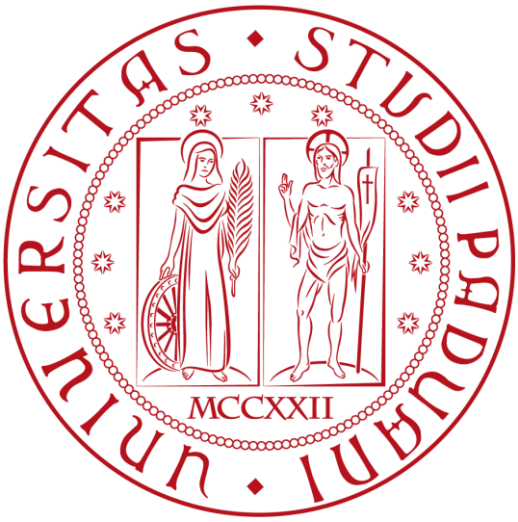
UNIVERSITA' DEGLI STUDI DI PADOVA

FACOLTA' DI INGEGNERIA INFORMATICA



TESI DI LAUREA MAGISTRALE

Sviluppo e sperimentazione di un Auditory Display per la deambulazione di non vedenti

Laureando: **Marabese Daniele**

Relatore: **Antonio Rodà**

Correlatore: **Federico Avanzini**

Correlatrice: **Serena Zanolla**

Data Laurea: 22-10-2012

Anno accademico 2012

ABSTRACT

Le persone affette da problemi alla vista, sia ipovedenti che non vedenti, hanno grosse difficoltà a muoversi in autonomia all'interno degli spazi indoor così come in quelli outdoor, soprattutto in quelli dove non sono soliti deambulare, perciò hanno la necessità di riconoscere gli oggetti circostanti tramite esplorazioni accurate dell'ambiente.
SoundingARM è per l'appunto un'applicazione che consente all'utente di esplorare velocemente uno spazio indoor semplicemente restando in piedi vicino alla porta e compiendo gesture intuitive quali il puntamento col dito, senza la necessità di indossare particolari sensori. SoundingARM permette alla persona non vedente di riconoscere immediatamente il tipo di ambiente in cui si trova, lo rende libero di muoversi in modo sicuro nello spazio circostante poiché gli consente di identificare rapidamente ostacoli e oggetti di interesse.
SoundingARM effettua il tracciamento dell'utente per mezzo del sensore Microsoft Kinect, ed in questo modo risulta essere non invasivo, economico, facile da usare e da installare in tutti gli ambienti. Il suo obiettivo è quello di aiutare i soggetti minorati della vista durante le fasi di esplorazione di un ambiente domestico (salotti, cucine, camere da letto) in modo che possano essere agevolati anche nei contesti sconosciuti come una stanza di hotel e uffici pubblici.
L'auditory display veicolato da SoundingARM fa uso di differenti icone uditive come auditory icons, earcons, speech icons, spearcons opportunamente ottimizzate, e di tecniche di audio 3D quali la spazializzazione binaurale che garantiscono il miglior feedback audio possibile e, nel contempo, un soundscape virtuale dettagliato e ricco di informazioni.
Sono stati inoltre condotti mirati test sperimentali con l'obiettivo di comprendere l'efficacia del sistema in relazione alle differenti categorie di suoni utilizzati e alle abilità dei soggetti non vedenti che ne fanno uso.

INDICE

Capitolo 1

AUDITORY DISPLAY COME AUSILIO PER NON VEDENTI

1.1 INTRODUZIONE

E' ben noto che le persone che hanno problemi di vista, siano essi ipovedenti o non vedenti, hanno sostanziali difficoltà a muoversi negli spazi indoor così come in quelli outdoor, perciò hanno bisogno di riconoscere gli oggetti circostanti tramite esplorazioni accurate dell'ambiente.
In breve, SoundingARM (Acoustic Representation of a Map), ovvero la rappresentazione acustica di una mappa, è un'applicazione capace di fornire in tempi rapidi una mappa acustica di un ambiente, sconosciuto o conosciuto, dando così la possibilità ad un utente non vedente di esplorare la stanza stando in piedi davanti alla porta, semplicemente muovendo le braccia in modo da puntare gli oggetti che gli stanno attorno.

1.1.1 Lavori correlati e Ausili disponibili sul mercato

Il panorama degli strumenti e dei sussidi sviluppati negli ultimi anni per le persone affette da menomazioni all'apparato visivo è molto ampio, e complice il fatto che nessuna azienda del settore è mai stata in grado di imporre il proprio prodotto sulle altre, al giorno d'oggi il mercato pullula di soluzioni concorrenti, ciascuna con i propri vantaggi e svantaggi, e l'utente non vedente deve essere in grado di dotarsi dello strumento che meglio si adatta alle proprie esigenze.
Partendo dalla tipologia cardine circa i software che agevolano l'interazione del soggetto non vedente con i personal computer, si evidenziano i cosiddetti screen-reader, ovvero applicativi in grado di riprodurre tramite sintesi vocale e su display Braille il contenuto di uno schermo: menu, finestre, pulsanti, e tutto ciò che un utente normodotato può vedere sul desktop. Tali software, tra i quali si ricordano: JAWS, Ultra Hall Reader, Window-Eyes, VoiceOver, SAToGO, NDVA, permettono la sostituzione integrale del visual display tramite l'integrazione di auditory display (sintesi vocali) e display tattili (barre Braille), e rendono possibili all'utente una serie di attività quali la videoscrittura, la navigazione web, in modo rapido ed efficace.
Altri applicativi software di rilievo, soprattutto per i soggetti ipovedenti, sono i cosiddetti magnifier, ovvero software che ingrandiscono i caratteri e gli oggetti del desktop con fattori di ingrandimento anche molto sostenuti (da 200% a 10000%), e spesso sono dotati di interessanti features quali sintesi vocali ottimizzate per la lettura di testi, evidenziazione del testo e dei puntatori attivi, impostazioni avanzate di contrasto, allineamento con cursore o mouse. Tra questi si ricordano: ZoomText, Magic, Windows Mangifier, Zoom(Apple OSX/IOS).

Per quanto riguarda l'acquisizione di documenti quali libri, giornali, o contenuti cartacei, esistono moltissimi software applicativi in grado di connettersi con lo scanner, effettuare l'acquisizione e l'OCR in modo da riconoscere i caratteri (ad esempio FineReader), e leggere direttamente quanto acquisito salvandolo anche in formato audio quale .mp3 o simili.
Significativi sono inoltre quei software quali WinGuido, che forniscono all'utente non vedente tutta una serie di informazioni molto utili quali orari e tragitti dei treni, prenotazioni ospedaliere, mail, contatti, calendario, agenda, quotidiani e molto altro ancora.
Quanto trattato sino ad ora fa riferimento ai sistemi con cui i non vedenti si interfacciano con i pc, col materiale didattico di studio, in un contesto sempre e comunque domestico o di ufficio, quindi statico e ben determinato.
In ambito mobile, e quindi smartphone, telefonia, messaggistica, navigazione, lo scenario non è molto differente da quello desktop, infatti si possono trovare software screen reader quali Talks, MobileSpeak, VoiceOver, NokiaTalk che permettono al non vedente di interfacciarsi col terminale, spedire messaggi, navigare in internet, scattare foto, in poche parole il dominio completo anche sul dispositivo mobile, sempre grazie all'integrazione di sintesi vocali che danno un feedback sonoro continuo circa l'area del display selezionata.
In generale, uno degli scenari di maggiore difficoltà per il non vedente è la deambulazione sia in ambiente indoor che outdoor. Per quanto riguarda gli scenari outdoor si possono trovare tutta una serie di sussidi basati su GPS come le mappe TeleAtlas, Talking Signs, che grazie al riconoscimento della posizione, riescono a guidare il soggetto fino al raggiungimento della destinazione desiderata, tenendo anche in considerazione parametri quali mezzi pubblici accessibili, strade pericolose da evitare e altro ancora.
Un altro utilissimo ausilio è AuxDeco, un tool sviluppato dalla Tokyo University in collaborazione con Eye PlusPlus Co che aiuta i non vedenti nella deambulazione indoor e outdoor. Questo dispositivo permette agli utenti di percepire con la fronte ciò che li circonda, per esempio una linea bianca continua, un gruppo di strisce pedonali lungo il percorso, una superfice lontana all'orizzonte, una persona o un'auto che sta sopraggiungendo nella stessa direzione del movimento. Questo device consente agli utenti di passeggiare in modo sicuro per le vie delle città, poiché fornisce feedback tattili di gran lunga più accurati e dettagliati rispetto a quelli che si hanno con il bastone. AuxDeco fa uso del senso del tatto presente sulla fronte, un po' come accade con il linguaggio Braille che sfrutta il tocco e la sensibilità delle dita.
Per quanto concerne scenari indoor, i sistemi di deambulazione di solito fanno uso di sensori a infrarossi, ultrasuoni, radio frequency identifier (RFID) o complicati sistemi di computer vision che permettono il riconoscimento degli oggetti e il tracciamento dell'individuo che si muove nelle stanze.

Tipologia	Ambito	Descrizione	Esempi
Screen-Reader	Produttività personale / lavoro / studio	Riproducono tramite sintesi vocale e su display Braille il contenuto di uno schermo: menu, finestre e pulsanti.	JAWS, Ultra Hal Text-to-Speech Reader, Window-Eyes, VoiceOver, SAToGO, NDVA
Magnifier	Produttività personale / lavoro / studio	Ingrandiscono i caratteri e gli oggetti del desktop (200%-10000%), dotati di sintesi vocali e impostazioni avanzate di contrasto, allineamento con cursore o mouse.	ZoomText, Magic, Windows Mangifier, Zoom(Apple OSX/IOS)
OCR / Text-Reader	Produttività personale / lavoro / studio	Acquisiscono tramite scanner documenti di testo, libri, giornali, effettuano il riconoscimento dei caratteri e riproducono il contenuto tramite sintesi vocale	ABBY FineReader, Nuance OmniPage Pro, ReadIris Pro
Mobile Apps	Telefonia / messaggistica / connettività	Screen Reader mobile che permettono di controllare smartphone e tablet per telefonare, messaggi posta web	Nuance Talks, MobileSpeak, VoiceOver, NokiaTalk
Navigazione GPS-based	Deambulazione outdoor	Guidano a destinazione l'utente tramite la posizione corrente e le mappe GPS	TeleAtlas, Talking Signs, AuxDeco
Navigazione Sensor-based	Deambulazione indoor	infrarossi, ultrasuoni, radio frequency identifier (RFID), computer-vision	AuxDeco

Figura 1 tabella riassuntiva degli ausili per non vedenti

1.1.2 SoundingARM

Al contrario dei sistemi sopra descritti, SoundingARM è un'applicazione che consente all'utente di esplorare velocemente uno spazio indoor, semplicemente restando in piedi vicino alla porta, e compiendo gesture intuitive quali il puntamento col dito, senza la necessità di indossare particolari sensori.

SoundingARM permette alla persona non vedente di riconoscere immediatamente il tipo di ambiente indoor in cui si trova, gli consente di muoversi in modo sicuro nello spazio circostante e di identificare in modo rapido uno specifico oggetto di interesse.

SoundingARM è stato sviluppato utilizzando il sensore Microsoft Kinect, ed in questo modo risulta essere non invasivo, economico, facile da usare e da installare in un ambiente indoor. Il suo obiettivo è quello di aiutare i soggetti non vedenti e ipovedenti durante le fasi di esplorazione di un ambiente domestico, per esempio i salotti, le cucine, le camere da letto, in modo che possano essere agevolati anche nei contesti sconosciuti come una stanza di hotel.

Le persone non vedenti o coloro che hanno grossi problemi di vista, devono concentrare l'attenzione sugli altri sensi per ottenere informazioni circa l'ambiente che li circonda, in particolar modo i ciechi assoluti dipendono esclusivamente dal tatto e dall'udito. In tale contesto, l'udito è il senso dominante che guida la percezione e l'azione.

La compensazione sonora consiste in una riorganizzazione funzionale e in una re-allocazione degli apparati percettivi della corteccia cerebrale in modo che le

funzioni acustiche e tattili migliorino le loro performance in un soggetto non vedente, complice uno sviluppo maggiore rispetto ad un individuo normodotato.
Le persone non vedenti dalla nascita hanno maggior esperienza nell'ascoltare e nell'interpretare l'ambiente tramite informazioni sonore rispetto alle persone che hanno perso la vista in seguito a un incidente o comunque in età adulta, questo perché sin dalla tenera età sono stati abituati a sostituire la vista con l'udito e l'integrazione degli altri sensi come il tatto e l'olfatto. L'udito, del resto, è uno dei sensi che consente differenti livelli di intensità della percezione, ed il range va dai suoni di sottofondo al parlato quotidiano.
SoundingARM è in grado di offrire una mappa acustica essenziale della stanza; l'obiettivo dell'utente potrebbe essere quello di cercare uno specifico oggetto, oppure apprendere quali altri oggetti caratterizzano l'ambiente circostante. Di solito, due sono i tipi di ricerca usati quando si esplora un'area:

- ✓ l'esplorazione lungo il perimetro della stanza, che fornisce informazioni circa la dimensione, il profilo dell'area e l'eventuale presenza di ostacoli o aperture laterali;
- ✓ l'esplorazione tramite una serie di movimenti in linea retta da destra a sinistra o dal basso all'alto alternando direzioni opposte in successione, che ricorda la navigazione effettuata con il bastone al fine di individuare ostacoli lungo il cammino (quest'ultima è simile all'approccio usato dai sonar o dai pipistrelli).

A livello pratico, quando l'utente punta il braccio per trovare un oggetto, il sistema risponde tramite un feedback acustico caratteristico dell'oggetto indicato, sotto forma per esempio di auditory icons. Uno dei punti di forza di questo approccio è che l'utente ottiene l'informazione della posizione dell'oggetto tramite la propria gesture, infatti l'oggetto è posizionato esattamente nella direzione lungo la quale il dito sta puntando.
Nella prima implementazione, il feedback audio era costituito dalla descrizione dell'oggetto veicolata per mezzo di una sintesi vocale che pronunciava il nome dell'oggetto puntato.
Nelle implementazioni successive, la voce sintetizzata è stata sostituita con auditory icons, in modo da fornire una descrizione più completa e intuitiva; inoltre è stata realizzata anche una versione del software che fa uso della spazializzazione binaurale in modo da veicolare una vera e propria ricostruzione tridimensionale della stanza tramite auditory display.

1.2 AUDITORY DISPLAY

Molte delle applicazioni di virtual reality fanno uso di display visuali, tattili, sonori e spaziali. L'interazione tra diversi sensi è fondamentale per dare la sensazione di un mondo reale ed immersivo, in grado di mettere a disposizione un'interazione completa con l'utente. Il numero di caratteristiche veicolabili tramite l'interazione sensoriale-visiva-tattile e acustica è notevolmente maggiore rispetto a quella trasmessa con la sola adozione di un sistema di visione ad altissima risoluzione.
Nonostante la maggior parte delle informazioni siamo riconducibili a display visivi, ricerche di Gaver ([1] e [2]) hanno dimostrato che il feedback acustico veicola un importante sottoinsieme di caratteristiche quali forma, dimensione, consistenza, resistenza, rigidità, concavità che sono legate alle vibrazioni emesse dagli oggetti che ci circondano e dalle relazioni spaziali che intercorrono tra di essi, per esempio contatto per impatto, caduta, strofinio, schiacciamento.
I recenti studi riguardanti la sintesi di suoni tramite ricostruzione dei modelli fisici correlati alle azioni dei soggetti nel mondo che li circonda, consentono di fornire svariate informazioni ulteriori in grado di descrivere in modo completo la scena tramite feedback acustico.

1.2.1 Approccio Ecologico

Echological ([3]) riflette due temi importanti:

- in primo luogo la percezione è una concreta realizzazione di un unico "sistema animale–ambiente", non semplicemente esseri viventi dotati di cervello, ma tutto quello che fa parte dell'ambiente è parte integrante della percezione.
- In secondo luogo l'obiettivo principale della percezione è guidare l'azione, e ciò implica che non possono essere trascurate le azioni degli stessi esseri viventi, e l'ambiente in cui le compiono.

Percezione diretta vs indiretta

Lo statement fondamentale è il seguente: ***"la percezione è diretta".***
Tuttavia, considerando per esempio la retina (per la vista), o la coclea (per l'udito), la percezione è strettamente legata a come gli apparati sensoriali recepiscono e trasferiscono le informazioni; nel caso del moto, non c'è sempre una corrispondenza diretta tra quanto viene percepito a livello sensoriale e quello che in realtà sta avvenendo fisicamente in un preciso momento.
La percezione sensoriale è dunque solo una percezione indiretta della realtà fisica.
Il ruolo della percezione è allora quello di fissare un certo input, e aggiungere interpretazioni significative, cosicché il cervello sia in grado di inferire cos'ha causato tale input e agire di conseguenza.
L'accuratezza, di conseguenza, è correlata direttamente alla capacità dell'individuo di riempire il gap tra l'azione percepita a livello sensoriale e

l'azione che avviene realmente a livello fisico; tale capacità richiede sostanzialmente un "***processamento inferenziale cognitivo***".
In contrasto con quanto appena espresso, la teoria della percezione diretta afferma che esiste una corrispondenza 1 a 1 tra i pattern sensoriali e gli aspetti della realtà fisica (Gibson [4]); quindi la realtà è pienamente specificata dalle stimolazioni sensoriali.
Una conseguenza importante è che qualsiasi cosa che può essere percepita, può anche essere misurata a livello fisico.

"Ascolto di ogni giorno" vs "Ascolto musicale"
Gaver ([1] e [2]) ha introdotto il concetto di "***suono di ogni giorno***" in contrapposizione con l'"ascolto musicale": un individuo, quando ascolta un suono, si concentra sul pitch, sulla rumorosità, sul timbro, e sulle relative variazioni nel tempo, tutti aspetti che hanno a che fare con il suono stesso; d'altro canto l'ascoltatore si può concentrare anche sulla sorgente del suono, per esempio un'auto che sopraggiunge alle spalle, quest'ultimo caso coincide con l'ascolto di eventi e non di suoni, questi hanno a che fare con il processo di produzione e con l'ambiente, e non con il suono di per sé.
L'approccio classico considera i suoni solamente in relazione alla frequenza, all'ampiezza, alla fase, e non prende in considerazione minimamente i contenuti informativi di livello superiore, quali la correlazione spaziale, ambientale e le connessioni con gli eventi.
L'"everyday listening" ha bisogno di un apposito framework in grado di tenere sotto controllo due differenti aspetti:

- in primo luogo gli attributi legati alla percezione, per esempio le caratteristiche rilevanti dell'evento che si possono discernere con l'ascolto, quindi "cosa sentiamo?".
- In secondo luogo è necessaria un'"ecological acoustic" che descrive le proprietà acustiche del suono, correlate alla sorgente sonora che lo ha generato e quindi "come lo sentiamo?".

Invarianti acustici
Qualsiasi sorgente sonora è legata alle interazioni con i materiali, per esempio parte dell'energia prodotta dal motore di un'auto produce vibrazioni che generano a loro volta un'onda di pressione acustica nell'aria circostante, che si muove insieme al veicolo, e dalla quale un ascoltatore è in grado di ricavare informazioni. In generale, i pattern di vibrazione prodotti dal contatto di materiali dipendono dalle forze di contatto, dalla durata del contatto, dalle variazioni temporali dell'interazione, dalla forma, dimensione, superficie e consistenza degli oggetti coinvolti.
Il suono da informazioni anche sull'ambiente nel quale è stato prodotto. L'orecchio di un ascoltatore è raggiunto anche da suoni riflessi e non solo da quelli diretti, così si verificano numerose sovrapposizioni e sfumature dello spettro sonoro. Anche il mezzo attraverso il quale avviene la propagazione del suono è molto importante perché influenza la dispersione e la dissipazione di energia e trasmette informazioni circa la distanza della sorgente (ad esempio l'effetto doopler prodotto quando sorgente sonora e ascoltatore si trovano in moto reciproco l'uno rispetto all'altro).

Molti invarianti acustici possono essere associati agli eventi sonori, in particolare molti attributi di un solido che vibra, comprese forma, dimensione e densità determinano la frequenza del suono prodotto.

Mappe dei suoni quotidiani

- Suoni generati da oggetti solidi, legati alle vibrazioni emesse dalle interazioni con l'ambiente, all'insieme di caratteristiche e attributi geometrici e fisici dell'oggetto stesso.
- Suoni aerodinamici causati dall'introduzione/modifica della differenza di pressione atmosferica, ad esempio un'esplosione, oppure una ventola che gira, o il vento che passa tramite dei fili.
- Suoni legati all'interazione con i liquidi, simili a quelli dei solidi, però con la differenza che le vibrazioni nei liquidi non sono udibili, e i suoni sono generati dalle cavità risonanti come le bolle che oscillano sulla superfice del liquido.
- Altri suoni sono legati alle ripetizioni successive, per esempio il rumore dei passi, e sono caratterizzati da strutture complesse che coinvolgono vincoli temporali e interazioni ben precise tra gli oggetti.
  - Basic level source: descrivono eventi riguardanti i solidi come deformazione, impatto, raschiatura, rotolamento;
  - Patterned source: che coinvolgono pattern di eventi temporali come la camminata, gli impatti;
  - Eventi composti: coinvolgono più eventi di base, ad esempio una porta che sbatte;
  - Eventi ibridi: inducono un ulteriore livello di complessità legato ai materiali coinvolti, ad esempio la caduta dell'acqua su una superfice che genera riverbero.

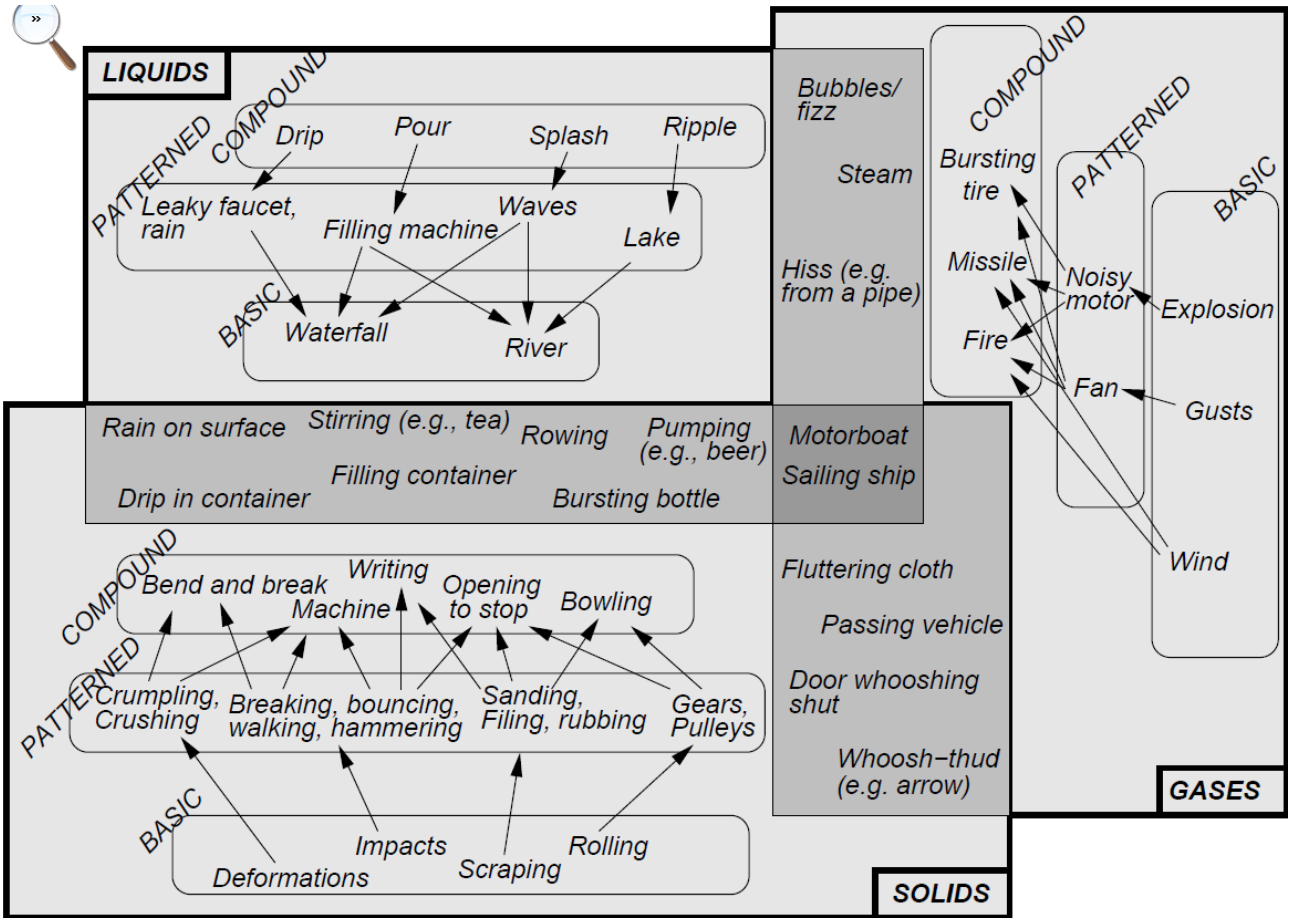


Figura 2 calssificazione ecologica dei suoni (Gaver)

Basic level source

I materiali possono essere caratterizzati usando un coefficiente di attrito interno, che misura l'anelasticità (es. acciaio-vetro, legno-gomma), la misura utilizza sia il fattore di qualità ***Q*** che il tempo di decadimento ***te*** della vibrazione: anelasticità decrescente corrisponde a un aumento di *Q* e *te*.

Altri studi condotti su barre di materiali differenti che vengono fatte vibrare a diverse frequenze, hanno evidenziato come sia la frequenza a cui vibrano il fattore che maggiormente discrimina un materiale dall'altro, mentre alle orecchie di un ascoltatore, i fattori come il tempo di decadimento e l'ampiezza passano decisamente in secondo piano.

Un altro aspetto significativo è la durezza con cui avvengono le collisioni, non è tanto la dimensione dell'oggetto percosso che influenza il suono, bensì le caratteristiche del percussore che lo genera e la forza di contatto impressa.

Quando si considerano contatti continui come quelli che intercorrono tra due oggetti che scivolano l'uno sull'altro, o il raschiamento di un oggetto, l'aspetto più rilevante è costituito dalle asperità della superficie di contatto. Le asperità sono percepite maggiormente utilizzando il tatto, infatti quando la pelle è a contatto con la superfice non si generano suoni, invece se si utilizza una sonda rigida per tastare la superficie, questa genera vibrazioni che consentono un'integrazione tra percezioni tattili e uditive.

Un altro ambito di interesse è la percezione delle caratteristiche geometriche degli oggetti tramite feedback acustico.

Patterned e compound source
Suoni come rimbalzi, rotture o "camminate di persone" sono composti e pervadono la vita quotidiana. Possono essere classificati in:
invarianti strutturali che specificano le proprietà degli oggetti, e
invarianti trasformazionali che specificano le interazioni e i cambiamenti.
La natura di questi invarianti è essenzialmente legata ad aspetti temporali.
Studi condotti su una moltitudine di categorie e tipologie di suoni, in particolare quelli non verbali, provenienti dal mondo reale, hanno evidenziato che il riconoscimento e l'identificazione degli stessi da parte di un ascoltatore dipendono in gran parte dall'esperienza individuale.
Tre sono le variabili che maggiormente influenzano l'identificazione:

- numero di picchi correlati,
- rapporto tra durata degli scoppi e durata complessiva,
- correlazione cross-channel;

questi sono tutti aspetti temporali che riflettono la periodicità, numero di silenzi e la coerenza dell'inviluppo tra i canali.

1.2.2 Percezione Multimodale

Gli esseri umani riescono ad avere una percezione molto buona tramite la combinazione e l'integrazione di più informazioni provenienti da differenti apparati sensoriali; alcuni studi ([3]) tendono ad affermare che la specializzazione sensoriale della percezione si manifesti gradualmente, mentre altri dicono che sin dal principio i vari sensi cooperino; l'unica certezza è che solo con l'esperienza l'essere umano riesce a combinare nel migliore dei modi le differenti percezioni.

Combinazione e Integrazione
La combinazione tende a massimizzare il numero di informazioni utili raccolte dagli apparati sensoriali mentre l'integrazione si concentra sul ridurre al minimo la variazione e massimizzare l'affidabilità della percezione complessiva.
La ***combinazione sensoriale*** riguarda i segnali della percezione che non sono ridondanti, oppure che si riferiscono a differenti coordinate spaziali, che sono quindi complementari e disgiunti: se un solo segnale non è sufficiente, si utilizza una percezione multimodale per avere una stima migliore della realtà che ci circonda.
Al contrario, l'***integrazione*** descrive l'interazione tra segnali ridondanti, per esempio quando si sente battere alla porta avviene l'integrazione tra più percezioni: acustica, visiva e propriocettiva (la **propriocezione** è la capacità di percepire e riconoscere la posizione del proprio corpo nello spazio e lo stato di contrazione dei propri muscoli, anche senza il supporto della vista).
Ci sono situazioni in cui è la vista che domina la percezione, mentre altre in cui prevalgono altre percezioni sensoriali: per quanto riguarda la localizzazione, la vista domina perché fornisce una descrizione più accurata rispetto all'udito.
Non è la modalità stessa della percezione o lo stimolo che dominano, la dominanza è determinata dalla stima e dall'affidabilità della stessa in relazione

ad una specifica modalità e stimolo ricevuto, quindi è più conveniente parlare di precisione della stima.

"Auditory capture" e illusioni

Molti studi sono stati fatti da psicologi per capire come avviene l'interazione sensoriale, in particolare, significativi sono gli esempi dei ventriloqui e gli esperimenti di McGurk ([5]) in cui quanto si vede è diverso da quello che si sente, e ciò porta a una percezione che può variare, non tanto nello spazio, bensì nel tempo. La finestra di sincronizzazione tra audio e video è fondamentale e non può superare una certa soglia, es. i 300ms nel caso del ventriloquo o dell'effetto McGurk.

In molti casi la sincronia permette di ottenere effetti particolari nella percezione, come testimoniano recenti esperimenti ([3]) in cui si fanno susseguire impulsi flash e impulsi sonori "beep", in particolare si manda prima un flash seguito da due beep, e in questo modo la percezione dell'essere umano risulta alterata in quanto vengono recepiti correttamente i 2 beep ma, a causa della sincronia, si genera l'illusione di 2 flash quando in realtà ne viene proiettato solo uno. Questo effetto è chiamato "***cattura audio***".

Altri studi che mettono in relazione suoni e luci che si ripetono nel tempo, hanno mostrato come l'alternanza di flash e beep porti ad un netto miglioramento nella qualità della percezione, mentre un miscuglio di essi, senza una netta alternanza temporale, porti ad un sostanziale degradamento della percezione.

Questi effetti di cattura o integrazione non sono limitati solo all'udito/vista, ma coinvolgono qualsiasi altro apparato di percezione. In particolare, studi recenti ([3]) hanno dimostrato come la percezione sia fortemente condizionata dalla presenza (anche secondaria) di un senso come la vista; esperimenti tattili-sonori svolti su persone vedenti, ipovedenti e non vedenti hanno mostrato come l'accuratezza nella percezione sia tanto maggiore, tanto minore è l'interferenza dovuta alle percezioni irrilevanti, come in questo caso la vista.

La finestra temporale circa l'integrazione audio-tattile è sensibilmente maggiore rispetto a quella audio-video.

Maggiore è la forza e l'affidabilità del segnale, meno influenzabile dovrebbe essere la percezione, inoltre i segnali sonori possono influenzare sia le percezioni tattili che quelle visive e questo indica che quando si contano il numero di eventi in sequenza, i segnali sonori sono più affidabili sia di quelli visivi che di quelli tattili.

Per quanto concerne la percezione tattile delle superfici, l'aspetto sonoro mediamente ha una scarsa rilevanza, tuttavia, quando la superficie viene esplorata tramite una sonda rigida e non più con le dita, l'effetto di contatto e i relativi suoni generati diventano più significativi, anche in relazione al fatto che la sensibilità tattile diminuisce nettamente, e ciò indica che, nonostante il tatto sia il senso dominante in questo contesto, la rilevanza dell'udito può essere modulata e l'integrazione diviene fondamentale.

Altri studi ([3]) riguardanti l'interazione tra suono e percezione tattile hanno mostrato come l'attenuazione delle alte frequenze induce una percezione maggiormente levigata della superfice, al contrario, un boost sonoro provoca un aumento di ruvidità della medesima superfice.

1.2.3 Percezione in movimento

In accordo all'approccio classico, la percezione è un "processo nel cervello" nel quale i cosiddetti "sistemi di percezione" costruiscono una rappresentazione interna del mondo e le azioni intraprese seguono come una funzione subordinata. Questo punto di vista usa l'assunzione che la percezione è l'input dal mondo alla mente dell'individuo, e l'azione è l'output dalla mente al mondo esterno; la cognizione è il processo che trasforma input in output. La percezione e l'azione sono strumentalmente correlate l'una all'altra da una relazione che coinvolge il pensiero.

La visione ecologica rifiuta la prima assunzione, ma è d'accordo con la visione strumentale, mentre altre teorie più recenti rifiutano entrambe le assunzioni perché sostengono sia impossibile disassociare percezione e azione in modo schematico, inoltre la percezione non è un processo nel cervello, bensì un'attività abile che fa parte dell'intero "sistema animale": infatti solo le creature dotate di opportune abilità corporali possono essere considerate percettori.

O'Regan e No**ë** ([6]) hanno presentato una concezione "***enattiva***" dell'esperienza, che è qualcosa che non accade all'interno dell'individuo, ma piuttosto è qualcosa che l'individuo "enagisce" quando esplora l'ambiente che lo circonda, in questo modo il soggetto degli stati mentali è incorporato, l'individuo è in stretta relazione con l'ambiente, ambiente e essere vivente sono accoppiati e s'influenzano reciprocamente. La percezione è pensata come un'attività che fa parte dell'individuo animale.

"Embodied" è usato per sottolineare due aspetti:

- in primo luogo la cognizione dipende dai differenti tipi di esperienze generate dalla capacità sensoriali;
- in secondo luogo queste capacità sensoriali sono loro stesse comprese in un ben determinato contesto psicologico, biologico e culturale. In sostanza, i processi sensoriali, motori, percezione e azione sono indivisibili per la cognizione.

Le contingenze sensoriali-motorie sono sostanzialmente le leggi che descrivono le interazioni, per esempio un individuo può dire che una spugna è soffice, sulla base di una moltitudine di pattern di contatto, premendo, strofinando, schiacciando la medesima spugna. Quando un individuo sa in modo implicito che sta osservando le contingenze associate alla morbidezza, egli è nel processo di sperimentazione della morbidezza.

O'Regan e No**ë** ([6]) hanno classificato gli input sensoriali secondo due criteri: "capacità corporali" e "capacità di allerta". Le prime riguardano l'attivazione della nervatura che dipende dai movimenti del corpo: la retina, la coclea, i meccanorecettori della pelle, i proprio-percettori possiedono capacità corporali perché qualsiasi movimento del corpo induce dei cambiamenti di posizione, che modificano i segnali percepiti. Le capacità corporali sono un importantissimo fattore che è in grado di indurre la distinzione tra esperienza sensoriale e non-sensoriale, come il pensiero e la memoria.

La capacità di allerta di un input sensoriale può causare comportamenti automatici in grado di catturare risorse di processamento cognitivo. Vista, udito, tatto e olfatto hanno quindi entrambe le proprietà sopra descritte.

Una delle principali obiezioni a questa teoria è che alcune percezioni avvengono anche senza alcuna forma di attività di esplorazione, però è pur sempre vero che la percezione avviene in modo non istantaneo, bensì coinvolge un intervallo temporale nel quale si presume che avvenga, in ogni caso, una qualche attività di processamento.
Un osservatore nota gli aspetti di una scena che in quel momento sono visivamente manipolati, ed è ragionevole che solo un sottoinsieme di elementi della scena (situati in una precisa posizione) possano essere percepiti in un dato momento.
Sempre in riferimento alla vista, Noë ([7]) introduce il concetto di "***cecità esperienziale***", da non confondere con la cecità dovuta a danneggiamenti o alterazioni dell'apparato visivo come cataratta o retinopatie, bensì una cecità dovuta all'inabilità della persona di integrare stimoli sensoriali con pattern di movimento e pensiero. A testimonianza di ciò basti pensare alle persone che si sottopongono ad interventi per la rimozione delle cataratte: subito dopo l'operazione, il paziente è consapevole di avere una percezione perché riceve molti impulsi sensoriali, però è comunque cieco perché non è in grado di vedere. Un altro esempio significativo è legato alla cecità dovuta a paralisi dei muscoli che muovono gli occhi: chi è affetto da questa patologia non è in grado di vedere, nonostante l'apparato visivo funzioni correttamente, e questo perché l'occhio non è in grado di posizionarsi correttamente, e il fenomeno della persistenza retinica impedisce la corretta visione. Fenomeni di questo tipo sono anche legati al concetto di fatica, correlato con tutti gli apparati e organi sensoriali del nostro corpo, si pensi al contatto continuo dei vestiti con la pelle, alla presenza di un anello al dito e così via.

1.2.4 Sostituzione sensoriale

La qualità di una modalità sensoriale non deriva dal particolare sensore in input o dall'attività neurale coinvolta, ma dalle leggi di abilità sensoriale-motoria esercitate. La differenza principale tra udire e vedere sta principalmente nel fatto che una persona vede se c'è un grande cambiamento nel segnale percepito in input; al contrario, una persona sente quando non succede niente mentre guarda, ma c'è una differenza sinistra-destra quando muove la testa. Questa linea di pensiero sottintende che sia possibile ottenere un'esperienza visiva da un input tattile o acustico, ovviamente seguendo le leggi sensoriali-motorie associate alla visione.
Il fenomeno della "***sostituzione sensoriale***" è per l'appunto coerente con questa visione. Bach-y-Rita ([8]) è stato il primo studioso che si è interessato all'argomento ed ha sviluppato uno strumento in grado di permettere ai non vedenti di vedere tramite percezioni tattili, in particolare matrici di elementi vibranti e di stimolatori elettro-cutanei in grado di rappresentare il grado e la distribuzione di luminanza catturata da una telecamera sulla pelle. Da notare che, nonostante i problemi dovuti alla bassa risoluzione, le persone non vedenti che provavano lo strumento riuscivano a percepire gli oggetti solamente quando potevano manipolare la posizione della telecamera, e non quando quest'ultima rimaneva fissa. L'idea che traspare, è che l'esperienza associata con una modalità sensoriale non è collegata direttamente all'hardware neurale, ma è piuttosto un esercizio dell'attività sensoriale

motoria: la visione è costituita dalla capacità di modificare attivamente le impressioni sensoriali in accordo con qualche legge.
Altri studi inerenti la sostituzione sonora sono stati realizzati tramite un particolare device che emette onde a ultrasuoni lungo differenti direzioni, molto simile al sistema di localizzazione dei pipistrelli, in grado di identificare gli ostacoli. L'idea è generare l'immagine dell'ambiente sulla base della direzione, distanza, e riflessione delle onde sonore che vengono riportate tramite un fattore di scala nelle frequenze udibili dagli umani. Nonostante non sia possibile realizzare una completa immagine virtuale della scena, i non vedenti per mezzo di questo sistema hanno la chiara percezione delle cose che gli stanno davanti, un po' come quando utilizzano il bastone per conoscere lo spazio circostante.
O'Regan e Noë ([6]) affermano che le sensazioni non hanno una locazione precisa e la posizione è un'astrazione costruita per tenere in considerazione l'invarianza delle contingenze sensoriali-motorie.
Meijer ([9]) ha condotto studi correlati alla trasformazione video-audio, in particolare, egli realizzò uno strumento in grado di campionare uno stream video ad un certo rate, e di convertirlo in uno spettrogramma in cui i livelli di grigio dell'immagine corrispondono a un'amplificazione parziale, fornendo così una rappresentazione dell'intera scena, e non solamente degli ostacoli. Nonostante questo tipo di percezione sia complicata, l'idea chiave sta ancora una volta nella possibilità che ha l'utente di manipolare direttamente il device per conoscere il mondo.

Bisogno di feedback multisensoriali
Tutti i moderni sistemi di simulazione, intrattenimento, virtual reality che si possono trovare dall'analisi medica, ai videogiochi, alle simulazioni di fluidi e solidi, fanno spiccato uso di una componente visiva (grafica di altissimo livello), mentre sono fortemente limitati sotto l'aspetto tattile (eccezion fatta per sistemi di tracciamento), e di sistemi acustici che si limitano alla distribuzione spaziale degli emettitori. Poter toccare, manipolare e ascoltare oltre che vedere, riesce a trasmettere un grado di immersione che altrimenti non sarebbe raggiungibile.
Secondo Hahn ([10]) il problema di generare i suoni in un ambiente virtuale è suddivisibile in tre sottocategorie:

- modellazione del suono,
- sincronizzazione del suono,
- rendering del suono.

Nel primo problema, i parametri associati col movimento, possono essere mappati direttamente con i parametri di controllo del suono, risultando in un'effettiva sincronizzazione tra i display visivi e acustici. Il sound rendering si riferisce al processo che genera segnali sonori a partire dai modelli degli oggetti e i loro movimenti nell'ambiente, sostanzialmente equivalente al processo di generazione delle immagini a partire dai modelli geometrici. L'energia sonora emessa deve essere tracciata in relazione all'ambiente, e il processamento del segnale sonoro potrebbe essere richiesto per considerare gli effetti prodotti al ricevitore.

“Global array”

La percezione diretta presuppone che non vi sia differenza tra quello che viene percepito e il mondo reale, però è innegabile che vi siano situazioni in cui qualcosa appare in movimento quando invece è fermo, dipende tutto dal sistema di riferimento adottato; infatti lo scenario osservato da una persona che sta ferma e una che viaggia in treno è sicuramente differente perché influenzato dalla dinamica relativa.

La ricerca in questo ambito mira a scoprire dei pattern detti “invarianti” che sono particolarmente rilevanti nel flusso ottico per la percezione e le conseguenti azioni intraprese dagli individui nel rispettivo ambiente. Considerazioni analoghe sono state fatte anche nei riguardi degli altri sensi, in particolare, si parla di “***global array***”, e le forme di energia come il flusso visivo o audio, sono componenti subordinate di un’entità di livello più alto.

Il concetto di global array introdotto da Stoffregen e Bardy ([11]) fornisce informazioni che possono ottimizzare la percezione e le performance, infatti gli umani possono riconoscere dei pattern global array e continuamente possono usare queste informazioni per la percezione ed il controllo, sia negli ambienti virtuali che nella vita quotidiana.

Per costruire il global array, gli sviluppatori devono capire quali global array già esistono e come sono strutturati, a partire dall’individuazione dei parametri principali legati alla fisica e alla generazione dei segnali stessi, in modo da poterli riprodurre in laboratorio. È quasi sempre cruciale supportare l’analisi acustica con una fisica approfondita dell’evento, perché fornisce sia informazioni rilevanti circa gli attributi sonori della sorgente, che sul suono stesso.

1.3 NEURONI SPECCHIO E PERCEZIONE

Alcune recenti scoperte in ambito neuro-scientifico hanno permesso di fornire una spiegazione in termini fisiologici e neouro-biochimici a molti aspetti della cognizione umana, quali la percezione, la comprensione degli atti e delle emozioni altrui, la capacità imitativa, le forme di comunicazione gestuali o vocali.
Di particolare rilievo nel panorama neuro-scientifico è stata la rivoluzionaria scoperta di due classi di neuroni visuo-motori, denominati rispettivamente: neuroni "canonici" e neuroni "specchio", che si contraddistinguono per la capacità di associare proprietà di carattere sensoriale (prevalentemente visivo) ad altre di carattere motorio. Tale scoperta ha contribuito a mettere in discussione la tradizionale concezione logico-astratta dei processi mentali e cognitivi, la quale prevedeva la netta separazione tra le aree sensoriali del cervello preposte alla codifica degli stimoli sensoriali visivi, somato-sensitivi e uditivi, e le aree motorie preposte all'organizzazione dei movimenti.
Il funzionamento di questi neuroni che codificano in un formato comune sia informazioni sensoriali che motorie, non consente di etichettare queste aree come "sensitive pure" e "motorie pure", bensì suggerisce l'esistenza di una forte correlazione tra il sistema motorio, la percezione e i processi cognitivi.
È stato appurato che la funzione primaria di questi neuroni è quella di rendere possibile il riconoscimento e quindi anche la comprensione del significato degli atti altrui, dove per atto si intende un insieme di movimenti finalizzati al compimento di uno scopo.
In linea con quanto affermano Giacomo Rizzolatti e Corrado Sinigaglia ([12]), protagonisti della scoperta dei neuroni specchio e autori del saggio "So quel che fai. Il cervello che agisce e i neuroni specchio":

> *È in questi atti, in quanto atti e non meri movimenti, che prende corpo la nostra esperienza dell'ambiente che ci circonda e che le cose assumono per noi immediatamente significato. Lo stesso rigido confine tra processi percettivi, cognitivi e motori finisce per rivelarsi in gran parte artificioso: non solo la percezione appare immersa nella dinamica dell'azione, risultando più articolata e composita di come in passato è stata pensata, ma il cervello che agisce è anche e soprattutto un cervello che comprende.*

Tale comprensione, che si realizza attraverso un'immediata trasposizione in termini motori di dati sensoriali, e prescindendo da ogni esplicita e deliberata azione teorica conoscitiva, è definita "comprensione *pragmatica"* e riguarda sia le possibilità di azione che abbiamo in un determinato contesto, sia gli eventi motori che in esso si realizzano.
Grazie alle ricerche condotte dal gruppo di neuro scienziati di Parma, si è potuto in seguito appurare come il sistema specchio svolga un ruolo importante anche nell'imitazione di atti, nel conseguente apprendimento tramite imitazione, e di come offra le basi neurofisiologiche per la creazione di quello spazio intersoggettivo condiviso che supporta le modalità di comunicazione linguistica ed empatica.

1.3.1 Le proprietà visuo-motorie dei neuroni

La scoperta dell'esistenza di neuroni con proprietà visuo-motorie prima nella scimmia, e successivamente nell'uomo, è avvenuta grazie all'analisi di alcuni semplici gesti della mano e della bocca.

I primi ad essere stati scoperti nell'area F5 della corteccia premotoria ventrale[1] della scimmia (che contiene rappresentazioni motorie della mano e della bocca), sono stati i cosiddetti "***neuroni canonici***", una particolare classe di neuroni di tipo visuo-motorio che si attivano sia durante l'esecuzione di specifici atti motori, riferiti a un oggetto tridimensionale, sia durante la semplice osservazione dell'oggetto stesso, intervenendo in maniera decisiva nel processo di trasformazione dell'informazione visiva di un oggetto negli atti motori necessari per interagire con esso.

Al secondo gruppo di neuroni, anch'essi primariamente individuati nell'area F5 della corteccia premotoria ventrale della scimmia, è stato invece dato il nome di "***neuroni specchio***", a causa della loro proprietà caratteristica di attivarsi sia quando il soggetto esegue una determinata azione in prima persona, sia quando costui osserva la medesima azione mentre è compiuta da altri individui. In buona sostanza, l'osservazione di un'azione provoca nell'osservatore l'attivazione dello stesso circuito nervoso deputato all'esecuzione dell'azione osservata. Questa codifica dell'informazione sensoriale in termini motori induce nell'osservatore un meccanismo di rispecchiamento che è responsabile dell'immediato riconoscimento dei gesti altrui e del loro significato.

Ciò che accomuna questi due tipi di neuroni è il fatto che entrambi generano in chi osserva una rappresentazione motoria interna dell'azione, indipendentemente dal fatto che la si stia effettivamente compiendo, che se ne stia anticipando l'effetto, o che si stia assistendo al suo compimento ad opera di altri.

La comprensione tanto delle proprie possibilità d'azione, quanto delle azioni altrui, non sarebbe esclusivamente frutto di un'operazione conoscitiva volontaria a livello cognitivo, ma avrebbe un carattere pragmatico. Si parla, infatti, di ***comprensione pragmatica*** in riferimento alla possibilità del cervello di capire il significato degli atti compiuti da altri sulla base delle proprie competenze motorie, senza che si renda necessaria una rielaborazione cognitiva dei dati percettivi.

Alla luce di queste scoperte, si è resa necessaria una profonda rivalutazione della funzione del sistema motorio, per decenni sottovalutato e ingiustamente relegato a compiti puramente esecutivi del movimento, privi di qualsiasi valenza percettiva o cognitiva. La corporeità inizia ad essere intesa come condizione necessaria per lo sviluppo dei processi cognitivi poiché l'elaborazione dei dati sensoriali correlati all'attivazione di un movimento

[1] L'area premotoria ventrale è parte della corteccia motoria secondaria la quale, a sua volta, assieme alla corteccia motoria primaria, alla corteccia associativa parietale posteriore e alla corteccia associativa prefrontale dorsolaterale, costituisce una delle principali aree corticali del sistema sensori-motorio. L'area premotoria ventrale, composta dalle aree F4 ed F5, è coinvolta soprattutto nell'elaborazione motoria di alto livello, come la preparazione e la programmazione di sequenze di movimenti e la coordinazione di movimenti complessi e bilaterali.

comincia già a livello del sistema neurale adibito al controllo del movimento stesso.
Queste due classi di neuroni visuo-motori risultano dunque parte costitutiva di un più ampio e complesso meccanismo neurale che coinvolge l'area frontale[2] e quella parietale[3] della corteccia cerebrale, codificando in un formato comune le informazioni sensoriali e motorie.

1.3.2 I Neuroni Canonici e il Circuito F5-AIP

L'analisi del comportamento dei neuroni dell'area F5 della scimmia ha permesso di appurare che essa è formata da neuroni motori che principalmente non codificano singoli movimenti, bensì ***atti motori***, (Di Pellegrino [22]) cioè movimenti coordinati da un fine specifico. Per esempio, i neuroni di F5 che si attivano nel momento in cui si afferra del cibo, si attivano indipendentemente dal fatto che l'atto sia compiuto con una delle mani o con la bocca. Ulteriore riscontro del fatto che questo gruppo di neuroni sia in grado di identificare lo scopo dell'azione, sta nella constatazione che uno stesso movimento come la flessione di un dito che attiva un neurone durante l'atto di afferrare, non lo attiva in un atto con finalità diversa come quello di grattare. Da ciò segue l'ipotesi che in F5 gli schemi motori formino un "vocabolario di atti" che comprende il repertorio di azioni eseguibili dal soggetto e che viene selettivamente attivato dalla percezione di oggetti tridimensionali.
Una caratteristica interessante dei neuroni di F5 è che un numero considerevole di essi si attiva anche in presenza di stimoli visivi. Da uno studio di Akira Murata et al ([13] e [14]) è emerso che i neuroni che si attivano durante l'esecuzione di un compito si distinguono in:

- "*neuroni motori*" che si attivano durante l'esecuzione dell'atto motorio (afferramento)
- "*neuroni visuo-motori*" che si attivano alla sola presentazione dell'oggetto, prima o in assenza di presa.

Inoltre, tutti i neuroni visuo-motori che rispondono ad uno specifico tipo di presa (selettività motoria), rispondono solo alla vista di oggetti per i quali quella presa risulta essere efficace (selettività visiva). Tali neuroni, chiamati "canonici", (poiché già dalla prima metà del Novecento se ne era ipotizzata l'esistenza nella corteccia premotoria), condividono le proprietà visuo-motorie con un gruppo di neuroni dell'area intraparietale anteriore (AIP), da cui ricevono gran parte dell'informazione visiva e con cui creano un circuito coinvolto nelle trasformazioni visuo-motorie necessarie per afferrare un oggetto.
Hideo Sakata e colleghi nel 1995 avevano precedentemente già riscontrato in AIP, insieme a neuroni a dominanza visiva, anche la presenza di neuroni a

[2] La corteccia frontale umana costituisce almeno 1/3 dell'intera superficie cerebrale. La parte più anteriore, che a sua volta può essere suddivisa in un'area dorsolaterale ed una regione orbito frontale, è denominata corteccia prefrontale ed ha diffuse connessioni col resto del cervello. È una struttura che media le abilità del pensiero astratto, organizza il comportamento in sequenze logiche e temporali ed inibisce le risposte inappropriate agli stimoli ambientali.
[3] L'area parietale posteriore, divisa in corteccia parietale inferiore e superiore, oltre ad essere deputata all'elaborazione di determinati aspetti dell'informazione sensoriale, è coinvolta in compiti di programmazione motoria in parte anche in collaborazione con le aree prefrontali.

dominanza motoria e visuo-motoria, successivamente divenuti rispettivamente neuroni motori e visuo-motori dell'area F5.
In seguito si è osservato che anche i neuroni di AIP mostrano selettività visiva e rispondono alla vista e alla presa di uno specifico oggetto o di un ristretto gruppo di oggetti: sferici, cubici, piatti... Da ciò si evince che i neuroni di AIP e di F5, attivi durante la presa di un oggetto, fanno parte di un circuito AIP-F5, la cui funzione è quella di trasformare le proprietà visive dell'oggetto in proprietà motorie finalizzate alla presa. Tale Ipotesi è stata confermata da esperimenti che hanno provato come l'inattivazione alternativa dell'area F5 e dell'area AIP comporti una serie di difficoltà nel conformare la mano alle caratteristiche dell'oggetto per la presa.

I Neuroni Canonici e il concetto di *affordance*

Attraverso i neuroni canonici, l'informazione visiva relativa ad un oggetto viene tradotta in informazione motoria, e più precisamente negli atti motori necessari per interagire con esso. Ciò significa che, tramite il meccanismo di attivazione di questi neuroni, l'uomo percepisce lo spazio attorno a sé, ed il contesto in termini di possibilità d'azione ancor prima che intervenga e in modo totalmente indipendente dall'elaborazione concettuale.
Il significato funzionale di questo meccanismo neuronale si può ritrovare nella nozione di "***affordance***", termine introdotto dallo psicologo della visione James J. Gibson ([4]) per definire le proprietà dell'ambiente che si offrono ad ogni essere umano o animale in termini di possibilità d'interazione. Gibson definisce le *affordances* come le possibilità di azione latenti nell'ambiente, misurabili oggettivamente e indipendenti dalla capacità dell'individuo di riconoscerle, ma in ogni caso riferite all'individuo e quindi dipendenti dalle sue capacità. Esse non sono riducibili alle proprietà fisiche dell'oggetto, ma si riferiscono ad opportunità pratiche che l'ambiente, e gli oggetti in esso compresi, offrono al determinato organismo che li percepisce e con essi interagisce. Si tratta dunque di una proprietà che non appartiene né all'oggetto né a colui che ne usufruisce, bensì è individuabile nella relazione che tra essi si instaura. Queste qualità percettive degli oggetti attivano gruppi specifici di neuroni dell'area intra-parietale anteriore e l'informazione così selettivamente elaborata viene trasmessa ai neuroni visuo-motori di F5, i quali non codificano più le singole *affordances* bensì gli atti motori ad esse corrispondenti.
Il concetto di *affordance* è successivamente stato rielaborato e distinto in: "affordances naturali" e "affordances culturali": queste ultime, a differenza di quelle naturali a cui si è fatto finora riferimento, riguardano oggetti di valore culturale nel senso di uso culturalmente determinato. Per esempio, se una sedia normalmente offre all'adulto la possibilità di sedersi e al bambino quella di accovacciarvisi sotto per nascondersi (entrambi esempi di *affordance* naturale dipendente dal rapporto possibilistico tra l'oggetto e il soggetto agente), per la stessa sedia esposta in un museo come opera d'arte non sarebbero *affordances* valide.
In entrambi i casi quello che le *affordances* suggeriscono è lo stretto rapporto intercorrente tra il soggetto agente, con l'ambiente fisico e socio-culturale.

1.3.3 I Neuroni Specchio nella scimmia, la comprensione di azioni

Attorno agli anni Novanta, ad opera del gruppo di ricerca di Giacomo Rizzolatti ([25]), fu scoperta una popolazione di neuroni nell'area premotoria F5 del cervello delle scimmie, caratterizzati dalla proprietà di attivarsi non solo quando la scimmia eseguiva azioni ben finalizzate con la mano (ad esempio afferrare un oggetto) o con la bocca, ma anche quando osservava le stesse azioni mentre venivano eseguite da un altro individuo, sia che si trattasse di un'altra scimmia o di un essere umano.

Analogamente ai neuroni canonici, la cui attivazione è funzionale all'identificazione e pre-attivazione delle potenzialità pragmatiche di un oggetto osservato, anche in questo caso l'informazione neuronale visiva relativa ad un determinato tipo di atto è in grado di generare nell'osservatore una rappresentazione interna che "rispecchia" il movimento osservato riconducendolo al suo repertorio di atti motori. Inoltre, come i neuroni canonici, anche i neuroni specchio sono inclusi in un circuito neurale più ampio che integra diverse aree corticali che ricevono l'informazione visiva dal Solco Temporale Superiore (STS) attraverso la mediazione del settore del lobo parietale inferiore, formato a sua volta dalle aree PF e PFG[4] che contengono i cosiddetti neuroni specchio parietali.

Non sorprende che la scoperta del sistema specchio sia ben presto stata paragonata a quella del DNA, poiché ha rivoluzionato il modo di concepire i meccanismi, non solo di produzione delle azioni, ma anche, e soprattutto, i meccanismi di comprensione delle azioni osservate, dimostrandone la base motoria.

Diversi studi hanno confermato che l'attivazione dei neuroni specchio è associata all'identificazione dello scopo dell'atto motorio, indipendentemente dai movimenti compiuti per eseguirlo. Si è osservato che i neuroni di F5 non codificano singoli movimenti, bensì atti motori, cioè sequenze di movimenti volti a un fine specifico. I neuroni possono essere così classificati in neuroni "afferrare con la mano", "manipolare", "collocare" e così via.

Ulteriori dati sperimentali sembrano suggerire che lo scopo dell'atto motorio sia l'unico elemento necessario per l'attivazione della risposta visiva di un neurone specchio, indipendentemente dall'effettore, dal momento che risposte analoghe sono prodotte sia quando l'atto dell'afferrare è eseguito dalla scimmia stessa con la mano o con la bocca, sia alla vista dello sperimentatore intento ad afferrare del cibo con uno strumento artificiale.

Per quanto riguarda le condizioni di attivazione non è stata rilevata alcuna attività del sistema specchio della scimmia in caso di visione di gesti intransitivi, cioè in assenza di un'effettiva interazione effettore-oggetto, né in caso di visione di movimenti che mimano l'azione. Durante uno studio più recente (Ferrari [23]) però, è stato possibile osservare l'esistenza di neuroni specchio correlati all'esecuzione o osservazione di azioni della bocca, infatti,

[4] L'attività del lobo parietale inferiore è legata a compiti sia motori che sensoriali. Più specificamente sembra coinvolta in compiti di trasformazione degli input da coordinate retinocentriche a coordinate egocentriche o centrate sull'arto, nell'integrazione sensoriale della programmazione motoria degli arti superiori, nei compiti di prensione e manipolazione e coordinazione **occhi - mano**.

oltre ai neuroni “ingestivi” (legati ad azioni transitive come leccare, mordere, masticare), esistono dei neuroni definiti “comunicativi” per la loro proprietà di attivarsi alla visione di azioni facciali intransitive facenti parte del repertorio comunicativo della scimmia, quali lo schiocco e la protrusione delle labbra e della lingua. Poiché tali neuroni sono in grado di associare una risposta motoria “ingestiva” a una risposta visiva “comunicativa”, si ritiene che essi abbiano svolto un ruolo significativo nella storia evolutiva circa lo sviluppo della comunicazione sociale.

Il sistema dei neuroni specchio è dunque alla base della capacità del soggetto di comprendere, in maniera implicita, il significato intenzionale delle azioni altrui; conferma di ciò è che la scimmia è in grado di intuire l’atto anche se non ne segue l’intera esecuzione. Questo è possibile perché, come sostiene Vittorio Gallese ([15]), le trasformazioni neuronali operate dal sistema specchio generano un atto “simulato” internamente, corrispondente in tutto e per tutto all’atto osservato. Tale meccanismo di simulazione interna, dal momento che non è riflessivo e si realizza a livello di attivazione neuro-motoria, viene definito “***simulazione incarnata***”.

Mediante la simulazione multimodale dell’azione da parte del sistema neuromotorio dell’osservatore, la parte non vista dell’azione può essere ricostruita e il suo scopo può essere implicitamente compreso. Questa proprietà spiega anche come sia possibile che l’atto compiuto da altri possa essere compreso anche quando non è visto, ma se ne percepisce solo il suono caratteristico. Lo studio di Kohler ([16]) ha rilevato la presenza dei cosiddetti neuroni “***audio-motori***”, codificanti azioni normalmente accompagnate da un suono caratteristico e riconoscibile, essi si attivano non solo in concomitanza con l’esecuzione o con l’osservazione delle suddette azioni, ma anche in caso di semplice ascolto del suono caratteristico da esse prodotto. La comprensione implicita, pragmatica, dell’atto parzialmente percepito attraverso una sola modalità sensoriale è resa possibile proprio grazie all’esistenza di rappresentazioni neuronali complete del movimento, in cui l’integrazione multimodale sensori-motoria è operata dal sistema specchio.

Altri studi suggeriscono un possibile coinvolgimento del sistema specchio nel riconoscimento dell’intenzione che sottende un’azione. Fogassi ([17]) ha documentato la diversa attivazione di parte dei neuroni del lobo parietale inferiore (LPI) in presenza dello stesso atto dell’afferrare inserito in azioni con fini diversi. Tale diversità di risposta è stata rilevata sia nel caso in cui fosse la scimmia stessa a compiere l’atto, sia nel caso in cui assistesse semplicemente alla sua esecuzione.

La capacità di predire lo scopo finale di un atto motorio incluso in una catena di atti ha portato alla formulazione dell’ipotesi secondo cui i neuroni di LPI che codificano singoli atti motori sono inclusi in “catene neuronali intenzionali predeterminate”, ognuna delle quali codifica una specifica azione. In questo modo, a seconda della catena motoria attivata dall’esecutore, l’osservatore è in grado di attivare lo schema motorio corrispondente e quindi di anticiparne l’intenzione.

I Neuroni Specchio nell'uomo

Nonostante l'impossibilità di rilevare l'attività di singoli neuroni, ma solo di intere aree cerebrali, l'esistenza nella nostra specie del meccanismo specchio è stata comunque provata grazie all'utilizzo di tecniche non invasive di neurofisiologia, quali l'EEG (elettroencefalogramma), la MEG (magnetoencefalografia) e la TMS (stimolazione magnetica transcranica), e di tecniche di *imaging* cerebrale quali la PET (tomografia ad emissione di positroni) e la fMRI (risonanza magnetica funzionale per immagini). Proprio grazie a queste ultime è stato possibile localizzare i neuroni specchio umani:

a) nella regione parieto-frontale (lobulo parietale inferiore, area premotoria ventrale compresa l'area di Broca e parte posteriore del giro frontale inferiore;

b) nell'insula e nel corpo cingolato anteriore.

I numerosi studi condotti dimostrano che il sistema dei neuroni specchio è più esteso nell'uomo rispetto che nella scimmia e, pur mantenendo inalterate le sue funzioni primarie deputate al controllo dell'esecuzione delle azioni e al riconoscimento e comprensione del significato degli atti altrui, si distingue per alcune caratteristiche dai risvolti funzionali molto importanti come la capacità di codifica sia di atti transitivi che intransitivi, e quella di attivazione anche in caso di atti mimati, in assenza quindi di effettiva interazione con l'oggetto. È provato, inoltre, che i neuroni specchio non si attivano solo all'osservazione di azioni eseguite con la mano, ma anche di quelle eseguite con la bocca, il piede e altre parti del corpo.

1.3.4 La comprensione delle azioni e delle intenzioni altrui

Per comprensione delle azioni altrui si intende la capacità di interpretare un insieme di movimenti nei termini di un atto motorio finalizzato. Analogamente alla scimmia, il sistema specchio umano ha la capacità di poter codificare non solo singoli atti, ma intere catene di atti accedendo così al significato non solo del tipo di atto eseguito, ma anche del suo scopo in termini di intenzionalità. Quanto già rilevato da Fogassi per i neuroni specchio della scimmia vale anche per l'uomo, ciò è stato provato da Iacoboni ([18]) in uno studio in cui ai partecipanti venivano presentate tre sequenze filmate diverse:

- nella prima erano presentati oggetti (una tazza, una teiera, un piatto con biscotti) la cui disposizione su una tavola suggeriva il contesto di una colazione da cominciare o già ultimata;
- nella seconda sequenza si scorgeva la mano di una persona che afferrava una tazza in assenza di contesto;
- nella terza si vedeva la stessa mano afferrare la tazza all'interno dei due contesti della prima sequenza, i quali suggerivano due diverse intenzioni sottostanti all'azione (bere o sparecchiare).

Si è visto che l'osservazione delle azioni nel loro contesto presentata nella terza condizione era associata a un sensibile incremento dell'attività di una parte del sistema specchio, e ciò testimonia che le aree premotorie dotate di proprietà caratteristiche dei neuroni specchio sono coinvolte anche nella comprensione delle intenzioni che hanno promosso una determinata azione.

In conclusione, i dati della scimmia e quelli dell'uomo mostrano che l'intenzione che sottende l'azione eseguita da altri viene compresa dal sistema motorio grazie al sistema dei neuroni specchio.[5]

Imitazione e Apprendimento

Sulla base di quanto evidenziato precedentemente circa il funzionamento del sistema specchio per la comprensione dell'azione e dell'intenzione contenuta, si può affermare che la risonanza esiste in quanto è presente un codice neurale (substrato) comune alla base della percezione e dellesecuzione. Questo meccanismo è cruciale anche per il comportamento imitativo.

Per imitazione si intende la capacità che un soggetto ha di replicare un atto eseguito da un altro soggetto. Tale atto può appartenere o meno al patrimonio motorio del soggetto imitante:

- nel primo caso il soggetto imitante si limita ad attivare un'immagine mentale dell'atto già posseduta e lo riproduce, anche in assenza della comprensione del suo significato (es. i neonati che imitano espressioni del viso degli adulti);
- nel secondo caso, trattandosi di imitazione di un nuovo schema d'azione, non è possibile "ripescare" nel proprio repertorio d'atti uno schema motorio già posseduto, ma occorre apprenderlo.

In entrambi i processi imitativi si assiste ad una attivazione dei neuroni specchio, i quali costituiscono parte del più complesso meccanismo che sottende la nostra capacità imitativa.

Mentre la ripetizione immediata di un'azione osservata è sostenuta quasi esclusivamente dal sistema specchio, l'apprendimento per imitazione richiede anche l'intervento del lobo prefrontale, in particolare dell'area 46 di Broodmann, e di alcune aree della corteccia mesiale anteriore. L'area 46 è generalmente associata a funzioni legate alla memoria di lavoro, ma in questo contesto riveste un ruolo anche nel combinare atti motori elementari in schemi motori più complessi. Durante il processo di apprendimento, infatti, i neuroni specchio sarebbero i responsabili della suddivisione dell'azione osservata in singoli elementi che vengono successivamente ricomposti in una sequenza idonea, in modo che l'azione riprodotta si avvicini il più possibile a quella osservata.

1.3.5 Dalla Comunicazione gestuale al Linguaggio

La localizzazione dei neuroni con proprietà specchio anche in corrispondenza di una delle aree classiche del linguaggio, quale è l'area di Broca[6], ha evidenziato un possibile coinvolgimento del sistema specchio nell'origine delle diverse

[5] Va sottolineato che questo tipo di comprensione non è in antitesi né preclude quella supportata dal ragionamento logico-deduttivo. Ne va però sottolineata la peculiarità di comprensione implicita e slegata da una modalità sensoriale precisa, ma riconducibile al repertorio motorio presente nel vocabolario d'atti di ogni individuo. Proprio grazie a questa proprietà l'informazione visiva può anche essere parziale, come in Umiltà et al. (2001) o sostituita da informazione sonora come in Kohler et al. (2002), dato che grazie alle rasformazioni messe in atto dal circuito F5-PFPFG assumono valenza di azione vera e propria.

[6] L'area di Broca è una parte dell'emisfero sinistro del cervello, localizzata nel piede della terza circonvoluzione frontale, le cui funzioni sono coinvolte nella elaborazione e comprensione del linguaggio.

modalità di comunicazione umana ed in particolare nell'evoluzione del linguaggio. Come riportato da Rizzolatti e Sinigaglia ([12]):

> *"Sappiamo che l'area di Broca […] possiede proprietà motorie non riconducibili esclusivamente a funzioni verbali e presenta un'organizzazione simile a quella dell'area omologa nella scimmia (area F5), attivandosi durante l'esecuzione di movimenti orofacciali, brachiomanuali e orolaringei. […] Ciò sembra suggerire che le origini del linguaggio andrebbero ricercate, prima ancora che nelle primitive forme di comunicazione vocale, nell'evoluzione di un sistema di comunicazione gestuale controllato dalle aree corticali laterali."*

Sulla base di queste evidenti somiglianze anatomo-funzionali, i due studiosi ipotizzano che le origini del linguaggio non siano riconducibili soltanto alle modificazioni avvenute nell'arco di migliaia di anni nell'apparato orolaringeo, ma che si possano far risalire al sistema brachio-manuale, il quale avrebbe a sua volta fornito sostegno a quello oro facciale:

> *"Proprio l'architettura anatomo-funzionale dell'area F5 (e dell'area di Broca), contraddistinta dalla presenza di rappresentazioni motorie differenti (orofacciali, orolaringei e brachiomanuali), lascia infatti supporre che la comunicazione interindividuale non si sia evoluta da una sola modalità motoria, bensì dall'integrazione progressiva di modalità diverse (gesti facciali, brachiomanuali e, infine, vocali), accompagnata dalla comparsa dei relativi neuroni specchio."*

L'ipotesi è che un iniziale linguaggio di "proto-segni" gestuali abbia iniziato ad evolversi in un linguaggio di matrice vocale a cui la gestualità è nel corso del tempo divenuta accessoria, nel momento in cui ai "proto-segni" furono associati dei suoni. In questo processo evolutivo il sistema specchio sarebbe intervenuto permettendo la creazione di una piattaforma comunicativa condivisa che rendesse comprensibile a tutti il valore dei nuovi gesti intransitivi carichi di significato simbolico sviluppatisi sulla base di precedenti gesti transitivi[7]. Il linguaggio vocale umano si sarebbe così evoluto tramite l'informazione trasmessa per via gestuale, e il sistema specchio sarebbe poi stato capace di codificare/decodificare questa informazione.
Più in generale, un coinvolgimento dei neuroni specchio nella comunicazione sociale è stato avvalorato da un recente studio di Rizzolatti e Sinigaglia ([12]) nel quale ad adulti umani veniva richiesto di osservare filmati in cui azioni bucco-facciali venivano eseguite rispettivamente da uomini, scimmie e cani (un uomo muove le labbra per parlare, una scimmia compie un movimento ritmico delle labbra con valenza affiliativa detto *lipsmacking*, un cane abbaia). L'osservazione di questi tre tipi di azioni comunicative induceva nell'osservatore risposte corticali differenziate in base alla specie che le eseguiva: le azioni appartenenti al repertorio comunicativo umano o simili

[7] Per gesto intransitivo si intende un gesto che non è rivolto all'interazione con un oggetto, al contrario del gesto transitivo che invece comporta la manipolazione o l'uso spontaneo di oggetti.

(come nel caso della scimmia) inducevano l'attivazione di regioni del sistema motorio dell'osservatore responsabili di azioni analoghe, mentre azioni comunicative estranee al repertorio umano (abbaiare), attivavano solo le aree visive senza indurre alcun fenomeno di risonanza motoria nel cervello dell'osservatore.

In ambito linguistico (Rizzolatti [24], Corballis [21]) Il meccanismo del sistema specchio costituisce il presupposto neuro-biologico necessario per fondare lo sviluppo delle funzioni cognitive e del linguaggio sulla base dell'esperienza corporea. Il linguaggio non è più considerato un sistema autosufficiente deputato alla "manipolazione" di concetti astratti e simbolici che non hanno alcun collegamento con il loro significato, ma una facoltà mentale legata al funzionamento complessivo della mente e del corpo.
Secondo l'approccio embodied[8], le stesse strutture nervose che presiedono all'organizzazione dell'esecuzione motoria delle azioni svolgono un ruolo decisivo anche nella comprensione semantica delle espressioni linguistiche che le descrivono.
Un'ipotesi verosimile sulla base dei risultati degli esperimenti di Buccino ([19]) è che l'ascolto di frasi che descrivono azioni motorie determini la modulazione del sistema dei neuroni specchio, il quale a sua volta influenzerebbe l'eccitabilità della corteccia motoria primaria e quindi l'esecuzione dei movimenti da essa controllati. Questo studio ha dimostrato che l'elaborazione di frasi descriventi azioni eseguite dalla mano o dal piede, attiva in modo specifico regioni diverse della corteccia motoria che controllano le azioni di questi stessi effettori. Questo tipo di attivazione selettiva della corteccia motoria e premotoria è stato riscontrato perfino in caso di lettura silenziosa o di ascolto di parole e frasi descriventi azioni della bocca, della mano o del piede.
Tutte queste evidenze indicano come il sistema dei neuroni specchio sia non solo coinvolto nella comprensione del significato delle azioni osservate, ma si attivi anche durante la comprensione di espressioni linguistiche descriventi le stesse azioni.

L'empatia

Il riconoscimento degli stati emotivi altrui sembra poggiarsi su un sistema di circuiti neurali che condividono le proprietà "specchio" rilevate nel caso della comprensione delle azioni.
La capacità di parti del cervello umano di attivarsi alla percezione delle emozioni altrui, espresse con moti del volto, gesti e suoni, e la capacità di codificare istantaneamente questa percezione in termini viscero-motori, rende ogni individuo capace di agire in base a quella che viene definita "***partecipazione empatica***". È stato provato che le stesse strutture cerebrali (la

[8] Secondo la teoria dell'*embodied cognition*! (cognizione incarnata) la cognizione si fonda sull'interazione del corpo con il mondo esterno e, di conseguenza, dipende dal tipo di esperienze rese possibili dal fatto di possedere un corpo dotato di particolari capacità percettive e motorie. Percezione e movimento sono strettamente intrecciati e costituiscono la matrice su cui si innestano la memoria, le emozioni, il linguaggio e ogni altro aspetto della vita. I concetti non sono riproduzioni astratte della realtà esterne né frutto delle nostre credenze, ma sono fondati sull'azione e variano in funzione del nostro corpo e del contesto.

corteccia premotoria ventrale, l'insula [9] e l'amigdala [10]) responsabili della realizzazione delle espressioni facciali corrispondenti a un'emozione si attivano anche durante l'osservazione e l'imitazione delle emozioni primarie universalmente condivise (felicità, tristezza, paura, disgusto, rabbia, sorpresa). Potrebbe essere la componente insulare del sistema specchio a fornire la base anatomica dell'empatia mediante un processo neurale che trasforma gli input sensoriali derivanti dal mondo sociale circostante in reazioni viscerali vissute in prima persona.
Uno studio di *brain imaging* condotto da Wicker ([12]) ha dimostrato come la regione anteriore dell'insula si attivi non solo nel provare disgusto ma anche alla sola vista di espressioni facciali di disgusto degli altri.
Tale capacità implica che l'osservazione di volti altrui esprimenti un'emozione determina l'attivazione dei neuroni specchio della corteccia premotoria, che successivamente inviano alle aree somato-sensoriali e all'insula l'informazione codificata in un formato simile a quello che inviano quando l'osservatore vive in prima persona quell'emozione. La conseguente attivazione delle aree sensoriali, analoga a quella che si avrebbe quando l'osservatore esprime spontaneamente quell'emozione, sarebbe alla base della comprensione delle reazioni emotive degli altri.

1.3.6 Il ruolo dei neuroni visuo-motori: memoria, apprendimento e didattica

Da quanto detto finora risulta che il sistema dei neuroni visuo-motori individuato nella corteccia cerebrale è coinvolto in processi cognitivi di alto livello, in particolare, nel riconoscimento percettivo di oggetti, azioni ed emozioni, e nella comprensione del loro significato. Percezione, comprensione e azione si trovano così unificate in un meccanismo unitario che supporta la comprensione immediata degli atti altrui, delle loro intenzioni e la condivisione delle emozioni.
Sono essenzialmente due i livelli su cui la pratica didattica può venir influenzata dal funzionamento dei neuroni visuo-motori:

- il primo si riferisce alla rilevanza dell'esperienza pratica, in particolare motoria, per il processo di acquisizione della lingua straniera ([20]) o di altre discipline interattive;
- il secondo ribadisce l'importanza della valorizzazione del contesto fisico, socio-culturale ed emotivo in cui si svolge l'azione didattica.

Occorre che siano create le condizioni per una partecipazione "mente e corpo" del discente, e che per questo l'ambiente, il contesto, siano ricchi di stimoli e che ne siano egualmente valorizzate tutte le componenti compresa quella fisica, sociale, culturale, emotiva ed empatica.

[9] L'insula si trova in profondità nella scissura laterale di Silvio, che separa la il lobo temporale da quello parietale inferiore. La funzione di questo lobo non è stata ancora chiarita del tutto, anche se sono descritte in letteratura implicazioni nella sensibilità gustativa e nelle funzioni effettrici viscerali. Collegata al sistema limbico, è sperimentalmente dimostrato che è coinvolta nell'esperienza del dolore e di un gran numero di emozioni di base come odio, pietà, felicità, disgusto, tristezza.
[10] L'amigdala è un centro del sistema limbico del cervello posto sopra il tronco cerebrale!Ritenuta da molti un centro di integrazione di processi neurologici superiori come le emozioni, è coinvolta anche nei sistemi della memoria emozionale, nel sistema di comparazione degli stimoli ricevuti con le esperienze passate e nell'elaborazione delle esperienze olfattive.

A tal proposito, esempi significativi sono i temi di ricerca proposti dal gruppo di informatica musicale della facoltà di ingegneria di Padova, come la Stanza Logomotoria, o la Fiaba Magica, che mirano all'integrazione di percezioni multimodali sonore e tattili, con lo scopo di rendere più partecipi i bambini nelle fasi di apprendimento tramite modalità di didattica spiccatamente interattive, conoscitive e in grado di stimolare collaborazione e integrazione socio culturale con i propri coetanei.

Il Contesto fisico, socio-culturale ed emotivo come risorsa

Parlando dei neuroni canonici si è visto come essi permettano la creazione di un ponte tra il soggetto agente e il contesto in cui opera, grazie al fatto che colgono in maniera immediata le possibilità d'azione contenute nell'ambiente circostante. L'esistenza di tali neuroni, che codificano atti completi e non singoli movimenti, fornisce una spiegazione del perché gli umani interagiscono con gli oggetti quasi sempre alla stessa maniera. La formazione di schemi motori standardizzati avviene secondo un processo di "rinforzo motorio" che privilegia e consolida l'atto più efficace per l'ottenimento di un determinato scopo. Così, ad esempio, la presa che risulta più adeguata per l'afferramento di un certo oggetto viene ripetuta con successo fino alla sua memorizzazione definitiva. Questo fa si che le nostre risposte comportamentali in un determinato contesto siano "automaticamente" predisposte dalla nostra esperienza e dalla capacità di anticipazione che ne deriva.

Inoltre, si può affermare che il coinvolgimento motorio, l'agire intenzionalmente in un contesto attraverso l'attivazione dei neuroni specchio (grazie alla loro proprietà di "fare proprie" le azioni e le esperienze altrui) contribuisce alla creazione di una traccia mnestica più stabile non solo in chi compie in prima persona l'azione, ma anche in chi osserva.

1.4 AUDITORY ICONS

Auditory icons

L'idea è utilizzare i suoni quotidiani e naturali per rappresentare le azioni e le attività tramite un'interfaccia.

Secondo Gaver ([1] e [2]), il nostro sistema uditivo prima di tutto è un tool per interagire con la vita di tutti i giorni, e l'ascolto musicale è solo una parte minima di tutto ciò che compete all'udito. Quando ascoltiamo, in realtà non percepiamo suoni, bensì eventi e sorgenti sonore, questo è denominato "ascolto di ogni giorno", in contrapposizione all'"ascolto musicale"; ad esempio, quando si sente un quartetto d'archi suonare, potremmo concentrare l'attenzione sulle sensazioni evocate, ed in questo caso si parla di ascolto musicale, ma allo stesso tempo, potremmo porre l'attenzione sulle caratteristiche dei suoni e l'identificazione degli strumenti, si parla allora di ascolto di tutti i giorni.

L'approccio ecologico proposto da Gaver tiene prima di tutto sotto controllo tutti gli aspetti legati alla sorgente del suono e l'interazione dell'onda sonora con l'ambiente circostante, quindi fa uso di concetti come la forza, la forma delle superfici, in contrapposizione all'approccio classico che invece si concentra sulla frequenza, l'ampiezza e la fase.

Il suono da informazioni circa l'interazione di materiali in una posizione ben precisa e in un ambiente altrettanto preciso.

Uno studio complementare di Gaver si focalizza invece sulle informazioni acustiche tramite le quali si ottengono informazioni circa gli eventi: il punto di partenza si basa sempre sulla differenza tra l'esperienza degli stessi suoni (ascolto musicale) e la percezione del suono prodotto dall'evento (ascolto di ogni giorno). Gaver propose parecchi algoritmi per sintetizzare su più dimensioni i suoni di tutti i giorni, secondo un approccio di sintesi e analisi, in cui vengono analizzati sia suoni che eventi per estrarre i dati necessari, ri-sintetizzati e comparati agli originali.

Mentre nella sintetizzazione degli strumenti musicali si vuole ottenere la riproposizione di un'identica percezione, per i suoni di ogni giorno l'importante è veicolare la stessa informazione riguardo un particolare aspetto dell'evento. Per evidenziare gli aspetti significativi e gli attributi della sorgente, l'analisi acustica dell'evento sonoro deve essere supportata da un'analisi fisica dell'evento stesso.

Gaver ha identificato 3 eventi sonori di base: impatto, raschiamento, gocciolamento, e porta tre esempi di eventi temporalmente complessi come rottura, rimbalzo e versamento.

Gaver definisce in modo chiaro gli obiettivi delle auditory icon come "veicolo che porta informazioni utili circa gli eventi dei computer".

Tutte le classi di suoni sono considerati sulla base dei loro attributi psico-acustici nella prospettiva di definizione di alcuni modelli fisici, o sulla base di considerazioni spettrali, con l'obiettivo di ottenere la sintesi e i parametri di controllo delle auditory icons. L'aspetto principale è che la non linearità della

relazione tra parametri fisici e risultati percettuali dovrebbe essere bypassata tramite una semplificazione di modelli.

Mapping

L'obiettivo principale del design sonoro è trovare un "mapping" effettivo tra i dati e gli oggetti sonori che si presume li rappresentino in modo che abbiano un significato compiuto sia a livello percettivo sia a livello cognitivo.
Kramer ([26]) ha descritto il ruolo delle strutture che mediano tra i dati e l'ascoltatore ovvero ciò che viene definito mapping. Il termine "***audificazione***" (audification) indica una traduzione diretta di una data onda sonora verso il dominio dell'udibile con gli obiettivi del monitoraggio e della comprensione (esempi di questo approccio si possono trovare nei sismografi, elettrocardiogrammi…).
Al contrario, nel processo di "***sonificazione***" (sonification), i dati sono utilizzati per controllare la generazione dei suoni, e la tecnica di generazione non è necessariamente in relazione diretta con i dati. Per esempio possiamo associare il pitch, e il timbro di una sorgente sonora a percussione con le variabili fisiche lette da un sensore in un motore di rendering.

Audificazione diretta

Il tipo di mapping più diretto è quello che prende i dati per alimentare i convertitori digitali-analogici direttamente, e restituire indietro i dati ad un certo sampling rate; questo può avvenire solamente se i dati sono serie temporali. L'idea di ascoltare i dati prodotti da un sismogramma per cercare i fenomeni rilevanti e aumentare l'apprendimento è piuttosto vecchia. L'audification, ovvero la trasposizione diretta dei dati in suono con un processamento minimale, ha senso in pochi casi come in presenza di dati sismici in quanto essi sono prodotti da fenomeni sismici, onde elastiche che sono simili alla propagazione dei terremoti tra le rocce e delle onde sonore nell'aria. In questo modo se i segnali sismici sono opportunamente trasposti in frequenza, suonano naturali alle nostre orecchie e possiamo utilizzare le nostre abilità per interpretare i rumori nelle condizioni di ogni giorno.
Una delle principali motivazioni nell'usare auditory display è che ci sono eventi importanti che sono difficili da individuare nei display visuali-temporali, ma sono facilmente individuati dalle orecchie. Ci sono molti problemi quando si prova a "sonificare" i suoni sismici, specialmente in relazione all'enorme range dinamico maggiore di 100dB e con frequenza di banda che sebbene sotto i 40Hz, si espande oltre le 17 ottave.
Per rendere udibili molti degli eventi rilevanti, i segnali registrati devono essere processati tramite gain control, compressione temporale, shift di frequenza o trasposizione, looping e stereo placement. Tutte queste tecniche sono descritte da Hayward ([27]), in particolare il raddoppio di frequenza, che lavora bene per onde sinusoidali, risulta però non libero da artefatti con i segnali reali.
E' importante che l'ascoltatore non percepisca flussi audio scorrelati a livello temporale, infatti Il problema dell'accurata localizzazione temporale dei diversi eventi è un fenomeno rilevante che dovrebbe essere sempre

considerato durante la fase di design dell'auditory display. L'audification funziona bene soprattutto quando viene integrata con i display visivi.

Mapping Naturalistico

In molti casi è possibile usare suoni naturali o meccanici per trasmettere informazioni di vari tipi, questo accade, in particolare, quando l'informazione è fisicamente correlata alla sorgente del suono di riferimento. Così l'esperienza quotidiana può essere sfruttata per interpretare i suoni stessi. Fitch e Kramer ([26] e [27]) hanno proposto di studiare i display acustici e visivi in modo che le azioni intraprese dipendano dalle differenti configurazioni delle variabili del sistema. Il sistema che viene monitorato è il corpo umano, e il sistema di riferimento è quello visivo. Il display acustico è progettato come un ibrido di suoni realistici e astratti. A tal proposito, fondamentali sono gli studi di Gaver per i suoni quotidiani, e di Kramer per i livelli di parametri imposti nel flusso audio sulla base dell'annidamento dei parametri.

Lo spazio è il concetto fondamentale per la visione, mentre il tempo è la dimensione principale dell'udito. I task e i dati considerati hanno una struttura temporale e sono inerentemente concorrenti; in questo modo si ottiene una sostanziale equivalenza di informazioni tra i display uditivo e visivo. Gli esperimenti condotti mostrano come gli utenti reagiscano più velocemente ai display uditivi rispetto a quelli visivi, inoltre gli utenti sono in grado di processare nello stesso momento un insieme di flussi audio e velocemente determinare le configurazioni salienti.

Oggetti Suonanti

Il focus non è nel suono di per sé stesso, ma l'idea è modellare la sorgente in termini di comportamenti fisici. Le tre linee guida principali sono: l'analisi della percezione, la cartoonificazione/semplificazione e il controllo dei parametri fisicamente interessanti.

"***Cartoonificazione***" si riferisce ai cartoon, e in particolare, alla tecnica per ridurre complesse informazioni audio e video nelle rispettive componenti essenziali. Queste componenti ridotte sono poi enfatizzate in modo da renderle più facili e immediatamente decodificabili. Uno dei principali vantaggi della visione cartoonificata della realtà è la facilità della realizzazione, ma anche un'intelligibilità aumentata.

Un esempio di cartoonificazione potrebbe essere il riempimento di una bottiglia d'acqua con un timbro che varia in modo enfatizzato man mano che la bottiglia si riempie e il livello del liquido cresce.

Il suono è progettato in accordo a una precisa scomposizione degli eventi fisici che intercorrono durante l'azione: innanzitutto ogni evento viene trattato separatamente, di conseguenza, l'identità di una vasta classe di suoni può essere definita e riprodotta conoscendo un insieme ristretto di modelli di interazioni fisiche di base, per esempio quelle relative alla sintesi di suoni di impatto e frizione. Questi ultimi possono essere trattati usando modelli deterministici, invece rotture, rotolamenti, e altri tipi di suoni complessi devono essere trattati con modelli stocastici. In altre parole, questi ultimi possono essere considerati come un insieme di fenomeni elementari, suoni di impatto e frizione, i cui comportamenti statistici presentano caratteristiche

ben precise, sia per la distribuzione temporale, sia per l'energia. Il rotolamento può essere considerato come un moto Browniano, il crash di una scatola metallica potrebbe essere pensato come una sequenza di eventi temporali di un processo di Poisson, rotture e crash possono essere rappresentati con processi a valanga o sistemi caotici che si auto-organizzano. La qualità di questi segnali può essere raffinata e migliorata tramite tecniche di signal processing.
Un'organizzazione ad alto livello nel tempo e nello spazio è fondamentale per allargare il vocabolario e i contenuti informativi che si vogliono veicolare con il suono stesso.
Secondo quanto sostenuto da Gaver ([1] e [2]), nell'ascolto quotidiano una persona dovrebbe essere in grado di sentire la dimensione, la forma e il materiale di un oggetto suonante. In generale, l'uomo è in grado di sentire qualcosa che non è la dimensione, la forma, la densità di un oggetto ma l'effetto della combinazione di questi attributi.

1.5 EARCONS, SPEECH ICONS, SPEARCONS

1.5.1 Alert e Warning

L'udito tende a essere naturalmente un senso strettamente correlato agli avvisi, in questo caso si dice siano le orecchie che guidano gli occhi. Gli "auditory warnings" sono applicati a 4 aree distinte: device personali, trasporti, ambito militare e controllo di ambienti; a queste quattro se ne aggiunge una quinta che comprende gli alert in ambito geografico ([38]).
L'approccio scientifico agli warnings acustici è diviso solitamente in due fasi: l'ascolto e l'apprendimento, influenzato dalla formazione, dalla struttura e dal numero di segnali coinvolti. Gli allarmi dovrebbero essere posizionati 15,25 dB sopra la soglia dell'ambiente.
Secondo l'approccio classico, allarmi e avvisi sono strettamente correlati al suono di una sirena, ma in realtà secondo studi più recenti, ci sono molti più tipi di allarmi che condizionano la vita di tutti i giorni.
Stanton e Edworthy ([28]) hanno distinto in AIA alarm iniated activities, attività scaturite da allarmi cioè dove è necessaria una risposta pronta all'evento, ed eventi critici dove invece è richiesto un adeguato ragionamento deduttivo; in generale, l'allarme deve garantire un perfetto bilanciamento tra la qualità sonora e l'impatto che ha sulle attività quotidiane.
L'uso di suoni evocativi per generare warning e alert è particolarmente significativo, in quanto permette di mappare direttamente sul suono il concetto di urgenza, un esempio sono i suoni generati dal monitor per gli ECG. Se il suono scelto per l'allarme non è sufficientemente concreto, ma tende a essere astratto, esso è poco evocativo e fallisce nel suo intento anche se l'utente finale è sottoposto a un lungo periodo di test.
Per capire come alcune proprietà acustiche dei suoni affliggano la stessa identificazione del fenomeno associato, è stata condotta un'analisi acustica su 41 suoni della vita di tutti i giorni. E' risultato chiaro come un fattore associato alle performance percettive sia l'unione di:

- armoniche nei suoni continui e
- pattern spettrali simili in gruppi di suoni non continui.

Questa unione è chiamata *Hst* e descrive una sorta di entropia spettro-temporale. Secondo questo approccio, gli warning sono strutturati come pattern spettrali simili in gruppi ripetuti; la ripetizione di un componente aumenta l'identificazione, mentre l'aggregazione di componenti differenti danneggia l'identificazione.
Guyot nei suoi studi ha cercato le relazioni tra cognizione e percezione tramite la categorizzazione dei suoni di tutti i giorni; in particolare l'astrazione avviene secondo tre livelli:

1. tipi di eccitazione,
2. movimento che produce il pattern acustico,
3. identificazione dell'evento.

Le categorie dei suoni utilizzati sono:

- correlate all'acqua;
- legate ai segnali e al pericolo;

- porte e suoni modulati;
- 2 o più componenti transitori.

1.5.2 Earcons

Blattner, Sumikawa e Greenberg ([29]) hanno introdotto il concetto di "earcon", ovvero messaggi audio non verbali che sono usati nelle interfacce dei pc per fornire informazioni all'utente circa gli oggetti, le interazioni o le operazioni intraprese. Questi messaggi sono chiamati motivi, brevi successioni di picchi arrangiati in modo da produrre un pattern tonale sufficientemente distinto per funzionare come un'entità individualmente riconoscibile.
Gli earcon devono essere imparati perché non c'è un collegamento intuitivo tra il suono e quello che rappresenta: gli earcon sono segnali musicali astratti, al contrario delle "auditory icon" proposte da Gaver, dove i suoni di tutti i giorni sono usati per costruire interfacce.
Brewster ([30]) ha presentato un approccio strutturato all'auditory display, definendo regole di composizione e un'organizzazione gerarchica dei parametri musicali come timbro, ritmo, registro… col fine di rappresentare la struttura gerarchica di un file system tramite earcon. Un'applicazione tipica sono le interfacce telefoniche in cui la navigazione visuale è un problema data l'assenza di display.
L'idea è quella di definire un insieme di suoni e regole di composizioni per dar luogo a cosiddetti atomi musicali cioè gli earcons, con la caratteristica chiave di essere facilmente distinguibili l'uno dall'altro.
Gli earcorns devono, per loro natura, essere facilmente riconoscibili e distinguibili, inoltre, la monofonia del segnale e la banda limitata inducono forti vincoli al design degli stessi (Hemenway [39]).
Un altro importante aspetto è correlato al concetto di memoria musicale; molti esperimenti hanno mostrato infatti come molti ascoltatori riconoscano più facilmente gli earcon dopo una settimana dal training, piuttosto che immediatamente dopo l'apprendimento.
In fine, il design degli earcons è problematico tanto maggiore è il numero di livelli gerarchici da rappresentare. Al crescere del numero di earcon contemporanei, cresce significativamente anche la difficoltà associata al loro riconoscimento, in particolare si è visto come usando timbri multipli o scaglionando gli attacchi, aumenti la capacità di identificare gli attributi.

Sonificazione

La sonificazione può essere considerata come l'equivalente audio delle rappresentazioni grafiche del dominio visivo. L'obiettivo principale della sonfiicazione è quello di definire un modo per rappresentare la realtà attraverso il suono.
Hermann e Ritter ([31]) hanno proposto un nuovo approccio alla sonificazione, partendo da un'indagine profonda circa il collegamento tra suono e significato. L'idea di base è trovare un modo di effettuare la sonificazione senza alcuna forma di ascolto musicale, né tanto meno di allenamento.
Il Model Base Sonification proposto prevede un modo naturale di interazione col sistema di sonificazione e consente lo sviluppo di auditory display per un qualsiasi insieme di dati. I due obiettivi sono stati raggiunti tramite l'utilizzo di

un oggetto virtuale nell'interazione e un modello sonoro parametrizzato come display uditivo. Questo approccio ha parecchi vantaggi: pochi parametri da ottimizzare, collegamento naturale tra suono e dati, apprendimento soft, interfaccia intuitiva, un controllo continuo e naturale.
Rath e Rocchesso ([32]) hanno spiegato come i continui feedback sonori prodotti dai modelli fisici si possano impiegare in modo fruttuoso nelle HCI. L'idea del feedback deriva dall'analisi del comportamento naturale nel quale ci si riferisce a suoni continui per avere informazioni circa quello che accade. I feedback sonori hanno il vantaggio che possono aumentare l'efficacia dell'interazione, senza distogliere l'attenzione dall'obiettivo. Un esempio significativo è costituito da una pallina rotante in una superficie, il cui suono fornisce informazioni circa direzione, velocità, superficie e dinamica del moto.
La caratteristica fondamentale di questo modello è la reattività e il comportamento dinamico: il modello di impatto usato produce complessi effetti transitori che dipendono dai parametri dell'interazione e dagli stati istantanei degli oggetti a contatto.
Questi risultati mostrano come il feedback continuo di un suono sviluppato con precisione possa essere utilizzato per la sostituzione sensoriale di un feedback tattile o visivo in molte interfacce.
Molti contesti multimodali possono beneficiare del modello cartoon sound per migliorare l'efficacia dell'interazione, per esempio i videogames o gli ambienti virtuali.

1.5.3 Speech icons

Il modo più intuitivo per veicolare informazioni circa un oggetto è quello di fornirne una semplice e rapida descrizione verbale, ed è per questo motivo che molti sistemi per non vedenti fanno uso proprio di interfacce speech-based, ovvero guidate dalla voce.
Tuttavia, utilizzare la voce può comportare alcuni problemi, per esempio legati alla bandwidth relativamente ristretta in frequenza; inoltre una parola o una breve frase è difficile da spazializzare e localizzare rispetto ad un suono di altro tipo, che di solito utilizza bandwidth in frequenza molto maggiori.
Oltre a ciò, il processamento della voce richiede molte risorse mentali; a tal proposito, basti pensare a quanto sia difficile sostenere una conversazione con un interlocutore mentre si ricevono altre sollecitazioni verbali.
Creare o sintetizzare speech icons è però molto più veloce e semplice rispetto alla sonificazione di auditory icons o di earcons, in quanto si possono utilizzare appositi software di sintesi vocale TTS text to speech.

1.5.4 Spearcons

Con l'obiettivo di trovare alternative per migliorare le performance e l'usabilità delle interfacce a menu, Walker et al ([33]-[37]) hanno sviluppato le spearcons ovvero un tipo di speech icons, sostanzialmente più veloci. Le spearcons usano frasi pronunciate in modo talmente veloce, da non essere più riconoscibili come voce. Basate sulle stesse semplici metodologie di creazione proprie delle speech icons, le spearcons possono essere facilmente create automaticamente usando un software text to speech e un algoritmo che aumenta la velocità della frase pronunciata. Ciascuna spearcon è unica a causa

della specifica frase sottostante, ed è per questo che le spearcons risultano da un lato distinte perché diverse tra loro, ma allo stesso tempo simili a tal punto che si possono formare famiglie di suoni correlati, un po' come accade con gli earcons. Palladino e Walker ([36]) hanno scoperto che l'apprendimento di un menu, quando si utilizzano spearcons, diviene molto più veloce, se confrontato con l'analogo earcons. Dal momento che il mapping tra spearcons e gli oggetti che essi rappresentano non è arbitrario, in generale si richiede un minor tempo di apprendimento.

1.5.5 Suoni ibridi

Tutti i metodi di sonificazione hanno vantaggi e svantaggi, perciò quando si sviluppa un auditory display, bisogna capire innanzitutto quanto apprendibili sono i suoni in questione, per poi ottenere un'interfaccia facile da apprendere ed utilizzare.
A tal proposito, è possibile combinare differenti tipi di suoni, ad esempio earcons e auditory icons, in differenti modi, al fine di generare suoni appropriati, in grado di coniugare le peculiarità di entrambe le tipologie, e di sopperire agli svantaggi dell'una, con i punti di forza dell'altra.

1.6 AUDIO 3D

Le applicazioni di mixed reality (MR) si appoggiano in particolare sui processi di rendering che considerano il mondo come punto di riferimento, invece di considerare l'ascoltatore come riferimento. Il grado di immersione e la definizione di una struttura spaziale sono qualità direttamente correlate al concetto di "virtuality continuum" introdotto da Milgram et al. ([42]) per i display visuali. Queste nozioni possono essere adattate ai display audio virtuali e alla realtà audio aumentata (AAR) includendo effetti sonori e sovrapposizioni sonore generate al computer al di sopra dei segnali audio acquisiti in real-time.
L'obiettivo è ottenere un modello che possa essere sviluppato per la rappresentazione sonora immersiva con differenti gradi di virtualità. Ci si concentra sui sistemi basati su cuffie per il rendering binaurale, nonostante si abbiano svantaggi quali invasività, risposta non piatta in frequenza, che però sono opportunamente controbilanciati da tutta una serie di vantaggi come l'eliminazione dei riverberi e altri effetti acustici dello spazio reale d'ascolto, la riduzione del rumore di sottofondo e la compatibilità con display audio adattabili. Con questo tipo di sistemi ogni orecchio riceve segnali distinti, semplificando notevolmente il desing del rendering audio 3D.
Al giorno d'oggi molti sistemi di mixed reality sono in grado di controllare fluentemente due dimensioni di uno spazio audio, per esempio sorgenti sonore posizionate sul piano orizzontale in relazione a un sistema di coordinate centrate sulla testa. Tecnologie di tracciamento della testa, risposte in frequenza con testa artificiale o modelli adattabili per la localizzazione orizzontale consentono un'accurata discriminazione tra sorgenti sonore posizionate attorno all'utente e all'interno del sottospazio definito.
La terza dimensione, elevazione o controllo verticale, richiede una caratterizzazione specifica dell'utente in modo da simulare l'effettiva percezione sul piano verticale principalmente a causa del profilo dell'orecchio esterno cioè della pinna ([43]).
L'approccio utilizzato prevede il processamento di un segnale monofonico dalla parte del ricevitore utilizzando un filtro low-order, in modo da ridurre il costo computazionale. Inoltre, grazie all'intrinseca bassa complessità, tale approccio può essere utilizzato per rappresentare scene con molti oggetti audio-video in differenti situazioni come videogiochi, scene cinematografiche, intrattenimento. Si può anche utilizzare in qualsiasi scenario che richiede un una spazializzazione molto realistica del suono e la riproduzione di suoni personalizzati.

1.6.1 Grado d'immersione DI e Deviazione dalle coordinate CSD

Una scena audio 3D, creata dalla riproduzione di suoni binaurali, può essere calcolata dal segnale proveniente da ciascuna sorgente sonora usando i metadati associati e poi sommando il segnale sinistro e quello destro per produrre il segnale stereo finale inviato alle cuffie. Questa architettura può

consentire anche una sostanziale scalabilità sulla base delle risorse computazionali a disposizione o della bandwidth disponibile.
Criteri psico-acustici possono definire le priorità del rendering delle sorgenti sonore e i relativi attributi come l'udibilità della sorgente. Più precisamente, in relazione alla quantità di bandwidth disponibile, la sorgente meno percepibile può essere rimossa dalla scena senza compromettere la qualità complessiva del rendering e l'esperienza offerta dal servizio.
Nelle applicazioni tipiche di virtual audio la testa dell'utente è il riferimento centrale per il rendering degli oggetti audio. Inizialmente la posizione della testa dell'utente stabilisce un sistema virtuale di coordinate e costruisce una mappa della scena audio virtuale. Nel caso di uno scenario caratterizzato da molti suoni ambientali, gli oggetti sono posizionati all'interno del mondo fisico che sta attorno all'utente e perciò concettualmente posizionati consistentemente a un sistema di coordinate fisiche. Posizionare oggetti audio virtuali in un ambiente variegato richiede la super-imposizione di un unico sistema di coordinate sull'altro.
Sulla base della specifica natura di un'applicazione, si possono usare parecchi settaggi per caratterizzare il sistema di coordinate per gli oggetti di virtual audio e la locazione degli stessi oggetti nell'ambiente.
Una semplice distinzione è la scelta di riferirsi a un sistema di posizionamento globale o a un sistema di coordinate locale. Una classificazione ideale può aiutare la definizione delle possibili applicazioni che usano tecnologie di audio spazializzato. In alcuni casi è necessario far coincidere i due sistemi di coordinate in modo che le sorgenti audio virtuali appaiano in una specifica locazione all'interno dell'ambiente fisico, mentre in altri contesti le sorgenti virtuali "galleggiano" da qualche parte attorno all'utente perché l'obiettivo è giungere a un livello concettuale di disgiunzione nell'interazione con l'utente.
La caratterizzazione si muove attorno ad uno spazio bidimensionale semplificato, definito in termini di gradi di immersione (DI) e sistema di deviazione di coordinate (CSD). Questa è una semplificazione dello spazio 3D di Milgram et al. ([42]), le corrispondenze sono effettuate considerando attentamente le tre entità coinvolte:

- ✓ il mondo reale,
- ✓ il mixed reality engine
- ✓ e l'ascoltatore.

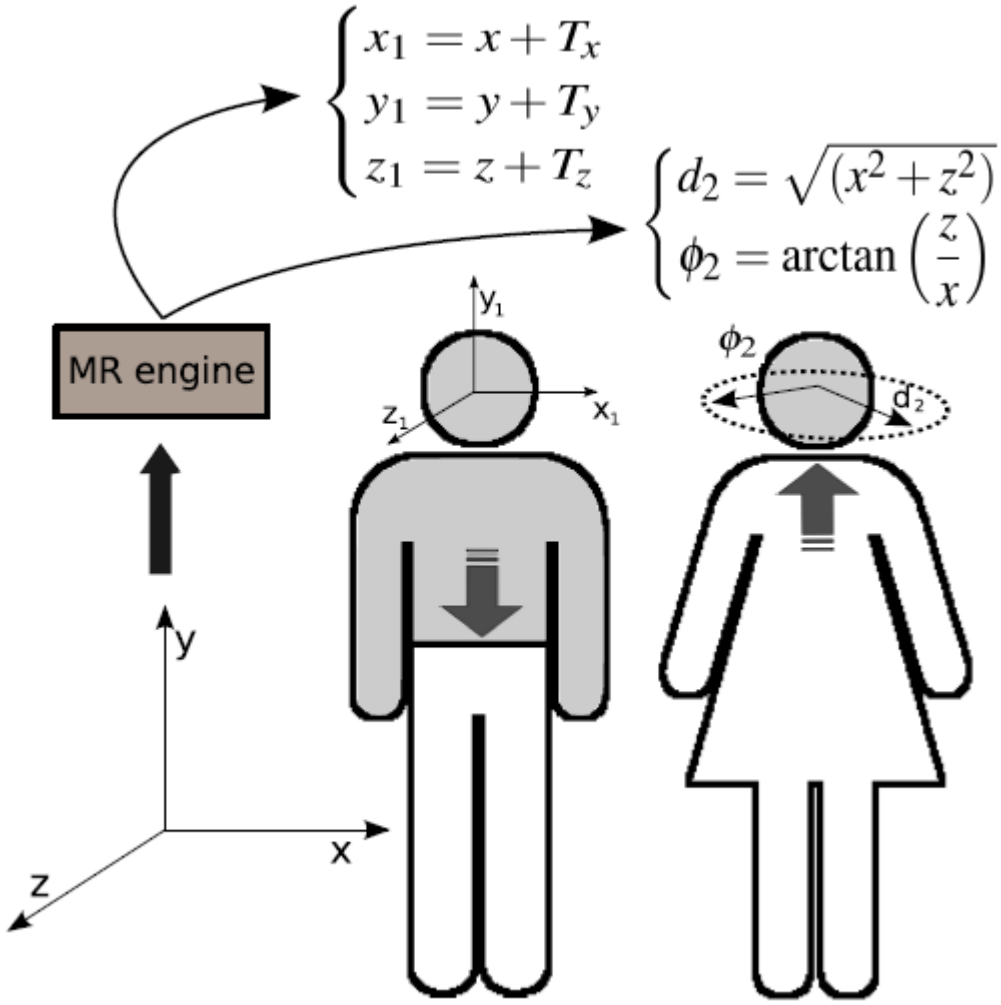


Figura 3 l'utente maschio alto grado DI e bassa CSD, l'utente femmina basso grado Di e alta CSD

L'MR engine è l'intermediatore tra la realtà e la rappresentazione percepita dall'ascoltatore; l'MR engine conosce ogni cosa circa la posizione degli oggetti nella realtà e può renderizzare la scena acustica sintetica come se l'ascoltatore percepisse un mondo coerente.
D'altro canto il problema del realismo sta nel fatto che le tecnologie coinvolte nell'MR engine e la complessità delle tassonomie per un sistema di questo tipo crescono considerevolmente. Il grado di immersione DI è definito in accordo alla seguente idea: quando un ascoltatore è circondato da suoni reali, tutto il suo corpo interagisce con l'onda acustica, per esempio una tecnologia con un buon livello di realismo è capace di monitorare l'intero corpo dell'ascoltatore in termini di posizione e orientazione.
In figura il soggetto sulla sinistra è un esempio di alto grado di immersione DI che corrisponde a un un'alta percentuale di corpo che viene tracciato nel sistema di coordinate virtuali, e fornisce una bassa deviazione di coordinate virtuali dal sistema di coordinate fisiche, in quanto frutto di una semplice traslazione. Al contrario, il soggetto sulla destra mostra un basso grado di immersione DI e un alto CSD, rappresentato da una testa grigia e uno spazio virtuale 2D centrato sull'ascoltatore.

L'utente femmina è immersa in uno scenario totalmente virtuale e non completamente tridimensionale, l'utente maschio rappresenta invece uno scenario di sovrapposizione completa tra mondo virtuale e mondo reale.

Coordinate Virtuali

Il caso più comune di un sistema di coordinate virtuali "galleggianti" è quello in cui l'unico punto di ancoraggio in relazione al quale l'evento è localizzato, è la testa dell'utente. Di solito le sorgenti virtuali dei suoni sono renderizzate in differenti direzioni e sono puramente virtuali: minimo grado d'immersione DI e massima deviazione CSD.

Per esempio, servizi di informazioni come le news, le e-mail, gli eventi calendario o altri tipi di messaggi possono essere posizionati nello spazio acustico virtuale attorno alla testa dell'ascoltatore. Eventi calendario sotto forma di messaggi vocali sono renderizzati in differenti direzioni sulla base della tabella dei tempi dell'agenda dell'utente, in modo che mezzogiorno appaia esattamente di fronte.

Le applicazioni di realtà virtuale immersiva usano anche specifiche coordinate, solitamente correlate alla geometria di una scena grafica; nei giochi per pc che usano tecnologie di spazializzazione audio, il sistema di coordinate virtuali è definito in accordo alla scena del gioco e qualche volta combinato con informazioni circa la locazione fisica di un utente come il tracciamento della testa via webcam o simili (PlayStation Move, Xbox Kinect).

La telepresenza è un altro caso di sistema di coordinate virtuali "galleggianti" ed è simile ai sistemi virtuali degli auditory display se focalizzati sull'esperienza d'immersione di un utente. Un caso interessante di mixed reality è la telepresenza[11] bidirezionale aumentata nella quale un segnale di telepresenza neurale è combinato con un ambiente pseudo-acustico. L'MR engine combina l'ambiente locale pseudo-acustico con un ambiente remoto pseudo-acustico col fine di produrre acusticamente l'ambiente dell'altra persona. In questo contesto la deviazione CSD correlata all'ambiente remoto è molto bassa.

In ambienti virtuali collaborativi, Benford et al. ([44]) hanno mostrato che segnali spaziali possono combinarsi con segnali audio-video in modo naturale al fine di aggiungere comunicazione. Il ben noto effetto "cocktail-party" mostra che le persone possono facilmente monitorare molti flussi audio spazializzati alla volta, selezionandosi e soffermandosi selettivamente su quelli

[11] La telepresenza si riferisce a un insieme di tecnologie che consentono ad un individuo di percepire come se fosse presente, di avere l'apparenza di essere presente, o di generare un effetto, tramite telerobotica, in un posto differente dalla sua reale posizione spaziale. La telepresenza richiede che i sensi dell'utente percepiscano degli stimoli che inducano la percezione di trovarsi in un altro luogo, in un altro contesto, anche completamente differente da quello in cui si trova realmente. Inoltre all'utente può essere consentito di operare nella locazione remota, modificando l'ambiente ed il contesto remoto come se stesse agendo dal vivo. In questo caso, la posizione dell'utente, i movimenti, le azioni, la voce devono essere catturati, trasmessi e riprodotti nella locazione remota affinché abbiano realmente effetto, perciò le informazioni devono viaggiare in entrambe le direzioni tra l'utente e la locazione remota e viceversa. Un'applicazione famosa è la telepresence videoconferencing, che consiste nel più alto livello di videoconferenza mai raggiunto grazie all'accuratezza della riproduzione dei segnali video e audio. Avanzamenti tecnologici sia agli apparati di rete che nella collaborazione mobile hanno esteso le potenzialità della telepresenza, e ora consentono di controllare le mani e gli arti da remoto, aprendo le porte a scenari di utilizzo quali quello medico ed ingegneristico.

di maggiore interesse. Nelle teleconferenze multi-party il posizionamento di ciascun interlocutore può essere fatto liberamente in una stanza virtuale.
Walker e Brewster ([45]) hanno esplorato l'uso dell'audio spazializzato nei mobile devices, ad esempio per risolvere problemi di confusione visiva nelle interfacce..

Coordinate Fisiche
Quando si posizionano oggetti audio virtuali in locazioni prestabilite del mondo fisico attorno all'utente, il sistema di coordinate usato per il rendering dei suoni virtuali deve coincidere con una mappa dell'ambiente virtuale. Idealmente bisognerebbe posizionare un oggetto audio virtuale il più possibile vicino a un oggetto fisico nel mondo reale. Messaggi audio localizzati vicini a un'opera d'arte esposta in un museo, così come un'introduzione ad un'esibizione, sono esempi di sistemi di audio guide. Un post-it acustico è collegato a un sistema di coordinate fisiche ben precise; un messaggio audio registrato viene riprodotto nelle orecchie di un visitatore quando costui è in una certa posizione del museo. La posizione e l'orientazione dell'utente sono rilevate ed aggiornate, e le caratteristiche acustiche dell'edificio sono monitorate allo stesso tempo, ottenendo così un alto livello di immersione DI, veicolando un soundscape dinamico tramite l'esecuzione di differenti messaggi sonori riprodotti tramite le cuffie wireless che indossa l'ascoltatore.
Tuti gli aspetti sopra citati sono molto importanti per le applicazioni mobile e i contesti dove non si può utilizzare la vista. I sistemi di aiuto alla navigazione rappresentano un contesto dove si fa un forte uso di queste tecnologie, infatti le mappe degli spazi fisici possono essere globali o locali. Ulteriori due esempi per sottolineare l'importanza dei sistemi di coordinate fisiche sono i display audio virtuali per le battaglie aeree nelle missioni di combattimento simulato e i sistemi di allarme anticollisione per i piloti aerei. In questi ultimi due contesti applicativi il sistema di coordinate associate si muove con l'aeroplano e in entrambi i casi si ottiene un basso livello di deviazione CSD, perciò l'obiettivo critico è il matching tra coordinate virtuali e fisiche.

Riproduzioni di suoni binaurali
Ci sono differenti tecniche per la localizzazione delle sorgenti sonore nello spazio che seguono diversi approcci; una prima distinzione riguarda il metodo di riproduzione del suono, ad esempio l'uso di grossi altoparlanti in contrapposizione all'uso di sistemi basati su cuffie.
Le tecniche binaurali si posizionano nel mezzo tra i due gruppi infatti la riproduzione binaurale può essere ottenuta sia con grossi altoparlanti, sia con cuffie e consente di ottenere un'autentica esperienza sonora se i timpani sono stimolati da segnali che portano sostanzialmente la stessa pressione come avviene nelle condizioni della vita reale.
Due ulteriori approcci che derivano esclusivamente dalla riproduzione tramite grossi altoparlanti sono:

- il tentativo di ricreare l'intero campo sonoro sopra una grande area d'ascolto;
- l'intento di introdurre solo gli elementi di cui ha bisogno il sistema audio al fine di percepire la locazione del suono.

Nonostante ciò, l'uso di cuffie in stretta correlazione con il tracciamento della testa garantisce un notevole grado di interattività, realismo e immersione che non sarebbe altrettanto facile raggiungere con sistemi multicanale o con la sintesi di onde, a causa di limitazioni dello spazio di lavoro dell'utente e di effetti acustici di real listening space.

1.6.2 Localizzazione del suono

Ci sono scenari e ambiti applicativi dove è opportuno considerare i ricevitori di onde acustiche come oggetti puntiformi, ad esempio i microfoni omnidirezionali. In altri casi però è opportuno caratterizzare con precisione forma, dimensione e peculiarità del ricevitore, come avviene per il sistema uditivo dell'essere umano, formato da due orecchie separate dalla testa ("un ostacolo") nel mezzo.
In tutte le considerazioni che vengono svolte di seguito si assume che i due segnali acustici di pressione che arrivano ai timpani dell'ascoltatore, contengano già tutte le informazioni di cui un essere umano ha bisogno per la percezione sonora. Per inciso, se due eventi sonori generano nei timpani gli stessi segnali di pressione, questi saranno percepiti dall'essere umano come lo stesso evento acustico.
Le informazioni spaziali trasportate da questi segnali riguardano la posizione della sorgente sonora in relazione alla posizione dell'ascoltatore.
Risulta dunque evidente come sia opportuno comprendere e simulare le modalità con cui il suono, nel suo percorso dalla sorgente al timpano, sia modificato dalle parti del corpo vicine alle orecchie come torso e spalle, pinna e canale uditivo.

Il campo acustico ai timpani

Gli attributi spaziali del campo sonoro sono codificati in attributi temporali e spettrali della pressione acustica nei timpani, attraverso l'effetto di un filtraggio di tre componenti: testa, orecchio esterno e torso-spalle.

Testa

Le nostre orecchie non sono oggetti isolati nello spazio, al contrario, sono posizionate, alla stessa altezza, sulle due parti opposte di un oggetto acusticamente rigido: la testa. La testa appunto, agisce come ostacolo alla libera propagazione del suono e produce due principali effetti:

1. introduce un ***interaural time difference ITD*** dal momento che un'onda acustica deve percorrere una distanza extra per raggiungere l'orecchia più lontana;
2. introduce un ***interaural level difference ILD*** poiché l'orecchia più lontana è acusticamente "in ombra" a causa della presenza della testa.

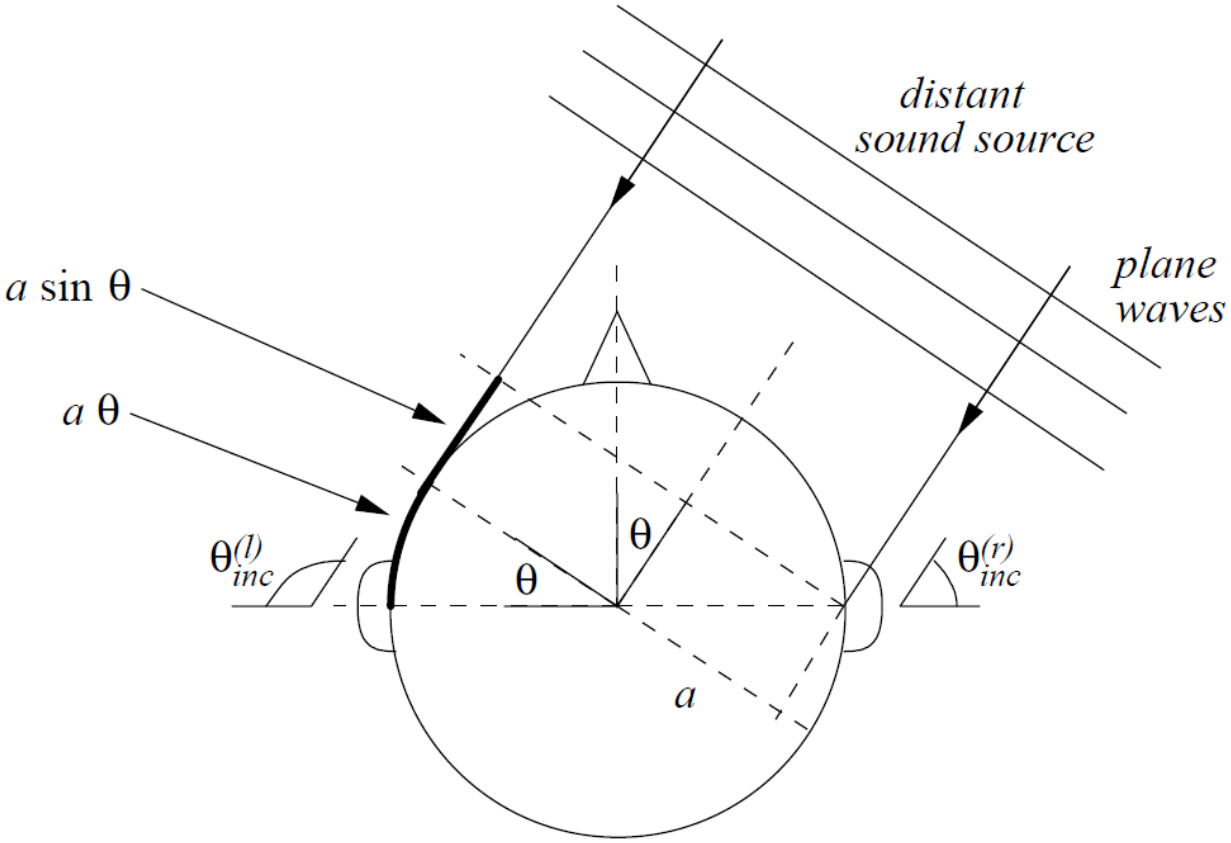


Figura 4 Stima dell'ITD nel caso di una sorgente distante (onde piane) e testa sferica

Una descrizione approssimata ma allo stesso tempo accurata dell'ITD, si può derivare da alcune semplici assunzioni: considerando il caso di sorgenti audio distanti e una testa sferica.
La prima semplificazione implica che le onde che colpiscono la testa sono onde piane, ed in questo modo la distanza extra Δx che deve percorrere un raggio per raggiungere l'orecchia più lontana è calcolata nel modo seguente: $ITD = \frac{\Delta x}{c}$, perciò $ITD \sim \frac{a}{c}(\theta + \text{sen}\theta)$, dove a è il raggio della testa, e θ è l'angolo azimuth che definisce la direzione del suono entrante nel piano orizzontale. Dunque ITD è 0 quando la sorgente è posta dritta davanti (θ=0), e raggiunge il suo massimo a $\frac{a}{c}(\frac{\pi}{2} + 1)$ quando la sorgente è completamente da una sola parte ($\theta = \frac{\pi}{2}$). Un caso realistico è una ITD di 0,6ms con un raggio a = 8,5cm.
Nonostante sia accettabile approssimare l'ITD come un parametro indipendente in frequenza, l'ILD è altamente dipendente dalla frequenza: a basse frequenze per lunghezze d'onda lunghe in relazione al diametro della testa, è difficile apprezzare differenze circa la pressione sonora percepita alle due orecchie, mentre alle alte frequenze le differenze divengono molto significative.
Anche le ILD possono essere studiate nel caso ideale di una testa sferica e di un raggio a, con una sorgente sonora posizionata ad una distanza r > a dal centro della sfera. Se si considera un punto sulla sfera, allora la diffrazione di un'onda acustica vista dal punto scelto è espressa tramite la funzione di trasferimento:

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

Dove *Pm* e *hm* sono rispettivamente il polinomio di Legendre di ordine m e la funzione sferica di Hankel, mentre *ϑinc* è l'angolo di incidenza. L'incidenza normale corrisponde a θinc = 0, mentre il punto della sfera opposto alla sorgente è a θinc = π.

A basse frequenze la funzione di trasferimento non è direzionalmente dipendente e l'ampiezza del modulo di *Hsphere* è essenzialmente unitaria per qualsiasi angolo di incidenza θinc. Quando μ supera 1, la dipendenza da θinc diviene evidente, la risposta aumenta attorno alla parte anteriore della sfera di circa 6dB. Il modulo di *Hsphere* è approssimativamente piatto quando θinc è 100 gradi, e progressivamente decresce attorno alla parte posteriore della sfera. Da notare come la risposta minima non si ha con θinc = π, al contrario, si genera un punto caratterizzato da un effetto "bright spot", dovuto al fatto che tutte le onde che si propagano attorno alla sfera arrivano in fase in quel preciso punto.

A frequenze veramente alte, il lobo del bright spot diviene estremamente stretto, e il retro della sfera si trova effettivamente in una sorta di ombra acustica. Infine, gli effetti di interferenza causati dalla propagazione delle onde in varie direzioni attorno alla sfera introduce increspature nella risposta che risultano prominenti nella parte in ombra.

Orecchio Esterno

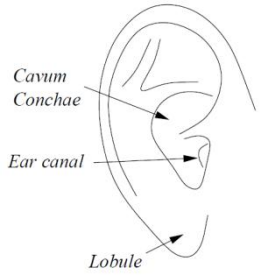


Figura 6 pinna

L'orecchio esterno è composto dalla pinna (la parte più esterna), e dal canale uditivo (che raggiunge il timpano); dietro al timpano ci sono rispettivamente l'orecchio medio e l'orecchio interno. La pinna è caratterizzata da una forma molto simile a un basso rilievo, con peculiarità che differiscono molto da individuo a individuo. La pinna, approssimativamente, può essere descritta come un canale di ampiezza costante, con pareti ad alta impedenza acustica. Dalla parte opposta alla pinna, il canale uditivo termina con la membrana del timpano. Il canale uditivo si comporta sostanzialmente come un risonatore monodimensionale. La pinna, invece, introduce effetti ben più complicati da modellare poiché agisce come un'antenna acustica. Le sue cavità risonanti amplificano alcune frequenze, e le sue caratteristiche geometriche comportano effetti di interferenza che attenuano altre frequenze. Inoltre, la risposta in frequenza dipende dalla direzione. Dal punto di vista acustico, agisce come un filtro la cui funzione di trasferimento dipende dalla distanza e dalla direzione della sorgente sonora in relazione all'orecchio (sinistro o destro).

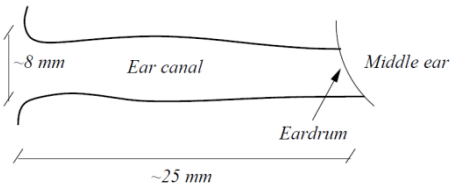


Figura 5 canale uditivo

Primo approccio: orecchio esterno come riflettore di suoni. Ci sono sempre almeno due percorsi dalla sorgente al canale uditivo: un percorso diretto e uno

più lungo caratterizzato dalla riflessione della pinna. A basse frequenze, la pinna essenzialmente accumula energia sonora addizionale e il segnali provenienti dalle due direzioni arrivano in fase. Alle alte frequenze, il segnale in ritardo è fuori fase rispetto al segnale diretto e si genera interferenza distruttiva. Si ha l'interferenza maggiore quando la differenza nella lunghezza del percorso è metà della lunghezza d'onda, e questo produce un "pinna notch".
Poiché la pinna è un riflettore soprattutto per i suoni che provengono dal davanti rispetto che a quelli che giungono da dietro, il notch risultante è più pronunciato per le sorgenti frontali piuttosto che per quelle posteriori, inoltre la lunghezza del percorso cambia con l'elevazione.
I modelli di riflessione utilizzati dipendono dalle dimensioni delle superfici riflettenti, che sono comparabili se non addirittura inferiori rispetto alla lunghezza d'onda acustica. I coefficienti di riflessione dovrebbero essere dipendenti dalla frequenza.

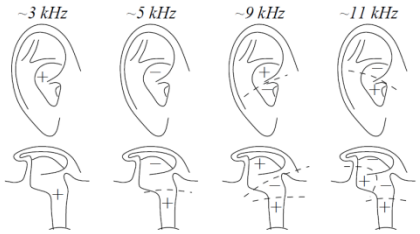


Figura 7 effetti dovuti alla pinna: risonanze

Secondo approccio: modal analysis. Un altro approccio è la modal analysis del risonatore orecchio esterno, tramite misurazioni di risposte in frequenza. La prima risonanza è quella di un canale chiuso-aperto, circa il 33% più lungo del canale uditivo. La pinna agisce come un prolungamento del canale con un effetto di apertura. La seconda risonanza è una risonanza della cavum concha: la distribuzione di pressione è simile a quello che si otterrebbe se il canale fosse tappato. Le altre risonanze, invece, sono associate alle onde longitudinali: queste non sono molto spaziate in ampiezza e dipendono dall'individuo.
La pinna e il canale uditivo formano dunque un sistema acustico di risonatori la cui risonanza dipende essenzialmente dalla direzione e dalla distanza delle sorgenti sonore.

Torso e Spalle

Il torso e le spalle condizionano le onde sonore incidenti in due modi:

- per prima cosa aggiungono riflessioni addizionali che si sommano con il suono diretto (figura a destra);
- in secondo luogo inducono un effetto di "shadowing" per i raggi acustici provenienti dal basso (figura sulla sinistra).

La geometria del torso è abbastanza complicata, tuttavia è possibile derivarne una descrizione semplificata considerando un torso ellittico al di sotto di una testa sferica. Questi tipi di approssimazioni sono spesso chiamate "snowman models".

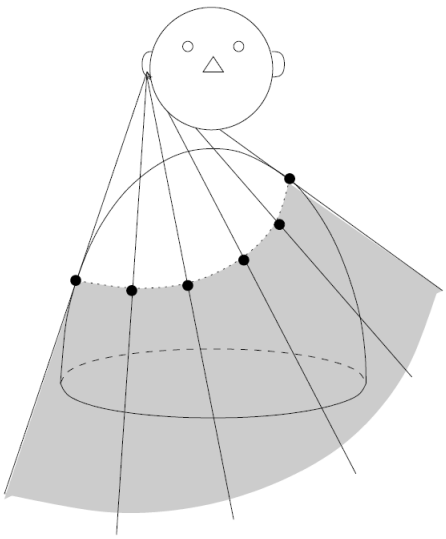


Figura 9 effetti dovuti al torso: shadowing

Se si misura la risposta impulsiva all'orecchio destro, considerata una sorgente in una ben determinata posizione, si può notare come l'impulso iniziale sia susseguito da una serie di impulsi successivi, i cui delay aumentano e poi decrescono con l'elevazione. Questi impulsi addizionali sono causati dalle riflessioni del torso.

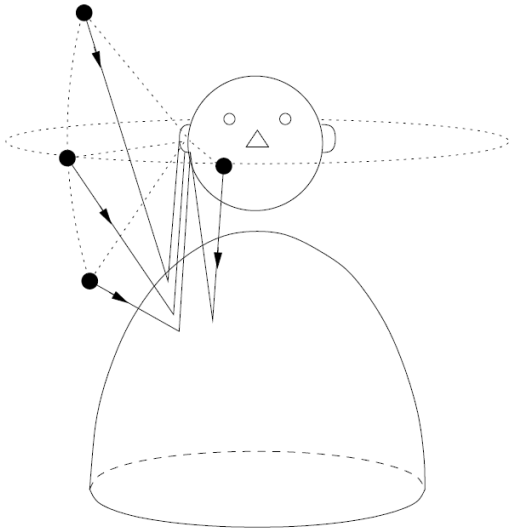


Figura 8 effetti dovuti al torso: riflessioni

Il delay tra il suono diretto e il raggio riflesso non varia molto se la sorgente sonora si muove su una circonferenza situata sul piano orizzontale, specialmente se il suo raggio è maggiore rispetto a quello della testa. In secondo luogo il delay varia considerevolmente se la sorgente sonora si muove verticalmente, in particolare gli impulsi riflessi che subiscono il maggior delay sono quelli delle sorgenti sonore posizionate in alto a destra rispetto all'ascoltatore. Se si considera che la distanza tra il canale uditivo e le spalle è di circa 16cm, allora un raggio riflesso da una sorgente in alto a destra rispetto ad un utente, deve percorrere una distanza extra di circa 32cm che corrisponde a un delay che è all'incirca di 1ms.

Nel dominio delle frequenze le riflessioni del torso agiscono come un filtro combinato che introduce notch periodici nello spettro. Le frequenze alle quali si incontrano i notch sono inversalmente correlate ai delay e così si produce un pattern che varia con l'elevazione della sorgente. Il notch frequency più basso corrisponde al delay più lungo. I delay più lunghi di un sesto di un millisecondo producono uno o più notch al di sotto dei 3KHz, che è all'incirca la frequenza più bassa nella quale si possono notare gli effetti introdotti dalla pinna.

Modellare gli effetti introdotti dal torso come riflessioni speculari significa considerare solo una parte della questione; in primo luogo la riflessione è un concetto correlato alle alte frequenze, in secondo luogo, non appena la sorgente diminuisce di elevazione, si raggiunge un punto di incidenza al di sotto del quale le riflessioni causate dal torso scompaiono, mentre compare un altro effetto collegato al torso che è lo shadowing.

I raggi disegnati dall'orecchio ai punti di tangenza attorno alla parte superiore del torso definiscono un cono di ombra. Lo specular reflection model non si applica al torso shadow cone; invece, la diffrazione e lo scattering producono comportamenti qualitativamente differenti, caratterizzati da una forte

attenuazione alle alte frequenze per lunghezze d'onda comparabili o più piccole rispetto alla taglia del torso.
Nonostante gli effetti introdotti da torso e spalle non siano forti come quelli introdotti dalla pinna, sono ugualmente importanti perché appaiono a basse frequenze, dove tipicamente i segnali sonori hanno la maggior parte della loro energia e dove la risposta della pinna è essenzialmente piatta. In termini di frequenze, i range frequenziali dove agisce il torso sono complementari rispetto a quelli della pinna.

1.6.3 Funzioni di trasferimento correlate alla testa

Tutti gli effetti introdotti sino ad ora sono lineari, e questo significa che:

- possono essere descritti tramite una funzione di trasferimento,
- si combinano additivamente.

In questo modo la pressione sonora prodotta da un'arbitraria sorgente nel timpano è determinata unicamente dalla risposta impulsiva della sorgente nel timpano. Questa è definita ***Head-Related Impulse Response HRIR***, e la sua trasformata di Fourier è chiamata ***Head Related Transfer Function HRTF***. L'HRTF cattura tutti gli effetti fisici che sono stati precedentemente analizzati.
L'HRTF è una funzione che comprende tre componenti spaziali e in frequenza. Considerando l'approssimazione della testa sferica, è consuetudine utilizzare le coordinate sferiche. In questo contesto le coordinate angolari verticali e orizzontali: elevation e azimuth, sono indicate con Φ e θ rispettivamente, mentre le coordinate radiali sono chiamate range e si indicano con r.
In letteratura si utilizzano due differenti sistemi di coordinate sferiche: il più popolare è chiamato ***vertical polar:*** in questo sistema l'azimuth è misurato come l'angolo compreso tra il piano YZ fino a un piano verticale che contiene la sorgente e l'asse Z, e l'elevazione è misurata come l'angolo che raggiunge il piano XY. Con questo sistema le superfici ad azimuth costante sono piane lungo l'asse Z, e le superfici ad elevazione costante sono coni concentrici lungo l'asse Z.
In alternativa c'è il sistema ***intraural polar***: in questo caso l'elevazione è misurata come l'angolo che è compreso tra il piano XY e un piano che contiene la sorgente e l'asse X, e l'azimuth è misurato come l'angolo sul piano YZ. In questo sistema le superfici ad elevazione costante sono piani lungo l'asse X, e le superfici ad azimuth costante sono coni concentrici sempre rispetto all'asse X. Un vantaggio di questo secondo approccio è che rende più facile esprimere le differenze inter-aurali a qualsiasi elevazione, in particolare i coni ad azimuth costante sono i luoghi dei punti che hanno uguali ILD e ITD per una testa sferica.

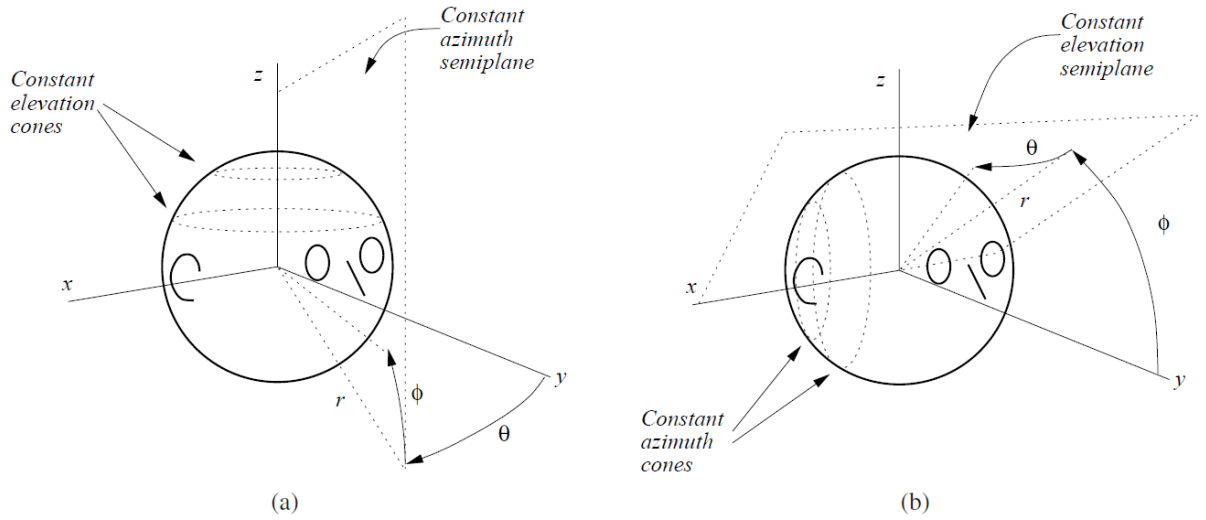


Figura 10 Spherical coordinate systems HRTFs: (a) vertical-polar coordinate system, (b) interaural-polar coordinate system.

L'HRTF si indica con $H^{(l),(r)}(r,\theta,\phi,\omega)$ dove *l* e *r* indicano l'HRTF all'orecchio sinistro e destro, τ tende a +∞ che in pratica accade per τ > 1m. La sorgente è chiamata il campo lontano; in questo caso si scrive: $H^{(l),(r)}(\theta,\phi,\omega)$.
Formalmente si deifnisce l'HRTF ad un'oreccchia come il rapporto dipendnete in frquenza tra il sound pressure level SPL $\Phi^{(l),(r)}(\theta,\phi,\omega)$ al timpano corrispondnete, ed il free-field SPL $\Phi_f(\omega)$ al centro della testa, come se l'ascoltatore fosse assente:

$$H^{(l)}(\theta,\phi,\omega)=\frac{\Phi^{(l)}(\theta,\phi,\omega)}{\Phi_f(\omega)},\qquad H^{(r)}(\theta,\phi,\omega)=\frac{\Phi^{(r)}(\theta,\phi,\omega)}{\Phi_f(\omega)}$$

1.6.4 Percezione della locazione della sorgente sonora

Molti effetti di contingenza e di interferenza possono influenzare la percezione uditiva della locazione della sorgente sonora.

Percezione Azimutale (piano orizzontale)
Il posizionamento orizzontale delle orecchie massimizza le differenze degli eventi sonori che accadono attorno all'ascoltatore rispetto a quelli che avvengono sotto a sopra, permettendo la percezione di sorgenti sonore situate a livello del terreno e al di fuori del campo visivo. ITD e ILD sono considerati i parametri chiave per la percezione azimutale.
Ad esempio, si consideri un'onda sinusoidale che raggiunge l'orecchio sinistro e quello destro. A basse frequenze, l'ITD shifta la forma d'onda di una frazione di ciclo. A livello qualitativo si può dire che se metà della lunghezza d'onda è più grande della dimensione della testa, allora il sistema uditivo può determinare la fase di queste forme d'onda in modo non ambiguo.
D'altro canto però, ad alte frequenze c'è ambiguità nell'ITD perché ci possono essere parecchi cicli di shift. Qualitativamente si può affermare che il punto critico è quello in cui mezza lunghezza d'onda diviene più corta della

dimensione della testa: per lunghezze d'onda più corte l'informazione circa la fase in relazione al tempo di arrivo alle orecchie non può più essere trasferita. Il punto critico in frequenza è di solito assunto essere un valore attorno agli 1.5KHz.
Per quanto riguarda l'ILD la situazione è rovesciata, infatti a basse frequenze la funzione di trasferimento della testa è pressoché piatta, e perciò c'è poca informazione ILD. Al contrario, ad alte frequenze l'ILD è più marcata e può diventare molto estesa.
Per questa ragione la Duplex Theory afferma che **l'ILD e l'ITD sono complementari rispetto alla percezione azimutale** e, considerate assieme, consentono un'azimuth perception anche al di fuori del range frequenziale dell'udibile.
Questo in realtà non è del tutto vero, infatti le timing information possono essere sfruttate per la percezione azimutale anche ad alte frequenze perché le differenze di timing in ampiezza si possono determinare; nuovamente si pensi ad un'onda sinusoidale che è modulata in ampiezza. Si può sfruttare un inviluppo ITD, chiamato anche ***Interaural Envelope Difference IED***, che opera sulla base dell'estrazione delle differenze temporali dell'udito, a partire dall'ampiezza dell'inviluppo, piuttosto che a partire dai timing della lunghezza d'onda che costituisce l'inviluppo stesso.

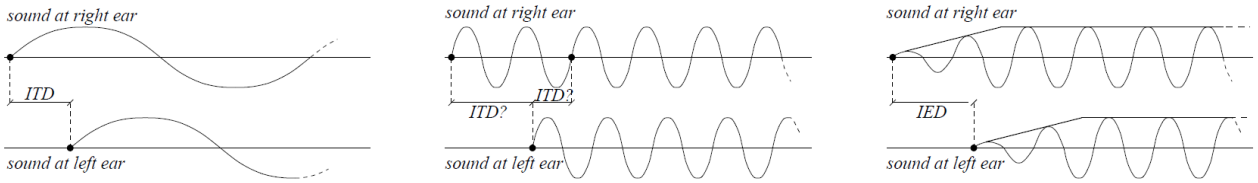


Figura 11 ITD non ambigua, ITD ambigua, IED

Tutto ciò è dimostrato dall'***Effetto Franssen***: se un'onda sinusoidale è improvvisamente fatta partire e viene inviata da un lato ad un altoparlante A tramite un filtro passa-alto, e dall'altro ad un altoparlante B tramite un filtro passa-basso, la maggior parte degli ascoltatori localizza il suono come proveniente solo da A.
Le informazioni fornite da ITD e ILD possono essere ambigue; se si considera la geometria sferica, una sorgente sonora posizionata di fronte ad un ascoltatore ad un certo θ, ed una seconda posizionata dietro, a $\pi - \theta$, forniscono gli stessi valori di ITD e ILD. In realtà ITD e ILD non saranno perfettamente identici perché:

- ✓ la testa umana non è esattamente sferica;
- ✓ ci sono asimmetrie e altre feature facciali;
- ✓ le orecchie non sono posizionate perfettamente ma cadono sotto e dietro l'asse X.

Nonostante ciò i valori saranno molto simili e, di fatto, si osserva spesso una confusione front/back: gli ascoltatori effettuano inversioni quando valutano l'azimuth, ed erroneamente posizionano sorgenti dietro invece di davanti, o viceversa. L'inversione anteriore accade più spesso di quella posteriore. Questa asimmetria potrebbe essere generata da una sorta di meccanismo di sopravvivenza ancestrale secondo cui un predatore, per esempio, poteva

essere udito anche senza essere visto, e quindi con buona probabilità doveva essere alle spalle.
La Duplex Theory opera essenzialmente in condizioni aneconiche, ma nella vita di ogni giorno i riverberi possono seriamente degradare in particolar modo le informazioni ITD. In una tipica stanza, le riflessioni iniziano ad arrivare pochi millisecondi dopo del suono diretto. Sotto una certa frequenza, la prima riflessione raggiunge l'orecchia prima che termini un intero periodo di oscillazione. Ancor prima che l'udito stimi la frequenza dell'onda sonora che sta per arrivare, e conseguentemente inferisca l'ITD, il numero di riflessioni nell'orecchio è cresciuto esponenzialmente e l'orecchio non è più in grado di calcolare l'ITD. Perciò suoni caratterizzati da parecchia energia nel range a bassa frequenza (all'incirca sotto i 250Hz) sono impossibili da localizzare all'interno di un ambiente pieno di riverberi. Al contrario, si può utilizzare l'IED perché i transitori iniziali garantiscono una localizzazione non ambigua, mentre lo steady-state signal è molto difficile da localizzare. Si può dunque concludere che l'energia ad alta frequenza è importante solo per la localizzazione in ambienti pieni di riverberi.

Lateralizzazione ed Esternalizzazione
I sistemi più semplici per il rendering spaziale dei suoni sono basati sulla manipolazione delle interaural cues, e sulla riproduzione degli auditory display in cuffia. Questi sistemi possono essere usati in applicazioni dove solo due dimensioni sono coinvolte, ovvero sul piano orizzontale. In questo contesto, il termine ***lateralizzazione*** è tipicamente usato per indicare un caso speciale di localizzazione dove la percezione dello spazio è "sentita dentro la testa", principalmente lungo l'asse inter-aurale, e per produrre la percezione si opera una manipolazione di ITD e/o di ILD tramite cuffie.
La lateralizzazione costituisce un esempio concreto della contrapposizione tra posizione della sorgente sonora virtuale e quella reale. Quando suoni identici monoaurali sono spediti alle cuffie stereo, l'ascoltatore non sente due suoni distinti provenienti dai trasduttori, ma al contrario, percepisce una singola sorgente sonora virtuale, che sembra essere posizionata al centro della testa. Dal momento che ITD e ILD sono aumentate, la posizione percepita della sorgente virtuale inizia a spostarsi verso un'orecchia, lungo una linea immaginaria. Una volta che si raggiunge un valore critico di ITD o ILD, la sorgente sonora percepita arresta il suo movimento lungo l'asse inter-aurale e viene posizionata in una delle due orecchie. Questo effetto è chiamato ***Inside-the-Head Localization IHL***. Conoscere questo effetto è importante perché la riproduzione tramite cuffia è migliore degli altoparlanti per trasmettere componenti acustiche virtuali in tre dimensioni.
Il raggiungimento dell'esternalizzazione del suono, per esempio la rimozione dell'effetto IHL, è in un certo qual modo il "Santo Graal" dei sistemi di spazializzazione audio basati su cuffie. Non è chiaro quali siano le cues addizionali che maggiormente influenzano il processo di produzione dell'esternalizzazione sonora, tuttavia è stato osservato da molti che l'esternalizzazione cresce quando la stimolazione approssima più da vicino una stimolazione naturale, e che, in particolar modo i riverberi (siano essi naturali o artificiali), possono aumentare drammaticamente l'esternalizzazione.

In generale, l'IHL non è una conseguenza inevitabile dell'ascolto in cuffia, perché suoni esternalizzati possono essere sentiti tramite cuffie in molti casi.

Percezione dell'Elevazione (piano verticale)

Sorgenti sonore posizionate da qualche parte su una superficie conica che si estende al di fuori dell'orecchia sferica, producono valori identici di ITD e ILD. Queste superfici, non a caso, vengono spesso chiamate "coni di confusione", ed estendono il concetto di front/back confusion precedentemente trattato nei riguardi del piano orizzontale.

Quanto descritto riguarda una situazione puramente teorica: in realtà ITD e ILD non saranno mai del tutto identiche sullo stesso cono, a causa delle feature facciali e delle asimmetrie precedentemente menzionate. Ad ogni modo, quando ITD e ILD sono molto simili in due locazioni, si può generare una potenziale confusione circa le posizioni assunte dalle due sorgenti in assenza di altre cues spaziali.

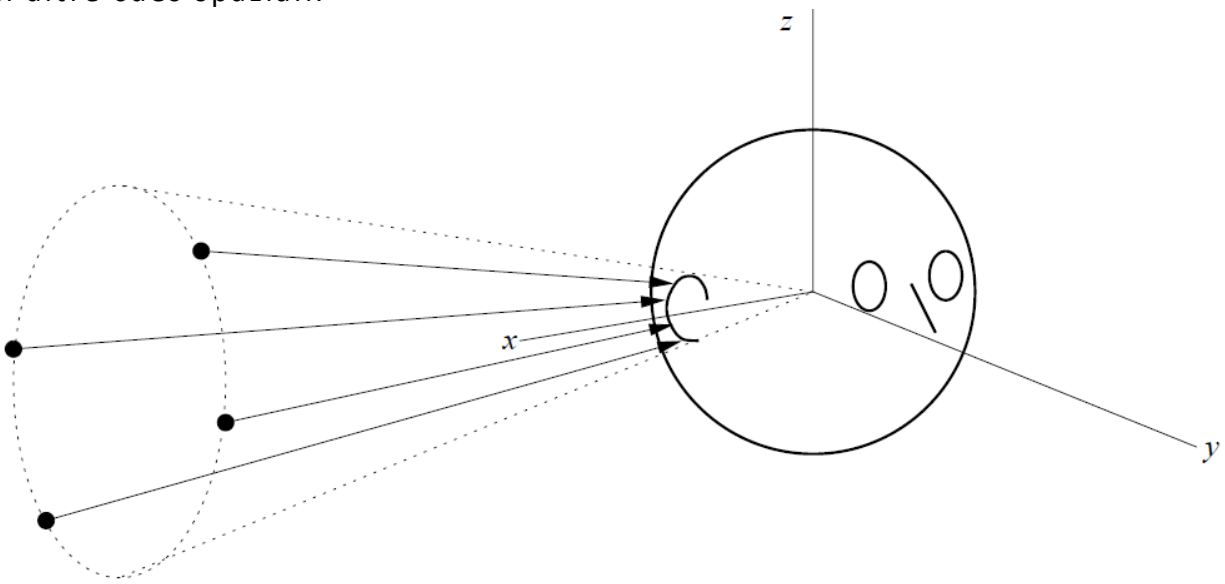


Figura 12 cono di confusione

Gli effetti direzionali della pinna possono rendere meno ambigua la confusione, e sono particolarmente importanti per la localizzazione verticale. Il ruolo della pinna per quanto concerne il miglioramento della localizzazione verticale può essere valutato sperimentalmente, ad esempio comparando le considerazioni fatte sotto normali condizioni, rispetto a quelle fatte in condizioni dove la pinna è bypassata o occlusa.

La localizzazione verticale, in effetti, può essere raggiunta anche quando un'orecchia è completamente occlusa. Questo prova che le cues spettrali prodotte dalla pinna lavorano principalmente in mono-aurale.

Ci sono molte teorie circa il ruolo delle pinna spectral cues, in generale tutte affermano che una cue fondamentale per l'elevazione coinvolge il movimento di spectral notch e di peak, che cambiano in base all'orientazione di una funzione che considera sia sorgente che ascoltatore.

Un modo per apprezzare le pinna spectral cues è quello di esaminare il caso di una sorgente sonora lungo il piano YZ dell'ascoltatore: questo è il luogo dei punti dove non solo ITD e IID sono nulle, ma anche le differenze spettrali tra le HRTF sinistra e destra sono nulle fintanto che la pinna sinistra e destra sono

uguali. Osservando la figura si può vedere uno spectral notch che è importante per la percezione dell'elevazione.

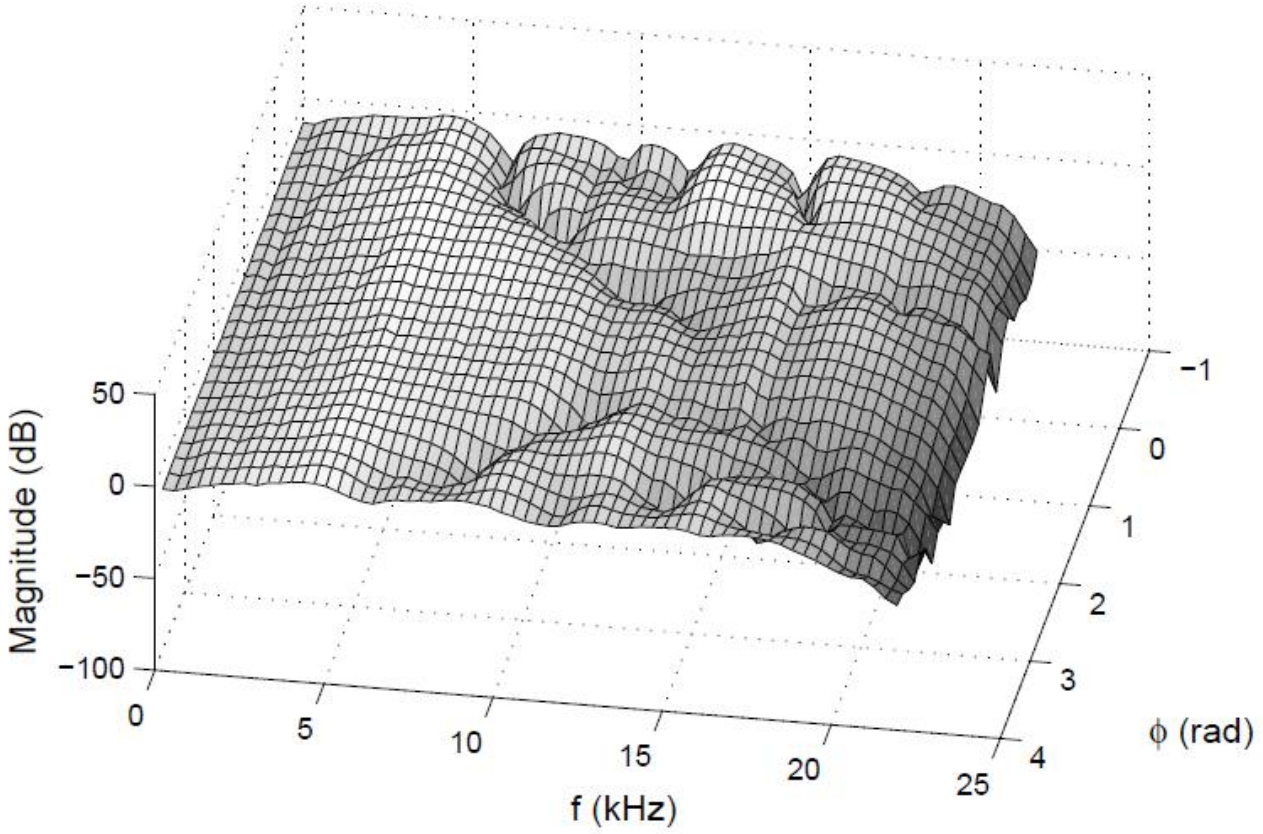


Figura 13 esempio di magnitudine di HRTF

In generale, senza un'estesa valutazione psicoacustica, è difficile capire quanto importanti sono questi cambiamenti e come funzionino in qualità di spatial cues. Non è chiaro se le localization cues derivino da una particolare feature spettrale come un peak o un notch, o dalla forma complessiva dello spettro. Inoltre si considera che una sorgente sonora debba contenere dell'energia nel range delle alte frequenze per avere una stima accurata dell'elevazione, poiché la pinna ha dimensioni limitate nello spazio e le lunghezze d'onda più lunghe della taglia della pinna non sono affette.
Mentre il ruolo della pinna nella localizzazione verticale è stato approfonditamente studiato, il ruolo del torso è meno conosciuto. Il torso disturba le onde sonore incidenti alle frequenze minori di quelle interessate dalla pinna; tuttavia gli effetti del torso sono relativamente negativi.

Percezione della Distanza

In riferimento a quanto evidenziato nei paragrafi precedenti, sembra essere chiaro come la stima della localizzazione azimutale sia abbastanza accurata, sicuramente di più della stima dell'elevazione, e in questo contesto, la stima della distanza risulta essere il task più difficile e impegnativo. La percezione della distanza, infatti, coinvolge un processo di integrazione di cues multiple, i cui contributi sono difficilmente separabili.
In assenza di altre informazioni, l'intensità è la cue primaria usata dagli ascoltatori per stimare la distanza, l'essere umano è portato a imparare dall'esperienza a correlare la dispersione fisica del suono in corrispondenza di

incrementi o riduzioni in intensità. In condizioni aneconiche, la riduzione di intensità sonora con l'aumentare della distanza è controllata dalla legge dell'inverso dei quadrati: ***l'intensità di una sorgente sonora omnidirezionale cala di circa 6dB per ogni raddoppio di distanza***.
Questa legge non è ben motivata a livello percettivo, essa esprime il rapporto dell'intensità di una sorgente rispetto a un livello di riferimento, invece la magnitudine dell'intensità percepita dall'essere umano è un concetto leggermente differente, si chiama ***loudness***. E' perciò preferibile utilizzare un mapping secondo cui, al raddoppio della stima relativa di una distanza, si ha un dimezzamento in termini di loudness, anziché un dimezzamento di intensity. Le due scale così ottenute sono però differenti.
Incrementi di intensità acustica (loudness) possono operare come distance cue solo in assenza di altre informazioni come i riverberi. In presenza di riverberi, infatti, l'intensità acustica complessiva alle orecchie dell'ascoltatore non cambia molto per sorgenti molto vicine o molto lontane: le scale dipendenti dalla distanza si applicano solo al suono diretto dove l'energia riflessa rimane approssimativamente costante.
Il cambio nella proporzione tra energia riflessa e diretta, chiamato ***rapporto R/D***, sembra essere un'ottima cue per la distanza rispetto alla scala di intensità. In particolare, si ha la sensazione di un cambiamento di distanza se l'intensità acustica complessiva rimane costante, ma il rapporto R/D è alterato. In alcuni casi, possibili variazioni al rapporto R/D possono essere limitate dalle dimensioni del particolare contesto ambientale, rendendo tale cue meno robusta, ad esempio in una piccola stanza il rapporto varia molto più rapidamente rispetto ad un ambiente molto grande.
La stima della distanza con stimoli aneconici è di solito peggiore rispetto a esperimenti compiuti in condizioni di riverbero ottimali. Molti risultati sperimentali mostrano una sottostima complessiva della distanza apparente di una sorgente in un ambiente aneconico, e ciò può essere spiegato dall'assenza di riverbero. Il riverbero fornisce la spazialità che permette all'ascoltatore di muoversi dal dominio dell'intensità acustica a quello della distanza, per esempio da un ascolto analitico ad un ascolto di tutti i giorni (everyday listening).
La percezione della distanza è anche condizionata dall'aspettazione (familiarity) della sorgente sonora. Se il suono è completamente sintetico, un ascoltatore tipicamente si focalizza sui cambiamenti parametrici di volume e rapporto R/D. D'altro canto, se la sorgente sonora è cognitivamente associata con un range di distanze tipico, quel range sarà più facilmente percepito rispetto a distanze inaspettate o non familiari. Questo è particolarmente vero per la voce: è facile simulare un mormorio distante 20cm dalle orecchie dell'ascoltatore, rispetto a simulare un'innaturale bisbiglio rumoroso a 10 m di distanza.
Gli effetti spettrali possono anch'essi alterare la percezione della distanza, nonostante abbiano un effetto meno importante rispetto alle cues precedentemente descritte. Le condizioni atmosferiche e l'assorbimento dell'aria sono ulteriori aspetti da tenere in considerazione: all'aumentare della distanza, le alte frequenze di un suono sono fortemente attenuate dall'umidità dell'aria e dalla temperatura. In assenza di altre cues, la forza applicata ad uno

stimolo a bassa frequenza viene interpretata come più distante se comparato ad uno stimolo non trattato.
Un secondo effetto spettrale è prodotto nel cosiddetto ***near field,*** per distanze minori di 1 m, in questo range non è possibile assumere che il fronte d'onda del suono sia planare e l'effetto della curvatura deve essere considerato. Questo fenomeno corrisponde al "darkening" (offuscamento del tono di colore) che accade quando una sorgente sonora si muove molto vicino all'orecchio di un soggetto che ascolta.
Tutte le cues finora descritte sono monoaurali. Una questione ancora aperta è se l'ascolto binaurale possa migliorare la percezione della distanza. Il modello della testa sferica mostra che in questo limite, cioè quello del near field, sia ILD che ITD a basse frequenze sono rafforzate specialmente per sorgenti sonore lateralizzate $\theta \sim \pi/2$. Questo effetto è chiamato ***auditory parallax.*** Grazie a questo effetto, l'accuratezza della stima di un suono proveniente da una parte, dovrebbe essere migliore se comparata alla percezione della distanza sul piano mediano.

Dynamic Cues
Tutte le considerazioni svolte nei paragrafi precedenti sono state effettuate sulla base dell'assunzione implicita di condizioni statiche, con ascoltatore e sorgente non in movimento. Tuttavia, nella percezione di ogni giorno si usano anche dynamic cues in aggiunta a quelle statiche. L'ascoltatore, spesso in modo inconscio ed involontario, si muove e modifica la posizione reciproca tra le sue orecchie e la sorgente sonora. L'essere umano, quando sente un suono e vuole localizzarlo, si muove in modo da minimizzare la differenza inter-aurale, usando la testa come una sorta di puntatore (gli animali, al contrario, usano una pinna mobile per lo stesso scopo).
Quando muovono la testa, gli ascoltatori integrano una combinazione di cambiamenti di ITD, ILD e spectral notch/peak (causati dal movimento della testa stessa nel tempo), e usano queste informazioni per migliorare l'abilità di localizzazione. L'esempio più chiaro è quello della confusione front/back, che è molto comune nei test di ascolto statico, e invece sparisce quando gli ascoltatori sono liberi di girare la testa. Un ascoltatore che sta cercando di localizzare una sorgente, ad esempio a $\theta = 30° \; \Phi = 0°$, probabilmente cerca di centrare l'immagine dell'udito muovendo la testa verso destra. Se il suono diviene molto centrato, significa che la differenza inter-aurale è minimizzata a seguito del movimento della testa, e che il suono proviene dal davanti. Se, al contrario, diviene notevolmente lateralizzato, per esempio arriva più forte e prima all'orecchio destro rispetto che all'orecchio sinistro, allora significa che la sorgente si trova nella parte posteriore, alle spalle dell'ascoltatore.
Le dynamic cues sono importanti anche per l'esternalizzazione. L'IHL è meno probabile che avvenga quando il movimento della testa è permesso, sempre per lo stesso motivo per cui la confusione front/back risulta essere scongiurata: le cues dinamiche derivanti dal movimento della testa sono usate per rendere le localizzazioni non ambigue, rispetto alle condizioni statiche.
Una situazione non desiderabile è quando la scena sonora è veicolata tramite cuffie senza un sistema di tracciamento del movimento della testa o del corpo, e l'ascoltatore è libero di muoversi: in questo caso, le dynamic cues sono

assenti e la scena ruota assieme all'utente, creando disagio e ostacolando l'esternalizzazione.
Quando si aggiungono anche cues visuali, per esempio un utente può muoversi all'interno di un ambiente virtuale immersivo e può vedere la sorgente virtuale del suono, è molto probabile che la combinazione di diverse tipologie di cues possa permettere l'esternalizzazione. In effetti, l'esternalizzazione può avvenire anche quando si sta ascoltando la televisione con un singolo auricolare, questo perché la vista è più affidabile dell'udito per quanto concerne la localizzazione spaziale e perciò il nostro cervello si fida maggiormente della vista piuttosto che del feedback sonoro (per ulteriori dettagli si fa riferimento al capitolo 1 dove si tratta il fenomeno della "cattura visiva").
In fine, il movimento dell'ascoltatore fornisce cues per la percezione della distanza. Una di queste cue è il movimento indotto dal rate di cambiamenti di intensità, chiamato ***acoustic τ***, secondo cui un ascoltatore in movimento in direzione del suono, può ricavare informazioni circa la distanza. Un'altra cue è la ***motion parallax***, che indica il rate di cambiamenti nella direzione angolare derivanti dalla traslazione dell'ascoltatore: per una sorgente molto vicina, un piccolo shift della testa causa grandi cambiamenti circa la direzione angolare, mentre per una sorgente sonora molto lontana il cambiamento è sostanzialmente nullo se confrontato con la quantità di shift effettuato. Il rate di cambiamento di ITD, ILD e degli spectral notch/peak risulta essere influenzato dalla distanza. Questa cue dinamica è simile alla controparte visuale: una sfera grande e lontana e una sfera piccola e vicina sembrano esattamente le stesse, ma se ci muoviamo cambiando la prospettiva, ci si rende subito conto della differenza di distanza che intercorre.

1.6.5 HRTF Generalizzate

Le funzioni di trasferimento correlate alla testa HRTF catturano le trasformazioni che un'onda sonora subisce nel suo percorso dalla sorgente al timpano, tipicamente a causa della diffrazione e della rifrazione sulla superficie della testa, pinna, torso e spalle dell'ascoltatore. Questa caratterizzazione consente un posizionamento virtuale delle sorgenti sonore nello spazio circostante: consistentemente con la posizione relativa alla testa dell'ascoltatore, il segnale è filtrato attraverso la corrispondente coppia di HRTF creando i segnali sinistro e destro che poi saranno riprodotti dalle cuffie.
In questo modo, si possono simulare spazi sonori 3D con un alto livello di immersione e si possono integrare in contesti di mixed reality. Tuttavia registrare individuali HRTFs per uno specifico ascoltatore richiede specifiche attrezzature molto costose e processamenti audio molto sensibili. Questi fattori rendono difficile utilizzare HRTF personalizzati in ambienti virtuali, considerando l'alto costo di altri componenti immersivi come sistemi di tracciamento, display montati sulla testa, e device tattili.
Per queste ragioni si utilizzano HRTFs generalizzate, chiamate anche HRTFs non individualizzate, nonostante si abbiano degli svantaggi tollerabili quali errori evidenti di localizzazione sonora, come la percezione non corretta dell'elevazione della sorgente, inversione anteriore posteriore, e perdita di espressività.

Una serie di esperimenti sono stati condotti da Wenzel et al. ([46]) per comprendere e valutare l'efficacia delle HRTF non individualizzate per display acustici virtuali. Sostanzialmente si ottiene una buona e accurata percezione angolare orizzontale sia nelle condizioni reali che nel rendering di suoni 3D; tuttavia gli esperimenti mostrano che l'uso di funzioni generalizzate aumenta il rate degli errori di inversione. Begault et al. ([47] e [48]) hanno comparato l'effetto di HRTF generalizzate e di HRTF individualizzate, in un sistema dotato di tracciamento della testa e applicando filtri di riverbero a un suono vocale. I risultati mostrano che il tracciamento della testa è cruciale per ridurre gli errori angolari e soprattutto per scongiurare inversioni, mentre la percezione azimutale nell'ascolto tramite generiche HRTF è marginalmente deteriorato se comparato con quello ottenibile da HRTF individualizzate, ed è bilanciato dall'introduzione di riverberi artificiali.
Riassumendo, mentre le HRTFs non individualizzate rappresentano una soluzione economica e diretta per fornire percezione 3D in cuffia, l'ascolto di suoni spazializzati non individualizzati è certamente caratterizzato da errori di localizzazione che non possono essere del tutto controbilanciati dall'aggiunta di segnali spettrali addizionali, specialmente in condizioni statiche. In particolare, l'elevazione non può essere caratterizzata tramite feature spettrali generalizzate. In conclusione, anche la posizione reciproca tra ascoltatore e sorgente sonora, e le caratteristiche antropomorfiche del corpo umano giocano un ruolo chiave nella caratterizzazione dell'HRTF.

1.6.6 3D Sound Rendering

Le tecniche sviluppate dipendono dal tipo di sistema che si intende utilizzare: il tipo di riproduttori (altoparlanti vs cuffie), il loro numero e la disposizione geometrica (sistemi stereo, 5.1, 7.1…).
I sistemi stereo sono i più semplici. Per posizionare un suono a sinistra o a destra il suo segnale deve essere mandato alla cassa corrispondente; se lo stesso segnale è mandato ad entrambi gli speaker, gli altoparlanti si dicono collegati in fase, e l'ascoltatore è all'incirca equidistante da essi. L'ascoltatore percepisce una sorgente fantasma posizionata a metà strada tra i due altoparlanti. Effettuando un cross-fading dei segnali da un altoparlante all'altro, si può veicolare l'impressione del movimento continuo della sorgente tra una cassa e l'altra. Con queste tecniche, tuttavia, la sorgente percepita non si muove mai al di fuori della linea che unisce i due altoparlanti.
I sistemi multicanale sono il livello successivo; l'idea è avere un canale separato per ciascuna direzione desiderata, con l'aggiunta anche delle direzioni sopra e sotto. Gli home-theatre commerciali sono basati proprio su questo concetto. In ambienti pieni di riverberi, un individuo può sfruttare le limitazioni della percezione e usare piccoli altoparlanti sparsi per la stanza, assieme ad un unico grande speaker, ovvero il subwoofer che fornisce i contenuti non direzionali a basse frequenze.
I sistemi basati su cuffie hanno alcuni svantaggi rispetto agli altoparlanti: le cuffie sono invasive e non sempre confortevoli da indossare per periodi di tempo prolungati; hanno risposte in frequenza non piatte che possono compromettere gli effetti di spazializzazione, tendono a veicolare l'impressione di eccessiva vicinanza della sorgente e non compensano il

movimento dell'ascoltatore senza che sia usato un sistema di tracciamento. D'altro canto però, hanno due vantaggi principali: in primo luogo eliminano i riverberi introdotti dallo spazio d'ascolto; in secondo luogo consentono di veicolare i segnali distinti a ciascuna orecchia, e ciò semplifica notevolmente il design del 3D rendering.
Al contrario, i sistemi di altoparlanti soffrono del problema del cross-talk, il suono emesso da un altoparlante verrà sempre sentito da entrambe le orecchie. Se si trascurano gli effetti introdotti dall'ambiente di ascolto, le condizioni di ascolto in cuffia possono essere sostanzialmente approssimate da degli speaker stereo usando tecniche di cancellazione del cross-talk, che tentano di pre-processare i segnali stereo in modo che il suono emesso da un altoparlante sia cancellato all'orecchia opposta a quella verso cui è diretto.
Usando queste tecniche la sorgente fantasma può essere posizionata significativamente al di fuori del segmento che unisce i due altoparlanti, e in particolare gli effetti di elevazione possono essere prodotti senza problemi. Il problema principale è che il risultato dipende da dove si trova l'ascoltatore rispetto agli speaker: la cancellazione del cross-talk si ottiene solo nelle vicinanze dello "sweet spot", una specifica posizione dell'ascoltatore assunta dal sistema.

Rendering basato su HRTF

L'idea generale nei sistemi audio 3D basati su HRTF è di usare HRIR e HRTF opportunamente misurate. Dato un segnale aneconico e una sorgente audio virtuale posizionata a (θ, Φ), i segnali destro e sinistro sono sintetizzati come segue:

- ritardando il segnale aneconico di un quantitativo appropriato di tempo, al fine di introdurre l'ITD desiderata e,
- effettuando la convoluzione tra il segnale aneconico e le corrispondenti risposte impulsive sinistra e destra correlate alla testa.

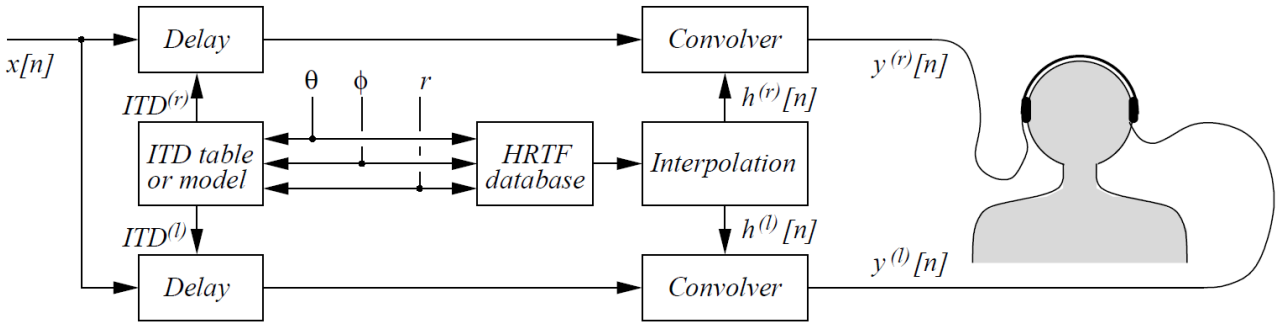


Figura 14 schema a blocchi di un sistema di rendering 3D basato su HRTF

Interpolazione

Le misurazioni delle HRTF possono essere fatte in un numero finito di locazioni, e quando una sorgente sonora deve essere renderizzata in una locazione intermedia, l'HRTF deve essere interpolata. Se non viene applicata l'interpolazione, si generano artefatti udibili come click e rumori nello spettro sonoro nel momento in cui la posizione della sorgente cambia.
Un modo diretto per effettuare l'interpolazione di samples HRIR è il metodo bilineare, che consiste nel calcolare la risposta dato un punto (θ, Φ) come una media pesata delle risposte misurate associate con i 4 punti più vicini. Più

precisamente, se l'insieme delle corrispondenti HRIR è stato misurato sopra una rete sferica con step *ϑgrid* e *Φgrid,* l' $\hat{h}$ stimata di HRIR in un punto arbitrario (θ, Φ) può essere calcolata come:

$$\hat{h}[n] = (1 - c_\theta)(1 - c_\phi)h_1[n] + c_\theta(1 - c_\phi)h_2[n] + c_\theta c_\phi h_3[n] + (1 - c_\theta)c_\phi h_4[n]$$

I parametri c_θ e c_ϕ si calcolano:

$$c_\theta = \frac{\theta \mod \theta_{\text{grid}}}{\theta_{\text{grid}}}, \qquad c_\phi = \frac{\phi \mod \phi_{\text{grid}}}{\phi_{\text{grid}}}$$

Se si utilizzano filtri nella forma pole-zero, si può calcolare l'interpolazione nel modo seguente:

$$H_k(z) = 1 + \sum_{m=1}^{q} b_{k,m} z^{-m} = \prod_{k=0}^{q} (1 - c_{k,m} z^{-1}), \qquad k = 1, 2,$$

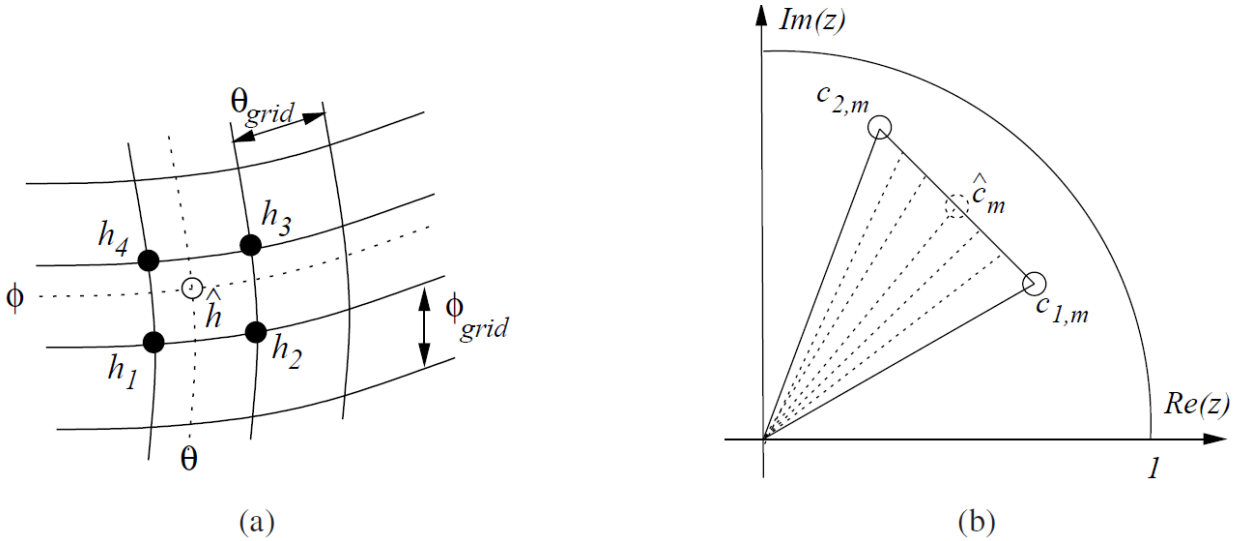


Figura 15 interpolazione bilineare (a), interpolazione pole-zero (b)

Rendering basato su Modelli Strutturali

Contrariamente all'approccio HRTF, l'approccio strutturale è basato sul modellamento degli effetti separati del torso, della testa, della pinna che sono combinati assieme per formare la Head Related Transfer Function.

L'HRTF è poi modellata come combinazione di filtri, ciascuno dei quali considera il contributo di una struttura anatomica. I parametri di ciascun blocco possono essere in principio correlati a misure antropometriche come la distanza inter-aurale o il diametro della concha, con il vantaggio che un modello generico HRTF può essere adattato per uno specifico ascoltatore e può tenere conto di effetti correlati alla postura.

Un altro vantaggio è che gli effetti della stanza possono essere incorporati nello schema di rendering, in particolare le riflessioni provenienti dall'ambiente possono essere integrate e processate con il modello dell'orecchio esterno, sulla base della direzione di ingresso. La scelta del modello della stanza è flessibile rispetto alla specifica applicazione e ha come

obiettivo non solo la riproduzione di caratteristiche di una stanza reale, ma anche aggiungere alcune esternalizzazioni.
Separare gli effetti delle varie strutture anatomiche in filtri perfettamente indipendenti è un'approssimazione euristica che non tiene conto delle interazioni dovute ai fenomeni di scattering tra una struttura e l'altra. Tuttavia, ricerche in questo campo hanno mostrato come i modelli strutturali forniscano buone approssimazioni di HRTF reali.

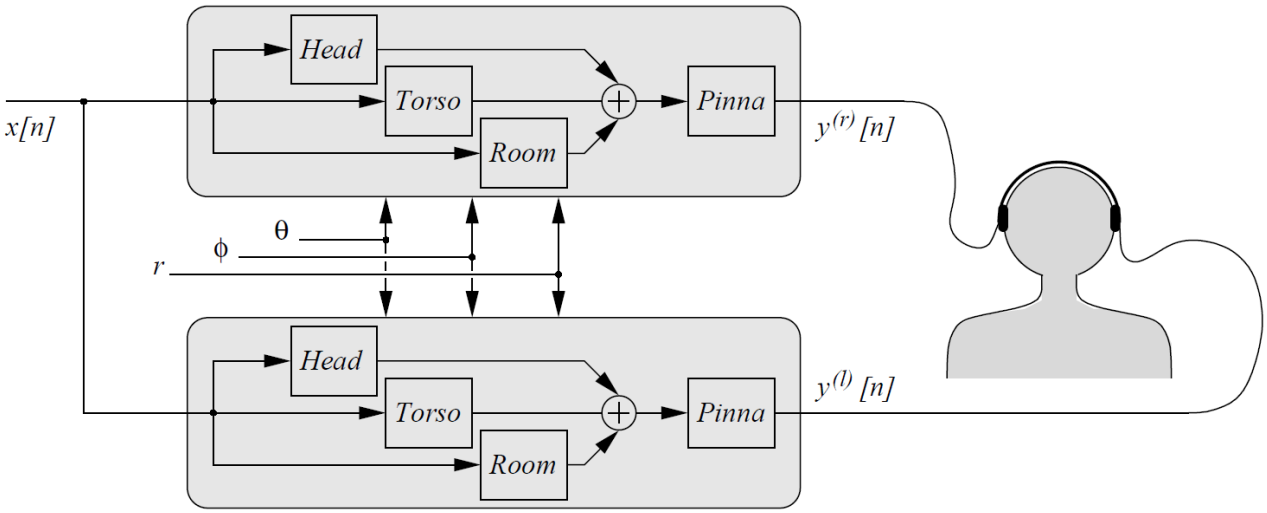


Figura 16 schema a blocchi di un sistema di rendering 3D basato su un modello strutturale

Ai fini pratici, è importante sottolineare che le tecniche discusse richiedono requisiti hardware minimali rispetto a quelli necessari per la manipolazione di video realistici o quelli adottati per realizzare altre tecniche di immersione audio 3D come sistemi multicanale e sintesi a campi d'onde (wavefield).

Modelli per la testa

Un filtro del primo ordine può fornire risultati ragionevoli se ben parametrizzato:

$$\tilde{H}_{\text{sphere}}(\theta_{inc}, \mu) = \frac{1 + \frac{j}{2}\mu \cdot \alpha(\theta_{inc})}{1 + \frac{j}{2}\mu}, \qquad 0 \leq \alpha(\theta_{inc}) \leq 2$$

α controlla la posizione dello zero al numeratore: per α = 2 il filtro da un boost di 6dB alle alte frequenze, che corrisponde al comportamento di *Hsphere* per ϑinc = 0, mentre per α < 1 si genera un effetto passa basso. Inoltre per ϑinc diverso da 0, il parametro α deve dipendere da ϑinc in modo non lineare:

$$\alpha(\theta_{inc}) = \left(1 + \frac{\alpha_{min}}{2}\right) + \left(1 - \frac{\alpha_{min}}{2}\right) \cos\left(\frac{\theta_{inc}}{\theta_{min}}\right)$$

Modelli per le riflessioni del torso e della pinna

Gli effetti principali da considerare circa il torso e la pinna sono le riflessioni. Sia il torso che la pinna sono modellate come filtri FIR, in cui ogni riflessione genera una comb series nello spettro.

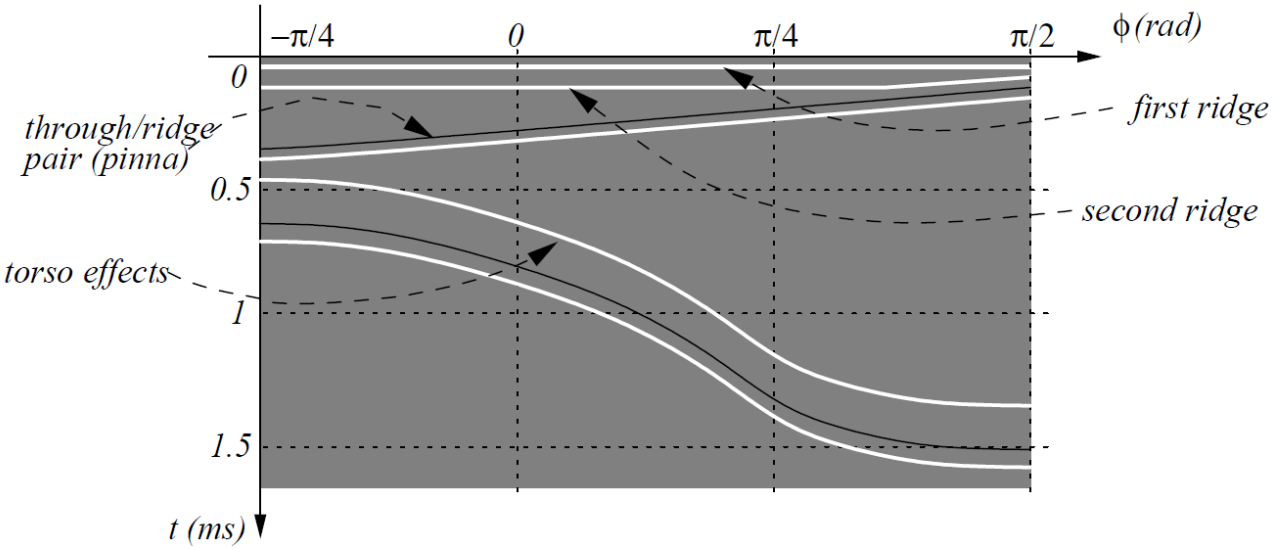


Figura 17 Rappresentazione schematica delle features dell'HRIR sul piano mediano. Le linee bianche e nere indicano ridges e throughs nella risposta

Per poter realizzare un modello per il torso e la pinna, tutto si riduce alla stima dei reflection delays e la loro dipendenza da θ e Φ, sia tramite l'analisi di HIHR / HRTF misurate, sia tramite simulazioni numeriche.

Il delay tra il suono diretto e quello riflesso dal torso è massimo sopra la testa e decresce con l'elevazione, come del resto ci si aspetta dalle considerazioni geometriche.

Gli echo introdotti dal torso variano molto con l'azimuth, al contrario, gli effetti introdotti dalla pinna mostrano una dipendenza dall'azimuth molto ridotta.

Gli effetti introdotti dal torso possono essere modellati con un singolo fractional delay $g^{(t)}F_{\tau^{(t)}}(\theta,\phi,z)$ dove $g^{(t)}$ è il coefficiente di riflessione del torso.

Gli effetti dovuti alla pinna sono più complicati da modellare perché è più difficile estrarre in modo automatico i parametri. In questo caso si procede ad un'analisi nel dominio frequenziale che consiste nell'identificazione di serie di notch nella HRTF.

Il delay $\tau_i^{(p)}(\phi)$ dell'i-esima riflessione della pinna causa periodici notch nello spettro con frequenza:

$$\omega_{i,n}^{(p)}(\phi) = 2\pi(2n+1)/\tau_i^{(p)}(\phi)$$

$$\tau_i^{(p)}(\phi) = 2d_i^{(p)}(\phi)/c$$

$$\omega_{i,0}^{(p)}(\phi) = \pi c/2d_i^{(p)}(\phi)$$

Gli effetti della pinna sono perciò modellabili con un set di n fractional delay filters

$$g_i^{(p)}F_{\tau_i^{(p)}}(\phi,z)$$

Modelli Personalizzati per la pinna

Un approccio strutturale completo richiede la personalizzazione di tutte le componenti introdotte precedentemente, l'attenzione, in questo paragrafo, viene riposta al "pinna block" ([43]). Le onde sonore che arrivano e procedono verso la testa di un ascoltatore, come già evidenziato nei precedenti paragrafi, in primo luogo devono attraversare una distanza extra per raggiungere l'orecchio più distante, e diventano acusticamente in ombra a causa della presenza della stessa testa; le differenze di tempo e livello tra i due segnali sonori che raggiungono le orecchie sinistra e destra sono le quantità binaurali: "ITD internal time difference" e "ILD internal level difference".

Per questo motivo, gli effetti acustici della testa sono modellati tramite delay lines e filtri passa-alto/basso in accordo alle dimensioni della testa dell'ascoltatore. Prima di entrare nel canale dell'orecchio, le onde sonore subiscono altre modifiche spettrali a causa dell'interazione con l'orecchio esterno che agisce sia come riflettore sonoro che come cassa di risonanza.

- ✓ *Riflessioni sopra il profilo della pinna*. In accordo con gli studi di Batteau, le onde sonore sono tipicamente riflesse dall'orecchio esterno finché la loro lunghezza d'onda è sufficientemente piccola se confrontata con la dimensione della pinna. L'interferenza tra le onde dirette e riflesse causa picchi appuntiti che appaiono nel dominio ad alte frequenze dello spettro del segnale ricevuto;
- ✓ *modalità risonanti nelle cavità della pinna*. In linea con quanto sostenuto da Shaw ([49]), poiché la concha[12] agisce come un risonatore, alcune bande di frequenza sia delle onde dirette che di quelle riflesse sono significativamente aumentate: l'amplificazione è correlata all'elevazione della sorgente.

Considerando questi due aspetti l'obiettivo è definire un modello che possa essere facilmente fuso con le molte soluzioni proposte in letteratura circa i blocchi per la testa, il torso le spalle e la stanza.

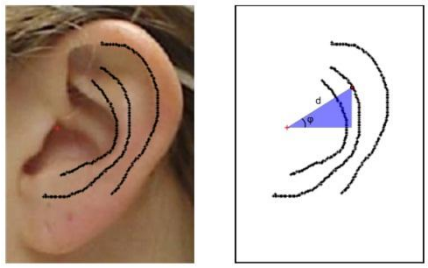


Figura 18 profili della pinna

Fissando la direzione della sorgente sonora in accordo con l'orientazione dell'ascoltatore, le più grandi differenze circa le HRTF di varie persone sono dovute alla forma e al profilo soggettivo della pinna. Il contributo della pinna, chiamato Pinna Related Transfer Function (PRTF) estrapolato dall'HRTF, mostra una serie di peak e notch in ampiezza.

Analisi condotte sulle parti riflesse hanno rivelato che, mentre le PRTFs generalmente mostrano strutture povere di notch quando la sorgente è sopra la testa, non appena l'elevazione cala, la locazione spettrale e la profondità delle notch in frequenza cresce a tal punto che differisce profondamente da soggetto a soggetto, e le loro evoluzioni possono essere correlate direttamente alla locazione dei punti di riflessione sopra le superfici della pinna.

[12] La concha, caratterizzata da un profilo a forma di catino, è la parte della pinna (padiglione auricolare esterno) più vicina al canale uditivo.

Assumendo che il coefficiente di tutte le riflessioni che avvengono all'interno della pinna sia negativo, la distanza extra percorsa dall'onda riflessa rispetto a quella diretta deve essere uguale a metà della lunghezza d'onda per ottenere un'interferenza distruttiva, e questo si traduce in una notch frequency che è inversamente proporzionale a tale distanza.

Modello Strutturale Completo

I componenti precedentemente analizzati possono essere combinati assieme per formare un semplice ma completo modello strutturale. Il significato di tale approccio si basa sul fatto che Il suono può raggiungere la pinna seguendo due percorsi principali: diffrazione attorno alla testa e riflessione dovuta al torso. In entrambi i casi, le onde sonore che raggiungono la pinna sono alterate dalla riflessioni prima di entrare nel canale uditivo.

Si possono introdurre molti miglioramenti al modello, per esempio le riflessioni del torso possono variare con l'elevazione. In questo modello i suoni diffratti dalla testa e i suoni riflessi dal torso sono processati tramite lo stesso modello della pinna. Questo non è interamente corretto poiché gli echi del torso arrivano all'orecchio da una direzione differente rispetto alla direzione del suono. D'altro canto però la rilevanza a livello percettivo delle riflessioni del torso non è chiara, perciò questa descrizione approssimata può essere considerata accettabile.

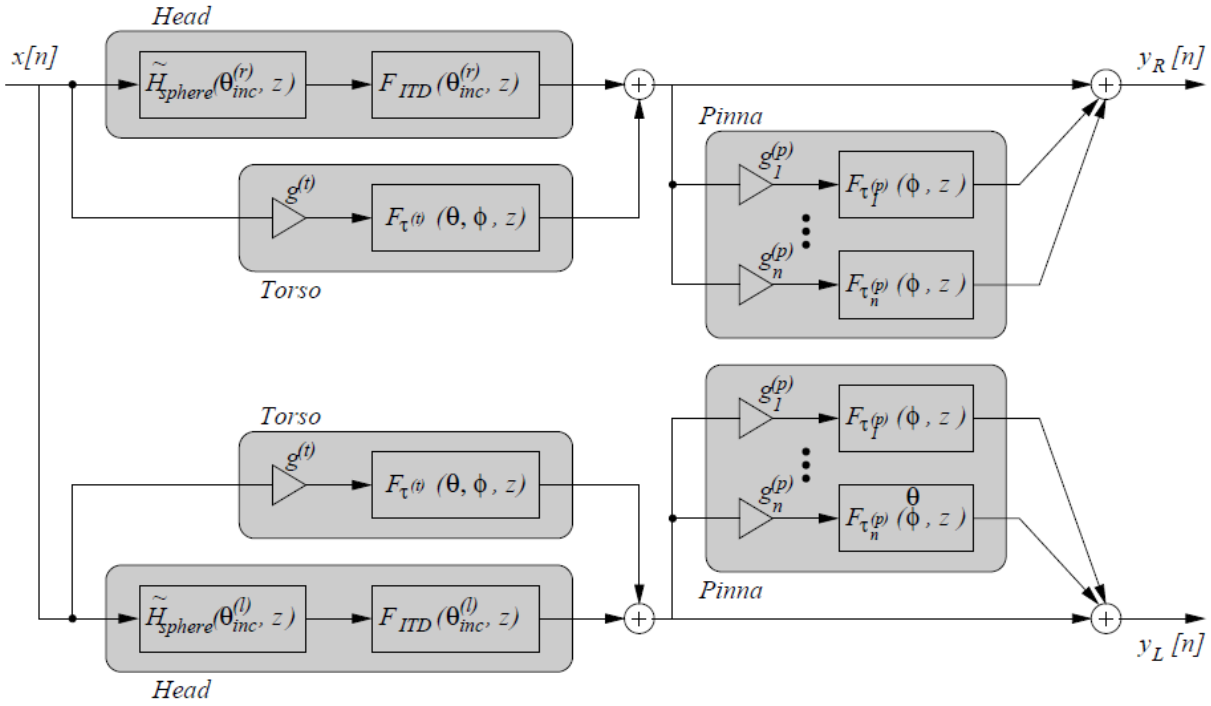


Figura 19 un modello strutturale completo

Approccio computazionalmente efficiente

Il modello sopra proposto ha come obiettivo quello di evitare computazioni costose e step temporali come, ad esempio, l'interpolazione delle HRTF su differenti locazioni spaziali, l'individuazione della HRTF non individualizzata

che produce il miglior fitting, o l'aggiunta di locazioni lontane artificiali, consentendo così un'implementazione in un ambiente real-time.
In tale contesto è opportuno evidenziare un'assunzione fondamentale: l'elevazione e l'azimuth sono tra loro ortogonali così i corrispettivi contributi sono separati in due parti distinte. Il controllo verticale è associato agli effetti acustici relativi alla pinna e quello orizzontale è delegato alla diffrazione della testa.
In effetti, un'osservazione informale delle differenti HRTF, rivela che il piano mediano di riflessione e i pattern di risonanza di solito variano molto lentamente quando il valore assoluto dell'azimuth è in aumento, specialmente fino ai 30°. Questa approssimazione permette di definire un'elevazione personalizzata e "cues azimutali" che mantengono il loro comportamento medio attraverso l'emi-sfera frontale.
L'elevazione della sorgente ϕ è l'unico parametro indipendente usato dal pinna block e consente la valutazione delle funzioni polinomiali dei peak e dei notch spectral form. Solo la notch center frequency è personalizzata sul profilo della pinna dell'utente, e perciò il corrispondente polinomio deve essere calcolato offline.

Le estensioni richieste per avere una completa esperienza surround binaurale portano verso un modello che considera le posizioni delle sorgenti dietro, sopra e sotto l'ascoltatore. L'immersione crescente del corpo dell'ascoltatore richiede l'inclusione dei contributi delle spalle e del torso che coinvolgono ulteriori pattern di riflessione e effetti di shadowing al modello, specialmente quando la sorgente è sotto l'ascoltatore. Nonostante ciò, questo primo stage d'implementazione è comunque in grado di garantire un concreto controllo 3D di una sorgente sonora in numerose applicazioni frontali come gli schermi sonificati.

Per maggiori informazioni sull'Audio 3D si fa riferimento a [51] Federico Avanzini, Sound in Space - Chapter 4 – dispense di informatica musicale UniPD, par 4.5-4.6.

Capitolo 2

SOUNDING ARM – architettura del sistema

SoundingARM è un'applicazione nata per supportare le persone affette da minorazioni più o meno accentuate della vista; può essere sfruttata per scoprire velocemente tutti gli elementi che caratterizzano gli ambienti familiari, per esempio può essere di aiuto a pazienti che hanno subito incidenti e tornano dopo un lungo ricovero in ospedale, oppure può essere utilizzata per esplorare in modo rapido e completo un ambiente sconosciuto come una stanza d'albergo, un appartamento di villeggiatura o altro ancora.
Le informazioni veicolate tramite i suoni sono molto importanti per sviluppare un vero e proprio senso di orientamento spaziale, utile per comprendere anche le distanze e gli ostacoli che si trovano nell'ambiente circostante. Molte persone con problemi di vista sono familiari ai sistemi basati su sintesi vocali o suoni contestuali come alerts, che fanno uso di informazioni associative oppure anche suoni realistici come auditory icons.
Dal punto di vista operativo, SoundingARM richiede, in prima istanza, una fase di installazione effettuata da personale qualificato il quale procede alla disposizione della Kinect all'interno della stanza, in modo che il sistema nel suo complesso sia posizionato nella locazione più opportuna; successivamente viene creato il file di configurazione, necessario all'applicazione per riconoscere i mobili e gli oggetti presenti nell'ambiente circostante. Di seguito si connette e si abilita la Kinect in modo che sia sempre pronta a catturare lo scheletro dell'utente non appena costui apre la porta ed inizia a puntare gli oggetti che lo circondano.

2.1 Hardware

Dal momento che l'applicazione fa uso principalmente dei dati provenienti dal sensore Microsoft Kinect, è importante che i requisiti hardware siano almeno quelli richiesti dal software Kinect SDK, evidenziati nelle appendici relative alla Microsoft Kinect.
Ai fini pratici, 3 sono state le piattaforme hardware sulle quali è stato installato e testato il sistema. Le caratteristiche sono le seguenti:

Configurazione 1

CPU	Intel Core i7 2630QM: 2.3Ghz, 4core/8thread, 8mb cache
Scheda grafica	Nvidia Geforce GT540, 96 stream processors, 128bit bus, 2Gb VRAM
Hard disk	SSD Samsung 830 256gb sata III
RAM	8Gb DDR3 Corsair 1600Mhz cl8
OS	Windows 7 Home premium 64-bit

Configurazione 2

CPU	Intel Core 2 Duo T7700: 2.4Ghz, 2core/2thread, 4mb cache
Scheda grafica	Nvidia Quadro FX 1600M, 32 stream processors, 128bit bus, 512Mb VRAM
Hard disk	Seagate Momentus XT 500Gb 7200rpm
RAM	4Gb DDR2 Corsair 800Mhz cl4
OS	Windows 7 Home premium 64-bit

Configurazione 3

CPU	Intel Core i5 2500k: 4.8Ghz, 4core/4thread, 6mb cache
Scheda grafica	Nvidia Geforce GTX 470, 448 stream processors, 320bit bus, 1.28Gb VRAM
SSD	Corsair Force GT 120Gb sata III
RAM	16Gb DDR3 Corsair 1600Mhz cl8
OS	Windows 7 Home premium 64-bit

2.2 Software

SoundingARM richiede essenzialmente:

- ✓ il Microsoft Kinect SDK (versione 1.0 o 1.5, esistono entrambe le versioni del codice),
- ✓ il software open source Pure Data Extended 0.42.5.

Figura 20 screenshot del setup di SoundingARM

La parte dell'applicazione relativa all'interfacciamento con la Kinect, il riconoscimento degli oggetti, la realizzazione del pacchetti OSC e l'invio tramite UDP è scritta in C++ tramite Microsoft Visual Studio 2010 che sfrutta il .Net Framework 4.0.

2.3 Architettura del sistema

- **Componenti di input**
  il sensore Kinect viene connesso tramite cavo USB al computer sul quale sono installati Windows 7 e il kinect SDK, ovvero le infrastrutture software che si occupano della gestione del riconoscimento e il tracciamento dello scheletro dell'utente.
  Per rendere consapevole l'applicazione e il sistema stesso della presenza degli oggetti, è necessaria l'introduzione di un file di configurazione (estensione .CONF) che include una serie di informazioni circa gli oggetti presenti e la loro posizione rispetto al sensore (che si suppone fisso).
- **Componenti di processamento**
  SoundingARM server application analizza lo scheletro dell'utente, ed è in grado di comprendere quale oggetto l'utente sta indicando.
  La patch Pure Data nella prima implementazione (quella con sintesi vocale text to speech), svolgeva il ruolo di "intermediario" tra l'applicazione e il server della sintesi vocale; nelle versioni successive, invece, la patch costituisce il "motore audio" che si occupa della riproduzione di tutti i suoni sui differenti canali audio.
- **Componenti di output**
  Nell'implementazione con la sintesi vocale si fa uso del server di sintesi vocale che sfrutta l'API Microsoft Speech, al contrario, nelle successive implementazioni, l'output è fornito direttamente dalla patch Pure Data che controlla gli speaker e le cuffie (adottate nel caso della spazializzazione binaurale).

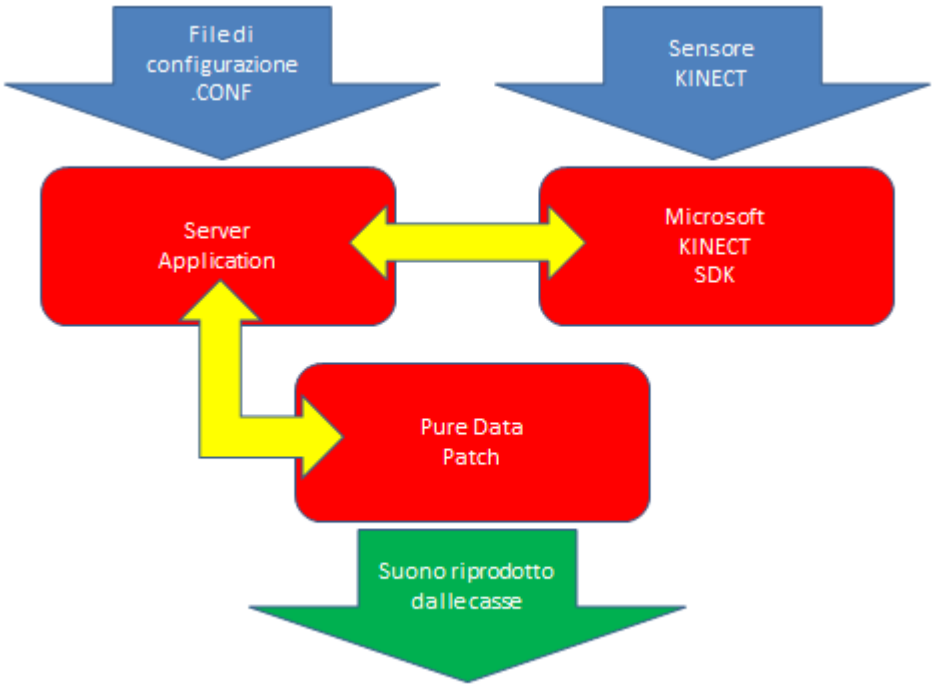


Figura 21 schema a blocchi dell'applicazione SoundingARM

Scendendo più nel dettaglio, Il file di configurazione contiene tutte le informazioni riguardanti gli oggetti presenti all'interno della stanza. Per esempio un frigorifero (un tavolo, una tv) è definito tramite il parallelepipedo che contiene l'oggetto stesso. Tale parallelepipedo ha i quattro vertici di base che si trovano sul piano del pavimento, l'altezza coincide con l'altezza dell'oggetto, e l'eventuale offset verticale coincide con l'altezza dell'oggetto dal pavimento.
E' opportuno precisare che I punti di base devono essere presi considerando il sistema di riferimento del sensore Kinect, inoltre nel file .CONF vanno inseriti tutti gli oggetti relativamente grandi e fissi, in modo da fornire all'utente un'idea dello spazio, cosicché possa discernere gli elementi di interesse dagli ostacoli; tutti gli oggetti che si muovono, come ad esempio le sedie, sono escluse dalla mappa, dal momento che l'applicazione non è in grado di gestire spostamenti dinamici degli oggetti o shift degli stessi rispetto alla configurazione iniziale. Modifiche alla disposizione spaziale degli oggetti della stanza richiedono, pertanto, modifiche corrispondenti ai relativi file di configurazione.
La parte principale di SoundingARM è costituita dall'application server, ovvero la componente software che si occupa di analizzare il file di configurazione, inizializzare una struttura dati che rappresenta tutti gli oggetti della stanza e successivamente chiudere il file .CONF, perciò i dati di configurazione non vengono modificati sin tanto che l'applicazione non viene terminata.
Durante il funzionamento, il programma traccia ininterrottamente la presenza di un utente, calcola la vista prospettica della stanza che l'utente ha di fronte a sé sulla base della posizione della testa, e controlla se l'utente sta indicando qualcosa oppure no. In caso affermativo, l'application server manda alla patch Pure Data il comando di apertura del file audio che corrisponde all'oggetto indicato, e la patch semplicemente inizia la riproduzione del file tramite il riproduttore audio adeguato.
La patch Pure Data in primo luogo agisce come interfaccia grafica in grado di controllare l'esecuzione dell'applicazione, in secondo luogo nella versione con sintesi vocale interfaccia il server application con il text to speech server tramite pacchetti OSC, mentre nelle successive implementazioni si occupa direttamente della riproduzione ed elaborazione di tutti i suoni tramite casse o cuffie binaurali.

Dettagli Implementativi

2.3.1 Componenti di input

```
// Estratto di un file .CONF
        [open frigorifero.wav]
        p1 = (-2510, -470);
        p2 = (-3110, -470);
        p3 = (-2510, 130);
        p4 = (-3110, 130);
        heightOfObject = 2500;
        heightFromFloor = 0;
```

Figura 22 esempio di file di .CONF

Il file di configurazione è caricato dall'utente durante la fase di avvio dell'applicazione. Per ciascuna stanza in cui viene utilizzato il software SoundingARM, si utilizza un file di configurazione differente. Il file deve seguire un enconding UTF-8, presenta una struttura sintattica ben precisa, e non deve includere errori in quanto deve poi essere analizzato dal parser dell'application server.
Per ogni oggetto mappato, innanzitutto si evidenzia il comando racchiuso tra parentesi [] contenente il prefisso "open" seguito dal nome del file audio corrispondente all'oggetto che deve poi essere aperto ed eseguito da Pure Data, ad esempio [open frigorifero.wav].
Vi sono poi gli attributi caratterizzanti dell'oggetto, inseriti sotto come coppia "chiave = valore"; in particolare si specificano i 4 vertici di base del parallelepipedo virtuale che racchiude l'oggetto ovvero p1, p2, p3 e p4; i valori tra parentesi (x,z) sono le coordinate che indicano la distanza dal sensore, rispettivamente lungo l'asse X e lungo l'asse Z, per esempio p1 = (-2510, -470); indica che p1 si trova a 2.51m a destra e 47cm dietro il sensore, le misure sono tutte passate in mm.
L'attributo heightOfObject è l'altezza dell'oggetto (2,5m per quanto concerne il frigorifero), invece l'attributo heightFromFloor indica l'altezza dell'oggetto dal pavimento (in questo caso 0 perché il frigorifero poggia per terra). In generale però oggetti come quadri, mensole, ripiani, televisori, orologi sono vincolati alla parete e non si trovano perciò a contatto diretto col pavimento e hanno una heightFromFloor diversa da 0.
L'elemento jolly [kinect] specifica l'altezza del sensore rispetto al pavimento, e per ottenere le migliori performance, questo valore deve essere il più possibile preciso.

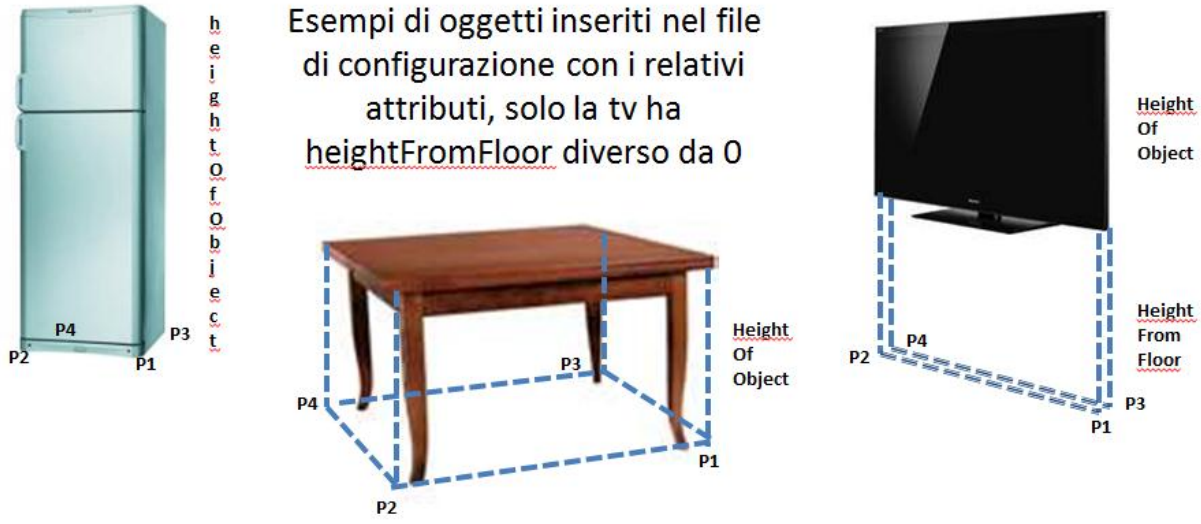


Figura 23 esempi di oggetti che compongono le stanze

2.3.2 Componenti di processamento

SoundingARM server application riceve in input il file con le informazioni riguardanti gli oggetti che compongono la stanza. Tali informazioni sono analizzate tramite un parser e salvate in RAM in una struttura dati personalizzata costituita dalle classi Room e Furniture. Una stanza (room) è

composta appunto da mobili/oggetti (furniture), mentre una Furniture rappresenta l'oggetto con gli attributi sopra descritti.

Creazione della stanza virtuale

Dopo aver effettuato il parsing del file di configurazione, SoundingARM conosce tutte le posizioni degli oggetti mappati. Quando un utente è identificato dal sensore, per comprendere con precisione quello che costui sta indicando, l'applicazione simula la vista prospettica dell'utente.

L'idea di fondo è quella di creare un piano di proiezione che mappa in un array bidimensionale le viste angolari che un utente ha circa un oggetto. ProjectionSphere è la classe che realizza tutto ciò, ed il richiamo alla sfera non è casuale, in quanto il piano si potrebbe pensare come una sfera che circonda la testa dell'utente e ha come raggio l'estensione del braccio. L'angolo solido che sottende l'oggetto è fattorizzato nelle sue componenti verticali e orizzontali.

Per maggiori dettagli circa l'implementazione della ProjectionSphere si rimanda ([52]) alla tesi di Nicola Scattolin: "A comparison between gesture tracking models and the development of an interactive mobility aid system for the visually impaired", in particolare al paragrafo 8.5.2.

Creazione della ProjectionSphere

Dopo che tutti gli angoli sono stati calcolati, un oggetto è rappresentato come una coppia (αxmin, αxmax) (αymin, αymax), che sono rispettivamente i bound orizzontale e verticale di una vista angolare dell'oggetto considerato. Lo spazio dei valori è compreso tra 0° e 180°. Questi valori sono usati per riempire un array bidimensionale, una matrice 180x180 di interi che rappresentano appunto la ProjectionSphere, ovvero i dati che simulano la vista della stanza secondo la prospettiva dell'utente.

Gli indici dell'array bidimensionale corrispondono alla vista angolare dell'utente da sinistra a destra in riga da 0 a 180, e dal basso all'alto in colonne da 0 a 180.

Ogni oggetto definito nel file di configurazione ha un numero identificativo ID, assegnato secondo l'ordine sequenziale crescente con cui viene letto il file di configurazione.

L'Esecuzione di SoundingARM

Se viene identificato un utente, il sensore Kinect inizia a tracciare la posizione della testa e calcola la ProjectionSphere secondo quanto detto in precedenza. Il calcolo utilizza parecchie risorse di CPU, inoltre l'utente potrebbe non rimanere fermo e muoversi di continuo.

Per ridurre il tempo di CPU occupato dal calcolo continuo della ProjectionSphere, sono considerati solo gli spostamenti della testa superiori a 10cm. Questo significa che l'utente può muoversi e camminare attorno, ma l'aggiornamento della proiezione prospettica della vista non avviene per i movimenti inferiori ai 10cm.

Questa misura costituisce un buon trade-off tra l'accuratezza del sensore e l'errore medio nella stima della profondità.

Durante il normale funzionamento, l'applicazione cerca di ottenere continuamente informazioni circa la testa, la mano sinistra e destra dell'utente. H è la posizione dell'utente, M la posizione della mano. SoundingARM calcola la grandezza dell'angolo solido generato dal vettore HM, vettore che unisce appunto la testa e la mano dell'utente. Tale vettore ha due componenti angolari: quella orizzontale, la cui magnitudine varia con la vista prospettica da sinistra a destra dell'utente, e quella verticale, la cui ampiezza varia dal basso all'alto.
Non appena entrambe le componenti sono conosciute, queste sono usate come indici per accedere alla matrice di ProjectionSphere che restituisce l'ID dell'oggetto puntato se e solo se l'utente sta indicando qualcosa, altrimenti restituisce -1.

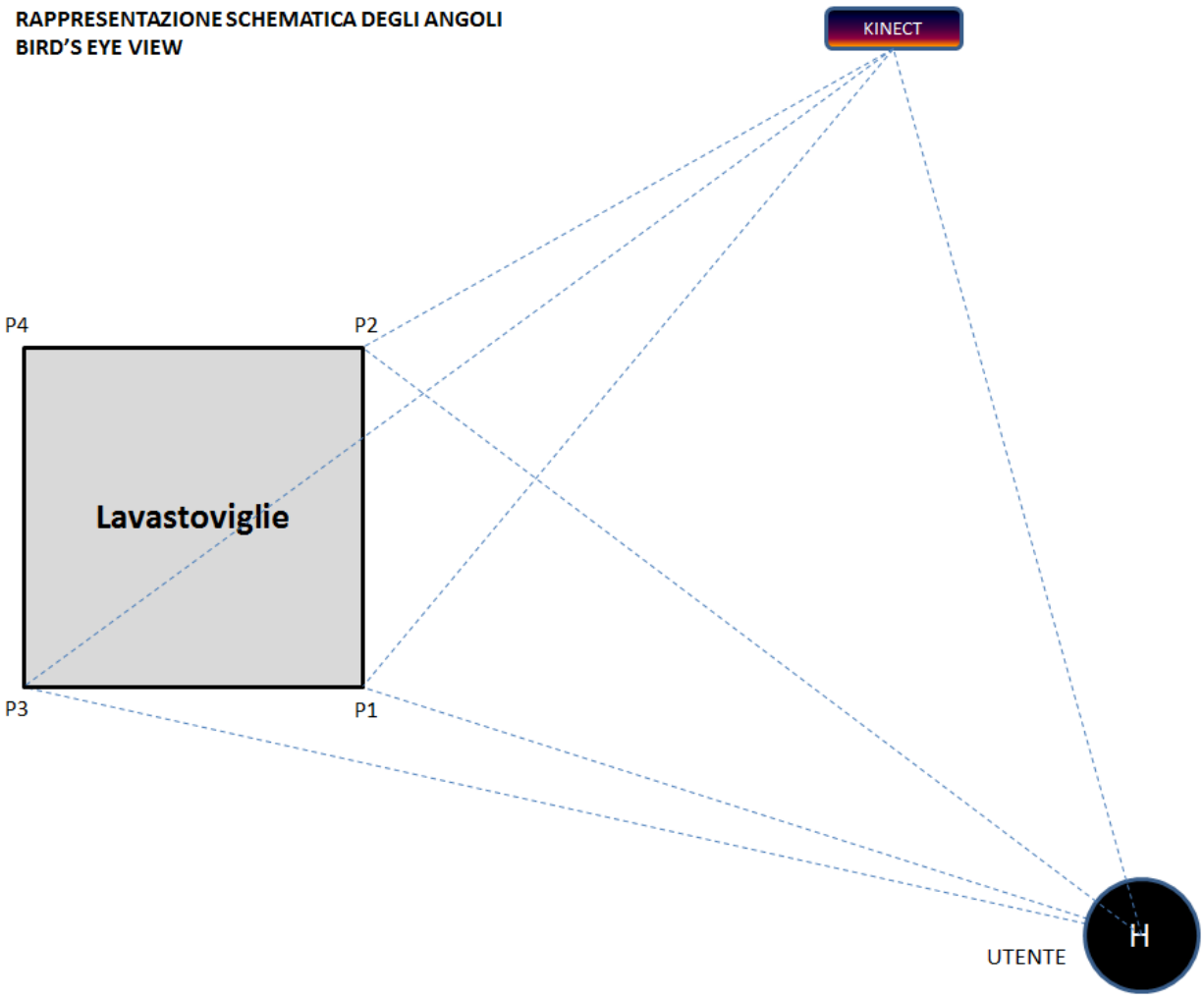


Figura 24 visione dall'alto: bird's eye view

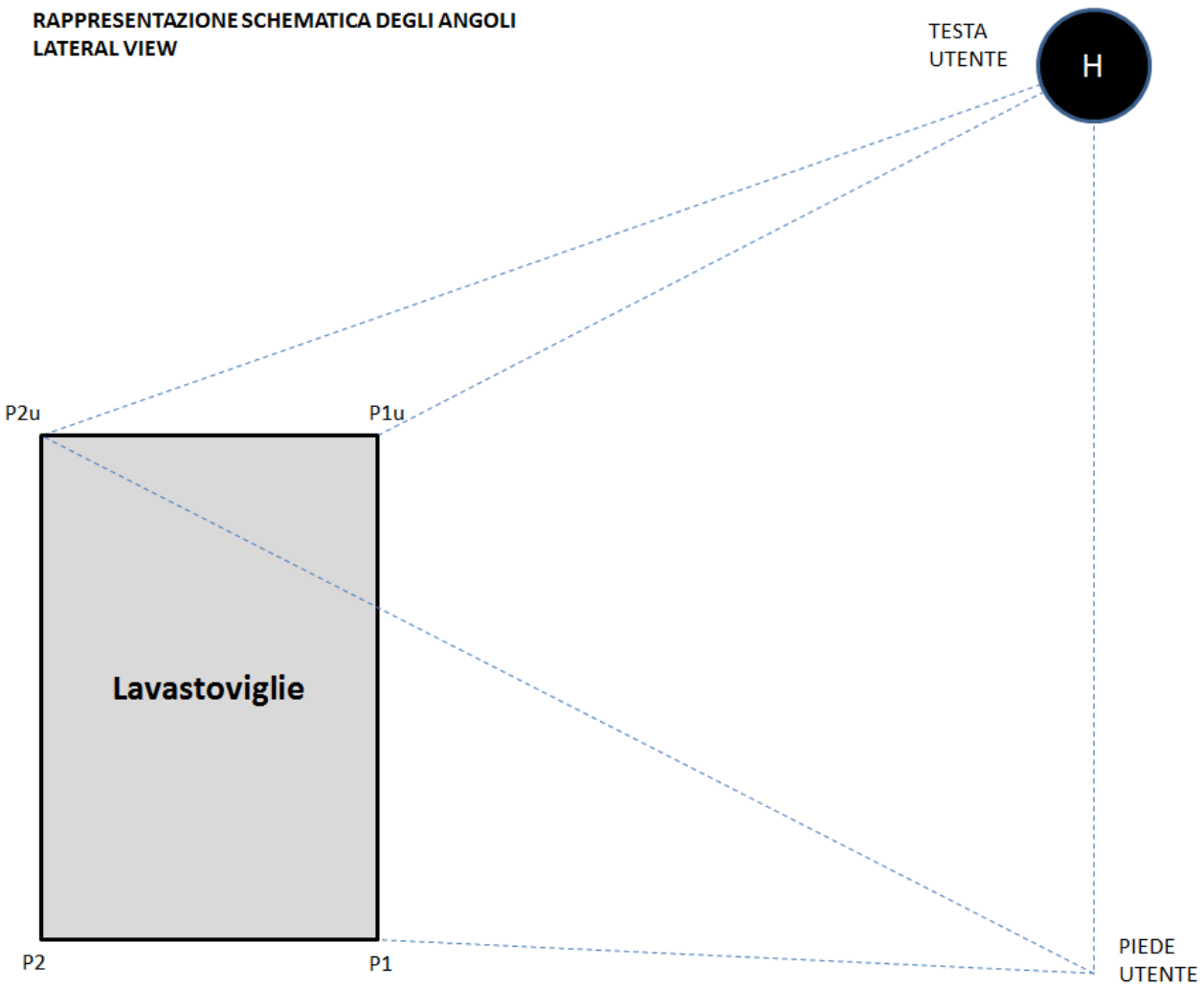


Figura 25 visione laterale: lateral view

Nella figura seguente, un soggetto è situato davanti ad una lavastoviglie, egli la vede sotto un certo angolo, la cui ampiezza è in funzione della sua posizione rispetto alla posizione della lavastoviglie. Questo angolo può essere fattorizzato in una componente orizzontale e in una verticale. Il piano verticale e orizzontale sono piani prospettici immaginari che definiscono l'orientazione della vista dell'utente. Nell'esempio la lavastoviglie ha un angolo solido racchiuso orizzontalmente tra [75°, 105°] e verticalmente tra [55°, 85°].

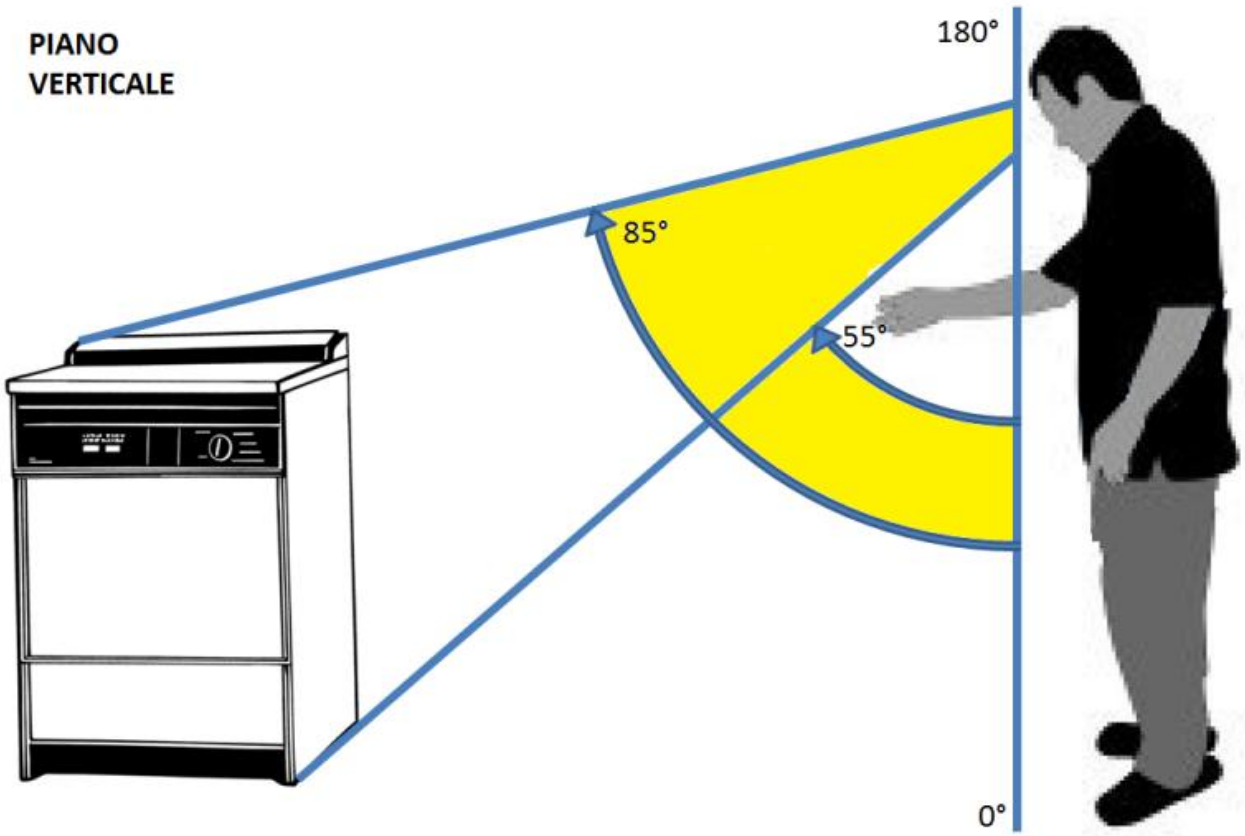


Figura 26 angoli coinvolti, visione laterale

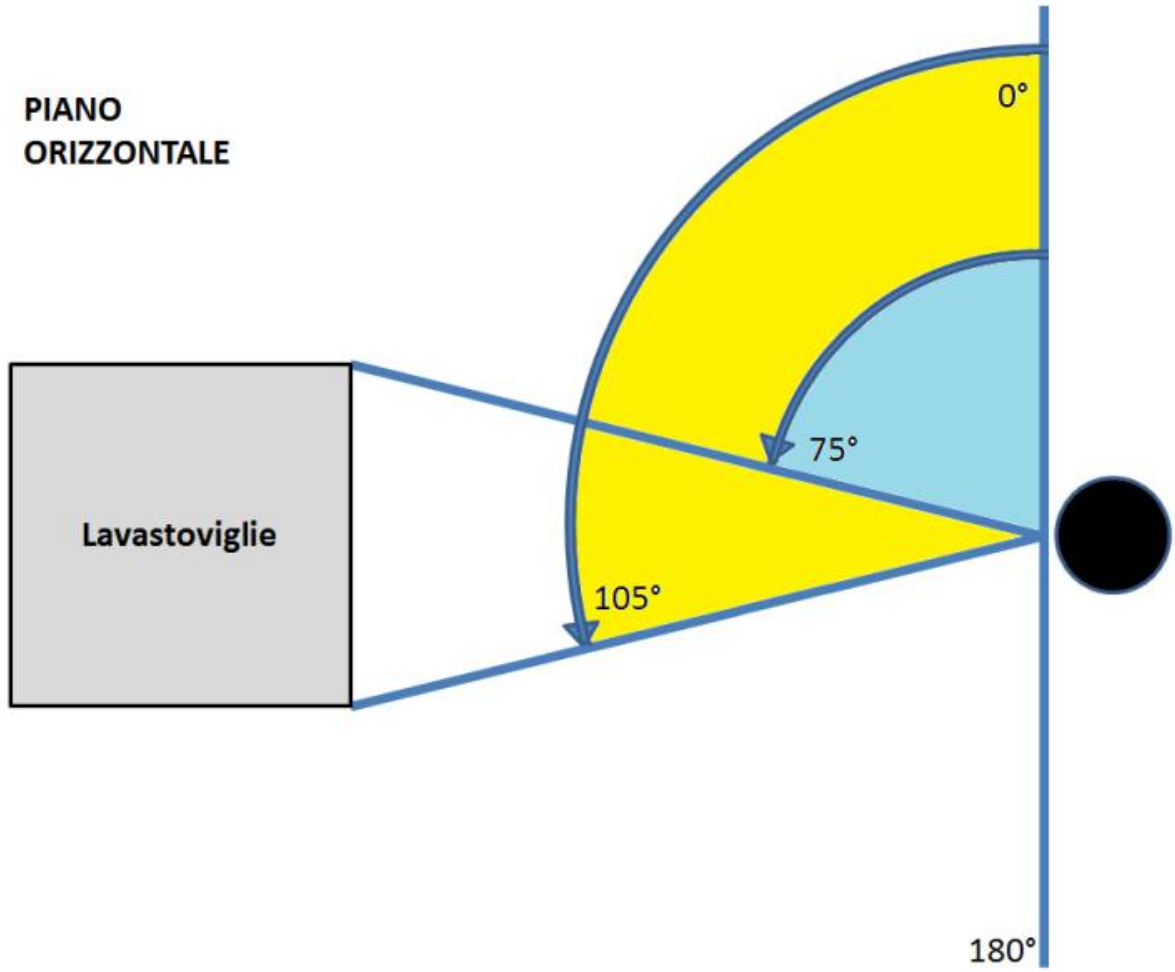


Figura 27 angoli coinvolti, visione verticale

Va sottolineato che SoundingARM non fa distinzione tra mano sinistra e destra, l'utente può infatti cambiare il braccio che sta utilizzando per puntare in qualsiasi momento senza comprometterne il corretto funzionamento.
Può succedere però che l'utente muova entrambe le mani nello stesso momento, per esempio con una sta puntando e con l'altra esegue piccoli movimenti involontari. Poiché il sistema non è in grado di distinguere quale sia il braccio da tracciare, è stata introdotta un'area ombra cilindrica che si estende per 35cm attorno al corpo dell'utente; in tale modo SoundingARM riesce a tracciare solamente le mani quando fuoriescono dall'area ombra.

2.3.3 Componenti di output

Nella prima implementazione, la patch "forwardava" semplicemente il nome dell'oggetto indicato a una sintesi vocale che a sua volta gestiva la pronuncia vera e propria della parola corrispondente.
Nelle implementazioni successive, che verranno dettagliatamente trattate nei prossimi capitoli, tutto l'output è gestito direttamente dalla patch pure data che si occupa di effettuare anche tutta una serie di rielaborazioni del segnale audio, in particolar modo nelle versioni polifoniche e in quelle dove si implementa la spazializzazione binaurale.

2.4 Patch Pure Data

SoundingARM server application lavora in modo molto simile a un processo in background, non è infatti dotata di interfaccia grafica, ed è stata programmata in modo che sia sempre in attesa di comandi spediti tramite il protocollo OSC over UDP sulla porta 7660.

Al contrario, la patch Pure Data fornisce una semplice interfaccia utente che permette di controllare SoundingARM anche grazie alle informazioni stampate sulla finestra console di Pure Data. Più precisamente, l'interfaccia è dotata di tasti che fanno partire l'applicazione, abilitano e disattivano il tracciamento dello scheletro e danno anche la possibilità di terminare la server application.

La SoundingARM server application trasmette tutti i dati utili alla patch Pure Data tramite pacchetti OSC che contengono informazioni circa le posizioni di testa e mani dell'utente, posizione dell'oggetto indicato, nome e attributi.

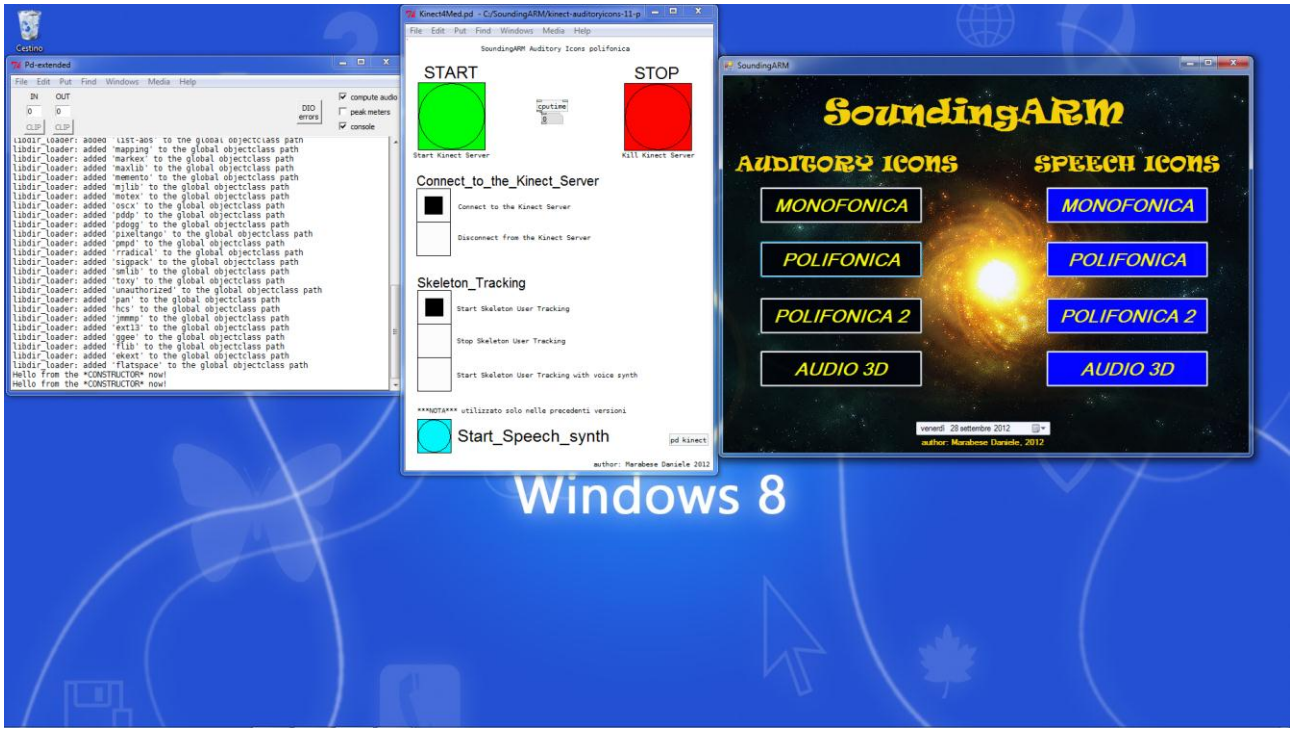


Figura 28 screenshot del desktop che mostra la console Pure Data a sinistra, l'interfaccia di SoundingARM al centro, e l'applicazione SoundingARM_Launcher che permette di lanciare la versione di SoundingARM che più si preferisce

Formato OSC

Open Sound Control OSC è un content format per lo scambio di messaggi tra computer, sintetizzatori di suoni e altri device multimediali che sono ottimizzati per l'interfacciamento con le moderne tecnologie di networking. OSC porta tutti i vantaggi delle moderne infrastrutture di networking all'interno del mondo degli strumenti musicali elettronici, favorendo l'interoperabilità, l'accuratezza, la flessibilità, la documentazione e l'organizzazione avanzata.

OSC è un content format sviluppato al CNMAT da Adrian Freed e Matt Wright, sostanzialmente comparabile a XML WDDX o JSON. Inizialmente pensato per condividere dati musicali (quali gestures, parametri e sequenze di note) tra strumenti musicali (soprattutto sintetizzatori), computer e altri devie

multimediali. OSC è divenuto ben presto un'alternativa allo standard MIDI in tutti quei contesti applicativi dove si esigono migliori performance, maggiori risoluzioni e un più ricco spazio dei parametri.
I messaggi OSC sono comunemente trasportati tramite internet o sottoreti domestiche per mezzo dei protocolli (UDP/IP, Ethernet). Al contrario, i messaggi OSC scambiati tra controller gestuali sono solitamente trasportati tramite interfacce seriali o USB per mezzo del protocollo SLIP.
OSC consente ai musicisti e agli sviluppatori un maggior controllo e una sostanziale flessibilità circa il tipo dei dati che possono spedire, permettendo così una comunicazione tra dispositivi e applicazioni ad un livello più alto.

Features OSC

- Schemi di nomi simbolici, dinamici, URL-style;
- Dati simbolici e numerici ad alta risoluzione;
- Pattern matching language che permette di specificare strutture dati multiple per un singolo messaggio;
- Time tags ad alta risoluzione;
- Gruppi di messaggi I cui effetti accadono contemporaneamente.

Ci sono dozzine di implementazioni di OSC, incluse quelle per ambiti di audio real-time, ambienti di media processing, tool di interazione web, sintetizzatori software, linguaggi di programmazione e corrispettivi device hardware. OSC recentemente ha allargato i campi di utilizzo, raggiungendo anche le interfacce per la robotica, il video processing, i sistemi di distribuzione di musica nelle WAN/LAN, e persino la complicata area dell'intercomunicazione tra processi.
OSC è utilizzato anche in moltissimi controller musicali, prodotti commerciali come il Jazz Mutant Lemur, il Monome, Native Instruments Reaktor e Cycling 74 Max/MSP.

Design OSC

I messaggi OSC contengono coppie nome/valore e un time tag opzionale. I valori sono nominati secondo un name space gerarchico che ricorda lo stile adottato dai filesystem UNIX o quello di una URL. I tipi di valore sono caratterizzati da una rappresentazione a stringa. I valori si rappresentano in forma binaria secondo un allineamento a 4-byte. I tipi supportati di default sono:

- interi con segno a 32-bit in complemento a 2;
- Numeri floating point a 32-bit (IEEE 754);
- Array di caratteri di 8 bit (stringhe C-style);
- Oggetti binary di dimensione arbitraria (come dati audio o video frame).

I vantaggi di OSC rispetto a MIDI sono in primo luogo la velocità e il throughput, ma anche la connettività di rete, la risoluzione dei tipi di dato, la semplificazione nella specificazione di path simboliche.

Open Sound Control (OSC) per Pure Data

Di seguito viene brevemente riportato l'Insieme di oggetti che si utilizzano per gestire I messaggi OSC in Pure Data; questi oggetti si limitano a convertire da PD message a OSC message (formato binario) e viceversa, per cui c'è bisogno di un insieme separato di oggetti che implementino il protocollo di trasporto OSC (layer 4), per esempio [udpsend]/[udpreceive] per mandare pacchetti OSC over UDP.

- **[packOSC]**
  converte un Pd-message in un OSC (binary) message. (utile se si vuole trasmettere messaggi OSC over UDP o altri protocolli che implementano la lunghezza variabile dei pacchetti).
- **[unpackOSC]**
  converte un OSC (binary) message in un Pd-message.
- **[routeOSC]**
  Instrada I messaggi OSC come fa l'oggetto route con I messaggi Pd, sulla base del primo elemento della lista.
- **[packOSCstream]**
  Converte un Pd-message in un OSC (binary) message adatto al trasporto sotto forma di stream (utile se si vuole trasmettere OSC over TCP/IP o line seriali).
- **[unpackOSCstream]**
  Converte un OSC (binary) message adatto allo streaming in un Pd-message.

Capitolo 3

SOUNDINGARM – progetto e implementazione dell'Auditory Display

3.1 IMPLEMENTAZIONE AUDITORY ICONS

In linea con il caso di studio preso in esame, ovvero la cucina, si è provveduto a registrare tutta una serie di suoni tipici dell'ambiente in questione, ed in particolare si è pensato fosse opportuno veicolare l'informazione in modo da generare un background sonoro che prevede l'esecuzione di una serie di suoni caratteristici della cucina, come il rumore dei piatti quando vengono lavati, l'apertura-chiusura del frigorifero, del forno, della lavastoviglie … e molti altri ancora, che vengono riprodotti non appena l'utente inizia a puntarli.
Così facendo, l'utente riesce ad ottenere molto velocemente più informazioni: innanzitutto ha ben chiaro il contesto in cui si trova, inoltre, tramite auditory icons, è consapevole sin da subito di cosa sta indicando e in che direzione tale oggetto si trova.
L'espressività garantita dalla scelta di opportune auditory icon è molto importante, perché consente all'utente non vedente di esplorare la stanza in modo rapido e deciso, senza bisogno di soffermarsi troppo a lungo nell'ascolto delle descrizioni di oggetti del quale non è interessato.

Patch Pure Data – Auditory icons

L'application server comunica con la patch Pure Data tramite pacchetti OSC over UDP sulla porta 7660.

3.1.1 Interfaccia Grafica Principale

All'apertura, la patch Pure Data si presenta con una finestra all'interno della quale si trovano tutti i comandi necessari per controllare SoundingARM, in particolare si distinguono 4 gruppi di pulsanti:

1. **avvio/chiusura SoundingARM server application**
   - ✓ **Start_kinect_server** avvia la SoundingARM server application (che come precedentemente descritto è priva di interfaccia grafica ma opera in background) la quale a sua volta apre la finestra Folder Browser all'interno della quale si seleziona l'opportuno file .CONF da caricare.
   - ✓ **Kill_Kinect_Server** termina SoundingARM server application.

2. **Connessione al server Kinect**
    - ✓ **Connect to the Kinect server** esegue la connessione tra il SoundingARM application Server e il Kinect Server che si occupa dell'interfacciamento e l'acquisizione dei dati provenienti dal sensore.
    - ✓ **Disconnect from the kinect server** disconnette l'applicazione dal Kinect Server.
3. **Tracciamento dello scheletro**
    - ✓ **Start Skeleton User Tracking** abilita il tracciamento dello scheletro dell'utente; sostanzialmente abilita la patch alla ricezione dei pacchetti OSC contenenti le informazioni circa la testa e le mani dell'utente;
    - ✓ **Stop Skeleton User Tracking** disabilita il tracciamento dello scheletro dell'utente;
    - ✓ **Start Skeleton User Tracking with voice synth** abilita il tracciamento dello scheletro dell'utente nella modalità con sintesi vocale (in questa modalità i pacchetti OSC che arrivano contengono meno informazioni, in particolare solo il nome dell'oggetto puntato e non tutte le altre coordinate della testa, della mano e gli attributi degli oggetti).
4. **Sintetizzatore vocale**
    - ✓ **Start_Speech_Synth** attiva il text to speech server, ovvero l'applicazione che riceve dalla patch Pure Data il nome dell'oggetto indicato e, tramite l'API Microsoft Speech, avvia la riproduzione vocale dell'oggetto selezionato.

NOTA: *questo tasto, così come la modalità precedente Start Skeleton User Tracking with voice synth erano utilizzate dalla prima versione di SoundingARM, quella che appunto utilizzava la sintesi vocale. Nelle versioni successive questa funzionalità è ancora presente, solamente per motivi di retro-compatibilità, però non viene, di fatto, più utilizzata.*

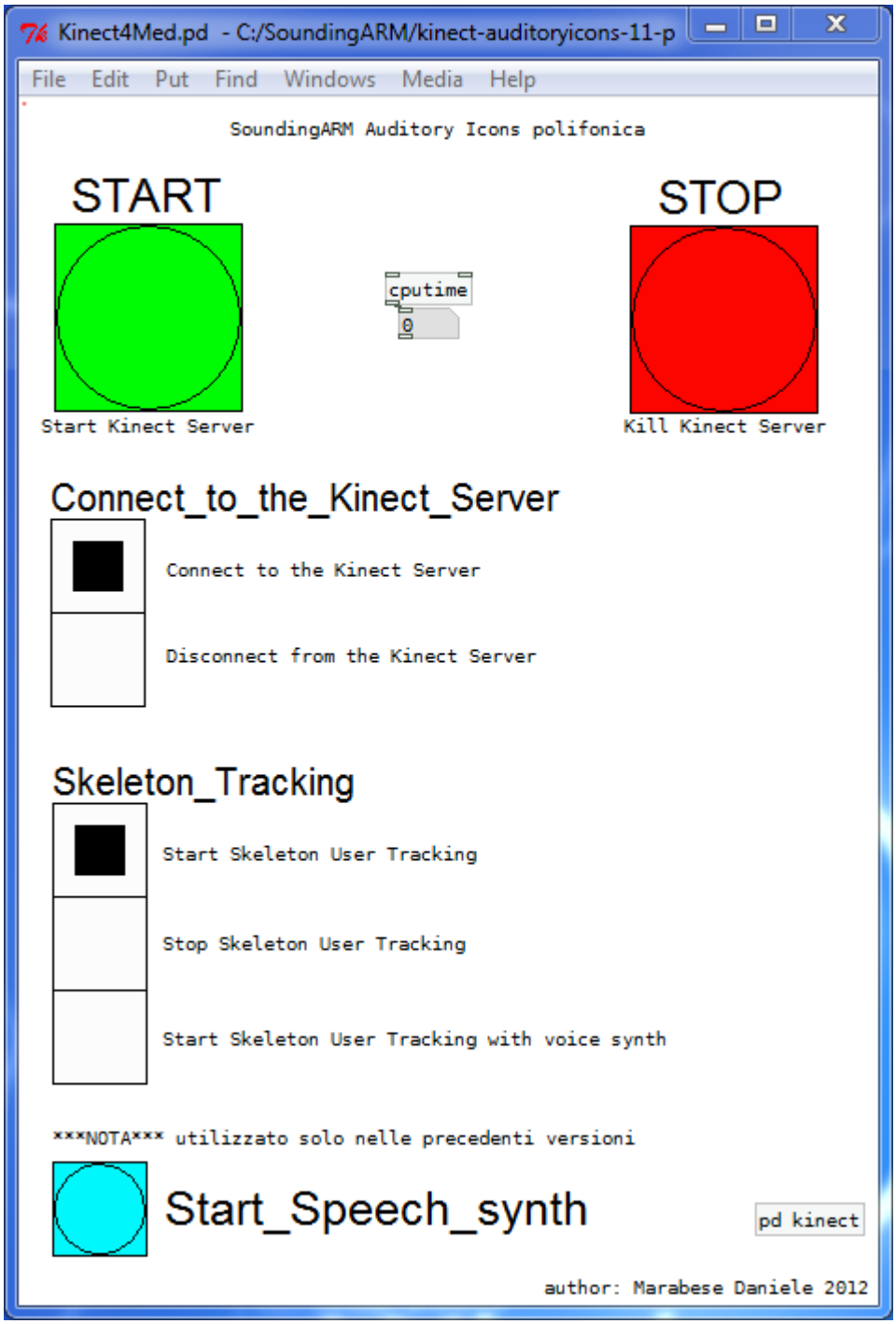


Figura 29 screenshot interfaccia utente: controlli

3.1.2 Subpatch kinect

In basso a destra si trova il collegamento alla subpatch Kinect, la quale contiene al suo interno tutto il meccanismo di ricezione ed invio dei pacchetti OSC provenienti dal SoundingARM application Server e verso il Text to Speech Server.

A sinistra si possono scorgere gli oggetti netreceive e la subpatch sapi_pd che sono collegati direttamente alla subpatch extra, che costituisce in un certo qual modo il “motore audio” delle nuove implementazioni.

A destra, invece, ci sono tutti gli oggetti direttamente connessi alla patch di controllo descritta in precedenza.

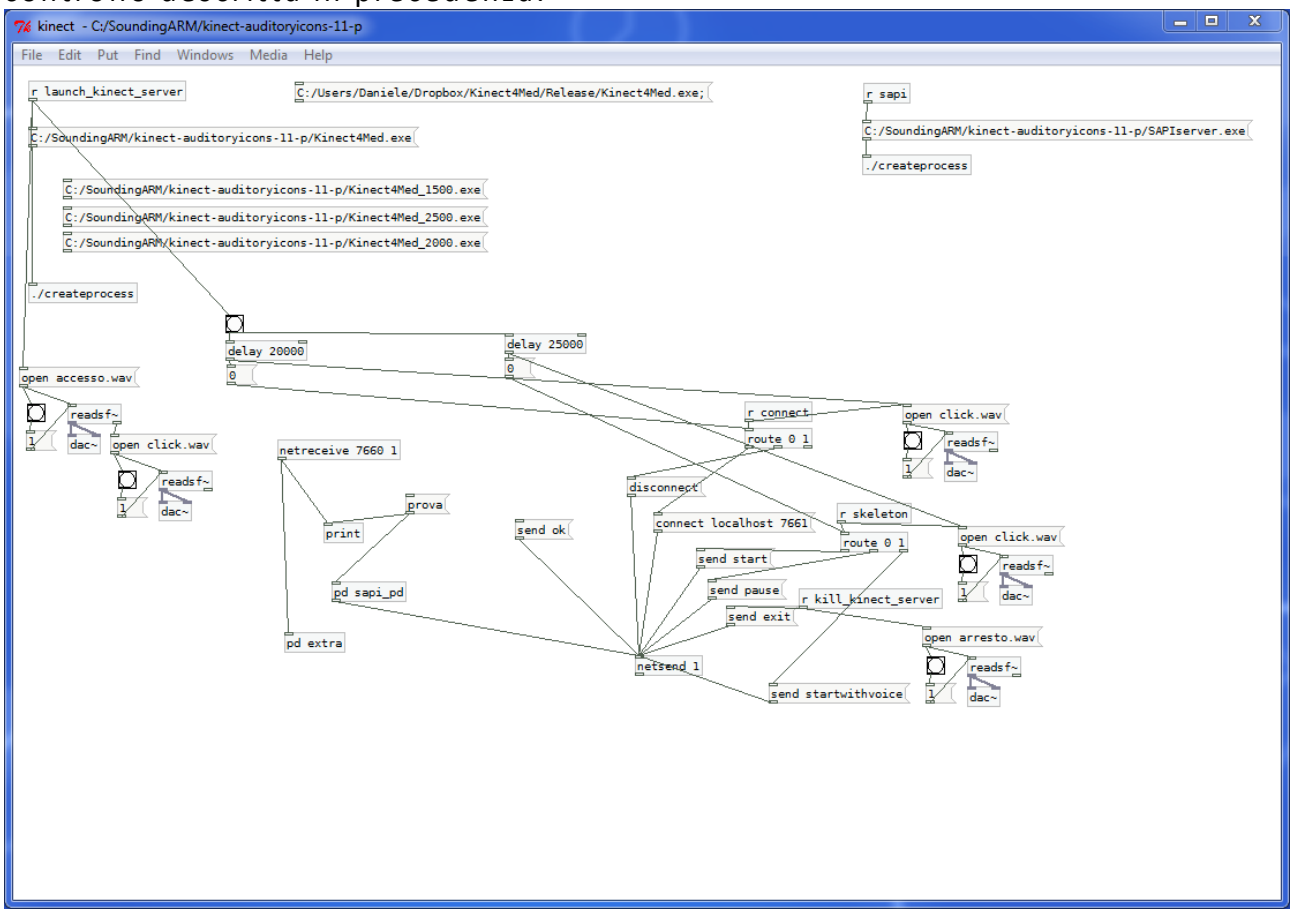


Figura 30 screenshot della subpatch kinect

3.1.3 Subpatch extra

Nella subpatch extra avviene la vera e propria analisi dei messaggi OSC provenienti dal SoundingARM application Server. Il formato di un messaggio OSC ricevuto dalla patch è il seguente:

> */id open frigorifero.wav, /p1 111 222, /p2 333 444, /p3 555 666, /p4 777 888, /height 999, /heightfromfloor 1010, /userposition 1111 2222 3333, /righthandposition 4444 5555 6666 /lefthandposition 7777 8888 9999*

In particolare il primo tag contiene il commando “open” seguito dal nome del file audio.wav da aprire che corrisponde all’oggetto indicato, p1 p2 p3 p4 sono le coordinate (X,Z) di base dell’oggetto indicato, height e heightfromfloor sono gli attributi dell’oggetto, userposition righthandposition e lefthandposition sono rispettivamente le coordinate della testa, della mano destra e sinistra dell’utente.

L’oggetto Pd [./osc/routeOSC /id /p1 /p2 /p3 /p4 /height /heightfromfloor /userposition /righthandposition /lefthandposition] è in grado di scompattare tutte le informazioni sopra descritte nei singoli “atomi” e di indirizzarli tramite i suoi 11 outlet verso gli altri oggetti della patch che si occupano della rielaborazione.

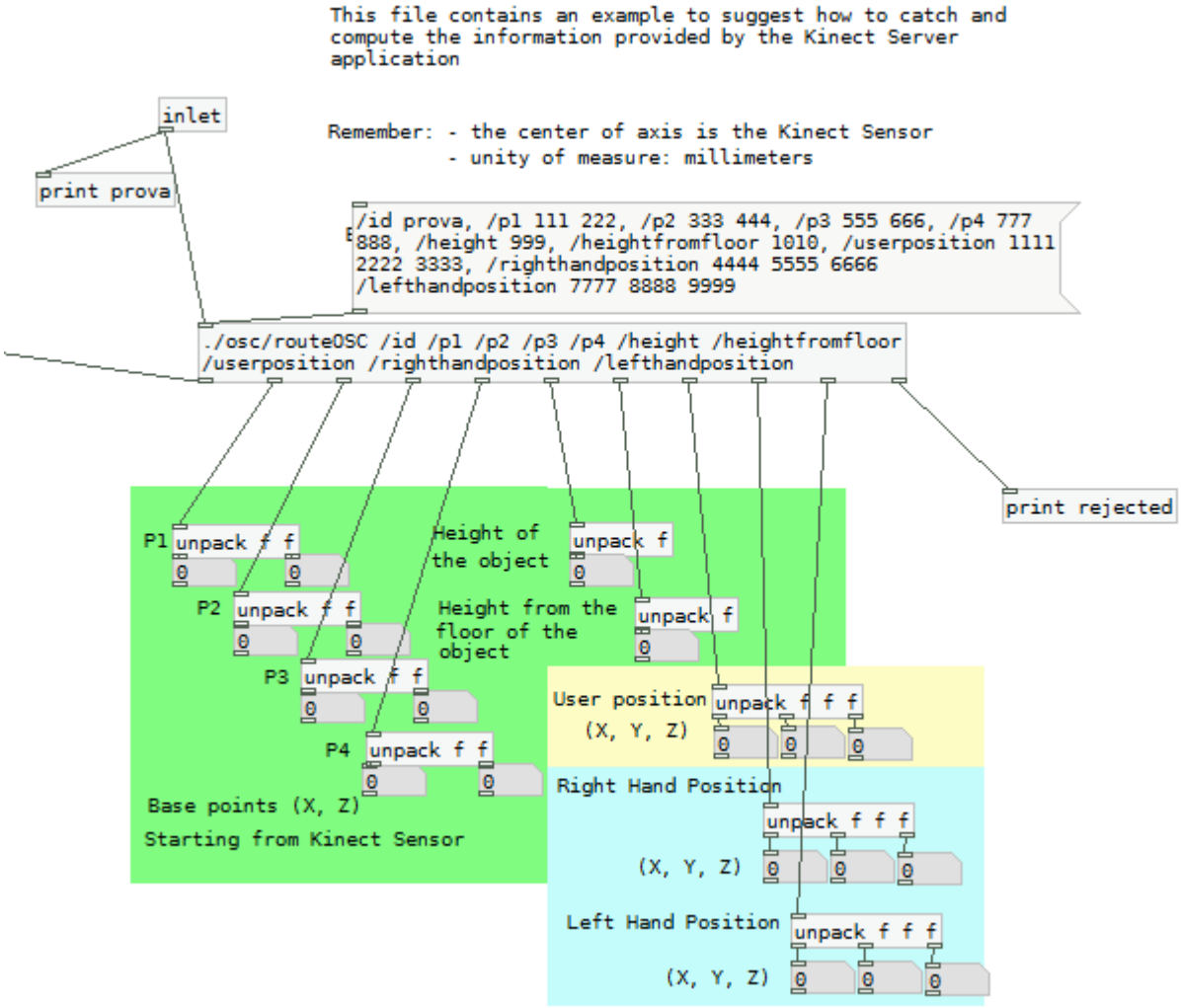


Figura 31 screenshot della subpatch extra, focus sul pacchetto OSC ricevuto

3.1.4 Versione Monofonica

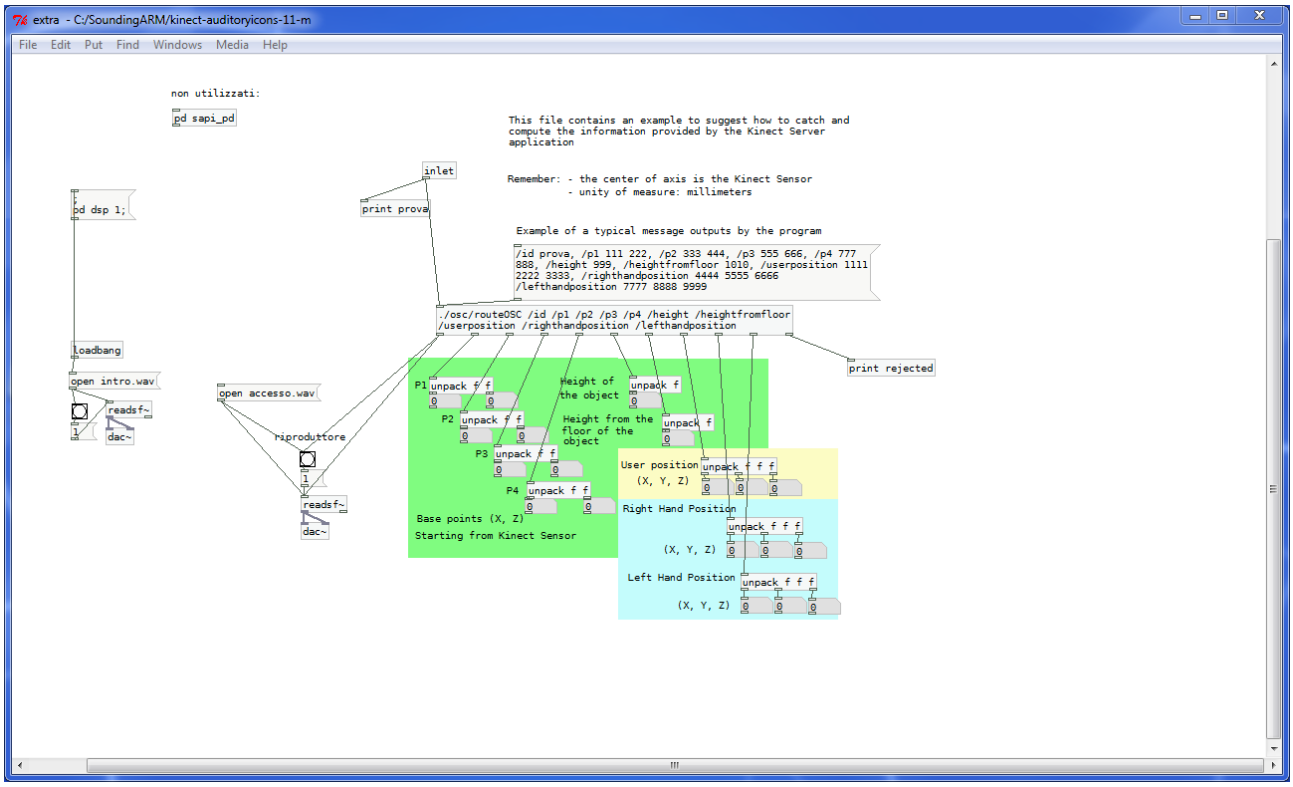


Figura 32 screenshot subpatch extra - versione monofonica

L'idea chiave delle versioni che sfruttano le auditory icons è quella di eseguire un appropriato file audio quando l'utente punta un determinato oggetto, per definizione di auditory icons, il suono deve richiamare l'oggetto in modo naturale, ed è per questo che, anche dal punto di vista dell'applicazione, l'apertura di suddetta traccia audio deve avvenire contestualmente all'azione di puntamento eseguita dall'utente.
Rispetto all'approccio con sintesi vocale, dove ad ogni oggetto corrispondeva un ID, al quale era associato un nome che successivamente veniva letto e pronunciato dalla sintesi vocale, nelle implementazioni che sfruttano le auditory icons, l'oggetto è ancora caratterizzato in modo univoco dall'ID, però ad esso non è tanto associato il nome (in quanto di nessuna utilità all'applicazione) bensì è associato direttamente il comando di apertura file che successivamente viene interpretato nel modo corretto dall'oggetto readsf~ che si occupa della riproduzione della traccia stessa.

L'oggetto readsf~
readsf~ è appunto l'oggetto Pure Data che si occupa di leggere e riprodurre file audio per mezzo di un oggetto dac~.
L'oggetto readsf~ legge un file audio e lo converte in segnale audio di output. Il file audio deve essere aperto qualche istante prima della riproduzione per mezzo del messaggio "open". Non appena readsf~ riceve "open", immediatamente inizia la lettura del file, ma la riproduzione vera e propria, e quindi l'uscita del segnale in output inizia solo dopo che readsf~ riceve un messaggio "1" che costituisce lo start del playback. Analogamente, un messaggio "0" corrisponde allo stop della riproduzione.

I file .wave, aiff, e altri formati vengono letti automaticamente, sebbene siano accettati solo samples composti di 2, 3 e 4 byte.
“open” richiede un nome di file o un percorso, un sample rate, e opzionalmente si possono specificare anche una dimensione di header da saltare, il numero di canali da usare nella riproduzione, un buffer di tot bytes per canale, e l’ordinamento big o little endian. Inoltre, al termine della riproduzione, readsf~ emette un bang in uscita dal proprio outlet di destra.

Implementazione
Da un punto di vista implementativo, la versione monofonica proposta fa uso di un oggetto readsf~ la cui inlet calda è direttamente collegata all’outlet sinistra dell’oggetto routeOSC, che si occupa della ricezione e dello scompattamento del messaggio OSC.
Non appena arriva un nuovo messaggio OSC, questo viene scompattato e dalla prima outlet di routeOSC partono contemporaneamente il messaggio contenente il comando open nomefile.wav diretto a readsf~, e un messaggio che aziona un bang che a sua volta invia un messaggio “1”, sempre nell’inlet di readsf~, che a questo punto attiva la riproduzione di nomefile.wav.
L’applicazione è strutturata in modo tale che non appena un utente punta qualche oggetto, il SoundingARM application Server va in sleep per 1500ms, dopo di che torna ad essere attivo; in questo modo si scongiura il rischio che un utente indichi continuamente lo stesso oggetto o oggetti differenti senza poter ascoltare il suono emesso dall’oggetto indicato poiché senza sleep l’applicazione entrerebbe in una sorta di loop nel quale l’application server manderebbe i dati del nuovo oggetto indicato, la patch li inoltrerebbe a readsf~ che inizierebbe la riproduzione, ma sarebbe costretto subito a interrompersi per eseguire un nuovo file audio che potrebbe essere lo stesso o un altro (dipende dal movimento dell’utente) senza però avere mai il tempo di eseguire nemmeno un istante del suddetto file.
Lo sleep time non è certo una soluzione molto dinamica perché entra in azione ogni qual volta venga indicato un oggetto, senza fare distinzione del fatto che l’utente stia indicando lo stesso oggetto nuovamente per riascoltare il suono o perché ha compiuto un gesto errato; però è anche una buona soluzione perché permette all’applicazione di veicolare un buon soundscape, senza compromettere la componente real-time del sistema.
A livello pratico, comunque, lo sleep time non è percepito minimamente dall’utente, in quanto 1500ms sono veramente pochi se si considera che ogni file audio, sebbene sia stato registrato e ottimizzato per partire nel modo più rapido possibile, ha comunque circa 250ms di silenzio all’inizio, una durata media di 3 secondi, e soprattutto bisogna tenere conto che l’utente ha interesse a scoprire la stanza nei minimi dettagli quindi compie gesti non rapidissimi, anche perché, in caso contrario, verrebbe a mancare una delle caratteristiche fondamentali del puntamento, ovvero la coincidenza tra la direzione verso la quale l’utente sta puntando e l’oggetto puntato con il relativo suono emesso.

Considerazioni

- **Vantaggi**

  Questo tipo di implementazione ha consentito di ottenere una versione di SoundingARM decisamente reattiva, molto precisa nel puntamento e riconoscimento degli oggetti, sia in senso orizzontale che verticale, con un feeling acustico che dipende in larga parte dai file audio utilizzati, piuttosto che dalle componenti di elaborazione audio del sistema, dato che queste ultime sono ridotte veramente al minimo.

- **Svantaggi**

  L'approccio auditory icons monofonico è indubbiamente veloce e preciso, però ha il difetto (soprattutto nei contesti dove gli "oggetti suonanti" sono molto ravvicinati nello spazio) di interrompere troppo spesso la riproduzione di un suono per far partire il successivo, di conseguenza, in taluni ambiti come quello della cucina, può risultare troppo reattivo, e richiede all'utente una notevole velocità nella percezione del suono.

3.1.5 Versione Polifonica

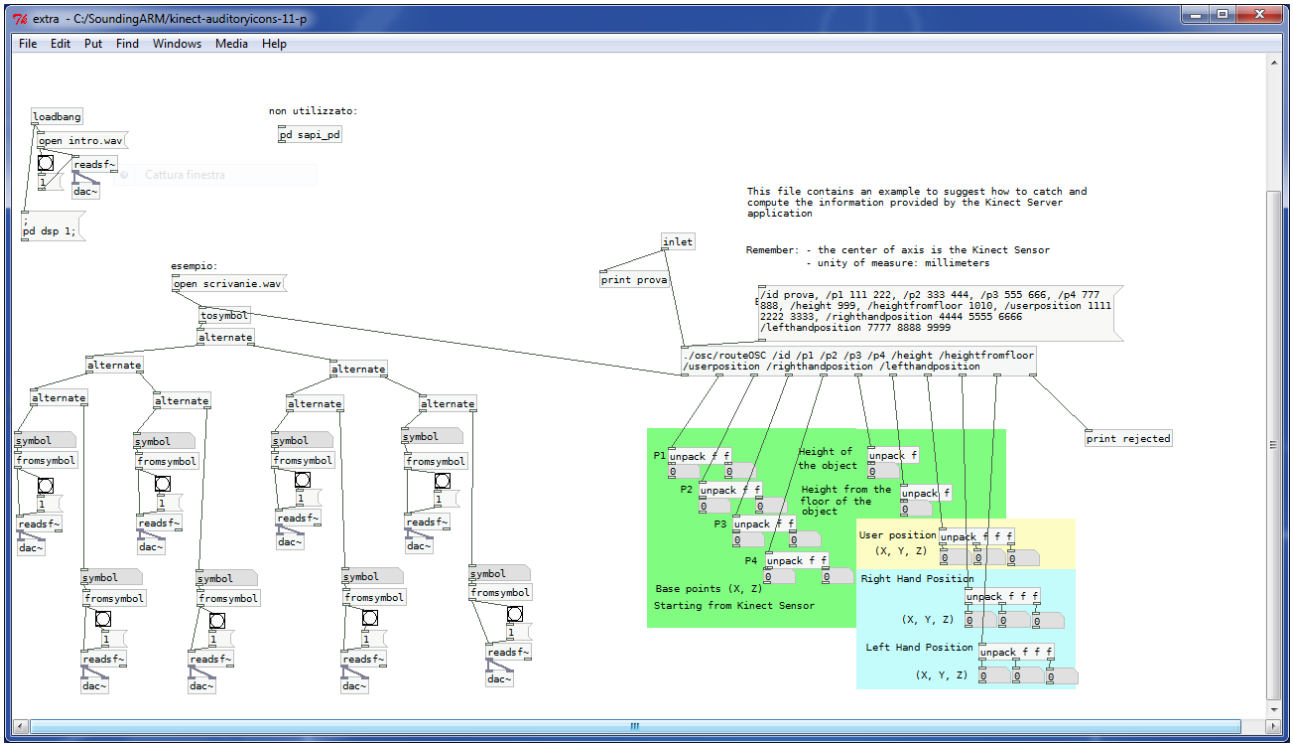


Figura 33 screenshot subpatch extra - versione polifonica

La prima versione polifonica proposta segue, di fatto, quanto già introdotto dalla versione monofonica, però ha come obiettivo quello di risolvere il difetto riscontrato nella precedente versione, ovvero la possibilità di riprodurre un solo file audio alla volta.

L'idea chiave è, per l'appunto, quella di utilizzare più oggetti readsf~ alla volta, una sorta di lettori di file audio paralleli (simile a quanto accaduto nel passaggio da processori single-core a processori multi-core), in particolare 8 readsf~ indipendenti l'uno dall'altro che hanno il compito di riprodurre ciascuno un file audio in modo da ottenere due effetti:

- ✓ in primo luogo se l'utente continua il movimento e punta più oggetti in successione, l'esecuzione del primo suono non viene interrotta per dar spazio al suono successivo, in quanto la riproduzione del secondo avviene su un readsf~ differente,
- ✓ in secondo luogo l'utente man mano che procede col puntamento degli oggetti, percepisce sia il suono dell'oggetto che sta puntando, sia la parte terminale del suono del precedente oggetto indicato, ottenendo una sorta di "effetto fade" complessivo, una sovrapposizione audio non distruttiva che può essere molto utile in contesti ricchi di oggetti suonanti, come gli scenari domestici della cucina, del bagno o del salotto. Per ottenere questo effetto bisogna porre molta attenzione in fase di registrazione dei suoni, poiché questi ultimi devono essere sufficientemente corti e non sovrapporsi in modo fastidioso.

Gli oggetti Pure Data tosymbol e fromsymbol

Dal momento che Pure Data di default non è in grado di riconoscere i caratteri di spazio all'interno dei symbols, è necessario utilizzare le librerie di PD-extended, in particolare Cycling'74 che mette a disposizione due oggetti: tosymbol e formsymbol che risolvono il problema.

tosymbol è un oggetto che riceve in ingresso un message Pure Data, quindi caratterizzato anche dalla presenza di spazi, e restituisce in uscita un symbol con gli spazi, sostanzialmente trasforma un messaggio con spazi in un symbol con spazi.

Fromsymbol fa esattamente il contrario, prende in ingresso un symbol, con o senza spazi, e restituisce in uscita un messagge equivalente.

L'oggetto alternate

alternate è un oggetto che alterna i messaggi che riceve nella sua inlet calda, facendoli uscire appunto alternativamente dall'outlet sinistra e destra.

alternate non esegue alcuna forma di rielaborazione o processamento dei messaggi che gli arrivano, semplicemente li fa passare attraverso, alternando tra uscita sinistra e destra, in particolare, il primo messaggio ricevuto (il numero 0) e tutti quelli con numerazione pari escono a sinistra, analogamente il secondo e tutti quelli con numerazione dispari escono a destra.

Implementazione

A livello implementativo, come anticipato, si utilizzano 8 readsf~ distinti. Rispetto all'implementazione monofonica però, non si possono collegare i singoli readsf~ direttamente all'outlet sinistra di routeOSC, in quanto tutti e quattro i riproduttori riceverebbero lo stesso comando e di conseguenza eseguirebbero all'unisono il medesimo file audio, ottenendo come risultato una sovrapposizione perfetta dello stesso suono, che però non porta all'effetto sperato, in quanto, non appena un utente indica l'oggetto successivo, tutti e quattro i riproduttori si fermano ed eseguono il nuovo suono.

L'idea è appunto quella di utilizzare l'oggetto alternate, in particolare, basterebbe un alternate da solo per gestire due readsf~ distinti, però numerose prove sperimentali hanno mostrato come, in scenari affollati di

“oggetti suonanti” (da 20 in su), convenga utilizzare 4-8 readsf~ in modo da garantire un buon soundscape globale.
Per poter usare 8 readsf~, è sufficiente collegare 7 alternate in configurazione a cascata, sostanzialmente un primo alternate è collegato direttamente alla prima outlet sinistra di routeOSC, e poi l’outlet sinistra del primo alternate è collegata a un secondo alternate, e l’outlet destra del primo alternate è collegata a un terzo alternate (2° e 3° alternate costituiscono il secondo livello). Ciascun alternate del secondo livello è collegato ad altri due alternate del terzo livello che a loro volta sono collegati ai riproduttori radsf~.
Così facendo si possono collegare tutti gli 8 readsf~ e, per l’appunto alternativamente, ognuno di essi eseguirà un suono scongiurando sovrapposizioni, perché di volta in volta, quando l’utente punta un nuovo oggetto, la riproduzione del suono precedente non si stoppa bensì procede sino al termine del file audio, e al contempo la riproduzione del nuovo oggetto appena indicato può iniziare su un readsf~ differente.
Osservando la figura precedente dove viene riportata la patch, si può notare come, in realtà, la prima outlet sinistra di routeOSC sia collegata a un oggetto tosymbol, e poi sia quest’ultimo a essere collegato all’alternate principale, analogamente gli alternate al terzo livello non sono collegati direttamente alle inlets di ingresso dei readsf~, bensì sono connesse a degli oggetti fromsymbol. Questi accorgimenti sono dovuti al fatto che readsf~ non accetta in ingresso nessun oggetto che non sia un Pd-message, quindi è stato necessario utilizzare questi oggetti di PD-Extended, che sono in grado di effettuare le conversioni, mantenendo gli spazi e i formati dati richiesti da readsf~.

Considerazioni

- **Vantaggi**
  Questa implementazione mantiene tutte le caratteristiche positive della versione monofonica, aggiungendo un sostanziale effetto di fade e una polifonia complessiva, grazie alla possibilità di poter eseguire fino a 8 suoni corrispondenti a 8 differenti oggetti alla volta, senza incorrere in fastidiose sovrapposizioni, e senza comportare bruschi arresti nella riproduzione, come invece accadeva con la versione precedentemente descritta.
- **Svantaggi**
  Essenzialmente questa versione non comporta svantaggi, in quanto rimane comunque molto reattiva e precisa. L’unico aspetto degno di nota è che, da un punto di vista computazionale, l’uso di risorse cresce, non tanto lato CPU bensì lato scheda audio.
  Inoltre va sottolineato che l’effetto di soundscape e di fade è determinato:
  - in primo luogo dal numero di oggetti suonanti e dalla loro disposizione spaziale: l’effetto si avverte molto in ambienti dove ci sono molti oggetti vicini tra loro (20 e più), meno in altri dove gli oggetti sono spazialmente lontani e disposti in uno spazio maggiore;
  - in secondo luogo la cura con cui si scelgono, registrano e modificano i suoni audio è cruciale, in particolare la durata dei suoni è stimata come trade-off tra rapidità di esecuzione ed espressività del suono in

relazione all'oggetto che rappresenta, al fine di ottenere riproduzioni polifoniche armoniose, prive di fastidiose sovrapposizioni distruttive che annullerebbero il soundscape percepito.

3.1.6 Versione Polifonica 2

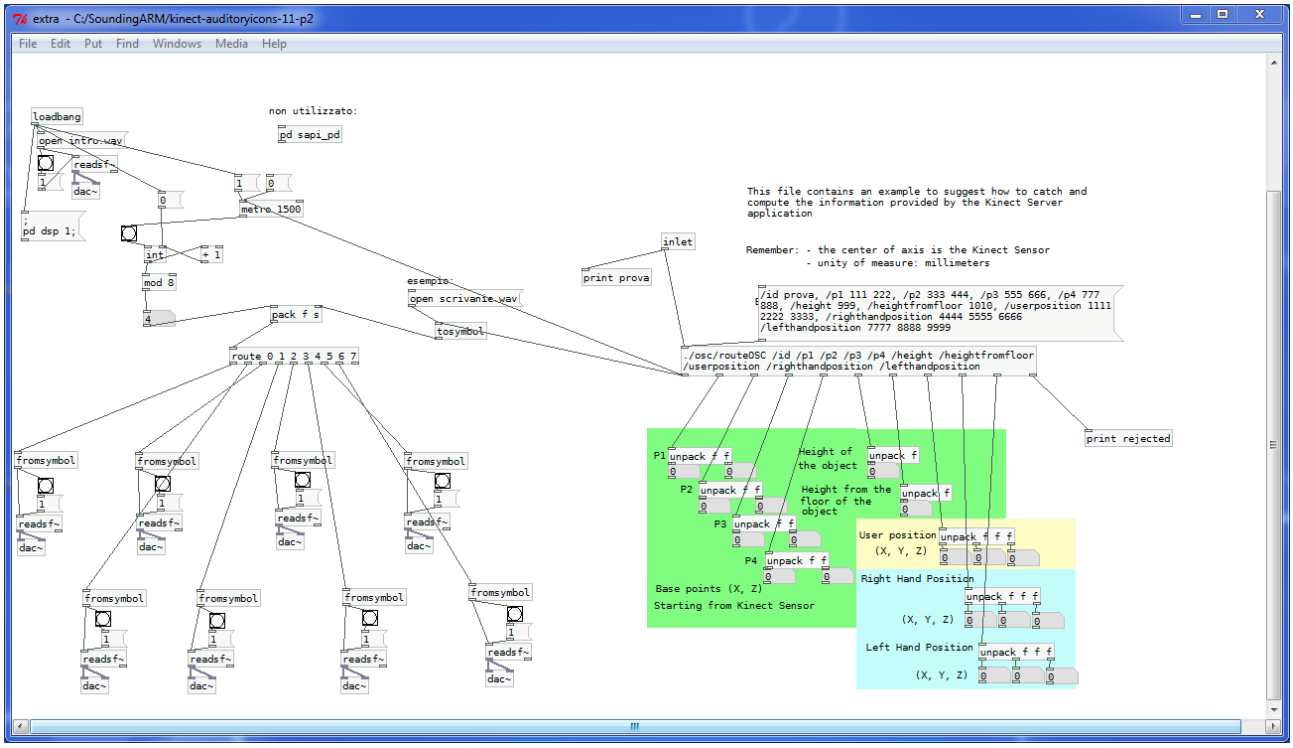


Figura 34 screenshot subpatch extra - versione polifonica 2

La polifonia può essere ottenuta utilizzando un approccio radicalmente differente rispetto al precedente, infatti, invece dell'oggetto alternate per direzionare i messaggi provenienti da routeOSC verso i differenti readsf~, si sfrutta una combinazione di più oggetti:

- un oggetto metro,
- un contatore,
- un oggetto pack,
- un oggetto route.

L'oggetto metro

L'oggetto metro produce in output un bang ogni n millisecondi, dove n è l'argomento dell'oggetto e rappresenta la distanza temporale fra un bang e l'altro, espressa in ms. Per funzionare ha bisogno che sia attivato attraverso un messaggio diverso da zero o un bang. Un messaggio uguale a zero oppure stop, spegne l'oggetto.

Il contatore

Attraverso l'oggetto metro si può costruire un contatore. La patch non fa altro che emettere a intervalli regolari (per esempio 500 ms) un numero intero a

partire da zero tale che il successivo sia maggiore del precedente di una unità. L'algoritmo necessita di un oggetto che permetta di conservare in memoria un numero intero, ovvero dell'oggetto int. L'inlet destro di int riceve un intero e lo memorizza finché un bang nell'inlet sinistro non lo costringe a inviarlo tramite il proprio l'outlet. All'apertura della patch l'oggetto int viene inizializzato con uno zero nella sua memoria (tramite loadbang). All'attivazione di metro, int riceve un bang nella sua entrata calda facendo uscire zero, che viene mandato all'uscita della patch e contemporaneamente sommato a 1, per mezzo dell'oggetto '+'. Quest'ultimo a sua volta spedisce il risultato, cioè uno, nell'entrata fredda di int. Il nuovo valore ("1") uscirà non appena sopraggiungerà il successivo bang prodotto da metro e così via.

L'oggetto pack

L'oggetto pack mette insieme diversi valori singoli detti "atomi", combinandoli in liste. Gli argomenti di pack inizializzano il valore del relativo inlet al valore segnato. Ad esempio [pack 2 19 25] avrà 3 inlet rispettivamente inizializzati con i valori 2, 19, 25. Un bang in entrata nell'inlet sinistro produrrà l'output della lista in memoria in quel momento. Il numero di argomenti di creazione determina il numero di inlet, mentre il tipo degli argomenti determina i tipi di messaggi che pack si aspetta di ricevere su ciascuna inlet.
Il modo migliore per inserire gli argomenti di creazione è definire esplicitamente il tipo di dati che pack si deve aspettare, per esempio floats, symbols, o pointers. Pack può essere creato con qualsiasi numero di argomenti di creazione, e automaticamente viene creata un'inlet per ogni argomento; il valore di ciascun argomento può essere modificato tramite l'apposita inlet oppure tutti simultaneamente possono essere aggiornati per mezzo di un messaggio o una lista.

L'oggetto route

Indirizza i messaggi in accordo al loro primo elemento. Route controlla il primo elemento del messaggio in ingresso e lo confronta con ciascuno dei propri argomenti di creazione, per esempio numeri o simboli ma non un misto dei due, a meno che il tipo dei dati non sia esplicitamente definito, e poi inoltra il messaggio tramite l'outlet appropriata.

- ✓ Se route trova un match e il messaggio contiene solo un elemento, allora route manda un bang attraverso la corrispondente outlet.
- ✓ Se route trova un match e il messaggio contiene più elementi, ad esempio una lista, allora tutti gli elementi della lista, eccetto il primo, sono mandati tramite il corrispondente outlet.
- ✓ Se route non trova nessun match, allora l'intero messaggio è mandato attraverso l'outlet più a destra che è chiamata la "rejection" outlet. In breve, il numero di outlet è il numero di argomenti + 1.

Implementazione

Rispetto alle precedenti implementazioni, la patch è sicuramente più complicata anche dal punto di vista grafico.

In questo caso, l'obiettivo è quello di direzionare i messaggi provenienti da routeOSC verso gli 8 readsf~ in modo che ciascuno esegua ciclicamente un suono differente, in modo sequenziale, senza sovrapposizioni.

L'idea di base è di intercettare il messaggio proveniente da routeOSC con all'interno il comando del file audio da aprire, aggiungerci un header costituito da un int che può essere 0, 1, 2, 3, 4, 5, 6, o 7, far passare il messaggio all'interno di un oggetto route il quale appunto direziona ciclicamente i messaggi verso gli 8 readsf~ (readsf~0, readsf~1, readsf~2, readsf~3, readsf~4, readsf~5, readsf~6 e readsf~7).

Da un punto di vista concettuale la patch è composta da due parti, la prima che si occupa del contatore e quindi dell'header da aggiungere, e la seconda che assembla e dirotta i messaggi sui riproduttori corrispondenti.

Innanzitutto l'oggetto [metro 1500] è l'elemento che regola l'iterazione del contatore, infatti ogni 1500ms emette un bang che spinge fuori da int il valore numerico contenuto (inizialmente inizializzato a 0 da loadbang), lo somma ad 1 e lo manda all'interno dell'oggetto [mod 8] che restituisce appunto il resto della divisione del numero per 8. Questo meccanismo, composto da metro + int + mod8 permette di realizzare un contatore che va da 0 a 7 e poi si resetta.

tosymbol intercetta il messaggio proveniente dall'outlet sinistra di routeOSC, lo converte in symbol e lo passa come secondo argomento a [pack f s] che è l'oggetto che giustappone l'header numerico proveniente dal contatore e poi spedisce il messaggio così modificato a route.

[route 0 1 2 3 4 5 6 7] è l'oggetto che indirizza i messaggi verso il corrispondente riproduttore sulla base dell'header numerico, in particolare a ogni argomento di creazione corrisponde un outlet direttamente collegata a un oggetto fromsymbol che riconverte da symbol a message e invia il comando di apertura file al corrispettivo readsf~.

Da un punto di vista implementativo, se si sceglie di utilizzare la patch nella forma appena descritta, tutti i meccanismi funzionano correttamente, però ci si scontra ben presto con un problema a dir poco fastidioso, infatti l'oggetto metro prosegue ad emettere bang ogni 1500ms, facendo iterare il contatore anche se l'utente non sta più puntando, e a causa del fatto che i l'ultimo messaggio spedito da routeOSC viene mantenuto lo stesso in memoria, l'ultimo suono riprodotto entra in loop in quanto viene continuamente eseguito ciclicamente su tutti i riproduttori readsf~ (dal momento che il contatore continua ad essere iterato da metro) finché non viene chiusa bruscamente l'applicazione.

Una soluzione non proprio elegante, ma ugualmente efficace, è stata quella di introdurre due oggetti fittizi all'interno della stanza, in corrispondenza di due zone non occupate da altri oggetti. In entrambi gli scenari di utilizzo, tali oggetti sono stati aggiunti nelle mappe approssimativamente al di sopra della testa dell'utente, in corrispondenza della spalla destra e della spalla sinistra. I due oggetti "jolly", non contengono un suono correlato e quindi quando vengono puntati non lanciano comandi "open filename.wav", bensì mandano rispettivamente i comandi di start ("1") e stop ("0") all'oggetto metro.

```
// Estratto di un file .CONF: oggetto jolly stop metro
        [stop]
        p1 = (-150, 2950);
        p2 = (-950, 2950);
        p3 = (-150, 3500);
        p4 = (-950, 3500);
        heightOfObject = 500;
        heightFromFloor = 2000;
```

```
// Estratto di un file .CONF: oggetto jolly start metro
        [1]
        p1 = (50, 2950);
        p2 = (800, 2950);
        p3 = (50, 3500);
        p4 = (800, 3500);
        heightOfObject = 500;
        heightFromFloor = 2000;
```

Figura 35 oggetti jolly start e stop

In questo modo, quando l'utente smette di puntare gli oggetti, semplicemente alza in alto la mano sinistra in modo da puntare l'oggetto corrispondente a stop, spegnendo così il metro, si arresta il contatore e l'applicazione intera si blocca. Per ripristinare il funzionamento dell'applicazione è sufficiente alzare la mano destra puntando l'oggetto start che riattiva il metro e con esso tutto il programma.
L'idea che sta sotto a questo approccio è che metro emette bang, e se metro non è attivo, non sono generati bang e conseguentemente l'applicazione non può proseguire nell'esecuzione solo con i messaggi che riceve, per cui entra in una sorta di stand-by finché metro non è riattivato.

Considerazioni

- **Vantaggi**

  Questa implementazione raggiunge prestazioni molto simili alla polifonica precedentemente descritta, sia in termini di precisione che di reattività.
  Dal punto di vista computazionale è leggermente più impegnativa poiché esegue una maggiore rielaborazione, ma nonostante ciò non impatta in modo significativo in real-time.
  Il vero vantaggio di questa versione è che consente all'utente di attivare e disattivare a piacimento la riproduzione dei suoni. Questo può essere particolarmente utile, per esempio nei contesti in cui si richieda del tempo per far entrare o accompagnare l'utente all'interno della stanza, e in tali frangenti sarebbe inopportuno che il sistema di tracciamento iniziasse a riconoscere i movimenti non tanto dell'utente bensì di chi lo sta accompagnando. In questo modo l'utente solamente quando è pronto può attivare la riproduzione dei suoni, e quando si ritiene soddisfatto dell'esplorazione può decidere di interrompere la riproduzione senza la

necessità di interfacciarsi con la patch, ma semplicemente compiendo un gesto con la mano.

- **Svantaggi**
  Se da un lato la possibilità di attivare e stoppare la riproduzione può essere vantaggiosa, di contro obbliga l'utente a eseguire operazioni supplementari che nella precedente implementazione non erano necessarie, in quanto l'applicazione restava sempre attiva e non c'era bisogno di attivarla. Per questo motivo si è pensato fosse opportuno mantenere entrambe le versioni polifoniche, in quanto ci possono essere contesti in cui è preferibile l'una piuttosto che l'altra.

3.2 IMPLEMENTAZIONE SPEECH ICONS, SPEARCONS, EARCONS

Patch Pure Data – Speech icons, Earcons

Grazie alla modularità con cui è stata sviluppata la patch Pure Data con auditory icons descritta in precedenza, per integrare gli earcons, le speech icons e le spearcons, è sufficiente sostituire i nuovi file audio contenenti appunto earcons e speech icons al posto di quelli utilizzati nell'implementazione auditory icons, dal momento che il meccanismo ed il funzionamento delle varie implementazioni rimane esattamente lo stesso e perciò non sono richieste ulteriori modifiche.

3.3 IMPLEMENTAZIONE SPAZIALIZZAZIONE BINAURALE

Negli ultimi anni, il suono 3D sta diventando sempre più una caratteristica importante nelle applicazioni di intrattenimento; il grado di coinvolgimento raggiunto tramite film e videogames si basa essenzialmente su effetti audio realistici e pervasivi, che possono essere considerati una simulazione virtuale di un vero suono ambientale.
Secondo quanto emerge in una delle definizioni di Virtual Reality, la simulazione non coinvolge solo un ambiente virtuale ma anche una vera e propria esperienza immersiva; invece di parlare di percezione basata sulla realtà, per Virtual Reality si intende una realtà alternativa basata sulla percezione. Un'esperienza immersiva si avvantaggia di ambienti che realisticamente riproducono le realtà che devono essere simulate.
Il problema di riprodurre audio 3D non è affatto triviale; con un sistema basato su cuffie standard, il suono sembra essere generato all'interno della testa dell'ascoltatore. Ciò è risolvibile tramite la spazializzazione binaurale che è in grado di veicolare una buona percezione realistica tridimensionale di una sorgente sonora S, posizionata da qualche parte attorno all'ascoltatore L.
Al giorno d'oggi molti progetti di audio-rendering usano la spazializzazione binaurale per riprodurre le sorgenti sonore animate S, e tengono sotto controllo la posizione dell'ascoltatore L (che spesso rimane fissa al fine di semplificare il modello).
Tuttavia, per fruire di un'esperienza immersiva, questo approccio non è sufficiente, infatti è necessario conoscere la posizione e l'orientazione dell'ascoltatore in relazione allo spazio virtuale per poter fornire un segnale consistente, cosicché le sorgenti sonore possano rimanere fisse nello spazio virtuale, indipendentemente dai movimenti della testa, come accade con l'ascolto quotidiano.
È opportuno perciò considerare un sistema capace di tracciare e determinare la posizione della testa dell'ascoltatore L rispetto allo spazio, e di modificare il segnale diretto alla cuffie sulla base di tale tracciamento. Il sistema è dunque in grado di verificare la posizione delle sorgenti S rispetto all'ascoltatore L e reagire ai movimenti dello stesso L.
I sistemi audio tipicamente usano tecnologie di head-tracking magnetici, complice il fatto che possono controllare 360° di rotazione e forniscono ottime performance. Sfortunatamente, a causa della necessità di hardware dedicato molto complesso, questi sistemi si possono utilizzare solamente per sperimentazione e ricerca. Però, grazie all'incremento della potenza di calcolo dei computer domestici, negli ultimi anni si sta sviluppando una nuova generazione di head-tracker ottici basati sull'utilizzo di webcams o sensori di profondità come la Microsoft kinect.
La soluzione adottata fa uso per l'appunto della Microsoft Kinect per registrare i dati relativi alla testa dell'ascoltatore, mentre il paper da cui si è tratto spunto: ([53]) "Head in space: a head tracking based bineural spatialization system" (Luca A. Ludovico, Davide A. Mauro, e Dario Pizzamiglio, LIM) usa

un'implementazione basata su un sistema di tracciamento via webcam e una spazializzazione binaurale che sfrutta la convoluzione dei segnali.

3.3.1 Spazializzazione Binaurale

La spazializzazione binaurale è una tecnica che mira alla riproduzione di un suono reale dell'ambiente usando esclusivamente due canali (come avviene, del resto, nella registrazione stereo). E' basata sull'assunzione che il nostro udito ha solo 2 ricevitori, le orecchie; perciò, inviando un segnale uguale (o quasi uguale) a quello che un ascoltatore riceverebbe in un ambiente reale, si riesce a indurre nell'ascoltatore la medesima percezione, ottenendo così un'esperienza audio veramente immersiva.
Il sistema audio gestisce vari task simultanei detti "cues" (principalmente basati sui parametri fisici del segnale di interesse) per ottenere una rappresentazione dell'acustica ambientale.
La spazializzazione binaurale può essere raggiunta tramite vari processi come ad esempio: le equalizzazioni e delay, o la convoluzione con la risposta impulsiva della testa HRIR.
L'ultimo approccio è quello di maggiore interesse. Al fine di ottenere questi impulsi, molti esperimenti che prevedono l'uso di dummy head[13] sono stati fatti nel tempo e sono stati creati database di risposte impulsive. Molte di esse usano una distanza fissata solitamente di 1 metro tra S sorgente e L ascoltatore.

Ai fini applicativi, si assume che l'head tracking sia una black box che restituisce all'applicazione un insieme di parametri riguardanti la posizione della testa, ad un rate ben preciso.
In input ci sono due insiemi di parametri che servono a definire:

1) la posizione dell'ascoltatore, e
2) la posizione della sorgente audio.

Considerando queste informazioni, insieme alla posizione del sensore di tracciamento, è possibile calcolare la posizione reciproca dell'ascoltatore rispetto alla sorgente, in termini di ***azimuth, elevazione e distanza***. Questo è tutto ciò di cui il sistema ha bisogno per scegliere quale risposta impulsiva usare per la spazializzazione.

[13] La dummy head è un manichino che reproduce la testa umana.

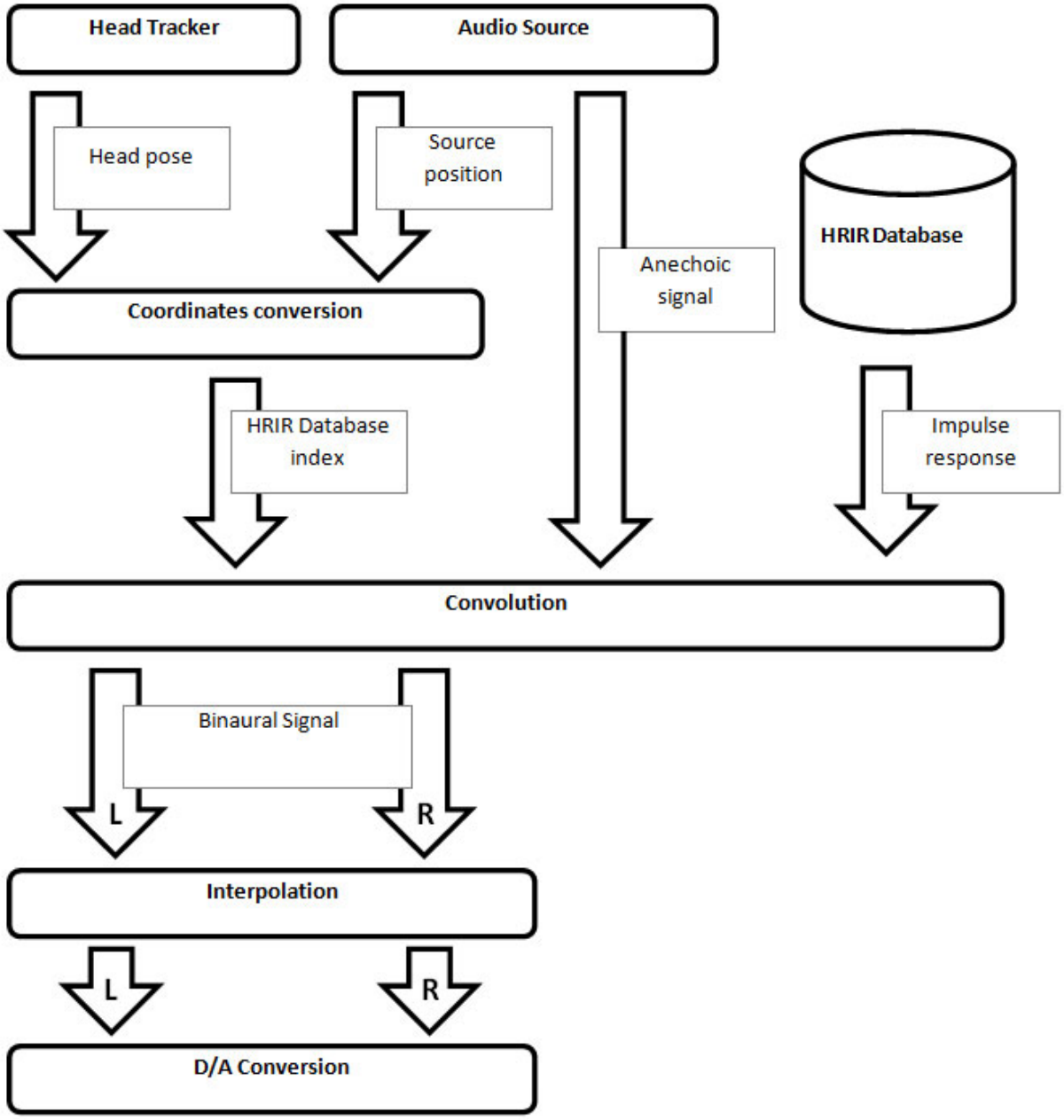


Figura 36 schema generale per la spazializzazione

Una volta che si ottiene la corretta HRIR dal database, è possibile eseguire la convoluzione tra un segnale audio mono in input e la risposta impulsiva stereo. Dal momento che le posizioni sia dell'ascoltatore che della sorgente possono cambiare nel tempo, si utilizza un apposito meccanismo di interpolazione che effettua lo switch tra due differenti HRIR per garantire dinamicità al sistema.

Sistema delle Coordinate

Il meccanismo di spazializzazione usa un sistema di coordinate sferiche che ha l'origine al centro della testa dell'ascoltatore. La sorgente è identificata da una misura di distanza e 2 angoli: l'azimuth sul piano orizzontale e l'elevazione sul piano mediano. Le distanze angolari sono espresse in gradi e salvate nella

patch tramite variabili integer, mentre la distanza è espressa in metri e ed il suo valore è salvato tramite un float.
Il tracciamento della testa presenta coordinate in forma cartesiana che hanno l'origine sul cono di proiezione della webcam. In questo modo la rappresentazione delle coordinate del sistema di spazializzazione e quelle del tracciamento della testa sono differenti e c'è bisogno di una procedura di conversione che, per prima cosa, esegue una rototraslazione del sistema al fine di ottenere le nuove coordinate di traslazione, sia della sorgente che della testa, dentro un sistema di riferimento rettangolare.
Rispetto ai contesti in cui si adotta una webcam, nella seguente implementazione si utilizza una kinect, perciò le coordinate del sensore sono già rettangolari, ed il centro del sistema coincide con il centro del sensore stesso, quindi non c'è bisogno della conversione intermedia in cui le coordinate della webcam sono trasformate nelle coordinate rettangolari.

Il passaggio alle coordinate sferiche da quelle cartesiane fa uso delle seguenti formule:

$$distance\ \rho = \sqrt{x^2 + y^2 + z^2}$$

$$azimuth\ \varphi = \arctan\left(\frac{z}{x}\right)$$

$$elevation\ \theta = \left(\frac{y}{\sqrt{x^2 + y^2 + z^2}}\right)$$

$$R = \begin{Vmatrix} \cos(R_y)\cos(R_z) & \cos(R_x)\sin(R_z) + \sin(R_x)\sin(R_y)\cos(R_z) & \sin(R_x)\sin(R_z) - \cos(R_x)\sin(R_y)\sin(R_z) \\ -\cos(R_y)\sin(R_z) & \cos(R_x)\cos(R_z) - \sin(R_x)\sin(R_y)\sin(R_z) & \sin(R_x)\cos(R_z) + \cos(R_x)\sin(R_y)\sin(R_z) \\ \sin(R_y) & -\sin(R_x)\cos(R_y) & \cos(R_x)\cos(R_y) \end{Vmatrix}$$

Il nuovo sistema di coordinate può essere impiegato per recuperare l'HRIR corretta dal database. Poiché il database include solamente HRIR misurate a una data distanza, si usano solo le componenti azimuth ed elevazione. Dal momento che non tutte le coppie azimuth-elevazione hanno un corrispondente, si sceglie la corrispondenza sulla base della minima distanza euclidea.

Convoluzione e Interpolazione
Il processo di convoluzione avviene tra un segnale aneconico (cioè che è in grado di assorbire le onde sonore senza rifletterle, privo di eco) e un'HRIR binaurale. Il paper "Head in space: a head tracking based bineural spatialization system" utilizza il CIPIC database, ovvero un insieme di risposte impulsive per 45 soggetti a 25 differenti valori di azimuth e 50 differenti valori di elevazione. Il processo è eseguito prima una volta per il canale sinistro e poi un'altra per il canale destro. Azimuth ed elevazione sono misurati nel database con una numerazione ad hoc

Uno dei problemi correlati all'uso di HRIR per la spazializzazione è l'interpolazione tra due segnali convoluti con due differenti impulsi. Questo è un caso comune per le applicazioni real-time perché quando ci si muove da un valore di azimuth ad un altro, gli impulsi sono molto differenti. Ne consegue che il segnale in output cambia bruscamente, affliggendo negativamente la qualità percepite dell'intero sistema. Una procedura di interpolazione basta sul

cross-fade è utile per limitare gli artefatti prodotti dallo switch tra un impulso e l'altro.
Negli ambienti real-time ogni operazione ridondante dovrebbe essere eliminata per migliorare la reattività e le performance. 40msec è il valore di tempo usato per definire la durata del crossfade scelto tra i sample.

Simulazione della Distanza
Una delle maggiori limitazioni del CIPIC database è quella che presenta candidati solo a una certa distanza. Per simulare l'effetto distanza si usa una semplice procedura, basata sulla legge dell'inverso del quadrato. L'espressione è la seguente:

$$20\ log_{10}\left(\frac{1}{distance}\right) dB$$

Il range dei valori di distanza è limitato dal sistema di tracciamento della testa tra 0.1 e 2 metri. Convenzionalmente 1metro identifica la distanza di riferimento delle risposte impulsive e in questo caso non si ha guadagno. Il processo potrebbe essere ulteriormente migliorato aggiungendo un filtro che simula l'assorbimento dell'aria o tramite l'utilizzo di un database dove HRIR sono misurate a varie distanze.

3.3.2 Patch Pure Data – Spazializzazione Binaurale

Il tracciamento della testa è inteso come una "black box", in quanto alla patch di Pure Data arrivano tutte le informazioni relative alle coordinate della testa tramite pacchetti OSC provenienti dall'applicazione che si occupa dell'interfacciamento con il sensore kinect.

La patch Pure Data che implementa la spazializzazione, è un'evoluzione di quella utilizzata per la riproduzione di auditory icons monofonica. L'interfaccia e la struttura di base rimangono esattamente le stesse, si aggiungono essenzialmente due componenti: una parte di oggetti PD destinati al calcolo del centroide dell'oggetto puntato, e una subpatch che si occupa dell'introduzione di tutti gli effetti di spazializzazione.

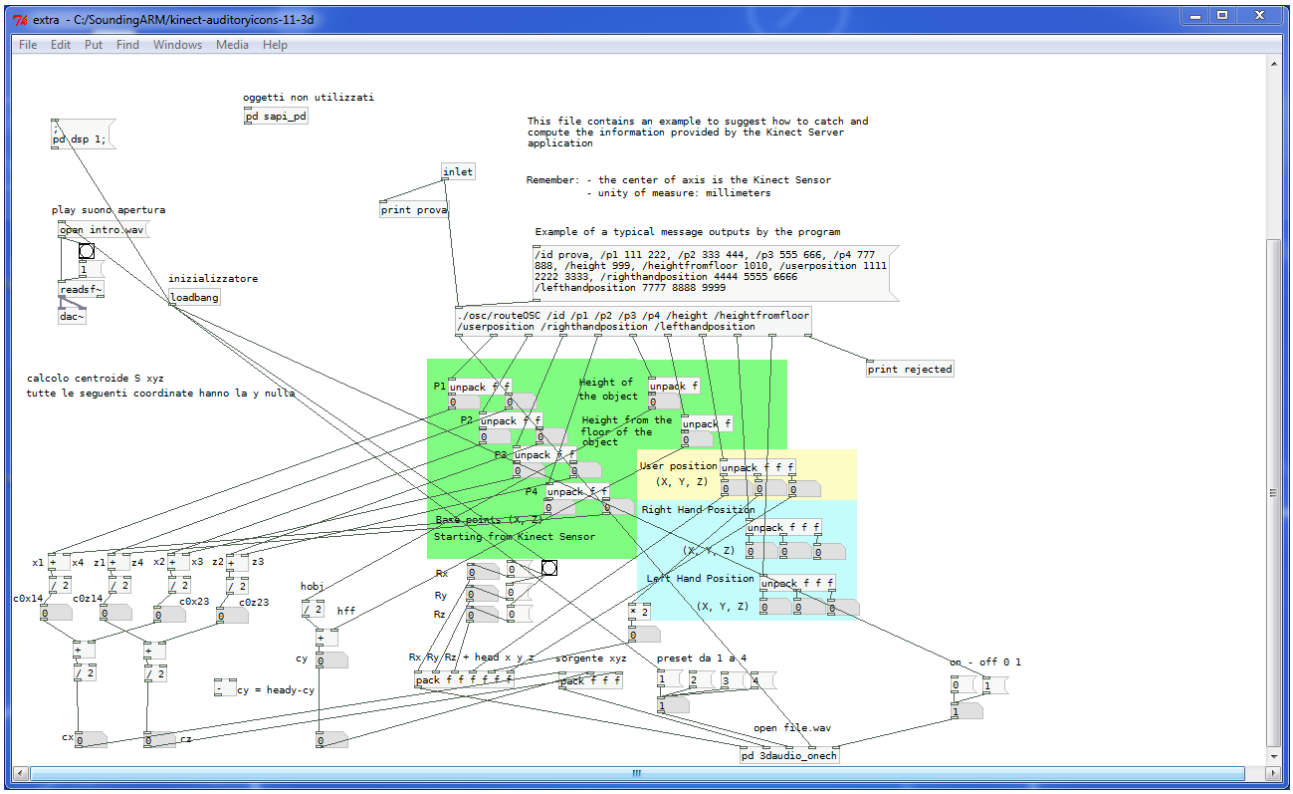


Figura 37 screenshot subpatch extra versione 3daudio (calcolo centroide in basso a sinistra, collegamenti a 3daudio_onech in basso al centro

Calcolo del Centroide

Dal momento che l'applicazione di spazializzazione necessita della posizione nello spazio della sorgente sonora, è opportuno calcolare il centroide corrispondente a ciascun oggetto indicato, ovvero l'intera sorgente sonora viene condensata in un unico punto nello spazio 3D, e per ottenere una mappatura la più fedele possibile, tale punto deve trovarsi esattamente al centro del parallelepipedo che comprende l'intero oggetto puntato.

A livello implementativo, tutte le informazioni riguardanti il parallelepipedo che ricopre l'oggetto puntato arrivano direttamente dal Kinect application Server sotto forma di messaggio OSC e vengono scompattati e resi disponibili per l'elaborazione dall'oggetto Pd routeOSC.

Per calcolare il centroide, l'idea è quella di trovare il centro del quadrilatero costituito dai punti di base P1,P2,P3,P4, facendo la media tra i corrispettivi valori di X e Z non adiacenti; mentre per quanto riguarda la Y, si sommano rispettivamente i valori degli attributi HeightOfObject e HeightFromFloor e si divide per 2 il risultato.

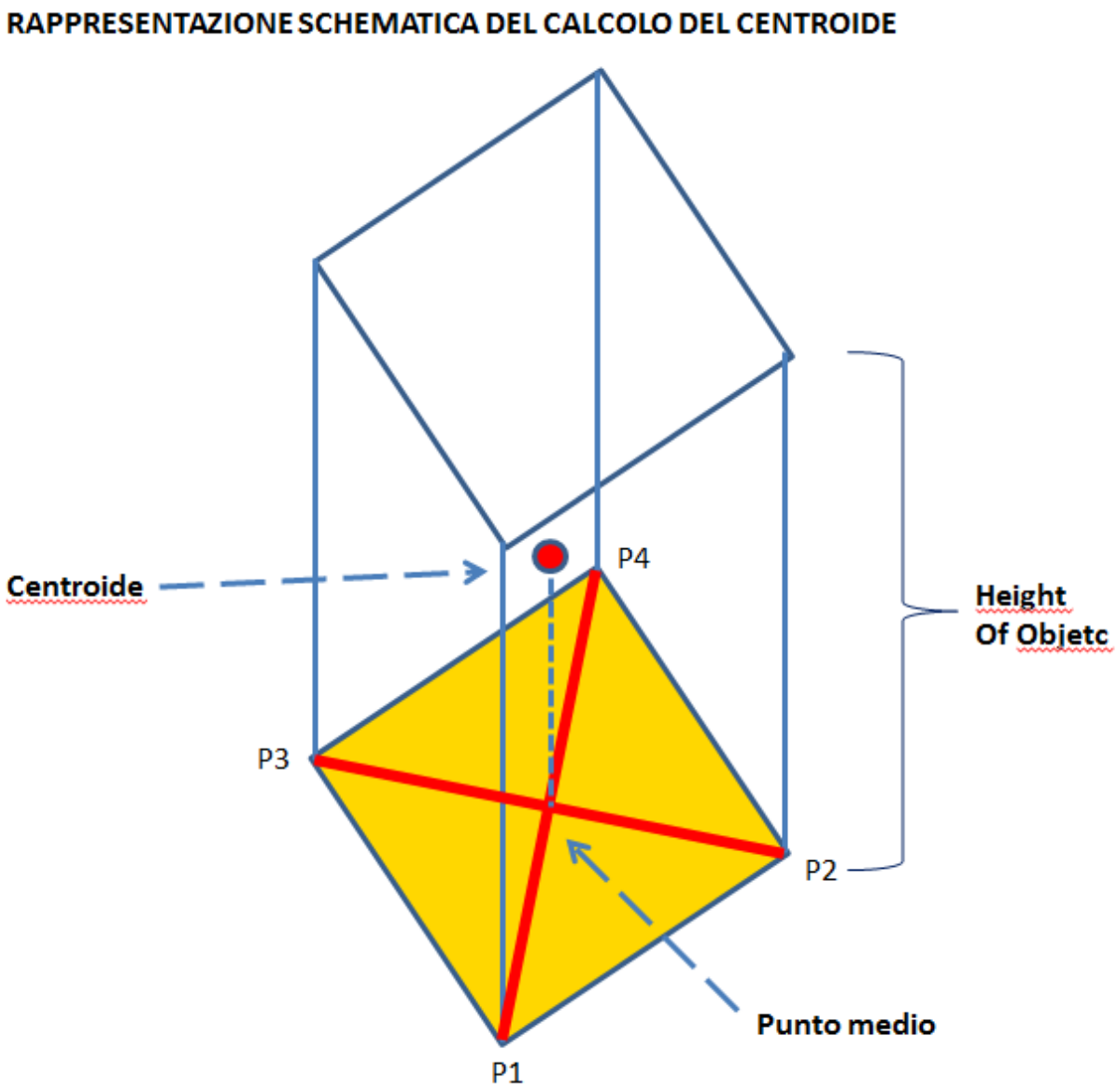


Figura 38 calcolo del centroide

Subpatch 3daudio_onech

Questa subpatch comprende l'intero algoritmo che permette di ottenere la spazializzazione binaurale, l'algoritmo che aggiunge l'effetto di riverbero e anche il riproduttore readsf~ utilizzato per riprodurre i segnali audio opportunamente rielaborati.

Innanzitutto si analizzano i 5 inlet della subpatch:

- ✓ **inlet 1:** riceve I dati provenienti da un oggetto pack che raggruppa nell'ordine i valori:
  - **RX RY RZ** ovvero i coefficienti di rotazione della testa (che nella seguente implementazione verranno sempre messi a 0 poiché il sensore di profondità di cui è dotata la kinect non è in grado di distinguere l'orientazione della testa in quanto si riduce a un solo punto);
  - **Userposition X Y Z**, ovvero le coordinate dell'utente che sta effettuando il puntamento in quel preciso momento.

- ✓ **inlet 2**: riceve in ingresso I dati provenienti da un oggetto pack che raggruppa i valori **sX sY sZ** che sono rispettivamente le coordinate X Y Z del centroide precedentemente calcolato, e rappresentano cioè la posizione della sorgente sonora.
- ✓ **inlet 3:** seleziona uno dei 4 preset differenti che si possono utilizzare per gestire il riverbero e la simulazione della distanza.
- ✓ **inlet 4:** riceve in ingresso, direttamente dalla outlet più a sinistra di routeOSC, il messaggio Pd contenente il commando "open filename.wav" corrispondente all'oggetto indicato.
- ✓ **inlet 5:** riceve il commando di on-off che abilita e disabilita il riproduttore readsf~.

La subpatch 3daudio_onech è molto ampia ed è conveniente procedere ad una rapida descrizione per sezioni.

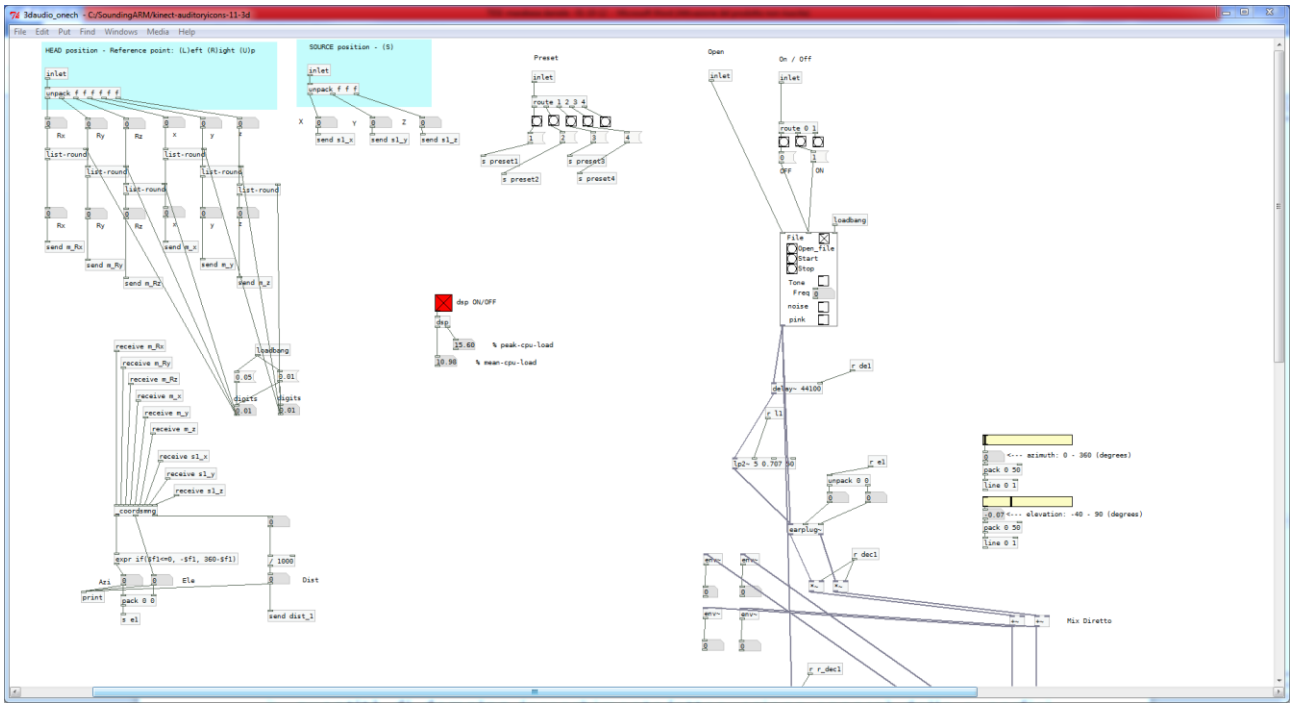


Figura 39 screenshot 3daudio_onech – parte in alto a sinistra

Nella parte in alto a sinistra gli inlet riguardanti RXRYRZ, userpositionXYZ e sourceposition sXsYsZ sono mandati a loro volta in input a un'altra subpatch _coordsmng che si occupa della conversione delle coordinate: da quelle cartesiane tridimensionali a quelle azimutali, in particolare, restituisce i 3 valori di: azimuth (°), elevazione (*) e distanza (metri).
Di seguito si riportano gli screenshot anche delle subpatch che sono comprese in _coordsmng.

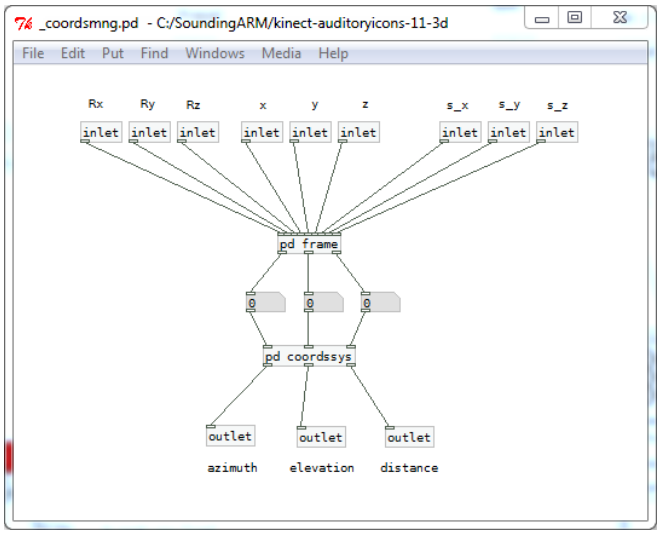


Figura 40 screenshot subpatch _coordsmng

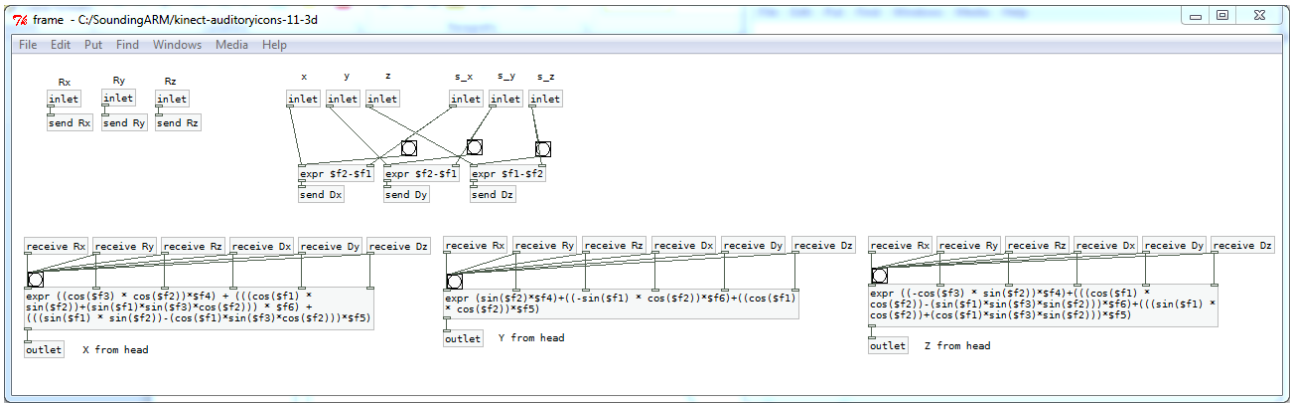


Figura 41 screenshot subpatch frame (subpatch di _coordsmng)

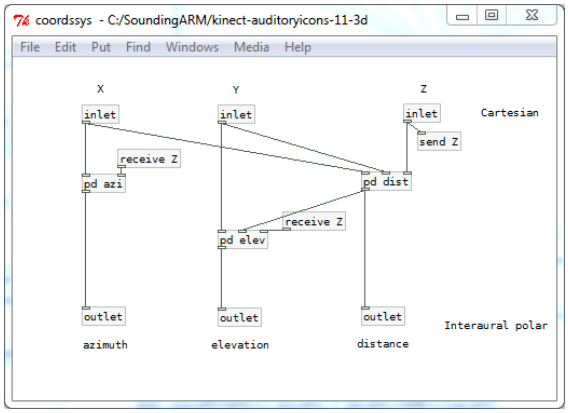


Figura 42 screenshot della subpatch coordssys (subpatch di _coordsmng)

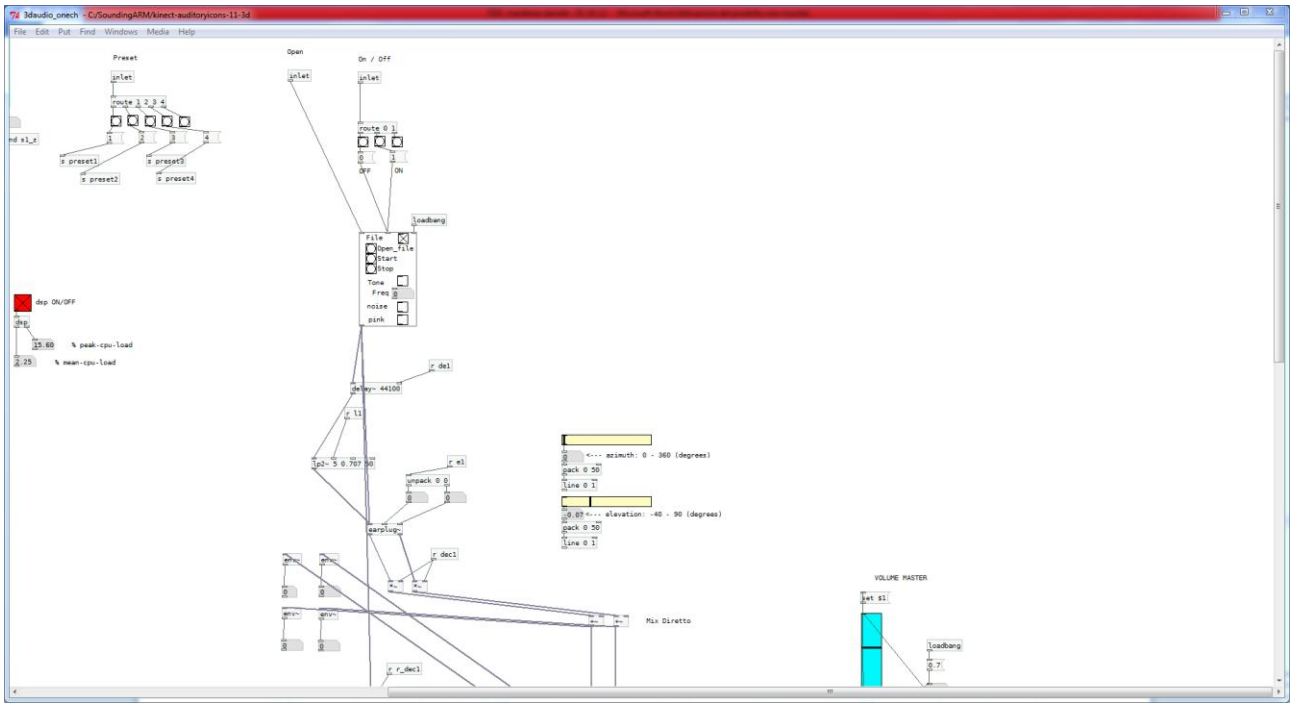


Figura 43 screenshot 3daudio_onech - parte in alto a destra

La parte destra della patch è occupata dal riproduttore di suoni e dalla sezione che si occupa dell'integrazione del riverbero nella riproduzione stessa.

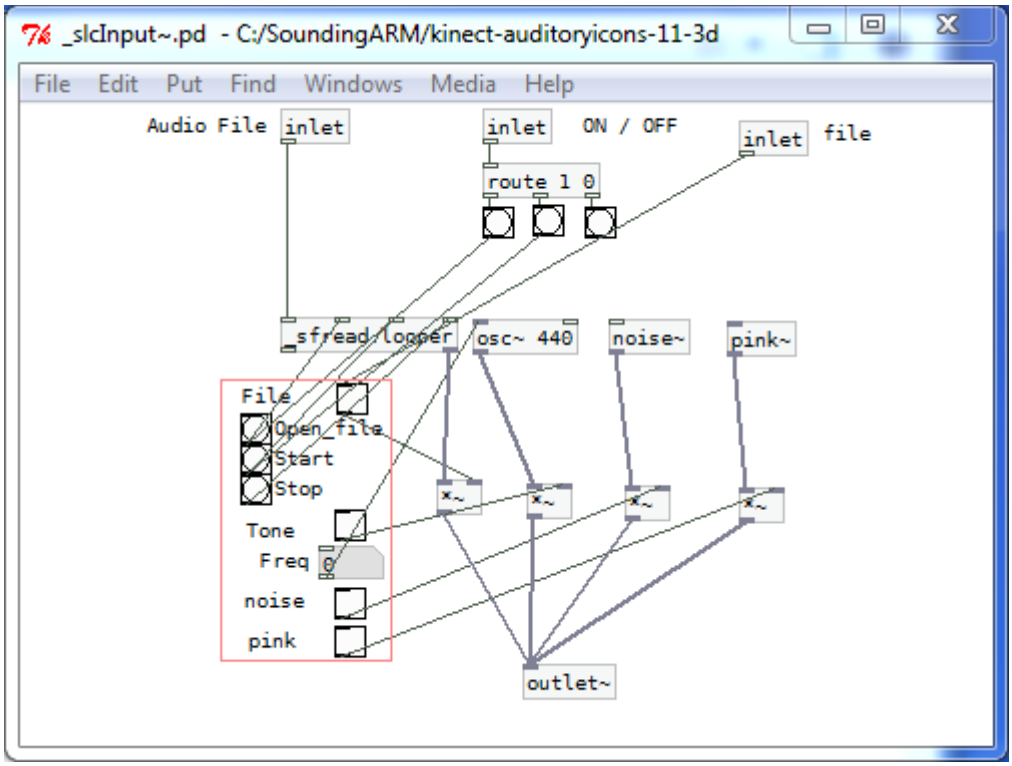


Figura 44 screenshot riproduttore file audio e noise

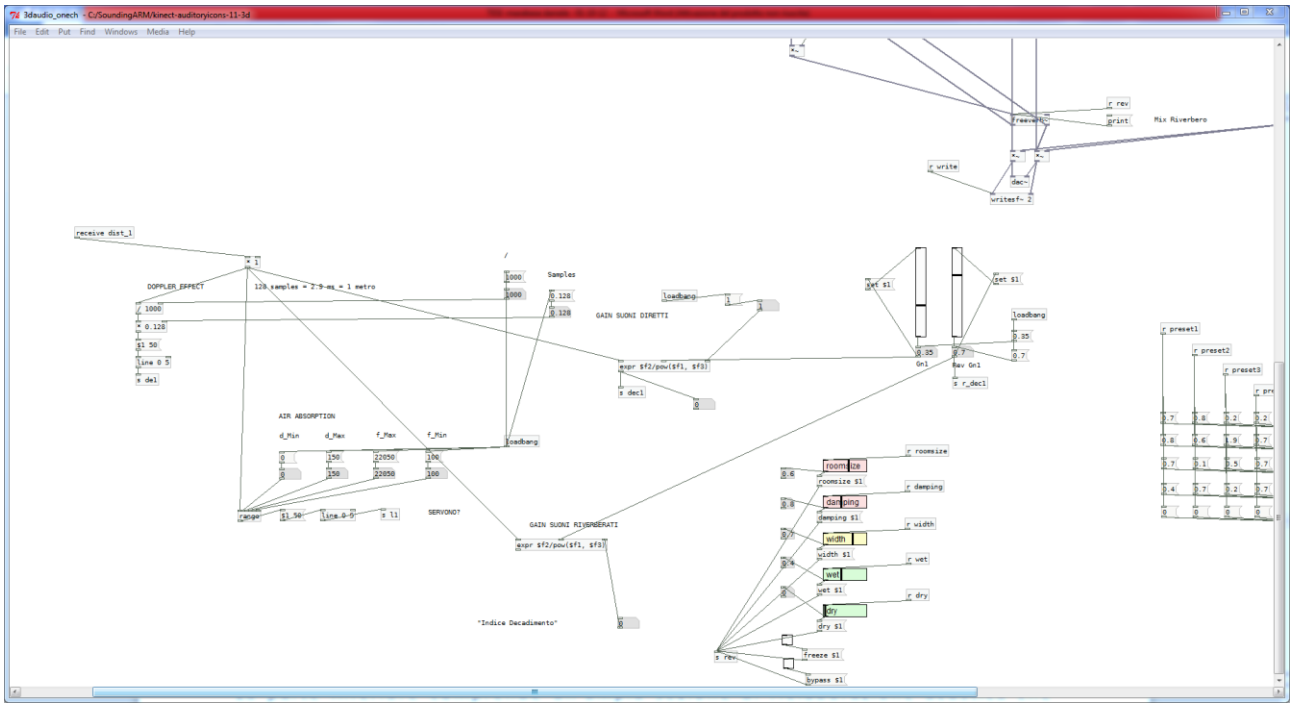


Figura 45 screenshot 3daudio_onech - parte in basso a sinistra

La parte inferiore comprende un'ampia sezione di rielaborazione acustica che permette di inserire differenti effetti quali: l'effetto doopler, la simulazione dell'assorbimento dell'aria con relativo indice di decadimento e guadagni dei suoni diretti e riverberati.

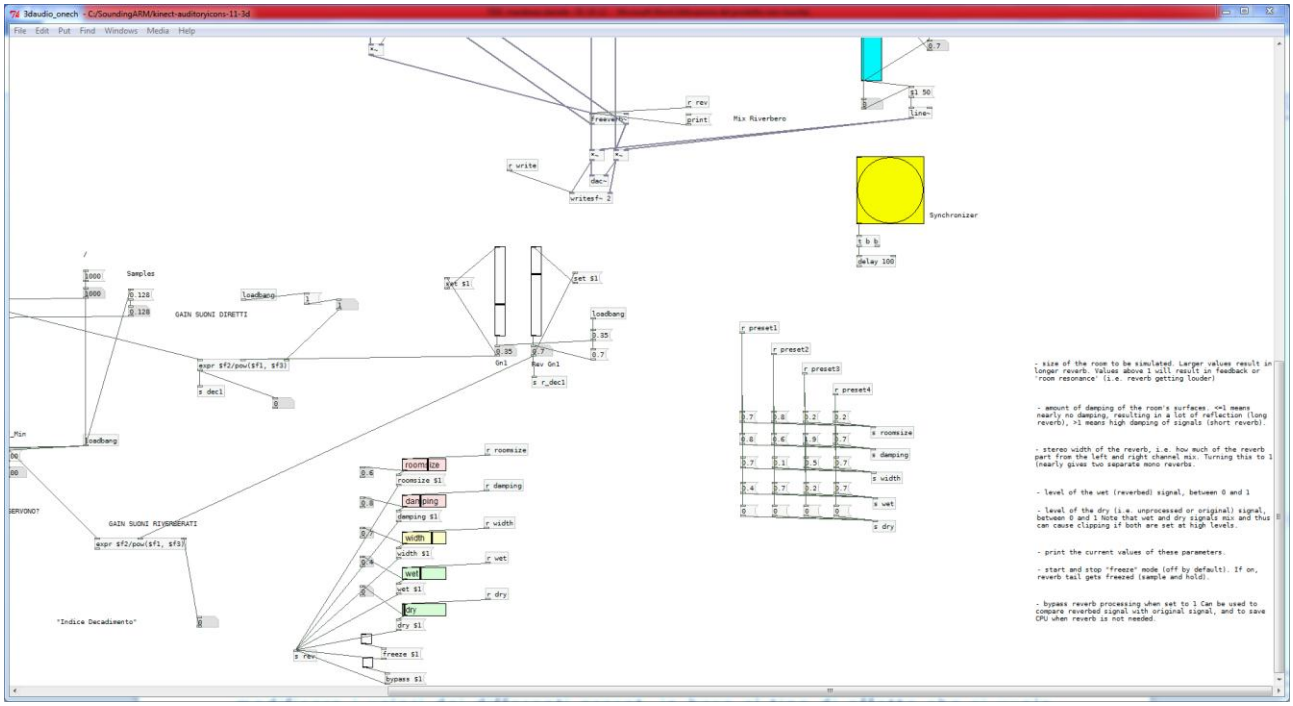


Figura 46 screenshot 3daudio_onech - parte in basso a destra

Infine, in basso a destra, si individua un'area dove si possono settare e modificare i valori dei differenti preset, in base al tipo di effetto che si vuole ottenere.

I parametri su cui si può intervenire per migliorare la qualità della spazializzazione sono i seguenti:

- **dimensione della stanza da simulare**, valori maggiori restituiscono reverberi più lunghi. I valori superiori a 1 comportano un effetto di risonanza della stanza (il riverbero si fa più forte).
- **quantità di damping** (assorbimento/attenuazione sonora) delle superfici della stanza:
  - <=1 significa pressoché assenza di damping, che comporta parecchia riflessione (riverbero lungo),
  - >1 significa alto livello di damping del segnale (riverbero corto).
- **larghezza stereo del riverbero**, permette di definire quanta parte del riverbero viene mixata a sinistra o a destra. Impostando questo parametro a 1 si ottengono due riverberi mono separati.
- **livello di wet reverber** del segnale, compresa tra 0 e 1.
- **livello di dry reverber** del segnale originale e non processato, compreso tra 0 e 1.
  Da notare che il mix di segnali wet e dry può causare fenomeni di clipping se entrambi settati ad alti livelli.
- **start & stop "freeze" mode** (disabilitata di default). Se on, la coda di riverbero viene bloccata.
- **bypass reverb processing**, quando viene settato a 1, salta il processamento che aggiunge il riverbero. Questo parametro può essere usato per comparare i segnali riverberati con quelli originali, e permette di risparmiare tempo di CPU quando il riverbero non è necessario.

Capitolo 4

SPERIMENTAZIONE E VALUTAZIONE DEL SISTEMA

4.1 PRECISIONE MICROSOFT KINECT

4.1.1 Stima della precisione della kinect: puntamento orizzontale

Per stabilire la precisione del sensore Kinect in uno scenario d'uso caratteristico di SoundingARM, sono stati condotti alcune semplici prove in grado di evidenziare il rumore e la precisione del tracciamento dello scheletro.
E' opportuno precisare che il codice utilizzato per effettuare questo tipo di misurazioni è esattamente il medesimo di SoundingARM, opportunamente modificato in modo che restituisca in un file .txt tutti i dati relativi alla posizione della testa, della mano sinistra e della mano destra.

Test 1

Il primo test ha avuto come obiettivo quello di misurare il rumore che affligge il sensore. A tal proposito, è opportuno specificare come i dati provenienti dalla Kinect siano già opportunamente filtrati e corretti tramite le procedure di smooth implementate dalle classi del Kinect SDK, di conseguenza i valori che l'applicazione riceve hanno subito in precedenza un pre-processamento che elimina quasi del tutto le fluttuazioni e gli errori sporadici.
In questo test un individuo si pone di fronte al sensore, ad una distanza approssimativamente compresa tra 2.7 e 3.5 metri, e per tutto il tempo in cui vengono catturati e salvati i dati relativi alle posizioni di testa e mani, egli rimane perfettamente fermo senza compiere gesto alcuno.
In condizioni di totale assenza di rumore, i valori raccolti dovrebbero sovrapporsi perfettamente in modo da raffigurare un unico punto nel grafico, in quanto il soggetto rimane fermo, e nel tempo l'applicazione non dovrebbe registrare alcuno spostamento.
In pratica, dai grafici seguenti (unità di misura in mm), si può vedere come si registrino lievi scostamenti, nell'ordine di pochi centimetri, in particolare:

- ✓ nel primo grafico si rappresenta lo scostamento lungo gli assi X e Y, con una varianza complessiva lungo X di 41mm;
- ✓ nel secondo si effettua un'analoga rappresentazione però sugli assi Y e Z, con una varianza complessiva lungo Z di 32mm.

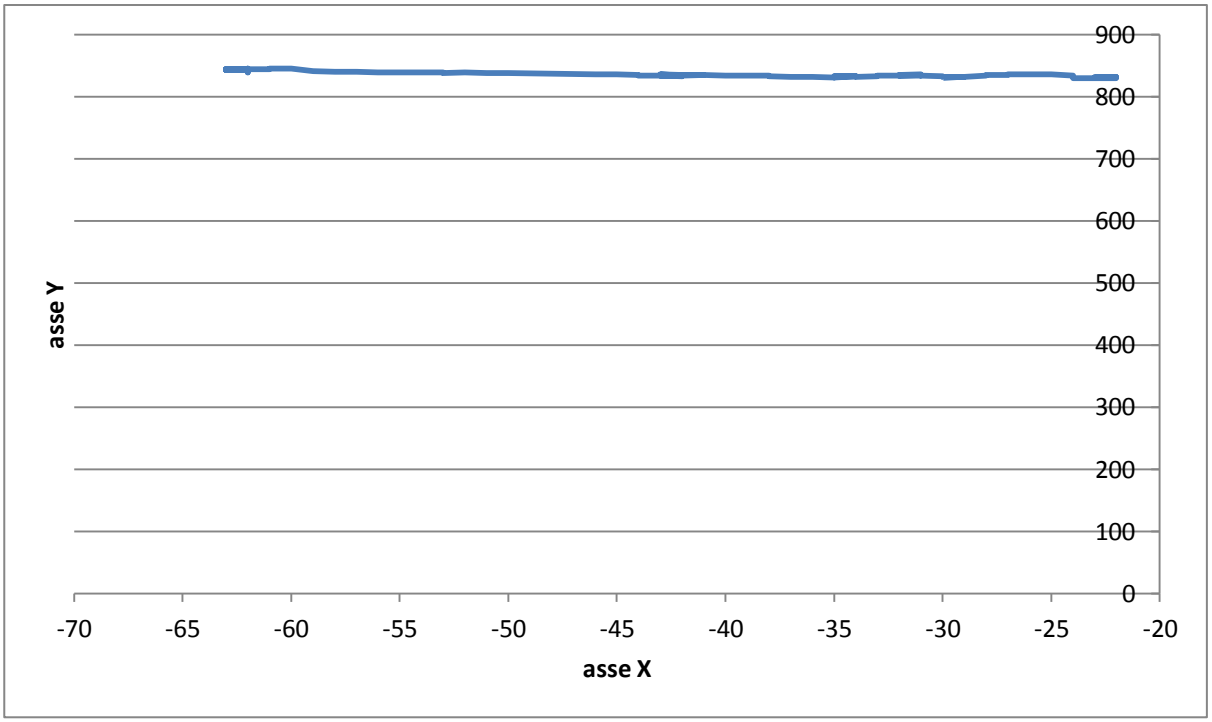


Figura 47 rumore lungo l'asse X

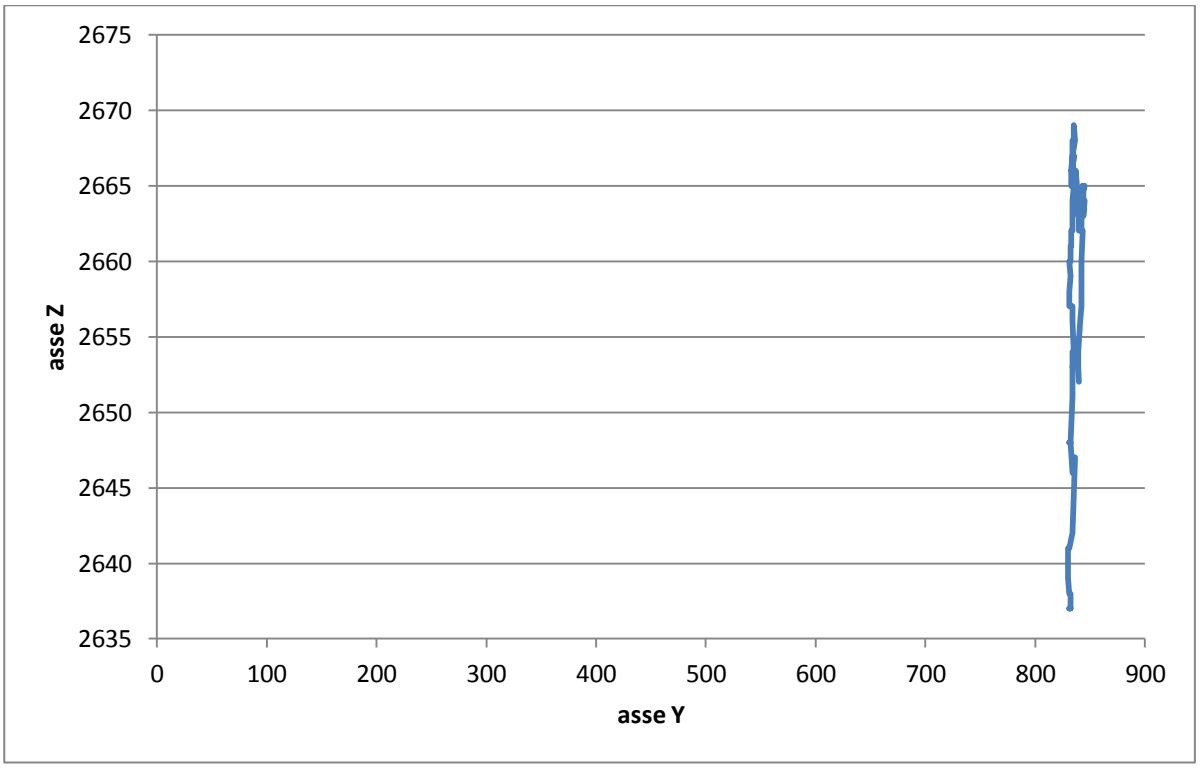


Figura 48 rumore lungo l'asse Z

Test 2

Nella prova successiva, l'individuo sta di fronte alla kinect, sempre in una posizione compresa tra 2.7 e 3.5 metri, e, cercando di mantenere il braccio destro il più rigido possibile, esegue un movimento di puntamento col medesimo braccio, da destra a sinistra, tracciando una sorta di semi-

circonferenza con centro la testa e raggio l'intero braccio in estensione, esattamente perpendicolare alla spalla.

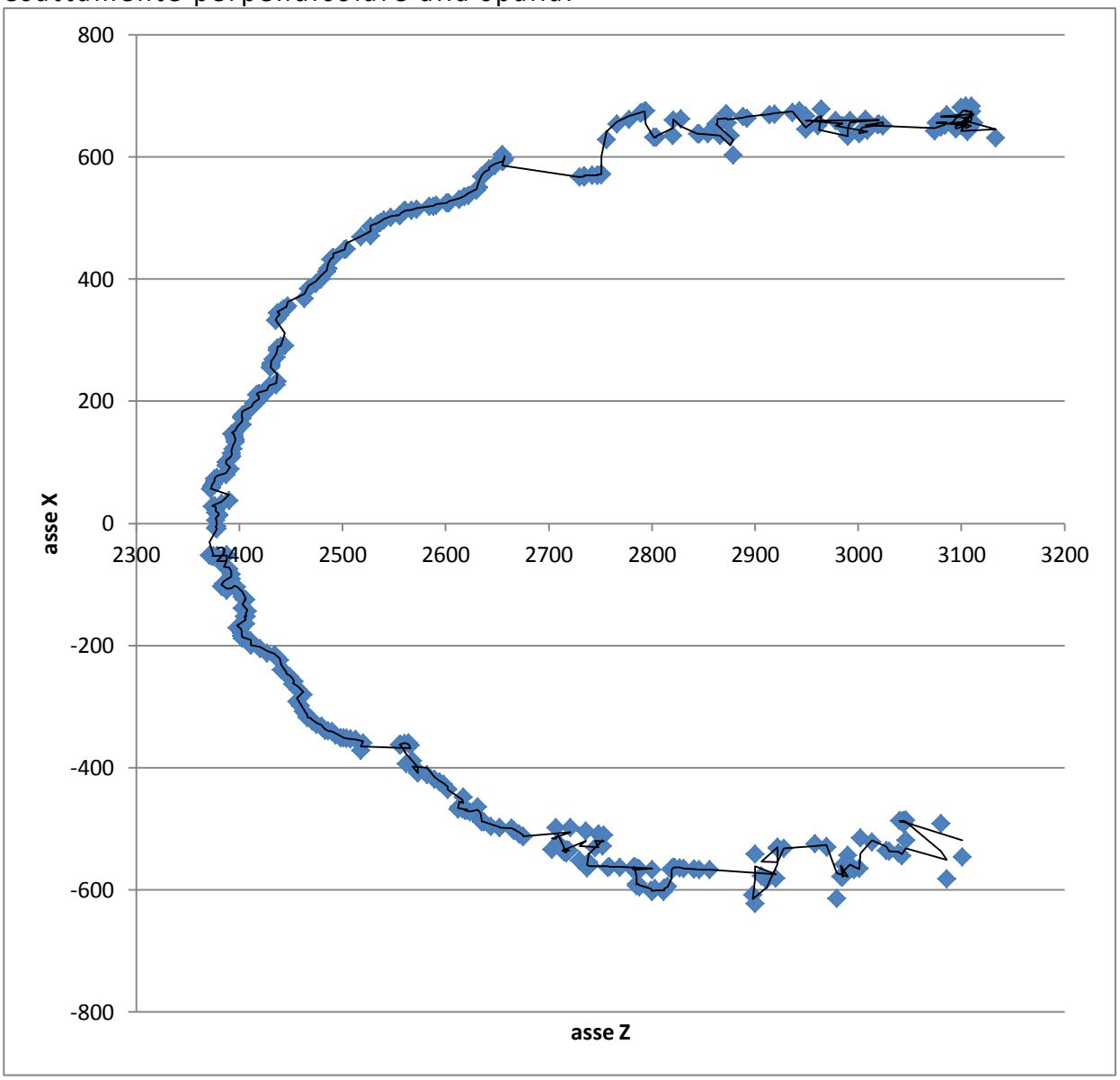


Figura 49 curva che approssima una semicirconferenza - braccio rigido

Osservando la scena dall'alto, quindi in un grafico Z-X, si può notare come, nonostante alcune oscillazioni dei valori, la figura disegnata, nel complesso, sia molto simile a quella di una semi-circonferenza (come evidenzia l'approssimazione a media mobile con periodo 2 in figura).
In tale contesto, si può notare come il sensore riesca a mantenere pressoché invariate la precisione sia sulla parte superiore che inferiore della semi-circonferenza, nonostante il fatto che, nella parte superiore, la Kinect effettui un tracciamento continuo e diretto di tutti i punti del braccio destro, mentre, nella parte inferiore, il sensore effettui un'interpolazione dei valori, poiché i punti del braccio destro si sovrappongono agli altri punti dello scheletro in prossimità del busto, del collo e delle spalle.

Test 3

L'ultimo test effettuato è essenzialmente uguale al precedente, con la differenza che il soggetto effettua sempre il movimento a semi-circonferenza da destra a sinistra ma senza tenere il braccio rigido e perpendicolare alla spalla. Questo scenario simula quanto accade realmente nelle fasi di utilizzo di SoundingARM.

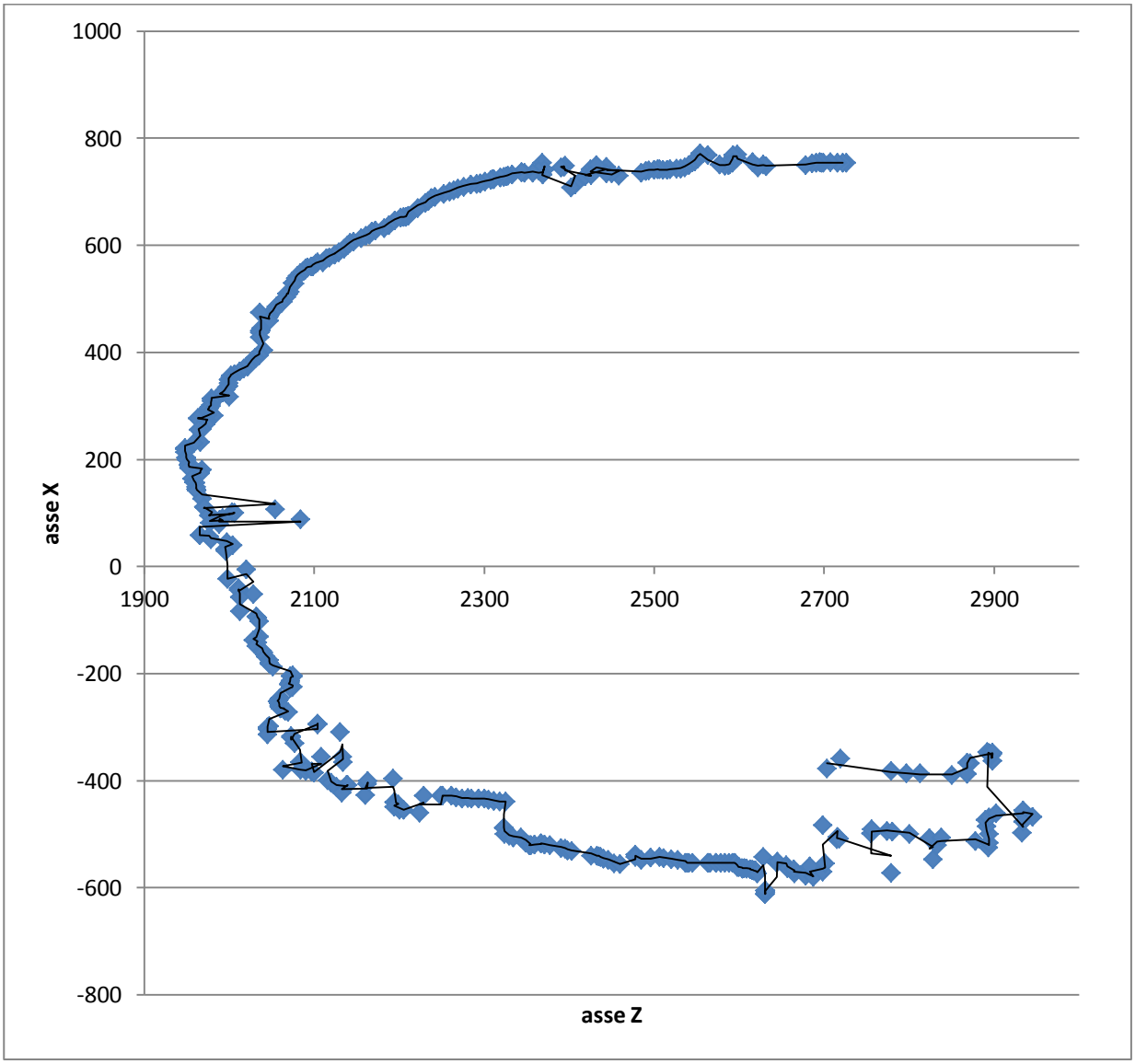


Figura 50 curva che approssima una semicirconferenza - braccio con movimento libero

I risultati mostrano essenzialmente un comportamento simile al caso precedente, con la differenza che la parte inferiore del grafico è caratterizzata da un andamento meno regolare, che si discosta a tratti da quello della semi-circonferenza, a testimonianza del fatto che con il braccio in posizione rilassata, l'interpolazione effettuata dal sensore è meno precisa.

4.1.2 Stima della precisione della Kinect: confronto con OptiTrack

Come evidenziato da Nicola Scattolin nella sua tesi ([52]): "A comparison between gesture tracking models and the development of an interactive mobility aid system for the visually impaired", per stabilire l'accuratezza del

sensore Kinect, sono stati effettuati degli esperimenti che hanno comparato tale sistema con l'OptiTrack, ritenuto il miglior sistema di tracciamento al mondo grazie a 12 camera a infrarossi che garantiscono una vista a 360° ad un rate fisso di 100Hz, e si sono poi confrontati i risultati di misura per determinare l'errore commesso dalla kinect nei vari scenari di test.
In estrema sintesi, gli esperimenti condotti sono stati i seguenti:

- l'utente esegue movimenti liberi in particolare con la mano;
- l'utente esegue movimenti pseudo-random simili ai precedenti ma i dati sono stati presi da differenti distanze;
- valutazione del delay della kinect.

L'accuratezza misurata circa la Kinect è di 4cm, in particolare l'errore lungo l'asse X è di circa 4cm, mentre lungo l'asse Z, a causa della triangolazione, è di circa 10cm, inoltre risulta evidente come non ci sia correlazione tra l'errore commesso e la distanza a cui si trova l'utente rispetto al sensore.

4.1.3 Stima dell'errore angolare del sensore per la sintesi binaurale

Sulla base di quanto affermato nel capitolo 2, SoundingARM calcola in modo dinamico la vista prospettica degli oggetti della stanza in relazione alla posizione dell'utente.
La discretizzazione introdotta dal programma è nell'ordine del grado °, infatti nella projectionSphere (ovvero la matrice bidimensionale 180x180) si considerano 180° in orizzontale e 180° in verticale, e sono ammissibili per il puntamento tutti gli angoli solidi compresi in tali bound.

Partendo da quanto affermato nel paragrafo 4.1.2, si è ritenuto interessante calcolare l'errore angolare introdotto dal sensore in relazione alla distanza, e in riferimento ai differenti assi cartesiani X, Y e Z.
Considerando costante l'errore del sensore, rispettivamente 4cm per gli assi X ed Y, 10cm per Z, e facendo variare la distanza dell'utente a step di 25cm, da un massimo di 5metri ad un minimo di 50cm, i passi per determinare l'errore angolare sono i seguenti:

1. si considera il triangolo rettangolo che ha come cateti:
   - l'errore (4cm o 10cm)
   - e la distanza dell'utente dal sensore (5m -> 50cm)
2. si determina l'ipotenusa del triangolo rettangolo;
3. chiamato α l'angolo che costituisce l'errore angolare di interesse, si utilizza la seguente formula:
$$sen\alpha = \frac{errore}{ipotenusa\ tr.}$$
4. si usa la funzione inversa:
$$arcsen\alpha$$
per trovare α.

Invece di considerare il triangolo rettangolo, si può centrare l'errore rispetto alla distanza, in particolare:

1. si considera il triangolo isoscele:
   - base 4 cm o 10cm,
   - altezza la distanza dell'utente dal sensore (5m -> 50cm)

2. si determina l'ipotenusa del triangolo rettangolo (metà isoscele);
3. chiamato $\alpha/2$ l'angolo che costituisce metà dell'errore angolare di interesse, si utilizza la seguente formula:

$$sen(\alpha/2) = \frac{errore}{ipotenusa\ tr.}$$

4. si usa la funzione inversa:

$$arcsen(\alpha/2)$$

per trovare $\alpha/2$.

5. infine si moltiplica $\frac{\alpha}{2} * 2$ per ottenere α.

Dal momento che l'errore compiuto dal sensore lungo l'asse X e Y è il medesimo, vengono riportate solo le tabelle per X, ma analoghi valori si possono ottenere applicando la stessa equazione anche sull'asse Y.

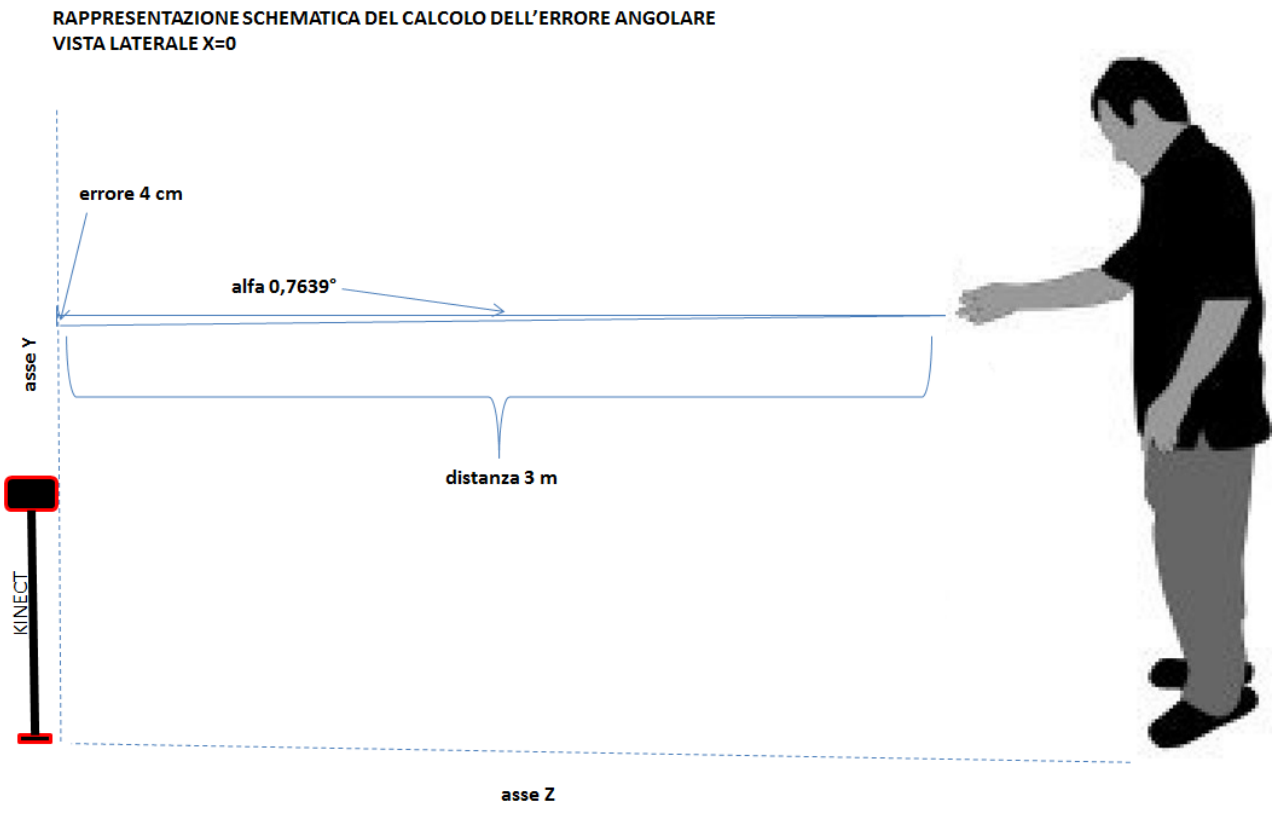


Figura 51 errore angolare - vista laterale

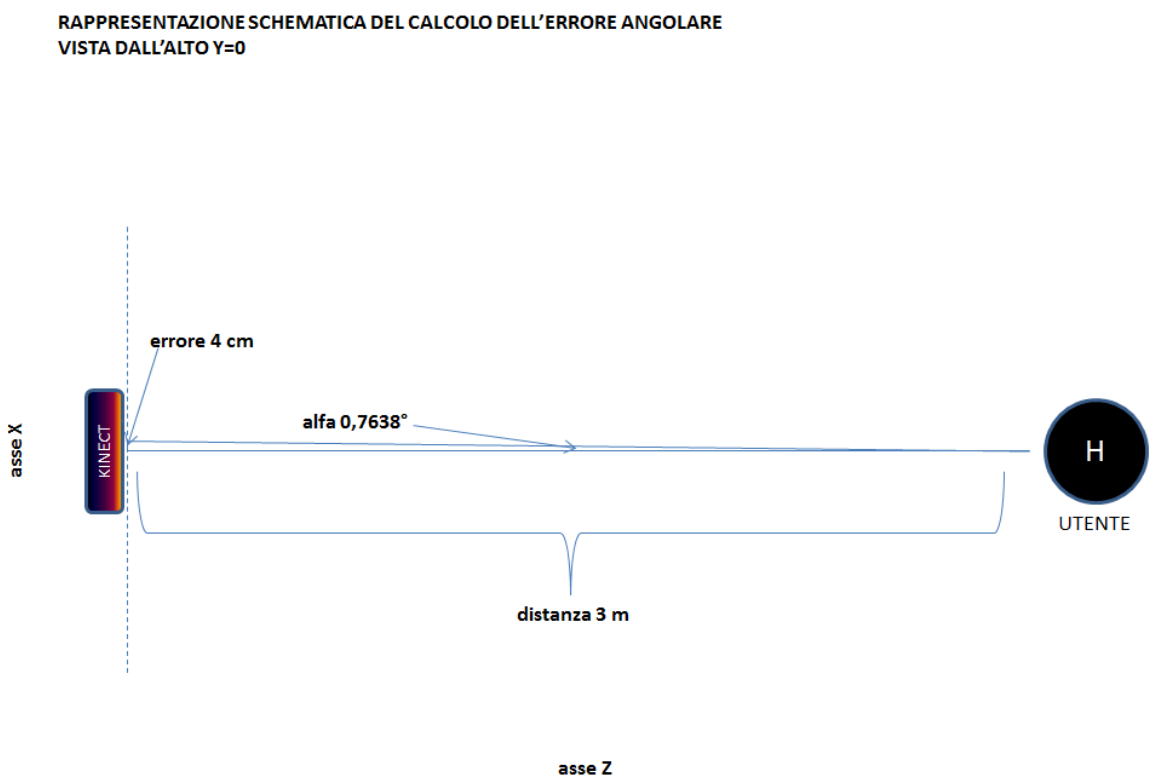


Figura 52 errore angolare - vista dall'alto - asse X – triangolo rettangolo

distanza (cm)	errore (cm)	3° lato tr. (cm)	sen-alfa	alfa °
500	4	500,016000	0,007999744	0,458356458
475	4	475,016842	0,008420754	0,48247937
450	4	450,017777	0,008888538	0,509282405
425	4	425,018823	0,009411348	0,539238474
400	4	400,020000	0,0099995	0,572938698
375	4	375,021333	0,01066606	0,611131804
350	4	350,022856	0,011427825	0,654780402
325	4	325,024614	0,01230676	0,705143221
300	4	300,026665	0,013332148	0,763898461
275	4	275,029089	0,014543916	0,83333439
250	4	250,031998	0,015997952	0,916654256
225	4	225,035553	0,017774969	1,018484348
200	4	200,039996	0,019996001	1,145762838
175	4	175,045708	0,022851174	1,309389819
150	4	150,053324	0,02665719	1,527525442
125	4	125,063984	0,031983629	1,832839506
100	4	100,079968	0,039968038	2,290610043
75	4	75,106591	0,053257643	3,052882515
50	4	50,159745	0,079745222	4,57392126

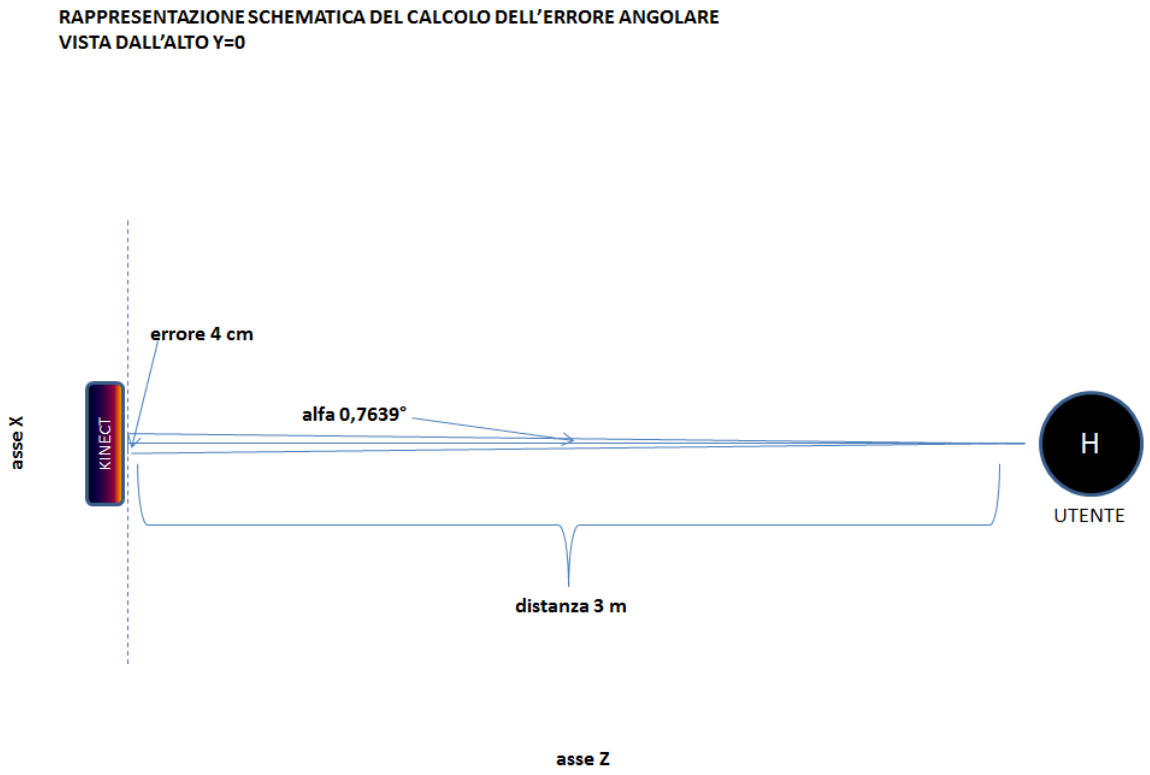


Figura 53 errore angolare - vista dall'alto - asse X – triangolo isoscele

distanza (cm)	errore (cm)	3° lato tr. (cm)	sen-alfa	alfa °/2	alfa °
500	2	500,004000	0,003999968	0,229181896	0,458363792
475	2	475,004211	0,004210489	0,241243962	0,482487924
450	2	450,004444	0,004444401	0,254646232	0,509292465
425	2	425,004706	0,00470583	0,269625207	0,539250415
400	2	400,005000	0,004999938	0,28647651	0,572953021
375	2	375,005333	0,005333257	0,305574593	0,611149187
350	2	350,005714	0,005714192	0,327400891	0,654801782
325	2	325,006154	0,00615373	0,352584962	0,705169923
300	2	300,006667	0,006666519	0,381966205	0,763932409
275	2	275,007273	0,007272535	0,416689232	0,833378464
250	2	250,008000	0,007999744	0,458356458	0,916712916
225	2	225,008889	0,008888538	0,509282405	1,01856481
200	2	200,010000	0,0099995	0,572938698	1,145877395
175	2	175,011428	0,011427825	0,654780402	1,309560805
150	2	150,013333	0,013332148	0,763898461	1,527796922
125	2	125,015999	0,015997952	0,916654256	1,833308513
100	2	100,019998	0,019996001	1,145762838	2,291525676
75	2	75,026662	0,02665719	1,527525442	3,055050884
50	2	50,039984	0,039968038	2,290610043	4,581220085

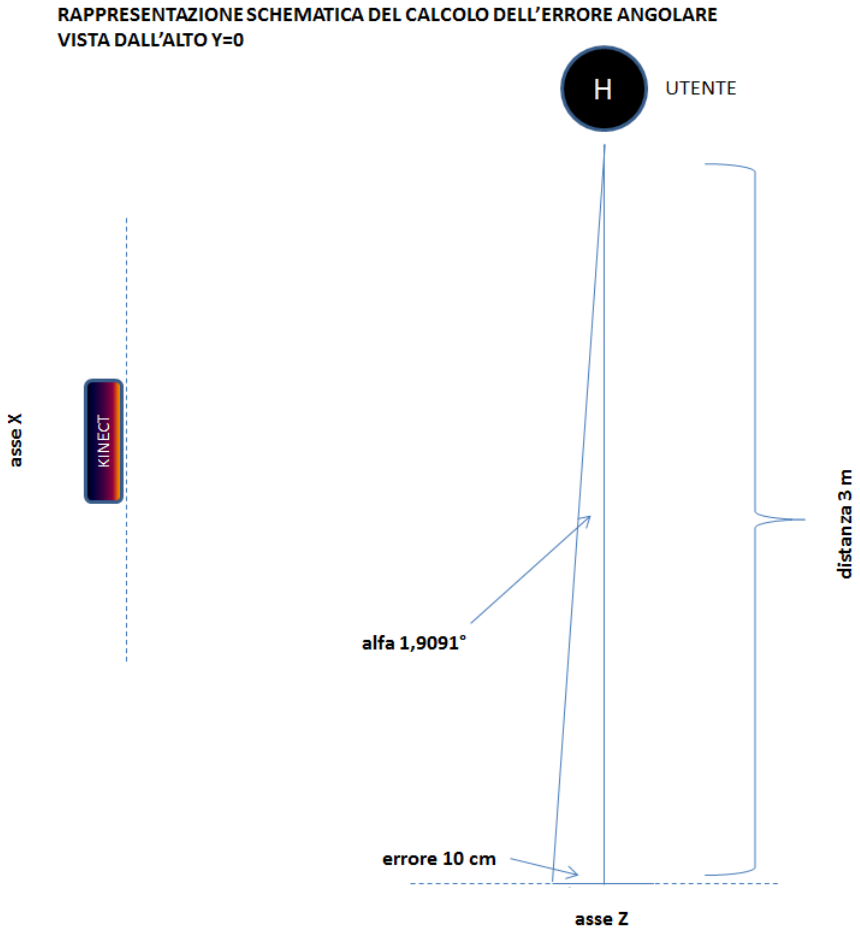


Figura 54 errore angolare - vista dall'alto - asse Z – triangolo rettangolo

distanza (cm)	errore (cm)	3° lato tr. (cm)	sen-alfa	alfa °
500	10	500,099990	0,019996001	1,145762838
475	10	475,105251	0,021047968	1,206048779
450	10	450,111097	0,022216737	1,27303002
425	10	425,117631	0,023522901	1,34788728
400	10	400,124980	0,024992191	1,432096184
375	10	375,133310	0,02665719	1,527525442
350	10	350,142828	0,028559774	1,636577042
325	10	325,153810	0,030754676	1,762391024
300	10	300,166620	0,03331483	1,909152433
275	10	275,181758	0,036339618	2,08256528
250	10	250,199920	0,039968038	2,290610043
225	10	225,222113	0,044400614	2,54480438
200	10	200,249844	0,049937617	2,862405226
175	10	175,285481	0,057049791	3,270487923
150	10	150,332964	0,066519011	3,814074834
125	10	125,399362	0,079745222	4,57392126
100	10	100,498756	0,099503719	5,710593137
75	10	75,663730	0,13216372	7,594643369
50	10	50,990195	0,196116135	11,30993247

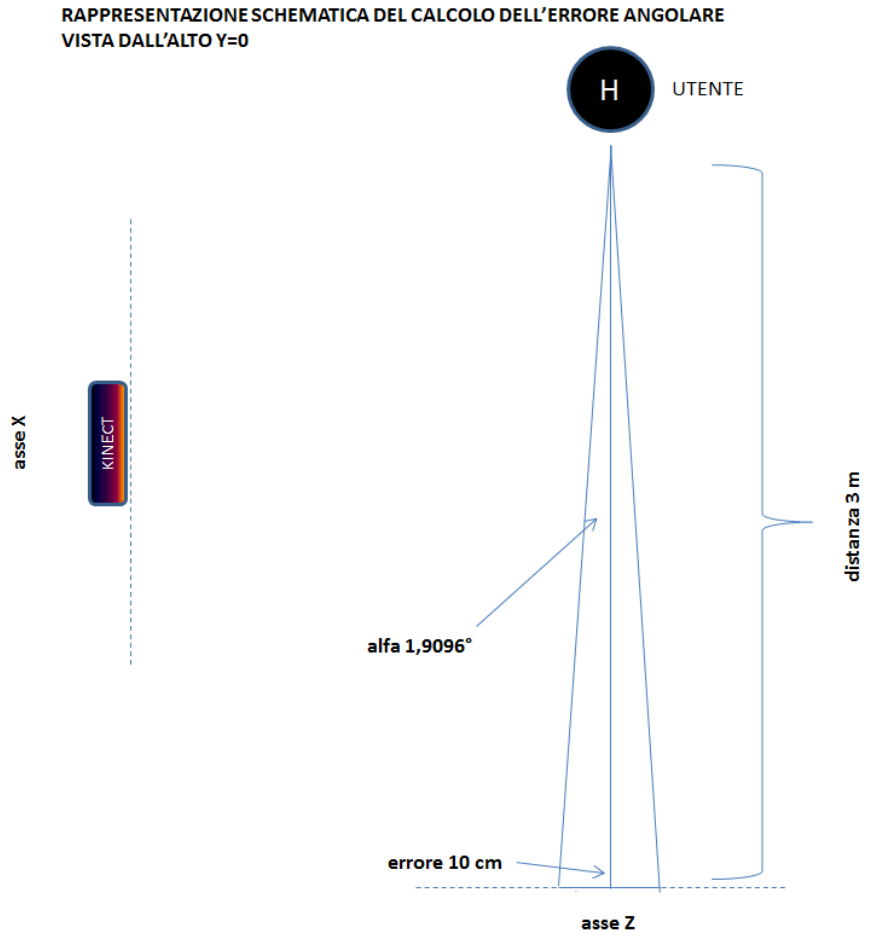


Figura 55 errore angolare - vista dall'alto - asse Z - triangolo isoscele

distanza (cm)	errore (cm)	3° lato tr. (cm)	sen-alfa	alfa °/2	alfa °
500	5	500,024999	0,0099995	0,572938698	1,145877395
475	5	475,026315	0,010525733	0,603091194	1,206182389
450	5	450,027777	0,011110425	0,636593576	1,273187152
425	5	425,029411	0,011763892	0,674036898	1,348073796
400	5	400,031249	0,012499024	0,716159945	1,432319891
375	5	375,033332	0,013332148	0,763898461	1,527796922
350	5	350,035712	0,014284257	0,818455462	1,636910923
325	5	325,038459	0,015382795	0,881403997	1,762807993
300	5	300,041664	0,016664352	0,954841254	1,909682508
275	5	275,045451	0,018178814	1,041626676	2,083253352
250	5	250,049995	0,019996001	1,145762838	2,291525676
225	5	225,055549	0,022216737	1,27303002	2,54606004
200	5	200,062490	0,024992191	1,432096184	2,864192368
175	5	175,071414	0,028559774	1,636577042	3,273154083
150	5	150,083310	0,03331483	1,909152433	3,818304866
125	5	125,099960	0,039968038	2,290610043	4,581220085
100	5	100,124922	0,049937617	2,862405226	5,724810452
75	5	75,166482	0,066519011	3,814074834	7,628149669
50	5	50,249378	0,099503719	5,710593137	11,42118627

Le misurazioni effettuate evidenziano come, al crescere della distanza tra sensore e utente, l'errore angolare commesso diminuisce, ed in particolare, se si considera una distanza media di utilizzo dell'applicazione che si concentra tra i 2.5m e i 3.5m, l'errore commesso lungo l'asse X ed Y è compreso tra 0.65° e 0.91°, quindi meno di 1°.
Per quanto riguarda l'asse Z, invece, l'errore nella stima della profondità di circa 10cm condiziona maggiormente la stima angolare, che è caratterizzata da un errore compreso tra 1.63° e 2.29°.

4.2 DISEGNO SPERIMENTALE Between subjects

4.2.1 Partecipanti

I partecipanti alla sperimentazione sono venti utenti non vedenti, tutti iscritti e frequentanti gli uffici dell'Unione Italiana dei Ciechi e degli Ipovedenti di Rovigo.
Per non introdurre ulteriori variabili di aleatorietà al sistema, i soggetti in questione sono stati accuratamente selezionati in modo da formare quattro gruppi di 5 persone sostanzialmente omogenei ed equivalenti sia da un punto di vista anagrafico e socio-culturale, sia percettivo e visuo-motorio.
In particolare due gruppi effettueranno il test1 e gli altri due gruppi il test2. Si registrano:
11 maschi e 9 femmine, di età compresa tra 18 e 71 anni (media = 41,65, dev standard = 18,38).

TEST 1

Gruppo sperimentale	Sesso	Età	Tipo di disabilità
Utente 1	maschio	47	Non vedente: perdita progressiva della vista a causa di forte trauma
Utente 2	femmina	27	Non vedente: perdita completa della vista a causa del diabete
Utente 3	femmina	59	Non vedente: perdita completa della vista a causa della degenerazione maculare legata all'età
Utente 4	maschio	52	Non vedente: perdita quasi completa della vista a causa del glaucoma
Utente 5	femmina	23	Non vedente: Cecità congenita

Gruppo di controllo	Sesso	Età	Tipo di disabilità
Utente 6	femmina	19	Non vedente: retinite pigmentosa disabilità cognitiva associata
Utente 7	femmina	47	Non vedente: Anoftalmia degenerativa congenita
Utente 8	maschio	65	Non vedente: degenerazione maculare legata all'età
Utente 9	maschio	71	Non vedente: Cecità congenita
Utente 10	maschio	24	Ipovedente: retinopatia grave, visione completamente sfuocata

TEST 2

Gruppo sperimentale	Sesso	Età	Tipo di disabilità
Utente 1	femmina	37	Non vedente: perdita progressiva della vista a causa del glaucoma
Utente 2	femmina	44	Non vedente: Cecità congenita
Utente 3	maschio	69	Non vedente: degenerazione maculare legata all'età
Utente 4	maschio	54	Non vedente: perdita quasi completa della vista dovuta al diabete mellito
Utente 5	maschio	20	Non vedente: cecità sopraggiunta a causa di un forte trauma

Gruppo di controllo	Sesso	Età	Tipo di disabilità
Utente 6	femmina	21	Ipovedente: percezione di luci ed ombre, disabilità cognitiva associata
Utente 7	maschio	28	Non vedente: cataratta degenerativa congenita
Utente 8	femmina	66	Non vedente: Cecità congenita
Utente 9	maschio	42	Non vedente: Cecità congenita
Utente 10	maschio	18	Ipovedente: retinopatia grave, visione completamente sfuocata

4.2.2 Equipaggiamento

- ✓ ***computer***

  Il computer utilizzato per la prova ha le seguenti caratteristiche:

CPU	Intel Core i7 2630QM: 2.3Ghz, 4core/8thread, 8mb cache
Scheda grafica	Nvidia Geforce GT540, 96 stream processors, 128bit bus, 2Gb VRAM
Hard disk	SSD Samsung 830 256gb sata III
RAM	8Gb DDR3 Corsair 1600Mhz cl8
OS	Windows 7 Home premium 64-bit

- ✓ ***audio***

  I suoni sono stati riprodotti tramite un paio di cuffie wireless Sennheizer

- ✓ ***Software***

  Il software utilizzato prevede:

  - Microsoft Kinect SDK 1.5;

- Pure Data 0.42 extended;
- SoundingARM versione auditory icons;
- SoundingARM versione speech icons.

4.2.3 Oggetti scelti

Gli oggetti che sono stati scelti per la sperimentazione sono tutti elementi dell'ambiente caratteristici di qualsiasi cucina, in particolare si è preferito considerarne un insieme variegato sia per dimensione sia per forma, in modo da influenzare il meno possibile l'esito finale.
Ogni oggetto è stato poi classificato secondo una categoria di alto livello, secondo dimensione e in base alla capacità di produrre o no direttamente un suono.

oggetto	Categoria	Tipo suono	Taglia
Macchinetta del caffè	Elettrodomestico	Auditory & speech	Medio
Carta Scottex	Utensile	Auditory & speech	Piccolo
Fornello	Elettrodomestico	Auditory & speech	Medio
Forno	Elettrodomestico	Auditory & speech	Grande
Piatti e stoviglie	utensili	Auditory & speech	Medio
Pattumiera	mobilio	Auditory & speech	Grande
Sedia	Mobilio	Auditory & speech	Grande
Posate	utensili	Auditory & speech	Piccolo

4.2.4 Realizzazione tracce audio

Tutte le tipologie di icons uditive di seguito descritte sono memorizzate sotto forma di file audio in formato .wav, per questa ragione possono essere facilmente scambiate e modificate in quanto indipendenti dal funzionamento dell'applicazione SoundingARM.

Realizzazione delle Auditory icon

Per alcuni oggetti non ci sono stati problemi ad identificare un suono da utilizzare come auditory icon in quanto l'oggetto di per sé stesso emette quel suono caratteristico (es. forno, macchinetta del caffè, fornello); per altri, invece, è stato opportuno veicolare un suono indiretto che comunque fosse correlato all'oggetto, per esempio il suono caratteristico emesso dal materiale di cui quell'oggetto è costituito (es. tavolo, sedie, piatti e stoviglie). In linea con quanto sostenuto da Keller e Stevens ([40]), spesso è opportuno mixare assieme più suoni differenti, ma ugualmente caratteristici dell'oggetto al fine di ottenere un'icona più diretta ed espressiva (es. differenti stoviglie a contatto).

Realizzazione delle Earcon

Le earcons sono pattern musicali che possono essere decomposti in cinque dimensioni: ritmo, pitch, timbro, registro e dinamiche (Brewster [41]). Complice la loro predisposizione a costruire gerarchie, il design delle earcons include la categorizzazione degli oggetti, perciò ogni earcon inizia con un suono di apertura che rappresenta la categoria del suono da cui proviene. Si è

scelto di utilizzare strumenti musicali differenti per ogni categoria e in particolare:

- tavoli → chitarra,
- porte → ding e dong,
- ostacoli → pianoforte,
- forno-fornello → drum e strumenti a percussione,
- oggetti correlati all'acqua → flauto,
- contenitori → organo.

Dopo il suono di categorizzazione, inizia il vero e proprio suono correlato all'oggetto. Ogni oggetto è stato rappresentato da una melodia e da un ritmo univoci. Dal momento che le melodie e gli strumenti scelti sono arbitrari, si è scelto il suono che meglio approssima la categoria di appartenenza.

Realizzazione delle Earcon-icon hybrid

A causa del fatto che le earcons spesso sono arbitrarie e poco associative, il loro apprendimento risulta essere critico. Le earcons però hanno il vantaggio che ogni oggetto è distinto, così per trarre beneficio dai pro e annullare i contro, si è scelto di usare delle earcon-icon ibride ([54]), che combinano il suono di apertura rappresentante la categoria di ogni oggetto con l'auditory icon dello specifico oggetto. In questo modo, ogni suono è specifico per ogni oggetto.

Ciascuna earcon-icon hybrid è composta da un suono di apertura che ne determina la categoria, e da un secondo suono, caratteristico dell'oggetto stesso, che lo rende unico, effettua cioè una sorta di identificazione dell'oggetto sulla base dell'auditory icon che lo caratterizza.

Realizzazione delle Size-hybrid

Col fine di veicolare anche l'impressione della dimensione di un oggetto, si è scelto di inserire un layer di suoni che contiene appunto informazioni circa la dimensione "size-hybrid" ([54]). Di fatto, in questo modo si realizza una classificazione basata sulla dimensione con quattro taglie: molto piccolo, piccolo, medio, grande. Per ogni categoria di taglia è stata composta una melodia univoca che differisce nel pitch e nella durata. Il suono che rappresenta gli oggetti grandi è caratterizzato da un pitch basso e una durata lunga, mentre un oggetto piccolo è corto e con pitch alti.

La dimensione dell'oggetto è stata aggiunta in coda al suono earcon-icon hybrid. Una variante può essere quella di eseguire il suono correlato alla dimensione, in parallelo all'icona sonora rappresentante l'oggetto, in modo da avere una sovrapposizione non distruttiva, che permette all'icona nel suo complesso di essere eseguita in un tempo significativamente minore, e l'intera applicazione risulta così più veloce e reattiva al puntamento.

Realizzazione delle Speech icons

Per quanto riguarda le speech icons, si è scelto di registrare in formato .wav la voce di una ragazza che pronuncia i nomi degli oggetti, invece di adottare la più consueta voce sintetizzata da un software text to speech, in quanto i soggetti non vedenti sovente si annoiano in presenza di voci artificiali, e preferiscono di gran lunga una voce umana cadenzata e più realistica.

Realizzazione delle Spearcons
Per realizzare le spearcons si utilizza un algoritmo a compressione logaritmica in MatLab come descritto da Palladino e Walker.

4.2.5 Procedura sperimentale

1. Raccolta dati partecipanti
2. Costituzione gruppo sperimentale e gruppo di controllo
3. Istruzioni
4. Esecuzione test
5. Sintesi dei risultati

4.2.6 PRIMO TEST: MAPPA SONORA

Creazione della mappa della stanza con SoundingARM.

Descrizione
Il test consiste nell'esplorazione di una cucina reale, situata all'interno della sede dell'Unione Italiana dei Ciechi e degli Ipovedenti di Rovigo.
Prendono parte al test un gruppo di 10 utenti opportunamente selezionati secondo le modalità descritte al paragrafo 4.2.1.
Cinque utenti costituiscono il gruppo sperimentale, ed utilizzano SoundingARM nella versione speech icons per esplorare e costruire una mappa spaziale della stanza, la più accurata possibile.
Gli altri cinque utenti, invece, costituiscono il gruppo di controllo, ed effettuano la medesima esplorazione senza l'utilizzo di SoundingARM, perciò sono liberi di deambulare per la stanza, e devono ricreare la mappa servendosi unicamente del bastone e delle loro capacità locomotorie e tattili, frutto dell'esperienza individuale.

TASK1: Ciascun utente, indipendentemente dal fatto che appartenga al gruppo sperimentale o di controllo, ha a disposizione tutto il tempo che vuole per effettuare l'esplorazione della cucina, dopo di che deve uscire e lasciar posto all'utente successivo.
Indipendentemente dalla modalità di esplorazione, a ciascun utente viene richiesto di cercare di individuare e ricordare la posizione del maggior numero possibile di oggetti caratteristici della cucina.
Dal momento che a tutti e dieci gli utenti non viene comunicato in precedenza quali sono gli oggetti da trovare e localizzare, è significativo, a test compiuto, verificare quali e quanti oggetti ciascun utente è stato in grado di identificare e localizzare nella stanza.

TASK2: SoundingARM viene spento per tutti e due i gruppi. Agli utenti viene chiesto di raggiungere (uno alla volta) 3 oggetti presenti nella stanza: carta-scottex, macchinetta del caffè e sedia. La richiesta di raggiungimento del primo oggetto viene effettuata quando l'utente si trova sulla soglia; raggiunto il primo oggetto all'utente viene chiesto di raggiungere il secondo oggetto e così anche per il terzo.

Per mezzo di questo secondo task si misura quanto tempo l'utente impiega a raggiungere ognuno degli oggetti: se la costruzione della mappa spaziale è avvenuta in modo efficace l'utente impiegherà poco tempo.

Gli obiettivi del test sono dunque due:
verificare quale dei due gruppi, in media, è in grado di riconoscere e posizionare nel modo corretto il maggior numero di oggetti;
verificare quale delle due metodologie di esplorazione risulta essere più efficace nella creazione di una mappa spaziale accurata.

Dati raccolti
TASK1: per ciascun utente si registrano:

- quali oggetti ha individuato,
- il numero di oggetti individuati in totale,
- il tempo complessivo impiegato per completare l'esplorazione.

esempio tabella raccolta dati:

Utente	Tipo esplorazione	Oggetti individuati	Numero oggetti trovati	Tempo impiegato
1	SoundingARM speech icons	Forno, sedia, scottex	3/8	1':22''
6	Locomotoria-tattile	Forno, mac.caffè, sedia, fornello	4/8	2':10''

I soggetti appartenenti al gruppo di controllo che non utilizzano SoundingARM devono verbalizzare esplicitamente l'identificazione dell'oggetto, in questo modo coloro che raccolgono i dati sperimentali hanno la conferma dell'avvenuta individuazione.

TASK2: Per ciascun utente si registra:

- il tempo per raggiungere il primo oggetto,
- il tempo per raggiungere il secondo oggetto,
- il tempo per raggiungere il terzo oggetto,
- il tempo complessivo somma dei tre precedentemente indicati.

esempio tabella raccolta dati:

Utente gruppo	Tempo 1° oggetto	Tempo 2° oggetto	Tempo 3° oggetto	Tempo totale
1 s	0':30''	0':20''	0':10''	1':00''
6 c	0':50''	0':10''	0':35''	1':35''

Variabili
Nel Task1 la prima variabile di interesse è il numero di oggetti individuati da ciascun utente, indipendentemente dal fatto che utilizzi SoundingARM o l'esplorazione diretta dell'ambiente. La seconda variabile d'interesse è il tempo impiegato per completare l'esploarazione della stanza.
Nel Task2, invece, la variabile di interesse è il tempo per raggiungere i 3 oggetti indicati.

Analisi dei risultati

Tabella Task1

TEST1	TASK1											
utente	gruppo	esplorazione	forno	posate	fornello	pattumiera	mac caffè	carta-scottex	piatti/stoviglie	sedia	totale	tempo totale (s)
1	sperimentale	speech icon	1	1	1	1	1	1	1	1	8	144
2	sperimentale	speech icon	1		1	1	1	1	1	1	7	140
3	sperimentale	speech icon	1	1	1	1	1	1	1	1	8	53
4	sperimentale	speech icon	1			1	1	1	1	1	6	110
5	sperimentale	speech icon	1	1	1	1	1	1	1	1	8	90
6	controllo	tattile	1	1	1		1	1	1	1	7	58
7	controllo	tattile	1	1	1	1	1	1	1	1	8	121
8	controllo	tattile	1	1	1	1	1	1	1	1	8	196
9	controllo	tattile	1	1	1	1	1	1	1	1	8	192
10	controllo	tattile	1	1	1	1	1	1	1	1	8	142

numero medio oggetti trovati
7,4
tempo medio (s) + dev st esplorazione
107,4
37,65

numero medio oggetti trovati
7,8
tempo medio (s) + dev st esplorazione
141,8
56,81

Come si può vedere dalla tabella precedente, entrambi i gruppi hanno ottenuto buoni risultati nell'esplorazione della cucina: il gruppo sperimentale raggiunge una media di 7.4 oggetti su 8 totali, e questo è un ottimo risultato in considerazione del fatto che oggetti come la carta scottex e la macchinetta del caffè hanno dimensioni piccole se confrontate agli altri oggetti e al volume complessivo della stanza.

Il gruppo di controllo che ha eseguito l'esplorazione tattile raggiunge una media complessiva di ben 7,8 oggetti individuati su 8 totali, un valore ancora maggiore di quello fatto registrare dal gruppo sperimentale, però è opportuno precisare che, per mezzo dell'esplorazione tattile, gli utenti sono liberi di muoversi e manipolare direttamente gli oggetti, e in questo modo, se si conduce un'esplorazione metodica dell'ambiente, è più semplice indivduare tutti gli oggetti.

Uno dei vantaggi evidenti che comporta l'utilizzo di SoundingARM è una sostanziale diminuzione del tempo medio necessario per completare l'esplorazione dell'ambiente, si registra infatti un tempo medio di 107,4 secondi per gli utenti del gruppo sperimentale, circa un 30% inferiore rispetto al tempo medio impiegato dagli utenti del gruppo di controllo.

Tabella Task2

TEST1	TASK2					
utente	gruppo	esplorazione	tempo scottex	tempo caffè	tempo sedia	tempo totale s
1	sperimentale	speech icon	83	14	20	117
2	sperimentale	speech icon	81	10	9	100
3	sperimentale	speech icon	N/D	18	16	N/D
4	sperimentale	speech icon	5	30	33	68
5	sperimentale	speech icon	5	81	27	113
6	controllo	tattile	24	32	9	65
7	controllo	tattile	16	7	65	88
8	controllo	tattile	N/D	15	29	N/D
9	controllo	tattile	13	17	6	36
10	controllo	tattile	4	18	6	28

Tempi medi:

scottex	caffè	sedia	totale
43,5	30,6	21	99,5
14,25	17,8	23	54,25

Analizzando i tempi impiegati dagli utenti dei due gruppi per raggiungere gli oggetti indicati, innanzitutto si può notare come, a parte due casi (1 per gruppo), tutti gli utenti sono stati in grado di raggiungere la posizione dell'oggetto che veniva loro chiesto, e questo testimonia che, indipendentemente dalla modalità di esplorazione utilizzata, gli utenti hanno avuto modo di crearsi una mappa abbastanza precisa degli oggetti e della loro dislocazione nella stanza.
Dalla tabella precedente si può vedere come gli utenti del gruppo di controllo siano stati, in media, molto più veloci: circa 3 volte a raggiungere il primo oggetto cioè la carta scottex, e circa 2 volte a raggiungere il secondo oggetto ovvero la macchinetta del caffè. Al contrario, per raggiungere il terzo oggetto (la sedia), hanno impiegato meno tempo gli utenti del gruppo sperimentale, 21 secondi contro i 23 del gruppo di controllo.
Questi dati evidenziano principalmente un aspetto: gli utenti che hanno potuto manipolare e "tastare con mano" l'ambiente e gli oggetti, hanno avuto modo di crearsi un'idea precisa non solo della dislocazione, ma anche delle distanze a cui si trovano gli oggetti, al contrario chi ha utilizzato SoundingARM è consapevole della direzione approssimativa in cui si trovano i singoli oggetti, però è più impacciato e timoroso nel momento in cui deve raggiungerli poichè non ha avuto la possibilità di interagire direttamente con l'ambiente.
La sedia è un oggetto in controtendenza rispetto agli altri fondamentalmente perché è l'unico oggetto a essere dislocato lungo la parete di sinistra della stanza, per cui è facile da raggiungere da qualsiasi punto.

Considerazioni

Considerando nel complesso il primo ed il secondo task, si può affermare che da un lato SoundingARM è un ottimo sostituto dell'esplorazione tattile perché permette all'udente di costruirsi una mappa sufficientemente precisa e dettagliata della stanza, in poco tempo e senza la necessità di dover deambulare in spazi sconosciuti.

D'altro canto però risulta evidente la superiorità dell'esplorazione tattile nel momento in cui si devono fisicamente raggiungere le posizioni in cui sono dislocati gli oggetti, e questo perché la possibilità di manipolazione e interazione con l'ambiente veicola informazioni multisensoriali che SoundingARM non trasmette.
In ottica di miglioramento del sistema, queste considerazioni suggeriscono come sia opportuno integrare l'auditory display di SoundingARM con informazioni multisensoriali in grado di riproporre quanto l'utente percepisce nel momento in cui effettua l'esplorazione locomotoria e tattile.

4.2.7 SECONDO TEST: AUDITORY-ICONS vs SPEECH-ICONS

Riconoscimento direzione dell'oggetto e tempo impiegato con le varie modalità

Descrizione

Questo test si suddivide in due fasi:

Fase1: ciascun utente ha 1 minuto di tempo per effettuare un'esplorazione della stanza utilizzando SoundingARM, in particolare:
- il gruppo sperimentale utilizza la versione con auditory icons,
- il gruppo di controllo utilizza la versione con speech icons.

Fase2: dopo aver effettuato l'esplorazione con SoundingARM acceso per entrambi i gruppi, a ciascun utente è richiesto di:
- indicare la direzione in cui è posizionato il cassetto delle posate e raggiungerlo;
- indicare la direzione in cui è posizionata la pattumiera e raggiungerla;
- indicare la direzione in cui sono posizionati i piatti e le stoviglie e raggiungerli.

Questo test ha come obiettivo quello di confrontare "sul campo" quale sia il miglior auditory display per apprendere la disposizione degli oggetti di una stanza.
In particolare si confrontano speech icons ed auditory icons nell'ambiente "cucina", in cui tutti gli oggetti sono caratterizzati da suoni specifici ed univoci che nella maggioranza dei casi veicolano più informazioni rispetto a quelle fornite da una voce che li descrive.

Dati raccolti

Per ciascun utente si raccolgono i seguenti dati:

1° oggetto: direzione individuata (si / no)
Tempo impiegato per raggiungerla;
2° oggetto: direzione individuata (si / no)
Tempo impiegato per raggiungerla;
3° oggetto: direzione individuata (si / no)
Tempo impiegato per raggiungerla;

Esempio tabella raccolta dati:

Utente	Tipo esplorazione	1*oggetto: direz/tempo	2*oggetto: direz/tempo	3*oggetto: direz/tempo
Utente 1	Speech icons	Si – 0':21''	No – N/D	Si – 0':10''
Utente 6	Auditory icons	no – N/D	Si – 0':25''	Si – 0':33''

Variabili

La variabile più significativa in questo test è il tempo che intercorre tra l'individuazione da parte dell'utente della direzione dell'oggetto richiesto e il momento in cui egli raggiunge tale oggetto.
Tra auditory icons e speech icons, risulterà preferibile la modalità che in media permette al soggetto di costruirsi la mappa migliore consentendogli di raggiungere l'oggetto richiesto nel minor tempo possibile.

Analisi dei risultati

TEST2	1 min fase1							
utente	gruppo	esplorazione	d.posate	t.posate	d.pattum	t.pattum	d.piatti	t.piatti
1	sperimentale	auditory icon	si	8	no	N/D	si	10
2	sperimentale	auditory icon	no	12	si	18	si	4
3	sperimentale	auditory icon	no	N/D	si	14	si	9
4	sperimentale	auditory icon	si	9	si	7	si	3
5	sperimentale	auditory icon	si	10	si	9	si	2
6	controllo	speech icon	si	1	si	2	si	1
7	controllo	speech icon	si	14	si	7	si	1
8	controllo	speech icon	si	8	si	11	si	3
9	controllo	speech icon	si	12	si	23	no	4
10	controllo	speech icon	si	7	si	9	si	2

direzioni medie

posate	pattumiera	piatti	totale
3 su 5	4 su 5	5 su 5	13 su 15
5 su 5	5 su 5	4 su 5	14 su 15

tempi medi (s)

posate	pattumiera	piatti	tempo medio oggetto
9,75	12	5,6	9,11
8,4	10,4	2,2	7

Considerando prima di tutto le direzioni degli oggetti, risulta evidente che sia SoundingARM con auditory icons che SoundingARM con speech icons sono sufficientemente precisi, e solamente in pochi casi la direzione è stata sbagliata. Quello che è curioso osservare è che in certe situazioni gli utenti non sono stati in grado di indicare la direzione dell'oggetto, però poi sono riusciti a raggiungerlo comunque.
Soprattutto nelle situazioni in cui ci sono parecchi oggetti adiacenti o abbastanza ravvicinati, gli utenti non vedenti hanno l'idea di dove si trova l'oggetto richiesto, e lo sanno raggiungere anche in tempi rapidi, però non sanno indicare con precisione l'oggetto stesso. Da quanto si è potuto osservare

durante il test, questo fenomeno è dovuto principalmente al fatto che gli utenti costruiscono la mappa sulla base della loro posizione iniziale, e tutti i feedback che ricevono dall'esplorazione sono mappati consistentemente a tale posizione. Nonostante all'inizio del test tutti gli utenti fossero stati posizionati con le spalle perfettamente allineate e parallele al sensore kinect, a causa dei movimenti di puntamento effettuati durante l'esplorazione, molti di loro inconsapevolmente ruotavano di poco le spalle e si spostavano dalla posizione di partenza; in questo modo erano convinti che un oggetto fosse in una certa direzione mentre in realtà si trovava leggermente più a destra o a sinistra perché, da una distanza di 2-3m, anche pochi gradi di rotazione comportano una sostanziale modifica nella direzione del puntamento.
Passando all'analisi dei tempi medi impiegati per raggiungere gli oggetti, gli utenti che hanno utilizzato le speech icons si confermano più veloci di coloro che hanno utilizzato le auditory icons, anche se i risultati non si discostano di molto, mediamente si registra un vantaggio di tempo medio del 20% a favore delle speech icons, eccezion fatta per il cassetto delle posate (speech icons 2,2 secondi, auditory icons 5,6 secondi), ma ciò non stupisce poi più di tanto perché con le speech icons gli utenti ricevevano un feedback molto preciso: "cassetto delle posate", mentre con le auditory icons percepivano un suono di forchette, cucchiai, coltelli e mestoli a contatto, chiaramente più difficile da associare ed interpretare.

Considerazioni
Il secondo test nel complesso fornisce indicazioni incoraggianti in merito alle performance di SoundingARM, il sistema risulta infatti essere preciso e reattivo in entrambi gli scenari, nonostante la versione auditory icons fornisca un miglior soundscape e la speech icons sia più chiara e immediata.
Tutti gli utenti che hanno avuto modo di provare il sistema si sono dimostrati molto interessati e hanno riportato impressioni positive circa il feedback acustico fornito: molti hanno affermato di preferire la versione con auditory icons perché decisamente più espressiva ed "evocativa" complici i suoni di ogni giorno, mentre altri hanno espresso il desiderio di poter effettuare prove ulteirori in modo da prendere confidenza col sistema.
Un altro aspetto che si è potuto osservare durante tutta la sperimentazione è che alcuni utenti sono naturalmente in grado di utilizzare a dovere il sistema in quanto puntano e sondano l'ambiente nel modo corretto, stendendo il braccio ben diritto davanti a loro e ruotandolo in modo da "toccare" idealmente gli oggetti per mezzo di un prolungamento dei propri arti; altri utenti, invece, si sono trovati in difficoltà poiché non riuscivano ad effettuare un buon puntamento, ad esempio muovevano il braccio in modo incerto ed impacciato, oppure indicavano in modo approssimativo, senza effettuare un vero panninga 360°, ed in questo modo il sistema di tracciamento aveva difficoltà nel calcolare la projectionSphere e interpretare quale oggetto fosse puntato perché in realtà l'utente non stava puntando affatto.
In linea con quanto appena osservato, si ritiene che per taluni utenti sia necessaria una fase di training che permetta loro di prendere confidenza col sistema e li renda capaci di utilizzare SoundingARM nel modo corretto.

Infine, un aspetto curioso in un certo qual modo di questa sperimentazione è che alcuni utenti sono stati molto abili ad esplorare la stanza, e hanno individuato velocemente tutti gli oggetti, però poi nei task successivi dove veniva chiesto loro di indicare e raggiungere gli oggetti, si sono trovati in difficoltà perché non si ricordavano più dove erano dislocati, e conseguentemente impiegavano molto più tempo per raggiungerli. L'abilità di esplorare un ambiente sconosciuto, dunque, non può prescindere anche da una buona memoria e da un discreto metodo analitico durante la fase di individuazione e posizionamento degli oggetti nella mappa virtuale che ciascun utente si costruisce dentro di sé.

Capitolo 5

CONCLUSIONI

Grazie alle opinioni e ai suggerimenti raccolti dalla sperimentazione, si può affermare che SoundingARM è un progetto software particolarmente apprezzato in quanto tutti gli utenti ne comprendono le potenzialità e lo considerano un importante sussidio, il cui utilizzo non deve precludere o sostituirsi del tutto all'esplorazione tattile, bensì può rendere più agevoli le fasi di ambientazione e riconoscimento della disposizione degli oggetti e degli ostacoli in un ambiente completamente sconosciuto.
Molto apprezzate sono state anche le auditory icons, una novità sostanziale per molti utenti, da anni abituati esclusivamente ai feedback sonori tramite speech icons. Nonostante il feedback veicolato dalla voce sia più immediato, il soundscape che si ottiene per mezzo dei "suoni di ogni giorno" è qualcosa di nuovo e al contempo immersivo ed evocativo di realtà e contesti ben noti per gli utenti, perciò si può pensare ad un'integrazione tra le varie tipologie di icone sonore in modo da associare a ciascun oggetto l'icona che meglio lo descrive, lo identifica e lo caratterizza nel minor tempo possibile.
I risultati sperimentali evidenziano come in alcuni contesti l'esplorazione tattile sia ancora preferibile, ed in un certo qual modo superiore se confrontata all'esplorazione tramite SoundingARM, però bisogna considerare due fattori: innanzitutto gli utenti che si sono maggiormente distinti nell'esplorazione locomotoria tattile hanno acquisito un'esperienza negli anni che ha permesso loro di sviluppare un vero e proprio "sesto senso", l'esperienza acquisita è una sorta di ausilio essa stessa; per cui, in particolar modo per gli utenti che hanno mostrato sostanziali difficoltà nel puntamento e nell'esecuzione delle gesture, si ritiene opportuno effettuare fasi di training ripetute nel tempo, così da insegnare loro il modo corretto in cui sondare virtualmente la stanza.
In secondo luogo coloro che effettuano l'esplorazione locomotoria tattile riescono ad ottenere molte più informazioni poiché integrano i feedback sensoriali provenienti anche dal tatto, dall'olfatto, dai proprio-percettori, e in un certo qual modo effettuano una vera esplorazione immersiva ed interattiva perché hanno la possibilità di manipolare gli oggetti e l'intera stanza. SoundingARM, al contrario, dal punto di vista acustico fornisce un auditory displey sensibilmente migliore perché ricco di icone sonore completamente assenti nell'esplorazione locomotoria-tattile, però non è in grado di veicolare tutte le altre informazioni addizionali sopra descritte. Sarebbe perciò auspicabile che nelle future release si potessero integrare componenti software in grado di simulare anche altre percezioni sensoriali, in modo da veicolare un vero e proprio global array che convoglia informazioni riguardanti la stanza e gli oggetti provenienti da differenti punti di vista.
Un primo passo verso questa direzione è la spazializzazione binaurale delle auditory icons, una tecnica di rielaborazione audio che, se ben ottimizzata,

permette di ricostruire un'autentica scena di audio 3D, con tutte le sorgenti sonore posizionate nell'ambiente circostante consistentemente alla posizione e all'orientazione della testa dell'ascoltatore (alto DI, bassa CSD paragrafo 3.3.1). Tramite la spazializzazione binaurale ed appositi filtri di riverbero, si possono veicolare anche molte altre informazioni riguardanti le dimensioni degli oggetti, la forma e la volumetria della stanza, la distanza degli oggetti dall'utente, realizzando così un auditory display sensibilmente più ricco ed immersivo.
Un altro aspetto su cui si può intervenire per migliorare l'usabilità del sistema è il processo di creazione del file di configurazione. Allo stato attuale, infatti, è necessario compilare manualmente il file .CONF e apportare di volta in volta eventuali modifiche nel caso in cui mutino il numero o le dislocazioni degli oggetti. Sarebbe auspicabile realizzare un meccanismo più diretto e intuitivo che permetta all'utente finale di configurare la stanza e i relativi oggetti con pochi click del mouse, in modo preciso ed accurato, senza la necessità di effettuare misure sperimentali e settaggi di parametri. A tal proposito si potrebbe pensare da un lato di utilizzare direttamente il sensore kinect per scattare delle range image per mezzo delle quali effettuare l'individuazione degli oggetti che compongono la stanza, con le relative misure e distanze relative, dall'altro di creare un apposito tool applicativo grafico che metta a disposizione una sorta di CAD 3D semplificato drag and drop, tramite il quale l'utente finale si limita ad inserire la kinect e gli oggetti nella stanza virtuale, ed il file .CONF viene generato automaticamente dal tool stesso.

APPENDICE A

MICROSOFT KINECT – quadro generale

A.1 Requisiti hardware

Figura 56 cavo USB per alimentazione e dati

Il dispositivo Kinect può essere connesso ad un normale PC tramite un apposito cavo usb, il quale, oltre che fornire la connettività USB, garantisce anche l'alimentazione al device.

Per poter sviluppare applicazioni che sfruttano le potenzialità del Kinect è necessario scaricare ed installare il Microsoft Kinect SDK. Per le capacità di voice recognition del dispositivo, è necessario isntallare anche lo Speech SDK.

I requisiti hardware da soddisfare sono i seguenti:

- Processore dual-core con cpu a 2.66GHz o superiore;
- Almeno 2GB di RAM;
- Scheda grafica compatibile con Windows 7 che supporti DirectX 9.0

A.2 Requisiti Software

Dal punto di vista software:

- Microsoft Visual Studio 2010 Express (o una qualsiasi altra versione);
- .NET Framework 4.0.
- Per la parte di grafica:
  - per lo sviluppo: Microsoft DirectX SDK;
  - per il runtime: DirectX 9.0 Runtime per l'esecuzione;
- Per le funzionalità Audio:
  - per lo sviluppo: Microsoft Speech Platform SDK versione 10.2 con Kinect for Windows Runtime Language Pack v.0.9;
  - per il runtime: Microsoft Speech Platform Runtime v.10.2.

A.3 Sensori on-board

Figura 57 immagine ravvicinata della Microsoft Kinect

- ✓ **Video Camera**: si tratta di una comune camera in grado di fornire immagini con risoluzione 640x480 a colori;
- ✓ **Sensori di Profondità**: rappresentano uno dei punti di forza del dispositivo e sono in grado di recuperare un'immagine in cui sono riportati i dati di distanza di ciò che appare davanti al Kinect;
- ✓ **Batteria di Microfoni**: si tratta di 4 microfoni in grado di fornire funzionalità di pulitura del suono, posizionamento della sorgente sonora, cancellazione dell'eco e altro ancora.
- ✓ **Inclinazione motorizzata**: il dispositivo può essere brandeggiato (da codice) con un angolo verticale di ±23 gradi rispetto al piano orizzontale.

Il Software Development Kit fornisce una libreria in grado di recuperare e gestire i flussi di dati provenienti dai sensori appena visti e può essere utilizzata sia in C++ che in uno dei linguaggi Managed del Framework .NET.

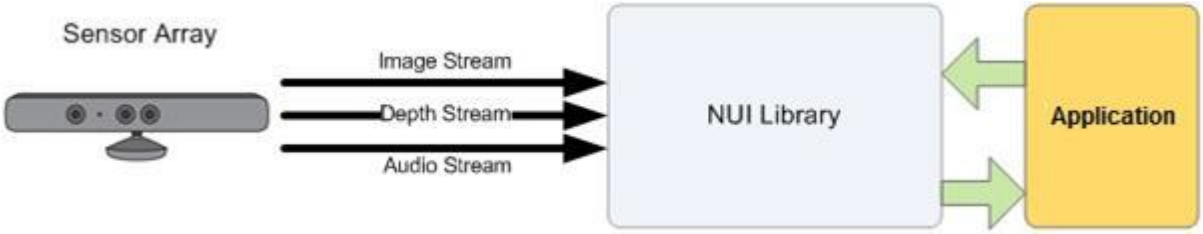


Figura 58 Stream di dati provenienti dal sensore

A.4 Struttura SDK

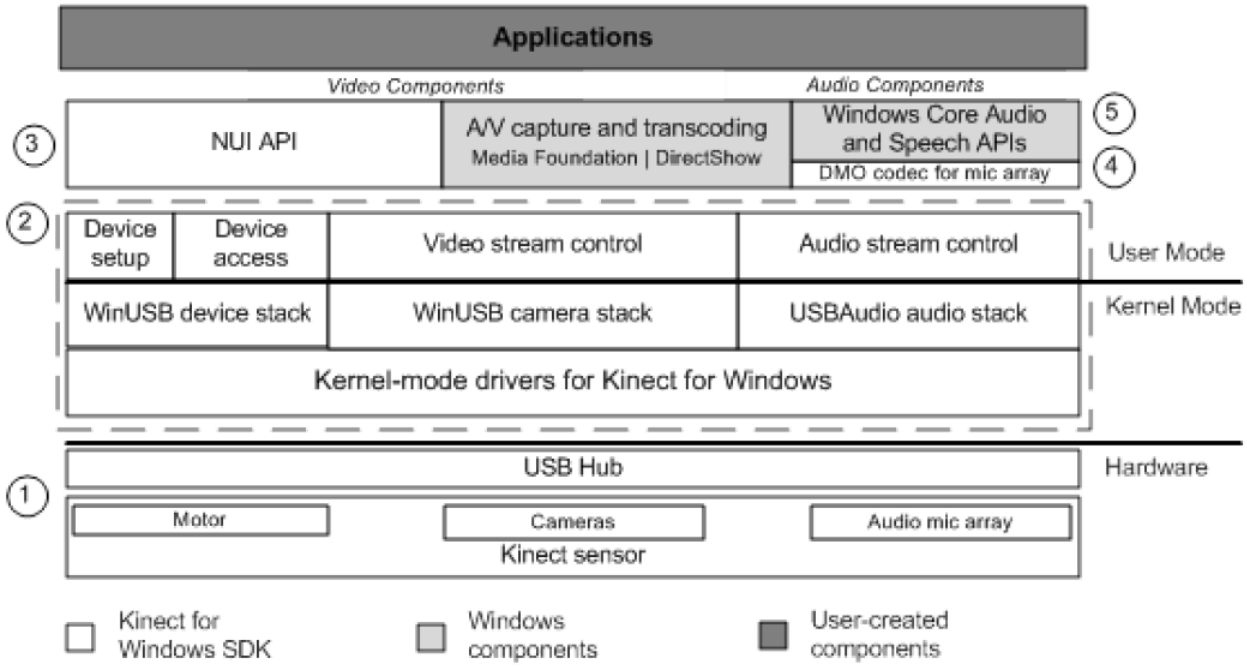


Figura 59 Struttura dell'SDK

1. *Hardware:* l'hardware comprende i sensori visti in precedenza è l'hub USB che permette il loro collegamento al pc.
2. *Microsoft Kinect drivers:* i driver di Windows 7 hanno le seguenti funzionalità:
   - Permettono l'accesso all'array di microfoni con le API Audio standard di Windows.
   - Forniscono gli stream della video camera e dei sensori di profondità.
   - Forniscono la possibilità di utilizzare più device contemporaneamente.
3. *NUI API:* un insieme di API che permettono di recuperare i dati dai sensori di immagine e di controllare il device stesso (ad esempio brandeggiare il dispositivo).
4. *KinectAudio DMO:* estende le funzionalità dell'array di microfoni supportato in Windows 7 per fornire le funzionalità di Beamforming (mappatura sonora dell'area) e localizzazione della sorgente sonora.
5. *Windows 7 standard APIs:* le API audio, speech e media presenti in Windows 7 e Microsoft Speech.

A.5 Funzionalità Video

Gli stream video e di profondità vengono gestiti entrambi grazie alla classe Runtime la quale permette anche di ottenere informazioni riguardo lo Skeletal Tracking. La classe Runtime mette a disposizione tre eventi che permettono di recuperare i frame video, di profondità e di skeletal tracking:

- **DepthFrameReady**: restituisce il frame di profondità disponibile proveniente dai sensori;
- **VideoFrameReady**: restituisce il frame video disponibile proveniente dalla camera;

- **SkeletonFrameReady**: restituisce il frame di Skeletalk Tracking elaborato dalla libreria dell'SDK e contenente le informazioni sugli "scheletri" dei player posti davanti al dispositivo.

La classe ImageFrame contiene l'immagine recuperata e le informazioni riguardanti il Timestamp , il tipo di immagine, le risoluzione, etc., etc. I bytes costituenti l'immagine sono contenuti all'interno della struttura PlanarImage nell'array Bits (array di bytes). La struttura PlanarImage contiene altresì i dati relativi ad altezza, larghezza e numero di bytes per ogni pixel:

- Immagine video di tipo Color: 4 bytes per pixel contenenti le componenti ARGB (Alfa, Red, Green e Blue);
- Immagine di tipo ColorYUV: 4 bytes per pixel contenenti le componenti YUV (ciano, magenta, giallo e nero);
- Immagine di tipo ColorYUV Raw: 2 bytes per pixel contenenti le componenti YUV (ciano, magenta, giallo e nero);
- Dati di profondità senza indice del player: 2 bytes per pixel (solo i primi 12 bit rappresentano la distanza, in millimetri del punto dal device);
- Dati di profondità con indice del player: 2 bytes per pixel (i 3 bit meno significativi contengono l'indice del player, i successivi 12 bit rappresentano la distanza, in millimetri, del punto dal device).

Per quanto riguarda la parte video, il dispositivo riprende le immagini sempre con risoluzione 1280x1024, le comprime in modo che si possa avere un trasferimento dati efficiente su bus USB, le trasferisce al pc dove la libreria le decomprime nella risoluzione scelta. L'algoritmo è comunque a perdita di informazione, seppur minima.

A.6 Lo Scheletro dell'utente

Peculiarità della classe SkeletonFrame è quella di contenere una collezione di oggetti SkeletonData (uno per ogni player identificato dal dispositivo). Gli SkeletonData altro non sono che classi che modellano gli "scheletri" dei player rilevati dal dispositivo.
Si definisce "scheletro" l'insieme di un certo numero di punti (20 per l'esattezza) caratteristici del corpo umano come la testa, le mani, i polsi, i piedi e così via. Ogni punto (Joint) è identificato da un id specifico (enumerazione JointID) ed espone la posizione che questo assume nello spazio (in termini di coordinate x, y e z) e lo stato di tracciamento del punto: tracciato, dedotto dalla posizione degli altri punti (es. nascosto da altri punti del corpo) o non tracciato. Lo stesso "scheletro" dispone dello stato di tracciamento in modo che possiamo selezionare quale player è realmente tracciato e quale no.

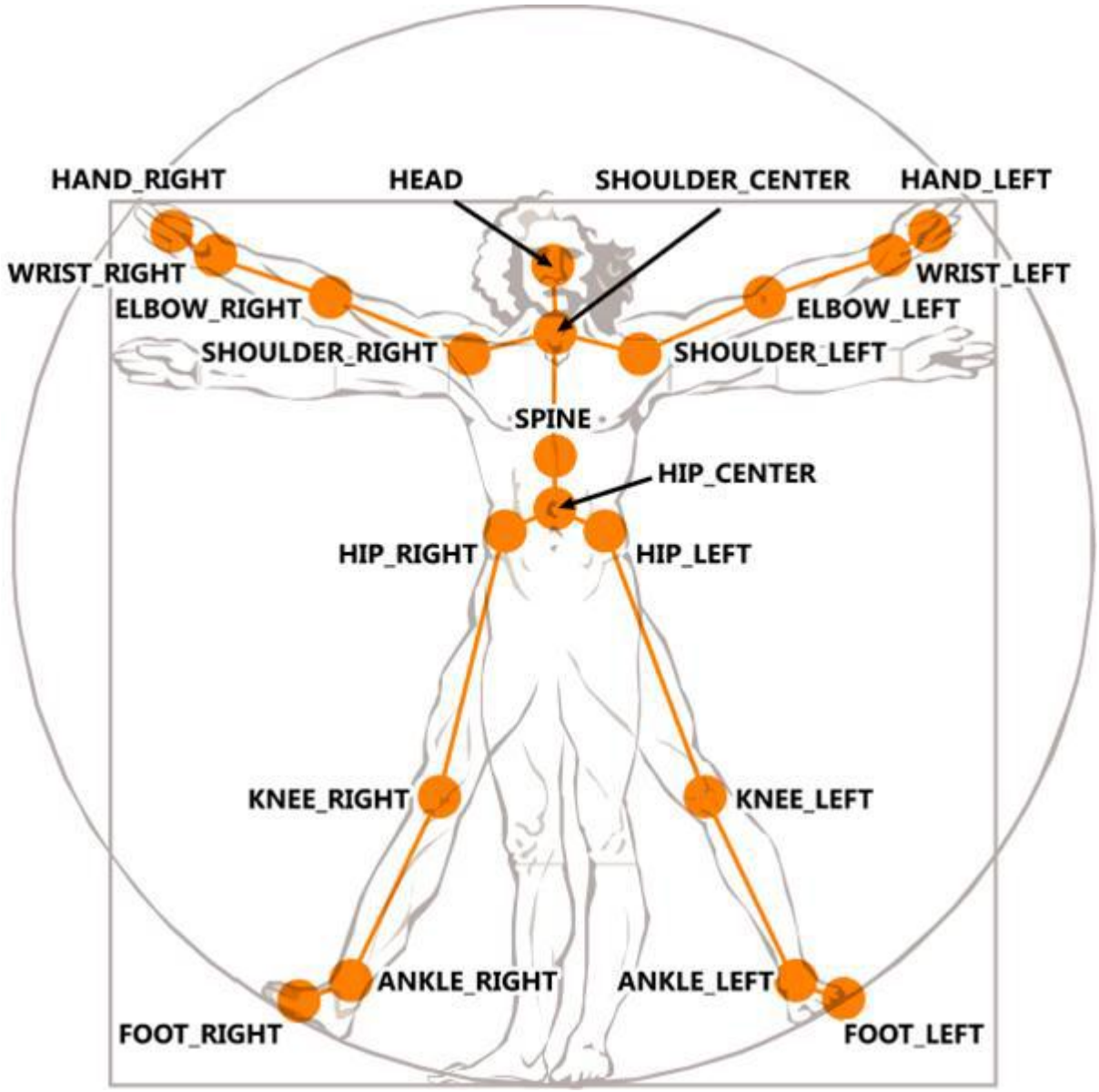


Figura 60 Scheletro tracciato dalla kinect e relativi punti salienti

A.7 Funzionalità Audio

Il device dispone di 4 microfoni disposti lungo la parte inferiore del case. Il sensore Kinect è in grado di fornire le seguenti funzionalità audio:

- eliminazione dell'eco (Multichannel Echo Cancellation MEC);
- posizionamento della sorgente sonora (Beamforming);
- soppressione o riduzione del rumore.
- In più, se si utilizzano le API fornite dal Kinect SDK in cooperazione con lo Speech SDK, si possono anche dotare le applicazioni del voice recognition.

La classe fondamentale per la gestione del flusso audio è la KinectAudioSource. Il metodo Start consente di avviare l'operazione di cattura dell'audio da parte del dispositivo e restituisce lo stream dati contenente l'audio.
La classe KinectAudioSource fornisce anche informazioni sulla localizzazione della sorgente sonora, ed in particolare lo fa in due modi distinti:

- **Evento BeamChanged**: sollevato dalla classe ogni qualvolta si accorge che la sorgente sonora ha cambiato posizione. L'argomento dell'evento contiene l'angolo orizzontale dove è posizionata la sorgente (con 0 se la sorgente è immediatamente di fronte al device);
- **Proprietà SoundSourcePosition**: la proprietà contiene la stessa informazione restituita dall'argomento del precedente evento.

La classe KinectAudioSource dispone di una serie di proprietà che consentono di agire sui parametri dei microfoni per modificare eco, soppressione del rumore e altri effetti.

APPENDICE B

MICROSOFT KINECT – caratteristiche del device

La Microsoft Kinect è basata su una tecnologia software sviluppata internamente da Rare, una sussidiaria di Microsoft Game Studios, e su una range camera (sviluppata da PrimeSense) che interpreta le informazioni 3D provenienti da un sensore che utilizza una luce infrarossa strutturata per ricostruire la scena 3D. Questo sistema scanner 3D è chiamato Light Coding ed impiega una variante delle ben note tecnologie di ricostruzione di immagini 3D.
La Kinect è stata introdotta sul mercato solamente come un device da gioco, connettibile unicamente alla Microsoft Xbox 360, come controller concorrente del Nintendo Wii e della Playstation Move, con l'obiettivo di immergere il giocatore sempre più nel vivo dell'azione, senza l'uso di gamepad o body sensors.
Il punto di forza del sensore Kinect è racchiuso nel fatto che permette al computer di conoscere direttamente la terza dimensione, ovvero la profondità dell'ambiente in cui è immerso il giocatore; inoltre è in grado di riconoscere quando l'utente sta parlando, sa identificare gli utenti, e sa interpretare i movimenti e i gesti compiuti da ciascuno di essi.
L'impatto del sensore Kinect ben presto si è espanso ben oltre l'industria videoludica, infatti, complice la sua grande disponibilità ed il prezzo contenuto, molti ricercatori hanno pensato di utilizzarlo per inventare nuovi modi per interagire con le macchine, dai sistemi informativi, al controllo robotico.

Dal punto di vista tecnico, Il sensore Kinect è una barra orizzontale connessa ad una piccola base motorizzata ed è disegnato per essere posizionato sopra o sotto un televisore. Il device è equipaggiato anche di una telecamera RGB, un sensore di profondità e un multy-array di microfoni che garantiscono il motion tracking 3D di tutto il corpo, il riconoscimento della faccia e della voce.
Nella Kinect l'assemblaggio dell'illuminazione e l'assemblaggio dell'immagine catturata sono controllati in accordo ad una relazione spaziale fissa. Questa configurazione, insieme alle tecniche di processamento usate dall'image processor, rendono possibile un mapping 3D usando solo l'immagine catturata, senza movimento relativo tra il sistema di illuminazione e il sistema di cattura, e senza il coinvolgimento di parti in movimento.
Per semplificare la computazione della mappa 3D e dei cambiamenti intercorsi in tale mappa dovuti al movimento di un oggetto, l'assemblaggio di illuminazione e immagine è strutturato in modo che l'asse ottico passante per il centro della fotocamera che cattura l'immagine, e lo spot coincidente con la sorgente luminosa, siano paralleli ad uno degli assi dell'immagine catturata dal sensore, convenzionalmente l'asse X, mentre l'asse Z corrisponde alla distanza dal device.

Uno shift lungo la direzione Z di un punto dell'oggetto, genera uno shift anche lungo l'asse X nello spot pattern osservato nell'immagine. Le coordinate Z dei punti dell'oggetto, così come gli shift di coordinate lungo Z nel tempo, possono perciò essere stimate a partire dalle divergenze osservate tra le misurazioni degli shift lungo X degli spot prodotti dall'immagine catturata dalla Kinect, e un'immagine di riferimento presa ad una distanza Z ben nota. Gli shift lungo Y possono essere trascurati.
Questo approccio, una sorta di triangolazione, è particolarmente appropriato per il mapping 3D che fa uso di pattern di spot incorrelato. Per generare la mappa 3D di un oggetto, gli image processors comparano i gruppi di spot in ciascuna area dell'immagine catturata con l'immagine di riferimento per trovare il gruppo che garantisce il miglior matching possibile. Lo shift relativo tra i matching group di spot nell'immagine fornisce lo shift lungo Z. Lo shift nello spot pattern può essere misurato usando la correlazione d'immagine o altri metodi di image matching ben noti.
Il sensore di profondità consiste di un proiettore laser di infrarossi combinato con un sensore CMOS monocromatico che cattura dati video in 3D sotto qualsiasi condizione di luce ambientale. Il range del sensore di profondità è regolabile e il software kinect è in grado di calibrarlo in modo automatico sulla base del gioco o dell'ambiente fisico del giocatore, in modo da adeguarsi alla presenza di mobili o altri tipi di ostacoli.
L'output video del sensore è di 30Hz, lo stream video RGB ha una risoluzione di 8bit VGA 640x480 pixel, con un filtro colore di Bayer, mentre il sensore di profondità percepisce video in VGA con 11bit cioè 2048 livelli di sensibilità. Il range di utilizzo della kinect, quando utilizzata con software Xbox è compreso tra 1.2 e 3.5 metri di distanza. L'area richiesta per giocare con la kinect è all'incirca di 6 metri quadri, un volume di 12 metri cubi, nonostante il sensore sia in grado di effettuare il tracciamento in un range più ampio, tra 0.7 e 6 metri. Il sensore ha un angolo di visione di 57° orizzontalmente e 43° verticalmente, mentre il pivot motorizzato può garantire ulteriori 27° sia in alto che in basso. Il campo orizzontale della kinect alla distanza minima di 0.8 metri è pari a circa 87cm, e il campo verticale a 63cm, che comporta una risoluzione di 1.3mm per pixel.
Come ampiamente descritto in precedenza, la kinect è formata essenzialmente da 3 tipi di input, una normale telecamera, un sensore di profondità e un array di microfoni. L'output è costituito da due matrici, una per la camera RGB e un'altra per la depth camera. La prima viene mostrata direttamente in modo grafico, in quanto trattasi di semplicissime successioni di immagini alfaRGB; la seconda è di solito utilizzata nel riconoscimento dello scheletro, può essere mappata come una range image, con una colorazione che esprime la profondità degli oggetti come riportato in figura.

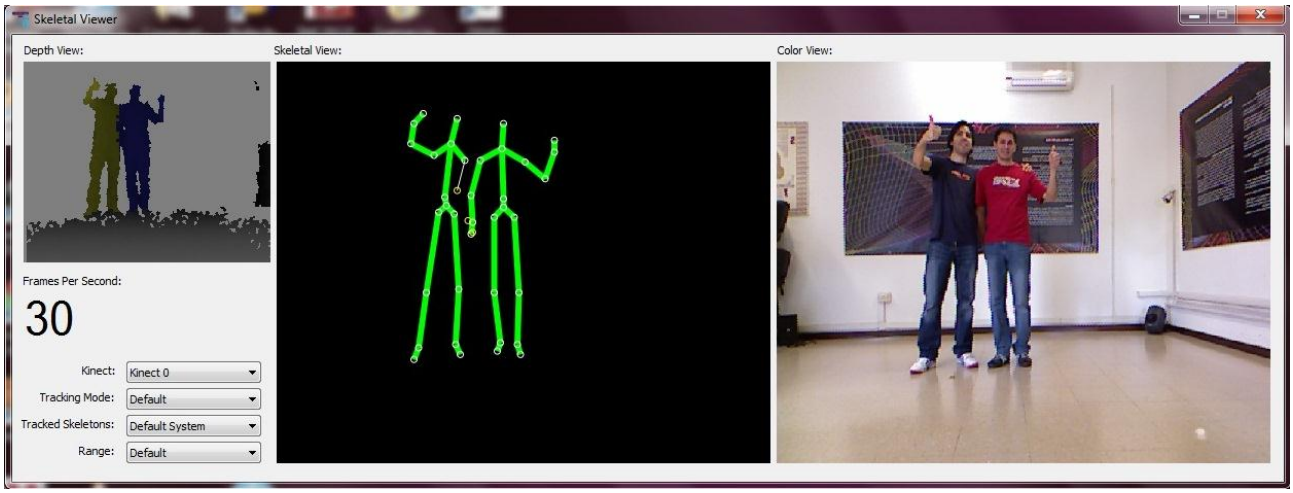


Figura 61 Flussi video kinect: Range image a sinistra, Skeleton tracking al centro, RGB image a destra

APPENDICE C

CASI DI STUDIO

C.1 CUCINA

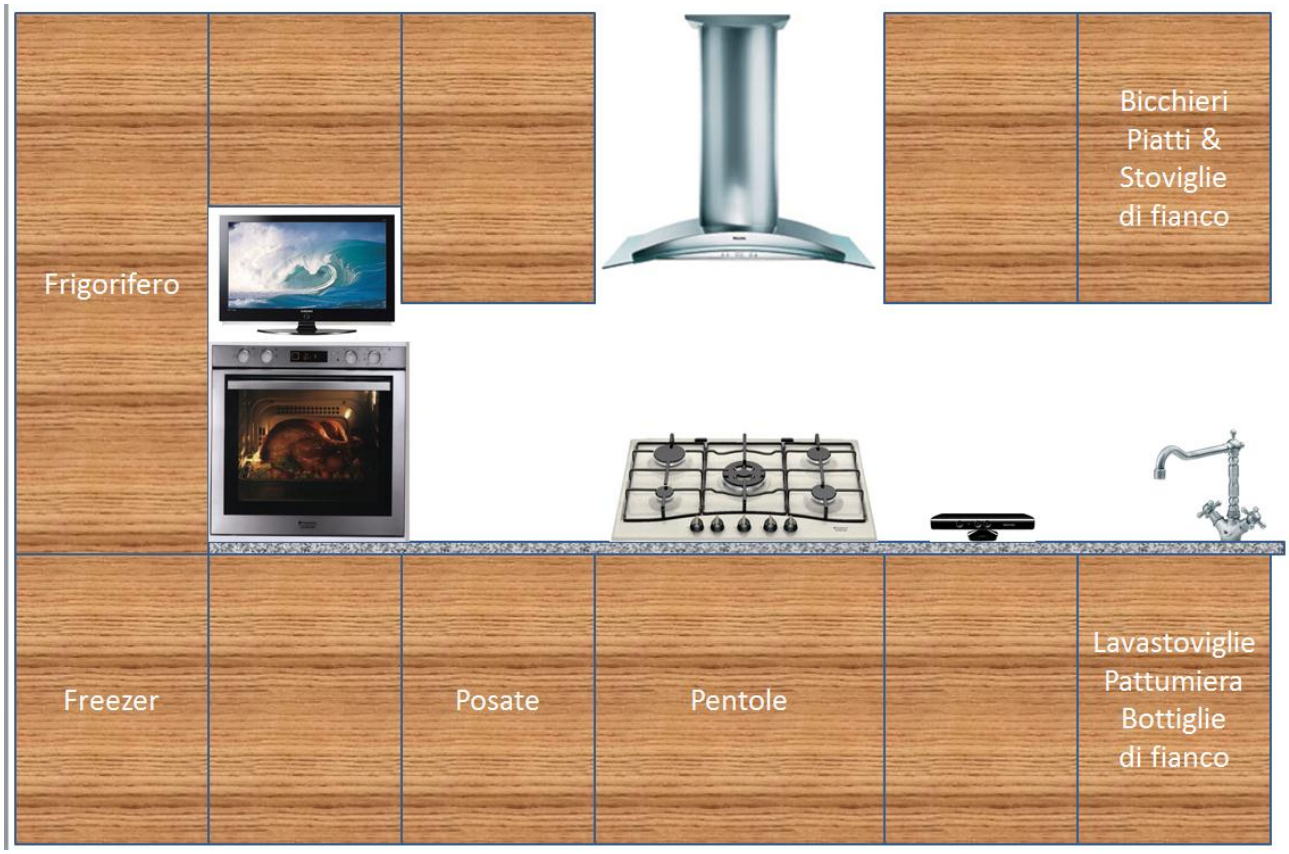


Figura 62 cucina lato A

Sulla base delle caratteristiche hardware di cui è equipaggiata la Microsoft Kinect, lo scenario tipico di utilizzo è quello indoor, e il contesto operativo è all'incirca quello di una stanza quadrata di lato 4 metri.
Il caso di studio che si è ritenuto più opportuno ai fini della caratterizzazione acustica e sperimentale è la cucina poiché, in primo luogo, è una stanza domestica caratterizzata da una moltitudine di "oggetti suonanti" e, soprattutto, è un ambiente dove l'utente non vedente ha maggiormente necessità di supporto per potersi muovere in autonomia e rendersi autosufficiente.

Figura 63 cucina lato B

Di seguito sono proposte alcune foto che raffigurano la cucina e gli oggetti che la caratterizano.



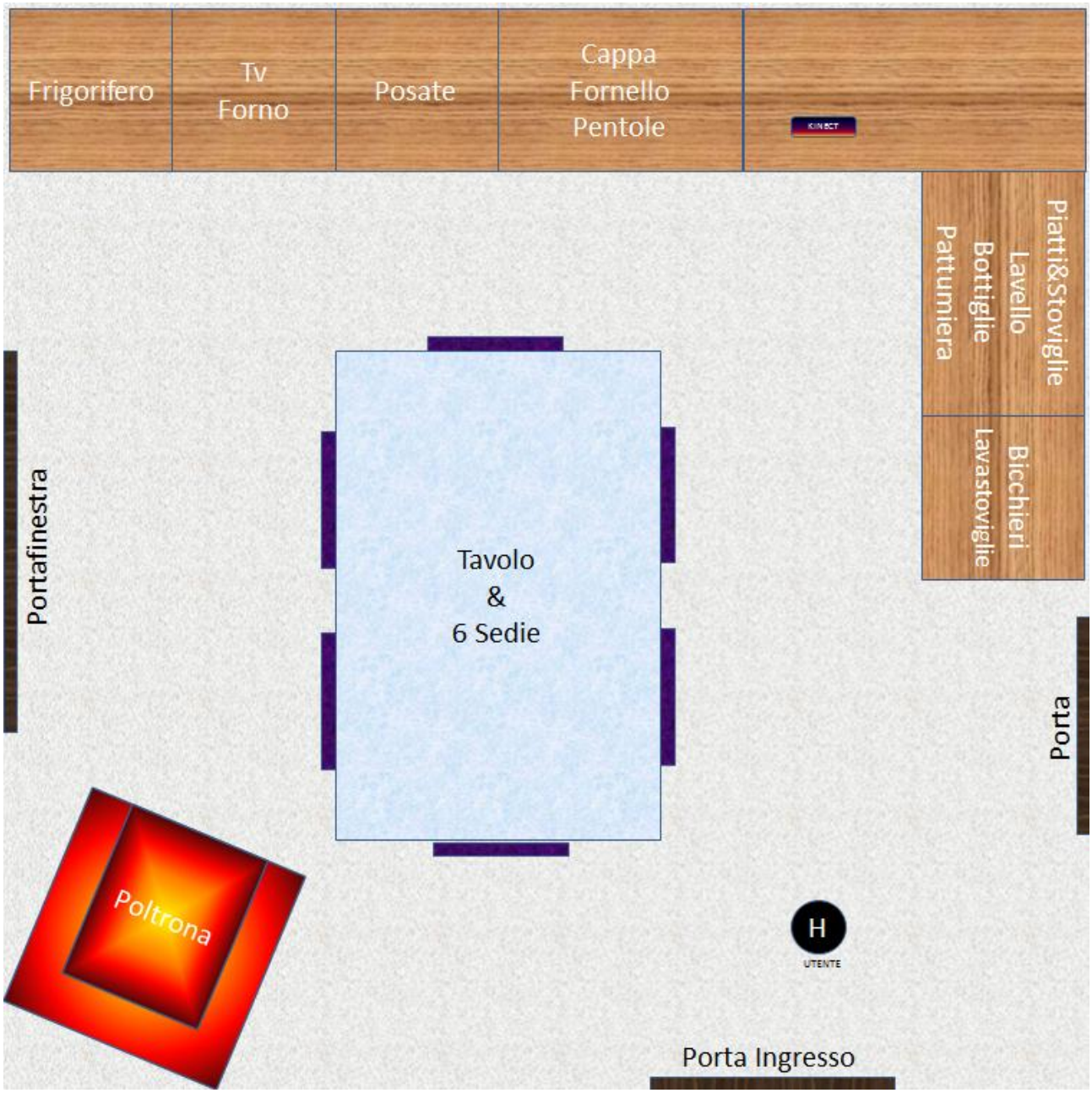



Figura 64 cucina vista dall'alto con tutti gli oggetti scelti

Oggetti individuabili

Di seguito viene proposta una tabella in cui vengono riassunti tutti gli oggetti individuabili all'interno della cucina, con una breve descrizione del materiale di cui sono composti, e del tipo di auditory icon utilizzata per caratterizzarli.

Oggetto	**Materiale**	**Suono**
Frigorifero	Legno-ferro	Porta che si apre-e chiude
Forno	Acciaio	Sportello che si apre, ventola che raffredda e timer che suona
Tv	Plastica-vetro	Sigla del TG5
Posate	acciaio	Tintinnio posate a contatto

Cappa	acciaio	Ventola di aspirazione alla massima potenza
Fornello	acciaio	Gas acceso, pentola sul fuoco
Pentole	acciaio	Coperchi che sbattono
Piatti & stoviglie	ceramica	Piatti a contatto che sbattono
Lavello	Acciaio	Rubinetto aperto, acqua che scende per lo scarico in modo rumoroso
Pattumiera	Ferro	Apertura chiusura pattumiera
Bicchieri	vetro	Tintinnio di bicchieri a contatto
Lavastoviglie	Legno-ferro	Fase di lavaggio
Porta soggiorno	legno	Tock tock su legno
Porta ingresso	Legno pesante	Porta che cigola con suono di serratura e chiavi
Tavolo	vetro	Tock tock su vetro
Sedie	legno	Sedia spostata, trascinata su pavimento
Poltrona	Legno e Stoffa imbottita	Poltrona spostata, trascinata pesantemente sul pavimento
Porta porticato	Legno-vetri	porta che si apre e chiude
Ventilatore	Legno-vetro	Pale che ruotano, flusso d'aria generato
Interruttore luce	plastica	Click dell'interruttore

C.2 LABORATORIO DEI-P

Dal momento che in fase implementativa e di sperimentazione è stato comodo utilizzare un caso di studio vicino all'ambiente di lavoro, si è pensato fosse opportuno testare l'applicazione anche in uno scenario reale come il laboratorio 2 del DEI-P, caratterizzato dalla presenza di tutta una serie di oggetti tipici dei luoghi di lavoro e degli uffici.

Oggetti individuabili

Di seguito viene proposta una tabella in cui vengono riassunti tutti gli oggetti individuabili all'interno del laboratorio, con una breve descrizione del materiale di cui sono composti, e del tipo di auditory icon utilizzata per caratterizzarli.

Oggetto	**Materiale**	**Suono**
Poster	Film plastica	Poster che sbatte contro la parete mosso dal vento
Bacheca	Legno	Rumore elettrico, bzzz di fondo
Armadietto di ferro	Ferro	Tock-tock-tock ripetuti su ferro
Finestra destra	Legno-vetro	Veneziana sollevata 2 volte
Scrivanie	Legno	Tock sul tavolo e apertura cassetti
Finestra centrale	Legno-vetro	Veneziana sollevata 2 volte
Scrivania	Legno	Tock-tock su legno
Finestra sinistra	Legno-vetro	Veneziana sollevata 2 volta

Armadio	Ferro-vetro	Apertura e chiusura porte ad anta
Altoparlanti	Legno	Emissione di un suono caratteristico di grandi altoparlanti, ricco di bassi
Porta	Legno	Clack maniglia e chiusura porta
Tastiere	Plastica	Motivetto suonato con la tastiera

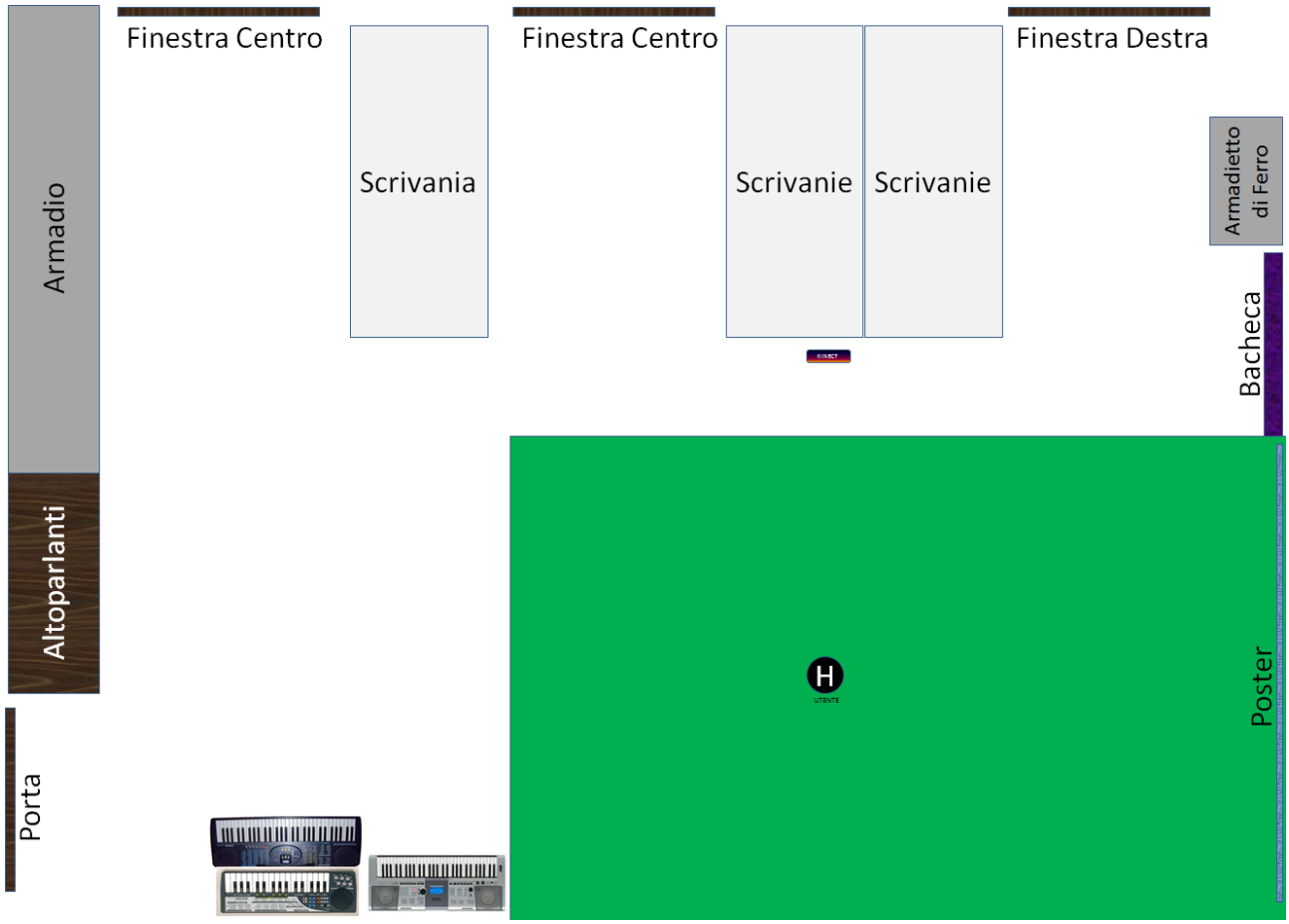
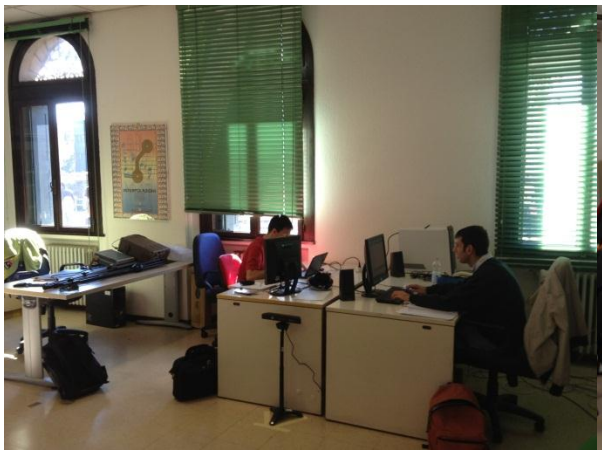


Figura 65 vista dall'alto DEI-P

Di seguito sono riportate alcune foto che raffigurano il laboratorio e gli oggetti che lo caratterizano.




BIBLIOGRAFIA

Bibliografia e testi di riferimento:

- [1] W. W. Gaver, "Auditory icons: Using sound in computer interfaces". Human-Computer Interaction, 1986.
- [2] W. W. Gaver. "What in the world do we hear? an ecological approach to auditory event perception". Ecological Psychology, 1993;
  e W. W. Gaver, "How do we hear in the world? explorations of ecological acoustics". Ecological Psychology, 1993.
- [3] Pietro Polotti e Davide Rocchesso, "Sound to Sense - Sense to Sound: A state of the art in Sound and Music Computing". 2008. (capitoli 8,9,10).
- [4] J. J. Gibson, "The ecological approach to visual perception". Lawrence Erlbaum Associates, Mahwah, 1986.
- [5] H. McGurk e J. MacDonald, "Hearing lips and seeing voices". Nature, 1976.
- [6] J. K. O'Regan e A. Noë, "A sensorimotor account of vision and visual consciousness". Behavioral and Brain Sciences, 2001.
- [7] A. Noë, "Action in perception". MIT press, Cambridge, Mass., 2005.
- [8] K. A. Kaczmarek, J. G. Webster, P. Bach-y-Rita e W. J. Tompkins, "Electrotactile and vibrotactile displays for sensory substitution systems". IEEE Trans. Biomedical Engineering, 1991.
- [9] P. B. L. Meijer, "An experimental system for auditory image representations". IEEE Trans. Biomedical Engineering, 1992.
- [10] J. K. Hahn, H. Fouad, L. Gritz e J. W. Lee, "Integrating sounds in virtual environments. Presence: Teleoperators and Virtual Environment". 1998.
- [11] T. A. Stoffregen e B. G. Bardy, "On specification and the senses". Behavioral and Brain Sciences, 2001.
- [12] G. Rizzolatti e C. Sinigaglia, "So quel che fai. Il cervello che agisce e i neuroni specchio". Milano, Raffaello Cortina Editore, 2006.
- [13] A. Murata, L. Fadiga, L. Fogassi, V. Gallese, V. Raos e G. Rizzolatti, "Obiect presentation in the ventral premotory cortex (area F5) in the monkey". 1997.
- [14] A. Murata, V. Gallese, G. Luppino, M. Kaseda e H. Sakata, "Selectivity for the shape, size and orientation of objects for grasping in neurons of monkey parietal area AIP". 2000.
- [15] V. Gallese, P. Migone e M.N. Eagle, "La simulazione incarnata: i neuroni specchio. Le basi neurofisiologiche dell'intersoggettività ed alcune implicazioni per la psicoanalisi". Psicoterapia e Scienze Umane, 2006.
- [16] E. Kohler et al., "Hearing sounds, understanding actions: action representation in mirror neurons". Science, 2002.
- [17] L. Fogassi et al., 2003.
- [18] M. Iacoboni et al., "Grasping the intentions of others with one's own mirror neuron system". PLoS Biology, 2005.

- [19] G. Buccino et al., “Listening to action-related sentences modulates the activity of the motor system: A combined TMS and behavioral study”. Cognitive Brain Research, 2005.
- [20] L. Ziglari, “Affordance and Second Language Acquisition”. European Journal of Scientific Research, 2008.
- [21] M.C. Corballis, “Mirror neurons and the evolution of language”. Brain & Language, 2009.
- [22] G. Di Pellegrino et al., “Understanding motor events – A neurophysiological study”. Experimental Brain Research, 1992.
- [23] P.F. Ferrari et al., “Mirror neurons responding to the observation of ingestive and communicative mouth actions in the monkey ventral premotor cortex”. European Journal of Neuroscience, 2003.
- [24] G. Rizzolatti e M.A. Arbib, “Language Within our Grasp”. Trends in Neuroscience, 1998.
- [25] G. Rizzolatti, e L. Craighero, “The Mirror – Neuron System”. Annual Review of Neuroscience, 2004.
- [26] S. Barrass e G. Kramer, “Using sonification”. Multimedia Systems, 1999.
- [27] C. Hayward, “Listening to the earth sing”. In G. Kramer editor, “Auditory Display: Sonification, Audification and Auditory Interfaces”. Addison-Wesley, Reading, Mass., 1994.
- [28] N.A. Stanton e J. Edworthy, “Auditory warnings and displays: an overview”. In N.A. Stanton e J. Edworthy editors, “Human Factors in Auditory Warnings”. Ashgate, Aldershot, UK, 1999.
- [29] M.M. Blattner, D.A. Sumikawa e R.M. Greenberg. “Earcons and icons: their structure and common design principles”. Human-Computer Interaction, 1989.
- [30] S.A. Brewster, “Using nonspeech sounds to provide navigation cues”. ACM Trans. on Computer-Human Interaction, 1998.
- [31] T. Hermann e H. Ritter, “Sound and meaning in auditory data display”. Proceedings of the IEEE, 2004.
- [32] M. Rath e D. Rocchesso, “Continuous sonic feedback from a rolling ball”. IEEE Multimedia, 2005.
- [33] B. N. Walker e G. Kramer, "Ecological psychoacoustics and auditory displays: Hearing, grouping, and meaning making". In J. Neuhoff editor, “Ecological psychoacoustics”. New York: Academic Press, 2004.
- [34] B. N. Walker, A. Nance e J. Lindsay, "Spearcons: Speech-based earcons improve navigation performance in auditory menus". International Conference on Auditory Display, London, 2006.
- [35] B. N. Walker e J. Lindsay, "Effect of beacon sounds on navigation performance in a virtual reality environment". Ninth International Conference on Auditory Display ICAD2003, Boston, 2003.
- [36] D. Palladino e B. N. Walker, "Learning rates for auditory menus enhanced with spearcons versus earcons". International Conference on Auditory Display, Montreal, 2007.

- [37] J. Wilson, B. N. Walker, J. Lindsay, C. Cambias e F. Dellaert, "SWAN: System for Wearable Audio Navigation". International Symposium on Wearable Computers, Boston, 2007.
- [38] R. D. Patterson, "Guidelines for auditory warning systems on Civil Aircraft". Civil Aviation Authority, London, 1982.
- [39] K. Hemenway, "Psychological issues in the use of icons in command menus". CHI'82 Conference on Human Factors in Computer Systems, New York, 1982.
- [40] P. Keller e C. Stevens, "Meaning from environmental sounds: Types of signal-referent relations and their effect on recognizing auditory icons". Journal of Experimental Psychology, 2004.
- [41] S. Brewster, P. C. Wright, e A.D.N. Edwards, "A detailed investigation into the effectiveness of earcons". First International Conference on Auditory Display, Santa Fe, New Mexico, 1992.
- [42] P. Milgram, H. Takemura, A. Utsumi e F. Kishino, “Augmented reality: A class of displays on the reality-virtuality continuum”.
- [43] M. Geronazzo, S. Spagnol e F. Avanzini, “Customized 3D sound for innovative interaction design”.
- [44] S. Benford e L. Fahlén, “A spatial model of interaction in large virtual environments”. European Conference on Computer-Supported Cooperative Work, Norwell, USA, 1993. Kluwer Academic Publishers.
- [45] A. Walker e S. Brewster, “Spatial audio in small screen device displays”. Personal and Ubiquitous Computing, 2000.
- [46] E.M. Wenzel, M. Arruda, D.J. Kistler e F.L. Wightman, “Localization using non-individualized head-related transfer functions”. The Journal of the Acoustical Society of America, 1993
- [47] D.R. Begault, E.M. Wenzel e 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”. Journal of the Audio Engineering Society. 2001.
- [48] D.R. Begault, “3-D sound for virtual reality and multimedia”. Academic Press Professional, San Diego, USA, 1994.
- [49] E.A.G. Shaw, “Binaural and Spatial Hearing in Real and Virtual Environments”. Chapter Acoustical features of human ear, R.H. Gilkey e T.R. Anderson, Lawrence Erlbaum Associates, Mahwah, USA, 1997.
- [50] D.W. Batteau “The role of the pinna in human localization”. Biological Sciences, London, 1967.
- [51] F. Avanzini, “Sound in Space” - Chapter 4 – Dispense di Informatica Musicale. (paragrafi 4.5-4.6). Università degli Studi di Padova.
- [52] Nicola Scattolin, tesi: “A comparison between gesture tracking models and the development of an interactive mobility aid system for the visually impaired”. (paragrafo 8.5.2). Università degli Studi di Padova, 2012.
- [53] L.A. Ludovico, D.A. Mauro, e D. Pizzamiglio, “Head in space: a head tracking based bineural spatialization system”. LIM - Laboratorio di Informatica Musicale, Dipartimento di Informatica e Comunicazione (DICo), Università degli Studi di Milano.

- [54] T. Dingler, J. Lindsay e B.N. Walker, “Learnability of sound cues for environmental features: auditory icons, earcons, spearcons and speech”. Ludwig-Maximilians-Universität München & School of Psychology Georgia Institute of Technology.
- [55] Fabiana Galiussi, tesi: “Dal gesto alla parola: implicazioni glottodidattiche del sistema dei neuroni specchio” università Ca’ Foscari, Venezia, 2009.

References Microsoft Kinect:

- [1] Kinect for Windows SDK Web Site: http://research.microsoft.com/en-us/um/redmond/projects/kinectsdk/
- [2] Channel9 Kinect for Windows SDK Quickstarts: http://dev9.channel9.msdn.com/Series/KinectSDKQuickstarts
- [3] Channel9 Coding4Fun: http://dev9.channel9.msdn.com/coding4fun/kinect