Journal of the Audio Engineering Society

# Spatial Audio for Audioconferencing in Mobile Devices: Investigating the Importance of Virtual Mobility and Private Communication and Optimizations.

--Manuscript Draft--

| | |
|---|---|
| **Manuscript Number:** | JAES-D-15-00079R2 |
| **Article Type:** | Research Paper |
| **Manuscript Classifications:** | 70: Signal processing; 80: Transmission and interconnection; 140: Perception and human factors |
| **Keywords:** | spatial audio; audioconferencing; mobile devices |
| **Corresponding Author:** | Emanuel Aguilera, M.D.<br>Universitat Politècnica de València<br>Valencia, Valencia SPAIN |
| **First Author:** | Emanuel Aguilera, M.D. |
| **Order of Authors:** | Emanuel Aguilera, M.D. |
| | Jose J. Lopez, Professor |
| | Jeremy R. Cooperstock, Professor |
| **Abstract:** | Audioconferencing systems are becoming increasingly sophisticated, seeking to improve immersion, intelligibility and sense of presence. In parallel, mobile devices are gaining traction for such applications, often supplanting desktops as the platform of choice, especially when the communication does not require video. This article describes our design of a mobile multiparty audioconference application and our research into the influence of spatial audio and interactivity on user experience with the application. In particular, we consider implementation tradeoffs, and investigate whether the full potential of spatial audio is realized simply by distributing the virtual locations of the participants according to some predetermined configuration. In addition, we analyze the utility of "whisper" mode functionality, in which a subset of participants can engage in an ad-hoc sidebar conversation privately from the remaining participants. Our results provide interesting guidelines of relevance to the development future audioconferencing systems. |
| **Full Title:** | Spatial Audio for Audioconferencing in Mobile Devices: Investigating the Importance of Virtual Mobility and Private Communication and Optimizations. |
| **Additional Information:** | |
| **Question** | **Response** |
| Has this article previously been published at a conference or convention? | No |
| Please enter the text of an appropriate cover letter to accompany your submission, briefly describing the article you want to submit and saying why you believe it is suitable for publication in the AES Journal. | Please enter your cover letter text here. |
| **Author Comments:** | |
| **Response to Reviewers:** | |

# Spatial Audio for Audioconferencing in Mobile Devices: Investigating the Importance of Virtual Mobility and Private Communication and Optimizations.

**EMANUEL AGUILERA**[1], *AES Member*, **JOSE J. LOPEZ**[1], *AES Senior Member*, **JEREMY R. COOPERSTOCK**[2]
(emagmar@iteam.upv.es)

[1]*Institute of Telecommunications and Multimedia Applic., Technical University of Valencia, Valencia, 46022, Spain*
[2] *Centre for Intelligent Machines, McGill University, Montreal, QC H3A 0E7, Canada*

Audioconferencing systems are becoming increasingly sophisticated, seeking to improve immersion, intelligibility and sense of presence. In parallel, mobile devices are gaining traction for such applications, often supplanting desktops as the platform of choice, especially when the communication does not require video. This article describes our design of a mobile multiparty audioconference application and our research into the influence of spatial audio and interactivity on user experience with the application. In particular, we consider implementation tradeoffs, and investigate whether the full potential of spatial audio is realized simply by distributing the virtual locations of the participants according to some predetermined configuration. In addition, we analyze the utility of "whisper" mode functionality, in which a subset of participants can engage in an ad-hoc sidebar conversation privately from the remaining participants. Our results provide interesting guidelines of relevance to the development future audioconferencing systems.

## 0 INTRODUCTION

With advances in hardware and telecommunication networks, audioconferencing systems continue to improve in terms of the user experience they offer. These systems have been widely adopted by business but are also used regularly by the general public. Setups range from dedicated installations and audioconference rooms to office desktops, and increasingly, mobile devices, which allow users to employ the technology outside of their offices and homes.

Various opportunities exist to make multi-party conference calls more realistic, interactive, and immersive. These include spatial audio, interactive manipulation of avatar positions, and "whisper mode".

With respect to the first, audio delivery through headphones can be used to render spatial audio [1, 2], allowing the human auditory system to locate, separate and understand multiple individuals even when they are speaking simultaneously (the "cocktail party effect"). Provided that the avatars of the participants are simulated to be spatially distributed, speech intelligibility is significantly improved. This motivates efforts to design and implement low-latency and low computational cost spatial audio-conferencing that runs on mobile devices with a limited power budget, as we describe in Section 2.

However, it is unclear whether the full potential of spatial audio is derived simply by distributing the sound sources around each user according to some predetermined configuration. In this regard, interactive, individual, dynamic control over avatar positioning may be an important factor in the user experience.

Finally, support for "whisper mode" functionality, in which a subset of participants can engage in an ad-hoc sidebar conversation privately from the remaining participants, may also enrich audioconference communication. Such capability is consistent with the affordances of real-world communication but is rarely provided in commercial audioconferencing systems.

We posit that spatial audio, interactive control, and "whisper mode" are useful in certain conversational contexts, e.g., negotiation and tasks requiring selective attention, but not all. In this article, we evaluated our hypothesis through a series of formal studies involving a total of 40 participants. The studies, described in Section 3, involved three separate experiments in different conversational contexts that cover the different interaction conditions we wished to evaluate.

The main contribution of this article is our exploration of the impact of the various factors on user experience of an audioconferencing system, outlined above. The analysis of our experimental results, presented in Section 4, confirm our hypothesis.

## 1 LITERATURE REVIEW

The importance of spatial arrangement in videoconferencing scenarios was described by Sellen and Buxton [3]. Their research demonstrated how spatial cues support natural interaction dynamics among multiple participants. In an informal study, Baldis found that when given the opportunity to adjust the position of four sound sources, users chose angles of approximately 20º and ± 60º around them [4].

Notably, more than half of the cues Sellen and Buxton describe relate to audio affordances (selectively listen to different, parallel conversations; make side comments to other participants; and hold parallel conversations). The system they designed to support such natural interaction in videoconferencing, Hydra, was based on physical avatars for each participant, consisting of a small camera, video display, and loudspeaker, all integrated in a single package. As another early solution, Billinghurst et al. describe a wearable conferencing space in which avatars of the remote participants appear around the user, who can turn to talk to specific individuals or even move about the virtual space [5]. Similarly, leveraging physical mobility, Kan et al. used GPS data of the conference participants to determine the spatial audio rendering between them [6].

Spatial audio is generally preferred by subjects over non-spatial monaural audio [1, 7]. It is known to improve speaker discrimination [4, 7] and intelligibility of conversations, leveraging the ability of the human auditory system to attend selectively to sounds from a specific direction, i.e., the "cocktail party effect". Under conditions in which avatar positions were fixed, Baldis found that supplying static visual representations of the participants, as avatars, was not helpful for comprehension or memory [4]. In contrast, other work demonstrated that memory of the details of the conversation improved when users could manipulate the avatar representations in the virtual world [1, 7]. More recently, Inkpen et al. found that the addition of video display of the individual participants also led to improved user experience, with subjects rating the conference to be of higher quality, demonstrating greater engagement, and finding it easier to keep track of who was speaking, especially in conjunction with monaural audio [8].

Achieving a desired spatial arrangement by physical movement of dedicated monitors, or walking around one's environment, as in the early systems of Sellen and Buxton, Billinghurst et al., and Kan et al., are unsuitable for typical conferencing applications. More practically, control over the user's position can be accomplished by manipulation of avatars, representing the conference participants, within a virtual world. Such virtual navigation has long been a staple of online virtual worlds such as Second Life. Spatial audioconferencing systems with graphical user interfaces that allow for movement of the user's own avatar to control the spatial rendering include software from Mingleverse[1] and Voxeet.[2] Another possibility is to allow for movement of the avatars of the other participants, such as used in the Conferencing3 system of Goose et al. [1].

A related aspect of real world interaction that is poorly supported in audio- or videoconferencing systems is the ability to speak privately with one or more participants in the context of a larger multi-party discussion, sometimes referred to as "whisper mode". More generally, Cohen describes the need for a taxonomy of inclusion/exclusion operators for audio sources and sinks in groupware, suggesting the terminology of "deafen" to disable a sink, and "confide" to concentrate on a particular sink by disabling others [9]. Fernando et al. describe a graphical representation for such attributes using colour and symbols [10] adjacent to the avatars, but the usability of their approach was not evaluated. Design issues for features related to selective attention or private conversations include selecting an easy-to-understand mechanism to specify the individual(s) with whom one wishes to enter into a private communication and an appropriate form of feedback to indicate to the participants when they are in this mode. For multi-party telephony interaction, this can be challenging,[3] but with the addition of a graphical user interface, selection of the "whisper" participants and indication of the state may be more straightforward.

## 2 EXPERIMENTAL PLATFORM

We developed a complete multi-party audioconferencing system that was used for this study. The system uses a client-server architecture in which the mobile device clients (smartphones and tablets) are represented in a virtual room, and communicate with each other via a server, as shown in Figure 1.

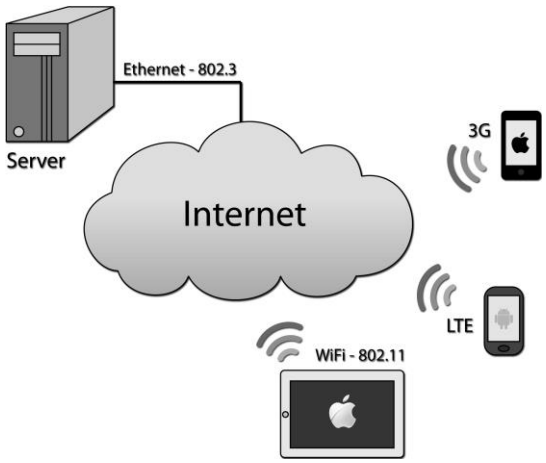



Fig. 1. Client-server architecture.

[1] The Mingleverse system, although now defunct, provided virtual rooms within which participants could position themselves for a spatial audioconferencing experience.

[2] http://www.voxeet.com/features.html

[3] See for example https://downloads.avaya.com/elmodocs2/avayaip/users/definity/def4624/6k_whsp.htm

As noted in Section 1, there exist several commercial audioconferencing applications that employ spatial audio, but this capability is rarely found in mobile systems. As we developed our own application for this purpose, several considerations had to be taken into account in our choice of spatial audio algorithm. These include the constraints of transmission bitrate, which are in turn dependent on the network and mobile standard used. Additionally, computational limitations, and minimization of energy requirements to extend battery life, are critical factors for mobile devices.

In the remainder of this section, we explain the preprocessing applied to the voice signal, the binaural techniques for rendering spatial audio, and our design decisions for the communications system and its user interface.

### 2.1. Audio processing for transmission

In order to achieve optimal audio quality and reduce bandwidth requirements for the transmission of the voice streams, we apply noise reduction, voice activity detection (VAD) and voice compression to the audio signal captured by the microphone. However, the typically expensive processing of echo cancellation is unnecessary since our system is used in conjunction with a stereo headset.

Noise reduction increases intelligibility when several speech signals are mixed, but also improves the subsequent VAD and voice codec performance. The voice activity detector works by calculating the probability of speech in each frame, and applying hysteresis thresholding. By discarding frames not containing speech, no effort is wasted on compression and transmission, nor on decompression and spatialization at the receivers, thereby reducing both CPU usage and bandwidth consumption. These benefits become more important as the number of participants, increases, since the speech probability per person decreases when more people are involved.

Compression of the voice signal reduces the bandwidth needs while maintaining maximum voice quality. For this purpose, we use the open source and low-latency Speex voice codec [11]. The configurable parameters of the codec are adjusted to achieve a trade-off between voice quality and computational cost on a smartphone. For example, we choose a wideband 16 kHz sampling rate rather than the narrowband 8 kHz rate since the modest increase in computational cost is warranted by the improvement in voice quality. The audio signal is divided into 20 ms frames of 16-bit samples (640 bytes/frame), which are reduced to 70 bytes/frame after Speex compression.

In order to improve the perceptual quality of the communications we employed the Packet Loss Concealment (PLC) tool provided by the Speex library. This technique allows interpolating voice packets that have been lost during transmission by means of voice model-based methods, with the purpose of minimizing gaps in the communication.

### 2.2. Spatialization

Traditional stereo rendering can place a sound at a point between the left and right loudspeakers, using a simple panning algorithm that applies a different gain to each channel. However this basic spatial sensation is unrealistic and fails to provide a high degree of immersion. As summarized by Rumsey [12], listeners prefer this simple mix when asked about sound quality but when source positioning is important, as in our case, binaural sound processing improves localization, justifying its use.

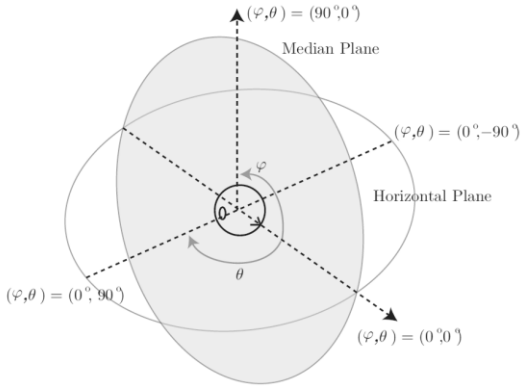



Figure 2: Binaural coordinate system.

Binaural sound spatialization tries to simulate the hearing function by using the same mechanisms employed by the human auditory system. Headphone reproduction helps to recreate sound pressure in the inner ear similar to that of an equivalent real situation. This makes it possible to discern the direction of sound sources based on the relative differences in the sound received by the two ears.

The most important cues for source localization in azimuth are interaural time difference (ITD) and interaural level difference (ILD) as proposed by Rayleigh in what is known as the “duplex theory” [13]. These effects and others are considered by the head-related transfer function (HRTF), which characterizes how sounds are perceived by the ear. The HRTF is a function of direction, distance and frequency.

One problem that arises in binaural spatialization using HRTF is related to the anthropomorphic differences between individuals. Differences in head size, interaural distance and pinna shape lead to different HRTF. A non-individualized HRTF reduces the accuracy in source localization and leads to front/back confusion. Moreover, elevation cues are directly related to the pinna shape and are difficult to reproduce. However, the source positioning requirements in our experiment and software are limited. We only need to place sources in the horizontal plane, in front of the listener With these restrictions, a universal HRTF can provide acceptable results, with the advantage that the software can be used by any person without individual adaptation.

Each speaker in the conference is treated as a monaural point source. Using the touchscreen interface of the mobile device, the user can place these sources at

arbitrary positions in the horizontal plane of a virtual room where all speakers participate. To synthesize these point sources at the desired spatial locations, we employ filtering with the head-related impulse response (HRIR), the time-domain version of the HRTF, to generate a binaural audio representation of the scene. This is done by convolving the audio signal in the time domain for both left and right ears, as shown in Equation (1).

$$L[n] = HRIR_L(\varphi,\theta,n) * s(n)$$
$$R[n] = HRIR_R(\varphi,\theta,n) * s(n) \quad (1)$$

where L[n] is the signal for the left ear, R[n] for the right ear, s(n) is the speaker's voice and the HRIR functions are specific to the arrival direction (φ,θ) of the speaker's voice (at each ear), as seen in Figure 2.

In practice, HRIR functions have a length between 128 and 512 samples. Convolving each monaural signal with the two HRIR vectors implies a significant computational cost. As a result, performing these operations in the frequency domain as FFT multiplication (using overlap-add or overlap-save algorithms) with the HRTF is often preferred for efficiency reasons.

Although we verified during development that current smartphones have sufficient power to carry out these filtering algorithms in real time, doing so requires significant power consumption, thereby reducing battery life. Moreover, these operations risk monopolizing computational resources that are needed for other tasks, and increasing latency in the communications.

As a more power-efficient alternative, we simplified the computation to provide a "good enough" synthesis. First, we considered the fact that the complex frequency pattern presented by the HRTF at frequencies above 5 kHz, due to the effects of the pinna, are only important for localization of elevated sources. Assuming that our sound sources are all located on the horizontal plane, these effects can be ignored. For practical purposes, this reduces the ILD to modeling the diffraction effect of the head. This effect is much less complex in frequency and can be modeled by low order IIR filters. Although head size and shape vary between individuals, the effect of these differences is less important than that of the pinna, allowing for the implementation of an (almost) listener-independent spatial audio system.

The implementation of the HRTF was split into two parts. First, the ILD was implemented by means of two IIR filters. Second, the ITD was achieved by adding a time delay between the left and right ear signals. This model was already used successfully by other researchers [14].

The design of the ILD IIR filters followed a similar procedure to that described in previous work [15], where the authors obtained a standard HRTF model by averaging the values from a database of real HRTF responses. Using the averaged response, a sixth order parametric IIR filter was adjusted for each azimuth, with the parameters linked to direction through a simple polynomial approximation.

The ITD delay between the two ears is also dependent on direction, and can be expressed by Equation (2):

$$ITD(s) = \frac{c}{r} * (\varphi + \sin\varphi) \quad (2)$$

where φ is the azimuth angle of arrival as shown in Figure 3, c is the speed of sound and r is the head radius.

The ITD delay, which is usually fractional, can be modeled efficiently using FIR or IIR filters. For our model, an eighth order IIR filter provides sufficient accuracy and an almost flat frequency response. However, for our final implementation, this filter was replaced by the use of a considerably simpler integer delay and fade-in/fade-out mixes between time windows when sources are in movement. Through subjective testing, we determined that the difference between fractional and integer delays is almost imperceptible for most people and the spatial sensation is not degraded. The details of the testing procedure are outside the scope of this article.

Finally, a simple distance attenuation algorithm was applied to recreate an approximate perception of distance.

### 2.3 Communications

As it is not feasible to modify telephony protocols, we designed proprietary transmission protocols over Internet Protocol (IP) for our audioconference call service. One protocol is used for session management, and a second for exchange of audio signal (voice) between participants. For both, a primary consideration was maintaining low-latency communication.

The audio stream consists of compressed voice packets that are sent between the terminals and the server. For this purpose, we use the UDP transport layer, which is typically preferred for real-time streaming [16]. In our case, we prioritize low latency, and thus, do not attempt to retransmit lost packets. Instead, we apply concealment techniques at reception to mask the effects of lost packets, based on algorithms provided by the Speex library [11]. The server then distributes the incoming streams to each of the other participants in the virtual room. To minimize overhead, packet headers are kept as small as possible, containing only a user identifier for the multi-party system and a sequence number for each packet. In case the activity detector does not find voice in a frame, the terminal sends only the header, with no audio data. This allows the receivers to maintain synchronization with little bandwidth use. Each voice stream produced by a terminal uses between 2.4 and 30.4 kbps of bandwidth, depending on the voice activity.

Terminals and server also need to exchange messages to manage the service and allow actions such as logging in, sending a user's contacts, inviting to a conversation or entering "whisper mode", described in the following section. TCP transport is used to ensure reliable delivery of these messages, which are short and require very little bandwidth. TCP is also used for activity of each terminal.

### 2.4 User interface

Usability of the system was considered paramount, and the user interface was designed to ensure easy visibility of system state and available actions. When users enter a

conversation, they see the other participants represented by their picture or avatar, and their name below. The avatars can be moved by finger-dragging them to the desired position, which immediately affects the spatial audio rendering as perceived only by the user applying the action. The system does not enforce consistency between spatial representations and thus, each participant is able to produce an individualized spatial arrangement.

Dragging an avatar into the semi-circular region in front of the user's own representation at the bottom-center of the screen invokes the "whisper mode" between the two individuals, as shown in Figure 3a. As visual feedback for the users engaged in a private conversation, their semi-circular regions turn red, as seen in Figure 3b. Other participants observe the names of the private conferees highlighted in red to inform them that a private conversation is taking place.

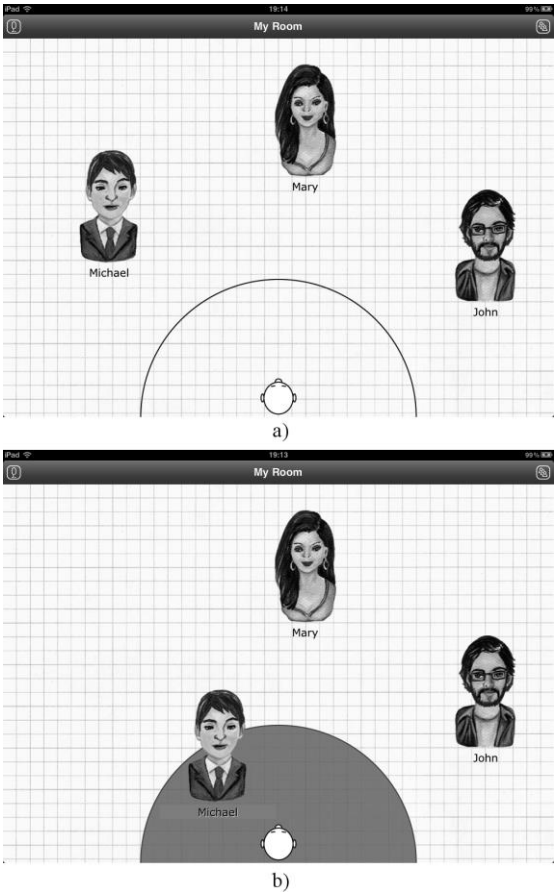



Fig.3. Audio-conferencing system screen display. (a) Your position is represented by the small head in the bottom-center of the screen. The semi-circle in front of you represents the "whisper zone".

## 3 METHOD

### 3.1 Equipment

For the following experiments, the aforementioned software was run on four iPad devices. Participants wore conventional Apple earbud headphones with built-in microphones. End-to-end audio latency was measured at 90 ms.

The audioconferencing server software, which ran on a dedicated, general-purpose PC, was instrumented to log the times during which each participant was speaking, the movements by each participant of all avatars, and the use of the whisper mode.

For network communications, all devices were connected to the shared university network. The server used an Ethernet connection for reliability, and the iPads connected over Wi-Fi, each on a different access point given their distributed locations. Some trials using 3G connectivity instead of Wi-Fi at preliminary tests provided non appreciable differences in quality."

### 3.2 Participants

Ten groups of four participants, drawn from the Technical University of Valencia, student and staff population took part in the experiment. These individuals ranged in age from 24 to 43 years old (mean 30.5), 25 male and 15 female. All participants self-reported as having normal hearing. Participants were recruited informally, and compensated for their time by token gifts and motivated to take the experiment seriously by the offer of additional prizes for the best "score" obtained on a combination of the experimental tasks.

As discussed below, a further test was later conducted with ten different participants, from a second university to evaluate the compare the effects of fixed and dynamic sound sources on user performance of identifying errors that were deliberately introduced. The subjects ranged from 21 to 41 years old (mean 27.6), six male and four female.

An issue raised during the debrief session with our pilot subjects was that their familiarity with each other may have facilitated their identification of everyone's voices. In such a situation, the importance of the visual display of the avatars to know who is speaking is quite likely reduced. To control for this effect, we screened participants to ensure that approximately half of the groups consisted entirely of individuals who were strangers to each other. The remaining groups contained at least a subset of participants with prior familiarity with each other.

### 3.3 Tasks

Our experiment was intended to determine the conditions under which the ability to change the arrangement of participants in an audioconference is considered useful. In a preparatory study, we asked participants to use our system to work on the "desert survival problem" [17] as a topic of group discussion. However, we found that participants almost immediately focused their visual attention on the list of items available to them and rarely attended to the graphical display of the audioconferencing interface. This precluded the possibility of serious investigation of the conditions of interest to us.

As a result, we discarded this option and opted to choose three other tasks that did not encourage a visual focus on information outside of the audioconferencing system's own graphical user interface display. Each task

was chosen based on our hypothesis of the importance of the different features we were testing.

The first task consisted of a group discussion on a topic of local interest, serving as a general-purpose audioconferencing activity.

The second task was a four-player instance of the pirate game [18], intended to simulate a business negotiation session. Participants were designated as greedy, rational pirates, who voted on the distribution of gold coins proposed by the most senior pirate. Unless the suggested allocation is accepted by 50% or more of the pirates, the proposing pirate is thrown overboard and the next senior pirate proposes a distribution. We relied on the fact that most of our participants were not familiar with the optimal solution to this game, thus ensuring a lively audioconferencing session, and we anticipated that "whisper mode" would prove desirable for the game.

The third task was a critical listening activity involving multiple musical instruments, in which participants assumed the role of a recording engineer who was asked to identify the performers who were playing with errors. This activity, illustrated in Figure 4, was inspired by our involvement in network distributed musical teaching and performance sessions, for which the ability to focus on the playing of an individual performer is considered valuable if not necessary. For this task, we predicted that most participants would appreciate the ability to rearrange the sound sources to facilitate concentration on a subset of one or more tracks.

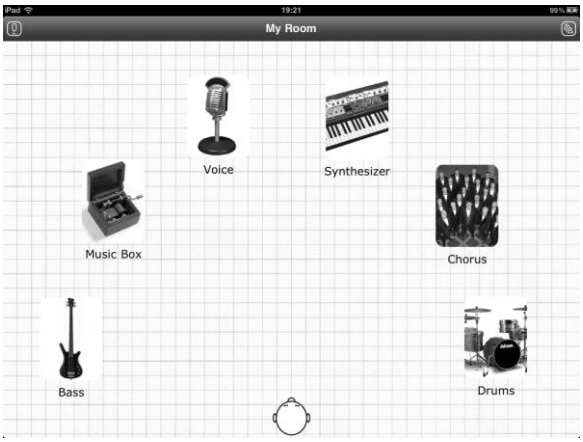



Fig.4. Recording engineer screen display.

To compare performance on this task when participants were not allowed to move the avatars representing the instruments, this task was later repeated with ten different participants, using a basic version of the system with fixed sources.

Participants were first briefed on the experiment and the different modes of operations that were to be presented. After moving to separate rooms, they were then asked to familiarize themselves with the system by trying the various features during a training period that lasted a maximum of 5 minutes. Once all were ready, the three tasks were presented in balanced order across groups. Each task was allowed to run for a maximum of ten minutes, following which, participants completed a brief questionnaire, consisting of six to eight questions related to their ease of participating in the task, their level of immersion, and the perceived value of the system's features. Following the questionnaire, the participants were instructed to proceed with the next task. After the end of the last round, a final questionnaire was administered to assess overall effects.

## 4 RESULTS

Following the experiment, the results were processed and analyzed according to each task. The remainder of this section presents the results, classified according to the task, whereas Section 5 discusses the two new interactive features introduced in this work.

### 4.1 Group Discussion

Following completion of the group discussion task, participants answered the questionnaire shown in Table 1. The questionnaire results are summarized in Figure 5. All factors were rated positively (score ≥ 5 out of 7), with the exception of the perceived value of the private discussion

Table 1. Questionnaire for Task 1

| # | Question | Scoring Scale |
|---|---|---|
| 1.1 | Did you feel that you were able to contribute usefully to the discussion? | 1 (Not at all) – 7 (Very much) |
| 1.2 | During the session, determining which individual was speaking was: | 1 (Difficult) - 7 (Easy) |
| 1.3 | Rate your overall comprehension of the session: | 1 (Poor) – 7 (Excellent) |
| 1.4 | How immersed did you feel in the session? | 1 (Not at all) – 7 (Completely) |
| 1.5 | How much of your attention was focused on determining who was speaking? | 1 (Most) – 7 (None) |
| 1.6 | How helpful was the location of the voices of the other individuals? | 1 (Not at all) – 7 (Very helpful) |
| 1.7 | How helpful was the ability to move the individuals around you? | 1 (Not at all) – 7 (Very helpful) |
| 1.8 | How helpful was the ability to have a "private" discussion with another individual? | 1 (Not at all) – 7 (Very helpful) |

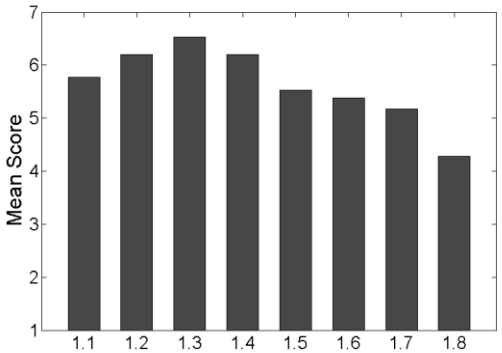



Fig.5. Mean scores of questionnaire from Task 1.

feature (mean rating of 4.3). The results indicated a somewhat higher perceived value for the ability to move the avatars of the other participants (mean rating of 5.175, and indeed, half of the participants made use of this feature to establish a preferred configuration during the training period.

In order to inform future implementations of the system with regard to users' habits, the server logged the coordinates of avatar positions throughout the experiment. These data were then filtered by a low-pass Gaussian filter, and points in low-density regions were removed, to obtain the point cloud of avatars positions, shown in Figure 6. As seen, there were five main areas of preferred avatar positions for the three on-screen avatars during the task. The central position is the most common, followed by two different angles for lateral positions, approximately symmetrical about the vertical axis. Avatars tend to be spread out with a medium distance around the listener, with corners avoided.

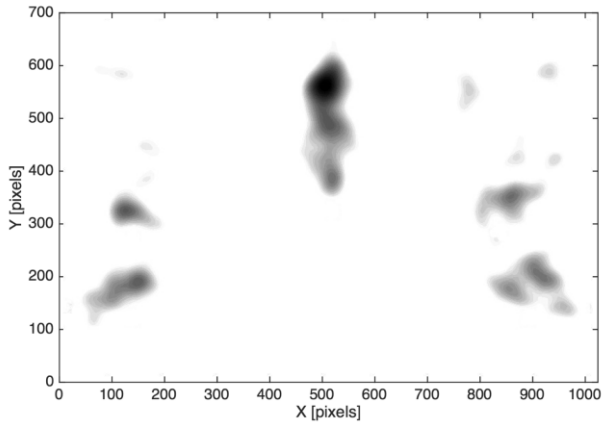



Fig.6. Cloud of positions used by participants for placing the three avatars in Task 1.

### 4.2 Negotiation Scenario

Following completion of the pirate game, participants answered the questionnaire shown in Table 2. The questionnaire results are summarized in Figure 7. All factors were rated positively (score $\geq$ 5 out of 7). As we anticipated, the most important difference from the first task was the increased perception of the value of the "whisper mode" private discussion feature, which was given the highest rating of all the questions (mean rating of 6.6).

During this activity, each participant entered into a private conversation, on average, every 1.6 minutes, distributed approximately equally throughout the sessions, which lasted 7.6 minutes on average. We did not find any correlation between the duration of the session and the number of private conversations that took place. Furthermore, there was no gender bias in the initiation or receipt of "whisper mode" requests.

### 4.3 Critical Listening

The third task was carried out in two different formats: dynamic sources, in which the participant had the freedom to move the avatars (in this case, of the musical instruments), and static sources, in which their positions were fixed throughout the experiment. This was done in order to ascertain objectively whether avatar movement was beneficial to the activity, independently of the participants' own perception of the value of this feature. The subjects who participated in the static sources version of this experiment did not carry out a test with the dynamic sources, so as to ensure they were not already familiar with the errors that had been introduced into the musical tracks.

Table 2. Questionnaire for Task 2

| # | Question | Scoring Scale |
|---|---|---|
| 2.1 | During the session, determining which individual was speaking was: | 1 (Difficult) - 7 (Easy) |
| 2.2 | How much of your attention was focused on determining who was speaking? | 1 (Most) – 7 (None) |
| 2.3 | How immersed did you feel in the session? | 1 (Not at all) – 7 (Completely) |
| 2.4 | How helpful was the location of the voices of the other individuals? | 1 (Not at all) – 7 (Very helpful) |
| 2.5 | How helpful was the ability to move the individuals around you? | 1 (Not at all) – 7 (Very helpful) |
| 2.6 | How helpful was the ability to have a "private" discussion with another individual? | 1 (Not at all) – 7 (Very helpful) |

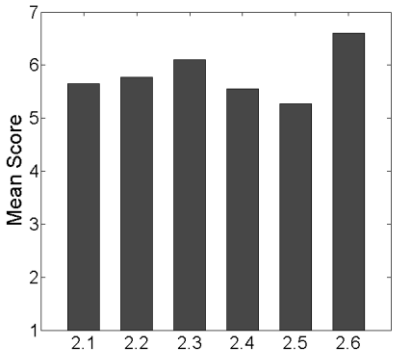



Fig.7. Mean scores of questionnaire from Task 2.

Following completion of the critical listening task, participants answered the questionnaire shown in Table 3. The questionnaire results for both dynamic and static sources are summarized in Figure 8. It may be noteworthy that all ratings given for the static sources version of the experiment were lower than that for the dynamic sources version. Consistent across both groups was the perceived difficulty of the task, which is unsurprising, with six instruments playing simultaneously.

In the dynamic sources conditions, participants moved the avatars frequently, with an average of 16.3 movements per minute; there was no significant difference between musicians and non-musicians in this regard. The artifacts that were inserted into the audio tracks consisted of three errors per minute in each of the

Table 3. Questionnaire for Task 3

| # | Question | Scoring Scale |
|---|---|---|
| 3.1 | During the session, recognizing errors in the musical performance was: | 1 (Difficult) - 7 (Easy) |
| 3.2 | During the session, focusing your attention on individual instruments was: | 1 (Difficult) - 7 (Easy) |
| 3.3 | How immersed did you feel in the session? | 1 (Not at all) – 7 (Completely) |
| 3.4 | How helpful was the separation of the instruments? | 1 (Not at all) – 7 (Very helpful) |
| 3.5 | How helpful was the ability to move the instruments around you? | 1 (Not at all) – 7 (Very helpful) |

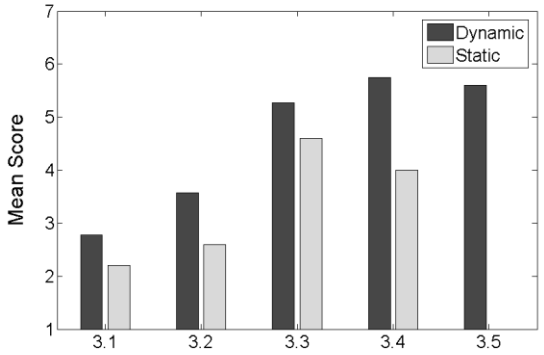



Fig.8. Mean scores of questionnaire from Task 3.

voice and music box tracks, and ten errors per minute in the bass track. Although it is questionable whether the rate at which the avatars were moved was helpful to the critical listening task, the perceived ability to recognize errors in the performance was rated higher by the participants who were able to move the sources during the experiment. However, since these two conditions involved different subject pools, we should be cautious about drawing conclusions from these differences.

Differences in the detection of the actual errors in the performance are discussed below.

## 5 DISCUSSION

This section discusses the statistical analysis of user behaviour of avatar movement, as well as the two new interactive features introduced in this work.

### 5.1 Avatar movement

The main difference between the first two tasks was the requirement of negotiation for the pirate game. Questionnaire results indicated that the ability to move the avatars of the other participants was rated similarly positively for the two tasks (Figures 5 and 7). However, there was considerably greater variance in the use of this feature for the first task compared to the second, as seen in Figure 9. This difference is likely explained by the need for avatar movement in the second task to invoke whisper mode, which was used by most participants. In contrast, avatar movement in the first task was entirely optional, and seemed to be used by participants more as an element of their experimentation with the system.

For the critical listening task, participants in the static sources condition performed better at detecting tone errors of the music box, whereas those in the dynamic sources condition were better at detecting distortion in the bass (Figure 10). However, ANOVA testing indicated that only the latter were significant ($F=10.935$, $fd=49$, $p=0.002$). Error detection was considered successful only when the identification was temporally accurate. Overall, error detection was slightly higher in the interactive (dynamic sources) condition. ANOVA testing indicated that false positives were also slightly higher, on average, in the dynamic sources condition ($F=4.114$, $fd=49$, $p=0.048$). The frequent movement of music avatars in the dynamic condition may have skewed perception of the associated musical content, resulting in this slightly higher incidence of false positives.

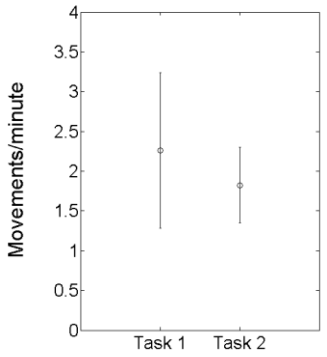



Fig.9. Use of avatar movement feature in the first two tasks. Values shown indicate the mean number of movements per minute and 95% confidence intervals.

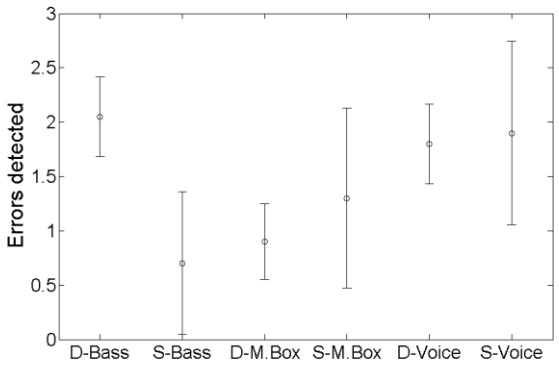



Fig.10. Detection of instrument errors in the critical listening task, with means and 95% confidence intervals.

### 5.2 Whisper mode

Given the nature of group discussion in the first task, private discussion was unnecessary, so the neutral rating for "whisper mode" feature in Figure 5 is unsurprising. We speculate that participants may have interpreted the negative end of the scale, "not at all helpful" as implying that the feature interfered with their task, and thus

considered a neutral rating as appropriate when they did not make use of the feature. In fact, only one out of the ten groups tested made any use of this feature, and no significant difference was observed in the answers of that group to the questionnaire.

One of the pilot participants noted that even with the GUI display offering visual feedback, he was surprised that he could suddenly no longer hear two of the other participants. In the absence of whisper mode capability in the pirate game, all but one of the pilot participants used the dynamic avatar control at the beginning of the session to verify that it was working, but then did not use it again. However, when whisper mode was enabled, all participants made use of this capability and commented that it was very important.

## 6 CONCLUSIONS

The use of binaural audio was demonstrated by prior research to improve the quality of multi-party audioconferencing. Our work analyzed the potential benefits of the added capabilities of avatar movement and sidebar, "whisper mode" conversations under different use contexts of discussion, negotiation, and critical listening tasks.

The results indicate that participants found avatar movement to be useful in all three contexts. However, approximately half of the participants only used this feature for initial positioning of the avatars in a desired configuration in the first task, and the movement of avatars did not appear to improve the quality of information obtained.

The ability to maintain sidebar conversations was used frequently and rated highly in the negotiation activity. However, participants made almost no use of this feature in the conversation activity, for which negotiation was not involved.

Importantly, we also demonstrated the viability of implementing an effective multi-party spatial audioconference on today's mobile devices. By optimizing the signal processing stages and employing an efficient communications protocol, we were able to achieve a stable, energy efficient, and highly usable interface, even with several simultaneous users.

## 7 ACKNOWLEDGEMENTS

The Spanish Ministry of Economy and Competitiveness supported this work under the project TEC2012-37945-C01. The experiments were carried out with the support of the Transatlantic Partnership for Excellence in Engineering (TEE). This support is gratefully acknowledged.

## 7 REFERENCES

[1] S. Goose, J. Riedlinger, and S. Kodlahalli, "Conferencing3: 3d audio conferencing and archiving services for handheld wireless devices", *Int. J. Wireless Mobile Comput*. vol. 1, pp. 5-13 (2005 Nov). http://dx.doi.org/10.1504/IJWMC.2005.008049

[2] S. Deo, M. Billinghurst, N. Adams, and J. Lehikioinen, "Experiments in spatial mobile audioconferencing", in Proceedings of the 4th international conference on mobile technology applications and systems and the 1st international symposium on Computer human interaction in mobile technology, vol. 7, pp. 447-451 (ACM Press, 2007). http://dx.doi.org/10.1145/1378063.1378133

[3] A. Sellen, B. Buxton, and J. Arnott, "Using spatial cues to improve videoconferencing", in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '92, pp. 651–652 (ACM New York, 1992). http://dx.doi.org/10.1145/142750.143070

[4] J. J. Baldis, "Effects of spatial audio on memory, comprehension, and preference during desktop conferences", in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '01, pp. 166–173 (ACM New York, 2001). http://dx.doi.org/ 10.1145/365024.365092

[5] M. Billinghurst, J. Bowskill, M. Jessop, and J. Morphett, "A wearable spatial conference space", in *Wearable Computers* (1998 Oct.). http://dx.doi.org/ 10.1109/ISWC.1998.729532

[6] A. Kan, G. Pope, C. Jin, and A. V. Schaik, "Mobile spatial audio communication system", in International Conference on Auditory Display (2004 June).

[7] R. Kilgore, M. Chignell, and P. Smith, "Spatialized audioconferencing: what are the benefits?", in Proceedings of the 2003 conference of the Centre for Advanced Studies on Collaborative research, CASCON '03, pp. 135–144 (IBM Press, 2003).

[8] K. Inkpen, R. Hegde, M. Czerwinski, and Z. Zhang, "Exploring spatialized audio & video for distributed conversations", in Proceedings of the 2010 ACM conference on Computer supported cooperative work, CSCW '10, pp. 95–98 (ACM New York, 2010).

[9] M. Cohen, "Exclude and include for audio sources and sinks: Analogs of mute & solo are deafen & attend", Presence: Teleoper. Virtual Environ. 9, 1, pp 84–96 (2000 Feb.). http://dx.doi.org/ 10.1162/105474600566637

[10] O. N. N. Fernando, M. Cohen, and A. D. Cheok, "Multipresence-enabled mobile spatial audio interfaces", in Proceedings of the 9th international conference on Entertainment computing, ICEC'10, pp. 434–436 (Springer-Verlag, Berlin, 2010).

[11] J.-M. Valin, "Speex: A free codec for free speech", in Australian National Linux Conference (2006).

[12] F. Rumsey, "Spatial audio processing", *J. Audio Eng. Soc.*, vol. 61, pp. 474-478 (2013 June).

[13] J. W. S. Rayleigh, "The Theory of Sound (2nd Edition)" (Dover Publications, New York, 1945).

[14] J. Mackenzie, J. Huopaniemi, V. Välimäki, and I. Kale, "Low-order modeling of head-related transfer functions using balanced model truncation", *IEEE Signal Process. Lett.*, Volume 4, no. 2, pp. 39-41 (1997 Feb.). http://dx.doi.org/10.1109/97.554467

[15] J. J. Lopez, M. Cobos, and B. Pueo, "Elevation in wave-field Synthesis using HRTF cues", *Acta Acustica*, vol. 96, pp. 340-350 (2010). http://dx.doi.org/10.3813/AAA.918283

[16] A. Xu, W. Woszczyk, A. Settel, B. Pennycook, R. Rowe, P. Galanter, J. Bary, G. Martin, J. Corey, and J. R. Cooperstock, "Real-time streaming of multichannel audio data over Internet", *J. Audio Eng. Soc*., vol 48, pp. 627-641 (2000 July).

[17] J. C. Lafferty, P. M. Eady, and J. Elmers, "The desert survival problem", *Human Synergistics*, Plymouth, Michigan: Experimental Learning Methods (1974).

[18] I. Stewart, "Mathematical recreations: a puzzle for pirates", Scientific American, vol 280, num. 5, pp. 98–99 (1999 May).

## THE AUTHORS

E. Aguilera

J.J. Lopez

J. Cooperstock

Emanuel Aguilera received a telecommunications engineering degree in 2004 and a M.S. degree in Artificial Intelligence, Pattern Recognition and Digital Image in 2011, both from the Technical University of Valencia, Spain. He is a researcher and senior programmer at the Institute of Telecommunications and Multimedia Applications (iTEAM), where he has been working since 2006 on the area of digital signal processing for audio, multimedia, virtual reality and mobile devices applications. He is interested in wave-field synthesis, image processing, real-time multimedia processing for telecommunications and audio applications for mobile platforms.

•

Jose Javier Lopez (Ph.D., Technical University of Valencia 1999). Since 1993, he has been involved in education and research at the Communications Department, Technical University of Valencia, where he is currently a Full Professor. His research activity is centered on digital audio processing in the areas of spatial audio, wave field synthesis, physical modeling of acoustic spaces, efficient filtering structures for loudspeaker correction, sound source separation, and development of multimedia software in real time. He has published more than 200 papers in international technical journals and at renowned conferences in the fields of audio and acoustics and has led more than 30 research projects. Dr. Lopez was workshop co-chair at the 118th Convention of the Audio Engineering Society in Barcelona and has been serving on the committee of the AES Spanish Section for 15 years. He is a full ASA member, AES member and IEEE senior member.

•

Jeremy Cooperstock (Ph.D., University of Toronto, 1996) is an associate professor in the department of Electrical and Computer Engineering, a member of the Centre for Intelligent Machines, and a founding member of the Centre for Interdisciplinary Research in Music Media and Technology at McGill University. He directs the Shared Reality Lab, which focuses on computer mediation to facilitate high-fidelity human communication and the synthesis of perceptually engaging, multimodal, immersive environments. Cooperstock's work on the Ultra-Videoconferencing system was recognized by an award for Most Innovative Use of New Technology from ACM/IEEE Supercomputing and a Distinction Award from the Audio Engineering Society. The research he supervised on the Autour project earned the Hochhausen Research Award from the Canadian National Institute for the Blind and an Impact Award from the Canadian Internet Registry Association, and his Real-Time Emergency Response project won the Gold Prize (brainstorm round) of the Mozilla Ignite Challenge. He is an associate editor of the Journal of the AES.

Figure 1

Click here to download Figure Figure 1.tif

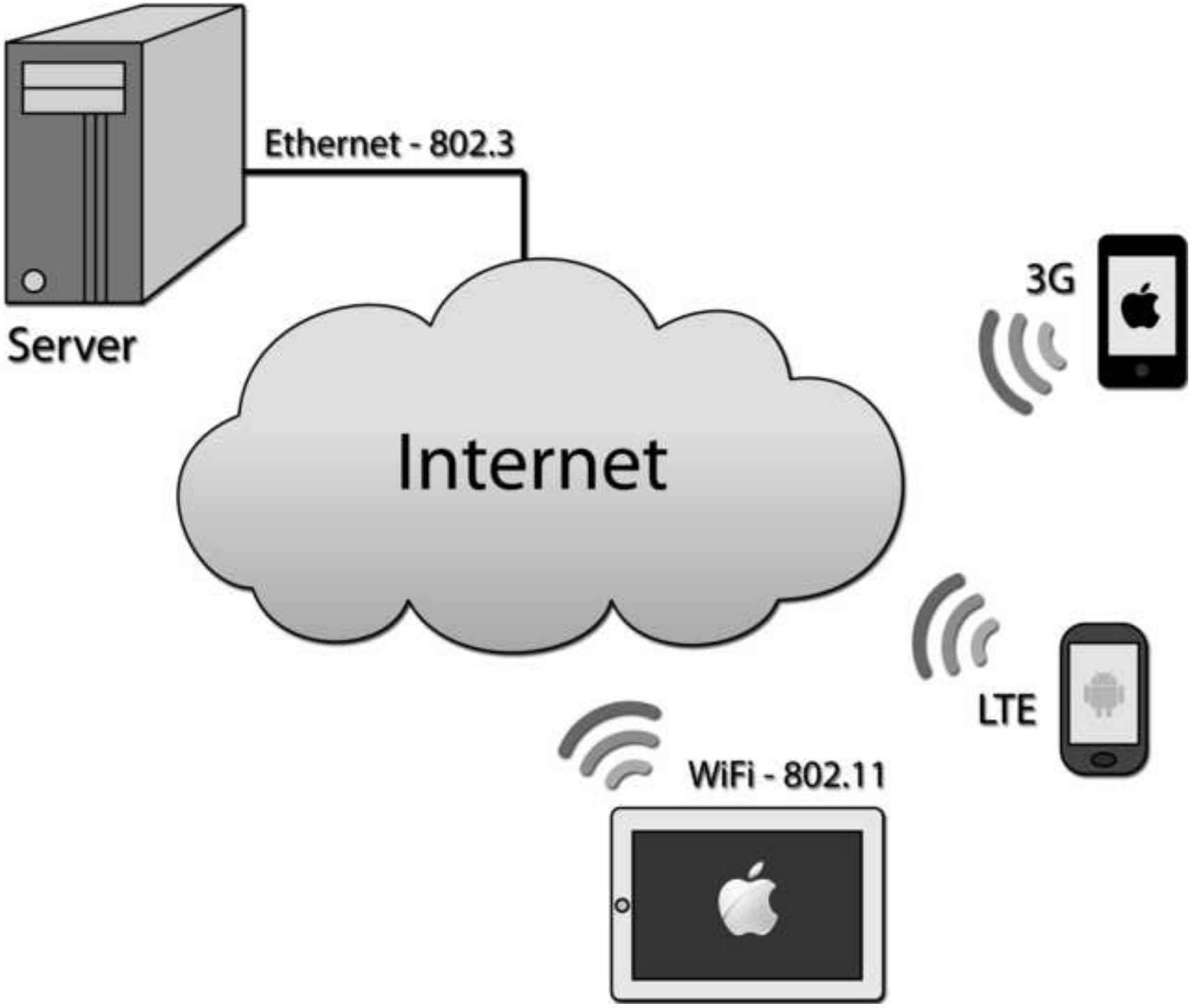


Figure 2

Click here to download Figure Figure 2.tif

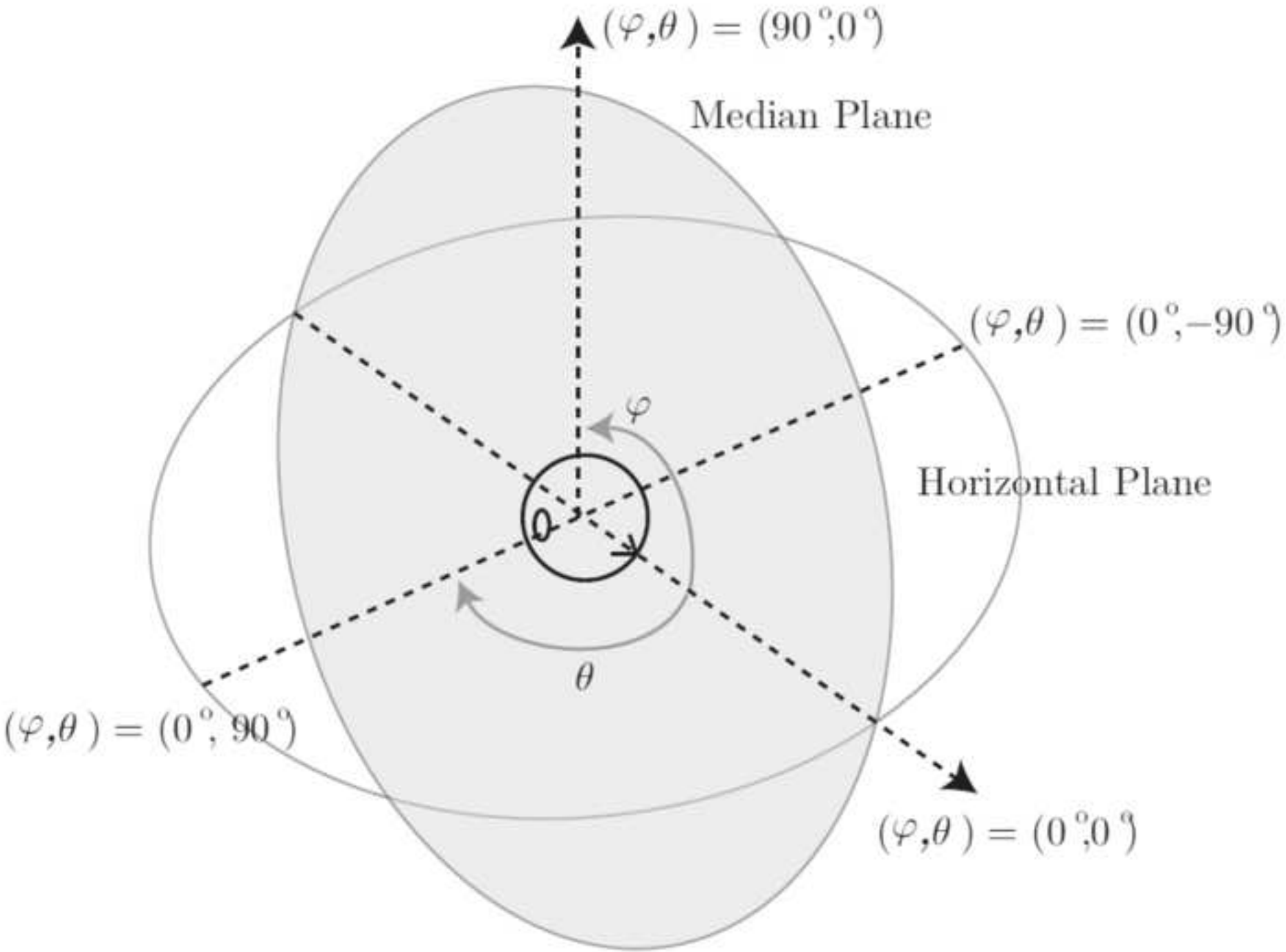
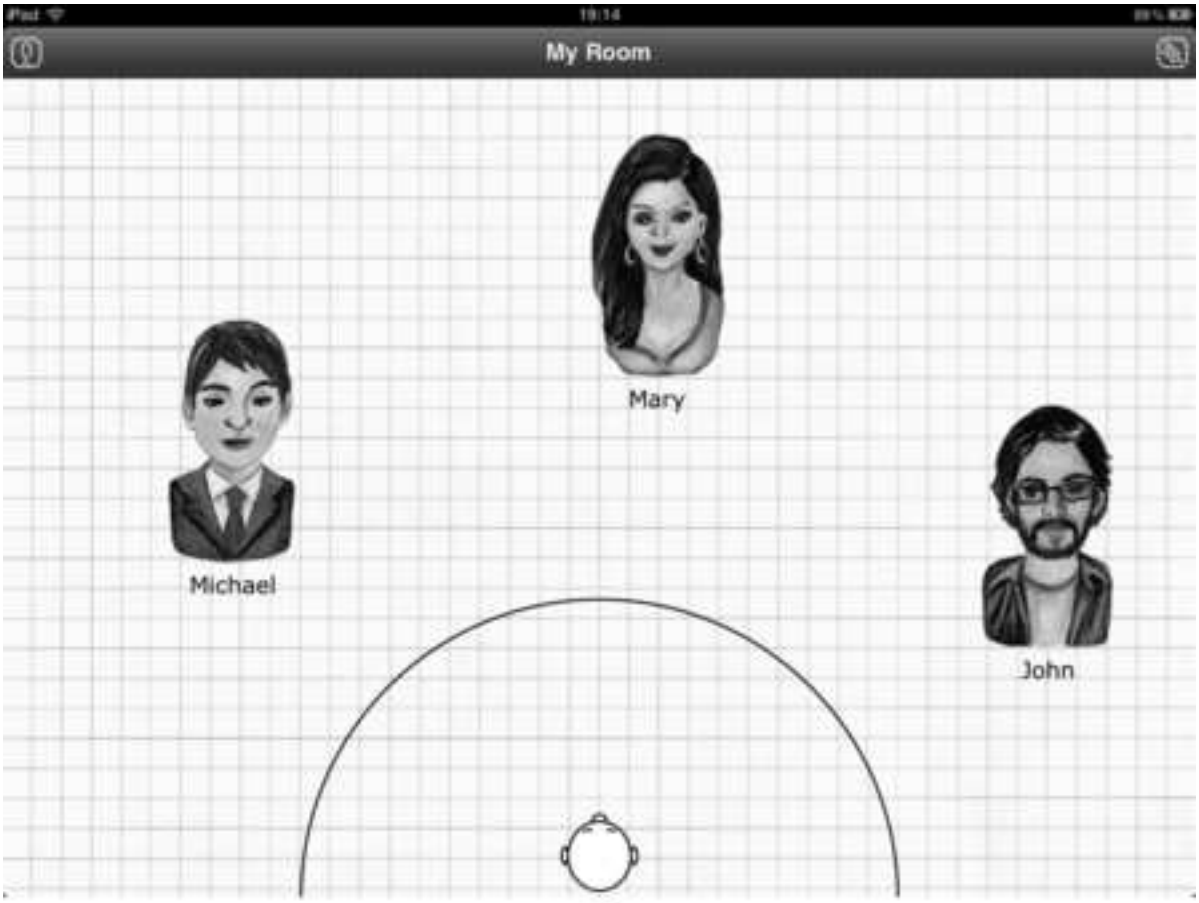
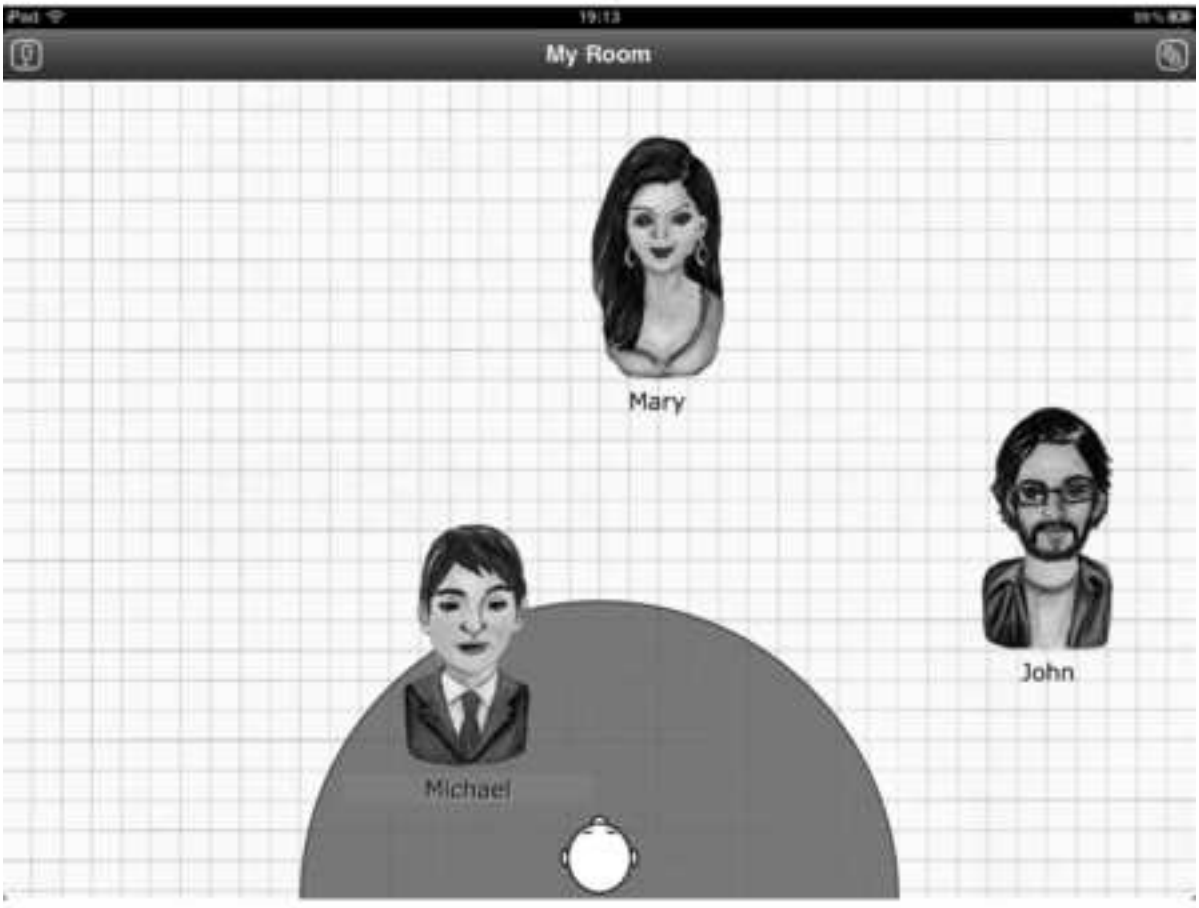


Figure 3






a)




b)

Figure 4

Click here to download Figure Figure 4.tif

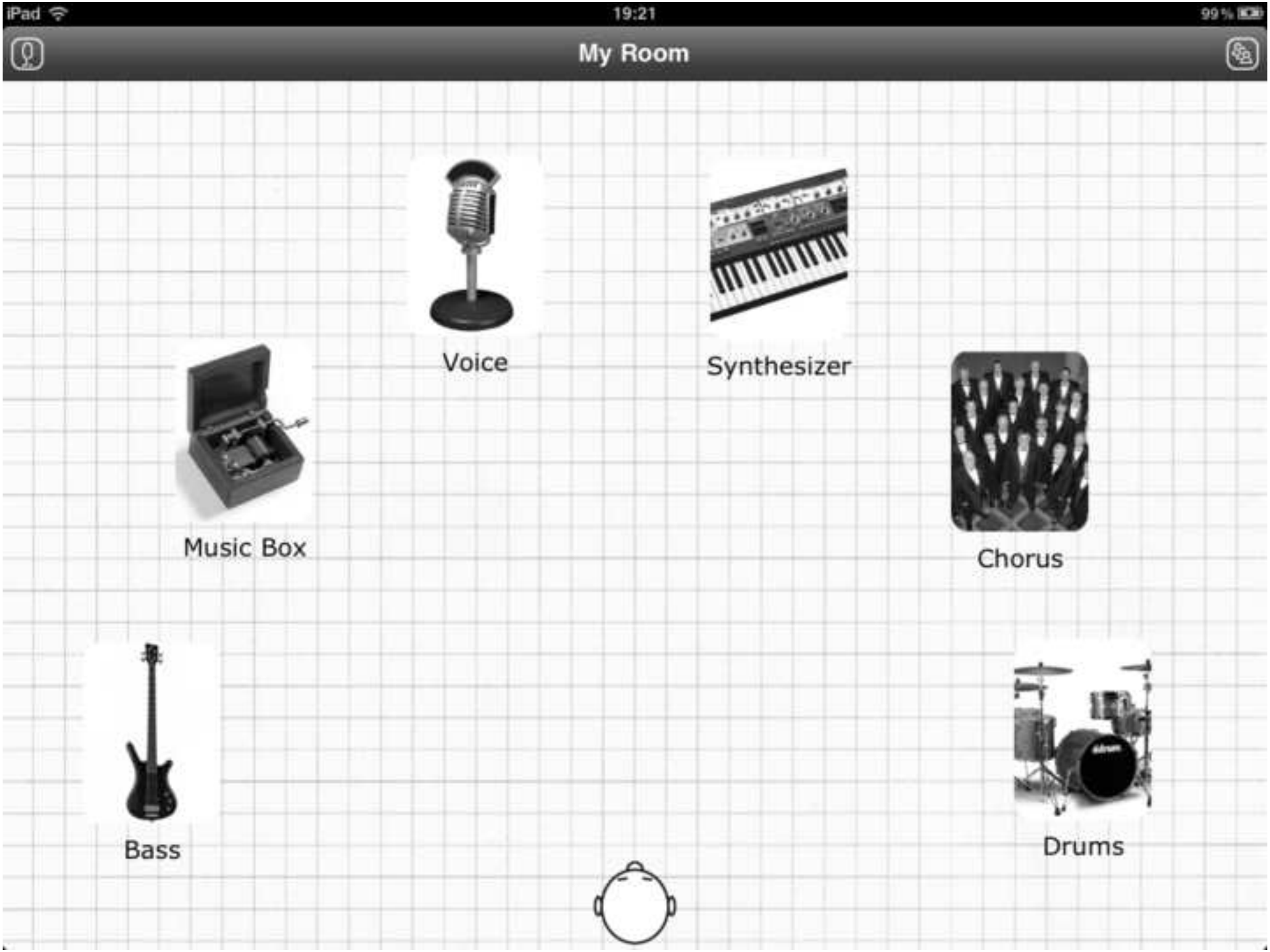


Figure 5

Click here to download Figure Figure 5.eps

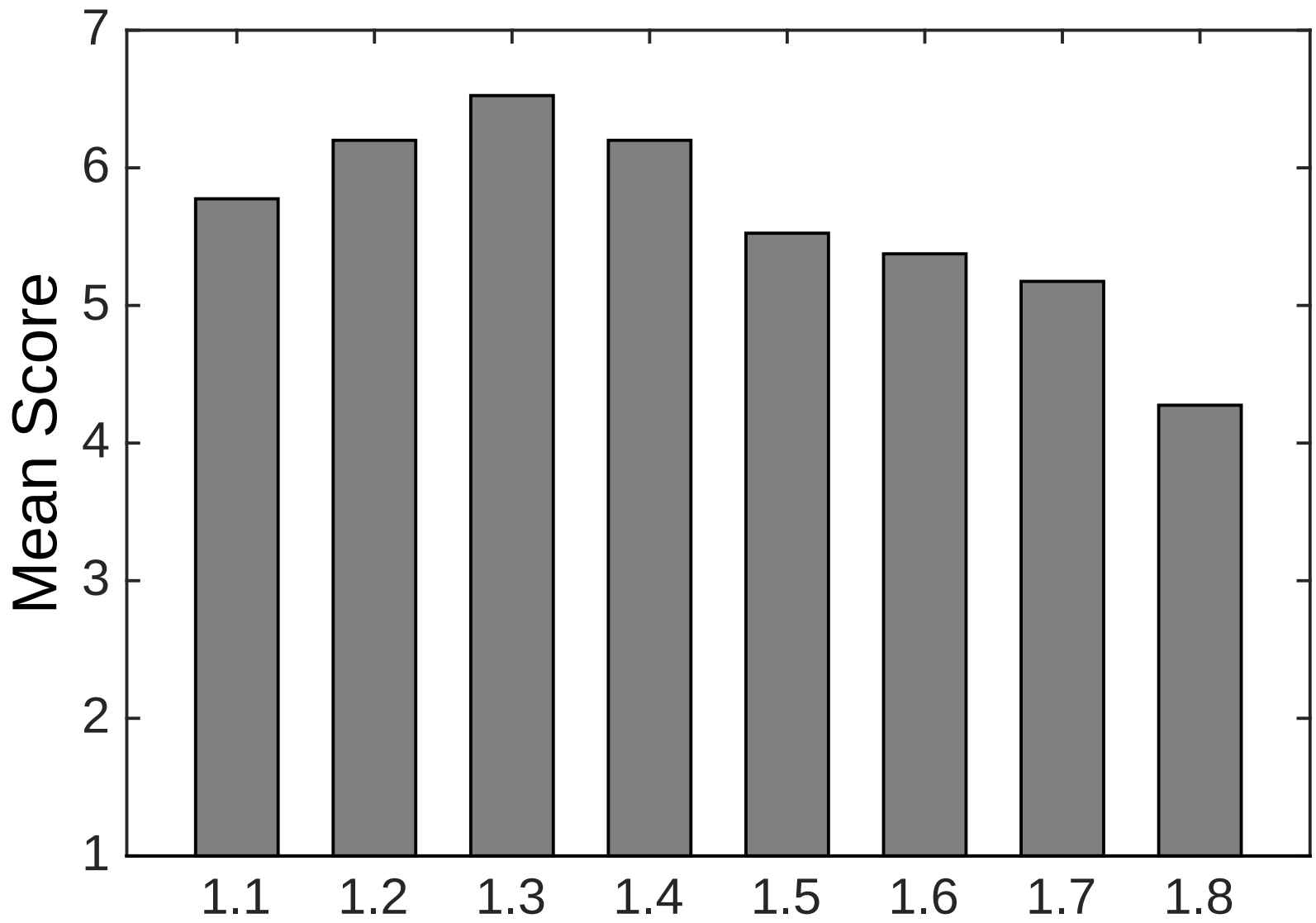


Figure 6

Click here to download Figure Figure 6.tif

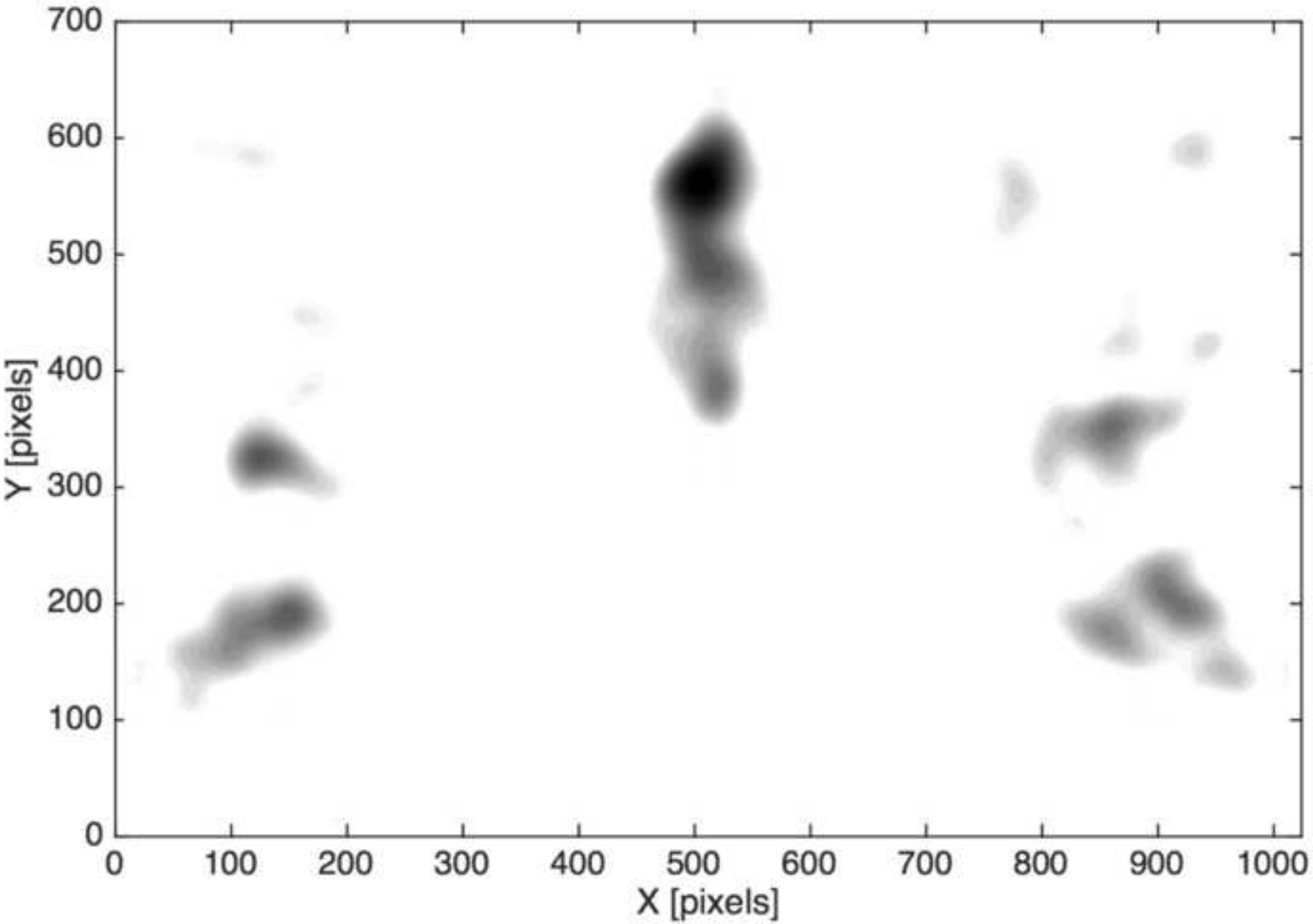


Figure 7

Click here to download Figure Figure 7.eps

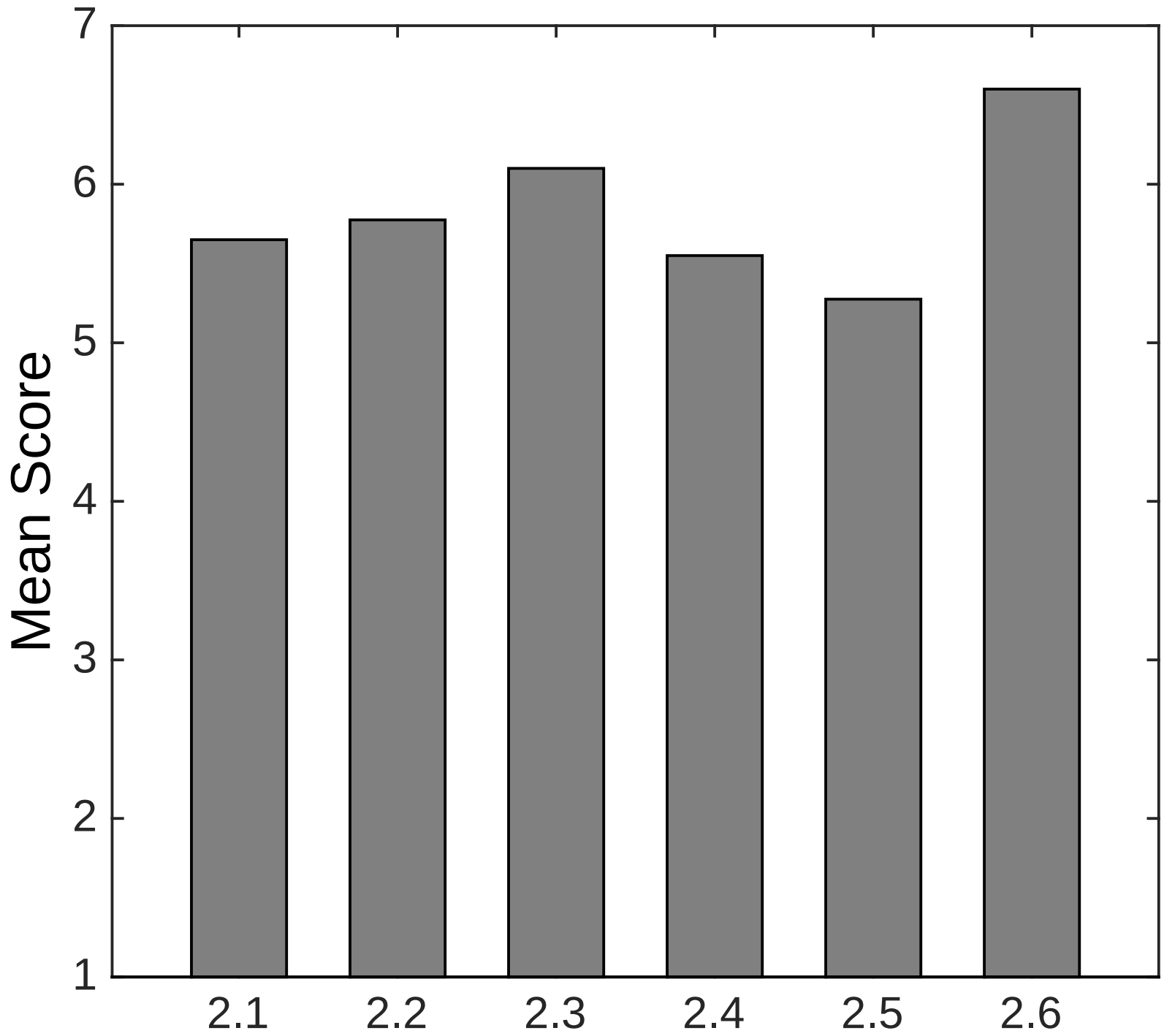


Figure 8

Click here to download Figure Figure 8.eps

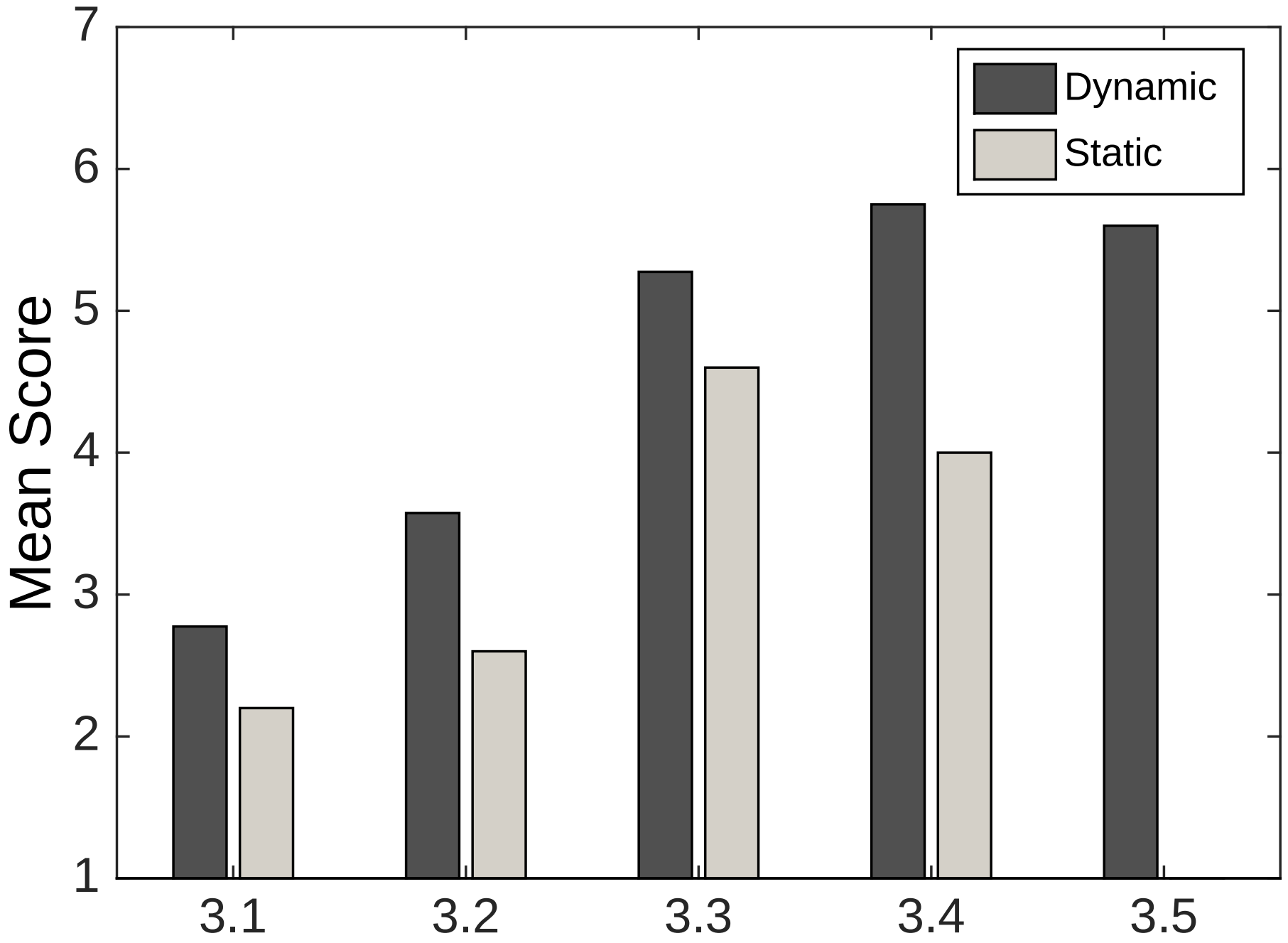


Figure 9

Click here to download Figure Figure 9.eps

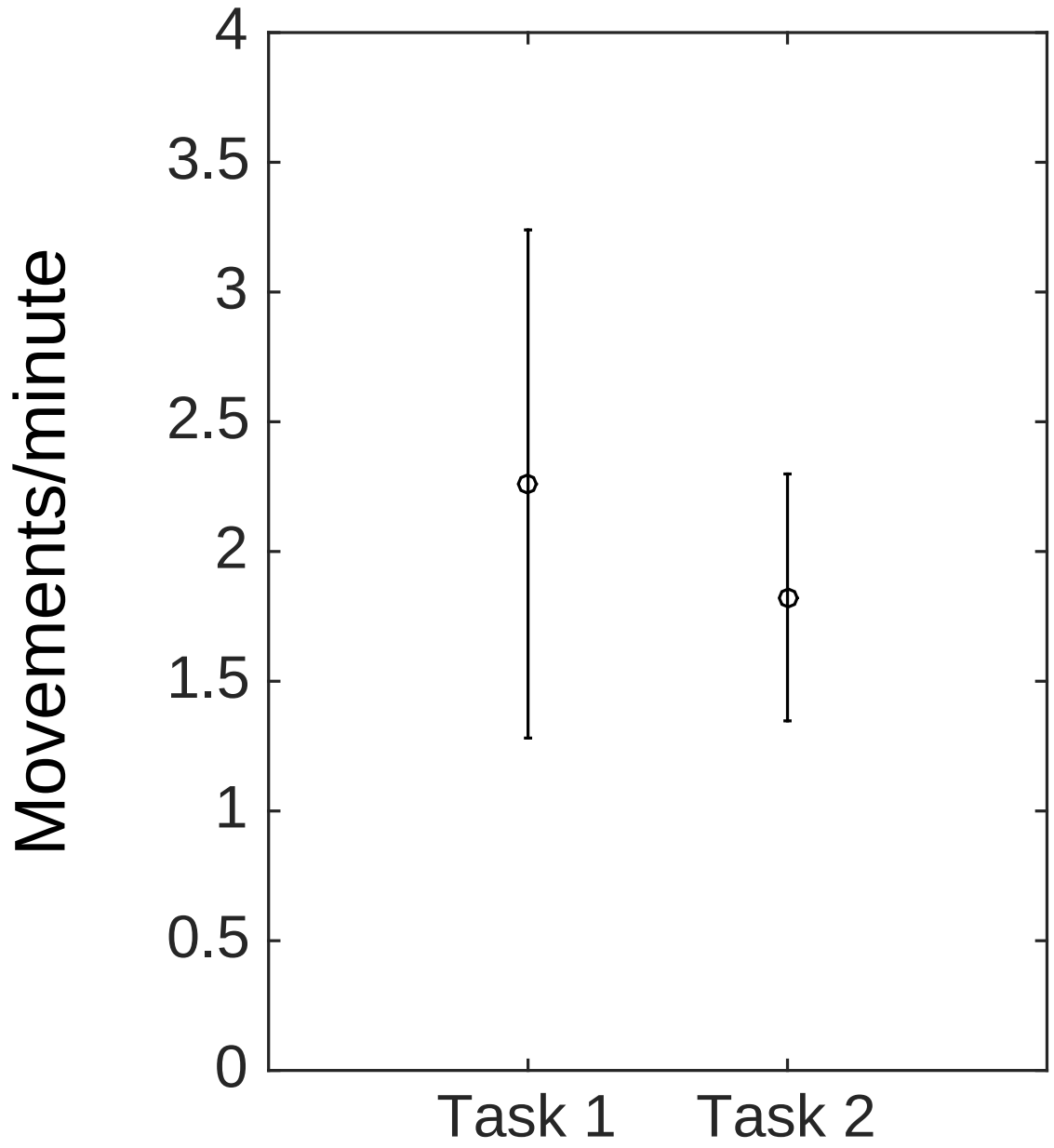
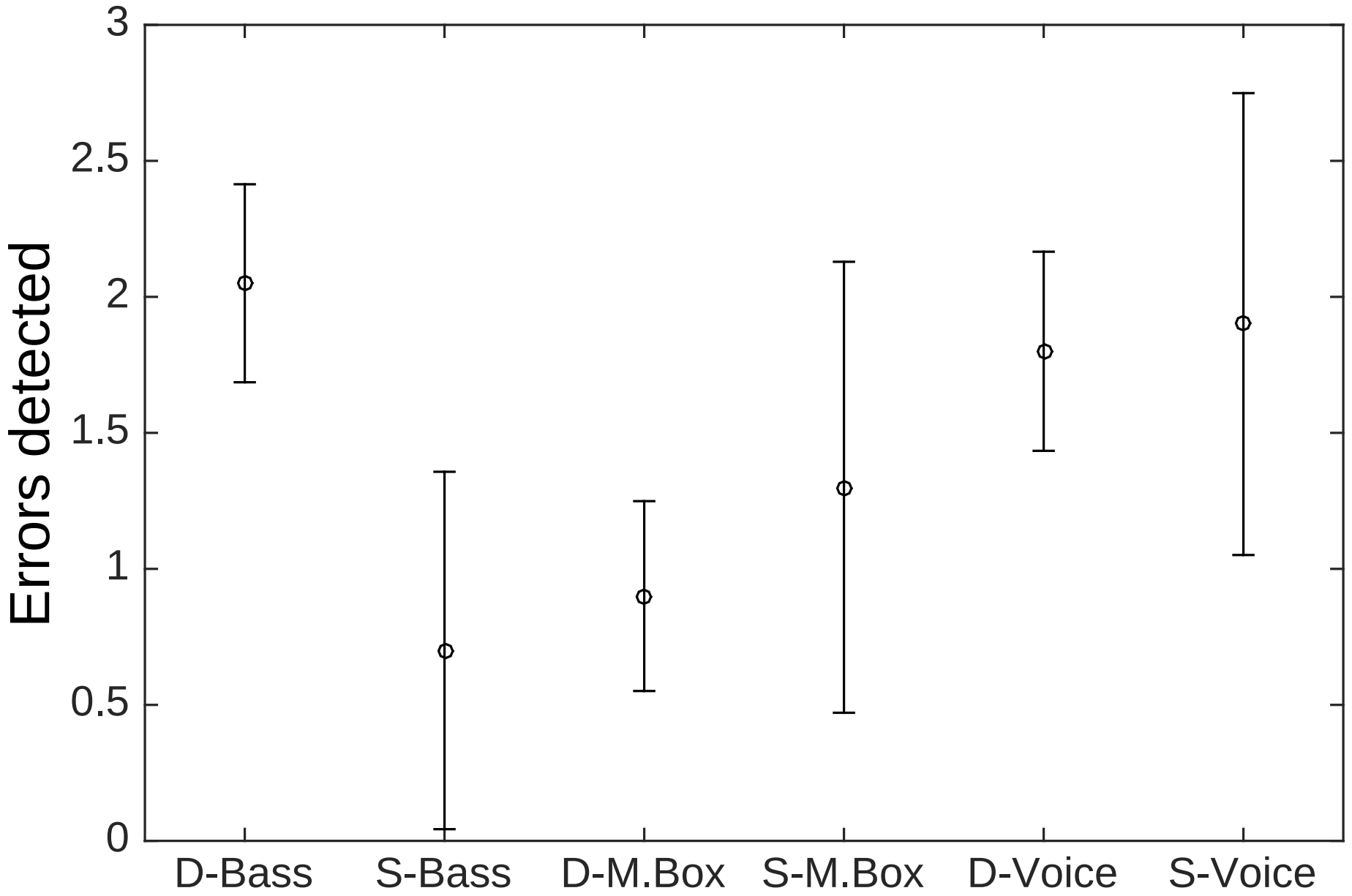


Figure 10

Click here to download Figure Figure 10.eps



Author photo 1

Click here to download Author photo Emanuel.tif

Author photo 2

Click here to download Author photo Jose Javier.tif

Author photo 3

Click here to download Author photo Jeremy.tif