MX-Ray: prototipizzazione

Giovanni Donati

27/05/2010

Contents

Chapter 1

Presentazione del progetto MX-Ray

1.1 Introduzione

Il progetto MX-Ray nasce dall'individuazione di un problema generato dalla prolificazione di reti sociali a sfondo musicale.
Tali reti sociali solitamente raccolgono grandi database di musica composti principalmente da artisti ben affermati sulla scena musicale ed in parte anche da musicisti appartenenti ad una scena underground ben poco conosciuta e dunque non in grado di possedere la dovuta visibilità da parte degli utenti della community.
Più in generale una social network contiene normalmente un enorme database di canzoni in formato digitale compresso, solitamente in mp3, rese accessibili all'ascolto del pubblico attraverso un interfaccia di riproduzione online simile in tutto ad un normale player come Winamp, Windows Media Player o VLC. Quando un utente accede al portale, interagisce con un interfaccia grafica composta da un campo di ricerca all'interno del quale egli può inserire il nome dell'artista che desidera ascoltare. Il risultato della ricerca è l'esecuzione di un brano dell'artista richiesto, al quale viene affiancato un campo con risultati alternativi alla ricerca effettuata, ossia i collegamenti ai brani di altre band simili a quella ricercata dall'utente e che potenzialmente esso possa gradire. A questo scopo si è reso necessario, con l'evoluzione di tali servizi web, classificare e catalogare la musica, generando delle etichette per i brani in modo da stabilirne il genere di appartenenza e quindi di poter creare delle associazioni tra i brani in modo da rendere possibile i feedback per gli utenti.
L'operazione di associazione di un etichetta ad un brano musicale è detta tagging, dall'inglese tag che significa bersaglio. Di fatto la tag identifica il gruppo di appartenenza di un brano all'interno di una vasta collezione, ed è quella che permette ad un brano di acquisire maggiore visibilità a seguito della ricerca da parte dell'utente.
Molti portali online ad oggi hanno utilizzato tecniche basate sul tagging manuale da parte degli utenti stessi, fiduciosi del fatto che gli utenti operassero debitamente in modo da fornire piu visibilità alle band emergenti iscritte come artista sul portale.

Questa supposizione si è rivelata errata.
Gli utenti normalmente accedono al portale per ascoltare le band più conosciute, e pur entrando in contatto con i contenuti musicali di nicchia non si soffermano a taggarli dopo l'ascolto, lasciandoli al di fuori dello spazio di visibilità offerto dal sistema di tagging.
Un esempio di portale di questo tipo, è LastFM.com che si era prefissato target, quello di promuovere la musica delle band emergenti, utilizzando la folksonomia come metodo di generazione delle tag. La folksonmia come già accennato prevede che gli utenti dopo l'ascolto forniscano una tag al brano.
Un altro metodo di operare è attraverso la tassonomia, ovvero l'artista al momento della pubblicazione del proprio contenuto musicale, fornisce una tag al suo brano, associandovi una o più parole chiave. Anche questo metodo è risultato poco efficace per raggiungere lo scopo voluto poichè gli artisti pur di acquisire visibilità elevata tra i risutalti di ricerca, tendono a generare tag copiose e completamente scorrelate dal proprio genere di appartenenza, con il risultato che acquisiscono si visibilità, ma da parte anche di utenti non assolutamente interessati ad ascoltarli.
Un progetto che ha cercato di risolvere questo problema è stato "The Music Genome Project", integrato poi nella web radio chiamata Pandora. Essenzialmente esso utilizza la qualità umana più efficiente per il riconoscimento di un genere musicale e la generazione di una tag, ossia l'ascolto. La classificazione avviene attraverso una ascolto del brano da parte di diversi esperti in campo musicale, durante alcune sedute che della durata stimata di circa 20 minuti di analisi per brano.
Sono state stabilite alcune caratteristiche distintive di un genere musicale come ad esempio il sesso del cantante solista, il numero di chitarre, la presenza o meno di suoni campionati e molte altre. Tali caratteristiche sono state nominate "geni" ed ammontano ad un numero di 150 per brani musicali in generale, mentre per brani rap/hip hop ammontano ad un numero nettamente superiore, circa 300, per via della presenza dei particolari suoni campionati tipici di questo genere.
Il sistema adottato da music genome project per effettuare poi la classificazione ed il riconoscimento è una rete neurale allenata in modo da poter prendere decisioni sui vettori di feature estratti manualmente.
Tuttavia, come si può facilmente intuire, è necessario un intervento umano per completare il funzionamento del sistema che richiede un certo tempo per l'analisi dei brani ed inoltre rischia di divenire soggettivo e quindi affetto da caratteristiche decisionali prettamente umane e quindi influenzabili in base al soggetto.
Da qui nasce l'idea di creare un sistema di tagging automatizzato ed in grado di ottenere comunque una buona efficienza, pur approssimando il comportamento del sistema orecchio-cervello di un essere umano. Ciò può essere reso possibile sfruttando l'analisi temporale e frequenziale di segnali audio allo scopo di estrarne caratteristiche utili alla catalogazione e classificazione di contenuti musicali, in modo da poter generare una tag completamente oggettiva e priva di interventi umani che richiedano risorse e tempo per essere completati.

1.2 Idea alla base di MX-Ray

L'orecchio umano è uno strumento molto complicato che interagisce con il cervello per poter effettuare complicate operazioni di analisi sia frequenziale che temporale su segnali audio. A livello base ciò che avviene all'interno del nostro cervello è un operazione di scomposizione del suono in diverse sottobande di frequenza (o ancora meglio in singole armoniche) e riesce dunque ad estrapolare qualsiasi caratteristica sonora in maniera molto precisa.
Ad esempio riesce immediatamente a capire il genere del brano, quanti strumenti suonano e quali, il tipo di spazio in cui è avvenuta l'esecuzione e molto altro. Questo è difficile per una macchina programmata, ma è possibile comunque, a partire dal file audio compresso, analizzare il segnale digitale risultante ed estrarre dei parametri informativi in grado di posizionare il brano all'interno di uno spazio multidimensionale in una ben nota posizione.
Per fare questo è necessario tener conto del fatto che il nostro cervello non basa il proprio funzionamento esclusivamente su fenomeni fisici, ma anche sulla propria esperienza passata e su fenomeni psicoacustici non facilmente riproducibili e tutt'ora inesplorati.
Il signal processing rimane comunque una buona base di partenza perchè può fornirci strumenti necessari ad estrarre diversi tipi di informazione dai file musicali e dunque avere una base per poter istruire una macchina ad effettuare determinati tipi di decisione.
Il signal processing permettere dunque come abbiamo detto di effettuare le operazioni di estrazione dei parametri decisionali, la discriminazione può poi essere fatta attraverso l'utilizzo di una rete neurale, ossia un programma "'allenato"' a pensare come un cervello, che effettua connessioni logiche tra i diversi parametri estratti e fornisce in output una decisione.

1.3 Perchè MX-Ray

Il nome del progetto è nato dalla natura alla base del medesimo. "'The music genome project"' fu chiamato così poichè cercava di classificare la musica fino agli aspetti più profondi della sua natura, ossia fino alla radice di quello che è stato definito il suo DNA. Di fatto tutto le caratteristiche estratte attraverso gli ascolti nell'ambito di tale progetto sono chiamate "'geni"'.
MX-Ray si pone un obiettivo di classificazione simile, ma automatizzato, ossia attraverso una macchina. MX-Ray funzionerà come una specie di macchina a raggi-X in grado di penetrare la "'superficie"' dei segnali audio e musicali, estraendone alcune caratteristiche sonore automaticamente, ed utilizzandole per etichettarli. In MX-Ray i parametri distintivi estratti saranno chiamati "'bones"' ovvero ossa.

1.4 Procedura dello sviluppo e piano di intervento

Il progetto è basato sullo svolgimento di cinque attività principali, ognuna delle quali è suddivisa in un certo numero di sotto-attività. Il diagramma di Gantt

fornito in allegato al piano d'intervento mostra lo svolgimento indicativo del piano d'intervento e le durate preventivate per ogni task.

Chapter 2

Attività introduttive

2.1 Considerazioni iniziali e consultazione letteratura

La prima attività del piano di intervento per la realizzazione del progetto MX-Ray è stata una fase di accrescimento delle competenze allo scopo di individuare le tecnologie più idonee a sviluppare il componente di analisi frequenziale per l'estrazione delle caratteristiche distintive dei brani musicali. Tale attività ha occupato all'incirca le prime venti giornate di lavoro, partendo da una consultazione dei lavori già presenti in letteratura in ambito di classificazione musicale, signal processing ed estrazione di caratteristiche da contenuti multimediali.
Durante tale fase sono state individuate diverse tecniche presentate in lavori scientifici antecedenti MX-Ray. Nell'ambito di questa parte del progetto, sono stati di particolare rilievo i lavori di George Tzanetakis, professore associato di Computer Scienze presso l'Università di Victoria in Canada, e Karin Kosina, dottoranda presso la Technische Universität di Vienna.
George Tanetzakis è stato uno dei pionieri nel campo della classificazione musicale automatizzata ed è autore di diversi paper scientifici. Karin Kosina è stata invece autrice di un progetto open source per la catalogazione automatica della musica, oggetto della sua tesi di laurea nell'anno 2002.
Altri lavori in campo scientifico sono stati presi in considerazione, sia per arricchire il mio bagaglio tecnico/scientifico, sia per comprendere quali tecnologie applicare in fase di progetto e sviluppo del sistema.
Per prima cosa verrà introdotto il funzionamento dei fenomeni ondulatori coinvolti nella produzione dei suoni, e i meccanismi legati all'ascolto da parte degli essere umani. Dopo di che verranno spiegate alcune tecniche base utilizzate nel signal processing per analizzare e scomporre segnali multimediali, come segnali audio, ed in fine saranno esposti i risultati più interessanti per quanto riguarda la ricerca effettuata in letteratura, e di cui si terrà conto nello sviluppo di MX-Ray.

2.2 Elementi di base per la comprensione del suono

L'attività umana dell'ascolto è fortemente influenzata non solo dall'analisi frequenziale che il sistema orecchio-cervello svolge automaticamente, ma anche dal modo di decodificare le informazioni che il nostro cervello adotta.
I suoni non provocano solamente eccitazioni meccaniche ma agiscono anche sul subconscio della persona creando sensazioni che hanno un loro peso sulla valutazione dei suoni. I meccanismi appena citati fanno si che i giudizi su un singolo brano musicale risultino soggettivi e dunque differenti da ascoltatore ad ascoltatore.
Il suono è un fenomeno fisico studiato da una branca della fisica chiamata acustica. Essa studia il suono in tutte le sue caratteristiche specificandone i meccanismi di creazione, propagazione e ricezione da parte dell'orecchio umano. Si tratta di onde di pressione acustica, determinate da cambiamenti di quest'ultima attraverso meccanismi di compressione e decompressione dell'aria, ossia da variazioni della pressione atmosferica rispetto ad una condizione di equilibrio.
La conoscenza della fisica del suono da sola non basta quando si vuole operare su una produzione musicale, poiché bisogna tenere conto di una vasta molteplicità di fattori che esulano dalla semplice natura del suono, ma anche dal funzionamento di apparecchiature elettroniche per l'elaborazione, dalle caratteristiche degli strumenti musicali in gioco ed in generale da ciò che fa parte del corredo di dispositivi presenti in uno studio di registrazione.
Tutto questo viene completato da un altro importante settore della scienza, ossia la psicoacustica.
Quest'ultima si occupa dello studio della percezione soggettiva umana dei suoni, o meglio è una sorta di studio sulla psicologia legata alla nostra percezione acustica.
In fase di produzione, la conoscenza della fisica del suono e dell'elettronica per il sound editing e il processing sono molto importanti ai fini di poter dare una certa forma ai brani, ma questo non basta, poiché bisogna tenere ben conto di come l'utenza percepisce il contenuto prodotto.
Ad esempio, incidendo una chitarra in mono (ossia su un unico canale audio centrale), attraverso l'editing è possibile crearne un immagine stereo, prelevando tale traccia, raddoppiandola su due canali mono e "pannandola" a destra su un canale ed a sinistra sull'altro canale (aggiungendo anche un ritardo di pochi millisecondi tra le due).
Questo crea nell'ascoltatore una sensazione stereofonica.
Nell'arco di questo processo sono coinvolti importanti fenomeni tra i quali: la riflessione dei suoni, l'anatomia dell'orecchio umano e l'interpretazione del cervello di questi fenomeni.
Volendo realizzare un sistema in grado di catalogare la musica, è necessario tenere conto di tutti questi aspetti, per cercare di approssimare il comportamento di un utente umano nel riconoscimento di diversi fattori della musica digitalizzata.
Ossia bisogna cercare di individuare quali siano le caratteristiche del comportamento umano e degli elementi di fisica, sound engineering e psicoacustica da utilizzare per poter implementare un software in grado di fare valutazioni simili all'essere umano ma in modo automatico.

Per fare questo ci si vuole avvalere di aspetti legati all'analisi frequenziale e temporale dei brani musicali ed utilzzare il signal processing per stabilire dei metodi in grado di estrarre determinate caratteristiche della musica ed effettuare la catalogazione attraverso determinati criteri dipendenti da tali caratteristiche.
Pensiamo innanzitutto a quello che accade tra l'incisione di un contenuto musicale, ed il nostro orecchio. Una produzione musicale giunge al nostro orecchio dopo un processo riassumibile attraverso lo schema in figura 2.1.
Quello che giunge a noi in forma di CD audio è in realtà il risultato di un lungo processo produttivo. Lo schema proposto ci rende chiaro quello che accade tra il momento in cui gli strumenti e le voci vengono incisi e il momento in cui ci troviamo comodamente sulla nostra poltrona a goderci il prodotto finale.
In partenza gli artisti si recano presso lo studio di registrazione, dove attraverso dei microfoni di alta qualità incidono gli strumenti.
In questa fase è molto importante il know how del sound engineer che effettua le riprese, poichè dalla scelta dei microfoni e delle tecniche di microfonaggio dipende molto la qualità del suono acquisito e dunque del prodotto finale.
I suoni vengono trasferiti attraverso schede audio apposite ad un interfaccia di memorizzazione. Ad oggi tale interfaccia è rappresentata dall'hard disk di un PC, che attualmente è lo strumento chiave utilizzato negli studi per effettuare l'editing dei suoni attraverso dei software.
Da questa working station il sound engineer sistema i risultati delle registrazioni, tagliando parti errate o eseguite male, copiando ed incollando le riprese migliori, mettendo a tempo la batteria ed applicando poi gli effetti desiderati a strumenti e voci. Queste sono quelle che vengono chiamate fasi di editing e missaggio.
A questo punto i brani vengono ascoltati attraverso i monitor da studio, che rendono possibile valutare la qualità del risultato ottenuto ed effettuare eventuali attività di ritocco dei suoni prima di esportare quello che viene chiamato "'master"'.
Il master rappresenta l'ultimo step di una produzione, e serve a rendere il prodotto ascoltabile in modo piacevole su qualsiasi impianto, e pronto per il mercato musicale.
A questo punto vi è una linea di demarcazione tra quello che è il lavoro effettuato in studio, e quello che ascoltiamo attraverso il nostro impianto casalingo, o comunque attraverso il nostro riproduttore musicale.
L'apparato utilizzato per la riproduzione dei suoni influisce molto sulla nostra percezione acustica, in quanto non tutti gli impianti audio posseggono la stessa qualità, non tutte le cuffie o casse sono ottimizzate per esaltare le frequenze dei brani, e sopratutto l'ambiente in cui si effettua l'ascolto varia le caratteristiche dei suoni che giungono a noi.
Non bisogna dimenticare che in tutto questo processo, gli ambienti in cui vengono effettuati gli ascolti e le registrazioni influenzano da un lato l'orecchio del sound engineer che in base ai suoni percepiti produce determinate azioni sul prodotto, e dall'altro lato l'ascoltatore che sperimenta diverse sensazioni in base al riverbero presente attorno a lui.
La parte terminale di questa catena è l'orecchio dell'ascoltatore. Da qui si può partire per cercare di capire come sentiamo, e come ascoltiamo la musica (che sono due concetti ben distinti) ed individuare i principali fenomeni coinvolti nel processo di analisi.
Questo per rendere possibile la modellazione del sistema in base al comportamento umano, e giungere ad una sua realizzazione.

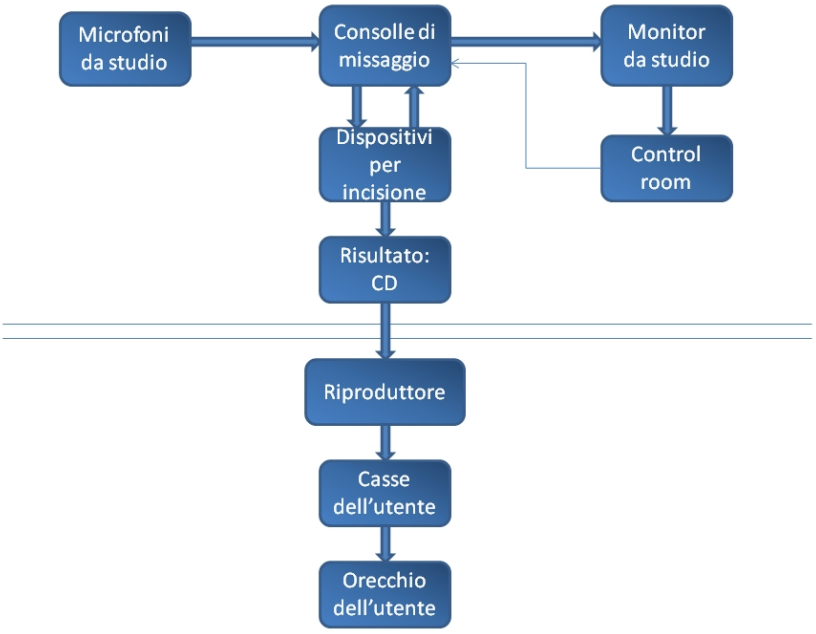

Figure 2.1: Catena di creazione di un CD musicale.

2.2.1 Fisica del suono

Il suono è un fenomeno fisico di tipo ondulatorio. In questa sottosezione di farà chiarezza su alcuni termini che verranno utilizzati nel seguito della trattazione. Il suono non è altro che una variazione periodica della pressione atmosferica, rispetto ad una condizione di equilibrio.
Le variazioni periodiche a cui si è accennato sono dette "'compressione"' e "'rarefazione"' dell'aria. Ad esempio se prendiamo un palloncino e lo facciamo scoppiare, inizialmente la pressione dell'aria all'interno di esso era maggiore di quella posseduta dall'aria al di fuori di esso.
In seguito allo scoppio le molecole d'aria ad alta pressione prima contenute all'interno del palloncino tendono a muoversi nello spazio circostante, sospinte dalla forza generata dallo sbalzo di pressione creatosi, ovvero si spostano verso la zona di bassa pressione circostante. Le molecole d'aria vicine vengono colpite da quelle che prima possedevano alta pressione ed acquisiscono l'energia sprigionata, e mentre nella zona di spazio adiacente l'aria di decomprime, o rarefà, nelle porzioni di spazio successive si comprime.
Il fenomeno si ripete uguale fintanto che tutta l'energia viene dispersa a causa dell'attrito e dissipata in parte in potenza acustica ed in parte in calore per effetto Joule.
Le masse d'aria spostate dunque sono soggette a un continuo passaggio tra gli stati di compressione e decompressione con una certa velocità. Tale velocità è la frequenza del suono generato, e siccome stiamo parlando di fenomeni periodici se ne evince che l'inverso di tale grandezza è il periodo, ovvero la durata temporale di uno dei cicli di compressione-decompressione.

A questo punto è meglio fare una precisazione dovuta, ovvero bisogna tenere bene a mente che non sono le molecole d'aria a spostarsi alla velocità del suono, ma solamente l'onda sonora si muove continuamente attraverso l'atmosfera come onda di di compressione ad alta pressione che continua a spingere verso zone di pressione piu bassa. Questo è il fenomeno con cui si propaga l'onda. La figura 2.2 mostra un esempio visivo del concetto appena esposto. E' possibile inoltre vedere cosa si intende per lunghezza d'onda (wavelength, in Inglese) ed ampiezza del suono (amplitude).

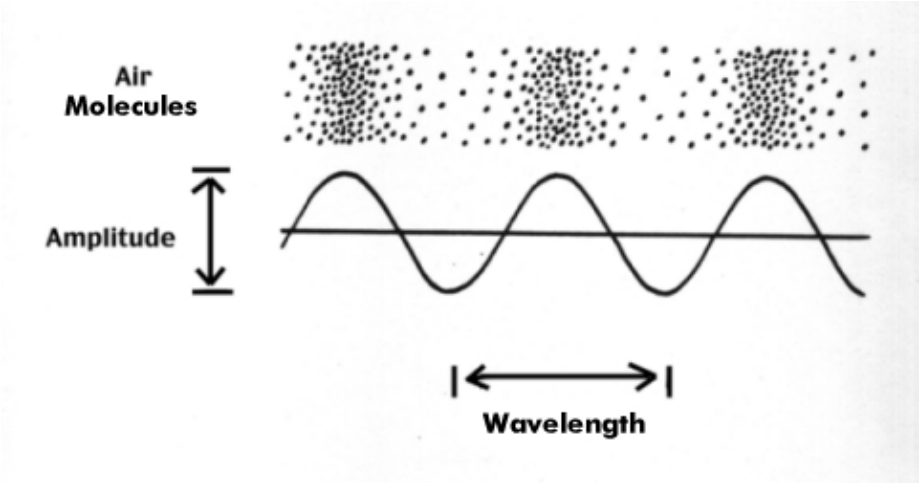

Figure 2.2: Andamento della pressione acustica, e periodo.

La forma d'onda di un onda sonora ha altre caratteristiche che verranno elencate sinteticamente per completezza:

- Fase: E' il punto in cui un onda sonora inizia, espresso come angolo in radianti.
- Contenuto armonico: sono le frequenze fondamentali contenute in un suono, e ci permettono di distinguere ad esempio uno strumento da un altro.
- Inviluppo: I suoni sono composti da tre fasi, chiamate: attacco (attack), decadimento (decay), mantenimento (sustain) e rilascio (release). Un possibile inviluppo è mostrato in figura 2.3. L'attacco è dato dal tempo che passa tra l'inizio del suono e il momento in cui raggiunge il suo massimo valore di volume. Il decadimento è la fase successiva in cui il volume scende fino ad un certo livello. Il mantenimento è il punto tra la fine del decadimento e l'ultimo istante in cui la nota rimane di volume pressochè costante. Per ultima abbiamo la fase di rilascio, in cui il suono decade progressivamente di volume fino a smorzarsi completamente.
- Velocità: è la velocità con cui l'onda si muove nello spazio. Circa 344 metri/secondo attraverso l'aria quando la temperatura è circa 20 gradi centigradi. Da questo evince che la velocità del suono nell'aria è influenzata anche dalla temperatura di quest'ultima.

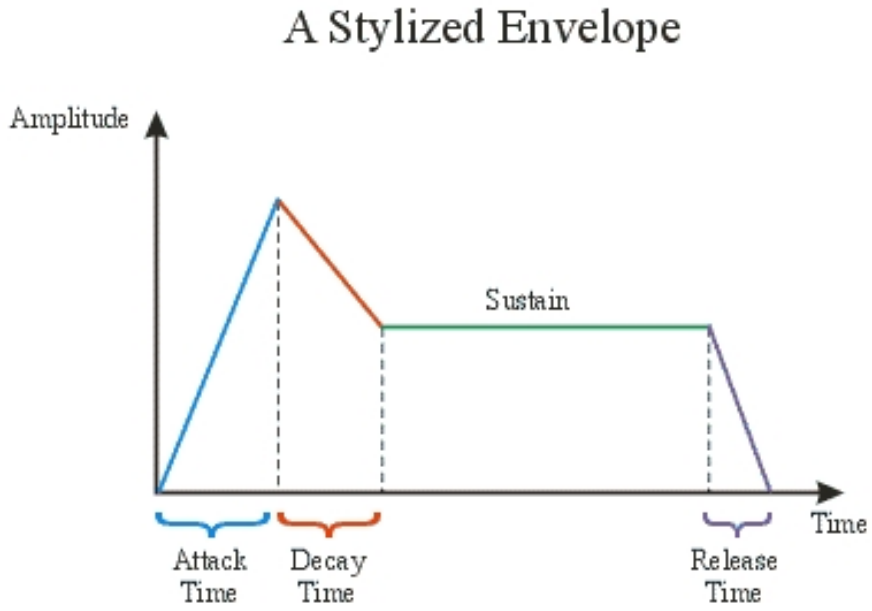

Figure 2.3: Inviluppo di un suono.

2.2.2 L'orecchio umano

Dopo aver introdotto alcune basi sulla natura dei suoni, viene fornita una breve spiegazione sul funzionamento di base di funzionamento dei nostri recettori acustici. In figura 2.4 sono mostrati l'orecchio esterno, medio ed interno.
L'orecchio umano non è altro che un trasduttore che converte onde di pressione acustica in una serie si segnali fisiologici interpretabili dal cervello. L'ascotlatore riceve onde di pressione che vengono raccolte dal padiglione auricolare (la cui dimensione influenza la nostra percezione dei suoni, piu è grande e meglio possiamo raccogliere il suono). Da quì attraverso il canale uditivo esterno, la sollecitazione acustica va ad interagire con tre piccole ossa chiamate martello, incudine e staffa, tre termini che ne ricordano la forma.
Questo sistema di ossa viene messo in vibrazione ed amplifica in qualche modo la sollecitazione trasferenso le vibrazioni acustiche alla coclea.
Quest'ultimo è un organo molto importante in quanto dalla sua forma si può comprendere la sensibilità dell'orecchio umano ad un vasto range di frequenze. Di fatto la coclea è una sorta di cono organico arrotolato su se stesso. Ad ogni sezione corrisponde quindi un diametro differente, e dunque diverse lunghezze d'onda vado ad abitarne la cavità.
Se immaginiamo questo organo srotolato, esso avrà una certa lunghezza, e l'imboccatura sarà quella con il diametro maggiore. Dato che l'imboccatura della coclea è più vicina all'orecchio esterno, sicuramente essa collezionerà maggior energia nella sua parte iniziale, e questa andrà in decadimento con la lunghezza dell'organo. Questo spiega come mai l'orecchio umano è più sensibile alla basse frequenze, e meno alle alte.
La coclea attraverso nervi acustici coverte vibrazioni meccaniche in impulsi elet-

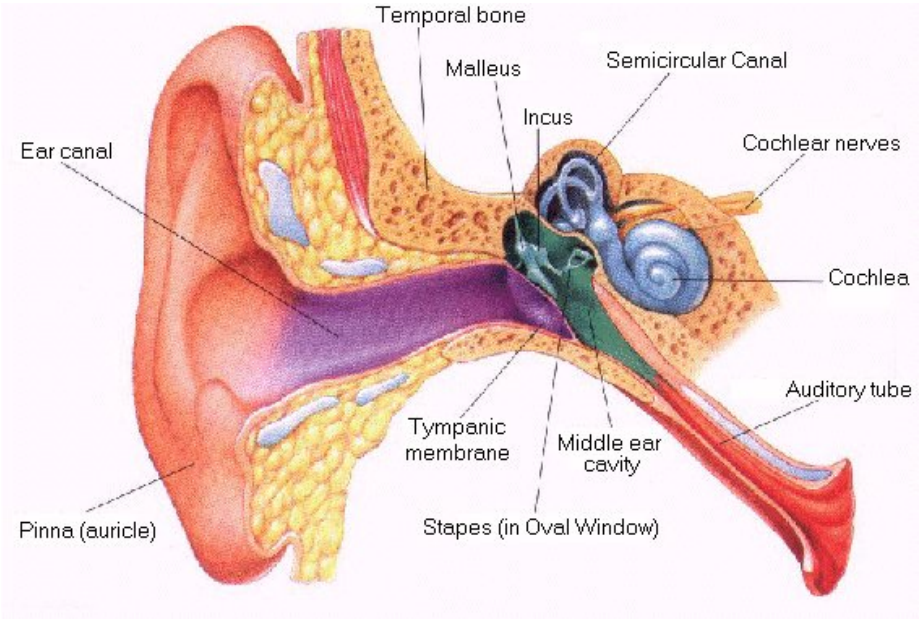

Figure 2.4: Orecchio umano.

trici interpretabili dal cervello.

2.2.3 Psicoacustica

Come anticipato, i fenomeni fisici che producono in noi sensazioni udibili sono perfettamente conosciuti poichè dipendenti da fattori fisico-fisiologici ben noti attraverso scienze come la fisica e la medicina che ci spiegano perfettamente quello che accade nel nostro sistema uditivo.
Quello che tutt'ora rimane inesplorato, o comunque poco conosciuto a livello scientifico è ciò che accade tra il momento in cui gli impulsi sonori ricevuti vengono trasformati in sensazioni coscienti.
In psicoacustica si distinguono due aspetti principali della materia:

- la capacità dell'udito che permette la valutazione fisica del suono
- la capacità di valutarne le variazioni.

Si rende dunque necessario ed importante definire il concetto di soglia.
Per soglia si intende il livello minimo di pressione acustica che il nostro orecchio può percepire. in psicoacustica non si parla di soglia assoluta, ma di soglia differenziale in quanto secondo la legge di Weber, è necessario un incremento dello stimolo che sia pari a una frazione costante del medesimo per poter produrre una percepibilità accettabile.
La non esistenza di un vero "'orecchio assoluto"' per altro suggerisce il motivo per cui quando si eseguono test acustici, i risultati forniti vengono sempre relazionati ad un insieme di persone, ovvero con significato statistico.
Diversi test in passato sono stati condotti per ottenere la soglia di udibilità media. A livello base l'esperimento per ottenere la risposta in frequenza dell'orecchio

umano è molto semplice. Si utilizzano toni puri, ossia suoni contenenti un unica frequenza ed emessi con una certa potenza acustica. Ad esempio prendendo un tono a 5000 Hz e riproducendolo a volumi sempre piu crescenti, è possibile verificare a quale livello di potenza acustica corrisponde, per quella determinata frequenza, corrisponde un'eccitazione udibile mediamente da un essere umano. La stessa cosa è stata effettuata su tutte le frequenze dello spettro udibile (circa da 20 Hz a 20000 Hz), e graficando su un piano cartesiano i risultati ottenuti con questo metodo sono state ottenute le cosidette curve isofoniche, e quindi il campo di udibilità mostrato in figura 2.5.

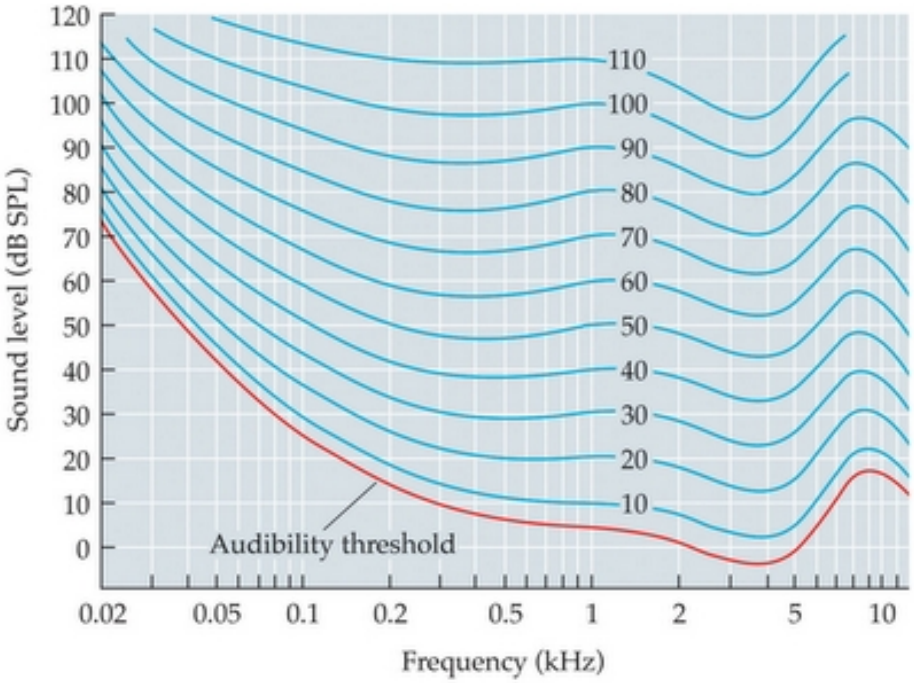

Figure 2.5: Curve isofoniche.

Le curve rappresentano la soglia di udibilità per tutto il campo di frequenze udibili, con toni di livello di pressione sonora diversi.
Osservando il grafico di figura 2.5 si può comprendere facilmente che il nostro orecchio può percepire sensazioni simili a frequenze ed entità di eccitazione differenti. Ad esempio un tono sinusoidale a frequenza 5 KHz con un ampiezza di 40 dB SPL (a 1 KHz), produce nell'orecchio umano la stessa sensazione sonora si un tono a 150 Hz ma di ampiezza 30 dB SPL (a 1 KHz).
A questo punto si può accennare un concetto fisico che evince da quest'ultimo ragionamento. L'energia prodotta da un fenomeno fisico, è legata al concetto di lavoro e di potenza, in questo caso si tratteràdi potenza acustica.
Quantisticamente, l'energia di fenomeni ondulatori (così come il suono) è legata al concetto di frequenza secondo la seguente relazione matematica:

$$E = h * v$$

dove h è la costante di Planck, mentre E è il valore di energia, collegabile alla potenza acustica in grado di provocare determinati stimoli nervosi al nostro apparato uditivo.
Assumiamo dunque di trovarci ad esempio sulla curva isofonica a 40 dB SPL (a 1 KHz), assumendo di trovarci ad ascoltare un tono a frequenza 5000 Hz. Tale tono sinusoidale produce nel nostro orecchio una ben precisa sensazione acustica. La stessa identica sensazione acustica la possiamo ottenere ascoltando un tono a frequenza 150 Hz sulla curva a 30 db SPL (a 1 KHz). In pratica è come dire che i toni a frequenza più bassa trasportano un energia maggiore dei toni a frequenze alte.
Questo è un fenomeno importantissimo e va sotto il nome di "'mascheramento in frequenza"'.
Praticamente toni a frequenze basse, possono andare a mascherare toni a frequenze più elevate poichè posseggono maggiore energia e sono percepiti dal nostro orecchio con lo stesso, o con maggior volume in relazione a suoni a frequenze più elevate.
Le curve isofoniche ci fanno dunque capire come risponde il nostro orecchio alle varie frequenza. Ovviamente tali curve non sono uguali per tutti gli individui e dipendono molto dall'età, da possibili danni subiti all'apparato uditivo o al semplice deterioramento dell'udito dovuto all'invecchiamento.
La figura 2.5 mostra anche alcuni importanti limiti del nostro apparato auditivo. Per prima troviamo la soglia di udibilità, ovvero non siamo in grado di percepire suoni al di sotto dei 16-20 Hz, e nemmeno al di sopra dei 20000 Hz. Dopo di che possiamo notare l'esistenza di diverse curve, relative a diversi livelli di eccitazioni. In generale quando il livello di pressione sonora aumenta, e quindi maggiore potenza acustica colpisce il nostro orecchio esterno, ci spostiamo sempre di piu verso sensazioni rumorose, fino a raggiungere i 120 dB che sono considerati la soglia del dolore, alla quale l'orecchio subisce forti danni, specialmente se l'esposizione a tali sollecitazioni avviene per un periodo prolungato.
Un altro risultato molto importante è la nostra percezione dell'altezza di una nota. Un prima precisazione dovuta va rivolta verso il termine "'altezza"'. Comununemente quando si sente parlare di altezza di una nota si tende a pensare al suo volume, questa definizione tuttavia è semplicemente un abuso di termine. In musica il termine "'altezza"' sta ad indicare il posizionamento della nota sulla scala musicale, e dunque si parla di note acute, oppure gravi.
L'altezza delle note dipende dalla frequenza con cui sono prodotte, dunque un aumento di frequenza è percepito dal nostro orecchio come un la produzione di un tono piu acuto, oppure piu grave. In inglese "'altezza"' è tradotto come "'pitch"'. Il pitch percepito in funzione della frequenza è espresso attraverso quella che viene chiamata scala di Mel, mostrata in figura 2.6. Come si può notare il pitch delle note viene misurato in Mel, questa unità di misura equivale a un millesimo del pitch di un tono puro la cui frequenza è di 1000 Hz, che raggiunge il volume di 40 dB al di sopra della soglia dell'ascoltatore.
Si nota inoltre come un aumento della frequenza espressa in Hz porti ad un aumento logaritmico dell'altezza percepita della nota. Tutto questo ha senso dato che la risposta dell'orecchio è proprio di questo tipo.
Un altro fenomeno a cui il nostro orecchio è soggetto è la formazione di suoni di combinazione. Quando udiamo due suoni a frequenze molto vicine, questi si intermodulano tra di loro, quasi come due emittendi radio che disturbano le

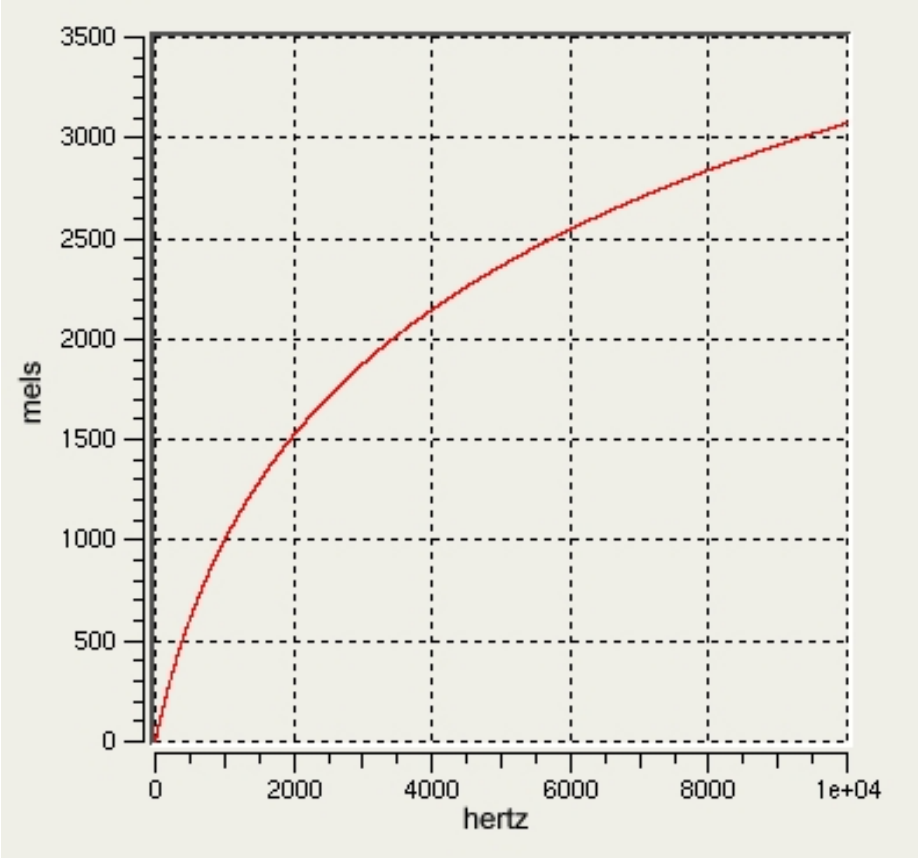

Figure 2.6: Scala di Mel: Relazione tra frequenza e altezza delle note.

proprie trasmissioni a vicenda. Quello che accade è che percepiamo un terzo suono, che ha natura di tipo differenziale, ossia a frequenza pari alla differenza tra le due percepite. Questo è quello che viene chiamato "'terzo suono di Tartini"'. In realtà si possono percepire anche suoni a frequenza pari alla somma dei due riprodotti.
Un altra caratteristica posseduta dal nostro orecchio è che prima che un suono possa essere percepito, non solo deve soddisfare a una certa soglia di "'volume"' ma deve anche avere dare il tempo al nostro sistema recettore di vincere l'inerzia della sorgente sonora e portatla dallo stato di riposo fino al suo regime ondulatorio. Vi sono poi fenomeni psicoacustici dovuti a caratteristiche collocabili nel tempo. In psicoacustica tutti i transitori di durata inferiore ai 30 ms sembrano avere attacchi di durate simili tra loro. Inoltre per percepire l'altezza del suono è necessario che, indipendente dalla sua altezza, esso abbia una durata minima di 10 ms.
Un suono percepito per una durata inferiore a questa soglia appare come un flebile rumore. Inoltre il nostro cervello tende a sottovalutare la durata dei suoni lunghi ed a sopravvalutare i suoni di durate brevi.
Altri fenomeni psicoacustici sono invece di ordine spaziale. Il fatto di avere non uno, ma due apparati recettori ci dota di alcune capacità particolari. Si parla di ascolto binaurale (monoaurale nel caso di un solo apparato di recezione funzionante). E' questa la base della stereofonia.
Il fatto di possedere due apparati uditivi rende possibile l'individuazione della direzione da cui una sorgente sonora emette. In pratica se la sorgente si trova in posizione centrale rispetto alle orecchie, i ritardi di propagazione del suono percepiti sono uguali e producono un eccitazione contemporanea o quasi. Il cervello interpreta tali segnali ed in questo caso capisce che il nostro corspo si trova in posizione frontale rispetto a tale sorgente. Se invece la posizione che assumiamo rispetto alla sorgente è asimettrica (ad esempio ci posizioniamo con un orecchi diretto verso di essa, e l'altro rivolto verso il lato opposto) i ritardi in gioco saranno differenti, dunque l'orecchio destro ad esempio riceve uno stimolo, mentre il sinistro lo riceve ma con un ritardo maggiore, al nostro cervello arrivano due impulsi che portano un informazione differenziale sulla quantità di tempo (e dunque sullo spazio) intercorsa tra la ricezione dei due segnali, e capisce che la sorgente è piu spostata da un lato piuttosto che da un altro.
Anche in questo punto torna in gioco la soggettività, in quanto ogni soggetto ha esperienze diverse alle spalle, e quindi valuta anche le distanze in modo diverso.
Altro effetto psicoacustico è quello che viene chiamato "'effetto party"', ovvero, la persona è in grado applicando un ascolto "'intenzionale"' di percepire un parlato all'interno di un ambiente altamente frequentato e con un certo livello di rumore di fondo, ed addirittura riconoscere la voce della persona. Dunque possediamo anche la capacità di "'rimuovere"' alcuni suoni ed enfatizzarne altri. Questi vantaggi sono possibili solamente effettuando un ascolto binaurale. Il principio di funzionamento è esattamente lo stesso degli occhi, con un solo occhio non è possibile percepire tutte le informazioni, ad esempio si perderebbe la percezione della profondità.
Di fatto anche nel caso dell'udito, un ascolto monoaurale non ci permetterebbe di posizionare dovutamente l'immagine stereo di un suono e quindi si potrebbe capire precisamente da dove provenga un suono.
Un altro effetto interessante è il rilevamento di "'falsi bassi"' durante un ascolto. Entriamo a contatto ogni giorno con questo fenomeno psicoacustico. E' noto che

dalla dimensione dei diffusori dipendono le frequenze riproducibili dal medesimo. Esse devono essere pari per lo meno alle lunghezze d'onda riproducibili con buona approssimazione. Dalla fisica si sa chè:

$$f = \frac{c}{\lambda}$$

dove f rappresenta la frequenza, del segnale audio in questo caso, c è la velocità della luce e λ è la lunghezza d'onda. Vi è di fatto una proporzionalità inversa tra la frequenza e la lunghezza d'onda. A frequenza basse corrisponde una lunghezza d'onda molto elevata, e viceversa per le frequenze alte. Da questo si può comprendere come mai i subwoofer per apprezzare le basse frequenze provenienti dai nostri impianti casalinghi, abbiano dimensioni maggiori dei diffusori "'normali"' che riproducono princpialmente frequenze medio-alte.
Come mai dunque quando si ascolta della musica attraverso le piccole cuffie dei lettori mp3, si riescono lo stesso a sentire le linee di basso?
La risposta è presto detta. Il nostro cervello sembra possedere una tendenza naturale a ricostruire le basse frequenze, in base alle informazioni contenute nella parte alta dello spettro acustico.
Ulteriore effetto delle strabilianti capacità che il nostro sistema orecchio-cervello è in grado di sviluppare, è quella di posizionare a determinate distanze, suoni riprodotti comunque in prossimità della nostra testa.
Ad esempio se disponessimo di una registrazione di un aeromobile in volo, e la facessimo riprodurre da un impianto stereo in modo che essa inizialmente sia completamente posizionata a destra e poi progressivamente passi al diffusore di sinistra al livello della nostra testa, posizionandoci al centro di questa situazione acustica, il nostro cervello collocherebbe direttamente tale sensazione acustica come proveniente da un livello di altezza piu elevato.
Questa introduzione sugli effetti psicoacustici serviva a dare un idea della complessità dei processi che coinvolgono l'ascolto da parte degli esseri umani, e che ne influenzano dunque il giudizio. Da qui evince la difficoltà di emulare il comportamento umano in termini di decisioni su un brano musicale, ma evincono anche le basi sulle quali fondare un software atto ad estrarre caratteristiche musicali e prendere decisioni in maniera simile.

2.3 Dalla psicoacustica all'analisi di segnali audio

Come è stato introdotto dalla sezione precedente l'analisi di segnali audio è un operazione che deve tener conto di molteplici fattori da parte di un essere umano. Ciò che bisogna capire è come relazionare il mondo psico-fisco al mondo dei calcolatori, in modo da poter ottenere risultati simili in merito alla classificazione di tali segnali attraverso il loro contenuto e la loro, per così dire, forma.
Per prima cosa, è stato detto che il nostro cervello è in grado di effettuare in maniera completamente gratuita l'analisi frequenziale di un brano musicale (e comunque di un suono nel senso più lato), dunque è immediato pensare che per svilppare un sistema di questo tipo possono essere utilizzati concetti legati allo sviluppo in serie di Fourier di segnali multimediali, ed al concetto di trasformata di Fourier, che ci permettono di rappresentare tali segnali nel dominio della frequenza, e dunque di poter fare valutazioni simili (con un certo margine

di approssimazione) a quelli effettuati dal nostro sistema orecchio-cervello.
Ogni suono possiede una forma d'onda (la rappresentazione grafica della sua evoluzione temporale), la quale può essere rappresentata nel tempo da una funzione matematica.
Supponiamo ad esempio di partire considerando un tono a frequenza fissa, il suono più semplice che potremmo immaginare, ovvero un suono con forma d'onda sinusoidale visualizzato in figura 2.7.

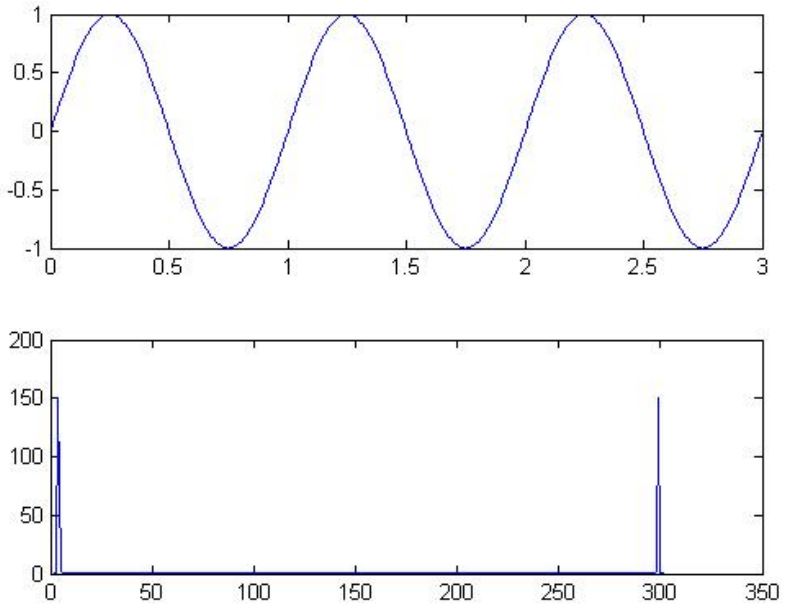

Figure 2.7: Sinusoide a frequenza 1 Hz, in tempo e frequenza

Attraverso la matematica sappiamo che è possibile esprimere tale segnale tramite la ben nota serie di Fourier, che permette di scrivere un segnale complesso come somma di tanti segnali elementari di tipo sinusoidali (e dunque periodici). Anche nel mondo dell'acustica vale tale approssimazione, dunque un suono è esprimibile come somma di suoni (componenti) elementari a frequenze multiple della fondamentale, ove la fondamentale è definita come la componente a frequenza piu bassa che compare nello spettro del segnale. Osservando la figura 2.7 possiamo vedere quello che si intende per spettro del segnale. Questo è ottenuto attraverso un altro strumento matematico ben noto, ovvero la trasformata di Fourier che ci permette di passare dal dominio temporale a quello frequenziale. Il passaggio tra rappresentazione temporale e rappresentazione dello spettro del segnale è il seguente:

- Segnale sinusoidale nel tempo:

$$x(t) = A * sin(2 * \pi * t)$$

  dove A è l'ampiezza del segnale sinusoidale.

- Segnale sinusoidale in frequenza:

$$X(f) = \int_{-\inf}^{\inf} x(t) * \exp(-2 * \pi * f * t)$$

quello che si ottiene da tale formula è lo spettro di ampiezza del segnale, che mostra come una sinusoide pura rappresenti una riga nello spettro.
Cosa accade per segnali più complessi?
Andando ad analizzare un suono più ricco, come ad esempio una canzone, ci rendiamo subito conto che i contributi di tutti gli strumenti si sommano (anche tra di loro) e generano una forma d'onda più complessa che si tradute in un segnale con un contenuto armonico in frequenza molto più ricco.
In figura 2.8 possiamo vedere quello che accade analizzando ad esempio il ritornello di un brano musicale.

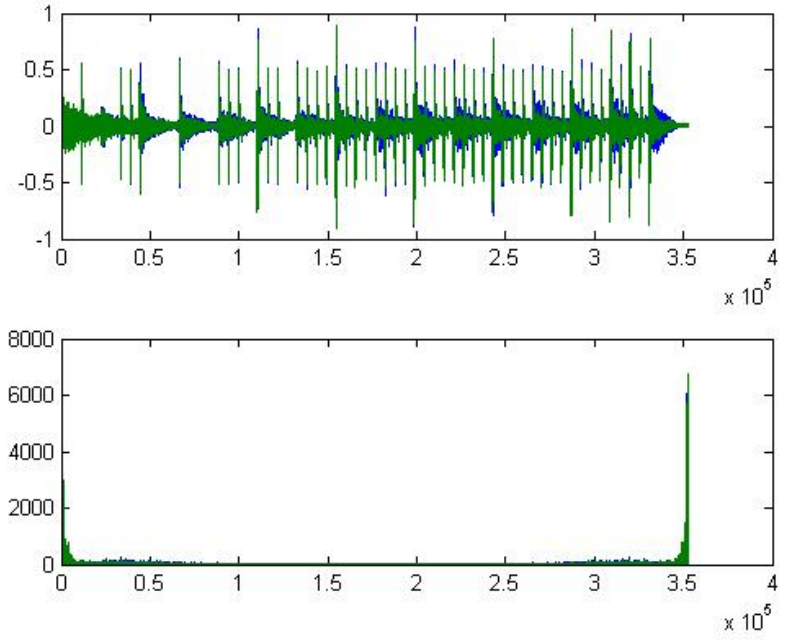

Figure 2.8: Ritornello di un brano musicale nel tempo ed in frequenza.

Come si può notare la forma d'onda si complica notevolmente, e lo spettro possiede molte più componenti all'interno di un certo range di frequenze. Rimarcando, in campo audio tale intervanno va da circa 20 Hz sino a 22050 Hz.
Questo tipo di rappresentazione dei segnali è di vitale importanza per l'analisi di contenuti audio, in particolare per determinare la timbrica, effettuare operazioni di fitraggio e scomposizione del segnale allo scopo di estrarne determinate caratteristiche che saranno introdotte più avanti in questa trattazione.
Dato che tutto questo sarà realizzato tramite un software per computer, sarà necessario coinvolgere metodi tempo discreti per attuare tali operazioni su file multimediali in formato digitale come gli mp3, i wav, gli au. La trasformata di Fourier è computazionalmente molto pesante da eseguire su un calcolatore

ma fortunatamente esistono tecniche per implementarla, che permettono di risparmiare notevoli risorse di calcolo e di comprimere di molto i tempi di esecuzione, si farà uso dunque di versioni tempo discrete per la realizzazione della trasformata che per semplicità verranno solamente elencate a questo punto della trattazione, si tratta di:

- Discrete Fourier Transform (DFT - Trasformata Di Fourier)
- Fast Fourier Transform (FFT - Trasformata di Fourier Veloce)
- Short Time Fourier Transform (STFT - Trasformata di Fourier su brevi intervalli di tempo)

Ottenuta la rappresentazione spettrale del segnale è possibile agire su di esso ad esempio per effettuare un filtraggio allo scopo di estrarre alcune caratteristiche, oppure effettuare operazioni elementari (e non) sulle sue componenti frequenziali.
Ad esempio, se consideriamo sempre il ritornello musicale rappresentato in figura 2.8 potremmo decidere, per qualche ragione, di rimuoverne le componenti frequenziali piu bassi, ad esempio al di sotto dei 3000 Hz (per esaltarne le componenti che il nostro orecchio percepisce meglio, poichè questa è circa la frequenza di risonanza del nostro apparato auditivo). A tale scopo è necessario progettare un filtro da applicare al suono attraverso il calcolatore. Si sorvolerà sulle tecniche di progettazione di filtri digitali, poichè non sono scopo di questa trattazione, ne verranno solo mostrati graficamente i risultati in figura 2.9.

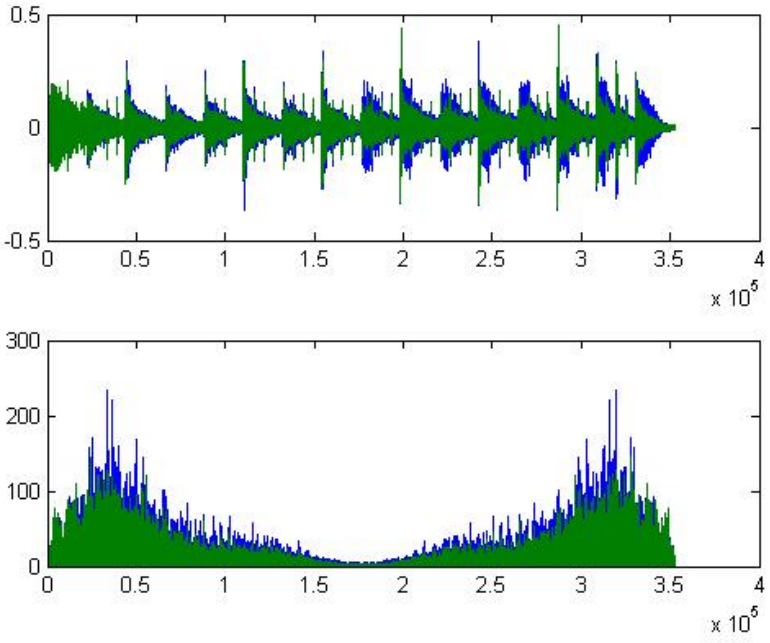

Figure 2.9: Filtraggio passa-alto su ritornello (oltre i 3 KHz)

Il tipo di filtro applicato in questo caso, non ha solamente soppresso le frequenze al di sotto dei 3 KHz, ma ha anche enfatizzato quelle superiori a tale soglia. Il risultato è un suono brillante, e completamente privo di basse frequenze, un pò come quello proveniente dall'altoparlante di un telefono.