A Metric Learning Based Solution for Non-Stationary Acoustic Source Classification

Guilherme Zucatelli and Ricardo R. Barioni

***Abstract*— In this work, a metric learning-based approach is proposed for non-stationary acoustic source classification. A classic time-frequency representation of acoustic signals is adopted as input of a convolutional neural network in order to generate embedded features of reduced size. The embedding generation is optimized on similarity constraints in order to maximize intra-class and minimize inter-class distances. Eight sources with different degrees of non-stationarity are selected for the acoustic source classification task. Experiments demonstrated that the proposed solution outperforms the baseline systems for all individual acoustic sources, leading to an increment in the average balanced accuracy of more than twenty percentage points.**

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

I. Introduction

The recognition of environmental sounds has been a topic with growing interest in signal processing and machine learning research areas [1]–[3]. In this regard, acoustic source classification presents a major role in a variety of applications such as surveillance systems, hearing aid devices, smart homes and robot navigation. Real urban acoustic sources, such as Jackhammer and Siren, usually change their temporal and spectral characteristics over time, which implies a natural non-stationary behavior [4] [5]. The non-stationarity represents a fundamental challenge for robust classification since the available training sources have to be sufficient to discriminate signals with varying statistics [6] [7].

Deep learning has emerged as an effective strategy for classification, regression and clustering tasks related to relevant unstructured data such as acoustic signals and images. In this context, metric learning has been successfully adopted in different scenarios from audio-visual emotion recognition to medical diagnosis and object detection [8]–[11]. Traditionally, the focus of this approach relies on learning the parameters of a pairwise distance function, or metric, based on a similarity optimization constraint [8]. For classification problems, this would imply learning an optimal strategy that maximizes intra-class and minimize inter-class distances.

In this work, a metric learning-based approach is proposed for non-stationary acoustic source classification. A deep convolutional neural network (CNN) is adopted to transform time-frequency representations of acoustic signals into embedded features of reduced size. The embedding generator network is optimized based on metric learning similarity constraints. Therefore, the CNN can identify similar characteristics on different representations of a target class and map them to adjacent embeddings. Moreover, acoustic signals from different classes lead to separated embedded features. As far as the authors are concerned, this is the first time that a metric learning-based approach is adopted specifically for the classification of non-stationary acoustic sources.

Several experiments are conducted to validate the proposed solution on a multi-class classification scenario. A total of eight acoustic sources with different non-stationary degrees are selected from the UrbanSound [12] database. The non-stationarity is objectively accessed based on the Index of Non-Stationarity (INS) [13]. The proposed solution is compared to two classical procedures based on the mel-frequency cepstral coefficients (MFCC) acoustic feature and classifiers gaussian mixture models (GMM) and support vector machine (SVM). Four distance functions are selected for the metric learning strategy considering different input audio lengths for the feature extraction and class representation. A Tukey's Honestly Significant Difference [14] statistical evaluation was performed to validate the most suitable model. As a result, the proposed solution achieves at least a 20.0 percentage points (p.p.) increment over the baseline approaches.

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 to correctly identify such acoustic sources and avoid classification errors. 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) [13] 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 [15].

This measure compares the target signal with stationary references called surrogates, adopting the symmetric Kullback-Leibler distance and log-spectral deviation [16]. Surrogate

The authors are with the Speech Signal Processing team at SiDi, Alphaville, Campinas, SP - Brazil, e-mail: {g.zucatelli, r.barioni}@sidi.org.br.

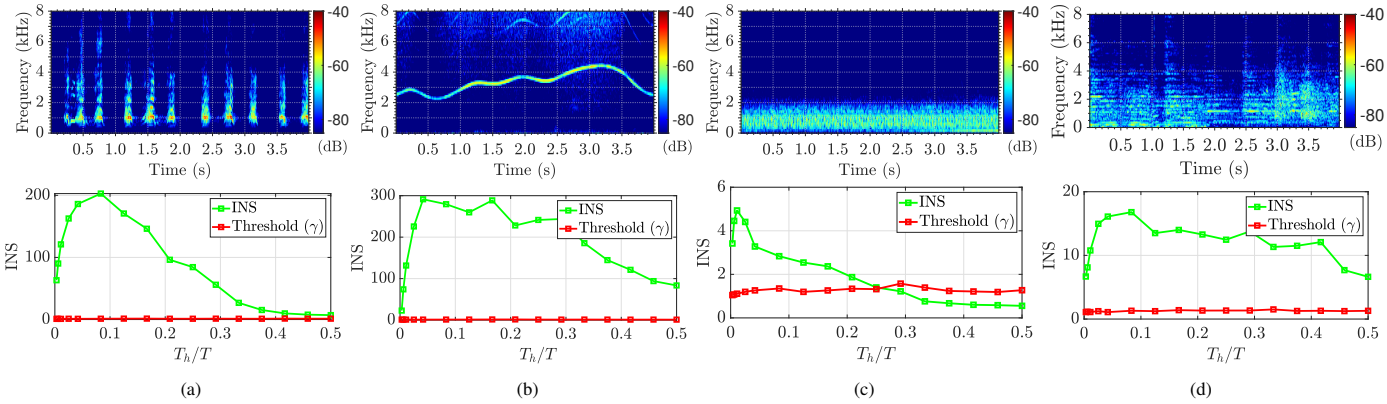

Fig. 1. Spectrograms and relative Index of Non-Stationarity (INS) for acoustic sources Dog Bark (a), Drilling (b), Engine Idling (c) and Street Music (d).

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 [13]. 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

$$\text{INS} \quad \begin{cases} \leq \gamma, & \text{signal is stationary} \\ > \gamma, & \text{signal is non-stationary.} \end{cases} \tag{2}$$

Figure 1 depicts the spectrogram and its relative INS for four different acoustic sources extracted from the UrbanSound database [12]. The maximum value of the INS index is superior to the non-stationary threshold γ in all cases, which means that all sources are non-stationary. In this example, both Dog Bark and Drilling can be characterized as highly non-stationary acoustic sources. Note that, the non-stationarity is accessed with only 4 seconds of signal duration.

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

The proposed approach adopts the MobileNet deep convolutional neural network architecture as its backbone [17]. Although many other network topologies could be considered, this particular CNN has a reduced memory footprint, small number of parameters, fast performance and low latency so that it can be applied in real-time applications. The last MobileNet layer is removed since the default model is directly used for classification tasks, whereas an embedding generation 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. Figure 2 depicts an overview of the proposed model.

A. Metric Learning for Acoustic Source Classification

Metric learning (originally called distance metric learning) is a machine learning approach whose main purpose is to, given a set of acoustic sources from different classes, learn a function that minimizes the distance for the same classes and maximizes the distance for different classes [18]. As a Deep Learning solution, metric learning does not need to directly optimize the distance function from the original sources. 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 so that they satisfy the Metric Learning constraints. After the training, the metric learning module is discarded, and the final model outputs the trained embeddings in a feature extractor fashion.

The main advantage of this approach is that the embedding generation is agnostic of the audio input class. This way, the trained model can be used to extract embeddings from acoustic sources whose classes do not exist on the training dataset. Therefore, reference data of seen and unseen classes can be registered (in a process called Enrollment) during execution time and input data are labeled according to the registered classes.

The most common metric learning approaches rely on minimizing the intra-class and maximizing the inter-class geodesic distance between normalized embeddings within the surface of the hypersphere $H \in \mathbb{R}^d$, where d is the size of the Embeddings. One important example is the Modified Softmax loss [19], which is defined as 3:

$$L_m = -\frac{1}{N}\sum_{i=1}^{N} log\frac{e^{s(cos\theta_{y_i})}}{e^{s(cos\theta_{y_i})} + \sum_{j=1, j\neq y_i}^{n} e^{s(cos\theta_j)}} \tag{3}$$

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 $W \in \mathbb{R}^{d\times n}$, where n is the number of classes being used during the training phase.

Although the Modified Softmax loss outputs separated embeddings from different classes, it still can produce uncertainty regarding frontiers between classes. To solve that, a margin value is enforced to the metric learning function in order to

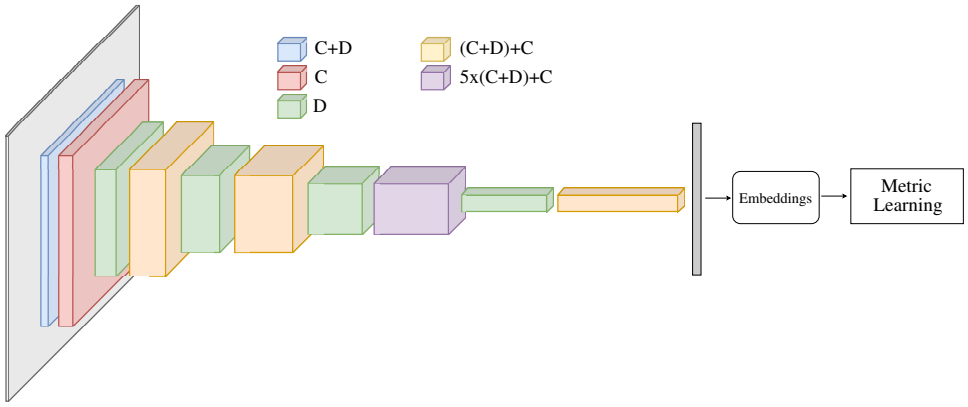

Fig. 2. The model approach. The colored cuboids denote the MobileNet blocks. The gray rectangle represents the input layer. C represents as a sequence of Conv + BatchNormalization + ReLU and D denotes a sequence of DepthwiseConv + BatchNormalization + ReLU. For each cuboid transition, either the number of channels doubles (thicker cuboid) or the channel dimensions decrease by half (more shrunken cuboid). This is followed by the Metric Learning module that is used during the training step.

vanish the frontiers' uncertainties. The margin penalties are commonly applied in three fashions: (i) multiplicative angular margin (known as SphereFace [20]), (ii) additive cosine margin (known as CosFace [21]) and (iii) additive angular margin (known as ArcFace [22]), as it is shown in Equations 4, 5 and 6, respectively:

$$L_s = -\frac{1}{N}\sum_{i=1}^{N} log\frac{e^{s(cos(m_s\theta_{y_i}))}}{e^{s(cos(m_s\theta_{y_i}))}+\sum_{j=1,j\neq y_i}^{n} e^{s(cos\theta_j)}} \quad (4)$$

$$L_c = -\frac{1}{N}\sum_{i=1}^{N} log\frac{e^{s(cos(\theta_{y_i})-m_c)}}{e^{s(cos(\theta_{y_i})-m_c)}+\sum_{j=1,j\neq y_i}^{n} e^{s(cos\theta_j)}} \quad (5)$$

$$L_a = -\frac{1}{N}\sum_{i=1}^{N} log\frac{e^{s(cos(\theta_{y_i}+m_a))}}{e^{s(cos(\theta_{y_i}+m_a))}+\sum_{j=1,j\neq y_i}^{n} e^{s(cos\theta_j)}} \quad (6)$$

Where m_s, m_c and m_a are the SphereFace, CosFace and ArcFace margin values, respectively. Although the margin constraint is applied either on the angle space (SphereFace and ArcFace) or cosine space (CosFace), both approaches aim to penalize the target logit so that it models the intra-class similarity and inter-class dissimilarity.

IV. Experiments and Results

In order to evaluate the proposed Metric Learning strategy for non-stationary acoustic source classification, eight sources were selected from the UrbanSound database [12]: Air Conditioner, Car Horn, Dog Bark, Drilling, Engine Idling, Jackhammer, Siren and Street Music. All sources from this database were manually checked and subjectively classified as *Foreground* or *Background*, related to the distance between the acoustic source and the actual recorder. Only 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 on tasks of scene or event classification, which are not the focus of the present work. Therefore, a total of 4810 audio files were adopted, each with a sample rate of 48 kHz and an average duration of 3.6 seconds.

Experiments are performed considering the classic acoustic feature MFCC with 25 coefficients extracted from 40 Mel bands. MFCC features are obtained on a per-frame basis with a window size of 21.3 ms and 50% frame overlap. The final feature matrix is composed of the MFCC and its summarized

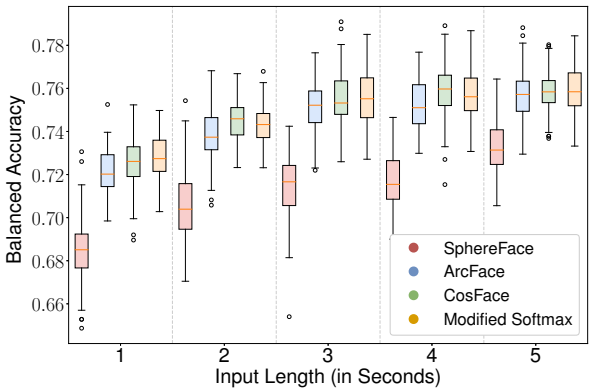

Fig. 3. Box-plot representation of metric learning approaches SphereFace, ArcFace, CosFace and Modified Softmax for five different audios input sizes from 1 second to 5 seconds.

statistics across five consecutive frames: minimum, maximum, median, mean, skewness, kurtosis and the mean and variance of the first and second derivative. This leads to a feature vector of 275 dimension per frame.

The Metric model was trained with SGD optimizer with a learning rate of 0.005. For each batch, a total of 32 audio samples 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 10-fold cross-validation procedure as designed in [12]. The comparative baseline methods are defined by the stochastic GMM models with 32 distributions and an SVM classifier with a linear kernel. Therefore, at every fold interaction, each tested audio source is associated with one of the available classes based on a maximum likelihood criterion or minimum hyperplane distance, respectively. For Metric Learning, audios are divided into non-overlapping segments. The Metric model is able to map each segment to its correspondent 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.

A first experiment is carried out to evaluate different strategies for the Metric Learning approach. A total of 2000

TABLE I

ACOUSTIC SOURCE CLASSIFICATION ACCURACIES (%) OBTAINED WITH MFCC + GMM.

Original Sources	Predicted Sources							
	Air.	Car.	Dog.	Dri.	Eng.	Jac.	Sir.	Str.
Air Conditioner	**16.6**	0	19.9	16.8	4.6	1.1	0.5	40.5
Car Horn	0.2	**52.2**	5.8	16.1	0.7	0.5	0.2	24.2
Dog Bark	0.3	0.4	**83.4**	4.8	0.5	0	0.3	10.3
Drilling	1.4	0.2	7.1	**61.4**	0.3	7.4	1.9	20.3
Engine Idling	6.5	2.2	12.2	11.0	**44.4**	1.0	0.1	22.6
Jackhammer	4.8	0	0	24.4	5.7	**35.8**	0	29.3
Siren	0.3	1.0	30.1	5.8	2.4	0.1	**38.2**	22.1
Street Music	0.5	1.3	7.0	5.4	1.1	0.1	1.3	**83.3**
Average Balanced Accuracy: **53.2%**								

TABLE II

ACOUSTIC SOURCE CLASSIFICATION ACCURACIES (%) OBTAINED WITH MFCC + SVM.

Original Sources	Predicted Sources							
	Air.	Car.	Dog.	Dri.	Eng.	Jac.	Sir.	Str.
Air Conditioner	**36.6**	0.5	9.9	19.5	22.2	6.2	0.0	5.1
Car Horn	12.4	**18.9**	5.8	26.8	10.3	6.5	3.7	15.6
Dog Bark	4.3	0.1	**70.2**	6.6	10.2	1.2	0.7	6.7
Drilling	7.7	0.2	2.7	**60.8**	7.0	13.2	0.1	8.3
Engine Idling	6.4	0	5.0	8.8	**69.2**	8.5	0.2	1.9
Jackhammer	2.9	0	0	19.8	25.2	**50.8**	0	1.3
Siren	12.2	0	19.7	6.1	16.6	2.4	**38.3**	4.7
Street Music	6.0	0.1	5.9	13.1	14.7	3.0	1.2	**56.0**
Average Balanced Accuracy: **56.1%**								

TABLE III

ACOUSTIC SOURCE CLASSIFICATION ACCURACIES (%) OBTAINED WITH METRIC LEARNING.

Original Sources	Predicted Sources							
	Air.	Car.	Dog.	Dri.	Eng.	Jac.	Sir.	Str.
Air Conditioner	**55.2**	0	0	15.3	10.0	10.0	0	9.5
Car Horn	0	**85**	6.5	3.3	0.7	0	0.7	3.9
Dog Bark	1.2	1.2	**93.5**	1.9	0.6	0	0	1.6
Drilling	2.9	1.3	1.8	**80.0**	4.2	6.4	0.4	2.9
Engine Idling	7.5	0.4	0.3	3.5	**73.6**	11.4	0	3.3
Jackhammer	0.4	0.3	0.4	29.1	4.9	**60.7**	0	4.1
Siren	0	1.5	5.2	1.1	8.6	2.6	**68.8**	12.3
Street Music	1.6	1	1.3	2.4	0.6	0.3	1.1	**91.7**
Average Balanced Accuracy: **76.1%**								

different training were considered for this purpose. Models are initialized with the same random parameters. The main objective here is to evaluate the classification accuracy for each of the four margins considered and audio input sizes varying from 1 second to 5 seconds. The balanced accuracy distribution for each of these cases is presented via box-plot in Figure 3. Note that, the adoption of the audio input length can have a significant impact on the balanced accuracy for all margins, with an average increment of at least 2 p.p. The CosFace and Modified Softmax margins have the highest overall results for all scenarios, followed by ArcFace and SphereFace, respectively.

The Tukey's Honestly Significant Difference [14] statistical evaluation is carried among different margins (i.e., CosFace versus Modified Softmax) and different input lengths (i.e., 2-seconds CosFace versus 3-seconds CosFace). The statistically significant differences are observed for most of the tested conditions. A relevant exception occurs for the highest balanced accuracy results of CosFace and Modified Softmax margins. In this case, there is no significant difference obtained for audio lengths higher than 3 seconds. This indicates that, for these

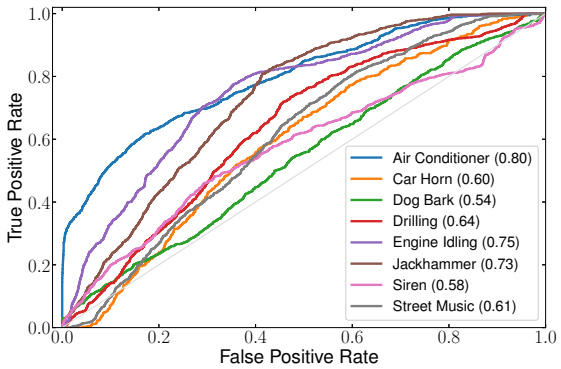

Fig. 4. ROC curve and respective AUC values for the baseline MFCC-GMM strategy.

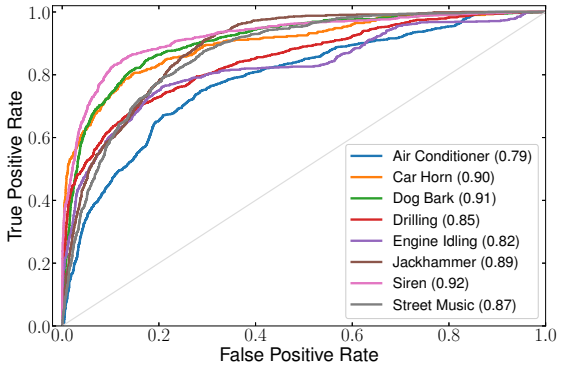

Fig. 5. ROC curve and respective AUC values for the baseline MFCC-SVM solution.

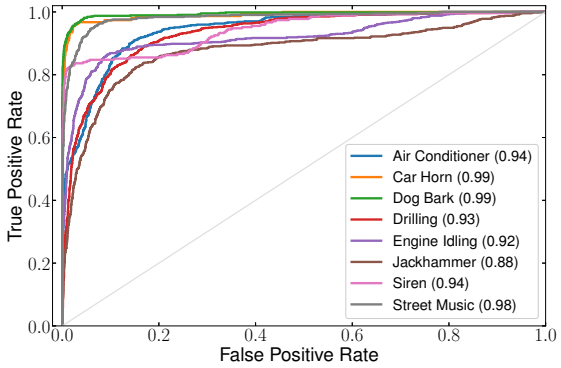

Fig. 6. ROC curve and respective AUC values for the proposed metric learning approach.

Metric Learning approaches, this input length is sufficient to characterize and discriminate non-stationary acoustic sources. Moreover, from the statistical point of view, both margins present similar balanced accuracy distribution for all scenarios considered. This way, a 5-seconds input Modified Softmax model is selected as means to represent the average Metric Learning behavior on further evaluations.

The evaluation of the non-stationary acoustic source classification for the baseline and proposed metric approach is carried out by comparison of the multi-class classification task. Tables I, II and III depicts the confusion matrix results for the MFCC-GMM, MFCC-SVM and the Metric Lerning model, respectively. Note that the proposed Metric Lerning strategy outperforms the baseline for all eight non-stationary acoustic sources, leading to an increment of 20.0 p.p. on the average classification accuracy from 56.1% to 76.1%, when comparing to MFCC-SVM. In this case, the highest

classification accuracy improvement is observed for the Car Horn source from 52.2% to 85.0%, which is equivalent to a 32.8 p.p. increase. Relevant accuracy gain is also observed for Siren and Drilling sources. In the first case, the improvement of 30.5 p.p. leads to an accuracy of 68.8%, whereas for the second source a 18.6 p.p. improvement can be observed, resulting in an accuracy of 80%.

Figures 4, 5 and 6 depict the true positive and false positive rates on a ROC curve for each source considering the MFCC-GMM and MFCC-SVM baseline and metric learning proposed solution, respectively. The true positive audios relate to a target acoustic class, whereas all seven remaining classes are considered for the false positive rate evaluation. The related Area Under Curve (AUC) is also detailed for each class. An area value close to 1 indicates that the particular class not only presented an accurate classification close to 100% but also significant discrimination among other acoustic sources.

Note that, for the baseline approaches, the Siren acoustic source presents the highest discrimination among classes for the MFCC-SVM approach, with an AUC of 0.92. This can be partially explained by the lower error occurrences related to this class (column Siren on Table II). On the other hand, note that the Street Music source acquired an AUC of only 0.61 for the MFCC-GMM approach even with a classification accuracy of 83.3%. This indicates that the MFCC-GMM approach was not able to fully discriminate this acoustic source from others on the multi-class classification task. Once more, this result can be partially justified by the higher error occurrences (Street Music column on Table I). As the main goal of the proposed metric learning method is to reach a higher classification accuracy for each acoustic source and a more discriminative representation among classes, this approach is able to achieve an average AUC value of 0.95. This indicates an increment of 0.08 compared to the MFCC-SVM baseline. Moreover, in Figure 6 the lower area under the curve is 0.88 for the Jackhammer acoustic source, which indicates that the metric learning method achieves good discrimination among classes. This result reinforces the capacity of the proposed metric learning approach for multi-class classification tasks.

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 average overall accuracy improvement of 21.5 percentage points.

Acknowledgement

The results presented in this paper entitled "A Metric Learning Based Solution for Non-Stationary Acoustic Source Classification" 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] 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.

[3] 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.

[4] 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)*, pp. 21–25, IEEE, 2021.

[5] 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.

[6] A. S. Iwashita and J. P. Papa, "An overview on concept drift learning," *IEEE access*, vol. 7, pp. 1532–1547, 2018.

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

[8] E. Xing, M. Jordan, S. J. Russell, and A. Ng, "Distance metric learning with application to clustering with side-information," *Advances in neural information processing systems*, vol. 15, 2002.

[9] E. Ghaleb, M. Popa, and S. Asteriadis, "Metric learning-based multimodal audio-visual emotion recognition," *IEEE MultiMedia*, vol. 27, no. 1, pp. 37–48, 2020.

[10] J. V. Sundgaard, J. Harte, P. Bray, S. Laugesen, Y. Kamide, C. Tanaka, R. R. Paulsen, and A. N. Christensen, "Deep metric learning for otitis media classification," *Medical Image Analysis*, vol. 71, p. 102034, 2021.

[11] X. You, Q. Li, D. Tao, W. Ou, and M. Gong, "Local metric learning for exemplar-based object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 8, pp. 1265–1276, 2014.

[12] 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*, pp. 1041–1044, 2014.

[13] 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.

[14] H. Abdi and L. J. Williams, "Tukey's honestly significant difference (hsd) test," *Encyclopedia of research design*, vol. 3, no. 1, pp. 1–5, 2010.

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

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

[17] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[18] L. Yang and R. Jin, "Distance metric learning: A comprehensive survey," *Michigan State Universiy*, vol. 2, no. 2, p. 4, 2006.

[19] R. Ranjan, C. Castillo, and R. Chellappa, "L2 constrained softmax loss for discriminative face verification," Oct. 3 2019. US Patent App. 15/938,898.

[20] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 212–220, 2017.

[21] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "Cosface: Large margin cosine loss for deep face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5265–5274, 2018.

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