Monitoramento e classificação sonora em UTI Neonatal usando redes neurais

Igor H. D. Fontes[1], **Arthur M. A. Melo** [1], **Andre L. L. Aquino**[1]

[1]Instituto de Computação – Universidade Federal de Alagoas

{ihdf, amam, alla}@laccan.ufal.br

Abstract. *Neonatal Intensive Care Units (NICUs) are specialized units to treat newborns with health complications. Many factors can influence treatment phases, including noise levels and sound sources. To provide a helpful tool to enable proper monitoring and feedback to medical staff, we performed correct sound classification in NICUs using Convolutional and Long Short-Term Memory (LSTM) Neural Networks. We focus on three audio classes: cry, human talks, and alerts from hospital machines (beep sounds). The results include the extraction of relevant sound features and comparisons between classifiers. State-of-the-art models for environmental sounds achieve, on average, 74.4% classification performance. Using the proposed models, we achieved up to 84% performance using the evaluation metrics.*

Resumo. *As Unidades de Terapia Intensiva Neonatais (UTINs) são unidades especializadas no tratamento de recém-nascidos com complicações de saúde. Muitos fatores podem influenciar as fases do tratamento, os quais incluem níveis de ruído e fontes sonoras. Para fornecer uma ferramenta útil para permitir o monitoramento e feedback adequados à equipe médica, realizamos uma correta classificação sonora em UTIs Neonatais usando redes neurais convolucionais e Long Short-Term Memory. Focamos em três classes de áudio: choro, conversas humanas e alertas de máquinas hospitalares (sons de bipe). Os resultados incluem a extração de features sonoras relevantes e comparações entre classificadores. Modelos do estado da arte para sons ambientais atingem, em média, 74,4% de performance na classificação. Utilizando os modelos propostos, alcançamos um desempenho de até 84% usando as métricas de avaliação.*

1. Introdução

Com o crescente apelo nas unidades neonatais para manter o atendimento o mais humanizado possível, o controle e a identificação do ruído nesses ambientes surgem com grande interesse dos profissionais de saúde e gestores desses estabelecimentos. A exposição a altos níveis de ruído por longos períodos pode resultar em aumento da pressão arterial e da frequência cardíaca.

Esse aspecto ambiental pode desequilibrar o comportamento humano, principalmente em recém-nascidos, em que qualquer alteração é capaz de causar déficits no desenvolvimento neuropsicomotor ao longo da vida. Portanto, é necessário manter o ambiente da UTI Neonatal (UTIN) o mais acolhedor e confortável possível, mantendo os níveis de ruído com o padrão de qualidade esperado e proporcionando um tratamento seguro e

eficaz. Além disso, a análise das fontes sonoras da UTINs é essencial para desenvolver ferramentas para monitorar o ambiente acústico e dar *feedback* à equipe médica.

Quando os recém-nascidos (RNs) vão para a unidade neonatal, encontram um ambiente significativamente diferente do que vivenciaram no ventre materno. Por exemplo, o nível sonoro costuma ser alto e as luzes são intensas e contínuas nas unidades neonatais. A fala dos profissionais durante os procedimentos de solicitação de materiais, exames, orientações comumente ultrapassa os chamados limites sonoros saudáveis, tornando-se ruído.

Os valores recomendados pelos Padrões de Acústica Brasileira [ABNT 2003] são 35-45 dB, que podem ir até 50 dB dependendo da situação. Para obter esse padrões é necessário usar uma abordagem muito abrangente, o que envolve uma mudança física e cultural, além de alterações no design, revestimentos, equipamentos, móveis, rotinas e cuidados com o recém-nascido. Apesar de reconhecer o problema e todas as iniciativas para amenizá-lo, a dificuldade em mensurar a influência desses fatores ao longo do tempo e verificação da eficácia de tratamentos alternativos ainda é um grande obstáculo. Desse modo, o estudo da classificação sonora e da detecção de eventos em ambientes hospitalares corresponde a um passo essencial para o desenvolvimento de um monitoramento acústico eficaz.

Devido à sensibilidade do bebê, níveis sonoros inadequados e eventos acústicos podem prejudicar o tratamento do paciente [Smith et al. 2018]. Para superar esses problemas, o monitoramento e classificação de sons não intrusivos em UTIs Neonatais podem fornecer ferramentas relevantes para avaliar o ruído ambiental e detectar eventos relevantes. No entanto, as abordagens disponíveis na literatura são, em geral, focadas em análises específicas de choro usando vários algoritmos. Além disso, não incluem alertas e caracterização de conversas. Em vez disso, esses estudos exploram a capacidade de distinguir a saúde e as condições emocionais de um bebê usando métodos automáticos [Bănică et al. 2016].

O problema da classificação sonora é explorado em diversos cenários acústicos usando abordagens de aprendizado supervisionado e não supervisionado. No entanto, faltam propostas de classificação sonora para descrever ambientes acústicos sensíveis, como hospitais. Soluções do estado da arte exploram a classificação de sons urbanos e eventos acústicos usando redes neurais e modelos pré-treinados para melhorar a performance dos classificadores. No entanto, o cenário acústico hospitalar ainda é pouco explorado. Nosso trabalho apresenta uma descrição acústica eficiente e uma proposta de classificação sonora em Unidades de Terapia Intensiva Neonatais. Apresentamos um esquema de monitoramento sonoro que inclui propostas de classificação baseadas em redes neurais LSTM e CNN [Lezhenin et al. 2019] que exploram diversas classes de áudio. Construímos um conjunto de dados privado (restrição imposta pelo hospital) composto por amostras sonoras coletadas em uma UTIN real do Hospital Universitário da Universidade Federal de Alagoas para executar os experimentos. Utilizando os modelos propostos, alcançamos um desempenho de até 84% usando as métricas de avaliação. Trabalhos que propõem a classificação de sons urbanos, os quais incluem sons de alerta e conversas, atingem até 74,4%, considerando F1-score e utilizam Log-Mel espectrogramas como entrada para o treinamento dos modelos [Adapa 2019] [Piczak 2015].

2. Trabalhos relacionados

As principais contribuições existentes para a classificação sonora em ambientes hospitalares concentram-se em sons específicos emitidos pelos pacientes. Em trabalhos recentes, diversos autores exploraram a análise e classificação do choro infantil. Ferretti et al. [Ferretti et al. 2018] usa uma abordagem baseada em redes neurais profundas para detectar choros em ambientes acústicos adversos. O método utiliza Log-Mel extraído de áudios como *features* de entrada e apresenta bom desempenho em comparação com outros algoritmos de classificação sonora. No entanto, outras classes sonoras relevantes em UTINs não são incluídas na avaliação.

[Salamon et al. 2014] explorou o uso de *Support Vector Machine* usando como *features* bandas Mel e MFCCs com acurácia de 70%. Devido à falta de conjuntos de dados públicos com sons de UTINs, alguns autores analisaram o desempenho da detecção de choro em conjuntos de dados sintéticos obtidos por simulação de cena acústica [Severini et al. 2019] ou usando dados coletados em campo. Em nosso trabalho, coletamos amostras de áudio de um ambiente real de UTIN, as quais incluem conversas humanas, alertas e amostras de choro. Dessa forma, propusemos a classificação dos sons predominantes em uma UTIN, ao invés de se concentrar em apenas uma classe sonora. Além disso, avaliamos o uso de CNN e LSTM, o que corresponde ao estado da arte no contexto de sons urbanos.

Existem estudos com propostas bem-sucedidas de classificação de sons urbanos [Salamon and Bello 2015], ambientais [Piczak 2015] e de sons de batimentos cardíacos [Tschannen et al. 2016]. [Sang et al. 2018] utiliza redes neurais recorrentes convolucionais com formas de onda brutas para realizar a classificação de sons urbanos. Os resultados mostram melhoria da acurácia em comparação com outras propostas. [Lezhenin et al. 2019] explorou a vantagem do LSTM no aprendizado de dependências temporais para melhorar o desempenho da classificação e superar modelos baseados em redes neurais convolucionais (CNN) em ambientes sonoros urbanos. [Deng et al. 2020] usou MFCCs com variações de CNNs para classificar sons de batimentos cardíacos. Para comparar o desempenho da rede LSTM com outras propostas de classificação, usamos uma rede neural convolucional com coeficientes cepstrais como entrada, que corresponde ao estado da arte em diversos domínios de classificação sonora.[Mushtaq et al. 2021] explora o uso de imagens espectrais e *features* relacionadas ao domínio visual e sonoro para classificar sons ambientais. [Adapa 2019] usa Redes Neurais Convolucionais (CNN) e Mel-espectrogramas para classificar amostras sonoras e apresenta 3 modelos, os quais incluem um modelo de regressão logística para múltiplas classes e 2 modelos baseados em CNNs usando uma forma adaptada de MobileNetV2 [Sandler et al. 2018]. Os modelos obtiveram de 66,4% a 74,5% de performance considerando a métrica F1-score.

3. Materiais e métodos

Usamos dados rotulados para desenvolver um classificador eficiente para sons de UTIs Neonatais com base em descritores sonoros relevantes e redes neurais. Os dados de treinamento e teste incluem amostras de áudio de um conjunto de dados formado por amostras coletadas em uma UTIN. Analisamos classificadores baseados em redes neurais convolucionais (CNN) e *Long Short-Term Memory* (LSTM) e comparamos o desempenho da classificação de eventos sonoros usando as métricas F1-score, precision, recall e acurácia [Hossin and Sulaiman 2015].

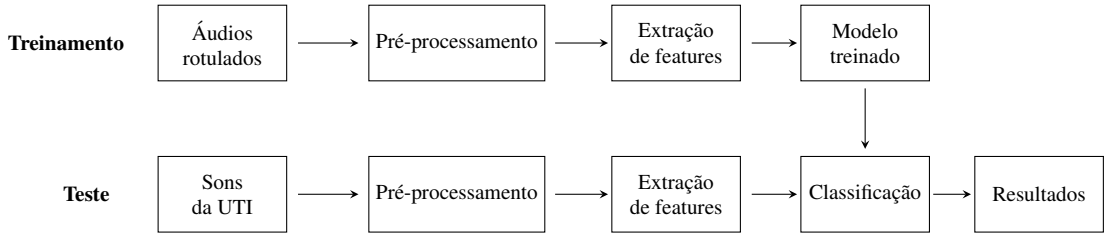

Figura 1. Esquema geral da metodologia

A Figura 1 mostra uma visão geral de nossa abordagem. No *Treinamento*, temos os seguintes passos: 1. *Áudios rotulados* corresponde as amostras de cada classe estudada, ou seja, choro, falas e alertas; 2. *Pré-processamento* realizado sobre cada amostra para uma duração fixa em segundos e em seguida re-amostramos os sinais de áudio para as mesmas frequências; 3. *Extração de features* permite identificarmos as *features* temporais e espectrais para uma distinção precisa entre amostras de diferentes classes de som; 4. *Modelo treinado* com LSTM. Para o *Teste*, temos os seguintes passos: 1. *Sons da UTI* coletados de uma Unidade de Terapia Intensiva Neonatal; 2. *Pré-processamento* onde separamos os áudios em amostras de mesma duração e frequência dos dados do treinamento. Além disso, rotulamos as amostras de teste para permitir a avaliação; 3. *Extração de features* repete a extração de *features* usada anteriormente durante o treinamento. 4. *Classificação* utiliza o modelo treinado anteriormente para classificar os sons coletados da UTIN por meio de diferentes classificadores. 5. *Resultados* avalia o desempenho da classificação de eventos sonoros comparando a LSTM e CNN.

Já que não há conjuntos de dados públicos de áudios rotulados com precisão e, em geral, os áudios disponíveis na Internet são de domínios genéricos, construímos um conjunto de dados com sons de uma UTIN para superar esse problema e testar a classificação adequadamente. Para o monitoramento da UTIN, foi adquirido e instalado um gravador digital profissional Zoom H1N no Hospital Universitário, onde realizamos a fase de monitoramento por 48 horas. Gravamos esses áudios a uma taxa de amostragem de 44,1 kHz, resolução de 16 bits e no formato WAV para oferecer melhor qualidade e resolução. A figura 2 ilustra amostras de som de 2 segundos das três classes modeladas: choro, conversa e alertas. Cada gráfico mostra as diferentes amplitudes da onda sonora ao longo do tempo.

O monitoramento sonoro fornecido pode melhorar as taxas de sucesso no tratamento de prematuros e fornecer *feedback* aos profissionais sobre a acústica do ambiente e o conforto sonoro dentro das salas. Fornecer amostras de áudio rotuladas em ambientes de UTIN pode ajudar a identificar as fontes sonoras que causam níveis de alta intensidade em decibéis (dB) e prejudicam a saúde do recém-nascido. Assim, uma classificação automática é altamente relevante para permitir o tratamento adequado.

Na etapa de *Pré-processamento* (Figura 1), reamostramos a representação do sinal para 22.050 Hz e usamos os áudios rotulados a cada 2 segundos para treinar e testar a classificação. Classificamos os sons da UTIN e comparamos o desempenho das redes neurais LSTM e CNN para classificar as amostras de áudio da UTIN.

Nossa abordagem usa vetores de *features* para treinar o classificador para a rotulação de amostras sonoras. Extraímos informações de amostras de áudio usando *fe-*

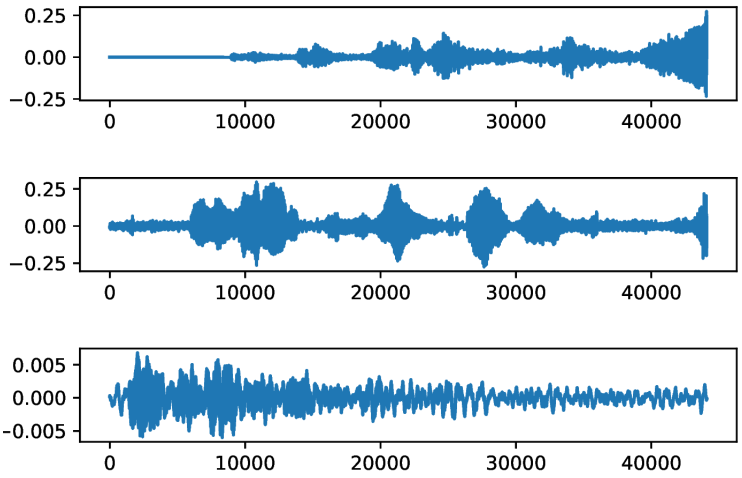

Figura 2. Gráfico da Amplitude vs. tempo - Choro, conversa e alerta

atures sonoras para converter a forma de onda de áudio clássica em uma representação vetorial. As *features* mais comuns são aquelas no domínio do tempo obtidas das formas de onda do áudio bruto. No entanto, convertemos o som do domínio do tempo para a frequência usando a transformada de Fourier e analisamos as propriedades sonoras relevantes para compor os vetores de treinamento do classificador. Selecionamos *features* sonoras entre diferentes descritores de tempo e domínio espectral. Entre as *features* temporais, extraímos a taxa de cruzamento do eixo zero ao longo do tempo (em inglês, ZCR), a raiz da média quadrática (RMS) e o desvio padrão no tempo (STD) [Giannakopoulos and Pikrakis 2014]. Essas características estão relacionadas ao comportamento da magnitude do sinal em cada amostra.

A taxa de cruzamento no zero (ZCR) mede a mudança de magnitude do sinal de negativo para positivo e vice-versa. ZCR é amplamente utilizado para determinar sinais de áudio com ou sem voz e classificar gêneros musicais [Ahrendt et al. 2004]. Por definição, temos:

$$ZCR = \frac{1}{2N} \sum_{k=1}^{N} |sgn(x_k) - sgn(x_{k-1})|, \tag{1}$$

onde N é o número de observações de magnitude em uma dada amostra de som, x é o sinal de áudio e x_k é a magnitude do sinal em um dado instante de tempo k e função sinal é dada por

$$sgn(x_k) = \begin{cases} 1, & \text{if } x_k \geq 0 \\ -1, & \text{if } x_k < 0. \end{cases} \tag{2}$$

Podemos definir a raiz da média quadrática (RMS) do sinal no domínio do tempo como

$$RMS = \sqrt{\frac{\sum_{k=1}^{N} |x_k|^2}{N}}, \tag{3}$$

O RMS é expresso em decibéis (dB) usando a seguinte conversão

$$c_{RMS} = 20\, log(RMS). \tag{4}$$

Também calculamos o STD para investigar a variação na magnitude do sinal. Por definição, temos

$$STD = \sqrt{\frac{\sum_{k=1}^{N} |x_k - \mu|}{N}}, \tag{5}$$

onde μ é a média das magnitudes do sinal, ou seja,

$$\mu = \frac{\sum_{k=1}^{N} |x_k|}{N}. \tag{6}$$

Além das *features* temporais, extraímos *features* espectrais relacionadas ao domínio da frequência para incluir *features* sonoras que os descritores temporais não identificaram. Seja f_k a frequência em Hz correspondente ao bin k, μ_k a frequência média e s_k o valor espectral no bin k, as *features* espectrais extraídas foram [Peeters 2004]:

- **Centróide espectral** que indica o centro de massa do sinal, relacionado às frequências de um som, e é definido por

$$\mu_1 = \frac{\sum_{k=b_1}^{b_2} f_k s_k}{\sum_{k=b_1}^{b_2} s_k}, \tag{7}$$

  onde b_1 e b_2 são as bordas da banda, em bins, usadas para calcular as *features* espectrais;
- **Dispersão espectral** que é o desvio padrão em torno do centróide espectral, indicando a dominância de um som, e é definido por

$$\mu_2 = \sqrt{\frac{\sum_{k=b_1}^{b_2} (f_k - \mu_1)^2 s_k}{\sum_{k=b_1}^{b_2} s_k}}; \tag{8}$$

- **Distorção espectral** que determina a simetria em torno do centroide, e é definida por

$$\mu_3 = \sqrt{\frac{\sum_{k=b_1}^{b_2} (f_k - \mu_1)^3 s_k}{(\mu_2)^3 \sum_{k=b_1}^{b_2} s_k}}; \tag{9}$$

- **Curtose espectral** que mede a planicidade do espectro em torno de seu centróide, indicando o pico de um espectro, e é definida por

$$\mu_4 = \sqrt{\frac{\sum_{k=b_1}^{b_2} (f_k - \mu_1)^4 s_k}{(\mu_2)^4 \sum_{k=b_1}^{b_2} s_k}}; \tag{10}$$

- **Diminuição espectral** que fornece a diminuição da amplitude espectral com base na percepção humana do som, e é definida por

$$decrease = \frac{\sum_{k=b_1+1}^{b_2} \frac{s_k - s_{b_1}}{k-1}}{\sum_{k=b_1+1}^{b_2} s_k}; \tag{11}$$

- **Crista espectral** que mede a razão entre o máximo do espectro e a média aritmética do espectro, e é definida por

$$crista = \frac{\max(s_{k\epsilon[b_1,b_2]})}{\frac{1}{b_2-b_1}\sum_{k=b_1}^{b_2} s_k}; \tag{12}$$

- **Coeficientes Cepstrais de Frequência Mel (MFCCs)** [Dave 2013] que considera o espectro sonoro para cada amostra como 13 MFCCs. Esses coeficientes são obtidos a partir da transformação de frequências usando a escala de Mel [Umesh et al. 1999], que é baseada em logaritmos e aproxima a percepção do som por um ser humano com base em um mapeamento de frequências para o tom percebido relacionado. A conversão das frequências da escala Hertz para Mel é dada por

$$Mel(f) = 2595\, log\left(1 + \frac{f}{700}\right), \tag{13}$$

onde f é a frequência. Adotamos a média do terceiro MFCC como traço perceptivo para caracterizar as classes sonoras. Isso facilita a detecção de *features* relevantes na faixa de intensidade típica para os sons presentes em uma UTI.

4. Resultados

Para avaliar o desempenho dos classificadores, usamos um processo de validação cruzada de 5 partições (*5-fold cross validation*). Cada partição de áudios contém amostras de som de todas as classes modeladas. Distribuímos as amostras de cada classe uniformemente nas partições para balancear os dados usados para teste e treinamento.

Treinamos os modelos usando 4 partições e usamos a partição restante como dado de teste. O processo foi repetido até que cada partição fosse testada. A distribuição das classes sonoras no conjunto de dados utilizado foram: choros 836; conversas 1.328; e alertas 1.020. Essa distribuição revela as conversas como a principal fonte sonora em UTINs, com 41,7% das amostras. Outra classe relevante é a dos sons de alerta, que aparece com uma contribuição de 32,03%. A ocorrência desses tipos de som são frequentes ao longo do dia, enquanto choros acontecem em um menor intervalo de tempo. Uma verificação frequente das contribuições das fontes sonoras pode facilitar a análise da eficiência de métodos adotados para redução de ruído.

O modelo usado para classificação é baseado numa rede neural LSTM, a qual corresponde a um tipo particular de Rede Neural Recorrente (RNN) e é composto por 2 camadas LSTM com 64 unidades em cada e uma camada densa em que usamos a função de ativação *softmax*. Para a entrada do modelo, utilizamos vetores de *features* conforme descrito na seção 3. Tais *features* englobam tanto características temporais quanto espectrais.

Para fins de comparação, implementamos um classificador sonoro baseado em Redes Neurais Convolucionais (CNN). Executamos 50 épocas usando como entrada 64 MFCCs extraídos das mesmas amostras de áudio usadas para treinar nosso classificador. Para o treinamento, usando a função de perda entropia cruzada categórica, a qual é minimizada usando o otimizador Adam.

Para extrair diversas *features* sonoras e processar as amostras de áudio, adotamos as bibliotecas `librosa` [McFee et al. 2015] e `pyACA` [Lerch 2012]. Cada biblioteca inclui ferramentas úteis para processamento, análise e visualização de áudio em diferentes domínios de som. CNN e LSTM foram implementados usando *Keras*, uma API de rede neural, e correspondem ao estado da arte da classificação sonora para sons urbanos. Nesse trabalho, comparamos a performance desses classificadores no contexto de sons hospitalares. As métricas que usamos na avaliação são precision, recall, F1-score e acurácia.

A análise a seguir apresenta uma avaliação exaustiva dos classificadores considerando as classes sonoras individualmente. A tabela 1 mostra os valores das métricas obtidas usando LSTM e CNN. A classificação das amostras de sons de choro alcançou o melhor desempenho em comparação com as demais classes de sons. Tanto a maior disponibilidade de amostras sonoras para treinamento quanto a capacidade das *features* sonoras escolhidas distinguirem o som humano favorecem a classificação correta. Na comparação entre os classificadores, a CNN supera a LSTM na detecção de choros quanto às métricas F1-score e recall, enquanto atingem valores semelhantes de precision. Isso indica a ocorrência de poucos falsos positivos nesse tipo de classe e destaca a capacidade das CNNs de classificação de sons humanos.

Tabela 1. Resultados da avaliação de métricas

Classe	**CNN**			**LSTM**		
	precision	*recall*	*F1-score*	*precision*	*recall*	*F1-score*
Alertas	0.6071	0.5833	0.5950	0.5670	0.6421	0.6022
Choro	0.8356	0.7439	0.7870	0.8403	0.6097	0.7067
Conversas	0.6068	0.6666	0.6353	0.6773	0.7234	0.6996

Considerando as 3 métricas utilizadas, a classificação de conversas atingiu até 72.32% no recall. Já para os alertas, temos um desempenho inferior em comparação com os demais tipos, o que pode ter ocorrido pela grande variabilidade de intensidade e tipos de alerta em UTIs, o que dificulta um treinamento eficaz. Levando em conta a acurácia para a classificação de conversas, atingimos 72.34% e 66.67% com LSTM e CNN, respectivamente, o que mostra uma vantagem da LSTM na detecção desse tipo de som em relação ao total de amostras.

Por fim, os resultados também mostram uma melhoria considerável obtida usando LSTM para a classificação de alertas e conversas. A vantagem da LSTM está na identificação de dependências temporais de forma mais eficiente do que as CNNs, o que pode favorecer em determinadas classes sonoras. Dessa forma, validamos a utilização de redes neurais no monitoramento e classificação de sons em UTINs com a identificação e caracterização das fontes sonoras nesse tipo de ambiente.

5. Conclusão e trabalhos futuros

Nossa abordagem forneceu uma classificação eficiente dos sons da UTIN com foco na detecção de choro, fala humana e alertas. Os dados usados para treinar e testar o modelo contêm amostras de dados coletados em campo num ambiente real. Além disso, realizamos extração de *features* sonoras de domínios de áudio espectrais e temporais. Final-

mente, com o objetivo de investigar várias técnicas de classificação, avaliamos o desempenho do uso de LSTM e CNN para classificação com acurácia, precision, recall e F1-score como métricas de avaliação. Para trabalhos futuros, podemos explorar a classificação de sons de manipulação de medicamentos e movimentação de mobília, que aparecem com menor frequência em UTINs. Além disso, incluir a detecção de eventos com múltiplos sons simultaneamente.

A avaliação de técnicas alternativas de aprendizagem profunda também pode ser útil para inserir novas classes de sons para classificação. Além disso, a criação de um conjunto de dados sintético e público de sons de UTIN pode ajudar nos esforços futuros para o monitoramento sonoro hospitalar de baixo custo. Por fim, investigar as *features* sonoras mais relevantes em ambientes hospitalares também é interessante para filtrar descritores de áudio decisivos para cada classe sonora e melhorar o desempenho dos algoritmos. Pretendemos testar classificadores alternativos baseados em redes neurais que identifiquem semelhanças mais complexas entre *features* sonoras e obtenham melhor performance na classificação sonora.

Referências

ABNT (2003). Nbr 10151: Acoustics - evaluation of noise in inhabited areas aiming the comfort of the community - procedure. *ABNT*, 1(1):1–4.

Adapa, S. (2019). Urban sound tagging using convolutional neural networks. *arXiv preprint arXiv:1909.12699*.

Ahrendt, P., Meng, A., and Larsen, J. (2004). Decision time horizon for music genre classification using short time features. In *2004 12th European Signal Processing Conference*, pages 1293–1296, Vienna, Austria. IEEE, IEEE.

Bănică, I.-A., Cucu, H., Buzo, A., Burileanu, D., and Burileanu, C. (2016). Automatic methods for infant cry classification. In *2016 International Conference on Communications (COMM)*, pages 51–54, Bucharest, Romania. IEEE, IEEE.

Dave, N. (2013). Feature extraction methods lpc, plp and mfcc in speech recognition. *International journal for advance research in engineering and technology*, 1(6):1–4.

Deng, M., Meng, T., Cao, J., Wang, S., Zhang, J., and Fan, H. (2020). Heart sound classification based on improved mfcc features and convolutional recurrent neural networks. *Neural Networks*, 130:22–32.

Ferretti, D., Severini, M., Principi, E., Cenci, A., and Squartini, S. (2018). Infant cry detection in adverse acoustic environments by using deep neural networks. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pages 992–996, Rome, Italy. IEEE, IEEE.

Giannakopoulos, T. and Pikrakis, A. (2014). *Introduction to audio analysis: a MATLAB® approach*. Academic Press, Oxford, UK.

Hossin, M. and Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. *International journal of data mining & knowledge management process*, 5(2):1.

Lerch, A. (2012). *An introduction to audio content analysis: Applications in signal processing and music informatics*. Wiley-IEEE Press, Hoboken, New Jersey.

Lezhenin, I., Bogach, N., and Pyshkin, E. (2019). Urban sound classification using long short-term memory neural network. In *2019 federated conference on computer science and information systems (FedCSIS)*, pages 57–60. IEEE.

McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., and Nieto, O. (2015). librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, volume 8, pages 18–25, Austin, Texas. Citeseer, Citeseer.

Mushtaq, Z., Su, S.-F., and Tran, Q.-V. (2021). Spectral images based environmental sound classification using cnn with meaningful data augmentation. *Applied Acoustics*, 172:107581.

Peeters, G. (2004). A large set of audio features for sound description (similarity and classification) in the cuidado project. *CUIDADO Ist Project Report*, 54(0):1–25.

Piczak, K. J. (2015). Environmental sound classification with convolutional neural networks. In *2015 IEEE 25th international workshop on machine learning for signal processing (MLSP)*, pages 1–6, Boston, USA. IEEE, IEEE.

Salamon, J. and Bello, J. P. (2015). Unsupervised feature learning for urban sound classification. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 171–175, South Brisbane, QLD, Australia. IEEE.

Salamon, J., Jacoby, C., and Bello, J. P. (2014). A dataset and taxonomy for urban sound research. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 1041–1044.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520.

Sang, J., Park, S., and Lee, J. (2018). Convolutional recurrent neural networks for urban sound classification using raw waveforms. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pages 2444–2448. IEEE.

Severini, M., Ferretti, D., Principi, E., and Squartini, S. (2019). Automatic detection of cry sounds in neonatal intensive care units by using deep learning and acoustic scene simulation. *IEEE Access*, 7:51982–51993.

Smith, S. W., Ortmann, A. J., and Clark, W. W. (2018). Noise in the neonatal intensive care unit: a new approach to examining acoustic events. *Noise & health*, 20(95):121.

Tschannen, M., Kramer, T., Marti, G., Heinzmann, M., and Wiatowski, T. (2016). Heart sound classification using deep structured features. In *2016 Computing in Cardiology Conference (CinC)*, pages 565–568, Vancouver, BC, Canada. IEEE, IEEE.

Umesh, S., Cohen, L., and Nelson, D. (1999). Fitting the mel scale. In *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No. 99CH36258)*, volume 1, pages 217–220, Phoenix, AZ, USA. IEEE, IEEE.