Robust Sound Event Recognition using Subband Power Distribution Image Feature

Pooja K J
Department of Telecommunication Engineering
Siddaganga Institute of Technology,
Tumakuru-572103, Karnataka, India

Usha L
Department of Telecommunication Engineering
Siddaganga Institute of Technology,
Tumakuru-572103, Karnataka, India

***Abstract*—This paper proposes a feature extraction and classification technique for the task of sound event recognition (SER) in a severe mismatched noise condition. An SER system that can achieve human like sound recognition performance has wide range of application which includes acoustic surveillance, bio-acoustical monitoring, health care application, environment context detection and machine hearing. The approach used here takes inspiration from both audio and image processing field and is based on transforming a sound into a two dimensional representation and then extracting image feature for classification. Spectrogram image feature is being usually used for this purpose. Here a novel method is used to improve the sound event classification in a severe noise condition which is called a subband power distribution (SPD) Image-a two dimensional representation that characterizes the spectral power over time in each frequency subband. In SPD, the high power reliable elements of spectrogram are transformed to localized region and hence can easily be separated from noise. An image feature is extracted from the SPD and missing feature classification method is developed which selects the signal from SPD which is not affected by noise. This is done based on nearest neighbor classifier (kNN) .The proposed method is tested on a database containing 50 sound classes over a challenging noise condition. The results shows that the SPD-IF is discriminative over the broad range of classes and robust in non stationary noise condition.**

Keywords- Sound event recognition; subband power distribution; Spectrogram; Missing feature theory; kNN classifier.

I. INTRODUCTION

The environment sound is rich in acoustic information beyond the speech signal that is mainly the focus of automatic speech recognition system. While speech is the most informative sound event, this paper focuses on general sound events such as door closing or bell ringing which provide information and context for the environment beyond that contained in speech.

Sound event classification is the task where in the audio content of short sound clip is assigned into one set of pre-trained classes. This has a wide range of important application such as acoustic surveillance [2], environmental sound [3, 4], bioacoustics monitoring [5] or in field of machine hearing. In most of these applications, the sound event occurs in the presence of challenging noise condition and signal to noise ratio (SNR) may even fall to 0dB. Typical systems used are often based on adaptation of common speech recognition system like Mel-Frequency Cepstral Coeffient (MFCC) and Hidden Markov Model (HMM) classifier. The most important difference between sound and speech signals is that sounds have more distinctive time frequency characteristics and therefore classification depends on characterizing the stochastic nature of the signal. To capture this variation the conventional frame based MFCCs has to be combined with complex recognizers like HMM. This requires a large amount of training data to perform well and the performance often degrades in the presence of mismatched noise conditions.

One efficient way to capture the time-frequency characteristics of sound is to extract features from the spectrogram of the sound which is called spectrogram image feature (SIF) [6].This is from the fact that humans can easily identify the signal in an image even in the severe background noise. In SIF, spectrogram is quantized and segmented, similar to pseudocolouring and partitioning in image processing. The feature is extracted from each time frequency block in terms of their central moments. Support Vector Machines (SVM) is used to perform the classification.

Due to the physical nature of most of the sounds, the spectrogram is sparse where the sound energy is concentrated in few of the localized frequency bands compared to diffuse noise which is spread evenly across the frequency spectrum. Hence in noisy condition the low power quantization are most affected. An improvement to this SIF approach therefore includes a missing feature framework to marginalize the regions affected by noise. However, due to the non stationary nature of the noise across time and frequency, developing a reliable missing feature mask is challenging [7]. Another drawback of SIF method is the sensitivity of the block distribution to time shifting .Due to this sound clips need to balanced but in real world situation, time shifting may occur due to variation in the performance of the detector which causes uncertainty to the onset and offset times in the sound clip.

To address these problems, a novel sound event image representation called the subband power distribution (SPD), which is invariant to effects of time shifting, is developed. The SPD captures the distribution of the log-spectral power of sound over time in each frequency subband. This can be visualized as a two dimensional image representation of frequency against normalized spectral power. As done in SIF, feature is extracted from the block wise central moments of the quantized and segmented SPD image. To improve the classification system, a missing feature classifier is developed which automatically selects the blocks that are unaffected by noise in SPD image representation. This is done by utilizing kNN with hellinger distance as a distance metric to measure the distribution distance between two image features.

II. SUBBAND POWER DISTRIBUTION IMAGE FEATURE (SPD-IF)

In this section, the proposed subband power distribution (SPD) image representation for robust sound event classification is presented and the overview of the proposed system is as shown in fig. 1 and is compared to previous method, SIF alongside. The subband power distribution is based on the distribution of the spectral power in each frequency subband over time. It is designed such that the reliable part of the signal is transformed to a continuous region of the SPD representation which makes it easily separable from the noise. As shown in fig 1, after contrast enhancement step to obtain the SPD image, image feature is extracted using the same approach as used for SIF. This process involves quantization and mapping which is analogous to pseudocolourmapping in image processing. After image feature extraction, a robust missing feature classification system is developed. This generates a missing feature mask based on the SPD and then marginalizes the features that are affected by noise. Classification is done using kNN with hellinger distance measure as it is found that it takes into account the distribution information captured in the SPD image feature. Hence it is more efficient way to measure the similarity between two SPD-IF compared to conventional Euclidean distance.

A. Subband power distribution (SPD) image algorithm

The SPD image is designed to represent the distribution of spectral power in each frequency subband over time starting from a time frequency spectrogram representation of the sound hence the algorithm starts from a time-frequency spectrogram representation of the sound, S(f,t).An example of this process is as shown in fig 2. Here probability distribution of log power spectrogram of a whistle sound is calculated for each subband, and then stacked together to form a two dimensional representation of frequency against normalized spectral power as shown in fig 2.b. The SPD then undergoes a contrast enhancement that gives the final SPD image as shown in fig 2.c. Contrast enhancement is performed to ensure that the important signal information is represented over the full[0,1] range of the image. This SPD which is obtained after contrast enhancement forms the basis for image feature extraction and classification. As the SPD captures the temporal distribution statistics, it is desirable for S(f,t) to have a high time resolution to better capture the distribution.

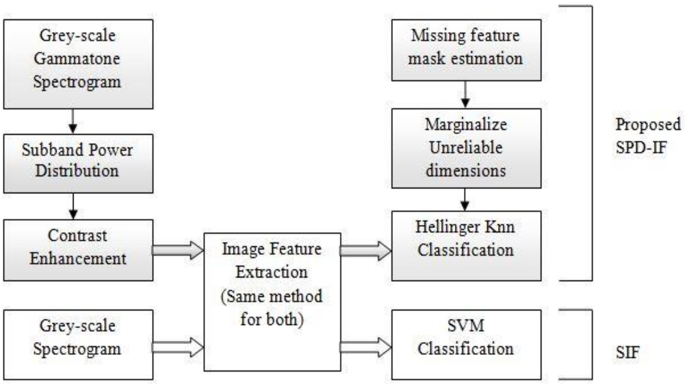

Grey-scale Gammatone Spectrogram
Subband Power Distribution
Contrast Enhancement
Grey-scale Spectrogram
Image Feature Extraction (Same method for both)
Missing feature mask estimation
Marginalize Unreliable dimensions
Hellinger Knn Classification
SVM Classification
Proposed SPD-IF
SIF

Fig. 1.Overview of the proposed SPD-IF approach, comparing to previous work on the SIF (the shaded boxes indicate the contributions in this paper).

Here, the gammatone filterbank decomposition is chosen for time frequency analysis which is derived from the cochlear filtering in the inner ear. This has the advantage that there is no trade off between time and frequency resolution which is common in conventional Short Time Fourier Transform (STFT) representation. Here, a bank of 50 filters is used, with the centre frequencies equally spaced between 100 and 8000Hz on an equivalent rectangular bandwidth (ERB) scale [8].

The SPD is based on the normalized log-power spectrogram G(f,t), given as:

$$G(f,t) = \frac{\log S(f,t)}{\max\limits_{f,t}(\log S(f,t))} \tag{1}$$

Log power is used to compress the dynamic range of the spectrogram to enhance the high power elements in the SPD.The values in G(f,t) that are less than zero i.e, G(f,t) < 0 are all set to zero which normalizes G(f,t) into grey scale image in the range [0,1]. This ensures that the relative volume of different sound clips is equalized and the high power elements are transformed to the same region of the SPD always. The SPD represents the distribution of power in each frequency subband of the normalized spectrogram over time as:

$$D_f(z) = P_{G_f}(z) \tag{2}$$

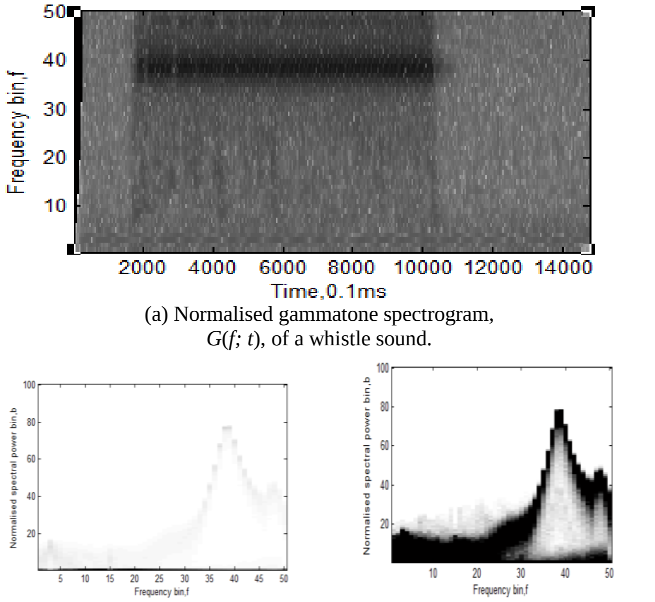
Where z represents the normalized spectral power, P is the probability density function and G_f is a random variable that represents the normalized spectrogram, G(f,t) in the frequency subband, f. The SPD then forms a two dimensional representation, D(f,z) which is obtained by stacking together each subband distribution over frequency, f. As the upper and lower bounds of the distribution are fixed, the distribution D(f,z) is estimated using non parametric approach based on the histogram for its speed and simplicity. Therefore SPD is given by:

$$H_R(f,b) = \frac{1}{t_{max}} \sum_t 1_b(G(f,t)) \tag{3}$$

Where t_{max} is equal to number of time samples in the segment and 1_b is a indicator function which equals to one for the b^{th} bin if the normalized log power spectral power G(f,t) lies within the range of the bin and it is zero otherwise. In this, a total of 100 bins are used with the bin edges equally spaced over the [0.6, 1] range of the normalized spectral power z. The values in $H_R(f,b)$ are raw probability distribution information for each frequency subband over time which are constrained to lie in the range

$$0 \le H_R(f,b) \le 1 \tag{4}$$

Although this implies that the $H_R(f,b)$ is a grey scale image already, it is found that most of the information is constrained only within a small region of dynamic range. This is due to the physical nature of many sound events which have an attenuating or non stationary spectrogram envelope. This means that for a high number of histogram bins, it is unlikely for the subband distribution density values in any one of the bin to be high.



(a) Normalised gammatone spectrogram, *G(f; t)*, of a whistle sound.

(b)Raw SPD, $H_R(f, b)$, formed by stacking the subband distribution information across frequency.

(c)Enhanced SPD Image , H(f,b),after contrast has Enhancement performed

Fig.2. Overview of generation of the SPD Image. The probability distribution is taken over each subband and is stacked to form the raw SPD in (b). This undergoes contrast enhancement, to give the SPD in (c).

Therefore, the contrast of the raw SPD is enhanced to produce the enhanced SPD image that enables the better extraction of important signal information for classification. In image processing this is referred as "contrast stretching" which is performed as follows, where h is an appropriate constant

$$H(f,b) = \begin{cases} H_R(f,b) \times h, & \text{if} \quad H_R(f,b) < \frac{1}{h} \\ 1, & \text{otherwise} \end{cases} \tag{5}$$

This operation does not affect fully stationary subbands as these are still assigned a high value in the enhanced SPD. Hence this step improves the classification over a broad range of sound classes. Empirically, using h=50 provides a sufficient enhancement in contrast.

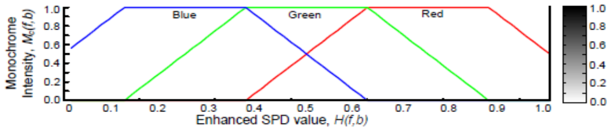
B. SPD image feature extraction

In this step, feature that characterizes the sound information in image is extracted from the two dimensional SPD image. The process starts by quantizing the dynamic range of the grayscale SPD image into different regions as shown in fig 3(a), Each of which maps to a monochrome images. The information in each monochrome image is represented separately in an image feature. This operation is the generalization of the pseudocolourmapping procedure in image processing, where grey scale intensities are quantized into red, green and blue (RGB) monochrome values. This mapping is denoted as,

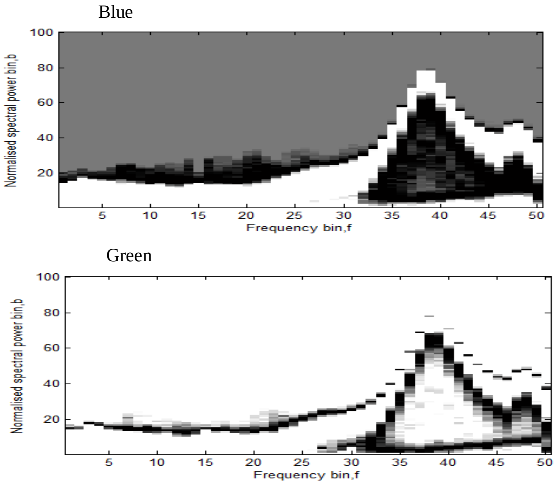
$$M_c(f,b) = h_c(H(f,b)) \qquad \forall c \in (c_1, c_2, ..c_N) \tag{6}$$

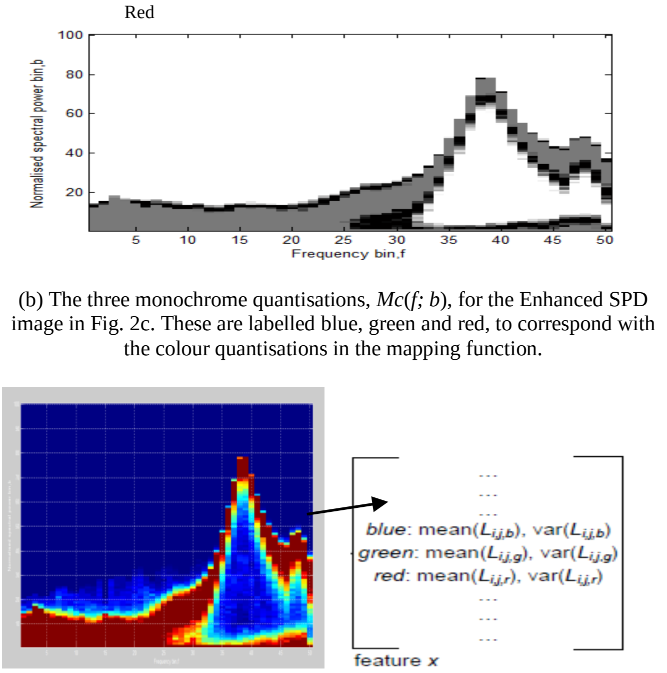
Where h_c is the nonlinear mapping function for mapping dimension c, considering c=3, represents the red, green and blue color dimensions. here a mapping function from image processing ,similar to 'Jet' color map in MATLAB is utilized, as it is found that such existing color maps provide a best suitable quantization of the dynamic range [6]. To characterize the information in each monochrome separately in an image feature, each monochrome image is partitioned into two dimensional local sub-blocks, with each block of size (P/D, Q/D).this gives a total of D^2 blocks as shown in fig 3.c. For clarity, the c notation for each mapping is dropped. Therefore each sub block can be written as a subset of pixel from the entire monochrome as

$$L_{i,j} \subset M(f,b) \tag{7}$$



(a) The quantisation and mapping function, *hc*.





(b) The three monochrome quantisations, *Mc(f; b)*, for the Enhanced SPD image in Fig. 2c. These are labelled blue, green and red, to correspond with the colour quantisations in the mapping function.

(c) Partitioning of the SPD image, and extraction of distribution statistics to generate the image feature, *x*. The SPD is shown here in colour, as a combination of the monochromes above.

Fig. 3: Overview of the image feature extraction, continuing from the whistle sound SPD in Fig. 2c.

Where i,j={1,2,…D} represents the indices of sub blocks and $L_{i,j}$ represents the region of monochrome image ,M(f,b) ,corresponding to the particular sub-block. Next, the image feature $x_{i,j}$ is extracted from each local sub block by utilizing the central moments in every block to capture distribution as:

$$\mu_k = E[(X - E[X])^k] \qquad (8)$$

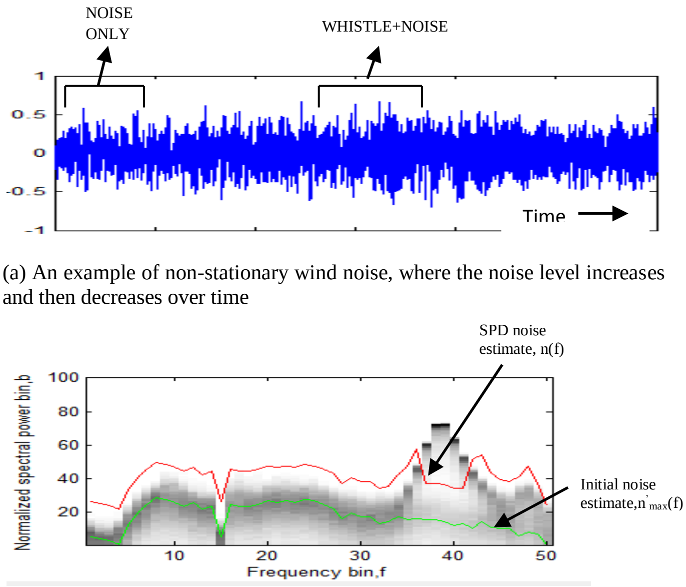
Where X is the distribution, E is the expectation operator and μ_k is the k^{th} moment about the mean. Particularly, the second and third central moments are used in this system. Feature $x_{i,j}$ from each local sub block is therefore:

$$x_{i,j} = \left\{ \mu(L_{i,j}),\ \ \sigma^2(L_{i,j}) \right\} \qquad (9)$$

Where $\mu(.)$ and $\sigma^2(.)$ are mean and variance respectively obtained from second and third central moments found from (8). Experimentally it is found that, partitioning each monochrome image dimension into D=10 blocks give a good tradeoff between feature vector size and performance. The total number of feature for single sound event is therefore becomes 10×10×3×2=600, since there are D^2=100 sub blocks, three monochrome mapping(c=3) and with two central moments.

C. Noise estimation based on Subband Power Distribution

Based on the SPD representation, a non stationary noise estimation approach is proposed. This is from the fact that despite of changes in the non stationary noise intensity, the noise distribution characteristics remain the same over time. In SPD representation, this change in intensity is approximated as a shift in the normalized spectral magnitude of the noise distribution. therefore, if a SPD from a segment containing only noise, which is denoted as H_N (f,b),is extracted ,it can be assumed that the noise in the SPD containing both noise and signal is represented as H_N(f,b+a).Therefore, the problem simplifies to estimating 'a'- which is a change in non-stationary noise intensity. This step is illustrated in fig 4 where the cross correlation between the noise SPD and noisy sound spd is performed. This enables to find a_{max},that corresponds to the highest correlation between two signals. In turn this helps to get the upper bound of the noise estimate in the clip, based on the initial estimate from the noise SPD.



(a) An example of non-stationary wind noise, where the noise level increases and then decreases over time

(b) Noisy Whistle sound SPD, *H(f,b)*. The SPD noise estimate, *n(f)*, is estimated based on the maximum subband cross-correlation.

Fig. 4: Overview of the SPD noise estimate approach

The process begins with a noise SPD H_N (f, b) found using (3), here t is replaced by noise only frames t_N. The upper bound of the noise in H_N (f, b) is then estimated as the maximum occupied bin for each frequency subband:

$$n_{max}(f) = \max_b (H_N(f,b) > 0) \qquad (10)$$

n_{max}(f) is smoothed to avoid sharp discontinuities across frequency using a moving average filter of order M given by

$$n'_{max}(f) = \frac{1}{M} \sum_{i=f-\frac{M}{2}}^{i=f+\frac{M}{2}} n_{max}(i) \qquad (11)$$

Then an SPD, H(f,b) from a noisy sound clip is taken and correlation (*) between H_N (f,b) and H(f,b) is performed to find the intensity difference, a_{max} with highest correlation. Since H(f,b) is a mixture of noise and signal ,the cross correlation is performed separately on each SPD subband, f, such that the highest correlation should occur between the two noise dominated subbands:

$$a_{max} = \max_{a}[H(f,b) * H_N(f,b+a)] \quad \forall f \qquad (12)$$

The final SPD noise estimate is then given by

$$n(f) = n'_{max}(f) + a_{max} \qquad (13)$$

D. SPD-IF missing feature classification

The approach used here is based on masking the unreliable SPD-IF dimensions by using the noise estimate done in previous section. Missing feature classification is performed with kNN using the hellinger distance to measure the distribution distance between image features. For the SPD of the noisy sound clip, there exists a boundary, ∂H between clean and noisy regions. The region in the SPD above this boundary is derived only from the signal:

$$\exists \ \partial H : \forall(f,b) > \partial H \rightarrow H(f,b) = H_r(f,b) \qquad (14)$$

$H_r(f,b)$ represents the unreliable region of the SPD. It can be noted that the reliable SPD boundary, ∂H can be approximated by the noise estimate in the clip, n(f), as found in the previous section. Therefore the reliable region of the SPD $H_r(f,b)$ can be found as:

$$H(f,b) \rightarrow \begin{cases} H_r(f,b), & \text{if } \ b > n(f) \\ H_u(f,b), & \text{otherwise} \end{cases} \qquad (15)$$

Where subscript r and u represents reliable and unreliable regions respectively .This mask is applied to SPD-IF feature. If any sub block of the SPD image, denoted by $L_{i,j}$ is intersected by the noise estimated, n(f),it must be assumed that whole block is unreliable. Because the feature $x_{i,j}$ is based on the distribution statistics of image pixels in $L_{i,j}$ and hence will be affected by noise .therefore, the sub blocks where all pixels belong to $H_r(f,b)$ are reliable as given below

$$x_{i,j} \rightarrow \begin{cases} x_r, & \forall L_{i,j} \in H_r(f,b) \\ x_u, & \text{otherwise} \end{cases} \qquad (16)$$

x_u are unreliable feature dimensions and can be marginalized as they do not contain useful signal information .for classification, kNN is used even though it is uncommon in acoustic field and is relatively common in image processing. kNN can achieve comparable performance with SVM [9].the advantage of using kNN in this system is that, it can be easily combined with a missing feature framework and this is not straightforward for the SVM classification. kNN also offers flexibility in the choice of distance measure. Here Hellinger distance is used over the conventional Euclidean distance. Hellinger distance used here is a measure of the similarity between two distribution derived from the data, which fits naturally with the image feature as this models the distribution of pixels in sub block of each monochrome image. Characterizing the distribution by the mean m, and variance of the image pixel distribution which is obtained by second and third central moments of the blocks, the Hellinger distance between two SPD-IF vectors can be written as:

$$d_H(x, x_T) = \sum_{k=1}^{N_r}\left(1 - \sqrt{\frac{2\sigma_k\sigma_{T,k}}{\sigma^2{}_k + \sigma^2{}_{T,k}}}\, e^{-\frac{1}{4}\frac{(\mu_k - \mu_{T,k})^2}{\sigma^2{}_k + \sigma^2{}_{T,k}}}\right)^{\frac{1}{2}} \qquad (17)$$

Where N_r signifies the number of reliable dimension, x_T is sample from training data.

III. EXPERIMENT

A. Sound event database

50 sound event classes are selected from the Real Word Computing Partnership (RWCP) Sound Scene Database in Real Acoustical Environment [10], which gives a selection of action, collision and characteristic sounds. The sound files in this database have high SNR and each sound file contains an isolates sound event. There is some silence before and after the sound. The database has wide range of sound event types, including metal, wooden and china impacts, friction sounds and other sounds such as whistle, bell and clock. Many of these sound events have sparse time frequency spectrogram representation meaning that the most of the power contained in a particular frequency band, while some others sounds such as sandpaper or buzzer have diffuse spectrogram representation. From each sound event, 80 sound clips are collected. In that 50 files are selected randomly for training and 30 for testing. Therefore, overall with 50 sound events, this gives 2500 and 1500 samples for training and testing respectively.

B. Experiment setup

Here the proposed SPD-IF is compared with previously used method i.e., Spectrogram Image Feature (SIF), which is based on a raw power STFT spectrogram and Support vector machine (SVM) classifier [6]

C. Noise condition

For each experiment, the classification accuracy is investigated in mismatched noise condition, using only clean samples for training. The average performance for each method is noted in clean and at 20, 10 and 0dB SNR for the "speech babble" noise environment.

IV. RESULT

The result of the experiments is shown in Table1. It can be seen that the SPD-IF using log power spectrogram gives a comparable result in clean condition with raw power SIF. The main advantage of SPD-IF are demonstrated in mismatched noise condition, where even at 0dB it achieves the accuracy of 90.14% compared to 80.95% for raw power SIF.

TABLE I
RESULTS SHOWING THE CLASSIFICATION ACCURACY FOR THE PROPOSED SPD-IF AVERAGED AGAINST SIF ACROSS "SPEECH BABBLE" NOISE CONDITION

Methods	Clean	20dB	10dB	0dB	Average
Proposed SPD-IF	98.46%	97.60 %	95.80 %	90.14 %	95.5%
SIF	91.13%	91.10 %	90.71 %	80.95 %	88.55%

V. CONCLUSION

This paper proposes a novel feature extraction and classification method for robust sound event classification, motivated by visual perception of sound through images. The proposed subband power distribution (SPD) image improves over the previous SIF, as SPD transforms the reliable signal components to a localized region of the image this makes it simple to combine the extracted feature with the missing feature classification. Results shows that proposed SPD-IF is robust to noise with classification accuracy of over 90.14% at 0dB noise and overall average of almost over 95.5%

REFERENCES

[1] J. Dennis, H. D. Tran, and E. S. Chng, "**Image Feature Representation of the Subband Power Distribution for Robust Sound Event Classification**," IEEE Transactions on Audio, Speech, and Language Processing, vol. 21, pp. 367-377, Feb.2013.

[2] L. Gerosa, G. Valenzise, F. Antonacci, M. Tagliasacchi, and A. Sarti, "**Scream and gunshot detection in noisy environments**," in 15th European Signal Processing Conference, Sep. 3-7, Poznan, Poland, 2007.

[3] S. Chu, S. Narayanan, and C. Kuo, "**Environmental sound recognition with time–frequency audio features**," Audio, Speech, and Language Processing, IEEE Transactions on, vol. 17, no. 6, pp. 1142–1158, 2009.

[4] B. Ghoraani and S. Krishnan, "**Time–Frequency Matrix Feature Extraction and Classification of Environmental Audio Signals**," Audio, Speech, and Language Processing, IEEE Transactions on, vol. 19, no. 7, pp. 2197–2209, 2011.

[5] F. Weninger and B. Schuller, "**Audio recognition in the wild: Static and dynamic classification on a real-world database of animal vocalizations**," in Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on. IEEE, 2011, pp. 337–340.

[6] J. Dennis, H. Tran, and H. Li, "**Spectrogram image feature for sound event classification in mismatched conditions**," Signal Processing Letters, IEEE, vol. 18, no. 2, pp. 130–133, 2011.

[7] B. Raj and R. Stern, "**Missing-feature approaches in speech recognition**," Signal Processing Magazine, IEEE, vol. 22, no. 5, pp. 101–116, 2005.

[8] M. Slaney, "**An efficient implementation of the Patterson-Holdsworth auditory filter bank**," Apple Computer, Tech. Rep, 1993.

[9] O. Boiman, E. Shechtman, and M. Irani, **"In defense of nearest-neighbor based image classification,"** in Computer Vision and Pattern Recognition, CVPR. IEEE, 2008, pp. 1–8.

[10] S. Nakamura, K. Hiyane, F. Asano, T. Nishiura, and T. Yamada, "**Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition**," in Proc. ICLRE, 2000, pp. 965–968.