UNIVERSITA' DEGLI STUDI DI NAPOLI

"FEDERICO II"

FACOLTA' DI SCIENZE MATEMATICHE FISICHE E NATURALI

TESI DI LAUREA IN FISICA

ALGORITMI PER LA DIVISIONE DEL SEGNALE VERBALE IN UNITA' SILLABICHE

RELATORI:

chiar.mo prof. G. Trautteur

chiar.mo dott. F. Cutugno

CANDIDATO:

Massimo Petrillo

Matr. 07/5881

Anno Accademico 1998-99

Indice

Introduzione

Il riconoscimento automatico del parlato (dall'inglese Automatic Speech Recognition d'ora in poi ASR), più in generale, l'elaborazione e la sintesi dei segnali vocali, sono aree disciplinari in grande fermento in questi anni. La ricerca tecnologica, infatti, è sempre più impegnata nella creazione di interfacce utente più versatili, nelle quali la voce è considerata uno strumento di ingresso/uscita molto semplice ed intuitivo per l'utente dei sistemi informatici. In questo ambito i sistemi di dettatura e i servizi di assistenza automatizzati dei *call-center*, sono solo alcuni degli esempi delle possibili applicazioni per le quali si cerca di migliorare le prestazioni.

I sistemi orientati alla applicazione però, sono stati finora poco utili a chiarire i meccanismi umani della comprensione del parlato, laddove gli sforzi sono stati rivolti più verso il raggiungimento di risultati "commerciali" a scapito degli aspetti più direttamente "cognitivi" che pure avrebbero potuto fornire valide integrazioni alle conoscenze fin qui sviluppate. A questo ha in parte provveduto in questi anni la

ricerca in ambito cognitivo, che si è a lungo interessata alla definizione di modelli che, a diversi livelli, hanno tentato di descrivere i meccanismi che sottostanno alla comprensione del messaggio verbale dell'uomo.

Ingegneria del parlato, linguistica e psicolinguistica hanno comunque colloquiato poco in questo tempo: avere tanti differenti approcci scientifici e culturali che mirano a rispondere agli stessi quesiti non ha finora deposto, almeno in questo caso, a favore della possibilità di trovare una soluzione organica ai problemi individuati.

La ricerca cognitiva, inoltre, è ulteriormente scissa in due branche, la fonetica percettiva e la psicolinguistica vera e propria. Le differenze tra le due branche potrebbero sembrare concentrate prevalentemente sul punto di partenza dell'indagine speculativa: laddove la prima sembra essere più sensibile alla specificità del codice acustico ed ai processi di decodifica necessari al riconoscimento di singole unità linguistiche, la seconda, considera poco rilevanti gli aspetti più direttamente collegati al codice per concentrarsi maggiormente sulle rappresentazioni mentali che, a livello cognitivo, costituiscono il primo stadio per l'elaborazione.

Né è ancora ben chiara l'influenza che la competenza linguistica può avere sulla percezione né quella che problemi specifici della natura del codice acustico possono avere sulla formazione delle rappresentazioni.

La pressoché totale assenza di considerazione nei confronti di queste problematiche nell'ambito della ricerca applicata, unita ai problemi specifici incontrati nello

sviluppo tecnologico, portano ad un attuale stato dell'arte dell'ASR poco ottimistico. I sistemi attualmente disponibili non sono in grado di andare oltre certi limiti come la dipendenza dal parlante, la limitazione del contesto lessicale, l'incapacità di riconoscere bene il parlato continuo eccetera .

In tempi recenti sono venuti alla luce molti studi che mirano al miglioramento dei sistemi ASR proponendo una maggiore integrazione fra le conoscenze specifiche dell'analisi di segnale con quelle derivanti dallo studio parallelo dei meccanismi cognitivi di riconoscimento del parlato e di quelli che derivano da un miglioramento delle conoscenze nell'ambito della fonetica strumentale [Rabiner & Juang, 1993].

Una delle conseguenze derivate da questo nuovo atteggiamento è quello di rimettere in discussione la durata temporale dell'unità minima di analisi dei sistemi per l'ASR che storicamente era sempre stata legata, in maniera più o meno diretta, a quella di singoli suoni della lingua: gli attuali sistemi di riconoscimento, infatti, operano su segmenti temporali piuttosto piccoli (circa 25ms). Le caratteristiche spettro-acustiche di ogni finestra vengono confrontate attraverso un'opportuna metrica, con quelle del repertorio di riferimento (solitamente derivato da un processo di "apprendimento"), al fine di individuare la classe di appartenenza (fonema) del suono esaminato. Più finestre consecutive vengono considerate come parte di un singolo suono linguistico se esse coincidono secondo la metrica usata. L'unità minima di analisi è quindi quella che, al variare del tempo, presenta delle caratteristiche acustiche stabili.

Ma questo approccio non è stato sufficiente per allestire dei sistemi di ASR versatili, in quanto la presunta stabilità acustica del segnale è spesso introvabile a causa dei fenomeni di variabilità presenti nel parlato. Quindi si stanno realizzando nuovi sistemi che invece si basano su unità linguistiche più grandi quali sono, appunto, le sillabe [Hauestein 1997] [Wu, *et alii* 1997].

Sebbene il ruolo della sillaba nei meccanismi della comprensione del parlato [Mehler, *et alii* 1981] [Cutler *et alii* 1986] fosse già da tempo noto, solo recentemente iniziano ad incontrarsi le prime proposte di segmentare il flusso continuo di segnale verbale in unità di dimensioni confrontabili con quelle di questa unità linguistica [Pfitzinger *et alii* 1996], [Reichl & Ruske, 1993], [Shastri, *et alii* 1999] come primo stadio di un sistema di riconoscimento automatico del parlato.

In particolare in questo modo risulta più facile esaminare anche le informazioni prosodiche, quelle che riguardano intonazione e velocità di eloquio, presenti nei segnali, le quali, pur contenendo, a parere dei linguisti, dati importanti per la comprensione del messaggio, sono state, spesso, ignorate dai sistemi sviluppati fino ad ora.

La disponibilità dei confini sillabici permette di aumentare, anche, le prestazioni di sistemi di riconoscimento più tradizionali, come mostrato in [Wu *et alii* 1997] poiché tali confini permettono di restringere notevolmente il campo di scelte tra cui deve operare il sistema di riconoscimento. Inoltre la possibilità di segmentare preventivamente il segnale in unità predefinite, come le sillabe, elimina il problema

della definizione della finestra di analisi poiché essa viene determinata in ogni punto del segnale in base alle sue proprietà acustiche.

I primi due capitoli del presente lavoro descrivono i concetti utili per focalizzare meglio il problema e per riassumere le nozioni necessarie a chi non si occupa di linguistica e per comprendere il senso di quando esposto nel corso della descrizione della procedura e dei risultati ottenuti. Mentre il primo avrà un impostazione più generica, il secondo capitolo affronterà più in dettaglio il concetto di sillaba sia dal punto di vista fonetico-acustico che fonologico-formale cercando di esprimere il rapporto tra i due punti di vista. Verranno passate in rassegna le diverse proposte di definizione per questo concetto, che è ancora oggetto di discussione tra fonologi e fonetisti.

Il terzo capitolo mostra la procedura proposta di segmentazione in sillabe. La segmentazione viene effettuata seguendo il profilo della curva di energia del segnale completo e della curva di energia del segnale filtrato con un filtro passa basso. La suddivisione del segnale viene effettuata individuando sulla curva di energia degli andamenti approssimativamente crescenti seguiti da un tratto decrescente. L'energia del segnale filtrato serve a selezionare ulteriormente le possibili sillabe e per stabilire meglio i loro confini. Il numero di parametri presente in questa procedura, pur essendo di diversi ordini di grandezza più piccolo rispetto ai pesi e alle soglie usati nelle rete neurali, ha reso necessario la definizione di una tecnica di determinazione

automatica, a partire da un corpus di addestramento, dei valori dei parametri che permettono di avere le migliori prestazioni da parte del sistema. A tal fine sarebbe stato necessaria la disponibilità di un corpus segmentato manualmente da linguisti esperti da usarsi come riferimento. Non essendo disponibile un siffatto corpus, si è provveduto a creane uno a partire da un corpus segmentato ed etichettato a livello fonetico, mediante una procedura di sillabificazione delle etichette fonetiche. Detto in altri termini la stringa di simboli è stata segmentata secondo le regole fonetiche-linguistiche ed è stata riportata sul segnale.

Le prestazioni del sistema vengono esaminate nel quarto capitolo, effettuando un confronto tra una segmentazione di riferimento effettuata da un fonetista contro quella prodotta dalla procedura su tre diversi corpora.

La segmentazione prodotta dal sistema riesce ad avere un tasso di errori complessivo mai superiore al 15% scendendo in alcuni casi fino al 7%. Gli errori commessi dalla procedura verranno classificati in modo da mettere in evidenza quelli che maggiormente contrastano con le regole della linguistica. Questa distinzione è dovuta al fatto che non tutte queste regole sono universalmente accettate, motivo per cui i risultati della procedura potrebbero essere usati dai linguisti per migliorare il loro accordo. Gli errori “più gravi” possono essere invece dovuti ad una implementazione che potrebbe essere ancora approssimativa o all’incertezza esistente nella segmentazione di riferimento. D’altra parte come mostrato in [Pfitzinger *et alii* 1996], l’accordo sulla segmentazione manuale in sillabe degli stessi

segnali da parte di più fonetisti non riesce mai ad andare oltre una percentuale del 96%.

I risultati ottenuti promettono, comunque, un uso soddisfacente nelle applicazioni di riconoscimento automatico del parlato.

Capitolo 1

Caratteristiche generali del segnale vocale ed aspetti linguistici[1]

1.1 Concetti introduttivi

La **linguistica** è lo studio scientifico del linguaggio. Tra le sue branche ci interessano in modo particolare la **fonologia** e la **fonetica**. La fonetica studia la produzione e la percezione dei suoni linguistici da parte degli esseri umani. Essa fornisce dei metodi per la descrizione, classificazione e trascrizione dei suoni. In fonetica vengono studiate le caratteristiche fisiche ed articolatorie, il suono pronunciato viene arbitrariamente diviso in porzioni, dette **foni**, riconducibili, sulla base dell'insieme delle loro caratteristiche acustiche, ad unità linguistiche minime. I vari foni possibili,

[1] L'impostazione di questo capitolo e la maggior parte delle immagini in esso presentate sono frutto dell'attenta lettura del manuale di fonetica di Albano Leoni e Maturi [1998] integrata per alcune definizioni da [Crystal, 1980].

vengono suddivisi in classi di equivalenza che vengono chiamate **fonemi** e che sono le unità minime della **fonologia**. Detto in altri termini, i fonemi sono le possibili classi di suoni previste da una lingua, mentre i foni sono la loro effettiva realizzazione.

La fonetica è, quindi, più legata al sostrato fisico, articolatorio ed acustico; mentre la fonologia è più astratta e, quindi, legata alle strutture linguistiche formali.

1.2 Anatomia

Per comprendere in pieno le caratteristiche acustiche del segnale vocale non si può prescindere dal meccanismo di produzione del parlato. Gli organi che permettono la fonazione fanno parte sia dell'apparato digerente che dall'apparato respiratorio. Il meccanismo principale, detto egressivo, comporta che l'aria uscita dai polmoni, incontrando degli ostacoli provochi, delle vibrazioni udibili come suoni.

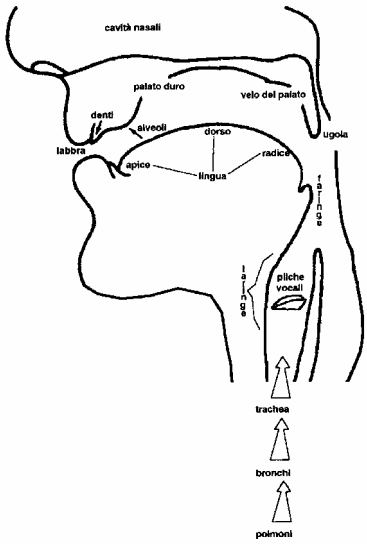

Figura 1.1 anatomia della fonazione.

La contrazione dei muscoli toracici genera una differenza di pressione tale che l'aria dall'interno dei polmoni viene convogliata attraverso i bronchi e la trachea verso la laringe dove la glottide costituisce la prima, eventuale, sorgente di vibrazioni. Essa è formata da

due pliche cartilaginee ricoperte da mucosa e dotate di piccoli muscoli che permettono di chiudere il passaggio dell'aria. L'aria passando attraverso le due pliche le mette in vibrazione producendo un suono periodico, questo viene poi filtrato dalla cassa di risonanza costituita dagli organi più esterni come la bocca, la faringe e le cavità nasali. I movimenti dei muscoli coinvolti nella produzione del parlato, **fonazione**, variano la forma della cassa di risonanza, – la risposta in frequenza e quindi il timbro del segnale glottico. Il velo del palato può mettere in comunicazione le cavità nasali con la faringe mentre la lingua, la mandibola e le labbra possono variare la forma della bocca. Altri meccanismi di articolazione sono *ingressivi* e *avulsivi*, in cui, rispettivamente, la fonazione avviene durante l'ispirazione o è completamente indipendente dalla respirazione.

1.3 Vocali

1.3.1 Caratteristiche articolatorie delle vocali

I foni prodotti mediante le vibrazioni delle pliche vocali senza nessuna restrizione al passaggio dell'aria nelle vie aeree superiori vengono detti vocali. La distinzione tra le vocali è dovuta alle posizioni reciproche degli organi mobili come la lingua, le labbra ed il velo del palato. Variando queste posizioni, si viene a formare di volta in volta una cassa di risonanza di forma diversa per il segnale glottico che modifica il timbro del suono prodotto dalle pliche vocali.

Dal punto di vista articolatorio le vocali vengono classificate secondo la forma assunta dalle labbra e la posizione della lingua. Le vocali che vengono pronunciate con un arrotondamento delle labbra vengono dette labializzate (ad es. le vocali nella parola *uomo*) mentre se le labbra sono distese si hanno le vocali non labializzate (ad es. le vocali nella parola *casa*)

I movimenti possibili della lingua foneticamente rilevanti sono:

- In senso orizzontale, limitatamente ai movimenti in avanti ed indietro[1]
- In senso verticale

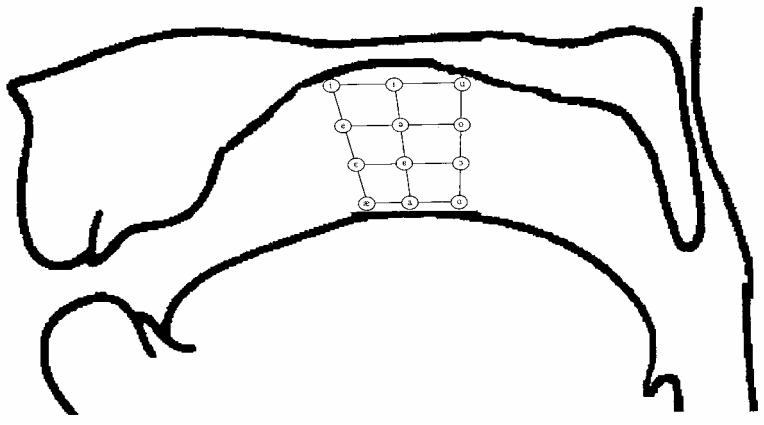

Figura 1.2 La posizione del centro della lingua distingue tra vocali alte e basse, e tra anteriori e posteriori.

Per classificare le vocali secondo la posizione della lingua si usa considerare la posizione di un ipotetico punto posizionato al centro del dorso della lingua. Le posizioni possibili di questo punto vengono rappresentate su un trapezio disposto tra la mandibola e la volta del palato. Le vocali pronunciate con la lingua abbassata vengono dette **basse** o **aperte** (ad esempio le vocali in *casa*), mentre se la lingua è

alzata si dicono **alte** o **chiuse** (ad esempio la vocale in *chi*) . Se la lingua è spostata in avanti si hanno le vocali **anteriori** (ad esempio le vocali in *linee*) mentre si dicono **posteriori** quando la lingua è spostata all'indietro (ad esempio le vocali in *buono*).

1.3.2 Caratteristiche acustiche delle vocali

Dal punto di vista acustico le vocali costituiscono delle porzioni del segnale con delle caratteristiche di periodicità. La frequenza fondamentale di queste porzione corrisponde alla frequenza di vibrazione delle pliche vocali, ed ha un importante valore prosodico. Infatti le variazioni della frequenza fondamentale tra varie vocali di una sequenza di parlato ne costituiscono l'intonazione. Una tentativo di caratterizzazione acustica viene fatta con lo studio delle caratteristiche spettro-acustiche del segnale: la glottide produce delle vibrazioni periodiche che vengono filtrate dal risonatore costituito dalle via aeree superiori. La loro diversa forma durante la pronuncia delle vocali corrisponde a diverse frequenze di risonanza. Nell'analisi spettrografica del segnale prodotto ciò comporta una diversa distribuzione dell'energia rispetto alle armoniche del segnale, le frequenze intorno alle quali viene concentrata l'energia vengono dette **formanti**.

---

[1] Non sono rilevanti i movimenti laterali destra-sinistra

Le prime due formanti sono quelle che maggiormente caratterizzano le vocali. La prima formante F_1 ha valori bassi per le vocali alte e valori alti quando la lingua è tenuta bassa, mentre F_2 è bassa per le vocali posteriori ed alta per le vocali anteriori.

1.4 Consonanti

1.4.1 Caratteristiche articolatorie delle consonanti

I foni prodotti creando delle restrizioni al flusso dell'aria vengono dette *consonanti*. La classificazione delle consonanti avviene attraverso tre parametri:

- Modo di articolazione
- Luogo di articolazione
- Vibrazione delle pliche vocali

Il modo di articolazione indica il tipo di ostacolo incontrato dall'aria e può essere di vari tipi:

- *Occlusivo,* quando si ha un'occlusione totale del flusso di aria causato dal contatto di due organi come le labbra o la lingua ed il palato, seguita dall'improvviso rilascio (es. le consonanti in *baco, patata*).
- *Fricativo,* nel caso dell'avvicinamento di due organi senza un vero e proprio contatto. In questo modo si produce una turbolenza che crea un rumore (es. le consonanti in *fase*).

- *Affricato,* se un ostacolo causato dallo stretto contatto di due organi viene rilasciato gradualmente generando comunque un rumore di frizione (es. le consonanti in *ciao, gigi*).
- *Nasale,* quando, in presenza di un'altra occlusione, il velo del palato viene abbassato in maniera da permette all'aria di defluire attraverso le cavità nasali, in questo modo si aggiunge un secondo risonatore (es. le consonanti in *mano*).
- *Laterale,* nel caso che un'occlusione centrale provocata dalla lingua permette il deflusso dell'aria lateralmente (es. le consonanti in *lello*).
- *Vibrante,* nel caso l'ostacolo prodotto da una debole occlusione intermittente (es. le consonanti in *raro*).
- *Approssimante,* per le articolazioni di incerta definizione perché a metà tra le vocali e le consonanti (es. il primo fono in *ieri*).

Il luogo di articolazione indica il tipo di ostacolo incontrato dall'aria, esempi per le consonanti in italiano sono:

- *Bilabiale,* contatto tra le labbra (es. le consonanti in *babà*).
- *Labiodentale,* tra denti e labbra (es. le consonanti in *vivo*)
- *Dentale,* con la lingua accostata agli incisivi superiori (es. le consonanti in *dato*)
- *Alveolare,* con la punta della lingua accostata alla base (alveolo) degli incisivi superiori (es. le consonanti in *sasso, zozzo*)
- *Palatale,* dorso della lingua a contatto con il palato (es. il primo fono in *gnomo*)

- *Velare*, dorso della lingua a contatto col velo del palato (es. le consonanti in *cocco*)

La vibrazione delle corde vocali distingue le consonanti *sonore* da quelle *sorde*.

Alla categoria delle consonanti vanno assimilate anche i suoni avulsivi e ingressivi.

Il quadro completo delle possibili classificazioni e riportato in figura 1.1 (figura allegata a fine capitolo).

1.4.2 Caratteristiche acustiche delle consonanti

Le caratteristiche acustiche delle consonanti variano moltissimo a seconda del tipo.

Le occlusive sono riconoscibili per un piccolo intervallo di silenzio corrispondente alla fase di occlusione, seguito, spesso, da un picco di energia corrispondente al rilascio dell'aria. Per distinguerle è utile osservare l'andamento delle formanti nelle vocali che seguono o precedono l'occlusiva. Come già detto le formanti sono in relazione con la posizione dei vari organi della fonazione, quindi, durante la pronuncia di una vocale seguita da un'occlusiva si comincia a preparare l'ostacolo per l'aria necessario per pronunciare l'occlusiva. Il movimento necessario provoca una variazione della forma della cassa di risonanza e quindi una variazione del valore delle formanti verso quelle che sarebbero le frequenze di risonanza corrispondenti alla consonante.

Le fricative sono caratterizzate dal cosiddetto rumore di frizione. Nel caso delle fricative sorde il suono è privo di una struttura formantica ma ha uno spettro che varia a seconda del luogo di articolazione. Le fricative sonore sono invece riconoscibili, oltre che per il rumore, anche per una propria, seppur debole, struttura formantica e per la notevole presenza di componenti a basse frequenza.

Nelle affricate è possibile distinguere una fase di occlusione seguita da frizione.

Le nasali hanno una struttura simile alle vocali, anche se più debole, in quanto non esiste un vero ostacolo al flusso dell'aria, il suono prodotto dalle corde vocali viene quindi modificato dalla cassa di risonanza formata dalle cavità nasali e dalla bocca. Per gli stessi motivi, anche le laterali hanno un comportamento simile alle vocali.

Le vibranti sono riconoscibili dall'intermittenza di piccoli intervalli di silenzio, corrispondenti alle occlusioni, e di rumore detti *spikes*.

1.5 Coarticolazione

Nel corso della produzione del parlato, i foni non vengono pronunciati isolatamente, ma vengono concatenati l'uno all'altro in rapida successione. Questo comporta che il passaggio da una configurazione all'altra degli organi della fonazione avvenga senza alcuna soluzione di continuità. In questo modo la produzione di ogni fono viene influenzata da quelli vicini dando luogo al fenomeno della coarticolazione. Per esempio nelle sequenze *fi*, *fu* e *fa* durante la pronuncia della fricativa la lingua e le labbra hanno già assunto la posizione relativa alla vocale successiva.

1.6 Sillabe

La definizione del concetto di *sillaba* è un argomento molto controverso. Essendo la fonologia un sistema formale, esistono delle regole di sillabificazione fonologiche ben definite. La stessa cosa non è possibile in fonetica, dove non è stato possibile trovare delle regolarità che permettano di descrivere in maniera coerente tale concetto.

In fonologia viene assegnato un valore, detto *sonorità*[1], ad ognuno dei fonemi della lingua studiata. Partendo da questi valori vengo stabilite delle regole in grado di sillabificare, senza incertezze, studiando l'andamento della sonorità.

In fonetica, sono stati numerosi i tentativi di definizione, sia di tipo articolatorio che acustico. Ma nessuna delle definizioni proposte è stata, tuttora, ritenuta soddisfacente. La natura sfuggente della sillaba ha portato qualcuno, come Kohler [Kohler, 1966], a proporre di abbandonare tale concetto anche in fonologia, cosa che non è stata fatta poiché è stato dimostrato il potere descrittivo superiore dei modelli fonologici che usano il concetto di sillaba ([Nespor, 1993]).

Agli spinosi problemi sommariamente presentati in questo paragrafo introduttivo sarà dedicato l'intero capitolo 2.

---

[1] La sonorità è una proprietà determinata, in realtà, dall'osservazione delle realizzazioni fonetiche dei fonemi, e corrisponde, grossomodo, articolatoriamente all'*apertura* del tratto vocalico e acusticamente all'intensità sonora .

1.7 Prosodia

Nello studio e nella comprensione del parlato sono molto importanti altri fattori psicolinguistici che sono in grado di mettere in risalto delle parti rispetto a delle altre o a dare un preciso significato a ciò che viene pronunciato. Un velocità di eloquio più bassa o una intensità più alta mettono in risalto quelle cose a cui bisogna fare più attenzione, mentre la diversa intonazione può dare significati diversi alla stessa sequenza di fonemi.

I parametri fondamentale nell'analisi prosodica di una sequenza di foni sono:

a) durata
b) intensità
c) altezza o pitch

1.7.1 Intensità

La sensazione di intensità con cui viene percepito un suono dipende principalmente dalla sua energia, ma essa dipende anche dalla distribuzione di quest'ultima nello spettro: per esempio suoni di pari energia ma di diversa frequenza fondamentale possono dare sensazioni diverse. Per questo motivo in psicoacustica sono state

compilate delle tabelle che fanno corrispondere ad ogni valore dell'energia e della frequenza fondamentale una stima della sensazione di intensità, misurata in *son*[1].

Spesso si preferisce, però, usare il rapporto tra l'energia del suono e quello di un suono campione universalmente accettato come tale, ottenendo quella che, comunemente, viene chiamata intensità relativa. Per ottenere valori più vicini alla sensazione uditiva si moltiplica per dieci il logaritmo decimale di questo rapporto, ottenendo il valore in decibel dell'intensità:

$$I = 10 \log \frac{E}{E_0}$$

Dove E indica l'energia del suono ed E_0 l'energia del suono campione.

1.7.2 Durata

La durata dei singoli segmenti è un fattore fondamentale nello studio della velocità di eloquio e nell'assegnazione dell'accento. L'allungamento della durata di alcuni segmenti, viene usata dai parlanti, infatti, per porre maggior risalto a parti di un discorso rispetto a delle altre (un esempio potrebbe essere la frase: *ti ho detto che devi stare ziiiiittooo*). In talune lingue, la durata dei segmenti ha anche una valenza semantica, due parole possono avere significati diversi a seconda della durata di un

[1] Ad un tono di frequenza 1000Hz ed energia 40dB corrisponde l'intensità di un *son*. La relazione tra

singolo fono (es. in inglese le parole *ship* [tr. fon. SAMPA /Sip/, nave] e sheep [/Si:p/, pecora] si differenziano nella pronuncia solo per la durata della vocale) .

Si possono considerare come segmenti sia i foni che le sillabe. Nella scelta del tipo di segmento da analizzare è molto importante considerare il metodo di riconoscimento. Poiché la fonologia sembra suggerire un metodo semplice per individuare le sillabe, si è preferito usare queste ultime.

La scala di sonorità cerca di rappresentare il grado di apertura di un fonema, cioè la larghezza del condotto vocale durante la sua pronuncia. Questa stima, in maniera del tutto empirica ed arbitraria, sembra essere in diretta corrispondenza con l'intensità, e quindi l'energia, dei foni corrispondenti. Per sillabare una sequenza di parlato, quindi, si potrebbe analizzare la curva di energia della stessa: i minimi indicherebbero i confini delle sillabe.

Questo tipo di segmentazione si basa su una corrispondenza energia-sonorità che non è stata dimostrata, e che, come vedremo, non sempre sussiste.

1.7.3 Frequenza fondamentale

La frequenza fondamentale corrisponde alla grandezza psicoacustica *altezza*, (*pitch* in inglese). Come nel caso dell'intensità, anche per l'altezza non esiste una proporzionalità diretta tra le due grandezze, motivo per cui sono state definite, sulla

---

decibel e *son* non è lineare e varia al variare della frequenza.

base dei risultati di alcuni test percettivi, delle tabelle di conversione dagli *Hertz* verso l'unità usata in psicoacustica, *mel*[1].

I musicisti, invece, sono soliti usare per descrivere lo stesso tipo di sensazione uditiva il concetto di **semitono**. La differenza di **altezza** tra due suoni viene qui espressa secondo la formula:

$$D = 12 \log_2 \frac{F_1}{F_2}$$

dove D rappresenta la distanza in semitoni corrispondente all'intervallo in Hz (F1-F2).

Questa scala presenta il vantaggio di esprimere una precisa relazione con la grandezza fisica *frequenza* e. allo stesso tempo, ben rappresenta le caratteristiche di sensibilità e potere risolutivo in frequenza tipiche del sistema uditivo.

1.7.4 Analisi prosodica

L'esame dei parametri prosodici permette di suddividere un testo parlato in unità prosodiche, dette *unità tonali*, che rappresentano delle sequenze prosodicamente indipendenti che possono coincidere con delle unità di informazioni non necessariamente coincidenti con frasi di senso compiuto. I confini delle unità tonali sono identificabili per la presenza di alcuni dei seguenti fenomeni:

- calo dell'intensità;
- calo della frequenza fondamentale;

---

[1] Un tono puro di 1000Hz mel corrisponde a *possiede* un pitch corrispondente a 1000*mel*. La

- rallentamento della velocità di eloquio;
- presenza di una pausa.

relazione tra Hertz e *mel* non è lineare.

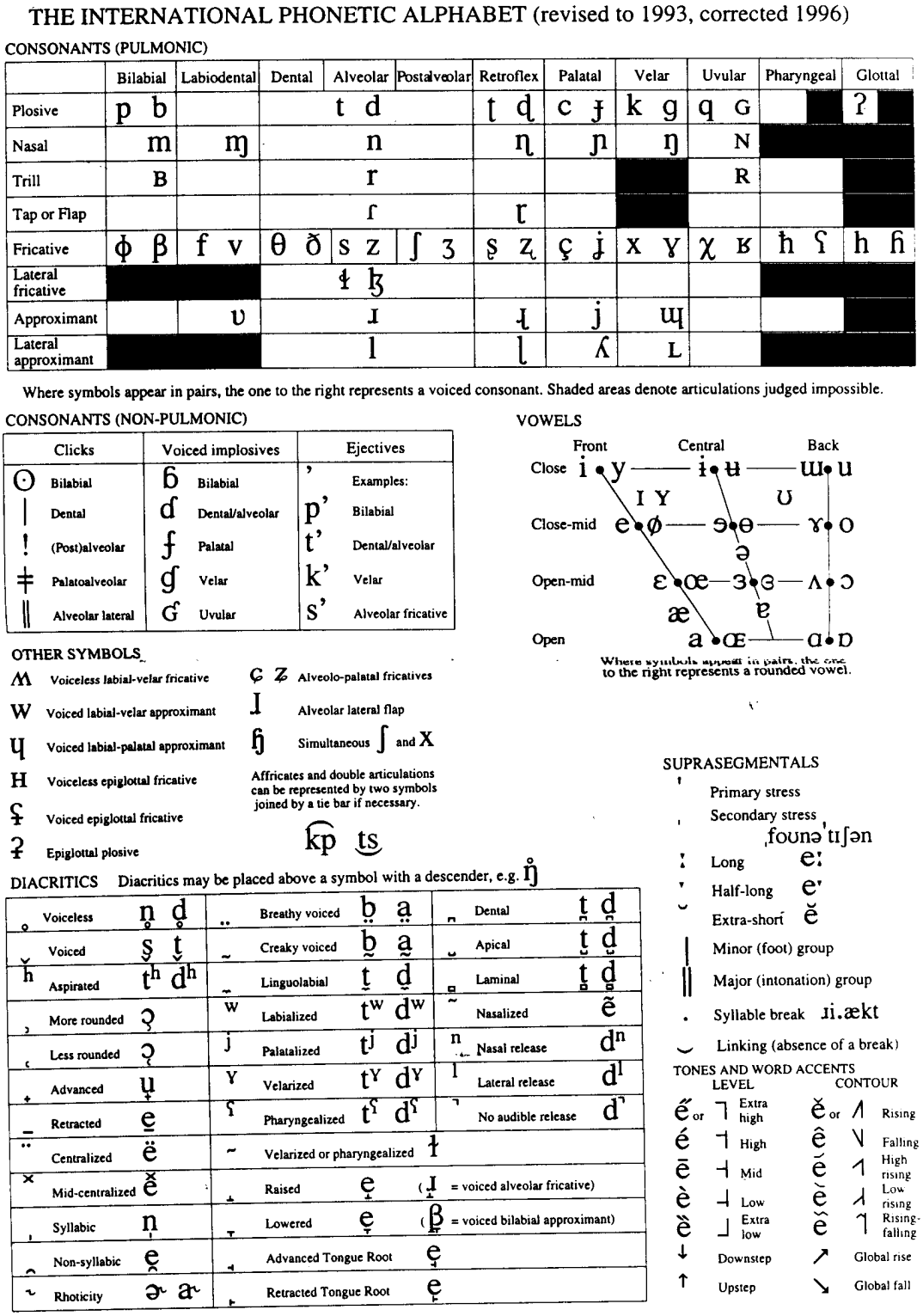
THE INTERNATIONAL PHONETIC ALPHABET (revised to 1993, corrected 1996)
CONSONANTS (PULMONIC)
Bilabial
Labiodental
Dental
Alveolar
Postalveolar
Retroflex
Palatal
Velar
Uvular
Pharyngeal
Glottal
Plosive
Nasal
Trill
Tap or Flap
Fricative
Lateral fricative
Approximant
Lateral approximant
Where symbols appear in pairs, the one to the right represents a voiced consonant. Shaded areas denote articulations judged impossible.
CONSONANTS (NON-PULMONIC)
Clicks
Bilabial
Dental
(Post)alveolar
Palatoalveolar
Alveolar lateral
Voiced implosives
Bilabial
Dental/alveolar
Palatal
Velar
Uvular
Ejectives
Examples:
Bilabial
Dental/alveolar
Velar
Alveolar fricative
VOWELS
Front
Central
Back
Close
Close-mid
Open-mid
Open
Where symbols appear in pairs, the one to the right represents a rounded vowel.
OTHER SYMBOLS
Voiceless labial-velar fricative
Voiced labial-velar approximant
Voiced labial-palatal approximant
Voiceless epiglottal fricative
Voiced epiglottal fricative
Epiglottal plosive
Alveolo-palatal fricatives
Alveolar lateral flap
Simultaneous ʃ and X
Affricates and double articulations can be represented by two symbols joined by a tie bar if necessary.
SUPRASEGMENTALS
Primary stress
Secondary stress
ˌfoʊnəˈtɪʃən
Long
Half-long
Extra-short
Minor (foot) group
Major (intonation) group
Syllable break ɹi.ækt
Linking (absence of a break)
TONES AND WORD ACCENTS
LEVEL
CONTOUR
Extra high
High
Mid
Low
Extra low
Rising
Falling
High rising
Low rising
Rising-falling
Downstep
Global rise
Upstep
Global fall
DIACRITICS Diacritics may be placed above a symbol with a descender, e.g. ŋ̊
Voiceless
Voiced
Aspirated
More rounded
Less rounded
Advanced
Retracted
Centralized
Mid-centralized
Syllabic
Non-syllabic
Rhoticity
Breathy voiced
Creaky voiced
Linguolabial
Labialized
Palatalized
Velarized
Pharyngealized
Velarized or pharyngealized
Raised
(ɹ̝ = voiced alveolar fricative)
Lowered
(β̞ = voiced bilabial approximant)
Advanced Tongue Root
Retracted Tongue Root
Dental
Apical
Laminal
Nasalized
Nasal release
Lateral release
No audible release

Figura 1.1

Capitolo 2

Le sillabe

2.1 Perché la sillaba?

Gli attuali sistemi di ASR usano modelli che vedono il parlato come una sequenza di parole a loro volta composte da unità minime corrispondenti ai foni. Poiché è difficile che ogni fono in una parola venga pronunciato in maniera standard, i sistemi ASR spesso sono basati su un lessico con più pronunce della stessa parola, questo per far corrispondere più sequenze acustiche ad un unica unità lessicale. Poiché ci sono, in pratica, diversi modi di pronunciare una parola, questo approccio standard aggiunge un livello di complessità e di ambiguità nel processo di decodifica., che, se modificato, può portare ad un miglioramento delle prestazioni di riconoscimento.

Un modo di migliorare le prestazioni è quello di provare a considerare come unità minime di analisi porzioni di segnale più ampie dei foni, per ridurre l'influenza della coarticolazione, e al tempo stesso più piccole delle parole per ridurre gli effetti della variabilità che si riscontra nella pronuncia dei suoni.

E' forse questo il motivo per cui, nell'ambito delle ricerche sull'ASR sta aumentando, negli ultimi anni, l'interesse verso il concetto di sillaba. Detto in altri termini, sembrerebbe che una maggiore *"comprensione della comprensione del parlato"* (trad. letterale da [Greenberg, 1996]), dovrebbe portare ad una scelta più mirata dell'*unità di partenza*[1] che sia quindi quanto più collegata con i vincoli fisici e percettivi che sono insiti nella struttura del linguaggio parlato.

Ci sono molti motivi per i quali la sillaba si trova ad assumere un ruolo di primaria importanza ne elenchiamo alcuni:

1) Il sistema uditivo umano analizza 200ms di parlato alla volta
2) ricerche psicolinguistiche hanno mostrato che, in compiti di identificazione di stimoli verbali, i tempi di reazione al riconoscimento delle sillabe sono più bassi che non per i fonemi [Mehler, *et alii* 1981] [Cutler *et alii* 1986];

[1] Si preferisce, qui, parlare di *unità di partenza*, piuttosto che di *unità minime*, poiché nulla vieta la possibilità, in seguito, di ulteriori segmentazioni.

3) i fenomeni di coarticolazione sono confinati, spesso, all’interno di una sillaba e in poche situazioni estendono il loro effetto a porzioni di dimensioni superiori;
4) il dominio dei lapsus è spesso di tipo sillabico e più raramente coinvolge il livello fonemico;
5) la sillaba sembra costituire unità elementare fondamentale nei processi di apprendimento infantile del linguaggio [Elliot 1981];
6) la sillaba, in quanto unità più stabile e vincolata, risente di meno degli effetti della variabilità e della coarticolazione: un’analisi sistematica [Greenberg, 1998] delle variazioni di pronuncia in un corpus di parlato spontaneo inglese (Switchboard) ha dimostrato l'esistenza di alcune sistematicità nella pronuncia delle sillabe, è stato infatti osservato che gli incipit delle sillabe sono pronunciati in maniera "canonica" molto più frequentemente del nucleo e della coda. La pronuncia accurata dell'inizio della sillaba, in media, potrebbe quindi costituire un elemento di identificazione utile nell'analisi della catena ininterrotta di parlato.

Sulla base di quanto fin qui affermato, dunque, un sistema di riconoscimento basato sulle sillabe potrebbe assomigliare maggiormente al processo di percezione dell’uomo e quindi avere più possibilità di successo.

Haustein [1997] ha provato a confrontare due sistemi di ASR, per altro simili, basati l’uno sulle sillabe e l’altro sui foni, notando dei notevoli vantaggi del primo

nell’analisi del parlato in presenza di rumore, suggerendo che un sistema basato sulle sillabe possa, appunto, assomigliare meglio al sistema uditivo umano. I sistemi analizzati da Haustein, non segmentavano preventivamente il segnale in sillabe o foni, ma cercavano di individuare in finestre di ampiezza fissa la presenza delle une o degli altri. Ma già quattro anni prima Reichl e Ruske [1993] tentarono di segmentare il segnale in sillabe con l’uso di reti neurali, dando l’avvio ad un nuovo modo di pensare nell’ambito dell’ASR. Attualmente il gruppo dell’ICSI di Berkeley [Greenberg, 1998] sta continuando su questa strada, sviluppando dei sistemi che, partendo dalla suddivisione del segnale in sillabe, giunga al riconoscimento delle unità linguistiche.

Questo lavoro si occuperà di trovare un algoritmo che suddivida il segnale in porzioni simili alle sillabe da usarsi come prima parte di un sistema di analisi del parlato che tenga conto, anche di informazioni prosodiche. L’output dell’algoritmo potrebbe quindi essere usato sia per il riconoscimento che come informazione prosodica a se stante.

Verrà fatta anche un’analisi qualitativa degli errori commessi dalla procedura, nel tentativo di ottenere delle informazioni importanti dal punto di vista linguistico. Insito in questa procedura c'è però un rischio di circolarità: il dato che deriva dal processo di segmentazione automatica deve essere comparato con i risultati di un processo di sillabificazione astratto basato su regole formali. A loro volta tali regole sono comunque state derivate dalla definizione che della sillaba si da a partire dalla

osservazione dei fenomeni fisici e linguistici. Non si deve quindi rischiare che il confronto fra le sillabe ricavate dal processo di segmentazione automatica e quelle derivate dalla applicazione delle regole si risolva in una identità e, allo stesso tempo, ci si porrà il problema di stabilire se eventuali differenze fra le unità prodotte e quelle attese debbano essere attribuite ad errori del processo automatico o a "falle" della teoria.

Detto in altri termini: ogni volta che il nostro sistema seleziona una porzione di segnale, nel caso in cui il contenuto fonetico di quella porzione non coincida con le aspettative sillabiche dovremmo concludere che è il sistema a commettere un errore o è la regola di sillabificazione che deve essere rivista?

2.2 Definizione della sillaba in Linguistica

Le possibili definizioni del concetto di sillaba in linguistica dovrebbero considerare gli aspetti formali che le caratteristiche articolatorie ed acustiche della produzione del parlato. In linea di principio, potrebbero non esserci delle convergenze tra le descrizioni proposte in fonologia e quelle proposte in fonetica, specialmente per quanto riguarda la individuazione dei confini e nel ruolo stesso della sillaba nei fenomeni prosodici.

2.3 La sillaba in Fonologia

La Fonologia classifica i suoni di una lingua attraverso una distinzione tra classi e varianti. Le idee fondamentali della Fonologia (così come sono state espresse nei "Traveaux du Circle Linguistique de Prague" , il Circolo di Praga, che dopo il congresso dell'Aja del 1928, riuniva eminenti linguisti europei attorno alla figura di N. Trubeckoj) si basano sull'uso dei fonemi come unità di analisi dei suoni della lingua, definiti in base ai loro tratti distintivi. Il sistema di Trubeckoj (edizione italiana [Trubeckoj, 1971]) si basava sulla dicotomia tra classi e varianti, dove le classi sono i fonemi e le varianti sono i foni. Poiché tra i fonemi è possibile determinare delle relazioni, come ad esempio le "opposizioni funzionali"[1], e possibile considerali come un sistema.

A questo punto la sillabificazione di una sequenza di fonemi viene eseguita attraverso l'individuazione di picchi vocalici, **nuclei**, attorno ai quali si combinano altri fonemi formando l'**attacco**, o **incipit**, come segmento iniziale e la **coda** come

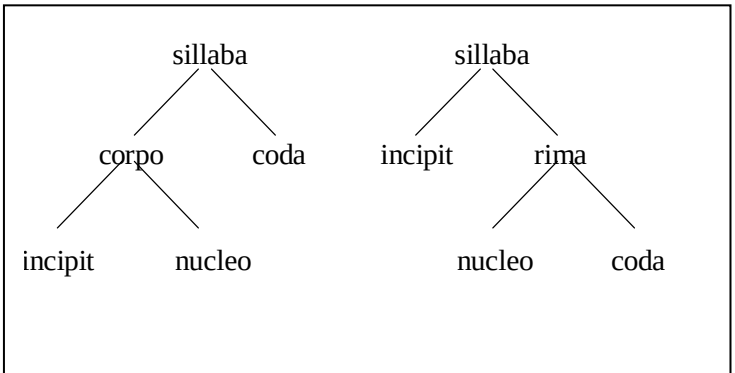

Figura 2.1 possibili strutture sillabiche

[1] Due fonemi vengono considerati distinti se esistono due parole che si distinguono solo per la presenza di uno di essi, esempio: *caso* e *casa.*

segmento finale. Nella definizione della sillaba dovranno essere presenti anche i criteri per individuare i fonemi che fanno da confine. La struttura della sillaba può quindi essere considerata sia come un incipit seguita da una **rima**, nucleo e coda, sia come un **corpo**, incipit e nucleo, seguita da una coda.

La struttura di base, secondo alcuni universale, della sillaba sarà dunque: (C)V(C), dove C indica una consonante od una approssimante, V indica una vocale, o comunque il nucleo della sillaba caratterizzato dal tratto [+ sillabico] e le parentesi tonde stanno a significare la non obbligatorietà dei segmenti di incipit e di coda. In Italiano le strutture sillabiche esistenti sono:

V, VC, CV, CVC, CCV, CCVC

Le regole di sillabificazione vengono, quindi, formalizzate in base al *principio della scala di sonorità* fondato sulle caratteristiche intrinseche della produzione dei suoni linguistici. A questo devono, però, affiancarsi regole di combinazione dei suoni, regole fonotattiche, che variano a seconda delle lingue.

2.3.1 La «scala di sonorità» e le regole fonotattiche

La scala di sonorità prevede l'assegnazione di un valore di sonorità per ciascun fonema, valori che sono tendenzialmente non arbitrari, almeno a livello

fonologico; così alle vocali viene assegnato il valore massimo, alle occlusive sorde quello minimo. Questi valori definiscono un gradiente per la classificazione dei fonemi sulla base della loro sonorità (vedi tabella 2.1)

Il principio della scala di sonorità permette di individuare i nuclei delle sillabe come i punti di massimo della sonorità.[1] Pur essendo di validità universale, questo principio non basta per individuare i confini, motivo per cui occorre utilizzare le regole fonotattiche.

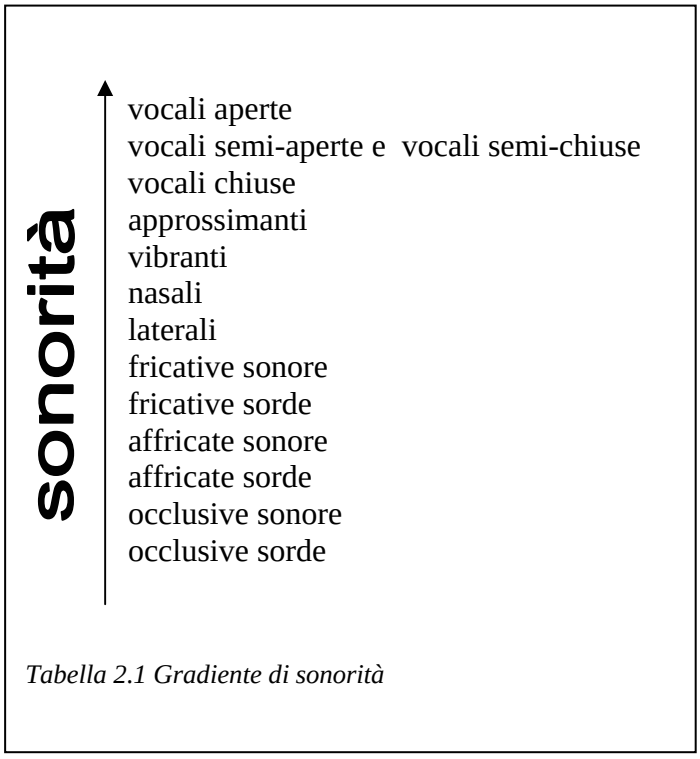

Tabella 2.1 Gradiente di sonorità

Secondo i criteri fonottatici della lingua italiana una sillaba può essere chiusa (può avere cioè come confine una coda consonantica) da una sonorante (nasale, laterale o vibrante), od anche da una consonate doppia (geminata), se si considera questa come la somma di due segmenti uguali (caso in cui anche occlusive, fricative

[1] Il concetto di sonorità viene spesso paragonato all'intensità media con cui viene prodotto un fonema o all'apertura del tratto vocalico

ed affricate possono occupare la posizione di coda sillabica); inoltre, considerando i foni approssimanti segmenti consonantici, la sillaba che contenga un dittongo discendente, formato da una vocale e da una approssimante, sarà una sillaba chiusa (C)VC.

Un altro principio di ispirazione fonotattica consiste nel fare riferimento ai gruppi consonantici ammessi nella lingua ai confini di parola per individuare i confini sillabici. Ad esempio, una parola come «apri», non può essere sillabificata come [ap.ri], dal momento che un'occlusiva sorda non può rappresentare il confine di una parola secondo le regole fonotattiche dell'italiano.

Altre due regole, da applicarsi in sequenza, di inspirazione fonottattiche sono quelle del **massimo attacco** e dell'**assegnazione in coda** per risolvere il caso in cui siano possibili diverse segmentazioni basate sui possibili inizi e fine di parole. Secondo il principio del "massimo attacco" vanno assegnati quanti più fonemi possibili all'incipit della sillaba successiva, rispettando comunque il principio della sonorità, mentre i fonemi che non è stato possibile assegnare all'incipit perché verrebbe violato il principio della sonorità, in base al all'assegnazione in coda, vanno a chiudere la sillaba precedente.

Queste regole, o principi, sono di natura strettamente fonologica, vale a dire, si fondano essenzialmente sulla distribuzione dei fonemi per una determinata lingua, guardando alla fonotassi ed alle regole che questa prevede per la combinazione dei fonemi. Concludendo: la sillaba definita fonotatticamente è funzione dei confini

possibili che la determinano. La definizione fonologica della sillaba, pertanto, sarà essenzialmente una definizione dei suoi confini, o meglio, una definizione dei principi che portano alla loro identificazione. La definizione di Pulgram [1970] «la sillaba è un'unità fonologica i cui confini sono fonotatticamente determinati» è forse la più onesta e completa, sotto questo punto di vista.

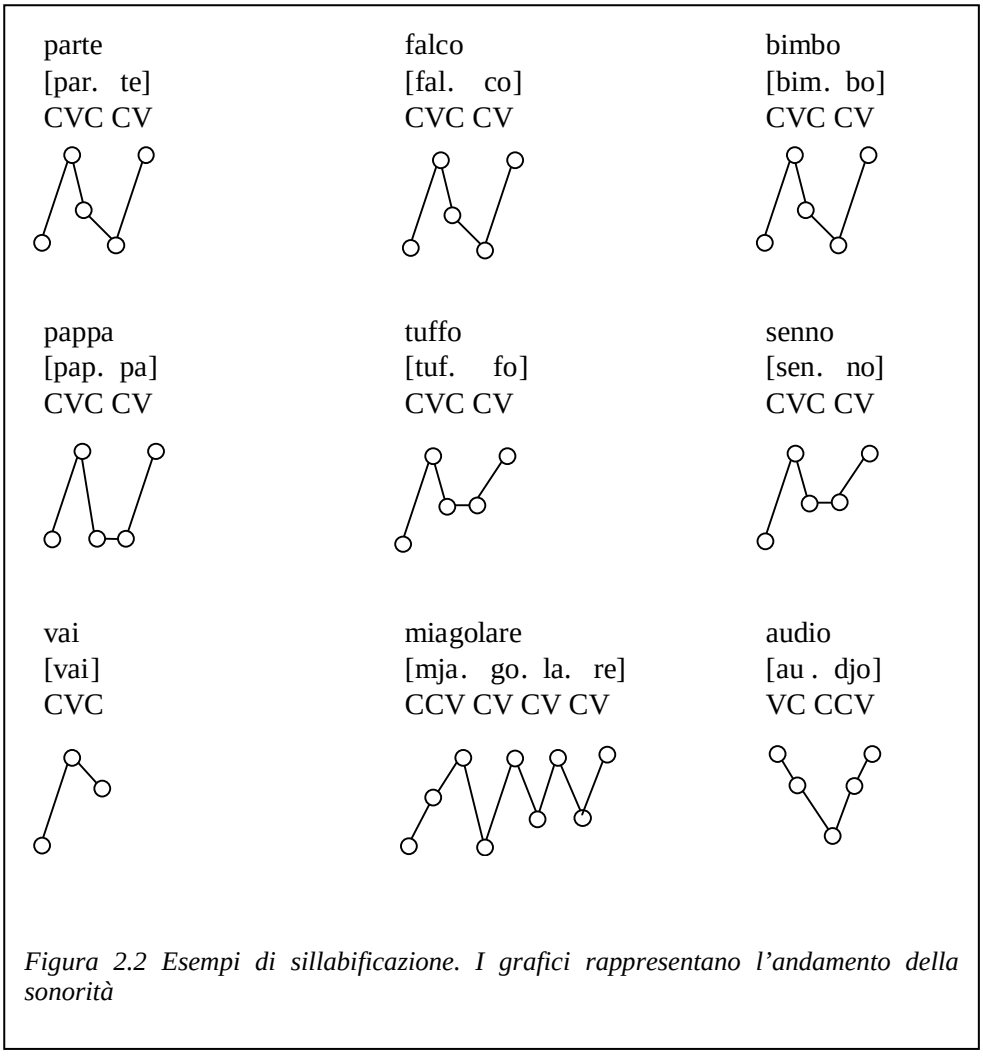

Figura 2.2 Esempi di sillabificazione. I grafici rappresentano l'andamento della sonorità

2.4 La sillaba in Fonetica

I tentativi di definizione della sillaba in fonetica sono essenzialmente basati sulle caratteristiche del segnale acustico o delle fisiologia dell'articolazione dei suoni linguistici. Nella storia della fonetica moderna sono stati presi in considerazione l'apertura del tratto articolatorio [De Saussure, 1967], la tensione ed il rilascio degli organi della fonazione, le pressione toracica: il cosiddetto "impulso toracico" [Stetson, 1951] e la variazione dell'attività neuromuscolare, cercando di definire la sillaba a partire da questi parametri.

Particolarmente interessante è la tesi secondo la quale è possibile individuare le sillabe seguendo l'andamento dell'intensità del suono. I picchi di intensità corrisponderebbero ai nuclei, mentre i minimi i punti di confine [Jespersen, 1920]. Questa formulazione trae ispirazione dalla scuola fonetica tedesca di fine ottocento cui si devono alcune tra le definizioni della «scala di sonorità» (più tardi anche in [Grammont 1946]), riprese poi nelle moderne teorie fonologiche. Privilegiando l'aspetto articolatorio su quello acustico, la sillaba può essere definita come unità articolatoria: i comandi motori sarebbero programmati ed attivati congiuntamente per una determinata sequenza fonica, corrispondente, appunto, alla sillaba [Koschevnikov & Chistovich, 1966].

In questo modo la definizione della sillaba tende a coincidere con l'identificazione e la definizione di un'unità minima di programmazione del parlato [Fry, 1964].

Per ciascuna di queste definizioni e per ciascuno dei criteri utilizzati sono però stati opposti controesempi, motivo per cui non esiste ancora una definizione universalmente accettata di sillaba in Fonetica (per una rassegna completa di proposte di definizioni e relative opposizioni si veda [Bertinetto, 1980]).

Il problema della definizione della sillaba è talmente complesso che, in passato, c'è stato chi ha messo in discussione la natura di concetto linguistico. Esemplare, in questo senso la critica di Kohler [1966]: «the syllable is unnecesary...impossible....harmful»; che partendo dalla, pretesa, sostanziale arbitrarietà della segmentazione e sul vizio di circolarità che introdurrebbe il riconoscimento di questa unità. Ma è un controesempio significativo osservare la capacità del parlante nativo di sillabare una parola, o una sequenza di parole in una frase, senza per questo padroneggiare una definizione formale di sillaba.

2.5 Sillabificazione del segnale

L'idea di Jespersen di segmentare secondo l'andamento dell'energia è quella che più di ogni altra sembra tener conto anche delle altre ipotesi di segmentazione. La segmentazione fonologica, basata sulla scale di sonorità, trae, implicitamente, spunto

da questa idea, mentre non sono da escludersi delle correlazioni con le ipotesi articolatorie: l'unità articolatoria e l'unità di programmazione non escludono che al loro interno avvenga un aumento dell'energia del segnale prodotto seguito da una discesa. Se mai, considerando l'unita articolatoria come l'insieme dei movimenti prodotti durante l'apertura e la successiva chiusura del tratto vocalico, è lecito aspettarsi un massimo dell'energia in corrispondenza della massima apertura.

Si può quindi partire da questa ipotesi per creare una procedura di sillabificazione del segnale. Benché essa sia molto semplice i migliori sistemi di sillabificazione, ad esempio [Reichl & Ruske, 1993] e [Shastri *et alii* 1999], fanno ricorso all'uso di reti neurali per distinguere i massimi dell'energia corrispondenti ai nuclei sillabici da quelli occasionali dovuti ad altri tipi di oscillazioni (per es. rumore o frequenza fondamentale). Sembra, però, eccessivo l'uso di strumenti così potenti per risolvere un problema che può essere affrontato con l'uso di tecniche più semplici. Viene qui proposto un sistema algoritmico molto più semplice. L'uso di tecniche di tipo procedurale promette di illustrare meglio i meccanismi di riconoscimento del parlato diversamente da quanto permesso dalle decine di migliaia di parametri di una rete neurale.

Capitolo 3

Segmentazione in sillabe

3.1 Descrizione degli strumenti utilizzati

Scopo di questo capitolo è la descrizione degli algoritmi di estrazione dei parametri prosodici dal segnale vocale. L'ambiente di sviluppo scelto è il Matlab. I linguaggi *interpretati*, quale è il Matlab, non riescono a produrre delle procedure molto veloci, ma la maggior flessibilità rispetto ai linguaggi *compilati* e la presenza di numerose librerie matematiche e di analisi dei segnali lo rendono più adatto alle nostre esigenze. Per esempio, con il Matlab è possibile provare alcune idee direttamente dalla riga di comando, mentre un linguaggio compilato avrebbe comportato la scrittura di un programma e la relativa compilazione.

I segnali utilizzati dalla procedura devono essere campionati alla frequenza di 22050Hz come compromesso tra l'esigenza di ridurre l'occupazione di memoria e conservare la qualità del segnale.

Nella descrizione che segue si utilizzerà il termine ***sillaba*** per indicare i segmenti previsti dalla fonologia o quelli usati come riferimento, il termine ***segmento*** sarà invece usato per indicare quelli prodotti dal sistema.

3.2 Calcolo della curva di energia

L'energia di una sequenza $s=\{s_0\ s_1\ \ldots\}$ viene definita come:

$$E = \sqrt{\sum_i s_i^2}$$

Questa espressione fornisce l'energia della sequenza completa e non il suo andamento nel tempo. Il presente lavoro è invece basato proprio su quest'ultimo dato. Per ottenerlo è necessario spezzare la sequenza in tante porzioni contigue di lunghezza fissa e poi calcolare l'energia di queste sottosequenze. La lunghezza di queste porzioni (finestre) è molto importante e deve essere decisa con una certa cura, poiché se essa è molto grande si perde molta informazione. Viceversa per segmenti molto piccoli si possono avere oscillazioni che

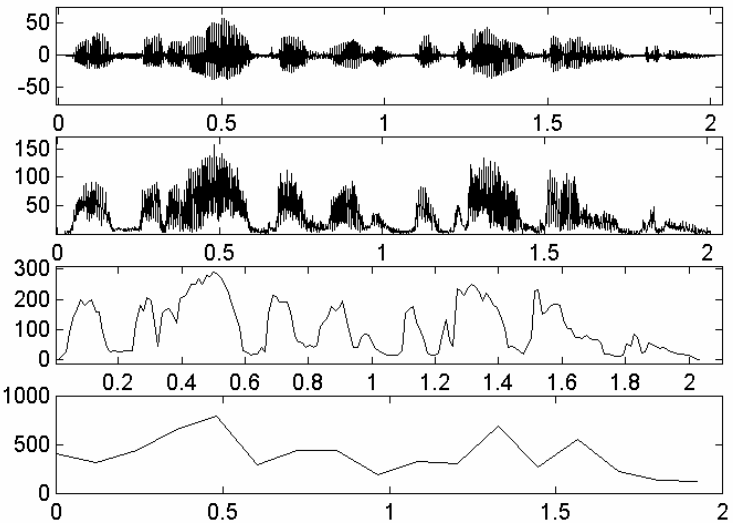

Figura 3.1 Variando la lunghezza della finestra usata per il calcolo dell'energia, si ottengono andamenti diversi: in ordine i grafici rappresentano il segnale originale e l'andamento dell'energia per finestre di 1.3ms, 13ms e 130ms.

non sono significative per i nostri scopi. Nel nostro caso la lunghezza della finestra dovrebbe essere inferiore alla durata minima di una sillaba e superiore al massimo periodo delle oscillazioni delle corde vocali in modo da poter seguire l'andamento dell'energia nelle sillabe senza osservare oscillazioni prive di interesse prosodico. Non è sempre possibile però rispettare questi due vincoli fra loro contrastanti, motivo per cui si è scelto di usare finestre della durata di 11ms. Se da un lato la durata delle sillabe è, praticamente, sempre superiore a questo valore, dall'altro dei maschi adulti possono avere un periodo di vibrazione delle pliche vocali superiore ad 11ms. La curva di energia così calcolata potrà quindi presentare delle oscillazioni indesiderate in presenza di voci con frequenza fondamentale molto bassa. La procedura di segmentazione dovrà quindi essere in grado di discriminare tra le oscillazioni utili da quelle dovute all'uso di una finestra troppo piccola.

Poiché l'intensità percepita dall'udito non è direttamente proporzionale all'energia del segnale, ma esiste, fra la sensazione di intensità e l'energia del segnale una corrispondenza approssimativamente logaritmica, si definisce, qui, l'**intensità** come il logaritmo naturale dell'energia. Ciò allo scopo di considerare solo le variazioni dell'energia che siano percettivamente significative.

Dal momento che siamo interessati principalmente ai rapporti tra i valori dell'intensità, la curva così definita non dipende dai valori assoluti.

3.3 Determinazione dei marker sillabici

La procedura di determinazione dei confini sillabici dipende da un insieme di venti parametri. La giustificazione dei valori di questi parametri costituirà un paragrafo a parte. Nella descrizione che segue useremo dei caratteri in ***grassetto corsivo*** per indicare i parametri usati.

I stadio: segmentazione preliminare. Dall'analisi dell'andamento della curva di intensità si deriva la segmentazione del segnale in sillabe proposta in questo progetto. La nostra analisi trae spunto dall'ipotesi che esiste una relazione diretta tra il parametro fonologico *sonorità* con la grandezza fisica *energia*. Anche se, come si vedrà nel capitolo successivo, questa relazione viene meno con una certa regolarità.

Secondo il principio della sonorità occorrerebbe trovare i massimi ed i minimi della curva di intensità per individuare le sillabe. La difficoltà consiste, ancora una volta, nella determinazione della finestra ideale per il calcolo della curva di energia, poiché finestre troppo piccole genererebbero dei segmenti che dal punto di vista fonologico non potrebbero

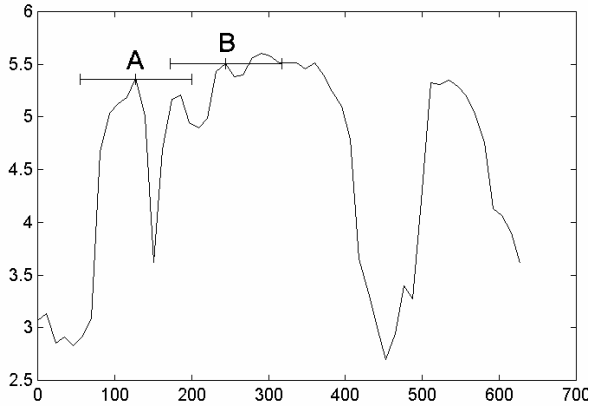

Figura 3.2 Esempio della prima selezione dei massimi. Quello indicato dalla lettera 'A' viene considerato nucleo vocalico poiché è un massimo assoluto all'interno della sua finestra, diversamente da quello indicato dalla 'B'

corrispondere ad una sillaba, mentre finestre troppo grandi accorperebbero segmenti previsti come distinti dalla fonologia.

Occorre stabilire, quindi, un metodo per individuare i massimi dell'intensità che effettivamente coincidono con i nuclei sillabici. Il nostro algoritmo, dunque prevede come primo passo l'individuazione di tutti i massimi relativi presenti nella curva di intensità, successivamente si considera per ogni massimo, una finestra, che chiameremo "finestra di segmentazione", con centro nel massimo in esame e un raggio di lunghezza ***passo***; i massimi che risultano assoluti all'interno di tale finestra vengono ritenuti nuclei vocalici. Anche se l'algoritmo e da considerarsi ancora sottospecificato già in questo modo si riescono ad ottenere buoni risultati.

Assumendo come confini tra i segmenti i minimi della curva di intensità compresi tra due massimi selezionati, gli errori che si verificano sono i seguenti:

1. Accorpamento di due sillabe in un solo segmento;
2. Consonanti fricative che formano segmento a sé;
3. Divisione in due segmenti di un'unica sillaba in corrispondenza di vocali toniche;
4. Confini non coincidenti con quelli previsti dalle regole fonologiche;

Per questo motivo è stato necessario inserire altri quattro stadi di elaborazione che riducono la presenza di questi errori.

II stadio: split. I primo tipo di errore viene affrontato andando a considerare i massimi ed i minimi scartati dalla prima analisi. Si considera un segmento alla volta e per ogni minimo relativo dell'intensità all'interno di esso si cerca di individuare un

ulteriore confine di sillaba. L'idea è di verificare se c'è una variazione di intensità considerevole nelle prossimità dei minimi scartati in precedenza. Per questo scopo si prendono due finestre contigue di lunghezza ***passosplit*** con punto di intersezione nel minimo esaminato. Si considera il valore del più piccolo dei massimi in queste due finestre, per esempio in figura 3.3 si sceglierebbe il valore di B, il rapporto tra questo valore e quello del minimo C deve essere superiore ad una certa soglia. Per facilitare la suddivisione di segmenti molto lunghi, per i quali la probabilità di contenere due sillabe è più alta, si è preferito usare una soglia più alta per i segmenti corti, ed una soglia più piccola per i segmenti lunghi; per questo motivo di considera la lunghezza del più piccolo dei segmenti ottenuti dalla

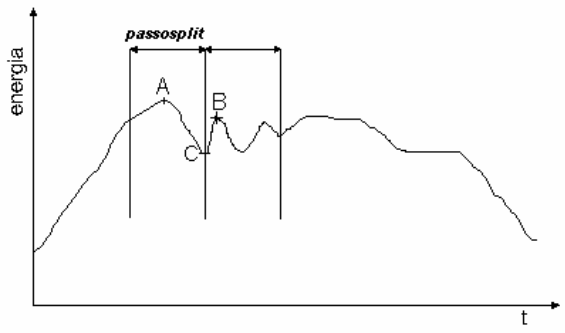

Figura 3.3 Split. La figura mostra la curva di intensità di un singolo segmento. Esso viene diviso in corrispondenza del minimo, indicato con C, se in prossimità del minimo stesso ci sono grosse variazioni dell'intensità.

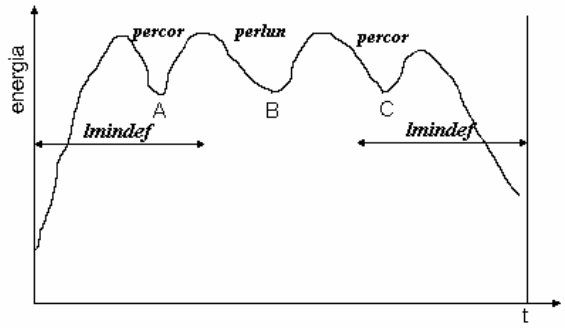

*Figura 3.4. I punti A B e C sono candidati come ulteriori confini di sillabe. La soglia sarà pari a **percor** se la durata del più corto dei due segmenti che verrebbero prodottiè inferiore a **lmindef**, A e C; altrimenti, se entrambi superano **lmindef**, sarà pari a **perlun**. In questo modo viene favorita la segmentazione in corrispondenza di C poiché **perlun** è inferiore a **percor**.*

suddivisione: se essa è più grande del valore di ***lmindef*** allora la soglia di riferimento scelta sarà ***perlun*** altrimenti si sceglierà il valore corrispondente a ***percor*** (maggiore di ***perlun***).

III stadio: accorpamento delle fricative. Il problema delle fricative che formano segmenti a sé viene affrontato considerando l'intensità del segnale ottenuto filtrando con filtro passa basso il segnale originario; per questo scopo viene usato un filtro, con frequenza di taglio di 1100Hz. In seguito si indicherà come **intensità residua** l'energia del segnale così filtrato, mentre l'**intensità totale** sarà l'intensità del segnale originale.

Un segmento si ritiene, quindi, costituito da una sola fricativa se calcolando in ogni suo punto la differenza in percentuale tra l'intensità totale e quella residua, mediamente essa risulta essere superiore al valore di ***fricmed*** e se, in coincidenza del massimo dell'intensità totale, ci sia almeno un rapporto tra intensità totale e residua superiore al valore contenuto in ***fricmax***. Una volta individuato un segmento costituito solo da una

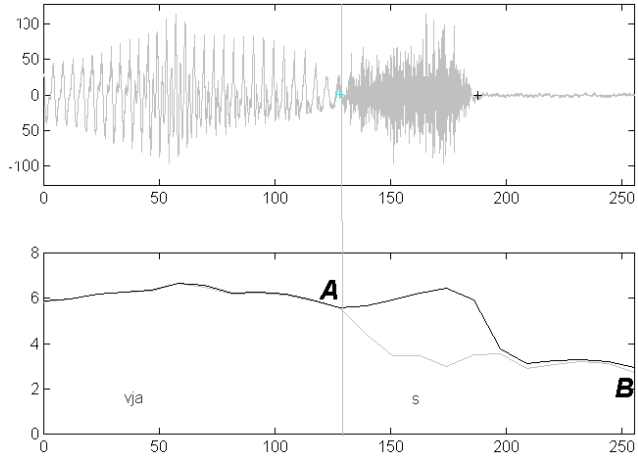

Figura 3.5 Accorpamento di una fricativa al segmento che la precede. In grigio l'intensità residua. La figura mostra due segmenti, l'andamento discendente dell'intensità residua del secondo viene individuato dal massimo assoluto nel primo punto (A) e dal rapporto tra l'intensità in B e in A.

fricativa, se il massimo dell'intensità residua si trova nel primo punto del segmento e, se tra l'ultimo ed il primo punto c'è un rapporto di ampiezze inferiore al valore di ***assfrsx*** allora il segmento viene messo in coda a quello che lo precede. In pratica se l'intensità residua ha un andamento decrescente si considera la fricativa come coda della sillaba precedente ritenendo che sia seguita da una pausa o da un fono occlusivo. Nel caso contrario essa viene assegnata all'attacco del segmento successivo.

IV Stadio: ricompattazione toniche. Il terzo tipo di errori è più difficile da trattare, ed è dovuto ai casi in cui risulti impossibile assicurare un flusso continuo e costante dell'aria attraverso le pliche vocali durante la produzione di una vocale sostenuta per un tempo molto alto (>300 ms). Fortunatamente, in questi casi la variazione di intensità è spesso piuttosto contenuta: il quarto stadio, dunque, esamina la differenza di intensità tra ogni minimo e i due massimi dei segmenti che separa, procedendo in questo caso all'unione dei due segmenti se il rapporto tra il più piccolo dei due massimi ed il minimo soddisfa almeno una delle seguenti condizioni:

- è inferiore a ***ricomass***
- è inferiore a ***ricomlun*** e, allo stesso tempo, la durata del segmento ottenuto è inferiore a ***lunglun***
- è inferiore a ***ricommed*** e la durata del segmento ottenuto é inferiore a ***lungmed***.

V Stadio: spostamento margini delle fricative. L'ultimo tipo di errori commessi dal primo stadio la non perfetta collocazione del confine di segmento, con la

conseguenza che un fono risulta trovarsi per metà in un segmento e per l'altra metà nel successivo. Questo problema riguarda principalmente le consonanti fricative e almeno per esse è stato trovato un modo di porre rimedio. Secondo le regole fonologiche le fricative vanno sempre all'inizio della sillaba a meno che non siano seguite da una consonante occlusiva. In questo stadio, quindi, si spostano tutti i confini di segmento verso sinistra se il rapporto tra l'intensità residua e quella totale non è inferiore a ***margfri***. La nuova posizione del confine verrà messa nel primo punto in cui si verifica questa condizione. Lo spostamento del confine deve essere effettuato solo se l'intensità del segnale è maggiore di ***occlus***, in modo da impedire lo spostamento in presenza di un'occlusiva.

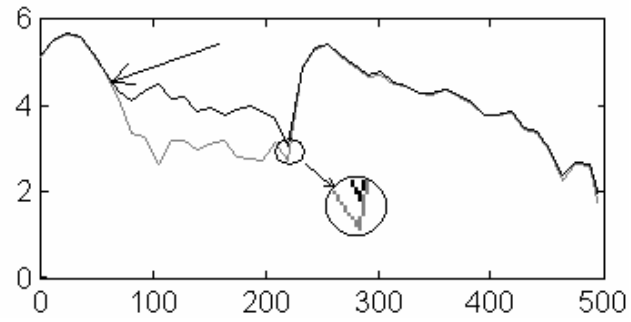

Figura 3.6 Spostamento margini delle fricative. La differenza tra l'intensità assoluta e residua nel punto di minimo, vedi parte evidenziata, permette di spostare il margine fino nella sua corretta posizione indicata in fiaura dalla freccia

Ulteriori analisi: eliminazione silenzi. I rumori di fondo e la respirazione del parlante possono generare dei massimi della curva di intensità durante le pause tali di indurre il sistema a trovare dei segmenti composti solo da silenzio: per questo motivo è stato opportuno introdurre un ulteriore stadio di elaborazione che accorpa i segmenti superflui ad uno dei segmenti adiacenti. Un segmento viene considerato composto solo da silenzio se è verificata almeno una delle seguenti condizioni:

- il numero di punti della curva di intensità che superano il valore di ***minonsil*** non è inferiore a ***leminsil*** ed il massimo nel segmento non supera ***minmaxsi***;
- il massimo dell'intensità del segmento non arriva ***minmaxs2*** e la lunghezza del segmento è minore di ***lmaxsile***.

Se una di queste due condizioni si verifica i segmento viene considerato composto solo da silenzio.

Riassumendo il segmentatore può essere rappresentato con lo schema a blocchi dei figura 3.7

Segmentazione preliminare	
Passo	raggio finestra di analisi intensità
Split	
Passosplit	rapporto (raggio per split)/raggio
Lmindef	discrimine tra lunghe e corte
Percor	soglia per split corte
Perlun	soglia per split lunghe
Assimilazione delle fricative	
Fricmax	minima differenza tra intensità ed intensità residua nel massimo
Fricmed	minima differenza media tra intensità ed intensità residua
Assfrsx	Minimo rapporto richiesto tra intensità filtrata agli estremi del segmento per assimilare il segmento al successivo
Ricompattazione toniche	
Ricomass	Ricompattazione indipendente da lunghezza
Ricommed	Ricompattazione per lunghezza media
Lungmed	Lunghezza media
Ricomlun	Ricompattazione per lunghezza lunga
Lunglun	Lunghezza lunga
Spostamento margini	
Margfri	Rapporto tra intensità filtrata e non per distinguere le fricative
Occlus	Rivela occlusione
Elimina silenzi	
Minonsil	Minimo valore dell'intensità per non essere in presenza di silenzio
Leminsil	Lunghezza minima sillaba effettiva
Minmaxsi	Il massimo dell'intensità non deve superare questo valore
Mimmaxs2	Limite superiore per il massimo di segmenti di silenzio corti
Lmaxsile	Limite lunghezza segmenti corti

Tabella 1 Tabella riassuntiva dei parametri usati

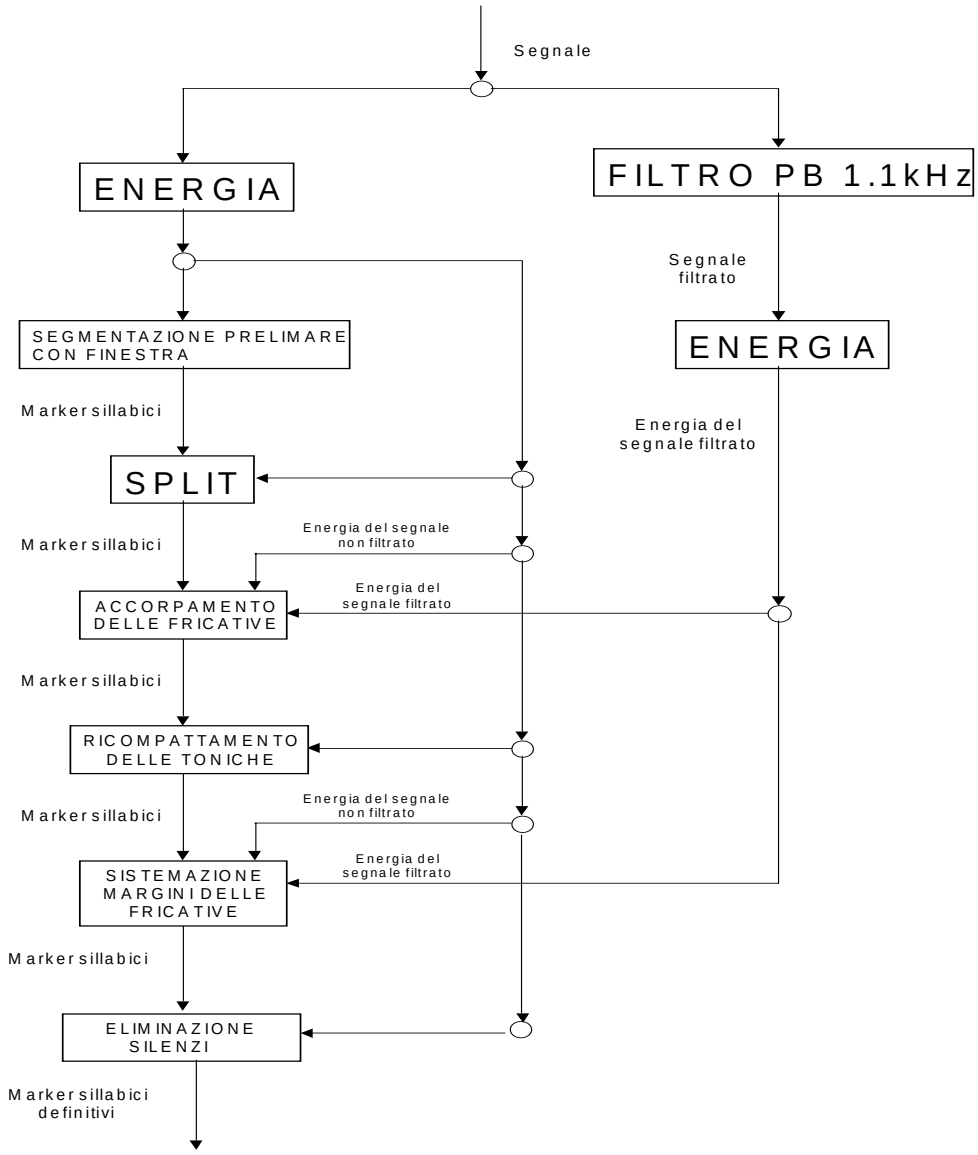

Figura 3.7 Schema a blocchi del segmentatore. I campioni del segnale vengono usati per calcolare l'energia del segnale filtrato e del segnale non filtrato. L'energia del segnale completo viene usata per una segmentazione preliminare con il metodo delle finestre, per separare sillabe contenute nello stesso segmento e ricompattae quelle divise. L'energia del segnale filtrato viene usata per trattare le fricative.

3.4 Determinazione dei parametri

Il numero di parametri usato è nettamente inferiore a quelli necessari in sistemi con lo stesso scopo utilizzanti le reti neurali. Tuttavia non è semplice determinare i valori dei parametri impiegati. Si è deciso, quindi, di usare un approccio, per alcuni aspetti, simile a quello necessario per le reti neurali. E stato quindi definito un corpus di parlato da usare per una procedura di apprendimento. Per questo scopo è stato usato

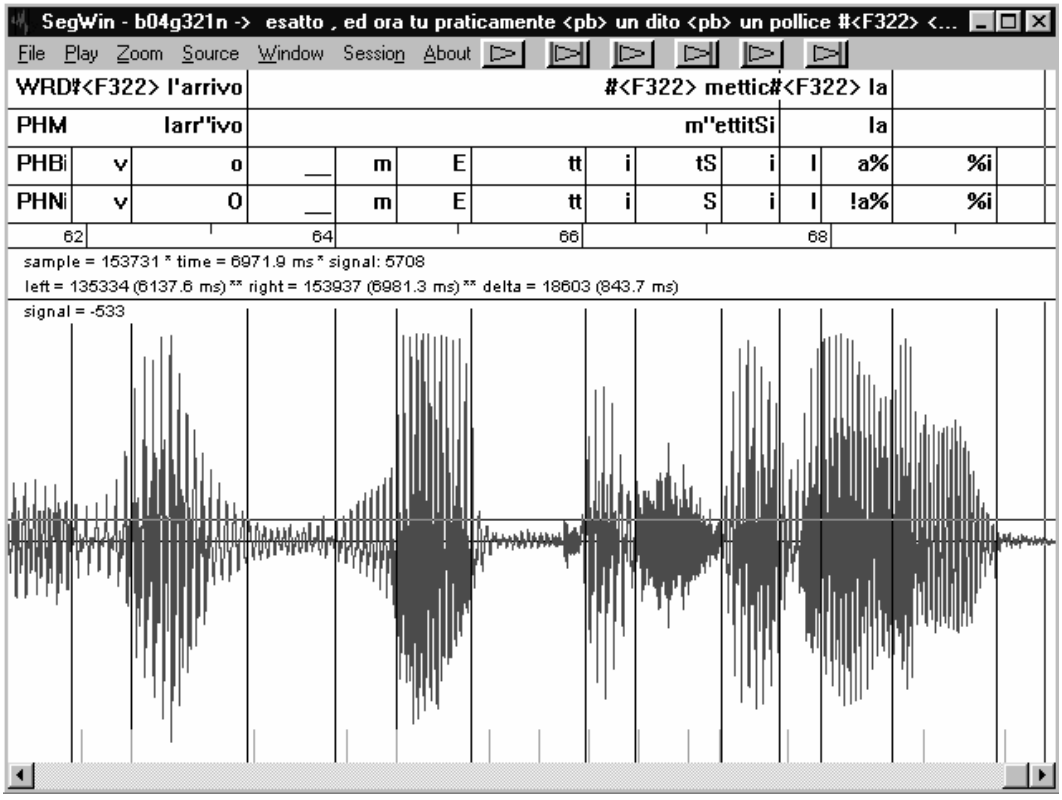

Figura 3.8 Il corpus di AVIP viene segmentato ed etichettato a livello fonetico e lessicale, la segmentazione in sillabe, necessaria come riferimento, viene ottenuta applicando le regole fonologiche alla stringa di simboli fonetici che accompagna ogni segnale.

un set di 35 turni di dialogo spontaneo estratto tra quelli disponibili nell'ambito di un progetto nazionale di ricerca mirante alla costituzione di un corpus di parlato

segmentato manualmente (Archivio Varietà Italiano Parlato, in seguito AVIP). Nell'ambito di questo progetto sono dunque disponibili segnali vocali campionati ed annotati a vari livelli (foni, fonema, parola); in figura 3.8 è mostrato il tipo di informazioni ricavabili da AVIP, come si può notare il segnale è segmentato ed etichettato a vari livelli ma non è diviso in sillabe. Il materiale a disposizione era costituito, oltre che dai segnali, da tabelle, come la tabella 2, che riportavano per ogni fono la sua trascrizione e l'ubicazione nel segnale. Sarebbe stata molto comoda una tabella simile con le sillabe al posto dei foni. Per questo motivo è stata scritta una procedura che divide automaticamente in sillabe la stringa ricavata dalla terza colonna delle tabelle, e da questa ricava, tramite il ricollegamento con i markers temporali contenuti nelle tabelle, la segmentazione di riferimento sul segnale.

0	1460	__
1461	34244	%__
34245	36124	p
36125	36515	4
36516	38153	o
38154	39362	s
39363	41990	E
41991	43132	g_f
43133	43705	%w
43706	44715	%e
44716	45346	E
45347	47595	a
47596	48476	%r
48477	50459	i
50460	51282	v\
51283	52891	a
52892	55397	f
55398	57605	i
57606	58952	n
58953	63927	O~

Tabella 2. L'etichettatura fonetica del segnale è costituita da tabelle che riportano; inizio, fine ed etichetta del fono

Una funzione di valutazione confronta la segmentazione del segnale con quella ricavata dalla segmentazione di riferimento ottenuta dalla segmentazione della stringa di simboli.

La valutazione dell'adeguatezza dei parametri ottenuti deve quindi essere fatta con un secondo corpus dello stesso tipo di quello usato per la fase di apprendimento, che denomineremo corpus di test.

Un terzo corpus, di natura differente dai primi due impiegati viene considerato per il terzo stadi delle procedure di valutazione come corpus di verifica.

3.4.1 Segmentazione della stringa fonetica

La segmentazione della stringa deve essere effettuato con criteri quanto più simili alle regole fonologiche. Essendo universalmente accettato il principio della scala di sonorità (di cui si è parlato nel capitolo 2), è necessario far corrispondere ad ogni simbolo della stringa il corrispondente valore della sonorità secondo i valori della tabella 3. Il valore numerico del singolo elemento è completamente arbitrario, quello che invece ricalca le idee consolidate della fonologia è l'ordine con cui vengono assegnati i valori ai simboli. Non conta, quindi, la differenza di sonorità tra i vari foni, ma solo il fatto che uno abbia una sonorità superiore, inferiore o uguale a quella di un altro fono. I nuclei delle sillabe corrispondono quindi ai massimi della sequenza numerica associata alla stringa. Il problema è, invece, la determinazione dei confini: a

Scala di sonorità

{	24	n	11
a	24	N	11
A	24	v	9
E	23	z	9
O	23	Z	9
6	22	S	8
@	21	f	7
e	21	h\	7
o	21	H\	7
i	19	s	7
u	19	S	7
j	18	dZ	6
L	18	dz	6
v\	18	ts	4
w	18	tS	4
B	16	?	3
D	16	b	3
G	16	d	3
4	14	g	3
l	14	k	1
r	14	p	1
J	11	t	1
m	11	__	0
M	11	+	0

Tabella 3.3 I foni sono indicati secondo la codifica SAMPA descritta in appendice.

differenza di quando avviene nella segmentazione del segnale, in questo caso ogni valore della sonorità corrisponde ad un intero fono, se questo è un minino sorge il problema di assegnarlo alla sillaba precedente o a quella di coda. Su questo aspetto le regolo fonologiche possono differire a seconda della teoria considerata. Il solo principio di sonorità non basta, in quanto qualunque si la posizione che si assegna al fono, il principio non sarà violato. E' necessario, quindi, introdurre delle regole di combinazione dei suoni all'interno della lingua considerata, dette *regole fonotattiche*, le quali variano da lingua a lingua. Nel caso dell'italiano si ha che le consonanti possono essere coda di sillaba solo se geminate o se sono delle sonoranti seguite da una consonate [Nespor, 1993]. Una convenzione che viene qui usata è quella di non dividere le geminate, questa scelta viene fatta per due motivi:

-la suddivisione delle occlusive non può avvenire in nessun modo al livello della segmentazione del segnale;

-anche senza implementare questa convenzione, le occlusive, fricative ed affricate non verrebbero, comunque, suddivise.

Per uniformità con questi casi si è deciso di lasciarle sempre insieme.

La pseudocodifica della segmentazione in sillabe dei simboli è la seguente.

```
Assegna ad ogni fono il corrispondente valore di sonorità;
Minimi:=nessuno;
Per ogni fono,
   se (   sonorità(fono)<sonorità(fono precedente)
       &
          sonorità(fono)<=sonorità(fono successivo)
      )
```

```
       allora
         aggiungi fono a minimi;
   fine se;
fine per;
per ogni minimo
     se (   sonorità(minimo)>=10  %sonorante
         &
            sonorità(fono successivo)<18 %consonante
         &
            sonorità(minimo)<>sonorità(fono successivo) %geminate
        )
         allora
            minimo va in coda alla sillaba precedente;
         altrimenti
            minimo va in testa alla sillaba successiva;
      fine se
fine per
```

3.4.2 Valutazione automatica della segmentazione

La segmentazione simbolica descritta precedentemente fornisce dei marker sul segnale da considerare come riferimento. Il sistema trova invece marker che possono essere diversi: per considerarli corretti, i marker prodotti dal sistema di segmentazione del segnale devono trovarsi tra i due massimi dell'intensità dei segmenti teorici corrispondenti. In altri termini si prendono due sillabe adiacenti, e si trovano i rispettivi massimi dell'intensità, se tra questi massimi non è compreso nessun confine di segmento allora le due sillabe sono state incluse nello stesso segmento, se invece si trova più di un confine allora una delle due sillabe è stata spezzata in due. La funzione di valutazione conta il numero di errori totali che vengono commessi

La pseudocodifica è la seguente:

```
trovati:=0;
troppi:=0;
```

```
per ogni coppia di sillabe adiacenti (a,b)
  ma:=posizione del massimo di intensità di a;
  mb:=posizione del massimo di intensità di b;
  cisono:=confini di segmento compresi tra ma e mb;
  se cisono>0,
        trovati:=trovati+1;
        se cisono>1,
           troppi:=troppi+1;
        fine se
  fine se
fine per;
persi:=numero(confini di sillabe)-2-trovati;
errori:=persi+troppi;
```

Il valore di questa funzione è solo una prima approssimazione delle reali prestazioni del sistema, ma comunque è sufficiente allo scopo di determinare il miglior insieme di parametri.

La differenza tra la funzione di valutazione e il numero reale di errori commessi dal sistema è dovuta essenzialmente a due fattori:

- fenomeni vocali non verbali (sospiri, risate, inspirazioni ecc.), disfluenze (parole interrotte, esitazioni, ecc.) e i rumori accidentali sono stati ignorati nella segmentazione da parte dei fonetisti, ma comunque vengono trattati dal sistema;
- vengono considerati errori solo la suddivisione di una sillaba teorica in due segmenti e la fusione di due sillabe in un solo segmento; non viene valutato quindi la precisione con cui vengono stabiliti i confini sillabici.

Si fa l'ipotesi che i problemi del primo tipo siano poco frequenti e si presentino un numero di volte simile per ogni set di parametri, in modo da essere ininfluenti nella valutazione del set di parametri.

3.4.3 Strategie di ricerca del migliore set di parametri

Una prima valutazione dei parametri è sta fatta in maniera grossolana procedendo per tentativi mirati e con alcune misure informali sulla curva di intensità. Il set ti parametri così determinato è riportato in tabella 4.

Benché non sia questo un metodo scientificamente corretto l'accuratezza della segmentazione prodotta con questi parametri è tutt'altro che insoddisfacente[1].

Parametri iniziali

Segmentazione preliminare		
passo	5	11*ms
split		
passosplit	*0.8*	
lmindef	8	11*ms
percor	1.07	
perlun	1.04	
assimilazione delle fricative		
Fricmax	1.05	
Fricmed	1.3	
Assfrsx	0.7	
Ricompattazione toniche		
Ricomass	1.02	
Ricommed	1.06	
Lungmed	19	11*ms
Ricomlun	1.035	
Lunglun	21	11*ms
Spostamento margini		
Margfri	1.08	
Occlus	3	Logaritmo dell'energia
Elimina silenzi		
Minonsil	3.9	Logaritmo dell'energia
Leminsil	3	11*ms
Minmaxsi	5.5	Logaritmo dell'energia
Mimmaxs2	4.6	Logaritmo dell'energia
Lmaxsile	6	11*ms

Tabella 4: parametri determinati manualmente, la terza colonna indica l'unità di misura del parametro.

[1] Nel capitolo successivo verranno mostrati i risultati che si otterrebbero usando il set di parametri iniziale sul corpus di apprendimento.

Il metodo ideale per determinare il miglior set di parametri sarebbe quello di fissare per ognuno di essi un set di valori plausibili e quindi di esaminare in maniera esaustiva tutte le combinazioni possibili. Ma usando solo tre possibili valori per ogni parametro si ottengono 3^{20}=3.486.784.401 possibili set.

Sono stati usati, quindi, tre diversi metodi per restringere il numero di prove da effettuare.

Un primo metodo, detto "semiesaustivo", ricalca, in parte, l'euristica usata nella stesura dell'algoritmo di segmentazione e nella determinazione dei parametri iniziali. Il principio base è quello di ottimizzare uno alla volta i vari stadi che compongono il sistema disabilitando quelli successivi.

Per ognuno dei parametri vengono stabiliti un insieme di valori possibili, quelli nella colonna "valori" della tabella 5, quindi si procede abilitando uno alla volta i moduli e provando i modo esaustivo tutte le

Parametro	min	max	step	valori	no
passo	3	7	1	3-4-5-6-7	
passosplit	0,5	1,5	0,15	0,5-0,8-1,2-1,4	
lmindef	5	12	1	5-8-10-13	
percor	1,02	1,1	0,015	1,05-1,07-1,09	1000
perlun	1,02	1,1	0,015	1,02-1,04-1,05	1000
fricmax	1,02	1,08	0,01	1,03-1,05-1,07	1000
fricmed	1,1	1,5	0,03	1,2-1,3-1,4	1000
assfrsx	0,7	0,7	0,03	0,7	
ricomass	1,01	1,04	0,005	1,01-1,02-1,04	1
ricommed	1,03	1,12	0,01	1,03-1,06-1,09	1
lungmed	15	25	2	16-19-22	
ricomlun	1,02	1,07	0,005	1,02-1,035-1,05	1
lunglun	15	25	2	18-21-24	
margfri	1,08	1,08	0,01	1,08	
occlus	3	3	0,1	3	
minonsil	3	5	0,3	3,5-3,9-4,3	0
leminsil	2	5	1	2-3-4	
minmaxsi	4	6	0,3	5-5,5-6	0
mimmaxs2	3	6	0,3	4,2-4,6-5	0
lmaxsile	4	8	1	5-6-7	

Tabella 5 Dati per la determinazione dei parametri

possibili configurazione del set di parametri associati al modulo stesso; vengono tenuti fissati quelli relativi ai moduli precedenti ai valori ottimali trovati, mentre i moduli successivi vengono disabilitati assegnando ai relativi parametri i valori nella colonna “no” in tabella 5. In questo modo vengono ottimizzati in sequenza i vari moduli.

Come si può facilmente calcolare è necessario provare 644 configurazioni[1] degli assegnamenti dei valori dei parametri contro un totale di 382.637.520 previsti dai valori della colonna “valori” della tabella 5.

In questo modo si esclude la possibilità che una impostazione non ottimale di due stadi, visti singolarmente, dia complessivamente un risultato migliore. Ciò è possibile poiché gli errori provocati da uno stadio potrebbero essere corretti da uno stadio successivo. In effetti questo è molto probabile: per esempio, lo stadio di split potrebbe suddividere più segmenti del dovuto, mentre lo stadio di ricompattazione potrebbe riunire i segmenti indebitamente divisi fino ad ottenere un comportamento migliore che con due stadi ottimizzati separatamente.

Per verificare questa ipotesi, il secondo metodo provato, detto “casuale”, consiste nel definire per ogni parametro un intervallo di variabilità e provare in maniera totalmente casuale un numero significativo di possibili configurazioni. Nelle colonne “min” e “max” della tabella 5 sono mostrati gli intervalli usati nelle nostre prove.

---

[1] Pari a 5+4*4*3*3+3*3*1+3*3*3*3*3+3*3*3*3*3, i segni di addizione sono dovuti al fatto che i vari stadi vengono ottimizzati in sequenza e non tutti insieme

Il terzo metodo, detto “semicasuale”, consiste nel far variare di poco i parametri iniziali fino a trovare una configurazione migliore di quella originaria. Questo aggiustamento viene ripetuto finché non viene trovata una configurazione che sia ancora migliore. La colonna “*step”* nella tabella 5 mostra le massime variazioni, rispetto alla migliore configurazione precedente, che possono assumere i singoli parametri in ogni tentativo. Nella tabella 4, per alcuni parametri è stato indicato un solo valore, questo avviene per quei parametri che servono a spostare i margini dei segmenti. Poiché questa operazione non influenza la funzione di valutazione, è inutile permettere una loro variazione. Per questo motivo essi sono stati determinati tramite opportune misure effettuate su alcuni segnali presi come campione.

3.4.4 Parametri trovati

La ricerca casuale ha riportato i risultati in tabella 6. Vengono riportati i cinque migliori risultati. Come si può osservare si ottengono risultati molto simili anche per valori molto diversi dei parametri, questo è possibile perché i vari stadi di elaborazione sono in grado di ovviare agli errori degli altri, questo lavoro di compensazione permette quindi anche con parametri piuttosto arbitrari, come quelli usati come punto di partenza, di avere una risposta del sistema soddisfacente.

Il metodo semicasuale è quello che riesce a dare i migliori risultati scendendo ad un punteggio di 113. Il metodo semiesaustivo, invece, riesce solo ad eguagliare i risultati del metodo casuale, ma ci sono ben 26 set di parametri per lo stesso risultato. La tabella 5 mostra, nelle ultime due colonne, i valori dei parametri ricavati con il metodo parzialmente casuale e di set scelto a caso tra quelli ricavati con l'ultimo.

Nome parametro	**casuale**					**semicasuale**	**semiesaustivo**
Passo	5	4	5	**4**	5	**5**	**5**
passosplit	1,00	1,14	1,11	**1,41**	1,26	**1,13**	**1,40**
lmindef	9	17	11	**8**	11	**8**	**8**
perlun	1,073	1,060	1,040	**1,067**	1,025	**1,039**	**1,050**
percor	1,071	1,061	1,023	**1,032**	1,051	**1,083**	**1,090**
fricmax	1,044	1,034	1,024	**1,026**	1,034	**1,044**	**1,030**
fricmed	1,29	1,39	1,45	**1,32**	1,33	**1,33**	**1,30**
assfrsx	0,7	0,7	0,7	**0,7**	0,7	**0,7**	**0,7**
ricomass	1,027	1,034	1,027	**1,026**	1,027	**1,020**	**1,020**
ricommed	1,069	1,044	1,063	**1,039**	1,106	**1,064**	**1,090**
lungmed	17	17	18	**20**	15	**13**	**16**
ricomlun	1,036	1,052	1,023	**1,051**	1,050	**1,041**	**1,020**
lunglun	17	17	24	**17**	17	**19**	**18**
margfri	1,08	1,08	1,08	**1,08**	1,08	**1,08**	**1,08**
occlus	3	3	3	**3**	3	**3**	**3**
minonsil	4,22	3,78	3,97	**4,59**	3,84	**4,58**	**3,90**
leminsil	5	3	3	**5**	3	**1**	**2**
minmaxsi	4,94	5,82	5,94	**4,60**	5,44	**6,44**	**6,00**
mimmaxs2	5,89	4,07	3,68	**4,95**	4,05	**3,84**	**4,60**
lmaxsile	8	5	8	**6**	4	**4**	**7**
risultati	**117**	**116**	**117**	**115**	**117**	**113**	**115**

Tabella 5

Capitolo 4

Analisi dei risultati

4.1 Descrizione dei corpora

Nel capitolo precedente abbiamo determinato il set di parametri che minimizza le differenze tra la segmentazione del segnale e quella di riferimento. Una volta deciso i valori da assegnare ai parametri è stato opportuno verificare manualmente il comportamento del sistema su tre diversi corpora: addestramento, test e verifica.

Il primo corpus è lo stesso usato per la determinazione dei parametri, in questo caso l'analisi manuale valuta esattamente il numero degli errori che vengono commessi. Il corpus di test serve a verificare che i parametri trovati permettano di avere dei buoni risultati anche con segnali diversi da quelli usati per l'addestramento. Il secondo corpus viene quindi preso in modo che sia molto simile al primo sia come stile di parlato che come condizioni di registrazione.

Il corpus di verifica serve a controllare che la procedura realizzata sia estensibile ad altri contesti, motivo per cui esso é stilisticamente diverso dai precedenti.

Tutti i corpora sono stati registrati in condizioni di rumore ottimale. Quelli di addestramento e verifica sono tratti tra quelli disponibili nell'ambito del progetto

AVIP. Lo stile di parlato si può indicare come semispontaneo, in quanto i parlanti sono a conoscenza del fatto che i loro dialoghi vengono registrati, ma messi davanti ad un'altro compito che assorbe tutta la loro attenzione dovrebbero esprimersi in maniera del tutto naturale. Il corpus di addestramento è composto da 35 turni di dialogo di quattro parlanti di ambo i sessi, mentre il corpus di test è composto da 25 turni di due parlanti di sesso differente.

Il terzo corpus viene definito per simulare l'impiego reale dell'applicazione con dei segnali molto diversi da quelli usati per il suo sviluppo. A questo scopo vengono usate quattro registrazioni di un minuto dall'inizio di altrettanti telegiornali regionali (Campania, Lazio, Lombardia, Toscana). Il parlato presente in queste registrazioni è da considerarsi molto controllato ma non al punto da considerarsi parlato di laboratorio. Questo corpus è stato precedentemente segmentato in sillabe da un esperto fonetista, motivo per cui si presta molto bene ai nostri scopi. Per i primi due corpora, invece, si ricorre alla suddivisione prodotta dalla procedura, descritta nel capitolo precedente, di segmentazione della stringa di simboli fonetici associata ai segnali, corretta manualmente in alcune parti dall'autore.

4.2 Classificazione dei possibili errori

La classificazione degli errori viene poi fatta secondo due schemi. Il primo schema ricalca il tipo di classificazione degli errori in uso negli ambiti di ASR, in esso si distinguono gli **errori di numero,** che variano il numero di segmenti rispetto al

numero di sillabe previste dai fonetisti, dagli **errori di margine,** quando l'errore è dovuto alla sbagliata attribuzione di un fono o di parte di esso. La distinzione tra questi è stata fatta perché gli errori di numero permettono di confrontare il nostro sistema con quelli realizzati da altri autori, in particolare [Pfitzinger *et alii* 1996] e da [Reichl & Ruske 1993].

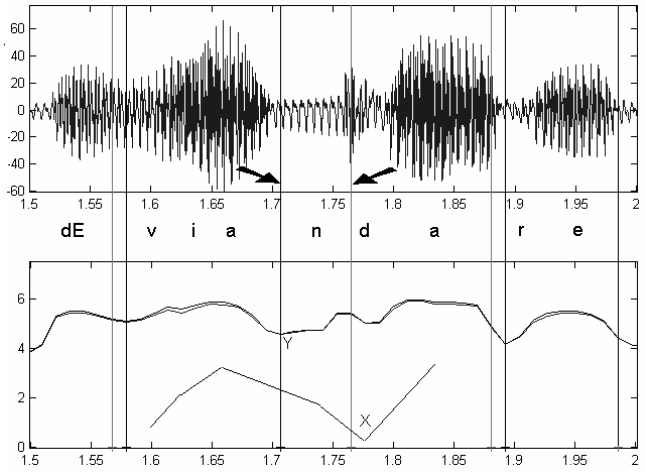

Figura 4.1 Esempio di errore di margine che risulta essere un errore di sonorità in senso stretto. Le linee verticali grigie rappresentano la segmentazione in sillabe prevista, mentre quelle nere rappresentano la segmentazione prodotta dal sistema. La spezzata sovrapposta al secondo grafico rappresenta l'andamento della sonorità (come dalla tabella 3.3); come si può notare il minimo di sonorità è presente sull'occlusiva (indicata con X sulla spezzata della sonorità), non sulla nasale, mentre il minimo di energia si trova su quest'ultima (nel punto indicato con Y sulla curva di energia).

L'altro schema di classificazione verifica l'aderenza dei segmenti prodotti con il principio teorico della scala di sonorità. In base a questo principio una sillaba deve essere composta da una sequenza di foni con sonorità prima crescente e poi decrescente. Quando il segmentatore produce un segmento composto da foni la cui sonorità oscilla più di una volta (per esempio nel segmento *come* la sonorità presenta due massimi in corrispondenza delle vocali) si ha un **errore di sonorità in senso stretto**. Se invece una sillaba viene divisa in più parti allora si ha un **errore di sonorità in senso lato**, poiché all'interno dei segmenti

prodotti non c'è alcun minimo di sonorità tra due massimi, ma è quindi viene violata la richiesta che i confini di sillaba debbano trovarsi in corrispondenza dei minimi di sonorità. Esempi di questi errori sono le sonoranti isolate, sillabe come *dro* divisa ulteriormente in *dr*|*o* .

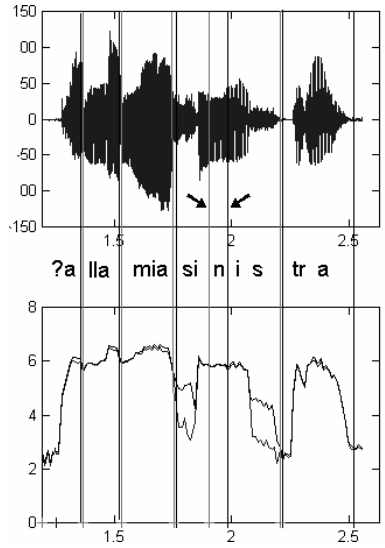

Figura 4.2 Errore di margine che risulta essere anche errore di confine

Quando un minimo di sonorità viene messo in una sillaba diversa da quella prevista dalle regole fonotattiche non si ha alcun tipo di violazione del principio di sonorità si ha quindi solo un **errore di confine**, ad esempio *cas*|*a*.

Gli errori definiti come "di margine" del primo schema non coincidono sempre con gli "errori di confine" del secondo schema poiché tra essi trovano posto anche alcuni

Errore	Descrizione	Esempio
Numero	Sillabe divise in due o accorpate	*Si*|*m*|*bo*|*lo, re*|*gale*
Margine	Confini non corrispondenti	*Sin*|*is*|*tra, do*|*rso*
Sonorità in senso stretto	Segmenti con più di un massimo di sonorità	*Con*|*sonan*|*ti, do*|*rso*
Sonorità in senso lato	Nessun minimo di sonorità tra due segmenti	*Se*|*n*|*so*
Confine	Il principio di sonorità viene rispettato	*Sin*|*is*|*tra,* controesempio *do*|*rso*

Tabella4.1 Schema riassuntivo delle due classificazioni degli errori

errori di sonorità. Ad esempio se la parola *dorso* viena suddivisa in *do*|*rso* si ottiene sia un errore di margine ma non un errore di confine, poiché secondo il secondo schema si ottiene un errore di sonorità in senso stretto in quanto la seconda *sillaba* risulta composta da un fono, "*s*", minimo di sonorità compreso tra due foni, "*r*" e "*o*", con sonorità più alta. Un esempio è riportato in figura 4.1.

4.3 Risultati

I risultati dati dalla funzione di valutazione non possono essere utilizzati per stimare l'accuratezza con cui la procedura segmenta in sillabe il segnale, motivo per cui si è reso necessario esaminare manualmente le segmentazioni prodotte, usando, per ognuno dei metodi per stabilire quale sia effettivamente il miglior set di parametri, oltre che per avere una ulteriore conferma della bontà della funzione. Il risultato è stato conforme a quello della funzione di valutazione, in quanto il set ricavato dal metodo semicasuale è risultato migliore degli altri. Come si può vedere, quindi, dalla tabella 4.2 sul corpus di apprendimento

	semicasuale		casuale		semiesaustivo		di partenza	
errori totali	92	14%	99	15%	99	15%	101	15%
errori di margine	26	4%	26	4%	25	4%	28	4%
errori di numero	66	10%	73	11%	74	11%	73	11%
son stretta	44	7%	45	7%	48	7%	40	6%
son lato	24	4%	32	5%	28	4%	34	5%
confine	24	4%	22	3%	23	3%	27	4%

Tabella 4.2 Risultati sul corpus di addestramento. Il corpus di addestramento è composto da 667 sillabe. Vengono riportati anche i risultati che si otterrebbero con il set di parametri di partenza: i risultati non sono molto diversi.

la procedura è riuscita ad avere una percentuale di errori del 14%. Da notare che anche con il set di parametri di partenza i risultati non sono stati molto peggiori, ciò è dovuto alla compensazione dei vari stadi operano tra di loro, cosa che permette di avere dei buoni risultati anche con parametri piuttosto arbitrari.

Errori	88	14%
errori di margine	18	3%
errori di numero	70	11%
son stretta	40	6%
son lato	34	5%
Confine	14	2%

Tabella 4.3 Risultati sul corpus di test. Il totale delle sillabe è 645

Come si può vedere in tabella 4.3, nel corpus di test i risultati sono stati confermati essendo rimasta invariata la percentuale di errori.

Le sorprese vengono invece dal corpus di verifica nel quale le prestazioni del sistema vengono, diversamente dalle aspettative, migliorate rispetto ai corpus di apprendimento. Nel caso del telegiornale della Lombardia si arriva ad una percentuale di errori dimezzata, nei telegiornali di Campania e Lazio si ottiene comunque un miglioramento delle prestazioni, mentre in quello della Toscana si ha solo un piccolo deterioramento dell'accuratezza con cui viene fatta la divisione in sillabe rispetto al parlato spontaneo.

	Lazio		Campania		Lombardia		Toscana	
totale sillabe	329		350		314		358	
errori	32	10%	30	9%	21	7%	53	15%
errori di margine	11	3%	6	2%	6	2%	8	2%
errori di numero	21	6%	24	7%	15	5%	45	13%
son stretta	13	4%	10	3%	10	3%	39	11%
son lato	10	3%	14	4%	5	2%	7	2%
confine	9	3%	6	2%	6	2%	7	2%

Tabella 4.4 Risultati sui corpora di verifica

Il miglioramento delle prestazioni del sistema con il corpus di verifica è spiegabile con il maggior controllo della pronuncia da parte dei parlanti che producono segnali più facilmente analizzabili.

4.3.1 Analisi quantitativa dei risultati

Il sistema qui proposto ottiene delle prestazioni comparabili se non superiori a quelli di altri sistemi più complessi. Il confronto con i sistemi proposti in [Reichl & Ruske, 1993] e in [Pfitzinger *et alii* 1996] può essere fatto considerando gli errori che vengono qui detti "errori di numero". La percentuale di errori dell'11% nel parlato spontaneo (sia esso di apprendimento o di test) è circa la metà di quella ottenuta dal sistema in [Pfitzinger *et alii* 1996] per altro molto simile alla impostazione qui proposta.[1] Gli errori di numero nei telegiornali di Lazio, Campania e Lombardia (rispettivamente 6%, 7% e 5%) risultano migliori dei risultati della rete neurale di [Reichl & Ruske 1993]. La complessa rete neurale di [Shastri *et alii* 1999] viene comunque superata dal peggiore dei nostri risultati (16% della rete neurale contro il 15% di errori complessivi nel telegiornale della Toscana).

[1] Il sistema di Pfitzinger è molto simile al primo stadio del nostro sistema.

4.3.2 Analisi qualitativa dei risultati

Il nostro sistema si basa sull'assunto di una corrispondenza diretta tra la grandezza fisica "energia" con il parametro fonologico "sonorità". L'analisi qualitativa dei risultati ha lo scopo di verificare fino a che punto questa relazione è soddisfatta, dove ciò non avviene e perché. L'elevata percentuale di accordo (mai sotto l'85%) tra la segmentazione fonologica e quella prodotta dal sistema esclude l'indipendenza tra questi due concetti che, quindi, comunque restano profondamente legati.

La presenza di errori di confine è irrilevante in questa analisi, in quanto essi mettono in discussione le regole fonotattiche che assegnano un fono minimo di sonorità ad una sillaba o all'altra ma non violano il principio di sonorità.

Gli interrogativi veri vengono posti dagli errori di sonorità (sia in senso stretto che lato). Nel caso degli errori di sonorità in senso stretto si ha che un segmento prodotto dal sistema presenta, di fatto, due massimi di sonorità. Questo avviene per tre possibili motivi:

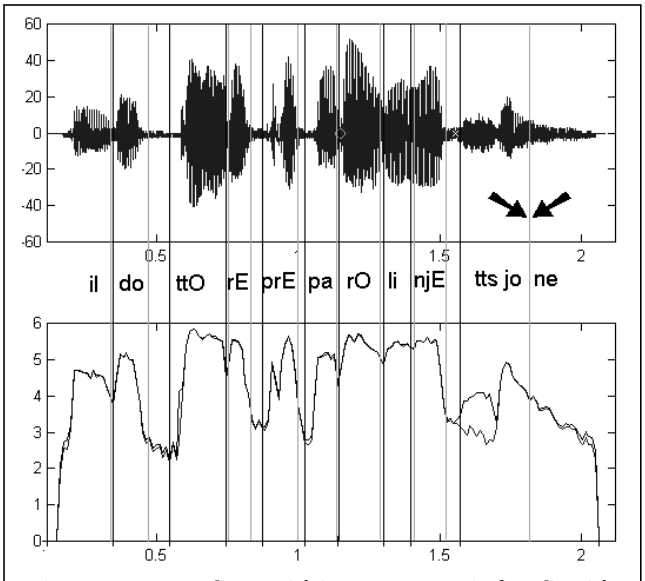

Figura 4.3 Errore di sonorità in senso stretto in fine di unità tonale. Le frecce indicano il confine di sillaba che non è stato trovato, poiché il minimo di energia non è presente.

- C'è un solo massimo dell'energia

- I due massimi sono molto vicini oppure non c'è molta differenza tra i valori dei massimi e quello del minimo tra essi compreso.

Il primo caso avviene in prossimità dei confini di unità tonale dove le variazioni di intensità mascherano, verso la fine, la variazione di energia che dovrebbe sussistere tra i foni. La figura 4.3 mostra un esempio di questo caso.

Il secondo caso avviene durante le brusche accelerazioni della velocità di eloquio, in questi casi l'energia presenta un diminuzione poco pronunciata in corrispondenza del minimo di sonorità, ed è quindi più difficile separare le due sillabe.

Gli errori di sonorità in senso lato riguardano soprattutto i foni sonoranti, cioè liquide, nasali e approssimanti. Questi foni vengono comunemente prodotti con un'energia molto superiore alle altre consonanti poiché nel loro modo di articolazione non esiste una vera e propria occlusione. Infatti tra gli errori di sonorità in senso lato sono presenti molte sonoranti che formano segmento a sé. In questo caso si comportano molto similmente a una vocale che forma sillaba a parte in quanto è presente un aumento dell'energia seguita da una sua discesa. Infatti gli errori di sonorità in senso lato sono essenzialmente costituiti da vocali toniche, che per la loro durata non mantengono un livello di energia costante per tutta la durata, ma presentano molte oscillazioni dovute all'irregolarità del flusso espiratorio del parlante (figura 4.4).

Le regole fonotattiche non sono state rispettate dagli errori di confine dovuti alla sbagliata attribuzioni di nasali, liquide, fricative e vibranti alla sillaba prevista dalla fonologia. Spesso esse sono state divise e messe metà in una sillaba e metà nell'altra; ciò è successo in particolare in presenza di geminate. Il valore numerico degli errori così generati da, però, conferma della nostra scelta di spostare tutta la geminata nell'attacco della sillaba che segue e, non secondo l'ortografia[1], divise tra le due sillabe.

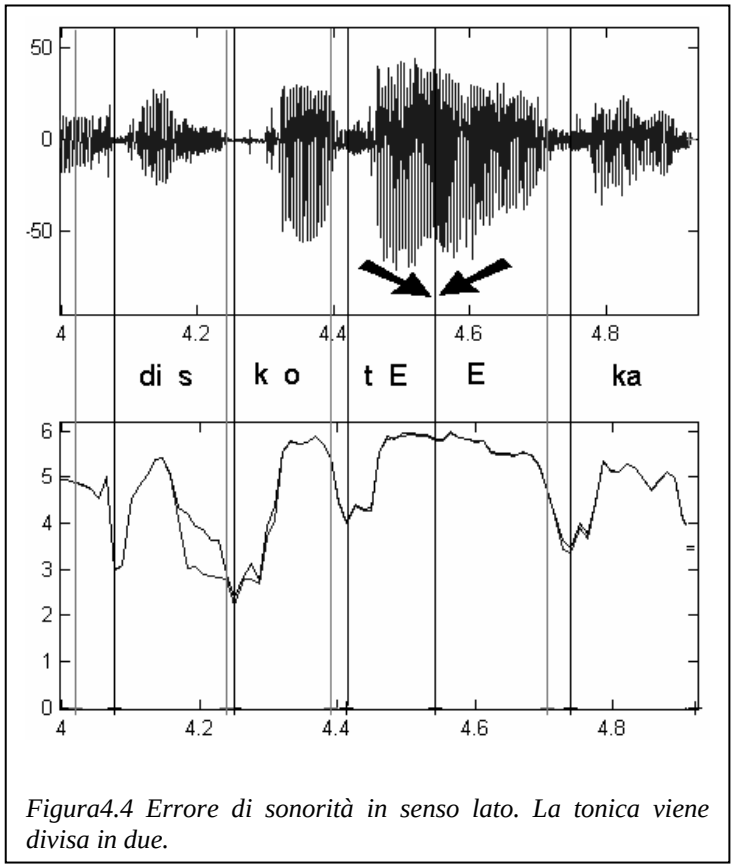

Figura4.4 Errore di sonorità in senso lato. La tonica viene divisa in due.

[1] Solo l'ortografia prevede, in maniera universalmente accettata, di considerare le consonanti geminate suddivise tra le due sillabe adiacenti, mentre è, invece, ancora aperta la discussione tra i linguisti sul come trattare le geminate.

4.4 Limiti del sistema e possibili soluzioni

In [Pfitzinger *et alii* 1996] viene indicato con la percentuale del 6% il limite minimo di errori di numero di un sistema di segmentazione automatica in sillabe. Si arriva a questo risultato facendo segmentare dei segnali di parlato continuo a diversi fonetisti e verificando l'accordo tra di loro. Queste differenze sono dovute alla natura soggettiva della percezione, secondo la quale un fono presente secondo un ascoltatore potrebbe non esserci secondo un altro o potrebbe essere etichettato diversamente. I risultati di Pfitzinger si riferiscono a sequenze di parlato in lingua tedesca ma non ci sono ragioni per dubitare che con altre lingue le cose vadano diversamente.

Rimangono ancora margini di miglioramento per un sistema di segmentazione automatica. Questo dato è sostenuto dal fatto che i fonologi sono attualmente orientati nell'indicare la sonorità come il grado di apertura del tratto vocalico, non, quindi dall'intensità del suono. A questa descrizione articolatoria della sonorità dovrebbe, tuttavia, corrispondere una grandezza acustica misurabile del segnale. In altri termini la segmentazione potrebbe dipendere anche dall'andamento della seconda formante, quella che appunto dipende dall'apertura del tratto vocalico. Questo approccio potrebbe essere usato in ulteriori versioni del sistema sviluppato, per il momento esso esula dai nostri scopi in quanto dovrebbe essere il primo stadio di ulteriori analisi del parlato che potranno quindi analizzare la seconda formante anche per altri scopi.

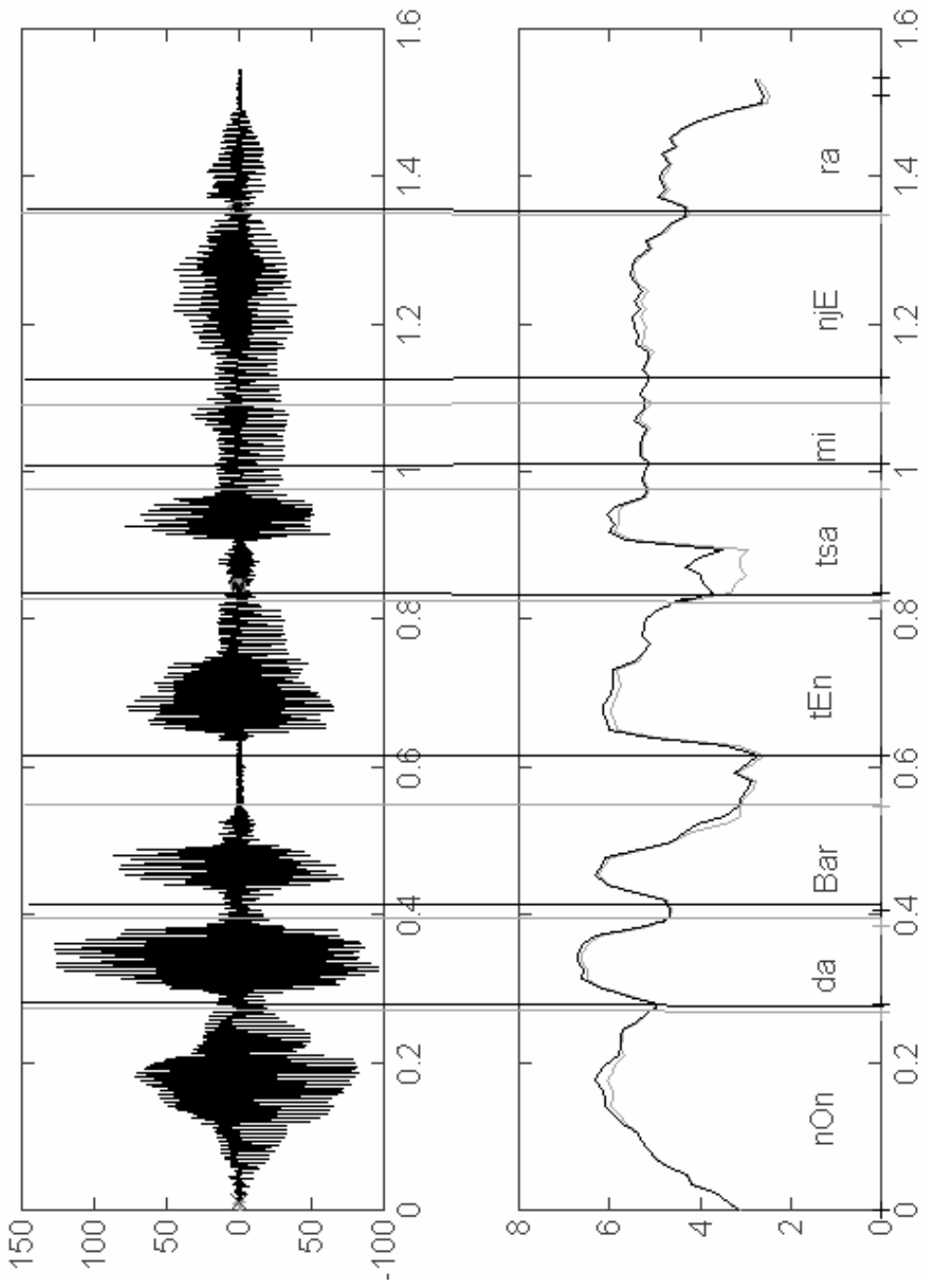

figura 4.5 Esempio di segmentazione eseguita correttamente.

Conclusioni

In questo lavoro si è riusciti a mostrare come, utilizzando semplici tecniche algoritmiche, si possa produrre una procedura in grado di segmentare il segnale verbale in sillabe. I dati usati sono solo le curve di energia del segnale e quella di una sua versione opportunamente filtrata. L'idea generale sulla quale si basa il metodo di segmentazione, è quella di individuare dei massimi e dei minimi nella curva di energia. E' stata necessaria, tuttavia, una selezione tra i massimi presenti sulla curva di energia al fine di non farsi ingannare da oscillazioni accidentali. Tuttavia, le condizioni secondo le quali vanno selezionati i massimi corrispondenti ai nuclei vocalici non sono banali. Questo è coerente con il fatto che in altri tentativi di segmentazione, sono state utilizzate tecniche molto complesse come ad esempio le reti neurali. In quei casi, vedi [Reichl & Ruske 1993] [Shastri *et alii* 1999], è stato inoltre necessario esaminare il segnale anche nel dominio della frequenza, aumentando, di conseguenza, la quantità di dati estratti.

I risultati ottenuti con il nostro sistema mostrano come sia possibile ottenere una buona segmentazione anche con dati più facilmente calcolabili. Non viene compiuta, a meno che per le fricative, alcuna azione di riconoscimento dei foni che costituiscono la sillaba, cosa che sarebbe stata implicitamente eseguita se fosse stato preso in considerazione lo spettro del segnale.

Il primo campo di applicazione in cui si potrebbe impiegare questo sistema di segmentazione è quello che lo vede come stadio iniziale di un riconoscitore automatico del parlato. L'approccio, qui presentato, cioè quello di segmentare il segnale in unità senza per questo individuarne l'identità, permetterebbe, invece, di ribaltare l'approccio con il quale normalmente si progetta un riconoscitore automatico del parlato, vale a dire riconoscere in primo luogo i foni che costituiscono le sillabe.

D'altro canto la segmentazione effettuata dal sistema non coincide strettamente con la divisione in sillabe eseguita da un linguista. Successive modifiche al sistema di segmentazione proposto, che migliorassero l'accordo con la divisione in sillabe di un fonetista, comporterebbero, necessariamente, l'analisi di altre caratteristiche acustiche del segnale andando nella direzione del riconoscimento di, almeno, alcuni foni. E' opportuno fare un esempio: supponiamo che in un certo punto del segnale accada che due sillabe separate da una sonorante siano incluse dal sistema in un solo segmento; un possibile successivo sviluppo della procedura potrebbe essere in grado di ricavare dallo spettro informazione sufficiente ad individuare la sonorante e ad

eseguire la segmentazione necessaria. Il sistema verrebbe così migliorato, ma solo grazie a dati che i linguisti chiamerebbero “segmentali”, cioè che riguardano un singolo fono.

Più in generale si potrebbe, quindi, immaginare un sistema di ASR modulare. Il primo stadio sarebbe costituito dal nostro sistema privato del trattamento delle fricative; il secondo, invece, potrebbe essere costituito proprio dal trattamento delle fricative, mentre il terzo si occuperebbe delle sonoranti.

Considerato che la maggior parte delle discordanze *gravi* (quelle, qui, definite come *errori di sonorità in senso stretto*) riguarda le nasali, i tre stadi sopra proposti permetterebbero di individuare all’interno della sillaba i tre elementi costitutivi (attacco, nucleo, coda). Resterebbero quindi da aggiungere degli stadi che etichettino[1] queste parti.

Questa ultima parte potrebbe considerare anche dati di carattere prosodico. Questi sarebbero estratti, in parte, anche dai risultati del segmentatore proposto, poiché la durata sillabica (banalmente ricavabile dal nostro sistema) è anch’essa un parametro prosodico legato al ritmo.

Un sistema di questo tipo uscirebbe dagli schemi consueti. Non sarebbe né un sistema del tipo "da piccolo al grande" né un sistema "dal grande al piccolo"[2]. Dati

---

[1] Si tenga presente che a questo punto si avrebbero già informazioni importanti per restringere il campo delle scelte per queste parti: l’attacco deve essere una consonante, il nucleo una vocale o un dittongo, la coda deve essere una sonorante o una fricativa.

[2] La terminologia informatica prediligerebbe per questi concetti i termini *top-down* e *bottom-up*. In termini cognitivi, invece, si ritiene più valido usare queste denominazioni per indicare processi che

"intermedi" (la segmentazione basata sulla sola curva di energia) sarebbero usati per riconoscere delle unità ad un livello segmentale (i foni), e allo stesso tempo verrebbero usate per ricavarne altri ad un livello prosodico[1].

L'uscita complessiva del riconoscitore dipenderebbe quindi dalla combinazione di questi tre livelli secondo lo schema riportato in figura 5.1

Si apre, a questo punto, un interrogativo interessante: la segmentazione che si otterrebbe con il nostro sistema, eliminando gli stadi che trattano le fricative, ha solo una somiglianza con il concetto di sillaba così come, finora, inteso dai linguisti. L'accordo tra la segmentazione automatica e quella manuale, mostrato nel quarto capitolo è dovuto solo all'uso di una tecnica di riconoscimento di alcune parti (le fricative) del segnale. Ciò significa che sono stati introdotti degli elementi di circolarità, quelli di cui si è parlato nel secondo capitolo, che fanno dipendere la segmentazione *fisica* (nella fattispecie quella automatica) dalle regole teoriche che, a loro volta, invece dovrebbero derivare dalle caratteristiche acustiche del parlato. La circolarità sarebbe ulteriormente accentuata se si cercasse di migliorare il sistema con stadi che risconoscono dei segmenti. D'altro canto le regole linguistiche di sillabificazione sono, implicitamente, derivate da tre diversi tipi di considerazioni:

---

partono da facoltà cognitive superiore per poi analizzare le informazioni percettive (*top-down*) o viceversa (*bottom-up*). La determinazione dell'unità di analisi percettiva è invece un problema diverso, il presente lavoro discute della possibilità di impiego di unità più grandi dei foni, ma rimane sempre collegata alle informazioni percettive (il segnale) e quindi, cognitivamente parlando, al *bottom*.

[1] Si noti, a questo proposito, che la maggior parte delle teorie correnti, sia nel settore dell'ASR che nelle scienze cognitive, ancora prediligono un approccio "piccolo-grande".

- articolatorie (apertura del tratto vocalico),
- acustiche (massimi di energia)
- programmazione del parlato (regole fonotattiche).

La segmentazione effettuata con la sola curva di energia, dipende, invece, solo dalle caratteristiche acustiche, quelle direttamente analizzate dal sistema uditivo umano. Potrebbe essere, quindi, utile considerare al posto delle sillabe due concetti, a priori, indipendenti:

- unità di programmazione del parlato, coincidente con il concetto di sillaba così come finora intesa dai linguisti;
- unità di riconoscimento (automatico o umano che sia) del parlato, coincidente con le segmentazioni del nostro sistema automatico.

I due meccanismi potrebbero essere più indipendenti di quanto finora ipotizzato. In altre parole la comprensione dei suoni non dipenderebbe dalle conoscenze implicite del parlante sull'articolazione degli stessi. In questo modo i processi di produzione e di comprensione del parlato risulterebbero separati.[1]

La distinzione tra questi due concetti potrebbe essere un elemento importante nella disputa tra i linguisti sulla definizione della sillaba descritta nel capitolo2.

---

[1] Questa affermazione è in contrasto con quanto sostenuto dalla *Motor theory* [Liberman *et alii,* 1967] secondo la quale i due meccanismi sono intimamente connessi. L'ipotesi che viene sostenuta è che le "conoscenze articolatorie" dell'ascoltatore possono entrare in gioco solo dopo che un'analisi del segnale acustico sia stata preventivamente effettuata. La *Motor theory* non ha mai fornito spiegazioni sul ruolo del segnale acustico e del sistema uditivo nel processo di riconoscimento dei suoni linguistici.

Nello schema di figura 5.1 viene, implicitamente, presentata un'ipotesi su come avviene il riconoscimento del parlato nell'uomo. In questa ipotesi è presente una separazione delle conoscenze attive tra fonetica percettiva e psicolinguistica (vedi introduzione); quest'ultima viene rappresentata dal blocco chiamato "composizione dei dati", mentre la fonetica percettiva è rappresentata dalla parte sinistra dello schema.

Dal punto di vista psicolinguistico, la distinzione tra le unità di programmazione e quelle di riconoscimento rafforzerebbe l'ipotesi esposta in [Cutugno, 1999] di asincronia tra parlante ed ascoltatore poiché i pacchetti inviati dal primo potrebbero non corrispondere a quelli ricevuti dal secondo. La produzione viene scandita dalle unità di programmazione, mentre la comprensione dalle unità di riconoscimento. Inoltre, una verifica dell'adeguatezza di questa distinzione potrebbe venire dalla ripetizione dell'esperimento di *syllable monitoring* di [Mehler *et alii* 1981] e [Cutler *et alii* 1986]. Questo esperimento è stato, inizialmente, eseguito su soggetti madrelingua inglese e consisteva nel verificare che una stessa sequenza fonica del tipo CV (Consonante-Vocale) presa dall'inizio di parole isolate venga riconosciuta meglio se essa costituisce l'intera prima sillaba piuttosto che solo il corpo di questa (ad esempio in inglese la seguenza *ba* veniva riconosciuta meglio se presa da *balance* che da *balcony*). Al risultato positivo per gli inglesi si sono contrapposti dei risultati diversi con altre lingue (vedi lo stesso [Cutler *et alii 1986*] per il francese, [Tabossi *et alii* 1995] per l'italiano e [Sebastien *et alii 1992*] per il castigliano). Si potrebbe, a

questo punto, proporre di usare al posto delle sillabe le unità di riconoscimento, se in questo modo i risultati di Mehler e Cutler venissero confermati anche per le altre lingue, l'utilità di questi nuovi concetti sarebbe indubbia.

Dal punto di vista metodologico, è risultato sorprendente il successo del procedura di aggiustamento fine dei parametri detto "semicasuale", secondo il quale è stato possibile migliorare il sistema provando a variare, di poco, in maniera totalmente casuale, tutti i parametri della migliore configurazione fin lì disponibile. Fissato il numero di "tentativi" pari a quelli effettuati in maniera più deterministica, i risultati sono stati sensibilmente migliori. Sarà opportuno verificare, in seguito, se la stessa tecnica può essere utile in altri problemi, in tal caso ci sarà un nuovo strumento a disposizione per ottimizzare l'assegnazione dei valori dei parametri in quei sistemi caratterizzati dall'avere un numero di parametri che supera le poche unità.

Lo schema di riconoscitore, a cui si è stato fatto cenno poco fa, è attualmente il canovaccio sul quale si basano altri lavori in svolgimento presso il Centro Interdipartimentale di Analisi e Sintesi dei Segnali (CIRASS). Da questi è già possibile ricavare le implementazioni di alcuni altri moduli, in particolare quelli del trattamento delle nasali e delle laterali, degli accenti e delle unità tonali.

I tratti nasali e laterali vengono individuati, così come è qui stato fatto con le fricative, confrontando la curve di energia del segnale completo con quella di una sua

versione filtrata con un taglia-banda. La possibilità di individuare le porzioni fricativi nasali e laterali permette di isolare le vocali ed applicare un modello secondo il quale dal prodotto di durata ed intensità delle vocali è possibile individuare le posizioni degli accenti. Le unità tonali vengono trovate dallo studio dell'andamento del pitch, dell'energia e delle durate. Tutti questi lavori sono basati soltanto su tecniche algoritmiche che permettono di ipotizzare , al tempo stesso, un modello dei processi cognitivi umani.

Riassumendo, si è cercato di affrontare un problema tipico dei sistemi di riconoscimento automatico del parlato da varie prospettive: linguistica, cognitiva ed informatica. Le prestazioni del sistema sono state esaminate anche dal punto di vista linguistico, dove è sorto il dubbio che il concetto di sillaba potesse non essere lo stesso sul piano della percezione e su quello della produzione del parlato: la sillaba acustico-percettiva non necessariamente deve coincidere con quella articolatoria, pur tuttavia essendo legate dalla necessità di sincronizzazione tra parlante ed ascoltatore. Dalla sillaba acustica-percettiva, quella che dovrebbe essere più direttamente legata ai meccanismi di comprensione umana, dovrebbero partire i tentativi di sviluppo di sistema di riconoscimento più simile a quello umano e per questo più versatili.

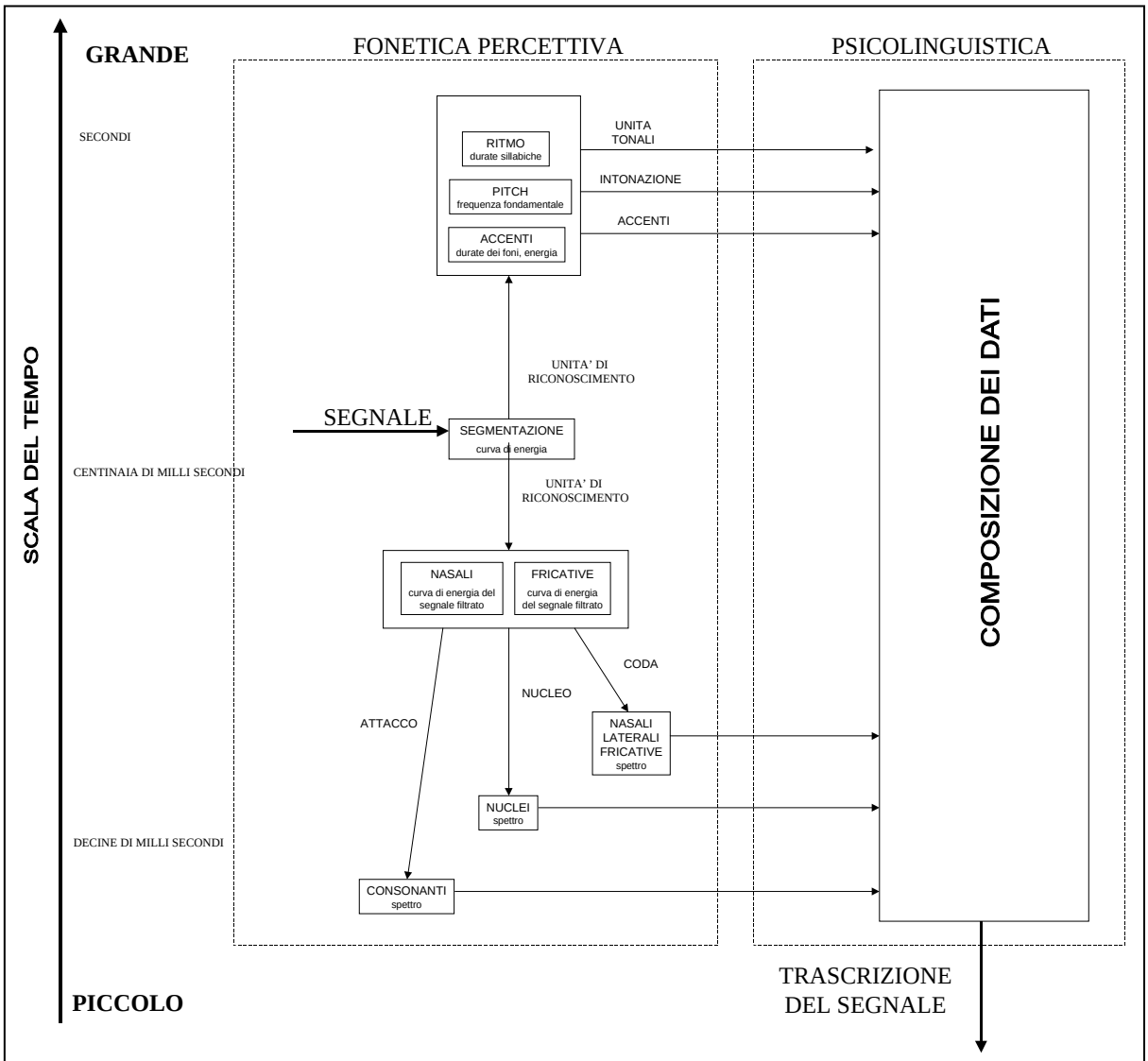

Figura 5.1 Schema del riconoscitore automatico del parlato proposto. Il segnale viene scomposto in unità (qui denominate come "unità di riconoscimento") che vengono usate sia a livello segmentale (cioè nella direzione delle sotto-unità) che a un livello soprasegmentale (cioè nella direzione di macro-unità che si ottengono dalla concatenazione delle unità di riconoscimento), l'informazione così raccolta viene esaminata da un blocco denominato "composizione dei dati" che ottiene la trascrizione del segnale integrando la conoscenza sul segnale acustico con quella proveniente da altre competenze(situazione pragmatica, contesto, grammatica eccetera).
In questo schema il ruolo svolto dall'intonazione è volutamente sottospecificato. Il ruolo che le variazioni della frequenza fondamentale giocano nei processi di riconoscimento del parlato, sia nei sistemi automatici sia nell'uomo, pur essendo un aspetto cruciale, non viene in dettaglio analizzato nel presente lavoro.

APPENDICE

Codici SAMPA usati

Simbolo SAMPA	descrizione	parola	trascrizione
p	occlusiva bilabiale sorda	pane	"pane
b	occlusiva bilabiale sonora	bara	"bara
t	occlusiva dentale sorda	tana	"tana
d	occlusiva dentale sonora	dado	"dado
k	occlusiva velare sorda	cane	"kane
g	occlusiva velare sonora	gatto	"gatto
ts	affricata dentale sorda (intervocalica sempre lunga)	zitto azione	"tsitto a’’ttsjone
dz	affricata dentale sonora (intervocalica sempre lunga)	zona azoto	"dzOna a’’ddzOto
tS	affricata palatale sorda	cena	"tSena
dZ	affricata palatale sonora	gita	"dZita
f	fricativa labiodentale sorda	fame	"fame
v	fricativa labiodentale sonora	vano	"vano
s	fricativa alveolare sorda	sano	"sano
z	fricativa alveolare sonora	zbaglio	"zbaLLo
S	fricativa palatale sorda (intervocalica sempre lunga)	scena, esce	"SEna, "ESSe
m	nasale bilabiale	mano	"mano
n	nasale dentale	nano	"nano
J	nasale palatale (intervocalica sempre lunga)	gnomo, bagno	"JOmo, "baJJo
r	liquida vibrante	rana	"rana
l	liquida laterale	lana	"lana
L	liquida palatale (intervocalica sempre lunga)	gli maglia	Li "maLLa
j	semivocale palatale	ieri	"jEri

w	semivocale labiovelare	uomo	"wOmo
B	allofono[1] approssimante dell'occlusiva bilabiale		
D	allofono approssimante dell'occlusiva dentale		
G	allofono approssimante dell'occlusiva velare		
Z	allofono fricativo dell'affricata palatale sonora /dZ/		
N	allofono velare della nasale		
M	allofono labiodentale della nasale		
4	allofono monovibrante di /r/		
v\	approssimante labiodentale		
h\	fricativa glottidale sonora		
H\	fricativa glottidale sorda		
?	colpo di glottide		
@	schwa, vocale centrale		
i	vocale anteriore alta	mite	"mite
e	vocale anteriore medio-alta	sera	"sera
E	vocale anteriore medio-bassa	meta	"mEta
a	vocale centrale bassa	rata	"rata
O	vocale postreriore medio-bassa	mora	"mOra
o	vocale postreriore medio-alta	voto	"voto
u	vocale postreriore alta	muto	"muto
C+C	consonante geminata (e lunga)	vacca bagno	"vakka "baJJo

[1] Due foni si dicono **allofoni** se appartengono allo stesso fonema

Bibliografia

Albano Leoni F., Maturi P. *Manuale di fonetica* La Nuova Italia Scientifica, Roma. 1995

Bertinetto P.M. *Strutture prosodiche dell'italiano* Accademia della Crusca, Firenze. 1981

Crystal D. *A dictionary of linguistics and phonetics* Basil Blackwell, Oxford. 1980

Cutler A., Mehler J., Norris D.G., Segui J., *The syllable's differing role in the segmentation of French and English* Journal of Memory and Language 25:385-400. 1986

Cutugno F. *Il tempo della voce* in Delmonte R., Bistrot A. (a c. di) *Aspetti computazionali in fonetica, linguistica e didattica delle lingue: modelli e algoritmi.* Atti delle IX Giornate del Gruppo di Fonetica Sperimentale, Venezia, pp. 231-242, 1999

Elliot *Child Language* A. J. Cambridge University Press. 1981

Grammont M., *Traité de phonétique* Delagrave, Parigi. 1946

Greenberg S. *Insights into spoken language gleaned from phonetic transcription of switchboard corpus* Proceedings of the Fourth International Conference on Spoken Language (ICSLP), Philadelphia, S24-27. 1996

Greenberg S. *Recognition in a key – towards a science of spoken language* in Proceedings of ICASSP98, Internantional Conference on Acoustics, Speech and Signal Processing, Seattle, pp. 1041-1045. 1998

Greenberg S.*Speaking in shorthand – a syllabe-centric perspective for undertanding pronunciation variation* Proceedings of the ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition 1998

Greenberg S.*Understanding speech understanding: towards a unified theory of speech perception* Proceedings of the ESCA Tutorial and Advanced Research Workshop on the Auditory Basis of Speech Perception, Kekrade (Netherlands), pp.47-56. 1996

Hauenstein A. *Using syllables in Hybrid HMN-ANN Recognition system* 5th European Conference on Speech Communication and Technology Rodi, , pp. 1203-1206, 1997.

Jespersen O. *Lehrbuch der Phonetic*, Leipzig e Berlin, B.G. Teubner. 1920

Kohler K.*Is the syllable a phonological universal?* Journal of Linguistics, 2, 1966

Kozhevnikov V., Chistovich L. *Speech: Articulation and perception*, Washington DC: U.S. Department of Commerce, Joint Publication Research Service. 1966

Liberman A.M., Cooper F.S., Shankweller D.P., Studdert-Kennedy M. *Perception of the speech code* Psichological Review. 1967

Mehler J., Dommergues J.Y., Frauenfelder U., Segui J. *The syllabe role in speech segmentation* Journal of Verbal Learning and Verbal Bejaviour 20:298-305. 1981

Nespor M. *Fonologia* Il Mulino, Bologna. 1993

Pfitzinger H.R., Burger S., Heid S. *Syllable detection in read and spontaneous speech* Proceedings of the Fourth International Conference on Spoken Language (ICSLP) 1996

Pulgram E. *Syllable, Word, Nexus, Cursus* Mouton, The Hague, 1970

Rabiner L., Juang B.H. *Fundamentals of speech recognition.* Prentice Hall 1993

Reichl W., Ruske G. *Syllabe segmentation of continuos speech with artificial neural networks* in Proceedings of Eurospeech93, 3rd European Conference on Speech Communication and Technology, Berlino, pp. 1771-1774. 1993

Saussure F. de *Course de linguitique générale*, Payot 1916 ed. it. a cura di T. De Mauro,

Corso di linguistica generale Laterza. 1967

Sebastien-Galles N., Dupox E., Segui J., Mehler J. *Contrasting effects in Catalan and Spanish* Journal of Memory and Language 31:18-32. 1992

Shastri L., Chang S., Greenberg S. *Syllabe detection and segmentation using temporal flow neural networks* Proceedings of the Fourteenth International Congress of Phonetic Sciences, San Francisco. 1999

Stetson R. H., *Motor Phonetics*, North-Holland, Amsterdam. 1951

Tabossi P., Burani C., Scott D. *Word identification in fluent speech* Journal of Memory and Language 34:440-467. 1995

Trubeckoj N. S. *Grundzüge der Phonologie*, Gottingen, 1958, ed. it. *Fondamenti di Fonologia*, Einaudi, Torino.1971

Wu S.L., Shire M., Greenberg S., N. Morgan *Integrating syllable boundary information into speech recognition* IEEE International Conference on Acoustics, Speech and Signal Processing, Munich, pp. 987-990. 1997

Ringraziamenti

Ringrazio prima di tutto i miei genitori, per la fiducia che hanno sempre riposto in me e per gli incoraggiamenti che non mi hanno mai fatto mancare, neanche nei momenti più difficili.

Un doveroso ringraziamento ai Proff. Giuseppe Trautteur e Federico Albano Leoni non solo per i puntuali consigli e suggerimenti utili per questo lavoro, ma per i numerosi insegnamenti cui devo gran parte della mia formazione scientifica.

Un ringraziamento speciale per tutti gli amici del Cirass per il tempo che mi hanno dedicato: Gianluca Passaro, Renata. Savy, Leandro D'Anna, Natale Lettieri, Claudia Crocco e Rosella Giordano.

Impossibile trovare le parole per ringraziare il *mio professore* il Dott. F. Cutugno, maestro, ma soprattutto amico, da molti anni.