BEAT TRACKING FOR SALSA MUSIC: ADAPTING AND BENCHMARKING MODELS USING A NEWLY INTRODUCED SALSA DATASET

Antonin Rapini
University of Kent, UK
aplr3@kent.ac.uk

Anna Jordanous
University of Kent, UK
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.

1. INTRODUCTION

Beat tracking is the temporal identification of beats—the basic rhythmic units of a song. Although it is a skill that comes naturally for most people [1], automatic beat tracking—the computational identification of beats from audio data—poses significant challenges for computational systems.

Downbeat tracking involves identifying the first beat of each measure in a musical piece and requires a deeper understanding of a song's musical structure, making it more challenging for computational models. In the current literature, downbeat tracking generally yields lower accuracy than beat tracking.

Current state-of-the-art beat tracking algorithms typically rely on machine learning models trained on large datasets predominantly featuring Western musical styles [2]. These models achieve high accuracy when evaluated on these same genres. However, their performance on genres such as Salsa remains largely unexplored.

Salsa music, known for its rich rhythmic structure characterized by syncopations and poly-rhythms, holds significant cultural importance and enjoys global popularity [3] [4]. The absence of annotated data for Salsa hinders the formal assessment and adaptation of these algorithms to such complex rhythmic patterns.

This study aims to bridge this gap by evaluating the adaptability of two state of the art beat tracking models: BeatNet [5] and BöckTCN [6], on Salsa music. We introduce a new beat-annotated Salsa dataset and explore training methods that minimize the need for extensive annotated data.

2. BACKGROUND

Salsa is known for its rich and dynamic rhythmic tapestry, reflecting the genre's deep roots in Afro-Cuban musical traditions, African rhythms, and the cultural fusion brought about by Latin American communities in New York City [3] [4]. Central to the genre is the clave pattern, a fundamental rhythmic motif that serves as the structural backbone, often alternating between 3-2 and 2-3 patterns within a 4/4 meter. Additionally, Salsa incorporates polyrhythms, where multiple rhythmic patterns are played simultaneously by different percussion instruments such as congas, timbales, and bongos. This layering of diverse rhythms results in off-beat accents and irregular syncopations. Furthermore, the variable tempo and expressive timing variations in Salsa performances add another layer of difficulty, requiring models to adapt to subtle fluctuations and maintain consistent beat detection. These features are not often represented in current beat tracking datasets, which predominantly focus on genres with more straightforward rhythms typically found in many Western music genres.

Annotating music is a time-consuming and arduous process [7]. The temporal nature of music means the manual annotation process takes at least the length of the annotated segment, often requiring multiple listens and minor corrections to achieve accurate beat placement. This labour-intensive process limits the availability of large, annotated datasets, particularly for genres like Salsa that have been overlooked in previous beat tracking research.

Addressing this lack of data for genres such as Salsa is crucial for progress toward music information retrieval systems that are more representative of diverse musical genres.

3. RELATED WORK

The adaptation of beat tracking models to non-Western musical genres has been a growing focus in Music Information Retrieval (MIR), as existing models trained primarily on Western music often struggle with different rhythmic structures. This section reviews three relevant studies that address these challenges.

Maia et al. [8] investigated adapting beat tracking models to Latin American music, specifically Samba and Candombe, using minimal annotated data and computational resources. They tested strategies including training from scratch, fine-tuning a pre-trained model, and applying data augmentation with a TCN model. Their approach demonstrates the potential for adapting beat tracking models to under-represented musical genres with limited annotated data. Building upon their strategies, our work focuses on Salsa music, which, like Samba and Candombe, features complex rhythmic structures. We explore similar methods to evaluate their effectiveness in the context of Salsa.

Fiocchi et al. [9] applied transfer learning to beat tracking in Greek folk music by utilizing a deep BLSTM-based RNN originally trained on popular Western music datasets. They collected and manually annotated a dataset of Greek folk music, which includes a variety of rhythms with irregular time signatures and tempo fluctuations. By freezing the lower layers of the pre-trained network and retraining the top layer on the Greek dataset, they achieved significant improvements compared to models trained from scratch on the limited data. Their results demonstrate the effectiveness of transfer learning in adapting beat tracking models to new genres with limited annotated data. This approach underscores the potential for leveraging existing models to handle diverse musical traditions without the need for extensive new datasets.

Pinto et al. [10] proposed a user-driven fine-tuning approach for beat tracking, aiming to enhance the performance of state-of-the-art models on specific challenging musical pieces. Their method involves adapting a pre-trained TCN by fine-tuning it using a small, user-annotated segment of the target piece. This approach allows the model to better handle expressive timing variations and complex rhythms without the need for large annotated datasets. They demonstrated significant improvements in beat tracking accuracy across various datasets. However, their approach is tailored to individual pieces, which limits its scalability and applicability to genre-wide adaptation. This raises questions about its effectiveness for broader applications where generalization across an entire genre is desired.

These studies provide valuable insights into the challenges and potential strategies for adapting beat tracking models to under-represented musical genres. However, limitations remain in achieving generalization across an entire genre with minimal annotated data. Our work extends these efforts by benchmarking multiple models and training conditions specifically for Salsa.

4. OBJECTIVES

This study sets out to establish a benchmark for beat tracking accuracy on Salsa music by evaluating two state-of-the-art models, Beatnet and BöckTCN, on a newly created beat-annotated Salsa dataset. This benchmark will allow future research to measure progress in developing more effective beat tracking systems for diverse musical genres.

Leveraging techniques such as transfer learning, we aim to optimise these models for Salsa music despite the scarcity of annotated data.

Additionally, this research introduces a novel beat and downbeat annotated Salsa dataset, with the objective of further improving beat tracking systems for diverse musical genres.

5. METHODOLOGY

We assess the accuracy of two prominent beat tracking models: BeatNet and BöckTCN, on an unseen Salsa test dataset created for this study. The models were trained under three distinct conditions:

5.1 Training Data

5.1.1 Other Datasets Used

The non-Salsa music datasets used (referred to as "Others") include: GTZAN [11], Ballroom [12], SMC [13], Beatles [14] and Rock corpus [15]

We were not able to obtain some of the datasets used in the original training of the two models, such as the Hainsworth dataset, due to accessibility constraints.

5.1.2 Salsa Dataset

The Salsa dataset comprises 40 tracks, which were divided into five folds of eight tracks for training. All tracks were used for evaluation, following the process described below.

5.2 Training Conditions

Three training conditions were used: (1) 'Others Only', using non-Salsa datasets; (2) 'Salsa Only', training solely on our Salsa dataset; and (3) 'Fine-Tuning', training with 'Others' and then fine-tuning with Salsa data.

We employed 5-fold cross-validation to measure the average F-measure accuracy of the models on the Salsa dataset. In each fold, the models were trained on 32 songs (with 10% used for validation) and evaluated on 8 unseen songs. This method ensures that every song in the dataset is used for testing exactly once.

5.3 Model Configurations

In all cases except fine-tuning, the models were trained with the original parameters presented in their respective papers or official implementations. For "Fine-Tuning" and "Salsa Only", we experimented with reduced learning rates ranging from 1×10^{-3} to 2×10^{-6}. We found that a learning rate of 5×10^{-4} resulted in a stable training process for both models, with consistent decreases in validation loss.

Training was conducted for a large initial number of epochs, and we monitored the validation loss throughout. The model checkpoint with the lowest validation loss was selected for evaluation.

The models were implemented using their official repositories when possible to ensure consistency with the original designs [16, 17]. Very minor changes were made to enable training with the datasets at our disposal.

5.3.1 Fine-Tuning Details

For fine-tuning, specific layers of each model were trained while others were frozen:

- **BeatNet**: The convolutional layers were frozen, and fine-tuning was applied to the LSTM layers and the final layer.
- **Böck TCN**: The convolutional layers were frozen, allowing fine-tuning of the Temporal Convolutional Network (TCN) layers.

5.4 Evaluation Metrics

We used the F-measure [7] as the primary metric to assess beat and downbeat tracking accuracy. For comparison, we include in Table Table 1 the average F-measure accuracy on popular music datasets previously reported for the two prominent beat tracking models we investigate in this study. A standard tolerance window of ± 70 milliseconds was applied when matching detected beats to the ground truth, accounting for slight timing variations and reflecting human perception of beat alignment.

5.5 Creation of the Salsa Dataset

5.5.1 Song Selection

The Salsa dataset compiled for this study consists of 40 tracks, selected to capture a diverse range of eras, regional styles, sub-genres, tempos, and instrumentation that characterize Salsa music. These tracks span various origins, with representation from areas such as Puerto Rico, Cuba, and the United States, and include popular sub-genres like Salsa Romántica, Salsa Dura, and Cuban Salsa. Release dates span from the 1970s to the 2020s. Tempos vary from 155 to 246 BPM, with an average around 191 BPM, calculated from the annotated beat intervals.

5.5.2 Beat Annotation Process

Beat annotations were created using the *Sonic Visualiser* software [18]. Each beat was manually placed through a combination of visual waveform inspection and auditory analysis to ensure precise timing. Subjective choices were made regarding the inclusion or exclusion of beats in certain sections; for instance, intros and outros with ambiguous or free rhythms were sometimes omitted to maintain annotation consistency.

The Salsa dataset can be accessed publicly via GitHub github.com/AntoninRap/Salsa-dataset.

6. RESULTS

Our findings reveal that accuracy increases with specialisation and can benefit from the general musical knowledge derived from training on other datasets. This is an unsurprising result to some extent, but it was useful to see that this consistent improvement could be obtained even with a small amount of genre-specific training data.

	BeatNet	BöckTCN
GTZAN	0.806	**0.885**
Ballroom	N/A	**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.

	BeatNet	BöckTCN
Fine-tuned	0.845	**0.771**
Salsa only	**0.855**	0.437
Others (base)	0.560	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.

	BeatNet	BöckTCN
Fine-tuned	**0.522**	**0.216**
Salsa only	0.516	0.052
Others (base)	0.215	0.042

Table 3. Average downbeat F-measure accuracy on the Salsa dataset of two prominent beat tracking models under the three training conditions outlined in the Methodology section.

Tables Table 2 and Table 3 present the average beat and downbeat F-measure accuracies, respectively.

BeatNet achieved its highest F-measure when trained solely on the Salsa dataset (0.855), slightly surpassing its performance when fine-tuned (0.845). The base BeatNet model scored significantly lower (0.560). For BöckTCN, fine-tuning resulted in the highest F-measure (0.771), outperforming the Salsa-only training (0.437) and the base model (0.420). These results suggest that incorporating Salsa data enhances beat tracking performance. BeatNet benefits more from training exclusively on Salsa data, while BöckTCN shows greater improvement through fine-tuning.

Both models exhibited lower F-measure scores for downbeat tracking compared to beat tracking. BeatNet achieved its highest downbeat F-measure when fine-tuned (0.522), closely followed by Salsa-only training (0.516). The base model had a considerably lower score (0.215). BöckTCN's best downbeat F-measure was 0.216 when fine-tuned; performance decreased with Salsa-only training (0.052) and was minimal for the base model (0.0*).

Fine-tuning improves downbeat detection, particularly for BeatNet, but overall accuracy remains low. The original BöckTCN Paper does not report any downbeat capabilities or results. These results were obtained using Ben Hayes's BöckTCN implementation [16]

A closer look into individual beat tracking results reveals that most models obtain higher accuracy on most songs in the dataset after training with Salsa specific data. Specifically, for BeatNet in the "Salsa Only" and "Fine-Tuning" training conditions, a majority of songs achieved higher-than-average F-measure scores compared to the overall average in their respective training conditions. The average accuracy was brought down by a few songs with significantly reduced performance. Interestingly, for both models, these particular songs actually achieved higher results with the base models and present differences in instrumentation compared to the rest of the tracks in the dataset. We explore these findings in more detail in the discussion section below.

7. DISCUSSION

This study established a baseline for beat and downbeat tracking in Salsa music using a new, small-scale dataset. Our results show that models trained with Salsa-specific data perform better than those trained on non-Salsa datasets, and, in the case of BeatNet, outperforms its accuracy obtained on Western music genres.

Upon analysing the outlier negative results presented in Table 4, it became apparent that the results were likely due to the difference in instrumentation and lack of strong rhythmic elements rather than the complexity of the rhythms. For instance, in challenging tracks such as "Venenosa", "Es Tu Mirada" and "Juntando Amores" rhythmic instruments are either less prominent or played more subtly. In "Venenosa", for example, the rhythm is primarily carried by a soft Tumbao on the conga, while the piano, guitar, bass, and vocals dominate the mix. One thing to note here is that in Salsa, the piano most often functions as a rhythmic instrument, rather than serving a primarily melodic role as it does in many Western genres.

"Es Tu Mirada" is widely enjoyed by Salsa dancers around the world; however, its instrumentation differs from traditional Salsa music and more akin to fusion of Cuban pop with traditional Cuban music elements. In this track, the rhythmic instruments are slightly muted compared to the prominent vocals and bass. Similarly, "Juntando Amores" blends Salsa rhythmic elements with flamenco guitar and is characterized by a very fast tempo. In this track, the rhythmic elements take a backseat to the prominent guitar. These deviations in instrumentation and emphasis on non-traditional elements may have impacted the models' performance on these three tracks.

The high beat tracking accuracy obtained on most of the dataset suggests that the models effectively specialized in traditional Salsa instrumentation. Training on 32 songs enabled the models to generalize effectively to 8 unseen songs with similar characteristics, and, notably, this generalization occurred consistently across all five folds in our cross-validation. However, the outlier results highlighted above seem to indicate that the models lost some of their ability to accurately track beats in songs whose instrumentation differed from traditional Salsa arrangements.

	Venenosa	Es Tu Mirada	Juntando Amores
BeatNet Fine-tuned	0.378	0.333	0.387
BeatNet Salsa only	0.324	0.249	0.442
BeatNet (others)	0.634	0.660	0.639
BöckTCN Fine-tuned	0.416	0.368	0.514
BöckTCN Salsa only	0.404	0.463	0.465
BöckTCN (others)	**0.974**	**0.662**	**0.657**

Table 4. Beat tracking F-measure accuracy on three songs with outlier accuracies for BeatNet and BöckTCN under the three training conditions outlined in the Methodology section.

For downbeat tracking, the results are more challenging to interpret. The F-measure accuracy varies significantly—from 0 to 1—and it is not clear yet why this variation occurs.

8. FUTURE WORK

The promising results obtained during this study with a limited amount of data highlight the need for further research. In following experiments we will focus on leveraging large amount of unannotated data, such as dance videos posted online, through techniques such as self-supervised learning. Through this research, we hope to further improve beat tracking accuracy by exploring multimodal data that integrates visual information from dance movements.

9. CONCLUSION

The study demonstrates that fine-tuning beat tracking models with genre-specific data can significantly improve accuracy for Salsa music. It also establishes a baseline for the performance of beat and downbeat tracking on this genre, providing a reference point for the efficacy of more intricate future methodologies. This work contributes to the ongoing efforts to develop beat tracking systems that better account for the rhythmic diversity found in global music genres, and with this objective in mind, introduces a new beat-annotated dataset of Salsa music.

10. REFERENCES

[1] J. Phillips-Silver and L. Trainor, "Feeling the beat: Movement influences infant rhythm perception," *Science*, vol. 308, no. 5727, p. 1430, 2005.

[2] S. Böck, Matthew, M. E. P. Davies, and P. Knees, "Multi-task learning of tempo and beat: Learning one to improve the other," in *The 20th International Society for Music Information Retrieval*, Delft, The Netherlands, 2019, pp. 486–493.

[3] P. Manuel, "Puerto rican music and cultural identity: Creative appropriation of cuban sources from danza to salsa," *Ethnomusicology*, vol. 38, no. 2, pp. 249–280, 1994.

[4] L. Waxer, *The City of Musical Memory: Salsa, Record Grooves, and Popular Culture in Cali, Colombia*. Middletown, CT: Wesleyan University Press, 2002.

[5] 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.

[6] 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.

[7] M. Davies, N. DeGara, and M. Plumbley, "Evaluation methods for musical audio beat tracking algorithms," Centre for Digital Music, Queen Mary University of London, London, UK, Tech. Rep. C4DM-TR-09-06, 2009.

[8] 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.

[9] D. Fiocchi, F. A. M. Buccoli, M. Zanoni, and A. Sarti, "Beat tracking using recurrent neural network: a transfer learning approach," in *2018 26th European Signal Processing Conference (EUSIPCO)*, Rome, Italy, 2018, pp. 1929–1933.

[10] A. Pinto, J. C. S. Böck, and M. Davies, "User-driven fine-tuning for beat tracking," *Electronics*, vol. 10, no. 13, pp. 10–13, June 2021.

[11] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 293–302, 2002.

[12] F. Gouyon, A. Klapuri, S. Dixon, M. Alonso, G. Tzanetakis, C. Uhle, and 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.

[13] A. Holzapfel, M. Davies, J. Zapata, J. Oliveira, and F. Gouyon, "Selective sampling for beat tracking evaluation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 9, pp. 2539–2548, 2012.

[14] M. E. P. Davies and M. Plumbley, "Context-dependent beat tracking of musical audio," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1009–1020, 2007.

[15] T. de Clerq and D. Temperley, "A corpus analysis of rock harmony," *Popular Music*, vol. 30, no. 1, pp. 47–70, 2011.

[16] H. Mojtaba. (2021) Beatnet. [Online]. Available: https://github.com/mjhydri/BeatNet

[17] B. Hayes. (2020) Beat tracking tcn. [Online]. Available: https://github.com/ben-hayes/beat-tracking-tcn

[18] C. Cannam, C. Landone, and M. Sandler, "Sonic visualiser: An open source application for viewing, analysing, and annotating music audio files," in *Proceedings of the ACM Multimedia 2010 International Conference*, 2010, pp. 1467–1468.