UNIVERSITÀ DEGLI STUDI DI MILANO
FACOLTÀ DI SCIENZE E TECNOLOGIE

CORSO DI LAUREA TRIENNALE IN INFORMATICA MUSICALE

TECNICHE DI SONIFICAZIONE DELLE IMMAGINI: IL CASO DI STUDIO DEL VOLTO UMANO

Relatore: Prof. Luca Andrea Ludovico
Correlatore: Prof. Giorgio Presti

Tesi di Laurea di:
Bonadede Davide
Matr. Nr. 828975

ANNO ACCADEMICO 2015-2016

Indice

Capitolo 1

Introduzione

La sonificazione è una tecnica per comunicare informazioni che utilizza il suono per descrivere dati e interazioni, senza usare i segnali vocali o musica [1], al fine di facilitarne l'interpretazione [2]. Come disciplina negli ultimi anni si sta diffondendo in maniera capillare in campi di applicazione molto diversi tra loro: biomedicina, interfacce per persone non vedenti, sismologia, interazione con computer desktop e dispositivi mobili eccetera. Il motivo per cui viene usato il suono per descrivere la realtà è che il nostro sistema uditivo è particolarmente efficiente nell'identificare sorgenti sonore e comprendere il significato, talvolta complesso, che esse vogliono comunicare a più livelli di astrazione, partendo dalla stessa informazione giunta al nostro orecchio, anche quando per esempio la scena uditiva e composta da tanti suoni diversi miscelati tra loro.

Questo elaborato tratta l'uso della sonificazione al fine di descrivere la struttura del volto umano, utilizzando tecnologie informatiche in grado di elaborare le immagini in movimento e la sintesi sonora.

L'obiettivo non è quello di descrivere le emozioni di una persona, bensì di trasdurre nel dominio audio le microespressioni facciali, lasciando all'utente il compito di interpretarle, semplicemente ascoltando il risultato prodotto dal software.

L'esposizione della tesi è suddivisa in varie fasi: nel secondo capitolo viene trattato il Sonification Space, di modo da chiarire il collocamento dell'elaborato al suo interno e definirne lo scopo in base ai risultati sonori, insieme alle varie tecniche di

sonificazione. Nel terzo capitolo vengono esposte le tecnologie utilizzate per la realizzazione del lavoro di tesi, nella fattispecie il modulo software FaceOSC, accompagnato da una trattazione sull'elaborazione digitale delle immagini e dei video, seguito dal linguaggio Pure Data. Nel Capitolo 4 viene descritto più nel dettaglio l'applicativo complessivamente, partendo dallo schema generale e arrivando spiegare ogni piccolo componente. Nel Capitolo 5 viene raccontata l'esperienza del Meet Me Tonight, la notte dei ricercatori, in occasione della quale si è potuto testare l'applicativo creato su un gran numero di persone durante due giornate intere, ai giardini di Indro Montanelli di Milano. Nel sesto e ultimo capitolo vengono esposte invece le conclusioni e i possibili sviluppi futuri, parlando anche delle possibili applicazioni pratiche del lavoro svolto.

Capitolo 2

Stato dell'arte

In questo capitolo viene definito l'auditory display per comprendere effettivamente il contesto in cui viene impiegata la sonificazione; successivamente vengono descritti il Sonification Space e le varie tecniche di sonificazione, per dare una panoramica più completa di come si può operare nel momento in cui si vogliono comunicare informazioni attraverso il suono. Infine verranno analizzate le modalità con cui si sonificano le immagini e i video.

2.1 Definizione di auditory display

Per auditory display si intende, sia il contesto in cui si propagano segnali acustici udibili, sia il complesso di tecnologie capaci di creare onde sonore e di diffonderle. Si parla quindi di dispositivi come altoparlanti, cuffie e apparecchi per la conduzione ossea, mentre per quanto riguarda il contesto in cui sono presenti gli ascoltatori, si stabiliscono i compiti che si prefigge la sonificazione e i vincoli entro cui operare. Inoltre l'auditory display, non comprende solo le interfacce che producono suono, ma anche quelle che, ricevendo dati tramite l'interazione, generano un controllo uditivo bidirezionale tra utente e le tecnologie presenti.

2.2 Tecniche di sonificazione e Sonification Space

Le tecniche di sonificazione si differenziano per i differenti approcci con cui si relaziona ciò che si vuole significare e la sua trasformazione in suono [1]. in questo paragrafo vengono esposte nel dettaglio le principali tecniche di sonificazione, facendo riferimenti anche ad esempi di usi nella realtà. Prima di affrontarle nello specifico è opportuno descrivere le caratteristiche fondamentali di una sonificazione, che Hermann riassume nei seguenti punti [3]:

- Il suono deve riflettere proprietà o relazioni oggettive dei dati in ingresso.
- La trasformazione dei dati deve essere sistematica, deve esserci cioè una precisa definizione di come il suono cambi in funzione dei dati.
- La sonificazione deve essere riproducibile, quindi dagli stessi dati e le stesse interazioni, il suono risultante deve essere strutturalmente identico.
- Il sistema infine deve essere concepito per essere usato con dati differenti e anche in ripetizione degli stessi dati in ingresso.

Le diverse tecniche di sonificazione possono essere raggruppate in 5 tipologie: Audificazione, Auditory Icons, Earcones, Parameter Mapping e Model-Based Sonification.

L'audificazione consiste nell'interpretare un segnale monodimensionale, o bidimensionale nel caso di set di dati multivariati, come un'ampiezza che varia periodicamente in funzione del tempo, per poterlo poi riprodurre attraverso degli altoparlanti e quindi destinarlo a degli ascoltatori.

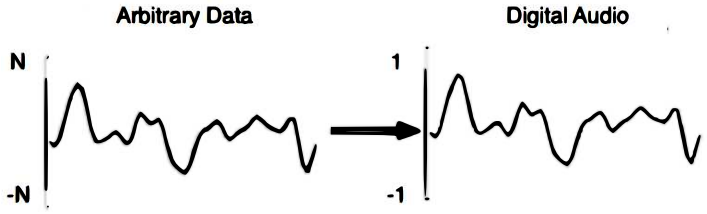

Figura 1: *Trasformazione di set di dati in segnale acustico.*

Storicamente fu l'invenzione del telefono ad essere definita come la prima trasformazione di un insieme di informazioni (segnali elettrici) in suono (e viceversa). Le applicazioni di questa tecnica sono diffuse in medicina (trasformazione di EEG in suono), in fisica (ad esempio audificazione dell oscillazioni quantistiche di un atomo di elio), in statistica e in ambiti più artistici (creazione di nuovi suoni nella computer music).

Le Auditory Icons invece sono dei suoni non verbali che sono familiari in quanto si ritrovano nella realtà di tutti i giorni e basano il loro significato sulle nostre precedenti esperienze. Esse sono comparse con l'avvento dei personal computer come corrispettivo sonoro delle icone visive che si trovano nel desktop. Per creare Auditory Icons si può far uso di un approccio basato sul segnale, il quale tenta di imitare dei suoni che si trovano ovunque quotidianamente (similmente allo scheumorfismo), oppure attraverso dei modelli fisici che descrivono l'oggetto che produce eventi acustici, con l'ausilio di strumenti matematici in tempo reale. Le applicazioni di questo modo di comunicare, oltre che all'interno dei pc, si trovano anche in ambito sanitario, con la traduzione in suono dei segni vitali (battito cardiaco, livello CO2, respirazione ecc.) e per dispositivi mobili (es. applicazioni che simulano suoni di oggetti).

Le Earcons sono dei piccoli frammenti musicali le cui proprietà sono associate ai differenti parametri dei dati da comunicare. La differenza cruciale con le Auditory Icons sta nel fatto che nelle Earcons non esiste un vera relazione tra il significato e il suono prodotto che lo vuole trasmettere. Inizialmente erano oggetto di ricerca all'interno di un contesto di applicazioni di sicurezza, come le unità di terapia intensiva, sotto forma di avvertimenti sonori. Più in generale vengono utilizzate anch'esse all'interno di un pc in maniera complementare alle Auditory Icons, ma si possono trovare per esempio anche all'interno di un aereo (segnali acustici per far interagire i passeggeri con gli assistenti di volo). Un altro esempio interessante riguarda i leitmotiv composti da Prokof'ev in "Pierino e il lupo", che associano ad ogni personaggio dell'opera una particolare e riconoscibile particella melodica.

Il Parameter Mapping traduce in suono le caratteristiche di un set di dati, mappandole con quelle di parametri acustici, come l'altezza, il timbro, la brillantezza ecc.. All'aumentare del numero di attributi sonori, aumenta la multidimensionalità

della sonificazione.

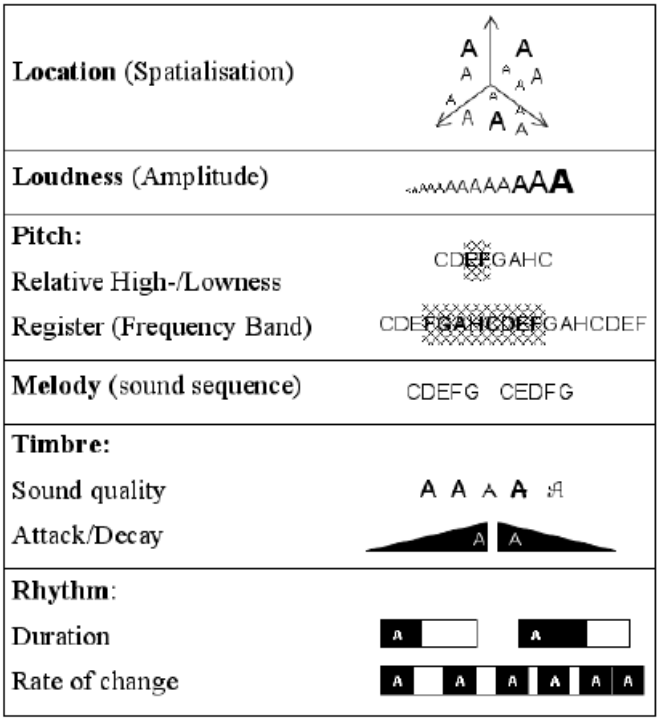

Figura 2: *Parameter Mapping Sonification.*

Nella realtà questa tecnica viene utilizzata per ricavare di dati, per tecnologie assistive e anche per scopi artistici (la traduzione in suono delle immagini, installazioni sonore e la mappatura di processi statistici e stocastici, come nell'opera orchestrale "Metastasi" di Iannis Xenakis).

Infine la sonificazione **Model Based** si focalizza su come le risposte acustiche sono generate in funzione delle azioni dell'utente, creando un sistema capace di generare segnali acustici con il quale interagire. Essa si differenzia dalla Parameter Mapping in quanto, al contrario di quest'ultima, qui viene descritta l'architettura di un sistema dinamico, non semplicemente le caratteristiche di un contesto descritte da una sensazione uditiva. Questo sistema è descritto da cinque blocchi fondamentali: lo spazio dei dati, lo spazio che definisce il modello, lo spazio sonoro e infine l'ascoltatore, che costituisce sia la destinazione dell'informazione sia il punto fondamentale

dell'interazione.

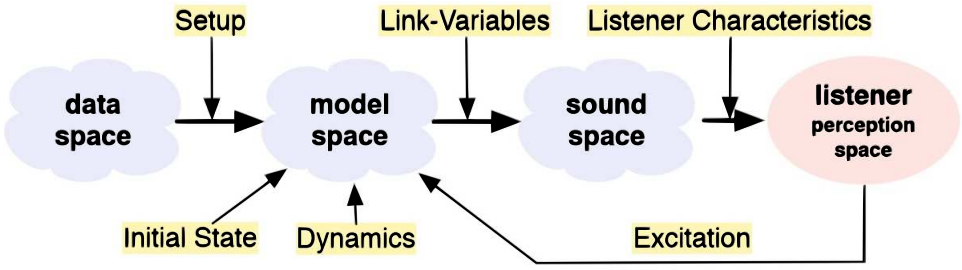

Figura 3: *Struttura Model-Based Sonification.*

Per quanto riguarda l'applicazione nella realtà, questa tecnica viene per esempio utilizzata in quei contesti in cui si vuole aumentare il grado di interazione con una macchina (ad esempio nei computer desktop).

Tutte le tecniche esposte di cui sopra, che descrivono esaurientemente il mondo della sonificazione, nel momento in cui si opera nella realtà spesso non sono applicate in maniera isolata. Per esempio la Parameter Mapping Sonification è particolarmente diffusa per la natura intrinseca del suono di essere multidimensionale, e quindi di riuscire a comunicare più informazioni, anche dinamiche, attraverso un solo evento sonoro, integrandosi quindi con altri metodi di diffusione di contenuti. Inoltre nella maggior parte dei casi in cui si vuole sonificare qualcosa, vi è un certo grado di interazione, ergo, da questo punto di vista, la Model-Based Sonification è altrettanto presente nelle realizzazioni comunicative. In sintesi, anche se ogni tecnica ha un suo specifico modo di operare, non esiste un confine netto oltre il quale si impone l'uso di una in particolare piuttosto che un'altra. Questo è una delle caratteristiche principali del Sonification Space.

Il Sonification Space è un strumento grafico in grado di aiutare l'utente finale a orientarsi sul modello di sonificazione in uso, basandosi sul risultato udibile finale [2].

Sull'asse x viene rappresentata la granularità temporale, mentre sull'asse y si trova il livello di astrazione dell'output sonoro (dal più basso, quello fisico, a quello più alto, ovvero il simbolico). Si può così suddividere questo spazio in più aree, che vanno dal

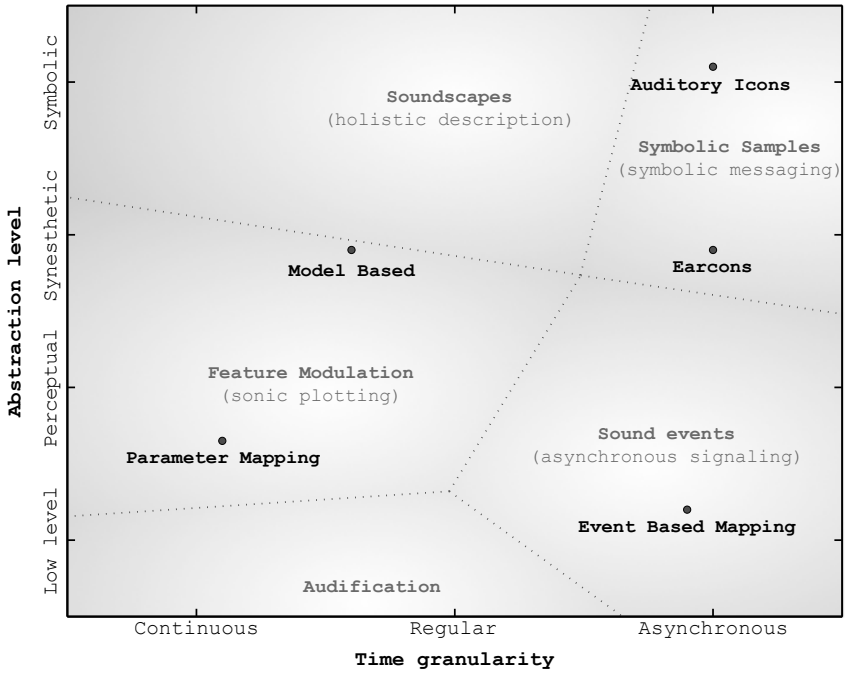

Figura 4: *Sonification Space.*

risultato più astratto e asincrono, quello riguardante i paesaggi sonori, a quello più basso e continuo che consiste nella traduzione diretta e immediata di un gruppo di dati in suono. Tuttavia è molto importante tenere presente che nonostante la suddivisione in aree dello spazio, esso è concepito per essere utilizzato come spazio continuo, per poter effettivamente collocare un tipo di sonificazione realizzata al suo interno senza vincoli particolari.

2.3 Sonificazione delle immagini e dei video

Come spiegato in [4], la sonificazione delle immagini consiste in un fenomeno sinestetico, che è capace di generare del suono che tenta di delineare degli elementi visivi, come colori o forme, fruiti contemporaneamente. Storicamente la prima macchina in grado di mettere in atto questo tipo di sinestesia, seppure attraverso un'ottica più artistica, fu il clavicembalo oculare, inventato dal padre gesuita Louis Bertrand Castel, che permetteva, tramite la pressione di un tasto, di far sollevare un pannello colorato

tramite l'associazione tra la scala utilizzata e lo spettro cromatico.

Figura 5: *Clavicembalo Oculare.*

Oggi esistono software come Coagula, il quale gestisce l'immagine come se fosse lo spettrogramma del suono prodotto, oppure come Audio Paint, che tratta ogni riga della figura in ingresso come un'oscillatore. Come affrontato in [5], in generale quando si intende descrivere attraverso eventi acustici un immagine, occorre considerare la natura del suono come fenomeno intrinsecamente legato al tempo, aspetto che per esempio non riguarda un singolo frame. Esistono dunque due modi di ricostruire il tempo partendo delle immagini: lo Scanning, che consiste nel passare la figura da sinistra verso destra a velocità fissata, concependola come lo spettro di cui si vuole ricostruire un suono, mentre il secondo è il Probing, dove c'è un puntatore che analizza man mano l'illustrazione seguendo le forme, i percorsi e la velocità in maniera arbitraria e flessibile. Entrambi i metodi possono essere combinati per fornire una sonificazione più ricca di significato.

Come esposto in [6], un'altro modo di descrivere figure mediante il suono è quello di assegnare a ogni colore uno strumento, sempre basandosi su associazioni di tipo

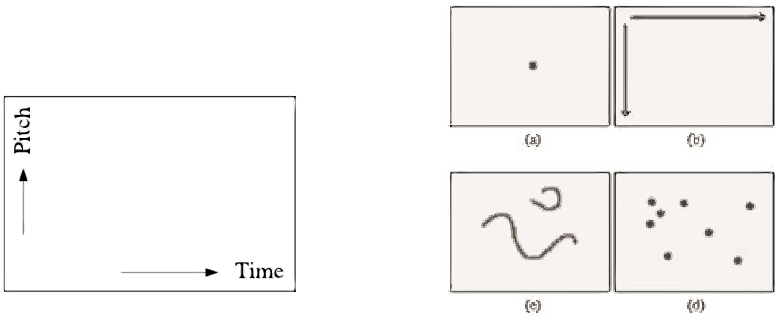

Figura 6: *Scanning e Probing.*

sensoriale, disponendoli all'interno di una griglia di Photoshop, la quale viene successivamente fatta coincidere con il piano roll di un DAW, trasformando così gruppi di pixel in note MIDI in successione, di modo riuscire a descrivere delle forme di un'opera, oltre alla semplice scansione di tutti i pixel, con l'uso del suono.

Partendo appunto dalla descrizione sonora di una raffigurazione, trattandola come lo spettro di quest'ultimo, si può applicare lo stesso procedimento ai video. I primi esperimenti in questo senso avevano il problema della sincronizzazione, risolto alterando temporalmente la parte figurativa e avvalendosi della trasformata di Fourier veloce per ridurre al minimo la latenza. Un esempio di trasformazione in eventi acustici di una sequenza di immagini in movimento, come sperimentato in [7], può essere ottenuta gestendo la colonna centrale dell'inquadratura come un insieme di valori da portare da RGB a una scala di grigi, che poi vengono utilizzati come parametri di un filtro che viene applicato su un rumore rosa, oppure si può utilizzare un metodo alternativo che consiste nel trattare sempre la colonna centrale come un frammento di piano roll, che permette l'invio di messaggi MIDI ad uno strumento reale per cui, per ogni pixel, le note più alte si trovano nella parte alta dei frame, mentre quelle più basso nella parte bassa, invece il tipo di strumento suonato è determinato dal valore della tinta e infine la brillantezza del colore indica il volume della nota generata. Una variante di questa tecnica si ottiene suddividendo sempre la medesima colonna in quattro parti di uguale altezza, all'interno delle quali vengono riprodotte note di diverse ottave, grazie al calcolo della distanza del colore di ogni pixel rispetto ai 12

assegnati all'interno di una tastiera colorata.

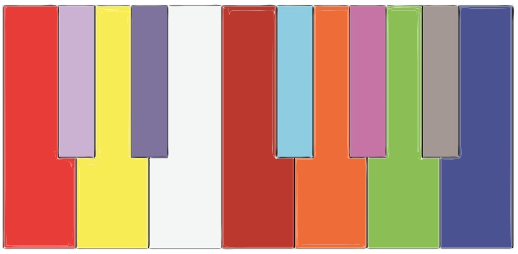

Figura 7: *Tastiera colorata.*

Queste ultime tre modalità sono concepite principalmente per i video in cui gli elementi dell'immagine scorrono continuamente da un lato all'altro (come lo scorrimento di un paesaggio), mentre un diverso modo di approcciare la sonificazione di immagini in movimento dove c'è un oggetto o una persona che si muove, come descritto in [8], può essere effettuato aggiungendo un altro strumento, chiamato Motiongrams, che altro non è che la rappresentazione nel tempo del movimento medio di ogni riga di pixel di un'inquadratura, risultando così come una raffigurazione ridotta del movimento, che può essere trasformata direttamente in spettro sonoro e successivamente in evento acustico con una IFFT.

Capitolo 3

Tecnologie Utilizzate

In questo capitolo vengono riportate tutte le tecnologie impiegate per la creazione dell'applicativo, assieme a una analisi di come si realizza il tracciamento del volto (Face Tracking), affinchè venga approfondito l'ambito di applicazione e la modalità di operare del modulo software FaceOSC. Verranno inoltre esposte le caratteristiche del linguaggio Pure Data e del protocollo OSC.

3.1 Face Tracking: usi e tecniche

Come analizzato in [9], il riconoscimento e il tracciamento del volto grazie a tecnologie informatiche, ha molte utilità in quanto il viso è un elemento essenziale all'interno delle interazioni umane. Esso può essere utilizzato per esempio per migliorare la vita delle persone disabili; come modo alternativo per utilizzare applicazioni, anche video ludiche, senza l'ausilio delle mani; in contesti di sicurezza dove si rende necessario autenticare le persone; all'interno di un'automobile per riconoscere quando l'autista ha una perdita di attenzione o un colpo di sonno ecc..

Di base il Face Tracking rappresenta un'ottima base di relazione tra uomo e macchina, per la quale si vuole sapere se in un'immagine sono presenti uno o più volti e la loro locazione. Un algoritmo fondamentale per riconoscere il volto umano è l'AAC, che sta per Active Appearence Model, il quale è in grado di far corrispondere un modello statistisco di una figura generata deformabile a una nuova immagine. La

modellazione si ottiene partendo da delle immagini piene di punti di riferimento etichettati che verranno poi uniti formando dei vettori, che a loro volta vengono trattati attraverso strumenti statistici; dopodichè le illustrazioni prese come esempio vengono deformate per ricavare una media e creare dei moduli che riescano a lavorare in maniera slegata da una sagoma fissa. La fase successiva consiste nella gestione di parametri che permettono all'AAC creato di adattarsi al meglio a nuove raffigurazioni dinamiche, cercando di basarsi su cambiamenti lineari, sempre con l'aiuto della statistica. Infine, una volta realizzato il modello, si cerca di irrobustirlo contro fattori come le condizioni di luce, le espressioni facciali e la posa della testa.

Un modo più avanzato per ottenere il tracciamento del volto è quello di utilizzare un prototipo tridimensionale del viso. Si differenziano nel campo due tipi di modello: rigido e non rigido. il primo porta con se il problema di non riuscire sempre a far fronte ai cambiamenti di espressione che può effettuare un volto nel tempo. Al contrario un modello non rigido è in grado di seguire i movimenti, come l'apertura e chiusura degli occhi e le variazioni della bocca. Quest'ultimo si ottiene costruendo una maschera in 3D dal primo frame del video contenente il viso e aggiornandolo continuamente utilizzando pochi moduli, adattandosi ai cambiamenti di blocchi di pixel.

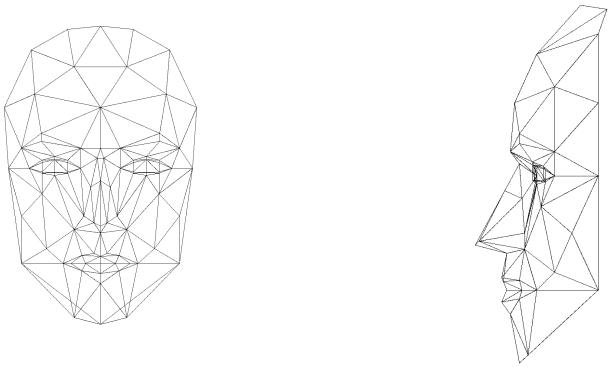

Figura 8: *Modellazione tridimensionale del volto.*

3.2 Tecnologie utilizzate

3.2.1 Applicativo FaceOSC

FaceOSC è un programma sviluppato da Kyle McDonald, partendo da un lavoro di Jason Saragih per il tracciamento del volto, con cui è stato sviluppato un applicativo chiamato ofxFaceTracker, che si appoggia su openFrameworks, un inseme di librerie in C++ open source per la creazioni di software [10]. Nella fattispecie FaceOSC invia un flusso messaggi attraverso un protocollo chiamato OSC, Open Sound Control, che verrà spiegato nel dettaglio. Il programma si basa su una modellazione tridimensionale dinamica del viso di una persona, comunicando in maniera costante come variano alcuni parametri. L'applicativo è stato creato partendo da OpenCV, Open Source Computer Vision Library, una libreria appunto che si basa sulla computazione visiva, anch'essa open source, generata per avere un'infrastruttura comune per tutti gli applicativi che fanno uso di computer vision [11].

3.2.2 Pure Data

Pure Data è un linguaggio di programmazione grafico che rende facile la creazione di applicativi che operano con i segnali [12]. Esso è anche un ambiente in continua evoluzione e un'alternativa gratuita a Max/MSP. Grazie alle molteplici interfacce multimediali è in grado di collegarsi a sensori, macchine e controller per gli scopi più disparati, come la generazione di installazioni interattive che si basano fondamentalmente sul suono. La sua struttura ricorda molto i moduli hardware capaci di processare segnali, inoltre, come linguaggio, tratta flussi di dati che vengono elaborati da elementi che possono essere collegati tra loro.

Gli elementi dell'ambiente di Pure data sono di 3 tipologie: Oggetti, Messaggi e Numeri. Tutti possono avere ingressi ed uscite (chiamati rispettivamente inlets e outlets), che permettono a tutti gli elementi di collegarsi tra loro tramite cavi che trasportano segnali o dati. Gli Oggetti sono dei blocchi in grado di elaborare i dati in ingresso e spedirli in uscita. I Messaggi sono di tipo alfanumerico e rappresentano dei dati che vengono inviati agli oggetti. Infine i Numeri sono solo costituiti da cifre e possono entrare e uscire dai blocchi.

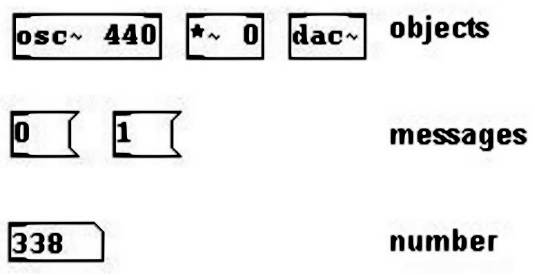

Figura 9: *Elementi di Pure Data.*

Un documento in Pure Data viene chiamato Patch, il quale è il modo vecchio di definire i sintetizzatori costruiti tramite unità modulari. Ogni oggetto durante l'esecuzione mantiene uno stato, che può cambiare di volta in volta, restando inattivo sino a quando arriva un dato in ingresso. Nel momento in cui viene mandata in esecuzione una Patch, la logica seguita dall'interprete di Pure Data, per quanto riguarda l'esecuzione delle operazioni, è quella di processare i vari rami da destra verso sinistra e di andare sempre in profondità.

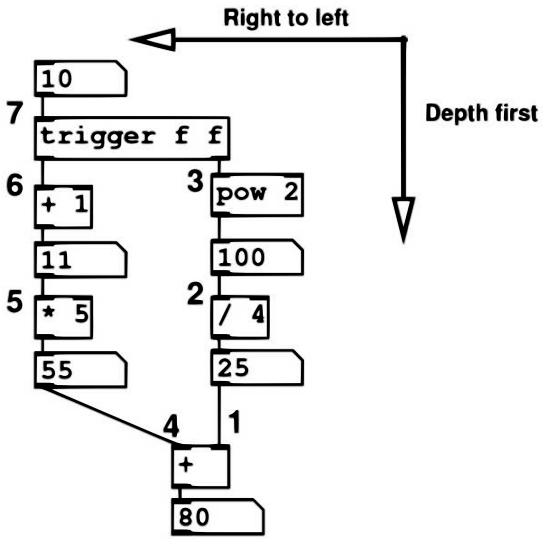

Figura 10: *Modo di processare dell'interprete di Pure Data.*

3.2.3 Protocollo OSC

Open Sound Control è un protocollo di comunicazione ottimizzato per la moderna tecnologia di interconnessione tra computer, sintetizzatori e altri dispositivi [13]. I vantaggi derivati dall'uso di OSC sono l'interoperabilità, la precisione e il miglioramento della documentazione e l'organizzazione, fornendo la gestione di audio e multimedia in tempo reale.

Come protocollo si struttura in messaggi, e presenta 5 tipi di dati:

- Int32: intero in complemento a 2 in 32 bit.
- OSC-timetag: etichetta temporale in virgola fissa a 64 bit.
- Float32: numero a virgola mobile a 32 bit.
- OSC-string: sequenza di caratteri ASCII non nulli, seguiti da uno o più caratteri nulli (massimo 4) per arrivare a un numero di bit multiplo di 32.
- OSC-blob: un contatore a 32-bit, seguito da diversi byte di dati e da un numero massimo di 3 bit nulli, in modo da formare un'array binario di lunghezza multipla di 32 bit.

L'unità di trasmissione di questo protocollo è il pacchetto OSC. Se un applicazione invia questi pacchetti viene definita OSC Client, mentre se li riceve viene definita OSC Server. Un pacchetto è costituito da un blocco contiguo di dati binari e dalla dimensione, espressa con un numero di 8 bit, il quale indica di quanti byte è composto. Si possono utilizzare sia il protocollo UDP, sia TCP, ma nel secondo caso, essendo orientato al flusso, bisogna fornire in anticipo la dimensione di ogni pacchetto trasmesso tramite un int32.

Le unità inviate in OSC hanno due tipi di contenuti, OSC Message e OSC Bundle, i quali sono distinguibili tramite il primo byte del pacchetto inviato. Il primo tipo contiente un OSC Address Pattern, un OSC Type Tag String e uno o più OSC Arguments, mentre il secondo tipo porta con se un OSC-string "#bundle", un OSC Time Tag, uno zero o più OSC Bundle Element.

Capitolo 4

Realizzazione del software

In questo capitolo viene descritto complessivamente l'applicativo, inizialmente dando una panoramica più generale per capirne il funzionamento; in seguito vengono spiegati i due moduli principali. La seconda parte del programma, essendo la più importante, viene esposta più nel dettaglio, in quanto è la parte nella quale viene realizzata la sonificazione.

4.1 Schema generale applicativo

Come già suggerito, l'elaborato è diviso in due grandi moduli:

- Il primo che tramite la Webcam genera le immagini in movimento, all'interno delle quali si trova un volto, il quale viene tracciato, grazie a una modellazione tridimensionale
- Il secondo modulo che provvede a processare matematicamente il flusso di dati inviato dal primo, per formare delle funzioni le quali creano, gestiscono e modulano suoni.

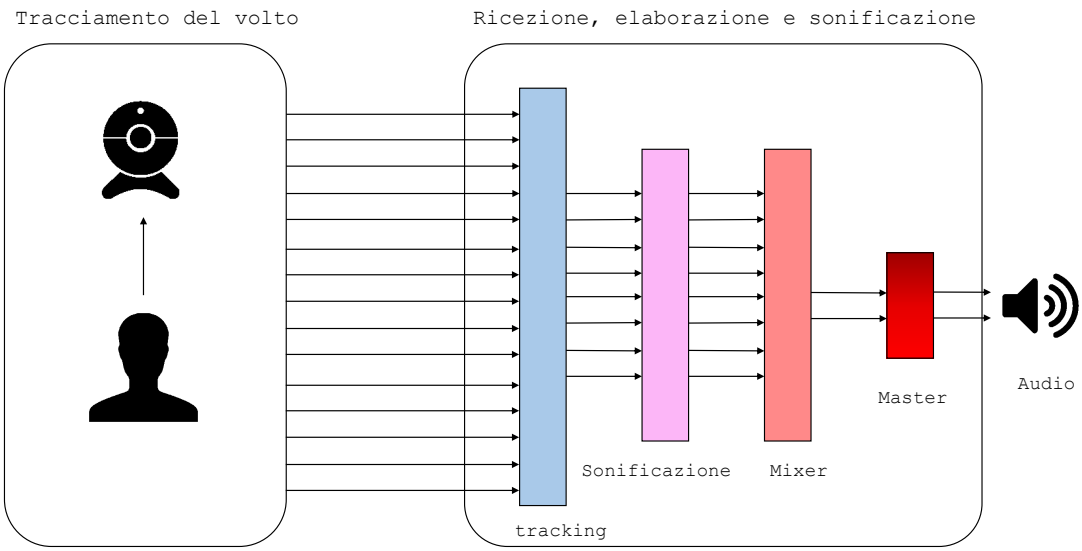

Figura 11: *Schema generale applicativo.*

Il secondo modulo è ulteriormente suddiviso in 4 parti:

- Il **Tracking**, uno per ogni gruppo di parametri (occhi, bocca e la disposizione volto all'interno dello spazio) che si occupa di definire il dominio dei dati che vengono raccolti dalla patch.
- La **Sonificazione** che rappresenta il nucleo del secondo modulo, in cui i dati si trasformano in suono.
- Il **Mixer**, all'interno del quale vengono gestiti autonomamente i livelli dei suoni creati, proprio come nelle piste del dispositivo omonimo analogico.
- Il **Master Volume**, il quale si occupa di regolare il livello del suono complessivo della sonificazione, abbassandolo completamente nel momento in cui non viene riconosciuto alcun viso e rialzandolo quando invece avviene un tracciamento.

4.2 Primo modulo: cattura dell'immagine e tracciamento del volto

Questa parte del software è stata sviluppata in gran parte da Jason Saragih, mentre il modulo che gestisce i parametri ricavati dal volto e il loro successivo invio all'interno di messaggi del protocollo OSC, è stato creato da Kyle Mcdonald. L'unico dispositivo necessario al funzionamento di questa parte del software è la Webcam, nella quale è possibile gestire parametri cruciali come l'esposizione, la saturazione, la messa a fuoco e il frame rate, particolarmente determinante per definire il numero massimo di dati trasmissibili al secondo. Tutto l'apparato è realizzato in C++, e nel momento in cui viene lanciato, ricava determinati parametri chiave del viso, individuando zone scure e ombre situate su di esso, quando una persona si posiziona davanti alla Webcam. Successivamente viene disegnato e applicato il modello tridimensionale, che a sua volta segue in un secondo momento tutti i movimenti effettuati dal volto dello stesso soggetto. Quest'ultima operazione viene realizzata grazie alla variazione di lunghezza dei lati di tutti i triangoli di cui è composta la maschera in funzione dei movimenti sottostanti.

4.2.1 FaceOSC

FaceOSC è composto da un metodo principale che si occupa di ricavare tutti i dati da ofxFaceTracker (nelle 3 dimensioni del modello), una volta riconosciuto un volto nelle immagini catturate dalla Webcam. Esso fa anche uso di altri metodi che compongono i messaggi da inviare, accompagnati da delle etichette che permettono di distinguere a destinazione da quale parte del viso provengono determinati parametri. Questi ultimi vengono tutti inseriti dentro un pacchetto OSC, il quale fa parte di un numero di altri pacchetti inviati al secondo, stabilito arbitrariamente tra i settaggi di base di FaceOSC, contenuti all'interno di un file XML.

4.3 Secondo modulo: Patch di Pure data

In questa parte del software, essendo scritta interamente in linguaggio Pure Data, vengono ripresi i concetti esposti a riguardo nel Capitolo 3, sulle tecnologie utilizzate,

seguiti da un'analisi di come sono stati ricevuti ed elaborati i dati sotto forma di flusso e infine da una descrizione delle due Sub patch principali.

4.3.1 Gestione dati in ingresso

Tutti i dati vengono creati e inviati dall'applicativo FaceOSC tramite il protocollo TCP, al fine di garantire una comunicazione orientata alla connessione. Essi sono suddivisi in più flussi, corrispondenti a determinati parametri del volto, i quali sono ricevuti dalla Patch di Pure Data e racchiusi all'interno di un dominio, di modo che i rispettivi valori non possano andare al di sopra di un certo limite e al di sotto di un altro, entrambi stabiliti registrando i minimi e i massimi raggiunti da ogni parametro testando ripetutamente l'applicativo di Face Tracking. Successivamente il flusso di informazioni viene convogliato all'interno della sub Patch Sonificazione.

4.3.2 Sub Patch Sonificazione

In questa parte del modulo di Pure Data vengono ricevuti tutti i dati, dopo essere stati delimitati entro certi confini, ma solo alcuni verranno sonificati, un quanto non tutti variano nel tempo in maniera stabile e sufficientemente palese per poter essere descritti in modo efficiente. La riproduzione della sonificazione avviene tramite dispositivo stereofonico, con l'obiettivo di presentare una descrizione più completa e immersiva. I parametri dinamici che sono stati tradotti nel dominio audio sono:

- L'altezza del sopracciglio, la quale va a modulare la frequenza centrale di un filtro passa banda che opera su un rumore bianco, generato da un oggetto specifico chiamato "noise~".
- La bocca: mentre la larghezza della stessa definisce l'intonazione fondamentale di un'onda modulata in frequenza, l'apertura verticale va a controllare sia l'indice di modulazione, sia la frequenza di taglio di 4 filtri passa basso in cascata applicati all'onda complessiva (in quanto la pendenza della curva del filtro è direttamente proporzionale al numero di filtri impiegati sullo stesso evento sonoro). Questa scelta di sintesi del suono è stata fatta per cercare di descrivere

la risonanza del cavo orale, nel momento in cui delle onde acustiche vi si riflettono all'interno. In questo frammento del software si fa uso anche di oggetti che manipolano gli "array", elementi particolari con i quali è possibile visualizzare l'onda sonora in tempo reale dentro un grafico che viene ripetutamente sovrascritto.

- La rotazione verticale della testa, descritta da un suono (un'Earcon) generato ripetutamente, il quale aumenta di velocità di riproduzione (e quindi anche di pitch) man mano che il volto si volge verso lo zenith, per suggerire il senso di elevazione. Questo effetto è ottenuto utilizzando un'onda a dente di sega, che assume valori da 0 a 44100, come un 'indice di riproduzione del campione scelto, permettendo così di gestire la rapidità di emissione del suono modulando sensibilmente il periodo del dente di sega.

- La rotazione orizzontale del viso viene comunicata anche qui da un'Earcon, la quale si attiva solo nel canale sinistro quando il volto ruota a sinistra, e il canale destro nel momento in cui ruota a destra, con un numero di riproduzioni del suono stesso nell'unità di tempo direttamente proporzionale alla rotazione.

- L'inclinazione laterale si traduce in una onda sinusoidale, la quale si disponde nell'immagine stereo in funzione del lato in cui viene individuata la pendenza del viso.

Unitamente a questi aspetti dinamici, viene descritta anche la parte del volto più statica, effettuando una istantanea nell'attimo stesso in cui viene riconosciuto per la prima volta il viso di una persona e combinando alcuni parametri scelti (bocca, mandibola e occhi), per cercare discriminare un viso dall'altro, generando un tappeto sonoro costante per l'intera durata della sonificazione di quel particolare volto. Quest'ultimo consiste in una successione di note delle quali cambia, di viso in viso, la scala musicale seguita dalle note, il loro inviluppo e il timbro (onda a dente di sega trattata con dei filtri passa basso, oppure un'onda sinusoidale).

4.3.3 Sub Patch Mixer e Master Volume

In queste due sub Patch vengono gestiti i volumi di ogni suono generato per cercare di miscelarli e rendere la sonificazione comprensibile a chi la ascolta, senza che un suono prevalga eccessivamente sugli altri. Tutti i controlli del Mixer, compreso quello del Master Volume, sono stati creati con due tipi di elementi speciali di Pure Data: lo slider e il VU meter. Mentre il primo serve a gestire graficamente il valore da inviare ad un oggetto collegato nel suo outlet, il secondo serve a visualizzare i valori in dB RMS del segnale collegato al suo inlet, esattamente come lo strumento analogico omonimo.

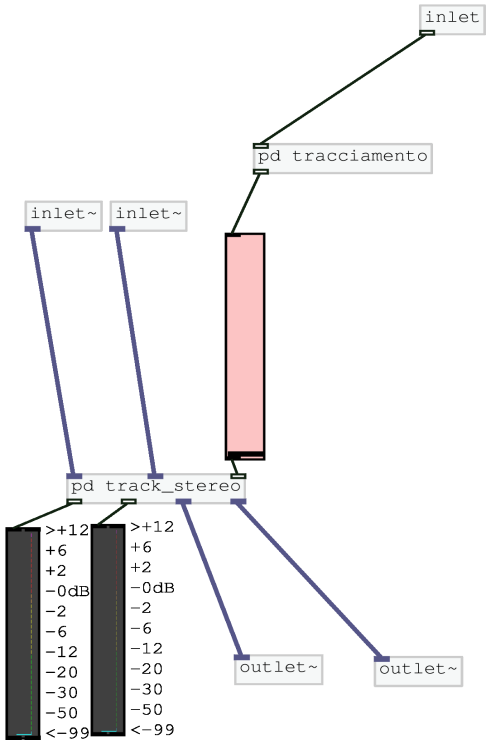

Figura 12: *all'interno della sub patch "track_stereo" vengono presi i segnali, left and right, e convertiti in dB RMS, così da poter essere correttamente interpretati dall'oggetto "VU meter".*

Attraverso un bang, ovvero un oggetto di Pure Data che funge da pulsante per inviare messaggi e numeri agli oggetti (oppure ad innescare dei processi), ogni volta che l'applicativo riconosce un volto nuovo, tutti i suoni si impostano con i volumi prestabiliti permanentemente dal software e viene generato il tappeto sonoro come spiegato sopra, con il proprio volume gestito all'interno di una singola pista del Mixer. Tramite un altro bang invece vengono portati tutti i volumi dei singoli suoni a 0, per avere un controllo sull'applicativo nel momento in cui si passa dal sonificare il viso di una persona a quello di un altra.

Il Master Volume invece controlla il volume complessivo della sonificazione, ed è il punto in cui è possibile invertire i due canali stereo di riproduzione, destro e sinistro, nel momento in cui l'uditore si trova nella collocazione speculare rispetto alla normale disposizione degli altoparlanti.

4.3.4 Collocazione della sonificazione creata all'interno del Sonification Space

L'applicativo creato può avere una collocazione particolare all'interno del Sonification Space, in quanto ogni parametro trasdotto nel dominio audio fa uso di un tipo diverso di sonificazione, oppure di una sua variazione particolare. Nel cercare di inserire tutte queste modalità di generazione sonora di ogni aspetto del volto sonificato all'interno dello spazio, è possibile identificare la posizione del software, costruendo una sorta di baricentro della figura risultante. I suoni ottenuti che fanno parte della sonificazione complessiva sono suddivisi in questo modo:

- La bocca: si rifà alla Model-Based Sonification, in quanto si comporta come un sistema puramente interattivo (rimane in silenzio quando la cavità orale rimane chiusa) e in parte alla Parameter Mapping Sonification, e per questo si può inserire nella zona di tempo continua e nella zona di astrazione percettiva.
- Il sopracciglio: esso è stato semplicemente mappato, di conseguenza ha la componente temporale continua, mentre quella di astrazione è poco più bassa di quella percettiva.

- La rotazione verticale: è un'Earcon con una componente di mappatura, proprio come la rotazione orizzontale, e per questo è possibile introdurre entrambe nella zona centrale dello spazio.
- L'inclinazione laterale: Rappresenta una tecnica di mappatura con una forte componente percettiva, collocabile nella zona in cui si trova il sopracciglio.
- Tappeto sonoro: consiste in un vero e proprio soundscape creato partendo da 3 parametri del viso, in maniera totalmente asincrona. Per questo si situa in alto a destra nello spazio.

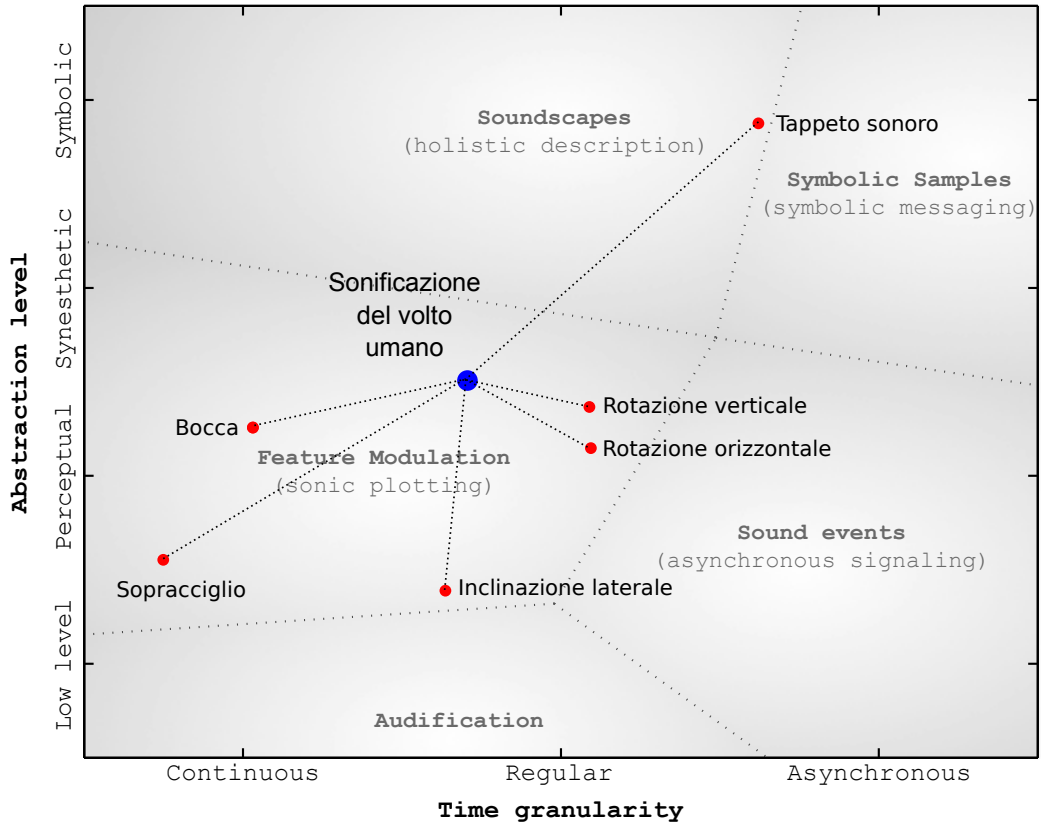

Figura 13: *Grafico della disposizione della sonificazione creata all'interno del Sonification Space.*

Capitolo 5

Esperienza al Meet Me Tonight

Nelle due giornate del 31 di settembre e 1 di ottobre 2016, in occasione dell'evento dedicato alla ricerca, allestito all'interno dei giardini di Indro Montanelli di Milano, si è potuto testare l'applicativo rendendolo disponibile al pubblico presente da provare. L'installazione comprendeva:

- Uno sgabello, sul quale la persona si sedeva avendo di fronte la Webcam sostenuta da un'asta che poggiava a terra.
- Un pc collegato alla Webcam su cui era in esecuzione il programma.
- Uno schermo ad altezza d'uomo situato alle spalle del visitatore, su cui veniva mostrata la Patch globale in Pure Data e il video catturato dalla telecamera.
- Gli altoparlanti, posti sotto lo schermo, i quali producevano la sonificazione stereofonica (invertendo i canali destro e sinistro all'interno della sub patch Master Volume, come spiegato nel Capitolo 4 sulla realizzazione del software).

Nella mattina del primo giorno il software è stato testato in maggior numero da alunni di scolaresche, le quali, soprattutto per quanto riguarda le elementari, hanno reagito in maniera molto positiva, giocando molto con l'aspetto interattivo dell'applicazione.

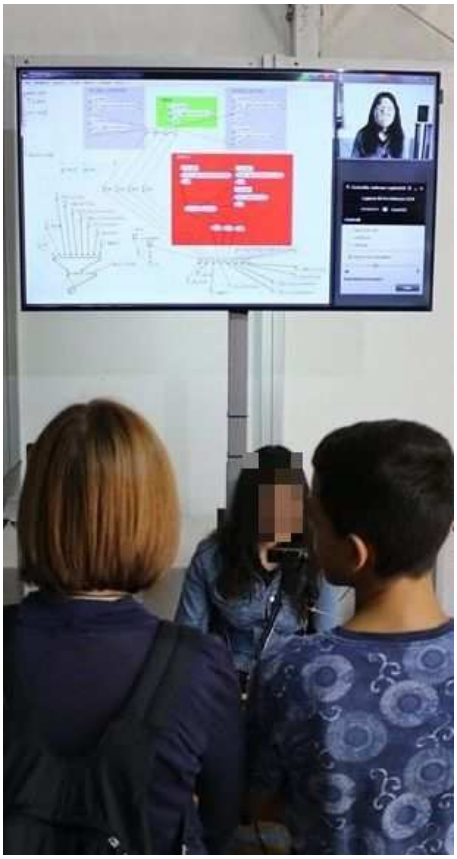

Figura 14: *Installazione dell'applicativo.*

Tra il pomeriggio del 31 e la sera dell'1 il programma è stato provato da un gran numero di persone molto diverse tra loro per età, sesso, nazionalità, titolo di studi ecc.. La quasi totalità dei soggetti è rimasta positivamente colpita dal funzionamento dell'applicativo. Una buona maggioranza, partendo dal presupposto che il risultato sonoro sarebbe stato musicalmente piacevole, è rimasta delusa, dopo aver realizzato che il suono uscito dai diffusori acustici durante il test era più orientato a una descrizione sonora convincente, piuttosto che una appagante in senso puramente estetico, ma dopo aver compreso l'obiettivo della sonificazione, hanno saputo apprezzare di più il risultato acustico dell'applicativo.

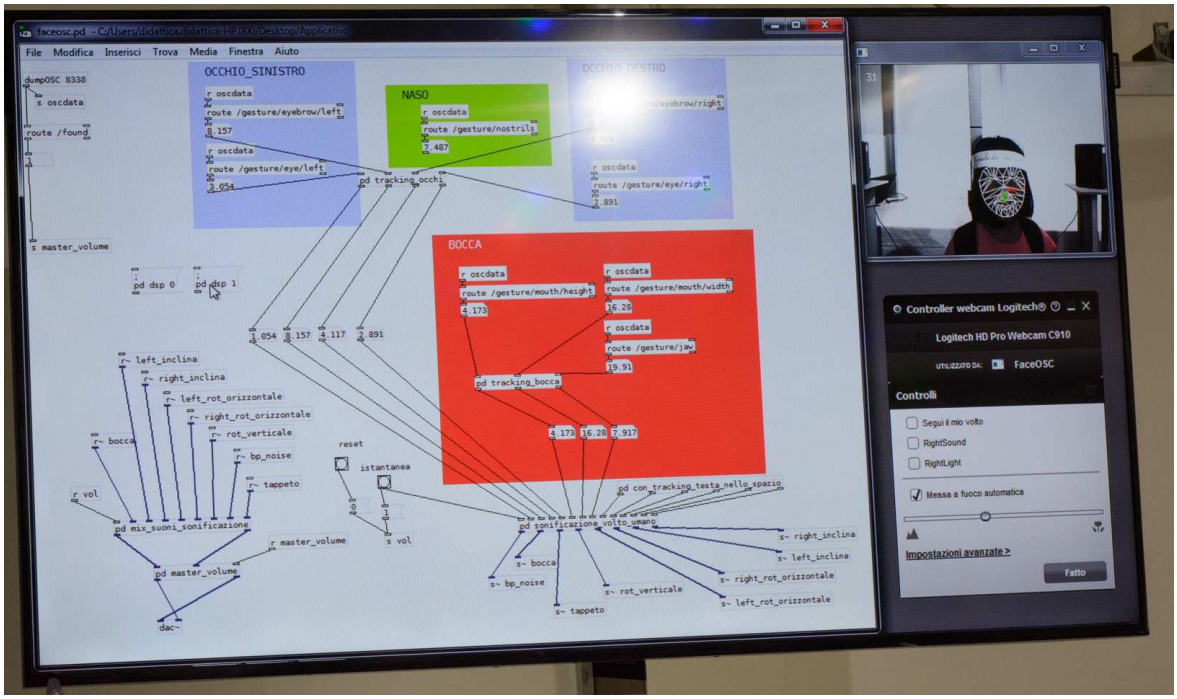

Figura 15: *Schermo installazione.*

Tutte le persone, comprese coloro le quali hanno mantenuto una certa perplessità prima e dopo aver ascoltato il suono del proprio volto o di quello altrui, si sono rivelate curiose (come ci si può aspettare in un contesto come quella della manifestazione in questione) chiedendo quali possono essere le applicazioni pratiche nella realtà di un simile lavoro, al di là di quella di ricerca accademica, rimanendo molto soddisfatte e

positivamente impressionate nell'ascoltare la risposta (contenuta nell'ultimo capitolo dell'elaborato, sulle conclusioni sviluppi futuri).

Si è potuto notare inoltre che un soggetto all'interno dello staff dello stand, verso la fine della seconda giornata, dopo un'esposizione acustica passiva ai risultati dei molteplici test condotti dal pubblico presente, per la durata dell'intera manifestazione, ha acquisito la capacità di comprendere i movimenti compiuti dal volto di una persona semplicemente ascoltando la sonificazione prodotta dal software. Ciò dimostra l'effettivo successo nell'impiego del suono come metodo di comunicazione, avvalorando la buona riuscita nell'intento dell'applicativo.

Capitolo 6

Conclusioni e sviluppi futuri

Questo capitolo analizza tutti gli aspetti che riguardano il funzionamento del software ultimato, riportando le problematiche incontrate e risolte, le possibili applicazioni nella realtà e gli sviluppi futuri, anche all'interno di un'ottica di miglioramento sostanziale del lavoro svolto.

6.1 Problematiche e risoluzione

Dal momento che l'applicativo si occupa di produrre una sonificazione utilizzando la sintesi del suono, le maggiori problematiche emerse su questo aspetto riguardano principalmente la scelta dei suoni da riprodurre (sia campionati, sia sintetizzati), che durante il corso della scrittura del programma sono stati cambiati più di una volta, in quanto i test eseguiti man mano hanno richiesto una maggior chiarezza nella comunicazione di un determinato parametro statico o dinamico.

Inoltre, nel processo di creazione del tappeto sonoro in Pure Data si è cercato di non trasdurre nel dominio audio le informazioni statiche del volto in maniera così dettagliata da farle predominare sui suoni generati dalla componente dinamica della sonificazione (e quindi l'aspetto più interattivo), in quanto l'obiettivo era quello di discriminare sensibilmente il risultato sonoro di un viso rispetto ad un altro.

Un aspetto importante riguarda il contesto in cui viene utilizzato l'applicativo, per cui i volumi scelti per ogni suono nel momento in cui si ascolta in cuffia, vengono

stravolti quando si passa a un ambiente all'interno del quale si fa uso di altoparlanti (come ad esempio è accaduto in occasione dell'evento Meet Me Tonight), con tutti le caratteristiche acustiche del luogo e fonti di disturbo.

6.2 Applicazioni nella realtà e sviluppi futuri

Una possibile applicazione nella realtà del lavoro svolto è quella di descrivere il volto umano ad una persona non vedente attraverso il suono, affinchè possa avere un'immagine mentale di come è fatto il viso di un soggetto all'interno di una foto, video o presente nello stesso luogo.

Più in generale, migliorando notevolmente la quantità e la qualità delle informazioni ricavate dal volto umano, l'applicativo può essere sfruttato in un contesto di identificazione; in un contesto di interazione con una macchina in movimento da pilotare, come ad esempio un aereo o un'automobile, al fine di agevolarne il compito, oppure di trasferire il controllo dagli arti superiori e inferiori alla testa, rendendo questi ultimi disponibili per effettuare altre operazioni secondarie. Nel portare tutti i suoni in un sistema musicalmente coerente è possibile utilizzare il viso come strumento musicale in grado di generare anche delle polifonie più o meno modulabili a seconda del livello di dettaglio del sistema; inoltre è possibile aumentare il numero di informazioni a livello emotivo che già traspaiono dal volto, sia in un attimo, sia in un intervallo di tempo arbitrario, a seconda di quanto si vuole descrivere dello stato emozionale di una persona.

Per un possibile sviluppo futuro, la prospettiva più interessante è quella di inserire il software all'interno delle tecnologie assistive per ipovedenti, in quanto esistono già delle start up che stanno lavorando a dei supporti hardware i quali integrano anche dei sensori capaci di riportare fedelmente la realtà circostante (paesaggi, oggetti, testi scritti ecc.) attraverso il suono, di modo da poter dare più mezzi possibili per rendere autonoma una persona disabile. Il risultato è ottenibile grazie a un training uditivo sostenuto dal soggetto non vedente, così da poter percepire agevolmente, tramite la sonificazione, ogni informazione veicolata dal dispositivo.

Bibliografia

[1] Thomas Hermann, Andy Hunt e John G Neuhoff. *The sonification handbook*. Logos Verlag Berlin, 2011.

[2] Luca A Ludovico e Giorgio Presti. «The sonification space: A reference system for sonification tasks». In: *International Journal of Human-Computer Studies* 85 (2016), pp. 72–77.

[3] Thomas Hermann. «Taxonomy and definitions for sonification and auditory display». In: (2008).

[4] Michele Mengucci, Francisco Medeiros e Miguel Amaral. «Image Sonification Application to Art and Performance». In: *proceedings of INTER-FACE International Conference on Live Interfaces, Lisbon*. 2014.

[5] Woon Seung Yeo e Jonathan Berger. «Application of image sonification methods to music». In: *Online document* (2005).

[6] Matteo Spadaro. «Sonificazione di dipinti e multimedialità in una pubblicazione EPUB3». Short Degree. Università degli studi di Milano, 2015.

[7] Tim Pohle e Peter Knees. «Real-Time Synaesthetic Sonification of Traveling Landscapes». In: *5TH INTERNATIONAL MOBILE MUSIC WORKSHOP 2008 13-15 MAy 2008, VIENNA, AUSTRIA*. 2008, p. 23.

[8] Alexander Refsum Jensenius. «Motion-sound interaction using sonification based on motiongrams». In: (2012).

[9] Nenad Markuš. «Overview of algorithms for face detection and tracking». In: ().

[10] Kyle Mcdonald. *https://github.com/kylemcdonald/ofxFaceTracker*. 2009.

[11] Nižnij Novgorod. *http://opencv.org/about.html*. 2016.

[12] Andy Farnell. *Designing sound*. Mit Press, 2010.

[13] Adrian Freed e Matt Wright. *http://opensoundcontrol.org/*. 2016.