Multimodal Assessment of Network Music Performance

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

***Abstract*—The most common method of assessing the Quality of Musician's Experience (QoME) in Network Music Performance (NMP) is to perform a subjective study, where the participants evaluate their experience via questionnaires. Translating experiences into metrics is not an exact science, though: in our recent study on the effects of audio delay and quality in the QoME of NMP, the responses had a high variance and were inconsistent. To strengthen our confidence in the results of the subjective study, we analyzed video recordings of the participants using machine learning. Specifically, we used Facial Expression Recognition (FER) to detect the emotions felt by the participants and then compare them with their questionnaire responses. In addition to pointing out interesting phenomena that were not apparent from the questionnaires, this multimodal analysis showed analogies between the emotions felt (as captured by FER) and the emotions expressed (as captured by the responses).**

***Index Terms*—NMP, QoME, Facial Expression Recognition, Emotion Detection.**

I. Introduction

During the recent pandemic, methods of real-time remote collaboration have come to the forefront of computing research. *Network Music Performance* (NMP), that is, the performance of music when musicians are connected over a network, is a classic case of such an application, which is however hampered by its need for ultra low delay communication. Even though all human-to-human communications have strict delay requirements, NMP is an outlier: while regular video conferencing can tolerate up to 100 ms of one way delay, in NMP delays of more than 25-30 ms are considered problematic [1]. These delay limits were derived either from artificial scenarios, such as participants trying to synchronize hand claps, or from small studies with real musicians. Interestingly, some studies have reported that actual musicians managed to cope with higher delays during real performances [2], indicating that the aesthetic experience of a music performance is a more complex phenomenon.

Although delay is the most crucial factor in NMP, when bandwidth is limited quality and delay can be traded off against each other. Single channel uncompressed CD quality audio (sampling at 44.1 KHz with 16 bits per channel) requires around 700 Kbps, which is close to the *uplink* capacity of ADSL connections (768 kbps – 1 Mbps), without even considering video. Since audio compression/decompression introduces delays, in bandwidth-limited network scenarios we either need to reduce the audio quality to allow for uncompressed audio to be sent with minimal delay, or compress the audio at the cost of increasing delay.

In order to evaluate more accurately the effects of both audio delay and audio quality on the *Quality of Musicians' Experience* (QoME) of NMP, we performed a large number of controlled experiments (22 subjects in total, the largest study we are aware of), where pairs of musicians played a musical piece of their choice under different delay and quality settings, completing a questionnaire at the end of each performance [3]. As in many other subjective studies, the questionnaire results exhibit a high degree of variance. In addition, even individual musicians did not provide consistent responses. For example, as the underlying delay grew, musicians did not consistently grade their perception of delay as being worse.

For this reason, we considered alternative methods of evaluating the QoME. Specifically, having recorded videos of our NMP sessions, we used machine learning techniques to extract the facial characteristics of the performers in order to determine their emotional state. Our goal was to assess whether the emotion analysis agreed with the subjective evaluation and whether it would uncover any interesting phenomena; essentially, this made the assessment multimodal. A longer term goal was to assess whether emotion analysis could be used in the future for QoME assessment, providing a more complete picture than what is possible by relying solely on questionnaires. To the best of our knowledge, this is the first study to apply emotion recognition and multimodal assessment to NMP experiments.

The outline of the rest of the paper is as follows. In Section II, we briefly present related work on assessing the effects of delay and quality on QoME. Section III describes the setup of our experimental scenarios. In Section IV we explain the method used to detect the musicians' emotions during the sessions as delay and quality are varied, while in Section V we present the results from this analysis and correlate them with the subjective evaluation. We summarize our findings and discuss future work in Section VI.

II. Related Work

Most studies of the QoME of NMP are subjective, that is, musicians respond to surveys evaluating their experience, while a parameter, such as audio delay or quality, is manipulated. For example, rhythmic hand clapping was used to investigate the effects of delay in the tempo in [4], [1], while musical instruments were used in [5], [6]. Some studies have also analyzed audio recordings of the NMP sessions, looking specifically at the variance of the performance tempo [6], [7].

An extended review of the studies related to emotion recognition through various sensors can be found in [8]. A person's emotional state may change depending on their subjective experience [9]. The emotional state of a person can be evaluated by varying environmental conditions; this evaluation can benefit from self reports, as well as from the data collected by sensing devices [10], [11].

The effects of music in generating emotions to listeners have been explored in multiple studies where listeners were asked to listen to musical pieces and data were gathered through electromyograms for zygomatics, skin conductance and heart rate [12]. In one of the few studies of emotions specific to NMP, the author asked six singers and a pianist to perform remotely, following a conductor through TV monitors. In parallel, data were gathered from wearable sensors measuring the performers' galvanic skin responses [13].

In [14] an extended review of previous works on emotion recognition is presented where multiple physiological signals are employed, such as EEG, electromyogram, electrocardiogram and skin conductance, to extract emotional information using various stimuli such us music, movies, robot actions. Gabrielsson and Juslin [15] employ Emotional Expression in music performance as an instrument to communicate emotions to listeners. In [16], the authors state that emotions are highly subjective and emotional changes can be observed for a very small time between 3 and 15 sec.

Ekman [17] states that facial and vocal expression, as well as gestures and posture, during emotion episodes are generally considered to be central motor components of emotion. On the other hand, Scherer [18] argues that the issue of emotions induced by music is a complex task and inappropriate measurements can miss essential aspects of the phenomenon or obtain biased data. Gabrielsson and Juslin [19] note that subjective strategies like rating sheets measure the subjective perception of *expressed* emotion rather than *felt* emotion.

Our work essentially focuses on correlating the *felt* emotion, which we try to detect via *Facial Expression Recognition* (FER), and the *expressed* emotion, which was evaluated via the questionnaires; it is a multimodal assesment, attempting to correlate the results from both methods. Although emotion analysis via FER is not a highly accurate method, our hope is that by considering both the qualitative results from the questionnaires and the quantitative results from emotion analysis we may derive a more accurate characterization of the QoME of NMP and, eventually, complement the questionnaires with automated assessment methods.

III. Experimental Setup

For our experiments, we used two visually and aurally isolated rooms on the same floor of our building. Musicians performed with their counterparts in separate rooms, while listening to them through headphones and seeing them through a 32" TV. We varied two underlying parameters: in Scenario A, audio delay varied while audio quality was fixed, while in Scenario B audio quality varied while audio delay was fixed. To conduct our experiments, we used the same topology with slightly different setups for each scenario.

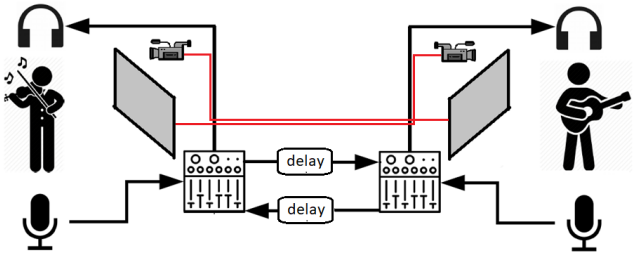

Fig. 1. Experimental Setup for Scenario A (variable delay).

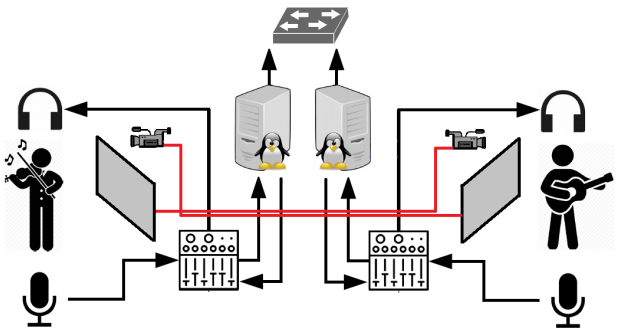

Fig. 2. Experimental Setup for Scenario B (variable quality).

In Scenario A, shown in Figure 1, an eight channel mixing console was used in each room for the necessary audio routing, monitoring and recording. Audio was captured by condenser microphones and closed type headphones were used by the musicians to listen to each other. A video camera was capturing and sending a composite (analog) video signal through the existing network cabling to the 32" TV of the other room (red lines in the figure); the camera was set up to provide a wide shot of the musician and his/her instrument, to help musical interaction. The network cables were patched directly to each other, without passing through any network equipment, providing us a direct analog connection between the camera and monitor. Our goal was to achieve the lowest possible visual delay between musicians, which was experimentally measured to be about 15 ms from the HD camera to the TV. The two mixing consoles were also connected through the existing network cabling, using direct cable patching, hence the audio signal was also transmitted in analog form from one room to the other. The reason for connecting them directly was to be able to achieve perfectly fixed audio delays even below 10 ms, which is impossible when computers and network devices intervene in the signal path. We used two AD-340 audio delay boxes by Audio Research between the two mixing consoles, via which we were able to set the audio delay in each direction to the desired value.

In Scenario B, the setup was modified to the one shown in Figure 2. The audio signals from the mixing consoles were fed to PCs running Linux, where our own NMP software [20] digitized and sent the audio streams. We used our software to manipulate the audio sampling rate, hence altering the audio quality; we did not compress the audio signal. The video setup

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

TABLE I
SCENARIO A: MM2ME DELAYS.

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

TABLE II
SCENARIO B: SAMPLING RATES.

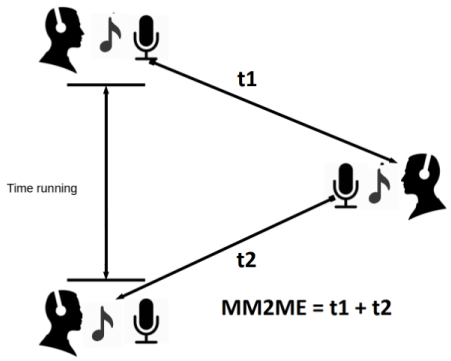

Fig. 3. My Mouth to My Ear delay.

was the same as in Scenario A, hence the delay was again 15 ms. The PCs were connected via the Fast Ethernet LAN of the building, with three Ethernet switches in the path; audio delay was experimentally measured to be about 10 ms in each direction.

Unlike most NMP studies which use *Mouth to Ear* (M2E) delay, which is the end-to-end delay between the microphone at one end and the speaker at the other end, in our work we use the *My Mouth to My Ear* (MM2ME) delay. As shown in Figure 3, MM2ME is the two-way counterpart to M2E, over which it has three advantages. First, when musicians play together, each musician plays one note and unconsciously expects to listen to the other musicians' note to play his next one, and so on. Second, measuring MM2ME delay accurately is much easier than measuring the M2E delay, as it can be done at one of the endpoints, by simply reflecting the transmitted sound at the other endpoint and comparing the input and output sound; in contrast, M2E needs to be measured at both endpoints, thus requiring perfectly synchronized clocks [21]. Third, MM2ME takes into account the possible asymmetry between the two directions of a connection.

The 22 musicians participating in the study performed in pairs (11 pairs in total), with each pair playing different musical instruments. Each pair of musicians played a one minute musical part of their choice, following their own tempo and repeating it ten (10) times, using a different MM2ME delay setting for each repetition; Table I shows the delays used. Then, the musicians performed the same musical piece ten (10) more times using a different audio sampling rate; Table II shows the rates used. No metronome or other synchronization aids were used.

After the end of each repetition, each musician was asked to answer an electronic questionnaire on a tablet. In this paper, we only consider the answers to three questions: *Perception of Satisfaction* (PoSat), which was graded on a 5 point Likert scale (from 1, not satisfied at all, to 5, very satisfied), *Perception of Audio Quality* (PoAQ), also graded on a 5 point Likert scale (from 1, very low quality, to 5, very high quality) and *Perception of Audio Delay* (PoAD), again graded on a 5 point Likert scale (from 1, very low delay, to 5, very high delay); results from the entire questionnaire are reported in [3].

Musicians were not informed about which variable was manipulated each time, or about the purpose of the experiment, and we randomly set the order in which the audio delay values and sampling rates were set for each repetition, as shown in Tables I and II. The main goal was to conduct an experiment that would allow us to evaluate multiple variables without bias or noise in the answers.

IV. EMOTION DETECTION WITH MACHINE LEARNING

To process the videos recorded during our NMP experiments, we turned to machine learning techniques which analyze the facial expressions of the participants in order to derive their emotions. *Deep Neural Networks* (DNNs) have become the standard in modern emotion detection, which is based on *Facial Expression Recognition* (FER) [22]. This process consists of three main stages: pre-processing, feature learning and feature classification. We will briefly present these stages and how they are implemented in the DeepFace system that we employed.

Since our videos were not recorded with the intention of performing FER as explained above, they exhibit considerable variations on background, illumination and head poses. In such *unconstrained* scenarios, pre-processing is required to align and normalize the visual semantic information conveyed by the face. The first step is to detect the face and then remove the background and non-face areas (face alignment phase). To avoid overfitting and ensure generality, DNNs require sufficient training data, which the publicly available datasets often fail to provide. Therefore, input samples are randomly cropped from the four corners and center of the image and then flipped horizontally, which can result in a dataset that is many times larger than the original training data. The final pre-processing step, face normalization, ameliorates variations in illumination and head poses that are likely to impair FER performance.

After pre-processing is completed, the feature learning stage is performed. Some of the most common DNNs that have been

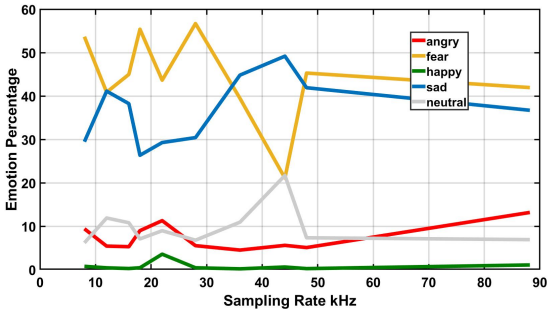

Fig. 4. Musician A's emotions vs. Sampling Rate (each point depicts averages from a single performance).

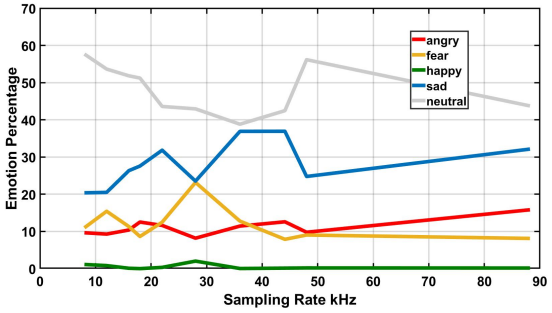

Fig. 5. Musician B's emotions vs. Sampling Rate (each point depicts averages from a single performance).

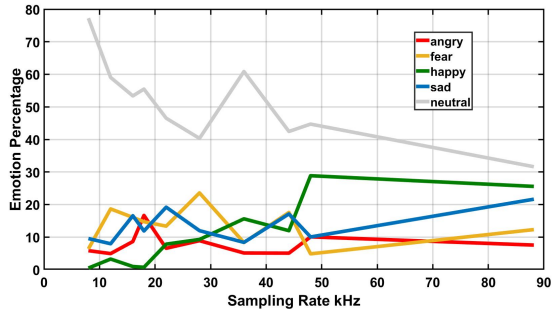

Fig. 6. Musician C's emotions vs. Sampling Rate (each point depicts averages from a single performance).

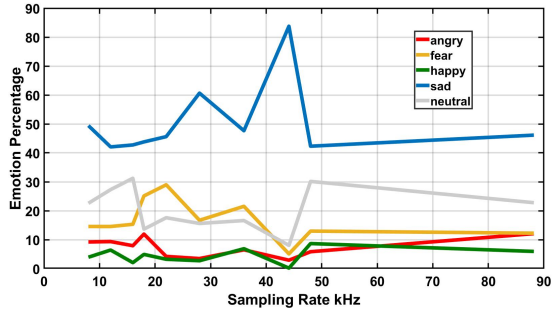

Fig. 7. Musician D's emotions vs. Sampling Rate (each point depicts averages from a single performance).

used for FER are *Convolutional Neural Networks* (CNNs), Deep Belief Networks, Deep Autoencoders, Recurrent Neural Networks and Generative Adversarial Networks. Finally, after the features have been extracted, the model has to classify a given face into one of the basic emotion categories. DNNs can perform this action in an end-to-end way, by adding a loss layer at the end of the network to regulate the back-propagation error, or alternatively employ a CNN as a feature extraction tool and then apply additional independent classifiers, such as Support Vector Machines or Random Forests, to the extracted features.

For this work, we used the DeepFace system to analyze the videos of the musicians[1]. DeepFace is an open-source face recognition and facial attribute analysis framework for python, mainly based on Keras and TensorFlow. According to [23] DeepFace can achieve more than 92% accuracy. To perform face detection, the *Multi-Task cascaded Convolutional Neural Network* (MTCNN) detector was utilized, since it seemed to outperform the other detectors supported by Deepface in this use case [24]. The output of the face recognition stage is a bounding box for the face (a 4 element vector), a 10 element vector for facial landmark localization and the positions of five facial landmarks, two for the eyes, two for the mouth and one for the nose [25]. The final step is to classify the given face into one of the basic emotion categories (anger, disgust, fear, happiness, sadness, surprise, and neutral). A fully connected CNN model, with three convolution layers is employed as a feature extraction tool.

[1]https://pypi.org/project/deepface/

The DeepFace system essentially examines each frame of a recorded video, detects a human face and decides which emotions are present, using a large set of images as a training model. Thus, for a 30 second video shot at 30 frames per second, 900 frames must be examined for emotion detection. For each frame the algorithm produces (estimates) a percentage value for each emotion. As an example, for a random frame a musician was found to be a/100 angry, d/100 disgusted, f/100 frightened, h/100 happy, sa/100 sad, su/100 surprised and n/100 neutral with SUM(a,d,f,h,sa,su,n)=100. When we report results for an entire session, we simply find the average fraction of each emotion across all video frames of the performance.

There are two issues with using DeepFace for the analysis of our video recordings. First, the videos were captured directly by the cameras used in the experiment, which were set up to support musical interaction, thus offering a wide shot of the musicians and their instruments. As a result, the videos are not ideal for facial recognition, as faces are a small part of the frame, they are usually shown in profile and they can be partially obscured by headphones, microphones, cables and musical instruments. Ideally, a separate pair of cameras would have focused on the performer's faces, to help with the analysis. Second, the emotions detected by the DeepFace system are generic, rather than those expected in an NMP scenario; for example, in NMP it is unlikely to experience disgust, but it is likely to experience frustration.

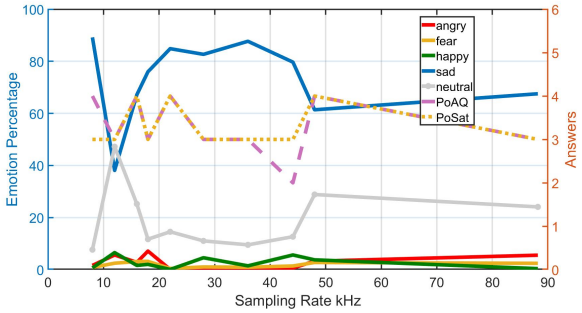

Fig. 8. Musician E's Emotions (left y axis, solid lines) and PoSat/PoAQ (right y axis, dashed lines) vs. Sampling Rate: PoSat follows neutrality.

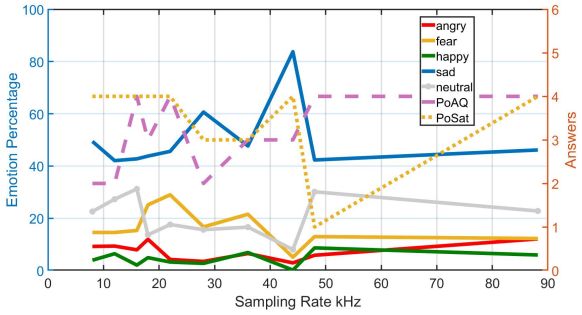

Fig. 9. Musician F's Emotions (left y axis, solid lines) and PoSat/PoAQ (right y axis, dashed lines) vs. Sampling Rate: PoSat follows sadness.

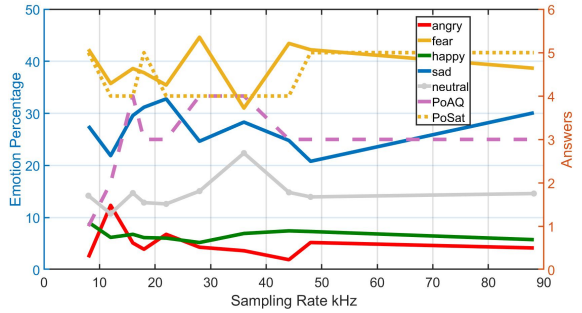

Fig. 10. Musician's G's Emotions (left y axi, solid lines) and PoSat/PoAQ (right y axis, dashed lines) vs. Sampling Rate: PoSat follows fear.

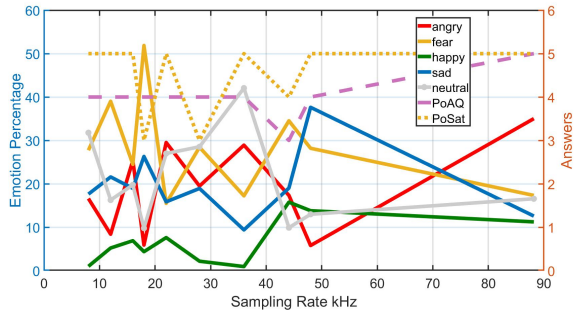

Fig. 11. Musician's H's Emotions (left y axis, solid lines) and PoSat/PoAQ (right y axis, dashed lines) vs. Sampling Rate: PoSat follows anger.

V. Analysis Results

We analyzed the results of both Scenario A (variable delay) and Scenario B (variable sampling rate) with DeepFace. A first observation is that each video analyzed by the algorithm revealed a different dominant emotion, depending on the musician. For example one musician was found to be mostly sad during all the sessions that he participated in, no matter the audio conditions he was exposed to. Similarly, another one was found to be mostly neutral and so on. This indicates that emotion detection through face analysis produces results that mix the general emotional state of a participant and the specific emotions induced by the NMP experiment; it would be unrealistic to expect participants to shut off all other emotions during their performance.

A second observation was that the emotional reactions when audio conditions changed were different for each musician. However, interesting points come up by looking at the results. For example, Figures 4, 5, 6 and 7 show the average percentages of each emotion for an entire performance for four random participants as the sampling rate is modified; note that we do not show disgust and surprise, as they were negligible. A sharp change in the emotions, either increasing or decreasing, occurs when the sampling rate changes from 44.1 to 48 kHz. Even though the change was different for each musician, it was common for most of the participants, indicating that this specific sampling rate change was noticeable to the participants.

Figures 8, 9, 10 and 11 show the average percentage of each emotion for an entire performance (left y axis, solid lines) and the scores of the PoSat and PoAQ subjective variables (right y axis, dotted lines) against the sampling rate, for four selected musicians. It is interesting to note that while the PoAQ line depicting the Perception of Audio Quality does not look like any of the emotion curves, the PoSat lines depicting the Perception of Satisfaction do: in Figure 8 PoSat follows neutrality, in Figure 9 PoSat follows sadness, in Figure 10 PoSat follows fear and in Figure 11 PoSat follows anger; note that since the two y axes have different scales, it is the trends (up/down) that matter rather than the absolute values. The matching is not perfect, it relates to a different emotion for different musicians and it is not so clear in every case, but it is intriguing that such a match does exist in many cases, as it indicates that the PoSat answers (the *expressed* emotion) do have a correlation with the emotions detected (the *felt* emotion), even though the relationship is not clear enough to allow us to make conclusions without the subjective analysis.

Figures 12 and 13 show the average values of emotions across all 22 participants, for each sampling rate and delay value, respectively, as well as the appropriate subjective variables, that is, PoAQ and PoSat when audio quality (i.e., the sampling rate) is modified and PoAD and PoSat when audio delay is modified. Neutrality and sadness are the dominant emotions in both scenarios. When the audio quality is modified, we can see in Figure 12 a clear disruption between 44.1 and 48 kHz, as mentioned above. Furthermore, we see an increase in happiness and anger and a decrease in sadness and fear as the sampling rate, and hence the audio quality, is increased. Looking at the subjective variables, both PoSat and

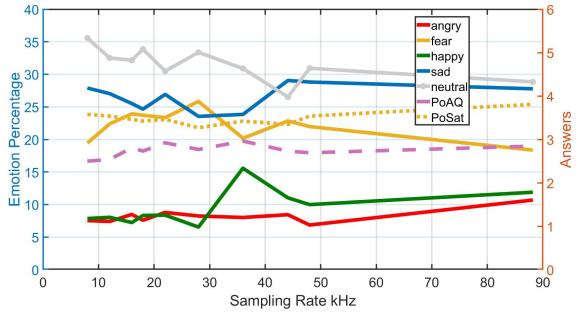

Fig. 12. Average values of emotions (left y axis, solid lines) and PoSat/PoAQ (right y axis, dashed lines) across all musicians vs. Sampling Rate.

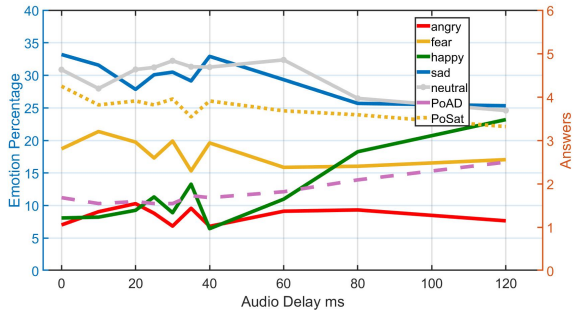

Fig. 13. Average values of emotions (left y axis, solid lines) and PoSat/PoAQ (right y axis, dashed lines) across all musicians vs. Audio Delay.

PoAQ only improve slightly with higher sampling rates, and they do not seem similar to any of the emotion curves.

On the other hand, in Figure 13 we can see a disruption as delay grows from 20 to 40 ms, where it starts becoming noticeable. Interestingly, at this point happiness starts to grow and sadness starts to drop; the reason is that as the musicians became unable to synchronize, they would often burst into laughter, which made the system detect hapiness! The implication here is that additional information is needed to interpret such results, beyond the curves. Looking again at the subjective variables, PoSat drops with increasing delay, while PoAD, the perception of Audio Delay, grows, which are both as expected. We also note that PoSat has a similar shape to the fear curve.

Looking at both Figure 12 and Figure 13, we can see that the audio quality has a much smaller effect on satisfaction (PoSat) than the audio delay: it seems that musicians detect delay changes (PoAD) easier than quality changes (PoAQ), with a corresponding effect on satisfaction. Their emotional responses are also stronger with delay changes, since after the discontinuity evident in both figures, the emotions change more abruptly with increasing delay than with increasing quality.

At the same time, while emotion analysis via FER, at least in our setup where the video was not captured with this intention, does show clear emotional responses for individual musicians and when averaging results across musicians, indicating that the subjective analysis does capture the felt emotions, it cannot *by itself* provide concrete results for the QoME of NMP: in addition to being rather inexact and not showing statistically significant correlation with the subjective results, it also suffers from unexpected responses (e.g., musicians laughing when losing sync). For this reason, this additional mode of assessment can be used to support, but not to replace the results of the subjective analysis.

VI. Summary

We conducted a set of NMP experiments, where the audio delay and audio quality between a pair of musicians was varied in a controlled manner for each session, with video from the sessions being recorded for later analysis. In our experiments, 22 musicians participated as pairs, playing a diverse set of musical instruments and performing in a variety of musical styles.

The analysis performed on the recorded video revealed that emotion detection via facial emotion recognition is not as conclusive as we would like, since each musician's emotional state cannot realistically be affected only by the NMP session and, sometimes, additional information is needed to interpret the trends exhibited in the results. However, this additional mode of assessing the experience of the musicians can be used to strengthen the conclusions drawn from subjective studies. It would be worthwhile in future experiments to use dedicated cameras for emotion analysis, focused on the faces of the performers, so as to facilitate emotion detection.

As future work, in addition to looking at alternative video analysis tools, we are planning to analyze the video data by grouping the experiments by instrument, style and tempo, to determine whether the emotional responses have a correlation with these factors. We are also looking for methods to normalize the emotion analysis results via the "subtraction" of the baseline emotional state of each musician, which is independent of the NMP session, so as to look only at the changes to the emotional state that can be attributed to the participation to the NMP experiment.

Acknowledgments

We would like to thank all the participating musicians for their patience during the experiments, as well as the fellows who helped with setting up and carrying out the experiments.

References

[1] N. Schuett, "The effects of latency on ensemble performance," Bachelor Thesis, CCRMA Department of Music, Stanford University, 2002.

[2] K. Tsioutas, G. Xylomenos, I. Doumanis, and C. Angelou, "Quality of musicians' experience in network music performance: A subjective evaluation," in *Audio Engineering Society Convention 148*, May 2020.

[3] K. Tsioutas, G. Xylomenos, and I. Doumanis, "An empirical evaluation of QoME for NMP," in *IFIP International Conference on New Technologies, Mobility and Security*, April 2021.

[4] C. Agastya, D. Mechanic, and N. S. Kothari, "Mouth-to-ear latency in popular VoIP clients," Department of Computer Science, Columbia University, Tech. Rep., 2009, CUCS-035-09.

[5] A. Carôt and C. Werner, "Fundamentals and principles of musical telepresence," *Journal of Science and Technology of the Arts*, vol. 1, pp. 26–37, May 2009.

[6] C. Rottondi, M. Buccoli, M. Zanoni, D. Garao, G. Verticale, and A. Sarti, "Feature-based analysis of the effects of packet delay on networked musical interactions," *Journal of the Audio Engineering Society*, vol. 63, pp. 864–875, November 2015.

[7] K. Tsioutas and G. Xylomenos, "Assessing the QoME of NMP via audio analysis tools," in *Proceedings of the International Conference on Signal Processing and Multimedia Applications (SIGMAP)*, July 2021.
[8] A. Raheel, M. Majid, M. Alnowami, and S. M. Anwar, "Physiological sensors based emotion recognition while experiencing tactile enhanced multimedia," *MDPI Sensors*, vol. 20, no. 14, p. 4037, Jul 2020.
[9] H. Gunes, B. Schuller, M. Pantic, and R. Cowie, "Emotion representation, analysis and synthesis in continuous space: A survey," in *IEEE International Conference on Automatic Face Gesture Recognition (FG)*, March 2011, pp. 827–834.
[10] C. L. Bethel, K. Salomon, R. R. Murphy, and J. L. Burke, "Survey of psychophysiology measurements applied to human-robot interaction," in *IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2007, pp. 732–737.
[11] A. Dzedzickis, A. Kaklauskas, and V. Bucinskas, "Human emotion recognition: Review of sensors and methods," *Sensors*, vol. 20, no. 3, 2020.
[12] P. Juslin and P. Laukka, "Expression, perception, and induction of musical emotions: A review and a questionnaire study of everyday listening," *Journal of New Music Research*, vol. 33, pp. 217–238, 09 2004.
[13] 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 *12th Annual International Workshop on Presence*. ISRP, 2009, pp. 1–9.
[14] J. M. Garcia-Garcia, V. M. R. Penichet, and M. D. Lozano, "Emotion detection: A technology review," in *International Conference on Human Computer Interaction*, 2017.
[15] A. Gabrielsson and P. N. Juslin, "Emotional expression in music performance: Between the performer's intention and the listener's experience," *Psychology of Music*, vol. 24, no. 1, pp. 68–91, 1996.
[16] J. Selvaraj, M. M, R. Nagarajan, and W. Khairunizam, "Physiological signals based human emotion recognition: a review," in *IEEE International Colloquium on Signal Processing and Its Applications (CSPA)*, March 2011.
[17] P. Ekman, "Facial expressions of emotion: New findings, new questions," *Psychological Science*, vol. 3, no. 1, pp. 34–38, 1992.
[18] K. Scherer, "Which emotions can be induced by music? what are the underlying mechanisms? and how can we measure them?" *Journal of New Music Research*, vol. 33, pp. 239–251, 09 2004.
[19] A. Gabrielsson and P. Juslin, "Emotional expression in music," in *Handbook of Affective Sciences*, R. J. Davidson, K. R. Scherer, and H. H. Goldsmith, Eds. Oxford University Press, 2003, pp. 503–534.
[20] K. Tsioutas, G. Xylomenos, and I. Doumanis, "Aretousa: A competitive audio streaming software for network music performance," in *Audio Engineering Society Convention 146*, March 2019.
[21] A. Carôt, C. Hoene, H. Busse, and C. Kuhr, "Results of the Fast-Music project - five contributions to the domain of distributed music," *IEEE Access*, vol. 8, pp. 47 925–47 951, 03 2020.
[22] S. Li and W. Deng, "Deep facial expression recognition: A survey," *CoRR*, vol. abs/1804.08348, 2018. [Online]. Available: http://arxiv.org/abs/1804.08348
[23] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *IEEE Conference on Computer Vision and Pattern Recognition*, September 2014.
[24] S. I. Serengil and A. Ozpinar, "Lightface: A hybrid deep face recognition framework," in *Innovations in Intelligent Systems and Applications Conference (ASYU)*, October 2020, pp. 1–5.
[25] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.