Image Feature Representation of the Subband Power Distribution for Robust Sound Event Classification

Jonathan Dennis, Student Member, IEEE, Huy Dat Tran, Member, IEEE, and Eng Siong Chng, Senior Member, IEEE

Abstract—The ability to automatically recognise a wide range of sound events in real-world conditions is an important part of applications such as acoustic surveillance and machine hearing. Our approach takes inspiration from both audio and image processing fields, and is based on transforming the sound into a two-dimensional representation, then extracting an image feature for classification. This provided the motivation for our previous work on the spectrogram image feature (SIF). In this paper, we propose a novel method to improve the sound event classification performance in severe mismatched noise conditions. This is based on the subband power distribution (SPD) image — a novel two-dimensional representation that characterises the spectral power distribution over time in each frequency subband. Here, the high-powered reliable elements of the spectrogram are transformed to a localised region of the SPD, hence can be easily separated from the noise. We then extract an image feature from the SPD, using the same approach as for the SIF, and develop a novel missing feature classification approach based on a nearest neighbour classifier ($k$NN). We carry out comprehensive experiments on a database of 50 environmental sound classes over a range of challenging noise conditions. The results demonstrate that the SPD-IF is both discriminative over the broad range of sound classes, and robust in severe non-stationary noise.

Index Terms—Sound event classification, subband power distribution, spectrogram, missing feature theory.

I. INTRODUCTION

SOUND event classification is the task of assigning the audio content of a short sound clip into one of a set of pre-trained classes. This has a wide range of important applications, such as acoustic surveillance [1], bioacoustic monitoring [2], environmental sounds [3,4], or in the more general field of machine hearing [5]. In many of these applications, the sound events occur in the presence of a wide variety of challenging noise conditions, and where the signal-to-noise ratio (SNR) may even approach 0dB.

In our previous work on the SIF [6], we were inspired by the concept of “spectrogram reading” [7] to develop a two-dimensional sound event classification approach based on the time-frequency spectrogram. This approach used image processing techniques, such as grey-scale normalisation, dynamic range quantisation and mapping, and pixel distribution statistics, to generate a robust representation of the visual information. Classification was performed using Support Vector Machines (SVM), where the most discriminative components, typically high-power quantisation regions of the dynamic range, should be assigned a higher weight. Due to the physical nature of many sounds, the spectrogram is sparse, with the sound energy concentrated in a small number of localised frequency bands, compared to the diffuse noise which is spread more evenly across the frequency spectrum. Therefore, in noisy conditions the low-power quantisations of the spectrogram are the most affected, while high-power regions remained relatively robust. An improvement to this previous SIF approach could therefore be to include a missing feature framework to marginalise the regions that are affected by the noise. However, missing feature mask estimation poses a practical problem, as the non-stationary nature of the noise across time, frequency and dynamic range makes developing a reliable mask challenging [8].

In this paper, we propose a novel sound event image representation, called the subband power distribution (SPD). The SPD captures the distribution of the sound’s log-spectral power over time in each subband, such that it can be visualised as a two-dimensional image representation of frequency against normalised power. The advantage of this representation over the spectrogram is that the reliable, high-power spectral components are mapped to a localised region of the SPD image. This simplifies the problem of missing feature mask estimation, as the SPD has a large, continuous region of the image that can be used for robust classification. We can then extract an image feature from the SPD, which we call the SPD-IF, using the approach previously proposed for the SIF. For classification of the SPD-IF, we propose a missing feature framework to marginalise the noisy region of the SPD image utilising $k$NN with the Hellinger distance, to measure the distribution distance between two image features. This offers an improvement over $k$NN with the conventional Euclidean distance, which was used in our initial presentation of the SPD approach in [9]. We also propose a method to estimate the noise mask based on the SPD representation, which can adapt to non-stationary noise environments and provide an improvement over conventional stationary noise estimates.

The rest of the paper is organised as follows. Section II first covers related work on sound event classification. Section III then details the proposed SPD image and missing feature classification. In Sections IV and V, we validate our approach through a series of experiments, before Section VI concludes the work.
II. RELATED WORK

A. Literature Review

The literature for sound event classification reveals that many previous approaches have been based on speech recognition techniques [10], where it is common to complement traditional speech features with others, such as from the MPEG-7 standard, to improve the performance [11,12]. More recently, techniques have emerged that are specifically designed for sound event classification, and these fall into two broad categories: temporal and time-frequency features.

An example of temporal feature extraction is found in [13], where features are developed that can capture the temporal evolution of the sound event using a small set of morphological descriptors. Another example is found in [14], where the authors propose to characterise the subband temporal envelope of the sound, and combine this with a pattern classification method based on the subband probabilistic distance SVM. In both cases, the use of temporal features was shown to give improvements over the baseline methods.

Examples of time-frequency features can be found in [3,4], where both approaches use the matching pursuit (MP) algorithm to perform a time-frequency decomposition. Features can then be extracted either directly from the MP decomposition, as in [3], or through a further decomposition of the sound spectrogram using non-negative matrix factorisation, further followed by feature extraction, as performed in [4]. A different approach is to characterise local spectro-temporal modulations using two-dimensional Gabor filter functions, and use feature selection to find a small set of functions that can characterise the acoustic information [15]. The output of these filters is used directly as a feature for a Hidden Markov Model (HMM) classifier, or combined with a more complex Tandem system [16]. There is also research in the area of image-based sound classification, which can be found in [17]–[20]. For example, in [17], a texture-of-texture approach is used, which recursively applies filters to spectrogram and MFCC images to capture certain shapes, such as horizontal and vertical lines in the image, can classify music into one of three genres. Another approach in [19] uses local image patches, which are randomly selected during training, to classify musical instruments. A feature is then extracted for classification based on the minimum matching energy of each patch over the testing spectrogram.

Our approach falls generally into the category of time-frequency methods, since we similarly aim to capture the joint information across two dimensions. However, unlike other approaches, we do not perform feature extraction directly on the spectrogram. Instead, we transform the sound event into the SPD image representation, before extracting an image feature. This step is advantageous, as firstly the SPD has fixed dimensions, which simplifies the extraction of the image feature signature. Secondly, unlike the spectrogram, the signal and noise are easily separated in the SPD, which allows us to develop a missing feature classification approach, which works well in severe noise conditions.

B. Conventional Missing Feature Techniques

As noise robustness has not be investigated in the context of image-based approaches, we choose to compare our missing feature SPD-IF against conventional missing feature methods from speech recognition. This makes for a better comparison, as these methods are based on the same concept of robust classification with reliable and unreliable data, and have a proven performance in noisy conditions [8,21]. A typical system uses a frame-by-frame feature, combined with HMM using Gaussian mixture models (GMM) for classification. Table I shows how two common missing feature methods compare to our SPD-IF kNN approach. Below we give a short summary of each approach [21]:

- **Imputation:** Here, the approach is to estimate values for the unreliable spectral components using the information available from the remaining reliable components. The approach uses replacement values drawn from the conditional distribution, given the reliable components.
- **Marginalisation:** Here, the approach is to compute the output probabilities of each state by integrating out the distribution of the unreliable components.

Both methods can also be bounded using the knowledge of the minimum and maximum possible value of each component. The minimum is typically taken to be zero, while the maximum is given by the observed spectral value.

The drawback of these two missing feature techniques is that they have a high computational cost. This is due to the fact that the missing feature calculations must be computed for each possible combination of state and observation. Another problem is that they must be applied in the spectral domain, which means the advantages of the cepstral domain, such as uncorrelated features, are lost. Despite this, both imputation and marginalisation provide a good baseline comparison for noisy sound classification, hence are implemented and compared in our experiments.

However, it has been shown that the performance of these methods depends greatly on the accuracy of the estimated mask. This mask determines which regions of the spectrogram are reliable signal, and which have been corrupted by noise. There are two common approaches [22]:

- **Local SNR:** this estimates the noise using regions of the spectrogram without any signal, and allows a local SNR to be calculated for each element in the spectrogram. Elements below a threshold are set as unreliable.
- **Classifier-based:** A feature set is extracted to represent the local SNR, and a classifier is used to decide the reliability of each element.
However, neither method may perform satisfactorily for sound event classification. For the first case, in 0dB noise, some noise elements may have similar sharp peaks and power as the signal. Therefore, such regions may be marked as reliable, and have a severe effect on the classification performance. In addition, the local SNR is often based on a stationary noise estimate, hence does not work well for non-stationary noise, which is more typical of realistic noise environments. The classifier-based method requires prior knowledge about the characteristics of the signal. While this works well for speech, it is difficult to design a feature to reliably capture the wide variety of sound event characteristics. Therefore, it is challenging to generate a missing feature mask which can be reliably combined with spectrogram-based approaches, and we found in preliminary experiments that this was the case for our previous SIF method. The advantage of the proposed two-dimensional SPD representation over the spectrogram, is that the reliable and unreliable areas of the SPD are separated simply by a line, as opposed to the time-frequency surface required for the spectrogram. The peaks in the spectrogram belonging to the noise, which may easily be mistaken for signal by a SNR-based noise estimate, will be scattered over time and frequency. Therefore, they will not contribute significantly to the distribution information in SPD, which will be dominated by the more continuous signal information, and as such the classification of the SPD will be more robust.

III. SUBBAND POWER DISTRIBUTION IMAGE FEATURE

In this section, we present our proposed SPD image representation for robust sound event classification [9]. An overview of the proposed system is shown in Fig. 1, and is compared alongside our previous SIF approach [6].

The SPD is based on the distribution of the subband spectral power over time, and is designed such that the reliable parts of the signal are transformed to a continuous region of the SPD representation, making them easily separable from the noise. As can be seen in Fig. 1, after a contrast enhancement step to produce the SPD image we extract an image feature (SPD-IF) using the same approach as previously used for the SIF. This process involves a quantisation and mapping of the dynamic range into a higher dimension, analogous to pseudo-colour mapping in image processing. We then develop a robust missing feature classification system, that generates a missing feature mask based on the SPD and then marginalises the noisy feature dimensions. For classification we use $k$NN with the Hellinger distance measure, as we found that this takes into account the distribution information captured in the SPD-IF, hence is a more natural way to measure the similarity compared to the Euclidean distance between features.

A. Subband Power Distribution Image Algorithm

Starting from a time-frequency spectrogram representation of the sound $S(f, t)$, the SPD image is designed to represent the distribution of spectral power in each frequency subband over time. An example of this process is shown in Fig. 2. Here, the probability distribution of the log-power spectrogram of a bell sound is calculated for each subband, and then stacked together to form a two-dimensional representation of frequency against normalised spectral power, as shown in Fig. 2c. The SPD then undergoes a contrast enhancement step to give the final SPD image in Fig. 2d, which is performed to ensure the important signal information is represented over the full $[0, 1]$ range of the image. The SPD image then forms the basis for our image feature extraction and classification approach.

As the SPD captures the long term temporal distribution statistics, it is desirable for $S(f, t)$ to have a high time resolution, to better capture the distribution. Here, we choose to use the gammatone filterbank decomposition as our model for time-frequency analysis, which is derived from the cochlear filtering in the inner ear [23]. This has the advantage that there is no tradeoff between time and frequency resolution as is common with the conventional short-time Fourier transform (STFT) representation. Here, we use a bank of 50 filters, with centre frequencies equally spaced between 100 and 8000 Hz on the equivalent rectangular bandwidth (ERB) scale [24].

The SPD is based on the normalised log-power spectrogram, $G(f, t)$, which we write as:

$$G(f, t) = \frac{\log S(f, t)}{\max_{f, t} (\log S(f, t))}$$

(1)

where log-power is used to compress the dynamic range of the spectrogram to enhance the high-power elements in the SPD. We set all $G(f, t) < 0$ equal to zero, which normalises $G(f, t)$ into a grey-scale image in the range $[0, 1]$. This ensures that the relative volume of different sound clips is equalised, and that the high-power elements are always transformed to the same region of the SPD.

The SPD represents the distribution of power in each frequency subband of the normalised spectrogram over time:

$$D_f(z) = P_{G_f}(z)$$

(2)

where $z$ is a variable representing the normalised spectral power, $P$ is the probability density function, and $G_f$ is a random variable representing the normalised spectrogram.
G(f, t) in the frequency subband, f. The SPD then forms a two-dimensional representation, D(f, z), by stacking together each subband distribution over frequency, f.

As the upper and lower bounds of the distribution are fixed, we choose to estimate the distribution, D(f, z), using a non-parametric approach based on the histogram, for its speed and simplicity. Therefore the SPD becomes:

\[ H_R(f, b) = \frac{1}{t_{max}} \sum_t I_b(G(f, t)) \quad (3) \]

where \( t_{max} \) is the number of time samples in the segment and \( I_b \) is the indicator function, which equals one for the \( b^{th} \) bin if the normalised log-spectral power \( G(f, t) \) lies within the range of the bin and is zero otherwise. In our experiments, we use a total of 100 bins, with the bin edges equally spaced over the [0, 1] range of the normalised spectral power, z.

The values in \( H_R(f, b) \) are the raw probability distribution information for each frequency subband over time, which are constrained to lie in the range:

\[ 0 \leq H_R(f, b) \leq 1. \quad (4) \]

Although this implies that \( H_R(f, b) \) is already a grey-scale image, we found that most of the information is contained within a small region of the dynamic range. This is due to the physical nature of many sound events, which have an attenuating or otherwise non-stationary spectrogram envelope. This means that, for a high enough number of histogram bins, it is unlikely for the subband distribution density values in any one bin to be high. Therefore, we propose to enhance the contrast of the raw SPD, to produce the enhanced SPD:

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

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

**B. Image Feature Extraction**

Given the two-dimensional SPD image, we want to extract a feature that characterises the sound information in the image for classification. Here we adopt the image feature approach from our previous work in [6], which was previously applied directly to the spectrogram, as shown in Fig. 1. The approach uses a quantisation and mapping of the dynamic range of the image, as shown in Fig. 3a, to produce a set of quantised “monochrome” images, such that the information from each monochrome is characterised separately in the image feature. The feature is then extracted through partitioning of the SPD into local sub-blocks and extraction of the image pixel distribution statistics, as shown in Fig. 3c. Although this image feature is relatively simple, it is well suited to capture the information in texture images, such as the spectrogram or SPD, which lack the geometrical structure that is the basis for other more popular features in object recognition, such as the scale-invariant feature transform (SIFT) [26].

The central idea here is to capture the image information jointly across the two dimensions of the image, and across the dynamic range, through a mapping similar to pseudo-colourmapping from image processing. As found in our previous work [6], this enables the quantised regions of the distribution to be characterised separately in the feature, such that noise fluctuations only affect a particular quantisation region, improving the overall robustness of the feature. Secondly, we found in practise that this step helps to reduce the distribution variance in each quantised region, increasing the discriminative capability and hence enhancing the classification results. Therefore, while we could extract an image feature directly from the grey-scale SPD, we found that the performance can be improved through this quantisation and mapping.

The first step therefore is to quantise and map the dynamic range of the SPD, \( H(f, b) \), into a set of monochrome images, \( M_c \), as follows:

\[ M_c(f, b) = h_c(H(f, b)) \quad \forall c \in (c_1, c_2, \ldots c_N). \quad (6) \]

where \( h_c \) is the nonlinear mapping function for mapping dimension \( c \). This operation can be seen as a generalisation of the pseudo-colourmapping operation from image processing.
where \( c = 3 \) would represent the red, green and blue colour dimensions. Here we utilise a mapping function from image processing, similar to that of the “Jet” colourmap in Matlab, as we found in our previous work that such existing colourmaps provide a suitable quantisation of the dynamic range [6]. This is shown graphically in Fig. 3a, where each of the three mapping dimensions can be written as:

\[
h_c(H(f, b)) = \begin{cases} 
  \frac{l_2 - l_1}{u_2 - u_1}, & \text{for } l_2 < H(f, b) < l_1 \\
  1, & \text{for } l_1 \leq H(f, b) \leq u_1 \\
  \frac{u_1 - u_2}{u_2 - H(f, b)}, & \text{for } u_1 < H(f, b) < u_2 \\
  0, & \text{otherwise.}
\end{cases} \tag{7}
\]

where the parameter set \( \{l_1, l_2, u_1, u_2\} \) for the three colour mappings are \( c_{\text{red}} = \{3, 5, 7, 9\} \), \( c_{\text{green}} = \{1, 3, 5, 7\} \), and \( c_{\text{blue}} = \{-1, 1, -3, 3\} \). It can be seen from Fig. 3b that each monochrome captures information from different regions of the dynamic range. In particular, the “blue” quantisation enhances the least stationary information from the spectrogram, but may be more susceptible to non-stationary noise fluctuations, while the “red” monochrome captures the more stationary information, which for the signal region is reliable.

To characterise the information in each monochrome, we now partition the image into two-dimensional local sub-blocks, each of size \((\frac{F}{2}, \frac{F}{2})\), giving a total of \( D^2 \) blocks, as shown in Fig. 3c. For clarity, we drop the \( c \) notation for each mapping, therefore each local sub-block can be written as a subset of pixels from the whole monochrome as:

\[
L_{i,j} \subset M(f, b) \tag{8}
\]

where \( i, j = [1, 2, \ldots, D] \) are the indices of the sub-blocks, and \( L_{i,j} \) represents the region of the monochrome image, \( M(f, b) \), corresponding to the specified sub-block. We note that overlapping blocks could be used, however we found this did not significantly improve the results.

Next, we extract the image feature, \( x_{i,j} \), from each local sub-block, based on the distribution of the image pixels:

\[
x_{i,j} = \{\mu(L_{i,j}), \sigma^2(L_{i,j})\}. \tag{9}
\]

where \( \mu(\cdot) \) and \( \sigma^2(\cdot) \) are the mean and variance respectively of the pixels in the sub-block, characterising the distribution as a Normal distribution. Experimentally, we found that partitioning each image dimension into \( D = 10 \) blocks gives a good tradeoff between performance and feature vector size. The number of features is therefore \( 10 \times 10 \times 2 = 200 \), and there are a hundred local blocks, three monochrome mappings (\( c = 3 \)), and we use two parameters (mean and variance) to represent the distribution of image pixels in each block.

### C. Noise Estimation based on the SPD

Here, we propose a non-stationary noise estimation approach based on the SPD representation. This idea is based on the observation that despite changes in the non-stationary noise intensity, the characteristics of the noise distribution remain the same over time. In the SPD representation, this change in intensity can be approximated as a shift in normalised spectral magnitude of the noise distribution. Hence, if we extract an SPD from a segment containing only noise, which we call \( H_N(f, b) \), we can assume that the noise in the SPD containing both noise and signal is represented as \( H_N(f, b + a) \). The problem therefore simplifies to estimating \( a \) – the change in the non-stationary noise intensity. This is illustrated in Fig. 4, where we perform the cross-correlation between the noise SPD in Fig. 4b, and the noisy sound SPD in Fig. 4c. This enables us to find \( a_{\text{max}} \), which corresponds to the highest correlation between the two. This enables us to get the upper bound of the noise estimate in the clip, based on an initial estimate from the noise SPD.

The processing begins with a noise SPD, \( H_N(f, b) \), found using (3), where \( t \) is replaced by the noise-only frames 2\( t_N \). The upper bound of the noise in \( H_N(f, b) \) is then estimated as the maximum occupied bin for each frequency subband:

\[
n_{\text{max}}(f) = \max_b(H_N(f, b) > 0) \tag{10}
\]

This is smoothed to avoid sharp discontinuities across frequency using a moving average filter, of order \( M \):

\[
n'_{{\text{max}}}(f) = \frac{1}{M} \sum_{i=f-M/2}^{i=f+M/2} n_{\text{max}}(i) \tag{11}
\]

Then, given an SPD, \( H(f, b) \) from a noisy sound event clip, we take the cross-correlation (\( * \)) between \( H_N(f, b) \) and
In the sound-clip SPD on the right, unlike the proposed SPD classification with

distributions, we perform the cross-correlation separately on each SPD subband, \( f \), such that the highest correlation should occur between two noise-dominated subbands:

\[
a_{\text{max}} = \max_a \left[ H(f, b) \ast H_N(f, b + a) \right] \quad \forall f
\]  

(12)

The final SPD noise estimate, \( n'(f) \), is then simply:

\[
n(f) = n'_{\text{max}}(f) + a_{\text{max}}.
\]  

(13)

Note that we do not limit \( a_{\text{max}} \) to be positive, as the noise intensity can both increase or decrease, as shown by the wind noise example in Fig. 4a.

In our experiments, we compare this with a conventional approach that estimates the upper bound of the noise during noise-only periods of the signal and assumes the noise to be stationary. Instead of basing the estimate simply on the mean, as in spectral subtraction [27], here we take the upper bound as the mean plus two standard-deviations plus a constant, which gave good results during our preliminary work [9]:

\[
n'(f) = \mu(G(f, t_N)) + 2\sigma(G(f, t_N)) + \Delta
\]  

(14)

where \( \mu(.) \) and \( \sigma(.) \) are the mean and standard deviation over the noise-only time frames, \( t_N \), and \( \Delta \) is a small constant added to account for unseen noise fluctuations. However, as this approach cannot be used to estimate the noise when the signal is present, it cannot track the noise well in non-stationary noise environments, where the noise intensity varies. This is demonstrated in Fig. 4, where the stationary noise estimate, shown by the dotted lines, does not track the noise in the sound-clip SPD on the right, unlike the proposed SPD noise estimate shown by the solid line.

D. SPD-IF Missing Feature Classification

Our proposed approach is based on masking the unreliable SPD-IF dimensions using the noise estimate proposed in the previous section. We then perform missing feature classification with kNN, using the Hellinger distance to naturally measure the distribution distance between image features.

To generate the missing feature mask, we first must understand how noise affects the SPD representation. We find that due to their physical characteristics, many sound events have a sparse time-frequency spectrogram representation, meaning that a large proportion of energy is contained in a few frequency bands. By comparison, many noise conditions are diffuse, with energy spread across the frequency spectrum. Hence, even for 0dB noise, there are spectral components that satisfy:

\[
s \gg n \Rightarrow \log(|s| + |n|) \approx \log |s|
\]  

(15)

where \( s \) is the signal, \( n \) is the noise, and the approximation can be made due to the MixMax principle [28].

By taking the distribution over time, as in the SPD, the sparse, high-power components form a continuous region of the image, as opposed to the spectrogram, where the reliable parts are isolated over time and frequency. For the noisy sound-clip SPD, \( H(f, b) \), there exists a boundary, \( \partial H \), between clean and noisy regions, where the SPD region above this boundary is derived only from the signal:

\[
\exists \partial H: \forall (f, b) > \partial H \rightarrow H(f, b) = H_r(f, b).
\]  

(16)

where \( H_r(f, b) \) refers to the reliable region of the SPD.

It can be seen that the reliable SPD boundary, \( \partial H \), can be approximated by the noise estimate in the clip, \( n(f) \), as found in the previous section, since this is also an upper bound on the noise distribution in the SPD. Therefore, the reliable region of the SPD, \( H_r(f, b) \), can be found as:

\[
H(f, b) \rightarrow \begin{cases} 
H_r(f, b), & \text{if } b > n(f) \\
H_u(f, b), & \text{otherwise.}
\end{cases}
\]  

(17)

where the subscripts \( r, u \) denote reliable and unreliable image regions respectively.

Now we can apply this mask to the SPD-IF feature. If a sub-block of the SPD image, denoted \( L_{ij} \) in (9), is intersected by the noise estimate, \( n(f) \), we must assume that whole block is unreliable. This is because the feature, \( x_{ij} \), is based on the distribution statistics of the image pixels in \( L_{ij} \), hence will be affected by the noise. Therefore, only sub-blocks where all pixels belong to \( H_r(f, b) \), are reliable, as follows:

\[
x_{ij} \rightarrow \begin{cases} 
x_r, & \forall L_{ij} \in H_r(f, b) \\
x_u, & \text{otherwise.}
\end{cases}
\]  

(18)

The unreliable feature dimensions, \( x_u \), can now be marginalised, as they do not contain useful signal information.
For classification, we propose to use kNN, which, although uncommon in the acoustic field, is relatively common in image processing, and can achieve comparable performance with SVM [29]. For our purpose, its advantage comes from the fact this it can be easily combined with a missing feature framework, which is not straightforward for the SVM classifier. It also offers flexibility in the choice of distance measure, hence we propose to use the Hellinger distance over the conventional Euclidean distance. The Hellinger distance is a measure of the similarity between two distributions derived from the data, which fits naturally with our image feature as this models the distribution of pixels in each monochrome image subblock. Among the probabilistic distances, the Hellinger distance has distribution of pixels in each monochrome image subblock. which fits naturally with our image feature as this models the similarity between two distributions derived from the data, hence provides a useful comparison.

### Hellinger Distance

The Hellinger distance, which is calculated as:

\[
H(x, x') = \sqrt{\frac{1}{2N_r} \sum_{k=1}^{N_r} \left( 1 - \frac{2 \sigma_k \sigma'_{T,k}}{\sigma_k^2 + \sigma'_{T,k}^2} e^{-\frac{1}{2} \left( \frac{(\mu_k - \mu'_{T,k})^2}{\sigma_k^2 + \sigma'_{T,k}^2} \right) } \right) ^{\frac{1}{2}}}
\]

where \(N_r\) is the number of reliable dimensions, and \(x, x'\) is a sample from the training data. For some feature dimensions however, the variance will be small or even zero. Therefore, we apply a floor to the observed variance, set at \(1e^{-3}\), such that a measure of the similarity between these dimensions can still be obtained.

For completeness, we carry out experiments to compare the proposed Hellinger distance to the conventional Euclidean distance, which is calculated as:

\[
E(x, x') = \frac{1}{N_r} \left[ \sum_{k=1}^{N_r} (x_T(k) - x'_T(k))^2 \right]^{\frac{1}{2}}
\]

This uses the mean and variance parameters as the feature vector, rather than using them to calculate the distribution distance, hence provides a useful comparison.

### IV. Experiments

#### A. Sound Event Database

A total of 50 sound event classes are selected from the Real Word Computing Partnership (RWCP) Sound Scene Database in Real Acoustical Environments [31], giving a selection of collision, action and characteristics sounds. The sound files have a high SNR, and each contains an isolated sound event, with some silence before and after the sound. There are a wide range of sound event types, including wooden, metal and china impacts, friction sounds, and others such as bells, phones ringing, and whistles. Many of the sound events have a sparse time-frequency spectrogram representation, with most of the power contained in a particular frequency band, while several others are more diffuse, such as buzzer or sandpaper.

For each run, from the 80 sound clips in each class, 50 files are randomly selected for training and 30 for testing. Overall, with 50 sound classes, this gives 2500 and 1500 for training and testing respectively, with experiments repeated over 5 runs.

#### B. Experiment Setup

The first experiment compares the proposed SPD-IF against conventional baseline methods. Here, the following classification methods are implemented and evaluated:

1. Proposed SPD-IF, based on the best performing approach using a log-power gammatone spectrogram and the kNN classifier with Hellinger distance measure and SPD noise mask from (13).
2. Spectrogram Image Feature (SIF), based on the best performing raw-power STFT spectrogram and SVM classifier from [6].
3. Baseline MFCC-HMM using 36-dimension frame-by-frame MFCCs, with 12 cepstral coefficients, without the zeroth component, plus their deltas and accelerations.

We implement the following:

- a) Plain MFCCs without noise reduction.
- b) Advanced Front End (AFE) noise reduction [32].
- c) Multi-conditional training method.

4. Baseline missing feature approaches, with 36-dimension Mel-frequency spectral coefficients (MFSCs) without deltas. We implement both:

   - a) Bounded Imputation.
   - b) Bounded Marginalisation.

5. Baseline multi-conditional Gabor-HMM, with the best performing 36 Gabor features selected in each run of the experiment using the Feature Finding Neural Network (FFNN) approach from [15].

Each of the HMM methods above uses 5 states and 6 Gaussian mixtures, and both training and testing is carried out with HTK [33], except for the missing feature methods, where testing is carried out with a local Matlab decoder. For the kNN classifier used for the proposed SPD-IF, we set the parameter \(k = 5\) for the class decision, based on preliminary experiments. This parameter is then used throughout.

A second set of experiments is designed to breakdown the SPD-IF method and compare each step of the process separately. Here, we use only the “Factory Floor” noise condition, as this was found to be the most difficult noise condition for each of the above methods. We compare the following:

1. SPD vs. Spectrogram image representation – to show the improvement gain over our previous work on the SIF.
2. Hellinger vs. Euclidean distance measure – to demonstrate the improvement when using the distribution distance between image features.
3. SPD vs. Stationary noise estimate – to show the improvement in non-stationary noise for the SPD method.

For the first case of SPD vs. Spectrogram, we ensure a fair comparison by generating a missing feature approach for the SIF, which we call MF-SIF. This uses the same kNN classification system as proposed for the SPD-IF, while the noise mask for the MF-SIF is derived in an analogous way to that of the SPD-IF, albeit across the time, frequency and dynamic range dimensions of the log-power STFT spectrogram.

#### C. Noise Conditions

For each experiment, except for the multi-conditional method, the classification accuracy is investigated in mis-
matched conditions, using only clean samples for training. The average performance for each method is reported in clean and at 20, 10 and 0 dB SNR for the following four noise environments: “Speech Babble”, “Destroyer Control Room”, “Factory Floor 1” and “Jet Cockpit 1”, obtained from the NOISEX92 database [34]. All four noises have most of the energy concentrated in the lower frequencies, and represent realistic non-stationary noise conditions.

For the multi-conditional MFCC-HMM and Gabor-HMM baseline methods, the system is trained on three out of the four noise environments with clean samples and those at 10dB SNR. Testing is then carried out on the remaining noise environment, across all four SNR conditions.

V. RESULTS AND DISCUSSION
A. Proposed SPD-IF vs. Baseline

A summary of the experimental results is shown in Fig. 5, while the complete set of results, with a breakdown for each SNR condition, can be found in Table II. These results demonstrate that the SPD-IF performs significantly better than the baseline results, with an average classification accuracy of 96.0% compared with 88.6% for the SIF, 88.0% for the multi-conditional MFCC-HMM, and 84.9% for the multi-conditional Gabor-HMM. For the individual noise conditions, it can be seen that the SPD-IF outperforms the baselines in almost every case. The only exception is in clean conditions, where the plain MFCC-HMM, with its sharp model, performs marginally better, but with a difference of less than 1%.

The performance of the proposed SPD-IF in clean conditions reflects the discriminative nature of the SPD representation for a wide variety of sound event classes, while the performance in noisy conditions demonstrates the ease of separation between the noise and signal regions of the SPD, for combination with the kNN missing feature approach. For example, in 0dB noise, the SPD-IF achieves an average accuracy of over 90%, which is around 10% higher than the next closest result for the SIF, and is nearly a 30% improvement over the HMM baseline techniques. Among the multi-conditional methods, the MFCC-HMM method performed better on average than the Gabor-HMM approach, although the Gabor method performed better in clean and 10dB noise condition. As the Gabor feature selection in training was carried out in these noise conditions, it appears that the method becomes tuned to these noise conditions, and did not generalise well to the other conditions.

One unexpected result was the performance of the MFSC-HMM missing feature methods, which did not perform as well as the multi-conditional training method. One of the factors contributing to this result may be that the missing feature approaches are limited to the spectral domain, rather than the cepstral domain. It was also found that the problem of modelling in the spectral domain was exacerbated by the sparse time-frequency spectrogram representation of the sound events, particularly those with a decaying time envelope, such as the bell sound in Figure 2a. Here, the variance of the decaying frequency components became very large, of the order of several hundred dB, such that the likelihood for certain mixtures could become very high when few feature components were missing. For the marginalisation approach, the converse then became true when these frequency dimensions were masked, as the integration favours mixtures with small variances that fall within the upper and lower bounds, such as those belonging only to background noise. This caused confusion among certain sound classes, even in clean conditions, resulting in the poorer overall performance.

B. Detailed Experiments on the SPD-IF

1) SPD vs. Spectrogram: The overall results are shown in Fig. 6, while the detailed results can be found in Table III. It can be seen that the proposed SPD-IF outperforms the MF-SIF method using the same missing feature kNN classification approach. The reason for this result is that the SPD representation has the advantage that the noise and signal form continuous regions of the image, making them easy to separate given an estimate of the noise. However, for the MF-SIF, the mask must be applied across both time, frequency, and the dynamic range mappings of the image feature, making it less reliable in non-stationary noise. In addition, the SIF has only a coarse dynamic range quantisation, with just three mapping dimensions as in pseudo-colouring in image processing. Hence we found that the noise mask could label some signal regions as unreliable when only a small amount of noise corrupted a region of the spectrogram. On the other hand, the SPD-IF has ten partitions across the dynamic range, hence the noise mask can better separate the observed noise and signal regions.

2) Hellinger vs. Euclidean Distance Measure: Finally, we compare the effect of utilising the Hellinger distance measure for the SPD-IF method. It can be seen by the bars on the right of Fig. 6, and for the last entry in Table III, that there is a
small but consistent improvement for the Hellinger distance measure. This is expected, as the image features represent the pixel distribution information, hence measuring the distribution distance, rather than the conventional Euclidean distance, gives a more representative comparison.

3) SPD vs. Stationary Noise Estimates: Comparing the two methods in Fig. 6, we can see that the proposed SPD noise estimate consistently outperforms the conventional stationary noise estimate. From Table III, we can see that the most significant improvement is found at 0dB. For example, for the SPD-IF with Hellinger kNN, the SPD noise estimate achieves a 6% absolute improvement over the stationary method. This highlights the ability of the SPD noise method to adapt to the non-stationary noise.

C. Discussion

1) Spectral Content of Sound Events: While the SPD captures the long-term temporal statistics of the sounds through the subband distribution, it does not explicitly model the temporal structure of the sound. Therefore, it may be possible to generate artificial sounds, such as simple upward and downward transients, that have similar SPD representations, but could be easily classified visually in the spectrogram or using an HMM approach. However, in practise, natural sound phenomena rarely conform to such simple examples, as they typically have distinct increase/release energy cycles, which may be different across frequency subbands. Therefore, classification using the SPD-IF is able to distinguish between a wide variety of natural sounds, which is demonstrated by our experiments on a large database containing 50 sound events. The SPD-IF classification also works well in noisy conditions, provided that the sound spectrogram contains at least a few characteristic, high-power components that can be mapped to give a reliable region of the SPD for classification. Due to the physical nature of sounds and noise, this should be the case down to very low SNRs, as the noise energy is diffuse across the spectrum. This is demonstrated by the performance of the SPD-IF in 0dB non-stationary noise conditions, where it achieves a classification accuracy of over 90%.

While not the focus of this paper, we also believe this approach can work for phone classification in speech, using a suitably chosen time-frequency representation that highlights the formant information. The choice of acoustic unit is important, as the SPD may not be suitable for isolated word recognition, if the whole word is treated as one unit. This is because the distribution over several phones in connected speech will result in different words containing mixtures of the same information, hence becoming harder to separate.

2) Image Feature vs. Frame-by-Frame Features: A question that might be asked is why we can compare the SPD-IF, a two-dimensional feature, with frame-by-frame approaches using MFCCs or Gabor features. If we simply compare the feature dimensions, then at first this appears valid, as the SPD-IF has 600 dimensions, while the frame-by-frame features have just 36. However, once combined with HMMs, the number of parameters increases dramatically. For the model used in these experiments, with 5 states and 6 Gaussians, we have 600 × 36 × 2 × 5 = 2160 parameters for each HMM model, where the 2 represents the mean and variance of each Gaussian. In addition, the Gabor features represent local two-dimensional spectral information in the spectrogram, which makes for an interesting comparison with the global two-dimensional information extracted in the SPD-IF. We also note that such methods are considered state-of-the-art for acoustic recognition tasks, particularly in speech, therefore we consider the experimental comparison in this paper to be valid.
3) Online Recognition System: Many applications of sound event classification require real-time performance. A typical system consists of a sound activity detector, followed by classification of the detected sound event segments. We note that our method is not limited to use any particular detection algorithm, although requires a set of noise-only frames to obtain the noise estimate. To investigate the performance of our approach, we implemented our proposed method in C#. Here, we found that the feature extraction and classification of the SPD-IF, with 50 sound classes, runs around 10 times faster than real-time. We found this comparable in speed to both Gabor-HMM and MFCC-HMM methods using HTK, although the Gabor method required slightly more computation for feature extraction. For the missing feature methods, we compared the speed of the marginalisation approach using the implementation in the CASA-Toolkit (CTK) [35]. This runs at around one times real-time, meaning a one second sound clip requires a further one second before the classification result becomes available.

One advantage of the SPD-IF for online systems is that retraining is exceptionally fast, as the kNN classification algorithm only requires feature extraction and storage. It also does not require a large number of samples for training, which makes the SPD-IF suitable for applications requiring a quick initial training setup, followed by additional training to be added later. This is not possible with HMM methods, which require the whole model to be recalculated, and in general require more time and data for training. In addition, while for simple kNN the amount of computation is proportional to the number of features extracted during training, there are algorithms which can speed up high-dimensional kNN searches. While these are not employed here, it can solve the potential problem with large training databases.

VI. CONCLUSION

This paper proposes a novel feature extraction and classification algorithm for robust sound event classification, motivated by the visual perception of sound through images. Our proposed subband power distribution (SPD) image improves upon our previous work on the spectrogram image feature (SIF), as it transforms the reliable signal components to a localised region of the image, making it simple to combine the extracted image feature with a missing feature classifier. We carried out experiments to validate our approach, using a comprehensive database of 50 sound event classes, across a range of noise conditions. Our results show that the SPD-IF is robust to noise, with a classification accuracy of over 90% in 0dB noise, and an overall average of almost 96%. This is improves on the conventional multi-conditional MFCC-HMM baseline, which achieved an average accuracy of 88%.

REFERENCES


Jonathan Dennis received the MEng in Engineering Science from the University of Oxford, U.K in 2007. He is currently persuing the Ph.D. degree in the School of Computer Engineering, Nanyang Technological University (NTU), Singapore. This is supported by the Institute for InfoComm Research (I2R), and the Agency for Science, Technology and Research (A*STAR) Graduate Academy, Singapore. His research interests include noise robust sound event recognition, pattern recognition and signal processing.

Huy Dat Tran received the M.Sc.Eng from the National Technical University of Ukraine (NTUU) in 1995 and Ph.D. degree from the National Academy of Sciences (NAS) of Ukraine in 2000, with both degrees specialized in acoustics. From 2000 to 2002 he did his postdoc research at the Institute of Hydromechanics, NAS of Ukraine. From 2002 to 2005 he had been a postdoc fellow at F.Itkura and K.Takeda Labs, Nagoya University, Japan. Since 2005 he has been a Senior Research Fellow at the Institute for InfoComm Research, Singapore. His research interests include acoustic and speech signal processing, sound recognition, and statistical methods for pattern recognition. He has participated and led a number of industrial projects of acoustic and speech applications. He serves as an Editorial member of the Open Acoustic Journal and a regular reviewer for prestigious journals including IEEE Transaction on Audio, Speech and Language Processing, Speech Communication, Signal Processing Letters, Neurocomputing and Pattern Recognition Letters.