Low-Complexity Environmental Sound Classification using Cadence Frequency Diagram and Chebychev Moments

Michael Neri
Dept. of Ind., Electr. and Mech. Engineering
Roma Tre University
Rome, Italy
michael.neri@uniroma3.it

Luca Pallotta
School of Engineering
University of Basilicata
Potenza, Italy
luca.pallotta@unibas.it

Marco Carli
Dept. of Ind., Electr. and Mech. Engineering
Roma Tre University
Rome, Italy
marco.carli@uniroma3.it

***Abstract*—The research conducted within the audio signal processing field is increasingly focusing on environmental sound classification. This paper presents a low-complexity Fully Convolutional Network composed of two parallel branches. These branches are responsible for extracting features from the Cadence Frequency Diagram representation and the Chebychev moments, respectively. By utilizing both domains of machine- and deep-learning, the proposed pipeline takes advantage of the unique characteristics of each. The key strength of this architecture lies in its reduced number of layers and parameters, as well as its ability to efficiently compute the Cadence Frequency Diagram and Chebychev moments. The effectiveness of the proposed pipeline is demonstrated through various tests conducted on two audio datasets, namely UrbanSound8K and ESC-50.**

***Index Terms*—Environmental Sound Classification, Deep Learning, Chebychev Moments, Audio Processing, Cadence Frequency Diagram**

I. Introduction

Environmental sound classification has gained significant attention in recent years due to its wide range of applications, which include audio anomaly detection and classification systems [1], audio and music tagging [2]–[4], as well as emotion classification [5]. In such contexts, a plethora of works aimed at classifying different audio patterns are designed. This mission is generally performed following two competing strategies which are referred to as machine- and deep-learning. All the methodologies that try to exploit handcrafted features belong to the former group, e.g., [6]–[9]. Differently, the frameworks categorized in the latter category utilize more sophisticated structures that elaborate on the incoming signal to automatically set the parameters in a deep neural network (DNN) which extracts features used in the classification process, e.g., [10]–[14]. Even though machine-learning model provide complete control of the extracted features and a limited computational complexity, each specific procedure requires strong theoretical expertise and difficulties in generalizing the developed methods. Conversely, deep-learning has the advantage of being more generally applicable with also higher classification performance. These results are however often paid in terms of a higher computational burden and the need for the availability of many data to train the network.

In this contribution, the main features of these two competing strategies are exploited to take advantage of the strengths of both and limit their weaknesses. To this aim, a fully convolutional network (FCN) is designed which encompasses two convolutional branches for extracting features from the cadence frequency diagram (CFD) representation and the Chebychev moments coefficients, respectively. Hence, the proposed pipeline's first branch consists of a few convolutional layers alternating with max-pooling layers that apply a proper transformation of the input CFD.

The CFD, known as cadence velocity diagram (CVD) in the radar context, was introduced in [15], [16] to improve micro-Doppler classification from radar signals. Its application to audio context has also shown interesting results in distinguishing audio sources as reported in [9]. Differently, the second branch (with only a reduced number of convolutional layers) directly operates on the Chebychev moments. More precisely, following the line of reasoning of [9], Chebychev moments are extracted starting from the incoming audio signals to form the feature vector. However, differently from [9] where Chebychev moments were concatenated to the Mel-frequency cepstrum coefficients (MFCC), here they are exploited alone and used as the input to the machine-learning-based branch of the developed network. Going into details, Chebychev moments [17] are computed projecting the CFD, in turn, derived from the Mel-spectrogram of the audio signals, into their related polynomials. The orthogonality properties of the Chebychev moments, together with the fact that they are defined in a discrete set, allow us to summarize in a feature vector the relevant information embedded in the CFD. As a matter of fact, they have already been widely used for classification purposes, like image or radar hand-gesture classification [18], [19]. The extracted image moments are then organized in a matrix form and used as input to the second branch of the designed FCN.

The main advantage of the proposed framework lies in its

limited computational complexity. In fact, the processing is performed on the reduced size CFD (efficiently obtained by means of the fast Fourier transform (FFT)) rather than the wider Mel-spectrogram. Additionally, the Chebychev polynomials can be a priori computed and stored since they only depend on the polynomial order and the CFD size. Beyond, the above attentions, the developed 2D FCN is characterized by few layers with also a very low number of parameters, if compared with state of the art (SOTA) architectures.

Performances have been assessed in terms of the average accuracy of performing cross-validation on two widely investigated datasets, viz. UrbanSound8K and ESC-50. Results show the effectiveness of the proposed approach in comparison with other existing machine learning approaches with higher computational complexity.

II. Proposed method

This section provides a description of each component of the proposed approach. First, the audio preprocessing step is illustrated in detail. Thereafter, an introduction on how the CFD and the Chebychev moments are computed from a generic image is given. Then, starting from the CFD and the extracted Chebychev moments from a time-frequency representation, a 2D FCN is designed for the classification of audio samples. Specifically, the neural network encompasses two branches, $f_{\mathrm{CFD}}(\cdot)$ and $f_{\mathrm{Cheb}}(\cdot)$, that are responsible for extracting features from the CFD representations and the Chebychev moments, respectively.

A. Audio preprocessing

Let $x[n] \in \mathbb{R}^{1\times L}$ be a single-channel audio with L samples. A preprocessing step is applied to $x[n]$ to obtain the complex short-time Fourier transform (STFT). The transform is computed with a Hanning window of length 32 ms and 50% overlap. Then, the Mel-Spectrogram $\Psi \in \mathbb{R}^{T\times N_{\mathrm{DFT}}}$ is computed by means of the Mel-filterbank on the squared magnitude of the STFT, where T and N_{DFT} are the time and frequency mel-bins, respectively.

B. Cadence Frequency Diagram

The CFD is devised in [15], [16] as an enhanced tool allowing to better distinguish micro-Doppler-based radar signals with respect to the use of the classic spectrogram. It consists of a discrete Fourier transform (DFT) applied to pass from a time-frequency to a cadence-frequency domain. In fact, performing a DFT of the spectrogram for each frequency bin allows to obtain the cadence information, that is the repetition cycle of each frequency involved in the original signal. As in [9], the CFD is computed from the Mel-spectrogram modulus used in place of the spectrogram in [15], [16], that is

$$\Delta(\xi,m) = \sum_{k=0}^{N_{\mathrm{CFD}}-1} |\Psi(k,m)|\, e^{-j2\pi k\xi/N_{\mathrm{CFD}}}, \quad m=0,\dots,N_{\mathrm{DFT}}, \tag{1}$$

where $|\cdot|$ denotes the modulus of its complex argument, ξ is the cadence frequency, N_{CFD} is the number of frequency bins used in the CFD computation, and N_{DFT} is the number of frequency bin involved in the Mel-spectrogram computation.

The extraction of the Chebychev moments occurs through the projection of the CFD into the orthogonal Chebychev polynomials that will be described in Subsection II-C by means of

$$C_{l,h} = \frac{1}{\bar\rho(l,N_{\mathrm{CFD}})\bar\rho(h,N_{\mathrm{DFT}})}\cdots \sum_{x=0}^{N_{\mathrm{DFT}}-1}\sum_{y=0}^{N_{\mathrm{CFD}}-1} \bar c_l(x)\bar c_h(y)|\overline{\Delta}(y,x)|, \tag{2}$$

with $\bar\rho$ the normalized amplitude factor, $\bar c_l(\cdot)$ and $\bar c_h(\cdot)$ the Chebychev polynomial of order l and h. Finally, $\overline{\Delta}(\cdot,\cdot)$ is the CFD normalized to be in the interval $[0,1]$.

C. Chebychev Polynomials and Moments

Let $f(x,y)$ be a non-negative real-defined image of size $L\times H$. The moments of order $l+h$ of $f(x,y)$ are defined as its projection on the monomials $x^l y^h$, by means of the following integral [20]:

$$M_{l,h} = \iint_{\mathbb{R}^2} x^l y^h f(x,y)\,dx\,dy. \tag{3}$$

In this work, Chebyshev polynomials are used as generating monomials $\{x^l y^h\}$ due to their orthogonality proprieties [17]. More specifically, Chebychev moments can be derived as the projection of the image $f(x,y)$ on the related polynomials (3) to a discrete polynomial set, that is:

$$C_{l,h} = \frac{1}{\bar\rho(l,L)\bar\rho(h,H)}\sum_{x=0}^{L-1}\sum_{y=0}^{H-1} c_l(x)c_h(y)f(x,y), \tag{4}$$

being $c_l(x)$ the Chebychev polynomial of order l that can be written as

$$\sum_{x=0}^{L-1} c_l(x)c_h(x) = \rho(l,L)\delta_{l,h}, \tag{5}$$

with $0\le l\le L-1$, $0\le h\le H-1$, and $\delta_{l,h}$ the Kronecker delta function that is equal to 1 when $l=h$ and 0 otherwise. Moreover, $\bar\rho$ is the normalized amplitude factor given by:

$$\bar\rho(l,L) = \frac{\rho(l,L)}{L^{2l}} = (2l)!\binom{L+l}{2l+1}\frac{1}{L^{2l}}. \tag{6}$$

In addition, the generalized hypergeometric function [21] of order (3, 2) is introduced, that is

$${}_3F_2\left(a_1,a_2,a_3;b_1,b_2;z\right) = \sum_{k=0}^{+\infty}\frac{(a_1)_k(a_2)_k(a_3)_k}{(b_1)_k(b_2)_k}\frac{z^k}{k}, \tag{7}$$

with $(a)_l$ denoting the Pochhammer symbol [21] given by

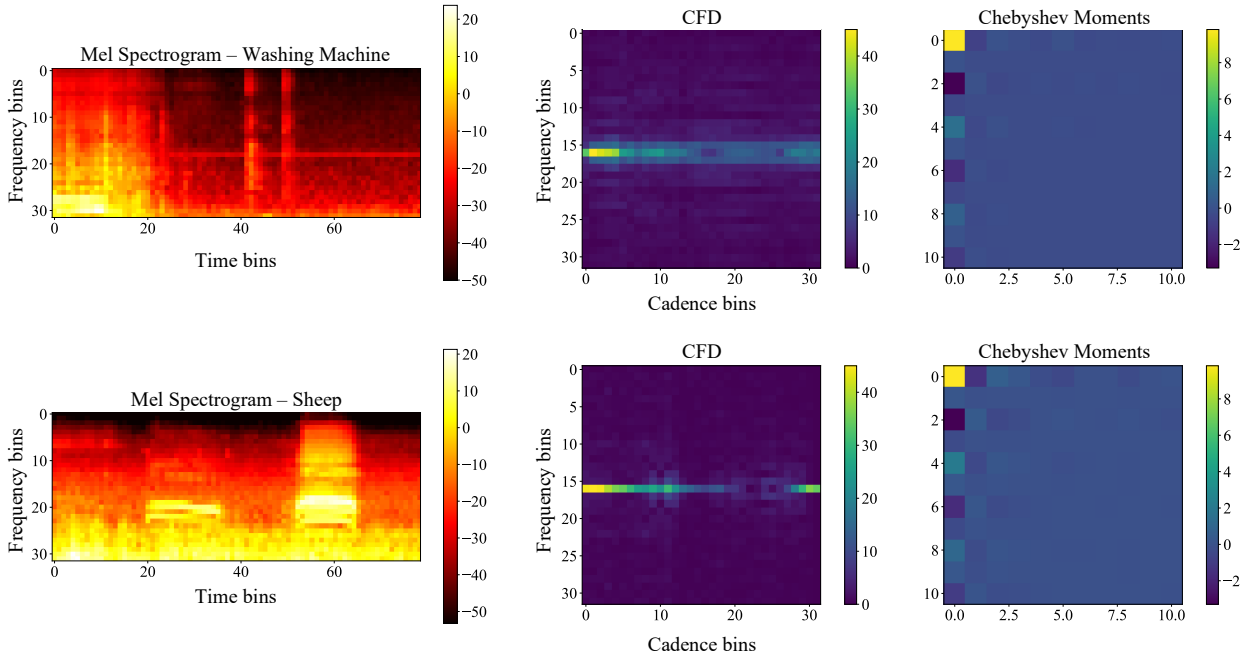

Fig. 1. Examples of Mel-spectrogram, CFD with $N_{\mathrm{CVD}} = 32$, and Chebyshev moments with $l = 10$ from two audios of the ESC-50 dataset.

$$(a)_l = a(a+1)\cdots(a+l-1) = \frac{\Gamma(a+l)}{\Gamma(a)}. \tag{8}$$

Finally, the Chebychev polynomials [17] can be expressed by means of the following equation

$$c_l(x) = (1-L)_l \, {}_3F_2\left(-l, -x, 1+l, ; 1, 1-L; 1\right), \tag{9}$$

where $x = 0, 1, 2, \ldots, L-1$.

Figure 1 depicts some examples of the Mel spectrogram, the CFD, and the Chebychev moments used for feature extraction and classification.

D. Fully Convolutional Network

Let $f(\cdot) : \mathbb{R}^{N_{\mathrm{CFD}} \times N_{\mathrm{CFD}}} \rightarrow \mathbb{R}^{N_{\mathrm{classes}}}$ be the FCN that processes the CFD representation obtained from (1) and predicts the class-wise probabilities, also denoted as class logits, $\hat{\mathbf{y}} \in \mathbb{R}^{N_{\mathrm{classes}}}$. Specifically, $f(\cdot)$ encompasses two neural branches (f_{CFD} and f_{Cheb}) that are responsible for processing the CFD and the Chebychev moments, respectively.

In more detail, the branch $f_{\mathrm{CFD}}(\cdot) : \mathbb{R}^{N_{\mathrm{CFD}} \times N_{\mathrm{CFD}}} \rightarrow \mathbb{R}^{N_{\mathrm{classes}}}$ consists of three convolutional blocks, labeled as $\mathrm{ConvBlock}(C_i)$, where C_i represents the number of output channels. These blocks are used for extracting spatial features from the 2D representation. Each block performs a 2D convolution with 3×3 kernels, followed by batch normalization [22] and the activation function called exponential linear unit (ELU) [23], which is defined as

$$\mathrm{ELU}(x) = \begin{cases} x, & x \geq 0 \\ \alpha(e^x - 1), & x < 0 \end{cases} \tag{10}$$

where α is set to 1 to avoid negative values saturation. After the first two blocks, an image downsampling process is performed using the $\mathrm{MaxPool}(2, 2)$ function. Then, a linear projection layer is employed to transform the output of the final convolutional block into a feature tensor with a number of channels equal to the number of classes N_{classes}. The idea is to have a feature map for each class. To this aim, a 1×1 convolutional layer is utilized. Finally, class logits $\hat{\mathbf{y}}_{\mathrm{CFD}} \in \mathbb{R}^{N_{\mathrm{classes}}}$ are obtained by means of the global average pooling (GAP) operator.

In parallel, the branch $f_{\mathrm{Cheb}}(\cdot) : \mathbb{R}^{N_{\mathrm{CFD}} \times N_{\mathrm{CFD}}} \rightarrow \mathbb{R}^{N_{\mathrm{classes}}}$ extracts the Chebychev moments from the normalized CFD and provides the class logits $\hat{\mathbf{y}}_{\mathrm{Cheb}} \in \mathbb{R}^{N_{\mathrm{classes}}}$. First, Chebychev moments are extracted from the CFD following the procedure detailed in Sec II-C. Then, the coefficients are arranged in a squared matrix of size $(l+1) \times (l+1)$ to be processed by the network $f_{\mathrm{Cheb}}(\cdot)$.

Regarding the architecture of the Chebychev branch, it is composed of 2 consecutive ConvBlock with 32 and 64 filters with size 3×3. Similarly to $f_{\mathrm{CFD}}(\cdot)$, a linear projection layer reduces the number of features maps to the number of classes, and the GAP layer maps to class logits.

Finally, the prediction of the approach $\hat{\mathbf{y}} \in \mathbb{R}^{N_{\mathrm{classes}}}$ is computed by element-wise multiplication, denoted as $\otimes$, between the two branch estimations as a soft-voting strategy

$$\hat{\mathbf{y}} = \hat{\mathbf{y}}_{\mathrm{CFD}} \otimes \hat{\mathbf{y}}_{\mathrm{Cheb}}. \tag{11}$$

This training procedure has been applied to provide coherence between the two proposed audio representations. The architec-

TABLE I
DESCRIPTION OF THE PROPOSED 2D FCN.

Input: normalized CFD $\overline{\Delta} \in \mathbb{R}^{N_{\mathrm{CFD}} \times N_{\mathrm{CFD}}}$

CVD branch $f_{\mathrm{CVD}}(\cdot)$	**Moments branch** $f_{\mathrm{Cheb}}(\cdot)$
$\mathrm{ConvBlock}(128)$	Chebychev moments extraction of order N_{Cheb}
$\mathrm{MaxPool}(2,2)$	$\mathrm{ConvBlock}(16)$
$\mathrm{ConvBlock}(128)$	$\mathrm{ConvBlock}(64)$
$\mathrm{MaxPool}(2,2)$	Projection to N_{classes} channels
$\mathrm{ConvBlock}(128)$	$\mathrm{GAP}(\cdot)$
Projection to N_{classes} channels	-
$\mathrm{GAP}(\cdot)$	-
Output: class logits $\hat{\boldsymbol{y}}_{\mathrm{CFD}}$	**Output**: class logits $\hat{\boldsymbol{y}}_{\mathrm{Cheb}}$

ture is trained by means of the Cross-Entropy loss between the predicted and the ground truth labels.

Table I provides a comprehensive overview of the architecture of the FCN. The whole neural network configuration has been tuned by exploiting a hyperparameters grid search optimization.

III. EXPERIMENTAL RESULTS

A. Dataset

To assess the performance of the proposed framework, a 10-fold and 5-fold cross-validation is applied to the UrbanSound8K [4] and ESC-50 [3] datasets, respectively. In particular, the UrbanSound8K dataset comprises 8732 audio files of at most 4 seconds of duration and divided into $N_{\mathrm{classes}} = 10$ classes. Differently, ESC-50 dataset contains 2000 short clips recorded at a sampling frequency of 44.1 kHz grouped into $N_{\mathrm{classes}} = 50$ classes.

The performance on these datasets is assessed by means of the accuracy metric, evaluating the number of perfect matches between predicted and ground truth labels.

B. Results varying the order of Chebychev moments L

Performance of the proposed approach on UrbanSound8K and ESC50 is depicted in Table II and Table III, respectively. The best results, which outperform the baselines, are obtained when the CFD is computed from the spectrogram with 32 mel bins. Moreover, it is notable that the best order of Chebychev moments depends on the scenario and on the amount of available data. In fact, with a smaller dataset, i.e., ESC50, the best performance is observed with a greater number of Chebychev coefficients than in the case of a larger dataset, i.e., UrbanSound8k.

Moreover, it is worth highlighting that the experiments have been carried out only with the Mel-spectrogram. Even though the approach is agnostic with respect to the audio representation, tests conducted on other time-frequency analyses, such as STFT and MFCC, yielded not converging training procedures.

C. Comparison with state of the art approaches

Table IV depicts the performance and the computational complexity (with respect to the number of learnable parameters) of well-known deep learning strategies, e.g., convolutional neural networks (CNNs) [1], [24] and Transformers [25], for audio classification without the use of additional training data such as AudioSet [2]. It is notable how the proposed approach is three orders of magnitude lower than well-known state-of-the-art models for sound recognition. Since the Chebychev polynomials only depend on the polynomial order as well as on N_{CVD}, they can be a priori computed. This is compliant with real-time applications of the proposed pipeline.

Moreover, regarding the computational complexity of deep neural networks, as explained in [26], having a smaller number of learnable parameters can help mitigate the risk of overfitting, especially when dealing with small datasets. Overfitting occurs when the model becomes too complex and starts to memorize noise or outliers in the training data. A simpler model with fewer parameters is less prone to overfitting. Additionally, with fewer parameters to update, the optimization process requires less computational resources and time. This

TABLE II
MEAN CLASSIFICATION ACCURACY WITH 95% CONFIDENCE INTERVAL ON THE URBANSOUND8K DATASET USING THE 10-FOLD CROSS-VALIDATION.

UrbanSound8K [4]

	Accuracy
Baseline	66.00
CFD Mel16 + order 10	0.72 ± 0.05
CFD Mel16 + order 15	0.73 ± 0.07
CFD Mel16 + order 20	0.67 ± 0.11
CFD Mel32 + order 10	0.72 ± 0.06
CFD Mel32 + order 15	$\mathbf{0.73 \pm 0.05}$
CFD Mel32 + order 20	0.67 ± 0.09
CFD Mel64 + order 10	0.73 ± 0.06
CFD Mel64 + order 15	0.72 ± 0.06
CFD Mel64 + order 20	0.71 ± 0.06

TABLE III
MEAN CLASSIFICATION ACCURACY WITH 95% CONFIDENCE INTERVAL ON THE ESC50 DATASET USING THE 5-FOLD CROSS-VALIDATION.

ESC50 [3]

	Accuracy
Baseline	44.30
CFD Mel16 + order 10	0.59 ± 0.05
CFD Mel16 + order 15	0.58 ± 0.02
CFD Mel16 + order 20	0.60 ± 0.03
CFD Mel32 + order 10	0.57 ± 0.04
CFD Mel32 + order 15	0.59 ± 0.06
CFD Mel32 + order 20	$\mathbf{0.62 \pm 0.05}$
CFD Mel64 + order 10	0.58 ± 0.04
CFD Mel64 + order 15	0.60 ± 0.04
CFD Mel64 + order 20	0.57 ± 0.02

TABLE IV

STUDY ON THE COMPUTATIONAL COMPLEXITY AND PERFORMANCE OF THE PROPOSED APPROACH WITH CFD ON 32-BINS MEL-SPECTROGRAM IN COMPARISON WITH STATE-OF-THE-ART ARCHITECTURES. WE DENOTE WITH ↑ WHEN THE PERFORMANCE IS BETTER WHEN THE METRIC IS HIGH AND ↓ OTHERWISE. DASH SYMBOL - MEANS NO EXPERIMENTS HAVE BEEN PROVIDED BY THE AUTHORS.

Model	Params (M) ↓	Acc ESC50 ↑	Acc USK8 ↑
CNN-based			
CNN14-PANN [1]	81.06	0.83	**0.79**
AemNet [24]	14.40	0.77	0.77
Transformer-based			
AST [25]	88.10	**0.87**	-
Proposed FCN	**0.32**	0.62	0.73

can be advantageous when working with limited computational capabilities or large datasets [26].

In addition, thanks to the CFD computation, the size of the input feature is lower than canonical time-frequency representations such as STFT and Mel-spectrogram. In fact, the configuration of the proposed approach that yields the best results encompasses a CFD of size 32×32. Instead, without this domain shift, a canonical time-frequency analysis that computes the same preprocessing yields a 32×64 spectrogram, increasing the overall forward step of neural networks.

However, a drawback of this approach is the loss of the time information. In fact, performing a DFT, for each frequency bin in the Mel-spectrogram, produces a new frequency-cadence domain, in which the cadence provides information about the amount of repetition of each frequency within the observed signal for all the observation time. Therefore, the original time information is integrated and hence somehow lost when the second DFT is applied.

IV. CONCLUSION

In this work, a new architecture that employs the CFD and the Chebychev moments for the classification of environmental sound is presented. Specifically, a low-complexity learning-based approach is designed for extracting features and classifying audio from a novel feature set. However, as mentioned in the discussion, the employed representation losses time information, making the architecture not suitable for tasks where time-wise classification is required such as sound event detection (SED) [14]. A possible improvement is to compute the CFD and Chebychev pipelines on sliding windows of the starting time-frequency representation. By doing so, it is possible to evaluate the features in a time-aware fashion. In conjunction, the employment of more advanced architecture attention-based, such as ViT [27], [28], and large-scale audio datasets, such as AudioSet [2], could improve the effectiveness of the proposed approach while always keeping an eye on the computational cost.

ACKNOWLEDGMENTS

The work of M. Neri and M. Carli has been partially supported by the H2020 ECSEL EU Project Intelligent Secure Trustable Things (INSECTT) and Italy, Grant Agreement Number 876038. The document reflects only the authors' view and the Commission is not responsible for any use that may be made of the information it contains.

REFERENCES

[1] K. Qiuqiang, C. Yin, I. Turab, W. Yuxuan, W. Wang, and M. D. Plumbley, "PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[2] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017.

[3] K. J. Piczak, "ESC: Dataset for Environmental Sound Classification," in *Proceedings of the 23rd ACM international conference on Multimedia*, 2015.

[4] J. Salamon, C. Jacoby, and J. P. Bello, "A Dataset and Taxonomy for Urban Sound Research," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014.

[5] S. Mao, P. C. Ching, and T. Lee, "Enhancing segment-based speech emotion recognition by iterative self-learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 123–134, 2022.

[6] S. Chu, S. Narayanan, and C.-C. J. Kuo, "Environmental Sound Recognition with Time–Frequency Audio Features," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1142–1158, 2009.

[7] J. T. Geiger and K. Helwani, "Improving Event Detection for Audio Surveillance using Gabor Filterbank Features," in *23rd European Signal Processing Conference (EUSIPCO)*. IEEE, 2015, pp. 714–718.

[8] D. M. Agrawal, H. B. Sailor, M. H. Soni, and H. A. Patil, "Novel TEO-based Gammatone Features for Environmental Sound Classification," in *25th European Signal Processing Conference (EUSIPCO)*. IEEE, 2017, pp. 1809–1813.

[9] L. Pallotta, M. Neri, M. Buongiorno, A. Neri, and G. Giunta, "A Machine Learning-Based Approach for Audio Signals Classification using Chebychev Moments and Mel-Coefficients," in *7th International Conference on Frontiers of Signal Processing (ICFSP)*, 2022.

[10] 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.

[11] B. Bahmei, E. Birmingham, and S. Arzanpour, "CNN-RNN and Data Augmentation Using Deep Convolutional Generative Adversarial Network for Environmental Sound Classification," *IEEE Signal Processing Letters*, vol. 29, pp. 682–686, 2022.

[12] H. Park and C. D. Yoo, "CNN-Based Learnable Gammatone Filterbank and Equal-Loudness Normalization for Environmental Sound Classification," *IEEE Signal Processing Letters*, vol. 27, pp. 411–415, 2020.

[13] H. Song, S. Deng, and J. Han, "Exploring Inter-Node Relations in CNNs for Environmental Sound Classification," *IEEE Signal Processing Letters*, vol. 29, pp. 154–158, 2022.

[14] M. Neri, F. Battisti, A. Neri, and M. Carli, "Sound event detection for human safety and security in noisy environments," *IEEE Access*, vol. 10, pp. 134230–134240, 2022.

[15] A. Ghaleb, L. Vignaud, and J. M. Nicolas, "Micro-Doppler Analysis of Wheels and Pedestrians in ISAR Imaging," *IET Signal Processing*, vol. 2, no. 3, 2008.

[16] S. Björklund, T. Johansson, and H. Petersson, "Evaluation of a micro-Doppler Classification Method on mm-Wave Data," in *IEEE Radar Conference*. IEEE, 2012.

[17] R. Mukundan, S. H. Ong, and P. A. Lee, "Image Analysis by Tchebichef Moments," *IEEE Transactions on Image Processing*, vol. 10, no. 9, pp. 1357–1364, 2001.

[18] S. P. Priyal and P. K. Bora, "A Study on Static Hand Gesture Recognition using Moments," in *International Conference on Signal Processing and Communications (SPCOM)*. IEEE, 2010.

[19] L. Pallotta, M. Cauli, C. Clemente, F. Fioranelli, G. Giunta, and A. Farina, "Classification of micro-Doppler Radar Hand-Gesture Signatures by means of Chebyshev Moments," in *IEEE 8th International Workshop on Metrology for AeroSpace (MetroAeroSpace)*. IEEE, 2021.

[20] M.-K. Hu, "Visual Pattern Recognition by Moment Invariants," *IRE Transactions on Information Theory*, vol. 8, no. 2, pp. 179–187, 1962.

[21] R. Diaz and E. Pariguan, "On Hypergeometric Functions and Pochhammer k-Symbol," *arXiv preprint math/0405596*, 2004.

[22] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning (ICML)*. pmlr, 2015.
[23] C. Djork-Arné, T. Unterthiner, and S. Hochreiter, "Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs)," in *International Conference on Learning Representations (ICLR)*, 2016.
[24] P. Lopez-Meyer, J. A. del Hoyo Ontiveros, H. Lu, and G. Stemmer, "Efficient End-to-End Audio Embeddings Generation for Audio Classification on Target Applications," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.
[25] Y. Gong, Y. Chung, and J. Glass, "AST: Audio Spectrogram Transformer," in *Interspeech*, 2021.
[26] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning (still) requires rethinking generalization," *Communications of the ACM*, vol. 64, no. 3, pp. 107–115, 2021.
[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.
[28] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, Z. Yang, Y. Zhang, and D. Tao, "A Survey on Vision Transformer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 87–110, 2023.