SLAM 1.0 PER WINDOWS

Piero Cosi
CSRF- Centro di Studio per le Ricerche di Fonetica C.N.R. Padova
Via G. Anghinoni, 10 - 35131 Padova (Italy)
Tel. 049 8274421, Fax 049 8274416, E-Mail cosi@csrf.pd.cnr.it

SOMMARIO

Viene descritta la prima versione ufficiale del sistema di segmentazione semiautomatica del segnale vocale denominato SLAM (dall'acronimo inglese *Segmentation and Labelling Automatic Module*), sviluppato in ambiente Windows per Personal Computer presso il CSRF. SLAM fornisce in modo automatico alcune ipotesi di segmentazione allo scopo di rendere più veloce il compito di esperti fonetisti nell'analizzare grossi corpora di segnale verbale. Sulla base della conoscenza ortografica del testo pronunciato gli esperti devono scegliere l'allineamento più opportuno fra quelli proposti automaticamente. Il sistema riceve in ingresso i parametri forniti da un modello del sistema uditivo periferico dimostratosi molto efficace nel codificare le informazioni contenute nel segnale vocale e si basa sulla teoria della segmentazione multi-livello per la costruzione delle ipotesi di segmentazione.

Oltre alla segmentazione del parlato, il sistema fornisce all'utente numerose possibilità di visualizzazione di vari parametri di analisi, tra cui vari tipi di 'spettrogramma', l'andamento della frequenza fondamentale, dell'energia e dello 'zero-crossing' ed altri ancora. All'utente sono inoltre consentite alcune elementari operazioni di *editing* del segnale quali 'taglia' e 'incolla' 'cancella' e molte altre. Il sistema è stato sviluppato in linguaggio di programmazione Microsoft C++©[1] per il sistema operativo Microsoft Windows 3.1© ma e' supportato anche in ambiente Windows 95© e WindowsNT 4.0© ed in particolari architetture Macintosh©. SLAM è disponibile a chi ne facesse esplicita richiesta ed è distribuito in rete dal CSRF.

INTRODUZIONE

Numerosi sono ormai i corpora di segnale vocale prodotti da varie organizzazioni e gruppi di ricerca nel campo delle scienze linguistiche e fonetiche in campo nazionale ed internazionale [1], [2], come è possibile osservare, infatti,

[1] Tutte le sigle seguite dal symbolo © nel testo si riferiscono ai copyrigth delle rispettive compagnie: Pentium, 486 (Intel Corp.), Windows-3.11, Windows-95. Windows-NT4.0 (Microsoft Corp.), C++7.0, Visual-C1.5 (Microsoft Corp.), SoundBlaster-16, AWE32, AWE64 (Creative Labs. Inc.), Oros AU22 DSP (Oros Inc.), Macintosh, PowerMacintosh Quadra 8600/200 (Apple Computer), Virtual-PC (Connectix).

in un interessante rassegna sull'ingegneria linguistica recentemente apparse in Internet [3]. Per trasformare questi corpora in materiali scientificamente utilizzabili in vari campi di ricerca applicata e non, finalizzati essenzialmente alle tematiche riguardanti l'analisi, la sintesi ed il riconoscimento automatico della voce, è ovviamente necessario trasformarli in database organizzati e facilmente consultabili. Una delle prime elaborazioni di cui tutti questi corpora necessitano e che vengono spesso tralasciate in fase di progettazione, è senza dubbio la segmentazione e l'etichettatura (dall'inglese *labelling*), a vari livelli (semantico, lessicale, ortografico, fonetico), del segnale verbale oggetto dei corpora stessi. Non infrequente è, infatti, la presenza nei vari laboratori di ricerca di corpora di segnale verbale la cui utilità scientifica, nonostante la loro complessità, è ridotta quasi a zero a causa della mancanza di questa necessaria elaborazione.

Purtroppo un primo grosso inconveniente introdotto dalla necessità di associare ai corpora un'opportuna etichettatura, e questo soprattutto a livello fonetico, risiede nel fatto che, normalmente, questa viene affidata all'opera manuale di esperti linguisti o fonetisti e, di conseguenza, costituisce un significativo collo di bottiglia a causa dell'enorme spreco di risorse, sia temporali che economiche, che una tale operazione necessariamente richiede.

Nonostante queste considerazioni ovvie e condivisibili a livello scientifico, raramente da parte degli enti preposti alla progettazione ed alla realizzazione dei vari corpora viene riposta quell'attenzione che dovrebbe invece permeare tutte le fasi di sviluppo dei corpora stessi. In altre parole alla fase di definizione dei corpora ed alla loro successiva acquisizione vengono affidate tutte le risorse, scientifiche ed economiche, dimenticandosi dell'organizzazione finale e quindi delle finalità essenziali dei corpora stessi che risiedono ovviamente nella loro effettiva utilizzazione per il progredire della ricerca scientifica nel campo dell'ingegneria linguistica.

Un altro problema di non facile soluzione è rappresentato dal fatto che l'etichettatura manuale è sempre caratterizzata da un elevata anche se controllabile soggettività [4], [5]. Infatti, nonostante l'ausilio di sempre più affidabili strumenti audio visivi, le divergenze nei valori di segmentazione dello stesso materiale vocale, prodotti manualmente da parte di più esperti, non potranno mai essere completamente eliminate. A causa delle diverse capacità percettive, sia visive che uditive, come anche dell'oggettiva difficoltà di definire una inequivocabile strategia comune, è evidente l'implicita incoerenza di un tale approccio manuale. Sulla base di queste considerazioni, l'interesse per la realizzazione di sistemi automatici di segmentazione e "labelling" è ovviamente elevatissimo. Tali sistemi automatici, oltre a minimizzare i tempi di esecuzione, renderebbero implicitamente coerenti i risultati della segmentazione. Infatti, gli eventuali errori di segmentazione risulterebbero facilmente identificabili e categorizzabili a causa della natura algoritmica delle procedure.

Il sistema descritto in questo lavoro, studiato e progettato per fornire una risposta pratica a tutti le difficoltà sopra elencate, fornisce in modo automatico alcune ipotesi di segmentazione allo scopo di ridurre al minimo il compito di esperti fonetisti nell'analizzare grossi corpora di segnale verbale. Nessun istante di segmentazione viene posizionato manualmente, salvo rari casi, e agli esperti

viene esclusivamente richiesta un'azione di supervisione sulle ipotesi di segmentazione prodotte automaticamente dal sistema. Gli esperti, infatti, devono scegliere, sulla base della conoscenza ortografica del testo pronunciato, l'allineamento più opportuno fra quelli proposti automaticamente, eventualmente eliminando "marker" sovrabbondanti.

DESCRIZIONE DEL SISTEMA

Il sistema di segmentazione di seguito descritto è stato denominato SLAM, dall'acronimo inglese Segmentation and Labelling Automatic Module. Slam ottiene la segmentazione del segnale verbale in ingresso essenzialmente in tre fasi. Nelle prime due fasi SLAM opera automaticamente sul segnale verbale mentre nella terza richiede in modo interattivo la collaborazione dell'utente. La prima fase corrisponde all'elaborazione digitale del segnale verbale, mentre la seconda si riferisce all'individuazione sul segnale di vari possibili confini di separazione fra le varie unità. A queste segue una terza ed ultima fase, dove, sulla base delle informazioni fornite dalle precedenti elaborazioni, viene esplicitamente richiesto l'intervento di un operatore esperto, generalmente un linguista o un fonetista, a cui viene richiesto di scegliere, fra le varie ipotesi di segmentazione proposte dal sistema, quella giudicata più corretta ed affidabile.

Elaborazione digitale del segnale verbale.

Sul segnale possono essere effettuate varie elaborazioni ed i risultati di queste elaborazioni possono essere poi opportunamente visualizzati. L'utente può scegliere fra vari tipi di analisi spettrale fra i quali: spettrogramma (FFT) a banda larga, stretta, o comunque a banda selezionabile a piacere, spettrogramma basato sull'analisi LPC, anche questa impostabile a piacere, e 'neurogramma' basato sull'analisi effettuata da un particolare modello del sistema uditivo periferico umano [6], realizzato presso il CSRF [7], dimostratosi assai efficace nel codificare le informazioni necessarie e sufficienti per una valida segmentazione fonetica anche in presenza di condizioni di registrazione non ottimali, cioè con segnali particolarmente rumorosi [8]. Questa è, infatti, la rappresentazione consigliata per sfruttare al massimo le potenzialità del programma.

Qualora si ritenessero utili in fase di analisi, altri parametri di interesse quali la frequenza fondamentale, l'energia e lo zero-crossing possono essere visualizzati su richiesta dell'utente. Sebbene lo scopo di SLAM sia essenzialmente la segmentazione e l'etichettatura di corpora di segnale verbale, SLAM può essere utilizzato anche per tutte le normale operazioni di 'editing' del segnale. SLAM è dotato, infatti, di numerose funzioni mediante le quali è possibile operare direttamente sul segnale visualizzato quali ad esempio: 'taglia', 'incolla', 'cancella', 'normalizza' ed altre ancora.

Creazione delle ipotesi di segmentazione.

Questa fase rappresenta è il vero e proprio 'nocciolo' di SLAM. L'algoritmo di segmentazione, motore di SLAM, è basato interamente sulla teoria della segmentazione multi-livello [9-11]. La filosofia alla base di questa teoria sottolinea che non esiste un unico livello di rappresentazione segmentale in grado di descrivere tutti gli eventi acustici di interesse presenti nel segnale vocale. Per ovviare a questa implicita difficoltà viene adottata una rappresentazione multi-livello la quale consente di evidenziare all'interno di un unica struttura sia i mutamenti rapidi che quelli graduali riscontrabili sul segnale. Il segnale vocale viene considerato come una sequenza temporale di segmenti acustici quasi stazionari. Le caratteristiche del segnale all'interno di tali segmenti sono considerate fra loro più simili di quelle fra segmenti adiacenti. La segmentazione, seguendo questa interpretazione, può essere considerata come un semplice problema di *local clustering* in cui le decisioni da prendere riguardano esclusivamente la somiglianza dei vari frame con i segmenti acustici immediatamente precedenti o successivi. A differenza di altre tecniche di segmentazione basate sulla ricerca dei massimi o dei minimi di particolari parametri di analisi, utilizzando solo misure relative di differenza acustica la tecnica sopra descritta risulta essere assai più robusta per quanto riguarda l'indipendenza dal parlatore, dal vocabolario ed anche dal rapporto segnale disturbo. La costruzione della struttura multi-livello [9], descritta algoritmicamente in Tabella 1 ed illustrata graficamente in Figura 1 e in Figura 2, può quindi riassumersi in due fasi. Nella prima fase, per ogni target frame viene calcolata una media dei vari componenti del vettore di analisi corrispondenti rispettivamente ad una finestra lunga Δ frame sia alla destra che alla sinistra del frame in esame. Vi è da sottolineare che, anche se il vettore di analisi può essere prodotto da qualsivoglia tecnica di elaborazione digitale, l'utilizzazione della tecnica di analisi basata su un modello del sistema uditivo periferico umano, a cui si faceva precedentemente riferimento, garantisce a SLAM i migliori risultati [8]. Mediante misure di distanza euclidea, viene calcolata la somiglianza del vettore di analisi corrispondente al frame in esame con le due medie corrispondenti alle due finestre sopra indicate e viene quindi presa una decisione per associare il frame alla finestra precedente o a quella successiva. Altre strategie, oltre a quella della sola distanza euclidea, possono essere adottate per la definizione della somiglianza, consentendo quindi alla procedura di poter adattare la sensibilità delle associazioni alla particolare condizione locale. Allorché tutti i frame sono stati analizzati si vengono così a creare vari segmenti acustici adiacenti. Queste regioni iniziali costituiscono la base per la seconda fase della costruzione di quella particolare struttura di segmentazione gerarchica (nota in letteratura come 'dendrogramma') suggerita dal fatto che il segnale vocale è spesso caratterizzato da eventi acustici molto rapidi, le cui caratteristiche si diversificano in modo molto netto da quelle corrispondenti al loro intorno. Questa particolare segmentazione gerarchica, incorporando alcuni vincoli temporali, risulta di particolare utilità nel valutare ed ordinare la significatività dei singoli eventi acustici. Il sistema di

clustering utilizzato per la costruzione della struttura multi-livello si basa essenzialmente sulla stessa tecnica utilizzata per ottenere i vari segmenti acustici adiacenti di base. In fatti, partendo dalle regioni precedentemente calcolate, ogni regione è associata con quella alla sua destra o alla sua sinistra sempre utilizzando una misura di distanza euclidea. La misura di distanza si applica in questa fase ad una media dei componenti dei vettori di analisi corrispondenti ai frame costituenti le varie regioni. Due regioni sono quindi unite fra loro quando possono essere associate sulla base del criterio appena esposto. La procedura viene ripetuta fino a che l'intera frase non può essere descritta da un singolo evento acustico. Mantenendo l'informazione relativa della distanza con cui due regioni si fondono fra loro si può quindi costruire una struttura di segmentazione multi-livello ('dendrogramma') come quella illustrata nelle Figure 3 e 4, relative all'analisi della frase "Susan ca[n't...]" (le ultime due consonanti non sono visualizzate) pronunciata da un parlante femminile inglese.

Scelta della segmentazione finale e successiva etichettatura.

E' questa la fase interattiva della procedura di segmentazione, in quanto l'intervento dell'utente viene esplicitamente richiesto da SLAM. In linea di principio la segmentazione finale potrebbe essere estratta anche in modo automatico, mediante tecniche di *pattern recognition*, ricercando il cammino di segmentazione ottimo all'interno della struttura multi-livello e avendo come informazione la trascrizione fonetica tipo della frase pronunciata in ingresso al sistema. SLAM però, attualmente, perviene alla segmentazione finale mediante un limitato intervento manuale dell'utente, consistente nel determinare la posizione, lungo l'asse verticale del dendrogramma, in cui scegliere i marker definitivi ed eventualmente eliminandone alcuni se sovrabbondanti.

In una fase successiva si etichetteranno i marker ottenuti mediante opportune etichette o *label* che rispecchieranno il particolare livello di segmentazione: (semantico, lessicale, ortografico, prosodico, fonetico).

IMPLEMENTAZIONE SOFTWARE

Originariamente, SLAM è stato sviluppato in ambiente Microsoft Windows 3.1© mediate il linguaggio di programmazione C++7.0©. La versione attuale a cui fa riferimento questo lavoro si riferisce invece all'ambiente di sviluppo Visual C++1.5©. SLAM è stato provato su *personal computer* basati su processori Intel© 486 e Intel Pentium©, in ambiente operativo Windows3.1© WindowsWorkgroup3.11©, Windows95© e Windows-NT4.0©, equipaggiati con schede SuperVGA dotate di almeno 256 colori[2], mouse a tre tasti[3] e almeno 4/8 Mbytes di RAM. SLAM è stato inoltre utilizzato in ambiente Macintosh su un

[2] Questo solo per sfruttare al massimo le potenzialità di visualizzazione dello spettrogramma.

[3] Il mouse a tre tasti è necessario soltanto per utilizzare le funzioni di etichettatura.

PowerMacintosh Quadra 8600/200© equipaggiato con Virtual-PC Windows95 Emulation Software©

Qualora siano attivate le funzionalità di registrazione e riproduzione audio, ovviamente indispensabili per le operazioni di segmentazione (non però per altre funzioni di *editing* o visualizzazione), SLAM utilizza le schede sonore Creative SoundBlaster16© o AWE32/64©, anche se potrebbero essere utilizzate altre interfacce A/D-D/A, come ad esempio, la vecchia scheda OROS-AU22© DSP, su cui inizialmente si basava la versione originaria. SLAM è in grado di eseguire moltissime operazioni sul segnale verbale come esemplificato nei menù illustrati in Figura 5. Oltre alla forma d'onda temporale del segnale verbale si possono visualizzare gli andamenti temporali di altri parametri di interesse come l'energia, la frequenza fondamentale, calcolata mediante due noti algoritmi denominati AMDF [12] e SIFT [13], e lo *zero-crossing*. Dal punto di vista spettrale SLAM fornisce la visualizzazione di uno spettrogramma basato sulle tecniche classiche di analisi quali FFT e LPC, con parametri impostabili a piacere, ma soprattutto su un modello del sistema uditivo periferico che, rispetto alle tecniche tradizionali, ha fornito risultati migliori in fase di segmentazione. Originariamente, ad ogni file di segnale verbale venivano associati e visualizzati vari file di analisi, creati *off-line da opportuni* programmi "satellite". Nella versione attuale, tulle le analisi possono essere attivate *on-line* all'interno di SLAM ed i relativi risultati possono essere immediatamente visualizzati.

Sul segnale verbale l'utente ha la possibilità di eseguire numerose operazioni. Il segnale, infatti, può essere registrato, ascoltato ed opportunamente visualizzato, mediante varie opzioni di ingrandimento e *scrolling.* Il segnale, inoltre. può essere modificato mediante semplici operazioni di editing quali ad esempio: *CUT*, *PASTE*, *CLEAR*, *COPY*, *FADE-in*/*out*, *NORMALIZE*, ed altre elementari operazioni matematiche. Muovendo, inoltre, il cursore, mediante un *mouse*, all'interno delle varie finestre di visualizzazione, i valori dei parametri di analisi associati alle varie rappresentazioni attive quali ad esempio l'ampiezza del segnale, la posizione temporale, l'energia, lo zero-crossing, la frequenza fondamentale o i valori corrispondenti alla corrente analisi in frequenza, sono immediatamente visualizzati all'utente. Pur riassumendo alcune delle principali funzioni di un semplice sistema di analisi e visualizzazione del segnale verbale, la funzione principale di SLAM, quella cioè per cui è stato sviluppato, rimane quella della segmentazione automatica. Per questa, viene calcolata e successivamente visualizzata la rappresentazione spettrale uditiva e sulla base di questa rappresentazione viene applicato l'algoritmo MLS, precedentemente descritto, con cui viene elaborato ed immediatamente visualizzato, in modo automatico, il dendrogramma delle ipotesi di segmentazione. L'utente ha quindi la possibilità di scegliere, sulla base delle ipotesi di segmentazione, fornite automaticamente da SLAM, quelle ritenute più affidabili semplicemente posizionando il cursore all'interno del dendrogramma, come illustrato graficamente in Figura 1. Operando in questo modo, gli istanti di segmentazione non sono mai posizionati dall'utente, ma esclusivamente dal sistema, che ne consente tuttavia, qualora richiesto esplicitamente in particolari casi difficili, anche una verifica manuale. Nella segmentazione con SLAM, la strategia da adottare è, preferibilmente, quella di

scegliere una segmentazione fine, in altre parole una sovrasegmentazione, e di modificarla successivamente, sulla base della trascrizione ortografica del segnale in analisi, eliminando eventualmente i possibili *marker* sovrabbondanti. Per quanto riguarda la precisione ottenibile mediante SLAM nella segmentazione di un semplice corpora si può far riferimento a [4].

Come già sottolineato, l'utilizzazione di un particolare algoritmo di analisi basato su un modello del sistema uditivo periferico umano ha sensibilmente ridotto, rispetto ad altre tecniche di analisi più tradizionali, basate su FFT o LPC, l'intervento manuale dell'operatore, specialmente in particolari condizioni di segnali rumorosi, come appare evidente dall'esame delle Figure 3 e 4.

Qualora si disponga di un mouse a tre tasti, in SLAM è inclusa inoltre la possibilità di etichettare con simboli SAMPA [14], automaticamente richiamabili a video, i marker ottenuti. Altra caratteristica di SLAM, esemplificata in Figura 6, è la possibilità di operare simultaneamente su più segnali. Questa possibilità è stata ottenuta utilizzando tecniche di programmazione facenti uso delle specifiche MDI (*Multiple Document Interface*) disponibili in ambiente Windows. Il numero dei segnali sui quali è possibile operare simultaneamente è limitato soltanto dalla memoria RAM disponibile.

RINGRAZIAMENTI

Questo lavoro è stato reso possibile esclusivamente grazie alla gentile collaborazione di S. Seneff e J.R. Glass del MIT (*Massachesets Institute of Technology*) di Boston. In particolare i loro suggerimenti per l'implementazione del modello uditivo denominato *Joint Synchrony/Mean-Rate (S/M-R) model of Auditory Speech Processing (ASP)* [6], e per lo sviluppo della strategia di segmentazione multilivello *MLS* [9] sono stati di fondamentale importanza.

CONCLUSIONI

L'utilizzazione di un modello del sistema uditivo periferico e di un sistema di analisi multi-livello hanno consentito la realizzazione di SLAM, un sistema semiautomatico di segmentazione del segnale vocale. L'utilizzazione di un tale sistema da parte di esperti fonetisti consente, in primo luogo, di ridurre enormemente i tempi di segmentazione di grosse basi dati vocali. Infatti gli istanti di segmentazione sono posizionati automaticamente dal sistema riducendo sensibilmente l'intervento umano. In secondo luogo, a causa della natura algoritmica del sistema, la coerenza dei risultati della segmentazione risulta ovviamente enfatizzata. Infatti gli eventuali errori di segmentazione risultano essere facilmente identificabili e categorizzabili a differenza di quelli umani. Eliminando infine l'intervento umano sulla decisione finale del livello di segmentazione voluto e sostituendolo con un algoritmo di ricerca dell'allineamento ottimo del segnale verbale con la sua corrispondente trascrizione, all'interno della struttura multi-livello, il sistema può diventare completamente automatico, non

solo per quanto riguarda la segmentazione o allineamento temporale, ma anche per quanto riguarda il labelling o l'etichettatura del segnale vocale.

BIBLIOGRAFIA

[1] LDC: Linguistic Data Consortium, Internet www page address: http://www.ldc. upenn.edu/.

[2] ELRA: European Language Resources Association, Internet www page address: http://www.icp.grenet.fr/ELRA/home.html.

[3] Ingegneria Linguistica, Internet www page address: http://comel.ing.uniroma1.it/ ~sandro/lengeng.htm.

[4] P. Cosi, D. Falavigna and M. Omologo (1991), “A Preliminary Statistical Evaluation of Manual and Automatic Segmentation Discrepancies”, Proceedings of EUROSPEECH-91, 2ndEuropean Conference on Speech Technology, Genova, 24-26 September, 1991, pp. 693-696.

[5] T. Lander, B. Oshika, J. Carlson, T. Durham, and T. Bailey (1996), "Analysis of Inter-Labeler (dis)agreement in Phonetic Transcriptions of Multiple Languages." Proceedings of the Acoustical Society of America, Waikiki, Hawaii, December 1996.

[6] S. Seneff (1988), "A Joint Synchrony/Mean-Rate Model of Auditory Speech Processing", Journal of Phonetics, January 1988, pp. 55-76.

[7] P. Cosi (1997), "SLAM v1.0 for Windows: a Simple PC-Based Tool for Segmentation and Labeling", Proceedings of ICSPAT-97, International Conference on Signal Processing Applications & Technology, San Diego, CA, USA, September 14-17, 1997, pp. 1714-1718.

[8] P. Cosi, "Ear Modelling for Speech Analysis and Recognition", ESCA Workshop-92, Sheffield, 7-9 Apr, 1992.

[9] J.R. Glass, "Finding Acoustic Regularities in Speech: Application to Phonetic Recognition", Ph. D. thesis, Massachusetts Institute of Technology, May 1988.

[10] J.R. Glass and V.W. Zue (1988), "Multi-Level Acoustic Segmentation of Continuous Speech", Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-88), New York, N.Y., April 11-14, 1988, pp. 429-432.

[11] V.W. Zue, J. Glass, M. Philips and S. Seneff, "Acoustic Segmentation and Phonetic Classification in the SUMMIT System", Proc. IEEE-ICASSP 1989, paper S8.1, pp. 389-392.

[12] M.J. Ross, H.L. Shaffer, A. Cohen, R. Freudberg and H.J. Manley (1974), “Average magnitude difference function pitch extractor”, IEEE Trans. Acoust. Speech Signal Process., Vol. ASSP-22, pp. 565-572

[13] J.D. Markel (1972), “The SIFT algorithm for fundamental frequency estimation”, IEEE Trans. Audio Elettroacoust., Vol. AU-20, pp. 367-377.

[14] A.J. Fourcin, G. Harland, W. Barry e W. Hazan eds. (1989), "Speech Input and Output Assessment, Multilingual Methods and Standards ", Ellis Horwood Books in Information Technology, 1989.

Fase 1

1) Trova i marker iniziali $\{b_i, \quad 0 \le i \le N\}, \quad t_i < t_j, \quad \forall i < j$

Fase 2

2) *Crea l' insieme iniziale di regioni*

$R_0 = \{r_0(i), \quad 0 \le i < N\}, \quad r_0(i) \equiv r(i, i+1)$

3) Crea l' insieme iniziale di distanze

$D_0 = \{d_0(i), \quad 0 \le i < N\}, \quad d_0(i) \equiv d(r_0(i), r_0(i+1))$

4) *Fino a che* $R_N = \{r_N(0)\} \equiv r(0, N)$

Per ogni k *tale che* $d_j(k-1) > d_j(k) < d_j(k+1)$

(a) $r_{j+1}(i) = r_j(i), \quad 0 \le i < k$

(b) $r_{j+1}(k) = merge(r_j(k), r_j(k+1))$

(c) $r_{j+1}(i) = r_j(i+1), \quad k < i < N - j - 1$

(d) $R_{j+1} = \{r_{j+1}(i), \quad 0 \le i < N - j - 1\}$

(e) $d_{j+1}(i) = d_j(i), \quad 0 \le i < k - 1$

(f) $d_{j+1}(k-1) = \max(d_j(k-1), d(r_j(k-1), r_{j+1}(k)))$

(g) $d_{j+1}(k) = \max(d_j(k+1), d(r_{j+1}(k), r_j(k+1)))$

(h) $d_{j+1}(i) = d_j(i+1), \quad k < i < N - j - 1$

(i) $D_{j+1} = \{d_{j+1}(i), \quad 0 \le i < N - j - 1\}$

Definizioni

- b_i *confine fra segmenti al tempo* t_i.
- *r(i, j) regione fra* t_i *e* t_j.
- $r_j(i)$ *regione i all' iterazione j.*
- *d(i, j) distanza fra le regioni i e j.*
- $d_j(i)$ *distanza i all' iterazione j.*
- *merge(r(i, j), r(j, k)) raggruppa due regioni adiacenti in un unica regione r(i, k) fra* t_i *e* t_k.
- *Le distanze* $d_j(-1)$ *e* $d_j(N - j)$ *sono infinite.*

Tabella 1. Descrizione algoritmica della procedura di segmentazione multi-livello.

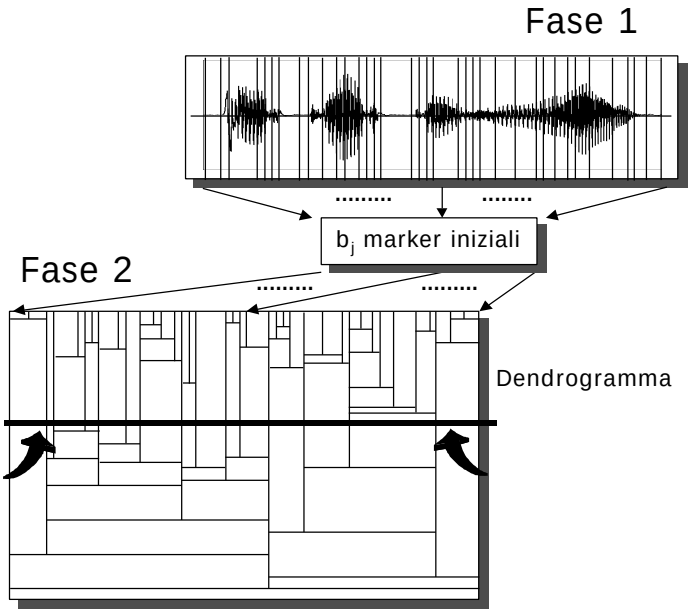

Figura 1. Rappresentazione schematica delle varie fasi dell'algoritmo di segmentazione multilivello.

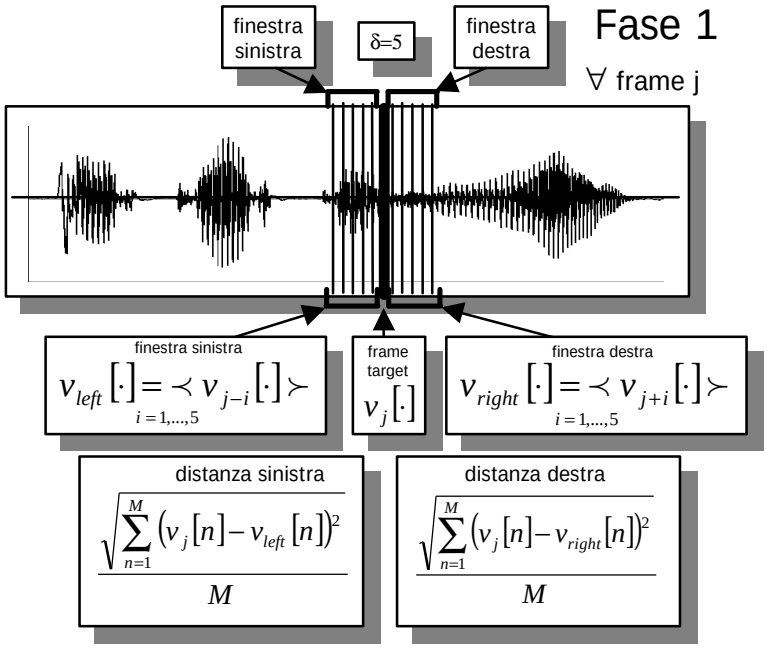

Figura 2. Illustrazione in dettaglio della prima fase dell'algoritmo di segmentazione multilivello.

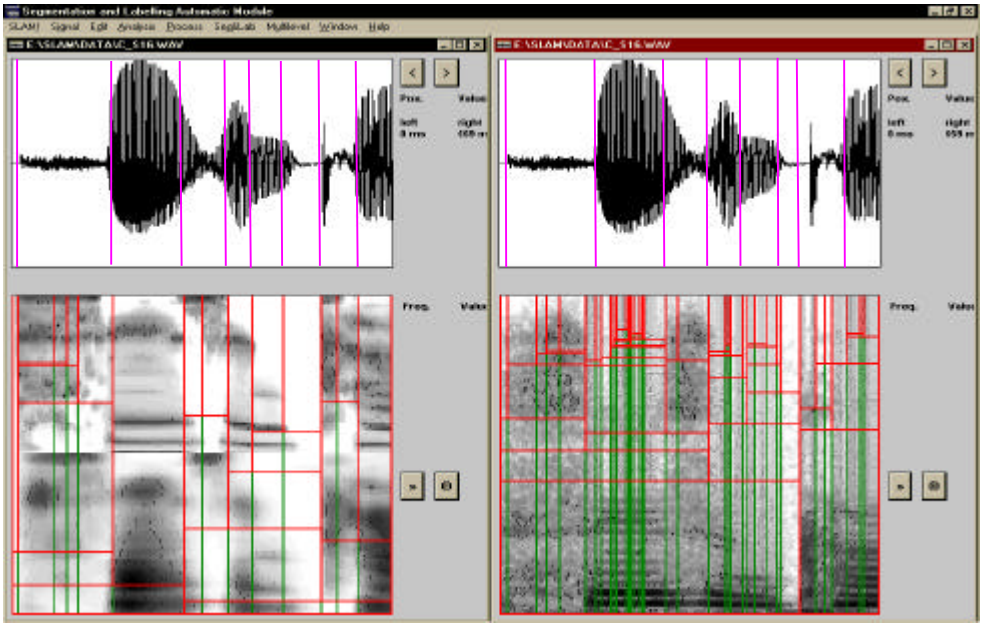

Figura 3. Applicazione di SLAM per la segmentazione della frase "Susan ca[n't...]" (le ultime due consonanti non sono visualizzate) pronunciata da un parlante femminile inglese. L'algoritmo MLS è applicato utilizzando un modello uditivo (sx) e una normale analisi FFT a banda stretta dx).

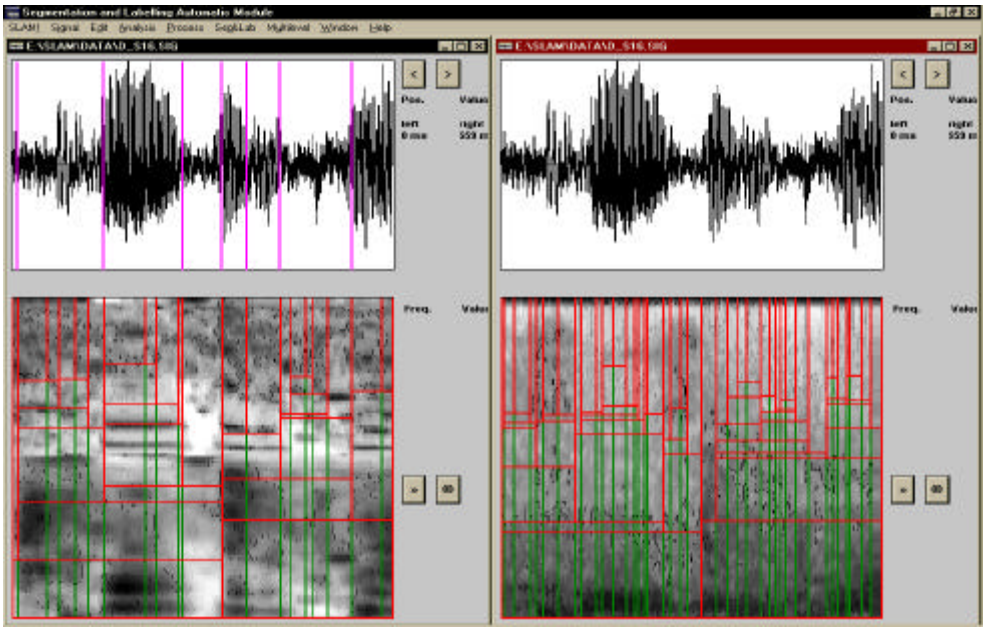

Figura 4. Applicazione di SLAM all'analisi della frase "Susan ca[n't...]" di Figura 3 in condizioni di segnale rumoroso. A sinistra è utilizzata la rappresentazione uditiva mentre a destra è utilizzata quella calcolata mediante FFT a banda stretta.

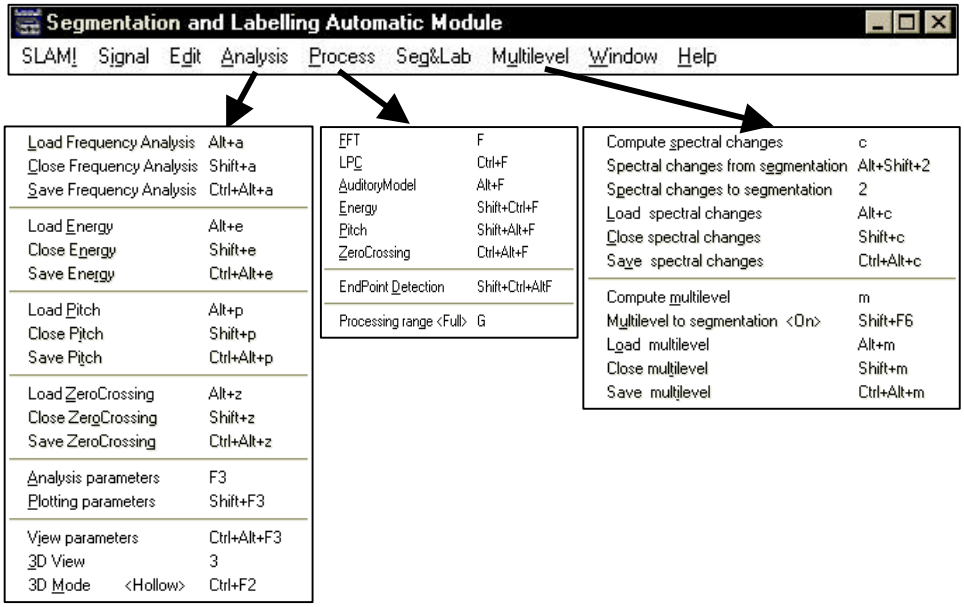

Figura 5. Menù principale e tre menù secondari di SLAM.

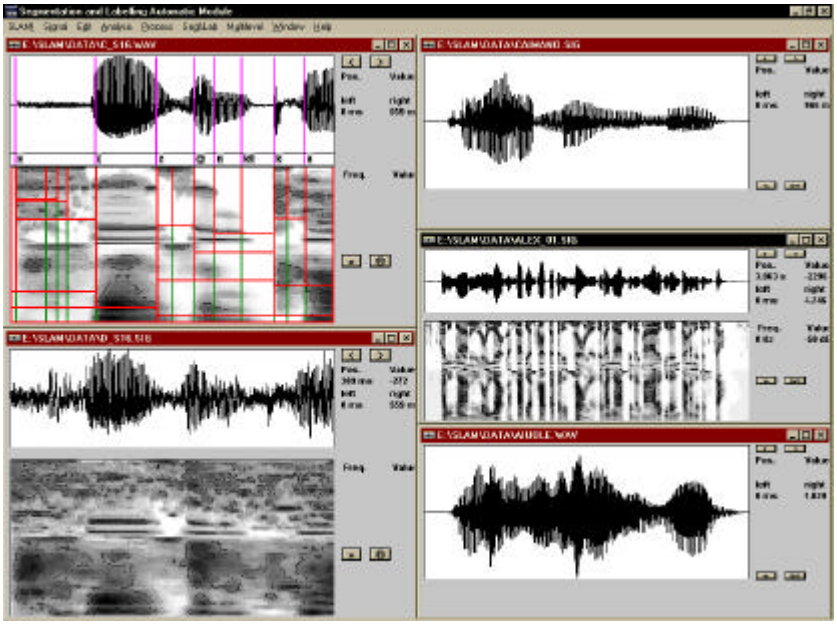

Figura 6. Esempio dell'applicazione simultanea di SLAM su più segnali.