New Csound Opcodes for Binaural Processing

Victor Lazzarini and Brian Carty
Sound and Digital Music Technology Group,
National University of Ireland, Maynooth,
Co. Kildare,
Ireland
victor.lazzarini, brian.m.carty@nuim.ie

Abstract
Although solutions to the challenge of binaural artificial recreation of audio spatialisation exist in the Computer Music domain, a review of the area suggests that a comprehensive, generic, accurate and efficient toolset is required. A number of Csound opcodes, using a Head Related Transfer Function based approach, are presented to satisfy this necessity. The process is a complex one, with perhaps the most significant difficulty being phase interpolation. Novel approaches (specifically methods using phase truncation and functionally derived itd respectively), as well as a method based on established digital signal processing methods (minimum phase plus delay) are implemented.

Keywords
HRTF, binaural, Csound

1. Sound Localisation

Binaural hearing is the term given to listening with two ears rather than one, and is the main factor involved in sound localisation. Interaural time and intensity differences (itd and iid respectively) can provide very accurate localisation cues. It is generally accepted that itd and iid work together to provide a well-defined spatial image, with itd working best for low frequencies and iid for high.

Monaural information (independent information from one ear) also plays an important role in sound localisation. The pinna and concha both have a non-linear frequency response over the audible spectrum, altering incoming sounds. These alterations vary with sound location.

2. Head Related Transfer Functions

Head Related Transfer Functions (HRTFs) are essentially functions that describe how a sound from a specific location is altered from source to tympanic membrane. The process of simulating an auditory location using HRTFs can be summarised thus:

- Record the impulse response of the left and right ear for the desired point in space.
- Analyse the frequency content of the sound you wish to spatialise.
- Impose the HRTF for the left and right ear on the sound using convolution.

As the physiology of everyone's ears is different, HRTFs vary considerably from subject to subject. However, certain consistencies can be observed and generalised/*non-individualised* HRTFs, recorded using a dummy head and torso model or a specific subject are frequently used. Results from [18] suggest that although non-individualised HRTFs are certainly a useful tool for binaural simulation, they can result in a distortion of the spectral characteristics used in front/back and elevation resolution.

It is also important to note that binaurally generated signals should be reproduced on headphones to avoid crosstalk and environmental and listener interactions with the source.

3. HRTF Interpolation

HRTF data sets are typically measured at discrete, equidistant points around a listener or dummy head, for example [3]. Therefore some form of interpolation is required for non measured points. The topic of HRTF interpolation is a multi faceted one, with many suggested and possible approaches, for example spatial frequency response surfaces (see [2]) and virtual loudspeaker multichannel approaches (see [15]).

The process of HRTF localisation outlined above describes the localisation of a source sound to one specific area of space. When other locations are required, however, the relevant HRTF data is needed. A fixed amount of points are typically recorded and stored. However, if a location is required that has not been measured, or if a sound is required to move smoothly from one location to another, some kind of averaging or interpolation must be done.

HRTF interpolation can be thought of as taking the two (or more for increased accuracy) nearest HRTF representations to a non-measured point in between, and deriving a new HRTF by averaging the known values with greater relative weighting(/s) on the nearer known point(/s).

Audio, or indeed any type of signals can be represented in several ways. Traditionally, audio is viewed, edited, processed and auditioned in the time domain. However, the frequency domain can provide more useful insights into the properties of the signal. Individual sinusoidal components of a signal, in this case a head related transfer function, can be examined, and their magnitude and phase can be extracted in the frequency domain. The former quantifies the relative strength of the signal at each frequency in the analysis, the latter, the phase/starting point of the component.

Frequency domain interpolation can generally give more accurate results than time domain techniques (see [4] and [14]). However, interpolating in the frequency domain poses the problem of phase interpolation. Phase values are closely related to itd, so are important in the localisation process. The linear interpolation of phase values is flawed. Phase, unlike magnitude, is a periodic quantity, measured in fractions of a full cycle. Uncertainty arises when trying to interpolate phases, as a phase value can be +/- any amount of full cycles. For example, in Figure 1, the first and second points have phase values of 10 and 50 degrees respectively. However, as phase is periodic, these may be 10 or 50 degrees plus any number of full cycles. Therefore interpolated phase may be 30 or 210 degrees, depending on whether the 50 degree phase represents 50 degrees or 410 degrees (50 degrees plus one cycle) respectively, for example.

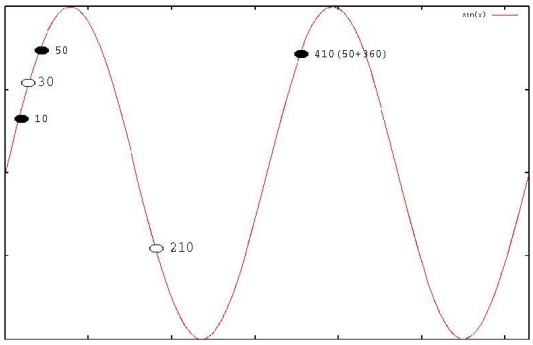

Figure 1: Phase Interpolation

4. Minimum Phase

Oppenheim and Schafer observe that any rational system function can be broken into a minimum phase and an all pass system [16]. An all pass system can de defined as one which has a magnitude response that is absolutely constant with respect to frequency [17]. Therefore, the magnitude of the minimum phase all pass decomposition is represented solely by the minimum phase system and the phase is reconstituted by both the allpass and minimum phase representations.

The system in question can thus be defined as:

$$H(z) = Hmin(z)Hap(z), \qquad (4.1)$$

where Hmin(z) is a minimum phase system, and Hap(z) is an all pass system.

Typically, magnitude and phase spectra are not related. A unique and, in this case, extremely useful property of minimum phase systems, however, is that phase values for each component frequency can be derived from the corresponding magnitude values, see [16] for details.

5. Minimum Phase and HRTFs

The significance of phase information and the auditory system's limitations in responding to changes in phase information has been investigated in depth, for example, see [9]. In

[11], it is observed that the auditory system approaches minimum phase. The authors decomposed their measured transfer functions into minimum phase and allpass functions in order to obtain a clear representation of phase without the above mentioned 2 pi ambiguity. While doing this, they realised that the minimum phase function appeared to contain almost all the detail of the phase spectrum and that the allpass phase approached linearity for the free field to ear canal function. The paper goes on to assert that the allpass component of the full HRTF (including the ear canal response, as defined by Begault[1]) exhibits a 'nearly linear' phase response up to 10 kHz. Therefore, the allpass component can be implemented as a simple time delay. This time delay can be realised using a time domain, frequency independent delay line, quite a simple and efficient process to implement. This observation of approximate minimum phase has become a key factor in binaural HRTF based processing, and has been used in several studies of HRTFs, many of which suggest HRTF models based on minimum phase plus delay decomposition, such as principal component analysis [6] and infinite impulse response models [8].

The minimum phase and (assumed linear) all pass decomposition allows a pair of HRTFs (for the left and right ears) to be broken down into 3 parts: a minimum phase representation of each empirical HRTF pair (left and right ear), and an interaural delay. The overall magnitude will be represented by that of the minimum phase filter; the overall phase will be the minimum phase phase spectrum (derivable from the magnitude spectrum) plus a frequency independent, linear delay. Thus phase interpolation is no longer a problem. Figure 2 shows an empirical impulse and its minimum phase representation in the time domain.

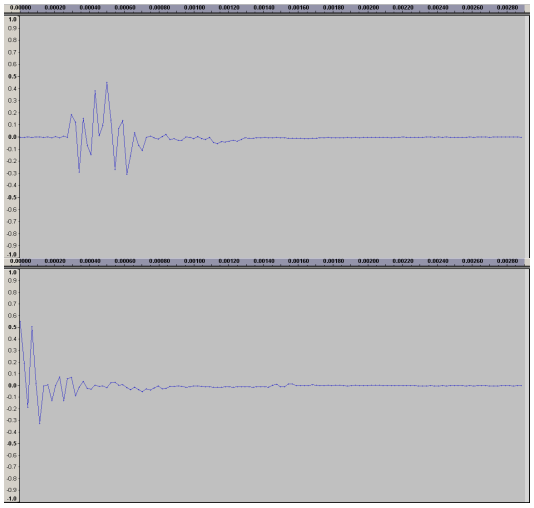

Figure 2: An Empirical HRTF (source at 0 degrees) and its Minimum Phase Representation

The process of HRTF interpolation consequently involves analysing each HRTF pair to find the relevant interaural delay and reducing them to minimum phase representations. The minimum phase magnitude values and extracted delay can then be linearly interpolated. Interpolated minimum phase phase spectra can be derived from interpolated magnitude spectra.

The description of HRTFs as minimum phase filters plus delays above is validated theoretically by work on decomposition of impulses in [11]. However, perhaps a more pertinent validity test from the point of view of a developer of artificial spatialisation tools involves psychophysical testing of a subject group. Kulkarni's seminal paper examining the sensitivity of human subjects to HRTF phase spectra [9] reports high coherence values between empirical and minimum phase plus delay data sets. However, coherence values were systematically worse at lower elevations and extremes of the horizontal plane. It is suggested that this is due to the shadowing effect of the head and interactions with the torso making the allpass delay non

[1] Begault defines the hrtf as 'the spectral filtering of a sound source before it reaches the ear drum that is caused primarily by the outer ear' in [1]. However, it is undesirable to use hrtfs that contain the auditory canal response of the dummy head in artificial localisation applications, as the listener, using headphones that transmit audio from the entrance of the ear canal, is then essentially listening through 2 auditory canals, that of the dummy head and their own. This is avoided in the MIT dataset used here through diffuse field equalization.

linear, a phenomenon discussed in [7]. This is supported by better performance at higher elevations, where there is less obstruction in the path to the contralateral (further from the source) ear. Phase error results enforce this assumption.

Psychophysical results point to a low frequency cue present at extremes of the horizontal plane, aiding the subject in distinguishing between min phase plus delay and empirical impulses. This perhaps suggests that modelling itd as a linear delay is not adequate; however, overall the study concludes that minimum phase plus delay models are sufficient for most locations, and that the finer structures of phase are not overly important, as long as the overall delay is approximated in accordance with that of the empirical.

Practicalities in implementation of a minimum phase based spatialisation system also need to be considered. The tiny delays extracted from the HRTF data set will often fall in between the sample by sample values used in a delay line. This is a consequence of digital delay lines and sampling. To remove abrupt changes in delay for moving sources, interpolated variable delay lines can be employed. As observed in [2], interpolated delay lines attenuate high frequencies, and are therefore not ideal. However, informal listening tests performed both in [2] and by the authors suggest that these artefacts are not significant.

Alternatives to the minimum phase approach are suggested that do not assume the approximations involved in modelling the HRTF as minimum phase plus delay. This essentially involves engaging more directly in the phase ambiguity problem. The minimum phase method employs complex digital signal processing of the HRTF data, and is quite computationally expensive. The approaches outlined below are intended to give spatially accurate and efficient processing without the necessity to perform complex analysis, compression or transformation of the data.

6. Current Csound Solution to Binaural HRTF Based Processing

The HRTFer opcode uses the MIT data set (see [3]) to spatialise the desired source sound.

aleft, aright **HRTFer** ainput, kangle, kelevation, "HRTFcompact"

HRTFer provides accurate spatialisation for static locations which correspond exactly to HRTF measured points. However, if a static point is required that has not been measured, the system simply chooses the nearest measured point. Considering the density of this data set these inaccuracies may be tolerable for certain situations. This approach causes more significant errors when a specific trajectory is desired for a source. A dynamic, rather than static source will skip from one nearest measured point to the next along a user defined trajectory. This staggered movement causes irregularities in the output, manifesting themselves as discontinuities, an undesirable result. The original authors of the opcode suggested a fade out of the old convolution result and a fade in of the new to reduce this noise. However, these crossfades have been disabled, as they cause dropouts in the output, leading to worse irregularities, which are assumed to be caused by an error in the source code. In tests performed by the authors, these crossfades, when implemented, reduce the irregularities to a degree depending on the frequency content of the source.

Another consequence of abruptly changing these complex filters (HRTFs) as a source travels along a defined trajectory is the sudden perceptual change in the output, which can be detrimental even in frequency rich sources (which may mask discontinuities to a certain extent). For example, in a trajectory going from 50 degrees above the listener to directly in front, the source will appear to jump downwards every 10 degrees, as this is the measurement increment. Clearly, this opcode could benefit from the addition of interpolation between measured points.

7. Novel Solutions to HRTF Binaural Processing

Two novel approaches to HRTF binaural processing are presented below.

7.1 Phase Truncation, Magnitude Interpolation

The first suggested approach can be summarised as magnitude interpolation and phase truncation. It can be simply defined as using interpolated spectral magnitude values, and the nearest known phase values to derive the impulse for each block of audio processed. This approach, as well as providing an adequate solution to HRTF interpolation for sources to be placed at non measured points, allows artefact free, user defined source trajectories.

The movement in the program is achieved by updating user defined angle and elevation values, according to where the source is moving from and to, every processing block.

The interpolation algorithm works by storing the four nearest HRTF values to the desired location, left and right below and above. Linear interpolation of the magnitude values is performed. This magnitude interpolation method derives an accurate intermediate/transitional fir filter, essentially boosting/attenuating spectral bands to a level that is proportionate to and sympathetic with the nearest measured points. For source trajectories adhering to minimum audible movement angle constraints (see [13]), noise introduced by filter magnitude values changing as the source moves is inaudible/tolerable.

The nearest measured phase value is used for intermediate filters. As the difference in measured points is often quite minimal, although always significant, it is proposed that choosing the phase of the nearest measured point will not have a significant adverse effect on the final spatialisation quality. As discussed above, studies have shown that phase does indeed play an important role in localisation, but as long as an accurate overall itd is maintained, users frequently cannot distinguish errors.

However, as the source trajectory moves closer to a different measured point, these phase values need to be updated. An abrupt switch of phase values will cause an audible inconsistency in the output. Although the severity of this inconsistency depends on the source material to be spatialised, a method to minimise it is desirable. The crossfade method suggested by the csound HRTFer opcode is considered and developed. Fades are performed when new, nearer *phase* values are available. This approach, coupled with magnitude interpolation, gives much smoother movement. The frequency content of the source defines the audibility of the switch between phase values. If a narrowband source is to be spatialised (i.e. a source with energy focused in a small number of narrow frequency bands), the switch will be quite obvious. However, more noisy, frequency rich sources may be able to mask the inconsistency caused by the new phase values to an extent related to the complexity of the source. It is with this in mind that a user definable, source specific solution is proposed. The user may choose to perform crossfades over a number of processing buffers. One buffer may be enough to mask unwanted inconsistencies for certain sources, whereas others may require up to 16 buffers to mask all artefacts.

The process of crossfading thus involves processing the old HRTF data with the new input data and fading out, while processing the new HRTF data with the input and fading in.

Another point to consider is that for very fast trajectories, the nearest measured phase values may be changing quite swiftly. However, considering the minimum audible movement angle (see [13]), and that only audible trajectory changes are desirable, the system is adequate. A related criticism, however, is that occasionally a path may be required which causes the angle and elevation index to change over the same crossfade. If the path involves a three dimensional trajectory, phase value updates may not be uniform. Users will be warned in this scenario, and can simply reduce the crossfade size. As mentioned previously, the spectral content of the input sound may mask discontinuities, so shorter crossfades will suffice in certain situations. Figure 3 illustrates the magnitude interpolation, phase truncation process for a moving source.

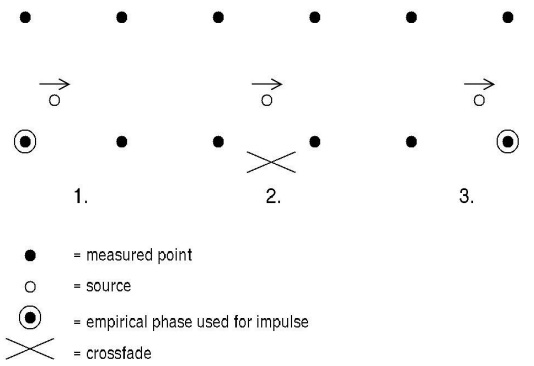

Figure 3: Magnitude Interpolation, Phase Truncation

A minimum phase based implementation is also developed using the model discussed in [9]. Essentially, magnitudes are interpolated as above, and phase is derived from these interpolated magnitudes. A linear allpass delay is inserted using a variable delay line. Minimum phase or phase truncation based processing can be chosen by the user as an optional parameter of the opcode developed.

7.2 Functional Phase Model

Another approach, essentially a hybrid of an empirical and modelled transfer function is presently suggested. As discussed above, spectral magnitude measurement, representation and interpolation is straightforward and easily realisable. Therefore empirical magnitude values are employed here. The difficulty of phase representation and interpolation is approached from a functional modelling point of view.

The main task in functional phase derivation is to model correctly the interaural phase difference, therefore deriving the correct interaural time difference. A basic, yet practical model for the head is to assume it approximates a sphere. The degree to which this phase simplicity will distort the spatial image is closely related to the discussion above on sensitivity to phase differences, which concluded that low frequency itd across frequency is the predominant phase cue (see [9]).

Mathematically, the itd for a particular source location, assuming a spherical head can be defined thus:

$$\frac{r(\vartheta + \sin \vartheta)}{c} \cos \phi \,, \qquad (7.2.1)$$

where r is the head (/sphere) radius, c is the speed of sound, ϑ is the angle and ϕ the elevation of the source. This formula is described as the Extended Woodworth/Schlosberg Formula in [12].

Successful use of this basic Woodworth model for HRTF phase modelling and a magnitude interpolation algorithm is reported in [19]. Some development and improvement is suggested here. As concluded in [9], low frequency consistency of empirical and employed itd is crucial for accurate modelling. Also, it is agreed in both [9] and [7] that higher frequency itd is not as significant, and specified in [7] that a Woodworth model can account for steady state high frequency itds. Also, physiologically, interaural phase difference based localisation breaks down above approximately 1500 Hz (see [13]). Therefore a low frequency, frequency dependent scaling factor is introduced as a more complete solution, requiring minimal extra processing. Essentially, itd is extracted from the empirical HRTFs for each low frequency band of interest. These values are then used as frequency dependent scaling factors in the synthesis of the phase spectrum for the desired HRTFs.

This model provides an accurate average low frequency itd for this particular dataset, and a steady Woodworth based itd for higher frequencies.

The values derived from this Extended Woodworth/Schlosberg non linearly low frequency scaled (/functional) model are then used as phase values. Phase values are calculated per frequency bin, with values of minus and plus half the itd for the ear nearer and further from the source respectively. Practically, negative phase values simply wrap around to the end of the impulse. It therefore appears that the nearer ear impulse happens after the further ear, which is an unnatural result. For this reason, the impulse is shifted in

time, by half the size of the buffer. The result is a time and phase accurate filter.

Phase interpolation for dynamic sources has been discussed and a solution presented in the form of phase truncation. However, with the Woodworth functional approach, a new phase can be derived for any location, and can be used and updated for each processing block of a dynamic source. This is an initially exciting prospect; however implementation illustrates undesirable noise, caused by phase updates, and phase not 'matching' magnitudes, as it does in minimum phase implementations.

The short time Fourier transform (stft, see [14]) is employed to avoid the irregularities introduced by changing modelled phase per processing block.

8. Csound Implementations

Three plugin opcodes are designed using the guidelines in [10], one allowing phase truncation or minimum phase binaural processing, and two based on the functional model. The reason for two opcodes based on the functional model is due to the efficiency with which a static source can be processed in comparison to the necessity of stft processing for a dynamic source. The phase truncation/minimum phase model allows the user to choose between minimum phase and phase truncation processing, the latter also allowing user defined crossfade sizes. Functional models allow choice of spherical head radius for itd calculation, and stft overlap for dynamic trajectories. All models allow sampling rates of 44.1, 48 and 96 kHz. Data files containing the HRTF data at the appropriate sampling rate, as well as minimum phase delay data are also required.

Despite the addition of magnitude interpolation, and algorithms for appropriate phase representation, the new, optimised opcodes perform favourably in comparison to the HRTFer opcode. For example, the phase truncation process takes an approximate average of .11 seconds of CPU time to spatialise 2 seconds of audio on a 0 to 90 degree trajectory. HRTFer takes an approximate average of .16 CPU seconds to perform the same operation. This figure is comparable with minimum phase processing time for the same trajectory. To place this source statically with the functional model takes just .07 CPU seconds, but to perform the trajectory above with the functional model takes .17 seconds due to the addition of the stft processing. Note: default opcode values were used for the above approximate average csound CPU time tests (crossfade over 8 buffers for phase truncation, head radius of 9cm for the functional models, overlap of 4 for the stft and sampling rate of 44.1 kHz for all).

9. Conclusion and Discussion of Methods Employed

As discussed in [9], HRTF phase data does not require exact accuracy. More specifically, maintaining low frequency interaural time delays appears to provide accurate phase data. The phase truncation method described maintains nearest measured phase data, thus meeting this criterion. The goal of the method: to use the data directly, is also achieved. A generic, user definable model is presented to allow for compromise between complex trajectories, narrow band sources and changing phase noise removal through variable length cross fades.

Minimum phase requires data preparation and knowledge of complex digital signal processing. Furthermore, casual listening tests show there is often an audible discrepancy between minimum phase and empirical data convolution for musical and test sources, although localisation is good for both. Phase truncation output appears to give a result more consistent with the empirical dataset as a whole.

Functional models introduced above assume the head is a sphere and will be accurate to this degree, but adding non linear low frequency scaling factors will reintroduce some of the finer phase detail involved in the non uniformly spherical shape of the head, the pinnae and the torso.

The functional model implementation provides a more mathematical approach, with the addition of the specifics of the data set used, implemented in an efficient and psychoacoustically consistent way. The importance of low frequency phase

information is preserved and applied to an efficient, simple model for phase. This provides a speedy solution for static sources; however dynamic sources require stft processing.

The binaural processing capabilities of csound have thus been updated and improved, using existing and novel approaches. Smooth, artefact free dynamic and static binaural processing is now realisable using the various techniques described above.

Acknowledgements

This work is supported by the Irish Research Council for Science, Engineering and Technology: funded by the National Development Plan and NUI Maynooth.

References

[1] Durand Begault. *3-D Sound for Virtual Reality and Multimedia*. AP Professional, London, 1994.

[2] Cheng and Wakefield. Moving Sound Source Synthesis for Binaural Electroacoustic Music Using Interpolated Head-Related Transfer Functions (HRTFs). *Computer Music Journal*, 25:4: 57–80, 2001.

[3] Gardner and Martin. HRTF Measurements of a KEMAR Dummy Head Microphone (http://sound.media.mit.edu/KEMAR.html, accessed July 2007) MIT, 1994.

[4] Hartung, Braaschand Sterbing. Comparison of Different Methods for the Interpolation of Head Related Transfer Functions. *AES 16th International Conference: Spatial Sound Reproduction*, 319-329, 1999.

[5] Jot, Larcher and Warusfel. Digital Signal Processing Issues in the Context of Binaural and Transaural Stereophony. *AES 98th Convention,* 1995.

[6] Kistler and Wightman. A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction. *Journal of the Acoustical Society of America* Volume 91(3): 1637-1647, 1992.

[7] Kuhn. Model for the interaural time difference in the azimuthal plane. *Journal of the Acoustical Society of America* Volume 62(1): 157-167, 1977.

[8] Kulkarni and Colburn. Infinite-impulse-response models of the head-related transfer function. *Journal of the Acoustical Society of America* Volume 115(4): 1714-1728, 2004.

[9] Kulkarni, Isabelle and Colburn. Sensitivity of Human Subjects to Head-Related Transfer-Function Phase Spectra. *Journal of the Acoustical Society of America.* Volume 105(5): 2821-2840, 1999.

[10] Lazzarini. Extensions to the Csound Language. *Linux Audio Conference*, 13-19, 2005.

[11] Mehrgardt and Mellert. Transformation Characteristics of the external human ear. *Journal of the Acoustical Society of America.* Volume 61(6), 1977.

[12] Minnaar, Plogsties, Olesen, Christensen and Moller. The Interaural Time Difference in Binaural Synthesis. *AES 108th Convention,* 2000.

[13] Moore. *An Introduction to the Psychology of Hearing* Elsevier Academic Press, London, 1977; 5th edn, 2004.

[14] Moore: *Elements of Computer Music.* Prentice-Hall, New Jersey, 1990.

[15] Noisternig, Musil, Sontacchi and Holdrich. 3D Binaural Sound Reproduction using a Virtual Ambisonic Approach. *IEEE Symposium on Virtual Environments*, 174-178, 2003.

[16] Oppenheim and Schafer: *Discrete-Time Signal Processing*. Prentice Hall , New Jersey, 1989; 2nd edn, 1999.

[17] Steiglitz. *A DSP Primer*. Addison-Wesley, Clifornia, 1996.

[18] Wenzel, Arruda, Kistler, and Wightman. Localization using non-individualized head related transfer functions. *Journal of the Acoustical Society of America* Volume 94(1): 111-123, 1993.

[19] Zotkin, Duraiswami and Davis. Rendering Localized Spatial Audio in a Virtual Auditory Space, *IEEE Transactions on Multimedia*, Volume 6(4), 553-564, 2004.