AUDIO AUGMENTED REALITY IN TELECOMMUNICATION THROUGH VIRTUAL AUDITORY DISPLAY

Hannes Gamper and Tapio Lokki

Aalto University School of Science and Technology
Department of Media Technology
P.O.Box 15400, FIN-00076 Aalto, Finland
[Hannes.Gamper,ktlokki]@tml.hut.fi

ABSTRACT

Audio communication in its most natural form, the face-to-face conversation, is binaural. Current telecommunication systems often provide only monaural audio, stripping it of spatial cues and thus deteriorating listening comfort and speech intelligibility. In this work, the application of binaural audio in telecommunication through audio augmented reality (AAR) is presented. AAR aims at augmenting auditory perception by embedding spatialised virtual audio content. Used in a telecommunication system, AAR enhances intelligibility and the sense of presence of the user. As a sample use case of AAR, a teleconference scenario is devised. The conference is recorded through a headset with integrated microphones, worn by one of the conference participants. Algorithms are presented to compensate for head movements and restore the spatial cues that encode the perceived directions of the conferees. To analyse the performance of the AAR system, a user study was conducted. Processing the binaural recording with the proposed algorithms places the virtual speakers at fixed directions. This improved the ability of test subjects to segregate the speakers significantly compared to an unprocessed recording. The proposed AAR system outperforms conventional telecommunication systems in terms of the speaker segregation by supporting spatial separation of binaurally recorded speakers.

1. INTRODUCTION

Audio augmented reality (AAR) aims at enhancing auditory perception through virtual audio content. Virtual auditory display (VAD) is used to present the content to the AAR user as an overlay of the acoustic environment. This principle is applicable to telecommunication systems as a new interface paradigm. Conventional telecommunication systems often provide the user only with a monaural audio stream, played back via a headset or a hand-held device. The term "monaural" refers to the fact that only one ear is necessary to interpret the auditory cues contained in the audio stream. However, face-to-face communication, which is considered the "gold standard" of communication [1, 2], is inherently binaural. In a face-to-face conversation, a listener is able to segregate multiple talkers based on their position, a phenomenon referred to as the "cocktail party effect" [3]. Monaural audio employed in conventional telecommunication systems, such as mobile phones and voice-over-IP (VoIP) softwares, does not support interaural cues and hence deteriorates the communication performance compared to face-to-face communication [4, 5].

AAR helps overcoming these limitations by embedding spatialised virtual audio into the auditory perception through VAD. The "cocktail party" principle holds also for a multi-party telecommunication scenario. Using VAD to separate the speech signals of participants spatially improves the listening comfort and intelligibility [6, 7]. In contrast to the sense of vision, auditory perception is not limited to a "field of view". The participants of a teleconference can thus be distributed all around the user, regardless of the orientation of the user. By registering the virtual speakers with the environment, the user can turn towards a conferee the same way as in a face-to-face conversation.

A major challenge in telecommunication lies in the physical distance itself, which puts limits to the naturalness of interaction with a remote end. Communication over distance suffers from a lack of "social presence", compared to face-to-face communication [8]. Through spatial audio, an AAR telecommunication system improves the sense of "presence" [9, 10] and "immersion" [6]. In this work, algorithms are presented to process binaural recordings and embed them into the auditory perception of a user. This serves as a proof-of-concept for employing AAR in a telecommunication scenario. A user study is conducted to analyse the ability of users to localise and segregate remote speakers with the proposed system.

2. EXPERIMENTAL SETUP

The basic principle of AAR is to augment, rather than replace, reality. Therefore, the transducer setup for AAR needs to be acoustically transparent to ensure unaltered perception of the real environment. If a headset is used for the reproduction of virtual audio content, acoustical transparency is achieved by capturing the real-world sounds at the ears of the user and playing them back through the earphones. Mixing these captured real-world sounds with virtual sounds is the basic working principle of "mic-through augmented reality" [11], which refers to the fact that the real world is perceived through microphones.

In this work, the MARA headset, introduced by Härmä et al. [12], is used. It consists of a pair of insert-earphones with integrated miniature microphones. Insert-earphones provide the advantage of leaving the pinnae of the listener uncovered, thus preserving the pinna cues, which are important for the localisation of real-world sounds [13]. Inserting the earphones into the ear canal minimises effects of the transmission paths from the earphone to the ear drum.

The MARA microphone signals provide a realistic representation of the acoustic environment [12]. In the proposed AAR telecommunication system, these signals are transmitted to the other end of the communication chain, where they are embedded

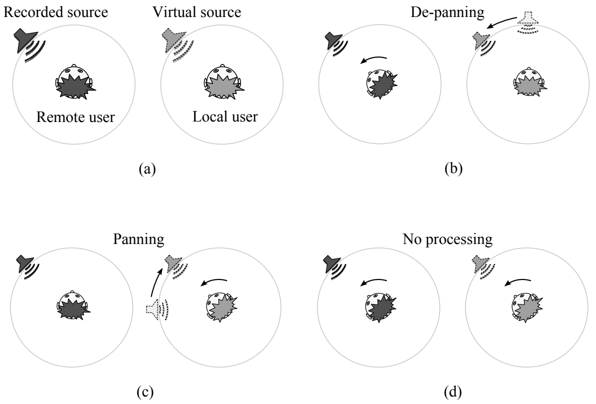

Figure 1: *De-panning and panning of binaural recordings.* (a) The source recorded at the remote end is perceived at the local end as a virtual source at the same direction. (b) De-panning is applied to compensate for head movements of the remote user. (c) Panning compensates for head movements of the local user, to register the virtual sources with the environment. (d) No processing is necessary if the head orientation of both users is the same, e.g. if both are facing the source.

into the auditory perception of a listener. The listener thus perceives the remote acoustic environment through the ears of a remote user as an overlay of the own acoustic environment. As a test case for the proposed system, a teleconference between the two ends is devised. The conference is held at the remote end, captured through a MARA headset worn by one of the participants (hereafter referred to as the “remote user”). The recording is transmitted to the user at the other end, i.e. the “local user” (cf. fig. 1a).

If both the remote and the local user keep their heads still, the local user perceives each remote conferee at a distinct direction. If the remote user rotates the head, however, the spatial cues of the virtual speakers change, which affects the perceived directions. The resulting lack of distinct spatial cues deteriorates the listening comfort and the speaker segregation ability of the local user.

To restore the spatial cues contained in the binaural recording, the head rotation at the remote end has to be compensated for. After compensation, the virtual speakers have a fixed direction relative to the local user, regardless of the head orientation of the remote user recording the conference. In a telecommunication scenario it might be desirable to employ virtual auditory display (VAD) registered with the environment for each virtual speakers. This allows the user for example to turn the head to look at a remote talker, which is a natural behaviour in face-to-face communication. In the following section, algorithms are presented to compensate for head movements of the remote user and register binaurally recorded speakers with the environment of the local user.

2.1. Compensation for head movements

To compensate for head movements during the recording, the binaural recording has to be processed such as to reposition the virtual speakers. Two measures need to be known for this “de-panning” process: The head orientation of the remote user and the position of the sources. For simplicity, the positions of the conference participants are assumed to be fixed. The head orientation of the remote user is tracked.

The aim of the de-panning process is to remove the alterations of the spatial cues introduced by head movement. These alterations occur both in the time domain and in the spectral domain. The following sections propose methods to remove or minimise these alterations.

2.1.1. Restoring interaural differences

The most important alteration of spatial cues caused by head movement during a binaural recording is a change in the time of arrival of the signal at both ears. This results in an altered interaural time difference (ITD). If, for simplicity, the ITD is assumed to be frequency-independent (cf. Wightman and Kistler [14]), it can be represented by a delay of one ear input signal with respect to the other. The head movement affects this delay. Thus, by delaying the binaural signals appropriately in the de-panning process, the ITD of a virtual speaker can be restored. $TD(\alpha)$ is the frequency-independent delay as a function of the angle of incidence (after Rocchesso [15]):

$$TD(\alpha) = \begin{cases} \frac{f_s}{\omega_0} \cdot [1 - \cos(\alpha)] & \text{if } |\alpha| < \frac{\pi}{2}, \\ \frac{f_s}{\omega_0} \cdot \left[|\alpha| - \frac{\pi}{2} + 1\right] & \text{else.} \end{cases} \tag{1}$$

with

$$\omega_0 = \frac{c}{r}, \tag{2}$$

where α denotes the angle of incidence, r the head radius (i.e. half the distance between the two ear entrances) and c the speed of sound. By delaying each signal with an appropriate $TD_{correction}$ factor, the influence of head rotation on the ITD is eliminated (cf. fig. 2, top graph).

In the frequency domain, head rotation affects the head-related transfer function (HRTF). Pinna and shoulder reflections introduce azimuth-dependent peaks and notches in the HRTF. In addition to direct sound, room reflections and diffuse sound make a compensation of HRTF alterations caused by head movements rather complex and impractical. As a simple approximation, the impact of head rotation on the spectrum of the ear input signals can be described in terms of variations of the interaural level difference (ILD). Rocchesso proposes a simple model for the head shadow effect as a one-pole/one-zero shelving filter [15]. From this model, an azimuth-dependent gain correction $LD_{correction}$ is derived to compensate for the ILD alterations caused by head movement. It is given by

$$LD_{correction}(\alpha) = 1.05 + 0.95\cos(\frac{6}{5}\alpha). \tag{3}$$

To achieve the gain correction, a high shelving filter is applied to each channel, with transfer function

$$H_{LD}(z) = k \cdot \frac{1 - qz^{-1}}{1 - pz^{-1}}, \tag{4}$$

where k is the filter gain, p is the pole and q is the zero of the filter. The pole of the filter is fixed [15]:

$$p_{hs} = \frac{1 - \frac{\omega_0}{f_s}}{1 + \frac{\omega_0}{f_s}}. \tag{5}$$

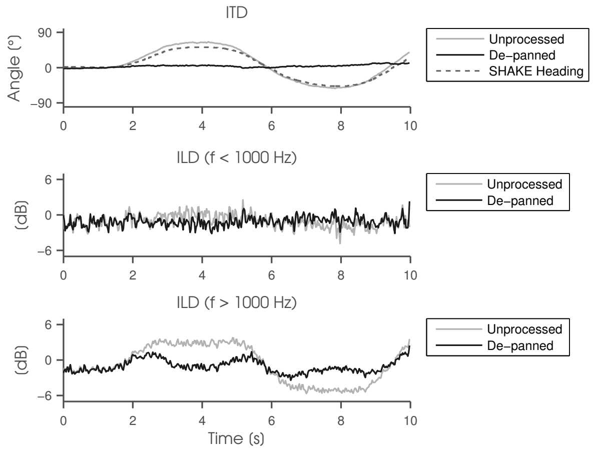

Figure 2: *Restoring interaural differences of a binaural recording.* Head movement (dashed line) causes ITD and ILD variation. The ITD is calculated as the maximum of the interaural cross correlation (IACC). Above 1000 Hz, head rotation causes ILD variations.

with f_s denoting the sampling frequency.
The gain k and zero q of the filter are chosen to meet the following two criteria: At low frequencies, the impact of head shadowing is negligible, thus the filter has a DC gain of unity. At high frequencies, the impact of head rotation on the head shadowing needs to be compensated for. At the Nyquist limit, the filter gain equals the value given by $LD_{correction}$:

$$H_{LD}(z)|_{z=1} = k \cdot \frac{1-q}{1-p} \overset{!}{=} 1, \qquad (6)$$

$$H_{LD}(z)|_{z=-1} = k \cdot \frac{1+q}{1+p} \overset{!}{=} LD_{correction}. \qquad (7)$$

Solving eq. (6) and eq. (7) for q and k yields

$$q = \frac{\phi - 1}{\phi + 1} \qquad (8)$$

for the filter zero q with

$$\phi = LD_{correction} \frac{1+p}{1-p} \qquad (9)$$

and

$$k = \frac{1-p}{1-q} \qquad (10)$$

for the filter gain k. Applying a gain factor $LD_{correction}$ to each channel via a separate shelving filter reduces the impact of head rotation on the ILD of the recorded binaural signals.

The effect of the de-panning algorithm on a binaural recording is shown in Fig. 2. The input signal is white noise, played back from a loudspeaker in a small office environment and recorded via a MARA headset. During the recording, the head orientation changes by ±60°. The resulting ITD variations are compensated for through appropriate delays, defined by $TD_{correction}$, applied to both channels. For frequencies below 1000 Hz, the ILD change due to head shadowing is negligible (cf. Fig. 2, middle graph). Above 1000 Hz, the de-panning algorithm compensates for the head shadowing effect (cf. Fig. 2, bottom graph).

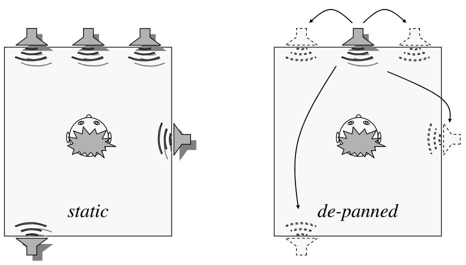

Figure 3: *Recording conditions for speaker localisation task.* For the *static* recording, the speech sample is played back from one of five different loudspeakers. For the *de-panned* recording, only one loudspeaker is used. The spatial separation of the speakers is obtained through de-panning.

2.2. Panning of binaural audio

To register the virtual speakers with the local environment, the head rotation of the local user has to be taken into account. The remote speakers are played back through VADs at fixed positions in the local environment by processing the binaural recording according to head movements of the local user. This "panning" process is analogous to the de-panning process described in the previous section. The head of the local participant is tracked and the spatial cues contained in the binaural recording are adjusted by tuning ITD and ILD. By combining the head orientations of the remote and the local user, the de-panning and the panning process are merged to a single processing stage. Instead of de-panning the recording to the original position (to compensate for head rotation of the remote user) and then panning it to the desired position (determined from the head orientation of the local user), the recording is directly panned to the desired position.

Merging de-panning and panning to a single process provides the advantage of eliminating redundant computations. Low latency is vital in an interactive telecommunication scenario. Processing the binaural audio in a single step has another major benefit: In a communication scenario it is natural for participants to turn towards the speaker. Therefore, the head orientations of both the remote and the local user are assumed to be similar, if the speaker is registered with the local user's environment. In this case, little or no processing is applied to the binaural recording (cf. Fig. 1d), as the actual source position, relative to the remote user, and the desired source position, determined from the head orientation of the local user, are similar or identical.

3. USER STUDY

To evaluate the performance of the proposed AAR telecommunication system under controlled conditions, a formal user study was conducted. 13 test subjects with normal hearing were used in a within-subjects design. 5 of the test subjects were students of the Department of Media Technology of the Helsinki University of Technology. Having vast experience in using and assessing spatial audio, they were classified as "professional listeners". The other 8 subjects had little or no experience with spatial audio, and were thus classified as "naïve listeners". The test subjects were presented with a binaural recording simulating a teleconference. The conference was recorded via a MARA headset in a room with a

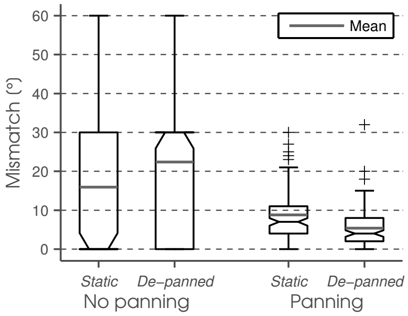

Figure 4: *Absolute angle mismatch.* The mean and median absolute angle mismatch is significantly higher without panning than with panning enabled.

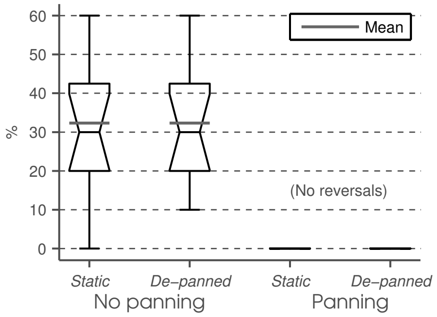

Figure 5: *Front–back reversals.* The *static* and *de-panned* condition yield the same mean reversal rate. No front–back reversal is observed with panning enabled, in either condition.

reverberation time of 0.3–0.5 s.

To analyse the performance in each test condition, mean and median values are analysed and compared using parametric ANOVA and non-parametric Friedman analysis. To compare two matched conditions, a paired t-test is performed. Judgements are based on 0.05 significance level. Results are given as p-values for rejecting the null hypothesis. For multiple comparisons, a post test with Tukey-Kramer correction is applied.

3.1. Localisation of virtual speakers

To test the ability of test subjects to localise a speaker recorded with the MARA headset, a recording was used consisting of ten repetitions of a male speech sample from the "Music for Archimedes" CD [16]. The sample duration is about 11 s, with 1 s of silence between each repetition. Two different conditions were tested: *static* and *de-panned*.

For the *static* condition, the binaural recording was made using five loudspeakers (cf. Fig. 3): three in front (at $30°$, $0°$ and $-30°$ azimuth), one to the right (at $-90°$), and one in the back (at $150°$). The anechoic speech sample was played from each loudspeaker, in random order, with each direction occurring twice.

For the *de-panned* condition, a situation was assumed where the remote participant recording the conference is turning towards the currently active speaker. To simulate this scenario, one loudspeaker in front of the MARA headset user at $0°$ azimuth was used for the recording. The speech sample was the same as in the *static* condition. The recorded sample was then de-panned to encode the interaural cues of the same azimuth angles as used in the *static* condition (i.e. $150°$, $30°$, $0°$, $-30°$ and $-90°$). The listener should thus perceive the speakers as emanating from these directions, even though they were recorded with the remote user facing them. Again, the order of the directions was randomised, with each direction occurring twice.

Presented with a binaural recording from the MARA headset, the test subjects were asked to identify the direction of the speakers from twelve possible directions in the horizontal plane, spaced $30°$. The test was conducted using an unprocessed *static* and a *de-panned* recording. To minimise learning effects, the order of the recordings was randomised among subjects.

In the second task, the head of the test subject was tracked, to register the virtual speakers with the environment. The test subjects were asked to turn towards the speakers to specify their directions. Again, this was tested in the *static* and *de-panned* condition, in random order.

3.1.1. Angle mismatch

An objective measure to determine the performance of test subjects to localise speakers from the MARA recording is the angle mismatch between the guess β of the test subject and the actual recording angle α. In the second part, where test subjects are asked to turn towards the speaker, the mismatch is calculated as the offset between the playback angle α and the head orientation β of the test subject.

The mismatch is compensated for front–back reversals and they are analysed separately. A front–back reversal occurs, when the test subject perceives the source as being in front when in fact it is in the back, and vice versa. The error due to the reversal is removed from the angle mismatch, as it would severely distort the measurement results [17].

Boxplots of the mean and median absolute angle mismatches in both subtasks are shown in Fig. 4. To compare performance under the two conditions in each subtask, a paired two-way analysis is performed on the absolute values of the angle mismatches. Applying a two-way ANOVA to the data of subtask I reveals that the mean absolute angle mismatch without panning is significantly smaller with the *static* recording ($15.9°$) than with the *de-panned* recording ($22.4°$), $F(1, 12) = 6.57$, $p_{Cond} = 0.0110$. The Friedman analysis yields an analogous result: The median of the absolute angle mismatch is significantly smaller with the *static* recording ($0°$) than with the *de-panned* recording ($30°$), $\chi^2(1, n = 13) = 6.13$, $p_{Cond} = 0.0133$. No significant difference between subjects is found (ANOVA: $F(1, 12) = 1.02$, $p_{Subj} = 0.4315$, Friedman: $\chi^2(12, n = 2) = 12.97$, $p_{Subj} = 0.3714$).

With panning enabled the order is reversed: The mean absolute angle mismatch is significantly lower in the *de-panned* condition ($5.4°$) than in the *static* condition ($8.8°$), $F(1, 12) = 14.42$, $p_{Cond} = 0.0002$. The Friedman analysis indicates a significantly smaller median with the *de-panned* recording ($4.0°$) than with the *static* recording ($7.0°$), $\chi^2(1, n = 13) = 16.83$, $p_{Cond} = 0.0000$. Again, no significant difference between subjects is found (ANOVA: $F(1, 12) = 1.59$, $p_{Subj} = 0.0952$, Friedman: $\chi^2(12, n = 2) = 15.87$, $p_{Subj} = 0.1972$). A Tukey-Kramer post test indicates a significantly higher mean absolute angle mismatch in the *de-panned* condition without panning than in

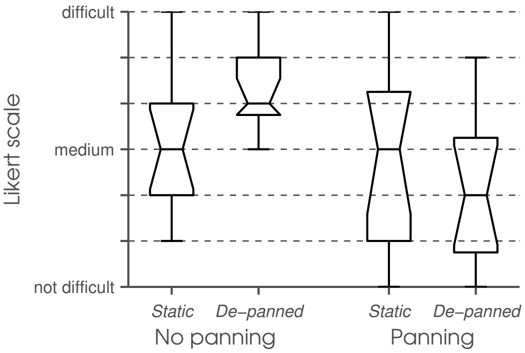

Figure 6: *Perceived difficulty of speaker localisation task.* Localising speakers in the *de-panned* condition without panning is perceived to be significantly more difficult than in the *de-panned* condition with panning enabled.

any of the other conditions.

3.1.2. Front–back reversals

The mean front–back reversal rate is equal in both tested recording conditions without panning: 32 percent (cf. Fig. 5). This is close to chance level, as 2 out of the 10 tested directions were at the extreme right (-90°), where no reversal can occur. Most of the reversals (83 percent in the *static* and 85 percent in the *de-panned* case) occurred when a source was mistakenly perceived to be in the back. The chance of this kind of error was increased by the fact that frontal source directions prevailed in the test. A Lilliefors normality test indicates that the error rates follow a normal distribution.

With panning enabled, no front–back reversal was observed. All test subjects managed to correctly identify whether a source was in front or in the back when asked to turn towards the virtual source.

3.1.3. Perceived difficulty

As a subjective measure, test subjects were asked to judge the perceived difficulty of each subtask. The difficulty was marked on a balanced seven-step Likert scale [18], ranging from *not difficult* to *difficult*, with *medium* marking the centre point. To compare the perceived difficulty of each subtask, a Friedman analysis is performed on the medians (cf. fig. 6). The null hypothesis is rejected for the first subtask, indicating that localisation in the *static* condition is perceived to be significantly less difficult than in the *de-panned* condition, $\chi^2(1, n = 13) = 7.36$, $p = 0.0067$. No significant difference between conditions is found in subtask II, $\chi^2(1, n = 13) = 1.6$, $p = 0.2059$. When comparing both subtasks, the Friedman analysis indicates a significant difference between all tests, $\chi^2(3, n = 13) = 14.5221$, $p = 0.0023$. A post test with Tukey-Kramer correction reveals the speaker localisation in the *de-panned* case without panning to be perceived significantly more difficult than speaker localisation in the *de-panned* case with panning.

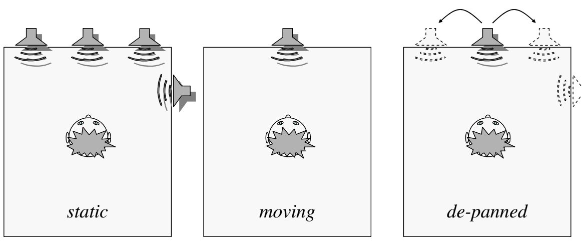

Figure 7: *Recording conditions for speaker segregation task.* For the *static* recording, a separate loudspeaker is used for each speaker. The *moving* and *de-panned* recordings are obtained using just one loudspeaker. De-panning is applied to separate the speakers on the *de-panned* recording spatially, thus simulating the speaker positions used in the *static* recording.

3.2. Segregation of virtual speakers

Speech samples from the TIMIT database [19] were recorded via the MARA headset. Two groups of four male speakers were chosen from the database. Twenty speech samples were recorded per tested condition, five from each speaker. The speakers talk in turns, in random order. The segregation performance is tested in three different conditions: *static*, *moving* and *de-panned* (cf. fig. 7). In the *static* condition, each speaker is assigned one of four different loudspeakers in the recording hall, at 60°, 30°, 0°, and at -30°. This simulates a situation where the conference participants are seated around a table with the MARA headset user. In the *moving* condition, just one loudspeaker in front of the MARA user is used. All four speakers are recorded at 0° azimuth. This simulates a situation where the MARA headset user turns towards the active speaker during the simulated conversation. For the *de-panned* condition, the same recording setup is used as in the *moving* condition, but de-panning is applied to each recorded speaker, to yield the same perceived speaker directions as in the *static* condition, in random order.

In each condition, the ability of test subjects to segregate the four speakers on the binaural recording is tested. The test subjects are presented with the binaural recordings and asked to mark the words of one of the four speakers in each condition. The *static* and *de-panned* condition are also tested with panning enabled, by tracking the head of the test subject. This way, the recorded speakers are registered with the environment, allowing the test subjects to turn towards them during the test.

The segregation task is repeated three times, to analyse the impact of learning effects on the performance. To counterbalance the order in which the conditions are presented, the order is governed by a *Latin square* [20], and randomised among subjects.

3.2.1. Error rates

The segregation performance is measured in terms of the number of correctly identified speaker turns. The most striking result is the speaker segregation performance in the *moving* condition, producing the highest error rates in all three rounds (cf. fig. 8). The differences between mean and median error rates are found to be significant (ANOVA: $F(4, 2) = 26.34$, $p_{Cond} = 0.0000$, Friedman: $\chi^2(4, n = 3) = 71.34$, $p_{Cond} = 0.0000$). Applying a Tukey-Kramer post test to the ANOVA results indicates that

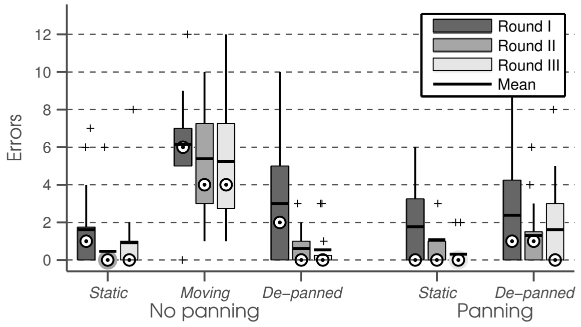

Figure 8: *Error rates of speaker segregation task* for identifying 5 turns of the speaker in question. The mean and median error rates in the *moving* condition are significantly higher than in all other conditions in round II and III. The performance of test subjects improved significantly from round I to round II.

the *moving* condition leads to significantly higher mean error rates compared to all other conditions, in all three rounds. No significant difference is found between the other four conditions, i.e. *static* and *de-panned* with and without panning. Similar conclusions can be drawn from a Tukey-Kramer post test of the Friedman analysis results. In rounds II and III, test subjects performed significantly worse in the *moving* condition than in all other conditions. No significant difference is found between the mean ranks of the *static* and *de-panned* conditions in any of the three rounds. Whether or not panning is used to register the virtual speakers with the environment has no significant impact on the performance, as indicated by a two-way ANOVA, $F(1,1) = 2.33$, $p_{Pan} = 0.1287$.

To identify learning effects, the segregation performance of all three rounds is compared. A two-way ANOVA indicates significant differences between the mean error rates in the three rounds, $F(2,4) = 5.96$, $p_{Rnd} = 0.0031$. A Friedman analysis yields analogous results regarding the median error rates, $\chi^2(2, n = 5) = 14.98$, $p_{Rnd} = 0.0006$. A Tukey-Kramer post test reveals a significant improvement of the segregation performance from round I to round II. No significant improvement from round II to round III is found. A two-way ANOVA indicates that no significant interaction effects exist between the test round and the test condition, $F(2,8) = 0.48$, $p_{Int} = 0.8699$. The improvement after round I is thus independent of the test condition.

3.2.2. Perceived difficulty

A Friedman analysis indicates a significant difference between the perceived difficulty of the five test conditions, $\chi^2(4, n = 11) = 74.44$, $p = 0.0000$. Two test subjects were removed from this analysis due to missing entries. A Tukey-Kramer post test reveals the *moving* condition to be perceived significantly more difficult than all other conditions, as depicted in Fig. 9. No significant difference is found between the other conditions.

3.3. Comments of test subjects

One of the most stated problems in the speaker localisation task was inside-the-head locatedness [13]. Test subjects reported difficulties to localise sound sources that were straight ahead, as they often lacked externalisation. This was said to be confusing. Some

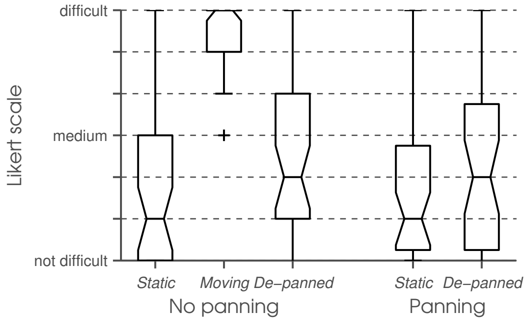

Figure 9: *Perceived difficulty of speaker segregation task.* The *moving* condition is perceived to be significantly more difficult by test subjects than all other conditions.

test subjects pointed out a lack of depth in the *de-panned* recording. Whereas the sound sources appeared to be positioned on a "clear circle" in the *static* recording, in the *de-panned* recording they seemed to be positioned on a "straight line", ranging from the far left to the far right of the listener. This made it more difficult to map sources to a virtual circle than in the *static* case.

Most test subjects pointed out difficulties to distinguish speakers in the *moving* recording. Some test subjects said they became more acquainted with the voice of the speaker in question towards the end of the test, and managed to segregate the speakers based on their accents or articulations. In the other test conditions test subjects reported to rely mainly on the direction when segregating different speakers.

Only one test subject named the head tracking as a helpful factor in the speaker segregation task. Another subject stated that turning towards the speaker in question made the segregation task indeed more difficult, as it was easier to localise and identify a speaker a bit off the centre. Yet another test subject named inside-the-head locatedness as a cue for segregating the speakers: After turning towards the speaker in question, that speaker was not externalised anymore, which clearly separated him from the other speakers in the recording.

4. DISCUSSION

The *static* case, made with several loudspeakers at fixed positions, and recorded without head movement, represents the "ideal" case of a binaural recording, preserving the spatial cues of all speakers. In the *de-panned* recording, simulating a situation where the MARA headset user moves the head during the recording, interaural cues are restored by compensating for the head movements through the de-panning algorithm. If no panning is applied during playback to register the recorded speakers with the environment, the *de-panned* recording yields a significantly larger mean and median absolute angle mismatch between the perceived and the actual direction of the recorded speakers than the *static* recording. This indicates that the de-panning algorithm cannot fully restore the spatial cues contained in the recording. Test subjects perceived localisation with the *de-panned* recording to be significantly more difficult than with the *static* recording. This may be related to the fact that some test subjects perceived the speakers in the *de-panned* recording to be positioned on a line, whilst in the *static* recording they appeared to reside on a circle around the listener, with distinct directions.

With head tracking and panning enabled, the mean and median absolute angle mismatch decreased significantly. Test subjects localised speakers significantly more accurately by turning towards them than by indicating their directions. The reduced localisation blur, defined as the minimum audible displacement [13], achieved by facing the virtual speakers implies that registering virtual sources with the environment through panning may lead to better spatial separability of the sources. This is seen as a major benefit in a telecommunication scenario. When comparing the two test conditions with panning enabled, the *de-panned* condition leads to a significantly better localisation performance. As test subjects turn towards the de-panned speaker, their head orientation approximately matches the head orientation during the recording, therefore nearly unprocessed audio is delivered to the test subjects (c.f. fig. 1d). Turning towards a virtual source recorded off the centre, as in the *static* case, increases the localisation blur significantly, as the panning algorithm fails to fully restore the spatial cues.

No effect of the recording condition on the number of front–back reversals is found. The *de-panned* recording does not yield a higher rate of reversals than the *static* recording. We assume front–back reversals to be mainly a result of the ambiguity of interaural cues in general, not of the processing involved in generating them. A more striking finding, however, is the fact that with head tracking and panning enabled, no front–back reversal occurred in any of the 260 observations. This is a strong argument for the hypothesis that panning improves the localisation performance. When a test subject turns the head to search for the virtual sound source, the interaural cues change accordingly, indicating unambiguously whether the source is in front or in the back. Even test subjects without any prior experience with spatial audio and head tracking instinctively interpreted these motional cues correctly.

Results of the speaker segregation task prove the importance of interaural cues to segregate multiple speakers. The *moving* condition, which contains little or no interaural cues to separate speakers, leads to significantly higher mean and median error rates than the *static* and *de-panned* cases, which contain natural or algorithmically restored interaural cues. Even after being presented with the same recording for the third time in round III, the median error rate of test subjects when trying to identify the 5 turns of the speaker in question is 4. Some subjects stated their choices in the *moving* case to be based on pure guessing, others marked no turn at all. In all other conditions the median error rate in round III drops to 0, indicating that more than 50 percent of the test subjects managed to identify all speaker turns correctly. The result is supported by the perceived difficulty, with the *moving* condition rated significantly more difficult than all other conditions. This underlines the importance of spatial cues to segregate multiple speakers in a telecommunication scenario.

No significant differences are found between the *static* and *de-panned* case regarding the speaker segregation. Whilst the de-panning has a negative effect on the speaker localisation, it does not deteriorate the speaker segregation performance. Compared to an unprocessed binaural recording with no or misleading interaural cues, such as the *moving* recording, de-panning significantly improves speaker segregation, and theoretically yields the same performance as the ideal case of a *static* recording devoid of head movements.

The segregation performance improved significantly from round I to round II. This is attributed to the fact that test subjects became acquainted with the test procedure and the a priori unfamiliar voices of the speakers used in the test. No significant improvement from round II to round III is found, indicating that learning effects vanish after round I.

5. CONCLUSIONS

An audio augmented reality (AAR) telecommunication system based on the transmission of binaural recordings from a MARA headset is presented. The binaural recordings preserve the spatial cues of recorded sound sources, yielding a listening experience similar to the natural auditory perception of an environment. Head movements distort the spatial cues and thus the perceived directions of the recorded sound sources. A de-panning algorithm is presented to restore the perceived directions. The localisation accuracy of virtual sources contained on a de-panned recording was analysed in a formal user study with 13 test subjects. After de-panning, test subjects were able to localise speakers in a binaural recording, though with a significant increase of the mean absolute angle mismatch compared to an unprocessed recording not distorted by head movements.

A panning algorithm adjusts the binaural playback according to head movements of the listener, to register the binaurally recorded sound sources with the environment. With panning enabled, the localisation performance of test subjects improved significantly. The test subjects interacted with the system intuitively, using head rotations to “search” for the virtual sources. No significant performance difference was found between subjects, even though about half of the test subjects had no previous experience with spatial audio or head tracking. These results imply that the proposed system is suitable also for “naïve” users. By registering the virtual sources with the environment, no front–back reversal occurred, i.e. all test subjects correctly determined whether a source was in front or in the back.

To analyse their ability to segregate multiple recorded speakers, test subjects were asked to identify speaker turns on a binaural recording. Interaural cues are shown to improve the segregation performance of test subjects significantly, compared to a recording with no interaural cues. In case of misleading spatial cues, i.e. arbitrary changes in the perceived directions of the sources due to head movements during the recording, the performance is expected to be even worse. No significant difference is found between the recordings containing interaural cues. The *de-panned* recording, in which the spatial cues are algorithmically restored, does not lead to a significantly worse performance than the ideal case, an unprocessed binaural recording of sound sources separated in space, devoid of head movements. In a telecommunication scenario, the de-panning algorithm restores the perceived directions of speakers and enhances the ability of a listener to segregate the participants of a meeting. This is assumed to improve the listening comfort and the ability to follow a remote conversation, which is a major argument for the use of AAR in a telecommunication scenario.

Transmitting a binaural recording of one’s environment via a MARA headset is a simple yet effective way to share auditory perception over distance. Tackling issues related to head movements with the algorithms proposed in this work allowed both experienced an inexperienced users to localise virtual sources on a binaural recording. This significantly improved the ability of test subjects to segregate multiple sources on the recording. Due to their simplicity, the proposed de-panning and panning algorithms run on a standard PC, with a responsiveness that was found to be sufficient for the test scenario. System lag was an issue only in the case of fast head movements, due to the limited update rate of the head tracking device. The processing is based on simple ITD (interaural time difference) and head shadowing models, hence the system does not require a dataset of head-related transfer functions (HRTFs). This makes it transferable and relatively robust against individual HRTF variations.

In terms of future research, parametrisation of the algorithms could account for individual differences among users and various recording environments and yield more accurate spatial cues. This might improve the localisation accuracy of virtual sources. A central aspect of AAR is the combination of real and virtual auditory content. An issue further to be investigated upon is the mixing of binaural recordings from a remote end with the pseudo-acoustic environment, perceived through the MARA headset. The biggest limitation of the proposed system is that it currently supports only one virtual source at a time, i.e. speakers talking in turns. To allow for multiple simultaneous speakers, a time-frequency decomposition approach as employed in Directional Audio Coding (DirAC) [21] could be integrated to the system, to segregate and process each speaker individually.

The proposed implementation of an AAR telecommunication system using VAD might serve as a valuable tool to enhance existing telecommunication systems and help overcome the gap to face-to-face communication.

6. ACKNOWLEDGMENT

The research leading to these results has received funding from Nokia Research Center [kamara2009], the Academy of Finland, project no. [119092] and the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement no. [203636].

7. REFERENCES

[1] B. A. Nardi and S. Whittaker, "The place of face-to-face communication in distributed work," in *IN P. HINDS AND S. KIESLER (EDS.), DISTRIBUTED WORK*. MIT Press, 2002, pp. 83–112.

[2] P. Rohde, P. M. Lewinsohn, and J. R. Seeley, "Comparability of Telephone and Face-to-Face Interviews in Assessing Axis I and II Disorders," *Am J Psychiatry*, vol. 154, no. 11, pp. 1593–1598, 1997.

[3] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *The Journal of the Acoustical Society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[4] R. W. Lindeman, D. Reiners, and A. Steed, "Practicing what we preach: IEEE VR 2009 virtual program committee meeting," *Computer Graphics and Applications, IEEE*, vol. 29, no. 2, pp. 80–83, March-April 2009.

[5] M. Billinghurst, H. Kato, K. Kiyokawa, D. Belcher, and I. Poupyrev, "Experiments with face-to-face collaborative ar interfaces," *Virtual Reality*, vol. 6, no. 3, pp. 107–121, 2002.

[6] B. Kapralos, M. R. Jenkin, and E. Milios, "Virtual audio systems," *Presence: Teleoper. Virtual Environ.*, vol. 17, no. 6, pp. 527–549, 2008.

[7] R. Drullman and A. W. Bronkhorst, "Multichannel speech intelligibility and talker recognition using monaural, binaural, and three-dimensional auditory presentation," *The Journal of the Acoustical Society of America*, vol. 107, no. 4, pp. 2224–2235, 2000.

[8] M. H. Bazerman, J. R. Curhan, D. A. Moore, and K. L. Valley, "Negotiation," *Annual Review of Psychology*, vol. 51, no. 1, pp. 279–314, 2000.

[9] R. Shilling and S. B. Cunningham, *Virtual auditory displays*, ser. Handbook of Virtual Environments. Mahwah NJ: Lawrence Erlbaum Associates, 2002.

[10] H. Lehnert and J. Blauert, "Virtual auditory environment," in *Advanced Robotics, 1991. 'Robots in Unstructured Environments', 91 ICAR., Fifth International Conference on*, June 1991, pp. 211–216 vol.1.

[11] R. Lindeman, H. Noma, and P. Goncalves de Barros, "An empirical study of hear-through augmented reality: Using bone conduction to deliver spatialized audio," in *Virtual Reality Conference, 2008. VR '08. IEEE*, March 2008, pp. 35–42.

[12] A. Härmä, J. Jakka, M. Tikander, M. Karjalainen, T. Lokki, J. Hiipakka, and G. Lorho, "Augmented reality audio for mobile and wearable appliances," *Journal of the Audio Engineering Society*, vol. 52, no. 6, pp. 618–639, 2004.

[13] J. Blauert, *Spatial Hearing - Revised Edition: The Psychophysics of Human Sound Localization*. The MIT Press, October 1996.

[14] F. L. Wightman and D. J. Kistler, "Factors affecting the relative salience of sound localization cues," in *Binaural and Spatial Hearing in Real and Virtual Environments*, R. H. Gilkey and T. R. Anderson, Eds. Mahwah, New Jersey: Lawrence Erlbaum Associates, 1997, pp. 1–23.

[15] U. Zölzer, Ed., *DAFX:Digital Audio Effects*. John Wiley & Sons, May 2002, ch. Spatial Effects by D. Rocchesso, pp. 137–200.

[16] Bang & Olufsen, "Music for Archimedes," CD B&O 101, 1992.

[17] E. M. Wenzel, M. Arruda, D. J. Kistler, and F. L. Wightman, "Localization using nonindividualized head-related transfer functions," *The Journal of the Acoustical Society of America*, vol. 94, no. 1, pp. 111–123, 1993.

[18] H. J. Gardner and M. A. Martin, "Analyzing ordinal scales in studies of virtual environments: Likert or lump it!" *Presence: Teleoper. Virtual Environ.*, vol. 16, no. 4, pp. 439–446, 2007.

[19] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus CDROM," 1993.

[20] N. Rapanos, "Latin squares and their partial transversals," in *Harvard College Mathematics Review*, S. D. Kominers, Ed. Harvard College, 2008, vol. 2, pp. 4–12.

[21] V. Pulkki and C. Faller, "Directional Audio Coding: Filterbank and STFT-Based Design," in *Preprint 120th Conv. Aud. Eng. Soc.*, 2006.