Escalogramas Wavelet Aplicados à Estimativa de Andamento Musical

Luiz Alberto G. Viana, Antonio C. L. Fernandes Júnior e Eduardo F. de Simas Filho

***Resumo*— A estimativa de andamento é uma das tarefas mais fundamentais da Recuperação da Informação Musical. Neste trabalho foi escolhida uma representação do sinal de áudio como uma imagem bidimensional, através do escalograma wavelet. As imagens foram utilizadas para treinar uma Rede Neural Convolucional, relacionando a imagem com um valor de andamento alvo. Foi aplicado aumento de dados para elevar a quantidades de exemplos na qual o modelo é treinado. Os resultados foram comparados com o estado da arte. Percebeu-se que os escalogramas wavelet são capazes de representar os sinais de áudio quando utilizados no treinamento de redes neurais convolucionais.**

***Palavras-Chave*— Sinais de Áudio, Andamento Musical, MIR, Wavelet, Escalograma, CNN, Aumento de Dados.**

***Abstract*— Musical tempo estimation is one of the most fundamental tasks in the Music Information Retrieval. In this article, a wavelet scalogram is used as a two-dimensional image representation of the audio signal. The images were used to train a Convolutional Neural Network relating an image to a target tempo value. Data augmentation was used to increase the number of examples on the training. The results were compared with the state-of-the-art. It was noticed that the wavelet scalograms are able to represent the audio signals when used in training of convolutional neural networks.**

***Keywords*— Audio Signals, Musical Tempo Estimation, MIR, Wavelet, Scalogram, CNN, Data Augmentation.**

I. Introdução

Com o avanço tecnológico dos últimos anos, os sistemas computacionais estão cada vez mais interligados através das redes de internet. O ser humano vem modificando a forma de interagir com os diversos tipos de sistemas e não é diferente com o consumo da música. Plataformas de *streaming* ganham mais espaço entre os usuários e apresentações no meio digital são comumente realizadas. Neste contexto, a extensa área de Recuperação da Informação Musical (MIR - *Musical Retrieval Information*) vem aumentado a sua importância através de tarefas como identificação de músicas, recomendações, transcrição de letras, classificação de gêneros, entre outras. Pode-se destacar a ISMIR - *International Society for Music Information Retrieval*, que é uma organização sem fins lucrativos que busca o avanço da pesquisa no campo da MIR [1].

O andamento musical é a velocidade com a qual uma peça musical é executada, em BPM (batidas por minuto), e a sua estimativa é uma das tarefas mais fundamentais da MIR [2]. Através dela é possível definir o andamento de uma peça musical, o que abre possibilidades para classificações e até mesmo automatizar um acompanhamento musical para uma performance ao vivo. A comunidade de MIR tem conduzido pesquisas ao longo dos últimos 25 anos [3]. Gouyon *et al.* [4] forneceu o primeiro método de validação em larga escala para algoritmos de indução de tempo que foi o critério comparativo entre os sistemas que participaram do *ISMIR Contest*. Schreiber *et al.* [3] argumentou que, apesar dos ótimos resultados conseguidos pelos trabalhos de Böck *et al.* [2], baseado em Redes Neurais BLSTM (*Bidirectional Long Short-Term Memory*) e Schreiber e Müller, que utiliza espectrogramas e os classifica com uma CNN (*Convolutional Neural Networks*) [5], o problema de estimativa de andamento musical ainda está em aberto.

Para construir modelos de Redes Neurais com sinais de áudio como entrada é necessário decidir como os dados serão representados. Alguns trabalhos utilizaram o sinal de áudio bruto, unidimensional, para pré-processamento e extração de atributos, como Fernandes Júnior [6]. Outros trabalhos optaram por uma representação bidimensional do sinal de áudio, gerando imagens para treinar uma Rede Neural Convolucional, como Schreiber e Müller [5], Gkiokas *et al.* [7] e Sun *et al.* [8]. Chen *et al.* [9] aplicaram os escalogramas wavelet e conseguiram bons resultados no problema de modelagem de cenas de áudio . Neste contexto, Mnasri *et al.* [10] realizaram uma ampla revisão em trabalhos que detectam anomalias em sinais de áudio através de *machine learning* para aplicações na indústria, medicina, reconhecimento de voz, processamento músical, entre outros.

Este trabalho visa contribuir para o problema da estimativa de andamento musical trazendo uma forma de representação do sinal de áudio, que ainda não foi utilizada para este tipo de tarefa, os escalogramas wavelets. Foi utilizada a transformada wavelet contínua para gerar escalogramas a partir de sinais de áudio de peças musicais. Os sinais de áudio possuem andamentos pré-definidos organizados em bancos de dados comumente utilizados na literatura. Os escalogramas gerados foram utilizados para o treinamento de uma CNN. Por fim, o método de validação cruzada k-*fold* foi utilizado para garantir a generalização do modelo proposto e os resultados foram comparados com outros trabalhos publicados.

II. Modelo Proposto

O modelo proposto para estimativa de andamento consiste em gerar um escalograma wavelet a partir de um sinal de áudio de uma peça musical com o andamento em BPM pré-definido. O objetivo é realizar um aprendizado supervisionado treinando a Rede Neural Convolucional com as imagens geradas.

Intuitivamente, o problema da estimativa de andamento aparenta ser um problema de regressão para um valor inteiro.

Luiz Alberto G. Viana, Antonio C. L. Fernandes Júnior e Eduardo F. de Simas Filho, Programa de Pós-Graduação em Engenharia Elétrica, Departamento de Engenharia Elétrica e Computação, Universidade Federal da Bahia, Salvador-BA, e-mails: luiz.guimaraes@ufba.br, antonio.lopes@ufba.br, eduardo.simas@ufba.br.

Baseado na abordagem de Schreiber e Müller [5], este trabalho opta por tratar como um problema de classificação. A justificativa é que a distribuição de probabilidade entre diversas classes nos permite julgar o quão confiável é aquela estimativa. Foram definidas classes de andamento entre 23 e 257 BPM com passos de 1 BPM, ou seja, 235 classes diferentes.

A. Bancos de Dados

1) Bancos de Dados para Treinamento.: Para que o modelo seja capaz de generalizar o problema da estimativa de andamento musical, é necessário treiná-lo com conjuntos de dados que contenham exemplos de diversos estilos musicais e diversas classes de andamento. Foram escolhidos os bancos LMD Tempo (3611 exemplos), MTG Tempo (1159 exemplos) e Extended Ballroom (3826 exemplos). Estes bancos de dados são estudados e referenciados em [3].

2) Bancos de Dados para Avaliação.: Os bancos de dados escolhidos para a avaliação do modelo são amplamente utilizados na literatura. Com isso, será possível comparar os resultados alcançados com outras publicações. Logicamente, exemplos utilizados no treinamento não estarão nos conjuntos de avaliação. Foram utilizados os conjuntos ACM Mirum (1410 exemplos), Ballroom (698 exemplos), GiantSteps Tempo (660 exemplos), GTzan (999 exemplos), Hainsworth (222 exemplos), ISMIR2004 (465 exemplos) e SMC Mirum (217 exemplos) [3]. Todos possuem um conjunto de obras musicais em arquivos *wav* com um andamento pré-definido em BPM.

B. Representação do Sinal de Áudio como Imagem

Sempre que algoritmos de aprendizado profundo são utilizados para solucionar problemas que envolvem sinais de áudio, a representação do sinal é um ponto importante a ser definido. Neste trabalho, para gerar os escalogramas, foi utilizado um *offset* $= 5$ s, ou seja, foram desprezados os 5 segundos iniciais do sinal de áudio. Isto é necessário porque, comumente, os segundos iniciais de uma peça musical possuem valores de andamento diferentes do andamento global. Após isso, os sinais foram convertidos para mono (*downmixing*) e subamostrados, para 11.025 Hz, valor suficiente para detectar andamentos acima de 646 BPM [5]. Como o andamento musical não é uma característica instantânea, é necessário que o escalograma represente um espaço de tempo suficiente. Por isso, foi escolhido o valor de 11,888 segundos para que o comprimento do vetor fique representado na base 2, otimizando as operações. O vetor resultante do áudio, após a subamostragem, possui 131072 amostras. As dimensões finais de 256 *pixels* no eixo horizontal e 40 *pixels* no eixo horizontal foram escolhidas de modo a garantir que a informação de andamento seja preservada. Desta forma, cada *pixel* irá representar uma janela de 512 amostras do sinal. O processo de geração do escalograma é mostrado na Figura 1.

A partir do vetor do sinal de áudio, pode-se gerar o escalograma wavelet aplicando inicialmente a Transformada Wavelet Contínua (CWT - *Countinuous Wavelet Transform*). Dado um sinal $f(t)$, sua CWT é definida como:

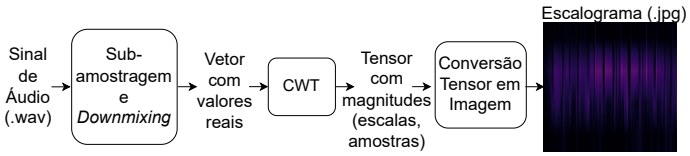

Fig. 1. Processo de geração do escalograma wavelet. O sinal de áudio passa pelos processos de subamostragem e *downmixing*, aplicação da CWT e por fim a conversão do tensor em imagem, resultando no escalograma.

$$\mathcal{W}_f^{\psi}(a,\tau) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} f(t)\psi^* \left(\frac{t-\tau}{a} \right) dt, \qquad (1)$$

em que o parâmetro a (>0) se refere à escala e τ à translação ou localização da função analisadora wavelet ψ, ou wavelet-mãe, sendo a e $\tau \in \mathbb{R}$. O parâmetro a controla a dilatação/contração da função analisadora wavelet. O asterisco superior em ψ^* denota o complexo conjugado da função ψ e $\mathcal{W}_f^{\psi}(a,\tau)$ é conhecido como coeficiente wavelet [13]. Existem diversas funções analisadoras wavelets que podem ser utilizadas na análise de sinais. Essa escolha pode alterar os resultados obtidos de forma a enfatizar uma certa característica do sinal analisado. A wavelet escolhida para este trabalho é conhecida como Chapéu Mexicano, e é definida como:

$$\psi(t) = \frac{2}{\sqrt{3}\sqrt[4]{\pi}} e^{-\frac{t^2}{2}} (1 - t^2). \qquad (2)$$

Esta wavelet analisadora foi escolhida como ponto de partida analisando visualmente os escalogramas gerados à partir de diferentes funções. A função Chapéu Mexicano gerou um escalograma com *pixels* mais intensos, destacando melhor os momentos de pulsos durante a peça musical.

Os escalogramas são gráficos que representam a visualização bidimensional dos coeficientes wavelets $\mathcal{W}_f^{\psi}(a,\tau)$. Estes podem ser visualizados por meio de um campo de isolinhas ou imagem [13]. A Figura 2 mostra um escalograma wavelet gerado a partir da forma de um sinal de áudio. Esse escalograma foi gerado após a aplicação da CWT utilizando a função analisadora wavelet Chapéu Mexicano, com quatro níveis de escala [1.3, 11, 45.5, 130]. O eixo horizontal do escalograma representa o tempo, em segundos, enquanto que o eixo vertical mostra a representação discreta dos níveis de escala. A wavelet Chapéu Mexicano foi escolhida analisando visualmente escalogramas gerados com diversos tipos de funções.

C. Arquitetura e Treinamento da Rede Convolucional

Testes preliminares mostraram que redes clássicas para classificação de imagens, como VGG16 e InceptionV3, não tiveram um bom desempenho para classificar os escalogramas. Em todos os casos, o modelo se especializava nos exemplos de treinamento e não conseguia generalizar para os exemplos de validação e teste. Por isso, optou-se por utilizar a CNN utilizada por Schreiber e Müller [5]. Esta rede conseguiu bons resultados utilizando espectrogramas-mel, e por isso espera-se que ela apresente um bom desempenho ao ser treinada com os escalogramas wavelet. O diferencial desta arquitetura é que

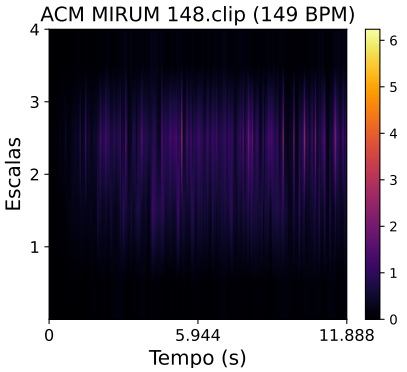

Fig. 2. Escalograma gerado a partir da CWT aplicada ao exemplo ACM Mirum 148.clip, mostrado na Figura 3. Foi utilizada a função analisadora wavelet Chapéu Mexicano com quatro níveis de escala [1.3, 11, 45.5, 130]. A paleta de cores mostra a intensidade de cada *pixel*, que representa os valores dos coeficientes wavelets.

todas as convoluções são do tipo *"same"*, ou seja, o *padding* é utilizado para que a imagem permaneça com a mesma dimensão, e o *stride* igual a um. Como os filtros possuem dimensão unitária ao longo do eixo vertical, o formato do tensor permanece inalterado ao longo do eixo do tempo, que é primordial para detecção do andamento. A arquitetura CNN utilizada pode ser observada na Figura 3.

Após as camadas convolucionais com filtros curtos tem-se os módulos multifiltros, na qual a estrutura pode ser observada na Figura 4. Esses módulos tem como objetivo reduzir a dimensionalidade ao longo do eixo das escalas, resumindo a informação e combinando o sinal com uma variedade de filtros que são capazes de detectar dependências temporais [5].

Para classificar os atributos gerados pelas camadas convolucionais são adicionadas duas camadas densas com 64 neurônios cada, com função de ativação ELU (*Exponential Linear Unit*). A camada de saída possui 235 neurônios, representando as 235 classes de andamento e com função de ativação *softmax*. Esta arquitetura resulta em uma CNN com um total 2.920.319 parâmetros, sendo 2.919.677 treináveis e 642 não treináveis.

Os bancos de dados de treinamento são unificados e aleatoriamente divididos em cinco partes, para utilização da validação cruzada k-*fold*, com k=5. Destas cinco partes, quatro são usadas para o treinamento, e uma parte dividida entre validação e teste, ou seja, 80% treinamento, 10% validação e 10% teste. Todos os tensores são normalizados antes de iniciar o treinamento do modelo. Cada valor de k representa um modelo treinado com diferentes exemplos dos conjuntos de dados.

A avaliação do desempenho de modelos de estimativa de andamento possui uma particularidade. É comum que mesmo um humano treinado estime um andamento de uma peça musical com valores múltiplos e submúltiplos de 2 ou 3 do andamento definido. Isto porque uma peça musical pode ser compatível com diferentes valores absolutos de andamento, a depender da forma dos elementos rítmicos que a compõe. Por isso, será utilizada a forma de avaliação escolhida por Schreiber e Müller [5], Fernandes Júnior [6] e também em diversos outros trabalhos.

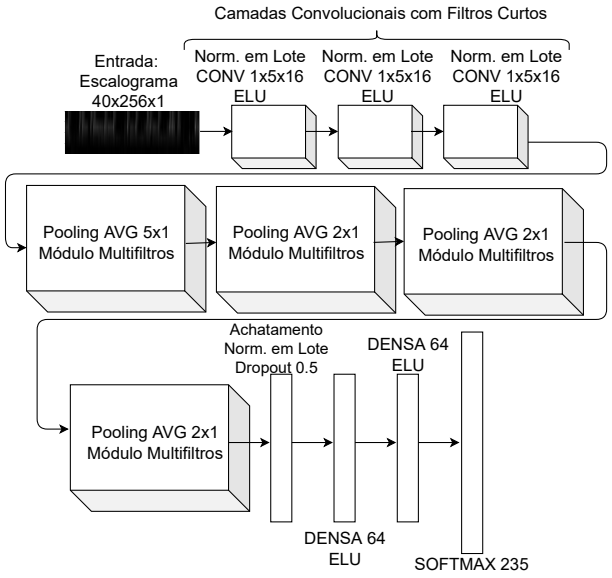

Fig. 3. Arquitetura Rede Neural Convolucional. Adaptado de [5]. É compostada pela camada de entrada seguida por três camadas convolucionais com filtros curtos, quatro módulos multifiltros e por fim o tensor é achatado e conectado a duas camadas totalmente conectadas. A camada de saída é uma softmax com 256 classes. No caso de uma regressão, haverá apenas um neurônio na camada de saída.

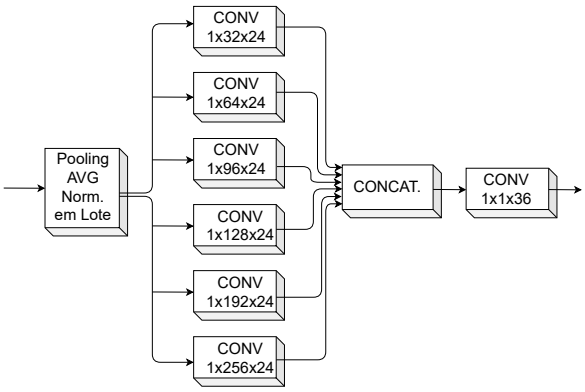

Fig. 4. Módulo Multifiltros. Sua característica principal são as convoluções em paralelo com diferentes dimensões de filtros. Todas as funções de ativação são ELU. Adaptado de [5].

A Acurácia 0, é definida como a acurácia real do modelo, quando a rede neural convolucional consegue prever exatamente o andamento (Γ) da peça musical, $\hat{\Gamma} = \Gamma$. A Acurácia 1, considera valores dentro de uma janela de precisão de 4%, $\hat{\Gamma} = \Gamma \pm 4\%$. Este critério leva em consideração que esta diferença mínima é imperceptível ao ouvido humano, e mesmo pessoas bem treinadas podem prever andamentos com a mesma margem de erro. Por fim, a Acurácia 2, considera também os submúltiplos(1/2 e 1/3) e múltiplos (2 e 3) para o valor real do andamento, dentro de uma janela de precisão de 4%, $\hat{\Gamma} = (\Gamma \pm 4\%)\, M$, onde $M \in \{\frac{1}{2}, 1, 2, 3\}$.

D. Aumento de Dados

É fundamental que o aumento de dados seja realizado apenas no conjunto de treinamento, e nunca na validação e

no teste. Como o banco de dados de treinamento possui 8596 exemplos e 80% é utilizado no treinamento, aproximadamente 6877 exemplos passam pelo aumento de dados. O primeiro aumento foi realizado simplesmente alterando o *offset* para 10,0 segundos durante a geração do escalograma, dobrando o número de exemplos.

O segundo aumento foi realizado alterando a dimensão do escalograma ao longo do eixo do tempo e, consequentemente, alterando o valor real do andamento, porém mantendo as características no eixo das escalas. O fator de alteração (F_a) foi escolhido aleatoriamente em um grupo de valores pré-definidos $F_a \in \{0.8, 0.85, 0.9, 0.95, 1.05, 1.1, 1.15, 1.2\}$. Após alterar a dimensão do escalograma, ele é reajustado para a dimensão padrão utilizando interpolação bilinear, e o valor de andamento é modificado utilizando o mesmo fator. Após este segundo aumento, o conjunto de treinamento tem sua quantidade original de exemplos triplicada.

III. Resultados

A CNN foi treinada utilizando a *categorical crossentropy* como função custo. Desta forma, o modelo tende a melhorar os resultados para a Acurácia 0 ao longo do treinamento. Porém, o resultado mais significativo é o da Acurácia 2 que considera os erros de oitava.

Um ótimo desempenho no conjunto de treinamento não significa que a CNN generaliza para os conjuntos de validação e teste, por isso o melhor modelo é o que possui um resultado equilibrado entre os conjuntos. Durante o treinamento, o conjunto foi dividido em 5 partes e, a cada um dos 5 modelos (k=1,2,3,4 e 5), uma das partes foi usada como validação. Os modelos foram treinados durante 15 épocas. Ao final, o melhor resultado para o conjunto de validação foi de 92,52% para Acurácia 2. Porém, em outro modelo o conjunto de validação atingiu um pico de 95,24%. Para o conjunto de teste, o melhor resultado foi uma Acurácia 2 de 92,30%.

O resultado final do treinamento, que é a média entre os valores de cada um dos modelos, foi de 98,41% para o conjunto de treinamento, 90,38% para o conjunto de validação e 91,48% para o conjunto de teste. Observando todos os valores de k, foi possível definir que o modelo k=1 possui o resultado mais equilibrado. Por isso, são apresentados gráficos que representam este melhor modelo.

As Figuras 5, 6, 7, mostram os gráficos de valores de andamento reais *versus* valores de andamento estimados, em BPM. O gráfico da Figura 5 mostra as previsões para o conjunto de treinamento, o gráfico da Figura 6 para o conjunto de validação e o gráfico da Figura 7 para o conjunto de teste. A reta vermelha representa quando a CNN prevê exatamente o valor de andamento real da peça musical. A zona sombreada em vermelho representa a margem de erro de 4%, permitida pelas Acurácias 1 e 2. As retas verdes representam os acertos de múltiplos e submúltiplos permitidos pela Acurácia 2, assim como a zona sombreada em verde. É possível observar, principalmente nos dados de treinamento, que os erros são concentrados nas retas onde $\hat{\Gamma} = 2\Gamma$ e $\hat{\Gamma} = (1/2)\Gamma$.

A. Avaliação do Modelo

Nesta etapa foram calculadas as Acurácias 0, 1 e 2 para que fossem comparadas com outros trabalhos. A última coluna das

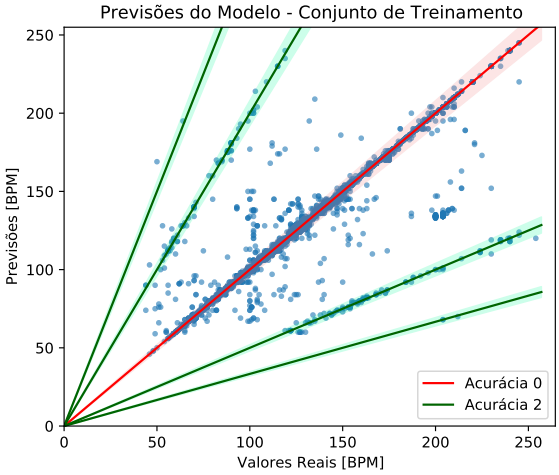

Fig. 5. Valores Reais *versus* Valores Estimados para o Conjunto de Treinamento - k=1.

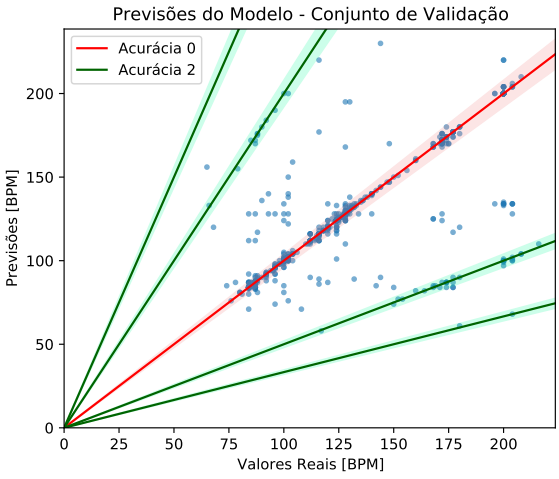

Fig. 6. Valores Reais *versus* Valores Estimados para o Conjunto de Validação - k=1.

Tabelas I, II e III mostram respectivamente as Acurácias 0, 1 e 2 para o modelo proposto (k=1). Destaca-se o desempenho nos conjuntos ACM Mirum, Ballroom, GiantSteps e o Combinados, com valores acima de 80%.

Em 2020, Schreiber *et al.* [3] apontou os trabalhos do Schreiber e Müller [5] e Böck *et al.* [2] como os principais trabalhos de estimativa de andamento musical. Por isso, as Tabelas I, II e III mostram respectivamente as Acurácias 0, 1 e 2 para os trabalhos citados em comparação com o modelo proposto. O modelo proposto neste trabalho está representado na tabela como k=1. O trabalho do Schreiber e Müller [5] utiliza um modelo parecido com o modelo proposto, porém com um maior aumento de dados e utilizando os espectrogramas como entrada do treinamento, enquanto que o trabalho do Böck *et al.* [2] utiliza redes neurais recorrentes e o áudio pré-processado como entrada.

Analisando a Tabela I é possível ver que o modelo proposto

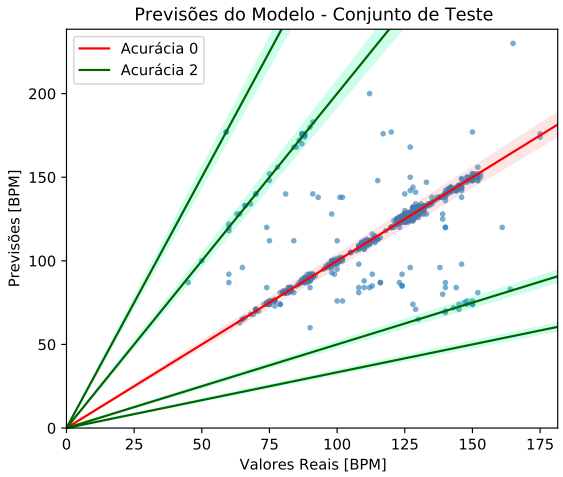

Fig. 7. Valores Reais *versus* Valores Estimados para o Conjunto de Teste - k=1.

se aproximou do melhor resultado de Acurácia 0 para o banco de dados Ballroom, chegando a 64,41%. Para a Acurácia 1 o modelo conseguiu bons resultados para o Ballroom e para o GiantSteps, superando o algoritmo do Böck *et al.* [2] para este último banco, conforme Tabela II. Para a Acurácia 2 o modelo obteve bons resultados para o ACM Mirum, Ballroom e GiantSteps, mas não conseguiu superar os trabalhos comparados, conforme Tabela III.

TABELA I
COMPARAÇÃO COM O ESTADO DA ARTE - ACURÁCIA 0

Acurácia 0	Schr [5]	Böck [2]	k=1
ACM Mirum	40,60%	29,40%	25,39%
Ballroom	67,90%	33,80%	64,41%
GiantSteps	59,80%	37,20%	16,96%
SMC	12,40%	17,10%	4,15%
ISMIR04	34,10%	27,20%	19,78%
Hainsworth	43,20%	33,80%	28,05%
GTzan	36,90%	32,20%	24,12%
Combinados	44,80%	31,20%	25,47%

TABELA II
COMPARAÇÃO COM O ESTADO DA ARTE - ACURÁCIA 1

Acurácia 1	Schr [5]	Böck [2]	k=1
ACM Mirum	79,50%	74,00%	66,24%
Ballroom	92,00%	84,00%	83,33%
GiantSteps	73,00%	58,90%	67,12%
SMC	33,60%	44,70%	18,43%
ISMIR04	60,60%	55,00%	45,81%
Hainsworth	77,00%	80,60%	64,25%
GTzan	69,40%	69,70%	54,65%
Combinados	74,20%	69,50%	60,41%

IV. CONCLUSÕES E TRABALHOS FUTUROS

Foi possível observar que o modelo proposto atingiu bons resultados nos conjuntos de dados utilizados no treinamento da rede neural convolucional, porém ainda não conseguiu atingir os resultados do estado da arte. Isto mostra que os escalogramas wavelets são capazes de representar os sinais de áudio, assim como os espectrogramas, porém abre possibilidades para manipular a representação podendo gerar resultados superiores em algumas aplicações. Como trabalho futuro, é possível variar os parâmetros de geração do escalograma wavelet, como a função analisadora, e a quantidade e valores de escala e observar se isto impactará nos resultados do modelo.

TABELA III
COMPARAÇÃO COM O ESTADO DA ARTE - ACURÁCIA 2

Acurácia 2	Schr [5]	Böck [2]	k=1
ACM Mirum	97,40%	97,70%	87,94%
Ballroom	98,40%	98,70%	90,68%
GiantSteps	89,30%	86,40%	82,88%
SMC	50,20%	67,30%	29,49%
ISMIR04	92,20%	95,00%	74,19%
Hainsworth	84,20%	89,20%	72,85%
GTzan	92,60%	95,00%	78,88%
Combinados	92,10%	93,60%	80,12%

Métodos alternativos de aumentos de dados também poderão aumentar o número de exemplos para apresentar ao modelo, fazendo com que os valores de acurácia se elevem. A arquitetura da rede também pode ser alterada, inserindo mais módulos multi-filtros ou realizando conexões entre as camadas.

REFERÊNCIAS

[1] *ISMIR - International Society for Music Information Retrieval*. Disponível em: `ismir.net`. Acesso em: 16 agosto 2022.

[2] S. Böck, F. Krebs, G. Widmer. "Accurate Tempo Estimation Based on Recurrent Neural Networks and Resonating Comb Filter". *16th International Society for Music Information Retrieval Conference*, pp. 486-493, 2015.

[3] Hndrik Schreiber, Julián Urbano , Meinard Müller. "Music Tempo Estimation: Are We Done Yet?". *Transactions of the International Society for Music Information Retrieval*, vol. 3, pp. 111, 2020.

[4] F. Gouyon, A. Klapuri, S. Dixon, M. Alonso, G. Tzanetakis, C. Uhle, P. Cano. "An Experimental Comparison of Audio Tempo Induction Algorithms". *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1832– 1844, 2006.

[5] Hndrik Schreiber, Meinard Müller. "A Single-step Approach to Musical Tempo Estimation Using a Convolutional Neural Network". *19th International Society for Music Information Retrieval Conference*, pp. 98-105, 2018.

[6] Antônio Carlos L. Fernandes Júnior. "Contribuições ao Problema de Extração de Tempo Musical". Tese (Doutorado), Campinas, São Paulo, Brasil, 2015.

[7] A. Gkyokas, V. Katsouros. "Convolutional Neural Network for Real-Time Beat Tracking: A Dance Robot Application". *18th International Society for Music Information Retrieval Conference*, 2017.

[8] X. Sun, Q. He, Y. Gao, W. Li. "Musical Tempo Estimation using a Multi-Scale Network". *22th International Society for Music Information Retrieval Conference, online*, 2021.

[9] H. Chen, P. Zhang, H. Bai, Q. Yuan, X. Bao, Y. Yan. "Deep Convolutional Neural Network With Scalogram for Audio Scene Modeling". *Interspeech 2018, Hyderabad*.

[10] Z. Mnasri, S. Rovetta, F. Masulli, A. Cabri. "Anomalous Sound Event Detection: A Survey of Machine Learning Based Methods and Applications". *Multimed Tools Appl 81*, pp. 5537–5586, 2022. doi:/10.1007/s11042-021-11817-9.

[11] UNC Vision Lab. *Large Scale Visual Recognition Challenge 2016 (ILSVRC2016)*. 2016. Disponível em: <http://image-net.org/challenges/LSVRC/2016/index>. Acesso em: 07 março 2021.

[12] A. Géron. "Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools and techniques to build intelligent Systems". *O'Reilly Media, Sebastopol*, CA, 2019.

[13] M. Domingues, O. Mendes, M. Kaibara, V. Menconi, E. Bernardes. "Explorando a Transformada Wavelet Contínua". *Revista Brasileira de Ensino de Física*, vol. 38, 2016. doi:10.1590/1806-9126-RBEF-2016-0019.