LEVERAGING REAL ELECTRIC GUITAR TONES AND EFFECTS TO IMPROVE ROBUSTNESS IN GUITAR TABLATURE TRANSCRIPTION MODELING

Hegel Pedroza
National Autonomous University of Mexico
Mexico City, Mexico

Wallace Abreu
Federal University of Rio de Janeiro
Rio de Janeiro, Rio de Janeiro, Brazil

Ryan M. Corey
Discovery Partners Institute & University of Illinois Chicago
Chicago, Illinois, USA

Iran R. Roman
New York University
New York, New York, USA

ABSTRACT

Guitar tablature transcription (GTT) aims at automatically generating symbolic representations from real solo guitar performances. Due to its applications in education and musicology, GTT has gained traction in recent years. However, GTT robustness has been limited due to the small size of available datasets. Researchers have recently used synthetic data that simulates guitar performances using pre-recorded or computer-generated tones, allowing for scalable and automatic data generation. The present study complements these efforts by demonstrating that GTT robustness can be improved by including synthetic training data created using recordings of real guitar tones played with different audio effects. We evaluate our approach on a new evaluation dataset with professional solo guitar performances that we composed and collected, featuring a wide array of tones, chords, and scales.

1. INTRODUCTION

Guitar tablature transcription (GTT), a form of automatic music transcription (AMT) [1, 2], involves transcribing real guitar performances into tablatures [3, 4]. Unlike standard Western notation, tablatures intuitively illustrate finger placements, and are thus of high relevance for music education [5, 6], musicological research [7], guitar performance theory [8–10], and general communication of artistic expression.

Significant advances in GTT have been achieved through deep learning models, primarily trained and evaluated using the GuitarSet dataset [11–18]. However, its limited size has led to poor generalization capabilities [19]. This is the well-known "domain-shift problem" in machine learning, which explains model failures in real-world applications due to discrepancies between the training conditions and actual usage environments [20]. Therefore, assessment of GTT model robustness in new domains is essential to understand their usefulness. Data augmentation is a common technique to mitigate "domain-shift". By systematically modifying existing data or metadata, additional examples can be generated and used as training data, thereby expanding a model's familiarity with a wider data distribution [21]. This technique has recently shown improvements in related tasks like sound event detection [22–25].

In the case of GTT, data augmentation can be carried out by taking existing tablatures to guide the temporal placement of pre-recorded or synthesized individual guitar tones to simulate a guitar performance. Zang *et al.* [19] recently used this technique with commercially-available guitar synthesizers.

We want to investigate whether diversifying the timbres present in training data by including audio effects improves GTT. Since recording and annotating solo guitar with audio effects is highly time-intensive, we draw inspiration from Zang et al.'s SynthTab [19] to scalably simulate guitar performances from guitar tablature. However, instead of using synthetic guitar tones (i.e. "MIDI"), we hypothesize that GTT robustness can be enhanced by two factors: the exclusive use of real guitar tones and audio effects to generate synthetic data for model training.

Finally, after training we assess model robustness using a new dataset of professional solo guitar performances that we composed and collected for this study. This new dataset features a diverse array of performance styles. Therefore, our key contributions are:

1. A scalable method to generate data to train GTT models using real recordings of guitar tones and audio effects.
2. A new dataset for evaluating GTT models.
3. A benchmark clearly demonstrating the benefits of our approach for model robustness.

We release our code to reproduce our data generation and model training, as well as the new dataset we collected for this study[1].

2. METHODS

2.1. Datasets for training and validation

Like Wiggins & Kim [12], we train TabCNN using GuitarSet [11] for training and cross-validation. Additionally, we introduced two synthetic datasets to expand the data used for model training: GuitarSetFX and GuitarProFX. Our approach maximizes tone diversity by randomly selecting tones from a large set of pre-recorded single guitar notes. Thus, in our synthetic solo guitar performances, melodies and chords consists of notes played with varied guitar tones and audio effects. This data generation strategy aligns with our hypothesis that such diversity will enhance the model's robustness, allowing it to concentrate on pitch content and guitar string+fret inference, while disregarding specific timbre qualities.

The guitar tone set we use for GuitarSetFX and GuitarProFX includes the clean tones and effects from EGFxSet[2], which used a 2004 Fender Stratocaster guitar [26]. Additionally, we recorded all possible clean notes in a 1978 Ibanez Performer 300 guitar, using

[1] `robust-guitar-tabs.github.io.`

[2] We exclude the tones processed through the "delay" effect due to audibly repeated tone onsets that would confuse our model.

Table 1: TabCNN performance on GuitarSet. Averages (± denotes the standard deviation) across the six hold-out folds. Top row: metrics by Wiggins & Kim [12] (we could reproduce their results). Bottom rows: outcomes when training data includes simulated tracks.

	Multi-pitch estimation			**Tablature estimation**			
	F$_1$	**P**	**R**	**F$_1$**	**P**	**R**	**TDR**
TabCNN [12]	0.826±0.025	0.900±0.016	0.764±0.043	**0.748**±0.047	**0.809**±0.029	0.696±0.061	0.899±0.033
+ GuitarSetFX	**0.837**±0.019	0.904±0.019	**0.785**±0.038	0.746±0.030	0.795±0.022	**0.708**±0.044	0.896±0.021
+ GuitarProFX	0.830±0.018	**0.908**±0.014	0.769±0.027	0.743±0.029	0.802±0.029	0.696±0.035	**0.900**±0.021

alternate-picking technique and captured through direct input via an Audient iD14 audio interface, also miked through an Orange CR-60 amplifier with a Shure sm57 microphone. The amplifier settings were adjusted with bass and treble knobs at 5 and plate reverb at 3.

GuitarSetFX reproduces the 360 guitar performance tracks of GuitarSet, while GuitarProFX comprises 360 randomly-chosen solo performance tracks from DadaGP [27]. All resulting audio clips are resampled to the 22050Hz sampling rate expected by TabCNN.

2.2. Model training, and validation

The GTT model we train is TabCNN, using the implementation in AMT tools [13][3]. More specifically, we train three different TabCNN models. The first one is a reproduction of the six-fold cross-validated training setup by Wiggins & Kim [12]. The second one follows the same cross-validation but duplicates the training data using the corresponding synthetic GuitarSetFX tracks. The third one also follows the same cross-validation but adds all the synthetic GuitarProFX tracks to the training split. All other aspects of the training setup remain the same as in the original TabCNN implementation, including model architecture, optimizer, learning rate, batch size, and validation data [12].

2.3. The EGSet12 evaluation set

Furthermore, we introduce EGSet12, a new evaluation set with twelve original solo electric guitar performances (31.65 seconds avg. duration, totaling 379.8 seconds). These pieces were composed by a professional musician and guitar player for this project, showcasing the full tonal range of the electric guitar across diverse melodies and chord complexities. EGSet12 encompasses a broad spectrum of styles, including pop, funk, jazz and twelve-tone, reflecting varied tonalities, keys, rhythms, and modes.

EGSet12 was performed by a single professional guitarist using a Sire T7 Telecaster guitar and a Yamaha B15 amplifier. The performance setup allowed the performer to freely select the guitar's volume and tone knobs, also allowing techniques like alternate picking, hybrid picking, and palm mute. We captured the performance using an ECM8000 microphone positioned 15 centimeters from the amplifier and connected to a UMC202 HD audio interface (original sampling rate of 48000Hz; resampled to 22050Hz for model inference). This recording setup differs significantly from those used in any of the training and validation datasets that we used, offering a new testing domain. EGSet12, offers realism, tone diversity, and varied playing styles, making it valuable for assessing GTT model robustness.

EGSet12 features a realistic noisy recording setup and diverse guitar tones and techniques. Other than the amplifier, its content was not processed using other guitar effects. Future research can process it through more effects to further study model robustness.

2.4. Metrics

Consistent with Wiggins & Kim [12], we use two types of metrics: multi-pitch and tablature estimation. Multi-pitch metrics assess model performance at the level of pitch estimation and can be thought of as independent of the guitar hardware. Tablature estimation metrics assess the model's ability to determine which specific string and fret produced a tone. Both are broken down by F$_1$ score, precision, and recall. Additionally, the tablature disambiguation rate (TDR) calculates how often a correctly-identified pitch gets assigned to the correct fret and string.

3. RESULTS

3.1. GuitarSet cross-validation

First, we assessed TabCNN performance on GuitarSet during its cross-validated training. Table 1 shows the results. In general, we observe slight improvements when the training set includes synthetic data with effect tones. However, these improvements are minor. This indicates that cross-validated performance of TabCNN on GuitarSet does not benefit (or suffer) much from the addition of synthetic data with guitar effects.

3.2. Model evaluation on the new EGSet12 test domain

Next, we assessed the three models on EGSet12, with results presented in Table 2. Although, model performance was comparable during cross-validation with GuitarSet (see Table 1), evaluation on EGSet12 showed that models trained with synthetic data using audio effects exhibited superior generalization. Notably, the F$_1$ score improved more than 10 percentage points for both multi-pitch and tablature estimation, primarily driven by major gains in model recall. Precision remained consistent across models for multi-pitch metrics but showed improvements of ∼9 points in GuitarProFX tablature estimation and slightly over 5 points in GuitarSetFx. Finally, the TDR also saw improvements of more than 10 points for the model trained using the GuitarProFX data.

3.3. Qualitative comparison of TabCNN models on EGSet12

Figure 1 shows a qualitative comparison between the TabCNN models we tested on EGSet12. The first two columns in Figure 1 feature musical excerpts showing that TabCNN trained with GuitarProFx demonstrated greater stability, meaning that it did not exhibit constant string jumping when trying to locate the guitar notes

[3] github.com/cwitkowitz/amt-tools.

Table 2: TabCNN performance on EGSet12. Each cell is a metric averaged across the twelve tracks (± denotes the standard deviation). Top row: performance as trained by Wiggins & Kim [12]. Bottom rows: performance when training data includes simulated tracks. "*" denotes a statistically-significant difference ($p < 0.05$ via t-test) compared to the model by Wiggins & Kim [12]. "$^{\diamond}$" denotes a marginally-significant difference ($0.1 > p > 0.05$). The underlying distributions of significant (or marginally-significant) comparisons are normal-shaped based on the D'Agostino and Pearson's test.

	Multi-pitch estimation			**Tablature estimation**			
	F$_1$	**P**	**R**	**F$_1$**	**P**	**R**	**TDR**
TabCNN [12]	0.638±0.060	0.819±0.080	0.530±0.067	0.447±0.071	0.565±0.089	0.375±0.067	0.695±0.075
+ GuitarSetFX	**0.740**±0.055*	0.835±0.085	**0.679**±0.052$^{\diamond}$	0.557±0.088	0.619±0.100*	0.518±0.084	0.755±0.106
+ GuitarProFX	0.719±0.061$^{\diamond}$	**0.839**±0.082	0.647±0.068	**0.585**±0.084	**0.658**±0.073*	**0.541**±0.087$^{\diamond}$	**0.819**±0.075$^{\diamond}$

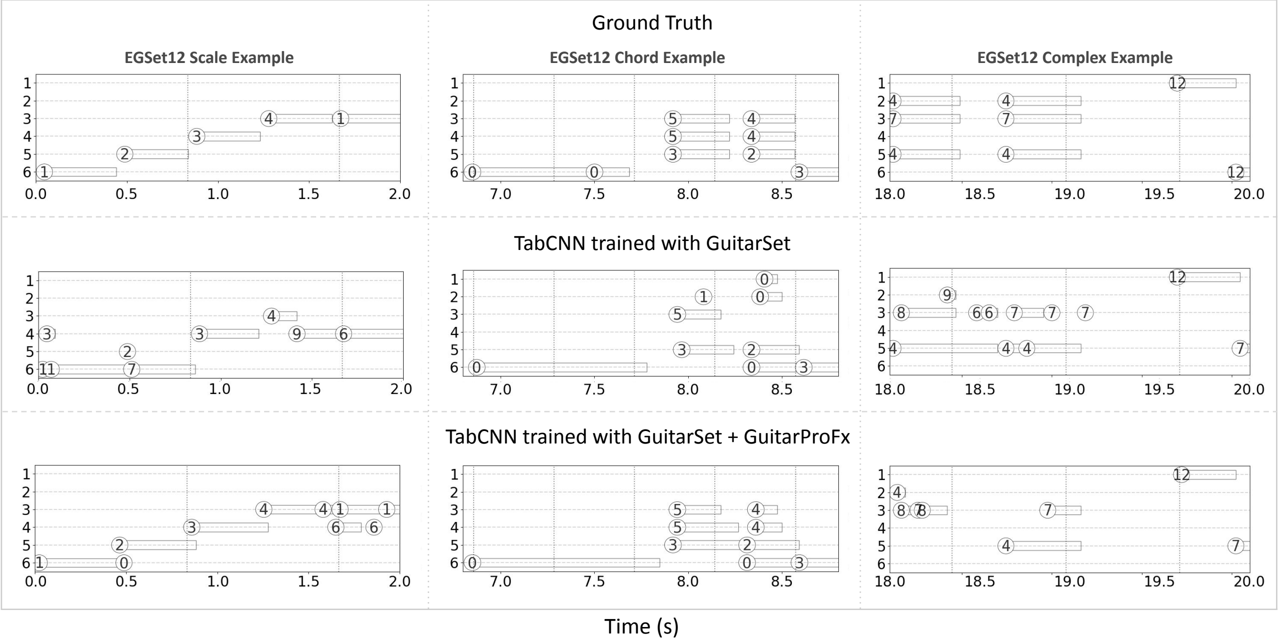

Figure 1: Each column is a two-second EGSet12 excerpt, comparing models trained using GuitarSet, with and without GuitarProFX, against ground truth. Each circled number is a tracked note on a specific guitar fret over time, the vertical lines indicate musical beats.

in the audio signal. Therefore, its predictions of chords and individual notes were more aligned with ground truth.

The third column is a more challenging example. It shows that both models' performance decreased when processing complex musical passages with harmonic content across multiple simultaneous strings and dissonances such as minor seconds. Additionally, both models struggled to make good predictions for the higher frets, as most training data features fret numbers below 12. However, in both cases, some pitch predictions were accurate as tablature was associated with possible fingerings in other frets.

4. DISCUSSION

Our results in Table 2, showing the benefits of synthetic training data, are consistent with empirical evidence from the sound event detection literature [22–25]. An interesting observation is the fact that the benefit size was not evident during the cross-validated training (see Table 1). This highlights the importance of evaluating GTT on new, structured domains and controlled scenarios (like EGSet12) to assess model robustness accurately.

Another point of discussion is how valid EGSet12 is as a test set for GTT. For example, considering EGSet12 uses an electric guitar, is it fair to use it to evaluate GTT models trained with an acoustic guitar dataset (i.e GuitarSet)?. We believe that this is the case since in our setup the electric guitar used in EGSet12 and all recording hardware (including microphone, amplifier, and interface) is different from not just GuitarSet, but also GuitarSetFX, and GuitarProFX. The fact that tablature estimation precision is higher on EGSet12 for models trained using GuitarSetFX or GuitarProFX (see table 2) may suggest an advantage due to the fact that these datasets used electric guitars. However, the more than 10-point increase in multi-pitch estimation F$_1$ indicates that model robustness is driven by correct inference of pitch content. Therefore, our evidence suggests that EGSet12 is a fair benchmark.

We studied how using real guitar tones processed through audio effects hardware to generate training data improves GTT model robustness. Therefore, our current study is inspired by SynthTab [19] and is not a comparison against it. It is worth noting, however, that we used considerably less synthetic data than SynthTab (we only synthesized 360 tracks to train each model, while SynthTab added

6,700 hours [19]). In future work, we will systematically explore the impacts of the various factors that go into simulating guitar performances, such as including real versus computer-generated tones and/or effects, the impact of delay effects that repeat the onset of a tone, and the amount of data used for training.

5. CONCLUSION

We have demonstrated the impact that using synthetic guitar performances as training data has on the robustness of GTT models. Specifically, we leveraged guitar tablatures to produce these performances using real recordings of electric guitar notes with a wide array of processing that included real audio effects hardware. We showed increased model robustness on multi-pitch and tablature prediction metrics via our proposed method. In the future, we look forward to enhancing datasets for GTT using our methodology and systematically studying all the parameters involved in the data generation, such as dataset size or tone diversity.

6. ACKNOWLEDGEMENTS

This work was supported by the IEEE Mentoring Experiences for Underrepresented Young Researchers (ME-UYR) and Mexico's National Scholarship for Graduate Studies from the National Council of Humanities, Sciences, and Technologies (CONAHCYT). The authors thank the funding sources and collaborators, particularly PhD candidate Elliot Hernandez, who proofread this manuscript.

7. REFERENCES

[1] E. Benetos, S. Dixon, Z. Duan, and S. Ewert, "Automatic music transcription: An overview," *IEEE Signal Processing Magazine*, vol. 36, no. 1, pp. 20–30, 2018.

[2] B. S. Gowrishankar and N. U. Bhajantri, "An exhaustive review of automatic music transcription techniques: Survey of music transcription techniques," in *International Conference on Signal Processing, Communication, Power and Embedded System*, 2016, pp. 140–152.

[3] G. Burlet and I. Fujinaga, "Robotaba guitar tablature transcription framework." in *ISMIR*, 2013, pp. 517–522.

[4] M. Paleari, B. Huet, A. Schutz, and D. Slock, "A multimodal approach to music transcription," 11 2008, pp. 93 – 96.

[5] D. E. Thompson, "Speaking their language: Guitar tablature in the middle school classroom," *General Music Today*, vol. 24, no. 3, pp. 53–57, 2011.

[6] E. Harrison, "Challenges facing guitar education," *Music educators journal*, vol. 97, no. 1, pp. 50–55, 2010.

[7] C. M. Gavito, "Oral transmission and the production of guitar tablature books in seventeenth-century italy," *Recercare*, pp. 185–208, 2015.

[8] J. De Souza, "Guitar thinking," *Soundboard Scholar*, vol. 7, no. 1, p. 1, 2022.

[9] ——, "Fretboard transformations," *Journal of Music Theory*, vol. 62, no. 1, pp. 1–39, 2018.

[10] T. Koozin, "Guitar voicing in pop-rock music: A performance-based analytical approach," *Music Theory Online*, vol. 17, 10 2011.

[11] Q. Xi, R. Bittner, J. Pauwels, X. Ye, and J. Bello, "Guitarset: A dataset for guitar transcription." in *ISMIR*, 2018.

[12] A. Wiggins and Y. Kim, "Guitar tablature estimation with a convolutional neural network." in *ISMIR*, 2019, pp. 284–291.

[13] F. Cwitkowitz, T. Hirvonen, and A. Klapuri, "Fretnet: Continuous-valed pitch contour streaming," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023, pp. 1–5.

[14] S. Kim, T. Hayashi, and T. Toda, "Note-level automatic guitar transcription using attention mechanism," in *European Signal Processing Conference*, 2022, pp. 229–233.

[15] G. Bastas, S. Koutoupis, M. Kaliakatsos-Papakostas, V. Katsouros, and P. Maragos, "A few-sample strategy for guitar tablature transcription on inharmonicity analysis & playability constraints," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022, pp. 771–775.

[16] G. Byambatsogt, L. Choimaa, and G. Koutaki, "Guitar chord sensing and recognition using multi-task learning and physical data augmentation with robotics," *Sensors*, vol. 20, no. 21, 2020.

[17] X. Riley, D. Edwards, and S. Dixon, "High resolution guitar transcription via domain adaptation," 2024.

[18] Y. Jadhav, A. Patel, R. H. Jhaveri, and R. Raut, "Transfer learning for audio waveform to guitar chord spectrograms using the convolution neural network," *Mobile Information Systems*, vol. 2022, p. 8544765, Aug 2022.

[19] Y. Zang, Y. Zhong, F. Cwitkowitz, and Z. Duan, "Synthtab: Leveraging synthesized data for guitar tablature transcription," 04 2024, pp. 1286–1290.

[20] J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, *Dataset shift in machine learning*. Mit Press, 2022.

[21] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of big data*, vol. 6, no. 1, pp. 1–48, 2019.

[22] J. Salamon, D. MacConnell, M. Cartwright, P. Li, and J. P. Bello, "Scaper: A library for soundscape synthesis and augmentation," in *Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, 2017, pp. 344–348.

[23] I. R. Roman, C. Ick, S. Ding, A. S. Roman, B. McFee, and J. P. Bello, "Spatial scaper: a library to simulate and augment soundscapes for sound event localization and detection in realistic rooms," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2024.

[24] A. S. Roman, B. Balamurugan, and R. Pothuganti, "Enhanced sound event localization and detection in real 360-degree audio-visual soundscapes," *arXiv preprint arXiv:2401.17129*, 2024.

[25] H. Dinkel, M. Wu, and K. Yu, "Towards duration robust weakly supervised sound event detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 887–900, 2021.

[26] H. Pedroza, G. Meza, and I. Roman, "Egfxset: Electric guitar tones processed through real effects of distortion, modulation, delay and reverb," *ISMIR LBD*, 2022.

[27] P. Sarmento, A. Kumar, C. Carr, Z. Zukowski, M. Barthet, and Y.-H. Yang, "Dadagp: a dataset of tokenized guitarpro songs for sequence models," in *ISMIR*, 2021.