Multimed Tools Appl (2017) 76:7421–7444
DOI 10.1007/s11042-016-3386-2

Ensemble audio segmentation for radio and television programmes

Paula Lopez-Otero[1] · **Laura Docio-Fernandez**[1] · **Carmen Garcia-Mateo**[1]

Received: 7 May 2015 / Revised: 3 January 2016 / Accepted: 23 February 2016 /
Published online: 9 March 2016

Abstract State-of-the-art audio segmentation strategies obtain good results when performing simple tasks but its performance is degraded when segmenting real-world scenarios such as radio and television programmes; this issue can be partially solved by performing a fusion of different audio segmentation strategies. Hence, a framework to perform decision-level fusion in the audio segmentation task is presented in this paper. First, the class-conditional probabilities of each audio segmentation strategy are estimated from a confusion matrix obtained by performing audio segmentation in a training dataset. Performance measures are extracted from these class-conditional probabilities, which are used to compute different estimates of the classifier's reliability; specifically, reliability estimates based on precision, recall, accuracy, F-score and mutual information were proposed. These reliability estimates are used as weights in a weighted majority voting fusion strategy. The validity of the proposed fusion scheme and reliability estimates was assessed in the framework of Albayzin 2010, 2012 and 2014 audio segmentation evaluations, which consisted in segmenting collections of radio and television programmes. The experimental results showed that this simple fusion strategy improves the performance achieved by the individual audio segmentation strategies and by other well-known decision-level fusion strategies.

Keywords Ensemble classification · Confusion matrix · Reliability estimation · Audio segmentation

✉ Paula Lopez-Otero
plopez@gts.uvigo.es

[1] AtlantTIC Research Center, Multimedia Technologies Group, University of Vigo, E.E. Telecomunicación, Campus Universitario de Vigo, S/N, C.P. 36310 Vigo, Spain

1 Introduction

Audio segmentation is a task consisting in dividing an audio signal into homogeneous segments according to some criteria; it encompasses simple tasks such as speech activity detection, which consists in detecting the speech parts of an audio stream, and more complex tasks such as the detection of other classes such as speech combined with background music or noise. Audio segmentation plays a crucial role in the performance achieved by subsequent speech technologies tasks. Automatic speech recognition (ASR) is usually preceded by a segmentation stage in order to improve the detection of sentences, to remove music and non-speech information that would cause insertions in the transcription and to detect overlapped speech [30]. Hence, its use is very common in broadcast news transcription and subtitling [22], but it is also applied to more complex tasks such as information retrieval of big collections of multimedia documents [23]. Besides these tasks, audio segmentation can be used to annotate and tag audio or multimedia documents in order to perform tasks such as supervising the payment of royalty fees related to music [33]. Another task in which audio segmentation has a paramount importance is the automatic classification of music collections by performer or gender, in which music and boundaries between songs must be detected before classification [40].

Audio segmentation strategies can be classified in two groups: strategies for segmentation by acoustic change detection and strategies for segmentation by classification. Strategies of the first type, also known as distance-based approaches, compare adjacent windows of audio by using a given distance measure, deciding whether there is an acoustic change between the windows in function of this distance. After the division of the audio stream into segments, these segments can be classified into the target classes. Common approaches for segmentation by acoustic change detection imply the segmentation of the audio using the Bayesian information criterion (BIC) [32] or its variants [1, 9]; after this segmentation stage, statistical modelling via Gaussian mixture models (GMMs) or support vector machines (SVM) are common approaches for the classification step [21].

The second type of audio segmentation strategies consists in, instead of looking for an acoustic change in a window of data, directly assigning a target class to an audio frame or a window of data, hence resulting in a sequence of homogeneous segments. These segmentation approaches, known as model-based strategies, use statistical models to represent the different acoustic classes; using these models, the audio stream is segmented by finding the most likely sequence of models. The most common strategy for segmentation by classification is Viterbi decoding [41], which is an algorithm that finds the most likely sequence of states given the input data [28, 38]. Another common approach is the use of GMMs for frame-by-frame classification of the audio streams [11, 28]. Other approaches for segmentation by classification are based on different machine learning algorithms such as neural networks [22] or support vector machines [27].

State-of-the-art audio segmentation techniques achieve a good performance when dealing with easy tasks in controlled situations, for example speech activity detection in clean conditions, but the detection of more complex acoustic classes is still a challenge. The different approaches for audio segmentation that can be found in the literature have partially solved the problem, but there is still room for improvement. A plausible solution to deal with degraded performance in difficult audio segmentation tasks consists in the integration of data and knowledge from different sources, which is known as data fusion [5] and aims at enhancing the strengths of different systems while dimming their weaknesses at the same time.

There are three main types of fusion techniques: feature-level fusion, score-level fusion and decision-level fusion. Feature-level fusion consists in the combination of different features extracted from the waveform in order to create high-dimensional feature vectors, which is commonly performed by concatenating and normalizing the different feature vectors [29]. This results in large feature vectors which often include redundant information that is increasing the computational cost of the system and even damaging its performance, which is commonly avoided by performing feature selection [13, 24] or projecting the data into a discriminative subspace [8, 10]. Score-level fusion consists in, given a set of systems that output a score for each target class when classifying an example, combining the scores in order to obtain a single score per class that leads to the final decision. These scores can be interpreted as a posteriori probabilities of the different classes, so they can be combined by means of combination rules [16]; scores can also be considered as fuzzy membership values, so fuzzy logic techniques [7] or belief functions [34] can be used to obtain a classification result. It is also possible to use the scores to train another classifier that will output the final classification decision.

The last type of fusion techniques, namely decision-level fusion, uses the classification decisions of different systems to decide which is the actual class of an example. The combination of different classifiers in this way is commonly referred to as ensemble classifier. Although combining different systems by making ensembles of classifiers may optimize strengths and minimize weaknesses, this is not always an easy task. Feature and score-level fusion seem to be the best choice, as the amount of information used in order to make a classification decision is greater than in the case of decision-level fusion, but they lack of some qualities that are present in a decision-level fusion scheme. As stated in [18], any classifier is capable of producing an output label, making this fusion level the most universal one. Moreover, this type of fusion scheme has great scalability as, in general, no re-training of the whole system is necessary anytime a new classifier or modality has to be integrated [31]. In addition, the application of decision-level fusion approaches is common in pattern recognition scenarios other than audio segmentation, as other types of fusion may not be feasible due to the incompatibility of the different systems. For example, in [37] and [39], this fact is discussed in the scenario of template-protected biometrics. Another example can be found in [12], where the incompatibility of the sensors used for mine detection only allows for a decision-level fusion.

The most straightforward combination method in ensemble classification is the majority voting, i.e. the class that was selected by more systems is the one that is chosen [26]. Another habitual version of decision-level fusion is the weighted majority voting [20], that requires an estimation of the weight assigned to each system, which is commonly based on the reliability of that system: a training set is classified using a given system, and the reliability of that system, i.e. its weight, is assigned depending on how accurately it classified the training examples. Performing a relaxation of the weighted majority voting scheme leads to the Naive Bayes (NB) approach [19], in which a reliability is computed for each class and for each classifier; this reliabilities are obtained by estimating the class-conditional probabilities computing the confusion matrix of the classifier when classifying a training set of examples. One advantage of this method is that it can be used in two-class problems [15] but applying it in multi-class scenarios is also possible [19]. On the other hand, this method only takes into account the true positives achieved by the classifier for that class; this do not necessarily lead to a good solution, as in this way the fusion strategy is not accounting for the errors committed by the classifier, it only accounts for the right choices.

In this work, we propose a framework to create ensembles of audio segmentation systems, that can be used when the only available output information about the individual audio

segmentation strategies is the start and end instants of the audio segments and the target class assigned to them. The class-conditional probabilities for each individual strategy are estimated by computing the confusion matrix resulting from performing audio segmentation on a training dataset. Performance measures extracted from the confusion matrices are then used to estimate the reliability of the individual strategies when classifying each of the target classes. Finally, a weighted majority voting scheme is used to combine the outputs of the different audio segmentation systems that comprise the ensemble. In this way, any audio segmentation approach can be part of this ensemble system, as it does not matter whether it outputs scores corresponding to the different classification decisions or just the label of the assigned class, given that only the output labels are needed to estimate the reliability. Besides this ensemble classification framework, we propose different reliability estimates that can be extracted from the class-conditional probabilities. These reliability estimates are built by combining the true/false positives and the true/false negatives obtained by the classifier on a training set, which leads to reliabilities that do not only account for the correct choices but also for the errors committed by the classifiers.

The proposed ensemble classification techniques and reliability estimates applied to audio segmentation are assessed in three different experimental frameworks, which are those defined for Albayzin 2010 [3], 2012 [35] and 2014 [25] audio segmentation evaluations (from now on, Albayzin 2010, 2012 and 2014 ASE). These evaluations consisted in performing audio segmentation in broadcast news domain.

The rest of this paper is organized as follows; an overview to ensemble classification is presented in Section 2; in Section 3 we describe the procedure to extract performance measures from a confusion matrix; after defining the performance measures, in Section 4 we propose different reliability estimates by combining them , and a theoretical comparison among this method and others based in the same principle can be found in Section 5 . With respect to the experimental validation of the proposed technique, first some aspects on the application of the proposed technique to the audio segmentation task are described in Section 6; next, the experimental frameworks used in the experiments are described in Section 7; to conclude, the experimental results are presented in Section 8, followed by a discussion in Section 9 and some conclusions and future work in Section 10.

2 Ensembles of classifiers

Two issues have to be taken into account when building an ensemble classifier: the generation of classifiers (where diversity plays a crucial role) and the design of the method for combining them [26]. Focusing on the latter issue, the most basic ensemble classification method is the majority voting (MV) scheme, which counts the votes assigned to each class in order to make a final decision. This scheme assumes that all the votes are equally important, i.e. all classifiers have the same weight in the voting. A relaxation of this constraint leads to the weighted majority voting (WMV) strategy, which assigns a different weight to each classifier depending on system performance, in a way that votes of top performing systems count more than those of other systems. However, not all the systems are equally good at classifying different classes, so a relaxation of the constraint that assumes that all the classes are equiprobable can be done; in this way, a different weight can be assigned to each classifier and each class according to their performance [19]. These weights can be considered as reliabilities, as they indicate how reliable a classifier is when classifying a specific class. There is another constraint in the latter approach that can be relaxed, namely the assumption of equal individual accuracies; relaxing this assumption leads to a Naive Bayes (NB)

combiner [19], in which a confusion matrix is obtained from classifying a set of examples and the weight for each classifier and class is equal to the corresponding class-conditional probability.

Formally, let $\mathrm{E} = \{\mathrm{e}_1, \ldots, \mathrm{e}_\mathrm{m}\}$ be an ensemble classifier composed of m classifiers (or ensembles), and let $\mathrm{C} = \{\mathrm{c}_1, \ldots, \mathrm{c}_\mathrm{n}\}$ be a set of n classes. Classifier E has to classify examples taking into account the individual decisions of each ensemble $\mathrm{e}_i,\ i \in 1, \ldots, \mathrm{m}$. To do so, each e_i has an associated reliability $\mathrm{r}_{\mathrm{e}_i,\mathrm{c}_j}$ for each of the n classes, which is an estimate of the classification performance of e_i when classifying class c_j. As stated above, this estimate can be made by observing the behaviour of the ensembles when classifying a set of examples [36]. Given that a WMV scheme is used to combine the different classifiers, the combination rule can be defined as:

$$c^* = arg \max_{\mathrm{c}_j \in \mathrm{C}} \sum_{\mathrm{e}_i \in \mathrm{E}} \mathrm{r}_{\mathrm{e}_i,\mathrm{c}_j} \delta_{\mathrm{e}_i,\mathrm{c}_j} \tag{1}$$

where $\delta_{\mathrm{e}_i,\mathrm{c}_j} = 1$ if e_i outputs c_j and $\delta_{\mathrm{e}_i,\mathrm{c}_j} = 0$ otherwise.

3 Estimating class-conditional probabilities and performance measures from a confusion matrix

A confusion matrix contains information about the actual and predicted classifications of a classification system, so it provides valuable information about the errors produced by the classifier. Let $\mathbf{CM}_{\mathrm{e}_i}$ be the confusion matrix of ensemble e_i, where $\mathbf{CM}_{\mathrm{e}_i}(\mathrm{c}_j|\mathrm{c}_k)$ is the number of examples of class c_k that were classified as class c_j by e_i. This matrix, once normalized by the number of classified examples, represents an estimate of the class-conditional probabilities of actual and predicted classes:

$$\mathbf{P}_{\mathrm{e}_i} = \frac{1}{\|\mathbf{CM}_{\mathrm{e}_i}\|} \mathbf{CM}_{\mathrm{e}_i} \tag{2}$$

where $\|\mathbf{CM}_{\mathrm{e}_i}\|$ is the cardinality of the confusion matrix, i.e. the number of examples used to build the matrix. Hence, $\mathbf{P}_{\mathrm{e}_i}(\mathrm{c}_j|\mathrm{c}_k)$ is the conditional probability that ensemble e_i classifies an example of class c_k as class c_j.

Four main performance measures can be extracted from the class-conditional probabilities of ensemble e_i for class c_j:

– The true positive $\mathrm{TP}_{\mathrm{e}_i,\mathrm{c}_j}$ represents the ratio of correctly classified examples of class c_j:

$$\mathrm{TP}_{\mathrm{e}_i,\mathrm{c}_j} = \mathbf{P}_{\mathrm{e}_i}(\mathrm{c}_j|\mathrm{c}_j) \tag{3}$$

– The true negative $\mathrm{TN}_{\mathrm{e}_i,\mathrm{c}_j}$ is the ratio of examples of class $\mathrm{c}_k, k \neq j$ not classified as c_j:

$$\mathrm{TN}_{\mathrm{e}_i,\mathrm{c}_j} = \sum_{\substack{x=1 \\ x \neq j}}^{\mathrm{n}} \sum_{\substack{y=1 \\ y \neq j}}^{\mathrm{n}} \mathbf{P}_{\mathrm{e}_i}(\mathrm{c}_x|\mathrm{c}_y) \tag{4}$$

– The false positive $\mathrm{FP}_{\mathrm{e}_i,\mathrm{c}_j}$ is the ratio of examples of class $\mathrm{c}_k, k \neq j$ classified as c_j

$$\mathrm{FP}_{\mathrm{e}_i,\mathrm{c}_j} = \sum_{\substack{x=1 \\ x \neq j}}^{\mathrm{n}} \mathbf{P}_{\mathrm{e}_i}(\mathrm{c}_j|\mathrm{c}_x) \tag{5}$$

– The false negative $\mathrm{FN}_{\mathrm{e}_i,\mathrm{c}_j}$ is the ratio of examples of class c_j classified as c_k, $k \neq j$:

$$\mathrm{FN}_{\mathrm{e}_i,\mathrm{c}_j} = \sum_{\substack{x=1 \\ x\neq j}}^{\mathrm{n}} \mathbf{P}_{\mathrm{e}_i}(\mathrm{c}_x|\mathrm{c}_j) \tag{6}$$

Figure 1, which shows the estimated class-conditional probabilities for one of the ensembles used in our experiments, exemplifies how to compute the four aforementioned performance measures in a five-class classification system for class c2.

4 Proposed reliability estimates

Given $\mathbf{P}_{\mathrm{e}_i}$, two different ways to estimate the reliability $\mathrm{r}_{\mathrm{e}_i,\mathrm{c}_j}$ of the ensemble system e_i for class c_j are described below.

4.1 Mutual information

The mutual information of e_i for class c_j represents the amount of information that e_i shares with the groundtruth g about class c_j. Its reliability is defined as:

$$\mathrm{r}_{\mathrm{MI}_{\mathrm{e}_i,\mathrm{c}_j}} = \mathrm{p}_{\mathrm{g},\mathrm{e}_i,\mathrm{c}_j} \cdot \log\left(\frac{\mathrm{p}_{\mathrm{g},\mathrm{e}_i,\mathrm{c}_j}}{\mathrm{p}_{\mathrm{e}_i,\mathrm{c}_j}\mathrm{p}_{\mathrm{g},\mathrm{c}_j}}\right) \tag{7}$$

where $\mathrm{p}_{\mathrm{g},\mathrm{e}_i,\mathrm{c}_j}$ is the probability that e_i correctly outputs class c_j:

$$\mathrm{p}_{\mathrm{g},\mathrm{e}_i,\mathrm{c}_j} = \mathrm{TP}_{\mathrm{e}_i,\mathrm{c}_j} \tag{8}$$

$\mathrm{p}_{\mathrm{e}_i,\mathrm{c}_j}$ is the probability that e_i outputs class c_j:

$$\mathrm{p}_{\mathrm{e}_i,\mathrm{c}_j} = \mathrm{TP}_{\mathrm{e}_i,\mathrm{c}_j} + \mathrm{FP}_{\mathrm{e}_i,\mathrm{c}_j} \tag{9}$$

and $\mathrm{p}_{\mathrm{g},\mathrm{c}_j}$ is the groundtruth probability of class c_j:

$$\mathrm{p}_{\mathrm{g},\mathrm{c}_j} = \mathrm{TP}_{\mathrm{e}_i,\mathrm{c}_j} + \mathrm{FN}_{\mathrm{e}_i,\mathrm{c}_j} \tag{10}$$

Substituting (9) and (10) into (7), the mutual information based reliability estimate $\mathrm{r}_{\mathrm{MI}_{\mathrm{e}_i,\mathrm{c}_j}}$ is defined as follows:

$$\mathrm{r}_{\mathrm{MI}_{\mathrm{e}_i,\mathrm{c}_j}} = \mathrm{TP}_{\mathrm{e}_i,\mathrm{c}_j} \log \frac{\mathrm{TP}_{\mathrm{e}_i,\mathrm{c}_j}}{(\mathrm{TP}_{\mathrm{e}_i,\mathrm{c}_j} + \mathrm{FP}_{\mathrm{e}_i,\mathrm{c}_j})(\mathrm{TP}_{\mathrm{e}_i,\mathrm{c}_j} + \mathrm{FN}_{\mathrm{e}_i,\mathrm{c}_j})} \tag{11}$$

		Predicted					
		c1	**c2**	**c3**	**c4**	**c5**	
Actual	**c1**	0.189	0.001	0.027	0.010	0.001	$\mathrm{TP}_{\mathrm{ei,c2}}$
	c2	0.001	0.079	0.005	0.001	0.001	$\mathrm{FN}_{\mathrm{ei,c2}}$
	c3	0.055	0.008	0.244	0.004	0.004	$\mathrm{FP}_{\mathrm{ei,c2}}$
	c4	0.022	0.006	0.087	0.129	0.045	$\mathrm{TN}_{\mathrm{ei,c2}}$
	c5	0.002	0.005	0.002	0.003	0.051	

Fig. 1 Example of computation of performance measures using the estimated class-conditional probabilities of a five-class classifier

4.2 Precision

Precision, which decreases when the false positive ratio increases, represents the relevance of the positively classified examples. A precision-based reliability estimate can be defined using the performance measures described in Section 3 as follows:

$$r_{PR_{e_i,c_j}} = \frac{TP_{e_i,c_j}}{TP_{e_i,c_j} + FP_{e_i,c_j}} \tag{12}$$

4.3 Recall

Recall, which decreases when the false negative ratio increases, measures the goodness of a system when detecting positives. Its corresponding reliability estimate can be defined as:

$$r_{RE_{e_i,c_j}} = \frac{TP_{e_i,c_j}}{TP_{e_i,c_j} + FN_{e_i,c_j}} \tag{13}$$

4.4 F-score

The F-score takes into account both the false positive and the false negative ratios of a classifier. It is defined as the harmonic mean between precision and recall.

$$\text{F-score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{14}$$

Combining (12) and (13), the F-score based reliability estimate $r_{F_{e_i,c_j}}$ is defined as follows:

$$\begin{aligned} r_{F_{e_i,c_j}} &= \frac{2 \cdot r_{PR_{e_i,c_j}} \cdot r_{RE_{e_i,c_j}}}{r_{PR_{e_i,c_j}} + r_{RE_{e_i,c_j}}} \\ &= \frac{2TP_{e_i,c_j}}{2TP_{e_i,c_j} + FP_{e_i,c_j} + FN_{e_i,c_j}} \end{aligned} \tag{15}$$

4.5 Accuracy

A reliability estimate can be defined such that a classifier is more or less reliable depending on its ability to classify a given class, i.e. depending on its accuracy:

$$r_{RE_{e_i,c_j}} = TP_{e_i,c_j} + TN_{e_i,c_j} \tag{16}$$

5 Comparison with other existing methods

As briefly discussed in Section 2, a simple way to combine classifiers is the WMV strategy, in which the accuracy of each classifier is used as its reliability. However, this strategy assumes that the accuracy of a classifier is the same for all the classes; this constraint is removed in the NB strategy, which uses the class-conditional probabilities defined in (2) as reliabilities since, in this way, the reliability depends on the performance when a classifier classifies a given class. However, the class-conditional probability only accounts for the true positives TP_{e_i,c_j}: this means that only the correct classifications are taken into account, without considering the errors, namely false positives or false negatives. This can lead to

the following situation: given a classifier e_i that always outputs class c_j, TP_{e_i,c_j} would be maximum, but FP_{e_i,c_j} would be very high. If the class-conditional probability is considered as a reliability estimate, a high reliability would be assigned to a classifier like this one regardless the fact that it always outputs the same class. Nevertheless, when using performance measures such as those proposed in Section 4, the false positives and false negatives are also considered, hence avoiding this type of issues.

Other of the methods to estimate reliabilities that can be found in the literature account for the behaviour of groups of classifiers, such as the behavioural knowledge space (BKS), which consists on learning which combination of the outputs of the classifiers produces which output [14]. This method and others such as the pairwise fusion matrix method described in [17] lead to a loss of scalability of the fusion strategy, as well as to an increase of the computational cost when the number of classifiers and classes is big. The fusion strategy proposed in this paper preserves the scalability and efficiency even when the number of classifiers and classes is high, as the confusion matrices are individually computed for each classifier without taking into account the output produced by the other classifiers that are part of the ensemble. In addition, the proposed method do not require an optimization procedure as happens in other fusion strategies that learn agreement or disagreement patterns among individual classifiers, which do not always lead to a great improvement in performance [31] and increases the dependence on the training data, as those observed patterns are expected to be found in the test data as well.

6 Use case: audio segmentation

This Section describes how the ensemble classification methods described in previous Sections were applied in the audio segmentation task. When combining audio segmentation outputs, it must be noted that the examples to be classified, i.e. the acoustic-homogeneous segments, are not predefined, so different audio segmentation strategies may output segments that start and end at different time instants, making it impossible to consider the segments as examples to be classified. Another option would consist in taking temporal windows of a given duration and combining the corresponding labels of the different systems, but this would lead to a resolution loss at segment boundaries. Hence, to overcome this problem, each frame is considered as an example to be classified; in this way, as the resolution of the system is the highest possible, the loss of information at segment boundaries is avoided by working in a frame-by-frame-basis.

A set of individual audio segmentation strategies must be defined in order to fuse their outputs and assess the proposed fusion technique. In these experiments, four different state-of-the-art audio segmentation approaches were used; two of them consist in segmentation followed by classification approaches, while the other two consist in segmentation by classification approaches:

- Audio segmentation strategy e_1: segmentation is done using the BIC strategy [32] following the classic growing-sliding window approach [6]. Classification is performed using an SVM with a lineal kernel [21].
- Audio segmentation strategy e_2: segmentation and classification are performed in the same way as in e_1, but in this case the SVM uses a radial basis function kernel.
- Audio segmentation strategy e_3: Viterbi decoding is used, and the different classes are modelled using GMMs [38].

– Audio segmentation strategy e_4: Viterbi decoding is used, but in this case the different classes are modelled using three-state HMMs.

The acoustic features used in the aforementioned audio segmentation strategies were 12 Mel-frequency cepstral coefficients, augmented with their energy, delta and acceleration coefficients. These features were extracted every 20 ms using a 10 ms sliding window.

It must be noted that the audio segmentation strategies are named e_i in order to be consistent with the notation followed in the previous Sections of this paper.

7 Experimental frameworks

This Section describes the experimental frameworks used to assess the proposed fusion strategy and the reliability estimates. As mentioned in the Introduction, three different experimental frameworks used in audio segmentation evaluations were used; specifically, those used in Albayzin 2010, 2012 and 2014 ASE. Two different sets of experiments were carried out: the first one consists in a 10-fold cross-validation experiment on each database, in order to analyse in detail the performance of the proposed techniques; the second one consists in performing the experiment defined for the different Albayzin evaluations, in order to validate the results of the cross-validation experiments as well as to compare the achieved results with the ones obtained in the aforementioned evaluations.

7.1 Albayzin 2010 ASE

Albayzin 2010 ASE consisted in performing audio segmentation into classes "speech" (sp), "music" (mu), "speech over music" (sp+mu), "speech over noise" (sp+no) and "other" (ot), where class "other" contained everything that did not belong in the other classes, and was not considered for evaluation [2].

The database used in Albayzin 2010 ASE was compiled for this evaluation, and it is composed of a set of broadcast news programmes in Catalan recorded from 3/24 TV channel [3]. This database consists of 24 sessions of different durations, which were divided into a training partition and a testing partition, as summarized in Table 1. The database includes speech and non-speech regions, and the speech regions can be clean speech or speech with some background information such as noise or music. Silence detection was automatically performed, while the different audio classes were manually labelled. The distribution of the different classes in the database, both for the evaluation experiment and the 10-fold cross-validation experiment, is represented in Fig. 2.

Table 1 Summary of the datasets of Albayzin 2010, 2012 and 2014 ASE

Database	Set	Number of sessions	Duration
Albayzin 2010	Train	16	57 h 27 min
	Test	8	30 h 4 min
Albayzin 2012	Train	32	5 h 17 min
	Test	72	18 h
Albayzin 2014	Train	20	21 h 16 min
	Test	15	15 h 38 min

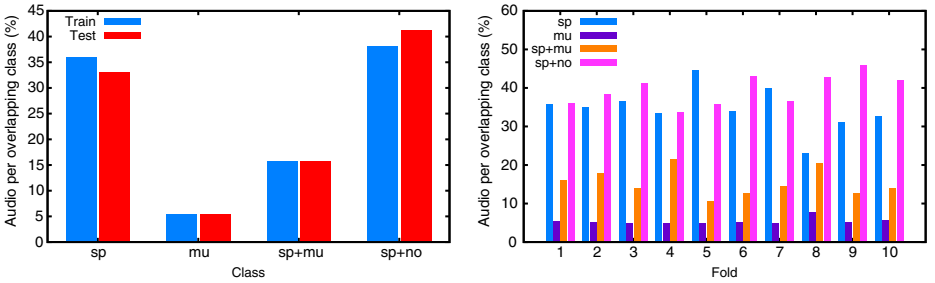

Fig. 2 Class distribution on Albayzin 2010 database. Percentage of audio of each class in the train and test partitions (*left*) and in the 10 folds (*right*) of Albayzin 2010 experiments

7.2 Albayzin 2012 ASE

Albayzin 2012 consisted in automatically detecting the presence or absence of classes speech, music and noise, being it possible that these classes appear either separately or simultaneously, even the three of them at the same time.

The training material of Albayzin 2012 ASE was the database used in Albayzin 2010 ASE, described above, while the development and test sets were composed by recordings of the Aragón Radio radio station. In these experiments, the development dataset was used for training in order to avoid data mismatch, so from now on, the train partition of Albayzin 2012 ASE refers to this development partition. A summary of the data partitions used in the experiments is presented in Table 1.

As commented above, the aim of Albayzin 2012 ASE was to detect classes speech (sp), music (mu) and noise (no), either individually or simultaneously; thus, this task can be also considered as a problem in which seven different classes must be detected, where these seven classes correspond to all the possible combinations of the three individual classes, namely speech (sp), music (mu), noise (no), speech with music (sp+mu), speech with noise (sp+no), music with noise (mu+no) and speech with music and noise (sp+mu+no); we will refer to these classes as overlapping classes. Figure 3 shows the amount of audio of each overlapping class in the different partitions both for the evaluation and cross-validation experiments. As shown in the Figure, the amount of audio of classes "no" and "mu+no" is negligible, so we decided to ignore it in order to avoid training models with such a small amount of data.

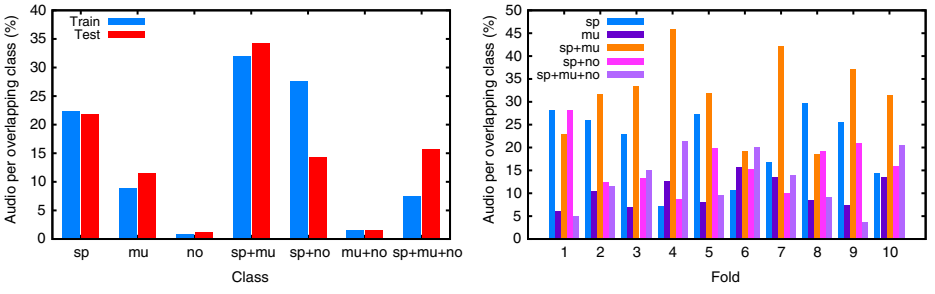

Fig. 3 Class distribution on Albayzin 2012 database. Percentage of audio of each class in the train and test partitions (*left*) and in the 10 folds (*right*) of Albayzin 2012 experiments

7.3 Albayzin 2014 ASE

The task to be performed on Albayzin 2014 ASE was the same as in Albayzin 2012 ASE; it consisted in detecting the presence or absence of classes speech, music and noise.

The data used in this evaluation consists in a combination of the databases used in Albayzin 2010 and Albayzin 2012 ASE, as well as environmental sounds extracted from different websites. The different databases can be merged or even overlapped, making the task more challenging [25].

Two partitions, were defined for Albayzin 2014 ASE, namely a training dataset and a test dataset, which are summarized in Table 1. Figure 4 shows the amount of audio of each overlapping class in these two partitions as well as in the 10 folds of the cross-validation experiment. Classes "no" and "mu+no" were not considered in this case either for the same reason as in Albayzin 2012 ASE.

7.4 Evaluation metric

A common metric used to measure the performance of an audio segmentation system is the segmentation error rate, which is defined as the ratio of the overall segmentation error time to the sum of the durations of the segments that are assigned to each class in an audio file [35]. Given a dataset Ω composed of different audio files, each file is divided into adjacent segments that are separated by change-points. The segmentation error time of a segment n is defined as:

$$\Xi(n) = T(n)\left[max(N_{ref}(n), N_{sys}(n)) - N_{correct}(n)\right] \tag{17}$$

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

The overall segmentation error rate is computed as

$$SER(\%) = \frac{\sum_{n\in\Omega} \Xi(n)}{\sum_{n\in\Omega} \left(T(n)N_{ref}(n)\right)} \tag{18}$$

This error measure includes three different types of error:

– Class error time (CET): amount of time that was assigned to an incorrect class.

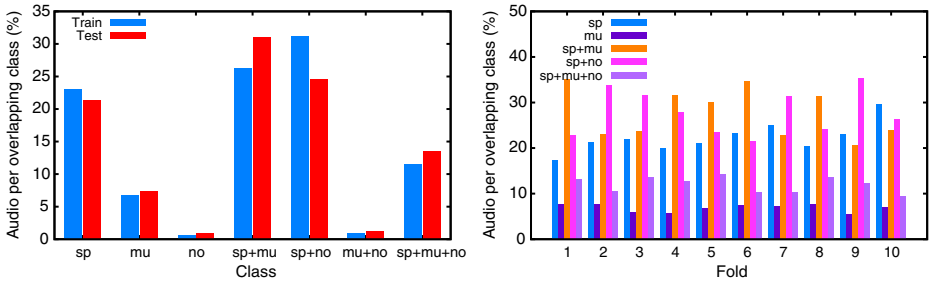

Fig. 4 Class distribution on Albayzin 2014 database. Percentage of audio of each class in the train and test partitions (*left*) and in the 10 folds (*right*) of Albayzin 2014 experiments

– Missed class time (MCT): amount of time that a class is present but not labelled by the audio segmentation system.
– False alarm class time (FACT): amount of time that a class is not present but it was labelled by the audio segmentation system.

A forgiveness collar of one second was considered in order to alleviate the influence of inconsistencies and errors in the manual annotations.

It must be noted that SER was the performance measure used in Albayzin 2012 and 2014 ASE, but not in Albayzin 2010 ASE, where a different metric was used [3]. In this paper, in order to have coherence between the different experimental frameworks, this metric was used in Albayzin 2010 ASE experiments as well.

8 Experimental results

In this Section, the performance of the fusion strategies for audio segmentation presented in this paper are assessed on the experimental frameworks described in Section 7.

As mentioned in Section 7, two different experiments are performed on each experimental framework, namely a 10-fold cross-validation experiment and the test experiment defined for the evaluations, namely evaluation experiments. Table 2 shows the performance achieved in the different experimental scenarios; these results, as well as those achieved when using a MV fusion strategy, must be outperformed by the proposed fusion techniques.

A reference system is necessary in order to establish a baseline to the performance of the proposed fusion strategy and the different reliability estimates; on the one hand, the fusion strategy must outperform the systems that are being combined; also, the proposed fusion strategy is expected to outperform simple fusion strategies. In the 10-fold cross validation experiment, a MV strategy is used as a baseline system; in the evaluation experiments, the MV system is used for comparison as well as a NB approach [18] and a WMV scheme, in which the optimal weights of the classifiers are selected based on the global accuracy (different to the per class accuracy as defined in Section 4) of each classifier as suggested in [18].

8.1 10-fold cross-validation experiments

Tables 3, 4 and 5 show the SER obtained on Albayzin 2010, 2012 and 2014 ASE audio segmentation experiments, respectively, when constructing ensemble audio segmentation systems of two, three and four individual systems using the different ensemble strategies

Table 2 SER (%) of the audio segmentation strategies on Albayzin 2010, 2012 and 2014 experimental frameworks when performing two different experiments: 10-fold cross-validation (CV) and test both on the training and test partitions defined for Albayzin ASE

	Albayzin 2010			Albayzin 2012			Albayzin 2014		
System	CV	Train	Test	CV	Train	Test	CV	Train	Test
e_1	19.41	19.32	22.89	20.39	14.32	20.39	24.23	21.36	24.23
e_2	19.62	18.03	21.67	21.55	16.45	21.55	24.34	21.39	24.35
e_3	24.15	23.43	23.55	23.16	23.13	23.18	30.37	28.07	30.48
e_4	20.92	19.71	20.08	22.73	15.06	22.72	26.27	23.03	26.31

Table 3 SER (%) on Albayzin 2010 ASE cross-validation experiments

Experiment	Best	MV	r_F	r_{MI}	r_{PR}	r_{RE}	r_{AC}
e_1&e_2	19.41	19.93	20.34	20.08	**18.70**	**18.94**	**18.95**
e_1&e_3	19.41	23.67	20.77	20.20	**18.86**	20.76	22.37
e_1&e_4	19.41	22.35	21.74	20.57	**18.13**	**18.55**	19.70
e_2&e_3	19.62	23.86	21.07	20.64	**19.00**	20.84	22.48
e_2&e_4	19.62	22.48	21.94	20.80	**18.32**	**18.58**	19.71
e_3&e_4	20.92	23.50	22.70	21.75	**20.78**	22.12	23.81
e_1&e_2&e_3	19.41	18.90	**18.89**	**18.82**	**18.51**	**18.69**	**18.88**
e_1&e_2&e_4	19.41	18.66	18.74	**18.62**	**18.35**	**18.47**	**18.56**
e_1&e_3&e_4	19.41	19.71	19.56	19.41	**18.79**	**19.23**	20.03
e_2&e_3&e_4	19.62	19.71	**19.53**	**19.43**	**18.80**	**19.24**	20.05
e_1&e_2&e_3&e_4	19.41	20.03	19.73	**19.06**	**17.57**	**17.69**	**18.30**

Results in boldface show those ensemble classifiers that outperformed both the MV strategy and the best individual classifier

proposed in this paper. Specifically, r_F, r_{MI}, r_{PR}, r_{RE} and r_{AC} stand for F-score, mutual information, precision, recall and accuracy reliability estimates, respectively. These strategies are compared with a MV ensemble strategy, in order to see if the proposed techniques outperform this simple fusion approach.

Comparing Tables 3, 4 and 5, it is straightforward to see that the best reliability estimate of the five presented in this paper is the one based on the precision, namely r_{PR}; it succeeded to outperform both the best individual audio segmentation strategy and the MV strategy in 23 out of 33 fusion experiments. The recall-based reliability estimate, namely r_{RE}, succeeded to outperform the baseline system in 18 out of 33 fusion experiments. The

Table 4 SER (%) on Albayzin 2012 ASE cross-validation experiments

Experiment	Best	MV	r_F	r_{MI}	r_{PR}	r_{RE}	r_{AC}
e_1&e_2	16.71	17.05	**16.50**	**16.48**	**16.49**	**16.49**	17.05
e_1&e_3	17.15	21.06	17.49	17.40	**16.61**	18.00	20.02
e_1&e_4	17.15	19.96	**17.03**	**16.97**	**15.89**	17.48	19.18
e_2&e_3	16.71	20.79	17.31	17.28	17.03	17.62	18.69
e_2&e_4	16.71	19.74	16.92	16.75	**15.96**	17.17	18.32
e_3&e_4	20.32	21.85	20.51	20.40	**19.65**	21.92	22.19
e_1&e_2&e_3	16.71	16.59	**16.26**	**16.25**	**16.09**	**16.24**	**16.43**
e_1&e_2&e_4	16.71	16.39	**16.07**	**16.04**	**15.91**	**16.03**	**16.18**
e_1&e_3&e_4	17.15	18.61	18.18	18.10	17.75	18.30	18.55
e_2&e_3&e_4	16.71	18.48	18.03	18.10	17.65	18.15	18.38
e_1&e_2&e_3&e_4	16.71	17.86	**16.21**	**16.20**	**15.65**	**15.99**	16.95

Results in boldface show those ensemble classifiers that outperformed both the MV strategy and the best individual classifier

Table 5 SER (%) on Albayzin 2014 ASE cross-validation experiments

Experiment	Best	MV	r_F	r_{MI}	r_{PR}	r_{RE}	r_{AC}
e_1&e_2	20.62	21.09	20.99	20.83	**19.96**	**20.37**	20.80
e_1&e_3	20.98	25.16	21.09	20.95	**20.05**	22.69	25.19
e_1&e_4	20.98	22.84	21.84	21.14	**19.39**	**20.24**	22.19
e_2&e_3	20.62	25.12	21.01	20.90	**19.83**	22.38	24.87
e_2&e_4	20.62	23.00	21.60	21.00	**18.92**	**20.19**	21.74
e_3&e_4	22.60	25.72	23.81	23.62	**22.42**	23.75	25.65
e_1&e_2&e_3	20.62	20.59	**19.99**	**20.02**	**19.52**	**19.76**	**20.27**
e_1&e_2&e_4	20.62	19.91	**19.67**	**19.65**	**18.93**	**19.10**	**19.60**
e_1&e_3&e_4	20.98	22.55	21.62	21.19	20.96	21.32	22.27
e_2&e_3&e_4	20.62	22.65	21.68	21.36	20.95	21.30	22.21
e_1&e_2&e_3&e_4	20.62	21.07	**19.72**	**19.57**	**18.91**	**19.62**	20.65

Results in boldface show those ensemble classifiers that outperformed both the MV strategy and the best individual classifier

worst performing reliability estimate was the accuracy, which only outperformed the baseline systems in 8 out of 33 experiments. It must be noted that, although the F-score reliability estimate is defined as a combination of r_{PR} and r_{RE}, its performance was poorer than those obtained by these reliability estimates individually.

Observing the results in function of the number of audio segmentation strategies included in the ensemble, it can be seen that, in general, the fusion of two systems is only successful when using the precision and recall reliability estimates. The experiment e_3&e_4 is specially noticeable, as in this case the two worst-performing systems are fused, and an improvement of the performance of the individual systems is only achieved with r_{PR}. The experiments where three systems are fused show a clear pattern; almost all the combinations of two good segmentation strategies, namely e_1 and e_2, with one of the worst-performing systems are successful no matter which reliability estimate is used; however, when the two worst performing systems are fused with one of the best-performing systems, ensemble performance is worst than individual performance in two out of three experimental frameworks (an improvement is obtained in Albayzin 2010 when using some reliability estimates). The fusion of four systems is generally successful, except in the case of the accuracy based reliability estimate, which only outperformed the baselines in Albayzin 2010 scenario.

Figures 5, 6 and 7 show the error matrices obtained when performing the fusion of four audio segmentation systems, both when using the individual systems and the fusion strategies. The first conclusion extracted from these matrices is that the majority voting strategy, which did not outperform the best individual system in any experimental framework, is not able to improve the accuracy of the overlapping classes when the accuracy of the individual systems is poor. The accuracy-based reliability obtained a remarkable improvement of accuracy of class sp+mu+no in Albayzin 2012 and Albayzin 2014 experiments but, in the latter scenario, the deterioration of the performance in other classes led to the overall performance to be below the baseline. The strategies that obtained the best overall performance, namely r_{PR} and r_{RE}, achieved noticeable improvements in some classes, which compensate slight reductions of accuracy in other classes.

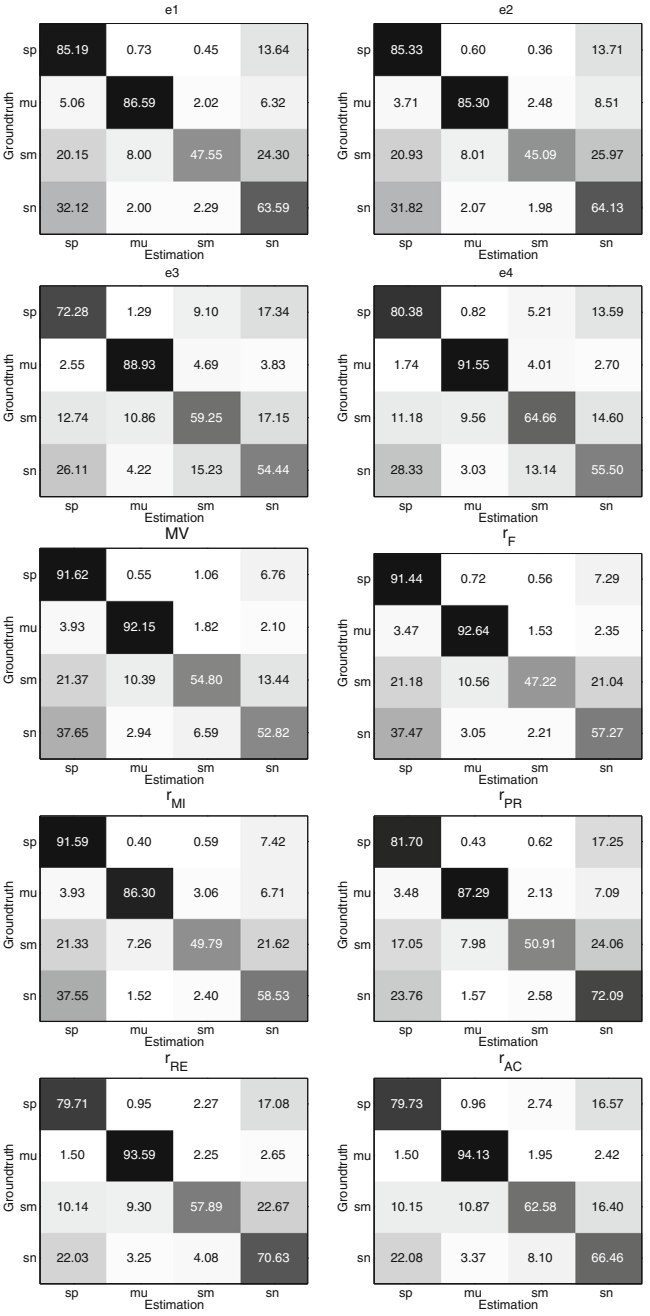

Fig. 5 Classification errors (%) between groundtruth and estimations in Albayzin 2010 ASE cross-validation experiments when classifying the target classes speech (sp), music (mu), speech with music (sm) and speech with noise (sn). The different error matrices correspond to the audio segmentation results obtained using the individual audio segmentation strategies e_1, e_2, e_3, e_4, and when performing fusion of these four systems using the majority voting strategy (MV) and the proposed fusion strategy with different reliability estimates r_F, r_{MI}, r_{PR}, r_{RE}, r_{AC}. Note that each row is normalized to 100 %

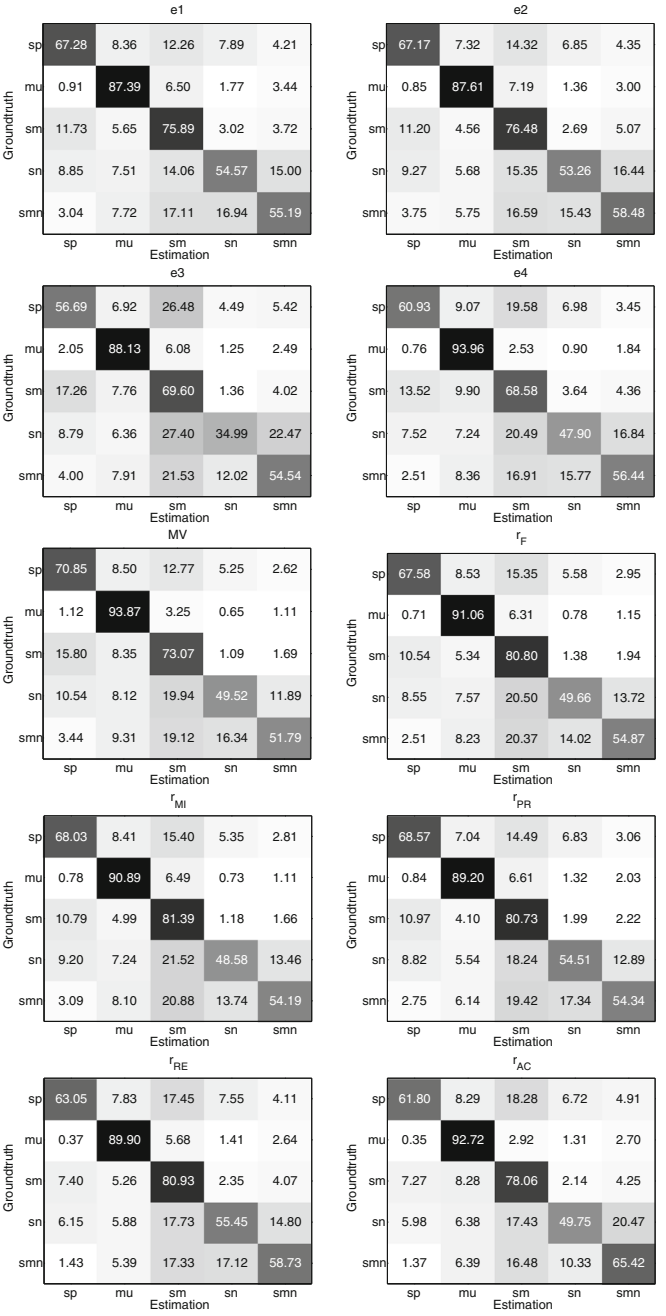

Fig. 6 Classification errors (%) between groundtruth and estimations in Albayzin 2012 ASE cross-validation experiments when classifying the target classes speech (sp), music (mu), speech with music (sm), speech with noise (sn) and speech with music and noise (smn). The different error matrices correspond to the audio segmentation results obtained using the individual audio segmentation strategies e_1, e_2, e_3, e_4, and when performing fusion of these four systems using the majority voting strategy (MV) and the proposed fusion strategy with different reliability estimates r_F, r_{MI}, r_{PR}, r_{RE}, r_{AC}. Note that each row is normalized to 100 %

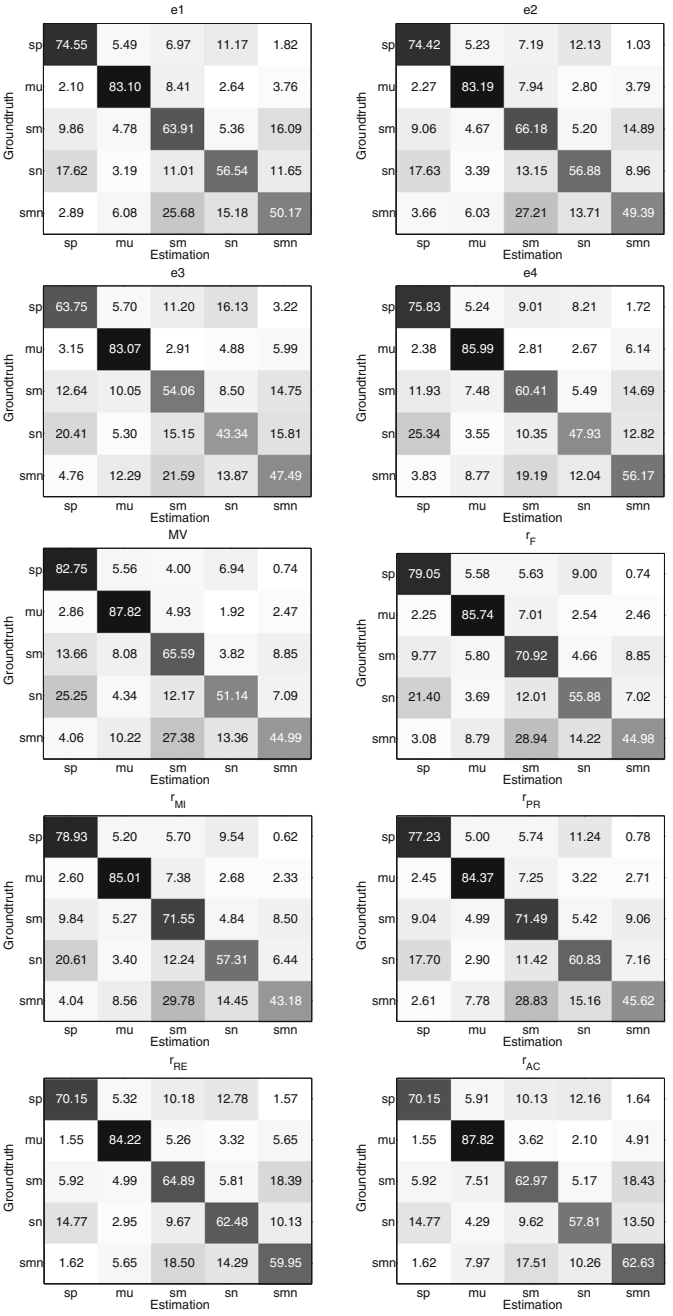

Fig. 7 Classification errors (%) between groundtruth and estimations in Albayzin 2014 ASE cross-validation experiments when classifying the target classes speech (sp), music (mu), speech with music (sm), speech with noise (sn) and speech with music and noise (smn). The different error matrices correspond to the audio segmentation results obtained using the individual audio segmentation strategies e_1, e_2, e_3, e_4, and when performing fusion of these four systems using the majority voting strategy (MV) and the proposed fusion strategy with different reliability estimates r_F, r_{MI}, r_{PR}, r_{RE}, r_{AC}. Note that each row is normalized to 100 %

8.2 Evaluation experiments

Table 6 shows the results achieved when performing the evaluation experiments of Albayzin 2010, 2012 and 2014 ASE. The SER for the train and test datasets is shown in the Table in order to find out whether the top-performing reliability estimate in the training dataset is also the top-performing in the test dataset, which would allow to decide which reliability estimate is more suitable for a given database. It must be noted that the train experiments are slightly biased, as the reliability estimates are applied in the same data that was used to compute them. Nevertheless, the Table shows that the results obtained on the training dataset succeed at predicting which reliability estimate would achieve the best performance in the test dataset. It must be noted that, in the Albayzin 2010 experiment, r_{PR} achieves the best performance in train but it is r_{RE} the one that achieves the best performance in test; however, the difference in SER between these two reliability estimates is negligible in both datasets, so any of them would be suitable for this experiment.

Table 6 also shows a comparison between the proposed strategies, the best individual classifier, and other well known ensemble classification techniques; the Table shows that the ensemble classifiers based on reliability estimates outperform these other approaches. These results suggest that assigning reliabilities in function of the class performs better than assigning a weight to the individual systems without taking into account the individual results achieved for each class as in the WMV scheme. NB takes into account the performance of each class but, theoretically, this method is only optimal when it is assumed that all the classes are equiprobable [19], and this assumption is not correct in this specific case.

Table 7 shows the SER achieved by the different systems that participated in Albayzin 2012 and 2014 ASE. Comparing the results achieved with the proposed fusion techniques to those obtained by the participants in the evaluation, it can be seen that both the individual audio segmentation systems and the ensemble strategies proposed in this work obtain state-of-art performance. Indeed, the SER achieved in Albayzin 2012 ASE is much lower than those obtained by the participants in the evaluation; results achieved by the fusion strategies

Table 6 SER (%) on Albayzin 2010, 2012 and 2014 ASE evaluation experiments using the proposed reliability estimates, the best individual system (Best), majority voting (MV), weighted majority voting (WMV) and Naive Bayes (NB)

Ensemble	Albayzin 2010		Albayzin 2012		Albayzin 2014	
	train	test	train	test	train	test
Best	18.03	20.08	14.32	20.39	21.36	24.23
MV	19.12	21.86	13.72	19.27	21.85	25.35
WMV	16.88	18.78	13.70	19.90	21.13	24.12
NB	17.00	20.46	12.97	19.06	20.09	22.95
r_F	18.56	21.65	12.39	19.04	21.14	24.76
r_{MI}	17.50	20.76	**11.93**	**18.91**	20.93	24.68
r_{PR}	**15.83**	18.61	12.10	18.98	**19.23**	**22.84**
r_{RE}	16.01	**18.49**	13.38	19.29	20.23	23.14
r_{AC}	17.22	19.21	13.85	19.36	21.22	23.66

Results in boldface show the top performing strategy for each experiment

Table 7 SER (%) achieved by the participants in Albayzin 2012 and 2014 ASE

System	Albayzin 2012	Albayzin 2014
1	26.34	20.68
2	33.30	31.59
3	40.01	33.93
4	39.55	20.80
5	26.53	29.13
6	28.12	22.52
7		30.67

in Albayzin 2014 ASE do not outperform the best systems submitted by the participants, but still they achieved better results than most of the participants' systems.

Results achieved by participants in Albayzin 2010 cannot be compared with those obtained in these experiments because, as mentioned in Section 7.4, the metric used in Albayzin 2010 ASE was not the SER, so results are not comparable. Nevertheless, according to [4], the best-performing system of this evaluation obtained a SER of 19.3 %, which is below the performance obtained by the precision-based and recall-based reliability estimates.

9 Discussion

The experimental results presented in Section 8 left some issues to take into consideration.

As defined in Section 4, r_F is defined as the harmonic mean of r_{PR} and r_{RE}, expecting r_F to have a better performance than r_{PR} and r_{RE}, as it takes two types of error into account at the same time. However, the experimental validation showed that precision and recall achieve better results when used individually, suggesting that the F-score is not the most suitable strategy to combine precision and recall in this ensemble classification scenario.

The results shown in Table 6 show that it is possible to select the best reliability estimate performing a previous fusion experiment in the training data, as in general the reliability estimate that obtained the best performance in the training dataset was the one that obtained the best performance in test as well. This did not happen in Albayzin 2010 evaluation experiment but, as mentioned above, the difference in performance between the reliability estimate selected in train and the one that performed better in test is negligible. Another mismatch in the results can be observed in Albayzin 2012 experiments, as this was the only experiment in which the top performing reliability estimate was not the same in the cross-validation and in the evaluation experiment. This might be caused by the distribution of the classes in the different folds of the cross-validation experiment because, as shown in Fig. 3, it is very changeable.

As mentioned in Section 2, the diversity of the classifiers, which represent how different the individual classifiers are when classifying the different examples, is an important fact to be taken into account when designing ensembles of classifiers [26]. The idea behind quantifying the diversity of the ensembles of classifiers is that there is no point in combining classifiers that make the same error on the same examples. Nevertheless, what actually has a paramount importance in the proposed ensemble strategy is the diversity of the reliability estimates for the different classes and individual systems. Let us consider an ensemble of classifiers and a reliability estimate; if the reliability of all the individual classifiers e_i when

classifying an example of class c_j is very similar, we have a system that tends to a majority voting strategy (which is equivalent to always having the same weight). Hence, what matters when dealing with reliability estimates is how diverse the reliabilities of the different classifiers and weights are. In fact, when performing a qualitative analysis of the proposed reliability estimates obtained from the confusion matrices of the individual classifiers, it can be observed that reliabilities are more diverse, both in terms of classifiers and classes, when using the precision estimate; this can explain why, in these specific experimental frameworks, performance is, in general, better when using the reliability estimate based on the precision. Nevertheless, in other scenarios or when using different individual classifiers, it is likely that other reliability estimate leads to a greater diversity and, hence, obtains a better performance.

10 Conclusions and future work

This paper presented a framework to perform decision-level fusion of audio segmentation outputs, based on the premise that there is strength in numbers, so a fusion strategy would help to dim the weaknesses of the different systems as well as enhancing their strengths. The fusion approach we proposed consisted in estimating the reliability of each audio segmentation strategy when classifying each of the possible classes, and using this reliability estimate as weights in a weighted majority voting fusion strategy. The estimation of the reliability was carried out by computing the confusion matrices of the different audio segmentation systems when classifying a set of examples from a development dataset. We proposed different reliability estimates, namely precision, recall, F-score, accuracy and mutual information.

The validity of the proposed fusion technique for audio segmentation was assessed in three different experimental frameworks based on radio and television programmes, which were defined for Albayzin 2010, 2012 and 2014 audio segmentation evaluations. The experimental results showed that an improvement of the audio segmentation results can be obtained by using this fusion paradigm, and a comparison to a majority voting scheme proved that the use of the reliability estimates as weights is advantageous for this task.

The reliability estimate that showed the best performance of the five proposed in this paper was the one based on the precision, followed by the one based on the recall. The reliability estimates based on accuracy, F-score and mutual information were able to outperform the baseline systems in some of the experiments, but they did not achieve a good overall performance when compared to precision and recall. Nevertheless, the mutual information-based reliability estimate showed promising results, so its formulation will be revisited in future work in order to try to improve the results presented in this paper.

The experimental validation suggested that, although not all the reliability estimates perform the same in different datasets, it is possible to select the most suitable one by performing a fusion experiment in the training data. This allows to empirically predict which reliability estimate will have the best performance beforehand. Nevertheless, the theoretical formulation of the proposed fusion strategy suggests that variability of the reliability estimates for different classifiers and classes is crucial in this fusion framework: if the reliabilities of the different classifiers and classes are very similar, our fusion strategy tends to a majority voting scheme, strategy that showed a poor performance in the experimental validation. Moreover, as mentioned in Section 9, the precision based reliability estimate seemed to be the most diverse in these experiments, fact that reinforces the assumption “more diversity

in the reliabilities leads to a better performance"; hence, in further work we will try to define a metric to quantize the goodness of different reliability estimates according to the data to be fused, in order to be able to know beforehand which reliability estimate will give the best fusion results without having to select it empirically.

The proposed experimental validation consisted in computing the confusion matrix using the training data of a given database and then using it to fuse the test outputs of the same database; nevertheless, it would be interesting to find out if it is possible to obtain the confusion matrix in a dataset and using it in a different one. In this way, the amount of training data as well as the variability of the data itself would increase, leading to a confusion matrix that reflects the performance of the audio segmentation strategy in a more general scenario. Thus, in future work, cross-corpus experiments will be performed in order to explore this possibility.

Lastly, we would like to emphasize the fact that the fusion strategy proposed in this paper is not only a strategy to fuse audio segmentation systems but a framework to design different fusion strategies that can be used for any classification task or pattern recognition problem. The two design decisions to be made in this system are which reliability estimate to use, as they can be different to the ones we proposed, and how to combine the different classifiers. We proposed a fusion strategy based on weighted majority voting, but in the future we plan to assess other strategies that make use of the reliability estimates in order to make a common decision.

Acknowledgments This work has been supported by the European Regional Development Fund, the Galician Regional Government (GRC2014/024, 'Consolidation of Research Units: AtlantTIC Project' CN2012/160) and the Spanish Government ('SpeechTech4All Project' TEC2012-38939-C03-01).

References

1. Anguera X, Hernando J (2004) XBIC: Nueva Medida para segmentación de locutor hacia el indexado automático de la señal de voz. In: III Jornadas en tecnología del habla, 237–242
2. Butko T, Nadeu C (2011) Audio segmentation of broadcast news in the albayzin-2010 evaluation: overview, results, and discussion. EURASIP Journal on Audio, Speech and Music Processing 2011(1)
3. Butko T, Nadeu C, Schulz H (2010) Albayzin-2010 audio segmentation evaluation: Evaluation setup and results. In: Proceedings of FALA 2010 - VI jornadas en tecnología del habla and II iberian SLTech workshop, 305–308
4. Castan D, Ortega A, Miguel A, Lleida E (2014) Audio segmentation-by-classification approach based on factor analysis in broadcast news domain. EURASIP Journal on Audio, Speech and Music Processing 2014(34)
5. Castanedo F (2013) A review of data fusion techniques. Sci World J:2013
6. Cettolo M, Vescovi M (2003) Efficient audio segmentation algorithms based on the BIC. In: Proceedings of ICASSP VI, 537–540
7. Cho S, Kim J (1995) Multiple network fusion using fuzzy logic. IEEE Trans Neural Netw 6(2):497–501
8. Comon P (1994) Independent component analysis - a new concept? Signal Process 36:287–314
9. Delacourt P, Kryze D, Wellekens CJ (2000) DISTBIC: a speaker-based segmentation for audio data indexing. Speech Comm 32(1-2):111–126
10. Do CT, Barras C, Lee VB, Sarkar AK (2013) Augmenting short-term cepstral features with long-term discriminative features for speaker verification of telephone data. In: Proceedings of interspeech, 2484–2488
11. Franco-Pedroso J, Gomez-Rincon E, Ramos D, Gonzalez-Rodriguez J (2014) ATVS-UAM system description for the albayzin 2014 audio segmentation evaluation. In: Proceedings of iberspeech 2014: VIII jornadas en tecnología del habla and IV iberian SLTech workshop, 247–252

12. Gunatilaka AH, Baertlein BA (2001) Feature-Level And Decision-Level fusion of noncoincidently sampled sensors for land mine detection. IEEE Trans Pattern Anal Mach Intell 23(6):577–589
13. Hall M (1998) Correlation-based feature subset selection for machine learning. Ph.D. Thesis, University of Waikato, Hamilton, New Zealand
14. Huang YS, Suen CY (1993) The Behavior-Knowledge space method for combination of multiple classifiers. In: Conference on Computer Vision and Pattern Recognition (CVPR), pp 347–352
15. Kasapoglu NG, Anfinsen SN, Eltoft T (2012) Fusion of optical and multifrequency PolSAR data for forest classification. In: Proceedings of IEEE International Geoscience and Remote Sensing Symposium (IGARSS), pp 3355–3358
16. Kittler J, Hatef M, Duln P, Matas J (1998) On combining classifiers. IEEE Trans Pattern Anal Mach Intell 20(3):226–239
17. Koa AH, Sabourina R, de Souza Britto Jr. A, Oliveira L (2007) Pairwise fusion matrix for combining classifiers. Pattern Recogn 40(8):2198–2210
18. Kuncheva LI (2004) Combining pattern classifiers: methods and algorithms. Wiley-Science
19. Kuncheva L, Rodriguez J (2014) A weighted voting framework for classifiers ensembles. Knowl Inf Syst 38(2)
20. Littlestone N, Warmuth M (1994) Weighted majority algorithm. Inf Comput:212–261
21. Lopez-Otero P, Docio-Fernandez L, Garcia-Mateo C (2014) GTM-UVIgo System for Albayzin 2014 Audio Segmentation Evaluation. In: Proceedings of iberspeech 2014: VIII jornadas en tecnología del habla and IV iberian SLTech workshop, 253–262
22. Meinedo H, Neto J (2005) A Stream-Based audio segmentation, classification and clustering Pre-Processing system for broadcast news using ANN models. In: Proceedings of interspeech, 237–240
23. Metze F, Rawat S, Wang Y (2014) Improved audio features for Large-Scale multimedia event detection. In: IEEE International conference on multimedia and expo, ICME, 1–6
24. Molina L (2002) Feature selection algorithms: a survey and experimental evaluation. In: Proceedings of IEEE international conference on data mining, 306–313
25. Ortega A, Castan D, Miguel A, Lleida E (2014) The albayzin 2014 audio segmentation evaluation. In: Proceedings of iberspeech: VIII jornadas en tecnología del habla and IV iberian SLTech workshop, 283–289
26. Polikar R (2006) Ensemble based systems in decision making. IEEE Circuits Syst Mag 6(3):21–45
27. Ramona M, Richard G (2009) Comparison of different strategies for a SVM-based audio segmentation. In: Proceedings of the european signal processing conference (EUSIPCO)
28. Rodriguez-Fuentes L, Penagarikano M, Varona A, Diez M, Bordel G (2012) GTTS Systems for the albayzin 2012 audio segmentation evaluation. In: Proceedings of iberspeech 2012: VII jornadas en tecnología del habla and III iberian SLTech workshop, 590–595
29. Ross A, Govindarajan R (2005) Feature level fusion using hand and face biometrics. In: Proceedings of SPIE conference on biometric technology for human identification II 5779, 196–204
30. Rybach D, Gollan C, Schlüter R, Ney H (2009) Audio segmentation for speech recognition using segment features. In: Proceedings of IEEE international conference on acoustics, speech and signal processing (ICASSP), 4197–4200
31. Schuller B, Metze F, Steidl S, Batliner A, Eyben F, Polzehl T (2010) Late fusion of individual engines for improved recognition of negative emotion in speech - learning vs. democratic vote. In: Proceedings of IEEE international conference on acoustics, speech and signal processing (ICASSP), 5230–5233
32. Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6:461–464
33. Seyerlehner K, Pohle T, Schedl M, Widmer G (2007) Automatic music detection in television productions. In: Proceedings of the 10th international conference on digital audio effects (DAFx-07)
34. Shafer G (1976) A mathematical theory of evidence. Princeton University Press, Princeton
35. Silvestre-Cerdà J, Giménez A, Andrés-Ferrer J, Civera J, Juan A (2012) Albayzin evaluation: the PRHLT-UPV audio segmentation system. In: Proceedings of iberspeech: VII jornadas en tecnología del habla and III iberian SLTech workshop, 596–600
36. Sokolova M, Lapalme G (2009) A systematic analysis of performance measures for classification tasks. Inf Process Manag 45(4):427–437
37. Tao Q, Veldhuis R (2009) Threshold-optimized decision-level fusion and its application to biometrics. Pattern Recogn 42:823–836
38. Tavarez D, Navas E, Alonso A, Erro D, Saratxaga I, Hernaez I (2014) Aholab audio segmentation system for albayzin 2014 evaluation campaign. In: Proceedings of iberspeech 2014: VIII jornadas en tecnología del habla and IV iberian SLTech workshop, 273–282
39. Tulys P, Akkermans A, Kevenaar T, Schrijen G, Bazen A, Veldhuis R (2005) Practical biometric authentication with template protection. In: Proceedings of 5th international conference on audio- and video-based personal authentication, 436–446

40. Tzanetakis G (2002) Manipulation, analysis and retrieval systems for audio signals. Ph.D. Thesis, Princeton University
41. Young SJ, Kershaw D, Odell J, Ollason D, Valtchev V, Woodland P (2006) The HTK book version 3.4, Cambridge University Press

Paula Lopez-Otero was born in A Coruña. She received the degree in Telecommunication Engineering and the PhD degree in 2008 and 2015, respectively, both from the Universidade de Vigo. During her PhD, advised by Carmen García Mateo and Laura Docío Fernández, she was a visiting researcher at University of Surrey (United Kingdom) and Technical University of Munich (Germany). She is currently a postdoctoral researcher at the Multimedia Technology Group. Her research interests focus on audio and speaker segmentation and clustering, spoken term detection and emotional state detection, fields in which she has published several research papers in international conference proceedings. She has participated in several evaluations related to these research topics, being part of the winning team of Albayzin 2014 Search on Speech Evaluation.

Laura Docio-Fernandez received her MSc degree and PhD in Telecommunications Engineering from the University of Vigo (Spain) in 1995 and 2001, respectively. She has participated in more than 15 research projects funded by national or international public institutions and companies. She is the author of more than 40 papers published in international conference proceedings. In 2002 she was a Postdoctoral Fellow in the International Computer Science Institute (ICSI) of Berkeley, USA. She is currently an Associate Professor in the Department of Signal Theory and Communications at the University of Vigo, Spain, and a member of the Multimedia Technologies Group (GTM). Her research interests lie in the broad field of speech and audio processing, especially the analysis, modelling and recognition of speech, speaker and audio signals in general. She has participated in several Albayzin Evaluation Campaigns organized by Spanish National Network on Speech Technology; specifically, she has participated in Albayzin 2010 and 2012 Audio Segmentation Evaluations, in Albayzin Speaker Diarization Evaluation and in Albayzin 2010 and 2012 Language Recognition Evaluations.

Carmen Garcia-Mateo (female) (M'87) received the Ph.D. degree in telecommunications engineering from the Technical University of Madrid, Vigo, Spain, in 1993. She is currently Full Professor, teaching digital signal processing and biometrics. Her research interests include signal and speech-based biometrics, audio segmentation and extraction of metadata from audiovisual contents. She is the Head of the GTM group and has been the leader of more than 30 research projects and contracts (She has been the leader of Spanish national projects on Speech Technologies and Biometric authentication). She has authored more than 100 research papers and advised 10 PhD thesis. She has served in many technical program committees. She served as Vice-Chancellor for Studies and EHEA Harmonization of the University of Vigo from 2006 to 2010. She is the Chair of the Signal Theory and Communications Department of the University of Vigo. She is the recipient of the 2014 "Maria Josefa Wonenburger" Award for her career trajectory.