Indicizzazione di sequenze video attraverso l'analisi dell'audio sottostante

A. Bugatti, R. Leonardi, P. Migliorati, e L.A. Rossi

Dip. di Elettronica per l'Automazione – Università degli Studi di Brescia

Email: (bugatti, leon,pier, rossi)@ing.unibs.it

Sommario

In questo lavoro viene proposto un approccio innovativo all'indicizzazione di sequenze video basato sull'analisi dello stream audio sottostante. I metodi implementati permettono di segmentare l'audio in un insieme di classi omogenee con elevato contenuto semantico. Ciò consente di individuare situazioni altamente significative da un punto di vista semantico all'interno di documenti multimediali utilizzando solamente informazioni estratte attraverso l'elaborazione del segnale audio. Per far questo è stato sviluppato un algoritmo di segmentazione dell'audio in quattro classi omogenee (silenzio, voce, musica, rumore) basato sull'analisi di semplici caratteristiche temporali e frequenziali.

1. Introduzione

Adesso, e ancora più in futuro, ognuno avrà a disposizione un'enorme quantità di materiale multimediale proveniente da un'immensa varietà di sorgenti digitali. C'è quindi il bisogno di fornire gli strumenti per un'efficace navigazione attraverso questi documenti multimediali e nello stesso tempo realizzare delle tecniche di indicizzazione che permettano di ricavare tutte le informazioni d'interesse ai fini di un'efficiente recupero. È in questo contesto che si colloca il processo di standardizzazione dell'IS0 (International Standard Organization), iniziato nel'ottobre del 1996, per la descrizione di documenti multimediali, denominato MPEG-7 "Multimedia Content Description Interface" [2,3].

Quindi l'obbiettivo di molte ricerche attuali in questo campo è quello di individuare possibili procedure automatiche che consentano l'indicizzazione automatica di materiale audio-video. Questo permetterebbe di creare indici in grado di caratterizzare la struttura temporale dei documenti multimediali da un punto di vista semantico. La segmentazione potrebbe avvenire individuando un certo grado di coerenza fra segmenti consecutivi e, basandosi su di esso, dovrebbe poter essere possibile costruire una rappresentazione gerarchica dell'informazione, così da creare qualcosa di analogo al sommario che si trova all'interno di un qualsiasi libro tecnico. Una tale descrizione appare adeguata per permettere la navigazione all'interno di documenti, grazie alla possibilità che offre di potersi muovere secondo vari livelli semantici di rappresentazione dell'informazione.

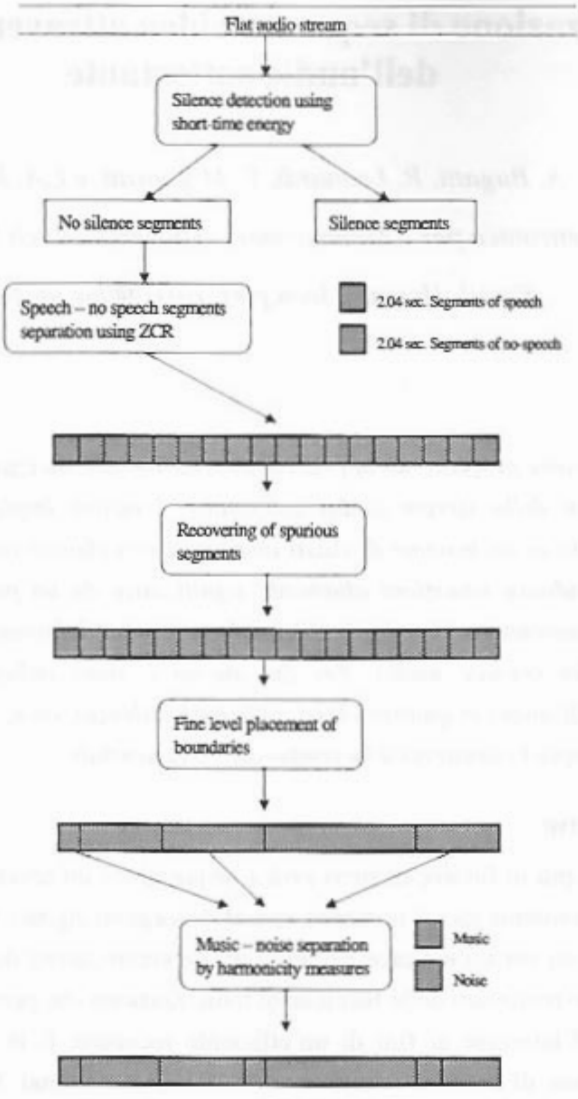

Fig. 1 L'algoritmo di segmentazione proposto.

2. L'indicizzazione di documenti multimediali attraverso un approccio basato sull'analisi dell'audio.

Tradizionalmente la creazione di indici di documenti multimediali è stata prevalentemente basata sul rilevamento automatico di cambi di camera fra shot e sui relativi effetti di editing video fra tali shot [1]. Questo tipo di approccio si è dimostrato generalmente soddisfacente da un punto di vista delle performance, fornendo una buona caratterizzazione di basso livello del contenuto visuale. Purtroppo però la semantica di questa segmentazione non risulta sufficientemente ricca, a causa dell'elevato numero di transizioni presenti all'interno di un qualunque programma.

In alternativa sono stati recentemente fatti degli sforzi per effettuare un'analisi dei documenti utilizzando congiuntamente le informazioni estratte sia dallo strem video che da quello audio. In [4,5] entrambi gli stream sono stati usati per identificare semplici tipi di scena che possono comporre un programma. L'analisi dello stream video può però essere molto pesante da un punto di vista computazionale (ad esempio si pensi all'analisi della correlazione fra due shot non consecutivi). Considerando l'alto livello di significato semantico che la sola analisi audio permette di ottenere, ci è parsa una buona soluzione quella di utilizzare solo quest'ultima per ottenere comunque degli indici caratterizzati da un elevato contenuto informativo. Quindi in questo lavoro proponiamo in insieme di metodi per l'indicizzazione che utilizzano solo lo stream audio: l'assunzione di base sulla quale si basa questo lavoro è che un segmento di un documento multimediale significativo da un punto di vista dell'audio lo è anche da un punto di vista del video.

3. L'algoritmo proposto

Seguendo quanto fatto in analoghe ricerche nel campo della segmentazione audio [6], il modello da noi proposto si compone di quattro classi distinte: silenzio, voce, parlato e musica. L'approccio adottato si basa su un'analisi a diversi livelli, nella quale per ogni livello è adottato un diverso algoritmo in modo tale da ottenere una segmentazione via via più definita [vedi Figura 1].

Come si può vedere il primo passo consiste nell'identificazione dei segmenti contenenti silenzio, che non subiranno poi alcuna ulteriore fase di analisi.

L'algoritmo da noi utilizzato per realizzare questa separazione si basa sull'utilizzo di un valore di soglia aggiornato dinamicamente. Considerando due classi rappresentate da vettori in $\Re_n$ lo scopo è di selezionare delle caratteristiche significative, in modo che vettori appartenenti a classi diverse possano essere associati a cluster diversi. La *short-time energy function* fornisce un'adeguata misura delle variazioni d'ampiezza del segnale ed è usata in parecchi lavori sull'individuazione di segmenti di silenzio. Per i nostri scopi questa feature risulta sufficientemente descrittiva e quindi il nostro problema di pattern classification si riduce al caso monodimensionale, dove ogni vettore è la misura della short time energy su segmenti della durata tipica compresa fra i 20 e i 40 millisecondi. Il classificatore divide lo spazio in due regioni R_1 e R_2: quando il vettore $\boldsymbol{x}$ cade all'interno della regione R_1 , $\boldsymbol{x}$ è classificato come appartenente alla classe $\mathbf{w}_1$, altrimenti $\boldsymbol{x}$ è classificato come appartenente alla classe $\mathbf{w}_2$. È chiaro che la scelta del valore della soglia che discrimina fra queste due regioni è fondamentale. L'algoritmo sviluppato stima il valore di questa soglia utilizzando la media del valore dell'energia di un numero sufficientemente alto di segmenti di silenzio e sommandogli un valore dipendente dalla varianza della stessa: i segmenti il cui valore cade al disotto di questo valore sono classificati poi come silenzio. Questa stima viene aggiornata dinamicamente per evitare errori dovuti ad un cambiamento della statistica del segnale: l'unica assunzione necessaria è che la statistica dell'energia del silenzio non abbia cambiamenti troppo bruschi nel tempo. Per migliorare le prestazioni viene utilizzata un'ulteriore caratteristica del silenzio: quando un frame è circondato da frame di silenzio la probabilità che anch'esso sia un frame di silenzio è significativamete più alta che nel caso contrario. Per tenere conto di ciò è stato introdotto l'utilizzo di una macchina a stati.

Una volta isolati i segmenti di silenzio, l'audio deve ancora essere ulteriormente separato in voce, musica e rumore.

Per far ciò il primo passo da noi adottato è stato quello di separare la voce dalle due rimanenti classi utilizzando una caratteristica tipica del segnale vocale. Quest'ultimo è infatti composto da una successione di suoni vocalizzati alternati a suoni non vocalizzati (tipicamente le consonanti). Mentre i primi sono caratterizzati da un alto contenuto energetico prevalentemente alle basse frequenze, i secondi hanno un comportamento noise-like, con l'energia maggiormente distribuita alle alte frequenze. Sia la musica che il rumore non presentano questa alternanza del comportamento energetico in frequenza. In analoghi lavori viene utilizzato lo Zero Crossing Rate, in quanto caratterizza in maniera significativa questo comportamento [vedi Figura 2] ed è semplice ed efficiente da un punto di vista computazionale.

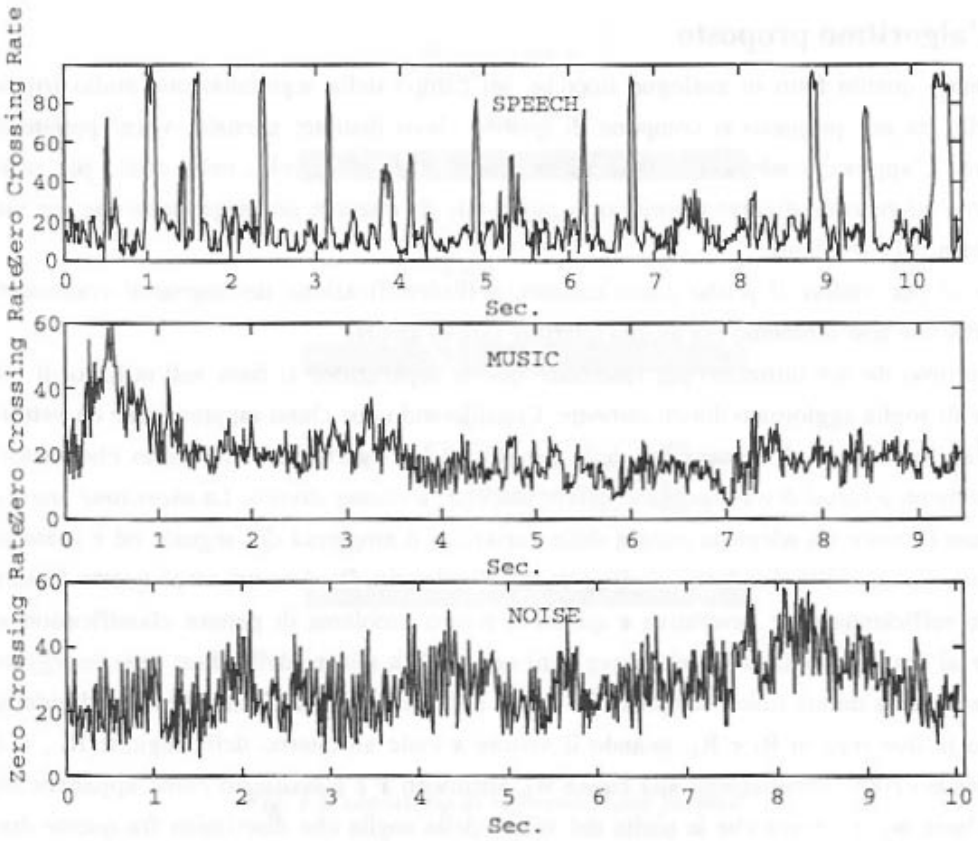

Fig. 2 Andamento dello ZCR su segmenti di voce, musica e rumore.

Per classificare i segmenti di voce il file audio viene partizionato in segmenti della durata di 2.04 secondi, ognuno dei quali è composto da 150 frames non sovrapposti. Questi valori permettono di avere un numero statisticamente significativo di frame e ogni frame (della lunghezza di circa 13 millisecondi) permette di ottenere un adeguato compromesso fra la quasi stazionarietà del segnale e un valore sufficiente per ottenere uno ZCR significativo.

Per ogni frame viene calcolato il valore dello ZCR, usando la definizione classica. L'insieme dei 150 valori che compongono un segmento sono usati per stimare le seguenti misure statistiche:

- Varianza: indica la dispersione rispetto al valore medio
- Momento del terzo ordine: indica il grado di asimmetria rispetto al valore medio
- Differenza tra il numero di valori al di sopra e al di sotto della media

Ogni segmento di 2.04 secondi risulta così associato ad un vettore 3-dimensionale. La separazione tra la classe del parlato e le altre due classi viene ottenuta utilizzando un classificatore gausiano multivariato.

Alla fine di questo procedimento otteniamo una serie di segmenti classificati come parlato o non-parlato. Basandoci sull'osservazione empirica che è più facile incontrare segmenti di parlato i cui vicini siano anch'essi segmenti appartenenti alla stessa classe, un'ulteriore aggiustamento viene fatto utilizzando un filtro mediano.

A questo punto i confini tra le differenti classi sono piazzati in posizioni fisse, a causa della natura dell'algoritmo precedentemente utilizzato. Per una più precisa collocazione degli stessi vengono ulteriormente processati i valori dello ZCR dei segmenti vicini appartenenti a classi diverse, ottenendo un nuovo segnale tramite l'applicazione della seguente formula:

$$y[n]=\frac{1}{P}\sum_{m=n-P/2}^{n+P/2}(x[m]-\bar{x}_n)^2 \quad \text{with} \quad P/2<n<300-P/2$$

dove x[n] è l'ennesimo valore dello ZCR dei due segmenti e xn è definito come:

$$\bar{x}_n=\frac{1}{P}\sum_{m=n-P/2}^{n+P/2}x[m]$$

Applicando un filtro passa-basso e successivamente un estrattore di picco a questo nuovo segnale si va a restimare la posizione del confine fra i due segmenti, ottenendo così un posizionamento migliore. L'ultimo passo consiste nella separazione in musica e rumore dei segmenti non ancora classificati. In generale la musica può essere considerata armonica, dove un suono armonico consiste in una serie di picchi in frequenza all'altezza della frequenza fondamentale e dei suoi multipli interi. In base a questa considerazione è possibile misurare un grado di armonicità per discriminare tra musica e rumore (difatti la maggior parte dei rumori sono non armonici). Per far ciò è stato applicato un semplice picth-detector che utilizza le misure dei picchi dell'autocorrelazione su finestre di 1024 campioni.

4. Risultati sperimentali

Vari pacchetti software sono stati sviluppati all'interno del nostro progetto. Un primo software produce la segmentazione del file-audio in segmenti di silenzio e non silenzio e i risultati ottenuti, verificati attraverso l'ascolto del file originale, non presentano errori significativi.

Un altro software per la separazione fra la musica e il parlato utilizzando l'algoritmo descritto sopra è stato sviluppato e testato sul materiale appartenente al *Content Set* di MPEG-7, ottenendo risultati con un'accuratezza superiore al 90%
I test sul software che integra tutte le fasi descritte nell'articolo sono ancora ad una fase preliminare, ma i primi risultati sembrano confermare la bontà di questa classificazione, avente un buon contenuto semantico.

5. Conclusioni

In questo lavoro è stato sviluppato un approccio alternativo per la segmentazione di documenti multimediali basato solo sull'analisi dello stream audio. I metodi proposti sono caratterizzati da una bassa complessità computazionale e sono così molto attraenti da questo punto di vista. I primi risultati mostrano che quest'approccio porta ad una segmentazione significativa, in accordo con la semantica del video associato.

6. Bibliografia

[1] N. Adami e R. Leonardi, "Identification of editing effects in image sequences by statistical modelling", *Proc. Picture Coding Symposium '99*, Portland, OR, U.S.A., Apr. 1999.

[2] MPEG Requirement Group. MPEG-7, "Context and objective", *ISO/IEC JTC1/SC29/WG11 N2460, MPEG98*, Atlantic City, USA, Oct. 1998.

[3] MPEG Requirement Group. MPEG-7, "Requirements", *ISO/IEC JTC1/SC29/WG11 N2461, MPEG98*, Atlantic City, USA, Oct. 1998.

[4] C. Saraceno e R. Leonardi, "Indexing audio-visual databases through a joint audio and video processing", *International Journal of Imaging Systems and Technology*, 9(5):320-331, Oct. 1998.

[5] C. Saraceno e R. Leonardi, "Identification of story units in audio-visual sequences by joint audio and video processing", *Proc. International Conference on Image Processing 1998*, Chicago, IL, U.S.A., Oct. 1998.

[6] T. Zhang e C.-C. Jay Kuo, "Audio-Guided Audiovisual Data Segmentation and Indexing", *IS&T/SPIE's Symposium on Electronic Imaging Science & Technology - Conference on Storage and Retrieval for Image and Video Databases VII*, SPIE Vol.3656, p316-327, San Jose, Jan. 1999