

Riconoscimento automatico dei brani in trasmissioni audio affette da rumore

Fabrizio Mondo
Università degli Studi di Palermo
Corso di Laurea in Ingegneria Informatica
AA 2009 – 2010
Relatore Prof. Marco Ortolani

Abstract — Il problema dell'attribuzione di un titolo ed un autore ad un brano musicale, in qualsiasi modo esso venga trasmesso è una tematica di forte interesse nell'ambito della tutela del copyright nelle trasmissioni e tra gli amatori musicali.

Questa tesi verte su più modalità di riconoscimento automatico di brani musicali in trasmissioni affette da rumore. La letteratura a riguardo espone molteplici modalità di risoluzione del problema. Le principali sono quelle espresse in [2] e in [3] che affrontano la questione in modalità fondamentalmente diverse, tramite i modelli nascosti di Markov il primo e tramite un grafico 3D il secondo. Nessuno dei due approcci risulta comunque perfettamente adeguato alle trasmissioni radiofoniche odierne o alla ricezione d'audio d'ambiente, ma l'utilizzo di grafici per la valutazione dei picchi d'ampiezza di un brano, valutato su esperimenti eterogenei, risulta mediamente più efficiente.

Index Terms — Algoritmo di Viterbi, Audio Fingerprinting, Modelli nascosti di Markov, Coefficienti di Mel Cepstrum, Trasformata di Fourier, Algoritmo Expectation/Maximization, Analisi Content-based.

I. INTRODUZIONE

Attribuire correttamente un titolo ed un autore ad un brano musicale trasmesso da un'emittente radiotelevisiva, sia su Internet che tramite onde elettromagnetiche, è diventato un problema di notevole interesse. Tale questione, specialmente nelle trasmissioni tradizionali, via onde elettromagnetiche, è non banale, in quanto ci si trova di fronte a dei flussi audio non scanditi, privi ovviamente di metadati di alcun tipo, corrotti dal rumore (sia acustico che elettronico) e di qualità ed ampiezza fortemente variabili. L'interesse per tale questione scaturisce non solo da problemi inerenti all'utilizzo indebito di opere dell'ingegno, come nel caso di trasmissioni senza licenza o plagii, ma anche dalla necessità di automatizzazione, costantemente richiesta dalle società mediatiche, nel riconoscere l'effettivo numero di brani musicali trasmessi in un periodo temporale predefinito. La SIAE (Società Italiana autori ed editori) e l'IMAIE (Istituto Mutualistico per la tutela degli Artisti Interpreti ed Esecutori) hanno posto come problema, degno di nota, l'identificare con maggiore

precisione gli autori delle trasmissioni radiofoniche in modo da potere definire degli ordini di importanza nelle vendite e nelle attribuzioni di riconoscimenti, dato il loro attuale carattere aleatorio. Inoltre tali società intendono utilizzare algoritmi di riconoscimento per permettere di evitare conflitti di competenza con altre società operanti nello stesso ambito. L'interesse per il riconoscimento rapido e affidabile della paternità di un brano è avvertibile anche tra le comunità di semplici amatori, che avrebbero così la possibilità di ordinare collezioni musicali contenute in supporti analogici, non basandosi esclusivamente sulla loro conoscenza del campo.

II. BACKGROUND

Il problema del riconoscimento dei brani musicali trasmessi sia in una rete dati che tramite onde elettromagnetiche è affrontato da diversi punti di vista in letteratura: tramite algoritmi di riconoscimento dei fonemi, oppure tramite il cosiddetto fingerprinting. In [1], si è evidenziata la bontà di un approccio tramite l'estrazione di caratteristiche di un brano, popolamento di un database e successivo confronto di dati. In [2] invece viene proposta una metodologia di risoluzione basata sempre sul fingerprinting, che rappresenta un generico procedimento di creazione di dati che distinguano univocamente un brano da un altro, ma con algoritmi di confronto diverso. La qualità degli algoritmi di fingerprinting e confronto (che rappresentano *de facto* la tecnica più diffusa per affrontare il problema del riconoscimento audio) risulta fondamentale per il successo di un determinato software costruito a questo scopo. La qualità di tale approccio varia molto in letteratura e nessun caso può dirsi completo in toto. In [3] si evidenziano in particolare le difficoltà di riconoscimento qualora l'ampiezza sonora del brano sia inferiore ad un limite minimo, mentre in [2] si considera un filtraggio lineare del brano ascoltato, opzione troppo semplicistica nella maggior parte dei casi. In [7] si considera l'ipotesi di potere effettuare un'attribuzione di titolo ed autore ad un brano effettuando anche un confronto tra un testo associato al brano presente in database e il testo riconosciuto nel brano. Spesso è il

testo dello stesso brano, specialmente se contiene parti vocali, oppure un commento se il brano è esclusivamente strumentale.

III. L'APPROCCIO

Effettuare un riconoscimento digitale di un brano musicale proveniente da una qualsiasi fonte sonora, consiste fondamentalmente nell'effettuare un confronto tra un quantitativo di dati prelevati in un determinato periodo di tempo ed un database che contiene un indicizzazione di dati. Riconoscere il brano significa quindi avere una corrispondenza tra i dati campionati e una entry di tale database. Esistono diversi modi per verificare tale corrispondenza. La prima fase, ovvero il popolamento del database, viene effettuato in [2] tramite i Coefficienti di Mel Cepstrum, corrispondenti all'anti trasformata di Fourier del logaritmo della trasformata di Fourier del segnale di partenza, mentre in [3] la fase di estrazione delle caratteristiche univoche del brano viene effettuata tramite costruzione di un grafico in tre dimensioni: tempo, frequenza e ampiezza. (Fig. 1). Tutte le metodologie di lavoro portano all'estrazione di dati univoci, ma esse differiscono fondamentalmente nella quantità di dati estratti per ciascun brano e sulla successiva velocità di confronto.

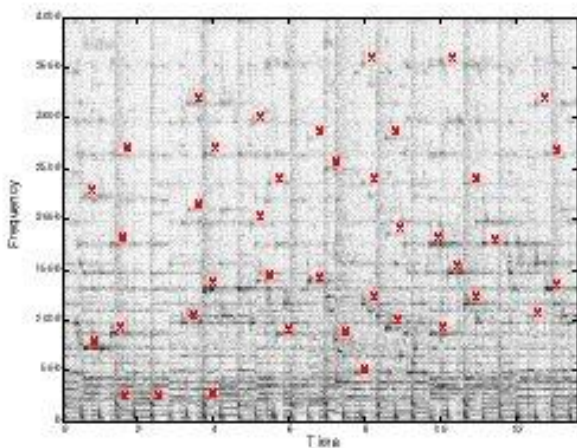


Fig. 1 – Mappa stellare dei punti di picco del grafico 3D presentato in [3]

Una delle operazioni eseguite da [2] per poter effettuare un riconoscimento maggiormente efficiente del brano è la stima della distorsione presente nel flusso audio che si sta analizzando. La scelta progettuale in questo caso è stata quella di proporre un filtro modellato da un sistema lineare tempo invariante (2) che risulta adeguato per flussi audio affetti da rumore continuo e costante sia in frequenza che nel tempo, ma non risulta tale nel caso in cui il rumore non sia bianco e non sia

distribuito uniformemente nel tempo.

Anche nel semplicistico caso di rumore rosa, ovvero di rumore in cui le componenti a bassa frequenza hanno potenza maggiore, le performance degradano.

$$CR(z) = 0.99 \frac{1 - z^{-1}}{1 - 0.98z^{-1}} \quad (2)$$

Filtro LTI proposto in [2]

La scelta di tale filtro è migliorabile se consideriamo come flusso audio da analizzare una qualsiasi trasmissione radiofonica. In tale caso infatti, il rumore è concentrato al 90% nella fase iniziale e finale di ogni brano e sovrasta il rumore termico costante. Le considerazioni presenti in [6], fanno supporre che un filtro con parametri dipendenti dal tempo realizzi filtraggi generalmente migliori.

Secondo [3] invece, è fondamentalmente inutile filtrare il rumore di fondo, filtrare il watermarking, ovvero l'inclusione di informazioni all'interno di un file multimediale o di altro genere, che può essere successivamente rilevato o estratto per trarre informazioni sulla sua origine e provenienza. Ma anche filtrare le voci degli speaker. Tutto questo perché, modificando la metodologia di estrazione delle caratteristiche concentrandosi esclusivamente sul rapporto frequenza/ampiezza in funzione del tempo, tali modifiche al brano d'origine non influenzano il corretto riconoscimento, tranne che ovviamente, in casi palesemente infattibili. E' curioso notare come il concetto di infattibilità nel riconoscimento sia correlato alla fisicità dell'essere umano.

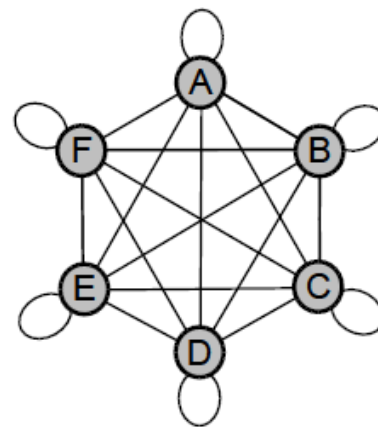


Fig. 2 - Modello HMM

Viene infatti riconosciuto (se presente in database) qualsiasi degradazione di un brano che consenta il riconoscimento da parte di un essere umano. Una delle fasi successive del processo di riconoscimento proposto in [2] è l'addestramento dei modelli nascosti di Markov (da ora in poi HMM) che rappresentano uno strumento fondamentale. Una sua rappresentazione è in Fig. 2

Un processo stocastico markoviano o processo di Markov è un processo stocastico nel quale la probabilità di transizione che determina il passaggio ad uno stato di sistema dipende unicamente dallo stato di sistema immediatamente precedente e non dal come si è giunti a tale stato. I modelli nascosti rappresentano generatori di audio generici. Nel caso di musica non modellabile come semplice somma di strumenti differenti, quale ad esempio la musica in commercio e le trasmissioni radiofoniche, un HMM è semplicemente un modello che massimizza la probabilità per la quale, se esso fosse davvero un generatore di suoni, allora genererebbe quel suono in particolare. Con gli HMM ci sono fondamentalmente due problemi canonici, risolti da due algoritmi. Il primo problema è quello di trovare la sequenza più probabile che potrebbe generare una data sequenza dell'uscita, risolto dall'algoritmo di Viterbi, mentre il secondo è trovare l'insieme più probabile per il quale si possano dichiarare le probabilità dell'uscita e di transizione, ovvero "addestrare" i parametri dell'HMM dato mediante il gruppo dei dati relativi alle sequenze. Il problema è risolto dall'algoritmo di Baum-Welch.

L'algoritmo di Viterbi è un algoritmo fondamentale nell'ingegneria delle telecomunicazioni ed ha radicalmente rivoluzionato il mondo della trasmissione numerica. Viene generalmente utilizzato per trovare la migliore sequenza di stati (detta Viterbi path) in una sequenza di eventi osservati in un processo Markoviano. L'algoritmo è tanto più prestante quanto il numero di passi è alto. Ovviamente maggiore è il numero di passi e maggiore è la lentezza nella decodifica e maggiore è il dispendio di risorse. La complessità di calcolo del decodificatore si può immaginare calcolando che per un codice con i stati e t passi di osservazione, si hanno $2^{i(t-1)}$ cammini possibili.

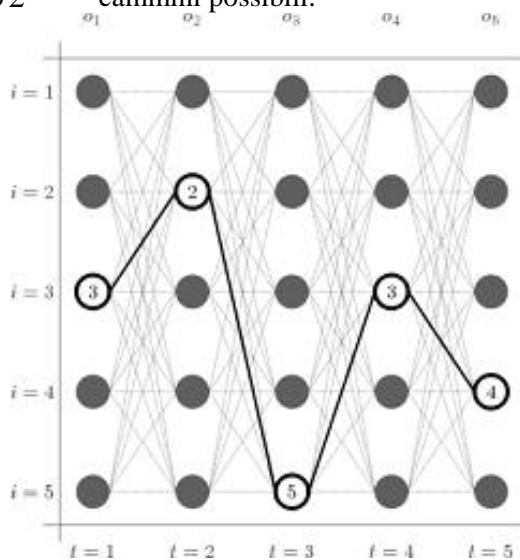


Fig. 3 – Esempio di movimento nel traliccio in un algoritmo di Viterbi

Ad ogni passo vi sono 2^i cammini che raggiungono ogni singolo stato. Di tutti i cammini uno solo sarà quello a distanza minima fino a quel passo. La ricerca della soluzione ottima con una tecnica esaustiva diventa rapidamente inapplicabile al crescere di i e t al di sopra di valori abbastanza bassi. Sono invece applicabili tecniche che riducono la complessità del problema applicando metodologie di programmazione dinamica. Il criterio di scelta tra le possibilità può essere la distanza minima di Hamming rispetto alla sequenza ricevuta oppure la distanza euclidea tra i vari segnali. L'algoritmo di Baum-Welch, detto anche algoritmo Forward Backward, invece, consta di una modificazione dell'algoritmo di Expectation Maximization, (da ora in poi EM). L'algoritmo EM permette l'apprendimento con set di dati incompleti o con set di dati generati da misture di distribuzioni di probabilità nel caso in cui non si conoscano né i parametri delle funzioni di distribuzione, né l'appartenenza di ciascun dato ad una data funzione di distribuzione. Il metodo utilizzato per ricavare i parametri di ciascun HMM, è modificare ulteriormente l'algoritmo EM, in modo che i dati incompleti siano non solo i parametri degli HMM ma anche la loro corretta sequenza per ciascun brano, trattandoli tramite la formula espressa in (3).

$$g(O|\lambda) = \int_{\Phi(O)} f(\phi|\lambda) d\phi \quad (3)$$

Relazione presente in [2] tra lo spazio dei campioni completo e quello incompleto. Procedura attuata per l'addestramento iterativo dei HMM

A differenza del riconoscimento vocale dove i fonemi sono definiti a priori per standard internazionale, è quasi impossibile ottenere degli standard ben definiti in ambito musicale. Per questo motivo ciascuna iterazione dell'algoritmo di Baum-Welch, modificato ad hoc per la causa, consente progressivamente di migliorare il set a disposizione, che è accettabile con una trentina di iterazioni per la musica con contenuto spettrale nella banda di frequenza compresa tra 200 e circa 8000 hertz, mentre richiede iterazioni supplementari, per contenuti spettrali rilevanti al di fuori della banda in frequenza della voce umana.

In [3] questa fase è sostituita dal semplice campionamento delle caratteristiche del brano e da un algoritmo che campioni tali caratteristiche in modo che esse siano robuste e tempo invarianti. In [2], la generazione della firma, consiste in un procedimento di realizzazione di una sequenza di HMM che identifichi un brano tra gli altri. Le firme sono generate usando l'algoritmo di Viterbi, che calcola il cammino dalla probabilità più alta tra HMM in un grafo di HMM

completo, ovvero dove da qualsiasi HMM si possa passare ad un qualsiasi altro. Tutte queste firme sono calcolate e immagazzinate in un database apposito. La complessità computazionale dell'algoritmo di Viterbi è dimostrabile essere linearmente dipendente dalla lunghezza del brano e dipendente dal quadrato del numero di HMM del grafo. (4)

$$O(L*Q^2) \quad (4)$$

Complessità computazionale dell'algoritmo di Viterbi applicato nel procedimento descritto in [2]

L'ultima fase del processo è l'identificazione vera e propria. In [2] viene utilizzato l'algoritmo di Viterbi nuovamente, ma questa volta per un grafo non completo, quale è il modello ciclico di HMM proposto. Tale modello proposto è costruito linkando tutte le firme in sequenza in un anello, dove ciascun HMM ha solo un link con se stesso e con il suo immediato successore, schema visibile in Fig. 4

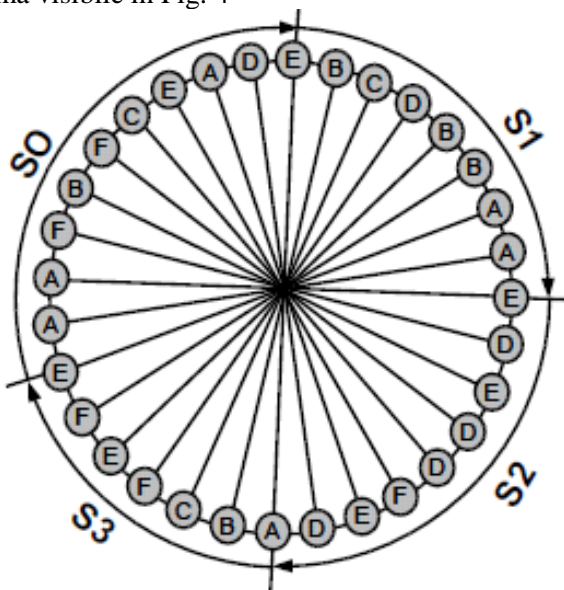


Fig. 4 (Modello di HMM proposto in [2] per un database di soli quattro brani)

L'algoritmo di Viterbi è periodicamente autorizzato a fare salti da una sezione ad un'altra, in modo da consentire passaggi da una sezione ad un'altra di un brano se non anche a brani diversi. L'unione delle due cose, lo schema ad anello e lo stesso algoritmo di Viterbi, consente di ottenere risultati nei casi di riconoscimento standard, ovvero il caso più semplice in cui vi è solo un brano, ma anche nei casi di missaggio, che devono ovviamente includere salti interni dovuti ai cambi di brano. La complessità computazionale dell'algoritmo di Viterbi per il grafo ad anello è lineare nella lunghezza del brano e lineare anche nel numero di

HMM presenti, il che implica un costo di ingrandimento del database di firme potenzialmente accettabile anche su lunghi periodi. Un diagramma di flusso delle operazioni effettuate dalla procedura in [2] è visibile in figura 5.

In [3] viene utilizzato l'hashing, ovvero un algoritmo che, partendo da un documento di qualsiasi dimensione, lo elabora e produce un codice di misura fissa. Il metodo d'elaborazione è tale che, se il documento fosse cambiato in qualunque sua parte, questo codice cambierebbe. Per esemplificare s'immagini un algoritmo che calcola il numero di lettere, il numero di parole o la frequenza d'ogni lettera in un testo, se cambia una qualsiasi lettera o parola anche il risultato cambia. Si può pensare al codice prodotto dall'algoritmo di hashing come appunto ad un'impronta del documento. Dall'impronta non è possibile risalire al documento, però se questo cambia, anche solo in minima parte, allora cambia anche l'impronta. L'hashing viene effettuato per ogni intorno del punto candidato a picco d'ampiezza. In questo modo, è più semplice creare un grafico di matching tra il flusso in ingresso e i campioni presenti in database, ma rende estremamente più complesso il riconoscimento di un brano in dipendenza dei movimenti del pitch. E' stato provato sperimentalmente che è sufficiente aumentare o diminuire del 10% il pitch di un brano, per ottenere falsi negativi o falsi positivi.

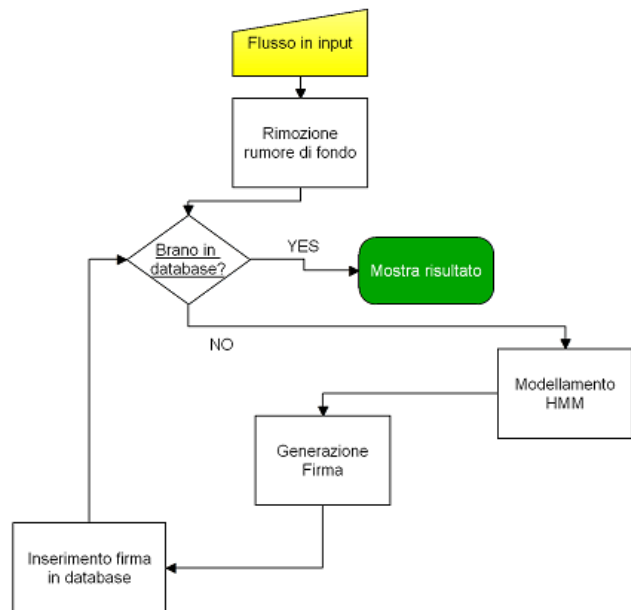


Fig 5 – Diagramma di flusso delle operazioni che conducono al riconoscimento di un brano descritte in [2]

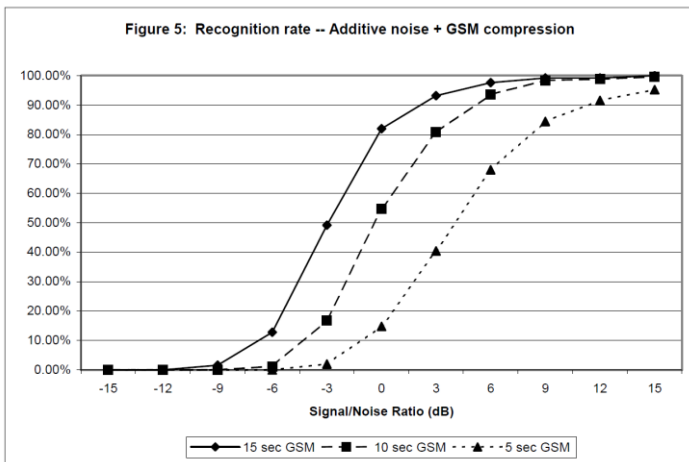


Fig. 6 – Tasso di riconoscimento in caso di rumore e di compressione GSM in [3]

In entrambi i propositi di risoluzione del problema, esistono delle sottoproblematiche non semplicemente risolvibili. I due problemi maggiormente rilevanti nella creazione del database delle firme e nell'effettiva messa in funzione dell'algoritmo, sono i cosiddetti falsi. Tali falsi possono essere positivi (ad esempio la presenza di multiple copie di uno stesso brano) e negativi (mancato riconoscimento di un brano presente in database). I primi possono nascere fondamentalmente in tre occasioni:

- **Stesso file:** Nel caso in cui un artista operi un duetto di uno stesso brano da lui precedentemente cantato come artista singolo o nel caso di inserimento di un brano all'interno di una compilation, il brano viene riconosciuto come differente rispetto alla versione originale.
- **Stesso brano:** Una versione live ed una versione studio di un brano, sono differenti solo qualitativamente, ma rappresentano lo stesso brano. Tali versioni sono registrate diversamente in database. Questo può o non può essere considerato un problema. In [2] non sono ammesse differenze di versioni tra gli stessi brani, mentre sarebbe stato più opportuno semplicemente creare un'altra entry.
- **Missaggio:** Un singolo brano causa falsi positivi se è composto da più parti dello stesso album, ovvero sia nel caso dei medley che in quello di un bootleg. In musica, un medley rappresenta una serie di canzoni (o solo parti di canzoni a volte sovrapposte), suonate una dopo l'altra senza interruzioni. Generalmente i temi che formano il medley sono connessi oltre

l'unione musicale diretta, a volte essendo tutti di uno stesso artista, album o di uno stesso genere. Il termine bootleg è entrato nell'uso gergale italiano per indicare un disco prodotto, distribuito o commercializzato, non necessariamente a fini di lucro, senza l'autorizzazione del detentore dei diritti d'autore. Spesso tali dischi sono registrazioni abusive eseguite ai concerti usando microfoni nascosti, ma non necessariamente la qualità della registrazione è di basso livello.

I falsi negativi avvengono invece per altre cause:

- **Eccessivo rumore:** Esiste un limite oltre il quale non è possibile, per nessun algoritmo esistente, potere trovare corrispondenze esatte.
- **Filtro non adeguato al brano:** Un filtraggio a priori che non risulti adeguato al brano da analizzare causa una maggiore difficoltà di riconoscimento.
- **Missaggi improvvisi in parti non canoniche del brano:** I passaggi tra brani, se non preventivati dai filtri, possono condurre a maggiori difficoltà di riconoscimento. Nei cosiddetti bootleg o nei medley tale problema è infatti notevolmente accentuato.
- **Eccessivi falsi positivi:** Una quantità eccessiva di falsi positivi (la presenza di multiple copie di uno stesso brano) causa un incremento dei falsi negativi in ricezione. Questa apparente contraddizione è giustificata da una maggiore difficoltà degli algoritmi nel capire quale sia il corretto autore e titolo da attribuire.

Gli esperimenti compiuti in [2] sono stati effettuati utilizzando 256 Modelli di Markov come limite minimo per ciascuna firma di un brano, con una media di 450 modelli. Sono state realizzate 3852 firme di brani, immagazzinate in un database, con sei secondi di periodicità per i link interni. Le sperimentazioni hanno portato alla realizzazione di un test bench che riassume globalmente il percorso di riconoscimento del brano, illustrato in Fig. 7.

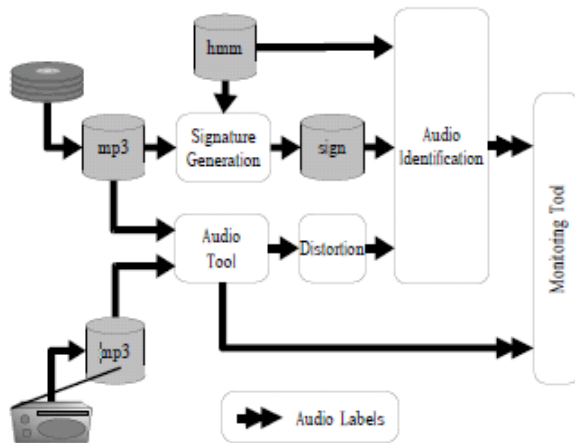


Fig. 7 - Schema del Test Bench

In [3], in un database di circa ventimila tracce implementato su un personal computer, il tempo di ricerca era dell'ordine delle centinaia di millisecondi, in dipendenza del settaggio dei parametri e dall'applicazione. L'applicazione può riconoscere un brano anche se pesantemente corrotto in qualche decimo di secondo, con un obiettivo di ottimizzazione di circa 1 millisecondo a richiesta. Gli esperimenti sono tuttavia stati condotti perturbando i brani con un insieme di rumori prestabilito, ponendo dei dubbi sulla veridicità dei risultati nei casi in cui i disturbi siano di natura non prevista in sede sperimentale. Uno di questi casi può porsi ad esempio nei casi in cui venisse effettuato il tentativo di riconoscimento di un brano in un ambiente con fonte sonora distante più di 5 metri dall'apparecchio ricevente (nel caso di Shazam, un cellulare) e con tale fonte costituita da diffusori di un brano digitale compresso. Tale contesto risulta quello più comune nella media utenza ed è dimostrabile essere il caso peggiore tra le condizioni di effettiva possibilità di riconoscimento di un flusso audio.

IV. CONSIDERAZIONI

La complessità computazionale, considerati anche i vari sottoalgoritmi, dei due approcci considerati in questa tesi, è minore in [3] in quanto complessità quasi lineare nel numero di brani in database $O(N \log N)$, di quanto in [2], che ha complessità quadratica

La metodologia di risoluzione del problema riconoscimento brani proposta in [2] è stata sperimentata tramite un Pentium-III con clock da 1GHz. Con tale piattaforma i risultati sono stati considerati tempo-realistici per un singolo flusso audio in ingresso. Nell'utilizzo comune, i software basati sulla proposta espressa in [3] sono risultati performanti in tempi

accettabili, considerato anche che una tale applicazione non ha la necessità di essere strettamente real time.

Gli esperimenti e gli sviluppi futuri di tali applicazioni devono portare ad ottenere un riconoscimento più sicuro nei casi di missaggi e di versioni live di brani memorizzati in database. Sebbene tale algoritmo non venga affetto da riduzioni nella risoluzione dei campioni di un brano sorgente, si è ancora lontani dall'aver una piena capacità di riconoscimento se la frequenza di campionamento è minore di 22050 hertz e si è sotto i 32 kbps.

V. CONCLUSIONI

La tesi proposta è stata realizzata per mettere a confronto diverse tipologie di risoluzione del problema riconoscimento brani in trasmissioni audio affette da rumore, focalizzando l'attenzione su due procedure in particolare. Il problema dell'attribuzione di paternità di un brano è un problema sensibile che ancora non può dirsi risolto. Esso fondamentalmente dipende dalla costruzione di un archivio di firme, la cui fattibilità viene spesso messa in mani alle comunità di utenti che utilizzano software di riconoscimento apposito e dalle società di tutela del diritto d'autore, che tendono a tutelare solo i loro iscritti. Questo implica che la maggior parte dei brani meno conosciuti risulta invisibile a tali algoritmi. Una proposta di miglioramento delle procedure di identificazione di artista e titolo è quella di basarsi su una vecchia idea ormai tramontata, il riconoscimento dei fonemi.

BIBLIOGRAFIA

- [1] Pedro Cano¹, Eloi Batlle¹, Harald Mayer² and Helmut Neuschmied, "Robust Sound Modeling for Song Detection in Broadcast Audio" (AES 112TH CONVENTION, MUNICH, GERMANY), May, 2002
- [2] Eloi Batlle, Jaume Masip, Enric Guaus, "Automatic Song Identification in Noisy Broadcast Audio" (Audiovisual Institute and Dept. of Technology, Pompeu Fabra University), 2002
- [3] Avery Li-Chun Wang, "An Industrial-Strength Audio Search Algorithm" in Conference on Music Information Retrieval (ISMIR), 2008
- [4] O. Izmirli, "Using a Spectral Flatness Based Feature for Audio Segmentation and Retrieval," in Proceedings International Symposium on Music Information Retrieval (2000).
- [5] B. Logan, "Mel Frequency Cepstral Coefficients for Music Modelling," in Proceedings International Symposium on Music Information Retrieval (2000).
- [6] R. A. Bates, "Reducing the Effects of Linear Channel Distortion on Continuous Speech Recognition," M.S. thesis, Col. Of Engineering, Boston University, 1996
- [7] E Brochu, N De Freitas, "Name That Song!: A Probabilistic Approach to Querying on Music and Text", Advances in neural information processing, 2003.