SINTESI DI SUONI MEDIANTE INVERSIONE DEL MODELLO UDITIVO DI LYON.

P.Cosi (1), E. Zovato (2)

1) - Centro di Studio per le Ricerche di Fonetica, C.N.R. Padova
2) - Dipartimento di Elettronica ed Informatica, Università di Padova

SOMMARIO

Viene descritto un sistema di analisi e sintesi di suoni sviluppato mediante un modello del sistema uditivo periferico originariamente realizzato da R.F. Lyon. Anche se non esaustivamente in tutti i dettagli matematici, viene descritta la teoria su cui il modello è basato e le problematiche che sottostanno al processo di inversione utilizzato per ottenere la sintesi. Viene infine presentato un esempio di sintesi ottenuto con la procedura descritta.

INTRODUZIONE.

L'analisi dei segnali realizzata mediante i modelli uditivi ci permette di simulare il modo in cui il nostro sistema uditivo percepisce e successivamente elabora gli stimoli provenienti dall'esterno. Di particolare importanza risulta la conoscenza delle operazioni che, a livello corticale, avvengono sugli stimoli di natura neurale. E' probabile che essi subiscano delle autocorrelazioni; in tal modo vengono messe in evidenza eventuali periodicità presenti negli stimoli stessi. E' plausibile che sia proprio questo tipo di informazione che ci consente di distinguere segnali con *pitch* diversi emessi simultaneamente, oppure di riconoscere suoni anche in contesti rumorosi. L'utilità dei modelli uditivi risiede nel fatto che essi forniscono un potente strumento di analisi e quindi sono particolarmente adatti in applicazioni di riconoscimento vocale [1]. Inoltre le informazioni che si ricavano dall'analisi possono essere utilizzate con vantaggio anche per effettuare riconoscimento di segnali in contesto rumoroso, oppure per la separazione di segnali emessi simultaneamente.

SISTEMA DI ANALISI E SINTESI.

Il processo di analisi è costituito da due fasi principali. La prima è l'elaborazione realizzata mediante il modello di coclea passiva di Lyon [2-3]. La coclea viene scomposta in un numero finito di tratti che coprono tutta la sua lunghezza. Il modello prevede pertanto tante uscite quanti sono questi tratti; ogni uscita rappresenta il tipo di

stimolo nervoso che si può verificare in corrispondenza di una data posizione lungo la coclea. Tutte queste uscite forniscono una rappresentazione dei segnali che viene detta cocleogramma. La seconda fase dell'analisi consiste nel calcolare le autocorrelazioni delle uscite del cocleogramma. La nuova rappresentazione che si ottiene prende il nome di correlogramma.

Il modello della coclea passiva di Lyon è composto da tre stadi in cascata: lo stadio di filtraggio, lo stadio di rilevamento di energia, ed infine quello di adattamento e compressione. Il primo stadio è realizzato mediante una serie di filtri del secondo ordine in cascata, il cui numero dipende dalla frequenza di campionamento, (e quindi dalla banda disponibile), e da altri parametri quali il coefficiente di sovrapposizione delle bande dei filtri. L'andamento della funzione di trasferimento di ogni filtro risulta essere un picco di risonanza seguito da uno di antirisonanza. Le frequenze di risonanza decrescono esponenzialmente man mano che si procede lungo la cascata. In tal modo la funzione di trasferimento complessiva data dalla cascata di più stadi è di tipo passa banda; procedendo lungo la cascata tale banda si riduce e la frequenza di centro banda si sposta verso le basse frequenze. Dinnanzi ai filtri del banco, ci sono altri due filtri che simulano gli effetti dovuti all'orecchio esterno e medio.

Lo stadio di rilevazione dell'energia descrive la trasformazione che avviene allorquando il movimento della membrana basilare collegata alla coclea, prodotto dall'onda che si propaga all'interno della stessa, causa la produzione di scariche nervose. Nel modello tale stadio è realizzato mediante un rettificatore ideale ad una semionda, (HWR). In realtà si tratta di una semplificazione in quanto non viene tenuto conto degli effetti di saturazione; tuttavia tale soluzione risulta di immediata implementazione.

Lo stadio di compressione è composto da quattro blocchi di controllo automatico di guadagno, (AGC), in cascata. Si tratta di guadagni moltiplicativi che variano il loro valore in funzione dell'ampiezza dell'ingresso. Il guadagno per cui viene moltiplicato il campione del segnale di un canale è funzione del valore del campione precedente del canale stesso e dei canali ad esso adiacenti, e di una costante di tempo. I quattro stadi hanno costanti di tempo decrescenti, in modo tale da simulare i diversi tempi di adattamento del nostro sistema uditivo; in tal modo il primo stadio reagisce all'ingresso più lentamente mentre quelli successivi reagiscono via via più velocemente.

Calcolando le autocorrelazioni a tempo breve, (S*hort Time Autocorrelation*), delle uscite del cocleogramma si ottiene il correlogramma. Esso è costituito da un numero di frame che dipende dalla ampiezza e sovrapposizione delle finestre di analisi. Ogni frame contiene le autocorrelazioni di una determinata finestra temporale dei segnali di tutti i canali. In tal modo si ha la possibilità di ottenere informazioni relative al contenuto spettrale del segnale e inoltre si ha una buona indicazione del ritardo di autocorrelazione per il quale i segnali dei vari canali hanno la stessa periodicità. In figura 1 è raffigurato il procedimento di analisi con il quale si ottiene il correlogramma e come da questo si ottiene la sintesi. E' inoltre illustrato un esempio di frame: l'asse orizzontale rappresenta il ritardo di autocorrelazione, mentre quello verticale le frequenze centrali dei canali.

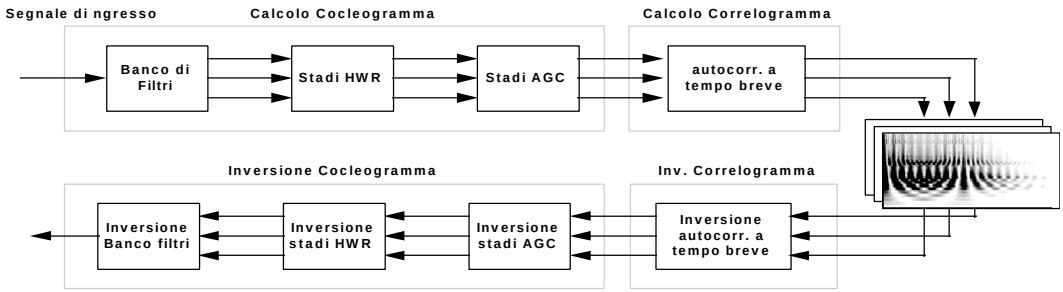

Fig.1 Schema di Analisi e Sintesi mediante modello uditivo di Lyon.

Nell'ottenere la sintesi mediante l'inversione del correlogramma prima, e del cocleogramma successivamente, si è seguita sostanzialmente la linea sviluppata da M. Slaney [4]. Le difficoltà insite in questo tipo di operazione sono dovute alle non linearità presenti nel modello di coclea passiva e all' inversione del correlogramma.

Dall'autocorrelazione di un segnale è possibile risalire al modulo della trasformata di Fourier del segnale stesso. Infatti, detta $R_{xx}(t)$ l'autocorrelazione del segnale *x(t)* e $X(w)$ la sua trasformata di Fourier, vale la seguente relazione:

$$\left|X(w)\right|^2 = \int_{-\infty}^{+\infty} R_{xx}(t) \cdot e^{-jwt}\, dt \qquad (1)$$

Quindi il problema consiste nel ricavare dei segnali a partire dai moduli delle trasformate di Fourier a tempo breve (*Short Time Fourier Transform*) di tutti i canali. Nessuna informazione è disponibile invece sulle fasi della trasformate. A tal scopo si è fatto uso dell'algoritmo iterativo di Griffin e Lim [5], il quale ricostruisce un segnale a partire da una sua stima iniziale (se disponibile), e dal modulo della sua STFT. Ad ogni iterazione viene diminuito l'errore quadratico tra il modulo della STFT della ricostruzione calcolata e il modulo della STFT nota. Con tale procedura si ricostruisce in qualche modo la fase, anche se il modulo non sarà più esattamente uguale a quello noto a priori. Tale algoritmo risulta essere particolarmente sensibile alla stima iniziale: una buona stima a fase non nulla permette di ridurre notevolmente il numero di iterazioni. A tal scopo le ricostruzioni dei vari canali sono state fatte in modo sequenziale partendo dal primo. La stima iniziale di ciascun canale è stata calcolata prendendo la ricostruzione del canale precedente e filtrandola con il filtro che in fase di analisi produce l'uscita che si va ora a ricostruire. Inoltre poichè è noto che i segnali da ricostruire possiedono solo la parte positiva, è stata effettuata anche una rettificazione della stima. Per il primo canale la stima iniziale è stata calcolata mediante la procedura di Roucos e Wilgus [6], ideata per ottenere stime iniziali efficienti per l'algoritmo di Griffin & Lim utilizzato in applicazioni di compressione della scala dei tempi. Tale procedura consiste nel sovrapporre e sommare le sequenze temporali, (ottenute mediante l'inversione delle sequenze del modulo della STFT), in modo sincronizzato cioè massimizzando la crosscorrelazione tra i dati. Questo espediente si rivela particolarmente efficace se applicato a segnali che presentano periodicità, come per esempio il parlato.

Ottenuta una ricostruzione del cocleogramma, si procede con la inversione dei tre stadi del modello di Lyon in ordine rovesciato. L'inversione degli stadi AGC consiste nel ricavare gli ingressi a partire dalle uscite divise per una quantità che in fase di analisi era il guadagno moltiplicativo. Quest'ultimo è esattamente ricavabile a partire dalle uscite per cui questa operazione non comporta particolari problemi.

L’inversione degli stadi HWR è più critica in quanto bisogna ricostruire la parte negativa dei segnali in modo coerente con le informazioni note a priori su di essi. A tal scopo si è fatto ancora ricorso ad una tecnica iterativa. Ad ogni iterazione si applicano alla stima di volta in volta trovata le caratteristiche note a priori del segnale. In tal caso esse sono i valori positivi del segnale e la sua banda di frequenze. Per quest’ultima proprietà bisogna operare un filtraggio della stima con un opportuno filtro passa banda. Nel sistema implementato si sono utilizzati gli stessi filtri utilizzati nell’analisi.

Infine si inverte il banco di filtri, cioè a partire dalle uscite di questo si deve ricostruire il segnale di ingresso. Viene calcolata una stima secondo l’espressione:

$$x'(t) = \sum_i y_i(-t) * h_i(t) \qquad (2)$$

dove $y_i(t)$ e $h_i(t)$ sono rispettivamente l’uscita e la risposta impulsiva del canale *i*-esimo, per cui per tale calcolo vengono utilizzati gli stessi filtri usati per l’analisi. In frequenza si ottiene la seguente espressione:

$$X'(w) = X(w) \cdot \sum_i \left| H_i(w) \right|^2 \qquad (3)$$

dove $X(w)$ è la trasformata di Fourier del segnale da ricostruire. Per ottenere il segnale voluto si deve pertanto effettuare un semplice filtraggio oppure si possono pesare opportunamente i segnali di uscita prima di filtrarli nuovamente.

In figura 2 è illustrato un confronto tra un segnale e la sintesi ottenuta dal suo correlogramma, ricavata mediante la procedura sopra descritta.

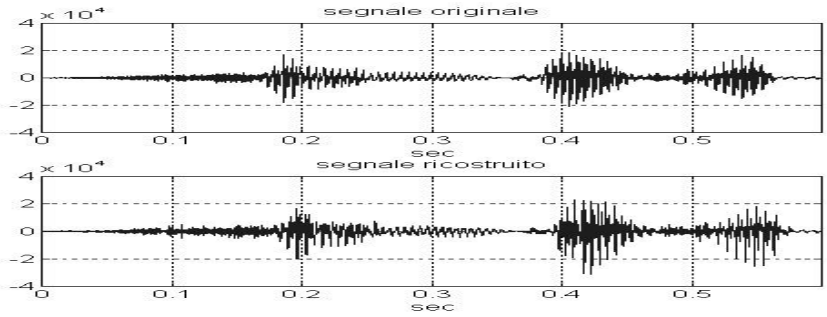

Fig.2 Confronto tra un segnale e la resintesi ottenuta dal suo correlogramma.

CONCLUSIONI.

Sono state condotte alcune prove di sintesi con le quali si è potuta valutare l’efficacia del sistema di inversione. E’stato subito rilevato che l’ inversione del cocleogramma avviene in modo pressocchè esatto con una percentuale di errore molto bassa. Maggiori errori invece si hanno nell’inversione del correlogramma. Tale fatto è imputabile alla completa assenza di informazione sulla fase. Prove di ascolto hanno tuttavia messo in evidenza che le sintesi prodotte risultano essere particolarmente fedeli ai segnali originali.

BIBLIOGRAFIA.

[1] Cosi P., “Auditory modelling for speech analysis and recognition”. in M. Cooke, S. Beet, M.Crawford (Eds.): *Visual representation of speech signals*, Wiley & Sons Chichester, 1993, pp. 205-212.

[2] Lyon R. F., “A Computational Model of Filtering, Detection, and Compression in the Cochlea.” *Proc IEEE-ICASSP*, 1982, 1282-1285.

[3] Slaney M., “*Lyon’s Cochlear Model*” (Techn. Rep. # 13) Apple Computer Inc. Cupertino, Ca., 1988.

[4] Slaney M., Naar D. and Lyon R.F., “Auditory Model Inversion for Sound Separation”, *Proc. IEEE-ICASSP*, Adelaide, 1994, II.77-80.

[5] Griffin D.W. and Lim J.S., “Signal Estimation from Modified Short-Time Fourier Transform”, *IEEE-ASSP*, 32, 1984, 236-243.

[6] Roucos S., and Wilgus A.M., “High Quality Time-Scale Modification for Speech”, *Proc. IEEE-ICASSP*, 1985, 493-496.