EXPLORING SUPERFRAME CO-OCCURRENCE FOR ACOUSTIC EVENT RECOGNITION

Huy Phan[1,2] *and Alfred Mertins*[1]

[1]Institute for Signal Processing, University of Lübeck, Germany
[2]Graduate School for Computing in Medicine and Life Sciences, University of Lübeck, Germany
Email: {phan, mertins}@isip.uni-luebeck.de

ABSTRACT

We introduce in this paper a concept of using acoustic superframes, a mid-level representation which can overcome the drawbacks of both global and simple frame-level representations for acoustic events. Through superframe-level recognition, we explore the phenomenon of superframe co-occurrence across different event categories and propose an efficient classification scheme that takes advantage of this feature sharing to improve the event-wise recognition power. We empirically show that our recognition system results in 2.7% classification error rate on the ITC-Irst database. This state-of-the-art performance demonstrates the efficiency of this proposed approach. Furthermore, we argue that this presentation can pretty much facilitate the event detection task compared to its counterparts, e.g. global and simple frame-level representations.

Index Terms— Acoustic event recognition, superframe, histogram, co-occurrence

1. INTRODUCTION

Detection of acoustic events is important in various applications [3–5]. However, building a robust acoustic event detection system, in which the category and the temporal location of events are determined, still remains a challenging task. The difficulty stems not only from how to discriminate events among different categories but also from the nature of overlapping events, the large intra-class variations in terms of event duration and sounds, as well as non-stationary background noise. Various attempts have been reported to tackle the problem. Most of them borrow the speech recognition framework where they employ simple frame-based presentation of the audio, and individual events are modelled as Hidden Makov Models (HMMs) to represent higher-level structure [1, 2]. However, HMMs require the training-data size to be large enough to estimate probabilistic distribution. On the other hand, the systems using discriminative models, e.g. Support Vector Machines (SVM) [5], and hybrid models, e.g. HMM-SVM combinations [6], have shown superior performance. A good-performance classifier telling apart events of different categories particularly plays an important role in such detection systems.

In literature, the recognition strategies employ models that work directly on global feature vectors derived from whole audio segments of the events [5, 7], which fail to capture local features as well as their temporal structure. Additionally, the simple frame level characterization, e.g. 30 ms, of audio can result in significantly inferior performance [8]. The work of [8] also shows that the events themselves embed temporal structures of acoustic units, and the occurrence patterns of these mid-level characterizations can be used for event recognition. Inspired by this, in this work, we introduce the concept of *acoustic superframe* and represent an event as a collection of superframes. Through studying the ambiguity of superframe-wise recognition we empirically show that the co-occurrence of superframes frequently happens among event categories. That is, different event categories share some common superframes. To the best knowledge of the authors, although the phenomenon is typical for acoustic signals, it has not been explored and utilized to enhance acoustic event recognition and detection system. We propose a classification scheme that takes this information into account to significantly boost the event discrimination power.

The rest of the paper will be organized as following. In Section 2, we introduce the concept of acoustic superframe and its presentation, followed by investigation of superframe co-occurrence phenomena through analysis of superframe-wise ambiguity. Section 3 will describe how to exploit superframe co-occurrence to improve acoustic event recognition. Next, we present the experimental results in Section 4, followed by the discussion and conclusion in Section 5.

2. EVENT SUPERFRAME AND ITS REPRESENTATION

2.1. The concept of acoustic superframe

The problem with the global presentations of acoustic events as in [5, 7] is that the local features and their temporal information of the events are lost. Also, these global feature pre-

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].

sentations do not facilitate event detection since we need to search on large temporal scale space due to the high variance of event duration. On another extreme, although the frame level presentation offers fine temporal resolution, it appears to be too noisy for high-accuracy recognition [8]. It is very common that these frame level presentations are combined to form a global presentation using some statistical measures, such as mean and standard deviation as in [5]. It raises a need for a mid-level presentation that can overcome the disadvantages of both global and frame-level presentations. A such presentation should: (1) sufficiently capture the signal distribution for the recognition task; (2) preserve the local features and their temporal structure of the events; (3) offer a satisfactory temporal resolution to ease the detection task.

We define a *superframe* as a 100 ms segment of acoustic signal. And a superframe contains multiple small frames, hence its name. The rational behind the adoption of this presentation are numerous:

- It is obvious that local event features are preserved and their structure can also be kept if we consider their temporal order.
- As can be shown in the next section, 100 ms segments alone are semantically acceptable for event recognition. By naively considering an event as a collection of superframes, the event-wise recognition can be noticeably improved with a simple voting scheme and close to the state-of-the-art system on the same dataset.
- For the event detection task, the detection error tolerance is usually set to 100 ms as in the most recent campaigns [10–12]. Hence, its temporal resolution is sufficient for event detection in superframe fashion. The temporal resolution can be further improved by overlapped sampling.

Therefore, the superframe representation meets the strict requirements above. By taking into account the superframe co-occurrence across event categories, our classification system sets state-of-the-art performance.

2.2. Acoustic features for superframe representation

For a superframe, we divide the audio signal into interleaved small frames of 20 ms with Hamming window and 50% overlap. In order to characterize a frame, we utilize the set of acoustic features suggested by Temko and Nadeu [5] since they have been proven to represent speech spectral structure well in CLEAR Evaluations for acoustic event recognition and detection task [10, 11]. They 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, 4 sub-band energies, spectral flux calculated for each sub-band, spectral centroid, and spectral bandwidth. It results in a 60-dimensional feature vector for each frame. In turn, the empirical mean and standard deviation of frame features are calculated to form a 120-dimensional feature vector to represent a superframe.

3. EVENT SUPERFRAME CO-OCCURRENCE

3.1. ITC-Irst acoustic event database

We use the database ITC-Irst of isolated meeting-room acoustic events [14] throughout the experiments of this paper. This database has originally been collected under the CHIL (Computer in the Human Interaction Loop) project [13]. Event detection and classification using this database have been extensively examined in recent CLEAR Evaluations [10, 11]. The data consists of 12 sessions each of which is of approximately 7-minute duration recorded by multiple microphones. We used only one channel, named *TABLE_1*, in our experiments.

The database contains 16 semantic classes of variable-length acoustic events. Each session contains around four repetitions of each of the event classes, resulting in about 36 examples of each event. The data labels are also provided with short intervals that contain instances of the labeled sound. We are only interested in 12 semantic classes that are investigated as in the CLEAR Evaluations including: door knock, door slam, steps, chair moving, spoon cup jingle, paper wrapping, key jingle, keyboard clicking, phone ring, applause, cough, and laugh. Many of them are subtle (low SNR, e.g. steps, chair moving, and keyboard typing) making the task more challenging. Following the setup of event classification in CLEAR Evaluations, we use the 9 sessions as training files and 3 remaining sessions as testing files in our experiments.

3.2. Part-wise recognition and part co-occurrence

We empirically study the superframe co-occurrence between different event categories through superframe-wise event classification. By analysis of the classification confusion matrix, we are able to show that different event categories share some common superframes at different levels. Foremost, the audio signal is down-sampled from 44.1 kHz to 16 kHz. Given the audio signal of an event, we divided it into multiple interleaved superframes with 75% overlap. Each superframe is represented by the features described previously and is labelled using the label of the event. As a result, an event is a collection of superframes. For the dataset we use the produced superframe-wise training, and testing data contains 74,322 and 25,078 samples respectively. This data is large enough to prevent most popular classification algorithms, such as non-linear SVM [15], from performing efficiently.

Fortunately, *Random Forest* [9] is particularly suitable for this purpose since it has been proven to be efficient for data with large number of samples and dimensions. The main idea behind Random Forest is to mitigate over-fitting and lack of generalization problems of decision-tree classifiers by: (1) injecting randomness into the training of the trees, and (2) combining the output of multiple randomized trees into a single classifier. Random Forests have been demonstrated to produce lower test errors than conventional decision trees

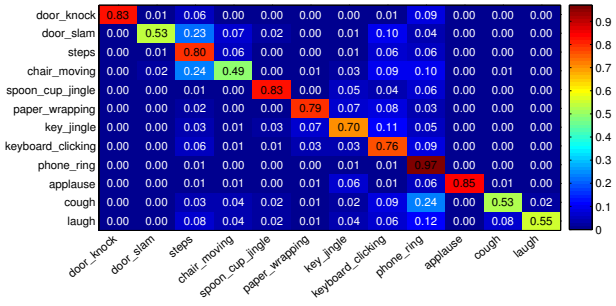

Fig. 1. Superframe-wise classification confusion matrix.

and performance comparable to SVMs in multi-class problems, such as [19], while maintaining high computational efficiency.

Let $\{(x_i, y_i)\}_{i=1,\ldots,N_{tr}}$ denote the training data and $\{(x_i, y_i)\}_{i=1,\ldots,N_{te}}$ denote the testing data where $x_i \in \mathcal{R}^D$ and $y_i \in \{1, \ldots, \mathcal{Y}\}$ denote the feature vector and label of the superframe i respectively. N_{tr} and N_{te} correspond to the cardinality of training and testing data. $D = 120$ and $\mathcal{Y} = 12$ are the dimensionality and the number of event categories, respectively. We train a random-forest classifier to classify the event superframes into 12 semantic classes using the training data and test the model with the testing data. We conservatively set the number of trees to $T = 500$ and choose a maximum depth of 30. The event categories are weighted by their inverse frequencies. The overall superframe-wise classification error is approximately 23.0%. This suggests that superframe presentation is informative for event recognition.

The superframe-wise classification confusion matrix is illustrated in Fig. 1. Each row of the matrix shows the testing probabilities in classifying the superframes of the corresponding event category as other categories. It can be seen that for every event categories, a certain amount of superframes are wrongly classified as other event categories. This suggests that different event categories show overlap in the feature space and have similar superframes. While the event duration is in the order of seconds, the ambiguity is understandable since it is not evident enough to tell apart between event superframes in a short duration of 100 ms. It is even more difficult for low SNR events such as 'steps' and 'chair moving'. It is also interesting to notice that 'door_slam' and 'chair_moving' superframes are most confused with 'steps' superframes because they are similar in short time. Furthermore, most of the events are regularly wrongly classified as 'phone_ring', especially the periodic events like 'cough', and 'laugh', owing to not only the periodicity of 'phone_ring' audio signal but also to its high variance of sounds.

3.3. Integration of superframe co-occurrence: from majority voting to accumulated histogram

The question now is how to fuse the superframe-wise recognition to accomplish the event-wise recognition. To combine

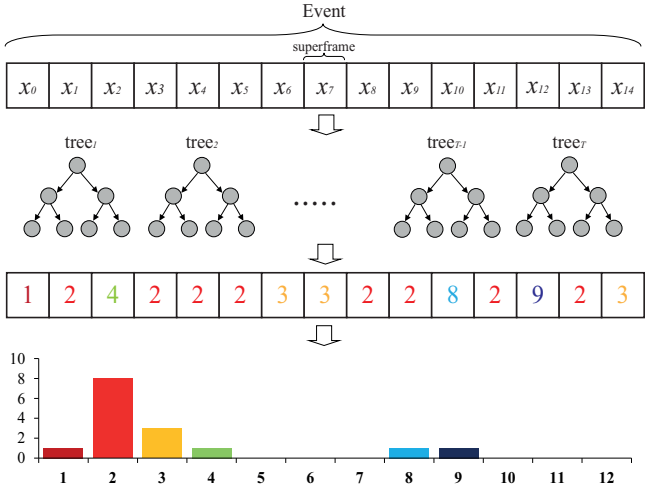

Fig. 2. Procedure to produce superframe histogram representation.

the superframe-wise recognition results, we employ *majority voting*:

$$\hat{y}_{event} = \underset{y \in \{1,\ldots,\mathcal{Y}\}}{\operatorname{argmax}} \sum_{p=1}^{P} \mathcal{I}(\hat{y}_p = y). \tag{1}$$

In (1), $\hat{y}_{event}$ and $\hat{y}_p$ denote the predicted labels of the final event and the superframe i where P is the number of superframes belonging to the event. $\mathcal{I}(\hat{y}_p = y)$ is the indicator function given by:

$$\mathcal{I}(\hat{y}_p = y) = \begin{cases} 1 & \text{if} \quad \hat{y}_p = y \\ 0 & \text{if} \quad \hat{y}_p \neq y. \end{cases} \tag{2}$$

As a result, the predicted label of the event is determined by the majority of its superframes' predicted labels. It magnificently reduces the overall classification error from 23.0% superframe-wise to 7.5% event-wise.

However, this voting scheme is efficient for the event categories with minor superframe sharing like 'applause' but not for those with relatively serious ambiguity such as 'laugh' and 'chair moving'. Instead of ignoring superframe sharing, we can take advantage of it to gain the evidence for event recognition. Intuitively, it is more informative to say "a 'chair_moving' event should contain a percent of 'chair moving' superframes and b percent of 'steps' superframe and so on" rather than "a 'chair_moving' event should only contain a percent of 'chair_moving' superframes".

The idea of taking event superframe co-occurrence into account is illustrated in Fig. 2. For each event consisting of P superframes $\{x_p\}_{p=1,\ldots,P}$ with respect to the predicted labels $\{\hat{y}_p\}_{p=1,\ldots,P}$ outputted by the superframe classifier, we accumulate all $\hat{y}_p$ into a label histogram $z \in \mathcal{R}_+^{\mathcal{Y}}$ with each element z_i given by

$$z_i = \sum_{p=1}^{P} \mathcal{I}(\hat{y}_p = i). \tag{3}$$

By this, we can keep all information about superframe co-occurrence in the event representation. Eventually, the obtained histogram vectors are used as feature vectors for the events. We will exploit them to learn an event-wise classifier for the event recognition task in the following section.

4. EXPERIMENTS

4.1. Creation of event-wise training and testing data

We need somehow to generate event-wise training and testing data using the procedure of forming histogram presentation in Fig. 2. For testing data, it is straightforward to use the random forest superframe classifier as in Section 3.2 to run over all the event audio signals in the testing files. However, it is more tricky for training data since they do not readily exist. We cannot simply run the superframe classifier learned from the training files to run over them again, because they are prone to overfitting. To overcome this, we conducted 9-fold sub-training on the 9 training files. Each time, we used 8 out of 9 files to train a superframe-wise classifier using the Random Forest algorithm and conducted superframe classifying with the remaining file. The superframe predicted labels are used to form the event-wise histogram representations for all events contained in that file as in Fig. 2. Finally, we concatenate event-wise histogram representations in all 9 runs and use them as training data.

4.2. Experiment

The histogram vectors are firstly normalized by l_1-norm. Using the event-wise trainining data, we employ the C-SVM classification algorithm [15] to learn two event-wise classifiers, SVM$_{hist+chi}$ and SVM$_{hist+int}$, with *Chi-square* and *histogram intersection* kernels [17], respectively. For normalized histogram based feature vectors $x, z \in \mathcal{R}_+^{\mathcal{Y}}$, Chi-square kernel $\mathcal{K}_{\chi^2}(x, z)$ and histogram intersection kernel $\mathcal{K}_{int}(x, z)$ are defined as

$$\mathcal{K}_{\chi^2}(x, z) = \sum_{i=1}^{\mathcal{Y}} \frac{2x_i z_i}{x_i + z_i}, \tag{4}$$

$$\mathcal{K}_{int}(x, z) = \sum_{i=1}^{\mathcal{Y}} \min(x_i, z_i). \tag{5}$$

While these kernels are very fast to evaluate, they are also particularly proven to be the best-suited kernels and most frequently used for histogram presentations [17]. We used *libSVM* [16] in our experiments. The parameter C of the SVM classifier was set to 1.0 for both SVM$_{hist+chi}$ and SVM$_{hist+int}$ since we found that the leave-one-out cross-validation error is always minimized around this value.

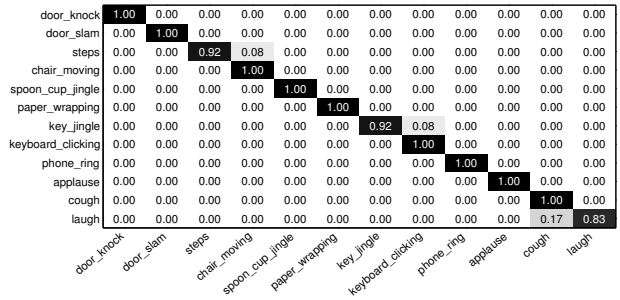

Fig. 3. Event-wise classification confusion matrix.

4.3. Results

We tested the classifiers SVM$_{hist+chi}$ and SVM$_{hist+int}$ on the event-wise testing data. The event-wise classification errors were significantly reduced to 2.7% and 3.4%, respectively, which significantly outperforms the majority voting scheme. This result quantifies the usefulness of the superframe sharing in event recognition. We tabulate the event-wise classification confusion matrix with the best classifier SVM$_{hist+chi}$ in Fig. 3. To demonstrate the efficiency of our approach, we further compare the performance in terms of recognition error rate of SVM$_{hist+chi}$ and SVM$_{hist+int}$ with:

- SVM$_{hist+linear}$, the event-wise SVM classifier learned from histogram presentations using linear kernel;
- SVM$_{hist+RBF}$, the event-wise SVM classifier learned from histogram presentations using non-linear RBF kernel;
- SVM$_{global+RBF}$, the event-wise SVM classifier trained on the global presentations with non-linear RBF kernel.

The linear and nonlinear RBF kernels [15] are given by (6) and (7), respectively:

$$\mathcal{K}_{linear}(x, z) = x^{\mathsf{T}} z, \tag{6}$$

$$\mathcal{K}_{RBF}(x, z) = e^{-\gamma ||x-z||^2}. \tag{7}$$

The setting and parameter search for SVM$_{hist+linear}$ are similar to what has been done for $SVM_{hist+chi}$ and SVM$_{hist+int}$. For SVM$_{global+RBF}$, to extract the event global presentation, we divide each event signal into 30 ms frames with Hamming window and 50% overlap. We utilize the same set of 60 features described in Section 2.2 to characterize each frame. The global feature vector of an event is produced by calculating empirical mean and standard deviation of its frame feature vectors. In addition, the global feature vectors are normalized into the range [-1;1]. The same grid parameter search is done for both SVM$_{hist+RBF}$ and SVM$_{global+RBF}$ with leave-one-out cross validation for the parameters C and γ. The coarse grid search, corresponding to $\log C \in [-5; 8]$ and $\log \gamma \in [-8; 3]$ with a common step of 1.0, is first performed, followed by the fine grid search over $\log C \in [-1; 1]$ and $\log \gamma \in [-1; 1]$ with a common step

Table 1. Comparison of classification error rates (in %) for different event classifiers.

	$\text{SVM}_{hist+chi}$	$\text{SVM}_{hist+int}$	$\text{SVM}_{hist+linear}$	$\text{SVM}_{hist+RBF}$	$\text{SVM}_{global+RBF}$	UPC-C	CMU-C1	ITC-C1
Error rate	**2.7%**	3.4%	4.8%	4.8 %	3.4%	4.1%	7.5%	12.3%

of 0.1 around the optimal coarse parameters. The classifiers are finally trained with the found optimal parameters on the training data and evaluated on the testing data. In addition, we also compare the performance with the systems submitted to CLEAR 2006 campaign [18] on the same dataset, including *UPC-C*, *CMU-C1*, and *ITC-C1*. The comparison results are shown in Table 1.

As can be seen, $\text{SVM}_{hist+chi}$ outperforms all the other systems and some with a large margin. The results also show that linear and RBF kernels are equally efficient for superframe histogram representation in our experiments while Chi-square kernel is the most efficient for this. We argue that the use of global features, which are deteriorated by averaging operator, explains the inferior result of $\text{SVM}_{global+RBF}$ compared to $\text{SVM}_{hist+chi}$.

5. DISCUSSION AND CONCLUSION

Although our approach achieves state-of-the-art performance on the event-recognition task by accumulating the superframe predicted labels into histogram representations, we believe that the performance can be further improved by considering the temporal order of the superframes. As argued, the superframe representation offers a satisfactory temporal resolution for the event-detection task, and superframe-wise detection would simplify the detection process, especially in real-time scenarios. However, we need to deal with the question of how to use superframe-wise detection results to determine the boundaries of the target event in time. These are worth further studying.

In conclusion, we presented in this paper the concept of acoustic superframe and study the phenomena of superframe co-occurrence across event categories. We empirically showed that taking advantage of this phenomenon into event-wise recognition can significantly improve the recognition model. Our classification system with histogram representation and Chi-square kernel yields state-of-the-art performance in terms of classification error rate on the ITC-Irst database.

REFERENCES

[1] K. Lee, D. Ellis, and A. Loui, "Detecting local semantic concepts in environmental sounds using markov model based clustering," in *ICASSP*, 2010.

[2] A. Mesaros, T. Heittola, A. Eronen, T. Virtanen, "Acoustic event detection in real life recordings," in *EUSIPCO*, 2010.

[3] J. Schröder, S. Wabnik, P.W.J. van Hengel, and S. Götze, "Detection and Classification of Acoustic Events for In-Home Care," *Ambient Assisted Living*, Springer, 2011.

[4] P.K. Atrey, N.C. Maddage, and M.S. Kankanhalli, "Audio based event detection for multimedia surveillance," in *ICASSP*, 2006.

[5] A. Temko and C. Nadeu, "Acoustic event detection in meeting-room environments," *Pattern Recognition Letters*, vol. 30, pp. 1281-1288, 2009.

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

[7] J. Dennis, H.D. Tran, and E.S. Chng, "Image Feature Representation of the Subband Power Distribution for Robust Sound Event Classification," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 21, no. 2, pp. 367-377, 2013.

[8] A. Kumar, P. Dighe, R. Singh, S. Chaudhuri, and B. Raj, "Audio event detection from acoustic unit occurrence patterns," in *ICASSP*, 2012.

[9] L. Breiman, "Random Forest," *Machine Learning*, vol. 45, pp. 5-32, 2001.

[10] "CLEAR 2006: Classification of Events, Activities and Relationships. Evaluation and Workshop," `http://isl.ira.uka.de/clear06`

[11] "CLEAR 2007: Classification of Events, Activities and Relationships. Evaluation and Workshop," `http://www.clear-evaluation.org/`

[12] "IEEE AASP Challenge 2013: Detection and Classification of Acoustic Scenes and Events," `http://c4dm.eecs.qmul.ac.uk/sceneseventschallenge/`

[13] "CHIL - Computers in the human interaction loop," `http://www.ipd.uka.de/CHIL/`

[14] C. Zieger and M. Omologo, "Acoustic event detection - ITC-Irst AED database," Internal ITC report, Tech. Rep., 2005.

[15] A.J. Smola, B. Schölkopf, "Learning with Kernels," *MIT Press*, 2002.

[16] C.C. Chang and C.J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. on Intelligent Systems and Technology*, vol. 2, no. 3, 2011.

[17] S. Maji, A.C. Berg, and J. Malik, "Efficient Classification for Additive Kernel SVMs," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 66-77, 2013.

[18] 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 Computer Science*, vol. 4122, pp. 311-322, 2007.

[19] A. Bosch, A. Zisserman, and X. Munoz, "Image classification using random forests and ferns," in *ICCV*, 2007.