A Mobile App for Geolocalized, Dynamic Binaural Synthesis

Markus Hädrich[1], Alexander Lindau[2], and Stefan Weinzierl[1]

[1] *Audio Communication Group, TU-Berlin, Einsteinufer 17c, 10587 Berlin, Germany*

E-mail: {markus.haedrich, stefan.weinzierl}@tu-berlin.de

[2] *Max Planck Institute for Empirical Aesthetics, Grüneburgweg 14, 60322 Frankfurt am Main, Germany*

E-mail: alexander.lindau@aesthetics.mpg.de

Abstract

Today, technical evolution and dissemination of smartphones have reached a point where powerful mobile computing has become virtually ubiquitous. Additionally, devices are equipped with a multitude of sensors and interfaces allowing for a – potentially continuous – acquisition of optical and acoustical data, the continued determination of bearing and position, and mobile network connectivity. Therefore, smartphones are increasingly used as a platform for mobile augmented or virtual reality applications. As a result, we report here on the development of an augmented audio application featuring Dynamic Binaural Synthesis (DBS) on iOS devices. The purpose of the app is to interactively render singular sound sources – *sound spots* virtually attached to fixed geographical positions – binaurally via headphones, while (currently) using the smartphone's sensors to indicate the user's viewing direction. With the help of a 2D map view, users may place multiple such sound spots in the actual (outdoor) environment - this way, for instance, creating augmented reality soundwalks. A proximity criterion is used to automatically switch the audio rendering process between different sound spots. During the realization of this application, special emphasis has been placed on finding suitable trade-offs between both a still cost-effective and plausible acoustic simulation. Thereby, obtaining an accurate continuous geographical position has emerged as a challenge.

Introduction

The goal was an implementation of an interactive augmented audio application, which is offering location-based audio content following Azuma's [1] characteristics of an augmented reality system, which are

a) combination of real and virtual content,

b) interactive real time control,

c) suitable content registration in 3D.

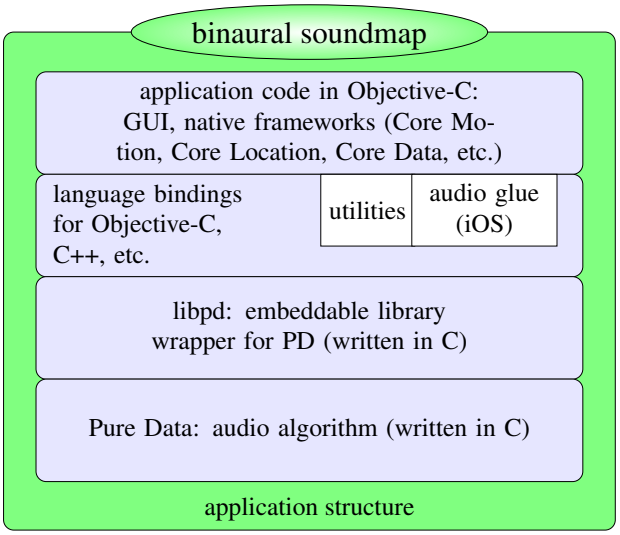
While point (a) is satisfied by the usage of acoustically permeable open headphones to merge the reality and the virtual channel, points (b) and (c) are met by implementing geo-localized DBS.

Slightly simplified, urban outdoor environments may acoustically be characterized as partially bounded free field conditions. Hence, such a sound field may be assumed to comprise a direct sound path (including distance damping and air absorption), a ground reflection path (additionally including surface scattering and absorption), and some diffuse reflection patterns produced by surrounding obstacles such as housings. Taking acoustic diffraction around obstacles into account would be much more computationally intensive. Therefore, diffraction is not included in our sound field model. The auditory impression of a sound source positioned at an arbitrary direction in free field may be evoked straight-forwardly by convolving anechoic audio signals with Head-Related Transfer Functions (HRTFs) or Head-Related Impulse Responses (HRIRs). In order to evoke a proper impression of distance, additional cues such as early and late reflections may be added while taking care of proper temporal and energetic alignment.

Implementation

As represented in figure 1, the augmented audio application was implemented for iOS using Objective-C, native frameworks and additional libraries.

Figure 1: Application structure, incl. the layer model by Brinkmann [2, p. 44, fig. 4-1]

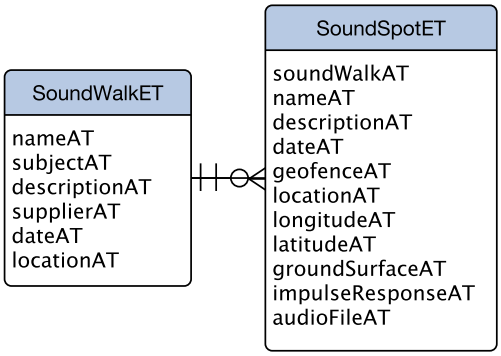


The auralization algorithm was prototyped in Pure Data (PD) and embedded in iOS using the open source API libpd. The communication of separate layers takes place between the immediate neighbour only and includes a

layer by layer type conversion. [2] The selected hardware platform, an iPhone 5s, has proved as suitable.

The application's data model in figure 2 includes two entities: a *sound walk* entity and a *sound spot* entity standing to each other in a one-to-many relationship. A sound spot represents an auralisation zone which is determined by its expansion and geographical position. The properties of such a sound spot include geographical information such as the coordinates of the position of the virtual sound source, and a geo-fence which is an adjustable virtual radius of no more than 30 m to limit the auralization zone. Further properties are audio data, such as the current audio file, the Outdoor Impulse Response (OIR), and the filter preset of the ground reflection – and meta data, such as name, context, and description. A sound walk represents a collection of sound spots which may have some contextual relationship and may be explored as a whole.

Figure 2: Data model: entities (ET) and their attributes (AT) in a one-to-many relationship (1:N)



The acoustical parameters of the simulation are defined by the sound spot model. To allow a real time interaction, the concept of the mobile dynamic auralization follows a plausible less than an exact approach. Thus, each sound spot is an omni-directional sound source neglecting potentially occluding obstacles in its vicinity. For smaller distances the direct sound and the ground reflection are the most important sound field components of such a sound source. As mentioned before, the model of both direct sound and ground reflection include the acoustic propagation path delay, distance-based and atmospheric damping.

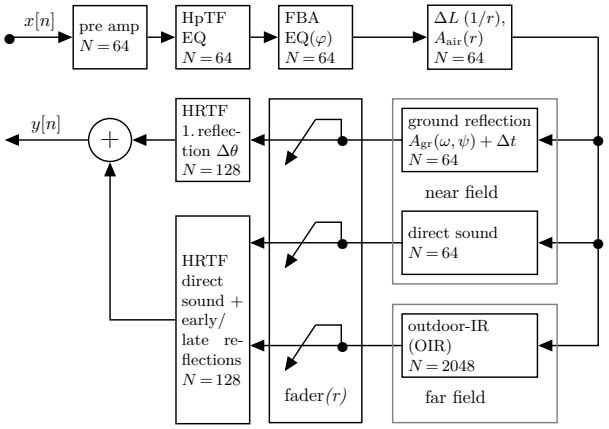
The ground reflection is constructed with a first order mirror image source and takes into account material and angle dependent absorption characteristics of the floor.

The calculated sound field characteristics are applied to the original audio content by using temporal delays and (frequency dependent) amplitude weightings. Finally, and in order to evoke a correct localization, audio streams for both direct sound and ground reflection are convolved separately with suitable HRTFs. Furthermore, these HRTFs are continuously updated in respect to the listener's current head orientation towards the active sound spot.

With increasing sound spot distances, the direct sound and the ground reflection path will become more and more similar in terms of direction of incidence and overall damping. Therefore, the mirror image source model is gradually replaced by a convolution with suitable OIRs that is completely acoustically effective close to the inner borders of the geo-fence. These OIRs contain only early and late reflections, so the direct sound part is not doubled. They are stored in a database and resemble various prototypic outdoor environments and distances during different weather conditions and are selected by the user during the sound spot setup.

The complete algorithmic structure as it has been implemented in the application is visualized in figure 3. Relevant elements of this algorithm are explained in more detail.

Figure 3: Audio synthesis signal flow chart



$A_{\mathrm{gr}}(\omega, \varphi)$ = angular- and frequency dependent ground attenuation

$A_{\mathrm{air}}(r)$ = distance- and frequency dependent air absorption (dissipation)

FBA EQ(φ) = notch filter with $f_m = 1.2\,\mathrm{kHz}$ and angular dependent attenuation to reduce front/back ambiguity

N = block size

For correct localization, the OIR convolution result is filtered with the HRTF, the direction of the direct sound path. Both situations – the near (direct sound and ground reflection path) and far field (OIR convolved path) – are simultaneously calculated and the results are mixed depending on the current distance to the virtual sound source.

When assuming a constant height of the virtual sound spot above ground, the angle of incidence of the resulting ground reflection depends on the distance to the listener. Moreover, this angle determines the actual reflection factor of the ground. Under the further assumption of a locally reacting surface for the floor, the reflection factor can be calculated as follows:

$$r_\theta = \frac{\sin\theta - \frac{Z_0}{Z_1}}{\sin\theta + \frac{Z_0}{Z_1}} \tag{1}$$

θ = angle of sound incidence

Z_0 = characteristic impedance of a plane wave resulting from $\rho_0 \cdot c_0 = 415.1$ [Pa · s/m]; $\vartheta = 20\,°\mathrm{C}$

Z_1 = ground impedance $R_1 + \mathrm{j}X_1$ (local reacting surface) [N · s /m^3]

The reflection factor r_θ in equation 1 depends on the ground impedance and can be calculated according to equation 2 which was postulated by Delany and Bazley [3, p. 107] and revised by Miki [4, p. 21]. Values for flow resistance σ may be found at Embleton [5] and [6, pp. 141] for prototypic surface types.

$$Z_1 = \rho_0 c_0 \left[1 + 5.50 \left(10^3 \, \frac{f}{\sigma} \right)^{-0.632} - \mathrm{j}\, 8.43 \left(10^3 \, \frac{f}{\sigma} \right)^{-0.632} \right] \quad (2)$$

σ = flow resistance in direction of propagation

In case of an interactive augmented reality scenario this requires a continuous calculation of the ground reflection factor and an according filtering. On basis of calculations of formula 1 and 2 for different ground surfaces (results showed in [7, p. 20, fig. 5]), conceived adaptive IIR filter with distance depending coefficients is approximating the exemplarily calculated angle-depending frequency response using biquadratic (or second–order) sections. Air absorption is calculated according to the ISO 9613-1 [8]. In the application, this effect is approximated with an adaptive biquadratic IIR filter, whose coefficients are also calculated accordingly to the current distance.

While interactivity with respect to head movements is a reliable measure against front-back confusion occurring with binaural signal presentation, a spectral approach towards treating this issue was implemented, too: According to Maxwell & Burkhard [9] an enhancement of specific direction-dependent spectral cues would help improving localization. The approach utilizes a filter which realizes a direction-dependent attenuation in the frequency range centered around $f_c = 1,2\,\mathrm{kHz}$ and with a maximum attenuation of $-10\,\mathrm{dB}$ at 0° azimuth angle. This frequency band was selected because changes here are prominent and run exactly oppositely to each other for frontal and rear sound incidence while only little changes are observed for other directions (such as, e.g., for sounds from above [10]). Further on, a notch filter was chosen for the realization because narrow band dips are less well detectable in broad band spectra than comparable peaks [11].

To meet the requirement of a flat transfer function of the headphone-ear-canal-interface [12], headphone equalization is mandatory. In practice, two problems do occur here: One, in a real-life application the particular headphone used (and therefore its specific frequency response) is typically unknown. Two, the actual headphone transfer function – especially the exact position of peaks and notches in the higher frequency range – depends to a large degree on the listeners individual morphology and on the exact wearing position of the headphones on the listener's head [13]. Hence, while referring to the (inverted) headphone target function of Møller [12, p. 225, fig. 5b], we implemented a shelving filter with a moderate attenuation above $f_m = 3\,\mathrm{kHz}$ targeting a compensation filter which is suitable for as many headphones as possible, at least in the intermediate frequency range.

Dynamic HRIR convolution is realized with the help of the PD extension `earplug~` by Xiang et al. [14]. This object uses a set with HRIRs from a KEMAR mannequin [15] for time domain convolution, where we measured a processing latency of 0.5 ms on an iPhone 5s. At the current state of development we use Apple's Core Motion Framework and the iPhone's internal sensor data for head-tracking. External head-trackers could be supported in the future. When used as head-tracker, the 50 Hz update rate of the iPhone's internal Inertial Measurement Unit (IMU) will add 20 ms to the overall latency with respect to head movements.

Our application allows real-life and virtual audio content to be perceived simultaneously in a mixed or augmented reality [16] audio application. According to Brungart [17], augmented reality scenarios will pose highest constraints on the maximum tolerable latency. Hence, Brungart found values as low as 30 to 40 ms just acceptable. In order to minimize the overall audio processing latency, we first classified all required processing steps in two categories: One where latency is not critical – and one where real-time capability is required. In the latter case, filters were realized either by using efficient second order Infinite Impulse Response (IIR) filters for dynamic audio processing [18], or by (dynamic) time-domain convolution of comparatively short FIR filters (HRIRs) [14]. This applies to the time-variable notch filter implemented for treating front-back-ambiguities and to the dynamic HRIR convolution. Thereby, the latter one — the computationally most expensive calculation — causes nearly no significant latency. On the other hand, for adding monoaural spatial cues by convolution with monophonic OIRs, we use the PD extension `convolve~` [19]. The `convolve~`-object induces a latency of about 40 ms. However, in view of the fact that human auditory distance perception is relatively poor, we did not consider it a severe shortcoming.

After falling short of a predefined geo-fence radius of a sound spot by the actual user position, the auralization starts immediately. In order to switch between multiple sound spots, the one next to the user's current position is determined using a nearest-neighbor criterion. Thereby, the calculation of the sound path length using coordinates delivered from the Assisted Global Positioning System (A-GPS) of the smartphone turned out to be especially problematic. Depending on the receiving conditions we measured a best-case positional accuracy up to ±5 m. However, in practice, the accuracy of the position data could be even worse which resulted in spatial discontinuities during audio rendering potentially exceeding the perceptual threshold for spatial discretization in virtual acoustic environments [20].

Summary and outlook

An augmented audio application using geo-localized Dynamic Binaural Synthesis (DBS) has been implemented for iOS devices using the open source real-time audio programming language PD embedded with libpd [21, 22]. Users of the software may place multiple virtual sound spots freely on a 2D map of their environment which is bounded by an adjustable border (geo-fence).

Additional scene parameters such as the audio content of sound spots and the prototypical outdoor environments may be defined. Geographical positioning and (interactive) user orientation are provided through the iPhone's IMUs and A-GPS modules. Thereby it was found that A-GPS provides real time positioning data which is applicable for the simulation of distant sound sources. However, for shorter distances, the limited accuracy and stability of the A-GPS position data resulted in audible spatial discontinuities during audio rendering.

As an alternative to A-GPS, Differential Global Positioning System (GDGPS) [23] could be taken into account, however, at the time of writing, it is not available for the general public [24]. Another alternative would be a combination of A-GPS and an Inertial Navigation System (INS), which uses sensor data fusion and drift compensation to hold the sound source position stable while the auralization is running. Because of the small radii of the sound spots, a recalibration of the A-GPS supported INS could be preferably performed outside of a current auralization zone.

In the near future, an external head-tracker will be supported. Additionally, the current computational costs (for now $\approx 20\,\%$ of the iPhone 5s' A7 cores) shall be reduced in order to allow for a future simultaneous simulation of multiple and moving virtual sound sources.

References

[1] Azuma, R. (1997): "A survey of augmented reality." In: *Presence: Teleoperators and Virtual Environments*, vol. 6, 355–385.

[2] Brinkmann, P.; Inc, G.; Kirn, P.; Lawler, R.; Mccormick, C.; Roth, M.; Steiner, H. (2011): "Embedding pure data with libpd." In: *in Proceedings of the Pure Data Convention, Weimar*.

[3] Delany, M.; Bazley, E. (1970): "Acoustical properties of fibrous absorbent materials." In: *Applied Acoustics*, **3**(2):105–116.

[4] Miki, Y. (1990): "Acoustical properties of porous materials. modifications of delany-bazley models." In: *Journal of the Acoustical Society of Japan (E)*, **11**(1):19–24.

[5] Embleton, T.F.W. (1983): "Effective flow resistivity of ground surfaces determined by acoustical measurements." In: *The Journal of the Acoustical Society of America*, **74**(4):1239.

[6] Attenborough, K.; Li, K.; Horoshenkov, K. (2006): *Predicting Outdoor Sound*. Taylor & Francis.

[7] Hädrich, M. (2015): *Konzeption und Implementierung einer mobilen App für geolokalisierte dynamische Binauralsynthese*. Master's thesis, Technische Universität Berlin.

[8] ISO (1996), "ISO 9613-1: Acoustics - Attenuation of sound during propagation outdoors - Part 1: Calculation of the absorption of sound by the atmosphere."

[9] Maxwell, R.J.; Burkhard, M.D. (1979): "Larger ear replica for KEMAR manikin." In: *The Journal of the Acoustical Society of America*, **65**(4):1055.

[10] Blauert, J.; Braasch, J. (2008): "Räumliches Hören." In: S. Weinzierl, ed., *Handbuch der Audiotechnik*, chap. 3, 87–121, Springer-Verlag.

[11] Bücklein, R. (1981): "The audibility of frequency response irregularities." In: *J. Audio Eng. Soc*, **29**(3):126–131.

[12] Møller, H.; Jensen, C.B.; Hammershøi, D.; Sørensen, M.F. (1995): "Design criteria for headphones." In: *J. Audio Eng. Soc*, **43**(4):218–232.

[13] Schärer, Z.; Lindau, A. (2009): "Evaluation of equalization methods for binaural signals." In: *Audio Engineering Society Convention 126*.

[14] Xiang, P.; Puckette, M.; Camargo, D. (2005): "Experiments on spatial gestures in binaural sound display." In: *Proceedings of ICAD*, International Conference on Auditory Display.

[15] Gardner, W.G. (1995): "HRTF measurements of a KEMAR." In: *The Journal of the Acoustical Society of America*, **97**(6):3907.

[16] Milgrim, P.; Takemura, H.; Utsumi, A.; Kishino, F. (1995): "Augmented reality: a class of displays on the reality-virtuality continuum." In: H. Das, ed., *Telemanipulator and Telepresence Technologies*, vol. 2351, Boston,MA: The International Society for Optical Engineering.

[17] Brungart, D.S.; Simpson, B.D.; Kordik, A.J. (2005): "The Detectability of Headtracker Latency in Virtual Audio Displays." In: *Eleventh Meeting of the International Conference on Auditory Display*, Limerick, Ireland.

[18] Moser-Booth, M. (2011), "biquad.mmb~.pd."
URL `https://github.com/dotmmb/mmb/blob/master/biquad.mmb%7E.pd`

[19] Brent, W. (2010), "Williambrent.com."
URL `http://williambrent.conflations.com/pages/research.html`

[20] Lindau, A.; Weinzierl, S. (2009): "On the spatial resolution of virtual acoustic environments for head movements in horizontal vertical and lateral direction." In: *Proc. of the EAA Symposium on Auralization*, Espoo.

[21] Brinkmann, P. (2012): *Making Musical Apps*. O'REILLY.

[22] Kirn, P. (2012), "About libpd."
URL `http://libpd.cc/about/`

[23] Yoaz Bar-Sever, J.P.L.C.I.o.T. (2013), "The global differential gps system."
URL `http://www.gdgps.net`

[24] OpenDGPS (2012), "Opendgps - differential gps for the rest of us."
URL `http://opendgps.de/about/`