Adaptive Semi-supervised Tree SVM for Sound Event Recognition in Home Environments

Ng Wen Zheng Terence*, Tran Huy Dat*, Huynh Thai Hoa* and Chng Eng Siong†
*Institute for Infocomm Research, A*STAR, Singapore
†School of Computer Engineering, Nanyang Technological University
wztng@i2r.a-star.edu.sg, hdtran@i2r.a-star.edu.sg, thhuynh@i2r.a-star.edu.sg, aseschng@ntu.edu.sg

***Abstract*—This paper addresses a problem in sound event recognition, more specifically for home environments in which training data is not readily available. Our proposed method is an extension of our previous method based on a robust semi-supervised Tree-SVM classifier. The key step in this paper is that the MFCC features are adapted using custom filters constructed at each classification node of the tree. This is shown to significantly improve the discriminative capability. Experimental results under realistic noisy environments demonstrate that our proposed framework outperforms conventional methods.**

I. Introduction

Sound event recognition (SER) is the task of understanding real-life events using sound information. In this paper, we place our emphasis on SER in home environments which has a wide range of important applications, such as acoustic surveillance [1], smart home automation [2] and healthcare monitoring systems [3]. For these applications however, assessing or collecting large databases has always been a big challenge. Therefore in such situations, active learning methods, particularly semi-supervised methods are usually the most appropriate solution for this limitation. These methods overcome the lack of data problem by carefully selecting unlabelled data from testing to use for retraining.

In our previous work in [4], we have successfully developed an effective classification scheme, called Semi-supervised Tree Super Vector Machine (SST-SVM). It was designed specifically for cough monitoring where limited training data is usually a problem. The advantages of SST-SVM over conventional classification methods can be summarized as below:

- Discriminative hierarchy classifier: Tree SVM built upon Fisher Linear Discriminant (FLD) [5] provides good discriminating capability.
- Self-learning capability: Semi-supervised self-update of Tree model by retraining from unlabelled test data, using a confidence metric.

In this paper, we would like to extend our previous scheme by designing an adaptive feature, referred to as Adaptive MFCC, where the features are adapted at each classification node of the SST-SVM. This feature is generated by constructing filters that weight the spectral coefficients of each frequency bin according to its ranking in separability measurements with FLD. These custom filters are subsequently used as preprocessing before conventional MFCC extraction.

We note that, a few similar works on supervised filter design for characterizing phoneme have been reported in [6] and [7]. However, those methods are based on specific characteristics (e.g. formants) or structures (e.g. phoneme connection) of speech signals and therefore cannot be applied to sound events in general.

With the integration of Adaptive MFCC into SST-SVM, we refer to the new framework as ASST-SVM. In this paper, ASST-SVM is applied to healthcare applications in home environment, particularly for the application of bathroom monitoring. The task is to recognise events occurring in the bathroom under different noisy conditions. This is an important healthcare application [8] as we can monitor and foresee dangerous situations which are about to happen. Evaluation is carried out on a comprehensive database, consisting of 2000 sound events clips from four sound event classes: 'Door', 'Flush', 'Speech' and 'Tap'. The experimental results proved the effectiveness of the proposed ASST-SVM method.

The organization of this paper is as follows: Section 2 introduces the proposed Adaptive MFCC feature. Section 3 describes the integration of Adaptive MFCC into SST-SVM. Section 4 then presents experimental results and discussions before Section 5 concludes the work.

II. The Adaptive MFCC feature

In this section, we introduce the Adaptive MFCC feature that will be integrated into SST-SVM. The central idea here is to design a specific filter for the MFCC feature to enhance the discriminative capability at each binary classification node. We exploit the fact that between two classes, the characteristic, high-power frequency components, which best discriminates the classes, usually occurs in different bins for each class. Our aim is to give higher weights to such frequency bins and lower weights to bins that are less discriminative. With the filters, our aim is to make the classes more discriminative, so it remains robust even when noise is present in the signal.

The algorithm to generate the proposed custom filter is shown in Algorithm 1, and proceeds as follows. Using clean training data from two classes i, j at each binary classification node, we first compute the Short-Time Discrete Fourier Transform (ST-DFT) for each sound clip:

$$X_m(k,t) = \left| \sum_{n=0}^{nFFT-1} x_{m,t}[n] w[n] \, e^{-i2\pi \frac{k}{nFFT} n} \right| \quad (1)$$

where m is the m^{th} sound clip, k is the frequency bin, t is the frame index, $nFFT$ is the number of samples per frame and $w(.)$ is the Hamming window. Each computed ST-DFT for each sound clip is normalised to the minimum number of frames of all sound clips between the two classes at the given node.

Next, we append each subband, k, into matrices M_i^k or M_j^k as follows:

$$M_i^k = \begin{bmatrix} X_1(k,t) \\ X_2(k,t) \\ \vdots \\ X_{n_i}(k,t) \end{bmatrix}, M_j^k = \begin{bmatrix} \hat{X}_1(k,t) \\ \hat{X}_2(k,t) \\ \vdots \\ \hat{X}_{n_j}(k,t) \end{bmatrix} \quad (2)$$

where n_i and n_j is the number of samples for class i and j respectively. Then, FLD analysis [5] is used to find its separability index $r_{ij}(k)$ between the subband matrices of the two classes. A higher value of r_{ij} means that it is highly discriminative and vice versa.

Finally, we normalise the separability values as follows:

$$W_{ij}(k) = \frac{r_{ij}(k)}{\sum_{k=1}^{\text{nFFT}} r_{ij}(k)}, \quad \forall k$$

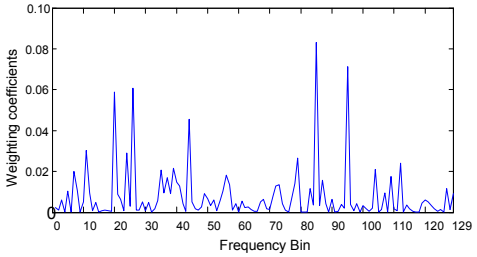
and the custom filter for each pair is $W_{ij} = \{W_{ij}(k)\}$. During training and testing, at each binary classification node comparing classes i and j, the spectrum is multiplied with W_{ij} before conventional MFCC extraction to give the proposed Adaptive MFCC feature for that node.

Figure 1 illustrates an example of how the custom filter appears for the classes considered in the bathroom monitoring application. Sub-bands with higher discriminability have been assigned higher weights and vice versa. Note that the choice of classification pairs are based on the structure of our final Tree-SVM, which will be presented in the next section. The effect of discriminative filter can be visualised as in Figure 2. Here, the spectrogram of a representative signal from class 'Door' is displayed and the changes can be observed before and after discriminative filtering.

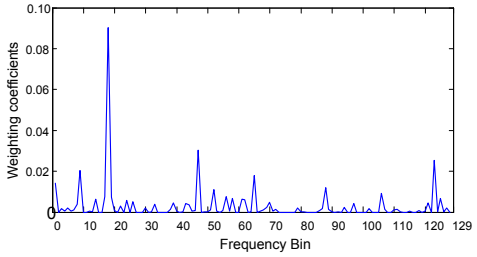
Algorithm 1 Designing a discriminative filter

Require: Clean training data from two classes $\{i; j\}$

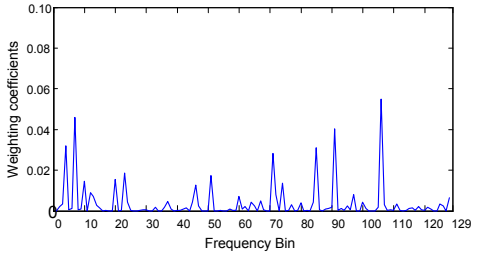
1: Compute ST-DFT for every sound clip $x_{m,t}[n] \rightarrow |X_m(k,t)|$.
2: Normalise the total number of frames for each sound clip.
3: **for all** subbands k, such that $1 \leq k \leq \frac{nFFT}{2} + 1$ **do**
4: Extract k-th subband $|X_m(k,t)|$ for every samples.
5: Append $|X_m(k,t)|$ into matrices M_i^k or M_j^k.
6: Do FLD analysis, get separability index $r_{ij}(k)$.
7: **end for**
8: Normalise $W_{ij}(k) = \frac{r_{ij}(k)}{\sum_k r_{ij}(k)}$ for all $1 \leq k \leq \frac{nFFT}{2} + 1$
9: **return** Discriminative filter $W_{ij} = \{W_{ij}(k)\}$



(a) Node 1: Class 1 = 'Speech', Class 2 = {'Door', 'Flush', 'Tap'}



(b) Node 2 : Class 1 = 'Door', Class 2 = {'Flush', 'Tap'}



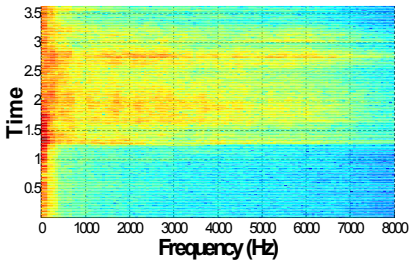
(c) Node 3 : Class 1 = 'Flush', Class 2 = 'Tap'

Figure 1. The discriminative filters at different binary nodes of Tree-SVM.

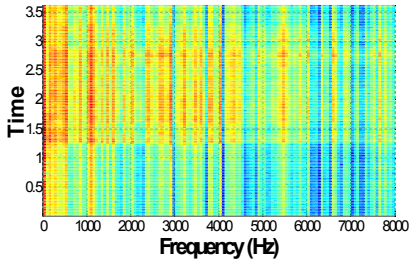
III. Integrating Adaptive MFCC feature into SST-SVM

The SST-SVM was introduced in our previous paper [4] to solve the problem of limited availability of training data. The method uses semi-supervised learning, as shown in Figure 3, where useful testing data based on a confidence metric can be used for retraining. The method's main strengths lie in the self-learning capability and the discriminative hierarchical tree-SVM classifier. We propose to extend this strength by introducing the Adaptive MFCC for each classification node of the SST-SVM. The goal is to further improve the discriminability of the original SST-SVM framework.

In the initial training phase, the new ASST-SVM is designed according to the Algorithm 2. For testing, each test sample will go through all tree junctions with the specific Adaptive MFCC and the corresponding SVM model until it is finally classified. A confidence metric is computed as the distance-to-hyperplane at the final junction where it is being recognized. This confidence metric is used as a threshold for semi-supervised training of the unlabelled test data. We note that in the retraining process, there is no redesign for both the Adaptive MFCC or structure of the ASST-SVM, only parameters of binary SVM models at each junctions are



(a) Original Spectrogram



(b) Spectrogram with custom filter

Figure 2. Effects of the discriminative filter applied on spectrogram of a 'Door' signal.

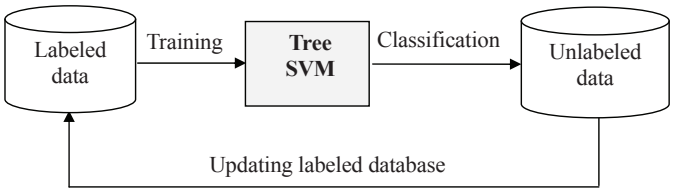


Figure 3. *Overview of the SST-SVM*

updated.

Algorithm 2 Design structure of a ASST-SVM

Require: Clean training data from all the classes

1: **for all** nodes of tree **do**
2: Search all possible binary groupings as in [4].
3: Design discriminative filter for each groupings as in Algorithm 1; apply it on training data.
4: Continue extract features and design node as in [4].
5: **end for**
6: **return** tree

Figure 4 shows an example of a final ASST-SVM for bathroom monitoring application, where the main interest is to classify the following four types of sound events: Door, Flush, Speech, Tap. From the figure, we observe that the structure generated by our algorithm agrees well with human perception: (1) human speech is separable with the other three, which are non-voiced sounds, (2) flushing and tap both involve with liquid running which sounds very different from the door sound. In addition, we note that at every junction of this tree structure, an adaptive MFCC feature with its unique filter is used for each binary classification node.

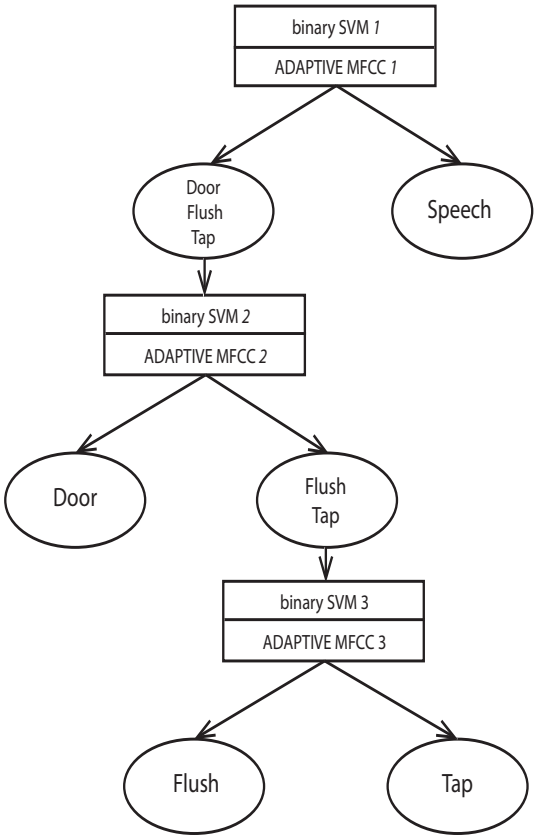


Figure 4. *ASST-SVM: SST-SVM with Adaptive MFCC.*

IV. Experiments

A. Databases

In this section we carry out experiments to validate the performance of our proposed system. Testing is performed on a comprehensive database consisting of sound events commonly occurring in the bathrooms. The classes are placed in a moderately sized bathroom with a single omni directional microphone at a sampling frequency of 16000 Hz. The microphones are recorded near the events to achieve a high signal-to-noise (SNR) ratio. We have chosen four classes related to bathroom sound events:

1) Door: Sound generated from door opening/closing
2) Flushing: Sound of water gushing out from water tank into the toilet bowl
3) Speech: Conversational human speech
4) Tap: Sound of water running from the tap into the basin at different speeds

Each class has a total 20 samples for training and 500 samples for testing. Note that we used a very low number of training samples to simulate our problem of lack of data. For evaluation, noise were added at different signal-to-noise ratio: clean, $15dB$, $10dB$, $5dB$. We have chosen a bathroom related noise namely sound generated from a washing machine, taken from Audio Pro European SFX Library [9].

Table I
Recognition results (classification accuracy in percentage %).

Method	Feature	Clean	15dB	10dB	5dB
OAO-SVM	MFCC	99.55	83.15	66.95	55.30
	Adapted MFCC	99.35	92.05	79.80	61.45
T-SVM	MFCC	99.55	96.00	81.00	63.70
	Adapted MFCC	99.30	95.30	87.95	77.50
SS-OAO-SVM	MFCC	99.60	92.85	84.25	68.00
	Adapted MFCC	99.50	97.55	89.85	82.55
SST-SVM	MFCC	**99.65**	83.50	97.35	68.95
	Adapted MFCC	99.50	**98.60**	**98.75**	**92.65**

B. Experiments setup

For a comprehensive evaluation, we compare our Tree SVM with conventional one-against-one SVM. Also, in both cases, we evaluate it using with and without semi-supervised training. In summary, the following four methods are used for evaluation:

1) OAO-SVM: One-against-One SVM
2) SS-OAO-SVM: Semi-Supervised OAO-SVM
3) T-SVM: Tree SVM
4) SST-SVM: Semi-Supervised Tree SVM

All four methods are evaluated with both features: conventional MFCCs and the proposed Adaptive MFCCs. The MFCC used is set up with standard configuration with 36-dimensions including deltas. The threshold of confidence metric for retraining is optimized through experiments. Lastly we note that all SVMs are setup with the linear kernel [10].

C. Results and discussions

Table I summarizes the empirical results of our experiments on the bathroom sound event database. Here it can be seen that in clean conditions, all the methods performed similarly well, with negligible improvements in comparison with each other. However in noisy conditions, it can be clearly seen that performances of the four methods follow a consistent pattern that reinforces our previous claims and findings in [4]. In particular, T-SVM always performs better than conventional OAO-SVM, with semi-supervised training always providing a boost in performance. Overall, SST-SVM with both tree structure and semi-supervised outperforms all classifiers.

Next, comparing the performance of the two features, our proposed Adaptive MFCC consistently outperforms the conventional MFCC with all classification methods. This is true under each mismatched SNR condition and this affirms that our proposed feature is more discriminative. Also, we observed that the improvements are greater with the presence of higher levels of noise. In particular, the best improvement is observed at 5dB of noise, where Adaptive MFCC is used with SST-SVM and provides an absolute improvement of 23.7%.

In summary, SST-SVM with Adaptive MFCC (ASST-SVM) is the top-performer among all classifiers and features in noisy conditions. The superior performance can attributed by the following factors: (1) Semi-Supervised learning - the mechanism for self-updating new useful information, keep up with changes in environments (2) Discriminative hierarchy classifier - provides higher discrimination power (3) Adaptive feature - further improves the discriminability, especially at low SNRs.

V. Conclusions

This paper is a extension of our robust method presented in [4]: semi-supervised learning with a discriminative classifier for sound event recognition in healthcare applications. The experimental results presented in this paper reaffirms the robustness of the method under problems of both noisy conditions and the lack of training data. Moreover, in this paper, the experimental results shows that the combination of the Adaptive MFCC with SST-SVM outperformed conventional classification methods as well as our previous method. We conclude that the robustness is further improved by our proposed Adaptive MFCC, under the ASST-SVM framework.

References

[1] C. Clavel, T. Ehrette, and G. Richard, "Events detection for an audio-based surveillance system, in Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on. IEEE, 2005, pp. 13061309.

[2] J.C.Wang, H.P. Lee, J.F.Wang, and C.B. Lin, "Robust environmental sound recognition for home automation, Automation Science and Engineering, IEEE Transactions on, vol. 5, no. 1, pp. 2531, 2008.

[3] Vacher, Michel, et al. "Speech and sound use in a remote monitoring system for health care." Text, Speech and Dialogue. Springer Berlin Heidelberg, 2006.

[4] T.H.Huynh, V.A.Tran and H.D.Tran, "Semi-Supervised Tree Support Vector Machine for Online Cough Recognition", in *Proc. 12th International Conference of the International Speech Communication Association*, 2011, pp.1637-1640.

[5] R. O Duda, P. E. Hart, D. H. Stork. Pattern Classication (2nd ed.),Wiley Interscience, MR1802993, ISBN 0-471-05669-3.

[6] T.Kinnunen, "Designing a Speaker-Discriminative Adaptive Filter Bank for Speaker Recognition", in *Proc. 7th International Conference on Spoken Language Processing*, 2002, pp. 2325-2328.

[7] S.H.Mohammadi, H.Sameti, A.Tavanaei and A.Soltani-Farani, "Filter-bank Design Based on Dependencies Between Frequency Components and Phoneme Characteristics", in *Proc. 19th European Signal Processing Conference*, 2011, pp. 2142-2145.

[8] Chen, Jianfeng, et al. "Bathroom activity monitoring based on sound." Pervasive Computing. Springer Berlin Heidelberg, 2005. 47-61.

[9] Audio Pro European SFX Library URL: http://www.soundideas.com

[10] G.Fung and O.L.Mangasarian, "Proximal support vector machine classifiers", in *Proc. of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2001, pp. 77-86.