The MusiNet project: Towards unraveling the full potential of Networked Music Performance systems

D. Akoumianakis*, C. Alexandraki†, V. Alexiou‡, C. Anagnostopoulou§, A. Eleftheriadis‡, V. Lalioti§,
A. Mouchtaris‖**, D. Pavlidi‖**, G. C. Polyzos††, P. Tsakalides‖**, G. Xylomenos††, P. Zervas†
* TEI of Crete, Dept. of Informatics Engineering, Heraklion 71500, Greece
† TEI of Crete, Dept. of Music Technology and Acoustics Engineering, Rethymnon 74100, Greece
‡ University of Athens, Dept. of Informatics and Telecommunications, Athens 15784, Greece
§ University of Athens, Dept. of Music Studies, Athens 15784, Greece
‖ University of Crete, Dept. of Computer Science, Heraklion 70013, Greece
** FORTH-ICS, Heraklion 70013, Greece
†† Athens University of Economics and Business, Dept. of Informatics, Athens 10434, Greece
Email: da@epp.teicrete.gr, chrisoula@staff.teicrete.gr, valexiou@di.uoa.gr, chrisa@music.uoa.gr,
eleft@di.uoa.gr, vlalioti@music.uoa.gr, mouchtar@csd.uoc.gr, pavlidi@csd.uoc.gr, polyzos@aueb.gr,
tsakalid@csd.uoc.gr, xgeorge@aueb.gr, pzervas@staff.teicrete.gr

***Abstract*—The MusiNet research project aims to provide a comprehensive architecture and a prototype implementation of a complete Networked Music Performance (NMP) system. In this paper we describe the current status of the project, focusing on critical decisions regarding the system's architecture and specifications, the low delay audio and video coding techniques to be employed, the media relay design, and the synchronous and asynchronous collaboration algorithms to be adopted.**

I. Introduction

Networked Music Performance (NMP) systems allow geographically distributed musicians to collaborate, or even perform a live concert, via computer networks. NMP systems can potentially advance music creativity, education, and cross-cultural interaction, offering a common platform for social engagement among musicians of any culture and background.

Although NMP experiments date back as early as the 1970s [1], the first practical NMP systems were realized only during the last decade due to the appearance of high speed research/educational computer networks such as Internet2 in the USA and GEANT in Europe. The technological challenges of NMP systems include the elimination of latencies during recording, transmission, reception and reproduction of the audiovisual information, the elimination of network bandwidth bottlenecks, the constraints in synchronizing the exchanged information, and the high sensitivity to network data loss. The most important hurdle is that the *Ensemble Performance Threshold* (EPT), that is, the maximum delay between any two endpoints so that NMP is possible, must remain below 25 msec [2], while in teleconferencing systems the acceptable end-to-end delay can be 150 msec or higher [3]. As a result, NMP systems require ultra low delay solutions for media coding, transmission, relaying and decoding, each one a very challenging task on its own.

Previously proposed NMP systems include Jacktrip [4], Distributed Immersive Performance (DIP) [5], SoundJack [6] and DIAMOUSES [7]. These systems either pose significant limitations to NMP interaction, or require excessive amounts of resources and direct access to high speed networks. In the MusiNet project we are developing technologies that will significantly reduce the resource requirements of NMP and studying in depth the sociological and musicological aspects of such systems. The central objective of the MusiNet project is to address the main challenges of NMP systems, so as to make them available to musicians who have access to commonly available computing and communication resources. The proposed research follows a comprehensive interdisciplinary approach which is expected to unfold challenges, reveal limitations and envision novel uses of NMP systems.

The remainder of this paper is structured as follows. Section II presents existing work on NMP technologies, while Section III presents the framework of the MusiNet system, describing its architectural infrastructure and specifications, as well as the technologies that will be employed by the system. We conclude in Section IV.

II. Background and Related Work

Audio coding schemes for NMP have been reported in [8] and [9]. The former uses the proprietary Fraunhofer *Ultra Low Delay* (ULD) codec, while the latter uses the open source "Wavpack" codec. Popular compression algorithms for audio signals, such as MP3 [10], AAC [11] and Dolby AC-3 [12] are not suitable for NMP as, even though they achieve high compression ratios, they experience relatively high algorithmic delays for NMP purposes.

It has been observed that visual cues are very important in music collaboration; consider, for example, the conductor's role [13]. Video is demanding with respect to data rates, synchronization and delays. An important improvement in this area is the use of the *Scalable Video Coding* (SVC) [14] scheme which enables data to be relayed between the endpoints by a *Selective Forwarding Unit* (SFU) [15], which does not process the data, as opposed to the traditional *Multipoint*

Control Unit (MCU) used in conferencing. The use of the SFU offers end-to-end video delays in the order of 100-180 msec for hundreds of users in a conferencing scenario.

The successful implementation of NMP systems additionally involves the exchange of context-dependent information. Depending on the scenario (e.g., rehearsal, concert, class, improvisation), the collaboration environment must make provisions for a series of "boundary artifacts" and associated collaboration practices. Boundary artifacts refer to objects of practice that span boundaries between different social worlds (i.e., involved parties such as performers, composers, audience) or the boundary between the material and the virtual, while maintaining certain meaning in each case. For example, in distributed music lessons, it is often necessary to present a shared electronic music score to the remote participants, providing them with a shared context of reference and an additional interaction means for organizing distributed collaborative work [16], [17]. The music score may be designed so as to be sensible to the participants' local actions, in order to afford tracking and sharing contextual information in the course of a music performance.

III. The MusiNet system

This section provides an overview of the MusiNet project. We first refer to the architectural infrastructure and then we proceed with real-time audio and video coding and network delay mitigation techniques. Finally, we refer to synchronous and asynchronous collaboration techniques.

A. System architecture and specifications

The MusiNet architecture has been designed to ensure that the MusiNet system will a) be easy and intuitive to use, i.e., without requiring specialized knowledge on behalf of musicians, b) support synchronous and asynchronous collaboration activities c) allow multipoint communications, being scalable to multiple dispersed musicians, d) support capturing and playback of multiple streams of audio and video at each network location, e) be implementable in multiple platforms (at least Linux and Mac OSX) and, finally, f) handle network traffic according to a number of QoS related parameters.

The final architecture comprises two main entities: the *MusiNet Client*, which is the software used by distributed musicians, and the *MusiNet Center*, which is the central entity controlling all types of connections during an NMP session. There are three types of connections between clients and the center, namely SIP, RTP and HTTP connections. SIP connections are signalling connections, used to register each client to the MusiNet center (i.e., announce its current IP address), communicate presence information (i.e., reveal the ability or willingness of each musician to communicate with other registered peers), as well as call other clients and accept or reject incoming calls. Once a connection of two or more peers has been established using SIP messages, a number of RTP connections are activated to transfer the actual media (i.e., audio and video) streams. Finally, HTTP connections from each client to the MusiNet Center are used to exchange textual data that are complementary to the media transfers.

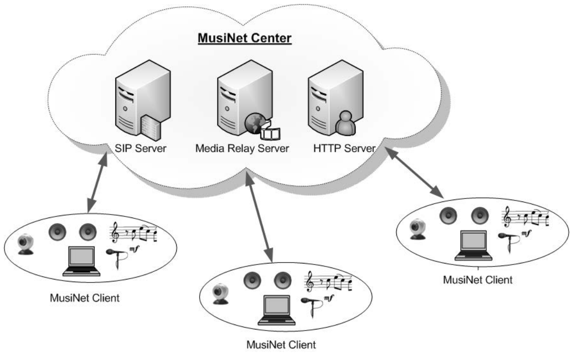

Fig. 1. The general architecture of MusiNet.

Consequently, as shown in Fig. 1, the MusiNet Center encapsulates three components: a SIP server, a Media Relay Server and an HTTP server. Besides signalling, the SIP server allows overcoming firewall issues via NAT traversal, which is a common problem in realistic NMP sessions. In addition, the SIP server controls the behavior of the Media Relay Server, based on the SIP messages exchanged with the clients.

The Media Relay Server is an SFU: it is responsible for selectively forwarding the media streams from each participating network node to the intended recipients. It does not process media streams in any way, other than by relaying them to the appropriate recipients. For example, MusiNet clients may not desire to receive the media streams produced by all other clients, in which case the Media Relay Server would selectively replicate the incoming streams towards the minimum number of clients. The alternative would be to exchange media streams directly among peers, in a peer-to-peer or mesh topology. The benefit of the Media Relay Server is that in this way, clients do not need to send their transmitted streams to multiple recipients (i.e., to all remaining peers) but only to a single destination (i.e., the media relay server), hence saving outbound bandwidth which is commonly low in network infrastructures available to consumers [18].

Finally, the HTTP server is responsible for maintaining supplementary session data, which may convey either information necessary for supporting musician collaboration on the graphical level, or any other type of supplementary information deemed appropriate or relevant. Furthermore, in cases that the online ensemble comprises members in different roles who decide to intertwine with or notify the performers in real time, such interventions should be facilitated in a manner that does not disturb (technically or otherwise) the actual performance.

Furthermore, MusiNet clients may communicate with external components (e.g., cloud services and dedicated toolkits) intended to facilitate pre- and/or post-performance negotiations between peers on material related to the actual performance (e.g., exchanging musical material, scheduling synchronous performances, etc.), as discussed in section III-F.

B. Real-time audio coding with spatial attributes

In a NMP session, where potentially groups of remote musicians are involved (i.e., more than one participant at each remote site), it is essential to allow musicians to interact in a natural and flexible manner. Natural interaction implies the reproduction of a physical acoustic scene, including perceptually relevant attributes, such as the direction of the audio sources. Flexible interaction underlines the ability of a participant to define his/her desired acoustic scene, e.g, the ability to focus on some of the audio sources and deemphasize/mute some others. This task demands the efficient recording, compression, transmission and reproduction of each physical site to the interaction venues. At the same time, the temporal threshold for feasible human audio interaction must be satisfied. In audio coding algorithms, this is determined by the audio frame size used, since the last sample has to be acquired before the processing of the first sample of a frame begins. The audio coding research community has proposed a few but very effective ultra-low delay coding solutions (i.e., using small frame sizes) during the past years. The most popular are the *Advanced Audio Coding-Low Delay* (AAC-LD) algorithm [19], the *Ultra Low Delay* (ULD) audio coding algorithm [20] and the *Constrained Energy Lapped Transform* (CELT) codec, recently merged into the Opus Codec [21].

According to objective and subjective evaluation measurements, CELT outperforms other ultra-low delay codecs with very good perceived audio quality and reasonable bit rate requirements (see [21] and its references). Moreover it is an open-source codec, with no proprietary rights. In the Opus context, the supported sampling rates are 8, 12, 16, 24 and 48 kHz and the algorithmic delay can vary from 2.5 to 20 ms. Bit rates from 6 to 510 kbps are possible. For example, for a frame size equal to 20 ms and stereo music encoding it is recommended to use bit rates in the range of 64 to 128 kbps.

For the recording and reproduction of the interacting venues we propose the use of spatial audio techniques, aiming at rendering the spatial attributes of each audio scene along with the audio data, thus achieving a more realistic audio impression as depicted in Fig. 2. For this purpose we intend to adopt the *Immersive Audio Communication System* (ImmACS) introduced in [22] and [23] for spatial audio attributes estimation and spatial audio reproduction, respectively.

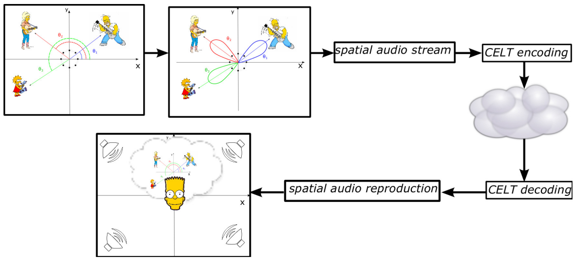

Fig. 2. Spatial audio recording and reproduction in MusiNet.

In [22] the authors propose a novel method for estimating the number and *directions of arrival* (DOAs) of the active sources in an audio scene, using a uniform circular array comprised of omnidirectional microphones. The basic assumption of the method is that the sources overlap in the *Short Time Fourier Transform* (STFT) domain, except in a few constant time zones where only one source is active. Based on this assumption, "single source zones" are identified and over these zones a single source DOA estimation algorithm is applied. The DOA estimates from all detected single-source zones form a histogram over which a matching pursuit inspired method is applied in order to accurately estimate the number of the active audio sources and their corresponding DOAs. The method of [23] is comprised by an analysis and a synthesis stage. At the analysis stage the number and the DOAs of the audio sources are estimated using the method of [22] and the estimated audio signals are efficiently downmixed in a single audio stream. The flow diagram of the analysis stage is depicted in Fig. 3. At the synthesis stage the signal of each loudspeaker is estimated with the *Vector Based Amplitude Panning* (VBAP) method [24], using as input the received downmixed audio stream and side information.

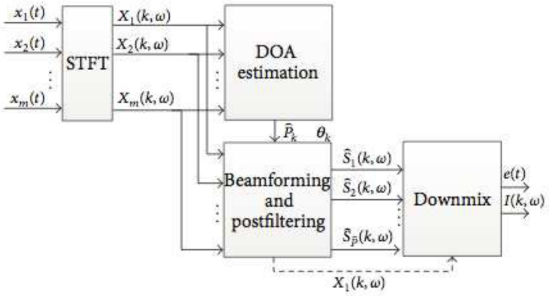

Fig. 3. Flow diagram of the analysis stage of [23].

Using ImmACS we can achieve efficient recording of the multiple audio sources along with their corresponding spatial information, producing a single downmixed audio stream and side information, thus reducing the bandwidth requirements of the NMP system. In order to incorporate spatial audio recording and reproduction in MusiNet, we have to take into account the constraints that each system sets. More specifically, we have to efficiently encode/decode the side information of [23] and transparently merge it into the CELT encoded downmixed audio in the RTP stream. We also need to synchronize the input frame rate of the two methods, since the CELT codec requires a high frame rate in order to satisfy the EPT, while [23] has been tested so far with frames of 46 ms. Depending on the NMP scenario, we may also need to consider applying an *Acoustic Echo Cancellation* (AEC) algorithm, for example, when the collaborating musicians use loudspeaker reproduction systems.

C. Delay estimation of real-time audio streams

As already mentioned, delay is the most significant parameter in enabling music collaboration. With the stringent requirements of NMP, it is important to quantify with high accuracy the delay across the entire path from capture, to encoding and transmission, decoding, as well as playback. The exact value of the delay that each participant is subject to can be used by the system in order to improve his/her experience. One example is tempo adaptation [25]. In various experiments that have been conducted, a general inverse relationship between tempo and delay was reported. For instance, slowing down the tempo of a particular musician who experiences a bigger delay than the rest of his group, can be beneficial and can enhance an ensemble's ability to play synchronously and with comfort.

In [26] the construction of inaudible pilot signals and associated receiver processing techniques are described, that can be used in audio communication, so that an accurate estimation of the signal's end-to-end delay can be obtained. In this work, the properties of the human auditory system are considered, and the pilot signal is structured so that it is a) acoustically untraceable and b) able to maximize the time accuracy of the delay estimation. An online proposal and a signal design that exploits more complicated perceptual processing taking part in the human auditory system, are currently under investigation, and would enhance its practical use.

Finally, given that each stream (audio and video) from each user is handled independently in our architecture, and does not get mixed with other streams (i.e., as in MCUs), assuming a collaboration scheme of N users, each subject becomes an endpoint where N-1 streams are received with N-1 different delay values. Applying the aforementioned delay estimation technique to audio streams, which are the most critical in a music ensemble, we will be able to quantify the maximum mix phase difference (the mix phase captures the differences in the mixes between different nodes) that enables the collaboration in a multi-point NMP. In the simplest case, assuming two users, extensive experimentation in the desirable and realistic environment could be done, in order to determine the EPT for a specific tempo, genre and instrumentation combination since, according to the bibliography, a global EPT cannot be defined, as its value is a function of these parameters.

D. Real-time video coding

Current practice shows that the visual component of communication is extremely important, particularly if the interaction lasts for more than 15 minutes. For short-duration communication, sound (voice) is a sufficient modality (e.g., traditional telephony or *Voice over IP* (VoIP) calls). For long-duration (e.g., 1 hour or more) collaborations however, video helps to keep the team focused on the task at hand. The state-of-the-art in multipoint video communication uses the H.264 SVC standard [14], [27], which has been incorporated into the H.264 standard as Annex G. SVC allows the encoding of video in a number of layers, in a pyramidal fashion. The encoder produces a base layer that offers a representation of the original signal at a certain fidelity. Then, a number of enhancement layers are constructed, increasing the quality in one of three dimensions: temporal, spatial, or SNR (quality). The pyramidal construction means that a particular layer requires all lower layers that it is referring to in order to be decoded. The different layers are multiplexed into a single bitstream, with appropriate framing that allows one to know which layer a particular set of data corresponds to, without having to decode video data.

SVC configurations that have proved to be particularly effective are those that use hierarchical P pictures for temporal scalability, as well as spatial scalability. Fig. 4 shows a typical case with three temporal layers and two spatial layers. The arrows indicate the direction of prediction. "B" indicates the base spatial layer and "S" indicates the spatial enhancement. The number following the spatial layer indication shows the temporal layer the particular layer belongs to.

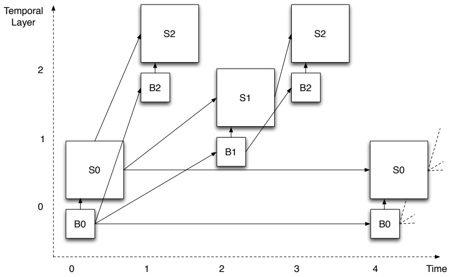

Fig. 4. Spatio-temporal coding structure in SVC.

The use of SVC together with an SFU to relay media [15] allows communication with a delay of less than 160 msec. The upper bound for interactive communication according to ITU-T Rec. G.114 is 180 msec. This delay is a considerable improvement compared to the delay of traditional MCUs, where the delay reaches 400 msec or more. Still, this is not sufficient for NMP applications: delay needs to be reduced by one more order of magnitude, to 25 msec or less. Towards this goal, we will investigate coding techniques for very high frame rate coding with a very large number of slices, so that the time required for digitization and encoding is minimized. The use of spatial and temporal scalability is necessary in order to enjoy SVC's error robustness benefits. The delays must be harmonized with the corresponding delays of the audio encoders, so that the total system delay is minimized.

Delay can be minimal in the case of video, as long as the packet size can be very large. This property is significant because it relates to the overhead associated with packet-based transport: with small packet sizes, the header overhead can become comparable in terms of bitrate to the actual information that is transmitted. With large packets, this overhead can become very small or negligible. We can significantly reduce the quality of the video signal in cases where available network bandwidth is scarce, as long as synchronization is maintained, since the human eye is more tolerant to loss than the human ear (especially in music applications).

E. Media relaying

The end-to-end delay in an NMP system includes network delays which are beyond the control of regular users. There is, however, another element of delay between the endpoints: the Media Relay Server which receives media from all NMP participants and relays them appropriately. In traditional conferencing, a MCU decodes all incoming media streams, composes them (e.g., mixing all audio streams or selecting only the current speaker, and composing all video streams into a single video pane), re-encodes the result and separately transmits it to each participant. In an NMP system, participants would prefer to choose how to mix the media streams themselves, e.g., the drummer of a rock band may want to hear the bass guitar more than the vocals. For this reason, in the MusiNet project an SFU is used as a Media Relay Server [15]: each participant indicates to the SFU which media streams it desires to receive, and the SFU simply replicates and forwards these streams. The operation of such a unit is shown in Fig. 5.

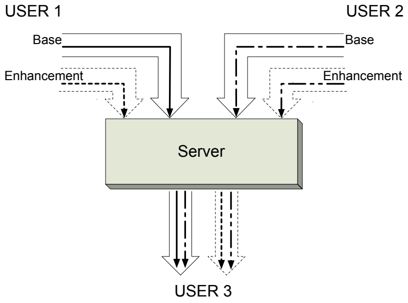

Fig. 5. The operation of the Media Relay Server (SFU).

Even this approach introduces SFU delays of up to 20 ms, which are acceptable for conferencing but not for NMP. Therefore, some NMP approaches employ direct communication between participants. This is inefficient with multiple participants, as it requires transmitting a copy of each media stream to every other participant. For this reason, in previous work we studied the option of avoiding the SFU by letting NMP participants directly multicast their media streams to all other participants [28]. While this does indeed reduce latency, it requires network protocols capable of multicasting data, which are not currently deployed on the Internet at large. As a result, the SFU remains essential for media routing, hence it makes sense to reduce its latency as much as possible.

In order for an SFU to route media packets, it receives signaling and media packets from the NMP participants. The signaling packets indicate which media streams are desired by a recipient, i.e., which video and audio streams the participant wants to receive; these are produced by the SIP server. The media packets contain the media streams produced by each participant and are sent directly to the SFU. Based on the signaling packets, the SFU maintains a data structure indicating how to treat the packets of each media stream, i.e., drop them, or replicate them for each participant that has requested them. The data flow in the SFU is therefore as depicted in Figure 6(a): packets are passed to the SFU process by the kernel, the SFU replicates them for each recipient, and the replicated packets are passed to the kernel for transmission. This requires context switching between the kernel and the application, as well as data copying which grows with the number of participants, as each packet transmission requires a separate system call. Both these activities (copying and system calls) are well-known sources of delay.

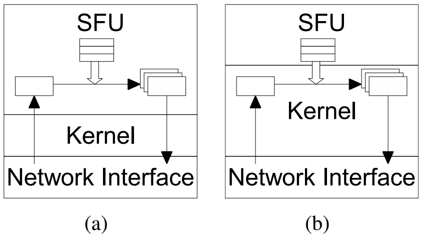

Fig. 6. (a) Regular SFU operation, (b) SFU with netmap.

The approach that we are currently working on is to use the netmap framework for packet handling at the application level [29]. In this case, the entire SFU resides at the user level, as shown in Figure 6(b), so both control signalling and user packets are handled at the user level by a program written in a high level language. However, no packet exchanges are needed between the kernel and the user level, since the SFU application can directly manipulate the media packets in kernel memory, thus reducing context switching overhead [30]. An optimized data structure is used to look up the required destinations for each packet. Our initial prototype exhibits very promising performance, offering far lower delays than a regular application level SFU, so we are currently experimenting with techniques to avoid packet copying within the kernel, e.g., use the same copy for multiple transmissions when a packet must be transmitted to multiple destinations.

F. Synchronous and asynchronous collaboration techniques

Since MusiNet is intended to facilitate collaboration, it should make provisions for developing a special vocabulary [31] which is multi-platform, modular, and extensible. The core functional scope of such a music vocabulary is to facilitate both synchronous and asynchronous interactions between remote peers either prior, during, or after NMP. Such a vocabulary should be designed to establish the language for co-engagement in a linguistic domain, that of music, as well as to exhibit quality attributes that alleviate proximity and thematic boundaries that prevail in distributed settings.

To this effect, we have used the JMusic (JM) library[1] which offers the primitive interaction components and supports qualities such as augmentation, extensibility and integration that are deemed as crucial for collaborative music making [32]. Specifically, augmentation will allow "injection" of additional special-purpose interaction facilities needed to bypass inherent structural limitations of the JMusic library. For instance, JMusic adopts a graphics-based interaction model which should be augmented both at the level of interaction objects and event handling. The extension of JMusic will allow for novel interaction facilities such as social awareness, translucence, activity monitoring and digital trace data that are important in collaborative settings. Finally, integration will establish bridges with functionalities offered by emerging

[1] http://sourceforge.net/projects/jmusic/

technologies, e.g., Google APIs and cloud services as means for sharing, synchronizing clients, collaborating and managing multiple views and interactive manifestations of music.

The specific design requirements for the above are being compiled by envisioning scenarios. The guiding principles for these scenarios rely on semiotic engineering principles that (a) qualify virtual work in relations to what is digitized versus what is virtualized and (b) anchor co-engagements in music practices as operations with, on, through, or within representations. This, in turn, brings to the forefront a socio-material design perspective which is tuned to the purpose so as to offer the required insights.

IV. Conclusion

We have provided an overview of the MusiNet project which is building an NMP system based on commodity computing and communication components. MusiNet is currently conducting research on many areas, including ultra-low delay audio coding, scalable video coding, low delay relaying and musician collaboration techniques. The eventual goal of the project is to integrate all these techniques into a unified prototype, using existing components for the endpoints and servers, augmented with our own research results.

Acknowledgment

This research has been co-financed by the European Union (European Social Fund–ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF)–Research Funding Program: THALIS–MusiNet (http://musinet.aueb.gr).

References

[1] G. Föllmer, "Electronic, aesthetic and social factors in net music," *Organized Sound*, vol. 10, no. 3, pp. 185–192, Dec. 2005.

[2] C. Chafe, M. Gurevich, G. Leslie, and S. Tyan, "Effect of time delay on ensemble accuracy," in *International Symposium on Musical Acoustics*, 2004.

[3] *Recommendation P.861: Objective Quality Measurement of Telephone-Band (300-3400 Hz) Speech Codecs*, International Telecommunications Union-Telecommunication Standardization Sector Std., Feb. 1998.

[4] J.-P. Cáceres and C. Chafe, "JackTrip: Under the hood of an engine for network audio," in *Proceedings of International Computer Music Conference*, 2009, p. 509–512.

[5] A. A. Sawchuk, E. Chew, R. Zimmermann, C. Papadopoulos, and C. Kyriakakis, "From remote media immersion to distributed immersive performance," in *ACM SIGMM Workshop on Experiential Telepresence (ETP)*, 2003, pp. 110–120.

[6] A. Carôt, A. Renaud, and V. B., "Network music performance (NMP) with Soundjack," in *NIME Conference*, 2006.

[7] C. Alexandraki, P. Koutlemanis, P. Gasteratos, N. Valsamakis, D. Akoumianakis, G. Milolidakis, G. Vellis, and K. N., "Towards the Implementation of a Generic Platform for Networked Music Performance: The DIAMOUSES approach," in *International Computer Music Conference (ICMC)*, Aug. 2008, pp. 251–258.

[8] U. Krämer, H. Jens, G. Schuller, S. Wabnik, A. Carôt, and C. Werner, "Network music performance with ultra-low-delay audio coding under unreliable network conditions," in *Audio Engineering Society Convention 123*, Oct. 2007.

[9] Z. Kurtisi and L. Wolf, "Using wavpack for real-time audio coding in interactive applications," in *IEEE International Conference on Multimedia and Expo*, June 2008, pp. 1381–1384.

[10] M. Bosi, K. Brandenburg, S. Quackenbush, L. Fielder, K. Akagiri, H. Fuchs, M. Dietz, J. Herre, G. Davidson, and Y. Oikowa, "ISO/IEC MPEG-2 Advanced Audio Coding," *Journal of Audio Engineering Society*, vol. 45, no. 10, pp. 789–814, 1997.

[11] K. Brandenburg, "MP3 and AAC Explained," in *AES International Conference on High-Quality Audio Coding*, August 1999.

[12] M. Davis, "The ac-3 multichannel coder," in *Convention of the Audio Engineering Society*, Oct 1993.

[13] C. Alexandraki and I. Kalantzis, "Requirements and application scenarios in the context of network based music collaboration," in *The AXMEDIS Conference i-maestro workshop*, Nov. 2007, pp. 39–46.

[14] Y.-K. Wang, M. Hannuksela, S. Pateux, A. Eleftheriadis, and S. Wenger, "System and transport interface of SVC," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 9, pp. 1149–1163, Sept 2007.

[15] A. Eleftheriadis, R. Civanlar, and O. Shapiro, "Multipoint videoconferencing with scalable video coding," *Journal of Zhejiang University SCIENCE A*, vol. 7, no. 5, pp. 696–705, April 2006.

[16] D. Akoumianakis, G. Vellis, I. Milolidakis, D. Kotsalis, and C. Alexandraki, "Distributed collective practices in collaborative music performance," in *International Conference on Digital Interactive Media in Entertainment and Arts (DIMEA)*, 2008, pp. 368–375.

[17] G. Hajdu, "Real-time composition and notation in network music environments," in *International Computer Music Conference (ICMC)*, 2008.

[18] C. Alexandraki and D. Akoumianakis, "Exploring new perspectives in network music performance: The DIAMOUSES framework," *Computer Music Journal*, vol. 34, no. 2, pp. 66–83, Jan. 2010.

[19] R. Geiger, M. Lutzky, M. Schmidt, and M. Schnell, "Structural analysis of low latency audio coding schemes," in *Audio Engineering Society Convention 119*, October 2005.

[20] S. Wabnik, G. Schuller, and F. Kraemer, "An error robust ultra low delay audio coder using an MA prediction model," in *IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP)*, 2009, pp. 5–8.

[21] J.-M. Valin, G. Maxwell, T. B. Terriberry, and K. Vos, "High-quality, low-delay music coding in the opus codec," in *Audio Engineering Society Convention 135*, Oct. 2013.

[22] D. Pavlidi, A. Griffin, M. Puigt, and A. Mouchtaris, "Real-time multiple sound source localization and counting using a circular microphone array," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2193–2206, 2013.

[23] A. Alexandridis, A. Griffin, and A. Mouchtaris, "Capturing and reproducing spatial audio based on a circular microphone array," *Journal of Electrical and Computer Engineering*, vol. 2013, Feb. 2013.

[24] V. Pulkki, "Virtual sound source positioning using vector base amplitude panning," *Journal of Audio Engineering Society*, vol. 45, no. 6, pp. 456–466, 1997.

[25] A. Barbosa, "Displaced soundscapes," Ph.D. dissertation, Pompeu Fabra University, 2006.

[26] V. Alexiou and A. Eleftheriadis, "Real-time high-resolution delay estimation in audio communication using inaudible pilot signals," in *Submitted to the International Symposium on Communications, Control and Signal Processing (ISCCSP)*, 16 December 2014.

[27] H. Schwartz, D. Marpe, and T. Wiegand, "Overview of the scalable video coding extension of the H.264/AVC standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 9, pp. 1103–1120, September 2007.

[28] C. Stais, Y. Thomas, G. Xylomenos, and C. Tsilopoulos, "Networked music performance over information-centric networks," in *IEEE IIMC Workshop*, 2013.

[29] L. Rizzo, "Netmap: a novel framework for fast packet i/o," in *USENIX Advanced Technology Conference*, 2012.

[30] G. Xylomenos, C. Tsilopoulos, Y. Thomas, and G. C. Polyzos, "Reduced switching delay for networked music performance," in *Packet Video Workshop (Poster Session)*, 2013.

[31] D. Akoumianakis and C. Alexandraki, "Collective practices in common information spaces: Insight from two case studies," *Human-Computer Interaction*, vol. 27, no. 4, pp. 311–351, Oct. 2012.

[32] D. Akoumianakis, "Managing universal accessibility in software-intensive projects," *Software Process: Improvement & Practice*, vol. 13, 2008.