Random Regression Forests for Acoustic Event Detection and Classification

Huy Phan, *Student Member, IEEE*, Marco Maaß, *Student Member, IEEE*, Radoslaw Mazur, *Member, IEEE*, and Alfred Mertins, *Senior Member, IEEE*

***Abstract*—Despite the success of the automatic speech recognition framework in its own application field, its adaptation to the problem of acoustic event detection has resulted in limited success. In this paper, instead of treating the problem similar to the segmentation and classification tasks in speech recognition, we pose it as a regression task and propose an approach based on random forest regression. Furthermore, event localization in time can be efficiently handled as a joint problem. We first decompose the training audio signals into multiple interleaved *superframes* which are annotated with the corresponding event class labels and their displacements to the temporal onsets and offsets of the events. For a specific event category, a random-forest regression model is learned using the displacement information. Given an unseen superframe, the learned regressor will output the continuous estimates of the onset and offset locations of the events. To deal with multiple event categories, prior to the category-specific regression phase, a superframe-wise recognition phase is performed to reject the background superframes and to classify the event superframes into different event categories. While jointly posing event detection and localization as a regression problem is novel, the superior performance on two databases ITC-Irst and UPC-TALP demonstrates the efficiency and potential of the proposed approach.**

***Index Terms*—Acoustic event detection, random forest, regression forest, superframe.**

I. Introduction

ACOUSTIC event (AE) classification and detection are important for many real-world applications such as ambient assisted living [1], security surveillance [2], meeting room transcription [3], [4], human-computer interaction [5]–[7], multimedia retrieval [8], and "machine hearing" [9] to name a few. It has been under great attention of the research community with many recent evaluation campaigns including CLEAR 2006 [10], CLEAR 2007 [11], and AASP CASA 2013 [12]. Acoustic event classification (AEC), which performs on segmented AEs, can be readily addressed with a large number of off-the-shelf classifiers and acoustic features [5]–[7], [13]. Compared to AEC, acoustic event detection (AED) is a more interesting, yet more difficult task, because we need to determine not only the identity of the sounds but also their positions in time. Up to now, the AED problem has been still largely unsolved. It is challenging due to large intra-class variations in terms of event durations and sounds, nonstationary background noise, as well as event overlap. Furthermore, for some applications (such as content-based multimedia indexing/retrieval, meeting-stage detection, etc.), it is vital to have a good temporal resolution of the detected AEs, i.e. localization problem. To the best knowledge of the authors, this problem has not been explicitly addressed in the literature.

Inspired by the success of speech recognition, the automatic speech recognition (ASR) framework [14] has been adapted for AED [11], [15]–[17]. This method can be divided into three stages. First, local features, e.g. Mel-frequency cepstral coefficients (MFCCs) [18], are extracted from small frames. The local feature vectors are then modeled by Gaussian Mixture Models (GMMs). Finally, the distributions of the feature vectors are learned given the feature vector sequences and the state sequences using Hidden Markov Models (HMMs). On testing, given an unseen feature vector sequence, the event is recognized with the maximum posterior probability. The ASR framework works well for speech in practice, but the results on AED have not been satisfactory [10]–[12]. First, unlike speech, the underlying sound event information is less structured, particularly as no sub-word dictionary exists in the same way as for languages. Moreover, while frame-based acoustic features are reliable for speech, AEs contain a wider range of characteristic and nonstationary effects which may not be captured in such frame-based features. Regarding to temporal localization, i.e. event boundary determination, the HMM-based sequence models cannot generalize well over highly variable durations which are usually the case for audio events. This is understandable since they rely on limited-duration models that assume exponentially distributed duration probabilities of each state.

Another common approach is based on a detection-by-classification scheme [3], [19]–[21]. This approach extracts global presentations for isolated events in training data. Classification models, e.g. Support Vector Machines (SVMs), are then trained to distinguish the events from background as well as classify them into different classes. Finally, the learned classifiers are used to detect AEs in continuous audio signals by sliding

Manuscript received April 28, 2014; revised October 16, 2014; accepted October 18, 2014. Date of publication November 07, 2014; date of current version January 14, 2015. This work was supported by the Graduate School for Computing in Medicine and Life Sciences funded by Germany's Excellence Initiative [DFG GSC 235/1]. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Bozena Kostek.

H. Phan and M. Maaß are with the Institute for Signal Processing and the Graduate School for Computing in Medicine and Life Sciences, University of Lübeck, 23538 Lübeck, Germany (e-mail: phan@isip.uni-luebeck.de; maasz@isip.uni-luebeck.de).

R. Mazur and A. Mertins are with the Institute for Signal Processing, University of Lübeck, D-23562 Lübeck, Germany (e-mail: mazur@isip.uni-luebeck.de; mertins@isip.uni-luebeck.de).

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the authors. This includes one AVI format movie clip, which is a demonstration of event onset and offset scoring for spoon_cup_jingle events. This material is 2.85 MB in size.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASLP.2014.2367814

window fashion. Typically, the audio segments need to be long enough, like one second long, in order to capture sufficient signal distribution so that they can be recognized individually. A post-processing step, e.g. median filter [19] or majority voting [22], is also necessary to smooth the intermittent label sequence. Although this approach is intuitive and straightforward to implement, it confronts one with two unsolved problems. First, these systems heavily depend on the quality of the classification models, which are far from perfect in practice. The noisy segmentation/classification results are considered as detection hypotheses and contribute to the detection error. Secondly, for the localization task, using long windows results in low temporal resolution of the detected AEs. In general, although this approach shows good performance on the AEC task, it is less efficient for the AED task compared to the HMM-based ASR framework [10].

The proposed system is able to overcome the above mentioned problems. It differs from the majority of contributions in the field in that it considers the joint problem of AE detection and localization as a regression problem and uses a random forest regression framework [23], [24] to deal with it. Motivated by the success of regression forests in various computer vision tasks, we adapt it for the AED task. We take advantage of the acoustic *superframes* proposed in [22], [25], which are able to be recognized independently at an acceptable accuracy. The training audio signals, containing multiple AE occurrences of different categories, are firstly divided into multiple interleaved superframes. Each superframe is associated with a class label and a two-dimensional displacement vector to the onset and offset of the corresponding AE. Thereafter, using the displacement vectors, category-specific regression forests are trained to map each event superframe to the continuous estimates of onset and offset locations of the events in time, i.e., we consider a multi-variate, continuous parameter estimation problem. In order to handle multi-class detection, before category-specific regression is performed, two classification models are learned using random forest classification [26]: one of them is to distinguish event superframes from background superframes and the other is to subsequently classify event superframes into different categories of interest. On testing, the learned classifiers are applied to recognize event superframes which are finally inputted to the category-corresponding regressor to detect and localize the events from test audio signals. We will show that our approach significantly outperforms the common competitive approaches in terms of detection error rate on two databases ITC-Irst and UPC-TALP. Besides that, by inducing the continuous estimates of event boundaries, the proposed system is invariant to event temporal scales.

In summary, our contributions are three-fold: (i) the formulation of the joint AE detection and localization as a regression problem; (ii) the development of a category-specific random forest architecture and learning method that leverages the random forest regression framework in order to detect and localize AEs in time; and (iii) advance the state-of-the-art significantly on the two databases ITC-Irst and UPC-TALP, decreasing the detection error rate by more than six percent and ten percent, respectively.

The rest of this paper is organized as follows. Some related works on AED are briefly presented in Section II. After that, we describe our algorithm to learn the multivariate regression forests in Section III and our AE detection and localization system in Section IV. The experimental setup and results are presented in Section V followed by the conclusion and future works in Section VI.

II. Related Works

The previous works on AED can be mentioned with different aspects. From the algorithmic viewpoint, two dominant trends have been seen. The first was based on HMMs with various topologies [4], [15], [16]. The detection task was accomplished in two ways: (1) the HMM-based events/background segmentation followed by the HMM event classification and (2) merging the segmentation and classification in one step with the standard ASR framework. The other trend exploits discriminative classifiers for both events/background segmentation and subsequent event classification [3], [19]. Beside SVMs, some other classification algorithms were also used, such as Gaussian Mixture Models (GMM) [27], Adaboost [28], and random forest [22]. In the recent international evaluation campaigns [10]–[12] for AED, most of the submission systems pursued these common directions. In another approach, by considering an AE as structured sequence of acoustic units [29] or I-vectors [30], the AE instances can be directly segmented from the audio signals.

In the work of Stork *et al.* [7] the events are modeled as ensembles of event frames. For every event category, the event instances in the training data are divided into multiple frames each of which maintains its displacement to the corresponding event center. The frames are then clustered using k-means to form category-specific codebooks. On testing, a frame recognized as event is matched to a learned codebook. Finally, the displacements of the frames stored in the codebook are used to vote for the event center. Their goal is to find the event centers under the assumption that all category-specific events are equal in duration to ease the localization. Yet, in practice, some categories experience large variations of intra-class duration. Furthermore, the model in [7] is data-based, requiring a large memory for storage. These drawbacks hinder this approach in many cases.

Regarding the representations, the traditional features for speech recognition like MFCCs [18] and log frequency filter bank parameters [3] have been prevalent. Various other features have also been developed and found useful for AED, for instance, spectro-temporal features based on spectrograms [31], [32], dictionaries induced by non-negative matrix factorization (NMF) [12], event exemplar-based features [33]. It is also worth mentioning that the works on relevant feature selection [4], [34] reported significant improvement on AED.

The target environments also get involved. The reason is that different environments (for example, kitchen rooms [7], bathrooms [35], car inside space [36], and meeting-rooms [3], [4]) may significantly vary in background noise characteristics, event overlapping, overlapping with speech, etc., and require tailored strategies to deal with. Further, multi-source [11], [37] and multi-modal fusion [38], [39], when available, can be utilized to cope with the ambient noise as well as compensate for low SNR events.

In this article, we tackle the joint AED/L problem with single-channel non-overlapped AEs in meeting-room environment using random regression forests. We firstly decompose the event instances into superframes which are associated with

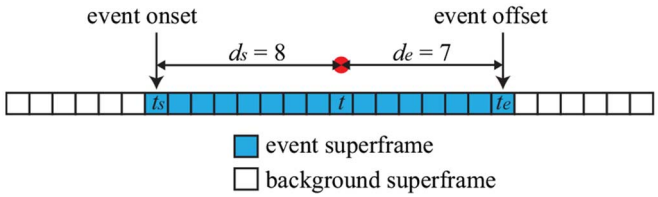

Fig. 1. Displacements of the superframe at the time index t to the onset t_s and the offset t_e of an AE.

their displacements to event onsets and offsets. Following the common extremely randomized trees framework [40], the random regression forests are constructed for every category. In testing, an event superframe, which is inputted into the corresponding regression forest, provides continuous estimates for event onset and offset positions. Event temporal scales are well handled in the proposed approach since we implicitly capture them in the regression forest models.

III. Multivariate Random Forest Regression

A. Random Forest Regression

A regression forest is an ensemble of different regression trees. Each of them plays the role of a nonlinear mapping from complex input spaces into continuous output spaces. The nonlinearity is achieved by dividing up the original problem into smaller ones, solvable with simple models. A split node in the tree maintains a test that is applied to a data sample to send it toward the left or the right child node. The tests are picked by some criteria to group the training samples into clusters where a good prediction can be achieved by simple models. These models are computed from the annotated data samples that reached the leaves and were stored there. While overfitting likely happens for a standard decision tree alone, an ensemble of randomly trained trees enjoys high generalization power [40].

B. Training

The training of our regressors is supervised and category-specific. Given a set of annotated superframes $\mathcal{S}^c = \{(\mathbf{x}_i, c, \mathbf{d}_i)\}$ of an event category $c \in \{1, \ldots, C\}$, each superframe $\mathbf{x} \in \mathbb{R}^M$ is associated with the class label c and a displacement vector $\mathbf{d} = (d_s, d_e) \in \mathbb{R}^2_+$. Here, M is the dimensionality of feature space and C denotes the number of event categories of interest. The values d_s and d_e, respectively, represent the displacements (in superframes) of the current superframe at the time index t to the onset t_s and offset t_e of the corresponding event, given as:

$$d_s = t - t_s, \tag{1}$$

$$d_e = t_e - t. \tag{2}$$

The displacement notations are illustrated in Fig. 1. Since we do not use the class label c for training category-specific regression forests, it can be safely ignored in this section. Our aim is to learn the clustering of superframes based on their features and their confidence in predicting the onsets and offsets of the events.

Generally, the tree construction for regression forests follows the common extremely randomized trees framework [40]. Each tree T in the forest $\mathcal{T} = \{T_i\}$ is constructed from a subset of superframes $S^c_T = \{(\mathbf{x}_i, \mathbf{d}_i)\}$ randomly sampled from $\mathcal{S}^c$. Starting from the root node, at each split node a large set of possible binary tests is randomly generated. A binary test $t_{f,\tau}$ on a data sample $(\mathbf{x}, \mathbf{d})$ is defined as

$$t_{f,\tau}(\mathbf{x}) = \begin{cases} 1, & \text{if} \quad \mathbf{x}^f > \tau \\ 0, & \text{otherwise,} \end{cases} \tag{3}$$

where $\mathbf{x}^f$ indicates the value of $\mathbf{x}$ at the feature channel $f \in \{1, \ldots, M\}$, and τ is a threshold. During the construction of the tree, at each split node, a pool of binary tests is generated with a randomly selected feature channel f and random values for τ generated in the range of $\mathbf{x}^f$. In our implementation, 20,000 random binary tests were considered for each split node. A test is selected from this pool to split the set of training samples S_l at a split node l into two sets: $S^{\text{right}}_l = \{(\mathbf{x}^{\text{right}}_i, \mathbf{d}^{\text{right}}_i)\}$ containing those samples satisfying the test and $S^{\text{left}}_l = \{(\mathbf{x}^{\text{left}}_i, \mathbf{d}^{\text{left}}_i)\}$ containing the rest of samples not satisfying the test:

$$S^{\text{right}}_l = \bigcup \{(\mathbf{x}, \mathbf{d}) \in S_l | t_{f,\tau}(\mathbf{x}) = 1\}, \tag{4}$$

$$S^{\text{left}}_l = \bigcup \{(\mathbf{x}, \mathbf{d}) \in S_l | t_{f,\tau}(\mathbf{x}) = 0\}. \tag{5}$$

S^{right}_l and S^{left}_l are sequentially sent to the right child and the left child, respectively. The data samples arriving at the nodes are evaluated by all binary tests in the pool, and the test maximizing a predefined measure is selected and assigned to the node. In this work, the test is selected to minimize *displacement uncertainty*, which is defined as

$$U = \sum \|\mathbf{d}^{\text{left}}_i - \bar{\mathbf{d}}^{\text{left}}\|^2_2 + \sum \|\mathbf{d}^{\text{right}}_i - \bar{\mathbf{d}}^{\text{right}}\|^2_2, \tag{6}$$

where $\bar{\mathbf{d}}$ denotes the mean displacement vectors over all superframes in the set. This measure corresponds to the impurity of the displacement vectors. A leaf node is created when the maximum depth $D_{\max}$ or a minimum number of remaining superframes $N_{\min}$ is reached.

After training, each split node remains associated with the feature channel f and the threshold τ of the selected binary test. At each leaf node, we store the learned mean offset $\bar{\mathbf{d}}$ and covariance matrix Γ of the displacement vectors, i.e. the parameters of a multivariate Gaussian distribution $\mathcal{N}(\bar{\mathbf{d}}, \mathbf{\Gamma})$:

$$\bar{\mathbf{d}} = (\bar{d}_s, \bar{d}_e), \tag{7}$$

$$\mathbf{\Gamma} = \begin{pmatrix} \Gamma_s & 0 \\ 0 & \Gamma_e \end{pmatrix}. \tag{8}$$

However, as it can be seen from the matrix $\mathbf{\Gamma}$, we do not consider covariance between the onset and offset displacements. That is, $\mathcal{N}(\bar{\mathbf{d}}, \mathbf{\Gamma})$ is equivalent to two univariate Gaussian distributions $\mathcal{N}(\bar{d}_s, \Gamma_s)$ and $\mathcal{N}(\bar{d}_e, \Gamma_e)$. Fig. 2 demonstrates such a regression tree.

C. Testing

Via the trained regression forest, a test superframe at the time index t can provide the estimates for the event onset and offset positions. At each split node, the stored binary test is applied to the superframe, sending it either to the right or left child until ending up at a leaf node. At a leaf node l, the superframe gives estimates of the displacement vector $\hat{\mathbf{d}}$ to the onset and offset

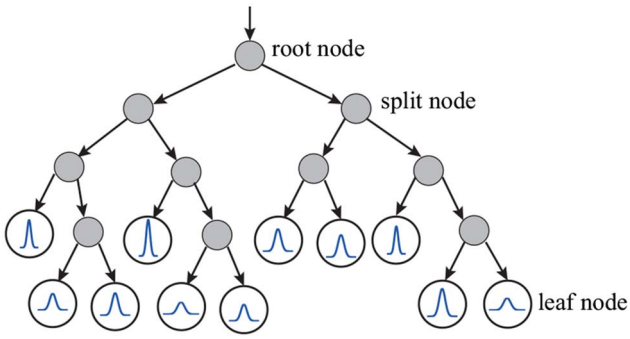

Fig. 2. Illustration of a random regression tree.

positions of the corresponding event in terms of the stored distribution $p(\hat{\mathbf{d}}|l) = \mathcal{N}(\hat{\mathbf{d}}; \bar{\mathbf{d}}, \mathbf{\Gamma})$. The posterior probabilities are summed up over all trees:

$$p(\hat{\mathbf{d}}) = \sum_{l \in \bar{\mathcal{L}}} p(\hat{\mathbf{d}}|l). \tag{9}$$

Here, $\bar{\mathcal{L}}$ is a subset of the corresponding leaf nodes. Owing to the fact that we do not consider covariance between the onset and offset displacements, $p(\hat{\mathbf{d}}|l)$ are explicitly equivalent to two separate distributions $p(\hat{d}_s|l) = \mathcal{N}(\hat{d}_s; \bar{d}_s, \Gamma_s)$ and $p(\hat{d}_e|l) = \mathcal{N}(\hat{d}_e; \bar{d}_e, \Gamma_e)$, respectively, leading to two separate posterior probabilities for onset and offset displacements:

$$p(\hat{d}_s) = \sum_{l \in \bar{\mathcal{L}}} p(\hat{d}_s|l), \tag{10}$$

$$p(\hat{d}_e) = \sum_{l \in \bar{\mathcal{L}}} p(\hat{d}_e|l). \tag{11}$$

Due to (1) and (2), the estimates of event onset and offset positions can be computed through the estimates of the displacements:

$$p(\hat{t}_s) = p(\hat{d}_s - t) = \sum_{l \in \bar{\mathcal{L}}} p((\hat{d}_s - t)|l), \tag{12}$$

$$p(\hat{t}_e) = p(\hat{d}_e + t) = \sum_{l \in \bar{\mathcal{L}}} p((\hat{d}_e + t)|l). \tag{13}$$

The expectations of $p(\hat{t}_s)$ and $p(\hat{t}_e)$ can indicate the onset and offset positions. That is, the location and duration of the corresponding AE in time are determined.

IV. Event Detection and Localization System

A. Acoustic Superframe and its Representation

In our system, it is essential that the AEs are decomposed into multiple parts, and each individual part is able to be recognized independently. Therefore, instead of using small frames, e.g. 30 ms long, we employ superframes as proposed in [22], [25]. A superframe is defined as a 100 ms long segment of the acoustic signal. It is a mid-level representation offering more discriminative power, hence being more reliable to be recognized independently. Furthermore, the detection error tolerance is usually set to 100 ms as in the most recent campaign [12], making its temporal resolution sufficient for AED in superframe fashion. The temporal resolution can be further improved by overlapping.

Superframes are divided into multiple interleaved small frames of 30 ms duration with Hamming window and 20 ms overlap. We utilize the set of 60 acoustic features suggested by Temko *et al.* in [3] to represent a small frame. These features have already been used in the CLEAR 2006/2007 challenges [10], [11], where they showed good discrimination power. Using the same feature set as the one used in the literature allows us to obtain a fair comparison between recognition engines. The feature set consists of: (1) 16 log-frequency filter bank parameters, along with the first and second time derivatives, and (2) the following set of features: zero-crossing rate, short time energy, four sub-band energies, spectral flux calculated for each sub-band, spectral centroid, and spectral bandwidth. Eventually, the empirical mean and the standard deviation of the frame feature vectors are calculated to form a 120-dimensional feature vector to represent the superframe.

B. System Description

Given training audio signals annotated with AEs of C categories of interest, we decompose each of them into interleaved superframes with an overlap of 90% of their duration to obtain the training set $\mathcal{S} = \{(\mathbf{x}_i, c, \mathbf{d}_i)\}$. The dense overlap is to ensure a high level of data correlation. Furthermore, the computational efficiency of decision trees allows us to do so. Each superframe, represented by a 120-dimensional feature vector, as described in Section IV-A, is annotated with the class label $c \in \{1, \ldots, C\}$ and the displacement vector $\mathbf{d} = (d_s, d_e)$. The background superframes are labelled with the class label 0, and no offset vectors are required.

The system consists of the following classification and regression models which are trained using the training data $\mathcal{S}$:

- M_{bg}: the classifier to distinguish foreground superframes from background ones. It outputs 0/1 if the input superframe is predicted as background/foreground.
- M_{ev}: the classifier to recognize superframes between different event categories. It outputs c if the predicted class label of the input superframe is c.
- R_c: the multivariate category-c regressor that estimates the temporal onsets and offsets of the events of category c given a test superframe. In total, C regressors are learned for C event categories.

The classifier M_{bg} to distinguish between possible events and background is applied first. Then the events are discriminated by the second classifier M_{ev}. By this scheme, we can avoid the problem of highly skewed training data. Both classifiers are based on random-forest classification [26] to take advantage of its computational efficiency. More importantly, random forest classification supports probability output which we will show to be very useful in our approach. For both classifiers, the number of random trees is conservatively set to 300. Due to dense overlapping of superframes, a large amount of data is generated. For the ITC-Irst database, the training and testing data contain 614,460 and 156,745 superframes, respectively. Those for

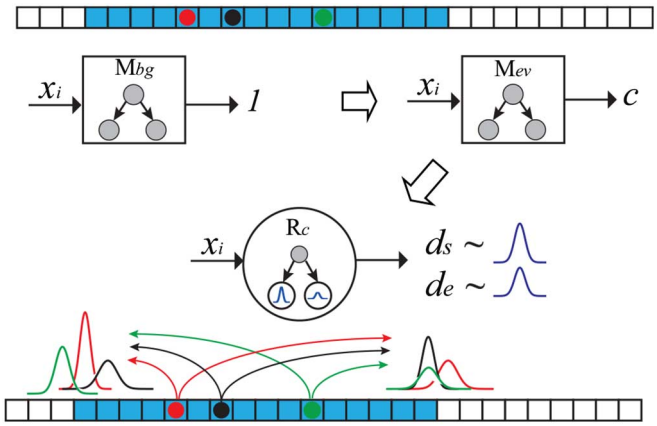

Fig. 3. Pipeline for event detection and localization with the learned models.

the UPC-TALP database are 397,914 and 196,554 superframes, respectively.

The regressors are trained with the random forest regression algorithm from Section III with ten random trees each. A randomly sampled subset containing 50% superframes of the category c training set $\mathcal{S}^c$ is used to train each random tree of R_c. In addition, we set the maximum depth $D_{\max} = 12$ and minimum number of superframes at leaf nodes $N_{\min} = 10$. This choice of parameters has been found experimentally. It yields a good compromise between under- and overfitting and computational cost. For example, for longer events, such as phone ringing or applause, a larger value for $D_{\max}$ should be used than for short events like chair moving. $D_{\max} = 12$ allows us to adequately model the longest-duration categories while not overfitting the short ones. The choice of $N_{\min} = 10$ is also large enough to avoid overfitting for short events and sufficient to approximate the mean and covariance of displacement vectors.

The pipeline of the AE detection and localization system is illustrated in Fig. 3. Given a test audio signal, we again divide it into multiple interleaved superframes as in the training phase. Afterwards, each superframe is fed into M_{bg} to test for background. If the superframe is recognized as foreground by M_{bg}, it is further fed into M_{ev} to predict the event class label. After the recognition phase, the superframes with predicted class label c are pushed through the regressor R_c to estimate the onset and offset positions of the AEs of category c in the audio signal.

C. Joint Event Detection and Localization

In order to detect and localize the AEs of category c, for each superframe at the time index t, we separately calculate the confidence of being event onset and offset by accumulating the posterior probabilities in (10) and (11) over the whole audio signal using the regressor R_c:

$$Z_s(t) = \sum_t (\mathbb{I}(\hat{c}_t = c) \cdot p(\hat{d}_s - t)), \tag{14}$$

$$Z_e(t) = \sum_t (\mathbb{I}(\hat{c}_t = c) \cdot p(\hat{d}_e + t)), \tag{15}$$

where $\hat{c}_t$ denotes the predicted class label of the superframe at the time index t and $\mathbb{I}$ is an indicator function given by

$$\mathbb{I}(z) = \begin{cases} 1 & \text{if} \quad z \text{ is true} \\ 0 & \text{if} \quad \text{otherwise.} \end{cases} \tag{16}$$

Moreover, we can further weight the scores with the confidence that a superframe's class label is predicted as c by the classifier M_{ev} as in (17) and (18), thanks to the probability output of random forest classification [26]:

$$Z_s(t) = \sum_t (w_t \cdot \mathbb{I}(\hat{c}_t = c) \cdot p(\hat{d}_s - t)), \tag{17}$$

$$Z_e(t) = \sum_t (w_t \cdot \mathbb{I}(\hat{c}_t = c) \cdot p(\hat{d}_e + t)). \tag{18}$$

Here, w_t is the probability that the predicted class label $\hat{c}_t$ equals c. By weighting the scores, the superframes recognized with higher confidence will contribute more into the scores.

In order to reduce the computation overhead during calculating the scores, we only evaluate the Gaussian distributions for the superframes in the displacement range of all superframes arriving at a leaf node during training. Moreover, we ignore the leaf nodes with the number of samples less than $N_{\min} = 10$. Eventually, the larger the scores of a superframe are, the higher confidence we have that the event onset and offset occur at it.

Typically, the audio signals should contain multiple AE occurrences, resulting in multiple peaks in both score spaces. Furthermore, since classifiers are generally not perfect, Z_s and Z_e are likely to be noisy, especially for AEs with low SNR. However, the peaks are expected to be dominant above the noise floor. In order to determine them, we normalize the scores Z_s and Z_e over all t into $[0; 1]$ by

$$\tilde{Z}_s(t) = Z_s(t) / \max(Z_s), \tag{19}$$

$$\tilde{Z}_e(t) = Z_e(t) / \max(Z_e), \tag{20}$$

and apply a cutoff threshold $\beta \in [0; 1]$ for both $\tilde{Z}_s$ and $\tilde{Z}_e$ to eliminate the noise below it:

$$\bar{Z}_s(t) = \tilde{Z}_s(t) \cdot \mathbb{I}(\tilde{Z}_s(t) \geq \beta), \tag{21}$$

$$\bar{Z}_e(t) = \tilde{Z}_e(t) \cdot \mathbb{I}(\tilde{Z}_e(t) \geq \beta). \tag{22}$$

Eventually, the peaks in $\bar{Z}_s$ and $\bar{Z}_e$ are determined as the maximum values in the connected positive regions. This idea is demonstrated in Fig. 4. for three different event categories in a test audio signal of the ITC-Irst database. The duration between a pair of peaks, a $\bar{Z}_s$ peak followed by a $\bar{Z}_e$ peak in temporal order, is considered as an event hypothesis. We impose a constraint that duration of the event hypotheses should not exceed twice the maximum duration of the AEs in the training audio signals.

V. Experiments

A. Evaluation Metrics

Following the CLEAR 2006 [10] and CLEAR 2007 [11] campaigns, we evaluate the proposed approach using three evaluation metrics: *Acoustic Event Error Rate (AEER)*, *AED-ACC*, and *AED-ER*.

AEER is computed as

$$\mathrm{AEER} = \frac{N_d + N_i + N_s}{N}, \tag{23}$$

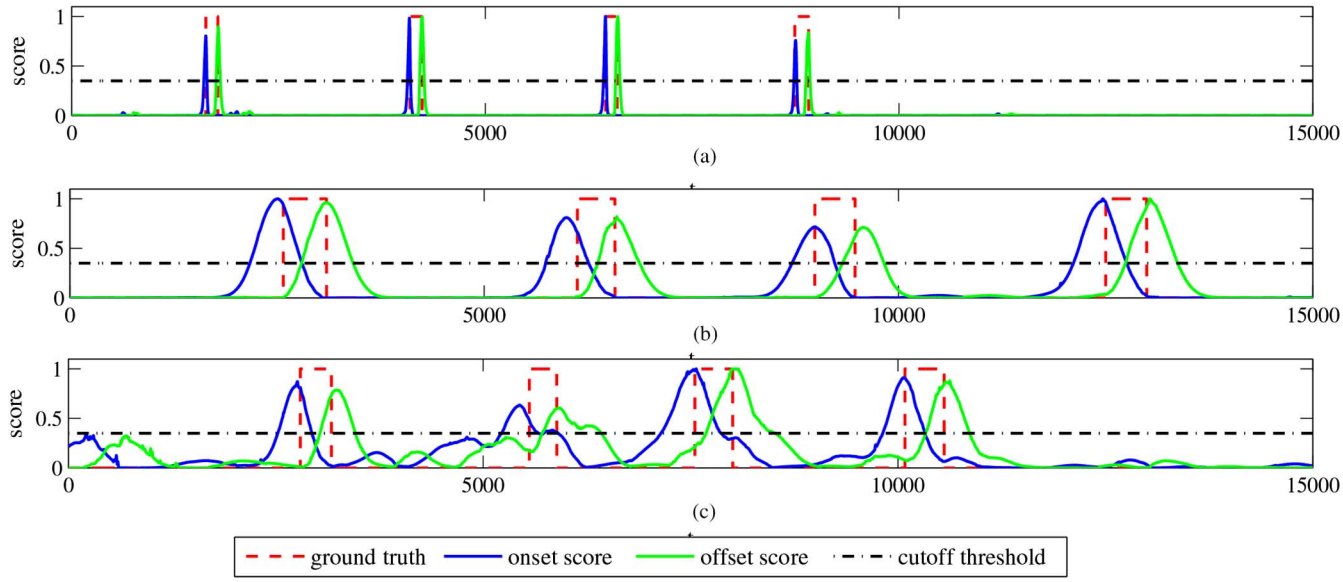

Fig. 4. Illustration of applying a common threshold to determine the score peaks on ITC-Irst database: (a) door slam AEs, (b) spoon cup jingle AEs, and (c) steps AEs.

where

- N = the number of ground-truth AEs to detect,
- N_d = the number of unmapped ground-truth AEs,
- N_i = the number of unmapped AE hypotheses,
- N_s = the number of mapped AE hypotheses with mismatched class labels.

A ground-truth AE is mapped as long as there exists at least one AE hypothesis whose center falls inside the interval of the ground-truth AE, and vice versa. A ground-truth AE is considered correctly detected if it is mapped by an AE hypothesis and their labels are matched.

AED-ACC is defined as the F-score measure:

$$\text{AED-ACC} \equiv \text{F-score} = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}, \qquad (24)$$

where

$$Precision = \frac{\text{the number of correct AE hypotheses}}{\text{the number of AE hypotheses}}, \qquad (25)$$

$$Recall = \frac{\text{the number of correctly detected ground-truth AEs}}{\text{the number of ground-truth AEs to detect}}. \qquad (26)$$

AED-ER is adapted from the NIST metric for speaker diarization [41] and is defined as

$$\text{AED-ER} = \frac{\sum_s \{L(s) \cdot (\max(N_\star(s), N_\triangleright(s)) - N_\diamond(s))\}}{\sum_s \{L(s) \cdot N_\star(s)\}}. \qquad (27)$$

This metric is evaluated on the audio segments that only contain event intervals, either hypothesized or ground-truth or both and is computed as the fraction of mismatching duration between AE hypotheses and ground-truth AEs. In (27), for a segment s:

- L = the duration of the segment,
- $N_\star$ = the number of ground-truth AEs,
- $N_\triangleright$ = the number of AE hypotheses,
- $N_\diamond$ = the number of ground-truth AEs matched by AE hypotheses.

The AEER and AED-ACC metrics focus on the detection of AE instances, and the temporal coincidence between the ground-truth and hypothesized AEs is not important. They are oriented for applications like real-time services for smart rooms, audio-based surveillance, etc. On the other hand, AED-ER focuses more on AE localization where a good temporal resolution of the detected AEs is important, making it suitable for applications like multimedia indexing/retrieval. AEER was used in the CLEAR 2006 evaluation whereas AED-ACC and AED-ER were used in CLEAR 2007. Note that AEER and AED-ER may exceed 100% because of the additional insertion errors.

B. Baseline Systems

In order to demonstrate the efficiency of the propose approach, we compare the performance of our systems, with both weighted and unweighted scores, to the performance of three baseline systems submitted to the CLEAR 2006 campaign [10]:

- **SVM**: this system pursues discriminative SVM classification for AED in detection-by-classification fashion with sliding window of 1 second and a 100 ms shift. The detection task is accomplished by two SVM classifiers: the first for event/background classification and the second for subsequent multi-class event classification. A median-filter of size 17 is applied on the binary sequences of decisions to eliminate too short silences or non-silences. Localization is carried out by considering the beginning and end of each detected event category. This system is the UPC-D submission in the campaign.
- $\mathbf{HMM}_1$: the detection strategy of this system is similar to the **SVM** system except that it uses HMMs as classification algorithms in lieu of discriminative SVMs. It is the CMU-D submission implemented by the CMU group.
- $\mathbf{HMM}_2$: different from the above two baseline systems, this system merges the event/background segmentation

TABLE I
ITC-IRST DATABASE OF NON-OVERLAPPED AES

Event category	#events		#superframes	
	Training	Testing	Training	Testing
door knock (kn)	35	12	5,977	1,983
door slam (ds)	39	12	6,263	2,076
steps (st)	38	12	17,866	4,810
chair moving (cm)	35	12	11,556	3,812
spoon cup jingle (cl)	36	12	21,989	7,065
paper wrapping (pw)	36	12	18,519	7,149
key jingle (kj)	36	12	23,655	8,421
keyboard typing (kt)	35	12	21,603	7,647
phone ring (pr)	66	23	38,824	12,316
applause (ap)	9	3	5,345	1,894
cough (co)	36	12	7,233	3,046
laugh (la)	36	12	7,003	2,459
door open	36	13	6,386	1,715
falling object	36	12	5,127	1,613
phone vibration	10	3	5,052	1,474
mimo pen buzz	36	12	24,144	9,528
unknown	17	9	2,647	2,033
Total	572	195	229,189	79,041

and event classification into a single step, as usually performed by the Viterbi search in common ASR framework. It is implemented by the ITC group and submitted as ITC-D system in the campaign.

All the baseline systems are single-channel. The comparison is only based on the AEER metric on both ITC-Irst and UPC-TALP databases as what have been done in CLEAR 2006. To our best knowledge, there have been no reports on the databases using the AED-ACC and AED-ER metrics. Nevertheless, we will present the results for further improvements and comparisons.

C. Experimental Results on ITC-Irst Database

The ITC-Irst database of non-overlapped AEs [42] was recorded with 32 microphones mounted in seven T-shaped arrays (with four microphones each) and four table microphones. It consists of twelve recording sessions with the AEs created by nine participants under the CHIL project [43]. There are totally 16 semantic event categories including door knock (kn), door slam (ds), steps (st), chair moving (cm), spoon cup jingle (cl), paper wrapping (pw), key jingle (kj), keyboard typing (kt), phone ring (pr), applause (ap), cough (co), laugh (la), mimo pen buzz, falling object, phone vibration, and unknown. Many of them are subtle (low SNR, e.g. steps, chair moving, and keyboard typing), making the task more challenging. Approximately 50 events were recorded for most of the event categories. The statistics for each event category are summarized in Table I. The database has been extensively examined in the CLEAR evaluations. Following the CLEAR 2006 setup, we only evaluate the first twelve classes. Nine recording sessions were employed as training files and three remaining sessions were employed as test files. Only one channel named *TABLE_I* was used.

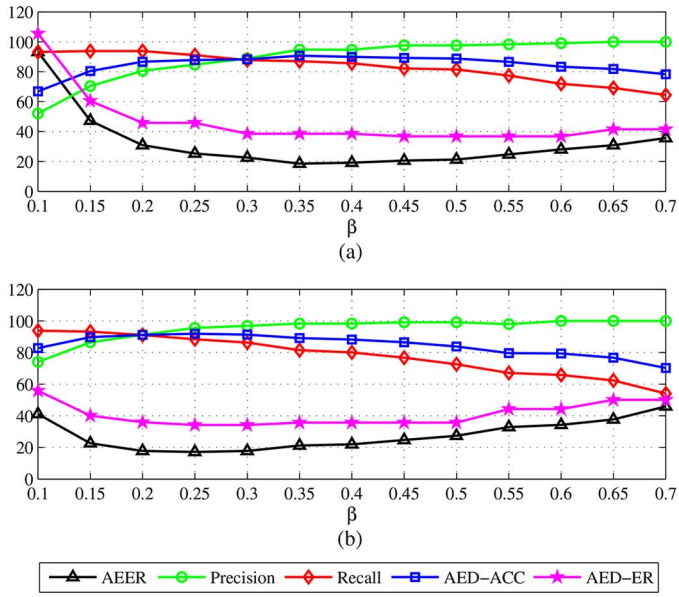

Fig. 5. Evaluation metrics over the parameter β for ITC-Irst database: (a) unweighted system, (b) weighted system.

TABLE II
AED PERFORMANCE COMPARISON WITH THE BASELINE SYSTEMS ON ITC-IRST DATABASE

	Our systems		SVM	HMM_1	HMM_2
	unweighted	weighted			
AEER (%)	**18.5**	**17.1**	64.6	45.2	23.6
AED-ACC (%)	90.7	91.8	N/A	N/A	N/A
AED-ER (%)	38.5	34.2			

First of all, the audio signals were downsampled to 16 kHz. Using training files, we trained the classifier M_{bg} to separate background superframes from event ones and M_{ev} to classify superframes among 16 semantic event categories. Twelve category-specific regressors were also trained for each of the twelve event categories of interest. The superframe-wise testing accuracies for M_{bg} and M_{ev} were 87.0% and 70.3%, respectively. The testing results of event detection and localization are shown in Fig. 5 with different values for cutoff threshold β from 0.1 to 0.7 with a step size of 0.05. For simplicity, we utilized the same cutoff threshold across all categories.

It can be seen from Fig. 5 that all the metrics show a similar behavior in both unweighted and weighted systems with increasing β. As expected, AED-ACC soars to the peak when β reaches the most appropriate value. After the peak, we saw a slow decline of AED-ACC. It is caused by fast decreasing of recall due to missed AEs although the quality of the AE hypotheses is improved. The AEER and AED-ER show the reversed patterns to AED-ACC because they are in the opposite sense of performance.

The highest performance in terms of overall detection error is obtained with $\mathrm{AEER} = 18.5\%$ at $\beta = 0.35$ and $\mathrm{AEER} = 17.1\%$ at $\beta = 0.25$ for unweighted and weighted systems, respectively. These results consistently outperform the baseline systems and some with a large margin, as is illustrated in

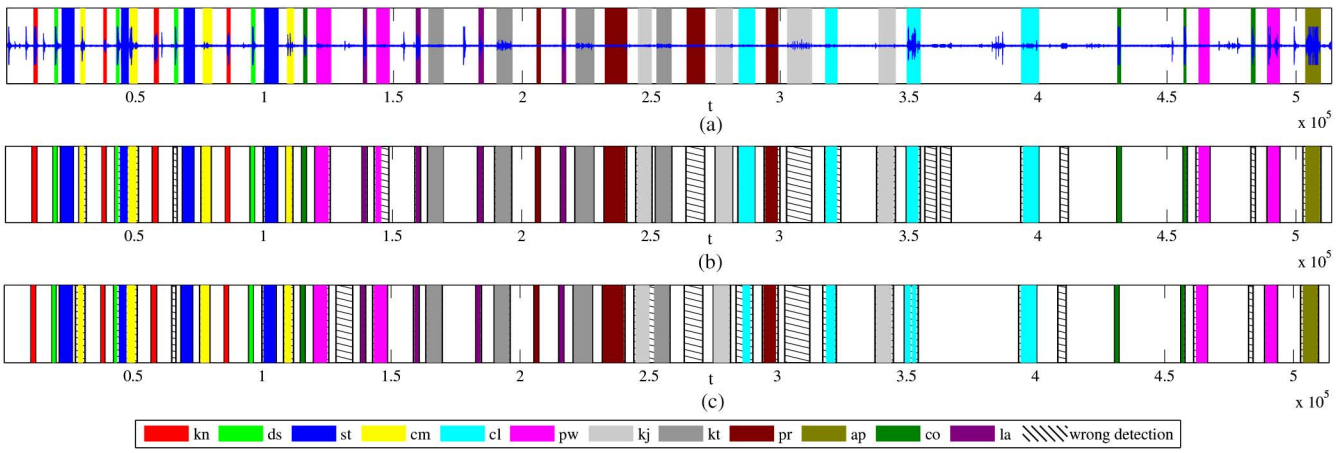

Fig. 6. Alignment of the AE localization results to the ground-truth AE durations on a test audio file of ITC-Irst database: (a) waveform and ground-truth AE durations, (b) localization results with unweighted system, and (c) localization results with weighted system.

TABLE III
AED PERFORMANCE FOR DIFFERENT CATEGORIES OF THE ITC-IRST DATABASE: THE UNWEIGHTED AND WEIGHTED SYSTEMS CORRESPOND TO $\beta = 0.35$ AND $\beta = 0.25$

		kn	ds	st	cm	cl	pw	kj	kt	pr	ap	co	la
AEER (%)	**unweighted**	0.0	16.7	16.7	58.3	0.0	0.0	8.3	16.7	39.1	0.0	25.0	16.7
	weighted	0.0	16.7	8.3	58.3	0.0	0.0	16.7	8.3	34.8	0.0	16.7	16.7
AED-ACC (%)	**unweighted**	100.0	90.9	91.7	81.5	100.0	100.0	95.7	91.7	75.7	100.0	87.0	90.9
	weighted	100.0	90.9	95.7	81.5	100.0	100.0	91.7	95.7	78.9	100.0	91.7	90.9
AED-ER (%)	**unweighted**	17.3	45.1	32.5	43.6	23.9	27.9	31.1	22.4	27.5	7.9	60.5	41.2
	weighted	17.6	44.6	26.6	40.7	27.8	29.3	30.7	24.4	28.7	9.8	59.5	41.1

TABLE IV
UPC-TALP DATABASE OF NON-OVERLAPPED AES

Event category	**#events**		**#superframes**	
	Training	**Testing**	**Training**	**Testing**
door knock (kn)	33	17	4,038	2,455
door slam (ds)	40	20	5,207	2,585
steps (st)	52	21	14,850	10,150
chair moving (cm)	51	25	14,590	7,054
spoon cup jingle (cl)	44	20	12,636	6,162
paper wrapping (pw)	60	24	19,432	10,617
key jingle (kj)	36	23	10,224	4,407
keyboard typing (kt)	46	20	13,190	6,255
phone ring (pr)	73	43	20,999	10,540
applause (ap)	40	20	13,834	7,459
cough (co)	44	21	5,445	3,123
laugh (la)	43	21	7,507	4,376
door open	40	20	4,552	2,142
unknown	83	42	5,284	3,571
Total	691	337	151,788	80,896

Table II. Noticeably, this is also the case with a wide range for β in Fig. 5. Compared to the best baseline system $\mathbf{HMM}_2$, the reductions of 5.1% and 6.5% were seen.

The detection and localization performances for different individual categories are reported in Table III. In Fig. 6, we also show the alignment of the localization results against the ground-truth duration on one of three test audio signals.

TABLE V
AED PERFORMANCE COMPARISON WITH THE BASELINE SYSTEMS ON UPC-TALP DATABASE

	Our systems		**SVM**	**$\mathbf{HMM}_1$**	**$\mathbf{HMM}_2$**
	unweighted	**weighted**			
AEER (%)	**24.7**	**22.9**	58.9	52.5	33.7
AED-ACC (%)	89.1	90.4	N/A	N/A	N/A
AED-ER (%)	38.14	39.79			

We also found that the main reason for wrong detection and localization is low SNR.

D. Experimental Results on UPC-TALP Database

The UPC-TALP database of non-overlapped AEs [44] was recorded in a meeting-room environment using 84 microphones: one array of 64 Mark III microphones, three T-shaped clusters (four microphones per cluster), four tabletop directional and four omni-directional microphones. It consists of three recording sessions performed by the same ten actors. The database includes 14 semantic classes: door knock (kn), door slam (ds), steps (st), chair moving (cm), spoon cup jingle (cl), paper wrapping (pw), key jingle (kj), keyboard typing (kt), phone ring (pr), applause (ap), cough (co), laugh (la), door open, and unknown. A summary of the dataset is shown in Table IV. About 60 sounds per class were recorded. Although this database is quite similar to the ITC-Irst database, it differs in the room arrangement, microphone setup, and acting positions. Therefore, it is useful to confirm the consistent efficiency

TABLE VI
AED PERFORMANCE FOR DIFFERENT CATEGORIES OF THE UPC-TALP DATABASE: THE UNWEIGHTED AND WEIGHTED SYSTEMS CORRESPOND TO $\beta = 0.5$ AND $\beta = 0.4$

		kn	ds	st	cm	cl	pw	kj	kt	pr	ap	co	la
AEER (%)	**unweighted**	17.7	5.0	33.3	36.0	0.0	20.8	30.4	10.0	46.5	0.0	19.0	33.3
	weighted	0.0	5.0	47.6	40.0	0.0	25.0	30.4	10.0	58.1	0.0	9.5	38.1
AED-ACC (%)	**unweighted**	97.0	97.4	88.9	91.4	100.0	90.9	87.8	97.4	73.3	100.0	89.5	85.0
	weighted	100.0	97.4	80.0	91.5	100.0	90.9	95.5	97.4	71.5	100.0	95.0	85.7
AED-ER (%)	**unweighted**	23.2	30.6	33.7	41.0	27.3	21.0	27.7	17.4	63.8	13.7	15.7	44.8
	weighted	21.0	30.5	60.3	29.8	22.0	31.1	26.5	12.2	51.9	15.2	8.1	59.19

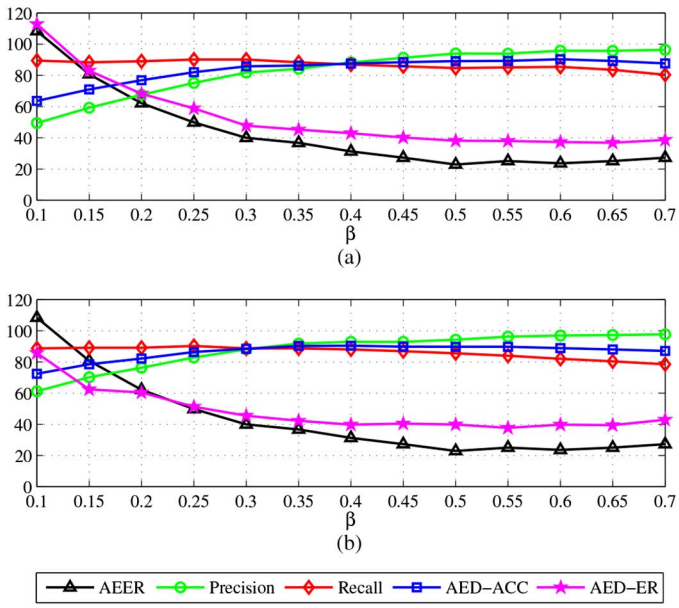

Fig. 7. Evaluation metrics over the parameter β for UPT-TALP database: (a) unweighted system, (b) weighted system.

of the proposed approach. Alike to the experiment on the ITC-Irst database, we evaluated the detection performance on the first twelve classes and considered the rest as background. The audio signals of the first two recording sessions were used for training, and testing was conducted on the remaining recording session. Only the third channel of the Mark III array was used for analysis.

The training procedure for the classifiers M_{bg}, M_{ev}, and twelve regressors R_c was repeated. M_{bg} was trained to recognize and reject background superframes from event ones and M_{ev} is to classify superframes among 14 semantic event categories. The superframe-wise testing accuracies for M_{bg} and M_{ev} are 91.7% and 74.1%, respectively. The overall detection and localization results are shown in Fig. 7 as functions of the common cutoff threshold β.

From Fig. 7, we can see a similar behavior of the AEER, AED-ACC, and AED-ER metrics as in the experiment with the ITC-Irst database. However, the optimal cutoff thresholds are noticeably different. With respect to the unweighted and weighted system, the optimal β is around 0.5 and 0.4 with $\mathrm{AEER} = 24.7\%$ and $\mathrm{AEER} = 22.9\%$. This dissimilarity will be discussed later in Section V-E. For the sake of comparison, AED results are given in Table V. As one can see, our systems enjoy the improvements of approximately 9% and 10.8% over the best baseline system $\mathbf{HMM}_2$.

The performance on individual categories with respect to the optimal cutoff-threshold values are further demonstrated in Table VI. Apart from the observation that the typical errors were caused by low-SNR events, in both experiments, the largest detection errors were seen with the 'phone ring' category, blaming to its high variance of sounds. For completeness, in Fig. 8, we also illustrate the alignment of the localization results against the ground-truth AEs on one of test audio signals.

E. Discussion

The rationale behind the state-of-the-art performance of the proposed approach can be explained by looking at some of its individual advantages over other approaches. First, while the common approaches, e.g. the HHM-based ASR framework and detection-by-classification approach, transfer the noisy segmentation/classification results into the final detection hypotheses, we can reject unreliable hypotheses by the cutoff threshold β. Second, longer frames can approximate nonstationary effects of audio events better than traditional short frames. One may argue that we then can use HMM models on sequences of superframes. However, on that viewpoint, our regression forests are even stronger. While HMMs can only capture dependencies between two consecutive frames, our approach can capture higher degrees of dependency (i.e. temporal structure) between superframes by maintaining displacements of a superframe to the event onset and offset. Last but not least, when the localization task is involved, unlike the detection-by-classification approach, the regression forests provide continuous estimates of event onset and offset positions, hence, implicitly capture event temporal-scale variations in the models.

Regression forests are different from other regression methods such as Support Vector Regression (SVR) [45]. While other methods model the mapping function as a whole, regression forests hierarchically split the regression problem into simpler smaller problems which are then modeled easily by simple models at the leaf nodes. With the tree construction algorithm proposed in the paper, we aim at clustering the training superframes into multiple clusters at the leaf nodes based on their features and their relative positions to event onsets and offsets. This means that we split the feature space into small regions whose relationships can be modelled easily. As already seen, we modeled the superframes in the same leaf node as Gaussian distributions. Another important aspect is that, unlike other regressors, which output point estimates, the

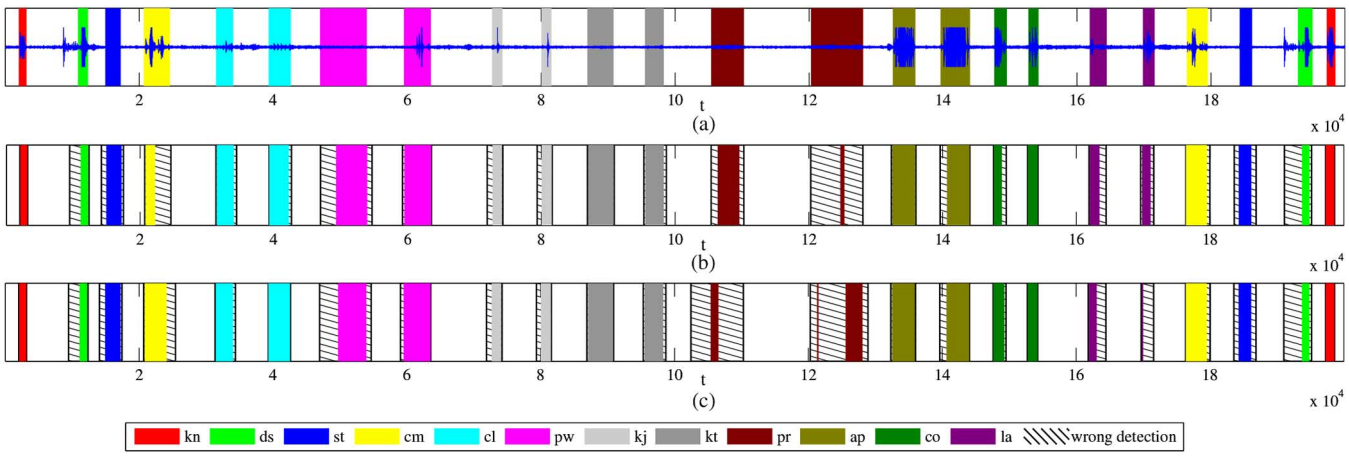

Fig. 8. Alignment of the AE localization results to the ground-truth AEs on one of the test audio files of UPC-TALP database: (a) waveform and ground-truth AE durations, (b) localization results with unweighted system, and (c) localization results with weighted system.

output of regression forests is a probability density function. It is much easier and more natural to sum up predicted probability densities obtained by all superframes to make predictions, while this cannot be done easily for point estimates.

Some observations about the importance of weighting scores can be inferred from the experimental results. First, in the experiments, the performance of the weighted systems are regularly better than those of the unweighted counterparts. Thus, favoring the superframes recognized with higher confidence can yield better results. This is a strong advantage of using the random forest classification [26] in our systems. Second, for the system with weighting, it is obvious that the optimal cutoff thresholds are significantly smaller than those for the unweighted systems. That means the performance converges faster to the optimum as the cutoff threshold increases. This observation suggests that the weighted systems produce a lower noise floor in the score spaces facilitating the peak determination.

It can be seen that the optimal cutoff thresholds were significantly different for the ITC-Irst and UPC-TALP databases. This is not about the approach itself but the data-dependency. In the ITC-Irst database, the audio files are much longer and contain more AE instances compared to the UPC-TALP database. This leads to the high variation of the maximum scores per file between two databases. As a result, this variation is transformed into dissimilarity in normalized score spaces via the normalization process. In practice, the optimal cutoff threshold value can be determined through cross-validation on the training data. Furthermore, in real-time AED scenarios where the score normalization becomes inappropriate, the cutoff threshold can be defined based on the absolute values of the scores which, again, can be found beforehand by cross-validation.

For the sake of simplicity, we utilized a common cutoff threshold for all event categories. However, it is more reasonable that different threshold values should be adapted for different event categories since their scoring spaces behave differently as illustrated in Fig. 4. Short events (like door slam) produce isolated peaks, periodic events (such as phone ring) lead to high-value plateaus, and low-SNR events (like steps) experience a significant noise floor. To be more specific, we show in Figs. 9 and 10, for ITC-Irst and UPC-TALP respectively,

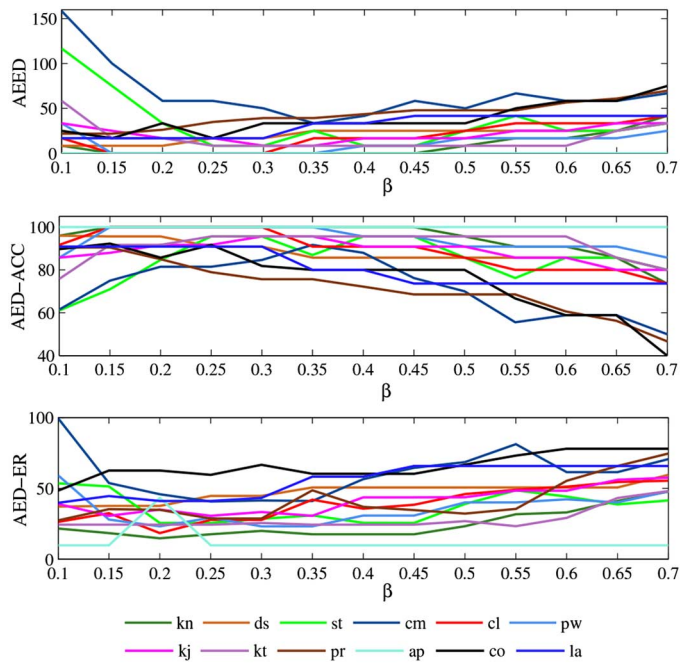

Fig. 9. Variation of category-specific results on the ITC-Irst database with respect to the parameter β in the weighted system.

the variations of twelve event categories on three evaluation metrics that correspond to different cutoff threshold values.

As the results indicate, our system is robust to short-term noise in form of wrongly recognized superframes. As reported, the recognition accuracies of the classifiers M_{bg} and M_{ev} are only at acceptable level and, in fact, they do not need to be perfect since we only need a portion of event superframes to be correctly recognized to estimate the onset and offset positions. In contrast, the performance of commonly adopted approaches strongly relies on the quality of the classifiers. In addition, this property also leads to the robustness to partial event overlapping and missing data, which are often the case in practice. Explicit background noise, such as the noise present in outdoor urban environments, may significantly degrade the performance of the algorithm. Thus, the proposed algorithm in its present form is mainly suitable for situations with reasonably low background

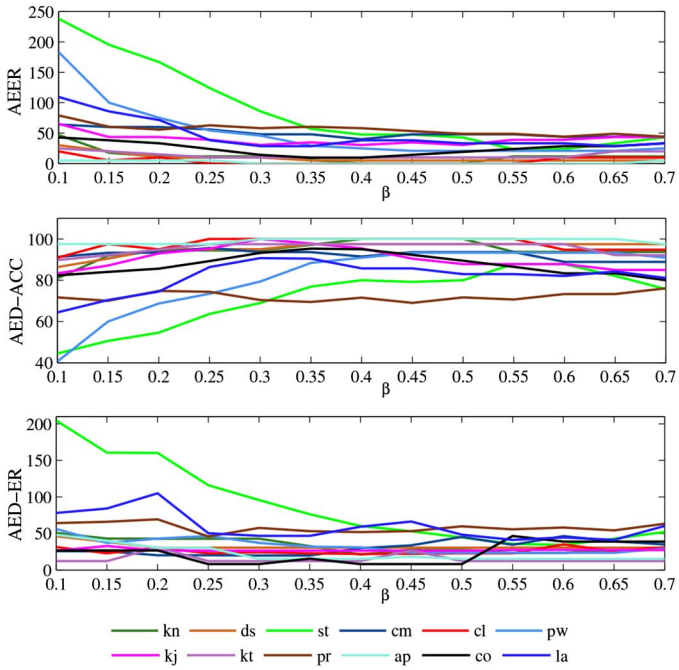

Fig. 10. Variation of category-specific results on the UPC-TALP database with respect to the parameter β in the weighted system.

noise, such as indoor recordings. To enhance the robustness, noise reduction techniques and source separation may be applied prior feature extraction, and more noise-robust features may be sought in future works.

It is also worth mentioning again the independence to event temporal scales of the proposed approach. Clearly, AEs in one category and across different categories can largely vary in their durations. Other approaches, like sliding windows [10] and event center detection [7], need to search on a huge temporal scale space to be able to localize the AEs. Our approach provides the continuous estimates for the onset and offset locations of the AEs. Therefore, we implicitly deal with this issue.

VI. Conclusions and Future Works

We proposed a novel approach for efficient automatic AE detection and localization based on regression forests. Using the concept of acoustic superframes, we trained two classifiers to recognize the superframes of background and different event categories of interest. Based on the random forest regression framework, we further learn category-specific regressors using the event superframes annotated with their displacements to the onsets and offsets of the events. On testing, after an event superframe is recognized, the corresponding regressor will provide the estimates of the onset and offset of the event hypothesis in time. The performance on the ITC-Irst and UPC-TALP databases exceeds those of three baseline systems by a large margin. This superior results demonstrate the efficiency and potential of the proposed approach.

The proposed method can be extended in different ways, offering room for further improvement. First, evaluation on databases with different degrees of event overlapping and speech-overlapping [3], [36] would be valuable for many applications. It is also useful for another evaluation for real-time AED scenarios. Second, this framework can be easily extended for multi-source fusion to account for low-SNR events. Third, the criteria used for selecting the binary tests at the split nodes of the decisive trees can be designed for the classification purpose. Consequently, both multi-class superframe classification and multi-class regression tasks can be done in the same decisive trees as in [23], [24], unifying all the tasks in the same forest model. This is especially meaningful when the number of event categories is significant large.

References

[1] J. Schröder, S. Wabnik, P. W. J. van Hengel, and S. Götze, "Detection and classification of acoustic events for in-home care," in *Ambient Assisted Living*. New York, NY, USA: Springer, 2011, pp. 181–195.

[2] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci, and A. Sarti, "Scream and gunshot detection and localization for audio-surveillance systems," in *Proc. IEEE Int. Conf. Adv. Video Signal Based Surveill.*, 2007, pp. 21–26.

[3] A. Temko and C. Nadeu, "Acoustic event detection in meeting-room environments," *Pattern Recogn. Lett.*, vol. 30, pp. 1281–1288, 2009.

[4] X. Zhuang, X. Zhou, M. A. Hasegawa-Johnson, and T. S. Huang, "Real-world acoustic event detection," *Pattern Recogn. Lett.*, vol. 31, no. 12, pp. 1543–1551, 2010.

[5] M. Janvier, X. Alameda-Pineda, L. Girin, and R. Horaud, "Sound-event recognition with a companion humanoid," in *Proc. 12th IEEE-RAS Int. Conf. Humanoid Robots (Humanoids)*, 2012, pp. 104–111.

[6] M. Janvier, X. Alameda-Pineda, L. Girin, and R. Horaud, "Sound representation and classification benchmark for domestic robots," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2014.

[7] J. A. Stork, L. Spinello, J. Silva, and K. O. Arras, "Audio-based human activity recognition using non-Markovian ensemble voting," in *Proc. IEEE Int. Symp. Robot Human Interactive Commun. (RO-MAN'12)*, 2012, pp. 509–514.

[8] M. Bugalho, J. Portelo, I. Trancoso, T. Pellegrini, and A. Abad, "Detecting audio events for semantic video search," in *Proc. Interspeech*, 2009.

[9] R. F. Lyon, "Machine hearing: An emerging field," *Signal Process. Mag.*, vol. 27, no. 5, pp. 131–139, Sep. 2010.

[10] A. Temko, R. Malkin, C. Zieger, D. Macho, C. Nadeu, and M. Omologo, "Clear evaluation of acoustic event detection and classification systems," *Lecture Notes in Comput. Sci.*, vol. 4122, pp. 311–322, 2007.

[11] A. Temko, C. Nadeu, D. Macho, R. Malkin, C. Zieger, and M. Omologo, "Acoustic event detection and classification," in *Computers in the Human Interaction Loop*. London, U.K.: Springer, 2009, pp. 61–73.

[12] D. Giannoulis, D. Stowell, E. Benetos, M. Rossignol, M. Lagrange, and M. D. Plumbley, "A database and challenge for acoustic scene classification and event detection," in *Proc. EUSIPCO*, 2013.

[13] J. Dennis, H. D. Tran, and E. S. Chng, "Image feature representation of the subband power distribution for robust sound event classification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 2, pp. 367–377, Feb. 2013.

[14] L. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.

[15] X. Zhou, X. Zhuang, M. Liu, H. Tang, M. Hasegawa-Johnson, and T. Huang, "HMM-based acoustic event detection with adaboost feature selection," *Lecture Notes Comput. Sci.*, vol. 4625, pp. 345–353, 2008.

[16] A. Mesaros, T. Heittola, and A. E. T. Virtanen, "Acoustic event detection in real life recordings," in *Proc. EUSIPCO*, 2010.

[17] A. Diment, T. Heittola, and T. Virtanen, "Sound event detection for office live and office synthetic AASP challenge," in *Proc. IEEE AASP Challenge on Detection Classif. Acoust. Scenes Events (WASPAA)*, 2013.

[18] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 1, pp. 72–83, Jan. 1995.

[19] A. Temko, C. Nadeu, and J.-I. Biel, "Acoustic event detection: SVM-based system and evaluation setup in CLEAR'07," *Lecture Notes Comput. Sci.*, vol. 4625, pp. 354–363, 2008.

[20] A. Plinge, R. Grzeszick, and G. Fink, "A bag-of-features approach to acoustic event detection," in *Proc. ICASSP*, 2014, pp. 3704–3708.

[21] G. Raboshchuk, C. Nadeu, O. Ghahabi, S. Solvez, B. M. noz Mahamud, A. R. de Veciana, and S. N. Hervas, "On the acoustic environment of a neonatal intensive care unit: Initial description, and detection of equipment alarms," in *Proc. Interspeech*, 2014.
[22] H. Phan and A. Mertins, "A voting-based technique for acoustic event-specific detection," in *Proc. 40th Annual German Congr. Acoust. (DAGA)*, 2014.
[23] A. Criminisi, J. Shotton, D. Robertson, and E. Konukoglu, "Regression forests for efficient anatomy detection and localization in computed tomography scans," *Med. Image Anal.*, vol. 17, no. 8, pp. 1293–1303, 2013.
[24] J. Gall, A. Yao, N. Razavi, L. V. G. Member, and V. Lempitsky, "Hough forests for object detection, tracking, and action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 11, pp. 2188–2202, Nov. 2011.
[25] B. Schuller, M. Wimmer, L. Mosenlechner, C. Kern, D. Arsic, and G. Rigoll, "Brute-forcing hierarchical functionals for paralinguistics: A waste of feature space?," in *Proc. ICASSP*, 2008, pp. 4501–4504.
[26] L. Breiman, "Random forest," *Mach. Learn.*, vol. 45, pp. 5–32, 2001.
[27] X. Zhuang, J. Huang, G. Potamianos, and M. Hasegawa-Johnson, "Acoustic fall detection using Gaussian mixture models and GMM supervectors," in *Proc. ICASSP*, 2009, pp. 69–72.
[28] Y. Lee, D. K. Han, and H. Ko, "Acoustic signal based abnormal event detection in indoor environment using multiclass adaboost," in *Proc. IEEE Int. Conf. Consumer Electron. (ICCE)*, Jan. 2013, pp. 322–323.
[29] A. Kumar, P. Dighe, R. Singh, S. Chaudhuri, and B. Raj, "Audio event detection from acoustic unit occurrence patterns," in *Proc. ICASSP*, 2012, pp. 489–492.
[30] Z. Huang, Y.-C. Cheng, K. Li, V. Hautamäki, and C.-H. Lee, "A blind segmentation approach to acoustic event detection based on I-vector," in *Proc. Interspeech*, 2013.
[31] C. V. Cotton and D. P. W. Ellis, "Spectral vs. spectro-temporal features for acoustic event detection," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA)*, 2011, pp. 69–72.
[32] J. Schröder, B. Cauchi, M. R. Schädler, N. Moritz, K. Adiloglu, J. Anemüller, S. Doclo, B. Kollmeier, and S. Goetze, "Acoustic event detection using signal enhancement and spectro-temporal feature extraction," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA)*, 2013.
[33] J. F. Gemmeke, L. Vuegen, B. Vanrumste, and H. Van hamme, "An exemplar-based NMF approach for audio event detection," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA)*, 2013, pp. 1–4.
[34] X. Zhuang, X. Zhou, T. S. Huang, and M. Hasegawa-johnson, "Feature analysis and selection for acoustic event detection," in *Proc. ICASSP*, 2008, pp. 17–20.
[35] J. Chen, J. Zhang, A. H. Kam, and L. Shue, "An automatic acoustic bathroom monitoring system," in *Proc. Int. Symp. Circuits Syst. (ISCAS)*, 2005, pp. 1750–1753.
[36] C. Müller, J.-I. Biel, E. Kim, and D. Rosario, "Speech-overlapped acoustic event detection for automotive applications," in *Proc. Interspeech*, 2008.
[37] T. Butko, "Feature selection for multimodal acoustic event detection," Ph.D. dissertation, Univ. Politecnica de Catalunya, Barcelona, Spain, 2011.
[38] T. Butko, C. Canton-Ferrer, C. Segura, X. Giró, C. Nadeu, J. Hernando, and J. R. Casas, "Acoustic event detection based on feature-level fusion of audio and video modalities," *EURASIP J. Adv. Signal Process.*, 2011, 485738.
[39] P.-S. Huang, X. Zhuang, and M. Hasegawa-Johnson, "Improving acoustic event detection using generalizable visual features and multi-modality modeling," in *Proc. ICASSP '11*, May 2011, pp. 349–352.
[40] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Mach. Learn.*, vol. 63, no. 1, pp. 3–42, 2006.
[41] NIST, "Spring 2007(RT-07) rich transcription meeting recognition evaluation plan," Tech. Rep., 2007.
[42] C. Zieger and M. Omologo, "Acoustic event detection - ITC-irst AED database," Internal ITC report, Tech. Rep., 2005.
[43] "CHIL. Computers in the human interaction loop," [Online]. Available: http://www.ipd.uka.de/CHIL/ Mar 2014
[44] A. Temko, D. Macho, C. Nadeu, and C. Segura, UPC-TALP database of isolated acoustic events Internal UPC Rep., Tech. Rep., 2005.
[45] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statist. Comput.*, vol. 14, no. 3, pp. 199–222, 2004.

Huy Phan (S'13) received the B.Sc. degree in computer science from the University of Science, Ho Chi Minh City, Vietnam, in 2007 and the M.Eng. in computer engineering from the Nanyang Technological University, Singapore, in 2012. From 2012 to 2013, he was with the University of Information Technology, Ho Chi Minh City, Vietnam, as a Lecturer. He is currently a Ph.D. student at the Graduate School for Computing in Medicine and Life Sciences, University of Lübeck, Lübeck, Germany, and a Research Associate at the Institute for Signal Processing, University of Lübeck. His research interests include audio/acoustic signal processing, pattern recognition, and machine learning, with a special focus on acoustic/audio event detection.

Marco Maaß (S'13) received the B.Sc. and M.Sc. degrees in computer science from the University of Lübeck, Lübeck, Germany, in 2010 and 2012, respectively. He is currently pursuing the Ph.D. degree at the Graduate School for Computing in Medicine and Life Sciences, University of Lübeck, and is a Research Associate at the Institute for Signal Processing, University of Lübeck. His research interests include machine learning, filter design, and image processing, with a special focus filter bank design, MRI reconstruction, and MPI reconstruction.

Radoslaw Mazur (S'09–M'11) was born in Wroclaw, Poland, in 1976. He received the Diplominformatiker degree from the University of Oldenburg, Oldenburg, Germany, in 2004 and the Dr.-Ing. degree in computer science from the University of Lübeck, Lübeck, Germany, in 2010. He was an Assistant Researcher in the Department of Physics, University of Oldenburg, from 2004 to 2006, and then joined the University of Lübeck. The current research interests are digital signal and audio processing, with a special focus on blind source separation.

Alfred Mertins (M'96–SM'08) received his Dipl.-Ing. degree from the University of Paderborn, Germany, in 1984 the Dr.-Ing. degree in Electrical Engineering and the Dr.-Ing. habil. degree in telecommunications from the Hamburg University of Technology, Germany, in 1991 and 1994, respectively. From 1986 to 1991, he was a Research Assistant at the Hamburg University of Technology, Germany, and from 1991 to 1995 he was a Senior Scientist at the Microelectronics Applications Center Hamburg, Germany. From 1996 to 1997, he was with the University of Kiel, Germany, and from 1997 to 1998 with the University of Western Australia. In 1998, he joined the University of Wollongong, where he was at last an Associate Professor of Electrical Engineering. From 2003 to 2006, he was a Professor in the Faculty of Mathematics and Science at the University of Oldenburg, Germany. In November 2006, he joined the University of Lübeck, Germany, where he is a Professor and Director of the Institute for Signal Processing. His research interests include speech, audio, and image processing, wavelets and filter banks, pattern recognition, and digital communications.