Estimativa de Andamento Musical Através de Escalogramas Wavelet e Redes Neurais Convolucionais

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 de Computação
Universidade Federal da Bahia
Salvador, Bahia
E-mails: luiz.guimaraes@ufba.br, antonio.lopes@ufba.br, eduardo.simas@ufba.br.

***Resumo*— O andamento musical é a velocidade com a qual uma peça musical é executada e a sua estimativa é uma das tarefas mais fundamentais da área de Recuperação da Informação Musical (MIR - *Musical Information Retrieval*). Com o objetivo de contribuir com este tipo de tarefa, este trabalho propõe um modelo para estimar o andamento a partir de um sinal de áudio musical. O sinal de áudio foi representado através do escalograma wavelet, que é uma imagem bidimensional. Foram testadas diferentes formas de geração do escalograma wavelet, variando a função wavelet analisadora e os níveis de escala. As imagens foram utilizadas para treinar uma Rede Neural Convolucional (CNN - *Convolutional Neural Network*) realizando um aprendizado supervisionado, relacionando a imagem com um valor de andamento alvo. O método de validação cruzada *k-fold* foi utilizado para gerar uma maior confiabilidade estatística do modelo proposto e definir o melhor resultado para as escolhas envolvendo os parâmetros de geração dos escalogramas. Foi implementado o aumento artificial de dados em tempo real, modificando os escalogramas durante a rotina de treinamento. Por fim, o modelo foi avaliado em bancos de dados amplamente utilizados na literatura e os resultados foram comparados ao estado da arte. Resultados compatíveis ao estado da arte foram atingidos em um dos bancos de dados de avaliação, o "GiantSteps", atingindo uma acurácia (Tipo 2 - ACC2) de 92,6% com as wavelets analisadoras Morlet e Shannon.**

***Palavras-Chave*— Andamento Musical, Wavelet, Escalograma, Rede Neural Convolucional, Aumento artificial de dados, MIR.**

***Abstract*— Musical tempo, the pace or speed at which a musical piece is performed, stands as a critical aspect within the realm of Musical Information Retrieval (MIR). In pursuit of enhancing this facet, this study presents a model for tempo estimation derived from a musical audio signal. The audio signal is represented through a wavelet scalogram, a two-dimensional image. Various methodologies for generating the wavelet scalogram were explored, involving variations in the analyzing wavelet function and scale levels. These images were harnessed to train a Convolutional Neural Network (CNN) using supervised learning, associating each image with a target tempo value. To bolster statistical reliability and determine optimal outcomes regarding scalogram generation parameters, the k-fold cross-validation method was deployed. Additionally, a real-time data augmentation technique was integrated, wherein scalograms were dynamically modified during the training process. Ultimately, the model underwent evaluation across widely referenced datasets in the literature, with resultant performances compared against the state of the art. Commensurate with the state of the art outcomes were achieved on the "GiantSteps"evaluation dataset, registering a Type 2 accuracy (ACC2) of 92.6% using Morlet and Shannon analyzing wavelets.**

***Keywords*— Audio Tempo Estimation, Musical Tempo, Wavelet, Scalogram, Convolutional Neural Network, Data Augmentation, MIR.**

I. Introdução

O andamento musical é a velocidade com a qual uma peça musical é executada, normalmente medida em BPM (batidas por minuto), e a sua estimativa é uma das tarefas mais fundamentais da extensa área de Recuperação da Informação Musical (MIR - *Musical Retrieval Information*) [1]. Através dela é possível definir o andamento de uma peça musical, o que abre possibilidades para classificação automática e até mesmo automatização de um acompanhamento musical para uma performance ao vivo.

Começando com os trabalhos de Goto e Muraoka [2] e Scheirer [3], a comunidade de MIR tem conduzido pesquisas sobre a estimativa de andamento ao longo dos últimos 25 anos [4]. Gouyon *et al.* [5] forneceram o primeiro método de validação em larga escala para algoritmos de indução de andamento que foi o critério comparativo entre os sistemas que participaram do ISMIR 2004 *Contest* [6]. Em 2011, Zapata e Goméz [7] atualizaram a visão geral sobre sistemas de estimativa de andamento. Diversos trabalhos recentes passaram a utilizar Redes Neurais combinadas com o pré-processamento do sinal de áudio. Böck e Schedl [8] introduziram um método para detecção de *onsets* baseado em Redes Neurais B-LSTM (*Bidirectional Long Short-Term Memory*). Gkiokas *et al.* [9] utilizaram Redes Neurais Convolucionais (CNN - *Convolutional Neural Networks*) para estimativa de andamento e *beat tracking* em tempo real.

A representação de sinais de áudio como imagens, em duas ou três dimensões, pode ser realizada de diversas formas utilizando abordagens tempo-frequência: MFCCs (*Mel-Frequency Cepstrum Coefficients*), forma de onda do sinal de áudio, espectrogramas, etc. Os espectrogramas, que são gerados a partir da Transformada de Fourier, se tornaram bastante populares porque eles geram bons resultados quando combinados com as Redes Neurais Convolucionais [10]. Uma alternativa para a utilização dos espectrogramas são os Escalogramas Wavelet.

Fernandes Júnior [11] utilizou a transformada wavelet para o pré-processamento do sinal de áudio no problema de estimativa de andamento musical, mostrando que a transformada wavelet possui boa capacidade de localizar eventos ao longo do tempo, quando comparada com a Transformada de Fourier. Chen *et al.* [12] aplicou os escalogramas wavelet e conseguiu bons resultados no problema de modelagem de cenas de áudio.

Em 2020, Schreiber *et al.* [4] argumentaram que, apesar dos ótimos resultados conseguidos pelos trabalhos de Böck *et al.* [1] e Schreiber e Müller [13], o problema de estimativa de andamento musical ainda está em aberto. Isto porque as métricas utilizadas para validação desses modelos não apresentam relação direta com aplicações reais e são focadas em eliminar os chamados erros de oitava (*octave erros*). Este fato não diminui a necessidade de estudo deste tipo de métrica, mas enfatiza que os resultados para outras métricas ainda podem ser melhorados. Entre trabalhos recentes publicados podem-se destacar Souza *et al.* [14], Sun *et al.* [15] e Quinton [16]. Souza *et al.* obtiveram bons resultados com exemplos musicais exclusivamente percussivos, utilizando redes B-LSTM. Sun *et al.* obtiveram bons resultados para a Acurácia 1, utilizando a técnica de multi-scale network. Quinton trouxe uma abordagem de método auto supervisionado, onde não é necessário que os exemplos do banco estejam rotulados. Em 2023, Morais *et al.* [17] explorou um modelo auto supervisionado para estimativa de *pitch*, adaptado para estimativa de andamento musical.

Neste contexto, este trabalho visa contribuir para o problema da estimativa de andamento musical, complementando o trabalho de Viana *et. al.* [18] na utilização dos escalogramas wavelet. Foi possível estudar diferentes formas para se gerar um escalograma e observar os impactos na estimativa de andamento. Foi utilizada a transformada contínua wavelet 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 serão utilizados como entradas para o treinamento de uma Rede Neural Convolucional (CNN). Como a quantidade de exemplos nos principais bancos de dados da literatura é limitada, foi utilizada a técnica de aumento de dados, na qual os exemplos reais sofrem modificações para que novos exemplos sejam gerados. Por fim, o método de validação cruzada k-fold foi utilizado para gerar uma maior confiabilidade estatística do modelo proposto, que foi avaliado analisando a relação entre os escalogramas e os resultados dos valores estimados. Os resultados se aproximaram do estado da arte e foi possível determinar qual wavelet se comportou melhor para os parâmetros estudados de geração dos escalogramas.

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, conforme Figura 1.

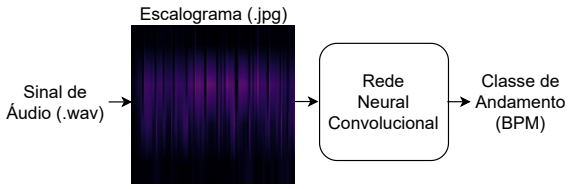

Fig. 1: Modelo proposto simplificado. O sinal de áudio é transformado em um escalograma wavelet. O escalograma é utilizado para o treinamento da rede neural convolucional que irá detectar o valor do andamento musical em BPM.

Intuitivamente, o problema da estimativa de andamento aparenta ser um problema de regressão para um valor inteiro. Baseado na abordagem de Schreiber e Müller [13], 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.

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 [4].

2) Bancos de Dados para Avaliação: Os bancos de dados escolhidos para a avaliação do modelo são amplamente utilizados na literatura. Com isso, foi possível comparar os resultados alcançados com outras publicações. Logicamente, exemplos utilizados no treinamento não estão 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) [4]. Todos possuem um conjunto de obras musicais em arquivos *wav* com um andamento pré-definido em BPM.

Observando todos os bancos de dados disponíveis, é possível determinar que os valores de andamento variam entre 23 e 257 BPM. Por isso, podem-se definir classes de andamento variando entre 23 e 257 BPM com passos de 1 BPM, ou seja, 235 classes diferentes. Porém, na seção II-C serão discutidos os chamados "erros de oitava", em que o modelo pode prever um múltiplo ou submúltiplo do valor de andamento original e ainda assim este ser considerado um acerto. Devido a este tipo de erro, optou-se por concentrar o treinamento no intervalo onde ocorre a maioria dos exemplos, resultando em classes entre 60 e 199 BPM, reduzindo para 140 classes diferentes. Este tipo de intervalo é chamado por alguns autores como *sweet octave* [19], que é o intervalo que possui mais peças musicais do que qualquer outro.

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 [13]. 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 vertical 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 2. 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:

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

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 [20]. Existem diversas funções analisadoras wavelets que podem ser utilizadas na análise de sinais. Ainda não se sabe qual função contínua wavelet se comporta melhor para o problema de estimativa de andamento musical. Por isso, neste trabalho foram investigadas algumas funções wavelets conhecidas na literatura, que são utilizadas nas mais diferentes aplicações, para verificar se existem variações significativas no resultado. Foram escolhidas as wavelets: Chapéu Mexicano, Morlet e Shannon [20].

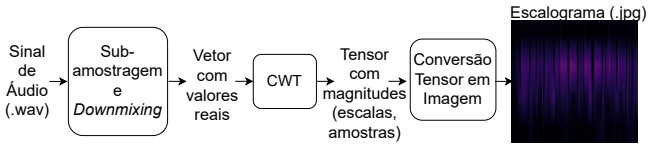

Fig. 2: Processo de geração do escalograma wavelet. O sinal de áudio passa pelos processos de subamostragem e *downmixing*, aplicação da CWT e por fim a conversão do tensor em imagem, resultando no escalograma.

O segundo parâmetro para aplicar a CWT são as escalas a a serem utilizadas. Elas podem variar em quantidade e em valores do parâmetro a da Equação 1. Cada valor escolhido para o parâmetro a refere-se a uma faixa de frequência na qual o sinal está sendo observado. Esta faixa de frequência de observação também varia a depender da wavelet utilizada e também da taxa de amostragem do sinal. As escalas foram escolhidas buscando um equilíbrio entre as regiões do espectro comumente utilizado para equalização, sendo elas as frequência subgraves (20-60 Hz), baixas (60-250 Hz), médias-baixas (250-2000 Hz), médias-altas (2000-6000 Hz) e altas (6000-20000 Hz).

A Tabela I mostra os diferentes parâmetros para geração dos escalograma que foram utilizados para treinar, validar e testar o modelo em diferentes experimentos. Na primeira coluna está o número do experimento e na segunda o rótulo utilizado.

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 [20]. A Figura 3 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.

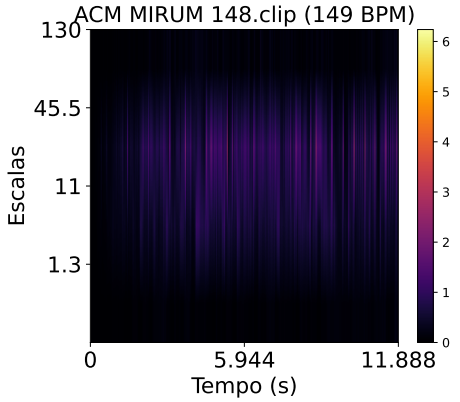

Fig. 3: Escalograma gerado a partir da CWT aplicada ao exemplo ACM Mirum 148.clip. Foi utilizada a função analisadora wavelet Chapéu Mexicano com quatro níveis de escala [1.3, 11, 45.5, 130]. A paleta de cores mostra a intensidade de cada *pixel*, que representa os valores dos coeficientes wavelets.

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 [13]. Esta rede conseguiu bons

TABELA I: Parâmetros de Geração dos Escalogramas

Exp.	Escalograma	Função Analisadora	Escalas	Frequências de Observação (Hz)
1	Mexh_4	Chapéu Mexicano	[1.3, 11, 45.5, 130]	[2120, 251, 61, 21]
2	Mexh_6	Chapéu Mexicano	[0.13, 0.45, 1.3, 11, 45.5, 130]	[21202, 6125, 2120, 251, 61, 21]
3	Mexh_40	Chapéu Mexicano	[1.3, ..., 130] (40 escalas)	[2120, ..., 21]
4	Morl_4	Morlet	[4.4, 35.5, 149, 400]	[2036, 252, 60, 22]
5	Morl_6	Morlet	[0.44, 1.49, 4.4, 35.5, 149, 400]	[20358, 6012, 2036, 252, 60, 22]
6	Morl_40	Morlet	[4.4, ..., 400] (40 escalas)	[2036, ..., 22]
7	Shan_4	Shannon	[1.51, 12, 50.5, 149]	[2008, 253, 60, 20]
8	Shan_6	Shannon	[0.15, 0.50, 1.51, 12, 50.5, 149]	[20213, 6064, 2008, 253, 60, 20]
9	Shan_40	Shannon	[1.51, ..., 149] (40 escalas)	[2008, ..., 20]

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 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 4.

Após as camadas convolucionais com filtros curtos tem-se os módulos multifiltros, para os quais a estrutura pode ser observada na Figura 5. 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 [13].

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 140 neurônios, representando as 140 classes de andamento e com função de ativação *softmax*. Esta arquitetura resulta em uma CNN com um total 2.840.392 parâmetros, sendo 2.839.750 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.

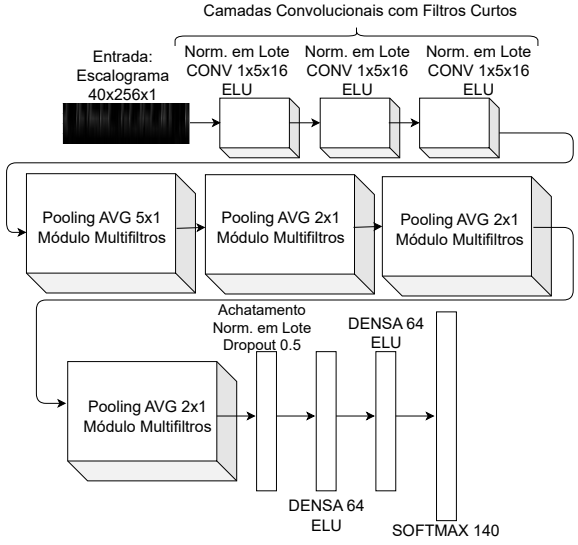

Fig. 4: Arquitetura da Rede Neural Convolucional. Adaptado de [13]. É compostada pela camada de entrada seguida por três camadas convolucionais com filtros curtos, quatro módulos multifiltros e por fim o tensor é achatado e conectado a duas camadas totalmente conectadas. A camada de saída é uma softmax com 140 classes.

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

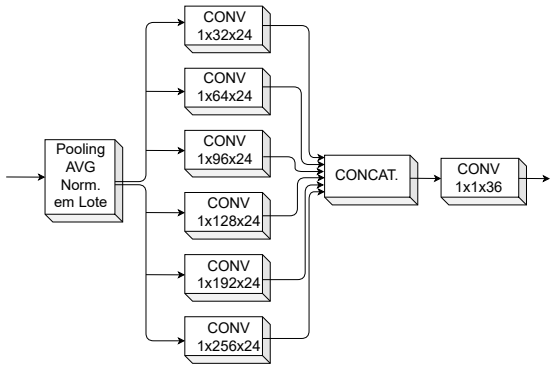

Fig. 5: Módulo Multifiltros. Sua característica principal são as convoluções em paralelo com diferentes dimensões de filtros. Todas as funções de ativação são ELU. Adaptado de [13].

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 [13], Fernandes Júnior [11] e também em diversos outros trabalhos.

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 $M \in \{\frac{1}{2}, 1, 2, 3\}$. Estes são os chamados "erros de oitava".

D. Aumento de Dados

O Aumento de Dados (ou *Data Augmentation*) é a técnica para aumentar artificialmente a quantidade de exemplos do conjunto de treinamento, gerando variantes realistas de cada exemplo. Isto reduz o *overfitting* o que a torna uma técnica de regularização [21].

Quando o problema envolve imagens é comum utilizar técnicas como rotação, deslocamento, redimensionamento, entre outras. Quando envolve alterações de iluminação, pode-se utilizar técnicas de variação de contraste ou variações de cores. Porém estas técnicas não se aplicam ao escalograma, pois ele deve ser gerado de um arquivo de áudio. Aplicar alguma destas técnicas iria distorcer a informação contida no escalograma.

Para realizar o aumento de dados com arquivos de áudio, pode-se utilizar técnicas de mudanças de frequências, mudança de velocidade, inserção de ruído, porém algumas delas não fazem sentido para o problema da estimativa de andamento musical. Por isso, inspirado pelo trabalho de Schreiber e Müller [13], optou-se por fazer compressões e expansões do escalograma ao longo do eixo horizontal, mantendo o eixo vertical sem alteração e ajustando o valor de anotação do andamento após a modificação. Ao expandir ou comprimir um escalograma, a velocidade de execução está variando proporcionalmente. Expandindo-o, a música fica mais lenta e o andamento diminui, enquanto que ao comprimir o escalograma, a música fica mais rápida e o andamento aumenta.

Este aumento de dados foi implementado de forma iterativa. A cada época os exemplos são alterados, criando novos exemplos e expondo a rede a uma quantidade muito maior de escalogramas. O método de compressão e expansão está exemplificado nos quatro escalogramas da Figura 6. Inicialmente é gerado um escalograma de toda a música, conforme Figura 6a utilizando o exemplo ACM MIRUM 169108.clip (124 BPM). Foi definido um fator de alteração (F_a) que é o valor utilizado para comprimir ou expandir o escalograma. O fator de alteração é escolhido aleatoriamente em um grupo de valores pré-definidos $F_a \in \{0.8, 0.85, 0.9, 0.95, 1, 1.05, 1.1, 1.15, 1.2\}$. Se $F_a = 1$, o escalograma permanece com magnitude horizontal inalterada e uma janela de 256 *pixels* por 40 *pixels*, com *offset* aleatório, é selecionada. O escalograma da Figura 6b mostra esta janela selecionada.

Para o caso de $F_a = 1.2$, toda a música sofre uma expansão, conforme o escalograma da Figura 6c. A classe do exemplo ACM MIRUM 169108.clip passa a ser 103 BPM. Da mesma forma, uma janela de 256 *pixels* por 40 *pixels*, com *offset* aleatório, é selecionada conforme apresentado na Figura 6d.

E. Previsão por Janelas Aleatórias

Ao estimar um valor de andamento em uma janela com 256 *pixels* horizontais, o modelo está prevendo um valor de andamento apenas para aquele intervalo de tempo específico, de 11,888 s. A maioria das músicas dos bancos de dados utilizados possuem um andamento constante ao logo de todo o sinal de áudio, porém em alguns momentos podem ocorrer variações. Caso a janela se posicione exatamente em uma destas variações, a previsão pode ser errada.

Para estimar o andamento global da música e evitar erros pelo posicionamento do local observado, utilizou-se uma estratégia de janelas aleatórias ao longo da música. Esta estratégia consiste em fazer previsões de 30 janelas escolhidas aleatoriamente em uma peça musical. Após a escolha aleatória das 30 janelas e todas as previsões, o modelo escolhe o valor de andamento com maior ocorrência e o considera como a estimativa de andamento global.

Após a definição do melhor modelo, a partir dos experimentos de variação dos parâmetros de geração dos escalogramas, foi realizado um novo treinamento com aplicação do *early stopping* com o objetivo de interromper o treinamento com um valor de acurácia 2 próximo ao valor máximo atingido pelo modelo ao longo de 50 épocas. A partir daí, foi realizada a previsão por janelas aleatórias nos bancos de dados de avaliação e os resultados foram comparados com o estado da arte.

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,

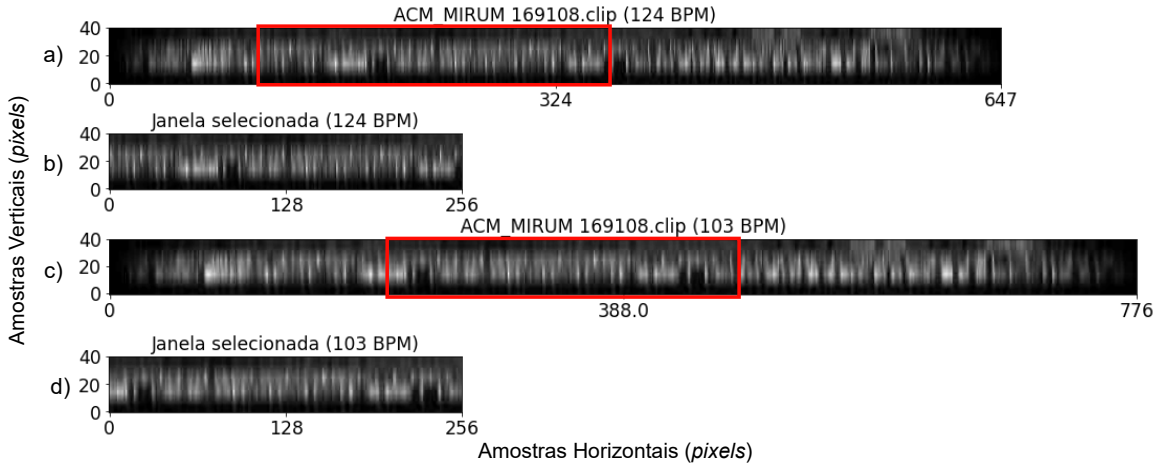

Fig. 6: Escalogramas do exemplo ACM MIRUM 169108.clip (124 BPM). A figura (a) mostra a música completa e uma janela escolhida aleatoriamente para o treinamento da rede. A figura (b) mostra o detalhe da janela escolhida no escalograma de 124 BPM. A figura (c) mostra a música completa expandida com $F_a = 1.2$ e uma janela escolhida aleatoriamente. A figura (d) mostra o detalhe da janela escolhida no escalograma de 103 BPM.

o resultado mais significativo é o da Acurácia 2 que considera os erros de oitava.

Os resultados obtidos nos experimentos realizados foram bastante próximos entre si, tanto em comportamento das curvas de treinamento e validação, como valores finais e máximos dos modelos. Por isso, a análise para definição do melhor modelo foi minuciosa, observando a média e desvio padrão dos principais indicadores. A Tabela II mostra os resultados do valor máximo de validação atingido durante o treinamento. Este indicador é importante, pois através dele pode-se retreinar a rede utilizando a técnica de *early stopping*, interrompendo o treinamento próximo ao valor máximo. Desta forma o modelo irá atingir melhores resultados nos conjuntos de avaliação. O melhor resultado foi conseguido com a wavelet Morlet, utilizando 40 níveis de escala (Morl_40). A Figura 7 mostra as curvas de treinamento e validação para os experimentos com melhor resultado (Morl_40) e pior resultado (Morl_6), ao longo das 50 épocas. É possível observar um comportamento similar, porém com o experimento Morl_6 oscilando mais e atingindo uma acurácia 2 final inferior.

TABELA II: Comparativo entre os valores máximos de Acurácia 2 no conjunto de validação, ao longo do treinamento

Experimento	Escalograma	Resultado Final (%)
1	Mexh_4	90,8 ± 0,7
2	Mexh_6	90,8 ± 0,5
3	Mexh_40	90,3 ± 0,7
4	Morl_4	90,8 ± 0,5
5	Morl_6	90,0 ± 0,8
6	Morl_40	**91,0 ± 0,6**
7	Shan_4	90,7 ± 0,4
8	Shan_6	90,5 ± 1,0
9	Shan_40	90,6 ± 0,3

Na Figura 8 é apresentando um gráfico *box plot*, onde estão sintetizados os resultados (*k-fold*) de validação e teste ao final de 50 épocas. Cada experimento está representado em cores diferentes e as colunas estão em ordem, primeiramente o conjunto de validação e posteriormente o conjunto de teste. Analisando este gráfico é possível afirmar que a wavelet Morlet, utilizando 40 níveis de escala (Morl_40), atingiu melhores resultados para o conjunto de validação. Enquanto a wavelet Shannon, também utilizando 40 níveis de escala (Shan_40), obteve melhores resultados no conjunto de teste.

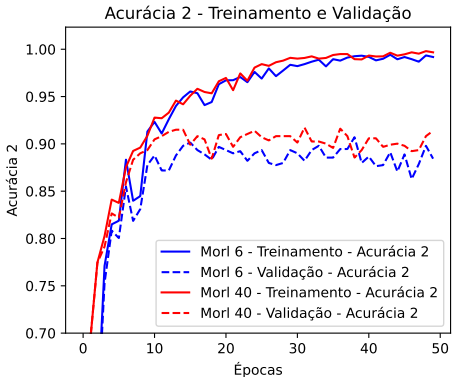

Fig. 7: Curvas de treinamento e validação para os experimentos com melhor resultado (Morl_40) e pior resultado (Morl_6).

Estes resultados levam a concluir que o modelo se comportou melhor com os escalogramas gerados a partir de 40 níveis de escala. Isto era previsto, pois estes são os escalogramas que apresentam melhor resolução, e consequentemente mais informação. A wavelet analisadora Chapéu Mexicano foi a que apresentou menos variações de intensidade, e isto pode ter relação com os resultados do experimento Mexh_40, que foram um pouco abaixo que os demais com 40 níveis de escala. É interessante notar que os resultados mais baixos foram com escalogramas gerados a partir de 6 níveis de escala.

Em 2020, Schreiber *et al.* [4] apontaram os trabalhos de

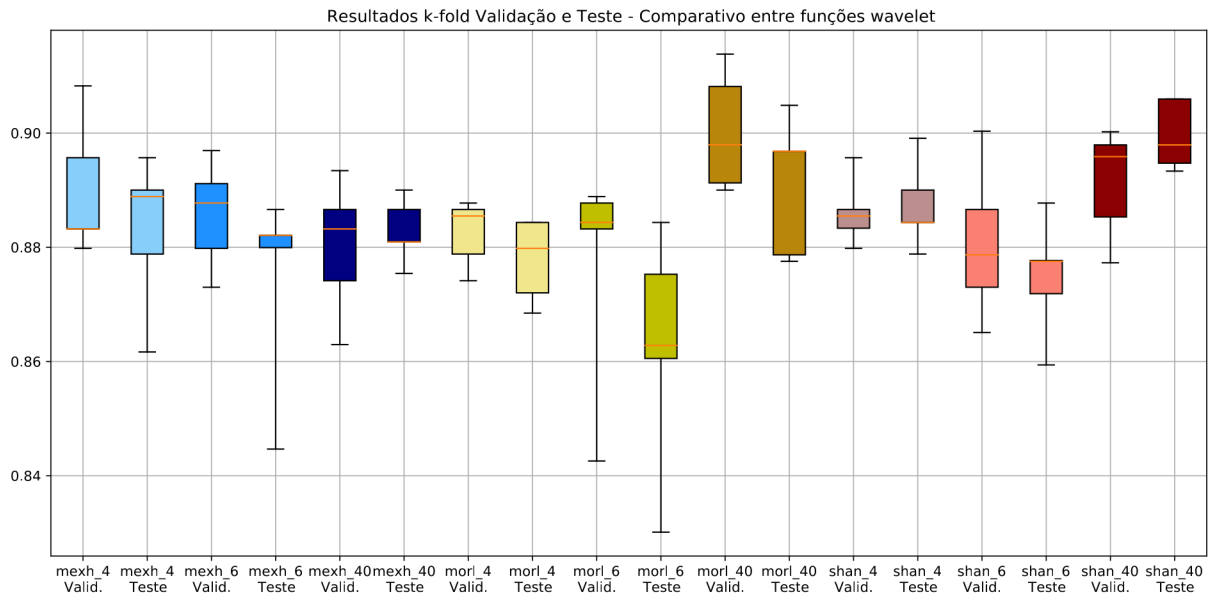

Fig. 8: *Box plot* comparativo entre os resultados do treinamento k-*fold* para os conjuntos de dados de validação e teste, para todos os experimentos.

Schreiber e Müller [13] e Böck *et al.*[1] como os principais trabalhos de estimativa de andamento musical. Até o momento, não há registros de um trabalho que tenha superado a performance destes algoritmos em todos os bancos de dados estudados. Por isso, estes trabalhos foram escolhidos para serem comparados com os modelos propostos. Além disso, foi realizada uma comparação com o trabalho publicado anteriormente por Viana *et al.* [18] para observar a evolução das novas técnicas estudadas.

Foram considerados dois modelos com métodos de geração do escalograma diferentes, Morl_40 e Shan_40. Ambos os modelos foram treinados com todos os bancos de dados de treinamento, e validados com todos os bancos de dados de avaliação combinados. Para o Morl_40, foi utilizado um *early stopping* para interromper o treinamento quando a acurácia 2 para o conjunto de validação fosse maior que 87,5%, enquanto que para o Shan_40 o treinamento seria interrompido quando a acurácia 2 para o conjunto de validação fosse maior que 88,0%. Para ambos os modelos o limite máximo de épocas foi 50.

No caso do modelo Morl_40 o treinamento durou as 50 épocas e a acurácia 2 para o conjunto de validação não ultrapassou 87,5%. Para o modelo Shan_40, o treinamento foi interrompido na 31ª época, com 88,0%. É importante enfatizar que durante o treinamento o método de previsão por janelas aleatórias não é utilizado.

A Tabela III mostra os resultados para acurácia 2 dos modelos Morl_40 e Shan_40 comparados ao estado da arte, respectivamente. Pode-se observar que os modelos propostos atingiram bons resultados com o banco de dados GiantSteps chegando a um resultado de 92,6% com o Morl_40 e 92,3% com o Shan_40, superando o estado da arte. Para o banco de dados "Combinados" o modelo Shan_40 conseguiu resultados próximos ao estado da arte, com uma acurácia 2 de 90,1%. Para o banco SMC Mirum, os resultados foram consideravelmente inferiores ao estado da arte. O SMC Mirum é composto por música clássica, estilo com poucos instrumentos percussivos e pouco representado nos bancos de dados de treinamento, o que justifica o resultado ruim.

O GiantSteps é composto por músicas eletrônicas, o que permite que o andamento da música seja rigorosamente constante. A maior dificuldade que um modelo encontra ao estimar um andamento com este estilo são os efeitos sonoros inseridos nestas peças musicais que pode interferir na percepção do andamento. O escalograma se mostrou eficaz em abstrair estes efeitos e permitir que o modelo classificasse corretamente o andamento.

Pode-se evidenciar a evolução nos resultados atingidos pelos experimentos Morl_40 e Shan_40 quando comparados aos resultados de Viana *et al.* [18]. Isto mostra a melhoria conseguida com a variação dos parâmetros de geração do escalograma e também das técnicas de aumento de dados e de previsão por janelas aleatórias.

TABELA III: Comparação com o estado da arte - Acurácia 2 (%)

Bancos de Dados de Avaliação	**Schr**	**Böck**	**Vian**	**Morl_40**	**Shan_40**
ACM Mirum	97,4	97,7	87,9	95,5	96,7
Ballroom	98,4	98,7	90,7	93,6	93,8
GiantSteps	89,3	86,4	82,9	**92,6**	**92,3**
GTzan	92,6	95,0	78,9	88,6	90,4
Hainsworth	84,2	89,2	72,9	83,7	81,4
ISMIR04	92,2	95,0	74,2	85,2	87,3
SMC	50,2	67,3	29,5	45,2	44,2
Combinados	92,1	93,6	80,1	89,6	90,1

IV. Conclusões e Trabalhos Futuros

Foi possível observar que os nove experimentos realizados com os diferentes tipos de escalograma se comportaram de maneira similar, apresentando resultados próximos. Ao se analisar o resultado final dos conjuntos de validação e teste para os experimentos, chegou-se a conclusão que os modelos treinados pelos escalogramas Morl_40 e Shan_40 apresentaram os melhores resultados.

Para a avaliação final do modelo e comparação com o estado da arte, foi aplicada a técnica de aumento de dados e a previsão por janelas aleatórias. Estas técnicas em conjunto aumentaram significativamente a acurácia 2 no conjunto de dados "Combinados", aproximadamente 10%. O modelo proposto Shan_40 atingiu uma acurácia 2 de 90,1% para o banco de dados "Combinados", se aproximando do estado da arte que é 93,6%. Destaca-se que ambos os modelos propostos superaram o estado da arte para o banco de dados "GiantSteps", e o modelo Morl_40 atingiu o melhor resultado para acurácia 2, 92,6%.

Por fim, pode-se concluir que os modelos propostos foram capazes de estimar o andamento musical com resultados similares ao estado da arte. A utilização dos escalogramas wavelet como representação dos sinais de áudio se mostrou vantajosa visto que é possível realizar ajustes nos parâmetros de geração de forma a investigar uma representação que se comporte melhor para um determinado tipo de problema. Porém, para o problema da estimativa de andamento musical, os escalogramas não apresentaram melhorias significativas quando comparado aos espectrogramas, que são a forma de representação mais utilizada. A técnica de aumento artificial de dados proposta se mostrou eficaz, comprovado pelos resultados da variação da quantidade de exemplos que eram submetidos a esta alteração. Da mesma forma, a técnica de janelas aleatórias para estimativa de andamento global se mostrou importante para melhoria dos resultados finais do modelo proposto, para a acurácia 2.

Algumas ideias para melhoria dos resultados para o problema de estimativa de andamento musical foram visualizadas. A principal delas é o desenvolvimento de um "decisor de oitavas" que poderia ser uma etapa final do modelo proposto ou incorporado à arquitetura da rede. Sua função seria decidir se a previsão do modelo foi feita na oitava certa ou não. Desta forma, seria possível observar melhorias dos resultados de acurácia 0 e acurácia 1.

Outra possibilidade de melhoria é investigar mais escalogramas gerados à partir de outros parâmetros. Percebeu-se que os escalogramas com mais níveis de escala apresentaram melhores resultados, por isso seria uma alternativa aumentar a quantidade de escalas dos escalogramas que atingiram os melhores desempenhos. Para isto, seria necessário alterar a camada de entrada da rede de forma a se adequar ao novo escalograma gerado. A arquitetura da rede também pode ser modificada e estruturas com mais camadas e mais conexões devem ser estudadas.

Agradecimentos

Os autores agradecem ao CNPq pelo apoio financeiro.

Referências

[1] 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.

[2] Masataka Goto and Yoichi Muraoka. *A Beat Tracking System for Acoustic Signals of Music*. In *Proceedings of the Second ACM International Conference on Multimedia*, volume 0, pages 365–372, San Francisco, CA, USA, 1994.

[3] Eric D. Scheirer. *Tempo and Beat Analysis of Acoustic Musical Signals*. *The Journal of the Acoustical Society of America*, 103(1): 588-601, 1998. doi: 10.1121/1.421129. URL: `https://doi.org/10.1121/1.421129`.

[4] 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.

[5] 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.

[6] International Society for Music Information Retrieval (ISMIR). *Copyright © ISMIR*. Disponível em: `ismir.net`. Acesso em: 20 dezembro, 2022.

[7] Jose R Zapata and Emilia Gómez. *Comparative Evaluation and Combination of Audio Tempo Estimation Approaches*. AES 42nd International Conference, Ilmenau, Germany, 2011.

[8] Sebastian Bock and Marcus Schedl. *Enhanced Beat Tracking With Context-Aware Neural Networks*. Proc. of the 14th Int. Conference on Digital Audio Effects (DAFx-11), Paris, France, 2011.

[9] 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.

[10] Kamalesh Palanisamy, Dipika Singhania, and Angela Yao. *Rethinking CNN Models for Audio Classification*. Department of Instrumentation and Control Engineering, National Institute of Technology, Tiruchirappalli, India, 2020.

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

[12] 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*.

[13] 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.

[14] Mila Souza, Pedro Moura, and Jean-Pierre Briot. *Tempo Estimation Via Neural Networks - A Comparative Analysis*. Em *Anais do XVIII Simpósio Brasileiro de Computação Musical*, Recife, 2021. Páginas: 17-24. Editora: SBC. Endereço: Porto Alegre, RS, Brasil. DOI: 10.5753/sbcm.2021.19420.

[15] Xiaoheng Sun, Qiqi He, Yongwei Gao, and Wei Li. *Musical Tempo Estimation Using a Multi-scale Network*. DOI: 10.48550/ARXIV.2109.01607. Ano: 2021.

[16] Elio Quinton. *Equivariant Self-Supervision for Musical Tempo Estimation*. DOI: 10.48550/ARXIV.2209.01478. Ano: 2022.

[17] Giovana Morais, Matthew E. P. Davies, Marcelo Queiroz, and Magdalena Fuentes, *Tempo vs. Pitch: Understanding Self-Supervised Tempo Estimation*, in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1-5, doi: 10.1109/ICASSP49357.2023.10095292.

[18] Luiz Alberto Viana, Eduardo Simas Filho, e Antonio Carlos Lopes Fernandes Junior. *Escalogramas Wavelet Aplicados à Estimativa de Andamento Musical*. XL Simpósio Brasileiro de Telecomunicações e Processamento de Sinais (SBrT2022), 2022. DOI: 10.14209/sbrt.2022.1570824871.

[19] Hendrik Schreiber and Meinard Müller. *A Post-Processing Procedure for Improving Music Tempo Estimates Using Supervised Learning*. Outubro de 2017. DOI: 10.5281/zenodo.1415045.

[20] 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.

[21] 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.