Improving Non-Stationary Acoustic Source Classification with Metric Learning

Guilherme Zucatelli and Ricardo R. Barioni

***Abstract*— In this work, the metric learning is adopted to improve the classification of non-stationary acoustic sources. The proposed strategy aims to overcome the statistical differences that arise from the non-stationary behavior by learning an optimal function that minimizes intra-class and maximizes inter-class distances. A convolutional neural network with metric learning module generates embedded features of reduced size. Several sources with different degrees of non-stationarity are selected for the acoustic source classification task. Experiments demonstrated that the proposed solution outperforms baseline systems for all individual acoustic sources, leading to increments in the average balanced accuracy, ROC and AUC in all scenarios.**

***Keywords*— non-stationary acoustic sources, multi-class classification, metric learning, deep learning**

I. Introduction

Non-stationarity is a significant challenge for recognizing environmental sounds in signal processing and machine learning research areas [1][2][3][4]. This is especially important for acoustic source classification systems, where limited training samples are usually adopted to discriminate signals with varying statistics over time [5][6]. Rather than fixing a stationarity assumption of acoustic signals, tackling the natural non-stationary behavior could lead to meaningful improvements in various applications such as surveillance systems, hearing aid devices, smart homes and robot navigation.

In a deep learning context, metric learning has been successfully adopted in different acoustic tasks from emotion recognition [7] and medical diagnosis [8] to speaker [9] and acoustic scene classification [10][11]. Acoustic scenes are usually composed of several different sources (Dog Bark, Street Music and Siren) and acoustic effects (i.e. echo and reverberation). This is fundamentally different from recognizing individual non-stationary acoustic sources. The mixture of signals and effects mitigates the non-stationarity of the target source as a practical consequence of the central limit theorem.

In this work, the metric learning strategy is explored to improve individual non-stationary acoustic source classification. The idea is extended from [12] to overcome the statistical differences that arise from the non-stationary behavior by learning an embedding generator network optimal strategy that minimizes intra-class and maximizes inter-class distances [13]. A deep convolutional neural network (CNN) is adopted to extract embedded features of reduced size from time-frequency representations of acoustic signals. Therefore, the CNN can learn similar characteristics on different representations of a target class and map them to adjacent embeddings. Moreover, acoustic sources from different classes lead to separated embedded features.

Several experiments are conducted to validate the proposed solution on a multi-class classification scenario. Acoustic sources with different non-stationary degrees are selected from the UrbanSound [14] and ESC-10 [15] databases. The non-stationarity is objectively accessed based on the Index of Non-Stationarity (INS) [16]. The proposed approach is compared to two baseline solutions: a classical support vector machine (SVM) and a CNN model with a softmax classification layer. As a result, the metric learning solution achieves at least an average 1.4 percentage points (p.p.) increment over the baseline approaches in a multi-class classification task. Moreover, the approach surpass competing methods for all individual non-stationary acoustic sources.

The contributions of this work can be summarized as:

1) Investigation of metric learning approach for non-stationary acoustic source classification.
2) Complete objective non-stationarity assessment for UrbanSound and ESC-10 acoustic sources.
3) Evaluation of the proposed strategy on multi-class classification and acoustic source verification tasks.

The remaining of this paper is organized as follows. In Section II it is described the non-stationarity of acoustic sources. The proposed metric learning strategy is presented in Section III. Experiments and results are described at Section IV. Finally, the conclusion is exposed at the end of this paper.

II. Non-Stationary Acoustic Sources

A key goal for environmental sound classification systems is to achieve a relevant and discriminative representation of each class. This can be challenging when dealing with acoustic sources due to their natural non-stationary behavior. In other words, acoustic sources commonly present temporal and spectral variations throughout time.

The Index of Non-Stationarity (INS) [16] is here defined to objectively examine the non-stationarity of acoustic sources. For a target signal $x(t)$, the INS is obtained considering its multitaper spectral representation $S_x(l, f)$ as

$$S_x(l, f) = \frac{1}{K} \sum_{k=1}^{K} S_x^{(h_k)}(l, f), \tag{1}$$

where l is the frame, f is the frequency bin and $S_x^{(h_k)}(l, f)$ is the spectrogram obtained considering the k-th Hermitian function $h_k(t)$ as the taper [17], i.e.,

$$S_x^{(h_k)}(l, f) = \left| \int x(s) h_k(s - l) e^{-j2\pi f s} \right|^2 \tag{2}$$

Guilherme Zucatelli and Ricardo R. Barioni are with the Speech Signal Processing Team at SAMSUNG Institute for Development in Informatics (SiDi), Campinas, São Paulo - Brazil, e-mail: {g.zucatelli, r.barioni}@sidi.org.br.

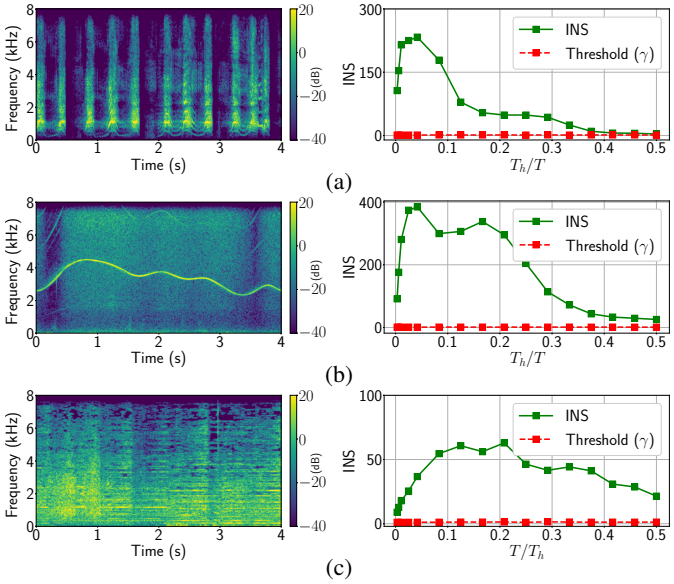


Fig. 1: *Spectrograms and relative INS for acoustic sources Dog Bark (a), Drilling (b), and Street Music (c).*

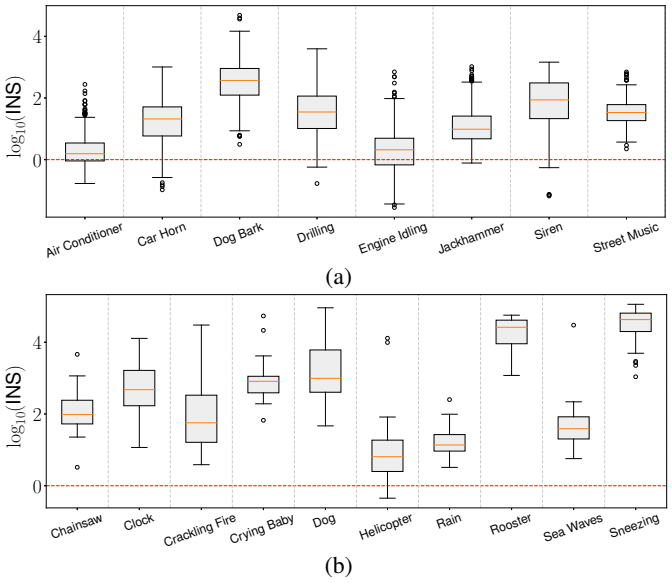


Fig. 2: *Distribution of INS maximum values over acoustic sources from UrbanSound (a) and ESC-10 (b). The red dotted line defines the non-stationary threshold (γ).*

for $h_k(t) = e^{-t^2/2}H_k(t)/\sqrt{\pi^{1/2}2^k k!}$, where $H_k(t)$ are Hermite polynomials that are obtained by recursion $H_k(t) = 2tH_{k-1}(t) - 2(k-2)H_{k-2}(t)$ for $k \geq 2$ and initializations of $H_0(t) = 1$ and $H_1(t) = 2t$.

This measure compares the target signal with stationary references called surrogates, adopting the symmetric Kullback-Leibler distance and log-spectral deviation [18]. Surrogate signals are generated by changing the phase of the spectral representation of $x(t)$ to realizations of a uniform distribution $\mathcal{U}[-\pi, \pi]$, which then guarantees their stationary behavior [16]. The comparison is carried out for different time scales T_h/T, where T_h is the short-time spectral analysis length and T is the total signal duration. For each length T_h, a threshold γ is defined to keep the stationarity assumption considering a 95% confidence degree as

$$INS \quad \begin{cases} \leq \gamma, & \textit{signal is stationary} \\ > \gamma, & \textit{signal is non-stationary}. \end{cases} \tag{3}$$

Fig. 1 depicts the spectrogram and the corresponding INS for three different acoustic sources extracted from the UrbanSound and ESC-10 databases [14][15]. The INS assessment was implemented in Python[1] and conducted in 14 time scales T_h/T. The maximum value of the INS is superior to the non-stationary threshold γ in all cases, which means that all sources are non-stationary. In Fig. 2 the INS maximum distribution for all acoustic sources available in a $\log_{10}$ scale is illustrated. The majority of cases present a maximum INS value above the non-stationarity threshold, reinforcing the assumption of acoustic sources' non-stationarity. For maximum INS values above 100 ($log_{10}(INS) > 2$) the signal is considered as highly non-stationary. It it important to notice that this is the case for higher quartiles of several distributions presented on Fig. 2, as well as for samples (a) and (b) from Fig. 1.

[1]Available at https://github.com/g-zucatelli/pyINS.

The focus of this work relies on the multi-class classification of non-stationary acoustic sources. In addition to the non-stationarity, each class is composed of a variety of audio sources, which challenges the definition of a straightforward classification strategy. This exemplifies the necessity of solutions that can correctly adopt the varying characteristics of acoustic sources to perform identification and discrimination from a multi-class perspective.

III. Backbone Architecture and Metric Learning

From a deep learning perspective, metric learning is designed to (1) aggregate similar characteristics on closer embedding regions, while (2) separating different features on the embedding space. Therefore, we hypothesize that this strategy might be adequate to learn similar characteristics of non-stationary acoustic sources on varying time-frequency representations. This way, the trained model is suited to deal with non-stationary acoustic behavior by selecting the relevant information of a class. Moreover, the embedding generator model could be used on enrollment steps, labelling unseen trained classes for acoustic recognition tasks.

The proposed approach adopts the MobileNet deep convolutional neural network architecture as its backbone [19]. Although other topologies could be considered, this particular CNN has a lower number of parameters, due to depthwise separable convolutions, while maintaining its performance on several applications. The last MobileNet layer is removed since the default model is directly used for classification tasks, whereas an embedding layer is considered in the metric learning strategy. To this end, an average pooling layer, a dense layer, and the metric learning module are respectively included as replacements. Fig. 3 depicts an overview of the proposed model.

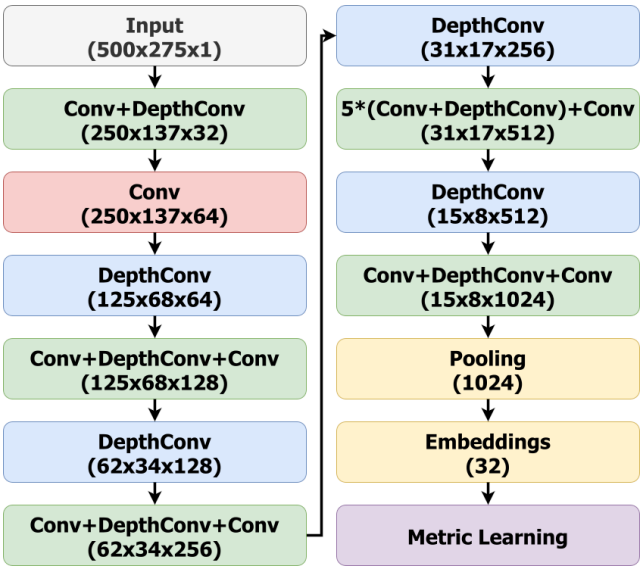


Fig. 3: *The model approach. The colored rectangles denote the MobileNet blocks. The gray rectangle represents the input layer. Conv represents as a sequence of Conv + BatchNormalization + ReLU and DepthConv denotes a sequence of DepthwiseConv + BatchNormalization + ReLU. For each block transition, either the number of channels doubles or the channel dimensions decrease by half. This is followed by the Metric Learning module that is used during the training step.*

A. Metric Learning for Acoustic Source Classification

Metric learning (former called distance metric learning) is a machine learning approach whose main purpose is to, given a set of inputs from different classes, learn a function that minimizes the distance for the same classes and maximizes the distance for different classes [20].

Given a set of m inputs $\{x_i\}_{i=1}^{m} \subseteq \mathbb{R}^n$, the goal is to find a positive semi-definite matrix A so that the to-be-learned distance metric $d(x, y)$ between points $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^n$ is small when x and y are from the same class and large otherwise, as in [13]:

$$d(x, y) = \sqrt{(x - y)^T A (x - y)}. \tag{4}$$

As a Deep Learning solution, metric learning does not need to directly optimize the distance function between the original inputs. Since a deep neural architecture is previously connected to the actual metric distance, the former aims to output a set of features called embeddings, which are then fed to the latter function as inputs. During the training step, the embeddings are updated to satisfy the Metric Learning constraints. After training, the metric learning module is discarded, and the final model outputs the trained embeddings in a feature extractor fashion.

Most deep metric learning approaches rely on minimizing the intra-class and maximizing the inter-class geodesic distance between embeddings of size d. Within the surface of the hypersphere $H \in \mathbb{R}^d$, and considering a total of n classes,

TABLE I: *Metric Learning Loss Function Parameters [21].*

Loss Function	α	β	γ
Modified Softmax	0	0	0
SphereFace	1.35	0	0
CosFace	0	0.1	0
ArcFace	0	0	0.1

the metric loss can be generically defined as

$$L = -\frac{1}{N}\sum_{i=1}^{N} log \frac{e^{s(cos(\alpha\theta_{y_i}+\gamma)-\beta)}}{e^{s(cos(\alpha\theta_{y_i}+\gamma)-\beta)} + \sum_{j=1,j\neq y_i}^{n} e^{s(cos\theta_j)}} \tag{5}$$

where s is the hypersphere radius (which is the norm of the Embedding vector), y_i is the class of instance i and θ_{y_i} denotes the angle between the Embeddings of instance i and the weights $W_j \in \mathbb{R}^d$ from $W \in \mathbb{R}^{d\times n}$. Parameters α and β are defined for each loss function as represented in TABLE I.

IV. Experiments and Results

In order to evaluate the proposed Metric Learning strategy for non-stationary acoustic source classification, eight sources are first selected from the UrbanSound database [14]: Air Conditioner, Car Horn, Dog Bark, Drilling, Engine Idling, Jackhammer, Siren and Street Music. Audios labeled as *Foreground* are considered to guarantee the task of multi-class source classification. The usage of other sources and *Background* audios would rather imply tasks of scene or impulse event classification, which are not the focus of the present work. In total, 4810 audio files at 16 kHz were adopted with an average duration of 3.6 seconds. Similarly, all of the ten sources available from the ESC-10 dataset are used on a multi-class source classification task. Since this dataset is already designed with foreground exposure and limited background noise, filtering classes are not required.

Experiments are performed considering the classic acoustic feature MFCC with 25 coefficients extracted every 21.3 ms and 50% frame overlap. The final feature matrix is composed of the MFCC and its summarized statistics as in [14], which leads to a feature vector of 275 dimension per frame. The Metric model was trained for all loss functions presented on TABLE I with an SGD optimizer and a learning rate of 0.005. For each batch, 32 audio samples of 5 seconds are considered. The training is performed for a total of 20 epochs, where the selected model is obtained based on the highest validation accuracy.

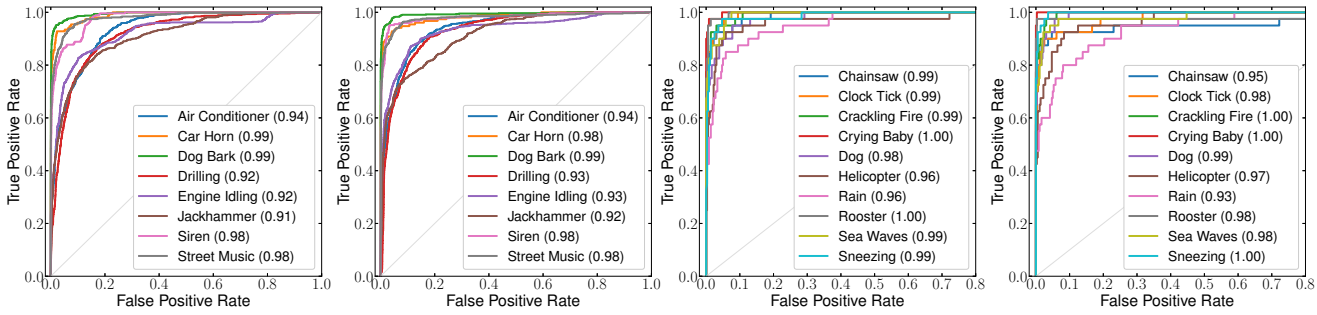
The evaluation is conducted in a multi-fold cross-validation procedure as designed in [14] and [15]. The comparative baseline methods are defined by the classical SVM classifier with a linear kernel and a CNN model with a softmax output classification layer. For Metric Learning, audios are divided into non-overlapping segments. The Metric model is able to map each segment to its corresponding 32-dimension embedding vector. The test occurs by calculating the average distance between the test embeddings and embedding centroids derived from training audio classes. Each test audio is therefore associated with the smallest average distance among the eight acoustic classes.

TABLE II: *Acoustic Source Classification Accuracies (%) for UrbanSound.*

Non-Stat. Sources (%)	Average $\log_{10}(INS_{max})$	Acoustic Source	MFCC-SVM	CNN	Modified Softmax	SphereFace	CosFace	ArcFace
100.0	2.52	Dog Bark	70.2	92.6	92.2	93.0	93.0	**93.3**
100.0	1.53	Street Music	90.0	89.3	88.6	89.4	90.9	**90.9**
99.5	1.08	Jackhammer	50.8	66.1	67.3	66.9	67.9	**68.3**
96.9	1.56	Drilling	79.9	76.3	80.3	80.5	80.8	**81.5**
90.8	1.19	Car Horn	79.7	79.7	77.1	64.7	79.1	**81.7**
90.3	1.79	Siren	70.6	72.5	71.4	**73.2**	68.4	71.0
62.7	0.28	Air Conditioner	49.9	54.3	53.3	**58.2**	54.1	56.6
56.4	0.28	Engine Idling	66.7	70.0	63.3	66.6	**70.1**	68.9
Average Balanced Accuracy			56.1	75.1	74.2	74.1	75.5	**76.5**

TABLE III: *Acoustic Source Classification Accuracies (%) for ESC-10.*

Non-Stat. Sources (%)	Average $\log_{10}(INS_{max})$	Acoustic Source	MFCC-SVM	CNN	Modified Softmax	SphereFace	CosFace	ArcFace
100.0	3.38	Sneezing	87.5	95.0	95.0	90.0	87.5	**97.5**
100.0	2.96	Dog	67.5	77.5	70.0	75.0	75.0	**82.5**
100.0	2.88	Rooster	20.0	90.0	92.5	**97.5**	92.5	92.5
100.0	2.87	Crying Baby	85.0	90.0	90.0	92.5	**95.0**	**95.0**
100.0	2.69	Clock	82.5	85.0	**92.5**	90.0	85.0	85.0
100.0	2.03	Chainsaw	65.0	87.5	85.0	**90.0**	87.5	**90.0**
100.0	1.70	Crackling Fire	85.0	90.0	90.0	**92.5**	90.0	90.0
100.0	1.59	Sea Waves	87.5	87.5	**90.0**	85.0	**90.0**	**90.0**
100.0	0.81	Rain	57.5	65.0	65.0	65.0	**72.5**	62.5
85.0	0.78	Helicopter	70.0	67.5	72.5	**75.0**	**75.0**	67.5
Average Balanced Accuracy			70.7	83.5	84.3	**85.3**	85.0	**85.3**



(a) (b) (c) (d)

Fig. 4: *ROC curve and AUC values for the CNN ((a) and (c)) and ArcFace ((b) and (d)), for UrbanSound and ESC-10, respectively.*

In TABLE II, it is presented the classification accuracy for each acoustic source on UrbanSound database. Note that sources are firstly sorted based on the percentage of non-stationary audios of a given class and secondly on the average $\log_{10}(INS_{max})$. As all signals of Dog Bark and Street Music were objectively accessed as non-stationary, these sources are placed at the top two rows. Engine Idling is the lowest rank, because only 56.4% of the available acoustic signals are non-stationary. The proposed Metric Learning with ArcFace strategy achieves the highest average classification accuracy of 76.5%. The metric learning outperforms the classic MFCC-SVM, leading to at least 18.0 p.p. increment for SphereFace and up to 20.4 p.p. for ArcFace.

In comparison with the competing CNN approach, the proposed solution reaches higher accuracy values for all individually non-stationary and highly non-stationary acoustic sources. In this case, the highest classification accuracy gain of 5.2 p.p. is observed for Drilling with ArcFace, which presents an objective measured non-stationary behavior on 96.9% of the database samples. It is important to note that the Metric Learning ArcFace strategy achieves the highest results for the five most non-stationary acoustics sources of

Dog Bark, Street Music, Jackhammer, Drilling and Car Horn, simultaneously. Moreover, the SphereFace outperforms on Siren and Air Conditioner with 73.2% and 58.2% respectively, while CosFace function presents the highest value of 70.1% for Engine Idling.

The classification results for the ESC-10 dataset is presented on TABLE III. In this case, ArcFace and SphereFace strategies achieve the highest average classification accuracy of 85.3%, which is a 1.8 p.p. increment over competing DNN. Note that, for all non-stationary acoustic sources, the highest accuracy is achieved by a Metric Learning solution. Furthermore, ArcFace overcomes the competing methods on the majority of cases, specially for the most non-stationary acoustic sources. These results reinforce the capacity of the proposed metric learning strategy to overcome the non-stationarity challenge on real acoustic source classification.

As a further comparison between the classic CNN and Metric Learning ArcFace proposed solution, Fig. 4 depict the ROC curve and Area Under the Curve (AUC) for each source verification task and datasets. The true positive audios relate to a target acoustic class, whereas all seven remaining classes are considered for the false positive rate evaluation.

In line with the previous result, the proposed approach achieves a better AUC performance on three non-stationary acoustic sources on UrbanSound (Jackhammer, Drilling and Engine Idling) with and increment from 0.92 up to 0.93. As the main goal of the proposed metric learning method is to reach a higher classification accuracy for each acoustic source, this approach is able to achieve an average AUC value of 0.96 which is 0.01 higher than the classic CNN strategy. Regarding the ESC-10, both methods reach a similar overall average AUC value. Moreover, in Fig. 4 the lower area under the curve is 0.92 for the Jackhammer on UrbanSound, which indicates that the metric method achieves good discrimination among classes.

V. Conclusion

In this work, it was proposed a metric learning-based approach for non-stationary acoustic source classification. The solution adopted a convolutional neural network for embedded feature generation with reduced size. The embedding generation was optimized on similarity constraints in order to maximize intra-class and minimize inter-class distances using the metric learning strategy. Experiments demonstrated that the proposed solution outperforms the baseline system accuracy for all non-stationarity acoustic sources, leading to an overall average accuracy improvement on two largely used acoustic sources datasets. Moreover, the proposed strategy is also able to achieve a higher AUC values for non-stationary acoustic source verification task.

Aknowledgements

The results presented in this paper entitled "Improving Non-Stationary Acoustic Source Classification with Metric Learning" have been developed as part of a project at SiDi, financed by Samsung Eletrônica da Amazonia Ltda., under the auspices of the Brazilian Federal Law of Informatics no. 8248/91.

References

[1] 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.

[2] Gregory Ditzler, Robi Polikar, and Nitesh Chawla, "An incremental learning algorithm for non-stationary environments and class imbalance," in *2010 20th International Conference on Pattern Recognition*. IEEE, 2010, pp. 2997–3000.

[3] 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.

[4] 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)*, May 2019, pp. 21–25.

[5] Justin Salamon and Juan Pablo 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 and R Coelho, "Adaptive learning with surrogate assisted training models using limited labeled acoustic sample sequences," in *2021 IEEE Statistical Signal Processing Workshop (SSP)*. IEEE, 2021, pp. 21–25.

[7] Esam Ghaleb, Mirela Popa, and Stylianos Asteriadis, "Metric learning-based multimodal audio-visual emotion recognition," *IEEE MultiMedia*, vol. 27, no. 1, pp. 37–48, 2020.

[8] Guiping Zhu, Mingzhu Ma, Yuwen Huang, Kuikui Wang, and Gongping Yang, "Dual-domain low-rank fusion deep metric learning for off-the-person ecg biometrics," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 2914–2918.

[9] Joao Monteiro, Md Jahangir Alam, and Tiago H Falk, "Combining speaker recognition and metric learning for speaker-dependent representation learning.," in *INTERSPEECH*, 2019, pp. 4015–4019.

[10] Shubhr Singh, Helen L Bear, and Emmanouil Benetos, "Prototypical networks for domain adaptation in acoustic scene classification," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 346–350.

[11] Jee-weon Jung, Hee-soo Heo, Hye-jin Shim, and Ha-jin Yu, "Dnn based multi-level feature ensemble for acoustic scene classification.," in *DCASE*, 2018, pp. 118–122.

[12] G. Zucatelli and R. Barioni, "A metric learning based solution for non-stationary acoustic source classification," in *XL Brazilian Conference of Telecommunication and Signal Processing*. SBrT, 2022.

[13] Eric Xing, Michael Jordan, Stuart J Russell, and Andrew Ng, "Distance metric learning with application to clustering with side-information," *Advances in neural information processing systems*, vol. 15, 2002.

[14] Justin Salamon, Christopher Jacoby, and Juan Pablo Bello, "A dataset and taxonomy for urban sound research," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 1041–1044.

[15] Karol J Piczak, "Esc: Dataset for environmental sound classification," in *Proceedings of the 23rd ACM international conference on Multimedia*, 2015, pp. 1015–1018.

[16] Pierre Borgnat, Patrick Flandrin, Paul Honeine, Cédric Richard, and Jun Xiao, "Testing stationarity with surrogates: A time-frequency approach," *IEEE Transactions on Signal Processing*, vol. 58, no. 7, pp. 3459–3470, 2010.

[17] F. Cakrak and P.J. Loughlin, "Multiple window time-varying spectral analysis," *IEEE Transactions on Signal Processing*, vol. 49, no. 2, pp. 448–453, 2001.

[18] Michèle Basseville, "Distance measures for signal processing and pattern recognition," *Signal Processing*, vol. 18, no. 4, pp. 349 – 369, 1989.

[19] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[20] Liu Yang and Rong Jin, "Distance metric learning: A comprehensive survey," *Michigan State Universiy*, vol. 2, no. 2, pp. 4, 2006.

[21] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4690–4699.