

Estimação de Múltiplos Pitches em Áudio Musical Polifônico Utilizando Rede Neural Convolutacional

Marcus F. S. da Silva¹, Waldir S. S. Júnior¹, Celso B. Carvalho¹, Frederico da S. Pinagê¹, Gabriel M. Araújo², Luiz C. G. Junior¹.

¹Universidade Federal do Amazonas e Centro de P&D em Elet. e Tec. Inf. (UFAM/CETELI), AM-Brasil Inf. (UFAM/CETELI)

²Centro Federal de Educação Tecnológica Celso Suckow da Fonseca (CEFET-RJ)

Emails: fabio.marcus@gmail.com, {ccarvalho,waldirjr}@ufam.edu.br, fredericopinage@ufam.edu.br, gabriel.araujo@cefet-rj.br, luiz.junior@ufam.edu.br

Resumo—Este trabalho propõe o CREME (*Convolutional Representation for Multipitch Estimation*), extensão do modelo CREPE para estimação de múltiplos *pitches* em sinais musicais polifônicos. A arquitetura é modificada pela inserção de uma camada densa de 4096 unidades e expansão da saída para 720 neurônios, cobrindo seis oitavas com resolução de 10 cents/bin. O treinamento emprega *transfer learning* a partir dos pesos do CREPE e utiliza a base *MDB-stem-synth-multi*, construída pela combinação de *stems* sintéticos com até duas frequências simultâneas por *frame*. Os resultados indicam RPA médio de 0,93 na validação polifônica e RPA > 0,90 na base *Bach10-mf0-synth*, com preservação do desempenho monofônico e variação média de $\approx 0,025$ RPA em relação ao CREPE original. O modelo demonstra potencial para aplicações em transcrição automática e análise de performances musicais polifônicas.

Palavras-Chave—Estimação de *pitch*, múltiplos *pitches*, redes neurais convolucionais, recuperação de informação musical, transcrição musical automática.

Abstract—This paper proposes CREME (*Convolutional Representation for Multipitch Estimation*), an extension of the CREPE model for estimating multiple pitches in polyphonic musical signals. The architecture is modified by adding a dense layer with 4096 units and expanding the output to 720 neurons covering six octaves at 10 cents/bin resolution. Training employs transfer learning from CREPE weights using the *MDB-stem-synth-multi* dataset, constructed by combining synthetic instrumental stems with up to two simultaneous frequencies per frame. Results show a mean RPA of 0.93 on polyphonic validation and RPA > 0.90 on the *Bach10-mf0-synth* dataset, while fully preserving monophonic performance with an average difference of only ≈ 0.025 RPA relative to the original CREPE. The model demonstrates potential for automatic music transcription and polyphonic performance analysis.

Keywords—Pitch estimation, multiple pitch estimation, convolutional neural networks, music information retrieval, automatic music transcription.

I. INTRODUÇÃO

A recuperação de informação musical (MIR, do inglês *Music Information Retrieval*) tem crescido devido à expansão de plataformas digitais de música como Spotify, Apple Music e YouTube Music, e à necessidade de técnicas eficientes para análise automática de sinais musicais [1], [2], [3] e [4]. Entre os principais problemas investigados nessa área, destaca-se a estimação de *pitch*, responsável pela identificação da frequência fundamental (F0) de notas musicais. Essa capacidade permite que pesquisadores transcrevam e estudem

extensos corpora musicais, descobrindo padrões e tendências na composição musical e nas práticas de performance em diferentes gêneros e períodos históricos.

Apesar dos avanços obtidos com métodos clássicos de processamento digital de sinais, a estimação de múltiplos *pitches* ainda representa um grande desafio devido às complexidades inerentes ao áudio musical, como a sobreposição de harmônicos de fontes distintas, presença de ruído e variações de timbres [4]. Nesse contexto, abordagens baseadas em aprendizado profundo, especialmente redes neurais convolucionais (CNNs), têm demonstrado desempenho superior em tarefas de estimação de *pitch*.

Dentre essas abordagens, é utilizado um modelo de rede neural convolutacional bem estabelecido para estimativa de *pitch* monofônica, denominado CREPE (do inglês, *Convolutional Representation for Pitch Estimation*) [5]. Neste trabalho, o modelo inicialmente desenvolvido para estimação monofônica terá sua arquitetura adaptada para estimar até dois *pitches* simultâneos no contexto de música polifônica. Como contribuição adicional, é construída a base de dados *MDB-stem-synth-multi*, gerada pela combinação de *stems* instrumentais da base *MDB-stem-synth* [6], contendo anotações precisas de múltiplos *pitches* por *frame*. Os resultados demonstram desempenho robusto em diversas situações, incluindo a capacidade de identificar corretamente *pitches* simultâneos em material não visto no treinamento, evidenciando o potencial da abordagem para aplicações em recuperação de informação musical e transcrição automática.

II. TRABALHOS RELACIONADOS

Os pesquisadores McLeod e Wyvill [7] avaliaram o desempenho dos algoritmos *YIN* e *pYIN* em diferentes instrumentos e idiomas, evidenciando sua eficiência na estimação de *pitch*. O algoritmo *YIN*, proposto por Alain de Cheveigné e Hideki Kawahara [8], realiza estimação diretamente no domínio do tempo utilizando funções de diferença cumulativa normalizada para reduzir ambiguidades associadas à autocorrelação e melhorar a identificação da frequência fundamental. Posteriormente, Matthias Mauch e Simon Dixon [9] propuseram o *pYIN*, uma extensão probabilística baseada em Modelos Ocultos de Markov (HMM) e no algoritmo de *Viterbi*, tornando a detecção de *pitch* mais robusta em sinais ruidosos e musicalmente complexos.

Além disso, as Redes Neurais Convolucionais (CNNs), descritas por *Goodfellow, Bengio e Courville* [10], representam uma das principais abordagens de aprendizado profundo para processamento de dados estruturados, sendo amplamente aplicadas em tarefas de reconhecimento de padrões em imagens, áudio e sinais temporais. Assim, no contexto de recuperação de informação musical (*MIR*), *Humphrey, Bello e LeCun* [11] destacaram que modelos baseados em aprendizado profundo apresentam maior capacidade de aprendizado automático de características relevantes quando comparados aos métodos clássicos de processamento digital de sinais (*DSP*), reduzindo a necessidade de extração manual de atributos e aumentando a robustez em diferentes cenários musicais.

Nesse contexto, *Kim, Salamon, Li e Bello* [5] propuseram o modelo *CREPE* (*Convolutional Representation for Pitch Estimation*), uma rede neural convolucional desenvolvida especificamente para a estimação da frequência fundamental (F_0) em sinais de áudio. Diferentemente dos métodos tradicionais, o *CREPE* opera diretamente no domínio do tempo, utilizando como entrada segmentos de áudio bruto amostrados em 16 kHz. A arquitetura do modelo é composta por múltiplas camadas convolucionais capazes de aprender representações hierárquicas do sinal de áudio, permitindo identificar padrões harmônicos relevantes para a detecção de *pitch*.

O modelo utiliza uma representação em escala logarítmica baseada em *cents*, na qual cada neurônio da camada de saída corresponde a uma faixa específica de frequência. A estimação final do *pitch* é obtida por meio da média ponderada das ativações da rede, possibilitando maior precisão na identificação da frequência fundamental. Além disso, o treinamento do *CREPE* utiliza técnicas como normalização em lote (*batch normalization*), *dropout* e otimização baseada no algoritmo *ADAM*, contribuindo para maior estabilidade e capacidade de generalização do modelo.

Em comparação aos métodos clássicos, como *YIN* e *pYIN*, o *CREPE* apresenta vantagens relacionadas à precisão, robustez a ruídos e capacidade de adaptação a diferentes instrumentos e condições acústicas. Dessa forma, o modelo tornou-se uma das principais referências para estimação monofônica de *pitch* baseada em aprendizado profundo.

III. METODOLOGIA

A. Bases de Dados

Os experimentos utilizam as seguintes bases: (i) *MDB-stem-synth* [6]: 230 *stems* instrumentais sintéticos com anotações perfeitas de F_0 , usada para *transfer learning*; (ii) *MDB-stem-synth-multi* (contribuição deste trabalho): base polifônica construída combinando pares de *stems* da mesma música, resultando em 453 arquivos com até duas frequências simultâneas por *frame* e anotações precisas, usada para treino e validação do CREME; (iii) *Traditional Flute Dataset* [12]: 30 fragmentos de flauta solo (2.245 eventos, C4–D7) com anotações manuais, usada para avaliar preservação monofônica; (iv) *Bach10-mf0-synth* [13]: 10 peças de Bach resintetizadas para quatro instrumentos (fagote, clarinete, saxofone, violino),

combinadas em pares para gerar 60 arquivos polifônicos de teste; (v) *RWC Music Database* [14]: gravações reais de violão e clarinete, usada para avaliar limites de resolução em *pitches* próximos.

O treinamento foi realizado exclusivamente com a base *MDB-stem-synth-multi*. Os testes com as demais bases foram feitos diretamente com os pesos resultantes do treinamento, sem *fine-tuning* adicional.

B. Modelo Base: CREPE

O CREPE [5] é uma rede neural convolucional estado da arte para estimação de *pitch* monofônica. Recebe janelas temporais (*frames*) de 1024 amostras de áudio bruto a 16 kHz (equivalentes a 64 ms de sinal) e produz um vetor de probabilidade de 360 componentes, cobrindo seis oitavas. A frequência estimada é calculada por média ponderada dos *bins* próximos ao pico. A avaliação utiliza a métrica RPA (*Raw Pitch Accuracy*) [15], que mede a fração de instâncias em que a frequência estimada está dentro de 50 *cents* da referência.

C. CREME: Extensão para Estimação Polifônica

Com base em estudos preliminares com o CREPE e o FCN [16], que confirmaram a robustez do CREPE em material musical inédito e o potencial de *transfer learning* com dados de domínio adequado, foi desenvolvido o CREME (*Convolutional Representation for Multipitch Estimation*) [17]. O modelo estende o CREPE para estimar no máximo duas frequências simultâneas por *frame*. A Figura 1 apresenta a arquitetura proposta.

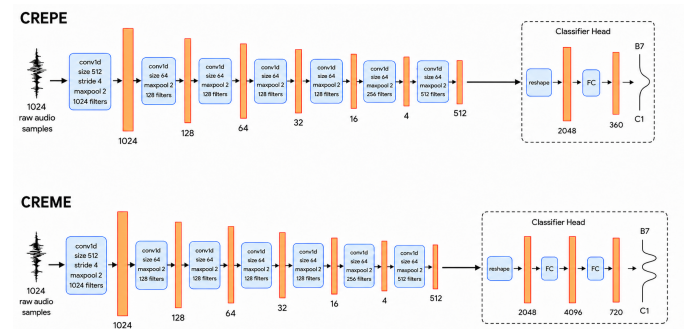


Fig. 1. Arquitetura do modelo CREME. A entrada de 1024 amostras corresponde a uma janela temporal de áudio bruto a 16 kHz.

A camada de entrada (1024 amostras) e as seis camadas convolucionais do CREPE foram mantidas, preservando a analogia estrutural com a FFT. As modificações foram: (i) inserção de uma camada densa de 4096 unidades, para capturar padrões hierárquicos polifônicos; e (ii) expansão da camada de saída de 360 para 720 neurônios (6 oitavas $\times$ 12 semitons $\times$ 10 *bins*/semitom, resolução de 10 *cents/bin*), aumentando a granularidade espectral para discriminar frequências próximas. Os rótulos de treinamento são gerados como soma de gaussianas independentes centradas em cada frequência presente no *frame*, generalizando a estratégia unimodal do CREPE para o

caso multimodal. A função de perda empregada é a entropia cruzada binária (BCE).

O treinamento emprega *transfer learning*: os pesos pré-treinados do CREPE (obtidos com bases de áudio musical diversas) são utilizados como inicialização, e o modelo é então treinado com a *MDB-stem-synth-multi*.

IV. PROCEDIMENTOS EXPERIMENTAIS

A. Configuração Experimental

Os treinamentos foram executados no *Google Colab* com GPUs NVIDIA A100/L4 (TensorFlow 2.10). Aplicaram-se técnicas de aumento de dados: deslocamento de *pitch* (Py-rubberband), adição de ruído em intensidade aleatória e randomização de *frames* entre arquivos distintos. A base *MDB-stem-synth-multi* foi organizada em sete grupos de 60 arquivos para treino, com alternância cíclica (*round robin*), e um grupo de 33 arquivos para validação. O CREME foi treinado por 500 épocas com *batches* de 32 amostras, taxa de aprendizado 0,0002 e otimizador Adam.

B. Resultados

Validação com *MDB-stem-synth-multi*: A Tabela I sintetiza o desempenho do CREME no conjunto de validação. Com RPA médio de $0,9341 \pm 0,0398$, o modelo demonstrou precisão consistente na estimação simultânea de duas frequências. A análise por indicadores de erro revelou que o *miss error* (frequências presentes não detectadas) é o tipo de falha dominante, enquanto alarmes falsos e erros de substituição permanecem baixos.

TABELA I

DESEMPENHO DO CREME NO CONJUNTO DE VALIDAÇÃO.

Indicador	Valor
RPA mínimo	0,848
RPA médio	$0,934 \pm 0,040$
RPA máximo	0,992
Erro dominante	<i>Miss error</i>

Preservação da Capacidade Monofônica: A Figura 2 compara o RPA do CREPE e do CREME para os 30 fragmentos do *Traditional Flute Dataset*. A variação média foi de $\approx 0,025$ RPA, evidenciando que o treinamento polifônico não prejudicou a estimação de um único *pitch*: o CREME manteve desempenho equivalente ao CREPE em cenário monofônico.

Generalização Polifônica (*Bach10-mf0-synth*): Para avaliar a generalização, utilizou-se a base *Bach10-mf0-synth* [13], com 60 arquivos sintéticos gerados pela combinação de pares instrumentais (fagote, clarinete, saxofone e violino). Esses dados não foram vistos no treinamento. O CREME superou $RPA > 0,90$ na maioria dos arquivos testados, confirmando boa capacidade de generalização a material polifônico inédito.

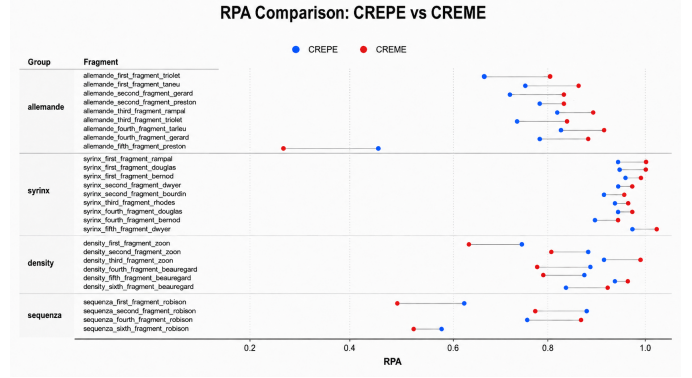


Fig. 2. Comparação de RPA: CREPE vs. CREME na base *Traditional Flute*.

TABELA II

LIMITES DE RESOLUÇÃO DO CREME NA BASE RWC.

Cenário	RPA (faixa)	Observação
Clarinete (D3 ref.)	0,49–0,99	Crescente com distância
Violão (E2 ref.)	0,22–0,82	Irregular; melhor em 5ª justa
Violão + Clarinete	$\approx 0,50$	Limitado pela complexidade

Limites de Resolução (RWC): Com gravações reais da base RWC [14], avaliou-se a capacidade de distinguir notas próximas utilizando os pesos treinados na *MDB-stem-synth-multi*, sem *fine-tuning*. A Tabela II sintetiza os resultados para os três cenários testados.

No clarinete, o RPA cresceu consistentemente de $\approx 0,49$ (1 semitom) a 0,99 (intervalos maiores), seguindo padrão esperado. No violão, os resultados foram irregulares: o melhor desempenho ocorreu no intervalo de quinta justa E2/B2 ($RPA = 0,82$), altamente consonante. A combinação de violão e clarinete resultou em $RPA \approx 0,50$, limitada pela complexidade acústica do violão (ressonâncias do corpo e vibração espontânea de cordas adjacentes).

V. CONCLUSÕES

Este trabalho apresentou o CREME, extensão do CREPE para estimação de até dois *pitches* simultâneos em sinais musicais polifônicos. O modelo alcançou RPA médio de 0,93 em validação polifônica e $RPA > 0,90$ em material não visto (*Bach10-mf0-synth*), demonstrando boa generalização. A capacidade monofônica foi integralmente preservada, com diferença média de $\approx 0,025$ RPA em relação ao CREPE na base *Traditional Flute Dataset*, confirmando que a extensão polifônica não comprometeu o desempenho em casos mais simples. O CREME tem potencial para aplicações em transcrição musical automática, reconhecimento de acordes e análise de performances polifônicas.

As principais limitações identificadas foram: (i) predominância do *miss error*, indicando tendência do modelo a subestimar frequências ativas; (ii) queda expressiva em intervalos dissonantes e combinações artificiais sem correlação harmônica; e (iii) dificuldades com instrumentos acusticamente complexos, como o violão. Como trabalhos futuros, propõe-se ampliar o modelo para mais de duas frequências simultâneas, diversificar

o conjunto de treinamento com combinações dissonantes e interferências acústicas reais, e avaliar o CREME em cenários com maior densidade polifônica.

REFERÊNCIAS

- [1] A. Chopra, A. Roy, and D. Herremans, “Mirflex: Music information retrieval feature library for extraction,” *arXiv preprint arXiv:2411.00469*, 2024, biblioteca modular para extração de características musicais usadas em recomendação, classificação e recuperação de informação musical.
- [2] Harvard Business School Digital Initiative, “Spotify’s machine learning strategy,” 2021, descrição das estratégias de recomendação do Spotify usando filtragem colaborativa, processamento de linguagem natural e análise de áudio. [Online]. Available: <https://digital.hbs.edu/platform-digit/submission/spotify-machine-learning-strategy/>
- [3] Flourish Music Group, “How machine learning is transforming music marketing and trends analysis,” 2023, aborda o uso de machine learning em marketing musical, previsão de tendências e personalização publicitária. [Online]. Available: <https://flourishmusicgroup.com/how-machine-learning-is-transforming-music-marketing-and-trends-analysis/>
- [4] J. Pons, “Estimating pitch in polyphonic music,” *arXiv preprint arXiv:1901.05079*, 2019, apresenta técnicas de estimativa de pitch em música polifônica, fundamentais para transcrição musical automatizada.
- [5] J. W. Kim, J. Salamon, P. Li, and J. P. Bello, “CREPE: A convolutional representation for pitch estimation,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 161–165.
- [6] J. Salamon, E. Gomez, and D. P. W. Ellis, “An analysis/synthesis framework for automatic F0 annotation of multitrack datasets,” in *Proceedings of the 18th International Society for Music Information Retrieval Conference*, 2017.
- [7] A. McLeod and G. Wyvill, “Evaluation of the yin pitch tracking algorithm across languages and instruments,” *The Journal of the Acoustical Society of America*, vol. 143, no. 3, pp. EL202–EL208, 2018.
- [8] A. De Cheveigné and H. Kawahara, “Yin, a fundamental frequency estimator for speech and music,” *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002.
- [9] M. Mauch and S. Dixon, “pyin: A fundamental frequency estimator using probabilistic threshold distributions,” in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 659–663.
- [10] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016, available at <http://www.deeplearningbook.org>.
- [11] E. J. Humphrey, J. P. Bello, and Y. LeCun, “Deep learning for music information retrieval: Promises, challenges, and opportunities,” *IEEE Signal Processing Magazine*, vol. 30, no. 2, pp. 82–86, 2013.
- [12] J. P. B. Brum, “Traditional flute dataset for score alignment,” <https://www.kaggle.com/jbraga/traditional-flute-dataset>, 2018.
- [13] Z. Duan, Y. Han, B. Pardo, and C. Zhang, “Multiple fundamental frequency estimation by modeling spectral peaks and non-peak regions,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 8, pp. 2121–2133, 2010.
- [14] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, “RWC music database: Popular, classical and jazz music databases,” in *Proceedings of the 3rd International Society for Music Information Retrieval Conference*, 2002, pp. 287–288.
- [15] C. Raffel, B. McFee, E. J. Humphrey, J. Salamon, O. Nieto, D. Liang, and D. P. W. Ellis, “mir_eval: A transparent implementation of common MIR metrics,” in *Proceedings of the 15th International Society for Music Information Retrieval Conference*, 2014.
- [16] L. Ardaillon and A. Roebel, “Fully-convolutional network for pitch estimation of speech signals,” in *Proceedings of Interspeech 2019*, 2019.
- [17] M. F. S. d. Silva, “Estimação de múltiplos *Pitches* em Áudio musical polifônico utilizando rede neural convolucional,” Dissertação de Mestrado, Universidade Federal do Amazonas, Manaus, AM, Brasil, 2025. [Online]. Available: <https://tede.ufam.edu.br/handle/tede/11342>