Pseudo-Color Cochleagram Image Feature and Sequential Feature Selection for Robust Acoustic Event Recognition

Roneel V Sharan[1] and Tom J Moir[2]

[1]School of Information Technology and Electrical Engineering, The University of Queensland, Brisbane, QLD 4072, Australia

[2]School of Engineering, Auckland University of Technology, Private Bag 92006, Auckland 1142, New Zealand

Email: r.sharan@uq.edu.au, tom.moir@aut.ac.nz

Corresponding Author: Roneel V Sharan (Email: r.sharan@uq.edu.au)

Abstract

This work proposes the use of pseudo-color cochleagram image of sound signals for feature extraction for robust acoustic event recognition. A cochleagram is a variation of the spectrogram. It utilizes a gammatone filter and has been shown to better reveal spectral information. We propose mapping of the grayscale cochleagram image to higher dimensional color space for improved characterization from environmental noise. The resulting time-frequency representation is referred as pseudo-color cochleagram image and the resulting feature, which captures the statistical distribution, as pseudo-color cochleagram image feature (PC-CIF). In addition, sequential backward feature selection is applied for selecting the most useful feature dimensions, thereby reducing the feature dimension and improving the classification performance. We evaluate the effectiveness of the proposed methods using two classifiers, *k*-nearest neighbor and support vector machines. The performance is evaluated on a dataset containing 50 sound classes, taken from the Real World Computing Partnership Sound Scene Database in Real Acoustical Environments, with the addition of environmental noise at various signal-to-noise ratios. The experimental results show that the proposed techniques give significant improvement in classification performance over baseline methods. The most improved results were observed at low signal-to-noise ratios.

Keywords: Acoustic event recognition, cochleagram, pseudo-color, sequential backward feature selection, support vector machines, time-frequency image

1.0 Introduction

Research in acoustic event recognition (AER), also referred as sound event recognition (SER), and its many applications has received a lot of attention in recent years. Some of these applications are audio surveillance [1], hearing improvement devices [2], and urban sound classification [3]. A key challenge in this field has been achieving robust AER, that is, improving the sound recognition rate in the presence of noise. The proposed techniques mostly revolve around finding robust features and/or classifiers. For example, use of wavelet features and one-class support vector machines (1-SVM) are proposed for robust audio surveillance in [1]. The use of time-frequency image derived features have been proposed in [4] and matching pursuit [5] for environmental sound recognition in [6]. More recently, deep learning methods have been applied for the same purpose as seen in [7].

The use of time-frequency image derived features, in particular, has been shown to be effective in achieving robust AER [4, 8, 9]. Every sound signal produces a unique texture which can be visualized using a time-frequency image. The addition of noise affects certain frequency bands and features which can accurately capture the non-noise manipulated frequency bands can help in differentiating sounds.

In [4], the sound signal spectrograms are divided into blocks and second and third central

moments are extracted as features in each block, referred as the spectrogram image feature (SIF). While extracting features from the grayscale spectrogram is a common technique, in [4], the grayscale spectrogram image is also quantized and mapped to a higher dimensional color map. The classification performance using features extracted from this pseudo-color quantized spectrograms were shown to be significantly more robust than the SIF and mel-frequency cepstral coefficients (MFCC), a commonly used feature in audio classification applications.

Spectrograms, however, have a disadvantage. The frequency components are equally distributed and have constant bandwidth in this conventional time-frequency distribution. Most sound signals generally have dominant frequency components in the lower frequency range and less frequency components in the upper frequency range. As such, the spectral information is not best revealed in this time-frequency representation.

In [10], the use cochleagrams are proposed over spectrograms for sound classification in an audio surveillance application. Cochleagrams utilize a gammatone filter which models the frequency selectivity property of the human cochlea. The resulting time-frequency image offers more frequency components in the lower frequency range with narrow bandwidth and less frequency components in the upper frequency range with wide bandwidth, thereby revealing more spectral information than the spectrogram image. The corresponding feature, referred as the cochleagram image feature (CIF), was shown to significantly outperform the SIF.

In this work, we extend the color mapping technique proposed for spectrograms in [4] to cochleagrams. That is, we propose the use of pseudo-color quantized cochleagrams for feature extraction over grayscale cochleagrams used in [10]. We follow the same feature extraction technique and refer this as the pseudo-color cochleagram image feature (PC-CIF).

The use of robust features and classifiers has been applied successfully in other similar applications, automatic speech recognition (ASR), in particular. However, not all feature dimensions are useful. Surprisingly, very little attention has been given to dimensionality reduction in AER applications. Most of the current feature extraction and classification methods employed in AER are inspired from ASR where feature dimension reduction has received significant attention.

The benefits of feature dimension reduction are manifold. Firstly, not all feature dimensions are relevant or useful and removing these often leads to improved classification performance. Secondly, feature dimension is one of the factors that affects computation time. A reduced feature dimension often leads to reduced computational costs. In addition, a reduced feature dimension also requires less storage space.

Feature dimension reduction techniques can be grouped into feature selection or feature extraction [11]. Feature selection algorithms can be put into three main categories [12]: wrappers, filters, and embedded. In the wrapper method, the aim is to search for a good subset of features using the feature subset selection algorithm where different combinations are evaluated based on the model accuracy using a predictive model [13]. Filter methods score each feature through application of some statistical measure. Features are then either selected or removed based on its score ranking. Embedded methods perform feature selection and classification simultaneously by using the learning algorithm to determine the most effective features during model creation.

Wrapper methods are directed at improving the classification performance and, therefore, are quite popular in various applications. One commonly used wrapper method is sequential feature selection (SFS). Two commonly used SFS methods are sequential forward feature

selection (SFFS) and sequential backward feature selection (SBFS), also referred as forward selection and backward elimination, respectively. The SFFS/SBFS algorithms start with an empty/full feature set and add/remove features until the maximum objective function is achieved [14]. Some recent applications of SFS, sometimes with modifications, include gene selection [15], identifying fish vocalizations [16], face recognition [17], and emotion detection using speech [18].

In this paper, we study the usefulness of feature dimension reduction in AER. In applications such as gene expression data analysis, the feature vector dimension can be extremely large, multiple thousands of dimensions as seen in [19]. A key reason for feature dimension reduction in such work is to reduce the computational costs. However, the feature dimension in most AER applications, such as [4, 8] and in this work as well, is a few hundred at most. As such, the primary aim of feature dimension reduction in our work is to achieve improvement in classification performance through removal of irrelevant feature dimensions. Also, our primary focus is studying the effect of feature dimension reduction on the classification performance in the presence of noise, that is, the robustness in AER. As such, we consider sequential feature selection for this purpose.

The classification performance of SFFS and SBFS has been compared in many works. In [20, 21], SBFS is determined to be better than SFFS. However, the results in [22] show that SBFS does not always outperform SFFS, SBFS is seen to be more appropriate when the number of features is large enough in [23], and no one method is seen to dominate in [24]. In general, the performance of a feature selection algorithm could be affected by a number of factors such as number of features, number of observations, application, objective function, etc. In this work, we performed some preliminary experiments under the same conditions to compare the classification performance of SFFS and SBFS. On occasions, we found the SFFS method to converge early and the SBFS to be more consistent. To keep the paper concise, we report results using the SBFS method only and refer interested readers to [20-24].

We study the usefulness of SBFS on individual features, MFCC, CIF, and PC-CIF, and combined feature sets, MFCC+CIF and MFCC+PC-CIF. In all cases, we compare the classification performance of raw features and the effect of SBFS under clean conditions and in the presence of noise at various signal-to-noise ratios (SNRs). Similar to [25], the performance is evaluated using two classifiers: *k*-nearest neighbor (KNN) and support vector machines (SVM). KNN is probably the simplest of all classifiers which has shown to be useful with linear cepstral and linear time-frequency image features, which produced the best results in [25]. SVM is a relatively new classification method which has gained widespread attention in AER applications, such as [16, 26], and shown to be especially useful on small datasets. Literature review on SVM, the multiclass classification methods for this binary classifier, and a comparison on its classification performance can be found in [25]. In addition, to further study the effect of feature dimension reduction, we evaluate the training and testing time for the proposed features with the KNN and SVM classifiers with and without feature dimension reduction.

The rest of this paper is organized as follows. Section 2 gives an overview of the baseline and proposed methods, which include grayscale cochleagram image color mapping and sequential backward feature selection methods. The experimental setup, experimental results, and related discussions are given in section 3 and conclusion and recommendations are given in section 4.

2.0 Overview of Feature Extraction, Classification, and Selection

2.1 Baseline Features

The baseline features for use in this work are linear MFCCs and CIF, the details for which can be found in [8]. While conventional cepstral coefficients apply log compression to the filter bank energies before computing DCT, linear cepstral coefficients, without any compression, were determined to be more noise robust and have a better overall classification performance for MFCCs in [8]. Therefore, only results using linear compression will be presented here.

The use of time-frequency image derived features for sound classification has been seen in a number of literature [4, 7, 8]. In [8], two types of time-frequency images were considered for feature extraction: spectrogram and cochleagram. Cochleagram image derived features were determined to give a better overall classification performance and also much more noise robust, therefore, only cochleagram image derived features are considered in this work.

The cochleagram image is divided into subbands and second and third central moments are extracted as features in each subband. The features from each subband are concatenated and the final feature vector referred as the cochleagram image feature (CIF).

2.2 Pseudo-Color Quantized Cochleagram Image

The procedure for pseudo-color quantization of the grayscale cochleagram images is same as for the grayscale spectrogram images in [4]. The grayscale cochleagram is quantized and then mapped onto the red, green, and blue (RGB) monochrome components. The mapping of the grayscale image to the monochrome image can be given as

$$m_c(k,t) = f_c\left(X(k,t)\right) \qquad \forall c \in (c_1, c_2, \ldots c_N) \tag{1}$$

where m_c is a monochrome image (*R*, *G*, or *B*), f a nonlinear mapping function, c the quantization regions, and $X(k,t)$ is the k^{th} harmonic in the t^{th} frame.

A colormap is essentially a color lookup table which in this work is used for mapping the grayscale intensity values. In this work we consider three commonly used colormaps: HSV, Jet, and Hot. The resulting pseudo-color mapped cochlegram images, with and without noise, and the corresponding colormaps for a sample sound signal are given in Figure 1. The construction sound signal used in this illustration is sanding of a piece of wood. For this illustration, the sound signal was divided into frames of size 1024 points with 50% overlap between frames. A total of 512 gammatone filters [8, 27] are used to reveal the frequency characteristics of the sound signal.

The procedure for computing PC-CIF is similar to CIF except that computation is now performed in all three monochrome images and the features concatenated. The advantage of the grayscale cochleagram over the grayscale spectrogram is the better spectral energy distribution as documented in [8]. In addition, the usefulness of time-frequency image feature extraction in achieving robust AER has also been explained in [4, 8, 25]. Here, we focus on the usefulness of the nonlinear higher dimensional color mapping of the grayscale cochleagram which is illustrated using the corresponding spectral energy distribution in Figure 2 for the HSV color space.

With the grayscale cochleagram, the first 200 frequency bins are significantly affected by noise. The noise has minimal or no effect from bin 250 onwards which could help in recognizing the sound signal. When mapped to the HSV colorspace, the noise affects bins up to about 300 for the Green color component. However, there is significantly less corruption for the Red and Blue color components. With the Blue color component, only two frequency bands are affected by noise with narrow bandwidth. The effect of noise on the Red component is between bins 1 and 200 but the noise effect is less than the grayscale image. As

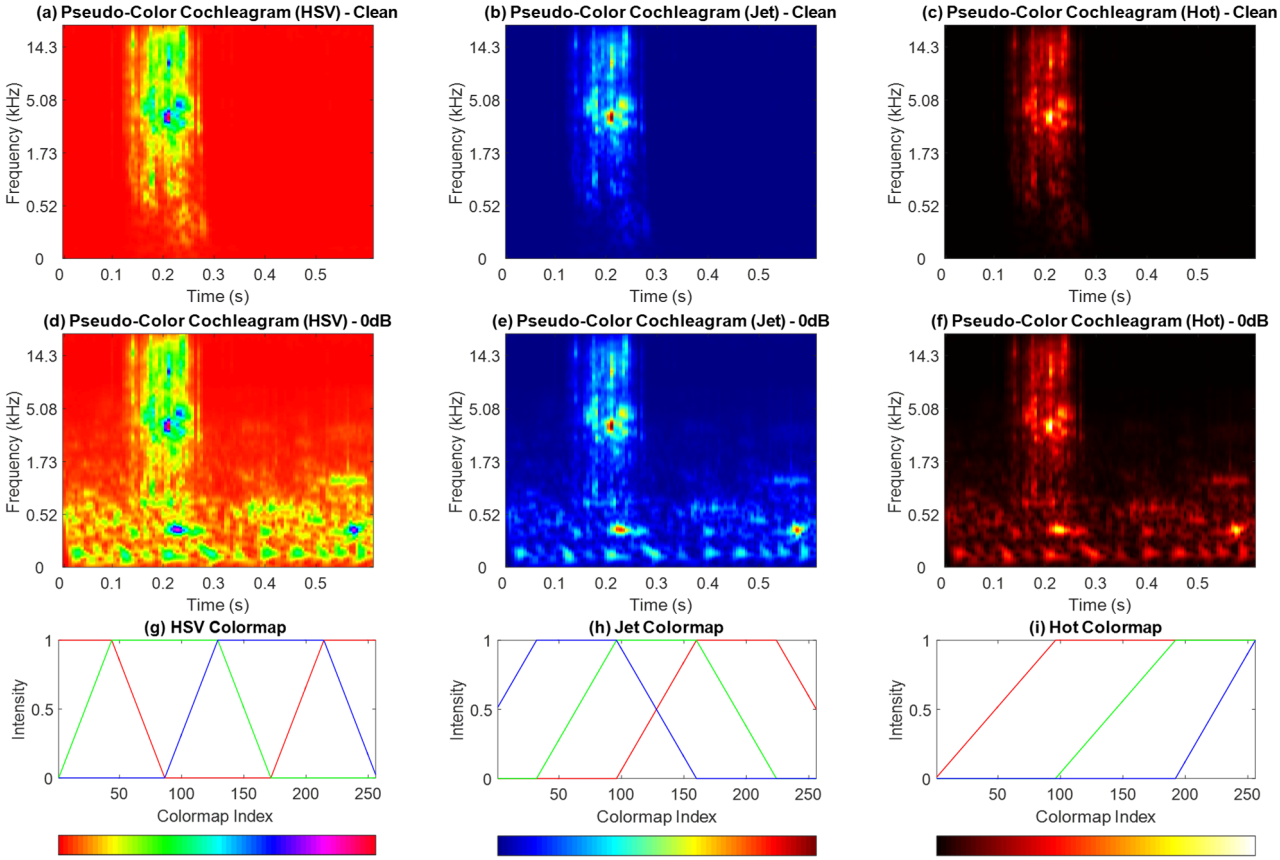

Figure 1: Pseudo-color cochleagram illustrations for a construction (sanding a piece of wood) sound signal along with the HSV, Jet, and Hot colormaps. (a) HSV pseudo-color cochleagram under clean conditions, (b) Jet pseudo-color cochleagram under clean conditions, (c) Hot pseudo-color cochleagram under clean conditions, (d) HSV pseudo-color cochleagram at 0dB SNR, (e) Jet pseudo-color cochleagram at 0dB SNR, (f) Hot pseudo-color cochleagram at 0dB SNR, (g) HSV colormap, (h) Jet colormap, and (i) Hot colormap.

such, in this case, the Red and Blue component images should be more useful for feature extraction then the grayscale image.

2.3 Sequential Backward Feature Selection

The aim of SBFS is to select a subset of features that maximize an objective function [28], the classification accuracy in this case. That is, given a feature set $F = \{f_v | v = 1, \dots, D\}$, search for a subset F_S, with $S < D$, that maximizes an objective function $J(F)$

$$F_S = \{x_{v1}, x_{v2}, \dots, x_{vS}\} = \arg\max_{S, v_S} J\{x_v \mid v = 1, \dots, D\}. \quad (2)$$

The following steps are followed in sequential backward feature selection.

Step 1: Use all the feature dimensions $T_{k=0} = F$, to calculate the average accuracy $J(T_0)$ using cross-validation, where T_k and J represent the feature set and average accuracy, respectively.

Step 2: One feature dimension is removed at a time and the average accuracy is calculated with the remaining features. At the end of this step, the feature removal of feature dimension f^- corresponding to the highest average accuracy,

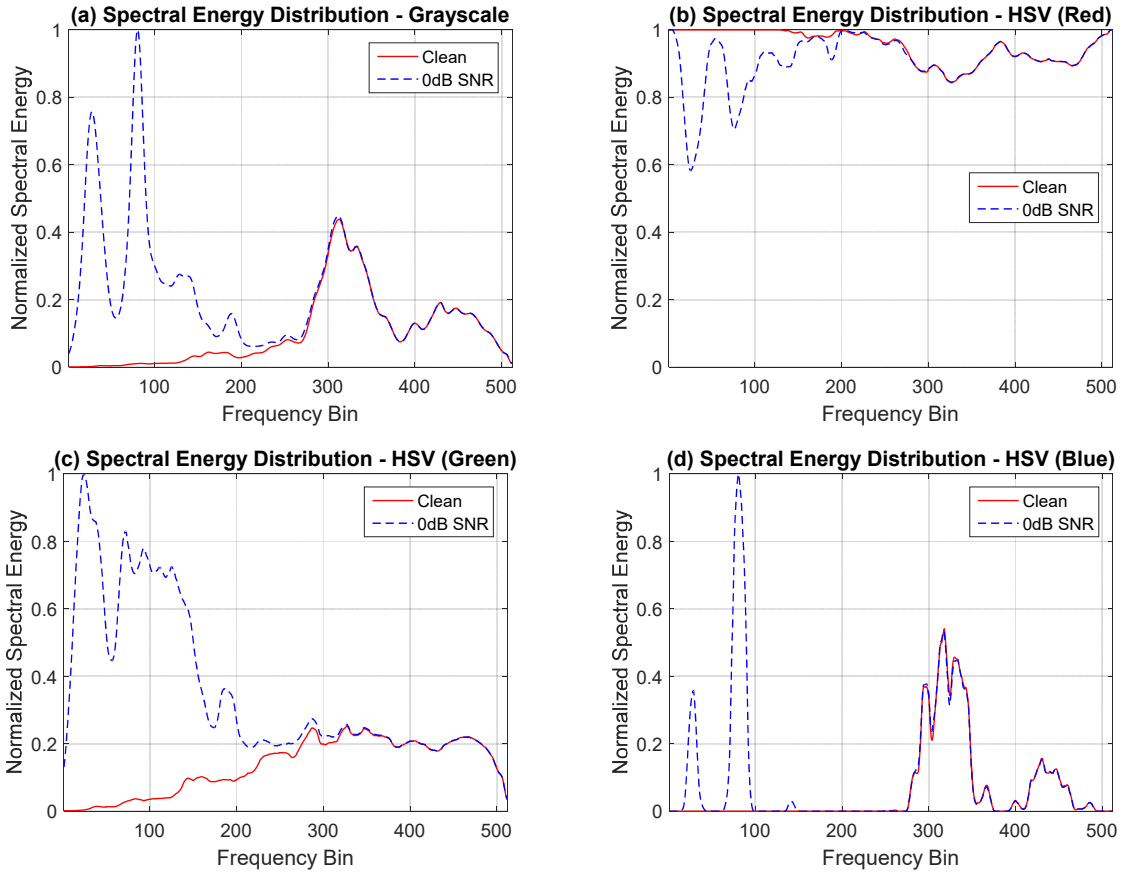

Figure 2: (a) Grayscale cocheagram spectral energy distribution, (b) HSV pseudo-color cochleagram energy distribution for the Red component, (c) HSV pseudo-color cochleagram energy distribution for the Green component, and (d) HSV pseudo-color cochleagram energy distribution for the Blue component.

$$f^{-} = \arg\max_{x \in T_k} \left[J\left(T_k - f\right) \right], \tag{3}$$

is permanently removed and the feature set is updated to

$$T_{k+1} = T_k - f^{-} \tag{4}$$

Step 3: Repeat step 2 with the remaining features. This process continues until no further improvement can be achieved in the classification performance.

Step 4: The SBFS process terminates when the classification accuracy plateaus and the remaining features are utilized in training and testing the final models.

3.0 Experimental Evaluation

An overview of the experimental setup is given first followed by the classification performance using the baseline features (MFCC, CIF, and MFCC+CIF), baseline features after SBFS, and then using the proposed features: PC-CIF and MFCC+PC-CIF. Finally, the training and testing time of the KNN and SVM classifiers before and after feature dimension reduction are compared.

3.1 Experimental Setup

The sound database has a total of 4000 files belonging to 50 classes, 80 files per class. The sound files are taken from the Real World Computing Partnership (RWCP) Sound Scene

Database (SSD) in Real Acoustical Environments [29] which has been used in other similar works such as [4, 7]. This is a large database and, therefore, the sound classes and files used in this work may or may not be similar to what is used in other works. Also, in [4, 7], the performance is evaluated under four noise levels but we use five noise levels. As such, direct comparison of results is difficult. All signals in the database have 16-bit resolution and a sampling frequency of 44100 Hz.

The classification performance is evaluated under three different noise environments taken from the NOISEX-92 database [30]: *speech babble*, *factory floor 1*, and *destroyer control room*. The performance is evaluated in clean conditions and at 20dB, 10dB, 5dB, and 0dB SNRs.

For all experiments, signal processing is carried out using a Hamming window of 1024 points (23.22 ms) with 50% overlap. For the CIF and PC-CIF, which utilize the gammatone filter, only results using the best performing ERB model [10] are reported. The classification accuracy is given in percentage as *number of correctly classified test samples* divided by the *total number of test samples*. For the KNN classifier, k values from 1 to 30 were experimented with for each feature set but only the best results are presented here. For the SVM classifier, nonlinear SVM with a Gaussian RBF kernel is used in all cases as it was found to give the best results. SVM parameters, the penalty parameter and the width of the Gaussian function [31], were tuned using Bayesian optimization [32]. In tuning the parameters, one set of parameters which gave the best average classification accuracy was selected. For all experimentations, the classifier is trained with 50 clean samples per class with the remaining 30 samples used for validating and testing the model under clean and noisy conditions. As such, a total of 2500 samples are used for training the classifiers and the remaining 1500 samples used for validating and testing the models.

3.2 Results Using Baseline Features

The classification accuracy values for MFCC, CIF, and MFCC+CIF with raw features using KNN and SVM classifiers are given in Table I.

In computing MFCC, the number of mel-filters was varied in the range 10-60 to determine the optimal number of mel-filters for each classifier. The variation of the average classification accuracy for KNN and SVM classifiers against the number of mel-filters is shown in Figure 3. In general, the best average classification accuracy for both classifiers was achieved when the number of mel-filters was in the range 13-35. For KNN, the highest average classification accuracy was achieved at $M = 22$ and $M = 20$ for SVM.

For MFCC, the feature vector for each frame in the sound signal is $3 \times M$ dimensional which includes the M cepstral coefficients and the first and second derivatives. The final feature vector is a concatenation of the mean and standard deviation values for each dimension. As such, the final MFCC feature vector dimension is $2 \times 3 \times M$ resulting in 132 dimension feature vector for KNN and 120 dimension feature vector for SVM.

For the CIF, various number of subbands were experimented with. The optimal number of subbands was chosen as 64 for both KNN and SVM classification methods as it was determined to give the best tradeoff between classification performance and feature dimension. The number of gammatone filters was set to 512 which means there are 8 frequency bins per subband. The second and third central moments are computed in each subband and concatenated to form the final feature vector which is 128 dimensional. Finally, for MFCC+CIF, the feature vector is 260 dimensional for KNN classification and 248 dimensional for SVM classification.

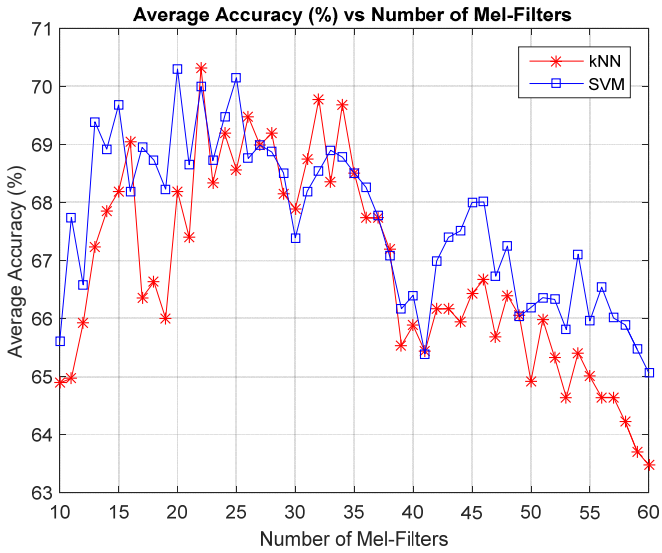

Figure 3: Variation in the average classification accuracy with increasing number of mel-filters

Table I: Classification accuracy values for MFCC, CIF, and MFCC+CIF using KNN and SVM classification with raw features

Feature	KNN						SVM					
	Clean	20dB	10dB	5dB	0dB	Average	Clean	20dB	10dB	5dB	0dB	Average
MFCC	82.07	81.76	78.82	69.04	39.91	70.32	85.53	85.02	78.73	65.11	37.09	70.30
CIF	85.20	84.40	81.53	73.73	53.40	75.65	91.13	90.18	85.16	75.49	52.49	78.89
MFCC+CIF	86.33	85.69	83.02	76.58	53.80	77.08	92.33	92.07	88.42	79.69	53.89	81.28

The average classification accuracy of the CIF is determined to be significantly better than MFCC with both KNN and SVM classification methods, +5.33% and +8.59%, respectively. The CIF is seen to outperform MFCC under all noise conditions with the most improved results at low SNR. At 0dB SNR, the improvement from MFCC to CIF is +13.49% and +15.40% with KNN and SVM classification methods, respectively. As such, it can be deduced that the CIF is significantly more noise robust than MFCC.

Further improvement in the average classification performance is observed with the combined feature set, MFCC+CIF. The improvement in the average classification accuracy over CIF, the best performing individual feature, is +1.43% and +2.39% for KNN and SVM classification methods, respectively. The classification accuracy is seen to improve under all noise conditions with both the classifiers.

In addition, in general, the average classification performance of the SVM classifier is seen to be marginally better than the SVM classifier. With MFCC, the average classification accuracy value using the KNN and SVM classifiers are almost same while the SVM classifier performs better with CIF and MFCC+CIF, +3.24% and +4.20%, respectively. The SVM classifier is particularly seen to perform well under clean and high SNR conditions. More such analysis on feature and classifier performance can be found in [8, 25].

3.3 Results for Baseline Features after SBFS

The classification accuracy results after applying SBFS are given in Table II. The results with

KNN classification show that there is significant improvement in average classification accuracy for all three feature sets over the results using raw features. The improvement in average classification value over the corresponding raw feature results are 10.99%, 5.12%, and 6.55% for MFCC, CIF, and MFCC+CIF, respectively. The classification accuracy values are observed to improve under all noise conditions for all three feature sets but the most improved results are at low SNR. At an average classification accuracy of 83.63%, the combined feature set, MFCC+CIF, is once again seen to give the best classification performance. The improvement in classification accuracy for MFCC+CIF is 3.47%, 3.58%, 4.22%, 6.53%, and 14.91% under clean conditions and at 20dB, 10dB, 5dB, and 0dB SNR, respectively.

A similar trend can also be observed with SVM classification. The improvement in average classification accuracy over the raw features is 9.76%, 5.87%, and 6.21% for MFCC, CIF, and MFCC+CIF, respectively. The average classification accuracy improves under all noise conditions with the most improved results at low SNR. The combined feature set, MFCC+CIF, once again outperforms the individual features. The improvement in classification accuracy value for MFCC+CIF are 0.67%, 0.33%, 1.98%, 7.47%, and 20.60% under clean conditions and at 20dB, 10dB, 5dB, and 0dB SNR, respectively.

In addition, the comparison of average classification accuracy values using KNN and SVM classification methods are observed to be similar to raw features. KNN performs slightly better, +1.25%, with MFCC while SVM performs marginally better, +3.99% and +3.86%, with CIF and MFCC+CIF, respectively. With MFCC, SVM is seen to perform slightly better under clean and high SNR conditions and KNN performing better at low SNR. For CIF and MFCC+CIF, however, the classification accuracy values using SVM classification are better than KNN classification under all noise conditions.

Table II: Classification accuracy values for MFCC, CIF, and MFCC+CIF using KNN and SVM classification after SBFS

Feature	KNN						SVM					
	Clean	20dB	10dB	5dB	0dB	Average	Clean	20dB	10dB	5dB	0dB	Average
MFCC	84.93	84.84	83.27	80.96	72.56	81.31	85.87	85.93	83.98	79.56	64.98	80.06
CIF	88.13	87.16	85.18	80.29	63.09	80.77	91.80	90.98	88.62	83.93	68.44	84.76
MFCC+CIF	89.80	89.27	87.24	83.11	68.71	83.63	93.00	92.40	90.40	87.16	74.49	87.49

The SBFS algorithm is targeted at improving the classification performance. It continues to remove feature dimensions until no further improvement can be achieved in the classification performance. This explains the usefulness of the SBFS algorithm which also makes it an exhaustive and time consuming process, a disadvantage if the raw feature dimension is large.

3.4 Results Using PC-CIF and SBFS

With the PC-CIF the final feature vector is three times CIF, that is, 384 dimensional. The classification accuracy results using PC-CIF and MFCC+PC-CIF are given in Table III. The presented results are after applying SBFS since this has already shown to be significantly more robust than the raw features. While experimentation was performed using all three colormaps, only the best results are presented here. The best results were achieved using the HSV and Hot colormaps for KNN and SVM classification methods, respectively.

For PC-CIF, an average classification accuracy of 84.96% and 88.80% are achieved using KNN and SVM classification methods, respectively. This is marginally better than the

average classification accuracy of 83.63% and 87.49% achieved using MFCC+CIF using KNN and SVM classification methods, respectively. As such, the PC-CIF on its own is able to match the classification performance of MFCC+CIF. In addition, it is significantly better than the average classification accuracy of 80.77% and 84.76% using CIF with KNN and SVM classifiers, respectively.

With the combined feature vector, the improvement over PC-CIF is –0.92% and +1.32% using KNN and SVM classifiers, respectively. The performance of KNN classifier drops slightly with feature combination indicating its unsuitability with high dimensional feature combination. With an average classification accuracy of 90.12%, the feature vector of MFCC+PC-CIF and SVM classification produces the best classification performance. Also, MFCC+PC-CIF produces improvement in classification accuracy over PC-CIF under all noise conditions with SVM classification.

Table III: Classification accuracy values for PC-CIF and MFCC+PC-CIF using KNN and SVM classification after SBFS

Feature	KNN						SVM					
	Clean	20dB	10dB	5dB	0dB	Average	Clean	20dB	10dB	5dB	0dB	Average
PC-CIF	92.93	91.49	87.24	82.69	70.44	84.96	95.07	94.44	92.49	87.78	74.22	88.80
MFCC+PC-CIF	91.73	90.27	86.04	81.91	70.27	84.04	95.33	94.78	93.38	89.76	77.33	90.12

3.5 Training and Testing Times

The normalized training and testing time of the KNN and SVM classifiers before and after feature dimension reduction using SBFS are shown in Figure 4. For KNN classifier, training and testing time reduce for all feature sets after applying SBFS. With the best performing feature set of MFCC+PC-CIF, the training and testing time are about 99% and 92% of the original training and testing time, respectively. A similar trend is also observed with SVM classification. For MFCC+PC-CIF, both the training and testing time are about 91% of the original training and testing time, respectively.

While the feature dimension of MFCC is significantly less than PC-CIF and MFCC+PC-CIF, the training time is observed to be the highest using SVM classification both before and after feature dimension reduction. Also, the training time is observed to be slightly higher after feature dimension reduction. To further investigate the reason for this, an experiment was performed to measure the average number of optimization iterations and the average number of support vectors for the three feature sets. The results for this are shown in Figure 5.

The average number of optimization iterations for MFCC is observed to be significantly higher than PC-CIF and MFCC+PC-CIF which could explain the increase in training time. Also, both the average number of optimization iterations and the average number of support vectors are observed to increase after feature dimension reduction thereby increasing the training time.

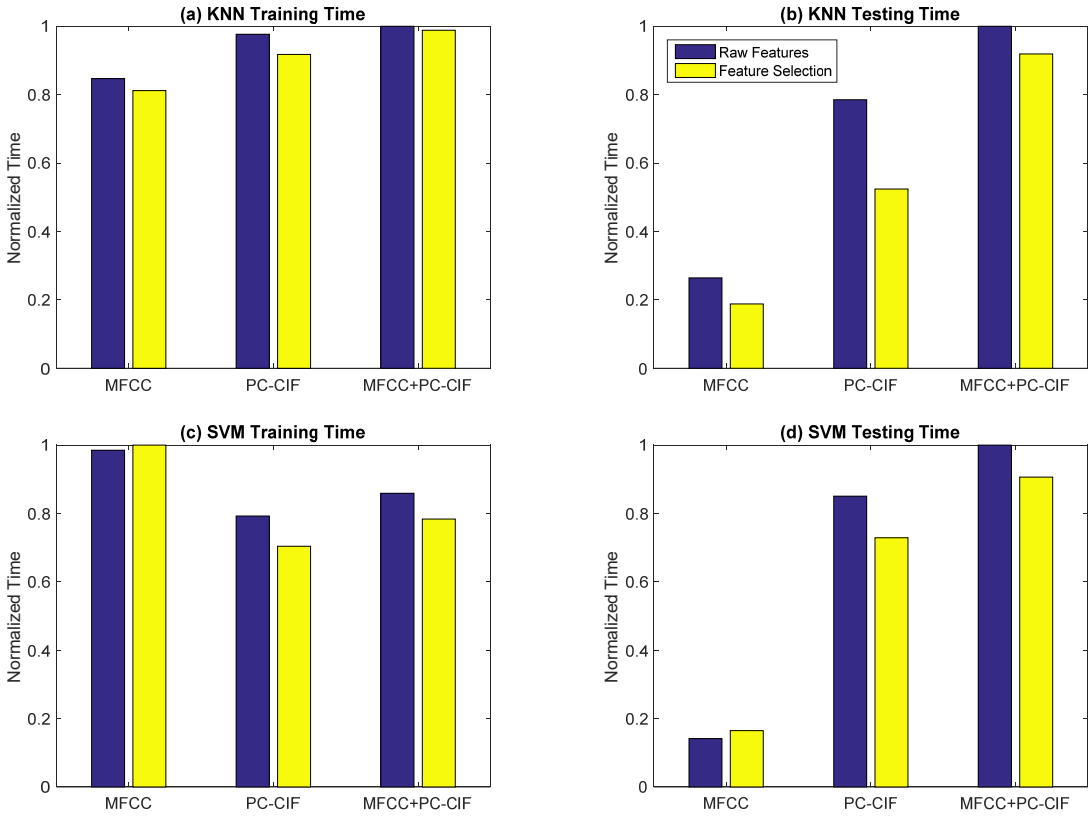

Figure 4: Comparison of training and testing time for raw features and after feature selection using KNN and SVM classification. (a) KNN training time, (b) KNN testing time, (c) SVM training time, and (d) SVM testing time.

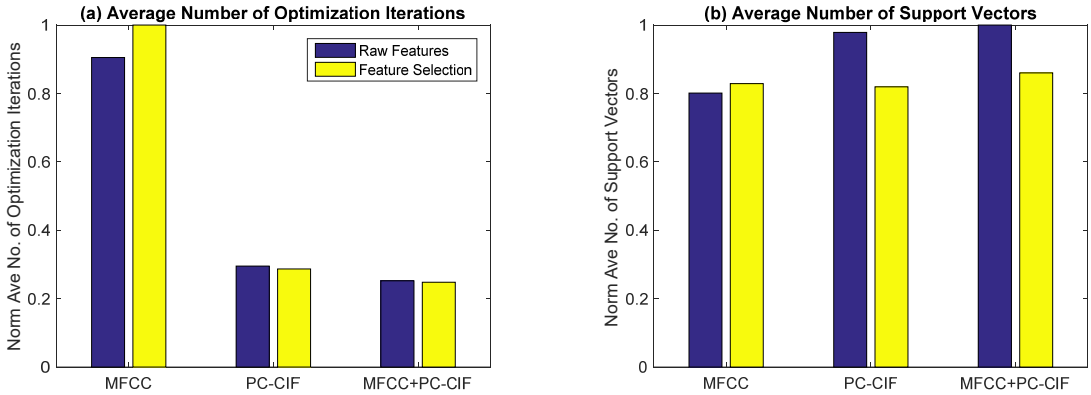

Figure 5: (a) Average number of optimization iterations and (b) average number of support vectors for the SVM classifier with raw features and after SBFS.

4.0 Conclusion

This work proposes the use of pseudo-color quantized cochleagram images for feature extraction and sequential backward feature selection for robust sound event recognition. The classification accuracy values using the proposed feature, PC-CIF, were seen to increase under both clean and noisy conditions with both KNN and SVM classification methods over the baseline features. Further improvement in classification performance was achieved with feature combination, MFCC+PC-CIF. The improvement in classification accuracy was seen

to increase from clean and high SNR conditions to low SNR conditions. As such, the most improved results were obtained at low SNR conditions.

SBFS is, however, a greedy algorithm making it very time consuming, especially if the input or raw feature dimension is high. This wasn't a significant disadvantage in this work since feature selection was supervised and the input feature vectors a few hundred at most. However, in future, effort to reduce this time will be considered. One way to achieve this would be using hybrid feature dimension reduction such as using PCA or another feature dimension reduction technique to select the subset of features from the original feature set and then applying SBFS to this subset features. It would also be important to test the performance of the developed models on an independent test dataset to ensure there is no feature selection bias and to get a true measure of the classification performance. At the moment feature selection is performed on the validation dataset with the assumption that any given test data would be statistically similar.

References

[1] A. Rabaoui, M. Davy, S. Rossignol, N. Ellouze, Using one-class SVMs and wavelets for audio surveillance, IEEE Transactions on Information Forensics and Security, 3 (4) (2008) 763-775.

[2] F. Saki, N. Kehtarnavaz, Real-time hierarchical classification of sound signals for hearing improvement devices, Applied Acoustics, 132 (2018) 26-32.

[3] J. Ye, T. Kobayashi, M. Murakawa, Urban sound event classification based on local and global features aggregation, Applied Acoustics, 117 (Part B) (2017) 246-256.

[4] J. Dennis, H.D. Tran, H. Li, Spectrogram image feature for sound event classification in mismatched conditions, IEEE Signal Processing Letters, 18 (2) (2011) 130-133.

[5] S.G. Mallat, Z. Zhang, Matching pursuits with time-frequency dictionaries, IEEE Transactions on Signal Processing, 41 (12) (1993) 3397-3415.

[6] S. Chu, S. Narayanan, C.C.J. Kuo, Environmental sound recognition with time-frequency audio features, IEEE Transactions on Audio, Speech, and Language Processing, 17 (6) (2009) 1142-1158.

[7] I. McLoughlin, H. Zhang, Z. Xie, Y. Song, W. Xiao, Robust sound event classification using deep neural networks, IEEE/ACM Transactions on Audio, Speech, and Language Processing, 23 (3) (2015) 540-552.

[8] R.V. Sharan, T.J. Moir, Subband time-frequency image texture features for robust audio surveillance, IEEE Transactions on Information Forensics and Security, 10 (12) (2015) 2605-2615.

[9] S. Wang, X. Zeng, Robust underwater noise targets classification using auditory inspired time–frequency analysis, Applied Acoustics, 78 (2014) 68-76.

[10] R.V. Sharan, T.J. Moir, Cochleagram image feature for improved robustness in sound recognition, in: Proceedings of the IEEE International Conference on Digital Signal Processing (DSP), Singapore, 2015, pp. 441-444.

[11] R. Gutierrez-Osuna, Pattern analysis for machine olfaction: A review, IEEE Sensors Journal, 2 (3) (2002) 189-202.

[12] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, Journal of Machine Learning Research, 3 (2003) 1157-1182.

[13] R. Kohavi, G.H. John, Wrappers for feature subset selection, Artificial Intelligence, 97 (1-2) (1997) 273-324.

[14] G. Chandrashekar, F. Sahin, A survey on feature selection methods, Computers & Electrical Engineering, 40 (1) (2014) 16-28.

[15] B. Liao, Y. Jiang, W. Liang, W. Zhu, L. Cai, Z. Cao, Gene selection using locality sensitive laplacian score, IEEE/ACM Transactions on Computational Biology and Bioinformatics, 11 (6) (2014) 1146-1156.
[16] F. Sattar, S. Cullis-Suzuki, F. Jin, Identification of fish vocalizations from ocean acoustic data, Applied Acoustics, 110 (2016) 248-255.
[17] N. Gu, M. Fan, L. Du, D. Ren, Efficient sequential feature selection based on adaptive eigenspace model, Neurocomputing, 161 (2015) 199-209.
[18] N. Semwal, A. Kumar, S. Narayanan, Automatic speech emotion detection system using multi-domain acoustic feature selection and classification models, in: IEEE International Conference on Identity, Security and Behavior Analysis (ISBA), New Delhi, 2017, pp. 1-6.
[19] S. Chao, C. Lihui, High dimensional gene expression data dimension reduction, in: IEEE Conference on Cybernetics and Intelligent Systems, 2004, pp. 451-455.
[20] J. Doak, An evaluation of feature selection methods and their application to computer security, University of California, Computer Science, 1992.
[21] C.M. Wang, S.C. Yang, P.C. Chung, Comparative evaluation of classifiers and feature selection methods for mass screening in digitized mammograms, in: IEEE/NLM Life Science Systems and Applications Workshop, 2006, pp. 1-2.
[22] D.W. Aha, R.L. Bankert, A comparative evaluation of sequential feature selection algorithms, in: D. Fisher, H.-J. Lenz (Eds.) Learning from Data: Artificial Intelligence and Statistics V, Springer New York, New York, 1996, pp. 199-206.
[23] Y.H. Chan, W.W.Y. Ng, D.S. Yeung, P.P.K. Chan, Empirical comparison of forward and backward search strategies in L-GEM based feature selection with RBFNN, in: International Conference on Machine Learning and Cybernetics, 2010, pp. 1524-1527.
[24] H. Hongwei, L. Cheng-Lin, H. Sako, Comparison of genetic algorithm and sequential search methods for classifier subset selection, in: Seventh International Conference on Document Analysis and Recognition, 2003, pp. 765-769.
[25] R.V. Sharan, T.J. Moir, Noise robust audio surveillance using reduced spectrogram image feature and one-against-all SVM, Neurocomputing, 158 (2015) 90-99.
[26] K. Lopatka, A. Czyzewski, Acceleration of decision making in sound event recognition employing supercomputing cluster, Information Sciences, 285 (2014) 223-236.
[27] M. Slaney, Auditory Toolbox for Matlab, Interval Research Corproation, Technical Report 1998-010, 1998.
[28] R. Gutierrez-Osuna, Lecture 11: Sequential feature selection, Texas A&M University, Lecture Notes.
[29] S. Nakamura, K. Hiyane, F. Asano, T. Nishiura, T. Yamada, Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition, in: Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC 2000), Athens, Greece, 2000, pp. 965–968.
[30] A. Varga, H.J.M. Steeneken, Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems, Speech Communication, 12 (3) (1993) 247-251.
[31] C. Cortes, V. Vapnik, Support-vector networks, Machine Learning, 20 (3) (1995) 273-297.
[32] J. Snoek, H. Larochelle, R.P. Adams, Practical bayesian optimization of machine learning algorithms, in: Advances in neural information processing systems, 2012, pp. 2951-2959.