A Robust Sound Event Recognition Framework Under TV Playing Conditions

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

***Abstract*—In this paper, we address the problem of performing sound event recognition tasks in the presence of television playing in a home environment. Our proposed framework consist of two modules: (1) a novel regression-based noise cancellation (RNC), a preprocessing which utilises a addition reference microphone placed near the television to reduce the noise. RNC learns an empirical mapping instead of the convention adaptive methods to achieve better noise reduction. (2) An improved subband power distribution image feature (iSPD-IF) which build on our existing classification framework by enhancing the feature extraction. A comprehensive experiment is carried out on our recorded data, which demonstrates high classification accuracy under severe television noise.**

I. Introduction

Sound event recognition (SER) is the task of understanding real-life events using sound information. This has a wide range of important applications in home environments, such as safety surveillance [1], [2] and home automation [3]. However, in home environments, listening to the radio and watching TV are frequent daily activities which can severely affect the performance of a sound event recognition system. The noise produced by these activities are different from stationary noise which has a fixed or slowly changing statistics. Examples of stationary noise include sounds generated from washing machine, air-con or vacuum cleaner. Conventional noise robust techniques for SER are mostly used to address such stationary noise. On the other hand, the noise produced by listening to the radio and watching TV is highly non-stationary and is regarded as a interference signal itself.

In this paper, we propose a dual microphone solution which utilize an additional microphone to effectively reduce non-stationary interferences, such as TV signals, and achieve high classification accuracy under severe noisy conditions. The proposed system consists of two modules: (1) the regression-based noise cancellation (RNC), a preprocessing for the television noise (2) A robust improved subband power distribution image feature (iSPD-IF) classification framework.

Our first novelty, the RNC, in contrast to conventional approaches using adaptive filters, such as least mean squares (LMS) [4], uses a short calibration to learn the empirical mapping from the reference to the classifying microphone. The advantage of our method is that it is free from the explicit assumptions of the statistical independence between noise and signal, also the performance of the system is not sensitive to the iterative updating parameters as in LMS.

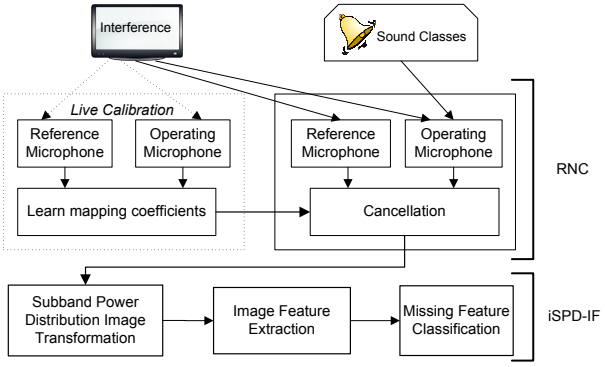

Fig. 1. Overview of proposed system.

The second novelty of our system is the iSPD-IF which is an extension of our robust classification framework, SPD-IF, proposed in [5]. We have shown that in the SPD-IF framework, the noise and signal can be localised and separated, hence producing superior performance by using a missing feature classifier. In this paper, we improve the SPD-IF method by replacing the histogram representations by a non-parametric windows (NP-Windows) method [6] in the feature extraction, which better represents the distribution of strongly dependent signals.

The overview of our proposed system is illustrated in Fig 1, the system consist of two microphones: (1) the microphone to capture the sound events for SER and (2) the reference microphone, which should be located near to the interference (TV or radio). The first few seconds of the recording are used in a short calibration, which learns the mapping function between the reference and the operating microphone. This enables the interference signal to be mapped and cancelled from the operating microphone which is then transformed into the SPD domain. Finally, the image feature is extracted, followed by a missing feature classification.

The rest of this paper is as follows: Section II introduces our proposed RNC method and compares it to the conventional methods, Section III recaps our previous work on SPD-IF and introduces our proposed modifications. Section IV then describes the experiments used to validate our approach, before Section V concludes the work.

II. Regression Based Cancellation

In this section, we present our proposed regression-based noise cancellation (RNC). Similar to the conventional methods, we employ a dual microphone approach which includes a reference microphone, that should be located nearer to the noise source (e.g. Television) and an operating microphone, which is used for the recognition. The first step is to remove the delay, d, between the microphones by using the autocorrelation function:

$$d = \max_{\tau} E[x(t)r(t+\tau)] \tag{1}$$

where $x(t)$ is the signal from operating microphone and $r(t)$ is the signal from reference microphone in time domain. In the absence of any sound event signals and with only the interference playing, the power spectrum of operating and reference microphones are related in following expression:

$$P_X[t,k] \approx |H[t,k]|^2 P_R[t,k] + P_N[t,k] \tag{2}$$

where H is the unknown frequency response characterizing the relative transfer function between the signal power at the operating, P_X and reference, P_R, microphones. Symbols t and k denote the frame index and bin, respectively. P_N denotes the noise power. Due to the window effect, $H[t,k]$ is not a constant but more like a random variable distributed around its mean value. For each fixed frequency bin, k, the aim is to find a function G such that:

$$\arg\min_{G} \int_t ||P_X[t,k] - G(P_R[t,k])||^2 dt \tag{3}$$

where $||.||$ represents the L_2 Euclidean norm.

In general, the function G can be any one-to-one function of any form, analytic or non-analytic. From the physical model as seen in equation 2, a natural choice of a function is the linear model, i.e:

$$P_X[t,k] = c_1(k)P_R[t,k] + c_2(k) \tag{4}$$

By utilising a short calibration period where only the interference source is active, the mapping coefficients c_1 and c_2 is learnt in each subband. A closed form solution is derived as follows [7]:

$$[c_1(k) \quad c_2(k)]^\intercal = (\bar{R}^\intercal \bar{R})^{-1} \bar{R}^\intercal \bar{X} \tag{5}$$

where $\bar{R} = [ones(M,1)^\intercal [P_R[1,k]\, P_R[2,k] \dots P_R[M,k]]^\intercal]$ and $\bar{X} = [P_X[1,k]\, P_X[2,k] \dots P_X[M,k]]^\intercal$ and M is the number of total consecutive frames. Since the samples with higher power are usually more reliable, additional weights is given on these samples and modify the equation as follows:

$$[c_1(k) \quad c_2(k)]^\intercal = (\bar{R}^\intercal W R)^{-1} \bar{R}^\intercal W \bar{X} \tag{6}$$

where W is the weighting function based on the power of the reference. Once the mapping function is learned, the desired output signal power, S, is obtained by cancelling the mapped interference from the noisy signal at the operating microphone:

$$P_S(t,k) = max\left(P_X[t,k] - \{c_1(k)P_R[t,k] + c_2(k)\}, 0\right) \tag{7}$$

A floor value of zero for the signal power is added to prevent over subtraction. Now, since the phase information from the operating microphone does not suffers as significantly compared to the magnitude in the presence of noise. This means that the estimated output signal can be reconstructed in time domain using the phase of the observed signal with an inverse FFT(IFFT).

$$s(t) = IFFT(P_S(t,k)\varphi[X(t,k)]) \tag{8}$$

where $\varphi[X(t,k)]$ is the phase of the observed signal and $c_1(k), c_2(k)$ are the mapping coefficients. The resulting signal may be transformed to any domain for any feature extraction method.

III. Subband Power Distribution Image Feature Classification

The noise cancellation module provides a significant reduction of the TV interference. However, it does not completely solve the mismatch in the classification, which still remains in the form of residual noise. In this section, we present our improved Subband Power Distribution Image Feature (iSPD-IF) classification which combines sophisticated normalization, robust component extraction and missing feature classification, and is an extension of our previous SPD-IF framework [5].

A. Subband Power Distribution Image Algorithm

The basic idea of the SPD is to transform the spectrogram into a new image representation, where the signal and noise are better localized and separable. This is done by following steps:

1) Normalise the auditory spectrogram into a grey-scale image
2) Transform each subband power series into its distribution, stacking them together to form a new image representation
3) Enhance the above image by using "contrast stretching" [8].

More details of this module is found in our previous work [5]. In the second step, we recall that the SPD represents the distribution of power, $D(f,z)$, in each frequency subband of the normalised spectrogram over time. In our previous work, we estimated this distribution using a conventional histogram. However, we note that the conventional probability density estimation (PDF) methods, such as histogram or Parzen kernel [9], assume the observations to be discrete independent and identically distributed (i.i.d) samples. For speech and sound, due to the modulation effects, the signal powers in each subband are often strongly correlated.

In this paper, we propose to replace the histogram, by employing a novel density method called non-parametric windows (NP-Windows) [6]. This approach treats the input as an analytical signal, approximated by interpolations between each consecutive pair of observation.The summary to calculate the PDF using NP-Windows [6] consists of three main steps:

1) Calculate the polynomial coefficients for the signal samples

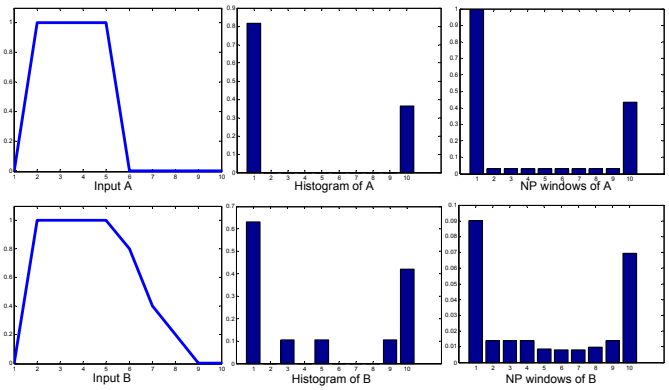

Fig. 2. The top row indicates input A and its density estimation using histogram and NP-Windows from left to right respectively. The bottom row indicates input B which is a reverberated version of signal B.

2) Calculate the PDF for each piecewise section, the signal is considered a function of a uniform random variable representing its domain
3) Populate the appropriate bins for each piecewise section.

To estimate the PDF in each subband, first we connect each adjacent data input with a straight line in the form $l_{f,i}(x) = a_{f,i}x + b_{f,i}$ where i represents the piecewise index and f represents the subband. For each piecewise straight line, we would assign a PDF, g, scaled to its gradient:

$$g_{f,i}(z) = \begin{cases} \frac{1}{|a_{f,i}|} & b_{f,i} \leq z \leq a_{f,i} + b_{f,i} \\ 0 & otherwise \end{cases} \quad (9)$$

Then, an arbitrary number of bins is chosen for the output histogram by summing up all the PDF that lies within the interval:

$$D(f, z) = \sum_i g_{f,i}(l_b \leq z \leq r_b) \quad (10)$$

where l_k and r_k are the respective left and right edges of a particular histogram bin, b.

With this formulation, we illustrate an example as seen in Fig.2 by considering a clean input A and its reverberated version, input B. It is seen that the histogram of signal A is mismatched with input B with many spikes. However by using NP-windows, the output distribution of both input A and B is smoothed and produces less mismatch between the clean and reverberated input.

B. Image Feature Extraction & Missing Feature Classification

After the improved SPD image using NP-windows is formed, an image feature based on the visual signature is extracted using spectrogram image feature (SIF) [10]. To extract the image feature, the two-dimensional SPD image is partitioned into 9x9 local sub-blocks, and then compute the image pixel distribution statistics. The image pixel distribution are inspired by the color layout which is described further in [10]. The final image feature is a 486 (2x3x9x9)-dimensional vector, using red, green and blue (RGB) quantisation regions with the second and third central moments to capture the distribution statistics.

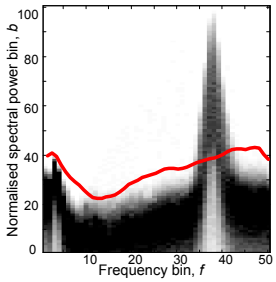

Fig. 3. An SPD image of a cancelled signal. The red line indicates the level noise estimate of the residual noise.

In Fig 3, it is seen that the noise masking effectively separates the residual noise and the signal into two regions. By choosing only a region containing the reliable parts, we then perform missing feature classification with kNN, using the Hellinger distance to measure the distribution distance between image features. Further details on the missing feature classification are found in our previous work [5].

IV. Experiments

In this section we carry out experiments to demonstrate the performance of our proposed system on a SER task.

Sound Database: For this, we select the following ten sound classes related to the home environment setting from the Real Word Computing Partnership (RWCP) Sound Scene Database [11]: horn, bells5, bottle1, phone4, whistle1, whistle3, clock2, ring, doorlock and trashbox. The sound files have a high SNR, and each contains an isolated sound, with some silence before and after the sound. For each class, 50 files are randomly selected for training and another 50 for testing. Overall, with 10 sound classes, this gives 500 clips for training and testing, with experiments repeated over 10 runs. For television noise, we recorded a half hour news segment from a local television channel.

Recordings: Both the sound classes and television signal are played back and recorded in our medium-sized lab (10m*4m*3m) with a reverberation time of approximately 400 milliseconds. They are played back through loudspeakers placed 4 meters apart in the middle of the room. The recording microphones are omni directional microphones (Shure-MX184): the operating microphone is placed randomly in between the speakers, while the reference microphone is placed next to the speaker playing back the television signal. The sampling rate of the recordings is 16000hz in 16bits resolution.

Experimental Methods: The following preprocessing methods are evaluated:

1) Proposed RNC method, based on subband power linear regression mapping.
2) Baseline FDAF method, based on state-of-the-art frequency domain adaptive filtering [12].

and the following classification methods are evaluated:

	MFCC-SVM	MFCC-GMM	MFCC-GMM-MVAN	MFCC-GMM-MVAN-Multi	SPD-IF	iSPD-IF
Accuracy	44.60	48.20	67.40	95.12	97.84	**98.52**

TABLE I
CLASSIFICATION RESULTS WITHOUT INTERFERENCE (%).

	MFCC-GMM-Multi			SPD-IF			iSPD-IF		
	-5db	0db	5db	-5db	0db	5db	-5db	0db	5db
No Processing	41.78	43.78	47.24	51.18	62.94	69.96	51.32	63.74	70.88
FDAF	49.80	50.64	50.88	87.64	90.06	91.20	90.54	91.50	92.54
RNC	61.56	65.44	66.00	89.38	92.80	94.98	**91.46**	**93.96**	**96.10**

TABLE II
CLASSIFICATION RESULTS WITH INTERFERENCE AT DIFFERENT SIR LEVELS (%).

1) SPD-IF, a robust sound classification framework introduced in [5]. The default parameters is the same as author's chosen parameters.
2) Proposed iSPD-IF, a extension of SPD-IF, using NP-Windows to improve the estimation of power densities.
3) MFCC-SVM: MFCC features modelled with one-against-one (OAO) SVM.
4) MFCC-GMM: MFCC features modelled with a GMM model
5) MFCC-GMM-MVAN: based on MFCC-GMM with mean, variance and arma normalisation (MVAN).
6) MFCC-GMM-MVAN-Multi: based on MFCC-GMM-MVAN with multi-conditional training using additive noise and also convoluting with random room impulse responses. The noise is added from random segments of the television signal at various noise levels.

For all various preprocessing and classification methods, frame lengths of 0.016s with frame shifts of 0.008s were used throughout. All MFCC features include deltas and delta-deltas, without the 0th coefficient and log energy to reduce mismatch due to loudness, giving a total of 36 dimensions. All methods using GMM are generated with 10 mixtures. For training, only the original clean samples from the CD are used in each classification method. For testing, each sound event is segmented using the ground truth time label for fair comparison, since the focus of the evaluation is solely on the classification and not the detection accuracy. Both training and testing are coded and evaluated using Matlab 2012b.

The performance for each method is reported in two tables: Table I - sound classes are recorded without interference, denoted as "Distant Microphone"; Table II - sound classes are recorded with TV playing simultaneously at a signal-to-interference ratio (SIR) of 5, 0 and -5 dB.

Results: Table I shows the results of the experiment conducted to evaluate the performance in the absence of television playing. This is an important step to find an upper bound performance for each of the classification method when compared to playing in the presence of television. Even though no interference is present, the sound events are played in various positions in the room and resulted the recorded test samples to be mismatched with the clean training due to the room's impulse response. This is reflected in the result for MFCC-SVM and MFCC-GMM which performs badly at 44.60% and 48.20% respectively.

Using MFCC-GMM-MVAN, MFCC-GMM with mean, variance and ARMA normalisation (MVAN) performed better at 67.40%, compensated slightly for the mismatch. However, the performance without interference was expected to perform at least 90% for a good lower bound for the experiment with interference. A multi-conditional training is added to MFCC-GMM-MVAN (MFCC-GMM-MVAN-Multi) where the training uses additive noise and convoluting with random room impulse responses performs much better at 95.12%. The multi-conditional training is effective in reducing the mismatch caused by the room's impulse response.

The proposed classifier iSPD-IF and the original SPD-IF both achieve a good baseline performance in the mismatched condition, with accuracies more than MFCC-GMM-MVAN-Multi. Also, even though the SPD-IF was not tested on convolution noise in our previous work, from this result we find that both the SPD-IF (97.84%) and iSPD-IF (98.52%) perform well with the recordings from the distant microphone. With these findings, only the top three feature-classifier combination with scores over 95% are used to evaluate in the next section, where interference noise is present. We will find out if the accuracy suffer and if so, which preprocessing methods will be most effective.

Next we discuss about the performance when the sound classes are recorded with interference in different SIR levels as shown in Table II. It is seen that the accuracy of each method reduces significantly when no pre-processing has been done before classification. However, we observe that that both SPD-IF and iSPD-IF outperforms the conventional MFCC-GMM-Multi method at all SIR levels. This is because SPD-IF and iSPD-IF are able to mask out some of the non-stationary noise from the interference. However, it cannot cope completely with the highly non-stationary TV noise.

The last two rows of Table II show the classification results after pre-processing the signals with the FDAF and RNC

methods. It is clearly seen that using the additional reference channel and applying pre-processing improves the results in all cases. The most significant result is that our proposed RNC outperforms FDAF for all three classification methods at all SIR levels. Another important observation is that the improvements for MFCC-GMM-Multi method are much less than for both SPD-IF methods. This is because the residual noise left from the preprocessing are effectively masked off in the both SPD-IF methods. However for MFCC-GMM-Multi method, the MFCC features are sensitive to the slight change caused by the residual noise, this causes mismatch and thus the less relative improvement.

Finally comparing our proposed iSPD-IF and the original SPD-IF in both Tables I and II, we observe that our proposed iSPD-IF consistently outperforms our original SPD-IF , with between 1-4% improvement. This is due to the use of NP-Windows to estimate the SPD, which reduces mismatch under reverberant conditions.

Overall, our proposed RNC preprocessing combined with the iSPD-IF gives a very good accuracy of above 90% for all SIR cases. In particular at 5db, the average accuracy was 96.1% which is less than 2.5% difference to the clean baseline. In addition, a significant advantage of the SPD-IF method, as compared to the conventional multi-conditional training, is the simplicity of using only the original clean signals for training which removes the possibility of room mismatch occurring, while maintaining a superior performance in the classification accuracy.

V. Conclusion

We propose a sophisticated sound event recognition framework in the presence of interferences such as TV, radio or music playing. The proposed method is a combination of two novel modules: the regression-based noise cancellation (RNC) and the improved subband power distribution image feature (iSPD-IF) for classification. The noise cancellation is able to greatly reduce the non-stationary interference, while the novel classification method is designed to transform the signal to a novel representation where the signal is separable from the residual noise. The proposed method has shown a significant improvement in a realistic noisy environment using only clean training. Particularly, we achieved more than 96% classification accuracy of ten sound classes under 5dB TV interference.

References

[1] Héctor Lozano, Inmaculada Hernáez, Artzai Picón, Javier Camarena, and Eva Navas, "Audio classification techniques in home environments for elderly/dependant people," in *Computers Helping People with Special Needs*, pp. 320–323. Springer, 2010.

[2] G Virone, D Istrate, M Vacher, N Noury, JF Serignat, and J Demongeot, "First steps in data fusion between a multichannel audio acquisition and an information system for home healthcare," in *Engineering in Medicine and Biology Society, 2003. Proceedings of the 25th Annual International Conference of the IEEE*. IEEE, 2003, vol. 2, pp. 1364–1367.

[3] J.C. Wang, H.P. Lee, J.F. Wang, and C.B. Lin, "Robust environmental sound recognition for home automation," *Automation Science and Engineering, IEEE Transactions on*, vol. 5, no. 1, pp. 25–31, 2008.

[4] S. Haykin, "Adaptive filter theory (ise)," 2003.

[5] J. Dennis, T. Dat, and E. Chng, "Image feature representation of the subband power distribution for robust sound event classification," *Audio, Speech, and Language Processing, IEEE Transactions on*, 2012.

[6] T. Kadir and M. Brady, "Non-parametric estimation of probability distributions from sampled signals," Tech. Rep., Technical report, OUEL, 2005.

[7] S.M. Kay, "Fundamentals of statistical signal processing, volume i: Estimation theory (v. 1)," 1993.

[8] R.C. Gonzalez and E. Richard, "Woods, digital image processing," 2002.

[9] E. Parzen, "On estimation of a probability density function and mode," *The annals of mathematical statistics*, vol. 33, no. 3, pp. 1065–1076, 1962.

[10] Jonathan Dennis, Huy Dat Tran, and Haizhou Li, "Spectrogram Image Feature for Sound Event Classification in Mismatched Conditions," *IEEE Signal Processing Letters*, vol. 18, no. 2, pp. 130–133, Feb. 2011.

[11] S. Nakamura, et al., "Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition," in *Proc. ICLRE*, 2000, pp. 965–968.

[12] J.J. Shynk, "Frequency-domain and multirate adaptive filtering," *Signal Processing Magazine, IEEE*, vol. 9, no. 1, pp. 14–37, 1992.