Adapting Beat Tracking Models for Salsa Music: Establishing a Baseline With a Novel Dataset

Antonin Rapini
(aplr3@kent.ac.uk)

Anna Jordanous
(A.K.Jordanous@kent.ac.uk)

Abstract: This study addresses the challenge of adapting current beat tracking algorithms, predominantly trained on Western music, to the rhythmic complexities of Salsa, a genre rich in syncopations and polyrhythms. Using training methods that minimise the need for extensive annotated data, we benchmark the adaptability of two established models: BeatNet and BöckTCN, on our newly introduced beat and downbeat annotated Salsa dataset. We find that, on Salsa music, models trained with Salsa largely outperform models trained without any Salsa, nearly matching the accuracy of these models on Western music. This research not only establishes a baseline for beat and downbeat tracking performance in Salsa music but also contributes to the broader goal of developing more adept music information retrieval systems. We also contribute a 40-song Salsa dataset for beat and downbeat tracking research in this genre.

Background

Beat tracking—the automatic detection of beats in audio recordings—is a fundamental task in music information retrieval (MIR). While current models perform well on Western music genres, they often struggle with styles that feature complex rhythms [1], such as Salsa.

Salsa music is rich in rhythmic diversity, characterized by syncopations, polyrhythms, and the clave pattern—a fundamental motif alternating between 3-2 and 2-3 patterns within a 4/4 meter. This complexity, along with variable tempos and expressive timing, presents significant challenges for beat tracking algorithms.

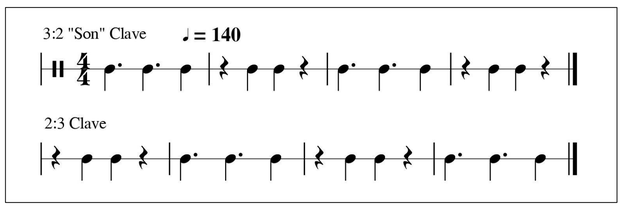

Figure 1. The Son clave, a fundamental rhythmic pattern serving as the underlying guiding structure for most Salsa music.

Objectives

This study aims to bridge the gap by:

- **Benchmarking** the beat and downbeat tracking accuracy on Salsa music.
- **Evaluating** the adaptability of two state-of-the-art models—**BeatNet** [2] and **BöckTCN** [3]—to Salsa's complex rhythms.
- **Contributing** a newly annotated Salsa dataset to the research community.

F-measure accuracy		
Model	GTZAN	Ballroom
BeatNet	0.806	N/A
TCN	0.885	0.962

Table 1. Reported average F-measure accuracy on popular music datasets of two prominent beat tracking models. BeatNet did not report any results for the Ballroom dataset.

Figure 2. Orquesta El Macabeo, A Puerto-Rican Salsa band featuring many traditional Salsa instruments, Congas, Piano, Timbales, Trumpets, Maracas…

Methodology

We conducted experiments using both existing and newly created datasets, evaluating two prominent beat tracking models under different training conditions.

Datasets

- **Non-Salsa Datasets ("Others"):**
  - GTZAN, Ballroom, SMC, Beatles, Rock Corpus.
- **New Salsa Dataset:**
  - **40 tracks** covering various eras, sub-genres, regional styles, and tempos (155–246 BPM).
  - **Annotations:** Manually labeled for beats and downbeats using Sonic Visualiser.

The Salsa Dataset

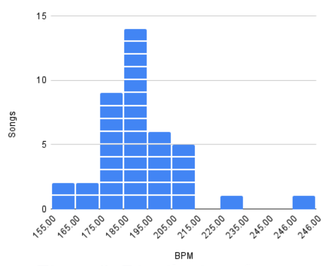

Figure 3. Distribution of tempos in the Salsa dataset

Most recent song:
Victimas Las Dos– Marc Anthony, LA INDIA (2021)

Least recent song:
Margie – Ray Barretto (1966)

Figure 4. Origins of songs in the Salsa dataset

Models Evaluated

- **BeatNet:** Combines convolutional neural networks with long short-term memory (LSTM) layers.
- **BöckTCN:** Employs temporal convolutional networks (TCNs) for sequence modeling.

Training Conditions

1. **Others Only:**
   1. Models trained exclusively on non-Salsa datasets.
2. **Salsa Only:**
   1. Models trained solely on the Salsa dataset.
3. **Fine-Tuning:**
   1. Models initially trained on "Others" and then fine-tuned with Salsa data.
   2. **Layer Adjustments:**
      1. **BeatNet:** Frozen convolutional layers; fine-tuned LSTM and output layers.
      2. **BöckTCN:** Frozen convolutional layers; fine-tuned TCN layers.

Cross-Validation

We used **5-fold cross-validation** to ensure robust evaluation, The 40 Salsa tracks were divided into five folds of eight tracks.

Evaluation Metric:

The **F-measure** was used to assess beat and downbeat tracking accuracy. A standard tolerance window of ±70 milliseconds was applied to accommodate human perception of timing variations.

Results

Key Observations:

- Training with Salsa data significantly improves beat tracking performance.
- **BeatNet** achieved the highest accuracy when trained exclusively on the Salsa dataset.
- **BöckTCN** showed substantial improvement through fine-tuning.

Key Observations:

- Downbeat tracking remains more challenging than beat tracking.
- Fine-tuning improves downbeat detection, particularly for **BeatNet**.
- **BöckTCN** exhibits limited downbeat tracking capabilities in this context.

F-measure accuracy			
Model	Fine-tuned	Salsa only	Others
BeatNet	0.845	0.855	0.560
TCN	0.771	0.437	0.420

Table 2. Average beat F-measure accuracy on the Salsa dataset of two prominent beat tracking models under the three training conditions outlined in the Methodology section.

Analysis of Challenging Tracks

Certain tracks presented difficulties for the models:

- **"Venenosa," "Es Tu Mirada," and "Juntando Amores":**
  - **Instrumentation Differences:** These songs feature less prominent rhythmic elements or non-traditional Salsa instrumentation.
  - **Model Performance:** Models underperformed on these tracks after Salsa-specific training, possibly due to overfitting to traditional Salsa rhythms.

Conclusion

Fine-tuning beat tracking models with genre-specific data significantly enhances accuracy for Salsa music. This study establishes a baseline for beat and downbeat tracking in Salsa, contributing to the development of beat tracking systems that better account for the rhythmic diversity of global music genres. We also introduce a new beat-annotated Salsa dataset to support future research.

Future Work

Using the baseline accuracy acquired from this experiment we plan to leverage large amounts of unannotated data, such as online dance videos, using self-supervised learning techniques. By integrating visual information from dance movements, we aim to improve beat tracking accuracy through multimodal data exploration.

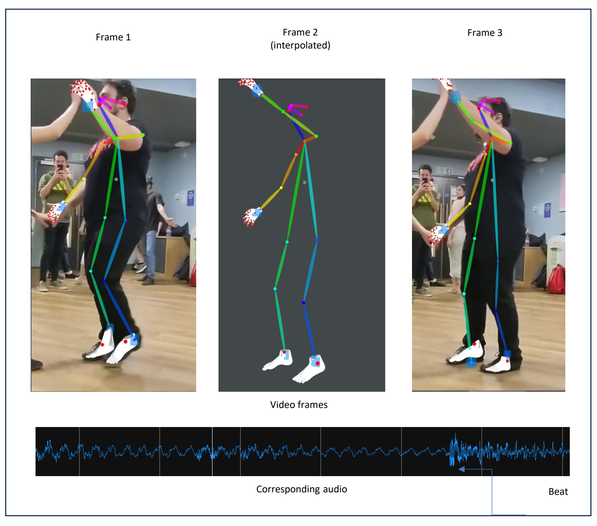

Figure 5. Frames of a dance video with corresponding audio. The estimation of a pose tracking algorithm is overlaid on the video frames. Frame 2 is an interpolation of the pose of the dancer between two video frames.

REFERENCES

1. L. Maia, M. Rocamora, L. Biscainho, and M. Fuentes, "Adapting meter tracking models to latin american music," in The 23rd International Society for Music Information Retrieval, Bengaluru, India, 2022, pp. 3– 11.
2. H. Mojtaba, F. Cwitkowitz, and Z. Duan, "Beatnet: A real-time music integrated beat and downbeat tracker," in The 22nd International Society for Music Information Retrieval, Online, 2021, pp. 270–277.
3. M. E. P. Davies and S. Böck, "Temporal convolutional networks for musical audio beat tracking," in 2019 27th European Signal Processing Conference (EUSIPCO), A Coruna, Spain, 2019, pp. 1–5.