Audio Engineering Society

Convention Paper 10714

Presented at the AES 156th Convention
2024 June 15–17, Madrid, Spain

This paper was peer-reviewed as a complete manuscript for presentation at this convention. This paper is available in the AES E-Library (http://www.aes.org/e-lib), all rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

Correlation-Based Surround-to-Immersive Upmixer

Thaddeus Páez[1], Sunil Bharitkar[2], and Ema Souza-Blanes[2]

[1]*Samsung Research Tijuana, Mexico*
[2]*DMS Audio Lab, Samsung Research America*

Correspondence should be addressed to Thaddeus Páez (t.paez@samsung.com)

ABSTRACT

This paper presents a novel surround-to-immersive upmixing approach, incorporating a generalized equal-level ambiance extraction algorithm for multichannel content. Additionally, a new side/rear channel extraction technique based on similarities in channel triads is presented. The proposed upmixer surpasses two commercially available upmixers in objective metrics and subjective listening tests, achieving higher spatial consistency while preserving high timbral quality.

1 Introduction

In recent years, both consumption and production of immersive consumer sound has increased [1], yet previously existing content released in stereo or surround format cannot directly take advantage of immersive sound systems. Upmixing is an audio processing technique in which the input content of m channels is mapped to n channels, where $n > m$ [2]. It allows the exploitation of larger sound systems than the legacy content was originally intended for. However, it is challenging to preserve the spatial scene and timbre of the input audio as closely as possible to what the mixing engineer intended without introducing artifacts in an unrealistic immersive sound environment. Consequently, upmixers often prioritize generating aesthetically pleasing results by redistributing certain extracted components rather than attempting full reconstruction.

From now on we will be using *surround* to refer to the 5.1 configuration with 6 channels, following [3], and *immersive* to refer to the 7.1.4 configuration, also known as the 'System J' in [4], comprised of 12 channels. There have been several algorithms proposed to perform upmixing from stereo to surround like [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], but there is not significant work in surround-to-immersive upmixing. Most of the upmixers perform a type of Primary-Ambience Decomposition (PAD), which is a common approach based on a simple signal model (e.g., Eq. 1), to then spatially render the primary component and decorrelate or redistribute the ambience component to fill the auditory space, and the result will be highly dependent on the chosen PAD's assumptions and its relation to the directional rendering process. Nevertheless, it is not straightforward when dealing with more than two input channels. Some studies have tried to circumvent that problem, such as applying pairwise stereo upmixing [16], cascaded intermediate channel upmixing [17], PCA-based PAD [18], statistical PAD [19], or perceptual-based ambience diffusion [20]. However, some problems may arise from these approaches; for instance, PCA might benefit from longer windows to better determine the principal components, unless a cross-correlation-based PCA implementation is used [18]. Another possible issue for classic PAD-based upmixing is when there is more than one primary sig-

nal coming at once from different points in the space. In this case, classic PAD will look just for the dominant signal [17]. While pairwise or cascaded upmixing could effectively address this issue, the constraint of extracting information solely from within a given pair of speakers may pose a limitation.

In the present work we propose a novel surround-to-immersive upmixer by generalizing the classic stereo level-based ambience extraction [8] to the Generalized Equal Levels Ambience Extraction (GELAE), and by introducing a correlation-based rear-channel extraction algorithm. Since the upmixing pipeline is not compute intensive it is suited to run in a real-time application, and both objective and subjective results show the proposed methodology surpasses spatial and timbral quality of two of the private upmixers found in the market.

This paper is structured as follows. Section 2 describes the methodology behind GELAE, the immersive side/rears extraction, and the listening tests. The results of objective metrics and subjective listening tests are shown in Section 3 and discussed in Section 4. Finally, Section 5 summarizes and concludes the paper.

2 Methods

2.1 Generalized Equal-Levels Ambience Extraction

The Equal-Levels Ambience Extraction (ELAE) algorithm, proposed in [8], assumes a simple signal model

$$\vec{X} = \vec{P} + \vec{A}, \tag{1}$$

where $\vec{X}$ is the time-frequency representation of a stereo signal, with $\vec{P}$ and $\vec{A}$ its corresponding primary and ambience components, respectively. Assuming, as well, that the primary components in both channels are correlated and the ambience components have equal energy

$$\|\vec{A}_L\| = \|\vec{A}_R\| = I_A, \tag{2}$$

it is possible to extract the ambience through a soft-mask α

$$\vec{A} = \alpha \vec{X} \tag{3}$$

with the help of the relation

$$I_A^2 = \frac{1}{2}\left(r_{LL} + r_{RR} - \sqrt{(r_{LL} - r_{RR})^2 + 4|r_{LR}|^2}\right), \tag{4}$$

where r_{ij} is the cross-correlation between signals from channels i and j. In the case of the application we are developing (i.e. surround-to-immersive upmixing) it would be useful to have an ambience extraction process in some part of the pipeline to be able to use such ambience to fill the immersive space.

There are at least two approaches to apply ELAE for surround mixes. For instance: i) Use ambience from L and R for the frontal scene and from Ls and Rs for the rear scene; ii) implementing pair-wise ambience extraction, where ELAE is applied iteratively to pairs of contiguous speakers. However, those methods could represent some issues such as: i) losing some ambience information from the laterals and skipping all ambience information from C; and ii) leading to overestimation when identical ambience signals are coming from more than two speakers at a time. Therefore, there is a need for a multi-channel (3 or more) ambience extraction method able to overcome such issues.

Here we extend ELAE's capacity to work with an arbitrary number of channels N through a Generalized Equal-Levels Ambience Extraction (GELAE) algorithm. First, let's take the assumption in Eq. 2 to all channels

$$\|\vec{A}_i\| = I_A \qquad i = 1, ..., N. \tag{5}$$

Now, notice that Eq. 4 holds for any pair of channels

$$I_A^2 = \frac{1}{2}\left(r_{ii} + r_{jj} - \sqrt{(r_{ii} - r_{jj})^2 + 4|r_{ij}|^2}\right), \tag{6}$$

where $i, j \in \{1, ..., N\}$ and $i \neq j$. However, we are interested in the solution of I_A that includes information from all channels, considering the constraint in Eq. 5. To achieve such result we add up I_A^2's from different channel-pairs and solve for I_A^2. Thus, we have the generalized equation

$$I_A^2 = \frac{1}{\binom{N}{2}}\left[\sum_{i=1,j=1}^{N-1,N}\left(2r_{ii} - \frac{1}{2}\sqrt{(r_{ii} - r_{jj})^2 + 4|r_{ij}|^2}\right)\right] \tag{7}$$

where $\binom{n}{k}$ is the binomial coefficient for n choose k. Finally, from Eq. 3 and Eq. 5, we can find the mask for each channel i with

$$\alpha_i = \frac{\|\vec{A}_i\|}{\|\vec{X}_i\|} = \frac{I_A}{\|\vec{X}_i\|}. \tag{8}$$

2.2 Immersive Rear Channels Extraction

To distinguish channels from the surround input to those of the immersive output, we will use the subscript 5 when referring to the former and the subscript 7 when referring to the latter. Therefore, C_5 is the center channel of the surround input while C_7 is the center channel of the immersive output. An illustration of the process to be described in this section can be found in Fig. 1.

We hypothesized that the considerable physical distance between the frontal channels (L_5, C_5, R_5) and the surround rear channels (Ls_5, Rs_5) might compromise the spatial image of the sound if information were extracted from L_5 and R_5 to be relocated in Ls_7, Lr_7 and Rs_7, Rr_7, respectively. Therefore, instead of drawing energy from the front channels, we exclusively utilize the energy already present in Ls_5 and Rs_5 to generate the new immersive side and rear channels. However, we take into account the similarities between the rear and front-lateral channels ($L_5 - Ls_5$ or $R_5 - Rs_5$), as well as between both rear channels ($Ls_5 - Rs_5$). By selecting channel triplets from the set $\{(L_5, Ls_5, Rs_5), (R_5, Rs_5, Ls_5)\}$ and comparing the front-lateral with the rear-contralateral similarities at each time step, we can extract the corresponding immersive rear and side channels.

We used cross-correlation as a similarity measure. For example, to construct Ls_7 and Lr_7 we take the first triplet (L_5, Ls_5, Rs_5). We compare the cross-correlation of Ls_5 and L_5 ($r_{Ls_5L_5}$) with the cross-correlation of Ls_5 and Rs_5 ($r_{Ls_5Rs_5}$), so when the former is larger the corresponding time-frequency tile would be assigned to Ls_7. On the other hand, if the latter is larger, that tile is assigned to belong to Lr_7. And in the case they are both equal, it is assigned to both Ls_7 and Lr_7. Such process can be done through masking, as well. The extracting mask for Ls_7 is defined as

$$\beta_{Ls_7} = \begin{cases} 1 & r_{Ls_5L_5} \geq r_{Ls_5Rs_5} \\ q & r_{Ls_5L_5} < r_{Ls_5Rs_5} \end{cases}, \tag{9}$$

where $q = 0.1259$ is a constant value corresponding to the attenuation (-18 dB) for each time-frequency tile. Similarly, the corresponding mask for Lr_7 is

$$\beta_{Lr_7} = \begin{cases} 1 & r_{Ls_5L_5} \leq r_{Ls_5Rs_5} \\ q & r_{Ls_5L_5} > r_{Ls_5Rs_5} \end{cases}. \tag{10}$$

The derivation of the masks for Rs_7 and Rr_7 taking the second triplet (R_5, Rs_5, Ls_5) is straightforward.

$$\beta_{Rs_7} = \begin{cases} 1 & r_{Rs_5R_5} \geq r_{Rs_5Ls_5} \\ q & r_{Rs_5R_5} < r_{Rs_5Ls_5} \end{cases}, \tag{11}$$

$$\beta_{Rr_7} = \begin{cases} 1 & r_{Rs_5R_5} \leq r_{Rs_5Ls_5} \\ q & r_{Rs_5R_5} > r_{Rs_5Ls_5} \end{cases}. \tag{12}$$

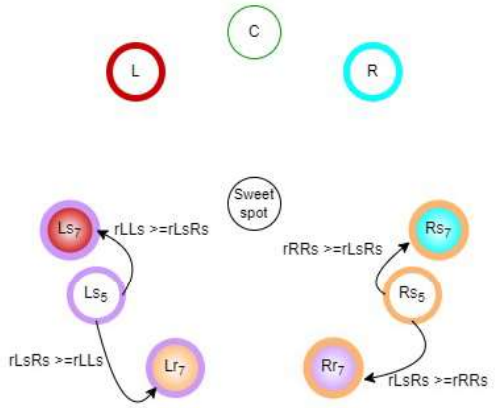


Fig. 1: Immersive Rears Extraction. The red color in Ls_7 represents the assignment of Ls_5 values when correlation of Ls_5 with L is larger than correlation with Rs. And the orange color in Lr_7 represents keeping Ls_5's values when correlation with Rs is larger than correlation with L. Similarly, the colors in Rs_7 and Rr_7 shows their dependence on correlation of Rs_5 with R and Ls.

2.3 Upmixing

A high-level flow diagram of the upmixing process is depicted in Fig. 2. Starting with a surround audio input, the first step is to get its time-frequency representation using the Short-Time Fourier Transform (STFT) with a Hanning window, $NFFT = 1024$, and 75% overlap. Afterwards, we extract the surround ambience and the primary components through GELAE (described in section 2.1). The extracted surround ambience is then diffused, randomly spread to generate 11 new channels, by multiplying a random matrix $O \in \mathbb{R}^{11x5}$ times the ambience vector in time t $A(t) \in \mathbb{C}^{Kx5}$, where K is the number of frequency bands in the STFT, similarly to

what is done in [19], taking care that rows of O have unit norm. In parallel, using the primary components from channels L_5, Ls_5, R_5 and Rs_5 we extract the new channels Ls_7, Lr_7, Rs_7 and Rr_7 with the algorithm described in Section 2.2. Additionally, we use the primary components of L_5, C_5 and R_5 as the new channels L_7, C_7 and R_7. We also used traditional stereo ELAE[8] on L_5 and R_5 to get frontal ambience sent to HFL_7 and HFR_7, and on Ls_5 and Rs_5 to send rear ambience to HRL_7 and HRR_7. Finally, we add up the diffused ambience channel-wise, and perform the Inverse STFT (ISTFT) to get the time-domain representation of the upmixed signal.

Additionally, a variation of the *proposed* upmixer was included in the test conditions. This variant, referred to as $proposed_{noELAE}$, consists of taking out the stereo ELAE (dashed lines in Fig. 2), used to send front/rear ambience to the height channels, to gain insight on whether or not is it relevant to preserve some of the spatial imagery even in the ambience components, and not only the diffused ambience across all the channels.

The upmixing pipeline was created using a frame-by-frame approach, where the cross-correlation was calculated at each time frame using a recursive method similar to [8]. Its Real-Time Factor[1] (RTF) is 0.37x, which was tested on a Core i7-7700HQ @ 2.8 GHz CPU with 16 GB of RAM. The total number of floating point operations (FLOPS) required for the whole process is approximately 2600, estimated through the *Counting the Floating Point Operations (FLOPS)* [21] MATLAB library.

2.4 Subjective evaluation

A listening test was conducted to assess the performance of the two proposed methods ($proposed_{noELAE}$ and *proposed*) against two commercially available upmixers (Dolby Surround and Neural:X).

The listening test audio material consisted of cinematic extracts, originally mixed in immersive format (7.1.4). These samples were downmixed to surround (5.1) and upmixed back to 7.1.4 using different methods to obtain

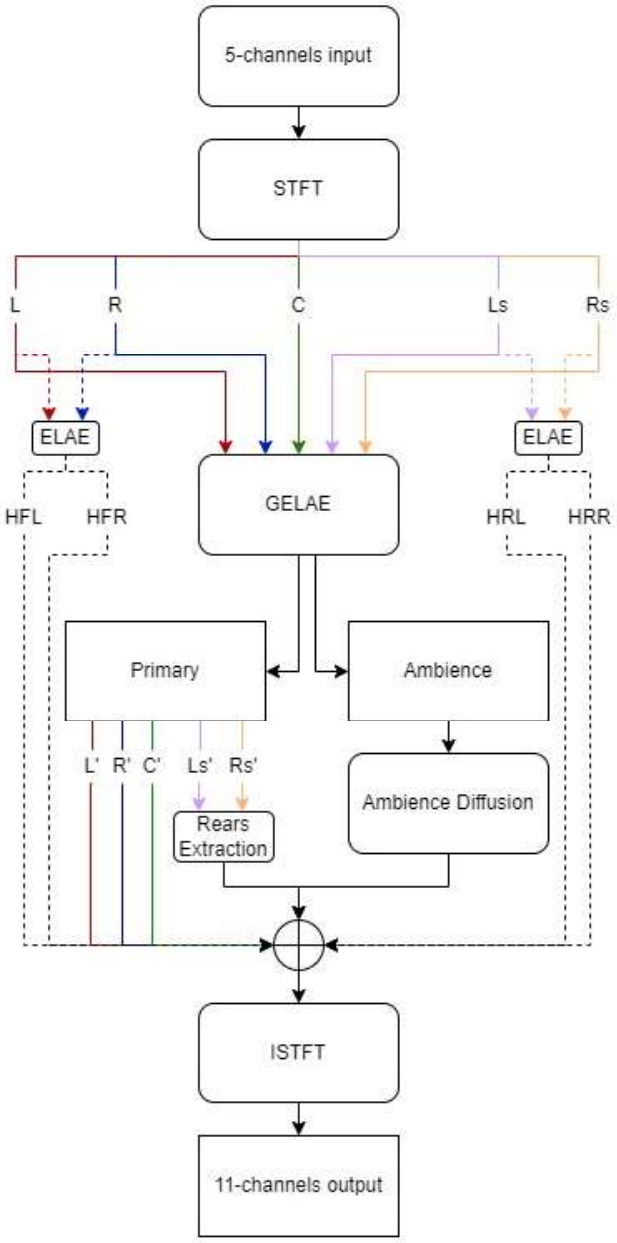


Fig. 2: Flow diagram of the *proposed* surround-to-immersive upmixer. The dashed lines show the paths avoided in the $proposed_{noELAE}$ variation.

[1]Real-Time Factor is defined as the ratio of the computation time λ of a given process to the duration d of the processed signal. $RTF = \frac{\lambda}{d}$.

the test stimuli. The downmixing was done through a passive matrix:

$$\begin{bmatrix} L_7 \\ R_7 \\ C_7 \\ Ls_7 \\ Rs_7 \\ Lr_7 \\ Rr_7 \\ HFL_7 \\ HFR_7 \\ HRL_7 \\ HRR_7 \end{bmatrix}^T \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ \frac{1}{\sqrt{2}} & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{\sqrt{2}} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{\sqrt{2}} & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{\sqrt{2}} \end{bmatrix} = \begin{bmatrix} L_5 \\ R_5 \\ C_5 \\ Ls_5 \\ Rs_5 \end{bmatrix}^T \tag{13}$$

where T represents transposition.

In the case of the proposed upmixers, the surround files were upmixed to 7.1.4 using the *proposed* and $proposed_{noELAE}$ methods previously defined. Regarding the commercial upmixers, the surround files were sent to an Emotiva RMC 1 pre-amplifier, first with the Dolby Digital Surround upmixer setting, and then with the DTS Neural:X upmixer setting. Each of the 12 channels of the resulting audio hardware output was converted to digital by an RME M-32 AD, and sent to a laptop running Nuendo 12 through an AVB network. All channels were then combined to obtain both Dolby Surround and Neural:X upmixed audio files.

A MUSHRA (Multiple Stimuli with Hidden Reference and Anchor) [22] methodology was adopted to compare the following conditions:

- hidden reference (*H_REF*): original 7.1.4 samples
- the *proposed* method (*Sys_A*)
- the $proposed_{noELAE}$ method (*Sys_B*)
- Dolby Digital Surround upmixer
- DTS Neural:X upmixer
- low anchor - 3.5kHz low-pass filtered version of the reference, downmixed to stereo (*L_ANC*).

The two commercial upmixers will be referred to as *Sys_C* and *Sys_D*, with no particular order.

The listening test was divided in two sessions, each focusing on one of the following rating attributes: "Timbral quality" and "Spatial quality". The order of presentation of the test sessions were evenly distributed among the panel of assessors. In each session, the listeners were asked to compare the listening conditions to the known reference and to rate them on the MUSHRA 100-point scale according to the corresponding attribute.

Table 1: Description of tracks used for benchmark.

Label	Content	Dur.
SAMP1	Concert scene. Male singer voice, with surrounding crowd clapping and singing.	10s
SAMP2	Gathering of people cheering, with surround motion of firework.	32s
SAMP3	Children playing in a forest, with sound effects, music and narration voice.	15s
SAMP4	Steam train whistles and sound effects.	15s
SAMP5	Family members arguing in a reverberant space, with steps and sound effects heard moving around and upstairs.	21s
SAMP6	Children running around in a forest, with elevated narration voice.	15s
SAMP7	Steam train whistling and getting in motion with many sound effects.	15s
SAMP8	Army marching with horn blaring.	21s

The selection of 8 cinematic samples used as listening test material is listed in Table 1. Samples 1 to 4 were used in the Timbral quality test, while samples 5 to 8 were presented in the Spatial quality test. Each test session included a repetition, resulting in a total of eight trials per test session.

Listening tests took place in a 7 m (L) x 5.33 m (W) x 3.05 m (H) listening room equipped with 11 loudspeakers and 4 subwoofers supporting 7.1.4 playback. The positions of the speakers and the listeners were based on the ITU-R BS.2051-2 [23] standard. The loudspeakers were leveled at the listener's position and met the ITU-R BS.1116-3 [24] specification in terms of room response curve within the 50 Hz -16 kHz frequency range. LFE channel was muted for all tracks, including references, and loudness was normalized to -23 dB.

3 Results

Results of objective metrics such as Scale-Invariant Signal-to-Distortion Ratio (SI-SDR)[25], Mean Squared Error (MSE) and Log-Spectral Distance (LSD) aggregated by taking the average across all channels are shown in Table 2.

With respect to the subjective evaluation, fourteen listeners took part in the experiment, ten of them being

Table 2: Average results of objective tests

Upmixer	SI-SDR↑	MSE (10^{-4})↓	LSD (10^{-1})↓
Sys_A	**-1.69 ±1.68**	**2.27 ±2.04**	**1.26 ±0.14**
Sys_B	-3.36 ±2.34	2.28 ±2.15	1.28 ±0.16
Sys_C	-9.05 ±2.27	3.78 ±2.64	1.33 ±0.09
Sys_D	-8.7 ±2.3	3.96 ±1.36	1.33 ±0.08

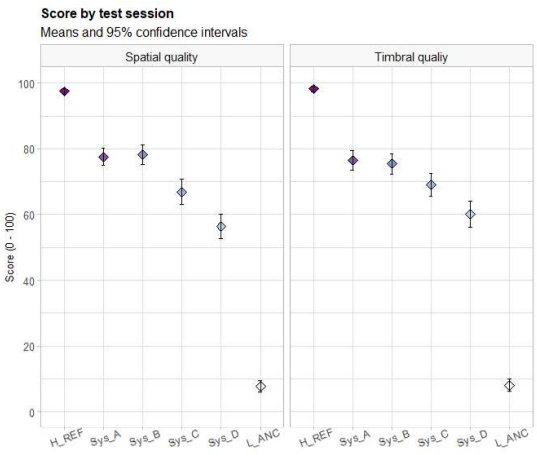


Fig. 3: Overall results

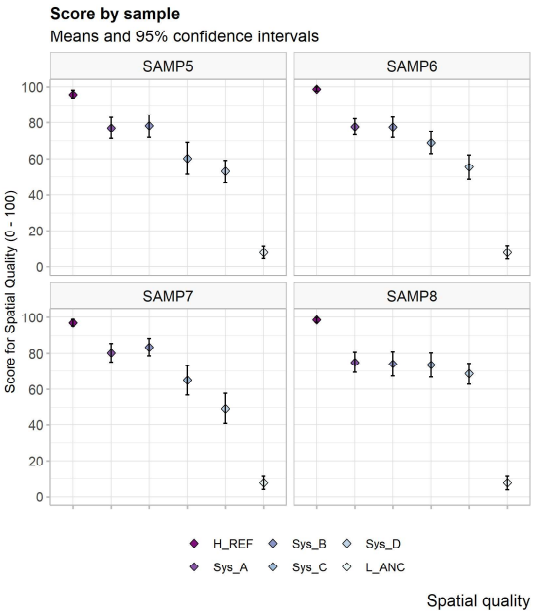


Fig. 4: Results by sample - spatial quality test

experts. The panel of assessors was screened following the MUSHRA and ITU-R BS.2300-0 [26] recommendations. Good reliability and discrimination skills were exhibited by the assessors, and the panel showed agreement.

The data failed to meet the ANOVA assumptions for normally distributed residuals (Shapiro-Wilk test: $p_{spatial} = 7.68 * 10^{-10}$, $p_{timbral} = 6.82 * 10^{-9}$) and homogeneous variance (Levene's test: $p_{spatial} = 1.55 * 10^{-28}$, $p_{timbral} = 1.76 * 10^{-25}$). A Friedman test was run to evaluate the effect of the Condition (6 levels) and Sample (4 levels) factors and their interaction, separately for each attribute. For both tests sessions, a significant effect of the condition factor was revealed ($p_{spatial} = 1.26 * 10^{-12}$, $p_{timbral} = 1.40 * 10^{-12}$), as well as the interaction Condition/Sample ($p_{spatial} = 2.42 * 10^{-41}$, $p_{timbral} = 5.58 * 10^{-41}$). The Sample factor was not found to have a significant effect on the scores ($p_{spatial} = 0.0519$, $p_{timbral} = 0.293$).

A Wilcoxon's signed rank test was used to perform pairwise post-hoc comparisons between conditions. Overall results are presented in Fig. 3. In both test sessions, all conditions were rated significantly lower than the hidden reference, and higher than the low anchor. No significant difference between the proposed methods (*Sys_A* and *Sys_B*) was found. In the spatial quality test, both commercial upmixers (*Sys_C* and *Sys_D*) were rated significantly lower than the proposed methods. In the timbral quality test, although the two commercial upmixers were rated lower than the proposed methods, only *Sys_D* was rated significantly lower.

A closer investigation of the interaction between the Condition and Sample factors revealed that this effect is most likely due to different degrees of discrimination showed by listeners for different samples, although the overall trend is maintained for all samples. In particular, conditions were less differentiated in scores for samples 2, 3 and 8.

Depending on the sample, the proposed upmixing methods performed as good or better than the commercial upmixers tested, for both spatial quality (see Fig. 4) and timbral quality (see Fig. 5) attributes. The two proposed methods were rated very closely. Comments written by the assessors were mostly focused on *Sys_C*

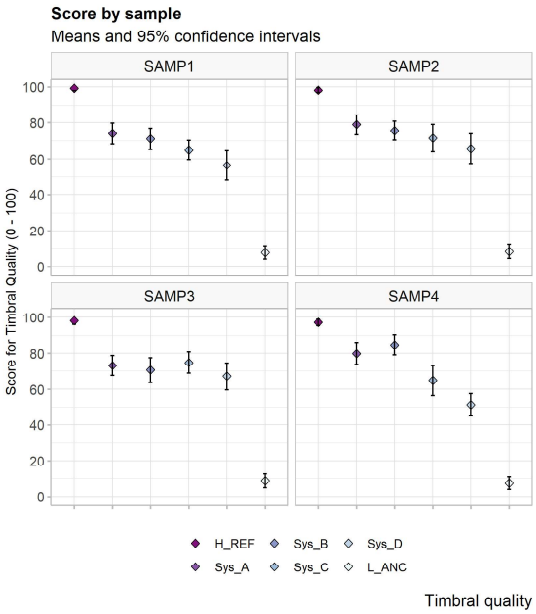


Fig. 5: Results by sample - timbral quality test

and *Sys_D*, with spatial smearing, shifts in localization, and lack of height content. The latter was also mentioned for the *Sys_B* condition.

4 Discussion

Downmixed tracks processed with the *proposed* upmixer tend to be closer to their original 7.1.4 version compared to the other tested upmixers. This is confirmed not only by objective metrics such as SDR, MSE, and LSD, displayed in Table 2, but also by the listening tests where the subjects had identified that the spatial rendering of the sounding sources within each track were placed more accurately while preserving timbre quality.

Through the Immersive Rear channels Extraction algorithm, it has been possible to retrieve a more consistent auditory spatial image, as shown in the listening tests results, by accurately positioning sound sources at the back and side of the user. An advantage of using channel triads instead of pairs is the capacity to extract and relocate spread sound sources that are originally intended to be played back in more than two loudspeakers in the 2D case, or three in the 3D case. Additionally, by limiting the source extraction of primary components to Ls_5 and Rs_5 the final upmixed output preserves more of the original spectral quality. Even more, by using a recursive estimation of the cross-correlation, the computation needed can be kept low, as well as the computation time, so the algorithm can run in real time.

On the other hand, the simultaneous multichannel ambience extraction from GELAE enables the diffusion to be more realistic than doing pairwise stereo ambience extraction and then redistributing such signal. Listening tests results show that sending anterior/posterior stereo ambience to height channels (Sys_A) does not improve the spatial immersive experience, suggesting that diffused surround ambience preserves more consistently the spatial quality of the audio scene than decorrelation and redistribution of a set of extracted stereo ambience, which is what Sys_C and Sys_D seem to be doing. Still, Sys_A achieved a slightly higher score on Timbral quality than Sys_B, since the extra processing of the stereo ambience yields additional artifacts. In addition, some of the assessors commented that Sys_B lacks of elevated content, suggesting a tradeoff between spatial and spectral quality. As future work it would be interesting to test an intermediate system with attenuated stereo ambience for the height channels.

Finally, it is important to notice a limitation associated with the MUSHRA test paradigm. Such test would tend to favor upmixing algorithms that closely resemble the inverse of the employed downmixing process, resulting in outputs that are closer to the reference material. Future research could explore alternative evaluation methodologies, including preference tests, to deliver a more comprehensive assessment of upmixing algorithms' performance across diverse scenarios.

5 Conclusion

In this work we presented a surround-to-immersive upmixer, including two novel techniques: i) a Generalized multi-channel Equal-Levels Ambience Extractor, which allows us to extract more consistently the ambience component from all channels at once, to then diffuse it to the new immersive space, and ii) an Immersive Rear channels Extractor that takes correlation information from channel triads, with each surround rear channel in the middle, to extract new immersive side and rear channels. The whole upmixing pipeline ran on a frame-by-frame basis in real time (RTF=0.37x). The

proposed upmixer surpasses other two commercial upmixers on spatial consistency and spectral quality in both objective metrics and subjective listening tests.

6 Aknowledgements

The authors would like to express their gratitude to the participants of the listening tests.

References

[1] ITU-R BS.2159-9, "Multichannel sound technology in home and broadcasting applications," Standard, International Telecommunication Union, Geneva, CH, 2022.

[2] Páez, T., Tejeda, C., Souza, E., Bharitkar, S., and Madrid, L., "Deep Learning Based Voice Extraction and Primary-Ambience Decomposition for Stereo to Surround Upmixing," *Proceedings of the 154 Audio Engineering Society Convention*, 2023.

[3] ITU-R BS.775-4, "Multichannel stereophonic sound system with and without accompanying picture," Standard, International Telecommunication Union, Geneva, CH, 2022.

[4] ITU-R BS.2051-3, "Advanced Sound System for Programme Production," Standard, International Telecommunication Union, Geneva, CH, 2022.

[5] Faller, C., "Matrix Surround Revisited," *Audio Engineering Society 30th International Conference*, 2007.

[6] Avendano, C. and Jot, J.-M., "Ambience extraction and synthesis from stereo signals for multichannel audio up-mix," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pp. II–1957, IEEE, 2002.

[7] Avendano, C. and Jot, J.-M., "Frequency domain techniques for stereo to multichannel upmix," in *Audio Engineering Society Conference: 22nd International Conference: Virtual, Synthetic, and Entertainment Audio*, Audio Engineering Society, 2002.

[8] Merimaa, J., Goodwin, M., and Jot, J.-M., "Correlation-Based Ambience Extraction from Stereo Recordings," in *Audio Engineering Society 123rd Convention*, Audio Engineering Society, 2007.

[9] Baek, Y.-H., Jeon, S.-W., Park, Y.-c., and Lee, S., "Efficient primary-ambient decomposition algorithm for audio upmix," in *Audio Engineering Society Convention 133*, Audio Engineering Society, 2012.

[10] Irwan, R. and Aarts, R. M., "Two-to-five channel sound processing," *Journal of the Audio Engineering Society*, 50(11), pp. 914–926, 2002.

[11] Ibrahim, K. M. and Allam, M., "Primary-ambient extraction in audio signals using adaptive weighting and principal component analysis," in *Proc. Sound and Music Computing Conference (SMC)*, pp. 227–232, 2016.

[12] Kraft, S. and Zölzer, U., "Stereo signal separation and upmixing by mid-side decomposition in the frequency-domain," in *18th International Conference on Digital Audio Effects (DAFx)*, 2015.

[13] Choi, J. and Chang, J.-H., "Exploiting Deep Neural Networks for Two-to-Five Channel Surround Decoder," *Journal of the Audio Engineering Society*, 68(12), pp. 938–949, 2021.

[14] Park, S. Y., Chun, C. J., and Kim, H. K., "Subband-based upmixing of stereo to 5.1-channel audio signals using deep neural networks," in *2016 International Conference on Information and Communication Technology Convergence (ICTC)*, pp. 377–380, IEEE, 2016.

[15] Yang, H., Wager, S., Russell, S., Luo, M., Kim, M., and Kim, W., "Upmixing via style transfer: a variational autoencoder for disentangling spatial images and musical content," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 426–430, IEEE, 2022.

[16] He, J. and Gan, W.-S., "Applying Primary Ambient Extraction for Immersive Spatial Audio Reproduction," in *Proceedings of APSIPA Annual Summit and Conference*, Asia-Pacific Signal and Information Processing Association, 2015.

[17] Faller, C., Altmann, L., Levison, J., and Schmidt, M., "Multi-Channel Ring Upmix," *Audio Engineering Society 134th Convention*, 2013.

[18] Goodwin, M. M. and Jot, J.-M., "Primary-ambient signal decomposition and vector-based localization for spatial audio coding and enhancement," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, volume 1, pp. I–9, IEEE, 2007.

[19] Vinton, M., McGrath, D., Robinson, C., and Brown, P., "Next Generation Surround Decoding and Upmixing for Consumer and Professional Applications," in *Audio Engineering Society Conference: 57th International Conference: The Future of Audio Entertainment Technology–Cinema, Television and the Internet*, Audio Engineering Society, 2015.

[20] Lee, H., "Perceptually Motivated 3D Diffuse Field Upmixing," *Audio Engineering Society Conference on Sound Field Control*, 2016.

[21] Qian, H., "Counting the Floating Point Operations (FLOPS)," 2024.

[22] ITU-R BS.1534-2, "Method for the Subjective Assessment of Intermediate Quality Level of Audio Systems," Standard, International Telecommunication Union, Geneva, CH, 2014.

[23] ITU-R BS.2051-2, "Advanced Sound System for Programme Production," Standard, International Telecommunication Union, Geneva, CH, 2018.

[24] ITU-R BS.1116-3, "Methods for the Subjective Assessment of Small Impairments in Audio Systems," Standard, International Telecommunication Union, Geneva, CH, 2015.

[25] Le Roux, J., Wisdom, S., Erdogan, H., and Hershey, J. R., "SDR half-baked or well done," in *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP*, IEEE, 2019.

[26] ITU-R BS.2300-0, "Methods for Assessor Screening," Standard, International Telecommunication Union, Geneva, CH, 2014.