Associazione Italiana di Acustica
34° Convegno Nazionale
Firenze, 13-15 giugno 2007

TEST D' ASCOLTO PER LA VALIDAZIONE DI RISPOSTE ALL' IMPULSO BINAURALI SINTETIZZATE MEDIANTE DUE DIFFERENTI TECNICHE.

Simone Campanini (1), Simone Fontana (1), Angelo Farina (1)

1) Laboratorio di Acustica ed Elettroacustica, Parma

1. Introduzione

Il programma Ramsete [1] fa parte della famiglia dei software previsionali utilizzati per la progettazione acustica delle sale. Attraverso un algoritmo di *pyramid-tracing* [2] è in grado di calcolare l'ecogramma (risposta energetica) della sala corrispondente alla propagazione sonora da un punto (sorgente) ad un altro (ricevitore) della sala considerata.

Per applicazioni quali ad esempio l'auralizzazione [3], è necessario risalire alla risposta all'impulso (IR, Impulse Response) della sala, a partire dall'ecogramma. Per rispondere a tale bisogno una possibilità è quella di utilizzare AudioConverter, software del pacchetto Ramsete, che opera una sintesi della IR a partire da bursts di rumore. Tale tecnica, descritta nel capitolo 2, presenta però alcuni inconvenienti uno dei quali è di natura estetica; altri inconvenienti sono di natura più quantitativa. Al fine di migliorare questi aspetti, si è introdotta una nuova tecnica di sintesi della IR a partire dall'ecogramma.

Un'altra possibilità di Audio Converter è di integrare i dati dell'ecogramma alle informazioni direzionali contenute nei primi raggi per ottenere delle risposte binaurali, adatte per un' auralizzazione binaurale, molto più realistica di una semplice auralizzazione mono, soprattutto grazie all'introduzione di una spiccata dimensione spaziale. Una volta ottenuta la risposta all'impulso binaurale (BIR, Binaural Impulse Response), l'effetto dell'auralizzazione può essere ascoltato attraverso un paio di buone cuffie, o tecniche più sofisticate, quali, ad esempio, lo stereo dipolo [11], che permettono un ascolto binaurale su due casse.

La sintesi di BIR è effettuata da AudioConverter utilizzando particolari filtri direzionali, chiamati HRTF (Head Related Tansfer Functions) [4]. I filtri utilizzati da AudioConverter erano qualli ottenuti a partire da misure sul manichino KEMAR [5]: nel presente studio filtri più accurati sono stati utilizzati per la sintesi delle BIR. In effetti si sono utilizzati filtri ottenuti a partire da misure su individui reali della base dati *Listen* dell'*IRCAM* [6]; tali filtri sono più lunghi e con una migliore equalizzazione. Inoltre una preselezione dei filtri HRTF da parte dell'ascoltatore, è stata realizzata in modo da permettergli un ascolto binaurale con le HRTF che, secondo la sua percezione, più si avvicinano alle sue.

Per verificare in maniera sperimentale i miglioramenti introdotti su AudioConverter, dei test percettivi preliminari sono stati effettuati su 5 persone, alla casa della Musica di Parma. In tali test, degli estratti audio ottenuti a partire da IR standard e IR migliorate sono stati riprodotti attraverso uno stereo dipolo e giudicati per capire se il miglioramento è percettibile.

2. Sintesi della risposta all'impulso

Il problema della sintesi di una risposta all'impulso si pone in modo sistematico ogniqualvolta vi sia il desiderio o la necessità di avere un riscontro uditivo del risultato di una simulazione acustica.

Ogni software di *ray* oppure *pyramid-tracing*, categoria, quest'ultima, cui appartiene Ramsete, applica di fatto l'approssimazione geometrica per descrivere la propagazione del campo sonoro. In effetti all'interno dell'ambiente vengono definiti almeno una *sorgente* irradiante una certa potenza con un ben determinato diagramma di direttività, almeno un *ricevitore* e le proprietà acustiche di tutte le superfici; l'algoritmo simula, quindi, la propagazione della suddetta potenza al trascorrere del tempo, annotando, istante per istante, quella intercettata dal ricevitore. L'insieme di tali dati forma il cosiddetto ecogramma; nel caso di Ramsete il risultato della simulazione è costituito da una matrice di dieci ecogrammi, uno per banda di ottava.

Non è di per sè possibile impiegare direttamente un ecogramma per un'operazione di convoluzione per una quantità di ragioni: innanzitutto il tempo di campionamento normalmente adottato per le simulazioni acustiche è piuttosto elevato, nell'ordine di 1 ms, fino a 10 ms. Un altro problema, è dovuto al fatto che in un ecogramma manca qualsiasi informazione di fase: i valori rilevati ed immagazzinati dal ricevitore sono infatti proporzionali alla densità di energia sonora presente in quel punto dell'ambiente in quell'istante, ossia sono numeri reali sempre maggiori di zero.

Il processo di sintesi di una IR udibile, quindi, deve operare una duplice interpolazione, sia nel dominio del tempo, sia nel dominio della frequenza, quest'ultima per il fatto che gli ecogrammi-campione sono solamente 10 per tutta la banda audio. In termini più pratici, nel momento in cui si porta il tempo di campionamento ad un valore adeguato a quello di un file audio, circa 20 microsecondi, si viene a creare una sequenza di zeri tra due campioni seguenti dell'ecogramma (figura 1). Si impone allora la necessità di colmare tale vuoto; una semplice operazione di interpolazione non è sufficiente poiché l'informazione di fase, non presente, deve essere in qualche modo ricreata dal nulla.

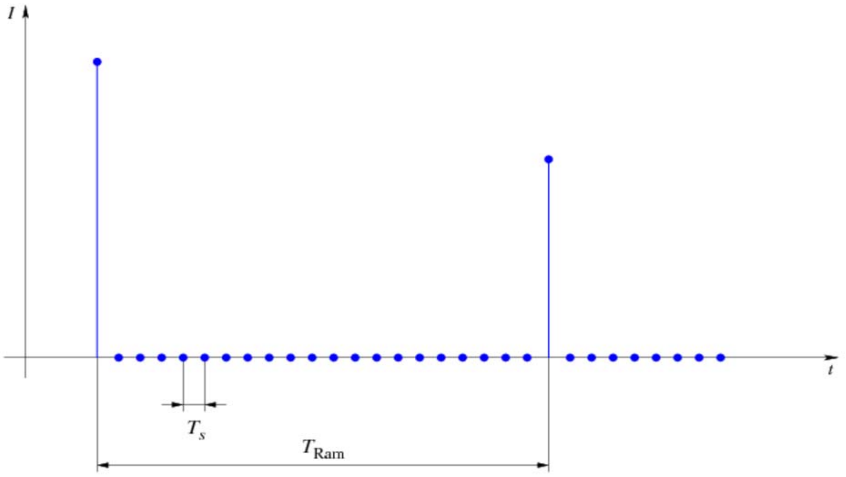

Figura 1 - Effetto del sovracampionamento: comparsa di 'zeri' tra due campioni dell'ecogramma.

In AudioConverter si utilizzano burst di rumore bianco (tecnica *noise-burst*), filtrati a seconda della banda di appartenenza: tale scelta é stata anche dettata dalla bande di tale burst, che ricopre interamente la sottobanda considerata.

Tuttavia, nonostante la ormai comprovata solidità ed attendibilità nel calcolo dei parametri acustici, le IR generate con l'algoritmo a rumore bianco presentano almeno due limiti: l'aspetto artefatto all'ascolto e la comparsa di innaturali “frastagliamenti” nei transitori quando impiegate per *Acoustic Quality Test*. Entrambi i problemi sono dovuti all'andamento fortemente discontinuo del rumore bianco.

Da quest'ultima osservazione è nata l'idea di sostituire il rumore con un altro tipo di segnale privo di discontinuità, ma comunque in grado di soddisfare alle necessità dell'applicazione. La scelta è caduta sulla somma di segnali sinusoidali, allo stato attuale 12 per ogni sottobanda (tecnica *sine-burst*); il comportamento stocastico del campo riverberante è stato mantenuto assegnando fase casuale ad ogni sinusoide generata nella coda della risposta.

Inoltre, per evitare la comparsa di discontinuità all'accostamento di due burst, questi vengono parzialmente sovrapposti nel tempo, previa moltiplicazione di ognuno per un'opportuna finestra del tipo *raised cosine flat-top*.

In conseguenza delle caratteristiche dei segnali sinusoidali è risultato assolutamente fondamentale assegnare ai burst una lunghezza variabile in funzione della banda: 100 ms per la banda centrata a 31.5 Hz, 50 ms per quella centrata a 63 Hz, e così via, mantenendo, però, una lunghezza di 1 ms per le ultime tre bande d'ottava. Così facendo, la maggior parte dei burst non risultava assegnata ad un unico campione dell'ecogramma, ma a più di uno, portando ad un problema nell'assegnazione di un'ampiezza al burst in questione: la soluzione, statisticamente ragionevole, è consistita nel prendere come ampiezza il valor medio delle radice quadrata del valore di ogni campione (esemplificazione in figura 2).

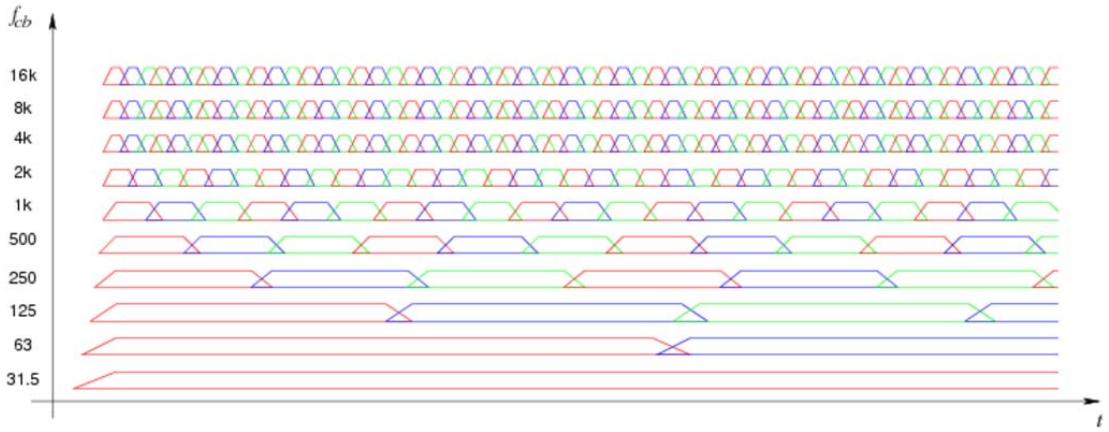

Figura 2 - Schema della costruzione della IR: ogni riga corrisponde ad uno dei dieci ecogrammi interpolato con burst (i 'trapezi') formati sommando 12 segnali sinusoidali. Il risultato è dato dalla somma delle righe.

3. Sintesi della risposta all'impulso binaurale

Quando si desideri produrre una IR spazialmente caratterizzata, come il caso binaurale, risulta essenziale disporre di informazioni geometriche quali la direzione del fronte d'onda incidente il ricettore ed il tempo di volo di quest'ultima, in modo da poter ricostruire in modo credibile la fase della riflessione in questione. La lungimiranza dei progettisti di Ramsete ha fatto sì che, a richiesta dell'utente, tale software produca un ulteriore file a corredo di quello contenente i risultati in forma di ecogrammi, nel quale, per ognuna delle prime due o tre decine di riflessioni, vengono riportate proprio le informazioni di cui sopra, oltre all'energia, sempre suddivisa nelle 10 bande, ad esse associate. Grazie a tutti questi dati è stato possibile estendere le funzionalità dell'algoritmo di sintesi nelle direzioni dei segnali multicanale tipici di almeno due sistemi di riproduzione audio spaziale: binaurale e B-Format. Nel presente articolo ci si occuperà solo del primo, tuttavia i concetti sono i medesimi in entrambi i casi.

La tecnica adottata consiste, sostanzialmente, nel dividere la IR in due parti: una *testa* ed una *coda riverberante*. La prima comprende i primi 80 ms [7] dall'onda diretta ed è quella che contiene *effettivamente* le informazioni spaziali cui il cervello umano è sensibile, la seconda è invece costituita dalla parte rimanente del segnale e può essere determinata tramite l'algoritmo precedentemente descritto. Ci si concentrerà, quindi, sull'algoritmo di sintesi della *testa* della IR, il quale è così riassumibile:

1. modellazione di ogni riflessione precoce tramite una delta di Dirac, filtrata con i valori di energia realmente immagazzinati dal ricevitore;
2. dati l'azimuth e l'elevazione della direzione di arrivo di ogni riflessione precoce, filtraggio delle stesse con le HRTF proprie di quegli angoli;
3. somma di tutti i segnali precedentemente ottenuti.

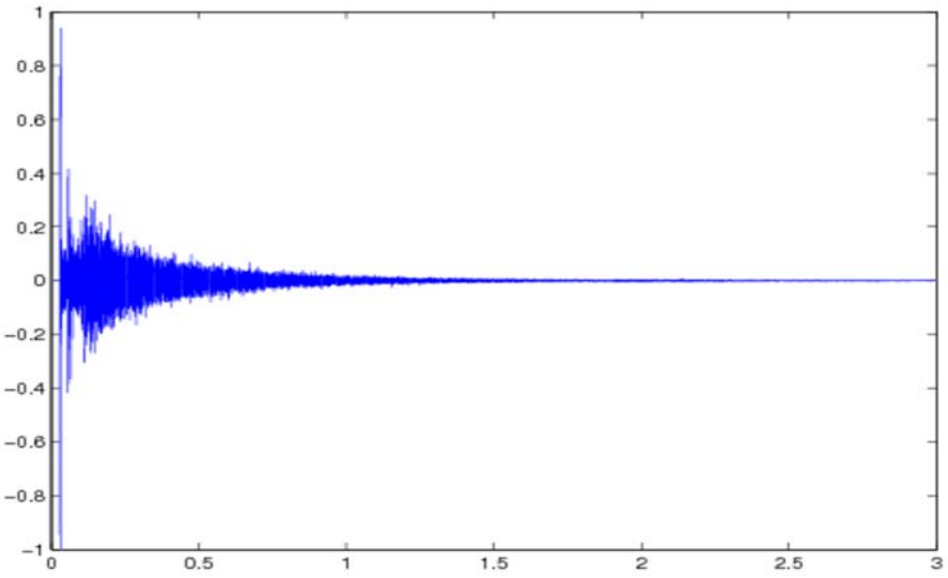

Figura 3 - Esempio di IR sintetizzata con la tecnica dei burst sinusoidali.

In particolare, si è deciso di impiegare i set di HRTF del progetto *Listen* dell'*IRCAM*, dopo averne riscontrata l'elevata qualità. Le *teste* delle IR ottenute seguendo il metodo precedente, tuttavia, mostravano ancora una certa artificiosità evidentemente dovuta alla modellazione delle prime riflessioni tramite impulsi ideali: per questa ragione si è aggiunta al software la possibilità compiere la stessa operazione mediante un impulso qualsiasi importato dall'esterno. Altri impulsi candidati sono in fase di studio.

In figura 3 si riporta una delle due risposte all'impulso binaurali ottenute con il metodo dei burst sinusoidali.

4. Personalizzazione delle HRTF

Le HRTF sono la trasformata di Fourier delle risposte all'impulso (Head Related Impulse Responses, HRIR) misurate in generale e quando possibile, tra i punti disposti su una sfera attorno all'individuo, e i timpani dell'individuo stesso. In questo modo si caratterizza il canale acustico tra la sorgente e il timpano, per una data posizione della sorgente. Si pensa che le HRIR rappresentino il mezzo con cui il sistema uditivo è capace di discriminare spazialmente delle sorgenti sonore. Misurandole, convolvendole con file audio non spazializzati e riproducendo i segnali ottenuti al timpano dell'ascoltatore, si possono quindi riprodurre artificialmente delle sorgenti disposte in una qualsiasi posizione della griglia di misura [4].

Tale processo di misura è particolarmente costoso in termini economici e di tempo, pertanto vari sistemi ([8],[9], e [10] per una lista delle tecniche attualmente studiate), sono stati proposti per evitarlo, pur garantendo una corrispondenza tra individuo e filtro. Questa tematica di ricerca, molto attuale, va sotto il nome di personalizzazione di HRTF.

Il sistema qui utilizzato è simile a quello proposto in [9]. Una sequenza audio formata da tre ripetizioni di un segnale a larga banda è stata sintetizzata in 4 diverse posizioni

nel piano azimutale (0, +/- 90 gradi, 180 gradi). La sintesi è stata fatta a partire dalle 4 HRTF stereo equalizzate corrispondenti a dette posizioni nella base dati *Listen* dell'*IRCAM*. Sono stati creati segnali di test per tutti gli individui contenuti nella suddetta base dati, in modo da avere la più ampia gamma di scelta possibile.

I files contenenti le sequenze binaurali sono stati presentati agli individui attraverso un sistema detto stereo dipolo, di cui parleremo brevemente nel capitolo successivo. La conoscenza a priori della posizione sintetizzata e la migliore o peggiore corrispondenza con la posizione percepita determina la scelta da parte dell'individuo del set di HRTF personalizzato, che sarà usato nella costruzione delle risposte binaurali.

5. Stereo dipolo

Le tecniche binaurali sono volte alla riproduzione al timpano dell'individuo dello stesso segnale che avrebbe ricevuto se presente all'atto della performance. Il segnale binaurale, ossia quello ai timpani dell'individuo, che sia registrato o sintetizzato, deve essere poi riprodotto. Il metodo più immediato è ovviamente quello dell'ascolto in cuffia.

Per questioni ergonomiche, ma anche percettive, si è cercato di permettere un ascolto binaurale anche su sistemi di casse, il cui principale problema è costituito dal crosstalk, ovverosia il fatto che il segnale di destra e di sinistra, a differenza dell'ascolto in cuffia, interferiscono se non tenuti separati. I sistemi che cercano di cancellare il crosstalk sono detti sistemi transaurali. Uno dei problemi di detti sistemi è il fatto che la cancellazione del crosstalk è efficace in un punto preciso dello spazio, ma si deteriora rapidamente per posizioni differenti. Tra i vari sistemi proposti, qui consideriamo lo stereo dipolo, che è meno sensibile a deboli movimenti dell'individuo, i quali possono risolvere tra l'altro la confusione front-back tipica dei sistemi binaurali.

Il processo di creazione dello stereo dipolo [11] prevede l'acquisizione di risposte all'impulso binaurali (nel nostro caso, tramite dummy head – Neumann Ku100), inversione della matrice delle risposte binaurali, e cross-convoluzione del segnale binaurale con la rete di filtri inversi.

L'inversione della matrice delle risposte binaurali si rivela critica per questioni di ordine timbrico, e di stabilità della cancellazione del crosstalk. In questa realizzazione dello stereo dipolo si è scelto:

- di cancellare il crosstalk solo dal fronte diretto, senza operare l'equalizzazione della stanza di riproduzione, contenuta nella coda riverberante della risposta.
- Di imporre la cancellazione del crosstalk in una banda ridotta, da 1000 a 15000, in modo da evitare i problemi dovuti al boosting delle basse frequenze implicito nella cancellazione del crosstalk e gli errori ad alte frequenze, dove la sensibilità della risposta all'impulso a piccoli movimenti è più importante

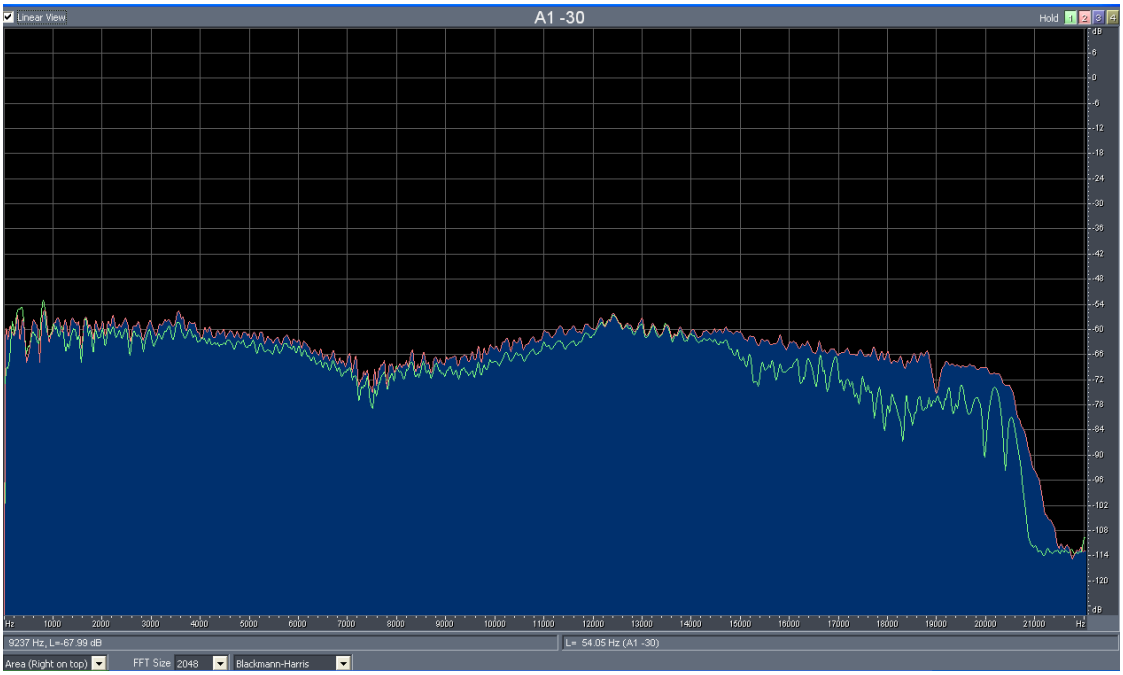

Figura 4 – Funzione di trasferimento diretta non trattata (rossa) e trattata (verde)

- Di non equalizzare la cassa, in modo da lasciare il più possibile inalterato il suono del sistema elettroacustico.

Tale soluzione si è rivelata efficace sia per la corretta riproduzione del timbro e stabile per deboli movimenti dell'ascoltatore.

Lo stereo dipolo è stato montato nella saletta di ascolto della Casa della Musica di Parma [12], e utilizza un paio di casse acustiche Genelec, una scheda M-audio, unitamente al software di processamento real-time AudioMulch e al convolutore Voxengo Pristine Space.

Analizzando la figura 4, che riporta la funzione di trasferimento corrispondente alla risposta all'impulso dalla cassa sinistra all'orecchio sinistro prima e dopo il trattamento si puo' vedere che le due funzioni sono all'incirca le stesse, a parte una leggera attenuazione delle alte frequenze per la risposta trattata, e che puo' attribuirsi al filtro che é comunque applicato anche a tale funzione.

Il risultato della cancellazione del crosstalk è presentato in figura 5. Si nota che il crosstalk è stato ridotto in media di circa 10 dB nella banda utile. Tale attenuazione è ampliamente sufficiente per riprodurre fedelmente sorgenti a banda larga con contenuto frequenziale significativo nella banda utile considerata.

6. Test di ascolto

I test di ascolto sono stati effettuati su 5 persone, musicisti e/o musicologi, alla Casa della Musica di Parma, con il sistema di stereo dipolo precedentemente descritto.

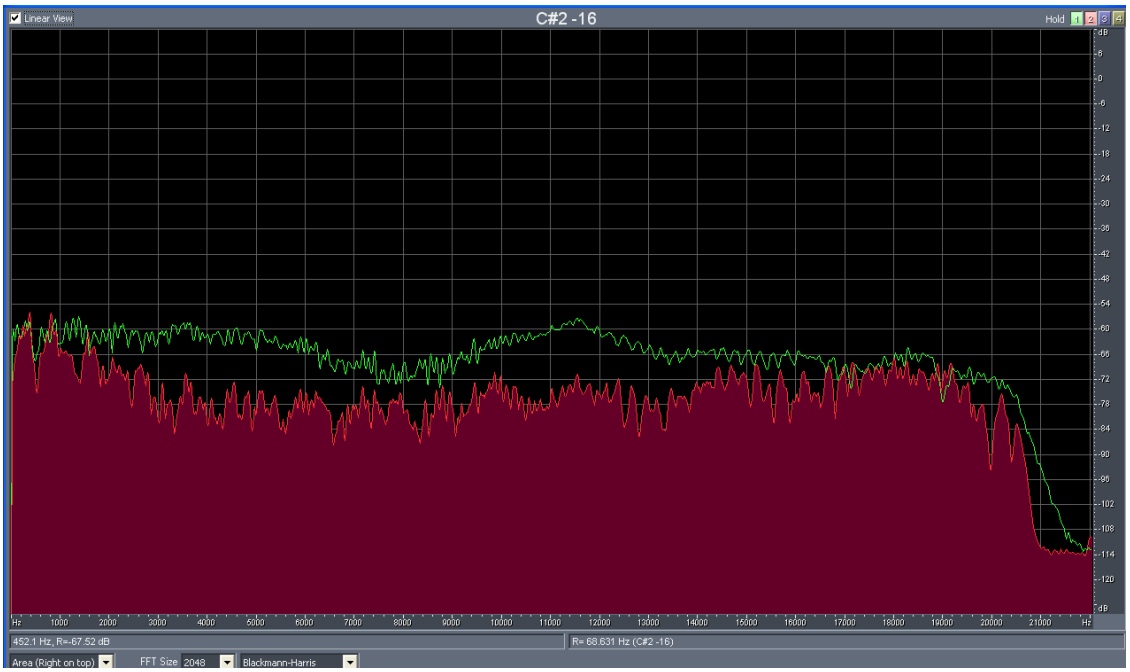

Figura 5 - Funzione di trasferimento del crosstalk non trattata (verde) e trattata (rosso)

Il test si è svolto come segue: la persona viene introdotta nella saletta di ascolto e il test di preselezione é compiuto per stabilire il set di HRTF personalizzato. In seguito sono fatti ascoltare al soggetto due brani anecoici (*Water Music* di *G.F.Handel e My funny valentine*, voce solista) convoluti con la risposta binaurale sintetizzata attraverso le due tecniche e con il set di HRTF prescelto. La risposta binaurale corrisponde a quella simulata a partire dal modello Ramsete della chiesa di San Vitale in Parma con sorgente localizzata sul presbiterio e molto disassata a sinistra e ricettore nel centro della navata. Gli individui possono passare da una versione all'altra in real time, potendo così paragonare le due tecniche in modo più semplice, e possono ascoltare le due versioni a piacimento, interrompendo e riprendendo l'ascolto secondo il loro giudizio.

Ai soggetti è stato chiesto di compilare un questionario cartaceo. I parametri da valutare sono stati formulati tenendo presente [13], e sono visibili in figura 6, insieme ai risultati.

7. Risultati

7.1 Personalizzazione di HRTF

Il file di test é stato ripetutamente presentato ai soggetti, selezionando di volta in volta un individuo di riferimento differente. Sebbene le posizioni laterali e frontale fossero generalmente ben identificate indipendentemente dal set di HRTF utilizzato, si verficavano talvolta fenomeni di rotazione (più che fenomeni di non-localizazzione, o localizzazione intracraniale) nella posizione percepita. Tale fenomeno si attribuisce a un non

allineamento dell'Interaural Time Difference (ITD) tra il soggetto e il set di HRTF di riferiento.

Se la corretta localizzazione delle suddette posizioni avveniva abbastanza di frequente, e i possibili set di HRTF personalizzati erano multipli, lo stesso non si può dire per la localizzazione della sorgente virtuale dietro all'individuo. In effetti la maggior parte dei set testati non garantivano una corretta percezione di questa sorgente virtuale che, veniva generalmente localizzata davanti. Questo effetto può in parte essere dovuto alla non perfetta cancellazione del cross-talk, ma senz'altro anche alla front-back confusion dovuta all'imperfetta corrispondenza tra le HRTF del soggetto e quelle del set di riferimento. In ogni caso tra tutti i set possibili, nei casi osservati si è arrivati, in modo più o meno flagrante, all'identificazione di un set ottimale, ascoltando fino a 40 diversi segnali test, corrispondenti quindi a 40 diversi tipi di orecchie!

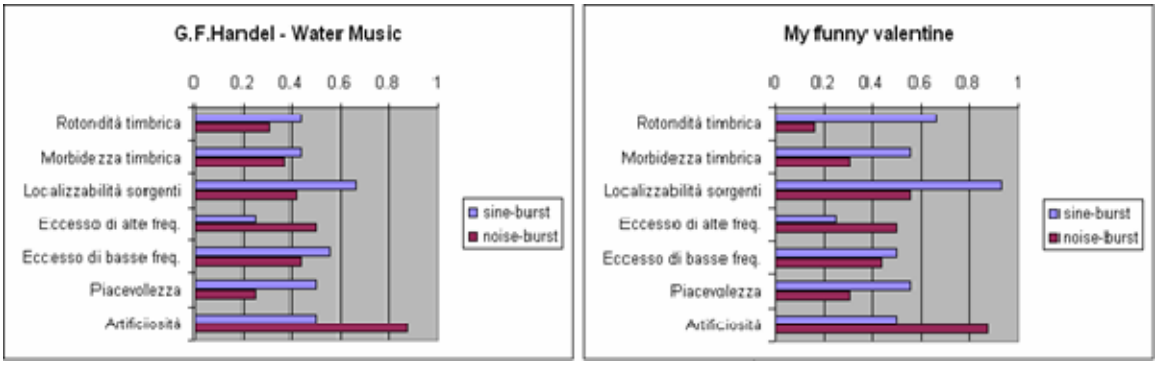

Figura 6 – Risultati del test: '1' è il punteggio massimo.

7.2 Valutazione degli algoritmi di sintesi

Le valutazioni richieste ai partecipanti al test erano volte soprattutto a confrontare la qualità dell'auralizzazione ottenuta con le risposte all'impulso dell'ambiente sintetizzate secondo i due metodi; alcune domande, inoltre, possono offrire utili indicazioni per lo sviluppo futuro del software. In sintesi

- **Rotondità** e **Durezza**: senz'altro uno dei problemi maggiori in cui incorrono i segnali audio sintetizzati è l'assenza di morbidezza e la comparsa di suoni sgradevolmente spigolosi: la tendenza osservata è di una maggior morbidezza rilevata nei segnali auralizzati con la IR *sine-burst* rispetto all'altra, soprattutto nel brano *My funny valentine*. Tuttavia non si può parlare di suoni rotondi e morbidi in nessuno dei due casi.
- **Localizzabilità**: assolutamente migliore la localizzabilità della sorgente nel caso del software presentato in queste pagine, in virtù anche del fatto di aver utilizzato set di HRTF di elevata qualità e di aver compiuto un test preliminare di selezione delle corrette HRTF per ogni individuo. Si evidenzia, in particolare, come nel caso *noise-burst* , talora la localizzazione fosse piuttosto aleatoria.
- **Eccesso di alte frequenze**: i soggetti hanno individuato un eccesso di alte frequenze relativi alla tecnica *sine-burst*, soprattutto *My funny valentine*.
- **Eccesso di basse frequenze**: il risultato non presenta differenze di particolare

rilievo tra le due tecniche.

- **Piacevolezza**: abbastanza rilevante è il divario nel caso del brano *My funny valentine*: l'auralizzazione con IR *sine-burst* è stata ritenuta più gradevole dell'altra. Risultato analogo, ma con differenza meno marcata per l'altro brano.
- **Naturalezza**: pur non potendo affermare di aver ricostruito una percezione completamente naturale, il miglioramento rispetto alla tecnica *noise-burst*, è sensibile e rilevato in entrambi i brani, nonostante le caratteristiche differenti di questi ultimi.

8. Conclusioni

Col presente lavoro si è voluto valutare quanto fosse valida la strada della sintesi di tipo *sine-burst* quando messa a confronto diretto con la tecnica *noise-burst*. Tenendo presente che la scrittura del software è iniziata solo nel dicembre del 2006, ci si può ritenere abbastanza soddisfatti dei risultati raggiunti, i quali, in sostanza, confermano l'intuizione ed incoraggiano il proseguimento dello sviluppo.

9. Bibliografia

[1] http://www.ramsete.com/

[2] A. Farina, "RAMSETE - a new Pyramid Tracer for medium and large scale acoustic problems", Proc. of EURO-NOISE 95 Conference, Lyon 21-23 march 1995

[3] M. Kleiner, B.I. Dalenback, P. Svensson, "Auralization: an overview". Journal of the Audio Engineering Society ,1993.

[4] Møller, Henrik; Sørensen, Michael Friis; Hammershøi, Dorte; Jensen, Clemen Boje, "Head-Related Transfer Functions of Human Subjects", Journal of the Audio Engineering Society ,1993.

[5] http://sound.media.mit.edu/KEMAR.html

[6] http://recherche.ircam.fr/equipes/salles/listen/

[7] Kittiphong Meesawat, Dorte Hammershøi, Aalborg University, Aalborg, Denmark, "The Time When the Reverberation Tail in a Binaural Room EImpulse Response Begins", AES 115 Convention, New York, 2003

[8] S. Fontana, A. Farina, Y. Grenier, "A system for Head Related Impulse Responses Rapid Measurement and Direct Customization",
120th Convention AES, Paris 2006 May 20-23.

[9] Bernhard U. Seeber and Hugo Fastl, "Subjective selection of non-individual transfer functions', Proceedings of the 2003 International Conference on Auditory Display, Boston, MA, USA, July 6-9,

[10] Dmitry N. Zotkin, Ramani Duraiswami, " Rendering localized Spatial Audio in a Virtual Auditory Space", IEEE Transaction on Multimedia, vol. 6, no. 4, August 2004

[11] O Kirkeby, PA. Nelson, H Hamada "The stereo dipole: A virtual source imaging system using two closely spaced loudspeakers", Journal of the Audio Engineering Society, 1998

[12] A. Farina, P. Martignon, A. Azzali, A. Capra "Listening Tests Performed Inside a Virtual Room Acoustic Simulator", I seminario Música Ciênca e Tecnologia "Acústica Musical", São Paolo do Brasil, 3-5 November 2004

[13] A. Farina "Acoustic quality of theatres: correlation between experimental measures and subjective evaluations", Applied Acoustics, Volume 62, Issue 8, Pages 889-1023 (August 2001).