Associazione Italiana di Acustica
41° Convegno Nazionale
Pisa, 17-19 giugno 2014

UN NUOVO APPROCCIO A MODELLI STRUTTURALI MISTI PER LA SINTESI E LA PERSONALIZZAZIONE DI HRTF

Michele Geronazzo (1), Simone Spagnol (1), Federico Avanzini (1)

1) Università di Padova, Dip. di Ing. dell'Informazione, <nome.cognome>@dei.unipd.it

1.Introduzione

Le caratteristiche spaziali del suono possono essere rese binauralmente attraverso la convoluzione di un segnale acustico di partenza con una coppia di "funzioni di trasferimento della testa", o Head-Related Transfer Functions (*HRTF*). Il segnale stereo che ne risulta viene riprodotto attraverso cuffie. Più precisamente, le *HRTF* sono le funzioni di trasferimento acustiche dalla sorgente sonora al timpano dell'ascoltatore, e dipendono quindi sia dalla frequenza che dalle coordinate spaziali della sorgente [1]. Misurare le *HRTF* di un singolo individuo è costoso in termini di tempo e risorse. In alternativa si usano spesso *HRTF* non individuali, misurate su appositi manichini con misure antropometriche medie, che hanno però lo svantaggio di una resa spaziale meno fedele.

Le tecniche proposte in letteratura per il design di *HRTF* sintetiche possono essere raggruppate in due famiglie: modelli poli-zeri [2], in cui le *HRTF* vengono approssimate da filtri razionali di basso ordine, e espansioni in serie [3], in cui si cercano rappresentazioni basate su combinazioni lineari di funzioni di base. Seguendo un approccio completamente diverso, i cosiddetti modelli strutturali [4] isolano e modellano tramite strutture filtranti gli effetti di ogni singola componente anatomica coinvolta – testa, padiglione auricolare (o *pinna*), canale uditivo, spalle/torso – e la *HRTF* completa viene costruita combinando tutti questi effetti.

Un recente tema di ricerca è quello della personalizzazione di *HRTF* per un singolo soggetto. Molti approcci si basano su espansioni in serie nelle quali ogni soggetto può auto-regolare i coefficienti della combinazione lineare [5], o semplicemente su selezione di *HRTF* non-individuali [6]. D'altro canto, l'approccio basato su modelli strutturali mostra grandi potenzialità: i parametri dei filtri associati a ciascuna componente anatomica potrebbero essere stimati da misure antropometriche dell'ascoltatore stesso [7].

In questo articolo presentiamo un nuovo approccio, che combina il paradigma dei modelli strutturali con altre tecniche di selezione di *HRTF*. Più precisamente, il nostro approccio basato su modellazione strutturale mista (o Mixed Structural Modeling, MSM), vede la *HRTF* complessiva come una combinazione di componenti strutturali

che possono essere sia sintetiche che misurate. In entrambi i casi queste possono essere personalizzate a partire da dati antropometrici individuali, i quali vengono usati o per determinare i valori dei parametri dei modelli, o per selezionare una specifica componente acustica all'interno di un database di risposte all'impulso.

2. Modelli strutturali misti per HRTF

Nel suo significato comunemente accettato, la definizione di "head-related transfer function" indica la funzione di trasferimento completa correlata a tutto il corpo, che comprende anche gli effetti acustici di parti del corpo diverse dalla testa. Sulla base di questa osservazione, vengono introdotte due ulteriori definizioni.

Def. 1 Una *head-related transfer function parziale* (*pHRTF*) contiene informazioni acustiche misurate isolando specifiche parti del corpo (es., relative al contributo del solo orecchio esterno) oppure stimate mediante tecniche DSP atte alla decomposizione di HRTF misurate [8].

Def. 2 Una *head-related transfer function sintetica e parziale* ($\widetilde{pHRTF}$) contiene le informazioni acustiche relative a specifiche parti del corpo siano esse modellate o computazionalmente generate attraverso simulazioni acustiche.

L'approccio presentato in questo articolo ha come obiettivo la costruzione di un modello strutturale completamente personalizzabile attraverso successivi raffinamenti, avendo come punto di partenza una selezione di *pHRTF* registrate e come punto di arrivo un modello di filtri totalmente sintetico. I passaggi intermedi comprendono alcune tra le possibili combinazioni di *pHRTF* selezionate e componenti sintetiche.

Sia $HRTF_i$ il set individuale di *HRTF* per un soggetto *i*. L'approccio di modellazione strutturale mista qui proposto fornisce una possibile approssimazione, $\widetilde{HRTF_i}$, tale che

$$HRTF_i \overset{MSM}{\leftrightarrow} \widetilde{HRTF_i}.$$

Tale approssimazione è costruita collegando *N* componenti, ovvero le *N pHRTF* relative a diverse parti del corpo. Nei modelli strutturali, *N* tipicamente è uguale a 3 (le componenti di testa, spalle/torso e orecchio), ma questo numero è relazionato a quali e quante di queste componenti vengano considerate indivisibili (ad esempio in una *HRTF* completa, $N = 1$), oppure ulteriormente separate (es. conca ed elice possono essere modellati separatamente) o estese supportando componenti aggiuntive (ad esempio, il contributo del canale uditivo o la risposta delle cuffie, elementi anch'essi strettamente correlati all'antropometria).

Ogni componente può essere scelta all'interno di tre differenti raggruppamenti:

1) componenti individuali (le *pHRTF* del soggetto *i*);
2) componenti selezionate (le *pHRTF* di soggetti diversi da *i*);
3) componenti modellate (le $\widetilde{pHRTF}$ sintetizzate).

L'approssimazione $\widetilde{HRTF_i}$ includerà *S* componenti selezionate, *I* componenti individuali e *M* componenti modellate:

$$\widetilde{HRTF_i} = \bigoplus_{k=1}^{S} pHRTF_{s_k^*} \circledast \bigoplus_{k=1}^{I} pHRTF_{i_k} \circledast \bigoplus_{k=1}^{M} \widetilde{pHRTF}_{m_k} \tag{1}$$

con $i, s \in \mathcal{S}$, $m \in \mathcal{M}$ e $I + S + M = N$, dove:

- $\mathcal{S}$ e $\mathcal{M}$ rappresentano le collezioni di soggetti e modelli per cui almeno una *pHRTF* o una $\widetilde{pHRTF}$ sia disponibile;
- $\circledast$ è l'operatore che si riferisce ad una rappresentazione a filtri, e può identificare una connessione in serie o in parallelo;

s_k e i_k indicano la k-esima componente parziale rispettivamente per un soggetto *s* e per il soggetto *i* presi in considerazione;
m_k è la k-esima componente modellata.

Le componenti selezionate in (1) sono in generale un sottoinsieme di *N* componenti scelte in base al seguente criterio di ottimizzazione:

$$\{s_k^*\} = \left\{ s \in \mathcal{S} - \{i\}, k = 1, ..., N \middle| \; s_k \text{ minimizes } e_k^{\mathbf{S}} \right\}. \tag{2}$$

dove:
$\mathcal{S}$ rappresenta una data tecnica di selezione;
$e_k^{\mathbf{S}}$ è l'errore di selezione associato alla componente k-esima.

Come caso particolare, per $S = M = 0$ e $I = N$ si ottiene la *HRTF* individuale misurata:

$$\widetilde{HRTF}_i = HRTF_i = \bigoplus_{k=1}^{I} pHRTF_{i_k}. \tag{3}$$

Adottando diverse combinazioni di *S*, *I* e *M*, il nostro formalismo può descrivere altre casistiche rilevanti e già proposte nella precedente letteratura scientifica:

- $S = N = 1$, $I = M = 0$ e utilizzo di un soggetto generico *s*: utilizzo indiscriminato di *HRTF* non individuali (ad esempio, quando sono disponibili solamente le *HRTF* di un manichino).
- $S = N = 1$, $I = M = 0$ e utilizzo di un soggetto s^* che minimizza un dato errore di selezione: selezione di *HRTF* [6].
- $M = N = 1$, $I = S = 0$ e utilizzo di un modello m^* che minimizza un dato errore di modellazione: modellazione di *HRTF* senza decomposizione strutturale [2].
- $M = N = 3$, $I = S = 0$ e utilizzo di modelli personalizzati m_k per ogni componente: modellazione strutturale di *HRTF* [4].

L'obiettivo dell'approccio MSM è duplice:

1. eliminare progressivamente tutte le componenti parziali individuali, cioè $I = 0$ e $S + M = N$;
2. fornire tecniche affidabili per la modellazione e selezione di *pHRTF*, e valutarne le combinazioni con l'obiettivo di ottenere un modello strutturale completo.

In definitiva, la soluzione ottimale corrisponde al caso $M = N$, $I = S = 0$:

$$\widetilde{HRTF}_i = \bigoplus_{k=1}^{M} p\widetilde{HRTF}_{m_k^*}. \tag{4}$$

Il processo di ricerca verso tale obiettivo considera un ampio gruppo di candidati tra gli MSM, ciascuno descritto da un insieme di parametri. La figura 1 schematizza il flusso di lavoro che determina lo sviluppo di uno specifico MSM all'interno dello spazio di tutte le possibili istanze del modello. Date le collezioni $\mathcal{S}$ e $\mathcal{M}$, assieme ad un insieme di soggetti di prova le cui *HRTF* siano note, la procedura di valutazione in figura fornisce il "migliore MSM", vale a dire la migliore combinazione tra componenti modellate e selezionate.

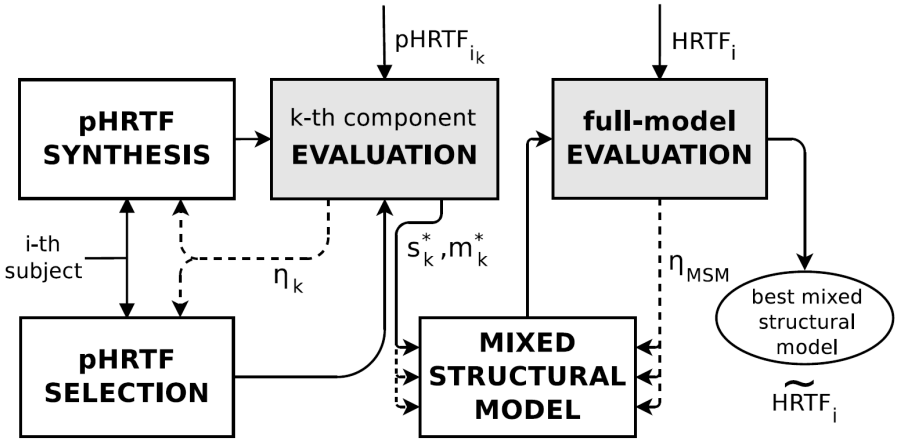

Figura 1 – Tipico flusso di lavoro verso la ricerca di un modello strutturale misto.

Una procedura di valutazione in due fasi, composta da una prima valutazione di ogni singola componente ed una del modello nella sua interezza, guida l'esclusione di alcune istanze e di alcune combinazioni di componenti. I due parametri di valutazione fondamentali che vengono considerati nella prima fase sono:

- *accuratezza* $\alpha_k \in [0, 1]$, definita come la correlazione tra le performance di localizzazione della singola *pHRTF* selezionata, modellata o individuale;
- *praticità* $\lambda_k \in [0, 1]$, che misura la facilità di gestione del modello o della procedura di selezione attraverso l'utilizzo di parametri individuali.

Per semplicità, l'accuratezza può essere misurata su una dimensionalità ridotta nello spazio di localizzazione (ad esempio, per l'orecchio l'errore di localizzazione può essere stimato solo sul piano mediano). Questi due parametri infine definiscono l'efficienza $\eta_k = \alpha_k \lambda_k$ di m_k preso in considerazione, da cui ci si propone di ottimizzare

$$\{m_k^*\} = \left\{ m \in \mathcal{M}, k = 1, ..., N \middle| \; m_k \text{ maximizes } \eta_k \right\}. \tag{5}$$

Il candidato m_k^* è quindi confrontato con il candidato s_k^*. Se s_k^* fornisce un'accuratezza maggiore verrà scelto come *k*-esima componente, altrimenti si sceglierà m_k^*. Successivamente, la valutazione completa del modello prende le migliori soluzioni per ogni *k*-esima componente strutturale in modo da testare gli effetti della combinazione e l'ortogonalità dei modelli all'interno di una scena virtuale tridimensionale che comprenda tutto lo spazio attorno all'ascoltatore. Vengono quindi utilizzati gli stessi criteri di valutazione impiegati nella procedura per la componente singola, determinando $\alpha_{MSM,}$ ossia la correlazione tra le performance di localizzazione globale delle risultanti $\widetilde{HRTF}_i$ e $HRTF_i$, e

$$\lambda_{MSM} = \prod_{k=1}^{N} \lambda_k. \tag{6}$$

La massimizzazione di $\eta_{MSM} = \alpha_{MSM} \lambda_{MSM}$ guida quindi il processo di combinazione per ogni successiva versione di MSM.

3. HRTF parziali ed esempi di MSM

Il fattore chiave che rende possibile la progettazione di MSM sta nel fatto che gli indicatori spaziali per la localizzazione del suono possano essere classificati a seconda della componente strutturale che li introduce. Ad ogni coordinata polare (azimut θ, elevazione φ e distanza r) corrispondono infatti uno o più indicatori dominanti associati ad una specifica parte del corpo e in un dato intervallo di frequenze. In particolare,

- gli indicatori di azimut e distanza sono associati alla testa a qualsiasi frequenza;
- gli indicatori di elevazione alle alte frequenze sono associati alla pinna;
- gli indicatori di elevazione alle basse frequenze sono associati a torso e spalle.

Di conseguenza, risulta naturale considerare i contributi di testa, pinna e torso come essenzialmente ortogonali, da trattare separatamente e combinare linearmente. Forniamo ora due semplici esempi di approccio strutturale misto nei quali tali componenti strutturali vengono considerate separatamente e selezionate da un database o modellate.

1. Esempio #1

Nella figura 2(d) sono riportate le risposte di ampiezza relative all'orecchio destro di un manichino KEMAR [9] senza padiglioni auricolari (*pinnaless*), nel piano orizzontale e fino a 5 kHz. Il range considerato per l'azimut è θ = [−80°, 80°], dove $\theta > 0$ corrisponde all'emisfero destro (e quindi la zona ipsilaterale). Nella figura si può facilmente riconoscere il diverso comportamento delle risposte in questa zona, nella quale le riflessioni dovute alle spalle si aggiungono al percorso diretto dell'onda sonora, rispetto alla zona controlaterale, nella quale i fenomeni di ombra acustica e diffrazione intorno alla testa attenuano in maniera significativa qualsiasi suono in ingresso.

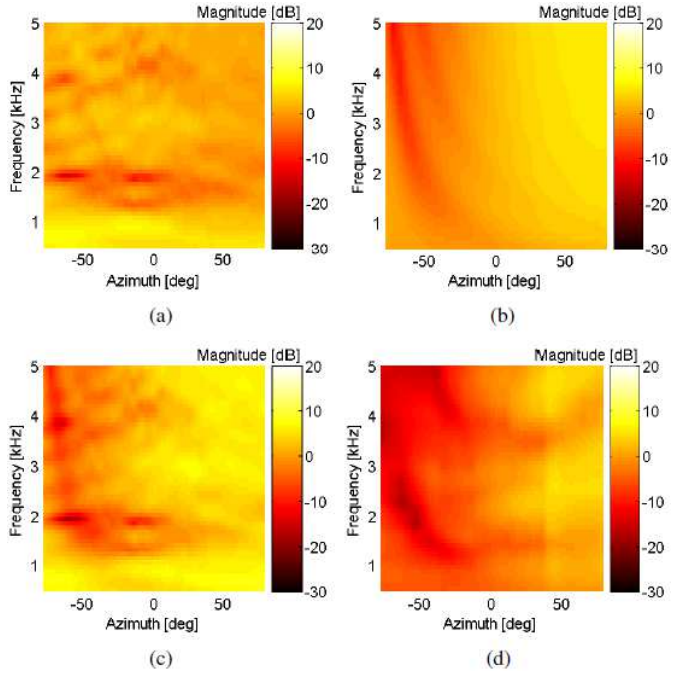

Figura 2 – HRTF per θ = [−80°, 80°], piano orizzontale, orecchio destro. (a) Risposta del torso del Soggetto CIPIC 156. (b) Testa sferica rigida. (c) Combinazione di (a) e (b). (d) Manichino KEMAR *pinnaless*.

Al fine di approssimare tale comportamento, i contributi di testa e spalle nella risposta *pinnaless* sono trattati separatamente e quindi combinati. Il modello sferico con rag-

gio personalizzato di Algazi *et al.* [10] rappresenta una scelta accreditata riguardo il contributo della testa. Dalle tre dimensioni della testa KEMAR (larghezza w_h, altezza h_h e profondità d_h) si può ottenere il raggio ottimo a_{opt} come

$$a_{opt} = 0.26w_h + 0.01h_h + 0.09d_h + 3.2 = 8.86. \quad \text{[cm]} \tag{7}$$

Il set di *HRTF* per la testa sferica viene quindi calcolato come in [11] fissando tale valore di a_{opt} oltre che $r = 1$ m. Le risposte risultanti sono riportate nella figura 2(b), dove possiamo identificare sia il sostanziale guadagno nella zona ipsilaterale che gli effetti di ombra acustica e diffrazione nella zona opposta. Quest'ultimo effetto è tuttavia molto meno marcato rispetto alle risposte misurate di riferimento, e potrebbe essere attribuito alle ovvie differenze tra una sfera ideale e la testa del manichino.

Il contributo di spalle e torso è invece estrapolato dalle *HRTF* del Soggetto 156 incluso nel database di *HRTF* CIPIC [12], il cui parametro "shoulder width" (larghezza da spalla a spalla) più si avvicina a quello del manichino KEMAR tra tutti i soggetti presenti. Nonostante la pinna modifichi anche le riflessioni dovute alle spalle, il contributo della stessa alle basse frequenze è trascurabile. Per tale ragione, la risposta del torso – ovvero le riflessioni sulle spalle – può essere isolata semplicemente compensando la risposta all'impulso completa con una versione finestrata della stessa (con finestra di Hann da 1 ms). La risultante risposta in ampiezza, tracciata nella figura 2(a), mostra una riflessione principale tra 1 e 2 kHz seguita da "armoniche" più deboli nella zona controlaterale.

In questa prima istanza di MSM presa in considerazione si ha $N = 2$, e in particolare $M = 1$, $S = 1$ e $I = 0$. I due contributi sono combinati semplicemente moltiplicando tra di loro le rispettive *HRTF*. Il risultato, riportato nella figura 2(c), rivela come il contributo della testa nella zona controlaterale non mascheri del tutto la debole riflessione della spalla come invece succedeva nella figura 2(d). Tuttavia, poiché viene utilizzata una risposta non individuale, il contributo del torso è ovviamente diverso tra una risposta e l'altra. Tuttavia, la risposta approssimata riesce a replicare fedelmente sia il notch più basso in frequenza che il comportamento complessivo della testa. Naturalmente, solo un test psicoacustico può certificare l'accuratezza di tale modello, il quale esibisce comunque un'elevata praticità: sono infatti sufficienti soltanto quattro quantità antropometriche scalari per personalizzarlo.

2. Esempio #2

Le risposte KEMAR *pinnaless* usate per il confronto nel precedente esempio vengono ora utilizzate come componente strutturale di un modello più completo comprendente il contributo acustico dell'orecchio. In tal caso l'obiettivo è ricreare le *HRTF* complete del manichino KEMAR con orecchie piccole (Soggetto 165 del database CIPIC) nella ragione frontale del piano mediano, ovverosia la regione nella quale l'effetto dell'orecchio esterno è maggiormente pronunciato [13].

Le risposte in ampiezza sul piano mediano per $\varphi = [-45°, 45°]$ dei soggetti KEMAR *pinnaless* e completo sono riportati rispettivamente nelle figure 3(b) e 3(d). Il confronto tra questi due grafici evidenzia l'effetto dell'orecchio esterno nel piano mediano, il quale maschera letteralmente i contributi di testa e torso con i suoi tre principali notch (situati approssimativamente a 6.5, 9 e 12.5 kHz a $\varphi = -45°$) e gli andamenti delle risonanze, la più importante delle quali cade attorno ai 4.5 kHz per tutte le elevazioni.

Il contributo della pinna è introdotto grazie a un apposito modello personalizzabile basato su due filtri *peak* e tre filtri *notch* [13]. In questo caso, i parametri dei filtri deri-

vano da un'analisi delle *HRTF* del Soggetto 165 [8]. Le funzioni di trasferimento del modello, riportate nella figura 3(a), riproducono accuratamente gli andamenti di picchi e notch della risposta originale.

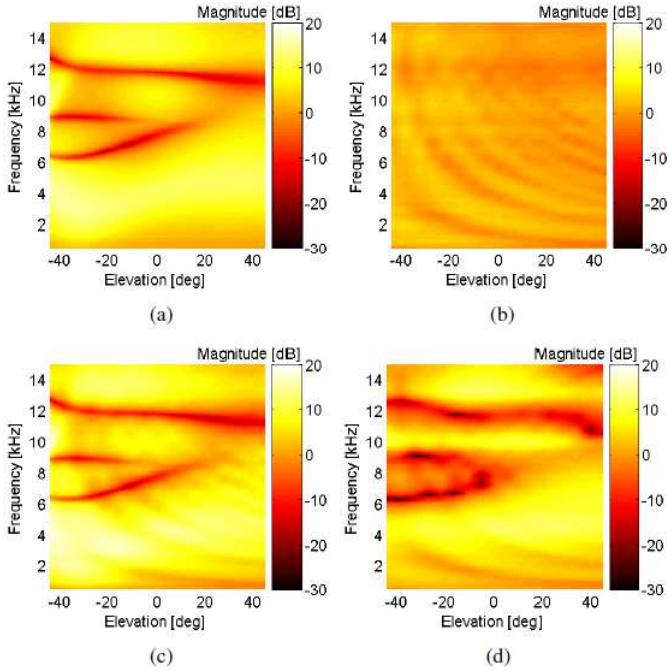

Figura 3 – HRTF per φ = [−45°, 45°], piano mediano, orecchio destro. (a) Modello di pinna del Soggetto CIPIC 165 (KEMAR con orecchio piccolo). (b) Manichino KEMAR *pinnaless*. (c) Combinazione di (a) e (b). (d) Risposta completa del Soggetto CIPIC 165.

In questa seconda istanza di MSM N = 2, e in particolare M = 1, S = 0 e I = 1. Le *HRTF* della KEMAR *pinnaless* vengono combinate con il modello di pinna; il risultato è il grafico riportato nella figura 3(c). Grazie all'utilizzo di contributi individuali – sia nella loro forma originale che in una forma approssimata – le differenze tra *HRTF* originali e approssimate sono visivamente minime. Ovviamente, nonostante il presunto elevato α_{MSM}, l'uso di contributi individuali fa sì che λ_{MSM} vada a zero.

4. Conclusioni

La formulazione MSM descritta in questo articolo permette di creare agilmente una combinazione di risposte acustiche e modelli sintetici, incorporando e sfruttando tale eterogeneità. La rigorosa formalizzazione dell'approccio MSM favorirà lo sviluppo di nuovi modelli sintetici di *pHRTF*, nonché di procedure di selezione di *pHRTF* ottenute tramite misure acustiche o tramite simulazioni numeriche.

Al fine di migliorare l'accuratezza dei modelli in termini di localizzazione in uno spazio 3D, sarà necessario valutare il grado di ortogonalità delle diverse componenti strutturali. Ciò implica una estensione dei modelli di *pHRTF* al di fuori della loro componente spaziale dominante. Esempi rilevanti in questo senso riguardano: (i) l'estensione del modello di pinna fuori dal piano mediano; (ii) l'inclusione di andamenti dipendenti dall'elevazione in risposte di teste non-sferiche e (iii) uno studio sul comportamento del torso nei limiti di campo vicino.

Miglioramenti in termini di localizzazione potrebbero essere ottenuti anche incrementando il numero di componenti strutturali. Ad esempio, il canale uditivo contribuisce ad approssimare la corretta pressione acustica al timpano, in condizioni di ascolto sia in campo libero che in cuffia. D'altro canto, l'aumento di istanze di modelli strutturali misti che richiedano di essere testate promuove l'uso di modelli uditivi complessi e affidabili, al fine di facilitare l'esclusione sistematica di modelli poco efficienti di *pHRTF* con la selezione delle istanze migliori.

5. Ringraziamenti

Questo lavoro è stato finanziato dal progetto di ricerca *PADVA* (Personal Auditory Displays for Virtual Acoustics), n. CPDA135702 dell'Università di Padova.

6. Bibliografia

[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] 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

[6] M. Geronazzo, S. Spagnol, A. Bedin, and F. Avanzini. *Enhancing vertical localization with image-guided selection of non-individual head-related transfer functions*. In Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP 2014), Firenze, Italy, May 2014

[7] 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

[8] 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

[9] 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

[10] 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

[11] 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

[12] http://interface.cipic.ucdavis.edu/sound/hrtf.html

[13] 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