Non-Stationarity Objective Assessment for Acoustic Source Classification

Guilherme Zucatelli[†‡], Ricardo Barioni[†] and Evandro Salles[‡]

***Abstract*— In this work, a new combined feature is proposed to improve the recognition of non-stationary acoustic sources. The idea is to overcome the non-stationarity problem on classification tasks due to mismatches that arise from natural statistics variations. The Index of Non-Stationarity (INS) is adopted to assess the non-stationarity behavior of acoustic signals on a frame-by-frame basis, generating a new feature vector. The evaluation is performed with the combined MFCC+INS feature. 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 the majority of individual acoustic sources, leading to significant increment in the average accuracy in all scenarios. Moreover, a single INS feature value is sufficient to obtain up to 2.7 percentage points gain on the average classification accuracy when compared to the baseline approach.**

***Keywords*— non-stationary assessment, acoustic features, acoustic sources, multi-class classification**

I. Introduction

Non-stationarity is a natural property observed on acoustic signals [1] [2]. This varying temporal and frequency statistics over time impose a significant challenge on classification systems, specially when there are limited number of training samples available [3]–[6]. The correct recognition of acoustic sources in this demanding scenario is beneficial to a variety of purposes such as hearing aid devices [7], robot navigation, smart homes, surveillance systems and applications on the Internet of Audio Things (IoAuT) [8].

Urban environments are commonly rich in acoustic events, which can be broadly divided into scenes and/or sources. Acoustic scenes are usually composed of several sources (Dog Bark, Street Music and Siren) and acoustic effects (i.e. echo and reverberation). This is fundamentally distinct from recognizing individual non-stationary acoustic sources. The mixture of signals and effects mitigate the non-stationarity of target source, which are essentially non-stationary. Therefore, tackling the natural non-stationary behavior is crucial to improve acoustic source classification.

In this work, the Index of Non-Stationarity (INS) [9] is proposed as a complementary feature to improve individual non-stationary acoustic source recognition. It is known that different acoustic sources present distinct degrees of non-stationarity [4] [5] [10]. The idea is to incorporate the non-stationary pattern, intrinsic to every source, as an acoustic feature for classification systems. This solution would overcome the statistical differences that arise from the non-stationary behavior by incorporating this information into a feature vector. To this end, the INS is extracted on a frame-by-frame basis and used to generate the new composed feature (MFCC+INS), where MFCC stands for the classical Mel-Frequency Cepstral Coefficients.

Several experiments are conducted to validate the proposed solution on multi-class classification. Scenarios are separated for varying signal duration from 1 up to 4 seconds. A total of eight acoustic sources with different non-stationary degrees are selected from the UrbanSound [11] database. The proposed approach is compared to a baseline solution wihout INS feature vector considering a classical SVM classifier. A single INS feature value is able to increase the average classification accuracy up to 2.7 p.p. (percentage points). Moreover, the proposed strategy surpass the baseline for the majority of individual non-stationary acoustic sources with a 10.1 p.p. maximum accuracy increase.

The contributions of this work can be summarized as:

1) Design of a new composed feature (MFCC+INS) for non-stationary acoustic source recognition.
2) Evaluation of the proposed strategy on four multi-class classification tasks.

The remaining of this paper is organized as follows. In Section II it is presented the INS measure and the overall scheme for the proposed combined solution. Experiments are described at Section III followed by results and discussions. Finally, the conclusion is exposed at the end of this paper.

II. Non-Stationary Assessment for Acoustic Source Classification

A main target for acoustic source classification systems is to find relevant and discriminative representations of each class. Meaningful features are essential to correctly identify sources and avoid misclassification. This can be particularly difficult for acoustic sources due to their non-stationary behavior, i.e., temporal and spectral variations throughout time.

The Index of Non-Stationarity (INS) [9] is here defined to objectively assess the non-stationarity of acoustic sources. Consider a target signal $x(t)$ and 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 for the k-th Hermitian function

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

[‡] The authors are with the Laboratory of Computer and Neural Systems (CISNE - PPGEE/Ufes), Goiabeiras, Vitoria, ES - Brazil. e-mail: evandro.salles@ufes.br

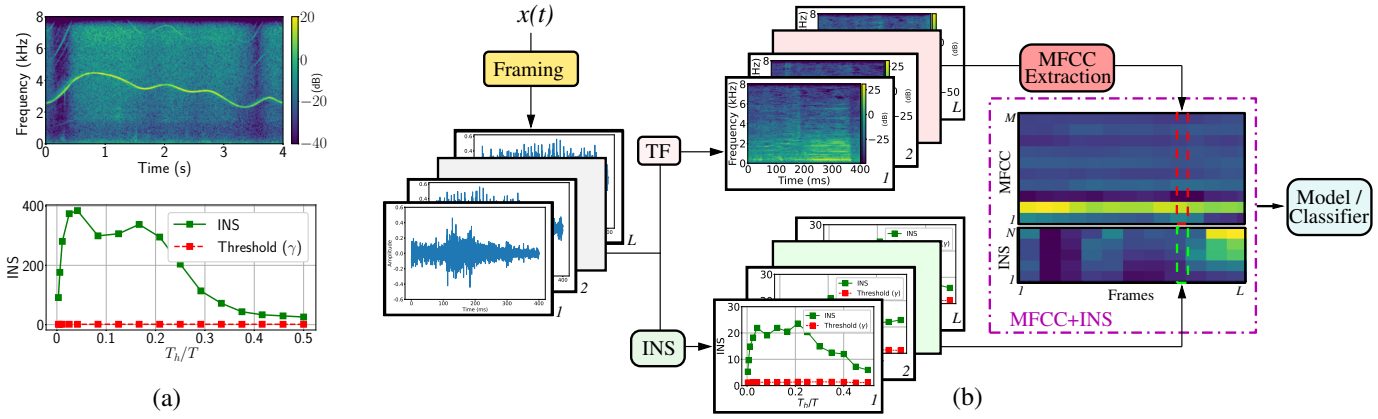

Fig. 1. (a) Spectrogram and Index of Non-Stationarity (INS) for sample Drilling source from UrbanSound. (b) Scheme of proposed combined feature MFCC+INS. For every MFCC frame, a corresponding INS vector is extracted as a direct assessment of the non-stationary behavior.

$h_k(t)$ as the taper [12]:

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

for

$$h_k(t) = e^{-t^2/2} H_k(t) / \sqrt{\pi^{1/2} 2^k k!}, \quad (3)$$

where $H_k(t)$ are Hermite polynomials that are obtained by recursion as

$$H_k(t) = 2tH_{k-1}(t) - 2(k-2)H_{k-2}(t), \quad (4)$$

for $k \geq 2$ and initializations of $H_0(t) = 1$ and $H_1(t) = 2t$.

The INS is a measure that compares the target signal with stationary references called surrogates, adopting the symmetric Kullback-Leibler distance and log-spectral deviation [13]. 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 [9].

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 $\gamma \approx 1$ is defined to keep the stationarity assumption considering a 95% confidence degree as

$$\text{INS} \quad \begin{cases} \leq \gamma, & \text{signal is stationary} \\ > \gamma, & \text{signal is non-stationary.} \end{cases} \quad (5)$$

For the INS assessment, a Python implementation[1] was adopted. Fig. 1 (a) depicts the spectrogram and the corresponding INS for the Drilling acoustic source extracted from the UrbanSound database [11]. The maximum INS value (green) is superior to the non-stationary threshold γ (red), which means that this source is non-stationary. Note that the degrees of non-stationarity is completely dependant of the observed T_h/T scales. Overall, non-stationarity is better assessed on smaller values of T_h/T. On the other hand, as the scale increases and approaches 0.5, acoustic sources present a decaying non-stationary degree, as illustrated in the current example.

Multi-class classification of non-stationary acoustic sources can be a demanding task, specially due to varying time and frequency statistics. In addition to the non-stationarity, each class is composed of a variety of acoustic founts, which challenges the definition of a straightforward classification strategy. This corroborates the necessity of solutions that can accurately adopt the varying characteristics of acoustic sources on a multi-class identification and discrimination perspective.

In this work, the usage of INS is proposed as meaningful acoustic features to discriminate non-stationarity acoustic sources. Fig. 1 (b) illustrates the scheme of designed composed feature. A target signal $x(t)$ is first divided into L overlapping frames $x_l(t)$. Each framed signal $x_l(t)$ is then transformed to a multi-taper time-frequency (TF) representation $S_x(l,f)$ with a corresponding non-stationary pattern, which is objectively assessed by the INS measure considering N different T_h/T scale values. From each TF representation, the state-of-the-art MFCC feature with M elements is extracted, leading to a feature matrix of dimension $L \times M$. Note that in this case the multi-taper spectral is obtained with a single taper.

For every extracted MFCC vector, the proposed approach calculates a reciprocal INS feature vector with size N, one value per observable scale T_h/T. By applying this procedure in all frames, a resulting $L \times N$ INS feature matrix is generated, measuring the non-stationary behavior of a target signal through time. Therefore, the proposed solution accounts for the new combined feature MFCC+INS of size $L \times (M+N)$. The new feature can then be easily incorporated to different models and classifiers. The MFCC+INS feature not only represent the signal into human auditory system perspective but also incorporates information regarding the non-stationarity of target sources which can be beneficial to acoustic source classification.

III. Experiments and Results

In order to evaluate the proposed INS objective measure as a feature strategy for non-stationary acoustic source classification, the following sources were selected from the UrbanSound database [11]: Air Conditioner, Car Horn, Dog Bark, Drilling, Engine Idling, Jackhammer, Siren and Street Music. All signals from this database were manually checked and subjectively classified as *Foreground* or *Background*, related to the distance between the acoustic source and the actual

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

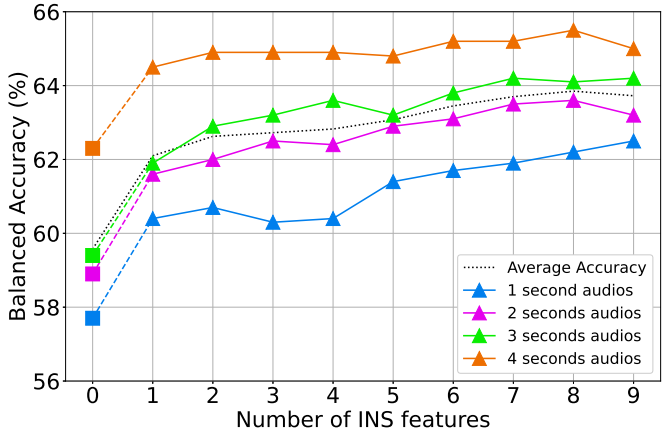

Fig. 2. Acoustic source classification accuracy for proposed MFCC+INS features. Squares represent the baseline MFCC only scenario. The average accuracy per number of INS features is also depicted as black dashed line.

TABLE I

NUMBER OF INS (# INS), CORRESPONDING T_h/T SCALE AND AVERAGE BALANCED ACCURACY (%).

# INS	0	**1**	2	3	4	5	6	7	**8**	9
T_h/T	-	**0.003**	0.006	0.012	0.025	0.042	0.083	0.125	**0.2**	0.4
(%)	59.6	**62.1**	62.6	62.8	62.8	63.1	63.5	63.7	**63.8**	63.7

recorder. *Background* signals can be composed of several acoustic sources and such signals would rather imply tasks of scene or impulse event classification, which are not the focus of the present work. Therefore, only audios labeled as *Foreground* are considered to guarantee the task of multi-class source classification. This leads to a total of 4810 acoustic signals sampled at 44.1 kHz. Experiments are conducted in a 10-fold cross-validation, as designed in [11].

Four different classification scenarios are adopted in the evaluation comprising signal segments of 1.0, 2.0, 3.0 and 4.0 seconds. Acoustic signals with smaller duration (i.e. 1.0s) are usually harder to correctly classify as the available information is reduced. The state-of-the-art acoustic feature MFCC and SVM classifier with is used as the baseline solution. The SVM considered a linear kernel with ℓ_2 penalty loss and unity regularization C. A total of 40 MFCC coefficients are extracted from 80 mel-scaled bands for every 400 ms frames and 50% overlap. The main idea is to progressively incorporate INS values with increasing T_h/T scales, composing the hybrid feature MFCC+INS and evaluate the INS as a feature for non-stationary acoustic sources.

In Fig. 2 it is depicted the balanced accuracy obtained for the baseline MFCC and proposed MFCC+INS feature. Note that the baseline is represented by square marks and correspond to the case where the number of INS feature is zero. TABLE I summarizes which INS scale T_h/T is defined for each INS value. The same scales are adopted for all signal durations. The proposed MFCC+INS feature is evaluated incrementally always considering the n-th smallest T_h/T scales. Note that a single INS value (40 MFCC + 1 INS) is sufficient to significantly increase the balanced accuracy in all scenarios. On average, a single INS feature is able to increase the balanced accuracy from 59.6% to 62.1%, which represents a 2.5 p.p. increment.

TABLE II

ACOUSTIC SOURCE CLASSIFICATION ACCURACIES (%) OBTAINED WITH 1S AUDIO SIGNALS AND 40 MFCC WITHOUT INS (BASELINE).

Original Sources	Predicted Sources							
	Air.	Car.	Dog.	Dri.	Eng.	Jac.	Sir.	Str.
Air Conditioner	**28.3**	0.9	0.9	2.3	41.7	10.5	0.0	15.5
Car Horn	1.7	**83.3**	0.0	1.7	8.3	0.0	0.0	5.0
Dog Bark	0.8	0.0	**81.6**	3.6	6.6	0.2	3.4	3.8
Drilling	5.8	0.1	2.2	**64.9**	8.9	12.6	0.0	5.5
Engine Idling	5.5	0.0	0.7	5.3	**60.9**	20.0	0.2	7.3
Jackhammer	12.9	0.0	0.0	28.4	12.1	**44.5**	0.0	2.1
Siren	11.8	0.0	16.5	1.2	6.7	0.0	**59.8**	3.9
Street Music	8.0	0.3	7.4	8.8	21.0	4.3	2.7	**47.4**
Average Balanced Accuracy: **57.7%**								

TABLE III

ACOUSTIC SOURCE CLASSIFICATION ACCURACIES (%) OBTAINED WITH 1S AUDIO SIGNALS AND 40 MFCC + 9 INS FEATURES.

Original Sources	Predicted Sources							
	Air.	Car.	Dog.	Dri.	Eng.	Jac.	Sir.	Str.
Air Conditioner	**34.1**	0.0	0.4	3.2	40.6	8.3	0.7	12.8
Car Horn	2.4	**83.1**	0.0	2.4	4.8	0.0	0.0	7.2
Dog Bark	1.3	0.0	**85.8**	4.1	3.7	0.2	0.6	4.4
Drilling	6.6	0.2	1.2	**62.9**	10.8	13.1	0.6	4.6
Engine Idling	7.3	0.1	0.5	4.7	**62.4**	18.1	3.3	3.5
Jackhammer	8.8	0.0	0.0	22.4	11.3	**54.4**	0.0	3.1
Siren	10.8	0.0	9.3	3.5	4.6	0.0	**69.9**	1.9
Street Music	6.9	1.1	8.0	11.4	18.2	4.5	2.4	**47.5**
Average Balanced Accuracy: **62.5%**								

Observe that, for most applications, an unit increase in the feature dimension would not imply on a substantial complexity growth, which is an encouraging evidence for INS features. Moreover, by continually adding INS of increasing T_h/T scales the average balanced accuracy is consistently improved, achieving its highest score of 63.8% with eight INS values, a 4.2 p.p. increment. It is important to notice that a balanced accuracy reduction is observed specially for $T_h/T = 0.4$, i.e, considering nine INS features. This is can be partially explained by the fact that acoustic sources usually present a reduction on its non-stationary behavior for higher T_h/T values [10] [4] [5]. Therefore the last incorporated feature does not hold the same discrimination power over features of reduced scale.

In order to evaluate the proposed MFCC+INS composed feature, the balanced accuracy for scenarios with 1 second and 4 seconds signals are selected. In TABLES II and III it is presented the confusion matrix obtained for the classification task of one second duration signals for the baseline and the proposed composed feature with 9 INS features, respectively. This corresponds to the most challenging scheme as the classification needs to be performed on a reduced size signal. In this case, it is observed the highest average balanced accuracy gain of 4.8 p.p. considering the proposed MFCC+INS feature. Aside from Car Horn and Drilling, the composed feature is able to improve the individual classification accuracy for all sources. This is particularly true for the highly non-stationary classes of Siren and Jackhammer, with a 10.1 p.p. and 9.9 p.p. accuracy increase, respectively. Note that acoustic sources Air Conditioner and Dog Bark also presented important accuracy gains corresponding to 5.8 p.p. and 4.2 p.p. This experiment indicates that the proposed INS feature is relevant to non-stationary acoustic source classification, even for the challeng-

TABLE IV

ACOUSTIC SOURCE CLASSIFICATION ACCURACIES (%) OBTAINED WITH 4S AUDIO SIGNALS AND 40 MFCC WITHOUT INS (BASELINE).

Original Sources	Predicted Sources							
	Air.	Car.	Dog.	Dri.	Eng.	Jac.	Sir.	Str.
Air Conditioner	**31.8**	0.4	0.7	2.8	42.2	8.8	0.0	13.3
Car Horn	1.9	**83.0**	0.0	0.0	5.7	0.0	0.0	9.4
Dog Bark	0.7	0.0	**84.8**	3.1	3.3	0.0	2.4	5.7
Drilling	6.4	0.0	2.0	**64.7**	9.7	10.9	0.0	6.3
Engine Idling	7.3	0.0	2.2	6.6	**66.6**	12.5	1.7	3.1
Jackhammer	15.3	0.0	0.0	23.0	7.2	**45.0**	0.0	9.5
Siren	10.3	0.0	15.1	1.2	5.2	0.0	**63.5**	4.8
Street Music	7.0	0.0	5.8	6.1	17.9	1.8	2.6	**58.9**
Average Balanced Accuracy: **62.3%**								

TABLE V

ACOUSTIC SOURCE CLASSIFICATION ACCURACIES (%) OBTAINED WITH 4S AUDIO SIGNALS AND 40 MFCC + 8 INS FEATURES.

Original Sources	Predicted Sources							
	Air.	Car.	Dog.	Dri.	Eng.	Jac.	Sir.	Str.
Air Conditioner	**33.7**	0.0	0.2	2.3	40.3	9.7	0.4	13.4
Car Horn	1.9	**81.1**	0.0	0.0	7.5	0.0	0.0	9.4
Dog Bark	0.5	0.0	**89.1**	1.7	2.4	0.0	1.0	5.5
Drilling	7.3	0.0	0.6	**66.8**	11.1	8.3	0.3	5.7
Engine Idling	10.4	0.0	0.2	6.0	**65.0**	12.4	3.2	2.7
Jackhammer	13.9	0.0	0.0	20.0	7.4	**51.1**	0.0	7.6
Siren	10.3	0.0	8.3	0.8	3.6	0.0	**73.0**	4.0
Street Music	4.3	0.0	4.8	8.0	14.2	2.2	2.4	**64.0**
Average Balanced Accuracy: **65.5%**								

ing scheme of reduced information.

Results obtained for signals of 4 seconds duration are presented in TABLES IV and V. In this scenario, the highest average balanced accuracy of 65.5% is achieved for MFCC + 8 INS, which corresponds to a 3.2 p.p. increment over the baseline. Similar to the previous condition, the proposed feature is able to increment the classification accuracy for most acoustic sources. The highest individual accuracy gain of 9.5 p.p. is noted for Siren. Sources Jackhammer, Street Music and Dog Bark reached increments of 6.1 p.p., 5.1 p.p. and 4.3 p.p. These results reinforces that the propose INS-based feature is able to incorporate meaningful information to non-stationary acoustic source classification.

IV. CONCLUSION

In this work, a new combined feature MFCC+INS was proposed to improve the recognition of non-stationary acoustic sources. The solution incorporated the INS into the classical MFCC, as a feature vector able to assess the non-stationary behavior of acoustic signals in a frame-by-frame basis. Experiments demonstrated that the new strategy outperforms the baseline systems for the majority of individual acoustic sources, leading to significant increment in the average balanced accuracy in all scenarios. Moreover, a single INS feature value was sufficient to obtain significant gain on the classification accuracy compared to the baseline solution.

ACKNOWLEDGMENT

The results presented in this paper entitled “Non-Stationarity Objective Assessment for Acoustic Source Classification" have been developed as part of a project at SiDi, financed by Samsung Eletrônica da Amazônia 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] 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.

[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)*. IEEE, 2021, pp. 21–25.

[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] 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.

[7] Kirsten Carola Wagener, Martin Hansen, and Carl Ludvigsen, “Recording and classification of the acoustic environment of hearing aid users,” *Journal of the American Academy of Audiology*, vol. 19, no. 04, pp. 348–370, 2008.

[8] Luca Turchet, György Fazekas, Mathieu Lagrange, Hossein S. Ghadikolaei, and Carlo Fischione, “The internet of audio things: State of the art, vision, and challenges,” *IEEE Internet of Things Journal*, vol. 7, no. 10, pp. 10233–10249, 2020.

[9] 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.

[10] Guilherme Zucatelli and Ricardo Barioni, “A metric learning based solution for non-stationary acoustic source classification,” in *XL Brazilian Conference of Telecommunication and Signal Processing*. SBrT, 2022.

[11] 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.

[12] 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.

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