Mixed structural modeling of head-related transfer functions for customized binaural audio delivery

Michele Geronazzo, Simone Spagnol, Federico Avanzini
Department of Information Engineering
University of Padova
Padova, Italy
{geronazzo,spagnols,avanzini}@dei.unipd.it

***Abstract*—A novel approach to the modeling of head-related transfer functions (HRTFs) for binaural audio rendering is formalized and described in this paper. Mixed structural modeling (MSM) can be seen as the generalization and extension of the structural modeling approach first defined by Brown and Duda back in 1998. Possible solutions for building partial HRTFs (pHRTFs) of the head, torso, and pinna of a specific listener are first described and then used in the construction of two possible mixed structural models of a KEMAR mannequin. Thanks to the flexibility of the MSM approach, an exponential number of solutions for building custom binaural audio displays can be considered and evaluated, the final aim of the process being the achievement of a HRTF model fully customizable by the listener.**

***Index Terms*—spatial hearing; binaural audio; HRTF**

I. Introduction

Spatial sound rendering is becoming increasingly important in several application domains, as it greatly enhances the effectiveness of auditory human-computer interfaces (particularly in cases where the visual interface is limited, as in mobile devices), and improves engagement and presence in augmented/virtual reality systems.

Binaural spatial sound is synthesized by convolving an anechoic sound signal with the left and right *Head-Related Transfer Functions (HRTFs)*, and by delivering the resulting stereo signal through headphones. HRTFs are defined as the frequency- and space-dependent acoustic transfer functions between the sound source and the eardrum [1]. Measuring individual HRTFs of a human subject is an expensive and time-consuming task. Alternatively, non-individualized HRTF sets can be recorded using "dummy heads" (mannequins with averaged anthropometric measures), at the expense of lower quality of the rendering and higher sound localization errors.

Several techniques for synthetic HRTF design have been proposed during the last two decades. These can be grouped into two main families: *pole-zero models* [2], in which the HRTF is approximated with low-order rational filters, and *series expansions* [3], in which the HRTF is represented as a weighted sum of simpler basis functions. On a different level of representation stand *structural HRTF models* [4]. In this approach, the effects of body components (head, pinnae, ear canals, shoulders/torso) are isolated and modeled separately with a corresponding filtering element. The global HRTF model is constructed by combining all the considered effects [5].

More recent research has focused on the problem of HRTF customization for individual subjects. Although most approaches use series expansions with self-tuning of weights [6], [7] or simply non-individualized HRTF selection [8], [9], [10], structural HRTF modeling remains the most attractive alternative in terms of both computational efficiency and physical interpretation: parameters of the rendering blocks can be estimated from real data, fitted to low-order filter structures, and related to anthropometric data [11], [12].

In this paper we propose a novel framework for synthetic HRTF design and customization, that combines the structural modeling paradigm with other HRTF selection techniques: namely, the *Mixed Structural Modeling (MSM)* approach regards the global HRTF as a combination of structural components, which can be chosen to be either synthetic or recorded components. In both cases, customization is based on individual anthropometric data, which are used to either fit the model parameters or to select a recorded component within a set of available responses.

The paper is organized as follows. Section II introduces the MSM formalism and presents a procedure for model evaluation. Section III discusses the main modeling and estimation techniques for the components of a MSM. Section IV provides two relevant examples of the proposed approach in which target responses are approximated using both modeled and selected components, customized according to individual data.

II. Mixed Structural HRTF Models

In its commonly accepted meaning, the term "head-related" transfer function indicates in fact the full "body-related" transfer function, that also includes acoustic effects of body parts different from the head. Based on this remark, we introduce two additional definitions.

Def. 1 A *partial head-related transfer function* ($pHRTF$) contains acoustic information either recorded by isolating specific body parts (e.g. pinna-related transfer functions [12]), or estimated through DSP techniques from the decomposition of recorded HRTFs. We refer to its inverse Fourier transform as *partial head-related impulse response* ($pHRIR$).

Def. 2 A *synthesized partial head-related transfer function*, $\widehat{pHRTF}$, contains modeled acoustic information related to

978-1-4673-5807-1/13/$31.00 ©2013 IEEE

specific body parts, or computationally generated through acoustic simulations. We refer to its inverse Fourier transform as *synthesized partial head-related impulse response* ($p\widetilde{HRIR}$).

The presented approach aims at building a completely customizable structural model through subsequent refinements, ranging from a selection of recorded $pHRTF$s to a totally synthetic filter model. Intermediate steps include mixtures of selected $pHRTF$s and synthetic components.

Let $HRTF_i$ be the individual HRTF set of a subject i. The *mixed structural modeling* (MSM) approach proposed here provides a possible approximation $\widetilde{HRTF}_i$:

$$HRTF_i \overset{MSM}{\leftrightarrow} \widetilde{HRTF}_i. \tag{1}$$

Such approximation is constructed by connecting N components, i.e. N pHRTFs related to different body parts. Typically in structural models N is equal to 3 (head, torso, and pinna components), but it depends on whether some of these components are merged (e.g. in a complete HRTF, $N = 1$), further decomposed (e.g. concha and helix are modeled separately) or supported by additional components (e.g. the ear canal contribution or headphones responses, which are also strictly related to anthropometry).

Each component can be chosen within three different sets:

1) individual components ($pHRTF$s of subject i);
2) selected components ($pHRTF$s of different subjects);
3) modeled components (synthesized $p\widetilde{HRTF}$s).

The approximation $\widetilde{HRTF}_i$ will include S selected components, I individual components, and M model components:

$$\widetilde{HRTF}_i = \bigoplus_{k=1}^{S} pHRTF_{s_k^*} \circledast \bigoplus_{k=1}^{I} pHRTF_{i_k} \circledast \bigoplus_{k=1}^{M} p\widetilde{HRTF}_{m_k} \tag{2}$$

where

$$i, s \in \mathcal{S}, \qquad m \in \mathcal{M}$$
$$I + S + M = N$$

The sets $\mathcal{S}$ and $\mathcal{M}$ represent the collections of subjects and models of which at least one $pHRTF$ or one $p\widetilde{HRTF}$ is available; s_k and i_k denote the k^{th} partial component for a subject s and for the target subject i, respectively; m_k denotes the k^{th} modeled component. The $\circledast$ operator relates to the filter representation, and denotes for each of its instances series or parallel filter connections.

Selected components in Eq. (2) are in general a subset of N components chosen based on the following optimization criterion:

$$\{s_k^*\} = \left\{ s \in \mathcal{S} - \{i\}, k = 1, ..., N \middle|\ s_k \text{ minimizes } e_k^{\mathbf{S}} \right\}. \tag{3}$$

Here $\mathbf{S}$ represents a given selection technique,[1] and $e_k^{\mathbf{S}}$ is the associated selection error for the k^{th} component.

[1]Here we also consider techniques based on series expansions with self-tuning of weights and perceptually-driven HRTF selections as candidate selection techniques, even if our focus lies on HRTF selection with respect to anthropometric features.

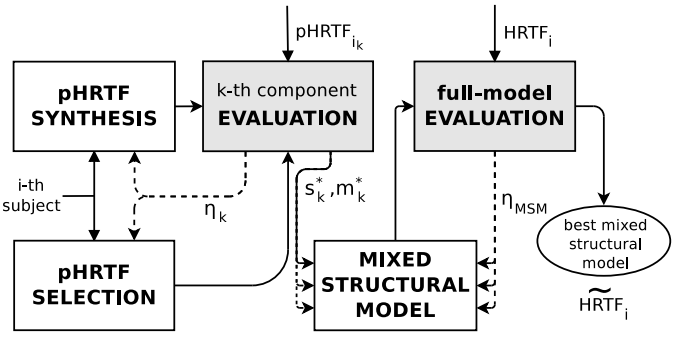


Fig. 1. Typical research workflow towards a mixed structural model.

As a particular case, $S = M = 0$ and $I = N$ yields:

$$\widetilde{HRTF}_i = HRTF_i = \bigoplus_{k=1}^{I} pHRTF_{i_k}. \tag{4}$$

Different combinations of S, I, M in our formalism include other relevant cases already proposed in previous literature:

- $S = N = 1$, $I = M = 0$ using a generic subject s: common use of non-individualized $HRTF$s (e.g., only mannequin HRTFs available).
- $S = N = 1$, $I = M = 0$ using one subject s^* that minimizes a given selection error: HRTF selection [10].
- $M = N = 1$, $I = S = 0$ using a model m^* that minimizes a given modeling error: direct HRTF modeling without structural decomposition [2].
- $M = N = 3$, $I = S = 0$ using customized models m_k for each component: structural HRTF modeling [4].

The goal of the MSM approach is twofold:

1) progressively remove all the individual partial components, i.e. $I = 0$, $S + M = N$;
2) provide reliable techniques to pHRTF modeling and pHRTF selection, and to evaluate their combinations [13] towards a complete structural model.

Ultimately, the optimal MSM solution corresponds to the case $M = N$, $I = S = 0$:

$$\widetilde{HRTF}_i = \bigoplus_{k=1}^{M} p\widetilde{HRTF}_{m_k^*}. \tag{5}$$

The process towards this case considers a wide group of candidate MSMs each described by a set of parameters. Fig. 1 depicts the workflow that leads to the construction of a specific MSM in the space of all possible model instances. Given the collections $\mathcal{S}, \mathcal{M}$, and given a test set of subjects with known HRTFs, the evaluation procedure in Fig. 1 provides the "best MSM", i.e. the best combination of modeled and selected components, including the relative balance between S and M.

A two-stage evaluation procedure, composed by a *single-component* and a *full-model* evaluation, guides the exclusion of certain instances and combinations of single components. The two fundamental evaluation parameters we consider in the first stage are:

- *accuracy* $\alpha_k \in [0,1]$, defined as the correlation between localization performances of the single $pHRTF_{s_k^*}$ or $p\widehat{HRTF}_{m_k}$ and $pHRTF_{i_k}$;
- *handiness* $\lambda_k \in [0,1]$, which measures the ease in feeding the single model or selection procedure with individual parameters.[2]

For simplicity, accuracy may be measured on a dimensionally reduced localization space (e.g., for the pinna the error may be measured only on the median plane). These two parameters ultimately define the *efficiency* $\eta_k = \alpha_k \lambda_k$ of the considered m_k, that we aim to maximize:

$$\{m_k^*\} = \{m \in \mathcal{M}, k = 1, ..., N |\ m_k \text{ maximizes } \eta_k\}. \quad (6)$$

The candidate m_k^* is then compared to the candidate s_k^*. If s_k^* provides an efficiency greater than η_k for m_k^*, it will be chosen as the k^{th} component, otherwise m_k^* will be chosen.

Subsequently, the full-model evaluation takes the best representative solutions of each k^{th} structural component in order to test the combined effects and the orthogonality of the models within full-space 3D virtual scenes. The same two evaluation criteria of the single-component evaluation procedure are used here, where α_{MSM} is the correlation between global localization performances of the resulting $\widehat{HRTF}_i$ and $HRTF_i$, while

$$\lambda_{MSM} = \prod_{k=1}^{N} \lambda_k. \quad (7)$$

The minimization of $\eta_{MSM} = \alpha_{MSM} \lambda_{MSM}$ leads the mixing process over subsequent versions of the MSM.

III. Partial Head-Related Transfer Functions

The key factor which allows the design of MSMs is that spatial cues for sound localization can be categorized according to the structural component that produces them. As a matter of fact, each polar coordinate (azimuth θ, elevation ϕ, and distance r) has one or more dominant cues associated to a specific body component in a given frequency range depending on its dimensions. In particular,

- azimuth and distance cues at all frequencies are associated to the head;
- elevation cues at high frequencies are associated to the pinnae;
- elevation cues at low frequencies are associated to the torso and shoulders.

In this section we will exhaustively describe how the three main components building a pHRTF behave and how such behaviour is approximated both in the literature and for our own MSMs.

[2]For instance, an acoustically measured individual HRTF implies $\lambda_k = 0$, while the use of a generic HRTF of a different subject has $\lambda_k = 1$ because no individualization is needed. All of the possible customization techniques ranging from the use of MRI scanning to the measurement of simple scalar anthropometric quantities have $\lambda_k = (0,1)$ in decreasing order of customization burden.

A. The head

1) Azimuth and distance cues: Azimuth cues can be reduced to two basic quantities thanks to the active role of the head in the differentiation of incoming sound waves, i.e.

- the *Interaural Time Difference* (ITD), defined as the temporal delay between sound waves at the two ears;
- the *Interaural Level Difference* (ILD), defined as the ratio between the instantaneous amplitudes of the same two sounds.

ITD is known to be frequency-independent below 500 Hz and above 3 kHz, with a theoretical ratio of low-frequency ITD versus high-frequency ITD of $3/2$, and slightly variable at middle range frequencies [14]. Conversely, frequency-dependent shadowing and diffraction effects introduced by the human head cause ILD to greatly depend on frequency.

Interaural cues are distance-independent when the source is in the so-called *far field* (approximately more than 1.5 m from the center of the head) where sound waves reaching the listener can be assumed to be plane. For such ranges, distance dependence can be approximated by a simple inverse square law. On the other hand, when the source is in the *near field* interaural cues exhibit a clear dependence on distance. By gradually approaching the sound source to the listener's head in the near field, it was observed that low-frequency gain is emphasized; ITD slightly increases; and ILD dramatically increases across the whole spectrum for lateral sources [15].

2) The spherical head model: The most recurring head model in the literature is the rigid sphere. The response related to a fixed observation point on the sphere's surface can be described by means of the following transfer function [16], based on Lord Rayleigh's diffraction formula [17]:[3]

$$H(\rho, \mu, \theta_{inc}) = -\frac{\rho}{\mu} e^{-i\mu\rho} \sum_{m=0}^{\infty} (2m+1) P_m(cos\theta_{inc}) \frac{h_m(\mu\rho)}{h'_m(\mu)}, \quad (8)$$

where a is the sphere radius, $\rho = r/a$ is normalized distance, θ_{inc} is the incidence angle (i.e. the angle between rays connecting the center of the sphere to the source and the observation point), and μ is normalized frequency, defined as

$$\mu = f \frac{2\pi a}{c}, \quad (9)$$

where c is the speed of sound.

A first-order approximation of the transfer function produced by Eq. 8 for $r \to \infty$ was proposed by Brown and Duda [4] as a minimum-phase analog filter. Near-field distance dependence can instead be accounted for through the filter structure H_{dist} we propose in [18], where spatial parameters ρ and θ_{inc} define all the inputs to its single components. Fig. 2 reports the whole spherical filter model structure, including a gain factor and parameters of a first-order shelving filter H_{sh} defined by tabulated coefficients of second-order rational

[3]Here P_m and h_m represent, respectively, the *Legendre polynomial* of degree *m* and the *m*th-order *spherical Hankel function*. h'_m is the derivative of h_m with respect to its argument.

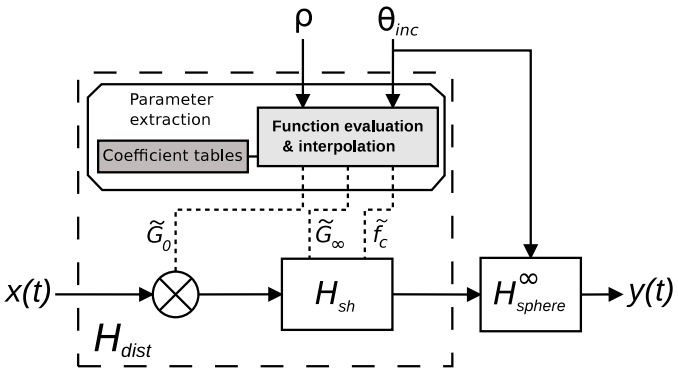


Fig. 2. A spherical head model including distance dependence in the near field.

functions, together with the digital counterpart of the Brown-Duda filter H^{∞}_{sphere}.

Typically, in spherical models the two observations points (i.e. the ear canals) are assumed to be diametrically opposed. As an alternative model, the spherical-head-with-offset-ears model described in [19] was obtained by displacing the ears backwards and downwards by a certain offset, introducing a nonlinear mapping between θ_{inc} and θ in the horizontal plane and elevation dependency on a cone of confusion. [4] Such model was found to provide a good approximation to elevation-dependent patterns both in the frequency and time domains, particularly replicating a peculiar X-shaped pattern along elevation (due to the superposition of two different propagation paths around the head) commonly seen in measured contralateral HRIRs.

Note that Eq. 8 is a function of head radius, a. This is a critical parameter: as an example, a sphere having the same volume of the head approximates its behaviour much better than a sphere with diameter equal to the interaural distance [20]. Hence, in order to fit the spherical head filter model to a specific listener, parametrization of a on the subject's anthropometry shall be performed. In [21] the ITD produced by spheres with different radii is compared to a number of real ITD measurements for a specific subject, and the best head radius for that subject is defined as the value that corresponds to the minimum mean least squares distance between the two estimates for different azimuth angles on the horizontal plane. A linear model for estimating the head radius given the three most relevant anthropometric parameters for the head (width w_h, height h_h, and depth d_h), is fitted to ITD-optimized radii of 45 different subjects through linear regression, yielding the optimal solution

$$a_{opt} = 0.26w_h + 0.01h_h + 0.09d_h + 3.2 \quad [cm]. \qquad (10)$$

This result highlights how head height is a relatively weak parameter in ITD definition with respect to head width and depth.

[4]The *cone of confusion* is defined as the set of spatial points producing the same ITD and ILD values for a spherical head.

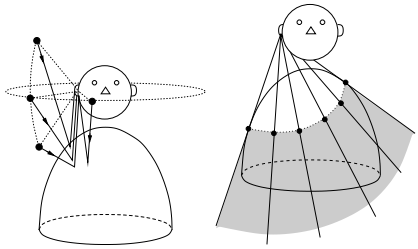


Fig. 3. Torso effects: shoulder reflections (a) and shadowing (b).

3) Alternative head models: The spherical model of the head provides an excellent approximation to the magnitude of a measured HRTF [22]. Still, it is far less accurate in predicting ITD, being the latter actually not constant around a cone of confusion, but variable by as much as 18% of the maximum interaural delay [23]. ITD estimation accuracy can be improved by considering an ellipsoidal head model that can account for the ITD variation and be adapted to individual listeners. A drawback of this formulation is that the analytical solution for the ITD is far complicated, and no explicit model for the ellipsoid-related transfer function was proposed.

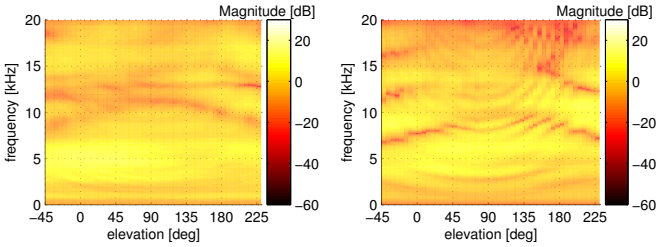
Conversely, models for the head as a prolate spheroid were studied in [24], [25] as the sole alternative analytical model to a sphere. Although adding nothing new in the ITD's point of view, comparison of spheroidal HRTFs against spherical HRTFs revealed a different behaviour in head-induced low-frequency ripples in the magnitude response at the contralateral ear, which is closer to responses of a KEMAR head [26] for the spheroidal case [27]. Still, this model has been very little studied, and consistent advantages over the spherical model have not been made clear.

B. Torso and shoulders

1) Low-frequency elevation cues: The effects of the torso and shoulders on the HRTF are relatively weak if compared to those due to the head and pinnae, and experiments to establish the perceptual importance of the relative cues have produced mixed results in general [4], [28], [19]. The torso introduces a shadowing effect for sound waves coming from below. Complementarily, shoulders disturb incident sound waves coming from all directions other than below at low frequencies. In particular, they provide a major reflection whose delay is proportional to the distance from the ear to the shoulder when the sound source is directly above the listener [29]. Fig. 3 schematically sketches the two torso effects.

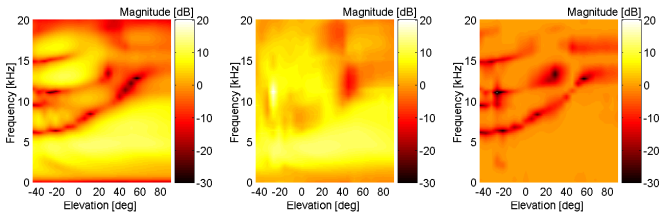
The shoulder reflection translates into a series of comb-filter notches in the frequency domain [30]. Nevertheless, the relative strength of this reflection with respect to the direct path of the sound wave seems to depend on both the subject's clothing and his/her upper torso dimensions. For instance, as Fig. 4 demonstrates, lateral HRTFs of the two CIPIC database [5] subjects having the smallest (a) and largest (b)

[5]http://interface.cipic.ucdavis.edu/



(a) (b)

Fig. 4. Left HRTFs of CIPIC subjects 018 (a) and 050 (b) for $\theta = -65°$.



(a) (b) (c)

Fig. 5. Separation of CIPIC Subject 165's left PRTFs (a) into a resonant (b) and a reflective (c) component.

shoulder width exhibit different behaviours of the shoulder reflection (represented as peculiar arch-shaped patterns along elevation).

Torso and shoulders are also commonly seen to perturb low-frequency ITD, even if it is questionable whether they may help in resolving localization ambiguities on a cone of confusion [29]. However, as Algazi *et al.* remarked in [19], when a signal is low-passed below 3 kHz elevation judgement is very poor in the sagittal plane if compared to a broadband source, but proportionally improves as the source is progressively moved away from the median plane, where performance is more accurate in the back than in the front. This result suggests the existence of low-frequency cues for elevation that, although being overall weak, are significant away from the median plane.

2) The snowman model: Similary to the head, in previous works the torso has been approximated by a sphere. Coaxial superposition of the two spheres of radius a and b, respectively, separated by a distance h that accounts for the neck, gives birth to the *snowman model* [30]. The far-field behaviour of the snowman model was studied in the frontal plane both by direct measurements on two rigid spheres and by computation through multipole reexpansion [31]. A structural head-and-torso model was also derived from the snowman model [30]; its structure distinguishes the two cases where the torso acts as a reflector or as a shadower, switching between the two filter sub-structures as soon as the source enters or leaves the torso shadow zone, respectively.

Additionally to the spherical model, an ellipsoidal model for the torso was studied in combination with the usual spherical head. This was done either by ray-tracing analysis [19] or through the BEM [31]. Such model is able to account for different torso reflection patterns; listening tests confirmed that this approximation and the corresponding measured HRTF gave similar results, showing larger correlations away from the median plane. Also, the ellipsoidal torso can be easily customized for a specific subject by directly defining control points for its three axes on the subject's torso [31].

C. The pinna

1) High-frequency elevation cues: Even though the torso provides weak elevation cues at low frequencies, vertical localization ability is mainly due to the presence of the pinnae [32]. The external ear plays an important role by introducing peaks and notches in the high-frequency spectrum of the HRTF, whose center frequency, amplitude, and bandwidth greatly depend on the elevation angle of the sound source [33], to a remarkably minor extent on azimuth [34], and are almost independent on distance between source and listener beyond a few centimeters from the ear [15]. It is generally considered that a sound source has to contain substantial energy in the high-frequency range for accurate judgement of elevation, because wavelengths longer than the size of the pinna are not affected: one could roughly state that the pinnae have a relatively little effect below 3 kHz.

The pinna can be seen both as a filter in the frequency domain [35] and a delay-and-add reflection system in the time domain [36] as long as typical pinna reflection delays for elevation angles are seen to produce spectral notches in the high-frequency range. Additionally to reflections, pinna resonances and diffraction inside the concha were also seen to contribute to HRTF spectral shaping. Shaw [37] identified six resonant modes of the pinna excited at different directions which clearly produce the most prominent HRTF spectral peaks, while Lopez-Poveda and Meddis [34] motivated the slight dependence of spectral notches on azimuth through a diffraction process that scatters the sound within the concha cavity, allowing reflections on the posterior wall of the concha to occur for any direction of the sound.

2) PRTF separation: In general, both pinna peaks and notches seem to play an important function in vertical localization of a sound source. However, a previous work of ours [38] highlighted that while the resonant component of the pinna-related counterpart of the HRTF (known as PRTF) exhibits a similar behaviour among different subjects, the reflective component of the PRTF comes along critically subject-dependent. This result was achieved by separating the resonant and reflective components through an ad-hoc designed algorithm [39], an instance of which can be appreciated in Fig. 5.[6]

Such an algorithm is essential to study these two contributions separately. An analysis-by-synthesis approach drives the algorithm towards the iterative compensation of the PRTF magnitude spectrum through a sequence of synthetic

[6]In these and in all of the following plots, magnitude values are linearly interpolated across the available azimuth/elevation angles to yield a 1-degree resolution.

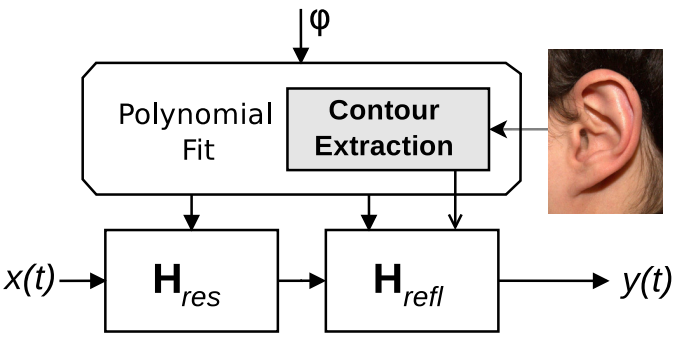


Fig. 6. The anthropometry-based structural PRTF model.

multi-notch filters until no local notches above a given amplitude threshold are left. Each multi-notch filter is fitted to the shape of the PRTF spectrum at the current iteration with its spectral envelope removed and subtracted to it, giving the spectrum for the next iteration. Eventually, when convergence is reached the spectrum contains the resonant component, while the reflective component is given by direct combination of all the calculated multi-notch filters.

3) Anthropometry-based pinna model: Different physical and structural models of the pinna have been proposed in the past, an exhaustive review of which can be found in [40]. Restricting our attention to points near the median plane, we propose a pinna filter realization that acts as a synthetic PRTF (schematically reported in Fig. 6), consisting of two second-order peak filters (filter structure H_{res}) and three second-order notch filters (filter structure H_{refl}) synthesizing two resonance modes and three pinna reflections respectively. The associated parameters (peak/notch central frequency, bandwidth, and gain) are computed by evaluating a number of elevation-dependent polynomial functions constructed from single or average PRTF measurements or derived from the subject's anthropometry [41].

As a matter of fact, in [40] we exploited a simple ray-tracing law to show that in median-plane frontal HRTFs the frequency of the spectral notches, each assumed to be caused by its own reflection path, is related to the shape of the concha, helix, and antihelix. This result allows direct parametrization of the reflective component of the pinna model onto the subject's anthropometry presented under the form of one or more side-view pictures of his/her head. Spectral distortion between real and synthesized PRTFs indicated that the approximation provided by the pinna model is objectively satisfactory.

IV. MSM Examples

In this last section we provide two basic examples of our mixed structural modeling approach. In the first one, frontal horizontal-plane HRTFs of a pinnaless KEMAR mannequin are approximated by the combination of a spherical head model parameterized on the mannequin's head dimensions and the selected torso response from the nearest subject in the CIPIC HRTF database with respect to the shoulder

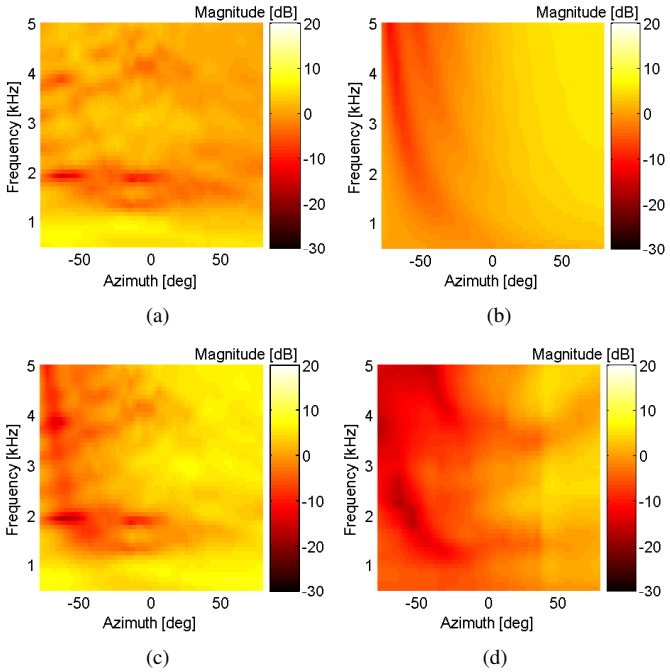


Fig. 7. Horizontal-plane right-ear HRTFs ($\theta = [-80°, 80°]$). (a) Extrapolated torso of CIPIC Subject 156. (b) Rigid spherical head. (c) Combination of (a) and (b). (d) Pinnaless KEMAR mannequin.

width parameter. In the second example, frontal median-plane HRTFs of a full KEMAR mannequin are derived from the application of our pinna model to the correlated recorded pinnaless responses.

A. Example #1

Right HRTF magnitudes of a pinnaless KEMAR mannequin in the horizontal plane up to 5 kHz are plotted in Fig. 7(d) for $\theta = [-80°, 80°]$, where $\theta > 0$ corresponds to the right hemisphere, hence the ipsilateral side. One can easily detect the different behaviour of the pinnaless mannequin in this zone, where shoulder reflections add up to the direct path of the sound wave, and in the contralateral side, where shadowing and diffraction by the head significantly attenuates any incoming sound.

In order to approximate such behaviour, the contributions of the head and shoulders to the pinnaless response are treated separately and then combined. Concerning the head, the spherical model with custom radius is the most straightforward choice. The optimal radius a^* for the KEMAR head is calculated as in Eq. 10, yielding $a^* = 8.86$ cm. A set of HRTFs from a spherical head are then derived from Eq. 8 by setting $\rho = 1$ m. These responses are reported in Fig. 7(b), where we can detect the substantial direct-path gain in the ipsilateral side and the effects of shadowing and diffraction in the contralateral side. The latter effect is however much shallower than in the pinnaless KEMAR responses and could be attributed to the intrinsic differences between an ideal sphere and a mannequin head, even though their gross behaviour is overall similar.

The shoulder's contribution is instead extrapolated from the

HRTFs of the CIPIC database subject (KEMAR excluded) whose shoulder width is the closest to the KEMAR's, i.e. Subject 156. Even though the pinna modifies shoulder reflections, its contribution to the low-frequency range is negligible. For this reason, the torso response - i.e. the shoulder reflection - is isolated by simply subtracting a windowed version of the HRIR (1-ms Hann window) to the full HRIR. The magnitude plot in Fig. 7(a) shows a main reflection between 1 and 2 kHz followed by fainter comb-like harmonics in the contralateral side.

In this first MSM instance $N = 2$, and in particular $M = 1$, $S = 1$, and $I = 0$. The two separate contributions are simply combined by convolving the related HRIRs. The result, reported in Fig. 7(c), reveals that the head contribution in the contralateral side fails to overshadow the weak shoulder reflection as it happened in Fig. 7(d). The torso contribution is of course different; this is the price to pay when a non-individual response is used. However, the approximated response succeeds in replicating the lowest frequency notch and the gross behavior of the head. Of course, only psychoacoustic tests can evaluate the accuracy of the approximated pinnaless KEMAR responses, subject, however, to the high handiness of both contributions (only 4 anthropometric scalar quantities are needed overall).

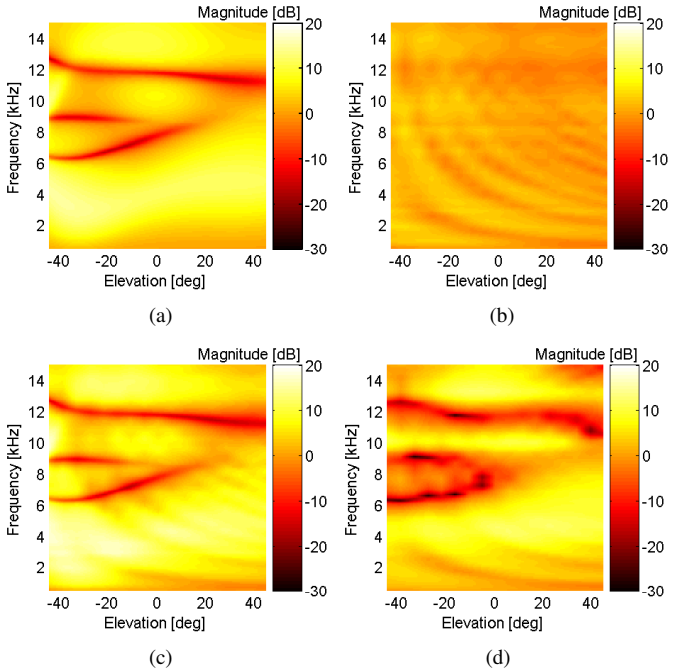


Fig. 8. Median-plane right-ear HRTFs ($\phi = [-45°, 45°]$). (a) Pinna model of CIPIC Subject 165 (KEMAR with small pinna). (b) Pinnaless KEMAR mannequin. (c) Combination of (a) and (b). (d) Subject 165, full response.

B. Example #2

The pinnaless KEMAR responses used for comparison in the previous example are now used as a structural component of a more complete model including the pinna of the subject. In this case we aim at recreating the full-body HRTFs of a KEMAR mannequin with small pinnae (i.e. Subject 165 of the CIPIC database) in the frontal side of the median plane, the region where the effect of the pinna and its subject intervariability is most prominent [42].

Median-plane HRTF magnitudes for $\phi = [-45°, 45°]$ of the pinnaless- and full-KEMAR mannequin are reported in Fig. 8(b) and Fig. 8(d) respectively. A quick comparison of these two plots reveals the massive effect of the pinna in the median plane, that literally overshadows the contributions of the head and torso with its three main notch ridges (beginning approximately at 6.5, 9, and 12.5 kHz) and its resonance patterns, the most prominent of which falls around 4.5 kHz at all elevations. The pinna contribution is provided by the filter model introduced in Section III-C3, with parameters of the characteristic peak and notch filters derived from an analysis of Subject 165's PRTFs (hence not taken from its anthropometry). Transfer functions of this model, reported in Fig. 8(a), accurately reproduce the peak/notch patterns of the original response.

In this second MSM instance $N = 2$, and in particular $M = 1$, $S = 0$, and $I = 1$. The pinnaless KEMAR HRIRs are fed to the pinna model yielding the approximated HRTF plot in Fig. 8(c). Thanks to the use of individual contributions - either in their original or modeled form - differences between the approximated and original HRTFs are visually forgettable. Of course, despite the allegedly high α_{MSM}, the use of individual contributions pushes λ_{MSM} to 0.

V. Developments and Future Perspectives

The approach presented in this paper answers both the requirements of structural modularity and integration of heterogeneous contributions. The mixed structural modeling formulation allows indeed an agile mixture of acoustic responses and synthetic models with the appealing aim of incorporating such diversity. The well-defined characterization facilitates the design of novel synthetic filter models and pHRTF selection processes possibly nourished by computer-simulated pHRTFs.

In order to improve the localization accuracy provided by the model in a full 3-D space, the degree of orthogonality among structural components has to be tested. This implies an extension of our pHRTF models outside their dominant spatial dimension. Among the possible options are an extension of the pinna model outside the median plane; the inclusion of elevation-dependent patterns in non-spherical head responses; and a study of the behaviour of the torso in the near field.

Localization error minimization can be also achieved by increasing the number of structural components. As an example, the ear canal contributes to the approximation of the correct pressure at the eardrum both in free-field and headphone listening conditions. On the other hand, the gradual increase of mixed structural model instances requires reliable and complex auditory models so as to facilitate the systematic exclusion of weak pHRTF models or selections in favour of the best instances.

ACKNOWLEDGMENT

The authors wish to thank Prof. Ralph Algazi for the pinnaless KEMAR responses and Fabrizio Granza for his help with the figures and the torso extrapolation procedure.

REFERENCES

[1] C. I. Cheng and G. H. Wakefield, "Introduction to head-related transfer functions (HRTFs): Representations of HRTFs in time, frequency, and space," *J. Audio Eng. Soc.*, vol. 49, no. 4, pp. 231–249, April 2001.

[2] E. C. Durant and G. H. Wakefield, "Efficient model fitting using a genetic algorithm: Pole-zero approximations of HRTFs," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 1, pp. 18–27, January 2002.

[3] D. J. Kistler and F. L. Wightman, "A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction," *J. Acoust. Soc. Am.*, vol. 91, no. 3, pp. 1637–1647, March 1992.

[4] C. P. Brown and R. O. Duda, "A structural model for binaural sound synthesis," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 5, pp. 476–488, September 1998.

[5] V. R. Algazi, R. O. Duda, R. P. Morrison, and D. M. Thompson, "Structural composition and decomposition of HRTFs," in *Proc. IEEE Work. Appl. Signal Process., Audio, Acoust.*, New Paltz, New York, USA, October 2001, pp. 103–106.

[6] S. Hwang, Y. Park, and Y. Park, "Modeling and customization of head-related impulse responses based on general basis functions in time domain," *Acta Acustica united with Acustica*, vol. 94, no. 6, pp. 965–980, November 2008.

[7] K. H. Shin and Y. Park, "Enhanced vertical perception through head-related impulse response customization based on pinna response tuning in the median plane," *IEICE Trans. Fundamentals*, vol. E91-A, no. 1, pp. 345–356, January 2008.

[8] B. U. Seeber and H. Fastl, "Subjective selection of non-individual head-related transfer functions," in *Proc. 2003 Int. Conf. Auditory Display (ICAD03)*, Boston, MA, USA, July 2003, pp. 259–262.

[9] R. H. Y. So, B. Ngan, A. Horner, J. Braasch, J. Blauert, and K. L. Leung, "Toward orthogonal non-individualised head-related transfer functions for forward and backward directional sound: Cluster analysis and an experimental study," *Ergonomics*, vol. 53, no. 6, pp. 767–781, June 2010.

[10] B. F. G. Katz and G. Parseihian, "Perceptually based head-related transfer function database optimization," *J. Acoust. Soc. Am.*, vol. 131, no. 2, pp. EL99–EL105, February 2012.

[11] V. C. Raykar, R. Duraiswami, and B. Yegnanarayana, "Extracting the frequencies of the pinna spectral notches in measured head related impulse responses," *J. Acoust. Soc. Am.*, vol. 118, no. 1, pp. 364–374, July 2005.

[12] P. Satarzadeh, R. V. Algazi, and R. O. Duda, "Physical and filter pinna models based on anthropometry," in *Proc. 122nd Conv. Audio Eng. Soc.*, Vienna, Austria, May 2007, pp. 718–737.

[13] M. Geronazzo, S. Spagnol, and F. Avanzini, "A modular framework for the analysis and synthesis of head-related transfer functions," in *Proc. 134th Conv. Audio Eng. Soc.*, Rome, Italy, May 2013.

[14] G. F. Kuhn, "Model for the interaural time differences in the azimuthal plane," *J. Acoust. Soc. Am.*, vol. 62, no. 1, pp. 157–167, July 1977.

[15] D. S. Brungart and W. M. Rabinowitz, "Auditory localization of nearby sources. Head-related transfer functions," *J. Acoust. Soc. Am.*, vol. 106, no. 3, pp. 1465–1479, September 1999.

[16] W. M. Rabinowitz, J. Maxwell, Y. Shao, and M. Wei, "Sound localization cues for a magnified head: Implications from sound diffraction about a rigid sphere," *Presence*, vol. 2, no. 2, pp. 125–129, Spring 1993.

[17] J. W. Strutt, "On the acoustic shadow of a sphere," *Phil. Trans.*, vol. 203, pp. 87–110, 1904.

[18] S. Spagnol, M. Geronazzo, and F. Avanzini, "Hearing distance: A low-cost model for near-field binaural effects," in *Proc. EUSIPCO 2012 Conf.*, Bucharest, Romania, September 2012, pp. 2005–2009.

[19] V. R. Algazi, C. Avendano, and R. O. Duda, "Elevation localization and head-related transfer function analysis at low frequencies," *J. Acoust. Soc. Am.*, vol. 109, no. 3, pp. 1110–1122, March 2001.

[20] B. F. G. Katz, "Boundary element method calculation of individual head-related transfer function. I. Rigid model calculation," *J. Acoust. Soc. Am.*, vol. 110, no. 5, pp. 2440–2448, November 2001.

[21] V. R. Algazi, C. Avendano, and R. O. Duda, "Estimation of a spherical-head model from anthropometry," *J. Audio Eng. Soc.*, vol. 49, no. 6, pp. 472–479, June 2001.

[22] P. Mokhtari, H. Takemoto, R. Nishimura, and H. Kato, "Acoustic simulation of KEMAR's HRTFs: Verification with measurements and the effects of modifying head shape and pinna concavity," in *Proc. Int. Work. Princ. Appl. Spatial Hearing (IWPASH 2009)*, Zao, Miyagi, Japan, November 2009.

[23] R. O. Duda, C. Avendano, and V. R. Algazi, "An adaptable ellipsoidal head model for the interaural time difference," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'99)*, Phoenix, AZ, USA, March 1999, pp. 965–968.

[24] R. W. Novy, "Characterizing elevation effects of a prolate spheroidal HRTF model," Master's thesis, San Jose State University, 1998.

[25] H. Jo, Y. Park, and Y. Park, "Approximation of head related transfer function using prolate spheroidal head model," in *Proc. 15th Int. Congr. Sound Vibr. (ICSV15)*, Daejeon, Korea, July 2008, pp. 2963–2970.

[26] M. D. Burkhard and R. M. Sachs, "Anthropometric manikin for acoustic research," *J. Acoust. Soc. Am.*, vol. 58, no. 1, pp. 214–222, July 1975.

[27] H. Jo, Y. Park, and Y. Park, "Optimization of spherical and spheroidal head model for head related transfer function customization: Magnitude comparison," in *Proc. Int. Conf. Control, Automat., Syst. (ICCAS 2008)*, Seoul, Korea, October 2008, pp. 251–254.

[28] F. Asano, Y. Suzuki, and T. Sone, "Role of spectral cues in median plane localization," *J. Acoust. Soc. Am.*, vol. 88, no. 1, pp. 159–168, July 1990.

[29] O. Kirkeby, E. T. Seppälä, A. Kärkkäinen, L. Kärkkäinen, and T. Huttunen, "Some effects of the torso on head-related transfer functions," in *Proc. 122nd Conv. Audio Eng. Soc.*, Vienna, Austria, May 2007, pp. 1045–1052. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=14015

[30] V. R. Algazi, R. O. Duda, and D. M. Thompson, "The use of head-and-torso models for improved spatial sound synthesis," in *Proc. 113th Conv. Audio Eng. Soc.*, Los Angeles, CA, USA, October 2002, pp. 1–18.

[31] V. R. Algazi, R. O. Duda, R. Duraiswami, N. A. Gumerov, and Z. Tang, "Approximating the head-related transfer function using simple geometric models of the head and torso," *J. Acoust. Soc. Am.*, vol. 112, no. 5, pp. 2053–2064, November 2002.

[32] M. B. Gardner and R. S. Gardner, "Problem of localization in the median plane: Effect of pinnae cavity occlusion," *J. Acoust. Soc. Am.*, vol. 53, no. 2, pp. 400–408, 1973.

[33] E. A. G. Shaw and R. Teranishi, "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, 1968.

[34] E. A. Lopez-Poveda and R. Meddis, "A physical model of sound diffraction and reflections in the human concha," *J. Acoust. Soc. Am.*, vol. 100, no. 5, pp. 3248–3259, November 1996.

[35] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*. Cambridge, MA, USA: MIT Press, 1983.

[36] D. W. Batteau, "The role of the pinna in human localization," *Proc. R. Soc. London. Series B, Biological Sciences*, vol. 168, no. 11, pp. 158–180, August 1967.

[37] E. A. G. Shaw, "Acoustical features of human ear," in *Binaural and Spatial Hearing in Real and Virtual Environments*. Mahwah, NJ, USA: R. H. Gilkey and T. R. Anderson, Lawrence Erlbaum Associates, 1997, pp. 25–47.

[38] S. Spagnol, M. Geronazzo, and F. Avanzini, "Fitting pinna-related transfer functions to anthropometry for binaural sound rendering," in *Proc. IEEE Int. Work. Multi. Signal Process. (MMSP'10)*, Saint-Malo, France, October 2010, pp. 194–199.

[39] M. Geronazzo, S. Spagnol, and F. Avanzini, "Estimation and modeling of pinna-related transfer functions," in *Proc. 13th Int. Conf. Digital Audio Effects (DAFx-10)*, Graz, Austria, September 2010, pp. 431–438.

[40] S. Spagnol, M. Geronazzo, and F. Avanzini, "On the relation between pinna reflection patterns and head-related transfer function features," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 3, pp. 508–520, March 2013.

[41] M. Geronazzo, S. Spagnol, and F. Avanzini, "A head-related transfer function model for real-time customized 3-D sound rendering," in *Proc. INTERPRET Work., SITIS 2011 Conf.*, Dijon, France, November-December 2011, pp. 174–179.

[42] S. Spagnol, M. Hiipakka, and V. Pulkki, "A single-azimuth pinna-related transfer function database," in *Proc. 14th Int. Conf. Digital Audio Effects (DAFx-11)*, Paris, France, September 2011, pp. 209–212.