An acoustic scene classification approach involving domestic violence using machine learning

Uma abordagem de classificação de cenas acústicas envolvendo violência doméstica utilizando aprendizado de máquina

Helton Santa Cruz Souto[1], Rafael Ferreira Mello[2], Ana Paula C. Furtado[1,2]

[1] Centro de Estudos e Sistemas Avançados do Recife (CESAR SCHOOL)
AV. Cais do Apolo, 77 – Recife, PE – Brasil

[2] Universidade Federal Rural de Pernambuco (UFRPE)
Recife, PE – Brasil

{hscs}@cesar.school, {rafel.mello,anapaula.furtado}@ufrpe.br

Abstract. *Classifying and detecting acoustic scenes is a rapidly developing area of research, as the signal produced by the sound of audio contains information that visual data cannot represent. In this paper we deal with the problem of detecting acoustic scenes involving domestic violence. To this end, we propose the use of a machine learning method using the SVM classifier to detect scenes of domestic violence of a man against a woman using sound. We present analysis of experiments with three different features extracted from the audios. As a result, we obtained the best performance using the MFCC feature achieving an accuracy of 73.14%.*

Resumo. *A classificação e detecção de cenas acústicas é uma área de pesquisa em rápido desenvolvimento, pois o sinal produzido pelo som de um áudio contém informações que dados visuais não podem representar. Neste artigo lidamos com o problema de detecção de cenas acústicas envolvendo violência doméstica. Para tanto, propomos a utilização de um método de aprendizado de máquina utilizando o classificador SVM para detectar cenas de violência doméstica de um homem contra uma mulher utilizando o som. Apresentamos análises de experimentos com três diferentes parâmetros extraídos dos áudios. Como resultado, obtemos o melhor desempenho utilizando o parâmetro MFCC conseguindo uma acurácia de 73,14%.*

1. Introdução

A violência sempre fez parte da humanidade e foi expressa de diferentes formas e, sendo assim, a redução da violência é um problema importante e, por isso, insiste-se na importância de reduzir a violência cotidiana pois o seu reconhecimento, feito apenas por um ser humano, não é eficaz e requer muitos recursos [Dorogyy et al. 2018]. E um dos tipos de violência muito discutido atualmente é a violência doméstica, que segundo [Oliveira et al. 2019], é um fenômeno grave que atinge os mais variados grupos.

Segundo [Andrade 2019], a violência contra a mulher é amplamente reconhecida como grave problema de saúde pública, com impactos na condição física e mental das vítimas. Ainda nesse mesmo artigo publicado pela revista Pesquisa FAPESP, em março de 2019 [Andrade 2019], em que foram analisados alguns estudos relacionados às vítimas de violência doméstica no Brasil, foi identificado que, em geral, as mulheres que sofrem violência doméstica estão mais propícias a desenvolver distúrbios psiquiátricos como ansiedade, depressão ou pensamentos suicidas e, brasileiras que registraram episódios de violência nos serviços públicos de saúde têm 151,5 vezes mais risco de morrer por suicídio decorrente de um quadro de depressão em comparação com a população feminina em geral.

Segundo [Mu et al. 2016], o sinal produzido pelo som de um áudio contém muitas informações que dados apenas visuais não podem representar, como por exemplo, gritos, explosões, palavras de abuso e até mesmo trechos sonoros demonstrando algum tipo de emoção.

A classificação e detecção de sons de um cenário em um ambiente é uma área de pesquisa em rápido desenvolvimento, e seu crescimento tem sido estimulado por campanhas emergentes de avaliação pública e conjuntos de dados que promovem o desenvolvimento ativo em áreas como classificação automática de cenas acústicas e detecção e classificação automática de eventos sonoros [Mesaros et al. 2018].

A classificação de cenas acústicas (CCA) baseia-se na premissa de que é possível fornecer um rótulo textual como uma caracterização geral de um local ou situação, que se supõe ser distinguível de outros com base em suas propriedades acústicas gerais [Mesaros et al. 2018]. O problema é tipicamente enquadrado como classificação supervisionada e geralmente envolve um número relativamente pequeno de classes [Mesaros et al. 2018]. Conforme mencionado por [Gharib et al. 2018], a CCA tenta classificar os sinais de áudio digital em categorias de cena mutuamente exclusivas.

Conforme mencionado por [Elizalde et al. 2016], o áudio desempenha um papel crítico na compreensão do ambiente ao nosso redor e isso torna a pesquisa de análise de conteúdo de áudio importante para tarefas relacionadas à interação multimídia, porém, ao contrário do campo da visão computacional, que tem uma variedade de conjuntos de dados padrão disponíveis publicamente, a análise de eventos / cenas de áudio não possui um conjunto de dados tão grande.

Segundo [Dorogyy et al. 2018], devem ser construídos sistemas que realizem a tarefa de reconhecimento automático de violência. E conforme mencionado por [Oliveira et al. 2019], a violência doméstica é um problema grave de violência, e devido a essa necessidade de reconhecimento automático, criamos um modelo inicial de aprendizado de máquina para detectar cenas de violência doméstica de um homem contra uma mulher utilizando informações de áudios no idioma português.

Neste estudo nos concentramos no problema de detecção de cenas acústicas envolvendo violência doméstica. Para tanto, propomos a utilização de um método de aprendizado de máquina utilizando o classificador SVM para detectar cenas de violência doméstica de um homem contra uma mulher. Apresentamos também análises para experimentos utilizando três diferentes parâmetros extraídos dos áudios.

2. Trabalhos relacionados

Em 2010, no trabalho [Wang et al. 2010], foi proposto o uso de máquina de vetor de suporte multiclasse (SVM) e busca por correspondência para reconhecimento de áudio de cenas de ação. Seis tipos de sons de cenas de ação (som de espada, som do clube, som desarmado, som quebrado, som de queda de metal e som de grito) foram usados para avaliar o algoritmo proposto. A acurácia de reconhecimento nas seis cenas de ação de áudio atingiu 92,6% [Guo et al. 2010]. Em 2012, foi proposto por [Hwang et al. 2012] uma ferramenta que modela a combinação do contexto de cena, evento e telefone, no qual, um histograma gaussiano é usado para representar um clipe de áudio com um pequeno número de parâmetros, e um classificador K-Vizinhos mais próximos (KNN) alcançou acurácia de 85,2% [Hwang et al. 2012]. Em 2014, [Su 2014] propôs um esquema de reconhecimento de cena auditiva que integra a análise dos dados de áudio da cena com o modelo de tópico *Latent Dirichlet Allocation* (LDA), e dez tipos de cenas de áudio (dentro do veículo, praia, estação de trem, rua, restaurante, auditório, floresta, chuva, parque e guerra) foram utilizadas para avaliar o método proposto e a acurácia do reconhecimento foi de 83,4% [Su 2014].

No trabalho de [Stowell et al. 2015], foi relatado o estado da arte na classificação automática de cenas de áudio e na detecção e classificação automática de eventos de áudio, e foram criados novos conjuntos de dados de áudio e sistemas de linha de base para um desafio estimulado em que uma das tarefas seria a de classificação de cenas acústicas, e os desempenhos dos sistemas submetidos para essa tarefa alcançaram acurácias entre 55% e 77%. Em [Elizalde et al. 2016], foram apresentadas diferentes abordagens para a classificação de cenas acústicas e detecção de eventos sonoros do *challenge on detection and classification of acoustic scenes and events* (DCASE) de 2016, e para a Tarefa 1 (classificação de cenas acústicas) obtiveram uma acurácia geral de 78,9% em comparação com a linha de base de 72,6%, utilizando MFCCs. Em (OO et al., 2018) foi apresentado um estudo comparativo de diferentes classificadores de aprendizado de máquina utilizando o MFCC para classificação de cenas acústicas, e obtiveram acurácias médias de 65,6%, 72,1%, 75,2% e 35,3% utilizando respectivamente os classificadores KNN, SVM, *Decision Tree* (DT) e LDA.

A Figura 1 representa o contexto da inteligência artificial relacionada com este trabalho, que contempla técnicas de aprendizado de máquina utilizadas para ajudar a combater um problema social, a partir da análise de áudios extraídos de cenas acústicas específicas que, no nosso caso, seria o problema da violência doméstica.

3. Metodologia

Nesta seção, apresentamos informações mais detalhadas sobre o nosso modelo proposto para a classificação de cenas de violência doméstica utilizando áudios.

Inicialmente, para gerar um modelo, extraímos diferentes parâmetros de sinais de áudio em domínios de frequência e tempo (que serão discutidos em detalhes nas seções 3.1 e 3.2) na tentativa de encontrar, para o nosso contexto, qual dos parâmetros utilizados são mais adequados para classificação de cenas acústicas envolvendo violência doméstica. Os três parâmetros utilizados são MFCC e Energia, do domínio da frequência, e o ZCR (Zero Rate Crossing) do domínio do tempo. O processo de extração dos parâmetros dos áudios foi realizado a partir de uma base de dados construída para realização dos experimentos deste trabalho. Essa base de dados e a forma como foi

construída serão detalhadas mais especificamente na seção 4.1. Após a fase de extração, iniciou-se a fase de treinamento do modelo, que foi onde realizamos os experimentos utilizando os dados extraídos para cada um dos parâmetros individualmente com o objetivo de treinarmos nossos modelos utilizando o classificador SVM (a utilização desse classificador é justificada na seção 3.3). Por fim, fazemos uma análise da acurácia (a utilização da acurácia como critério para a análise será justificada na seção 4.2) para cada um dos modelos treinados com a finalidade de identificar os parâmetros de áudios e os parâmetros do SVM que influenciam positivamente ou negativamente na geração de um melhor modelo de classificação para o contexto em questão (essa análise será tratada na seção 4.2). As etapas realizadas para manipulação dos experimentos foram:

1. Montagem da base de dados de áudios;
2. Extração dos parâmetros dos sinais de áudios da base construída;
3. Treinamento do modelo utilizando o classificador SVM para cada um dos parâmetros extraídos;
4. Análise de todos os modelos gerados na etapa anterior utilizando como base a acurácia em comparação com outros trabalhos já realizados.

3.1. Parâmetros utilizados

Neste artigo, utilizamos três parâmetros, o MFCC, Energia e o ZCR. Os MFCCs são os parâmetros acústicos de baixo nível mais populares em tarefas de processamento de áudio e fala [Mu et al. 2016]. O MFCC é uma representação do espectro de potência de curto prazo de um som, com base em uma transformação de cosseno linear de um espectro de potência de log em uma escala Mel de frequência não linear [Sarman et al. 2018]. Quando o MFCC é comparado a outros métodos de extração de parâmetros amplamente utilizados, ele tem um desempenho melhor do que outros parâmetros de domínio de frequência como *Linear Prediction Cepstral Coefficients* (LPCC) e *Relative spectral-perceptual linear prediction* (Rasta-PLP) [Sarman et al. 2018].

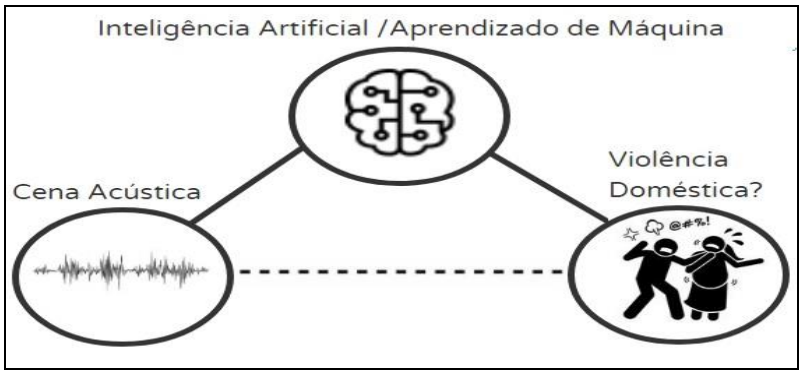

Figura 1. Contexto de Inteligência artificial relacionado com esse trabalho

Outro parâmetro utilizado foi Energia, que é a intensidade sonora percebida pelo ouvido humano, indica a amplitude do sinal em um determinado intervalo, e podem ser utilizados para determinar estados emocionais de locutores, onde altos valores de energia estão normalmente correlacionados com emoções de alta excitação [Ooi et al. 2014]. E por último, utilizamos também o ZCR, que é um parâmetro do domínio do tempo que representa a taxa de alterações de sinal ao longo do tempo e é amplamente

utilizado em aplicações de processamento de voz e aplicações de classificação [Sarman et al. 2018].

3.2. Extração de Parâmetros

Segundo [Giannakopoulos 2015], uma técnica comum utilizada na análise de áudio é o processamento da sequência de parâmetros de médio prazo, de forma que o sinal de áudio é primeiro dividido em janelas de médio prazo (segmentos), que podem ser sobrepostas ou não, e para cada segmento é realizado um processamento de curto prazo. A sequência de parâmetros de cada segmento de médio prazo é usada para calcular estatísticas de parâmetros como por exemplo o valor médio e desvio padrão de cada parâmetro específico. Portanto, cada segmento de médio prazo é representado por um conjunto de estatísticas.

Para o processamento de curto prazo a ser utilizado, o sinal é dividido em janelas de curto prazo (quadros) e para cada quadro são calculados os parâmetros desejados, que resulta em uma sequência de vetores de parâmetros de curto prazo [Giannakopoulos 2015].

Os tamanhos das janelas de curto prazo e de médio prazo indicados como mais adequadas por [Giannakopoulos 2015] são de 20 a 100 ms para as de curto prazo e de 1 a 10 s para as de médio prazo, com uma sobreposição de 50%. Nos nossos experimentos, utilizamos algumas combinações de janelas conforme informações da Tabela 1, considerando uma sobreposição de 50% para ambas as janelas.

Tabela 1. Combinações de janelas utilizadas nos experimentos

Experimento	Janela de curto prazo	Janela de médio prazo
1	20 ms	1 s
2	50 ms	1 s
3	60 ms	5 s
4	100 ms	5 s
5	100 ms	10 s

Para os experimentos demonstrados nesse artigo foi utilizada como base a biblioteca *pyAudioAnalisys* criada por [Giannakopoulos 2015] que faz uso de outras bibliotecas como por exemplo *numpy* e *scipy* para extração de parâmetros, e *scikit-learn* para tarefa de treinamento utilizando classificadores.

3.3. Classificador utilizado

Não é possível afirmar qual o melhor método de classificação existente, uma vez que é difícil comparar resultados na literatura já que estes dependem das bases de dados utilizadas, das condições de independência, do conjunto de parâmetros, entre outros fatores, ou seja, são muitas variáveis envolvidas de forma independentes para isolarmos o problema do algoritmo de classificação [Iriya 2014].

SVM é uma técnica utilizada em reconhecimento de padrões para decisões binárias, que tem sido largamente utilizada para o reconhecimento de emoções através

da voz [Iriya 2014]. Como nossa abordagem considera uma classificação binária em informar a partir de um sinal de áudio, que representa uma cena, se existe ou não violência doméstica de um homem contra uma mulher, utilizaremos o SVM como classificador nos nossos experimentos iniciais. Segundo [Schuller et al. 2011], o SVM fornece boas propriedades de generalização e é considerado como uma espécie de classificador de ponta [Schuller et al. 2011], e em estudo realizado por [Fernández-Delgado et al. 2014] que avaliou 179 classificadores diferentes, o SVM obteve um dos melhores desempenhos.

Utilizamos para treinamento os valores (0.001, 0.01, 0.5, 1.0, 5.0, 10.0, 20.0) como entradas para o parâmetro C, necessário para execução do SVM, conforme indicado por [Giannakopoulos 2015]. Segundo [Pedregosa et al. 2011], o parâmetro C informa à otimização do SVM quanto você deseja evitar classificar erroneamente cada exemplo de treinamento. E nos resultados serão apresentados os valores desse parâmetro com melhor desempenho para cada experimento. Utilizamos 90% dos dados da base para treinamento e 10% para testes.

Nossa abordagem proposta para obtenção de um modelo de classificação de cenas de violência doméstica é demonstrada na Figura 2.

4. Resultados e Avaliações

Essa seção contém informações relacionadas à base de dados construída e os resultados experimentais. Avaliamos a eficácia da nossa proposta utilizando a acurácia como métrica conforme descrito na seção 4.2.

4.1. Base de dados de Áudios

Tendo em vista que nosso objetivo, conforme descrito na seção 1, é desenvolver o modelo com aplicação para a língua portuguesa, e não foi obtido êxito na pesquisa por bancos de dados de áudios adequados para esse tipo de cena e linguagem específicas, optou-se por montarmos nossa própria base de dados utilizando trechos de vídeos da ficção extraídos manualmente de filmes e vídeos que utilizavam a língua portuguesa como idioma, visando cumprir esse objetivo da linguagem. Vários trabalhos na área de análise de cenas através do áudio já utilizaram base de dados extraídos da ficção como [Clavel et al. 2008], [Sarman et al. 2018] e [Mu et al. 2016]. Segundo [Clavel et al. 2008], a ficção fornece gravações de manifestações emocionais e oferece um amplo escopo de retratos de emoções verossímeis e, além disso, as emoções são expressas por atores habilidosos nas interações interpessoais que, somando-se ao contexto do roteiro de cada filme, favorece a identificação de atores com personagens reais e tende a provocar emoções genuínas tornando-se assim um material relevante para esse tipo de aplicação.

Nossa base de dados contém 124 áudios, com durações entre 21 segundos e 4 minutos, com cada um deles representando uma cena particular de conversa entre um homem e uma mulher. Os áudios foram divididos e categorizados em duas classes distintas: a primeira representa o contexto de algum tipo de violência do homem contra a mulher, seja ela sexual, psicológica ou física contendo um total de 62 áudios; a segunda representa qualquer outro contexto de conversa diferente da violência doméstica sugerida na primeira classe, entre um homem e uma mulher, como por

exemplo: uma conversa durante o jantar, uma discussão sobre o comportamento de um filho sem ocorrer agressão). Para essa classe, foram levantados 62 áudios.

4.2. Resultados dos experimentos

Existem várias métricas possíveis para apresentar os resultados da classificação [Sarman et al. 2018]. Segundo [Goodfellow et al. 2016], para a tarefa de classificação, geralmente é calculada a acurácia do modelo, ou seja, a proporção de exemplos que são corretamente classificados. Dessa forma, utilizamos a acurácia para avaliar os resultados de nossos experimentos.

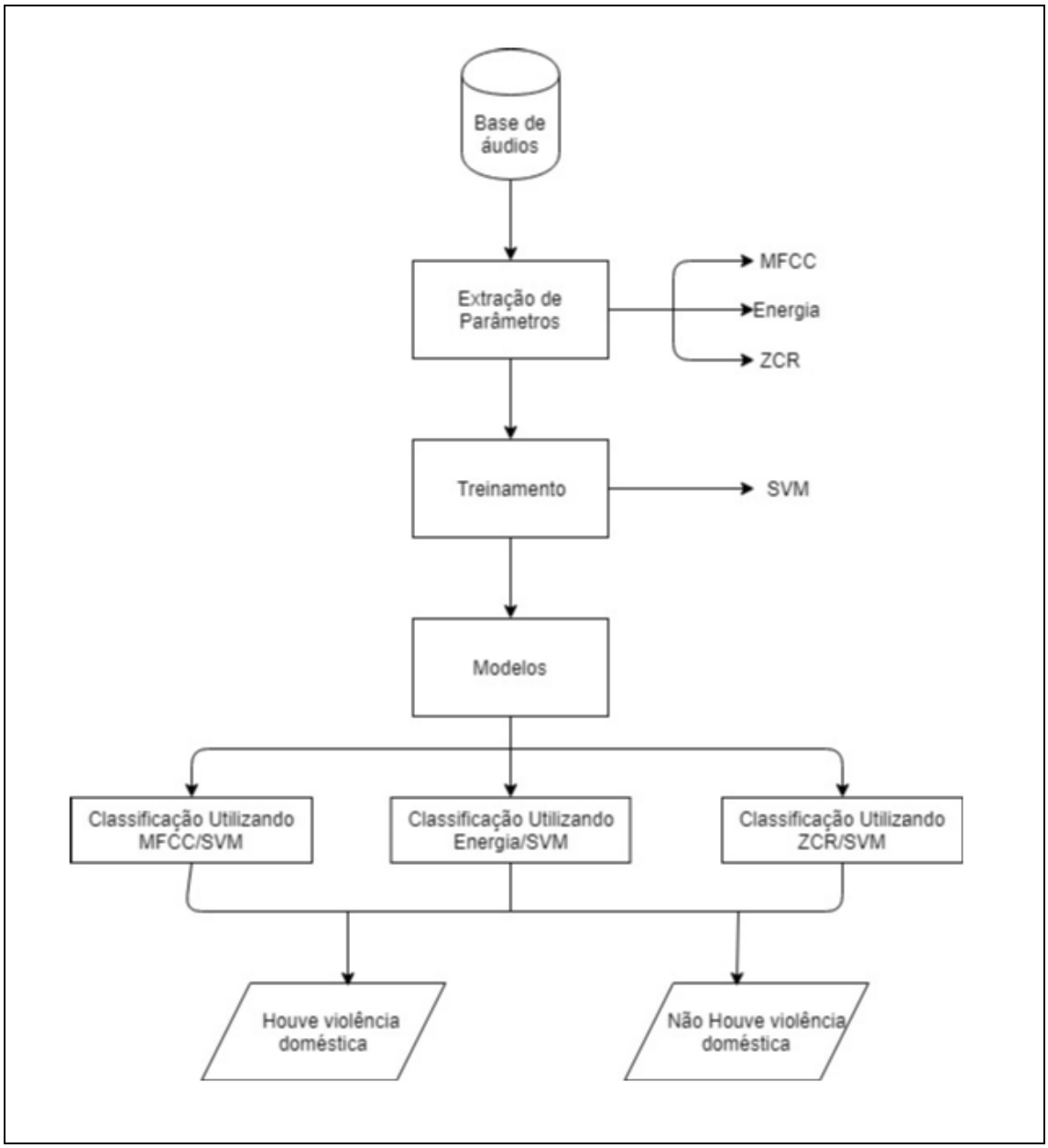

Figura 2. Diagrama da abordagem proposta

A Tabela 2 apresenta os resultados das acurácias para cada um dos parâmetros MFCC, Energia e ZCR utilizando a janela de curto e médio prazos com valores iguais a 20 ms e 1 s respectivamente. A Tabela 3 apresenta os resultados das acurácias utilizando na mesma sequência anterior valores de janelas de 50 ms e 1 s. Seguindo, a Tabela 4 apresenta os resultados das acurácias utilizando valores das janelas de 60 ms e 5 s. A Tabela 5 apresenta os resultados das acurácias utilizando janelas de curto e médio prazos com valores de 100 ms e 5 s. E por último, a Tabela 6 apresenta os resultados das acurácias utilizando valores de 100 ms e 10 s para as janelas de curto e médio prazos respectivamente. Todas as tabelas mencionadas acima também apresentam o melhor valor do parâmetro C utilizado pelo classificador SVM em cada experimento.

Comparando os resultados dos experimentos realizados, verificamos que utilizando o parâmetro ZCR extraído dos áudios foram alcançados os valores mais baixos atingindo uma acurácia média de 66,54%. Obtemos os melhores resultados utilizando o parâmetro MFCC com uma acurácia média de 73,14%, demonstrando ser superior na identificação de cenas de violência doméstica em relação ao ZCR e Energia, que obteve acurácia média de 71,3%. Acreditamos que o melhor desempenho do MFCC seja pelo fato de o objetivo desses coeficientes ser o de simular o comportamento de ouvidos humanos aplicando a análise cepstral no sinal, conforme levantado por [Kishore et al. 2013], e também pelo fato de fornecerem uma variabilidade entre classes razoavelmente alta para permitir a discriminação de classes por muitas abordagens diferentes de aprendizado de máquina conforme mencionado por [Mesaros et al. 2018].

Tabela 2. Resultados dos experimentos utilizando os três parâmetros do áudio isoladamente, com janelas de curto e médio com valores iguais a 20 ms e 1 s respectivamente

Parâmetro de Áudio	Parâmetro C (SVM)	Acurácia
MFCC	10.0	70,2%
ZCR	5.0	66,3%
Energia	20.0	72,4%

Tabela 3. Resultados dos experimentos utilizando os três parâmetros do áudio isoladamente, com janelas de curto e médio prazos com valores iguais a 50 ms e 1 s respectivamente

Parâmetro de Áudio	Parâmetro C (SVM)	Acurácia
MFCC	1.0	73%
ZCR	0.5	65,5%
Energia	10.0	71,7%

Com relação à Energia ter também obtido resultados bons em relação ao ZCR, pode ser justificado pelo fato de em uma cena de violência doméstica o estado emocional dos envolvidos poderem estar bem alterados justificando o fato de que a curva do parâmetro Energia está associado a emoções de alta excitação, segundo [Ooi et al. 2014].

Tabela 4. Resultados dos experimentos utilizando os três parâmetros do áudio isoladamente, com janelas de curto e médio com valores iguais a 60 ms e 5 s respectivamente

Parâmetro de Áudio	Parâmetro C (SVM)	Acurácia
MFCC	0.01	73,6%
ZCR	1.0	65,5%
Energia	10.0	71,2%

Tabela 5. Resultados dos experimentos utilizando os três parâmetros do áudio isoladamente, com janelas de curto e médio com valores iguais a 100 ms e 5 s respectivamente

Parâmetro de Áudio	Parâmetro C (SVM)	Acurácia
MFCC	0.01%	74,2%
ZCR	10.0	67,7%
Energia	20.0	70,4%

Não foi possível identificar uma grande influência geral nos resultados dos experimentos em relação aos valores dos tamanhos das janelas por não apresentar uma tendência em nenhum dos experimentos. Em alguns casos pode ter favorecido um parâmetro, mas pode ter influenciado negativamente no outro. Então, por esses experimentos realizados não foi possível identificar qual melhor janela seria indicada para extração de um parâmetro do áudio específico para ser usado nesse tipo de classificação. Com relação ao parâmetro C do SVM identificamos que para o MFCC as melhores acurácias utilizaram o valor de 0.01, demonstrando uma tendência de melhor valor combinado com o MFCC.

Tabela 6. Resultados dos experimentos utilizando os três parâmetros do áudio isoladamente, com janelas de curto e médio com valores iguais a 100 ms e 10 s respectivamente

Parâmetro de Áudio	Parâmetro C (SVM)	Acurácia
MFCC	0.01	74,7%
ZCR	1.0	67,7%
Energia	20.0	70,8%

5. Conclusão

Neste artigo, propomos um método de aprendizado de máquina baseado em parâmetros extraídos de áudios para classificar cenas envolvendo violência doméstica de um homem contra uma mulher. Avaliamos nossos experimentos utilizando a acurácia, conforme sugerido por [Goodfellow et al. 2016]. Observamos que o parâmetro de áudio

MFCC demonstrou-se ser superior na coleta de informações do áudio para identificação de cenas de violência doméstica em comparação com os outros parâmetros ZCR e Energia, alcançando uma acurácia média de 73,14%. Além disso, pelo fato de até o momento dessa pesquisa não termos encontrados em nossas nenhum estudo dessa natureza sendo realizado, abordando especificamente cenas de violência doméstica, consideramos que nossa abordagem apresenta bons resultados se comparados com outros trabalhos envolvendo apenas reconhecimento de algum outro tipo de cena acústica, utilizando parâmetros de áudios, e que obtiveram resultados de acurácias não tão distantes dos nossos, como por exemplo em [Kotti et al. 2008] com 82%, [Stowell et al. 2015] apresentando sistemas com acurácias entre 55% e 77% , [Yang et al. 2016] com 79,9%, [Elizalde et al. 2016] com 78%, e [OO et al. 2018] que obteve nos experimentos acurácias médias de 65,6%, 72,1%, 75,2% e 35,3% utilizando classificadores diferentes.

Para trabalhos futuros, poderemos utilizar uma abordagem de utilização de diferentes classificadores utilizando os parâmetros MFCC e Energia isoladamente e em conjunto para o aprendizado de máquina, pois demonstraram ter melhores resultados do que ZCR, considerando a ampliação do número de amostras de áudios em nossa base de dados, pois acreditamos que essa evolução da base pode ampliar a capacidade de aprendizado de máquina e consequentemente alcançar uma melhor acurácia como resultado.

Referências

Kotti, M., Ververidis, D., Evangelopoulos, G., Panagakis, I., Kotropoulos, C.,Maragos, P., Pitas, I. (2008) “Audio-Assisted Movie Dialogue Detection”, In: IEEE Transactions on Circuits and Systems for Video Technology, vol. 18, no. 11, pp. 1618-1627, Nov. 2008.

Clavel, C., Vasilescu, I.,Devillers, L., Richard, G., Ehrette, T. (2008) “Fear-type emotion recognition for future audio-based surveillance systems”. Speech Communication. Volume 50, Issue 6, June 2008, Pages 487-503.

Guo, F., Shan, S., Wang, X. (2010) “Using One-Class SVMs and MP for Audio Recognition of Action Scenes”, Second International Workshop on Education Technology and Computer Science, 2010,401-404.

Schuller, B., Batliner, A., Steidl, S., Seppi, D. (2011) “Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge”. Speech Communication, vol. 53, no. 9/10, pp. 1062–1087, 2011.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É. (2011) “Scikit-learn: Machine Learning in Python”, JMLR 12, pp. 2825-2830, 2011.

Hwang, K., Lee, S. (2012) “Environmental Audio Scene and Activity Recognition through Mobile-based Crowdsourcing”, In: IEEE Transactions on Consumer Electronics, 2012,58(2):700-705.

Kishore, K. V. K., Satish, P. K. (2013) “Emotion recognition in speech using MFCC and wavelet features”, In: 3rd IEEE International Advance Computing Conference (IACC), Ghaziabad, 2013, pp. 842-847.

Su, F. (2014) “Auditory scene analysis and recognition with LDA topic model”, In: IEEE International Conference on Multimedia and Expo,2014,1-6.

Ooi, C. S., Seng, K. P., Ang, L. M., Chew, L. W. (2014) “A new approach of audio emotion recognition”. Expert Systems with Applications 41 (13) (2014) 5858–5869.

Fernández-Delgado, M., Cernadas, E., Barro, S., and Amorim, D. (2014) “Do we need hundreds of classifiers to solve real world classification problems?”, The Journal of Machine Learning Research, 15(1), 3133-3181.

Iriya, R. (2014) “Análise de sinais de voz para reconhecimento de emoções”, Dissertação (Mestrado) — Curso de Engenharia e Sistemas Eletrônicos, Universidade de São Paulo, 2014.

Giannakopoulos, T. (2015) “pyAudioAnalysis: An Open-Source Python Library for Audio Signal Analysis”, PLoS ONE 10(12): e0144610. doi:10.1371/ journal.pone.0144610.

Stowell, D., Giannoulis, D., Benetos, E., Lagrange, M., Plumbley, M. D. (2015) “Detection and Classification of Acoustic Scenes and Events”, In: IEEE Transactions on Multimedia, vol. 17, no. 10, pp. 1733-1746, Oct. 2015.

Goodfellow, I., Bengio, Y., Courville, A. (2016) “Deep Learning”. MIT Press http://www.deeplearningbook.org.

Yang, J., Cai, M., Li M., Jin, H. (2016) “Movie audio scene recognition based on WFST”, In: International Conference on Audio, Language and Image Processing (ICALIP), Shanghai, 2016, pp. 77-80.

Elizalde, B., Kumar, A., Shah, A., Badlani, R., Vincent, E., Raj, B., Lane, I. (2016) “Experiments on the DCASE Challenge 2016: Acoustic scene classification and sound event detection in real life recording”, In: Proc. Workshop Detection Classification Acoust. Scenes Events, Budapest, Hungary, Sep. 2016, pp. 20-24.

Mu, G., Cao, H., Jin, Q. (2016) “Violent Scene Detection Using Convolutional Neural Networks and Deep Audio Features”. In: Tan T., Li X., Chen X., Zhou J., Yang J., Cheng H. (eds) Pattern Recognition. CCPR 2016. Communications in Computer and Information Science, vol 663. Springer, Singapore.

Mesaros, A., Heittola, T., Benetos, E., Foster, P., Lagrange, M., Virtanen, T., Plumbley, M. D. (2018) “Detection and classification of acoustic scenes and events: Outcome of the dcase 2016 challenge”, In: IEEE/ACM Transactions on Audio Speech and Language Processing vol. 26 no. 2 pp. 379-393 Feb 2018.

Gharib, S., Derrar, H., Niizumi, D., Senttula, T., Tommola, J., Heittola, T., Virtanen, T., Huttunen, H. (2018) “Acoustic scene classification: A competition review”. In: IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP), pages 1–6. IEEE, 2018.

OO, M. M. (2018) “Comparative Study of MFCC Feature with Different Machine Learning Techniques in Acoustic Scene Classification”, In: International Journal of Research and Engineering ISSN: 2348-7860 (O) | 2348-7852 (P) | Vol. 5 No. 7 | July 2018 | PP. 439-444.

Sarman, S. e Sert, M. (2018) “Audio based violent scene classification using ensemble learning”, In: 6th International Symposium on Digital Forensic and Security (ISDFS), Antalya, 2018, pp. 1-5.

Dorogyy, Y., Kolisnichenko, V., Levchenko, K. (2018) “Violent Crime Detection System”, In: IEEE 13th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT), Lviv, 2018, pp. 352-355.

Andrade, R. O. (2019) “Faces da violência doméstica”, Revista Pesquisa FAPESP. Edição 277, Março, 2019. Disponível em: <http://revistapesquisa.fapesp.br/2019/03/07/faces-da-violencia-domestica/>. Acesso em: 07 de maio de 2019.

Oliveira, C. A. B., Alencar, L. N., Cardena, R. R., Moreira, K. F. A., Pereira, P. P. S., Fernandes, D. E. R. (2019) “Perfil da vítima e características da violência contra a mulher no estado de Rondônia”, Brasil. Rev Cuid. 2019; 10(1): e573.