SVILUPPO DI UN SISTEMA DI *KEYWORD SPOTTING* PER L'INDICIZZAZIONE AUTOMATICA DEI DOCUMENTI AUDIO

Graziano Tisato*, Piero Cosi*, Isabella Gagliardi°

(*) Istituto di Scienze e Tecnologie della Cognizione del CNR
Via Martiri della Libertà, 2 - 35127 Padova, Italia
tisato@pd.istc.cnr.it, cosi@pd.istc.cnr.it

(°) Istituto per le Tecnologie della Costruzione del CNR
Via Bassini 15 - 20133 Milano, Italia
gagliardi@itc.cnr.it

1. SOMMARIO

L'applicazione del *Keyword Spotting* (*KWS*) presentata in questo lavoro rientra nel campo più generale dell'*Information Retrieval* (*IR*), e in quelli più specifici dello *Spoken Document Retrieval (SDR)*, dell'*Automatic Speech Recognition (ASR)* e del *Large Vocabulary Continous Speech Recognition (LVCSR)*.

Il *Keyword Spotting* è sostanzialmente un processo di *speech-to-text* del tutto simile al riconoscimento del parlato, in questo caso limitato solo all'individuazione di determinate parole chiave all'interno di un flusso audio continuo.

Il campo applicativo del *KWS* va dall'indicizzazione dei documenti audio, alla loro categorizzazione, ai sistemi di comando vocale, al rilevamento di particolari eventi, alla consultazione vocale delle basi dati, ad esempio per i motori di ricerca del *Web*, ecc. L'insieme delle applicazioni del *KWS* è anche sommariamente definito come *Speech Analytics*.

In generale, il *KWS* può rappresentare un valido aiuto nell'interazione uomo-macchina, permettendo l'uso del linguaggio naturale nella comunicazione.

Lo sviluppo e il miglioramento delle tecniche di riconoscimento del parlato e del *KWS*, assieme con la riduzione del tempo di elaborazione, sceso ormai al di sotto del tempo reale, ha esteso il campo applicativo dell'*IR*, in passato limitato ai documenti testuali, anche ai documenti audio.

Per dare un'idea del progresso ottenuto, mentre nel 1997, con un database di addestramento di 150 ore di parlato, si riusciva ad ottenere un errore sul riconoscimento delle parole cercate del 22%, già nel 2004 con *corpora* molto più grossi si scendeva al 9-10%.

La sfida attuale riguarda le tipologie di parlato che tradizionalmente ottengono i risultati peggiori, e cioè:

- Parlato rumoroso (telefonia, conferenze, ecc.).
- Parlato conversazionale.
- Variabilità di stili e accento dei parlatori.

Per quanto riguarda il *Keyword Spotting*, l'estrazione di determinate parole chiave può essere il primo passo di procedure di elaborazione tipiche dell'*IR*, che sono tradizionalmente basate sul testo, per ottenere le informazioni volute.

In certe applicazioni, in effetti, può essere di maggior utilità estrarre la presenza di parole significative dal punto di vista semantico, piuttosto che ricavare l'intera sequenza del

parlato, in modo da lanciare una azione appropriata. In questi casi, l'interesse è dato dalla velocità con cui si ottiene la risposta ad una interrogazione. Si tenga comunque presente che la velocità di elaborazione del *KWS* non è attualmente molto diversa da quella del riconoscimento del parlato continuo.

Una obiezione a questo approccio potrebbe essere quella che gli *ASR* non sono infallibili e sfornano un gran numero di parole errate, che possono avere una qualche parentela fonetica con l'originale. Si potrebbe pensare che lo scambio di parole nel processo di *Keyword Spotting* possa pregiudicare il funzionamento di un sistema di *IR*, che dipende forzatamente dalla correttezza delle stesse.

La scoperta, per certi versi sorprendente, fatta in questi ultimi anni è che l'influenza di questi errori sulle prestazioni complessive di un sistema di *IR* è molto limitata per la naturale ridondanza delle parole chiave relative ad un certo argomento. In effetti, è molto improbabile che tutte le occorrenze di una certa parola o dell'insieme delle parole chiave siano contemporaneamente scambiate con parole errate o semplicemente ignorate.

Ad esempio, con una percentuale di parole errate (*WER*) che passi dallo 0% al 40%, l'efficacia del sistema *IR* nell'individuare un documento secondo certi criteri diminuisce solo del 10% [Ng, 2000], [Allan, 2002]: si veda, ad esempio, gli esperimenti fatti dal 1997 (*TREC-6*) al 2000 (*TREC-9*) dalla *NIST Text REtrieval Conference* (*TREC*) oppure nel 1998 dalla *Topic Detection and Tracking* (*TDT*).

Questo spiega l'interesse che può rivestire l'utilizzo del *Keyword Spotting* nel campo dell' *Information Retrieval.*

La relazione si articola in questi argomenti:

- Introduzione al *Keyword Spotting* (Cap. 2)
- Possibili applicazioni (Cap. 3)
- Architetture implementate in questo lavoro e che si basano sull'azione contemporanea di due canali di riconoscimento (Cap. 4):
    - Il primo è un tipico *ASR* basato su un Modello Acustico (*AM*) e su un Modello Statistico del Linguaggio (*LM*).
    - Il secondo implementa una Grammatica a Stati Finiti (*GSF*), che non necessita della modellazione di un *LM* e permette la ricerca di una parola qualsiasi.
- Misure di valutazione della performance di un sistema di *WKS* (Cap. 5).
- Le caratteristiche dell'interfaccia grafico realizzato, per permettere la configurazione in una forma interattiva e rapida dei parametri dell'*ASR* utilizzato (Sonic – CSLR dell'Università del Colorado), la visualizzazione e la verifica immediata dei risultati della ricerca delle parole (Cap. 6).
- La valutazione dei risultati ottenuti, che nel caso del *Keyword Spotting* presenta un certo grado di complessità, dal momento che dipendono dai documenti scelti per il test e dalla velocità che si vuole imporre all'*ASR*. Sui test disponibili, i risultati della precisione delle parole riconosciute correttamente (60%) va giudicato con una certa indulgenza, considerando che si è utilizzato in questa prima fase del lavoro un Modello Acustico, *speaker independent*, ricavato da un corpus (*APASCI*) non conversazionale, e dunque non adatto ai documenti analizzati (Cap. 7-8).
- Le prospettive future che prevedono fra l'altro l'addestramento di un Modello Acustico su parlato conversazionale, l'utilizzo di parser semantici, l'uso di tecniche di adattamento (*Vocal Tract Length Normalization*, *Structured Maximum a Posterior*

Linear Regression, ecc.), dovrebbero contribuire a migliorare sensibilmente le prestazioni dell'*ASR* (Cap. 9).

2. INFORMATION RETRIEVAL E KEYWORD SPOTTING

Lo sviluppo scientifico e tecnologico ha reso evidente l'importanza della conoscenza e del trattamento della conoscenza nel mondo contemporaneo.

La quantità, la complessità e la varietà della conoscenza umana aumenta esponenzialmente assieme ai documenti, archivi, *database*, ecc. che la contengono a vari livelli gerarchici.

Ora, mentre ci si rende conto facilmente della difficoltà, ma, diciamo anche, dell'impossibilità e dell'inutilità di memorizzare ogni cosa nella propria testa, diventa invece ogni giorno più vitale sapere dove trovare e come accedere alle diverse forme della conoscenza disponibile.

L'approccio meno difficoltoso e costoso è quindi quello di trovare mezzi veloci ed efficaci per il recupero dell'informazione nel momento in cui si rende necessaria, piuttosto che cercare di stipare *una tantum* nel cervello una quantità di nozioni immensa.

In definitiva, l'aspetto rilevante della questione è il raggiungimento dello scopo (ottenere l'informazione voluta), e non il contenitore o la modalità (con cui l'informazione è ricavata).

L'azione combinata di queste esigenze e dei progressi delle scienze e delle tecnologie informatiche e di telecomunicazione ha portato alla attuale civiltà dell'informazione.

In particolare, per quanto riguarda l'*Information Retrieval*, si è assistito fin dal primo dopoguerra (vedi il *Memex* di *Bush* del 1945 [Buckland, 1992]) ad un proliferare degli studi teorici e delle implementazioni pratiche relative al recupero di informazioni da fonti testuali. I motori di ricerca sul *Web* sono la punta dell'*iceberg* di questa attività.

Oltre a questa tipologia statica di documenti, esiste però una miniera di informazioni, potenzialmente inesauribile, costituita dai documenti multimediali (archivi audio e video, trasmissioni radio e TV, lezioni accademiche, ecc.) che si sono accumulati negli anni passati e che aumentano vertiginosamente di giorno in giorno.

Una ulteriore categoria (storicamente la prima) candidata alla ricerca di informazioni è costituita dalle intercettazioni delle conversazioni telefoniche per scopi militari, di sicurezza nazionale e di spionaggio industriale (leggi *Echelon*) [Cernocky et al., 2007], ma anche per la supervisione delle conversazioni cliente-azienda, ecc.

Rispetto ai documenti testuali, i documenti multimediali pongono grossi problemi legati alla:

Difficoltà di indicizzarli, classificarli, riassumerli, ecc., data la quantità enorme che è prodotta ogni giorno.

Difficoltà per la ricerca e la consultazione, dato che per loro natura questi documenti hanno un accesso sequenziale.

Il *Keyword Spotting* (*KWS*) è una prima risposta all'esigenza di estrazione di informazioni da documenti multimediali e consiste nell'individuazione di un certo insieme di parole (*Keywords*) rilevanti dal punto di vista semantico [Higgins et al., 1985], [Rose et al., 1990], [Foote et al., 1995.].

Essendo l'informazione ricavata da documenti audio il *KWS* può anche essere chiamato con il nome di *Acoustic Keyword Spotting*.

Come si è detto nel sommario, il *KWS* è un settore di ricerca tipicamente interdisciplinare che fa riferimento a:

- *Information Retrieval (IR)* (che a sua volta si ispira alle scienze cognitive, alla linguistica, alla semiotica, alla scienza dell'informazione, *data mining*, ecc.)
- Spoken Document Retrieval (SDR)
- Automatic Speech Recognition (ASR)
- Large Vocabulary Continous Speech Recognition (LVCSR).

In effetti il *KWS* può essere realizzato in due processi distinti:

- *Speech-to-text*: Questa prima fase è simile al riconoscimento automatico del parlato per cui dalla voce in ingresso si ottiene una sequenza di parole, con la sostanziale differenza che nel *KWS* la ricerca può essere limitata solo a determinate parole chiave (o ad una loro combinazione). Generalmente dunque il vocabolario di un sistema di *KWS* consiste di un numero limitato di parole.
- *Information Retrieval*: Nella seconda fase si sfruttano tecnologie di *Information Retrieval* già sviluppate da alcuni decenni per ricavare le informazioni volute dalla sequenza delle parole in uscita dall'*ASR*.

Gli sviluppi del *KWS* sono relativamente recenti, dal momento che dipendono strettamente dalle prestazioni dei sistemi di riconoscimento automatico del parlato e che solo recentemente sono arrivati a *performance* e a tempi di risposta soddisfacenti.

Un contributo fondamentale al *KWS* basato su *Hidden Markov Model* (*HMM*) è dovuto al lavoro di Rose e Paul, fra i primi ad utilizzare l'algoritmo di Viterbi [Rose et al. 1990]. Questo approccio di calcolo della probabilità delle parole in uscita garantiva che le ipotesi generate per le varie parole non si sovrapponessero, ed limitava la generazione dei cosiddetti *falsi allarmi*, e cioè l'individuazione errata di parole chiave [James, 1995].

Assieme ai lavori teorici e pratici che si sono succeduti negli anni, si deve ricordare l'indispensabile attività dovuta al *National Institute of Standards and Technology* (*NIST*), che è un'agenzia del governo degli Stati Uniti d'America, alla *Defence Advanced Research Projects Agency* (*DARPA*) e all'*Advanced Research and Development Activity* (*ARDA*) del Dipartimento della Difesa degli USA. A partire dal 1992 il *NIST* sponsorizzò una serie di conferenze (*Text Retrieval Conference* - *TREC*) fondamentali per stabilire i criteri di valutazione oggettiva dei sistemi di *Information Retrieval*. In particolare, fra il 1997 e il 2000, la conferenza si dedicò alla valutazione dei documenti di parlato. Fra il 1998 e il 2004, il *NIST* sviluppò un programma di ricerca intitolato *Topic Detection and Tracking* (*TDT*) che faceva parte del *Translingual Information Detection, Extraction, and Summarization* (*TIDES*), sponsorizzato dal *DARPA*. Un altro importante progetto, intitolato *Spoken Term Detection* (*STD*), fu lanciato nel 2006 per la ricerca su archivi audio eterogenei in tre lingue (Arabo, Inglese, e Cinese), che non sia dipendente dal vocabolario (le keyword della ricerca non sono note a priori dal riconoscitore).

3. APPLICAZIONI DI KEYWORD SPOTTING

Come si è detto nei capitoli precedenti, il *KWS* ha un suo ambito preciso nell'*Information Retrieval*, per cui non interessa trascrivere automaticamente l'intera sequenza del documento parlato, ma recuperare solo le parole che lo possano qualificare in base a criteri voluti.

Il vantaggio non trascurabile di questo approccio è dato dalla velocità con cui si ottiene il risultato, visto che può essere di vari ordini di grandezza inferiore ad una completa decodifica con un normale *ASR*. In effetti, i tempi di esecuzione di un *ASR* crescono (anche

se non-linearmente) con l'aumentare del numero di parole cercate, per cui con un *KWS* di qualche centinaio di parole il tempo può essere di molto inferiore alla durata del brano in analisi (vedi cap. 7-8), mentre un riconoscitore che usi un vocabolario di decine di migliaia di parole può impiegare un tempo decine di volte il tempo reale, rendendo almeno per il momento improponibile questa strada di fronte alla enorme quantità dei documenti candidati ad una trascrizione.

I campi di applicazione del *KWS* sono i seguenti:

Indicizzazione dei documenti audio, per cui si individuano le parole rilevanti in un documento e si ricava la loro posizione temporale.

Categorizzazione (catalogazione, classificazione) dei documenti audio, per cui determinate parole chiave servono a collocare opportunamente il documento secondo criteri e algoritmi tipici del *NLP* (*Natural Language Processing*), e dell'analisi statistica e semantica.

Sistemi di comando vocale, per cui il sistema risponde a determinate parole o gruppi di parole con l'azione appropriata (ad esempio nei servizi di risposta automatica delle banche, ecc.).

Rilevamento di particolari eventi, simile al punto precedente, ma senza implicazione di interattività con l'utente.

Intercettazioni su conversazioni telefoniche e comunicazioni radio e video.

Consultazione delle basi dati (ad es. per i motori di ricerca del *Web*).

Supervisione di chiamate ai *Call Center*, per cui il verificarsi di determinate circostanza (parole o numero di ripetizioni di una parola) in una conversazione fra cliente e assistente del centralino di una ditta provocano l'intervento di un responsabile addetto a questo compito.

4. ARCHITETTURA DEL KEYWORD SPOTTING

Qualunque sia il disegno implementativo di un *KWS*, il compito è solo apparentemente più semplice di un *ASR* normale. In effetti le parole da cercare non sono isolate, ma immerse in una catena fonica continua che subisce profonde modificazione per il contesto, per i fenomeni di coarticolazione, e per le caratteristiche individuali del parlante (età, sesso, varianti dialettali, ecc.).

A questo si deve aggiungere il rumore introdotto dal mezzo di comunicazione e dall'ambiente.

I problemi e le soluzioni adottate sono del tutto simili a quelle del normale riconoscimento, e cioè:

Identificare inizio e fine delle parole.

Estrarre quelle che con la probabilità più alta corrispondono alle parole volute.

Formalmente si tratta di un processo di decodifica, per cui data la sequenza vettoriale O dei coefficienti *HMM* corrispondenti al parlato analizzato, si calcola la probabilità che una parola W (anch'essa rappresentata da una *HMM*) produca la sequenza osservata O, e cioè P(W|O).

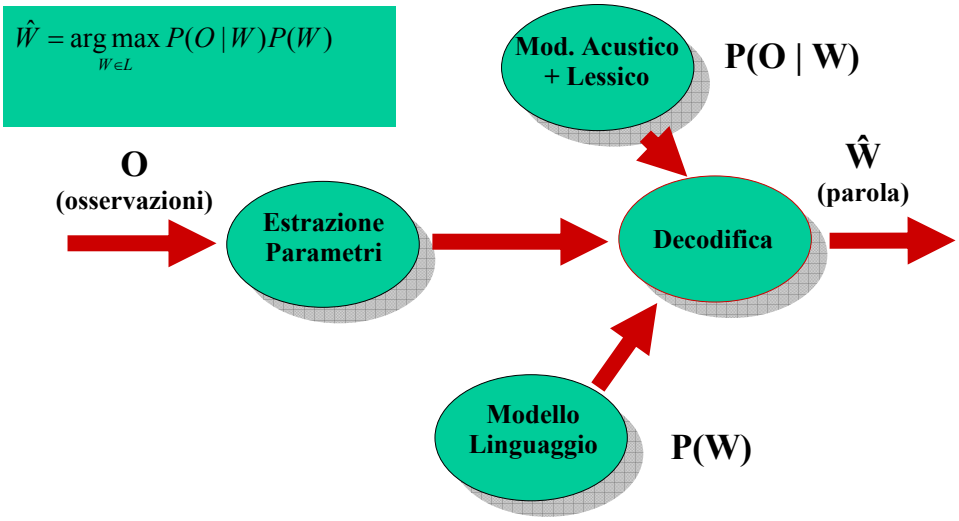

Fig. 1 – Schema di *ASR* con Modello Acustico e Modello del Linguaggio.

Una stima diretta di P(W|O) è in realtà poco praticabile, per cui si ricorre ad una riformulazione del problema seconda la regola di Bayes:

$\hat{W} = \text{argmax}_W P(W|O) = \text{argmax}_W P(O|W) P(W) / P(O) = \text{argmax}_W P(O|W) P(W)$

dove P(W) è la probabilità a priori di osservare la parola indipendentemente dal reale input vocale che si stia analizzando e corrisponde ad una modellazione opportuna di quella particolare tipologia di linguaggio che si vuole analizzare (Fig. 1).

Nella seconda parte dell'equazione, P(O|W) / P(O) si riduce al solo termine P(O|W), dal momento che P(O) non dipende dalla variabile W. P(O|W) calcola la probabilità di osservare un certo insieme di vettori cepstrali O all'occorrenza di una parola W. Questa quantità deriva dall'addestramento di un Modello Acustico su un certo insieme di campioni del parlato.

Il *Keyword Spotting* consiste dunque nel cercare la sequenza di stati di una certa parola chiave Ŵ, che con maggior probabilità produca il parlato osservato O. La sequenza più probabile si ottiene massimizzando l'equazione vista.

Nel *KWS* sono definite due classi distinte di modelli [Higgins et al., 1985], [Rose et al., 1990], [Manos, 1996], [Pellom, 2001], [Pellom et al., 2003]:

Il primo modello riguarda le *keywords* che sono rappresentate dalla loro sequenza fonetica. Questa modellizzazione è unica e caratteristica, come si vede nella Fig. 2, ad esempio, per la parola inglese /find/.

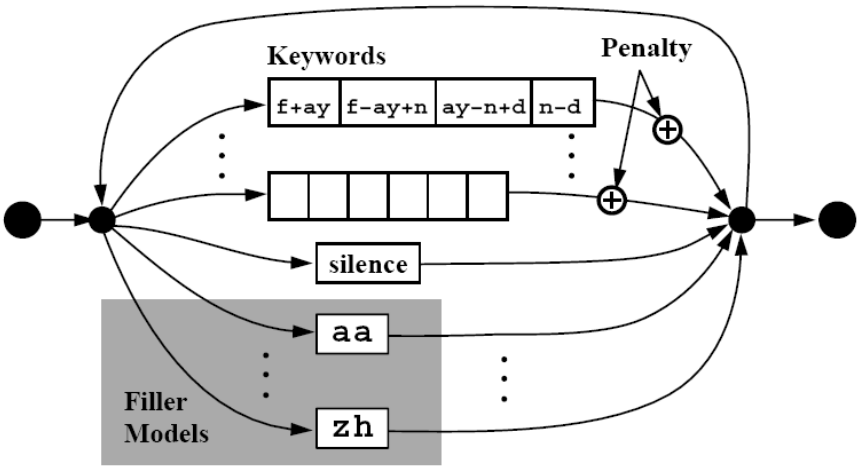

Fig. 2 – Modellazione delle parole chiave e dei *filler/garbage* nel *KWS* [Foote et al., 1995].

Il secondo modello riguarda i *filler/garbage*, di tutto ciò che deve essere eliminato come riempimento o spazzatura, e tutto ciò che non è una *keyword* compresa nel vocabolario delle parole chiave. Possono comparire in qualsiasi momento nel corso di una frase e sono assoggettati solo ad un parametro che ne controlla il comportamento (ad esempio, in Sonic, si tratta del *filler_penalty*).

È importante disporre di una qualche metodologia per la modellazione acustica di questi oggetti, in modo da ridurre i falsi allarmi. Nel caso più elementare questi modelli si riducono a una serie di singoli foni (Fig. 2). In altri casi si può volere la modellazione esplicita di tutte le parole da escludere. Anche fenomeni paralinguistici come la risata, il respiro, la tosse, ecc., possono essere modellati opportunamente nello stesso modo.

Sono stati tentati approcci diversi per la modellazione dei *filler/garbage*: ad esempio, nello *sliding window* si utilizzano tecniche di programmazione dinamica per estrarre le parole cercate [Wilpon et al., 1990], [Junkawitsch et al. 1996], [Silaghi et al. 2005]. Ogni percorso dinamico è considerato come l'occorrenza di una possibile parola chiave. È possibile rimuovere le sovrapposizioni eventuali di parole uguali e normalizzare in un secondo tempo i punteggi assegnati. Infine, opportune soglie permettono di ricavare le parole volute.

A differenza del modello *filler/garbage*, questa tecnica non richiede l'esplicita modellizzazione delle parti del parlato che non interessano la ricerca. Si calcola invece con Viterbi un allineamento che comincia in ogni finestra, dal momento che la parola chiave può essere presente in un punto qualsiasi. Si ricava il punteggio di tutti gli allineamenti che iniziano e finiscono in quel *frame*, e si assegna alla Misura di Confidenza per una certa *keyword*, il punteggio dell'allineamento migliore [Hacioglu et al., 2002].

A questo punto, si considera che una certa parola chiave sia effettivamente presente, se la misura della confidenza del modello di quella parola indica una alta probabilità di allineamento.

L'architettura di un sistema di *KWS* può essere classificata secondo tre principali categorie:

- *KWS* basato su *Large Vocabulary Continuous Speech Recognition* (*LVCSR*) (cap. 4.1).
- *KWS* basato su un riconoscimento di tipo fonetico (cap. 4.2).
- *KWS* basato su una Grammatica a Stati Finiti (cap. 4.3).

4.1 - KWS basato su LVCSR

Come risulta evidente dai capitoli precedenti, l'approccio più diretto al *KWS* che si possa pensare è quello di ricorrere ad un normale *ASR* del tipo *Large Vocabulary Continuous Speech Recognition*, tradizionalmente dotato di un Modello Acustico e di un Modello di Linguaggio, secondo lo schema di Fig. 1. [Rohlicek, 1989]. La soluzione banale è quella di analizzare l'uscita del riconoscitore per scoprire le eventuali parole chiave presenti.

Nella prima fase, che chiameremo di *speech-to-text*, un riconoscitore di parlato continuo ricava l'informazione testuale dai documenti (Fig. 3).

Nella seconda fase, si elabora l'uscita testuale del riconoscitore con i metodi tipici di *Information Retrieval* (Fig. 3).

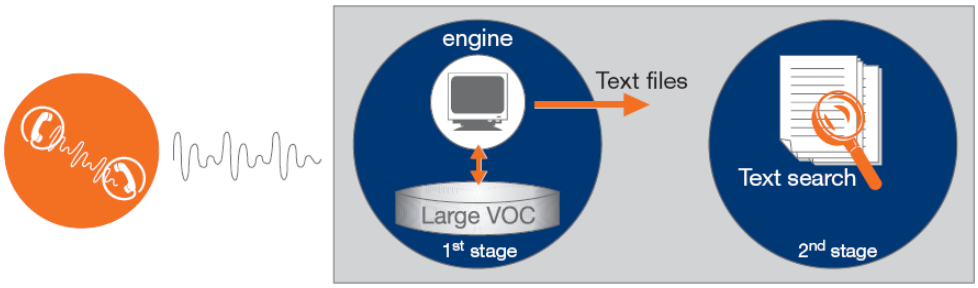

Fig. 3 – Sistema di *Keyword Spotting* basato su un *LVCSR* [Alon, 2005].

I vantaggi di questo approccio, nell'ipotesi che l'errore totale sulle parole (*WER*) sia basso, sono molti:

- Si può ottenere, volendo, la trascrizione completa dei documenti, e dunque una maniera economica per l'archiviazione.
- Le ricerche successive, che sfruttano il testo decodificato, sono ovviamente molto più veloci.
- Non c'è necessità di determinare a priori le parole chiave (tutte quelle presenti nel vocabolario sono candidate possibili).
- Si possono cercare nuove parole (sempre che siano presenti nel vocabolario) senza dover ricorrere ad un'altra fase di elaborazione.
- Con questa soluzione si ottengono i risultati migliori rispetto alle altre soluzioni, dal momento che il Modello del Linguaggio provvede ad eliminare molti errori.

Gli svantaggi sono dovuti alla poco flessibilità del sistema per le parole non comprese nel vocabolario:

- Il *LVCSR* comporta la necessità di creare un Modello del Linguaggio adatto al dominio di competenza (e questo può essere penalizzante).
- Il vocabolario previsto per il *LVCSR* non riesce a esaurire tutte le parole chiave che l'utente può richiedere (problema del *out of vocabulary words – OOV*). Quando si presenta questa necessità è necessario ridefinire il Modello del Linguaggio e rielaborare i documenti su cui si fa la ricerca.
- Come si era detto nel cap. 3, quando ci sia l'esigenza di risposte rapide e si voglia trattare documenti di grandi dimensioni (decine di migliaia di parole) e in grande quantità, l'utilizzo di un *ASR* completo è penalizzante per il carico computazionale (ma ormai non più così proibitivo con le macchine attuali).

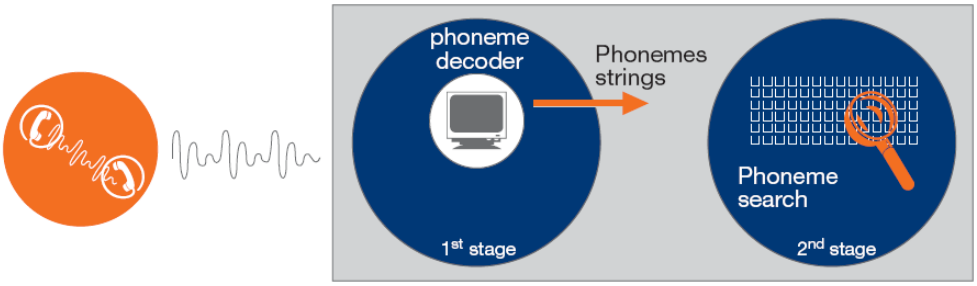

Fig. 4 – *Keyword Spotting* basato su riconoscimento fonetico [Alon, 2005]

4.2 - KWS con Riconoscimento Fonetico

Anche in questo caso, si tratta di un sistema con due processi distinti.

Nel primo lavora un riconoscitore fonetico che fornisce in uscita il flusso dei fonemi riconosciuti, senza tentare l'individuazione delle parole (Fig. 4). L'*ASR* non ha quindi necessità di un Modello del Linguaggio L'uscita dell'*ASR* è una stringa fonetica generalmente soggetta ad errori (con l'effetto di propagarsi nella successiva conversione a testo) (Fig. 4).

Il secondo passo tenta il recupero delle parole chiave la cui descrizione fonetica coincide in qualche punto con la sequenza ottenuta dall'*ASR* e prosegue con l'elaborazione di *Information Retrieval*.

Il vantaggio più rilevante di questa tipologia di *KWS* è che le parole chiave non devono essere predefinite: si può cercare un qualsiasi termine (nomi, parole straniere, ecc.) di cui si dia la trascrizione fonetica.

Gli svantaggi sono dovuti alla mancanza di un Modello del Linguaggio nella fase di decodifica che provveda ad eliminare in partenza molte delle possibili alternative errate della stringa fonetica.

Anche questo approccio, come quello basato su *LVCSR*, non è adatto nel caso si vogliano tempi di risposta rapidi e nel caso di grande quantità di documenti, poiché la ricerca sulla stringa fonetica è più complessa e pesante dal punto di vista computazionale che quella fatta su un testo di parole.

4.3 - KWS Basato su Grammatica GSF

In questo caso l'*ASR* restringe la sua ricerca ad una lista limitata di parole (o una loro combinazione). Il Modello del Linguaggio vero e proprio è sostituito da un modello a uni-

grammi (le parole chiave hanno tutte la stessa probabilità) o ad una Grammatica a Stati Finiti che stabilisce le possibili relazioni fra le parole ammesse (Fig. 5).

Vantaggi:

- Le parole sono fissate a priori e senza restrizioni: si possono inserire parole non comuni, parole straniere, ecc. Non si pone il problema *OOV*. Si può lanciare una ricerca contemporaneamente su parole in lingue diverse.
- Non c'è alcuna dipendenza dal contesto come avviene in un *ASR* normale e non c'è necessità di modellazione del linguaggio.

Si ottengono *performance* migliori rispetto agli altri approcci. La velocità è più elevata di un normale *ASR* e dipende dal numero di parole cercate (può essere inferiore di molte volte al tempo reale anche su un normale *PC*).

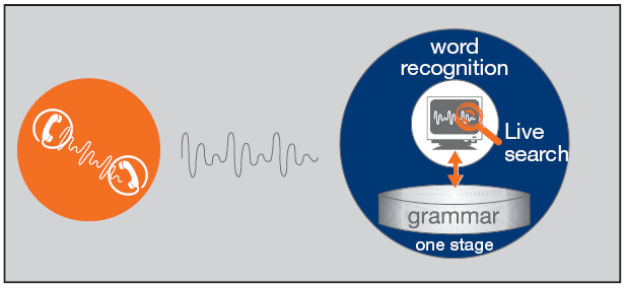

Fig. 5 – *Keyword Spotting* basato con una Grammatica a Stati Finiti [Alon, 2005]

Gli svantaggi dipendono anche in questo caso dall'assenza di *LM*, ma, come si è detto nel sommario, gli errori incidono in maniera poco rilevante sulle *performance* totali, essendo compensati spesso dalla ridondanza naturalmente presente nei documenti.

5. MISURE DI VALUTAZIONE DEL KWS

Questo è un aspetto critico per il *Keyword Spotting*, poiché le prestazioni di un sistema dipendono dalla scelta delle parole (ci sono parole più o meno facili da riconoscere), dal numero di parole chiave presenti e dalla tipologia dei campioni analizzati.

Come si è anticipato, il *NIST* ha compiuto sforzi notevoli per la formalizzazione di questo aspetto rilevante del *KWS*.

Le performance del *KWS* possono essere definite da varie misure:

Precisione o ***Word Correct Rate*** (*WCR*): percentuale delle parole individuate correttamente rispetto a tutte quelle effettivamente rilevate, oppure in altre parole la *proporzione rilevante delle parole trovate* (Fig. 6).

Word Error Rate (*WER*): percentuale delle parole errate (Sub+Del+Ins) rispetto a tutte quelle effettivamente rilevate, oppure in altre parole la *proporzione errata delle parole trovate* (Fig. 6).

Le parole errate comprendono:

Sub - % di parole sostituite con altre (*False alarm*)

Del - % di parole cancellate erroneamente (*False reject*)

Ins - % di parole inserite per errore (*False alarm*)

	Words	*Keys*	*Dur(s)*	*Elap.(s)*	*RT Ratio*	*WCR*	*Sub*	*Del*	*Ins*	*WER*
Results with LM	829	150	379	428.27	**1.13**	**95.4**	0.0	4.6	0.0	**4.6**
Results with GFS	829	150	379	152.83	**0.40**	**60.9**	14.9	24.1	8.0	**47.1**

Fig. 6 – *WCR* e *WER* per un test di *Keyword Spotting* con Modello del Linguaggio rispetto ad una Grammatica a Stati Finiti.

Fra le altre misure adottate nel *KWS*:

Accuratezza: (Corr – Ins) / (Corr + Sub + Del)

e cioè la percentuale fra la differenza parole_corrette e parole_inserite rispetto a tutte le parole chiave effettivamente presenti nel documento.

Figura di Merito: Un differente approccio per la misura della performance di un *KWS* è basato sulla Figura di Merito (FOM), e cioè la media di parole correttamente rilevate per un certo numero di falsi allarmi per ora.

Recall: Corr / (Corr + Sub + Del)

e cioè la percentuale fra parole corrette e tutte le parole chiave effettivamente presenti nel documento. In genere si usa plottare i valori della Precisione rispetto al *Recall* (Fig. 7).

Mean Average Precision (*mAP*) si calcola facendo la media aritmetica dei valori di Precisione Media rispetto ai punti individuati dal *recall* per tutti i test compiuti sui documenti (conversazioni, ecc.) (Fig. 8).

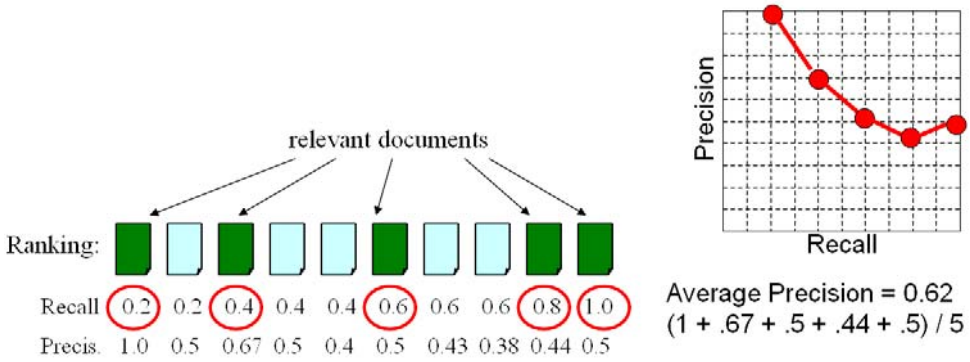

Fig. 7 – Grafico della Precisione rispetto al *Recall* e calcolo della Precisione Media [Lavrenko, 2005].

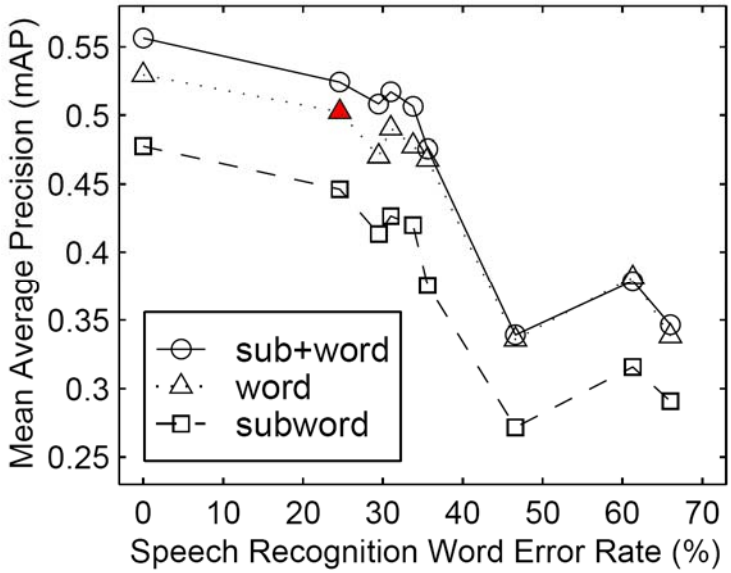

Fig. 8 – *Mean Average Precision* (*mAP*) rispetto a *Word Error Rate* (*WER*): si vede che anche con indici di errore *WER* abbastanza alti la *performance* media di precisione cala in maniera minima. Ad es., si passa da *mAP* =0.5295 per le trascrizioni di riferimento (per definizione corrette cioè *WER*=0%) ad un *mAP* =0.5025 per un *WER*=24.6% nel punto contrassegnato con un triangolo rosso [Ng, 2000].

Questo parametro è importante nella valutazione di vari sistemi di *KWS:* in effetti, se si plottano i dati di *mAP* rispetto al *Word Error Rate* (*WER*), si scopre in che modo varia la efficienza del sistema nell'individuare documenti, conversazioni, ecc. contenenti le parole desiderate. Nella Fig. 8 si può notare come si passa da *mAP* =0.5295 per le trascrizioni di riferimento (per definizione corrette *WER*=0%) ad una *Mean Average Precision*=0.5025 per un *WER*=24.6% nel caso di test su documenti reali (punto contrassegnato con un triangolo rosso), con un peggioramento di solo il 5% circa. Questo sta ad indicare come malgrado l'aumento notevole del *WER*, i documenti siano ancora individuati con successo [Ng, 2000].

La spiegazione della scarsa sensibilità del *KWS* agli errori del motore di riconoscimento va cercata, come si era visto nel sommario, nella naturale ridondanza di informazione presente in un documento [Allan, 2002].

6. CARATTERISTICHE DEL SISTEMA REALIZZATO

L'interfaccia grafica, realizzata per questo lavoro, ***Keyword Spotting Interface***, è stata progettata con lo scopo di facilitare le operazioni di impostazione dei parametri e la loro variazione, e soprattutto di consentire una rapida e efficace verifica dei risultati ottenuti (Fig. 9).

Il programma è scritto in **TCL/Tk** (http://www.activestate.com/Products/ActiveTcl) per compatibilità con vari sistemi operativi ed anche per la rapidità di sviluppo consentita.

Le caratteristiche principali del sistema sono:

- *Speaker Independent*: il *training* del Modello Acustico è stato fatto sul corpus APASCI elaborato dall'ITC-IRST e distribuito da Elda www.elda.org/catalogue/en/speech/S0039.html
- *Keyword Spotting su parlato continuo*: non c'è alcuna condizione restrittiva nell'estrazione delle parole (ad esempio che debbano essere isolate o che debbano essere limitate ad un certo numero) e sull'introduzione di nuove parole (*out of vocabulary words* - *OOV*).
- Trascrizione fonetica automatica delle parole da ricercare (da un database oppure con una conversione *letter-to-sound* per gli *OOV*).
- Il motore di riconoscimento utilizzato è Sonic [Pellom, 2001], [Pellom et al., 2003] cslr.colorado.edu/beginweb/speech_recognition/sonic.html
- *Compatibilità con Sonic*: Si è mantenuta la completa compatibilità con Sonic, per cui il funzionamento è garantito anche in assenza di interfaccia.
- *Compatibilità con vari Sistemi Operativi*: La nuova interfaccia può girare su tutti i sistemi operativi su cui già funzionava Sonic (in realtà finora è stato testato solo su Linux e Microsoft Windows XP).
- *Indipendenza da Sonic*: L'interfaccia grafica non dipende dal motore di riconoscimento scelto in questo caso (e cioè Sonic), ma può funzionare in modo trasparente con un altro motore *ASR* (ad es. Sphinx).
- *Interfaccia user-friendly*: L'interfaccia grafica realizzata facilita notevolmente il lavoro di test e modifica dei parametri dell'*ASR* (Fig. 10), e in modo particolare la visualizzazione e la verifica dei risultati con Sclite (Fig. 11). È anche possibile una verifica manuale, in assenza di trascrizione, dal momento che si può ascoltare sequenzialmente, o isolatamente, nel file sonoro i punti corrispondenti alle parole individuate, e constatarne la correttezza (Fig. 9). È previsto in futuro anche la contemporanea visualizzazione dei *frame* video relativi ad un eventuale filmato su cui si faccia il *Keyword Spotting*.

Il sistema si compone di due parti:

- *Sonic_batch*: Il riconoscitore vero e proprio, che sfrutta una versione opportunamente modificata del programma scritto in **c** da B. Pellom per poter estrarre le parole chiave volute secondo determinati criteri [Pellom, 2001].
- *Keyword Spotting Interface*: L'interfaccia di cui si è appena parlato (Fig. 9-11).

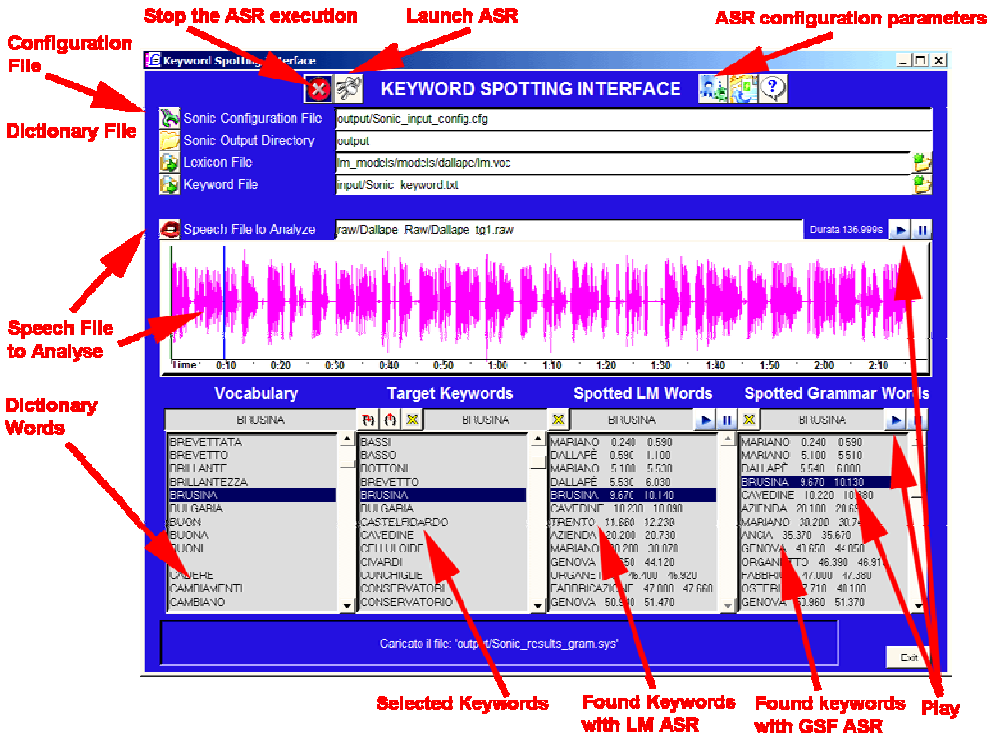

Fig. 9 – Keyword Spotting Interface: Finestra Principale

Parametri di Configurazione

I parametri di configurazione dell'*ASR* sono assegnati in fase di inizializzazione di Sonic, ma possono essere modificati dinamicamente nella finestra visibile in Fig. 10.

Il loro aggiustamento è molto utile per ottenere il miglior compromesso possibile fra risultati e tempo di esecuzione.

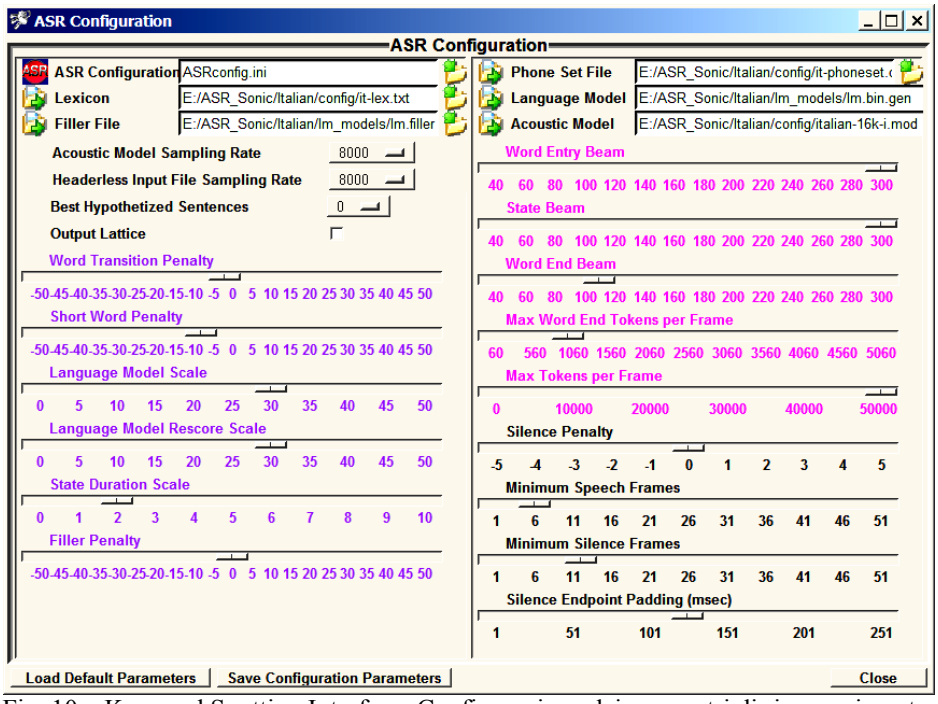

Fig. 10 – Keyword Spotting Interface: Configurazione dei parametri di riconoscimento.

I parametri più rilevanti rispetto alle prestazioni dell'*ASR* sono i seguenti:

`Word Transition Penalty:` Penalizzazione nella transizione da una parola ad un'altra. Con valori decrescenti si aumentano gli errori di cancellazione di parole, mentre si diminuiscono le parole inserite per errore.

`Short Word Penalty:` Fattore di penalizzazione applicato alle parole corte (tre fonemi o meno). Con valori che diminuiscono, si privilegiano l'individuazione di parole corte.

`Language Model Scale:` Peso assegnato al Modello del Linguaggio nella valutazione probabilistica delle parole. Diminuendo i valori di questo parametro, aumentano in modo rilevante i tempi di esecuzione e gli errori di rilevamento delle parole chiave.

`Max Active States:` Numero massimo di stati contemporaneamente attivi in ogni finestra di analisi. Aumentando il suo valore migliorano le prestazioni del riconoscitore, ma cresce (di molto) il tempo di esecuzione e l'occupazione di memoria.

`Word Entry Beam:` Soglia di potatura dei rami ammessi nella ricerca al primo stato di ogni parola. Valori bassi migliorano la velocità di esecuzione, ma peggiorano il grado di accuratezza del rilevamento delle parole.

`State Beam:` Soglia di potatura dei rami attivi nella ricerca per gli stati successivi al primo (ed escluso l'ultimo). Valori bassi migliorano

la velocità di esecuzione e deprimono l'accuratezza dei risultati ottenuti.

`Word End Beam:` Soglia di potatura dei rami attivi nella ricerca per l'ultimo stato di una parola. Valori bassi migliorano la velocità di esecuzione a scapito dei risultati ottenuti.

`Max Word End Tokens Per Frame:` Massimo numero di parole considerate nel passaggio da un *frame* all'altro. Valori limitati restringono la ricerca e migliorano i tempi di esecuzione, ma peggiorano i risultati.

`Keyword Threshold:` Soglia di confidenza che controlla l'accettazione delle parole chiave nel *Keyword Spotting*.

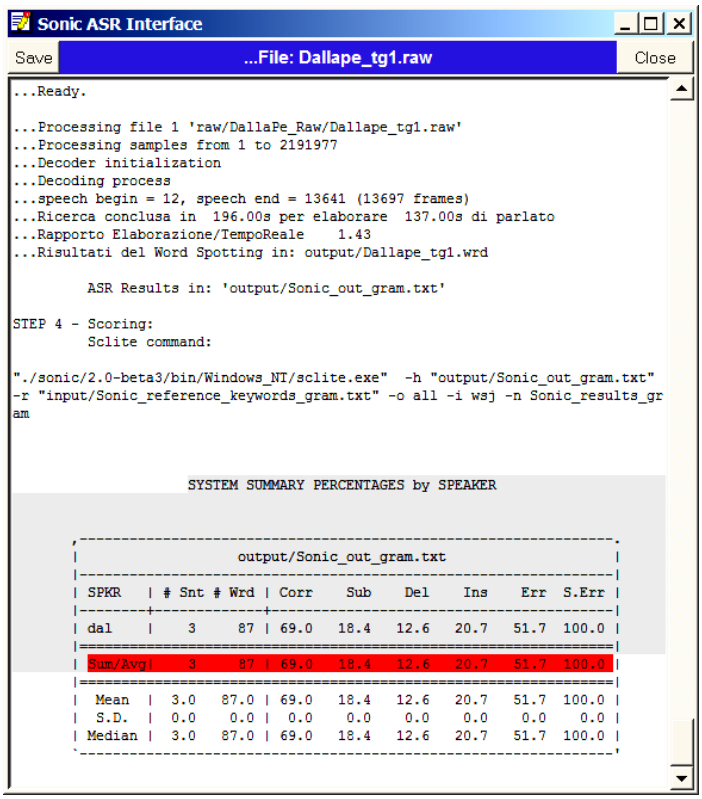

Fig. 11 – Keyword Spotting Interface: Risultati del decodificatore con Sclite.

7. TEST GSF SU CAMPIONI CLIPS TV E RADIO

Il sistema descritto in precedenza è stato testato su 10 brani estratti (a caso) dal database CLIPS (*Corpora e Lessici di Italiano Parlato e Scritto* www.clips.unina.it) e relativi a parlato televisivo e radiofonico (divulgazione, cultura, intrattenimento, ecc.) per complessivi 35 m. di parlato.

Il Modello Acustico utilizzato proviene invece dal corpus *APASCI* che è parlato microfonico su frasi lette, tipologicamente abbastanza diverso dal CLIPS.

Il criterio per la scelta delle parole chiave da cercare è stato quello di selezionare tutti i sostantivi presenti nei vari brani, che non fossero verbi, aggettivi, articoli, ecc. per un totale di 856 keywords, e cioè il 14% del totale di 5936 parole.

Il test è stato fatto solo con una Grammatica a Stati Finiti comprendente le parole selezionate per ogni brano. Non si è potuto ricavare un Modello di Linguaggio adeguato per la disparità e la limitata disponibilità di materiale trascritto.

I risultati sono i seguenti (Fig. 12):

- *Word Correct Rate (WCR)*= 60 %
- *Word Error Rate (WER)* = 47.3 %
- *Recall* = 74.2 %
- *Real-Time Ratio* = 0.3 %

Speaker	***Words***	***Keys***	***Dur(s)***	***Elap.(s)***	***RT Ratio***	***WCR***	***Sub***	***Del***	***Ins***	***WER***
LeccecorpusRDit14L	294	40	144.81	37.13	0.26	52.50	10.00	37.50	0.00	47.50
MilanocorpusTVit02M	453	43	145.49	38.70	0.27	72.10	14.00	14.00	7.00	34.90
NazcorpusRDdc02Z	1091	102	317.00	101.70	0.32	64.70	10.80	24.50	10.80	46.10
NazcorpusRDis01Z	717	33	306.27	72.05	0.24	51.50	15.20	33.30	6.10	54.50
ParmacorpusTVit02E	616	79	221.36	61.36	0.28	60.80	24.10	15.20	11.40	50.60
PerugiacorpusRDis020	388	108	135.27	45.84	0.34	87.00	2.80	10.20	8.30	21.30
PerugiacorpusTVis040	446	125	158.12	60.36	0.38	60.80	13.60	25.60	10.40	49.60
PerugiacorpusTVit040	454	76	149.78	43.75	0.29	36.80	11.80	51.30	3.90	67.10
TorinocorpusTVit02T	452	59	191.31	53.02	0.28	52.50	23.70	23.70	6.80	54.20
VeneziacorpusRDit01V	1025	191	329.18	123.08	0.37	53.40	17.30	29.30	4.70	51.30
	Words	***Keys***	***Dur(s)***	***Elap.(s)***	***RT Ratio***	***WCR***	***Sub***	***Del***	***Ins***	***WER***
Total	5936	856	2098.5	636.99	**0.30**	**60.00**	14.10	25.80	7.40	**47.30**

Fig. 12 – Risultati del sistema di *KWS* con Sclite su 10 brani CLIPS.

Come si vede, i valori di *WCR* e *WER*, pur non entusiasmanti, sono molto promettenti nella prospettiva di poter modellare opportunamente il linguaggio. L'indice di *Recall*, invece, e cioè la percentuale fra parole corrette e tutte le parole chiave effettivamente presenti nel documento è abbastanza elevato (74.2 %).

8. TEST SU BRANO CONVERSAZIONALE CON LM + GSF

Il secondo esperimento è stato condotto su un parlato conversazionale abbastanza rumoroso (intervista a Amleto Dallapè).

Il Modello Acustico, come nel caso precedente, è stato allenato su *APASCI* (quindi non adatto a questo brano).

Il Modello del Linguaggio è stato addestrato su circa 50 m del brano disponibile, lasciando al test del riconoscimento 7 m circa.

Come parole chiave da ricercare sono state scelte 150 sostantivi presenti nel brano (trascurando verbi, aggettivi, articoli, ecc.) e ritenuti più rilevanti per il contenuto dell'intervista. Queste parole corrispondono al 12.8% circa del totale di 1176 parole.

I risultati con il Modello del Linguaggio sono stati i seguenti (Fig. 13):

- *Word Correct Rate (WCR)*= 32.0 %
- *Word Error Rate (WER)* = 68.1 %
- *Recall* = 32.0 %
- *Real-Time Ratio* = 0.95 %

I risultati con la Grammatica a Stati Finiti sono stati i seguenti (Fig. 13):

- *Word Correct Rate (WCR)* = 24.0 %
- *Word Error Rate (WER)* = 76.0 %
- *Recall* = 31.6 %
- *Real-Time Ratio* = 0.45 %

	Words	***Keys***	***Dur(s)***	***Elap.(s)***	***RT Ratio***	***WCR***	***Sub***	***Del***	***Ins***	***WER***
Results with LM	1176	150	437	415.00	**0.95**	**32.0**	16.0	52.0	0.0	**68.1**
Results with GFS	1176	150	437	197.35	**0.45**	**24.0**	60.0	16.0	0.0	**76.0**

Fig. 13 – Risultati del sistema di *KWS* con Sclite su una intervista di A. Dallapè.

Inutile dire che i risultati così modesti sono dovuti ad un Modello Acustico inadeguato alla tipologia del brano e che il test va ripetuto non appena addestrato l'AM su un corpus conversazionale.

Se questo esperimento è ripetuto con una registrazione microfonica, adatta all'*AM* allenato su *APASCI*, in cui si ripete più o meno con le stesse parole una parte (6 m. circa) dell'intervista, allora i risultati migliorano in maniera più soddisfacente (Fig. 6):

I risultati con il Modello del Linguaggio sono stati i seguenti (Fig. 6):

- *Word Correct Rate (WCR)* = 95.4 %
- *Word Error Rate (WER)* = 4.6 %
- *Recall* = 95.4 %
- *Real-Time Ratio* = 1.13 %

I risultati con la Grammatica a Stati Finiti sono stati i seguenti (Fig. 6):

- *Word Correct Rate (WCR)* = 60.9 %
- *Word Error Rate (WER)* = 47.1 %

- *Recall* = 72.6 %
- *Real-Time Ratio* = 0.4 %

9. CONCLUSIONI E SVILUPPI FUTURI

È stato implementato un sistema di *KWS* che sfrutta un doppio canale di riconoscimento:

- Il primo è un Large Vocabulary Continous Speech Recognition basato su un AM e su un LM.
- Il secondo implementa una *GSF*, che non necessita della modellazione di un *LM* e permette la ricerca di una parola qualsiasi.

È stata creata una interfaccia che facilita l'impostazione dei parametri e la visualizzazione e la verifica dei risultati dei due canali del *KWS*.

I primi risultati ottenuti sono incoraggianti e in prospettiva molto promettenti con l'*AM* adeguato. I miglioramenti che si pensa di apportare sono i seguenti:

- Addestramento con modelli acustici conversazionali.
- Integrazione nel KWS di parser semantici del tipo di Phoenix (per i quali Sonic è già predisposto e che dovrebbe migliorare migliorare notevolmente le performance).
- Convergenza automatica (fatta ora fuori linea) dei parametri alla prestazione ottimale (ad es. per ridurre i falsi allarmi o aumentare la correttezza delle parole).
- Visualizzazione sincrona dei frame video relativi in un eventuale filmato su cui si faccia il *Keyword Spotting*.

RIFERIMENTI

Allan J. (2002), "Perspectives on Information Retrieval and Speech"**,** *Information Retrieval Techniques for Speech Applications*, Coden, Brown and Srinivasan, editors. Springer-Verlag Lecture Notes in Computer Science, Vol. 2273, pp. 1-10.

Alon G. (2005), *Key-Word Spotting-The Base Technology for Speech Analytics*, White Paper, NSC - Natural Speech Communication Ltd.

Buckland, M. (1992), "Emanuel Goldberg, Electronic Document Retrieval, And Vannevar Bush's Memex". *Journal of the American Society for Information Science* 43, n. 4, pp. 284–294.

Cernocky J. et al. (2007), "Search in speech for public security and defense", *Proc. IEEE Workshop on Signal Processing Applications for Public Security and Forensics*, SAFE '07, pp. 1-7

Foote J., Jones G., Jones K., Young S. (1995), "Talker- Independent Keyword Spotting for Information Retrieval", *Proc. Eurospeech* 1995, Vol. 3, pp. 2145-2149.

Garofolo J., Voorhees E., Stanford V., Sparck Jones K. (1998), "TREC-6 1997 spoken document retrieval track overview and results". *Proc. of TREC-6* (1997), pp. 83-92, NIST special publication 500-240.

Garofolo J., Voorhees E., Auzanne C., Stanford V., Lund B. (1999), "1998 TREC-7 spoken document retrieval track overview and results". *Proc. TREC-7* (1998), pp. 79-89, NIST special publication 500-242.

Garofolo J., Auzanne C., Voorhees E. (2000), "The TREC spoken document retrieval track: A success story". *Proc. of TREC-8* (1999). NIST special publication 500-246.

Garofolo J., Lard J., Voorhees E. (2001), *2000 TREC-9 spoken document retrieval track*, Powerpoint presentation trec.nist.gov

Gelin P. (1997), Détection de mots clés dans un flux de parole : application à l'indexation de documents multimédia, Thèse EPFL, n. 1658.

Higgins A., Wohlford R. (1985), "Keyword Recognition Using Template Concatenation", *Proc. ICASSP 1985*, pp. 1233- 1236.

James D. (1995), The application of classical information retrieval techniques to spoken documents, Ph.D. Thesis, University of Cambridge.

Junkawitsch J., Neubauer L., Hoge H., Ruske G. (1996), "A new keyword spotting algorithm with pre-calculated optimal thresholds", Proc. *ICSLP* Vol. 4, pp. 2067.2070.

Lavrenko V., (2005), "Information Retrieval", Powerpoint presentation www.clsp.jhu.edu/ws2005/calendar/documents/LavrenkoJuly5.PPT

Manos A. S. (1996), A Study on Out-of-Vocabulary Word Modeling for a Segment-Based Keyword Spotting System, M.S. Thesis, Massachusetts Institute of Technology.

Myers C.S., Rabiner L.R., Rosenberg A. (1980)., "An Investigation of the Use of Dynamic Time Warping for Word Spotting and Connected Word Recognition", *Proc. ICASSP 1980*, pp. 173-177.

Ng K. (2000), "Information Fusion For Spoken Document Retrieval", *Proc. International Conference on Acoustics, Speech,. and Signal Processing*, pp. 2405–2408

Pellom B. (2001), SONIC: The University of Colorado Continuous Speech Recognizer, University of Colorado, Tech Report #TR-CSLR-2001-01

Pellom B., Hacioglu K. (2003), "Recent Improvements in the CU SONIC ASR System for Noisy Speech: The SPINE Task", in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2003*, Vol. 1, pp. I 4-7.

Rohlicek J., Russel W., Roukos S., Gish H. (1989), "Continuous Hidden Markov Modeling for Speaker-Independent Word-Spotting", in *Proc. ICASSP 1989*, pp. 627-630.

Rose R., Paul D. (1990), "A Hidden Markov Model Based Keyword Recognition System", in *Proc. ICASSP 1990*, pp. 129-132.

Silaghi M., Vargiya R. (2005): "A new evaluation criteria for keyword spotting techniques and a new algorithm", in Proc. *Interspeech 2005*, pp. 1593-1596.

Szöke I., Schwarz P., Matejka P., Burget L., Fapso M., Karafiát M., Cernocký J. (2005), "Comparison of Keyword Spotting Approaches for Informal Continuous Speech", in *Proc. Eurospeech* 05.

Wilpon J., Rabiner L., Lee C., Goldman E, (1990), "Automatic Recognition of Keywords in Unconstrained Speech Using Hidden Markov Models", *IEEE Trans. ASSP*, Vo138. No. 11, pp. 1870-1878.