applied sciences

MDPI

Article

An AI-Driven Multi-Feature Approach for Synchronisation and QoE Assessment in Network Music Performance

Ioannis Doumanis [1,*], Kostantinos Tsioutas [2] and George Xylomenos [2]

[1] School of Engineering and Computing, University of Central Lancashire, Preston PR1 1JN, UK
[2] School of Information Sciences and Technology, Athens University of Economics and Business, 10434 Athens, Greece; ktsioutas@aueb.gr (K.T.); xgeorge@aueb.gr (G.X.)
* Correspondence: idoumanis@lancashire.ac.uk

Featured Application

Objective synchrony monitoring for Network Music Performance platforms, remote music education, rehearsal diagnostics, and adaptive audio-network systems.

Abstract

Network Music Performance (NMP) refers to remote musical collaboration over a network in applications such as music education, music production, and live performance. In NMP, synchronisation is a critical factor in musicians' Quality of Experience (QoE). This interpersonal coordination of musical actions is highly sensitive to variable network conditions, particularly to end-to-end delay and signal degradation. Existing evaluations rely mainly on subjective questionnaires or isolated objective descriptors, creating a gap for a unified metric that quantifies synchrony directly from performance signals. To address this gap, we propose the Objective Synchrony Index (OSI), an AI-driven metric that quantifies ensemble synchrony from paired NMP recordings. We computed OSI using a two-tower multi-task convolutional recurrent neural network (CRNN) that estimates synchrony-relevant descriptors from paired Musician A and Musician B audio streams. We introduce two OSI variants: timing-OSI, which captures temporal coordination through offsets, onsets, beats, and tempo coherence; and ensemble-OSI, which extends this formulation by integrating chord agreement and signal fidelity to reflect structural and perceptual aspects of ensemble interaction. We evaluated OSI using recordings from two NMP studies in which eleven pairs of musicians performed under systematically varied delay and sampling-rate conditions. After each performance, musicians completed QoE questionnaires, allowing us to relate OSI and its components to subjective ratings using repeated-measures correlation. Results showed that, under delay, timing-OSI decreases as latency increases and demonstrates construct validity against subjective QoE measures. Higher synchrony-OSI was associated with greater perceived synchronisation and satisfaction, and with lower perceived delay, irritation, and effort to follow a partner. These relationships were most consistent for offset synchrony and most selective for onset synchrony, while beat and tempo remained relatively stable. Under audio-quality degradation, ensemble-OSI remained relatively stable across sampling rates and did not significantly track subjective QoE as a single predictor. Instead, modest component-level associations suggested that satisfaction was higher when temporal stability and fidelity were preserved, whereas irritation was more closely related to reduced chord agreement. Together, these findings support timing-OSI as a promising objective synchrony metric for delay-impaired NMP, while showing that the extended ensemble-OSI requires further perceptual calibration for audio-quality degradations.

Academic Editors: Douglas O'Shaughnessy and Zong Woo Geem

Received: 17 March 2026
Revised: 27 May 2026
Accepted: 9 June 2026
Published: 11 June 2026

Keywords: Network Music Performance (NMP); deep learning (DL); audio delay; audio quality; convolutional recurrent neural networks (CRNN)

1. Introduction

Network Music Performance (NMP) enables geographically separated musicians to perform together in real time [1]. However, unlike many interactive media applications, NMP is constrained by exceptionally strict requirements on end-to-end delay and audio quality. Even small amounts of latency (on the order of only a few tens of milliseconds) can rapidly degrade coordination because musicians must maintain a shared tempo and coherent timing without being co-located [2]. In parallel, audio degradations introduced by compression, packet loss, or bandwidth constraints can affect timbral fidelity and the clarity of musical cues that performers rely on during ensemble interaction.

Most NMP research has therefore focused on Quality of Service (QoS), examining novel network architectures, codecs, and system frameworks designed to reduce latency. Rottondi et al. [3] provide a comprehensive overview of the technical challenges in achieving low latency under realistic wide-area network conditions. Although state-of-the-art tools such as Soundjack [4] and Aretousa [5] achieve low delays in controlled local-area network settings, delays increase substantially in remote settings due to Internet routing and queuing effects [6]. However, technical improvements alone do not resolve the Quality of Experience (QoE) problem. This paper focuses on a central aspect of QoE in NMP: ensemble synchrony**.** It remains difficult to quantify how network conditions translate into measurable changes in musicians' coordination.

Existing evaluations of synchrony and QoE in NMP settings often rely on subjective questionnaires [7], whereas objective assessments typically examine individual descriptors (e.g., tempo evolution or delay-related timing changes) in isolation [8,9]. Subjective ratings are difficult to generalise and diagnose, particularly under NMP conditions, as musicians adapt their strategies during performance in response to variable network conditions. Furthermore, objective approaches in Machine Learning (ML) and Music Information Retrieval (MIR) have typically not measured synchrony as a unified construct. Instead, they approach synchrony indirectly through expressive-performance modelling, temporal alignment, or the estimation of constituent synchrony-related components. As a result, there is a need for an objective, reproducible metric that quantifies synchrony holistically and tracks its evolution during performance under varying network conditions.

To fill this gap, we propose the Objective Synchrony Index (OSI), an AI-driven, multicomponent metric that quantifies ensemble synchrony from paired NMP recordings. The OSI models synchrony as an interaction phenomenon by decomposing musical coordination into interpretable components, including global timing offset and onset-level alignment, and combining them into a bounded composite score. Rather than reducing performance quality to a single network statistic, OSI operates on musicians' performance signals, enabling direct comparison of coordination across conditions.

We evaluated OSI under two experimental manipulations. In the delay study, we configured OSI to emphasise timing coordination and assessed its construct validity by relating OSI and its components to subjective QoE responses. In the quality study, we extended OSI to include additional descriptors intended to reflect how signal degradation affects ensemble interaction beyond timing. To compute these components consistently from audio, we developed an AI pipeline based on a two-tower multi-task CRNN that estimates timing- and harmony-related features from Musician A/Musician B NMP pairs.

The remainder of the paper is organised as follows. Section 2 positions OSI within prior work on NMP synchrony, objective measurement, and QoE assessment. Section 3

https://doi.org/10.3390/app16125919

details the OSI definitions and the AI pipeline used to compute OSI components from NMP recordings. Section 4 presents the objective OSI results under delay and quality conditions and examines their correspondence with subjective QoE. Section 5 discusses implications for evaluation practice in NMP, limitations of the current formulation, and directions for improving OSI as a measure of synchrony.

2. Background and Related Work

2.1. Synchrony as a Core Evaluation Problem in Networked Music Performance (NMP)

Synchrony, the temporal coordination between geographically distributed musicians, remains a central evaluation challenge in NMP. Research in NMP consistently frames synchrony as a core evaluation target because it is one of the first musical properties to fail as network latency and jitter increase [3]. Although audio latency can directly affect synchrony (how musicians hear each other), quality degradation affects the perception of synchrony more than the timing coordination itself [10].

Most NMP studies have focused primarily on the effects of delay, while the effect of audio quality degradation has received much less attention [2,11]. Prior work has identified an Ensemble Performance Threshold (EPT) [12] of roughly 20–30 ms one-way delay, beyond which ensemble coordination becomes notably difficult. Increased delay can alter musicians' interaction strategies as they attempt to compensate for shifted event timing, effects that are audible in the audio itself. In contrast, audio degradation primarily affects how comfortable musicians feel during performance and is therefore more commonly reflected in subjective experience [13]. However, as musicians can partially adapt their performance strategies under impairment, subjective questionnaires alone may not fully capture the underlying coordination dynamics. Together, these findings motivate evaluation approaches that quantify synchrony from performance signals rather than relying exclusively on post hoc questionnaires.

2.2. What Objective "Synchrony" Entails in Musical Coordination

In NMP, objective synchrony is a multi-layer coordination/interaction problem in which musicians continuously anticipate, attend to, and adapt to one another's actions in real time while pursuing shared performance goals [14]. Coordination is supported by a shared representation of the musical performance and by each musician's predictive models that simulate joint actions and outcomes, enabling rapid correction when timing diverges.

Objective synchrony includes multiple timing-based cues, ranging from global timing offset between musicians (the overall lag or lead between their audio streams), through micro-timing alignment of musical events (onsets, i.e., the moments when notes begin), to beat-level entrainment and longer-horizon tempo and phrasing coherence. Beyond timing, synchrony may also depend on the clarity of musical structure and the fidelity of the transmitted signal. For example, chord agreement can indicate whether performers maintain a consistent harmonic context, while signal fidelity reflects how clearly musical cues such as pitch, timbre, and transient attacks are preserved. Together, these elements capture complementary aspects of musicians' coordination in an NMP environment: timing alignment, structural coherence, and perceptual clarity [11,14–22].

2.3. Design Requirements for the Objective Synchrony Index (OSI)

We argue that a unified synchrony metric should include all components discussed above for three reasons: (1) synchrony in an NMP is inherently multidimensional, (2) individual descriptors capture only partial aspects of coordination, and (3) a unified metric enables consistent comparison across empirical NMP studies.

First, synchrony in NMP unfolds across multiple temporal and non-temporal cues. We have identified different layers of timing (onset alignment, global offset, beat entrainment, and tempo coherence) alongside non-temporal cues related to harmonic coherence and the clarity of the transmitted signal. Together, these cues shape how musicians coordinate with one another and how they perceive the interaction. Evaluating synchrony using only one descriptor (e.g., onset synchrony) risks overlooking important aspects of NMP. A composite weighted metric is therefore better suited to capturing the complementary nature of synchrony than any single timing variable alone.

Second, individual synchrony measures often differ across varying network and audio conditions. For example, as delay increases, onset and offset alignment may deteriorate while beat and tempo remain relatively stable; under audio-quality degradation, timing cues may remain intact while harmonic intelligibility and perceptual clarity change. Assessing overall coordination is difficult when these descriptors are evaluated independently, because each captures only one facet of the musical interaction. Combining them into a weighted index enables us to incorporate each component's relative contribution to NMP coordination, producing a single interpretable score while preserving the ability to inspect the individual components.

Third, a unified metric facilitates systematic evaluation and comparison in NMP research. Empirical studies often involve multiple network conditions (e.g., levels of delay), diverse musician groups (e.g., different levels of experience), and varied instrument types [11]. Researchers, therefore, need a consistent way to compare performance outcomes across experimental conditions. We propose the Objective Synchrony Index (OSI) in two variants: timing-OSI and ensemble-OSI**.** The first variant captures the timing-based components of coordination between two musicians (onset, offset, beat, and tempo) [14,15]. The second variant extends this formulation by adding chord agreement and signal fidelity, incorporating structural coherence and perceptual clarity alongside timing [18–21].

2.4. Evaluating Synchrony Using Subjective QoE Methods

As synchrony is a core component of the musicians' Quality of Experience (QoE), researchers frequently evaluate it using subjective QoE methods (e.g., electronic questionnaires) [11]. These methods, however, have well-known limitations (e.g., high response variance, individual differences in scale use), especially in NMP environments, where musical interaction is highly dynamic and adaptive. During an NMP, musicians may change their coordination strategy multiple times as conditions fluctuate. As a result, the same experimental condition can be experienced differently across participants and musical contexts. Subjective QoE methods, therefore, provide only a partial view of the musicians' experience and are rarely sufficient on their own to explain why experience changes or which aspects of coordination fail. We used the subjective QoE questionnaires collected in our empirical studies to evaluate the construct validity of the OSI variants (timing-OSI and ensemble-OSI) by computing statistical correlations between the items and OSI (including its components).

2.5. Using Machine Learning (ML) to Measure Synchrony

Machine Learning (ML) has been used in Music Information Retrieval (MIR) to measure musical synchrony mostly indirectly, rather than as a single end-to-end synchrony score. In the traditional ML literature, researchers have generally sought to model how musicians depend on one another during performance. For example, Marchini et al. [23] used ML to predict how each musician in a quartet shapes their performance when they are playing in an ensemble. Using separate regression models (trees, SVMs, and k-Nearest Neighbours) for each quartet member, they predicted note-level expressive parameters, such as timing deviations (how much a note is played earlier or later than notated) and

intensity (loudness). A key contribution of their work was the inclusion of cross-voice features, which incorporated information about what other musicians were playing at the same time. Their results showed that expressive performance modelling benefits significantly from inter-musician contextual information, supporting the idea that ensemble performance is shaped by musician interaction rather than by each musician acting alone.

In the MIR deep learning (DL) literature, researchers have shown that neural and recurrent models perform well on several core OSI-related tasks, including onset detection, beat tracking, and chord recognition. Böck et al. [8,9] demonstrated that a recurrent neural network can successfully detect onsets in a single audio stream in real time, and later extended their approach by combining an RNN with probabilistic decoding (RNN + DBN) for beat/downbeat tracking across diverse musical styles. Similarly, Zhou and Lerch showed that deep learning can successfully detect chords, outperforming earlier chord-recognition systems [19].

More recent MIR work in the DL literature has begun to move beyond single-stream subtask estimation towards ensemble alignment and analysis. Zeitler et al. [24] use raw onset and frame predictions from automatic music transcription (AMT) models for high-resolution audio-to-audio and audio-to-score alignment (i.e., aligning a recording with musical notation), showing that AMT-derived representations can be used to estimate temporal alignment between musical signals. In a related direction, Cheston et al. [25] introduced the Jazz Trio Database (JTD), a dataset of jazz piano trio recordings with automatically generated performer-level onset, beat, and downbeat annotations, together with MIDI for the piano soloist, enabling analyses of ensemble timing and inter-performer synchronisation.

To the best of our knowledge, however, no existing MIR work (ML or DL) directly computes OSI and its associated components as a unified synchrony metric. In Section 3.2.2, we present the Multi-Task Synchrony Network (MTSN), a two-tower Convolutional Recurrent Neural Network (CRNN) that estimates synchrony-relevant descriptors from asynchronously paired audio streams. The model extends prior work in several ways.

First, unlike the models of Böck et al. [8,9] and Zhou and Lerch [19], which operate on individual audio recordings, we have designed MTSN as a relational, multi-task model. Using a shared convolutional front end and separate rhythm and harmony towers, it jointly predicts multiple descriptors from paired NMP audio recordings via dedicated single-stream heads (onset, beat, tempo, chord identity, and fidelity), along with offset computations explicitly for each pair of musicians.

Second, our design aligns conceptually with the findings of Marchini et al. on the value of inter-voice contextual information. However, we have extended this relational principle to NMP. Rather than using cross-voice information to predict musical parameters for a sole performer, our model uses paired audio streams to estimate synchrony-relevant descriptors (offset, onset, beat, tempo, chord identity) and fidelity, which are then integrated into the final OSI descriptor.

Finally, OSI extends the score of Zeitler et al. [24] synchronisation work from temporal alignment between recordings, or between audio and musical score, of the same musical piece to a broader multi-dimensional dyad-level synchrony assessment of interacting performers integrating multiple timing cues and, in the ensemble variant, additional harmonic and signal fidelity components. The Jazz Trio Database (JDT) [25] is the closest recent work in its focus on performer-level ensemble analysis, but it provides annotations and analyses rather than a unified synchrony metric of the kind targeted by OSI.

3. Methods

3.1. Overview of the Objective Synchrony Index (OSI)

The Objective Synchrony Index (OSI) is a unified metric designed to quantify the degree of synchronisation between two (or more) musicians engaged in NMP. This produces a single composite score [0, 1], where values closer to 1 indicate tighter ensemble coordination. We built OSI from multiple synchrony-related components (e.g., global offset, onset alignment, beat/tempo coherence), each normalised to a component score $c_i \in [0, 1]$. OSI combines these descriptors using a normalised weighted average, where the denominator sums only the active weights for the specific OSI variant. The weights are not learned from the data; instead, we have manually defined them based on theoretical constraints of ensemble interaction.

$$OSI(general) = \text{clip}\left(\frac{\sum_{i\in A} w_i c_i}{\sum_{i\in A} w_i}, 0,1\right)$$

We implemented two variants of OSI, corresponding to the two empirical studies discussed in Section 3.4. In both variants, we specified the OSI coefficients as theory-driven, heuristic design weights to reflect the expected relative importance of each synchrony component under the corresponding experimental manipulation. In the delay variant, the theory emphasises temporal coordination [10,15,17] over beat and tempo. Thus, we assigned most of the weight to global offset (Offset = 0.25) and onset synchrony (Onset = 0.50), while down-weighting beat and tempo (beat = 0.05 and Tempo = 0.15) as supporting measures of rhythmic structure. We hypothesised that increased delay would worsen musicians' temporal coordination, thereby reducing OSI. In the quality variant, OSI assesses whether and how signal degradation affects synchrony within the performers' auditory feedback loop. The literature also supports keeping timing as the dominant block, with a lower emphasis on harmonic and perceptual clarity [14,20]. For this reason, this variant includes two related measures: (1) a timing-core OSI, computed from the offset, onset, beat, and tempo components (Offset = 0.20, Onset = 0.35, Beat = 0.15, Tempo = 0.20); and (2) an extended (ensemble) OSI, which additionally incorporates model-estimated chord agreement and an audio-fidelity term (Chord = 0.20, Fidelity = 0.10).

3.2. AI Pipeline for OSI Computation

We developed a multi-step AI pipeline that computes OSI scores for curated Musician A/Musician B NMP audio pairs grouped by two experimental conditions (delay and quality). The pipeline processes paired recordings that represent the musician's transmitted and received audio streams in NMP. For each pair, the two recordings are first standardised using a DSP frontend (see Section 3.2.1) to reduce recording artefacts and unwanted variability that can confound feature extraction. Each cleaned recording is then loaded at a fixed sample rate (44.1 kHz), downmixed to mono at load time, and transformed into a power Mel spectrogram using a Short-Time Fourier Transform (STFT) configuration (FFT = 2048, hop = 256, 80 mel bands spanning 30 Hz to 17 kHz).

The Multi-Task Synchrony Network (MTSN) is a two-tower multi-task Convolutional Recurrent Neural Network (CRNN) [26] that processes Mel spectrograms. The model has six output heads aligned with the OSI components: Beat, Onset, Tempo, Chord Identity, and Fidelity are estimated independently for each stream in the NMP pair, while the offset head estimates the relative lag between the paired inputs. These outputs form the basis for OSI computation. The model-derived descriptors (e.g., beat and onset probability sequences) are first time-aligned using the lag estimated by the offset head, and OSI is then computed as a weighted combination of bounded-similarity components.

We trained the CRNN using three datasets: the Artificial Audio Multitrack (AAM) [27] for music structure (beats, onsets, tempo, and chords), the Perceived Music Quality

Dataset (PMQD) [28] for musical fidelity, and a synthetically generated offset-pairs dataset derived from AAM tracks. For the synthetic offset-pairs dataset, we sampled fixed-length windows from the same track mix with a known time shift. The trained model computes all OSI-related features simultaneously through a combined learning signal.

When we first applied the pre-trained model to NMP Musician A/Musician B pairs, we observed that offset estimation did not transfer reliably to the NMP domain. Offset estimation requires robustness to cross-instrument differences, stream-specific capture, and experimental processing. To address this, we calibrated the offset estimation using a dedicated NMP-B offset-calibration dataset comprising live NMP Musician A/Musician B pairs under different delay conditions, separate from the NMP-A evaluation data used for the reported OSI analyses. During fine-tuning, we kept the shared feature extractor and all non-offset heads frozen to avoid drift in the other OSI components. The resulting offset-calibrated model was used to compute OSI variants for the NMP audio pairs under both experimental conditions.

3.2.1. Pre-Processing and Source Material

To prepare the data for OSI computation, we pre-processed 22 NMP Musician A/Musician B audio pairs (44 WAV files) using a DSP frontend before running the two-tower CRNN. The pipeline consists of two DSP stages: (1) signal cleaning and standardisation, and (2) feature extraction. In the first stage, we processed each recording independently using the same cleaning procedure to reduce recording artefacts and minimise unwanted variation across pairs. We began by applying stem separation (demucs, htdemucs) to suppress vocals by remixing the drums, bass, and other stems. After this, we applied the following conditional signal repair operations:

- Normalised programme loudness to −15 LUFS when the integrated loudness falls outside the range −18 to −12 LUFS;
- Applied denoising when the estimated SNR is below 22 dB;
- Dynamic levelling when the crest factor exceeds 20 dB; and
- Performed de-clicking when more than 10 single-sample pops are detected.

Once cleaning was complete, we re-encoded all processed audio to a common waveform format (44.1 kHz, 16-bit PCM WAV) to standardise the input for feature extraction.

In the second stage, we loaded each cleaned audio file at a fixed sample rate of 44.1 kHz and downmixed it to mono to produce a uniform input representation for the model. We then converted each waveform into a power Mel spectrogram using an STFT with a 2048-point FFT and 256-sample hop, mapping to 80 mel bands spanning 30 Hz to 17 kHz.

3.2.2. Multi-Task Synchrony Network (MTSN)

The architecture of MTSN follows a two-tower multi-task CRNN design (see Figure 1), with a shared convolutional frontend that takes a Mel spectrogram (80 mel bins over time) as input. In this shared stem, it applies a 2D convolution (1 input channel to 16 feature channels, 3 × 3 kernel) to learn local time–frequency patterns. The convolution scans small 3 × 3 regions of the spectrogram and produces sixteen feature maps that capture different acoustic cues. The model then applies frequency-only max pooling (2 × 1), which down-samples the mel axis from 80 to 40 while preserving full time resolution. This pooling step considers pairs of neighbouring mel bins and retains the strongest activation in each pair, without pooling across time (pool size 1 along the time axis).

After the shared musical feature extraction stage, the model splits into two parallel towers (see Figure 1). The rhythm tower focuses on timing-related information, such as onsets, beats, tempo, and offset representations, while the harmony tower captures pitch structure and harmonic context. This separation allows the model to treat temporal

structure and spectral harmonic properties of sound differently, while still benefiting from the shared low-level acoustic features learned in the earlier stage.

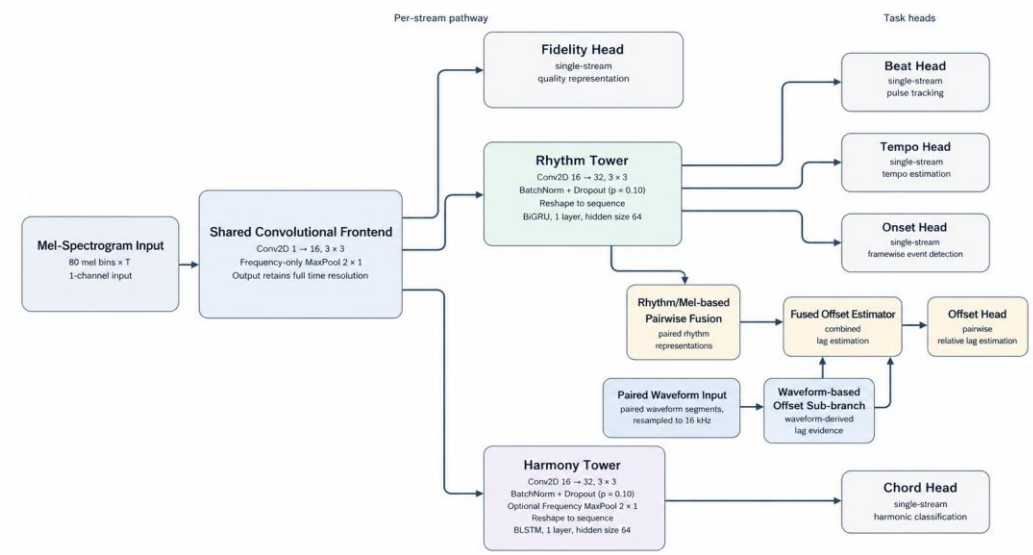

Figure 1. Multi-Task Synchrony Network (MTSN): a two-tower CRNN with shared convolutional features and task-specific rhythm and harmony branches.

The Rhythm Tower applies a 2D convolution that expands the feature channels from sixteen to thirty-two (3 × 3 kernel), followed by batch normalisation and dropout ($p = 0.10$). Batch normalisation stabilises intermediate time–frequency representations across heterogeneous training datasets. At the same time, dropout reduces overfitting and promotes feature representations that generalise across tasks (beat/onset/tempo/chord/fidelity) and transfer more robustly to NMP inference for OSI computation. It then reshapes the resulting feature map into a sequence of mel-frame vectors and passes them through a bidirectional Gated Recurrent Unit (GRU) (one-layer, hidden size 64). This recurrent layer aggregates time–frequency musical features into coherent rhythmic patterns, needed for beat, tempo, and offset estimation.

The Harmony Tower follows a similar structure. It begins with a 2D convolutional block (16 → 32 channels) with batch normalisation and dropout. We optionally apply a frequency-only max-pooling layer (2 × 1), reducing the frequency resolution from 40 to 20 to improve robustness to timbral differences across instruments, though at the cost of reduced spectral detail for chord discrimination. It then feeds the sequence into a bidirectional Long Short-Term Memory (LSTM) network (one layer, hidden size 64), which models longer-term harmonic dependencies, such as stable harmonic structures and chord transitions.

In its current configuration, the CRNN model includes six heads: five single-stream heads for onset, beat, tempo, chord, and fidelity, and one pairwise head for offset estimation. The single-stream head processes each audio stream of an NMP pair, while the offset head takes paired inputs and estimates the relative lag between them.

The onset head detects distinct musical events such as a note being struck, a drum hit, or a chord change. The beat head is the structural equivalent of the onset head but tuned for beat-level periodic events. The tempo head estimates audio speed (how fast events are happening), while the chord head decides which of the 26 types of chords musicians play. The fidelity head estimates a quality score for each NMP audio pair, and the offset head estimates the delay between them.

3.2.3. Training Procedure

We trained the MTSN model in two stages. In the first stage, we performed multi-source multi-task pre-training using the AAM structure dataset, a synthetic AAM offset-pairs dataset, and the PMQD fidelity dataset. Each input consisted of a fixed-length 512-frame Mel spectrogram segment computed at a sampling rate of 44.1 kHz using $n_{\text{fft}} = 2048$, a 256-sample hop size, 80 mel bins, and a frequency range from 30 Hz to 17 kHz. For batches drawn from AAM, the model optimised four tasks at once: chord classification over 26 classes using cross-entropy loss [29], beat detection using frame-wise binary cross-entropy with logits, onset detection using 31-channel frame-wise binary cross-entropy with logits, and tempo regression using mean-squared error [30]. Since onset labels are sparse in frame-wise music representations, we used an automatically estimated positive-class weight to reduce class imbalance. We normalised tempo targets by dividing by 240 BPM. For PMQD batches, the fidelity head minimised mean-squared error against a quality target normalised to [0, 1] using $(\text{rating} - 1)/4$. For synthetic offset batches, the offset head learned to predict the lag between two windows over a bounded set of delay classes.

At each training step, the model samples a data source, generates predictions for the relevant head(s), computes the corresponding losses, and aggregates them into a single weighted objective. We drew 60% batches from AAM to train beat, onset, tempo, and chord classification, 20% from synthetically generated offset pairs to train timing offset estimation, and 20% from PMQD to train the fidelity head. We used batch sizes of 64 for AAM, 64 for offset pairs, and 128 for PMQD.

We optimised all parameters jointly with the AdamW [31] optimiser using a fixed learning rate of 3×10^{-4} and weight decay of 1×10^{-4}. We clipped gradients to a global norm of 1.0 to stabilise parameter updates and used mixed precision training to improve computational efficiency. We trained for 10 epochs, each with 2000 optimisation steps, and selected the checkpoint with the lowest aggregate training loss to initialise the offset-head fine-tuning in the second training stage.

3.2.4. NMP-B Offset Calibration

When we first applied the Stage-1 MTSN to live NMP recordings (NMP-B dataset), the offset head produced unstable lag estimates. It frequently collapsed to 0 ms or to values near the search boundary. Because OSI heavily penalises large misalignments and saturates beyond 150 ms, these failures reduced the offset sub-score and, consequently, lowered the overall OSI.

We considered two complementary alignment strategies. A physics-style [32] delay estimation is effective when the two recordings share enough of the same underlying waveform (e.g., loopback of instruments through headphones), enabling direct delay estimation. A structure-based alignment [33] aligns recordings using shared rhythmic landmarks (onsets/beat patterns) across different instruments in the NMP audio pairs. However, NMP recordings mix multiple regimes and confounds, such as capture topology and pipeline processing, which can create multiple lag candidates and make it difficult to identify the true delay under either approach.

For stage 2, we extended the offset estimator with a waveform-based sub-branch and calibrated the resulting fused-offset component on NMP-B, a dataset reserved exclusively for offset calibration and completely disjoint from the live NMP dataset (NMP-A) used in OSI computations. We collected NMP-B in a different NMP study with 18 musicians and seven delay conditions (12, 22, 32, 42, 52, 62 and 72 ms). We used these nominal delay conditions as supervised offset targets for each Musician A/Musician B NMP pair and converted them into the model's frame-based lag representation. We excluded pairs when Musician A/Musician B files could not be resolved and restricted the remaining windows to the model's offset-calibration lag window, approximately ±160 ms.

To build the offset-calibration dataset, we partitioned NMP-B at the session level before window extraction, using a fixed-seed greedy grouped split that kept delay-condition histograms approximately balanced. In the final run, this split assigned 8 of the 10 NMP-B sessions to calibration training and 2 sessions to validation, yielding 108 directed training pairs and 28 directed validation pairs. We assigned every window from a given session to the same partition, which prevented leakage between calibration training and validation. We used the held-out validation split to select the fused-offset and waveform-refinement epoch counts. After model selection, we refit the final offset-calibrated checkpoint on the full NMP-B calibration corpus, comprising 136 resolved directed Musician A/Musician B pairs.

After partitioning NMP-B, we computed and cached a Mel spectrogram for each unique NMP recording to avoid recomputing features during training. For each Musician A/Musician B pair, we then sampled multiple fixed-length examples consisting of paired 512-frame Mel spectrogram windows and the corresponding waveform segments resampled to 16 kHz. Each sampled example retained the frame-based offset target derived from the pair's protocol-labelled delay condition. To focus training on musically informative regions, we computed a simple activity curve from mel energy and used it as a proxy reliability signal to weight training toward aligned windows more likely to contain timing landmarks.

We trained the offset component to predict delay as a discrete lag class. Rather than treating the supervised delay target as a single exact class, we represented it as a Gaussian-shaped target distribution [34] over nearby lag classes, so that predictions close to the nominal lag were penalised less than predictions further from the target. We combined this soft Gaussian target with a standard hard cross-entropy term [29] on the exact nominal lag class, allowing the model to retain a sharp lag estimate rather than a broad, ambiguous range. In the fused-offset calibration pass, we trained the fused, waveform, and rhythm offset outputs against these soft/hard lag targets, with a reliability weighting from the waveform-based offset gate. During the waveform-refinement pass, we also applied synthetic lag augmentation by adding a random extra lag within the model's calibrated lag window, approximately ±160 ms, and updating the target accordingly, while ensuring that both audio windows remained valid within the bounds of each recording.

We calibrated the offset component on the calibration dataset (NMP-B) in two steps. In the first step, we trained the fused offset component on the training partition and, after each epoch, evaluated the lag-prediction mean absolute error (MAE) on the validation partition. The best fused-offset checkpoint occurred at epoch 1, with a validation MAE of 2.858 frames, approximately 16.6 ms. In the second step, starting from this selected fused-offset checkpoint, we refined only the waveform-based offset branch using the same training/validation split and again selected the checkpoint with the lowest validation MAE. The best waveform-refinement checkpoint occurred at epoch 3, with a validation MAE of 5.286 frames, approximately 30.7 ms. We then used the selected fused-offset and waveform-refinement epoch counts to refit the offset calibration on the full NMP-B corpus. Specifically, we reran the fused-offset calibration on all 136 directed NMP-B pairs, followed by waveform-branch refinement on the same full calibration set. We used the resulting full-data refit checkpoint as the final model for OSI inference.

3.2.5. OSI Inference, Alignment, and Component Equations

For each curated NMP Musician A/Musician B pair, we computed OSI by extracting model features from each recording, estimating the relative lag between the pair, aligning the time-dependent descriptors on each Mel spectrogram frame, and then computing bounded-similarity scores for each OSI component. For the time-based components (offset, onset, beat, and tempo), we defined saturation thresholds that determine what counts

as similarity for each component. Differences at or above the corresponding threshold are clipped to a component score of 0 (maximal dissimilarity). In contrast, differences below the threshold are mapped linearly to 1, with small differences (e.g., within 5% of the threshold) producing scores close to 1 (maximal similarity). Below, we describe how we computed the components of the timing-core OSI:

We compute offset synchrony from the signed lag Δt_{ms} predicted by the offset head in milliseconds. For frame-wise descriptor alignment, we convert this lag into an integer frame shift s.

$$c_{\text{offset}} = \text{clip}\left(1 - \frac{\min(|\Delta t_{\text{ms}}|, \tau_{\text{offset}})}{\tau_{\text{offset}}}, 0,1\right)$$

$$h_{\text{ms}} = 1000 \cdot \frac{H}{f_s}$$

$$s = \text{round}\left(\frac{\Delta t_{\text{ms}}}{h_{\text{ms}}}\right)$$

In the offset formula, clip clamps the results to [0, 1]. Δt is the estimated effective lag between an audio Musician A/Musician B audio pair (typically in milliseconds), obtained from the offset head of the model. We set $\tau_{\text{offset}} = 150$ ms to saturate the offset component slightly beyond the maximum experimental delay of 120 ms. Given the STFT hop size H = 256 samples and sampling rate f_s = 44,100 Hz, each mel spectrogram frame corresponds to $h_{\text{ms}} \approx 5.80$ ms. Therefore, a lag of 150 ms corresponds to approximately 26 Mel spectrogram frames.

We compute the offset score c_{offset} from Δt_{ms} while we use the integer shift $s$to align frame-wise descriptor sequences before computing beat, tempo, onset, and chord agreement. An NMP audio pair has high offset synchrony when Δt_{ms} is small, and low offset synchrony when Δt_{ms} approaches the 150 ms threshold.

Beat synchrony measures frame-wise similarity between the aligned beat probability sequences

$$\text{c}_{\text{beat}} = \text{clip}\left(1 - \frac{1}{\text{T}}\sum_{\text{t=1}}^{\text{T}} \left|\text{b}_{\text{A}}[\text{t}] - \text{b}_{\text{B}}^{(\text{s})}[\text{t}]\right|, 0,1\right)$$

In the beat synchronisation formula, $\text{b}_{\text{A}}[\text{t}] - \text{b}_{\text{B}}^{(\text{s})}[\text{t}]$ is the difference between the beat probabilities of the in/out audio streams at frame (t) after applying the estimated effective lag (s). The estimated (s) reflect various latency factors, such as experimental delay and baseline end-to-end system latency, as well as the musician's compensatory behaviour (e.g., lead/lag between the streams resulting from the experimental delay).

An NMP audio pair $(\text{b}_{\text{A}}[\text{t}], \text{b}_{\text{B}}^{(\text{s})}[\text{t}])$ has a high beat synchrony when the probability curves agree over time, and low beat synchrony as their mean absolute difference increases.

Tempo synchrony uses a clipped, normalised tempo difference, averaged over time

$$c_{tempo} = \frac{1}{T}\sum_{t=1}^{T} clip\left(1 - \frac{\left|T_A[t] - T_B^{(s)}[t]\right|}{\tau_{tempo}}, 0,1\right)$$

In the tempo synchrony formula, $T_A[t] - T_B^{(s)}[t]$ is the absolute tempo disagreement of the in/out audio streams at frame (t) after alignment by the estimated shift (s). Absolute means only the magnitude of the difference matters. We set $\tau_{\text{tempo}} = 40$ BPM so that tempo similarity saturates only under large tempo divergences, and remains robust to model-estimation noise.

An NMP audio pair exhibits low tempo synchrony when the tempo difference exceeds the threshold (dt ≥ 40 BPM), and high tempo synchrony when differences remain small relative to 40 BPM (dt < 40 BPM).

Onset synchrony measures the alignment between onset time sequences.

$$c_{\text{onset}} = \text{clip}\left(\frac{2\left|M_{\tau_{\text{onset}}}\right|}{|O_A| + \left|O_B^{(s)}\right|} \left(1 - \frac{\text{median}(d_{ij})}{\tau_{\text{onset}}} \right), 0, 1 \right)$$

$$d_{ij} = \left| t_i^{(A)} - t_j^{(B,s)} \right|, \quad (i,j) \in M_{\tau_{onset}}$$

In the onset synchrony formula, O_A is the set of onset times extracted from the input stream A, and $O_B^{(s)}$ is the set of onset times extracted from the output stream B after applying the alignment shift s. $\left|M_{\tau_{\text{onset}}}\right|$ denotes the set of one-to-one matched onset pairs between the two streams (input A and aligned output $(B, s$) where matches are only permitted when onset times fall within the tolerance window τ_{onset}. For each matched pair $\left| t_i^{(A)} - t_j^{(B,s)} \right|, (i,j) \in M_{\tau_{onset}}$ provided the absolute time difference after alignment, i.e., how far apart are two matched notes in time. We set $\tau_{\text{onset}} = 120$ ms to match the maximum delay condition and to account for onset-detection uncertainty (e.g., small fluctuations in detected onset times caused by the model or signal processing).

An NMP audio pair has high onset synchrony when most onset events coincide (after alignment) and remain close in time. Conversely, a pair has low onset synchrony when few onsets can be matched, and they are far apart in time.

For the extended (ensemble) OSI, we computed the following:

Chord agreement measures the frame-wise match rate between predicted chord labels.

$$c_{chord} = \frac{1}{T} \sum_{t=1}^{T} I\left[k_A[t] = k_B^{(s)}[t] \right]$$

In the chord agreement formula, $k_A[t]$ is the predicted chord class ID for input stream A at frame t, and $k_B^{(s)}[t]$ is the predicted chord class ID for the output stream B at frame t, after applying the alignment shift s. An NMP audio pair achieves high chord agreement when both streams predict the same chord for most aligned frames and low chord agreement when the predicted chord labels differ for most frames.

Fidelity is a clip-level perceptual quality estimate predicted by the model that reflects the perceived audio quality of the received stream for each musician after transmission and processing.

$$c_{fidelity} = clip\left(\frac{f(x_{A\to B}^{out}) + f(x_{B\to A}^{out})}{2}, 0,1 \right)$$

In the fidelity agreement formula, $(x_{A\to B}^{out})$ is the received audio corresponding to the transmission from A to B. This stream contains the degradation in the receiver's sampling rate. $x_{B\to A}^{out}$ is the received audio corresponding to the transmission from B to A. This stream also includes the sampling-rate degradation experienced by the transmitter.

We computed the OSI and its associated component scores (above) for all curated NMP Musician A/Musician B pairs and aggregated the results by experiment. For the delay study (Study A in Section 3.4.1), we computed condition-level means for each delay value (ms). For the quality study (Study B in Section 3.4.1), we computed condition-level means for each sampling-rate condition (Hz).

3.2.6. Analyzing the Sensitivity of OSI Thresholds

We performed a sensitivity analysis of OSI to evaluate the hand-set scoring thresholds for the offset, onset and tempo components ($\tau_{\text{offset}} = 150$ ms, $\tau_{\text{onset}} = 120$ ms, and $\tau_{\text{tempo}} = 40$ BPM). We varied the value of each threshold independently ($\tau_{\text{offset}} \in$

{120,135,165,180} ms, $\tau_{\text{tempo}} \in \{32,36,44,48\}$ BPM, $\tau_{\text{onset}} \in \{96,108,132,144\}$ ms) and recomputed OSI while holding the model's outputs fixed.

We also varied the onset-event extraction settings to evaluate how sensitive OSI is to the way the model detects musical events. Specifically, we altered the initial peak detection threshold $\theta_{\text{onset}} \in 0.20,0.25,0.35,0.40$, which determines the model's confidence in treating a time point as a valid onset (lower values are more permissive, and higher values are more conservative). We then varied the minimum inter-peak gap $g_{\text{onset}} \in 30,40,60,70$ ms, which controls how closely spaced peaks must be before they are treated as separate events (small gaps are treated as separate events, while large gaps are treated as a single event). In addition, we adjusted the temporal smoothing window $w \in \{1,5\}$ frames, which affects how much short-term fluctuation is suppressed before onset detection (small values correspond to minimal smoothing while larger values produce a smoother curve). Finally, we tested alternative rules for combining instrument-level signals (31 instrument channels) into a single onset curve, including top-K aggregation (baseline), maximum, mean, and logical OR.

To evaluate robustness, we measured rank stability relative to baseline OSI using Spearman's ϱ and computed absolute score deviation using MAE, separately for the delay-based OSI and quality-based OSI variants (timing-core and ensemble).

3.3. Correlation of OSI to Subjective QoE

To evaluate the validity of OSI (and its constituent components) as a metric for QoE assessment, we used statistical correlations between OSI scores and musicians' questionnaire responses collected in our experimental studies (see Study A and Study B in Section 3.4.1). We correlated OSI and its components (offset, onset, beat, tempo, chord, and fidelity) with subjective ratings from questions assessing the musicians' QoE. Because both studies used a repeated-measures design, we analysed associations using the repeated-measures correlation (rmcorr), which quantifies within-musician covariation while controlling for between-musician baseline differences. This approach ensures that correlations reflect how changes in objective synchrony within a musician's track change their subjective experience, rather than being driven by individual differences in rating behaviour.

For the delay condition (Study A), we examined the correlations between subjective ratings and the timing-based OSI and its components. For the quality condition (Study B), we analysed both the timing-core OSI and the extended ensemble OSI (which additionally incorporates chord agreement and fidelity). This approach enabled us to test whether subjective QoE tracks temporal coordination alone or benefits from the inclusion of harmonic and signal-quality descriptors. We assessed robustness using Spearman-based repeated-measures correlation, complemented by cluster bootstrap confidence intervals at the musician level. Where applicable, we also fitted random-intercept mixed-effects models to obtain interpretable effect sizes while accounting for repeated observations within musicians.

3.4. Experimental Design

We used two separate rooms on the same floor of the university building for our experiments. Musicians performed in pairs, each listening to and watching the other through headphones and a 32″ monitor. Across both experiments, we used the same general topology but with slightly different setups for each scenario.

3.4.1. Factors and Conditions

In the first experiment (Scenario A), we varied the audio delay (Table 1) while keeping audio quality fixed. In the second experiment (Scenario B), we varied audio quality while keeping the audio delay fixed (Table 2).

Most studies on the impact of delay on NMP measure the Mouth-to-Ear (M2E) delay, which is the time between the microphone at one end and the headphones at the other. Measuring M2E requires precise clock synchronisation between those endpoints, which is only feasible using specialised equipment [35]. For this reason, in our study, we consider the My-Mouth-to-My-Ear (MM2ME) delay. MM2ME delay is the two-way version of the M2E delay (see Figure 2). Instead of measuring one-way transmission, however, it measures round-trip time: from when a musician plays a note into the microphone over the network to the other musician, and back through the headphones as a reply. Because musicians rely on quick feedback from fellow musicians to play in sync, MM2ME provides a meaningful measure of delay in the interaction loop.

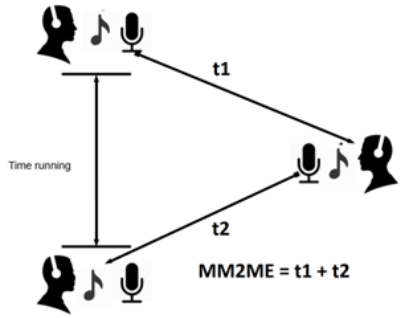

Figure 2. My-Mouth-to-My-Ear-delay.

Table 1. Scenario A: MM2ME delays.

Repetition	1	2	3	4	5	6	7	8	9	10
MM2ME delay (ms)	10	25	35	30	20	0	40	60	80	120

Table 2. Scenario B: MM2ME sampling rate.

Repetition	1	2	3	4	5	6	7	8	9	10
Sampling Rate (kHz)	44.1	36	28	22	16	12	8	18	48	88.2

3.4.2. Participants and Musical Material

In a within-subjects repeated-measures design, we asked 22 musicians to play a musical passage of their choice in pairs, using the instrument of their choice. Within each scenario, each pair selected a musical passage of their choice and performed the same one-minute passage 10 times, using the same instrument(s), under the variable delay conditions shown in Table 1 or the variable quality conditions shown in Table 2. By allowing musicians to use pieces they already knew and preferred, we minimised practice effects and increased confidence that our data accurately measured the impact of delay and quality on the musicians' synchrony and QoE. Although allowing musicians to use a piece and instrument of their choice introduced potential variability between pairs in repertoire and instrumentation, these factors were held constant within each pair across repeated conditions within a scenario.

3.4.3. Apparatus and Testbed

In Scenario A (see Figure 3), we used an eight-channel mixing console to route, monitor, and record audio in each room. Each musician listened to and watched the other through closed-type headphones and real-time HD video. We captured performances using condenser microphones. To achieve the lowest possible delay between musicians, we patched the monitor cables (shown as red lines in Figure 3) directly between the two

rooms, bypassing all network equipment. The visual delay between the HD camera and the TV monitor was 15 ms.

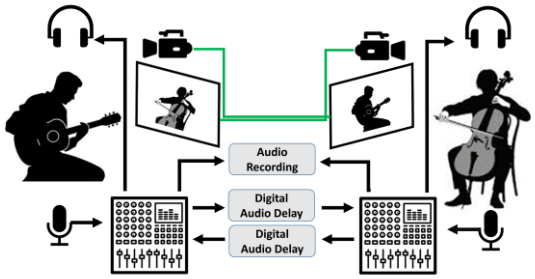

Figure 3. Experimental setup for Scenario A for controlled MM2ME delay between two remotely interacting musicians.

We also connected the two mixing consoles using direct cable patching, achieving a 10 ms delay, which is difficult to achieve when computers and network devices sit in the signal path. To manipulate the delay in either direction, we used AD-340 audio delay units made by Hall Research, Tustin, CA, USA.

We changed the configuration in the second study (Scenario B), as shown in Figure 4. In this setup, we routed the audio signals from the mixing consoles to two Linux-based PCs running with i7 processors and 12 GB RAM. Both PCs ran our in-house Aretousa software, version 1.0 [5] to capture and play back audio streams from the mixing consoles at the desired sampling rate (see Table 2). We did not modify the video configuration; the visual delay remained at 15 ms in both directions.

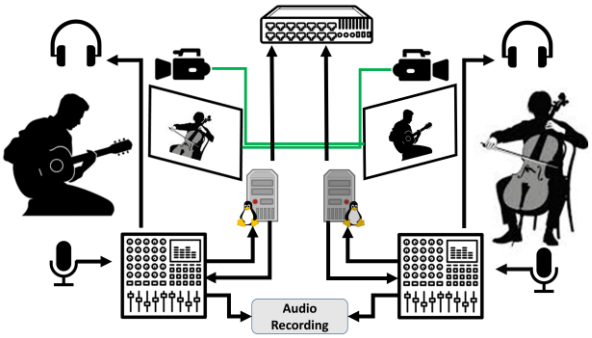

Figure 4. Experimental setup for Scenario B (variable quality), showing two remotely interacting musicians with controlled audio quality conditions while the MM2ME delay remained fixed.

3.4.4. Procedure

In Study A, each pair of musicians played the same one-minute passage at their own tempo and repeated it 10 times under different MM2ME delay conditions (see Table 1). In Study B, the same pair of musicians repeated the same one-minute passage 10 times under different sampling-rate conditions (see Table 2), while the audio delay remained fixed. At the end of each repetition, musicians completed a study-specific questionnaire assessing various aspects of their QoE, including synchrony. In the delay study (Study A), musicians rated their anxiety, perceived audio/video quality, irritation, perception of delay, satisfaction, perceived synchronisation, and the extent to which they actively tried to follow their partner. In the quality study (Study B), musicians rated their anxiety, irritation, perceived audio quality, and overall satisfaction. During the studies, we observed stable

performance across all musical groups, which gave us greater confidence that we had accurately measured the impact of delay and quality on their QoE.

4. Results

4.1. Sensitivity Analysis of OSI Thresholds

Across the three threshold sweeps ($\tau_{\text{offset}} \in \{120,135,165,180\}$ ms , $\tau_{\text{tempo}} \in \{32,36,44,48\}\, BPM$, $\tau_{\text{onset}} \in \{96,108,132,144\}$ ms) OSI remained highly stable, with nearly identical pairwise rankings relative to the baseline. In the quality study outputs, the worst Spearman correlation was $\varrho = 0.991$ for timing-OSI and $\varrho = 0.996$ for ensemble-OSI, with maximum mean absolute deviations of 0.0135 and 0.0102, respectively. The delay-study OSI output showed similarly high stability across the same tolerance sweeps, with a worst-case Spearman correlation of $\varrho = 0.994$ and a maximum mean absolute deviation of 0.0095.

For the three scoring-tolerance sweeps, the largest threshold-related deviations came from the tempo-tolerance sweep, while the offset- and onset-tolerance sweeps remained highly stable. Varying the onset-detection threshold (0.2–0.4) had negligible effects on the quality-study outputs: thresholds from 0.20 to 0.35 produced no measured change, and the 0.40 threshold produced only very small deviations (timing-OSI: MAE = 0.0005; ensemble-OSI: MAE = 0.0004). In the delay-study output, average deviations were also small (MAE ≤ 0.0018), and rankings remained highly stable ($\rho \geq 0.998$), reflecting the relative robustness of the onset extractor under both delay and quality conditions.

However, OSI was more sensitive to the structural choices for aggregating onset events. Alternative onset collapse rules led to larger deviations than the scoring-tolerance sweeps. For the quality-study timing-OSI, aggregation-rule variants produced MAE values of 0.0131–0.0394, while ensemble-OSI showed MAE values of 0.0098–0.0295. The delay-study OSI output showed a similar pattern, with aggregation-rule variants producing MAE values of 0.0190–0.0469. These changes indicate greater sensitivity to how onset-class evidence is combined, although the median shifts were mixed in direction rather than consistently inflated. Changes in the inter-peak gap produced only minor average effects (quality timing-OSI: MAE ≤ 0.0013; quality ensemble-OSI: MAE ≤ 0.0010; delay-study OSI: MAE ≤ 0.0004), while changes to the smoothing window produced larger but still non-catastrophic changes (quality timing-OSI: MAE = 0.0220–0.0270; quality ensemble-OSI: MAE = 0.0165–0.0202; delay-study OSI: MAE = 0.0287–0.0312).

4.2. Objective OSI Under Delay Conditions

We computed timing-OSI for the delay study (Study A) across 10 delay levels (0, 10, 20, 25, 30, 35, 40, 60, 80, and 120 ms, 11 musician pairs). Mean OSI values ranged from 0.365 to 0.459 across conditions, with the highest mean OSI observed at 10 ms (OSI = 0.459) and the lowest at 120 ms (OSI = 0.365) (see Figure 5). A Spearman rank correlation across the 10 conditions showed a strong negative association between delay and mean OSI ($\varrho = -0.52$). This reduction was primarily driven by the offset component (offset mean range: 0.680–0.958; $\varrho = -0.62$), followed by the onset component (onset mean range: 0.111–0.147; $\varrho = -0.37$). Beat synchrony remained consistently high across delay levels (0.962–0.970), while tempo synchrony varied modestly (0.432–0.510). These results indicate that under increasing delay, OSI is dominated by onset- and offset-level changes rather than beat- or tempo-based components.

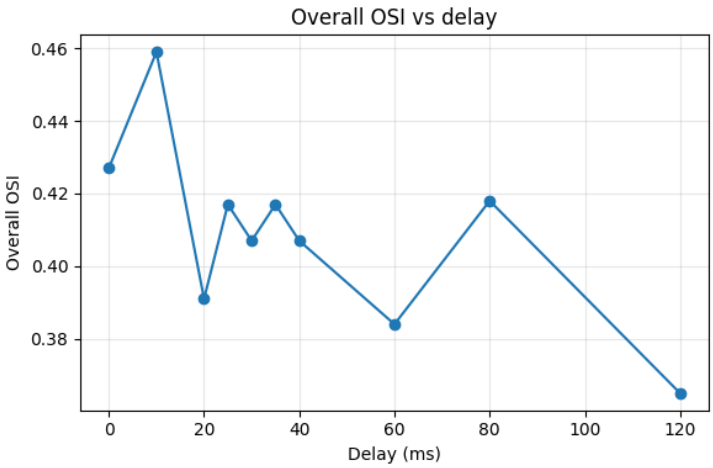

Figure 5. Mean timing-OSI across MM2ME delay conditions (Study A).

4.3. Objective OSI Under Quality Conditions

We computed ensemble-OSI (timing-OSI and extended ensemble-OSI) for the quality study (Study B) across 10 sampling-rate conditions (8 kHz to 88.2 kHz; n = 11 pairs per level). Across the 8 kHz to 48 kHz range, ensemble-OSI was relatively stable (0.536–0.576), and timing-OSI showed similarly limited variation (0.536–0.576), with no monotonic improvement as the sampling rate increased. Figure 6 shows that the highest mean ensemble-OSI occurred at 48 kHz (0.586), while the lowest occurred at 88.2 kHz (0.533), indicating that higher sampling rates did not correspond to stronger objective synchrony. Across conditions, beat synchrony remained consistently high (0.973–0.976), tempo synchrony varied moderately (0.513–0.602), and onset synchrony remained lower than the other timing components (0.131–0.225).

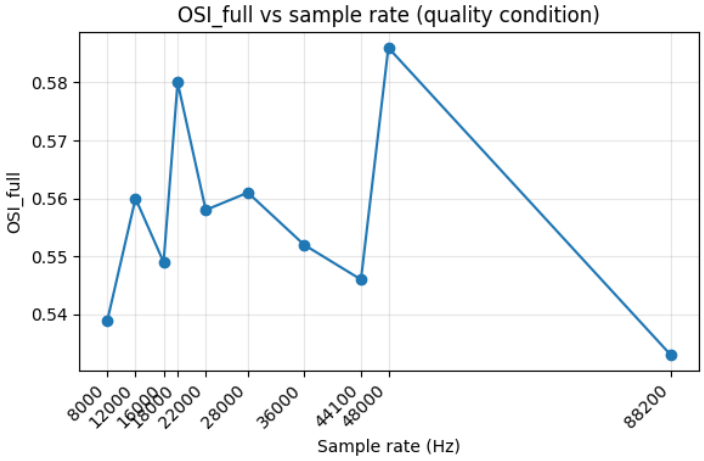

Figure 6. Mean ensemble-OSI across sampling rate conditions (Study B).

The ensemble components showed that chord agreement varied more substantially than fidelity (chord: 0.472–0.642; fidelity: 0.664–0.679), indicating that differences between timing-OSI and ensemble-OSI are more strongly influenced by chord agreement than by the fidelity score.

4.4. Linking OSI to Subjective QoE

To validate the OSI variants (timing-OSI and ensemble-OSI), we correlated their estimates (and components) with subjective QoE ratings. We computed repeated-measures correlations (rmcorr) between all questionnaire items and the objective OSI predictors. For

each study, we assessed within-musician covariation between the corresponding OSI and subjective ratings. This approach quantifies whether OSI can track musicians' ratings, while controlling for individual differences in baseline rating behaviour. Below, we report rmcorr values with 95% cluster-bootstrap confidence intervals and associated p-values. Furthermore, for the quality study, we conducted a descriptive condition-level analysis of the OSI outputs to examine which sampling-rate manipulations best preserved synchrony.

In the delay study (Study A), we found significant associations between timing-OSI and multiple QoE items (see Figure 7). Timing-OSI showed a strong association with perceived synchronisation (Q6: r = 0.330, $p \leq 0.001$, 95% CI [0.203, 0.445]) and a negative association with reported effort to follow (Q7: r = −0.276, $p < 0.001$, 95% CI [−0.388, −0.146]). It also showed negative associations with irritation (Q3: r = −0.216, $p = 0.002$, 95% CI [−0.290, −0.112]) and perceived delay (Q4: r = −0.231, $p = 0.001$, 95% CI [−0.362, −0.085]), as well as a weaker positive association with satisfaction (Q5: r = 0.148, $p = 0.037$, 95% CI [0.001, 0.280]).

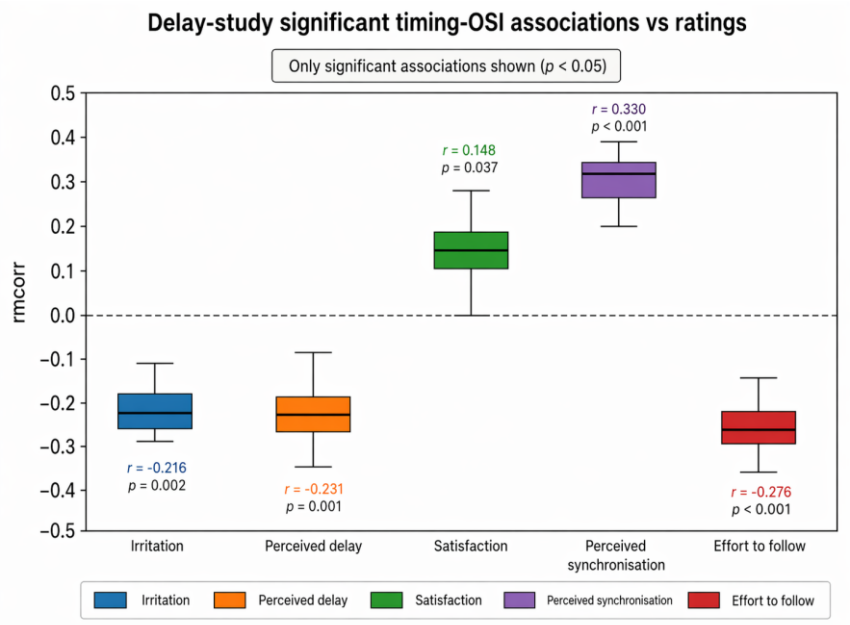

Figure 7. Repeated measures correlations between timing-OSI and subjective QoE ratings (Study A).

We also found strong component-level associations for offset and onset synchrony. Offset showed the strongest positive association with perceived synchronisation (Q6: r = 0.379, $p \leq 0.001$) and a significant positive association with satisfaction (Q5: r = 0.187, $p = 0.008$). Offset also showed significant negative associations with irritation (Q3: r = −0.260, $p < 0.001$), perceived delay (Q4: r = −0.284, $p < 0.001$), and reported effort to follow (Q7: r = −0.300, $p < 0.001$). Onset synchrony also showed significant but weaker associations with synchronisation and effort (Q6: r = 0.168, $p = 0.018$; Q7: r = −0.173, $p = 0.014$). Beat and tempo components did not show reliable associations with the subjective ratings.

In the quality study (Study B), ensemble-OSI did not show significant associations with any subjective rating (all $p \geq 0.065$), indicating that the ensemble-OSI, as defined, does not strongly track perceived QoE under sampling-rate manipulations. However, we observed modest but significant relationships for the following components (see Figure 8).

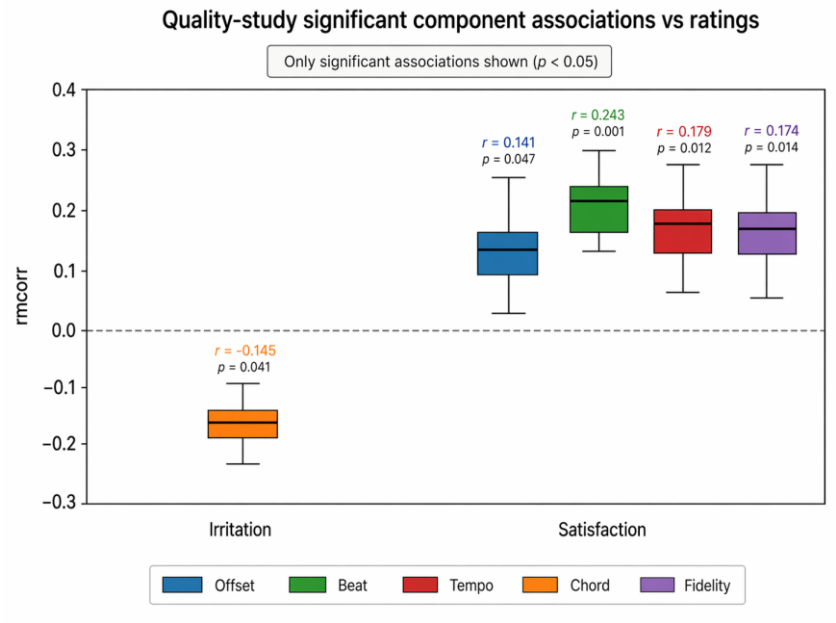

Figure 8. Component-level repeated-measures correlations between OSI components (Offset, Beat, Tempo, Chord, Fidelity) and subjective QoE ratings (Study B).

Offset, Beat, and Tempo synchrony showed a small positive association with satisfaction (r = 0.141, $p = 0.047$; r = 0.243, $p = 0.001$; r = 0.179, $p = 0.012$). When combined into a temporal synchrony composite, these components remained relatively strong across several conditions, including 8 kHz, 16 kHz, 18 kHz, 36 kHz, and 44.1 kHz, and were weakest at 88.2 kHz and 12 kHz. Fidelity showed a positive association with satisfaction (r = 0.174, $p = 0.014$). Chord agreement was negatively associated with irritation (r = −0.145, $p = 0.041$). Fidelity was highest at 12 kHz and remained relatively stable across conditions. In contrast, chord agreement varied more substantially, peaking at 16 kHz and reaching its lowest values at 22 kHz, 88.2 kHz, and 28 kHz. Onset synchrony did not show meaningful associations with any of the subjective ratings (all $p \geq 0.172$).

5. Discussion

Overall, the results for the single OSI predictors suggest that timing-OSI showed a strong negative association with experimental delay. In contrast, the ensemble-OSI did not show a monotonic improvement with increasing sampling rate. We expected this divergence, as OSI variants are weighted primarily toward temporal coordination cues. In the delay study (Study A), the systematic delay manipulation (10 levels, 0–120 ms) introduced phase-error into the musicians' auditory feedback loop, limiting their ability to make rapid timing corrections [2,10] and reducing tight coordination. In contrast, in the quality study (Study B), the sampling-rate manipulation primarily reduced spectral detail (fine structure and, indirectly, harmonic clarity) while preserving the temporal envelope cues that drive timing (onset and periodic pulse cues supporting beat and tempo, and, indirectly, cues used for offset estimation). As a result, the timing components of OSI do not show a clear monotonic trend across sampling-rate bands, keeping ensemble-OSI relatively stable.

At the component level, timing-OSI in Study A shows that onset and offset capture reduced local timing coordination under delay, while beat and tempo reflect higher-level metrical stability, i.e., maintenance of a shared pulse and overall tempo coherence [2,14,15]. These findings are consistent with musicians maintaining a shared pulse even when fine-grained event alignment has degraded. In Study B, the timing components (offset, onset, beat, tempo) remained stable across sampling-rate manipulation, and the ensemble terms (chord agreement and fidelity) did not increase monotonically with

sampling rate. Together, these component findings help explain why timing-OSI tracked delay effects more clearly than ensemble-OSI tracked sampling-rate changes.

The subjective (QoE) results reinforce our interpretation above. Timing-OSI correlated positively with perceived synchronisation and satisfaction. It correlated negatively with perceived delay, irritation, and reported effort to follow a partner; these relationships were most consistent for offset synchrony and more selective for onset synchrony. These findings align with the literature, which shows that temporal delay disrupts performers' ability to maintain synchronisation [36], while onset cues remain perceptual cues for evaluating ensemble togetherness [37]. The association with effort is also consistent with prior work: onset misalignment increases the predictive and corrective demands placed on musicians, which have been linked to measurable increases in mental effort in rhythmic tasks [38]. Because beat and tempo synchrony remained high across delay conditions and exhibited limited variance, they had limited explanatory power for subjective ratings.

In the quality study, Ensemble-OSI as a single predictor did not show a significant association with any subjective (QoE) item. One possible interpretation for the positive associations between satisfaction and offset, beat, and tempo synchrony is that musicians experienced temporal coordination as broadly usable across most sampling-rate manipulations. In this context, the positive association between fidelity and satisfaction suggests modest sensitivity to signal clarity: trials with slightly better acoustic detail were perceived as more satisfying. In contrast, the negative association between chord agreement and irritation suggests that reduced harmonic coherence may have made the ensemble sound more perceptually unstable, leading to irritation. However, given the weak effect sizes and the absence of a significant overall Ensemble-OSI–QoE association, we consider these patterns exploratory rather than conclusive.

The component-level findings above should be interpreted in light of a limitation in the current MTSN model. Although we calibrated the offset head using a separate NMP-B dataset, the non-offset heads were not directly validated against NMP-specific labelled annotations. To the best of our knowledge, no publicly available NMP-specific labelled corpus currently provides the frame-, event-, segment-, and clip-level annotations required to validate these heads directly. Creating such a corpus would be a substantial task. For example, validating the onset head would require paired musicians to perform under controlled NMP delay and audio-quality conditions, with separate Musician A/Musician B recordings manually annotated for onset events across instruments, musical roles, overlapping musical textures, performer adaptation strategies, and transmission-induced degradation. Therefore, we treated the onset, beat, tempo, chord, and fidelity outputs as model-derived descriptors within the OSI framework, rather than as independently validated NMP-domain measures. Component-level interpretations based on these model-derived descriptors should remain cautious, since changes in their values may reflect synchrony-related variation but may also be influenced by domain shift arising from live capture, instrument differences, room acoustics, network processing, and signal degradation.

Finally, the sensitivity analysis indicates that both timing-OSI and ensemble-OSI are numerically stable to moderate variations in the scoring tolerances ($\tau_{\text{offset}}, \tau_{\text{onset}}, \tau_{\text{tempo}}$), while the largest sensitivity arises from structural choices in onset-event definition. These results indicate that OSI is robust as a metric under reasonable threshold variations, but also that onset aggregation is a design-critical factor in the metric definition.

6. Conclusions and Future Work

This paper presents an AI-driven objective metric to measure synchrony, a critical aspect of the Musician's QoE in Network Music Performance (NMP) environments. The findings for each OSI variant highlight two important implications for its design. First, timing-OSI showed the clearest construct validity against subjective QoE under delay,

primarily because it measures timing misalignment (especially global lag and onset-level asynchrony) that musicians can readily notice and must actively compensate for during the study (Study A). At the component level, the main design implication is that offset synchrony should remain central to timing-OSI, while onset synchrony should be treated as a more selective cue for perceived synchronisation and effort. Second, ensemble-OSI did not significantly track subjective QoE as a single predictor under sampling-rate manipulation, although modest component-level effects suggested that satisfaction was higher when temporal stability and fidelity were preserved, whereas irritation was more closely related to reduced chord agreement. Ensemble-OSI may therefore require additional perceptual descriptors such as harmonic clarity (the intelligibility of pitch and chord structure) [39] and transient-audibility (the perceptibility of note attacks and other rhythmic transients under degradation or masking) [40]. This would enable it to capture not only whether two streams are temporally aligned, but also whether the cues required for musicians to perceive and act on that alignment remain audible under degraded audio conditions.

Future work will focus on releasing an open-source version of MTSN for both post hoc and real-time OSI analysis. We plan to design a low-latency streaming variant of the current architecture using sliding-window inference and shorter analysis windows to enable incremental estimation of the main OSI descriptors: onset, beat, tempo, chord, and fidelity. In parallel, we are exploring approaches for adapting the offset head and the OSI score-aggregation strategy for real-time deployment. Pairwise offset estimation is inherently more difficult than the remaining descriptors because it requires direct comparison between two streams and reliable alignment between them. We will also prioritise NMP-specific validation and further calibration of the onset component, since the sensitivity analysis showed that OSI is robust to scoring-threshold variation but more sensitive to structural choices in onset-event extraction and aggregation. We will also validate and, where necessary, recalibrate the remaining non-offset heads (beat, tempo, chord, and fidelity) using NMP-specific labelled data before treating them as standalone NMP-domain measures. Finally, we will explore which perceptual descriptors could enrich ensemble-OSI and help it better capture synchrony under degraded audio conditions.

Using the updated MTSN as a core synchrony mechanism for musical performance, we plan to conduct a larger empirical study with 50 musicians (25 pairs). This study will help us gain deeper insights into musicians' perceived synchronisation and enable a more robust validation of OSI against subjective QoE, particularly for the weaker ensemble-OSI component effects observed under audio degradation.

The second direction of future work focuses on learning the relative importance of OSI components directly from the data rather than using fixed weights. During asynchronous analysis, instead of specifying the relative importance of offset, onset, beat, tempo, chord, and fidelity, the extended framework will include a regression model to learn how each component contributes to predicting synchrony. Finally, we plan to introduce instrument- and role-aware modelling into the OSI framework. In NMP settings, musicians rely on different auditory cues for synchronisation depending on their instrument and musical role; for example, percussive instruments tend to prioritise transient timing cues, while harmonic instruments rely more on pitch and timbral clarity. Similarly, different musical roles (e.g., accompaniment or melodic) require musicians to rely on different coordination cues.

We are currently building a prototype as a musician-specific feedback companion for real-time NMP environments. The prototype enables musicians to manually select their instrument type and role via a user interface before the performance. It also displays the descriptor-level and aggregate synchrony scores relative to the ensemble (two or more musicians), based on the learned OSI coefficients.

Author Contributions: Conceptualization, I.D.; Methodology, I.D. and K.T.; Software, I.D.; Validation, I.D.; Formal analysis, I.D.; Investigation, I.D. and K.T.; Resources, I.D.; Data curation, I.D.; Writing—original draft, I.D. and K.T.; Writing—review and editing, I.D., K.T. and G.X.; Visualization, I.D.; Supervision, I.D.; Project administration, I.D.; Funding acquisition, I.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The questionnaire responses used in the study are openly available in GitHub, at https://github.com/mmlab-aueb/nmp/tree/master/Datasets/audio, accessed on 8 June 2026. The audio recordings, the MTSN model and the data generated for this study are available on request from the authors.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Lazzaro, J.; Wawrzynek, J. A case for network musical performance. In Proceedings of the 11th International Workshop on Network and Operating Systems Support for Digital Audio and Video, Braunschweig, Germany, 28–30 May 2001; pp. 157–166.
2. Tsioutas, K.; Xylomenos, G. Assessing the effects of delay to NMP via audio analysis. *SN Comput. Sci.* **2022**, *4*, 126.
3. Rottondi, C.; Chafe, C.; Allocchio, C.; Sarti, A. An overview on networked music performance technologies. *IEEE Access* **2016**, *4*, 8823–8843.
4. Carôt, A.; Werner, C. Distributed network music workshop with soundjack. In Proceedings of the 25th Tonmeistertagung, Leipzig, Germany, 13–16 November 2008. Available online: http://www.carot.de/Docs/TMT08.pdf (accessed on 8 June 2026).
5. Tsioutas, K.; Xylomenos, G.; Doumanis, I. Aretousa: A competitive audio streaming software for network music performance. In Proceedings of the Audio Engineering Society Convention, Dublin, Ireland, 20–23 March 2019; Volume 14.
6. Lakiotakis, E.; Liaskos, C.; Dimitropoulos, X. Improving networked music performance systems using application-network collaboration. *Concurr. Comput. Pract. Exp.* **2019**, *31*, e4730. https://doi.org/10.1002/cpe.4730.
7. Comanducci, L. Intelligent Networked Music Performance Experiences. In *Special Topics in Information Technology*; Riva, C.G., Ed.; Springer International Publishing: Cham, Switzerland, 2023; pp. 119–130.
8. Böck, S.; Arzt, A.; Krebs, F.; Schedl, M. Online real-time onset detection with recurrent neural networks. In Proceedings of the Proceedings of the 15th International Conference on Digital Audio Effects (DAFx-12), York, UK, 17–21 September 2012; pp. 17–21.
9. Böck, S.; Krebs, F.; Widmer, G. Joint Beat and Downbeat Tracking with Recurrent Neural Networks. In Proceedings of the ISMIR, New York, NY, USA, 7–11 August 2016; pp. 255–261.
10. Washburn, A.; Wright, M.J.; Chafe, C.; Fujioka, T. Temporal Coordination in Piano Duet Networked Music Performance (NMP): Interactions Between Acoustic Transmission Latency and Musical Role Asymmetries. *Front. Psychol.* **2021**, *12*, 707090. https://doi.org/10.3389/fpsyg.2021.707090.
11. Tsioutas, K.; Xylomenos, G.; Doumanis, I. Impact of Audio Delay and Quality in Network Music Performance. *Future Internet* **2025**, *17*, 337.
12. Schuett, N. The Effects of Latency on Ensemble Performance. Bachelor's Thesis, CCRMA Department of Music, Stanford University, Stanford, CA, USA, 2002**.**
13. Goot, D.K.; Chaubey, H.; Hsu, T.Y.; Deal, W.S. A perceptual evaluation of music real-time communication applications. *IEEE Access* **2023**, *11*, 46860–46870.
14. Keller, P.E.; Novembre, G.; Loehr, J. Musical Ensemble Performance: Representing Self, Other and Joint Action Outcomes. In *Shared Representations: Sensorimotor Foundations of Social Life*; Obhi, S.S., Cross, E.S., Eds.; Cambridge Social Neuroscience; Cambridge University Press: Cambridge, UK, 2016; pp. 280–310.
15. Repp, B.H.; Su, Y.H. Sensorimotor synchronization: A review of recent research (2006–2012). *Psychon. Bull. Rev.* **2013**, *20*, 403–452. https://doi.org/10.3758/s13423-012-0371-2.
16. Timmers, R.; Endo, S.; Bradbury, A.; Wing, A.M. Synchronization and leadership in string quartet performance: A case study of auditory and visual cues. *Front. Psychol.* **2014**, *5*, 645.
17. D'Amario, S.; Daffern, H.; Bailes, F. Perception of synchronization in singing ensembles. *PLoS ONE* **2019**, *14*, e0218162.

18. Harte, C.; Sandler, M.; Gasser, M. Detecting harmonic change in musical audio. In *Proceedings of the 1st ACM Workshop on Audio and Music Computing Multimedia*; ACM: New York, NY, USA, 2006; pp. 21–26.
19. Zhou, X.; Lerch, A. Chord detection using deep learning. In Proceedings of the 16th ISMIR Conference, Málaga, Spain, 26–30 October 2015; Volume 53, p. 152.
20. McAdams, S. The perceptual representation of timbre. In *Timbre: Acoustics, Perception, and Cognition*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 23–57.
21. Siedenburg, K.; Saitis, C.; McAdams, S. The present, past, and future of timbre research. In *TIMBRE: Acoustics, Perception, and Cognition*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 1–19.
22. Tsioutas, K.; Xylomenos, G.; Doumanis, I.; Angelou, C. Quality of musicians' experience in network music performance: A subjective evaluation. In Proceedings of the Audio Engineering Society Convention, Online, 2–5 June 2020.
23. Marchini, M.; Ramirez, R.; Papiotis, P.; Maestre, E. The sense of ensemble: A machine learning approach to expressive performance modelling in string quartets. *J. New Music. Res.* **2014**, *43*, 303–317.
24. Zeitler, J.; Maman, B.; Müller, M. Robust and Accurate Audio Synchronization Using Raw Features From Transcription Models. In Proceedings of the ISMIR, Online, 10–14 November 2024; pp. 120–127.
25. Cheston, H.; Schlichting, J.L.; Cross, I.; Harrison, P.M. Jazz trio database: Automated annotation of jazz piano trio recordings processed using audio source separation. *Trans. Int. Soc. Music. Inf. Retr.* **2024**, *7*, 144–158.
26. Heydari, M.; Cwitkowitz, F.; Duan, Z. BeatNet: CRNN and Particle Filtering for Online Joint Beat, Downbeat and Meter Tracking. 2021. Available online: https://archives.ismir.net/ismir2021/paper/000033.pdf (accessed on 8 June 2026).
27. Ostermann, F.; Vatolkin, I.; Ebeling, M. AAM: A dataset of Artificial Audio Multitracks for diverse music information retrieval tasks. *EURASIP J. Audio Speech Music. Process.* **2023**, *2023*, 13. https://doi.org/10.1186/s13636-023-00278-7.
28. Hilmkil, A.; Thomé, C.; Arpteg, A. Perceiving music quality with GANs. *arXiv* **2020**, arXiv:2006.06287.
29. Yang, Z.; Li, Z.; Gong, Y.; Zhang, T.; Lao, S.; Yuan, C.; Li, Y. Rethinking knowledge distillation via cross-entropy. *arXiv* **2022**, arXiv:2208.10139**.**
30. Terven, J.; Cordova-Esparza, D.-M.; Romero-González, J.-A.; Ramírez-Pedraza, A.; Chavez-Urbiola, E.A. A comprehensive survey of loss functions and metrics in deep learning. *Artif. Intell. Rev.* **2025**, *58*, 195.
31. Loshchilov, I.; Hutter, F. Decoupled weight decay regularization. *arXiv* **2017**, arXiv:1711.05101.
32. Fallon, M.; Godsill, S.; Blake, A. Joint acoustic source location and orientation estimation using sequential Monte Carlo. In Proceedings of the International Conference on Digital Audio Effects (DAFx-06), Montreal, QC, Canada, 18–20 September 2006; pp. 203–208.
33. Müller, M. Music Synchronization. In *Fundamentals of Music Processing: Using Python and Jupyter Notebooks*; Müller, M., Ed.; Springer International Publishing: Cham, Switzerland, 2021; pp. 119–170.
34. Xu, H.; Liu, X.; Zhao, Q.; Ma, Y.; Yan, C.; Dai, F. Gaussian label distribution learning for spherical image object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 1033–1042.
35. Lindborg, P.; Kwan, N.A. Audio Quality Moderates Localization Accuracy: Two Distinct Perceptual Effects? In *Proceedings of the Audio Engineering Society Convention 138*; Audio Engineering Society: New York, NY, USA, 2015.
36. Chafe, C.; Caceres, J.-P.; Gurevich, M. Effect of temporal separation on synchronization in rhythmic performance. *Perception* **2010**, *39*, 982–992.
37. Wing, A.M.; Endo, S.; Yates, T.; Bradbury, A. Perception of string quartet synchronization. *Front. Psychol.* **2014**, *5*, 1115. https://doi.org/10.3389/fpsyg.2014.01115.
38. Skaansar, J.F.; Laeng, B.; Danielsen, A. Microtiming and Mental Effort: Onset Asynchronies in Musical Rhythm Modulate Pupil Size. *Music. Percept.* **2019**, *37*, 111–133. https://doi.org/10.1525/mp.2019.37.2.111.
39. Yost, W.A. Pitch perception. *Atten. Percept. Psychophys.* **2009**, *71*, 1701–1715. https://doi.org/10.3758/APP.71.8.1701.
40. Parker, A.; Fenton, S. Musical Mix Clarity Prediction Using Decomposition and Perceptual Masking Thresholds. *Appl. Sci.* **2021**, *11*, 9578. https://doi.org/10.3390/app11209578.