I GOT RHYTHM, SO FOLLOW ME MORE: MODELING SCORE-DEPENDENT TIMING SYNCHRONIZATION IN A PIANO DUET

Akira MAEZAWA (akira.maezawa@music.yamaha.com) [1]

[1] **Yamaha Corporation**, Hamamatsu, Japan

ABSTRACT

Automatic accompaniment systems synchronizing software playback with human musicians require a natural, easy-to-manipulate music synchronization model. However, existing methods cannot synchronize in a way aware of the musical context. This paper presents an interpretable model of the sensorimotor synchronization (SMS) model based on the user's music score. The model maps the music score to an easy-to-interpret linear model of phase and period correction using a deep neural network, allowing timing synchronization to be aware of the music score. The evaluation shows the proposed method achieves lower estimation error than a model that is unaware of the music score or one that is linearly dependent on the score.

1. INTRODUCTION

Automatic accompaniment is a technique that enables a machine to accompany a human musician's performance by synchronizing its playback to the human performer. It helps accompany users who cannot play in time or wish to express themselves through artistic timing fluctuations. For the machine to provide comfortable accompaniment for humans, the machine should follow the musician and synchronize naturally, as a human accompanist would do. This is challenging because playing in a musical ensemble is an intricate art: multiple musicians interact with each other in a way that is dependent on the musical context, perception, and performance. In practice, it is also a challenge to provide an expressive model whose behavior is easy to analyze and manipulate.

Existing accompaniment systems lacked a simple yet expressive accompaniment playback model that is aware of musical contexts. Earlier methods required the composer to specify how smoothly it should play an accompaniment snippet in response to the user's performance [1]. Some methods have used heuristics of performance of music score to adjust simple parameters for accompaniment playback [2, 3], but such a heuristic approach can potentially miss intricate factors in timing synchronization. Many models use rehearsals to learn a synchronization model [4, 5] but cannot generalize outside the rehearsed pieces. A recent method has arrived at a simple and interpretable model inspired by sensorimotor synchronization (SMS) [6], but it has fixed parameters or the response throughout the piece, unaware of the musical contexts.

This paper proposes a method for timing synchronization that takes into account musical contexts, can generalize across different pieces, and is simple to interpret and adjust. Inspired by models of SMS [7] that are simple and robust, we present an extension of such a model that is conditioned on a music score. The proposed SMS model is aware of the music score thanks to the conditioning on the score, which allows generalization through training. It is also interpretable and allows intuitive manual modifications thanks to the simple formulation of the SMS model.

2. RELATED WORK

2.1 Automatic accompaniment

In automatic accompaniment systems [1,3,4,6,8], the playback of the machines is synchronized to the human player. A model of feasible playback time sequence is assumed, and its parameters are learned through rehearsals. For example, a dynamic Bayesian network describing the timing of a performance of a known piece of music [5] or a linear dynamical system using score and performance features [4] can be used to train how each performer plays a given piece of music. Typically, these are confined to a given piece of music.

To generalize synchronization to new pieces, designing a general model of timing synchronization is necessary. A line of work uses performance features such as IOI stability to compute who leads an ensemble [2]. Another line of work uses the music score features such as note densities [3] for timing synchronization. These use features that have been heuristically designed. To circumvent the manual design of hand-crafted features, our previous work trained a deep neural network to predict the tempo curve given the current performance and the music score [9]. In this paper we apply this approach for timing coordination.

2.2 Music performance rendering

A related task is music performance rendering, which generates expressive parameters such as the velocity and onset timing deviations given a music score. Here, deep neural networks have been successful and modeling the mapping between the score and the performance [10–13]. We use deep neural networks to learn the sensorimotor synchronization parameters, which can be used to control machine playback in a human music ensemble.

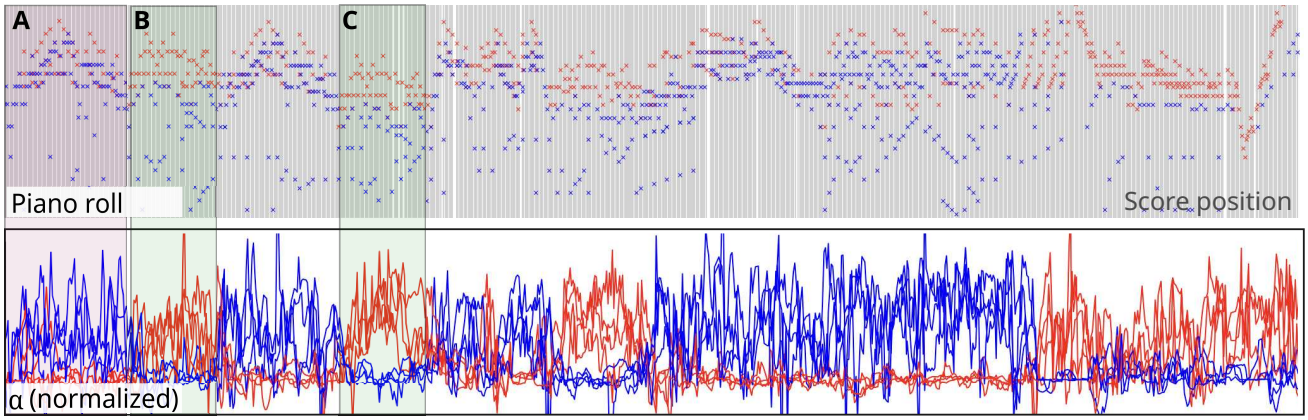

Figure 1. An example of normalized linear phase correction parameter α over four takes and the corresponding piano-roll (piano duet arrangement of Mendelssohn's "Song without words"). The red line represents piano 1, and the blue line represents piano 2. The roles of the two parts are interchanged in segments "A" and "B." Segments "B" and "C" are modulations of each other.

2.3 Sensorimotor Synchronization

Sensorimotor synchronization (SMS) is the phenomenon of human movement coordination in response to an external rhythm, which depends on various factors [14]. For computational modeling of SMS, many studies use linear models [7, 15, 16]. These explain period and phase work in terms of the asynchrony between humans and external stimuli, typically via low-order models like ARMA(1,2) under a restricted parameter space. Some studies implement such cognition-inspired models for SMS for automatic accompaniment systems [2, 6, 17]. Whereas these models assume that SMS model parameters are independent of the musical context, given that SMS depends on the rhythm [18], we hypothesize that the model parameters are better expressed by conditioning on the music score.

3. OUR METHOD

In our previous research [9], we have employed deep learning to extract features for estimating the tempo curve governed by a linear model. It offered the flexibility of a deep neural network while also providing an interpretable system. In this paper, we extend this idea to learning synchronization as well.

3.1 Model of synchronization

We introduce the synchronization model used in this paper and briefly explore the validity of assuming its parameters depend on the score.

3.1.1 Linear phase/period correction

We first consider the period correction model [16], which extends the linear phase correction model to handle tempo changes. It considers a case where a user is given beat stimuli to tap with an interval of S_k seconds for the kth stimulus, and the user would tap along with an error of A_k seconds. The model considers asynchrony to arise from (1) the motor noise $M_k \sim \mathcal{N}(0, \sigma_M^2)$, (2) the discrepancy between S_k and the expected beat period T_k, a probabilistic value with an underlying beat period t_k, $T_k \sim \mathcal{N}(t_k, \sigma_T^2)$, and (3) a corrective process based on recent asynchronies, that corrects beat phase and period according to parameters α and β, respectively:

$$A_{k+1} = (1-\alpha)A_k + M_{k+1} - M_k + T_k - S_k \quad (1)$$

$$t_k = t_{k-1} - \beta A_k. \quad (2)$$

For this study, we modify the original formulation to allow the inter-stimulus interval S_k and the corresponding T_k to scale by the notated note values. We thus describe the asynchrony also in terms of the notated note values d_k, as follows:

$$A_{k+1} = (1-\alpha)A_k + M_{k+1} - M_k + d_k(T_k - S_k). \quad (3)$$

The inter-stimulus interval is clear in tapping literature, but it is not so clear for automatic accompaniment, where one part may be silent. As an approximation, we take the scheduled playback times of the sequencer when using an automatic accompaniment system or the smoothed playback position of a recording when analyzing a human duet for training.

3.1.2 Does synchronization depend on musical context?

To demonstrate the effect the music score has on the linear synchronization parameters, we plot the period correction parameter α against four takes of an identical piano duet in Figure 1, played by one pair of advanced pianists. The figure shows that the region with α taking on more significant values is consistent with the musical role as implied by the music score (by comparing (A) and (B)) and is consistent with musically similar parts (by comparing (B) and (C)). It thus suggests that the music score is useful for inferring how performers synchronize with each other.

3.2 Parameter prediction model

We consider two models for predicting the parameters α, β, σ_M, σ_T. In both cases, the model takes as the input a two-channel binary piano-roll, $X(c, t, p) \in$

$\{-0.5, 0.5\}^{C\times T\times P}$, where c is the channel index, t is the frame index, p is the pitch index. Given a music score, the human part is assigned to $c = 0$ and the machine part to $c = 1$. The frame index t is computed by the number of 32nd notes elapsed since the beginning, and the pitch index p is computed by taking the pitch of each note event and taking the floor after dividing it by 8, chosen for mathematical convenience, with $P = 16$. The feature thus represents a rough indication of the timing and the register of the current note events.

First, we consider a *linear regression model*, which estimates the parameters by multiple regression that takes the piano roll centered about the current position, with a radius of two beats. For the kth onset time t_k, we compute X_k as a slice of $X(c, t, p)$ centered about t_k with a window of two beats, *i.e.*, from $t = t_k - 16$ to $t_k + 16$. In other words, for the input feature X_k for kth note onset time, we estimate the parameters as follows:

$$[\alpha_k, \beta_k, \sigma_{M,k}, \sigma_{T,k}]^T = W\mathrm{vec}(X_k) + b. \qquad (4)$$

where $W \in \mathbb{R}^{4\times \mathrm{length}(\mathrm{vec}(X_k))}$ and $b \in \mathbb{R}^4$.

Second, we consider a *deep neural network* (DNN) for parameter estimation. It also accepts the piano roll X centered about the current position with a radius of two beats. For the kth onset in the score, we split X_k into X_k^{pre} and X_k^{post}, which are the X_k computed before and after the onset frame t_k. The two features X_k^{pre} and X_k^{post}, respectively, then undergo four layers of a series of convolutional networks with a kernel size of 2 with a dilation factor of 2^l, where l is the zero-indexed layer index, followed by batch normalization, exponential linear unit (ELU) activation and dropout with dropout probability of 0.05. The number of output channels of each layer is $[128, 256, 256, 256]$. The output is average pooled over the frame and pitch axes to yield a pair of 256-dimensional vectors c_k^{pre} and c_k^{post}. These are concatenated with the piano roll evaluated exactly at the kth onset frame, $X(:, t_k, :)$, to arrive at a context vector c_k. It then undergoes three layers of linear layers with output channels of 1000, 200, and 4 with dropout and ELU nonlinearity, except the last layer, to arrive at $[\alpha, \beta, \sigma_M, \sigma_T]$.

4. EVALUATION

4.1 Dataset

We have acquired a dataset of piano duets played by an advanced-level piano duet group. Four-hands arrangements were obtained for the fourteen pieces listed in Table 1, published by Print Gakufu [1] . For each piece, the pianists had time to study the score beforehand and rehearse before recording. Each piece was taken at least twice, once asking the player to play normally and another with a different expression. Some pieces were asked to be played to exaggerate expressions or cues for synchronization. An annotator with musical training aligned the takes to the score so that the click track associated with the score sounds appropriate for the performance. The data was recorded in two recording sessions, approximately one year apart.

[1] Print Gakufu: `https://www.print-gakufu.com/`

Composer	Title	Genre	Takes
Hakase	Jonetsu Tairiku	Popular	4
Menken	Beauty and the Beast	Popular	4
Bizet	Carmen	Classical	2
Borodin	Polovetsian Dances	Classical	5
Brahms	Hungarian Dance 5	Classical	3
Debussy	Bateau	Classical	2
Elgar	Salut d'Amor	Classical	2
Faure	Berceuse	Classical	2
Mascagni	Cavalleria Rusticana	Classical	3
Mendelssohn	Song Without Words	Classical	4
Pachelbel	Canon	Classical	3
Saint-Saens	Le Cygne	Classical	2
Tchaikovsky	Barcarolle	Classical	5
Tchaikovsky	Troika	Classical	6

Table 1. List of the recorded pieces and the number of takes.

To extract the synchronization parameters, we assumed that one player serves as the human while another serves as the timekeeper. For each distinct note onset event written in the score, a corresponding human onset was obtained by manually aligning to the score and finding the mean onset time of the notes played within a window of the 32nd note of the aligned onset time, repeatedly increasing this radius by 1.1x if no onset exists in the window, and ignoring the note if the window exceeds 0.3 seconds. When the two players do not play simultaneously, we assume that, for the player with no onset, the onset timing is linearly interpolated between the adjacent note onset times of that player. The parameters are estimated using the bGLS method for phase correction models [7], considering the introduction of note values in our formulation. The estimation requires multiple asynchronies, so we compute the parameter by aggregating the current asynchrony and ten neighboring asynchronies. We further smooth the parameters by a moving average window with 4 note onset events to smooth outlier estimates.

4.2 Experiment 1: Parameter estimation accuracy for piece-level cross-validation

We evaluate both the proposed linear (denoted as method **linear**) and DNN models (**DNN**) and two baseline methods that do assume a relationship between the score and SMS parameters. The first baseline estimates a single set of parameters α, β, σ_M, σ_T using the entire *training* data (**uniform-train**), and the second baseline estimates a single set from the entire *validation* data (**uniform-valid**), which can be considered as the theoretical maximal performing method, assuming the independence of SMS parameters on the score.

The models were trained with leave-one-out cross-validation (14-fold cross-validation) at a piece level so that the piece used for training would not be used for evaluation. The models were trained to minimize the mean absolute error (MAE) using Adam. The MAEs of the estimated parameters were then computed for each of the validation folds, and their averages were taken as the metric.

Table 2 shows the result. The proposed method consistently outperforms the baselines. Namely, for α, the MAE

Condition	α	β	σ_T	σ_M
Uniform-train	7.48E-2	2.63E-3	1.89E-2	2.74E-2
Uniform-valid	7.27E-2	2.30E-3	1.81E-2	2.45E-2
Linear	6.90E-2	2.18E-3	1.52E-2	2.46E-2
DNN	**6.68E-2**	**1.70E-3**	**1.38E-2**	**2.33E-2**

Table 2. MAE of the synchronization parameter estimates (piece-level cross-validation).

Condition	α	β	σ_T	σ_M
Uniform-train	7.43E-2	2.85E-3	2.05E-2	2.76E-2
Uniform-valid	7.43E-2	2.84E-3	1.71E-2	2.74E-2
Linear	6.56E-2	**1.77E-3**	1.39E-2	2.42E-2
DNN	**5.65E-2**	1.80E-3	**1.19E-2**	**1.82E-2**

Table 3. MAE of the synchronization parameter estimates (player-level cross-validation).

of the **DNN** method is reduced by 11% compared to uniform, β by 35%, σ_T by 27%, and σ_M by 15%. Incorporating the score is advantageous, as can be seen by comparing **linear**, a simple linear model, and **uniform-valid**, the minimum attainable MAE when ignoring the score. More elaborate models like **DNN** can attain even lower performance. The results are significant for almost all pieces and parameters, using a two-sided t-test with a significance level of 0.01. For condition **linear**, one piece did not show significance for α, one piece for β, one piece for σ_M, and three pieces for σ_T. For condition **DNN**, two pieces did not show significance for σ_M and three pieces for σ_T.

Figure 2 shows an example of the outputs. The ground-truth data contain large variabilities throughout the piece, which cannot be expressed with a constant estimate. Both linear and DNN models capture some of the variabilities, but the linear model tends to revert to the uniform baseline, whereas the DNN model can have lower values.

4.3 Experiment 2: Parameter estimation accuracy for player-level cross-validation

We have repeated the previous experiment, except the cross-validation was performed at the player level. In other words, the piano roll and the SMS parameters obtained from one pianist in the duo were used to train each model to evaluate the MAE of the data from the other pianist and vice versa.

Table 3 shows the result. A similar tendency as the previous experiment can be seen, where the score-dependent parameter estimates have lower MAEs. The differences in the MAE were significant for all pairs of conditions, using a t-test with a significance level of 0.01.

The drop in MAE for **linear** and **DNN** is more pronounced here, suggesting the importance of learning the score over the individual player to estimate the SMS parameters. To elaborate, it is common in a piano duet for the two players to switch musical roles, resulting in the two parts having similar scores dispersed throughout the piece. Thus during training, the model in this experiment was allowed to see a score similar to that in the validation data but not the player in the validation set, whereas the model in the previous experiment was allowed to see the performer in the validation data but not the validation piece. Thus, the fact that MAE is lower for this experiment suggests that it is the learning of the music score (by an arbitrary pianist) that is more important than the learning of the individual player (by an arbitrary piece).

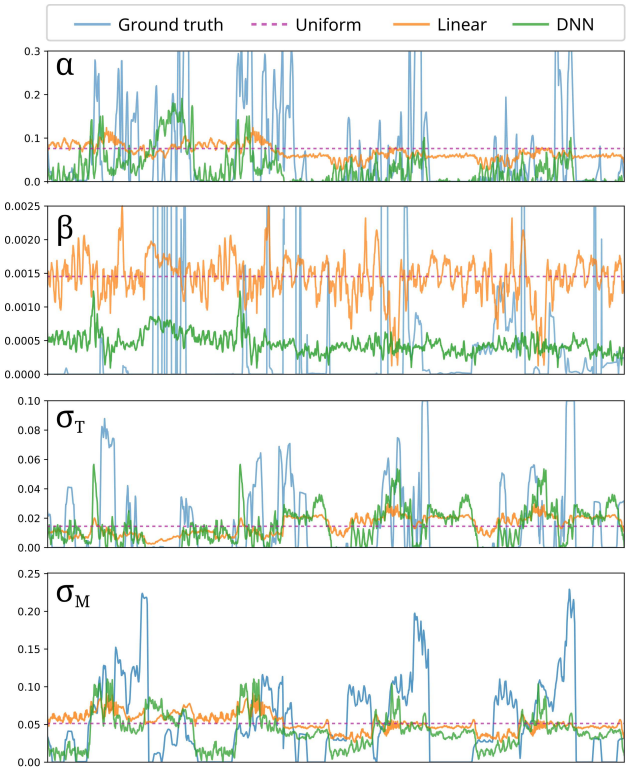

Figure 2. An example of estimated outputs versus the ground truth. Here, "Uniform" refers to condition **uniform-train**, which is almost identical to **uniform-valid** for this piece.

5. DISCUSSION

We have presented a model of SMS that is conditioned on the music score, allowing computers to be mindful of the music score when synchronizing with a human player. A deep neural network model has shown potential in this direction. Due to the simplicity of the parameters, it is possible to assess the model's behavior and combine the insights from the music cognition literature. For example, it has been reported that the behavior of human performers remains consistent at a range of α between 0 and 1 [19], so it is sensible to clip the generated values to ensure an accompaniment system behaves stably for human performers.

The model presented in this method has been utilized as a stochastic state space model for automatic accompaniment systems in a few professional piano duet performances and music installations for the general public. The system, in general, performed stably as long as the score follower functioned properly and the user played more or less within the rehearsed tempo. Strong asynchrony occurs when the user (1) plays with lots of abrupt timing pauses, such as when sight-reading or has motor disabilities, or (2) shows

prominent tempo expression, such as in a ralletando. For the first issue, a better model of non-expert piano performance will become necessary. For the second issue, combining a long-term model of tempo prediction with a local SMS model might be a possible direction.

Our study has a few limitations. First, it cannot express variability in the SMS parameters within a given performance instance. While it is reasonable to assume that the general approach to synchronization between human players remains consistent for a particular music score segment, no two performances are temporally identical, and therefore, no two performances have identical time coordination. This suggests that SMS is also a function of the instantaneous interaction or some auto-regressive stochastic process, which our model currently ignores. Second, since we only had access to one pair of pianists, the across-player cross-validation represents the behavior of players in a single ensemble; it may be possible that different pairs of pianists, or the pianist used in this study paired with another pianist may behave differently. For example, beginners with little exposure to music will mostly likely have different motor noise σ_M, or have different α due to the incapability to attend and listen to others. Our result nonetheless showed that, given a piano duo, it is possible to generalize timing synchronization to unseen pieces by learning from their performances.

6. CONCLUSION

This paper presented an SMS model conditioned on the music score. We demonstrated the effectiveness of incorporating the music score to improve the quality of SMS parameter estimation. Future work includes incorporating music performance features and improving synchronization robustness by piano players of wide skill levels.

7. REFERENCES

[1] A. Cont, J. Echeveste, J.-L. Giavitto, and F. Jacquemard, "Correct Automatic Accompaniment Despite Machine Listening or Human Errors in Antescofo," in *Proc. International Computer Music Conference 2012*, Sep. 2012.

[2] T. Mizumoto, T. Ogata, and H. G. Okuno, "Who is the leader in a multiperson ensemble? Multiperson human-robot ensemble model with leaderness," in *Proc. 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Oct. 2012, pp. 1413–1419.

[3] A. Maezawa and K. Yamamoto, "MuEns: A Multimodal Human-Machine Music Ensemble for Live Concert Performance," *Proc. 2017 CHI Conference on Human Factors in Computing Systems*, pp. 4290–4301, 2017. [Online]. Available: http://doi.acm.org/10.1145/3025453.3025505

[4] G. Xia, Y. Wang, R. Dannenberg, and G. Gordon, "Spectral Learning for Expressive Interactive," in *Proc. 16th International Society for Music Information Retrieval Conference*, 2015, pp. 816–822.

[5] C. Raphael, "A Bayesian Network for Real-Time Musical Accompaniment," in *Advances in Neural Information Processing Systems*, 2001, pp. 1433–1439.

[6] C. Cancino-Chacón, S. Peter, P. Hu, E. Karystinaios, F. Henkel, F. Foscarin, and G. Widmer, "The accompanion: combining reactivity, robustness, and musical expressivity in an automatic piano accompanist," in *Proc. International Joint Conference on Artificail Intelligence 2023*, 2023.

[7] N. Jacoby, N. Tishby, B. H. Repp, M. Ahissar, and P. E. Keller, "Parameter estimation of linear sensorimotor synchronization models: Phase correction, period correction, and ensemble synchronization," *Timing Time Perception*, vol. 3, no. 1-2, pp. 52 – 87, 2015.

[8] C. Raphael, "Music Plus One: A System for Expressive and Flexible Musical Accompaniment," in *Proc. International Computer Music Association*, 2001, pp. 159–162.

[9] A. Maezawa, "Deep Linear Autoregressive Model for Interpretable Prediction of Expressive Tempo," in *Proc. 16th Sound and Music Computing Conference*, 2019.

[10] D. Jeong, T. Kwon, Y. Kim, K. Lee, and J. Nam, "VirtuosoNet: A hierarchical RNN-based system for modeling expressive piano performance," in *Proc. International Society for Music Information Retrieval Conference*, 2019.

[11] A. Maezawa, K. Yamamoto, and T. Fujishima, "Rendering music performance with interpretation variations using conditional variational RNN," in *Proc. International Society for Music Information Retrieval Conference*, 2019.

[12] S. Rhyu, S. Kim, and K. Lee, "Sketching the Expression: Flexible Rendering of Expressive Piano Performance with Self-Supervised Learning," in *Proc. International Society for Music Information Retrieval Conference*, 2022.

[13] I. Borovik and V. Viro, "ScorePerformer: Expressive piano performance rendering with fine-grained control," in *Proc. International Society for Music Information Retrieval Conference*, 2023.

[14] 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

[15] D. Vorberg and H.-H. Schulze, "Linear phase-correction in synchronization: Predictions, parameter estimation, and simulations," *Journal of Mathematical Psychology*, vol. 46, no. 1, pp. 56–87, 2002.

[16] H.-H. Schulze, A. Cordes, and D. Vorberg, "Keeping synchrony while tempo changes: Accelerando and ritardando," *Music Perception: An Interdisciplinary Journal*, vol. 22, no. 3, pp. 461–477, 2005.

[17] K. Armstrong, J.-X. Huang, T.-C. Hung, J.-H. Huang, and Y.-W. Liu, "Real-Time Piano Accompaniment Using Kuramoto Model for Human-Like Synchronization," in *Proc. 16th International Symposium on Computer Music Multidisciplinary Research*, 2023. [Online]. Available: https://zenodo.org/records/10113439

[18] J. S. Snyder, E. E. Hannon, E. W. Large, and M. H. Christiansen, "Synchronization and Continuation Tapping to Complex Meters," *Music Perception*, vol. 24, no. 2, pp. 135–145, 2006.

[19] B. H. Repp and P. E. Keller, "Sensorimotor synchronization with adaptively timed sequences," *Human Movement Science*, vol. 27, no. 3, pp. 423–456, 2008.