Quality of Musicians' Experience in Network Music Performance: A Subjective Evaluation

Konstantinos Tsioutas*, George Xylomenos*, Ioannis Doumanis† and Christos Angelou*
*Athens University of Economics and Business, Department of Informatics, Greece
†University of Central Lancashire, School of Physical Science and Computing, United Kingdom

***Abstract*—In Network Music Performance (NMP), audio quality and audio delay are considered to be the most critical variables affecting the Quality of Musicians' Experience (QoME). In order to quantify the extent to which these parameters affect QoME, we executed a pilot study where eight musicians performed music in pairs in a controlled NMP setting and were asked to evaluate eight variables related to perception, while the end-to-end delay and the quality of the exchanged audio were varied. We present the design and execution of this experiment and discuss its results and their implications for the tolerance of musicians to increased delay and degraded audio quality.**

I. Introduction

The *Quality of Musicians' Experience* (QoME) in *Network Music Performance* (NMP) is a complex function which depends on many factors, including technical, environmental and psycho-acoustic variables. As a first step in evaluating the *QoME* of NMP sessions based on our previously proposed framework [1], we performed a pilot study in order to understand how audio quality and audio delay affect how musicians perceive the quality of an NMP session. While experimental NMP setups over research networks can support the highest possible audio quality at the minimum possible delay, for musicians relying on publicly available Internet connections, bandwidth limitations require either reducing audio quality or introducing audio compression which increases delay. It is, therefore, important to quantify the tolerance of musicians to audio delay and quality, in order to achieve a tradeoff between them.

Although objective measures, such as network or coding delay, clearly do affect QoME, is is not clear how the parameters of an actual NMP scenario lead to a specific QoME outcome, that is, how musicians perceive the effects of a combination of many underlying parameters to their performance. For this reason, in our study we varied two parameters only (delay and quality), using questionnaires to assess QoME in a subjective manner. As discussed in [1] there are many other factors affecting QoME in NMP; our eventual goal is to perform experiments covering more of them, using the same methodology.

The outline of the rest of the paper is as follows. In Section II, we briefly present related work on assessing musicians' experience. In Section III we present some variables indicative of QoME as well as the questionnaire used to assess them. Section IV describes the setup of our experimental scenarios, while in Section V we present and analyze the results. We summarize our conclusions and discuss future work in Section VI.

II. Related work

A large amount of research touches upon QoME evaluation, looking at it from different perspectives. Some research focuses on the experience of the audience and what feelings are generated to individuals when listening to music. For example, [2] is a study on the connection of emotion in music performance with emotional intelligence: 24 students were asked to complete listening tests, trying to identify the intended emotions in performances of classical piano music.

Nevertheless, in NMP the subject of interest is the musician; there may not even be an audience. Past research has considered how network latency affects the musicians behaviour. Kubacki [3] noted that as latency increases, a musician tends to slow down her tempo. Chafe [4] reached the same conclusion, experimenting with musicians who clapped their hands, indicating that when the latency was below 11 ms, the tempo was accelerating. Bartlette [5] experimented with two pairs of musicians performing two Mozart duets while isolated visually and connected through microphones and headphones. The latency introduced varied from 0 to 200 ms and the musicians rated the performances as non musical and non interactive for delays greater than 100 ms. Driessen [6] also experimented with two musicians in a clapping session and confirmed that the tempo slows down as delay grows.

Olmos [7] experimented with two opera singers and a conductor over a network, evaluating two bio-metric measures, the *Galvanic Skin Response* (GSR) and the number of *Skin Conductance Responses* (SCR), using software for behavior recording along with questionnaires. These tests can reveal far more about QoME in NMP than simple QoS measurements.

The diversity of works found in the literature indicates the different aspects of QoME that can be studied. Our goal is to take a holistic view of the field, assessing multiple objective and subjective factors and their influence on QoME, through a comprehensive measurement campaign with actual NMP sessions. This paper is a first step towards this goal, isolating two objective parameters and assessing their influence on QoME, as part of a small pilot study.

III. Variables based on perception

In this study we focus on the human perception of audio phenomena like audio delay and audio quality. Instead of gathering a single *Mean Opinion Score*, we constructed an eight

question survey covering different aspects of perception. We then defined corresponding variables based on the *Perception of* statement and evaluated these variables by performing our survey on pairs of musicians participating in NMP sessions.

A. The questionnaire

The questionnaire was formed in such a way that it could be easily answered after each NMP session. Each musician just had to choose a score in a Likert scale for each of the eight questions, by touching the suitable button on a smartphone. The questions were:

1) Evaluate the audio quality during the last music performance.
2) Evaluate the degree of synchronization during the last music performance.
3) Evaluate the degree of sound delay you perceived during your last music performance.
4) Evaluate the degree of your music and emotional expression during your last musical performance.
5) Evaluate the degree of audio clicks you experienced during the last music performance.
6) Evaluate your satisfaction during the last music performance.
7) My partner played very well (disagree, agree).
8) Generally, I was trying to follow my partner in rhythm (disagree, agree).

We can group these questions as follows: Questions 1, 2, 3 and 5 are strongly correlated to the QoS of the system, since the audio quality and audio delay are configured by us. Questions 4 and 6 cover musical and emotional expression and satisfaction. Finally, through questions 7 and 8 we try to assess the extent of dependence of a musician's experience on the performance of their counterpart. We further explain the perceptual variables corresponding to these questions in the following paragraphs.

B. Perception of Audio Quality

An interesting question raised in our research was "How do musicians perceive audio quality?" When a musician plays an acoustic instrument, she can hear its entire sound spectrum; we consider this to be "perfect" quality. When a musician performs through an analog audio system, where audio is just amplified by analog machines, she experiences the amplified sound of her instrument. In that case, there are frequencies that may be louder, therefore some equalization may be necessary for the musician to hear a natural sound. In the case of studio recording, where analog to digital and digital to analog conversions are taking place with the ultra low latency of an audio interface, the musician experiences perfect sound, since studio recordings use a high sampling frequency (e.g., 88.2 kHz). In the case of NMP, things are quite different. In conferencing applications, the sampling frequency can be as low as 8 HKz, with some (possibly lossy) audio coding applied to the audio signal. Therefore, we need to consider what is the minimal audio bandwidth that a musician perceives as acceptable and how her experience is affected when audio quality is poor. Thus, we propose the *Perception of Audio Quality* (PoAQ) variable. Musicians were asked to evaluate in a 1 to 5 Likert Scale the perceived audio quality of their performance, without being given detailed instructions about the levels, beyond "higher is better".

C. Perception of Synchronization Degree

Achieving synchronization is a critical issue in NMP which is strongly dependent on the audio delay. Many studies have been conducted which conclude that when one way delay is below 25 ms [3], [6], [5], musicians can synchronise. When delay grows beyond 25 ms, musicians start to slow down their tempo. Slowing down does not mean that one cannot play, since each musician tries to synchronize with the others. We propose the *Perception of Synchronization Degree* (PoSD) as a variable to be evaluated in a Likert Scale from 1 (cannot synchronize at all) to 5 (can synchronize perfectly).

D. Perception of Audio Delay

Next to the audio quality, the most critical variable in NMP is audio delay. Most of the studies in NMP evaluate *Mouth to Ear* delay, that is, the delay from the mouth (or, in our case, from the musical instrument) of one musician to the ear of another. As we proposed in [8] *My Mouth to My Ear* (MM2ME) delay, which the round-trip analogue to M2E delay, is a more appropriate metric for NMP, as when musicians play together, each musician plays one note and unconsciously expects to listen to the other musician's note to play her next note, and so on. In addition, measuring MM2ME delay accurately is much easier than measuring the M2E delay, as it can be done at one endpoint, by simply reflecting the transmitted sound at the other endpoint; M2E needs to be measured at both endpoints, thus requiring synchronized clocks. A critical variable is then the *Perception of Audio Delay* (PAD). We examine how the participants perceive audio delay in a Likert Scale from 1 (no delay) to 5 (too much delay).

E. Perception of Musical and Emotional Expression

A musician's experience in NMP is strongly correlated to her musical and emotional expression during the performance. Thus we propose *Perception of Musical and Emotional Expression* (PoMEE), a variable that we evaluate through using a 1 to 5 Likert Scale (higher is better). Through this variable we examine how a musician's musical and emotional expression could be affected by audio delay or audio quality variations.

F. Perception of Clicks

The *Perception of Clicks* (PoC) variable reflects audible errors in the audio signals perceived as clicks by the musicians. It is assessed in a 5-point Likert scale from 1 (no clicks) to 5 (too many clicks), and is meant to identify sessions where audio quality was affected by lost packets. Packet losses in audio cause signal interruptions which are perceived as clicks. When such artefacts are present, audio quality suffers for reasons unrelated to delay or sampling.

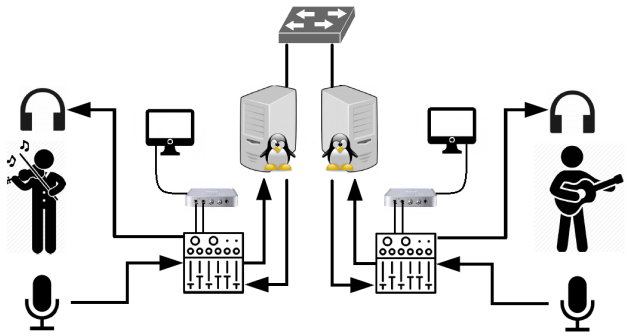

Fig. 1. Technical setup of the experiments

G. Perception of Satisfaction

The *Perception of Satisfaction* (PoSat) is similar to the *Mean Opinion Score* (MOS), which is widely used to evaluate Quality of Experience in similar studies. It is measured in a 5 point Likert scale (higher is better). As satisfaction is a very complex phenomenon, in this study we complement this metric with many other subjective variables, to better understand what leads to a satisfying NMP session.

H. Perception of My Partner's Performance

With this variable we search for correlations between a musician's performance and the performance of her partner. As proposed in [1], the QoME of each musician is strongly correlated to the QoME of the other. Thus, *Perception of my Partner's Performance* (PoMPP) evaluates in a scale from -3 (very bad performance) to +3 (very good performance) how each musician assesses her partner's performance.

I. Tempo Dependencies

Through the final question we try to assess the musical behaviour of each musician regarding the tempo. Specifically, we examine whether a musician depends on her partner's tempo and if she tries to follow him or not, during an NMP session. This variable is also assessed in a scale from -3 (fully disagree) to +3 (fully agree), reflecting whether the musician tried to follow her partner or not.

IV. Experimental Setup

We used our prototype streaming software Aretousa [8] to stream audio and configure the necessary parameters for our experiments. The network topology was a peer to peer architecture with two computers interconnected with a fast Ethernet switch. As shown in Figure 1, an eight channel mixing console with an auxiliary output, a condenser microphone and closed type headphones were used by each of the two musicians participating. At each endpoint, the computer used for audio capture and playout was complemented by a separate computer for recording, which used an external audio interface connected to the console, to avoid delaying the time-critical audio capture and playout operations

We designed two experimental scenarios, scenario A and scenario B. Eight musicians participated in pairs, with each pair playing different musical instruments. They could listen to each other through Aretousa and they also had visual contact via a WebRTC application, although the video connection had a high latency due to coding. The baseline MM2ME (that is, from microphone to headphones, and back) delay of our setup was measured to be 34 ms. In both scenarios, each pair of musicians played a one minute musical part of their choice, repeating it ten (10) times for each scenario. After the end of each repetition each musician was asked to answer an electronic questionnaire using a smartphone. In scenario A, audio delay was manipulated using `tc-netem`[1] while audio quality was kept at 88.2 kHz. In scenario B, audio quality was varied by modifying the sampling rate using Aretousa, while MM2ME delay was kept at 34 ms (the lowest possible).

As shown in Table I for Scenario A, the delay values were used in a random order and not in an increasing one. The MM2ME (two-way) delay varied from 34 ms (the minimum poossible in our setup) to 114 ms; M2E (one-way) delay can be estimated as half of those values. In scenario B, as shown in Table II, we also used the various sampling rates in a random order. The lowest value of 8 kHz results in very poor audio quality, similar to voice telephony, while the highest value of 88.2 kHz results in perfect (studio quality) audio. The instruments played by each pair of musicians are shown in Table III, while Table IV shows the sex, age and experience (in years) of each participating musician.

The questionnaire was the same for each repetition and each scenario. Therefore, musicians played the same piece 20 times (10 for each scenario) and answered the same questions at the end of each session. Musicians were not informed about which variable was manipulated each time, or what the purpose of the experiment was. The goal was to allow us to evaluate multiple variables without bias in the answers.

V. Results and Discussion

In this section we discuss the responses to our questionnaire, first for scenario A (variable delay) and then for scenario B (variable quality).

A. Scenario A

Below we discuss the results shown in the eight plots from Figure 2 to Figure 9. In scenario A, we manipulated delay using `tc-netem` as mentioned in Section IV. The MM2ME delay was configured as shown in Table I in random order, while the sampling frequency corresponding to audio quality was kept at 88.2 kHz. All the graphs were created with `matlab`[2] and show the average of the (10) responses to each question.

1) Figure 2 presents the responses to the question regarding the *Perception of audio quality*. We would expect the answers to vary between 4 and 5, corresponding to very good audio quality, since the sampling rate was constant and high. Unexpectedly, at some low delays we got low quality scores (2–3). A possible explanation

[1] https://www.linux.org/docs/man8/tc-netem.html
[2] https://www.mathworks.com/products/matlab.html

Repetition	1	2	3	4	5	6	7	8	9	10
MM2ME delay (ms)	34	44	54	39	59	64	36	52	69	114

TABLE I
SCENARIO A: MM2ME DELAYS.

Repetition	1	2	3	4	5	6	7	8	9	10
Sampling rate (kHz)	88.2	44.1	22	68	16	8	10	32	25	38

TABLE II
SCENARIO B: SAMPLING FREQUENCIES.

Duet 1	Duet 2	Duet 3	Duet 4
Bouzouki	Electric Bass	Oud	Accordion
Folk Guitar	Cahon	Folk Guitar	Mandolin

TABLE III
INSTRUMENTS PLAYED BY THE MUSICIANS.

Musician	1	2	3	4	5	6	7	8
Sex	M	M	M	M	M	M	M	F
Age	25	32	28	29	31	33	45	28
Experience	¡12	¡12	¡12	¡12	¡12	¡12	¿12	¡6

TABLE IV
AGE, SEX AND EXPERIENCE OF EACH MUSICIAN.

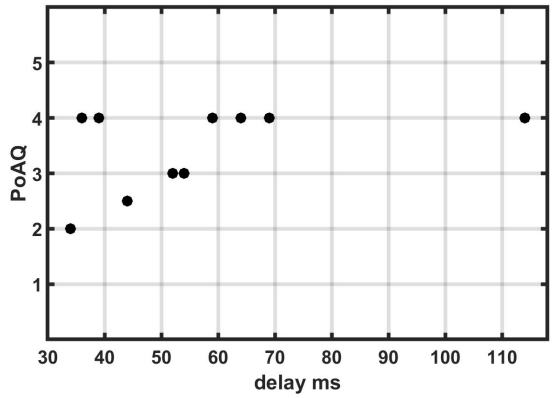

Fig. 2. Perceived Audio Quality vs. MM2ME delay.

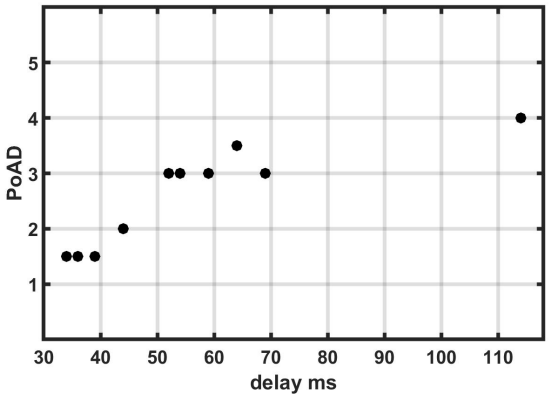

Fig. 4. Perceived Audio Delay vs. MM2ME delay.

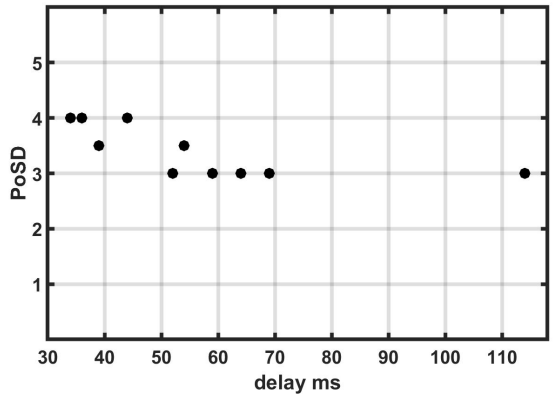

Fig. 3. Perceived Synch Degree vs. MM2ME delay.

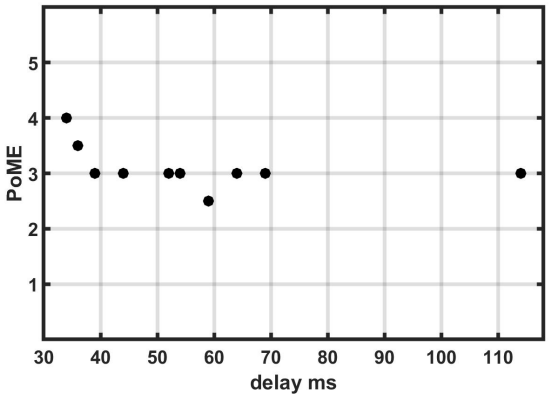

Fig. 5. Perceived Expression vs. MM2ME delay.

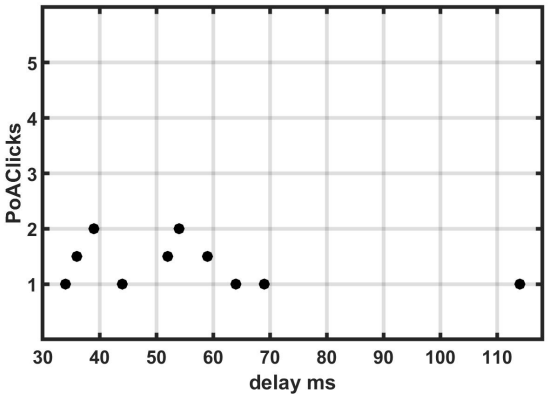

Fig. 6. Perceived Audio Clicks vs. MM2ME delay.

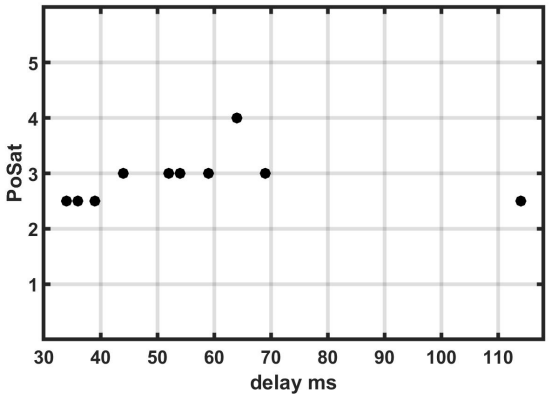

Fig. 7. Perceived Satisfaction vs. MM2ME delay.

for this phenomenon is that audible artefacts (see below) affected the audio quality.

2) Figure 3 presents the responses to the question regarding the *Perception of synchronization degree*. As shown, answers vary between 3 (fair synchronization) and 4 (good synchronization). As the delay increases, scores decrease, but do not fall below 3 even for the highest value of delay. This is an indication that the participants make an effort to synchronize even when the delay is large. Furthermore, in the range from 60 to 114 ms delay changes do not seem to affect the perception of synchronization.
3) Figure 4 shows the relation between the *Perception of audio delay* and the actual delay. Answers vary between 1.5 and 4. Interestingly, delay changes from 40 to 70 ms are more perceivable than in the range from 70 to 114 ms.
4) Figure 5 shows the responses to the *Perception of musical expression* question. There is a sharp decline in the delay range from 34 to 45 ms, with scores dropping from 4 to 3, something that we would expect, but further delay increases do not reduce the score.
5) Figure 6 shows the results for the *Perceived audio clicks* question. Clicks were mostly due to lost samples at the two audio interfaces; they were caused by QoS issues in the Aretousa software, rather than the QoME. The answers were between 1 which corresponds to zero audio clicks and 2 corresponding to few audio clicks. The audio clicks present at the lower delay values may explain the low scores in the perception of quality results, shown in Figure 2.
6) One of the most critical variables in our experiment is the *Perceived Satisfaction*, shown in Figure 7. This variable is similar to *Mean Opinions Score*. The answers range from 2.5 to 3, thus being almost constant. We would expect these values to decrease as delay increases, as in the perceived expression results, shown in Figure 4. An explanation for this behavior may be the clicks heard at lower delay values, shown in Figure 6.
7) The results from the *Perception of my partners performance* question are shown in Figure 8, where the

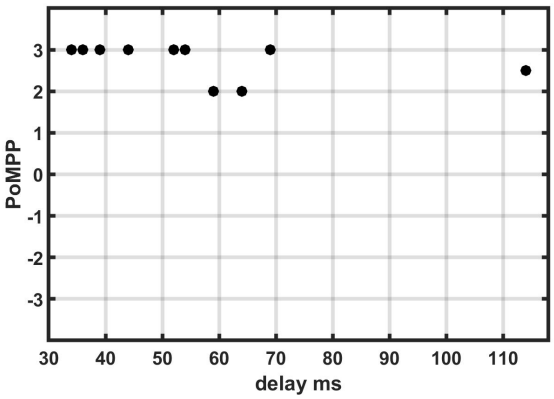

Fig. 8. My partner played well vs. MM2ME delay.

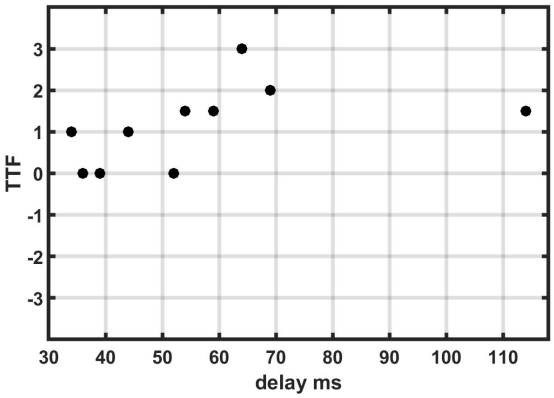

Fig. 9. I was trying to follow vs. MM2ME delay.

participants had to choose between very good (3) and very bad (-3). Answers range from 2 to 3, indicating that regardless of the conditions, each participant believes that her partner was performing well.

8) In Figure 9 we show whether each musician was trying to follow her partner in terms of tempo. The average value of the answers varies in range from 0 to 3 (fully agree). We observe an increase from 0 to 3 as delay increases from 35 to 70 ms. This is an indication that the participants were making a bigger effort to follow their partner with higher delays, something to be expected.

B. Scenario B

Below we discuss the results shown in the eight plots from Figure 10 to Figure 17. In this scenario we manipulated audio quality using Aretousa as mentioned in Section IV, using the sampling rates shown in Table II. The audio delay in this scenario was kept as low as possible (34 ms two-way). Again, we plot the average value from all the experiments for each sampling rate.

1) We can see the results for the *Perception of Audio Quality* question in Figure 10. The answers vary from 1, corresponding to very poor audio quality, to 4, corresponding to good audio quality. Although there is a decrease down to 2.5 in the range 20–45 kHz, the rest of the answers for sampling frequencies above 15 kHz are around 4. This indicates that there is a threshold in the frequency of 15 kHz, above which participants perceived audio quality as good enough.
2) With a sampling frequency of 8 kHz, there was a perceivable amount of delay inserted due to the resampling filter used in the Aretousa software. Therefore, the actual audio delay ended up being more than 34 ms. This explains the results in Figure 11, where *Synchronization degree* is perceived as bad (1) with low sampling frequencies. It increases with sampling frequencies above 16 kHz, where the delay actually falls to 34 ms; in this range, the scores are around 4, indicating good synchronization.
3) For the same reason as in the previous question, the *Perception of Audio Delay*, shown in Figure 12, shows higher perceived delays at low sampling rates. Above 16 kHz, the answers are around 1.5, therefore variations in sampling frequency do not seem to affect the perception of audio delay.
4) Figure 13 shows the answers for the evaluation of the *Perception of Musical and Emotional expression*. Again, there are issues at low sampling rates. There is a threshold in the frequency of 25 kHz, above which there are no variations in the answers, which range from 3.5 to 4, indicating high levels of expression.
5) In this scenario, when sampling frequency was manipulated there were no noticeable audio clicks as shown in Figure 14. The answers vary between 1 (no clicks) to 1.5 (very few clicks).
6) The *Perception of Satisfaction* variable was found to change as the sampling frequency was increased, as shown in Figure 15. The answers are in the range between 2 (low satisfaction) and 3.5 (average) satisfaction, with the highest satisfaction scores found only at the two highest sampling frequencies (68 and 88.2 kHz).
7) The responses to the *Perception of My Partners' Performance* question in this scenario varied between 2 and 3 (good performance) as shown in Figure 16. They do not seem to be consistently affected by the sampling frequency changes.
8) Finally, Figure 17 shows whether the musician was

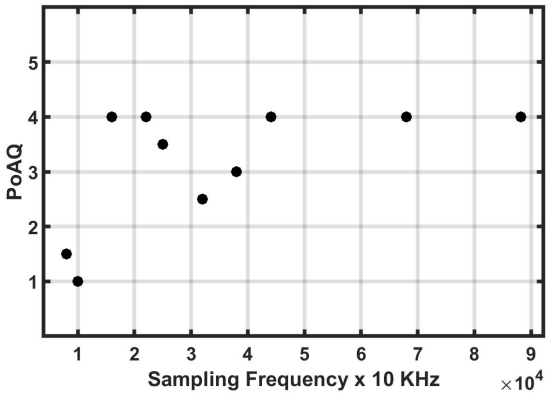

Fig. 10. Perceived Audio Quality vs. Sampling Rate.

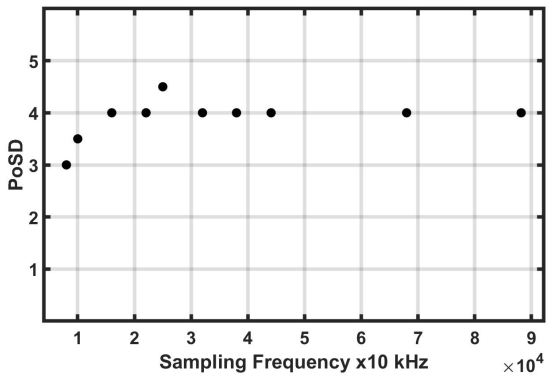

Fig. 11. Perceived Synch Degree vs. Sampling Rate.

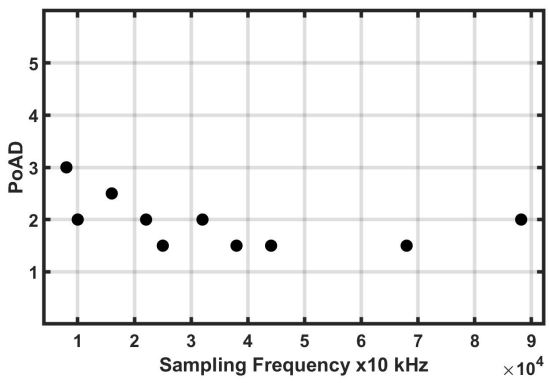

Fig. 12. Perceived Audio Delay vs. Sampling Rate.

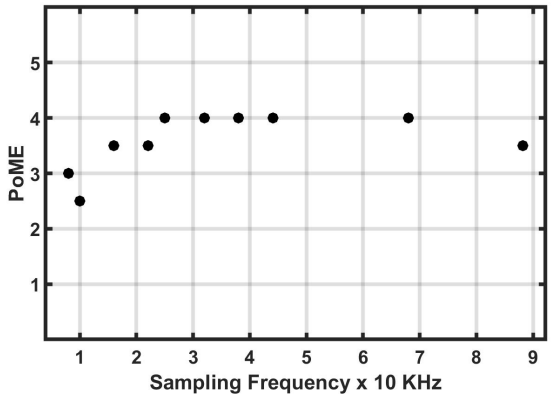

Fig. 13. Perceived Expression vs. Sampling Rate.

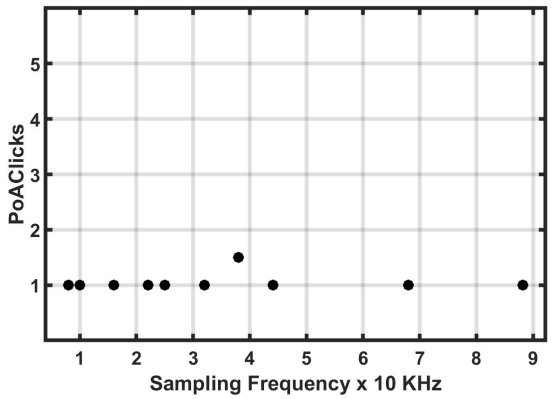

Fig. 14. Perceived Audio Clicks vs. Sampling Rate.

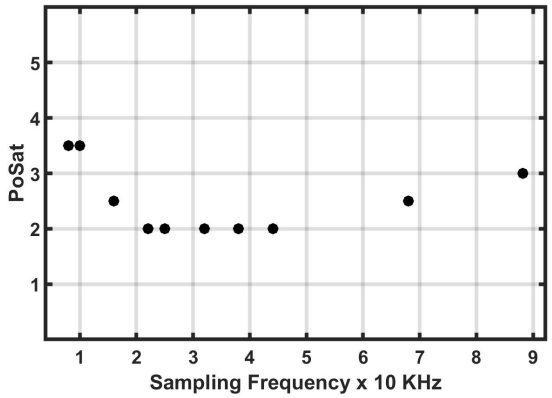

Fig. 15. Perceived Satisfaction vs. Sampling Rate.

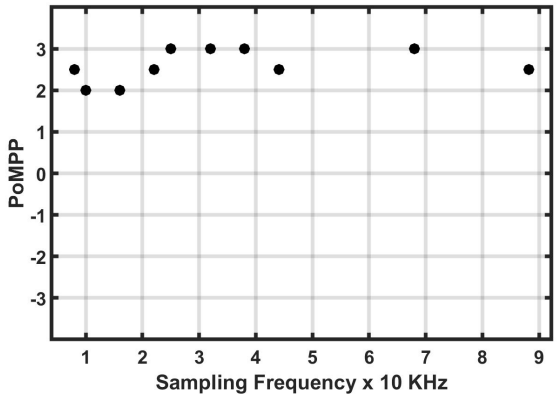

Fig. 16. My partner played well vs. Sampling Rate.

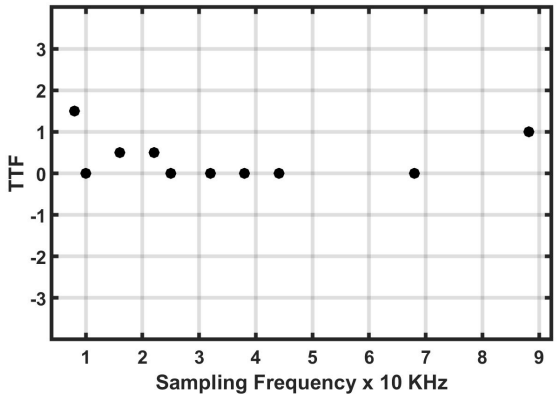

Fig. 17. I was trying to follow vs. Sampling Rate.

trying to follow her partner. Due to the inserted delay because of the audio buffer interfaces for the lowest sampling frequencies, we observe a average value of 1.5 corresponding to almost agree. In the range between 16 and 68 kHz, the average value of the answers is equal to 0, corresponding to neutral. As a conclusion we can say that tempo is not affected by sampling frequency changes.

VI. Conclusions and Future Work

We conducted a pilot NMP study in two sets of experiments, varying audio delay in the first set and audio quality in the second set. The eight musicians that participated in duets in both scenarios, performing the same piece 10 times for each scenario, responded to an eight-question survey after each performance, producing 10 sets of responses for delay and 10 for quality.

A general conclusion from the results is that the participants could not distinguish between noise, audio clicks, audio quality and other audio impairments. For example, when resampling was introduced as a result of a bad configuration, a phenomenon that occurs in low sampling frequencies, musicians perceived that change as noise, rather than delay. The same thing happened with audio clicks, which influenced the delay scores, although they introduced quality issues. In

general, when something unusual occurs, it is not easy to predict how the musicians will classify it.

From the results of scenario A (delay variation), perception of audio delay (PoAD), perception of synchronization degree (PoSD) and the try to follow (TTF) variables were found to be strongly correlated to audio delay variations. There is a threshold around 60 ms for the MM2ME delay above which further increases do not influence these variables, indicating that the acceptable delay range is around this threshold. This is consistent with previous work that indicates that the one-way delay threshold (half of MM2ME) is around 25 ms [3], [6], [5]. However, as we had no measurements in the 70 to 110 ms range, the acceptable threshold may be higher than revealed by our experiments. Other variables were found to be almost unaffected by the audio delay variations, although quality issues at lower delays may have affected the corresponding results.

From the results of scenario B (sampling frequency variation), it seems that at sampling frequencies higher than 16 kHz the scores were nearly the same in most variables, indicating that audio quality may not be affected above this threshold. Therefore, NMP could use sampling rates of 22 or 25 kHz to save on bit rate, without affecting the QoME. However, the resampling issues we faced at the lowest sampling rates which led to artificial delays, do now allow a clear interpretation of results at the low end of the sampling range.

For our future work, we plan to undertake a more comprehensive study, using a larger numbers of musicians in order to gather more measurement points so as to increase the confidence in our results. To avoid noise in the experiments, we are working on fixing the audio issues (clicks) of our software. In future measurements, we will also adjust the range of sampling frequencies to avoid the resampling issues we faced, as these influence the QoME results in an unintended manner.

VII. Acknowledgments

We would like to thank all the participating musicians for their patience during the experiments. This research has been partly financed by the Research Centre of Athens University of Economics and Business, Greece.

References

[1] K. Tsioutas, I. Doumanis, and G. Xylomenos, "A framework for understanding and defining quality of musicians' experience in network music performance environments," in *Audio Engineering Society Convention 146*, Mar 2019. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=20333

[2] J. E. Resnicow, P. Salovey, and B. H. Repp, "Is recognition of emotion in music performance an aspect of emotional intelligence?" *Music Perception: An Interdisciplinary Journal*, vol. 22, no. 1, pp. 145–158, 2004. [Online]. Available: http://mp.ucpress.edu/content/22/1/145

[3] K. Kubacki, "Jazz musicians: creating service experience in live performance," *International Journal of Contemporary Hospitality Management*, vol. 20, no. 4, pp. 303–313, 2008.

[4] C. Chafe and M. Gurevich, "Network time delay and ensemble accuracy: Effects of latency, asymmetry," in *Audio Engineering Society Convention 117*, Oct 2004. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=12865

[5] C. Bartlette, D. Headlam, M. Bocko, and G. Velikic, "Effect of network latency on interactive musical performance," *Music Perception: An Interdisciplinary Journal*, vol. 24, no. 1, pp. 49–62, 2006. [Online]. Available: http://www.jstor.org/stable/10.1525/mp.2006.24.1.49

[6] P. F. Driessen, T. E. Darcie, and B. Pillay, "The effects of network delay on tempo in musical performance," *Computer Music Journal*, vol. 35, no. 1, pp. 76–89, Mar. 2011. [Online]. Available: http://dx.doi.org/10.1162/COMJ_a_00041

[7] A. Olmos, M. Brulé, N. Bouillot, M. Benovoy, J. Blum, H. Sun, N. W. Lund, and J. R. Cooperstock, "Exploring the role of latency and orchestra placement on the networked performance of a distributed opera," in *Proceedings of the 12th Annual International Workshop on Presence*, 2009.

[8] K. Tsioutas, G. Xylomenos, and I. Doumanis, "Aretousa: A competitive audio streaming software for network music performance," in *Audio Engineering Society Convention 146*, Mar 2019. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=20334