INTRODUZIONE AL PHASE VOCODER

Rajmil Fischman (da Organized Sound N. 2, 97)
Traduzione e rielaborazione di E. Giordani

Laboratorio Elettronico di Musica Sperimentale
Conservatorio Statale di Musica G. Rossini - Pesaro

Il Phase Vocoder (PVOC) è un processo matematico per convertire un segnale campionato in una rappresentazione spettrale tempo-variante. Esso è computazionalmente efficiente quando viene implementato con la trasformata rapida di Fourier (FFT). Se il suono viene ri-sintetizzato dai dati di analisi, l'uscita sarà virtualmente identica al suono originale. Il PVOC offre alcuni vantaggi in più rispetto al filtro eterodina nell`analisi tempo-variante dei suoni.

Premessa storica

Il termine *vocoder* deriva dall'unione delle parole "Voice" e "CODER", il nome di un dispositivo disegnato per ridurre la larghezza di banda necessaria per una soddisfacente trasmissione del parlato sulle linee telefoniche. L'idea era di far passare il segnale del parlato attraverso un insieme di filtri passa-banda contigui, così che l'uscita combinata di tali filtri in un certo punto del tempo avrebbe approssimato lo spettro del segnale d'ingresso. In teoria, attraverso la trasmissione di soli pochi coefficienti dei filtri, si sarebbe potuto ottenere un certo risparmio di informazione trasmessa. In pratica tale risparmio non fu possibile perchè erano necessari un numero molto grande di canali (e quindi di filtri) al fine di preservare l'intelleggibilità del segnale. C'era anche un problema ulteriore nella ricostruzione poichè l'informazione di fase di ciascun filtro non veniva inviata e quindi l'informazione ricostruita non era mai identica all'originale, indipendentemente dal numero di filtri usati.

Il *phase vocoder*, sviluppato come una estensione del concetto originale di vocoder, preserva l'informazione di fase e consente all'ingresso e all'uscita di essere praticamente identici. Tale procedimento fu sviluppato negli anni da vari scienziati tra cui vale la pena di ricordare Schafer e Rabiner (1973) , Portnoff (1976,1978). Holtzman nel 1980 migliorò notevolmente la tecnica col risultato che la velocità di trasmissione fu aumentata (in termini di tempo di calcolo) consentendo allo stesso tempo di mantenere identici l'ingresso e l'uscita.

Phase Vocoder Musicale

In campo musicale il Phase Vocoder (PV) viene impiegato come uno strumento di analisi e ri-sintesi del suono . Esso deriva da un'estensione dei principi dell'analisi di Fourier (1768-1830). Come è già noto, l'analisi di Fourier afferma che ogni segnale periodico può essere rappresentato dalla somma di onde sinusoidali dette *armoniche,* dove ogni harmonica possiede una specifica ampiezza frequenza e fase. Inoltre se la frequenza del segnale è f, le frequenze delle armoniche saranno multiple di f (ovvero f1, 2f1, 3f1, 4f1, etc.). L'insieme delle armoniche che rappresentano un particolere segnale è definito come il suo *spettro*. Per questa ragione, è possibile riprodurre un segnale periodico noto usando una appropriata combinazione di armoniche ottenute dalla sua

analisi. È anche possibile immaginare nuove sonorità e sintetizzarle inventando nuove combinazioni spettrali.

Tuttavia, molti suoni naturali non sono periodici. Infatti, è proprio questa mancanza di periodicità stretta che conferisce ad essi una caratteristica di vivacità. Quindi, l'analisi dei suoni naturali richiede un miglioramento del metodo classico di Fourier. Un modo di trattare questo problema consiste di immaginare che il segnale cambi il suo periodo ad ogni istante; cioè, ad un certo momento, può essere usato un insieme differente di armoniche al fine di rappresentarlo. Alternativamente, è possibile considerare che il segnale sia composto di un particolare insieme di sinusoidi, dette genericamente *componenti parziali,* le cui ampiezze e frequenze cambiano nel tempo. Indipendentemente del metodo adottato, il risultato finale è uno spettro che cambia nel tempo. In questo caso, si dice che il segnale possiede uno *spettro dinamico* (o spettro *tempo-variante*).

Il modello tempo-variante di Fourier può essere usato al fine di ottenere tecniche che possono cambiare le caratteristiche spettrali dei suoni. Per questo proposito, un segnale può essere analizzato attraverso la sua decomposizione in un insieme di sinusoidi. Queste poi possono essere processate da meccanismi che alterino le loro ampiezze e frequenza. Infine, le componenti processate possono essere impiegate al fine di ri-sintetizzare un nuovo segnale.

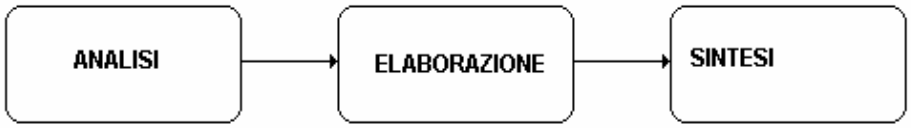

Figura 1. Tipico ciclo di processamento di Fourier.

Processo di analisi

Il primo passo di un ciclo di processamento di Fourier è l'analisi. La sua implementazione presenta due problemi che devono essere risolti. In primo luogo, il meccanismo di analisi deve essere in grado di individuare ciascuna componente singola in un certo istante. In secondo luogo, deve tenere in conto del fatto che lo spettro cambia nel corso del tempo.

Un passo in avanti verso la risoluzione di questo problema può consistere in un dispositivo che è in grado di riconoscere e isolare la singola componente. Ciò può essere raggiunto attrafverso un filtro *passa-banda* . Se una certa componente di frequenza f_i cade all'interno della banda passante bw del filtro e tutte le altre componenti sono al di fuori di tale banda, solo tale componente passerà. Se f_i è presente, la sua ampiezza sarà ottenuta direttamente all'uscita del filtro.

Lo stesso processo può essere applicato alle componenti rimanenti usando un banco di filtri le cui frequenze centrali sono distribuite su tutta l'estensione delle frequenze.

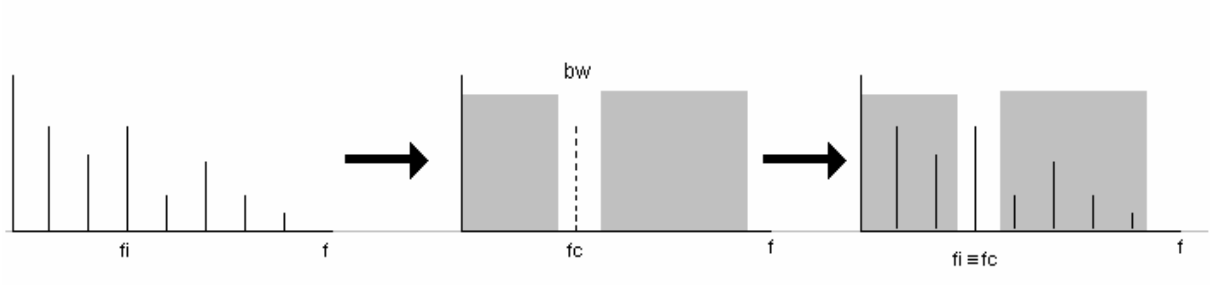

Figura 2. Impiego di un filtro passa-banda per isolare una componente sinusoidale a frequenza fi

Quando il segnale viene fatto passare attraverso il banco di filtri, lúscita di ciascun filtro lascia passare una componente differente.

Idealmente, il secondo problema può essere risolto attraverso la valutazione dello spettro ad ogni istante usando il banco di filtri. In pratica, lo spettro può essere campionato ad intervalli regolari, così spesso come è necessrio per individuare i cambiamenti significativi.

Individuazione delle componenti

La determinazione della larghezza di banda dei filtri richiede qualche attenzione. Se essa è troppo grande in confronto con la spaziatura tra le parziali, passeranno attraverso di essa più componenti. D'altra parte, larghezze di banda più strette richiedono un numero maggiore di filtri al fine di coprire l'intera estensione dello spettro, aumentando così il numero di calcoli da eseguire.

In ambiente numerico è noto che, detta *sr* (sampling rate) la frequenza di campionamento, è possibile esprimere correttamente un segnale qualunque attraverso la conoscenza di suoi *sr/2* campioni, dove sr/2 è detta *Frequenza di Nyquist*.

Per un segnale che appare all'uscita di un certo filtro il periodo $\tau = 1/f$ (dove f è la frequenza del segnale stesso) è definito come:

$$\tau = n\ \tau_s$$

dove $\tau_s = 1/sr$ è l'intervallo di tempo tra due campioni successivi

e $n =$ numero di campioni per periodo

La frequenza f del segnale sarà allora:

$$f = sr/n$$

e quindi

$$n = sr/f$$

Quindi dobbiamo assicurare che sia processato un numero sufficiente di campioni per essere in grado di individuare la parziale a frequenza più bassa. Per esempio, se la frequenza più grave è 40 Hz e stiamo campionando a 4 kHz, allora necessitiamo di almeno 4000 / 40 = 100 campioni per tale componente. Tuttavia, se il sr = 40 kHz tale quantità sarà 10 volte maggiore, cioè 1000.
Un banco di filtri può essere implementato usando una *Fast Fourier Transform (FFT)* . Questo procedimento converte un blocco di 2^N campioni in una rappresentazione spettrale che è concettualmente equivalente all'uscita di un banco di filtri passa-banda linearmente distribuiti come mostrato nella figura 3

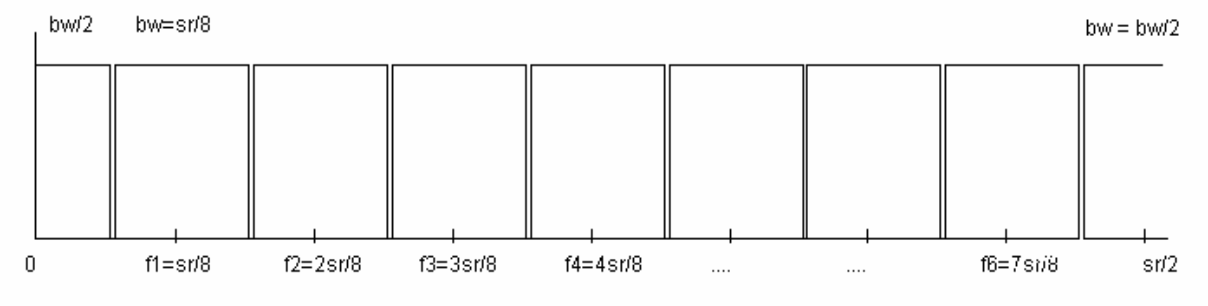

Figura 3. Un banco di filtri che rappresentano una FFT su 16 campioni

Tipicamente, tutti i filtri hanno la stessa larghezza di banda bw, ad eccezione della prima, che è un passa-basso con frequenza di taglio pari a bw/2, e dell'ultima, che è un passa-alto con frequenza di taglio pari a sr/2 - bw/2. I filtri rappresentano quindi i *canali* . È importante notare che l'algoritmo impiegato (*radix 2 FFT*) consente di analizzare blocchi di campioni pari a potennze di due (2,4,8,16,32,....).
Al fine di di rendere efficace l'impiego del banco di filtri, la frequenza della parziale a frequenza più bassa deve coincidere con la frequenza centrale del primo filtro passa-banda. Se il numero di campioni è 2^N allora la frequenza centrale del primo filtro è :

$$f1 = sr/2^N = sr/(2 * 2^{N-1}) = Nyquist/2^{N-1}$$

A questo punto possiamo calcolare la larghezza di banda dei filtri se si assume che la frequenza centrale è esattamente nel centro del filtro passa-banda. Si ha quindi che:

n_{bp} = (Nyquist - largh. Banda Low Pass - largh. Banda High Pass)/ largh. Banda

ovvero:

$n_{bp} = (sr/2 - bw/2 - bw/2) / bw = (sr/2 - bw) / bw = (sr/2 - sr/2^N) / sr/2^N = 2^{N-1} - 1$

dove n_{bp} = numero di filtri passa-banda

In conclusione, se consideriamo la somma dei due filtri passa-basso e passa-alto come un ulteriore passa-banda possiamo affermare che:

Il numero di campioni usati al fine di analizzare un segnale determina il numero e la larghezza dei filtri passa-banda. Usando una FFT a 2^N campioni, si divide lo spettro in 2^{N-1} canali di larghezza pari a Nyquist/2^{N-1}

Individuazione dei cambiamenti spettrali

Come detto in precedenza, al fine di inseguire i cambiamenti dello spettro di un segnale che muta nel tempo, è necessario campionare tale spettro ad intervalli appropriati.

Il metodo più semplice consiste nell'applicazione della FFT a gruppi consecutivi di campioni. Se queste istantanee spettrali sono grandi, e regolarmente intervallate, una grande quantità di dettagli tempo-varianti possono andare perduti. D'altra parte, se le istantanee sono prese troppo ravvicinate le une alle altre, il numero di campioni in ciascuna istantanea può diventare troppo piccolo per produrre una risoluzione accettabile tra le componenti dello spettro. Le istantanee spettrali sono definite come i *frames* e la frequenza alla quale viene campionato lo spettro (numero di frames / sec) è detta *frequenza di analisi* .

Un propblema ulteriore sorge con l'isolamento di gruppi di campioni dal segnale originale. All'interno del segnale, il primo e l'ultimo campione del gruppo sono preceduti e seguiti da altri campioni. Tuttavia, quando i gruppi di campioni vegono analizzati in isolamento, questi campioni sono preceduti e seguitri da nulla; in pratica i gruppi appaiono iniziare e finire in modo brusco, producendo transitori apparenti nel segnale (si possono udire come dei clicks). Quindi lo spettro derivato dall'FFT sarà gravemente distorto dalla presenza dei bordi di ciascun gruppo.

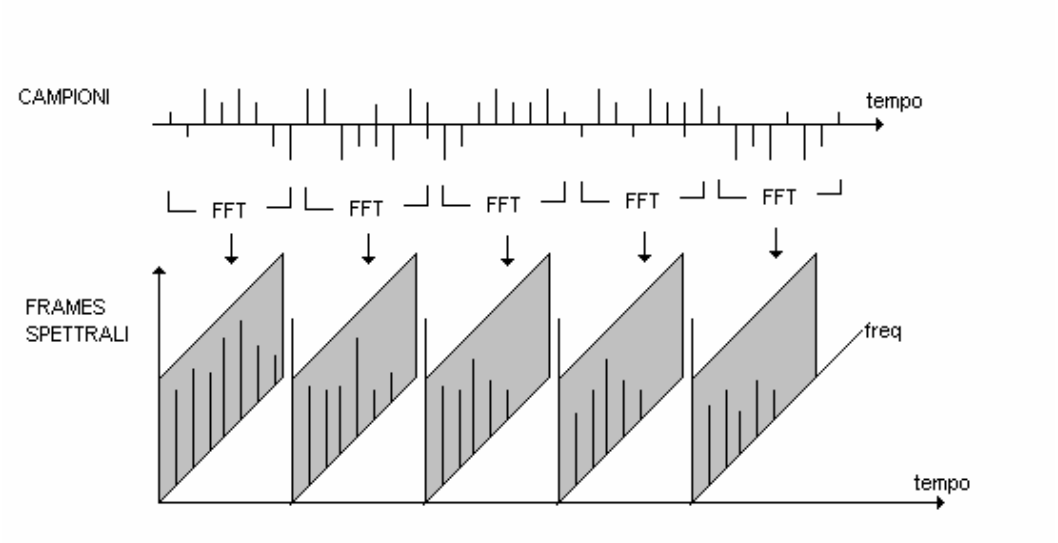

Figura 4 Trasformata di Fourier applicata a gruppi di 8 campioni successivi e contigui

Questo problema può essere attenuato se pochi campioni precedenti e seguenti sono rispettivamente trattati con una assolvenza e una dissolvenza (fade-in, fade out), ammorbidendo i bordi. In altre parole, i gruppi sono sovrapposti e ad esssi applicato un particolare inviluppo. La sovrapposizione inoltre aumenta la risoluzione della FFT, dal momento che è possibile analizzare più campioni nello stesso intervallo temporale.
Il processo di isolare un gruppo consecutivo di campioni è denominato *finestratura* (*windowing*), perchè ciò assomiglia all'azione di nascondere il segnale dietro una parete, impiegando una finestra scorrevole che restringe la vista ad un numero limitato di campioni alla volta. La forma dell'inviluppo applicata ai campioni determina il tipo di finestra (*window type*) e il numero di campioni la lunghezza della finestra (*window length*) . LA procedura che consiste di una FFT con finestra applicata ad intervalli regolari di tempo è conosciuta con il nome di *Short Time Fourier Transform* (STFT).

I tipi più comuni di finestra impiegate nella STFT corrispondono ai seguenti nomi: Bartlett, Blackmann, Hanning, Hamming e Kaiser. La prima di queste è un triangolo mentre Hanning è un coseno rialzato. Hamming e Blackmann sono varianti della forma a campana del coseno e Kaiser è una finestra molto più complessa che consente di eseguire controlli accurati sui filtri passa-banda della FFT. Se non viene applicato alcun inviluppo, la finestra è detta implicitamente rettangolare.

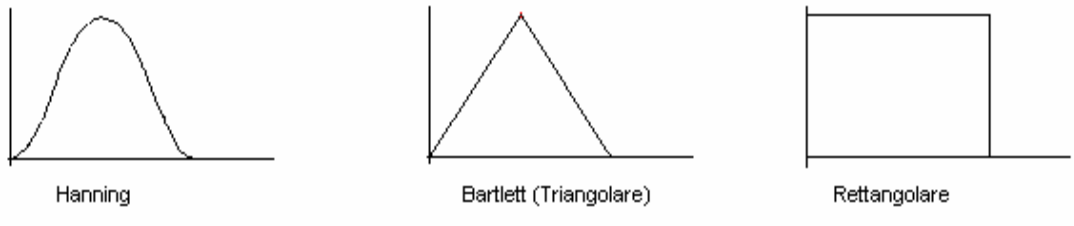

Figura 5 Tipi di finestre per la STFT

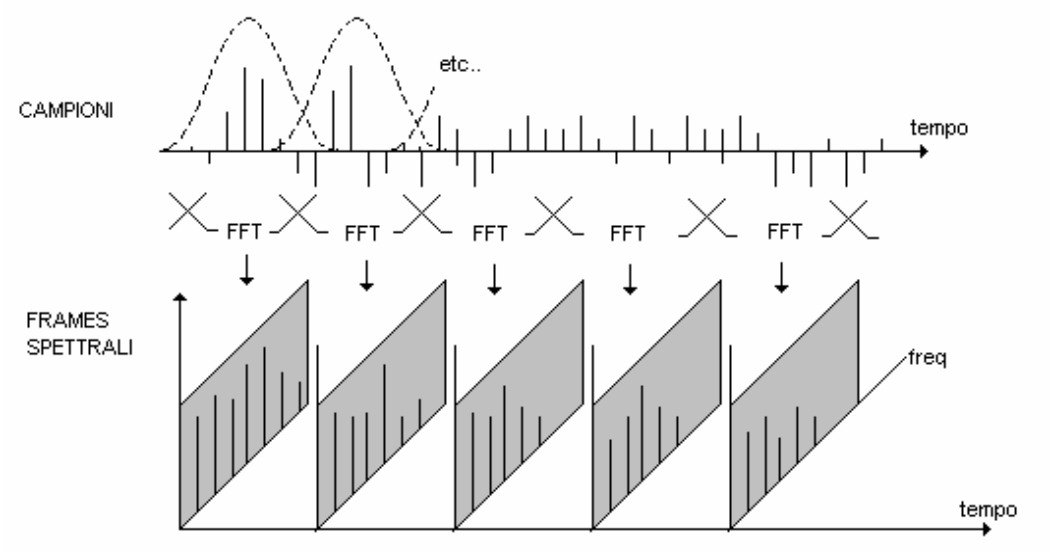

Figura 6 Trasformata di Fourier applicata a gruppi di 8 campioni successivi e contigui con *overlap* (sovrapposizione) e *windowing* (finestratura)

La quantità di overlap è misurata usando un *fattore di decimazione,* che indica il numero di sovrapposizioni attraverso la durata di una singola finestra. In figura 7 sono mostrati fattori di decimazione pari a 2 e a 4.

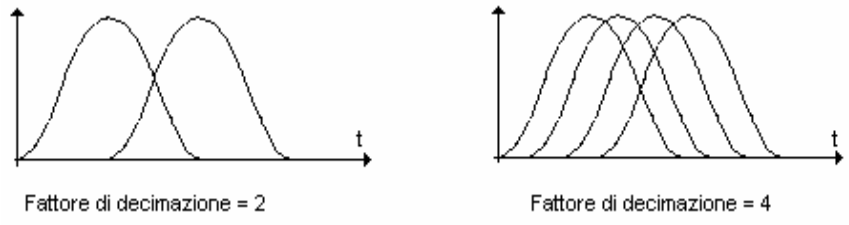

Figura 7. Esempi di overlap multipli di finestra

È possibile trovare una relazione tra la lunghezza della finestra e la frequenza di campionamento originale del segnale. Se la finestra non viene sovrapposta (fattore di decimazione D = 1), la frequenza di campionamento di analisi (asr) è semplicemente:

$$\text{asr} = \text{sr} / (\text{ lunghezza finestra}) = \text{sr} / 2^N$$

Se D > 1 (ovvero c'è sovrapposizione) allora l'asr viene di fatto aumentato. Per esempio, se D = 2, gli istanti di partenza delle finestre sono separati dalla metà della lunghezza della finestra stessa. Quindi la frequenza di analisi raddoppia in confronto con quella del caso senza sovrapposizione. Si può infatti scrivere che:

$$asr = D\ (sr / 2^N\ )$$

e quindi affermare che:

La lunghezza della finestra ed il fattore di decimazione determinano la risoluzione della STFT usata per misurare l'evoluzione spettrale di un segnale.
I filtri considerati all'inizio della trattazione, sono una idealizzazione della risposta in frequenza desiderata. Nei filtri reali, la transizione tra la banda passante e quella attenuata non è immediata : presenta una certa pendenza. Inoltre, la banda attenuata possiede una ondulazione (ripple) che permette ad una piccola parte del segnale di passare indisturbato. Questo fenomeno è chiamato perdita spettrale (*spectral leakage*) la cui quantità dipende dal tipo di finestra impiegata.

Fase di processamento

L'insieme di frames spettrali prodotti dalla STFT, fornisce una descrizione del modo in cui lo spettro di segnale cambia nel tempo. Questo tipo di informazione consente dir elaborare procedure che non sono possibili con altri mezzi. Per illustrare questo, possiamo immaginare un suono periodico che deve essere trasposto di un'ottava sopra. Questo potrebbe essere ottenuto campionando il suono ad una certa frequenza di campionamento e riproducendolo al doppio della sua frequenza di campionamento. Se dal punto di vista della frequenza ciò è corretto, porta però inevitabilmente al dimezzamento della durata effettiva del suono. D'altra parte, la durata di una STFT dipende dal numero di frames per secondo usati nell'analisi. Quindi se tale numero no viene alterato, la durata rimarrà la stessa. Nel caso in esame, tutto ciò di cui si necessita è di trasporre ciascun frame di un'ottava, la qual cosa viene realizzata moltiplicando per 2 ciascuna componente all'interno dei frames. Generalizzando questo processo, è possibile trasporre un segnale di qualunque rapporto senza alterare la sua durata. Ancora, l'altezza può variare nel tempo variando concordemente il rapporto di trasposizione.

Inoltre, il numero di frames può essere manipolato al fine di espandere temporalmente (time - stretch) un suono, alterando la sua durata senza per altro alterarne la sua frequenza. Per esempio, la sua durata può essere raddoppiata ripetendo ciascun frame, o dimezzandolo, omettendo ogni altro. L'espansione temporale può risultare come un effetto simile all'eco, che può divenire molto pronunciato, se il fattore di moltiplicazione diviene particolarmente alto. Molto spesso si ricorre a tecniche di interpolazione per prevenire questo fenomeno fastidioso.

Fase di risintesi

La FFT ha una sua controparte - la trasformata inversa di Fourier (IFFT) - che converte l'informazione spettrale in campioni. Quindi è possibile ottenere una procedura che trasformi le informazioni dell'analisi in un segnale campionato, ricostruito attraverso l'algoritmo IFFT.
Il successo di un processo specifico, dipende dalla scelta appropriata dei parametri della STFT. Abbiamo già visto che è importante assicurare che la larghezza di banda dei filtri FFT sia tale da garantire la necessaria risoluzione frequenziale. Data una risoluzione desiderata Δf, il numero di filtri necessari può essere determinato da:

$$n \geq Nyquist/\Delta f$$

per esempio, per ottenere una risoluzione di 10 Hz ad una frequenza di campionamento di 44.1 kHz, (Nyquisy = 22.05 kHz), il numero di filtri deve essere uguale o maggiore di 22050/10 = 2205: il numero potenza di 2 più vicino è 2048.

Un altro aspetto che determina la risoluzione spettrale è il fatto che il sistema uditivo lavora su scala logaritmica mentre la STFT è lineare. Ciò risulta in una scarsa risoluzione alle basse frequenze.

Seguendo l'esempio sopra, una risoluzione di 10 Hz può essere appropriata per segnali con componenti sopra i 100 Hz, nel qual caso l'errore risultante (5 Hz), è lo 0.5 %, corrispondente a meno di 1/12 di semitono temperato. Tuttavia, per una componente a 100 Hz, l'errore sale al 5% (vicino ad un semitono). A 50 Hz, l'errore è circa 2 semitoni e così via.

Infine, siccome il numero di canali è determinato dalla lunghezza della finestra di analisi, un grande numero di filtri richiederà lunghezze di finestra maggiori, producendosi così un minor numero di frames/sec. Ciò influenza negativamente la precisione dei cambiamenti spettrali. Nell'esempio sopre, il numero di campioni richiesti per ottenere 2048 (2 11) filtri è 4096. Assumendo un fattore di decimazione pari a 8, il numero di frame per secondo è 86. Raddoppiando la lunghezza della finestra questo numero passa a 43. Usando l'ultima finestra e dimezzando il fattore di decimazione si arriva a 22 frames /sec, che potrebbe essere troppo basso per evidenziare spettri velocemente variabili (transitori e attacchi).

Per sintetizzare, il numero di filtri scelti, è un compromesso tra la risoluzione frequenziale e temporale. La prima è determinata dalla larghezza di banda dei canali e la seconda dal numero di frames per unità di tempo.

Normalmente, suoni lentamente variabili sono meglio analizzati usando più alte risoluzioni spettrali, in particolare se essi sono quasi-periodici e non contengono alti contenuti di frequenza. Le basse frequenze richiedono pure elevate risoluzioni spettrali. I processi di time-stretching, possono favorire alte risoluzioni temporali, mentre pure operazioni spettrali, come la trasposizione, possono richiedere accurate descrizioni in frequenza.

In pratica, nelle applicazioni musicali, usando frequenze di campionamento di 44.1 kHz, possono essere richiesti da un minimo di 256 ad un massimo di 4096 canali.

Phase Vocoder e Csound

Tra i programmi di utilità di Csound, trova posto un applicativo di tipo *a linea di comando* denominato PVANAL che implementa l'analisi di file audio attraverso il processo STFT ed è stato scritto originariamente da Dan Ellis (Media Lab - MIT). La sintassi di lancio del programma è la seguente:

csound -U pvanal [flags] file-ingresso file-uscita oppure
pvanal [flags] infilename outfilename

Pvanal converte un file audio in una serie di frames STFT ad intervalli di tempo regolari (una rappresentazione nel dominio della frequenza). Il fil d'uscita può essere usato dal modulo di sintesi **pvoc**, per generare frammenti audio basati sul campione originale, con la possibilità di modificare la

scala dei tempi e della frequenza in modo arbitrario e dinamico. Vi sono anche un numero di nuove unità pvoc che sono in grado di impiegare tali files di uscita. Queste nuove unità sono state scritte da Richard Karpen e incluse nella versione di Bath (Csound 3.44). L'analisi è condizionata da un certo numero di flags (spazio opzionale tra il flag e il suo argomento).

FLAGS

-s<srate> frequenza di campionamento del file origine. Questo valore sostituisce la frequenza scritta nell'header del file, che altrimenti viene assunta. Se non sono presenti , il default è 10000.

-c<channel> numero del canale. Il default è 1.

-b<begin> tempo d'inizio (beginning time) in secondi del segmento audio da analizzare. Il default è 0.0

-d<duration> durata in secondi del segmento audio da analizzare. Il default di 0.0 significa fino alla fine del file.

-n<frmsiz> grandezza del frame STFT , ovvero il numero do campioni in ciascun frame di analisi di Fourier. Deve essere una potenza di due, nel range da 16 a 16384. Per ottenere un risultato pulito, un frame deve essere più lungo della durata del periodo della frequenza più grave presente nel segnale. Tuttavia, frames molto lunghi conducono ad una frammentazione temporale o riverberazione. La larghezza di banda di ciascun filtro STFT è determinato da freq. camp./ lunghezza del frame . Il valore di default di framesize è la più piccola potenza di due che corrisponde a più di 20 millisecondi della sorgente (es.. 256 punti @ 10 kHz, forniscono un frame di 25.6 ms).

-w<windfact> Fattore di sovrapposizione della finestre (window). Questo controlla il numero di trasformate di Fourier per secondo. Nella sintesi, pvoc interpolerà tra i frames, ma troppi pochi frames genereranno una distorsione udibile; troppi frames condurranno ad un file d'analisi enorme. Un buon compromesso per tale parametro è 4, significando che ciascun punto di ingresso compare in 4 finestre di uscita, o inversamente che l'offset tra frames successivi di STFT è framesize/4. Il valore di default è 4. Non si può usare questo flag unitamente al flag -h.

-h<hopsize> È l'offset tra i frames STFT. Contrariamente al precedente, specifica l'incremento in campioni tra successivi frames di analisi. Non si può usare questo flag unitamente al flag -w.

Esempio

pvanal asound pvfile

analizzerà il file "asound" usando i valori di default per frmsiz e windfact per produrre il file "pvfile" utilizzabile da pvoc.

FILES

Il file di uscita possiede uno speciale header contenente dettagli del file sorgente, il frame rate di analisi e l'overlap. I dati dei frames di analisi sono memorizzati come numeri decimali (float), con l'ampiezza e la frequenza (in Hz) per i primi primi N/2 + 1 filtri di Fourier di ciascun frame ciascuno. La frequenza viene codificata come incremento di fase così che in tale modo che per armoniche di forte intensità fornisce una buona indicazione della frequenza effettiva. Per basse ampiezza o per valori rapidamente variabili, questa indicazione è meno significante.

DIAGNOSTICA

Stampa il numero totale di frames, e i frames completati ogni 20.