UN SISTEMA MULTIMODALE PER LA SEGMENTAZIONE DI TELEGIORNALI BASATO SULL'INDIVIDUAZIONE AUTOMATICA DEGLI SPEAKER

Leandro D'Anna[1], Gennaro Percannella[2], Carlo Sansone[3], Mario Vento[2]
[1]Dipartimento di Studi Linguistici e Letterari, Università di Salerno
Via Ponte don Melillo, 1 – Fisciano (SA), Italy
[2]Dipartimento di Ingegneria dell'Informazione e Ingegneria Elettrica, Università di Salerno
Via Ponte don Melillo, 1 – Fisciano (SA), Italy
[3]Dipartimento di Informatica e Sistemistica, Università di Napoli "Federico II"
Via Claudio, 21 – Napoli, Italy
ldanna, pergen,mvento@unisa.it;carlosan@unina.it;

ABSTRACT

Story segmentation is a basic step towards effective news video indexing.

All the solutions to story segmentation may be basically ascribed to two approaches. According to the first, segmentation is accomplished by directly finding story boundaries. Such boundaries are typically obtained by looking for the occurrences of some specific event (a sequence of black frames, the co-occurrence of a silence in the audio track and a shot boundary in the video track, etc.), or an abrupt change of some features at a high semantic level, as a topic switch.

The other approach performs story segmentation according to the following news program model assumption: given that each shot of the news video can be classified as an anchor shot or a news report shot, then a story is obtained by linking each anchor shot with all successive shots until another anchor shot, or the end of the news video, occurs.

According to the above news story model, automatic anchor shot detection (ASD) becomes the most challenging problem to partition a news video into stories.

In the scientific literature there are many papers that propose ASD algorithms. The majority exploits only video information, but in the last years the use of audio as an additional source of information for video segmentation has been rapidly raised up. Several proposals use audio features for directly individuating news boundaries, by means of a silence or a speaker change detector, in order to strengthen or to weaken the boundaries provided by the analysis based on ASD video techniques. Common drawbacks of such approaches are the use of supervised model-based techniques, that are not general enough, as they require an a priori definition and construction of an anchor shot model, and the unsystematic use of audio information. This is not effective due to its incoherence with video and then yields to a misleading shot classification.

In order to overcome these limitations, in this paper we present an algorithm for anchor shot detection with two audio/video stages.

The proposed algorithm first creates a set of templates in an unsupervised way. Each template represents a different anchor shot model within a video. In the second stage, a video similarity metric is used in order to retrieve a set of candidate anchor shots, which might have been missed by the first stage, and classify them by evaluating the audio similarity with respect to the templates. Finally the shots are classified by comparing them, from an audio point of view, with all the templates, when the news video is presented by only one anchorperson and with the closest template when two anchorpersons are present.

In order to choose the most suitable comparison without any a priori information about the actual number of anchorpersons, we also propose an automatic selector, based only on the audio track that is able to classify a news video as presented by one or two anchorpersons.

This automatic selector is based on the observation that each speaker is typically characterized by a specific distribution of the fundamental frequency (f0): the idea is then to verify if the f0 distribution calculated on all the audio samples belonging to a single speaker can be approximated by a given distribution (Log-logistic).

In the affirmative case, there is only one speaker; otherwise the number of speakers is at least two.

In particular, the audio similarity is calculated by comparing a candidate shot with all the templates extracted by the first stage. This is the best choice if the news video is presented by only one anchorperson.

On the contrary, a different criterion for defining audio similarity should be used when a news video is presented by two anchorpersons. In this case, in fact, only the audio track of the closest template (from an audio point of view), among all the extracted ones, should be used within the comparison with a candidate anchor shot. This is due to the fact that different templates can now refer to different anchorpersons and then their audio tracks no longer belong to the same speaker.

If the number of anchorpersons is known, the correct criterion to be applied can be a priori chosen. It happens if a broadcaster would employ the proposed method for analyzing all the editions of its news videos. In the general case, however, archiving companies work with large quantities of videos from different sources; moreover, we are interested in developing a completely unsupervised method.

The overall method proposed here has been tested on a database composed by several videos from two of the main Italian broadcasters. The performance of our method was evaluated in terms of Precision, Recall and F-measure. Moreover, we also compared our algorithm with respect to other state-of-the art anchor shot detection algorithm, achieving a significant performance improvement and so demonstrating its effectiveness.

SOMMARIO

La segmentazione in story è il passo fondamentale per una efficace indicizzazione dei telegiornali.

Tutte le soluzioni presenti in letteratura per la segmentazione possono essere suddivise in due tipologie di approccio al problema.

Secondo il primo approccio, la segmentazione viene realizzata contestualmente alla determinazione dei confini tra una story ed un'altra. Tali confine sono ottenuti, tipicamente, cercando le occorrenze o di alcuni specifici eventi (ad esempio una sequenza di frame neri, la presenza contemporanea di un silenzio nella parte audio e di un confine di shot nella traccia video, etc..) o il brusco cambiamento di features di livello più alto quali quelle di tipo semantico legate ad un cambiamento di topic.

L'altro approccio, invece, realizza la segmentazione in story sulla base di un modello di telegiornale assunto a priori: supposto che ogni shot di un telegiornale sia classificato come anchor shot o news report, allora una story è ottenuta collegando ogni anchor shot a tutti i successivi shot finché o un nuovo anchor shot viene individuato o termina il telegiornale.

Secondo questo modello di story, quindi, l'individuazione automatica degli anchor shot (ASD) diviene il problema fondamentale per la suddivisione di un telegiornale in story.

Nella letteratura scientifica ci sono molti lavori in cui vengono proposti algoritmi per l'ASD. La maggioranza di essi sfrutta solo le informazioni video, ma negli ultimi anni si è sempre più affermata anche l'importanza dell'audio come una sorgente suppletiva di informazioni.

In alcuni lavori le features audio vengono usate direttamente per l'individuazione dei confini di una notizia attraverso l'uso dei silenzi o di opportune funzioni per l'individuazione del cambio di speaker allo scopo, poi, di rafforzare o indebolire le indicazioni sulla presenza di confini ottenute mediante un'analisi in video.

Tuttavia questi approcci hanno due grossi inconvenienti: il primo è dato dal fatto che usano tecniche model based supervisionate e quindi non abbastanza generiche, il secondo è legato al fatto che le informazioni audio non sono utilizzate in maniera sistematica e quindi in maniera alquanto arbitraria.

Tutto questo causa una errata classificazione degli shot in tutti quei casi in cui c'è una incoerenza tra le due fonti informative.

Allo scopo di superare queste limitazioni, in questo articolo presenteremo un algoritmo per l'individuazione degli anchor shot a due stadi integrati audio/video.

L'algoritmo proposto inizialmente crea un insieme di template di anchor in maniera non supervisionata. Ogni template rappresenta un differente modello di anchor shot all'interno del video. Successivamente, attraverso una metrica di similitudine sulla base del video, si costruisce un insieme di shot possibili candidati ad essere degli anchor shot. Allo scopo di individuare possibili anchor shot mancanti, ogni elemento di questo insieme viene, poi, confrontato con tutti i template di cui sopra in cui è presente un solo anchor e con il template più vicino nel caso in cui ci siano più anchor sulla base di un criterio di confronto in audio.

Allo scopo di scegliere il più opportune metodo di confronto senza usare informazioni a priori, abbiamo sviluppato un modulo per la selezione automatica del numero di speaker sulla base della sola traccia audio.

Questo modulo si basa sulla osservazione che ogni speaker è tipicamente caratterizzato da una specifica distribuzione della frequenza fondamentale (f0): l'idea è che per verificare che una certa distribuzione della f0, calcolata a partire dall'insieme dei campioni audio di un certo shot, appartenga ad un solo speaker è che essa debba essere approssimata da una certa distribuzione (Log-Logistica). In caso affermativo tutti i campioni audio apparterranno ad un solo speaker altrimenti ad almeno due differenti.

In particolare la similarità in audio è calcolata confrontando lo shot candidato con tutti i templates estratti durante il primo stadio. Questa è la scelta migliore nel caso in cui nel video ci sia solo un anchor. Al contrario un criterio differente per definire la similarità audio deve essere usato quando nel video sono presenti più di un anchor. In questo caso, infatti, solo la traccia audio del template più vicino (dal punto di vista audio) potrà essere utilizzato durante il confronto poiché template differenti possono ora riferirsi a differenti anchor e quindi anche le loro tracce audio non saranno più di un solo speaker. Occorre notare che se il numero di anchor è conosciuto, il criterio corretto può essere applicato a priori senza usare il modulo proposto. Ciò avviene quando l'emittente televisiva vuole impiegare il metodo proposto per analizzare tutte le edizioni delle sue news. Nel caso più in generale, però, le aziende lavorano con grandi insiemi di telegiornali di differenti emittenti: in questi casi un metodo non supervisionato quale quello proposto risulta essere di gran lunga preferibile.

Il sistema proposto è stato sperimentato su di un database costituito da numerosi video di emittenti italiane. Le prestazioni dello stesso sono state valutate in termini di Precision,

Recall e F-measure. Infine abbiamo confrontato il nostro algoritmo con altri algoritmi allo stato dell'arte ottenendo significative performance e dimostrando così l'efficienza dell'algoritmo proposto.

1. PREMESSA

Una sequenza video è una ricca sorgente di informazione contenente parlato, musica, immagini, testo, ecc., quindi, per la natura stessa dei dati multimediali è difficile realizzare ricerche approfondite in vasti archivi basandosi sulle classiche interrogazioni testuali. Tutto questo significa che le risorse multimediali dovrebbero essere invece indicizzate, memorizzate e recuperate in un modo molto simile all'attività di processamento di tali informazioni da parte del cervello umano. In quest'ottica, una prima fase prevede la costruzione di indici di materiale audio/video prevalentemente realizzata manualmente mediante associazione di un limitato numero di parole chiave (keyword) all'oggetto di interesse. Una delle problematiche più interessanti da affrontare in questo settore sia da parte dell'industria dell'ICT e sia della comunità scientifica internazionale diventa allora quella dell'indicizzazione automatica dei contenuti multimediali messi a disposizione. Infatti, nonostante l'indicizzazione manuale sia correntemente il metodo più accurato, il processo automatico risulta molto appetibile in quanto fornirebbe uno strumento specialistico affidabile e a costo minimo. Tuttavia i processi finora sviluppati sono molto dispendiosi sia per l'elevato sforzo computazionale richiesto sia per lo spreco della risorsa tempo.

La fase finale di questo processo prevede l'uso di un sistema di elaborazione automatico in grado di selezionare in maniera diretta, attraverso un opportuno motore di ricerca applicato ad un database di informazioni "metadatate", le informazioni audio-video di particolare interesse in modo semplice ed intuitivo. Per la creazione automatica di tali database, è necessario, quindi, un utilizzo di opportuni sistemi automatici, basati su metodi innovativi, per analizzare e classificare i dati multimediali sia dal punto di vista dell'audio che dal punto di vista del video, con tecniche di analisi multimodale.

In particolare il nostro interesse è in particolar modo focalizzato al "news video processing", ovvero all'analisi, identificazione e catalogazione dei filmati di telegiornali. Infatti le "broadcast news" sono molto preziose per gli analisti governativi, per i fornitori di informazioni (es. le agenzie ANSA) e per gli utenti finali e, poiché nuovi eventi accadono tutti i giorni nel mondo intero, è umanamente "utopico" che una persona possa tener traccia contemporaneamente tutte le notizie di tutte le emittenti televisive. Per realizzare tutto ciò è quindi necessario predisporre di un sistema di "News Video Segmentation" capace cioè di segmentare un filmato di telegiornale in unità semantiche omogenee: le notizie.

2. DESCRIZIONE DELLA RICERCA SVOLTA

Il presente lavoro si occupa dell'analisi automatica dei filmati di telegiornale allo scopo di suddividere un TG in news stories, ossia in singole notizie coerenti da un punto di vista semantico. Nel caso particolare dei filmati di telegiornale considerando la sola parte video, una prima semplice caratterizzazione è data dall'alternarsi di scene relative alla presenza del solo cronista del giornale e di scene relative ai servizi.

Una prima separazione delle scene del filmato potrebbe dunque essere fatta riconoscendo la sola presenza del cronista (vedi figura 1); in tal modo, indirettamente, si ritroverebbero tutte le parti che appartengono ai singoli servizi, ottenendo così una prima

separazione del filmato in due distinte classi, che potremmo definire classe Cronista (Anchor shot) e classe News (Non anchor shot).

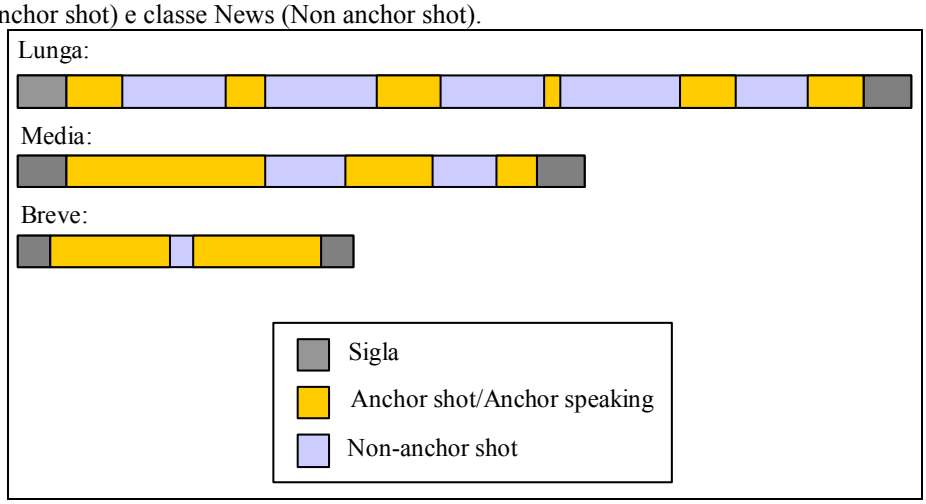

Figura 1: Esempi di suddivisione in anchor e news in un tg di diversa durata

Da un punto di vista dell'audio, un filmato di un telegiornale può essere strutturato come nella figura 2.

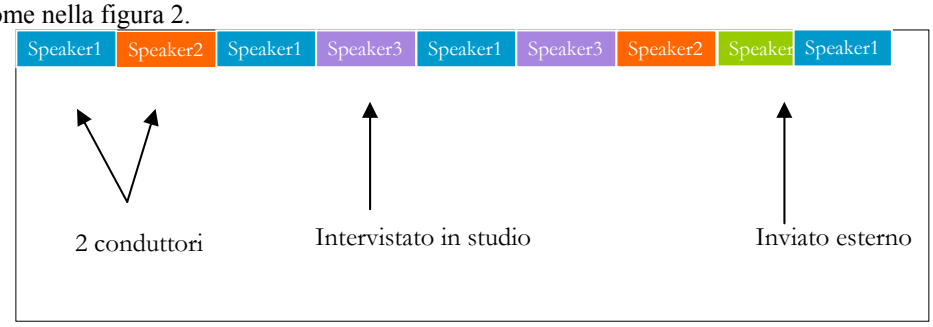

Figura 2: Esempio di suddivisione in turni dell'audio di un telegiornale

In questo caso si ha l'alternanza di diversi speaker che hanno una presenza nel filmato in percentuale molto diversa a seconda che siano speaker da studio o inviati esterni o intervistati in studio ad esempio. Tale alternanza suddivide la traccia audio in porzioni omogenee, dette anche turni, ognuna caratterizzata dalla presenza di un solo speaker.

In realtà, la singola segmentazione in video ed in audio è soggetta ad errori che impediscono di segmentare e classificare ogni segmento in maniera non ambigua. Inoltre anche per i segmenti univocamente identificati sia in video che in audio non sempre avviene una corrispondenza tra i confini temporali individuati dalla due fonti informative. Tutto questo rende impreciso il processo di costruzione di archivi di dati multimediali con la conseguenza che l'indicizzazione degli stessi non è sempre soddisfacente.

Dallo studio dello stato dell'arte, che è stato di fatto il punto di partenza del presente lavoro, possiamo suddividere i vari approcci per la segmentazione automatica di filmati di telegiornali in tre grosse categorie distinte in base a quale fonte informativa viene utilizzata (video e/o audio). Da ora in poi, escludiamo dalla nostra analisi tutti quei sistemi che utilizzano anche il testo (titoli e sottotitoli) essendo legati direttamente alla lingua utilizzata

nel telegiornale. All'interno dell'ultima categoria, ossia quelle che integra il video con l'audio, vi possono essere due approcci all'integrazione. Il primo prevede l'integrazione dell'informazione a priori ossia effettuando la stessa prima di effettuare la vera e propria segmentazione mentre il secondo approccio si basa sull'integrazione a posteriori delle singole segmentazioni indipendenti effettuate in video ed in audio.

3. DESCRIZIONE DELL'APPROCCIO

L'approccio integrato, utilizzato dal nostro sistema, nasce dall'esigenza di superare alcune delle grosse limitazioni presenti nell' approccio video e nell'approccio audio. In particolare nel primo caso risulta difficile associare allo stesso speaker inquadrature diverse. In questi casi un sistema basato sul video potrebbe considerare la porzione di filmato con un inquadratura diversa, se statisticamente poco occorrente, addirittura come quella appartenente ad un diverso speaker o anche ad un inviato da esterno. D'altro canto uno speaker da studio che commenta un filmato senza essere inquadrato sarebbe completamente omesso dal sistema basato sul video. Per quanto riguarda i sistemi basati solo sull'audio, poi, i maggiori problemi dipendono dalla estrema complessità delle caratteristiche acustiche. Questo significa che se da una parte si riesce abbastanza facilmente a distinguere il parlato dalla musica o dal rumore, diviene molto complesso identificare parlatori dello stesso sesso utilizzando clip di lunghezza non molto elevata e senza modelli addestrati a priori ed indipendentemente dalla lingua del parlante.

Queste considerazioni hanno quindi spinto molti gruppi di ricerca ad utilizzare una tecnica di segmentazione multimodale ossia che fa uso di entrambi i canali informativi servendosi così dei punti di forza di ognuno di essi. L'approccio utilizzato dal nostro sistema è dunque del tipo integrato a priori e si basa su un utilizzo integrato delle informazioni sulla segmentazione provenienti dai due canali. Una prima grossolana segmentazione in video viene successivamente raffinata da una successiva segmentazione in audio che a sua volta viene effettuata sulla base della precedente segmentazione in video. Tutto questo può essere ripetuto più volte fino ad ottenere la miglior segmentazione possibile senza alcuna scelta arbitraria.

In realtà, molto spesso è la componente audio che riesce a risolvere un'ambiguità presente in video. Consideriamo, ad esempio, il caso in cui in video compare un'esplosione caratterizzata da una rapida variazione della luminosità. La stessa scena può presentarsi durante un filmato lanciato durante un telegiornale o durante una pubblicità o durante un film di guerra. La sola analisi in video è quindi incapace di attribuire a quali tra queste categorie appartiene la scena in questione. Invece se consideriamo la stessa scena in audio potremmo notare che nel primo caso ci sarà sicuramente del parlato prima della scena in questione e questo parlato avrà una musica di sottofondo nel secondo caso. Nel terzo caso, poi avremmo una porzione abbastanza lunga in video caratterizzata da un volume molto più elevato rispetto alla media per evidenziare una scena d'azione. Quindi spesso l'audio è una fonte informativa da cui non si può prescindere per l'analisi automatica dei filmati.

In letteratura vi è utilizzata un'ampia e consolidata serie di features per risolvere il problema dell'anchor shot detection in audio. Sotto questo punto di vista, esiste una prima differenziazione tra features che possono essere di tipo più generale e meno precise nel descrivere uno speaker e features più specifiche ma meno generali che descrivono bene un singolo parlante. Le prime nascono con l'obiettivo di caratterizzare maggiormente una certa classe di suoni (ad esempio silenzio, parlato, rumore, musica,…) e sono legate più alle caratteristiche fisiche del segnale audio visto come un insieme di onde acustiche caratterizzate da una energia, una durata e uno spettro armonico ad esempio. Nel caso

specifico del parlato, oggetto della nostra ricerca, esse tendono a classificare più un gruppo di speaker che uno speaker in sé. Per quanto riguarda invece le features più specifiche, esse tendono a descrivere dettagliatamente un singolo parlante e la sua articolazione dei suoni fondamentali (foni). Tale features modellano la sorgente acustica, ossia le corde vocali e la conformazione degli organi fonatori, mediante una serie di coefficienti più o meno legati alle caratteristiche fisiche della stessa (quali spessore delle corde vocali, frequenze di risonanza, lunghezza del tratto glottideo, etc.). In tal modo si riesce a rappresentare il singolo parlante in maniera molto accurata. Quando, però, il numero degli speaker aumenta tali features risultano sempre più insoddisfacenti per effettuare una corretta discriminazione degli stessi.

Un ulteriore differenziazione delle features nasce da due considerazioni empiriche: in media un uomo utilizzando solo l'udito è capace di discriminare con una elevata precisione tutte le porzioni in cui compare uno stesso speaker all'interno di edizioni differenti di un telegiornale. Questo presuppone che questa capacità umana possa integrare informazioni provenienti da vari livelli temporali di analisi estendendo sia le informazioni specifiche a livello di suoni elementari (foni) e sia quelle legate alla semantica di quanto detto. Questo impone la necessità di individuare una varietà nella scelta delle features che vanno da quelle locali (20ms), calcolate su finestre di analisi a livello dei singoli foni, a features globali calcolate su finestre di analisi di dimensioni molto maggiori (>200ms).

4. DESCRIZIONE DEL SISTEMA

Il sistema sviluppato nasce con lo scopo di migliorare le prestazioni di un sistema già realizzato dal Mivia Lab dell'Università di Salerno basato anche esso sull'analisi del video e dell'audio con sole features di tipo locale. Tale sistema, quindi, rappresenta il sistema di base rispetto alla quale il sistema sviluppato si muove. La struttura generale del sistema in questione è riportato in figura 3 (per una descrizione dettagliata vedi De Santo et alii, 2005).

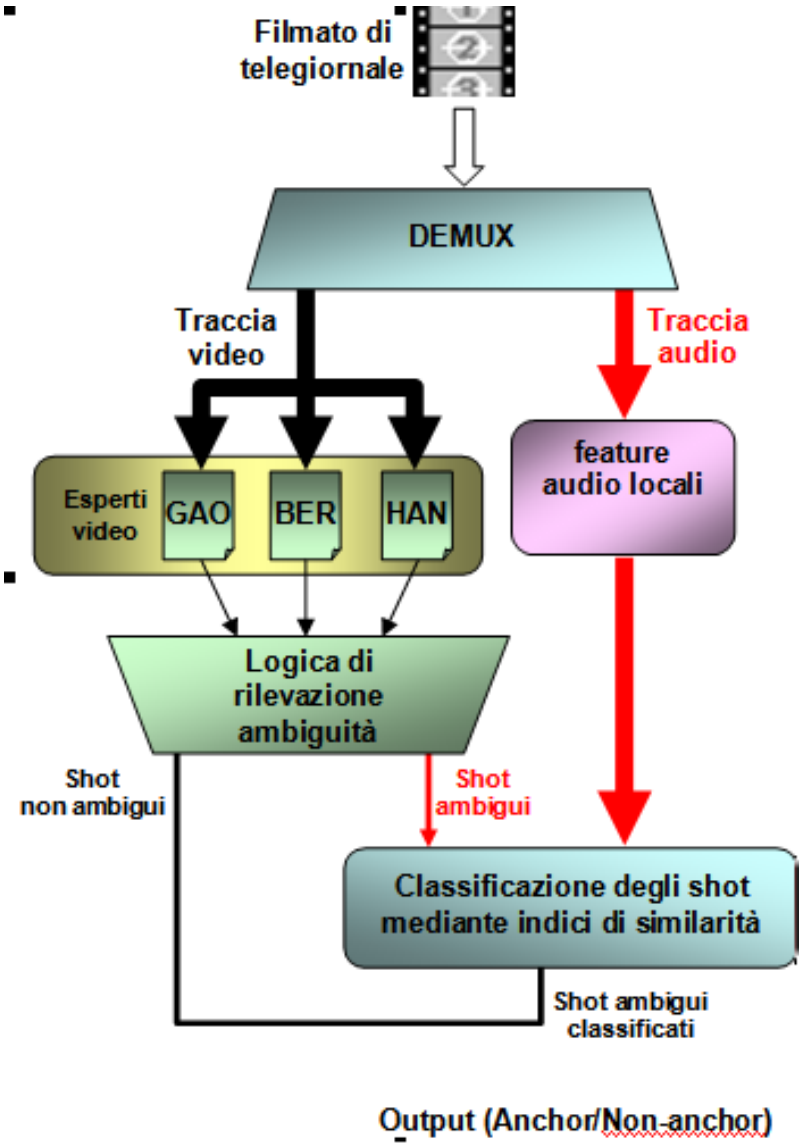

Figura 3: Architettura del sistema di base

Come si può osservare esso propone una modalità di analisi integrata a priori degli shot ambigui in video. Quindi, a partire da analisi video basata su multiesperti, si suddivide l'insieme degli shot di un telegiornale in tre classi: anchor shot certi (accordo totale multiesperti), news report certi (accordo totale multiesperti) e anchor shot dubbi.

A questo proposito l'approccio per l'analisi dell'audio utilizzato dal sistema di base si basa sull'ipotesi forte che le variazioni delle features locali siano specifiche del singolo speaker nel senso che ogni parlante sia caratterizzato da specifiche variazioni, diverse da quelle di un altro speaker, e che tali variazioni, quindi, permetterebbero di distinguerli. In realtà le features su base locale, proprio per il loro carattere, tendono a essere troppo

sensibili alle variazioni legate al contenuto sia sul piano fonetico che di parola. Tutto questo comporta una riduzione della capacità di discriminazione delle stesse.

Per superare tale limite nel sistema sviluppato è stato necessario introdurre sia una finestra di analisi di dimensioni temporali medie (100ms) o grandi (>1s) rispetto a quelle adoperate per le features locali e sia delle features globali più legate alle caratteristiche fisiologiche dello speaker e meno dipendenti dal contenuto linguistico. Dunque nel sistema sviluppato è stato sostituito il modulo per la classificazione dei frame ambigui basato sulle sole features locali con un altro che utilizza sia feature di tipo locale che feature di tipo globale per migliorare i risultati ottenuti dal sistema base. Inoltre, allo scopo di ridurre i tempi di computazione, è prevista un'analisi delle features più mirata, nel senso che tale analisi sarà limitata di volta in volta alle sole porzioni dell'audio interessanti per una certa feature.

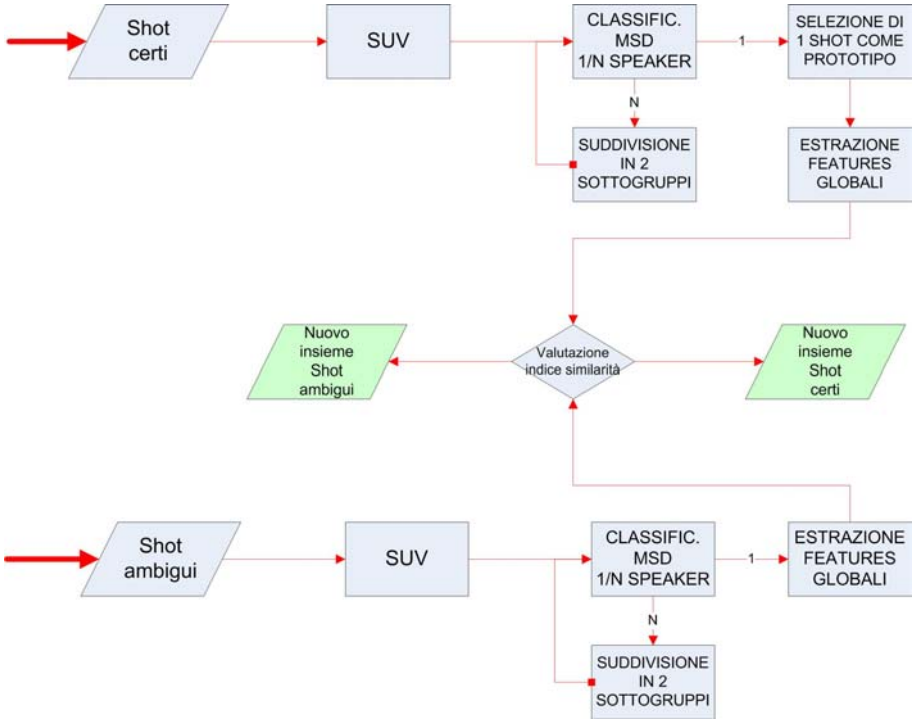

Figura 4: Architettura del sistema sviluppato

Nella figura 4, vi è uno schema che illustra l'architettura del sistema sviluppato (per una descrizione dettagliata dei moduli SUV e MSD vedi D'Anna et alii 2006). In particolare, esso agisce a partire da due set di anchor shot, uno certo l'altro ambiguo, individuati dal classificatore video ed il suo obiettivo è quello di raffinare i due insiemi di partenza fornendo così una segmentazione più accurata. Per far questo, si applica ai due insiemi prima un algoritmo (vedi figura 5) per l'individuazione delle porzioni salienti in audio (SUV)

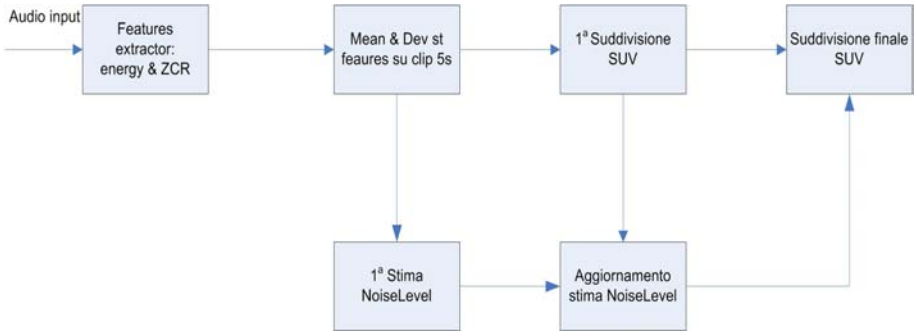

Figura 5: Schema di funzionamento dell'algoritmo SUV

e poi, sulla base della feature locale f0, si applica ad ogni shot un classificatore (MSD) che permette di discriminare se in quello shot c'è un solo speaker o più di uno (vedi figura 6).

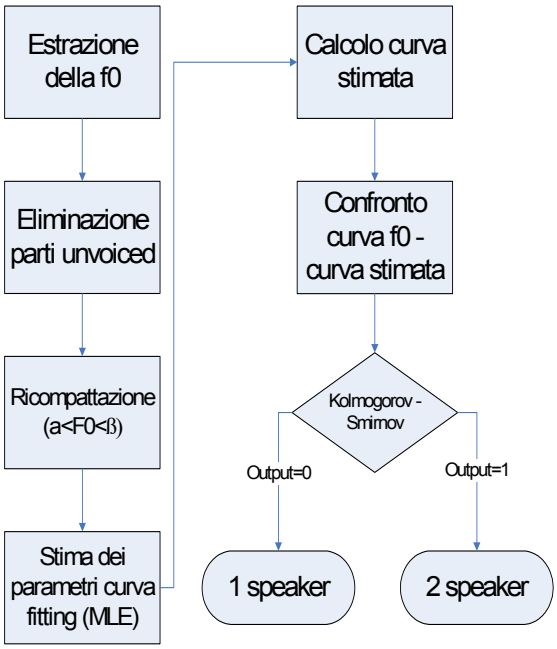

Figura 6: Schema a blocchi dell'algoritmo MSD.

Sulla base di questi risultati, per gli anchor shot certi, si costruiscono tutta una serie di prototipi relativamente a quegli shot in cui è stato identificata la presenza di un solo speaker. Per tutti gli altri shot, si suddivide ognuno in 2 sottogruppi di frame riapplicando l'algoritmo MSD. Il medesimo algoritmo si applica all'insieme degli anchor shot dubbi con l'obiettivo di separare in questi ultimi quelli in cui c'è un solo speaker in audio da quelli in cui c'è più di uno speaker. A questo punto, dopo aver estratto il set di features globali, per ognuno dei prototipi individuato sia negli anchor shot certi che in quelli dubbi, si effettua una valutazione dell'indice di similarità fissando un prototipo nell'insieme degli anchor

shot certi e confrontando quest'ultimo con tutti quelli individuati nell'insieme degli anchor shot dubbi. Nel caso in cui questo confronto dia esito positivo, il prototipo estratto dall'insieme degli anchor shot dubbi verrà associato all'insieme degli anchor shot certi ingrandendo questo insieme e riducendo l'insieme degli anchor shot dubbi.

Il processo verrà ripetuto per tutti i prototipi individuati nell'insieme degli anchor shot certi.

Il nostro sistema, dunque, permette sia di migliorare le prestazioni del sistema di base fornendo un insieme di anchor shot certi più ricco rispetto a quello del sistema di base e sia, grazie alla possibilità di individuare le porzioni salienti per effettuare il confronto di speaker, di ridurre in maniera considerevole (40%) i tempi di computazione (vedi D'Anna et alii, 2006).

5. CORPUS UTILIZZATO

Il dataset utilizzato durante tutta la sperimentazione è composto da filmati (per un totale di circa 16 ore) relativi a varie edizioni dei telegiornali delle emittenti nazionali italiane RAI 1 (26 edizioni del TG1) e CANALE 5 (16 edizioni del TG5) con una copertura in termini di ore di trasmissione approssimativamente uguale per ciascuna delle due emittenti. Le sessioni di registrazione sono avvenute nell'arco di tempo che va da giugno 2003 ad agosto 2004.

Riportiamo due tabella riassuntive per entrambe le reti, con l'elenco delle edizioni acquisite e le relative durate.

Orario	Data	Durata	Durata in frame	Numero di shot	Anchor shot	Anchor
11.30	25/06/2003	0.06.27	9675	66	7	Cristina Guerra
11.30	26/06/2003	0.05.05	7625	46	7	Cristina Guerra
11.30	27/06/2003	0.05.23	8075	62	8	Cristina Guerra
13.30	23/06/2003	0.25.45	38625	320	18	Tiziana Ferrario
0.00	10/07/2003	0.09.59	14975	83	12	Puccio Corona
0.55	14/07/2003	0.22.03	33075	200	15	Puccio Corona
11.30	14/07/2003	0.05.20	8000	61	7	Stefano Ziantoni
11.30	15/07/2003	0.05.31	8275	59	7	Stefano Ziantoni
20.00	10/07/2003	0.32.54	49350	379	22	Paolo di Giannantonio
20.00	11/07/2003	0.32.02	48050	345	21	Paolo di Giannantonio
20.00	14/07/2003	0.32.17	48425	334	19	Lilli Gruber
0.15	25/07/2003	0.24.42	37050	263	21	Paolo Giani
11.30	17/07/2003	0.05.30	8250	66	9	Stefano Ziantoni
11.30	23/07/2003	0.05.23	8075	81	9	Cristina Guerra
13.30	22/07/2003	0.24.37	36925	281	18	Tiziana Ferrario

13.30	23/07/2003	0.26.17	39425	291	18	Tiziana Ferrario
13.30	26/07/2003	0.27.27	41175	344	17	Tiziana Ferrario
13.30	29/07/2003	0.26.38	39950	277	17	Francesco Giorgino
17.00	25/07/2003	0.09.31	14275	83	13	Manuela de Luca
17.00	26/07/2003	0.09.57	14925	127	11	Manuela de Luca
22.50	11/07/2003	0.07.31	11275	66	10	Manuela Lucchini
22.50	14/07/2003	0.07.51	11775	86	10	Manuela Lucchini
22.50	25/07/2003	0.10.57	16425	105	13	Raffaele Genah
17.00	02/09/2003	0.10.27	15675	128	12	Manuela de Luca
17.00	04/09/2003	0.11.27	17175	122	11	Manuela de Luca
22.50	02/09/2003	0.05.34	8350	69	8	Raffaele Genah
		6.36.35	**594875**	**4004**	**340**	

Tabella 1: Dataset RAI1

Orario	**Data**	**Durata**	**Durata in frame**	**Numero di shot**	**Anchor shot**	**Anchor**
1.00	01/07/2003	0.24.45	37125	197	13	Giuseppe Brindisi
20.00	01/07/2003	0.34.27	51675	354	16	Lamberto Sposini
20.00	07/07/2003	0.37.02	55550	317	15	Didi Leoni & Fabrizio Summonte
1.00	28/12/2003	0.19.40	29500	163	9	Gianluigi Gualtieri
20.00	27/11/2003	0.35.36	53400	341	15	Lamberto Sposini
1.00	12/02/2004	0.24.54	37350	180	9	Paolo di Mizio
1.00	13/02/2004	0.24.46	37150	171	10	Paolo di Mizio
20.00	12/02/2004	0.30.20	45500	269	14	Cesara Buonamici
13.00	17/07/2004	0.34.34	51850	304	22	Didi Leoni & Fabrizio Summonte
13.00	19/07/2004	0.35.12	52800	242	21	Barbara Petri&Fabrizio Summonte
20.00	10/07/2004	0.31.08	46700	235	13	Enrico Mentana
20.00	03/07/2004	0.30.20	45500	279	11	Cesara Buonamici
20.00	11/07/2004	0.30.39	45975	282	14	Enrico Mentana

20.00	12/07/2004	0.32.55	49375	298	13	Lamberto Sposini
20.00	20/02/2004	0.37.31	56275	308	16	Lamberto Sposini
13.00	05/08/2004	0.35.04	52600	268	18	Didi Leoni & Fabrizio Summonte
		8.45.53	**788825**	**3979**	**229**	

Tabella 2: Dataset Canale 5

Il dataset di Canale 5 ha la particolarità di contenere cinque edizioni del TG interessate dalla presenza di due anchorperson in studio, tipicamente un uomo e una donna.

Il dataset di RAI1 ha invece la particolarità di essere composto da filmati di durata eterogenea. Infatti, distinguiamo in esso le cosiddette edizioni "ordinarie" e quelle "ridotte", queste ultime caratterizzate da situazioni in cui l'anchor-person commenta "a voce" le immagini dei servizi News.

Una prima distinzione è stata effettuata, quindi, in termini di durata:

Edizioni brevi (durata < 15 min): 16 edizioni RAI

Edizioni medio-lunghe (durata > 15 min): tutte le edizioni di C5 e il rimanente delle edizioni RAI

Per mostrare la variabilità degli anchor shot al variare sia dell'emittente televisiva sia della particolare edizione del telegiornale, si riportano nella tabella 3 degli esempi di shot in cui è presente l'anchor, classificandoli sulla base dell'inquadratura di quest'ultimo nella scena.

 Anchor in primo piano	 Anchor in primo piano	 Anchor in primo piano

 Anchor dietro alla scrivania con piccolo monitor a lato	 Anchor dietro alla scrivania con piccolo monitor a lato	 Anchor dietro alla scrivania con piccolo monitor a lato
 Anchor con ampio monitor	 Anchor con video e inviato	 Anchor con video e inviato
 Due anchor	 Anchor in piedi sullo sfondo	 Panoramica studio

Tabella 3: Diverse inquadrature dell'anchor nei vari shot per Tg della Rai e Canale 5

I filmati relativi al dataset sono stati registrati via satellite tramite un decoder satellitare con hard disk interno (a 6000 Kbit/sec in formato MPEG-2) e trasferiti su PC tramite scheda di acquisizione video (Pinnacle DC10+). La qualità finale dei filmati è stata ridotta a 2250 Kbit/sec, 25 frame/sec, ad una risoluzione di 720*576 pixel, in formato MPEG1, con audio stereo a 44.100 Hz (384 Kbit/sec).

La traccia audio del Tg viene estratta, dal filmato MPEG, mediante l'applicazione VirtualDub: essa è nel formato WAVE a 44.1kHz e 16 bit stereo. Poiché in effetti i due canali della traccia audio si equivalgono, per ridurre i tempi di calcolo nell'estrazione delle feature (che vedremo più avanti) è stato reputato necessario trasformare in mono la traccia stessa mediante l'uso del software Cool Edit Pro 2.0.

6. RISULTATI

Modulo SUV

Allo scopo di verificare l'ipotesi che le parti più salienti e caratteristiche di uno speaker sono quelle voiced, abbiamo fatto un confronto delle prestazioni tra il sistema di base e il sistema di base con il mascheramento utilizzando sia l'intero dataset di Rai1 che quello di Canale 5. Nelle tabelle 4 e 5 sono mostrati i risultati ottenuti.

RAI1	**Video senza masch.**	**Video con masch.**
Precision	95,0%	96,1%
Recall	88,5%	87,2%
F	91,6%	91,4%

Tabella 4: Confronto prestazioni del sistema di base senza e con mascheramento sul dataset di RAI 1

Canale5	**Video senza masch.**	**Video con masch.**
Precision	87,3%	86,8%
Recall	84,3%	85,8%
F	85,8%	86,3%

Tabella 5: Confronto prestazioni del sistema di base senza e con mascheramento sul dataset di Canale 5

Possiamo osservare che il sistema di base funziona meglio sui telegiornali di RAI1 caratterizzati dalla presenza di un solo anchor in studio rispetto a quelli di Canale 5 in cui sono presenti più anchor. Questo è un comportamento del tutto generale perché il primo dataset ha un carattere più omogeneo rispetto al secondo. Per quanto riguarda, poi, gli indici di prestazione possiamo notare che essi rimangono sostanzialmente identici. Tutto questo è chiaramente indicativo del fatto che il processo di mascheramento non fa perdere informazioni a riguardo del parlante e che quindi tale processo effettivamente individua le parti salienti per discriminare speaker differenti. Da notare che i tempi di computazione si sono ridotti del 40% pur fornendo risultati sostanzialmente identici da un punto di vista dei risultati.

Modulo MSD

Allo scopo di valutare le prestazioni del modulo MSD (multiple speaker detector) e la sua capacità di discriminare uno speaker da più speaker differenti abbiamo utilizzato il dataset dei 16 telegiornali di Canale5 essendo l'unico tra i due disponibili con la caratteristica di avere casi in cui c'è uno speaker o a volte due speaker come conduttori. Nella tabella 6 sono mostrati nel caso migliore, ossia fissata la distribuzione e l'intervallo di ricompattazione, gli errori medi ottenuti dal sistema distinguendo il sesso nei casi in cui c'è uno speaker.

% Errore medio	M (1sp)	F (1sp)	M+F (2sp)
AUTOCOR	15%	28%	38%
CEPSTRUM	5%	29%	53%
AUTOCOR OTTIM.	15%	21%	38%

Tabella 6: Errore medio percentuale nel caso migliore.

Come si può osservare la tecnica del cepstrum fornisce i migliori risultati nel caso di speaker maschili con il 5% di errore in media, mentre fornisce le prestazioni peggiori nei telegiornali con due speaker con un errore del 53%. Invece la tecnica della autocorrelazione in entrambi le versioni ottiene l'errore medio più basso, pari al 38%, con due speaker. Nel caso di speaker femminile i miglior risultati sono ottenuti dall'autocorrelazione ottimizzata con un errore medio pari al 21%.

Sistema complessivo

Concludiamo questo paragrafo con un'analisi dei risultati ottenuti dal sistema complessivo confrontando gli stessi con quelli ottenuti dal sistema di base inteso come quel sistema che utilizza il video e le feature audio di livello locale. Nella tabella 7 sono indicati i migliori risultati ottenuti dal sistema di base e dal nuovo sistema proposto sia nel caso del dataset di Rai 1 sia nel caso del dataset di Canale 5.

Sistema di base Rai 1 (340 Anchor shot)				Sistema sviluppato Rai1 (340 Anchor shot)				Diff. Relat.
correct	301	Precision	0,9495	correct	333	Precision	0,9708	2,1%
false	16	Recall	0,8853	false	10	Recall	0,9794	10,4%
miss	39	F	0,9163	miss	7	F	0,9751	5,9%
Sistema di base Canale 5 (229 Anchor shot)				Sistema sviluppato Canale 5 (229 Anchor shot)				Diff. Relat.
correct	193	Precision	0,9495	correct	216	Precision	0,9391	6,6%
false	28	Recall	0,8853	false	14	Recall	0,9432	10,0%
miss	36	F	0,9163	miss	13	F	0,9412	8,3%

Tabella 7: Risultati del sistema di base e del sistema sviluppato sui due dataset

I risultati tengono conto del numero di anchor shot correttamente identificati (correct), del numero di anchor shot erroneamente identificati (false) e del numero di anchor shot non individuati (miss) e sono espressi secondo i tre classici parametri di valutazione: precision, recall e figura di merito F.

Per quanto riguarda il numero di anchor shot correttamente individuati, le prestazioni del sistema proposto migliorano quelle del sistema di base diminuendo anche il numero di anchor shot omessi. Inoltre si osserva una riduzione del numero di falsi positivi introdotti dal sistema. Nell'ultima colonna della tabella 7 vi è confronto tra i due sistemi per i due dataset utilizzati.

Nel caso del dataset di RAI 1, osserviamo un moderato aumento della precision pari al 2,1% a fronte di un cospicuo aumento del 10,4 % della recall. Questo indica che il sistema proposto tende a diminuire in maniera maggiore il numero dei anchor shot missed rispetto

al numero di falsi positivi introdotti. Per quanto riguarda invece il dataset di canale 5, si nota un miglioramento sia della precision del 6,6% che della recall del 10%. Questo indica che il sistema proposto, nel caso di un dataset più difficoltoso, ottiene prestazioni migliori aumentando sia il numero degli anchor shot correttamente individuati sia diminuendo i falsi positivi che il numero di missed.

	DATASET	PRECISION	RECALL	F
(Chen et alii, 2002)	2h	89%	92%	90,4%
(Perez-Freire et alii, 2004)	13h	90,5%	90,5%	90,5%
(Lan et alii, 2004)	5h	92,6%	90,5%	91,5%
Sistema di base (De Santo et alii, 2005)	6h	95%	88,5%	91,6%
(Iyengar et alii, 2000)	1,5h	89,2%	95%	92%
Sistema proposto	6h	97,1%	97,9%	97,5%
(Kim et alii, 2005)	6h	97,6%	99,7%	98,6%

Tabella 8: Confronto con vari sistemi sviluppati in letteratura

Da un confronto del valore della figura di merito F dei migliori sistemi presenti in letteratura che impiegano un approccio integrato audio-video a priori (vedi tabella 8), osserviamo che il sistema preposto si comporta in maniera molto soddisfacente collocandosi al secondo posto assoluto su un dataset omogeneo, per quanto riguarda la durata, a quello utilizzato dal miglior sistema attualmente in letteratura (Kim et alii, 2005).

Riepilogando, da quanto detto, emerge chiaramente che sia l'utilizzo dell'analisi audio su scala globale e sia l'utilizzo delle tecniche di mascheramento e di individuazione di speaker multipli, permette di migliorare i risultati del sistema di base sia in termini di velocità di computazione che di indici di performance. Tali miglioramenti sono stati più cospicui nel caso del dataset di Canale 5 che è caratterizzato da un maggiore difficoltà intrinseca rispetto a quelli ottenuti con il dataset di RAI1.

7. CONCLUSIONI

Il sistema realizzato ha avuto l'obiettivo di risolvere alcuni dei problemi del sistema di base per la segmentazione in anchor shot. Innanzitutto è stato affrontato il problema dei tempi di computazione. A tale scopo è stato sviluppato e testato un modulo per la selezione delle porzioni salienti in audio, intese come quelle porzioni che effettivamente caratterizzano un anchor rispetto ad un altro. I risultati raggiunti dal sistema originale con l'utilizzo di questo modulo sono praticamente identici a quelli ottenuti dal sistema di base ma riducendo i tempi di computazione del 40% circa. Successivamente, allo scopo di migliorare i risultati del sistema originale, è stato individuato un nuovo set di features audio globali. Tali features sono caratterizzate da un livello di granularità più grande rispetto alle features locali e catturano caratteristiche dell'anchor più generiche ed indipendenti da ciò che viene pronunciato dallo stesso.

Infine è stato sviluppato un modulo per individuare la presenza di un anchor o di più anchor sulla sola base audio che usa il profilo della frequenza fondamentale sull'intero shot e che permette di eliminare alcune ambiguità introdotte dal sistema originale. Dai risultati ottenuti, quindi, il sistema proposto si dimostra più robusto ed affidabile in generale del sistema di base ottenendo dei risultati di tutto rispetto nel panorama della letteratura di settore.

BIBLIOGRAFIA

De Santo M., Percannella G., Sansone C., Vento M. (2005), "Combining experts for Anchorperson Shot Detection in News Videos", in *Pattern Analysis and Applications*, Springer-Verlag, Berlin. Vol. 7 - 4, pp. 447- 460.

D'Anna L., Percannella G., Sansone C., Vento M. (2006), "Un sistema automatico per la caratterizzazione degli speaker in flussi multimediali" in Atti del 2° convegno AISV, 30 novembre -2 dicembre 2005 Salerno, EDK Editore, Padova, pag.718-730.

Chen S.-C., Shyu M.-L., Liao W., Zhang C. (2002), "Scene change detection by audio and video clues", in Proceedings of ICME 2002, pp.365-368.

Perez-Freire L., Garcia-Mateo C. (2004), "A multimedia approach for audio segmentation in TV broadcast news", in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 1, pp. 369-372.

Iyengar G., Neti C. (2000), "Speaker change detection using joint audio-visual statistics", RIAO (Recherche d'Information Assistée par Ordinateur), Paris.

Lan D.-J., Ma Y.-F., Zhang H.-J. (2004), "Multi-level anchorperson detection using multimodal association", in Proceedings of 17th International Conference on Pattern Recognition, vol. 3, pp. 890-893.

Kim S.-K., Hwang D. S., Kim J-Y. and Seo Y-S. (2005), "An Effective News Anchorperson Shot Detection Method Based on Adaptive Audio/Visual Model Generation", in *Lecture Notes in Computer Science*, vol.3568, pp. 276-285, Springer Verlag, Berlin.