A Dynamic Systems Approach to Modeling Human–Machine Rhythm Interaction

Zhongju Yuan, Wannes Van Ransbeeck, Geraint A. Wiggins, and Dick Botteldooren, *Senior Member, IEEE*

***Abstract*—Rhythm is an inherent aspect of human behavior, present from infancy and embedded in cultural practices. At the core of rhythm perception lies meter anticipation, a spontaneous process in the human brain that typically occurs before actual beats. This anticipation can be framed as a time series prediction problem. From the perspective of human embodied system behavior, although many models have been developed for time series prediction, most prioritize accuracy over biological realism, contrasting with the natural imprecision of human internal clocks. Neuroscientific evidence, such as infants' natural meter synchronization, underscores the need for biologically plausible models. Therefore, we propose a neuron oscillator-based dynamic system that simulates human behavior during meter perception. The model introduces two tunable parameters for local and global adjustments, fine-tuning the oscillation combinations to emulate human-like rhythmic behavior. The experiments are conducted under three common scenarios encountered during human-machine interaction, demonstrating that the proposed model can exhibit human-like reactions. Additionally, experiments involving human-machine and interhuman interactions show that the model successfully replicates real-world rhythmic behavior, advancing toward more natural and synchronized human-machine rhythm interaction.**

***Index Terms*—Dynamic system, human-machine interaction, reservoir computing, time series prediction.**

I. Introduction

RHYTHM is a fundamental aspect of human behavior, evident from early infancy [1] to diverse cultural expressions [2]. It represents periodic patterns or cycles [3], forming the foundation of structured sequences like musical notes. Rhythm perception involves the *tactus*, a basic beat within a specific frequency rangeand the *meter*, a perceived temporal structure encompassing the tactus frequency [4], [5].

Received 31 October 2024; revised 16 January 2025; accepted 25 February 2025. Date of publication 25 March 2025; date of current version 24 April 2025. This work was supported in part by BOF under Grant BOF/24J/2021/246; in part by the WithMe FWO under Grant 3G043020; and in part by the Flemish AI Research Programme. This article was recommended by Associate Editor J. Sun. *(Corresponding authors: Geraint A. Wiggins; Dick Botteldooren.)*

Zhongju Yuan, Wannes Van Ransbeeck, and Dick Botteldooren are with the WAVES Research Group, Ghent University, 9000 Gent, Belgium (e-mail: zhongju.yuan@ugent.be; wannes.vanransbeeck@ugent.be; dick.botteldooren@ugent.be).

Geraint A. Wiggins is with the AI Lab, Vrije Universiteit Brussel, 1050 Brussel, Belgium, and also with the Electrical Engineering and Computer Sciences, Queen Mary University of London, E1 4NS London, U.K. (e-mail: geraint.wiggins@vub.be).

This article has supplementary material provided by the authors and color versions of one or more figures available at https://doi.org/10.1109/TCYB.2025.3547216.

Digital Object Identifier 10.1109/TCYB.2025.3547216

Humans synchronize rhythms by predicting timing and stress patterns, a process essential for coordinated interactions [6], [7]. Predictive coding models can simulate this ability, offering insights into the neural mechanisms underpinning rhythm synchronization [8], [9]. Neural resonance theory (NRT) [10] posits that rhythm arises naturally from coupled nonlinear oscillators, inspiring oscillator-based models for rhythm perception.

However, popular sequential learning models, such as transformer-based architectures, can not be viewed as oscillator-based models, where neurons operate independently within each layer. And commonly used time series prediction models, such as long short-term memory (LSTM) networks and gated recurrent units (GRUs), exhibit internal oscillatory behavior that predominantly falls within a limited frequency range. In other words, their oscillatory dynamics are typically confined to low frequencies, which contrasts with the broader range of frequencies observed in NRT.

Reservoir computing is a promising framework for constructing biologically plausible oscillator networks [11]. These networks efficiently handle time-series data by leveraging fixed internal connections and trainable output layers [12], [13]. Building on the biologically inspired principles of reservoir computing and NRT, recent studies have explored synaptic plasticity as a key mechanism for tuning network dynamics [14], [15]. Notably, prior research has explored advanced echo state network (ESN) architectures by integrating structural and intrinsic plasticity mechanisms, such as synaptic adaptation and multiclustered configurations, to improve learning capabilities and performance in dynamic and nonlinear systems. However, these approaches are not well-suited for the rhythmic prediction tasks addressed in this article. And the plasticity tuning mechanisms rely on pruning or predefined clustering rules that lack transparency. To overcome these limitations, we propose a physics-inspired model grounded in wave equations.

Our model employs a wave-equation-based connection matrix to generate oscillators across human-relevant frequencies. This design enhances biological plausibility and interpretability [16]. Our model features sparse connectivity, local delays, and frequency ranges aligned with human neural dynamics. Key parameters regulate oscillation persistence and stimulus tuning, facilitating rhythmic behaviors that mirror human capabilities.

For single-channel rhythm perception tasks, our model achieves predictive coding of meter frequencies within human-perceptible ranges [17], [18]. A fine-tuning mechanism adjusts

response timing, while delayed feedback sustains rhythmic output, enabling closed-loop behavior akin to human interaction [19]. In dual-channel scenarios, the model adapts to synchronization and deviations in collaborative rhythms.

Tests comparing the model against human experiments demonstrated realistic imperfections during interactions, highlighting its capacity to simulate diverse rhythmic behaviors. Customizable settings further align the model with individual tendencies in human-to-human interactions.

Our primary contributions are:

1) We propose the first biologically plausible, physics-based oscillation model for simulating human rhythm perception and interaction tasks.
2) We introduce an effective fine-tuning method to improve the models oscillation combination performance, along with a delayed feedback mechanism that sustains rhythmic output even after the input has stopped.
3) The model-generated behaviors closely resemble human behavior in real-world human-machine rhythm interaction scenarios.

II. Related Work

Whether it is an infant's interaction with rhythm [20], [21] or the performance of a professional symphony [22], rhythm perception has been a longstanding subject of interest in neuroscience [23], [24]. Notably, the tendency of embodied systems to anticipate the beat earlier than the actual beat timing [18] has framed this problem as a predictive coding task. This section is structured from two distinct perspectives aimed at addressing the problem: the neuroscience approach Section II-A and the deep learning approach Section II-B designed for prediction tasks.

A. Neuroscience Perspective

A growing body of research has explored how the human brain processes and produces musical rhythm, combining insights from neural imaging and neuronal dynamics. Neuronal oscillations in specific frequency ranges have been implicated in rhythm perception and production. For example, in [25], magnetoencephalography (MEG) was used to demonstrate that beta oscillations are closely linked to auditory-motor synchronization, highlighting their critical role in timekeeping during rhythmic perception. Further, neuroimaging studies employing functional magnetic resonance imaging (fMRI) and electroencephalography (EEG) have investigated the brain regions involved in rhythm perception and production. These studies identified the basal ganglia as a key region for regular, beat-based rhythm processing, while the cerebellum plays a crucial role in managing more complex rhythmic sequences [26]. Similarly, [27] demonstrated the involvement of the dorsal premotor cortex in coordinating motor responses to rhythmic stimuli, suggesting that a broader network of motor and auditory regions underpins rhythm perception and production.

From a neuron-level perspective, studies have investigated how groups of neurons encode rhythmic patterns. Neuropsychological and neuroimaging research has revealed distinct pathways originating from the primary auditory cortex, projecting to various targets [28], [29]. Some neurons are specifically tuned to the fundamental frequencies of complex tones, providing insights into how rhythmic structures are processed granularly [30].

Additionally, some studies have examined EEG activity during human tapping tasks involving binary and ternary meter rhythms [31], [32], revealing specific frequency activations associated with each meter type. However, there is a lack of comprehensive work on building neural network models to simulate the oscillator networks in the human brain, especially in the context of human-machine rhythm interaction.

B. Deep Learning Perspective

Deep learning has significantly advanced time series forecasting [33], [34], with Transformer architectures [35] achieving state-of-the-art performance in various applications [36]. Transformer-based models, such as LogSparse Transformer [37] and FEDformer [38], employ encoder-decoder structures and innovations like convolutional self-attention and frequency decomposition. Despite their success, these models are costly to train and lack interpretability.

Recurrent neural networks (RNNs), including LSTM [39] and GRU [40], are effective for time series prediction due to their temporal memory capabilities [41]. Enhancements like attention mechanisms and diffusion models improve their predictive accuracy [42]. However, these networks lack biological interpretability, as they do not exhibit oscillator characteristics, which are essential for rhythm perception in neuroscience [43], [44]. Oscillator networks better capture cortical music entrainment [45].

Reservoir computing, with its inherent oscillatory properties, excels in time series prediction [46]. Fine-tuning parameters like leaky rate and spectral radius enables accurate predictions of chaotic systems [46], [47]. Structured reservoirs, such as antisymmetric matrices [48], delay line models [49], concentric reservoirs [50], and hierarchical ESNs [51], further enhance performance by capturing higher-order features. Most reservoir adjustments rely on eigenvalues and basic mathematical properties.

III. Methods

The primary aim of this model is to predict rhythmic beat occurrences during interactions, capturing human-like precision and imprecision in a biologically plausible manner.

We propose a single-layer oscillator-based neural network that models the temporal dynamics of rhythms. This type of network inherently exhibits multiple damped resonances, allowing the replication of slow temporal behaviors through the interactions of fast-responding neurons. This property aligns with biological plausibility, serving as the core framework for the rhythm interaction task (outlined in Section III-A).

Our approach introduces two key innovations. First, we develop a topology-preserving neural network architecture (detailed in Section III-B). This design ensures that the flow of the structure and the generation of connectivity weights

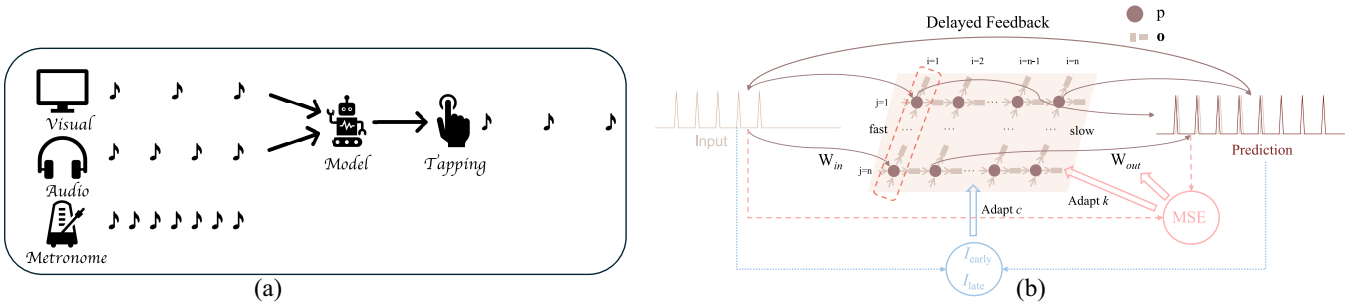


Fig. 1. Illustration of a typical task for the model and the model structure. (a) Illustration of a typical task: the model is primed with one rhythm (noted as visual here, referring to the human priming further on in this article) while exposed to another rhythm based on the same meter; its task is to predict the beats of the primed rhythm and continue doing so after the primer fades. (b) Flow of information in the model follows a 2-D structure. The neural network has two components p, and $\mathbf{o}$. The indices, i and j refer to spatial locations of the neurons. Rhythmic inputs are provided to the first row of neurons, denoted as p, at the fast end ($i = 1$). The models output is a linear combination of the activities from all neurons in p. The model aims to predict Δt ahead. The internal connections and output weights are adapted according to different measurements, and a feedback loop delayed by Δt is utilized to refine predictions.

(described in Section III-B1) are maintained throughout. The new architecture allows network tuning during application through parameters that have biologically interpretable and predictable physical impacts (Section III-B2).

Second, we implement a rapid adaptation mechanism to improve the models ability to learn new tasks, particularly in two-channel interaction scenarios. With the proposed delayed feedback, the model will continue interaction even without input. This enhancement, which optimizes the response of the model to new tasks, is discussed in Section III-C. To introduce greater flexibility during interactions, real human behaviors are utilized to customize the output weight matrix, as described in Section III-D.

A. Framework for Rhythm Interaction Task

A typical task for our model is illustrated in Fig. 1(a): beat a rhythm based on another aurally presented rhythm with the same underlying meter. To instruct the model what rhythm it is required to beat, a preparation phase is needed. As this will be done through visual stimulation in the human experiment, we refer to it as the visual input in Fig. 1(a), yet as far as the model is concerned it consists of a second periodic signal envelope. Computer models are inherently very fast and could simply follow, but for the system to be human-like and even implementable in a robot it needs to be predictive. Here, we selected a forward prediction of $\Delta t = 200$ ms.

To create this predictive behavior, an oscillating system is needed, as shown in Fig. 1(b). A network of connected neurons is chosen as a complex resonating system containing thousands of degrees of freedom. To select the desired rhythm, the internal states of this system are combined using trainable output weights. Training is performed during the preparation phase and could represent the biological process of selecting the appropriate envelope following component from the disentangling of the rhythm by the auditory system.

The outcome of this training process is a generative model designed to simulate sequential human behavior in rhythmic cognitive tasks. Central to the proposed model is a latent dynamical system characterized by state variables $\mathbf{h}_t$. The evolution of the latent state is governed by the following dynamics:

$$\mathbf{h}_t = f_\theta(\mathbf{h}_{t-1}, \mathbf{x}_t, \xi_t) \tag{1}$$

where $\mathbf{x}_t$ represents the inputs (primer, auditory reference, feedback) at time t, ξ_t is the noise term at time t, and f_θ denotes the dynamics function. Here, f_θ is modelled as a highly expressive physical-inspired neural network structure (Section III-B), θ denotes all of the parameters of the proposed model.

The representation of the input employs a smooth pulse to accentuate the occurrence or lack of a beat at any given moment. The injection of zero-mean Gaussian noise ξ_t at each time step is essential. From a physical perspective, the noise will trigger the oscillating system to exhibit "natural" metronomes. Additive noise is also biologically plausible due to spontaneous emission of the (auditory) neurons. As when stimulation is absent, some activity is expected to emerge. This could lead to spontaneous oscillations at natural eigenfrequencies of the system. During quiet epochs, the system could thus move to such behavior just like humans tend to move toward their "own beat". Here, it is assumed that spontaneous emission by neurons could trigger such behavior. This would lead to an additional noise term. Connections in our model do not represent individual neurons but groups of neurons. Hence, inhibition after recent firing and spontaneous firing are expected to interact in a complex way. Hence, we make the noise independent of the current state of the neuron and with zero mean. It is due to the nonlinearity introduced by $f(\cdot)$, the noise will have less pronounced effects during time intervals where the activation is high. According to the findings in [52], internal oscillations in the mammalian brain exhibit a flat power spectrum below 10 Hz, which corresponds to Gaussian white noise. This frequency range (0–10 Hz) falls within the band of human-perceptible rhythmic frequencies.

Model outputs, denoted by $\mathbf{y}_t$, are produced through a linear combination acting as a representation of mixed selectivity [13], derived from the activity of neurons within the model. The motor behavior is not explicitly modelled yet it is assumed that this introduces a delay of Δt. Therefore, the model is trained to predict upcoming beats Δt ahead of their occurrence. For training the output layer weights, we employ the mean

squared error (MSE) metric to ascertain the proximity of the model's predictions to the target behavior.

B. Neural Dynamic System Structure

1) Connectivity Weights: The proposed model is built on the combination of spontaneously generated internal oscillations, which predict outcomes in a recurrent manner. Its hidden states change according to the current input and the hidden states from the previous time step, which follows the equations:

$$\begin{aligned}\mathbf{h}_t &= (1-\alpha)\mathbf{h}_{t-1} \\ &\quad + \alpha f(\mathbf{W}_{\text{in}}\mathbf{x}_t + \mathbf{W}\mathbf{h}_{t-1} + \xi_t) \\ \hat{\mathbf{y}}_{t+\Delta t} &= \mathbf{W}_{\text{out}}\mathbf{h}_t \end{aligned} \tag{2}$$

where $\mathbf{W}$ is a sparse matrix defining the connectivity of the network, $\mathbf{W}_{\text{in}}$ is the input weight, and $\mathbf{W}_{\text{out}}$ is the output weight matrix, and α is the leakage rate of the model. $f(\cdot)$ is a nonlinear function, for which $\tanh(\cdot)$ is used in this article. $\mathbf{x}_t$ is the input signal at time step t, and $\mathbf{h}_t$ is the hidden state at time step t, and $\hat{\mathbf{y}}_{t+\Delta t}$ is the output of the model at time step t, which is the prediction after Δt ms. The expected prediction should satisfy $\hat{\mathbf{y}}_{t+\Delta t} = \mathbf{x}_{t+\Delta t}$, and $\Delta t = n\delta t$. n is the number of time steps predicted ahead, and δt denotes the simulation time step (equals to the downsample factor in this article).

The input weight matrix $\mathbf{W}_{\text{in}}$ is randomly generated and remains unchanged throughout the process. As illustrated in Fig. 1(b), the input is applied only to the first row ($i = 1$). The Gaussian noise ξ_t is regenerated every time step. The connectivity weight matrix $\mathbf{W}$, represented by arrows connecting neurons in Fig. 1(b) and (c), is typically adjusted to ensure that all its eigenvalues lie within the unit circle in the z-domain [47], thus ensuring the stability of the dynamic system in the linear, low-amplitude regime. The output weight matrix $\mathbf{W}_{\text{out}}$ is connected to all p neurons, as depicted in Fig. 1(b). In this study, we employ stochastic gradient descent (SGD) to minimize the MSE between the prediction $\hat{\mathbf{y}}_{t+\Delta t}$ and target $\mathbf{x}_{t+\Delta t}$ signals during training that is based on a large number of rhythmic beats that could theoretically be encountered in music.

For the problem at hand, predicting the occurrence of rhythmic beats in a human-like way, the poles in the z-domain of the $\mathbf{W}$ matrix and thus the resonances in the random reservoir are not optimally placed: (1) they span a frequency range that does not match human capabilities; (2) many of them are too much damped. To overcome this problem, we propose a novel oscillator-based structure designed following physical principles. To simplify the tuning of $\mathbf{W}$, we design it based on a 2-D finite-difference time-domain (2-D-FDTD) computational approximation of the linearized Euler equations [53] for wave propagation in a medium with randomly generated properties. Because this system results in local connections, it has a clear topology which allows crafting connections from input or outputs to areas showing specific dynamics. Starting from the wave equations (3) where c is the wave speed and k is a damping (amplification if negative) factor, and p and $\mathbf{o}$ are proportional to pressure and velocity

$$\begin{aligned}\frac{\partial p}{\partial t} + c^2\nabla\cdot\mathbf{o} &= 0 \\ \frac{\partial \mathbf{o}}{\partial t} + \mathbf{k}\cdot\mathbf{o} + \nabla p &= 0. \end{aligned} \tag{3}$$

The simplest FDTD model, a staggered grid, central differences, and an explicit time stepping approximation of these equations leads to their discretised form

$$\begin{aligned} p_{i,j}(t+\delta t) &= p_{i,j}(t) + c_{i,j}^2\delta t * \left(o_{x,i+1,j} - o_{x,i,j}\right)/\delta x \\ &\quad + c_{i,j}^2\delta t * \left(o_{y,i,j+1} - o_{y,i,j}\right)/\delta y \\ o_{x,i,j}(t+\delta t/2) &= \frac{1-k_{i,j}\delta t/2}{1+k_{i,j}\delta t/2}o_{x,i,j}(t-\delta t/2) \\ &\quad + \frac{\delta t}{\delta x\left(1+k_{i,j}\delta t/2\right)}\left(p_{i,j} - p_{i-1,j}\right) \end{aligned} \tag{4}$$

where the indices, i and j refer to spatial locations and the time dependence has been omitted on the right hand side of the equation. A similar equation holds for $\mathbf{o}_y$. Stability is guaranteed by keeping the Courant number, which relates the δt to δx and δy, smaller than 1.

The two groups of unknowns could be interpreted as two types of artificial neurons in a network as in Fig. 1(b): one is the primary neuron denoted as $p_{i,j}$, and the other is the intermediate neuron, labeled as $o_{x,i,j}$ or $o_{y,i,j}$. These can be grouped in a hidden state matrix $\mathbf{x}$ like in (2). As p, $\mathbf{o}$ are coupled locally and sparsely the coupling matrix $\mathbf{A}$ derived from (4) will also be sparse. For a comprehensive derivation of the $\mathbf{A}$ matrix, please refer to Section II of the supplementary material.

The weight matrix $\mathbf{W}$ of the network is computed by

$$\mathbf{W} = (\mathbf{A} - (1-\alpha)\cdot\mathbf{I})/\alpha \tag{5}$$

where $\mathbf{I}$ is the identity matrix. In this way, the update equations of the network (1) become very similar to the FDTD update equations. It implies very strong symmetry constraints on the $\mathbf{W}$ matrix. The local value of c determines how strongly the p-neuron responds to inputs from surrounding o-neurons and together with the coupling to its neighbors this can result in local resonances, where the physics equivalent learns that small c correspond to low-frequency resonances. By introducing a gradient in c on top of the random value, the slow (low-frequency resonances) and fast (high-frequency resonances) ends of the 2-D network can thus be realized, as shown in Fig. 1(b). The variable k determines the amount of information transferred between the p neurons that the o-neuron connects. Increasing k results in more strongly damped resonances.

2) Two Tunable Parameters: Following training of $\mathbf{W}_{\text{out}}$, the output layer can thus identify a suitable combination of correct oscillators, thereby providing an initial estimate of the target beat periodicity and timing. To synchronize the prediction with the target beats in a human-like way in real time, an adaptation method of c and k is introduced.

Some prior models for human rhythm perception use a few resonators and account for the fact that humans adjust

their timing when performing synchronization tasks by either slowing down or speeding up based on the accumulated error indicating whether they are too early or too late. This has been modeled by introducing a slowly changing value that affects the resonance frequencies. In our model, the physical interpretation of the reservoir is used to create a similar effect. The random wave speed c is tuned by multiplying it with a factor that makes the reservoir as a whole to speed up (increasing c) or slow down (decreasing c). Thus, a model is obtained that almost instantaneously tunes in by resonances being excited, and than slowly improves synchronization by changing the resonance frequencies of this system [54].

The parameter k controls the responsiveness of neurons in the network, drawing inspiration from the excitation and inhibition mechanisms in the human brain. By tuning this parameter for every neuron in the network, specialization in responding to specific beat frequencies can be increased. Our dynamic selection mechanism excites the top-n neurons that contribute most to the output and increases their responsiveness while inhibiting all others (global inhibition). In this way, the reservoir "remembers" recent inputs and could even become generative by continuing a beat after the input has stopped. Physically, k adjusts the local damping rate of the system. Excitation reduces the damping rate, enhancing the contribution of active neurons, while inhibition increases the damping rate to suppress less contributive neurons.

If the output of our model focuses on a single channel, we adapt the internal structure by tuning the global parameters c and k as depicted in Fig. 1(b). Specifically, we introduce two loss functions, I_{early} and I_{late}, which are designed to determine whether the propagation speed parameter c should be increased or decreased. Additionally, we evaluate each neurons contribution by measuring changes in the MSE. This information enables us to amplify or dampen activity in the most and least contributive regions by adjusting the parameter k.

A prediction of an upcoming beat can fail in to ways: "too early" or 'too late', hence the error is split in two parts. In both cases, there is usually an overlap between the sound envelopes corresponding to a beat in the prediction and target. As the peak is artificially extended as a smooth peak, the slope of the peaks is employed to calculate the error I_{early} and I_{late} of the prediction $\hat{\mathbf{y}}_{t+\Delta t}$ and target $\mathbf{y}_{t+\Delta t} = \mathbf{x}_{t+\Delta t}$ signals.

If the prediction is descending while the target is ascending, we consider the prediction to be too early. Otherwise, if the prediction is ascending while the target is descending, the prediction is too late. Both values I_{early} and I_{late} are updated until an update_step is reached, and the connectivity weights are changed. They are reinitialized to 0 when the interval ends, as shown in Algorithm 1. To ensure proximity in amplitude between the target and prediction within the same time window, a moving average and a softmax normalization are first applied to both the target and prediction values

$$\mathbf{y}_{\text{norm}}(t) = \frac{\mathbf{y}_t - \mathbf{y}_{\text{mean}}}{\mathbf{y}_{\text{softmax}}(t)} \tag{6}$$

$$\hat{\mathbf{y}}_{\text{norm}}(t) = \frac{\hat{\mathbf{y}}_t - \hat{\mathbf{y}}_{\text{mean}}}{\hat{\mathbf{y}}_{\text{softmax}}(t)} \tag{7}$$

Algorithm 1 Calculate Loss Function to Adapt c

```
Init: batch_size, update_step, threshold_sum, threshold
for i in batch_size do
    ε_early = 0, ε_late = 0, δ_sum = 0, δ_c = 0.02
    for t in update_step do
        if y_norm(t) > max(ŷ_norm(t), 0) then
            if y_norm(t) − y_norm(t − 1) > 0 then
                if ŷ_norm(t) − ŷ_norm(t − 1) < 0 then
                    I_early(t) = I_early(t − 1) + 1
                end if
            else if y_norm(t) − y_norm(t − 1) < 0 then
                if ŷ_norm(t) − ŷ_norm(t − 1) > 0 then
                    I_late(t) = I_late(t − 1) + 1
                end if
            end if
        end if
        ε_early = ε_early + δ_early · I_early(t)
        ε_late = ε_late + δ_late · I_late(t)
        if δ_sum < threshold_sum then
            if ε_early − ε_late < threshold then
                δ_sum = δ_sum + δ_c
                c* = 1 + δ_c
            else
                δ_sum = δ_sum − δ_c
                c* = 1 − δ_c
            end if
        else
            No update.
        end if
    end for
end for
```

where

$$\mathbf{y}_{\text{softmax}}(t) = \ln\left(\int_0^t e^{\mathbf{y}_{t'}} e^{\frac{t'-t}{\tau}} dt'\right) \tag{8}$$

$$\hat{\mathbf{y}}_{\text{softmax}}(t) = \ln\left(\int_0^t e^{\hat{\mathbf{y}}_{t'}} e^{\frac{t'-t}{\tau}} dt'\right) \tag{9}$$

where τ is an exponential averaging time constant spanning multiple interbeat intervals (IBIs).

By weighted comparison of I_{early} and I_{late}, the decision is made to increase or decrease the speed-up factor δ_c by a fixed amount, as shown in Algorithm 1. If the prediction is too early, we decrease all elements of c proportionally to their values; if it is too late, we increase c. In this way, the entire oscillation in the model slows down or speeds up.

The proposed dynamical selection (DS) mechanism controls the damping of network oscillations by modulating the poles of the W matrix. Key regions of the dynamic system, which are critical for beat prediction, are amplified by reducing k, while less significant oscillations are damped. Each neuron is masked in turn, generating masked outputs, and its MSE with respect to the target is computed in each time window. Neurons with the largest MSE reductions are identified as most critical, while those with minimal reductions are deemed less significant. The parameter k is adjusted around these neurons, enhancing or diminishing their activity to focus attention on specific

rhythmic behaviors. Since k changes slowly, this mechanism maintains focus even during rhythmic pattern transitions, emphasizing its utility for dynamic rhythmic processing.

C. Fine-Tuning and Continuation

Having completed the model training, a first approximation to the output W_{out} is obtained. However, predictions for task illustrated in Fig. 1(b) will involve two inputs in interaction task, lack accuracy. In this study, distinct combinations of beat multiples under the same meter are regarded as separate tasks. During the fine tuning process, fast adaptation is applied to the model's output layer, enabling the model to learn new combinations of multiples at different meters swiftly.

Due to the well-established initialization, the model, prior to fast adaptation, can accurately output beats at appropriate positions for tasks within the perceptible frequency range for humans. However, the amplitudes of these beats are inaccurately predicted, exhibiting significant errors compared to the target beats. Additionally, predictions sometimes include extra beats between two adjacent target beats, further contributing to discrepancies between predictions and targets.

To facilitate fast adaptation of the output weights to new tasks, we store the fixed time step predictions and targets [55]. We compute the MSE between them and perform a few steps of weight updates in the direction that minimizes the MSE. To account for the impact of the additional peak between adjacent target beats on the computed MSE, a penalty is applied. Since the beats between input signals consist entirely of zeros, the MSE penalizes all extra beats equally. Therefore, we replace the zero segments between two adjacent beats with the negative portion of a sine wave with a 0.5 phase, matching the amplitude of the input peak.

During the fine-tuning phase, the adjustment formula for the model's output weights is

$$\mathbf{W}_{\text{out}} = \mathbf{W}_{\text{out}} - \text{lr} * \frac{\partial \text{MSE}}{\partial \mathbf{W}_{\text{out}}} \tag{10}$$

where the lr is the defined learning rate.

Once the visual reference halts (Fig. 1), fine-tuning ends. Subsequently, the model's prediction persists for an additional beat. Its detection initiates a feedback loop, and the prediction replaces the visual reference. Notably, there is no need to modify the input weight, as shown in (11). Consequently, the closed-loop model acquires the capacity to learn from its own sound and sustain the acquired rhythm

$$\begin{aligned} \mathbf{h}_t &= (1-\alpha)\mathbf{h}_{t-1} \\ &\quad + \alpha f\left(\mathbf{W}_{\text{in}}(\mathbf{x}_t + \hat{\mathbf{y}}_{t-\Delta t}) + \mathbf{W}\mathbf{h}_t + \mathbf{b}_t + \xi_t\right) \\ \hat{\mathbf{y}}_{t+\Delta t} &= \mathbf{W}_{\text{out}}\mathbf{h}_t \end{aligned} \tag{11}$$

where $\hat{\mathbf{y}}_{t-\Delta t}$ denotes the feedback from the output, and $\Delta t = n\delta t$ indicates that the model is predicting Δt ms into the future.

D. Customization During Human Behavior Simulation

When two humans interact musically, they can adopt different behaviors. In broad terms they can act as a follower or as a leader in the interaction. When comparing the proposed model with interaction experiments [Fig. 4(b)], this is included by customizing the update learning rate. To this end, we utilize the Wasserstein distance $W(\cdot,\cdot)$ between the interbeat interval generated by Participant 1 (P1) and the primer interval, and between the interbeat interval generated by P1 and participant 2 (P2).

The Wasserstein distance [56] measures the amount of "work" required to move one set of timings to another set, effectively identifying the interbeat interval distribution similarity. The equation is as follows:

$$W\left(P_{\text{pred}}, P_{\text{target}}\right) = \min_{\gamma \in \Gamma\left(P_{\text{pred}}, P_{\text{target}}\right)} \int |d_{\text{pred}} - d_{\text{target}}|\, d\gamma \tag{12}$$

where P_{pred} and P_{target} are the interbeat interval distribution of human prediction and visual reference, respectively, d_{target} represents the set of interbeat intervals from the participant, d_{pred} denotes the set of interbeat intervals from the reference or another subject, $\min_{\gamma \in \Gamma(P_{\text{pred}}, P_{\text{target}})}$ signifies the minimum over all joint distributions γ with marginals P_{pred}, P_{target}, $|\cdot|$ denotes the absolute time difference between two interbeat intervals, and $d\gamma(d_{\text{pred}}, d_{\text{target}})$ represents the joint distribution.

The update learning rates are defined as

$$\beta * \frac{1}{W\left(P_{\text{pred}}, P_{\text{target}}\right)} * \exp\left(-\frac{t - t_0}{\tau}\right) \tag{13}$$

where $\beta = 0.001$ is a scaling factor, $W(P_{\text{pred}}, P_{\text{target}})$ represents the Earth mover's distance (EMD) between the predicted event timings and the target event timings, t and t_0 denote the current time and the time of the first target event, and $\exp(-[t - t_0/\tau])$ is an exponential decay term.

The purpose of this equation is to control the update rate of W_{out}. The term $W(P_{\text{pred}}, P_{\text{target}})$ helps to determine how closely the historical human behavior aligns with visual reference and audio reference, allowing W_{out} to prioritize the corresponding oscillators in the neural network. To prevent overly fast adaptation, we introduce a small scaling factor β. Assuming that participants interactions remain relatively stable, we apply an exponential decay term $\exp(-[t - t_0/\tau])$, where τ is the decay time constant, set to 6 ms.

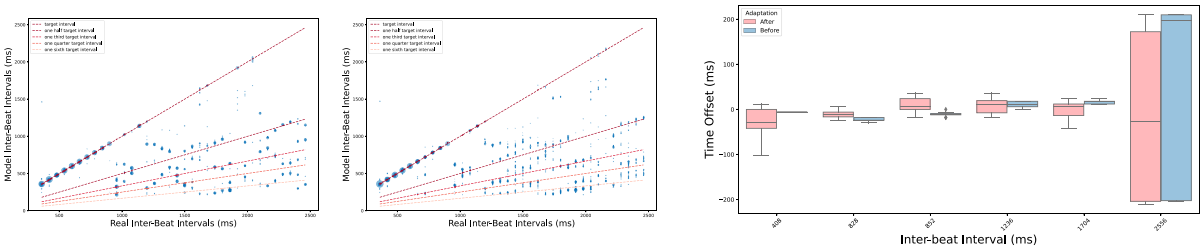
IV. Results

A. Experimental Setup

The pretraining phase utilized MATLAB-generated data comprising two-channel rhythms with varying frequency combinations, while the complete model was implemented in PyTorch. Beat timings, defined as the maximum points of smoothed peaks, were fed into the model.

The proposed model was evaluated using single-channel rhythms to assess prediction performance and dual-channel rhythms to test its efficiency in rhythm interaction tasks. Subsequently, the model was applied to real human interaction scenarios. The experimental setup is detailed below.

1) Single-Channel Experiments: In human perception, lower-level dynamics (e.g., grouping beats) involve sensory tissues such as the cerebellum, while higher cognitive processes handle complex tasks [8]. These processes enable rhythm



(a) (b)

Fig. 2. Proposed oscillator-based system demonstrates internal meter perception. (a) Comparison of the difference in IBI discrete bubble distribution before and after adaptation of global parameters. The larger the size of the bubbles, the higher the probability that the IBI will be concentrated in this area. (b) Comparison of the time offset before and after adaptation of global parameters c and k.

generation within specific frequency ranges and anticipation of future beats. Slower rhythms are subdivided into faster meters.

The model's frequency perception was tested on rhythms with IBIs from 400 to 3000 ms, in 50-ms increments, using a time-step δt of 6 ms. Pretraining used a large dataset of plausible rhythms, and testing was conducted without fine-tuning and with/without tuning c and k for synchronization. Results are shown in Section IV-B.

2) Dual-Channel Experiments: Dual-channel input was introduced to evaluate the models ability to interact with another rhythm. Channels were based on the same metronome but differed in beat grouping, denoted as $i : j$ $(i, j = 1, 2, 3)$. Two metronomes were selected: 72 BPM (low frequency) and 144 BPM (high frequency). Channel 1 (reference) ceased earlier than Channel 2, representing a participant or computer signal.

The model was compared with the Temporal Fusion Transformer [57] and the GRU in terms of prediction accuracy and internal state dynamics (Section IV-C). Three scenarios relevant to human experiments: skip-one, skip-a-while, and gradually increasing IBIs, are discussed in Section IV-D.

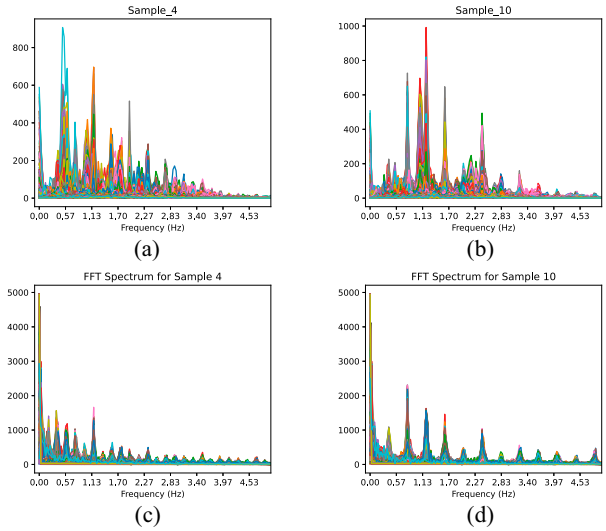
3) Real Human Experiments: The model was applied to human-computer and interhuman interaction settings. In the human-computer setup [Fig. 4(a)], the model received synchronized visual (Channel 1) and audio (Channel 2) inputs from a computer-based metronome. Visual cues were rotating balls, while audio cues were short beats. The visual metronome persisted for 30 s, while audio faded out after 15 s. The model also received human tapping signals recorded via a MIDI device, with beat timings extracted from smoothed peaks.

In interhuman interactions [Fig. 4(b)], Channel 2 was replaced by another human participant. The model simulated human behavior in both setups, simulating Participant 1 (P1) in interhuman interactions. Results are presented in Section IV-E.

B. Analysis of Single-Channel Input Results

As shown in Fig. 2(a), the model predicts rhythms accurately for IBIs below 1500 ms, with minimal variance between predicted beats. For longer intervals, predictions incorporate subdivisions, aligning with fractions (e.g., one-half or one-third) of the target intervals, similar to human behavior when IBIs exceed rhythmic prediction limits.

After tuning parameters c and k for synchronization (Fig. 2(a), right), the model achieves higher precision but



(a) (b) (c) (d)

Fig. 3. FFT spectrum of our model's and GRU's hidden states, for 2 representative examples. (a) 2 3, 1.2Hz (Ours). (b) 2 3, 2.4Hz (Ours). (c) 2 3, 1.2Hz (GRU). (d) 2 3, 2.4Hz (GRU).

generates more subdivisions, resembling internal human counting between beats. Fig. 2(b) compares time offset ratios between predictions and targets for six intervals. Without synchronization, the offset ratio remains within 4%, except for the 2556 ms interval, demonstrating the model's baseline accuracy. Synchronization reduces lead or lag errors, producing mean values close to zero or slightly ahead of the targetfeatures desirable for real-world tapping tasks.

These results suggest that the synchronization algorithm promotes human-like rhythmic behavior, maintaining accuracy while introducing variability. The tunable parameters c and k effectively regulate timing and directionality, drawing parallels with neural traveling waves and oscillations that dynamically coordinate timing and connectivity in human cognition and behavior [58], [59].

C. Comparison With Other Machine Learning Models

1) Single-Channel Comparison: In [14], the authors propose an ESN with a Synaptic Plasticity Rule, training output weights using ridge regression. We reproduced this model on our single-channel dataset and tested it. While it predicts beats effectively with input signals, it fails to generate an

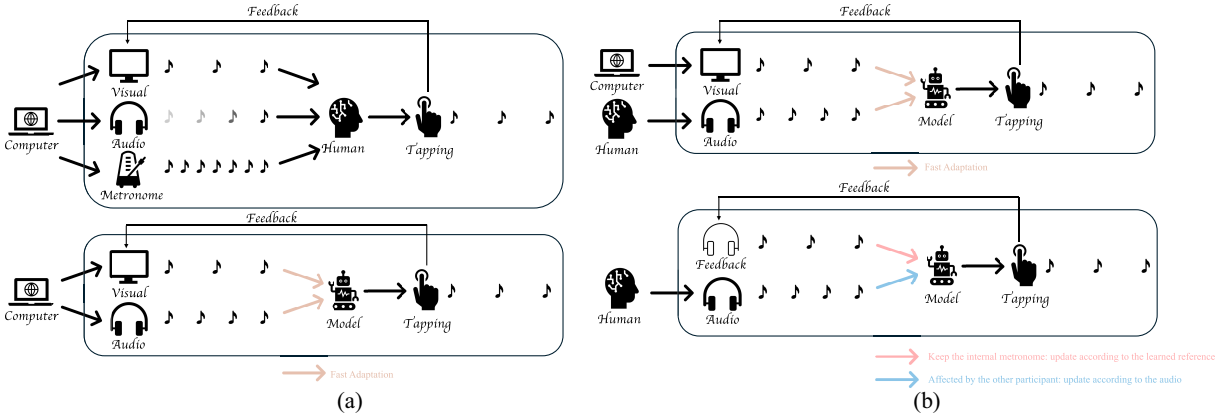


Fig. 4. *Display of interaction procedure.* (a) *Human-computer interaction experiment.* The upper part depicts the procedure of the experiment with people. During a learning phase, the participant is guided by a visual reference that gradually fades, while a different audio rhythm gradually fades in from a computer, along with a shared metronome. The participant is expected to tap the rhythm in sync with the visual reference, even after the latter has stopped. The lower part outlines the procedure for applying our model to a similar experimental setting. Given the model's proven metronome perception, the metronome is not included as an input. During learning, the input mirrors the human learning procedure, while the output weight undergoes *fine-tuning* to the dual input. After the visual reference stops, the model *feeds back* its own tapping. (b) *Interhuman interaction procedure.* The upper part illustrates the model output weight updating procedure, which remains consistent with the human-computer interaction. The lower part demonstrates the customization added to the model's output weight. The model is affected by itself and the other channel to varying degrees.

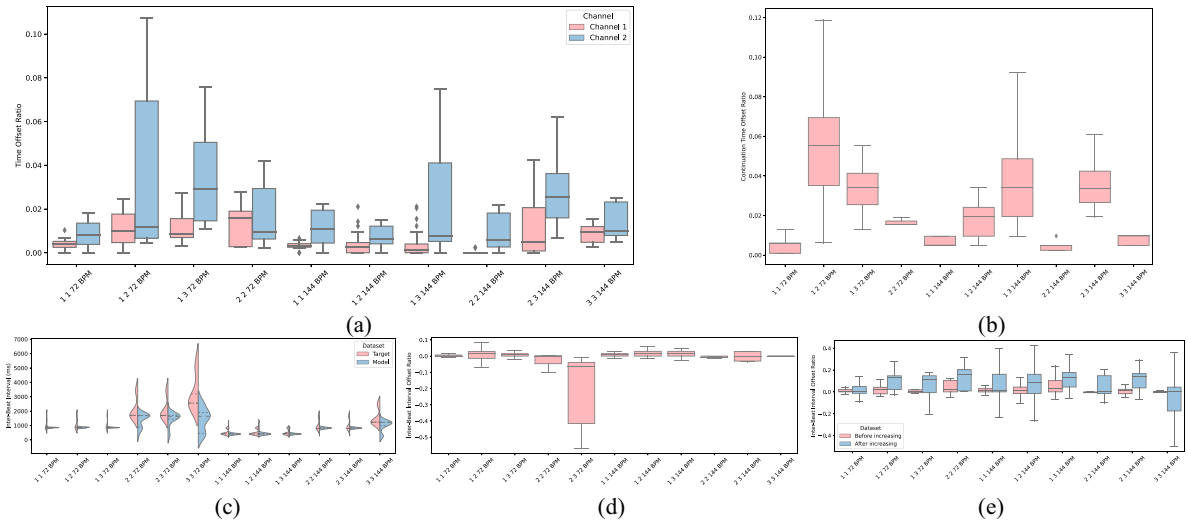


Fig. 5. *Dual-channel interaction of the proposed oscillator-based system.* (a) Time offset ratios for both channels in all combinations are compared. (b) Effect of closed-loop on the continuation task for channel 1 in all samples. (c) Comparison of the distribution between the skip-one target IBI and the prediction's distribution for all samples. (d) Illustration of the skip-a-while prediction's IBI error ratio for all samples. (e) After channel 2 increases the IBI by 2%, a comparison of the prediction's IBI error ratio before and after the increase for all samples.

additional beat when input ceases, diverging from human-like behavior (Fig. S1). To better compare with our model, we replaced ridge regression with backpropagation and MSE. Although this improved the output shape, the improvement mainly reflected the input signals peak structure rather than the models predictive ability (Fig. S2).

Chen et al. [15] presented a multicluster ESN for trajectory prediction, distinct from our application. Adapting it to our task caused issues with connectivity matrix inversion due to input signal sparsity. Using a pseudo-inverse during training resolved functionality but led to output convergence to a fixed value, failing to capture task dynamics effectively.

2) Dual-Channel Comparison: Fig. 5(a) shows the time offset errors before and after Fast Adaptation, highlighting significantly reduced mean errors and variance across all samples. The mean errors remain below Δt, demonstrating precise and stable predictions even for challenging rhythmic patterns. The continuation task further illustrates the model's ability to sustain rhythms after the reference halts, with continuation errors consistently below 58 ms and relative errors around 2%–5%, as shown in Fig. 5(b).

Comparative analysis with the temporal fusion transformer (TFT) [57] indicates that TFT maintains a stable -58 ms offset (Fig. S3), reflecting its reliance on the onset shape of beats rather than rhythmic oscillations. GRU performance aligns with our model for rhythms within human perception (Fig. S4) but diverges for complex patterns and slower tempos, as seen in Fig. S7 and Fig. S8.

Fast Fourier transform (FFT) spectra of internal states reveal critical differences between GRU and our oscillator-based model. As shown in Fig. 3 (Refer to the complete version in

Figs. S5 and S6 for more details.), for the 2:3 sample, the GRU exhibits low-frequency components below the rhythm's fundamental frequency (e.g., under 1.2 Hz), alongside higher harmonics. These low-frequency components likely stem from GRUs gating mechanism attempting to learn absolute timing over longer durations, rather than oscillatory behavior. This contrasts with our model, which avoids such low-frequency artifacts, focusing on biologically plausible rhythms. The GRU spectrum also includes higher-frequency responses around 4 Hz, typically perceived as modulation frequencies in human systems rather than rhythms suitable for motor synchronization.

In comparison, our model's states naturally generate oscillations that align closely with the 2:3 rhythm, balancing the fundamental frequency and its harmonics. These oscillations reflect internal "counting" mechanisms, enabling the model to track slow rhythms effectively. For the 2:3 case, the inclusion of additional harmonics by our model supports rhythmic stability while preserving a human-like response to slow and complex beats.

Unlike GRUs learned temporal patterns through gated updates, our models oscillatory behavior arises inherently from coupled oscillators, which dynamically interact to produce rhythmic patterns. This architectural design closely mirrors biological systems, where neural oscillations like alpha and beta waves are integral to attention and motor coordination. Consequently, our model not only achieves human-like rhythmic behavior but also exhibits superior biological plausibility for real-world rhythmic tasks.

D. Simulation of Human-AI Interaction Across Three Designed Scenarios

The approach proposed can be conceptualized as a challenge within the realm of human-computer interaction. To theoretically gauge our model's capacity in potential real-life applications, we describe three distinct scenarios: such as the human reference skipping a beat; the human excluding portions of a reference; and the computerized metronome enhancing the IBI by 2% with every beat. Note that, these changes start after the fast adaptation.

To evaluate whether the model has learned to predict, we employ the skip-one and skip-a-while tasks. If the model can accurately fill the skipped beats or time intervals in a predictive manner, it indicates that our model has indeed learned the rhythm. Additionally, during human and model interaction, human behavior typically exhibits some variability. It is challenging to expect all beats generated by humans to have precise timing. Therefore, introducing some variability in channel 2 by increasing the IBI by 2% with every beat and observing the resulting behavior change in channel 1 would be reasonable.

It is important to note that due to the low frequency of the metronome, beats occurring every 3 times the metronome interval may be challenging to perceive as rhythmic. Consequently, in samples such as 3:3 72 BPM, accurately predicting both channels becomes difficult. Moreover, samples like 2:3 72 BPM perform even worse due to the inherent difficulty of the 2:3 combination task. However, 1:3 72 BPM performs relatively better, primarily because channel 1 can act as a metronome, aiding channel 2 in counting beats.

Following Fast Adaptation, we input the first channel with a signal that skips one beat every two beats, while keeping the second channel unchanged. As shown in Fig. 5(c), the IBI distribution exhibits two peaks, representing the basic interval and double basic interval. However, the distribution of the model's predicted IBI only displays the basic interval. This indicates that the model has not learned the skip-one pattern, resulting in it naturally filling the gap.

When skipping input for the first channel temporarily but retaining its delayed feedback, the generated time intervals during the pause do not deviate significantly. With the exception of outliers in each sample and the most challenging sample involving the second channel with three multiples, the majority of IBI errors are under 3%.

After learning the provided rhythms in both channels, the IBI of the second channel begins to increase continuously and evenly. In Fig. 5(d), we compare the IBI error ratio of the first channel before and after the change in the second channel. It is evident that our model is affected by the changing channel. Prior to the increase, the prediction's IBI error is below 4%. However, after the interval increase, the variances also escalate, particularly for high-frequency rhythms. As illustrated in Fig. 5(e), the IBI error ratio shifts toward the negative range, indicating that with the increasing IBI in the second channel, the prediction in the first channel also tends to elongate the interval. This phenomenon is observed in human behavior, where individuals tend to subconsciously adjust their pace to match that of the person they are interacting with if the other person becomes slower.

E. Comparison With Actual Human Performance Across Two Settings

1) Human-Computer Interaction: Due to the simplicity of predicting the 1:1 combination, we conducted experiments using the 2:2, 2:3, 3:2, and 3:3 combinations instead. We replaced the participants in channel 1 with our model. Subsequently, we analyzed the distribution of IBIs between the beats generated by real participants and those generated by the model. To ensure the robustness of the findings, we tested four participants.

The experimental results are illustrated in supplementary materials Fig. S9. From the graph, it is evident that different participants exhibit varying performances when subjected to fixed computer audio inputs. In simpler tasks such as 2:2 and 3:3, the differences between participants are not significant, and the distributions can be maintained around the target interval. However, not every beat falls precisely on the target position, demonstrating some variance. This behavior mirrors that of our model, where the distribution's mean aligns with the target interval, albeit with less variance than that of the participants. In more complex tasks like 2:3 and 3:2, there is a greater disparity in performance among different participants. Some participants can sustain the learned rhythm even after the reference stops, while others are more influenced by the

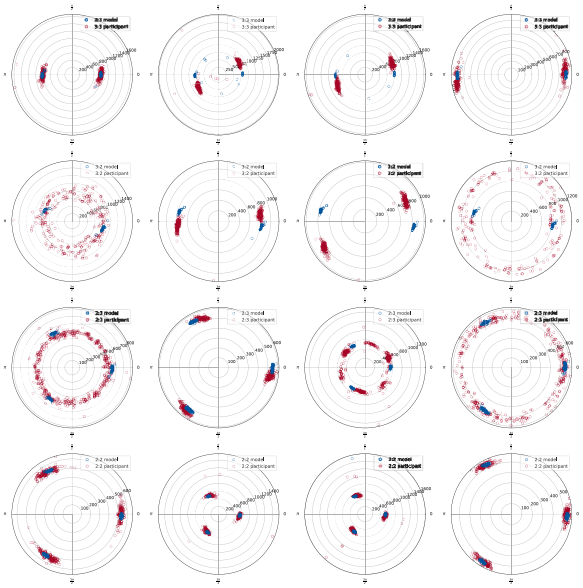


Fig. 6. *Comparison of the proposed oscillator-based systems interaction with a human to that of human-computer interaction.* Each subplot in a row represents a different participant interacting with the same rhythm combination. The subplots in the same row correspond to the same group of participants, illustrating their varied interactions with the given rhythm. Within each subplot, four groups of participants are shown alongside their respective models. Hollow circles indicate the beats dropped by both participants and models. The angle of each hollow circle denotes the phase shift relative to the cycle (6 * metronome IBI), while the distance from the center indicates the interval between consecutive beats.

computer audio, causing the distribution peaks to shift toward the interval of the other channel. Since our model maintains the same input throughout the initial 30 s, the distribution does not vary significantly during this period. However, due to the inherent randomness integrated into the model, the distributions of the four instances are not entirely identical.

To visualize the behavior patterns of humans and the AI model, we plotted the event timings from each group in each scenario in a circular arrangement, as shown in Fig. 6. Hollow circles represent beats missed by both participants and models. The angle of each hollow circle indicates the phase shift relative to the cycle (6 * metronome IBI), while the distance from the center represents the interval between consecutive beats. The model demonstrates patterns with high accuracy and low variance. In the simpler cases (2:2 and 3:3), participants also capture the pattern but with greater variance. Compared to the participants, the model exhibits smaller variance and is less prone to phase shifts. In more complex cases, different participants display varying interactions under the same conditions. For example, in the 3:2 scenario, one participant tends to generate additional beats between two consecutive beats in the visual reference, while another misses some reference beats. The other two participants can capture the pattern but exhibit larger phase shifts, indicating they follow the visual references. In the 2:3 scenario, participants show better pattern recognition but tend to strike one event later and the next one earlier. As the figures show no clear tendency for human behavior patterns to drift toward audio channel patterns, it is unnecessary to adapt the model's output weight according to the other channel.

To analyze frequency interactions across various conditions between human participants and our model, we used FFT on the time series outputs. These comparisons are detailed in supplementary materials Fig. S10. Participants show diverse frequency spectra in rhythmic tasks. Simpler tasks (2:2, 3:3) yield distinct frequency peaks, while challenging tasks (2:3, 3:2) result in more scattered spectra. The model's performance remains consistent across tasks. In the 2:2 and 3:3 tasks, human FFT spectra overlap significantly with the model's. However, in the 2:3 and 3:2 tasks, participants' responses vary, leading to scattered spectra. Some participants (e.g., participants 1 and 4 in 3:2) show less discernible frequency patterns, while high performers (e.g., participants 2 and 3 in 3:2) create subdivisions to maintain learned frequencies. Similar but less pronounced patterns are seen in the 2:3 task, indicating that higher frequency rhythms are easier to sustain. The model generally follows the visual reference frequency, occasionally showing subdivisions similar to those in humans.

2) Interhuman Interaction: To assess the participant's control over the rhythm, we compared their actual experimental data, encompassing all tapping IBIs, with two sets of data: the IBIs synchronized with channel 1's visual reference and those aligned with channel 2, originating from another participant's real IBIs. Through this comparison, we derived two distances, reflecting how closely the participant's taps align with each channel. These distances serve as indicators of the participants ability to control the rhythm and their susceptibility to external influences from the other participant. Accordingly, the output weight is adjusted based on the measurements, as shown in (13).

The results of comparing the four participants and their corresponding performances are depicted in supplementary materials Fig. S11. The model is utilized to simulate the behavior of Participant 1 (P1). The combination of two channels remains consistent with the human-computer interaction setup. There are four groups of participants, and our model simulates the behavior of P1 in each group. Across most combinations and groups, our model successfully emulates the participants' behavior. The IBI distributions between our simulation and the beats generated by the participants are similar.

To compare the interaction patterns between individuals with different rhythms, a dot plot is shown in Fig. 7. Participants typically exhibit unpredictable and variable behaviors during each experiment. As a result, the hollow dots plotted in Fig. 6 do not display clear patterns, with angles distributed throughout the entire cycle. Therefore, we replaced the fixed-interval reference with Participant 2's behavior, resulting in Fig. 7, where the cycle is adapted dynamically.

The basic metronome IBI is calculated as the average of two intervals generated by P2. For both the model and P1's behavior, we identify the closest beat generate by P2 and used its previous two intervals to calculate the cycle for both the model and P1. The cycle is updated for each event generated by the model and participants. To examine pattern shifts with the adaptive cycle, event timings are grouped according to the task, and each group is initialized as a new phase

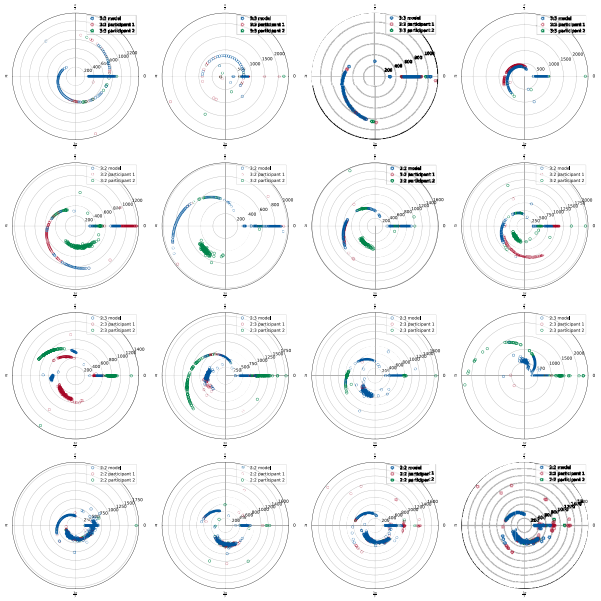


Fig. 7. *Comparison of the proposed oscillator-based systems interaction with a human to that of interhuman interaction.* Each subplot in a row represents a different participant interacting with the same rhythm combination. The subplots in the same row correspond to the same group of participants, illustrating their varied interactions with the given rhythm. Within each subplot, four groups of participants are shown alongside their respective models. Hollow circles indicate the beats dropped by both participants and models. The angle of each hollow circle denotes the phase shift relative to the cycle (6 * metronome IBI), while the distance from the center indicates the interval between consecutive beats.

starting from 0. Consequently, clearer patterns relative to P2 are revealed in Fig. 7. Similar to the models behavior shown in Fig. 6, when the task is 2, the dots appear around three angles: 0, $(1/3)\pi$, and $[2/3]\pi$. For task 3, the dots are located around 0 and $(1/2)\pi$. By introducing grouped phase initialization, the beats at the beginning of the cycle are consistently mapped to 0 degrees, though the intervals between them vary. As the radius of the dots at 0 degrees increases, the dots at other angles diverge away from the center, otherwise, the dots converge toward the center.

In all rhythm combinations, some participants tend to play faster than the reference, while others excel at maintaining their reference rhythm with closer mean IBIs and smaller variances. Different participants are influenced by each other in varying ways. For example, in the 3:3 combination, nearly all participants are significantly affected by each other, as evidenced by their interval distributions closely resembling one another. However, in the 2:3 group 2, P1 is not heavily influenced by the other participant, while the other participant struggles to maintain the rhythm behavior. P1 remains consistent with the learned rhythm despite external influences. Hence, employing the proposed varying degrees of updating the output weight, the model does not adhere rigidly to performance, akin to human-computer interaction. Instead, it exhibits differences across various combinations and participant groups. While the model may not replicate the exact IBI distribution of P1, it still demonstrates significant overlap. In all instances, the model's outputs exhibit mean values closely resembling P1's performance. However, during the updating of the output weight, occasional beats occurring between two beats are observed, mirroring findings from human-computer interaction experiments. This phenomenon suggests that humans may sometimes need to count metronome beats to maintain their internally learned rhythm.

In most cases, P1 exhibits three types of behavior: playing based on their own visual reference or following their partner. In the 2:2 and 3:3 cases, both the model and P1 typically display similar behaviors, as evidenced by the statistically similar IBI distributions shown in Fig. 7. In Groups 1, 3, and 4 for task 2:2, P1 tends to play faster than the visual reference after it disappears, with their partners showing similar behavior. The dots show the tendency to move toward the center. In Group 2, they are more likely to adhere to the presented visual reference, as the dots radium is more consistent. But the phase shift is different from each other. Consequently, there is a greater overlap between the model and both participants. For Task 3:3, the models in all groups exhibit similar behavior to P1, as indicated by the substantial overlap between their dots. In Group 1, the P1 is likely to change their behavior patterns during the experiment. In Group 2, the participant adheres more closely to the visual reference rhythm but with greater variance. In Groups 3 and 4, participants are likely to play faster during the experiment. In harder tasks, the interaction will be more complicated so these tasks will be analyzed separately. In the first group, P2 gradually increased their speed, while P1 exhibited fluctuations between fast and slow speeds. Consequently, the model is expected to be relatively accurate, resulting in increased phase variations. In the second group, P2 continued to increase their speed, while P1s performance closely matched the model, although not deterministically. Both participants positions remained around their respective targets. In the third group, both P2 and P1 increased their speeds slightly, but neither deviated far from their targets, with the model consistently centered around the target. In the fourth group, P2s speed increased gradually but not significantly, while P1s speed decreased. The model, influenced by P2, also increased slightly in speed but not as markedly as P1s decrease. In the first group, P2 exhibited fluctuating speeds, sometimes fast and sometimes slow, resulting in a radius variation around 800ms. P1 gradually increased their speed, but P2 maintained a 2:3 ratio. The model adhered to the rhythm of the visual reference, leading to a phase pattern with a certain angular deviation. Consequently, the models distribution around 0, $(1/3)\pi$, and $[2/3]\pi$ was more concentrated compared to P1. In the second and third groups, although the phase patterns of the model and P1 were more similar, P2s behavior was entirely different. This difference is evident from the radius distribution at angle 0. In the second group, P2s speed gradually decreased, deviating significantly from their visual reference. In the third group, P2s speed increased, but with a smaller deviation from the visual reference. In the fourth group, P2 exhibited completely random behavior patterns, resulting in many outliers at angle 0. However, P1 and the model displayed a certain degree of similarity, with a higher overlap in their behavior.

The FFT is applied to outputs from both participants and our model, with comparisons shown in Fig. S12 of the

supplementary materials. Compared to interactions with a computer, spectra indicate increased noise levels, especially in 2:2 and 3:3 scenarios. Participants' frequency spectra vary with rhythmic tasks, revealing diverse interaction patterns. Our model exhibits broader variations due to tailored output weight updates. In familiar scenarios like 2:2 and 3:3, participants' spectra converge with the model's, though some participants adopt novel patterns, causing spectral differences. In tasks 2:3 and 3:2, participants' frequencies have clearer peaks, resulting in less noisy model spectra. The model's output mirrors its counterpart's frequencies, such as a minor peak aligning with P2 in the 2:3 task involving groups 2 and 3.

V. Conclusion

In this study, we developed a biologically inspired dynamic system for simulating human-like rhythmic cognition, integrating predictive coding to align the model closely with human behavior. The weight matrix, derived from wave equation discretization, provides transparency and control over propagation speed and damping rate. Fine-tuning the output weights enhanced the model's adaptability to new rhythmic combinations, enabling richer, spontaneous, and human-like oscillatory behavior, distinguishing it from conventional deep learning methods.

The model successfully replicated human-like nuances in poly-rhythm tasks within the human perception range, including prediction timing deviations and subharmonic neuron activity. Interaction tests with human participants confirmed its generalization capabilities for fundamental rhythmic patterns, supported by a novel visualization-based measurement.

Future work will focus on refining experimental protocols, such as incorporating paired interaction experiments with diverse participants, and exploring the influence of various musical genres to enhance model evaluation and neural response control. These efforts aim to deepen our understanding of rhythm cognition and advance human-machine interaction in rhythmic perception and production.

References

[1] I. Winkler, G. P. Háden, O. Ladinig, I. Sziller, and H. Honing, "Newborn infants detect the beat in music," *Proc. Nat. Acad. Sci.*, vol. 106, no. 7, pp. 2468–2471, 2009.

[2] N. Jacoby and J. H. McDermott, "Integer ratio priors on musical rhythm revealed cross-culturally by iterated reproduction," *Curr. Biol.*, vol. 27, no. 3, pp. 359–370, 2017.

[3] R. G. Brown, *Smoothing, Forecasting and Prediction of Discrete Time Series*. Chelmsford, MA, USA: Courier Corp., 2004.

[4] E. W. Large, J. A. Herrera, and M. J. Velasco, "Neural networks for beat perception in musical rhythm," *Front. Syst. Neurosci.*, vol. 9, p. 159, Nov. 2015.

[5] S. Feld, *A Generative Theory of Tonal Music*. Cambridge, MA, USA: MIT Press, 1984.

[6] C. Palmer and A. P. Demos, "Are we in time? How predictive coding and dynamical systems explain musical synchrony," *Curr. Direct. Psychol. Sci.*, vol. 31, no. 2, pp. 147–153, 2022.

[7] O. A. Heggli, I. Konvalinka, M. L. Kringelbach, and P. Vuust, "Musical interaction is influenced by underlying predictive models and musical expertise," *Sci. Rep.*, vol. 9, no. 1, 2019, Art. no. 11048.

[8] D. Kudithipudi et al., "Biological underpinnings for lifelong learning machines," *Nat. Mach. Intell.*, vol. 4, no. 3, pp. 196–210, 2022.

[9] J. O'Byrne and K. Jerbi, "How critical is brain criticality?" *Trends Neurosci.*, vol. 45, no. 11, pp. 820–837, 2022.

[10] P. Tichko, J. C. Kim, and E. W. Large, "A dynamical, radically embodied, and ecological theory of rhythm development," *Front. Psychol.*, vol. 13, Feb. 2022, Art. no. 653696.

[11] M. Inubushi and K. Yoshimura, "Reservoir computing beyond memory-nonlinearity trade-off," *Sci. Rep.*, vol. 7, no. 1, 2017, Art. no. 10199.

[12] M. Xu and M. Han, "Adaptive elastic echo state network for multivariate time series prediction," *IEEE Trans. Cybern.*, vol. 46, no. 10, pp. 2173–2183, Oct. 2016.

[13] M. Rigotti et al., "The importance of mixed selectivity in complex cognitive tasks," *Nature*, vol. 497, no. 7451, pp. 585–590, 2013.

[14] X. Wang, Y. Jin, and K. Hao, "Computational modeling of structural synaptic plasticity in echo state networks," *IEEE Trans. Cybern.*, vol. 52, no. 10, pp. 11254–11266, Oct. 2021.

[15] Q. Chen, X. Li, A. Zhang, and Y. Song, "Neuroadaptive tracking control of affine nonlinear systems using echo state networks embedded with multiclustered structure and intrinsic plasticity," *IEEE Trans. Cybern.*, vol. 54, no. 2, pp. 1133–1142, Feb. 2022.

[16] G. R. Yang, M. R. Joglekar, H. F. Song, W. T. Newsome, and X.-J. Wang, "Task representations in neural networks trained to perform many cognitive tasks," *Nat. Neurosci.*, vol. 22, no. 2, pp. 297–306, 2019.

[17] E. W. Large, "Resonating to musical rhythm: Theory and experiment," in *The Psychology of Time*. Oxfordshire, U.K.: Routledge, 2008, pp. 189–231.

[18] J. London, *Hearing in Time: Psychological Aspects of Musical Meter*. Oxford, U.K.: Oxford Univ., 2012.

[19] T. E. Matthews, M. A. Witek, J. L. Thibodeau, P. Vuust, and V. B. Penhune, "Perceived motor synchrony with the beat is more strongly related to groove than measured synchrony," *Music Percept., Interdiscipl. J.*, vol. 39, no. 5, pp. 423–442, 2022.

[20] L. Demany, B. McKenzie, and E. Vurpillot, "Rhythm perception in early infancy," *Nature*, vol. 266, no. 5604, pp. 718–719, 1977.

[21] S. E. Trehub and L. A. Thorpe, "Infants' perception of rhythm: Categorization of auditory sequences by temporal structure," *Can. J. Psychol.*, vol. 43, no. 2, p. 217, 1989.

[22] J. S. Snyder, R. L. Gordon, and E. E. Hannon, " Theoretical and empirical advances in understanding musical rhythm, beat and metre," *Nat. Rev. Psychol.*, vol. 3, pp. 449–462, Jul. 2024.

[23] O. F. Vander Elst, N. H. Foster, P. Vuust, P. E. Keller, and M. L. Kringelbach, "The neuroscience of dance: A conceptual framework and systematic review," *Neurosci. Biobehav. Rev.*, vol. 150, Jul. 2023, Art. no. 105197.

[24] S. A. Whitton, B. Sreenan, and F. Jiang, "The contribution of auditory imagery and visual rhythm perception to sensorimotor synchronization with external and imagined rhythm," *J. Exp. Psychol., General*, vol. 153, pp. 1861–1872, Jul. 2024.

[25] O. Abbasi and J. Gross, "Beta-band oscillations play an essential role in motor–auditory interactions," *Human Brain Map.*, vol. 41, no. 3, pp. 656–665, 2020.

[26] J. A. Grahn and M. Brett, "Rhythm and beat perception in motor areas of the brain," *J. Cogn. Neurosci.*, vol. 19, no. 5, pp. 893–906, 2007.

[27] J. Ross, J. Iversen, and R. Balasubramaniam, "Dorsal premotor contributions to auditory rhythm perception: Causal transcranial magnetic stimulation studies of interval, tempo, and phase," bioRxiv, Preprint, 2018.

[28] J. P. Rauschecker and B. Tian, "Mechanisms and streams for processing of what and where in auditory cortex," *Proc. Nat. Acad. Sci.*, vol. 97, no. 22, pp. 11800–11806, 2000.

[29] Y. Ito, T. I. Shiramatsu, N. Ishida, K. Oshima, K. Magami, and H. Takahashi, "Spontaneous beat synchronization in rats: Neural dynamics and motor entrainment," *Sci. Adv.*, vol. 8, no. 45, 2022, Art. no. eabo7019.

[30] D. Bendor and X. Wang, "The neuronal representation of pitch in primate auditory cortex," *Nature*, vol. 436, no. 7054, pp. 1161–1165, 2005.

[31] T.-H. Z. Cheng, S. C. Creel, and J. R. Iversen, "How do you feel the rhythm: Dynamic motor-auditory interactions are involved in the imagination of hierarchical timing," *J. Neurosci.*, vol. 42, no. 3, pp. 500–512, 2022.

[32] M. Rosso, B. Moens, M. Leman, and L. Moumdjian, "Neural entrainment underpins sensorimotor synchronization to dynamic rhythmic stimuli," *NeuroImage*, vol. 277, Aug. 2023, Art. no. 120226.

[33] M. Jin et al., "A survey on graph neural networks for time series: Forecasting, classification, imputation, and anomaly detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 12, pp. 10466–10485, Dec. 2024.

[34] Y. Liang et al., "Foundation models for time series analysis: A tutorial and survey," in *Proc. 30th ACM SIGKDD Conf. Knowl. Discov. Data Min.*, 2024, pp. 6555–6565.

[35] A. Vaswani, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1–11.

[36] Q. Wen et al., "Transformers in time series: A survey," in *Proc. 32nd Int. Joint Conf. Artif. Intell.*, 2023, pp. 6778–6786.

[37] S. Li et al., "Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–11.

[38] T. Zhou, Z. Ma, Q. Wen, X. Wang, L. Sun, and R. Jin, "Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 27268–27286.

[39] S. Hochreiter, "Long short-term memory," in *Neural Computation*. Cambridge, MA, USA: MIT Press, 1997.

[40] K. Cho et al., "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proc. Conf. Empir. Methods Nat. Lang. Process. (EMNLP)*, Oct. 2014, pp. 1724–1734. [Online]. Available: https://aclanthology.org/D14-1179

[41] Z. Chen, M. Ma, T. Li, H. Wang, and C. Li, "Long sequence time-series forecasting with deep learning: A survey," *Inf. Fusion*, vol. 97, Sep. 2023, Art. no. 101819.

[42] T. Gangopadhyay, S. Y. Tan, Z. Jiang, R. Meng, and S. Sarkar, "Spatiotemporal attention for multivariate time series prediction and interpretation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2021, pp. 3560–3564.

[43] E. W. Large, "Neurodynamics of music," in *Music Perception*. New York, NY, USA: Springer, 2010, pp. 201–231.

[44] J. Bhattacharya and H. Petsche, "Phase synchrony analysis of eeg during music perception reveals changes in functional connectivity due to musical expertise," *Signal Process.*, vol. 85, no. 11, pp. 2161–2177, 2005.

[45] K. B. Doelling, M. F. Assaneo, D. Bevilacqua, B. Pesaran, and D. Poeppel, "An oscillator model better predicts cortical entrainment to music," *Proc. Nat. Acad. Sci.*, vol. 116, no. 20, pp. 10113–10121, 2019.

[46] G. Tanaka, T. Matsumori, H. Yoshida, and K. Aihara, "Reservoir computing with diverse timescales for prediction of multiscale dynamics," *Phys. Rev. Res.*, vol. 4, no. 3, 2022, Art. no. L032014.

[47] L. Manneschi, M. O. Ellis, G. Gigante, A. C. Lin, P. Del Giudice, and E. Vasilaki, "Exploiting multiple timescales in hierarchical echo state networks," *Front. Appl. Math. Stat.*, vol. 6, Feb. 2021, Art. no. 616658.

[48] C. Gallicchio, "Euler state networks: Nondissipative reservoir computing," 2022, *arXiv:2203.09382*.

[49] A. Rodan and P. Tino, "Minimum complexity echo state network," *IEEE Trans. Neural Netw.*, vol. 22, no. 1, pp. 131–144, Jan. 2011.

[50] D. Bacciu and A. Bongiorno, "Concentric ESN: Assessing the effect of modularity in cycle reservoirs," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, 2018, pp. 1–8.

[51] C. Gallicchio, A. Micheli, and L. Pedrelli, "Deep reservoir computing: A critical experimental analysis," *Neurocomputing*, vol. 268, pp. 87–99, Dec. 2017.

[52] Z. W. Davis, L. Muller, and J. H. Reynolds, "Spontaneous spiking is governed by broadband fluctuations," *J. Neurosci.*, vol. 42, no. 26, pp. 5159–5172, 2022.

[53] D. Botteldooren, "Finite-difference time-domain simulation of low-frequency room acoustic problems," *J. Acoust. Soc. America*, vol. 98, no. 6, pp. 3302–3308, 1995.

[54] B. Hutcheon and Y. Yarom, "Resonance, oscillation and the intrinsic frequency preferences of neurons," *Trends Neurosci.*, vol. 23, no. 5, pp. 216–222, 2000.

[55] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 1126–1135.

[56] S. Vallender, "Calculation of the Wasserstein distance between probability distributions on the line," *Theory Probab. Appl.*, vol. 18, no. 4, pp. 784–786, 1974.

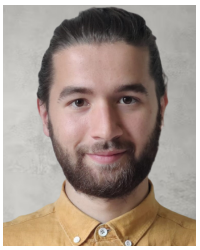
[57] B. Lim, S. Ö. Arık, N. Loeff, and T. Pfister, "Temporal fusion transformers for interpretable multi-horizon time series forecasting," *Int. J. Forecast.*, vol. 37, no. 4, pp. 1748–1764, 2021.

[58] U. R. Mohan, H. Zhang, B. Ermentrout, and J. Jacobs, "The direction of theta and alpha travelling waves modulates human memory processing," *Nature Human Behav.*, vol. 8, pp. 1124–1135, Mar. 2024.

[59] A. K. Engel, P. Fries, and W. Singer, "Dynamic predictions: Oscillations and synchrony in top–down processing," *Nat. Rev. Neurosci.*, vol. 2, no. 10, pp. 704–716, 2001.

Zhongju Yuan received the B.Sc. degree in applied mathematics from Central South University, Changsha, China, in 2020, the M.Sc. degree in materials and chemistry science from the Southern University of Science and Technology, Shenzhen, China, in 2022. She is currently pursuing the Ph.D. degree in computer science and engineering with the Waves Research Group, Ghent University, Ghent, Belgium.

Her research interests include bio-inspired neural networks for auditory processing and human–machine interaction.



Wannes Van Ransbeeck received the B.Sc. degree in electro-mechanical engineering from Ghent University, Ghent, Belgium, in 2019, and the Master of Science degree (Hons.) from Danish Technical University, Kongens Lyngby, Denmark, in 2021. He is currently pursuing the Ph.D. degree in biomedical engineering with Ghent University.

His research delves into the intersection of engineering, music, and brain activity, encompassing room and instrument acoustics, electronic system design, and human–music interaction.

Geraint A. Wiggins received the M.A. degree in mathematics and computer science from Corpus Christi College, Cambridge, U.K, in 1984, and the Ph.D. degree in artificial intelligence and in musical composition from the University of Edinburgh, Edinburgh, U.K., in 1991 and 2005, respectively.

He is Professor of Computational Creativity eith the Artificial Intelligence Lab, Vrije Universiteit Brussel, Brussel, Belgium. Previously, he was Head of the School of Electronic Engineering and Computer Science with the Queen Mary University of London, London, U.K., and remains a part time Professor of that institution. He has worked in artificial intelligence, computer music, and cognitive science since 1984. He was one of the founders of the research field of computational creativity, which studies creative intelligence, and was the first in the world to hold a named Professorial Chair in that research area. His current work relates to cognitive architectures that explicate the relationship between perception, learning and creativity, in sequential domains, most notably language, and music.auditory processing.

Dr. Wiggins is a Consulting Editor of *Music Perception* and an Editorial Board Member of the *Journal of New Music Research*. From 2000 to 2004, he chaired the Society for the Study of Artificial Intelligence and the Simulation of Behaviour, the U.K. learned society for Artificial Intelligence and Cognitive Science, and from 2004 to 2014, he chaired the international Association for Computational Creativity. He is the Founding Chair of the Dilys Trust, a charity that helps intellectually able but economically disadvantaged young people in the U.K. attend top class universities. He is a Fellow of the Royal Society of Arts.

Dick Botteldooren (Senior Member, IEEE) received the M.Sc. degree in electronic engineering and the Ph.D. degree in applied science from Ghent University, Ghent, Belgium, in 1986 and 1990, respectively.

He is a Full Professor with Ghent University, where he leads research on Acoustics and teaches a variety of courses related to sound and computational methods. In 1992, he became interested in acoustics and in particular environmental sound. He has made research contributions in the field of acoustic modeling, noise mapping, environmental sensor networks, computational intelligence, modeling perception of environmental sound, health impacts of sound, biomonitoring, urban sound planning, soundscapes, and noise policy support. This work was reported in over 200 journal publications and several hundred conference contributions. Based on his expertise, he was an Advisor for national and international health councils, and noise policy makers.

Prof. Botteldooren is currently the President of the European Acoustics Association. From 2004 to 2013, he was an Editor-in-Chief of *Acta Acustica* united with Acustica, the Journal of the European Acoustics Association. Until 2018, he was the President of the Belgian Acoustical Society. From 2015 to 2018, he was an I-INCE Vice-President for Europe and Africa. He is a Fellow of the Acoustical Society of America and the Institute of Acoustics and Vibration.