On the Evaluation of Perceived Spatial Immersion in the Application of Automatic Upmixing for 3D Surround Sound Systems

Alexander Poets, Stephan Preihs, Jürgen Peissig
Leibniz Universität Hannover, Institut für Kommunikationstechnik
Appelstr. 9A, 30167 Hannover, Email: poets@ikt.uni-hannover.de

Introduction

Although home cinema and entertainment systems capable of reproducing 3D multichannel surround formats are widely in use, the availability of most musical content is restricted to the two-channel stereo format. To bridge this gap, automatic (blind) upmixing algorithms can be employed. The present work investigates whether an automatic upmix from a stereo version is perceived similar to a manually created (dedicated) 3D surround mix in terms of spatial immersion which is supposed to be reflected by the spatial qualities of a virtual (acoustic) environment. Accordingly, we developed an upmixer that converts from stereo to 5.1.4 multichannel surround. Along with a commercial upmixer, the proposed design is evaluated and compared to dedicated surround mixes in the context of a listening experiment. Participating subjects were required to rate the perceived spatial quality of an upmixed test condition in relation to a manually mixed reference using SAQI descriptors. Our results give insights under which conditions an upmix can be perceived as being similar in spatial quality to a dedicated surround mix. We also show that, generally, the performance of an upmixer depends on the spatial cues embedded in the source material and the manual mix to which the upmix is compared.

Upmixing from Stereo to 5.1.4

Upmixing is generally understood as the process of synthesizing n channels from a m-channel source signal, where $m < n$. In the case of *blind* upmixing, the input signals are neither (matrix-)encoded nor is there any auxiliary information that helps guide the upmix process. This allows arbitrary source material to be adapted for different speaker setups, e.g., 3D surround sound systems.

The architecture of the proposed stereo-to-5.1.4 upmixer is illustrated in Figure 1. In the input stage, the incoming audio samples are gain-adjusted and buffered to a consistent frame size that can be used with a *fast Fourier transform* (FFT). After having been transformed to the frequency domain, the stereo signal is re-panned to incorporate the additional center speaker in the front into the mix. To generate the signals for the left and right surround channels, the stereo signal is decomposed into amplitude-panned direct sound (primary) components and uncorrelated diffuse sound (ambient) components, a procedure commonly referred to as *primary-ambient extraction* (PAE). Whereas the primary components can be discarded, the ambient components are subject to further processing before eventually being fed to the surround channels. Both the re-panning and ambience extraction modules rely on the subband cross- and autocorrelations of the input stereo signal which are computed in a preceding analysis stage. Using an inverse FFT and the *weighted overlap-add* (WOLA) method, the processed short-time spectra are finally transformed back into the time

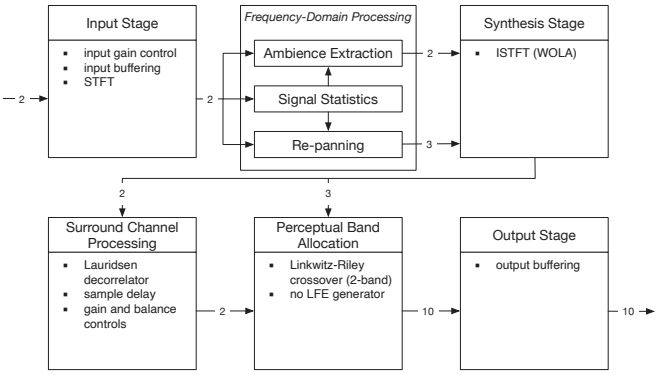

Figure 1: Processing scheme of the proposed upmixer that converts from two-channel stereo to 5.1.4 multichannel surround.

domain. The extracted ambient signal components are then passed through a so-called Lauridsen decorrelator [2] which essentially corresponds to a pair of complementary two-tap *finite impulse response* (FIR) comb filters. This way, the *apparent source width* (ASW) of the ambient sound sources is further increased. Additionally, to prevent source delocalization due to the precedence effect, the decorrelated signals are delayed by a few samples. Along with the signals for the three front channels which have been obtained in the re-panning stage, the modified ambient components are further processed in a subsequent module to create the height channels. This involves a 2D-to-3D upmixing technique named *perceptual band allocation* (PBA) [4] that exploits the so-called pitch-height effect (i.e., the observation that the predominant frequency of a source affects its perceived vertical position). Using a two-band crossover filter in Linkwitz–Riley topology, the signals are split into lower and upper frequency bands which are then routed to the main and height speakers, respectively. Finally, the outgoing samples are written to an output buffer. The proposed design has been implemented in MATLAB and compiled to a *Virtual Studio Technology* (VST) plugin using the *Audio Toolbox* and its code generation capabilities. This allows the plugin to be hosted in any VST-compatible *digital audio workstation* (DAW).

In addition to the proposed design, it was considered appropriate to also include a commercial solution in the experiment. For that purpose, the *Auro-Matic Pro 3D* upmixer (v3.0.4) from Auro Technologies was selected which is conveniently available as an AAX plugin. It can be argued that the Auro upmixer is well suited to complement the comparison as it relies on a different principle of operation than the proposed design: The Auro upmixer apparently uses reverberators to create an artificial spatial impression [6], whereas the proposed design exploits the spatial cues already present in the source material.

Experimental Design

With respect to the evaluation of immersion in the context of this experiment, some considerations are necessary. It is widely acknowledged that the concept of "immersion" must not be regarded as unidimensional but rather recognized as multifaceted. Therefore, a variety of attempts to define and classify the facets of immersion have been made. Borrowing from the work of Zhang et al. [8], we are distinguishing between spatial and emotional immersion within the scope of this paper. While *spatial* immersion is supposed to be elicited by the spatial qualities of a virtual (acoustic) environment, *emotional* immersion is considered to refer to the feeling of being "emotionally aroused and absorbed" by the presented scene. For the purpose of the experiment, we constrained the assessment to the spatial dimension of immersion. Accordingly, we chose to employ the *Spatial Audio Quality Inventory* (SAQI) [5] which is designed to assess the perceived spatial quality of a stimulus in relation to a reference. The SAQI comprises 48 verbal descriptors of which the six descriptors *clarity*, *degree-of-liking*, *height*, *localizability*, *naturalness* and *presence* were deemed the most relevant in the context of this experiment. To investigate how an automatic upmix from a stereo version compares to a manually created surround mix with respect to these descriptors, a set of appropriate stimuli is required. Consequently, excerpts from eight musical pieces in different genres (all between 33 and 68 seconds in length) have been selected that were available in both stereo and dedicated 5.1.4 versions. A detailed description of the stimuli can be found in [1]. As required by the SAQI, a reference condition had to be selected in relation to which participating subjects rate the anchor and test conditions. In this experiment, the manually mixed 5.1.4 (*3D*) version was chosen to constitute the reference, whereas the stereo version served as anchor. As for the test conditions, upmixes from the stereo versions were created using both our in-house development as well as the commercial software by Auro Technologies (subsequently referred to as *IKTUpmix3D* and *AMPro3D*, respectively). Consequently, for each of the eight musical pieces, three conditions had to be compared to the reference, which amounts to a total of 24 comparisons per subject. When creating the AMPro3D versions, the factory settings of the Auro upmixer have been used. The parameters of the IKT audio plugin were selected based on what we liked best in an informal listening session.

In the experiment, the subjects were presented two conditions at a time, i.e., the reference and a condition under test (which was either the stereo anchor or one of the upmixed versions). After having listened intently to both conditions, the subjects were asked to rate the perceived quality of the condition under test relative to the reference. Each of the verbal descriptors had to be rated on a continuous bipolar scale that ranges between ±3, where a value of zero marks the middle ("no difference perceived"). Whereas a positive difference rating indicates that the subject perceived the condition under test to be higher in spatial quality than the reference, a negative difference rating translates to the opposite. To control for possible confounding factors, the musical pieces and conditions were presented in random order. The order of the verbal descriptors was also randomized across subjects, but not varied between comparisons to reduce the cognitive load. The questionnaire which allowed the participants to control the audio playback and enter their ratings was implemented in MATLAB using the *App Designer*. A screenshot of the *graphical user interface* (GUI) is provided in Figure 2. In the implementation, the audio playback was handled by sending *Open Sound Control* (OSC) commands to a remote REAPER session containing the preprocessed and loudness-normalized stimuli. To reduce the difficulty of the comparison, the application enabled subjects to instantaneously switch between the presented conditions during playback.

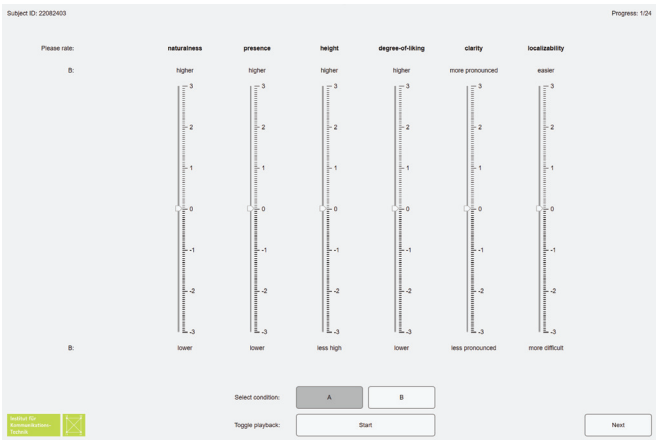

Figure 2: A screenshot of the questionnaire GUI in MATLAB. "A" refers to the reference, whereas the condition under test is labeled "B".

Results and Conclusions

The listening experiment was conducted over the span of two weeks in August 2022 at the institute's *Immersive Media Lab* (IML) [3]. The speaker arrangement that was used complies with Rec. ITU-R BS.2051-3 A and D (for stereo and 5.1.4, respectively). Participating subjects were seated in the sweet spot position and provided with a tablet computer which ran the questionnaire application. The test instructions and circumscriptions of the qualitative descriptors were handed out in written form and again read out loud by the experimenter. Due to the technical nature of the SAQI, only subjects with a background in audio participated in the experiment. In total, 14 subjects (13 M, 1 F) aged between 23 and 36 years (Median: 27, SD: 5.29) participated in the experiment. Most of the subjects were either students or employees of the institute. All subjects reported normal hearing and nine of them had participated in listening experiments before.

The statistical analysis was carried out using R 4.2.2 including the `stats` and `coin` packages. All statistical tests were evaluated at the $\alpha = 0.05$ significance level. The test procedure described in the following has been applied to each pair of qualitative descriptor and musical piece. First, a Shapiro–Wilk test of normality was employed to determine whether parametric or non-parametric methods have to be applied. In case the test indicated that a normal distribution of the dependent variable (i.e., the rating) could be assumed for each level of the within-subjects factor (i.e., the condition under test), a one-way *repeated measures analysis of variance* (rANOVA) was applied, and a Friedman test otherwise. As for the rANOVA, violations of the assumption of sphericity were identified using Mauchly's test and Greenhouse–Geisser corrections applied where appropriate. In the context of this experiment, the null hypothesis H_0 of the rANOVA and Friedman tests can be for-

mulated as follows: “There is no difference in mean ratings for a specific qualitative descriptor between the stereo, AMPro3D and IKTUpmix3D versions of a specific musical piece.” If H_0 could be rejected, a post-hoc test was further carried out. For that purpose, either a paired t-test (in case of the rANOVA) or a Wilcoxon signed-rank test (in case of the Friedman test) was used. To control the *family-wise error rate* (FWER), the p-values of the post-hoc multiple comparisons were adjusted using Holm’s method.

Figure 3 shows the observed mean ratings and corresponding 95% *confidence intervals* (CIs) for each qualitative descriptor. The CIs were estimated using *bias-corrected and accelerated* (BCa) bootstrapping with 10 000 repetitions. As can be seen from Figure 3a, only the AMPro3D and stereo versions of Laudate show statistically significant differences ($p < 0.05$) in mean ratings for the descriptor *clarity* which refers to the “impression of how clearly different elements in a scene can be distinguished from each other, how well various properties of individual scene elements can be detected” [5]. Generally, the anchor and test conditions have similar means and, on average, were perceived to be worse in clarity than the manually mixed 3D reference condition. It should be noted, however, that for some musical pieces (Laudate, Mellow and School, namely) the AMPro3D version has mean ratings just slightly below zero and, therefore, performs similar to the reference condition. In the case of Walkuere, all versions were rated slightly better than the reference condition on average. The worst average ratings can be observed for the musical piece Hantel. All in all, these findings allow the conclusion that, in terms of clarity, an upmixer can achieve results similar to a manually created 5.1.4 mix, especially when considering the performance of the AMPro3D version for the musical pieces Laudate, Mellow, School and Walkuere.

Figure 3b illustrates the results for the descriptor *degree-of-liking* which expresses “the perceived overall difference with respect to the degree of enjoyment or displeasure” [5]. Statistically significant differences in the mean ratings can be observed for the musical pieces Laudate, Mellow and Wunderschoen. Walkuere is the only musical piece for which, on average, the anchor and test conditions were liked better than the reference. Other than that, according to the mean ratings, the manually mixed 3D reference was always found to be more pleasant and enjoyable than the stereo and upmixed versions.

The results for the descriptor *height*, which denotes the “perceived extent of a sound source in vertical direction” [5], are shown in Figure 3c. In several cases, the stereo anchor was rated significantly lower (i.e., worse) than one or both of the AMPro3D and IKTUpmix3D test conditions. Consequently, the anchor condition was often correctly identified as such. This observation allows to draw the conclusion that both upmixers were able to add a height dimension to the perceived spatial image. The upmixed versions often have mean ratings greater than zero, which implies that, in those cases, the upmixes were found to extend farther into the vertical direction than the manually mixed 3D reference. The highest ratings were provided for the upmixed versions of Walkuere.

In Figure 3d, the results for the descriptor *localizability* are illustrated. The localizability of the sources in a mix depends on whether they appear diffuse or their spatial extent is perceived to be clearly delimited. For most muscial pieces, the mean ratings do not differ substantially between versions, with an exception: Apparently, the sound sources in the stereo version of Wunderschoen were significantly easier to localize than those in the upmixed versions as compared to the 3D reference condition. On average, the anchor and test conditions were found be worse in localizability than the reference. This is especially true for the stereo and upmixed versions of the musical piece Hantel which received the overall lowest ratings. The overall highest average ratings were given to the stereo and upmixed versions of Laudate which were perceived to be similar in localizability to the manually mixed 3D reference condition. From these findings can be concluded that it generally depends on the musical piece whether an upmixer is comparable to a manually created surround mix in terms of localizability.

Figure 3e shows the results for the descriptor *naturalness* which corresponds to the “impression that a signal is in accordance with the expectation/former experience of an equivalent signal” [5]. As can be seen, the only statistically significant difference in mean ratings occurred between the stereo and IKTUpmix3D versions of Wunderschoen: On average, the stereo anchor was perceived as more natural than the IKTUpmix3D version and found to be on par with the reference condition in terms of naturalness. The overall highest average ratings were provided for the anchor and test conditions of the musical piece Walkuere. Generally, the distribution of mean ratings suggests that the anchor and test conditions were perceived as less natural than the manually mixed 3D reference.

Figure 3f illustrates the results of the descriptor *presence* which translates to the “impression of being inside a presented scene or to be spatially integrated into the scene” [5]. Several statistically significant differences in mean ratings can be observed between the stereo anchor and either or both of the upmixed versions. Most often, the upmixed versions were rated significantly higher than the stereo anchor. Therefore, it can be concluded that upmixers are generally capable of improving the listening experience in terms of presence as compared to the source material. Sometimes, one of the upmixed versions was perceived to be on par with the reference condition (i.e., for Hantel, School and Wunderschoen). In the case of Walkuere, both the AMPro3D and IKTUpmix3D versions were clearly preferred over the manually mixed 3D reference condition.

Discussion

The observations made in the previous section allow the conclusion that an upmix can be perceived as being similar in spatial quality to a dedicated surround mix. Generally, it depends on the spatial cues embedded in the source material and the manual surround mix to which the upmix is compared. In the case of the musical piece Walkuere, the stereo and upmixed versions were, on average, most often preferred over the manually mixed 3D reference condition. This phenomenon may be explained by the fact that the sound engineers did not have the individual tracks of Walkuere when creating the 5.1.4 mix. Consequently, the 3D version of Walkuere cannot be considered a dedicated surround mix but rather a “manual upmix”. With respect to the musical piece Hantel, the stereo anchor and test conditions received exceptionally low ratings in localizability when considering the other musical pieces. This may be due

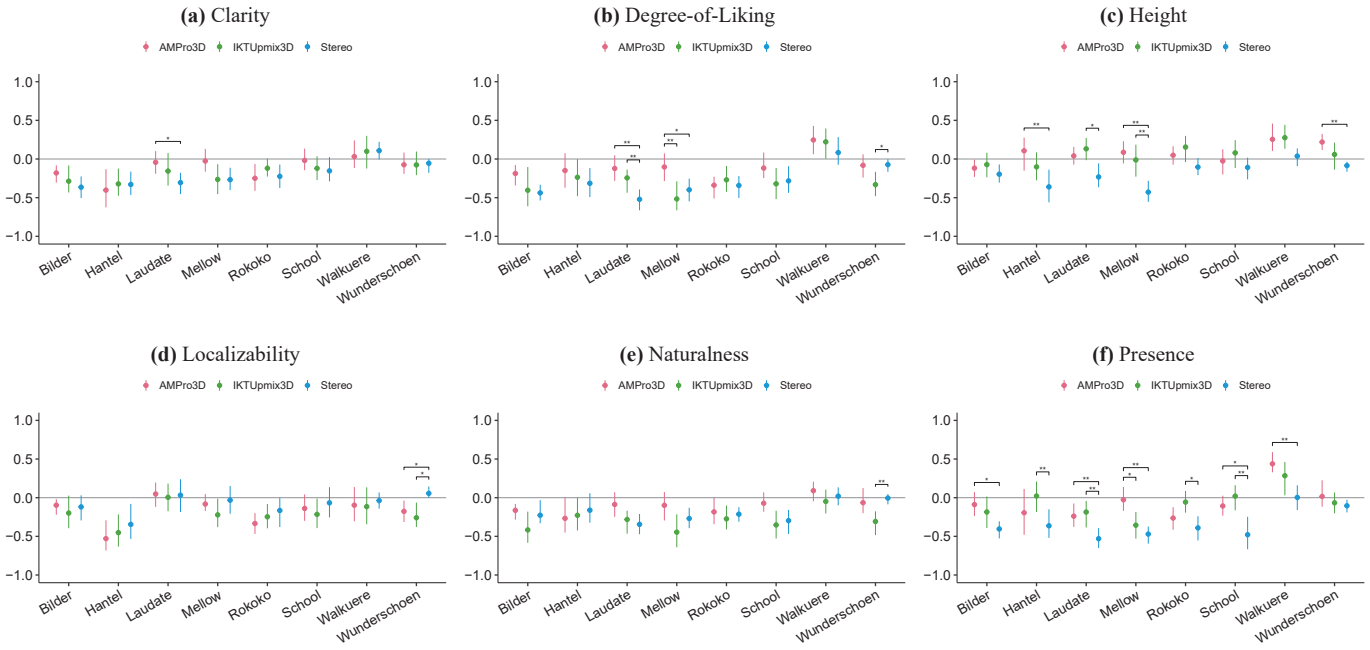

Figure 3: Mean ratings and 95% bootstrap confidence intervals for each qualitative descriptor. The x-axes indicate the musical piece and the y-axes the rating. Rating values have been normalized to the range between ±1. A rating of zero denotes that the subject did not perceive a difference between the reference and the presented condition with respect to a specific descriptor. A rating below zero indicates that the condition was found to be worse in perceptual quality than the reference. Statistically significant differences in mean ratings between conditions are denoted by the ∗ symbol, where a single star ∗, two stars ∗∗ and three stars ∗∗∗ indicate a p-value of less than 0.05, 0.01 and 0.001, respectively.

to the fact that the manually created 5.1.4 reference is a very sophisticated surround mix that uses 3D auto-panning effects which of course an upmixer cannot replicate based solely on the spatial cues embedded in the stereo version.

Except for the height and presence descriptors, the stereo condition was most often perceived as not being significantly worse in spatial quality than the test conditions and, in some cases, even preferred over the upmixed versions. Therefore, it is debatable whether the stereo version is suited to serve as the anchor condition. Due to the optimal listening conditions in the IML, stereo recordings already sound very good and appear to be very rich and detailed in spatial information. Thus, conducting the experiment in a more forgiving and less optimal listening room would probably give much different results which might arguably be more representative for, e.g., a home cinema environment.

Finally, as there are not a lot of significant differences in mean ratings between the AMPro3D and IKTUpmix3D versions of any musical piece, the experiment did not really reveal an overall preference for one or the other. To that end, more participants would have been necessary which was also indicated by the result of an *a priori* power analysis.

Outlook

In this experiment, the evaluation was constrained to the spatial dimension of immersion. Therefore, a future listening experiment should be conducted to also assess the emotional dimension of immersion in the context of the present research question, e.g., by employing items from the *Immersive Music Experience Inventory* (IMEI) [7]. Furthermore, a hidden reference should be included (i.e., as in the MUSHRA test methodology) to not only be able to detect significant differences between the anchor and test conditions but also with respect to the reference. Finally, listening positions off the sweet spot should be evaluated to also address other listening scenarios of practical relevance (when listening with multiple people at once, e.g., at a party).

References

[1] Jakob Bergner et al. “Identification of Discriminative Acoustic Dimensions in Stereo, Surround and 3D Music Reproduction”. In: *J. Audio Eng. Soc.* (2023). Accepted for publication.

[2] Roy Irwan and Ronald M. Aarts. “Two-to-Five Channel Sound Processing”. In: *J. Audio Eng. Soc.* 50.11 (2002).

[3] Institut für Kommunikationstechnik. *Immersive Media Lab*. URL: `https://go.lu-h.de/iml`.

[4] Hyunkook Lee. “2D-to-3D Ambience Upmixing based on Perceptual Band Allocation”. In: *J. Audio Eng. Soc.* 63.10 (2015).

[5] Alexander Lindau. *Spatial Audio Quality Inventory (SAQI)*. Test Manual. Technische Universität Berlin, 2015. DOI: `10.14279/depositonce-1.2`.

[6] Wilfried Van Baelen and Ralph Kessler. *Converter and method for converting an audio signal*. Patent WO2010057997. 2010.

[7] Yves Wycisk et al. “Wrapped into sound: Development of the Immersive Music Experience Inventory (IMEI)”. In: *Frontiers in Psychology* 13 (2022).

[8] Chenyan Zhang et al. “Spatial immersion versus emotional immersion, which is more immersive?” In: *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*. 2017.