Random Forest Classification based Acoustic Event Detection Utilizing Contextual-Information and Bottleneck Features

Xianjun Xia, Roberto Togneri, Ferdous Sohel, David Huang

PII: S0031-3203(18)30115-8
DOI: 10.1016/j.patcog.2018.03.025
Reference: PR 6503

To appear in: *Pattern Recognition*

Received date: 18 March 2017
Revised date: 22 February 2018
Accepted date: 27 March 2018

Please cite this article as: Xianjun Xia, Roberto Togneri, Ferdous Sohel, David Huang, Random Forest Classification based Acoustic Event Detection Utilizing Contextual-Information and Bottleneck Features, *Pattern Recognition* (2018), doi: 10.1016/j.patcog.2018.03.025

Highlights

- A random forest classification based acoustic event detection system was constructed as the baseline system.
- Contextual information was employed to cope with the acoustic signals with long duration.
- Global bottleneck features were employed in the acoustic event detection system to utilize the prior knowledge of the event category information.
- Category-specific bottleneck features were employed in the acoustic event detection system to utilize the prior knowledge of the event boundary information.
- Evaluations on the UPC-TALP and ITC-IRST databases of highly

Random Forest Classification based Acoustic Event Detection Utilizing Contextual-Information and Bottleneck Features

Xianjun Xia[a,*], Roberto Togneri[a], Ferdous Sohel[b], David Huang[a]

[a] *School of Electrical, Electronic and Computer Engineering, The University of Western Australia, 35 Stirling Highway Perth WA 6009, AUSTRALIA*
[b] *School of Engineering and Information Technology, Murdoch University, 90 South St, Murdoch WA 6150, AUSTRALIA*

Abstract

The variety of event categories and event boundary information have resulted in limited success for acoustic event detection systems. To deal with this, we propose to utilize the long contextual information, low-dimensional discriminant global bottleneck features and category-specific bottleneck features. By concatenating several adjacent frames together, the use of contextual information makes it easier to cope with acoustic signals with long duration. Global and category-specific bottleneck features can extract the prior knowledge of the event category and boundary, which is ideally matched by the task of an event detection system. Evaluations on the UPC-TALP and ITC-IRST databases of highly variable acoustic events demonstrate the effectiveness of the proposed approaches by achieving a 5.30% and 4.44% absolute error rate improvement respectively compared to the state of art technique.

Keywords: acoustic event detection, contextual information, global bottleneck features, category-specific bottleneck features

*Corresponding author
Email address: Xianjun.Xia@research.uwa.edu.au (Xianjun Xia)

1. Introduction

Acoustic event detection (AED) deals with the event category and the localization of the acoustic events and is of great importance in many real-world applications such as security [1, 2, 3], life assistance [4, 5, 6] and human-computer interaction [7, 8]. Intra-class variations and spectral-temporal properties across classes pose challenges to acoustic event detection. Intra-class variations include different duration for the same acoustic sound type and non-stationary backgrounds. Spectral-temporal properties across classes include impulse-like sounds (e.g., door slam), tonal events (e.g., phone ring) and noise-like events (e.g., printer sound). Many works [9, 10, 11, 12, 13] have been carried out to address such challenges. The CLEAR [14] and DCASE [15, 16] challenge have attempted to capture the wide range of variations in the design of the AED corpora [17, 18].

The popular features used in AED systems are frame based features [8, 19], such as Mel-Frequency Cepstral Coefficients (MFCCs) and log frequency filter bank parameters, which have been demonstrated to represent the speech spectral structure well. However, the non-speech acoustic events contain a wide range of characteristic and non-stationary effects which may not be captured in such frame based features [20]. Frame based features do not represent the contextual information which has shown its effectiveness in acoustic signal processing systems [21]. The works in [22, 23, 24, 25, 26] used short frame based acoustic features. Moreover, these frame based acoustic features are extracted without any prior knowledge of the target events, which has been shown to be useful in [27, 28]. In [20], a 100ms long window was used while extracting the acoustic features and significant improvements have been achieved. Although the contextual information was used in [20, 21, 29], it was extracted without any prior knowledge of the event category or the event internal boundary information.

Inspired by the successful applications of the random forest technique [30, 31, 32, 33, 34] and the Deep Belief networks (DBNs) [35, 36] in the area

of pattern recognition and to utilize the contextual information as well as the prior knowledge of the acoustic events, this paper extends our previous and random forest based work of [28]. In [28], the importance of prior knowledge of the acoustic event category was verified on a random forest regression based AED system. This paper is different from the existing methods in the sense that contextual information is combined with DBN based global and category-specific bottleneck features derived from the prior knowledge of the event category and the event boundary.

For contextual information, it enables the feature space with a strong ability to describe the acoustic signals with an even longer duration. Real-world acoustic events, especially those presented periodically in time such as "phone vibration" and "alarm" cannot be represented effectively if the window length of acoustic features is shorter than the length of the basic unit within an acoustic event (each periodical vibration in "phone vibration" or each repeat sound in "alarm"). By adopting the contextual acoustic features, acoustic events which are highly variable can be represented more effectively in time. However, when more contextual acoustic features are adopted, the feature dimension increases. Our proposed global bottleneck features are bottleneck layer outputs of a deep belief network trained with the event category as the outputs and they can reduce the input dimension as well as utilize the prior knowledge of the event category information. We also propose category-specific bottleneck features which are the bottleneck layer outputs of category-specific deep belief networks trained with discretized event boundary information as the outputs. Category-specific bottleneck features reduce the input feature dimension and make full use of the prior knowledge of the event internal boundary information. Experimental results show the superior performance of AED system using contextual information, global and category-specific bottleneck features.

The rest of the paper is organized as follows. A brief review of related works is given in Section 2. In Section 3, the random forest classification based acoustic event detection system is introduced. Our proposed algorithms are described in Section 4. In Section 5 we provide the experimental results and analysis followed

by conclusion and future work in Section 6.

2. Related work

In [37], the local spectrogram features and the generalised Hough transform were utilized in the AED system. Authors in [38] investigated the use of biologically-inspired features, derived from a filter-bank of two-dimensional Gabor functions. In [39], spectral band selection based features are used. A novel approach for classifying acoustic events was proposed based on a bag of features (BOF) approach in [40]. In [41], different acoustic features, such as log-frequency filter bank coefficients, audio spectrum envelope (ASE), audio spectrum flatness (ASF), audio spectrum centroid (ASC), audio spectrum spread (ASS), spectral flux, spectral roll-off frequency and zero crossing rate were introduced and analyzed. Although various feature representations are explored, the acoustic features were extracted without contextual information, event type and event internal boundary information. To address this, different bottleneck features amalgamating contextual information are proposed in this paper to improve the detection performance.

The statistical machine learning algorithms such as Gaussian Mix Model (GMM) [42], Hidden Markov Model (HMM) [43, 44], Support Vector Machine (SVM) [45, 46, 47] and Non-negative Matrix Factorization (NMF) [22, 23] are routinely used to perform the classification task. The highly confusable non-speech sounds were detected using the fuzzy integral (FI) [10] which showed comparable results to the high performing SVM feature-level fusion in [10]. In [20], the authors proposed a technique for the joint detection and localization of non-overlapping acoustic events using random forest regressors. Multi-variable random forest regressors are learned for each event category to map each frame to continuously estimate the onset and offset time of the events. Cakir et al. [29] proposed to use multi label feed-forward deep neural networks for polyphonic sound event detection. They used Deep Neural Networks (DNNs) to learn a mapping between features and sound events. Heittola et al. [48] proposed two

iterative approaches based on the Expectation Maximum (EM) algorithm [49] to select the most likely stream to contain the target sound: one by always selecting the most likely stream and the other by gradually eliminating the most unlikely streams from the training data.

According to [20, 50], the random forest technique outperformed both HMM and SVM methods and has been shown to be the state of art approach on the non-overlapping acoustic events. In this paper, the random forest classification based AED system [50] is adopted as our benchmark system, upon which global and category-specific bottleneck features with contextual information are explored.

3. Random forest classification based AED system

Fig. 1 depicts the flowchart of the random forest classification based acoustic event detection system. As shown in Fig. 1, the system consists of three modules, namely feature representation, frame position discretization and event category detection and localization.

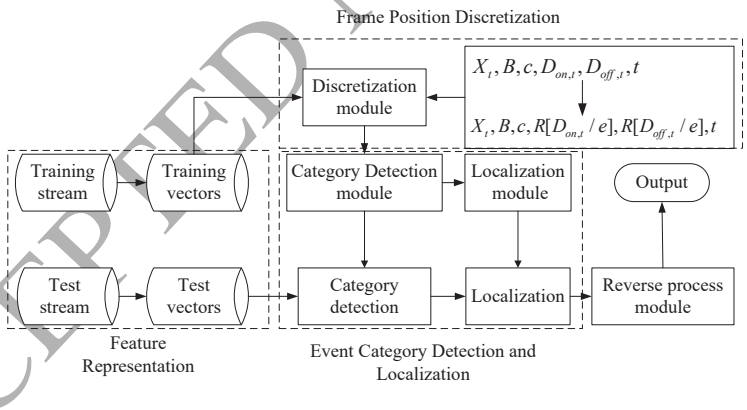

Figure 1: The flowchart of random forest classification based AED system.

3.1. Feature representation

For the training corpus, training vectors can initially be expressed as:

$$\boldsymbol{F}_t = \{\boldsymbol{X}_t, B, c, D_{on}, D_{off}, t\} \tag{1}$$

Here, $\boldsymbol{X}_t$ denotes the acoustic feature at frame t, B is a binary result representing whether the frame is silent or not, t is the time index and $c \in \{1, ..., C\}$ denotes the event category, where C is the number of event categories of interest. The event boundary information D_{on} and D_{off} denote the distances (number of frames) from the current frame to the start and end positions of the acoustic event that the current frame belongs to. The definition of D_{on} and D_{off} for the frame under consideration is shown in Fig. 2. This waveform is an event segment with start and end points where the selected part of the waveform is the current frame at time t. During testing, B, c, D_{on} and D_{off} need to be detected.

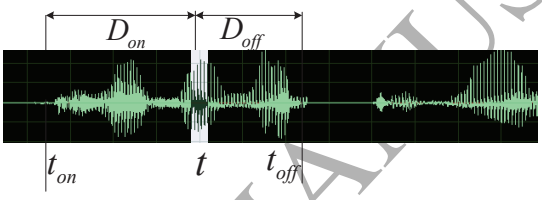

Figure 2: An illustration of D_{on} and D_{off}.

3.2. Frame position discretization

In the random forest classification based system, the frame positions are discretized with e as the discretization step. This is a reasonable approach because the final detected acoustic events are usually evaluated with an error tolerance and we choose e to match this. As displayed in Fig. 1, the event onset D_{on} and offset D_{off} are discretized with e as the interval first. After the discretization, the feature space for the training the category-specific localizer becomes:

$$\boldsymbol{F}_t^D = \{\boldsymbol{X}_t, B, c, R[\frac{D_{on}}{e}], R[\frac{D_{off}}{e}], t\} \tag{2}$$

Here, R is the rounding operation and e is the discretization step.

3.3. Event category detection and localization

The background random forest classifier (M_{bg}) and event category random forest classifier (M_{ev}) are trained first. The background and event category detection are treated as two-class and multi-class classification problems respectively using $\boldsymbol{F}_t$. At the stage of localization, $\boldsymbol{F}_t^D$ is adopted as the input feature space. With $\boldsymbol{F}_t^D$ as the input, the category-specific localization random forest classifiers $M_{c,on}$ and $M_{c,off}$ are trained. The output targets for $M_{c,on}$ and $M_{c,off}$ are $R[D_{on}/e]$ and $R[D_{off}/e]$ respectively. In the random forest classification based AED system, localization classifiers $M_{c,on}$ and $M_{c,off}$ are trained for each event category $c \in \{1, ..., C\}$. Training for random forest classifiers in this work is supervised and the Gini criterion [51] which focuses on minimizing the probability of misclassification is adopted as the splitting criterion while training.

Upon testing, with acoustic features $\boldsymbol{X}_t$ as the input, the background and event category are firstly detected. Assume that the detected event category for the test frame is $\hat{c}$, localization classifiers $M_{\hat{c},on}$ and $M_{\hat{c},off}$ are then used to output the boundary information o_{on} and o_{off}. The reverse process module ultimately converts the boundary information into the event onset $\hat{D}_{on}$ and offset $\hat{D}_{off}$. The detected event onset and offset at frame index m is expressed as:

$$\hat{D}_{on,m} = m - o_{on} \times e \tag{3}$$

$$\hat{D}_{off,m} = m + o_{off} \times e \tag{4}$$

When the event category, event onset and offset are detected at each frame index, the respective localization probability distribution across time are established as follows:

$$p_{on}(t) = \sum_{m=1}^{m=T} P_{on,m} f_{on,m}(t) \tag{5}$$

$$p_{off}(t) = \sum_{m=1}^{m=T} P_{off,m} f_{off,m}(t) \tag{6}$$

Here, T is the total number of test frames, and $P_{on,m}$ and $P_{off,m}$ are the output probability of the localization classifiers $M_{\hat{c},on}$ and $M_{\hat{c},off}$ respectively. The output probabilities are byproducts of the random forest classifiers. The $f_{on,m}(t)$ and $f_{off,m}(t)$ are defined as:

$$f_{on,m}(t) = \begin{cases} 1 & t = m - \hat{D}_{on,m} \\ 0 & \text{else} \end{cases} \tag{7}$$

$$f_{off,m}(t) = \begin{cases} 1 & t = m + \hat{D}_{off,m} \\ 0 & \text{else} \end{cases} \tag{8}$$

Fig. 3 shows the final detection process. As shown in Fig. 3, each test frame index m corresponds to an event category $\hat{c}$, and output probabilities $P_{on,m}$ and $P_{off,m}$ from $M_{\hat{c},on}$ and $M_{\hat{c},off}$ respectively. Each test frame index also corresponds to the event onset detection $\hat{D}_{on,m}$ and offset detection $\hat{D}_{off,m}$. The peaks of the localization distributions over the whole acoustic signal ($p_{on}(t_1)$ and $p_{off}(t_2)$) determine the ultimate acoustic event beginning time $\hat{t}_{on}$ and end time $\hat{t}_{off}$, which are expressed as:

$$\hat{t}_{on} = \underset{t}{\operatorname{argmax}}\, p_{on}(t) \tag{9}$$

$$\hat{t}_{off} = \underset{t}{\operatorname{argmax}}\, p_{off}(t) \tag{10}$$

4. Proposed AED algorithms

There are two types of bottleneck features in the proposed acoustic event detection system, the global bottleneck features (BN_G) and category-specific bottleneck features (BN_{CS}). The global and category-specific bottleneck

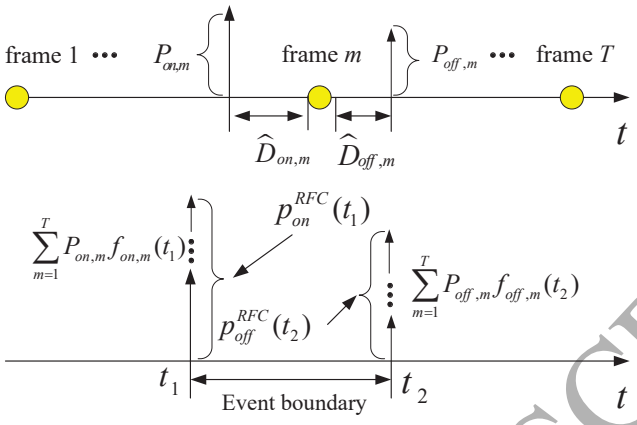

Figure 3: Determination of the acoustic event boundaries.

features are then combined with the acoustic features. Contextual information is defined as the sequence of multi-frame acoustic features:

$$\Delta_t = \{\boldsymbol{X}_{t-K}, ..., \boldsymbol{X}_t, ..., \boldsymbol{X}_{t+K}\} \tag{11}$$

where t is the frame index and K determines how many consecutive frames are utilized.

4.1. Global bottleneck features

Fig. 4 is the flowchart of the global bottleneck features fusion process in the proposed acoustic event detection system. As illustrated in Fig. 4, a global deep belief network Λ_G is trained with the contextual information Δ_t as the input and the event category as the output (the unit number for the output is the number of classes of interest). Several RBMs [52] acting as the building blocks for each layer are used to pre-train the initial weights of the network. The Λ_G models the joint distribution between Δ_t and the pth hidden layer h^p as follows:

$$p(\Delta_t, h^0, ..., h^\ell) = (\prod_{p=1}^{\ell} p(h^p|h^{p-1}))p(\Delta_t, h^0) \tag{12}$$

where $p(\Delta_t, h^0)$ means the visible-hidden joint distribution at the bottom-level RBM and $p(h^p|h^{p-1})$ denotes the conditional distribution for the hidden units

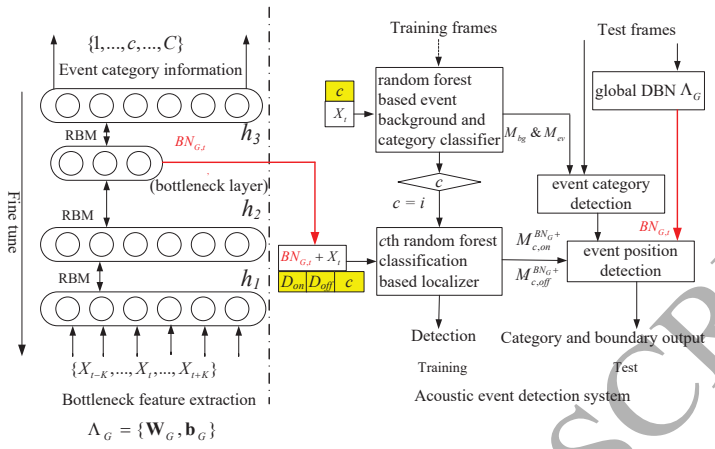

Figure 4: Flowchart of AED system utilizing global bottleneck features

conditioned on the visible units of RBM at level p. Training for the deep belief network proceeds as follows:

1) Train the first layer as an RBM with the acoustic feature Δ_t, as the visible input data Δ_t, to provide the estimate for $p(h^0|\Delta_t)$.

2) Utilize the first RBM output $p(h^0|\Delta_t)$ as the input data to the second layer, to train the second RBM to provide an estimate for $p(h^1|h^0)$.

3) Train the pth RBM layer from h^{p-1} to h^p, with the output of layer h^{p-1} as the input, to provide estimate $p(h^p|h^{p-1})$, then iterate for the desired number of layers, including the bottleneck layer.

A randomly initialized softmax layer is added to the top and backpropagation is adopted to optimize all the weights $\mathbf{W}_G$ and bias variable $\mathbf{b}_G$ by minimizing the following cross-entropy $J(\mathbf{W}, \mathbf{b})$:

$$\{\mathbf{W}_G, \mathbf{b}_G\} = \underset{\mathbf{W},\mathbf{b}}{\operatorname{argmin}} \{J(\mathbf{W}, \mathbf{b})\} \tag{13}$$

$$J(\mathbf{W}, \mathbf{b}) = -\frac{1}{M} \sum_{m=1}^{m=M} J(\mathbf{W}, \mathbf{b}; \Delta_m, c_m) \tag{14}$$

$$J(\mathbf{W}, \mathbf{b}; \Delta_m, c_m) = f(c_m = c) \log \frac{e^{z_m}}{\sum_{n=1}^{n=C} e^{z_n}} \tag{15}$$

Here, m denotes the training vector index and M is the total number of training vectors, c_m is the observed event category for the frame m and $f(v)$ is equal to 1 if v is true. z_i is the output of the final layer for the ith event category where x is the output of the second last layer, the w_i^T are the weights connecting the second last layer to the final layer and the b_i is the ith bias value for the final layer. The feature space of the proposed acoustic event detection system is defined as:

$$\boldsymbol{\Omega}_{G,t} = \{\boldsymbol{X}_t, \boldsymbol{BN}_{G,t}, B, c, R[\frac{D_{on}}{e}], R[\frac{D_{off}}{e}], t\} \tag{16}$$

where $\boldsymbol{BN}_{G,t}$ denotes the output of the bottleneck layer at frame t using the optimized weights and bias. The $\boldsymbol{BN}_{G,t}$ is expressed as:

$$\boldsymbol{BN}_{G,t} = \mathbf{W}_{l-1,l}^T \cdot \mathbf{O}_{l-1} + \mathbf{b}_l \tag{17}$$

where $\mathbf{W}_{l-1,l}^T$ is the optimized weights that collect the layer l-1 and the bottleneck layer l, the $\mathbf{O}_{l-1,t}$ is the output of the layer l-1 at frame t and $\mathbf{b}_{l,t}$ is the bias value for the bottleneck layer.

As shown in Fig. 4, bottleneck features $\boldsymbol{BN}_{G,t}$ and acoustic features $\boldsymbol{X}_t$ are concatenated to be the input feature space of the random forest classification based acoustic event detection system. With the newly constructed feature space $\boldsymbol{\Omega}_{G,t}$, new onset and offset random forest classifiers $M_{c,on}^{BN_G}$ and $M_{c,off}^{BN_G}$ are trained. During testing, the test vector and test vector context are fed into the pre-trained global neural network Λ_G to generate the global bottleneck features. The generated bottleneck features are then combined with the acoustic features as the input to the onset and offset random forest classifiers $M_{c,on}^{BN_G}$ and $M_{c,off}^{BN_G}$.

4.2. Category-specific bottleneck features

The mechanism for combining the category-specific bottleneck features with the acoustic features in the proposed system is shown in Fig. 5. Two components constitute the proposed AED system, namely the category-specific bottleneck feature extraction and acoustic event detection. For the category-specific

bottleneck feature extraction, a deep belief model Λ_c for each event category c is trained with contextual features as the input and discretized event boundary information $R[D_{on}/e]/R[D_{off}/e]$ as the outputs. A total number of C category-specific deep belief models are trained. The training for the category-specific deep belief network is identical to the training process of the global deep belief network Λ_G except for the cost function during the training. When the cth category-specific deep neural network Λ_c is trained, weights $\mathbf{W}_c$ and $\mathbf{b}_c$ are optimized by minimizing the following softmax cross-entropy $J(\mathbf{W}, \mathbf{b})$:

$$\{\mathbf{W}_C, \mathbf{b}_C\} = \underset{\mathbf{W},\mathbf{b}}{\operatorname{argmin}} \{J(\mathbf{W}, \mathbf{b})\} \tag{18}$$

$$\begin{aligned} J(\mathbf{W}, \mathbf{b}) = -\frac{1}{M}\{ &\sum_{m=1}^{m=M} J(\mathbf{W}, \mathbf{b}; \Delta_m, R^m[D_{on}/e]) \\ &+ \sum_{m=1}^{m=M} J(\mathbf{W}, \mathbf{b}; \Delta_m, R^m[D_{off}/e])\} \end{aligned} \tag{19}$$

For the acoustic event detection, acoustic features and category labels are used to train the background classifier M_{bg} and event category classifier M_{ev}. Then acoustic features and category-specific bottleneck features are concatenated to construct the input space of the onset and offset random forest classifiers $M_{c,on}^{BN_{CS}}$ and $M_{c,off}^{BN_{CS}}$. The feature space of the AED system with category-specific bottleneck features can be expressed as:

$$\boldsymbol{\Omega}_{CS,t} = \{\boldsymbol{X}_t, \boldsymbol{BN}_{CS,t}, B, c, R[\frac{D_{on}}{e}], R[\frac{D_{off}}{e}], t\} \tag{20}$$

where $\boldsymbol{BN}_{CS,t}$ is the output of the bottleneck layer of the category-specific neural network at frame t.

For testing, the acoustic features are extracted first, for which background and event category are detected. The category-specific bottleneck features are detected at the same time. Given the acoustic features, the detected event category and the obtained category-specific bottleneck features, the localization task is then performed.

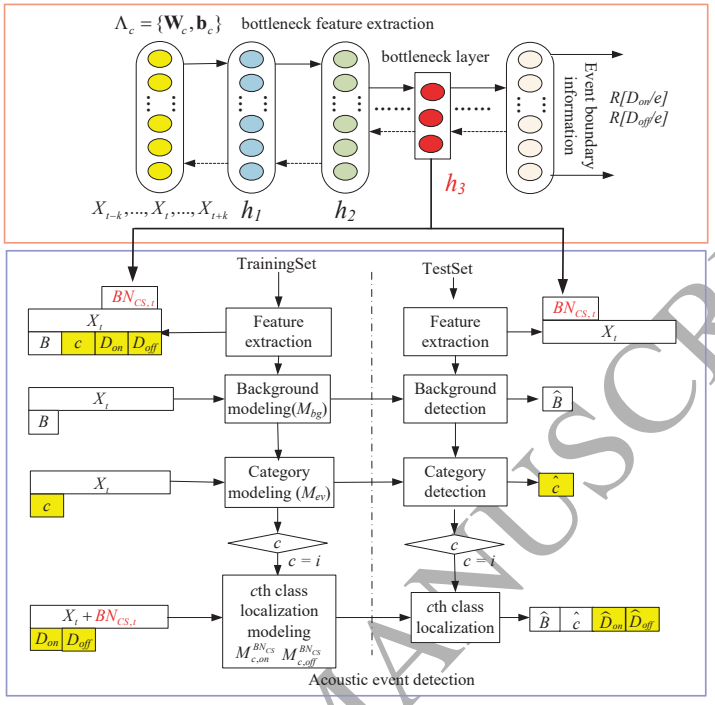

Figure 5: AED system utilizing the category-specific bottleneck features.

5. Experiments and analysis

5.1. System definition

This work implements and evaluates three types of acoustic event detection systems with different sets of input feature spaces.

5.1.1. Baseline system (BS)

The baseline system uses feature space $\boldsymbol{F}_t^D$ to train the random forest classifiers $M_{c,on}$ and $M_{c,off}$. The input acoustic feature to each of the random forest classifiers is $\boldsymbol{X}_t$. If the input feature space of BS is projected onto an m-dimensional feature space (e.g. using PCA), the system will be denoted BSm.

5.1.2. Extended system (ES)

In the extended system, the feature space is extended from $\boldsymbol{F}_t^D$ to $\boldsymbol{I}_t$ which is expressed as:

$$\boldsymbol{I}_t = \{\boldsymbol{X}_{t-K}, ..., \boldsymbol{X}_t, ..., \boldsymbol{X}_{t+K}, B, c, R[\frac{D_{on}}{e}], R[\frac{D_{off}}{e}\} \quad (21)$$

Here $\boldsymbol{X}_{t-K}$ and $\boldsymbol{X}_{t+K}$ denote the acoustic features of frame index $t-K$ and $t+K$ respectively. If the input feature space of ES is projected onto an m-dimensional feature space, the system will be denoted ESm.

5.1.3. Combined system (CS)

In the combined systems, the global and category-specific bottleneck features are combined with acoustic features $\boldsymbol{X}_t$ to train the random forest classifiers $M_{c,on}^{BN_G}/M_{c,off}^{BN_G}$ and $M_{c,on}^{BN_{CS}}/M_{c,off}^{BN_{CS}}$.

If an AED system adopts the b-dimensional global bottleneck features together with the acoustic feature $\boldsymbol{X}_t$ as its input feature space, it will be denoted CS_bG in this work. The system will be denoted CS_bE if the b-dimensional category-specific bottleneck features combined with the acoustic feature are used as the input feature space.

Table 1 provides the details of the feature space, dimension and side information being utilised for the three types of AED systems. Here, CI, EC and EB are abbreviations for contextual information, event category information and event boundary information respectively. Here, K denotes how many surrounding frames are utilized, m denotes the reduced feature dimension after using PCA, b is the bottleneck feature dimension and $Dim(\boldsymbol{X}_t)$ indicates the dimension of $\boldsymbol{X}_t$.

5.2. Database

Our proposed systems are tested on the two popular UPC-TALP [17] and ITC-IRST [18] databases.

The UPC-TALP database contains a set of isolated acoustic events that occur in a meeting room environment for the CHIL (Computers in the Human Interaction Loop) acoustic event detection task. Data was recorded at 44.1kHz, 24-bit precision. There are approximately 60 types of sounds for each of the sound classes. The acoustic events in the database are door

Table 1 Feature components for different systems.

System	Feature dimension	CI	EC	EB
BS	$Dim(\boldsymbol{X}_t)$	NO	NO	NO
ES	$Dim(\boldsymbol{X}_t).(2.K+1)$	YES	NO	NO
BSm	m	NO	NO	NO
ESm	m	YES	NO	NO
CS_bG	$Dim(\boldsymbol{X}_t)+b$	YES	YES	NO
CS_bE	$Dim(\boldsymbol{X}_t)+b$	YES	NO	YES

open(do), steps(st), door slam(ds), chair moving(cm), spoon-cup jingle(sc), paper wrapping(pw), key jingle(kj), keyboard clicking(kc), phone ringing(pr), applause(ap), cough(co), laugh(la), door knock(kn) and unknown(un).

The ITC-IRST database was also produced within the CHIL project. Different sets of isolated acoustic events in meeting room environments were recorded. These recordings are non-overlapping and 16 classes in total exist within this database: door knock(dk), door open(do), door slam(ds), steps(st), chair moving(cm), cough(co), paper wrapping(pw), falling object(fo), laugh(la), keyboard clicking(kc), phone ringing(pr), key jingle(kj), spoon-cup jingle(sc), phone vibration(pv), MIMIO pen buzz(mb) and applause(ap).

For the UPC-TALP and ITC-IRST databases, there are 1028 and 767 acoustic events, 5% of which are randomly chosen as the test sets and the total duration for the two databases are approximately 6.46 and 8.56 hours respectively. Details on the composition of the databases can be found in [20].

5.3. Acoustic signal representation

The acoustic representation of [7, 20] was adopted in our work and Table 2 provides the details of the feature components used. The log spectral parameters with their first and second derivatives, zero-crossing rate, spectral bandwidth, sub-band energies, spectral flux for each band and the short time energy are concatenated together to form the 60 dimensional feature for each frame (i.e.

$\mathrm{Dim}(\boldsymbol{X}_t) = 60$ in Table 1). The length is set to 100ms which was used in [20]. To utilize the correlation of acoustic features, each frame is divided into overlapped sub-frames of 30ms duration using a Hamming window with a 20ms overlap.

Table 2 Frame based acoustic representation.

Feature type	Dimension
16 log-spectral parameters, 1st, 2nd derivative	48
zero-crossing rate, spectral bandwidth	2
spectral centroid	1
4 sub-band energies	4
spectral flux for each sub-band	4
short time energy	1

5.4. Metrics

In our work, frame based correctness (AED-AC) is used to represent the correctness of the trained models and acoustic event detection error rate (AED-ER) is used to represent the detection performance.

The frame based correctness of the system is computed as:

$$AED\text{-}AC = \frac{N_d}{N} \tag{22}$$

where N is the total number of test frames and N_d is the number of frames correctly detected by the classifiers..

The acoustic event detection error rate is adopted from the NIST metric for speaker diarization and used as an evaluation metric in [20, 28, 50]. The AED-ER is defined as:

$$AED\text{-}ER = \frac{\sum_s l(s).[max(N_*(s), N_{\triangleright}(s) - N_{\diamond}(s))]}{\sum_s l(s).N_*(s)} \tag{23}$$

Here, for a segment s, $l(s)$ is the duration of the segment s, N_* is the number of manually labeled acoustic events, $N_{\diamond}$ is the number of correctly detected

acoustic events and $N_{\triangleright}$ is the number of total detections.

The AED-AC adopted in this paper is calculated frame wise and used to evaluate the performance of the trained acoustic models. The AED-ER includes the correctly and incorrectly detected events and is adopted as the final evaluation metric.

5.5. Configurations

During the process of the random forest training, the maximum tree depth and the minimum number of remaining frames within a leaf node are set to 12 and 10 respectively. According to the experimental results in [20], a tree depth of 12 allows for adequately modelling the long duration categories while not over-fitting the short ones. A minimum number of 10 frames within a leaf node is large enough to avoid over-fitting for short events. In the extended system ES and combined system CS, k is set to 3. The detection error tolerance e in this work is set to 100ms, which is a commonly used value as in [14, 20].

5.6. Results on the UPC-TALP database

In this section, the performance of different AED systems will compared and analyzed on the UPC-TALP database.

5.6.1. Performance of the BS system

To begin with, the baseline system BS using acoustic features of the current frame to train the localization random forest classifiers is constructed. The mean AED-AC and AED-ER across all the acoustic events for system BS are 61.96% and 29.46% respectively (see Fig. 6 for the individual event performance).

5.6.2. Performance of ES system

To demonstrate that long contextual information is good for an AED system, system ES with 420-dimensional features ($k = 3$ in $\boldsymbol{I}_t$) is compared with system BS in Fig. 6. The mean AED-AC and AED-ER across all the acoustic events for system ES is 70.37% and 25.79% respectively. The AED-AC and AED-ER for each event is shown in Fig. 6. As illustrated in Fig. 6, contextual

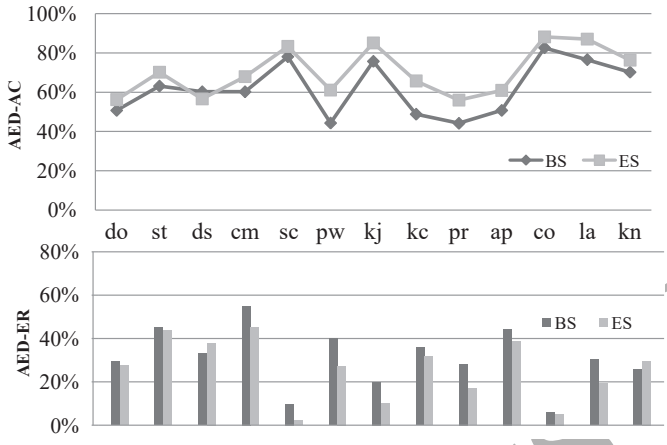

Figure 6: AED-AC and AED-ER for system BS and ES (UPC-TALP).

information contributes to a higher classification accuracy and lower acoustic event detection error.

To make a comparison between system BS and ES with an equal input feature dimension, the input feature space of BS and ES are both projected onto a 20-dimensional feature space. Principal component analysis (PCA) is adopted to reduce the input dimension. After reducing the dimension, system BS and ES become $BS20$ and $ES20$. The AED-AC and AED-ER results of $BS20$ and $ES20$ are displayed in Fig. 7. The higher detection accuracies and lower detection error rates for $ES20$ further verify that contextual information contributes to the performance of an acoustic event detection system.

5.6.3. Performance of system with global bottleneck features

When global bottleneck features $BN_{G,t}$ are combined with acoustic features, the dimension of the global bottleneck features (b in system CS_bG) is firstly optimized. We did this by varying the number of bottleneck layer units in the global deep belief network to maximise the category classification accuracy (percentage of frames with correctly detected event category). Table 3 shows the converged category classification accuracy using different number of bottleneck layer units. As shown in Table 3, the neural network with 20-dimensional

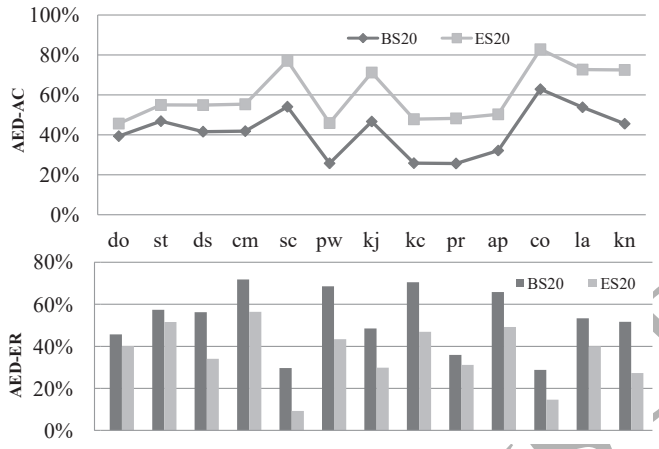

Figure 7: The AED-AC and AED-ER for system $BS20$ and $ES20$ (UPC-TALP).

global bottleneck features achieves the highest category classification accuracy.

The acoustic features and the 20-dimensional global bottleneck feature are

Table 3 The category classification accuracy using different number of bottleneck layer units.

	$b = 5$	$b = 10$	$b = 20$	$b = 50$	$b = 100$	$b = 200$
Accuracy	75.6%	78.3%	**79.2%**	78.4%	78.1%	76.7%

then concatenated to construct the system CS_20G. In the system CS_20G, the mean AED-AC and AED-ER across all the acoustic events are 69.12% and 27.19% respectively. Fig. 8 shows the results for each acoustic event compared to the BS system.

To further demonstrate the importance of global bottleneck features, the feature importance of acoustic features and global bottleneck features is shown in Fig. 9 for various acoustic events. Here, feature importance is the byproduct of the random forest classifier $M_{c,on}^{BN_G}$ and $M_{c,off}^{BN_G}$ during the process of splitting [53] at the stage of localization. In splitting, the decrease in the Gini node impurity [51] is recorded for each variable. The Gini node impurity was then used to guide the split. Averaging all decreases in the Gini impurity in the forest yields the variable importance. The importance of all variables is normalized

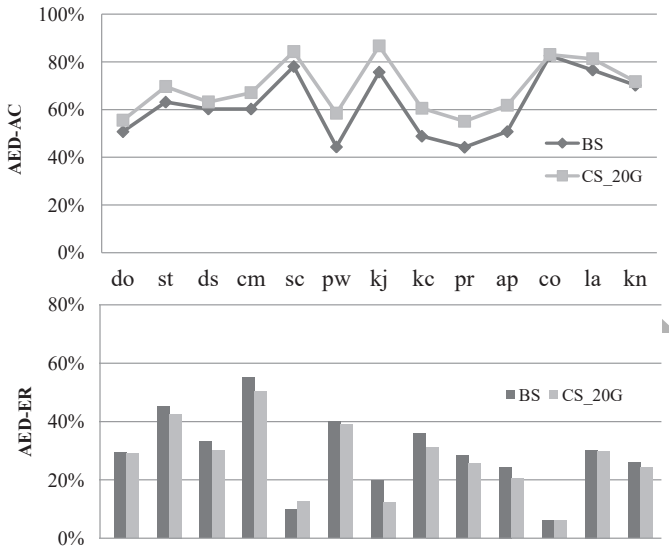

Figure 8: The AED-AC and AED-ER for system BS and CS_20G (UPC-TALP database).

from 0 to 1 by dividing by the largest importance score. A variable with a larger importance score indicates the variable is more important. The solid and dashed line segments denote the average mean importance of acoustic features and global bottleneck features respectively. As illustrated in Fig. 9, the global bottleneck features (dimension 61 to 80) achieved a higher mean importance than that of the acoustic features (dimension 1 to 60).

5.6.4. Performance of system with category-specific bottleneck features

When category-specific bottleneck features $BN_{CS,t}$ are combined with the acoustic features, the dimension of category-specific bottleneck features (b in the system CS_bE) also has to be optimized. We varied the number of units in the bottleneck layer and the minimum event detection error is adopted as the criteria in the selection of b. Fig. 10 shows that the converged AED-ER is achieved when b reaches 100. Thus the 100-dimensional category-specific bottleneck features are combined with the acoustic features $\boldsymbol{X}_t$. The mean AED-AC and AED-ER across all the acoustic events are 71.60% and 24.90%

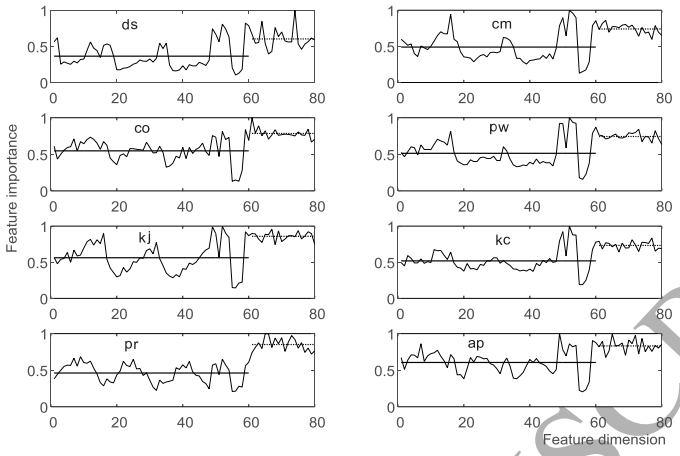

Figure 9: Importance of the global bottleneck features (UPC-TALP).

for system CS_100E. Detection results for each acoustic event are shown in Fig. 11 compared to the baseline AED system BS.

To show the importance of category-specific bottleneck features (dimension

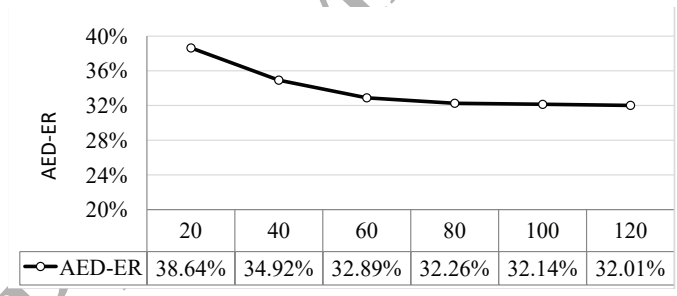

Figure 10: The AED-ER for systems with different sets of $BN_{CS,t}$ (UPC-TALP database).

61 to 160), the feature importance of various acoustic events are displayed in Fig. 12. Here, the feature importance is the byproduct of the random forest classifier $M_{c,on}^{BN_{CS}}$ and $M_{c,off}^{BN_{CS}}$. As shown in Fig. 12, in all cases, the average importance values of the category-specific bottleneck features consistently outperform that of acoustic features. Furthermore, as displayed in Fig. 12, the variation of category-specific bottleneck feature importance is smaller than the variation of acoustic feature importance, which indicates that category-specific bottleneck

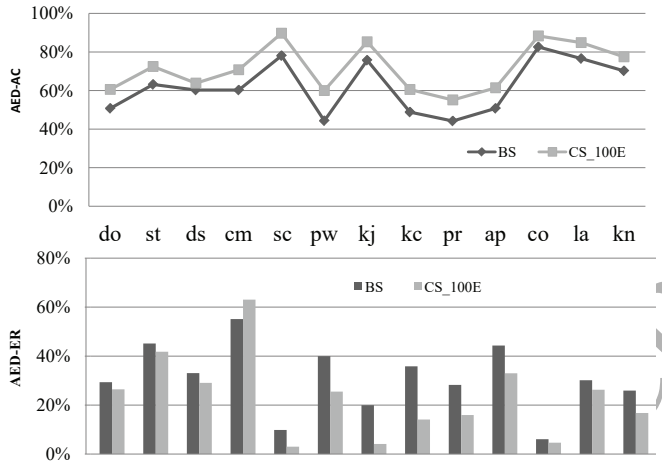

Figure 11: The AED-AC and AED-ER for system BS and CS_100E (UPC-TALP).

features provide consistent importance across all bottleneck units.

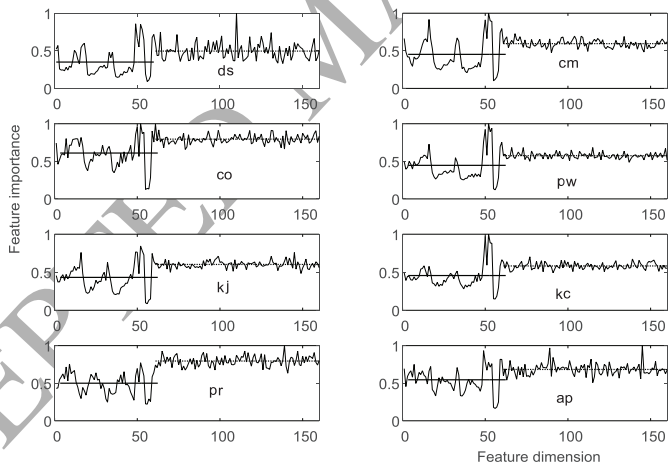

Figure 12: Importance of the category-specific bottleneck features (UPC-TALP).

5.7. Results on ITC-IRST database

In this section, performance of different AED systems will be compared and analyzed on the ITC-IRST database.

5.7.1. Performance of the BS system

For the baseline acoustic event detection system, the mean AED-AC and AED-ER are 59.51% and 30.59% respectively (for the individual event recognition refer to Fig. 13).

5.7.2. Performance of ES system

The mean AED-AC and AED-ER across all the acoustic events for *ES* are 67.69% and 27.16% respectively. The AED-AC and AED-ER for each acoustic event is shown in Fig. 13 for both *BS* and *ES* systems. Similar to the trend with the UPC-TALP database, higher classification accuracy and lower detection error are also achieved for the *ES* on the ITC-IRST database. These better detection results demonstrate the effectiveness of multi-frame contextual information.

To further verify the importance of contextual information on the AED

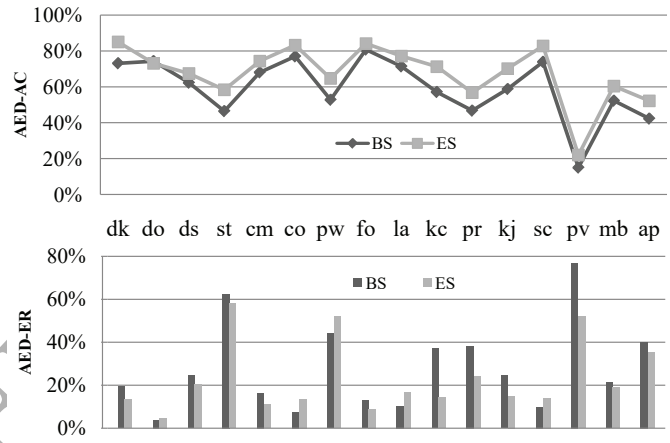

Figure 13: The AED-AC and AED-ER for system *BS* and *ES* (ITC-IRST).

system with equivalent input dimension, the input feature space of *BS* and *ES* are also projected onto a 20-dimensional feature space using PCA. The detection error rates for each acoustic event are displayed in Fig. 14 which demonstrates that contextual information assists the acoustic event detection by providing more discriminant features.

Table 4 The category classification accuracy using different number of bottleneck layer units on the ITC-IRST database

	$b = 5$	$b = 10$	$b = 20$	$b = 50$	$b = 100$	$b = 200$
Accuracy	66.2%	65.3%	**71.2%**	65.8%	64.6%	63.5%

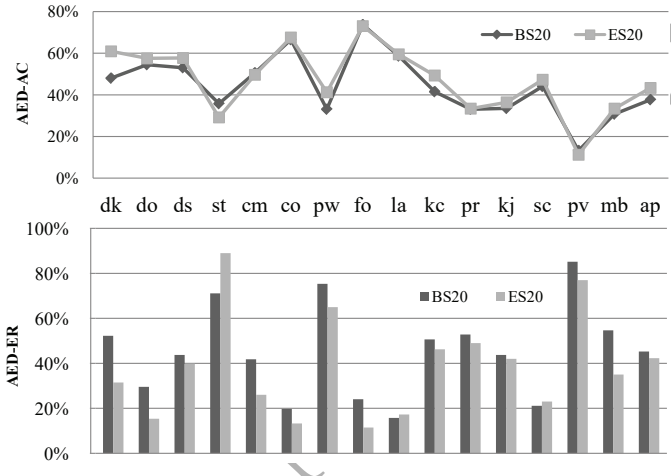

Figure 14: The AED-AC and AED-ER for system $BS20$ and $ES20$ (ITC-IRST).

5.7.3. Performance of system with global bottleneck features

When the global bottleneck features $BN_{G,t}$ are combined with the acoustic features, the dimension of global bottleneck features (b in the system CS_bG) is also optimized. From Table 4 where the number of bottleneck layer units is varied in the deep belief network, we can set b to 20 to maximise the category classification accuracy. The 20-dimensional bottleneck features are then combined with the acoustic features to construct the CS_20G system. The mean AED-AC and AED-ER are 65.50% and 28.17% for CS_20G. Fig. 15 shows the detection error rate of each acoustic event compared to the system BS.

The importance of global bottleneck features (dimension 61 to 80) of

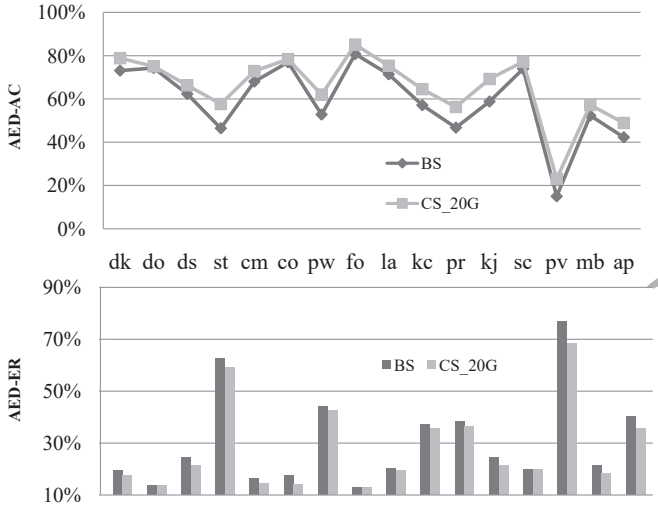

Figure 15: The AED-AC and AED-ER for system BS and CS_20G (ITC-IRST).

randomly chosen acoustic events are shown in Fig. 16. As displayed in Fig. 16, the global bottleneck features achieved higher average mean importance than the acoustic features.

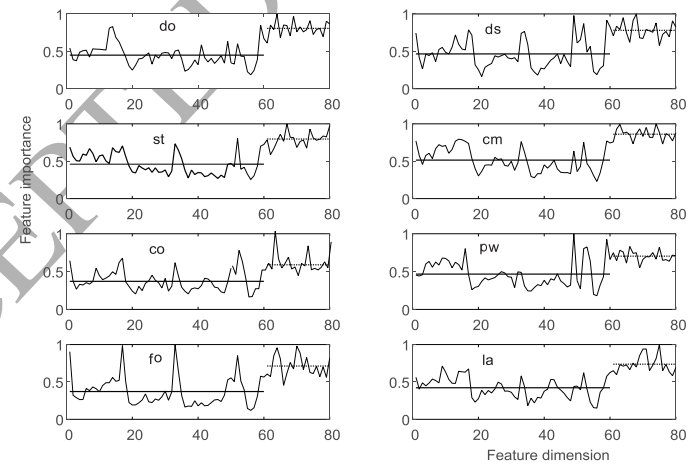

Figure 16: Importance of the global bottleneck features (ITC-IRST).

5.7.4. Performance of system with category-specific bottleneck features

When the category-specific bottleneck features $BN_{CS,t}$ are combined with acoustic features, the dimension of category-specific bottleneck features (b in the system CS_bE) needs to be optimized. We varied the number of units in the bottleneck layer and the minimum event detection error is adopted as the criteria in the selection of b. Fig. 17 shows the converged detection error when b reaches 60. Then the 60-dimensional category-specific bottleneck features are combined with the acoustic features $\boldsymbol{X}_t$. The mean AED-AC and AED-ER are 71.6% and 26.15% for the system CS_60E. The detection results for each acoustic event are displayed in Fig. 18 compared with the baseline system BS. The higher mean importance of the category-specific bottleneck features (dimension 61 to 120) are evident from Fig. 19.

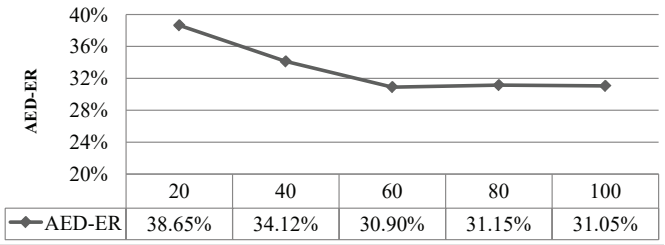

Figure 17: The AED-ER for systems with different sets of $BN_{CS,t}$ (ITC-IRST).

5.8. Discussion and analysis

Table 5 summarises the overall AED-ERs for systems using different approaches. Here, system RFR and SVM are systems which used random forest and SVM techniques. The CS_G and CS_E are our proposed random forest classification based systems using global bottleneck and category-specific bottleneck features respectively. From Table 5 system CS_E (CS_100E and CS_60E for the UPC-TALP and ITC-IRST database respectively) using category-specific bottleneck features together with acoustic features achieved the lowest detection error. To further demonstrate the efficiency of the proposed

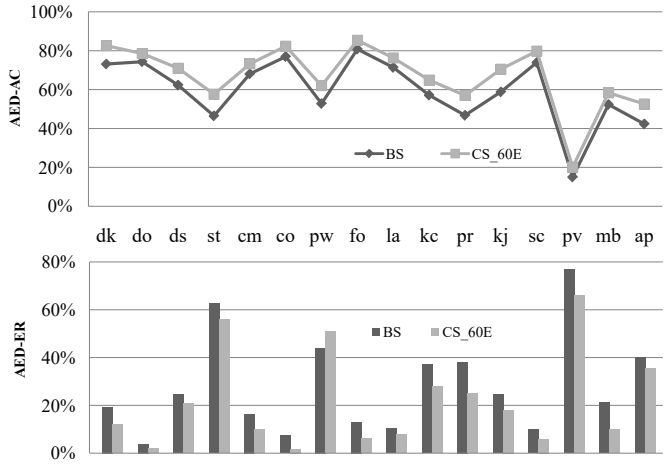

Figure 18: The AED-AC and AED-ER for system BS and CS_60E (ITC-IRST).

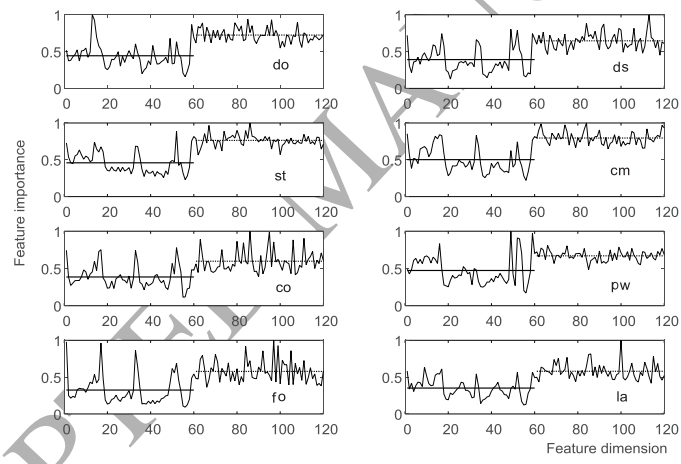

Figure 19: Importance of the category-specific bottleneck features (ITC-IRST).

features, support vector machine is combined with the best proposed bottleneck features (category specific bottleneck features) to construct the system $CS_{E\text{-}SVM}$, which outperformed the system SVM. Performance improvements can be attributed to the following factors:

Firstly, contextual information assists in the category classification and event localization. The contextual information helps to utilize the

Table 5 The AED-ER for systems using different approaches

RF based	RFR [20]	BS [50]	ES	CS_G	CS_E
UPC-TALP	38.14%	30.20%	25.79%	27.19%	**24.90%**
ITC-IRST	34.20%	30.59%	27.16%	28.17%	**26.15%**
SVM based	SVM	$CS_{E\text{-}SVM}$			
UPC-TALP	44.12%	37.10%			
ITC-IRST	38.70%	32.18%			

acoustic information from a longer duration rather than the frame based acoustic features. For some periodically acoustic events, such as the "phone vibration" and "applause", the contextual information can effectively capture the periodical information from the acoustic signals.

However, as more contextual information is used, the higher the input dimension will be. The use of the global bottleneck features reduces the input dimension but captures the important contextual information. Moreover, the global bottleneck features are trained with the acoustic event type as the output of the neural network, which makes the extracted bottleneck features more acoustic event discriminant. The resultant features are better able to compactly represent the complex unstructured acoustic events.

The category-specific bottleneck features are derived with the contextual information as the neural network input and the discretized acoustic event positions as the neural network output. This makes the category-specific bottleneck features amalgamate the contextual information and the acoustic event localization information. Category-specific bottleneck features embed rich information of the acoustic event boundaries and provide much more class-specific discriminant features for the final event localization.

6. Conclusion and future work

This paper proposes to utilize the prior knowledge of the acoustic event category and boundary information along with contextual information. Global

and category-specific bottleneck features are employed to construct a more discriminative feature space. We show that our proposed system achieves state of the art performance. However, only the available prior knowledge of the event category and boundary information are utilized in this paper. Additional prior knowledge, such as prior acoustic event duration, prior distribution of acoustic event signals and discriminative differences between specific acoustic events are areas for consideration in future work.

7. ACKNOWLEDGMENT

This work was supported by the International Postgraduate Research Scholarship (IPRS) from the University of Western Australia. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Tesla K40 GPU used for this research.

8. References

References

[1] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci, A. Sarti, Scream and gunshot detection and localization for audio-surveillance systems, in: International Conference on Advanced Video and Signal Based Surveillance (AVSS), IEEE, 2007, pp. 21–26.

[2] C. Clavel, T. Ehrette, G. Richard, Events detection for an audio-based surveillance system, in: International Conference on Multimedia and Expo (ICME), IEEE, 2005, pp. 1306–1309.

[3] J. Schröder, S. Goetze, V. Grutzmacher, J. Anemüller, Automatic acoustic siren detection in traffic noise by part-based models., in: International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2013, pp. 493–497.

[4] F. Jin, F. Sattar, S. Krishnan, Log-frequency spectrogram for respiratory sound monitoring, in: International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2012, pp. 597–600.

[5] J. Schroeder, S. Wabnik, H. Van, W. J. Peter, S. Goetze, Detection and classification of acoustic events for in-home care, in: Ambient Assisted Living, Springer, 2011, pp. 181–195.

[6] S. Päßler, W. J. Fischer, Food intake monitoring: automated chew event detection in chewing sounds, IEEE journal of biomedical and health informatics 18 (1) (2014) 278–289.

[7] A. Temko, C. Nadeu, Acoustic event detection in meeting-room environments, Pattern Recognition Letters 30 (14) (2009) 1281–1288.

[8] X. D. Zhuang, Z. Xi, A. H. J. Mark, S. H. Thomas, Real-world acoustic event detection, Pattern Recognition Letters 31 (12) (2010) 1543–1551.

[9] P. Foggia, N. Petkov, A. Saggese, N. Strisciuglio, M. Vento, Reliable detection of audio events in highly noisy environments, Pattern Recognition Letters 65 (2015) 22–28.

[10] A. Temko, D. Macho, C. Nadeu, Fuzzy integral based information fusion for classification of highly confusable non-speech sounds, Pattern Recognition 41 (5) (2008) 1814–1823.

[11] X. J. Xia, R. Togneri, F. Sohel, D. Huang, Confidence based acoustic event detection, in: International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2018, p. accepted on 29 Jan.

[12] X. Xia, R. Togneri, F. Sohel, D. Huang, Frame wise dynamic threshold based polyphonic acoustic event detection, in: Proc. Interspeech, ISCA, 2017, pp. 474–478.

[13] J. D. Krijnders, M. E. Niessen, T. C. Andringa, Sound event recognition through expectancy-based evaluation of signal-driven hypotheses, Pattern Recognition Letters 31 (12) (2010) 1552–1559.

[14] A. Temko, R. Malkin, C. Zieger, D. Macho, C. Nadeu, M. Omologo, CLEAR evaluation of acoustic event detection and classification systems, in: International Evaluation Workshop on Classification of Events, Activities and Relationships, Springer, 2006, pp. 311–322.

[15] D. Giannoulis, S. Dan, B. Emmanouil, R. Mathias, L. Mathieu, D. P. Mark, A database and challenge for acoustic scene classification and event detection, in: European Signal Processing Conference (EUSIPCO), IEEE, 2013, pp. 1–5.

[16] S. Adavanne, T. Virtanen, A report on sound event detection with different binaural features, Tech. rep., DCASE Challenge (September 2017).

[17] A. Temko, D. Macho, C. Nadeu, C. Segura, UPC-TALP database of isolated acoustic events, Tech. rep., Internal UPC report (2005).

[18] Z. Christian, O. Maurizio, Acoustic event detection ITC-IRST AED database, Tech. rep., Internal ITC report (2005).

[19] M. E. Niessen, T. L. M. V. Kasteren, A. Merentitis, Hierarchical modeling using automated sub-clustering for sound event recognition, in: Workshop on Applications of Signal Processing to Audio and Acoustics, IEEE, 2013, pp. 1–4.

[20] H. Phan, M. Maaß, R. Mazur, A. Mertins, Random regression forests for acoustic event detection and classification, IEEE/ACM Transactions on Audio, Speech, and Language Processing 23 (1) (2015) 20–31.

[21] O. Gencoglu, T. Virtanen, H. Huttunen, Recognition of acoustic events using deep neural networks, in: European Signal Processing Conference (EUSIPCO), IEEE, 2014, pp. 506–510.

[22] D. Arnaud, C. Arshia, L. Guillaume, Real-time detection of overlapping sound events with non-negative matrix factorization, in: Matrix Information Geometry, Springer, 2013, pp. 341–371.

[23] T. Komatsu, Y. Senda, R. Kondo, Acoustic event detection based on non-negative matrix factorization with mixtures of local dictionaries and activation aggregation, in: International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2016, pp. 2259–2263.

[24] I. Choi, K. Kwon, S. H. Bae, N. S. Kim, DNN-based sound event detection with exemplar-based approach for noise reduction, Tech. rep., DCASE Challenge (September 2016).

[25] T. Hayashi, S. J. Watanabe, T. Toda, T. Hori, J. L. Roux, K. Takeda, Bidirectional LSTM-HMM hybrid system for polyphonic sound event detection, Tech. rep., DCASE Challenge (September 2016).

[26] T. H. Vu, J. C. Wang, Acoustic scene and event recognition using recurrent neural networks, Tech. rep., DCASE Challenge (September 2016).

[27] Z. R. Feng, Q. Zhou, J. Zhang, P. Jiang, X. W. Yang, A target guided subband filter for acoustic event detection in noisy environments using wavelet packets, IEEE/ACM Transactions on Audio, Speech, and Language Processing 23 (2) (2015) 361–372.

[28] X. J. Xia, R. Togneri, F. Sohel, D. Huang, Random forest regression based acoustic event detection with bottleneck features, in: IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2017, pp. 157–162.

[29] E. Cakir, T. Heittola, H. Huttunen, T. Virtanen, Polyphonic sound event detection using multi label deep neural networks, in: International Joint Conference on Neural Networks (IJCNN), IEEE, 2015, pp. 1–7.

[30] J. Xia, S. Zhang, G. Cai, L. Li, Q. Pan, J. Yan, G. Ning, Adjusted weight voting algorithm for random forests in handling missing values, Pattern Recognition 69 (2017) 52–60.

[31] C. Hu, Y. Chen, L. Hu, X. Peng, A novel random forests based class incremental learning method for activity recognition, Pattern Recognition 78 (2018) 277–290.

[32] D. Ravì, M. Bober, G. M. Farinella, M. Guarnera, S. Battiato, Semantic segmentation of images exploiting DCT based features and random forest, Pattern Recognition 52 (2016) 260–273.

[33] J. Zhang, Y. Chen, E. Bekkers, M. Wang, B. Dashtbozorg, B. M. ter Haar Romeny, Retinal vessel delineation using a brain-inspired wavelet transform and random forest, Pattern Recognition 69 (2017) 107–123.

[34] D. Ni, X. Ji, M. Wu, W. Wang, X. Deng, Z. Hu, T. Wang, D. Shen, J.-Z. Cheng, H. Wang, Automatic cystocele severity grading in transperineal ultrasound by random forest regression, Pattern Recognition 63 (2017) 551–560.

[35] N. Lopes, B. Ribeiro, Towards adaptive learning with improved convergence of deep belief networks on graphics processing units, Pattern recognition 47 (2014) 114–127.

[36] Z. Zhao, L. Jiao, J. Zhao, J. Gu, J. Zhao, Discriminant deep belief network for high-resolution SAR image classification, Pattern Recognition 61 (2017) 686–701.

[37] J. Dennis, H. D. Tran, E. S. Chng, Overlapping sound event recognition using local spectrogram features and the generalised hough transform, Pattern Recognition Letters 34 (9) (2013) 1085–1093.

[38] J. Schröder, S. Goetze, J. Anemüller, Spectro-temporal Gabor filterbank features for acoustic event detection, IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP) 23 (12) (2015) 2198–2208.

[39] J. Ludeña-Choez, A. Gallardo-Antolín, Acoustic event classification using spectral band selection and nonnegative matrix factorization-based features, Expert Systems with Applications 46 (2016) 77–86.

[40] A. Plinge, R. Grzeszick, G. A. Fink, A bag-of-features approach to acoustic event detection, in: International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2014, pp. 3704–3708.

[41] E. Kiktova-Vozarikova, J. Juhar, A. Cizmar, Feature selection for acoustic events detection, Multimedia Tools and Applications 74 (12) (2015) 4213–4233.

[42] Z. Xiaodan, J. Huang, G. Potamianos, M. Hasegawa-Johnson, Acoustic fall detection using gaussian mixture models and GMM supervectors, in: International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2009, pp. 69–72.

[43] Y. Yang, J. Jiang, Bi-weighted ensemble via HMM-based approaches for temporal data clustering, Pattern Recognition 76 (2018) 391–403.

[44] J. Schröder, B. Cauchi, M. R. Schädler, N. Moritz, K. Adiloglu, J. Anemüller, S. Doclo, B. Kollmeier, S. Goetze, Acoustic event detection using signal enhancement and spectro-temporal feature extraction, in: Workshop on Applicat. Signal Process. Audio Acoust.(WASPAA), IEEE, 2013.

[45] W. Nogueira, G. Roma, P. Herrera, Automatic event classification using front end single channel noise reduction, MFCC features and a support vector machine classifier, IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (2013) 1–2.

[46] Y. Liu, K. Wen, Q. Gao, X. Gao, F. Nie, SVM based multi-label learning with missing labels for image annotation, Pattern Recognition 78 (2018) 307–317.

[47] A. Temko, C. Nadeu, Classification of acoustic events using SVM-based clustering schemes, Pattern Recognition 39 (4) (2006) 682–694.

[48] T. Heittola, A. Mesaros, T. Virtanen, M. Gabbouj, Supervised model training for overlapping sound events based on unsupervised source separation., in: International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013, pp. 8677–8681.

[49] J. Yu, C. Chaomurilige, M.-S. Yang, On convergence and parameter selection of the EM and DA-EM algorithms for gaussian mixtures, Pattern Recognition 77 (2018) 188–203.

[50] X. J. Xia, R. Togneri, F. Sohel, D. Huang, Random forest classification based acoustic event detection, in: IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2017, pp. 163–168.

[51] P. N. Tan, M. Steinbach, V. Kumar, Classification: basic concepts, decision trees, and model evaluation, Introduction to data mining 1 (2006) 145–205.

[52] G. E. Hinton, S. Osindero, Y. W. Teh, A fast learning algorithm for deep belief nets, Neural computation 18 (7) (2006) 1527–1554.

[53] K. J. Archer, R. V. Kimes, Empirical characterization of random forest variable importance measures, Computational Statistics & Data Analysis 52 (4) (2008) 2249–2260.

Xianjun Xia received the Bachelor degree in Electronic Engineering from the Hefei University of Technology (HFUT) in 2011 and Master degree in Signal and Information Processing from the University of Science and Technology of China (USTC) in 2014. Now he is a Ph.D. candidate in the School of Electrical, Electronic and Computer Engineering from The University of Western Australia (UWA) since 2015. His research interests include acoustic event detection, machine learning, speech and audio signal processing.

Roberto Togneri received the Ph.D. degree in 1989 from the University of Western Australia. He joined the School of Electrical, Electronic and Computer Engineering at The University of Western Australia in 1988, where he is now an Associate Professor. He leads the Signal Processing and Recognition Lab and his research activities in signal processing and pattern recognition include: feature extraction and enhancement of audio signals, statistical and neural network models for speech and speaker recognition, and audio-visual recognition and biometrics. He has published over 150 refereed journal and conference papers in the areas of signal processing and recognition, the chief investigator on three Australian Research Council Discovery Project research grants, and was an Associate Editor for *IEEE Signal Processing Magazine Lecture Notes and IEEE Transactions on Speech, Audio and Language Processing from 2012 to 2016.*

Ferdous Sohel received Ph.D. degree from Monash University, Australia in 2009. He is currently a Senior Lecturer in Engineering and Information Technology at Murdoch University, Australia. Prior to joining Murdoch University, he was a Research Assistant Professor/Research Fellow at the School of Computer Science and Software Engineering, The University of Western Australia from January 2008 to mid-2015. His research interests include computer vision, image processing, pattern recognition, multimodal biometrics, scene understanding, robotics, and video coding. He is a recipient of the prestigious Discovery Early Career Research Award (DECRA) funded by the Australian Research Council. He is also a recipient of the Early Career Investigators award (UWA) and the best PhD thesis medal form Monash University.

David Huang received the B. E. E. E. and M. E. E. E. degree in electronic engineering from Tsinghua University, Beijing, China, in 1996 and 1999, respectively, and the Ph.D. degree in electrical and electronic engineering from the Hong Kong University of Science and Technology (HKUST), Kowloon, Hong Kong, in 2004. He joined the School of Electrical, Electronic and Computer Engineering at the University of Western Australia in 2005 as a lecturer, and has been promoted to be a professor with the same school since 2011. Before joining UWA, he was a lecturer at Tsinghua University. He served as an Editor (2011-2015) for the IEEE Wireless Communications Letters, an Editor (2005-2011) for the IEEE Transactions on Wireless Communications, and the Editorial Assistant (2002-2004) to the IEEE Transactions on Wireless Communications. His research interest is signal processing, digital communications and artificial intelligence.