

# Automatic Structure Detection for Popular Music

Namunu C.Maddage

Institute for Infocomm Research

IEEE January – March 2006

Sebastiano Vascon (matr. 788442)

svascon@dsi.unive.it

Corso di Sistemi MultiMediali

A.A. 2009/2010

## Introduzione

In questo paper [1] viene proposto un nuovo approccio per l'estrazione del contenuto strutturale di un brano musicale. L'idea alla base, assente nei precedenti studi, è nel combinare elementi di conoscenza musicale di alto livello (come la struttura di un particolare genere musicale) ed elementi di basso livello come l'analisi dei segnali al fine di estrarre la struttura di un brano. Questo studio si basa fortemente su euristiche raccolte nell'ambito della musica Pop in lingua Inglese, pertanto risulta inutile, o comunque non porterà a risultati apprezzabili, se applicato a contesti diversi.

Le informazioni di una canzone quali il beat, il tempo, le linee melodiche ed armoniche, le parti vocali e strumentali e le *parti strutturali* come l'introduzione, il ritornello, il bridge ed il verso sono essenziali alla comprensione del contenuto di brano. Tali informazioni sono utilizzate in molte applicazioni come nella generazione di anteprime, nella trascrizione di partiture, nel riconoscimento automatico del testo di un brano, nella ricerca di musica e nello streaming audio.

I passi seguiti per la generazione della struttura di un brano sono quattro:

### 1. Estrazione del ritmo e BSS

In questa fase (vedi illustrazione 2) viene estratta la durata della nota più corta basandosi sugli onsets<sup>1</sup> e il beat della canzone. Successivamente il brano viene segmentato con la tecnica BSS<sup>2</sup>.

La struttura armonica di un segnale musicale è rappresentata per ottave, questo viene quindi decomposto in 8 sotto-bande dove ogni sotto-banda ha un suo range di frequenze (vedi illustrazione 1).

Ogni scomposizione viene segmentata in frame da

1 L'onset rappresenta l'istante in cui un evento ha inizio, in ambito musicale l'onset di una nota è l'istante nel quale la nota ha inizio.

2 B.S.S. è l'acronimo per *Beat Space Segmentation*, il brano viene diviso in segmenti con lunghezza pari alla durata della nota più breve.

60ms con 50% di sovrapposizione, che significa avere in ogni frame la metà del contenuto musicale del frame precedente.

Subband	01	02	03	04	05	06	07	08
Octave scale	-B1	C2 - B2	C3 - B3	C4 - B4	C5 - B5	C6 - B6	C7 - B7	C8 - B8
Frequency	64 - 128	128 - 256	256 - 512	512 - 1024	1024 - 2048	2048 - 4096	4096 - 8192	
12 pitch-class notes	C	65.406	130.813	261.626	523.251	1046.502	2093.004	4186.008
	C#	69.296	138.591	277.183	554.365	1108.730	2217.460	4434.920
	D	73.416	146.832	293.665	587.330	1174.659	2349.318	4698.636
	D#	77.782	155.563	311.127	622.254	1244.508	2489.016	4978.032
	E	82.407	164.814	329.628	659.255	1318.510	2637.02	5274.04
	F	87.307	174.614	349.228	698.456	1396.913	2793.826	5587.652
	F#	92.499	184.997	369.994	739.989	1479.987	2959.956	5919.912
	G	97.999	195.998	391.995	793.991	1567.982	3135.964	6271.928
	G#	103.826	207.652	415.305	830.609	1661.219	3322.438	6644.876
	A	110.000	220.000	440.00	880.000	1760.000	3520.000	7040.000
	A#	116.541	233.082	466.164	932.328	1864.655	3729.310	7458.62
	B	123.471	246.942	493.883	987.767	1975.533	395.066	7902.132

Illustrazione 1: Frequenze fondamentali (F0) delle note musicali e le sotto-bande delle ottave

La musica popolare situa la maggior parte del contenuto musicale nelle bande comprese tra la 1a e la 4a, si procede dunque ad una analisi con un metodo simile a quello proposto da Duxbury et. al [2] nel quale si misurano i *transienti*<sup>3</sup> della frequenza nelle bande 1a - 4a mentre per le bande 5a - 8a si utilizzano le transienti dell'energia. Questo è dovuto al fatto che frequenze basse hanno transienti più marcate nei valori di frequenza mentre frequenze alte hanno valori di transiente più evidenti nell'energia. Gli onsets finali vengono calcolati con una somma pesata degli onsets rilevati nelle varie sotto-bande. La stima della lunghezza del primo *interbeat*<sup>4</sup> viene calcolata con l'*autocorrelazione*<sup>5</sup> tra gli onsets rilevati precedentemente.

La teoria musicale ci insegna come un brano sia composto da alternanze tra beat forti e beat deboli (*accenti*), viene quindi utilizzata la *programmazione dinamica*<sup>6</sup> al fine di riconoscere correttamente pattern composti da sequenze di beat forti e deboli.

Una volta ottenuta la lunghezza dell'*interbeat* il brano viene segmentato e i silenzi vengono rimossi. Successivamente le parti segmentate non silenziose verranno analizzate per individuare gli accordi e le regioni con componenti vocali (cantate).

### 2. Estrazione degli accordi del brano

Un accordo è costruito su un insieme di almeno due note che vengono suonate contemporaneamente.

3 Immediatamente dopo l'onset di una nota abbiamo la fase transiente. In questa fase si riscontrano picchi di frequenza e di energia casuali che tendono negli istanti successivi a decadere fino a portare a regime la nota che noi udiamo.

4 Distanza temporale che separa un beat dal successivo

5 L'autocorrelazione definisce il grado di dipendenza tra i valori assunti da una funzione campionata nel suo dominio in ascissa. È utile per cercare in un segnale dei pattern che si ripetono, in modo tale da determinare la presenza di un segnale periodico, in questo caso il beat della canzone.

6 È una tecnica di progettazione di algoritmi basata sulla divisione del problema in sotto-problemi e sull'utilizzo di sotto-strutture ottimali.

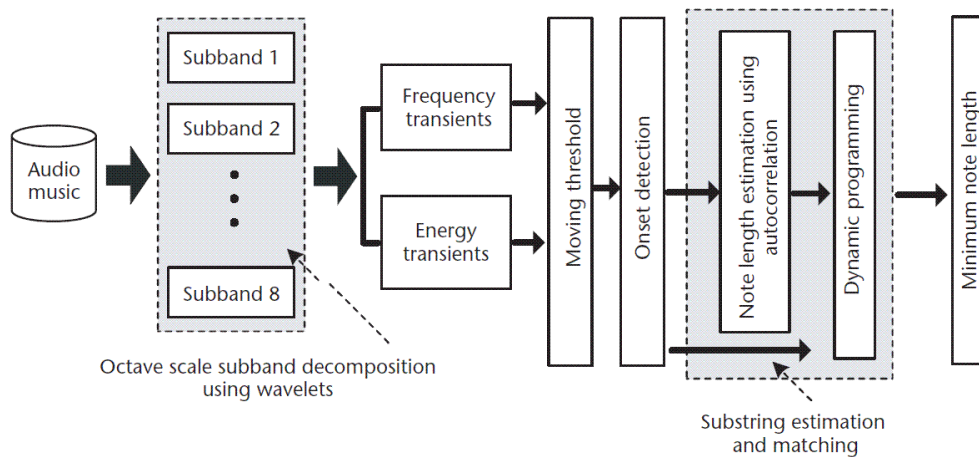


Illustrazione 2: Estrazione della lunghezza della nota più breve. Immagine tratta da [1]

Una nota è composta da una frequenza fondamentale (chiamata  $F0$ ) e dalle sue armoniche (multipli della frequenza fondamentale), l'idea è quindi di rilevare le frequenze fondamentali delle note, suonate in un certo istante, al fine di definire la struttura dell'accordo. Questo risulta essenziale per identificare le regioni che hanno la stessa melodia (*melody-based similarity region*) e quindi stessa progressione di accordi, come ad esempio il verso di una canzone. Gli accordi sono modellati tramite HMM<sup>7</sup> basati sul *Pitch Class Profile* [3], un estrattore di feature molto sensibile alle frequenze fondamentali.

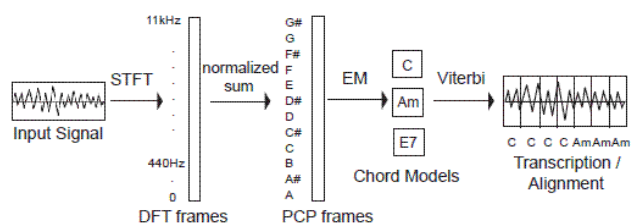


Illustrazione 3: Processo di estrazione del vettore di feature PCP. Immagine tratta da [3].

Sono stati impiegati 48 HMM che modellano le 4 principali tipologie d'accordo (maggiore, minore, aumentato e diminuito) e per ogni tipologia le 12 possibili toniche<sup>8</sup>. Questo non evita una serie di problemi dovuti alla fisica del suono. Ad esempio gli strumenti a corda presentano una forte componente sul 3° armonico che si sovrappone con l'ottavo semitono dell'ottava superiore. Ad esempio il terzo armonico della nota Do3 (C3) è simile alla nota situata ad 8 semitoni sopra il Do3, il Sol3, nella sua ottava superiore quindi il Sol4 (G4).

$$F0[C]_{03} = 130,813 \quad F0[G]_{04} = 391,995$$

$$TerzoArmonico[C]_{03} = 3 * F0[C]_{03} = 392,439$$

$$TerzoArmonico[C]_{03} - F0[G]_{04} = 0,444$$

Si possono quindi ottenere valutazioni errate della nota e conseguentemente trovare delle ambiguità negli accordi rilevati. Per risolvere questo genere di problemi vengono utilizzate euristiche raccolte sia nel brano corrente che nel genere musicale, cercando

di determinare la chiave<sup>9</sup> della canzone. Risulta sufficiente l'analisi di 16 battute per determinare la chiave di un brano [4]. Da questa analisi riusciamo a disambiguare coppie di accordi, in quanto solo uno dei due apparterrà alla scala del brano. Le regole di sostituzione degli accordi seguono lo schema:

1. Se l'accordo fuori chiave ha valore di osservazione superiore ad una soglia questo viene sostituito con l'accordo nella chiave la cui osservazione è immediatamente inferiore all'accordo errato.
2. Se non sono presenti accordi in chiave sopra la soglia allora viene assegnato il precedente accordo buono.

Ora il problema è sincronizzare i cambi d'accordo con il beat del brano. Si considera quindi il segnale musicale semi-stazionario in un intervallo di tempo pari all'interbeat rilevato al punto 1 e vengono applicate le seguenti euristiche:

1. Gli accordi variano sul tempo di beat piuttosto che in altre posizioni.
2. Gli accordi tendono a cambiare sulle minime<sup>10</sup>
3. Gli accordi tendono a cambiare all'inizio di una battuta.

Utilizzando questi accorgimenti gli accordi vengono ancorati alla griglia del brano.

### 3. Rilevamento delle parti vocali

Le parti cantate rivestono un ruolo particolarmente importante nell'analisi della struttura di un brano. Grazie alla loro individuazione riusciamo a distinguere regioni dal contenuto musicale simile. Ad esempio può succedere che il verso ed il ritornello (chorus) abbiano gli stessi accordi ma sicuramente avranno parti vocali differenti, mentre la parte vocale del ritornello sarà sempre lo stesso. Viene utilizzata SVM<sup>11</sup> per classificare i singoli frame

7 Hidden Markov Model

8 La nota principale dell'accordo

9 La chiave del brano (o tonalità) rappresenta l'insieme di note sulle quali esso è costruito. La chiave dunque induce una scala sulla quale sono costruiti gli accordi del brano. Ad esempio la scala di Do (Do, Re, Mi, Fa, Sol, La, Si) produce gli accordi: DoMaj7 - Rem7 - Mim7 - Fmaj7 - Sol7 - Lam7 - Sim5b.

10 Una minima è una nota la cui durata è la metà di una battuta.

11 Le Support Vector Machine appartengono alla famiglia dei classificatori lineari e rappresentano un insieme di metodi di apprendimento supervisionato per la regressione e la

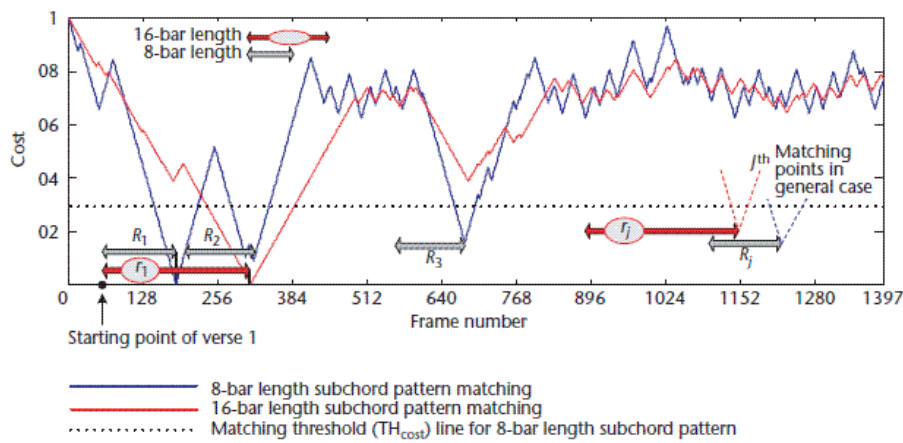


Illustrazione 4: Identificazione delle similarità nelle regioni (pattern matching) di un brano.  
Figura tratta da [1].

in “vocali” o “strumentali”. Il training della SVM viene fatto sull'estrazione degli “Octave Scale Cepstral Coefficients” che risultano particolarmente sensibili alla linea vocale [5].

#### 4. Rilevazione della struttura della canzone

La struttura del brano viene rilevata a partire dall'estrazione di regioni simili per melodia (*melody-based similarity region*) e per contenuto (*content-based similarity region*), successivamente vengono applicate delle euristiche ricavate dal genere musicale e quindi inferita la struttura finale.

##### Individuazione di regioni con melodia simile

La ripetizione di pattern di accordi da origine alle regioni con similarità melodica. Utilizzando algoritmi di pattern matching di programmazione dinamica rileviamo le similarità all'interno del brano. Il matching viene fatto per lunghezza di battute pari a 8 e 16. Nell'illustrazione 4 possiamo vedere come le regioni R1, R2 ed R3 abbiano lo stesso contenuto di accordi e lunghezza pari ad 8 battute, facendo il matching con lunghezza pari a 16 non riusciamo a trovare nessuna corrispondenza.

##### Individuazione di regioni con contenuto simile

Due ritornelli, della medesima canzone, hanno lo stesso contenuto vocale e differenze non troppo significative nella parte melodica (solitamente vengono aggiunti strumenti ad ogni esecuzione del ritornello per movimentare il brano). I frame marcati come vocali nella BSS vengono nuovamente segmentati in sub-frame da 30 ms con 50% di sovrapposizione. Vengono dunque estratti 20 OCSS per ogni sub-frame. Viene calcolata la distanza e la dissimilarità tra i vettori di coefficienti utilizzando

$$dist_{R_i, R_j}(k) = \frac{|V_i(k) - V_j(k)|}{|V_i(k) * |V_j(k)|} \quad i \neq j$$

$$dissimilarity(R_i, R_j) = \sum_{k=1}^n \frac{dist_{R_i, R_j}(k)}{n}$$

I valori di dissimilarità vengono computati per ogni

classificazione di pattern.

coppia di regioni e normalizzati. Sperimentalmente si è visto che una soglia  $TH_{smir}=0,3896$  permette una buona individuazione delle regioni con contenuti simili. Le coppie di regioni con valore di dissimilarità inferiore a  $TH_{smir}$  sono marcate come simili per contenuto.

##### Euristiche per la struttura finale

Le seguenti euristiche vengono utilizzate per dare la forma definitiva alla struttura del brano:

1. Pattern di versi e ritornelli:
  1. Intro – Verso 1 – Rit – Verso 2 – Rit – Rit – Outro
  2. Intro – Verso 1 – Verso 2 – Rit – Verso 3 – Rit – Bridge – Rit – Rit – Outro
  3. Intro – Verso 1 – Verso 2 – Rit – Verso 3 – Bridge – Rit – Rit – Outro
2. Il numero minimo di versi è due e di ritornelli è tre.
3. Il verso ed il ritornello hanno durata di 8 o 16 battute
4. Il bridge ha durata pari a 8 o 16 battute

La tonalità del brano, come detto in precedenza, determina gli accordi che lo comporranno. Brani popolari in lingua Inglese, nei quali si hanno più tonalità sono assai rari.

Introduzione può essere vocale o strumentale e precede sempre il primo verso. Per la ricerca di versi e ritornello si ipotizza una lunghezza iniziale pari ad 8 battute e si effettua un pattern matching sul brano, in base al risultato e se non si ricade in nessuno dei pattern del punto 1 si può decidere se allargare la finestra a 16 battute e ripetere l'operazione. Il bridge è una parte del brano nella quale la progressione di accordi segue una chiave diversa da quella della canzone. Pertanto nell'analisi del brano una diversa chiave può evidenziare la presenza del bridge e comunque deve trovarsi in una delle posizioni del punto 1.2 o 1.3. La fine della canzone (outro) si calcola partendo dall'ultimo ritornello e la sua lunghezza sarà pari alla distanza tra la fine dell'ultimo ritornello e la fine del brano.

##### Risultati sperimentali

Sono stati scelti 50 brani di artisti famosi della musica popolare in lingua Inglese e su questi

effettuati test di *chord detection* (rilevazione di accordi), *singing-voice boundary detection* (rilevazione delle parti vocali e cantate) ed estrazione della struttura del brano. Si sono utilizzati due metodi per verificare la bontà dei risultati (vedi tabella 1):

#### 1. Identification accuracy (I.A.)

Rapporto percentuale sul numero di parti individuate rispetto al numero effettivo. Ad esempio se sappiamo che una canzone ha 3 ritornelli e il sistema ne identifica solo 2 avremo una identification accuracy pari a  $2/3 * 100 = 66,6\%$ .

#### 2. Detection accuracy (D.A.)

Riguarda l'accuratezza percentuale media con la quale vengono rilevate le varie parti di una canzone.

	Intro	Verso	Rit.	INST	Bridge	M.Eight	Outro
I.A.	100	84,29	86,51	90,78	76,19	88,96	97,38
D.A.	96,22	77,48	79,16	80,18	71,48	83,73	87,86

Tabella 1: Risultati sperimentali. Dati tratti da [1].

Come si può notare abbiamo un'elevata precisione nel rilevamento dell'intro e dell'outro mentre troviamo una scarsa accuratezza nel rilevamento del bridge il resto è più che soddisfacente.

## Applicazioni

Come accennato nella parte introduttiva, la conoscenza della struttura di un brano trova applicazioni in svariati campi:

### Trascrizione musicale e identificazione del testo di un brano:

L'identificazione del ritmo e delle regioni vocali/strumentali rappresentano la prima fase sia della trascrizione musicale che dell'estrazione del testo. Sapendo che le frasi in un brano vengono costruite sulla sua struttura ritmica possiamo utilizzare la BSS per identificare i limiti delle parole, allo stesso modo sapendo che la parte melodica di un brano è costruita sulla progressione di accordi risulta più semplice identificare le note che compongono la melodia.

### Sunto di un brano:

La creazione di sunti concisi e dall'alto contenuto informativo è essenziale per applicazioni su larga scala. Attualmente tali sunti sono creati a mano utilizzando le parti di un brano più ripetitive in modo che il pubblico se lo ricordi più facilmente. Con questo metodo è possibile, ad esempio, identificare in modo facile il ritornello di una canzone che è la parte di più rapida memorizzazione e associazione per l'ascoltatore. Quindi siamo in

grado di creare dei buoni sunti.

### Ricerca di brani musicali (M.I.R.<sup>12</sup>):

Il numero sempre più elevato di grandi archivi musicali richiede un metodo facile e rapido per la ricerca e la consultazione dei loro contenuti. Sono possibili dunque ricerche per similarità di parti di brano, ad esempio il ritornello è spesso conosciuto da chi cerca un brano pertanto la ricerca può essere effettuata direttamente su quell'attributo. Oppure la ricerca per accordi diventa estremamente semplice, estraendo la struttura con questo sistema si ottiene la progressione di accordi per ogni parte del brano, sulla quale si possono effettuare le ricerche volute.

## Conclusioni

Il metodo proposto si pone ad un livello superiore rispetto ai predecessori, mescolando informazioni musicali di alto livello, euristiche su un genere musicale e componenti di basso livello dell'analisi dei segnali i risultati ottenuti sono incoraggianti. L'approccio mira ad estrarre gli ingredienti della struttura di un brano, essenziali allo sviluppo di altre applicazioni.

### Conclusioni personali

Ho trovato il lavoro molto ben documentato e sicuramente interessante. Da musicista l'idea di ottenere la struttura di un brano in modo automatico mi affascina e mi semplificherebbe non poco la vita. D'altro canto non ho trovato un'applicazione dell'autore per testare il contenuto del paper, non ho trovato un esempio di ipotetico "output" come non ho trovato studi rispetto ad altri generi musicali. Sarebbe interessante capire se tale metodologia, variando le euristiche, si plasmi correttamente ad altri generi musicali e questo non solo in via teorica.

## Riferimenti

- [1] Namunu C.Maddage - *Automatic Structure Detection for Popular Music* - Institute for Infocomm Research, January - March 2006
- [2] C. Duxbury, M. Sandle, and M. Davies - *A Hybrid Approach to Musical Note Onset Detection* - Proc.Int. Conf. Digital Audio Effects, 2002
- [3] Alexander Sheh and Daniel P.W. Ellis - *Chord Segmentation and Recognition using EM-Trained Hidden Markov Models* - Proc.Int. Conf. Music information retrieval, 2003
- [4] A. Shenoy, R. Mohapatra, and Y. Wang - *Key Detection of Acoustic Musical Signals* - Proc. IEEE Int'l Conf. Multimedia and Expo, 2004
- [5] Changsheng Xu, Namunu C. Maddage, Xi Shao, Gi Tian - *Content-Adaptive Digital Music Watermarking Based on Music Structure Analysis* - Institute for Infocomm Research, 2007