# STIMA DI FEATURE SPETTRALI DI HRTF MEDIANTE MODELLI ANTROPOMETRICI NON LINEARI PER LA RESA DI AUDIO 3D

**Simone Spagnol**
University of Iceland
Reykjavik, Islanda
spagnols@hi.is

**Silvio Galesso, Federico Avanzini**
Università di Padova
Padova, Italia
avanzini@dei.unipd.it

## SOMMARIO

La relazione tra i parametri antropometrici di un soggetto umano e le feature tipiche delle Head-Related Transfer Function (HRTF), in particolare quelle collegabili al padiglione auricolare (o *pinna*), non è compresa appieno. In questo articolo applichiamo tecniche di elaborazione del segnale per estrarre le frequenze del primo notch dovuto alla pinna (conosciuto come $N_1$) nella porzione frontale del piano mediano e costruiamo un modello basato su una rete neurale artificiale che relazioni le frequenze stesse a 13 diversi parametri antropometrici della pinna, alcuni dei quali dipendono dall'elevazione della sorgente sonora. I risultati mostrano una corrispondenza incoraggiante tra l'antropometria e le feature spettrali, la quale conferma la possibilità di predire la frequenza centrale del notch a partire da una semplice fotografia dell'orecchio.

## 1. INTRODUZIONE

La maggior parte delle tecniche per la resa binaurale del suono fa affidamento sull'utilizzo di Head-Related Transfer Function (HRTF), ovvero filtri che catturano gli effetti acustici del corpo umano [1]. Le HRTF permettono una simulazione realistica del segnale che giunge all'ingresso del canale uditivo in funzione della posizione della sorgente sonora nello spazio. La soluzione ideale, ovvero quella che meglio approssima l'ascolto spaziale reale, richiede l'utilizzo di HRTF individuali misurate acusticamente sull'ascoltatore stesso [2]. Tuttavia, la misurazione acustica di HRTF richiede una strumentazione generalmente costosa e procedure di registrazione invasive [3]. Tale è la ragione per cui in pratica viene spesso favorito l'utilizzo di HRTF non individuali, misurate su manichini o su altri individui. Lo svantaggio di tali HRTF è che difficilmente sono relazionabili all'antropometria dell'ascoltatore, in particolare quella della pinna. Ciò si traduce in probabili e frequenti errori di localizzazione, quali inversioni *front/back*, errata percezione dell'elevazione, e localizzazione all'interno della testa [4].

Diverse tecniche per il design di HRTF sintetiche sono state proposte nelle ultime due decadi per affrontare tali problemi. Una delle più promettenti è la modellazione strutturale [5]. Secondo questo approccio, gli effetti acustici più rilevanti per la percezione spaziale del suono (ritardi ed effetti d'ombra dovuti alla testa, riflessioni sui bordi della pinna e sulle spalle, eccetera) vengono separati e modellati ciascuno con una struttura di filtri digitali. I vantaggi di tale approccio rispetto ad altre tecniche per la resa binaurale del suono sono la possibilità di personalizzare le HRTF sull'ascoltatore, grazie all'acquisizione di quantità antropometriche (raggio della testa, forma della pinna, larghezza delle spalle, e così via), e l'efficienza computazionale, in quanto ogni modello è sottostrutturato in diversi blocchi ciascuno dei quali simula un singolo effetto acustico. Ciononostante, la precedente letteratura riguardante la relazione tra effetti acustici ed antropometria - tra cui applicazioni di metodi di regressione su database di HRTF [6, 7, 8, 9] - hanno prodotto risultati controversi, evidenziando come molte di queste relazioni non siano tuttora comprese appieno.

Possiamo ipotizzare due principali cause del fallimento di questi studi. In primo luogo, nessuna informazione risultante dalla conoscenza pregressa delle componenti strutturali responsabili degli indicatori di localizzazione è presa in considerazione: molti lavori applicano ciecamente tecniche di apprendimento automatico su lunghi vettori di *feature* antropometriche, i quali includono molti parametri irrilevanti. In secondo luogo, l'intera HRTF o una versione ridotta dimensionalmente della stessa (ad esempio tramite *Principal Component Analysis*) viene usata come l'insieme di variabili *target*, senza l'applicazione di alcun passo di *preprocessing* che estragga le caratteristiche locali più importanti. In particolare, è risaputo che i minimi (notch) e i massimi (picchi) locali nella funzione di trasferimento sono essenziali per la percezione sonora della dimensione spaziale più "individuale", ossia l'elevazione [10].

L'obiettivo principale di questo articolo è esplorare la relazione tra le frequenze centrali dei notch in un insieme di HRTF misurate nella porzione frontale del piano mediano e un insieme di parametri antropometrici sotto forma sia di misure unidimensionali della pinna (altezza della pinna, larghezza della conca, eccetera) che di misure variabili con l'elevazione della sorgente sonora (ovvero le distanze tra il canale uditivo e i contorni della pinna). Il punto di partenza dell'articolo (riportato in Sezione 2) è un lavoro precedente [11] che suggerisce come le frequenze dei principali minimi spettrali nelle HRTF siano strettamente legate alla forma della pinna, in particolare alle sopracitate distanze. La Sezione 3 descrive la metodologia di estrazione di

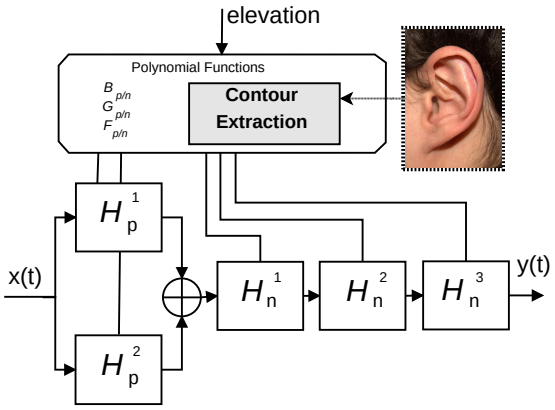


**Figura 1**. Rappresentazione schematica del modello strutturale della pinna.

*feature* e il modello di regressione non lineare utilizzato, mentre i risultati sono riportati e discussi nella Sezione 4.

## 2. UN MODELLO STRUTTURALE DI PINNA

Le differenze più marcate tra le HRTF di due diversi soggetti sono principalmente imputabili alle diverse caratteristiche della pinna (forma, grandezza e orientamento), talmente varie da rendere la pinna oggetto di recenti studi di identificazione biometrica [12]. La pinna ha un ruolo fondamentale nel determinare la forma spettrale delle HRTF grazie a due fenomeni acustici principali: riflessioni (sui bordi principali) e risonanze (nelle cavità). Di conseguenza, l'HRTF presenta nella sua ampiezza una sequenza di amplificazioni (picchi) in corrispondenza delle frequenze di risonanza e di nette attenuazioni (*notch*) in corrispondenza delle frequenze in cui si ha massima interferenza distruttiva tra onde dirette e onde riflesse. È proprio la collocazione spettrale di tali picchi e notch a costituire un indicatore fondamentale per la caratterizzazione della posizione spaziale della sorgente sonora, e in particolare della sua elevazione [10].

Nella letteratura dedicata troviamo diverse proposte modellistiche per rendere sinteticamente la componente della HRTF relativa alla pinna, conosciuta come *Pinna-Related Transfer Function* (PRTF). Tuttavia, tali modelli soffrono di evidenti limiti: la presenza della sola componente riflettente [13], la validità in regioni spaziali eccessivamente ristrette [14], e/o l'assenza di una parametrizzazione esplicita sull'antropometria dell'ascoltatore [15]. In un lavoro precedente [11] gli autori hanno proposto un modello strutturale di pinna costituito da due blocchi di filtri, il blocco *risonante* e il blocco *riflettente*. Il blocco risonante è costituito da due filtri *peak* del secondo ordine posti in parallelo; le uscite di tale blocco vengono sommate per essere inviate all'ingresso del blocco riflettente, costituito dalla cascata di 3 filtri *notch* del secondo ordine. Il modello è riportato schematicamente in Figura 1.

Nell'ambito dello stesso lavoro, gli autori hanno studiato la relazione tra le frequenze centrali dei notch presenti nelle PRTF e la geometria della pinna. A tal fine è stata utilizzata una procedura di *ray-tracing* su immagini 2D della pinna [1] per mappare i punti di riflessione a una certa distanza dal punto di riferimento del canale uditivo, univocamente determinata dalla frequenza centrale di ogni notch. Gli autori hanno verificato che l'utilizzo di coefficienti di riflessione negativi risulta un fattore chiave nella determinazione della frequenza centrale stessa. La relazione tra frequenza del notch e distanza del punto di riflessione dal canale uditivo è quindi completamente riassunta nella semplice equazione

$$D_i(\phi) = \frac{c}{2F_i(\phi)}, \tag{1}$$

dove la costante $c$ rappresenta la velocità del suono, $\phi$ è l'elevazione di cui si considera la PRTF, $F_i$ è la frequenza centrale dell'$i$-esimo notch $N_i$, e $D_i$ è la distanza tra il corrispondente punto di riflessione e l'ingresso del canale uditivo. Dai punti così trovati e mappati su immagini di pinne di un *pool* di soggetti sperimentali, gli autori hanno notato l'ottima corrispondenza tra i punti di riflessione e i tre contorni principali della pinna, ovvero bordo dell'elice, antielice/parete interna della conca, e bordo esterno della conca.

Partendo da tali risultati, gli autori in [16] hanno realizzato la procedura inversa: a partire dall'immagine della pinna sono stati tracciati i tre sopracitati contorni e, attraverso semplici calcoli trigonometrici, trasformati in coppie di coordinate polari $(D_i(\phi), \phi)$ rispetto al canale uditivo. Dall'Equazione 1 si ricavano quindi le frequenze centrali dei notch ad ogni elevazione $\phi$ desiderata e per ognuno dei tre contorni. L'unico parametro indipendente usato dal modello è infatti l'elevazione della sorgente sonora virtuale, alla quale sono associate tre funzioni polinomiali che interpolano le frequenze centrali ricavate dai tre contorni (vedi Figura 1) per un certo passo di campionamento $\Delta\phi$. Per ciò che riguarda i parametri di larghezza di banda e profondità dei notch stessi, non è stata trovata una relazione con la forma della pinna. Gli autori hanno precedentemente utilizzato per questi due parametri valori medi di una popolazione di soggetti sperimentali [16], così come per i parametri caratteristici delle risonanze.

## 3. METODOLOGIA

L'insieme di dati di partenza è costituito dalle Head-Related Impulse Response (HRIR) incluse nel database di HRTF CIPIC [17], un database di dominio pubblico di HRIR misurate ad alta risoluzione spaziale in $1250$ diverse direzioni per $45$ diversi soggetti. Poiché in questo lavoro consideriamo l'antropometria dei soggetti sia sotto forma di dati numerici (i parametri antropometrici inclusi nel database CIPIC) che di una fotografia della pinna sinistra o destra, abbiamo selezionato i $33$ soggetti per i quali questi dati sono disponibili per intero.

Prendiamo come riferimento il sistema di coordinate utilizzato nel database CIPIC, ovvero il sistema interaurale

[1] Il *ray-tracing* viene approssimativamente effettuato sulla proiezione dei contorni della pinna su di un piano parallelo al piano mediano del soggetto sperimentale.

**Tabella 1**. Coefficienti di correlazione di Pearson tra i parametri antropometrici della pinna.

| parametro | descrizione | $d_2$ | $d_3$ | $d_4$ | $d_5$ | $d_6$ | $d_7$ | $d_8$ | $\theta_1$ | $\theta_2$ |
|---|---|---|---|---|---|---|---|---|---|---|
| $d_1$ | altezza conca inferiore | -0.06 | 0.10 | 0.19 | 0.51 | 0.21 | 0.23 | 0.24 | -0.04 | 0.20 |
| $d_2$ | altezza conca superiore | | -0.02 | 0.33 | 0.47 | 0.13 | 0.11 | 0.30 | 0.02 | -0.11 |
| $d_3$ | larghezza conca inferiore | | | 0.03 | 0.19 | 0.47 | 0.59 | 0.27 | 0.23 | 0.02 |
| $d_4$ | altezza fossa triangolare | | | | 0.67 | 0.53 | 0.03 | 0.30 | -0.06 | -0.28 |
| $d_5$ | altezza pinna | | | | | 0.52 | 0.22 | 0.44 | -0.11 | 0.00 |
| $d_6$ | larghezza pinna | | | | | | 0.16 | 0.45 | 0.09 | -0.24 |
| $d_7$ | larghezza incisura intertragica | | | | | | | 0.15 | -0.09 | 0.03 |
| $d_8$ | profondità conca inferiore | | | | | | | | 0.01 | 0.14 |
| $\theta_1$ | angolo di rotazione | | | | | | | | | -0.12 |
| $\theta_2$ | angolo di apertura | | | | | | | | | |

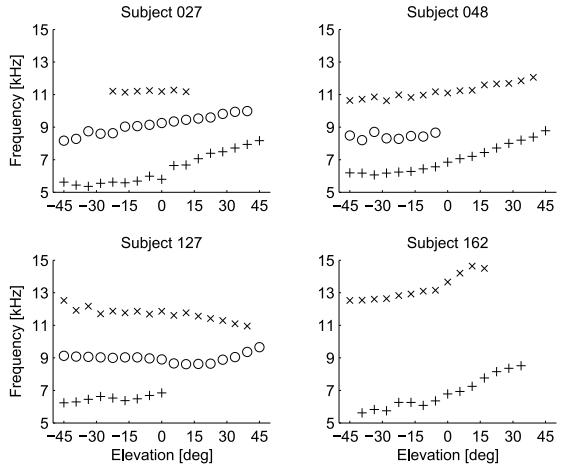


**Figura 2**. Notch estratti e raggruppati per quattro soggetti CIPIC (N1:+, N2:∘, N3:×).

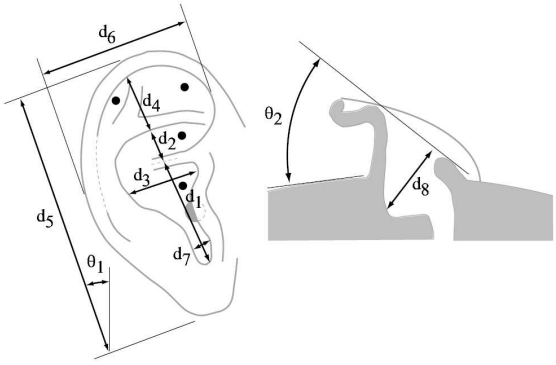


**Figura 3**. I 10 parametri antropometrici della pinna inclusi nel database CIPIC (figura riprodotta da [17]).

polare, restringendo l'analisi alla porzione frontale del piano mediano (azimut $\theta = 0°$), con l'elevazione $\phi$ che varia tra $\phi = -45°$ e $\phi = 45°$ a passi di $5,625$ gradi (17 HRIR per soggetto). La ragione per considerare il solo piano mediano è dovuta alla constatazione che variazioni relative di azimut fino a $\Delta\theta = 30°$ per una data elevazione producono modeste variazioni spettrali nella PRTF [18], permettendoci di generalizzare il modello a un intervallo di valori più ampio per l'azimut. Le elevazioni maggiori di $45°$ sono state invece scartate a causa della tipica mancanza di notch nelle HRTF corrispondenti [19].

### 3.1 Estrazione delle frequenze dei notch

Al fine di ottenere le frequenze dei notch più marcati, abbiamo applicato l'algoritmo di elaborazione del segnale di Raykar *et al.* [18] ad ogni HRIR disponibile. In sintesi, l'algoritmo prevede il calcolo della funzione di autocorrelazione del residuo della predizione lineare della HRIR ed estrae le frequenze dei notch come i minimi locali della funzione di *group delay* risultanti al di sotto di una determinata soglia (fissata euristicamente a $-0.5$ campioni). Successivamente, i notch appartenenti alla riflessione sullo stesso contorno sono stati raggruppati lungo le diverse elevazioni $\phi$ attraverso l'algoritmo di *partial tracking* di McAulay e Quatieri [20]. L'intervallo di *matching* per l'algoritmo è stato fissato a $\Delta = 1$ kHz. La stessa procedura di estrazione e raggruppamento di notch è stata utilizzata in un lavoro precedente [21].

Abbiamo quindi considerato solo i raggruppamenti (o *track*) con almeno 3 notch. Nei casi in cui un soggetto presenti più di 3 track che soddisfino tale requisito, abbiamo considerato i 3 track più lunghi ed etichettato ogni frequenza del notch con $F_1$, $F_2$ e $F_3$ in ordine crescente di frequenza media. Nei casi in cui un soggetto presenti meno di 3 track (9 casi su 33), le etichette sono state assegnate con un criterio di vicinanza alla frequenza mediana del track calcolata su tutti i soggetti con 3 track. In Figura 2 sono riportati i track di quattro soggetti rappresentativi. Complessivamente, la procedura applicata ai 33 soggetti ha prodotto 367 diversi campioni per $F_1$, 401 per $F_2$ e 303 per $F_3$.

### 3.2 Estrazione di feature antropometriche

Trentasette misure antropometriche unidimensionali per ognuno dei 33 soggetti sono disponibili nel database CIPIC: 17 per la testa e il torso e 10 per ognuna delle due pinne. Poiché il *focus* di questo lavoro è sui notch spettrali, consideriamo soltanto i parametri della pinna in esame, riportati nella Figura 3 e descritti nella Tabella 1. La tabella ri-

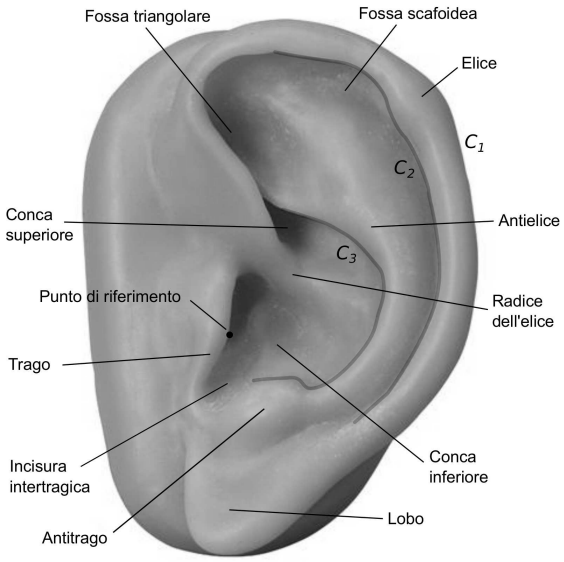


**Figura 4**. Anatomia della pinna e i tre contorni estratti: $C_1$, $C_2$ e $C_3$.

porta inoltre i valori di correlazione tra parametri calcolati sui soggetti CIPIC. Generalmente tali valori sono modesti, tolti quelli per le misure sovrapposte (ad esempio $d_4 - d_5$, $d_3 - d_7$), e ciò denota un grado di ortogonalità accettabile tra i parametri antropometrici.

Inoltre, abbiamo estratto parametri dipendenti dall'elevazione strettamente correlati all'idea di *ray tracing* sviluppata nella precedente Sezione 2. I tre contorni corrispondenti al bordo esterno dell'elice, al bordo interno dell'elice, e al bordo della conca inferiore/antitrago (vedi Figura 4) sono stati dapprima tracciati a mano con l'aiuto di una tavoletta grafica e memorizzati come sequenze di pixel.[2] Successivamente, il punto di massima protuberanza del trago è stato scelto come punto di riferimento del canale uditivo (vedi ancora Figura 4) per il calcolo dei parametri di distanza. Per ogni angolo di elevazione $\phi \in [-45, 45]$, le distanze in pixel tra il punto di riferimento e il punto che interseca ogni contorno della pinna lungo il raggio che ha origine nel punto di riferimento con pendenza $-\phi$ sono infine convertite in centimetri facendo affidamento sul metro riportato nelle fotografie accanto alla pinna e memorizzate come $r_i(\phi)$, dove $i \in \{1, 2, 3\}$ si riferisce al contorno $C_i$ associato.

A causa della forma approssimativamente ellittica della pinna, i tre parametri $r_1$, $r_2$ e $r_3$ sono strettamente correlati. Precisamente, il coefficiente di correlazione tra $r_1$ e $r_2$ è $0.95$, ed entrambi correlano con un valore di $0.75$ con $r_3$.

[2] Precisiamo che questi tre contorni non corrispondono univocamente ai tre contorni ipotizzati responsabili delle riflessioni nella precedente Sezione 2. La ragione di tale scelta è pratica: i contorni considerati possono essere estratti automaticamente in maniera robusta attraverso tecniche di elaborazione dell'immagine basate su rilevamento di profondità con dispositivi multi-flash [22]. Tuttavia, in questo lavoro l'estrazione manuale dei contorni si è resa necessaria a causa della disponibilità di singole immagini della pinna e della limitata efficacia di tecniche basate sull'intensità (e.g. Canny, Sobel) su immagini a basso contrasto.

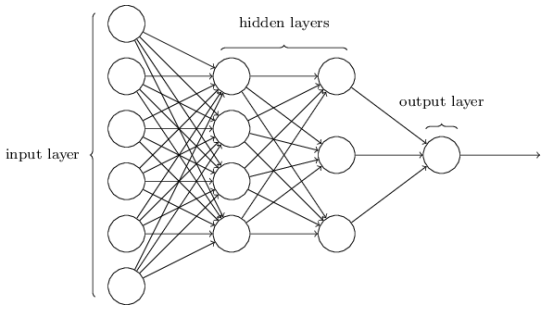


**Figura 5**. Rappresentazione di un percettrone multistrato.

L'elevata correlazione tra $r_1$ e $r_2$ può anche essere spiegata dal fatto che, semplicemente, entrambi i parametri si riferiscono all'elice.

### 3.3 Modello di regressione

Si vuole ora studiare la relazione tra le frequenze $F_1$, $F_2$ e $F_3$ e i tredici parametri misurati ($d_i$, $i = 1 \dots 8$, $\theta_j$, $j = 1 \dots 2$, $r_k(\phi)$, $k = 1 \dots 3$, $\phi \in [-45, 45]$) sui dati sopra descritti. Il problema può essere facilmente configurato come un problema di regressione, dove gli attributi sono i parametri antropometrici e le variabili target sono le frequenze dei notch. In un lavoro precedente [23] gli autori hanno verificato sullo stesso insieme di dati tramite tecniche di regressione lineare che i 13 parametri antropometrici considerati non sono sufficienti per stimare con un errore accettabile i notch $N_2$ e $N_3$. Test preliminari condotti utilizzando il modello non lineare descritto di seguito al posto della regressione lineare hanno evidenziato la medesima difficoltà di apprendimento. Di conseguenza, l'analisi seguente si concentra sulla stima di $N_1$.

Il modello non lineare utilizzato per la regressione è il percettrone multistrato (vedi Figura 5), un tipo di rete neurale artificiale elementare ma molto utilizzato la cui struttura è caratterizzata da tre o più strati di neuroni artificiali (nodi) in cascata. Di questi strati il primo è di input, l'ultimo è di output, mentre quelli tra essi compresi sono detti "strati nascosti" *(hidden layers)*. Ciascun nodo è connesso con un certo peso sinaptico $w_{ij}$ a ogni nodo dello strato successivo, con il verso della connessione sempre diretto verso l'uscita. Le componenti in ingresso (nel nostro caso i vettori di attributi) vengono trasmesse ai nodi di input, che calcolano la propria funzione di attivazione (non lineare) e ne trasmettono il valore allo strato successivo; la propagazione prosegue fino ai nodi di output (nel nostro caso singolo e relativo a $F_1$) [24, 25].

Poiché tutti i neuroni di una rete hanno la stessa funzione di attivazione, gli elementi parametrici variabili sono i pesi sinaptici. La fase di apprendimento supervisionato del percettrone, eseguita fornendo un *training set*, è un problema di ricerca del minimo di una funzione di errore tra le stime fornite in uscita a partire dagli ingressi e i valori reali delle variabili in uscita, e si affronta con la tecnica della *discesa del gradiente*. Una volta calcolata l'uscita della rete per una determinata istanza in ingresso, si procede al calcolo della variazione del peso di ciascuna sinapsi, proce-

dendo a ritroso, a partire dal nodo di uscita *(error backpropagation)* [25]. Tale operazione viene svolta ciclicamente, con ogni ciclo chiamato *epoca di apprendimento*.

Mentre il numero di nodi di input e output è fissato, è possibile intevenire sul numero e sulla dimensione degli strati interni: in generale, maggiore è la complessità della rete nascosta e più la rete neurale è efficace nella stima di funzioni articolate e di caratteristiche particolarmente astratte presenti nel dataset (*deep learning*). Oltre un certo livello di complessità, tuttavia, la rete tende all'*overfitting*. Nel nostro caso, avendo un insieme di dati per l'addestramento limitato e non dovendo stimare caratteristiche di alto livello, è stato impiegato un solo strato nascosto. Il numero di neuroni nascosti è stato invece stabilito in base al numero di nodi di input, come vedremo nella sezione successiva.

La motivazione dell'utilizzo del percettrone multistrato è duplice. In primo luogo, l'acquisizione di HRTF e dei parametri antropometrici è soggetta a errore di misurazione. L'impiego di una rete neurale, se correttamente dimensionata, permette di ridurre l'effetto di tale rumore. In secondo luogo, la scelta del percettrone multistrato permette di costruire un modello dalla struttura arbitraria e articolata, in grado di stimare eventuali nonlinearità.

## 4. REGRESSIONE ANTROPOMETRICA

Vengono ora descritti tre diversi modelli di regressione non lineare che si differenziano per gli attributi considerati in ingresso e quindi anche per il numero di nodi negli strati di input e nascosto. Il primo modello $M_1$ considera i dieci parametri CIPIC, il secondo modello $M_2$ i parametri di distanza estratti, e il terzo modello $M_3$ la combinazione dei due insiemi.

Per l'addestramento e la validazione dei modelli vengono usati rispettivamente un *training set* e un *test set* ricavati partizionando le $367$ istanze di partenza di $N_1$ per soggetti. Infatti, poiché il modello deve stimare caratteristiche di un soggetto nuovo, è fondamentale per una corretta validazione che i dati vengano partizionati senza separare le istanze relative allo stesso soggetto. Sono quindi state scelte $20$ diverse combinazioni di $7$ soggetti in modo tale che ogni test set fosse di circa il $25\%$ delle istanze totali, lasciando il restante $75\%$ nel training set. Il modello viene quindi sottoposto a *cross-validazione*, considerando la media delle statistiche risultanti da ogni test set.

La sperimentazione è stata condotta all'interno dell'ambiente *WEKA*, una piattaforma open source per l'apprendimento automatico caratterizzata da una vasta collezione di strumenti statistici ed algoritmi per l'analisi dei dati [24]. Tali strumenti sono stati utilizzati mediante API Java al fine di automatizzare e raffinare l'analisi.

### 4.1 Primo modello

I parametri del modello $M_1$ sono indipendenti dall'elevazione della sorgente sonora, che è pertanto necessario includere come attributo indipendente. I vettori di attributi includono quindi un indice intero di elevazione $n \in \{1, ..., 17\}$ e i $10$ parametri CIPIC $d_i$, $i = 1 \dots 8$, $\theta_j$, $j = 1 \dots 2$. Il percettrone è dunque vincolato ad avere $11$ nodi di ingresso e uno di uscita ($F_1$). É inoltre stato scelto euristicamente uno strato nascosto con $2$ soli nodi.

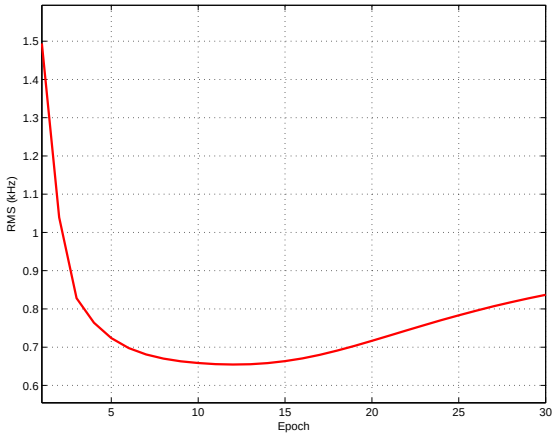


**Figura 6**. Progresso dell'addestramento di $M_1$ sul test set che include i soggetti CIPIC $003$, $010$, $050$, $058$, $127$, $137$ e $152$.

In Figura 6 viene riportato il processo di apprendimento in termini di errore RMS rispetto al numero di epoche di addestramento, con un test set rappresentativo. L'errore diminuisce progressivamente fino a raggiungere un livello ottimale di apprendimento oltre il quale si verifica overfitting. A questo punto il modello perde progressivamente capacità di generalizzazione e quindi l'errore di predizione sul test set torna a crescere.

Le prestazioni medie di $M_1$ sono: RMSE minimo di $604$ Hz e correlazione $r = 0.88$. I valori minimi dell'RMSE sono tuttavia molto diversi tra loro: si va da $425$ Hz a $1$ kHz con una deviazione standard di $148$ Hz nelle $20$ istanze considerate. Tutte le istanze considerate raggiungono il minimo valore dell'RMSE entro $40$ epoche di apprendimento.

### 4.2 Secondo modello

I parametri del modello $M_2$, le distanze $r_k(\phi)$, $k = 1 \dots 3$, $\phi \in [-45, 45]$, sono al contrario dipendenti dall'elevazione della sorgente sonora, che non è stata pertanto considerata. In questo caso i nodi di input del percettrone sono necessariamente 3, uno per ogni distanza. La rete ottimale risulta ancora composta da uno strato di 2 nodi interni nascosti. Poiché in generale la proporzionalità tra le frequenze dei notch e i parametri antropometrici estratti è inversa, prima di applicare l'operazione di regressione è stato considerato il reciproco di questi ultimi. La Figura 7 mostra a titolo esemplificativo il fit polinomiale di quarto grado delle distanze $r_1(\phi)$ dipendentemente dalla frequenza di $N_1$.

I valori medi di RMSE e correlazione sono rispettivamente $670$ Hz e $r = 0.84$. La differenza tra gli errori minimi è stavolta più contenuta, e spazia tra i $500$ e gli $800$ Hz con una deviazione standard di $92$ Hz. Tuttavia il percettrone, a causa del numero molto ridotto di nodi, ri-

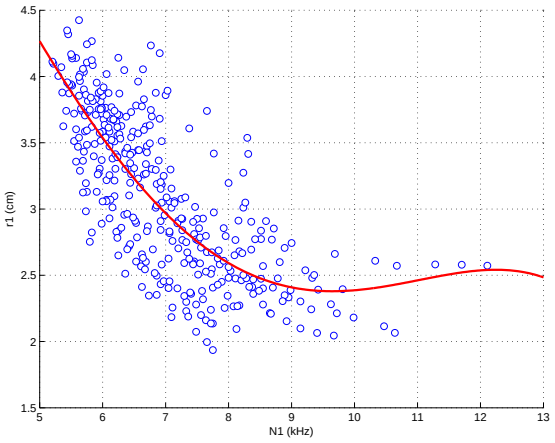


**Figura 7**. Grafico di dispersione tra $F_1$ e $r_1(\phi)$. La curva rappresenta il fit polinomiale di quarto grado dei dati.

chiede fino a $400$ epoche di apprendimento per assestarsi sulle prestazioni migliori.

### 4.3 Terzo modello

Avendo in ingresso tutti i parametri a disposizione meno l'elevazione, il modello $M_3$ ha $13$ nodi di input. La struttura interna scelta è in questo caso composta da $3$ nodi nascosti, uno in più rispetto ai precedenti modelli.

Le prestazioni complessive di questo modello risultano (prevedibilmente) migliori rispetto a quelle dei precedenti, in particolare in termini di RMSE, che scende a $554$ Hz con una deviazione standard di $84$ Hz, mentre la correlazione è $r = 0.88$. I tempi di apprendimento ottimali non superano mai le $70$ epoche. Il valore RMSE risulta inoltre leggermente inferiore al valore ottenuto con la regressione lineare in [23] ($590$ Hz). Questo risultato suggerisce che l'introduzione di una componente non lineare abbia aumentato la precisione della stima, senza tuttavia migliorarne drasticamente i risultati. Tuttavia, occorre specificare che in questo lavoro è stata utilizzata una differente procedura di validazione dei modelli.

Il modello $M_3$ introduce errori significativi se utilizzato per predire $F_1$, ed è risaputo che lo spostamento in frequenza degli indicatori di elevazione ha un impatto notevole sull'accuratezza della localizzazione. Ad esempio, spostamenti di $1$ kHz di $N_1$ possono corrispondere ad un aumento o una diminuzione dell'angolo di elevazione di $20^\circ$ od oltre [26]. Tuttavia, la letteratura [27] suggerisce che due notch ad alta frequenza (attorno a $8$ kHz) che differiscano soltanto nel valore della frequenza centrale possano essere percepiti come distinti solo se tale differenza è almeno pari al $10\%$ circa del valore di frequenza inferiore.

Se teniamo inoltre in considerazione il fatto che la procedura utilizzata contiene inevitabilmente errori, ad esempio l'estrazione completamente automatica delle frequenze dei notch, o il piazzamento arbitrario del punto di riferimento del canale uditivo necessario per calcolare le distanze $r_k$, possiamo concludere che la stima di $F_1$ a partire dai parametri antropometrici considerati è possibile.

## 5. CONCLUSIONI

In questo articolo abbiamo studiato l'applicazione di un modello di regressione non lineare per stimare la frequenza dei notch nelle HRTF relative alla porzione frontale del piano mediano a partire da dati antropometrici. I risultati per il notch $N_1$ mostrano una corrispondenza incoraggiante tra l'antropometria e le feature spettrali, e ciò conferma la possibilità di predire la frequenza centrale a partire da una semplice fotografia dell'orecchio.

Tuttavia, occorre precisare che la quantità limitata dei dati sperimentali a disposizione può avere influito sui risultati. In tale situazione, per evitare l'overfitting è necessario ridurre il numero di epoche di apprendimento. Ne consegue che il modello risultante resta in una certa misura dipendente dai pesi iniziali della rete, stabiliti in modo pseudocasuale. La qualità dei dati sperimentali può inoltre avere avuto un impatto altrettanto importante. I modelli sono stati addestrati e validati sul database CIPIC, che sebbene sia il database di HRTF più utilizzato dalla comunità scientifica soffre di errori di misura (ad esempio evidenti asimmetrie sul piano orizzontale [28]) e mancanza di documentazione (ad esempio non è specificato con esattezza il punto di misurazione delle HRTF).

Per questi motivi, i risultati ottenuti vanno interpretati come un'indicazione sulle potenzialità dell'approccio sperimentale, in prospettiva tanto dell'utilizzo di un database più recente e documentato [19, 29] quanto di avanzamenti nelle tecniche di estrazione dei contorni della pinna.

## 6. RINGRAZIAMENTI

Questo lavoro è stato finanziato dal progetto di ricerca PADVA (Personal Auditory Displays for Virtual Acoustics), n. CPDA135702 dell'Università di Padova e dal programma di ricerca e innovazione dell'Unione Europea Horizon 2020 in virtù del contratto di sovvenzione n.643636. [3]

## 7. 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, pp. 231–249, April 2001.

[2] D. R. Begault, E. M. Wenzel, and M. R. Anderson, "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, pp. 904–916, October 2001.

[3] B. Xie, *Head-Related Transfer Function and Virtual Auditory Display*. Plantation, FL, USA: J.Ross Publishing, 2nd ed., June 2013.

[4] H. Møller, M. F. Sørensen, C. B. Jensen, and D. Hammershøi, "Binaural technique: Do we need individual

[3] This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 643636. `http://www.soundofvision.net/`

recordings?," *J. Audio Eng. Soc.*, vol. 44, pp. 451–469, June 1996.

[5] C. P. Brown and R. O. Duda, "A structural model for binaural sound synthesis," *IEEE Trans. Speech Audio Process.*, vol. 6, pp. 476–488, September 1998.

[6] L. Li and Q. Huang, "HRTF personalization modeling based on RBF neural network," in *Proc. 38th IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP 2013)*, (Vancouver, BC, Canada), pp. 3707–3710, May 2013.

[7] Q. Huang and L. Li, "Modeling individual HRTF tensor using high-order partial least squares," *EURASIP J. Adv. Signal Process.*, vol. 2014, pp. 1–14, May 2014.

[8] P. Bilinski, J. Ahrens, M. R. P. Thomas, I. J. Tashev, and J. C. Platt, "HRTF magnitude synthesis via sparse representation of anthropometric features," in *Proc. 39th IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP 2014)*, (Firenze, Italy), pp. 4501–4505, May 2014.

[9] F. Grijalva, L. Martini, S. Goldenstein, and D. Florencio, "Anthropometric-based customization of head-related transfer functions using Isomap in the horizontal plane," in *Proc. 39th IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP 2014)*, (Firenze, Italy), pp. 4506–4510, May 2014.

[10] S. K. Roffler and R. A. Butler, "Factors that influence the localization of sound in the vertical plane," *J. Acoust. Soc. Am.*, vol. 43, pp. 1255–1259, June 1968.

[11] 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, pp. 508–519, March 2013.

[12] A. Abaza, A. Ross, C. Hebert, M. A. F. Harrison, and M. S. Nixon, "A survey on ear biometrics," *ACM Trans. Embedded Computing Systems*, vol. 9, pp. 39:1–39:33, March 2010.

[13] A. J. Watkins, "Psychoacoustical aspects of synthesized vertical locale cues," *J. Acoust. Soc. Am.*, vol. 63, pp. 1152–1165, April 1978.

[14] 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), pp. 718–737, May 2007.

[15] K. J. Faller II, A. Barreto, and M. Adjouadi, "Augmented Hankel total least-squares decomposition of head-related transfer functions," *J. Audio Eng. Soc.*, vol. 58, pp. 3–21, January/February 2010.

[16] 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), pp. 174–179, December 2011.

[17] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF database," in *Proc. IEEE Work. Appl. Signal Process., Audio, Acoust.*, (New Paltz, New York, USA), pp. 1–4, October 2001.

[18] 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, pp. 364–374, July 2005.

[19] 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), pp. 209–212, September 2011.

[20] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 34, pp. 744–754, August 1986.

[21] S. Spagnol, "On distance dependence of pinna spectral patterns in head-related transfer functions," *J. Acoust. Soc. Am.*, vol. 137, pp. EL58–EL64, January 2015.

[22] S. Spagnol, D. Rocchesso, M. Geronazzo, and F. Avanzini, "Automatic extraction of pinna edges for binaural audio customization," in *Proc. IEEE Int. Work. Multi. Signal Process. (MMSP 2013)*, (Pula, Italy), pp. 301–306, October 2013.

[23] S. Spagnol and F. Avanzini, "Frequency estimation of the first pinna notch in head-related transfer functions with a linear anthropometric model," in *Proc. 18th Int. Conf. Digital Audio Effects (DAFx-15)*, (Trondheim, Norway), pp. 231–236, December 2015.

[24] R. Rojas, *Neural Networks: A Systematic Introduction*. Berlin: Springer-Verlag, 1996.

[25] B. Widrow and M. A. Lehr, "30 years of adaptive neural networks: Perceptron, Madaline, and backpropagation," *Proc. IEEE*, vol. 78, pp. 1415–1442, September 1990.

[26] J. Hebrank and D. Wright, "Spectral cues used in the localization of sound sources on the median plane," *J. Acoust. Soc. Am.*, vol. 56, pp. 1829–1834, December 1974.

[27] B. C. J. Moore, S. R. Oldfield, and G. J. Dooley, "Detection and discrimination of spectral peaks and notches at 1 and 8 kHz," *J. Acoust. Soc. Am.*, vol. 85, pp. 820–836, February 1989.

[28] S. Spagnol and F. Avanzini, "Anthropometric tuning of a spherical head model for binaural virtual acoustics based on interaural level differences," in *Proc. 21st Int. Conf. Auditory Display (ICAD 2015)*, (Graz, Austria), pp. 204–209, July 2015.

[29] C. Jin, P. Guillon, N. Epain, R. Zolfaghari, A. van Schaik, A. I. Tew, C. Hetherington, and J. Thorpe, "Creating the Sydney York morphological and acoustic recordings of ears database," *IEEE Trans. Multimedia*, vol. 16, pp. 37–46, January 2014.