Quality of Musicians' Experience in Network Music Performance

Author:
Konstantinos TSIOUTAS

Supervisor:
Prof. George XYLOMENOS
Co-Supervisor:
Prof. George POLYZOS
Co-Supervisor:
Prof. Vasilios SIRIS

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy

in the

Department of Informatics
School of Information Science and Technology

February 28, 2022

Abstract of thesis entitled

Quality of Musicians' Experience in Network Music Performance

Submitted by

Konstantinos TSIOUTAS

for the degree of Doctor of Philosophy

at Athens University of Economics and Business

in February, 2022

Abstract The increased use of tele-presence and tele-conferencing facilities, whether due to the need to isolate during a pandemic, or due to the desire to avoid costly and time consuming travel, prompted a renewed interest in Network Music Performance (NMP), where musicians collaborate remotely over the Internet in real time. Although the Internet has made dramatic leaps in capacity since the first NMP systems were created in the 20th century, the delays involved when communicating over the Internet, whether due to the physical distance between the endpoints, or due to the unpredictable nature of network traffic, are an important hindrance to the widespread use of NMP applications.

The main question that this thesis attempts to answer is how much delay humans are able to tolerate for NMP to be acceptable. To achieve this goal, we first identify the factors influencing the Quality of Musicians' Experience (QoME) during NMP. Out of these factors, we single out audio delay, which makes or breaks a performance. We also consider audio quality, as it may be reduced to save bandwidth, without resorting to delay-inducing audio compression. A review of the literature shows that past work on evaluating the human tolerance to delay during NMP either employs a scenario where music is not performed, that is, synchronization of hand claps, or involves a very small number of experiments, thus having low statistical significance.

Before embarking on a large scale study of NMP with actual musical performances, we first performed two exploratory studies. The first study tested our experimental setup, including the software and hardware employed, so as to ensure that the testing environment was acceptable to musicians and that we could gather accurate data without interruptions. The second study tested our assessment method, which consisted of questionnaires answered by each participant at the end of every performance, with a small number of musicians.

Based on these studies, we then designed and carried out the largest NMP study to date with actual musicians performing real musical pieces. In this study, we varied either audio delay or audio quality in a systematic manner, gathering up answers to

a fine-tuned questionnaire for QoME assessment. This subjective evaluation revealed that after crossing a quality threshold, further increasing audio quality had no discernible effects to QoME, indicating that when bandwidth is limited, we can sacrifice (up to a point) audio quality to reduce the required bitrate, without resorting to compression. On the other hand, we found that varying delay did have a statistically significant effect to QoME. More importantly though, our results indicate that the delay threshold up to which NMP is feasible is closer to 40 ms, rather than the 25-30 ms previously considered acceptable.

Having recorded audio and video from all sessions, we complemented this subjective study with three additional evaluation methods, making our work the first multimodal study of the QoME for NMP. First, we performed tempo analysis on the recorded audio, to assess the highest delay at which the musicians could maintain a steady tempo; the results from this study confirmed that delays of up to 40 ms are acceptable for NMP, as indicated by the subjective study. Second, we analyzed the audio features of the recordings, finding that delay had a larger impact on percussive instruments and musicians performing rhythm parts; this result confirmed similar results from a previous, but much smaller study. Third, we analyzed the video recordings in order to detect the emotions felt by the musicians using machine learning methods, finding that as audio delay or audio quality was varied there were clear disruptions in the emotions of the musicians; while these results are intriguing, they were not clear enough to substitute the subjective analysis.

Quality of Musicians' Experience in Network Music Performance

by

Konstantinos TSIOUTAS

A Thesis Submitted in Partial Fulfilment
of the Requirements for the Degree of Doctor of Philosophy

at the

Athens University of Economics and Business
February, 2022

Αυτή η διατριβή είναι αφιερωμένη
στον πατέρα μου Γεώργιο,
τη μητέρα μου Μαρία,
τον αδερφό μου Ιωάννη, τη σύζυγό του Μαριάν
και τα δύο υπέροχα παιδιά του,
τον Γιώργο και τη Μυρτώ
και στη βαφτιστήρα μου Μαριαλένα.

This dissertation is dedicated to my father George, my mother Maria, my brother Ioannis, his wife Marian and his two lovely kids, George and Myrto and to my god-daughter Marialena.

Acknowledgements

First, I would like to thank, my supervisor, Professor George Xylomenos, who believed in me, gave me space to expand my knowledge and supported me until the last moment. Many special thanks to my two co-supervisors, Professor George C. Polyzos, head of the Mobile Multimedia Laboratory at the Athens University of Economics and Business, and Professor Vasilios A. Siris.

Special thanks to professor Konstantinos Tiligadis from the Ionian University, who also supported me and gave me hope to continue my work.

Many thanks to, my colleagues in the Mobile Multimedia Laboratory, Yiannis Thomas, Livia Chatzieleftheriou, Iakovos Pittaras, Nikos Fotiou, Ioannis Karakonstantis, Merkouris Karaliopoulos who stood by me in all the good and the difficult times.

I would also like to thank Dr. Alexandros Eleftheriadis, for his reference letter to professor Xylomenos, as well as professor Andreas Mniestris from Ionian University who gave me the chance to attend my first master's course in Music Technology, where this journey started.

I would like to thank all the participating musicians for their patience during the experiments, as well as the fellows who helped me setting up and carrying out the experiments thus, Christos Angelou, Konstantinos Rantzos, Victor Mastela and Stergios Gkatsis.

Finally, I owe a great thanks to the staff in the Network Operating Center of AUEB, who helped me set up a customized connectivity setup for my experiments. Without them, the experiments would not have been conducted.

Konstantinos TSIOUTAS
Athens University of Economics and Business
February 28, 2022

JOURNALS:

[1] **Konstantinos Tsioutas** and George Xylomenos, 'Assessing the Effects of Delay to NMP via Audio Analysis," invited to the *Springer Nature Computer Science Journal* (under review)

[2] **Konstantinos Tsioutas** and George Xylomenos, "On the Impact of Audio Characteristics to the QoME of NMP," in *Journal of Audio Engineering Society*, vol. 69, no. 12, pp. 914–923, 2021.

CONFERENCES:

[1] **K. Tsioutas**, K. Ratzos, G. Xylomenos and I. Doumanis, "Multimodal Assessment of Network Music Performance," in *Companion Publication of the 2021 International Conference on Multimodal Interaction (ICMI '21 Companion), October 18–22, 2021, Montréal, QC, Canada*

[2] **Konstantinos Tsioutas**, Ioannis Doumanis, and George Xylomenos, "Assessing the QoME of NMP via Audio Analysis Tools," in *18th International Conference on Signal Processing and Multimedia Applications, January 2021*

[3] **Konstantinos Tsioutas**, Ioannis Doumanis, and George Xylomenos, "An empirical evaluation of QoME for NMP," in *2021 11th IFIP International Conference on New Technologies, Mobility and Security (NTMS), April 2021*

[4] **Konstantinos 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.*, URL:http://www.aes.org/e-lib/browse.cfm?elib=20774.

[5] **Konstantinos Tsioutas**, Ioannis Doumanis, and George Xylomenos, "A Framework for Understanding and Defining Quality of Musicians Experience in Network Music Performance Environments," in *Audio Engineering Society Convention 146. Mar. 2019.*, URL:http://www.aes.org/e-lib/browse.cfm?elib=20333.

[6] **Konstantinos Tsioutas**, G. Xylomenos, and I. Doumanis. "Aretousa: A Competitive Audio Streaming Software for Network Music Performance". in *Audio Engineering Society Convention 146. Mar. 2019.*, URL:http://www.aes.org/e-lib/browse.cfm?elib=20334.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In this introductory chapter, we first discuss the motivation for the research performed as part of the thesis, and then identify the questions we intend to answer and our objectives. We then explain how our work extends past work in the field and present our research methodology. Finally, we present the contributions of our research, and present the outline of the rest of the thesis.

1.1 Research Motivation

Starting in 2020, the world experienced the SARS-CoV-19 global pandemic, which spread to every country and changed our everyday life. This pandemic changed the way we communicate, the way we work and the way we meet. One outcome of the pandemic was that in order to maintain social distancing, as many everyday activities as possible were performed remotely. Thus, every job, task and communication between individuals that could be performed remotely, without physical presence in the same space, took place in this manner for many months, using a variety of tools.

Education was strongly affected by the pandemic. Educational activities like school lessons and university lectures were conducted remotely via various teleconference platforms since the pandemic began. School teachers, students, university and academic staff had to work remotely for extended periods of time. People that were not familiar with computers and Internet usage had to quickly learn how to use these tools in order to carry out their roles. In such a situation, global telecommunications and the Internet were suddenly the most important (and only) tools for carrying out everyday activities in conditions of global and total isolation. Among Internet services, the use of teleconferencing rapidly increased, as expected. Business partners had to meet using their web camera and headphones, isolated in their home. Teachers had to log in to virtual classrooms where students were logged in.

As teleconferencing services were widely used, they were stress tested during the pandemic, both in terms of performance and in terms of resilience. Many commercial

platforms offer real-time remote communication for companies, governments, universities, military, and the general public. Microsoft Teams[1], Skype[2], Zoom[3], Cisco Webex[4] and other platforms supported high-quality multi-party video and audio transmission for millions of users and thousands of sessions simultaneously. All these platforms had to support real-time video streaming, audio streaming, instant messaging, presentation uploading, file sharing, user authentication, cloud storage and even more features.

Cultural activities, such as music concerts and theatrical shows, were immediately canceled after the pandemic began. Alternate ways had to be invented so that they could be conducted in a way that would prevent crowding and virus spreading. Hence, music concerts and theater performances were broadcast remotely to avert crowding. In addition, many music artists were driven to broadcast their music events via YouTube[5] while others uploaded music videos on Facebook[6].

Music education was also adversely affected by the pandemic. In music education, the teacher and student have to be situated in the same room to interact musically, allowing the teacher to correct and guide the student. Together, they perform many repeated music exercises to improve the student's proficiency. During the pandemic, lessons also had to be conducted remotely. Music teachers and students were forced to use conferencing systems to attend their lessons, losing the ability to perform together, due to aforementioned delays that prevented synchronization.

Many videos were published where multiple musicians appeared in separate windows, performing together a musical piece, perfectly synchronized, even though they were situated in their own homes. As Duffy [73] reports, *...social media is currently awash with virtual choirs and orchestras, with the now familiar Zoom-style composite of all the members seemingly performing together.*. But could these videos be produced in real time or they are product of editing? Actually, in all cases a musical director guided a band, a choir, or an orchestra to record a basic track and then sent it to the rest of the members. Then, the members recorded their parts, based on that basic track and sent it back to the director. Finally, an editor fitted all the recorded pieces together to create a Zoom-like video. These videos were products of offline synchronization, rather than live performances using teleconferencing platforms, as the delay of conferencing systems made remote synchronization impossible.

Although standard conferencing tools exhibit very high delays for music performance, performing music via the Internet, commonly known as *Network Music Performance* (NMP), is actually feasible, under specific conditions. NMP is far more demanding than regular conferencing, requiring very fast and ultra low delay bidirectional audio (and, possibly, video) transmission. To put this into perspective, consider that the tolerable limit to audio delay for audio conferencing is considered to be at 100–150 ms,

[1] https://www.microsoft.com/en-us/microsoft-teams/group-chat-software
[2] https://www.skype.com/
[3] https://zoom.us/
[4] https://www.webex.com/
[5] https://www.youtube.com
[6] https://www.facebook.com

based on many subjective studies. For NMP, the delay tolerance is much lower; what that tolerance is, is a basic research question of this dissertation, but previous work puts it at 25–30 ms. Since conferencing tools are oriented to voice communication, where the acceptable delays are higher, the commercial video-conferencing platforms mentioned above cannot support remote music sessions, remote live concerts, rehearsals or any other NMP activities.

NMP of course has applications in all kinds of music performance, beyond education. Thousands of musicians around the world would benefit if they were able to perform music remotely with their partners, without travelling to the same place. Music concerts and all kinds of music events could take place remotely and be broadcast to the world if NMP was feasible. Similarly, music recordings, where musicians need to play together at least at a first stage (the so called basic tracks, over which each musician can later overdub additional or improved performances) could take place remotely.

Even without the pandemic restrictions, NMP could be of great help for many music education scenarios. Many countries and regions have rich musical traditions that are gradually becoming harder to maintain, since only a few people are interested in learning some traditional instruments, and even fewer musicians are available to teach them. With NMP, these teachers could reach students everywhere, thus preventing traditional music from fading away. More generally, the ability to perform together for rehearsals, recording or concerts, would reduce the need to commute or travel and its environmental impact.

1.2 Questions and Objectives

In order to explain the goals of this dissertation, the role of delay in musical performance must be explained. Let us consider the simplest case where two musicians are situated in the same room with a physical distance of 1–2 m. The goal is to perform together a music piece. A performance can be considered successful if the musical piece is performed in a synchronized manner, with a constant tempo and without mistakes. In such conditions, delay is roughly 3.5–7 ms, considering that the speed of sound is 343 m/s, from the moment the first musician plays a note, to the moment the second musician will listen to the note and respond with his note. Every musician can trivially confirm that under these conditions, performances can be successful.

Let us now consider a symphonic orchestra performing in a concert hall, where the distances between musicians are in the range of 1–20 m. In this case, the audio delay between them can be up to 70 ms. In such a case, musicians have a hard time synchronizing. To achieve synchronization, a conductor is needed, providing a tempo reference via hand movements; since light travels much faster than sound, musicians can use this visual cue to synchronize, despite the higher sound delays.

When using a network to connect the musicians, although the optical, electrical or radio signals travel with speeds closer to the speed of light, the large physical distances between the participants, as well as many other factors due to network switching, contribute to high end-to-end delays (see Chapter 3). It has been reported that when performers use hand claps, the delay above which performance is evaluated as not successful or not collaborative lies between 25 and 30 ms. This threshold is called the *Ensemble Performance Threshold, EPT* [77].

The basic goal of this dissertation is the examination of the conditions under which the *Quality of Musicians' Experience* (QoME) of an NMP session is evaluated by humans as satisfactory, not only for hand claps, but for real musical performances with actual instruments. That is, our main goal is to determine the tolerance of musicians to delay in NMP sessions, by estimating the upper limit to delay that allows performances to be satisfactory. A secondary goal is to determine the tolerance of musicians to audio quality in NMP, by estimating a lower limit for audio quality to be acceptable; this allows us to determine how much we can drop the bandwidth requirements of NMP without resorting to audio coding and its delays.

Assessing the QoME of NMP requires planning, executing and analyzing the results of a large number of NMP sessions under carefully controlled experimental conditions. Each of these steps requires some advances to the state of the art, therefore we can state in more detail the objectives of our thesis as follows:

- To develop a model for the QoME of NMP and identify the main parameters influencing it, so as to drive the experiments.
- To design and implement an experimental setup that allows gathering a wealth of data under controlled conditions.
- To experimentally evaluate musicians' tolerance to audio delay and audio quality using subjective methods.
- To investigate alternative, non subjective, methods for the evaluation of QoME, based on audio and video analysis.
- To combine results from multiple modes of analysis in order to determine the audio delay and quality tolerance limits of musicians for NMP.

1.3 Progress Beyond the State of the Art

Various research groups have created customized tools for NMP. The CCRMA group of Stanford University has developed JackTrip, a software for real-time audio collaboration. The LOLA project is an ongoing collaborative European effort on the field of NMP. SoundJack is another browser based platform for remote musical performance which is widely used in practice.

Numerous studies have been conducted in the field of NMP, focusing basically on the evaluation of the *Quality of Service* (QoS) offered by various network architectures and topologies, considering metrics such as bandwidth and delay. Many studies have also been conducted evaluating the audio delay threshold above which performance becomes infeasible. Most of these studies, though, asked people to remotely perform hand claps rather than real instruments.

Only a few studies focus on the musicians' experience while performing. The evaluation of users' experience is a new concept that concerns the media and computing industries, called *Quality of Experience* (QoE). Qoe can be defined as *the degree of delight or annoyance of the user of an application or service. It results from the fulfillment of his or her expectations with respect to the utility and / or enjoyment of the application or service in the light of the user's personality and current state* [57]. QoE then assesses the overall performance as perceived by the user at the application layer, whereas QoS is the overall measured performance of the end-to-end transport or network layer. That is, QoE is subjective (how do I rate a performance?), while QoS is objective (what was the audio delay of the performance?).

The evaluation of QoE turns to be a very complex task as, in general, it cannot be directly estimated from the underlying QoS: it depends on human factors (perceptions, expectations, emotions, needs), while taking into account various aspects, such as technology, service, context of use, etc. In the NMP context, we talk about QoME, which incorporates the intricacies of musical performance and its specific needs.

Our work aims to advance the state of the art in the evaluation of NMP, by first creating a framework for the evaluation of the QoME of NMP and identifying the main factors that influence it, and then creating a tool to help perform controlled NMP experiments. We devise methods for QoME assessment, including not only the traditional subjective methods for QoE evaluation, such as questionnaires, but also objective methods, based on audio and video analysis. After undertaking a pilot study to fine tune our experimental process, we complete the largest NMP to date with real musicians and musical performances, and apply a wide range of analysis methods to the measurements. We finally combine the results from these multiple analysis modes in order to derive more concrete conclusions.

1.4 Research methodology

Multiple research methods were employed in this thesis. The first step was to search the literature to understand the methods used in previous NMP studies. A clear outcome of this analysis was that in most previous studies, hand claps were used. Therefore, we noted a lack of large NMP studies with actual instruments and performances.

To address this, we proposed a theoretical framework to model the NMP ecosystem, which includes all factors affecting a musicians' experience, when she/he participates in NMP sessions, and narrowing them down to the most important ones. A software for real-time ultra low delay multi-directional audio streaming was developed for experimental purposes, allowing us to modify audio quality at will. It was combined with an ultra low delay experimental setup, using audio delay boxes to modify end-to-end delay. We then designed controlled experiments where audio delay and audio quality were manipulated.

Based on this background, a pilot study was conducted in order to evaluate the experimental procedure and the software and hardware employed. Afterwards, a large study with real musicians was completed, with audio and video being recorded from every session. Subjective evaluation was applied using electronic questionnaires that musicians could easily answer. The questionnaires were designed in order to subjectively evaluate multiple QoME variables.

In addition to the questionnaires, all audio recordings were analyzed to extract tempo variations and signal analysis was performed to extract audio characteristics and correlate them with the subjective results. Finally, the videos of the participants were analyzed with suitable software in order to analyze their emotions so as to assess their experience. The results of these objective methods were combined with the subjective results to strengthen our conclusions.

1.5 Contributions of the thesis

This thesis proposes *a theoretical framework as a tool to investigate the QoME ecosystem*. This framework includes parameters that affect musicians' experience and performance. In a cross-disciplinary domain, this framework should be considered as a guiding tool for designing experiments aiming to search for relations between multiple factors.

Guided by the framework, we designed and implemented extensive experiments with multiple participants to assess the impact of audio delay and quality to NMP. Despite the fact that many studies report specific ranges of audio delays under which interactive performance with hand claps is successful, other studies using musical instruments reported a wider range of delays where interactive performance was feasible, albeit with a few subjects. In our main study, 22 musicians played various instruments in different musical styles; *this was the largest NMP study that we are aware of*. We then calculated objectively and subjectively the effect of a wide range of delayed values. *The subjective results show that interactive music performance is feasible for longer delays than with hand claps, that is, up to 40 ms*. We also performed a tempo analysis of the audio recordings, in order to assess whether musicians could hold a steady tempo under different delay values. *The objective analysis of the audio recordings further verifies that music performance is feasible with one way delays of up to 40 ms.*.

Another significant contribution of our work is the role of audio quality in networked musical experience. Although many studies on the field of music technology assume that audio should be always perfect in music applications, we explore the role of audio quality in the field of interactive musical experience. To accomplish that, we experimented with audio quality variations and asked the participants to evaluate the perceived audio quality. We found that *the reduction in audio quality up to a certain minimum, which leads to reduced bitrate requirements, did not significantly affect the ability of musicians to synchronize.*

In previous studies on the impact of delay to NMP, the performances were analyzed in order to evaluate the effects in the performance's tempo; *our analysis showed that tempo drops as delay is increased in real musical performances over NMP, in the same manner as already observed with hand claps*. However, this analysis also confirmed that *musicians can synchronize with delays of up to 40 ms, in the sense that they can find and maintain a steady tempo*, thus confirming the subjective analysis.

But how is the delay perceived by the musicians when performing? Can they distinguish between small delay variations in the range of some tens of milliseconds? Is this perception dependent on the musical instrument, or the music tempo and the music genre? We found that *perception of audio delay turns to be related to the musicians' experience, the instrument's audio features like audio spectrum, audio envelope, as well as to the music tempo, music genre and other music features*, as found by both the subjective analysis and its correlation with the audio features in the musical performances. On the other hand, the analysis of the questionnaires on the impact of visual contact between the musicians in NMP showed that *musicians focus on the audio rather than the visual contact during their performance*.

Finally, we analyzed the video recordings of the NMP sessions with machine learning toolkits, which allow the extraction of facial features from the videos, and in turn the derivation of the emotions felt by the musicians during their performance under different delay and quality settings. While the videos of the performances were not captured with the intention of performing emotion analysis, hence they are not ideal for facial recognition, we found that *emotion analysis based on video processing with machine learning tools reveals discontinuities in the felt emotions, even though the results are not clear enough to make subjective analysis redundant*.

1.6 Outline of the rest of this thesis

The next part of this thesis, Chapter 2, presents a literature review, discussing past work in creating NMP systems, assessing the QoME of NMP against delay, recognizing emotions and previous QoME frameworks. Then, in Chapter 3 we discuss the architecture of NMP systems and identify the sources of delay in NMP, as well as providing background on delay measurements, audio quality and audio characteristics.

In Chapter 4 we present our QoME framework and identify the parameters that we will be controlling in the experiments to follow. Chapter 5 describes the software we created in order to control the parameters of our experiments, as well as a small validation study that we performed and its results. Then, Chapter 6 presents our pilot study with a few musicians, which focused on testing an extended questionnaire and improving our experimental setup.

Chapter 7 describes our main experimental study and discusses the subjective results from the questionnaires and their statistical significance. Then, in Chapter 8 we explain how we recovered the tempo from the audio recordings and the results from its analysis. We then go over the analysis of the audio characteristics of the recordings against the controlled variables in Chapter 9. We finish the evaluation with the emotion analysis of the recorded videos in Chapter 10.

We close the thesis with our conclusions, which combine the results from all the evaluation modes used in the thesis, and some ideas for future work in the field in Chapter 11.

Chapter 2

Literature Review

In this chapter, we provide an overview of the state of the art in Network Music Performance (NMP), by first presenting some of the main systems that have been developed in the past 20 years, and then by classifying a large number of existing NMP systems along various dimensions. We then discuss previous work on the impact of audio delay, splitting them into experiments with hand claps and experiments with musicians, and also considering the impact of audio quality to NMP. We then present research on emotion recognition, as we also use this technique to analyze the results of our experiments. Finally, we summarize the assessment methods used in previous work, and present an existing framework for QoME evaluation.

2.1 NMP Evolution and State of the Art

A review of the state of the art in NMP circa 2016 can be found in [72]. NMP can be considered to start with John Cage in 1939 and his *Imaginary Landscape No. 4 for Twelve radios*. Twelve performers on a theater stage held a portable radio receiver each and tuned it to different frequencies, receiving twelve different broadcasts, while manipulating the volume level. This performance is considered to be the first remote performance, since the audio material was broadcast and the performers manipulated it.

The *League of Automatic Music Composers* was the first project using networked computers for music. A group of electronic music experimentalists formed a band in San Francisco and connected multiple computers and electronic circuits over a network in order to perform music. The group was later named *The Hub* and experimented with interconnection between the east and west coast of US, sending only messages rather than pure sound, due to the limited bandwidth available at that time.

In the early 2000s, Chris Chafe led the *The SoundWire* project at the Center for Computer Research in Music and Acoustics (CCRMA) at Stanford University [20]. In Soundwire the Internet2 backbone was used to transmit high quality uncompressed audio over large distances in US, using tools written in C++. The group developed and

still supports JackTrip[1], a multi-platform software for uncompressed, ultra low delay, bidirectional audio streaming.

Starting in the mid 2000s, Alexander Carôt created Soundjack [14, 89] at the International School for New Media (ISNM) in the University of Lübeck, Germany, moving later to Anhalt University of Applied Sciences, Germany, where he led the FAST project. Soundjack was inspired by the SoundWire project, and began as a standalone application, later moving to a browser based application[2]. It uses UDP/IP and audio compression with the Opus codec[3], allowing multiple users to perform music together in real-time, but in limited distances.

In the early to mid 2010s, the Greek DIAMOUSES project[4] developed an integrated collaboration environment for NMP, focusing on the creation of a community toolkit that allowed sharing artefacts between the participants, such as musical scores [3, 4]. Part of that team continued with the MusiNet project[5] which worked on developing open source and ultra low delay software for audio and video communication, for both clients and servers [1].

In the mid to late 2010s, the EU funded the LOLA project [26], for *LOw Latency audiovisual streaming system*. The LOLA project employed special (and expensive) hardware and software to support NMP; the software is provided for free to researchers and under a shareware model to the public.[6] It supports ultra low delay uncompressed audio and video streaming over (at least) Fast Ethernet connections.

2.2 NMP Systems

We will now present a more detailed taxonomy of existing NMP systems. A summary of these systems is given in Table 2.1. The proposed systems can be categorized based on various criteria, as discussed in the following subsections.

2.2.1 Network Architecture

Two different architectures have been considered for NMP, peer to peer [81, 1, 89, 26, 11, 3, 32, 20] where the endpoints communicated directly with each other, and client-server [74, 65, 81, 46, 90, 1, 43, 11, 3] where the server receives and retransmits the audio streams from/to the endpoints. As these studies report, the peer to peer architecture reduces the delay of audio transmission since in the client - server topology

[1] https://ccrma.stanford.edu/groups/soundwire/software/
[2] https://www.soundjack.eu/
[3] https://opus-codec.org/
[4] https://istl.hmu.gr/activities/research-projects/diamouses/
[5] http://musinet.aueb.gr/
[6] https://conts.it/art/lola-project/old-lola-project-web-site/lola-low-latency-audio-visual-streaming-system

Table 2.1: Network Music Performance Platforms

Authors	Name	Architecture	Range	Protocols	Data	Chnls	Sync	Codec
Saputra et al. [74]	BeatME	Client-Server	LAN WLAN	UDP OSC	MIDI	16 in 1 out	none	raw
Kurtisi et al. [38, 53]	-	Client- Server	LAN	RTP UDP (stream) TCP (data)	audio	N.A.	NTP	ADPCM FLAC (real time) MP3, MPEG4 (on demand)
Renwik et al. [65]	Sourcenode	Client-Server	LAN	UDP	MIDI	N.A.	none	raw
Stais et al. [81]	-	Client-Server or P2P	WAN	N.A.	audio	2	NTP	raw
Kapur et al. [46]	Gigapopr	Client-Server	WAN	UDP	audio video MIDI	N.A.	N.A.	raw
Wozniewski [90]	Audioscape	Client-Server	WLAN	N.A.	audio	1 in, 2 out	GPS	raw
Chew et al. [21, 53, 23, 75, 92]	-	Client-Server	WAN	RTP RTSP UDP	audio video MIDI	16	GPS CDMA	MPEG1-4
Alexandraki et al. [1, 2]	Musinet	Client-Server or P2P	WAN	SIP RTP HTTP	audio	any	none	Opus
Cârot et al. [14]	Soundjack	P2P	WAN	UDP	audio video	8	external	ULD Opus JPEG video
Drioli et al. [26]	LOLA	P2P	WAN	TCP (ctrl) UDP (data)	audio	8	N.A.	raw
Lazzaro et al. [56]	-	Client-Server (Control) P2P (media)	WAN WLAN	RTP/RTCP UDP (data) SIP	MIDI	16	RTP/RTCP sync tool	MPEG4
El-Shimy et al. [80]	-	P2P	LAN		audio video	N.A.	N.A.	
Fischer et al. [31]	Jamulus	Client-Server	WAN	UDP	audio	2	none	Opus
Caceres et al. [12, 11]	JackTrip	Client-Server or P2P	WAN	UDP	audio	any	software based audio resampling	raw
Akoumianakis et al. [3, 4]	Diamouses	Client-Server or P2P	WAN	RTP TCP/UDP	audio video	any	internal metronome	raw audio MPEG video
Gabrielli et al. [32]	WeMust	P2P	LAN WLAN	TCP UDP	audio MIDI	12	software based audio resampling	raw CELT
Kauer et al. [47]	Jamberry	P2P	WAN	UDP	audio	2	external master clock	Opus
Chafe et al. [20]	StreamBD	P2P	WLAN	UDP,TCP	audio	any	none	raw

audio streams have to pass through the NMP server; however, the client - server architecture can more easily support large numbers of participants, as there is no need for each client to setup connections to each other client, given the lack of support on the Internet for wide area multicast.

2.2.2 Network Range

Some frameworks operate over LANs, usually a Fast Ethernet switched LAN, while others support WAN features. There are also studies where the framework operated over Wireless LANs, which are similar to LANs in delay, although they suffer from higher jitter and some loss.

2.2.3 Network Protocols

A simple way to produce an audio stream is by using the RTP, UDP, IP and Ethernet protocols. RTP allows the application to put the received audio packets in the right order and UDP is the *fast but unreliable* transport protocol most suitable for NMP. In many frameworks, additional mechanisms were developed to manipulate control signals and other types of communications, often using TCP. The SIP protocol was used for session initialization and management in [3].

2.2.4 Data type

Besides actual audio (whether compressed or uncompressed), MIDI was used in some studies. MIDI data streams are much lighter in terms of size than audio and can be transferred faster than audio, albeit at the cost of only allowing a symbolic representation of music to be exchanged. MIDI is however useful for driving remote instruments, by using a MIDI controller (a keyboard or similar) at one end, and a sound producing device at the other end.

2.2.5 Visual Contact

In many studies, visual contact through video streaming was used. The usefulness of visual contact in NMP has not been deeply investigated, however. Many studies report that the human brain focuses better with aural rather than visual stimuli. These studies investigate the sensorimotor synchronization, which refers to the coordination of movement with an externally presented rhythmic stimulus. During sensorimotor synchronization with unimodal stimuli (e.g., auditory or visual), performance is generally better when the stimuli consist of auditory beeps rather than flashes of light [67, 68, 66]. Studies assessing the influence of multisensory stimuli on sensorimotor synchronization ability generally indicate that synchronization performance is enhanced

(i.e., variability is reduced) when bimodal stimuli are presented, compared with performance when presented with either of the constituent unimodal stimuli.

2.2.6 Multi stream synchronization

A very important issue in NMP is stream synchronization. Considering that multiple audio streams arrive at different times at an end point because they were possibly delayed, the streaming application is responsible for synchronizing and mixing them in a peer to peer system. The server is responsible for the same job in a client - server system. Synchronization mechanisms can be structured in multiple ways, for example, by using timestamps and synchronized clocks. The Network Time Protocol was used in some cases as shown in Table 2.1.

2.2.7 Audio format

Regarding the audio format, raw (PCM) audio is used in [74, 65, 81, 46, 90, 89, 26, 11, 3, 20] where LAN based scenarios were considered, thus providing ample bandwidth at small distances. In cases of WANs, the lack of bandwidth (at least for older studies) forced the use of compression using the *Ultra Low Delay* (ULD) mode of the Opus codec[7] [2, 89, 43, 47, 32].

2.3 Impact of Audio Delay to NMP

Studies on synchronization during human interaction have long concluded that delay is a critical factor for synchronization to occur; for NMP many studies have indicated that human tolerance to delay is far lower than that for teleconferencing, with participants reducing their tempo to compensate for higher delays.

To examine these effects, many studies used performers trying to synchronize hand claps. Hand claps have a very short and simple audio envelope, so it easy to record and analyze tempo variations, even by visual observation of the recorded waveforms. Some studies have used musical instruments, but these were harder to characterize. We summarize both types of study in Table 2.2 and discuss them in more detail in the following subsections.

2.3.1 Studies using hand claps

Schuett [77] investigates the effect of delay in tempo proposing and evaluating the *Ensemble Performance Threshold* (EPT), which is the amount of the one way delay above which clapping performers cannot synchronize. Two performers participated in this

[7] https://opus-codec.org/

Table 2.2: NMP Studies with Delay Tolerance Tests

Authors	Hand claps	Instruments	Rhythm Pattern	BPM range	M2E delay	Musical piece	Musical features	Quality metrics
Chafe Gurevich et al. [40, 19]	✓		✓	60 - 120	1 - 77			BPM trend BPM slope
Farner et al. [30]	✓		✓	86 - 94	6-67		Reverberation, subjects' musical training	BPM trend, slope, initial value, imprecision asymmetry Subjective Rating
Driessen et al. [25]	✓		✓	90	30 - 90			steady state BPM time difference subjective rating
Bartlette et al. [8]		violin, cello, flute, clarinet		NA	6 - 206	Twelve Duets k.487, No.2 and No. 5 (W.A. Mozart)	Prior practice	Mean Pacing Mean Regularity Mean Asymmetry Subjective Rating
Carôt et al.[16]		percus, bass, sax	✓	60 - 160	1 - 75	"The days of wine and roses" (Mancini, Mercer)	Reverberation, self - delay	Subjective rating
Barbosa et al.[7]		violin, cello	✓	80	25 - 120		attack time	BPM trend, BPM dynamic time warping analysis
Olmos et al. [61]		piano, singers conductor			15 - 135 (audio), 60 - 180 (video)	"IL core di vono.." (W.A. Mozart) "Ah! Voi signor" (G. Verdi) "Bess you are my woman" (G. Gershwin)		Subjective rating, galvanic skin response, skin conductance response, BPM dynamic time warping analysis, BPM curvature points
Chew et al. [22, 21]		piano		46 - 160	0 - 150	Sonata for Piano Four Hands (F. Poulenc)	prior practice	Subjective rating BPM segmental analysis
Rottondi et al. [71]		piano guitar clarinet violin		80 - 132	15 - 75	"Bolero" (M. Ravel), "Master blaster" (S. Wonder) "Yellow submarine" (The Beatles)	rhythmic complexity spectrum statistical moments musical part	BPM trend, BPM slope, subjective rating
Kobayashi et al. [50]		MIDI piano			0 - 100	demonstrative monodic music		onset global and local phase difference

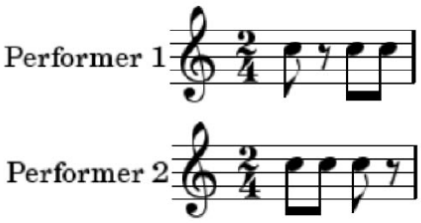

Figure 2.1: Rhythm pattern used in many hand clapping studies.

experiment, using the rhythmic pattern shown in Figure 2.1. They were informed of the amount of delay as it was increased, until the experiment was stopped at 100 ms. They participated in five scenarios with starting tempo and delay variations. The main findings of this research were:

1. If the delay was greater than 30 ms, the tempo would begin to slow down. This threshold was the EPT for impulsive music.
2. A strategy of leader - follower was used by the performers to maintain a steady tempo when the one way delay was between 50 and 70 ms.
3. EPT varies depending on the type of music (speed, style, attack times of instruments, etc).
4. When delay is between 10-20 ms, it may be providing a stabilizing effect on the tempo. A delay of 10-20 ms may be better for ensemble performance than 0 ms of delay.

Gurevich et al. [18][40] used seventeen pairs (34 performers) in clapping sessions under varied time delays. Each duo performed twelve trials. The subjects were located in two acoustically-isolated rooms. The authors reported that for delays shorter than 11.5 ms, 74% of the performances sped up. At delays of 14 ms and above, 85% slowed down. No correlation with the starting tempo was found in the range sampled.

Driessen et al. [25] experimented with two musicians who performed a clapping session together, with varying delays. The musicians were asked to follow a metronome that was set at 90 *Beats per Minute* (BPM) and to clap in rhythm for at least 60 seconds. The musicians also answered a subjective questionnaire about their experience after each session. The session had seven trials with total delays between 30 ms and 90 ms, in 10 ms delay increments and in a random order. The authors reported that the tempo of two musicians performing together at a distance, with network delay and without any external tempo reference, will slow down as the delay is increased. They also reported that the amount by which the tempo decreases is approximated by just over half (0.58) of the tempo times the delay in seconds, so that a tempo of 90 BPM with a delay of 60 ms will slow down to about 87 BPM.

Farner et al. [30] asked eleven pairs (22 subjects) to clap together for at least seven measures of a simple complementary rhythmic pattern. They used delays from 6 to

68 ms (corresponding to distances of 2 to 23 m). The tempo was found to decrease more rapidly for higher delays, and the relation was approximately linear. In addition, the tempo tended to increase for the shortest delay. The subjective evaluation showed that participants evaluated as good the trials where delay was short. Above 25 ms, the tempo variations increased, so this value was considered to be the delay tolerance threshold that many other studies also reported.

Chafe et al. [17] examined performances by twenty-four pairs (48 performers) of clappers under different delay conditions. The subjects performed a clapping rhythm from separate sound-isolated rooms, via headphones and without visual contact. One-way time delays between pairs were manipulated electronically in the range of 3 to 78 ms. The goal was to quantify the envelope of time delay within which two individuals produce synchronous performances. The authors report that for delays between 10 and 25 ms performance was natural. For delays lower than 10 ms, tempo accelerated while for delays over 25 ms delay decelerated.

2.3.2 Studies using musical instruments

According to the previous section, the delay threshold for NMP was 25–30 ms for hand claps following a certain rhythmic pattern. Furthermore, multiple strategies were employed by the subjects to cope with delay, such as slowing down the tempo. Musical instruments, however, have quite different audio envelope characteristics, compared to hand claps. In addition, musical performances are not as simple as hand clapping sessions, therefore it is conceivable that their delay tolerance may be different.

Barbosa et al. [6] investigated the self delay feedback effect. The self delay feedback is the delay a musician experiences if she/he listens to her/his sound delayed. This is a problematic situation and prevents musicians from synchronising. The author asked four musicians to play bass, percussion, piano and guitar. In a studio setup, musicians would listen to the feed-back from their own instruments trough headphones with delay. Their performance was synchronized with a metronome over several takes with different tempi. For each take, the feed-back delay was increased until the musician was not able to keep up a synchronous performance. The authors report that regardless of the instrumental skills or the music instrument, all musicians were able to tolerate more feed-back delay for slower tempos. They conclude that tempo and latency have a reverse relationship. No subjective evaluation was applied.

Barbosa et al. [7], investigates the *Perceptual Attack Time* (PAT), that is, how the attack time of notes affects tempi. He used two musicians performing cello and violin and analyzed the recordings. The delay introduced varied from 0 to 180 ms. He used the rhythm pattern shown in Figure 2.1 but with actual notes and a starting metronome set to 80 BPM. He conducted two experiments, one with slow attack from the musicians, and another with sharp attack. The analysis of the audio files show that tempo is generally higher in the sharp attack experiment than the slow attack. In both cases

it decreases with latency and starts at about 75 BPM (lower than the 80 BPM of the starting metronome). No subjective evaluation was applied.

Bartlette et al. [8] asked two pairs of musicians (4 musicians) to perform two Mozart duets while isolated visually and connected through microphones and headphones. Two clarinets were in one pair, and two stringed instruments (violin and viola) were in the other pair. Different levels of one way latency (0, 20, 40, 50, 80, 100, 120, 150, and 200 ms) were introduced into the performing environment (musicians heard themselves in real time and only the other part was heard delayed); the musicians performed the duets under these conditions and rated their musicality and level of interactivity. The author introduced four aspects of expression in an interactive performance: *pacing*, *regularity* (within parts), *coordination* (between parts) as objective, and *musicality* as subjectively evaluated by the subjects. *Pacing* denotes the tempo of a musical performance. *Regularity* denotes timing within parts, which may be characterized by quasi-isochrony, or nearly metronomic note timing. *Coordination* denotes timing between parts, thus mean asynchrony. Finally, *musicality* is measured by the participants' responses, with higher ratings given for more musical and interactive performances. Although the musicians chose different strategies to handle the latency, both duets were strongly affected by latency at and above 100 ms. At these levels of delay, the musicians rated the performances as neither musical nor interactive, and they reported that they played as individuals and listened less and less to one another.

Chew et al. [21][22] asked two pianists to perform Pulenc's sonata for two pianos. This sonata has three movements (parts) which should be played in three different tempi of 46, 132, 160 BPM. Audio delay was inserted in the range of 0 – 150 ms. The musicians were placed in the same room having visual contact, and they heard each other's sound delayed. After each repetition, they answered three questions regarding the ease of playing, the ease of creating musical interpretation and the ease to adapt in the condition. Authors report that in the fast tempo part (Prelude), of 132 BPM, the participants had problem in synchronization above 150 ms. Both players agreed that adaptation was possible below 50 ms. In the slow tempo part of 46 BPM (Rustique), synchronization was possible under 75 ms of delay, as both players agreed. In the very fast movement of 160 BPM (Final), difficulties appear even in the 10 ms audio delay. Players mentioned that they could overcome delay issues under 50 ms by practicing.

Cârot et al. [16], asked five professional drummers to perform (one at a time) with a single professional bass player as the rhythmical counterpart. This way a direct comparison of each rhythm section constellation is possible. The one way delay was in the range of 0 – 70 ms. The experiments were performed at speeds of 60, 100, 120 and 160 BPM and the delay between the two players was increased from 0 ms in steps of 5 ms until one of the player felt uncomfortable, when the players tended to slow down. Subjective evaluation was applied, and the players had to evaluate the actual delay situation as "excellent", "tolerable" and "not tolerable". The author reports that overall delay thresholds ranges between 0 and 65 ms. He also reports that the most important

observation is that the players do not exhibit a common latency acceptance value.

Olmos et al. [61] worked with six singers, one conductor and one pianist in order to simulate an orchestra placement. The singers were divided into three groups, each of which performed one of the following pieces: "Il core vi dono...", from Mozart's Cosi fan tutte (mezzosoprano and baritone voices); "Ah! – Voi signor" from Verdi's La Traviata (soprano, tenor and bass-baritone voices); and "Bess you are my woman" from Gershwin's Porgy and Bess (soprano and bass baritone voices). The music pieces were selected for their varying rhythm complexity. Six different combinations of perceived audio and video delays were selected in order to simulate the various latency conditions between Montreal, New York, San Francisco and Tromsø. Each isolated room contained two speakers, two cameras and two CRT monitors, with each monitor/camera/speaker set representing the audio and video from a different location. The singers were able to see and hear each other through the video monitors and speakers at all times. After each performance, the singers were asked to complete a questionnaire, rating their experience. Ratings for the following questions were obtained on a Likert scale of 1-7:

1. How satisfied were you with the performance?
2. How would you rate your emotional connection with the remote singer?
3. How would you rate your emotional connection with the conductor?
4. How important was the audio?
5. How important was the video?

The authors report that the singers managed to cope with delay under all conditions. They also report a feeling of "disconnect" felt by the singers between what they heard and the events to which they reacted. An important observation was that the conductor turned out to be very important for synchronization. The authors also report a finding that is the inverse of what was found in other studies: as delay increased, the tempo increased too. A possible explanation for this was the role of the conductor.

Rottondi et al. [71], asked eight musicians to participate in NMP experiments. The musicians had at least eight years of musical experience and were grouped in seven pairs; some performed in more than one pair (e.g., one clarinetist performed twice). The instruments the participants played were acoustic, classical and electric guitar, electric piano, keyboard(strings), clarinet and drums. Each repetition was characterized by different tempo and network settings in terms of reference BPM, network latency, and jitter. After each session, the participants evaluated two subjective parameters, the quality of their interactive performance in a five-valued range (1="very poor", 5="very good") and the perceived network delay, within a four-valued range (1="intolerable", 4="none"). In case the players spontaneously aborted their performance within the first 50 seconds, perceived delay was set to 1 and perceived quality was set to 0 by default. The two ratings are considered as subjective quality parameters. The authors

applied audio recording analysis to evaluate six audio parameters: spectral entropy, spectral flatness, spectral spread, spectral centroid, spectral skewness and spectral kurtosis. The authors report that the noisiness of the instrument, which is captured by spectral entropy, flatness and spread, has a relevant impact on the perceived delay. They also report that perceived delay is strongly affected by the timbral and rhythmic characteristics of the combination of instruments and parts. Finally, they report 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. larger than 75 ms) do not prevent networked musical interaction, but they limit the selection of the instrument/part combinations. The authors concluded that the quality of the musical experience is not only a function of the total delay, but it also depends on many other factors like the audio characteristics of the instruments, the role of the musician, the music genre etc. Our larger study, presented later in this thesis, confirms and extends these observations.

Delle Monache et al. [24] asked ten musicians (five duos, five males, five females, age ranging from 14 to 29) to perform an exercise, with each duo repeating it under six different conditions of emulated network delay (28, 33, 50, 67, 80, 134 ms). The sequence of delays was randomized for each duo. The musicians performed mandolin, accordion, guitar, percussion, harp, flute and alto sax. The setup included audio contact via microphones and loudspeakers and visual contact via cameras and large video monitors. The participants were asked to fill in a 5-item questionnaire after each single repetition, and a general 27-item questionnaire at the end of each session. Further comments were collected at the end of the test. The 5-item questionnaire was as follows:

1. The sense of playing in the remote environment was compelling.
2. The delay affected the sense of involvement.
3. I was able to anticipate the musical outcome in response to my performance in the remote environment.
4. How much delay did you experience between your actions and expected outcomes?
5. The delay affected the quality of my performance.

The answers regarding *The sense of playing in the remote environment was compelling* and the *The delay affected the sense of involvement* revealed a negative effect of delay to musicians' involvement in the environment. Another observation was that for higher delays, musicians could not understand who or which was responsible for playing out of time. Finally, the authors found that the musicians did not focus on the TV monitor, but focused on the audio signal. As mentioned above, the visual contact in NMP has not been extensively examined until today; this is one of the few studies to examine whether musicians rely more on audio or video.

2.4 Impact of Audio Quality to NMP

A variety of objective methods for evaluating audio quality is reviewed and examined in [83]. Metrics such as *Basic Audio Quality* (BAQ), *Perceptual Evaluation of Audio Quality* (PEAQ) and others [41] [84] [63] [42] [88] [9] are widely used for audio quality evaluation. Most of these methods rely on test subjects comparing a reference audio passage, considered as perfect, with the examined audio passage, and evaluating the quality of the examined passage in a Likert scale from 1 to 5. For specific applications (e.g., voice telephony) there are also methods of deriving such metrics based on an objective analysis of the audio passages, that is, converting the QoS to QoE.

There are no studies that we are aware of considering the effects of audio quality on NMP; most NMP assessment studies focus on delay, choosing an audio quality level appropriate for the network setup. Since in NMP systems and studies compression is avoided to prevent inflated delays, we varied the audio sampling rate as a means of assessing the effects of audio quality. While this approach has a direct effect on the transmission bitrate and does not introduce any delays, rather than changing the quality across the audible spectrum, it basically cuts off the frequencies above the Nyquist limit (half the sampling rate).

As mentioned above, some studies correlate the spectral characteristics of audio recordings made during NMP experiments with the QoME of the musicians [71]; we have also performed such an analysis on our experiments (see Chapter 9). These studies however are not concerned with the effects of audio quality on NMP, but on the effects of specific audio features on the perception of delay.

2.5 Emotion Recognition

A novel part of our work is the application of emotion recognition methods to the videos recorded during NMP sessions in order to assess the tolerance of musicians to delay and quality variations. An extended review of studies related to emotion recognition through various sensors can be found in [64]. A person's emotional state may change depending on their subjective experience [39] and 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, 28].

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 [45]. 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 [61].

In [35] 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 [34] employ Emotional Expression in music performance as an instrument to communicate emotions to listeners. In [78], 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 [29] 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 [76] 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 [33] note that subjective strategies like rating sheets measure the subjective perception of *expressed* emotion rather than *felt* emotion. Our work in automated emotion detection, reported later in this thesis, aims to bridge this gap between the expressed and felt emotion, by correlating the responses of the participants to the questionnaires with the emotions revealed by their facial expressions during each performance.

2.6 Assessing QoME

2.6.1 Assessment methods

In most of the studies assessing the satisfaction of the musicians, the audio delay was manipulated in a range between a few ms to 120 ms (one way). The mechanisms for delay manipulation included tools for network emulation like netem.[8] Furthermore, the starting tempo was also a variable in some studies. To set the starting tempo, metronomes were used.

The main method of evaluation was via subjective variables, that is, by having participants respond to questionnaires rating their experience with the NMP sessions. In a few studies, audio recordings were also analyzed to extract tempo variations; the so-called BPM trend or slope, showing how tempo slows down as delay is increased. In [71], audio spectral features were extracted using suitable tools from the recordings.

In addition to deploying all these methods (questionnaires, tempo analysis and audio feature extraction), this dissertation adds another analysis method, emotion analysis based on recorded videos of the musicians participating in the sessions. This method uses machine learning tools for facial feature extraction, which are then processed to detect the emotions felt by the participants. In addition, our work is the first to consider the effects of audio quality to NMP; this supplements the delay analysis, since

[8] https://wiki.linuxfoundation.org/networking/netem

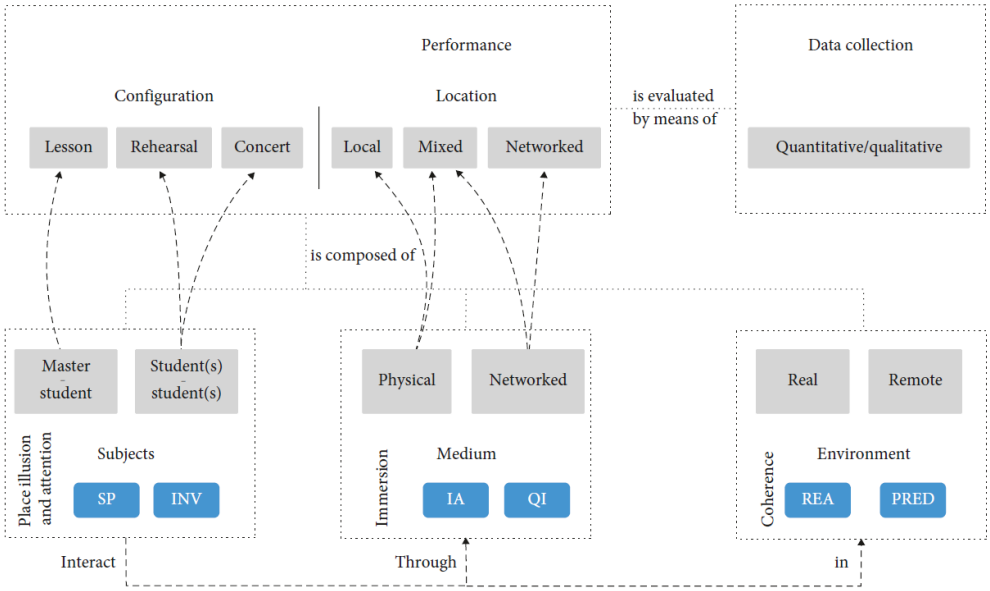

Figure 2.2: A conceptual framework for NMP [37].

lowering the quality can save bandwidth, thus allowing NMP to take place in limited bandwidth environments without introducing coding and its resulting delay.

2.6.2 Quality of Experience Frameworks

The only *Quality of Experience* (QoE) framework designed specifically for NMP that we are aware of, was proposed by Delle Monache et al. [37]; it is outlined in Figure 2.2. The framework is intended for chamber music practice and learning scenarios. This framework uses **Performance** as the basic component. The performance is characterized by the **Configuration** feature that includes *lesson, rehearsal or concert*, and the **Location** feature that can be *local, mixed or networked*. The performance can be *evaluated by means of* (connection between components) the **Data collection** component, either *quantitatively or qualitatively*. Additionally, the performance *is composed of* three major components, thus the **Subjects**, the **Medium** and the **Environment**. The **Subjects** component includes the *Master-Student* or *Student(s)-Student(s)* constructs. The **Medium** component can be *Physical* or *Networked*. The **Environment** component can be *Real* or *Remote*. These components are connected as follows: the **Subjects**, thus the master and the students, can *interact through* (connection between components) the **Medium** (thus physically or networked), or can *interact in* (connection between components) the **Environment** (thus real or remote). Other connections of the framework suggest that **subjects**, thus the master and the students, can participate in a *lesson* as a master-student session, or in a *rehearsal* and a *concert* as student-student.

According to the authors, the presence experience is composed by three major groups of constructs, that is, components: (i) the spatial-constructive and attention

facets of the experience of being there are, respectively, operationalized into spatial presence and involvement components, (ii) the coherence of the scenario, that is, the reasonableness of the events primed to the user, is reflected into the perceived realness and predictability components, (iii) the quality of the system's technology is distilled into the two components of the interface awareness, that is, distraction factors and the quality of immersion.

In more detail, for the first component of the experience, *Spatial presence* (SP) refers to the emerging relation between the mediated environment as a space and the user's own body. The sense of "being in place" is related to the role of the active body in constructing a spatial-functional model of the surrounding environment. *Involvement* (INV) or flow is a recurring concept in the presence literature and retains the attention side of the presence experience.

For the second component of the experience, *Perceived realness* (REA) encompasses reality judgments with respect to the meaningfulness and coherence of the scenario, as a function of the system's ability to provide stimuli which are internally consistent. *Predictability* (PRED) refers to the possibility to anticipate what will happen next, in terms of activation of motor representations as a consequence of perceiving while playing.

Finally, for the third component, *Interface awareness* (IA) takes into account distraction factors, that is, the obtrusion of control and display devices in terms of interference in the task performance. *Quality of immersion* (QI), is a component related to the presentation of the stimuli and can be defined as the set of valid actions supported by the mediating environment.

Chapter 3

NMP and Audio

In this chapter we discuss how audio is captured, processed, transmitted and played back in NMP systems. Our goal is to explain the sources of audio delay and audio quality degradation in the context of NMP, since these factors are central to our work. We also describe two additional characteristics of audio, the performance tempo and the audio envelope, as these are explored in later sections.

3.1 The NMP audio path

As audio delay is the basic problem in NMP, it is important to examine the sources of audio delay in the NMP path. A detailed analysis of the audio signal path in NMP is reported in [15] and shown in Figure 3.1. As shown in the figure, sound is generated in the source, passes through a number of blocks and arrives at the ear and the brain of the listener. The darker the color of the block, the bigger the amount of delay it introduces. The path includes:

1. Sound propagates from the source to the microphone.
2. The acoustic wave is transduced to an electrical signal in the microphone.
3. The signal is transmitted through the microphone connector.
4. The signal is filtered to avoid aliasing, and then it is sampled and quantized.
5. Filtering, including audio compression, may take place here. Alternatively, it may take place after the following step, just before packetizing the data.
6. The sampled data are stored in the buffer of the sound card until they are collected by a program.
7. The packetization stage where UDP, IP and Ethernet packets are constructed from the samples.
8. Network propagation, transmission and routing procedure.
9. Depacketization in the receiver's side.

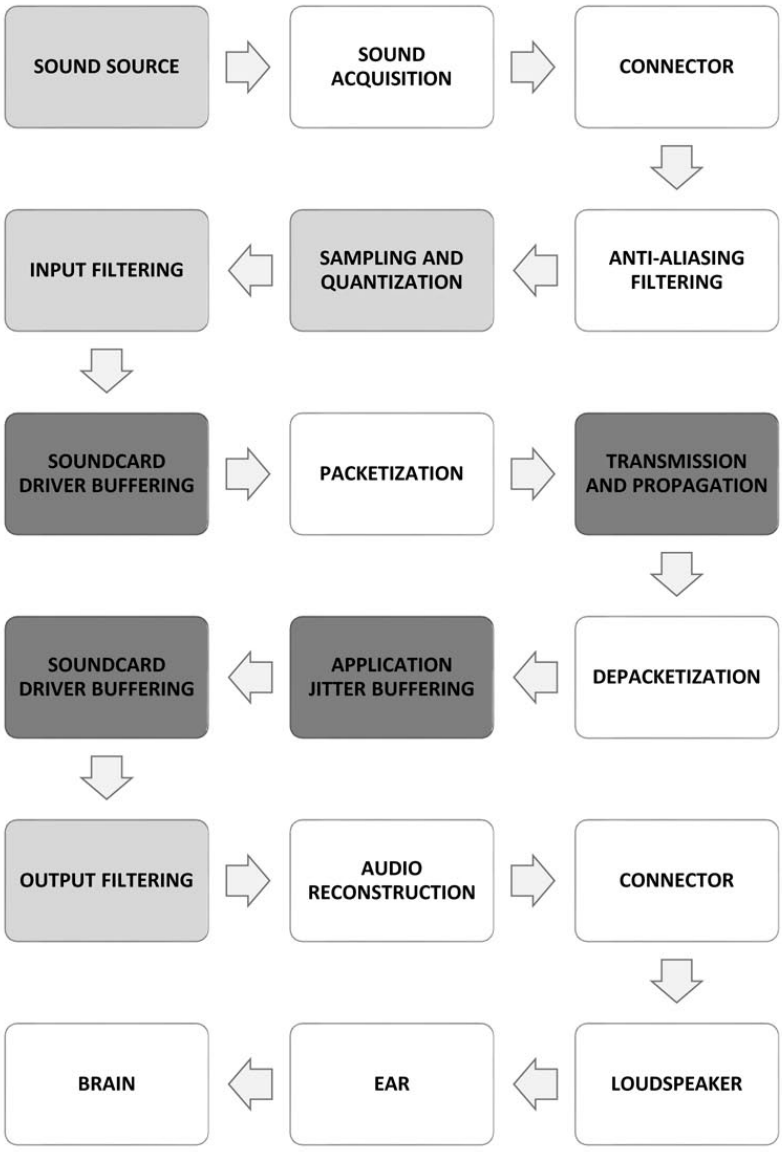

Figure 3.1: Audio delay sources in NMP [15].

10. The data are possibly held in a dejittering buffer, and then await processing at the buffer of the audio card.
11. Filtering, including audio decompression, may take place here.
12. Digital to analog conversion in the receiver's side
13. Audio transmission through the headphones' connector.
14. The electric signal is transduced to an acoustic wave.
15. Sound propagation from the speaker to the ear.

The basic sources of audio delay will be explained in the following sections.

3.1.1 Sound propagation outside the system

Since sound travels with a speed of 343 m/sec in 20 degrees Celsius, for a one meter distance between two musicians, the propagation delay is 1/343 of a second, thus almost 3 ms; this grows to more than 23 ms for a distance of 8 meters, which is noticeable to the human brain, explaining the need for a conductor in a large orchestra. When audio is captured by a microphone, it is normally placed quite close to the source so as to capture the sound more accurately; as a side effect, this minimizes delay. At the receiving end, if the sound is heard by loudspeakers, the overall delay is increased due to the distance of the listener from the loudspeaker; in this case, it makes sense to use headphones to also minimize this delay.

3.1.2 Acoustic wave to electrical signal and vice versa

The transduction of the acoustic wave to electrical signal and the opposite procedure, which are taking place in the coils of the microphone and the headphones, are considered to be instantaneous.

3.1.3 Signal transmission from and to the sound card

The electric signal is fed to the sound card through the connectors and the audio cable. This procedure is considered to be instantaneous, as the electrical wave travels at almost the speed of light and the distances involved are quite short. The same thing happens with the sound from the output of the sound card to the headphones, through the connectors and the cables.

3.1.4 ADC/DAC and buffering

Analog to Digital (ADC) and Digital to Analog (DAC) conversions are the first noticeable source of audio delay in NMP. At the stage of the analog to digital conversion,

the analog signal will pass through an analog anti-aliasing filter. Then it will be sampled at the rate of 8 - 96 KHz and each sample will be quantized, resulting in 8 to 24 bits/sample. Since computers deal with audio not sample by sample but in blocks of samples, the buffering stage introduces a noticeable amount of delay. Thus, as shown in Figure 3.2 there is a driver input buffer which is fed by samples from the sound card input; conversely, there is an output buffer which sends samples to the sound card so to be converted to analog signal. The input driver needs to await for a number of samples to be gathered before passing them to the application; conversely, the application passes a number of samples to the output buffer for playback. This procedure introduces the so-called audio buffering delay. If the size of the buffer is P samples and the sampling rate is R samples/second, then the introduced delay is calculated as $D = P / R$ s. The larger the size of the buffer is, the more stable is the system, as fewer interrupts are generated, but the delay is also larger. For example for $P = 256$ *bytes* and $R = 44100$ *samples/sec* $D = 256 / 44100 = 5.8$ *ms* is the delay for the input buffer, thus in order to listen to the input audio directly a total delay of 11 ms is introduced. By doubling the buffer size, total delay is doubled to 22 ms which is quite noticeable, even without any other delays.

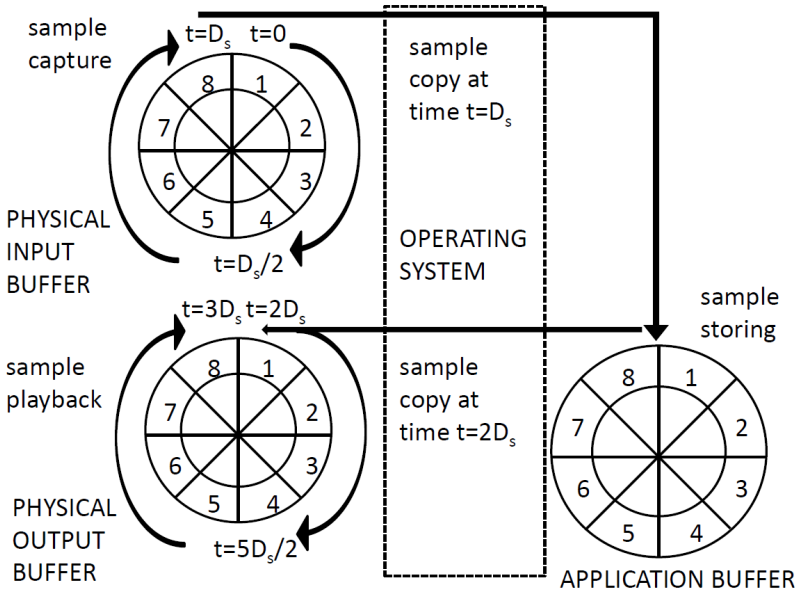

Figure 3.2: Circular audio buffer [15].

3.1.5 Compression and decompression

Audio compression is widely used in telecommunications to save bandwidth. Most compression algorithms for general audio use sub-band coding and psycho - perceptual models to reduce the number of bits a signal is represented with. The more compressed a signal is, the less amount of bandwidth it needs to be transmitted. On the other hand though, the algorithmic delay of the codec is higher for more compressed

signals, as larger buffers are used to exploit similarities in the signal. Both compression and decompression thus introduce additional delays. For this reason, the MP3 and AAC codecs commonly used for music suffer from excessive delays and are unsuitable for NMP. Some widely used low delay compression algorithms are Opus, especially in the ultra low delay mode [87, 86], ULD [51] and WavPack [54]. These algorithms can achieve delays as low as 4 – 8 ms.

3.1.6 Packetization and depacketization

In NMP applications, the streamed audio is either uncompressed, thus pure PCM audio, or compressed by a low delay codec. In both cases, the next step is to construct Ethernet frames using suitable protocols. The simplest way to produce an audio stream is to use the RTP, UDP and IP protocols to construct the Ethernet packets. Each of the four protocols encapsulate a header in the final packet. The Ethernet packet will travel through the Internet to arrive at the receiver. At the receiver side, depacketization will take place and the audio data will be extracted and driven to the sound card and the headphones or the loudspeakers. The procedures of packetization and de-packetization in both sides introduce a very small amount of delay, which is considered to be nearly zero, assuming that a single buffer of sampled (and, possibly, encoded) data is transmitted in each packet. In this case, the main element of delay is the time it takes to fill the sampling buffer.

3.1.7 Network transmission and queueing

The other major source of audio delay in NMP is the network delay, which consists of three parts:

- Propagation delay: The time required for a signal to propagate from one end of the circuit to the other. If we consider that a signal travels with almost the speed of light (0.7*c) then this delay is negligible for small distances, but noticeable for large distances.
- Transmission delay: In networking, this delay refers to the time required for all the packets' bits to be pushed on the wire. This delay depends on the speed of the channel, so it can be negligible for very fast networks, but significant for low speed networks.
- Queueing delay: The most uncertain part of the network delay is the queuing delay at the routers on the path. Due to the unpredictable nature of traffic on the Internet, a packet may have to wait in multiple transmission queues as it moves from router to router on the path to its destination.

3.2 Measuring Audio Delay

The NMP path, analyzed above, starts from the source and ends at the ear of the listener. If we consider that the source is (say) the mouth of a singer and the ear is the ear of a guitarist, this path can be named as *Mouth to Ear* (M2E) and the time required for the sound to travel through this path is the M2E delay.

In the case of NMP, M2E is mentioned in most of the studies, although in practice, it is usually estimated and not measured. Mouth to ear delay in this context refers to the time it takes from the moment the first musician will generate a sound until the moment that the second musician will hear this sound. In [16], it is mentioned as the *One Way Delay* (OWD) and in [72] as *Over-all One-way Source-to-Ear delay* (OOSE).

Unlike most NMP studies which use M2E delay, in our work we use the *My Mouth to My Ear* (MM2ME) delay. As shown in Figure 3.3, 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 endpoint, by simply reflecting the transmitted sound at the other endpoint; in contrast, M2E needs to be measured at both endpoints, thus requiring perfectly synchronized clocks [13]. Third, MM2ME takes into account the possible asymmetry between the two directions of a connection.

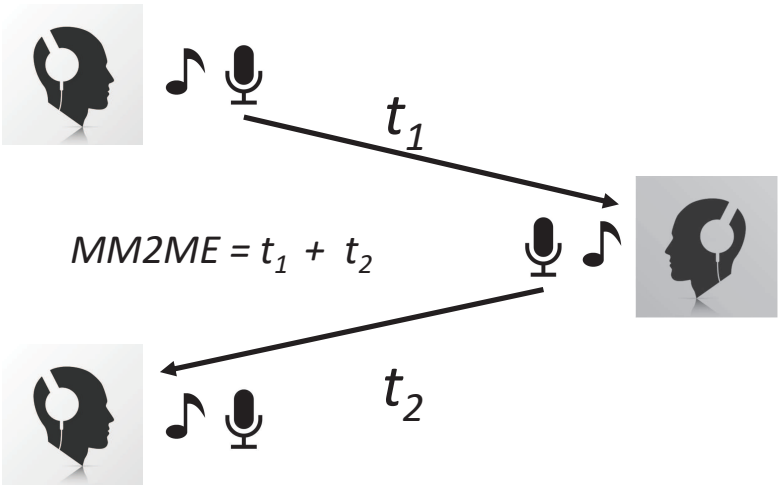

Figure 3.3: My Mouth to My Ear delay.

Of course, since most of the literature uses one way, M2E, delay, we clearly indicate which delay we are using, converting from one to the other when appropriate, assuming that delay is symmetric, i.e., the same in both directions; in all our tests, it was actually symmetric. This is not always the case, as asymmetry is the rule in ADSL

connections for example, but it is most likely what was assumed for the delay measurements given in the literature, where LANs were used for the last hop, therefore this assumption makes comparisons with past work more accurate.

3.3 Audio Quality in NMP

The second crucial, although not obvious, factor in NMP is audio quality. In telecommunications and, more particularly, in telephony, where human voice is transmitted, audio quality is intentionally downgraded to save bandwidth. Fortunately, the ear is most sensitive in a specific range of frequencies, so a portion of the voice bandwidth is enough for the listener to recognize it. Thus, audio quality reduction is widely used in telecommunications to save bandwidth, filtering frequencies beyond a threshold (3.4 kHz for standard telephony, 7 kHz for wideband tepehony). In addition, audio compression is used in mobile telephony, as well as in most commercial teleconference platforms, since delay in not so crucial in speech communication.

An NMP system needs a wider set of frequencies, so as to allow the timbre of each instrument to be discernible. Generally, sound must be rich in quality and as natural as possible. The listener must hear the full bandwidth of an instrument and any case of frequency filtering or noticeably compressed audio, or any kind of distortion can be evaluated as not satisfactory. Since for human hearing the highest audible frequencies do not exceed 20 kHz, any sampling rate above 40 kHz is sufficient to perfectly reconstruct an audio signal. We can therefore consider CD quality digital audio, with a 44.1 kHz sample rate and 16 bits per sample (and per channel, for multi channel audio) to be perfect. The 48 kHz sampling frequency is also commonly used in movie production, stemming from its use in Digital Audio Tape. For studio applications such as recording, mixing and editing, we often use double those frequencies (88.2 kHz or 96 kHz).

As mentioned above, the first big source of delay is related to sampling and audio buffering. Computer systems can perform satisfactorily when using a sample buffer size of 256 Bytes which, as mentioned above, introduces 5.8 ms of delay. The buffer size of 256 Bytes though is very small and often phenomena like over-runs or under-runs can occur. These lead to lost audio samples, which are heard as clicks in the audio signal. Audio clicks are very annoying and in many cases they are evaluated as distortion. To avoid audio clicks, the audio buffer size must be greater than 512 Bytes, which results in higher audio delays. Hence, there is always a trade-off between audio quality and audio delay in networked musical performance.

When algorithmic compression like Opus is used, audio quality is reduced. Although many perceptual studies have evaluated the quality of the compression codecs with high rates, when bandwidth is limited, the high compression ratio needed to fit the signal into the channel results in poor audio quality and lowers the QoME.

UDP is used in audio streaming applications due to its ability to transmit periodically, regardless of flow or congestion control. As it does not have mechanisms for recovering lost packets, audio samples can be lost in the NMP path. These lost packets are perceived as clicks and audio gaps. Although the RTP protocol is responsible for putting the packets in the right order, the Internet path can cause serious distortions to the audio signal, as many studies report.

3.4 Other Audio Characteristics

3.4.1 Performance Tempo

Tempo refers to the amount of beats during a minute, that a music performance, rehearsal or concert follows. It is usually given in *Beats per Minute* (BPM). In simple words, the tempo describes how fast or slow a song is performed by the musicians. It can be set by a metronome, a conductor or by the rhythm instruments of an ensemble (for example, the drummer in a pop band).

The tempo is important in NMP, since in order for musicians to play in a synchronized manner, they must follow the same tempo. Many studies have indicated that when delay is increased, the participants tend to reduce their tempo to compensate; interestingly, they also increase their tempo if the delay is less than what is expected (see Chapter 2).

3.4.2 Audio Envelope

The audio envelope is the outline of the waveform for a recording of an instruments' note. The so called ADSR envelope is different for every instrument. ADSR stands for *Attack, Decay, Sustain, Release* and describes the shape of the envelope for these four time stages. An example of an ADSR envelope is shown in figure 3.4. We can see that when a note is played, the amplitude of the sound rises fast (attack) and then drops a bit (decay); it then stays for a period at the same level (sustain) and finally it drops to zero (release).

Different audio envelopes are associated with different instruments, as shown in Figure 3.5; some instruments do not even go through some of the phases. The hand clapping technique used in many studies, when recorded, results an ADSR envelope similar to the percussive one shown in the figure, which has a very sharp attack and a quick release. In general, studies have found that instruments with envelopes closer to percussive ones, are more susceptible to delay variations, an important issue for NMP. Since percussive instruments are commonly used to maintain the rhythmic pattern of a performance, this influences the overall ability of the musicians to synchronize.

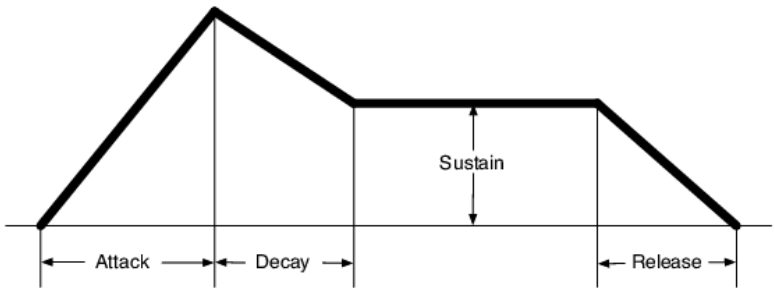

Figure 3.4: Attack, Decay, Sustain, Release time stages of a note

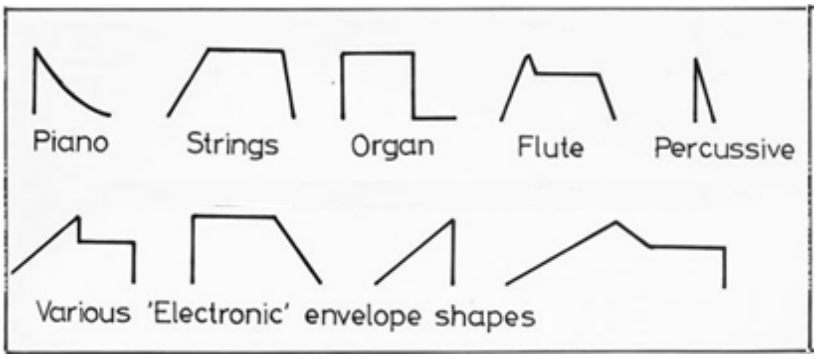

Figure 3.5: Various ADSR envelopes.

Chapter 4

Quality of Musicians' Experience

The main goal of our approach is to evaluate the quality of experience for musicians who participate in NMP sessions, by taking into account a large set of objective and subjective factors. For this reason, we first look at the different aspects of QoME and then introduce a framework specifically for the evaluation of the QoE of musical performances, that is, the QoME. At the end of the chapter, we narrow down the scope of our study to the variables that we controlled in our studies, in order to assess the QoME, and the variables that were used to group the results of the participants in the experiments for further analysis.

4.1 The Different Aspects of QoME

A large amount of research touches upon QoE evaluation for music in general, looking at it from different perspectives, but without taking a holistic view of its aspects. When discussing live music and concerts where a rock band or an orchestra performs music for the audience, everyone focuses on the experience of the audience. "Was the concert good?" "Did you like the music?" "Did you have fun?" are some questions individuals can answer after such an experience. Through these questions the experience can be evaluated subjectively, since the audience is a set of individuals.

There is also existing research related to music's emotional expression which focuses on the experience of the audience and what feelings individuals experience when listening to music. For example, Resnikow et al. [69] provide a study on the connection of emotion in music performance with emotional intelligence, where 24 students were asked to complete listening tests, trying to identify the induced emotions during performances of classical piano music.

Nevertheless, when talking about NMP the subject of interest is the musician himself; there may not even be an audience. Hence, the experience of the musician during the performance needs to be explored. Olmos et al. [61] experimented with two opera singers and a conductor over a network and evaluated two biometric measures, the *Galvanic Skin Response* (GSR) and the number of *Skin Conductance Responses* (SCR), using

software for behavior recording alongside questionnaires. Furthermore, Kubacki [52] offers a wide understanding of jazz musicians experience in the creation of live performances, using the method of interviews, concluding that their live performance is an experience created by the product itself. Finally, Geeves et al. [36] investigate the famous basist's Jeremy Kelshaw's performance experience via an interview with him.

Extensive research has also focused on *Music Performance Anxiety* (MPA). Matei and Ginsborg [59] provide a study on the MPA of classical music performers in professional performances where interviews were used. Research reported by Kenny et al. [48] explored the interrelationships between occupational stress, perfectionism, aspiration, and MPA in a group of elite operatic chorus artists employed full-time by a national opera company. Osborne and Franklin [62] examine the theoretical adequacy of establishing MPA as a subtype of social phobia. But is MPA the only aspect of music performance experience, and if other factors affect music performance, which are they? Furthermore, is the musician's mood a piece of the puzzle for his overall performance? Even if the session is a rock bands' studio rehearsal, is music performance affected by the musicians' psychological state, the room and its characteristics? If, on the other hand, the session is a duet of highly trained classical musicians in a rehearsal room, do all the above parameters affect their performance?

The diverse types of work found in the literature indicate the different aspects of QoME that can be studied. Our research 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. Such a framework is proposed by Kilkki [49] to describe QoE for communications ecosystems. Similarly, Rojas-Mendizabal et al. [70] explores the QoE for e-health ecosystems. citerehman explores the ecosystem for users of video streaming services. Finally, Geronazzo et al. [37] has proposed a framework for the QoE of NMP for a specific music genre (chamber music), presented in Chapter 2. We draw upon all these sources to come up with our own framework.

4.2 The Proposed NMP framework

By examining the ecosystem of NMP, as well as previous work on QoE assessment, we propose a framework consisting of four components for assessing the QoME in NMP sessions. Figure 4.1 shows some of the individual factors affecting a musical performance; these need to be extended with additional factors, to account for the extra complexities involved in NMP. The first component of our framework is the *physical space* where the musician is performing, which is described by its acoustic characteristics, such as reverberation time, resonance, and the room's impulse response for each range of sound frequencies. The second component is the *user state*, which includes both transient characteristics, such as anger or sadness, and long-term personality traits, such as enthusiasm or ambition. The third component is the *technical equipment* used for the performance, which includes the user interface, computers, networks, equipment and

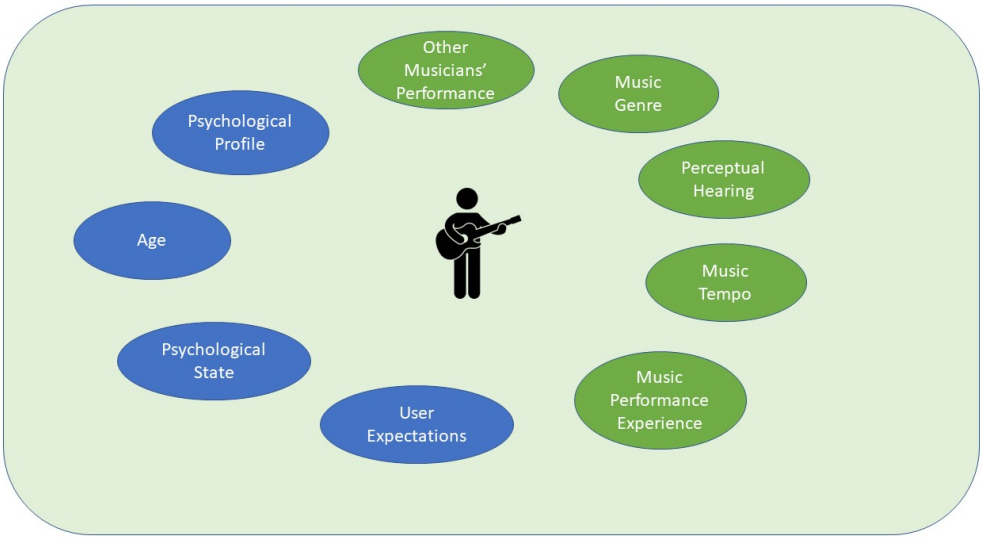

Figure 4.1: Factors affecting a musical performance.

all the technical aspects involved. The fourth component is the *musical context* of the performance, including aspects such as genre, tempo, instruments and musical scales involved, as well as performance characteristics related to the user, including experience and expectations. Each component includes the context and variables that interact with the user and affect QoME.

Our framework can be considered as a specialization of the Qualinet model for QoE, which identifies three categories of QoE *Influence Factors* (IFs) [57]: Human IFs, which are similar to our User State, System IFs, which are similar to our Technical Equipment, and Context IFs, which we split into Physical Space and Musical Context. We split context into these two parts since they are largely orthogonal, that is, the same studio can be used for very different musical performances.

The proposed NMP framework is outlined in Figure 4.2, which shows the environment of two musicians performing while placed in separate rooms connected via the Internet, computers and an NMP server; this is the technical equipment part of the framework. As shown, the QoS for the technical part can be evaluated in many ways, using objective metrics, such as latency, jitter and audio quality.

Unlike QoS, the quality of each musician's experience (QoME1 and QoME2) is a function of multiple variables, a subset of which is expressed by the QoS. The *Environment Acoustic Variables* (EAV) represent the physical space part of the framework, while the *Music Performance Variables* (MPV) represent the musical context part of the framework. For the user state part of the framework, we consider two aspects: the *Psychological State* (PS) refers to transient aspects of the performer, like the mood of the day and happiness or sadness at the time of the performance, while the *Psychological Profile* (PP) refers to the musician's overall personality and his fixed traits. We discuss

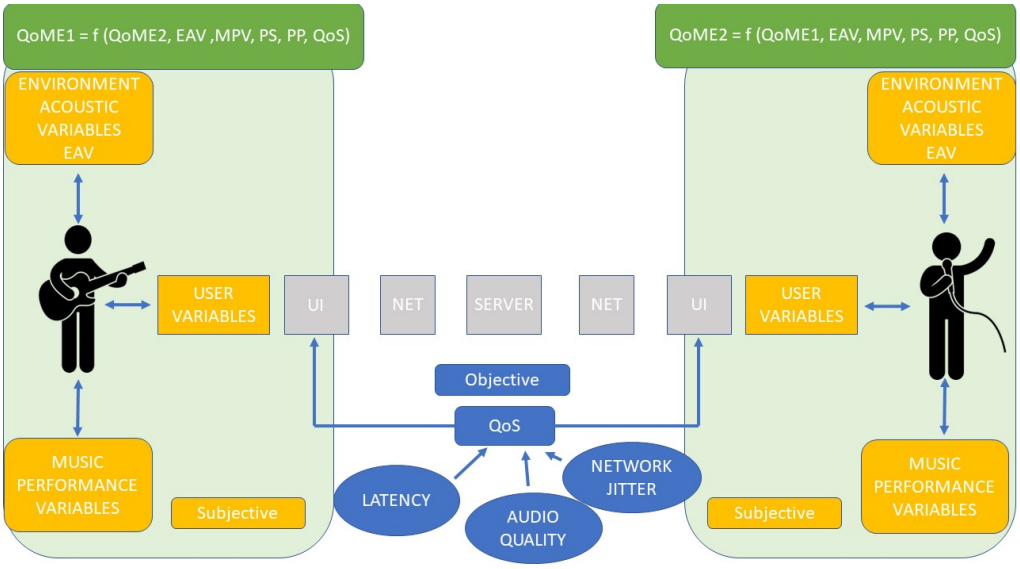

Figure 4.2: Network Music Performance Framework.

all these classes of variables in the following subsections.

4.2.1 QoS variables

The technical components of the ecosystem start with the *User Interface* (UI) offered by the NMP system, include the computer, the network and the server and end with the other user's UI. This component, considered as a system, has variables which affect QoS. Network latency, audio quality and network jitter are three parameters which are strongly correlated to the QoS. There is existing research related to how network latency affects QoS for NMP systems and, furthermore, how the musician's performance is affected by the audio latency, which concludes that as audio latency increases the musicians slow down their tempo. The audio quality on the other hand is a parameter that can be configured in such a way that latency can be reduced. Changing, for example, the audio sample rate to a value of less than 44.1 kHz, the audio bandwidth is reduced, which is perceived as lower audio quality by the musicians, but allows using lower bitrates without inducing compression delays. Finally, since the UDP protocol is used, there is always a percentage of packet loss, perceived as noticeable clicks by the musician, something also undesirable. Finally, delay jitter can lead to some packets arriving to late to be useful, thus turning them into losses; a jitter buffer would avoid these issues, but at the expense of increasing delay. Hence, these three variables, network latency, audio quality and network jitter are strongly correlated to the performance and will be taken into account in our study.

4.2.2 Music Performance Variables

Music Performance Variables are aspects related to the musical context. The first variable in this section is the musician's performance experience. Experience is an aspect that affects performance in musical sessions: a more experienced musician can perform with more confidence than a less experienced one, adapting to more difficult performance conditions. On the other hand, a more experienced one may have greater expectations from the NMP system. Another critical variable in this section is the performance of the other participating musicians: each musician's performance is directly affected by his peer, especially in musical genres where joint improvisation takes place. Additionally, perceptual hearing and music tempo are aspects that have a strong correlation to the performance.

4.2.3 User State Variables

In this section, aspects related to a musician's personality are discussed. A key piece of the puzzle in the discussed ecosystem is the musician's current psychological state. For example, a happy (in life) musician is expected to perform with high energy and enthusiasm, unlike an unhappy one who would probably perform with lower energy due to personal problems and the things that make him unhappy. There are many aspects that could describe the psychological state of a musician like anger, happiness, sadness, depression, boredom and many others, which may have a bearing on performance. One of our objectives in this thesis is to evaluate these parameters in real time, by using emotional recognition captured through video recordings of the sessions.

The musician's personality is another factor that plays an important role in performance. Personality aspects like aggressiveness, passivity, enthusiasm, patience, greediness could affect performance in general. For example, a soloist tends to play the central role in an orchestra or a band, something crucial in our case, as it tends to affect the performance of other musicians, too.

4.2.4 Environment Acoustic Variables

A subjective parameter that affects performance and overall experience is the way musicians perceive sound. As is well known, any individual perceives sound in a very specific psychoacoustic way, different from others. For example, older people perceive a narrower band of frequencies than younger ones. In addition, the acoustic profile of the room, such as reverberation time, resonance, and the room's impulse response for each range of sound frequencies, change the audio that the performers experience. These variables are related to the construction of the room, whether it is a home studio, a professional studio or any other type of room.

4.2.5 Quality of Experience as a function

Based on the above analysis of the parameters of the ecosystem, we come to the conclusion that QoME is correlated to all of them, as well as the QoME of the peer, that is, QoME can be expressed as the function

$$QoME_1 = f(QoME_2, PS, PP, EAV, MPV, QoS)$$

Where $QoME_1$ stands for Quality of Experience of the first musician, $QoME_2$ refers to the Quality of Experience of the second one, PS stands for the psychological state, PP stands for the psychological profile, EAV includes the environment acoustic variables, MPV includes the music performance experience introduced above and QoS includes the metrics for the technological aspects of the NMP system.

4.3 Variables under study

Since QoME is a function of so many variables, it would be unrealistic to examine all of them as part of a single dissertation. For this reason, we have focused on our work on the most limiting factor, audio delay, which determines whether NMP is feasible or not for a specific technical context (e.g., for a specific network path and its delay). Secondarily, we study audio quality as a means of reducing the audio bitrate, so as to avoid increasing delay due to compression and decompression. That is, we are only interested in how much we can drop quality, before it starts affecting the QoME.

Although these are both QoS variables, our intent is to assess their relationship with the QoME. As a result, our studies are informed by the QoME framework in many ways. In the subjective analysis, we include in our questionnaire many Music Performance Variables, such as the perception of the performance of the other musician, while also noting the effect of a musician's experience. In the emotions study, we consider a User Variable, the emotional state of each musician, and the effect that it has on the emotions expressed during the study. Finally, we consider grouping musicians based on various aspects, including parameters such as experience (in years), to detect correlations between these aspects and the QoME.

Chapter 5

Software for NMP Experimentation

Based on the QoME framework presented in Chapter 4, we decided to design and execute our own experiments to assess QoME in a realistic setting, that is, with real musicians performing real musical pieces. Our goal was to investigate correlations among the variables introduced and the aspects discussed in the previous chapter, using a larger number of participants than previous NMP studies, a more extensive questionnaire, and a wider range of assessment methods.

In order to carry out the experiments, we needed to be able to separately control the audio delay and the audio quality for each experiment, in an exact manner, and monitor the performance of the NMP system in real time. While existing NMP software does an excellent job at handling NMP sessions, and various tools are available to control the underlying variables and monitor performance, the need to quickly change settings for each repetition of our experiments led us to design our own software tool for NMP experimentation. In a nutshell, our tool aims to simplify the experimental configuration for the benefit of the researchers, rather than simplify the use of the system for the benefit of the participants.

In this chapter, we begin with a presentation of the software that we developed for NMP experimentation, explaining the reasons for its development. We then discuss the validation testing that we performed, and then describe a small subjective study that we carried out with real musicians, in order to evaluate our full experimental setup (including our software) in a small scale.

5.1 The Aretousa Tool

5.1.1 Motivation and Implementation

Among the frameworks used for NMP, JackTrip [11] is the most common choice for end users. It is free to download and simple to use, but it only supports direct connections

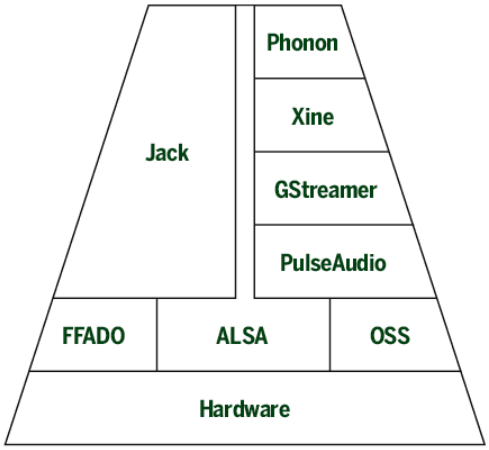

Figure 5.1: Gstreamer and JackTrip API's stack.

between two musicians. JackTrip uses the Linux ALSA[1] drivers and provides ultra low latency with (uncompressed) PCM audio. The user can listen to his own sound, along with the sound from his peer, and configure the audio buffer size and a few other driver properties. Most importantly, JackTrip is fast, as it was built to reduce latency as much as possible.

While JackTrip is an excellent NMP system, it aims for simplicity and speed in operation, as it targets musicians actually participating in NMP sessions. Our desire was instead to evaluate the musicians' QoME during NMP as a function of several parameters, including the audio format, network topology, audio buffer size, Ethernet packet size, audio compression parameters. This requires a far more flexible system, that could sacrifice simplicity, but should strive to keep latency low. This is the motivation for developing our own NMP system, Aretousa, which is used by the researchers to control the experiment; the musicians do not see our software at all.

To avoid low-level coding that would make the prototype difficult to implement and maintain, Aretousa is based on the GStreamer[2] and GTK[3] open frameworks. It supports the initialization, configuration, control and mix of multiple outgoing and incoming audio streams. GStreamer provides tools to build audio pipelines that capture audio from the sound card's input, make all the necessary transformations to the audio format, segment audio to packets, add necessary protocol headers like RTP and UDP and, finally, send it to the network using the `udpsink` plugin. At the other end, the `udpsrc` plugin, receives incoming UDP packets at a UDP port chosen by the user, strips the protocol headers and sends audio samples to the sound card's output. These pipelines are constructed by plugins connected serially via sources and sinks. The element that captures audio is `pulsesrc`, while the element that plays out the audio is

[1] Advanced Linux Sound Architecture
[2] https://gstreamer.freedesktop.org/
[3] https://www.gtk.org/

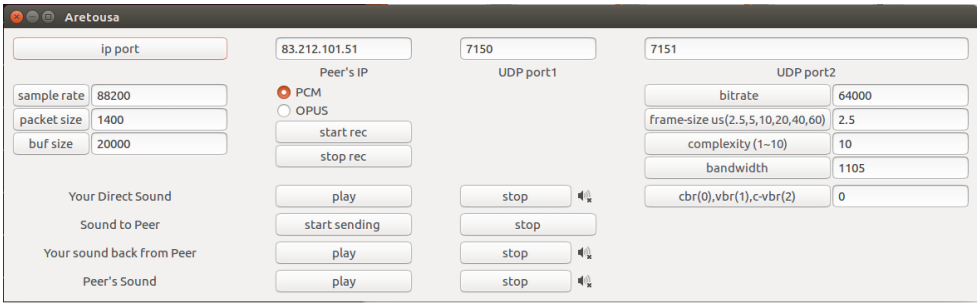

Figure 5.2: Aretousa's client UI.

`pulsesink`; these use the Pulse audio API which in turn uses the ALSA drivers to capture audio. GTK is used for the *User Interface* (UI).

As shown in Figure 5.1, while JackTrip uses directly the ALSA layer, the GStreamer framework sits on top of PulseAudio *and* ALSA. Gstreamer provides a lot of facilities to the programmer on top of the simpler interface offered by ALSA. For example, we can construct a simple pipeline using GStreamer command line tools to listen to our own audio, for example, `gst-launch-1.0 pulsesrc ! pulsesink`. Using this script, the user will experience an audio delay of over 200 ms. Each of these plugins is followed by parameters which can be configured for experimental purposes. By constructing a pipeline which streams audio to a certain host, we can implement a topology to evaluate latency for the path that the stream will follow.

Aretousa supports both peer-to-peer and server-based architectures, as well as both uncompressed PCM audio and the Opus codec for audio compression. Furthermore, the user can configure parameters such as the IP address of the NMP server in server-based mode, or the other peer's IP address in peer-to-peer mode, as well as the necessary UDP ports, as shown in Figure 5.2. Aretousa also allows configuring the audio buffer size, the Ethernet packet size, the Opus bit-rate and the Opus audio bandwidth, among other settings. The musician can listen to his own sound directly, his sound coming back from the peer (echo) and/or the peer's sound, controlling the sound level via volume sliders.

Aretousa provides the option of recording incoming and outgoing streams to separate WAV files, allowing the user to mix and/or analyze them offline, using audio processing software. We only used this facility for testing and delay measurement purposes, as we explain below; during the actual NMP sessions we used a separate machine to record audio, to avoid overloading the system used by the musicians.

Although Aretousa can be used by ordinary musicians, its target group is experimenters wanting to test different parameters and setups. To test server-based configurations, we have built a simple GStreamer-based server that operates as a *Selective Forwarding Unit* (SFU), that is, it selectively forwards packets, but it does not decode and re-encode them, to avoid adding delay. An SFU is very useful when more than two musicians want to collaborate, since it can replicate the stream coming from each

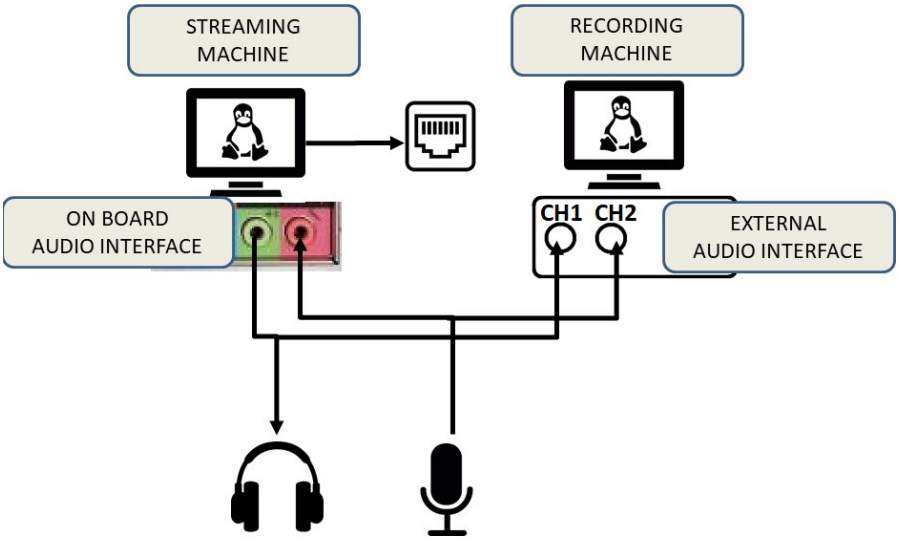

Figure 5.3: NMP endpoint configuration.

musician to all other participants in the session, without any processing [5]. We did not experiment with NMP sessions involving more than two participants in this thesis, but it is available as an option for future work.

5.1.2 Validation Setup

To illustrate the capabilities of Aretousa and demonstrate that, despite its flexibility and its prototype nature, it is comparable to JackTrip in terms of delay, we tested both systems and measured their *My Mouth to My Ear* (MM2ME) delay, as well as the round trip network delay between the testing endpoints.

The round trip network delay can be measured by using RTP time stamps in the packets and reflecting them back to their source. The one-way M2E delay also includes the time needed to capture, encode (optionally), packetize and then depacketize, decode (optionally) and playout samples; the two-way MM2ME delay requires repeating this process for each direction of communication. To assess such delays, we can send an audio stream to the peer, reflect the audio stream back to the sender, and compare the outgoing and incoming streams.

We ran tests using both Aretousa and JackTrip with uncompressed (PCM) audio. With Aretousa, we also tested compressed audio using the Opus codec, with varying parameters. We experimented with various Ethernet packet sizes and audio buffer sizes, so as to test their impact on audio delay, but also on audio quality (recall from Chapter 3 that very small buffers may lead to missing samples). The computers used for the experiments ran Ubuntu 16.04 with i7 processors and 12 GB of RAM; we used the on board sound card of each machine for audio capture and playback.

Figure 5.4: Topology using an NMP server.

As shown in Figure 5.3, at each endpoint the computer used for streaming was complemented by a separate computer for recording, which used an external audio interface, to avoid delaying the audio capture and playout operations due to the recording process. A mixing console with an auxiliary output, a condenser microphone and closed type headphones were used by each of the two musicians participating. The microphone, which captured physical audio, was routed to the audio input of the streaming machine for transmission using the Aretousa software. In parallel, it was routed via the console to the recording computer, using channel 1 of the external audio interface. The audio output of the streaming machine was directed to the musician's headphones, and was also routed to the recording computer, where it was recorded using channel 2 of the external audio interface. As a result, the recording combined what the musician produced (channel 1) and what the musician heard (channel 2).

As mentioned above, using Aretousa, the user can monitor or mute (a) his own sound, (b) the other peer's sound and (c) his audio echoed back from the peer. By monitoring his own sound, the user experiences an audio delay introduced by the streaming machine's audio buffers used to capture and playback sound. By monitoring his audio echoed back from the other peer, he also experiences delays due to packetization, network transmission and reception, depacketization at the peer, and then repacketization, transmission, reception and depacketization at his end.

In our first scenario, we played back the captured sound directly, so as to assess the delays due to audio buffering: the audio was captured and then played out directly in the same machine, in a loopback configuration. In our second scenario, we connected the two peers directly: the audio stream was sent from one client to the other, played out, captured again, and reflected back. This scenario captures the full MM2ME delay. The third scenario was the same as the second, but using an NMP server between the endpoints, as shown in Figure 5.4; the server did not perform any processing, operating as an SFU that passed through all packets without any processing. This scenario was only tested with Aretousa, as JackTrip does not support NMP servers. The NMP server was installed at the GRNET cloud[4], thus outside the AUEB network, therefore the transmission path included a number of hops (and the resulting delay) from each endpoint to the NMP server.

[4]https://grnet.gr/en/

To calculate MM2ME delay, we used hand claps, as they are easy to spot in audio processing programs: we simply needed to measure the distance between the peaks (the claps) among the two channels representing the sound sent and received. We then asked musicians to perform using the setup of Figure 5.4, without of course reflecting the audio of each musician. After each session, they were asked to answer a survey with questions about sound quality, clicks, audio interrupts, delay perceived, ability to synchronize, ability to express feelings and the overall testing procedure.

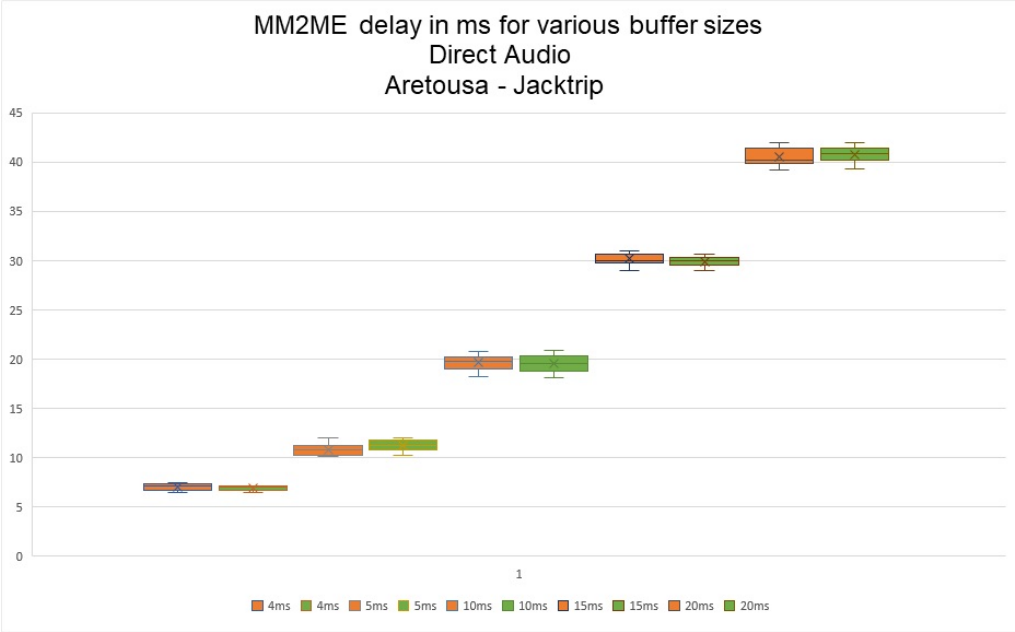

Figure 5.5: MM2ME delay with loopback connection for various buffer sizes.

5.1.3 Validation Results

Figure 5.5 shows the MM2ME delay measured in the first scenario (loopback audio in a single machine) as the audio capture buffer grows from 4 ms to 20 ms worth of audio, with uncompressed sound (PCM); orange boxes represent the mean/min/max and variance with Aretousa and green boxes the same metrics for JackTrip. Since in this scenario, audio is captured and played out once, these numbers are the minimum possible one way delays due to the audio system (that is, without networking delays). It is clear that beyond 10 ms of audio buffering, it is impossible to achieve the latencies required for NMP, if we consider the EPT to be 25-30 ms. Note that the delays of JackTrip and Aretousa are nearly the same in all cases.

Figure 5.6 shows the MM2ME delay measured in the second scenario (direct connection between two peers); this is the real-world MM2ME delay in a configuration without intervening servers. If we consider as an upper limit of MM2ME twice the limit for M2E, that is, 2x25 ms to 2x30 ms, these results indicate that the audio capture

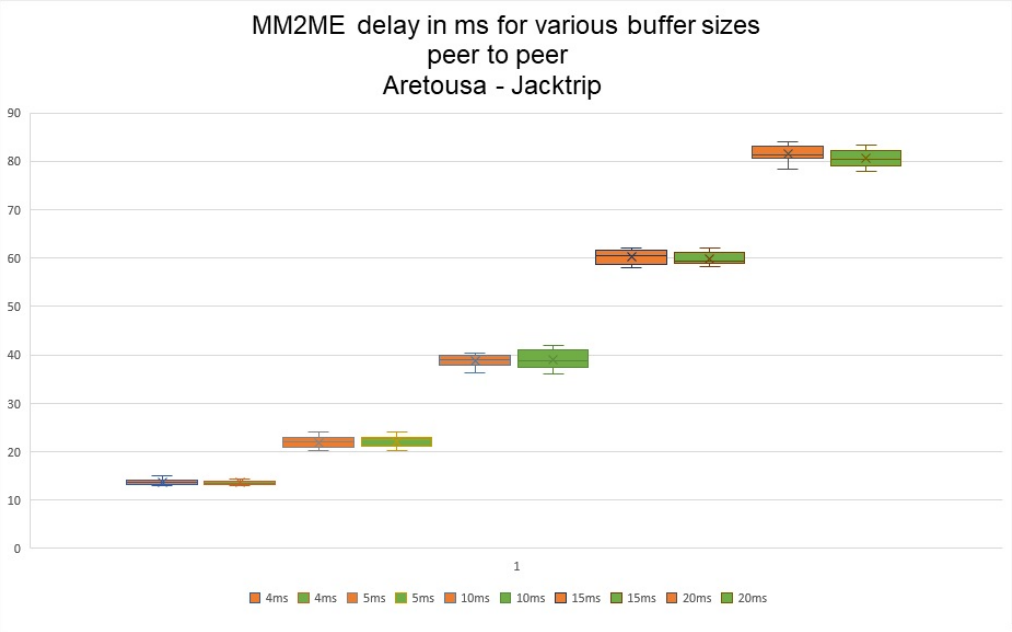

Figure 5.6: MM2ME delay with peer to peer connection for various buffer sizes.

buffer needs to be kept to no more than 10 ms. Again, Aretousa and JackTrip perform virtually the same in each case.

Figure 5.7 shows the MM2ME delay measured in the third scenario (two peers with a server in between, with the server located in the GRNET cloud), as the capture buffer size grows. This time, we only use Aretousa, since JackTrip is not compatible with servers. In addition to uncompressed audio (PCM), we show the MM2ME delay when using the Opus codec with a frame size of 2 us and 20 us. We also show for reference the network delay, as measured by the RTP timestamps of echoed packets; anything above this line, is due to the audio system. Assuming again that the MM2ME limit is around 2x25 ms to 2x30 ms, PCM audio can handle up to 10 ms of audio buffering, while Opus cannot handle more than 5 ms, due to its additional coding delay, which is 20 ms or more. We also note that the trip to the server and back has added around 7 ms of delay, by comparing the PCM delays at 10 ms of audio buffering between Figure 5.6 (for Aretousa) and Figure 5.7.

The MM2ME delay calculations show that over a high-speed research network, such as GRNET, and with a server acting as a simple forwarder for each stream, the largest fraction of the audio delay is due to the audio system, thus dominating the total MM2ME delay. The network delay when the server was included was very low, around 7 ms as calculated above; we also confirmed this value with Wireshark. Aretousa did not add any perceivable delay with uncompressed PCM audio, but coding with Opus added at least 20 ms of delay, mandating the use of a reduced audio buffer size to make NMP possible.

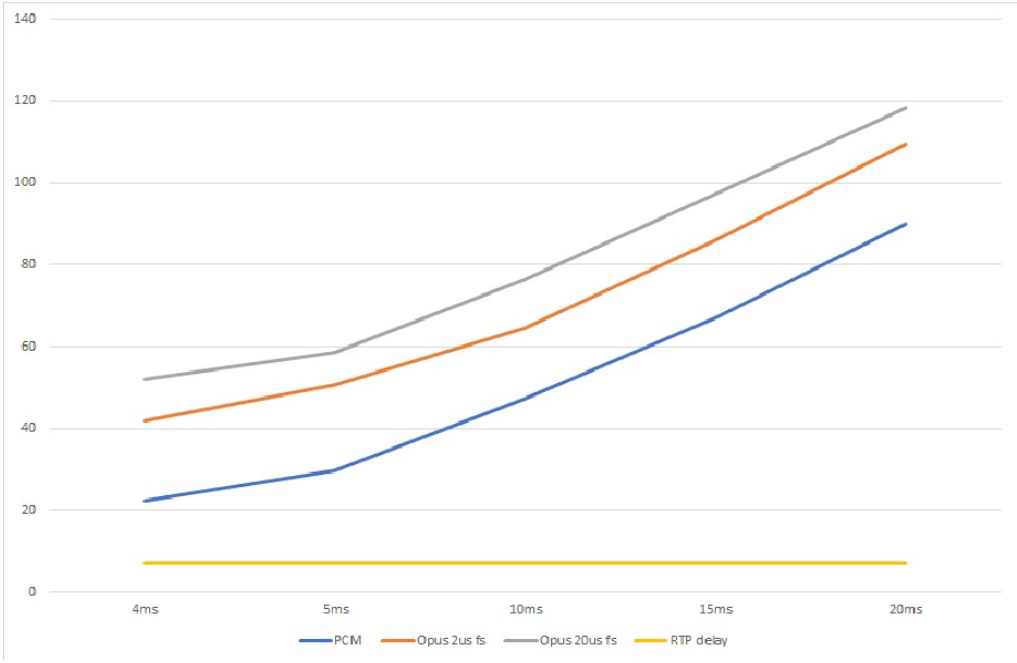

Figure 5.7: MM2ME delay via a server for various buffer sizes.

5.2 Subjective Evaluation

In order to test our setup and our experimental process, we asked a few musicians to perform using Aretousa, over the topology shown in Figure 5.4, with the two endpoints (and corresponding musicians) connected via an NMP server located at the GRNET cloud; the server did not perform any processing, it only forwarded packets. Each endpoint was connected to a small mixing console, a condenser microphone for the instrument and closed-type headphones for the musician. The computers used for the experiments ran Ubuntu 16.04 with i7 processors and 12 GB of RAM; we used the onboard sound card of each machine for audio capture and playback, with uncompressed (PCM) audio exchanged between the endpoints.

We conducted two sessions with real musicians performing over the system and used questionnaires to evaluate subjectively their experience, while varying the audio buffer size. The buffer size values used were 5, 10, 15, 20, 25, 35 and 50 ms. In the first session, two musicians played acoustic guitar and bouzouki (a traditional instrument). In the second session, two other musicians participated, playing the guitar and bouzouki again. After each performance (with a different audio buffer size), the musicians were asked to answer the following questions, using a 5 point Likert scale:

1. Evaluate the sound quality during the last musical performance (1 is very bad quality, 5 is very good quality).
2. Evaluate the degree of synchronization during the last musical performance (1 is no synchronization, 5 is perfect synchronization).

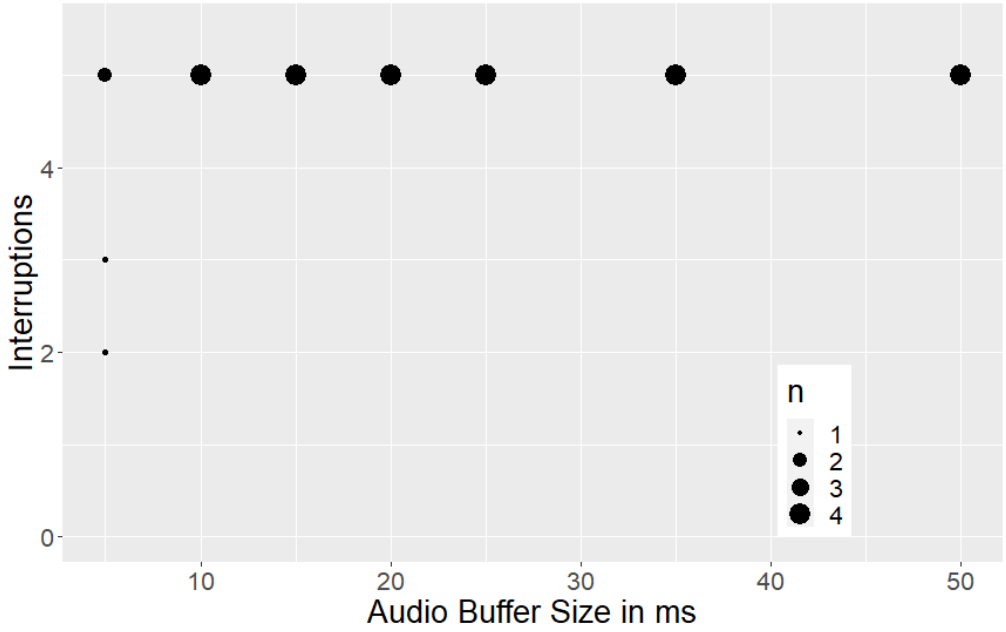

Figure 5.8: Evaluation of Audio Interruptions vs. Audio Buffer Size.

3. Evaluate the degree of audio delay you experienced during the last musical performance (1 is too much delay, 5 is no delay).
4. Evaluate the degree of your musical and emotional expression during the last musical performance (1 is no expression, 5 is excellent expression).
5. Evaluate the degree of interruptions in the sound during the last musical performance (1 is too many interruptions, 5 is no interruptions).
6. Evaluate your degree of satisfaction during the last performance (1 is very low satisfaction, 5 is very high satisfaction).

Note that higher values are better in terms of perceived quality, in all questions: a score of 5 means good quality, perfect synchronization, no delay, excellent expression, no interruptions and very high satisfaction.

We show the results for each question in Figures 5.12 too 5.11, combining the results from all four participants. Each graph shows the responses for each audio buffer value with dots, with larger dots representing more responses.

It is interesting to note that interruptions (Figure 5.8) were evaluated as nonexistent in all cases, except for the lowest buffer size, while the perception of synchronization (Figure 5.9) and the perception of musical and emotional expression (Figure 5.10) did not show a clear correlation to the buffer-size changes. While the sample size was too small to draw statistically valid conclusions, these results indicate that the perception of QoME can be quite different from what a simple QoS analysis shows.

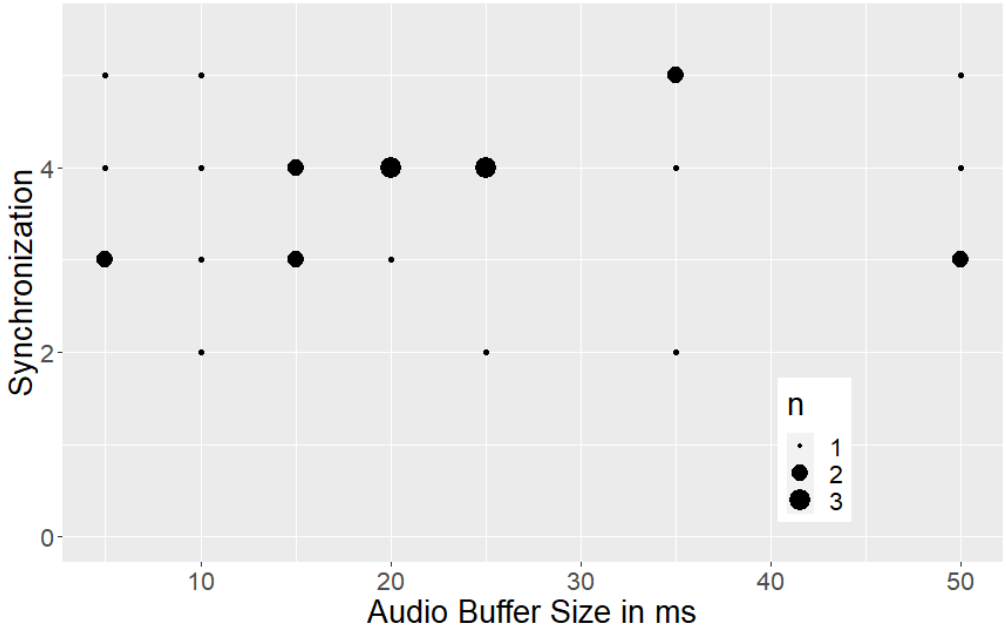

Figure 5.9: Evaluation of Synchronization Degree vs Audio Buffer Size.

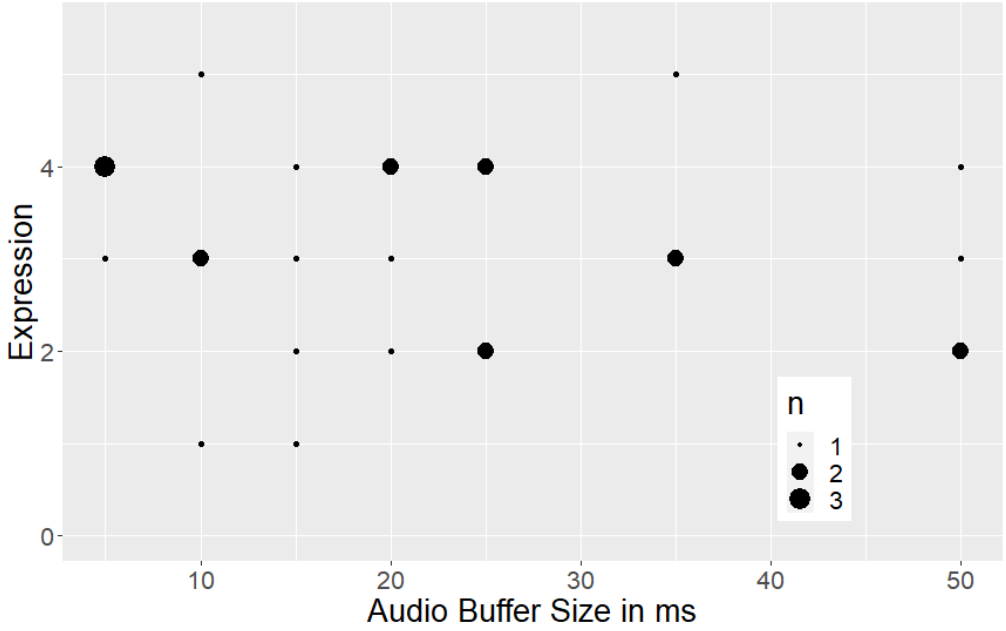

Figure 5.10: Evaluation of Musical and Emotional Expression vs Audio Buffer Size.

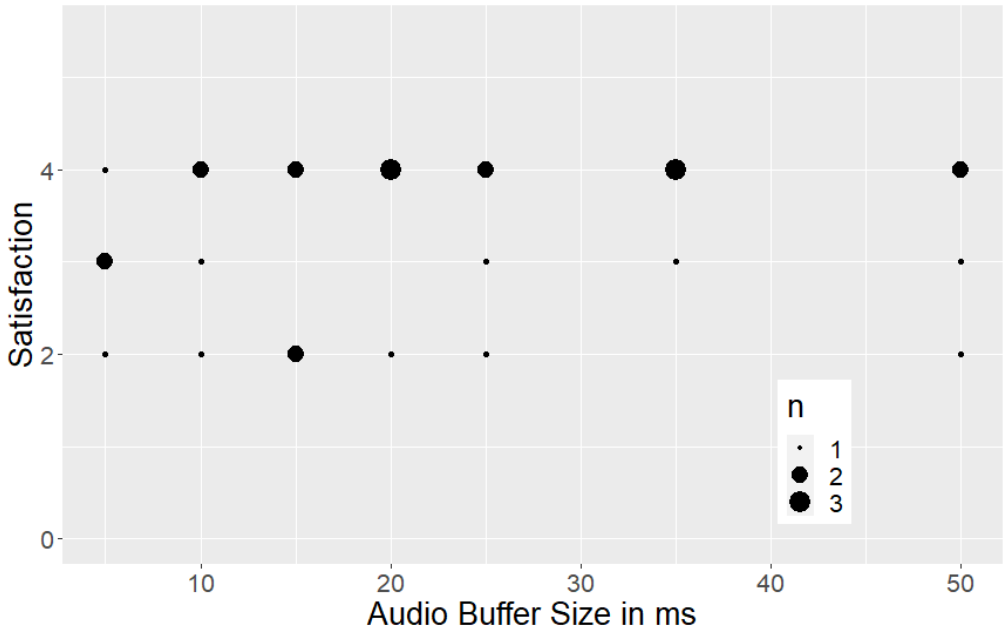

Figure 5.11: Evaluation of Satisfaction vs. Audio Buffer Size.

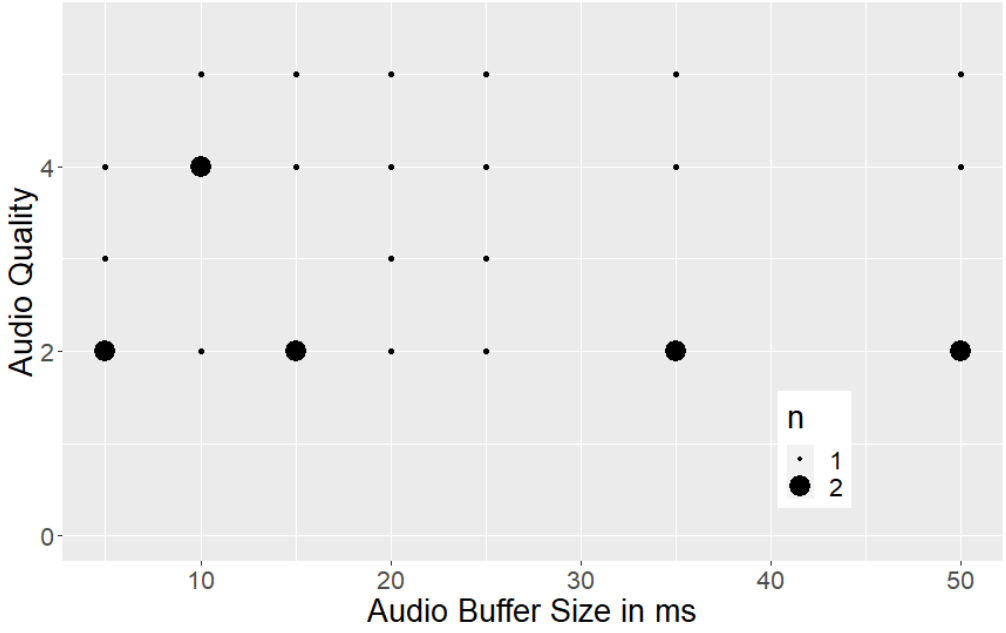

Figure 5.12: Evaluation of Audio Quality vs Audio Buffer Size.

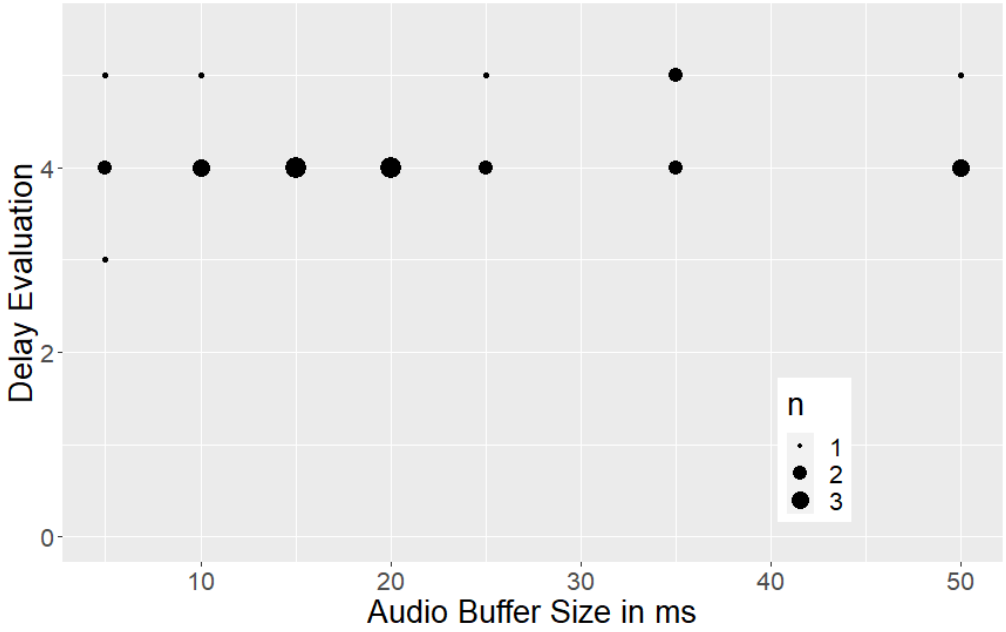

Figure 5.13: Evaluation of Audio Delay vs Audio Buffer Size.

In terms of the actual testing setup, the results indicate that our software performed acceptably under actual NMP conditions, with the perception of satisfaction being quite high (Figure 5.11), therefore it could be used for the larger studies described in the following chapters. Even though the quality (Figure 5.12) was found to be lacking, it is interesting to note that while the buffer size clearly influenced delay, as shown in the validation tests, the musicians rated the delay (Figure 5.13) nearly the same way in all tests; indeed, they rated the delay to be very low. This indicates that in an actual setup, musicians may interpret delay issues as quality issues.

Chapter 6

Pilot Study

Having verified (see Chapter 5) that our NMP software, Aretousa, with specific settings for packet and buffer sizes had acceptable performance, and that our overall setup was adequate for running experiments, we proceeded to design a subjective study. Running studies with real musicians, however, requires a lot of work: the organizers need to locate, co-ordinate and monitor the participants, while the participants need to spend considerable time to reach the testing site and perform a number of experiments with multiple repetitions. It is therefore advisable to begin with a pilot study, in order to assess at a preliminary stage the testing setup and the usefulness of the results.

The goal of our pilot study was to test a questionnaire with a large number of questions, in order to select those that would be used for the main study (see Chapter 7). We employed a moderate number of participants, in order to gain experience with setting up identical conditions for each experiment, and gain additional confidence in the subjective results. Furthermore, in the pilot study we tested not only variable audio delay values, but also variable audio quality levels, something not attempted in previous NMP studies.

In this chapter we first explain the motivation for the creation of the questionnaire used in the study, emphasizing the QoME aspect that each question wants to cover. Then we present our experimental setup, focusing on the changes from our validation test and the preliminary subjective study discussed in Chapter 5. Finally, we present and analyze the results of the pilot study, for both audio delay and audio quality.

6.1 Evaluation Variables

In the pilot study, we focus on the human perception of audio phenomena like audio delay and audio quality. Instead of a single *Mean Opinion Score*, we constructed an eight question survey covering different *QoE Features*, that is, *characteristic of an individual's experience of a service which contributes to its quality* [57], so as to provide a fuller picture

of the QoME in NMP. We then defined eight corresponding variables based on the *Perception of* statement and evaluated these variables by performing our survey on pairs of musicians participating in NMP sessions.

6.1.1 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 appropriate 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 audio delay you perceived during your last music performance.
4. Evaluate the extent of your musical 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 other musician's performance.

6.1.2 Perception of Audio Quality

An interesting question raised in our research was "How do musicians perceive audio quality?" For example, when a musician plays an acoustic instrument, she can hear the entire sound spectrum of it. Thus, the instrument's natural sound is considered to be perfect. When a musician performs through an analog audio system – where audio is just amplified by analog machines – then she will experience the amplified sound of her instrument. In that case, there are frequencies that may be louder and some equalizing is 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 recording interface, the musician experiences perfect sound since recordings use a sampling frequency of 88.2 or 96 kHz.

In the case of NMP, things can be quite different. Higher sampling frequencies translate into higher bit rates, which cannot be reduced by compression without substantially increasing delay. Therefore, we need to consider what sampling rates lead to acceptable audio quality and how a musician's experience is affected when audio quality is poor. Thus, we propose the *Perception of Audio Quality* (PoAQ) variable and search for its possible correlation with the actual audio quality and audio delay (recall from the previous subjective study that musicians often mistook higher delays for lower quality). Musicians were asked to evaluate in a 1 to 5 Likert Scale the perceived audio quality (higher is better).

6.1.3 Perception of Synchronization Degree

Achieving synchronization is a critical issue in NMP, and previous studies indicate that it is strongly dependent on the audio delay. Many studies have been conducted which conclude that when M2E delay (half of MM2ME) is below 25 ms, musicians can synchronize. When it increases above 25 ms, musicians start to slow down their tempo. Note that slowing down does not mean that one cannot play, since each musician tries to synchronize with the others; if the peers manage to synchronize at a slower tempo, they may still find the experience enjoyable. 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 fully synchronize).

6.1.4 Perception of Audio Delay

The most critical variable in NMP is audio delay; even quality, is subservient to delay in our work. Indeed, understanding the tolerance of real musicians to delay is the main goal of this thesis, Although our goal is to assess satisfaction, it is important to understand how delay is perceived and how this perception is related to with the actual delay in an experiment. An important variable is therefore the *Perception of Audio Delay* (PoAD). We examine how the participants perceive audio delay in a Likert Scale from 1 (no delay) to 5 (too much delay).

6.1.5 Perception of Musical and Emotional Expression

A musician's experience in NMP is strongly correlated to his musical and emotional expression during the performance, which can be hindered if the audio delay or quality prevent the musician from fully focusing on the music. 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 the

musician's musical and emotional expression could be affected by audio delay or audio quality variations.

6.1.6 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 artifacts are present, audio quality suffers for reasons unrelated to delay or sampling artefacts.

6.1.7 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. 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. It is also assessed in a 5-point Likert scale.

6.1.8 Perception of My Partner's Performance

With this variable, we search for correlations between a musician's performance and the performance of his partner. 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) to +3 (very good) how each musician assesses his partner's performance.

6.1.9 I was Trying to Follow

Through the final question, we try to assess the musical behavior 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 followed her partner or not.

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 6.1: 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 6.2: Sampling frequencies.

6.2 Experimental Setup

We used our prototype streaming software Aretousa (see Chapter 5) to stream audio and configure the necessary parameters for our experiments. The network topology was a peer to peer architecture with two computers and a fast Ethernet switch in the middle (without an intervening server). As shown in Figure 6.1, at each endpoint the computer running Aretousa was complemented by a separate computer for recording, which used an external audio interface to capture the audio; the intention was to avoid delaying the computer used for audio capture and playback with any other tasks, so as to minimize delay. 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. The sampling buffer size was set to 10 ms, as this setting offered good quality with reasonable delay. We experimentally found that the MM2ME delay in our setup was 34 ms (equivalent to 17 ms M2E); 20 ms of that (10 ms one way) was due to the sampling buffer.

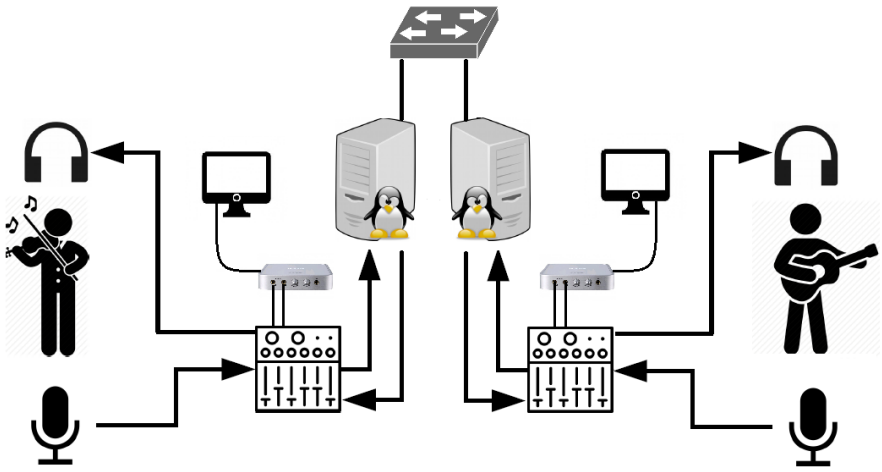

Figure 6.1: Experimental topology.

Eight musicians participated in pairs in the experiments, with each pair playing different musical instruments. The musicians could listen to each other only through Aretousa. Unlike in the initial subjective study, they also had visual contact with each other, via a WebRTC application, since studies have shown that both aural and visual cues are used during music performance. Note, however, that the video connection had a high latency due to the coding used by WebRTC and the cameras.

For each set of experiments, each pair of musicians played a one minute musical part of their choice, repeating it ten (10) times. 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 `netem` while audio quality was kept at 88.2 kHz. In Scenario B, audio quality was varied by modifying the sampling rate using Aretousa, while audio delay was kept at the minimum delay possible (34 ms).

As shown in Table 6.1 for Scenario A, the values of delay were set in a random order and not in an increasing one. In Scenario B, we also configured the various sampling rates in a random order, as shown in Table 6.2. In both cases, this was intended to prevent bias in the answers, as with strictly increasing or decreasing variables users tend to detect the pattern and expect what will happen next. The instruments played by each pair of musicians are shown in Table 6.3, while Table 6.4 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 had to play the same piece twenty times (ten for each experimental set) and answer the same questions. Musicians were not informed about which variable was manipulated each time, or about the purpose of the experiment. The main goal was to conduct an experiment that would allow us to evaluate multiple variables without bias or noise in the answers.

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

Table 6.3: 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 6.4: Age, Sex and Experience of each musician.

6.3 Scenario A: Variable Audio Delay

In this section we discuss the results shown in the eight plots from Figure 6.2 to Figure 6.9, for Scenario A, where we manipulated delay, using the MM2ME values shown in Table 6.1. The sampling frequency was kept constant at 88.2 kHz, providing studio-level audio quality. Each graph shows the responses for each delay value as dots, with larger dots representing more responses.

The graph of Figure 6.2 represents the results of the first question regarding the *Perception of audio quality* versus delay. We would expect the answers to vary between 4

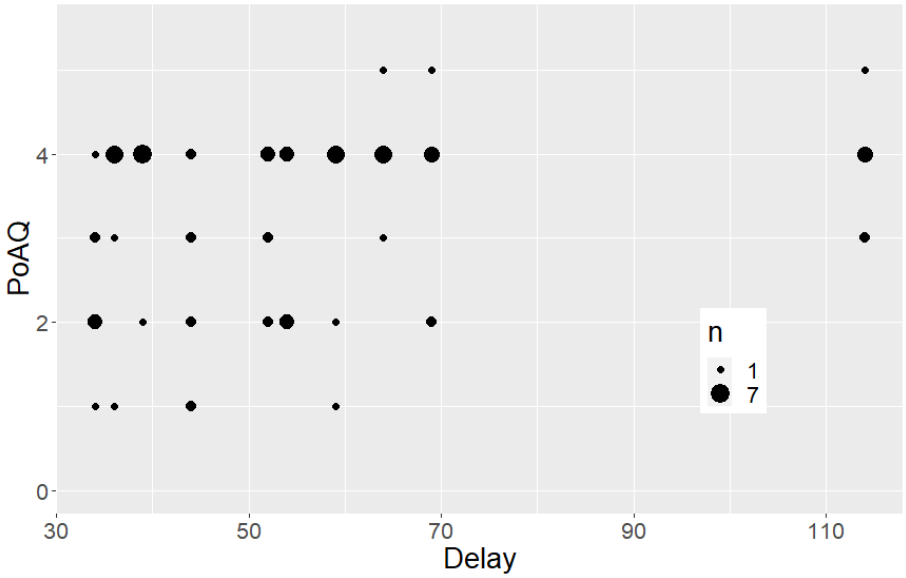

Figure 6.2: Perceived Audio Quality against delay.

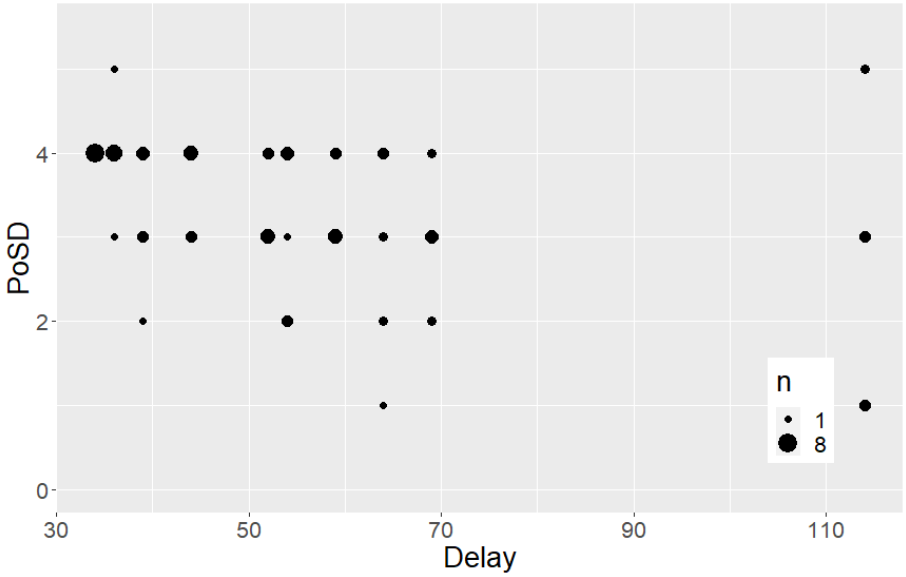

Figure 6.3: Perceived Synchronization Degree against delay.

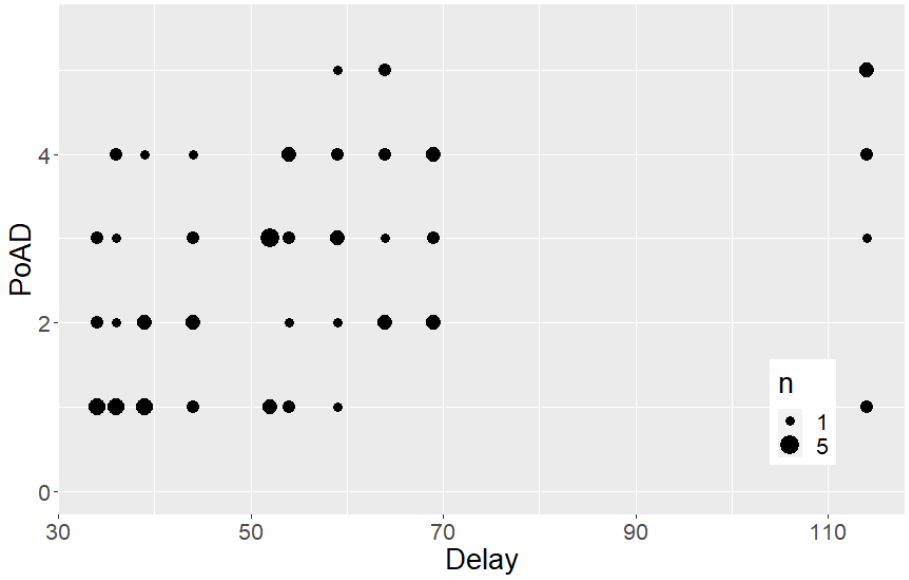

Figure 6.4: Perceived Audio Delay against delay.

and 5 corresponding to very good audio quality, since the sampling rate was constant and high. Instead, we can see that most answers were clustered around 4, with many answers at lower values.

In Figure 6.3 we show the answers to the *Perception of synchronization degree* question. As shown, answers vary between 3 (fair synchronization) and 4 (good synchronization). As the delay increases, answers decrease, but they are close to 3 even for high delay values. This is an indication that the participants make an effort to synchronize, even when delays are high. Furthermore, in the range from 50 ms to 70 ms, delay changes do not seem to dramatically affect their perception of synchronization.

In Figure 6.4 the relation between *Perception of audio delay* and the actual delay is shown. Answers vary between 1 and 4. There is a positive trend, indicating that the delay changes are perceived as such by the musicians, even though the variance is quite high, implying that the perception of delay is not very exact.

Figure 6.5 shows the answers for the *Perception of musical and emotional expression* versus audio delay changes. As shown, there are more responses with lower scores as delay grows, indicating a very small negative trend; the variance of the responses is again quite wide.

The graph in Figure 6.6 shows the answers for the *Perceived audio clicks*. Audio clicks were heard mostly as a result of lost samples at the two audio interfaces, and it was a question rather related to QoS of the Aretousa software than the QoME. Most answers were between 1 which corresponds to zero audio clicks and 2 corresponding to few audio clicks.

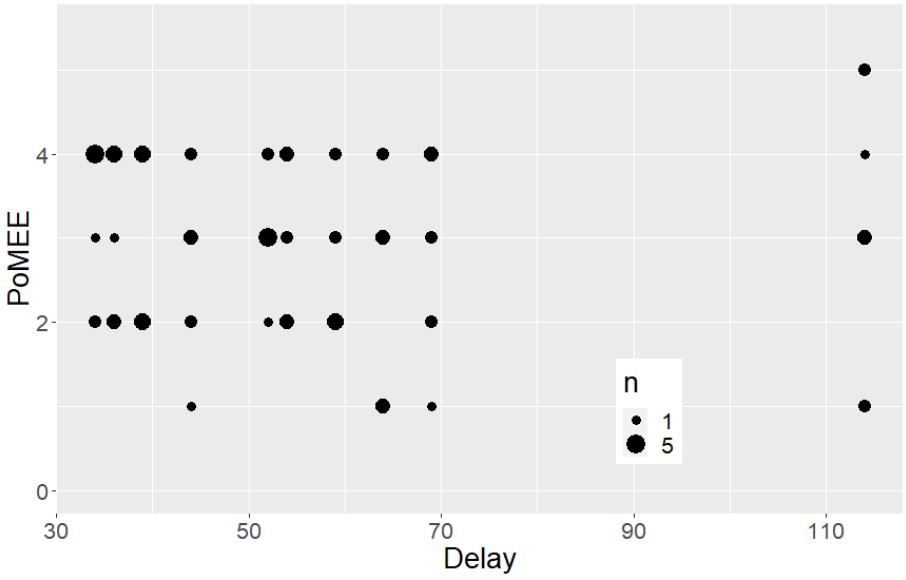

Figure 6.5: Perceived Musical and Emotional Expression against delay.

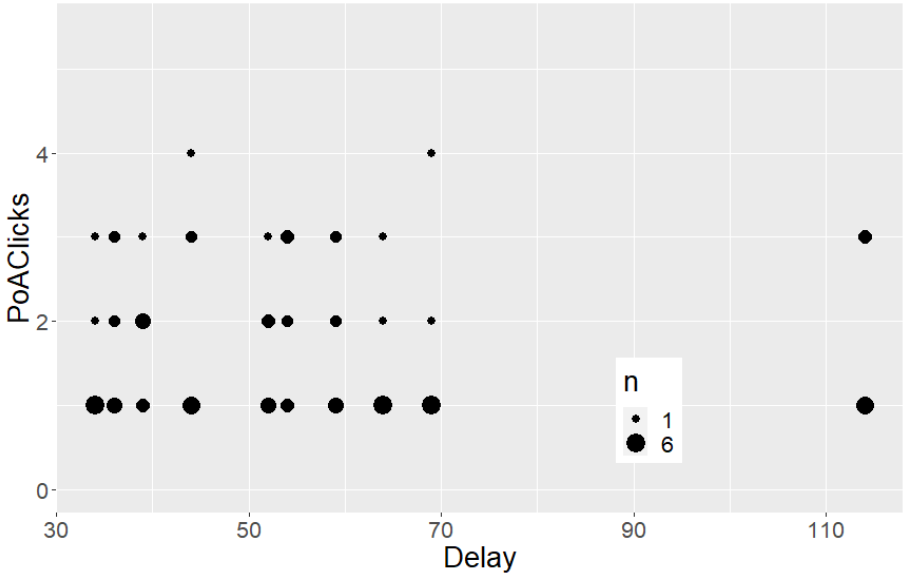

Figure 6.6: Perceived Audio Clicks against delay.

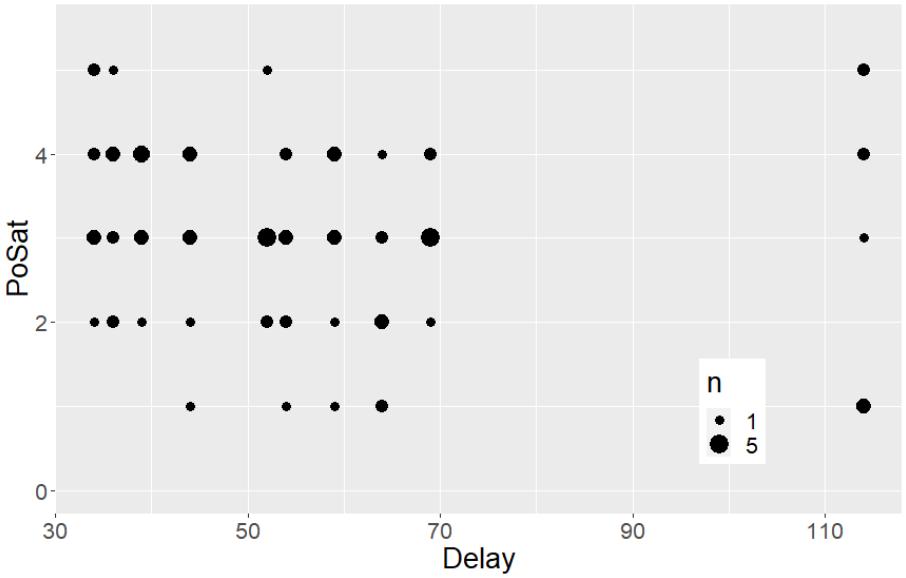

Figure 6.7: Perceived Satisfaction against delay.

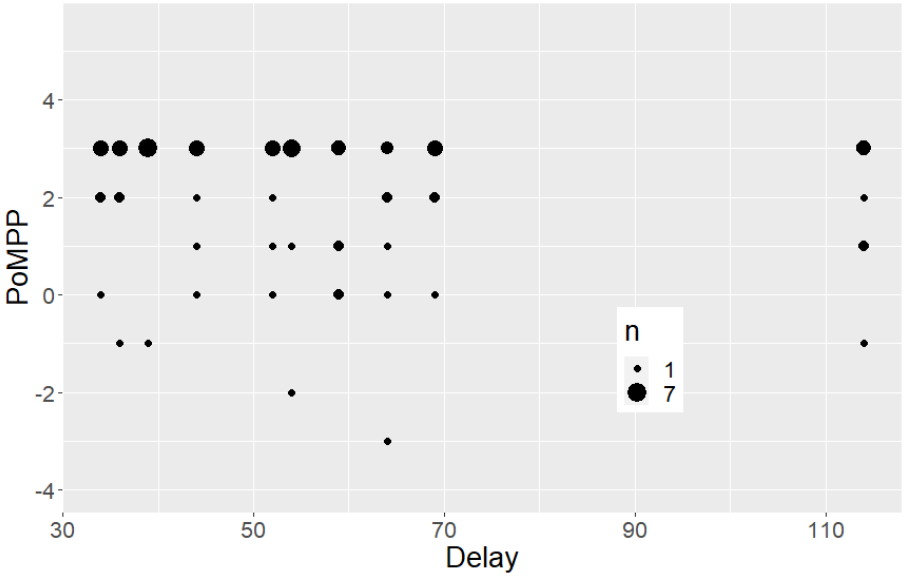

Figure 6.8: Perception of My Partners' Performance against delay.

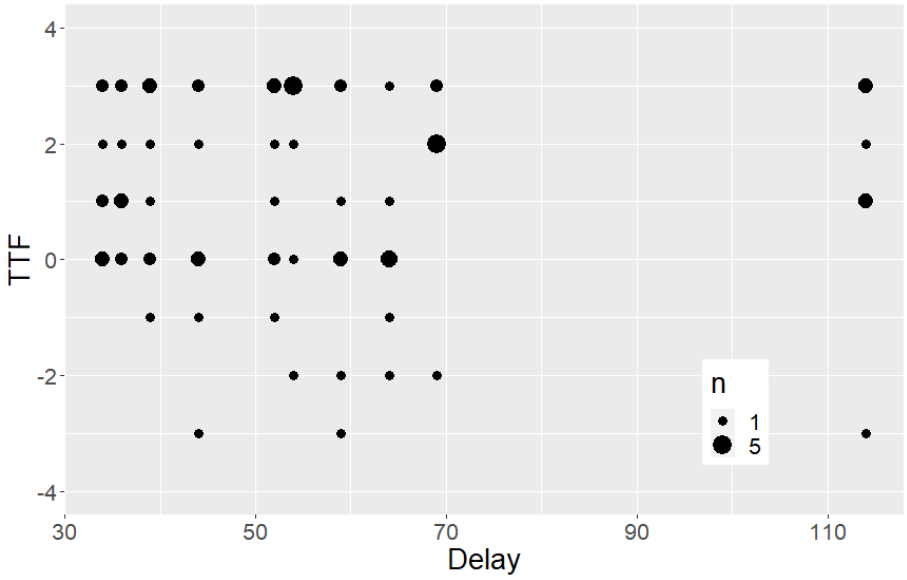

Figure 6.9: I was Trying to Follow against delay.

One of the most critical variables in our experiment is the *Perceived Satisfaction* per audio delay, which is shown in the graph of Figure 6.7. This variable is similar to the *Mean Opinion Score*. The answers vary from 2 to 4 for most delays. Again, in the range from 50 ms to 70 ms delay does not seem to greatly influence the answers. Although there is a slightly negative trend, the variance is quite high, as shown from the dispersion of the dots.

The *Perception of my partners' performance* is evaluated as shown in graph of Figure 6.8, where the participants had to choose between very good (3) and very bad (-3). Answers vary between 2 (corresponding to good) and 3 (corresponding to very good). This is an indication that regardless of the conditions, each participant believes that his partner was performing well.

Finally, in Figure 6.9 we can see the results for the *I was Trying to Follow my partner* (TTF) question. The median value of the answers vary in range from 0 to 3 (fully agree), indicating that musicians generally tried to follow their partners throughout the experiments.

6.4 Scenario B: Variable Audio Quality

We then discuss the results shown in the eight plots from Figure 6.10 to Figure 6.17. In this scenario, we manipulated audio quality using Aretousa, by using the sampling rates shown in Table 6.2. The audio delay in this scenario was kept as low as possible, therefore it was 34 ms (MM2ME). The lower value for the sample rate was set to 8 kHz corresponding to very poor audio (voice telephony) and the highest value was set to

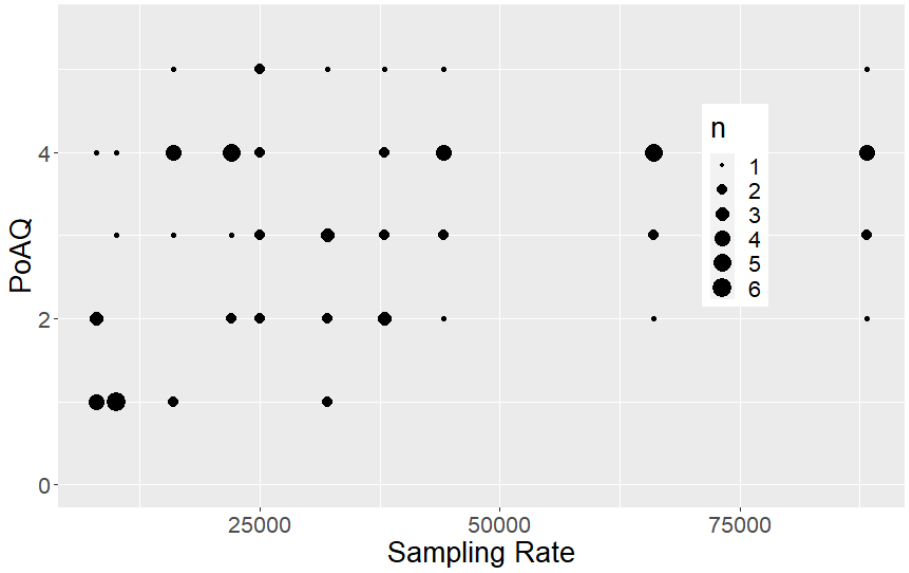

Figure 6.10: Perceived Audio Quality against quality.

88.2 kHz corresponding to very good audio quality (studio recording). The figures use the same format as in the previous section, but with quality rather than delay on the horizontal axis.

We can see the results for the *Perception of Audio Quality* variable in the graph shown in Figure 6.10. The answers vary in the range between 1 corresponding to very poor audio quality and 5, corresponding to excellent audio quality. Although the responses are widely dispersed, there is a clear positive trend, indicating that the perception of quality actually does follow the real quality of the audio. We note that above around 16 KHz, most participants perceive quality as good.

While the sampling frequency was 8 KHz, which corresponds to very poor audio quality, there was a perceivable amount of delay inserted due to the re-sampling filter used in the Aretousa software. This increased delay to more than 34 ms. This explains the graph in Figure 6.11 where the *Perceived Synchronization Degree* was rated as quite low by some musicians at the lowest sampling frequencies. However, it increases above 16 kHz where the delay becomes again 34 ms. In the range from 16 kHz to 88.2 kHz the answers are generally high.

The answers for the *Perception of Audio Delay* variable while sampling frequency varies, are shown in Figure 6.12. Due to the increase of delay at low sampling rates mentioned in the previous question,the answers are higher than what they should be for Sampling Frequencies below 16 kHz, something that we would expect. Above that, the answers are generally in the same (low) range, so variations in sampling frequency beyond that do not seem to affect the Perception of audio delay.

The graph shown in Figure 6.13 shows the answers for the evaluation of the *Perception of Musical and Emotional expression*. The median answers range from 2.5 to 4,

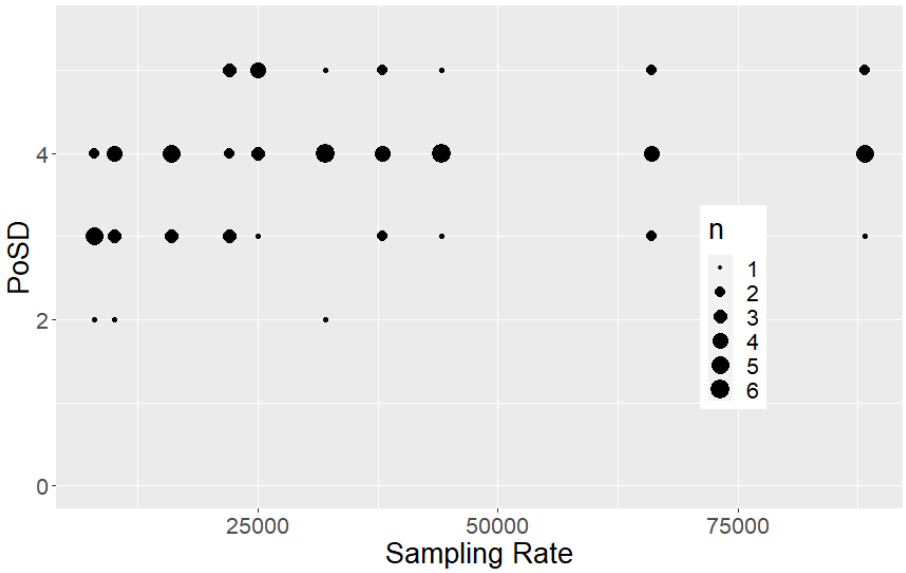

Figure 6.11: Perceived Synchronization Degree against quality.

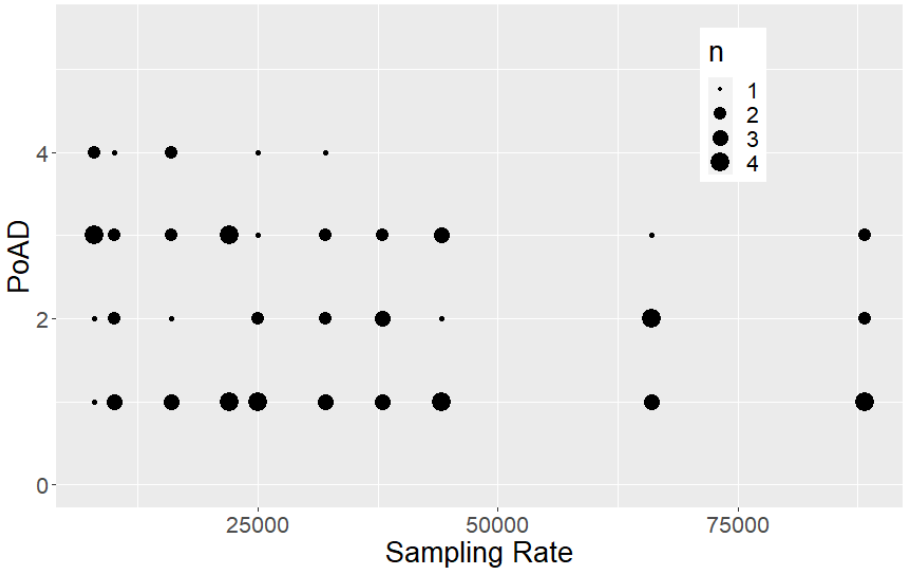

Figure 6.12: Perceived Audio Delay against quality.

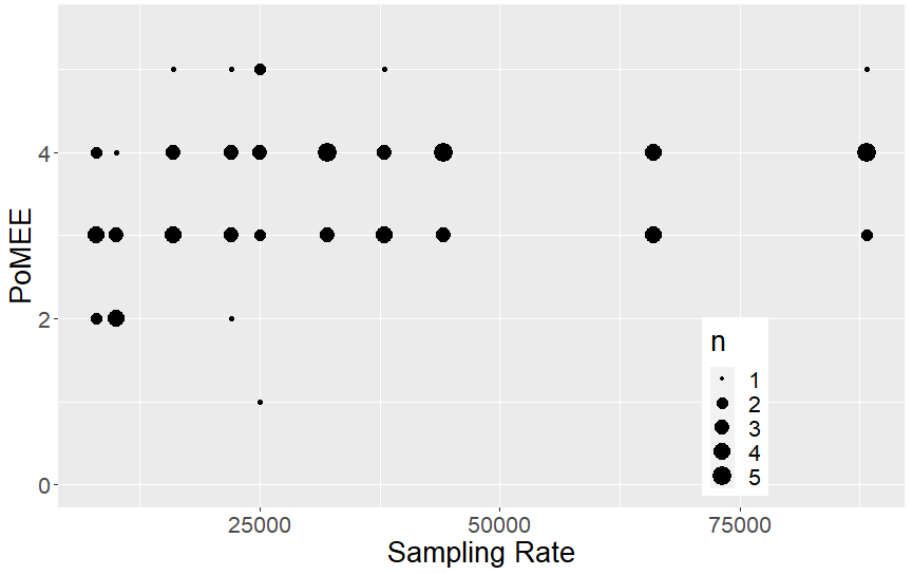

Figure 6.13: Perceived Emotional Expression against quality.

and we can observe a slightly positive correlation between the two variables. Additionally, there is also a threshold in the frequency of 16 kHz above which there are no real variation in the answers.

In this scenario, where sampling frequency was manipulated, there were no noticeable audio clicks as shown in the graph of Figure 6.14. The majority of the musicians perceived no clicks at all, responding with 1 in this question.

As in the delay experiments, the *Perception of Satisfaction* (PoSat) variable was found to increase as Sampling Frequency grew, as shown in Figure 6.15. The answers are in the range between 2 corresponding to little satisfaction and 4 corresponding to good satisfaction. In the range between 20 kHz and 45 kHz, the values are nearly the same, growing only slightly beyond that, which indicates that satisfaction is not greatly influenced by increasing the sampling rates beyond 20 kHz.

In the *Perception of My Partners' Performance* (PoMPP) question, most answers vary between 2 and 3 (fully agree), with only a few at 1, as shown in the graph of Figure 6.16 and there is little change depending on the sampling frequency. This indicates that musicians were satisfied with their partners at all quality levels.

Finally, the *I was Trying To Follow (TTF) my partner* results are shown in Figure 6.17. Due to the added delay issues for sampling frequencies below 20 kHz pointed out above, we observe more positive values at the lowest sampling rates. In the range between 16 kHz and 68 kHz, the median value of the answers is equal to 0 corresponding to neutral. As a conclusion, we can say that the TTF variable is not affected by sampling frequency changes, as all the participants were almost neutral is this question.

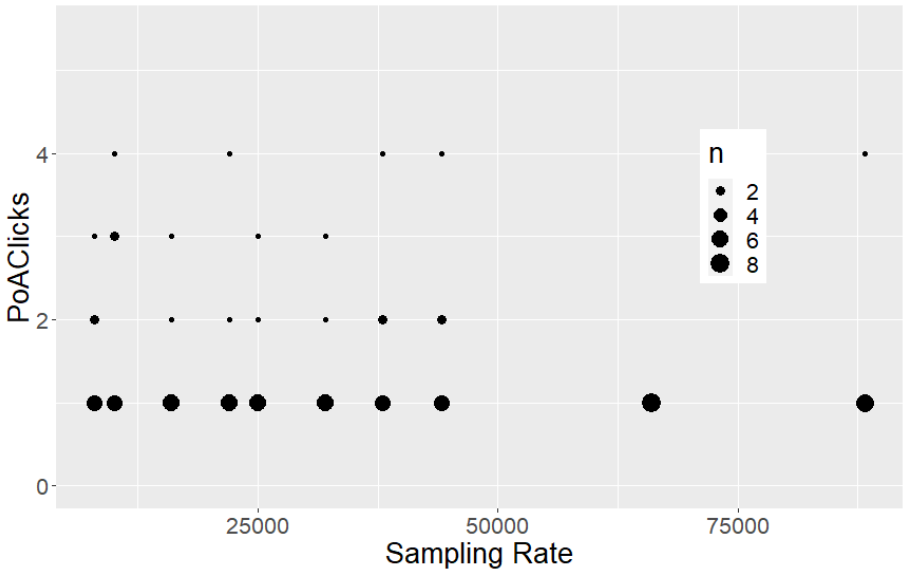

Figure 6.14: Perceived Audio Clicks against quality.

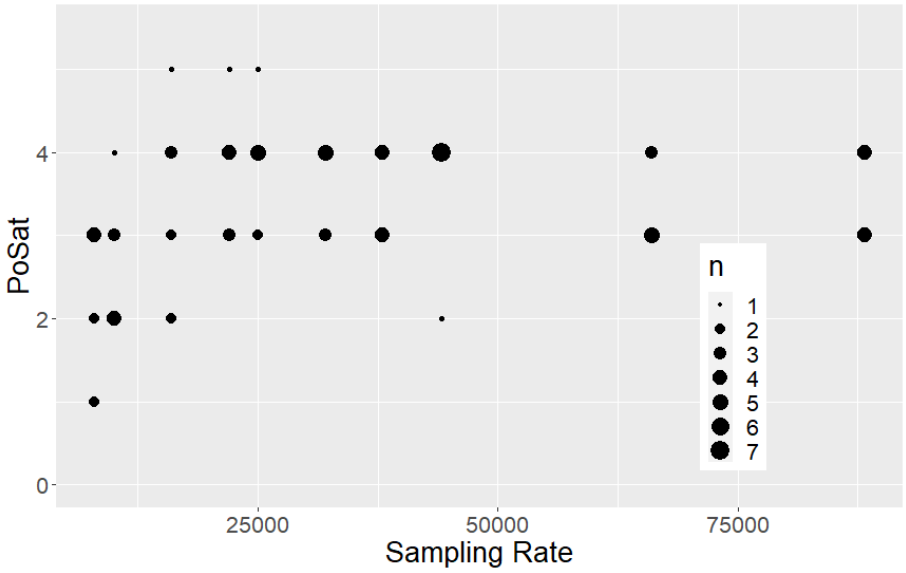

Figure 6.15: Perceived Satisfaction against quality.

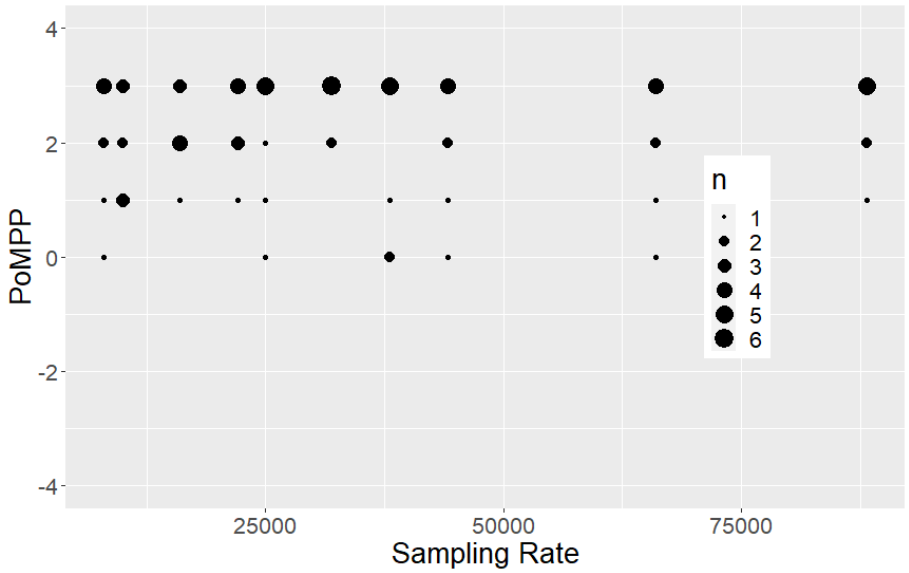

Figure 6.16: Perception of My Partners' Performance against quality.

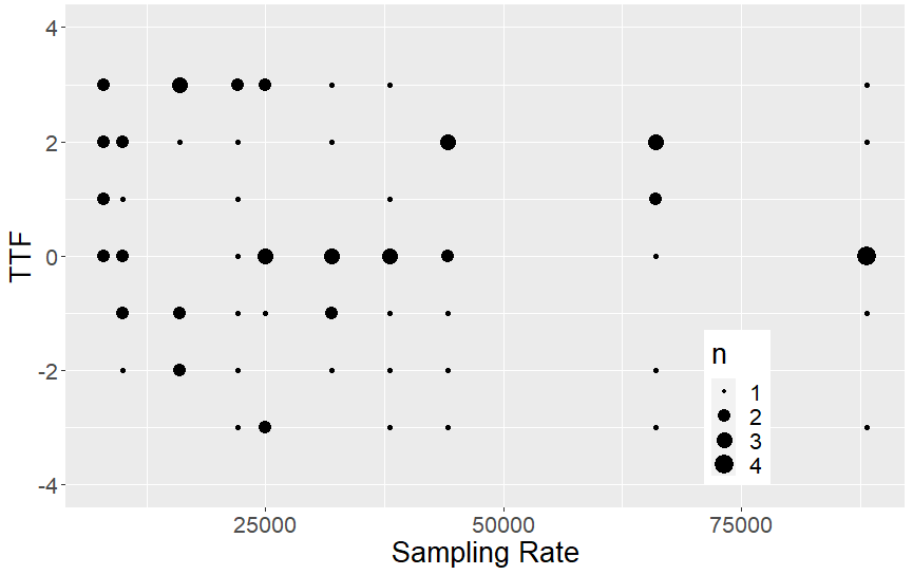

Figure 6.17: I was Trying to Follow against quality.

6.5 Summary of Results

In the pilot study, we conducted two sets of NMP experiments, first manipulating delay, and then manipulating quality, using an eight-question survey, which was answered by the eight musicians that participated as duets in both scenarios.

Before analyzing the results, an important observation is that the participants could not distinguish between noise, audio clicks, audio quality and other audio impairments. For example, when sometimes by accident re-sampling was introduced as a result of bad configuration, a phenomenon that only influenced the experiments with variable audio quality and at the lowest sampling frequencies, the musicians perceived that change as noise. The same thing happened with audio clicks. In general, anything unusual, whether noise or delay, was described as noise.

As we found from the results of Scenario A, the *Perception of Audio Delay* (PoAD) and the *Perception of Synchronization Degree* (PoSD) are correlated to audio delay variations. On the other hand, the *Perception of Musical and Emotional Expression* (PoMEE) was found to be almost unaffected by the audio delay variations. Interestingly, the *Perceived Satisfaction* was found to be only mildly affected, with small variations and a median value around 3, a fact that we did not expect; we expected, instead, the MOS score to drop more with increasing delay. Finally, the *Perception of My Partners Performance* PoMPP and the *I was trying to follow my Partner* (TTF) variables were also found to vary only slightly as delay grew.

In Scenario B where audio quality (via the sampling frequency) was manipulated, the *Perception of Audio Quality* (PoAQ), was found be affected less at frequencies above 16 KHz. Furthermore *Perceived Musical and Emotional Expression* (PoMME) were found to be similar to PoAQ. Most of the other perception variables were not greatly influenced by the changes in audio quality, even though the results at the lowest sampling rates were influenced by problems with our software. This indicates that human perception of audio quality may not be significantly affected above the 16 KHz threshold, for the instruments used in the pilot, which may allow reducing the bitrate to save bandwidth, with no appreciable loss in QoME. Of course, the issues with our software at the lowest sampling rates reduce our confidence in these observations.

Chapter 7

Main Study and Subjective Analysis

Following the pilot study presented in Chapter 6, we organized a larger study, where 22 musicians participated in pairs (11 duets in total). Based on the experiences with the pilot, we revised the topology of the experimental setup and the methods of data gathering, with the goal of achieving lower and more finely controlled delays, and we modified our NMP software to avoid the issues observed in the pilot. Furthermore, we reduced the length of the questionnaire and used a different set of questions for each variable under study (audio delay or audio quality) to avoid tiring the musicians. Finally, we ensured that in addition to the questionnaires, which we analyze in this chapter, we kept both audio and video recordings of all experiments for future study; we show results from such analyses in the following chapters of the thesis.

In this chapter, we describe this study, focusing on the setup, the execution and the analysis of the subjective data. Specifically, we first describe in detail our experimental setup, detailing the changes from the pilot, and then discuss the revised questionnaire, indicating the rationale behind it. Then, we analyze the results from the questionnaires, including the statistical significance of the results, for both the delay-based and the quality-based experiments. To the best of our knowledge, this study is the largest subjective study on the effects of audio delay to the QoME of actual musicians in NMP scenarios, as well as the first study on the effects of audio quality to QoME.

7.1 Experimental Setup

For the main study, we used two visually and aurally isolated rooms on the same floor of the main AUEB building: Room A41 and Office A410 on the fourth floor of the Antoniadou wing. Musicians performed with their counterparts in separate rooms, while listening to them through headphones and seeing them through a 32″ screen. We varied two underlying parameters: in Scenario A, audio delay varied while audio

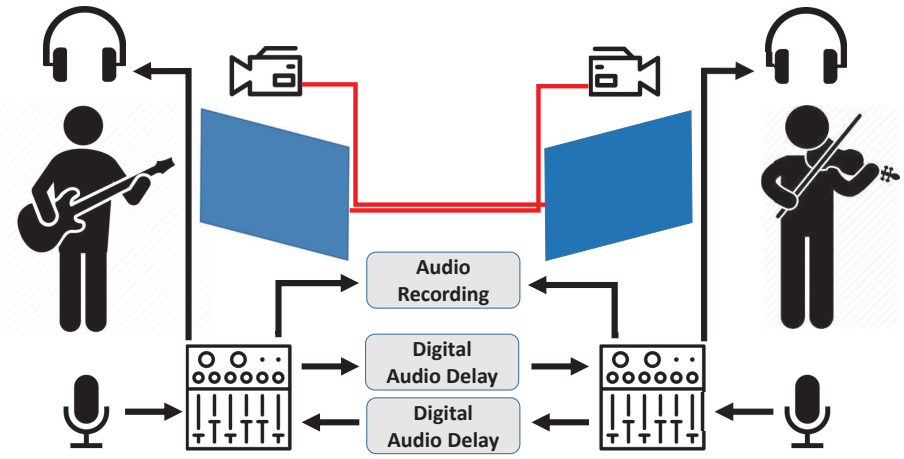

Figure 7.1: Experimental Setup for Scenario A (variable delay).

quality was fixed, while in Scenario B audio quality varied while audio delay was fixed, as in the pilot study.

To conduct our experiments, we used the same general topology, but with slightly different setups for each scenario. As in the pilot, we did *not* use a server between the two endpoints; we connected them directly in a peer to peer mode, in order to reduce the minimum delay of the system. Since the minimum MM2ME delay in the pilot was 34 ms (17 ms one way), we took additional steps in the main study to reduce the baseline delay, as explained in the next section.

7.1.1 Experimental Topologies

In Scenario A, shown in Figure 7.1, 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 video signal through the existing network cabling to the 32" screen of the other room (red lines in the figure). The network cables were patched directly to each other, without passing through *any* network equipment; we basically used one pair of the UTP cables to transmit the composite video signal in analog mode.

We used composite video in order to achieve the lowest possible visual delay between musicians; with the analog signal we did not have to wait for entire frames to be captured before transmission and received before display. We experimentally measured the round trip video delay by placing a smartphone with a running chronometer in front of the camera in one room, and turning the video camera to the video monitor in the other room, essentially reflecting the transmitted image back to the first room.

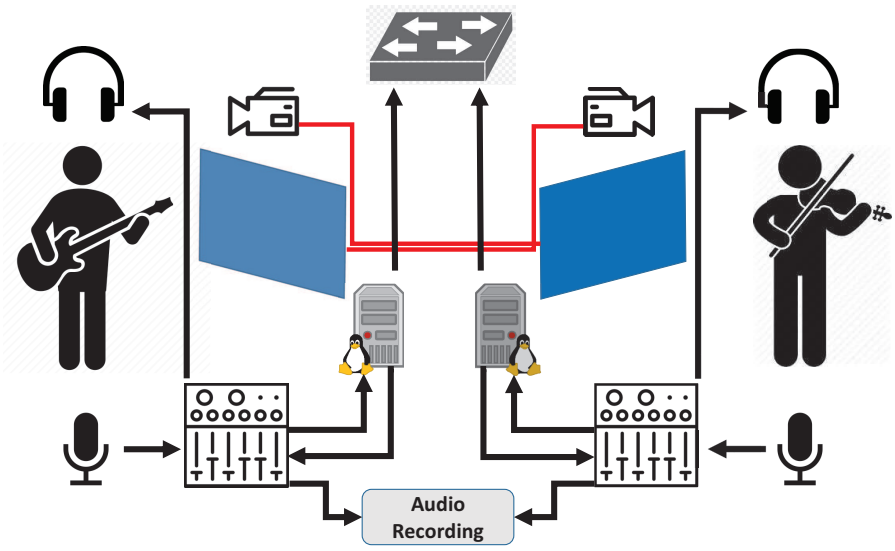

Figure 7.2: Experimental Setup for Scenario B (variable quality).

We then recorded with another smartphone's camera both the chronometer and its reflected image, and analyzed the video in a video editor, finding out that the round trip delay was 30 ms, therefore the one way delay for the video was 15 ms.

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 added two AD-340 audio delay boxes by Audio Research to the signal path, one in each direction, via which we were able to set the audio delay in each direction to the desired value at a very fine grain.

In Scenario B, the setup was changed as shown in Figure 7.2. The audio signals from the mixing consoles were fed to PCs running Linux with i7 processors and 12 GBytes of RAM, where our own software (Aretousa) sent and received the audio streams. We used Aretousa to manipulate the audio sampling rate, hence altering the audio quality; we did not compress the audio signal.

To avoid the audio issues that we noted during the pilot with Aretousa at the lowest sampling rates, which were due to resampling operations in the libraries, we modified Aretousa to run directly over the PulseAudio layer, bypassing the GStreamer framework entirely. As GStreamer is not really optimized for delay (which is not as important for streaming applications as it is for NMP), this change reduced the delay induced by the software itself and solved the audio issues at low sampling rates; the capabilities of the software did not change.

The PCs in Scenario B were connected via the Fast Ethernet LAN of the building, with three Ethernet switches in the path between the endpoints; audio delay was measured to be about 10 ms in each direction, or 20 ms MM2ME, compared to 34 ms MM2ME in the pilot, due to the lowest latency induced by the Aretousa software. Since we used a sampling buffer size of 10 ms, our setup provided the absolutely minimum delay possible in a computer-based setup. The video setup was the same as in Scenario A (direct cable connection in analog mode), hence the delay was again 15 ms.

Essentially, the topology of Scenario A bypasses computing and networking equipment as far as possible. Although this may seem counterintuitive in a study of NMP, which takes place over a network using computers, it make sense from the viewpoint of measuring QoME. In Scenario A, we are interested in assessing the effects of audio delay, therefore the ability to finely control delay and use very low values (even lower than those expected in real life music performance) allow us to focus on the delay itself, rather than on the ability of the software and the network to minimize delays and their jitter. By accurately deriving the limits to human tolerance to delay in NMP, we can then go back to an actual setup with computers and networks and assess whether NMP is realistic with that setup, by measuring its delays.

In Scenario B on the other hand, we need to modify the audio, hence the need to employ Aretousa to control the sampling rate. After the optimizations in Aretousa though, we measured one way delays of 10 ms in a Fast Ethernet topology without quality issues (e.g., clicks from missing sample buffers or dropped frames). This is close to the delay for musicians performing in the same room; to be exact, this is the delay two musicians would experience if they had a distance of 3.43 m between them. Therefore, with this setup we could concentrate on assessing the effects of audio quality, without paying any more attention to delay.

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 7.1: 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 7.2: Scenario B: Sampling rates.

7.1.2 Experimental Procedure

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 7.1 shows the delays used. Then, the musicians performed the same musical piece ten (10) more times

using a different audio sampling rate; Table 7.2 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; the questionnaire was the same for each repetition, but depended on whether we modified the audio delay or the sampling rate (see Section 6.1).

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 7.1 and 7.2. The main goal was to conduct an experiment that would allow us to evaluate multiple variables without bias or noise in the answers.

The MM2ME delay values used range from 0 ms to 120 ms (equivalent to 0 ms to 60 ms in one direction). Considering that sound travels 3.43 m in 10 ms, two musicians located in the same room would expect an MM2ME delay of 5 ms to 10 ms; an orchestra on a large stage could expect much higher delays, hence the need for a conductor providing a visual synchronization aid. We therefore tested a range of higher MM2ME delays to see until which point the QoME was still acceptable, but also lower delays to see how delay was perceived by the musicians. It should be noted that the delays are quite different than in the pilot study, where the range was 34 ms to 114 ms; the removal of computing and networking equipment from the path, allowed us to use far lower delays than what was possible in the pilot study. On the other hand, the highest delay did not change significantly, as evidence from the pilots showed that 114 ms was already too high for synchronization in an NMP context. Finally, seeing the changes in QoME at around 60 to 70 ms, we added the 80 ms delay as an intermediate step before the highest delay of 120 ms.

For audio quality, we tested some standard sampling rates for digital audio, including those for voice telephony (8 kHz), CD Audio (44.1 kHz) and Digital Audio Tape (48 kHz); we added 88.2 kHz which is considered studio level quality, and various sampling rates between voice telephony and CD Audio. In all cases we used 16 bit samples with linear sampling (as in CD Audio, but unlike regular voice telephony where 8 bit samples with logarithmic sampling are used). For single channel audio, these produce bitrates from 128 kbps (at 8 kHz) to 1.411 Mbps (at 88.2 kHz), an order of magnitude difference. The range was practically the same as in the pilot study, since these sampling rates are more or less standardized in the audio world. Note however that in this scenario the delay was lower than in the pilot study (20 ms rather than 34 ms MM2ME), thus ensuring that delay was not an issue during the tests.

7.2 Subjective Evaluation Design

Based on the results of the pilot study, which used an extended questionnaire with four pairs of participants, we refined the questionnaire for the main study reported in this

chapter, which involves eleven pairs of participants. The questionnaire was designed so that it could be easily answered after each individual NMP session. Each musician simply had to choose a score in a 5 point Likert scale for each question by touching a "button" on a tablet. We used the same scale in all questions, unlike in the pilot, where two questions were graded on a scale from -3 to +3, and the rest on a scale of 1 to 5, to make the questionnaire more uniform. We also rephrased the questions so that 1 and 5 would mean similar things in all cases (1 meaning less and 5 meaning more), in order to avoid confusing the participants.

To keep the questionnaire short, we used a different set of questions depending on the scenario, that is, we used questions more relevant to delay in Scenario A and questions more relevant to quality in Scenario B. Since each questionnaire had to be filled in after every repetition, we wanted to avoid boring the musicians, as this leads to more random answers. For the same reason, we started with the delay experiments, where we had more questions to ask, and continued with the quality experiments, where the questionnaire was much shorter.

We elaborate upon the questions in the remainder of this section, dividing them into three parts: questions common to both Scenarios, questions used only in Scenario A and questions used only in Scenario B. We explain why new questions were added or existing ones modified, and close with an explanation of why some questions were dropped entirely.

7.2.1 Questions common to both Scenarios

Evaluate your satisfaction. The *Perception of Satisfaction* (PoSat) is basically the MOS metric which is widely used to evaluate Quality of Experience in subjective studies. As satisfaction is a very complex phenomenon, in this study we complement this metric with many other subjective variables. We used a 5 point Likert Scale (higher is better) for this variable. As the MOS metric makes sense in both scenarios, the question was asked for both delay and quality variations.

Did you feel anxiety? and *Did you feel irritation?* As many musicians are not keen with technology, there is a possibility that anxiety and irritation may emerge during NMP sessions. Anxiety may be a result of unfamiliarity with the equipment used for NMP, while high audio delay or poor audio quality may irritate the musicians. We investigate the existence of these phenomena using a 5 point Likert scale (higher is worse). These questions did not exist in the pilot questionnaires, they were added after performing a more extensive literature survey that revealed that musicians often felt uncomfortable with the unfamiliar setups of NMP. We used these questions on both scenarios, since the setups used were different, hence they could have different effects on the performers.

7.2.2 Questions only for Scenario A

Evaluate the degree of delay you perceived. The *Perception of Audio Delay* (PoAD) variable asks the participants to grade the delay in a Likert scale from 1 (no delay) to 5 (too much delay). Even though we controlled the delay ourselves, it is important to understand how musicians perceive delay during their performance, as it reveals how perception of this particular aspect is related to the ground truth.

Evaluate the degree of synchronization. Achieving synchronization is a critical issue in NMP, with past work indicating its strong dependence on delay and the extremely low tolerance of musicians to delay. For the *Perception of Synchronization Degree* (PoSD) variable, musicians provided responses in a Likert Scale from 1 (cannot synchronize at all) to 5 (can fully synchronize).

To what degree did you follow your partner? With the *I was Trying to Follow my Partner* (TTF) variable, we examine whether a musician tries to follow her partner's tempo or not. Our previous work has indicated that as musicians find it harder to synchronize with increasing delay, they try to follow their partner. This variable is assessed in a scale from 1 (not at all) to 5 (I followed a lot). It should be noted that we modified the question and the scale compared to the pilot study (the original question was "I tried to follow my partner" and the response was from -3, fully disagree, to +3, fully agree), to avoid the different scale and the need to explain it to the musicians.

Did you focus on audio or video? The use of visual contact is an aspect that needs to be examined in NMP [71], as musicians often use visual cues for synchronization during actual performances. We used an ultra low delay camera/monitor setup and asked the musicians whether they mostly focused on audio or video contact; recall that the video delay was fixed and only the audio delay was varied. This variable is assessed on scale from 1 (only video) to 5 (only audio). This is also a new question, prompted again by the more extensive literature survey. Evaluating the contribution of video was only made possible by the low delay video setup used in the main study; the setup of the pilot also used cameras and monitors, but the video feeds exhibited very large delays due to the use of video compression, making video less useful for synchronization.

7.2.3 Questions only for Scenario B

Evaluate the audio quality. For the *Perception of Audio Quality* (PoAQ) variable, musicians were asked to evaluate in a 1 to 5 Likert scale the perceived audio quality (higher is better). Again, although we controlled the audio quality ourselves, it is important to understand how musicians perceive the quality of the audio, depending on the sampling rate used. As mentioned in Chapter 5, in earlier experiments we found that participants sometimes confused audio delay and audio quality.

7.2.4 Questions removed after the pilot study

The *Perception of Clicks* question was dropped, since in Scenario A we did not use any computing or networking equipment that could alter the audio, while in Scenario B the optimizations made to Aretouse allowed perfect audio transfer at a low delay. In a sense, the *Did you feel Irritation?* question replaced the perception of clicks, as it covered any irritating phenomena rather than specifically audio issues.

The *Perception of Musical and Emotional Expression* and the *Perception of my Partner's Performance* questions were dropped, as in the pilot study we found that the answers were not substantially affected by the delay and quality variations, hence making them redundant. We preferred to replace them with the *Did you feel Irritation?* and *Did you feel Anxiety?* questions, issues that we had not explored previously. Finally, the addition of the *Did you focus on audio or video?* question allowed us to see whether visual or aural contact was more important, something not explored in the pilots.

Table 7.3: Performance details for each duet (duets 1–6).

Duet No	1	2	3	4	5	6
Genre	Folk	Folk	Rock	Rock	Funk	Funk
Instrument A	Piano	Piano	El. Guitar	El. Bass	Organ	El. Bass
Instrument B	Santouri	Oud	El. Guitar	El. Guitar	El. Guitar	Percussion

Table 7.4: Performance details for each duet (duets 7–11).

Duet No	7	8	9	10	11
Genre	Rock	Rock	Classic	Folk	Folk
Instrument A	El. Bass	El. Guitar	Flute	Ac. Guitar	Lute
Instrument B	Ac. Guitar	Violin	Violin	Bouzouki	Violin

7.3 Evaluation Results

We conducted experimental sessions with 22 musicians (11 pairs). The musicians performed with a variety of instruments, including piano, acoustic guitar, electric guitar, electric bass, violin and flute, as well as traditional instruments including the lute, toumberleki, santouri and oud. In Tables 7.3 and 7.4 we first show the music genre of the piece performed by each duet, and then we show the instrument played by each musician.

We present below graphs with the results for both experimental scenarios. In each figure, we show dots representing individual answers for each MM2ME delay value for Scenario A and each sampling rate value for Scenario B; the size of the dots corresponds to the number of answers for each delay or sampling rate.

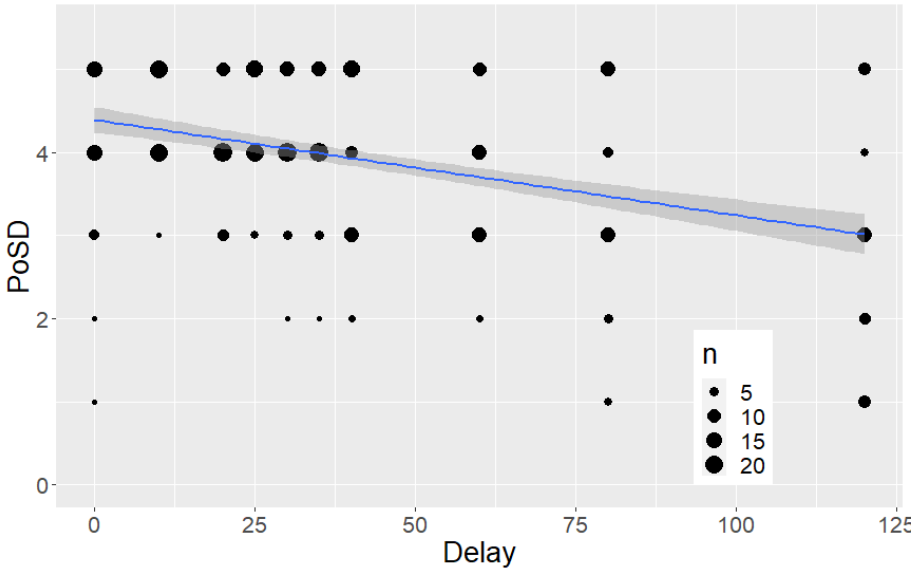

Figure 7.3: Perception of Synchronization Degree (N = 22, R^2 = 0.9002).

We also performed linear regression of the data values against the delay or quality, and show the results with a blue line; a shaded area indicates the 95% confidence intervals for the linear regression, that is, the area within which the true regression line lies. We also report for each figure the R^2 measure for the regression line, which indicates how much of the variation of the dependent variable is explained by the independent variable. This is often interpreted as a percentage, so a value of 1 means that the regression model explains 100% of the variation (it is a perfect fit), while a value of 0.5 means that the regression model explains only 50% of the variation. All graphs were created using RSTUDIO. Unless otherwise indicated, the graphs reflect results from the 22 participants mentioned above.

7.3.1 Scenario A: Variable Audio Delay

Figure 7.3 shows the answers to the *Perception of Synchronization Degree* (PoSD) question against the MM2ME delay. As shown in the graph, the fitted line has a clear negative slope as delay increases, starting at 4.5 and ending at 3; it seems that musicians perceived that they could not synchronize with delays of more than 80 ms, rather than the 50–60 ms reported in the literature (equivalent to 25–30 ms one way), since the average scores are 3.75 at 60 ms and 3.5 at 80 ms. The regression line is a good fit for the data, explaining 90% of the variation.

Figure 7.4 shows the results for the *Perception of Audio Delay* (PoAD) variable. A slightly increasing slope is observed, with average scores ranging from 1.5 to 2.5. Again, the differences between MM2ME delays of 60 ms and 80 ms were very small (slightly less and slightly more than 2, respectively), indicating that one way delays of 40 ms are quite acceptable. When we focus only on the sessions where pianists were

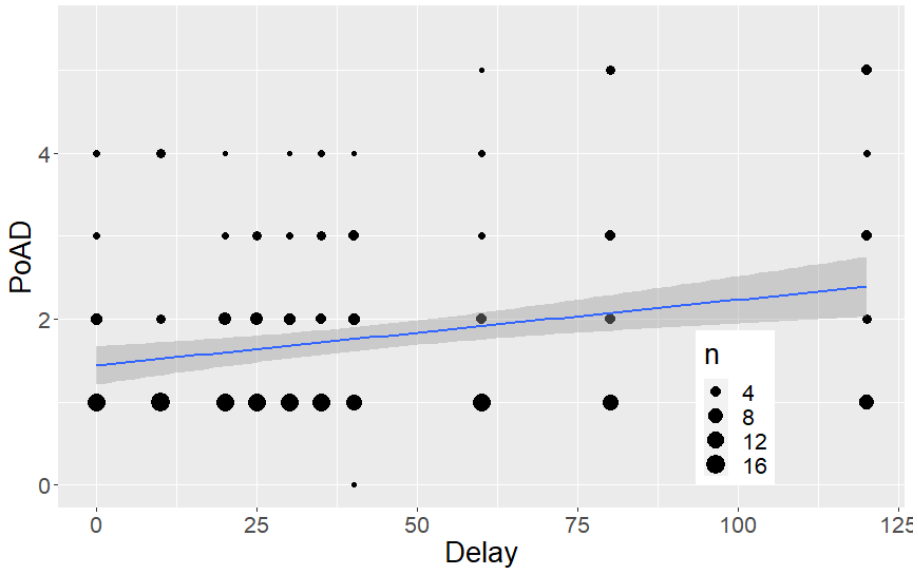

Figure 7.4: Perception of Audio Delay (N=22, R^2 = 0.8657).

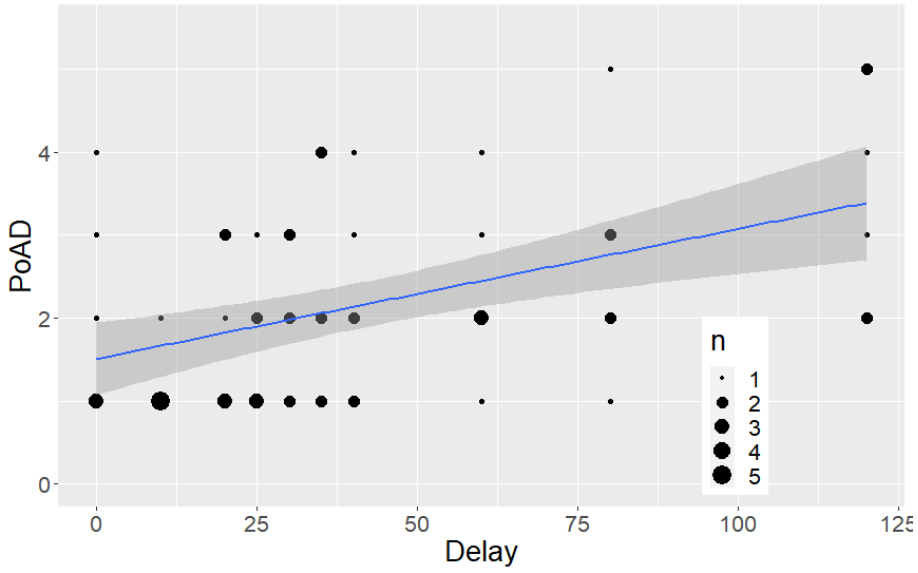

Figure 7.5: Perception of Audio Delay (Pianists and partners, N=6, R^2 = 0.8122).

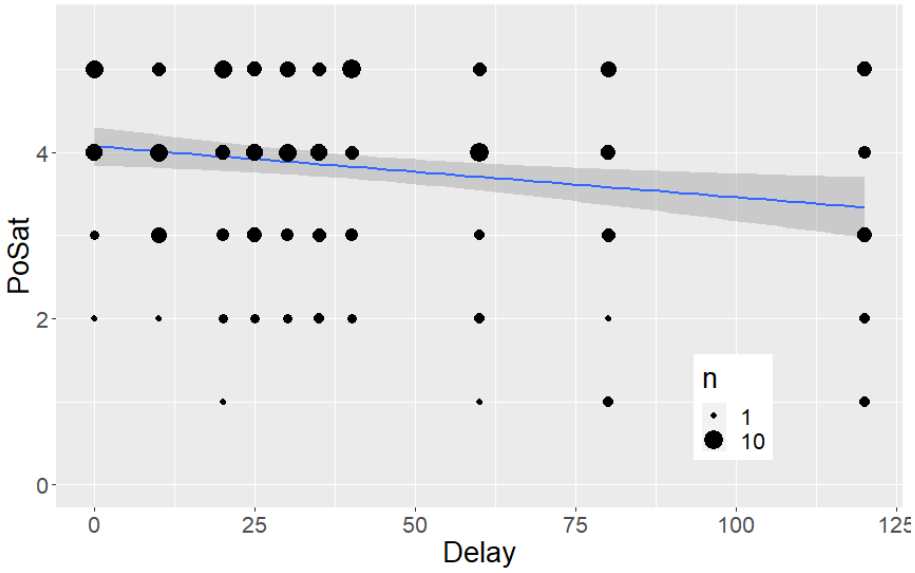

Figure 7.6: Perception of Satisfaction (N=22, R^2 = 0.8843).

involved (duets 1, 2 and 5), the responses from the pianists and their partners (6 participants), shown in Figure 7.5, indicate a higher rate of growth on the average scores, from 1.5 to 3.5. This observation is evidence that perception of delay depends on the setting; specifically, when pianists are involved, the performances seem to be more sensitive to delay. The regression line is a good fit for the entire data, explaining 86% of the variation, but it is slightly worse for the pianists and partners only, explaining 81% of the variation.

Maybe the most critical variable in our experiment is the *Perception of Satisfaction* (PoSat), which is basically the MOS metric. As shown in Figure 7.6, the fitted line has a small negative slope, starting from slightly above 4 and ending at around 3.5, again with a very small difference between 60 ms and 80 ms. When we focus again on pianists and their partners, Figure 7.7 shows that they were more influenced by increasing delay, as the slope of the line is steeper, starting from 4.75 and ending at slightly less than 3. Note also that while with all musicians the answers were widely spread for all delay values, pianists and their partners gave a more narrow range of answers at lower delay values, which grows as delay increases. The regression line is a good fit both for the entire data, explaining 88% of the variation, and for the pianists and partners only, explaining 89% of the variation.

In Figure 7.8 we can see the results for the *I was Trying to Follow my partner* (TTF) question. The fitted line has a similar positive rate to the perceived audio delay in Figure 6.4, but starting from 2.75 and ending at 4. For delays of up to 60 ms we have very similar (and wide) result ranges, indicating that the musicians are split between leading and following, as would be expected in a non NMP scenario. With the highest delay of 120 ms though, the fitted line tends to reach 4, showing that synchronization

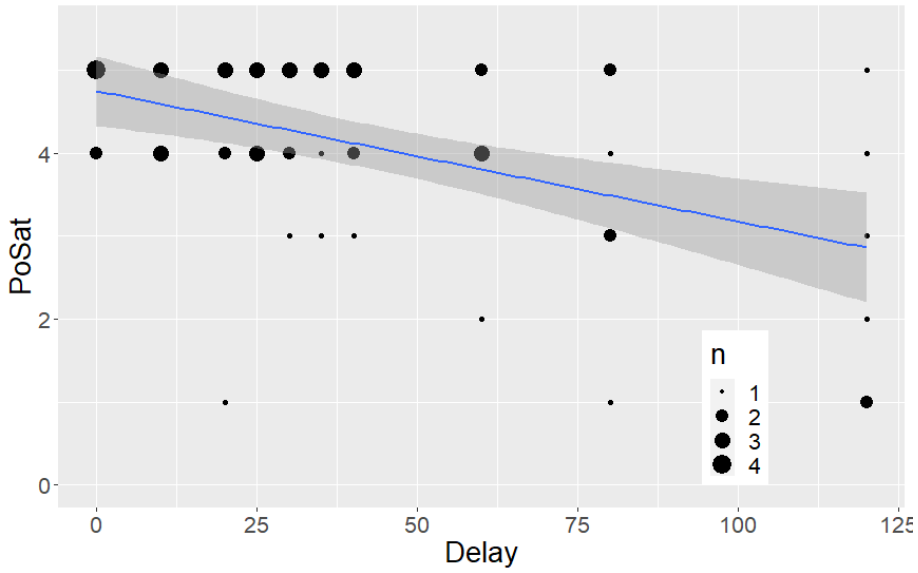

Figure 7.7: Perception of Satisfaction (Pianists and Partners, N=6, R^2 = 0.8909).

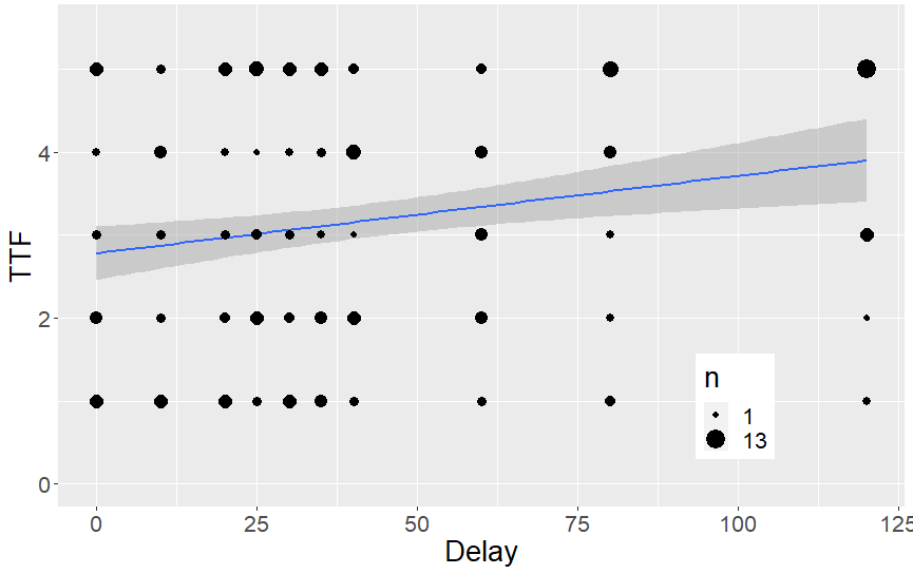

Figure 7.8: I was Trying to Follow my partner (N=22, R^2 = 0.8694).

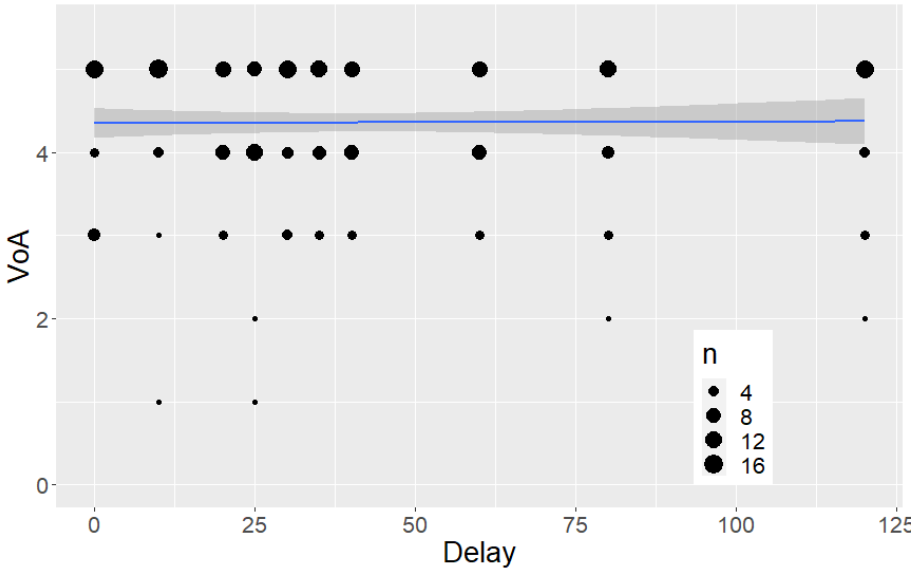

Figure 7.9: Focus on audio or video (N=22, R^2 = 0.0032).

is harder and each musician tries to follow the other one. The regression line is again a good fit for the data, explaining 87% of the variation.

Regarding the emphasis on audio or video, Figure 7.9 shows the answers to the *Did you focus on audio or video?* question. The results indicate a strong preference to audio contact, with the fitted line almost horizontal at 4.25 (mostly audio to only audio). This was despite the fact that the one way video delay was fixed to a low 15 ms, while the audio delay increased up to 60 ms (120 ms MM2ME). This indicates that musicians are mostly based on aural and not visual cues for synchronization and amplifies previous findings related to sensorimotor synchronization and its preference to aural cues. The regression line does not explain the variation, as the line is practically horizontal.

Figure 7.10 shows the results for the *Did you feel anxiety?* question, indicating that anxiety was non existent, with the fitted line is almost horizontal at 1.25; the regression does not explain the variation, as the line is flat. This is complemented by Figure 7.11 which shows the results for the *Did you feel irritation?* question, showing that irritation was non existent (1) to very low (1.75), even at the highest delay values. The regression line is a good fit for the data, explaining 98% of the variation. These results indicate that the participants felt comfortable and did not find the NMP experience frustrating. This allows us to place more trust on the other variables, as our previous work has indicated that when participants are uncomfortable with their NMP experience, they tend to provide lower scores to unrelated questions.

To test whether our results have statistical significance, we performed ANOVA for repeated measures with delay as the independent variable and *Perception of Synchronization Degree* (PoSD), *Perception of Audio Delay* (PoAD), *Perception of Satisfaction* (PoSat) and *I was Trying to Follow my partner* (TTF) as the dependent variable; we did not test

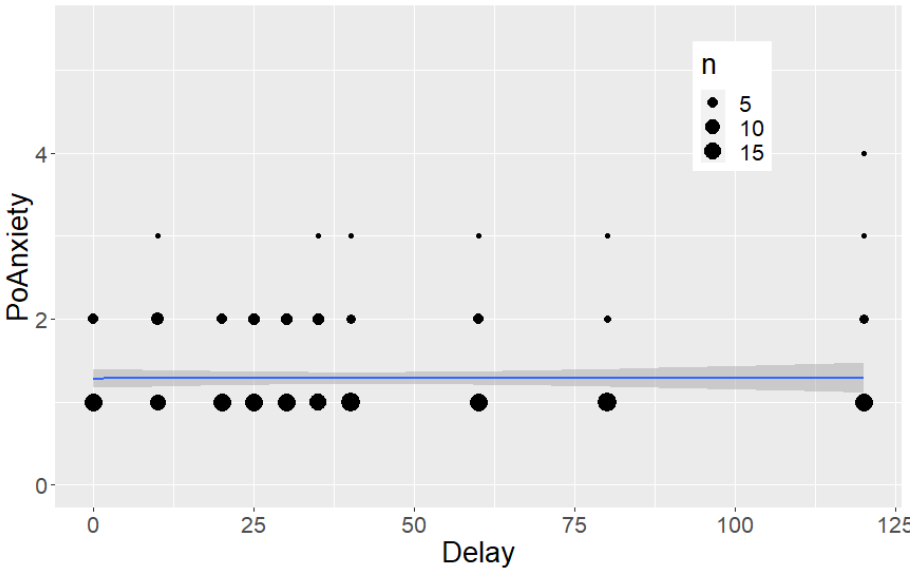

Figure 7.10: Perception of Anxiety (N=22, R^2 = 0.0887).

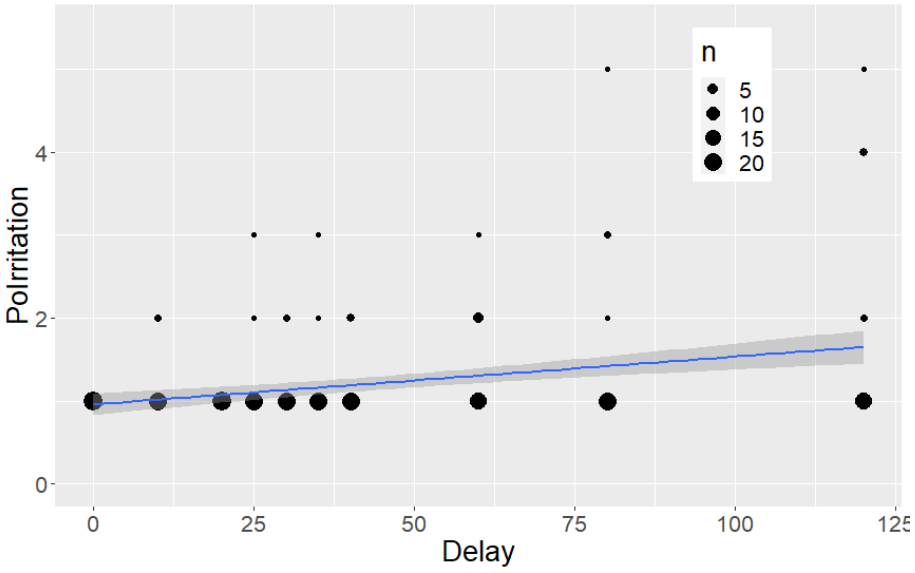

Figure 7.11: Perception of Irritation (N=22, R^2 = 0.9817).

Dependent Var	PoSD	PoAD	PoSat	TTF
Independent Var	delay	delay	delay	delay
Sample Size	22	22	22	22
p = 0.05	0.001	0.013	0.819	0.002

Table 7.5: ANOVA analysis: Delay vs. Subjective Results.

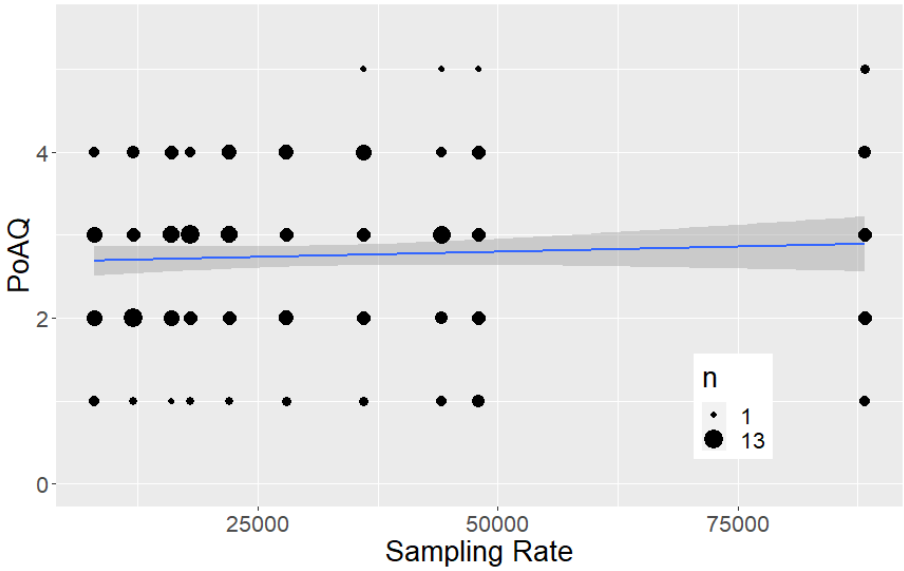

Figure 7.12: Perception of Audio Quality (N=22, R^2 = 0.2585).

the *Did you focus on audio or video?*, *Did you feel anxiety?* and *Did you feel Irritation?* variables, as it was clear that they were not influenced by delay. Table 7.5 shows the results for the entire set of 22 participants. Most of the p values are lower than 0.05, indicating a strong probability of correlation with delay. The only exception is the PoSat variable, which may be due to the fact that the Perception of Satisfaction was quite high even for the highest delay tested (see Figure 7.6).

7.3.2 Scenario B: Variable Audio Quality

The study of the effects of audio quality was a secondary issue in our work, since we only wanted to assess how much quality loss musicians can tolerate in NMP without reducing their QoME. For this reason, the audio quality tests followed the audio delay tests and used a shorter questionnaire. Our goal was to test how much we can shrink the bit rate by reducing the sampling rate, before QoME drops. The alternative would be to resort to audio compression and decompression, which inflate delay, thus also reducing the QoME, as seen from the delay tests.

Figure 7.12 shows the results for the *Perception of Audio Quality* (PoAQ) question against the sampling rate. The fitted line has a slightly positive slope as sampling rate increases, being slightly under 3 for all sampling rates, indicating that the musicians

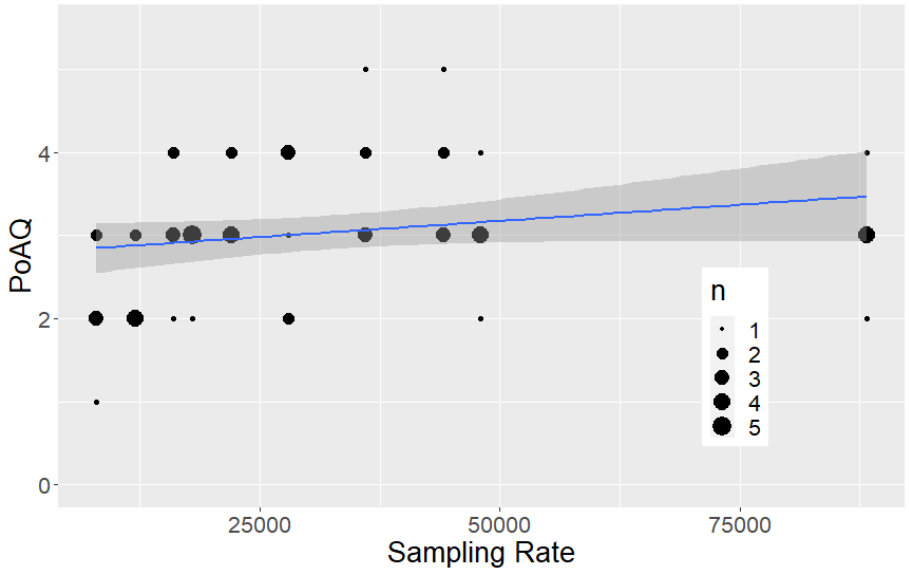

Figure 7.13: Perception of Audio Quality (Pianists and partners, N = 6, R^2 = 0.1394).

did notice a quality improvement, but it was not very significant. When we only focus on pianists and their partners, as we did in the previous section, Figure 7.13 shows that the fitted line has a more prominent positive slope, indicating that pianists and their partners were more aware of the variance in quality. The regression lines are not a good fit for the data, explaining only 26% of the variation for the entire set and only 14% for pianists and partners.

Figure 7.14 shows the results for the *Perception of Satisfaction* (PoSat) question, which is again the MOS metric, but depending on the sampling rate this time. The fitted curve has a positive slope as the sampling rate is increased, but again it is not too prominent, ranging from 3.25 to 3.5. Interestingly, when we focus on pianists and their partners, the slope for this variable is nearly zero, as shown in Figure 7.15. Again, we can conclude that the audio quality did not have a sizable effect on the musicians. The regression lines explain even less of the variation than with PoAQ (14% for the entire data set, 1% for pianists and partners only).

To ensure that musicians were not distracted by the setup, we asked again whether they felt anxiety (Figure 7.16) and irritation (Figure 7.17) during the sessions; the curves have a very slight negative slope, but the average results are always close to 1, indicating no anxiety and irritation during the experiments. The regression line can only explain some of the variation (54% for Anxiety and 11% for Irritation).

Dependent Var	PoAQ	PoSat
Independent Var	quality	quality
Sample Size	22	22
p = 0.05	0.371	0.48

Table 7.6: ANOVA analysis: Quality vs. Subjective Results.

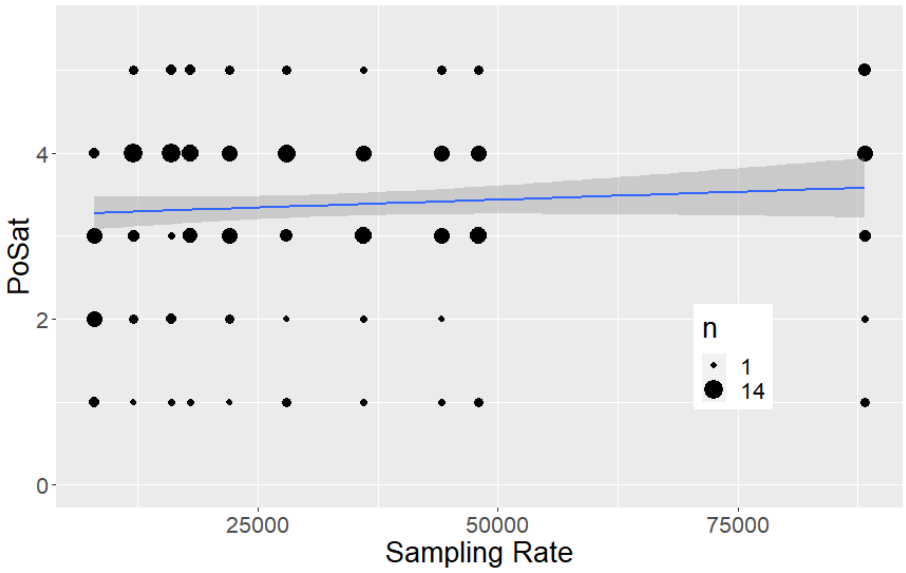

Figure 7.14: Perception of Satisfaction (N=22, R^2 = 0.1391).

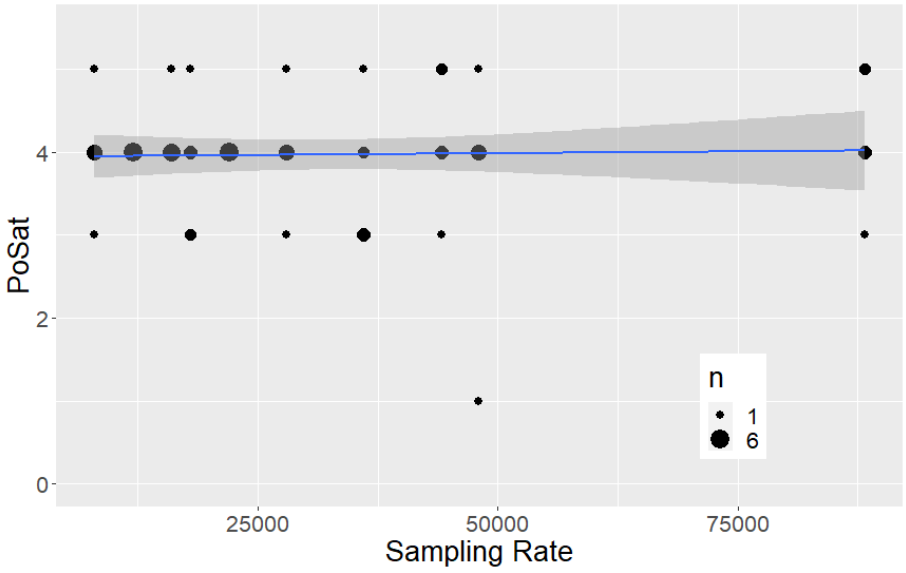

Figure 7.15: Perception of Satisfaction (Pianists and partners, N = 6, R^2 = 0.0117).

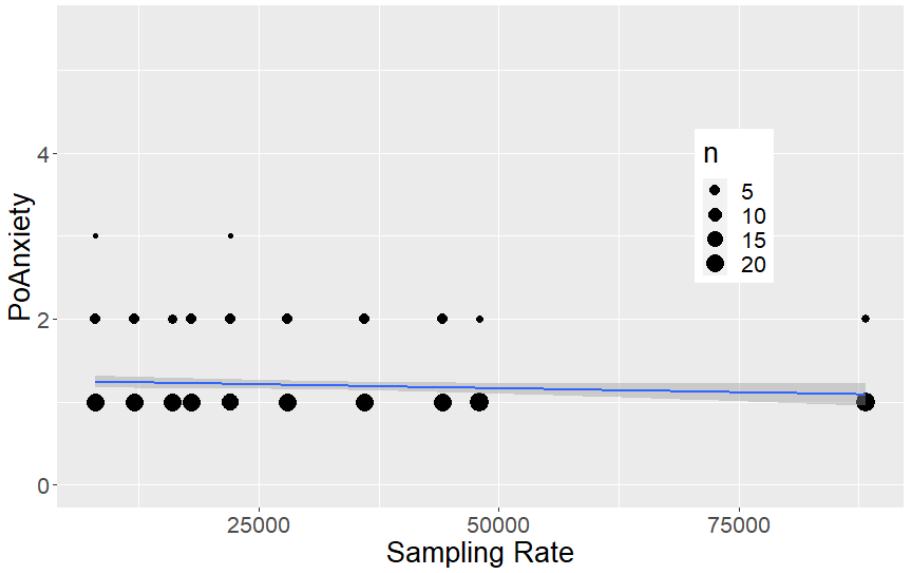

Figure 7.16: Perception of Anxiety (N=22, R^2 = 0.5381).

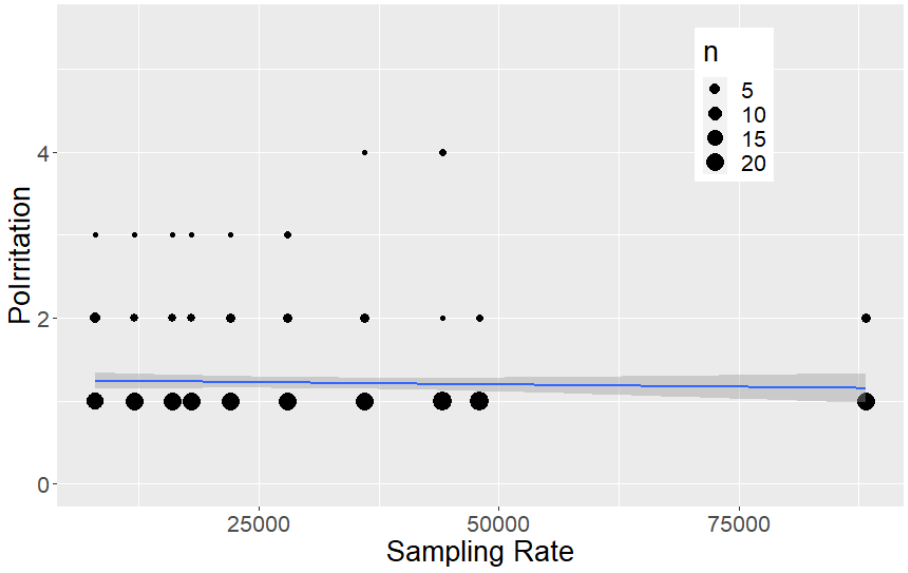

Figure 7.17: Perception of Irritation (N=22, R^2 = 0.1128).

To round up the evaluation, we performed ANOVA for repeated measures with quality (i.e., sampling rate) as the independent variable and *Perception of Audio Quality* (PoAD) and *Perception of Satisfaction* (PoSat) as the dependent variable; again, we did not test the *Did you feel anxiety?* and *Did you feel Irritation?* variables, as it was clear that they were not influenced by quality. The ANOVA test shown in Table 7.6, did not expose statistical significance for the dependent variables PoAQ and PoSat, with the p value being greater than 0.05 in both cases. This was to be expected from the graphs, since the slopes in both cases are very small, indicating the audio quality can be sacrificed in NMP without reducing the satisfaction of the participating musicians.

7.4 Summary of Results

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. In the experiments reported in this chapter, 22 musicians participated as pairs, playing a diverse set of musical instruments, constituting the largest NMP study with actual musical performances that we are aware of.

The results indicate that even though increasing delay does have a statistically significant effect on the QoME of the participating musicians, the range of acceptable delays is larger than previously reported based on hand clapping experiments. The musicians participating in our study considered the performances to be synchronized and satisfactory with one way (M2E) delays of up to 40 ms (or two way, MM2ME, delays of up to 80 ms). This means that the acceptable delay threshold is closer to 40 ms over a wide range of instruments and musical pieces, rather than the 25-30 ms widely cited in the literature. However, the results can vary depending on the context: in the duets where pianists were involved, delay had a more negative effect on the results.

We should note that our variable delay study was unusual in that our technical setup ensured perfectly fixed delays. In a realistic NMP scenario, the unpredictability of the computers and, more importantly, the network path, cause constant, and sometimes dramatic, delay fluctuations. By using a buffer at the receiver we can even out these fluctuations, at the cost of increasing the overall delay. We can therefore consider our results for a specific delay value to be valid for an NMP session where the delay *after buffering* is equal to our tested delay value.

On the other hand, reducing quality by lowering the sampling rate did not have a statistically significant effect on the QoME of the musicians, even though we considered an order of magnitude reduction (from 88.2 kHz to 8 kHz). This also held for the smaller group of pianists and their partners. This is important as it implies that when the available bitrate is limited, we can reduce the sampling rate to save bandwidth with no appreciable loss in QoME, rather than introduce an audio codec and its resultant delay.

For the audio quality study, we should note that although we used a wide variety of instruments, we did not test everything; it is possible that with some other instruments, the participants may be more sensitive to quality. On the other hand, in an NMP scenario it is conceivable that the participants may be more deterred by delays or audio clicks, rather than by the degradation of quality caused by lower sampling rates.

Chapter 8

Tempo Analysis

The subjective analysis of the questionnaires gathered during our main NMP study, as well as the responses from the pilot study, indicated that not only different musicians perceive the same conditions in quite different ways, even the responses from the same musicians are not always consistent with the underlying parameters; for example, their assessment of delay does not always follow the actual delay in the experiments. The results from these subjective evaluations thus exhibit a high variance, which makes drawing concrete conclusions harder. Specifically, the subjective analysis for Scenario A of the main study where delay was varied, indicated that the QoME of the musicians did not drop significantly when the MM2ME delay grew from 60 to 80 ms (or, from 30 to 40 ms one way). The question arises, however, whether musicians can actually synchronize at this delay setting. In contrast, the results from Scenario B of the main study where quality was varied, were more conclusive and did not warrant further exploration.

Having recorded audio from all the experiments, we decided to examine whether the performers could reach and maintain a steady tempo during their performances, by looking at the evolution of the tempo during each performance. Previous studies of tempo in NMP relied on hand claps, which have a simple audio signature, making it easy to note how the tempo evolves by simply looking at the waveform of the recordings. With real musicians however, this is not possible. Even worse, since each duet selected their own musical piece and tempo, we did not even know what the intended tempo was, so we had to recover all relevant information from the actual recordings, unlike in the study by Rottondi et al. [71] where the intended tempo was known. For this reason, we used a signal analysis toolkit to detect, as far as possible, the tempo of the performances, using only the recorded audio.

In this chapter we first describe the tools that we used to extract tempo measurements from the audio recordings of Scenario A (variable audio delay) of the main study. We then discuss the results of this analysis, in order to see how it correlates with delay, and then summarize our findings, with a focus on the verification of the findings of the subjective study, in terms of the maximum acceptable delay limits for NMP.

8.1 Method of Analysis

We analyzed the audio recordings using the MIRToolbox [55]. To determine the tempo at a period of time, we start with the *event density*, which estimates the average number of note onsets per second as follows:

$$E = \frac{O}{T} \tag{8.1}$$

where E is event density, O is the number of note onsets and T is the duration of the musical piece. The MIRToolbox estimates how the music tempo, measured in *Beats per Minute* (BPM), varies over time, by detecting the note onsets via signal processing of the audio. The analysis is not perfect, as it depends on each instrument's sonic signature and manner of playing, but it is revealing, especially for instruments with very clear sonic signatures, for example percussive ones, or with performances where the instrument plays a rhythmic pattern. We performed this analysis for the audio recording of each side of each NMP performance.

These results are not easily amenable to numerical summarization, since musicians adapt their playing over time as they listen to each other; as a result, each performance leaves a unique time-varying imprint, and we have 220 of them (each of the 22 musicians performed their piece 10 times, while we varied the audio delay). However, when presented visually, they show interesting trends. The figures in the following section show how the tempo (in BPM) varies over time (in seconds) for different musical instruments; each figure shows one such curve for each delay value, corresponding to one performance by a single musician, with progressively lighter curves corresponding to increasing MM2ME delays.

To reduce visual clutter, we only show results at 40 ms intervals, that is, with 0, 40, 80 and 120 ms MM2ME delays. The rationale behind using only 4 out of the 10 delay values is that they represent very low delay (lower than what is natural in a music performance), reasonable delay (the delay of a moderately large room or studio space), high delay (specifically, the delay limit that seemed acceptable in the subjective study) and very high delay (the delay limit that seemed unacceptable in the subjective study). The instruments used by each duet and the music genre of the performance are given in Tables 7.3 and 7.4 in the previous chapter; we also note in each figure whether the musician played a rhythm or a solo part.

8.2 Evaluation Results

Figures 8.1 and 8.2 show the delay variation for each instrument of duet 1. We can see that with a delay of 0 ms (shown with the darkest line), which is unnaturally low (less than what would occur even in a small room), both musicians actually speed up their tempo in the first part of the performance, as reported in previous studies. As

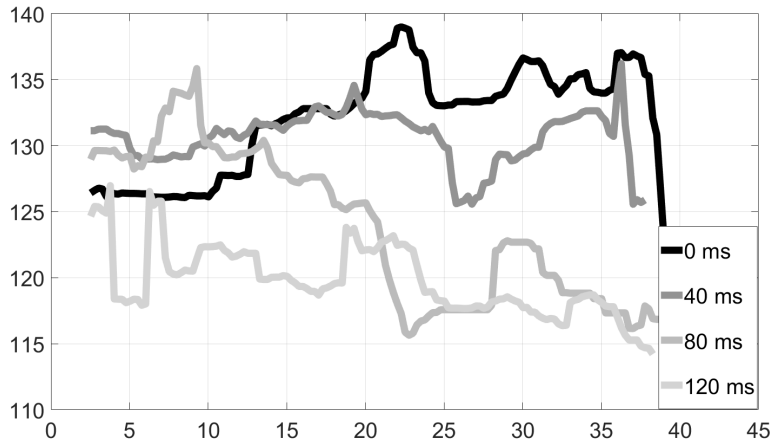

Figure 8.1: Tempo variation over time: Duet 1, Piano-Rhythm-Folk.

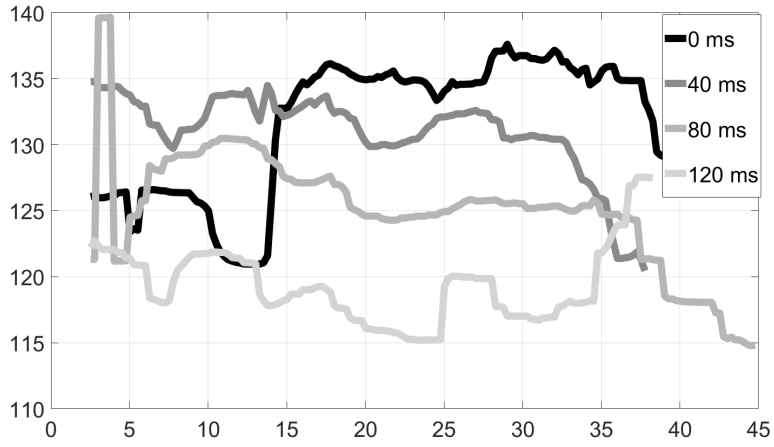

Figure 8.2: Tempo variation over time: Duet 1, Santouri-Solo-Folk.

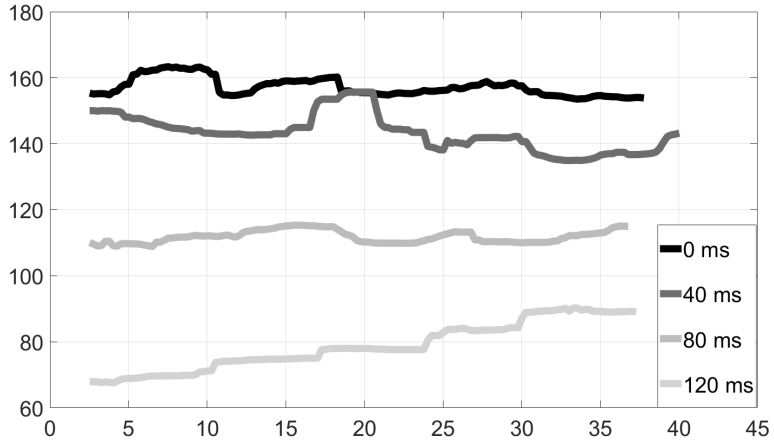

Figure 8.3: Tempo variation over time: Duet 2, Piano-Rhythm-Folk.

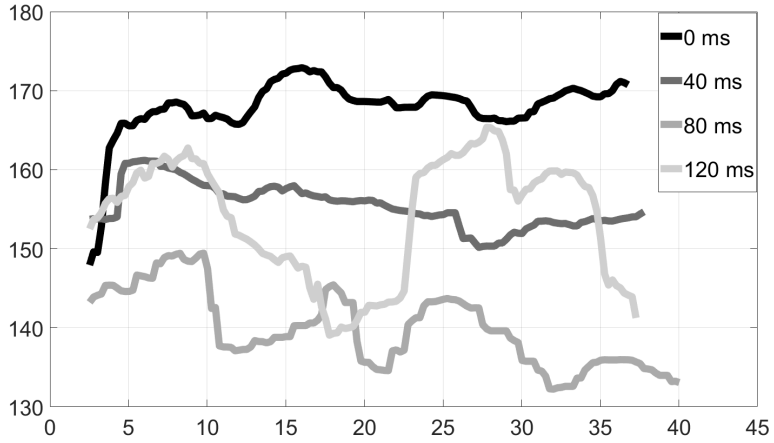

Figure 8.4: Tempo variation over time: Duet 2, Oud-Solo-Folk.

the delay grows, the tempo slows down, but the musicians have a hard time keeping a steady tempo at all delay values, as evidenced from the ups and downs in the curves.

While in duet 1 the musicians had trouble keeping a steady rhythm, in duet 2, Figures 8.3 and 8.4 show a different situation: the instrument playing the rhythm part, in this case the piano, is visibly affected by delay, since as the delay grows, the tempo drops; however, the tempo is steady in all but the highest delay value. The instrument playing the solo part however, in this case the oud, shows larger tempo variations, even though the tempo does generally drop with growing delay, and is most unstable with a delay of 120 ms. Of course, due to the method we are using to detect the tempo (note onsets), solo parts where musicians play more freely and improvise are harder to characterize precisely in terms of tempo. That is, their tempo may not have been recovered from the recording with perfect accuracy.

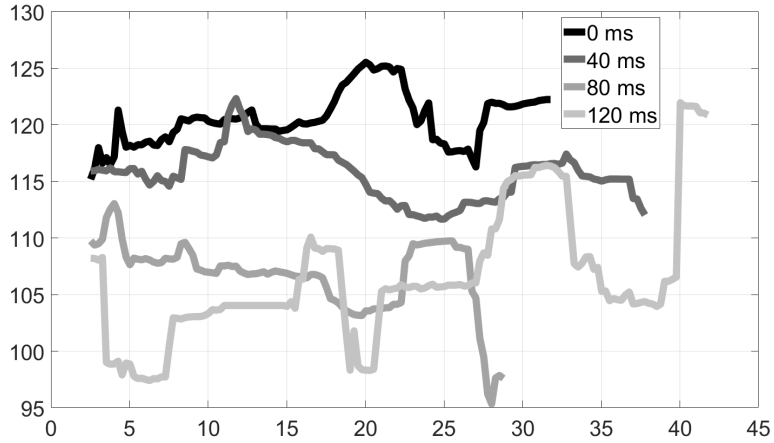

Figure 8.5: Tempo variation over time: Duet 3, Electric Guitar-Rhythm-Rock.

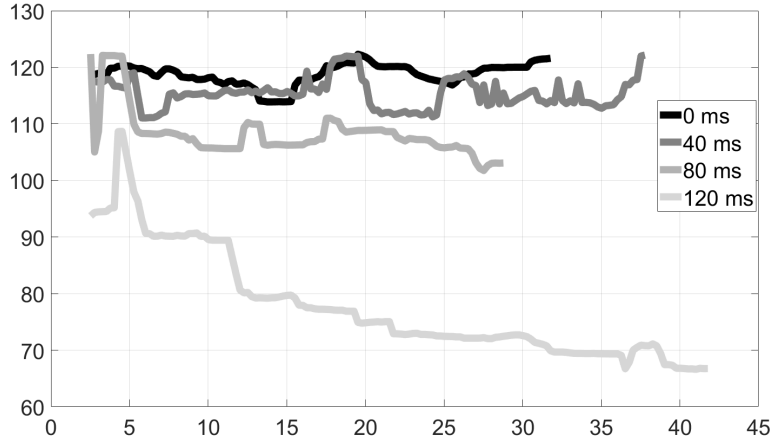

Figure 8.6: Tempo variation over time: Duet 3, Electric Guitar-Rhythm-Rock.

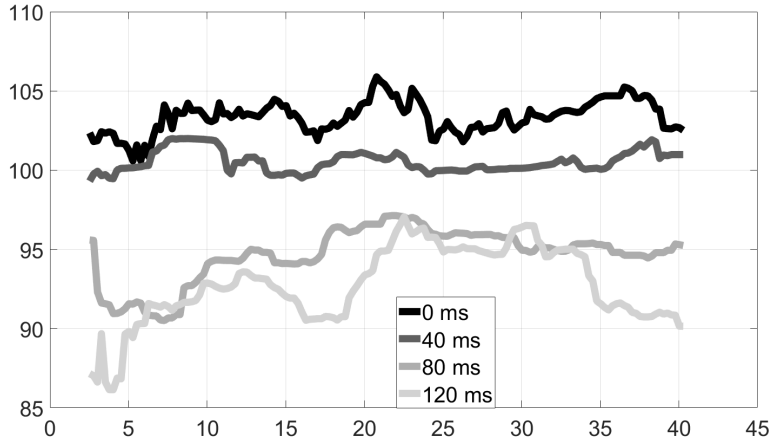

Figure 8.7: Tempo variation over time: Duet 5, Organ-Rhythm-Funk.

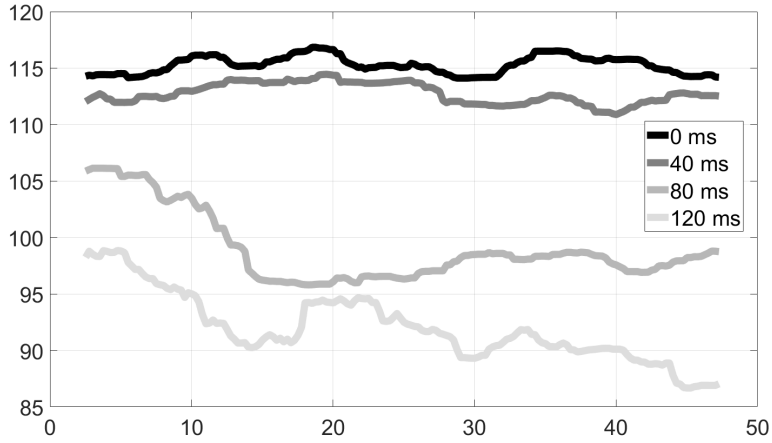

Figure 8.8: Tempo variation over time: Duet 6, Percussion-Rhythm-Rock.

In duet 3 where both musicians have a rhythm role, we can see in Figures 8.5 and 8.6 that both exhibit tempo variations, however, the musicians do manage to keep a relatively steady tempo, except for the highest delay value of 120 ms. Again, the tempo tends to drop with higher delays. Note that the performance ends at different time points for each delay value; they finish at the same time, of course, for each delay value.

The difficulty of keeping a steady tempo at higher delays is also apparent in Figure 8.7 which shows the rhythm instrument of duet 5 (organ); again, tempo drops with higher delays, and has wild variations at a delay of 120 ms. With the percussion instrument of duet 6, arguably the most rhythmic of instruments and the easiest in terms of automated tempo detection, as shown in Figure 8.8, the beat is noticeably slower for higher delays, and hard to keep steady when delay reaches 120 ms.

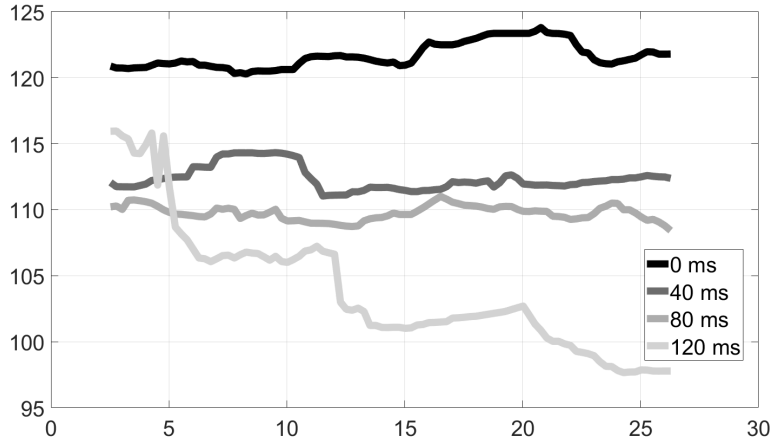

Figure 8.9: Tempo variation over time: Duet 7, Bass-Rhythm-Rock.

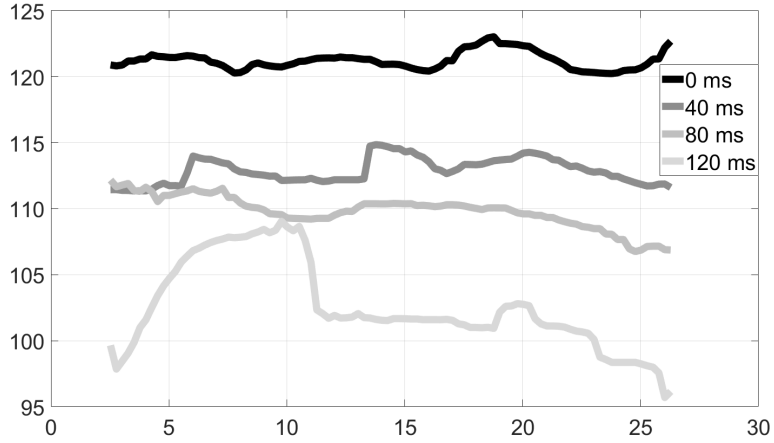

Figure 8.10: Tempo variation over time: Duet 7, Acoustic Guitar-Rhythm-Rock.

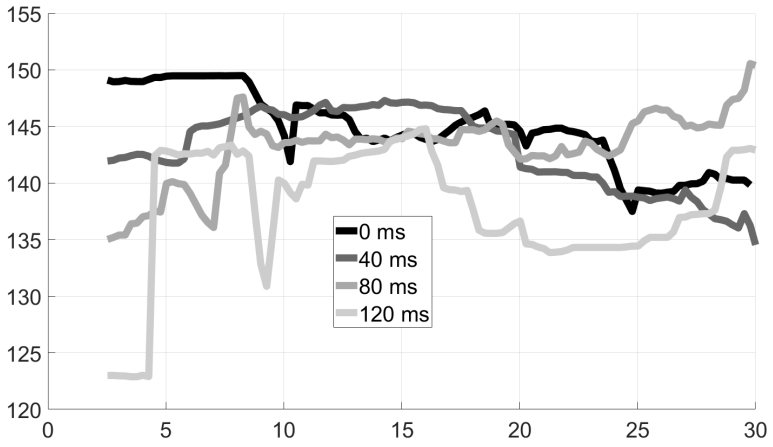

Figure 8.11: Tempo variation over time: Duet 8, Electric Guitar-Rhythm-Rock.

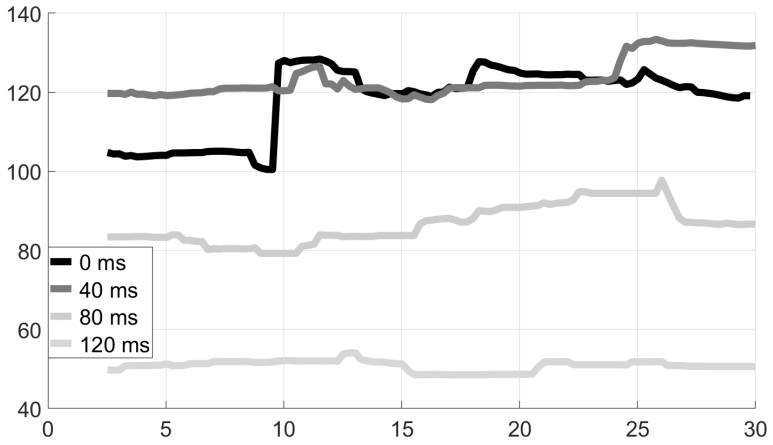

Figure 8.12: Tempo variation over time: Duet 8, Violin-Solo-Rock

There are also cases where both sides of a duet can keep the same rhythm, as with duet 7, shown in Figures 8.9 and 8.10: the rhythm is steady with delays of up to 80 ms; there is a very slight reduction in tempo from 40 to 80 ms, but at 120 ms the tempo either slows down continuously or varies wildly.

Duet 8 is unusual, in that the rhythm instrument (guitar), shown in Figure 8.11 has an unsteady tempo, while the solo instrument (violin), shown in Figure 8.12 has a very steady tempo, despite the visible slowdown at delays of 80 and 120 ms. The reason for this is the very different expertise levels of the musicians: the violinist was a 45-year-old professional musician, while the guitarist was a 23-year-old amateur one. Hence, the violinist's solo tempo was found to be more stable than the guitarist's, even though it was the guitarist who was supposed to keep a stable rhythm with the guitar. This is an indication that more experienced musicians may manage to partially compensate for delay by adapting their performance.

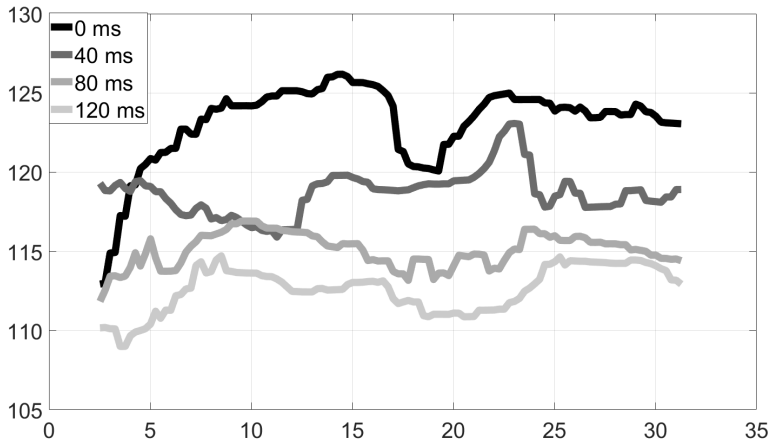

Figure 8.13: Tempo variation over time: Duet 10, Acoustic Guitar-Rhythm-Folk.

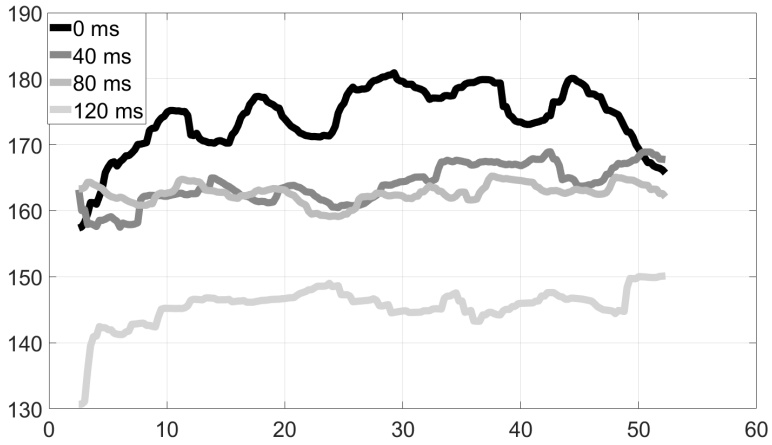

Figure 8.14: Tempo variation over time: Duet 11, Lute-Rhythm-Folk.

Finally, the rhythm instruments of duet 10 and duet 11, shown in Figures 8.13 and 8.14 further verify the previous observations of tempo speedup at the unusually low delay of 0 ms, good tempo stability at 40 and 80 ms, albeit at a slight reduction of tempo at 80 ms, and high variations at 120 ms.

From these figures, we can make the following general observations:

1. At the (unnaturally) low delay of 0 ms, musicians tend to speed up their tempo in the beginning of the session.
2. As delays rise beyond 40 ms, musicians adapt by slowing down the tempo of their performance.
3. Instruments performing rhythmic parts are more clearly affected by delay, as shown by their more visibly delineated curves.
4. Percussive instruments, which generally have a rhythmic role, are the most sensitive to delay.
5. In most cases, musicians manage to keep a steady tempo at delays of up to 80 ms.
6. At a delay of 120 ms performances break down, exhibiting either continuously slowing or wildly varying tempos.

These observations verify past work which found that musicians who perform percussive instruments suffer more from delay than others. Indeed, the hand clap experiments, where the rhythmic patterns are very clear, fall in the same category. Of course, these instruments, with their easy to detect sonic signatures and their clear temporal patterns, are ideal for this type of analysis. We can further observe that this is true for instruments having a rhythmic role in a duet. Although solo instruments seem to follow more irregular tempos, we must keep in mind that this may be an artifact of our audio analysis which relies on a steady production of note onsets; with improvisational parts, performers are expected to more often deviate from the base rhythmic pattern.

The most interesting observation however is that the limits to tolerance can vary considerably; most musicians could achieve a stable tempo at MM2ME delays of 80 ms, corresponding to a one way delay of 40 rather than 20–30 ms, higher than what was previously considered the limit to synchronization, even though this may come at the cost of a minor slow down in the performing tempo.

Finally, it should also be noted that we performed an ANOVA analysis for repeated measures of the average tempo scores for each session and for delays of 0, 40, 80 and 120 ms (MM2ME) and the p value was computed equal to 0.007 ($p < 0.05$). This indicates a strong statistical significance in the delay/Tempo/tempo relationship, that is, the calculated tempos were found to be statistically correlated with the delay values, in the sense that higher delays did lead to slower tempos.

8.3 Summary of Results

The analysis performed on the recorded audio indicates that musicians tend to slow down their tempo as delays grow, an effect made very clear with percussive instruments and quite clear with instruments playing rhythmic parts. However, in nearly all of the experiments they can synchronize and maintain a stable tempo with MM2ME delays of up to 80 ms (equivalent to 40 ms one way), indicating that the delay tolerance of actual musicians performing in NMP scenarios is higher than previously thought, that is, the EPT is closer to 40 rather than 20–30 ms. Indeed, musicians, especially more experience ones, try to adapt to higher delays by slowing down their tempo. This conclusion is aligned with the results of the subjective analysis of the QoME questionnaires for the variable delay experiments, reported in Chapter 7, which also indicated that one way delays of 40 ms can be handled by the musicians participating in an NMP session. We can therefore conclude, based on both analyses, that the EPT for real musicians performing actual musical pieces can be up to 40 ms.

Chapter 9

Audio Features Analysis

In Chapters 7 and 8 we established a clear correlation between delay and QoME in NMP and found that the acceptable delay limit for NMP is close to 40 ms one way, based on both subjective analysis and on tempo analysis. We also found evidence that these results depend on the instrument played, the role of the musician (rhythm or solo) and other features. In this chapter, we extract the audio features of the instruments for each of the performances in the main study for Scenario A (variable audio delay), following the methodology of Rottondi et al. [71], classify them into three ranges for each audio feature, and then attempt to see how delay influences the QoME metrics and the performance tempo for each group. We repeat the same analysis using the music genre and the musician's role to group the performances, rather than the audio features of the instruments.

Despite using the same basic methodology with [71], our study employs a larger set of participants, allowing the extraction of statistically significant results, with an extended QoME questionnaire and automated tempo measurements as the dependent variables, in an extremely accurate audio delay emulation environment, not subject to any jitter. In addition, we perform the same type of analysis by grouping performances based on the musical genre and the role of the musician to see how they influence the relationship of delay with QoME and tempo.

In the remainder of this chapter, we first explain the audio features extracted from the audio and the methods we used to classify them into ranges, and then discuss the results of the analysis of the audio features against delay, tempo and other variables. We close the chapter with a summary of the results and their implications for NMP.

9.1 Audio Characteristics

Audio feature analysis is used in a wide range of studies [27, 44, 60, 71, 85] with a diverse set of goals. The study by Rottondi et al. [71] focuses specifically on how audio features influence the perception of delay in NMP scenarios, so our work follows the

same methodology in order to lead to comparable results. The first step of this methodology is to calculate six audio features for each performance which are based on the spectral characteristics of the audio signal. Specifically, we perform a *Short Time Fourier Transform* (STFT) on the audio signal, that is, a Fourier Transform over each frame of audio samples. The STFT produces for each frame l a set of K frequency bins, where the magnitude of frequency bin $f(k)$ at frame l is equal to $S_l(k)$. Each audio feature is an expression over these frequencies and their magnitudes.

The *Spectral Centroid* (SC) corresponds to the "center of mass" of the frequency spectrum and is defined as follows:

$$SC_l = \frac{\sum_{k=1}^{K} f(k)S_l(k)}{\sum_{k=1}^{K} S_l(k)} \tag{9.1}$$

where l is the frame index; $S_l(k)$ is the spectrum magnitude computed at the k-th frequency bin at the l-th frame; $f(k)$ is the frequency corresponding to the k-th bin; and K is the total number of frequency bins. The SC is the first moment of the distribution and it shows where on the frequency axis the energy is concentrated, therefore it 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 around the SC:

$$SSp_l = \sqrt{\frac{\sum_{k=1}^{K} (f(k) - SC_l)^2 S_l(k)}{\sum_{k=1}^{K} S_l(k)}} \tag{9.2}$$

where SC_l is the spectral centroid at the l-th frame. The SSp characterizes the compactness of the distribution of the spectrum around the SC. A spread out distribution of the frequency components is characteristic of noisy sounds. For this reason, the SSp tends to measure the *noisiness* of a sound source.

The *Spectral Skewness* (SSk) is the third moment of the distribution and it captures the symmetry of its frequency distribution around the SC:

$$SSk_l = \frac{\sum_{k=1}^{K} (S_l(k) - SC_l)^3}{KSSp_l{}^3} \tag{9.3}$$

where SC_l is the SC and SSp_l is the SSp at the l-th frame. A positive value of SSk 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. Perfect symmetry corresponds to an SSk value of zero.

The *Spectral Kurtosis* (SK) is the fourth moment of the distribution and it describes the size of the tails of the distribution around the SC:

$$SK_l = \frac{\sum_{k=1}^{K} (S_l(k) - SC_l)^4}{KSSp_l{}^4} - 3 \tag{9.4}$$

where SC_l is the SC and SSp_l is the SSp at the l-th frame. Positive SK values indicate that distributions have relatively large tails, while distributions with small tails have negative SK, and normal distributions have zero SK. This is why the SK can be interpreted as a description of the deviation from normality. The offset -3 is a correction term that sets the SK of the normal distribution equal to zero.

The *Spectral Entropy* (SE) is a measure of the flatness of the spectrum, while the *Spectral Flatness* (SF) estimates the similarity of the source to a flat shape. Since white noise is characterized by a flat spectrum, SE and SF reflect the noisiness of an audio source. To be more exact, SE measures the flatness of the magnitude spectrum by applying Shannon's entropy definition:

$$SE_l = -\frac{\sum_{k=1}^{K} S_l(k) \log S_l(k)}{\log K} \tag{9.5}$$

A totally flat spectrum corresponds to maximum uncertainty where the entropy is maximal. On the other hand, a spectrum presenting only one very sharp peak with a flat and low background corresponds to minimum uncertainty, as the output will be entirely characterized by that peak. The SF on the other hand measures the similarity between the spectrum of the signal frame and a flat shape inside a predefined frequency band. Higher values of SF correspond to noisy sounds and vice versa. It is defined as the ratio between the geometric mean and the arithmetic mean of the spectrum:

$$SF_l = \frac{\sqrt[K]{\prod_{k=0}^{K-1} S_l(k)}}{\sum_{k=0}^{K-1} S_l(k)} \tag{9.6}$$

Starting from the audio recordings of the main study, we used the MIRToolbox [55] to extract these features for each instrument, based on the NMP session recordings for Scenario A, where delay was varied; we used the same tool in Chapter 8 to calculate the average tempo of each performance. Since each session was repeated 10 times, we calculated the average value for each metric across these repetitions.

As shown in Tables 9.1 and 9.2, even these averaged values can vary depending on the performance; for example, there were two sessions with electric piano, each

showing slightly different audio features. The same phenomenon occurred with the multiple electric guitar sessions. In practice, the way of playing each of the instruments influences its timbral metrics: arpeggios, solos and rhythm playing lead to different scores. In addition, the exact settings of the electric instruments and their amplifiers (e.g., adding distortion to an electric guitar) also influence their audio features.

Table 9.1: Musical genres and instruments played by each duet and their audio features (duets 1–6).

Duet No	1	2	3	4	5	6
Genre	Folk	Folk	Rock	Rock	Funk	Funk
Instrument A	Piano	Piano	El. Guitar	El. Bass	Organ	El. Bass
SC	5642	6082	7469	1669	1606	6041
SSp	6106	6452	7252	4442.3	2676	6952
SSk	1.26	1.15	0.6	3.24	3.76	0.92
SK	3.44	3.01	1.89	12.89	20.3	2.29
SF	0.47	0.44	0.61	0.13	0.10	0.48
SE	0.93	0.92	0.94	0.74	0.83	0.90
Instrument B	Santouri	Oud	El. Guitar	El. Guitar	El. Guitar	Percussion
SC	2384	1382	1010	1346	1456	1959
SSp	2505	2587	2868	3215	2357	2901
SSk	2.28	3.88	4.69	3.98	5.65	3.05
SK	11.07	21.03	25.9	19.6	40.82	14.34
SF	0.09	0.09	0.08	0.11	0.06	0.13
SE	0.85	0.80	0.75	0.79	0.82	0.86

Table 9.2: Musical genres and instruments played by each duet and their audio features (duets 7–11).

Duet No	7	8	9	10	11
Genre	Rock	Rock	Classic	Folk	Folk
Instrument A	El. Bass	El. Guitar	Flute	Ac. Guitar	Lute
SC	3459	5598	6971	3569	6249
SSp	6122	4120	6518	3884	4566
SSk	1.72	0.3	0.8	1.03	1.32
SK	4.58	1.83	2.32	2.77	4.31
SF	0.29	0.12	0.56	0.05	0.46
SE	0.81	0.93	0.94	0.89	0.94
Instrument B	Ac. Guitar	Violin	Violin	Bouzouki	Violin
SC	2807	2762	2185.9	2750	3386
SSp	4410	2782	2322	3394	4771
SSk	2.18	2.73	3.5	2.19	1.82
SK	7.35	13.3	21.05	8.68	5.56
SF	0.26	0.11	0.08	0.17	0.31
SE	0.86	0.85	0.81	0.87	0.75

In order to assess the correlation between the spectral characteristics and the responses to the QoME questionnaires we used the methodology proposed in [71], which we explain below. First, after calculating the average metrics for each of the 22 performances, we found the minimum and maximum value across all performances. The

Table 9.3: Classification ranges for the audio features.

Feature	SC	SSp	SSk	SK	SF	SE
Low	1010-3163	2322-3993	0.30-2.09	1.80-14.9	0.05-0.24	0.74-0.81
Middle	3164-5317	3994-5665	2.10-3.87	15.0-28.0	0.25-0.42	0.82-0.87
High	5318-7470	5666-7338	3.88-5.66	28.1-41.0	0.43-0.61	0.88-0.95

range for each metric was then divided in three equal parts, which we labeled as low, middle and high. For example, SE ranged from 0.74 to 0.95, so we divided this range to 0.74–0.81 (low), 0.82–0.87 (middle) and 0.88–0.95 (high). Table 9.3 shows these three ranges for each audio feature.

We then assigned each performance to one of these ranges and correlated each range and delay value with each of the QoME variables plus the performance tempo. We used the same approach of grouping performances and looking at their correlation with delay plus QoME or tempo, by looking at two musical features. First, we considered whether each musician performed the rhythm or the solo part of the duet, since previous work indicates that rhythmic parts are more sensitive to delay variations. Second, we considered the musical genre of each performance, which reflects the rhythmic structure of each musical piece.

9.2 Evaluation Results

In this section, we expand upon this analysis by looking at how each of the main subjective variables, that is, Perception of Satisfaction (PoSat), Perception of Audio Delay (PoAD), Perception of Synchronization Degree (PoSD) and I was Trying to Follow my partner (TTF) (see Chapter 7 for a detailed explanation of the subjective variables) was influenced not only by delay, but also by the audio features of the instruments and the musical features of the performances; we perform the same analysis for an objective variable, the observed tempo. We did not use the three subjective variables that were found to be nearly constant across the delay values (Perception of Anxiety, Perception of Irritation and Preference for Audio or Visual Contact).

To this end, we plotted one boxplot per characteristic, showing one set of boxes for each range of that characteristic (low, middle and high) and one box for each of the main MM2ME delays (0, 40, 80 and 120 ms); we used only these delays, as in Chapter 8, to avoid clutter and highlight only the most interesting delay settings. The boxes include a black horizontal line corresponding to the median value, with the entire box corresponding to 50% of the values (from the 25th to the 75th percentile). The whiskers (vertical lines above and below the box) show the minimum and maximum values excluding outliers: if we define IQR, the interquartile range, to be the difference between the 75th and the 25th percentile (that is, the height of the box), any value more than 1.5 times the IQR away from the box edges, is an outlier, shown as a dot in the plot.

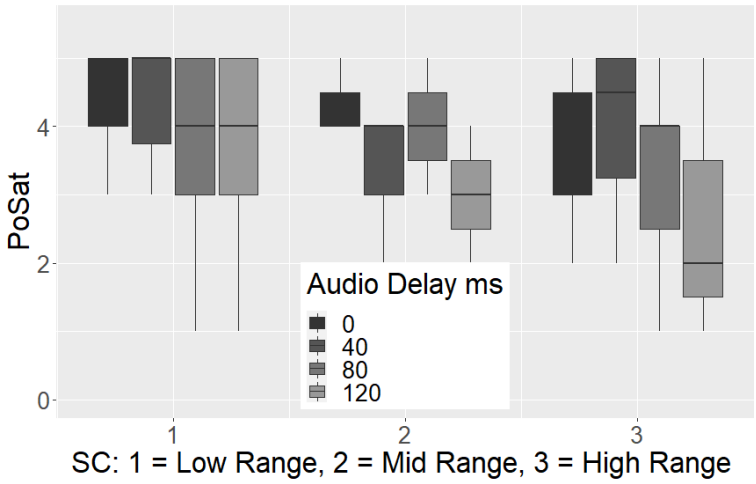

Figure 9.1: PoSat against delay and Spectral Centroid (SC).

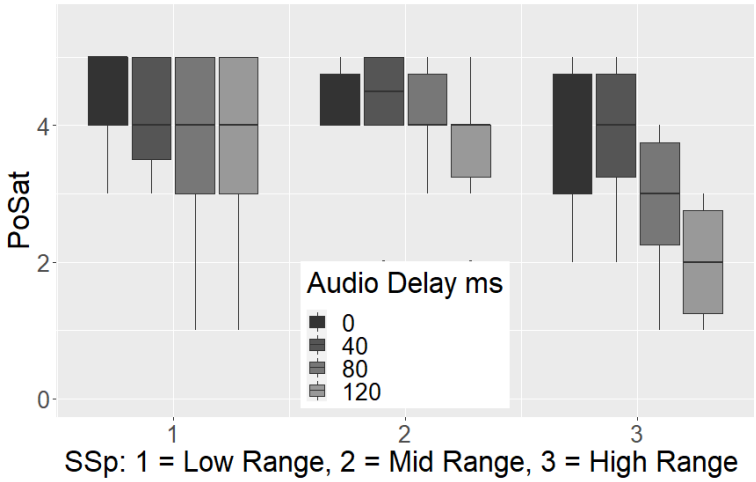

Figure 9.2: PoSat against delay and Spectral Spread (SSp).

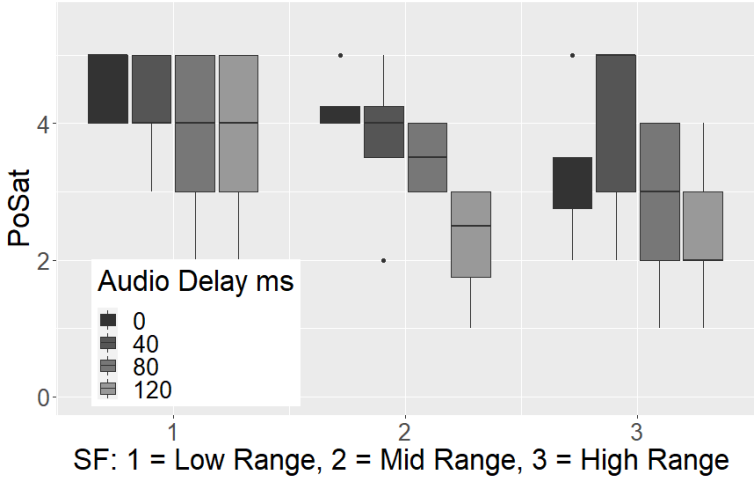

Figure 9.3: PoSat against delay and Spectral Flatness (SF).

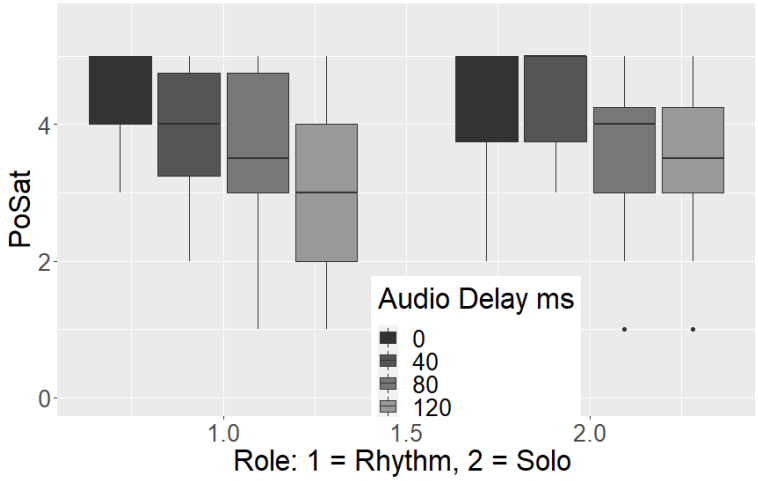

Figure 9.4: PoSat against delay and Rhythm or Solo.

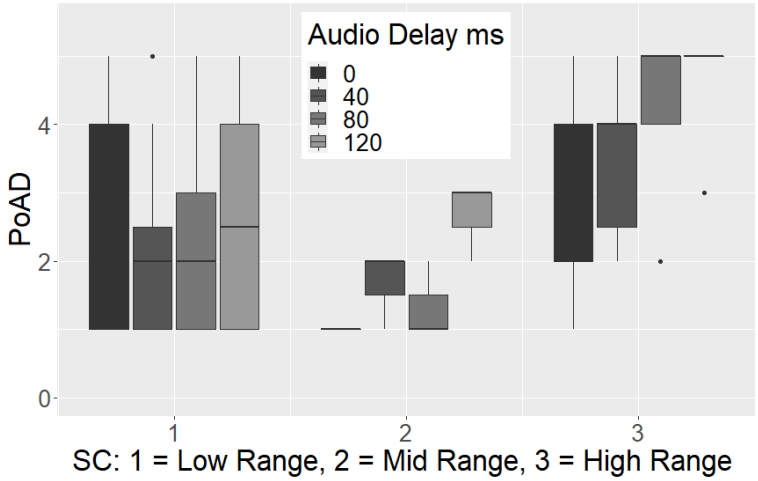

Figure 9.5: PoAD against delay and Spectral Centroid (SC).

Although the PoSat variable did not have a statistically significant correlation to delay, interesting observations arise when we look at the audio characteristics. PoSat was mostly affected by delay for instruments in the middle and high ranges of the Spectral Centroid (Figure 9.1), meaning instruments with brighter sounds; it was also mostly affected by delay for instruments in the high range of the Spectral Spread (Figure 9.2) and in the middle and high ranges of the Spectral Flatness (Figure 9.3), meaning noisier instruments. In addition, PoSat was more influenced by delay for rhythm rather than for solo performances (Figure 9.4).

Regarding PoAD, its dependence on the actual delay was higher for instruments in the middle and high ranges of the Spectral Centroid (Figure 9.5), while the Spectral Spread was a factor for instruments in the high range only (Figure 9.6) and the Spectral Skewness was a factor for instruments in the low and high ranges (Figure 9.7). Finally,

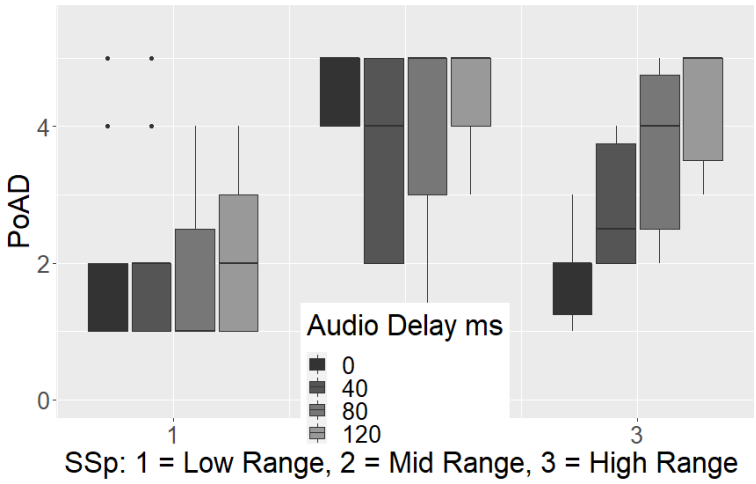

Figure 9.6: PoAD against delay and Spectral Spread (SSp).

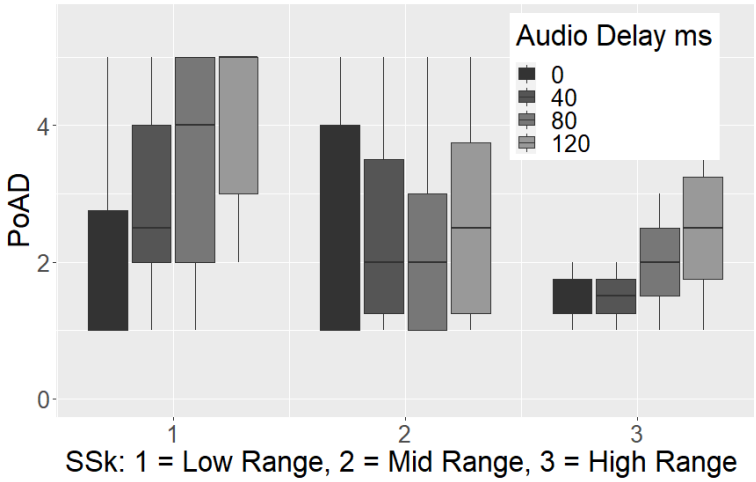

Figure 9.7: PoAD against delay and Spectral Skewness (SSk).

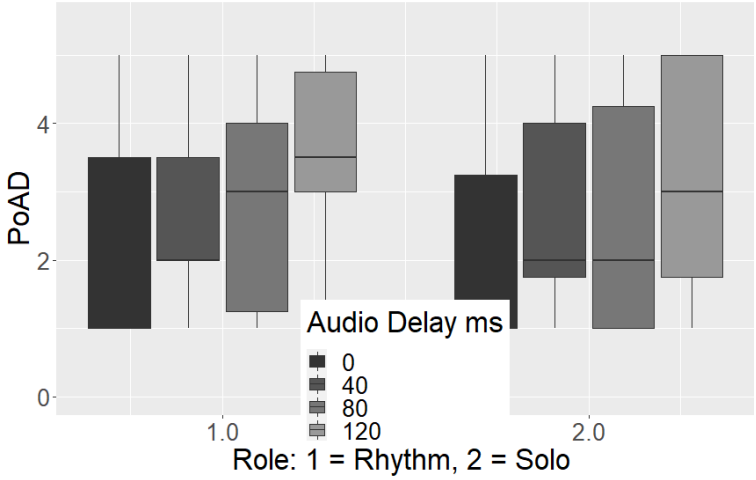

Figure 9.8: PoAD against delay and Rhythm or Solo.

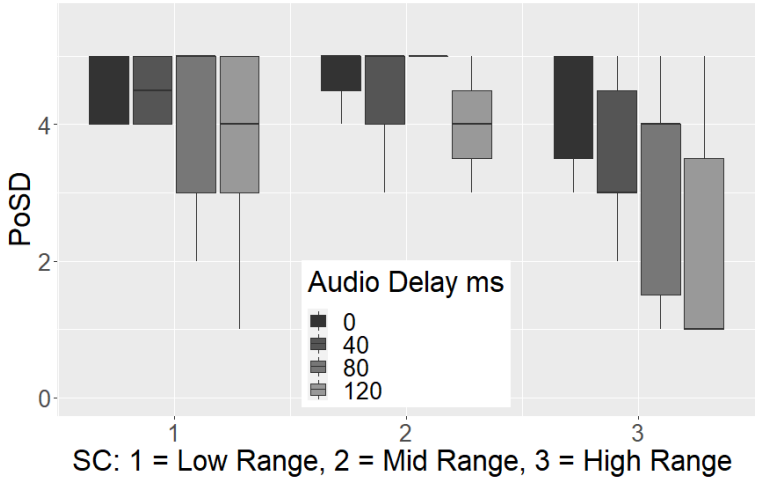

Figure 9.9: PoSD against delay and Spectral Centroid (SC)

PoAD (like PoSat) was more influenced by delay for rhythm rather than for solo performances (Figure 9.8).

PoSD was one of the most crucial metrics in our experiments, as it characterizes the perception of synchronization by the musicians themselves. It was mostly influenced by delay for instruments in the high range of the Spectral Centroid (Figure 9.9), with very similar results for the Spectral Spread and Spectral Flatness (not shown). It was also influenced by instruments in the middle and high ranges of Spectral Entropy (Figure 9.10). Again, it was more influenced by delay for rhythm rather than for solo parts (not shown).

The TTF metric was influenced by nearly all the audio and musical features. Regarding audio features, it was mostly affected by delay for instruments in the low and high ranges of the Spectral Centroid (Figure 9.11) and Spectral Spread (not shown), in the high range of Spectral Skewness (Figure 9.12) and in the middle and high ranges of Spectral Entropy (Figure 9.13) and Spectral Flatness (not shown). It was again more influenced by instruments playing rhythm parts (Figure 9.14), as was PoSat, PoAD and PoSD. Finally, it was more affected by delay with folk and rock performances, and less by funk and classical (Figure 9.15); it should be noted that the folk pieces performed were all intended for dancing, hence highly rhythmic.

Finally, the only objective variable assessed, tempo, was found to be mostly influenced by delay for instruments in the high range of Spectral Flatness (Figure 9.16), while amongst the four musical genres performed, only folk music (which was highly rythmic, as stated above) had an influence on the tempo (Figure 9.17), with the other genres being either insensitive (classical) or showing no clear correlation pattern (rock and funk).

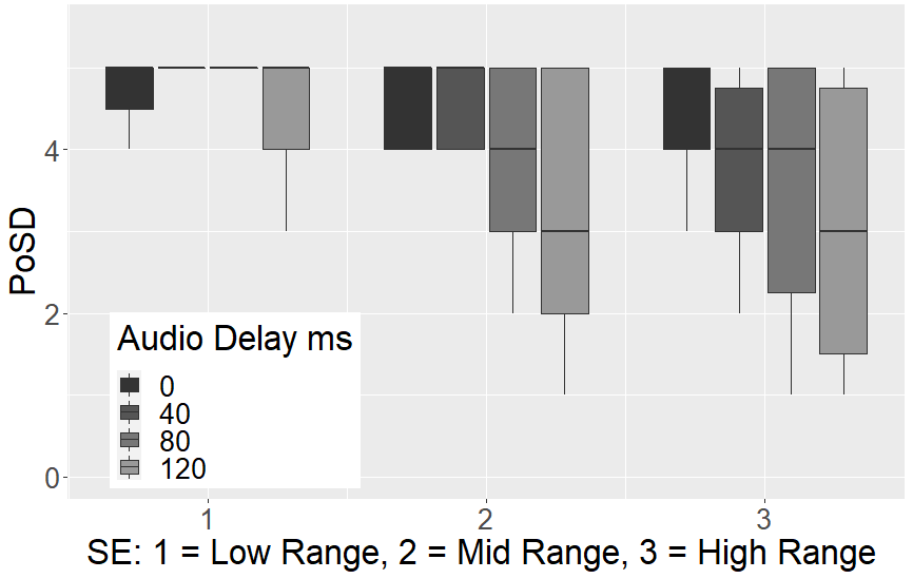

Figure 9.10: PoSD against delay and Spectral Entropy (SE)

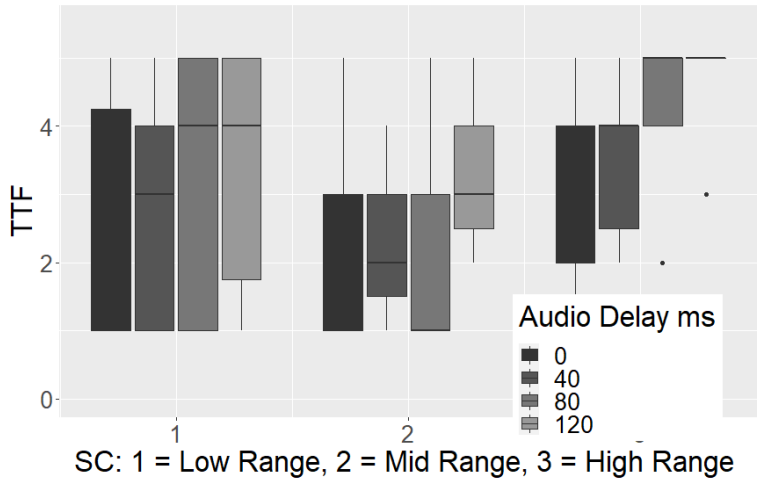

Figure 9.11: TTF against delay and Spectral Centroid (SC).

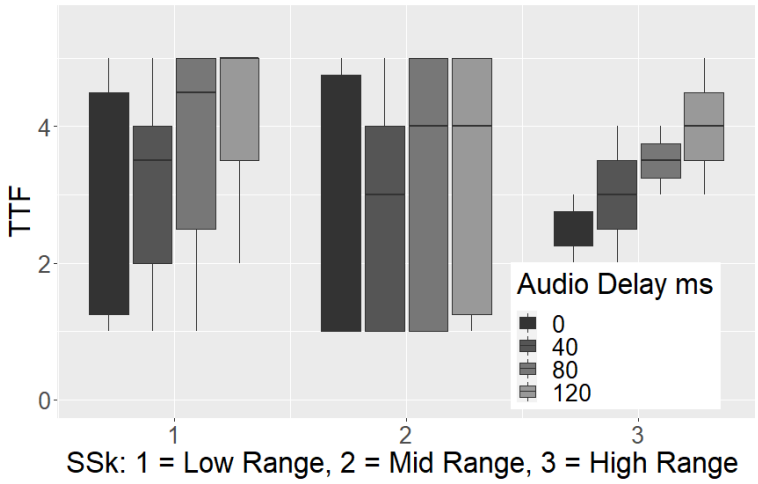

Figure 9.12: TTF against delay and Spectral Skewness (SSk)

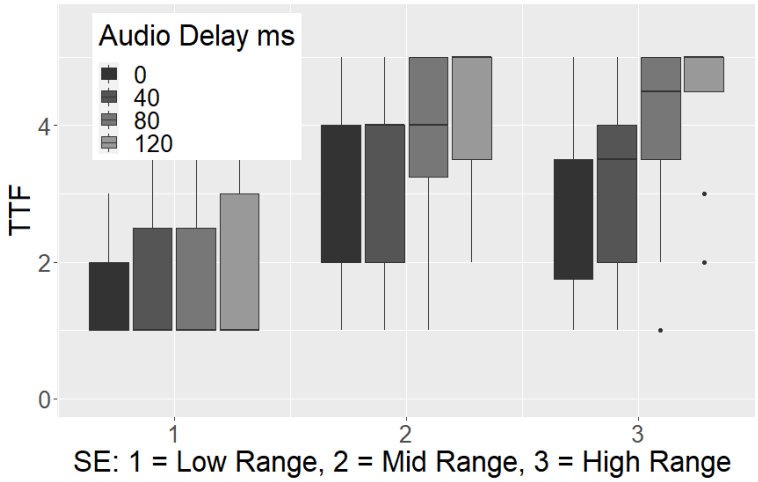

Figure 9.13: TTF against delay and Spectral Entropy (SE).

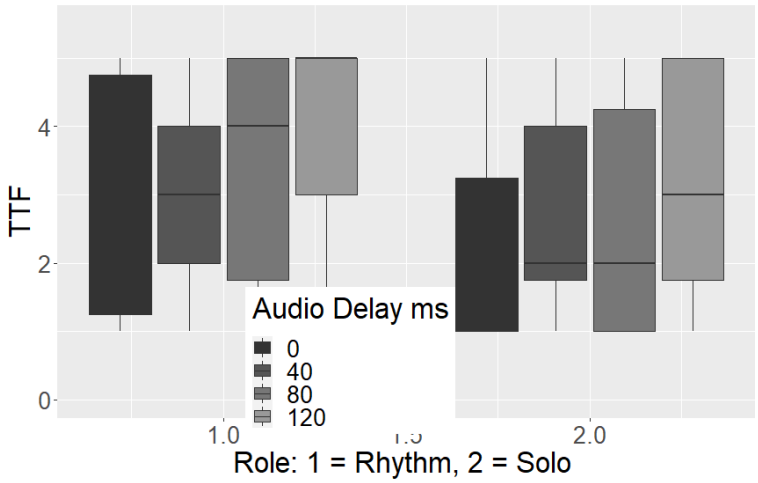

Figure 9.14: TTF against delay and Rhythm or Solo

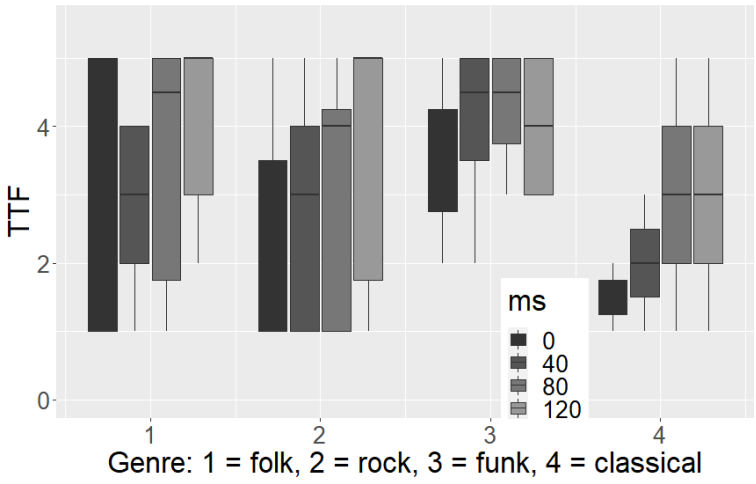

Figure 9.15: TTF against delay and Music Genre.

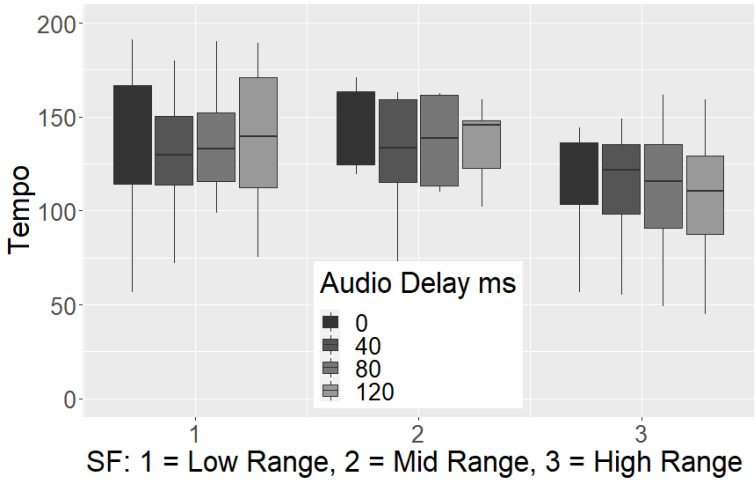

Figure 9.16: Tempo against delay and Spectral Flatness (SF).

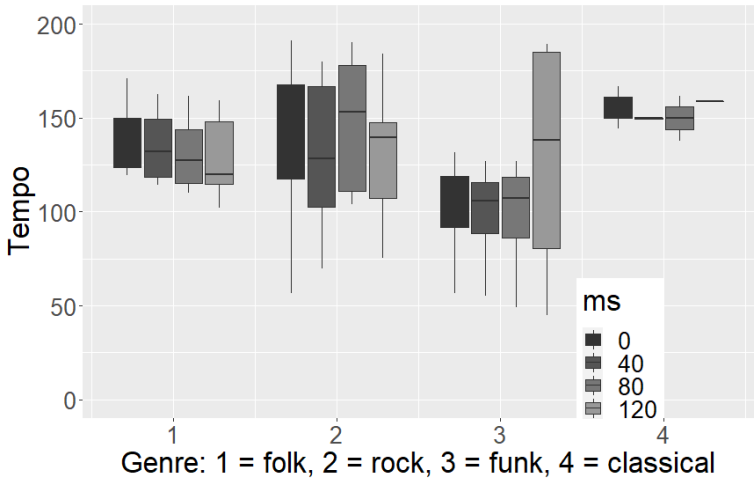

Figure 9.17: Tempo against delay and Music Genre.

9.3 Summary of Results

In this chapter, we analyzed the audio features of the instruments performed, the musical features of the performance and the tempo variations observed, in Scenario A (variable audio quality) of our main study. We assessed the correlation between delay and four variables characterizing the perception of an NMP session (perception of satisfaction, audio delay, synchronization and following the partner) plus the performance tempo. We also checked whether the effects of delay on those variables were dependent on two musical features of the performances, the role of the musician and the genre of the piece performed.

All the QoME variables (PoSat, PoAD, PSD and TTF) were more affected by delay with brighter and noisier instruments, performers that had a rhythm role and musical pieces with a more rhythmic structure. On the other hand, the effects of the audio and musical features on the performance tempo are not that clear. While our results are in line with previous studies of real musical performances (for example, [71] and [24]), our study shows that delay is more detrimental to rhythmic performances *in general*, not just on performances with faster initial tempos. Our results therefore lend additional credibility to previous results in the literature, due to the large sample size of our experiments, as well as a clearer picture of QoME due to the consideration of multiple variables and the use of both audio and musical features of the instruments used and the pieces performed.

Chapter 10

Emotion Analysis

Having analyzed the questionnaires from the main study, as shown in Chapter 7, the tempo evolution as reflected in the recordings of the sessions, as shown in Chapter 8 and the audio features of the recordings, as shown in Chapter 9, we now turn to an analysis of the videos captured during the main study. Specifically, we employ 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. 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 to NMP experiments.

In this chapter we describe the methods we used to recognize the emotions of the participating musicians based on the recorded videos, which are based on facial feature extraction with machine learning techniques, and then discuss the results of the emotion analysis for both scenarios (variable delay and variable quality) and how they compare with the previous analyses reported in this thesis, before summarizing our conclusions.

10.1 Emotion Recognition with Machine Learning

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 [33]; it is a multimodal assessment, 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 an automated (and more objective) assessment method.

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 based on FER [58]. 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 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 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 [82] 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 [79]. 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 [91]. 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.

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

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

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.

A second issue is that the emotions detected by the DeepFace system are generic, rather than those expected in an NMP scenario; for example, in NMP it is rather unlikely for a participant to experience disgust, but it is quite likely for the musician to experience frustration.

10.2 Evaluation Results

We analyzed the results of both Scenario A (variable audio delay) and Scenario B (variable audio quality) with DeepFace, using the videos recorded during the main study described in Chapter 7. Since this is just a preliminary analysis (and the first of its kind), we decided to work with the videos from both scenarios, in order to uncover any interesting details.

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. Furthermore, each musical piece induces in itself emotions to the participants, but since each participant played the same piece repeatedly, it was hard to distinguish the emotions due to the music itself and those due to the specific performance.

A second observation was that the emotional reactions when audio conditions changed were different for each musician. However, interesting points come up by

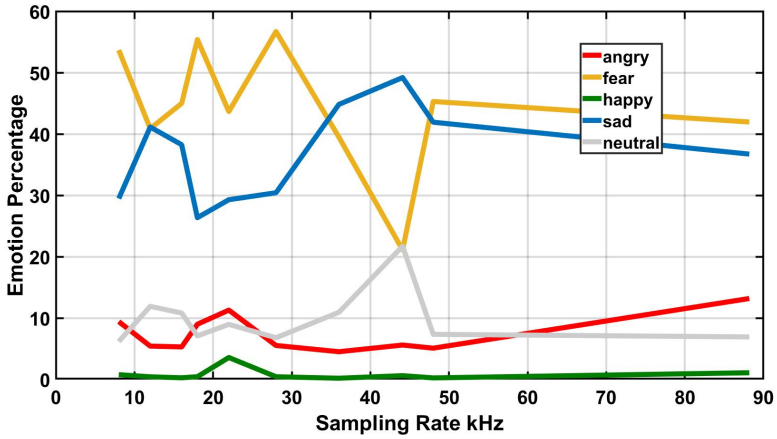

Figure 10.1: Musician A's emotions vs. Sampling Rate.

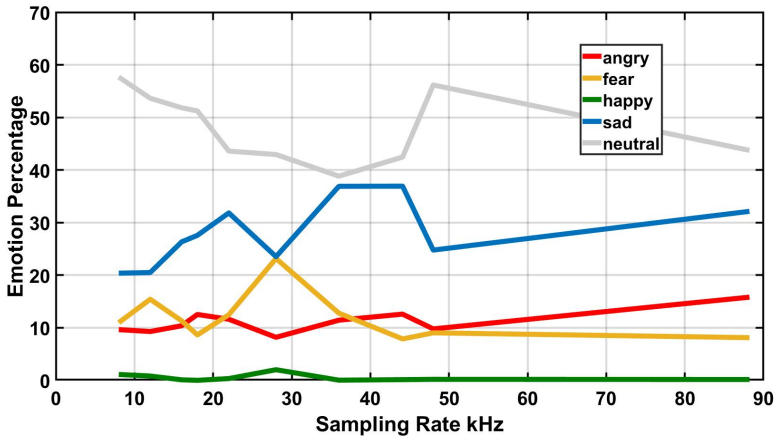

Figure 10.2: Musician B's emotions vs. Sampling Rate.

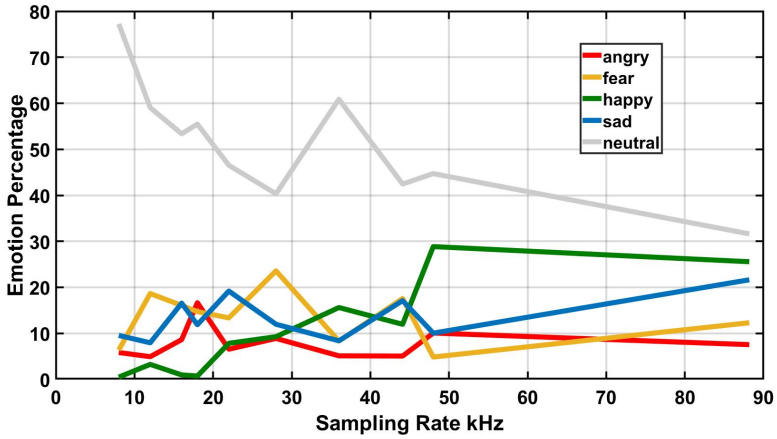

Figure 10.3: Musician C's emotions vs. Sampling Rate.

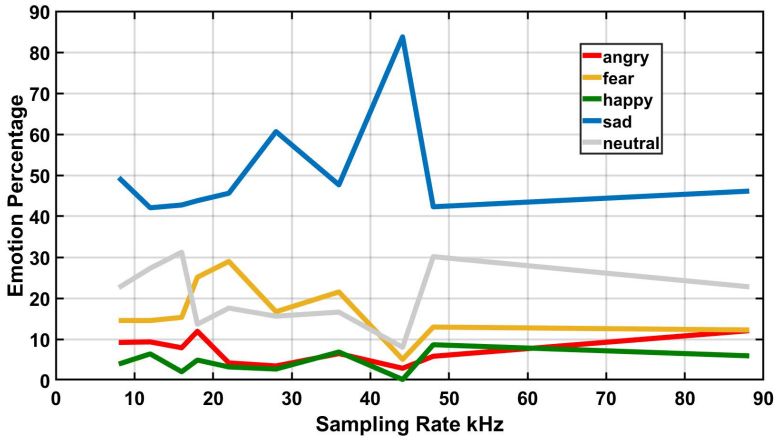

Figure 10.4: Musician D's emotions vs. Sampling Rate.

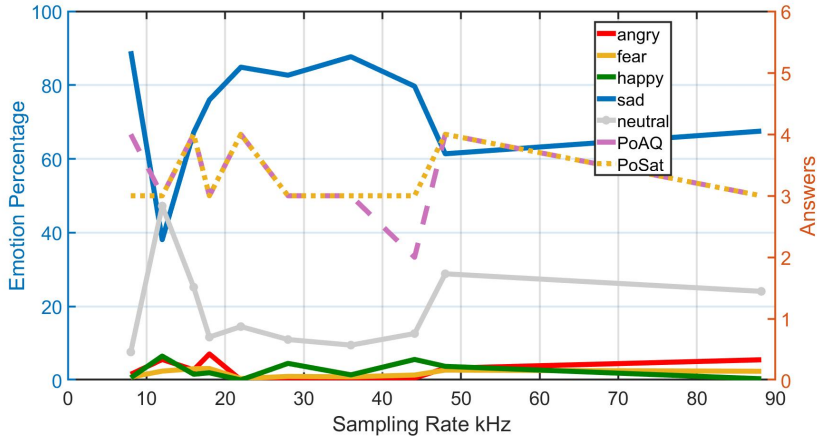

Figure 10.5: Musician E's Emotions and PoSat/PoAQ vs. Sampling Rate: PoSat follows neutrality.

looking at the results. For example, Figures 10.1, 10.2, 10.3 and 10.4 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 at or around the sampling rate of 44.1 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. We surmise that this may be due to the fact that 44.1 kHz was the very first sampling rate used in the variable quality tests (see Chapter 7), so it was the first performance of each participant in Scenario B, thus being more surprising to the musicians.

Figures 10.5, 10.6, 10.7 and 10.8 show the average percentage of each emotion for an entire performance (left y-axis, solid lines) and the scores of the PoSat and PoAQ

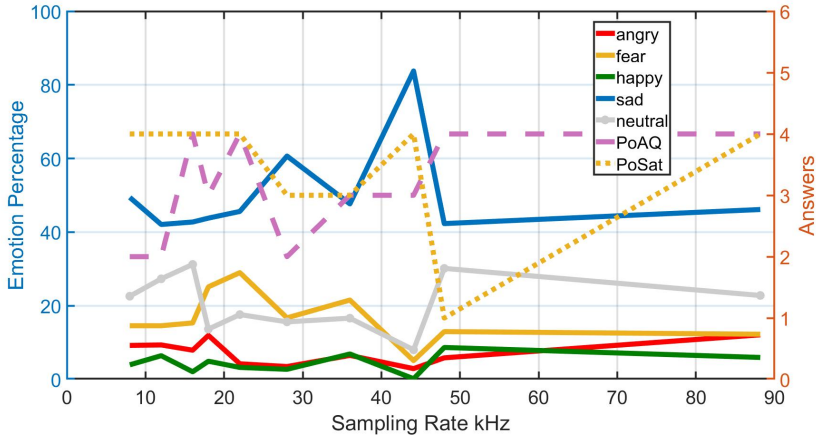

Figure 10.6: Musician F's Emotions and PoSat/PoAQ vs. Sampling Rate: PoSat follows sadness.

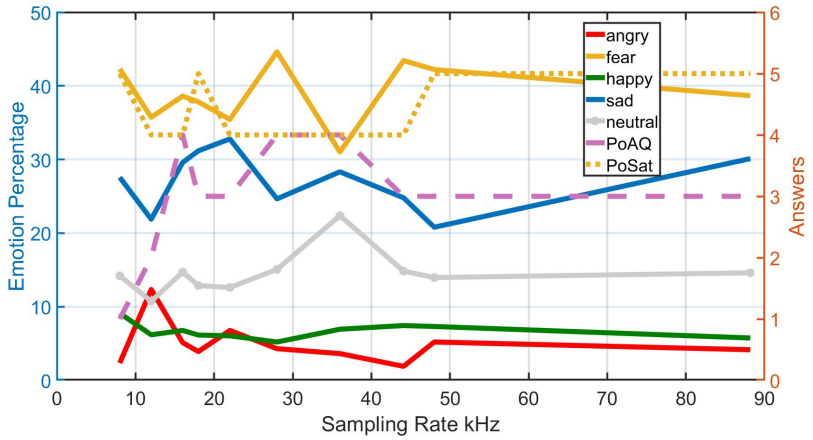

Figure 10.7: Musician's G's Emotions and PoSat/PoAQ vs. Sampling Rate: PoSat follows fear.

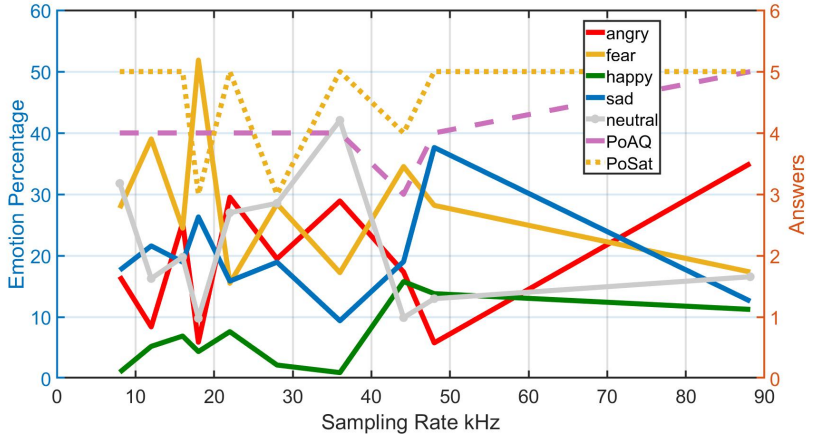

Figure 10.8: Musician's H's Emotions and PoSat/PoAQ vs. Sampling Rate: PoSat follows anger.

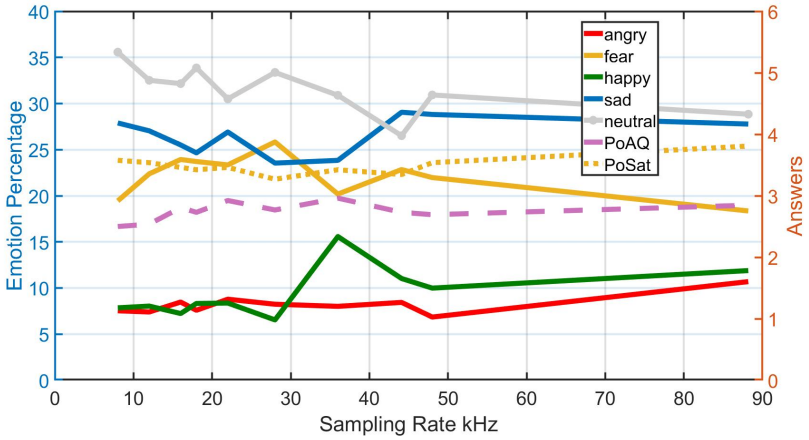

Figure 10.9: Average values of emotions and PoSat/PoAQ across all musicians vs. Sampling Rate.

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 10.5 PoSat follows neutrality, in Figure 10.6 PoSat follows sadness, in Figure 10.7 PoSat follows fear and in Figure 10.8 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 10.9 and 10.10 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 10.9 the disruption at 44.1 kHz which was 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 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 10.10 we can see a disruption as delay grows from 30 to 40 ms, where it starts becoming noticeable; unlike with sampling rate, where the disruption may be due to the order of the experiments, in the variable delay experiments the delays of 30 and 40 ms were used in the fourth and seventh repetition (see Chapter 7), respectively, so they are unlikely to be an artefact of the experimental sequence.

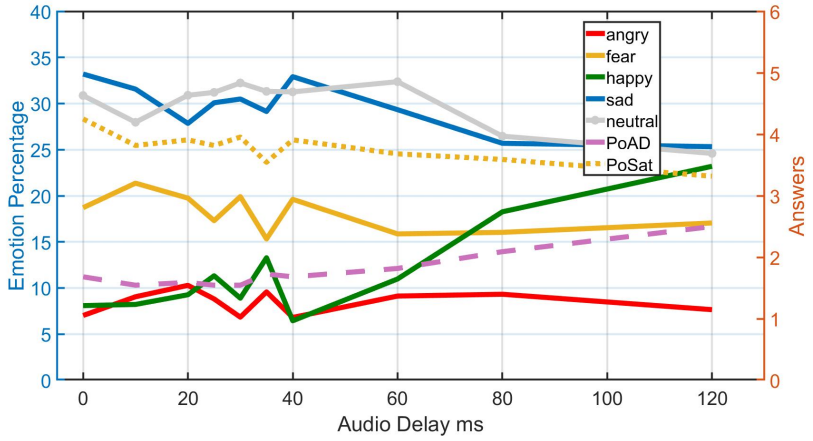

Figure 10.10: Average values of emotions and PoSat/PoAQ across all musicians vs. Audio Delay.

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 happiness! 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 10.9 and Figure 10.10, 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.

10.3 Summary of Results

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 performance in the NMP session, excluding his or her general emotional state and the emotions induced by the musical piece performed. In addition, in some cases further information is needed to interpret the trends exhibited in the results (e.g., the presence of laughter when the performance fails).

However, this additional mode of assessing the experience of the musicians can be used to strengthen the conclusions drawn from subjective studies. In our experiments, discontinuities in the emotional response in the variable audio experiments can be attributed to the ordering of the repetitions, while the discontinuities in the variable delay experiments are more likely due to an increasing perception of delay and the problems it causes to synchronization. We also found that the emotional response of the musicians is more correlated with the variance in audio delay, rather than with the variance in audio quality, which is in line with the results of the subjective analysis.

While it seems unlikely that emotion analysis will be able to provide conclusions for NMP experiments by itself, it is a useful additional mode of analysis. To make emotion analysis more useful, 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 this camera angle would better match the training sets used in the machine learning process.

Chapter 11

Conclusions and Future Work

In this chapter we first summarize the conclusions from the studies reported in the preceding chapters, focusing on the main NMP study and the multiple modes of evaluation that we employed, from the traditional subjective analysis, to the rarer tempo and audio features analyses, up to the first ever video-based emotion analysis of NMP sessions. Our goal is to bring together all our conclusions and see how they reinforce each other.

We then discuss directions for future work, based not only on the conclusions, but also on the limitations of the present study. We cover all the modes of analysis we employed, presenting ideas for extensions and additional work.

11.1 Conclusions

The results from the subjective analysis of the questionnaires indicate that even though increasing delay does have a statistically significant effect on the Quality of Musicians' Experience, the range of acceptable delays is larger than what previous studies found, based on hand clapping experiments. Indeed, most participants in our study considered the performances to be synchronized and satisfactory with one way (M2E) delays of up to 40 ms, which is higher than the previously considered limit of 25 to 30 ms. With real musicians and musical performances, the evaluation metrics did not significantly diverge when the delay grew from 30 to 40 ms. This is, however, the average case; in specific situations, for example in the duets where pianists were involved, delay had a more negative effect on the results. Note also that as delay was perfectly fixed in our setup, it corresponds to the increased delay after the receiver's de-jittering buffer in a real NMP system.

On the other hand, reducing quality by lowering the sampling rate did not have a statistically significant effect on the Quality of Musicians' Experience, even when we reduced the sampling rate (and the resulting bitrate) by an order of magnitude; the subjective scores are practically the same as long as the sampling rate is 16 kHz or higher. This implies than when bandwidth restrictions exist, we can reduce the sampling rate

and the resulting audio quality, without a penalty in the perception of the performance. If we instead used compression to save bandwidth, it would significantly inflate delays, even with the lowest latency codecs like Opus, as we found in our validation study. This does not imply that a sampling rate of 16 kHz is sufficient to enjoy music in general, but that it is sufficient for a pair of musicians to enjoy an NMP session, at least for the (quite wide) range of instruments tested. This reduction allows reducing the required bit rate to 35% of what is required for Audio CD level quality, which uses a sampling rate of 44.1 kHz.

The analysis of the tempo of the recordings confirmed previous studies showing that musicians tend to slow down their tempo as delays grow, especially with percussive instruments and rhythm parts; it also showed that with the (unnaturally) low delay of 0 ms, musicians tend to speed up their tempo. At the same time, the tempo analysis confirmed the results of the subjective analysis on the limits to delay tolerance, as musicians in nearly all cases could synchronize with one way delays of up to 40 ms, in the sense that they could reach and maintain a stable tempo. Indeed, musicians, especially more experienced ones, tried to adapt to higher delays by slowing down their tempo.

Moving to the audio feature analysis where we correlated audio and musical features of the recordings with delay and the subjective variables and tempo, we found that the Quality of Musicians' Experience was more affected by delay with brighter and noisier instruments, performers that had a rhythm role and musical pieces with a more rhythmic structure; these are consistent with the findings of the subjective and the tempo analysis. On the other hand, the effects of the audio and musical features on the performance tempo are not that clear. While our results are in line with previous studies of real musical performances, our study shows that delay is more detrimental to rhythmic performances in general, not just on performances with faster initial tempos. In addition, the larger number of participants and the wider range of instruments, genres and tempos, lend additional weight to such conclusions.

Finally, the emotion analysis performed on the video recordings of the NMP sessions, the first of its kind, 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, excluding their pre-existing emotional state and the emotions induced by the music itself. Indeed, 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, even though it cannot replace them. For example, we found that the emotional response of the musicians was stronger to audio delay changes than to audio quality changes, which is in line with the findings of the subjective analysis, where we also found that audio delay is more influential than audio quality to the Quality of Musicians' Experience.

11.2 Ongoing and Future Work

Following up on our subjective analysis, we are looking into more detail at the relationship between the instrument, style, tempo and the quality of experience. The analysis in this thesis indicates that all these factors do play a role in the experience of the musicians during NMP. It is clear however, that when taking subsets of our data set, the data points are too few to derive statistically significant results. Ideally, other researchers will duplicate our setup to add more observations to our data set, so that larger subsets with common characteristics can be formed for analysis.

Regarding the main study and the subjective analysis, we have a few pointers for other researchers: they could focus on audio delay, since audio quality does not seem to influence the end results that much; they could encourage the musicians to perform in fewer styles or encourage the reuse of the same musical pieces by multiple duets, so as to gather more data points with similar underlying characteristics; and they could test one way delays between 35 ms and 45 ms to more accurately pinpoint the exact limits to delay tolerance in NMP.

Similarly, while we are digging deeper on the audio data collected, focusing on issues such as the dependence of tempo variations on other factors, such as the style of music performed and the target tempo of each piece, the number of data points that we have for each subset are limited, and would require additional experiments with the same setup to lead to more conclusive results.

In the tempo analysis area, we would suggest to other researchers to record the intended tempo (in BPM) of the performances, either by asking the performers, or by having them perform an initial test run in the same room with a metronome, so as to pinpoint their desired BPM. This would allow the researchers to have a firm basis for comparisons with the tempo revealed by the tempo analysis. Of course, the metronome should be avoided during the actual tests, to allow the musicians to perform with each other, rather than with a common reference.

In the audio features analysis, other than gathering more samples, we would encourage other researchers to seek out more musicians with percussive or very rhythmic instruments (such as bass), so as to be able to better determine the influence of instruments with these audio features to the Quality of Musicians' Experience in NMP.

The area where we expect to come up with more interesting results is the emotion analysis based on the videos, where we are looking at data analysis from additional, more powerful, tools. Unfortunately, since the videos were not recorded with this goal in mind, they are not ideal for face recognition purposes, as the camera does not focus on the faces, which are often obscured by the equipment needed for the performance.

In the emotion analysis, in addition to using cameras pointing directly at the face of performers, separate from the cameras used for synchronization, so as to have clearer data, it would be worthwhile if additional instruments could be used, such as EEG

headsets, to detect emotions through additional channels. Of course, care should be taken to avoid the setup becoming too cumbersome and irritating to the musicians.

Finally, since the emotions felt by the musicians are a mix of their pre-existing emotional state, the emotions induced by the musical piece and the emotions induced by the performance, it would be interesting to try to account for the non-NMP emotions by calibrating the system before the performance. One idea would be to measure the emotions of the musicians before the performance, to assess their pre-existing condition, and also measure them during non-NMP performance, to assess the effects of a regular performance.

Bibliography

[1] D. Akoumianakis, C. Alexandraki, V. Alexiou, C. Anagnostopoulou, A. Eleftheriadis, V. Lalioti, Y. Mastorakis, A. Modas, A. Mouchtaris, D. Pavlidi, G. C. Polyzos, P. Tsakalides, G. Xylomenos, and P. Zervas. "The MusiNet project: Addressing the challenges in Networked Music Performance systems". In: *6th International Conference on Information, Intelligence, Systems and Applications (IISA)*. July 2015. DOI: 10.1109/IISA.2015.7388002.

[2] D. Akoumianakis, C. Alexandraki, V. Alexiou, C. Anagnostopoulou, A. Eleftheriadis, V. Lalioti, A. Mouchtaris, D. Pavlidi, G. C. Polyzos, P. Tsakalides, G. Xylomenos, and P. Zervas. "The MusiNet project: Towards unraveling the full potential of Networked Music Performance systems". In: *5th International Conference on Information, Intelligence, Systems and Applications (IISA)*. July 2014, pp. 1–6. DOI: 10.1109/IISA.2014.6878779.

[3] C. Alexandraki and D. Akoumianakis. "Exploring new perspectives in network music performance: The DIAMOUSES framework". In: *Computer Music Journal* 34 (June 2010), pp. 66–83. DOI: 10.1162/comj.2010.34.2.66.

[4] C. Alexandraki, P. Koutlemanis, P. Gasteratos, N. Valsamakis, and G. Milolidakis. "Towards the implementation of a generic platform for networked music performance: The Diamouses approach". In: (Aug. 2008).

[5] G. Baltas and G. Xylomenos. "Evaluating the impact of network I/O on ultra-low delay packet switching". In: *2015 IEEE Symposium on Computers and Communication (ISCC)*. July 2015, pp. 397–402. DOI: 10.1109/ISCC.2015.7405547.

[6] Á. Barbosa, J. Cardoso, and G. Geiger. "Network Latency Adaptive Tempo in the Public Sound Objects System". In: *Proceedings the International Conference on New Interfaces for Musical Expression (NIME)*. Jan. 2005, pp. 184–187.

[7] Á. Barbosa and J. Cordeiro. "The Influence of Perceptual Attack Times in Networked Music Performance". In: *Audio Engineering Society Conference: 44th International Conference: Audio Networking*. Nov. 2011. URL: http://www.aes.org/e-lib/browse.cfm?elib=16133.

[8] C. Bartlette, D. Headlam, M. Bocko, and G. Velikic. "Effect of Network Latency on Interactive Musical Performance". In: *Music Perception: An Interdisciplinary Journal* 24.1 (2006), pp. 49–62. ISSN: 07307829, 15338312. DOI: 10.1525/mp.2006.24.1.49.

[9] J. A. Bergstra and C. A. Middelburg. *ITU-T Recommendation G.107 : The E-Model, a computational model for use in transmission planning*. Tech. rep. ITU Telecommunication Standardization Sector, 2003.

[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. DOI: 10.1109/ROMAN.2007.4415182.

[11] J.-P. Cáceres and C. Chafe. "JackTrip: Under the Hood of an Engine for Network Audio". In: *Proceedings of International Computer Music Conference*. International Computer Music Association. San Francisco, California: International Computer Music Association, 2009, 509–512.

[12] J.P. Cáceres and C. Chafe. "JackTrip/SoundWIRE meets server farm". In: *Computer Music Journal* 34 (Sept. 2010), pp. 29–34. DOI: 10.1162/COMJ_a_00001.

[13] A. Carôt, C. Hoene, H. Busse, and C. Kuhr. "Results of The Fast-Music Project: Five Contributions to The Domain of Distributed Music". In: *IEEE Access* (Mar. 2020). DOI: 10.1109/ACCESS.2020.2979362.

[14] A. Carôt, U. Kramer, and G. Schuller. "Network Music Performance (NMP) in Narrow Band Networks". In: *Audio Engineering Society Convention 120*. May 2006.

[15] A. Carôt and C. Werner. "Fundamentals and Principles of Musical Telepresence". In: *Journal of Science and Technology of the Arts* 1 (May 2009). DOI: 10.7559/citarj.v1i1.6.

[16] A. Carôt, C. Werner, and T. Fischinger. "Towards a Comprehensive Cognitive Analysis of Delay-Influenced Rhythmical Interaction". In: *International Computer Music Conference*. 2009.

[17] C. Chafe, J.-P. Cáceres, and M. Gurevich. "Effect of temporal separation on synchronization in rhythmic performance". In: *Perception* 39 (Jan. 2010), pp. 982–92. DOI: 10.1068/p6465.

[18] C. Chafe and M. Gurevich. "Network Time Delay and Ensemble Accuracy: Effects of Latency, Asymmetry". In: *Audio Engineering Society Convention 117*. Oct. 2004. URL: http://www.aes.org/e-lib/browse.cfm?elib=12865.

[19] C. Chafe, M. Gurevich, G. Leslie, and S. Tyan. "Effect of Time Delay on Ensemble Accuracy". In: *Proc Intl. Soc. Musical Acoustics; Nara* (Aug. 2009).

[20] C. Chafe, S. Wilson, A. Leistikow, D. Chisholm, and G. Scavone. "A simplified approach to high quality music and sound over IP". In: *COST-G6 Conference on Digital Audio Effects (DAFx-00)*. 2000.

[21] E. Chew, A. Sawchuk, C. Tanoue, and R. Zimmermann. "Segmental tempo analysis of performances in user-centered experiments in the distributed immersive performance project". In: *Proceedings of the Sound and Music Computing Conference*. Jan. 2005.

[22] E. Chew, R. Zimmermann, A. Sawchuk, C. Kyriakakis, C. Papadopoulos, A.R.J. Francois, G.-J. Kim, A. Rizzo, and A. Volk. "Musical Interaction at a Distance : Distributed Immersive Performance". In: *Proceedings of the MusicNetwork Fourth Open Workshop on Integration of Music in Multimedia Applications*. 2004.

[23] E. Chew, R. Zimmermann, A. Sawchuk, C. Papadopoulos, C. Kyriakakis, C. Tanoue, D. Desai, M. Pawar, R. Sinha, and W. Meyer. "A Second Report on the User Experiments in the Distributed Immersive Performance Project". In: *Proceedings of*

the 5th Open Workshop of MUSICNETWORK: Integration of Music in Multimedia Applications. Jan. 2005.

[24] S. Delle Monache, M. Buccoli, L. Comanducci, A. Sarti, G. Cospito, E. Pietrocola, and F. Berbenni. "Time is not on my side: Network latency, presence and performance in remote music interaction". In: *Colloquium on Music Informatics-Machine Sounds, Sound Machines*. 2018.

[25] P.F. Driessen, T.E. Darcie, and B. Pillay. "The Effects of Network Delay on Tempo in Musical Performance". In: *Computer Music Journal* 35.1 (Mar. 2011), pp. 76–89. DOI: 10.1162/COMJ_a_00041.

[26] C. Drioli and N. Buso. "Networked Performances and Natural Interaction via LOLA: Low Latency High Quality A/V Streaming System". In: *Lecture Notes in Computer Science* 7990 (Jan. 2013), pp. 240–250. DOI: 10.1007/978-3-642-40050-6_21.

[27] K. Drossos, R. Kotsakis, G. Kalliris, and A. Floros. "Sound events and emotions: Investigating the relation of rhythmic characteristics and arousal". In: *International Conference on Information, Intelligence, Systems and Applications (IISA)*. 2013, pp. 1–6. DOI: 10.1109/IISA.2013.6623709.

[28] A. Dzedzickis, A. Kaklauskas, and V. Bucinskas. "Human Emotion Recognition: Review of Sensors and Methods". In: *Sensors* 20.3 (2020). ISSN: 1424-8220. DOI: 10.3390/s20030592.

[29] P. Ekman. "Facial Expressions of Emotion: New Findings, New Questions". In: *Psychological Science* 3.1 (1992), pp. 34–38.

[30] S. Farner, A. Solvang, A. Sæbo, and U.P. Svensson. "Ensemble Hand-Clapping Experiments under the Influence of Delay and Various Acoustic Environments". In: *Journal of the AES* 57.12 (2009), pp. 1028–1041.

[31] V. Fischer. *Case Study: Performing Band Rehearsals on the Internet With Jamulus*. URL: https://jamulus.io/PerformingBandRehearsalsontheInternetWithJamulus.pdf.

[32] L. Gabrielli and S. Squartini. *Wireless Networked Music Performance*. Dec. 2016. ISBN: 978-981-10-0334-9. DOI: 10.1007/978-981-10-0335-6.

[33] A. Gabrielsson and P. Juslin. "Emotional expression in music". In: *Handbook of Affective Sciences*. Ed. by R. J. Davidson, K. R. Scherer, and H. H. Goldsmith. Oxford University Press, 2003, pp. 503–534.

[34] A. Gabrielsson and P. Juslin. "Emotional Expression in Music Performance: Between the Performer's Intention and the Listener's Experience". In: *Psychology of Music* 24.1 (1996), pp. 68–91.

[35] J.M. Garcia-Garcia, V.M.R. Penichet, and M.D. Lozano. "Emotion Detection: A Technology Review". In: *International Conference on Human Computer Interaction*. Cancun, Mexico, 2017.

[36] A. Geeves, D. Mcllwain, J. Sutton, and W. Christensen. "Expanding Expertise: Investigating a Musician's Experience of Music Performance". In: *Proceedings of the 9th Conference of the Australasian Society for Cognitive Science*. 2010, pp. 106–113. DOI: 10.5096/ASCS200917.

[37] M. Geronazzo, S. Delle Monache, L. Comanducci, M. Buccoli, Massimiliano Zanoni, A. Sarti, E. Pietrocola, F. Berbenni, and G. Cospito. "A Presence- and Performance-Driven Framework to Investigate Interactive Networked Music Learning Scenarios". In: *Wireless Communications and Mobile Computing* 2019 (2019), p. 4593853. ISSN: 1530-8669.

[38] X. Gu, M. Dick, Z. Kurtisi, U. Noyer, and L. Wolf. "Network-centric music performance: practice and experiments". In: *IEEE Communications Magazine* 43.6 (2005), pp. 86–93. DOI: 10.1109/MCOM.2005.1452835.

[39] 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)*. Mar. 2011, pp. 827–834.

[40] M. Gurevich, C. Chafe, G. Leslie, and S. Tyan. "Simulation of Networked Ensemble Performance with Varying Time Delays: Characterization of Ensemble Accuracy". In: *International Computer Music Conference*. 2004.

[41] ITU-R. *BS.1534-3: Method for the Subjective Assessment of Intermediate Quality Level of Audio Systems*. Tech. rep. ITU Radiocommunication Standardization Sector, 2015.

[42] ITU-T. *P.862: Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-end Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs*. Tech. rep. ITU Telecommunication Standardization Sector, 2001.

[43] jamulus. *Play music online. With friends. For free.* https://llcon.sourceforge.io/. May 2013.

[44] K. Jensen and Tue H. Andersen. "Real-time beat estimation using feature extraction". In: *International Symposium on Computer Music Modeling and Retrieval*. Springer. 2003, pp. 13–22.

[45] P. Juslin and P. Laukka. "Expression, Perception, and Induction of Musical Emotions: A Review and a Questionnaire Study of Everyday Listening". In: *Journal of New Music Research* 33 (Sept. 2004), pp. 217–238. DOI: 10.1080/0929821042000317813.

[46] A. Kapur, G. Wang, P. Davidson, and P. Cook. "Interactive Network Performance: a dream worth dreaming?" In: *Organised Sound* 10 (2005), pp. 209–219.

[47] F. Kauer, M. Fink, and U. Zölzer. "The JamBerry - A Stand-Alone Device for Networked Music Performance Based on the Raspberry Pi". In: *Linux Audio Conference*. May 2014.

[48] D. Kenny, P. Davis, and J. Oates. "Music performance anxiety and occupational stress amongst opera chorus artists and their relationship with state and trait anxiety and perfectionism". In: *Journal of anxiety disorders* 18 (Feb. 2004), pp. 757–77. DOI: 10.1016/j.janxdis.2003.09.004.

[49] K. Kilkki. "Quality of Experience in Communications Ecosystem." In: *Journal of Universal Computer Science* 14 (Jan. 2008), pp. 615–624.

[50] Y. Kobayashi, Y. Nagata, and Y. Miyake. "Analysis of network ensemble with time lag". In: *Proceedings 2003 IEEE International Symposium on Computational Intelligence in Robotics and Automation. Computational Intelligence in Robotics and Automation for the New Millennium (Cat. No.03EX694)*. Vol. 1. 2003, 336–341 vol.1. DOI: 10.1109/CIRA.2003.1222112.

[51] U. Kramer, G. Schuller, S. Wabnik, J Klier, and J. Hirschfeld. "Ultra Low Delay audio coding with constant bit rate". In: (Jan. 2004).

[52] K. Kubacki. "Jazz musicians: creating service experience in live performance". In: *International Journal of Contemporary Hospitality Management* 20.4 (2008), pp. 303–313. DOI: 10.1108/09596110810873516.

[53] Z. Kurtisi, X. Gu, and L. Wolf. "Enabling Network-Centric Music Performance in Wide-Area Networks". In: *Commun. ACM* 49.11 (Nov. 2006), pp. 52–54. ISSN: 0001-0782. DOI: 10.1145/1167838.1167862.

[54] Z. Kurtisi and L. Wolf. "Using wavpack for real-time audio coding in interactive applications". In: *2008 IEEE International Conference on multimedia and Expo*. 2008, pp. 1381–1384. DOI: 10.1109/ICME.2008.4607701.

[55] O. Lartillot and P. Toiviainen. "MIR in Matlab (II): A Toolbox for Musical Feature Extraction from Audio". In: *International Conference on Music Information Retrieval (ISMIR)*. 2007, pp. 287–288.

[56] J. Lazzaro and J. Wawrzynek. "A Case for Network Musical Performance". In: *Proceedings of the 11th International Workshop on Network and Operating Systems Support for Digital Audio and Video*. New York, NY, USA: Association for Computing Machinery, 2001, pp. 157–166. ISBN: 1581133707. DOI: 10.1145/378344.378367.

[57] P. Le Callet, S. Möller, and A. Perkis. *Qualinet White Paper on Definitions of Quality of Experience*. Tech. rep. European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003), 2012.

[58] S. Li and W. Deng. "Deep Facial Expression Recognition: A Survey". In: *CoRR* abs/1804.08348 (2018). arXiv: 1804.08348.

[59] R. Matei and J. Ginsborg. "Music performance anxiety in classical musicians: what we know about what works". In: *British Journal of Psychiatry International* 14 (May 2017), pp. 33–35. DOI: 10.1192/S2056474000001744.

[60] H. Misra, S. Ikbal, H. Bourlard, and H. Hermansky. "Spectral Entropy Based Feature for Robust ASR". In: *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. 2004. ISBN: 0-7803-8484-9. DOI: 10.1109/ICASSP.2004.1325955.

[61] A. Olmos, M. Brulé, N. Bouillot, M. Benovoy, J. Blum, H. Sun, N. Windfeld Lund, and J.R. Cooperstock. "Exploring the role of latency and orchestra placement on the networked performance of a distributed opera". In: *Annual International Workshop on Presence*. 2009.

[62] M. Osborne and J. Franklin. "Cognitive processes in music performance anxiety". In: *Australian Journal of Psychology* 54 (Aug. 2002), pp. 86–93. DOI: 10.1080/00049530210001706543.

[63] S.R Quackenbush. "Objective Measures of Speech Quality (Subjective)". PhD thesis. USA: Georgia Institute of Technology, 1985.

[64] A. Raheel, M. Majid, M. Alnowami, and S.M. Anwar. "Physiological Sensors Based Emotion Recognition While Experiencing Tactile Enhanced Multimedia". eng. In: *MDPI Sensors* 20.14 (July 2020), p. 4037.

[65] R. Renwick. "Sourcenode: A Network Sourced Approach to Network Music Performance (NMP)". In: *Proceedings of the International Computer Music Conference (ICMC)*. Sept. 2011. DOI: 10.13140/RG.2.1.1479.6000.

[66] B.H. Repp. "Sensorimotor synchronization: A review of the tapping literature". In: *Psychonomic Bulletin & Review* 12.6 (Dec. 2005), pp. 969–992. ISSN: 1531-5320. DOI: 10.3758/BF03206433.

[67] B.H. Repp and A. Penel. "Auditory dominance in temporal processing: new evidence from synchronization with simultaneous visual and auditory sequences." eng. In: *Journal of experimental psychology. Human perception and performance* 28 (5 Oct. 2002), pp. 1085–99.

[68] B.H. Repp and A. Penel. "Rhythmic movement is attracted more strongly to auditory than to visual rhythms." eng. In: *Psychological research* 68 (4 Aug. 2004), pp. 252–70.

[69] J.E. Resnicow, P. Salovey, and B.H. Repp. "Is Recognition of Emotion in Music Performance an Aspect of Emotional Intelligence?" In: *Music Perception: An Interdisciplinary Journal* 22.1 (2004), pp. 145–158. ISSN: 0730-7829. DOI: 10.1525/mp.2004.22.1.145. eprint: http://mp.ucpress.edu/content/22/1/145.full.pdf. URL: http://mp.ucpress.edu/content/22/1/145.

[70] V. Rojas-Mendizabal, A. Serrano-Santoyo, R. Conte, and A. Gomez-Gonzalez. "Toward a Model for Quality of Experience and Quality of Service in e-health Ecosystems". In: *Procedia Technology* 9 (Dec. 2013), pp. 968–974. DOI: 10.1016/j.protcy.2013.12.108.

[71] 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". In: *Journal of the AES* 63 (2015), pp. 864–875. DOI: 10.17743/jaes.2015.0074.

[72] C. Rottondi, C. Chafe, C. Allocchio, and A. Sarti. "An Overview on Networked Music Performance Technologies". In: *IEEE Access* 4 (2016), pp. 8823–8843. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2016.2628440.

[73] D. Sam. *Real-time online music performance: fact or fiction?* https://musicinteractiononline.wordpress.com/2020/06/18/real-time-online-music-performance-fact-or-fiction/. June 2020.

[74] R. Saputra and A. Prihatmanto. "Design and implementation of BeatME as a Networked Music Performance (NMP) system". In: *Proceedings of the 2012 International Conference on System Engineering and Technology (ICSET)*. Sept. 2012, pp. 1–6. ISBN: 978-1-4673-2375-8. DOI: 10.1109/ICSEngT.2012.6339349.

[75] A. Sawchuk, E. Chew, R. Zimmermann, C. Papadopoulos, and C. Kyriakakis. "From remote media immersion to Distributed Immersive Performance". In: *Proceedings of the 2003 ACM SIGMM workshop on Experiential Ttelepresence (ETP)*. 2003.

[76] K. Scherer. "Which Emotions Can be Induced by Music? What Are the Underlying Mechanisms? And How Can We Measure Them?" In: *Journal of New Music Research* 33 (Sept. 2004), pp. 239–251. DOI: 10.1080/0929821042000317822.

[77] N. Schuett. "The Effects of Latency on Ensemble Performance". MA thesis. CCRMA Department of Music, Stanford University, 2002.

[78] J. Selvaraj, M. Murugappan, R Nagarajan, and W. Khairunizam. "Physiological signals based human emotion Recognition: a review". In: *IEEE International Colloquium on Signal Processing and Its Applications (CSPA)*. Mar. 2011.

[79] S. I. Serengil and A. Ozpinar. "LightFace: A Hybrid Deep Face Recognition Framework". In: *Innovations in Intelligent Systems and Applications Conference (ASYU)*. Oct. 2020, pp. 1–5.

[80] D. El-Shimy and J. Cooperstock. "Reactive Environment for Network Music Performance". In: *Proceedings the International Conference on New Interfaces for Musical Expression (NIME)*. May 2013.

[81] C. Stais, Y. Thomas, G. Xylomenos, and C. Tsilopoulos. "Networked music performance over information-centric networks". In: *2013 IEEE International Conference on Communications Workshops (ICC)*. June 2013, pp. 647–651. DOI: 10.1109/ICCW.2013.6649313.

[82] Y. Taigman, M. Yang, M.A. Ranzato, and L. Wolf. "DeepFace: Closing the Gap to Human-Level Performance in Face Verification". In: *IEEE Conference on Computer Vision and Pattern Recognition*. Sept. 2014.

[83] M. Torcoli, T. Kastner, and J. Herre. "Objective Measures of Perceptual Audio Quality Reviewed: An Evaluation of Their Application Domain Dependence". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), pp. 1530–1541. DOI: 10.1109/TASLP.2021.3069302.

[84] J. Tribolet, P. Noll, B. McDermott, and R. Crochiere. "A study of complexity and quality of speech waveform coders". In: *ICASSP '78. IEEE International Conference on Acoustics, Speech, and Signal Processing*. Vol. 3. 1978, pp. 586–590. DOI: 10.1109/ICASSP.1978.1170567.

[85] G. Tzanetakis and P. Cook. "Musical genre classification of audio signals". In: *IEEE Transactions on Speech and Audio Processing* 10.5 (2002), pp. 293–302. DOI: 10.1109/TSA.2002.800560.

[86] J.M. Valin, T. Terriberry, and G. Maxwell. "A full-bandwidth audio codec with low complexity and very low delay". In: *17th European Signal Processing Conference* (2009), pp. 1254–1258.

[87] J.M Valin, G. Maxwell, T. Terriberry, and K. Vos. "High-Quality, Low-Delay Music Coding in the Opus Codec". In: *135th Audio Engineering Society Convention 2013* (Jan. 2013), pp. 73–82.

[88] E. Vincent, M. Jafari, and M. Plumbley. "Preliminary guidelines for subjective evaluation of audio source separation algorithms". In: *UK ICA Research Network Workshop*. Sept. 2006.

[89] C. Werner. "Distributed Network Music Workshop with Soundjack". In: *Proceedings of the 25th Tonmeistertagung*. 2008.

[90] M. Wozniewski, N. Bouillot, Z. Settel, and J. Cooperstock. "Large-Scale Mobile Audio Environments for Collaborative Musical Interaction". In: *Proceedings the International Conference on New Interfaces for Musical Expression (NIME)*. 2008.

[91] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. "Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks". In: *IEEE Signal Processing Letters* 23.10 (2016), pp. 1499–1503. DOI: 10.1109/LSP.2016.2603342.

[92] R. Zimmermann, E. Chew, S. Ay, and M. Pawar. "Distributed musical performances: Architecture and stream management." In: *ACM Transactions on Multimedia Computing, Communications, and Applications* 4 (Jan. 2008).