LUCAS FERREIRA PAIVA

ESTIMAÇÃO AUTOMÁTICA DE RITMO PARA AUXILIAR SURDOS NO APRENDIZADO DA DANÇA DO FORRÓ

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Ciência da Computação, para obtenção do título de *Magister Scientiae*.

Orientador: Rodolpho Vilela Alves Neves

Coorientador: Leonardo Bonato Felix

VIÇOSA - MINAS GERAIS
2022

Ficha catalográfica elaborada pela Biblioteca Central da Universidade Federal de Viçosa - Campus Viçosa

T

P149e
2022

Paiva, Lucas Ferreira, 1996-

Estimação automática de ritmo para auxiliar surdos no aprendizado da dança do forró / Lucas Ferreira Paiva. – Viçosa, MG, 2022.

1 dissertação eletrônica (89 f.): il. (algumas color.).

Texto em português e inglês.

Orientador: Rodolpho Vilela Alves Neves.

Dissertação (mestrado) - Universidade Federal de Viçosa, Departamento de Informática, 2022.

Inclui bibliografia.

DOI: https://doi.org/10.47328/ufvbbt.2023.165

Modo de acesso: World Wide Web.

1. Redes neurais (Computação). 2. Aprendizado do computador. 3. Forró (Música) - Banco de dados. 4. Surdos - Meios de comunicação. 5. Integração social. I. Neves, Rodolpho Vilela Alves, 1987-. II. Universidade Federal de Viçosa. Departamento de Informática. Programa de Pós-Graduação em Ciência da Computação. III. Título.

CDD 23. ed. 006.32

Bibliotecário(a) responsável: Euzébio Luiz Pinto CRB-6/3317

LUCAS FERREIRA PAIVA

ESTIMAÇÃO AUTOMÁTICA DE RITMO PARA AUXILIAR SURDOS NO APRENDIZADO DA DANÇA DO FORRÓ

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Ciência da Computação, para obtenção do título de *Magister Scientiae.*

APROVADA: 21 de dezembro de 2022.

Assentimento:

gov.br Documento assinado digitalmente
LUCAS FERREIRA PAIVA
Data: 24/04/2023 12:20:42-0300
Verifique em https://validar.iti.gov.br

_____
Lucas Ferreira Paiva
Autor

gov.br Documento assinado digitalmente
RODOLPHO VILELA ALVES NEVES
Data: 24/04/2023 12:32:54-0300
Verifique em https://validar.iti.gov.br

_____
Rodolpho Vilela Alves Neves
Orientador

Este trabalho é dedicado a todas as pessoas amantes do forró, Surdos e ouvintes.

AGRADECIMENTOS

Agradeço primeiramente minha mãe, Gisléia, pelo exemplo de força, carinho e suporte durante toda minha vida. Ao meu pai, José Antônio e minha irmã Aline por todo apoio e incentivo. Também a toda minha família, avós e avôs, tias e tios, primas e primos, por toda fé, força e esperança confiadas a mim. Em especial, minha companheira Elizabeth por toda colaboração, carinho e incentivo. Agradeço a todos os amigos espirituais que seguiram me fortalecendo, protegendo e intuindo em toda essa jornada de aprendizagem técnica e científica, mas também espiritual e de autoconhecimento.

Agradeço a todos os integrantes do NIAS (Núcleo Interdisciplinar de Análise de Sinais) e amigos da pós que contribuíram direta e indiretamente com a realização desta pesquisa. Em especial aos professores e orientadores Rodolpho Vilela e Leonardo Bonato, agradeço por toda motivação, paciência, conhecimento e tempo dedicado. Agradeço também a todos os participantes da pesquisa por terem dançado todas as músicas utilizadas nesse trabalho, suas contribuições viabilizaram esta pesquisa.

Finalmente, agradeço aos professores e funcionários do departamento de Engenharia Elétrica e do Programa de Pós Graduação em Ciência da Computação, por todo suporte ao longo do mestrado. Agradeço também aos demais funcionários, da UFV e terceirizados, por toda estrutura oferecida. A dedicação de vocês é essencial para este e todos os trabalho realizados na UFV.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

"Não há saber mais ou saber menos: há saberes diferentes."
(Paulo Freire)

RESUMO

FERREIRA-PAIVA, Lucas, M.Sc., Universidade Federal de Viçosa, dezembro de 2022. **Estimação Automática de Ritmo para Auxiliar Surdos no Aprendizado da Dança do Forró**. Orientador: Rodolpho Vilela Alves Neves. Coorientador: Leonardo Bonato Felix.

Os Surdos e deficientes auditivos representam cerca de 5% da população mundial. Apesar disso, não gozam das mesmas oportunidades dos ouvintes. Um exemplo disso é o acesso à música e à dança, onde mesmo sendo tão capazes quanto os ouvintes para identificarem sentimentos e seguirem o ritmo de músicas, são estigmatizados como não musicais. Esse estigma acaba por resultar em poucas iniciativas voltadas para a criação de tecnologias que potencializem o contato dos surdos com a musicalidade. O principal trabalho encontrado na literatura consistiu na criação de um modelo baseado em redes neurais capaz de estimar o ritmo de músicas de forró para passar esse ritmo por vibração para Surdos. Apesar de resultados promissores terem sidos encontrados, limitações no banco de dados como pequeno número de amostras, falta de diversidade e imprecisão nas anotações do ritmo, inviabilizam a implantação da abordagem. Neste trabalho são apresentadas iniciativas para viabilizar a construção de um modelo para sinalizar o ritmo para Surdos por meio de vibração. A primeira abordagem foi a adição de músicas com ruído real de um espaço de dança junto com o banco de dados do trabalho anterior, o modelo baseado em redes neurais treinado alcançou erro percentual médio menor que 7%. Apesar deste banco de dados ser ainda pequeno e com anotações manuais, foi observado potencial do modelo para ser utilizado em condições reais. Devido a isso foram realizados trabalhos na direção de aumentar o banco de dados. Um deles é uma revisão de literatura para encontrar técnicas de aumento de dados de áudio. Foram encontradas 30 técnicas usadas em variadas tarefas de classificação de áudio, aumentando em até 30 pp a acurácia dos modelos para *datasets* pequenos. Além do aumento artificial dos dados, foram realizados trabalhos para a criação de um novo banco de dados, com instâncias suficientes para treinar modelos convolucionais. Inicialmente foi criado o Forroset, um banco de dados com 2977 músicas de forró, contendo 40 informações diferentes, dentre elas, os arquivos de áudio em MP3, a popularidade e o BPM, fornecidos pelo Spotify. Por fim, para adicionar a duração do passo base às músicas do Forroset, foi realizado um experimento onde 9 pessoas se filmaram dançando 380 músicas no total e tiveram a duração do passo base estimada por um modelo de visão computacional proposto. Esse modelo conseguiu estimar a duração do passo base com erro percentual médio inferior a 3%. Além das anotações da duração do passo base, foram adicionadas ao

Forroset, versões com ruído domestico das músicas dançadas, obtidas através dos áudios dos vídeos, criando assim o Forroset+. Estas iniciativas poderão possibilitar o treinamento de modelos com maior capacidade de generalização. Devido ao maior número de músicas será possível também a utilização de redes neurais profundas como redes convolucionais e recorrentes. Apesar da estrutura fornecida propiciar o treinamento e validação destes modelos, não foram realizados experimentos para verificar se de fato os esforços foram suficientes. Portanto, técnicas desenvolvidas em trabalhos futuros (e.g., redes neurais profundas) deverão ser comparadas aos modelos aqui utilizados, além da avaliação das técnicas de aumento de dados para áudio.

Palavras-chave: Aprendizado de Máquinas. Banco de Dados. Dança. Inclusão.

ABSTRACT

FERREIRA-PAIVA, Lucas, M.Sc., Universidade Federal de Viçosa, December, 2022. **Automatic Rhythm Estimation to Assist the Deaf in Forró Dance Learning**. Advisor: Rodolpho Vilela Alves Neves. Co-advisor: Leonardo Bonato Felix.

Despite making up a sizable portion of the population, hearing impaired and Deaf people do not have the same opportunities as hearing people. Access to music and dancing is an example of this, when people are stereotyped as unmusical despite their ability to perceive sentiments and follow musical rhythms on par with listeners. Few efforts are consequently carried out to develop technologies that improve the deaf people's interaction with music. The key contribution to the literature involves developing a neural network-based model that could estimate the forró music's rhythm and pass it by vibration to Deaf people. Despite the promising results, the approach is challenging to put into practice due to the database's constraints, including its size, lack of diversity, and imprecision in rhythm notes. Initiatives for the development of a model to vibrate and signal the rhythm for Deaf people are presented in this study. The first approach was to add songs with actual noise to the prior work's database. The model based on neural networks was shown to achieve an average percentage error of less than 7% even in a real noise scenario of a dancing space. Regardless of the fact that the database is currently limited and has manual annotations, the model has the potential to be employed in real-world scenarios. As a result, work was done to expand the database. The first step is to conduct a literature review to identify audio data augmentation techniques. We discovered 30 techniques used in various audio classification tasks, which increased model accuracy by up to 30 pp for small datasets. A new database with enough instances to train convolutional models was created in addition to artificially increasing the data. Initially, Forroset was created, a dataset containing 2977 forró songs and 40 different pieces of information from Spotify, such as audio files, popularity, and BPM. Finally, in order to add the duration of the base step to the Forroset songs, 9 people filmed themselves dancing to 380 songs in total and had the duration of the base step estimated by a proposed computer vision model. With an average percentage error of less than 3%, this model was able to estimate the duration of the base step. In addition to the annotations of the duration of the base step, versions of the danced songs with domestic noise were added to Forroset; these recordings were obtained via the audio of the videos, resulting in Forroset+. These initiatives may make it possible to train models with greater generalization capacity. Because there will be more songs, deep neural networks such as convolutional and recurrent networks will be possible to use. Despite the provided

structure for training and validation of these models, no experiments were conducted to determine whether the efforts were sufficient. As a result, in future works, the models already used with deep approaches, as well as the evaluation of data augmentation techniques for audio, should be compared.

Keywords: Machine Learning. Dataset. Dance. Inclusion.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

Capítulo 1

Introdução

Segundo a Organização Mundial de Saúde, 430 milhões de pessoas possuem perda de audição em algum grau [1]. Apesar de representarem uma parcela significativa da sociedade, barreiras na comunicação fazem com que pessoas surdas e com deficiência auditiva sejam mais suscetíveis a adoecerem mentalmente que ouvintes [2, 3]. A dança pode ser usada para reverter esse quadro, uma vez que possui forte poder de inclusão e pode oferecer maior capacidade para o indivíduo se expressar, independente da forma que é explorada [4, 5], além de possuir benefícios terapêuticos [6]. Apesar dos surdos serem erroneamente estigmatizados como indivíduos não musicais [7], quando comparados com ouvintes, são igualmente capazes de seguirem o ritmo da música através de estímulos táteis [8] e chegam a ser superiores na detecção de emoções em músicas através de vibração [9].

Como é possível sentir a vibração da música [10], principalmente os sons mais graves, a ideia de passar o ritmo por vibração possui pelo menos três décadas [11] e tem sido explorada atualmente [12, 13, 14, 15]. Apesar disso, um estudo recente identificou 83 instrumentos musicais digitais inclusivos e, destes, apenas 5 eram destinados à pessoas S/DA (Surdos / Deficientes Auditivos) [16], mostrando que a inserção de surdos em contextos musicais precisa ser mais explorada.

No contexto brasileiro o forró se destaca como um dos principais gêneros musicais, popular em todas as camadas socioeconômicas e com sua matriz tradicional reconhecida como Patrimônio Cultural Imaterial do Brasil [17]. Além disso, é um dos gêneros mais tocados no Spotify no país [18], a maior plataforma de *streaming* de música do mundo.

Com objetivo de otimizar o contato de surdos com o forró Paiva et al. [19] apresentou um modelo baseado em redes neurais para estimação do ritmo, visando embarcar esse modelo em um celular para passar o ritmo de músicas por vibração. O modelo se mostrou promissor, uma vez que apresentou erro percentual médio inferior a 4% para a estimação da duração do compasso de músicas sem ruído. No entanto, verificou-se a necessidade de aumentar o banco de dados para aumentar a generalização do modelo e de automatizar o processo de aquisição da duração do passo base (DPB) para mitigar possíveis erros humanos.

Extensos bancos de dados que refletem situações às quais os modelos de aprendizado de máquinas terão que lidar durante sua aplicação são essenciais para garantir a aplicabilidade do modelo [20]. Foram encontrados variados bancos de dados anotados na literatura [21, 22, 23], no entanto, nenhum deles contêm músicas de forró com anotação de ritmo, justificando a necessidade da criação de um banco de dados inédito. Além disso, técnicas de aumento de dados vêm sendo empregadas para aumentar o desempenho de modelos de classificação de sons ambientes [24, 25] e de gêneros musicais [26, 27], o que sustenta sua utilização. Por fim, a precisão na aquisição da DPB é essencial para que o modelo aprenda com exemplos corretos, visto que o processo é otimizado para aproximar as respostas da rede ao gabarito oferecido [28]. O mapeamento de movimentos humanos a partir de gravações de vídeo tem sido amplamente explorado com visão computacional [29, 30, 31] indicando potencial para mensurar, de forma automática, a duração do passo base através de vídeos de pessoas dançando.

1.1 Objetivos gerais e específicos

O objetivo geral deste trabalho é viabilizar a criação de um modelo capaz de estimar o ritmo de músicas de forró. Esse objetivo foi dividido em duas frentes de trabalho: avaliar modelos baseados em redes neurais para estimar a duração do compasso; e criar bancos de dados que permitam o ajuste dos modelos profundos. Os objetivos específicos são:

1. Criar um banco de dados com músicas de forró com ruídos reais;
2. Aumentar o banco de dados sem ruído;
3. Avaliar o uso de ruído branco para simular o ruído real;
4. Ajustar e avaliar o modelo proposto por [19] para um cenário real;
5. Revisar técnicas de aumento de dados com potencial para serem utilizadas em dados de áudio;
6. Implementar um algoritmo baseado em visão computacional para mensurar a duração do passo base a partir de vídeo; e
7. Usar o algoritmo de visão computacional para criar um banco dados grande o suficiente para treinar modelos profundos.

1.2 Organização dos capítulos

A estrutura desta dissertação segue o formato "artigos científicos" estabelecido pelo Conselho Técnico de Pós Graduação da Universidade Federal de Viçosa [32]. O formato é dividido em: introdução geral (Capítulo 1), artigos científicos (Capítulos 2, 3, 4, 5) e conclusões gerais (Capítulo 6).

No Capítulo 2 é apresentado o trabalho "Towards a device for helping deaf people to dance: estimation of "forró" bar length using artificial neural network", publicado na revista IEEE Latin America Transactions [33]. Neste artigo tem se a expansão do banco de dados apresentado em Paiva et al. [19], com a regravação de 37 das 40 músicas inicias com ruídos de um espaço de dança e adição de 42 novas músicas com gravação de estúdio. O modelo proposto por Paiva et al. [19] foi avaliado para o cenário real, além disso foi avaliado o uso do ruído branco para substituir o ruído real durante o treinamento.

No Capítulo 3 é apresentada uma revisão de técnicas de aumento de dados utilizadas em dados de áudio. O trabalho é intitulado "A Survey of Data Augmentation for Audio Classification" e foi apresentado no XXIV Congresso Brasileiro de Automática (CBA 2022) [34]. O contexto da classificação foi utilizado visando selecionar as melhores técnicas de aumento devido à vasta literatura existente.

Com intuito de avaliar redes neurais profundas, no Capítulo 4 é apresentado um banco de dados com aproximadamente 3 mil músicas de forró. Este banco de dados foi descrito em um artigo que foi apresentado no 17th Ibero-American Conference on Artificial Intelligence (Iberamia'2022), com o título "Forroset: A multipurpose dataset of Brazilian Forró music" [35]. O banco de dados foi extraído das base de dados do Spotify[1] e da plataforma Vagalume[2] e contém informações editoriais, características dos áudios, informações de ritmo, letras das músicas e arquivos de áudio.

A fim de complementar o Forroset, no Capítulo 5 é apresentado um algoritmo baseado em visão computacional para a anotação automática das músicas a partir de vídeos domésticos de pessoas dançando. O artigo é intitulado "Automatic forró rhythm estimation from home videos" e será submetido ao periódico Journal of the Brazilian Computer Society (JBCS).

Por fim, no Capítulo 6, são apresentadas considerações finais a respeito dos artigos, limitações do trabalho desenvolvido e oportunidades futuras de pesquisa.

[1] https://developer.spotify.com/documentation/web-api/
[2] https://api.vagalume.com.br/

Referências

[1] WHO. Deafness and hearing loss, mar 2021. URL https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss.

[2] Johannes Fellinger, Daniel Holzinger, and Robert Pollard. Mental health of deaf people. *The Lancet*, 379(9820):1037–1044, 2012. ISSN 0140-6736. doi: 10.1016/S0140-6736(11)61143-4.

[3] Meghan L Fox, Tyler G James, and Steven L Barnett. Suicidal behaviors and help-seeking attitudes among deaf and hard-of-hearing college students. *Suicide and Life-Threatening Behavior*, 50(2):387–396, 2020. doi: 10.1111/sltb.12595.

[4] Kate Elswit. So you think you can dance does dance studies. *TDR/The Drama Review*, 56(1):133–142, 2012.

[5] Theresa Purcell Cone. Teaching dance for access, inclusion, and equity, 2015.

[6] Sabine C Koch, Roxana F. F. Riege, Katharina Tisborn, Jacelyn Biondo, Lily Martin, and Andreas Beelmann. Effects of dance movement therapy and dance on health-related psychological outcomes. a meta-analysis update. *Frontiers in Psychology*, 10:1806, 2019. ISSN 1664-1078. doi: 10.3389/fpsyg.2019.01806.

[7] N. Haguiara-Cervellini. *A musicalidade do surdo: representação e estigma*. Plexus Editora, 2003.

[8] Pauline Tranchant, Martha M Shiell, Marcello Giordano, Alexis Nadeau, Isabelle Peretz, and Robert J Zatorre. Feeling the beat: Bouncing synchronization to vibrotactile music in hearing and early deaf people. *Frontiers in Neuroscience*, 11:507, 2017. ISSN 1662-453X. doi: 10.3389/fnins.2017.00507.

[9] Andréanne Sharp, B. A. Bacon, and F. Champoux. Enhanced tactile identification of musical emotion in the deaf. *Experimental Brain Research*, 238(5):1229–1236, 2020. ISSN 14321106. doi: 10.1007/s00221-020-05789-9.

[10] Edith Van Dyck, Dirk Moelants, Michiel Demey, Alexander Deweppe, Pieter Coussement, and Marc Leman. The impact of the bass drum on human dance movement. *Music Perception*, 30(4):349–359, 2013. ISSN 0730-7829. doi: 10.1525/mp.2013.30.4.349.

[11] Masashi Ezawa. Rhythm perception equipment for skin vibratory stimulation. *IEEE Engineering in Medicine and Biology Magazine*, 7(3):30–34, 1988. ISSN 07395175. doi: 10.1109/51.7932.

[12] Yudi Dong, Jian Liu, Yingying Chen, and Woo Y. Lee. Salsaasst: Beat counting system empowered by mobile devices to assist salsa dancers. In *2017 IEEE 14th International Conference on Mobile Ad Hoc and Sensor Systems*, pages 81–89, 2017. ISBN 9781538623237. doi: 10.1109/MASS.2017.25.

[13] H Florian, Adrian Mocanu, Cristian Vlasin, José Machado, Vitor Carvalho, Filomena Soares, Adina Astilean, and Camelia Avram. Deaf people feeling music rhythm by using a sensing and actuating device. *Sensors and Actuators A: Physical*, 267:431–442, 2017. ISSN 0924-4247. doi: 10.1016/j.sna.2017.10.034.

[14] The level of self-esteem of deaf children: Can participating in dance lessons with vibrational headphones improve it? *Arts in Psychotherapy*, 64(October 2017):34–38, 2019. ISSN 18735878. doi: 10.1016/j.aip.2019.03.004.

[15] Feeling the beat: Bouncing synchronization to vibrotactile music in hearing and early deaf people. *Frontiers in Neuroscience*, 11(SEP):507, sep 2017. doi: 10.3389/fnins.2017.00507.

[16] Emma Frid. Accessible digital musical instruments-a review of musical interfaces in inclusive music practice. *Multimodal Technologies and Interaction*, 3(3):57, 2019. ISSN 2414-4088. doi: 10.3390/mti3030057.

[17] IPHAN. Matrizes Tradicionais do Forró recebem título de Patrimônio Cultural do Brasil, 2021. URL https://bit.ly/iphan-forro.

[18] Maria Luiza Botelho Mondelli, Luiz M. R. Gadelha Jr., and Artur Ziviani. O que os países escutam: Analisando a rede de gêneros musicais ao redor do mundo. In *Anais do VII Brazilian Workshop on Social Network Analysis and Mining*, 2018.

[19] Lucas F Paiva, Hugo G Lopes, Leonardo B Felix, and Rodolpho VA Neves. Estimação do compasso musical do forró utilizando rede perceptron multicamadas. In *Anais do Congresso Brasileiro de Automática*, volume 2, 2020. doi: 10.48011/asba.v2i1.1331.

[20] C. W. Wu, C. Dittmar, C. Southall, R. Vogl, G. Widmer, J. Hockman, M. Muller, and A. Lerch. A review of automatic drum transcription. *IEEE/ACM Transactions on Audio Speech and Language Processing*, 26:1457–1483, 2018.

[21] Jort F. Gemmeke, Daniel P.W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio Set: An ontology and human-labeled dataset for audio events. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pages 776–780. Institute of Electrical and Electronics Engineers Inc., jun 2017. ISBN 9781509041176. doi: 10.1109/ICASSP.2017.7952261.

[22] Michaël Defferrard, Kirell Benzi, Pierre Vandergheynst, and Xavier Bresson. FMA: A dataset for music analysis. *Proceedings of the 18th International Society for Music Information Retrieval Conference, ISMIR 2017*, pages 316–323, 2017. URL https://arxiv.org/abs/1612.01840.

[23] Alastair Porter, Dmitry Bogdanov, Robert Kaye, Roman Tsukanov, and Xavier Serra. AcousticBrainz: A community platform for gathering music information obtained from audio. In *Proceedings of the 16th International Society for Music Information Retrieval Conference, ISMIR 2015*, pages 786–792, 2015. ISBN 9788460688532. URL https://github.com/metabrainz/.

[24] J. Salamon and J.P. Bello. Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Processing Letters*, 24: 279–283, 2017.

[25] Zohaib Mushtaq and Shun Feng Su. Environmental sound classification using a regularized deep convolutional neural network with data augmentation. *Applied Acoustics*, 167:107389, 2020.

[26] R. L. Aguiar, M. G. Y.M.G. Costa, and C.N. Silla. Exploring data augmentation to improve music genre classification with convnets. *Proceedings of the International Joint Conference on Neural Networks*, pages 1–8, 2018.

[27] Rémi Mignot and Geoffroy Peeters. An Analysis of the Effect of Data Augmentation Methods: Experiments for a Musical Genre Classification Task. *Transactions of the International Society for Music Information Retrieval*, 2(1):97–110, dec 2019. ISSN 2514-3298. doi: 10.5334/tismir.26. URL http://transactions.ismir.net/articles/10.5334/tismir.26/.

[28] Ivan Nunes da Silva, Danilo Hernane Spatti, and Rogério Andrade Flauzino. *Redes Neurais Artificiais para Engenharia e Ciências Aplicadas*. Artliber, São Paulo, 2 edition, 2016.

[29] Shian-Ru Ke, Hoang Le Uyen Thuc, Yong-Jin Lee, Jenq-Neng Hwang, Jang-Hee Yoo, and Kyoung-Ho Choi. A Review on Video-Based Human Activity Recognition. *Computers*, (2):88–131, 2013. doi: 10.3390/computers2020088. URL www.mdpi.com/journal/computers.

[30] Xiaohui Yuan, Longbo Kong, Dengchao Feng, and Zhenchun Wei. Automatic feature point detection and tracking of human actions in time-of-flight videos. *IEEE/CAA Journal of Automatica Sinica*, 4(4):677–685, oct 2017. ISSN 23299274. doi: 10.1109/JAS.2017.7510625.

[31] Romero Morais, Vuong Le, Truyen Tran, Budhaditya Saha, Moussa Mansour, and Svetha Venkatesh. Learning Regularity in Skeleton Trajectories for Anomaly Detection in Videos. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2019-June, pages 11988–11996. IEEE, jun 2019. ISBN 978-1-7281-3293-8. doi: 10.1109/CVPR.2019.01227. URL https://ieeexplore.ieee.org/document/8953884/.

[32] UFV, Conselho Técnico de Pós Graduação da Universidade Federal de Viçosa. Normas de redação de teses e dissertações. https://ppg.ufv.br/wp-content/uploads/2012/08/Normas-gerais-de-Teses-e-Dissertac%CC%A7o%CC%83es-12.pdf, 2019. Accessed: 22-10-2022.

[33] Lucas Ferreira-Paiva, Hugo Gonçalves Lopes, Elizabeth Regina Alfaro-Espinoza, Leonardo Bonato Félix, and Rodolpho Vilela Alves Neves. Towards a device for helping deaf people to dance: estimation of forro bar length using artificial neural network. *IEEE Latin America Transactions*, 20(6):970–976, 2022.

[34] Lucas Ferreira-Paiva, Elizabeth R. Alfaro-Espinoza, Vinícius Martins Almeida, Amanda Bomfim Moitinho, Leonardo Bonato Felix, and Rodolpho Vilela Alves Neves. A survey of data augmentation for audio classification. Aceito para publicação no Congresso Brasileiro de Automática - CBA 2022, .

[35] Lucas Ferreira-Paiva, Elizabeth R. Alfaro-Espinoza, Pablo de Souza Vieira Santana, Vinícius Martins Almeida, Amanda Bomfim Moitinho, Leonardo Bonato Felix, and Rodolpho Vilela Alves Neves. Forroset: A multipurpose dataset of brazilian forró music. Aceito para publicação no Advances in Artificial Intelligence - IBERAMIA 2022, .

Capítulo 2

Towards a device for helping deaf people to dance: estimation of “forró” bar length using artificial neural network

A dança tem o potencial de melhorar a qualidade de vida das pessoas, além de auxiliar na diminuição da depressão e da ansiedade. No entanto, a falta de tecnologias capazes de explorar sentidos alternativos de audição limita os efeitos benéficos da música e da dança aos ouvintes. A fim de encontrar um modelo capaz de ser implementado em dispositivos acessíveis, este trabalho avaliou o uso de um modelo baseado em redes neurais para estimar o comprimento do compasso musical do forró. As variações do modelo foram treinadas para sete conjuntos de dados compostos por misturas de amostras de música sem ruído, com ruído real e com ruído branco. Para cada conjunto de dados, a melhor variação foi selecionada e estas foram avaliadas para as mesmas amostras de ruído real. As variações do modelo que foram apresentadas para amostras com ruído real no treinamento estimaram a duração do compasso com um erro percentual médio menor que 7% na etapa de teste, sendo significativamente melhor o modelo treinado apenas com amostras contendo ruído real. O modelo avaliado foi capaz de estimar a duração do compasso musical do forró, mesmo em cenários reais, desde que apresentado a este cenário durante o treinamento. O aumento da diversidade do banco de dados e o uso de técnicas de aumento de dados podem levar a melhorias na generalização do modelo. A simplicidade do modelo avaliado e sua capacidade de aprendizado quando devidamente treinado, indicam seu potencial para ser utilizado, em tempo real, em um dispositivo móvel para passar o ritmo da música forró para surdos e deficientes auditivos.

2.1 Introdução

A perda de audição afeta cerca de 430 milhões de pessoas no mundo e tende a atingir 700 milhões de pessoas até 2050 [1]. Até meados do Séc. XX, a história das pessoas surdas e com deficiência auditiva (S/DA) foi marcada pela imposição do ora-

lismo e proibição da comunicação gestual [2, 3]. Atualmente, a comunicação gestual é reconhecida como Língua de Sinais (LS) [4]. Dentro da população S/DA, pode se caracterizar como "surdos" os indivíduos que se reconhecem como parte da "Cultura Surda", logo falantes da LS [5].

Para muitos surdos, a comunicação por LS é preferível à recuperação da audição por aparelhos auditivos e implantes cocleares [6]. A dificuldade de comunicação entre as pessoas S/DA com seus pais e cuidadores, precariza situações do cotidiano, como a alimentação [7] e resulta em maior frequência de doenças mentais e intenção suicida que pessoas ouvintes [8, 9]. Entender as demandas das pessoas com deficiências faz parte dos desafios da engenharia de reabilitação [10] e a dança pode ser um dos caminhos devido ao seu potencial de inclusão [11].

Um estudo qualitativo com crianças mostrou que a dança possibilita ver a diferença como algo comum e a apreciar a diversidade [12]. Na mesma linha, meta análises sobre o efeito da dança e dançaterapia na saúde mostraram que a dança pode produzir aumento na qualidade de vida, bem-estar, humor, afeto, imagem corporal e resultados clínicos, além de diminuição da depressão e ansiedade [13, 14]. Ainda, a participação de crianças surdas em aulas de dança com o uso de fones vibratórios proporciona melhora da autoestima [15].

A percepção da música não se restringe à modalidade auditiva uma vez que as baixas frequências da música geram vibrações que podem ser sentidas no corpo ou através de objetos [16]. Devido a essa possibilidade, há pelo menos 30 anos já se estuda estratégias para passar o ritmo das músicas por vibração [17]. Pode-se destacar o trabalho de Dong et al. [18] que propôs um sistema de detecção de batida para auxiliar dançarinos de Salsa com comandos de voz/vibração e o trabalho de Florian et al. [19] que apresentou um protótipo para ajudar pessoas surdas a sentirem o ritmo da música através de luz e vibração.

Pessoas S/DA são igualmente capazes de sincronizar com o ritmo da música através de estímulos táteis [20] e podem detectar emoções em músicas através de vibração com desempenho superior a ouvintes [21]. Em contra partida, um estudo recente identificou 83 instrumentos musicais digitais inclusivos e, destes, apenas 5 eram destinados a pessoas S/DA [22], mostrando que a inserção de surdos em contextos de música precisa ser mais explorada.

No Brasil, o forró, uma festa que virou um estilo musical, é dançado por todas as camadas da sociedade. Os passos básicos do forró são realizados ao longo de dois compassos [23], portanto, essa componente permite sinalizar a distância temporal entre o início e o fim de um passo, indicando a velocidade para se dançar a música. Devido a essa relação, a duração do compasso é utilizada em aplicações que almejam obter o ritmo de músicas de forró [24].

Visando a inclusão do público surdo em atividades culturais envolvendo o forró,

um trabalho prévio do nosso grupo de pesquisa [25] propôs um modelo computacional que estima a duração do compasso de músicas de forró utilizando uma rede *perceptron* multicamadas (PMC). No entanto, para que o modelo não tenha queda brusca de desempenho, antes de ser usado para passar o ritmo de músicas de forró para surdos por estímulos táteis, é necessário que seja treinado e validado para um cenário com ruídos reais [26].

Não foram encontrados trabalhos além de [25] que estimaram a duração do compasso diretamente, somente do BPM [27, 28, 29, 30, 31], da estrutura de divisão do compasso [32] e de ambas métricas simultaneamente [33]. Essas métricas até podem ser combinadas para se obter a duração de um compasso [24], mas exigiria que ambas fossem estimadas corretamente.

Portanto, o objetivo deste trabalho é ajustar e avaliar o modelo de [25], para músicas com ruídos reais. As contribuições em relação ao trabalho anterior são: (i) a ampliação do banco de dados proposto, com o acréscimo de 42 músicas, no qual foi levantado as respectivas durações dos compassos; (ii) a criação de um banco de dados com músicas de forró com ruídos de um espaço de dança; (iii) a avaliação do uso de ruído branco para simular o ruído real; e (iv) o ajuste e avaliação do modelo proposto por [25] para um cenário real.

2.2 Criação do banco de dados

Com a parceria de uma instrutora de forró e um projeto presente no campus de Viçosa-MG da Universidade Federal de Viçosa foram selecionadas músicas populares nos eventos de forró no campus. Foram acrescentadas 42 músicas ao banco de Paiva et al. [25], totalizando 82 músicas, majoritariamente de Forró Pé-de-serra e Forró Universitário, segundo caracterização de Junior and Volp [34], com diversidade rítmica, de músicas "mais lentas" às músicas "mais rápidas". Os títulos das músicas selecionadas, os respectivos interpretes da versão escolhida e a duração do compasso medida para cada música podem ser observados no repositório GitHub disponível em `https://github.com/NIASUFV/ForAll`.

2.2.1 Geração dos *datasets*

Na Fig. 2.1 é apresentado um fluxograma com o processo de divisão do banco de dados para a composição dos *datasets* de músicas sem ruídos, com ruídos reais e com ruído branco. As letras S, R e B, foram utilizadas para indicar Sem Ruído, Ruído Real e Ruído Branco, respectivamente, conforme listado a seguir:

- **S** — Sem Ruído

- **R** — Ruído Real
- **B** — Ruído Branco
- **SR** — Sem Ruído + Ruído Real
- **SB** — Sem Ruído + Ruído Branco
- **RB** — Ruído Real + Ruído Branco
- **SRB** — Sem Ruído + Ruído Real + Ruído Branco

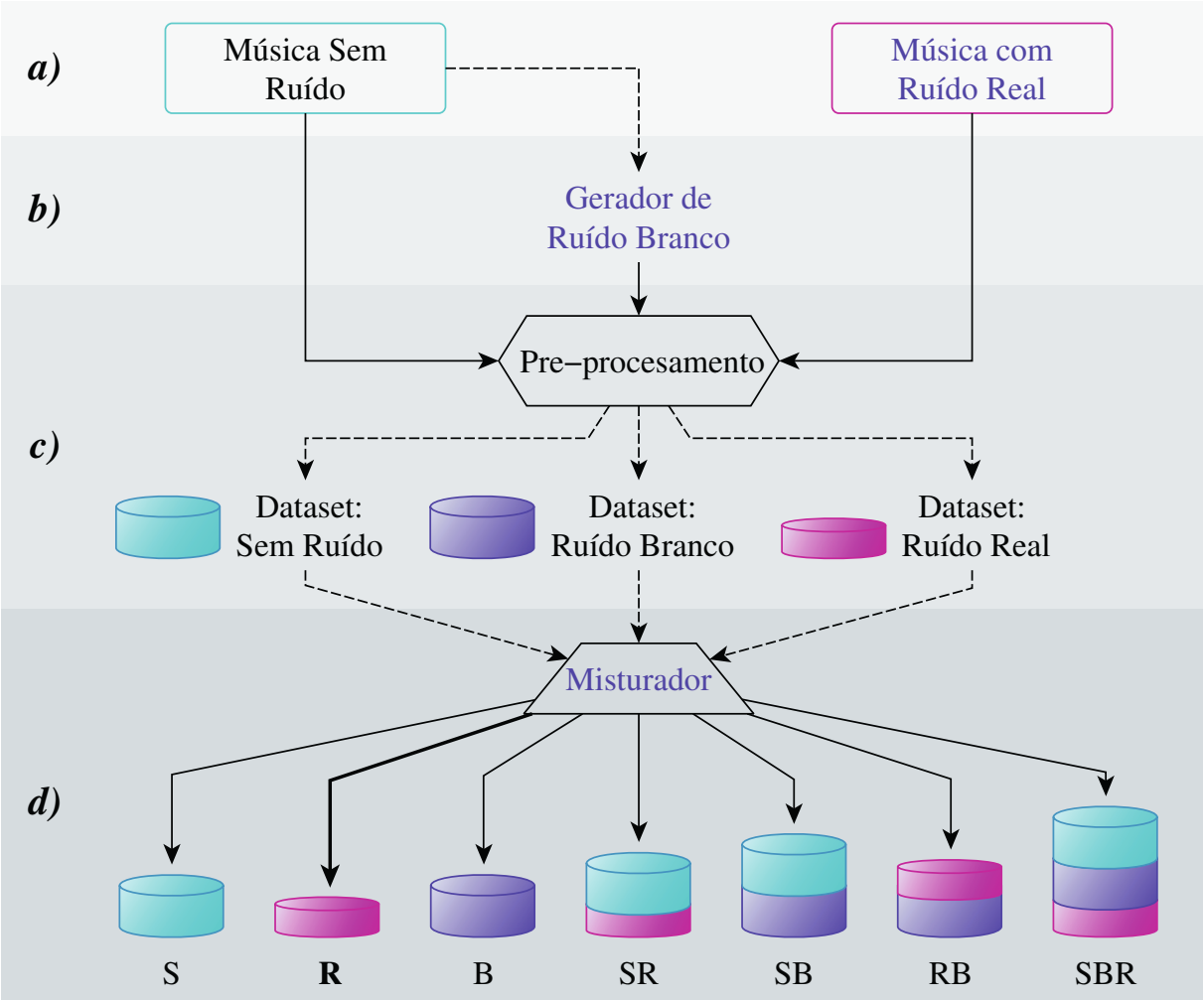


Figura 2.1: Processo de criação dos *datasets* a partir das músicas sem ruído e músicas com ruído real de um espaço de dança.

Músicas sem Ruído O banco de dados sem ruído é composto pelos 82 arquivos de áudio das músicas selecionadas em MP3, a uma taxa de amostragem de 44,1 kHz.

Músicas com Ruído Real O modelo avaliado será utilizado em um aplicativo móvel que ficará transmitindo o ritmo da música para os surdos por meio de estímulos táteis, seja pela vibração do celular ou por um dispositivo auxiliar de forma similar aos trabalhos de [18] e [35]. Portanto, se fez necessário um banco de dados real para minimizar a queda significativa do desempenho de modelos de recuperação musical criados em *Closed World* quando submetidos a dados de *Real World* [26].

De forma a obter um banco de dados com ruídos similares à aplicação, 37 faixas foram reproduzidas em um espaço de dança e regravadas por um celular Samsung Galaxy J5 localizado no bolso de um dançarino. As gravações foram feitas no formato WAV, a uma taxa de amostragem de 44,1 kHz, formando o banco "Músicas com Ruído Real".

Gerador de ruído branco Foi adicionado ruído branco às músicas sem ruído e o nível de ruído foi ajustado a um nível de relação sinal ruído de 30 dB. A inserção do ruído foi feita nas 82 músicas do banco de dados.

Preprocessamento Todos os arquivos de áudio receberam o mesmo tratamento independentemente de ter ruído ou não. Todas as etapas estão enumeradas a seguir.

1. Recorte dos 20 segundos iniciais e finais para remover os períodos de silêncio da música;
2. Segmentação dos arquivos em trechos de três segundos com sobreposição de um segundo; e
3. Normalização de cada trecho dividindo-se pelo valor eficaz do trecho, para eliminação do efeito de volume.

Misturador Partindo do pressuposto que apresentando amostras com e sem ruído durante o treinamento da rede neural a rede aprende a priorizar as entradas fundamentais, todas as combinações possíveis foram feitas com os três bancos de dados existentes. Os bancos de dados mistos foram criados concatenando as amostras de um, dois ou três *datasets* presentes.

2.2.2 Anotação dos dados

A extração do tempo do compasso foi feita de forma indireta a partir do tempo gasto para a execução de um passo base frente e trás (PBFT). O PBFT pode ser descrito em oito posições (P0-P7), como apresentado na Fig. 2.2, e é realizado ao longo de dois compassos sendo repetido indefinidamente [24].

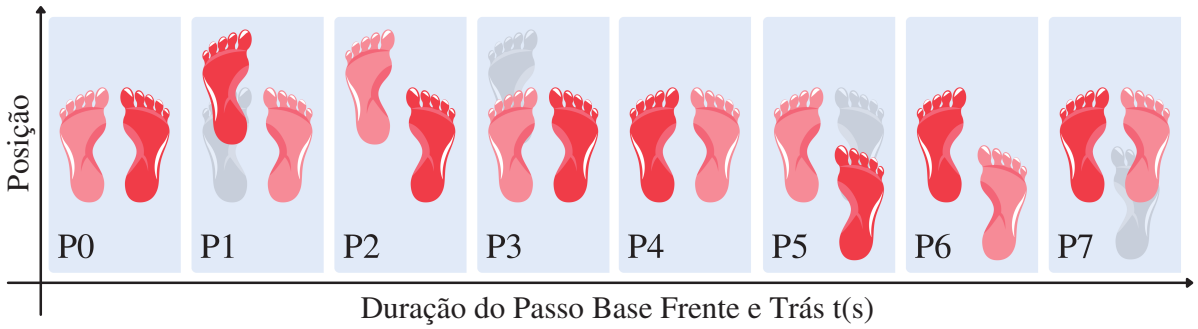


Figura 2.2: Ilustração de um passo base completo segundo [24]. O vermelho em tom mais escuro indica o pé que está recebendo a maior parte do peso do corpo.

Para medir o tempo do compasso de cada música selecionada, uma instrutora de forró dançou todas as 82 músicas realizando somente o PBFT. A medição consistiu em cronometrar o tempo gasto para realizar um passo base completo, ou seja, toda vez que a instrutora pisava com o pé direito a frente, o cronômetro era reiniciado manualmente. Foram feitas 20 medições para cada música, adotando-se a média como sendo o tempo de execução do passo base da música observada.

Finalmente, o tempo de duração do compasso se deu pela divisão do tempo de duração do PBFT por dois. Assumiu-se que a duração do compasso não variou ao longo da música, desta forma cada amostra está associada à duração do compasso da música a que pertence.

2.3 Criação do modelo de estimação

O modelo proposto por [25], apresentado na Fig. 2.3, possui um etapa de extração de características, por meio da Transformada Rápida de Fourrier, e uma etapa de estimação do compasso, por meio de uma rede PMC com uma camada oculta. O modelo foi treinado para cada um dos *datasets* e avaliado com validação *k-fold*, para definir o número de neurônios na camada oculta otimizado do modelo para cada *dataset*. Os melhores modelos foram avaliados com o *dataset* de ruído real (R), a fim de definir o *dataset* mais apropriado para capacitar o modelo a trabalhar em cenários reais.

2.3.1 Características para alimentação do modelo

Cada amostra do banco de dados consiste em um par de entradas-saída. As entradas consistem em 3 s de música de forró. As características selecionadas são as principais componentes do espectro de frequência mais grave da zabumba (50 a 300 Hz), retiradas a partir do espectro completo de cada trecho segmentado da música. O espectro é dividido em 25 faixas com espaçamento de 10 Hz. Assim, cada entrada corresponde à

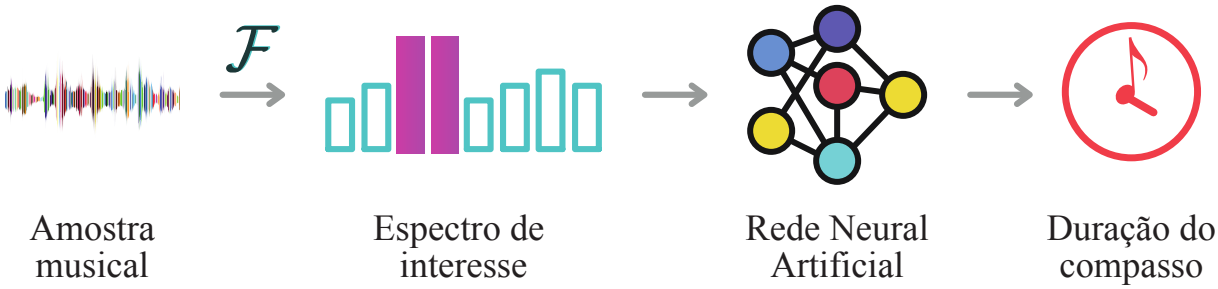


Figura 2.3: Modelo para a estimação do ritmo de músicas de forró pelo compasso musical implementado por [25].

média das amplitudes das componentes nas frequências correspondente à respectiva faixa (50-60, 60-70, ..., 290-300 Hz).

Quanto maior a amplitude da componente na frequência, maior a ocorrência do som caracterizado por esta frequência. Portanto, para uma amostra de 3 s, é esperado que músicas mais rápidas possuam maiores amplitudes das componentes na faixa da zabumba do que músicas mais lentas. Partindo destes efeitos, o modelo proposto por [25] mostrou-se capaz de estimar a duração do compasso para músicas de forró sem ruído.

2.3.2 Estrutura da rede

A rede neural escolhida foi o perceptron multicamadas com uma única camada oculta, devido a sua característica de aproximação universal de funções [36]. Outra característica desta rede é mapear qualquer função contínua no espaço das funções reais, desde que seja utilizada uma função de ativação contínua e limitada em sua imagem [36]. A seguir estão listados os hiper-parâmetros para qual o modelo foi treinado.

- Número de entradas: 25;
- Número de saídas: 1;
- Número de neurônios na camada oculta: [11, 12, ..., 51];
- Função de ativação da camada oculta: Tangente hiperbólica;
- Função de ativação da camada de saída: Linear; e
- Taxa de aprendizagem: 0,001.

Todos os hiper-paramentos são fixos, exceto o número de neurônios na camada oculta que variou de 11 a 51 conforme o critério de Fletcher-Gloss [36].

2.3.3 Treinamento da PMC

Conforme apresentado na Fig. 2.1, foram criados sete *datasets* para treinar as redes: *dataset* sem ruído (S), com ruído real (R), com ruído branco (B) e com as combinações destes (SN, SB, RB e SRB). Para cada *dataset* foram treinadas 40 variações do modelo mudando o número de neurônios na camada oculta.

A estimação de um parâmetro utilizando treinamento supervisionado consiste em apresentar para a rede um par de entradas e saídas, de modo que o treinamento possa modelar essa relação de entrada e saída [36]. O método de Levenberg–Marquardt foi utilizado para otimizar o treinamento. O método *early stopping* foi implementado para interromper o treinamento precocemente e evitar *overfitting*. Para escolha do melhor modelo foi utilizada validação cruzada *k-fold* com k=7, onde 6 *folds* foram reservados para o treinamento, o que equivale aproximadamente a 86% das amostras de cada *dataset* e um *fold* para teste. Para cada uma das sete iterações da validação *k-fold*, foram feitos dez treinamentos sorteando novos pesos iniciais, abrindo oportunidade de condições iniciais mais favoráveis ao aprendizado da rede. Ao todo foram feitos 19.600 treinamentos.

2.3.4 Desempenho dos modelos

A medida de desempenho utilizada em todos os testes foi o erro percentual médio (EPM) entre a resposta esperada e a saída da rede. Também foi observado o desvio padrão DP do EPM encontrado para cada *fold* da validação *k-fold* para avaliar a variância do erro do modelo.

Amostras da mesma natureza do treino Nesta etapa os modelos foram avaliados a partir dos desempenhos na fase de teste, que contou com amostras da mesma natureza do *dataset* de treino. Para todos os *datasets* avaliou-se o efeito do número de neurônios na camada oculta da rede PCM no desempenho no teste. O modelo com maior desempenho para cada *dataset* foi selecionado como candidato a melhor modelo para estimar a duração do compasso no cenário real.

Amostras com ruído real Os modelos selecionados com melhor desempenho na fase de teste quando treinados e testados com os *datasets* S, B, SN, SB, RB e SRB, tiveram seus desempenhos avaliados quando submetidos às amostras de teste. Esta mesmas amostras foram utilizadas para testar os modelos treinados com o *dataset* de ruído real.

Como todos os modelos foram avaliados com o mesmo conjunto de amostras nesta etapa, o teste *t student* foi utilizado comparando todos os modelos por pares a fim

de avaliar a significância de eventuais diferenças encontradas. Este teste possibilita selecionar o melhor modelo para estimar a duração do compasso para amostras reais.

Nessa etapa, além de definir qual o melhor modelo para estimar o compasso em um cenário real, é possível observar o quanto cada modelo altera o desempenho quando apresentado a um banco de dados de natureza diferente da qual foi treinado. Permitindo avaliar a necessidade de se utilizar um *dataset* fiel ao cenário que o modelo será aplicado.

2.4 Resultados

Os 82 arquivos de música sem ruído somados aos 37 arquivos de áudio gravados com ruído real resultaram em 9.632 amostras de três segundos. As durações dos compassos das músicas selecionadas variaram de 1,002 s até 1,92 s.

2.4.1 Desempenho de teste para cada *dataset*

Os desempenhos das redes na etapa de teste possibilitou avaliar o modelo proposto por [25] variando o número de neurônios na camada oculta, além de selecionar a configuração de rede mais adaptada ao problema para cada *dataset*.

Na Fig. 2.4, é possível observar que para todos os *datasets* houve queda acentuada do EPM entre 11 e 30 neurônios e a partir dessa faixa ocorre uma estabilização do erro, indicando que os benefícios de aumentar o número de neurônios na camada oculta ficam cada vez menores. Para uma implementação em dispositivo móvel ou embarcado, é mais vantajoso escolher uma rede com menor número de neurônios de camadas ocultas sem perda efetiva de desempenho.

O número de neurônios na camada oculta que ocasionou em melhor desempenho do modelo e os desempenhos para cada caso, destacado por uma estrela vermelha na Fig. 2.4, são apresentados na Tabela 2.1.

Os resultados mostrados na Tabela 2.1 indicam que o modelo avaliado foi capaz de estimar a duração do compasso de músicas de forró com erro percentual médio EPM menor que 6%, sustentando a potencialidade do modelo proposto por [25] para embarcar um aplicativo móvel que passe o ritmo de músicas de forró para surdos. Além disso, o desvio padrão da validação *k-fold* foi menor que 0,2% mostrando baixa variação de desempenhos entre as partições (*folds*). O modelo de [25] foi avaliado somente para músicas sem ruído. A rede com melhor desempenho contou com 87 neurônios e obteve EPM=3,408%. Para o *dataset* sem ruído, o melhor modelo encontrado neste trabalho possui 48 neurônios na camada oculta e EPM=5,059%, que equivale a um aumento relativo no EPM de 48%. No presente estudo o *dataset* sem ruído foi acrescido de 42 músicas o que tornou o *dataset* atual mais genérico e mais

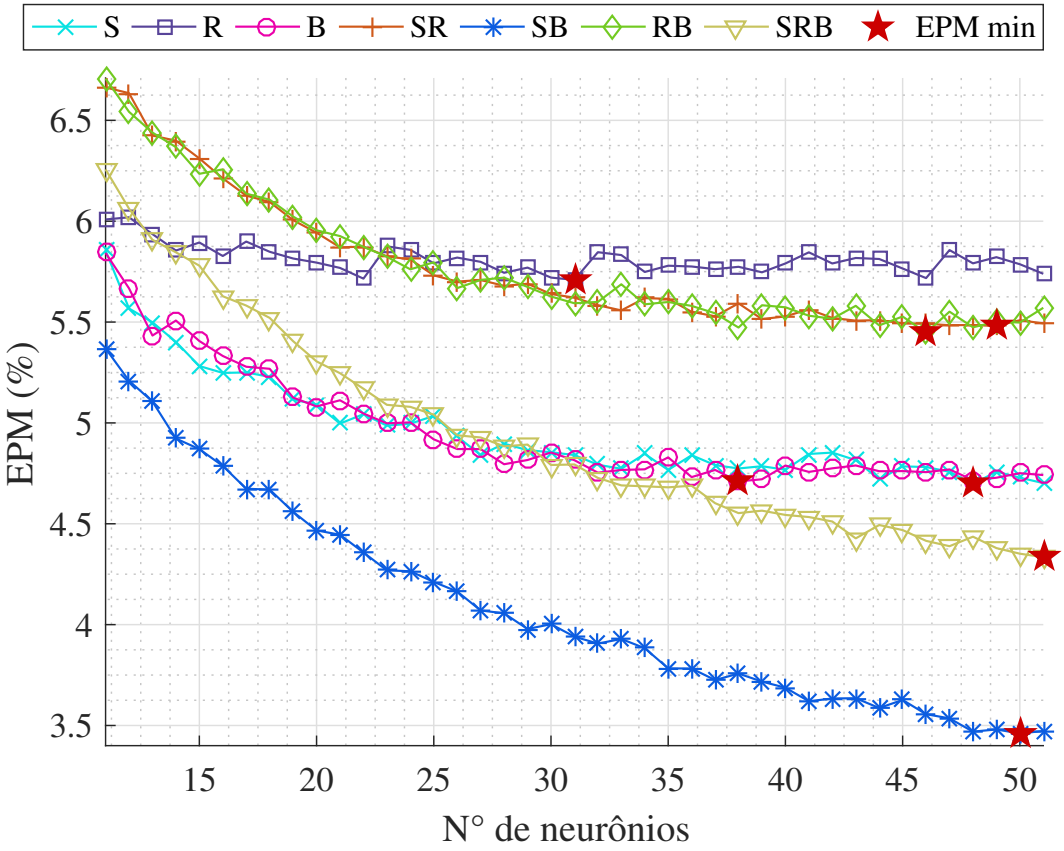


Figura 2.4: Variação do EPM no teste para cada *dataset* em função do número de neurônios na camada oculta. A estrela vermelha indica a rede com menor EPM para cada *dataset*.

Tabela 2.1: Número de amostras utilizadas para o teste, número de neurônios na camada oculta e desempenho das melhores redes, para cada *dataset* de acordo com o menor EPM e o desvio padrão (DP) obtidos com a validação *k-fold*.

Dataset	Amostras	Neurônios	EPM(%)	DP(%)
S	921	48	5,059	0,047
R	458*	31	5,541	0,116
B	921	38	4,600	0,074
SR	1313	49	5,344	0,163
SB	1843	50	3,209	0,116
RB	1313	46	5,313	0,145
SRB	2235	51	4,176	0,180
[25]	425	87	3,408	-

*As amostras de teste do dataset R em destaque também foram utilizadas para avaliar o desempenho dos demais modelos em cenário real.

complexo. A diversidade do novo *dataset* foi o principal motivo encontrado para justificar a aparente perda de desempenho do modelo, quando comparado com [25]. O *dataset* que o modelo avaliado apresentou menor erro foi o composto por áudios sem

ruído e com ruído branco SB, com EPM=3,209%, enquanto que o *dataset* que o modelo apresentou pior desempenho foi o composto por músicas gravadas com ruído real R, com EPM=5,541%. A princípio, estes resultados dão a entender que apresentar arquivos com variados ruídos facilitaria o modelo identificar as componentes principais, no entanto essa hipótese é inconsistente. O comportamento apresentado pelos modelos treinados com N e B, exibidos na Fig. 2.4, são praticamente coincidentes. O mesmo acontece com as curvas de erro dos modelos treinados com SR e RB.

Suspeita-se que o acréscimo de ruído branco aos arquivos sem ruído, não alterou de forma relevante as características espectrais na faixa de interesse e que os *datasets* S e B sejam muito parecidos. Portanto, uma amostra do *dataset* NB utilizada no treinamento pode ter uma amostra correspondente no teste, o que resultou no melhor desempenho exibido na Fig. 2.4.

2.4.2 Desempenho para o cenário real

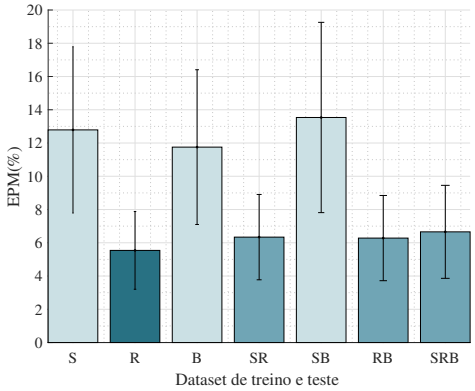
Na Fig. 2.5, pode-se observar que o modelo cujas amostras de treino pertenciam ao *dataset* R obteve o melhor desempenho quando avaliado com amostras diferentes do mesmo *dataset*. O EPM=5,541% foi significativamente menor que o dos demais ($p < 0.05$), mostrando que, para estimar a duração do compasso de músicas com ruído real, o melhor *dataset* deve conter estritamente músicas com ruídos reais.

O modelo treinado com o *dataset* SR apresentou EPM=6,227% (Fig. 2.5a), desempenho que não é significativamente diferente dos modelos treinados com RB e SRB. Apesar de apresentar o segundo melhor desempenho, o modelo treinado com SR pode ser uma alternativa para aumentar o banco de dados, visto que as músicas sem ruído são mais fáceis de obter. Vale destacar que o ruído real, apresentado neste trabalho pelo *dataset* R, é específico de uma situação de dança que é muito ruidosa. É possível que em um cenário com menos ruído, o modelo treinado com SR possa ter maior capacidade de generalização.

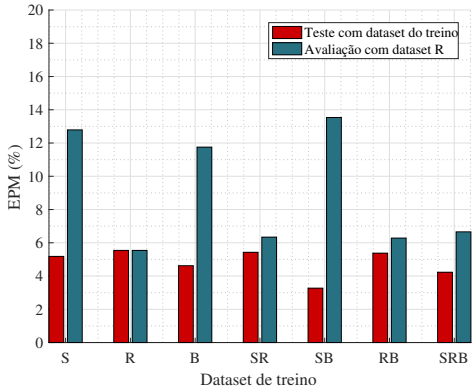
Os *datasets* menos eficientes no treinamento dos modelos foram S, B e SB com EPM>11%, mostrando a necessidade de inserção de músicas com ruídos reais no banco de dados. Isto reforça a suspeita de que a inserção do ruído branco não fez mudanças significativas no *dataset* e que S e B são muito semelhantes.

Na Fig. 2.5b, destaca-se a queda de desempenho de todos os modelos ao serem submetidos ao *dataset* exclusivo de ruídos reais. Esse fenômeno era esperado conforme discutido por [26], ressaltando a importância de utilizar bancos de dados reais para avaliar a aplicabilidade do modelo em mundo real.

O modelo que mais sofreu com os ruídos reais foi o treinado com SB que foi o modelo com menor EPM na etapa de teste. O erro deste modelo aumentou de 3,209% para 12,950%, que equivale à um aumento de aproximadamente 300%. Este aumento



(a) Resultado dos modelos quando apresentados ao *dataset* R. As barras com cores diferentes possuem valores significativamente diferentes ($p < 0.05$).



(b) Comparação do desempenho obtido pelas melhores redes quando apresentadas ao *dataset* R em relação à etapa de teste.

Figura 2.5: Desempenho dos modelos *dataset* e ruídos reais e sua comparação com os desempenhos obtidos na etapa de teste.

indica a baixa capacidade de generalização do modelo, fruto da pouca diversidade do *dataset* SB.

O modelo treinado com R não apresentou queda de desempenho (barra verde idêntica à vermelha na Fig. 2.5b) porque o *dataset* utilizado para ambos os testes foram os mesmos, conforme explicado anteriormente. Todos os modelos apresentados a amostras com ruído real durante o treino alcançaram erro médio absoluto menor que 100 ms, onde o modelo treinado com o dataset R, poderá passar o ritmo de músicas de forró para surdos com erro de 78 ms para cima ou para baixo do valor de compasso da música.

2.5 Desempenho do estado da arte em estimação de componentes musicais

Na Tabela 2.2 são apresentados trabalhos que estimaram componentes musicais, as componentes estimadas por eles e os desempenhos dos melhores modelos. As componentes estimadas foram BPM e estrutura do compasso (EC). Para a Acurácia 1 e o F1, foram considerados acertos diferenças de até 4%.

O resultado do modelo proposto possui Acurácia1 inferior a todos os trabalhos revisados, os melhores desempenhos foram 79% para a estimação do BPM [27] e 74,5% para estrutura do compasso [29], conforme apresentado na Tabela 2.2, sendo difícil inferir o desempenho final de uma eventual combinação dos modelos apresentados nesses trabalhos. Quando se observa as métricas de regressão, o modelo avaliado apresenta desempenhos mais otimistas, com EPM 5,541% e R2=0,771. No entanto,

Tabela 2.2: Paralelo do desempenho do modelo citado com a literatura.

Ref	Saída	Acu1	F1	EPM	R2
[27]	BPM	**79**	-	-	-
[29]	EC	-	82	-	-
[30]	BPM	73,4	-	-	-
[31]	BPM	77,41	-	-	-
[33]	BPM	84,6	-	-	-
	EC	**74,5**	-	-	-
Avaliado	DC	48,035	64,897	5,541	0,771

os trabalhos anteriores trataram a tarefa de estimação do BPM como uma tarefa de classificação binária, portanto, não foram apresentados desempenhos para métricas de regressão, impossibilitando uma comparação mais justa com a literatura.

2.6 Conclusão

Este trabalho propôs a avaliação de um modelo baseado em rede PMC para futuramente embarcar um aplicativo que passe o ritmo de músicas de forró para surdos em um espaço de dança por meio de vibração. O modelo avaliado foi treinado com sete composições de *dataset* e a apresentou erro menor que 6% em todos os cenários. Os modelos que não foram apresentados à amostras com ruídos reais durante o treinamento foram significativamente mais afetados, com erro se aproximando a 15% quando testados com amostras reais. Os modelos que foram apresentados às músicas com ruídos reais durante o treino mantiveram erro inferior a 7% na fase de teste, mas foram inferiores ao modelo treinado somente com amostras com ruídos reais, que obteve erro significativamente menor de 5,541% que equivale a um erro médio absoluto de 78 ms.

Foi comprovada a capacidade do modelo avaliado de aprender a duração do compasso de músicas de forró em variados cenários, desde que tenha sido apresentado a amostras da mesma natureza durante o treino. Além disso, o ruído branco foi ineficaz para simular os ruídos de um espaço de dança. Portanto, o uso de músicas com ruídos reais no treinamento foi essencial para capacitar o modelo a estimar o compasso neste cenário, sendo o mais indicado para embarcar, futuramente, o aplicativo para auxiliar surdos a dançarem.

O presente trabalho apresenta duas fortes limitações. A primeira delas está relacionada ao número reduzido de músicas no banco de dados, principalmente do *dataset* com ruído real, que teve metade do tamanho. A fim de aumentar a diversidade das amostras de treino e melhorar a generalização do modelo treinado, é necessário acrescentar músicas no banco de dados. Além disso, uso de técnicas de *data augmen-*

tation podem ser exploradas, bem como, acréscimo outras situações do cotidiano no *dataset* com ruídos reais, como músicas reproduzidas por computador ou celular em ambientes domésticos.

A segunda limitação consiste no método adotado para a seleção do número de neurônios e a validação do modelo. A abordagem proposta permitiu que amostras diferentes de uma mesma música fossem utilizadas nos *datasets* de treino, validação e teste. Desta forma, o modelo pode estar reconhecendo a música na fase de teste e estar associando com a duração do compasso das amostras que estiveram no treino. Neste caso, o modelo poderá ter desempenho menor quando apresentado a uma música que não teve amostras compondo os dados de treino. Para avaliar essa hipótese será necessário testar o modelo com músicas novas e avaliar possível queda de desempenho.

Em trabalhos futuros será necessário avaliar o quanto o erro obtido poderá atrapalhar a performance dos surdos dançando e qual limiar de erro será necessário alcançar para que a aplicação contribua de forma positiva para a experiência dos surdos com o forró.

Referencias

[1] WHO. Deafness and hearing loss, mar 2021. URL https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss.

[2] Douglas C Baynton. *Forbidden signs: American culture and the campaign against sign language*. University of Chicago Press, 1996.

[3] Maria Aparecida Leite Soares. *A educação do surdo no Brasil*. Editora Autores Associados, 2015. ISBN 978-85-7496-345-7.

[4] Ronice Müller de Quadros and Lodenir Becker Karnopp. *Língua de sinais brasileira: estudos lingüísticos.* Artmed, 2007. ISBN 978-85-363-1174-6.

[5] Carol A. Padden and Tom Humphries. *Inside deaf culture*. Harvard University Press, 2009. ISBN 978-0-674-015067.

[6] Bonnie Poitras Tucker. Deaf culture, cochlear implants, and elective disability. *Hastings Center Report*, 28(4):6–14, 1998. doi: 10.2307/3528607.

[7] Poorna Kushalnagar, Christopher J Moreland, Abbi Simons, and Tara Holcomb. Communication barrier in family linked to increased risks for food insecurity among deaf people who use american sign language. *Public Health Nutrition*, 21 (5):912–916, 2018. doi: 10.1017/S1368980017002865.

[8] Johannes Fellinger, Daniel Holzinger, and Robert Pollard. Mental health of deaf people. *The Lancet*, 379(9820):1037–1044, 2012. ISSN 0140-6736. doi: 10.1016/S0140-6736(11)61143-4.

[9] Meghan L Fox, Tyler G James, and Steven L Barnett. Suicidal behaviors and help-seeking attitudes among deaf and hard-of-hearing college students. *Suicide and Life-Threatening Behavior*, 50(2):387–396, 2020. doi: 10.1111/sltb.12595.

[10] Rory A. Cooper and Rosemarie Cooper. Rehabilitation engineering: A perspective on the past 40-years and thoughts for the future. *Medical Engineering and Physics*, 72:3–12, 2019. ISSN 1350-4533. doi: 10.1016/j.medengphy.2019.08.011.

[11] Gaetano Raiola. Inclusion in sport dance and self perception. *Sport Science*, 8(1): 99–102, 2015.

[12] Michelle R Zitomer. Children's perceptions of disability in the context of elementary school dance education. *Revue phénEPS/PHEnex Journal*, 8(2), 2016.

[13] Sabine Koch, Teresa Kunz, Sissy Lykou, and Robyn Cruz. Effects of dance movement therapy and dance on health-related psychological outcomes: A meta-analysis. *The Arts in Psychotherapy*, 41(1):46–64, 2014. ISSN 0197-4556. doi: 10.1016/j.aip.2013.10.004.

[14] Sabine C Koch, Roxana F. F. Riege, Katharina Tisborn, Jacelyn Biondo, Lily Martin, and Andreas Beelmann. Effects of dance movement therapy and dance on health-related psychological outcomes. a meta-analysis update. *Frontiers in Psychology*, 10:1806, 2019. ISSN 1664-1078. doi: 10.3389/fpsyg.2019.01806.

[15] Fu Hai Frank Wu and Jyh Shing Roger Jang. A supervised learning method for tempo estimation of musical audio. In *2014 22nd Mediterranean Conference on Control and Automation*, pages 599–604, 2014. ISBN 9781479959006. doi: 10.1109/MED.2014.6961438.

[16] Edith Van Dyck, Dirk Moelants, Michiel Demey, Alexander Deweppe, Pieter Coussement, and Marc Leman. The impact of the bass drum on human dance movement. *Music Perception*, 30(4):349–359, 2013. ISSN 0730-7829. doi: 10.1525/mp.2013.30.4.349.

[17] Masashi Ezawa. Rhythm perception equipment for skin vibratory stimulation. *IEEE Engineering in Medicine and Biology Magazine*, 7(3):30–34, 1988. ISSN 07395175. doi: 10.1109/51.7932.

[18] Yudi Dong, Jian Liu, Yingying Chen, and Woo Y. Lee. Salsaasst: Beat counting system empowered by mobile devices to assist salsa dancers. In *2017 IEEE 14th*

International Conference on Mobile Ad Hoc and Sensor Systems, pages 81–89, 2017. ISBN 9781538623237. doi: 10.1109/MASS.2017.25.

[19] H Florian, Adrian Mocanu, Cristian Vlasin, José Machado, Vitor Carvalho, Filomena Soares, Adina Astilean, and Camelia Avram. Deaf people feeling music rhythm by using a sensing and actuating device. *Sensors and Actuators A: Physical*, 267:431–442, 2017. ISSN 0924-4247. doi: 10.1016/j.sna.2017.10.034.

[20] Pauline Tranchant, Martha M Shiell, Marcello Giordano, Alexis Nadeau, Isabelle Peretz, and Robert J Zatorre. Feeling the beat: Bouncing synchronization to vibrotactile music in hearing and early deaf people. *Frontiers in Neuroscience*, 11: 507, 2017. ISSN 1662-453X. doi: 10.3389/fnins.2017.00507.

[21] Andréanne Sharp, B. A. Bacon, and F. Champoux. Enhanced tactile identification of musical emotion in the deaf. *Experimental Brain Research*, 238(5):1229–1236, 2020. ISSN 14321106. doi: 10.1007/s00221-020-05789-9.

[22] Emma Frid. Accessible digital musical instruments-a review of musical interfaces in inclusive music practice. *Multimodal Technologies and Interaction*, 3(3):57, 2019. ISSN 2414-4088. doi: 10.3390/mti3030057.

[23] Antonio Carlos de Quadros Junior, Ellen Cristina Fontes, Romualdo Dias, and Catia Mary Volp. Caracterização do xote e do baião dançados no interior do estado de são paulo. *Movimento*, 15(3):233–247, 2009. ISSN 1982-8918. doi: 10.22456/1982-8918.2347.

[24] Augusto Dias Pereira Dos Santos, Lie Ming Tang, Lian Loke, and Roberto Martinez-Maldonado. You are off the beat! is accelerometer data enough for measuring dance rhythm? In *ACM International Conference Proceeding Series*, 2018. ISBN 9781450365048. doi: 10.1145/3212721.3212724.

[25] Lucas F Paiva, Hugo G Lopes, Leonardo B Felix, and Rodolpho VA Neves. Estimação do compasso musical do forró utilizando rede perceptron multicamadas. In *Anais do Congresso Brasileiro de Automática*, volume 2, 2020. doi: 10.48011/asba.v2i1.1331.

[26] C. W. Wu, C. Dittmar, C. Southall, R. Vogl, G. Widmer, J. Hockman, M. Muller, and A. Lerch. A review of automatic drum transcription. *IEEE/ACM Transactions on Audio Speech and Language Processing*, 26:1457–1483, 2018.

[27] A.J. Eronen and A.P. Klapuri. Music tempo estimation with *k*-nn regression. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(1):50–57, 2010. ISSN 1558-7916. doi: 10.1109/TASL.2009.2023165.

[28] Fu Hai Frank Wu and Jyh Shing Roger Jang. A supervised learning method for tempo estimation of musical audio. In *22nd Mediterranean Conference on Control and Automation*, pages 599–604, 2014. ISBN 9781479959006. doi: 10.1109/MED.2014.6961438.

[29] Elio Quinton, Mark Sandler, and Christopher Harte. Extraction of metrical structure from music recordings. In *18th International Conference on Digital Audio Effects*, pages 1–7, 2015.

[30] Sebastian Böck, Florian Krebs, and Gerhard Widmer. Accurate tempo estimation based on recurrent neural networks and resonating comb filters. In *16th International Society for Music Information Retrieval Conference*, pages 625–631, 2015. ISBN 9788460688532.

[31] Hendrik Schreiber and Meinard Müller. A post-processing procedure for improving music tempo estimates using supervised learning. In *18th International Society for Music Information Retrieval Conference*, pages 235–242, 2017. ISBN 9789811151798.

[32] Sankalp Gulati, Vishweshwara Rao, and Preeti Rao. Meter detection from audio for indian music. In *Speech, Sound and Music Processing: Embracing Research in India*, volume 7172, pages 34–43, 2012. ISBN 9783642319792. doi: 10.1007/978-3-642-31980-8_3.

[33] C Uhle and J Herre. Estimation of tempo, micro time and time signature from percussive music. In *Proc. of the 6th Int. Conference on Digital Audio Effects*, pages 1–6, 2003.

[34] Antonio Carlos de Quadros Junior and Catia Mary Volp. Forró universitário: a tradução do forró nordestino no sudeste brasileiro. *Motriz*, 11(2):127–120, 2005. ISSN 1980-6574. doi: 10.5016/171.

[35] Josef Roth, Jan Ehlers, Christopher Getschmann, and Florian Echtler. Tempowatch: A wearable music control interface for dance instructors. In *Proceedings of the Fifteenth International Conference on Tangible, Embedded, and Embodied Interaction*, 2021. ISBN 9781450382137. doi: 10.1145/3430524.3442461.

[36] Ivan Nunes da Silva, Danilo Hernane Spatti, and Rogério Andrade Flauzino. *Redes Neurais Artificiais para Engenharia e Ciências Aplicadas*. Artliber, São Paulo, 2 edition, 2016.

Chapter 3

A Survey of Data Augmentation for Audio Classification

One of the most effective methods for reducing overfitting in deep learning models for audio classification is data augmentation. The range of techniques available, as well as a lack of understanding of the most efficient ones, can result in severe time and processing power costs. This survey covers numerous techniques, tools, and datasets for offline data augmentation to assist in the selection and implementation of data augmentation strategies to improve audio classification models in Environmental Sound Classification, Music Information Retrieval, and Automatic Speech Recognition. Finally, we present a short review of papers that apply data augmentation in Environmental Sound Classification which indicates that the use of spectrogram and audio augmentation has considerable potential for improving the performance of convolutional models, especially for small datasets with increases in accuracy of up to 30%. However, the accuracy gains achieved may be insufficient to justify the additional computer burden depending on the application. Furthermore, the usage of image data augmentation is unsuitable for audio data.

3.1 Introduction

Data Augmentation (DA) is defined as the creation of new data by adding deformations to increase the variety of the data so that these deformations do not change their semantic value. DA application for audio signals, including natural, and non-natural sounds, can be categorized accordingly to where the DA techniques are applied: the raw audio or to its spectrogram. Classification is the most important task in Environmental Sound Classification (ESC) and is highly noted in Music Informational Retrieval (MIR) and Automatic Speech Classification (ASR).

In ESC, the applications include urban noise recognition and mitigation [1, 2], and identification of animals by their sounds [3, 4]. Genre [5], instruments [6], and emotion [7] classification are prominent areas in MIR, while in ASR, speech commands classification [8, 9] is a traditional task. Convolutional Neural Network (CNN) is the

most widely model used in audio applications [5, 6, 10, 11, 4, 12, 13, 14].

However, when faced with small datasets, CNN's capacity for information retention becomes a flaw; the models memorize the training data and lose performance on new data [15]. To tackle the issue of overfitting, DA techniques can be used to improve the performance of the model [12, 5, 13, 14].

These explored techniques can be implemented in a variety of programming languages using multiple deformations. The variety of options available can make the DA application in audio a challenging task, resulting in excessive use of time and computational power, and a decline in model performance, especially for newcomers.

In order to enhance the process of learning and applying data augmentation strategies, this survey aims: (i) to provide an overview of the most used strategies to current augment audio data research; (ii) to present the main techniques for each data augmentation tool and packages; (iii) to discuss open datasets for implementing and validating CNN models and data augmentation techniques; and (iv) to thoroughly examine the advantages and shortcomings of data augmentation for convolutional audio classification models using ESC research articles as a case of study.

3.2 Offline Data Augmentation

Data Augmentation can be applied to audio samples directly or after the spectrogram has been extracted. Regardless of where the augmentation occurs, the samples generated receive the same annotation as the original sample. Fig. 3.1 depicts the usage of convolutional models with supervised training for audio classification [4, 14]. In this approach, the augmentation can be done to audio files and spectrograms are applied to both the original and augmented samples (Fig. 3.1a). Another possibility is to convert the original samples into spectrograms and then image or spectrograms augmentation techniques are used (Fig. 3.1b). Following this process, the spectrograms utilized to train the models are made up of both original and augmented data (Fig. 3.1c). In the test step (Fig. 3.1d), the trained model is applied to the new data without deformations.

3.3 Data Augmentation for Audio Classification

This section presents the deformation techniques, deployed in recent papers that used data augmentation for audio classification in ESC, MIR, and ASR. As the nomenclature of the procedures employed differed throughout the examined papers and tools, a standardization of the terms used was sought. In this section we organized the techniques into three major groups: Audio Data Augmentation (ADA), Image Data

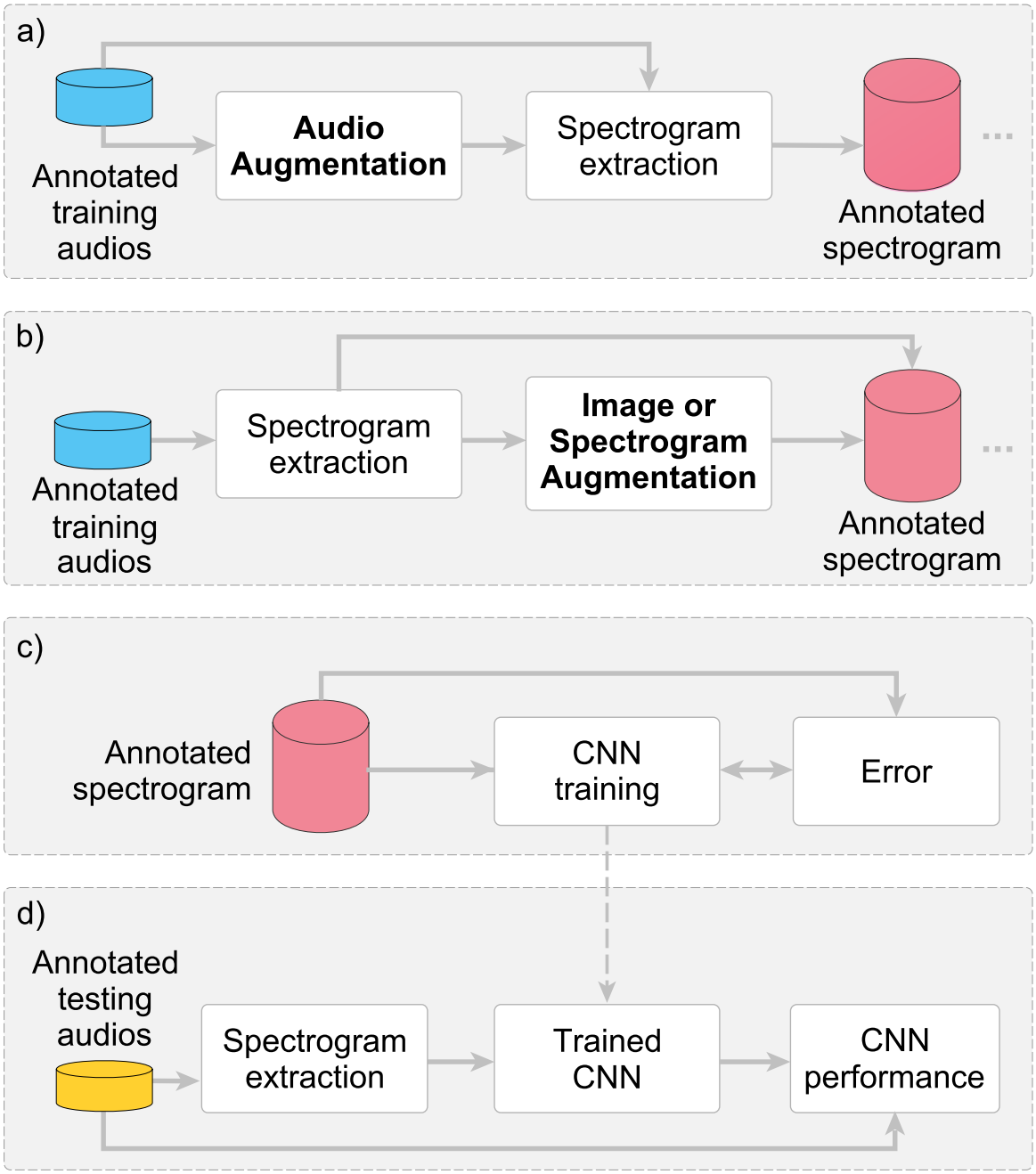


Figure 3.1: Offline augmentation approach. a) Audio augmentation. b) Image/spectrogram augmentation. c) CNN training with added augmented data. d) Validating the model with new data without augmented data.

Augmentation (IDA), and Spectrogram Data Augmentation (SDA).

3.3.1 Audio Data Augmentation

Audio augmentation approaches introduce deformations directly to the raw audio, with the spectrogram generated from the previous ones [12, 5, 4, 13, 14]. An example of ADA technique is shown in Fig. 3.2.

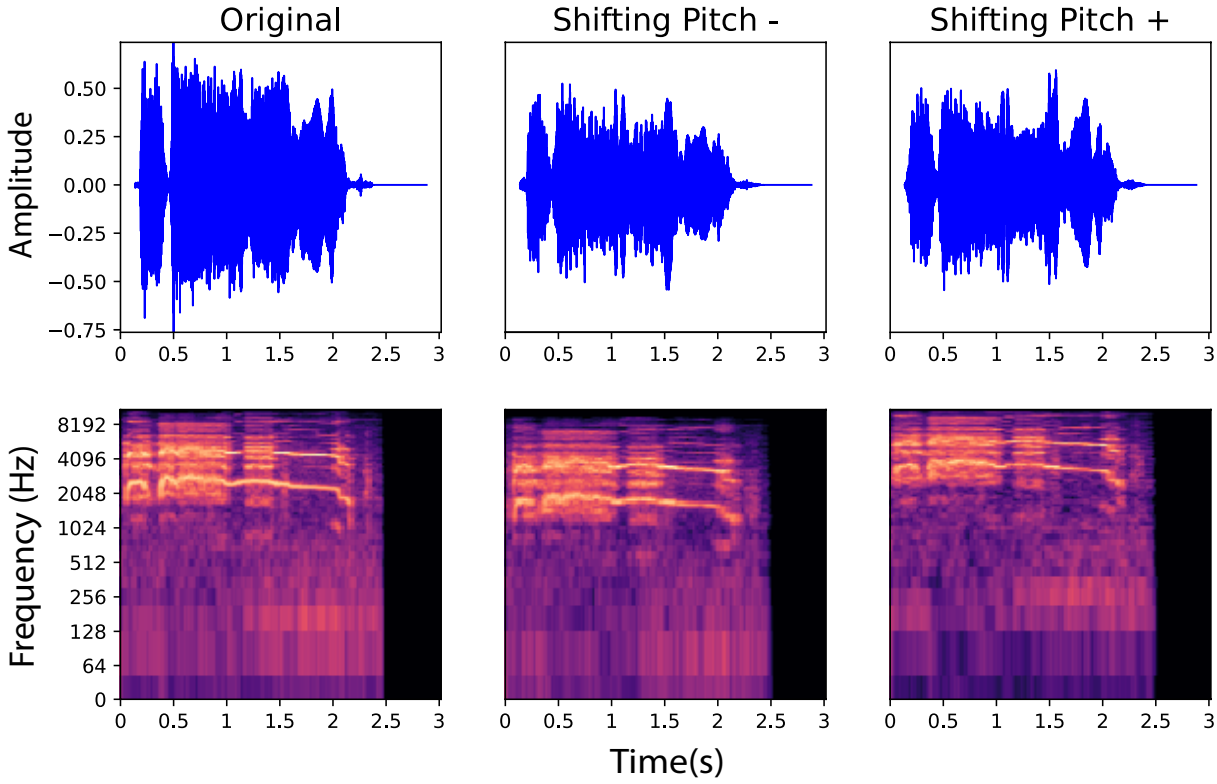


Figure 3.2: The decrease (-) and increase (+) in pitch for a rooster crowing and its effects on the log-mel spectrogram.

Shifting Pitch (SP)

In this technique, the pitch of each audio signal in the datasets is increased or decreased by a factor and the duration remains the same as shown in Fig. 3.2.

Time Stretching (TS)

It slows down or speeds up an audio sample by a preset ratio without altering the pitch drastically. In Mushtaq and Su [13], the authors used 1.2 and 0.7 to produce quicker and slower samples.

Volume Adjustment (VA)

It is done by varying the loudness of the audio file.

- *Loudness (L).* It increases and decreases the volume of all samples at a random or fixed rate. For example, -10 and +10 dB were used in Aguiar et al. [5].
- *Dynamic Range Compression (DRC).* It distorts samples by altering the loudness of the original sample using different noises. Salamon and Bello [12] used music standards, film standards, speech, and radio.

Noise (N)

It introduces noises into the original samples.

- *Background Noise (BN)*. It consists of mixing the original audios with everyday noise. Salamon and Bello [12] combined the original audio with noises from street workers, street traffic, street people, and parks.
- *Synthetic Noise (SN)*. It creates new samples by combining audio and synthetic noise. White noise, for example, as seen in Mushtaq and Su [13].

Silence Trimming (ST)

It eliminates the silence present at the start and end of each sample.

Time Shifting (TiS)

It shifts the audio to the left/right by a random factor.

SpeedUp (SU)

The signal is re-sampled at a preset sampling rate and later returned at the original sampling rate, resulting in a speed change.

Wow Resampling (WR)

The resampling frequency oscillates around the original sampling rate with a given frequency and amplitude, similar to SP, but with the intensity changing over time. The transformation is provided in (3.1) where x is the input signal. Nanni et al. [4] have used the amplitude $a_m = 3$ and the fundamental frequency $f_m = 2$.

$$F(x) = x + a_m \frac{sin(2\pi f_m x)}{2\pi f_m} \tag{3.1}$$

Clipping (C)

The audio sample is normalized so that a specific amount of points are saturated. The out-of-range samples are then clipped.

Harmonic Distortion (HD)

The transformation $sin(x)$ is applied many times, resulting in a saturation effect.

Impulse Response (IR)

An audio signal is convolved with a unitary response.

Filter (F)

It applies several kinds of filtering to the input audio. These are some common filters: band-pass, band-stop, high-pass, high-shelf, low-pass, low-shelf, and peaking filter.

Random Mask (RM)

The random frequencies or audio parts are masked.

MP3 Compression (MC)

This function compresses the audio to reduce its quality by using an encoder.

Inversion (I)

It can be performed on the y-axis, multiplying the audio by -1 or inverting the audio along the x-axis.

Peak Normalization (PN)

The highest signal level in the song is set to 0 dBFS. The loudest level must be at [-1,1].

Tangent Distortion (TD)

It adds distortion to guitars changing the timbre of the sound when applied hyperbolic tangent function.

3.3.2 Image Data Augmentation

In this section, it is presented some traditional deformations techniques used in computer vision applications as data augmentation, which has been used for audio classification [4, 13].

Flip (F)

The rows and columns of pixels can be reversed, as presented in Fig. 3.2.

Zoom Range (ZR)

It applies zooms randomly and augments new pixels around the image.

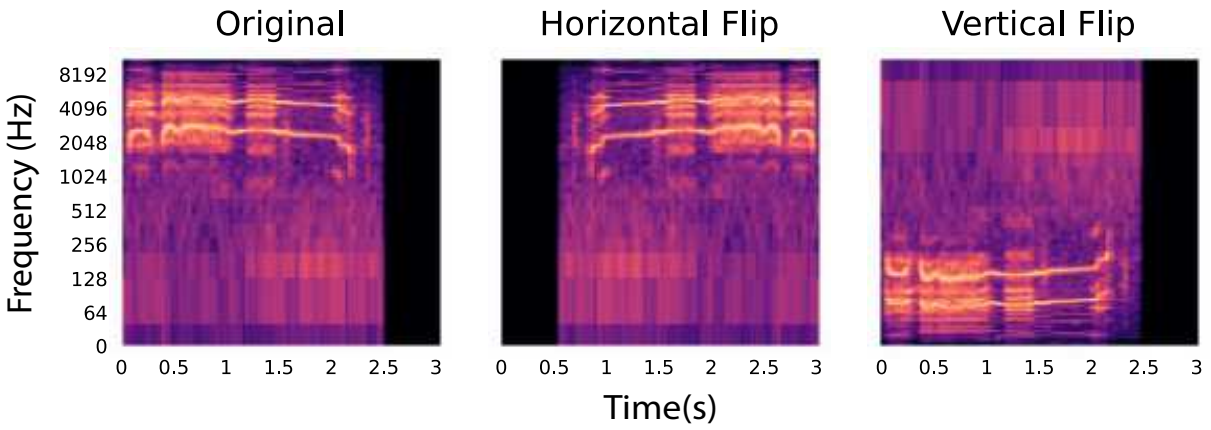


Figure 3.3: Horizontal and vertical flip for the rooster crowing present in Fig. 3.2.

Shift (S)

The pixels are moved in one direction that can either be vertically or horizontally, maintaining the size of the image.

Rotation Angle (RA)

The image is randomly rotated clockwise in the range from 0° to 360°.

Brightness Range (BR)

Both randomly darkening and brightening can augment the brightness of the image. This technique simulates the VA technique, which was presented previously.

Shear Range (SR)

It causes distortions along an axis that simulate the visualization of an object from different perspectives.

The S and SR techniques must be applied with low intensity since these deformations can change the semantic value of the signal. On the other hand, F, ZR, and RA techniques completely distort the signal and cannot be considered data augmentation techniques in the context of audio signals.

3.3.3 Spectrogram Data Augmentation

This class of data augmentation is similar to image augmentation because it is applied to spectral images, however, they are selected especially for audio applications [16, 4]. Fig. 3.4 shows an example of a random mask technique.

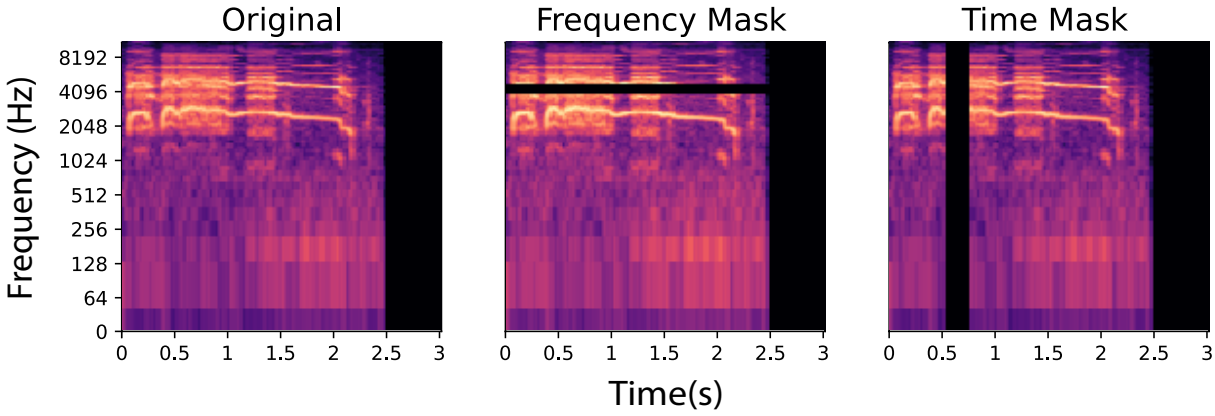


Figure 3.4: Frequency and time mask for the rooster crowing present in Fig. 3.2.

Spectrogram Random Shifts (SRS)

This technique randomly applies pitch shift and time shift simultaneously.

Spectrogram Sound Mix (SSM)

SSM creates a new image by summing the two random spectrograms of the same class.

Vocal Tract Length Normalization (VTLN)

It's an ASR technique that distorts the spectrum in the direction of a medium-level vocal treatment. In ASR, this technique is used to remove the variability that exists between the two vocal tracts length from each speaker [17].

Equalized Mixture (EM)

The weighted average of two randomly picked spectrograms with the same label [18].

Spectrogram Time Shift (STS)

It is a change in time that consists of dividing the spectrum into two parts and later restoring them in a reverse order.

Spectrogram Random Mask (SRM)

It consists of removing portions of spectrograms. The frequency and time mask are shown in Fig. 3.4.

Spectrogram Channel Shuffle (SCS)

It shuffles the channels of the spectrogram. By using this function, bias may be mitigated.

3.4 Augment Data Tools

A set of data augmentation tools is presented in this section to help in their implementation. The functionalities available for each data audio augmentation library are listed in Table 3.1. In addition, all of the librosa, Keras, and audiomentations techniques were implemented in a GitHub-hosted Jupyter notebook[1].

Table 3.1: Audio and image augmentation techniques are available for each of the presented tools.

Tool	ADA	IDA/SDA
librosa [19]	TS, SP, ST	
MUDA [20]	TS, SP, BN, DRC	
SoX [21]	TS, SP, ST, L, SN, BN, ST, SU, F	
audiogmenter [16]	SP, L, DRC, SN, SU, WR, C, HD, IR, F	SRS, SSM, VTLN, EM, STS, SRM
Keras [22]		F, ZR, S, RA, BR, SR
audiomentations [23]	SP, TS, VA, L, BN, SN, ST, TiS, C, IR, F, RM, MC, I, PN, TD	SRM, SCS

3.4.1 librosa: Python Audio and Music Analysis

The librosa Python package[2] was designed to evaluate audio and music signals [19]. Its 0.8.1 version includes three techniques to perform offline data augmentation. It is widely used for the extraction of musical characteristics and data augmentation in audio [13, 14].

3.4.2 MUDA: A Software for Increasing Musical Data

The Musical Data Augmentation package[3] implements annotation-based musical data augmentation [20]. This software is based on JAMS (JSON annotated music specification) and it enables the creation of custom deformations as the tracking of data provenance [12]. Its 0.4.1 version provides four DA techniques that can be customized to enlarge audio files.

[1] https://github.com/lucas-fpaiva/survey-audio-aug
[2] https://github.com/librosa/librosa
[3] https://github.com/bmcfee/muda

3.4.3 Audiogmenter: A MATLAB Tool for Augmenting Audio Data

Audiogmenter[4] was the first MATLAB library designed specifically for audio data augmentation [16]. It includes 23 data augmentation methods, such as image and audio augmentation. The library is free, but MATLAB is a paid software.

3.4.4 SoX: the Swiss Army Knife of Audio Manipulation

Sound eXchange [21] is a cross-platform command-line tool that can convert various audio file formats to others. It can apply various effects to audio files. In addition, pysox[5] is a Python wrapper for this tool [24], which is widely used in combination with SoX. SoX was also used for DA by Aguiar et al. [5] in the implementation of MUDA and audiogmenter packages. SoX 14.4.2 is the newest version.

3.4.5 Keras: An API for Deep Learning in Python

Keras [22] is a Python-based machine learning package that can be used in TensorFlow and in other programming languages, such as R. It allows quick experimentation and results. The package includes image augmentation functionalities that can also be applied to spectrograms for audio problems [14].

3.4.6 Audiomentations: A Python library for Audio Data Augmentation

Audiomentations[6] is a Python package [23] that can be used in Tensorflow/Keras or Pytorch training pipelines. Its 0.20.0 version has 28 ADA and two SDA techniques.

3.5 Datasets

In this section, we briefly introduce ESC, MIR, and ASR popular open datasets for audio classification tasks.

3.5.1 Urbansound8k

Urbansound8k[7] or US8K [25] is a subset of 4-second audio clips containing 8732 audio files. It is not uniformly distributed across its ten folds. The folds are air conditioner,

[4] https://github.com/LorisNanni/Audiogmenter
[5] https://github.com/rabitt/pysox
[6] https://github.com/iver56/audiomentations
[7] https://urbansounddataset.weebly.com/

dog bark, car horn, children playing, gunshot, engine idling, street music, siren, jackhammer, and drilling. It is derived from UrbanSound which was created by manually filtering and labeling the tracks from Freesound [26].

3.5.2 ESC-10 and ESC-50

ESC-10 and ESC-50 are part of the ESC dataset[8] [27] of urban environment 5-second audio recordings. Both datasets are extracted from Freesound [26], and organized into 5 uniformly sized cross-validation folds. ESC-10 has 400 clip recordings with a total time duration of 33 minutes and a clip rate of 50 clips per fold. These folds include sounds like a dog barking, a crackling fire, baby cries, rain, sneezing, a rooster, sea waves, a helicopter, a chainsaw, and a clock ticking. ESC-50 is more complex than the other due to its 50 folds, which are divided into five major categories. Sounds of animal, non-speech human, urban or outdoor noises, indoor noises, and various natural soundscapes are included in these fold. It contains 2000 sound clips and runs for 168 minutes.

3.5.3 CatSound

The CatSound Classification dataset[9] [28, 29] contains over three hours of domestic cat 4-second audio recordings divided into ten folds, every each with over 300 samples. The folds are resting, warning, angry, defending, fighting, happy, hunting mind, mating, mother call, and paining.

3.5.4 Audio Set

Audio Set [30] helps in the development of audio event recognition systems. It is an ensemble of 632 audio folds in a hierarchy over 6 levels of manually annotated ten-second audio files, containing 1,789,621 ten-second video segments from YouTube. Its seven main folds are human sounds, animal sounds, natural sounds, music, sounds of things, source-ambiguous sounds, and channel, environment, and background.

3.5.5 Speech Commands

The Speech Commands dataset [31] has 3.8 GB of 105,829 one-second long utterances of 35 short words recording by 2,618 members of the Artificial Intelligence Yourself community and stored in WAVE format file[10]. Their 35-word categories include actions, numbers, people names, animal names and objects.

[8] https://github.com/karolpiczak/paper-2015-esc-dataset
[9] https://zenodo.org/record/4724180
[10] http://download.tensorflow.org/data/speech_commands_v0.02.tar.gz

3.5.6 FMA

The Free Music Archive or FMA[11] is a large dataset suitable for evaluating several tasks in MIR [32]. FMA consists of 917 GB and 343 days of audio from 106,574 tracks from 16,341 artists and 14,854 albums, arranged in 16 genres and 145 subgenres with a track, album and artist metadata.

3.5.7 Nsynth

NSynth[12] holds 306,043 four-second prerecorded notes of 1006 instruments, ranging over a standard MIDI piano with an average of 4.75 unique velocities per pitch [33]. Nsynth metadata is subdivided into sources according to the instrument sound, family of the note instrument, and sonic qualities of the note. The eleven label families are bass, brass, flute, guitar, keyboard, mallet, organ, reed, string, synth lead, and vocal.

3.6 Data Augmentation in Environmental Sound Classification

We chose the ESC area due to its wide application in intelligent systems and the availability of current works that allow us to exemplify the benefits and challenges of using data augmentation in audio data. Table 3.2 presents the references found in literature, the analyzed datasets, the used augmentation techniques and the tool used. All of the works presented used accuracy as a performance measure, where $accuracy = \frac{VP+VN}{N}$, with VP representing the true positives, VN the true negatives and N the number of samples.

Table 3.2: Environmental classification sounds works. *BIRD's dataset is unavailable.

Works	Dataset	Techniques	Tool
[12]	US8K	ADA, NoAug	MUDA
[4]	BIRD*, CAT	ADA, SDA, IDA, NoAug	Audiogmenter
[13]	ESC-10, ESC-50, US8K	ADA, NoAug	librosa
[14]	ESC-10, ESC-50, US8K	ADA, IDA	librosa

Salamon and Bello [12] presented a CNN architecture for ambient sound classification that includes three convolutional layers interleaved with two maxpooling operations and two dense layers. Four audio augmentation techniques were used (TS, SP, DRC, and BN), yielding to five additional training datasets. Seven experiments

[11] https://github.com/mdeff/fma
[12] https://magenta.tensorflow.org/datasets/nsynth

were carried out for each model to evaluate the effects of each approach individually, collectively, and without data augmentation. The CNN model trained using all techniques together performed the best with an accuracy of 79% while the model without data augmentation reached 73%.

To classify bird and cat sounds, Nanni et al. [4] used pre-trained convolutional networks and ensemble techniques with a combination of five CNNs called "fusion". GoogleNet [18] and VGGnet [34] were the pre-trained networks utilized. The research examined the use of audio and image augmentation methods. For both datasets, fusion produced the best results. Table 3.3 shows the average accuracy of the fusion models for each data augmentation approach.

Table 3.3: Fusion model performance for each dataset and data augmentation approach in Nanni et al. [4].

Approach	Techniques	CAT	BIRD
NoAug	-	87.36	95.81
ADA1	TS, SP, L, SN, TiS	89.22	96.16
ADA2	WR, C, SN, SU, HD	89.05	**96.56**
IDA	F, ZR, RA, S	82.71	92.89
SDA	SRS, SSM, VTLM, STS, SRM	**91.73**	94.30

Mushtaq and Su [13] used SP, TS, and SN to improve CNNs models to classify ambient sounds. Mel, MFCC, and Log-Mel were tested as spectral extraction techniques. Two CNNs with five layers were proposed, one with and one without maxpooling. The best accuracy for all datasets was obtained by combining the model without maxpooling, the Log-Mel spectrogram, and ADA.

The accuracy of the best model for each dataset, the absolute accuracy gains and the cost of using audio augmentation over time are shown in Table 3.4. The costs were analyzed by (3.2), where T_{Aug} is the training duration in seconds with data augmentation and T_{NoAug} is the time used to train the model without using data augmentation.

$$Cost = \frac{T_{Aug}}{T_{NoAug}} \tag{3.2}$$

Mushtaq et al. [14] has tested new ways to solve the problems presented in Mushtaq and Su [13]. Log-Mel was used to extract spectrograms and CNNs with seven (CNN-7) and nine layers were applied. Also, pre-trained models with millions of images using FastAi (https://docs.fast.ai/vision.learner.html) to get the weights of ResNet [35], DenseNet [36], SqueezeNet [37], AlexNet [38], and VGG [34]. The research explores the usage of image augmentation (ZR, S, BR, RA, sR and F) and audio augmentation (SP, Ts and TS) for all convolutional models studied.

Table 3.4: The best model's performance (Accuracy, Gains and Costs) for each dataset in Mushtaq and Su [13].

	ESC-10	ESC-50	Us8k
No Aug (%)	81.25	57.00	94.14
Aug (%)	94.94	89.28	95.37
Gains (pp)	13.69	**32.28**	1.23
Costs	5.31	5.45	5.94

CNN-7 has the best accuracy among the proposed CNNs, and ResNet-152 was the best of the pre-trained models. The performance of CNN-7 and ResNet-152 for the three datasets and absolute gains in accuracy between audio and image augmentation are shown in Table 3.5.

Table 3.5: Performance of the best models in Mushtaq et al. [14] for each dataset.

	ESC-10	ESC-50	US8k
TAA CNN-7 (%)	77.86	40.46	69.13
NAA CNN-7 (%)	93.5	96.1	95.05
Gains (pp)	15.64	**55.64**	25.92
TAA ResNet-152 (%)	95.23	87.49	98.29
NAA ResNet-152 (%)	**99.04**	97.30	**99.05**
Gains (pp)	3.81	**9.81**	1.21

3.6.1 Spectrograms Methods

All convolutional models presented received the sound spectrogram, which describes the variation in the intensity of the spectral components over time as an input. The techniques used were Linear Spectrogram, Mel Spectrogram, Frequency Cepstral Coefficient, Log-Mel Spectrogram and Discrete Gabor Transform. Log-Mel presented a better performance than both Mel Spectrogram and Frequency Cepstral Coefficient with gains of up to 18 pp of accuracy [13].

3.6.2 Data Augmentation Techniques Comparison

If available in a group, ADA and SDA approaches outperform IDA techniques overall [14, 4]. The ESC-50 dataset had the greatest difference regarding performance between audio and image augmentation. Both models, ResNet-152 and CNN-7 increased 9.81 pp and 55 pp, respectively [14]. Furthermore, when the individual improvements are measured, the SP approach has made the most relevant contributions, whereas the BN was in charge of the least ones [12].

3.6.3 Data Augmentation versus No Augmentation

When comparing performance with and without data augmentation, DA approaches had a greater performance in most of the found papers. The results of Mushtaq and Su [13] presented in Table 3.4, demonstrate the advantages of ADA for the three datasets, especially for ESC-50, which has more classes than ESC-10 and US8K, with an absolute increase of 32 pp. The authors state that the low gain for the dataset US8k is due to the large number of training samples in the original dataset that has enough data diversity to avoid overfitting.

Salamon and Bello [12] have trained a CNN model using all techniques together that resulted in an absolute gain of 6 pp above the model trained without data augmentation. The most favored classes were idle engine and jackhammer, whereas DRC and BN impair the air conditioning sound class predictions. For even better outcomes, the authors recommend using the conditional ADA approach.

As shown in Table 3.3, the best performances for both datasets were obtained using augmentation techniques, with the highest gain for the CAT dataset using spectrogram augmentation [4]. However, the utilization of image augmentation deteriorated the performance of the models. The authors justified that the use of techniques such as reflection, when applied to spectrograms, drastically changes the sound and its semantic value. This result suggests that the use of traditional IDA techniques is not suitable for application in spectrograms.

3.6.4 Trade Off between Cost and Performance

The offline data augmentation is a costly approach because for each transformation in an original sample, a new one is created and this reflects in training time. The use of six ADA strategies increased training time by 5 to 6 times [13]. For ESC-10 and ESC-50, this cost is compensated with gains of 13.69 pp and 32.28 pp respectively. However, for US8k dataset, the gain was less than 2 pp. Therefore, it is necessary to consider the size of the dataset studied and the accuracy required for the respective task to define whether or not to use any data augmentation technique.

3.6.5 Transfer Learning and Ensemble Methods

The use of very deep pre-trained CNN models may provide more accuracy than convolutional networks with few layers [14]. When transfer learning and audio augmentation were used, the ResNet-152 accuracy outperformed all other networks, achieving 99% for ESC-10 and US8k using ADA [14]. In addition, the use of pre-trained ensemble models obtained higher accuracy than a unique model [4].

3.7 Conclusions

Throughout this paper, several offline DA methods for audio and image were presented to improve the performance of CNNs in audio classification. In addition, we provide a list of audio DA tools and open datasets for ESC, MIR, and ASR tasks. In the ESC analyzed works, audio and spectrogram augmentation were more effective than without augmentation approach, with gains in accuracy reaching 50 pp. In addition, the Shifting Pitch technique was the one that individually provided the greatest increase regarding accuracy.

Shortcomings were found in the choice of deformations to perform data augmentation, since traditional techniques applied in computer vision tasks were used. In this way, these setbacks changed the semantic value of the audio signals when applied to the spectrograms, resulting in a worse of the models' performance. On the other hand, the effect of DA techniques may vary according to the sound type, standing out the SP as the most efficient audio augmentation technique. Furthermore, DA is especially important for small and complex datasets, with improvements of over 30 pp.

The choice of augmentation technique depends on each specific application. This paper has shown some shortcuts to be followed in the absence of computational power necessary to test each technique within a reasonable time budget. Future works will focus on including more papers, as well as different methods such as online data augmentation and deep learning applied to data augmentation.

References

[1] J.Cao, M.Cao, J.Wang, C.Yin, D.Wang, and P. Vidal. Urban noise recognition with convolutional neural network. *Multimedia Tools and Applications*, 78:29021–29041, 2019.

[2] J.P. Bello, C. Silva, O. Nov, R.L. Dubois, A. Arora, J.Salamon, C. Mydlarz, and H. Doraiswamy. Sonyc: A system for monitoring, analyzing, and mitigating urban noise pollution. *Communications of the ACM*, 62:68–77, 2019.

[3] Y. R. Pandeya and J. Lee. Domestic cat sound classification using transfer learning. *The International Journal of Fuzzy Logic and Intelligent Systems*, 18:154–160, 2018.

[4] L. Nanni, G. Maguolo, and M. Paci. Data augmentation approaches for improving animal audio classification. *Ecological Informatics*, 57:101084, 2020.

[5] R. L. Aguiar, M. G. Y.M.G. Costa, and C.N. Silla. Exploring data augmentation to improve music genre classification with convnets. *Proceedings of the International Joint Conference on Neural Networks*, pages 1–8, 2018.

[6] C. W. Wu, C. Dittmar, C. Southall, R. Vogl, G. Widmer, J. Hockman, M. Muller, and A. Lerch. A review of automatic drum transcription. *IEEE/ACM Transactions on Audio Speech and Language Processing*, 26:1457–1483, 2018.

[7] Y. Seo and J. Huh. Automatic emotion-based music classification for supporting intelligent iot applications. *Electronics*, 8:164, 2019.

[8] Roman A Solovyev, Maxim Vakhrushev, Alexander Radionov, Irina I Romanova, Aleksandr A Amerikanov, Vladimir Aliev, and Alexey A Shvets. Deep learning approaches for understanding simple speech commands. In *IEEE 40th International Conference on Electronics and Nanotechnology*, pages 688–693, 2020.

[9] Juan P Dominguez-Morales, Qian Liu, Robert James, Daniel Gutierrez-Galan, Angel Jimenez-Fernandez, Simon Davidson, and Steve Furber. Deep spiking neural network model for time-variant signals classification: a real-time speech recognition approach. In *2018 International Joint Conference on Neural Networks*, pages 1–8, 2018.

[10] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W.T. Freeman, and M. Rubinstein. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *ACM Transactions on Graphics*, 37, 2018.

[11] Sharath Adavanne, Konstantinos Drossos, Emre Çakir, and Tuomas Virtanen. Stacked convolutional and recurrent neural networks for bird audio detection. In *2017 25th European Signal Processing Conference*, pages 1729–1733, 2017. ISBN 9780992862671. doi: 10.23919/EUSIPCO.2017.8081505.

[12] J. Salamon and J.P. Bello. Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Processing Letters*, 24: 279–283, 2017.

[13] Zohaib Mushtaq and Shun Feng Su. Environmental sound classification using a regularized deep convolutional neural network with data augmentation. *Applied Acoustics*, 167:107389, 2020.

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

[15] C. Shorten and T. M. Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6, 2019.

[16] G. Maguolo, M. Paci, L. Nanni, and L. Bonan. Audiogmenter: a matlab toolbox for audio data augmentation. *Applied Computing and Informatics*, 2021.

[17] N. Jaitly and G.E. Hinton. Vocal tract length perturbation (vtlp) improves speech recognition. In *Proceedings of the 30th International Conference on Machine Learning*, volume 90, pages 42–51, 2013.

[18] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015. ISBN 9781467369640. doi: 10.1109/CVPR.2015.7298594.

[19] Brian McFee, Colin Raffel, Dawen Liang, Daniel Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th Python in Science Conference*, pages 18–24, 2015. doi: 10.25080/Majora-7b98e3ed-003.

[20] Brian McFee, Eric J Humphrey, and Juan P Bello. A software framework for musical data augmentation. In *Proceedings of the 16th International Society for Music Information Retrieval Conference*, pages 248–254, 2015. ISBN 9788460688532.

[21] Chris Bagwell. Sox - sound exchange | homepage, 2010. URL http://sox.sourceforge.net/.

[22] François Chollet. Keras: the python deep learning api, 2021. URL https://keras.io/.

[23] Iver Jordal. Audiomentations, 2021.

[24] Rachel Bittner, Eric Humphrey, and Juan Bello. pysox: Leveraging the audio signal processing power of sox in python. In *Proceedings of the International Society for Music Information Retrieval Conference Late Breaking and Demo Papers*, 2016.

[25] 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*, page 1041–1044, 2014. ISBN 9781450330633. doi: 10.1145/2647868.2655045.

[26] Frederic Font, Gerard Roma, and Xavier Serra. Freesound technical demo. In *Proceedings of the 21st ACM International Conference on Multimedia*, page 411–412, 2013. ISBN 9781450324045. doi: 10.1145/2502081.2502245.

[27] Karol J. Piczak. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM International Conference on Multimedia*, page 1015–1018, 2015. ISBN 9781450334594. doi: 10.1145/2733373.2806390.

[28] Y.R. Pandeya and J. Lee. Domestic cat sound classification using transfer learning. *The International Journal of Fuzzy Logic and Intelligent Systems*, 18:154–160, 2018.

[29] Y.R. Pandeya, D. Kim, and J. Lee. Domestic cat sound classification using learned features from deep neural nets. *Applied Sciences*, 8, 2018.

[30] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 776–780, 2017.

[31] P. Warden. Speech commands: A dataset for limited-vocabulary speech recognition. *arXiv preprint arXiv:1804.03209*, 2018.

[32] Michaël Defferrard, Kirell Benzi, Pierre Vandergheynst, and Xavier Bresson. Fma: A dataset for music analysis. In *18th International Society for Music Information Retrieval Conference*, 2017.

[33] Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Mohammad Norouzi, Douglas Eck, and Karen Simonyan. Neural audio synthesis of musical notes with wavenet autoencoders. In *International Conference on Machine Learning*, pages 1068–1077, 2017.

[34] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations*, 2015.

[35] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. ISBN 9781467388504. doi: 10.1109/CVPR.2016.90.

[36] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2261–2269, 2017. doi: 10.1109/CVPR.2017.243.

[37] F.N. Iandola, S. Han, M.W. Moskewicz, K. Ashraf, W.J. Dally, and K. Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and <0.5mb model size. *arXiv preprint arXiv:1602.07360*, 2016.

[38] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25, 2012.

Chapter 4

Forroset: A multipurpose dataset of Brazilian Forró music

Forró is an important genre that has been developing the cultural identity of Brazil and it is one of the most consumed by Brazilians on Spotify. However, the lack of datasets and their specificity leads to less research about this genre. In order to overcome this issue, we presente a set of data roughly compounded by 3000 songs named Forroset, which provides editorial information, audio features, information of rhythm, and audio files from Spotify. Furthermore, over 1400 lyrics of songs were obtained from the Vagalume platform. When Forroset is compared to other datasets, it was seen that our dataset is more powerful regarding the diversity of information heading to comprehensive problems resolution.

4.1 Introduction

Forró is an important Brazilian musical genre that is popular throughout all socioeconomic layers and has its traditional matrix recognized as the Intangible Cultural Heritage of Brazil [1]. With its early origins in a party, it has contributed to the construction of northeastern and national identity for more than a century. Furthermore, forró is becoming one of the most played genres on Spotify in Brazil [2], the world's largest streaming music platform.

Forró could be classified in three musical genres: "Pé-de-Serra", "Universitário", and "Eletrônico" [3]. Except for "Forró Eletrônico", the core instrumental structures are the same: the zabumba, triangle, and accordion compose this basic structure, while the singer's voice completes the musicality and rhythm.

Forró has been explored in several fields, including applications to aid individuals in dancing [4, 5, 6, 7, 8], genre recognition [9, 10], and the evaluation of paradigm-breaking in a musical context [11]. The following gaps detected in the forró datasets identified that limit research findings in this genre are: *(i)* An unavailability or discontinuity of data access; *(ii)* failure to follow FAIR Data principles (Findable, Accessible, Interoperable, and Reusable) in dataset construction and sharing; and *(iii)* the dataset

Table 4.1: Forró related datasets. FAIR issues identified: F1 (not persistent identifier), A1 (non low-level protocol), I2 (no documentation), I3 (non qualified cross-reference), R1.1(unknown licence), R1.2. (undetailed provenance).

Name	Tracks	Features	Genres	FAIR
LMD	313	artist, title, genre, MP3 file	10	not available
BSL	1000	artist, title, genre, lyrics	9	F1, I2, I3, R1.1, R1.2
BLD	11862	artist, title, genre, lyrics	14	F1, A1, R1.1
FVD	27352	artist, title, year, mfcc, similarity network	1	A1, I3

information is insufficient to be used in different applications.

To handle these limitations, Forroset is introduced, a dataset that contains extensive information on over 2900 forró music. The present dataset offers information ranging from the song's authors to technical details such as the beat position and length. Forroset also includes over 1400 song lyrics and gives MP3 files for all dataset songs, allowing users to develop new applications with forró.

4.2 Related forró datasets

Four datasets containing forró music were identified. Their descriptions, as well as their FAIR's shortcomings, are presented in Table 4.1. The LMD - Latin Music Database [12] is the precursor and has prompted several musical works of genre classification that includes forró music. This dataset, despite being very important to investigating different Latin American genres, is no longer available. The BSL - Brazilian Songs Lyrics [13] on Kaggle and the BLD - "Brazilian Lyrics Dataset" [9], are two datasets that include extensive lyrics content collected from Vagalume. Lastly, the FVD - Forró em Vinil Dataset [14] has a large number of songs that include spectral and other metadata. This dataset lacks audio files, which may be obtained manually from the "Forró em Vinil" website[1].

There is a large volume of songs in the available datasets, but provide few features and application opportunities. BSL and BLD, for example, are aimed at classifying music genres from the lyrics of the songs. Furthermore, no available dataset actually provides audio files of the songs, as the FVD songs need to be searched on the website. It is worth mentioning that the latter does not have songs after the year 2000, therefore, it does not contain songs related to the rebirth of forró with the emergence of Forró Universitário. Finally, all datasets have FAIR's shortcomings.

[1] https://www.forroemvinil.com/

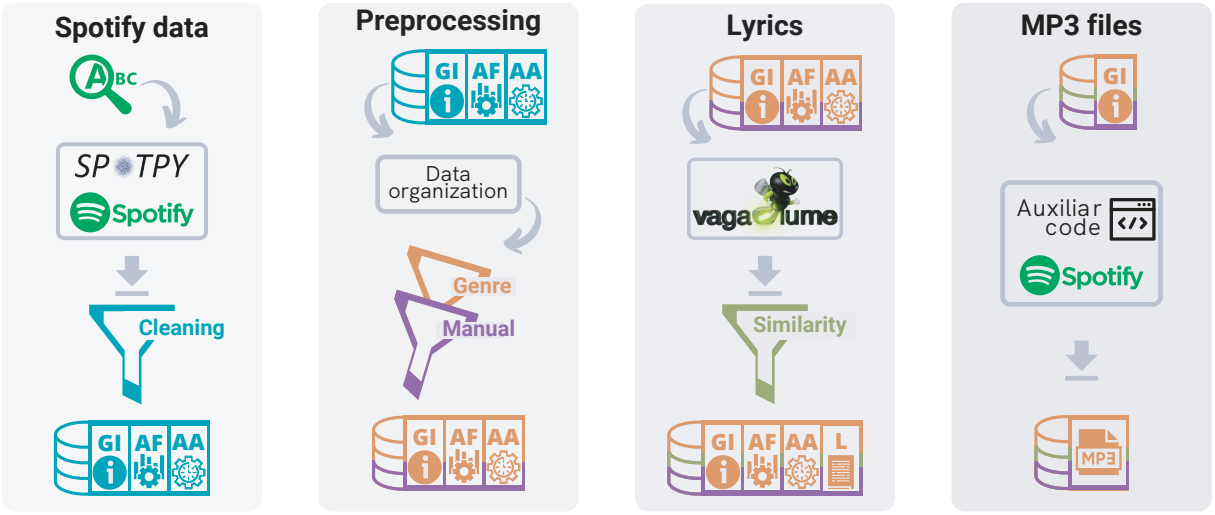


Figure 4.1: Main steps for obtaining Spotify and Vagalume data.

4.3 Forroset creation

Forroset contains six kinds of information: General Information (GI), Audio Features (AF), Audio Analysis (AA), Filters and Organization (FO), Lyrics (L) and MP3 files. File sets as GI, AF, and MP3 were collected via Spotify API[2] and Spotipy package[3], the AA through Spotify API, Spotipy and Librosa package[4], the Lyrics using Vagalume API[5] and package[6]. Furthermore, we obtained FO by Spotify data transformations and manual annotation. Figure 4.1 depicts a concise representation of the Forroset development.

4.3.1 Spotify data

The Spotify tabular data collection was carried out in Spotify's database. To search for songs, we chose 5 keywords that refer to Forró Pé-de-Serra e o Forró Universitário: "Forró Pé de Serra", "Forró Universitário", "Forró Tradicional", "Xote", and "Baião"; in addition to the names of 20 forró artists/groups. In the next step, songs with the same identifier, identical names, explicit content, non-playable, and without popularity and tempo values were completely removed from the dataset.

GI was collected through a keyword search, while AF and part of AA from the Spotify database via song identifiers were obtained separately. All Spotify tabular data is given for the entire song, although the audio files supplied only contain 30 seconds of a track which it is unknown if it is the beginning, middle or end of the

[2] https://developer.spotify.com/documentation/web-api/
[3] https://spotipy.readthedocs.io/en/2.19.0/
[4] https://librosa.org/doc/latest/index.html
[5] https://api.vagalume.com.br/
[6] https://github.com/diegoteixeir4/python-vagalume

song. This aspect prevents the application of the dataset for beat tracking since it is not possible to match the beat annotations of the entire track with the 30-second sample.

To overcome this setback, the librosa package for audio file beat recognition was used. In order to enable the algorithm's convergence, it was provided to the algorithm an initial stance of the song's tempo collected by Spotify to estimate the beats. Finally, a discrepancy score consisting of the error between the average distance of librosa and Spotify annotated beats was created.

4.3.2 Preprocessing data

To facilitate the use of the dataset in machine learning tasks, the data was divided into 20 randomly separated folds while maintaining a uniform distribution in terms of tempo and popularity. The tempo of the songs were divided into 10 bins of width of 11, except for the first and last bins, which were larger due to the lack of songs in the extremes.

We observed that Spotify only provides the genres by the artists/bands for each song and existing genres out of the scope ascribed to Forró authors. For example, even though Falamansa is one of the Forró Universitário's precursor bands [3], the Spotify genres ascribed to Falamansa artist include "axe", "brazilian reggae", "pagode", "sertanejo" e "sertanejo universitario".

To filter out songs that do not fit the scope, a metric called "genre filt¨ was developed, which calculates the ratio of the artist's genres that fit the scope on a scale from 0 to 1. The songs having a score of less than 0.3 on this scale were automatically excluded. Additionally, the songs from the 20 forró bands used for keyword selection stage, weren't filtered and a score of 1.1 was assigned to them.

Another point that was addressed is the presence of tracks with more than one song, which can be a problem, as it is not possible to identify from which part of the track the 30s is. Therefore, we manually evaluated the 100 most popular songs (according to Spotify popularity score), from each tempo bin. A binary score called "manual filt" was established, and the tracks that are in the scope received a score of 1. On the other hand, the tracks that weren´t recognized as Forró Pé-de-Serra/Universitário, or with more than one song per track, were given a score of zero.

4.3.3 Vagalume Lyrics

Vagalume is a Brazilian website for music lyrics. The search for the songs lyrics was performed using the information of the song and artist/group name provided by

Spotify. In cases when the search was successful, we picked up the track artist, track title, lyric and its Vagalume´s access link.

To ensure that the expected song and the received lyrics are consistent, the similarity was calculated using the python package difflib[7]. When the track title and artist similarities were less than 90%, the lyrics were manually evaluated, removing all confirmed divergent lyrics.

4.3.4 Getting MP3 files

Forroset contains a python helper code that uses the preview URL present in the GI to automatically get 30 second samples for all Forroset songs. Samples are downloaded in MP3 format at 22.05 kHz to a specified directory.

4.3.5 Ethics of Data Collection

Regarding the ambiguity found in the Terms of Services (TOS) of Social Media [15], as required by the Vagalume API TOS, the link to the song's lyrics on the Vagalume website is provided. Furthermore, according to the section *IV.3.a.i* of the Spotify API TOS[8], the GI, AF and AA data, necessary to operate Forroset, are compiled.

Furthermore, it is argued in this paper that Forroset was constructed in an ethical manner [16], considering that the data collected is publicly accessible on both sites and it is worth mentioning that the data does not contain sensitive personal information of any kind. Moreover, with this paper and the Completed Transparency Report, we are providing the detailed description of the aims, construction details, limitations and applications of Forroset.

4.3.6 FAIR principles Implementation

An attempt to ensure that Forroset adheres to all FAIR principles[9] was done, making it available on GitHub and Zenodo platforms, satisfying the Findable and Accessible requirements. To join to the Interoperable and Reusable principles, the cross-reference identifiers for data retrieved from Spotify and Vagalume, as well as documentation related to the dataset's usage are provided. Finally, the Reusable criteria was fulfilled by assigning the dataset the Creative Commons Attribution License (CC BY 4.0).

[7] https://docs.python.org/3/library/difflib.html
[8] https://developer.spotify.com/terms/
[9] https://www.go-fair.org/fair-principles/

Table 4.2: Forroset tabular groups information.

Infos	Features	Tracks
GI	Track, track id, artist, artist id, popularity, album, album id, uri, track year, duration and preview url	2977
AF	Energy, liveness, tempo, valence, acousticness, instrumentalness, key, time signature, danceability, loudness, speechiness and mode	2977
AA	Beats start, duration and confidence; bars start, duration and confidence; tatums start, duration and confidence. Librosa beats start and discrepancy	2977 2976
FO	Tempo bins, tempo bins max, genre filt, folds. Manual filt	2977 1000
L	Lyrics	1415

4.4 Data details

Forroset comprises a tabular file with 40 columns and a python code that downloads the Spotify previews in MP3 format from the URLs provided in the "preview url" column of the dataset. All information in GI, FO, AF, AA, and L groups are present in Table 4.2.

4.4.1 General Information

The search performed on 2021/04/11 returned 9043 tracks. Following data cleaning, 5373 tracks from 1032 different artists were found. After removing artist with a high propensity for having songs that were out of scope, Forroset had 2977 tracks from 82 distinct artists. The Forroset artists are shown in Fig 4.2, where the word cloud indicates all of them and the bar graph depicts the top 10 artists with the most tracks.

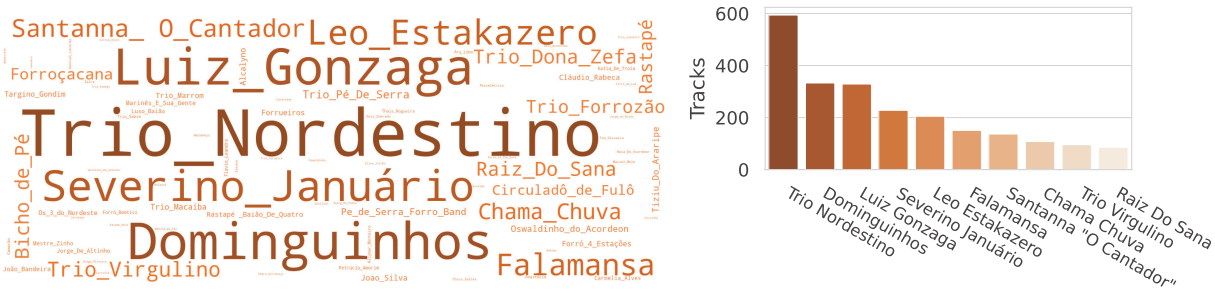


Figure 4.2: Artists ranked according to their number of tracks on Forroset.

Figure 4.3 shows the total songs by popularity and year. In addition, the number of songs per album is displayed in ascending order. Histograms are shown for all (in orange), songs with lyrics (in green), and songs that have been manually evaluated (in purple). Forroset contains 502 albums, of which 260 were manually analyzed, and the songs with lyrics are from 284 different albums.

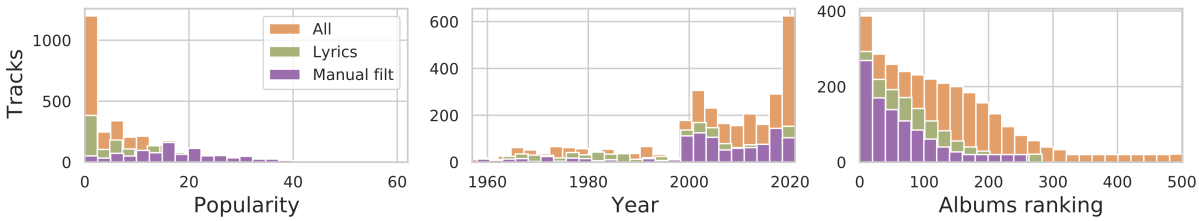


Figure 4.3: Histograms with popularity, albums, and year for each subset.

4.4.2 Spotify features

For each song, the dataset includes the feature information provided by the Spotify API. The number of songs per feature bin is presented in Figure 4.4. Histograms are shown for all AF information group and subsets.

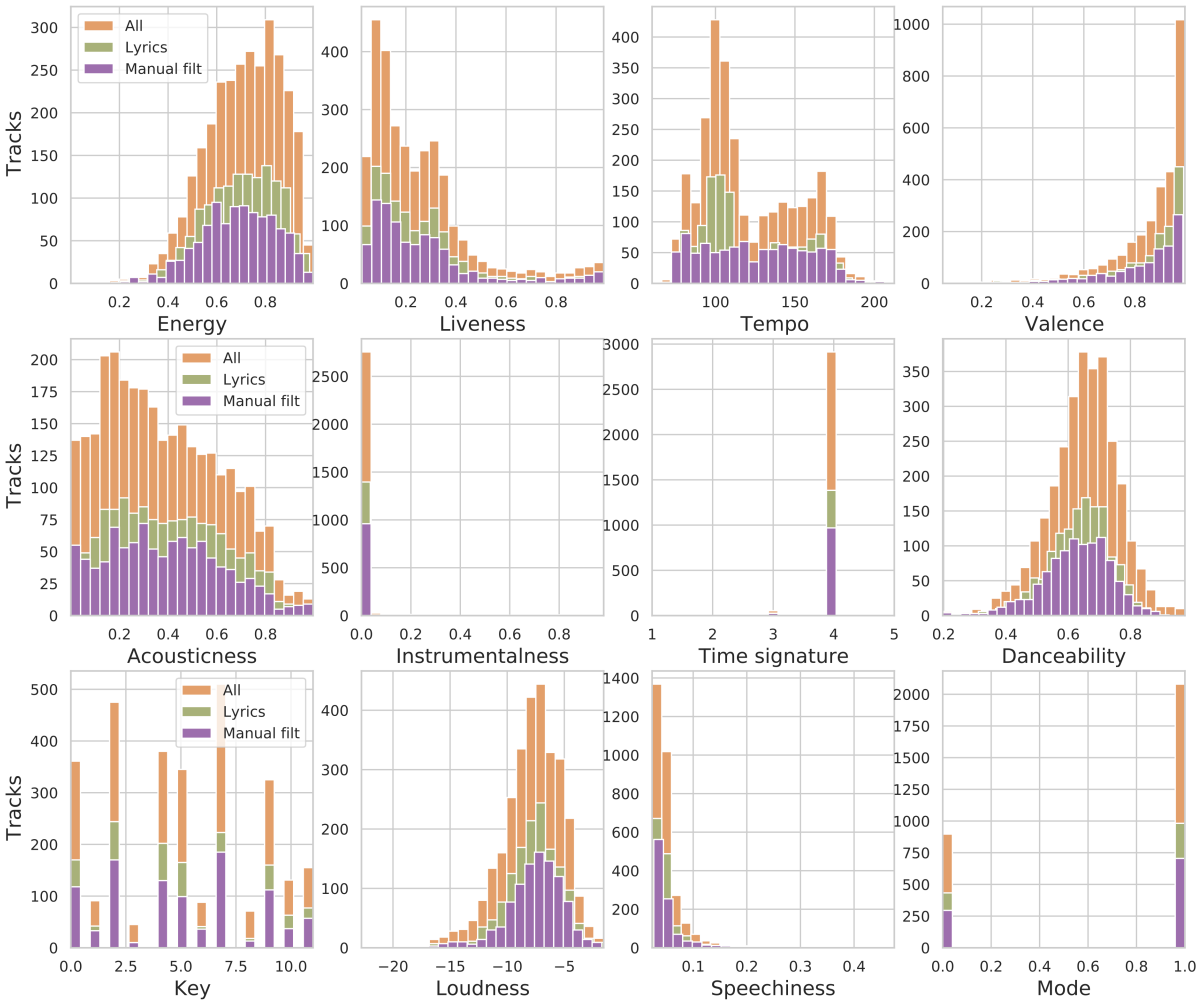


Figure 4.4: Histograms of Forroset Audio Features for all subsets.

The audio features collected via the Spotify API are briefly described in the following lines. More information is available in the API documentation.

- **Energy:** A perceptual measure of intensity and activity.

- **Liveness:** Higher liveness numbers indicate a greater likelihood that the track was performed live.
- **Tempo:** The overall estimated tempo of a track in beats per minute (BPM).
- **Valence:** The musical optimism given by a song.
- **Acousticness:** The degree to which the music is acoustic. This feature predicts whether or not a recording has no voices.
- **Time signature:** A notation standard that specifies the number of beats in each bar.
- **Danceability:** It describes a track's suitability for dancing.
- **Key:** The key of the track is given by matching the integers to pitches using standard Pitch Class notation.
- **Loudness:** The overall volume of a track measured in decibels (dB).
- **Speechiness:** Detects the presence of spoken words in a track.
- **Mode:** A track's mode denotes the type of scale from which its melodic content is formed.

4.4.3 Spotify audio analysis

Forroset includes the beginnings and durations of each bar, beat, and tatum, over the entire song, for all songs. Each metric is assigned to a confidence level ranging from 0 to 1, with 1 being the highest level of confidence. In addition, the beat starts for the 30s audio samples estimated by Librosa and a discrepancy score is provided. Figure 4.5 depicts the occurrences of the three events during a 15-second extract of Falamansa's song Xote dos Milagres, the most popular track in the dataset.

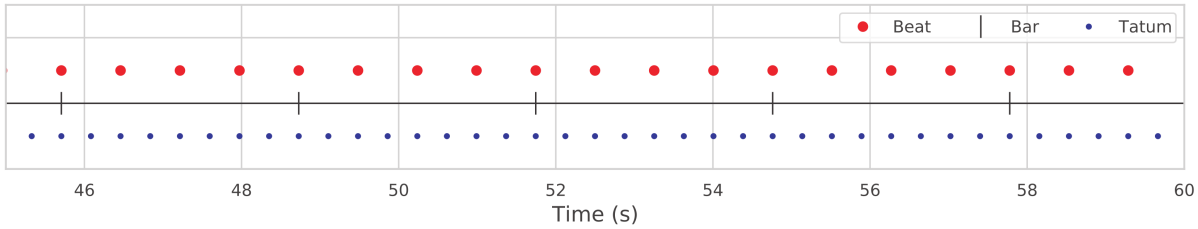


Figure 4.5: Bar, beat and tatum of Falamansa's song Xote dos Milagres of Forroset.

The following is a brief description of the described rhythmic structures:

- **Bar:** A time signature is a method of quantitatively organizing the sounds of a musical composition into beats and pauses. A bar in Forró is often made up of four beats.
- **Beat:** A beat is the fundamental time unit of music. Typically, beats are multiples of tatums.
- **Tatum**: A tatum is the lowest regular pulse train that a listener infers instinctively from the time of observed musical events.

4.4.4 Filters and organization

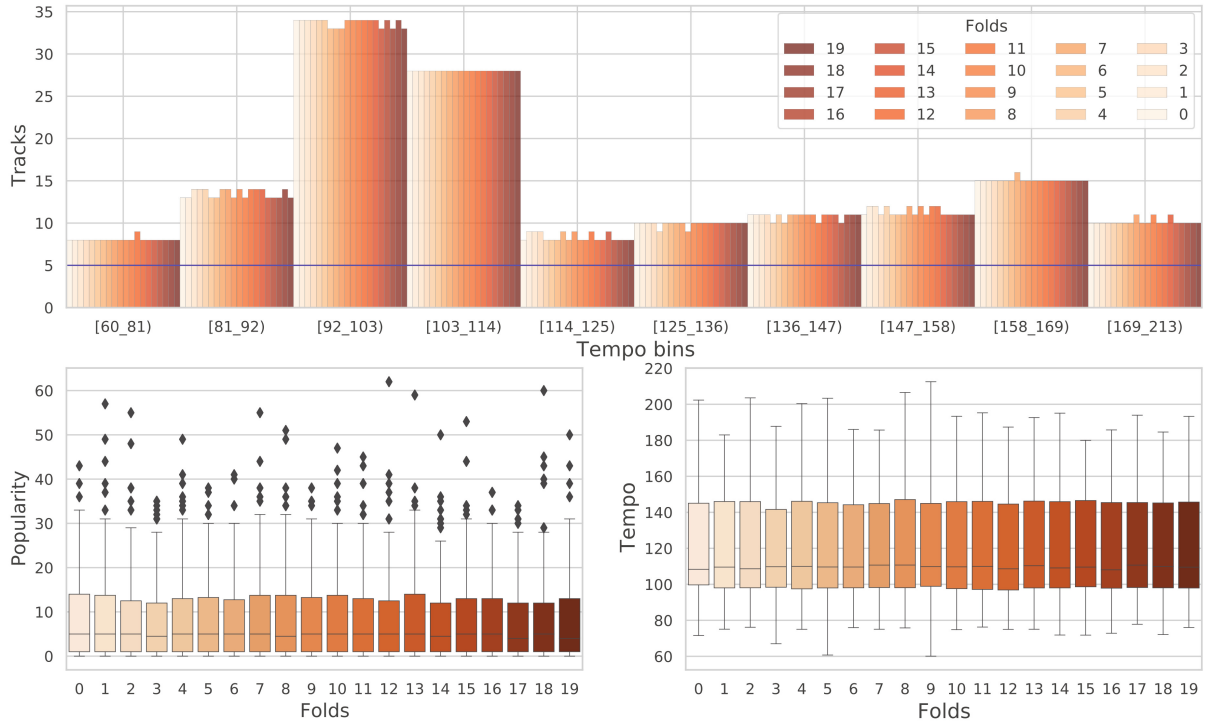


Figure 4.6: Separating and balancing folds over time and popularity. The blue line on the top histogram represents the 50 songs for fold that have been manually reviewed.

Regarding the 1000 tracks manually analyzed, 848 are from the Forró Universitário and Forró Pé-de-Serra scope. The histogram at the top of Figure 4.6 shows how songs are divided into 20 balanced folds based on tempo. At the bottom, the box-plots compare the popularity and tempo of each fold, allowing the balance of the folds to be seen concerning these characteristics.

4.4.5 Lyrics

The search for title and artist in Vagalume's API, performed on 2021/10/12, yielded lyrics for 1435 tracks in which 20 are manually removed. In a unique list format, the

title, artist, lyric, and url were created. Figures 4.3 and 4.4 highlight the information for songs that have lyrics.

4.4.6 MP3 files

Using our auxiliary code, the 2959-audio-files were automatically downloaded. It took roughly six hours to complete the access to all the audios in the dataset. This extra time is due to the use of delays between queries to avoid overloading Spotify's platform.

4.5 Forroset's Potential Applications

Forroset was created with the goal of supporting Forró research. Information from multiple sources and kinds were organized around a unique identifier per song to allow its usage in several tasks, as well as the use of over one type of information to explore the same task. The variety of information allows the contribution to areas that have already been explored for forró such as Music Information Retrieval and Dance Teaching, in addition, it can motivate the beginning of research in unprecedented areas such as the Music Industry.

4.5.1 Forró Industry

Identifying a future hit song is a task of great interest to the music industry. However, the success of a song can be related to several factors, making this a task widely studied [17, 18, 19]. In Forroset, the Spotify popularity score is provided and can be estimated using the information contained in GI, AF and AA, the lyrics and the audio files. This information could be used together or separately, allowing, besides predicting the popularity, the selection of the best information. In this way, music recommendation is a another complex task that Forroset can be used for. Because most of the data come from Spotify, Forroset can be highly valuable for Spotify-integrated apps such as the one shown in Álvarez et al. (2020) [20].

4.5.2 Dance teaching

The use of computer models to help people dance forró has been recently explored. One of the approaches consists of building applications that help teachers to assess how their students are dancing [6, 7, 8]. In this case, Forrosset contributes by providing the rhythmic information present in AA and AF. A facilitator is that researchers will be able to play the same versions of songs present in Forroset through Spotify.

Another recent initiative is the use of forró rhythm prediction models to pass this rhythm through tactile stimuli for deaf people [4, 5]. For this application, Forroset can contribute with the audio files and the respective rhythmic notes of tempo, bar and beat. Furthermore, it allows the application of deep learning models, previously impossible due to the absence of annotated datasets [4]. For both applications presented, the manual filter can be especially important, since they need songs with a well-defined rhythmic structure, as in the case of Forró Pé-de-Serra and Universitário.

4.5.3 Music Information Retrieval

Forroset is useful in beat tracking as it provides audio files and annotated beats. In this case, features can be extracted from the audio data to adjust neural network models [21]. This approach can be similarly used for tempo estimation [22]. Forroset has useful information for classifying musical genres such as the audio files [23], the lyrics of the songs [9] and the Spotify audio features [24]. It can be added to other datasets for the same purpose.

4.6 Conclusions

This paper described Forroset, a Forró dataset designed to promote studies specially on the musical information recognition, dance teaching and music industry from this valuable Brazilian genre, particularly the subgenres Forró Universitário and Forró Pé-de-Serra. The dataset includes multiple audio information received from Spotify, song lyrics obtained via Vagalume, and information relating to manual and automatic filters. When compared to the other Forró datasets found, Forroset has fewer tracks. However, it contains the most diverse information, allowing for more comprehensive problem-solving.

The main limitations are the small number of tracks, the non-manual rating of all of them, and the approximate measures in the beat annotation. In future works, it will be useful to manually analyze the remaining songs, and expand the dataset to other Forró classes, such as adding identifiers for each subgenre.

4.7 Availability

Forroset data and code can be accessed at https://github.com/lucas-fpaiva/Forroset.

References

[1] IPHAN. Matrizes Tradicionais do Forró recebem título de Patrimônio Cultural do Brasil, 2021. URL https://bit.ly/iphan-forro.

[2] Maria Luiza Botelho Mondelli, Luiz M. R. Gadelha Jr., and Artur Ziviani. O que os países escutam: Analisando a rede de gêneros musicais ao redor do mundo. In *Anais do VII Brazilian Workshop on Social Network Analysis and Mining*, 2018.

[3] Antonio Carlos Quadros Junior and Catia Mary Volp. Forró universitário: a tradução do forró nordestino no sudeste brasileiro. *Motriz. Journal of Physical Education. UNESP*, pages 117–120, 2005.

[4] Lucas Ferreira-Paiva, Hugo Gonçalves Lopes, Elizabeth Regina Alfaro-Espinoza, Leonardo Bonato Félix, and Rodolpho Vilela Alves Neves. Towards a device for helping deaf people to dance: estimation of forro bar length using artificial neural network. *IEEE Latin America Transactions*, 20(6):970–976, 2022.

[5] Lucas F Paiva, Hugo G Lopes, Leonardo B Felix, and Rodolpho VA Neves. Estimação do compasso musical do forró utilizando rede perceptron multicamadas. In *Anais do Congresso Brasileiro de Automática*, volume 2, 2020. doi: 10.48011/asba.v2i1.1331.

[6] Augusto Dias Pereira dos Santos, Lian Loke, Kalina Yacef, and Roberto Martinez-Maldonado. Enriching teachers' assessments of rhythmic forró dance skills by modelling motion sensor data. *International Journal of Human-Computer Studies*, 161:102776, 2022.

[7] Augusto Dias Pereira dos Santos, Lie Ming Tang, Lian Loke, and Roberto Martinez-Maldonado. You are off the beat! is accelerometer data enough for measuring dance rhythm? In *Proceedings of the 5th International Conference on Movement and Computing*, 2018.

[8] Augusto Dias Pereira dos Santos, Kalina Yacef, and Roberto Martinez-Maldonado. Let's dance: How to build a user model for dance students using wearable technology. In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization*, page 183–191, 2017.

[9] Raul de Araújo Lima, Rômulo César Costa de Sousa, Hélio Lopes, and Simone Diniz Junqueira Barbosa. Brazilian lyrics-based music genre classification using a blstm network. In *International Conference on Artificial Intelligence and Soft Computing*, volume 12415, pages 525–534, 2020.

[10] Tlacael Miguel Esparza, Juan Pablo Bello, and Eric J Humphrey. From genre classification to rhythm similarity: Computational and musicological insights. *Journal of New Music Research*, 44(1):39–57, 2015.

[11] Felipe Falcão Nazareno Andrade, Flávio Figueiredo, Diego Silva, and Fabio Morais. Measuring disruption in song similarity networks. In *Proc. International Society for Music Information Retrieval*, 2020.

[12] Carlos Nascimento Silla Jr, Alessandro L Koerich, and Celso AA Kaestner. The latin music database. In *Proc. International Society for Music Information Retrieval*, pages 451–456, 2008.

[13] Jorge Luiz Figueira. Brazilian songs lyrics, 2018. URL https://bit.ly/kaggle-BSL.

[14] Felipe Vieira Falcão. Dataset forró em vinil, 2021.

[15] Casey Fiesler, Nathan Beard, and Brian C Keegan. No robots, spiders, or scrapers: Legal and ethical regulation of data collection methods in social media terms of service. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 187–196, 2020.

[16] Luke Gelinas, Robin Pierce, Sabune Winkler, I Glenn Cohen, Holly Fernandez Lynch, and Barbara E Bierer. Using social media as a research recruitment tool: ethical issues and recommendations. *The American Journal of Bioethics*, 17(3):3–14, 2017.

[17] Yutong Ge, Jiaqian Wu, and Yutong Sun. Popularity prediction of music based on factor extraction and model blending. In *2020 2nd International Conference on Economic Management and Model Engineering*, pages 1062–1065, 2020.

[18] Junghyuk Lee and Jong-Seok Lee. Music popularity: Metrics, characteristics, and audio-based prediction. *IEEE Transactions on Multimedia*, 20(11):3173–3182, 2018.

[19] Eva Zangerle, Michael Vötter, Ramona Huber, and Yi-Hsuan Yang. Hit song prediction: Leveraging low-and high-level audio features. In *Proc. International Society for Music Information Retrieval*, pages 319–326, 2019.

[20] P. Álvarez, F. J. Zarazaga-Soria, and S. Baldassarri. Mobile music recommendations for runners based on location and emotions: The dj-running system. *Pervasive and Mobile Computing*, 67, 2020.

[21] EP MatthewDavies and Sebastian Böck. Temporal convolutional networks for musical audio beat tracking. In *2019 27th European Signal Processing Conference*, pages 1–5, 2019.

[22] Hendrik Schreiber and Meinard Müller. A single-step approach to musical tempo estimation using a convolutional neural network. In *Proc. International Society for Music Information Retrieval*, pages 98–105, 2018.

[23] Caifeng Liu, Lin Feng, Guochao Liu, Huibing Wang, and Shenglan Liu. Bottom-up broadcast neural network for music genre classification. *Multimedia Tools and Applications*, 80(5):7313–7331, 2021.

[24] Kehan Luo. Machine learning approach for genre prediction on spotify top ranking songs. Master's thesis, University of North Carolina, 2018.

Capítulo 5

Automatic forró rhythm estimation from home videos

Os vários benefícios da dança e a importância cultural do forró no Brasil tem motivado a realização de estudos visando auxiliar pessoas a dançarem forró. Estes estudos se dividem em abordagens para auxiliar professores a ensinarem seu alunos e iniciativas visando passar o ritmo da música por vibração para surdos. Em ambos, os bancos de dados de forró e modelos de estimação de ritmo são essenciais. Visando auxiliar estas pesquisas e o aparecimento de novas, este trabalho apresenta: *(i)* o ForrosetV, um banco de dados de vídeos de pessoas dançando músicas de forró do *dataset* Forroset; *(ii)* um modelo de visão computacional capaz de estimar a duração do passo base dos vídeos de teste do ForrosetV; e *(iii)* o Forroset+, uma nova versão do Forroset onde são adicionados músicas com ruído doméstico e a anotação da duração do passo base.

5.1 Introdução

A dança pode ser entendida como uma conexão mente-corpo integrada, na qual o dançarino precisa alcançar e manter a sincronia durante o fluxo de movimento do corpo [1]. A dança tem uma forte capacidade de expressão, sendo um meio importante de inclusão social [2]. No contexto brasileiro, o estilo de dança forró é popular em todas as camadas socioeconômicas e tem sua matriz tradicional reconhecida como Patrimônio Cultural Imaterial do Brasil [3], contribuindo para a construção da identidade nordestina e nacional há mais de um século.

O passo base do forró é o Passo Base Frente e Trás (PBFT) que pode ser descrito em oito posições (P0-P7), conforme ilustrado na Figura 5.1. O PBFT é realizado ao longo de dois compassos sendo repetidos indefinidamente ao longo de uma dança [4, 5]. Desta forma, a duração do compasso da música, multiplicado por dois, fornece o tempo em que um passo base deve ser realizado.

O uso de modelos computacionais para auxiliar pessoas a dançarem forró tem sido recentemente explorado, especialmente as abordagens baseadas na estimação da duração do passo base (DPB) [6, 7] e na construção de aplicações que auxiliam

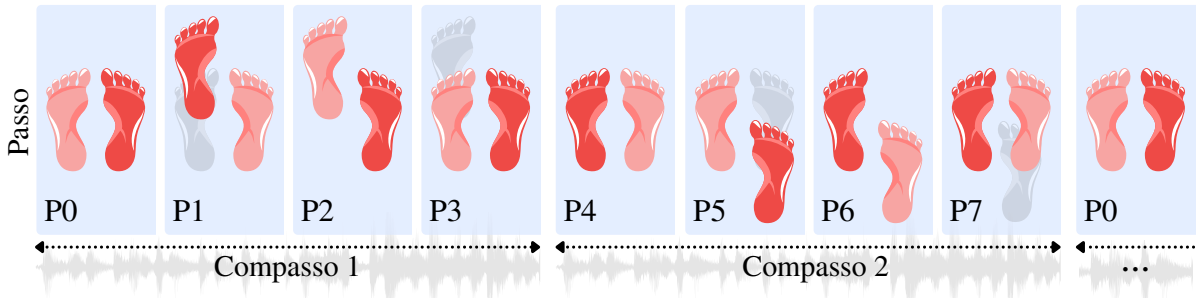


Figura 5.1: Para cada posição, o pé em vermelho escuro representa a posição do pé de sustentação, que está recebendo o peso do corpo, enquanto que o pé em vermelho claro representa a posição do pé que está apenas tocando o chão.

professores a avaliarem o quão bem seus alunos estão dançando [8, 9, 10]. No entanto, encontram-se dificuldades associadas a estas abordagens, como a complexidade da estrutura exigida para a coleta dos dados, o que reflete no pequeno número de *datasets* existentes que podem ser insuficientes para representar o escopo de músicas de forró.

Para contornar esse problema, é apresentado o ForrosetV, constituído por vídeos domésticos de pessoas dançando músicas do Forroset. Além disso, foi implementado um algoritmo baseado em visão computacional capaz de estimar a duração do passo base desses vídeos. Por fim, os áudios dos vídeos do ForrosetV foram extraídos para a criação de um conjunto de músicas com ruído real e o modelo foi utilizado para a anotação da duração do passo base das músicas do Forroset dançadas. Com a adição das músicas com ruídos reais e as anotações da DPB ao Forroset, foi criado o Forroset+.

5.2 Sistemas para auxiliar o ensino da dança do forró

Para o forró, já existem trabalhos que contém tecnologias assistivas de ritmo como Santos et al. [8], que apresenta um protótipo de aplicativo para dispositivos móveis. O mesmo é focado no aprendizado da dança, retornando ao usuário informações do ritmo dançado para que o dançarino possa sincronizar seus passos com a dança.

Entretanto, Santos et al. [9] revela os potenciais e limitações de usar apenas um acelerômetro (como um dispositivo móvel) para avaliação da qualidade da dança. Mesmo que este dispositivo detecte se o dançarino está no ritmo ou não, a avaliação de um profissional é mais capaz de identificar o motivo da falta de ritmo do aluno.

Em vista disso, dos Santos et al. [11] traz uma alternativa de avaliar a dança através da anotação manual de vídeos, onde é possível registrar maiores informações de movimento que reduzem o tempo de análise da dança pelos instrutores de forró ao fazer um histórico da evolução do aprendiz, complementando o ensino da dança.

Para obter uma abordagem automática, Santos et al. [10] complementa o trabalho

anterior [9] ao extrair características adicionais do acelerômetro para detectar além do tempo de realização do passo, a pausa, a transferência de peso e a largura do passo, mostrando as melhores características para os melhores modelos de classificação dessas habilidades na dança para auxiliar alunos e professores no aprendizado do forró.

5.3 ForrosetV

Para a anotação da DPB foi criado o *dataset* ForrosetV, que consiste em vídeos de pessoas com experiência em forró dançando músicas do Forroset. A criação do ForrosetV pode ser divida em três etapas: a criação das *playlists* de forró a serem dançadas; o recrutamento dos participantes; e a coleta dos vídeos.

5.3.1 *Playlists* de forró

As músicas utilizadas neste trabalho são pertencentes ao *dataset* Forroset [12]. O Forroset fornece informações editoriais, características de áudio, ritmo, amostras de áudio e link da respectiva música no Spotify. Além disso, 848 músicas tiveram sua aderência ao forró testada manualmente. Essa parcela do Forroset foi utilizada para criar as *playlists* que foram dançadas pelos participantes. As 848 músicas foram divididas em 10 faixas de velocidade, que variaram de 60 a 213 BPM, além disso as músicas de cada faixa foram ranqueadas pela popularidade no Spotify.

Para a criação das *playlists*, as duas músicas mais populares de cada faixa de BPM foram colocadas em todas as *playlists*, enquanto que as demais foram divididas para cada *playlist* até que todas as *playlists* ficassem com 60 músicas. Foram construídas 19 playlists, contendo 20 músicas repetidas em todas as *playlists* e 40 distintas.

A divisão das *playlists* foi realizada de forma a manter homogeneidade destas em relação ao BPM e a popularidade, conforme apresentado na Figura 5.2. As *playlists* criadas foram disponibilizadas no Spotify usando sua API e a biblioteca spotipy [13]. Como o Forroset fornece o identificador das músicas no Spotify, foi possível garantir que a versão das músicas dançadas pelos participantes é a mesma presente no Forroset.

5.3.2 Participantes

Por se tratar de um experimento com seres humanos, foi necessário a solicitação de permissão ao Comitê de Ética em Pesquisa com Seres Humanos da Universidade Federal de Viçosa. O projeto foi aprovado e pode ser encontrado sob título “Estimação de compasso musical para auxiliar surdos no aprendizado da dança do forró” e iden-

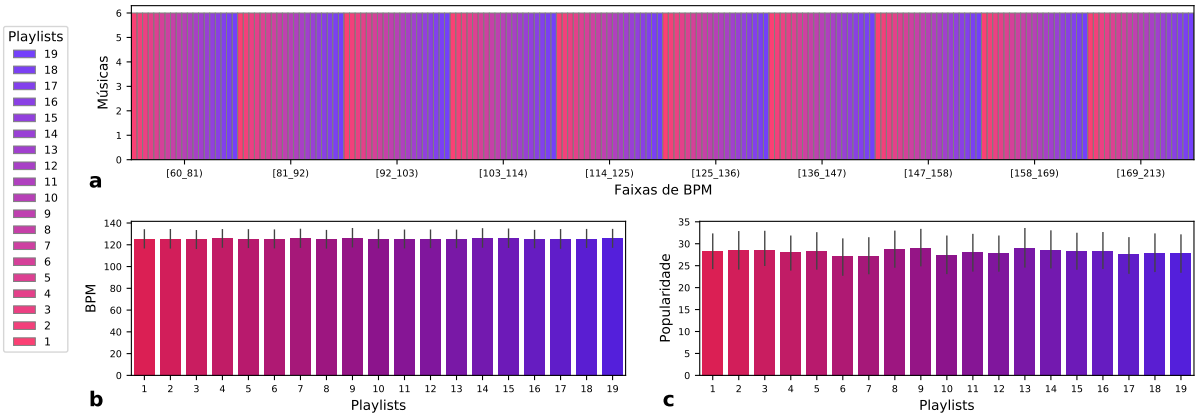


Figura 5.2: BPM e popularidade das *playlists* utilizadas no trabalho. a) As *playlists* foram construídas contendo seis músicas por faixa de BPM. b) Média e desvio padrão do BPM das *playlists* por faixa. c) Média e desvio padrão da popularidade das músicas por *playlist*.

tificação CAAE: 49804221.5.0000.5153 na Plataforma Brasil[1]. Para o recrutamento dos participantes foi realizada uma chamada para a comunidade do Forró de varias partes do Brasil, via Facebook e WhatsApp.

A pesquisa contou com 9 participantes, sete mulheres e dois homens. Suas idades variaram de 23 a 40 anos, com média e desvio padrão de 28,67±5,63 anos. Todos os participantes tinham algum nível de experiência com a dança do forró, que variou de 2 a 10 anos, com média e desvio padrão de 6,11±2,57 anos. Além disso, todos os participantes relataram ouvir músicas de forró pelo menos uma vez por mês e somente três relataram não estarem dançando com nenhuma frequência próximo ao momento da coleta.

5.3.3 Coleta dos vídeos

Cada participante recebeu o link de sua *playlist* de música para ser acessada pelo Spotify. Os participantes foram orientados a repetirem 20 passos base (Figura 5.1) para cada música de sua *playlist*, buscando realizar cada passo sempre no mesmo lugar. Além disso, foi solicitado que gravassem os vídeos do joelho para baixo, com roupas e sapatos contrastando com um fundo estático. O mais frequente foi o contraste entre as calças e sapatos escuros com o fundo claro.

As gravações foram realizadas em suas próprias casas, através de celular pessoal. Cada vídeo recebeu uma identificação que é composta por um identificador do participante, que preserva seu anonimato; e um identificador da música, disponibilizado pelo Spotify (Exemplo: 5yiB5J6Aw4ektO8oqI9Y7b_ P1). Na Figura 5.3, são apresentados alguns quadros de três vídeos de uma mesma música para três participantes.

[1] https://plataformabrasil.saude.gov.br/login.jsf

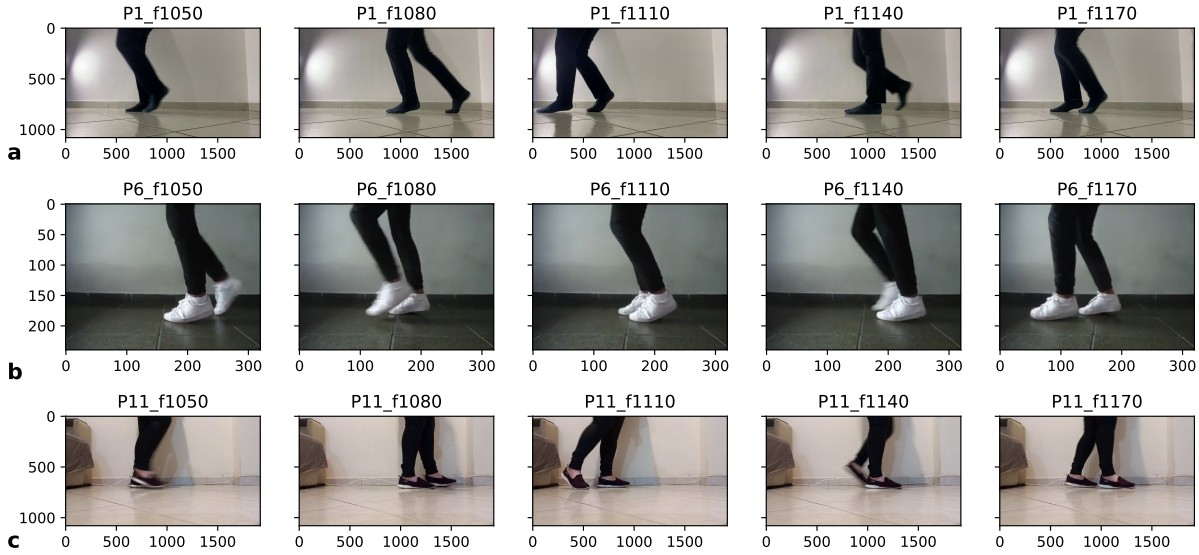


Figura 5.3: Exemplos de quadros de vídeo do banco de dados com participantes dançando a música "Avisa" da Banda Falamansa. a) participante P1. b) participante P6. c) participante P11

Os vídeos equivalentes às 20 músicas que foram dançadas por todos os participantes foram anotados manualmente com o objetivo de identificar eventuais discrepâncias entre os participantes, além de avaliar o desempenho do modelo desenvolvido para a anotação automática. A anotação consistiu em medir o tempo necessário para cada participante realizar 20 passos e calcular a média.

5.4 Estimação da duração do PBFT

O modelo proposto neste trabalho se baseia na variação das intensidades dos píxeis do vídeo causado pelo movimento do corpo dançando. Em uma gravação com fundo claro e uma pessoa dançando com roupas escuras é observado que o aparecimento do pé em uma determinada região irá diminuir a intensidade dos píxeis desta região. Enquanto que, quando o pé é retirado, espera-se aumento na intensidade dos píxeis da região.

As etapas do algoritmo proposto para a estimação da DPB são apresentadas na Figura 5.4, onde a entrada é o vídeo de um participante dançando. São realizadas duas etapas de pré-processamento: a extração dos quadros e a transformação para a escala de cinza. Posteriormente, os quadros são divididos em regiões e para cada região é calculada a intensidade média dos píxeis. Ao final desta etapa, o vídeo é resumido a um conjunto de séries temporais, com a variação da intensidades dos píxeis para cada região.

A próxima etapa consiste em definir quais regiões do vídeo possuem informações do movimento, para isso são escolhidas as regiões com maior desvio padrão na in-

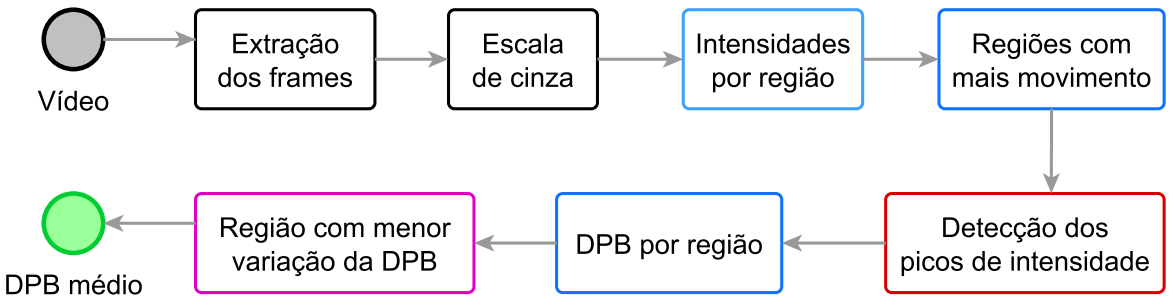


Figura 5.4: Principais etapas do algoritmo proposto.

tensidade dos píxeis. Para estas regiões são identificados os picos de intensidade, o que permite calcular as distâncias entre os picos, obtendo a duração do passo base ao longo dos quadros para cada região.

Por fim, a última etapa consiste em definir quais das regiões captura melhor a duração do passo base. Idealmente, a duração do passo base deveria ser medida a partir das regiões onde ocorrem os passos P1 ou P5 (Figura 5.1). Para isso é selecionada a região com menor variação da DPB. Com isto a DPB final é definida pela média da DPB ao longo dos quadros da região selecionada.

Um exemplo da aplicação do modelo é apresentado na Figura 5.5. Neste exemplo o algoritmo consegue identificar corretamente as regiões, detectar os picos e a região final escolhida é onde ocorre o passo P5.

5.4.1 Ajuste e avaliação do algoritmo

O algoritmo proposto possui quatro parâmetros que precisam ser ajustados: a divisão das regiões (*Reg*), o número de regiões (K) candidatas a serem utilizadas para o cálculo da duração do passo base, a proeminência (*Promi*) mínima que um pico candidato deve ter para ser detectado e o coeficiente de dispersão (*Disp*) a ser utilizado para a escolha da melhor região.

Para o ajuste e avaliação do modelo, os vídeos das 20 músicas dançadas por todos os participantes tiveram as DPBs anotadas manualmente. Os vídeos foram reproduzidos e o tempo gasto para a execução de 20 passos foi utilizado para obter a DPB média para um total de 180 vídeos. Destes, 90 foram utilizados para a realização de uma busca exaustiva para seleção das melhores combinações de parâmetros. Os 90 restantes foram utilizados para a avaliação do desempenho do algoritmo proposto.

A separação do banco de dados foi aleatória, mas garantindo que para cada participante, tivesse uma música no teste e outra no ajuste de cada faixa de BPM (Figura 5.2). Os valores possíveis para cada parâmetro utilizado no treino são apresentados a seguir:

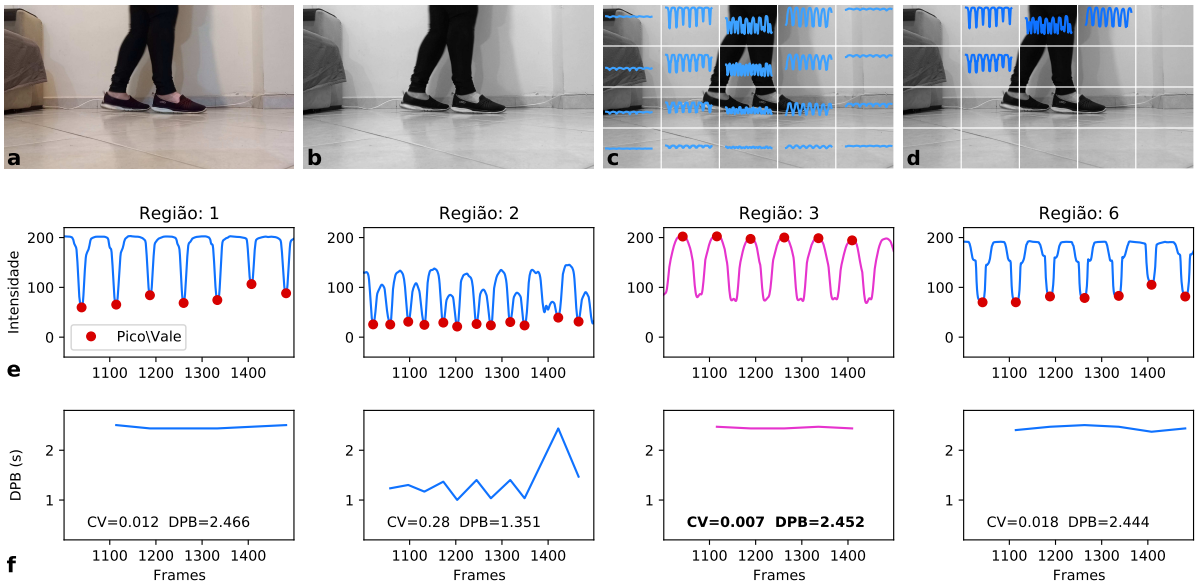


Figura 5.5: Exemplo de aplicação do modelo proposto para um trecho do vídeo do participante 11 dançando a música "Avisa"da banda Falamansa. Para o exemplo foram utilizados como parâmetros: $K = 4$, $Promi = 37$, $Reg = 5x4$ e $Disp = CV$. a) Quadro original do vídeo. b) Quadro transformado para escala de cinza. c) Divisão do vídeo em 20 regiões e apresentação da série temporal das intensidades médias de cada região ao longo dos quadros. d) Seleção das 4 regiões com maior desvio padrão em relação à intensidade média dos píxeis da região. e) Detecção de picos para as 4 regiões. f) Cálculo da DPB ao longo dos quadros e seleção da região com menor CV da DPB. A região selecionada está apresentada em rosa em e) e f).

- $Reg \in [‘5x4’, ‘6x4’, ‘7x5’, ‘7x6’]$
- $K \in [1, 2, ..., 7]$
- $Promi \in [1, 2, ..., 150]$
- $Disp \in [‘std’, ‘CV’]$, onde CV refere-se ao coeficiente de variação, definido por $CV = std/\mu$.

5.5 Forroset+

O Forroset+ surge da complementação do Forroset com as anotações da DPB e com a adição de versões com ruído doméstico das músicas do Forroset dançadas pelos participantes.

5.5.1 Inserindo a DPB

As DPBs das músicas dançadas pelos participantes foram adicionadas ao Forroset de forma que as 20 músicas dançadas por todos os participantes possuam nove anotações

manuais e as demais 360 possuem somente anotações automáticas usando o algoritmo de visão computacional proposto.

Apesar de todos os participantes recrutados terem pelo menos dois anos de experiência dançando forró, foi adicionada uma etapa de validação para identificar eventuais participantes que poderiam estar dançando de forma diferente da maioria. Logo, definiu-se as hipóteses nula e alternativa:

H_0 : Não há diferença significativa entre os participantes.
H_A : Há diferença significativa entre os participantes.

O primeiro passo consistiu em verificar a normalidade da média da DPB de cada participante a partir dos testes de D'Agostino e Pearson's, Shapiro e Lilliefors [14]. Em caso de normalidade foi realizado o teste de Análise de Variância (ANOVA), separando cada música como um bloco, pois as variações originadas pelas músicas são perturbações independentes dos participantes [15]. Seguido do teste t para amostras dependentes, aplicado par-a-par entre os participantes, para verificar quais participantes diferiram da maioria.

Não sendo constatada a normalidade dos dados, seguiu-se por uma abordagem não paramétrica utilizando Kruskal-Wallis, substituindo a ANOVA [16], e Wilcoxon, substituindo o teste t [15].

5.5.2 Músicas com ruídos domésticos

As músicas com ruído doméstico são referentes aos áudios dos vídeos gravados pelos participantes. Os mesmos foram orientados a reproduzir as músicas com um dispositivo deferente do utilizado para gravar os vídeos e sem a utilização de fones de ouvido, permitindo a aquisição das músicas durante a gravação. Portanto, semelhante à DPB, têm-se nove versões com ruído doméstico das músicas dançadas por todos os participantes e uma versão das demais músicas.

5.6 Resultados

Dos nove participantes da pesquisa, sete dançaram todas as músicas e dois deixaram de dançar uma música cada. Desta forma o ForrosetV é constituído de 538 vídeos com durações que variaram de um a três minutos.

5.6.1 Busca exaustiva dos parâmetros

Durante a busca exaustiva foram avaliadas 8400 combinações de parâmetros. O efeito da variação dos parâmetros *K*, *Disp* e *Promi* para cada divisão de região é apresentado na Figura 5.6. É possível observar que o parâmetro que mais afeta o desempenho dos modelos é a proeminência, que apresenta uma curva em 'U', com os melhores valores entre 50 e 100. Isso é esperado, uma vez que para uma baixa exigência do valor de proeminência, qualquer oscilação pode ser considerada um pico. Por outro lado, quando são utilizados valores elevados, picos relevantes podem não são considerados, chegando ao extremo de nenhum pico ser encontrado. Isso é observado na Figura 5.6c., com as curvas terminando antes de 150, valor máximo avaliado para proeminência.

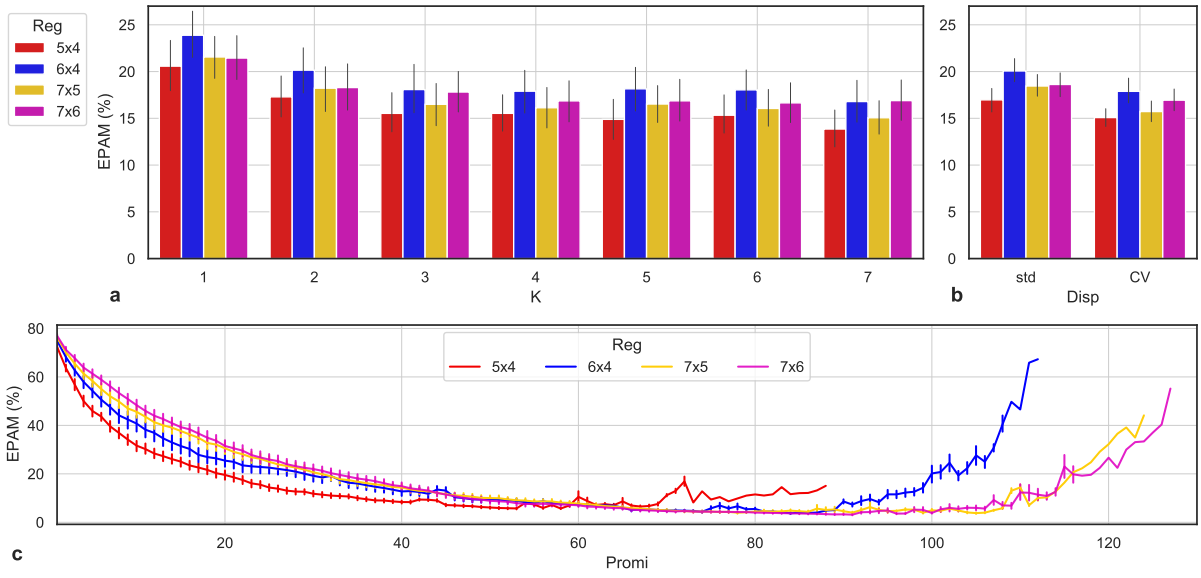


Figura 5.6: Média e desvio padrão do EPAM a partir da variação dos parâmetros com o banco de dados em função do critério de divisão das regiões do vídeo. a) Variação do valor de K. b) Variação do coeficiente de dispersão do EPM para a escolha da melhor região. c) Efeito da variação da proeminência no erro dos modelos.

5.6.2 Modelos selecionados

Todos os 5 melhores resultados foram encontrados a partir da separação do vídeo em 24 regiões (6 divisões no eixo x e 4 no eixo y). Devido a isso, foram acrescentados também a melhor combinação de cada uma das demais possibilidades de divisão. Portanto, foram selecionados oito combinações de parâmetros, ou modelos, para a avaliação final. Na Figura 5.7a. são apresentados os desempenhos dos oito melhores modelos para os dez vídeos de teste de cada participante.

Pode-se observar, variação no desempenho dos modelos a depender do partici-

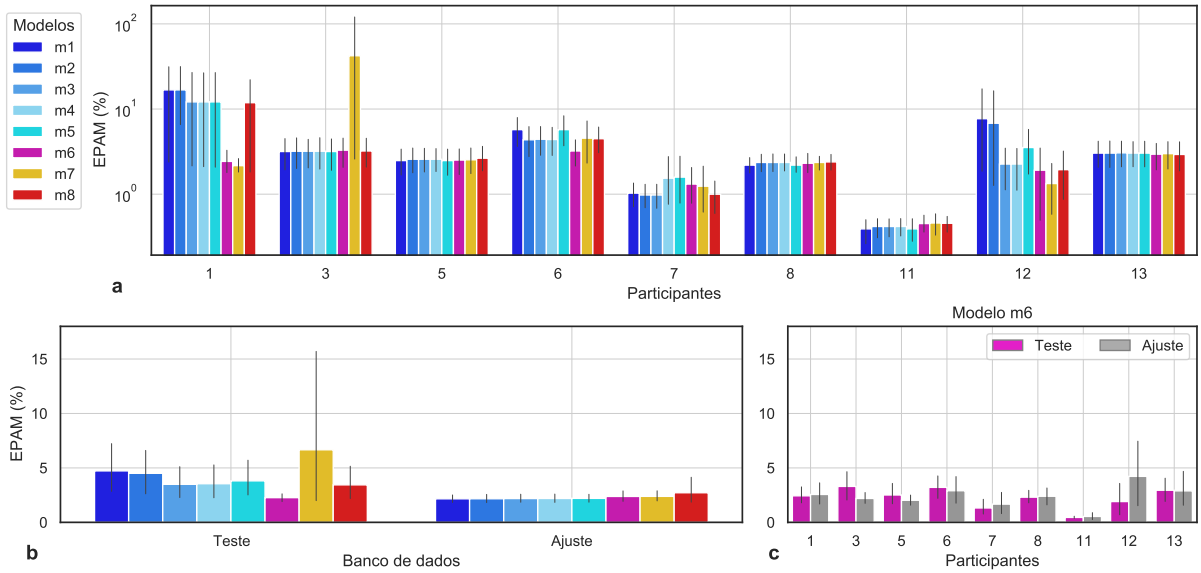


Figura 5.7: EPAM dos oito melhores modelos selecionados na etapa de treino. a) Desempenho de todos os modelos em relação à todos os participantes para os dados de teste em escala logarítmica. b) Comparação entre os erros de treino e teste para todos os modelos. c) Comparação entre os eros de treino e teste por participante para o modelo m6, modelo com menor erro para os dados de teste.

pante. Essa variação pode ser explicada por diversos fatores, como contraste entre o participante e o fundo, variação da posição do participante ao longo da dança, presença de sombras e irregularidades no tamanho dos passos realizados. Outro fenômeno esperado é a queda de desempenho dos modelos do treino para o teste, conforme apresentado na Figura 5.7b. Enquanto que no treino todos modelos tiverem EPAM menor que 3%, os erros chegaram a quase 7% no teste. A exceção é o modelo m6 com erro ligeiramente inferior no teste.

O modelo m6 foi o que atingiu menor EPAM durante o teste, conforme observado na Figura 5.7c. O EPAM para todos os participantes foi inferior a 4%, com comportamento muito próximo ao desempenho no treino e com pouca variação entre participantes. Portanto, além de ter o menor EPAM, o conjunto de parâmetros que constitui modelo m6 também foi capaz de lidar com as variações nos vídeos de um mesmo participante e entre participantes.

Os desempenhos dos oito modelos no treino e teste e os valores dos parâmetros para cada modelo são apresentados na Tabela 5.1. O modelo m6 obteve $EPAM = 2,268\%$, obtido com $Reg = '7x6'$, $Promi = 71$, $K = 7$ e usando coeficiente de variação para seleção da melhor região.

O desempenho do modelo m6 no teste é detalhado na Figura 5.8. Onde é possível observar aproximação entre os valores esperados e os preditos. Com erros percentuais absolutos (EPA) inferiores a 8% para todos os vídeos. Com EPAM de aproximadamente 2%, fica evidenciado a capacidade do modelo de estimar a duração do passo

Tabela 5.1: Apresentação dos oito modelos selecionados durante o treino e seus respectivos EPAM no treino e no teste.

					EPAM	(%)
Modelo	Reg	Promi	K	Disp	Ajuste	Teste
m1	6x4	68	3	CV	**2.161**	4.726
m2	6x4	68	4	CV	2.166	4.504
m3	6x4	69	4	CV	2.185	3.493
m4	6x4	70	4	CV	2.193	3.548
m5	6x4	70	3	CV	2.202	3.810
m6	7x6	71	7	CV	2.379	**2.268**
m7	7x5	78	7	CV	2.388	6.663
m8	5x4	37	7	CV	2.718	3.442

base do forró a partir de vídeos domésticos.

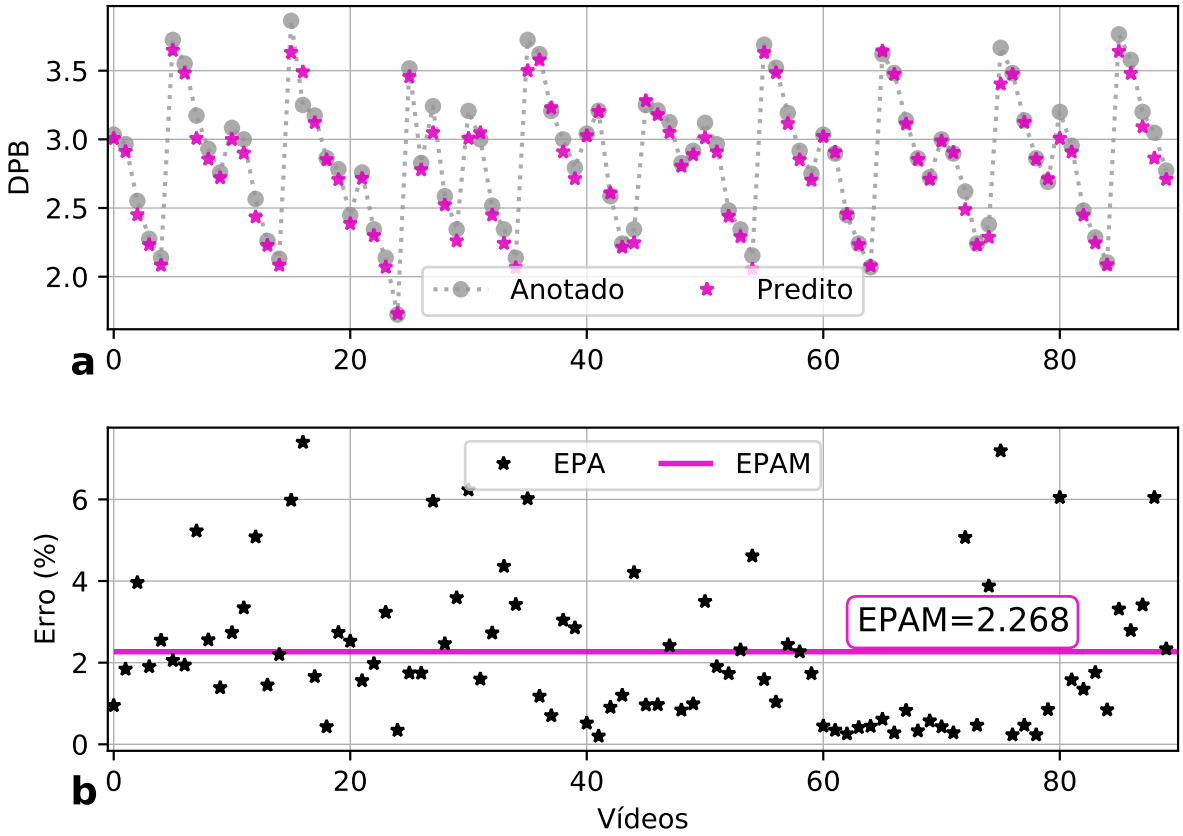


Figura 5.8: Detalhamento do desempenho do modelo m6. a) Comparação entre os valores preditos e anotados para todos os vídeos. b) Erro percentual absoluto (EPA) do modelo para todos os vídeos.

O erro apresentado pode ser explicado por falhas em alguns dos pressupostos necessários para o funcionamento do modelo como: preciso contraste entre a pessoa e o fundo; pouca variação do fundo; e que para um mesmo vídeo os passos sejam realizados aproximadamente no mesmo lugar.

5.6.3 Validação dos participantes

Para os três testes de normalidade aplicados, não foi encontrada diferença das distribuições das DPBs para a distribuição normal para o nível de significância (α) de 5%. Portanto, a ANOVA pôde ser aplicada para a comparação entre os participantes. O $p - valor$ resultante da ANOVA foi de $6,58 \cdot 10^{-15}$, indicando que há diferença significativa entre os participantes com o $\alpha = 5\%$.

Ao fazer o teste t par-a-par, para apenas um dos participantes foi encontrada diferença significativa entre todos os demais. Ao refazer a ANOVA removendo este participante do teste, o p-valor aumentou para $0,42$, indicando que não foi encontrada diferença significativa entre os demais dançarinos. Ao verificar as DPBs para cada música anotada, nota-se que este participante dançou mais rápido que todos os demais em 75% das músicas. Vale ressaltar que este participante também é o que possui menos prática em dança de forró em relação aos demais, tendo o mesmo 2 anos de prática, enquanto que os demais possuem mais de 4 anos experiência.

5.6.4 Utilização do Forroset+

Para a utilização do Forroset+ para a estimação da DPB através de modelos de aprendizado de máquinas é sugerido a utilização de validação cruzada k-fold. Desta forma as 40 músicas dançadas por cada participante são os folds. Com essas músicas pode ser realizada uma busca de hiperparâmetros usando um fold como validação e os demais para treino, por exemplo.

Desta forma, as 20 músicas dançadas por todos os participantes e com anotações manuais seriam usadas como teste. Com isso, evita-se o acúmulo de erros de predição do modelo de visão computacional e tem-se menor efeito do erro do especialista, uma vez que poderá ser utilizada a média ou mediana dos participantes.

Por fim, é sugerido a não utilização das músicas dançadas pelo participante P5 uma vez seu entendimento do ritmo divergiu dos demais, aparentemente, por falta de experiência com a dança.

5.7 Conclusão

Neste trabalho são apresentados o ForrosetV, com vídeos domésticos de pessoas dançando forró. Um algoritmo de visão computacional capaz de estimar a duração do passo base desses vídeos com erro inferior a 3% e o Forroset+ uma extensão do Forroset, com a adição da duração do passo base e músicas com ruídos domésticos.

Devido ao baixo erro e somente a necessidade de uma câmera comum para a gravação dos vídeos, o modelo poder ser utilizado para a avaliação de ritmo em

turmas de forró. Este algoritmo pode auxiliar professores e alunos durante o processo de aprendizado da dança do forró. No entanto, somente foi abordado um dos passos do forró, que deve ser realizado em um mesmo lugar e com o participante dançando sozinho. Em trabalhos futuros poderão ser adicionados outros passos de forró ao ForrosetV, bem como a utilização de vídeos com a presença de condutor e conduzido.

Por fim, o Forroset+ supre duas importantes lacunas apontadas pela literatura de estimação de ritmo de forró, sendo elas, a falta de banco de dados anotados e a falta de banco de dados com músicas que representem o cenário real da dança. Além disso, foram oferecidas orientações de uso do Forroset+ que facilita a replicação e a comparação de experimentos a serem realizados com o *dataset*.

Referências

[1] Mila Parrish. Technology in dance education. In *International handbook of research in arts education*, pages 1381–1397. Springer, 2007.

[2] Jordan Mino-Roy, Juliette St-Jean, Oliverio Lemus-Folgar, Katherine Caron, Ozalée Constant-Nolett, Jean-Philippe Després, and Camille Gauthier-Boudreault. Effects of music, dance and drama therapies for people with an intellectual disability: A scoping review. *British Journal of Learning Disabilities*, 50(3):385–401, 2022.

[3] IPHAN. Matrizes Tradicionais do Forró recebem título de Patrimônio Cultural do Brasil, 2021. URL https://bit.ly/iphan-forro.

[4] Antonio Carlos de Quadros Junior, Ellen Cristina Fontes, Romualdo Dias, and Catia Mary Volp. Caracterização do xote e do baião dançados no interior do estado de são paulo. *Movimento*, 15(3):233–247, 2009. ISSN 1982-8918. doi: 10.22456/1982-8918.2347.

[5] Augusto Dias Pereira Dos Santos, Lie Ming Tang, Lian Loke, and Roberto Martinez-Maldonado. You are off the beat! is accelerometer data enough for measuring dance rhythm? In *ACM International Conference Proceeding Series*, 2018. ISBN 9781450365048. doi: 10.1145/3212721.3212724.

[6] Lucas F Paiva, Hugo G Lopes, Leonardo B Felix, and Rodolpho VA Neves. Estimação do compasso musical do forró utilizando rede perceptron multicamadas. In *Anais do Congresso Brasileiro de Automática*, volume 2, 2020. doi: 10.48011/asba.v2i1.1331.

[7] Lucas Ferreira-Paiva, Hugo Gonçalves Lopes, Elizabeth Regina Alfaro-Espinoza, Leonardo Bonato Félix, and Rodolpho Vilela Alves Neves. Towards a device

for helping deaf people to dance: estimation of forro bar length using artificial neural network. *IEEE Latin America Transactions*, 20(6):970–976, 2022.

[8] Augusto Dias Pereira dos Santos, Kalina Yacef, and Roberto Martinez-Maldonado. Let's dance: How to build a user model for dance students using wearable technology. In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization*, page 183–191, 2017.

[9] Augusto Dias Pereira dos Santos, Lie Ming Tang, Lian Loke, and Roberto Martinez-Maldonado. You are off the beat! is accelerometer data enough for measuring dance rhythm? In *Proceedings of the 5th International Conference on Movement and Computing*, 2018.

[10] Augusto Dias Pereira dos Santos, Lian Loke, Kalina Yacef, and Roberto Martinez-Maldonado. Enriching teachers' assessments of rhythmic forró dance skills by modelling motion sensor data. 2022.

[11] Augusto Dias Pereira dos Santos, Lian Loke, and Roberto Martinez-Maldonado. Exploring video annotation as a tool to support dance teaching. In *Proceedings of the 30th Australian Conference on Computer-Human Interaction*, OzCHI '18, page 448–452, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450361880. doi: 10.1145/3292147.3292194. URL https://doi.org/10.1145/3292147.3292194.

[12] Lucas Ferreira-Paiva, Elizabeth R. Alfaro-Espinoza, Pablo de Souza Vieira Santana, Vinícius Martins Almeida, Amanda Bomfim Moitinho, Leonardo Bonato Felix, and Rodolpho Vilela Alves Neves. Forroset: A multipurpose dataset of brazilian forró music. Aceito para publicação no Advances in Artificial Intelligence - IBERAMIA 2022.

[13] Spotify. Web API Reference | Spotify for Developers, 2018. URL https://developer.spotify.com/documentation/web-api/reference/#reference-indexhttps://developer.spotify.com/documentation/web-api/reference/.

[14] Bee Wah Yap and Chiaw Hock Sim. Comparisons of various types of normality tests. *Journal of Statistical Computation and Simulation*, 81(12):2141–2155, 2011.

[15] Michael J Crawley. *The R book*. John Wiley & Sons, 2013.

[16] Patrick E McKight and Julius Najab. Kruskal-wallis test. *The corsini encyclopedia of psychology*, pages 1–1, 2010.

Capítulo 6

Conclusão

Neste trabalho foram apresentadas quatro iniciativas que buscaram criar condições para aplicação de modelos de redes neurais para a estimação de ritmo de músicas de forró. A componente rítmica estudada é o compasso uma vez que a maioria dos passos do forró são realizados ao longo de dois compassos. Conforme apresentado no Capítulo 2, para que o bom desempenho encontrado no trabalho de Paiva et al. [1], se mantenha em músicas com ruídos reais de um espaço de dança, é necessário que pelo menos parte das amostras de treinamento possuam ruído real. Além disso, a tentativa de utilizar ruídos branco para simular os ruídos reais não foi bem sucedida.

Um dos principais problemas enfrentados no trabalho exposto no Capítulo 2 foi o tamanho do banco de dados. A solução deste problema foi enfrentada pela revisão de técnicas de aumento de dados realizada no Capítulo 3 e pelo levantamento de um novo banco de dados realizado nos Capítulos 4 e 5.

A revisão mostrou que técnicas de aumento de dados podem acarretar aumentos expressivos de desempenho em modelos convolucionais. No entanto, a melhora do desempenho é acompanhada de aumento de tempo para treinamento dos modelos. Além disso, foi observado que nem todas as técnicas são efetivas, podendo ter aumentos pouco relevantes e até prejudicar o desempenho dos modelos. O principal problema é a utilização de deformações que alteram a natureza da amostra, como rotação de espectrograma, por exemplo.

Os Capítulos 4 e 5 fornecem o Forroset+, um banco de dados com anotações para 380 músicas em duas versões: gravação de estúdio e ruídos domésticos. Com o recrutamento de novos participantes este número poderá aumentar para 740, fazendo viável a implementação de modelos neurais profundos. Portanto, a partir da revisão de literatura e dos bancos de dados criados, têm se a principal contribuição deste trabalho, que é fornecer uma estrutura que propicie a utilização de modelos profundos para a predição de ritmo de músicas de forró. Esta estrutura poderá viabilizar a construção de um modelo que possa ser embarcado em um aplicativo móvel. Além disso, quando o aplicativo for implementado, a estratégia adotada no Capítulo 5 poderá ser usada para avaliar o ritmo de surdos dançando em vídeos, visando avaliar a efetividade do aplicativo.

A principal limitação do trabalho consiste na ausência de avaliação de modelos profundos com o banco de dados criado e utilizando as técnicas de aumento de dados estudadas, impossibilitando confirmar se a estrutura criada será de fato suficiente para alcançar o objetivo proposto. Portanto, como trabalho futuro têm-se a avaliação do modelo proposto por Paiva et al. [1] com o novo banco de dados, além da avaliação de modelos convolucionais e modelos convolucionais com transferência de aprendizado. Ambos os modelos podem ser avaliados em conjunto com as principais técnicas de aumento de dados apresentadas no Capítulo 3. Além disso, pode-se utilizar as músicas já coletadas e o modelo de visão computacional proposto para aumentar o banco de dados, o que poderá acarretar em maior capacidade de generalização dos modelos a serem avaliados.

Referencias

[1] Lucas F Paiva, Hugo G Lopes, Leonardo B Felix, and Rodolpho VA Neves. Estimação do compasso musical do forró utilizando rede perceptron multicamadas. In *Anais do Congresso Brasileiro de Automática*, volume 2, 2020. doi: 10.48011/asba.v2i1.1331.