CAPTURE, REPRESENTATION, AND RENDERING OF 3D AUDIO FOR VIRTUAL AND AUGMENTED REALITY

Ivan J. Tashev

Microsoft Research, One Microsoft Way, Redmond WA 98052
e-mail: ivantash@microsoft.com
USA

Abstract: Devices for augmented and virtual reality (AR/VR) find applications in gaming, entertainment, education, design and science. An AR/VR headset consists of a head-mounted display, headphones or loudspeakers, and a computing platform. The spatial audio system plays integral party of achieving the realism. In this article we present an overview of our work on technologies for the 3D audio targeting AR/VR scenarios.

Key words: spatial audio, augmented reality, virtual reality.

1. INTRODUCTION

Virtual reality (VR) is a set of technologies to create a perception in the human for being in another place, both visually and acoustically [1]. With first attempts dating as early as in 1950s, today it is a well-developed area in computer science and technologies. In augmented reality (AR) the human remains in its own environment, to which audio-visual objects are added (augmented) [2]. If we consider AR and VR as two extremes of the reality-virtuality continuum, then commonly used term is mixed reality (MR). The MR devices consist of head mounted displays to create the visual portion of MR, headphones for reproducing the spatial audio, system for human posture tracking, components of the user interface and computing platform. In this article we will make an overview of technologies for creating spatial audio experience in the MR systems. As such they need input from the head position and orientation tracking system, with requirements usually much lower than needed for the visual part.

Spatial audio is set of devices and signal processing algorithms which can create the perception in listener that the audio comes from any desired position (direction, elevation, and distance). Also, it is called 3D audio. Work on spatial audio reproduction starts with stereophonic (2 channels) audio rendering and continues through quadrophonic (4 channels) to surround sound systems today (5, 7, and more channels). These approaches are from the group of channel-based representation of the

spatial audio. The channel-based approach is widely used in cinema and entertainment. In VR/AR devices today more common are sound object-based and sound field-based representations. Both representations are expected to be independent of the sound capture and sound rendering systems.

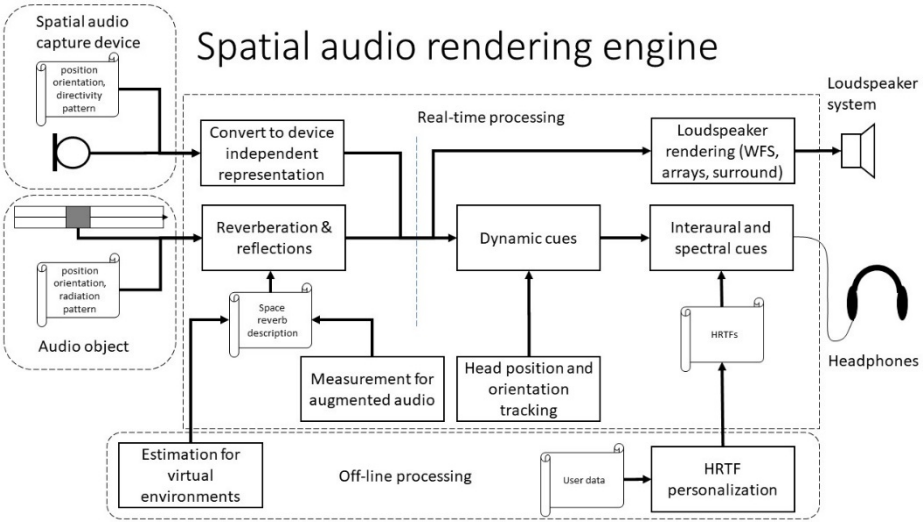


Fig. 1. Block diagram of spatial audio rendering engine.

2. SPATIAL AUDIO REPRESENTATION

A block diagram of a spatial audio rendering engine is shown in fig. 1 with inputs these two spatial audio representations.

2.1. Object-based representation

In object-based representation each sound source is represented by sound stream and properties (position, orientation, radiation pattern). The sound streams are either recorded in studio environment or generated synthetically.

The audio in the sound stream is dry and to sound realistic proper reverberation should be added by convolving with a room impulse response. The impulse responses are difficult to manipulate, they use substantial amount of memory and can't be interpolated. Therefore, the reverberation process in the room is represented by a set of reverberation parameters and synthesized from them during applying the reverberation. Such parameters are reverberation time [3], room volume [4], echo density [5], etc.

The impulse response characterizes the reverberation between two points in the space, with given radiation pattern and orientation of the source and directivity pattern and orientation of the receiver [6]. Measuring these in a dense grid with all combinations of radiation and directivity patterns and their orientation is practically an

impossible task. The spatial impulse responses can be measured using a spherical loudspeaker array and a spherical microphone array and later used to apply the radiation pattern and orientation of the source and the directivity pattern and orientation of the receiver [7].

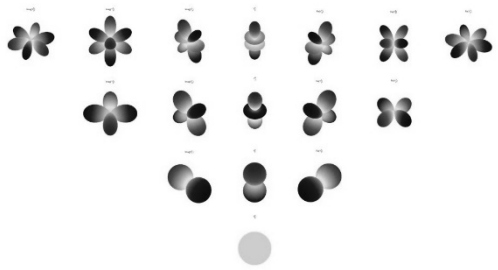


Fig. 2. Real part of the spherical harmonics

In VR environments typically we have the space geometry and information of the acoustical characteristics of the materials. This allows estimation of the reverberation parameters to happen off-line [8] for a grid of points and do interpolation on the fly in runtime. This approach is productized in Project Acoustics [9].

In AR scenarios the reverberation parameters should be measured in real-time using the integrated into the device microphones and properly reflected in the reverberation utilizing the approaches above.

In MR computer game we can have hundreds of sound objects and the computational costs increase linearly with the number of objects.

2.2. Sound field-based representation

In the sound field representation, the spatial audio is described as a decomposition of basis functions, one of the most commonly used are spherical harmonics. Given listening point, for each frequency bin the sound field intensity as function of the direction and elevation can be described as a closed surface $\mathrm{H}(\beta,\alpha,\omega)$, function of the direction α, the elevation β and the angular frequency ω. This surface can be represented by the coefficients $\breve{\mathrm{H}}_n^m(\omega)$ of a series of spherical harmonics $Y_n^m(\beta,\alpha)$ as [10]:

$$\mathrm{H}(\beta,\alpha,\omega)=\sum_{n=0}^{\infty}\sum_{m=-n}^{n}\breve{\mathrm{H}}_n^m(\omega)Y_n^m(\beta,\alpha) \tag{1}$$

We may apply the Helmholtz reciprocity principle and assume that we observe the sound pressure of a sound source on the surface of a sphere [11]. The surface spherical harmonics $Y_n^m(\beta,\alpha)$ are a complete and orthonormal set and may be defined as:

$$Y_n^m(\beta,\alpha)=(-1)^m\sqrt{\frac{(2n+1)}{4\pi}\frac{(n-|m|)!}{(n+|m|)!}}P_n^{|m|}(\cos\beta)e^{im\alpha} \tag{2}$$

where $P_n^m(\cdot)$ denotes *m*-th order of the associated Legendre function of *n*-th degree.

The eq. (1) converges at certain point and can be approximated by limiting the order *n* to *N* instead of infinity. The real parts of the spherical harmonics coefficients up to fourth order are shown in fig. 2. Spherical harmonics decomposition reflects

the physics of the sound wave propagation - note that the zeroth and first order spherical harmonics are the directivity patterns of the omnidirectional and pressure gradient microphones.

Using eq. (1) we can convert the sound pressure level in given set of points on a rigid sphere into a spherical harmonics' representation. In fig. 3 a) and b) are shown experimental 16-channel and 64-channel spherical microphone arrays, designed to do exactly that. Considering that it is quite possible that the rendering loudspeaker system could be a 5.1 or 7.1 surround sound, where the loudspeakers are in one plane, we can reduce the complexity and the amount of processed data by using cylindrical harmonics, which describe a height invariant sound field. The math is quite like the spherical harmonics and [12] provides ways to convert between them. Fig. 3 c) shows a 16-channel cylindrical microphone array designed to decompose the sound field to cylindrical harmonics representation.

(a) 16 channels 115 mm spherical *(b) 64 channels 200 mm spherical* *(c) 16 channels 115 mm cylindrical*

Fig. 3. Microphone arrays to capture ambisonics audio

Note that obtaining spherical or cylindrical decomposition of the sound field doesn't has to happen with spherical or cylindrical microphone arrays. A microphone system of four microphones: one omnidirectional and three perpendicular figure-8 microphones, placed closely together, performs first order sound field decomposition. The obtained four audio channels are directly the output of eq. (1), where n is limited to 1. This is also known as B-format audio. The same effect, with certain limitations, can be achieved with four omnidirectional microphones, placed not in the same plane. Then we can use beam pattern synthesis approach [13] to synthesize beams with directivity patterns directly matching the spherical harmonics coefficients in eq. (2). Obviously, this approach is limited by the number of microphones, their self-noise and the microphone array geometry.

The decomposed with spherical harmonics sound is a multichannel audio stream, independent of the sound capture device. Converting object-based audio representation to sound field-based representation is also straightforward. The number of channels increases with the order of the spherical harmonics $(N+1)^2$. For third order representations we must have 16 channels, and for seventh order (considered beyond the resolution of human hearing) – 64 channels. This is substantial intensity of the

audio stream. One of the ways to mitigate this is to apply audio compression algorithms, commonly used in entertainment. Another approach is to use a specialized spatial audio compression algorithm, such as DirAC [14] and its further developments. In this spatial audio compression is assumed that in each frequency bin of each audio frame the sound can represented as a single point source and isotropic component.

While more computationally expensive, the sound field representation of the spatial audio scales up better, with pretty much the same computational cost regardless of the number of sound objects.

3. HRTF AND PERSONALIZATION

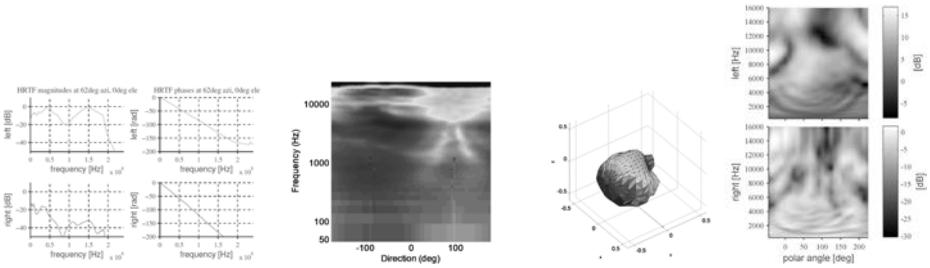
The perception of direction and distance in humans using audio is denoted as spatial hearing. The sound within the audible frequency and dynamic range and is delivered to both ears. To determine the direction and distance to the sound source the brain uses auditory localization cues: interaural time and level differences (ITD and ILD), spectral differences, reverberation and reflections, dynamic and multiradial cues. Also, a lot of prior knowledge about the sounds is utilized. The first two spatial cues are frequently denoted as head-related transfer functions (HRTF). If we have an audio signal $x(t)$ at given position in space, the process of propagation of the sound wave to the human head and the entrances of the ear canals can be modelled with two impulse responses: $h_L(t)$ and $x_R(t)$, as shown in fig. 4 a). The sounds at the entrances of the ears are convolution of the source signal with the impulse responses: $x_L(t) = s(t) * h_L(t)$ and $x_R(t) = s(t) * h_R(t)$. The set of these pairs of filters for all possible points in space is called HRTF.

(a) Sound propagation to the ears *(b) HRTF measurement setup in Microsoft Research* *(c) HRTF measurement points*

Fig. 4. Head related transfer functions measurement

HRTF describe the acoustic path from the sound source to the ear entrances. The pairs contain all interaural and spectral localization cues and these filters are function of the sound location (direction, elevation, and distance). We can consider the HRTF

distance independent for radii greater than one meter. Several representations of HRTF are shown in fig. 5.



(a) frequency for direction 62° and elevation 0° *(b) direction and frequency for elevation 0°* *(c) direction and elevation for 1000 Hz* *(d) elevation and frequency for direction 90°*

Fig. 5. HRTF are function of the direction, elevation, and frequency.

The propagation of the sound from the source to the entrances of the ear canals is affected by the head and torso geometry and the features of the pinna. As they are different in different people the HRTF are considered individual. The personal HRTF can be obtained via direct measurement or via personalization of a generic HRTF.

3.1. Measuring HRTF

The HRTF model the process of sound wave propagation up to the entrances of the ear canals. To measure someone's HRTF is enough to place small microphones at the entrances of the blocked ear canals and record known sounds coming from various directions. This must happen in anechoic environment, to prevent sound reflections from the walls, ceiling and floor to affect the measured impulse responses. Fig. 4 b) shows HRTF measurement setup built in Microsoft Research. The test subject sits in the chair and the head is fixed at the origin of the coordinate system. An arc with 16 loudspeakers rotates around the head and for several minutes HRTF are measured in 400 points in the space shown in fig. 4 c). Note that due to space limitations the bottom part of the points is missing. They are obtained via interpolation, as described in [15]. The same approach is used to interpolate the measured HRTF to 2048 equidistant Fliege points [16].

3.2. HRTF database in Microsoft Research

To enable future work on HRTF personalization in Microsoft Research was built an HRTF database, which consist of data for more than 300 subjects. Besides measured in 400 locations HRTF, using the setup in fig. 4 b), from all subjects was collected anthropometric information. They filled a questionnaire with data about their hat size (head circumvention), shirt collar size (neck circumvention), shirt size (torso size), jeans size (waist circumvention and leg length), and even shoe size. Also, direct anthropometric measurements were conducted to all of them using devices

shown in fig. 6 a). These measurements include head circumvention, width, depth, height; neck and chest circumvention; major bones length in the arms and legs; and the distance between eye pupils. For further work on the ear pinna features the head of all subjects was scanned with sub-millimeter accuracy, as shown in fig. 6 b).

(a) anthropometric measurement devices

(b) high resolution head scan

Fig. 6. HRTF dataset

3.3. HRTF personalization: direct estimation

Given high resolution 3D scan of the head, the estimation of HRTF is a solvable problem. There are numerous released programs for doing so, but all of them are highly computationally expensive, especially in the upper part of the frequency band where we have to have high density of the evaluation points. In search for fast and easy way to generate personalized HRTF was proposed to use magnitudes of a generic HRTF and adjust only the ITD as function of the direction and elevation [16]. The general idea is to use ray tracing for estimation of the propagation path length and from there the interaural time difference as shown in fig. 7. This method doesn't account for the magnitude variations based on the head shape and time delay altered by pinnae features.

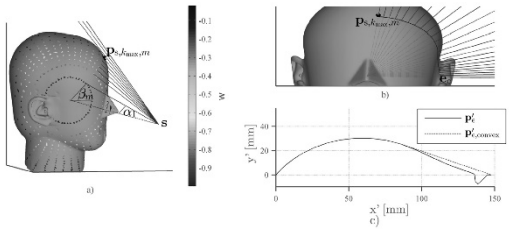


Fig. 7. HRTF personalization via direct estimation

3.4. HRTF personalization: anthropometrics based

The propagation of the sound wave is a process that depends only on the shape of the head and pinnae features. This means that two people with similarly shaped heads and pinnae will have similar HRTF. The general idea behind HRTF personalization using anthropometric data is, given certain number of anthropometric measures of the person, to find from the dataset several persons with similar anthropometries and synthesize the personalized HRTF from these HRTF. The process for personalization of magnitudes [18] is illustrated in fig. 8. Given set of anthropomet-

ric measurements of the person, they are represented as a combination of the anthropometries from the database using non-negative sparse representation. This results in a set of non-zero weights for several people from the database with closest anthropometries. Then we synthesize the magnitudes of the personalized HRTF as a weighted sum of the HRTF magnitudes using the same weights. The problem with phases is more complex because of the small variations caused by different pinnae features. One of the ways to solve it is to ignore these phase variations and assume constant across frequencies time delay for each direction and elevation.

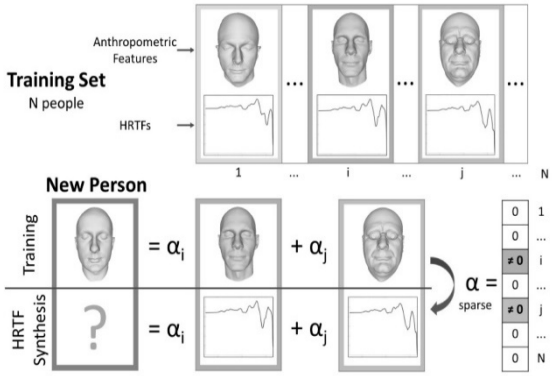


Fig. 8. HRTF personalization based on anthropometric Measurements

Fig. 9 shows the delay contour averaged from all persons in the dataset. Then the personalization of the HRTF phases is reduced to finding a scale factor for this contour – larger than one for people with bigger heads, smaller than one for people with smaller heads. This approach uses the same non-negative sparse representation and is described in [19]. Obviously, more anthropometric measurements mean better match of the personalized HRTF. Both approaches relay on the assumption that the same weights from anthropometric measures representation can be used for HRTF magnitudes and ITD scaling factor.

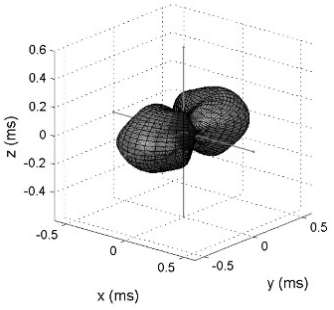


Fig. 9. Average interaural Delay contour

3.5. HRTF personalization: parametrization based

The main problem with the anthropometric measurements approach is obtaining them from the subject. Measuring large number of them can be as annoying and time consuming as direct measurement of the HRTF. These anthropometries can be obtained by processing of a high-quality 3D head scan. A more robust and generic solution is interpolation of the 3D head scan using spherical harmonic transforms (fig. 10) and use a distance metric for finding the closest HRTF from the database.

The results from HRTF personalization using spherical harmonics transform, spherical Fourier-Bessel transform, and spherical oscillator transform are presented in [20].

Rotating the subject in front of a depth camera, such as the one in Kinect, is an easy and quick way make a coarse 3D head scan. The precision of the obtained anthropometries is lower, and this affects the personalized HRTF. Fitting an ellipsoid or sphere over this head scan can be done relatively precise (fig. 11). The parameters of the ellipsoid then can be used for synthesizing of the personalized HRTF. The number of these parameters is small and not enough for synthesizing of the HRTF magnitudes but works well for estimation of the scaling factor [21].

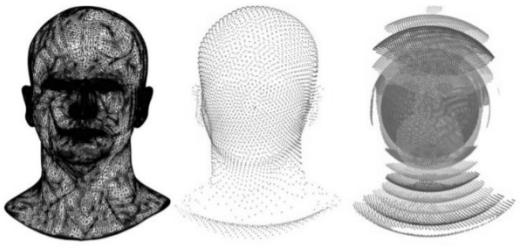


Fig. 10. Interpolation with spherical harmonics: original scan (left), raytracing intersections (center), and coarsely sampled example (right)

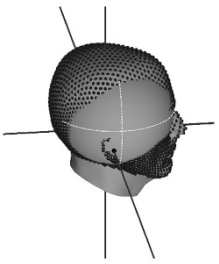


Fig. 11. Sphere fitted to3D head scan

3.6. HRTF personalization: average face based

A simple 3D picture of the face, or incomplete scan, can be obtained with a depth camera for a fraction of the second. Direct estimation of anthropometries might not be possible, and parametrization can be imprecise, thus hurting the quality of the personalized HRTF. Another approach is to use a face template, lay it over the incomplete head scan, and start deforming it to fit the incomplete scan. Fig. 12 shows the average face deformed over the incomplete scan. The deformation parameters can be used as input features for a machine learning-based estimator of the personalized HRTF. This approach is used in [22] to obtain the ITD scaling factor with good results.

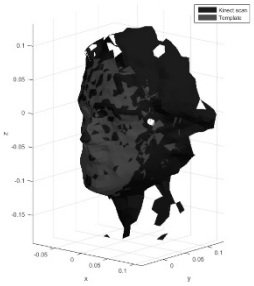


Fig. 12. Average face fitted to depth image

3.7. HRTF personalization: evaluation

Human spatial hearing is a complex process that involves the two hearing organs and the human brain. HRTF are part of this process, some of the HRTF features are

more important for spatial localization of the sound source that the others. This is illustrated in [23], where a convolutional neural network is trained to perform sound source localization using binaural recordings. Backpropagation of the weights uncover which features of the HRTF are more important for the localization task. In addition, not all directions are equality important, and this is reflected in the accuracy of sound source localization by humans. To date we do not have psychoacoustics-based distance measure between two HRTF to use as a quality metric of the HRTF personalization. In scientific literature are used numerous objective metrics with the understanding that they do not reflect the perceptual distance.

3.8. HRTF personalization: evaluation

Human spatial hearing is a complex process that involves the two hearing organs and the human brain. HRTF are part of this process, some of the HRTF features are more important for spatial localization of the sound source that the others. This is illustrated in [23], where a convolutional neural network is trained to perform sound source localization using binaural recordings. Backpropagation of the weights uncover which features of the HRTF are more important for the localization task. In addition, not all directions are equality important, and this is reflected in the accuracy of sound source localization by humans. To date we do not have psychoacoustics-based distance measure between two HRTF to use as a quality metric of the HRTF personalization. In scientific literature are used numerous objective metrics with the understanding that they do not reflect the perceptual distance.

Table 1. Comparison of HRTF personalization approaches.

Algorithm	Features	LSD, *dB*	RMSE, *ms*
HRTF Magnitude Synthesis via Sparse Representation of Anthropometric Features	92	4.5	
	21	6.7	
HRTF Phase Synthesis via Sparse Representation of Anthropometric Features	47		0.083
Estimation of Multipath Propagation Delays and Interaural Time Differences from 3-D Head Scans	Precise 3D head scan		0.019
Anthropometric Parametrization of a Spherical Scatterer ITD Model with Arbitrary Ear Angles	Head scan		0.049
Interaural Time Delay Personalization Using Incomplete Head Scans	3D Face scan		0.036

The most commonly used distance metric is log-spectral distance (LSD), followed by the ITD match. A comparison of the discussed methods for HRTF personalization is presented in Table 1. The sparse representation approach provides good fit of 4.5 *dB* with 92 anthropometric features, which degrades to 6.7 *dB* when using 21 anthropometric features. The same approach for phases, using 42 features, gives 0.083 *ms* ITD error. Multipath estimation from a precise head scan performs the best,

as it estimates exactly the ITDs. The other two methods for estimating of the ITD contour also provide satisfactory results. Which one to be used depends on what user data is available. In some cases, specific issues with personalized HRTF must be addressed, such as front-back or up-down confusions, or elevation perception [24], [25]. In all cases the quality of the HRTF personalization must be verified by conduction listening tests.

4. SPATIAL AUDIO RENDERING

Sound rendering systems vary from 832-channels wave field synthesis in [26], through modern movie theaters equipped with Dolby Atmos systems, through home theaters with 5.1 surround sound, or even only stereo speakers. Another class of rendering system at home are the loudspeaker arrays. The advantage is that they are typically mounted under the TV set, there is no need to install loudspeakers on the sides or in the back. The initial designs used focused beams of sound, pointed at the walls. These systems counted on one or two reflections to mimic the side and back loudspeakers. Others just focused the sound to reduce the noise pollution [26]. Sending two separate beams for the left and the right ear of a head in front of the loudspeaker array [28], allows creation of headphones-sans-headphones and utilization of the spatial audio techniques. With the existing loudspeaker array technologies today, it is possible to design crosstalk cancellation enough good to utilize the spatial audio techniques and create usable spatial audio system. The necessary head orientation and tracking system can be implemented using a depth camera as the one in Kinect.

Any of these loudspeaker systems can be used for rendering of spatial audio representations, described in section 2. In most of the cases we do not have enough dense set of loudspeakers to avoid spatial aliasing problems. The same is valid for representation of the sound field using high order ambisonics. These deficiencies of the rendering system, or of the model, cause listening artifacts. They can be partially compensated for by using various technics. One example for such compensation for spatial audio represented with high order ambisonics is presented in [29]. The spatial aliasing is the reason that in many cases wave field synthesis systems, or loudspeaker arrays, do the processing in two bands. The lower band, where we do not have spatial aliasing, is typically the better performing algorithm. In the upper frequency band is used algorithm more robust to this problem [30].

In VR/AR devices the sound rendering system is typically limited to headphones for the left and the right ear. This requires compensation for the head rotation (to keep the sound scene still) and applying generic or personalized HRTF. Mixed reality devices have head orientations sensor anyway – they must compensate for head rotation in the video channel as well. The precision of these head orientation sensors is way above the required for audio; in the case of augmented reality devices it is below one-pixel resolution.

Even in everyday scenarios, such as listening to stereo music, the spatial audio rendering approaches can bring value. The stereo sound is designed to be listened to through a pair of loudspeakers in front of the listener, approximately 90° apart. When a stereo sound is played through a pair of headphones the sound scene is perceived as inside the head. Plus, rotation of the head causes rotation of the entire audio scene. If we assume that the two loudspeakers are two audio objects in front of the listener and use spatial audio approaches for rendering the stereo music, the audio scene is perceived in front of the listener. Such externalization improves greatly the experience of listening to stereo music. To compensate for head movements, we will need a head orientation sensor. It can be either external (by head orientation tracking with a depth camera, such as Kinect), or internal (integrated into the headphones inertial measurement unit). Headphones with integrated head orientation sensors are already on the market, together with spatial audio rendering engines. They can render stereo, spatial audio (5.1 and 7.1), and 3D audio, greatly improving the listing experience.

5. CONCLUSIONS

In this article we discussed technologies for capturing, representation, and rendering of spatial audio. In general, this is a set of technologies to make the listener to perceive the sound coming from any desired direction. Also known as 3D audio, it found its increased demand with the advancement of mixed reality devices. The technologies for end-to-end implementation of the spatial audio exist, some of them already found their way to commercial products. Still, many of them need further polishing and development. While mainstream in mixed realty scenarios, the spatial audio has yet to find its way to film and art. The audio community lacks good tools for authoring and editing of spatial audio. Applications of the spatial audio techniques go way beyond the mixed reality scenarios, they are applicable to everyday scenarios, such as listening to stereo music and watching movies with surround sound. They are especially valuable in mobile devices, where can compensate the small screen with high quality spatial sound.

REFERENCES

[1] L. B. Rosenberg, “The Use of Virtual Fixtures As Perceptual Overlays to Enhance Operator Performance in Remote Environments,” Technical Report AL-TR-0089, USAF Armstrong Laboratory, Wright-Patterson AFB OH, 1992.

[2] Ronald Azuma, *Presence: Teleoperators and Virtual Environments*, vol. 6, chapter A Survey of Augmented Reality, pp. 355–35, August 1997.

[3] Hannes Gamper and Ivan Tashev, “Blind reverberation time estimation using a convolutional neural network,” in *International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, September 2018.

[4] Andrea Genovese, Hannes Gamper, Ville Pulkki, Nikunj Raghuvanshi, and Ivan Tashev, “Blind room volume estimation from single-channel noisy speech,” in *Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, May 2019.

[5] Helena Peic Tukuljac, Ville Pulkki, Hannes Gamper, Keith Godin, Ivan Tashev, and Nikunj Raghuvanshi, "A sparsity measure for echo density growth in general environments," in *Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, May 2019.

[6] Ivan J. Tashev, Hannes Gamper, and Lyle Corbin, "Modelling and estimation of the spatial impulse response in reverberant conditions," *The Journal of the Acoustical Society of America*, vol. 141, pp. 3749, 2017.

[7] Hannes Gamper, Keith Godin, Nikunj Raghuvanshi, and Ivan J. Tashev, "Characterizing acoustic environments using spherical loudspeaker and microphone arrays," *The Journal of the Acoustical Society of America*, vol. 144, pp. 1881, 2018.

[8] Nikunj Raghuvanshi and John Snyder, "Parametric directional coding for precomputed sound propagation," *ACM Trans. Graph.*, vol. 37, no. 4, pp. 108:1–108:14, July 2018.

[9] Microsoft, "Project Acoustics," https://docs.microsoft.com/enus/azure/ cognitive-services/acoustics/what-is-acoustics. Retrieved June 17, 2019.

[10] N. A. Gumerov and R. Duraiswami, *Fast Multipole Methods for the Helmholtz Equation in Three Dimensions*, Elsevier, Oxford, 2004.

[11] D. N. Zotkin, R. Duraiswami, and N. A. Gumerov, "Regularized HRTF fitting using spherical harmonics," in *Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, October 2009, pp. 257–260.

[12] Mark Thomas, Jens Ahrens, and Ivan Tashev, "A method for converting between cylindrical and spherical harmonic representations of sound fields," in *Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2014.

[13] Harry L. van Trees, *Optimum Array Processing*, Wiley, 2002.

[14] V. Pulkki, M.-V. Laitinen, J. Vilkamo, J. Ahonen, T. Lokki, and T. Pihlajamaki, "Directional audio coding - perception-based reproduction of spatial sound," in *International Workshop on the Principles and Applications of Spatial Hearing*, Zao, Miyagi, Japan, November 2009.

[15] Jens Ahrens, Mark Thomas, and Ivan Tashev, "HRTF magnitude modeling using a non-regularized least-squares fit of spherical harmonics coefficients on incomplete data," in *Annual Summit and Conference APSIPA*, December 2012.

[16] Jörg Fliege and Ulrike Maier, "The distribution of points on the sphere and corresponding cubature formulae," *IMA Journal of Numerical Analysis*, vol. 19, no. 2, pp. 317–334, 1999.

[17] Hannes Gamper, Mark Thomas, and Ivan J. Tashev, "Estimation of multipath propagation delays and interaural time differences from 3-D head scans," in *Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015.

[18] Piotr Bilinski, Jens Ahrens, Mark R.P. Thomas, Ivan J. Tashev, and John C. Platt, "HRTF magnitude synthesis via sparse representation of anthropometric features," in *Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2014.

[19] Ivan Tashev, "HRTF phase synthesis via sparse representation of anthropometric features," in *Information Theory and Applications Workshop*, February 2014.

[20] Archontis Politis, Mark Thomas, Hannes Gamper, and Ivan Tashev, "Applications of 3d spherical transforms to personalization of head related transfer functions," in *Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, March 2016.

[21] Hannes Gamper, Mark Thomas, and Ivan Tashev, "Anthropometric parameterisation of a spherical scatterer ITD model with arbitrary ear angles," in *Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, October 2015.

[22] Hannes Gamper, David Johnston, and Ivan Tashev, "Interaural time delay personalisation using incomplete head scans," in *Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, March 2017.

[23] Etienne Thuillier, Hannes Gamper, and Ivan Tashev, "Spatial audio feature discovery with convolutional neural networks," in *Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018.

[24] Vani G. Rajendran and Hannes Gamper, "Spectral manipulation improves elevation perception with non-individualized head-related transfer functions," *Journal of the Acoustical Society of America (JASA)*, vol. 145, no. 3, pp. 222–228, March 2019.

[25] Hannes Gamper and Ivan J. Tashev, "Improving the perception of a sound source's polar angle in mixed reality," *Journal of the Acoustical Society of America*, vol. 143, April 2018.

[26] C. Moldrzyk, A. Goertz, M. Makarskiand S. Feistel, and S. Weinzierland W. Ahnert, "Wellenfeldsynthese fr einen groen Hrsaal," in *Fortschritte der Akustik*, 2007.

[27] Ivan Tashev, Jasha Droppo, Mike Seltzer, and Alex Acero, "Robust design of wideband loudspeaker arrays," in *Int. Conf. on Audio, Speech and Signal Processing (ICASSP)*, April 2008.

[28] E. Choueiri, "Optimal crosstalk cancellation for binaural audio with two loudspeakers," Princeton University. [Online]. Available: http://www.princeton.edu/3D3A/Publications/BAC-CHPaperV4d.pdf.

[29] Christoph Hold, Hannes Gamper, Ville Pulkki, Nikunj Raghuvanshi, and Ivan Tashev, "Improving binaural ambisonics decoding by spherical harmonics domain tapering and coloration compensation," in *Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2019.

[30] Jens Ahrens, Mark Thomas, and Ivan Tashev, "Gentle acoustic crosstalk cancelation using the spectral division method and ambiophonics," in *Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, October 2013.

Information about the author:

Ivan J. Tashev received his Diploma Engineer degree in Electronic Engineering (1984) and PhD in Computer Science (1990) from the Technical University of Sofia, Bulgaria. He was assistant professor in the same university when joined Microsoft in 1998. Currently Ivan Tashev is a partner software architect and leads the Audio and Acoustics Group in Microsoft Research Labs in Redmond, USA. His research interests include audio signal processing; machine learning; multichannel transducers.

Manuscript received on 15 June 2019