Wireless Acoustic Tracking for Extended Range Telepresence

Ferdinand Packi, Frederik Beutler, and Uwe D. Hanebeck
Intelligent Sensor-Actuator-Systems Laboratory (ISAS),
Institute for Anthropomatics, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany.
ferdinand.packi@kit.edu, frederik.beutler@kit.edu, uwe.hanebeck@ieee.org.

***Abstract*—Telepresence systems enable a user to experience virtual or distant environments by providing sensory feedback. Appropriate devices include head mounted displays (HMD) for visual perception, headphones for auditory response, or even haptic displays for tactile sensation and force feedback. While most common designs use dedicated input devices like joysticks or a space mouse, the approach followed in the present work takes the user's position and viewing direction as an input, as he walks freely in his local surroundings. This is achieved by using acoustic tracking, where the user's pose (position and orientation) is estimated on the basis of ranges measured between a set of ceiling-mounted loudspeakers and a microphone array fixed on the user's HMD. To allow for natural user motion, a wearable, fully wireless telepresence system is introduced. The increase in comfort compared to wired solutions is obvious, as the user's awareness of distracting cables is taken away during walking. Lightweight design and small dimensions contribute to ergonomics, as the whole assembly fits well into a small backpack.**

I. Introduction

Telepresence systems are a means of putting a user into place to explore virtual environments or remote areas. Conceivable environments can either be entirely computer generated or established by a teleoperator, who acts as a proxy for the human user. In the latter case, visual or even multimodal (auditory, tactile) perception in a remote environment is transmitted over a distance, and output to the user by suitable interfaces, such as head mounted displays (HMD) and haptic displays [1]. A lineup of teleoperators developed and used at the ISAS laboratory is shown in Fig. 2 on the following page. Either way, the aim is to allow a human user to move freely and intuitively through arbitrarily large remote environments. Unlike many other approaches, the present system is controlled by the user's motion itself. Position, walking speed, and viewing direction are retrieved through the use of an acoustic tracking system. The remote scene is brought to the user via an HMD. Computation, scene rendering, and tracking is carried out on the Mobile Telepresence Unit, worn by the user in a backpack-like case. Applications lie in the field of exploring hazardous areas or forbidding places, such as nuclear facilities, military zones, mining and construction work, or areas exhibiting extreme climate conditions. Among the more friendly uses are educational scenarios, like virtual sightseeing, space exploration, or training for evacuations. During research, also multimodal gaming scenarios were developed,

978-1-4244-5864-6/10$26.00

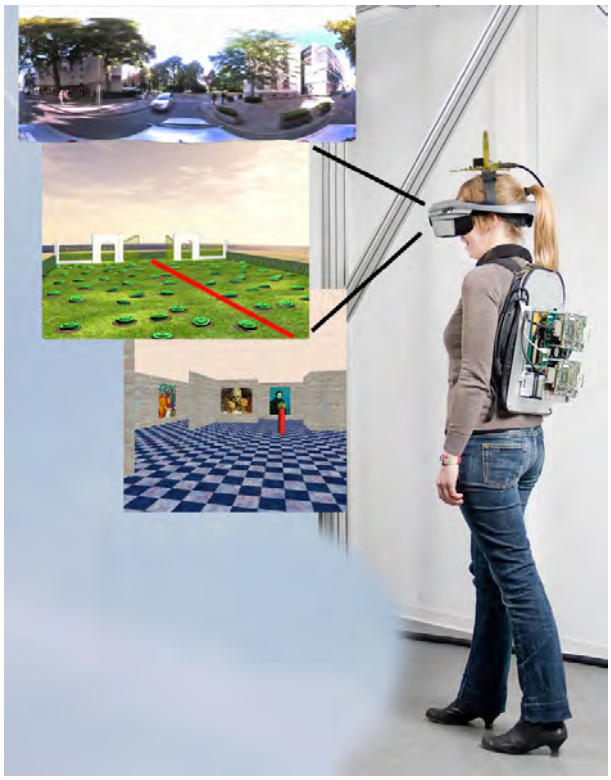

Fig. 1. User wearing the telepresence system. The depicted scenarios are (top to bottom): A spherical street panorama, a minesweeper game, and a virtual museum tour.

like PamCan, a Pacman™ clone or an excursion through the university cafeteria using the Quake™ engine. Fig. 1 shows a user wearing the telepresence system given exemplary scenic visualizations.

A. Requirements

To ensure a high degree of immersion into the remote environment, the user's motion in his local environment itself serves as an input to define position and viewing direction within target space. Therefore, a suitable indoor tracking system to determine the user's pose (head position and orientation) needs to be developed to meet the high demands regarding accuracy and latency. A major challenge is the fully wireless design, involving autonomous power supply, communication, and synchronization of the decentralized timebases.

Fig. 2. Teleoperators involved in the telepresence environment at the ISAS laboratory. Left: The Omnibase [2], an omnidirectional robot platform with pan-tilt stereo cameras. Top right: Radio controlled, teleoperated off-road vehicle. Bottom right: 6DOF miniature walking robots.

Moreover, ergonomic aspects need to be adressed, such as size and weight, to make the user feel comfortable wearing the mobile components.

B. Related Work

As stated earlier, the user's motion itself shall act as an input to the telepresence system. To determine the user's pose, a suitable tracking method has to be chosen. Relating to physics, a variety of approaches have explored all imaginable media, such as mechanical linkages, optics (marker based or image based), acoustics (time of flight, phase), radio (received signal strength, multilateration, triangulation), and (electro-)magnetic. Moreover, there is a choice between global (absolute) positioning and local systems, that measure parameters relative to their own coordinate system. In practice, a hybrid use of both principles is common, to overcome their respective deficiencies. As GPS is too imprecise for our purpose and mostly not available in indoor environments, alternative systems need to be discussed. Systems involving mechanical linkages suffer from limited workspace and poor ergonomics and therefore, are not qualified for pose tracking.

Among optical systems, [3] showed a head-tracking-system working with an array of ceiling-mounted infrared LEDs and specially designed optical sensors (lateral effect photo-diodes, LEPD) attached to the HMD. The IR-LEDs were switched on sequentially, while at each time step, a 2D-measurement was taken by each of the sensors. Using those measurements, pose was estimated by photogrammetric techniques. Following this approach, the HiBall head tracking system was introduced in the late 1990's. Again, the signal processing was user-worn, a backpack contained the electronics. Today's realizations of the HiBall system allow position updates at a rate of 3 kHz, with sub-millimeter accuracy. During development [4], it has become a 6 DOF pose tracker.

A principal disadvantage of optical systems is the necessity of line-of-sight (LOS) for taking measurements. As occlusions between emitter-receiver-pairs occur frequently in indoor environments, a different way to determine ranges has to be found. One of the more recent approaches of a wireless acoustic tracking system is [5], where a body-worn tracking system based on time-of-flight measurements from ultrasonic transducers at 40 kHz is proposed. The acoustic signals are emitted sequentially in order to separate the sound sources within the receiver. Therefore, a sufficiently long period (e.g., 8 ms) of silence has to be taken between the subsequent emission to let the previous pulse decease. Thus, using eight sources, the latency sums up to 8 ms, yielding an update rate slightly more than 10 Hz. An inertial measurement unit is provided to propagate movements between the acoustic measurements. Distance accuracies of up to 2 mm are achieved, however, an offline calibration of the sensors needs to be performed.

Since we are interested in minimum latency, the signals should be emitted concurrently. To allow for a separation of the sources, the signals need to be distinguishable. In [6], an acoustic tracking system is proposed, which uses spread spectrum methods to identify the signal emitters. Bandspreading the signals also yields a better rejection of false measurements. In order to increase robustness to occlusions, the property of diffraction in low frequency sound is exploited, which on the other hand causes disturbing audible noise to the user.

Another approach using acoustic signals and time-of-flight measurements has been presented in [7]. Here, the sensors are room-fixed microphones, whereas the emitter (loudspeaker) is aboard a mobile robot. Emitter and sensors are synchronized by exchange of radio pulses, so time-of-flight measurements of sound signals can be performed. In normal operation, the robot initiates a position update every 10 s, so we could talk of an update rate of 0.1 Hz, which is sufficient for the task of moving a robot at a designated speed of one meter per minute. Accuracy is bounded to about 4 cm, mostly due to the low sampling rate of 10 kHz for the 5 kHz acoustic signal.

In order to obtain not only the user's position, but also the orientation, we need to extend our research to tracking of extended objects. In Constellation™ [8], a number of ultrasonic transmitters is fixed on the ceiling, while the user carries an HMD exposing a set of ultrasonic microphones. TOF-measurements are made available by synchronizing stationary and mobile units using infrared light pulses. The beacons (emitters) emit characteric codes, the receivers can select the most proximate among the whole constellation. Then, the TOFs are used in a trilateration method to define the user's 3D pose. Using the InertiaCube™, an inertial measurement system by Intersense, loss of line-of-sight is compensated.

A combination of radio and ultrasound (BLUPS, bluetooth and ultrasound) is used in [9]. Again, TOF measurements are taken, yet they are processed using multilateration techniques. In partial LOS surroundings (room size of $5 \cdot 5\ m^2$), they still achieve 6 cm accuracy in 3D positions at update rates of up to 2 Hz, using least-median-of-squares estimation.

A different way of exploiting the different propagation speeds of (ultra-)sound and radio is CRICKET, described in [10]. Distances are obtained by simultaneously and periodi-

cally emitting radio frequency (RF) and ultrasound signals, and measuring the lag in receiving times. To distinguish among the beacons, different IDs are included in the RF packets. On the other hand, the ultrasonic signal is only a narrowband pulse. The resulting positional accuracy is 10 cm, orientation accuracy adds up to 3°.

A system previously developed at the ISAS laboratory [11] uses measurements between several stationary loudspeakers and a mobile microphone array. Here, the loudspeakers emit distinguishable audible sound signals. The system anticipates some basic ideas of this present work, like bandspreading signals by multi-carrier spread spectrum (chirp signals) and concurrent emission of those generated sound sequences over a set of loudspeakers. As in the present design, the hardware is worn by the user. However, wireless operation is not possible, so the user's liberty of action is reduced by cabling.

C. Main Contributions

The design of a wireless wide-range telepresence system allows users to freely move in a remote/virtual environment generated aboard the body-worn hardware assembly. Position and viewing direction of the user are determined in real-time by an embedded acoustic tracking system and are transformed on-line according to a Motion Compression [2] algorithm. The remote scene is then output in stereo vision on a binocular HMD. The all-wireless realization of the acoustic tracking system required special considerations, as the intended time-of-flight measurements require a tight synchronization of the stationary and the body-worn subsystem. For this purpose, radio signals are exchanged periodically to correct the drifting local clocks. The system is ergonomic to the user as the acoustic signals used for distance measurements are inaudible. Also, the lightweight (about 4 kg) design and the small dimensions of about $25 \cdot 18 \cdot 11\ cm^3$ contribute to the sensation of being immersed into the remote environment.

II. Telepresence System

A feeling of immersion into distant environments can be achieved by providing a multimodal response to the user. The setup currently developed at the ISAS lab allows for visual, acoustic, and even haptic feedback to accomplish the task. Yet, the present work focuses on the visual component, being considered the primary of human senses. However, most of the methods involved can be applied to process acoustic and haptic feedback as well, such as pose estimation of a user's head or hand. The telepresence system is designed as body-centric, similar to common egocentric computer games. Position and viewing direction are determined by acoustic tracking, which will be described in detail in Section III.

For visualization, an HMD with high resolution stereo displays (2x SXGA) is employed, generating a 3D impression of remote or virtual reality environments. Image processing happens locally on the body-worn computer system, featuring sufficient computation power to achieve the required frame rates, as described in Section IV.

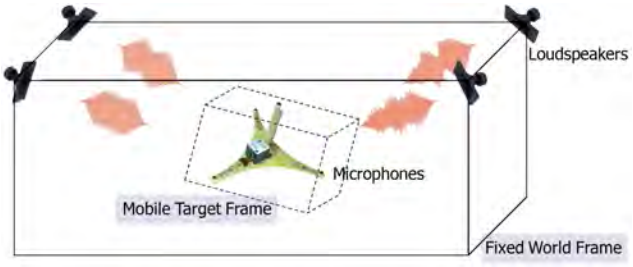

Fig. 3. Pose estimation (position and orientation) in 3D is performed using distances measured between stationary beacons (loudspeakers) and mobile sensors (microphones), given known geometry within their own coordinate system.

One issue among common telepresence systems is the limited operating range. Due to the reduced user space compared to the distant or virtual space, a simple solution is to introduce a scaling factor for all movements. As this seems quite unsatisfying, most approaches avoid to take the user's movements as input and thus, rely on well-established input devices like joystick or space-mouse.

A more sophisticated way of dealing with limited user space is the use of Motion Compression (MC) [2]. As stated so far, wide range telepresence requires a way to extend the limited local surroundings towards large-scale or even virtually infinite environments, which can be achieved by performing a suitable transformation. MC consists of three modules. In the first module (path prediction), the possible path of the user in the target environment is predicted (target path). This predicted target path has then to be mapped in the user environment (user path) by using the second module called path transformation. This module provides a nonlinear mapping between the target path and the user path, where the path length and the current turning angles are preserved. In the last module (user guidance), the user is guided on the calculated user path. To mediate between user and target space, a server program runs either on the body-worn computer system, or on a stationary computer system. It provides all functionality to exchange user/target positions. More details on Motion Compression can be found in [2].

III. Localization and Tracking

Extended range telepresence in the above stated sense requires on-line localization and tracking of the user's pose (position and orientation). In order to maintain a relation between the user and the globally fixed coordinate system, an absolute positioning method is applied, based on distance measurements between ceiling-mounted loudspeakers and body-worn microphones, as illustrated in Fig. 3. Moreover, the tracking can be supported by dead-reckoning methods, using various sources like gyroscope and accelerometer measurements. The combination of both principles yields a higher update rate and increased robustness, as the acoustic tracking suffers from failures due to massive occlusions or disturbing influences from different sources.

A. Wireless Acoustic Tracking

In the design presented, acoustic tracking is performed by measuring distances between a number of fixed loudspeakers and a user-worn microphone array. Several basic principles of operation are available, depending on the actual setup and other requirements, like the expected precision or whether emitters and receivers can be synchronized. If synchronization is not available, multilateration techniques can be applied, operating mostly on time differences of arrival (TDOAs). However, in our case, sending and receiving units are synchronized by exchanging radio signals periodically. By knowing the emission time, the receiving unit can measure the time of flight (TOF) of acoustic signals, corresponding directly to the desired ranges between loudspeakers and microphones, given the propagation speed (speed of sound in air). All loudspeakers involved concurrently emit distinguishable sound signals of a certain length. Microphones listen for these signals, and the underlying signal processing outputs the time elapsed until the first appearance of those signals within the receiving buffer by simply subtracting the known emitting timestamp from the receiving timestamp. This process repeats periodically to the update rate specified.

To allow for a distinction between different signal sources on the medium, channel encoding is necessary. Again, several possibilities exist that deal with sharing the common channel, adapted from communications engineering. Spreading the spectrum yields increased robustness towards narrowband disturbance. Another prerequisite is the property of inaudibility, which can be achieved by conditioning the signals to fit the spectrum above 20 kHz.

1) Bandspreading: Sending sequences are generated according to their correlation properties. Desirable sequences exhibit a narrow auto correlation peak, whereas cross correlation with all other sequences results in low values. A choice was to be made among various methods of Code Division Multiple Access (CDMA). Several pseudo noise codes have been tested to modulate a narrowband carrier, incorporating Direct Sequence Spread Spectrum (DSSS) signals. To increase diversity, especially regarding the bandpass-filtering ahead (for inaudibility), Multi-Carrier Spread Spectrum (MC-SS) has been exercised. After all, even uniformly distributed noise was adequate to constitute the unique signal sequences. Unlike telecommunication systems, no information needs to be transmitted over the spread spectrum carrier, so the whole signal can be stored into lookup-tables on the participating units.

2) Inaudible Signal Sequences: A favorable aspect of the acousting tracking system is the property of inaudibility. While the ultrasonic spectrum disqualifies itself for exhibiting high directivity and strong absorption through the air, the frequency band slightly above the audible range bears sufficiently low attenuation and almost omnidirectional emitting properties. Forming the spectrum to fit the inaudible range above 20 kHz involves bandpass-filtering. A high-order FIR-filter is applied to the bandspreaded signals to sufficiently suppress everything outside the cut-off frequencies. Unfortunately, some of the above gained properties (diversity and robustness) are decreased, due to narrowing the effective bandwidth. Yet, the transformation into bandlimited noise is necessary to improve ergonomics. Experiments show adequate performance even with a residual bandwidth of 5 kHz.

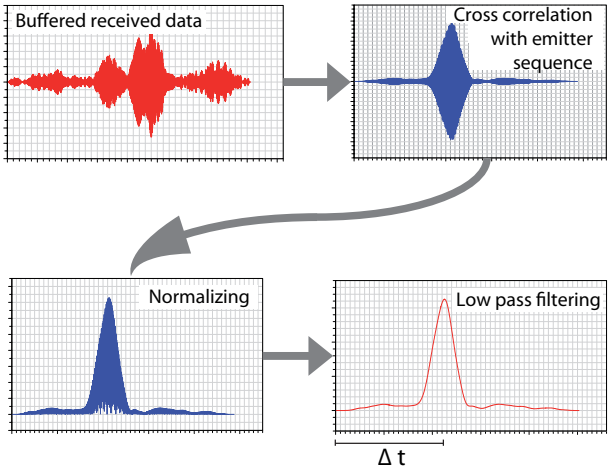

Fig. 4. Each microphone has a dedicated receiving buffer to store a sufficiently long (speed of sound times the update rate divided by the maximum credible distance) input sequence. Within the matched filter, the sequence is then convolved with the reverse of all sending signals involved, yielding a significant peak on consensus. The time elapsed between sending timestamp and peak detected adds up to the time of flight.

3) Distance Measurements: Retrieving distances between all pairs of m loudspeakers and n microphones is the basis for later pose estimation of the extended object, i.e., the user's head. Each loudspeaker emits a characteristic sequence of a certain length, e.g., 192 samples, which equals 2 ms at 96 kHz sampling rate. Due to the simultaneous emission of all loudspeakers involved, no additional time lag is introduced between measurements. The receiving unit stores the input of each microphone in a separate buffer, containing the overlaid versions of all emitted signal sequences. It then evaluates similarity with each of the characteristic sending signals, e.g., by using a matched filter. Eventually, low pass filtering is applied to eliminate ambiguity due to ripple on the correlation function. The peak correlation corresponds to the zero-shifted original sending signal above the recorded input sequence, its timestamp reveils the time of flight, since the beginning of recording conincides with the signal's emission time. These steps are illustrated in Fig. 4. Finally, after every update cycle, a set of $m \cdot n$ distances is determined, which are then used for pose estimation.

4) Synchronization: Time-of-flight measurements require sound signal emitters and receiving units to be synchronized. The participating timebases are aligned to a master timebase, which in this case is the signal generator unit. It periodically emits a radio pulse which is then received by the mobile tracking unit. Pulse latency stabilizes to a certain value (expectancy) with time, while maintaining a consider-

able variance. Wiener filtering is applied to compensate the variability. Between the pulse intervals, the clocks involved continue running autonomously at the same frequency. If synchronization pulses are lost, the clocks remain untouched, keeping up a semi-synchronized state, with increasing mutual offset due to drifting crystal oscillators. After an appropriate radio pulse is detected, the offsets are again reset to zero. A schematic diagram of the actual synchronization procedure is given in Fig. 5.

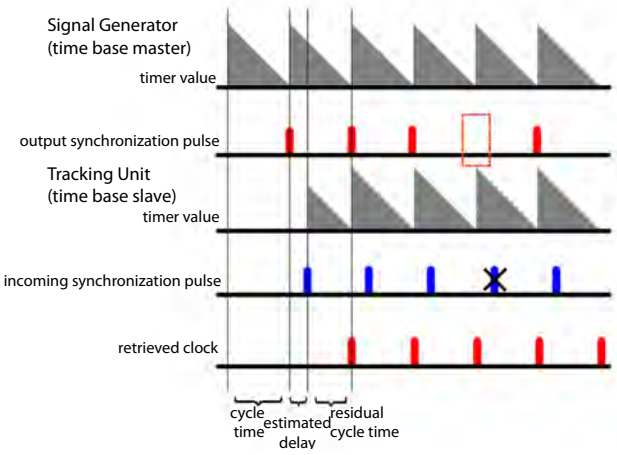

Fig. 5. Scheme of synchronization - The incoming synchronization pulse sets the receiving timer value according to the difference of full cycle time and estimated system-dependent delay. The original timestamp is retrieved within the Tracking Unit as the cyclic countdown timer reaches zero (reset). The method is tolerant to missing sending pulses as they are compensated by the underlying timer activity.

5) Common error sources: Acoustic tracking is subject to several types of disturbances, namely reflections, scattering, and directivity. The former effects are mainly caused by even surfaces and furniture and result in multipath propagation. The latter is a frequency dependent effect, which narrows the angle of aperture with increasing frequency. Furthermore, reverberant sound can corrupt the useful signal. On the other hand, the technology is widely robust to occlusion and shadowing, which is a remarkable advantage over optical systems, where line-of-sight is necessary.

B. Pose estimation

Given range measurements from m loudspeakers (world fixed) to n microphones (body fixed), pose estimation of an extended object can be performed following different approaches, such as closed-form solutions [12], gradient descent algorithms [13], or state estimators [14]. For an unambiguous pose specification in 3D, a configuration of at least 3 loudspeakers and 3 microphones is necessary. In this work, the closed-form solution presented in [12] is applied, which relies on decoupling computation of position and orientation. By knowledge of the emitter and receiver geometries, and given all $m \cdot n$ ranges, the target points can be expressed with respect to the world coordinate system, and vice versa. Thus, translation and rotation between the target and the world frame can be determined using a weighted least squares estimate. This method has been applied in this work for the reason of simplicity. However, the use of recursive stochastic state estimators such as Extended Kalman Filter or Unscented Kalman Filter could be considered to improve accuracy and efficiency, as described lately in [14].

Fig. 6. Signal Generator Unit used for sound signal generation and amplification as well as radio pulse emission. The orange circle highlights the 2.4 GHz radio module.

IV. System Design

The system hardware is composed of heterogeneous modular components, which are interconnected in a suitable manner to exhibit a fully integrated, high performance, yet affordable mobile telepresence system. Off-the-shelf hardware was chosen where appropriate, e.g., for graphics production (scene rendering, visualization) and Motion Compression (transformations and database management). However, for the challenging tasks of audio signal processing and motion tracking, dedicated embedded hardware was deployed to ensure excellent signal treatment even at high sampling rates and multi-channel operation. The overall design is tailored to meet low power requirements, which allows battery-supplied operation of over one and a half hour on a single battery pack. Ergonomic demands are satisfied by the fully wireless, wearable assembly featuring small dimensions and lightweight design. Communication between stationary and mobile modules is performed using different wireless technologies, namely WLAN and XBEE (IEEE 802.15.4), as illustrated in Fig. 8 on the next page.

The system consists of a stationary and a mobile subsystem. The stationary unit accounts for audio signal generation and amplification, as illustrated in Fig. 6. It outputs these generated signals over a set of ceiling-mounted loudspeakers. The mobile subsystem is typically integrated into a backpack, which is worn by the user (See Fig. 7 on the next page). It incorporates two principal sections: An embedded system supported by a digital signal processor (referred to as *Embedded Tracking Unit*) and a regular computer system. To complete the setup needed for telepresent interaction, an HMD is connected to the system to visualize the remote/virtual scene to the user. It has a microphone array attached which interfaces to the Embedded

Fig. 7. Mobile Telepresence Unit consisting of an Embedded Tracking Unit and a regular computer. Distance measurements and pose estimation are performed by DSP on the dedicated hardware, whereas scene rendering and gaming engine are located in the computer. Communication between the boards is done via USB, remote control over WLAN. Like in Fig. 6 on the preceding page, an XBEE radio module is furnished to allow for tight synchronization with the signal emitters.

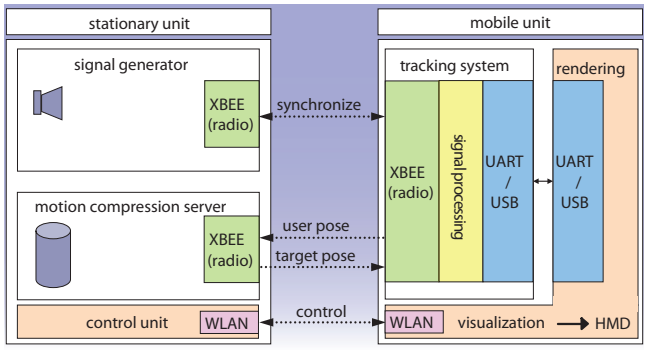

Fig. 8. Communication between distributed components of the telepresence system. The stationary unit consists of the sound signal generator and a server to transform posture information from user environment into target environment and vice versa. The Mobile Telepresence Unit includes sound signal receiving and processing as well as generating the video data to be passed to the HMD.

Tracking Unit to receive the signals emitted by the fixed loudspeaker array for acoustic localization. The embedded components are connected to the computer system by USB for control and data exchange.

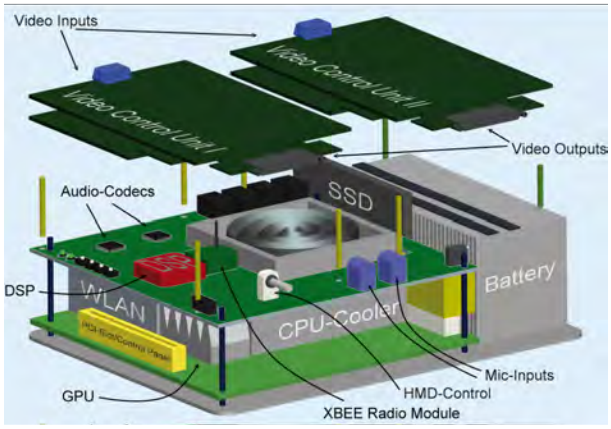

Fig. 9. 3D model of the Mobile Telepresence Unit. The video control units can be detached if not needed (only required for NVIS HMDs).

A. Signal Generator Unit

The purpose of the signal generator unit is to define and periodically emit unique and distinguishable audio signals. It comprises multiple channels to independently address up to eight loudspeakers.

At the heart of the signal generator stands a digital signal processor (DSP) Blackfin™ BF533 by Analog Devices clocked at 675 MHz which seamlessly feeds two integrated digital amplifiers of type TAS5704 by Texas Instruments via the I^2S audio interface. Each of them handles four channels of audio at 48 kHz sampling rate while providing eightfold oversampling. The output power of 10 W per channel is more than sufficient even for larger rooms (with tweeters operating beyond 20 kHz). A radio module (Maxstream XBEE) is attached to provide a periodic time pulse that other units can synchronize with (in particular the Embedded Tracking Unit).

B. Mobile Telepresence Unit

The Mobile Telepresence Unit is a lightweight assembly of embedded and standard desktop components, forming a complete user-wearable telepresence system. The Embedded Tracking Unit is based on the DSP Blackfin™ BF533. It accounts for audio signal recording and processing, pose estimation and tracking. The underlying standard computer hardware is concerned with rendering scenarios to be visualized on the user's HMD as well as holding scenic information and transforming positions/paths according to the Motion Compression algorithm. A detailed view of the assembly is shown in Fig. 9.

1) Computer System: Following the guidelines of energy-conscious design, a Mini-ITX mainboard with integrated graphics adapter and power saving AMD 240e CPU were deployed, completed by a solid state disk and an ultra-efficient power supply. Yet, it is capable of providing dual head graphics output to support the binocular stereo displays of the high resolution HMD (NVIS nVisor SX60). The operating system is a custom Ubuntu Linux which can be controlled remotely via onboard WLAN (IEEE 802.11n).

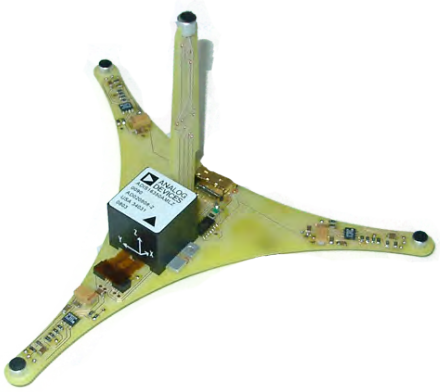

Fig. 10. The user-worn microphone array exhibits four microphones with onboard preamplification. The tetrahedral design is optimal with respect to 3D pose determination. The latest assembly has an inertial measurement unit supplied.

2) Embedded Tracking Unit: Consisting primarily of the above mentioned DSP and two Audio Codecs of type AD1938 by Analog Devices, the main activity of the Embedded Tracking Unit is to process and evaluate sound measurements buffered from the microphone arrays. The Codecs exhibit a sampling rate of up to 192 kHz, yet we use only 96 kHz in order to save buffer memory and computation power. A total of eight channels can be handled so far, extensible to sixteen using a stackable extension board. Same as in the Signal Generator, a radio module (Maxstream XBEE) operating in the 2.4 GHz-band is used to receive synchronizing pulses emitted by the Signal Generator Unit. It conforms to IEEE 802.15.4 standard and serves also as a transmitter of user pose values to surrounding receivers (e.g., the haptic display subsystem).

3) Microphone Arrays: After a generation of planar four-channel microphone arrays has passed, a new design of a tetrahedral microphone array has been developed (see Fig. 10) to improve accuracy in orientation. To avoid losses in signal quality, preamplifiers are directly integrated on the PCB in immediate vicinity of the microphones, featuring low-noise operational amplifier design. The array is attached on top of the HMD to receive surrounding loudspeaker signals with a minimum of extrinsic disturbances (like self-occlusion).

The latest progress was to furnish an inertial measurement unit (IMU) of type ADIS16350 by Analog Devices to increase robustness and update rate compared to the acoustics-only tracking. Experiments shown so far do not yet comprise the fusion with these accelerometer and gyroscope outputs, but they will be taken into account shortly.

4) Power Supply: A battery pack has been specifically designed to meet the required voltage and capacity under the given spatial conditions. Four serially connected blocks of eight parallel lithium-polymer-cells each are fit to the bottom of the integrated telepresence system, forming a 32-cell 14.8 V, 8 Ah battery pack with variable charging schemes (serially balanced or quick parallel charge). The total power consumption of the Mobile Telepresence Unit in full operation mode adds up to 5 A, which allows for about 1.5 hour runs on a single battery charge.

V. Experimental Results

A. Synchronization

The core clocks of the digital amplifier unit (clock master) and the embedded tracking unit (clock slave) are synchronized by periodically emitted radio signals. If the synchronization pulse is stalled, the drift increases without limit with a mean value of about 0.08 ppm/s at room temperature. This can be observed in Fig. 11, where the measured distances increase over time although a static scene was recorded. The synchronization signal is sent every 100 ms by the clock master, which for simplicity coincides with the update rate adjusted to 10 Hz. The latency of the pulse arrival is subject to variability with a standard deviation of 66 μs as shown in Fig. 12, which effects a ripple in distances of about 2.2 cm. To prevent fluctuations in distance measurements, filtering needs to be applied over the incoming synchronization pulses.

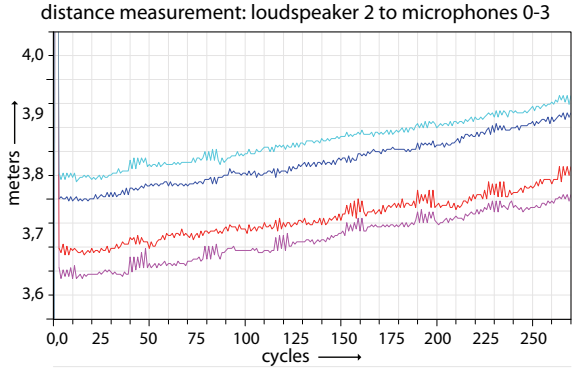

Fig. 11. Impact of drift between signal generator (clock master) and mobile tracking unit (clock slave) on the measured ranges between one speaker and four microphones.

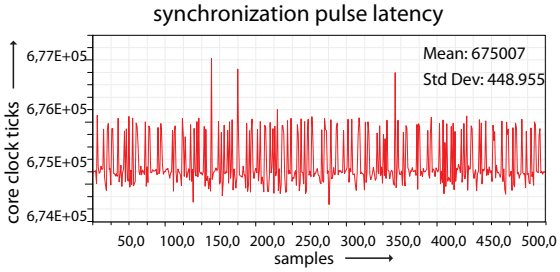

Fig. 12. Ripple on the incoming synchronization pulse: Timing is derived from the core clock running at 675 MHz. Mean value is 675000, corresponding to 1/1000 of a regular timer interval or 100 ms. Standard deviation in receiving pulse timestamps is 449, which equals 66 μs at 10Hz update rate (449 samples/(675000 $\cdot$ update rate) $=$ 66 μs). Multiplied by the speed of sound, we get a variability of 2.2 cm which adds up onto the distance measurement error.

B. Position measurements

In Fig. 13 on the following page, a fixed scene is shown to illustrate the average deviation of the mean

value in the x-y-plane. Mostly, the standard deviation is around 0.5 cm. Considering the smallest step that is possible in theory (propagation speed times the samping period, $344.8\frac{\text{m}}{\text{s}} \cdot \frac{1}{96000}\text{s} \approx 0.34$ cm), this is a satisfiable value. Yet, the error in orientation sums up to 0.04 rad, which equals $2.2°$, due to the short baselength of about 15 cm between the microphones. Sensitivity analyses showed, that 1 cm error in distances leads to $4°$ error in orientation, a linear dependence could be noticed between distance and orientation error. In

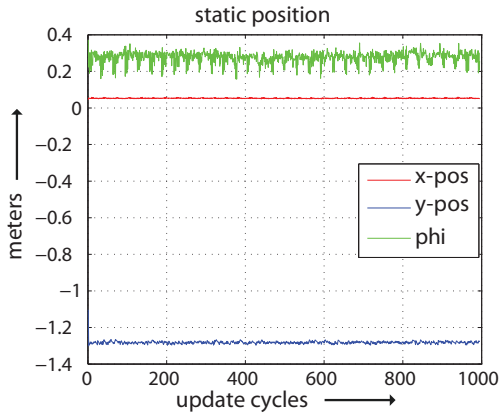

Fig. 13. Static position measurement. With the microphones fixed, positions were recorded over 1000 samples (which equals 100 seconds, at 10 Hz update rate). The positional standard deviation is well under 1 cm. The value for orientation ϕ is given in radians.

Fig. 14, a scatter plot is shown, corresponding to the static scene measurements of Fig. 13. Next, in Fig. 16, a straight

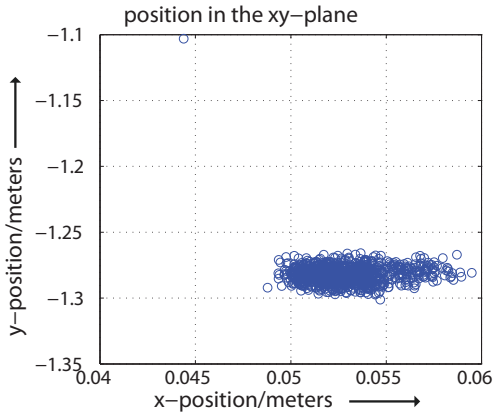

Fig. 14. The corresponding scatter plot to the data in Fig. 13 shows the positions in the xy-plane.

path has been recorded. The target was mounted on a carriage moving on a rail at constant speed, as shown in Fig. 15. This experiment shows that the measurements taken are very close to the true trajectory. Yet, several outliers occurred, due to occlusions, reflections on even surfaces, and high background noise. In Fig. 17, a scatter plot is added, where the resulting planar trajectory can be seen. As these experiments show, efficient outlier detection and handling is necessary, in order to get a smooth trajectory. Under ideal surroundings, with no

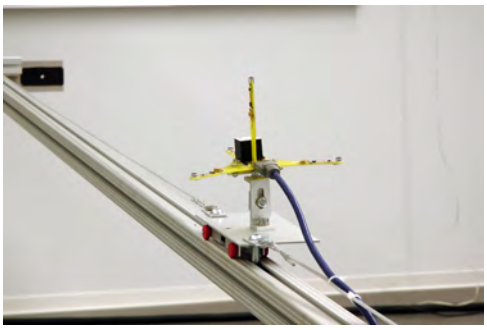

Fig. 15. Experimental setup for straight trajectories. The microphone array is mounted on a carriage driven by a DC-motor.

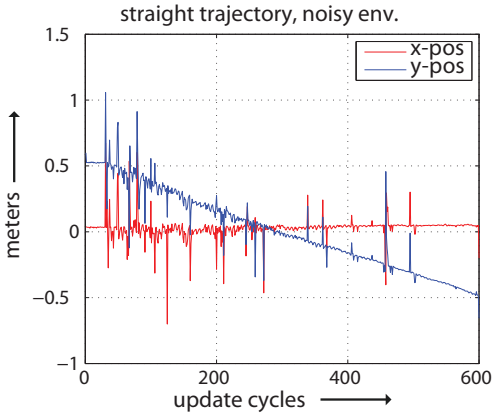

Fig. 16. Straight trajectory, generated by a carriage on a track moving at constant speed for 60 seconds in noisy environment.

background noise whatsoever, trajectories can be recorded that bear no outliers at all, as shown in Fig. 18 on the following page and Fig. 19 on the next page.

VI. Conclusions

Through the wireless acoustic tracking system presented in this work, a significant step towards natural immersion into remote/virtual environments has been taken. At an update

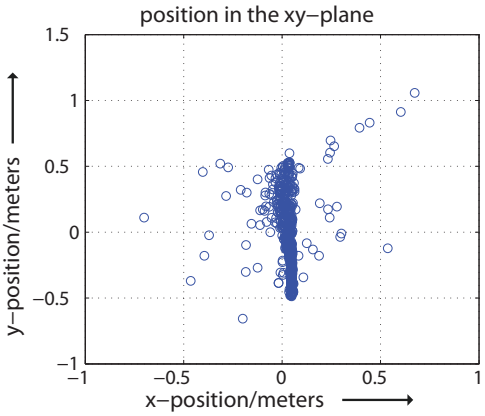

Fig. 17. The corresponding scatter plot to the above Figure shows the positions in the xy-plane.

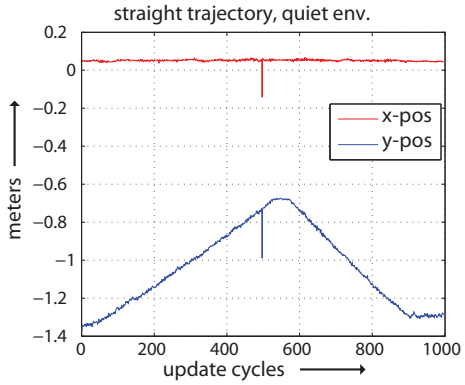

Fig. 18. Trajectory along a rail aligned with the y-axis. The carriage moved back and forth. Due to the quiet environment, only one outlier occurred during the 100 s run.

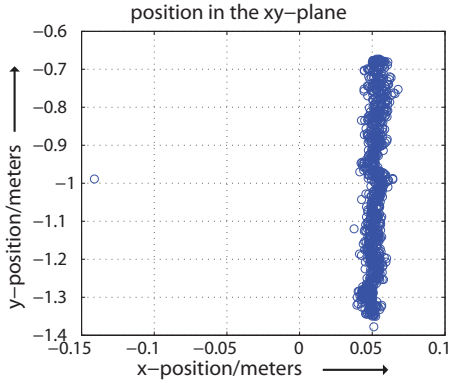

Fig. 19. The corresponding scatter plot to the above figure shows the trajectory in the x-y-plane.

rate of currently 10 Hz, a user's pose can be tracked with a precision of better than one centimeter in position and 2.2° in orientation, depending on the base length (distance between microphones) of the microphone array.

In combination with the Motion Compression algorithm presented in [2], a user can explore a wide range scenario by naturally walking in his local environment. All equipment needed is integrated in a lightweight (approx. 4 kg) backpack running on battery. The number of trackable objects is not limited by the technology, i.e., not only the user's head, but also the hands can be tracked. Tracking of multiple users is feasible, however, Motion Compression must be deactivated, to avoid collisions between participants walking in the same room.

As computational effort increases linearly with the number of objects, additional DSPs can be stacked via an extension slot. The modular system design allows for a variety of extensions, like a haptic display (as described in [1]) or teleoperators (like the omnibase [13]). Operation is ergonomic due to the absence of disturbing cables as found in prior setups. By performing acoustic distance measurements in the inaudible domain above 20 kHz, the user is not distracted by noise and can therefore better react to ambient sound originated in the target environment.

For future work, the tracking system will undergo several improvements involving sensor data fusion with inertial measurements, as already prepared in hardware (Fig. 10 on page 7). The intended increase in update rate and robustness will improve the system's behavior concerning fast head movements (changes of viewing direction) or failing acoustic measurements caused by occlusion or other unwanted effects. Furthermore, improvements can be made on the algorithmic side to deal with multipath propagation, which is a permanent nuisance in indoor environments.

References

[1] P. Rößler, T. Armstrong, O. Hessel, M. Mende, and U. D. Hanebeck, "A Novel Haptic Interface for Free Locomotion in Extended Range Telepresence Scenarios," in *Proceedings of the 3rd International Conference on Informatics in Control, Automation and Robotics (ICINCO 2006)*, Setúbal, Portugal, Aug. 2006, pp. 148–153.

[2] P. Rößler and U. D. Hanebeck, "Simultaneous Motion Compression for Multi–User Extended Range Telepresence," in *Proceedings of the 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2006)*, Beijing, China, Oct. 2006, pp. 5189–5194.

[3] M. Ward, R. Azuma, R. Bennett, S. Gottschalk, and H. Fuchs, "A demonstrated optical tracker with scalable work area for head-mounted display systems," in *Proceedings of the 1992 Symposium on Interactive 3D Graphics*, 1992, pp. 43–52.

[4] G. Welch and G. Bishop, "Scaat: incremental tracking with incomplete information," in *SIGGRAPH '97: Proceedings of the 24th annual conference on Computer graphics and interactive techniques*. New York, NY, USA: ACM, 1997.

[5] D. Vlasic, R. Adelsberger, G. Vannucci, J. Barnwell, M. Gross, W. Matusik, and J. Popović, "Practical motion capture in everyday surroundings," in *SIGGRAPH '07: Proceedings of the 30th annual conference on Computer graphics and interactive techniques*. New York, NY, USA: ACM, 2007.

[6] N. M. Vallidis, "WHISPER: A spread spectrum approach to occlusion in acoustic tracking," Ph.D. dissertation, University of North Carolina, Chapel Hill, 2002.

[7] W. Daniel, S. Ralf, and S. Matthias, "Low-cost sonic-based indoor localization for mobile robots," in *Proceedings of the 3rd Workshop on Positioning, Navigation and Communication (WPNC'06)*, March 2006.

[8] E. Foxlin, M. Harrington, and G. Pfeifer, "Constellation: A wide-range wireless motiontracking system for augmented reality and virtual set applications," in *SIGGRAPH '98: Proceedings of the 25th annual conference on Computer graphics and interactive techniques*. New York, NY, USA: ACM, 1998, pp. 371–378.

[9] C. R. and Marco A. and Guerrero J. J. and Falcó J., "Robust estimator for non-line-of-sight error mitigation in indoor localization," vol. 2006. New York, NY, United States: Hindawi Publishing Corp., January 2006, pp. 156–156.

[10] N. B. Priyantha, "The cricket indoor location system," Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, MA, USA, 2005.

[11] F. Beutler and U. D. Hanebeck, "The Probabilistic Instantaneous Matching Algorithm," in *Proceedings of the 2006 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI 2006)*, Heidelberg, Germany, Sep. 2006, pp. 311–316.

[12] ——, "Closed-Form Range-Based Posture Estimation Based on Decoupling Translation and Orientation," in *Proceedings of the 2005 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2005)*, vol. 4, Philadelphia, Pennsylvania, Mar. 2005, pp. 989–992.

[13] P. Rößler, F. Beutler, U. D. Hanebeck, and N. Nitzsche, "Motion Compression Applied to Guidance of a Mobile Teleoperator," in *Proceedings of the 2005 IEEE International Conference on Intelligent Robots and Systems (IROS 2005)*, Edmonton, Canada, Aug. 2005, pp. 2495–2500.

[14] F. Beutler, M. F. Huber, and U. D. Hanebeck, "Semi-Analytic Stochastic Linearization for Range-Based Pose Tracking," in *Proceedings of the 2010 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI 2010)*, Salt Lake City, Utah, Sep. 2010.