector of dimension $M{\cdot}D$ is obtained.

Segmentation approach: detection-and-classification.

The BIC algorithm is used in the detection of the segment boundaries. The classification of each segment is performed using support vector machines.

System 3

Features: frame-based 7 MFCCs plus shifted delta coefficients (SDC).

Segmentation approach: HMM-based.

The acoustic modeling is performed using a five-state HMM with full connected state transitions. Each state corresponds to one AC modeled by GMM with 1024 mixtures. Given a vector of observations, the Viterbi decoding algorithm is applied to obtain a sequence of HMM states. A *mode* filter (i.e., a filter that replaces a current state with *mode* of its neighboring states) is applied to avoid spurious changes between states.

System 4

Features: frame-based 16 frequency-filtered (FF) log filter-bank energies with their first time derivatives. Mean subtraction is applied at the segment level. A wrapper-based feature selection technique is used for finding the most discriminative features for each AC individually.

Segmentation approach: HMM-based.

The acoustic modeling is performed using five HMMs with one emitting state and 64 Gaussians per state. Each HMM corresponds to one acoustic class. A hierarchical organization of binary HMM detectors is used. First, the audio stream is pre-segmented using a silence detector. Then non-silence portions are segmented into "Music"/"non-Music"; the "non-Music" portions are further segmented into "Speech over music"/"non-Speech over music"; the "non-Speech over music" portions are further segmented into "Speech over noise"/"non-Speech over noise"; and, finally, the "non-Speech over noise" portions are segmented into "Speech"/"Other."

System 5

Features: frame-based 12 PLPs plus local energy and their first and second derivatives (delta and delta-delta).

Segmentation approach: HMM-based.

The acoustic modeling is performed using five HMMs with one emitting state and 64 Gaussians per state. Each HMM corresponds to one AC.

System 6

Features: frame-based 16 MFCCs including zero (energy) coefficient, plus eight perceptual coefficients (e.g., zero-crossing rate, spectral centroid, spectral roll-off, etc.) and their first-time derivatives.

Segmentation approach: mixed, detection-by-classification, and HMM-based.

An hierarchical organization of the detection process is used. First, silence and music are located using a repetition detector system based on fingerprinting (detection-by-classification). In the proposed fingerprinting system, a 32-bit binary pattern is computed for each frame of about 200 ms; spectral analysis is performed with a mel-scaled filter-bank with 32 channels, and the resulting spectrogram is binarized into a 32-bit pattern, choosing 1, essentially, when there is a spectral peak. The detection strategy consists in counting the number of matching bits between the signature and the audio binary patterns in each frame, and when this number is above a threshold, an acoustic class is detected. Second, a hybrid HMM/MLP segmentation is applied to the audio segments which are not classified as either music or silence. Each AC is modeled via a 10-state HMM with left-to-right state transitions.

System 7

Features: frame-based 13 MFCCs plus their first and second derivatives (delta and delta-delta). In addition, the mean, the variance, and the skewness of the first MFCC are calculated.

Segmentation approach: detection-and-classification.

The BIC algorithm is used to detect the segment boundaries. Classification is performed with a hierarchical organization of detectors and using GMMs combined with a binary decision tree. First, the audio stream, which is pre-segmented with a silence detector, is classified into "Music"/"non-Music" segments; and the "non-Music" ones are further classified into "Speech over music"/"Speech"/"Speech over noise."

System 8

Features: frame-based 13 MFCCs including zero (energy) coefficient. Cepstral mean subtraction was not applied.

Segmentation approach: detection-by-classification.

Each class is modeled by a GMM with 1024 mixtures. For each frame, the class yielding the highest likelihood is chosen. A *mode* filter is applied to smooth the decisions along time.

Results

Table 2 presents the final scores from the eight systems. The error rate is presented for each evaluated class individually, together with the average score over all the evaluated classes. It is noted that no participant was using any additional data for training the acoustic models apart from the data provided for the evaluation.

As can be observed in Table 2, "Music" is the best-detected class among all the systems. The system that obtained the best average score (30.22%), system 1, also got the highest score individually for each class.

The distribution of the miss and the false alarm errors from all the systems is presented in Figure 1. This plot shows a clear unbalance between misses and false alarms for the classes "Speech" and "Speech over music."

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

According to the confusion matrix, the most common errors are the confusions between "Music" and "Speech over music," between "Speech over music" and "Speech over noise," and also between "Speech" and "Speech over noise." Indeed, the two components of each of those pairs of classes have very similar acoustic content. Another interesting observation is the low proportion (almost 0%) of confusions between "Speech" and "Music." The second row of the confusion matrix indicates that 26.5% of the hypothesized speech is in fact "Speech over noise." This is the main reason of the high proportion of false alarms for the class "Speech" (Figure 1b). Actually, for many "Speech over noise" audio segments the level of noise in background is extremely low so that the detection systems usually confuse "Speech over noise" with "Speech."

In Figure 2, we present cumulative distributions of duration of testing segments. The solid curve corresponds to the segments incorrectly detected by the audio segmentation systems for the whole set of participants. The dashed curve corresponds to the cumulative distribution of the ground truth segments. Each point (x, y) of this plot shows the percentage y of segments with duration less than x seconds.

Table 2 Results of the audio segmentation evaluation

	Error rate				
systems	mu	sp	sm	sn	Average
1	19.21	39.52	24.97	37.19	30.22
2	22.41	41.80	27.47	40.93	33.15
3	31.01	40.42	33.39	39.80	36.15
4	26.40	44.20	33.88	41.52	36.50
5	23.65	45.07	36.95	45.21	37.72
6	21.43	48.03	51.66	48.49	42.40
7	28.14	51.06	48.78	51.51	44.87
8	26.94	52.76	47.75	52.93	45.09

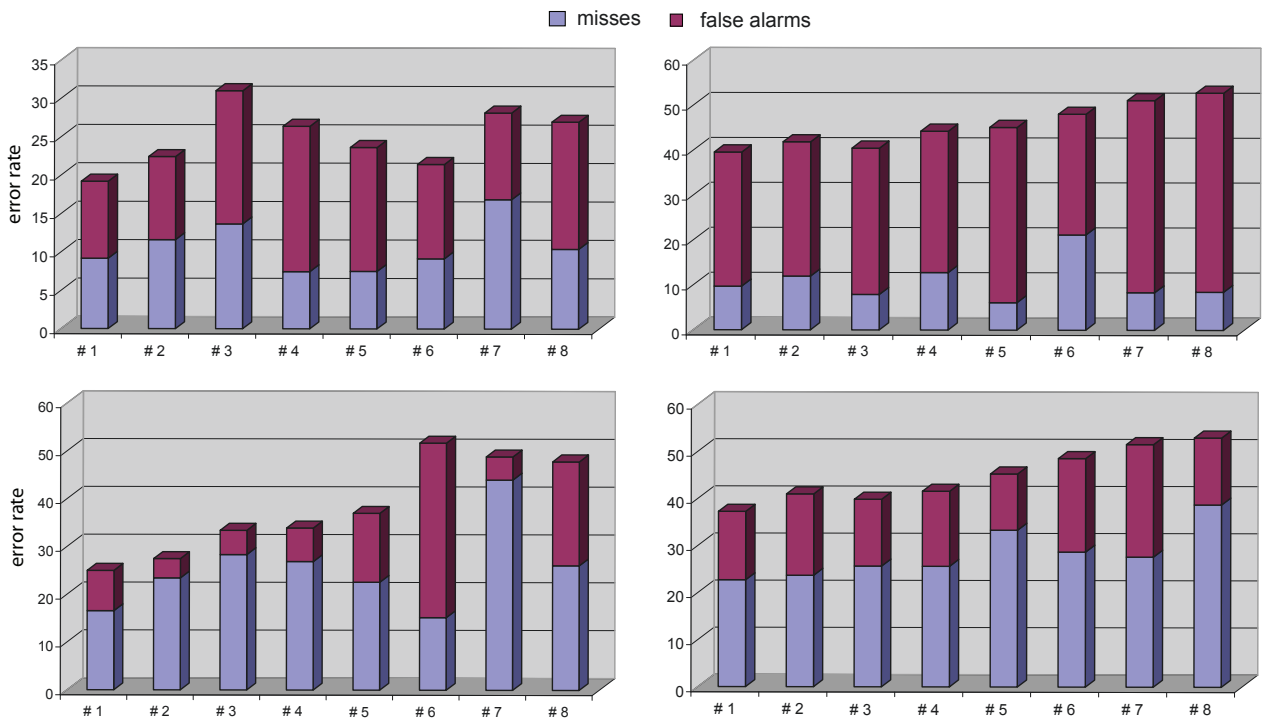

Figure 1 Distribution of errors across the eight systems and for each acoustic class.

According to this plot, more than 50% of the total amount of errors is shorter than 14 s. For comparison, according to the ground truth labels, 50% of audio is represented by segments of duration less than 26 s. Therefore, on average, the duration of erroneous segments is almost twice shorter than that of the ground truth segments.

In Figure 3, we compare the error distribution for three types of segments in the testing database: *very difficult*, *difficult*, and *misclassified by the best*. As illustrated in Figure 3a, *very difficult* are those segments which are totally included in error segments from eight systems. *Difficult* segments are those which are included in error segments from at least seven systems. Finally, *misclassified by the best* are those segments where the winner system in evaluation produced errors. The graphical distribution of those three types of segments is displayed in Figure 3b.

The error distribution for those segments, displayed in Figure 3, shows the degree of difficulty of the audio segmentation task. On average, only 6.98% of the segments in the testing database are *very difficult*. The rest of the segments were detected correctly at least by one detection system. Comparing this number with the final score from the winner system (30.22%), we conclude that there is still a large margin to improve the audio segmentation performance.

Table 3 Confusion matrix of acoustic classes

	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

Figure 4 shows a grouping of the errors which are shared by all the eight segmentation systems. The groups were defined after listening to all the segments which are defined as *very difficult*, and are longer than 5 s. Seven different types of error were distinguished, and the rest were included in *Other*.

According to the plot in Figure 4, a large percentage of shared errors was provoked by the presence of either a low level of sound in the background (23%) or overlapped speech (21%), while the annotator mistakes caused only 8% of the total amount of shared errors.

Discussion

By analyzing both the submitted audio segmentation systems and the corresponding segmentation results, several observations can be extracted which are outlined in the following.

The conventional use of ASR features for the audio segmentation task

Historically, there have been no features specifically designed for the audio segmentation task. In the current

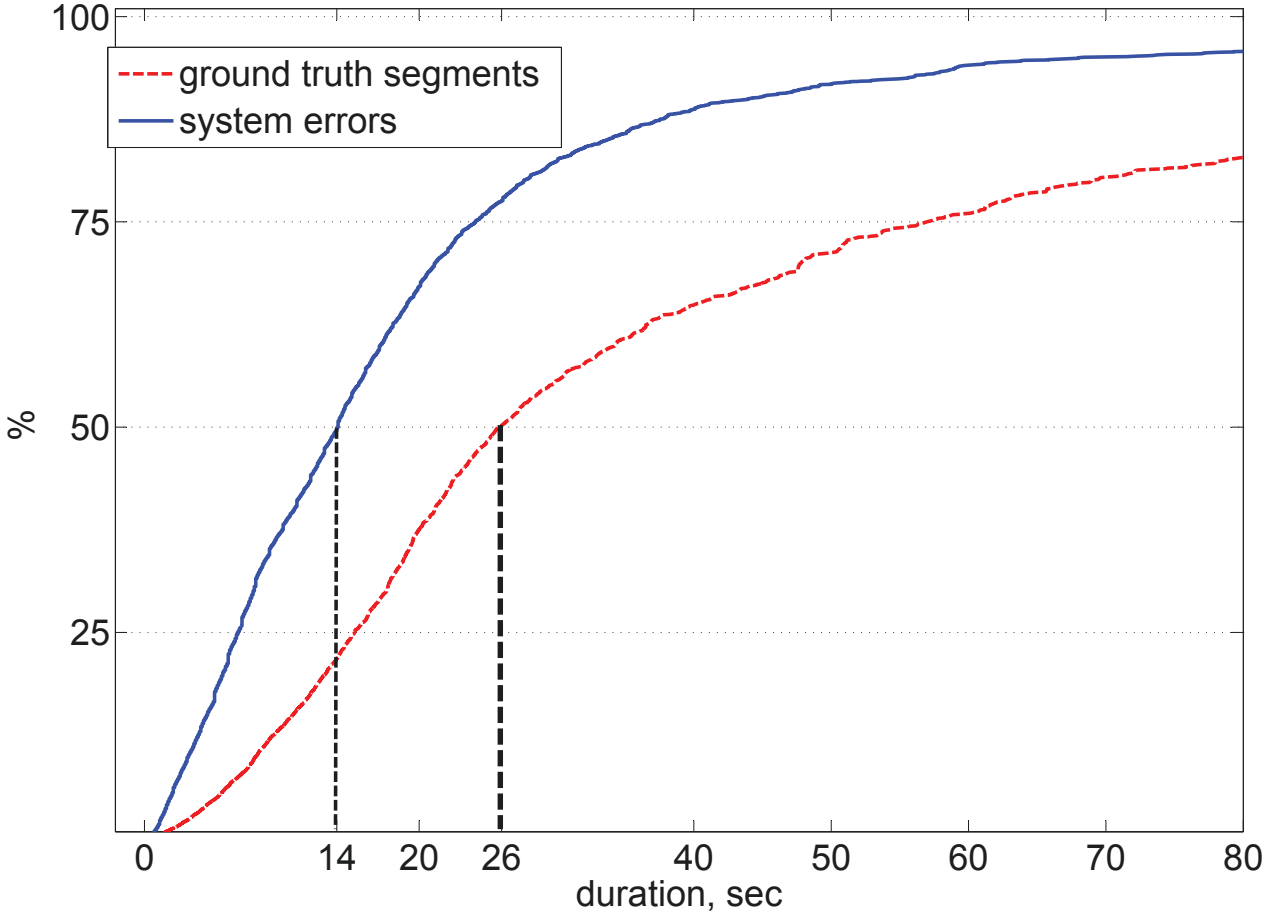

Figure 2 Cumulative distribution of segments in terms of duration.

evaluation, all the systems used features that were designed for the ASR task, like MFCC, PLP, or FF. A few systems combined the ASR features with other perceptual feature sets, but they could not report any significant improvement (for details, see [21]).

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

The best two audio segmentation systems parameterized the audio signal using segment-based features. The system 1 used the mean and variance along 1-s segments;

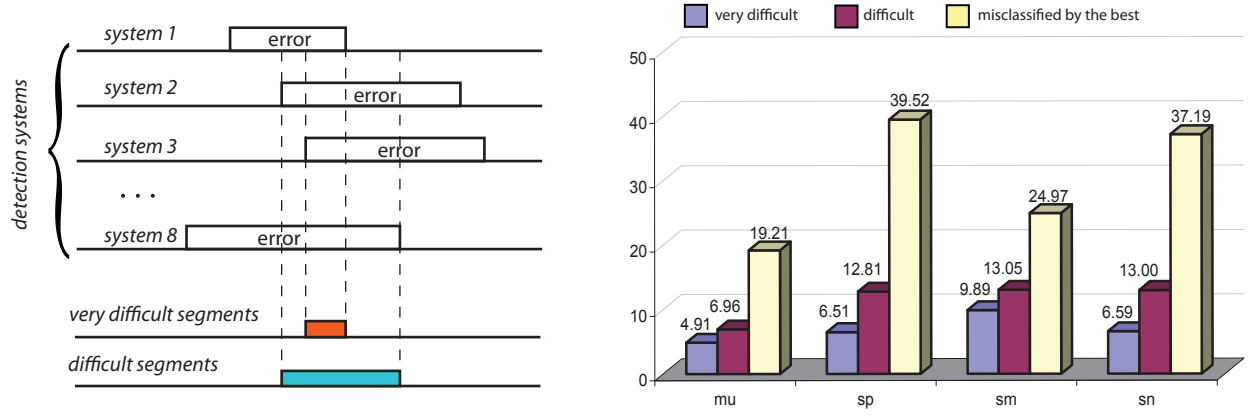

Figure 3 Distribution of error segments in the testing database. **(a)** Illustration of "difficult" and "very difficult" segments; **(b)** Error distribution of "difficult," "very difficult," and "misclassified by the best" segments.

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

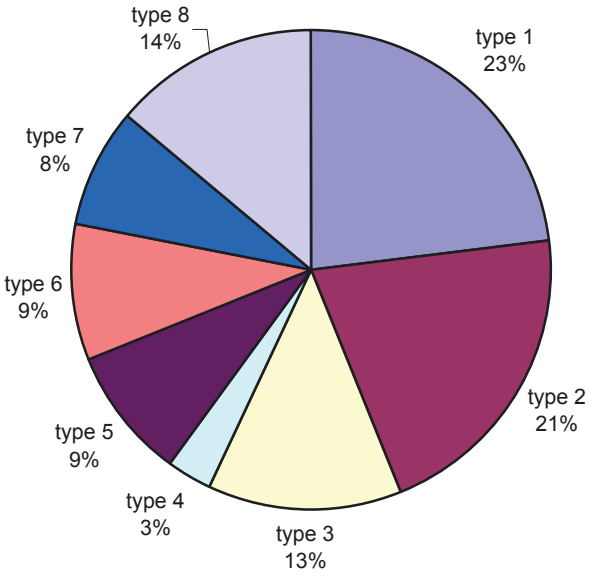

Figure 4 Percentages of distribution of the different types of shared errors.

the system 2 used a super-vector approach to parameterize along even longer segments. It is noted that the third best system used SDC coefficients, which take into account a long audio context. 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.

The majority of the audio segmentation 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. In addition, more parameters for tuning are required, which makes the system task dependent.

The hierarchical detection approach seems to be effective

Four research groups reported an improvement when using a hierarchical organization of the detection process. One of the most important decisions when using this kind of architecture lies in the orderings of the detection modules, since some of them may benefit greatly from the previous detection of certain classes. Those four audio segmentation systems detect the easiest classes ("Music" and silence, which is included in "Other") 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.

The fingerprinting approach for music detection seems to be effective

Finding of repetitions with fingerprinting seems to be useful in audio segmentation of BN due to the omnipresence of advertisements, jingles, and even repeated programs. The system 6, which used that approach, got the second best result for the class "Music."

Challenge of the audio segmentation task

Only 6.98% of the audio segments were detected incorrectly by all the audio segmentation systems. The rest of audio was recognized correctly by at least one detection system. Comparing this number with the score obtained by the winner system (30.22%), we conclude that there is still a large margin for improvement of segmentation results. Taking into account that the main source of mistakes are confusions between "Music" and "Speech over music," between "Speech over music" and "Speech over noise," as well as between "Speech" and "Speech over noise." Future research efforts should be devoted to improved detection of background sounds.

Complementarity of different segmentation systems

The segmentation results from different systems are complementary up to some extent, so that the combination of them yields improvement in accuracy. A simple majority voting fusion scheme of the best three systems reduces the average score to 28.60%, and the fusion of the best five systems, to 29.19%. Comparing these

numbers with the score obtained by the winner system (30.22%), we conclude that post-processing of the segmentation results from different segmentation systems is beneficial.

Applicability of the systems to work in real time

Unlike many speech recognition or speaker diarization systems, whose performances drop drastically when operating in real time, the described audio segmentation systems can work in real time due to their relative simplicity. In fact, four participants reported timing results (systems 3, 4, 5, and 8) and the total CPU time, computed by adding CPU times for feature extraction and audio segmentation, falls below 1 × RT (real-time factor).

Conclusion

In this article, first of all, a new, large, freely available and recently recorded BN database, which can be used for the audio segmentation task, has been presented, along with the setup and the specific metric used in the reported audio segmentation evaluation. Then, we have presented the audio segmentation systems, and the results from the eight different research groups which participated in the Albayzín-2010 evaluation, and compared their approaches and techniques.

All the presented systems used typical speech recognition features (MFCC, PLP, or FF), and most systems employed HMM-based Viterbi decoding for segmentation. The best two results were obtained by the systems that exploited segment-based features. Four presented systems reported an improvement by using a hierarchical organization of the detection process, so that the detection of the easiest classes (like "Music," in our task) at the beginning of the detection process is beneficial. Owing to the omnipresence of repeated programs and sounds in the BN data, the detection of repetitions seems to be effective for music segmentation.

It is also worth mentioning that the segmentation results from different systems are complementary up to some extent; in fact, a 1.62% absolute improvement is achieved in this article, when using a simple majority voting fusion of the best three systems. By analyzing the shared segmentation errors from all the submitted systems, we conclude that a large percentage of errors was induced either by the presence of a low level of sound in the background (23%) or by the overlapping speech (21%), while the annotator mistakes accounted for only 8% of the total amount of shared errors. On average, only 6.98% of the segments in the testing database are *very difficult*, in the sense that they were not detected correctly by any of the systems. Comparing this number with the score obtained by the winner system (30.22%), we conclude that there is still a large margin for improving the audio segmentation results.

Endnotes

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

Abbreviations

AC: acoustic classes; ASR: automatic speech recognition; BN: broadcast news; FF: frequency-filtered; SDC: shifted delta coefficients;

Acknowledgements

This study has been funded by the Spanish project SAPIRE (TEC2007-65470) and SARAI (TEC2010-21040-C02-01). The authors wish to thank their 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 Henrik Schulz for managing the collection of the database and help during its annotation. The first author is partially supported by a grant from the Catalan autonomous government.

Competing interests

The authors declare that they have no competing interests.

Received: 11 January 2011 Accepted: 17 June 2011
Published: 17 June 2011

References

1. DS Pallet, A look at NIST's benchmark ASR tests: past, present, and future. *Technical Report, National Institute of Standards and Technology (NIST)* (Gaithersburg, MD, USA, 2003)
2. J Saunders, Real-time discrimination of broadcast speech/music, in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, **2**, *993–996 (1996)*
3. E Scheirer, M Slaney, Construction and evaluation of a robust multifeature speech/music discriminator, in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1997
4. T Zhang, C-C Kuo, Hierarchical classification of audio data for archiving and retrieving, in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, **6**, 3001–3004 (1999)
5. L Lu, HJ Zhang, H Jiang, Content analysis for audio classification and segmentation, in *IEEE Transactions on Speech and Audio Processing*, **10**(7), 504–516 (2002)
6. K El-Maleh, M Klein, G Petrucci, P Kabal, Speech/music discrimination for multimedia applications, in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, **6**, 2445–2448 (2000)
7. W Chou, L Gu, Robust singing detection in speech/music discriminator design, in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, **2**, 1331–1334 (2001)
8. S Srinivasan, D Petkovic, D Ponceleon, Toward robust features for classifying audio in the cue video system, in *Proceedings 7th ACM International Conference on Multimedia*, 393–400 (1999)
9. A Bugatti, A Flammini, P Migliorati, Audio classification in speech and music: a comparison between a statistical and a neural approach. EURASIP J. Appl. Signal Process, **2002**(4), 372–378 (2002)
10. J Ajmera, I McCowan, H Bourlard, Speech/music segmentation using entropy and dynamism features in a HMM classification framework. Speech Commun. **40**(3), 351–363 (2003)
11. T Izumitani, R Mukai, K Kashino, A background music detection method based on robust feature extraction, in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 13–16 (2008)
12. P Dhanalakshmi, S Palanivel, V Ramalingam, Classification of audio signals using SVM and RBFNN, in *Proceedings Expert Systems with Applications*, **36**(2), 6069–6075 (2009)
13. S Lefèvre, N Vincent, A two level strategy for audio segmentation, *Digital Signal Processing*, **21**(2), 270–277 (2011)
14. M Exposito, G Galan, R Reyes, V Candeas, Audio coding improvement using evolutionary speech/music discrimination, in *Proceedings IEEE Conference on Fuzzy Systems*, 1–6 (2007)
15. X Huang, A Acero, H-W Hon, *Spoken Language Processing: A Guide to Theory, Algorithm and System Development* (Prentice Hall, 2001)

16. C Clavel, T Ehrette, G Richard, Events detection for an audio-based surveillance system, in *Proceedings IEEE International Conference on Multimedia and Expo*, 2005
17. S Kiranyaz, AF Qureshi, M Gabbouj, A generic audio classification and segmentation approach for multimedia indexing and retrieval, in *Proceedings of the European Workshop on the Integration of Knowledge, Semantics and Digital Media Technology, 55–62 (2004)*
18. O Lartillot, P Toiviainen, MIR in Matlab (II): a toolbox for musical feature extraction from audio, in *Proceedings International Conference on Music Information Retrieval*, 2007
19. S Pfeiffer, Pause concepts for audio segmentation at different semantic levels, in *Proceedings ACM International Conference on Multimedia*, 187–193 (2001)
20. SS Chen, PS Gopalkrishnan, Speaker, environment and channel change detection and clustering via the Bayesian information criterion, in *Proceedings of DARPA Broadcast News Transcription and Understanding Workshop*, 127–132 (1998)
21. FALA 2010 "VI Jornadas en Tecnología del Habla" and II Iberian SLTech Workshop, http://fala2010.uvigo.es/images/proceedings/index.html

doi:10.1186/1687-4722-2011-1
Cite this article as: Butko and Nadeu: **Audio segmentation of broadcast news in the Albayzin-2010 evaluation: overview, results, and discussion.** *EURASIP Journal on Audio, Speech, and Music Processing* 2011 **2011**:1.