Preliminary Analysis of an Immersive Low-Latency Audio System for Remote Therapeutic Intervention

Valeria Bruschi
Dept. of Information Engineering
Università Politecnica delle Marche
Ancona, Italy
v.bruschi@staff.univpm.it

Michael Fioretti
Dept. of Information Engineering
Università Politecnica delle Marche
Ancona, Italy
m.fioretti@pm.univpm.it

Giuseppe Bergamino
Dept. of Information Engineering
Università Politecnica delle Marche
Ancona, Italy
g.bergamino@pm.univpm.it

Alessandro Terenzi
Dept. of Information Engineering
Università Politecnica delle Marche
Ancona, Italy
a.terenzi@staff.univpm.it

Grazia Iadarola
Dept. of Information Engineering
Università Politecnica delle Marche
Ancona, Italy
g.iadarola@staff.univpm.it

Leonardo Gabrielli
Dept. of Information Engineering
Università Politecnica delle Marche
Ancona, Italy
l.gabrielli@staff.univpm.it

Susanna Spinsante
Dept. of Information Engineering
Università Politecnica delle Marche
Ancona, Italy
s.spinsante@staff.univpm.it

Stefania Cecchi
Dept. of Information Engineering
Università Politecnica delle Marche
Ancona, Italy
s.cecchi@staff.univpm.it

Stefano Squartini
Dept. of Information Engineering
Università Politecnica delle Marche
Ancona, Italy
s.squartini@staff.univpm.it

***Abstract*—Facilitating consistent social involvement is an important factor in enhancing the quality of life for elderly people. As mobility challenges and transportation gaps leave many older individuals confined to their homes, there's a growing need to create opportunities for remote assistance and social connection within the comfort of their own homes. The paper addresses such a scenario by tackling some of the challenges related to an immersive low-latency audio system designed to connect elderly users to social operators, for therapeutic interventions, such as rehabilitative exercises or music-assisted therapies. The system uses Ambisonic technology to deliver spatial audio to enhance realism and deepen user immersion and low latency communication to allow for a real-time interaction. To assess its impact, the study analyses physiological responses, specifically skin conductance, as a potential indicator of user engagement. Furthermore, subjective listening tests are conducted to study how the operator is impaired by the network in evaluating the person's motor skills and engagement. This test allows to determine acceptable latency thresholds, beyond which some music therapy interventions become unfeasible.**

***Index Terms*—low-latency audio system, immersive audio system, social interaction**

This research has received funding from the project Vitality – Project Code ECS00000041, CUP I33C22001330007 - funded under the National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.5 - 'Creation and strengthening of innovation ecosystems,' construction of 'territorial leaders in R&D' – Innovation Ecosystems - Project 'Innovation, digitalization and sustainability for the diffused economy in Central Italy – VITALITY' Call for tender No. 3277 of 30/12/2021, and Concession Decree No. 0001057.23-06-2022 of Italian Ministry of University funded by the European Union – NextGenerationEU.

I. Introduction

Over the past decade, tele-medicine has evolved beyond the classical remit of clinical-data management—monitoring, diagnostics and follow-up—to encompass services that address the patient holistically. In particular, remote music therapy can extend the benefits of music interventions to subjects who are home-bound or geographically isolated, maintaining therapeutic continuity while eliminating travel burdens [1]. By harnessing rhythm, melody, harmony and timbre, structured interventions that target physical, emotional, social and cognitive outcomes can be designed successfully. Empirical evidence substantiates its efficacy: Hole et al. [2] report significant reductions in peri-operative and hospital anxiety; Lin et al. [3] demonstrate concurrent gains in motor and cognitive performance among neurological and psychiatric cohorts; and Raglio et al. [4], together with de Witte et al. [5], describe marked decreases in depressive symptomatology.

Delivering sessions online, however, imposes a set of technical constraints. Limited audio fidelity and the loss of spatial cues may reduce immersion, while network latency has emerged as the principal obstacle to synchronous, sensorimotor activities [6]. Recent surveys consistently rank latency at the top of therapists' concerns: Cole et al. [7] observed a sharp decline in the use of rhythmic auditory stimulation once therapy moved online, attributing the drop to delay and poor sound quality; Kantorová et al. [8] likewise identify lag-induced rhythm mismatches as the predominant technological drawback; and Agres et al. [9] highlight how regional dis-

parities in network infrastructure exacerbate the problem. A recent scoping review of online choir initiatives for elderly clients further identifies transmission delay as a recurring technical barrier, suggesting that even activities thought to be relatively tolerant of timing error, such as group singing, are susceptible to latency-induced disengagement [10]. Although previous researches describe latency as a barrier, they seldom specify quantitative limits, leaving a critical evidence gap, which must be addressed.

To further improve the experience, immersive reproduction can enhance therapeutic interventions and foster more meaningful connections between elderly individuals and social operators by creating emotionally engaging environments. Immersive audio technology simulates spatial soundscapes using virtual sound sources, aiming to replicate how humans naturally localize audio. This localization depends on head-related transfer functions (HRTFs), which describe how sound travels around the head and torso to reach the ears [11], [12]. Two principal methods support 3D audio design: multichannel loudspeaker setups and binaural synthesis using headphones [13]. While headphones allow precise channel separation, prolonged use may be uncomfortable. Alternatively, stereo loudspeakers may introduce crosstalk, i.e., the interference between audio channels, which can be corrected with crosstalk cancellation (CTC) algorithms, including HRTF inversion [14] and adaptive techniques like least mean square (LMS) [15]. Another widespread technique is the recursive ambiophonic crosstalk elimination (RACE) algorithm, which eliminates crosstalk without prior HRTF knowledge [16]. The use of multichannel systems, which involve multiple loudspeakers arranged around the listener, can create a more immersive auditory experience. In this context, ambisonics technology builds on loudspeaker arrays to reproduce full-sphere acoustic fields [17], enhancing spatial accuracy. The emotional impact of music on listeners is widely recognized and supported by research [18]–[21]. Emotional arousal can be objectively measured via electrodermal activity (EDA), which tracks fluctuations in skin conductance (SC) due to eccrine sweat gland responses triggered by the sympathetic nervous system [22], [23]. EDA captures both slow background levels, named skin conductance level (SCL), and rapid emotional spikes, named skin conductance response (SCR) [24], [25]. Acoustic stimuli such as music or speech affect both autonomic and cognitive pathways, influencing emotional states [26], [27]. EDA is usually measured through exosomatic techniques, with electrodes located on the body skin, making it a feasible acquisition modality by wearable emotion-sensing devices [28], [29].

This paper presents advancements on an on-going research project aiming at delivering an immersive low-latency audio system to elderly individuals and social care professionals for therapeutic intervention. The system, first described in [30] is composed of a typical Networked Music Performance (NMP) low-latency audio infrastructure [31], [32], but leverages Ambisonic technology, to enhance immersion. At a broader level, this work aligns with the Internet of Sounds research agenda, which frames networked, intelligent audio systems and their applications across domains—including health and well-being [33].

In this work we provide further details on the 3D audio processing and we tackle two research questions. First, we monitor physiological signals from the subjects, specifically skin conductance, as a marker of emotional and cognitive engagement, in order to gain insights on the listening experience. Secondly, we examine how transmission latency affects the operator's ability to accurately assess the patient's motor–rhythmic skills. A series of controlled experiments involving multiple participants was conducted to examine both engagement responses and acceptable latency thresholds. Differently from previous works that deal with round-trip latency in the context of a musical interaction, here we address the task of allowing a therapist to play a backing track to the patient and assess her/his ability to clap in time in presence of network delay and jitter.

The paper is organized as follows. Section II presents the system by analyzing its two core components. Section III details the hardware setup used for the physiological validation and discusses the results. Section IV investigates the latency impact, describing the experiments and the results. Section V discusses the results, and Section VI concludes the paper.

II. System Description

This section describes the proposed architecture for remote therapy applications. It can be declined to match different space and budget constraints. First of all, the immersive system can be scaled from a fully fledged multichannel system to binaural, depending on the patient location, whether the sessions take place in a specialized clinic or at home. Furthermore, the audio networking can be installed to work in the same facility using a wired LAN (this is the case of patients which cannot leave their room or hospital ward) or using the Internet to connect distant places. These two cases are called intra-facility and extra-facility, in the following. More details about the audio and network connections follow.

A. Immersive Audio System

The immersive audio system employed in this work is shown in Figure 1. The input stereo signal is converted into a $M = 5$ channel through an upmixer following the approach of [34]. The upmixer scheme is reported in Figure 2. The central signal y_{C} and the rear surround signal y_{RS} are computed as

$$y_{\mathrm{C}} = (x_{\mathrm{L}} + x_{\mathrm{R}})/\sqrt{2}, \tag{1}$$

$$y_{\mathrm{RS}} = (x_{\mathrm{L}} - x_{\mathrm{R}})/\sqrt{2}, \tag{2}$$

where x_{L} and x_{R} define the left and the right channel of the input stereo signal, respectively. A band-pass filter (BPF) with cut-off frequencies of $100\,\mathrm{Hz}$ and $10\,\mathrm{kHz}$ is applied to the central signal. The rear surround signal y_{RS} is delayed by $\Delta_{\mathrm{RS}} = 15$ ms and filtered by a low-pass filter (LPF) with cut-off frequency of $15\,\mathrm{kHz}$. The rear signals, y_4 and y_5, are derived by introducing a 180-degree phase shift,

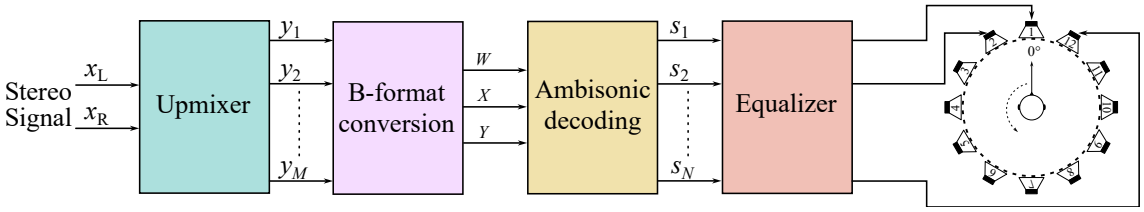

Fig. 1. Scheme of the immersive audio system.

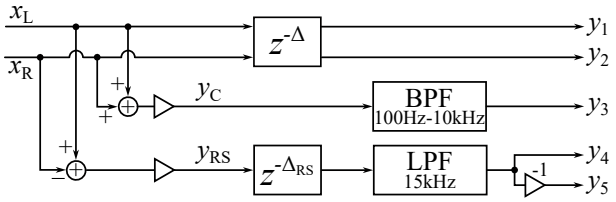

Fig. 2. Scheme of the upmixer.

achieved by inverting one of the two channels. The delay Δ (in samples), applied to the stereo signal, is used to synchronize the outputs. Both filters (BPF and LPF) are finite impulse response (FIR) types, each comprising 128 coefficients. The upmixer directs each output signal toward a specific azimuth angle by applying first-order ambisonic (FOA) encoding. The frontal stereo signals, y_1 and y_2, are oriented at $\theta_1 = 45°$ and $\theta_2 = 315°$, respectively. The center signal, y_3, is directed to $\theta_3 = 0°$. The rear signals, y_4 and y_5, are positioned at $\theta_4 = 135°$ and $\theta_5 = 225°$, respectively. Each signal y_m, with $m = 1, \ldots, M$, is first converted to B-format following the approach in [17], and subsequently combined to compute the WXY components as:

$$\begin{aligned} W &= \sum_{m=1}^{M} G y_m, \\ X &= \sum_{m=1}^{M} y_m \cos(\theta_m), \\ Y &= \sum_{m=1}^{M} y_m \sin(\theta_m), \end{aligned} \tag{3}$$

where $G = 1/\sqrt{2}$. The output signals s_n, where $n = 1, \ldots, N$ and $N = 12$ (representing the number of loudspeakers), are generated using the ambisonic decoder as follows:

$$\begin{aligned} s_n = G_1 W + G_2 X \cos\left[\frac{2\pi}{N}(n-1)\right] \\ + G_2 Y \sin\left[\frac{2\pi}{N}(n-1)\right], \end{aligned} \tag{4}$$

where $G_1 = 1/\sqrt{2}$ and $G_2 = 1/2$. To address the perceptual limitations associated with lower-order ambisonics [35], each signal is equalized before reproduction to emphasize mid-to-high frequency content, thereby enhancing clarity, spatial spaciousness, and timbral detail. The equalization stage comprises a cascaded configuration of two second-order filters: a shelving high-pass filter with a cutoff frequency of 3 kHz and a gain of 3 dB, followed by a parametric band-pass filter centered at 2.4 kHz, with a bandwidth of 4.8 kHz and a gain of 3 dB.

Although the system described above is optimized for multichannel ambisonic reproduction, its architecture is also adaptable for binaural playback, making it a practical solution in space-limited environments or budget-constrained scenarios, where installing multiple loudspeakers is not feasible.

B. High-bandwidth Low-latency Audio System

The ambisonic playback chain described above can be deployed in two distinct remote interaction scenarios. In the intra-facility case, therapist and client occupy separate rooms within the same building and are linked by a dedicated wired LAN, where end-to-end delay can be kept well below the 15 ms range. In the extra-facility case, the endpoints are geographically distant and connect over the public Internet. Latency is inevitably higher and few homes can host a full loudspeaker array. To preserve immersion under these constraints, the ambisonic stream can be down-mixed—either to a reduced loudspeaker layout or to a binaural feed—so spatial cues remain intact while hardware demands stay modest. To support real-time audio in both settings (local or remote) we screened several transport stacks against four criteria: (i) one-way latency, (ii) channel scalability, (iii) ease of start-up and usability, and (iv) the possibility of adding a synchronous video feed. We ultimately selected two candidates:

- Audinate's Dante® Virtual Soundcard with proprietary hardware. Dante yielded one-way latencies as low as 6 ms in a local wired configuration providing high channel scalability and robust session recovery. This solution is best suited for lab-based or clinical environments requiring robust multichannel audio.
- JackTrip™, an open source alternative used with standard USB sound cards. JackTrip provided comparable performance under optimal network conditions with fewer integration features. The system is more appropriate for home-based setups especially when paired with binaural rendering.

In an Internet-based link, latency can largely impair the outcomes of the session, therefore it is of uttermost importance to evaluate how this affects clinical practice. While experimental works, such as [36], have thoroughly explored the impact of latency in NMP, there is no specific quantitative evidence related to music therapy session. Section IV attempts to

quantify the delay limits that remain acceptable for music-therapy applications.

III. Physiological Experiments

A. Hardware setup

The experiments were conducted in a semi-anechoic chamber, where sound was reproduced through twelve Genelec 6010A loudspeakers uniformly distributed along the main circumference of a spherical structure at a height of $1.5\,\mathrm{m}$ from the ground. These loudspeakers were connected to a MOTU 24 I/O sound card, which interfaced with a PC running the NU-Tech software [37] to control the audio stimuli. The audio tracks were processed and played back at a sampling rate of $48\,\mathrm{kHz}$. The listener stood at the center of the sphere in an upright position, equipped with an Empatica Embrace-Plus smartwatch [38]. It is an advanced multi-sensory wrist-wearable device, capable of monitoring various physiological signals, including SC, measured in microSiemens (μS), in the range 0.01 μS to 100 μS. Based on the device manual provided by the manufacturer, the SC sensor operates with a sampling frequency of 4 Hz. The SC sensor available on the device silicone bracelet exploits two 316L stainless steel electrodes to detect small changes in electrical conductance at the surface of the ventral wrist skin.

A total of 10 participants (5 males and 5 females), aged between 23 and 45 years (mean=28.8, SD=6.6), took part in the experiments. The test started with a three-minute baseline period, during which no audio stimulus was presented and the participant's skin conductance (SC) levels were monitored. Following this initial phase, the listening session proceeded with the alternating reproduction of the original stereo track, played through the two frontal loudspeakers (loudspeakers 2 and 12 in Figure 1), and the ambisonic version as follows:

- 30 seconds of stereo track (reference);
- 30 seconds of ambisonic track;
- 30 seconds of stereo track (reference);
- 30 seconds of ambisonic track.

The same playback sequence was repeated across five musical genres, pop, R&B, rock, gospel, and classical, presented in a randomized order. The tracks selected for the study are detailed in Table I. Following the playback phase, Empatica recorded a one-minute baseline measurement. The entire experimental session lasted fourteen minutes. While the time length chosen for the audio stimuli ($30\,\mathrm{s}$) may appear short, in SC research, audio stimuli have a length from few ms up to several seconds. When the target of interest is capturing the stress-related or emotional reaction of a subject, up to 60 seconds-long stimuli are used, with typical length of 5 to 6 s [39].

B. SC signal processing

Within the skin conductance signal, the number of SCR peaks is targeted as an indicator of the subject's reactions to stimuli [40]. It is possible to also consider the frequency of SCR peaks, intended as the number of detected peaks over a given time interval. In this study, as the different experimental phases included in the test protocol have controlled time duration, the number of SCR peaks counted is considered. To be related to a specific stimulus, an SCR peak shall be typically identified not later than 3 or 4 s from it. Following the removal of the SCL baseline component from each acquired SC signal, by a proper sliding median filter, a classical Trough-To-Peak (TTP) [41] method can be applied to extract the number of SCR peaks. The TTP algorithm processes the SCR signal component and identifies peaks as the events in which the signal amplitude goes from a local minimum to a local maximum, within a fixed time window (4 s in this study). A minimum amplitude variation criterion of $0.01\,\mu\mathrm{S}$ is applied to identify a peak [42]. Fig. 3 shows an example of the SC signals acquired during each session that has a total duration of $840\,\mathrm{s}$. The vertical magenta lines mark the separation among the five different musical genres reproduction. Within each genre reproduction interval, two 30-seconds long stereo and two 30-seconds long ambisonic tracks alternate. In Fig. 4, the results of the TTP algorithm on the same signal represented in Fig. 3 are shown: the SCR component is separated from the SCL one, and SCR peaks are counted against a threshold of $0.01\,\mu\mathrm{S}$. For a better visualization of the SCR signal component and the corresponding peaks detected by the TTP algorithm, Fig. 5 provides a zoomed view on a signal portion.

TABLE I
List of the soundtracks used in the experiment

Genre	Track	Artist
R&B	I.G.Y. (What a Beautiful World)	Donald Fagen
Pop	You're The One That I Want	Lo-Fang
Rock	Rock 'N Roll Train	AC/DC
Gospel	These Bones	The Fairfield Four
Classical	The Nutcracker Op 71 Act 1	Tchaikovsky

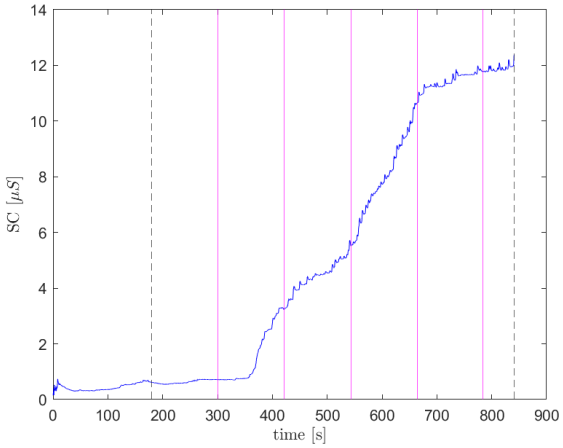

Fig. 3. SC signal collected from one of the subject participating to the test, in which the vertical magenta lines mark the separation among the five different musical genres reproduction. Grey vertical dotted lines mark the end of the initial 3-minutes long baseline phase (left), and the end of the final 1-minute long baseline phase (right).

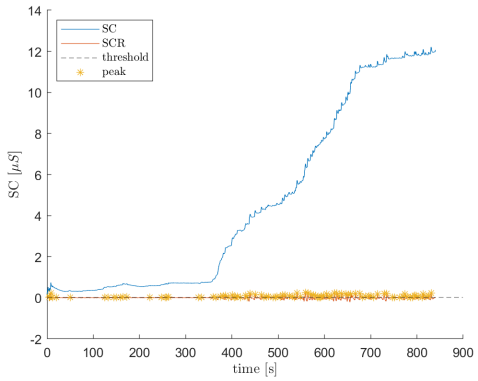

Fig. 4. The results of the TTP algorithm on the same signal shown in Fig. 3: the SCR component is separated from the SCL one, and SCR peaks are counted against a threshold of 0.01 μS.

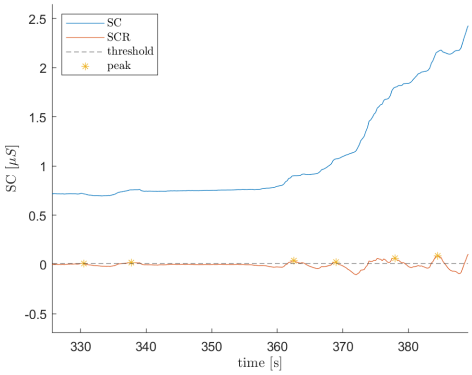

Fig. 5. A zoomed detail of the SCR signal from Fig. 4.

C. Experimental Results

Following acquisition and segmentation over the test phases of the SC signals, for each of the 10 subjects involved in the experiments, processing in the time domain was performed as described in III-B. A variation in the number of SCR peaks among the experiment phases in which music stimuli have been administered, and also with respect to those found during the initial baseline phase associated to the absence of stimulation and relaxation of the subject, can be taken as an indicator of stress or excitement. The number of SCR peaks counted during the final baseline phase reflects the longer-lasting effects of the experiments on each subject.

Table II presents the number of SCR peaks detected for each subject across the various test phases. The majority of participants (7 out of 10) exhibits a higher number of peaks during the initial baseline than the final one. While the two phases have a different time duration, it is anyway reasonable to assume a kind of anticipatory stress in the participants, which is not reflected at the end of the experiment, potentially indicating that the latter does not induce stress in the involved subjects. In particular, subject 10 features a quite evident emotional involvement at the beginning of the experiment (with an average of more than 10 SCR peaks per minute during the initial baseline), and a marked reduction following the playback test (1 SCR peak during the 1-minute long final baseline). Excluding the baseline phases, the maximum number of peaks recorded during the reproduction of tracks for each subject is highlighted in bold within the table. Analyzing the results, no consistent pattern emerges, and it is not possible to define a common effect across subjects. Notably, subjects 8 and 9 exhibited zero peaks throughout nearly all phases, including the baselines. At the end of every experiment, participants were asked to describe their impressions and emotional responses. All subjects perceived the transition between the stereo and ambisonic tracks, and consistently described that ambisonic format delivered a more immersive experience. It was perceived as more realistic, with sound arriving from different directions. In contrast, the stereo track was described as frontally confined, lacking spatial depth. The difference was particularly noticeable with certain music genres: several participants identified gospel and pop music as the most engaging, especially when experienced in ambisonic format.

IV. Latency Experiments

Qualitative feedback presented in the previous section shows that all participants noticed the switch from stereo to ambisonics and described it as markedly more immersive and engaging. However, this perceptual gain, could rapidly degrade under adverse network conditions. The following experiments are designed to find what latency values make interactive exercises unfeasible. For the remainder of the paper latency is meant as round-trip latency.

Let us consider the scenario of a hand clapping task (or other tempo-keeping tasks, using any percussion instrument) in a remote session. The operator is required to have a basic musical training, to be capable of judging the patient performance and elicit specific musical or speech interaction based on it. However, the network delay impairs the operator in doing so, even when the two subjects are not interacting together, as in the simpler case where the operator is playing back an audio track and the patient must follow it. Figure 6 depicts this scenario. As can be seen, the audio track is received by the patient with some delay. The patient then tries to clap in time with some cues (e.g. quarter beats), but some of the claps are correctly in time (R), some are wrong (W). These audio signals are sent back to the operator who may receive them not aligned with the audio track. Therefore judgment on the patient's tempo keeping is impaired to the point that wrong claps may be perceived as correctly timed and viceversa.

Please note that we are not assuming a constant network delay, but we are considering the more general case of a jittery delay, which poses further challenges to the operator in understanding whether the patient is keeping a steady tempo and whether the claps are at the right time. The operator does not know what latency separates them and whether the patient decided to clap on the quarter beats or rather on other music intervals that are equally coherent with the rhythm track.

TABLE II

SCR PEAKS COUNT AT EACH TEST PHASE, FOR EACH SUBJECT. THE LABELS IN THE FORM Sx AND Ax (WHERE $x = 1, 2$) IDENTIFY THE FIRST OR SECOND STEREO (S) OR AMBISONIC (A) TRACK, FOR EACH GENRE PRESENTED TO EACH SUBJECT. FOR EACH SUBJECT, THE HIGHEST NUMBER OF DETECTED PEAKS DURING TRACK PLAYBACK (EXCLUDING BASELINES) IS HIGHLIGHTED.

Test phase		**Sbj1**	**Sbj2**	**Sbj3**	**Sbj4**	**Sbj5**	**Sbj6**	**Sbj7**	**Sbj8**	**Sbj9**	**Sbj10**
Initial baseline		13	10	13	15	0	38	3	0	0	32
R&B	S1	0	2	2	3	0	6	2	0	0	2
	A1	1	**3**	1	6	0	6	1	0	0	2
	S2	4	**3**	3	5	0	**7**	1	0	0	**4**
	A2	0	2	5	3	1	4	1	0	0	3
Pop	S1	0	1	6	5	2	2	0	0	0	3
	A1	1	0	1	6	2	4	0	0	0	2
	S2	4	1	**8**	5	1	5	0	0	0	3
	A2	4	2	2	**9**	2	4	0	0	0	2
Rock	S1	3	0	5	2	**3**	5	1	0	0	0
	A1	4	0	1	3	0	3	1	0	0	0
	S2	4	0	2	3	1	4	1	0	0	0
	A2	4	0	4	4	0	4	0	0	**1**	2
Gospel	S1	3	2	3	4	2	6	0	0	0	2
	A1	**6**	1	3	5	0	5	0	0	0	3
	S2	4	2	2	5	1	3	0	0	0	3
	A2	4	**3**	4	4	1	5	0	1	0	1
Classical	S1	4	0	1	4	2	5	2	0	0	1
	A1	3	0	4	3	**3**	4	2	0	0	1
	S2	3	1	1	4	1	3	0	**2**	**1**	2
	A2	4	0	0	4	1	4	1	1	0	0
Final baseline		9	6	6	9	2	10	5	4	0	1

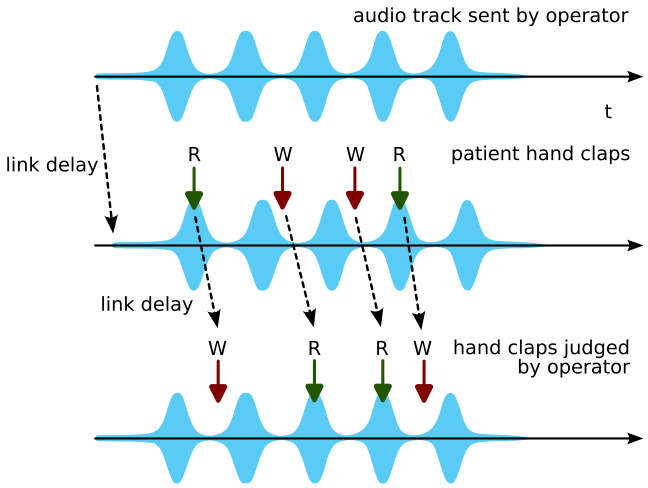

Fig. 6. An example of the backing track clapping task affected by link delay. The patient can clap in time with the beats (R) or at the wrong time (W), however the round-trip link delay may shift the claps inducing the operator to classify correct ones as wrong and viceversa.

A. Experiment Design

To understand how network delay affects the operator judgment of the client's performance, we designed a simulated experiment that mimics a remote music therapy scenario. Importantly, no actual patients or therapist participated, and no clinical intervention was performed. Instead we created a synthetic dataset of hand clapping performances with controlled tempo-keeping ability and simulated network latencies. The experiments consisted of listening tests where subjects must judge synthetically generated tracks, simulating a remote session with patients having different tempo-keeping capacity and with varying network latency.

Some of the criteria are taken directly from the established Individualized Music Therapy Assessment Profile (IMTAP) framework [43], ensuring continuity with real-world music-therapy practice. IMTAP organises objectives into domains (Motor, Social–Emotional, Cognitive, Musicality) and subsections with graded skills (Level 1 to Level 4). Within the Musicality–Tempo subsection we selected three items that map naturally to remote assessment:

- **Tolerates changing tempo (code MUS–B.i)**. This item probes whether the client can keep playing or moving when the beat speeds up or slows down. We implemented three backing tracks at 55, 75 and 85 beats per minute (BPM), representing slow, moderate and fast tempi that older participants can possibly follow without undue physical or cognitive strain [44]. Furthermore, varying the tempo allows us to test how a fixed network latency occupies a specific fraction of the inter-beat interval, amplifying its perceptual impact.
- **Conscious body movement in tempo (code MUS–B.iv).** The client is asked to align simple limb movements with the beat. Because only gross motor timing is required, in our test the performance is simulated by a hand-clapping audio sample, providing a controlled motor-rhythmic cue for the therapist to evaluate.
- **Plays in tempo of therapist within 1–4 measures (code MUS–B.v)**. The original item asks the client to match the therapist's tempo for up to 4 measures. In our implementation, the therapist plays a backing track that the patient must follow.

To remove confounding variables and avoid involving patients and operators in this preliminary analysis, we decided

to perform experiments based on a synthetic dataset of hand-clapping tempo keeping with controllable patient skills and network delay. The dataset is employed to conduct subjective listening tests, where users are asked to rate the tempo keeping task of different audio file. Specifically, the performance is simulated by placing a hand-clap sample on every quarter note, yielding 16 claps aligned with the backing track (4/4 meter). This item provides the most demanding synchronization skill and therefore the most sensitive probe of latency effects. Because the study did not involve real participants in a therapeutic context, no ethics approval was required under our institution's guidelines. The results should be interpreted as a technical feasibility and perceptual sensitivity study, not a clinical trial.

B. Dataset creation

To simulate realistic variability in human rhythmic performance, we developed a controlled dataset of hand clapping tracks exhibiting four graded levels of temporal perturbation. The goal was to model increasingly inaccurate rhythmic behavior that could plausibly occur in remote music therapy scenarios, and then assess how such deviations impact the therapist's perception of beat alignment.

The generation of the perturbed hand clapping tracks was based on a "perfect" reference version, constructed as a sequence of clap samples temporally aligned with metronomic beat positions of pre-recorded music tracks at several fixed tempos (55, 75, 85 BPM). Following the IMTAP grading system, we introduced structured temporal disturbances to this idealized sequence to create four distinct performance levels, labeled from Level 4 (best) to Level 1 (worst). These transformations were implemented programmatically using parameterized time-domain manipulations informed by perceptual and motor-control literature [45], though to our knowledge, no directly analogous method has been reported in the existing body of work.

The following time-domain modifications were employed:

- **Jitter**: Gaussian-distributed timing variability was applied independently to each beat, simulating natural inconsistency in motor timing. Two levels of jitter were defined: a low standard deviation ($15\,\mathrm{ms}$) used in Level 4, and a higher standard deviation ($50\,\mathrm{ms}$) used in Levels 1–3.
- **Offset**: Levels 1 through 3 incorporated a fixed offset of $220\,\mathrm{ms}$ to simulate the delayed auditory-motor responses commonly observed in elderly or neurologically impaired individuals. This value was chosen to model a transition from anticipatory to reaction-based timing strategies, consistently late claps irrespective of intention. Although we used this approach to capture extreme synchronization challenges, it likely compressed perceptual differences between the three degraded levels. We acknowledge this limitation and suggest that future studies disentangle offset from other perturbations to isolate their contribution to listener judgments.
- **Drift**: A gradual shift in beat timing was introduced via a linearly increasing temporal stretch. For Levels 1–2, a moderate drift rate (3.5%) was applied, modeling scenarios in which the participant subtly falls out of phase with the music over time.
- **Missing beats**: In Level 1 only, a proportion of beat events were randomly omitted at a fixed probability (33%). This transformation was included to simulate sporadic disengagement or intentional interpretive variation by the performer. However, upon auditory inspection, this modification proved to be perceptually ambiguous: some missing beats were interpreted not as performance degradation but as expressive choices or syncopation. For this reason, the perceptual impact of this manipulation is expected to be more nuanced and may vary significantly across listeners.

Each transformation was scaled according to the underlying tempo to preserve a musically equivalent effect across tracks. For instance, jitter and offset values were normalized relative to a reference tempo of 60 BPM and rescaled proportionally for faster or slower tracks. Drift was adapted using a hybrid scaling rule that preserved its perceptual slope while maintaining consistency across BPM conditions. Table III summarizes which variations were applied to each simulated performance.

TABLE III
SUMMARY OF ALTERATION TECHNIQUES APPLIED TO OBTAIN 4 DIFFERENT DEGRADED LEVEL OF CLAPPING PERFORMANCE

Rank	Jitter (ms)	Offset (ms)	Drift (%)	Miss (%)
Level 4	±15	X	X	X
Level 3	±50	+220	X	X
Level 2	±50	+220	3.5	X
Level 1	±50	+220	3.5	33

C. Listening Tests

To investigate how simulated latency impairs the listener's ability to assess rhythmic performance accuracy, we designed a perceptual evaluation experiment using an adapted version of a Multiple Stimuli with Hidden Reference and Anchor (MUSHRA)-like interface. We employed the WebMUSHRA framework for its ease of deployment and randomized presentation capabilities [46], though we note that our protocol does not strictly conform to the ITU-R BS.1534 standard, as we are not evaluating audio quality, but rather perceptual performance accuracy in a beat-tracking task.

Each participant was presented with 12 randomized trials, corresponding to 3 base tracks (each at a different tempo: 55, 75, and 85 BPM) combined with 4 levels of simulated latency ($6\,\mathrm{ms}$, $15\,\mathrm{ms}$, $50\,\mathrm{ms}$, and $200\,\mathrm{ms}$). The latency values were chosen with reference to the clapping study by Chafe et al. [36], which maps distinct synchronization behaviors as delay increase.

Within each trial, the participant was asked to evaluate five clapping performances: one reference performance with

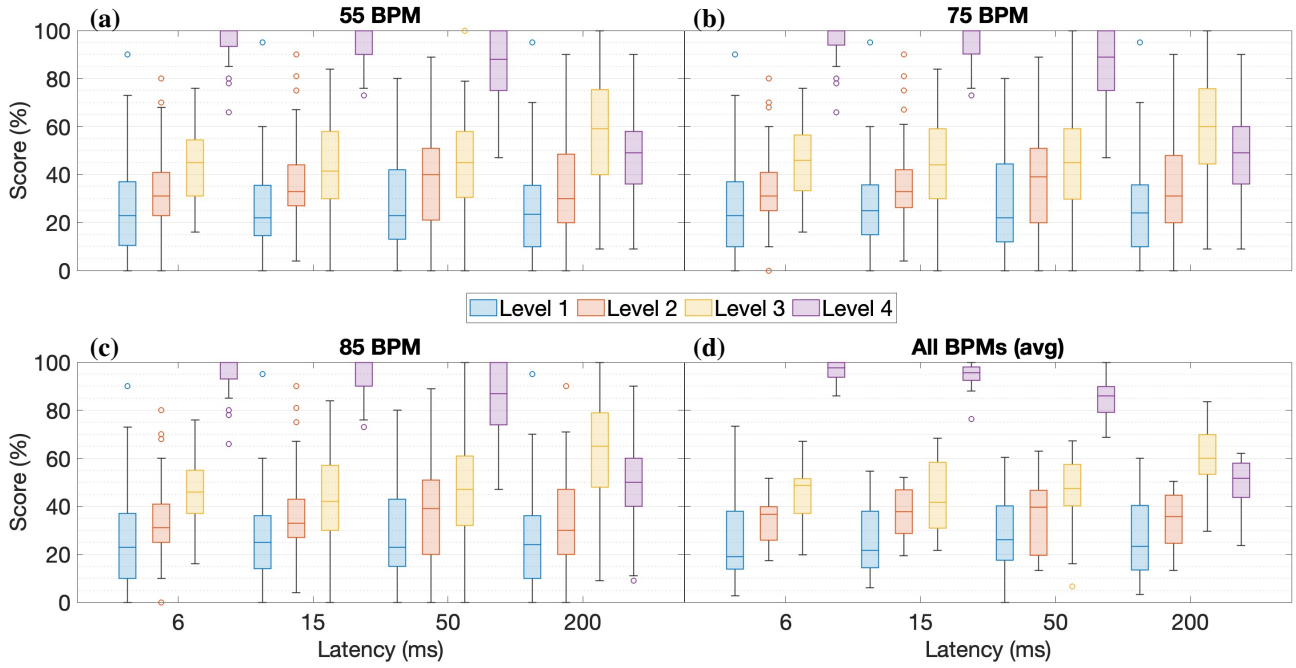

Fig. 7. Boxplots showing score distributions across four simulated performance levels and four latency conditions (6, 15, 50, 200 ms). Panel (a) reports results for the 55 BPM track, (b) for 75 BPM, and (c) for 85 BPM. Panel (d) shows average scores per participant for each latency and performance level, averaged across all tempi.

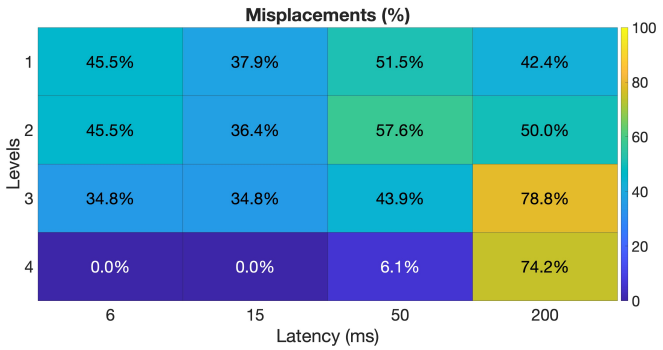

Fig. 8. Heatmap showing misidentification rates across latency values and performance levels, averaged over all tempi. As latency exceeds 50 ms, near-perfect performances (Level 4) are increasingly mistaken for degraded ones. At 200 ms latency, participants failed to correctly identify the near-perfect performance in up to 74% of trials.

perfect clapping, always presented with 0 ms latency and four degraded clapping performances (Levels 1–4), each subjected to the current trial's simulated latency. All stimuli were rendered as audio mixes of a backing track combined with hand-clap overlays. Participants were instructed to rate the rhythmic accuracy of each performance on a continuous scale from 0 (inaccurate) to 100 (perfectly accurate). No guidance was given on how to interpret specific numerical values, in order to reflect subjective and spontaneous perceptual judgments. To minimize learning effects and bias, both the order of pages (tempo and latency combinations) and the positions of stimuli within each page were fully randomized per participant.

D. Results

A total of 22 volunteers were recruited among university students and staff; participants were on average 30 years old (SD ± 7), and no formal musical or music-therapy training was required for inclusion. The test was conducted in a quiet environment with headphones at a comfortable level. All data were collected anonymously. Three participants were excluded because they rated the reference stimulus with a score of 90 or below in at least 25% of trials; all subsequent analyses therefore refer to the remaining 19 participants.

The distribution of scores for all trials is shown in Figure 7. Within the lower-latencies range and across all tempi, the median scores respect the intended levels hierarchy, confirming that listeners could discriminate the four rhythmic proficiencies. As latency increases, the median score for the most accurate performance (Level 4) decreases significantly. The distribution also gets larger suggesting greater uncertainty in evaluating the same performance when a large enough latency is introduced. When the simulated latency reaches 200 ms the Level 4 performance shows a median result below that of the Level 3 performance for 75 and 85 BPM tracks indicating that the participants were unable to discriminate between good and degraded performances. To further investigate this effect we considered the score distribution for a given BPM and simulated latency for each participant and mapped it to a positional rating system where each performance would be given a score of 1 to 4 depending on its relative position to

the other performances in the distribution. Figure 8 shows how many times (percentage) a given track was given a positional score which doesn't reflect its level of performance. We found that when the simulated latency reaches $200\,\mathrm{ms}$, the near perfect clapping performance was given the highest score only in about 1/4 of the trials.

V. Discussion

The experiments conducted in this study provides new evidences that are worth discussing, although in some respects further work is necessary.

Our latency-based experiments show that the ability to correctly judge a tempo-keeping task degrades sensibly with increasing latency, as expected. With a latency up to $50\,\mathrm{ms}$ the error is limited to 15% for all BPMs, while with a $200\,\mathrm{ms}$ latency the performance was frequently misjudged receiving a lower median score and often being confused with degraded performances. Taken together, our findings indicate that music interventions based on rhythmic exercises remain reliable as long as latency does not exceed $50\,\mathrm{ms}$. Beyond this threshold the operator's assessment capability rapidly degrades, with a pronounced drop at tempi above 75 BPM. Although we expect the score to degrade approximately linearly between $50\,\mathrm{ms}$ and $200\,\mathrm{ms}$, further experiments are needed to gather further data in that latency range. However, in practical terms, we can conclude that remote sessions with drumming or hand-clapping should be conducted only when a sub-50 ms network link is available; when not guaranteed, shifting to slower tempi or to different tasks offers a viable, more latency-tolerant alternative. To support this adaptive approach, remote-therapy platforms could report delay in musically meaningful terms rather than raw milliseconds—for example, by expressing the current latency as a tempo-dependent tolerance and expected rhythmic offset (e.g., "reliable up to 80 BPM; eighth-note misalignment at 120 BPM")—so that practitioners can immediately judge feasibility and adjust the exercise.

On the other hand, physiological experiments did not give definitive results, mainly due to the consistent changes across participants in skin conductance, particularly in the contrast between the initial and final baseline phases. Most subjects exhibited higher SCR peak activity before the listening session than at the end, suggesting that the auditory experience did not provoke increased physiological arousal. However, two participants recorded no SCR peaks throughout, possibly indicating either limited responsiveness or a consistently calm state. Despite the absence of strong trends in peak distribution during the playback phases, subjective feedback showed consistent responses: all listeners described the ambisonic reproduction as more immersive and spatially realistic, while the stereo version was viewed as frontally constrained, lacking the enveloping sensation. This perceptual distinction was particularly pronounced in certain genres, namely gospel and pop, which were often reported as the most emotionally engaging in the ambisonic reproduction.

VI. Conclusions

This paper described the advancements of a remote therapy system based on an audio networking infrastructure and immersive sound playback algorithms. Details are given for the audio processing architecture and experimental protocols are designed to address two current challenges. One is the search for an objective method to assess the status of the subject using SC signals while the other one is the need to quantify what latency values allow certain (non strictly interactive) tasks, such as tempo-keeping or following.

Concerning the first objective, an experimental setup was described, which shows how subjects react to changes in the listening experience (silence, stereo and ambisonic music reproduction). In line with previous study, each subject reacts differently, therefore there is no simple guideline to draw, however, some subjects react to the listening experience in a predictable way, which opens up for interpretation of their state, with future evidences coming in the field.

As to the second objective, we identified a robust trend in the ability of the operator to correctly judge the tempo-keeping exercise of the patient. Up to $50\,\mathrm{ms}$ there seems to be no real impairment in the operator's ability. This means that the feedback that can be given to the patient is valid, and can help build a good communication between the two. We also found that a patient with a jitter of $\pm\ 50\,\mathrm{ms}$ and an offset of $200\,\mathrm{ms}$ is given a moderately low score by most of our subject, (median value: 50%) when latency does not impair significantly the experience ($6\,\mathrm{ms}$).

Although significant for planning the next step of the project, which will consist of a real-world experience with a cohort of patients, the experiments require further work to do.

The four identified network latencies were taken from the literature as a starting point. However, more experiments need to be conducted in the range between 50 and $200\,\mathrm{ms}$ to see how the rating degrades and with more granularity between IMTAP Level 3 and 4. This will help operators to choose the right intervention depending on the network latency, or inversely, it will inform the development of software tools that enable a correct identification of the patient's level independently from the network latency.

Physiological tests need to further expand by increasing the size of the listening panel and including patients in their clinical environment. Furthermore, headphones should be included too, in order to evaluate their efficacy and the responses they elicit.

References

[1] C. Vinciguerra and A. Federico, "Neurological music therapy during the COVID-19 outbreak: updates and future challenges," *Neurological Sciences*, vol. 43, pp. 3473–3478, Jun. 2022.

[2] J. Hole, M. Hirsch, E. Ball, and C. Meads, "Music as an aid for post-operative recovery in adults: a systematic review and meta-analysiss," *The Lancet*, vol. 386, pp. 1659–1671, Oct. 2015.

[3] S.-T. Lin, P. Yang, C.-Y. Lai, Y.-Y. Su, Y.-C. Yeh, M.-F. Huang, and C.-C. Chen, "Mental health implications of music: Insight from neuroscientific and clinical studies," *Harvard review of psychiatry*, vol. 19, no. 1, pp. 34–46, 2011.

[4] A. Raglio, L. Attardo, G. Gontero, S. Rollino, E. Groppo, and E. Granieri, "Effects of music and music therapy on mood in neurological patients," *World journal of psychiatry*, vol. 5, no. 1, p. 68, 2015.

[5] M. de Witte, S. Aalbers, A. Vink, S. Friederichs, A. Knapen, T. Pelgrim, A. Lampit, F. A. Baker, and S. van Hooren, "Music therapy for the treatment of anxiety: a systematic review with multilevel meta-analyses," *EClinicalMedicine*, vol. 84, 2025.

[6] A. Clements-Cortés, M. Pranjić, D. Knott, M. Mercadal-Brotons, A. Fuller, L. Kelly, I. Selvarajah, and R. Vaudreuil, "International music therapists' perceptions and experiences in telehealth music therapy provision," *International Journal of Environmental Research and Public Health*, vol. 20, p. 5580, Aug. 2023.

[7] L. P. Cole, T. L. Henechowicz, K. Kang, M. Pranjić, N. M. Richard, G. L. Tian, and C. Hurt-Thaut, "Neurologic music therapy via telehealth: A survey of clinician experiences, trends, and recommendations during the covid-19 pandemic," *Frontiers in Neuroscience*, vol. 15, p. 648489, 2021.

[8] L. Kantorova, J. Kantor, B. Hořejší, A. Gilboa, Z. Svobodova, M. Lipskỳ, J. Marečková, and M. Klugar, "Adaptation of music therapists' practice to the outset of the covid-19 pandemic—going virtual: A scoping review," *International Journal of Environmental Research and Public Health*, vol. 18, no. 10, p. 5138, 2021.

[9] K. R. Agres, K. Foubert, and S. Sridhar, "Music therapy during covid-19: Changes to the practice, use of technology, and what to carry forward in the future," *Frontiers in Psychology*, vol. 12, p. 647790, 2021.

[10] B. Dowson and J. Schneider, "Online singing groups for people with dementia: scoping review," *Public Health*, vol. 194, pp. 196–201, 2021.

[11] S. Bharitkar and C. Kyriakakis, *Immersive Audio Signal Processing*. Springer, 2006.

[12] K. Iida, *Head-related transfer function and acoustic virtual reality*. Springer, 2019.

[13] W. G. Gardner, *3-D Audio using Loudspeakers*. Kluwer Academic Publishers, 1998.

[14] B. B. Bauer, "Stereophonic Earphones and Binaural Loudspeakers," *J. Audio Eng. Soc.*, vol. 9, no. 2, pp. 148–151, Apr. 1961.

[15] L. Lim and C. Kyriakakis, "Multirate Adaptive Filtering for Immersive Audio," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 5, Salt Lake City, UT, USA, May 2001, pp. 3357–3360.

[16] R. Glasgal, "360° localization via 4.x RACE processing," in *Proc. of 123rd Audio Engineering Society Convention*, New York, USA, Oct. 2007.

[17] R. K. Furness, "Ambisonics-an overview," in *Audio Engineering Society Conference: 8th International Conference: The Sound of Audio*. Audio Engineering Society, 1990.

[18] A. Poli, S. Cecchi, S. Spinsante, A. Terenzi, and F. Bettarelli, "A preliminary study on the correlation between subjective sound quality perception and physiological parameters," in *Audio Engineering Society Convention 150*. Audio Engineering Society, 2021.

[19] Z. Baracskai and S. Finn, "Relaxation effects of binaural phenomena," in *Audio Engineering Society Conference: 52nd International Conference: Sound Field Control-Engineering and Perception*. Audio Engineering Society, 2013.

[20] A. Greco, G. Valenza, L. Citi, and E. P. Scilingo, "Arousal and valence recognition of affective sounds based on electrodermal activity," *IEEE Sensors Journal*, vol. 17, no. 3, pp. 716–725, 2016.

[21] G. Iadarola, A. Poli, and S. Spinsante, "Analysis of galvanic skin response to acoustic stimuli by wearable devices," in *2021 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*. IEEE, 2021, pp. 1–6.

[22] Healey, J.A. and Picard, R.W., "Detecting stress during real-world driving tasks using physiological sensors," *IEEE Transactions on Intelligent Transportation Systems*, vol. 6, no. 2, pp. 156–166, 2005.

[23] W. Boucsein, *Parameters of Phasic Electrodermal Activity*. Springer, 2012, pp. 151–158.

[24] Y. Can, N. Chalabianloo, D. Ekiz, and C. Ersoy, "Continuous stress detection using wearable sensors in real life: Algorithmic programming contest case study," *Sensors*, vol. 19, 2019.

[25] D. Williams, C.-Y. Wu, V. Hodge, D. Murphy, and P. Cowling, "A psychometric evaluation of emotional responses to horror music," in *Audio Engineering Society Convention 146*. Audio Engineering Society, 2019.

[26] P. N. Juslin and J. Sloboda, *Handbook of music and emotion: Theory, research, applications*. Oxford University Press, New York, 2011.

[27] M. L. Chanda and D. J. Levitin, "The neurochemistry of music," *Trends in cognitive sciences*, vol. 17, no. 4, pp. 179–193, 2013.

[28] A. J. Hudspeth, "How the ear's works work," *Nature*, vol. 341, no. 6241, pp. 397–404, 1989.

[29] J. M. Appler and L. V. Goodrich, "Connecting the ear to the brain: molecular mechanisms of auditory circuit assembly," *Progress in neurobiology*, vol. 93, no. 4, pp. 488–508, 2011.

[30] V. Bruschi, A. Terenzi, N. A. Dourou, L. Gabrielli, M. Fioretti, G. Bergamino, S. Cecchi, and S. Squartini, "An immersive low-latency audio system for social interaction with elderly people," in *2025 Immersive and 3D Audio: from Architecture to Automotive (I3DA) - accepted for publication*. IEEE, 2025.

[31] L. Gabrielli and S. Squartini, "Wireless networked music performance," in *Wireless Networked Music Performance*. Springer, 2015, pp. 53–92.

[32] C. Rottondi, C. Chafe, C. Allocchio, and A. Sarti, "An overview on networked music performance technologies," *IEEE Access*, vol. 4, pp. 8823–8843, Dec. 2016.

[33] 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. 11 264–11 292, Jul. 2023.

[34] M. R. Bai and G.-Y. Shih, "Upmixing and downmixing two-channel stereo audio for consumer electronics," *IEEE Transactions on Consumer Electronics*, vol. 53, no. 3, pp. 1011–1019, 2007.

[35] I. Engel, C. Henry, S. V. Amengual Garí, P. W. Robinson, and L. Picinali, "Perceptual implications of different ambisonics-based methods for binaural reverberation," *The Journal of the Acoustical Society of America*, vol. 149, no. 2, pp. 895–910, 2021.

[36] C. Chafe, J.-P. Caceres, and M. Gurevich, "Effect of temporal separation on synchronization in rhythmic performance," *Perception*, vol. 39, no. 7, pp. 982–992, 2010.

[37] A. Lattanzi, F. Bettarelli, and S. Cecchi, "Nu-tech: The entry tool of the hartes toolchain for algorithms design," in *Proc. 124th Audio Engineering Society Convention*, 2008, pp. 1–8.

[38] https://www.empatica.com/en-eu/embraceplus/.

[39] E. Gatti, E. Calzolari, E. Maggioni, and M. Obrist, "Emotional ratings and skin conductance response to visual, auditory and haptic stimuli," *Scientific data*, vol. 5, p. 180120, June 2018. [Online]. Available: https://europepmc.org/articles/PMC6018518

[40] S. Lui and D. Grunberg, "Using skin conductance to evaluate the effect of music silence to relieve and intensify arousal," in *2017 International Conference on Orange Technologies (ICOT)*, 2017, pp. 91–94.

[41] M. Kuhn, A. M. V. Gerlicher, and T. B. Lonsdorf, "Navigating the manyverse of skin conductance response quantification approaches – A direct comparison of trough-to-peak, baseline correction, and model-based approaches in Ledalab and PsPM," *Psychophysiology*, vol. 59, no. 9, p. e14058, 2022, e14058 PsyP-2021-0638.R1.

[42] G. Iadarola, A. Poli, and S. Spinsante, "Reconstruction of Galvanic Skin Response Peaks via Sparse Representation," in *2021 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*, 2021, pp. 1–6.

[43] H. T. Baxter, J. A. Berghofer, L. MacEwan, J. Nelson, K. Peters, and P. Roberts, *The individualized music therapy assessment profile: IMTAP*. Jessica Kingsley Publishers, 2007.

[44] J. D. McAuley, M. R. Jones, S. Holub, H. M. Johnston, and N. S. Miller, "The time of our lives: life span development of timing and event tracking." *Journal of Experimental Psychology: General*, vol. 135, no. 3, p. 348, 2006.

[45] J. Cannon, A. Cardinaux, L. Bungert, C. Li, and P. Sinha, "Reduced precision of motor and perceptual rhythmic timing in autistic adults," *Frontiers in Human Neuroscience*, vol. 10, Jul. 2024.

[46] M. Schoeffler, S. Bartoschek, F.-R. Stöter, M. Roess, S. Westphal, B. Edler, and J. Herre, "webmushra—a comprehensive framework for web-based listening tests," *Journal of Open Research Software*, vol. 6, no. 1, 2018.