PULSE CLARITY METRICS DEVELOPED FROM A DEEP LEARNING BEAT TRACKING MODEL

Nicolás Pironio[1] **Diego Fernández Slezak**[1,2] **Martín A. Miguel**[1,2]

[1] Universidad de Buenos Aires. Facultad de Ciencias Exactas y Naturales. Departamento de Computación. Buenos Aires, Argentina.

[2] CONICET-Universidad de Buenos Aires. Instituto de Investigación en Ciencias de la Computación (ICC). Buenos Aires, Argentina.

npironio@dc.uba.ar

ABSTRACT

In this paper we present novel pulse clarity metrics based on different sections of a state-of-the-art beat tracking model. Said model consists of two sections: a recurrent neural network that estimates beat probabilities for audio and a dynamic Bayesian network (DBN) that determines beat moments from the neural network's output. We obtained pulse clarity metrics by analyzing periodical behavior from neuron activation values and we interpreted the probability distribution computed by the DBN as the model's certainty. To analyze whether the inner workings of the model provide new insight into pulse clarity, we also proposed reference metrics using the output of both networks. We evaluated the pulse clarity metrics over a wide range of stimulus types such as songs and mono-tonal rhythms, obtaining comparable results to previous models. These results suggest that adapting a model from a related task is feasible for the pulse clarity problem. Additionally, results of the evaluation of pulse clarity models on multiple datasets showed that, with some variability, both ours and previous work generalized well beyond their original training datasets.

1. INTRODUCTION

In music, the pulse refers to the underlying regular rhythmic pattern in a song, usually expressed by listeners by tapping their foot. In western notation, the pulse takes on an especially relevant role, given that location and duration of rhythmic events are described with respect to it. Listeners can extract a pulse from the acoustic surface and infer a meter structure that may enable them to adjust their own behaviour to it (e.g. dance accordingly to a song) [1].

The strength with which the feeling of the pulse emerges in a listener is not necessarily the same for all music. The concept of pulse clarity refers to such subjective experience. As the pulse is relevant for temporal organization, pulse clarity facilitates a listener's understanding of a song, affecting the musical experience. Musical cognition experiments have used pulse clarity as a high-level musical feature, usually correlating it to human responses. For example, it has been related with degree and variability of movement [2, 3], as well as with specific neural responses to different musical stimuli [4]. Pulse clarity has been seen to be influenced by different rhythmic structure characteristics (e.g. syncopation), as it affects participant's beat tapping variability [5–7].

In the mentioned experiments, pulse clarity is estimated from the musical input using the model from Lartillot et al. [8]. In their work, the authors present various descriptors for audio recordings based on the analysis of an onset detection curve and the periodicities it presents. The model consists of the best descriptor, which was selected based on experimental results where participants rated the pulse clarity of movie soundtrack excerpts. Miguel et al. [9] proposed another pulse clarity model for symbolic representations of rhythmic passages. Said model outputs a beat congruence score over time which is interpreted as a pulse clarity metric. The model's score was evaluated by comparing it to human beat tapping variability data over songs and achieved comparable results to the Lartillot et al. [8] model.

A closely related problem in the Music Information Retrieval discipline is the beat tracking task, which consists of determining the pulse moments for a musical excerpt [10]. This task has a long and varied history of models, with most recent and proficient ones making use of novel techniques such as deep learning. Since pulse clarity can be thought of as the difficulty of performing beat tracking, our work proposes a transfer learning approach using data from beat tracking models to estimate pulse clarity in musical excerpts [11]. Here we develop a methodology for interpreting the beat tracking deep learning architecture presented by Krebs et al. [12] and Bock et al. [13] which has proved its effectiveness on the beat tracking task. This

family of models was selected because of its state-of-the-art performance and its use of a dynamic Bayesian network, which estimates probabilities of different beat interpretations. We theorized these estimations to be useful to approximate pulse clarity.

In this exploratory analysis, we propose a series of pulse clarity metrics based on different sections from the architecture. We evaluated the proposed pulse clarity metrics by calculating the Spearman rank correlation coefficient against pulse clarity interpretations from three empirical datasets: the movie soundtrack excerpts used in [8], where pulse clarity was self reported, the musical excerpts from the MIREX Beat tracking train dataset, where pulse clarity was calculated from variability in the tapping data, and the rhythmic passages from Miguel et al. [14] where pulse clarity was both reported by participants and calculated from the tapping data. We also present results for the Lartillot et al. [8] and Miguel et al. [9] models over the datasets as reference. Evaluation results show that there is relevant information for the pulse clarity problem in the associated beat tracking task as our metrics performed as well as previous pulse clarity models.

In the next section the general architecture from Krebs et al. [12] and Bock et al. [13] is presented, as well as the corresponding developed interpretations. The evaluation section describes the datasets in detail, explains how pulse clarity was estimated from tapping data and shows the evaluation methodology and the obtained results. We conclude by arguing that, based on our results, repurposing a related task model is reasonable for pulse clarity estimation and suggest further evaluation of the existing models.

2. PULSE CLARITY MODEL

In this section we briefly review the model architectures presented by Krebs et al. [12] and Bock et al. [13] for the beat tracking task, highlighting the key features relevant for the pulse clarity metrics we derived. We then describe each metric proposed, categorized by which aspects of the beat tracking model were considered for its definition.

2.1 Beat tracking model

The beat tracking architecture consists of three steps: audio preprocessing, estimation of beat probability for a given audio frame, and selecting beat moments given the beat probabilities. Bock et al. [13] presents a modification to the original architecture from Krebs et al. [12] to also take into account downbeat tracking. As these are fairly similar in respect to their architecture, all metrics developed are implemented from both the beat and downbeat models. Here we describe the downbeat model from Bock et al. [13] (as it is implemented by the authors in the `madmom` package [15], version 0.16.1) and clarify the significant differences between the two models when necessary.

In the audio preprocessing stage, different magnitude spectrograms are computed from the audio signal. These in turn are used as the input for a neural network ensemble. Each network in the ensemble is a recurrent neural network (RNN) that estimates the beat probability of an audio frame. These RNN have three hidden recurrent bidirectional layers, each with 25 long-short term memory (LSTM) cells [16]. The network's output consists of three neurons with a softmax activation function, representing the probability distribution over the *beat*, *downbeat* and *no beat* classes for a given audio frame. In the case of the Krebs et al. model, the probability distribution represents only the *beat* and *no beat* classes. The final beat activation function of the ensemble is computed as the average between each individual network's output.

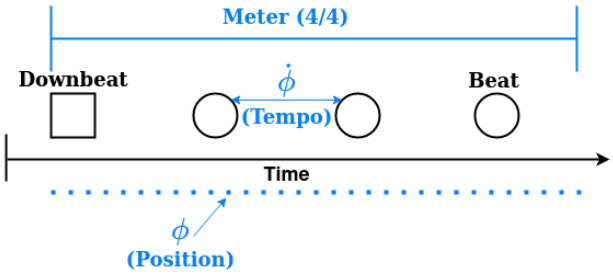

Figure 1. Example 4/4 bar with the position, tempo and meter state space variables associated.

Lastly, a dynamic Bayesian network (DBN) is used to determine the sets of beat and downbeat moments, given the probabilities output by the RNN ensemble. Conceptually, the sequence of audio frames is associated with a Markov chain of latent variables and the output of the RNN is used as the observations. The latent variables state space consists of a set of possible bar positions, tempi and time signatures - only the first two variables are considered in the Krebs et al. [12] model. Using the Viterbi algorithm, the most probable sequence of variables is determined and from it the sets of beats and downbeats moments are extracted.

2.2 Dynamic Bayesian Network based metrics

Compared to a deep learning architecture, the Bayesian dynamic network has a clearer interpretation. This is most notable by the use of a state space and transition model that encodes knowledge of the task at hand. In the analyzed model, the state space S of the DBN encodes the position within the bar (ϕ), the tempo ($\dot{\phi}$) and the time signature for each audio frame (3/4, 4/4). In Figure 1 we depict how these variables are related to an example 4/4 bar.

In a DBN, scores proportional to the probability of the most likely state sequences are calculated using the Viterbi algorithm. We define D_s as the scores for full sequences (analyzing the entire input) for each possible ending state s. From this distribution we make two possible interpretations of pulse clarity. The first one considers the probability estimate of the most probable state, $\max(D_s)$, which we interpret as the level of confidence with which the last variable is determined. We name this metric **Viterbi max**. The second aims to capture the uncertainty over D_s by computing its entropy: $H(D_s)$. We call this metric **Viterbi entropy**. As entropy is lower for a more concentrated dis-

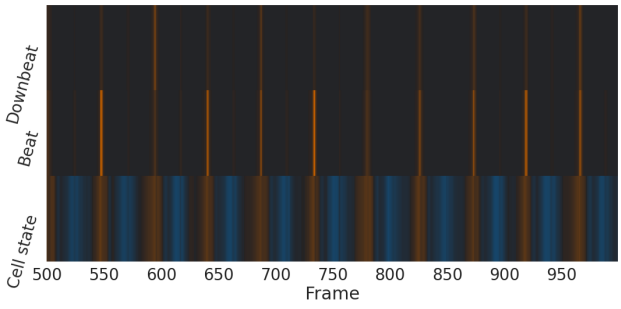

Figure 2. Average cell state activation for the last layer neurons in a 5 second excerpt of a song, contrasted with the beat activation function of the network. Orange colors are positive values, while blue colors are negative.

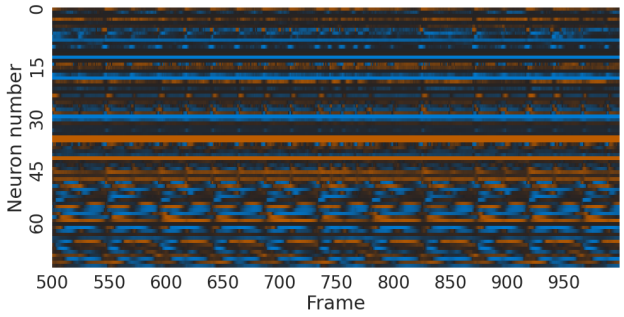

Figure 3. Sample output values of the forward layers neurons for a 5 second excerpt of a song.

tribution, a low value for this metric can be interpreted as a clearer decision made by the DBN.

2.3 Recurrent network based metrics

We decided to further explore whether the activations of the recurrent neural networks could be used to estimate pulse clarity. In this section of the architecture, we lacked a clear interpretation of the inner states of the deep learning model. Yet, given that the pulse of a song is an inherently periodical pattern, we hypothesized the RNN would present periodical activations at different levels of the network. When observing the network activations, these patterns were present in the form of activation peaks. With this in mind, we analyzed these activations from three different points of view, considering only a single trained network from the ensemble for simplicity.

Firstly, we considered the cell state values, which act as an "internal memory" of the LSTM cells, for each audio frame. Specifically, for each frame, the mean cell state activation of the 50 neurons in the last hidden layer is computed, separating the positive and negative values into separate series. Then, for each series, a peak picking process is performed, where the width for every peak is obtained. Averaging these widths results in the **Cell state precision** metric. In Figure 2 we can observe how the peak values of the average cell state activations for the last layer neurons tend to align to the final output of the network. The average width was interpreted as the confidence with which the network determines the beat probability for a frame: the wider a peak is, the lower the certainty.

Now focusing on capturing periodicity at a higher level, we turn to consider the cell activations (the output of the LSTM's) through time. Figure 3 shows the output series for a subset of neurons, in which periodical activation patterns can be interpreted for each neuron independently. This behavior motivated looking into possible periodicities between the output series of different cells and each cell with itself.

In the case of considering different series, for each pair of neurons i, j, the maximum absolute correlation between the output series o_i, o_j value is computed, considering all possible lags. We define the **neurons cross correlation** as the sum of each of these values, as depicted in Equation 1. With this metric we aim to determine the degree of coordination between neurons.

$$\text{NCC} = \sum_{o_i \neq o_j} \max_{lag \in [0, |o_i|]} \left| corr(o_i, o_j, lag) \right| \tag{1}$$

When considering each series against itself, we define **autocorrelation periodicity** as the average of the maximum autocorrelation values for each cell (Equation 2). These maximum autocorrelation values are obtained considering only lags representing periodicities between 40 and 330 BPM. Each autocorrelation value is divided by the size of the signal overlap, considering the lag. Compared to the neuron's cross correlation metric, the autocorrelation periodicity is less strict, as it only tries to capture if every neuron's output has a periodic pattern with itself.

$$\text{ACP} = \frac{1}{N} \sum_{i}^{N} \max_{lag \in L} \frac{A_{i,lag}}{\text{overlap}(o_i, lag)} \tag{2}$$

Where:

$$\begin{aligned} A_{i,lag} &= \left| corr(o_i, o_i, lag) \right| \\ \text{overlap}(o_i, lag) &= |o_i| - lag \\ L &= [40bpm, 330bpm] \end{aligned} \tag{3}$$

2.4 Output-based metrics

As the intention was to determine if there was relevant information from within the model for the pulse clarity task, we chose to develop reference metrics based on the output both from the RNN and the DBN. Subsequently, we analyze if the metrics derived from the inner workings of the networks surpass the performance of the output-based metrics. Using the beat activation function output by the RNN, two interpretations of pulse clarity were computed. First, we consider the average probability for beat moments as an indication of the overall certainty of the output. The **peak average** is obtained by applying a peak picking process to the beat probability signal and then averaging the peak probabilities. Second, using the previous calculated peaks as beat moments, we define the **RNN entropy** as the entropy of the inter-beat interval distribution. This concept is the same as the one used in the calculations of tapping variability presented in the Evaluation section. Analogously,

using the beat moments outputted by the DBN, we compute the **DBN entropy** as the entropy over the inter-beat interval distribution.

3. EVALUATION

In this section we evaluate the performance of the proposed metrics. We will present the considered datasets in detail, which vary in their type of stimuli and annotations and, as such, we clarify the pulse clarity interpretations of the annotated data specific to each one. Then the evaluation process is described, in which for each pulse clarity metric the absolute Spearman rank correlation coefficient is computed against each dataset. Using this coefficient as a performance score, a ranking of metrics is obtained.

3.1 Datasets

Three datasets were used for evaluation, which we named the MIREX, rhythms and soundtracks datasets. The *MIREX* dataset [17] has its origins in the beat tracking task, developed for the evaluation of models, with the intent to compile difficult-to-track musical excerpts. It consists of 30-second excerpts of 20 varied style songs. These excerpts have a stable tempo and present a varied distribution of tempi values. 40 beat annotations are available for each song.

The *rhythms* dataset was developed with the purpose of capturing the subjective pulse experience. To this end, we carried out an experiment were participants listened to 33 rhyhtmic passages of varying rhythmic complexity and were instructed to tap to a self selected beat. Participants were allowed to stop tapping if the beat was not clear enough or change their selected beat mid-trial. After each trial, participants rated how difficult the tapping task was with values between 1 (easy) and 5 (hard). The stimuli consisted of 11 rhythms from [14], 7 from [15], 5 were isochronous beats at 150, 200, 250, 500, 800 ms inter-beat intervals and 10 were new. 7 of the new stimuli were presented in increasing complexity order at the beginning of the experiment to familiarize the participants with the task. All other stimuli were randomized. With the exception of the isochronous stimuli, presentation inter-beat intervals varied between 450 and 550 ms avoiding having the same IBI in two consecutive trials. Each stimulus consisted of repeating a short rhythm the number of times required to last a minimum of 24 seconds. From 35 total participants, 30 remained after filtering participants that were deemed to not understand the concept of beat. They were selected as participants who replicated the stimulus instead of defining a beat in more than three trials. 6 participants were female, and 26 were male. Overall average age was 28.27 (sd = 7.94) and overall mean musical training was 4.85 years (sd = 3.90). For our evaluation, we will not consider the 5 isochronous stimuli in the dataset as these were not intended to evaluate rhythmic complexity.

Lastly, we use the *soundtracks* (ST) dataset used in [8], which is composed of 100 five-second excerpts of movie soundtracks, selected to cover a wide range of pulse clarity scenarios. From these, 15 excerpts were discarded as some metrics couldn't be computed for them because they provided too few beat events. Each track was rated by 25 musically trained participants in its beat clarity on a scale from 1 to 9, labeled from "unclear" to "clear". The mean clarity score is provided in the dataset [18].

Tracks in the MIREX and *rhythms* datasets have more than one annotation for each track. To obtain a single value for each track and category, the empiric pulse clarity value for a track is considered as the mean response of the subjects. Previously, pulse clarity values (tapping variability and self-reported) were z-standardized within participants.

As there are various types of annotations, we consider different interpretations of pulse clarity for each dataset. For the *rhythms* dataset we consider the answers to the tapping difficulty question and in the *soundtracks* dataset we use the "pulse clarity" reported answers. For both the MIREX and Experimental datasets, human tapping annotations are available. These consist of a list of moments in time where the person felt the underlying pulse. Using this information, we propose a tapping variability metric, "inter-tap-interval entropy" (ITI-E), to act as a proxy for pulse clarity. The computation is as follows: first the difference between subsequent taps is obtained. These differences are considered samples from the underlying subject's inter-tap interval distribution. A Gaussian kernel density estimator with a bandwidth of 5ms is fitted over the samples and 400 equidistant points considered from the range of 8 and 320 BPM are evaluated to obtain the density estimation. Lastly, the entropy over the density estimation is calculated as a means to capture its variability. In trials where less than five taps were produced, the metric was not considered reliable. For these cases, it was considered the participant had an unclear pulse precept and decided not to tap. The entropy value was replaced with the maximum entropy found for the participant. This methodology was verified by correlating the obtained values for the *rhythms* dataset and the reported tapping difficulty answers, obtaining an r coefficient of 0.81 with $p < 0.001$.

Calculating the entropy over the distribution as opposed to an approach based on standard deviation calculation was chosen because we consider the possibility that the distribution could be multi-modal, meaning that a subject tapped to two or more possible interpretations of the pulse in the same stimulus. In said case, the standard deviation would result in high variability, when in fact it could be argued that the tapping was precise for more than one pulse interpretation.

3.2 Model selection

For the evaluation of the proposed metrics we classify them in categories and select the best scoring metric in each one. The categories considered were DBN metrics, RNN metrics, output metrics and comparison metrics. This last one is comprised of the pulse clarity model proposed by Lartillot in [8] (`mirpulseclarity` function from MIRToolbox 1.7.2 [19] with parameters for `model=1`), the congruency score from the THT model presented in [9]

	MIREX	Rythms		ST	*Avg*
	ITI-E	Conf	ITI-E	PC	
Comp	0.550†	0.783	0.690	0.570	*0.648*
DBN	0.412‡	**0.912**	**0.775**	**0.860**	***0.740***
Out	**0.791**	0.858	0.769	0.388†	*0.701*
RNN	0.632	0.627	0.621	0.372†	*0.563*

Table 1. Test set scores for each category in every dataset. The selected metrics in the training process were: **THT-congruency** (Comp), **Viterbi max** (beat version) (DBN), **DBN entropy** (beat version) (Out), **cell state precision** (RNN). All correlations have p-values below 0.001, except for those with † where $0.05 > p > 0.001$, and ‡ where $0.1 > p > 0.05$.

and an additional interpretation of it called **THT entropy**, which is the same as **DBN entropy** but using the beat moments outputted by the THT model. These categories allow the evaluation of the two main questions this work proposed. The first one being if the interpretations made of the beat tracking model are comparable to the previous pulse clarity models. Furthermore it is of interest to know if the metrics derived from internal behaviour surpass the ones obtained by interpreting the output of the deep learning model.

We separate the data into train and test subsets to select the best metric per category. The train set is obtained by randomly selecting half of the *soundtracks* dataset. Selecting train data only from this dataset is motivated by the fact that the other two have few stimuli and partitioning them would increase the probability of losing representability in the subsets. In the training process we selected one metric from each category as the one that scored consistently higher when considering the correlation on 10 subsets of 80% of the training data. In Table 1, we report the test set scores for each metric selected in the training process.

From Table 1 we can observe that there is not a best model overall. Nonetheless it is remarkable that in two of the three datasets the DBN metrics section has the highest score, achieving r values over 0.77 with the **Viterbi Max** metric (2nd row on Table 1). Averaging all test set scores, this metric performed best. Comparing against the Comparison section, no proposed metric provides better results over all datasets. When comparing the metrics from the inner behaviour of the model versus those calculated using the output, neither the RNN or DBN categories surpass the Output metric selected in all datasets. This category was the second with highest average score with the **DBN entropy** metric, providing similar results to the **Viterbi max** metric.

4. CONCLUSIONS

In this paper we showed possible interpretations of pulse clarity from the inner workings of a beat tracking model. The developed metrics achieved comparable results with respect to previous works over distinct datasets, showing that there is relevant information in the analyzed beat tracking models. Specifically, the intuitive DBN based metrics performed considerably better compared to the RNN metrics. Nevertheless, when comparing the inner calculations of the model with simple transformations of the model's output, the inner calculations did not consistently yield better results. This indicates that, although useful, inspecting the inner behavior of the model may not be strictly necessary.

We evaluated the pulse clarity models on very different stimulus types: songs, rhythms and movie soundtracks. Results showed that both the developed and comparison models performed well on all datasets, even on those that had different types of stimulus than their original training data. This indicates that models can generalize well from one type of stimulus to another. Yet, some variability was present across datasets, inviting further evaluation of pulse clarity models on a broader spectrum of musical stimuli.

The pulse clarity metrics obtained from the beat tracking model presented correlations comparable with those of previous work. Given a more extensive dataset, with more musical styles and exploring more dimensions of pulse clarity, relevant pulse clarity predictors can be obtained. These would provide new tools for the musical psychology community. Overall, we propose that developments in models for music perception tasks can be repurposed for related tasks.

5. OPEN PRACTICES STATEMENT

A reference implementation written in `Python` is publicly available at `https://github.com/nPironio/maipc`, including all the metrics presented in this work.

6. REFERENCES

[1] W. T. Fitch, "Rhythmic cognition in humans and animals: distinguishing meter and pulse perception," *Frontiers in Systems Neuroscience*, vol. 7, 2013. [Online]. Available: https://doi.org/10.3389%2Ffnsys.2013.00068

[2] V. E. Gonzalez-Sanchez, A. Zelechowska, and A. R. Jensenius, "Correspondences between music and involuntary human micromotion during standstill," *Frontiers in Psychology*, vol. 9, p. 1382, 2018. [Online]. Available: https://www.frontiersin.org/article/10.3389/fpsyg.2018.01382

[3] B. Burger, M. R. Thompson, G. Luck, S. Saarikallio, and P. Toiviainen, "Music Moves Us: Beat-Related Musical Features Influence Regularity of Music-Induced Movement," no. July, 2012, pp. 183–187. [Online]. Available: http://icmpc-escom2012.web.auth.gr/sites/default/files/papers/183_Proc.pdf

[4] W. Trost, S. Frühholz, T. Cochrane, Y. Cojan, and P. Vuilleumier, "Temporal dynamics of musical emotions examined through intersubject synchrony

of brain activity," *Social Cognitive and Affective Neuroscience*, vol. 10, no. 12, pp. 1705–1721, 05 2015. [Online]. Available: https://doi.org/10.1093/scan/nsv060

[5] W. T. Fitch and A. J. Rosenfeld, "Perception and production of syncopated rhythms," *Music Perception: An Interdisciplinary Journal*, vol. 25, no. 1, pp. 43–58, 2007.

[6] B. H. Repp and Y.-H. Su, "Sensorimotor synchronization: A review of recent research (2006–2012)," *Psychonomic Bulletin & Review*, vol. 20, no. 3, pp. 403–452, Jun 2013. [Online]. Available: https://doi.org/10.3758/s13423-012-0371-2

[7] A. D. Patel, J. R. Iversen, Y. Chen, and B. H. Repp, "The influence of metricality and modality on synchronization with a beat," *Experimental Brain Research*, vol. 163, no. 2, pp. 226–238, May 2005. [Online]. Available: https://doi.org/10.1007/s00221-004-2159-8

[8] O. Lartillot, T. Eerola, P. Toiviainen, and J. Fornari, "Multi-feature modeling of pulse clarity: Design, validation, and optimization," in *In Proceedings of the International Symposium on Music Information Retrieval*, 2008.

[9] M. A. Miguel, M. Sigman, and D. Fernandez Slezak, "From beat tracking to beat expectation: Cognitive-based beat tracking for capturing pulse clarity through time," *PLOS ONE*, vol. 15, no. 11, pp. 1–22, 11 2020. [Online]. Available: https://doi.org/10.1371/journal.pone.0242207

[10] M. MCKINNEY, D. Moelants, M. DAVIES, and A. KLAPURI, "Evaluation of audio beat tracking and music tempo extraction algorithms," *JOURNAL OF NEW MUSIC RESEARCH*, vol. 36, no. 1, pp. 1–16, 2007. [Online]. Available: http://dx.doi.org/10.1080/09298210701653252

[11] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *Journal of Big Data*, vol. 3, no. 1, p. 9, May 2016. [Online]. Available: https://doi.org/10.1186/s40537-016-0043-6

[12] F. Krebs, S. Böck, and G. Widmer, "An Efficient State-Space Model for Joint Tempo and Meter Tracking." in *Proceedings of the 16th International Society for Music Information Retrieval Conference*. Málaga, Spain: ISMIR, Oct. 2015, pp. 72–78. [Online]. Available: https://doi.org/10.5281/zenodo.1414966

[13] S. Böck, F. Krebs, and G. Widmer, "Joint beat and downbeat tracking with recurrent neural networks," in *Proceedings of the 17th International Society for Music Information Retrieval Conference, ISMIR 2016, New York City, United States, August 7-11, 2016*, M. I. Mandel, J. Devaney, D. Turnbull, and G. Tzanetakis, Eds., 2016, pp. 255–261. [Online]. Available: https://wp.nyu.edu/ismir2016/wp-content/uploads/sites/2294/2016/07/186_Paper.pdf

[14] M. Miguel, M. Sigman, and D. F. Slezak, "Tapping to your own beat: experimental setup for exploring subjective tacti distribution and pulse clarity," Oct 2019. [Online]. Available: osf.io/7sqaw

[15] S. Böck, F. Korzeniowski, J. Schlüter, F. Krebs, and G. Widmer, "Madmom: A new python audio and music signal processing library," in *Proceedings of the 24th ACM international conference on Multimedia*. ACM, 2016, pp. 1174–1178.

[16] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[17] "Mirex beat tracking training dataset." 2006, https://www.music-ir.org/mirex/wiki/2019:Audio_Beat_Tracking.

[18] T. Eerola, "Ground-truth data for selected perceptual features," Jul 2020. [Online]. Available: osf.io/6wqh5

[19] O. Lartillot and P. Toiviainen, "A matlab toolbox for musical feature extraction from audio," in *International conference on digital audio effects*, vol. 237. Bordeaux, 2007, p. 244.