EEG Hyperscanning in the Internet of Sounds: Low-Delay Real-time multi-modal Transmission using the OVBOX

Giso Grimm
Department of Medical Physics and Acoustics
Carl von Ossietzky Universität
Oldenburg, Germany
g.grimm@uol.de
ORCID: 0000-0002-9685-4507

Mareike Daeglau
Neuropsychology Lab
Carl von Ossietzky Universität
Oldenburg, Germany
mareike.daeglau@uol.de

Volker Hohmann
Department of Medical Physics and Acoustics
and Cluster of Excellence "Hearing4all"
Carl von Ossietzky Universität
Oldenburg, Germany
volker.hohmann@uol.de
ORCID: 0000-0001-7056-1880

Stefan Debener
Neuropsychology Lab
and Cluster of Excellence "Hearing4all"
Carl von Ossietzky Universität
Oldenburg, Germany
stefan.debener@uol.de

***Abstract*—The OVBOX (ORLANDOviols consort box) is a tool for immersive network audio transmission with low delay and interactive spatial rendering on each client. In this paper, we demonstrate its extension to the transmission of electroencephalography (EEG) signals, which forms the basis for low-delay hyperscanning in telepresence scenarios. In addition to microphone signals, motion data, EEG signals and any other data can be exchanged in the Open Sound Control (OSC) or Lab Streaming Layer (LSL) protocol. This makes it possible to analyse several modalities simultaneously over the Internet. The methods used are described, a performance analysis and example hyperscanning data of a motor/audio task are presented.**

***Index Terms*—low delay network audio, hyperscanning, multi-modal data transmission**

I. Introduction

The study of social interaction is an important area of research in psychology and neuropsychology, as well as in applied fields like hearing aid research. One specific focus is on interpersonal synchrony [1], where studies have shown that both behavior and physiological characteristics can synchronize between people during interactions like conversations. Social interactions help people understand each other and form social structures such as families [2]. While the social nature of humans has been known for a long time, the study of brain activities during social interactions has only begun in the last decade [3]. New research in social neuroscience suggests that studying the brain activities of people interacting can give insights into their mental processes [4]. Methods like hyperscanning and pseudo-hyperscanning are used to measure how brains synchronize. Hyperscanning records brain activities of participants at the same time, while pseudo-hyperscanning does this for one person at a time [5]. The first study on brain activities between people using electroencephalography (EEG) was by Duane and Behrendt (1965), who recorded the brain activities of identical twins at the same time and compared their EEGs [6]. Recent reviews have looked at current methods used in hyperscanning studies, showing the progress and challenges in measuring brain coupling [7].

In traditional hyperscanning studies investigating interpersonal synchrony, the test subjects are typically in close proximity to each other so that several EEG systems can be used to record simultaneously [8]. If the test subjects are together in the same room, it is not easy to manipulate individual aspects of the interpersonal parameters, such as movement behaviour, time delay of the speech signal, or similar aspects of the communication. To overcome this problem, earlier studies were carried out with communication situations in telepresence [9]–[11]. If telepresence is to be used as a model for face-to-face communication, then similar criteria apply to the time delay requirements from mouth to ear (end-to-end delay) as for networked music performances (NMP). It is therefore recommended that, rather than utilising conventional video conferencing tools, specialised audio transmission tools with minimal delay be employed, as they are also employed for NMP. This makes the research of acoustic and multimodal communication via telepresence an application of the Internet of Sounds [12]. Several studies on EEG hyperscanning in

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 352015383 – SFB 1330 and SPP 2236 project 444761144.

musicians have been performed locally [13]–[15]. However, the integration of EEG transmission into a NMP system is a novelty in the field of the Internet of Sounds. In the communication research studies via telepresence by [9]–[11], a system for audio communication in low audio latency via the Internet, the OVBOX [16], [17], was used. This system was originally developed to enable rhythmic music communication via the internet. In the terminology of [12] it is a "Sound Thing" in an Internet of Music Things (IoMT), together with a session management service dedicated to NMP. In addition to audio data, this system is also able to exchange other data such as head movement data with low latency. A number of similar tools for NMP are available (see, for example, [18]–[21]). For an overview of these tools, see also [12]. The OVBOX system was used in this study, as well as in previous studies conducted by the authors [9]–[11], [22], due to its status as an extremely open and extensible system, coupled with the use of a low-delay interactive spatial audio engine for the rendering of immersive interactive audio. In addition, a method for the transmission of motion data for the remote rendering of moving sources has already been implemented in this tool [16]. This extension forms the basis for the addition of further modalities.

The objective of this study is to investigate whether the integration of EEG hyperscanning with telepresence into the Internet of sounds is feasible, using the OVBOX system. Previous studies have demonstrated that an internet connection can be employed to perform hyperscanning with magnetoencephalography (MEG) [23]. In their study, audiovisual communication between two MEG laboratories was achieved with a latency of just over 200 milliseconds. A common time stamp was generated via the Global Positioning System (GPS) and the data was recorded in a decentralised manner.

The specific aim of the present study was to transmit the EEG signals in real time and to achieve transmission times required for rhythmic interactions, which is below 50 ms end-to-end delay [24]. Different methods of synchronising the recording time were compared. Four different communication conditions, each with two test subjects, served as proof of concept. The two laboratories where the participants were placed were physically separated from each other and only connected via the Internet.

II. Methods

A. Apparatus

An overview of the network structure, distributed over several laboratories and servers, is shown in Figure 1. The first laboratory, hereafter designated as 'Lab 1', was situated within the Neuropsychology laboratory at the Carl von Ossietzky Universität Oldenburg. For reference, please refer to Figure 2. The second laboratory, designated as 'Lab 2', was a room within the Auditory Signal Processing group at the same university (see Fig. 3). The distance between the two laboratories was approximately 2 km. In both laboratories, a 32-channel EEG system with a mbt Smarting Pro amplifier was employed to measure the EEG signal of the test participants, configured to a

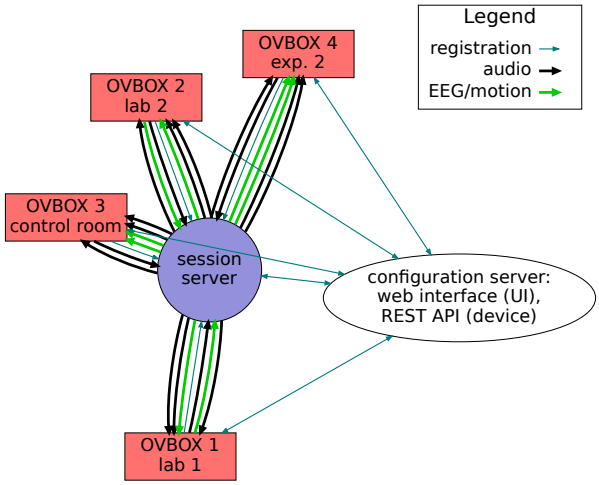

Fig. 1. Network structure of the experiment. The OVBOX systems in Lab 1 and Lab 2 transmitted audio and EEG data, and the device in Lab 2 also transmitted motion data. The OVBOX systems in the control room and experimenter 2 were used for data logging and voice communication. All OVBOX systems were configured via a central configuration server. All data was routed through a session server.

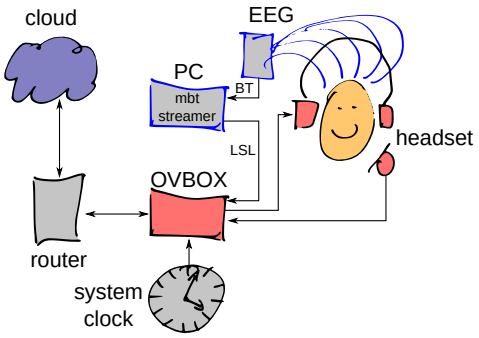

Fig. 2. Setup in laboratory 1: A test participant listens or speaks in this laboratory. The electroencephalography (EEG) is recorded with a Bluetooth EEG amplifier and made available in the local network as a Lab Streaming Layer (LSL) stream using the manufacturer's streaming software. The OVBOX sends the EEG signal, microphone signal and system clock to the session server ('cloud').

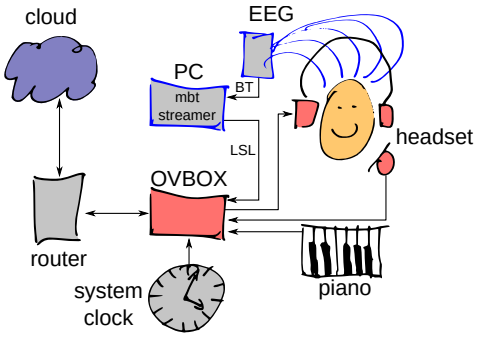

Fig. 3. Setup in laboratory 2: The same setup as in laboratory 1, except for an additional e-piano, which is sent as a second audio stream.

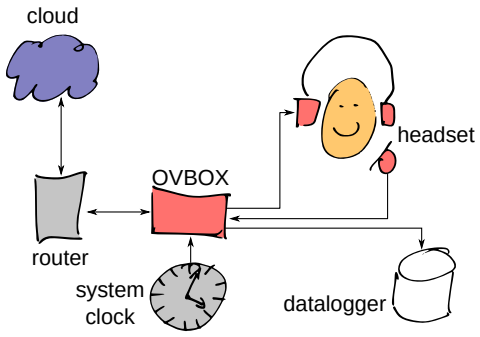

Fig. 4. Setup in the control room: This is the control room in which the experimenter is monitoring and recording the data, and can provide instructions to the test participants.

sampling rate of 250 Hz. The EEG system in laboratory 2 was configured to transmit 10 additional channels, which included data pertaining to head movements. The EEG stream was converted from Bluetooth to the Lab Streaming Layer (LSL) [25] using the streamer software of the amplifier, running on a dedicated computer. LSL is a network protocol for streaming EEG and other data within a local network. The use of LSL was essential with the amplifier in question, as the software does not offer any other interface and the manufacturer does not provide an SDK for direct access to the amplifier's Bluetooth protocol. This LSL stream was converted into a series of Open Sound Control (OSC) [26] messages with User Datagram Protocol (UDP) transport layer for transmission to the remote laboratories. By using UDP transport, the generic UDP data forwarding method of the OVBOX system could be used to transmit the EEG data. UDP is preferred over TCP or other higher level protocols such as Virtual Private Networks (VPN) or WebRTC because, due to the lack of error correction, significantly lower transmission delays and jitter can be achieved at the cost of a slightly higher risk of packet loss; see also [16] for discussion. For the purposes of speech communication and multimodal data transmission, an OVBOX system was used, comprising a Raspberry Pi 4B computer and a Focusrite Scarlett solo USB audio interface, to which an AKG HSD271 headset was connected. In the second laboratory, a second audio channel, recorded from the output of an electric piano (Yamaha P35B), was transmitted.

The data were recorded on a third system in the control room. This recording setup was situated in the same physical space as that of 'Lab 2'. It consisted of an OVBOX system combined with the datalogging sytem of the acoustic simulation and laboratory software TASCAR [27]. This system was running on a desktop computer. The experimenter, who was monitoring the system and controlling the data recording, was able to communicate with the test participants via a headset connected to the OVBOX system. A fourth OVBOX system, again running on a Raspberry Pi 4B, was used to facilitate communication between a second experimenter giving instructions and the test participants. While this device was not strictly necessary in this setup, it demonstrates that it is in principle possible to extend the setup to include more participants.

In this setup, the OVBOX systems were configured to use a central session server that acts as a selective forwarding unit (SFU) (see [16] for an overview). In this way, any UDP data can be exchanged between the clients. All data exchanged between the labs was routed through this server, which was located in Frankfurt am Main, Germany, which is about 350 km as airline distance from Lab 1 and Lab 2. In this way, all data travelled approximately 700 km.

B. Recording conditions

In order to test whether EEG hyperscanning using the OVBOX system is possible in general and whether low-delay critical human interaction can be analysed using this system, four different recording conditions were tested. As the results serve only as a proof of concept, only one participant in each laboratory was tested, and no statistical analysis of the data was applied.

In the first recording condition, the task of the participant in laboratory 1 was to play a note on a piano at irregular times. The task of the participant in laboratory 2 was to listen to this tone. The hypothesis is that motor potentials can be observed before the tone in participant 1 and auditory evoked potentials after the tone in both participants.

The second recording condition consisted of a clapping task. The task of one participant - the leader - was to clap at a rate of approximately 60 claps per minute. The task of the other participant - the follower - was to join in and clap along. Data were recorded for both combinations of leader and follower roles. It is hypothesised that this system will allow synchronised joint clapping and that the recorded EEG systems will show synchronised neural activity between the participants, reflecting the interpersonal synchrony achieved during the clapping task.

The third recording condition was a forced turn taking situation. The task of the participant in laboratory 1 was to name a city. The participant's task in laboratory 2 was to name a city beginning with the same letter as quickly as possible. This task was designed to test whether automatic labelling of speech activities is possible.

The fourth recording condition was a free conversation about superficial topics between the two participants. The aim of this recording condition was to analyse turn-taking timing.

C. Overview of clocks and temporal alignment of datalogging

Many clocks are involved in this set-up. A primary timeline is required to synchronise the data from all labs. The timeline of the data acquisition system is a monotonic clock in the control room, which is driven by the sample clock of the control room's audio system and counts the seconds since the start of the measurement. This timeline, referred to below as t_{rec}, was chosen as the primary timeline.

In addition, the system clock of each OVBOX system was sampled every 2 ms and transmitted to the data logger. On each system, the clock is controlled by a subset of the network time

protocol (NTP), which uses the automatically configured time server for synchronisation.

The time stamps of the EEG system were embedded in the EEG LSL streams. These time stamps have an arbitrary offset and are monotonically increasing. Similarly, the local session time of each OVBOX system was added to the sound pressure level streams, which is the monotonically increasing time in seconds since session start, based on the clock of the audio system.

The different clock times may drift apart. However, since all of these times have low jitter, they can be used to compensate for the jitter that occurs when transmitting over the Internet (for a detailed discussion of jitter in Internet transmission, see [16]).

D. End-to-End delay and network jitter measurement

Two different methods were used to estimate the end-to-end delay τ_{network}. In the first method, the round-trip time of the network connection (often called ping time) is measured. The end-to-end network delay between the laboratories is half of the round-trip time that is regularly measured in the OVBOX. This method assumes that the end-to-end delay is symmetrical, regardless of the direction of signal flow.

In the second method, the system time of the labs is used: The end-to-end network delay between a laboratory and the data recording system is the difference between the system time of the recording system and the system time of the laboratory at a given recording time. It is assumed that the system clocks are well synchronised and show neither drift nor an absolute difference.

The network jitter is the variance of the round-trip time, here characterized by the difference between the 99% quantile and the minimum.

E. Jitter and delay compensation

Because data is sent over a network link shared by many other independent processes and users, the arrival time of periodically sampled signals will be non-periodic and, in the worst case, not even monotonic. In order to compensate for the non-deterministic transmission delay, the original sampling time of the sampling device is added to each data packet (or, in the case of audio, it is periodically transmitted separately). The audio signals are de-jittered in real time using the methods and tools of [28] to achieve minimum stable audio delay. For this purpose, a 13 ms jitter compensation buffer was used. Jitter and delay of all non-audio data are compensated after completion of the data recording. The LSL streams contain additional jitter due to the blockwise transmission of data, which can be treated in the same way as network transmission jitter.

The sampling time t_{send}, with negligible jitter, arbitrary offset and drift, is compared to the recording time on the primary timeline t_{rec}, which includes jitter due to network transmission or buffering. First, t_{send} and $t_{\text{rec}} - t_{\text{send}}$ are divided into blocks of P samples, corresponding to two seconds, based on the nominal sampling rate. In each completely filled block k, the sending times $t_{\text{send},k}$ and receiving times $t_{\text{rec},k}$ are calculated as:

$$t_{\text{send},k} = \text{median}\left\{t_{\text{send}}\right\}_P \tag{1}$$

$$t_{\text{rec},k} = \underbrace{\operatorname{argmin}\left\{t_{\text{rec}} - t_{\text{send}}\right\}_P}_{\text{transmission time}} + t_{\text{send},k} \tag{2}$$

Now a linear function $C(t) = c_{\text{drift}}t + c_{\text{offset}}$ is fitted to all pairs $(t_{\text{send},k}, t_{\text{rec},k})$. The coefficients of this function include the clock offset c_{offset} and the drift c_{drift} between the primary recording clock and the sending clock. Finally, the jitter and delay compensated recording time on the primary timeline is returned for each sample:

$$t_{\text{rec,comp}} = C(t_{\text{send}}) - \tau_{\text{network}} \tag{3}$$

Here, τ_{network} is the estimated end-to-end network delay. Events which occurred at the same $t_{\text{rec,comp}}$ took place simultaneously.

In this study, the clocks of the non-audio modalities were synchronised post-hoc. For methods of online clock synchronisation in distributed systems see [29].

F. Onset detection to generate trigger signals

In two of the four recording conditions, the acoustic signal was used as the trigger for epoching the EEG data, instead of transmitting external trigger signals. The C-weighted L_{eq} short-term sound level of the microphones as well as the attached piano in Lab 1 was measured locally in 50 ms time windows every 4 ms. These values were sent to the other labs for analysis.

To detect onsets, the level changes ΔL of two consecutive samples were measured and the criterion c was calculated:

$$c = \begin{cases} 1 & \Delta L > 10\,\text{dB/sample} \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

An onset was detected when c changed from 0 to 1, which is the first sample at which a level rise of more than 10 dB/sample was measured.

G. EEG processing and analysis

The EEG data were preprocessed using MATLAB and eeglab [30] as described in [31]. In brief, the EEG data were band-pass filtered between 1 Hz and 40 Hz, independent component analysis (ICA) was performed and components reflecting artefacts were removed.

After epoching with the trigger signals from the recording condition in each task specific dataset, a principal component analysis (PCA) was performed on the reconstructed EEG data. The sum of the first four principal components normalized by the root-mean-square (RMS) of the PCA coefficients was used for further analysis.

H. Automated speech activity labeling

The C-weighted short-term speech levels were the basis for automatic speech activity labelling, which is required for automatic analysis of communication behaviour. First, the level distribution was analysed to find a robust level threshold to indicate speech activity, independent of background noise

and microphone calibration. Raw speech activity was detected whenever short-term levels were above the activity threshold. The rest of the analysis was similar to that described by Heldner and Edlung [32]. In particular, silence intervals up to 180 ms were bridged and utterances shorter than a minimum utterance duration of 90 ms were removed.

The detection of turns and speech overlap follows the definition of Petersen et al. [33]. First, gaps of up to one second within a speaker were removed. Then, short utterances during a longer utterance of another interlocutor (typically back-channel information) were removed. Finally, the start and end times of each utterance were collected.

III. Results

A. End-to-end network delay

The end-to-end network delay τ_{network} measured with method 1 (i.e., based on the round trip times) was 12.3 ms from Lab 1 to the control room, and 11.7 ms from Lab 2 to the control room. The jitter differed slightly between the different measurements and ranged from 1 to 5 ms between Lab 1 and the control room, and was always 0.4 ms from Lab 2 to the control room. All times were measured via the central session server outside the university campus. The difference can be explained by different hardware and amount of network hubs between the laboratories and the session server.

Using method 2, that is by comparison of the system clocks, the end-to-end network delay between Lab 1 and the control room was 8.9 ms, 8.4 ms, 8.0 ms and 7.8 ms in the four different conditions. The values for the connection from Lab 2 to the control room were 9.3 ms, 9.9 ms, 10.1 ms and 10.1 ms in the four conditions. The absolute drift of the system clocks was below 2.8 μs/s in all conditions (mean drift: -0.68 μs/s, RMS: 1.95 μs/s).

The fact that the estimated values using method 2 are condition- and therefore time-dependent indicates that the clocks are not perfectly synchronized and that a clock drift between the system clocks remains despite NTP synchronisation. The absolute difference between method 1 and method 2 was up to 4.5 ms. The source of this difference is not clear, see below for a discussion.

B. Bandwidth and packet loss

The transmission bandwidth consisted of the audio bandwidth as well as the transmission bandwidth for LSL and audio level data. The upload bandwidth at Lab 1 was 1.27 MBit/s, the upload bandwidth at Lab 2 was 2.11 MBit/s, see Table I for details. In total 2 out of 175712 samples were lost in the transmission from Lab 1 to the control room, and one sample out of 175707 samples was lost from Lab 2 to the control room.

The OVBOX tries to detect packet errors and discards all incoming UDP packets with an invalid session key or format. Lost and out-of-order packets are detected by a 16-bit packet counter. Simple correction methods are applied whenever possible. The rate of lost or discarded audio packets was not explicitly recorded during the session, but based on the measured jitter information, the rate of discarded audio packets is estimated to be around 0.3% for the communication between Lab 1 and Lab 2 (jitter up to 5 ms) and much lower for all other connections (jitter below 1 ms).

TABLE I
UPLOAD BANDWIDTH OF THE OVBOX SYSTEMS IN LAB 1 AND LAB 2.

	Lab 1	**Lab 2**
audio signal 16 Bit/sample, 48 kHz	768 kBit/s 1 ch.	1.536 MBit/s 2 ch.
EEG signal 32 Bit/sample, 250 Hz	264 kBit/s 32 ch. + time	346.5 kBit/s 42 ch. + time
input levels 32 Bit/sample, 125 Hz	8 kBit/s 1 ch. + time	12 kBit/s 2 ch. + time
output levels 32 Bit/sample, 125 Hz	12 kBit/s 2 ch. + time	12 kBit/s 2 ch. + time
system clock 64 Bit/sample, 500 Hz	64 kBit/s 1 ch + time	64 kBit/s 1 ch + time
motion sensor 32 Bit/sample, 100 Hz		9.6 kBit/s 3 ch
overhead packaging, RTT measurement	158 kBit/s	130 kBit/s
total	1.27 MBit/s	2.11 MBit/s

C. Recording condition 1: Response to tone

In Figure 5 the data of the first recording condition is displayed. The participant in Lab 2 was playing a tone on a piano at irregular time intervals and the other participant was listening. In participant of Lab 2, a motor related potential can be found during the movement of the finger. Additionally, a typical auditory evoked potential [34] can be seen in both participants. The N1 (104 ms) and P2 (172 ms) of the participant in Lab 2 (who is playing the piano) is earlier than the N1 (136 ms) and P2 (216 ms) of participant in Lab 1. This time shift (N1: 32 ms, P2: 44 ms) is slightly larger than the transmission delay of the acoustic signal (24 ms) and may be explained by the older age of the participant in Lab 1 (56 versus 25 years) [35], but this is purely speculative due to the small number of participants.

D. Recording condition 2: Hand clapping

The second recording condition was a clapping task. The acoustic timing data are provided in Table II. This condition consisted of two sub-tasks: in the first task, the participant in Lab 2 was instructed to take lead during the clapping, and in the second task the participant in Lab 1 was instructed to be the leader. The tempo was 65.5 bpm in task 1 and 64.7 bpm in task 2. The lag of the follower clap relative to the leader clap was negative in task 1 and positive in task 2, indicating a potential exchange of roles despite the instruction (see also Figure 6, bottom panel). The lead clapper had a smaller clap jitter than the follower (mean difference 6.4 ms), and when Lab 2 was leading, the clap jitter was generally smaller (mean difference 5.4 ms).

In the clapping recording condition, both participants clapped simultaneously, with some interpersonal delay and intrapersonal jitter. EEG partitioning was triggered at alternative times, either at the clapping times of the nominal leader

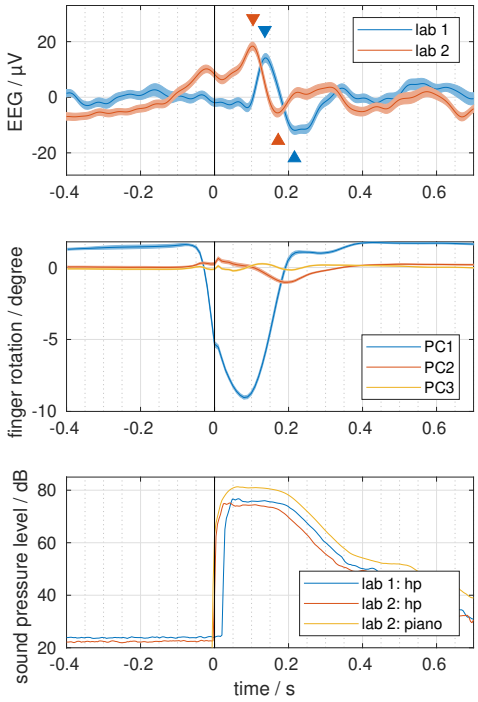

Fig. 5. EEG data (top panel), finger movement data (middle panel) and sound levels (bottom panel) of the response-to-tone recording condition. In participant of Lab 2, a motor-related potential (top panel) can be seen during the finger movement (middle panel). Additionally, a typical auditory evoked potential can be seen in both participants. The N1 and P2 of participant in Lab 1 (listener) are delayed. The delay is larger than the transmission delay of the audio signal (bottom panel).

TABLE II
CLAP TIME ANALYSIS OF RECORDING CONDITION 2.

Measure	Lab 2 has lead		Lab 1 has lead	
	lead (2)	*follow (1)*	*lead (1)*	*follow (2)*
clap period	916.1 ms		928.1 ms	
follower lag[a]	–3.8 ms		33.1 ms	
clap jitter	36.4 ms	44.6 ms	32.8 ms	37.4 ms
number of epochs	79		70	

[a]Negative values: follower is earlier.

(Figure 6, top panel) or at the clapping times of the follower (Figure 6, middle panel). In the EEG of the person to whom the trigger was synchronised, a salient N1 and P2 can be detected, independent of the role. In the first task, Lab 1 showed an N1 at 76 ms and a P2 at 168 ms, whereas Lab 2 showed an N1 at 88 ms and a P2 at 196 ms. In the second task, Lab 1 showed an N1 at 76 ms and a P2 at 164 ms, whereas Lab 2 showed an N1 at 80 ms and a P2 at 200 ms.

E. Recording condition 3: Forced turn-taking

In the forced turn-taking recording condition, the timing of the utterances was analysed. The speech levels and speech activity states are depicted in Figure 7. The average gap duration between 'call' and 'response' was 1.52 s (standard deviation: 0.96 s), and the average gap duration between 'response' and 'call' was 0.42 s (standard deviation: 0.14 s).

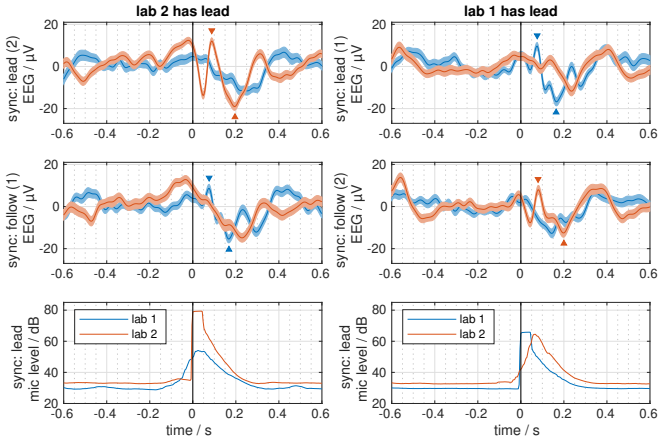

Fig. 6. Average EEG signal (top and middle panels) and average sound level (bottom panels) in the clapping recording condition. In the first task (left panels) the participant from Lab 2 was instructed to lead the clapping, whereas in the second task the participant from Lab 1 was instructed to lead. In the top panels, epoching was synchronised to the clapping times of the nominal leader, whereas in the middle panels it was synchronised to the clapping times of the follower.

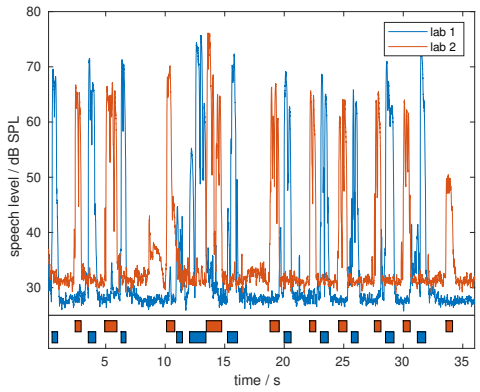

Fig. 7. Speech levels and identified voice activity regions in the forced turn taking recording condition.

F. Recording condition 3: Free conversation

The speech levels with automatic utterance and turn taking annotation is depicted in Figure 8. The total duration of the conversation was 83.5 s, during which participant 1 spoke for 27.5 s and participant 2 for 48.7 s, corresponding to 33.0% and 58.4%, respectively. Back-channel utterances were only performed by participant 1, for four times during the conversation. The median turn take times were 432 ms (interquartile range 78 ms to 656 ms), but also negative turn take times were observed.

IV. DISCUSSION

The purpose of this study was to demonstrate the feasibility of integrating EEG hyperscanning into the Internet of Sounds using the OVBOX system. Therefore, only a single data set was recorded on two subjects and no statistical analysis of the data was performed.

It could be shown that the system clocks showed a difference of up to 4.5 ms. This is probably due to the fact that

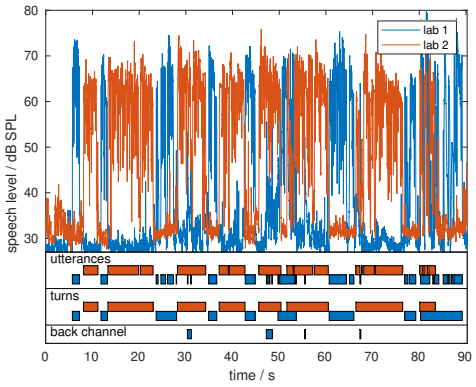

Fig. 8. Speech levels and identified voice activity regions in the free conversation recording condition are displayed. The automatic classification of utterances, turns, and back-channel utterances is shown at the bottom. It is evident that the participant from Lab 2 dominated the conversation, as indicated by their total speech activity time and the fact that only the participant from Lab 1 provided back-channel utterances.

the full NTP system was not used, but only a simplified client for consumer-grade systems. This system uses automatically detected servers for synchronisation, which can be either the nearest router or a remote server, depending on the local network setup. Using a full NTP system with identical time servers would probably result in better synchronisation of the system time. Nevertheless, the system time is sufficiently accurate to generate LTC time code for the synchronisation of video cameras, for example. A poor quality router or other hardware in Lab 1 could also explain the much higher jitter in all connections to Lab 1, which is not observed in the communication between Lab 2 and the control room.

Online jitter compensation was only performed for the audio signal. In the case of real-time control, e.g. for simulation of head movements based on remotely measured signals, as in [11], or future developments of brain-computer interfaces (BCIs) in this system, the most recent sample can be used to achieve minimal delay. However, if a re-synthesis of the sensor data is required, for example for temporal analysis, jitter compensation would also have to be performed for the sensor data.

No video transmission is integrated in the OVBOX system. The reason for this is that the system has been optimised for minimal delay, originally for music applications where delay is particularly critical. This initially limits the applications to purely acoustic communication. However, the system can also be used for audiovisual communication. One possibility is the use of virtual reality, and real-time animation of virtual avatars, controlled by the remotely recorded motion and audio signals, such as in [9], [11], [22]. Video transmission is also possible with external tools, but the video signal is potentially delayed compared to the acoustic signal.

Using telepresence with real-time animated virtual avatars, this system relates to the vision of the Musical Metaverse (MM) [36], at least in some aspects. In its original use as an NMP tool, the OVBOX fulfils some of the criteria of the MM, such as providing an interactive multi-user performance space with features such as real-time remote source movement, shared immersive spatial environments with shared background sound fields. The first live performance using the tool also contains some aspects[1]; here the live stream was sent as an interactive 360 degree video, rendered in a game engine, and the spatially separated musicians heard each other in the same spatial arrangement as presented to the audience. In the hearing research applications using the OVBOX, however, the use of virtual and mixed reality is intended to provide full control rather than a framework for MM.

The average turn-take timing, which was measured here using the OVBOX system, is 432 ms, which is within the range of the literature [32], [37]. This is an indication that the delay in the audio system is sufficiently small to enable natural communication. The presence of speech overlap and back-channel interjections are further indications that natural communication is possible with this system. This is a basic requirement for use in communication research. Conventional video conferencing systems with significantly higher delays and noise suppression make natural communication behaviour more difficult and are therefore less suitable.

With this open development, we aim to contribute to the advancement of open science. By fostering transparency and accessibility in our research processes, and by sharing our tools and methods, we aspire to enhance the reproducibility and reliability of scientific research.

V. Conclusions

In this study, it was shown that low-delay hyperscanning experiments with EEG via telepresence are possible with the OVBOX system, which is a novelty in the field of the Internet of Sounds. Specifically, it was possible to record motor and auditory evoked potentials in a task where one participant played a note on a piano and the second participant listened. Due to the low transmission delay of the acoustic signals, a joint clapping task can be performed with interpersonally synchronised EEG responses. Natural turn-taking behaviour was also observed, demonstrating the suitability of the system for communication-related experiments. Low-delay, high-bandwidth data transmission with delay and jitter compensation makes the system suitable for multimodal analysis.

Acknowledgment

Thanks to Melanie Klapprott, Lisa Straetmanns and Thorge Haupt for supporting the measurement.

References

[1] A. Marzoratti and T. M. Evans, "Measurement of interpersonal physiological synchrony in dyads: A review of timing parameters used in the literature," *Cognitive, Affective, & Behavioral Neuroscience*, vol. 22, pp. 1215–1230, May 2022.

[2] J. Decety and C. Lamm, "The role of the right temporoparietal junction in social interaction: How low-level computational processes contribute to meta-cognition," *The Neuroscientist*, vol. 13, pp. 580–593, Oct. 2007.

[3] R. Hari and M. V. Kujala, "Brain basis of human social interaction: From concepts to brain imaging," *Physiological Reviews*, vol. 89, pp. 453–479, Apr. 2009.

[1] https://www.youtube.com/watch?v=p_GSfDVmFfg

[4] R. Hari, L. Henriksson, S. Malinen, and L. Parkkonen, "Centrality of social interaction in human brain function," *Neuron*, vol. 88, pp. 181–193, Oct. 2015.

[5] L. Schoot, P. Hagoort, and K. Segaert, "What can we learn from a two-brain approach to verbal interaction?," *Neuroscience & Biobehavioral Reviews*, vol. 68, pp. 454–459, Sept. 2016.

[6] T. D. Duane and T. Behrendt, "Extrasensory electroencephalographic induction between identical twins," *Science*, vol. 150, pp. 367–367, Oct. 1965.

[7] U. Hakim, S. De Felice, P. Pinti, X. Zhang, J. Noah, Y. Ono, P. Burgess, A. Hamilton, J. Hirsch, and I. Tachtsidis, "Quantification of inter-brain coupling: A review of current methods used in haemodynamic and electrophysiological hyperscanning studies," *NeuroImage*, vol. 280, p. 120354, Oct. 2023.

[8] P. Barraza, G. Dumas, H. Liu, G. Blanco-Gomez, M. I. van den Heuvel, M. Baart, and A. Pérez, "Implementing EEG hyperscanning setups," *MethodsX*, vol. 6, pp. 428–436, 2019.

[9] A. Kothe, V. Hohmann, and G. Grimm, "Influence of natural head movements on communication behavior in virtual multi-talker conversations," in *24. Jahrestagung der Deutschen Gesellschaft für Audiologie*, German Medical Science GMS Publishing House, 2022.

[10] G. Grimm, H. Kayser, A. Kothe, and V. Hohmann, "Evaluation of behavior-controlled hearing devices in the lab using interactive turn-taking conversations," in *Proceedings of the 10th Convention of the European Acoustics Association, Forum Acusticum 2023*, European Acoustics Association, Jan. 2023.

[11] G. Grimm, A. Kothe, and V. Hohmann, "Effect of head motion animation on immersion and conversational benefit in turn-taking conversations via telepresence in audiovisual virtual environments," in *Proceedings of the 10th Convention of the European Acoustics Association Forum Acusticum 2023*, FA2023, European Acoustics Association, Jan. 2023.

[12] L. Turchet, M. Lagrange, C. Rottondi, G. Fazekas, N. Peters, J. Østergaard, F. Font, T. Bäckström, and C. Fischione, "The internet of sounds: Convergent trends, insights, and future directions," *IEEE Internet of Things Journal*, vol. 10, pp. 11264–11292, July 2023.

[13] U. Lindenberger, S.-C. Li, W. Gruber, and V. Müller, "Brains swinging in concert: cortical phase synchronization while playing guitar," *BMC Neuroscience*, vol. 10, Mar. 2009.

[14] A. Zamm, S. Debener, A. R. Bauer, M. G. Bleichner, A. P. Demos, and C. Palmer, "Amplitude envelope correlations measure synchronous cortical oscillations in performing musicians," *Annals of the New York Academy of Sciences*, vol. 1423, pp. 251–263, May 2018.

[15] A. Zamm, C. Palmer, A.-K. R. Bauer, M. G. Bleichner, A. P. Demos, and S. Debener, "Behavioral and neural dynamics of interpersonal synchrony between performing musicians: A wireless eeg hyperscanning study," *Frontiers in Human Neuroscience*, vol. 15, Sept. 2021.

[16] G. Grimm, "Interactive low delay music and speech communication via network connections (OVBOX)," *Acta Acoustica*, vol. 8, pp. 1–7, Apr. 2024.

[17] G. Grimm, "ovbox - ORLANDOviols consort box." https://ovbox.de/, 2024.

[25] C. Kothe, S. Y. Shirazi, T. Stenner, D. Medine, C. Boulay, M. I. Grivich, T. Mullen, A. Delorme, and S. Makeig, "The Lab Streaming Layer for synchronized multimodal recording," Feb. 2024.

[18] C. Chafe and S. Oshiro, "Jacktrip on raspberry pi," in *LAC-Linux Audio Conference*, 2019.

[19] A. Carôt, "Soundjack." https://www.soundjack.eu/, accessed 2023.

[20] C. Drioli, C. Allocchio, and N. Buso, *Networked Performances and Natural Interaction via LOLA: Low Latency High Quality A/V Streaming System*, pp. 240–250. Springer Berlin Heidelberg, 2013.

[21] L. Turchet and C. Fischione, "Elk audio os: An open source operating system for the internet of musical things," *ACM Transactions on Internet of Things*, vol. 2, pp. 1–18, Mar. 2021.

[22] M. Hartwig, V. Hohmann, and G. Grimm, "Speaking with avatars – influence of social interaction on movement behavior in interactive hearing experiments," pp. 94–98, 2021.

[23] A. Zhdanov, J. Nurminen, P. Baess, L. Hirvenkari, V. Jousmäki, J. P. Mäkelä, A. Mandel, L. Meronen, R. Hari, and L. Parkkonen, "An Internet-based real-time audiovisual link for dual MEG recordings," *PLOS ONE*, vol. 10, p. e0128485, June 2015.

[24] C. Rottondi, M. Buccoli, and M. Zanoni, "Feature-based analysis of the effects of packet delay on networked musical interactions," *Journal of the Audio Engineering Society*, vol. 63, pp. 864–875, dec 2015.

[26] M. Wright, "Open sound control: an enabling technology for musical networking," *Organised Sound*, vol. 10, pp. 193–200, Nov. 2005.

[27] G. Grimm, J. Luberadzka, and V. Hohmann, "A toolbox for rendering virtual acoustic environments in the context of audiology," *Acta Acustica united with Acustica*, vol. 105, pp. 566–578, May 2019.

[28] F. Adriaensen, "Controlling adaptive resampling," in *Linux audio conference, Stanford, USA*, 2012.

[29] Y.-C. Wu, Q. Chaudhari, and E. Serpedin, "Clock synchronization of wireless sensor networks," *IEEE Signal Processing Magazine*, vol. 28, pp. 124–138, Jan. 2011.

[30] A. Delorme and S. Makeig, "EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis," *Journal of Neuroscience Methods*, vol. 134, pp. 9–21, Mar. 2004.

[31] M. Rosenkranz, B. Holtze, M. Jaeger, and S. Debener, "EEG-based intersubject correlations reflect selective attention in a competing speaker scenario," *Frontiers in Neuroscience*, vol. 15, June 2021.

[32] M. Heldner and J. Edlund, "Pauses, gaps and overlaps in conversations," *Journal of Phonetics*, vol. 38, pp. 555–568, Oct. 2010.

[33] E. B. Petersen, E. Walravens, and A. K. Pedersen, "Backchannel behavior in conversations and how it is affected by hearingloss, noise, and hearing aids," in *Forum Acusticum*, (Turin), 2023.

[34] T. W. Picton, S. A. Hillyard, H. I. Krausz, and R. Galambos, "Human auditory evoked potentials. i: Evaluation of components," *Electroencephalography and Clinical Neurophysiology*, vol. 36, pp. 179–190, Jan. 1974.

[35] A. Kok, "Age-related changes in involuntary and voluntary attention as reflected in components of the event-related potential (erp)," *Biological Psychology*, vol. 54, pp. 107–143, Oct. 2000.

[36] L. Turchet, "Musical metaverse: vision, opportunities, and challenges," *Personal and Ubiquitous Computing*, vol. 27, pp. 1811–1827, Jan. 2023.

[37] S. C. Levinson and F. Torreira, "Timing in turn-taking and its implications for processing models of language," *Frontiers in Psychology*, vol. 6, June 2015.