ADAPTIVE LEARNING WITH SURROGATE ASSISTED TRAINING MODELS USING LIMITED LABELED ACOUSTIC SAMPLE SEQUENCES

G. Zucatelli and R. Coelho

Laboratory of Acoustic Signal Processing (lasp.ime.eb.br)
Military Institute of Engineering (IME), Rio de Janeiro, Brazil
(e-mail: zucatelli@ime.eb.br, coelho@ime.eb.br)

ABSTRACT

In this paper, an adaptive learning solution based on surrogate models is investigated under reverberant scenarios. A new surrogate selection criteria is proposed, leading to a higher discrimination among models. The method is evaluated considering a classic source classification approach with ROC and AUC analysis. Furthermore, the Bhattacharrya distance is adopted to measure the separability of selected signals in the feature domain, whereas the sparse coding capability of each selected model is evaluated with the K-SVD. Results show that the proposed solution improves classification accuracy and class separability while providing a reduction on sparse coding reconstruction error for all scenarios. Further experiments with pH feature vector fusion improved the classification accuracy of adaptive learning solutions.

1. INTRODUCTION

Reverberation is ubiquitous in urban acoustic environments. The acoustic reflection on walls and objects may alter target signals in indoor or outdoor scenarios. This condition changes characteristics of speech [1, 2, 3] and can also impact acoustic source classification [4, 5, 6]. The recognition of environmental sound has been a topic of great concern in the signal processing and machine learning research areas [7, 8, 9, 10]. Typical applications include surveillance, hearing aid, smart home and robot navigation. A key challenge for such systems is to select relevant observations among limited labeled acoustic samples that are able to represent the natural phenomena and than guarantee a robust classification under temporal and spectral distortions such as reverberation.

The sound propagation is usually described by the room impulse response (RIR) which is typically characterized by the reverberation time (T_{60}) and the direct-to-reverberant ratio (DRR). These parameters describe the effect duration and intensity relative to the direct signal, respectively. By selecting the most informative signals, active learning (AL) solutions [6, 10] can be good strategies to overcome inevitable mismatches caused by a diversity of acoustic environments. Furthermore, more informative signals lead to a better discrimination among acoustic sources.

The adaptive learning with surrogate assistance (ALSS) solution proposed in [6] is divided in two main steps: surrogate signals generation and selection. In this work, a new surrogate selection criteria is adopted to increase the discrimination capacity of acoustic sources under real reverberation scenarios, named the modified ALSS (ALSS$_{\text{mod}}$). The learning approach requires no human effort for unlabeled data and is implemented on reverberation free signals available as training data for acoustic models. The goal is to overcome the mismatch of tested reverberated signals by selecting more discriminative ones using only the information provided by the training data. Surrogates consider the Kurtosis ratio (K), the power spectral density (PSD) and the index of nonstationary (INS) [11] of labeled data to create new stochastic models.

The active learning experiments are conducted considering eight acoustic sources with different nonstationarity degree for four real reverberation scenarios. A classical classification procedure based on mel-frequency cepstral coefficients (MFCC) and Gaussian mixture models (GMM) is first adopted for evaluation. Results show an increment of 12.5 percentage points (p.p.) above the reverberation free scenario for a room with small T_{60}. For the most challenge reverberant condition, the ALSS$_{\text{mod}}$ increases the average classification rate and achieves values similar to the reverberation free environment. The ROC curve and the area under curve (AUC) are analyzed for all rooms. AUC increments are observed especially for the most nonstationary acoustic sources. Moreover, the Bhattacharrya distance (Bd) [12] is selected as a measure of separability between chosen signals in the MFCC domain. Selected matrices are also evaluated in terms of sparse coding reconstruction error for the dictionary learning technique K-SVD [13]. An increase in Bd and a reconstruction error reduction is assessed in all scenarios confirming the discrimination capacity of the AL solution. Finally, the pH feature vector [14][15] is also investigated for source classification, leading to a further 3.5 p.p. average accuracy improvement.

This work was partially supported by the National Council for Scientific and Technological Development (CNPq) 308155/2019-0 and Fundação de Amparo Pesquisa do Estado do Rio de Janeiro (FAPERJ) 203075/2016. This work is also supported by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Grant Code 001.

2. ADAPTIVE LEARNING WITH SURROGATE ASSISTANCE

The ALSS_{mod} technique is here presented considering the stages of generation and selection of surrogate signals. The first stage follows the steps of [6] and is succeeded by the proposed surrogate selection criteria. The main goal of ALSS_{mod} is to select an acoustic model λ_c that better discriminates a target class c among others and leads to a more robust scheme under acoustic distortions such as reverberation.

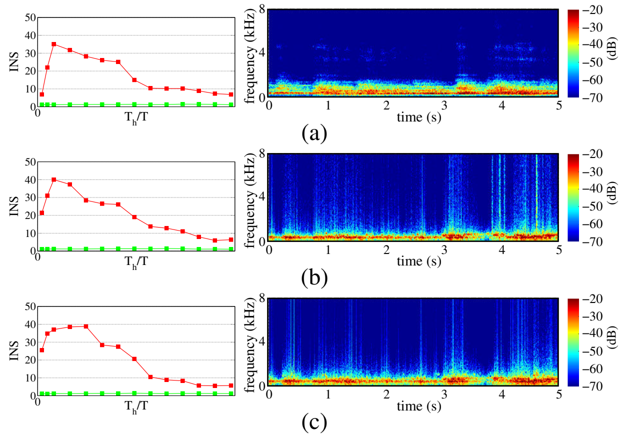
For that purpose, consider a set of training acoustic signals $\{\Phi_c^0 | c = 1, \dots, C\}$, one for each of C classes. A feature matrix Y_c^0 is extracted from Φ_c^0 and then used to obtain an initial acoustic model λ_c^0. The learning solution is implemented based on the generation of M nonstationary surrogate signals $\{\Psi_c^m | m = 1, \dots, M\}$ using the statistics of training samples. Then, a set of matrices Y_c^m is extracted from these surrogates, leading to a new set of acoustic models λ_c^m. The most informative acoustic model λ_c is finally obtained by comparing λ_c^m to the original model λ_c^0 depending on some criteria.

The surrogate generation works on a frame-by-frame basis. Given a reference signal $\{x(t)\}$, a division on Q short-time frames is performed with 50% overlapping. For each frame q, the algorithm is divided in three steps:

1. Generation of a random sequence of uncorrelated samples $\{y_q(t)\}$ with amplitude distribution defined by the Kurtosis ratio of $\{x_q(t)\}$ as in [16],
2. The $\{y_q(t)\}$ is passed through a finite impulse response (FIR) filter computed based on the target PSD [17, 18, 19] to obtain artificial samples $\{\bar{y}_q(t)\}$,
3. Short-time segments are adjusted depending on the target signal INS value and are than concatenated to form a single surrogate $\{y(t)\}$.

Possible PSD peaks are incorporated to Step 2 as in [6]. The PSD peak detection is performed on the basis of the sum of moving average with the standard deviation of L neighboring points multiplied by a factor F. Therefore, surrogate signals present short-time amplitude distribution, PSD decay and nonstationary behavior similar to a target signal $\{x(t)\}$. Fig. 1 depicts INS values and spectrograms from a real signal and two surrogates. These surrogates are obtained with parameters: $L = 16$ and $F = 1.6$ for Fig. 1 (b) and $L = 64$ and $F = 2.0$ for Fig. 1 (c). INS values are calculated considering scales of T_h/T where $T = 5$ s is the total duration and T_h is the signal length on analysis. The green dashed line represents the stationary threshold $\gamma \approx 1$. Note that the INS behavior is considerably similar for real signal and surrogates and the spectrograms energy are primarily concentrated at the same spectral regions.

A new surrogate selection criteria is proposed in this work for the ALSS_{mod}. Different from ALSS [6], in this paper the average of all classes classification rate R^{Γ} is considered for the learning process, where Γ stands for the number of models replaced at a given moment. Let R_c represent the percentage of correctly recognized trials for a class c and R_c^m the accuracy rate for a new model λ_c^m obtained from a surrogate m. A model replacement $(\Gamma + 1)$ occurs for λ_c^m, if λ_c^m simultaneously increases the average classification rate $(R^{\Gamma+1} > R^{\Gamma})$ and the model accuracy rate $(R_c^m > R_c)$. For this case:



(a)

(b)

(c)

Fig. 1. INS and spectrogram of (a) Dogs source and its surrogates (b)-(c) generated with different parameters L and F.

$$\lambda_c \leftarrow \lambda_c^{\hat{m}}, \quad \text{where } \hat{m} = \max_{1 \le m \le M} R^{\Gamma+1}. \tag{1}$$

The learning procedure is considered adaptive for a novel set of surrogate can be created whenever a new data set is available. Furthermore, the criteria adopted in ALSS_{mod} select the most discriminative models searching for the maximum average classification rate considering all surrogates.

3. SPARSE CODING FOR FEATURE MATRICES

The discrimination power of the learning solution can also be applied to sparse coding approaches. To this end, the K-SVD [13] dictionary learning is adopted considering feature matrices. Given a matrix of interest $\mathbf{Y}$, the sparse coding procedure aims to better express each column of $\mathbf{Y}$ as a linear combination of T_0 atoms of a dictionary $\mathbf{D}$. Therefore, the K-SVD objective function can be written as $\min_{\mathbf{D},\mathbf{X}} ||\mathbf{Y} - \mathbf{D}\mathbf{X}||_F^2$ subjected to $||\mathbf{x}_i|| \le T_0 \; \forall i$, where $\mathbf{x}_i$ is the ith column of $\mathbf{X}$. The K-SVD solves the minimization problem by updating each column of $\mathbf{D}$ and its relevant coefficients on $\mathbf{X}$ through a generalization of the k-means. In this work, a dictionary $\mathbf{D}_c$ is learned from each feature matrix after the surrogate selection. The idea is to reconstruct the reverberated signals feature matrices using the $\mathbf{D}_c$ vector space with the smallest possible reconstruction error.

4. EXPERIMENTS AND RESULTS

In order to evaluate the proposed ALSS_{mod} technique in reverberant scenarios, four RIRs at 22050 Hz were selected from the AIR [20] and LASP_RIR[1] databases. Rooms Meeting, LASP1, LASP2 and Stairway present T_{60} and DRR values of $\{0.36, 0.65, 0.79, 1.00\}$ and $\{2.7, -3.1, -4.3, -3.4\}$, respectively. The Meeting room holds the smallest T_{60} and

[1] Available at lasp.ime.eb.br

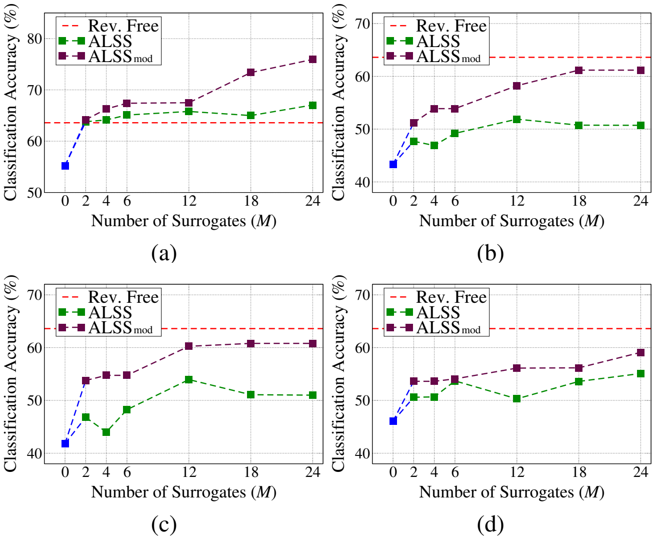


Fig. 2. The ALSS and ALSS$_{\text{mod}}$ classification accuracy with MFCC-GMM for rooms Meeting (a), LASP1 (b), LASP2 (c) and Stairway (d).

highest DRR values. On the other hand, the LASP2 and Stairway are the most challenging condition with highest T_{60} and lowest DRR. Experiments are conducted considering a total of eight acoustic sources[1]: Chainsaw, Dogs, Fan, Rain, Shower, Siren, Subway and Waterfall. Three different 5-second signals are adopted for each source for acoustic model generation, surrogate selection and reverberated tests. The learning process is implemented with 2, 4, 6, 12, 18 and 24 surrogates for $L \in [16, 64]$ and $F \in [0.6, 2.0]$ as in [6].

A multi-class classification experiment is first designed to assess the ALSS and ALSS$_{\text{mod}}$ improvement on average accuracy in reverberant scenarios. Feature matrices are composed of 12 MFCC vectors, extracted every 20 ms with 50% overlapping. A GMM model is implemented with five components. The classification is performed between each model and feature vector obtained from test signals according to the maximum likelihood criterion.

Fig. 2 illustrates both adaptive learning solutions for the four reverberant rooms. The blue squares represents the scheme without learning. As a reference, the classification accuracy for a reverberation free environment without any learning procedure is represented as a red dashed line. Note that reverberation decreases the accuracy values from 63.6 down to 41.8 for the LASP2 room. Because of its surrogate selection criteria, the ALSS$_{\text{mod}}$ progressive improves the classification for all conditions and outperforms the ALSS in most reverberant scenarios. The proposed method achieves an average accuracy around 60 for the three most challenging cases for 24 surrogates. Furthermore, the ALSS$_{\text{mod}}$ is able to improve the classification up to 76.1 for the Meeting room, which is an increment of 12.5 p.p. compared to a reverberant free environment and 9.5 p.p. higher than the ALSS accuracy. This result highlights the ability of the proposed technique to better discriminate audio classes under reverberation. The use of 24 surrogates suffices to reach the best result in most cases, therefore it is adopted in the remaining of this work.

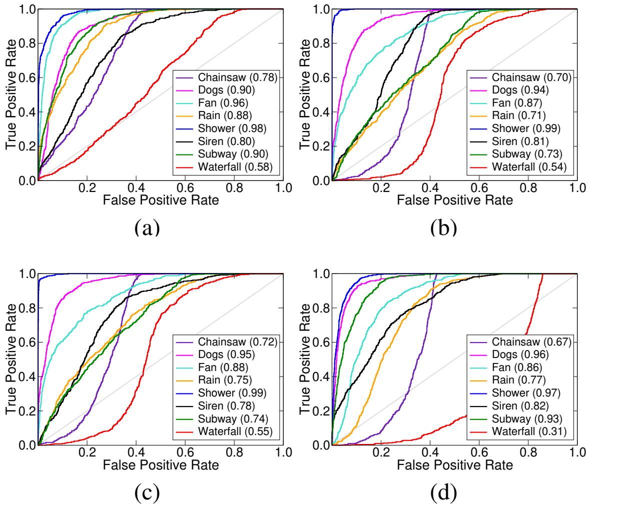


Fig. 3. ROC curves and AUC without learning for rooms Meeting (a), LASP$_1$ (b), LASP$_2$ (c) and Stairway (d).

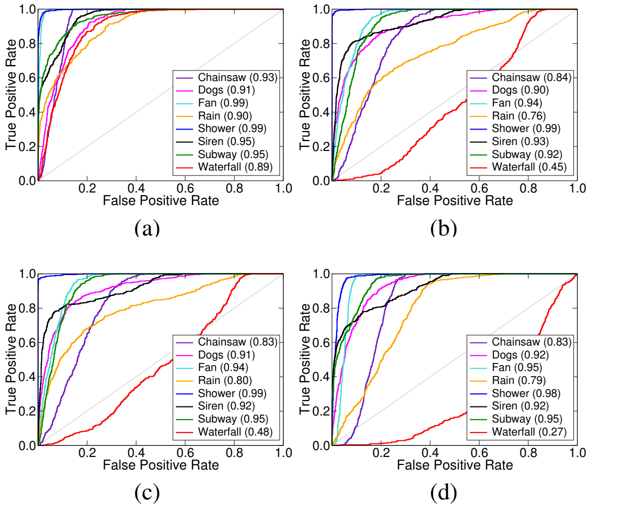


Fig. 4. ROC curves and AUC with ALSS$_{\text{mod}}$ for rooms Meeting (a), LASP$_1$ (b), LASP$_2$ (c) and Stairway (d).

Figs. 3 and 4 depict the true positive and false positive rates on a ROC curve for each source considering reverberated tests without any learning approach and with the ALSS$_{\text{mod}}$ solution, respectively. These rates are of great interest to evaluate the discrimination between sources. Moreover, the Area Under Curve (AUC) is also detailed for each class. The false positive rate for a particular class is measured considering test signals belonging to all the other classes. Note that the most nonstationary sources Siren and Chainsaw present relatively small AUC without the ALSS$_{\text{mod}}$ solution. The proposed technique improves the classification of all sources for the Meeting room. With the learning approach all AUC are greater than 0.88 and the average AUC increases from 0.85 up to 0.94. Considering the other rooms almost all AUC values are improved, with the exception of sources Dogs and Waterfall. As the most discriminative models are selected based on the overall classification rate, this configuration led

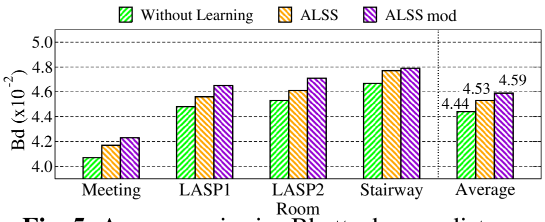


Fig. 5. Average pairwise Bhattacharrya distance.

to the best average result. For rooms LASP1, LASP2 and Stairway, the respective mean AUC of 0.79, 0.80 and 0.79 are increased to 0.84, 0.85 and 0.83 with ALSS$_{\text{mod}}$. It is important to notice that the AUC values for the most nonstationary sources (Chainsaw and Siren) significantly increases for all reverberant rooms. The highest AUC increment for the Chainsaw source is obtained on the Stairway room with values from 0.67 to 0.83. For the Siren source, the values varied from 0.80 up to 0.95 for the Meeting room.

As the ALSS$_{\text{mod}}$ aims to obtain the most discriminative model for a class, it is essential to examine this condition for all selected signals. The Bhattacharrya distance (Bd) is considered to assess the separability of classes in the MFCC domain for both learning techniques. Given different classes c_1 and c_2, the Bd for two Gaussian distribution is defined by

$$Bd = \frac{1}{2}\ln\frac{\frac{|\Sigma_1+\Sigma_2|}{2}}{|\Sigma_1|^{\frac{1}{2}}|\Sigma_2|^{\frac{1}{2}}} + \frac{1}{8}(\mu_1-\mu_2)^T\left(\frac{\Sigma_1+\Sigma_2}{2}\right)^{-1}(\mu_1-\mu_2), \quad (2)$$

where μ_c and Σ_c accounts for the mean vector and covariance matrix of class c, respectively. Figure 5 shows the total average distance computed pairwise and normalized across observations for each room. Learning methods increase the separability of classes in all scenarios. The highest distance gain among classes is observed for the LASP2 room with ΔBd= $0.18\text{x}10^{-2}$ for the ALSS$_{\text{mod}}$. LASP1 and Meeting rooms led to improvements of $0.17\text{x}10^{-2}$ and $0.16\text{x}10^{-2}$ for the ALSS$_{\text{mod}}$ and $0.07\text{x}10^{-2}$ and $0.11\text{x}10^{-2}$ for the ALSS. On average, the separability of classes raised from $4.44\text{x}10^{-2}$ to $4.56\text{x}10^{-2}$ with the ALSS$_{\text{mod}}$, while attaining $4.53\text{x}10^{-2}$ for the ALSS. Theses results reinforce the ability of the proposed technique to select discriminative signals.

Informative models can also be beneficial to sparse coding. In this work, the K-SVD algorithm generates sparse vectors using the surrogate MFCC matrices selected surrogates by ALSS$_{\text{mod}}$. For each acoustic source, the K-SVD is set to 80 iterations to learn a 12x12 dictionary $\mathbf{D}_c$. The OMP was applied with $T_0 = 6$ nonzero elements. Learned dictionaries are employed to reconstruct each test observation feature.

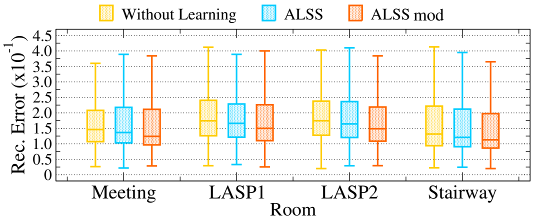
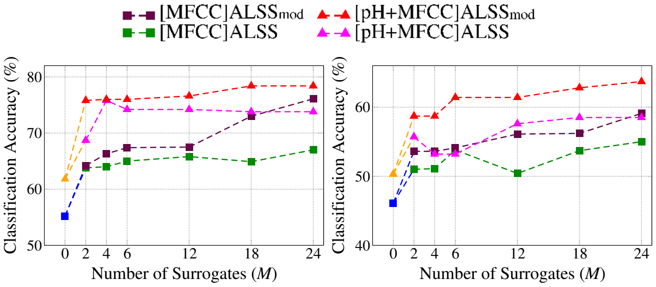


Fig. 6. The K-SVD reconstruction error.



(a) (b)

Fig. 7. Average classification accuracy considering MFCC and pH+MFCC for rooms Meeting (a) and Stairway (b).

Figure 6 illustrates the reconstruction error of reverberated test feature for each room. The adaptive learning technique reduces the reconstruction error in all scenarios. The maximal average decline is achieved by ALSS$_{\text{mod}}$ for rooms Meeting and LASP2 with values varying from 1.46 to 1.24 and 1.74 to 1.49, which represent a 15% error reduction. The comparable ALSS values are 1.37 and 1.64. Moreover, the ALSS$_{\text{mod}}$ attains a 14% reduction for rooms LASP1 and Stairway, followed by 5% and 8% for the ALSS solution. This result corroborates on the ALSS$_{\text{mod}}$ capacity to select discriminative, separated and informative models resulting on a more robust scheme regarding the reverberation effect. As preliminary experiments on the subject, it is also presented in this paper the usage of 7 pH feature vectors [14][15] for nonstationary acoustic sources classification under reverberation. The pH consists of a vector of Hurst exponent (H) values that express the time-dependence or scaling degree of acoustic signals. Therefore this feature could be used to discriminate non-stationary acoustic sources.

Fig. 7, illustrates the learning procedure for rooms Meeting (a) and Stairway (b) with the pH+MFCC feature fusion. The ALSS$_{\text{mod}}$ with pH+MFCC leads to the highest classification accuracy of 78.4 and 63.7 for these rooms. Which represent a 3.5 p.p. average classification improvement. Furthermore, both ALSS and ALSS$_{\text{mod}}$ with pH outperform the learning without pH for most of the surrogate cases. This indicates that the pH feature vector could be incorporated on future surrogate learning approaches as its adoption leads to classification accuracy that surpass current learning solutions.

5. CONCLUSION

An adaptive learning method was investigated for the selection of surrogate models to increase their discrimination power under real reverberation distortions. A new surrogate selection criteria was proposed and evaluated considering the MFCC-GMM classification with average rate, ROC curve and AUC values. The Bhattacharrya distance and the K-SVD were adopted to assess separability and derive sparse coding information of selected signals. Results showed that the ALSS$_{\text{mod}}$ improved classification accuracy and class separability providing a reduction on sparse coding reconstruction error for all scenarios. It was also shown that, the pH feature vector can significantly increase the nonstationary source classification accuracy under reverberation.

6. REFERENCES

[1] R. Bolt and A. MacDonald, “Theory of speech masking by reverberation,” *The Journal of the Acoustical Society of America*, vol. 21, no. 6, pp. 577–580, 1949.

[2] A. Nabelek, “Communication in noisy and reverberant environments,” *Acoustical factors affecting hearing aid performance*, pp. 15–28, 1993.

[3] G. Zucatelli and R. Coelho, “Adaptive reverberation absorption using non-stationary masking components detection for intelligibility improvement,” *IEEE Signal Processing Letters*, vol. 27, pp. 1–5, 2020.

[4] S. Zubair, F. Yan, and W. Wang, “Dictionary learning based sparse coefficients for audio classification with max and average pooling,” *Digital Signal Processing*, vol. 23, no. 3, pp. 960–970, 2013.

[5] J. Salamon and J. P. Bello, “Deep convolutional neural networks and data augmentation for environmental sound classification,” *IEEE Signal Processing Letters*, vol. 24, no. 3, pp. 279–283, 2017.

[6] G. Zucatelli, R. Coelho, and L. Zão, “Adaptive learning with surrogate assisted training models for acoustic source classification,” *IEEE Sensors Letters*, vol. 3, no. 6, pp. 1–4, 2019.

[7] J. Naranjo-Alcazar, S. Perez-Castanos, P. Zuccarello, and M. Cobos, “Acoustic scene classification with squeeze-excitation residual networks,” *IEEE Access*, vol. 8, pp. 112287–112296, 2020.

[8] A. Khamparia, D. Gupta, N. G. Nguyen, A. Khanna, B. Pandey, and P. Tiwari, “Sound classification using convolutional neural network and tensor deep stacking network,” *IEEE Access*, vol. 7, pp. 7717–7727, 2019.

[9] E. Fonseca, M. Plakal, D. P. W. Ellis, F. Font, X. Favory, and X. Serra, “Learning sound event classifiers from web audio with noisy labels,” in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 21–25, May 2019.

[10] Z. Shuyang, T. Heittola, and T. Virtanen, “Active learning for sound event classification by clustering unlabeled data,” in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 751–755, IEEE, 2017.

[11] P. Borgnat, P. Flandrin, P. Honeine, C. Richard, and J. Xiao, “Testing stationarity with surrogates: A time-frequency approach,” *IEEE Transactions on Signal Processing*, vol. 58, no. 7, pp. 3459–3470, 2010.

[12] K. Fukunaga, *Introduction to statistical pattern recognition*. Elsevier, 2013.

[13] M. Aharon, M. Elad, and A. Bruckstein, “K-svd: An algorithm for designing overcomplete dictionaries for sparse representation,” *IEEE Transactions on signal processing*, vol. 54, no. 11, pp. 4311–4322, 2006.

[14] R. Sant’Ana, R. Coelho, and A. Alcaim, “Text-independent speaker recognition based on the hurst parameter and the multidimensional fractional brownian motion model,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 3, pp. 931–940, 2006.

[15] L. Zão, D. Cavalcante, and R. Coelho, “Time-frequency feature and ams-gmm mask for acoustic emotion classification,” *IEEE signal processing letters*, vol. 21, no. 5, pp. 620–624, 2014.

[16] R. J. Webster, “A random number generator for ocean noise statistics,” *IEEE journal of oceanic engineering*, vol. 19, no. 1, pp. 134–137, 1994.

[17] M. A. Al-Alaoui, “Novel digital integrator and differentiator,” *Electronics Letters*, vol. 29, pp. 376–378, Feb 1993.

[18] L. Zão and R. Coelho, “Generation of coloured acoustic noise samples with non-gaussian distributions,” *IET Signal Processing*, vol. 6, pp. 684–688, Sep. 2012.

[19] R. Santana and R. Coelho, “Low-frequency ambient noise generator with application to automatic speaker classification,” *EURASIP Journal on Advances in Signal Processing*, vol. 2012, no. 1, p. 175, 2012.

[20] M. Jeub, M. Schafer, and P. Vary, “A binaural room impulse response database for the evaluation of dereverberation algorithms,” in *2009 16th International Conference on Digital Signal Processing*, pp. 1–5, IEEE, 2009.