Feature-based Analysis of the Effects of Packet Delay on Networked Musical Interactions*

Cristina Rottondi, Michele Buccoli, Massimiliano Zanoni, Dario Garao, Giacomo Verticale, and Augusto Sarti

{cristinaemma.rottondi,michele.buccoli,massimiliano.zanoni,dario.garao,giacomo.verticale,augusto.sarti}@polimi.it

Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Piazza Leonardo da Vinci, 32, Milano, Italy

Networked Music Performance (NMP) is a mediated interactional modality with a tremendous potential impact on professional and amateur musicians, as it enables real-time interaction from remote locations. One of the known limiting factors of distributed networked performances is the impact of the unavoidable packet delay and jitter introduced by IP networks, which make it difficult to keep a stable tempo during the performance. This paper investigates the tolerance of remotely interacting musicians towards adverse network conditions. We do so for various musical instruments and music genres, as a function of rhythmic complexity and tempo. In order to conduct this analysis, we implemented a testbed for psycho-acoustic analysis emulating the behavior of a real IP network in terms of variable transmission delay and jitter, and we quantitatively evaluated the impact of such parameters on the trend of the tempo maintained during the performance and on the perceptual quality of the musical interaction.

0 Introduction

Low-latency communication systems are a key requirement for a wide category of innovative applications, ranging from virtual machine mobility for cloud computing services to video conferencing and telepresence, real-time financial and business transactions. Among the many applications that are crucially dependent on latency minimization, that of Networked Music Performance (NMP) promises to revolutionize interactive music fruition (e.g. remote rehearsals, music teaching) by allowing remote players to interact with each other in a musical performance from remote physical locations through an Internet connection over a telecommunication network. Though computer-based systems enabling music performance have been investigated starting from the '70s (see [2] for an historical overview), in the past two decades the massive diffusion of Internet greatly widened the opportunities for new forms of musical interactions. Several experimental software applications have been recently developed, which support both real-time and latency-accepting NMP [27, 5]. Moreover, different network architectures have been investigated as enabling design paradigms for NMP systems, ranging from client-server [29, 19] and master-slave [28] to decentralized peer-to-peer infrastructures [31, 11].

In order to reproduce realistic environmental conditions for NMP, several technical, psycho-cognitive and musicological issues must be addressed. In particular, at network level, very strict requirements in terms of latency and jitter must be satisfied to keep the one-way end-to-end transmission delay below a few tens of milliseconds. Typically the delay tolerance is estimated to be 20-30 ms [8] (corresponding to the time that the sound field takes to cover a distance of 8-9 m), which has been shown to correspond to the maximum physical separation beyond which keeping a common tempo for rhythmic music interaction without conductor becomes difficult. In NMP, the overall delay experienced by the players includes multiple contributions due to different stages of the audio signal transmission: the first is the processing delay introduced by the audio acquisition, processing, and packetization; the second is the pure propagation delay over the physical transmission medium; the third is the data processing delay introduced by the intermediate network nodes traversed by the audio data along their path from source to destination, the fourth is the playout buffering which might be required to compensate the effects of jitter in order to provide sufficiently low packet losses to ensure a target audio quality level.

*To whom correspondence should be addressed Email:cristinaemma.rottondi@polimi.it

Some preliminary studies on the delay tolerance for live musical interactions have already appeared: in [20, 12, 14, 13] the authors evaluated the trend of tempo variations (measured in Beats Per Minute - BPM) while performing predefined rhythmic patterns through hand clapping, in different latency conditions. A similar analysis was integrated with an evaluation of the musicians' subjective rating of the performance quality in [9].

The sensitivity to delay and the quality of the musical experience in the context of NMP is influenced by several additional factors [4]: in [3], the authors investigate the correlation between perceptual attack times of the instrument and the sensitivity to network delay, concluding that instruments with a slow attack (e.g. strings) usually tolerate higher latency. In [15], the authors investigate the correlation between accepted network latency and the genres of pattern-based music.

To the best of our knowledge, a quantitative study of the sensitivity to delay and quality and its dependency on additional parameters such as the rhythmic complexity of the performed piece, the timbral characteristics of the instruments and the type of musical part that is being performed (e.g. melody, chord comping, sustained harmony) has not yet been proposed in the literature. Therefore, taking advantage of the feature-based analysis proposed in [23, 32] in this study we provide an evaluation of the impact of network conditions on the quality of the musical experience, according to the type of the instruments and to some characteristics of the performance. As far as the type of instrument is concerned we adopt a timbral feature-based representation, whereas we exploit musical part, Event Density [24] and Rhythmic Complexity [26] of the performed pieces to characterize the performance.

It is also worth noticing that all the previous studies in the literature conducted in controlled network environments do not consider the effect of packet jitter on the end-to-end latency and assume that each packet carrying audio data experiences exactly the same delay. Such assumption is quite unrealistic for actual telecommunication networks, in which jitter is by far not negligible and must be compensated by the receiver's buffer. The buffer must be sized to strike a balance between the additional packet delay due to queuing at the receiver side and the audio glitches due to buffer overruns and underruns. For this reason, we implemented a testbed for psycho-acoustic analysis which emulates the behavior of a real IP network in terms of variable transmission delay and jitter by generating random packet delays according to any desired statistical distribution. In our testbed, we opted for a peer-to-peer solution based on the publicly available SoundJack software [7], which also implements a direct real-time evaluation of the experienced one-way end-to-end latency (thus including processing, buffering and playout delays).

The remainder of the paper is organized as follows: Section 1 provides a brief definition of the extracted musical features. Our experimental testbed is described in Section 3. Results obtained from the performed psycho-acoustic tests are analyzed in section 4. Finally, we draw our conclusions in Section 5.

1 Background

Numerical attributes that can be extracted from musical audio signals are typically categorized into timbral and rhythmic features. Timbral features typically describe some short-time properties of the sound (e.g. spectral information, or zero-crossings), whereas rhythmic content tend to describe longer-term properties (e.g. beat, tempo, pitch changes). As the timbral characteristics of a musical instrument are reflected in the spectral distribution of the generated musical signal, spectral features are widely used in the literature for the characterization of a musical instrument (see [23, 33, 25, 10] for a comprehensive feature enlisting and mathematical definition).

In [21] the authors show that the timbre of the instruments can highly influence the emotional state elicited in the player or the audience during a performance. The experiment was conducted by asking musicians to play the same list of melodies using instruments with different timbres in presence of an audience. Players and members of the audience were then asked to express the emotion perceived during each performance. According to the study, the sound qualities of the instruments have a significant impact on the perceptual quality of the performance. In particular, noisy sounds can alter the perception of note transients and durations. For these reasons it worth analyzing the impact of instrument timbre on the quality of the musical experience and the sensitivity to delay.

In [26], the authors provide a definition of the perceived rhythmic complexity as the capability of the listener to clearly perceive a repetitive pattern in the rhythm and to decompose the pattern into a simpler structure. Playing a piece with a complex rhythmic pattern requires the musicians to strongly keep the same tempo, i.e., to be *synchronized* during the execution. In the real-time NMP, however, the musicians perceive the note onsets of the other players as shifted in time, due to the effect of the introduced delay. Consequently, it is important to address how the subjectively perceived rhythmic complexity level influences the tolerance to the delay introduced by NMP.

In [3], the authors show that the human auditory system focuses on onsets produced by instruments with a short or almost impulsive attack time (i.e., the time that the instrument takes to reach its maximum loudness), whereas it tends to perceive less immediately those onsets associated to instruments with a slow attack. The impact of delay on the synchronism of the performance is therefore expected to be more clearly perceivable when using musical instruments with a fast attack, than with instruments with a slower attack time. This means that the choice of musical instrument matters in presence of network delay. In practice, however, musicians tend to adjust their playing technique according to the specific attack time of the played instrument. For example, organ

players are used to naturally compensating the delay elapsed between the pressure of the keyboard keys and the sound emission at the pipes, as well as the time that the sound takes to travel back from the pipes to the musician. To a smaller extent this is also true for piano players. In this case the delay between pressing a key and detecting the corresponding note onset varies between 30 and 100 ms, depending on sound loudness and musical articulation (e.g. *legato, staccato*) [1]. For some categories of instruments, it has been shown that the expressive intention and direction of the musician (i.e., subjective artistic and interpretation choices, which are in turn affected by a particular emotional state during the performance) can have a significant impact on sonological parameters such as attack, sustain and decay time [6]. This is why in this study we do not evaluate the impact of the instrument attack time on the performance interplay, and consider this attack time simply as part of the overall delay perceived by the musician.

One specific aspect that needs to be addressed is the role played by a musical instrument in a performance. In western music, some instruments have a more pronounced "leading" role than others. For example, drums usually have the task of producing regular patterns that rhythmically lead all the other voices. The rhythmic evolution of the melodic line, on the other hand, is more dependent on the personal interpretation of the musician, and therefore its onsets tend to deviate from this reference timing. Drum players are therefore expected to keep a steady tempo even when the other musicians are playing off-tempo. This, of course, needs to be accounted for in the evaluation of the sensitivity to the delay.

As far as the latency perception is concerned, studies [27, 9] introduce three metrics: Ensemble Performance Threshold (LPD), Personal Beat Shift Range (PBSR), and Ensemble Delay Accepted Limit (EDAL). The LPD indicates the maximum delay that allows musicians to play in synchronization without noticeable latency perception. As discussed in Section 0, a consistent body of literature estimates the value of the LPD around 25 ms. Conversely, the PBSR evaluates the personal tolerance of a single musician to deviations from perfect onset synchronization (i.e., to inter-onset delays), when performing with a counterpart. This value highly depends on the musicians' training level, performing style and personal inclination. Finally, the EDAL defines the latency tolerance threshold for a specific musical performance, depending on the PBSR of each player and on several contingent factors such as the musical piece to be performed, the reference tempo, and the instruments being played. In our experimental results, for the evaluation of latency perception we focus on metric similar to the EDAL, i.e. the musicians' subjective rate of the delay perception, considering a wide variety of performances characterized by different combinations of instruments, reference tempo, rythmical complexity and musical role of the played parts.

2 Description of Musical Features

In this Section we discuss which timbral and rhythmic features are relevant for the analysis at hand. The timbral features that we use for analyzing recorded audio tracks are extracted using the MIRToolbox [24].

As for rhythmic features that are characteristic of the rhythmic complexity of the musical piece, we either computed them manually or extracted them directly from the MIDI symbolic data (score) using the MIDI Toolbox [18].

2.1 Rythmic Features

Despite the fact that several metrics have been proposed in the literature for characterizing the rhythmic content of a score [30], there is little consensus on one that can be seen as commonly accepted in the Music Information Retrieval (MIR) community. In this study, in order to capture the rhythmic complexity of the played parts, we consider the following rhytmic features: *Event Density* (*ED*) [24] and *Rythmic Complexity* (*RC*) [26]. Given the score of a part of a musical piece, the *Event Density* estimates the average number of note onsets per second [24]:

$$ED = \frac{NO}{T}, \tag{1}$$

where NO is the number of onsets and T is the duration of the musical piece. It is worth highlighting that an onset is defined as a temporal event, regardless of the number of notes that are simultaneously played. A chord is therefore counted as a single onset and the ED values for polyphonic parts are comparable with the ones computed over monophonic parts. The ED is thus a metric quantifying the degree of the density of notes. In this study the *note* also refers to drum and percussion events.

The *rythmic complexity* (*RC*) provides a numerical evaluation of the complexity, degree of surprise and unpredictability of the part [17]. The rhythmic complexity for a music piece is computed as a weighted mean of several factors:

$$RC = w_1\, H + w_2\, ED + w_3\, \sigma + w_4\, A, \tag{2}$$

where H and σ are the entropy and the standard deviation of the distribution of the duration of the notes in the score, respectively, $w_1, \ldots, w_4$ are the correspondent weights [16] and A is a measure of phenomenal accent synchrony from inferred metrical hierarchy. Musicians use to hierarchically decompose a rhythmic pattern into strong accents (that usually occur in the first or second beat of a bar) and weak accents (in the other beats of a bar). However, composers can place accents in different locations of the metric to provide their composition of a specific expression and to induce specific emotions. This is the case of the syncopation, where strong and weak accents are located on the upbeats. The (mis)alignment between the metrical hierarchy and accent synchrony, which affects the rhythmic complexity of a music piece, is captured by the metric A.

Both rhythmical features depend uniquely on the reference BPM (δ) and on the rhythmic figures that are prescribed in the score, therefore *ED* and *RC* can be extracted directly from the MIDI (symbolic) representation of the musical scores, given the reference BPM value.

2.2 Timbral Features

As far as timbral characterization is concerned, we consider a set of features that describe the shape of the magnitude spectrum. From the Short Time Fourier Transform (STFT) of the sound signal [23] we compute the first four moments of the magnitude spectrum. These moments are, in fact, widely used Spectrum descriptors [23, 33, 32], known as *Spectral Centroid* (*SC*), *Spectral Spread* (*SSp*), *Spectral Skewness* (*SSk*), and *Spectral Kurtosis* (*SK*). We also captured two additional spectral features which describe the noisiness of the sound: *Spectral Flatness* (*SF*), and *Spectral Entropy* (*SE*).

The *Spectral Centroid* (*SC*) corresponds to the "center of gravity" (first moment) of the magnitude spectrum:

$$F_{SC_l} = \frac{\sum_{k=1}^{K} f(k) S_l(k)}{\sum_{k=1}^{K} S_l(k)} , \tag{3}$$

where l is the frame index; $S_l(k)$ is the Magnitude Spectrum computed at the k-th frequency bin; $f(k)$ is the frequency corresponding to the k-th bin; and K is the total number of frequency bins. The spectral centroid gives us an idea of where on the frequency axis the energy is, therefore it somehow captures the *brightness* of the sound.

The *Spectral Spread* (*SSp*) is the second moment of the distribution and it measures the standard deviation of the magnitude spectrum from the Spectral Centroid:

$$F_{SSp_l} = \sqrt{\frac{\sum_{k=1}^{K} (f(k) - F_{SC_l})^2 S_l(k)}{\sum_{k=1}^{K} S_l(k)}} . \tag{4}$$

The *SSp* describes the compactness of the magnitude spectrum around the Spectral Centroid. A spread out distribution of the frequency components is characteristic of noisy sounds. For this reason, the Spectral Spread tends to measure the noisiness of a sound source.

The *Spectral Skewness* (*SSk*) is the third moment of the magnitude spectrum and captures the symmetry of its frequency distribution:

$$F_{SSk_l} = \frac{\sum_{k=1}^{K} (S_l(k) - F_{SC_l})^3}{K F_{SS_l}^3} , \tag{5}$$

where F_{SC_l} is the Spectral Centroid at the l-th frame (see eq.(3)) and F_{SS_l} is the Spectral Spread at the l-th frame (see eq. (4)). A positive value of Spectral Skewness corresponds to an asymmetric concentration of the spectrum energy towards higher frequency bins, which implies the presence of a long tail on lower frequencies. Vice versa, negative *SSk* coefficients represent a skewed distribution towards lower frequencies, with a long tail towards higher frequencies. The perfect symmetry corresponds to the zero *SSk* value.

The *Spectral Kurtosis* (*SK*) is the fourth moment of the distribution and describes the size of the tails of the distribution of the Magnitude Spectrum values:

$$F_{SK_l} = \frac{\sum_{k=1}^{K} (S_l(k) - F_{SP_l})^4}{K F_{SS_l}^4} - 3 , \tag{6}$$

where F_{SC_l} is the Spectral Centroid at the $l-th$ frame (see eq. (3)) and F_{SS_l} is the Spectral Spread at the l-th frame (see eq.(4)). Positive Spectral Skewness values indicate that the distributions have relatively large tails, distributions with small tail have negative kurtosis, and normal distributions have zero kurtosis. This is why the Spectral Kurtosis can be interpreted as a description of the deviation from normality. The offset -3 in eq. (6), in fact, is a correction term that sets the kurtosis of the normal distribution equal to zero [24].

As previously discussed, capturing the *noisiness* of the audio signal is crucial for investigating the impact of timbre on delay tolerance. For this reason, we introduce consider two additional features: *Spectral Entropy* (*SE*) [24], and *Spectral Flatness* (*SF*) [23]. As a white-noise is characterized by a flat Spectrum, spectral features devoted to capturing the noisiness of a sound should provide a sort of comparison with respect to the flat shape.

The *Spectral Entropy* (*SE*) is a measure of the flatness of the magnitude spectrum by applying the Shannon's entropy definition commonly used in information theory context:

$$F_{SE_l} = -\frac{\sum_{k=1}^{K} S_l(k) \log S_l(k)}{\log K} . \tag{7}$$

A totally flat magnitude spectrum corresponds to the maximum uncertainty and the entropy is maximal. On the other hand, the configuration with the spectrum presenting only one very sharp peak and a flat and low background corresponds to the case with minimum uncertainty, as the output will be entirely governed by that peak.

The *Spectral Flatness* (*SF*) estimates the similarity between the magnitude spectrum of the signal frame and the flat shape inside a predefined frequency band. Higher values of Spectral Flatness correspond to noisy sounds and vice versa. Mathematically it is defined as the ratio between the geometric mean and the arithmetic mean of the magnitude spectrum:

$$F_{SF_l} = \frac{\sqrt[K]{\prod_{k=0}^{K-1} S_l(k)}}{\sum_{k=1}^{K} S_l(k)} . \tag{8}$$

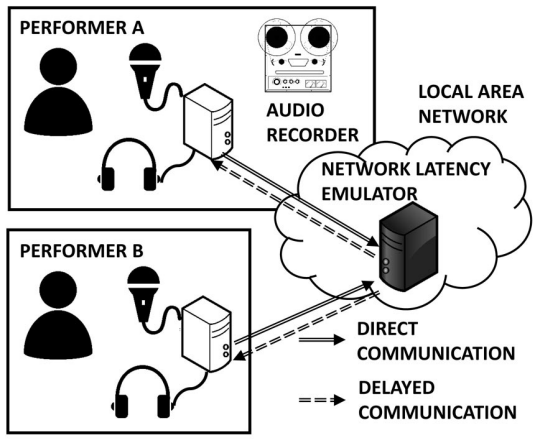


Fig. 1: Testbed setup

3 Experiment Description

3.1 Testbed Setup

As depicted in Figure 1, our experiments involved pairs of musicians playing in two phono-insulated near-anechoic rooms (sound rooms), to avoid any audio feedback or visual contact.[1] The musicians were also forbidden to verbally communicate to their counterpart during the performance. Each room was equipped with a desktop PC running the Soundjack software. Each PC was connected to an external sound card via high-speed connection (FireWire and AES/EBU) operating at a sampling rate of 48 kHz. The sound card was connected to high-quality headphones and microphones. An additional PC (with two network interfaces) running the WANem network emulator [22] was placed in between. The network interfaces of the three PCs were connected to each other through a Fast Ethernet switch. The PCs of the sound rooms were configured to communicate exclusively through the interfaces of the WANem emulator, thus preventing any direct communication between them.

Each musician was able to hear his/her own instrument as well as the instrument of the other player through headphones. The two audio signals were transmitted through the Local Area Network of the building. During the experiments, all the involved LAN segments were free of other traffic. The audio tracks were recorded as follows: the audio data generated by *performer A* were recorded directly after the electric transduction of the microphone, whereas the audio data generated by *performer B* were recorded from the SoundJack feedback after propagation of the audio stream through the network, i.e. as heard through *performer A*'s headsets.

[1] Visual contact was provided not even by means of video streaming because video processing time is larger than audio processing time and would have increased the minimum achievable end-to-end delay.

3.2 Scores and Network Parameters

We considered three pieces of different rhythmic complexity: "Yellow Submarine" (by The Beatles) at different values of BPM (88,110,130), "Bolero" (by Maurice Ravel), and "Master Blaster" (by Stevie Wonder), arranged for four different parts: main melody (M), chord comping (CC), sustained harmony (SH), and drums (D). From the score of every part of each piece we extracted the rhythmic characterization in terms of *ED* and *RC*, which are reported in Table 1. More specifically, the average *ED* has been manually computed, whereas the average *RC* has been computed based on the MIDI representation of the music scores using the MIDI Toolbox, which sets the weights in Formula (2) to $w_1 = 0.7,\ w_2 = 0.2,\ w_3 = 0.5,\ w_4 = 0.5$. Scores were released to the testers in advance[2].

Our experiments involved 8 musicians with at least 8 years of musical experience, all with semi-professional or professional training level, each playing one of the 7 different instruments reported in Table 3. The musicians were grouped in 7 pairs according to the combinations listed in Table 2. Note that some musicians performed in more than one pair (e.g., one clarinetist performed twice, i.e. in pairs 5 and 7, whereas the pianist played electric piano and keyboard in pairs 2,3,4 and 7). For a given pair, each musician performed only one of the four parts for each of the three considered musical pieces, as detailed in Table 2. Musicians in pairs 5 and 6 had regularly performed together in the last years, where the remaining pairs had never played together before. However, in order to avoid biases due to prior common performances, all the pairs were allowed to practice together in the testbed environment until they felt sufficiently confident. Before participating to our experiments, none of the players had ever experienced networked music interactions.

We consider the timbral features as properties of the instrument and we do not track their evolution during the performances. For each instrument, we compose an audio file with a representative selection of recordings of its timbre. For example, the timbral characterization of the drums included the recording of each percussive instrument from the drum set, whereas the characterization of guitar included different kinds of playing techniques, like chords played as arpeggio and plucked strings. The features were then extracted by means of the MIRToolbox and their average values computed over each recording are reported in Table 3.

The recording procedure was repeated several times for each piece. As reported in Table 4, each recording was characterized by different tempo and network settings in terms of reference BPM (δ), network latency and jitter. The two latter parameters were set by assigning each IP packet a random delay T_{net}, statistically characterized by

[2] Scores are publicly available at http://home.deib.polimi.it/buccoli/netmusic/

Table 1: Rythmic characterization of the musical pieces performed during the tests

		Bolero	Master Blaster	Yellow S. (88 BPM)	Yellow S. (110 BPM)	Yellow S. (130 BPM)
M	ED	2.1407	2.1667	1.5253	1.9067	2.2880
	RC	5.5337	5.5627	5.4160	5.7094	6.0567
CC	ED	1.3222	2.6542	1.8333	2.2917	2.7500
	RC	3.4516	6.8903	5.3064	5.6455	5.9592
SH	ED	0.3778	0.5778	0.8213	1.0267	1.2320
	RC	2.9364	5.3444	3.8062	4.0208	4.2237
D	ED	2.0148	4.3514	1.5253	1.9067	2.2880
	RC	6.0285	5.7255	4.5228	4.7548	4.9767

Table 2: Combination of parts played in each experiment session. M: main melody; CC: chord comping; SH: sustained harmony; D: drums

Id	Instrument A	Part A	Instrument B	Part B
1	Acoustic Guitar	M	Classic Guitar	CC
2	Electric Piano	M	Drums	D
3	Keyboard (strings)	SH	Drums	D
4	Keyboard (strings)	SH	Electric Guitar	CC
5	Clarinet	M	Clarinet	M
6	Eletric Guitar	CC	Drums	D
7	Keyboard (strings)	SH	Clarinet	M

Table 3: Timbral characterization for each instrument

Instrument	*SC*	*SSp*	*SSk*	*SK*	*SF*	*SE*
Ac. Guitar	2047	4109	2.76	10.25	0.19	0.76
Clarinet	1686	2272	4.85	31.81	0.07	0.731
Cl. Guitar	3263	4680	1.57	4.43	0.22	0.841
Drums	7903	7289	0.35	1.57	0.61	0.936
El. Guitar	1848	2522	3.70	23.46	0.09	0.818
El. Piano	2101	4251	3.16	12.26	0.16	0.734
Keyboard	1655	3065	4.39	23.77	0.1	0.733

Table 4: Tested network parameters and tempo settings

Piece	δ [BPM]	μ [ms]	σ [ms]
Yellow Submarine	88,110,132	20,30,40,50,60	1
Bolero	68	20,30,40,50,60	1
Master Blaster	132	20,30,40,50,60	1

independent identically distributed Gaussian random variables with mean μ and standard deviation σ. The payload of each packet contained 128 16-bit-long audio samples, corresponding to a duration of 2.67 ms. For the considered values of μ and σ, we set the receiver buffer size to 4 packets (i.e. 512 audio samples) and measured the number of buffer overruns/underruns during each recording. The overall probability of overrun/underrun events turned out to be smaller than 1%. This value is representative of realistic traffic conditions of a telecommunication network. Note that overruns/underruns generate glitches (e.i, distortions in the reproduction of the received audio signal) which affect the overall audio quality perceived by the musicians.

Note also that, as T_{net} accounts only for the emulated network delay, the additional latency T_{proc} introduced by the audio acquisition and the audio rendering processes must be taken into account in the computation of the one-way overall delay time $T_{tot} = T_{net} + T_{proc}$. More specifically, the processing time T_{proc} includes: in-air sound propagation from instrument to microphone; transduction from the acoustic wave to electric signal in the microphone; signal transmission through the microphone's wire; analog to digital conversion of the sender's sound card, internal data buffering of the sound card; processing time of the sender's PC to packetize the audio data prior to transmission; processing time of the receiver's PC to depacketize the received audio data; queuing time of the audio data in the application buffer at the receiver side; digital to analog conversion of the receiver's sound card; transmission of the signal through the headphone's electric wire; transduction from electric signal to acoustic wave in the headphones. We experimentally evaluated T_{proc} by measuring the end-to-end delay T_{tot} when setting $\mu = 0$ and $\sigma = 0$. The measured time[3] was $T_{proc} = 15$ ms.

During each recording session, the order of the proposed network configurations was randomly chosen and was kept undisclosed to the testers, in order to avoid biasing or conditioning. Two measures of metronome beats at the reference BPM were played before the beginning of each performance. At the end of each performance, the testers were asked to express a rating, Q_{perc}, of the quality of their interactive performance[4] within a five-valued range (1="very poor", 5="very good") and of the perceived network delay, D_{perc}, within a four-valued range (1="intolerable", 4="none"). In case the players spontaneously aborted their performance within the first 50 s, D_perc was set to 1 and Q_{perc} was set to 0 by default. The two ratings are considered as subjective quality parameters.

4 Numerical Results

We are interested in assessing the "trend" of the BPM, and particularly, the tendency to slow down or accelerate. We compute a metric of this trend by means of the linear regression of the BPM over the BPM measurements. In order to do so, we considered the first 50 s of each performance and we manually annotated the BPM over time. We then divided the audio track into 5-second windows, with a 50% overlap (2.5 s), with a total $N = 20$ windows. For each time window, we computed the average BPM, resulting in the BPM trend $b(t_n)$, with

[3]The T_{proc} estimated in our experiments is larger than the one reported in [27]. This is mainly due to the use of generic sound card drivers, which increased the processing time of SoundJack.

[4]Note that the metric Q_{perc} is not related to the audio quality experienced by the musicians, but only to the evaluation of the overall satisfaction of their experience and interaction with the counterpart.

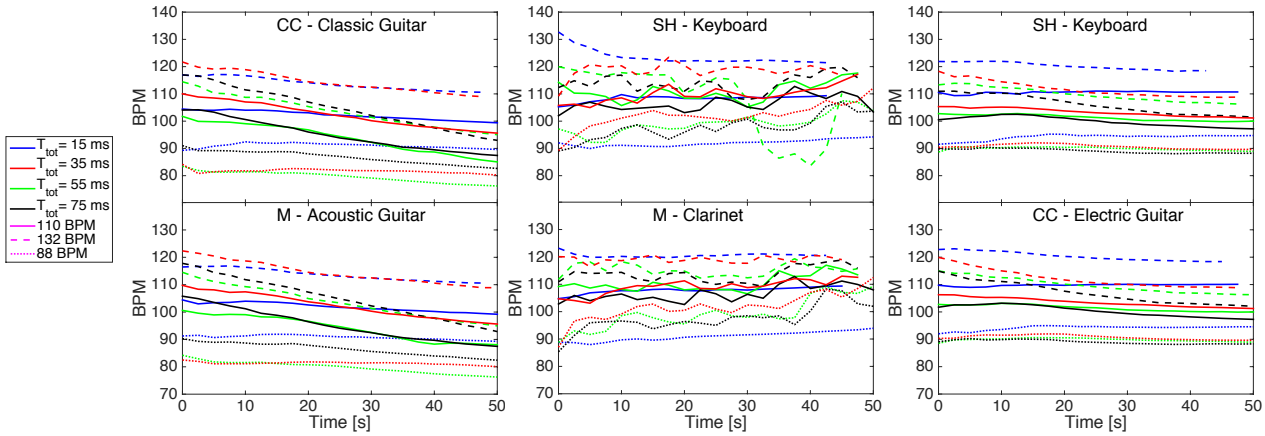


Fig. 2: BPM trend over time when playing "Yellow Submarine", for different combinations of parts and instruments (*performer A* on top, *performer B* on bottom), for various values of end-to-end delays T_{tot} (identified by the color) and reference BPM δ, identified by the type of line (solid, dashed, dotted).

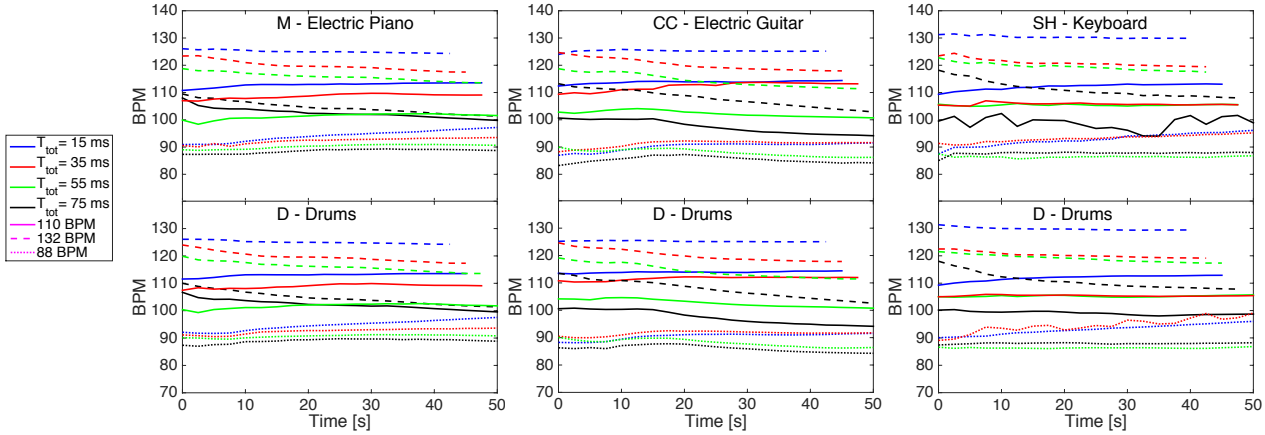


Fig. 3: BPM trend over time when playing "Yellow Submarine" with Drums, combined with different parts and instruments (*performer A* on top, *performer B* on bottom), for various values of end-to-end delays T_{tot} (identified by the color) and reference BPM δ, identified by the type of line (solid, dashed, dotted).

$t_n = n \cdot 2.5$ s and $n = 1,2,...,N$. We finally estimate the intercept β and the slope κ with linear regression:

$$\underset{\kappa,\beta}{\operatorname{argmin}} \frac{1}{N} \sum_{n=1}^{N} \left(b(t_n) - (\kappa t_n + \beta)\right)^2. \qquad (9)$$

In our experiments, the average Mean Square Error was about 1.75%.

In the remainder of the paper we consider the slope κ as an objective metric for the evaluation of the performance quality: $\kappa = 0$ means steady tempo; $\kappa > 0$ means that the musician is accelerating; $\kappa < 0$ means that the musician is slowing down and thus is unable to keep up with the tempo.

4.1 Preliminary Qualitative Results

We begin with some qualitative comments on the trend of the BPM curve $b(t_n)$ extracted from the execution of "Yellow Submarine" for different combinations of instruments and parts, various values of T_{tot} in the range between 15 and 75 ms and three different values of δ (as reported in Table 4). The lower bound of the tested delay values (i.e. $T_{tot} = 15$ ms) is obtained by setting $T_{net} = 0$ ms, meaning that no network delay is added to the unavoidable processing time T_{proc}. For values of T_{tot} above 75 ms (i.e., when $T_{net} = 60$ ms), a considerable amount of executions were aborted by the musicians due to the extreme difficulty of maintaining synchronization. Therefore, we limit our analysis to delay ranges which allowed every pair of musician to perform the piece uninterruptedly for at least one minute. Results reported in Figures 2 and 3 show that in all the considered recordings an initial deceleration occurs in the first few seconds, when the players adjust their tempo until they find a balance which allows them to reach the required degree of synchronization. Such initial deceleration is nearly absent for small network end-to-end delays and reference BPM, but it becomes much more pronounced for large values of T_{tot} and δ. In particular, the scenario with $\delta = 132$ BPM

and $T_{tot} = 75$ ms causes an initial tempo reduction of 12-20 BPM in all the tested combinations of instruments and parts. In addition, as shown in Figure 2, combining typically non-homorhythmic parts such as Melody (M) and Chord Comping (CC) or M and Sustained Harmony (SH) leads either to a tendency to constantly decelerate (see Figure 2, left-hand side), which is more pronounced for large δ , or to a "saw tooth" pattern in which the players periodically try to compensate the tempo reduction (Figure 2, middle). Note that, in the latter case, there is no such pattern in the benchmark scenarios with $T_{tot} = 15$ ms. The difference in the behavior of SH and CC when interacting with M is also due to the type of rhythmic interplay that takes place. Chord Comping, in fact, tends to closely follow and match the tempo of the Melody, while Sustained Harmony is a steady accompaniment ("pad") with more relaxed on-time constrains. As M is expected to meander off-tempo, it is harder for SH and M to stay in sync, and adjustments happen in bursts.

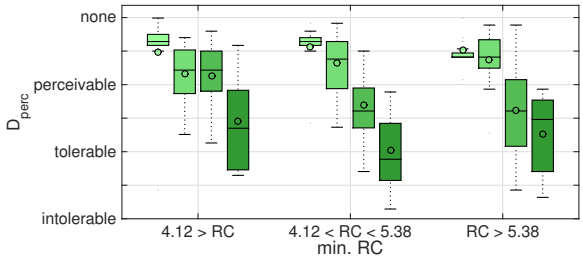
When two homo-rhyrthmic parts (those that are expected to keep a steady tempo, such as CC and SH) are combined, $b(t_n)$ tends to remain almost constant (see Figure 2, on the right-hand side, where a slight negative slope occurs only at $\delta = 132$ BPM). A similar behavior is observed when M, CC or SH combines with Drums (See Figure 3), despite the fact that the two parts are not always homo-rhythmic. This is due to the fact that drums tend to have a very specific rhythmic "leading role" in western music, therefore the other musicians generally tend to follow the drummer.

Based on the above results, we conclude that the choice of the combination of instruments and parts has a significant impact on the capability of the musicians to keep a steady tempo. In the next Subsection, we will give a more in-depth analysis of the impact of single rhythmic and timbral features characterizing the specific combination of parts and instruments on the subjective and objective performance quality metrics.

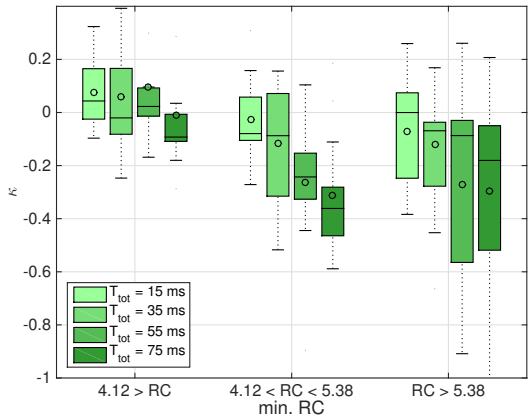
4.2 Dependency of Quality Metrics on Rythmic Features

We now analyze the impact of different end-to-end delays T_{tot} on the subjective quality metric D_{perc} and on the BPM slope κ, for various values of the rythmic and timbral features described in Section 2. The interaction quality rating Q_{perc} resulted to be strongly correlated to D_{perc}, therefore for the sake of brevity we do not report such results.

For every recording, we consider the maximum and minimum values of each feature among the two parts and instruments played by the musicians. For example, in test session 2 (see Table 2) when performing "Bolero", the minimum event density (ED) is 2.01 (ED of the D part) and the maximum ED is 2.14 (ED of the M part). Conversely, the minimum rhythmic complexity (RC) is 5.53 (on the M part) and the maximum RC is 6.03 (on the D part).



(a) Subjective Perception of Delay D_{perc}



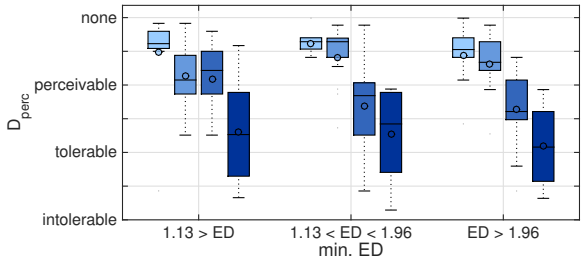
(b) Average BPM Linear Slope κ

Fig. 4: Dependence of κ and D_{perc} on the minimum Rythmic Complexity RC for different values of T_{tot}

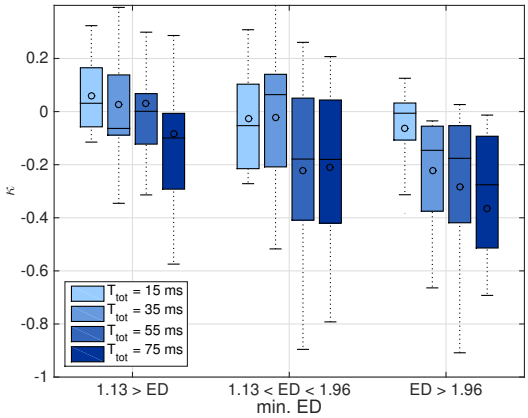
Figure 4a reports the subjective delay perception D_{perc} attributed by the pairs of testers to their performances, for different values of T_{tot}, as a function of the minimum RC between the two parts. For the sake of clarity, only four values of T_{tot} are reported, where $T_{tot} = T_{proc} = 15$ ms is considered as benchmark. Results shows that, for a given value of minimum RC, the average D_{perc} decreases when T_{tot} increases. Moreover, for a given T_{tot}, increasing RC also has a negative impact on the average quality rating. However, the reduction of D_{perc} is more relevant for large values of T_{tot}.

Let us now take a look at how the average BPM linear slope κ is affected by the minimum rhythmic complexity RC. For large values of RC (see Figure 4b), we found slightly negative values of κ (which denote a tendency to slow down) even in the benchmark scenario. As expected, the need of synchronism increases when musicians are playing more complex parts and the lack of typical synchronization cues, such as eye-contact, affect the performance even in absence of network delay. However, negative slopes tend to become much steeper for large values of T_{tot}, which suggests that the tolerance to the delay decreases for more complex musical pieces.

Similar conclusions can be drawn on the dependence of the perceived delay D_{perc} and the objective metrics κ on the minimum ED, as depicted in Figure 5, due to the non-negligible correlation that exists between RC and ED. These conclusions remain substantially unvaried if,

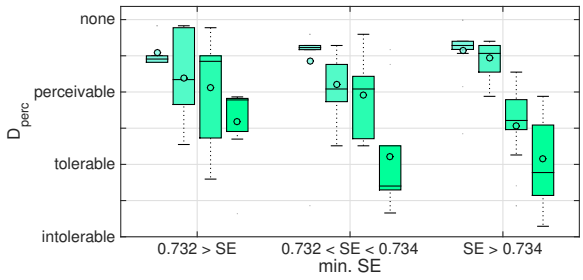


(a) Subjective Perception of Delay D_{perc}

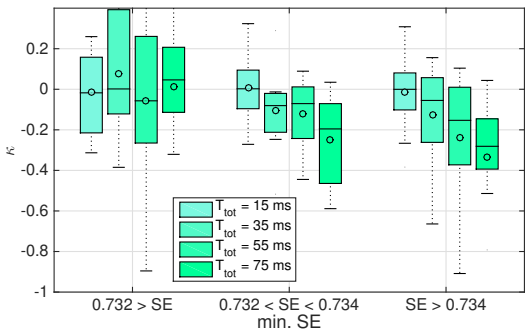


(b) Average BPM Linear Slope κ

Fig. 5: Dependence of κ and D_{perc} on the minimum Event Density ED for different values of T_{tot}

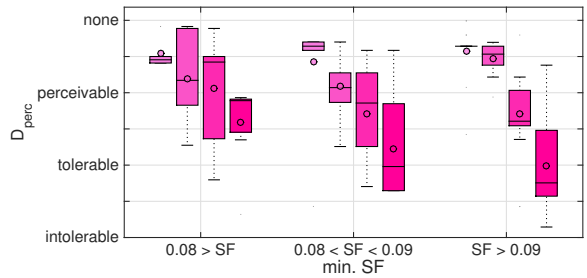


(a) Subjective Perception of Delay D_{perc}

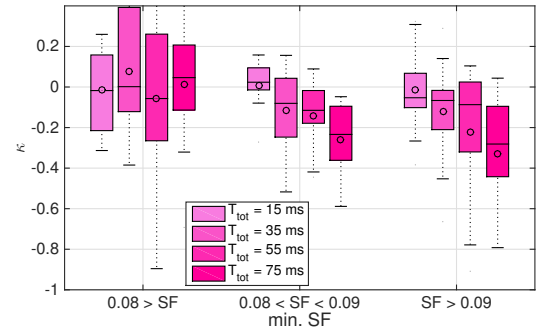


(b) Average BPM Linear Slope κ

Fig. 6: Dependence of κ and D_{perc} on the minimum Spectral Entropy SE for different values of T_{tot}

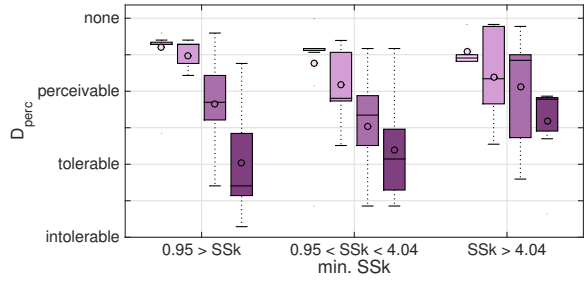


(a) Subjective Perception of Delay D_{perc}

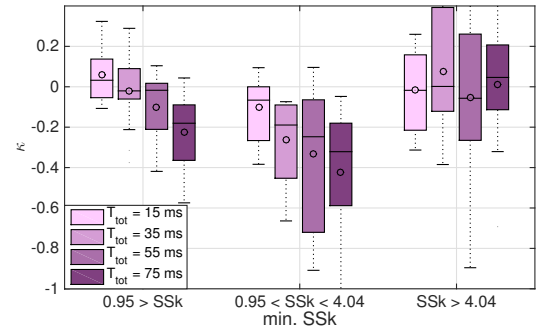


(b) Average BPM Linear Slope κ

Fig. 7: Dependence of κ and D_{perc} on the minimum Spectral Flatness SF for different values of T_{tot}



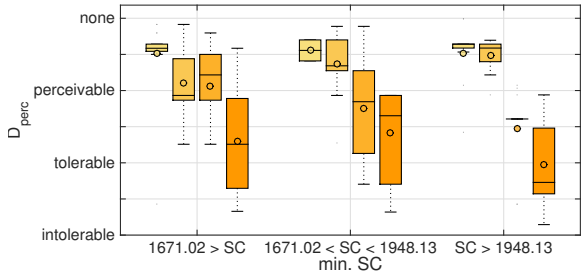
(a) Subjective Perception of Delay D_{perc}



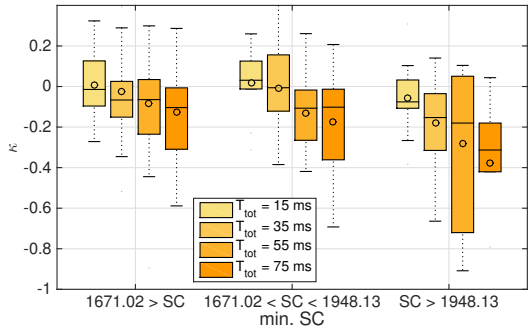
(b) Average BPM Linear Slope κ

Fig. 8: Dependence of κ and D_{perc} on the minimum Spectral Skewness SSk for different values of T_{tot}

instead of considering the minimum values of RC and ED, we consider the maxima.



(a) Subjective Perception of Delay D_{perc}



(b) Average BPM Linear Slope κ

Fig. 9: Dependence of κ and D_{perc} on the minimum Spectral Centroid *SC* for different values of T_{tot}

4.3 Dependency of Quality Metrics on Timbral Features

As far as timbral features are concerned, we observe that the noisiness of the instrument, which is captured by Spectral Entropy, Flatness and Spread, has a relevant impact on the perceived delay D_{perc}. For example, in Figures 6 and 7 we show D_{perc} and κ are affected by Spectral Entropy (*SE*) and Spectral Flatness (*SF*). We consider the minimum Entropy and minimum Flatness between the two involved instruments. Focusing on the objective metric κ (see Figures 6b and 7b), we notice that as the *SE* and the *SF* increase, the tempo slowdown becomes more relevant. This impact is negligible for low network delays, but it grows significantly for fairly large values of T_{tot}. Similar considerations are valid for D_perc, as reported in Figure 6a. Analogous findings also apply to the dependency of the quality metrics on the Spectral Spread (these results are not shown here for reasons of space). Conversely, when considering the impact of Spectral Skewness (*SSk*) and Sperctral Kurtosis (*SK*) on the performance metrics, we notice that, for a given delay T_{tot}, a change in their values does not cause the quality to perceivably worsen (see Figure 8, results on *SK* not reported for conciseness).

Finally, when looking at the influence of the Spectral Centroid *SC* (i.e., of sound brightness) on the subjective quality metrics, results reported in Figure 9a show that the perceptual metric D_{perc} does not exhibit significant fluctuations due to a varying *SC*. However, for large values of *SC*, a slight tendency to decelerate emerges in Figure 9b, which shows the impact of *SC* on the objective quality metric κ.

It is also worth noticing that D_{perc} is not necessarily an indicator of quality degradation of the performance, but only on the musicians' subjective perception of the end-to-end delay. However, results reported in Figures 3-8b show that such perception is strongly affected by the timbral and rhythmic characteristics of the combination of instruments and parts. For example, in Figure 8a, the perceived network delay D_{perc} is larger for large values of *SSk* and T_{tot} than the value we would have in the case of low delays. This leads us to think that the musicians' capability of estimating the network delay is biased by the perceived interaction quality of the performance. This means that large network delays (i.e., $T_{tot} \geq 75ms$) do not prevent networked musical interaction, but they limit the selection of the instrument/part combinations. Thus, the resulting experience can be satisfactory if the performer is willing to trade flexibility and perceived interaction quality with the convenience of playing over the network.

5 Conclusion

This article proposes an extensive evaluation of the quality of Networked Music Performances (NMPs) as a function of numerous parameters, some concerning telecommunication network delays and conditions, others involving rhythmic and timbral descriptors of the musical instruments involved. The analysis goes as far as considering the influence of the role of the instrument on such quality metrics. In order to conduct this analysis, we implemented a testbed for psycho-acoustic tests, which emulates the behavior of a real telecommunication network in terms of variable transmission delay and jitter, and we quantitatively evaluated the impact of the various performance parameters on the trend of the tempo that the musicians were able to keep during the performance, as well as on the perceived quality of the musical interaction. We found that the possibility of enjoying an interactive networked musical performance is not only a function of the total network delay, but it also depends on the role and the timbral characteristics of the involved musical instruments, as well as the rhythmic complexity of the performance.

In particular, the paper provides evidence to the following main findings. When playing more rhythmically complex pieces, musicians exhibit a more pronounced tendency to decelerate when the network latency is higher. Nonetheless, the rhythmical complexity does not significantly worsen their perception of the delay and of the interaction quality. Among the timbral features, instruments with a higher Spectral Entropy and Spectral Flatness (such as guitars and drums) lead to larger tempo slowdown in case of higher network delays. In addition, they also amplify the negative impact of network delay on the perceived delay and interaction quality.

6 Acknowledgments

The authors are grateful to Prof. Alexander Carôt for his generous assistance with the SoundJack software, and to all the people that were involved in the psycho-perceptual tests, particularly Antonio Canclini, Bruno Di Giorgi, Lorenzo Dainelli, Niccoló Dainelli, Davide Pradolini and Simone Pradolini.

7 REFERENCES

[1] Anders Askenfelt and Erik V. Jansson. From touch to string vibrations. i: Timing in the grand piano action. *The Journal of the Acoustical Society of America*, 88(1), 1990.

[2] Álvaro Barbosa. Displaced soundscapes: A survey of network systems for music and sonic art creation. *Leonardo Music Journal*, 13:53–59, 2003.

[3] Álvaro Barbosa and João Cordeiro. The influence of perceptual attack times in networked music performance. In *Proceedings of the Audio Engineering Society Conference: 44th International Conference: Audio Networking*, San Diego, 2011. Audio Engineering Society.

[4] Nicolas Bouillot. njam user experiments: Enabling remote musical interaction from milliseconds to seconds. In *Proceedings of the International Conference on New Interface for Musical Expression (NIME'07)*, pages 142 – 147, New York, January 2007.

[5] Nicolas Bouillot and Jeremy R Cooperstock. Challenges and performance of high-fidelity audio streaming for interactive performances. In *Proceedings of the 9th international conference on New Interfaces for Musical Expression (NIME'09)*, Pittsburgh, 2009.

[6] Sergio Canazza, Giovanni De Poli, Stefano Rinaldin, and Alvise Vidolin. Sonological analysis of clarinet expressivity. In Marc Leman, editor, *Music, Gestalt, and Computing*, volume 1317 of *Lecture Notes in Computer Science*, pages 431–440. Springer Berlin Heidelberg, 1997.

[7] Alexander Carôt and Christian Werner. Distributed network music workshop with soundjack. *Proceedings of the 25th Tonmeistertagung, Leipzig, Germany*, 2008.

[8] Alexander Carôt and Christian Werner. Fundamentals and principles of musical telepresence. *Journal of Science and Technology of the Arts*, 1(1):26–37, 2009.

[9] Alexander Carôt, Christian Werner, and Timo Fischinger. *Towards a comprehensive cognitive analysis of delay-influenced rhythmical interaction*. Ann Arbor, MI: MPublishing, University of Michigan Library, 2009.

[10] Michael Casey. Mpeg-7 sound recognition tools. In *IEEE Transactions on Circuits and Systems for Video Technology*, volume 11, pages 737–747, 2001.

[11] Chris Chafe. Living with net lag. In *Audio Engineering Society Conference: 43rd International Conference: Audio for Wirelessly Networked Personal Devices*. Audio Engineering Society, 2011.

[12] Chris Chafe, Juan-Pablo Cáceres, and Michael Gurevich. Effect of temporal separation on synchronization in rhythmic performance. *Perception*, 39(7):982–992, 2010.

[13] Chris Chafe and Michael Gurevich. Network time delay and ensemble accuracy: Effects of latency, asymmetry. In *Audio Engineering Society Convention 117*. Audio Engineering Society, 2004.

[14] Chris Chafe, Michael Gurevich, Grace Leslie, and Sean Tyan. Effect of time delay on ensemble accuracy. In *Proceedings of the International Symposium on Musical Acoustics*, volume 31, 2004.

[15] Elaine Chew, Alexander Sawchuk, Roger Zimmerman, Vely Stoyanova, Ilia Tosheff, Christos Kyriakakis, Christos Papadopoulos, Alexandre Francois, and Anja Volk. Distributed immersive performance. In *Proceedings of the Annual National Association of of the Schools of Music (NASM)*, San Diego, CA, 2004.

[16] Tuomas Eerola. *The Dynamics of Musical Expectancy: Cross-Cultural and Statistical Approaches to Melodic Expectations*. PhD thesis, 2003.

[17] Tuomas Eerola and Adrian C. North. Expectancy-based model of melodic complexity. In *Proceedings of the Sixth International Conference on Music Perception and Cognition*, Keele, Staffordshire, UK, 2000.

[18] Tuomas Eerola and Petri Toiviainen. *MIDI Toolbox: MATLAB Tools for Music Research*. University of Jyväskylä, Jyväskylä, Finland, 2004.

[19] Xiaoyuan Gu, Matthias Dick, Zefir Kurtisi, Ulf Noyer, and Lars Wolf. Network-centric music performance: Practice and experiments. *Communications Magazine, IEEE*, 43(6):86–93, 2005.

[20] Michael Gurevich, Chris Chafe, Grace Leslie, and Sean Tyan. Simulation of networked ensemble performance with varying time delays: Characterization of ensemble accuracy. In *Proceedings of the 2004 International Computer Music Conference, Miami, USA*, 2004.

[21] Julia C. Hailstone, Rohani Omar, Susie M. D. Henley, Chris Frost, Michael G. Kenward, and Jason D. Warren. It's not what you play, it's how you play it: Timbre affects perception of emotion in music. *The Quarterly Journal of Experimental Psychology*, 62(11):2141–2155, 2009. PMID: 19391047.

[22] Hemanta Kumar Kalitay and Manoj K. Nambiarz. Designing wanem: A wide area network emulator tool. In *Proceeedings of the Third International Conference on Communication Systems and Networks (COMSNETS)*, pages 1–4, Jan 2011.

[23] Hyoung-Gook Kim, Nicolas Moreau, and Thomas Sikora. *MPEG-7 Audio and Beyond: Audio Content Indexing and Retrieval*. Wiley, 2005.

[24] Olivier Lartillot and Petri Toiviainen. A matlab toolbox for musical feature extraction from audio. In *Proceedings of the 10th International Conference on Digital Audio Effects (DAFx-07)*, 2007.

[25] Ewa Łukasik. Long term cepstral coefficients for violin identification. In *Proceedings of the Audio Engineering Society Convention 128 (AES128)*, 2010.

[26] Dirk-Jan Povel and Peter Essens. Perception of temporal patterns. *Music perception*, pages 411–440, 1985.

[27] Alain Renaud, Alexander Carôt, and Pedro Rebelo. Networked music performance: State of the art. In *Proceedings of the Audio Engineering Society Conference: 30th International Conference: Intelligent Audio Environments*, Finland, March 1517 2007. Audio Engineering Society.

[28] Robin Renwick. *SOURCENODE: A Network Sourched Approach to Network Music Performacne (NMP)*. Ann Arbor, MI: MPublishing, University of Michigan Library, 2012.

[29] Randy Erfa Saputra and Ary Setijadi Prihatmanto. Design and implementation of beatme as a networked music performance (nmp) system. In *Proceedings of the International Conference on System Engineering and Technology (ICSET)*, pages 1–6. IEEE, 2012.

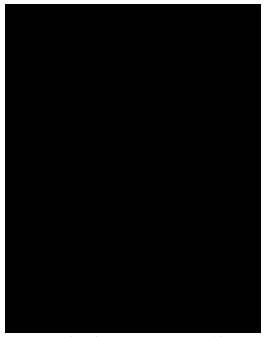
[30] Ilya Shmulevich and Dirk-Jan Povel. Complexity measures of musical rhythms. *Rhythm perception and production*, pages 239–244, 2000.

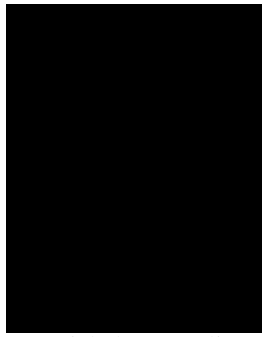
[31] Charilaos Stais, Yannis Thomas, George Xylomenos, and Christos Tsilopoulos. Networked music performance over information-centric networks. In *Proceedings of the IEEE International Conference on Communications Workshops (ICC)*, pages 647–651. IEEE, 2013.

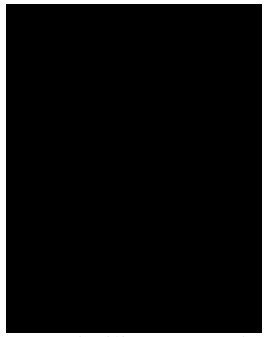
[32] Massimiliano Zanoni, Daniele Ciminieri, Augusto Sarti, and Stefano Tubaro. Searching for dominant high-level features for music information retrieval. In *Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*, pages 2025–2029. IEEE, 2012.

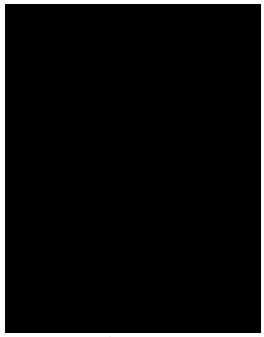
[33] Massimiliano Zanoni, Francesco Setragno, and Augusto Sarti. The violin ontology. In *Proceedings of the 9th Conference on Interdisciplinary Musicology (CIM14)*, Berlin, Germany,, 2014.

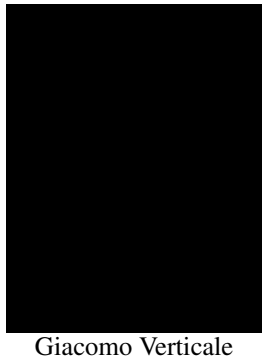
THE AUTHORS

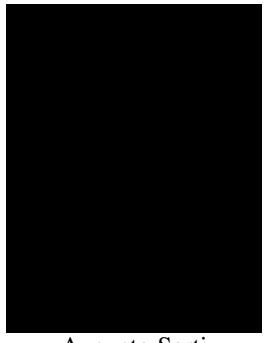

Cristina Rottondi


Michele Buccoli


Massimiliano Zanoni


Dario Garao


Giacomo Verticale


Augusto Sarti

Cristina Rottondi received both Master and Ph.D. Degrees cum laude in Telecommunications Engineering from Politecnico di Milano in 2010 and 2014 respectively. She is currently postdoctoral researcher at the Department of Electronics, Information, and Bioengineering (DEIB) of Politecnico di Milano. Her research interests include cryptography, communication security, design and planning of optical networks, and networked music performance.

●

Michele Buccoli received the B.Sc. degree in Computer Engineering from Universitá di Pisa in 2010 and the M.Sc. degree in Computer Engineering from Politecnico di Milano in 2013, defending a thesis on a music search engine based on textual queries in natural language. Since November 2013 he is a PhD Student in the Image and Sound Processing Group (ISPG) of Politecnico di Milano.

His research field is Music Information Retrieval, with a focus on the modelling of high-level (semantic) music descriptors and on the development of novel automatic paradigms for research, browsing and music description. He also works on the application of Music Information Retrieval techniques to studies on networked music performance and bootleg detection.

●

Massimiliano Zanoni is postdoctoral researcher in the Image and Sound Processing Group (ISPG) at the Department of Electronics, Information and Bioengineering (DEIB) of Politecnico di Milano. He received a Master degree in Computer Science from Alma Mater Studiorum University of Bologna and a Ph.D. degree in 2013 from Politecnico di Milano. His main research interests include Music Information Retrieval, Music Emotion Recognition, ontology-based information management for modeling musical instruments knowledge and feature-based analysis of musical instruments.

●

Dario G. Garao received his Master Degree in Telecommunication Engineering from Politecnico di Milano in 2011, defending a thesis on multistage integrated optical networks. Since 2012 he is a Ph.D. student in Information Technology at Department of Electronics, Information, and Bioengineering (DEIB) of Politecnico di Milano. His main research interests include optical networks and fault tolerant optical switching systems.

●

Giacomo Verticale is Researcher at Politecnico di Milano, Italy. He is co-head of the Broadband Optical Networks, Security and Advanced Internet (BONSAI) Laboratory in the Department of Elec- tronics, Information, and Bioengineering (DEIB). Before joining Politecnico di Milano, he was with the CEFRIEL research center. He graduated in 2003 at Politecnico di Milano defending a thesis on the performance of packet transmission in 3G mobile networks. His research interests are in network security and in performance evaluation of network protocols.

●

Augusto Sarti received the M.S. and the Ph.D. degrees in electronic engineering, both from the University of Padua, Italy, in 1988 and 1993, respectively, with research on nonlinear system modeling and inversion. His graduate studies included a joint graduate program with the University of California, Berkeley. In 1993, he joined the Politecnico di Milano, Milan, Italy, where he is currently an Associate Professor. In 2013, he also joined the University of California, Davis, as an Adjunct Professor. His research interests are in the area of multimedia signal processing, with particular focus on sound analysis, synthesis and processing; space-time audio processing; geometrical acoustics; and music information extraction.

He has also worked on problems of multidimensional signal processing, image analysis and 3D vision. He coauthored well over 200 scientific publications on international journals and congresses as well as numerous patents in the multimedia signal processing area. He coordinates the activities of the Musical Acoustics Lab and of the Sound and Music Computing Lab of the Politecnico di Milano. He has been the promoter/coordinator and/or contributor to numerous (20+) European projects. He is an active member of the IEEE Technical Committee on Audio and Acoustics Signal Processing, and is in the Editorial Boards of IEEE Signal Processing Letters and of IEEE Tr. on Audio, Speech and Language Processing.