A hybrid approach for three-dimensional sound spatialization

Algorithmen in Akustik und Computermusik 2, SE

Fabio Kaiser

Supervision: Dr. Franz ZOTTER and Dipl.Ing. Matthias FRANK

Graz, May 3, 2011

kunst uni graz

institut für elektronische musik und akustik

Abstract

Three-dimensional sound spatialization is widely done by using techniques like Ambisonics or Vector-Base Amplitude Panning. Both methods have their advantages and disadvantages. This makes the decision of what technique may be used dependent on the purpose. In this report a hybrid approach is presented combining the two techniques and by doing this it is supposed to overcome the main disadvantages of each method. The implementation of this new concept in Pure Data is shown at the end.

Contents

1 Introduction

The desire to spatialize sound in three dimensions and in big scale setups has been raised by composer in the 50's and 70's[1]. Since these days researchers suggested numerous techniques for this purpose which can be summarized into three categories:

– Inter-channel level difference
– Inter-channel time difference
– Soundfield reproduction

Inter-channel level difference techniques try to spatialize sound by reproducing coherent signals out of two or more loudspeakers which produces the perception of phantom sources anywhere in between the loudspeakers. Techniques using this is Stereo, Surround Sound, Vector Base Amplitude Panning (VBAP) and first order Ambisonics to name but a few.

Inter-channel time difference techniques use time delays to spatialize sound. The standard recordign techniques used for stereo reproduction and the space unit generator can be named here.

Sound field reproduction techniques try to reconstruct a soundfield of a real acoustical scene. Wavefield synthesis, Higher-order Ambisonics (HOA) and Boundary Surface Control are to name here.

Needs for big scale spatialization can not only be found in electro-acoustic music but also in sound for video games, virtual reality applications, auralization, research in spatial hearing and data sonification.

In this report we will take a closer look on HOA and VBAP. Both are widely used techniques and they have both advantages and deficiencies. However, the properties can be seen to complement one another which leads to the idea of a new approach based on the combination of two is presented here.

HOA and VBAP are discussed in detail and their properties are outlined. With that knowledge in mind, a hybrid approach is presented. At the end of the report the implementation of the concept in Pure-Data is revealed.

1. See Edgar Varése and his Poém Electronique in the Philipps Pavillion at the World Expo 1958 in Brussels and Karlheinz Stockhausen at the german pavillion at the World Expo 1970 in Tokyo.

2 Ambisonics

Ambisonics was introduced in the 70's by Gerzon [2]. By that time the concept was restricted to first order spherical harmonics and so could be seen as an inter-channel level difference panning technique. Later the ideas were extended to higher-order ambisonics (HOA) [1]. This extension now attempts to reproduce real existing sound fields and can be seen as a special case of acoustic holophony.

The starting point for the derivation of ambisonics is the solution of the Helmholtz equation

$$(\Delta + k^2)p = 0 \tag{1}$$

where p is the sound pressure, Δ the Laplace-operator and k the wavenumber.

The solution in spherical coordinates yields two parts, the radial part and the angular part. The radial part consists of the spherical bessel and hankel functions and determines the distance dependencies of the sound field. The angular part comprises the dependencies of the sound field in azimuth and elevation. These angular solutions are called spherical harmonics (SH) and can be seen as modes of vibration of a sphere. They state a infinite series of modes of vibration (orders) where e.g., the order zero determines the DC component and the first order represents the vibration of a dipole in x-,y- and z-direction. The spherical harmonics until order five are shown in Fig. 1.

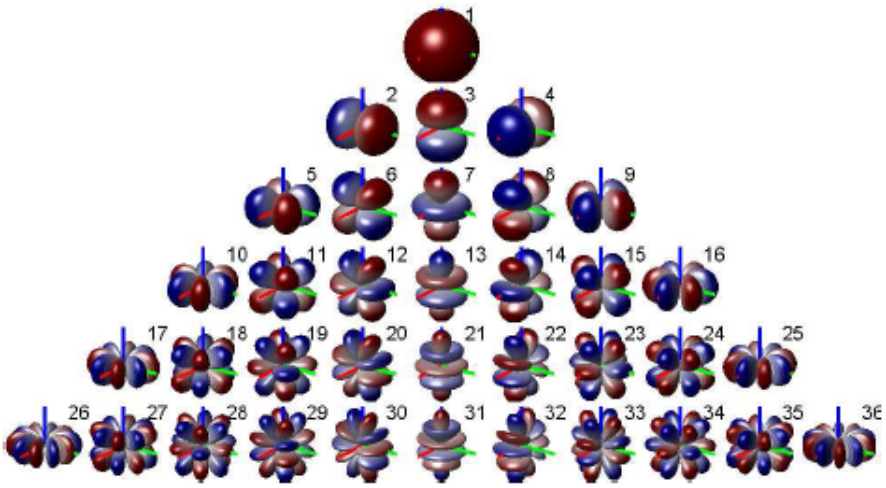


Figure 1: Spherical harmonics up to fifth order [6]

2.1 Ambisonics formulation

Originally the derivation of Ambisonics i.e., how to reproduce a given sound field with a given distribution of sound sources, was based on the assumption that every sound source emits a plane wave. It follows that the radial solution drops and the derivation is facilitated. However, because this is being far from what can be done in reality an

alternative approach to the derivation of Ambisonics has been shown by Zotter [5]. The approach is to assume the sources for reproduction to be a continuous angular distribution of point-sources at finite distance. If continuously spread sources excite the Helmholtz equation at a certain radius r_l with a continuous distribution of source strength $f(\theta)$, the nonhomogeneous Helmholtz equation is written as

$$(\Delta + k^2)p = -\frac{\delta(r - r_l)}{r^2} f(\boldsymbol{\theta}) \tag{2}$$

where $\delta(r - r_l)$ is the Dirac delta distribution and $\boldsymbol{\theta}$ is the Cartesian unit vector in spherical coordinates.

The solution can be found as

$$p = -\sum_{n=0}^{\infty} \sum_{m=-n}^{n} ikj_n(kr)h_n(kr_L)Y_n^m(\boldsymbol{\theta})\phi_{nm} \tag{3}$$

where $j_n(kr)$ and $h_n(kr)$ are the spherical Bessel and Hankel functions and $Y_n^m(\boldsymbol{\theta})$ is a spherical harmonic (see Fig. 1) (cf. [?]). ϕ_{nm} are the coefficients of the SH transform of $f(\boldsymbol{\theta})$

$$\phi_{nm} = \int_{S^2} f(\boldsymbol{\theta})Y_n^m(\boldsymbol{\theta})d\boldsymbol{\theta}. \tag{4}$$

Further a point source can be expressed as a spatial Dirac delta function $f(\theta) = \delta(1 - \boldsymbol{\theta^T}\boldsymbol{\theta}_0)$ pointing at $\boldsymbol{\theta}_0$. The SH transform yields

$$\phi_{nm} = \int_S \delta(1 - \boldsymbol{\theta^T}\boldsymbol{\theta}_0)Y_n^m(\boldsymbol{\theta})d\boldsymbol{\theta} = Y_n^m(\boldsymbol{\theta}_0). \tag{5}$$

In practice the reproduction of $f(\boldsymbol{\theta})$ is achieved by using loudspeakers placed at discrete positions. It is assumed that each speaker emits the field of a point source driven by the gains g_l. The non-homogeneous Helmholtz equation then can be written as

$$(\Delta + k^2)p = -\frac{\delta(r - r_l)}{r^2} \sum_{l=1}^{L} g_l \delta(1 - \boldsymbol{\theta^T}\boldsymbol{\theta}_0) \tag{6}$$

where $\hat{f}(\theta) = \sum_{l=1}^{L} g_l \delta(1 - \boldsymbol{\theta^T}\boldsymbol{\theta}_0)$ is a discrete angular source strength distribution.

Direct comparison of the continuous and the discrete source strength distribution yields the gains g_l necessary for reproducing accurately the sound field of $f(\theta)$

$$f(\boldsymbol{\theta}) = \hat{f}(\boldsymbol{\theta}). \tag{7}$$

The comparison in the SH transform domain yields

$$\phi_{nm} = \sum_{l=1}^{L} g_l Y_n^m(\boldsymbol{\theta}_l). \tag{8}$$

This is refered to as *modal source strength matching*.

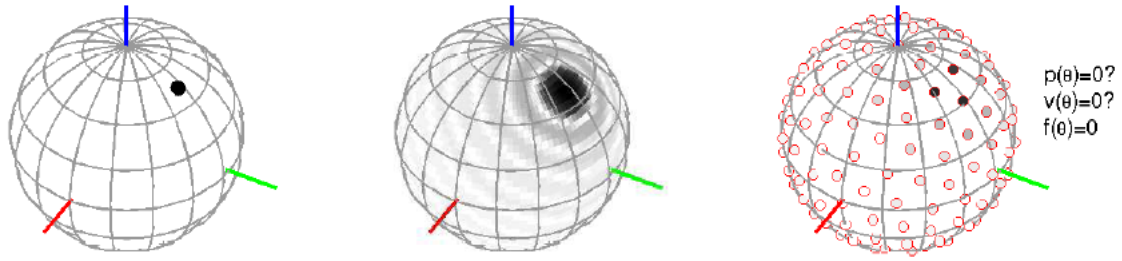
From Eq. 8 it can be seen that this equation is only solvable if the location of the point source θ_S and the location of the loudspeaker θ_l incident. In practice this can not always be guaranteed. However, if an angular band-limitation is assumed and applied to the SH transform the matching condition becomes

$$B_N\{f(\boldsymbol{\theta})\} = B_N\{\hat{f}(\boldsymbol{\theta})\} \tag{9}$$

where B_N indicates the band-limitation to order N. The truncation of the order of the SH transform has the effect that the source width of the point source is bigger and that side lobes evolve. This can be seen equivalent to the windowing effect of the Fourier transform on time signals.

Here this is an advantageous property because it reduces the loss of directional information when using discrete sources. Therefore the angular band-limitation solves the problem of finite discrete reproduction sources. Fig. 2 (a) shows the SH transform of a point-source and Fig. 2 (b) its truncated version. In Fig. 2 (c) a discrete distribution is shown.

However, the two steps angular band-limitation and angular sampling create two undesired effects. On the one hand a reduced range of accurate sound field reproduction (or the sweet spot) and on the other hand spatial aliasing. Note that these two aspects of Ambisonics have to be treated separately. This is not further investigated here and it is referred to [5].



(a) Continuous distribution (b) Angularly band-limited distr. (c) Discretized distribution

Figure 2: Spherical harmonic transform of a point-source.

2.2 Decoding

The modal source strength matching condition of Eq. 8 is normally formulated in vector/matrix notation

$$\phi_N = \boldsymbol{Y}_N \boldsymbol{g}. \tag{10}$$

In order to get the gain vector $\boldsymbol{g}$ for reproduction the inversion of matrix $\boldsymbol{Y}_N$ is necessary. Direct inversion is often impossible because the number of loudspeakers has to match

the number of SH, which is possible for just a few setups. However, a matrix fulfilling Eq. 10 can be found

$$\boldsymbol{g} = \boldsymbol{D}\phi_N \tag{11}$$

where $\boldsymbol{D}$ is the so-called decoder matrix and is the right inverse of $\boldsymbol{Y}_N$

$$\boldsymbol{D} = \boldsymbol{Y}_N^T(\boldsymbol{Y}_N\boldsymbol{Y}_N^T)^{-1}. \tag{12}$$

The numerical stability of $\boldsymbol{D}$ is determined by the inversion of the Gram-matrix $\boldsymbol{G}_d = \boldsymbol{Y}_N\boldsymbol{Y}_N^T$. This leads to the question: *When is the inversion of the Gram-matrix a problem?* Or: *When is the inversion of the Gram-matrix NOT a problem?*

As described in the previous section what we want is to reproduce a continuous distribution of point sources via an available distribution of loudspeakers, which at the moment can only be discrete. Intuitively, one can imagine that a distribution of loudspeakers solely on one half of the sphere, i.e. a hemisphere, cannot excite the modes of vibration on the other half of the sphere. In general a nonuniform distribution will lead to errors in the calculation of the loudspeaker gains, because the orthogonality of the band-limited spherical harmonics is distorted. The inversion might not be possible at all. This states a big problem in the use of ambisonics and careful decoder design has to be employed, cp. [6].

Still there is hope. Distributions which uniformally sample the sphere exist. The so called platonic solids are shown in Fig. 3. These sampling schemes with $L = \{4, 6, 8, 12, 20\}$ points and the orders $N = \{1, 1, 1, 2, 2\}$ provide that the Gram-matrix is equal to the scaled identity matrix

$$\boldsymbol{Y}_N\boldsymbol{Y}_N^T = \boldsymbol{I}\frac{4\pi}{L}. \tag{13}$$

Further there exist distributions that provide regular sampling of the sphere. The so called t-designs are shown in Fig. 4. The platonic solids state a subgroup of these.

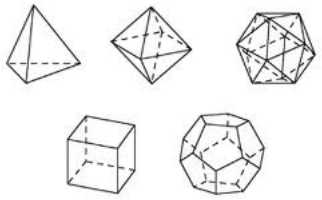


Figure 3: Platonic solids

2.3 Ambisonics features and performance

A feature of Ambisonics is that recording techniques are readily available. Since the beginnings of Ambisonics a first-order microphone was used (the Soundfield microphone)

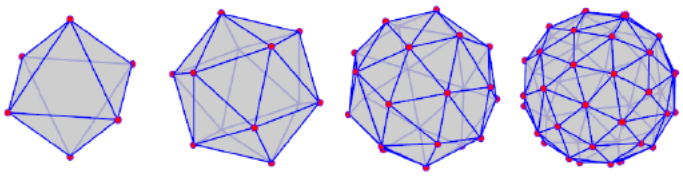


Figure 4: t-designs for t=3,5,7,9

and nowadays higher-order spherical microphones are becoming more common place. This means that a real acoustical scene can be recorded and reproduced by an Ambisonics system.

Further Ambisonics is praised for it's smoothness in the spatialization which comes from the fact that the angular spread of the sound source to be spatialized does not vary with the location.

3 VBAP

Vector-Base Amplitude Panning (VBAP) is a 2D or 3D spatialization techniques introduced by Pulkki in 1997 [3]. It uses a triangulation of the convex hull of the prevalent loudspeaker arrangement. If the intended sound source location matches the area of a triangle, coherent signals with different gains are played back by this speaker triplet and a phantom source is created. Depending on the desired location of the phantom source the gains of the loudspeakers are calculated by vector analysis. Psychoacoustically valid results are only achieved if the aperture between loudspeakers does not exceed $90°$.

3.1 VBAP formulation

Three loudspeakers are placed in arbitrary positions on the surface of the unit sphere, Fig. 5. Each loudspeaker is represented by a three-dimensional unit vector in Cartesian coordinates

$$\boldsymbol{l}_n = [l_{xn} \;\; l_{yn} \;\; l_{zn}]^T \tag{14}$$

where the index $n = 1, 2, 3$. The direction of the phantom source is written as

$$\boldsymbol{p} = [p_x \;\; p_y \;\; p_z]^T. \tag{15}$$

The virtual source vector $\boldsymbol{p}$ is now expressed as a linear combination of the loudspeaker vectors

$$\boldsymbol{p} = g_1\boldsymbol{l}_1 + g_2\boldsymbol{l}_2 + g_3\boldsymbol{l}_3 \tag{16}$$

in vector notation

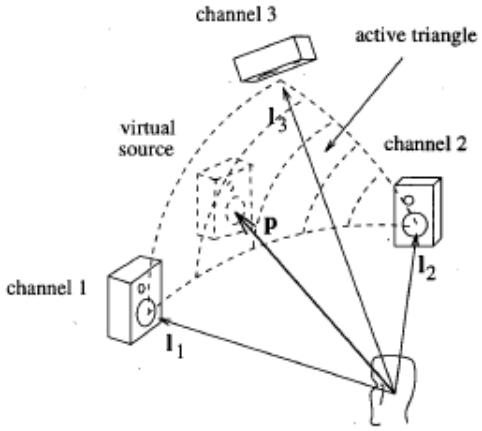


Figure 5: Three-dimensional setup of loudspeakers. A phantom source is created inside the area of the triplet by playing back coherent signals with different amplitudes [3].

$$\boldsymbol{p} = \boldsymbol{g}\boldsymbol{L} \tag{17}$$

where $\boldsymbol{g} = [g_1 \;\; g_2 \;\; g_3]^T$ is the gain vector and $\boldsymbol{L} = [\boldsymbol{l}_1 \;\; \boldsymbol{l}_2 \;\; \boldsymbol{l}_3]^T$ is the matrix comprising the unit vectors pointing to the loudspeakers positions.

The speaker gains are then calculated by inverting $\boldsymbol{L}$

$$\boldsymbol{g} = \boldsymbol{p}^T \boldsymbol{L}^{-1}. \tag{18}$$

The power of the three speakers is normalized to a constant value

$$g_1^2 + g_2^2 + g_3^2 = C \tag{19}$$

which leads to a scaling of the gains

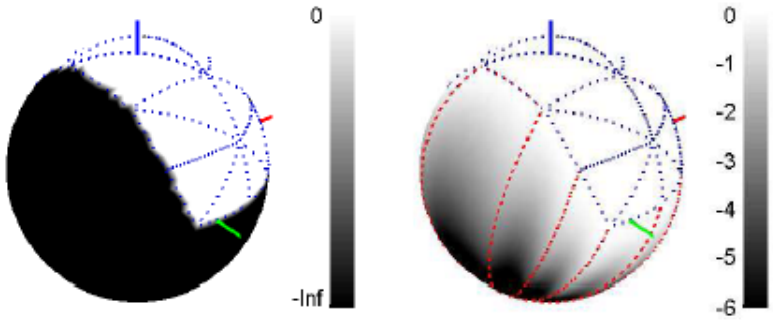
$$\boldsymbol{g}_s = \frac{\sqrt{\boldsymbol{C}}\boldsymbol{g}}{\sqrt{g_1^2 + g_2^2 + g_3^2}}. \tag{20}$$

As mentioned before, if more than three loudspeakers are to be used a triangulation is necessary. The correct triplet is then chosen if all loudspeaker gains for that triplet are positive.

3.2 VBAP extension

As virtual source positions outside the triangulation of the loudspeaker arrangement are not taken into account the signal level drops to zero for those sources. This is obvious and one may ask why this is mentioned but in the context of the presented technique here it is a disadvantageous behavior. In order to achieve a smoother transition an imaginary

loudspeaker is introduced. Fig. 6 depicts the idea. The position of the additional loudspeaker is calculated so that it includes the listener (origin) and the signal of this speaker is then omitted in the end. The calculation can be roughly explained as follows. The normal vector to each surface spanned by a triangle defined from the triangulation is calculated and it's size is compared to a certain margin (e.g. $90°$ as mentioned before). If it exceeds that margin the normal vector is saved and added to previously calculated ones. The resulting vector is normalized and then it states the position of the imaginary loudspeaker $\boldsymbol{\theta}_{L+1}$.



(a) standard VBAP-mapping (b) customized VBAP-mapping (*imaginary loudspeaker*)

Figure 6: Visualization of the effect of the imaginary loudspeaker [4].

3.3 VBAP features and performance

As mentioned VBAP doesn't allow the positioning of sound sources outside the active triangle. If a sound source is located exactly at the location of one speaker only this speaker is active. Equivalently a source on the line between two speakers activates only these two. For all other positions the three speakers are active. This leads to a modulation of the spread of a virtual source depending on it's position. The main advantage is that VBAP allows for nearly arbitrary loudspeaker arrangement.

4 A hybrid approach

The here presented new concept for three-dimensional spatialization technique was first proposed in [4]. It is a hybrid approach combing the techniques explained in the previous chapters.

On the one hand there is the advantage of t-design Ambisonics providing a constant angular power-spread σ_E and on the other hand VBAP supporting arbitrary loudspeaker arrangements $\boldsymbol{\theta_l}$. As both approaches provide constant power over all possible virtual source positions $\boldsymbol{\theta_s}$ a combination of these two methods results in a technique having it all.

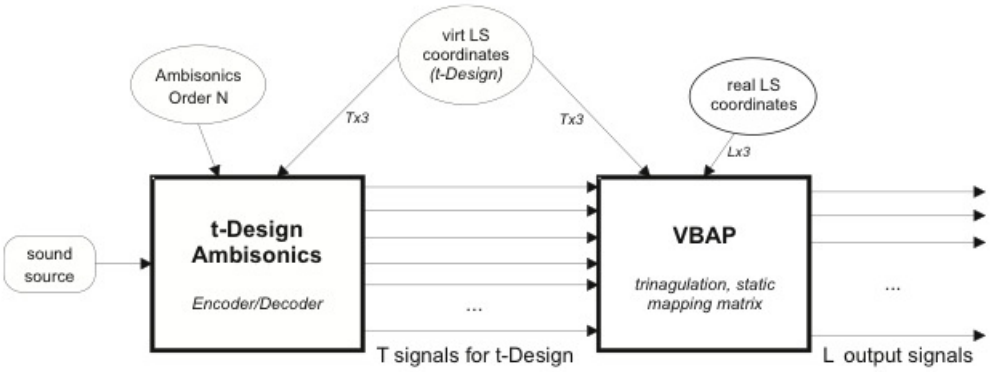


Figure 7: Block schema of the presented hybrid approach. The round windows state initializations where the cartesian coordinates of the loudspeakers are in matrix form of $T \times 3$ and $L \times 3$.

The spatialization process is explained in the following and illustrated in Fig. **??**. The Ambisonics block is initialized with the coordinates of the virtual loudspeakers $\boldsymbol{\theta_t}$ (t-design) and the ambisonics order N. An optimal combination of the t-design and the order is if $t \geq 2N + 1$ [4]. The VBAP block is initialized with the coordinates of the virtual loudspeakers $\boldsymbol{\theta_t}$ and the real loudspeakers $\boldsymbol{\theta_l}$. The additional imaginary loudspeaker is calculated and added if necessary. After the triangulation of the real LS rig the VBAP gains for each virtual LS is calculated and put into a big 'mapping' matrix $\boldsymbol{M} = T \times (L + 1)$. Only three elements in each row of $\boldsymbol{M}$ are non-zero and the values are static.

Now a virtual sound source with position $\boldsymbol{\theta_l}$ can be spatialized. It is first transformed into the spherical harmonics domain resulting in $(N+1)^2$ encoder signals which are then decoded to the virtual LS rig. The output of this block are T signals. These signals are then mapped onto the triangulation of the real LS rig, that means the signals are multiplied with the "mapping" matrix $\boldsymbol{M}$. The output is L signals which can be sent to the real loudspeakers.

It should be mentioned that an increasing ambisonics order leads to a more narrow point source in the spherical harmonics domain, as described in sec. 2. This means

that the ambisonics order should not be chosen too high for the prevalent loudspeaker arrangement because otherwise the properties of the system approach the properties of a VBAP system.

4.1 Evaluation

In order to evaluate the concept objectively two performance measures are suggested in [4]. The total signal power

$$E = \sum_{l=1}^{L} g_l^2 \tag{21}$$

and it's angular spread around the virtual source position $\boldsymbol{\theta}_S$

$$\sigma_E = arccos(\sum_{l=1}^{L} \langle \boldsymbol{\theta}_l, \boldsymbol{\theta}_S \rangle \frac{g_l^2}{E}) \tag{22}$$

Fig. 8 depicts the two performance measures for VBAP. The total signal power is constant but the angular spread varies dependent on the source location.

In Fig. 9 the two measures for t-design Ambisonics mapped onto a system of loudspeakers via VBAP are shown. The total signal power and the angular spread are nearly constant over all source locations. The big disadvantage though, is that a loudspeaker setup according to the t-designs is hardly realizable. Either because of the vast amount of loudspeakers necessary or because of the difficulty to place the loudspeakers at the right positions.

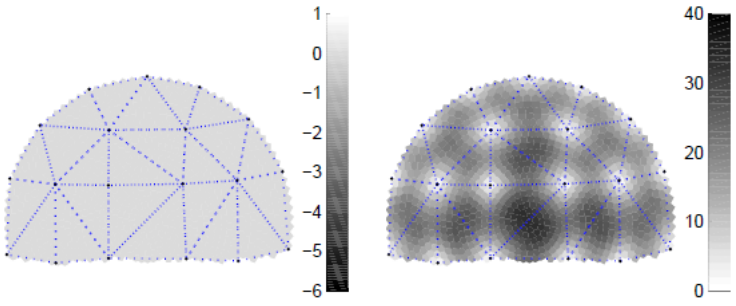


Figure 8: VBAP: The total signal power E (left) and the angular spread measure σ_E (right) for a sample triangulation. [4].

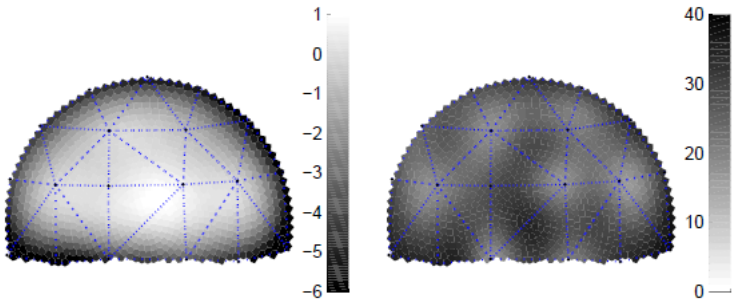


Figure 9: Ambisonics: The total signal power E (left) in dB and the angular spread measure σ_E (right) in degree for a virtual t-design using L=180 loudspeakers and the order N = 4 [4].

5 Implementation in Pure-Data

The presented concept was implemented in the real-time graphical programming environment Pure-Data (PD). The program was written in a modular way. Several objects or abstractions take a certain input, execute the calculations and output data in a certain format. The data formats are mostly chosen to be matrices. This is made possible by the use of the PD external *iemmatrix*. Fig. **??** shows the main patch with the different modules.

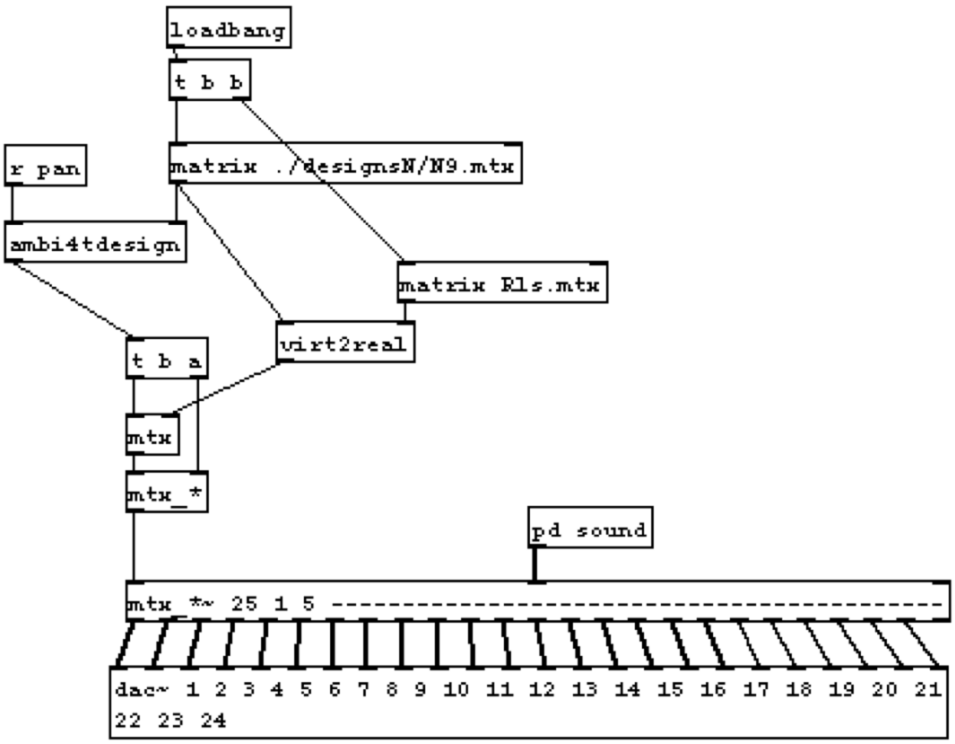


Figure 10: The main patch of the implementation showing the different modules.

The *virt2real* and the *ambi4tdesign* are the main objects in the patch. The *virt2real*object takes as input the virtual and the real loudspeaker coordinates (cartesian). The format is the matrix format de??ned by the iemmatrix external. After computing the triangulation and the imaginery extra loudspeaker it then calculates the static mapping matrix $\boldsymbol{M}$ and outputs it. Here it is stored in the *mtx* object. The mapping matrix doesn't change unless the real or virtual loudspeaker setup changes.

The *ambi4tdesign* object takes as input the virtual source position coordinates (spherical) and the virtual loudspeaker coordinates (cartesian), which are then converted to spherical coordinates. The decoder matrix is calculated out of the virtual loudspeaker coordinates and the encoder matrix (or vector, dependent on the number of sources) is calculated from the virtual source coordinates. As explained in sec. 2 the inversion of the decoder matrix can be done by simple transposition and weighting of the matrix. The matrix multiplication with the encoder matrix then yields a matrix carrying the so called Ambisonics gains. These gains are then calculated with the mapping matrix which gives the final loudspeaker gains. Applying these gains to the loudspeakers yields a sound source spatialized at the desired virtual source location. Note in Fig. **??** that the last output is not connected anywhere. This is the signal for the imaginary loudspeaker.

The implementation was generally straightforward and the *iemmatrix* external greatly facilitated programming in PD.

6 Conclusions

A hybrid approach for three-dimensional sound spatialization was discussed and the implementation in the real-time graphical programming environment Pure-Data was revealed.

The system's basis on the one side is an ideal version of Higher-Order Ambisonics using a virtual loudspeaker setup and on the other side a Vector-Base Amplitude Panning system consisting of the real available loudspeakers used for the reproduction of the virtual loudspeaker signals.

The advantage of Ambisonics is that it provides constant source power and constant source spread over all positions. On the other hand VBAP allows for arbitrary loudspeaker setups. The advantages of both systems are combined into the new system.

This system is also capable of reproducing recordings made with the more and more available spatial microphone techniques. These techniques provide Ambisonics encoded signals and therefore can also be used with this hybrid system.

Furthermore an advantage arises when on thinks of concert situation. In concerts for electro- acoustic music the pieces can sometimes require different reproduction setups. Such situation can be overcome with the presented system because it can simulate any desired loudspeaker setup over any available loudspeaker setup. Of course the setups should not be vastly unequal in size and shape.

As the presented system is a new approach there exist no psychoacoustic evaluation by the time of writing. In order to be able to compare such a system with established ones this has to be accomplished.

References

[1] J. Daniel, "Représentation de champs acoustiques, application à la transmission et à la repruduction de scènes sonor complexes dans un contexte multimèdia," Ph.D. dissertation, Université Paris 6, 2001.

[2] M. A. Gerzon, "Periphony: With-height sound reproduction," *J. Audio Eng. Soc*, vol. 21, no. 1, pp. 2–10, 1973. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=2012

[3] V. Pulkki, "Virtual sound source positioning using vector base amplitude panning," *J. Audio Eng. Soc.*, vol. 45, no. 6, 1997.

[4] F. Zotter, M. Frank, and A. Sontacchi, "The virtual t-design ambisonics-rig using vbap," *1st EAA - EuroRegio 2010, Congress on Sound and Vibration*, 2010.

[5] F. Zotter, H. Pomberger, and M. Frank, "An alternative ambisonics formulation: Modal source strength matching and the effect of spatial aliasing," *AES 126th Convention, Munich, Germany*, 2009.

[6] F. Zotter, H. Pomberger, and M. Noisternig, "Ambisonic decoding with and without mode-matching: A case study using the hemisphere," *Proc. of the 2nd International Symposium on Ambisonics and Spherical Acoustics*, 2010.