TEMPO ESTIMATION USING COMBINED MEL-SPECTROGRAM AND MEL-SCALOGRAM INPUTS

Luiz A. G. Viana [1]   Antonio Carlos L. Fernandes Júnior [1]   Eduardo F. de Simas Filho [1]

[1]Programa de Pós-Graduação em Engenharias Elétrica e de Computação (PPGEEC), Departamento de Engenharia Elétrica e de Computação (DEEC), Universidade Federal da Bahia

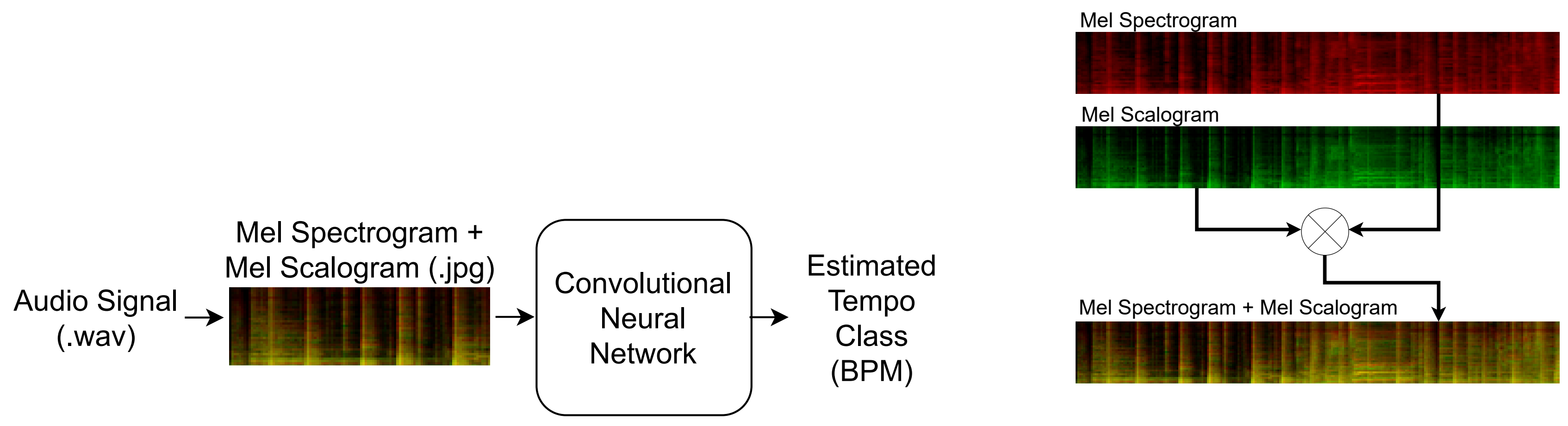
Introduction

- Musical tempo is the speed of a piece, measured in beats per minute (BPM).
- It is one of the fundamental tasks in Music Information Retrieval (MIR).
- The MIR community has been conducting researches on tempo estimation for the past 25 years.
- Although the works of [1] and [2] have achieved excellent results, the problem of musical tempo estimation is still open.
- This work proposes a new method:
  1. A novel three-dimensional audio representation combining mel-spectrograms and mel-scalograms.
  2. Training a convolutional neural network (CNN) with these representations.
  3. Validation using multiple datasets with robust cross-validation techniques.

Proposed Model

This study introduces a new approach for musical tempo estimation by creating a custom three-dimensional image from audio signals. The image combines Mel-Spectrogram and Mel-Scalogram.

The custom images are used to train a Convolutional Neural Network (CNN) for tempo estimation. Intuitively, the tempo estimation problem appears to be a regression problem for an integer value. However, this work treats it as a classification problem, assigning a class for each integer BPM value within a specific BPM range.


Audio Signal (.wav)
Mel Spectrogram + Mel Scalogram (.jpg)
Convolutional Neural Network
Estimated Tempo Class (BPM)
Mel Spectrogram
Mel Scalogram
Mel Spectrogram + Mel Scalogram

(a) Proposed model (b) Custom image

Figure 1. Simplified diagram of the proposed model.

Datasets:

- **Training:** LMD Tempo (3611 examples), MTG Tempo (1159 examples), Extended Ballroom (3826 examples).
- **Evaluation:** ACM Mirum (1410 examples), Ballroom (698), GiantSteps Tempo (660), GTZAN (999), Hainsworth (222), ISMIR2004 (465), SMC Mirum (217).

The training datasets cover diverse musical styles, while the evaluation datasets enable comparison with existing methods in the literature.

Audio Signal Representation

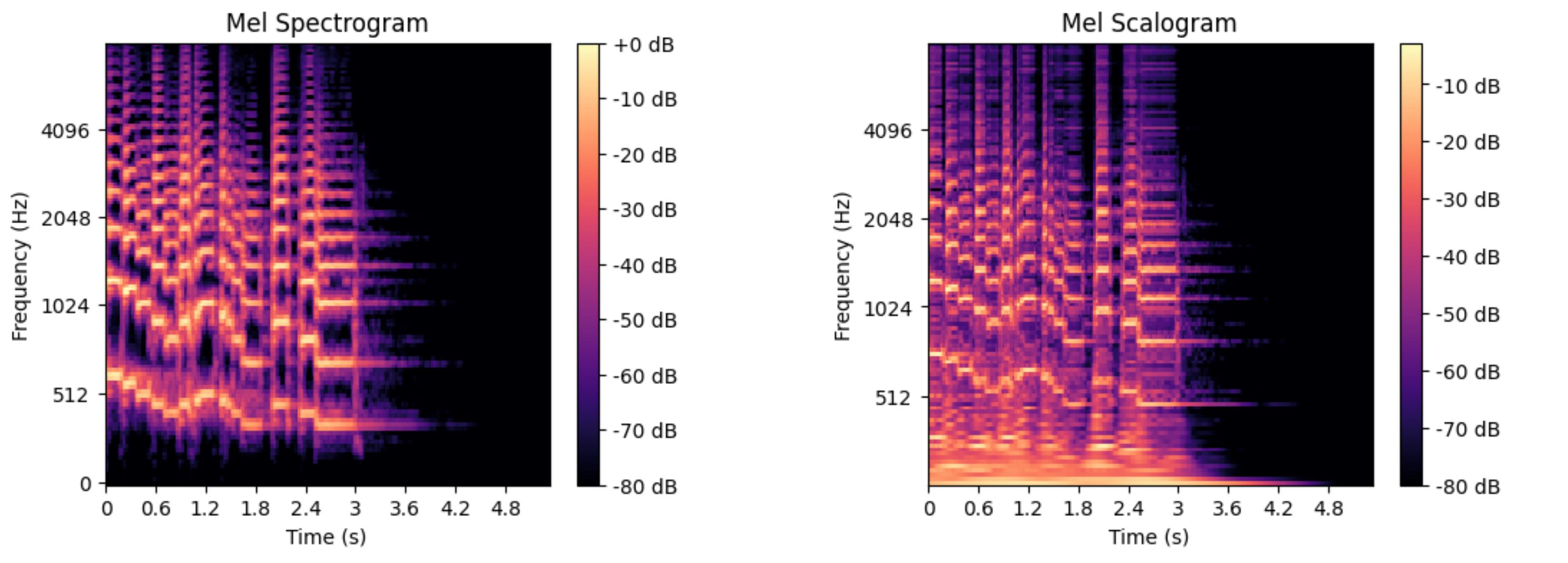
To prepare audio signals for the model, the following preprocessing steps were applied:

- Discarded the initial 5 seconds of audio to avoid tempo variations at the start.
- Converted signals to mono and downsampled to 11025 Hz, sufficient to detect tempos up to 646 BPM.
- Selected a duration of 11.888 seconds, resulting in an audio vector of 131072 samples.

The audio signal was represented as a pair of matrices:

- **Mel-Spectrogram**: Combines the Short-Time Fourier Transform (STFT) with a Mel scale conversion.
- **Mel-Scalogram**: Derived from the Continuous Wavelet Transform (CWT), using the complex Morlet wavelet.

These representations were combined into a single image of dimensions (256, 40, 2) for training the model. The figures below show examples of the mel-spectrogram and mel-scalogram:



(a) Mel-Spectrogram (b) Mel-Scalogram

Figure 2. Comparison between Mel-Spectrogram (a) and Mel-Scalogram (b) representations.

Continuous Wavelet Transform

Given a signal $f(t)$, its CWT is defined as follows:

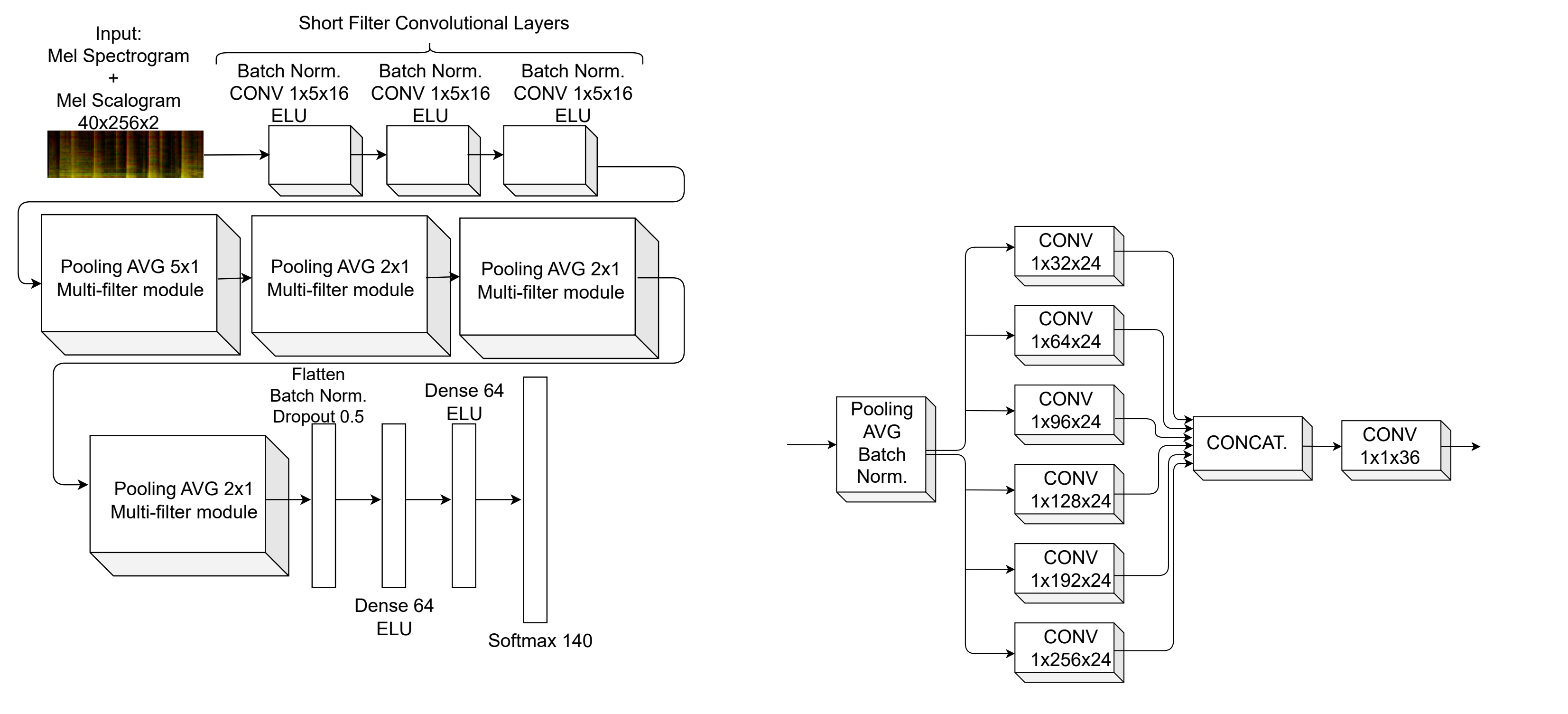
$$W_f^{\psi}(a,\tau) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} f(t)\psi^{*}\left(\frac{t-\tau}{a}\right) dt \qquad (1)$$

where the parameter a (>0) refers to the scale, and τ represents the translation or location of the mother wavelet function $\psi(t)$. Both a and $\tau \in \mathbb{R}$. The parameter a controls the dilation/contraction of the mother wavelet function. The superscript asterisk in $\psi^{*}(\cdot)$ denotes the complex conjugate of the function $\psi(\cdot)$, and $W_f^{\psi}(a,\tau)$ is known as the wavelet coefficient.

Convolutional Neural Network

The convolutional neural network (CNN) was designed to estimate tempo using the custom audio signal representation.

- **Input Layer:** Accepts images of size (256, 40, 2), with mel-spectrogram and mel-scalogram as channels.
- **Convolutional Layers:** Three layers with short filters (1x5) and Exponential Linear Unit (ELU) activations. **Multi-Filter Modules:** Parallel convolutions with various filter sizes to capture temporal patterns.
- **Output Layer:** A softmax layer with 140 classes representing BPM ranges (60-199 BPM).
- 5-fold cross-validation to ensure robust evaluation.
- Data augmentation by compressing and expanding custom images to simulate tempo variations.
- Evaluation metrics: Accuracy 0 ($\hat{\Gamma} = \Gamma$), Accuracy 1 ($\hat{\Gamma} = \Gamma \pm 4\%$), and Accuracy 2 ($\hat{\Gamma} = (\Gamma \pm 4\%)\, M$).



(a) CNN Architecture (b) Multi-Filter Module

Figure 3. (a) Diagram of the CNN architecture, adapted from [2]. (b) Multi-Filter Module with parallel convolutions of varying sizes.

Results

The proposed model was evaluated in two main aspects: the effectiveness of combining mel-spectrograms and mel-scalograms, and its performance compared to state-of-the-art methods.

Comparison of Representations: The table below shows the Accuracy 2 results for validation and test sets when using mel-spectrograms, mel-scalograms, and the combined representation.

Table 1. Accuracy 2 for Different Representations

Experiment	Validation (%)	Test (%)
Mel-Scalogram	93.5 ± 0.3	90.4 ± 1.1
Mel-Spectrogram	94.1 ± 0.4	91.2 ± 0.8
Mel-Spec+Scal	94.2 ± 0.6	92.0 ± 0.7

Comparison with State-of-the-Art: The performance of the proposed model was compared with state-of-the-art approaches. The table below presents Accuracy 2 results across multiple datasets.

Table 2. Comparison with the State-of-the-Art -Accuracy 2 (%)

Evaluation Datasets	Schr [2]	Böck [1]	Böck [3]	Mel-Spec+Scal
ACM Mirum	97,4	97,7	97,7	97,4
Ballroom	98,4	98,7	-	94,8
GiantSteps	89,3	86,4	95,8	93,4
GTZAN	92,6	95,0	93,9	92,2
Hainsworth	84,2	89,2	-	84,2
ISMIR04	92,2	95,0	-	89,2
SMC	50,2	67,3	-	52,5
Combined	92,1	93,6	-	91,4

- Combining mel-spectrogram and mel-scalogram, the model's performance improves further, achieving an accuracy of 94.2% on validation and 92.0% on the test set.
- The proposed model achieved competitive results with state-of-the-art methods, surpassing some in datasets like GiantSteps.

Conclusion

This study evaluated a new representation method for musical tempo estimation. Key findings include:

- **Mel-Scalograms:** Improved the model's performance (Accuracy 2) compared to using only mel-spectrograms, capturing unique temporal features crucial for tempo estimation.
- **Proposed Model:** Achieved results comparable to state-of-the-art methods, surpassing them in specific datasets like GiantSteps.

Future Work: Explore additional audio signal representations to further improve model inputs; Experiment with alternative CNN architectures to optimize performance; Enhance data augmentation methods to improve model generalization.

References

[1] S. Böck, F. Krebs, and G. Widmer, "Accurate tempo estimation based on recurrent neural networks and resonating comb filter," in *Proceedings of the 16th ISMIR Conference*, pp. 486–493, 2015.

[2] H. Schreiber and M. Müller, "A single-step approach to musical tempo estimation using a convolutional neural network," in *Proceedings of the 19th ISMIR Conference*, pp. 98–105, 2018.

[3] S. Böck, M. Davies, and P. Knees, "Multi-task learning of tempo and beat: Learning one to improve the other," in *Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR)* (A. Flexer, G. Peeters, J. Urbano, and A. Volk, eds.), pp. 486–493, Zenodo, 2019.

Acknowledgments: The authors would like to thank CAPES and CNPq.

https://lamir-workshop.github.io/   1st Latin American Music Information Retrieval (LAMIR) workshop, from December 9 to 11, 2024, in Rio de Janeiro   luiz.guimaraes@ufba.br