A Machine Learning-based Approach for Audio Signals Classification using Chebychev Moments and Mel-Coefficients

Luca Pallotta, Michael Neri, Martino Buongiorno, Alessandro Neri, and Gaetano Giunta
Department of Industrial, Electronic, and Mechanical Engineering
via Vito Volterra, 62, Roma Tre University
Rome, Italy
e-mail: {luca.pallotta, michael.neri, alessandro.neri, gaetano.giunta}@uniroma3.it, mar.buongiorno1@stud.uniroma3.it

***Abstract*—This paper proposes a machine learning-based architecture for audio signals classification based on a joint exploitation of the Chebychev moments and the Mel-Frequency Cepstrum Coefficients. The procedure starts with the computation of the Mel-spectrogram of the recorded audio signals; then, Chebychev moments are obtained projecting the Cadence Frequency Diagram derived from the Mel-spectrogram into the base of Chebychev moments. These moments are then concatenated with the Mel-Frequency Cepstrum Coefficients to form the final feature vector. By doing so, the architecture exploits the peculiarities of the discrete Chebychev moments such as their symmetry characteristics. The effectiveness of the procedure is assessed on two challenging datasets, UrbanSound8K and ESC-50.**

***Index Terms*—audio classification, Chebychev moments, cadence velocity diagram, MFCC, machine learning**

I. Introduction

Audio signals recognition consists of extracting relevant features from a sound recording in order to categorize it into semantic classes. In the recent years, this topic has attracted the research community because of the several possible applications. Audio signals recognition, also called audio pattern recognition, involves demanding tasks such as audio and music tagging [1]–[3], emotion classification [4] and audio anomaly detection and classification systems [5]. The aforementioned challenges are partially solved by introducing deep learning- and/or machine learning-based algorithms, exploiting traditional time-frequency representation as a input. However, the audio annotation process is time-demanding and prone to error, leading to a lack of large and high quality-annotated audio datasets. For this reason, the research community focuses also on developing new feature sets for training Artificial Intelligence (AI)-based models (the interested reader could refer for instance to the following non-exhaustive list of references [6]–[8]).

In this paper, an architecture based on a machine learning approach is designed to automatically classify audio signals by introducing a new feature set. Specifically, the proposed framework is based on a construction of a feature vector that arises from the concatenation of two distinct feature vectors:

- the Chebychev moments [9] which are extracted from the Cadence Frequency Diagram (CFD) and derived from the Mel-spectrogram of the recorded audios;
- the Mel-Frequency Cepstrum Coefficients (MFCC).

Even though the MFCC have been widely applied for audio classification, it worth to underline the benefits that could be obtained by the use of the Chebychev moments. These moments have been already used for hand-gesture classification in images or radar [10], [11], but, to the authors' best knowledge, they are not yet exploited in audio signals classification. Chebychev moments, differently from other image moments [12], are defined on a discrete set, hence, they can be implemented without performing any approximation. Moreover, they can be used in real-time applications thanks to their symmetry property that allows to reduce the number of moments to derive and, consequently, the overall computational time. Performances have been assessed in terms of the average accuracy metric performing a cross-validation on two challenging and widely investigated datasets, viz. UrbanSound8K and ESC-50. Results show the effectiveness of the proposed architecture in comparison with other existing machine learning approaches.

The paper is organized as follows: the proposed architecture for audio signals classification is deeply described in Section II together with details about Chebychev moments. Then, the effectiveness of the proposed method is assessed on the two challenging UrbanSound8K and ESC-50 datasets, whose results are discussed in Section III. Finally, Section IV draws the conclusions and provides some useful hints for future developments.

II. Audio Signals Recognition Algorithm

This section is devoted to the description of the proposed algorithm for classification of audio signals. In particular, the developed procedure is based on a machine learning approach, in which some peculiar features are extracted from the audio signals after some processing in order to discriminate them.

This work 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.

In summary, the method extracts two families of features, viz. Chebychev moments [9] and MFCC, that are then concatenated to form the final feature vector that feeds the classifier. The improvements provided by the use of Chebychev moments are strictly related to their intrinsic discrete nature that allows not to perform approximations. Moreover, their symmetric characteristics, as well as the possibility of a priori storing the polynomials, allow also to reduce the computation complexity of entire system.

A. Algorithm Description

The architecture of the proposed machine learning-based algorithm, described in the present section, is schematically illustrated in Figure 1.

The starting point of the method is the raw audio signal acquisition associated with one of possible sound event, and recorded by means of a classic microphone. This signal indicated as $a[n]$, $n = 0, \ldots, N-1$ comprises N samples depending on the used sampling frequency as well as on its time duration. Then, the signal is firstly processed to derive its Mel-spectrogram [13]. This is performed following some few steps; the signal $a[n]$ is firstly divided into short overlapped blocks of samples through the use of a smoothing window function $w[\cdot]$ whose size rules the trade-off between temporal and frequency resolution, then its Short-Time Fourier Transform (STFT) is computed as

$$\text{STFT}\{a[n]\}(k,m) = \sum_{n=0}^{N-1} a[n]w[n-k]e^{-j2\pi mn/N_{\text{DFT}}}, \quad (1)$$

where $k = 0, \ldots, K-1$, with K the number of time frames, whereas $m = 0, \ldots, N_{\text{DFT}} - 1$ denotes the frequency bin index. It is worth to underline that the smoothing window is applied to each frame that is then Fourier transformed through a specific number of points representing the spectrogram size in the frequency variable. If the number of frequency bins to compute the Discrete Fourier Transform (DFT) is a power of 2, then the efficient Fast Fourier Transform (FFT) algorithm can be used. At the next step, the modulus of STFT is given as input to the Mel filter bank, whose output is summed up to finally form the Mel-spectrogram diagram.

The next step in the proposed pipeline consists in a transformation of the Mel-spectrogram into a new domain referred to as CFD. In particular, in [14], [15] the Cadence Velocity Diagram (CVD) is derived to improve the extraction of micro-Doppler features from the spectrogram of radar signals. In fact, this transformed domain provides information about the repetition cycle of each frequency involved in the signal, that is dubbed cadence frequency. Therefore, following the line of reasoning of [14], [15], the CFD is computed as an additional domain to be investigated together with the Mel-spectrogram. More specifically, it is herein evaluated performing the DFT (through the efficient FFT algorithm) of the Mel-spectrogram modulus (in place of the classic spectrogram used in [14], [15]) for each frequency bin. Now, indicating with Ψ the above-mentioned Mel-spectrogram, the CFD is computed as:

$$\Xi(\xi, m) = \sum_{k=0}^{K-1} |\Psi(k,m)|\, e^{-j2\pi k\xi/K}, \quad (2)$$

where ξ is the cadence frequency.

Once the complex-valued CFD is computed, we firstly take the logarithm of its modulus and we normalized it in interval $[0, 1]$ to be compliant with the extraction procedure of Chebychev moments. More in detail, the normalized logarithm of the CFD modulus, say $\overline{\Delta}$, is projected in the orthogonal basis of the Chebychev polynomials (more insights about Chebychev polynomials and moments are provided in Subsection II-B) through the following operation

$$C_{l,h} = \frac{1}{\bar{\rho}(l,L)\bar{\rho}(h,H)} \sum_{x=0}^{N_{\text{DFT}}-1} \sum_{y=0}^{N_{\text{CVD}}-1} \bar{c}_l(x)\bar{c}_h(y)\overline{\Delta}(y,x), \quad (3)$$

where N_{CVD} denotes the number of frequency bins used to compute the CFD, $\bar{\rho}$ is a normalized amplitude factor described in Subsection II-B, and $\bar{c}_l(\cdot)$ is the Chebychev polynomial of order l. It is herein worth to underline that since the Chebychev polynomials only depend on the polynomial order (a priori set) as well as on N_{CVD} (this point is better detailed in Subsection II-B), they can be a priori computed. This is compliant with real-time applications of the proposed pipeline.

Finally, the feature vector $\boldsymbol{f}_1$ is constructed with the above moments as

$$\boldsymbol{f}_1 = [C_{0,0}, C_{0,1}, \ldots, C_{l,h}]^T. \quad (4)$$

From inspection of Figure 1 it is also evident that after the Mel-spectrogram computation, the MFCC are also extracted. In particular, they are obtained as the amplitudes of the Discrete Cosine Transform (DCT) of the logarithm of the Mel-spectrogram. Then, the feature vector $\boldsymbol{f}_2$ is constructed taking the mean value of each MFCC over time, say $\overline{\text{MFCC}}$, that is

$$\boldsymbol{f}_2 = \left[\overline{\text{MFCC}}_1, \overline{\text{MFCC}}_2, \ldots, \overline{\text{MFCC}}_{N_{\text{DFT}}}\right]^T. \quad (5)$$

Then, the feature vector $\boldsymbol{f}$ used to train the classifier is obtained by concatenation of the above mentioned feature vectors $\boldsymbol{f}_1$ and $\boldsymbol{f}_2$ as

$$\boldsymbol{f} = \left[\boldsymbol{f}_1^T, \boldsymbol{f}_2^T\right]^T. \quad (6)$$

Finally, the audio classification is carried out by machine learning-based classifier such as k-Nearest Neighbour (k-NN) and Random Forest (RF).

B. Theory of Chebychev Moments

Before starting, it is worth recalling that the moments of order or degree $l + h$, $M_{l,h}$, of a non-negative real-defined image of size $L \times H$, $f(x, y)$, are defined as its projection on the monomials $x^l y^h$, by means of the integral [12]:

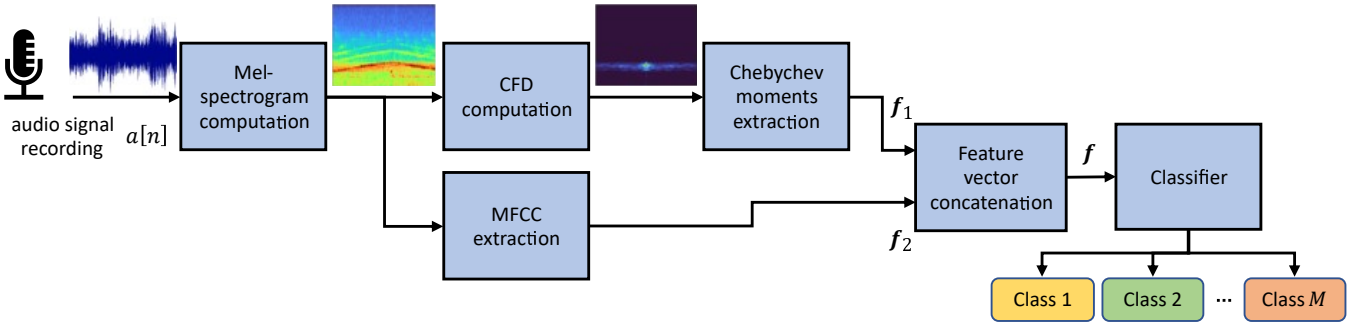


Figure 1. Block scheme of the proposed architecture for audio signals classification.

$$M_{l,h} = \iint_{\mathbb{R}^2} x^l y^h f(x,y)\, dx\, dy. \tag{7}$$

In general, the monomials $\{x^l y^h\}$ used in (7), do not share the orthogonality condition, hence producing non-orthogonal moments. However, this drawback is overcame by the Chebychev polynomials described in [9]. More in detail, Chebychev polynomials of order l describe a set of orthogonal functions sharing useful characteristics [9] and can be cast in the following form

$$c_l(x) = (1-L)_l\, {}_3F_2\left(-l, -x, 1+l, ; 1, 1-L; 1\right), \tag{8}$$

where $x = 0, 1, 2, \ldots, L-1$. The term $(a)_l$ is the Pochhammer symbol [16] defined as

$$(a)_l = a(a+1)\cdots(a+l-1) = \frac{\Gamma(a+l)}{\Gamma(a)}, \tag{9}$$

whereas

$${}_3F_2\left(a_1, a_2, a_3; b_1, b_2; z\right) = \sum_{k=0}^{+\infty} \frac{(a_1)_k (a_2)_k (a_3)_k}{(b_1)_k (b_2)_k} \frac{z^k}{k}. \tag{10}$$

It is now worth noticing that (8) can be also rewritten in a more simple form as:

$$c_l(x) = l! \sum_{k=0}^{l} (-1)^{l-k} \binom{L-1-k}{l-k} \binom{l+k}{l} \binom{x}{k}. \tag{11}$$

As already said, Chebychev polynomials share the orthogonality condition that reduces to the following expression

$$\sum_{x=0}^{L-1} c_l(x) c_h(x) = \rho(l,L)\delta_{l,h}, \tag{12}$$

where $0 \le l, h \le L-1$ and $\delta_{l,h}$ is the Kronecker delta function that is equal to 1 when $l = h$ and 0 otherwise, whereas the term $\rho(l,L)$ is an amplitude factor defined as:

$$\rho(l,L) = (2l)! \binom{L+l}{2l+1}. \tag{13}$$

Moreover, to ensure the numerical stability for the moments computation, the scaled Chebychev polynomials are considered instead, that is:

$$\bar{c}_l(x) = \frac{c_l(x)}{L^l}. \tag{14}$$

Hence, the Chebychev moments are obtained projecting the image $f(x,y)$ of size $L \times H$ on the specific polynomials given in (14):

$$C_{l,h} = \frac{1}{\bar{\rho}(l,L)\bar{\rho}(h,H)} \sum_{x=0}^{L-1} \sum_{y=0}^{H-1} \bar{c}_l(x) \bar{c}_h(y) f(x,y), \tag{15}$$

with $\bar{\rho}$ the normalized amplitude factor defined as:

$$\bar{\rho}(l,L) = \frac{\rho(l,L)}{L^{2l}}. \tag{16}$$

III. Performance Assessment and Results

In this section we show the effectiveness of the proposed architecture based on the joint exploitation of both Chebychev moments and MFCC to automatically distinguish among different audio sources. Tests are conducted on two publicly available databases, viz. UrbanSound8K [3] and ESC-50 [2]. In particular, the UrbanSound8K dataset comprises 8732 audio files of at most 4 seconds of duration and divided into the following 10 different classes: *air conditioner, car horn, children playing, dog bark, drilling, engine idling, gun shot, jackhammer, siren and street music*. As to the ESC-50 dataset, it has 2000 short clips recorded at a sampling frequency of 44.1 kHz grouped into 50 classes of various common sound events: *dog, rain, crying baby, door knock, helicopter, rooster, sea waves, sneezing, mouse click, chainsaw, pig, crackling fire, clapping, keyboard typing, siren, cow, crickets, breathing, door, wood creaks, car horn, frog, chirping birds, coughing, can opening, engine, cat, water drops, footsteps, washing machine, train, hen, wind, laughing, vacuum cleaner, church bells, insects, pouring water, brushing teeth, clock alarm, airplane, sheep, toilet flush, snoring, clock tick, fireworks, crow, thunderstorm, drinking, glass breaking, hand saw*.

Then, to assess the performance of the proposed framework, a 10-fold and 5-fold cross-validation is applied on the

Table I
MEAN CLASSIFICATION ACCURACY (%) FOR EACH FEATURE SET ON THE URBANSOUND8K DATASET USING THE 10-FOLD CROSS-VALIDATION AND TWO DIFFERENT CLASSIFIERS, VIZ. K-NN AND RF.

UrbanSound8K [3]

	k-NN	RF
Baseline [3]	55.00	66.00
MFCC	37.82	50.91
pseudo-Zernike order 20 [17]	38.39	60.05
Chebychev order 10	37.37	63.65
Chebychev order 20	37.70	62.13
Chebychev order 10 + MFCC (ours)	40.40	**68.55**
Chebychev order 20 + MFCC (ours)	40.10	67.35

Table II
MEAN CLASSIFICATION ACCURACY (%) FOR EACH FEATURE SET ON THE ESC-50 DATASET USING THE 5-FOLD CROSS-VALIDATION AND TWO DIFFERENT CLASSIFIERS, VIZ. K-NN AND RF.

ESC-50 [2]

	k-NN	RF
Baseline [2]	32.20	44.30
MFCC	18.15	31.60
pseudo-Zernike order 20 [17]	17.85	40.50
Chebychev order 10	13.45	45.05
Chebychev order 20	13.80	42.45
Chebychev order 10 + MFCC (ours)	16.85	**52.15**
Chebychev order 20 + MFCC (ours)	15.00	50.30

UrbanSound8K and on ESC-50 datasets, respectively. As to classifier, both a k-NN with the parameter k set equal to 11 and a RF with 500 trees are used. The settings of classifiers are the result of a grid search over a finite set of hyper-parameters. Results of tests on UrbanSound8K and ESC-50 are reported in terms of average accuracy in Table I and Table II, respectively, for the proposed algorithm considering two different values for the moments order, i.e., 10 and 20. More in detail, let TP be the number of true positive, FP be the number of false positive, FN be the number of false negatives, and TN be the number of true negatives, the average classification accuracy of the system with N_{class} audio classes is defined as follows:

$$Acc = \frac{TP + TN}{TP + FP + TN + FN}. \tag{17}$$

Since the classifier always outputs an audio class and each recording contains an event, i.e. $TN = FN = 0$, Equation (17) becomes:

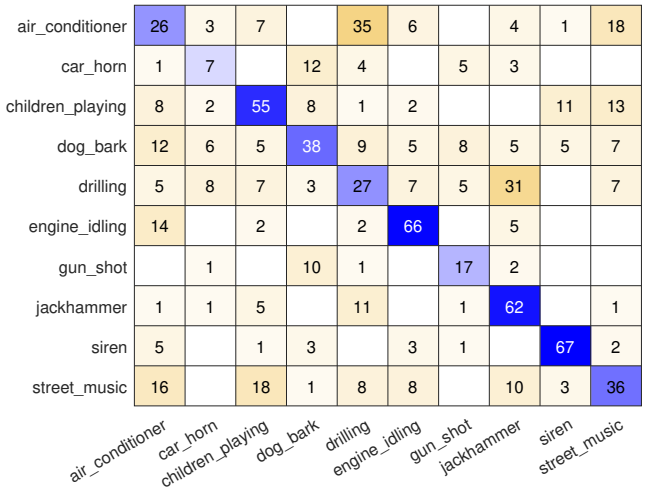
$$Acc = \frac{TP}{TP + FP} \tag{18}$$

Analogously, if the squared $N_{class} \times N_{class}$ confusion matrix CF is available, the average accuracy can be evaluated also as:

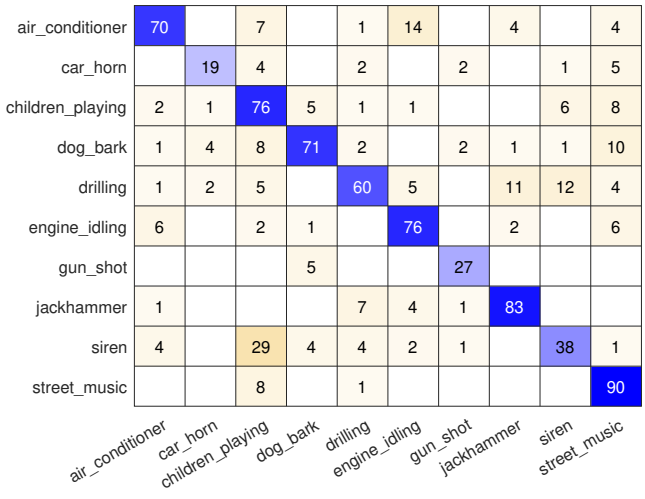
$$Acc = \frac{\sum_{i=1}^{N_{class}} CF(i,i)}{\sum_{i=1}^{N_class} \sum_{j=1}^{N_{class}} CF(i,j)}. \tag{19}$$

For comparison purposes, the results obtained applying other feature sets on both UrbanSound8K and ESC-50 classification are also reported, such as MFCC and Chebychev moments of order 10 and 20, separately.

To corroborate further the results analyzed in terms of average accuracy, Figure 2 shows the confusion matrix of the proposed method with Chebychev order equal to 10 for the case showing the highest accuracy in the cross-validation procedure. Specifically, with the k-NN the maximum value for the accuracy is 49.14%, whereas for the RF it reaches 72.97%. The figures refer to the tests conducted on UrbanSound8K with both the k-NN and RF used as classifier. Differently from the accuracy metric, the confusion matrix allows to better understand which classes are confused with each other. In fact, we can observe that more challenging cases are car horn and siren, whereas the best discrimination is observed in street music, air conditioner, dog bark, and engine idling.



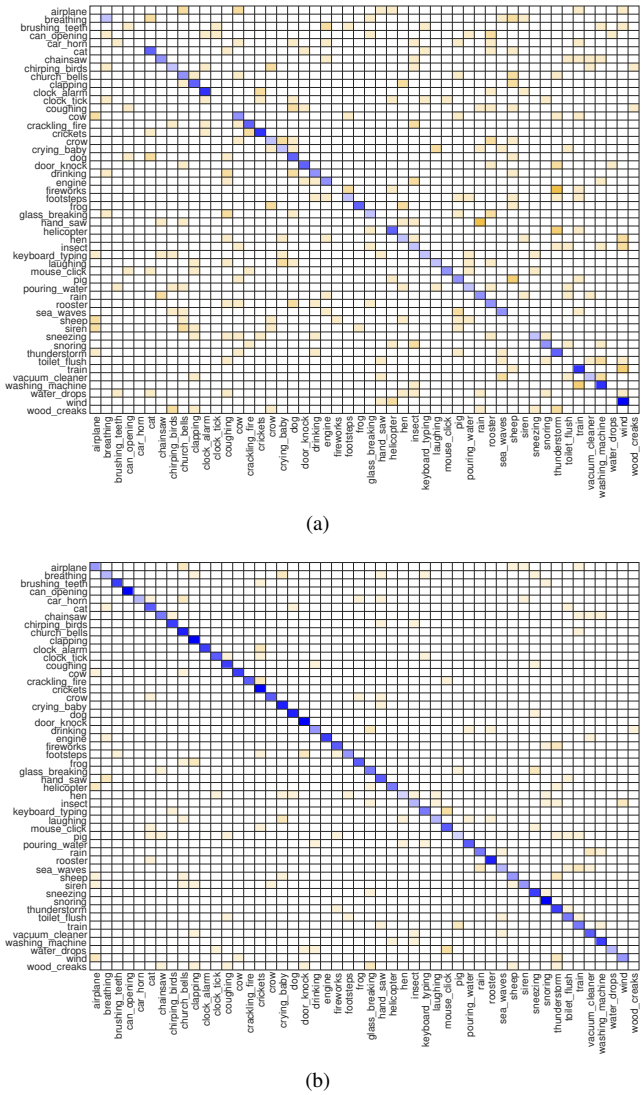
(a)



(b)

Figure 2. Confusion matrix of the proposed method (with Chebychev order equal to 10) for the 10 classes of the UrbanSound8K dataset. Subplots refer to the used classifier, i.e., (a) k-NN, and (b) RF.

Additionally, Figure 3 shows the confusion matrix of the proposed method with Chebychev order equal to 10 for the case showing the highest accuracy in the cross-validation procedure. This figure refers to tests conducted on the database ESC-50 again with both the k-NN and RF classifier. Similarly, in spite of the very high number of classes (i.e., 50), the effectiveness of the proposed framework can be also appreciated in terms of its discriminative capabilities. As a matter of fact, the proposed method allows to reach an accuracy of 20.00% with the k-NN and 58.00% with the RF.



(a)

(b)

Figure 3. Confusion matrix of the proposed method (with Chebychev order 10) for the 50 classes of the ESC-50 dataset. Subplots refer to the used classifier, i.e., (a) k-NN, and (b) RF.

IV. Conclusions

In this paper, an architecture based on a machine learning approach is devised and analyzed with the aim of automatically discriminating different audio signal sources. The proposed framework is based on a concatenation of two different feature vectors. The former is obtained as the Chebychev moments extracted from the CFD that is in turn obtained from the Mel-spectrogram of the incoming audio, and the second that comprises the well-known MFCC. Hence, the proposed procedure has a low computational complexity thanks to the symmetry property of the discrete Chebychev moments as well as the fast computation of the CFD with the FFT algorithm. Tests conducted on UrbanSound8K and ESC-50 datasets have shown interesting results demonstrating the effectiveness of the proposed pipeline. Finally, the use of Chebychev moments with deep learning-based features for improving the accuracy of the system is left as a future research work.

References

[1] J. F. Gemmeke et al., "Audio Set: An Ontology and Human-Labeled Dataset for Audio Events," in *ICASSP*, 2017.

[2] K. J. Piczak, "ESC: Dataset for Environmental Sound Classification," in *Proceedings of the 23rd ACM international conference on Multimedia*, 2015, pp. 1015–1018.

[3] 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, pp. 1041–1044.

[4] 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.

[5] 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.

[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] 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.

[10] S. P. Priyal and P. K. Bora, "A Study on Static Hand Gesture Recognition using Moments," in *2010 International Conference on Signal Processing and Communications (SPCOM)*. IEEE, 2010, pp. 1–5.

[11] 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, pp. 182–187.

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

[13] L. Rabiner and R. Schafer, *Theory and Applications of Digital Speech Processing*, Prentice Hall Press, 2010.

[14] 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, pp. 301–311, 2008.

[15] S. Björklund, T. Johansson, and H. Petersson, "Evaluation of a micro-Doppler Classification Method on mm-Wave Data," in *2012 IEEE Radar Conference*. IEEE, 2012, pp. 0934–0939.

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

[17] C. Clemente, L. Pallotta, A. De Maio, J. J. Soraghan, and A. Farina, "A Novel Algorithm for Radar Classification based on Doppler Characteristics Exploiting Orthogonal Pseudo-Zernike Polynomials," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 51, no. 1, pp. 417–430, 2015.