Dynamic Modeling of the Pinna for Audio Spatialization

ARMANDO BARRETO[1,2] and NAVARUN GUPTA[1]
[1]Electrical & Computer Engineering Department
[2]Biomedical Engineering Department
Florida International University
Miami, Florida, 33174
USA
{barretoa,ngupta02}@fiu.edu http://dsplab.eng.fiu.edu

Abstract: - This paper proposes a dynamic model for the human pinna, or outer ear. This dynamic model is needed to complete the implementation of structural models used for digital audio spatialization. The paper describes the rationale followed to derive this model and the process by which the parameters of the model can be instantiated for any given prospective user of the audio spatialization system. The results obtained with the model demonstrate that this is a plausible modeling solution for the acoustic performance of the pinna, when the sound sources are in the frontal plane.

Key-Words: - Dynamic Model, HRTF, HRIR, Pinna, Prony's Method, Sound Spatialization

1 Introduction

Due to the proliferation of Personal Computers (PCs), video games and virtual reality systems, there is an increasing interest in perfecting the process of sound spatialization, i.e., the process by which a digital sound (such as a "wave" file), originally lacking any directionality, is made to appear for a listener as if originating from a source at a definite virtual location [4].

Currently, many sound spatialization systems use the concept of Head-Related Transfer Functions (HRTFs), as their basis. HRTFs represent the transformation undergone by the sound signals, as they travel from their source to both of the listener's eardrums. This transformation is due to the interaction of sound waves with the torso, shoulder, head and outer ear of a listener [12]. Therefore, the two components of these HRTF pairs (left and right) are typically different from each other, and pairs corresponding to sound sources at different locations around the listener are different. Furthermore, since the physical elements that determine the transformations of the sounds reaching the listener's eardrums (i.e., the listener's head, torso and pinnae), are somewhat different for different listeners, so should be their HRTF sets [5].

Current sound spatialization systems use HRTFs, represented by their corresponding impulse response sequences, the Head-Related Impulse Responses, (HRIRs) to process, by convolution, a single-channel digital audio signal, resulting in the two components (left and right) of a binaural spatialized sound. When these two channels are delivered to the listener through headphones, the sound will seem to emanate from the source location corresponding to the HRIR pair used for the spatialization process. To have the ability to "virtually place" a sound in many locations around a listener, sound spatialization systems must have access to a library of HRIR pairs. These HRIR pairs are commonly obtained by empirical measurement in an anechoic chamber. A miniature microphone is placed at the entrance of the ear canal of the experimental subject and broad-band audio signals (Golay or MLS Codes) are played from a small speaker, meant to emulate a point source. The discrete-time sequences obtained by digitizing the sounds collected by the ear microphones are used as the HRIRs associated with the speaker location used, for the subject participating in the measurement. Typically, this experiment is repeated at the combinations of many (e.g., 12) azimuth angles, θ, (angle between the direction in front of the subject and the line that connects the subject location with the source location), and many (e.g., 6) elevation angles, ϕ, (angle between the line between subject and source, and the horizontal plane, at ear-level).

Unfortunately, the facilities, equipment and expertise required to measure the HRIRs of a subject, in the fashion described above, make the determination of HRIRs for each potential user of the sound spatialization system impractical for all but the high-end, purpose-specific systems [3]. For most consumer-grade applications, sound spatialization systems resort to the use of "generic" transfer functions, measured from a manikin with "average" physical characteristics [8], which, evidently is a fundamentally imperfect approach. Furthermore, the recording of empirical HRIR sequences from a given subject, or a manikin, lacks any flexibility, since it is impossible to adjust the 128-point or 256-point HRIR sequences recorded experimentally to make them more adequate to a specific intended user of the spatialization system.

Recently, an alternative approach to digital sound spatialization has emerged, in which the audio signal being spatialized is not just directly convolved with pre-recorded left and right HRIRs to emulate a certain azimuth and elevation. Instead, the digital audio sequence is processed by a pair of dynamic systems, which model the physical effects of the head, the torso and the outer ear on the sound, as it travels from a source location to the listener's eardrums. These "structural" spatialization models [6][7], typically omit the involvement of the shoulders in the spatialization process, as it has been recognized to be comparatively minor [10]. The impact of the head is modeled through two independent model blocks. The first is a "Head-Shadow" module, which basically re-sizes the sound depending on how indirectly it reaches the eardrum represented by the model. This establishes the amplitude unbalance between eardrums, or Interaural Intensity Difference (IID), which is an important cue for localization in the horizontal plane. Each of the two branches (left and right) of structural models also includes a variable delay which models the Interaural Time Difference (ITD), i.e., the difference in the arrival time of a sound wavefront to the left and right eardrums. This is the second important cue for sound source localization in the horizontal plane.

Both IID and ITD depend mostly on the distances between the source location and both eardrums. Therefore, these quantities are shared by many points in the space around the listener. The geometric locus of all these points is a cone that has its axis aligned with the inter-aural axis and its apex in the listener's ear. Because sources located in this cone cannot be distinguished by IID and ITD emulation alone, this cone has been called "Cone of Confusion" [Mills, 1972]. In particular, in the frontal plane (i.e., the vertical plane that contains the inter-aural axis), the cone of confusion causes the miss-localization of virtual sounds emulated as originating above ear level, which are mistakenly perceived as originating in a symmetrical location below ear level, incurring in a "reversal" of the perceived elevation.

However, it is well established that humans can discern the elevation of sound sources, particularly in the frontal plane. In fact, it is known that such elevation determination can be made on a monaural basis, i.e., listening with just one ear [17]. Therefore, a successful structural model for digital audio spatialization must incorporate in each one of its branches a pinna sub-model, capable of establishing monaural cues for the discernment of localization, particularly elevation.

2 Problem Formulation

According to the previous observations, this paper reports our work in defining a plausible pinna dynamic model, to complete a structural model for audio spatialization. More importantly, this paper outlines the mechanism used to instantiate the parameters of the pinna model from measured HRIRs. This systematic approach to the instantiation of the model proposed ensures the availability of a model instance for each azimuth and elevation, and for each subject. Moreover, this duality also enables the assessment of the effectiveness of the model, as compared to the measured HRIRs.

2.1 Previous Pinna Models

In his study of the auditory localization properties of the outer ear, or pinna, Mills distinguishes two types of effects [13]:

- Pinna Shadows, representing the higher level of attenuation imposed on sounds originating behind the listener [9].
- Pinna Reflections, representing the indirect paths followed by sound, towards the entrance of the ear canal.

Mills refers to the initial model proposed by Batteau [2], including a direct path (unity gain) in parallel with two delayed paths. In each of these delayed paths sound is affected by a reflection coefficient, ρ and lags by a delay τ. Batteau's model is illustrated in Figure 1.

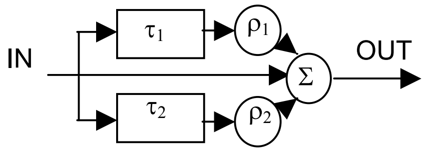

Figure 1: Batteau'sl Reflective Pinna Model

Using a 5x model of a human pinna, Batteau recorded the responses measured in the ear canal for impulses at various azimuths (θ) an elevations (ϕ), and "identified a waveform peak that was delayed by amounts that varied systematically with azimuth; another feature varied, although less distinctly, with elevation" [13]. Scaled back to the normal size ear, "the delay of the first reflection dropped from 80 to 10 μs as the azimuth angle of the source increased from 10° to 100°. The delay of the second reflection increased from 100 to over 300 μs when the source moved from above via the side to below." [10]. Batteau developed a three-path acoustic coupler to simulate his proposed pinna model, in which the delays were such that the shorter delay τ_1, encoded azimuth (θ) as:

$$\tau_1 = (s_1/c)\,(1 + \cos\theta) \qquad (1)$$

and the longer delay, τ_2, encoded elevation (ϕ) as:

$$\tau 2 = (2\,s_1/c) + (s_2/c)\,(1 + \sin\phi) \qquad (2)$$

In these equations s_1 and s_2 are the separations between two holes in the coupler and a third, reference hole.

Watkins [17], was able to verify the perception of vertical displacement of a sound source when he implemented Batteau's model, assigning unity value to all the gains in the system ($\rho 1 = \rho 2 = 1$), and keeping the shorter delay, supposed to encode azimuth, constant, ($\tau_1 = 15$ μs), while varying the second delay, τ_2, between, 100 and 300 μs. Most importantly, he performed experiments confirming the existence of a relationship between delay and perceived elevation in both cases.

Recently, Brown and Duda [6][7] used a model inspired in Batteau's, augmented to comprise not two but five delay-and-add branches, for which the delays where determined according to an equation that mixes both of Batteau's equations:

$$\tau_k\,(\theta, \phi) = A_k \cos(\,\theta / 2)\,\sin\,(D_k\,(90^{o} - \phi)) + B_k\,, \quad \text{for} \quad k = 1,\ldots,5 \qquad (3)$$

In these expressions A_k, B_k and D_k are constants, which, along with the five values of the reflective coefficients, ρ_k, were set in the model in an *ad hoc* fashion. In fact, all the constants in this model were kept the same when performing localization tests with two subjects, but the five D_k parameters had to undergo *ad hoc* adjustment for a third subject. It is also noteworthy that they used 3 reflective coefficients with positive values and assigned negative values to the other 2. This ad hoc procedure to instantiate the pinna model highlights the need to establish a systemic method to obtain an appropriate pinna model.

2.2 Resonant Component in the Pinna

In addition to the physically-plausible sound reflections that must be present in a pinna model, observation of numerous measured HRIRs indicated the likely presence of a resonant component that must also be included in a dynamic model for the pinna. For example, Figure 2 shows an HRIR sequence obtained from the MIT database collected by Gardner and Martin [8]. A dashed line has been drawn in this figure to outline the seemingly exponential decay in this HRIR. Additionally, there is evidence in the literature that several cavities of the outer ear (e.g., Concha Cavum, Fossa) act as resonators [10][16].

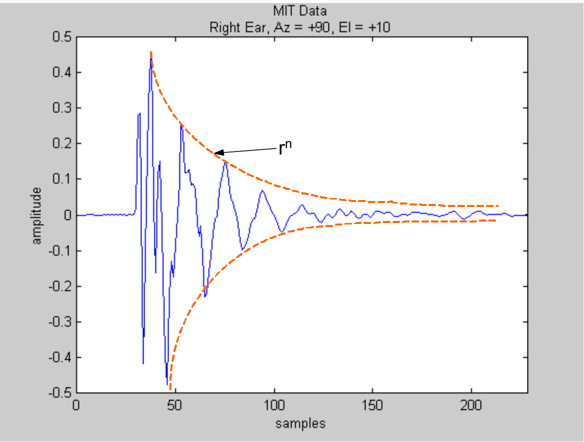

Figure 2: HRIR illustrating resonant components

2.3 Proposed Pinna Model

According to the discussion above, we propose a model that is both resonant and reflective, allowing for a resonant effect in a "direct path" component of the HRIR, but also in up to three indirect-path components or "echoes". This model is diagrammed in Figure 3.

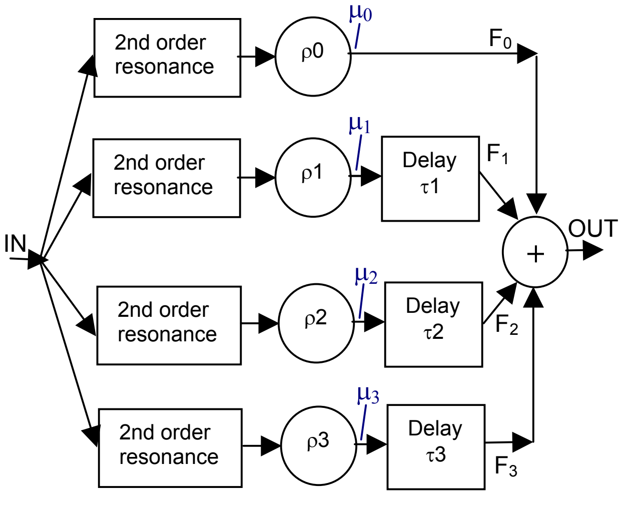

Figure 3: Proposed Pinna Model

It should be noted that this pinna model allows for the "direct-path", F_0, and each one of the "echoes", F_1, F_2, and F_3, to be affected by a different equivalent resonance.

Accordingly, HRIRs are envisioned as the impulse response of this model, which will be the superposition of four damped sinusoidals (the impulse responses of each of the 2nd order resonators), characterized by their frequency, f, and damping factor, σ. These damped sinusoidals are altered in their amplitude according to the ρ_k parameters, and delayed according to the τ_k parameters.

Thus, the instantiation of this proposed model will require the identification of the f_k, σ_κ, ρ_κ and τ_κ values, to characterize the parameters of the model that successfully approximates an HRIR collected for a given azimuth and elevation, through the output provided by the pinna model

3 Problem Solution

The deconstruction of the HRIR sequence into the four proposed components F_0, F_1, F_3 and F_4, affected by the appropriate values of their τ parameters (see Figure 3), is a challenging deconvolution process. The main difficulty in this operation is the fact that the several damped oscillations are mixed together, partially overlapping in time, in the measured HRIR.

3.1 Sequential Prony Modeling

This problem was addressed by the sequential application of Prony's modeling algorithm [11][14][15], to partial segments of the response. Prony's method approximates a given signal $\mu(t)$ as the superposition of p damped sinusoidals:

$$\mu(t) = \sum_{j=1}^{p} \rho_j e^{(\sigma_j t)} \sin(2\pi f_j t + \xi_j) \qquad (4)$$

However, Prony's method assumes that all p components start at the same instant (the beginning of the segment under modeling). So, the overall model had to be obtained in an iterative process that involved (starting with $t_0 = 0$):

[a] Apply Prony's method for the fitting of a single damped sinusoidal to progressively larger windows of observation, which extend from the beginning of the current modeling segment (t_m). Monitor the average residual error as the window is increased in size.

[b] When a sudden increase in modeling error becomes apparent, reduce the window size to the value that yielded the minimum in the modeling error prior to its sudden increase. Label that latency as t_{m+1}.

[c] Extrapolate the single damped sinusoidal modeled by Prony's method between t_m and t_{m+1}, to extend for 256 points. This sequence will model the m-th replica of the resonant response, $\mu_m(n)$. Assign ρ_m = as the peak amplitude of $\mu_m(n)$.

[d] Subtract $\mu_m(n)$ from the current modeling segment. Shift the resulting difference sequence t_{m+1} samples to the left (discarding samples that are re-assigned to negative time indices, and filling with zeros on the right of the sequence), so that the sample labeled t_{m+1} will now be at the beginning of the 256-sample shifted difference. This sequence will be the new modeling segment that will be used in the next iteration.

[e] Repeat from step [a].

Each completion of the iteration indicated above yields:

- A damped sinusoidal sequence, which is the m-th replica of the resonator response, $\mu_m(n)$
- The intensity of one of the echoes in the pinna model, ρ_m ,
- The overall latency of that same echo, as $\tau_m = t_0 + t_1 \ldots + t_m$.

Corresponding to the physical situation being modeled, it is observed the every new echo found is less and less significant to the overall composition of the HRIR under modeling. In order to keep the number the parameters in the model low, it was decided to include only the first four components (m = 0, 1, 2, 3) in the modeling process, as indicated in Figure 3. This decision was based on the average value of the τ_3 latencies found, which was approximately 300 μs, coincident with the upper bound of elevation-related delays observed by Mills [13] and Han [10]. To verify the adequacy of the sequential decomposition described above, a reconstructed waveform, R(n), can be synthesized by the superposition of the shifted sinusoidals obtained from each Prony stage:

$$R(n) = F_0(n) + F_1(n) + F_2(n) + F_3(n) \qquad (5)$$

$$R(n)=\mu_0(n)+\mu_1(n-\tau_1)+\mu_2(n-\tau_2)+\mu_3(n-\tau_3) \qquad (6)$$

Figure 4 shows the four F components, obtained using this method, from the measured HRIR displayed in Figure 5. Figure 5 also shows (dashed line) the reconstructed R(n), superimposed to the original HRIR.

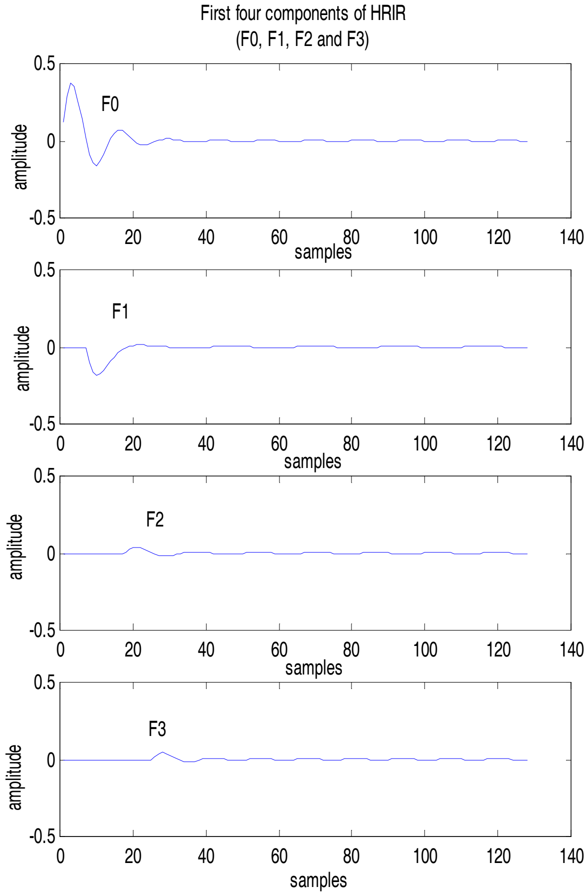

Figure 4: Example of HRIR Deconstruction

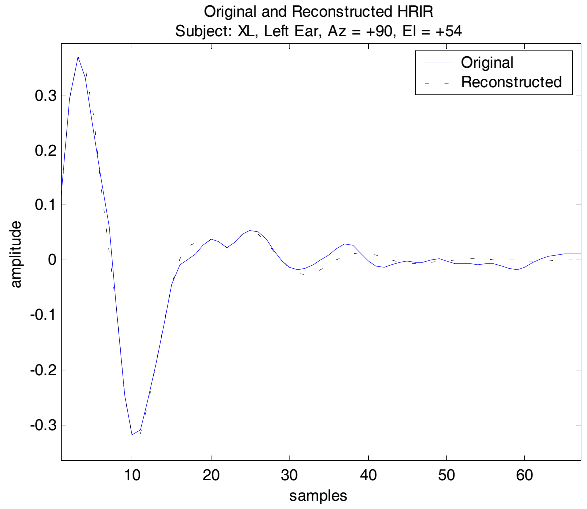

Figure 5: Example of HRIR approximation through the superposition of F_0, F_1, F_2 and F_3.

3.2 Model Evaluation and Results

The goodness of fit of each model output (superposition of the four model partial outputs, F_0, F_1 , F_2 and F_3), to the original measured HRIRs was assessed in terms of the ratio of the sum of squares (SS) of the error to the sum of squares (SS) of the original HRIR:

$$\text{Match}=\{1-[SS\{HRIR(n)–R(n)\}/SS\{HRIR(n)\}]\} \quad (7)$$

This “match” figure was expressed as a percentage. For example, the close approximation shown in Figure 5, along with the original HRIR, achieves a match figure of 97.8%.

In order to evaluate the viability of the model, HRIRs for the right and left ears of 15 volunteers, for sources in the frontal plane (ipsilateral to the ear considered), and at elevations of ϕ = -36°, -18°, 0°, 18°, 36°, and 54°, were modeled with the sequential Prony procedure outlined above. The original HRIRs from the process were obtained as minimum-phase, 256-point impulse response sequences, using the HeadZap HRTF system, by AuSIM 3D [1], at a sampling rate of 96 KHz. A reconstructed HRIR, R(n), sequence was created from each of the models formed, and the level of match of this R(n) to the original, measured HRIR was determined.

Overall, for the 180 HRIRs reconstructed in this study, the average percentage of match obtained was 93%. Table I shows the percentage of match obtained for HRIRs, grouped by elevation angle.

Table I: Average match between Reconstructed and Original HRIRs

Elevation	54°	36°	18°	0°	-18°	-36°
% Match	95%	95%	92%	93%	93%	91%

The percentages of match in this table range between 91% and 95%. This indicates that the proposed model is, in fact, capable of creating modeled HRIR sequences that closely resemble the ones that would be obtained by empirical measurement, provided it is instantiated with the proper model parameters. It should also be recalled that the model reconstruction process is being limited to involve only the first four components, which may leave some later features of the HRIR unrepresented. Overall, this suggests that the model proposed is, in fact, a viable one.

4 Conclusion

This paper has presented a proposed functional model of the pinna, to be used as the output block in a structural sound spatialization model. The definition of the model, containing second order resonances and "echoes" at varied latencies, was introduced and justified in terms of the sequential Prony deconstruction from the measured HRIRs of 15 subjects, at 6 different elevations. The values of the model parameters that emerge from this process were used to confirm that HRIRs reconstructed by instantiating the model with those recovered parameter values yielded high (Match > 90%) levels of agreement with the corresponding measured HRIRs.

The dynamic model of the pinna presented in this paper may serve as the basis for a customizable spatialization system, as the model parameters for each azimuth and elevation must be related to anthropometric features of the intended listener's outer ears. Work is underway to establish predictive equations that will allow the instantiation of the model parameters, given a small set of anthropometric measurements of a prospective user of the sound spatialization system.

5 Acknowledgements

This work was sponsored by NSF grants EIA-9906600 and HRD-0317692, and ONR grant N00014-99-1-0952.

6 References

[1] AuSIM, Inc., "HeadZap: AuSIM3D HRTF Measurement System Manual". AuSIM, Inc., 4962 El Camino Real, Suite 101, Los Altos, CA 94022, 2000.

[2] Batteau, D. W., "The Role of the Pinna in Human Localization", Proc. R. Soc. Lon. B., vol. 168, pp. 158-180, 1967.

[3] Begault, D. R., "A head-up auditory display for TCAS advisories." Human Factors, 35, 707-717, 1993.

[4] Begault, D., "3-D Sound for Virtual Reality and Multimedia", Academic Press, 1994.

[5] Begault, D. R., Wenzel, E. M. and Anderson, M. R., "Direct Comparison of the Impact of Head Tracking, Reverberation, and Individualized Head-Related Transfer Functions on the Spatial Perception of a Virtual Speech Source", J. Audio Eng. Soc., Vol. 49, No. 10, pp. 904-916, 2001.

[6] Brown, C. P. and Duda, R. O., "An Efficient HRTF Model for 3-D Sound," Proc. 1997 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, Mohonk, NY, Oct. 1997.

[7] Brown, C. P. and Duda, R. O., "A Structural Model for Binaural Sound Synthesis", IEEE Trans. Speech and Audio Processing, Vol. 6, No. 5, pp.476-488, September 1998.

[8] Gardner, W. G. and Martin, K. D., "HRTF measurements of a KEMAR". J. Acoust. Soc. Am., 97(6), 1995.

[13] Gupta, N., Ordonez, C. and Barreto, A. "The Effect of Pinna Protrusion Angle in the Localization of Virtual Sound in the Horizontal Plane" (Abstract) J. Acoustic Society of America, Vol. 110, No. 5, Part 2, November 2001, p. 2679.

[9] Han, H.L., "Measuring a Dummy Head in Search of Pinna Cues", J. Audio Eng., Soc., Vol. 42, No. 1 / 2 , pp. 15 – 37, 1994.

[10] Kahn, M., Mackisack, M. S., Osborne, M. R.,Smyth, G. K., "On the consistency of Prony's method and related algorithms", J. Comput. Graph. Statist., vol. 1, pp. 329-349, 1992.

[11] Kendall, G. S., "A 3-D Sound Primer: Directional Hearing and Stereo Reproduction", Computer Music Journal, 19:4, pp. 23-46, Winter 1995.

[12] Mills, A. W., "Auditory Localization," in Foundations of Modern Auditory Theory, Vol. II (J, V. Tobias, Ed.), pp. 303-348, Academic Press, New York, 1972.

[13] Osborne, M. R., and Smyth, G. K., "A modified Prony algorithm for fitting sums of exponential functions", SIAM J. Sci. Statist. Comput., vol. 16, pp. 119-138, 1995.

[14] Parks and C.S. Burrus, Digital Filter Design, John Wiley and Sons, p226, 1987.

[15] Shaw, E. A. G., Teranishi, R., "Sound pressure generated in an external-ear replica and real human ears by a nearby point source", J. Acoust. Soc. Am., vol. 44, No. 1, pp. 240-249, 1967.

[16] Watkins, A. J., "Psychoacoustical Aspects of Synthesized Vertical Locale Cues", J. Acoust. Soc. Am., vol. 63, no. 4, pp. 1152-1165, 1978.

[17] Wenzel, E. M., Arruda, M., Kistler, D. J., and Wightman, F. L., "Localization Using Nonindividualized Head-Related Transfer Functions", J. Acoust. Soc. Am., vol. 94, pp. 111-123, 1993.