INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Data Augmentation, Missing Feature Mask and Kernel Classification for Through-The-Wall Acoustic Surveillance

Huy Dat Tran, Wen Zheng Terence Ng, Yi Ren Leng

Institute for Infocomm Research A*STAR, Singapore

hdtran,wztng,yrleng@i2r.a-star.edu.sg

Abstract

This paper deals with sound event classification from poor quality signals in the context of "through-the-wall" (TTW) surveillance. The task is extremely challenging due to the high level of distortion and attenuation caused by complex sound propagation and modulation effect from signal acquisition. Another problem, facing in TTW surveillance, is the lack of comprehensive training data as the recording is much more complicated than conventional approaches using audio microphones. To address that challenge, we employ a recurrent neural network, particularly the Long Short-Term Memory (LSTM) encoder, to transform conventional clean and noisy audio signals into TTW signals to augment additional training data. Furthermore, a novel missing feature mask kernel classification is developed to optimize the classification accuracy of TTW sound event classification. Particularly, Wasserstein distance is calculated from reliable intersection regions between pair-wise sound image representations and embedded into a probabilistic distance Support Vector Machine (SVM) kernel to optimize the TTW data separation. The proposed missing feature mask kernel allows effective training with inhomogeneously distorted data and the experimental results show promising results on TTW audio recordings, outperforming several state-of-art methods.

Index Terms: Though-the-wall, acoustic surveillance, sound event classification, missing feature mask, kernel classification.

1. Introduction

Audio technologies have been reaching a new phase where applications are going beyond the controlled recording conditions. The new generation of robotic and Internet-of-Things (IoT) devices comes with massive deployments of low-cost microphones hence the problem of poor quality signals has to be addressed by algorithm developments.

This paper discusses one of the most challenging tasks dealing with poor quality audio signals: through-the-wall (TTW) acoustic surveillances [1] which aims to remotely assess the sound information inside an enclosed space. It has been identified as a major area to be developed by the US National Institute of Justice (NIJ) and Defence Advanced Project Research Agency (DARPA)[2] for military and security operations. Examples of applications are hostage rescue, building clearance and area search where situational awareness is critical. Currently, most TTW acoustic surveillance works are focused on sensor designs [3]-[4]-[5] and signal enhancement [6]-[7]. In our work, we move forward to the task of understanding sound events through TTW recordings. In particular, this paper addresses the sound event classification problem in two modes of recording: wall-attached [3] and laser [5] microphones.

Unlike conventional audio recording, TTW acoustic sensors capture sound information through generated vibrations on the surface from the external surface. This makes the output signals extremely sensitive to noise and distortion. Another practical problem of TTW sound event classification is the difficulty in collecting training data. There are no standardized recording devices and recording characteristics vary among the available devices hence we must consider using small sets of data recorded on a single device together with data augmentation.

To solve this problem, we introduce a recurrent neural network (LSTM) encoder to learn an empirical transformation from clean and noisy conventional audio signals to TTW signals and use that to generate additional training samples for the classification. Furthermore, we develop a missing feature kernel classification method to optimize the classification performances on the given inhomogeneous and distorted data. The key point here is the use of intersection of reliable regions from 2D-representations of sound in the kernel calculations, enabling training with naturally inhomogeneous data. Another novelty of the method is the employment of Wasserstein distance in the probabilistic kernel design which shows better results with non-Gaussian inputs. The experimental results are promising results and outperformed several state-of-art methods. Fig.1 illustrates the overview of processing components of the proposed method.

The organization of the rest of the paper is as follows. In Sec. 2, we describe the procedure for augmenting training data with a LSTM encoder. Next, in Sec. 3, we will introduce the missing feature mask. In Sec.4, we will give details on the Wasserstein distance kernel classification with missing feature mask kernel. In Sec.5, we report experimental evaluations of our proposed method and a series of state-of-art methods. Finally, Sec. 6 concludes the work and adds some remarks on the future development of the topic.

2. Data augmentation

The first issue we face is the huge mismatch between clean and recorded samples due to the effects of attenuations, distortions and (de)-modulations. Training models trained on clean samples get totally meaningless results on actual recordings. The lack of standards on devices is also a big problem as the signal quality varies greatly with recording systems, making conventional multi-condition training ineffective. A logical solution for the practical operation of TTW acoustic surveillance systems could be a fast calibration/training session on each actual recording device.

In this paper, we propose a data augmentation approach which uses small amounts of actual TTW recordings to generate simulated training data from conventional audio signals. Unlike conventional audio signals, where noisy data can be simulated by convolving clean signals to an acoustic impulse response and then adding noise samples, TTW audio is not easy to model due to complex sound propagation, transformation and modulation effects. To solve this problem, we employ a recur-

http://dx.doi.org/10.21437/Interspeech.2017-685

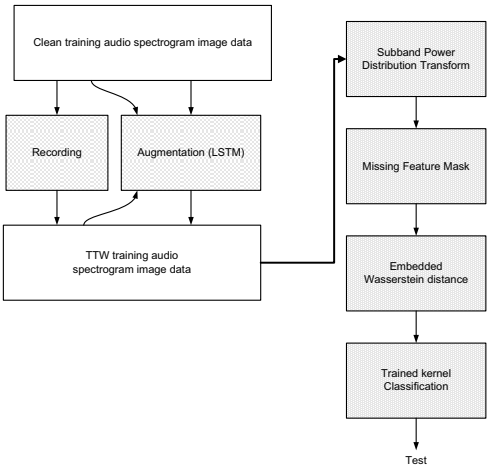

Figure 1: *Overview of proposed system.*

rent neural network, particularly, a Long Short-Term Memory (LSTM) regression, as an encoder, to learn the transformation from clean and noise spectrograms, obtained from conventional audio recordings, to the TTW spectrogram.

The left side of Fig.1 illustrates the data augmentation process. First, a small set of TTW audio is played back and recorded. Non-even TTW audio is also recorded as background noises. Both clean and randomly segmented noise spectrograms serve as input to the LSTM regression, having referenced TTW audio spectrogram as its output. Once the encoder is trained, other clean samples will be passed through the encoder to generate a more comprehensive training data set. To avoid overfitting bias, testing samples are recorded in different locations from training. The network regression consists of one LSTM layer (input 2x513, output 513 dimension) followed by three feed forward layers (513 dimension). ReLU activation function was applied on the first two FF layers while the last layer was linear. Min square errors objective function was used in training the encoder. Window length of 1024 and shift of 512 were used in FFT step. That configuration was used in speech denoising [8].

3. Missing Feature Mask

In this section, we present Subband Power Distribution (SPD) missing feature mask. The basic idea of SPD is to transform sound spectrograms into a new image representation, where the signal region is localized and hence easy to be separated [9].

The signal processing for SPD consists of three steps. First, the auditory spectrogram is normalized into a gray-scale image, noted by

$$G(f,t) = \frac{\log S(f,t) - \min\left(\log S(f,t)\right)}{\max\left(\log S(f,t)\right)} \tag{1}$$

where f represents the center frequencies of the filters, t is the time index. Second, the power values in each subband is transformed into its empirical distribution, for example using Pazen window, denoted as:

$$H(f,z) = \frac{1}{T}\sum_{t=1}^{T} K_h(z - G(z,t)) \tag{2}$$

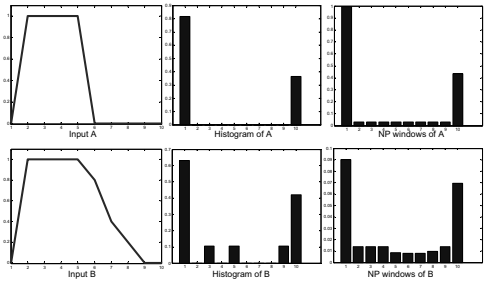

Figure 2: *The top row indicates input A and its density estimation using histogram and NP-Windows from left to right respectively. The bottom row indicates input B which is a distorted version of signal.*

where T is the number of frames and K_h is the kernel function. The matrix $H(f,z)$ therefore represents empirical distribution of spectrogram over time and frequency and is bounded in the range $0 \leq H(k,b) \leq 1$. Finally, contrast stretching is employed to enhance the contrast of the SPD image [9]. An improved version of SPD, which overcomes the weakness of statistical independence assumption of subband powers, employs non-parametric windows (NW) [10]. This method, in contrast to Pazen windows, treats each subband power series as an analytical signal. To estimate the NW-PDF in each subband, first we connect each adjacent data input to form a straight line: $l_{f,i}(x) = a_{f,i}x + b_{f,i}$ where i represents the piecewise index and k represents the subband. For each piecewise straight line, a NW-PDF, g, is assigned scaled to its gradient:

$$g_{f,i}(z) = \begin{cases} \frac{1}{|a_{f,i}|} & b_{f,i} \leq z \leq a_{f,i} \\ 0 & otherwise \end{cases} \tag{3}$$

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

$$D(f,z) = \sum_{i} g_{f,i}(l_b \leq z \leq r_b) \tag{4}$$

where l_b and r_b are the respective left and right edges of a particular histogram bin, b. Figure 2 illustrates an example of different density estimation of a clean input and its distorted version. It suggests that the NW estimations are closer and hence less mismatch than that of conventional histogram estimation.

Thanks to the resonating nature of sound signals, which translates into the sparse form of spectrograms, it can be expected that SPD transforms sound harmonic components into peaking patches in the right side of SPD image representation and that is isolated from noise patches which lie more to the left side. With an increasing amount of noise, the noise patch boundary will shift towards the right but the peaking patches representing the resonating components of sounds are still separable. That can be seen in Fig. 3, which illustrates the sound spectrum and its SPD representation without and with noise (at 0dB SNR). The region to the right of the yellow dotted line are similar for the noisy SPD and its clean counterpart in Fig. 3. This provides the motivation to find the boundaries of the noise mask, in order to perform classification using only the reliable area. To estimate the separation point in each frequency bin, we first utilize a SPD image from a segment containing only noise, denoted as $I_N(f,z)$. The upper bound of the distribution in $I_N(f,z)$ forms an estimate of the noise boundary, $n_{max}(f)$,

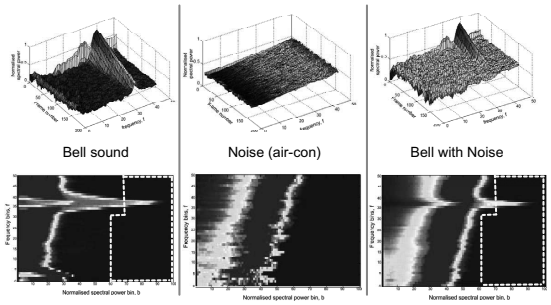

Figure 3: *Examples to illustrate the SPD representation of clean signal, noise and noisy, respectively. Top row represents the spectrogram surface and bottom row represents the SPD representation (bins-frequency). The dotted yellow lines show the reliable part of SPDs.*

in the SPD and is estimated as the maximum occupied bin for each frequency subband:

$$n_{max}(f) = \underset{z}{\operatorname{argmax}}(I_N(f,z) > 0) \tag{5}$$

Given an actual noisy segment/clip, we need to estimate the actual noise boundary from Eq 5. This can be done with the assumption that the noise subband power distribution vaaries less than its intensity. Accordingly, the subband noise intensity variation can be found by maximizing the cross-correlation between $I_N(f,z)$ and $I(f,z)$ in each subband, noted as

$$a_{max} = \max_{a} \left[I(f,z) \star I_N(f,z+a) \right] \quad , \quad \forall f \tag{6}$$

The final noise mask estimate, $n(k)$ is then simply derived as:

$$n(f) = n_{max}(f) + a_{max}. \tag{7}$$

4. Missing Feature Kernel Classification

As discussed above, a big problem facing in training of TTW sound event classification is the large variation of distortions in recording. To deal with that, we propose a novel method to integrate missing feature masks into classification systems.

4.1. Kernel classification

Starting with linear SVM, considering the problem of designing a separating hyperplane for m vectors $\mathbf{x}_i \in \mathbf{R^n}\ i = 1, 2, ..., m$. Each point $\mathbf{x}_i \in \mathbf{R^n}$ belongs to one of two classes, by its label $y_i \in \{1, -1\}\ , i = 1, 2, ..., m$. The goal of linear support vector machines is to find an optimal separating hyperplane $f(\mathbf{x}) = \mathbf{w^T x} + b$, which maximizes the margin, i.e. $\frac{2}{\|\mathbf{w}\|^2}$,or equivalently minimizes $\|\mathbf{w}\|^2$. For the solution, the non-negative variable ξ_i is introduced so that the soft margin can be found by quadratic programming:

$$\begin{aligned} &\min_{(\mathbf{w},\mathbf{b},\xi)} \left(\tfrac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{n}\xi_i \right) \\ &\text{s.t. } y_i\left(\mathbf{w}^T\mathbf{x}_i + b\right) \geq 1 - \xi_i\ ; i = 1, 2, ..., n \\ &\xi_i \geq 0, \end{aligned} \tag{8}$$

where the term $\sum_{i=1}^{n}\xi_i$ denotes the upper bound of the misclassification from the training samples and C is a regularization coefficient. There are several ways to solve this optimization problem, all return the form of separating hyperplane which can be generalized into a kernel form as follows:

$$f(\mathbf{x}) = \sum_{i=1}^{m} \alpha_i K(\mathbf{x}, \mathbf{x}_i) + b. \tag{9}$$

In the implementation, all the optimization methods lead to the calculation of the kernel matrix $\mathbf{K}(\mathbf{x}_i, \mathbf{x}_j)$ [11].

4.2. Missing feature mask kernel classification

The key idea of this paper is to integrate missing feature mask into kernel machine to optimize the training with missing feature samples. It is done by introducing a geometrical intersection area of reliable parts from each pair of samples in the kernel calculation noted as

$$\mathrm{K}(\mathbf{x}_i, \mathbf{x}_j) \rightarrow \mathrm{K}\left(\mathbf{x}_i^{R_i \cap R_j}, \mathbf{x}_j^{R_i \cap R_j}\right), \tag{10}$$

where R_i and R_j denote the reliable parts of sample $\mathbf{x}_\mathrm{i}$ and $\mathbf{x}_\mathrm{j}$, respectively. Similarly, the testing will be carried out with the use of intersection areas between testing and training SPDs:

$$f(\mathbf{x}) = \sum_{i=1}^{m} \alpha_i K\left(\mathbf{x}^{R \cap R_j}, \mathbf{x}_i^{R \cap R_j}\right) + b. \tag{11}$$

First, SPD image is partitioned into 9x9 local sub-blocks. Then missing feature mask is applied. For each pair of sound images, the intersection between reliable parts are indexed using the above partition. The missing feature kernel (MFK) is calculated as the distance between images' pixel distributions from the intersection area, following by Eq 10. Compared to our preliminary version presented in [18], the Wasserstein distance [19] is used instead of Hellinger distance and performed better in the experiments. We think the Wasserstein distance could better measure distance for non-Gaussian distributions by going beyond the calculations for second moments. It is noted by

$$\mathbf{W}_p(\mu, \vartheta) = \inf \mathbf{E}\left([d(X,Y))\right]^p, \tag{12}$$

where (μ, ϑ) denote cluster point sets of samples (X, Y). Particularly, we use $p = 3$ in our experiments.

5. Experiments

5.1. Task

The proposed methods are tested for the TTW acoustic surveillance task. Two setups of wall-attached microphone (WAM) and laser acoustic microphone (LAM) are investigated.

5.2. Sound Database

The original sound samples includes clips from normal speech and 4 aggressive events: crying, gunshot, explosion and banging [12]. These sound classes are chosen to align with an industrial collaboration in security surveillance. The original clips are sampled at 44kHz and manually balanced in advance with lengths varying from 2 to 5 seconds. For each sound class, 500 samples are prepared for experiments. The clips are split into two sets: 300 clips for training and 200 clips for testing. To reduce overfitting, training and testing recordings were carried out in two different rooms: one in an office and the other in a housing estate. As a reference, the dataset achieved 95.60% accuracy in the clean condition. To simulate practical operational situations, only 100 clips were played back and recorded as noisy signals, the rest of the 200 clean training samples were encoded using the trained LSTM network described in Sec. 2 to simulate the noisy signals.

Table 1: *Classification accuracy results (%).*

	MFCC-GMM	NMF-SVM	CNN	SPD-MFK	SPD-MFK-old
WAM	38.60	72.30	77.10	**82.50**	**79.80**
LAM	33.80	69.30	72.70	**76.80**	**75.40**

Table 2: *Confusion matrix from the* **SPD-MFK** *method (samples).*

	normal speech	crying	gunshot	explosions	banging
normal speech	149	46	0	0	5
crying	32	162	0	3	3
gunshot	0	0	140	41	19
explosions	0	0	12	184	4
banging	0	0	0	10	190

5.3. Recordings

The original sound clips were played inside enclosed rooms. For the WAM task, a cold gold piezo contact microphone [13] is attached outside on a cement and brick wall of estimated 20 cm thickness. For the LAM, a DIY set-up following [14] is adopted, i.e. a laser source is pointed to a glass window to capture the vibrations which carry sound information inside the rooms. In both cases, signals were recorded at 16kHz sampling rate. The recorded signals are rather weak with severe noises, distortions and attenuations. The signals' estimated SNR [15] is ranging from -3dB to 6dB.

5.4. Experimental Methods

Several state-of-art methods were evaluated together with the proposed method on TTW audio signals.

1. **MFCC-GMM**: conventional 39-dimension MFCC feature with optimized 16-components GMM classifier.
2. **NMF-SVM**: Non-negative matrix factorization (NMF) is trained with all available training data to get the code books dictionary. Each sample is decomposed into that and the weights are then used as features with a linear SVM classifier [16].
3. **CNN**: Applying conventional image classification on spectrogram images [17]: 5 convolutional layers with 5x5 windows, 16,32,64,128,5 outputs, and a final softmax layer.
4. **SPD-MFK**: proposed SPD missing feature mask kernel classification using data augmentation for training

For all various preprocessing and classification methods, frame length of 1024 with frame shifts of 256 were used throughout. All of the above methods use 100 recorded samples and 200 simulated samples in training and other 200 samples in testing. Training and testing samples are recorded in different buildings. To verify the effectiveness of the data augmentation approach, we also evaluate the proposed method with only 100 recorded samples and name that as **SPD-MFK-orig**. Preliminary results of **SPD-MFK-orig** was presented in [18]. Compared to the earlier work, data augmentation was applied to improve the training and Wasserstain distance is employed in the kernel classification, improving the classification accuracy. Data recorded with laser acoustic microphone (LAM) has also been added into the investigation.

5.5. Results and discussions

Table 1 summarizes the classification accuracies from the evaluated methods. As expected, conventional MFCC-GMM yielded a poorest result showing its not suitability for the task. It seems that the high level of distortions transforms MFCC into wrong components and the GMM is unable to handle the mismatch between samples. NMF-SVM learns the code books through training and hence could project signals into a better dictionary for good results. However the normal SVM classifier is unable to handle the mismatch between training and testing conditions. The spectrogram image CNN is presently the most popular method in the literature due its universality and simplicity in evaluation. The method achieved promising results on TTW audio even without large training datasets. However, as a common standard of TTW recording is absent and collecting large datasets on each recording device is impractical, the use of deep learning for the task may not be an advantageous. The proposed SPD mask with missing feature kernel classification performs the best, achieving 82.57% and 79.83% on average for classification accuracy. This is expected as the proposed method is designed specially to address the problems of TTW audio signals. The intersection missing feature mask kernel is the key to boost the classification accuracy as it manages to filter out the non-reliable parts of spectrogram which affects the classification accuracy. The data augmentation is promising as it helped to bring up the accuracy compared to the **SPD-MKF-orig**. It is interesting to note that the improvement was greater with LAM recording which is the more challenging task. Although the LSTM encoder is unlikely to be optimized with just 100 samples of training, it could add more variations of "closed" samples into training which appears to help. The confusion matrix from the best method is shown in Table 2. Very good results were obtained with banging and explosion classes, both yields higher than 90% classification accuracy.

6. Conclusions

We propose a novel missing feature kernel machine classification method specially designed for sound event classification task with low quality signals. The method combines data augmentation in training and missing feature mask kernel classification to enable reasonable classification accuracy under challenging through-the-wall (TTW) surveillance conditions.

7. References

[1] David D. Ferris, Jr. and Nicholas C. Currie "Survey of current technologies for through-the-wall surveillance (TWS)", Proc. SPIE 3577, Sensors, C3I, Information, and Training Technologies for Law Enforcement, 62 (January 7, 1999); doi:10.1117/12.336988

[2] Christopher A. Miles "Through-the-Wall Surveillance: A New Technology for Saving Lives", NIJ Journal Issue No. 258, October 2007

[3] Felber, F. (2015, May). Demonstration of novel high-power acoustic through-the-wall sensor. In SPIE Defense+ Security (pp. 945603-945603). International Society for Optics and Photonics

[4] Abe Davis and Michael Rubinstein and Neal Wadhwa and Gautham Mysore and Fredo Durand and William T. Freeman "The Visual Microphone: Passive Recovery of Sound from Video", ACM Transactions on Graphics, 2014, 33, 14, 79:1-79:10

[5] Muscatell Ralph P, US patent US4412105: Laser microphone, Oct 1983

[6] Renhua Peng, Chengshi Zheng and Xiaodong Li, "Bandwidth extension for speech acquired by laser Doppler vibrometer with an auxiliary microphone", Information Communications and Signal Processing (ICICS) 2015 10th International Conference on, pp. 1-4, 2015

[7] Yekutiel Avargel, and Israel Cohen, "Speech measurements using a laser Doppler vibrometer sensor: Application to speech enhancement", in proc. 2011 Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA), 2011

[8] Weninger F., Erdogan H., Watanabe S., Vincent E., Hershey J., Schuller B.,"Speech Enhancement with LSTM Recurrent Neural Networks and its Applications to Noise Robust ASR", in Proc. of 12th International Conference on Latent Variable Analysis and Signal Separation, 2015, pp. 91-99

[9] J. Dennis, T. H. Dat, and E. Chng, "Image feature representation of the subband power distribution for robust sound event classification," IEEE Transactions on Audio, Speech, and Language Processing, pp. 367 - 377, 2012

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

[11] T. H. Dat and H. Li "Sound Event Recognition With Probabilistic Distance SVMs,", IEEE Transactions on Audio, Speech, and Language Processing, , 1556-1568, 2011

[12] Sound Effect Collections, http://www.sound-ideas.com/

[13] Cold Gold contact microphone http://www.contactmicrophones.com/

[14] Laser microphone http://www.williamson-labs.com/lasermic.htm/

[15] T. H .Dat, K. Takeda, F. Itakura, "On-line Gaussian mixture modeling in the log-power domain for signal-to-noise ratio estimation and speech enhancement," Speech Communication, 1515-1527, 2006.

[16] Tatsuya K.,Yuzo S., Reishi K., "Acoustic Event Detection based on Non-negative Matrix Factorization of Local Dictionaries and activation aggregation", in Proc. ICASSP 2016, pp.2259-2262.

[17] Hamid E-Z.,Berhard L., Matthias D., Gerhard W., " CP-JKU Submissions for DCASE-2016: A Hybrid Approach Using Binaural I-Vectors and Deep Convolutional Neural Networks", in Proc. of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016).

[18] Tran Huy Dat, Jonathan William Dennis, and Ng Wen Zheng Terence, " Missing Feature Kernel and Nonparametric Window Subband Power Distribution for Robust Sound Event Classification,", in Proc. Speech and Computer (SPECOM), 2015, pp. 277-284.

[19] Ruschendorf, L. "Wassertein metric", in Hazewinkel, M. Encyclopaedia of Mathematics, 2001, ISBN 978-1-55608-010-4