Análise e Classificação de Múltiplas Cenas Acústicas Urbanas

M. Alves e R. Coelho

***Resumo*— Este trabalho apresenta uma análise detalhada e classificação das seguintes cenas acústicas urbanas: *Escritório, Parque, Rua Tranquila, Restaurante* e *Estação de Metrô*. A análise inclui a investigação do índice de não-estacionariedade (*index of nonstationarity* – INS) seguida da distância de Bhattacharyya ($\mathbf{B}_d$) que é empregada como medida de separabilidade destes sinais. Posteriormente, a classificação dessas cenas é realizada com diferentes atributos: Coeficientes Mel Cepstrais (MFCC), Parâmetros de Fourier (FP) e *Matching Pursuit* (MP). O atributo que possiu o melhor desempenho de classificação foi a fusão entre o MFCC e o FP. Os resultados demonstram que a partir do INS e $\mathbf{B}_d$ é possível ser feita uma pré-classificação das cenas coerente com a classificação feita com os diferentes atributos.**

***Palavras-Chave*— Cenas Acústicas, Índice de Não Estacionaridade, Classificação, Atributos.**

***Abstract*— This work presents a deep analysis and classification of the following urban acoustic scenes: *Office, Park, Quiet Street, Restaurant* and *Tube Station*. The analysis includes the investigation of the Index of Nonstationarity – INS followed by the Bhattacharyya distance (B_d) that is employed as a separability measure of these signals. Latter, the classification of these scenes is done with different features: Mel-Cepstral Coefficients (MFCC), Fourier Parameters (FP) and Matching Pursuit (MP). The feature that had the best classification performance was the fusion between MFCC and FP. The results show that from the INS and B_d it is possible pre-classify the scenes consistently with the classification made with the different attributes.**

***Keywords*— Acoustic Scenes,Index of Non-Stationarity, Classification, Features.**

I. Introdução

Recentemente, a classificação de cenas acústicas urbanas tornou-se um tópico muito atraente para a área de pesquisa [7],[1]. A análise apurada das cenas acarreta um aprimoramento do desempenho de diversas aplicações como, por exemplo, em sistemas de monitoramento de idosos [14] ou de reconhecimento de ambientes para navegação de robôs [15]. Em outras palavras, equipamentos podem mudar suas configurações e até mesmo reconhecer situações de risco ao identificar o ambiente ao seu redor pelo som que ele produz.

O principal desafio da pesquisa se refere à multiplicidade de sinais componentes de cada cena acústica. Cada cena é composta por um conjunto fontes acústicas que possuem comportamento variado. Além disso, cenas de um mesmo ambiente, por exemplo, um Restaurante, podem conter distintos sinais de diferentes fontes e continuar sendo considerado um Restaurante. Portanto, uma análise detalhada é necessária para a discriminação destas componentes e consequentemente obter uma melhor acurácia na classificação das cenas.

M. Alves, mestranda do Programa de Pós-Graduação em Engenharia Elétrica, Instituto Militar de Engenharia (PPgEE/IME) e bolsista do CNPq; R. Coelho, Laboratório de Processamento de Sinais Acústicos, Instituto Militar de Engenharia (IME); E-mails: marilia.alves@ime.eb.br, coelho@ime.eb.br.
Este trabalho foi parcialmente financiado pelo CNPq/307866/2015-7.

Nos últimos anos, A sequência de desafios *Detection and Classification of Acoustic Scenes and Events* (D-CASE) [1],[3] foi criada com o intuito estimular a atenção dos cientistas para a classificação de cenas acústicas. Por outro lado, existem pesquisadores que utilizam outras abordagens para classificação das cenas acústicas, como é o exemplo de [4] que propõe um atributo baseado em arranjo de microfones distribuídos ou em [5] que apresenta um atributo de tempo-frequência.

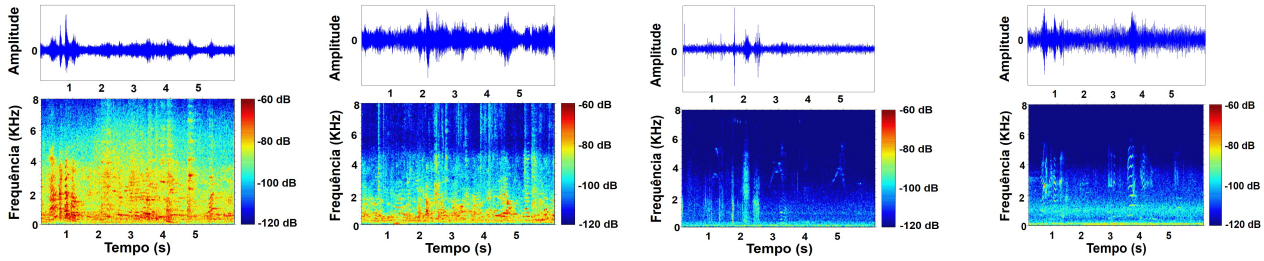
Primeiramente, neste artigo é desenvolvida uma análise aprofundada de cada cena acústica. Esta análise tem o intuito de estudar de forma detalhada os diversos comportamentos temporais e espectrais de cada cena. Esta fase é dividida em duas partes, primeiro é feito um estudo do grau de estacionaridade de cada cena utilizando o Índice de Não Estacionaridade (INS) [10],[17],e então classificando em : Moderamente Não-Estacionária, Não-Estacionária ou Altamente Não-Estacionária. Posteriormente, a distância de Bhattacharyya (B_d) [12] é calculada para quantificar o grau de separação entre as cenas. A partir dessas duas medidas, é feita uma pré-classicação entre as cenas acústicas urbanas.

A segunda fase é classificação das cenas acústicas urbanas utilizando diferentes atributos. Primeiramente é utlizado *Mel-Cepstral Coefficients* (MFCC) e *Gaussian Mixture Model* (GMM) como *baseline* para a classificação [1],[9],[18],[19]. Depois foram propostos dois atributos para esta aplicação: *Fourier Parameters* (FP) que em [8] foi apresentado para a classificação de emoções, e o *Matching Pursuit* (MP) que em [5] foi usado como atributo de tempo-frequência para o reconhecimento sons ambientais.

O restante deste trabalho está organizado da seguinte forma: Na Seção II são apresentadas as principais técnicas utilizadas para a análise das cenas acústicas urbanas: INS e Distância de Bhattacharyya. Em seguida, na Seção III é feito uma breve descrição dos atributos (FP, MFCC e MP) e classificador (GMM) que serão utilizados para a classificação das cenas. Na Seção IV são apresentados os resultados obtidos nos experimentos realizados. Por fim, a Seção V conclui este trabalho.

II. Análise de Cenas Acústicas

Nesta Seção, é apresentada a análise das cenas acústicas urbanas. Na Fig. 1 é possível observar os sinais e espectogramas de algumas cenas acústicas selecionadas da base de dados. Nota-se pelos espectogramas que as cenas possuem comportamento não estacionário e que existe uma grande variabilidade entre elas, mesmo em cenas rotuladas da mesma forma. Sendo assim, neste estudo é proposto o INS e a distância B_d para a análise das cenas.



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

Fig. 1. Emxemplo de de cenas acústicas urbanas e seus respectivos espectrogramas: (a) Restaurante [1], (b) Restaurante [2] , (c) Parque [1] e (d) Parque [2].

A. *INS*

O INS é uma medida tempo-frequência que analisa objetivamente a não estacionariedade de um sinal [10]. Seus autores definem um sinal como sendo estacionário em relação a uma escala de observação seu espectro local de tempo curto em diferentes instantes de tempo for estatisticamente similar ao seu espectro global. O teste de estacionariedade é realizado pela comparação de componentes espectrais do sinal com referenciais estacionários, chamados surrogates, obtidos do próprio sinal. Para tanto, os espectrogramas do sinal e dos surrogates são obtidos por meio da Transformada de Fourier de Tempo Curto (*Short Time Fourier Transform* – STFT). Então, a distância Kullback-Leibler (KL) [11] é usada para medir a divergência entre o espectro de curto tempo do sinal analisado e seu espectro global. Finalmente, o INS á dado pela razão entre esta distância e a KL correspondente aos valores obtidos dos referenciais estacionários. Em [10], os autores consideram que a distribuição do valores da KL são aproximados por uma distribuição Gamma. Por isso, um limiar γ, com 95% de precisão, pode ser definido para o teste de estacionariedade para cada janela de tempo Th. Desta forma, o sinal é considerado não estacionário se o valor de INS estiver acima deste limiar. Ou seja,

$$INS \begin{cases} \leq \gamma, & \text{x(t) é estacionário;} \\ > \gamma, & \text{x(t) é não-estacionário.} \end{cases} \quad (1)$$

B. *Distância de Bhattacharyya*

A Distância de Bhattacharyya (*Battacharyya distance – Bd*) foi proposta em [12], inicialmente para aplicações de seleção de sinais. Esta é uma medida da dissimilaridade entre duas distribuições de probabilidade, $p_1(x)$ and $p_2(x)$. Assim, Bd pode ser definida como

$$Bd = -ln \int \sqrt{p_1(x)p_2(x)}dx, \quad (2)$$

A distância Bhattaharyya é utilizada para comparação entre as distribuições das amostras das cenas acústicas urbanas.

III. Classificação

Nessa Seção, primeiramente, são descritos brevemente os atributos MFCC, FP e MP. Por último, o GMM, que é adotado como classificador.

A. *MFCC*

Para a obtenção de cada atributos o áudio deve passar por uma fase de pré-processamento e então ser dividido em quadros de curta duração de tempo . A extração dos atributos é realizada em cada quadro resultando em uma matriz de atributos para cada áudio. No caso dos Coeficientes Mel Cepstrais [13], inicialmente é realizada transformada rápida de Fourier (FFT) em cada quadro para então ser convertido para a escala Mel. Considerando K o número de filtros no banco de filtros na frequência Mel e E_j a saída de log-energia do filtro de ordem j, os coeficientes MFCC são calculados como

$$MFCC_i = \sum_{j=1}^{J} E_j cos\left[i\left(j-\frac{1}{2}\right)\frac{\pi}{K}\right], i = 1, 2, ..., D \quad (3)$$

,onde D é o número de coeficientes cepstrais.

B. *FP*

Outro atributo que é utilizado para caracterização de sinais acústicos é o Parâmetro de Fourier (FP) que foi proposto em [8] para a reconhecimento de emoções através da voz. Na análise de Fourier, um sinal $x(t)$ é decomposto em suas harmônicas pelo vetor $X[k]$. Para um conjunto de M harmônicas de $x(t)$, o vetor de atributos FP é dado por:

$$FP = [X_1[k], X_2[k], ..., X_p[k]], 1 \leq p \leq M \quad (4)$$

Desda forma o FP se comporta como atributo local (vetor de atributos para cada quadro). Porém, ele também pode ser utilizado com atributo global quando é extraído características estatísticas do vetor FP como média, mediana, valor máximo, valor mínimo ou desvio padrão.

C. *MP*

O *Matching Pursuit* (MP) é um algoritmo utilizado para obter um representação esparsa de sinal baseada em átomos que compõem um dicionário mais que completo [6]. Dado o sinal x e o dicionário $D = [d_1, d_2, ...]$, o MP obtém uma representação esparsa de x em D seguindo os seguintes passos:

- Inicializar o resíduo da iteração 0 por $R^0x = x$, onde R^i é o resíduo na iteração i
- De $t = 1$ até T
  - Selecionar o átomo como maior produto interno com o resíduo por

$$d_t = max < R^{t-1}x, d_i > . \quad (5)$$

  - Atualizar o resíduo

$$R^t = R^{t-1}x - \alpha_t d_t, \quad (6)$$

onde $\alpha_t =< R^{t-1}x, d_t >$ é o coeficiente de projeção de $R^{t-1}x$ em d_t

- A projeção de x em D é dada por

$$\hat{x} = \sum_{i=1}^{T} \alpha_i d_i \quad (7)$$

Um possível critério de parada para esse algoritmo é fixar o número de iterações (átomos). Outra possibilidade é utilizar a energia do sinal residual, onde a o algoritmo para quando $||R^{t-1}x||^2 <$ Limiar.

A utilização do MP foi proposta em [5] como atributo para o reconhecimento de sons ambientes, para isso foi utilizado um dicionário em tempo-frequência construído a partir da funções de Gabor (chamados de átomos de Gabor) que consiste em:

$$g_{s,u,\omega,\theta} = \frac{K_{s,u,\omega,\theta}}{\sqrt{s}} e^{-\pi(n-u)^2/s^2} cos[2\pi\omega(n-u)+\theta], \quad (8)$$

onde s, u, ω e θ são os fatores de escala, localização, frequência e fase de cada átomo, respectivamente, e $K_{s,u,\omega,\theta}$ é a constante de normalização para que $||g_{s,u,\omega,\theta}||^2 = 1$.

Os seguintes parâmetro foram escolhidos s: $s = 2^p(1 < p < 8)$,$u = [064128192]$, $\omega = Kj^{2,6}$(com $1 \leq\leq 35, K = 0,5x35^{2,6i}$ de modo que ω varia entre 0 e 0,5),e $\theta = 0$, com cada átomo de tamanho $N = 256$ assim como os quadros do sinal que tem tamanho de 256 amostras (11.6 ms) a uma frequência de amostragem de 22,05 kHz.

A precisão da classificação não é muito afetada para quantidade de átomos muito grande ($T > 5$) então o número de átomos é fixado em $T = 5$ átomos como critério de parada do algoritmo. Os parâmetros selecionados são a média e a variância dos parâmetros de escala e frequência dos 5 átomos selecionados, isto é, $[\mu_s, \mu_\omega, \sigma_s, \sigma_\omega]$ são referidos como os atributos MP. Os parâmetros de localização e fase são ignorados.

D. GMM

O classificador escolhido para a classificação das cenas acústicas foi o Modelo de Mistura Gaussiana (GMM) proposto inicialmente por [9]. O GMM (λ_C) de cada cena C é definida como a combinação linear das componentes Gaussianas

$$p(\vec{x}|\lambda_C) = \sum_{n=1}^{M} p_n b_n(x) \quad (9)$$

onde x é um vetor de atributos de dimensão D, p_n são os pesos das misturas, onde $\sum_{n=1}^{M} p_n = 1$ e $b_n(x)$ são as densidades das Gaussianas com vetor de médias μ_n e matriz covariância K_n. Como isso, o GMM de cada cena pode ser parametrizado como $\lambda_C = [p_n, \mu_n, K_n | n = 1, ..., M]$.

Na fase de treinamento do classificador, os parâmetros de λ_C são estimados para maximizar a função de verossimilhança

$$p(X|\lambda_C) = \prod_{t=1}^{Q} p(x_t|\lambda_C) \quad (10)$$

onde a matriz de atributos X é composta por Q vetores de atributos x_t extraídos de cada quadro do segmento de treinamento disponível para cena C. Para os testes, a regra de decisão classificação é baseada no critério máxima log-verossimilhança [9]. Isto significa que o cena identificada é aquele que maximiza a soma

$$\hat{C} = argmax_C \sum_{t=1}^{Q} logp(x_t|\lambda_C). \quad (11)$$

IV. Resultados e Discussão

Diversos experimentos foram realizados utilizando um subconjunto da base de dados disponibilizada pelo D-CASE/2013 [1]. O D-CASE é um desafio que vem sendo amplamente referenciado na literatura de processamento de sinais. Ele consiste em duas tarefas, a classificação de cenas acústicas e detecção de eventos acústicos. Um banco de dados de 10 diferentes cenas acústicas com 20 gravações por cena é disponibilizado para classificação, cada uma das gravações possui 30 segundos duração e frequência de amostragem de $F_s = 44,1$KHz. Neste trabalho foram escolhidas 5 cenas acústicas desta base, estas cenas foram rotuladas em: *Escritório, Parque, Rua Tranquila, Restaurante* e *Estação de Metrô*.

A. Análise do INS versus Espectrogramas

Os valores de INS de um trecho de 30 segundos cada cena acústica pode ser observado nas Fig. 2, seguidos pelos seus respectivos espectogramas na Fig. 3. A escala T_h/T refere-se a razão entre as janelas de observação (T_h) do INS e o tamanho todo (T) do sinal. Os pontos em verde representam o limiar de estacionariedade, enquanto que os pontos em vermelho representam os valores de INS encontrados em cada escala de tempo.

A partir dos valores dos picos dos gráficos de INS (INS_p) é possível classificar as cenas de acordo com seu comportamento em relação a estacionariedade. Para as cenas estudadas é possível dividir em 3 classes:

- Moderamente Não-Estacionário ($\gamma \leq INS_p < 40$): As cenas *Parque* (b) e *Rua Tranquila* (c) são enquadradas nesta classe.
- Não-Estacionário ($40 \leq INS_p < 100$): Apenas a cena *Escritório*(a) pertence a esta classe.
- Altamente Não-Estacionário ($INS_p \geq 100$): *Restaurante* (d) e *Estação de Metrô* (e) fazem parte desta última classe.

B. Histogramas e B_d das cenas urbanas

Os histogramas das cenas analisadas podem ser vistos na Fig. 4. Observe que a cena *Escritório* se diferencia das demais cenas pela concentração elevada de ocorrências na origem e baixa variância entre seus valores. Já as cenas *Rua Tranquila* e *Restaurante* se destacam pela sua alta variância.

A separabilidade entre os histogramas das cenas urbanas pode ser medida pela distância Bhattaharyya (B_d). A TABELA I apresenta as B_ds entre todas cenas utilizadas. Nota-se que a cena que mais se separa das demais é *Escritório*. Por outro lado, os menores valores de B_d são *Parque* $\longleftrightarrow$ *Rua Tranquila* e *Restaurante* $\longleftrightarrow$ *Estação de Metrô*.

Uma pré-classificação estocástica das cenas acústicas urbanas é realizada a partir da análise do grau de estacionaridade de cada cena e da B_d entre elas , como é mostrado na TABELA II .

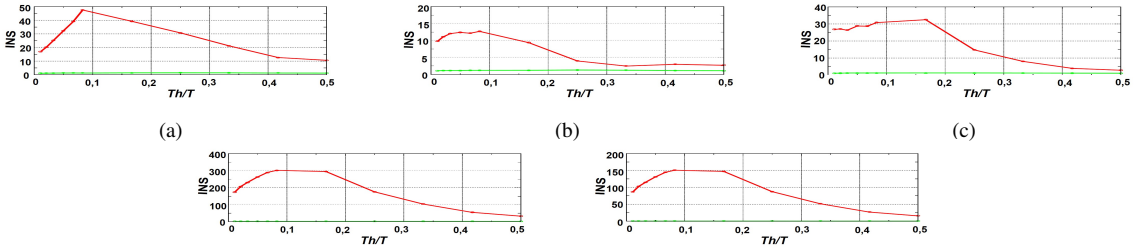


Fig. 2. Valores de INS das cenas acústicas. (a) Escritório, (b) Parque, (c) Rua Tranquila, (d) Restaurante, (e) Estação de Metrô.

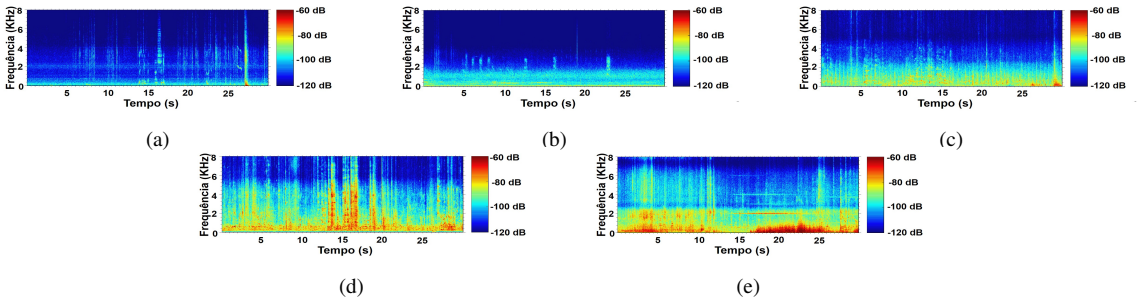


Fig. 3. Espectrograma dos sinais de cenas acústicas.

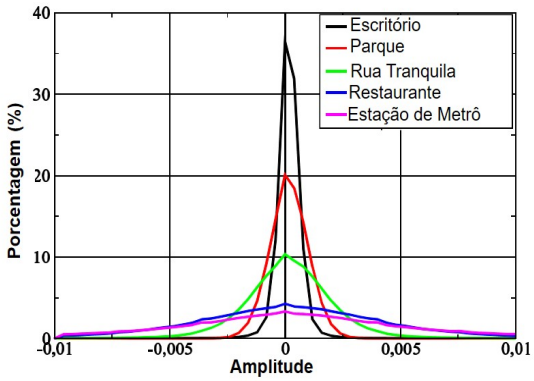


Fig. 4. Histogramas das cenas acústicas urbanas

TABELA I

RESULTADOS DA DISTÂNCIA BHATTAHARYYA (B_d) ENTRE AS CENAS ACÚSTICAS URBANAS

Cenas Acústicas	*Bd*
Escritório⟷ Restaurante	1,33102
Escritório⟷ Estação de Metrô	1,06880
Escritório⟷ Rua Tranquila	0,31752
Escritório⟷ Parque	0,27654
Rua Tranquila⟷ Estação de Metrô	0,50224
Rua Tranquila⟷ Restaurante	0,75395
Parque⟷ Estação de Metrô	0,54426
Parque⟷ Restaurante	0,79747
Parque⟷ Rua Tranquila	0,00271
Restaurante⟷ Estação de Metrô	0,06807

TABELA II

PRÉ-CLASSIFICAÇÃO CENAS URBANAS ACÚSTICAS

Classe	Segundo INS_p	*Bd* mínima
Parque e *Rua Tranquila*	Moderadamente Não-Estacionário	0,00271
Restaurante e *Estação de Metrô*	Altamente Não-Estacionário	0,06807
Escritório	Não-Estacionário	0,00000

V. CLASSIFICAÇÃO DAS CENAS ACÚSTICAS URBANAS

No experimento foram utilizados 5 cenas acústicas urbanas que podem ser rotuladas como *Escritório, Parque, Rua Tranquila, Restaurante* e *Estação de Metrô*. Primeiramente, os áudios foram re-amostrados para uma frequência de amostragem de 22,05KHz. Foram utilizados 100 segmentos de 4 segundos para cada cena, e então foi utilizado o método de 4-*fold cross-validation* na divisão do /textitdataset entre teste e treino, onde 3 conjuntos foram usados para treino e 1 conjunto foi utilizado para teste. Esta separação de segmentos em treino e teste foi feita de forma que arquivos segmentados do mesmo áudio de origem não fossem usados simultaneamente para teste e treino. Os atributos foram calculados a partir de janelas retangulares de 11,6ms (256 amostras) com sobreposição de 50% [5].

Os atributos descritos na Sessão III foram utilizados de maneira isolada e também em conjunto. A Fig. 5 apresenta o desempenho desses atributos: MFCC (12), FP (40), MFCC+FP (12+40), MP (4), MFCC+MP (12+4). Em todos os casos foi utilizado GMM com 5 Gaussianas como classificador.

Uma análise mais aprofundada da classificação pode ser observada na TABELA II pela matriz confusão dos 3 atributos que obtiveram a melhor performance (MFCC, MFCC+FP e MFCC+MP). Fica claro pela Fig. 5 e TABELA III que o MP isolado ou combinado com MFCC não obteve um bom desempenho como atributo para classificação das cenas acústicas urbanas utilizadas no trabalho. Por outro lado, o FP também não teve uma acurácia média melhor que o MFCC, porém com a fusão entre os dois atributos (MFCC+FP) houve uma melhora no desempenho, mostrando que o FP agrega valor ao MFCC.

Também foi possível observar a partir da TABELA III que a

maiores confusões estão entre as cenas *Parque* e *Rua Tranquila* e entre *Restaurante* e *Estação de Metrô*, confirmando a pré-classificação da TABELA II gerada a partir dos valores de INS e B_d . Em outras palavras, as cenas na mesma classe possuem características similares em relação a estacionaridade se confundem naturalmente.

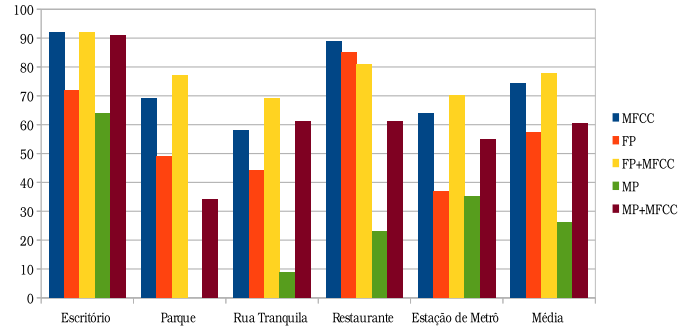


Fig. 5. Taxas de acurácia geral comparando 5 cenas utilizando MFCC, FP, MFCC+FP, MP e MFCC+MP.

TABELA III
MATRIZ CONFUSÃO UTILIZANDO MFCC, MFCC+FP E MFCC+MP COMO ATRIBUTOS PARA CLASSIFICAÇÃO

	Cena Real	Cena Classificada				
		Escrit.	Parq.	Rua	Rest.	Est.M.
MFCC	Escritório	**92**	2	0	6	0
	Parque	5	**69**	19	7	0
	Rua Tranq.	2	27	**58**	5	8
	Restaurante	0	0	0	**89**	11
	Est. de Metrô	5	7	9	15	**64**
	Acurácia Média de Classificação:**74,4%**					
FP+MFCC	Cena Real	Cena Classificada				
		Escrit.	Parq.	Rua	Rest.	Est.M.
	Escritório	**92**	3	5	0	0
	Parque	8	**77**	11	1	3
	Rua Tranq.	5	19	**69**	4	3
	Restaurante	0	0	1	**81**	18
	Est. de Metrô	2	5	6	17	**70**
	Acurácia Média de Classificação:**77,8%**					
FP+MFCC	Cena Real	Cena Classificada				
		Escrit.	Parq.	Rua	Rest.	Est.M.
	Escritório	**91**	3	2	4	0
	Parque	26	**34**	29	8	3
	Rua Tranq.	8	19	**61**	2	10
	Restaurante	5	2	0	**61**	32
	Est. de Metrô	13	6	10	16	**55**
	Acurácia Média de Classificação:**60,4%**					

VI. CONCLUSÃO

Este trabalho apresentou a análise e classificação de sinais acústicos de cinco diferentes cenas acústicas urbanas. Primeiramente a análise das cenas foi feita a partir da estimação do INS, sendo assim possível realizar uma pré-classificação das mesmas. Em seguida foi feito o cálculo da B_d entre cada uma das cenas que confirmou a pré-classificação obtida com o INS. A segunda etapa do trabalho foi a classificação das cenas acústicas utilizando diferentes atributos. O MFCC obteve o melhor desempenho como atributo isolado, porém ao realizar a fusão dele com o FP foi possível obter uma melhora no desempenho total de classificação. A pré-classificação obtida a partir do INS e B_d mostrou a similaridade entre as cenas da mesma classe, fato que foi confirmado pela matriz confusão gerada posteriormente.

REFERÊNCIAS

[1] D. Giannoulis, E. Benetos, D. Stowell, M. Rossignol, M. Lagrange, and M.D.Plumbley, “Detection and classification of acoustic scenes and events:an IEEE AASP challenge,” in Proc. 2013 Applications of Signal Processing to Audio and Acoustics (WASPAA), 2013 IEEE Workshop on, pp. 1–4.

[2] D. Barchiesi, D. Giannoulis, D. Stowell, and M. D. Plumbley, “Acoustic scene classification: Classifying environments from the sounds they produce,” IEEE IEEE Signal Processing Magazine , vol. 32, no. 3, pp. 16–34, May 2015.

[3] A. Mesaros, T. Heittola, and T. Virtanen, ”Tut database for acoustic scene classification and sound event detection”,In 24rd European Signal Processing Conference 2016 (EUSIPCO 2016). Budapest, Hungary, 2016.

[4] K. Imoto, N. Ono, ”Spatial cepstrum as a spatial feature using distributed microphone array for acoustic scene analysis”,IEEE/ACM Trans. Audio Speech Language Process., vol. 25, no. 6, pp. 1335-1343, 2017.

[5] S. Chu, S. Narayanan, and C. C. Kuo, “Environmental sound recognition with time-frequency audio features,” IEEE Trans. Audio, Speech, Lang. Process., vol. 17, no. 6, pp. 1142–1158, Aug. 2009

[6] S. Mallat and Z. Zhang, “Matching pursuits with time–frequency dictionaries,” IEEE Trans. Signal Process., vol. 41, no. 12, pp. 3397–3415, Dec. 1993

[7] D. Wang and G. J. Brown, Computational Auditory Scene Analysis: Principles, Algorithms, and Applications. Piscataway, NJ: IEEE Press, 2006.

[8] K. Wang, N. An, B. N. Li, Y. Zhang, L. Li, ”Speech emotion recognition using Fourier parameters”, EEE Transactions on Affective Computing , vol. 6, no. 1, pp. 69-75, Jan./Mar. 2015.

[9] D. A. Reynolds and R. C. Rose. ”Robust text-independent speaker identification using Gaussian mixture speaker models.” Speech and Audio Processing, IEEE Transactions, 1995

[10] 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, pp. 3459–3470, July 2010.

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

[12] T. Kailath, “The divergence and bhattacharyya distance measures in signal selection,” IEEE Transactions on Communication Technology, vol. 15, no. 1, pp. 52–60, 1967.

[13] S. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” IEEE Trans. Acoust., Speech, Signal Process., vol. ASSP-28, no. 4, pp. 357–366, Aug. 1980

[14] Y. Peng, C. Lin, M. Sun, and K. Tsai, “Healthcare audio event classification using hidden Markov models and hierarchical hidden Markov models,” in Proc. IEEE Int. Conf. Multimedia Expo, 2009, pp. 1218–1221.

[15] S. Chu, S. Narayanan, C.-C. Jay Kuo, and M. J. Matari, “Where am I? Scene recognition for mobile robots using audio features,” in Proc. IEEE Int. Conf. Multimedia and Expo., 2006, pp. 885–888.

[16] R. Radhakrishnan, A. Divakaran, and P. Smaragdis, “Audio analysis for surveillance applications,” in Proc. IEEE Workshop Applications of Signal Processing to Audio and Acoustics (WASPAA), 2005, pp. 158–161.

[17] E. Dranka e R. Coelho, ”Robust Maximum Likelihood Acoustic Energy Based Source Localization in Correlated Noisy Sensing Environments”, IEEE Journal of Selected Topics in Signal Processing, v. 9, n. 2, pp. 259-267, March 2015.

[18] A. Venturini, L. Zão e R. Coelho, ”On Speech Features Fusion, α-Integration Gaussian Modeling and Multi-Style Training for Noise Robust Speaker Classification”, IEEE/ACM Transactions on Audio, Speech and Language Processing, v. 22, n. 12, pp. 1951-1964, December 2014.

[19] R. Sant Ana, R. Coelho and A. Alcaim, ”Text-Independent Speaker Recognition Based on the Hurst Parameter and the Multi-Dimensional Fractional Brownian Motion Model”, IEEE Transactions on Audio, Speech and Language Processing, v. 14, n. 3, pp. 931-940, May 2006.