Università degli studi di Roma
"La Sapienza"

Facoltà di Ingegneria
Corso di laurea : Ingegneria Elettronica
Dipartimento di Scienza e Tecnica dell'Informazione e della Comunicazione

Tesi di Laurea in Comunicazioni Elettriche
Anno Accademico 1998/1999

ANALISI ACUSTICA E SINTESI DELLE CONSONANTI AFFRICATE SINGOLE E GEMINATE IN ITALIANO

Relatore:
Prof. Maria Gabriella Di Benedetto

Laureando:
Simone Faluschi
Matr. N. 09092053

INDICE

VOLUME I

CAPITOLO 2
IL FENOMENO DELLA GEMINAZIONE E LE CONSONANTI AFFRICATE

CAPITOLO 3
LA BASE DI DATI, IL SOFTWARE E GLI STRUMENTI STATISTICI

CAPITOLO 4
L'ANALISI ACUSTICA DELLE CONSONANTI AFFRICATE: METODOLOGIA E RISULTATI

CAPITOLO 5
SINTESI DELLE CONSONANTI AFFRICATE

CAPITOLO 6
CONFRONTI E CONCLUSIONI

VOLUME II

INTRODUZIONE

La presente tesi è stata svolta presso il Laboratorio Voce del Dipartimento INFOCOM della Facoltà di Ingegneria dell'Università "La Sapienza" di Roma.

Scopo della ricerca è stato lo studio analitico delle caratteristiche delle consonanti affricate italiane [τΣ, δZ, ts, dz,] e delle vocali [a, i, u] coarticolate con esse e la loro sintesi. In particolare, nel corso del lavoro, sono stati sviluppati i seguenti punti:

- studio delle caratteristiche acustiche delle pronunce con particolare riferimento al fenomeno della geminazione
- sintesi delle pronunce tramite il sintetizzatore HLsyn

I campi di applicazione dei risultati ottenuti da lavori sul segnale vocale come il presente sono molteplici: la conoscenza approfondita del segnale vocale permette la realizzazione di algoritmi di compressione sempre migliori, facilitando le possibilità di comunicazione vocale a distanza. Anche il progetto di riconoscitori vocali non può prescindere da studi acustici sul segnale vocale. I risultati di queste analisi si rivelano poi fondamentali per l'implementazione di sintetizzatori vocali, sempre più presenti nelle nuove applicazioni tecnologiche.

Per la realizzazione di tale lavoro sono stati attuati i seguenti passi:

1. Organizzazione della base dati e delle pronunce disponibili.
2. Fase di studio teorico dei segnali vocali corrispondenti alle pronunce della base dati, atto ad individuare le caratteristiche delle consonanti sotto esame e il modo di operare l'analisi futura.
3. Scelta dei parametri caratteristici da estrarre durante il corso dell'analisi.
4. Misurazione di tutti i parametri nel dominio del tempo e nel dominio della frequenza.
5. Sviluppo di software di supporto per l'estrazione automatica di altri parametri utili per l'analisi.
6. Studio teorico dei test statistici necessari all'analisi dei dati acquisiti.
7. Ricerca e studio di software che implementassero i test statistici scelti per l'analisi.
8. Analisi statistica dei dati ottenuti dalle misure.
9. Interpretazione dei risultati ottenuti al punto precedente e formulazioni di ipotesi.
10. Classificazione delle consonanti singole/geminate sulla base delle ipotesi fatte.
11. Studio dei principi di sintesi del segnale vocale.
12. Studio del funzionamento del sintetizzatore.
13. Sintesi delle pronunce di consonanti affricate.
14. Confronto con altri lavori in letteratura riguardanti la geminazione in italiano ed in altre lingue.

La tesi è stata strutturata come segue:

Nel primo capitolo, vengono descritte la produzione della voce attraverso l'apparato fonatorio e la percezione attraverso l'organo dell'udito. Sono date anche le nozioni fondamentali di acustica, di fonologia ed, infine, di elaborazione numerica del segnale vocale.

Nel secondo capitolo viene trattato a livello teorico il fenomeno della geminazione, uno degli argomenti centrali di tutta la tesi, e viene data una descrizione particolareggiata delle consonanti affricate, oggetto del presente studio.

Nel terzo capitolo, di preparazione all'analisi, sono descritte la struttura della base dati e gli strumenti software usati nel corso della tesi, con particolare riferimento a UNICE. Vengono inoltre richiamati i principi teorici delle analisi statistiche utilizzate per l'analisi dei dati.

Il quarto capitolo descrive quindi l'analisi acustica delle consonanti affricate nel tempo, in frequenza e dal punto di vista energetico. In questo capitolo sono inoltre riportate le ipotesi formulate sulla base dell'analisi statistica condotta ed i risultati ottenuti.

Nel quinto capitolo vengono descritti i principi fondamentali della sintesi, il funzionamento del sintetizzatore HLsyn e come ciò sia stato applicato allo scopo particolare di sintetizzare le consonanti affricate.

In ultimo, il capitolo sei riguarda il confronto dei risultati ottenuti in questo studio con quelli di altri lavori sulla geminazione (sia sulla lingua italiana sia su altre lingue). Sempre in questo ultimo capitolo vengono forniti alcuni spunti per lavori futuri sulla voce.

Le appendici sono parte integrante e fondamentale di tutta la tesi: esse raccolgono tutti i dati relativi alle misure effettuate con le loro medie e statistiche.

In particolare, nelle appendici A, B e C sono raccolti, rispettivamente, i dati dell'analisi temporale, dell'analisi energetica e dell'analisi in frequenza. Nell'appendice D sono raccolti i listati dei programmi scritti in C utilizzati, infine nell'appendice E sono riportati i risultati completi dell'analisi statistica condotta sui dati.

Tutto il materiale descritto: la base dati, i programmi C, i dati relativi a tutte le misure, le pronunce sintetizzate ecc. sono archiviati su cd-rom e sono disponibili presso il Laboratorio Voce del Dipartimento INFOCOM.

CAPITOLO 1

LA VOCE: FISIOLOGIA, FONETICA, ACUSTICA ED INGEGNERIA

INTRODUZIONE

La voce è indubbiamente la più antica forma di comunicazione possibile tra gli esseri umani ed è ancora quella maggiormente utilizzata. Per questo motivo è facile rendersi conto che vi sono tantissimi aspetti legati alla voce e molte scienze hanno a che fare con essa. In questo primo capitolo saranno quindi esaminati brevemente gli aspetti principali legati alla voce.

Nel primo paragrafo verranno dati dei cenni di fisiologia umana per ciò che concerne gli apparati di percezione e di produzione; nel secondo paragrafo è trattato l'aspetto linguistico, in particolare quello fonetico, della lingua italiana. Nel terzo paragrafo sono dati cenni di fisica acustica. Infine si accennerà ad alcune tecniche ingegneristiche usate per l'analisi del segnale vocale.

1.1 CENNI DI FISIOLOGIA

1.1.1 L'organo dell'udito

Esaminiamo la struttura propriamente anatomica dell'orecchio e i complicati processi di fisiologia neurologica per mezzo dei quali le vibrazioni sonore sono trasmesse, attraverso il nervo uditivo, al cervello, dove vengono interpretate come suoni.

L'orecchio consiste di tre parti:

- **Orecchio esterno,** che comprende il **padiglione**, visibile esteriormente, e il **condotto uditivo esterno,** che fa capo alla membrana del timpano; questa parte dell'orecchio raccoglie e dirige i movimenti vibratori dell'aria.
- **Orecchio medio,** o **cassa del timpano**, che trasforma le vibrazioni dell'aria in vibrazioni liquide; esso consiste di una cassa piena d'aria e comunica con la parte posteriore della cavità delle fosse nasali attraverso la **tromba di Eustachio.** Il timpano ha la forma di un cilindro le cui basi presentano la convessità dell'una rivolta verso l'altra: queste due basi, distanti 3-6 millimetri (alla circonferenza), sono la **membrana del timpano** e il setto dell'orecchio interno. Queste due pareti e la catena di ossicini che le unisce costituiscono il meccanismo di trasmissione delle vibrazioni sonore all'orecchio interno. La membrana del timpano ha uno spessore di un decimo di millimetro; quanto alla forma, è approssimativamente quella di un cerchio con un diametro verticale che va da 10 a 11 millimetri. Benché sia tanto sottile, la membrana del timpano è resistentissima grazie allo strato interno di tessuto fibroso posto fra la pelle del condotto uditivo esterno e la mucosa che riveste interamente la cassa del timpano.
- **Orecchio interno,** la cui parete racchiude gli organi della percezione uditiva. In questa parete sono praticati due fori: la **finestra rotonda,** che ha un diametro di 1,5-2 millimetri ed è chiusa da una membrana simile a quella del timpano, e la **finestra ovale,** cui fa capo la catena di ossicini: il **martello,** l'**incudine** e la **staffa.** Questa catena trasmette le vibrazioni dell'aria al liquido dell'orecchio interno, che è molto più denso dell'aria. L'equilibrio fra il liquido, l'aria interna e l'aria esterna è mantenuto dai muscoli dell'orecchio medio e da quelli della tromba di Eustachio. E' il gioco della staffa e della membrana della finestra rotonda che determina il movimento del liquido dell'orecchio interno il quale, a sua volta, mette in movimento la membrana basilare in punti dipendenti dalla frequenza dello stimolo sonoro.

E' dunque nell'orecchio interno che si compie quel fenomeno che chiamiamo audizione; ne sono centro le cavità ossee che per la loro forma sono dette **labirinto:** il **vestibolo,** i **canali semi-circolari** e la **chiocciola**.

Il vestibolo, che è in comunicazione verso l'esterno con la cassa del timpano, verso l'interno con i canali semicircolari e la chiocciola, ha forma ovale ed è lungo 6 millimetri, largo 3 e alto da 4 a 5. Dei canali, due sono verticali; uno, quello superiore, di 15 millimetri, è disposto perpendicolarmente all'asse della rocca petrosa (l'osso temporale in cui è scavato il labirinto), l'altro, quello posteriore, di 18 millimetri parallelamente a quest'ultima; il terzo canale, quello esterno, di 12 millimetri, è orizzontale.

La chiocciola consiste di tre sezioni: un nucleo, detto **columnella** alto circa 3 millimetri, forato da canaletti che accolgono il nervo uditivo (**canale afferente**, **canale spirale** e **canale efferente**); un tubo cilindrico aperto a una base e chiuso all'altra estremità dopo che s'è avvolto a spirale tre volte attorno al nucleo; terza, infine, una lamella ossea che con il suo bordo interno divide il tubo cilindrico in due rampe di cui una comunica con la cassa del timpano, l'altra col vestibolo. Il nervo uditivo si dipana nel condotto uditivo interno; il labirinto è in comunicazione con il cervello attraverso l'**acquedotto del vestibolo**.

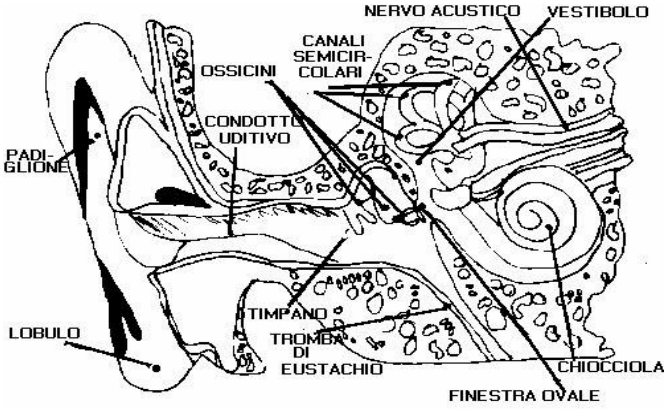


Fig. 1.1 Schematizzazione dell'organo dell'udito.

Le cavità del labirinto contengono un sistema di sacche e di tubi membranosi aderenti a una parte della parete dei canali ossei cui sono ancorati mediante sostegni fibrosi; le sacche sono contenute nei vestiboli, i tubi nelle cavità cilindriche. Questi condotti galleggiano in un liquido, la **perilinfa**, e sono pieni di un altro liquido, l'**endolinfa**. Le sacche del vestibolo sono in comunicazione fra loro mediante il canale endolinfatico dell'acquedotto vestibolare. Nelle sacche e nei canali sono collocati gli organi sensoriali.

Là dove il nervo uditivo sbocca nelle due sacche vestibolari (**utricolo** e **sacculo**)**,** la mucosa di rivestimento mostra tre tipi di formazioni cellulari: cellule **basali**, cellule **di sostegno** e cellule **sensoriali**. Nell'utricolo, nel sacculo e nelle ampolle, si trovano dei piccoli cristalli di carbonato di calcio.

Il canale cocleare è appoggiato alla parete del tubo cilindrico, cui è trattenuto dal legamento spirale, e alla lamina spirale, mediante la fasciola striata; esso sta dunque a cavallo delle due rampe della chiocciola da cui è separato mediante la **membrana di Reissmer** e la **membrana basilare**.

In perfetto equilibrio sulla membrana basilare si trovano gli organi uditivi. La mucosa del canale cocleare, al livello della parte interna della membrana basilare e in corrispondenza del punto in cui sboccano le ramificazioni terminali del nervo uditivo che spuntano dai **foramina nervina** della fasciola striata, si solleva a formare l'**organo del Corti**, il centro del quale è occupato da una serie di arcate. Le fibre nervose passano fra i pilastri che le sostengono. Ai due lati delle arcate si trovano le file delle **cellule uditive**, di cui 3.300 sono interne e 18.000 sono esterne, le quali presentano le **ciglia uditive** disposte a ferro di cavallo; le sovrasta la **membrana del Corti**.

Le ampolle su cui si innestano gli archi dei canali semicircolari sono considerate organi del senso dello spazio e dell'equilibrio; la percezione uditiva ha sede nelle vescicole del vestibolo e nella chiocciola. Le prime recepirebbero, pare, le vibrazioni aperiodiche che chiamiamo rumori, mentre le vibrazioni regolari, periodiche, ecciterebbero gli organi della chiocciola e ivi sarebbero percepiti come dei toni o suoni musicali.

Quando un'onda sonora colpisce la membrana del timpano mettendola in vibrazione, il movimento è trasmesso attraverso gli ossicini fino alla finestra ovale. I movimenti della staffa creano una pressione sulla perilinfa del vestibolo e questo scuotimento della perilinfa è a sua volta trasmesso attraverso la membrana di Reissmer all'endolinfa del canale cocleare così da provocare uno spostamento verso il basso sia della membrana basilare che della membrana reticolare e dell'organo del Corti.

Non si conosce ancora in tutti i suoi dettagli la maniera in cui funziona la chiocciola, tuttavia è stato stabilito con sicurezza che si ha uno spostamento massimo della posizione della membrana basilare ad ogni tono puro e che la posizione di questo spostamento varia al variare della frequenza dell'onda sonora che produce lo stimolo. Le onde ad alta frequenza causano uno spostamento massimo della membrana basilare fin vicino la finestra ovale alla base della coclea e le onde a bassa frequenza causano uno spostamento massimo verso la cupola della chiocciola. Quando la coclea è influenzata dalle vibrazioni di un'onda complessa, la membrana basilare viene spostata a dei punti corrispondenti alle frequenze delle componenti dell'onda. A ciascun punto di spostamento le ciglia dell'organo del Corti vengono scosse.

La ricerca dei fatti fisiologici e neurofisiologici che stanno dietro all'audizione, al livello dell'orecchio interno e a quello della corteccia, cioè fin nel centro uditivo del cervello, compete a diverse discipline; quel che interessa la fonetica è soprattutto il modo in cui l'orecchio reagisce ai diversi parametri fisici (frequenza, ampiezza, complessità, periodicità) dell'onda sonora che trasmette il messaggio linguisticamente formato. Il primo problema è pertanto di sapere qual è la gamma di frequenze e di ampiezze all'interno della quale l'orecchio è sensibile alle vibrazioni e alle differenze vibratorie.

1.1.2 Gli apparati di produzione della voce

L'apparato fonatorio dell'essere umano uomo è un insieme composto da un certo numero di organi la funzione primaria dei quali è, per tutti, una funzione eminentemente biologica: la respirazione, la deglutizione, ecc. L'apparato fonatorio umano è un adattamento ai fini comunicativi di organi la cui funzione è stata in origine, e resta tuttora, diversa. Si usa distinguere nell'apparato di fonazione le seguenti parti e funzioni:

- la realizzazione di una **corrente d'aria** che nell'assoluta maggioranza dei casi è una corrente espiratoria da parte dell'apparato respiratorio,
- la **sorgente sonora** responsabile delle vibrazioni periodiche utilizzate per la differenziazione fonetica (il tono glottidale): la laringe,
- e i **risuonatori** o cavità sopraglottidali.

Apparato respiratorio

La **respirazione,** addominale o costale a seconda dei casi, è una condizione essenziale per la formazione dei suoni del linguaggio ma contribuisce ben poco a differenziarli e non c'è bisogno di descriverla.

La **laringe** è una specie di scatola cartilaginea che forma la parte superiore della trachea; essa è composta di quattro cartilagini: la cricoide che ha forma di anello e ne costituisce la base, il corpo tiroide che è attaccato alla cricoide per mezzo di due corna, aperte verso l'alto e all'indietro, e le aritenoidi, due piccole piramidi poggiate sul castone della cricoide in modo da poter essere mosse mediante un sistema di muscoli.

La parte posteriore delle aritenoidi (l'apofisi muscolare) è il punto di appoggio dei muscoli che muovono le aritenoidi e comandano così l'apertura e la chiusura della glottide, cioè lo spazio circoscritto dalle due corde vocali e dai loro prolungamenti nelle apofisi vocali.

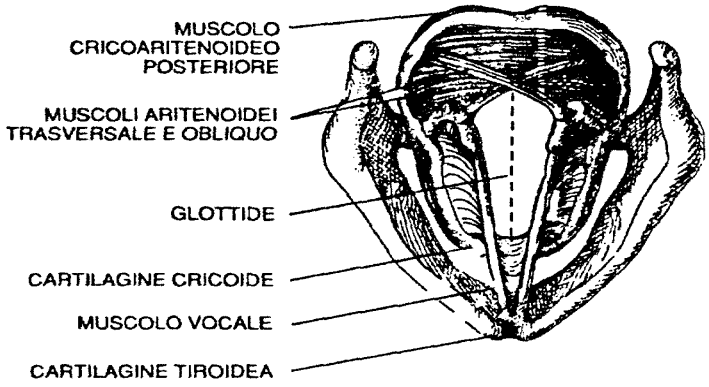


Fig. 1.2 Sezione longitudinale della laringe.

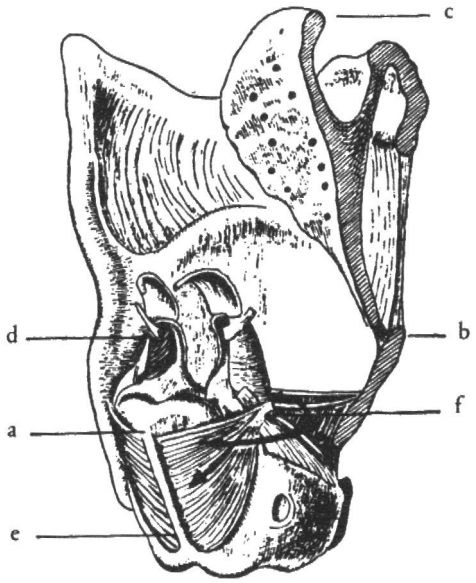


Fig. 1.3 La laringe vista da dietro. a: cartilagine cricoidea; b: cartilagine tiroidea; c: epiglottide; d: aritenoide (sinistra); e, f: muscoli (le frecce indicano le direzioni di movimento).

Tutte le pareti interne della laringe sono rivestite di una mucosa; questo tessuto forma sui lati dell'interno del corpo tiroide due coppie di pieghe che formano due rilievi orizzontali nella laringe. Sono queste pieghe che vengono chiamate **corde vocali** e **false corde vocali**.

Le corde vocali sono un muscolo rivestito di mucosa formato da cinque strati di tessuto con proprietà meccaniche differenti, che servono ad assicurarne una vibrazione corretta. Nell'uomo sono lunghe circa 23 mm e nella donna 18 mm, mentre l'apertura media glottale è di circa 5 mm^2 con picchi tipici dell'ordine di 15 mm^2.

Le tasche che si formano entro queste due pieghe si chiamano **ventricoli di Morgagni**. Le corde vocali si riuniscono in avanti nell'angolo della tiroidea; dietro esse sono attaccate alle apofisi vocali delle aritenoidi. Le aritenoidi sono attaccate al castone della cricoidea e sono mobili in più di una direzione: verso l'esterno, in posizione di riposo, verso l'interno, per chiudere la glottide, e verso l'alto e verso il

basso. In posizione di riposo esse si trovano a una certa distanza l'una dall'altra in modo che formano un triangolo col vertice nell'angolo della tiroide.

Il meccanismo che muove le aritenoidi è stato studiato e descritto dall'anatomista svedese Bertil Sonesson. E'grazie a questi movimenti delle aritenoidi realizzati mediante un sistema di muscoli che può essere variata la forma della glottide (cfr. fig. 1.4). Si distinguono quattro posizioni principali della glottide (cfr. fig. 1.5):

- la prima, triangolare, è utilizzata durante la normale respirazione;
- la seconda, pentagonale, è quella della respirazione profonda;
- la terza, con i bordi dei labbri incollati uno all'altro, ma con le aritenoidi separate, è quella che si adopera nel bisbiglio (infatti i suoni bisbigliati si formano al passaggio dell'aria attraverso lo stretto canale fra le aritenoidi);
- la quarta posizione della glottide è quella della fonazione: la glottide è chiusa in tutta la sua lunghezza e l'aria in uscita passa con una serie di scosse fra i bordi vibranti delle corde vocali.

Infine è possibile far assumere alle corde vocali una quinta posizione: i bordi possono essere appoggiati uno sull'altro e la conseguenza è una chiusura completa (occlusione) del passaggio dell'aria, questa posizione caratterizza la consonante detta colpo di glottide.

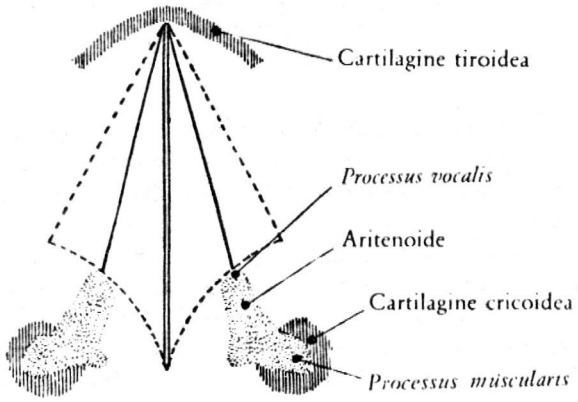


Fig. 1.4. Disegno schematico del meccanismo di apertura e chiusura della glottide. Le due linee più grosse indicano la posizione delle corde vocali durante la respirazione normale, le linee tratteggiate più grosse la posizione durante la respirazione profonda. Le due linee verticali sottili indicano la posizione di fonazione. Le linee tratteggiate sottili indicano la direzione del movimento delle aritenoidi quando la glottide cambia forma. (Da I.Tarneaud).

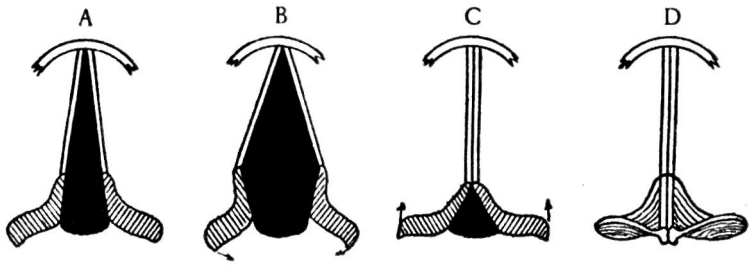


Fig. 1.5. Disegno schematico delle differenti posizioni della glottide: A respirazione normale, B respirazione profonda, C bisbiglio (le corde vocali sono chiuse ma il passaggio fra le aritenoidi resta libero), D fonazione. (Da J. Forchhammcr).

E' dunque grazie alle cartilagini aritenoidi e ai muscoli che ne comandano i movimenti che è possibile far variare la forma, la posizione e la tensione delle corde vocali interessate che possono vibrare o no al passaggio dell'aria attraverso la glottide. Il muscolo cricotiroideo, ad esempio, contribuisce al controllo dell'altezza dei suoni emessi quando le corde vibrano, variandone la tensione longitudinale e provocando così una loro deformazione. La variazione di tensione comporta una modifica delle frequenze di vibrazione delle corde vocali. E' noto, infatti, che le frequenze proprie di risonanza di una corda di lunghezza l soggetta ad una tensione T e fissata agli estremi, sono date dalla:

$$\nu = \frac{n}{2l}\sqrt{\frac{T}{\mu}} \qquad \text{n = 1, 2, 3, ...} \qquad (1.1)$$

ove μ rappresenta la densità lineare della corda. La laringe ha una tendenza naturale ad alzarsi e abbassarsi proporzionalmente all'ampiezza del suono prodotto, compromettendo così la sua emissione con qualità vocali costanti. Ciò può essere evitato impiegando i muscoli estrinseci per cercare di mantenere stazionaria la posizione dello scheletro cartilagineo.

Le **CAVITÀ SOPRAGLOTTIDALI** sono la **faringe**, **la cavità orale** e le **fosse nasali**.

La **cavità faringea** si estende fino alla glottide e può essere compressa ritraendo la radice della lingua verso la parete della faringe. Mediamente la lunghezza dell'intero condotto vocale è di 17 cm negli uomini.

La **cavità nasale** è principalmente ossea e quindi la sua forma è fissa. Essa può essere isolata dal resto del condotto vocale sollevando il **velo palatino** o **palato molle**. Così facendo, si solleva il diaframma rinovelare che mette in comunicazione la cavità nasale con quelle orale e faringale. Quando il condotto vocale è in posizione di riposo, il velo pende, estendendosi verso il basso, e il diaframma rinovelare è dunque aperto. Durante la produzione della maggior parte dei suoni linguistici, il velo è sollevato ed il diaframma è chiuso ma, nel caso di suoni nasali o nasalizzati, esso rimane aperto in modo che l'aria possa passare attraverso la cavità nasale per uscire dalle narici. Nell'uomo la cavità nasale ha una lunghezza e un volume medi rispettivamente di circa 12 cm e 60 cm^3.

La **cavità orale** si trova essenzialmente tra la lingua ed il palato e termina alle labbra. Essa può assumere un grandissimo numero di conformazioni diverse a causa del movimento della mandibola, delle labbra, della lingua e del velo palatino (organi fonatori mobili). Gli organi fonatori fissi sono i denti, gli alveoli ed il palato.

La cavità formata dalla protrusione e dall'arrotondamento delle labbra la si può considerare come quarto risuonatore. E' essenzialmente grazie ai movimenti della lingua che è possibile cambiare la forma e il volume, e di conseguenza l'effetto risuonatore, della faringe e della cavità boccale. Dal punto di vista delle possibilità articolatorie, bisogna distinguere fra il dorso e l'apice della lingua (articolazioni dorsali e apicali). La volta della cavità orale presenta le seguenti regioni (fra parentesi le denominazioni rispettive delle articolazioni che vi si formano):

- i denti (dentali),
- gli alveoli (alveolari),
- il palato duro (palatali, distinte in prepalatali, mediopalatali e postpalatali)
- il palato molle, o velo palatino (velari), con l'ugola o *uvula* (uvulari).

1. labbra
2. denti
3. gengive (alveoli)
4. palato duro
5. palato molle (velo)
6. uvula
7. punta della lingua (apice)
8. parte anteriore della lingua
9. parte posteriore della lingua
10. laringe
11. epiglottide
12. corde vocali

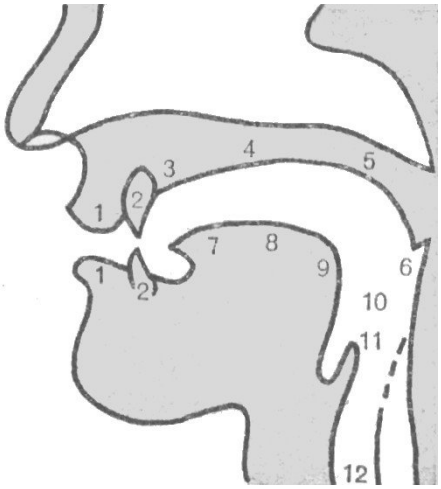


Fig. 1.6 .

Dietro si ha infine la parete posteriore della faringe (faringali). Un'articolazione con la partecipazione delle fosse nasali è detta nasale, o nasalizzata. Le articolazioni realizzate mediante le labbra sono dette labiali e più particolarmente, bilabiali se sono in gioco tutt'e due le labbra, labiodentali se il labbro inferiore va a toccate gli incisivi superiori, o il contrario, come accade talvolta. E' servendosi di combinazioni di questi termini che si arriva a definire abbastanza esattamente la maggior parte dei tipi articolatori che sono impiegati nel linguaggio: apico-dentali, dorso-palatali, dorso-velari, ecc., composti nei quali il primo termine indica l'organo articolante, il secondo il punto di articolazione come vedremo più dettagliatamente nel prossimo paragrafo.

1.2 LA SCIENZA DELLA FONETICA

1.2.1 Generalità

La **fonetica** è la scienza che si occupa dello studio della lingua parlata. Esistono diversi approcci allo studio di questa scienza: quello **articolatorio**, che studia la produzione dei suoni in funzione degli organi fonatori, quello **uditivo** o **percettivo**, che studia le modalità di acquisizione ed elaborazione delle informazioni fonetiche da parte del cervello umano, quello **funzionale** (fonologia) che analizza la struttura di un sistema fonologico dato o i principi generali della determinazione e della descrizione dei fonemi, interessandosi anche al valore e alla funzione che i suoni hanno in relazione con il loro significato. Altro approccio di interesse per noi è quello **acustico**, che studia strumentalmente le caratteristiche fisiche dei suoni.

I linguaggi in uso nel mondo sono composti ad alto livello da **morfemi**, che sono unità portatrici di significato (ad esempio la parola tavolino è articolata nei morfemi *tavol*, *in*, e *o*, con /tavol/ che ci dà l'informazione denotativa sull'oggetto, /in/ sul fatto che ci si sta riferendo ad esso con un diminutivo e /o/ sul suo genere, maschile, e numero, singolare), e dai cosiddetti **fonemi** a basso livello. I fonemi sono le unità minime distintive non dotate di senso che, combinandosi fra loro, permettono di formare le unità portatrici di significato o morfemi.

L'insieme dei fonemi di una lingua costituisce il complesso dei suoni elementari previsti dalle sue regole di pronuncia. Le realizzazioni foniche di un fonema sono dette **allòfoni**; ve ne sono teoricamente infiniti, in funzione delle caratteristiche dei diversi parlatori: loro età, sesso, stato d'animo, provenienza, etc. Una delle principali cause della diversità di realizzazione di un fonema da parte di uno stesso parlatore, anche a pochi secondi di distanza, è rappresentata dall'influenza dei fonemi confinanti nella sequenza pronunciata: si parla in tal caso del fenomeno della **coarticolazione**.

Anche nella lingua italiana si trovano numerosissimi allofoni o realizzazioni concrete di un solo suono (basti pensare alla /s/ pronunziata da un settentrionale, da un toscano o da un meridionale); tuttavia i fonemi nell'italiano sono soltanto 28.

Per individuare i fonemi bisogna ricorrere alla **prova linguistica di commutazione:** se esistono almeno due parole in italiano il cui significato varia esclusivamente per la sostituzione di un suono, allora diremo che quel suono è un fonema del sistema fonologico della nostra lingua. Così, nella sequenza *...atto,* potremo avere le coppie *gatto-matto,* o *fatto* e *ratto,* cioè dei significanti diversi, differenziati dai fonemi /g/, /m/, /f/, /r/.

Dato che le lettere del nostro alfabeto sono soltanto ventuno, vuol dire che i segni di trascrizione o **grafemi** non corrispondono esattamente ai suoni e le lettere non coincidono con i fonemi: una lettera può servire per più di un fonema o, viceversa, uno stesso fonema è trascritto con più grafemi; vi sono inoltre dei fonemi trascritti con due o tre lettere (i **digrammi** e i **trigrammi).**

Fonemi		Lettere
/a/		a
/b/		b
/č/	(cero)	c (digramma *ci* e *ch*)
/k/	(casa)	
/d/		d
/é/	(néro)	e
/è/	(bène)	
/f/		f
/ǧ/	(*g*ita)	g (digramma *gi* e *gh*)
/ɣ/	(*g*ara)	
—		h
/i/		i
/l/		l
/λ/	(fo*gli*o)	— (digramma *gl* e trigramma *gli*)
/m/		m
/n/		n
/ɲ/	(*gn*omo)	— (digramma *gn*)
/ó/	(póllo)	o
/ò/	(pòco)	
/p/		p
—		q
/r/		r
/s/	(suono)	s
/ʂ/	(caso)	
/ʃ/	(*sc*emo)	— (digramma *sc* e trigramma *sci*)
/t/		t
/u/		u
/v/		v
/z/	(pazzo)	z
/ʐ/	(zona)	

Tab. 1.1 Lettere e fonemi italiani

Le opposizioni fra /s/ sorda (*suono, casa* nella pron. toscana) e /s/ sonora (*smania, rosa* e *casa* nella pron. settentrionale) e fra /z/ sorda (*pazzo, zio* nella pron. toscana) e /z/ sonora (*zero, zio* nella pron. settentrionale) non sono sicuramente avvertite nei vari tipi di italiano regionale: così pure le opposizioni fra vocali aperte e chiuse: /é/ chiusa ed /è/ aperta non sono sempre distinte (si veda la pronuncia settentrionale di bene, vento, pesca con la *e* chiusa); ancora meno sentita la differenza fra /ó/ chiusa e /ò/ aperta, anche negli omografi come *bótte* (recipiente) e *bòtte* (percosse). Pertanto il numero dei fonemi con funzione realmente distintiva nell'italiano contemporaneo è di 24.

Per un uso puramente legato alla fonetica è stato creato, ed è oramai standardizzato, il metodo della trascrizione fonetica, che prevede l'uso di un set di caratteri diverso da quello dell'alfabeto, contenente un carattere per ciascuno dei fonemi (non degli allòfoni) previsti dalle lingue in uso. Una descrizione grafica standard dei suoni delle varie lingue è rappresentata dal sistema International Phonetic Alphabet (I.P.A.).

1.2.2 La fonetica articolatoria

Ogni suono linguistico è compreso in una delle due classi principali chiamate tradizionalmente vocali e consonanti. Riservando l'uso di questi termini al senso più scientifico della fonetica funzionale, in questo contesto si useranno i termini vocoidi e contoidi. Per lo studio dell'articolazione di tutti i fonemi ci si serve di diagrammi che mostrano la posizione dei vari organi coinvolti. In particolare, per i vocoidi si

usa il **trapezio fonetico**, e per i contoidi lo **spaccato sagittale** (sezione di profilo) dell'apparato fonatorio[1].

Articolazione dei vocoidi

Si possono definire **vocoidi** (in termini articolatori) quei suoni sonori, che sono prodotti dall'aria (proveniente dalla glottide) che non incontra alcuna ostruzione (nemmeno parziale) tra gli organi fonatori, né restringimenti tali da produrne la frizione. Il suono caratteristico di ciascun vocoide dipende soprattutto dalle posizioni assunte da due organi fonatori: lingua e labbra. In particolare, dipende dal sollevamento/abbassamento e avanzamento/arretramento della lingua (che può quindi muoversi in uno spazio schematizzato come bidimensionale) e dall'arrotondamento o meno delle labbra (spazio unidimensionale). Le possibili posizioni verticali della lingua rispetto al palato sono cinque: *alto*, *medioalto*, *medio*, *mediobasso* e *basso*; quelle orizzontali sono tre: *palatale*, *prevelare e velare* (o anteriore, cenrale, posteriore). La figura 1.7 mostra, invece, i particolari della posizione delle labbra durante l'articolazione delle tre vocali estreme italiane [i, a, u].

Il **trapezio fonetico** può ben rappresentare, schematicamente, uno spazio tridimensionale dove far "muovere" i vocoidi: sull'asse orizzontale e su quello verticale si rappresenta la rispettiva posizione della lingua[2], mentre un punto disegnato arrotondato o no rappresenta la posizione delle labbra. Nella figura 1.8 è disegnato il trapezio fonetico con i sette vocoidi propri dell'italiano.

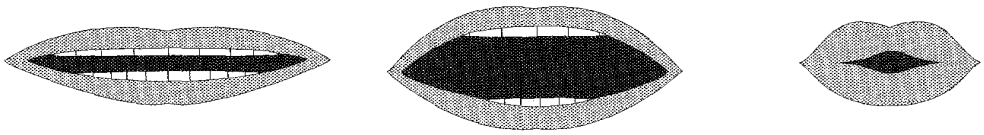


Fig. 1.7 Posizione delle labbra nelle tre articolazioni vocaliche estreme dell'italiano neutro: Labbra non arrotondate per la vocale alta anteriore [i] Labbra non arrotondate per la vocale bassa centrale [a] Labbra arrotondate per la vocale alta posteriore [u] (Canepari, 1992).

[1] Per descrivere adeguatamente le articolazioni di certe consonanti, il metodo fonetico accosta utilmente agli spaccati "sagittali", anche spaccati "ortogonali" (sezioni orizzontali normali al profilo) e spaccati "trasversali" (sezioni verticali di prospetto).

[2] Poiché i movimenti orizzontali della lingua in posizione bassa sono meno ampi, il campo dei possibili punti di articolazione viene racchiuso in un trapezio.

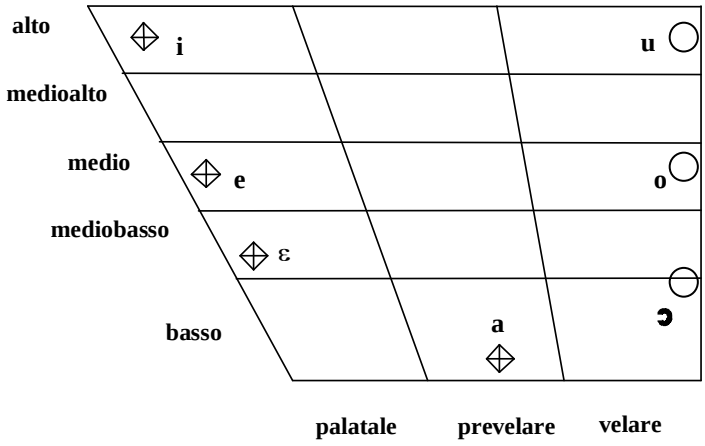


Fig. 1.8 Trapezio fonetico dell'Italiano (Canepari, 1979).

Articolazione dei contoidi

Si possono definire **contoidi** tutti quei suoni che non hanno le caratteristiche dei vocoidi. Infatti, nella produzione della maggior parte dei fenomeni consonantici si ha la formazione di costrizioni al passaggio dell'aria a causa dell'accostamento degli organi mobili contro le altre parti del condotto vocale.

La posizione in cui si forma la costrizione è detta **punto di articolazione** e se ne possono individuare diversi come mostrato in figura 1.9. Per quanto riguarda i punti di articolazione, in italiano, ce ne sono otto fondamentali individuabili:

- **Labiovelari**, che interessano labbra, dorso della lingua e velo palatino (p. es. il fonema /w/ di **u**omo);
- **Bilabiale**, in cui, per realizzare il modo di articolazione, vengono usate entrambe le labbra (p. es. i fonemi /p/ di **p**a**p**a, /b/ di **b**iro, /m/ di **m**ano);
- **Labiodentale**, che prevede l'uso del labbro inferiore e dei denti superiori (p. es. i fonemi /f/ e /v/ **f**a**v**o);
- **Dentale**, in cui sono interessati la punta della lingua e i denti superiori (p. es. i fonemi /s/, /ts/, /d/ e /t/ di **s**en**z**a **d**i **t**e,/dz/ di **z**ero, /z/ di o**s**are);
- **Alveolare**, realizzato con la punta della lingua e gli alveoli che prendono parte all'articolazione (p. es. i fonemi /r/ di **r**ane, /l/ di **l**ana, /n/ di **n**a**n**a);
- **Alveopalatale**, con la lingua alta e con la punta in zona intermedia tra alveoli e palato (p. es. i fonemi /τΣ/ di **c**inta, /δZ/ di **g**iro e /Σ/ di **sc**immia);
- **Palatale**, con il dorso della lingua ed il palato (p. es. i fonemi /j/ di **i**eri, /λ/ di **gl**i, /ɲ/ di le**gn**o);
- **Velare**, con il dorso della lingua ed il velo (p. es. i fonemi /k/ e /g/ di **c**an**g**uro).

Altri punti di articolazione vengono usati nelle realizzazioni allofoniche, tra i quali è di interesse il punto **prevelare** (p. es. i fonemi /k/ e /g/ seguiti dal fonema /i/ vengono realizzati, a causa dell'effetto

della coarticolazione, sul punto di articolazione prevelare, come in **ch**ina e **gh**iro). Rispetto al punto d'articolazione velare, in tal caso, la parte interessata risulta più spostata verso il palato.

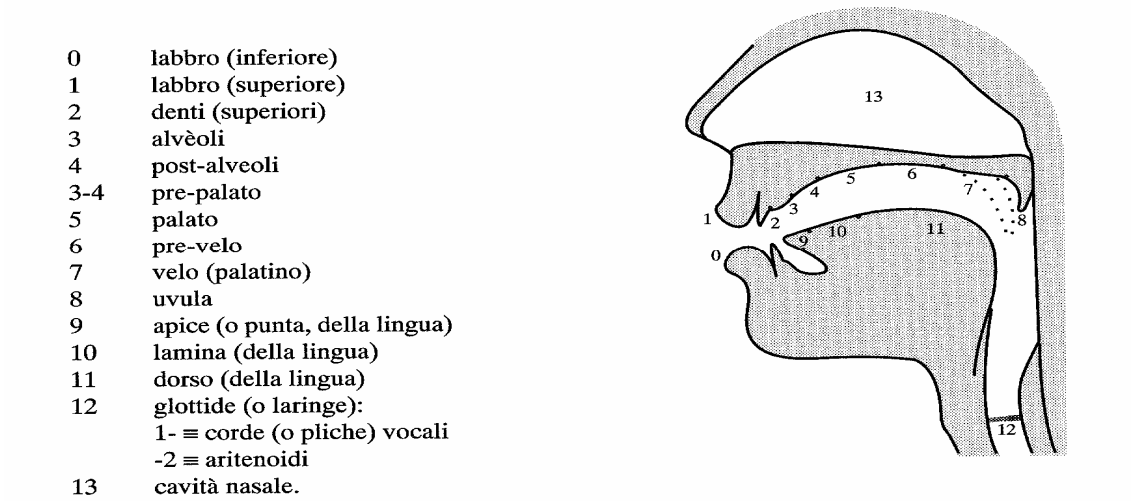


Fig. 1.9 Punti di articolazione (Canepari, 1992).

Il modo in cui la costrizione si realizza è detto **modo di articolazione**. Si distinguono, secondo questo aspetto, i seguenti gruppi di contoidi:

- **Occlusivi**, realizzati bloccando completamente il flusso d'aria, portando a contatto due organi fonatori e rilasciando in seguito velocemente tale costrizione (p. es. i fonemi /t/ e /p/ di **t**i**p**o, /k/ e /d/ di **d**i**c**o);
- **Fricativi**, prodotti operando un'occlusione non completa, che causa una particolare frizione dell'aria uscente (p. es. i fonemi /f/ e /v/ di **f**a**v**o, /s/ di **s**ano, /z/ di o**s**are, /ʃ/ di **sc**ena);
- **Affricati**, realizzati da un'occlusione seguita immediatamente da una frizione: si noti che non si tratta di un fonema occlusivo seguito da uno fricativo, il passaggio è rapidissimo e dà luogo ad un suono originale (p. es. i fonemi /ts/ di **z**ucchero, /dz/ di **z**aino, /τΣ/ di **c**ima, /δZ/ di **g**iugno);
- **Nasali**, prodotti occludendo il tratto vocale orale ma senza tenere il velo schiacciato sulla parete faringale retrostante come per gli altri, in modo che l'aria fluisca dal naso (p. es. i fonemi /m/ e /n/ di **m**a**n**o);
- **Laterali**, realizzati bloccando il flusso d'aria al centro della bocca ma lasciandola fluire lateralmente (p. es. i fonemi /l/ e /λ/ di **l**u**gl**io);
- **Vibranti** (mono o poli vibranti), prodotti mediante la vibrazione di un organo mobile (p. es. il fonema /r/ di **r**osa).
- **Approssimanti**, in cui la frizione è molto lieve, al punto che talvolta vengono indicati con il termine di semivocali o di semiconsonanti (p. es. il fonema /j/ di **i**eri e il fonema /w/ di **u**omo);

Suddivisioni dei fonemi di questo tipo permettono di costruire tabelle dove i fonemi sono raggruppati per tratti distintivi misti, come quella per l'italiano, riportata in tabella 1.2.

MODO DI ARTICOLAZIONE	PUNTO DI ARTICOLAZIONE							
	Labio-velari	Bilabiali	Labio-dentali	Dentali	Alveolari	Alveo-palatali	Palatali	Velari
Approssimanti	w						j	
Fricativi			f, v	s, z		Σ		
Affricati				ts, dz		τΣ, δZ		
Occlusivi		p, b		t, d				k, g
Vibranti					r			
Laterali					l		λ	
Nasali		m			n		ɲ	

Tab. 1.2 Tabella dei contoidi italiani (Muljacic, 1972)

Quindi, dal punto di vista della fonetica articolatoria, le consonanti si distinguono sulla base delle loro tre componenti indispensabili: il tipo di fonazione (sorda o sonora) su cui torneremo nel paragrafo 1.3.2, il modo di articolazione e il punto di articolazione. Per evidenziare il fatto che questo tipo di classificazione non è l'unico possibile, vedremo nel prossimo paragrafo il confronto con la classificazione operata tramite la fonetica binarista.

1.2.3 La fonetica Binarista

Secondo la **fonetica binarista**, dovuta al fonetista Jakobson, esistono una dozzina di tratti distintivi di natura binaria (o opposizioni); cioè, per ogni fonema (qualsiasi lingua esso appartenga), si può univocamente dire se presenta o meno tale tratto distintivo. Tali tratti possono essere scelti in vari modi, ma comunque sempre secondo canoni della fonetica acustica più che della fonetica articolatoria, cioè basandosi sull'analisi strumentale (spettrogrammi, ecc.) dei suoni di una lingua[3]. Una volta individuato l'insieme di tratti giudicato sufficiente a rappresentare l'intero sistema linguistico che si vuole descrivere, una sua rappresentazione alquanto compatta ed esplicativa è data dalla matrice binaria associata a tale sistema. Si tratta di una matrice con una riga per ciascun tratto distintivo e una colonna per ciascun fonema, e con il segno "+" o "-" agli incroci. Alcuni tratti distintivi sono detti *pertinenti* per un fonema, e sono quelli che bastano ad individuarlo univocamente all'interno del sistema linguistico cui appartiene; altri sono detti *ridondanti*, e servono a facilitare la "decodifica" del suono da parte dell'ascoltatore,

[3] Anche se il presupposto su cui si basa la scuola binarista, cioè che ogni realtà linguistica si identifichi tramite una successione di scelte binarie, può apparire più una costruzione ideale che una reale rappresentazione dei processi cognitivi del cervello umano, essa opera una sistematizzazione della fonetica molto utile metodologicamente.

qualora l'informazione connessa con i tratti pertinenti sia degradata. In tale secondo caso, nell'incrocio corrispondente, spesso si lascia la casella vuota o il simbolo viene indicato tra parentesi.

Il binarismo maturo cerca di evitare ad ogni costo i casi di mancata opposizione binaria. In ogni caso, i trenta fonemi italiani (includendo in questo contesto anche le semivocali [*j, w*]) possono specificarsi con undici coppie di tratti distintivi intrinseci (o TDI) [4], come mostrato in tabella 1.3. La media dei TDI per fonema è di 5,8.

FONEMI:	p	b	f	v	t	d	ts	dz	s	z	k	g	c	ɟ	ʃ	m	n	ɲ	l	ʎ	r	i	e	ɛ	a	ɔ	o	u	j	w
1 *Vocalico*	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	+	+	+	+	+	+	+	-	-
2 *Consonantico*	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-	-	-	-	-
3 *Nasale*	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+												
4 *Compatto*	-	-	-	-	-	-	-	-	-	-	+	+	+	+	+				-	+	-	-	-	-	+	-	-	-		
5 *Diffuso*																						+	-	-		-	-	+		
6 *Grave*	+	+	+	+	-	-	-	-	-	-	+	+	-	-	-	+	-	-				-	-	-		+	+	+	-	+
7 *Acuto*																	-	+												
8 *Teso*																							+	-		-	+			
9 *Sonoro*	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-															
10 *Continuo*	-	-	+	+	-	-	-	-	+	+			-		+				+		-									
11 *Stridulo*					-	-	+	+																						

Tab. 1.3 I fonemi italiani e i loro TDI secondo (Muljacic,1972).

1.2.4 Gli elementi prosodici

I fonemi da soli non descrivono completamente i "*suoni*" di una lingua, pertanto vanno considerate anche altre caratteristiche che agiscono su tutta la frase, trasmettendo informazione e completando la descrizione del processo di produzione propriamente detto.

Questi altri elementi prendono nome di **caratteristiche soprasegmentali** e sono molto difficili da definire e formalizzare da un punto di vista linguistico. Alcuni esempi sono il tono, l'accento e l'intonazione. Il **tono** non è presente in tutte le lingue, ma solo in quelle, come il cinese mandarino, in cui modifica il significato lessicale e grammaticale delle parole. Esso interessa l'altezza relativa delle parole e delle sillabe all'interno di una frase. L'**accento** si manifesta nel porre in risalto alcune sillabe rispetto alle altre all'interno di una stessa parola, combinando vari fattori quali l'intensità dell'emissione, la lunghezza (durata nel tempo) e l'altezza dei suoni. L'**intonazione** è una combinazione di alcuni fenomeni di carattere locale, come l'accentazione, la durata e l'intensità dei foni pronunciati, e di alcuni fenomeni di carattere globale, che coinvolgono tutta la frase. Tra questi, la differente modulazione della frequenza fondamentale usata per cambiare significato ad una stessa frase, come avviene ad esempio per differenziare una frase affermativa da una interrogativa o per comunicare le nostre emozioni

[4] Secondo la teoria binarista ci sono due tipi di *tratti distintivi*: *prosodici* e *intrinseci*. I primi si possono avere solo sul nucleo sillabico fonologico e si raggruppano in tre classi: altezza, intensità e durata. I secondi, invece, si possono classificare in dodici opposizioni la cui suddivisione e classificazione viene modificata quasi in ogni nuova opera che presenti questa teoria, anche da parte dello stesso Jakobson. In questa sede, ci occuperemo esclusivamente dei tratti distintivi intrinseci e gli altri non saranno più menzionati nel seguito.

all'ascoltatore. Questi contorni "melodici", chiamati anche contorni prosodici, sono caratteristici di ogni lingua, alla stregua dei suoni e delle regole grammaticali. Essi danno un gran contributo alla comprensione delle frasi e sono un aspetto fondamentale della naturalezza della voce umana.

1.3 IL SUONO E L'ACUSTICA DEL SEGNALE VOCALE

Quel che abbiamo l'abitudine di chiamare suono non è altro, in realtà, che una variazione della pressione atmosferica registrata dal nostro apparato uditivo mediante il timpano. I movimenti di questa membrana sono trasmessi dagli ossicini dell'orecchio medio all'orecchio interno dove, a condizione che si trovino all'interno del campo di sensibilità dell'orecchio[5], essi diventano segnali che vengono ricevuti dal cervello. Queste variazioni della pressione atmosferica hanno la forma di onde che si propagano nell'aria o, in certi casi, attraverso mezzi diversi, liquidi o corpi solidi; l'osso, per esempio, è un buon conduttore delle onde sonore. Le onde si propagano, nell'aria e alla temperatura di 0°, con una velocità di circa 330 metri al secondo, velocità che varia leggermente in rapporto alla pressione e alla temperatura: a 20°, per esempio, la velocità è di 344 metri al secondo. Queste variazioni di pressione sono dovute all'impulso esercitato sulle particelle dell'aria, che vengono smosse dal loro stato di quiete; il fenomeno inizia sempre con uno stimolo meccanico che mette in vibrazione una massa qualunque, un corpo solido, una certa porzione di un corpo gassoso.

L'energia sonora si propaga nello spazio per onde sferiche e quindi decresce con il quadrato della distanza; in ogni caso, quello che si intende con **segnale vocale acustico** è l'andamento temporale della variazione di pressione acustica nella zona limitrofa ad una persona che parla e perciò, con ottima approssimazione, si può considerare trascurabile la perdita di energia e unidimensionale il segnale generato.

Secondo la teoria acustica della **produzione del segnale vocale**, proposta la prima volta da (Fant, 1960) ed ancora oggi generalmente accettata, il segnale acustico viene generato facendo fluire l'aria nella laringe e/o in altre ostruzioni create nel condotto vocale. Le turbolenze che ne scaturiscono danno origine ad un segnale caratterizzato da un ampio contenuto armonico. Questo viene infine modificato tramite l'azione di filtraggio operata dal condotto vocale.

[5] Come si sa, l'uomo non percepisce tutte le vibrazioni come suoni. Nella musica il limite inferiore è di circa 25 Hz (anche se la frequenza più bassa che sia stata percepita è di 11Hz); mentre il limite superiore varia a seconda dell'età e da individuo a individuo. Un bambino può sentire frequenze fino a 20.000 Hz; in età avanzata non si sentono più le frequenze al di sopra di 12.000-13.000 Hz. Tutte le frequenze utilizzate dal linguaggio umano si trovano al disotto di 10.000 Hz.

1.3.1 Lo spettro acustico

E' noto da tempo che l'udito avverte principalmente le differenze di frequenza e quelle di ampiezza di oscillazione, ma non quelle di fase. Pertanto, nella maggioranza dei casi, i fenomeni sonori che differiscono fra loro soltanto per le relazioni di fase tra le loro componenti armoniche, vanno considerati come un solo fenomeno sonoro agli effetti dell'ascolto (Franchina, Marietti, 1994)[6]. Si rivela perciò assai utile una rappresentazione grafica del tipo di quella di fig. 1.10, nella quale compaiono soltanto le frequenze delle varie componenti sinusoidali e le corrispondenti ampiezze. L'insieme delle righe dei grafici come quello di fig. 1.10 prende il nome di **spettro acustico**. La prima riga a sinistra rappresenta l'armonica fondamentale (frequenza f1); le altre righe corrispondono alle frequenze $f_2 = 2f_1$ (seconda armonica), $f_3 = 3f_1$ (terza armonica) ecc.

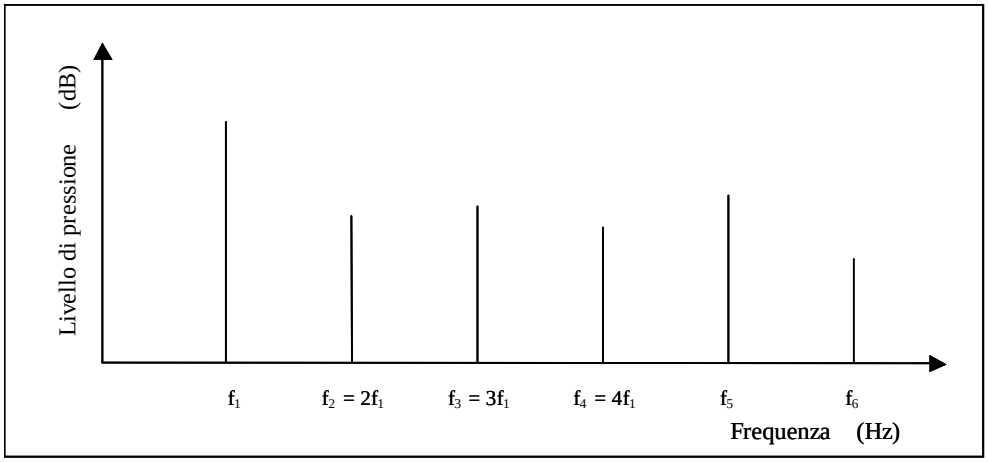


Fig. 1.10 Spettro acustico di un suono complesso.

Queste considerazioni si applicano integralmente soltanto ai fenomeni oscillatori periodici in regime stazionario, condizione quasi mai realizzata nella realtà. Il linguaggio parlato, infatti, è proprio un caso di fenomeno acustico costituito da un gran numero di suoni diversi di breve durata, che si susseguono in rapida successione. Mentre un suono isolato inizia, di regola, con un breve periodo transitorio di attacco ed ha termine con un periodo transitorio di estinzione, nel linguaggio parlato i diversi suoni si succedono senza soluzione di continuità, cosicché il transitorio di estinzione di ciascuno di essi si connette con quello di attacco del suono successivo in modo da costituire quasi un unico transitorio.[7]

Comunque, anche per i fenomeni sonori del tipo ora detto, la rappresentazione mediante lo spettro acustico può riuscire utile, purché si tenga conto in qualche modo dell'evoluzione delle caratteristiche spettrali nel corso del tempo (si ritornerà su quest'argomento nell'ultimo paragrafo).

---

[6] Questa affermazione va fatta comunque con cautela; infatti, alle relazioni di fase sono legati, in modo più o meno evidente, alcuni importanti aspetti della sensazione uditiva, come la identificazione della direzione di provenienza del suono, come il timbro e la stessa intensità soggettiva, che un tempo si pensava ne fossero indipendenti.

[7] Nel linguaggio parlato, i suoni elementari (foni) aventi carattere relativamente stazionario (vocali, semivocali e alcune consonanti quali [*n, m*]) si alternano con altri suoni consonantici aventi il carattere di brevi transitori (esplosive [*p, b, t, d* ecc.])

1.3.2 Suoni sordi e suoni sonori.

Durante la respirazione, il flusso d'aria non incontra ostacoli nel passaggio dalle corde vocali che si trovano in posizione allargata al condotto vocale che è privo di costrizioni. Acusticamente non si percepisce alcun suono. Saranno ora presi in esame i due principali modi di funzionamento dell'apparato di produzione della voce e, a partire da questi, si descriveranno le caratteristiche distintive dei diversi tipi suoni che siamo in grado di produrre e le conseguenti caratteristiche del relativo segnale acustico generato.

Suoni sordi

Le corde vocali possono essere tenute separate tra di loro cosicché l'aria può passare liberamente attraverso la glottide senza far vibrare le corde vocali. Se c'è però la presenza di una costrizione o di un'improvvisa apertura lungo il tratto vocale, si genera l'emissione di suoni chiamati sordi o non vocalizzati, provocati dal moto turbolento del flusso d'aria a valle dell'ostacolo. Acusticamente si percepisce un suono con caratteristiche "rumorose" ad ampio spettro. A seconda della posizione assunta dagli organi mobili del tratto vocale, sono soggetti ad ulteriori classificazioni (per es., sibilanti o plosive, con ulteriore suddivisione a seconda della posizione della costrizione o dell'improvvisa apertura del condotto).

Come esempio di suoni sordi riportiamo le consonanti [p t k f s Σ] in *pane, tondo ,corre, ferro ,sale, scena.*

Suoni sonori

Per la produzione dei suoni sonori, inizialmente le corde vocali sono a contatto l'una con l'altra a causa delle forze presenti e quindi la glottide è chiusa. Quando i polmoni espellono aria, la pressione[8] sotto la glottide aumenta fino a valori che consentono l'allontanamento progressivo delle corde vocali a partire dal basso. Un ulteriore aumento di pressione causa l'apertura della glottide con conseguente passaggio di aria. Le forze elastiche e di altro tipo resistono alla separazione del margine superiore delle corde, ma il flusso d'aria le sovrasta (fig. 1.11).

La legge di Bernoulli asserisce che quando un fluido passa attraverso una strozzatura la pressione ivi presente è minore che nelle sezioni a monte e a valle. Tale riduzione di pressione, accompagnata dalle proprietà elastiche dei tessuti, tende a richiudere le corde vocali. Nel frattempo la pressione sotto la glottide diminuisce anch'essa, dato che la glottide si è aperta per far uscire l'aria. A causa di questi fenomeni, i margini inferiori delle corde vocali cominciano a chiudersi quasi immediatamente, anche se quelli superiori si stanno ancora aprendo.

[8]Generalmente il valore della pressione dell'aria proveniente dai polmoni al livello glottale è di 7 cm H_2O per il parlato normale, 2 cm H_2O per un parlato appena percettibile, e di 20 cm H_2O per un parlato a voce molto alta.

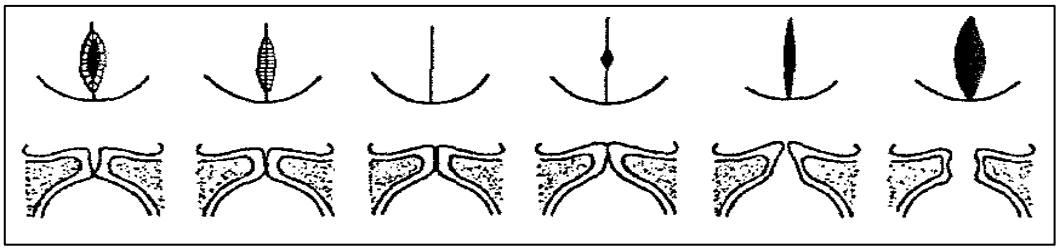


Fig. 1.11 Rappresentazione schematica dello stato di affrontamento delle corde vocali. Parte superiore: sezione longitudinale delle corde vocali (la mancanza del contatto è evidenziata in colore nero); parte inferiore: sezione trasversale.

Questo fatto riduce ulteriormente la forza esercitata dal flusso d'aria e i margini superiori delle corde vocali ritornano allora nella posizione iniziale e chiudono la glottide[9]. A questo punto l'aria torna ad accumularsi al di sotto della glottide e il ciclo così si ripete, alternando le fasi di apertura e di chiusura delle corde vocali[10].

I **suoni sonori** sono dunque quelli prodotti da questo funzionamento delle corde vocali; naturalmente il suono così prodotto può subire modifiche passando attraverso il resto del condotto vocale. Esempio di suoni sonori sono le consonanti [b, d, g, v, z, Z, dZ] di *bene, due, gara ,vetta ,usi, agile* (pronunciato alla toscana); inoltre in italiano sono sempre sonore [m, M, n, ɲ , N, r, l, ʎ] come in *mese, anfora, notte, bagno, àncora, rosa, lupo, figlio*. Le vocali sono tutte suoni sonori.

1.3.3 La frequenza fondamentale o pitch

Il singolo ciclo descritto per i suoni sonori si indica con il nome di **ciclo di fonazione** o **ciclo glottale**, mentre la frequenza con cui vibrano le corde vocali è chiamata **frequenza fondamentale** (F_0) o **pitch**, e la durata del singolo ciclo è detta **periodo di pitch**.

La frequenza fondamentale dell'emissione vocale di un parlatore, il cosiddetto "tono naturale", dipende dalle caratteristiche fisiche delle corde vocali. Varia quindi da parlatore a parlatore e può essere modificata con azioni fisiche, da parte del parlatore, variando il livello di tensione delle corde.

Mediamente il volume d'aria che attraversa il condotto vocale è pari a 1 cm^3/ciclo glottale. Il rapporto tra la durata della fase di apertura delle corde vocali e la durata dell'intero ciclo è variabile tra 0,3 e 0,7. Il valore del rapporto dipende dall'intensità, dalla frequenza con cui vibrano le corde vocali e da quanto è addestrato il soggetto. Infatti, i cantanti professionisti riescono ad ottenere i valori della velocità del volume d'aria minori, ad intensità costante, e a realizzare in questo modo un maggior rendimento nella conversione pressione - suono.

---

[9]Generalmente tra le corde vocali si realizza un contatto, quando si verifica la chiusura della glottide, per uno spessore di circa 2-5 mm.

[10]Il ciclo può anche avere luogo con le corde vocali inizialmente non in contatto tra loro. La pressione dovuta all'effetto di Bernoulli in questo caso fa dapprima avvicinare le corde; la fine della fonazione può avvenire in due modi, a seconda che le corde vocali si rilassino o che vengano forzate a rimanere unite: nel primo caso la vibrazione si esaurisce gradualmente e le corde vocali non si toccano per gli ultimi cicli; nel secondo la vibrazione cessa immediatamente e si ha chiusura glottale anche nell'ultimo ciclo.

Le corde vocali non imprimono quindi energia all'aria vibrando come le corde di un violino, ma aprendo e chiudendo la glottide, creando "sbuffi" d'aria nell'apparato vocale. L'improvvisa cessazione del flusso d'aria a causa del rapido accostarsi delle corde vocali produce una vibrazione acustica che risuona nel condotto vocale. Tale meccanismo è simile a quello che dà origine al suono prodotto sbattendo le mani. L'istante in cui avviene la completa chiusura della glottide è chiamato **istante di epoch**. Anche se è all'istante di *epoch* che viene prodotto il maggior contributo all'energia sonora responsabile dell'emissione della voce, un altro contributo di minor entità viene dall'aprirsi delle corde vocali che si verifica più lentamente della loro chiusura (Strube, 1974).

L'intensità vocale, o volume, dipende da quanta energia viene impartita dalle vibrazioni delle corde vocali all'aria nell'apparato vocale. Quando la pressione dell'aria aumenta, l'ampiezza delle vibrazioni cresce perché le corde vocali si allargano maggiormente e si richiudono più bruscamente; di conseguenza, durante ciascun ciclo di fonazione, il flusso d'aria attraverso la laringe si interrompe più nettamente e l'intensità del suono prodotto cresce.

L'andamento nel tempo della velocità del volume d'aria, per una voce di intensità normale, è un segnale quasi periodico di forma approssimativamente triangolare caratterizzata da due istanti di discontinuità, uno iniziale ed uno finale, che rappresentano rispettivamente gli istanti di apertura glottale e di *epoch*[11]. Data la natura periodica, il suo spettro è a righe, le cui componenti periodiche sono multipli interi della frequenza fondamentale. L'inviluppo dello spettro presenta un'attenuazione nelle alte frequenze di circa 12dB/ottava, anche se vi possono essere grandi differenze nelle altezze delle armoniche da soggetto a soggetto e, per lo stesso soggetto, passando da un periodo di *pitch* all'altro. Mediamente, per i soggetti che leggono un testo, l'intervallo di variazione della frequenza fondamentale di rado supera un'ottava nel corso della lettura. Poiché gli uomini hanno corde vocali più lunghe (tra i 20 e 25mm) delle donne e dei bambini (tra i 15 e 20 mm), il loro *pitch* è generalmente più basso. In tabella 1.4 sono illustrate le frequenze fondamentali che la voce può avere nel corso del parlato normale (nel caso del canto la frequenza fondamentale può variare approssimativamente tra i 40Hz e i 1800Hz).

Soggetto	$F_{o\ minima}$ (Hz)	$F_{o\ media}$ (Hz)	$F_{o\ massima}$ (Hz)
Uomini	50	125	200
Donne	150	225	350
Bambini	200	300	500

Tab. 1.4 Valori della frequenza fondamentale minima, media e massima per soggetti adulti maschili, femminili e per bambini (M.I.T., 1986)

Comunque la frequenza fondamentale normalmente può variare al massimo dell'1%/ms, il che corrisponde, ad esempio, ad un cambiamento del 2% per periodi di pitch adiacenti per F_o=500 Hz e del

[11]Le forze aerodinamiche responsabili delle oscillazioni delle corde vocali sono influenzate dal tratto sopra - glottale. Ciò causa un leggero ritardo dell'andamento nel tempo della velocità del volume d'aria rispetto all'andamento dell'aria nella glottide.

20% per F_o=50 Hz. Chiaramente la frequenza di pitch può essere modificata dal parlatore agendo sul livello di tensione delle corde vocali.

1.3.4 Frequenze Formanti

I suoni sonori sono caratterizzati, oltre che dalla F_0 anche dalle frequenze formanti. Vediamo, come abbiamo fatto nel precedente paragrafo per la F_0 , qual è l'origine fisica delle formanti.

Un risonatore acustico è un sistema fisico che presenta la capacità di alterare la natura di un suono che lo attraversa. Più precisamente nel passaggio di un segnale acustico nel risonatore, alcune frequenze componenti sono attenuate, altre, nelle regioni di risonanza, vengono invece amplificate e irradiate quindi con maggior ampiezza. Per quanto riguarda la voce, le frequenze di risonanza sono dette **frequenze formanti**, e sono determinate dalla forma del condotto vocale che dipende dalla posizione degli organi mobili, dall'età e dal sesso dell'individuo. Donne e bambini hanno un apparato vocale più breve degli uomini e di conseguenza i valori delle frequenze formanti saranno più elevati[12]. Ad esempio, se si schematizza il condotto vocale in posizione "neutrale", come per la vocale /*u*/ nella parola inglese "*but*", assimilandolo ad un tubo uniforme senza perdite chiuso ad un'estremità (la glottide) e aperto all'altra (le labbra), le frequenze di risonanza *v* delle onde stazionarie che vi si generano assumono i valori dati dall'espressione:

$$v = \frac{c}{4l}(2n+1) \qquad \text{n = 1, 2, 3, ...} \tag{1.2}$$

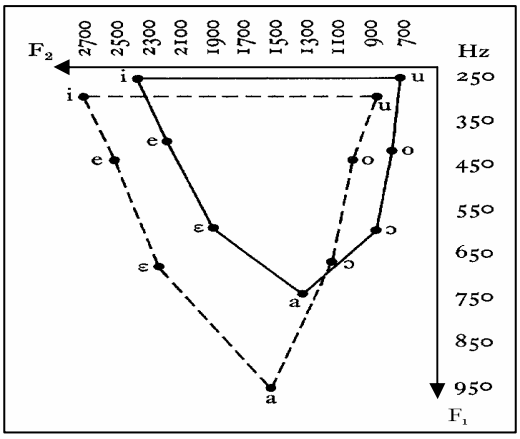


Fig. 1.12 Medie delle prime due formanti dei sette vocoidi tonici italiani: voci maschili (linea continua) e femminili (linea tratteggiata) sovrapposte. (Canepari, 1979).

dove *l* è la lunghezza del condotto vocale (mediamente 17 cm) e *c* la velocità delle onde elastiche nell'aria (circa 340 m/s). Per questi valori si hanno i seguenti valori di *v*: 500 Hz, 1500 Hz, 2500 Hz, ecc.

[12]Un'altra causa da cui dipende la lunghezza e la forma del condotto vocale, e quindi le caratteristiche delle frequenze formanti, è la frequenza fondamentale usata durante l'eloquio. Infatti, i suoi cambiamenti causano un abbassamento od un sollevamento dello scheletro cartilagineo della laringe, provocando perciò una modifica delle lunghezza del condotto vocale.

Tali valori corrispondono ai valori delle frequenze formanti. Per suoni diversi il condotto vocale assume configurazioni differenti, quindi si hanno valori differenti delle frequenze formanti, ciascuno caratteristico di ogni suono.

Vediamo infine più nel dettaglio come il timbro dei vocoidi dipende dalle singole formanti. Per i vocoidi sono fondamentali le prime due formanti (F1 e F2) contando dal basso dopo la fondamentale. Le formanti superiori servono soprattutto per le caratteristiche individuali della voce. Per i vocoidi F1 è bassa (250 Hz circa, per una voce maschile) se sono alti come [i] e [u], alta (intorno ai 750/800 Hz) se sono bassi come [a]. La F1 si sposta gradualmente tra questi due estremi, inversamente all'elevazione della lingua. Invece F2 è determinata dalla lunghezza della cavità orale: più essa è lunga, più F2 è bassa; se poi s'arrotondano le labbra ,come per la [u], la cavità si allunga ulteriormente facendo abbassare F2 ancora di più.

Nella figura 1.12, sono mostrate le medie delle prime due formanti delle vocali italiane così come riportato dal Canepari.

1.3.5 Caratteristiche acustiche generali della voce emessa

La conoscenza delle principali caratteristiche acustiche del linguaggio parlato è un dato preliminare indispensabile nella tecnica delle telecomunicazioni. Menzioniamo brevemente alcuni risultati medi sperimentali.

- La potenza vocale media a lungo termine[13] di un parlatore è dell'ordine di 20 μW con un livello di voce moderato (68 dB è il corrispondente livello di pressione acustica alla distanza di un metro). La massima escursione è compresa fra pochi μW (voce bassa) e oltre 1mW (voce urlata), corrispondente ad un intervallo di circa 24dB;
- Lo spettro acustico medio a lungo termine mostra che i livelli di voce più elevati si hanno nella banda 200÷400 Hz, mentre per frequenze più elevate il livello spettrale di voce decresce di circa 10 dB per ottava.
- La dinamica della voce è di circa 40 dB nel caso di un discorso tenuto a un livello normale.
- Il ritmo di fonazione medio, ossia la rapidità con la quale si succedono gli elementi fonetici nel discorso, si aggira intorno agli 8÷10 fonemi per secondo.

1.3.6 Caratteristiche acustiche della sensazione uditiva

Si espongono ora alcune caratteristiche dell'apparato percettivo umano. Tali caratteristiche devono essere tenute sempre presenti nel formulare conclusioni, per non incorrere nell'errore di dare importanza ad aspetti colti visivamente sullo spettrogramma, che però l'orecchio percepisce diversamente (o per nulla!) e che quindi non hanno rilevanza percettiva.

All'interno dell'orecchio vi sono una molteplicità di fibre nervose sensibili alla pressione dell'aria, e in grado di trasformare le onde sonore del segnale acustico in segnale elettrico inviato al cervello. Tali fibre

[13] Per media a lungo termine si intende quella che si riferisce a un intervallo di tempo comprendente parecchi fonemi, senza pause di silenzio tra frasi diverse.

sono in genere sensibili ad una frequenza ben precisa, detta **frequenza caratteristica**, con una banda passante di 100÷150 Hz; fibre vicine hanno frequenze caratteristiche vicine. Ma la caratteristica più importante da rilevare è che il loro funzionamento non è perfettamente lineare, nel senso che componenti a frequenza vicina vengono percepite dando luogo a componenti spurie con frequenza di intermodulazione tra le due originali. Ciò dà luogo al cosiddetto *effetto centro di gravità spettrale*, cioè due formanti a distanza inferiore di 300 Hz vengono percepite come una sola, avente frequenza intermedia tra le due (e spostata verso quella a maggior contenuto energetico). Per compensare il fenomeno della non linearità è stata proposta una scala alternativa a quella delle frequenze per descrivere il segnale vocale, la cui unità di misura è il *Bark*, e la formula di conversione è la seguente:

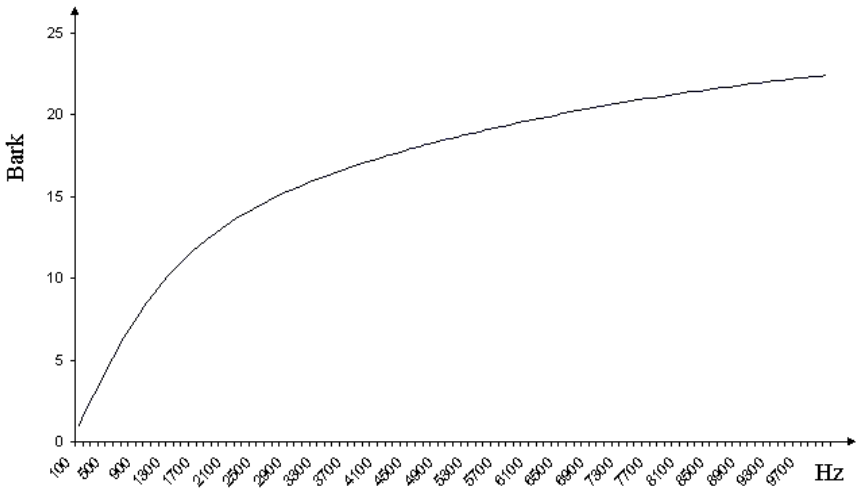
$$Bark = 13 \cdot \text{arctg}(0.76 \cdot f_{khz}) + 3.5 \cdot \text{arctg}\left(\frac{f_{khz}}{7.5}\right)^2 \qquad (1.3)$$

L'effetto della trasformazione è una compressione dei valori in frequenza (5kHz = 18.54B), con una maggiore conformità alle caratteristiche percettive non lineari dell'orecchio umano come si vede in figura 1.13a.

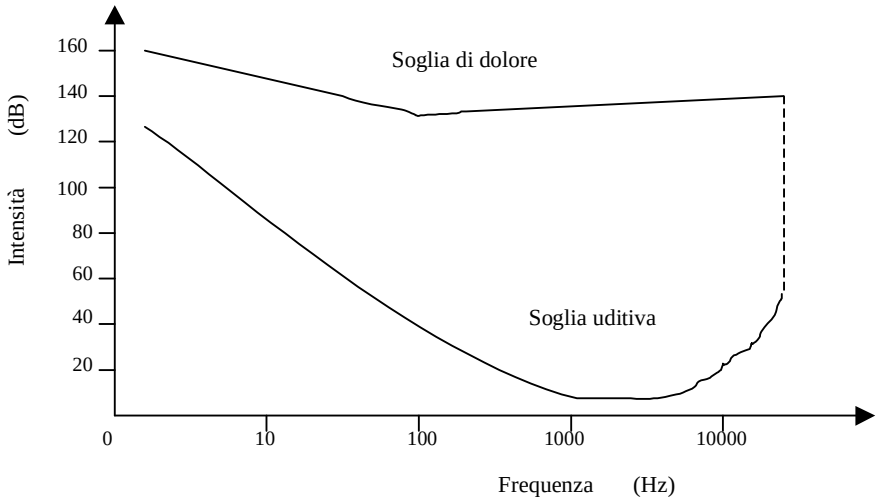
Un altro fenomeno da tenere presente è l'*adattamento*, per cui la risposta ad un suono stazionario è stazionaria per un po', per poi decadere con una costante di decadimento τ di circa 30 ms. Tale caratteristica suggerisce l'idea che il cervello preferisce individuare l'informazione nelle variazioni del segnale in arrivo. Conseguenza dell'adattamento è un altro fenomeno simile, detto del *mascheramento posteriore*, per cui l'orecchio sottoposto ad un suono di test prolungato, poi ad una pausa e poi ad una breve riproposizione del suono, fornisce stavolta una risposta alquanto debole.

Facendo riferimento al caso più semplice, e cioè a quello dei toni puri in regime stazionario, si possono inoltre individuare le seguenti caratteristiche:

- **Altezza tonale**, caratteristica per la quale i suoni si distinguono in più o meno gravi o acuti. E' legata essenzialmente alla frequenza dell'oscillazione;
- **Intensità soggettiva**. E' legata in modo essenziale sia al livello di pressione dell'onda sinusoidale, sia alla sua frequenza. Il conseguente comportamento dell'udito umano per i suoni puri è illustrato dall'audiogramma normale ottenuto costruendo sperimentalmente, per diversi valori di intensità, le cosiddette curve isofoniche (ovvero di isointensità soggettiva). L'andamento di queste curve (Raccomandazione Internazionale ISO/R226) mostra che, perché una vibrazione sia percepita come suono, bisogna che raggiunga un certo valore minimo di intensità (soglia inferiore di udibilità); al contrario esiste un valore massimo di tollerabilità dell'orecchio, sorpassato il quale si ha una sensazione di sofferenza (soglia del dolore). Inoltre, la sensibilità dell'udito è maggiore per le frequenze acustiche medie (fra qualche centinaio e qualche migliaio di Hz) che ai due estremi della banda acustica, e che nel campo dei toni gravi molto intensi la sensibilità dell'udito cresce con la pressione acustica più rapidamente che nella restante parte dell'area di udibilità. Un'idea dell'andamento di tali curve è dato in fig. 1.13b.
- **Timbro**, caratteristica per la quale suoni di stessa altezza e stessa intensità possono essere assai spesso facilmente distinti (ad esempio una stessa nota musicale emessa con uguale intensità da due diversi strumenti musicali). E' legata principalmente alla struttura spettrale del suono complesso ma anche ad altri parametri fra cui l'intensità globale.



a)



b)

Fig. 1.13 a) Conversione di scala Hz/Bark b)Il campo di sensibilità dell'orecchio umano alle vibrazioni.

1.4 L'INGEGNERIA: IL SEGNALE VOCALE ELETTRICO E LA SUA ELABORAZIONE

L'elaborazione analogica, e ancor più quella digitale del segnale vocale elettrico hanno portato grandi cambiamenti nella nostra vita quotidiana: si pensi a tutti i sistemi di telefonia e di comunicazione vocale, ai riconoscitori vocali che ormai sono a corredo di molti apparecchi hi-tech e dei computer (soprattutto negli USA), ai sintetizzatori vocali. Per questo motivo, ma anche per rendere più chiara la descrizione del lavoro svolto per la presente tesi, trattiamo in questo paragrafo i fondamenti dell'approccio ingegneristico al segnale vocale.

1.4.1 I sistemi numerici di elaborazione del segnale

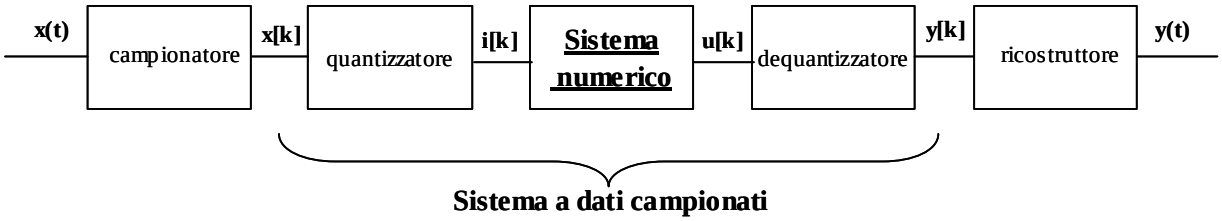


Fig. 1.14 Struttura di elaborazione per segnali unidimensionali.

Nel terzo paragrafo, si è definito il segnale vocale acustico come l'andamento temporale della variazione di pressione acustica nella zona limitrofa al parlatore. Questo segnale, per essere elaborato, viene trasdotto da un microfono, che lo trasforma in un segnale elettrico chiamato anch'esso vocale. La qualità del segnale riprodotto dipende quindi, in primo luogo, dalle caratteristiche del microfono. In pratica, il trasduttore di sorgente si limita a generare un segnale elettrico in qualche modo proporzionale a quello prodotto dalla sorgente. In questa sede, si farà conto che non ci sia perdita di segnale né degradazione di esso nel passaggio dalla forma d'onda acustica a quella elettrica (trasduttore ideale); nel seguito, ci si riferirà indifferentemente all'una o all'altra forma con il nome generico di segnale vocale.

La figura 1.14 rappresenta un generico sistema di comunicazione numerico. Nel caso più generale (e certamente per la voce), il segnale completamente numerico (sia in tempo che in ampiezza) i[k], che entra nel sistema di elaborazione numerica (ad esempio un calcolatore o un DSP), deriva dal rispettivo segnale analogico sul quale si è operato un campionamento e una quantizzazione. Nel passaggio dal sistema analogico a quello a dati campionati, sotto l'ipotesi che x(t) sia limitato in banda e che si siano rispettate le condizioni del teorema del campionamento[14], non c'è alcuna degradazione (almeno teorica) del segnale. I

[14] Per rappresentare un segnale limitato in banda con banda pari a W, è sufficiente estrarre i campioni del segnale alla Frequenza di Nyquist pari a $F_N = 2W$ (quindi con un periodo $T = 1/2W$). Questa è la minima frequenza richiesta per ricostruire correttamente il segnale, valida, ovviamente, solo per un campionamento ideale.

segnali limitati nel tempo hanno banda infinita e quindi si ha comunque una perdita di qualità nel segnale campionato. In pratica si sceglie la frequenza di campionamento a seconda della banda di frequenze che contiene informazioni importanti per la specifica applicazione per cui è progettato il sistema di elaborazione. La qualità del segnale riprodotto dipende quindi anche dalla frequenza di campionamento scelta[15]. Il segnale vocale è sempre di banda base e può ritenersi membro di un processo aleatorio spesso assumibile come stazionario ed ergodico (caratterizzato quindi da proprietà comuni a tutti i membri del processo quali larghezza di banda, spettro di densità di potenza, ecc.). Inoltre, a questa categoria di segnali è applicabile il teorema del campionamento (P. Mandarini, 1990), e dunque ciascuno di essi è rappresentabile completamente attraverso la sequenza dei suoi campioni, presi a distanza temporale opportuna (si ricorda che, al massimo, la voce umana copre solo i primi 10kHz della banda acustica).

Per quanto riguarda il passaggio dal sistema a dati campionati a quello completamente numerico, è inevitabile una degradazione del segnale già in linea teorica (il cosiddetto rumore di quantizzazione). Questo è causato dal dover necessariamente usare un numero finito di registri di memorizzazione o una lunghezza di parola finita, rispettivamente per un'elaborazione via hardware o via software. Ciò nonostante, l'elaborazione numerica presenta dei vantaggi notevolissimi e, addirittura, a parità di costi, spesso superiore anche come qualità a quella puramente analogica.

1.4.2 Un modello per la generazione del segnale vocale

Alla base dell'approccio ingegneristico allo studio di fenomeni fisici c'è spesso la creazione di un modello del sistema. Riportiamo in figura 1.15 lo schema a blocchi dell'apparato fonatorio umano generalmente accettato. Il filtro digitale H(z) tiene conto dell'influenza esercitata dall'atteggiamento assunto dagli organi fonatori ed è in genere una funzione con soli poli (anche se tale ipotesi non è verificata ad esempio nella produzione di suoni nasali). In pratica, tale influenza corrisponde a modificare le frequenze di risonanza delle cavità del tratto vocale, che hanno l'effetto di far assumere allo spettro del segnale uscente una forma particolare, esaltandone energeticamente alcune bande di frequenza rispetto ad altre. L'amplificatore pilotato dal parametro G_0 tiene conto del livello energetico del segnale.

[15] Ad esempio, nelle comunicazioni, spesso, il segnale vocale deve essere trasdotto, trasmesso e riprodotto, al solo scopo di rendere completamente riconoscibile il significato della locuzione e l'identità del parlatore, e ciò definisce una particolare esigenza di qualità che qualifica il segnale vocale come telefonico (nella pratica, quello con banda compresa tra 300 e 3400 Hz).

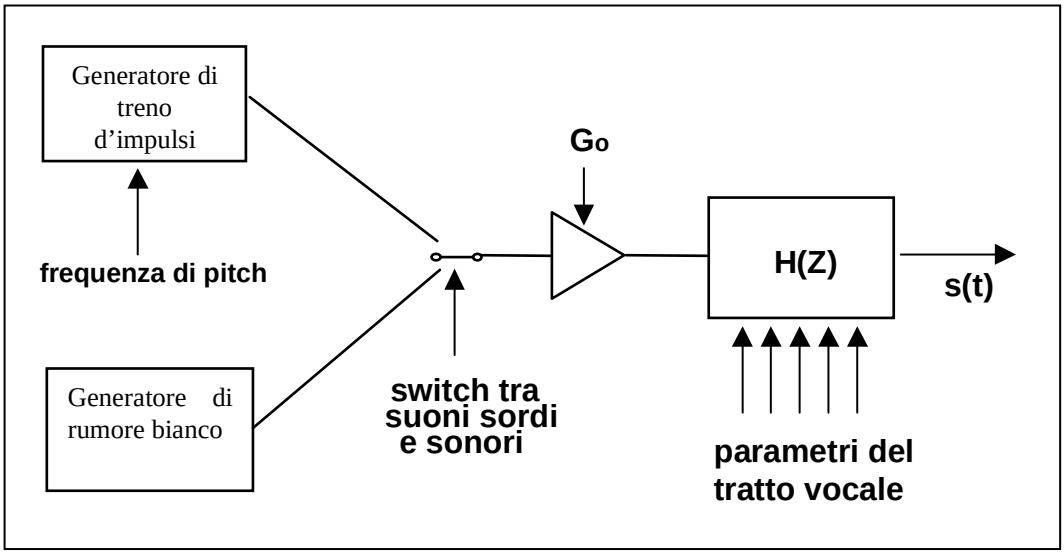


Fig. 1.15 Modello tempo-discreto dell'apparato fonatorio per la generazione di parlato.

1.4.3 Sottocampionamento e sovracampionamento

Si è ritenuto utile riportare le due tecniche di base utilizzate per modificare la frequenza di campionamento. Esse, infatti, si sono rivelate essenziali (vedi paragrafo 3.2.4) per digitalizzare la base dati usata per l'analisi delle consonanti nasali.

Si definisce **sottocampionamento** l'operazione atta a ridurre il tasso di campionamento di un segnale. La tecnica prevede, ovviamente, una *decimazione* dei campioni nel tempo. Questo comporta, in frequenza, che la banda del segnale originale dopo sottocampionamento, aumenti proporzionalmente col fattore di decimazione M.

Le formule (1.4) esprimono il legame tra una sequenza e il rispettivo segnale analogico dal quale è derivata con campionamento di periodo T. Come si vede, lo spettro della sequenza è periodico di periodo 2π e l'asse delle "frequenze analogiche" Ω si trasforma nel nuovo asse delle "frequenze numeriche" secondo la relazione $\omega = \Omega T$, quindi, come mostrato in figura 1.16, perché la sequenza non sia affetta da *aliasing* occorre che sia $\Omega_0 \leq \pi/T$ (teorema del campionamento).

$$x[n] = x_a(nT) \qquad X(e^{j\omega}) = \frac{1}{T}\sum_{k=-\infty}^{+\infty} X_a\left(j\frac{\omega}{T} - j2\pi\frac{k}{T}\right) \tag{1.4}$$

$$x_d[n] = x[M \cdot n] = x_a(M \cdot nT) \qquad X_d(e^{j\omega}) = \frac{1}{M}\sum_{i=0}^{M-1} X\left(e^{j\left(\frac{\omega}{M} - i\frac{2\pi}{M}\right)}\right) \tag{1.5}$$

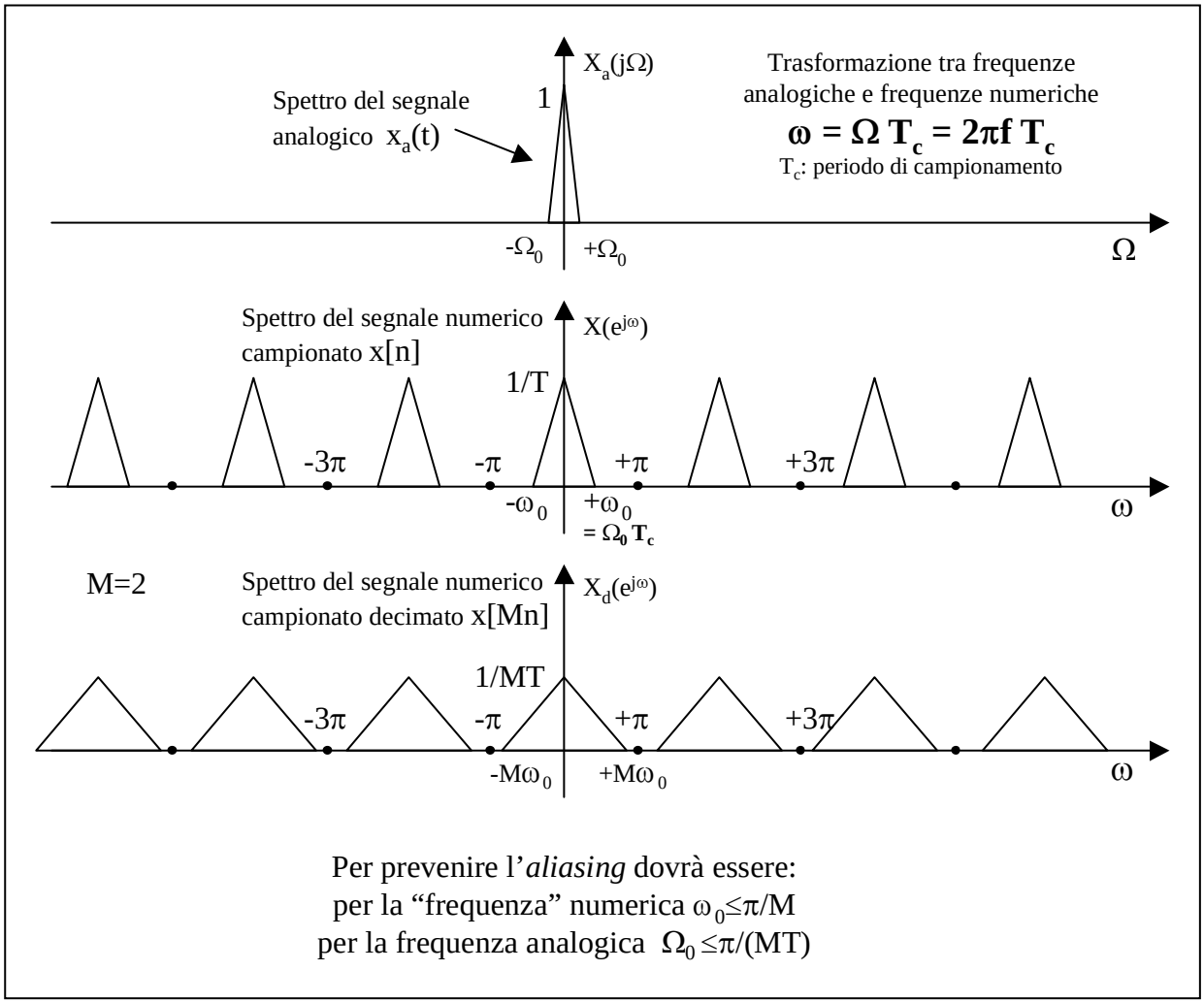


Fig. 1.16 Tecnica di sottocampionamento: legame (in frequenza) tra una sequenza, il segnale analogico dal quale è stata campionata e tra la stessa sequenza e decimata di un fattore M.

Le formule (1.5) esprimono, invece, il legame tra la stessa sequenza di prima e quella decimata di un fattore M (costruita cioè dalla prima prendendo un solo campione ogni M). Anche in questo caso il legame è chiaro: lo spettro originale viene espanso sull'asse ω di un fattore M. Perché non sussista aliasing la condizione è stavolta $\omega_0 \leq \pi/M$. Si veda la figura 1.16 a tal proposito.

L'operazione inversa della decimazione è l'*interpolazione*, detto anche **sovracampionamento**, che prevede l'inserzione di (L-1) campioni fittizi pari a 0 tra ciascuna coppia di campioni consecutivi della sequenza. Nel dominio della frequenza l'effetto di questa operazione consiste nel distanziare le repliche dello spettro a distanza L x (distanza originaria).

1.4.4 Lo studio nel dominio della frequenza: l'analisi spettrale

Il segnale vocale può essere utilmente studiato con vari approcci, per dedurne le caratteristiche e associarle ai vari fonemi e addirittura ai vari modi di articolazione. Le tecniche più usate sono quelle che prevedono lo studio nel dominio del tempo o nel dominio della frequenza, effettuando eventualmente elaborazioni ulteriori tese a evidenziare alcune proprietà particolari del segnale.

Come noto la trasformata di Fourier di un segnale s(t) è detta *spettro* del segnale, per cui per quanto riguarda lo studio in frequenza si parla in genere di **analisi spettrale** del segnale. La tecnica seguita in questa sede prevede il campionamento del segnale s(t) e il suo studio tramite elaborazioni di tipo numerico (Trasformata discreta di Fourier, DFT). Ricordiamo brevemente l'espressione matematica della DFT, che prevede, usando una finestra (detta normalmente **frame**) di N campioni del segnale s(t), il calcolo di N campioni in frequenza della trasformata di Fourier F[s(t)], nella banda propria del segnale:

$$DFT(s(nT)) = S(k) = \sum_{n=0}^{N-1} s(nT) e^{-\frac{j2\pi knT}{N}}$$

$$s(nT) = \frac{1}{N} \sum_{k=0}^{N-1} S(k) e^{\frac{j2\pi knT}{N}} \qquad \text{per k (n), da 0 a (N-1)} \qquad (1.6)$$

dove T è l'inverso della frequenza di campionamento. Se la banda del segnale s(t) è B e si è scelta una frequenza di campionamento T=1/2B, gli S(k) sono i campioni della sua trasformata continua di Fourier a distanza B/N.

Per l'analisi del segnale vocale, nella scelta della lunghezza del frame, occorre tener presente che bisogna eseguire un'analisi in intervalli di tempo sufficientemente brevi, da poter associare le caratteristiche del segnale a quelle del condotto vocale, ma sufficientemente lunghi perché le caratteristiche del segnale possano essere considerate stazionarie in tale intervallo, con sufficiente approssimazione. Si deve inoltre tenere presente il principio generale in base al quale la risoluzione in frequenza è tanto migliore quanto più grande è il frame scelto. Se, infatti, come detto, N è il numero di campioni di un frame analizzato e se indichiamo con f_c la frequenza di campionamento utilizzata per il segnale, la risoluzione in frequenza che si ha quando vengono calcolati gli spettri è data dalla formula:

$$Risoluzione\_frequenziale = \frac{f_c}{N} \qquad (1.7)$$

Per trovare un compromesso tra le due esigenze opposte di località dello spettro e di risoluzione in frequenza, si usa un **fattore di sovrapposizione S** tra frame adiacenti non nullo, ma compreso tra 0 e 1.

In pratica, ogni N·(1-S) campioni è analizzata una finestra di segnale lunga N campioni[16]. I parametri su cui si può agire sono quindi la *dimensione dei frame* N che determina la *risoluzione in frequenza* dello spettro (tanti campioni vi sono in un frame e tanti campioni vi sono nella DFT di quel frame) ed il *fattore di sovrapposizione* S dei frame che influisce sulla *risoluzione temporale* dello spettrogramma (più i frame sono sovrapposti, più frame vi saranno in un segnale lungo T).

Nell'analisi del segnale vocale risulta particolarmente utile l'osservazione dell'evoluzione temporale delle caratteristiche spettrali di un segnale. Ciò è possibile tramite lo **spettrogramma** ottenuto affiancando gli spettri locali di finestre contigue. Nello spettrogramma, la grandezza riportata in ordinata è la *frequenza*, sulle ascisse è riportato il *tempo* (tutti i vari frame analizzati) mentre l'*ampiezza* dello spettro è data dall'annerimento, maggiore o minore, sul disegno. Se N, ampiezza della finestra, è "grande" (128 o 256 campioni) lo *spettrogramma* viene detto **narrow band**, in quanto il passo di approssimazione della trasformata di Fourier è piccolo (circa 40 o circa 20Hz rispettivamente, per B=5kHz), se "breve" (16 o 32 campioni) viene invece detto **wide band**, avendo passo di approssimazione grande (circa 300 o circa 150Hz per B=5kHz). Chiaramente per i motivi precedentemente illustrati, lo spettrogramma wide band riesce a mostrare caratteristiche di breve durata del segnale al prezzo di una minore accuratezza nel campionamento in frequenza. Lo spettro narrow band è quindi particolarmente adatto per l'analisi dei segmenti fonici; lo spettro wide band, invece, è molto utile nello studio dei contoidi e nell'analisi delle caratteristiche individuali della voce, come il tono e l'intonazione.

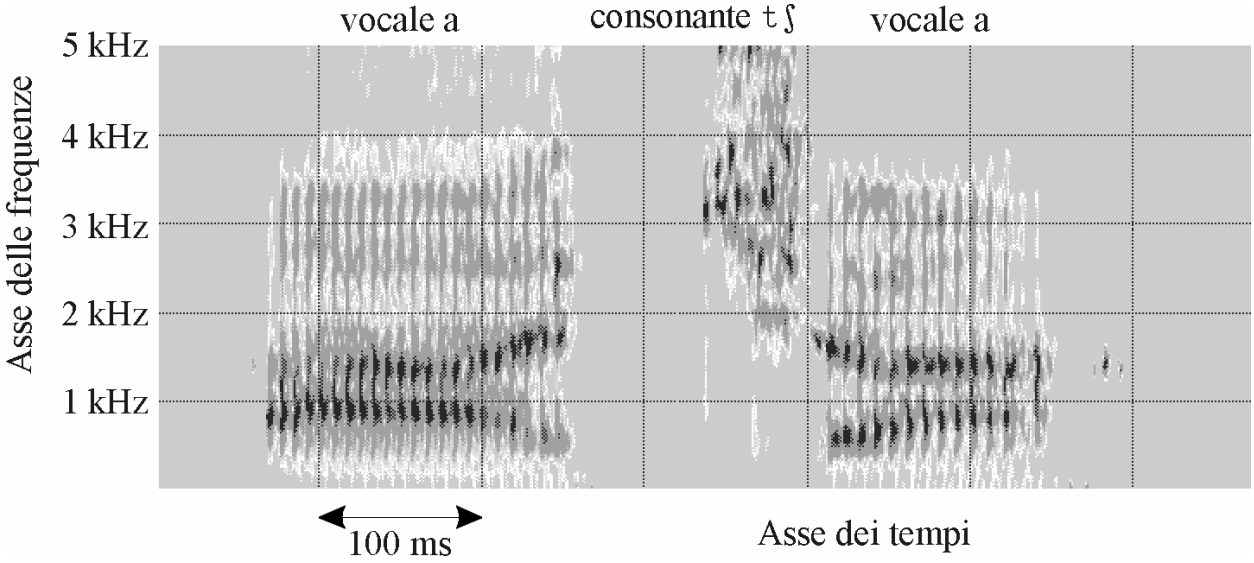


Fig. 1.17 Esempio di spettrogramma.

Esistono vari approcci nell'analisi del segnale vocale. Oltre l'analisi spettrale attuata mediante DFT, altre tipiche elaborazioni possibili sono la costruzione della funzione di *zero crossing*, per evidenziare i

[16] Se, ad esempio, il numero di campioni per finestra è pari a 128 e il fattore di sovrapposizione è 3/4, il risultato è che ogni 32 campioni (128·1/4) viene analizzato un tratto di segnale lungo 128.

momenti di silenzio fonetico, gli algoritmi di *pitch tracking*, il calcolo dell'energia locale del segnale, l'analisi LPC (*Linear Predictive Coding*), che aiuta molto nell'individuazione delle formanti, l'autocorrelazione, l'estrazione dei parametri statistici classici (covarianza, valore medio, ...). Sulle analisi tramite FFT e LPC in particolare conviene soffermarsi, visto l'uso estensivo che se ne farà nel seguito.

Analisi con la FFT

L'analisi in frequenza del segnale vocale può essere condotta eseguendo direttamente la FFT delle sequenze di campioni contenuti in ogni frame. Poiché la FFT di una sequenza di lunghezza "m" è la trasformata di Fourier del segnale periodico di periodo "m", ottenuto replicando la sequenza di durata finita (figura 1.18), essa conterrà delle componenti in frequenza spurie, non legate al segnale originario, ma semplicemente introdotte dalle brusche variazioni di ampiezza dovute alle repliche della sequenza di durata "m".

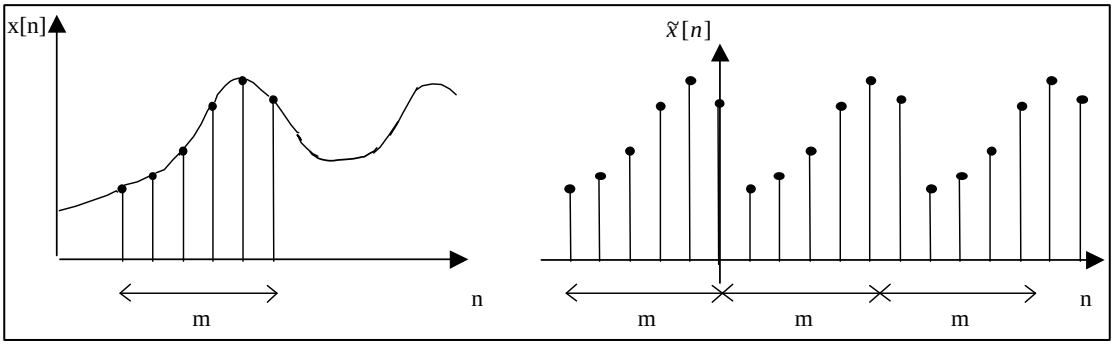


Fig. 1.18 Sequenza x[n] di durata finita "m" e corrispondente sequenza periodica di periodo "m".

Per prevenire la formazione di queste frequenze spurie, il tratto di segnale contenuto nel frame di cui si vuole calcolare la DFT, viene modulato con un'opportuna finestra, che attenua il segnale agli estremi dell'intervallo. La funzione di modulazione impiegata è la **finestra di Hamming** (o del coseno rialzato), la cui espressione è:

$$w(n) = 0.54 - 0.46 \cdot \cos\left(\frac{2\pi n}{N-1}\right) \qquad \text{con } 0 \le n \le N\text{-}1 \qquad (1.8)$$

che moltiplicata per il tratto di segnale contenuto nel frame, ne preserva la parte centrale. Con l'impiego della finestratura, si rende ancora più necessario lo slittamento di ciascun frame di almeno N/2 campioni (se N è il numero di campioni per frame), per non perdere le informazioni del segnale agli estremi del frame stesso. Infatti, con questa accortezza, i campioni che si trovano attenuati agli estremi di un frame, risulteranno praticamente inalterati all'interno di quelli immediatamente precedente e successivo.

Un'ulteriore operazione da compiere, prima di visualizzare la DFT del segnale, è quella di **preenfasi**, ottenuta con un filtro la cui risposta impulsiva è: $h(n) = \delta(n) - \alpha\, \delta(n-1)$. L'effetto che si desidera ottenere è quello di una *enfatizzazione* dello spettro tramite una trasformazione tesa ad esaltare l'importanza del contenuto energetico in alta frequenza, altrimenti poco visibile graficamente (ma non per questo meno importante dal punto di vista percettivo). Con un valore di α pari a 0.95, come comunemente si usa, le basse frequenze vengono attenuate notevolmente (fino a oltre 20 dB), mentre al limite della banda di lavoro si ha un'amplificazione di circa 6 dB.

L'andamento del modulo della funzione di trasferimento del filtro di preenfasi è riportato in figura 1.19.

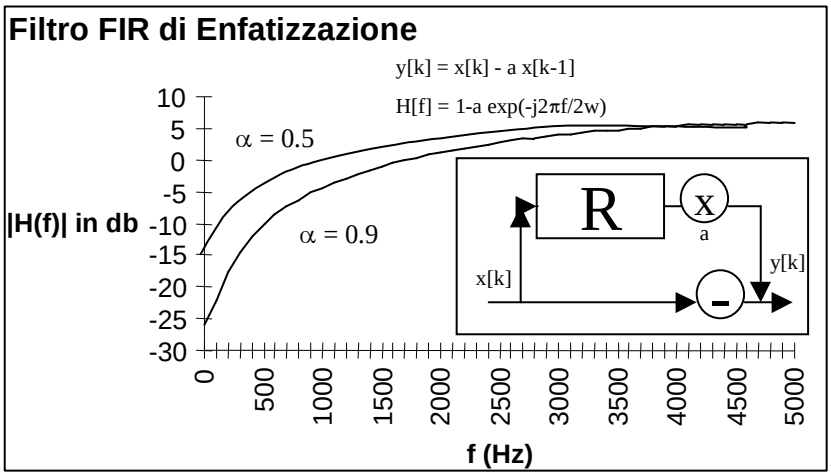


Fig. 1.19 Funzione di trasferimento del filtro di preenfasi.

Nella figura 1.20 è riportata, come esempio, la rappresentazione di un frame di segnale sinusoidale modulato con la finestra di Hamming. La grandezza nella parte inferiore della figura, è ovviamente il modulo quadrato, previa preenfasi, della DFT del segnale in questione; mentre lo spettro di un segnale perfettamente sinusoidale è formato da un'unica riga, lo spettro della sinusoide "finestrata" ha una banda più larga.

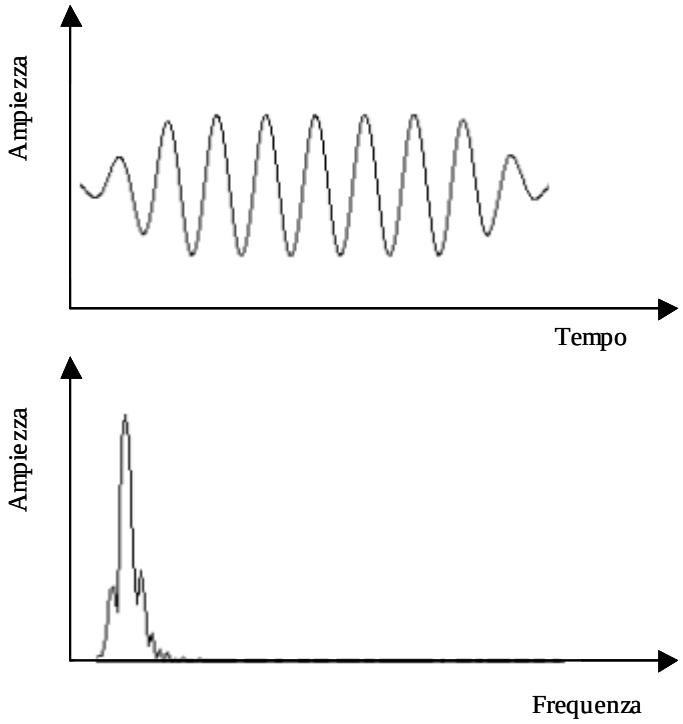


Fig. 1.20 Sinusoide finestrata secondo Hamming e sua DFT preenfatizzata.

Analisi LPC

Una delle più efficaci tecniche di analisi del segnale vocale è quella della predizione lineare. L'importanza di tale metodo risiede nella capacità di fornire una stima accurata dei parametri del tratto vocale e delle frequenze formanti, e nella sua velocità di calcolo.

Il problema fondamentale della predizione lineare è quello di esprimere il generico campione del segnale come una combinazione lineare dei "p" campioni immediatamente precedenti:

$$\tilde{s}(n) = \sum_{j=1}^{p} a_j s(n-j) \tag{1.9}$$

I coefficienti incogniti a_j della combinazione lineare prendono nome di **coefficienti di predizione**. Il problema della determinazione dei coefficienti incogniti è affrontato con il criterio di minimizzazione dell'errore quadratico medio di predizione.

Tale errore è definito come

$$E_{n,m} = \sum_{i=0}^{m} e_n^2(i) = \sum_{i=0}^{m} \left(s(n+i) - \tilde{s}(n+i)\right)^2 \tag{1.10}$$

dove n è il primo campione della finestra di ampiezza m. Sostituendo nella precedente relazione, l'espressione del campione predetto, si ottiene

$$E_{n,m} = \sum_{i=0}^{m} \left(s(n+i) - \sum_{j=1}^{p} a_j s(n+i-j) \right)^2 \tag{1.11}$$

La minimizzazione dell'errore quadratico medio, si ottiene imponendo uguali a zero le sue derivate parziali rispetto alle p incognite a_j per j = 1, 2, …, p. Così facendo si ottiene un sistema lineare di p equazioni in p incognite, che, risolto, dà proprio i coefficienti cercati.

$$\sum_{i=0}^{m} s(n+i-k) \cdot s(n+m) = \sum_{j=0}^{p} a_j \sum_{i=0}^{m} s(n+i-k) \cdot s(n+i-j) \tag{1.12}$$

Ricordando il modello di figura 1.15, chiamando u(n) il segnale in ingresso all'amplificatore, quello in uscita sarà $G_0u(n)$, ricordando poi che $s(z)=G_0u(z)H(z)$, si può stabilire l'uguaglianza $s(z)/H(z)=G_0u(z)$. Come già specificato la H(z) è una funzione di soli poli e perciò può essere scritta

$$H(z) = \frac{1}{1 - \sum_{k=1}^{p'} \alpha_k \cdot z^{-k}} \tag{1.13}$$

per cui antitrasformando l'uguaglianza impostata sopra si ottiene:

$$s(n) = \sum_{k=1}^{p'} \alpha_k \cdot s(n-k) + G_o \cdot u(n) \tag{1.14}$$

Il risultato fondamentale di questa analisi è il seguente: si può dimostrare che se un segnale vocale è generato con una sintesi come quella descritta dall'equazione (1.14), allora i p coefficienti che minimizzano l'errore quadratico medio in una finestra di larghezza m coincidono con gli α_k coefficienti del filtro che modella il tratto vocale. Questo importante risultato comporta anche che l'errore commesso usando l'approssimazione lineare di cui sopra è pari a $G_0u(n)$, cioè un treno di impulsi, di piccola ampiezza per la maggior parte del tempo. Inoltre ciò comporta che il calcolo dei coefficienti di predizione

fornisce automaticamente i parametri del filtro H(z), che con G_0 e il bit di selezione per i suoni sordi o sonori costituisce una rappresentazione completa del segnale vocale, frame per frame. Come sempre accade in questi algoritmi, esiste un trade-off tra la complessità di calcolo e l'accuratezza della rappresentazione: più sono i poli della H(z) (cioè i coefficienti dell'approssimazione tramite combinazione lineare), maggiore è la complessità di calcolo per la soluzione del sistema lineare (1.12).

La funzione $H(z)G_0$ fornisce lo spettro del segnale approssimante, ovvero uno spettro approssimato del segnale s(t). Tale spettro viene detto LPC e presenta la caratteristica utile di essere molto fedele nell'individuare i massimi dello spettro reale (ma poco per quanto riguarda i suoi minimi). Inoltre il parametro p permette di controllare la precisione dell'approssimazione, nel senso che un p elevato permette di evidenziare nello spettro LPC anche massimi vicini dello spettro, che altrimenti sarebbero stati fusi in un unico picco situato in una regione intermedia. Tenendo presente queste considerazioni, risulta evidente come le osservazioni fatte durante l'analisi LPC vadano sempre interpretate tenendo presente i limiti e le approssimazioni descritte. Tuttavia tale analisi riesce particolarmente utile nel processo di individuazione delle formanti (si tratta infatti di picchi ben distanziati), e costituisce un utile strumento anche per l'individuazione del pitch.

CAPITOLO 2

IL FENOMENO DELLA GEMINAZIONE E LE CONSONANTI AFFRICATE

INTRODUZIONE

Il fenomeno della geminazione è una caratteristica molto rara nelle lingue. Tra le lingue che presentano questo fenomeno, l'Italiano[1] è quella col maggior numero di parlatori e probabilmente quella che ne fa l'uso più estensivo. Per questo la lingua italiana presenta un grande interesse per chi si occupa di questo argomento in modo scientifico.

In questo capitolo si definirà il problema della geminazione e si traccerà un quadro della situazione degli studi condotti su questo fenomeno. Infine si illustreranno in dettaglio le caratteristiche fonetiche ed acustiche delle consonanti affricate nell'italiano.

2.1 LA GEMINAZIONE

In Italiano, vi sono diverse **coppie minime**, ossia coppie di parole dal diverso significato che possono essere distinte solo per la presenza o l'assenza della geminazione in una delle consonanti. Un esempio di ciò è dato dalla coppia minima *pane, panne*. Il Malmberg dà la seguente definizione: "Se una consonante è scissa in due parti da una frontiera sillabica, la chiamiamo *geminata*" (Malmberg, 1974).

[1] Pochissime lingue hanno, come l'Italiano, molte geminate. L'Hindi e il Finnico sono tra queste. Il Francese conosce consonanti geminate solo in qualche raro caso, mentre né l'Inglese, né il Tedesco, né lo Spagnolo e il Portoghese possiedono questa caratteristica.

Nel seguito della presente tesi si userà la seguente terminologia: con **rafforzamento sintattico** o **geminazione** si indicherà il fenomeno fonetico, mentre con la parola **raddoppiamento** si userà per indicare l'espediente grafico che serve a trascriverlo. Inoltre sarà detta **singola** la consonante che non subisce il rafforzamento sintattico e **geminata** quella che lo subisce.

Con **pronuncia singola** e **pronuncia geminata** si indicheranno tutti gli effetti che comporta la geminazione sull'intera parola (in particolare sulla consonante e sui fonemi adiacenti).

Verrà ora esaminato il fenomeno della geminazione dal punto di vista grammaticale, fonetico ed acustico-ingegneristico.

2.1.1 La geminazione dal punto di vista grammaticale

Non esiste, come appena accennato, una "corrispondenza biunivoca" tra pronuncia geminata di una consonante e corrispondente trascrizione grafica. Forme come *accorrere*, *eccellere*, *accanto*, sono pronunciate [ak ‘kor re re, et·‘tʃel le re, ak·‘kan:to] e anche la grafia ne tiene conto. Invece, nei casi di *a capire*, *va via*, *tu sai*, ecc., si vede agire lo stesso principio a livello di pronuncia per cui sarà [a kka'pi:re, va v·‘vi·a, tu s·‘sa·i], ma questa volta la grafia non ne tiene conto. Questo fenomeno è comunque giustificato in quanto non si parla pronunciando parole staccate, come potrebbe far supporre la scrittura, bensì emettendo intere fonie che formano la cosiddetta "catena parlata" (Canepari, 1979).

Per quanto riguarda la pronuncia all'interno delle frasi, il rafforzamento sintattico è prodotto da alcune forme uscenti in vocale e legate, semanticamente e foneticamente, alla parola seguente, che comincia con una delle consonanti che possono ricorrere geminate anche all'interno delle parole. Si riassumono le principali forme che si pronunciano rafforzate:

- La vocale a, e i monosillabi "forti" da, su, tra, fra (p. es. tra noi, fra mesi, …).
- I monosillabi che hanno accento grafico, come dà, dì, là, già, giù, sé, ciò, più ecc. (p. es. dà tutto, già lo vedo, ciò fu fatto, …).
- I verbi ho, ha, do, fa, fu, va (p. es. do tutto, fa male, …).
- Le parole che, chi, qui, qua, se, ma, o, e, tu ecc. (p. es. chi sa!, qui sotto, …).
- I polisillabi tronchi, con l'accento sull'ultima sillaba, come perché, poiché, però, andò, caffè, farà ecc.
- I quattro bisillabi piani come, dove, sopra, qualche (p. es. *sopra tutto*).

E' importante, infine, vedere quali forme non producono il rafforzamento sintattico. Esse sono, i monosillabi "deboli" la, le, lo, i, li; i monosillabi apostrofati nella scrittura come di', va' ecc. o le esclamazioni; inoltre di, ne, me, mi, te, ti, se, si, ce, ci, ve, vi, glie, gli.

Si espongono ora sinteticamente le regole della geminazione nell'italiano:

- Non si raddoppiano mai le consonanti iniziali e finali.
- Dinanzi a -ione, g e z non si raddoppiano mai (p. es. ragione, azione, …).
- Non si raddoppiano sc, gn, gl, mentre, per rafforzare ch e gh si raddoppiano solo la c e la g (p. es. ricche, agghiacciante, …).
- Il raddoppiamento di q è cq (tranne soqquadro).
- Si raddoppiano i prefissi a, e, o, da, se, su, so, ra, fra, sopra, sovra, contra (ma non contro!) ecc. (p. es. sebbene, supporre, frattanto, …).

2.1.2 La geminazione dal punto di vista fonetico

I suoni del linguaggio si distinguono gli uni dagli altri non solo per i loro tratti puramente qualitativi, ma anche per quel che concerne la "*quantità*". Già all'inizio del secolo, fonetisti come E.A.Mayer avevano intuito l'importanza linguistica degli aspetti quantitativi come la lunghezza o durata di un fonema, o anche l'intensità (energia) articolatoria. Una vocale, per esempio, è generalmente più lunga davanti ad una spirante che davanti ad un'occlusiva o davanti ad una sonora che davanti ad una sorda, più lunga anche davanti a [r] che davanti alle nasali e a [l] (Malmberg, 1974). Ancora, una vocale anteriore è spesso un po' più breve di una vocale posteriore. Per le consonanti valgono regole simili. Una sorda è normalmente più lunga di una sonora e così via. Tutti questi esempi fanno pensare che la misura di quantità relative, basate cioè sul confronto dei risultati ottenuti per differenti suoni nella stessa posizione o per lo stesso suono in posizioni diverse, sia forse molto più interessante delle misure assolute; inoltre, non tutte le variazioni di quantità "misurabili" hanno un valore linguistico propriamente detto, nel senso che non tutte portano differenze di significato. Perciò, l'osservazione condotta sulla reazione percettiva dell'uomo può dare l'auspicabile e definitiva oggettività rispetto al valore linguistico di quantità misurabili come energia e lunghezza. Queste considerazioni sono alla base delle teorie sulla geminazione.

Abbiamo introdotto il fenomeno del rafforzamento sintattico parlando di coppie minime (p. es. *fato* vs. *fatto, casa* vs. *cassa, eco* vs. *ecco,* ecc.) ma in Italiano, anche per le così dette coppie sub-minime (p. es. *l'ho dato* vs. *lodato, tra monti* vs. *tramonti, né gare* vs. *negare*), solo una corretta pronuncia del fenomeno del rafforzamento permette di eliminare i conflitti omofonici.

Secondo Muljacic, i fonemi consonantici che possono ricorrere singoli o geminati sono quindici. Essi sono: [f, v, s, p, t, k, b, d, g, m, n, l, r, τΣ, δZ]. Nelle descrizioni dell'italiano, i fonetisti si combattono su due punti di vista diametralmente opposti riguardo alla geminazione già dalla fine degli anni trenta, e la polemica non pare ancora esaurita (Muljacic, 1972). Un primo gruppo di studiosi (detti anche **monofonematisti**) si rende fautore di una *classe speciale* composta di quindici fonemi chiamati lunghi, rafforzati o intensi o, recentemente, anche tesi. Un secondo gruppo, tra i quali il Muljacic stesso, (chiamati, per contrapposizione, **bifonematisti**), considerano invece che quello che distingue una singola da una geminata, non è l'opposizione tra un fonema semplice e uno rafforzato, ma la presenza di un *fonema in più*. In pratica, una geminata sarebbe una consonante singola ripetuta due volte. Secondo i bifonematisti, quindi, l'ortografia denoterebbe (sebbene in modo imperfetto) lo stato di fatto fonologico.

In più di uno studio è stato messo in evidenza che le consonanti geminate influenzano con la loro presenza anche gli altri fonemi delle sillabe cui appartengono, anche se non è ancora chiarissimo quando e come questo avvenga.

Da quanto detto finora si capisce che il problema della geminazione è decisamente complesso e comprende aspetti diversi e interdipendenti.

2.1.3 La geminazione dal punto di vista acustico-ingegneristico

L'approccio ingegneristico agli studi linguistici è relativamente recente rispetto alle decennali (e anche secolari) tradizioni della fonetica. Inoltre gli studi ingegneristici sulla geminazione hanno anche diverse

finalità rispetto a quelli fonetici. Infatti gli studi acustici sono finalizzati in particolare al **riconoscimento automatico** della presenza o meno della geminazione e alla **produzione** con voce sintetica di una consonante geminata, tutto dopo l'analisi di risultati sperimentali. A questo scopo sono stati utilizzati una base di dati costituita da pronunce singole e geminate appositamente costruita, programmi di analisi e sintesi del segnale vocale e programmi di analisi statistica.

C'è da notare che comunque questi studi possono essere utili anche per risolvere problemi più strettamente teorici connessi al problema della geminazione.

Lavori precedenti a questo con simile impostazione e metodologia sono riportati in bibliografia (A.Vannucci 1993; R.Rossetti, 1993; F.Argiolas, 1995; F.Macrì 1995; M.Giovanardi, 1998; A.Esposito e M.G. Di Benedetto, 1999; M.G. Di Benedetto e M.Mattei, 1999). Si ricordano inoltre le precedenti tesi svolte presso il Laboratorio Voce del Dipartimento INFOCOM dell'Università di Roma "La Sapienza" e facenti parte, come questa, del "Progetto GEMMA". Esse si sono occupate delle consonanti occlusive [p, b, t, d, c, g] ([A.Vannucci 1993; R.Rossetti, 1993), delle consonati liquide [l, r] (F.Argiolas, 1995; F.Macrì, 1995), delle consonanti fricative [f, v, s, z, ʃ](M.Giovanardi, 1998) e delle consonanti nasali [m, n](M.Mattei, 1999) e sono state un importante base di partenza per l'impostazione del presente lavoro.

2.2 LE CONSONANTI AFFRICATE IN ITALIANO

I contoidi affricati italiani sono [ts, dz, tΣ, dZ] in *zia, zona, cera, gita.* Si può facilmente verificare che l'aria viene bloccata e poi liberata, eppure si sente che non sono suoni netti ed esplosivi come gli *occlusivi*. Consistono di un momento occlusivo subito seguito da uno fricativo, in quanto la lingua, togliendo l'occlusione, invece di passare direttamente al vocoide seguente, resta nella posizione del fricativo con la stessa articolazione corrispondente all'elemento occlusivo. Nel caso di *zia* e *zona* l'elemento fricativo è rispettivamente [s, z], mentre per *cera* e *gita* è [Σ, Z].

E' importante insistere sul fatto che l'elemento occlusivo degli affricati è *omorganico* (cioè con la stessa articolazione corrispondente) rispetto al fricativo che ne costituisce la parte finale. Troppo spesso si dice che, per esempio, [tΣ] è costituito da [t] seguito da [Σ], per quanto strettamente unito. Invece [t] sta solo a rappresentare la fase occlusiva determinata dall'accostamento completo degli organi nel punto d'articolazione specifico del contoide fricativo seguente per cui nella fase occlusiva di [tΣ] la lingua è disposta come per [Σ]. L'unico movimento che la lingua compie nel passare alla fase fricativa consiste nel togliere l'occlusione, spostandosi solo di quel tanto che basta per lasciar passare l'aria e causarne la frizione costringendola nello stretto passaggio del contoide fricativo. Discorso analogo vale per gli altri contoidi affricati.

I quattro contoidi affricati possono essere classificati in due diversi modi, ossia considerando le loro caratteristiche rispetto alla proprietà di essere **sordi** o **sonori** e rispetto al punto di articolazione. Infatti le consonanti [ts, tΣ] sono classificate come sorde in quanto durante la loro pronuncia (sia della fase occlusiva che della fase fricativa) non c'è vibrazione delle corde vocali, cosa che invece avviene in [dz,dZ].

Per quanto riguarda la classificazione in base al punto di articolazione, [tΣ, dZ] sono affricati **alveopalatali** (punto di articolazione della lingua tra alveoli e palato) mentre [ts, dz] sono affricati **dentali** (punto di articolazione della lingua sui denti). Questa seconda classificazione è ben illustrata dalle figure

seguenti che mostrano in sezione trasversale il condotto orale con la posizione che assume la lingua nei due diversi modi di articolazione.

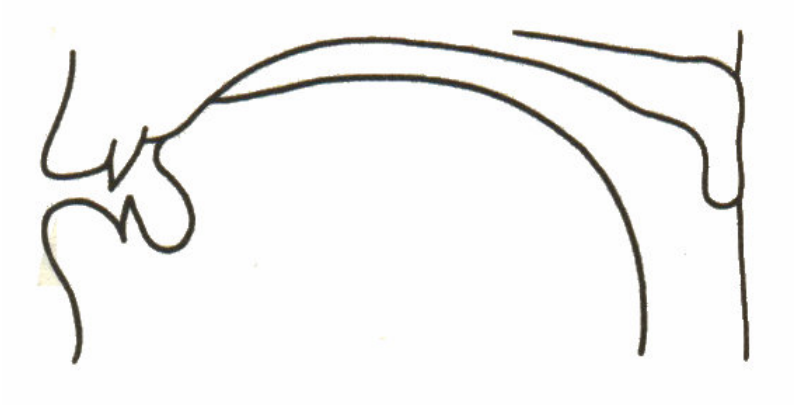


Fig. 2.1 Punto di articolazione alveopalatale delle consonanti affricate [tΣ,dZ].

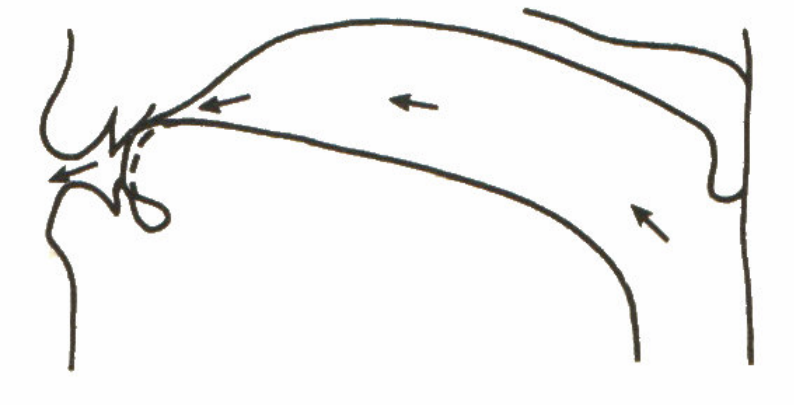


Fig. 2.2 Punto di articolazione dentale delle consonanti affricate [ts,dz]. La parte tratteggiata mostra la fase fricativa della consonante.

CAPITOLO 3

LA BASE DATI, IL SOFTWARE E GLI STRUMENTI STATISTICI

INTRODUZIONE

Nel presente capitolo si illustrerà dettagliatamente come è stata costruita la base di dati utilizzata in questo lavoro e gli strumenti utilizzati a tale scopo. Successivamente verranno descritte le potenzialità del software di analisi del segnale vocale UNICE e come esse siano state utilizzate per i particolari scopi della presente tesi. Verranno poi illustrati gli altri software utilizzati e infine verrà data una esauriente spiegazione di quali strumenti di analisi statistica sono stati utilizzati in sede di analisi.

3.1 LA BASE DATI

Per caratterizzare i quattro fonemi affricati italiani nelle loro versioni sia singole sia geminate, si è resa necessaria la raccolta di un certo numero di pronunce, in modo da poter realizzare un'opportuna base di dati in italiano. Questa base di dati è stata formata non soltanto per lo studio delle consonanti affricate ma è stata costruita per consentire l'analisi di tutti i fonemi presenti nella lingua italiana.

La base dati, infatti, comprende anche pronunce relative ai fonemi consonantici occlusivi, liquidi, fricativi e nasali. Tutte queste classi di consonanti sono già state oggetto di tesi presso il laboratorio Voce del Dipartimento INFOCOM dell'Università "La Sapienza" di Roma. Sulle classi di fonemi occlusivi e liquidi è stato eseguito anche un lavoro percezione mentre per quanto riguarda i fonemi fricativi oltre ad un lavoro di analisi è stato eseguito un lavoro di sintesi (si rimanda alle tesi citate in bibliografia per ulteriori approfindimenti).

Il presente lavoro si pone pertanto a conclusione del progetto GEMMA sulla geminazione delle consonanti italiane. Alla fine della tesi verranno perciò effettuati i confronti con i risultati raggiunti negli altri lavori.

3.1.1 Criteri di scelta dei parlatori e particolarità della base di dati delle consonanti affricate

Dato che il presente lavoro non è finalizzato direttamente al riconoscimento vocale si è ritenuto sufficiente un numero di sei parlatori, in quanto non è necessaria particolare attenzione alle sfumature di pronuncia che differenziano un soggetto rispetto ad un altro. Un primo criterio di scelta riguarda il sesso dei soggetti. Per garantire il massimo equilibrio si sono scelti tre uomini e tre donne. L'età è compresa tra i 20 e i 50 anni. Non si sono scelti parlatori più giovani in quanto proprio a circa venti anni si può ritenere la voce di un soggetto completamente formata. Si è inoltre fatta attenzione che i soggetti non presentassero particolari difetti di pronuncia o inflessioni dialettali. La scelta è quindi ricaduta su persone con un grado di istruzione medio-alto.

Nella tabella 3.1 riportiamo i dati relativi ai 6 parlatori (in ordine alfabetico) che si sono prestati per l'esperimento.

Parlatore (Sesso)	**Luogo/data di nascita**	**Formaz. Fonetica della famiglia**	**Frequentazione delle scuole**	**Professione**
AI (m)	Salerno 3/9/1967	Salerno	Roma	Studente universitario
AV (f)	Roma 13/6/1968	Potenza/Roma	Roma	Studente universitario
EZ (f)	Roma 21/4/1967	Lombardia	Roma	Studente universitario
GD (f)	Napoli 16/3/1958	Napoli	Parigi	Professore universitario
FM (m)	Roma 18/4/1967	Roma	Roma	Studente universitario
PM (m)	Roma 13/2/40	Napoli	Napoli	Professore universitario

Tab. 3.1 Dati relativi ai 6 parlatori che hanno contribuito alla formazione della base di dati. Nelle cinque colonne ci sono, rispettivamente, il nome, il cognome e il sesso, il luogo e la data di nascita, il luogo di formazione fonetica della famiglia, il luogo in cui si sono frequentate le scuole primarie e secondarie (dove quindi si è appresa la lingua), la professione.

A ciascuno dei parlatori appena elencati è stato chiesto di emettere un certo numero di pronunce di parole contenenti i fonemi [tΣ], [dZ], [ts], [dz]. Queste parole, in forma di segnale elettrico trasdotto da microfono, sono state memorizzate su nastri magnetici. Le parole scelte sono delle sequenze fonetiche prive di contenuto semantico, per ottenere una pronuncia il più possibile *neutra*. In particolare la scelta delle sequenze è stata influenzata soprattutto dagli aspetti in cui si articola lo studio: la coarticolazione e la geminazione. Per quanto riguarda la geminazione, si è scelto di far pronunciare ai soggetti tutte le

parole contenenti le consonanti affricate sia in versione singola che geminata. Per quanto riguarda invece la coarticolazione, si è limitato lo studio al solo caso vocalico, scegliendo le tre vocali che costituiscono gli estremi del trapezio fonetico: [a], [i] ed [u].[1]

Per ogni parola della base dati, per ciascun parlatore, sono state registrate tre versioni, al fine di ottenere in sede di elaborazione dei dati, dei valori medi non alterati da eventuali fenomeni aleatori. Sono state eliminate e quindi fatte ripetere alcune pronunce palesemente scorrette.

Considerando che la base dati doveva contenere pronunce relative alle quattro consonanti affricate, sono state registrate: *(4 fonemi)x(2 geminazioni)x(6 parlatori)x(3 ripetizioni)x(3 vocali) = (432 parole)*. Le parole sono state costruite secondo la struttura VCV per le versioni singole e VCCV per le corrispondenti geminate, cioè vocale-consonante-vocale, tipica dell'italiano (che solo raramente prevede parole terminanti con consonante). L'accento delle parole è stato posto sulla prima vocale, visto che la stragrande maggioranza delle parole italiane è piana. L'elenco completo delle parole componenti la base dati delle affricate italiane è mostrato nella tabella 3.2.

		Consonante							
Vocale		τΣ		δZ		τσ		δζ	
	a	ατΣα	αττΣα	αδZα	αδδZα	ατσα	αττσα	αδζα	αδδζα
	i	ιτΣι	ιττΣι	ιδZι	ιδδZι	ιτσι	ιττσι	ιδζι	ιδδζι
	u	υτΣυ	υττΣυ	υδZυ	υδδZυ	υτσυ	υττσυ	υδζυ	υδδζυ

Tab. 3.2 Elenco completo delle pronunce relative alla base dati sulle consonanti affricate italiane

3.1.2 La registrazione della base dati: modalità e strumentazione

Le registrazioni delle pronunce sono state effettuate presso il laboratorio Voce del dipartimento INFOCOM, il quale è appositamente attrezzato per questi scopi. Infatti nel suddetto laboratorio sono presenti una camera silente, un microfono omnidirezionale, un impianto stereo e il personal computer con la scheda di acquisizione ed elaborazione del segnale vocale di cui si parlerà dettagliatamente in seguito. Ci si è avvalsi dei suggerimenti di esperti e dei preziosi consigli pratici contenuti nel testo "Microphones" (Clifford, 1986). Per la registrazione delle pronunce ci si è serviti di supporto magnetico costituito da cassette TDK SA, nastri di buona qualità che presentano una risposta in frequenza praticamente piatta fino oltre 10 kHz. In un secondo tempo si è provveduto alla digitalizzazione e archiviazione delle pronunce.

Si ritiene a questo punto opportuno fare delle precisazioni e spiegare perché si sono usati dei nastri magnetici per la registrazione dei fonemi. Evidentemente se le registrazioni fossero state effettuate direttamente sul computer sarebbero state sicuramente più "pulite" e silenziose; bisogna tuttavia tenere

[1] Ad essere precisi, va osservato (vedi figura 1.8) che il trapezio fonetico ha, ovviamente, quattro vertici, e che mentre la /i/ e la /u/ italiane si trovano proprio in corrispondenza dei due superiori, altrettanto non può dirsi per la /a/. Essa, infatti, si trova al centro dei due vertici inferiori del trapezio fonetico, i quali rappresentano due vocali leggermente diverse da quella tipica italiana, una è palatale e l'altra è velare.

presente anche che le sedute di registrazione sono risultate molto stancanti per i parlatori e quindi, per evitare che le pronunce fossero affette dal fattore "stanchezza" si è cercato di velocizzare il più possibile le operazioni. Scartando per questo motivo la possibilità di registrare, controllare e catalogare contestualmente ciascuna pronuncia, una buona procedura sarebbe stata sicuramente quella di registrare di continuo l'intera seduta di registrazione sull'Hard Disk del computer per poi andare a scegliere e catalogare in un secondo tempo le pronunce corrette. Quando è stato registrato il database (1992) non si disponeva di HD così capienti e nemmeno di registratori digitali a costo contenuto e quindi, considerato anche che il fruscio introdotto dal nastro non risultava così fastidioso per gli scopi preposti, si è scelto di registrare l'intera seduta e di digitalizzare le parole in un secondo tempo.

Si rimanda alle tesi citate in bibliografia per altri particolari riguardanti le modalità di esecuzione delle registrazioni (R.Rossetti, 1993; A.Vannucci, 1993).

Vediamo ora più in dettaglio le caratteristiche tecniche del materiale utilizzato:

- Camera silente: Mini Cabina Amplisilence della Amplifon, con pareti interne fonoassorbenti per eliminare il riverbero della voce e una capacità di abbattimento dei rumori esterni di circa 30 dB alle frequenze di interesse.
- Microfono: SONY ECM 144, omnidirezionale (per catturare il suono proveniente da qualsiasi direzione), con risposta in frequenza piatta fino a 15kHz, mono, della tipologia a condensatore, con -55.3 dBm/mbar di sensibilità (potenza del segnale generato, in dB, in presenza di un suono di 1 mbar di pressione acustica). La scelta di questo particolare strumento è stata guidata dalla consultazione del testo "Microphones" (Clifford, 1986), testo assolutamente esauriente in materia.
- Impianto stereo: KENWOOD KT-48L con possibilità di regolazione del volume di registrazione (caratteristica che assicura l'assenza del dispositivo di regolazione automatica del volume d'ingresso, di cui sono dotati molti moderni apparecchi stereo, e che opera un filtraggio imprevedibile del segnale d'ingresso, al fine di evitare la saturazione della dinamica del dispositivo e del nastro).

Il collegamento del microfono (interno alla cabina) e l'impianto stereo (esterno) avviene tramite l'apposito pannello posto sul fronte della cabina stessa. In questo modo si tiene la porta sigillata e si isola il soggetto (e il microfono) dai rumori esterni e si evita il riverbero della voce dello stesso parlatore grazie alle pareti fonoassorbenti della cabina. Inoltre il vetro trasparente della cabina permette un contatto visivo tra il parlatore e chi registra la seduta. Grazie a tale caratteristica si sono realizzate le registrazioni mostrando ai soggetti dell'esperimento le parole da pronunciare mediante cartelli. Essendo mono il segnale prodotto dal microfono durante la registrazione, è stato sfruttato un solo canale dello stereo e una sola pista delle cassette magnetiche, senza che ciò abbia avuto conseguenze sulla qualità della registrazione. Nelle fasi di riascolto in cuffia il segnale è stato ridistribuito su entrambi i canali, in modo da permettere un ascolto più chiaro e naturale.

Le registrazioni effettuate su nastro magnetico sono poi state digitalizzate usando il software UNICE, utilizzato anche in seguito per l'analisi dei segnali.

3.2 UNICE: IL SOFTWARE PER L'ANALISI DEL SEGNALE VOCALE

UNICE è un software per l'ambiente MS-DOS progettato e realizzato dalla società francese Vecsys. Il programma sfrutta le routine del sistema di gestione della scheda per PC-IBM **AU21** prodotta dalla OROS. Questa è dotata di un chip di campionamento e tenuta a 16 bit, capace di lavorare fino ad una frequenza massima di 128kHz, di un filtro analogico con banda pari a 20kHz e del chip per il DSP **TMS320C25** della Texas Instruments. Per quanto riguarda l'interfacciamento con l'ambiente esterno, la scheda dispone di: un ingresso microfonico (MIC), un ingresso e un'uscita per il collegamento diretto con un sistema di riproduzione, registrazione e amplificazione del segnale audio (LINE IN / LINE OUT) e, infine, un'uscita per la cuffia (PHONES) (si rimanda per ulteriori specifiche tecniche al manuale di riferimento OROS citato in bibliografia). Grazie a questo dispositivo hardware è possibile ottenere una velocità di elaborazione che consente di visualizzare gli spettrogrammi in tempo reale.

Le principali funzioni di UNICE sono:

- Registrazione (da microfono o da ingresso esterno) e digitalizzazione di un segnale analogico.
- Visualizzazione dell'andamento del segnale nel tempo
- Ascolto in cuffia o su supporto esterno
- Visualizzazione e calcolo degli spettri e degli spettrogrammi in tempo reale con differenti tecniche (FFT a banda stretta e larga, LPC).
- Visualizzazione e calcolo della frequenza di pitch.
- Visualizzazione dell'energia a breve termine del segnale.

Il programma UNICE è descritto sommariamente nel relativo manuale di utilizzo (Vecsys, 1989). Si cercherà qui di seguito di metterne in luce le caratteristiche più rilevanti e le potenzialità maggiormente utili per il presente lavoro.

3.2.1 L'analisi temporale con UNICE

UNICE memorizza il segnale digitale in due file separati, con lo stesso nome ma con estensioni diverse. Il primo, con estensione **.sig** (da signal), contiene i dati veri e propri, ossia i campioni, mentre il secondo, con estensione **.key**, contiene le informazioni necessarie all'interpretazione dei dati. Il formato adottato per i file .sig, consiste in una semplice sequenza di campioni (tanti quant'è la frequenza di campionamento adottata moltiplicata per la durata del segnale in secondi[2]), ognuno dei quali è rappresentato con 16 bit in complemento a 2, senza alcuna intestazione. Il file .key che viene automaticamente creato, contiene la frequenza di campionamento, il numero di campioni ed eventuali segmentazioni ed etichettature. In pratica, l'insieme dei due file equivale al più conosciuto e utilizzato formato **wave** [.wav], che ha, però, intestazione e campioni del segnale in un unico file. La struttura di un file *.key* è mostrata in figura 3.1. Come detto in esso sono memorizzate anche le segmentazioni e le

[2] L'unico vincolo per il numero di campioni del segnale è che deve essere un multiplo intero di 128, dato che, come si spiegherà fra poco, Unice divide il segnale in frame di 128 campioni l'uno.

relative etichettature: UNICE permette, infatti, semplicemente con il mouse, di segmentare il segnale esattamente in corrispondenza di un ben preciso campione, evidenziandolo nella forma d'onda temporale con una barra rossa verticale. Questa possibilità si è rilevata utilissima per l'analisi temporale e si è rivelato molto comodo anche il fatto che la segmentazione sia memorizzata nel file *.key*.

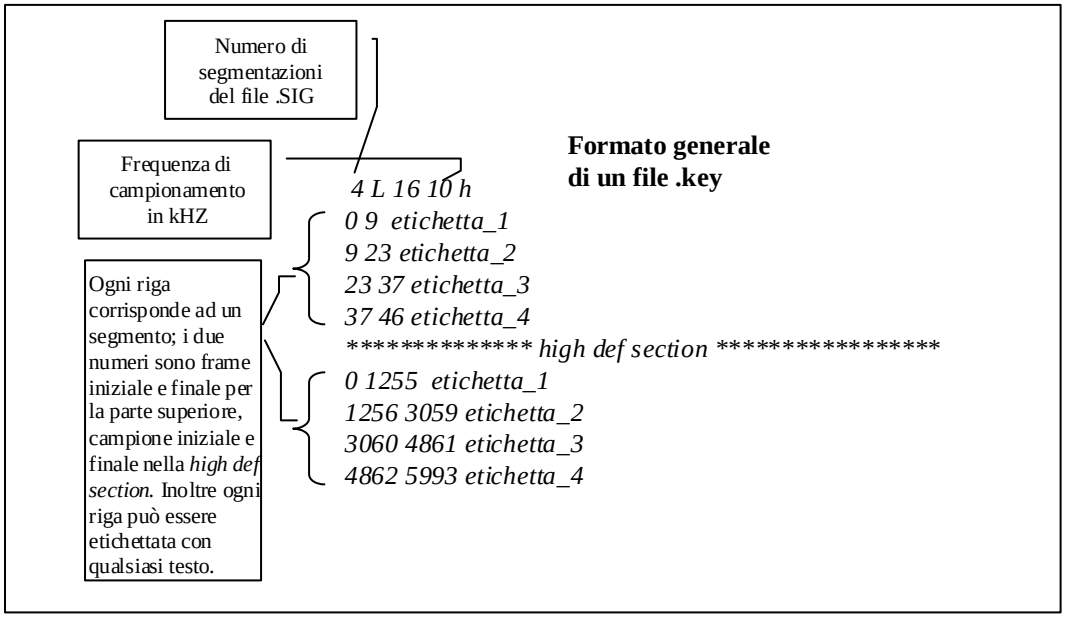


Fig. 3.1 Formato generale di un file .key usato da UNICE allo scopo di memorizzare le informazioni relative alle caratteristiche e alle segmentazioni di un file di voce.

3.2.2 Il metodo della *"short-time analysis"*

Si è già avuto modo di rilevare nel primo capitolo che, nello studio del segnale vocale interessano soprattutto le caratteristiche locali, in modo da poter associare le variazioni del tratto vocale alle variazioni del segnale nel tempo e in frequenza. Sarebbe di scarsa utilità conoscere l'energia oppure la FFT di un segnale nella sua totalità. Per questo, quella che si usa in genere è la tecnica chiamata short-time analysis (Rabiner e Schafer, 1978) con la quale si prende in considerazione di volta in volta, solo una sequenza di campioni relativi ad una parte del segnale. Matematicamente la sequenza può essere rappresentata come

$$Q_n = \sum_{m=-\infty}^{\infty} T[x(m)] \cdot w_N(n-m) \qquad (3.1)$$

dove T[ ] rappresenta una generica trasformazione (lineare o non lineare) operata sul segnale vocale che può dipendere da alcuni parametri e $w_N(n)$ è una finestra rettangolare di ampiezza N (cioè, con soli N campioni pari a 1 e tutti gli altri identicamente nulli) traslata in corrispondenza del campione di indice n. Esempi di analisi di questo tipo sono la FFT narrow-band o wide-band, l'analisi LPC, l'analisi condotta con la short-time energy. In quest'ultimo caso, ad esempio, si ha

$$E_n = \sum_{m=n-N+1}^{n} x^2(m) \qquad (3.2)$$

dove l'operazione T[ ] è semplicemente il quadrato. En rappresenta l'energia del segnale, considerato per soli N campioni consecutivi alla volta. La "finestra" di analisi viene ogni volta traslata in avanti di un campione ed En di nuovo calcolata. Il suo significato è alquanto diverso della semplice energia totale del segnale, ottenuta quadrando e sommando tutti i campioni. Un esempio dell'andamento dell'energia a breve termine per un segnale vocale è mostrato in figura 3.2.

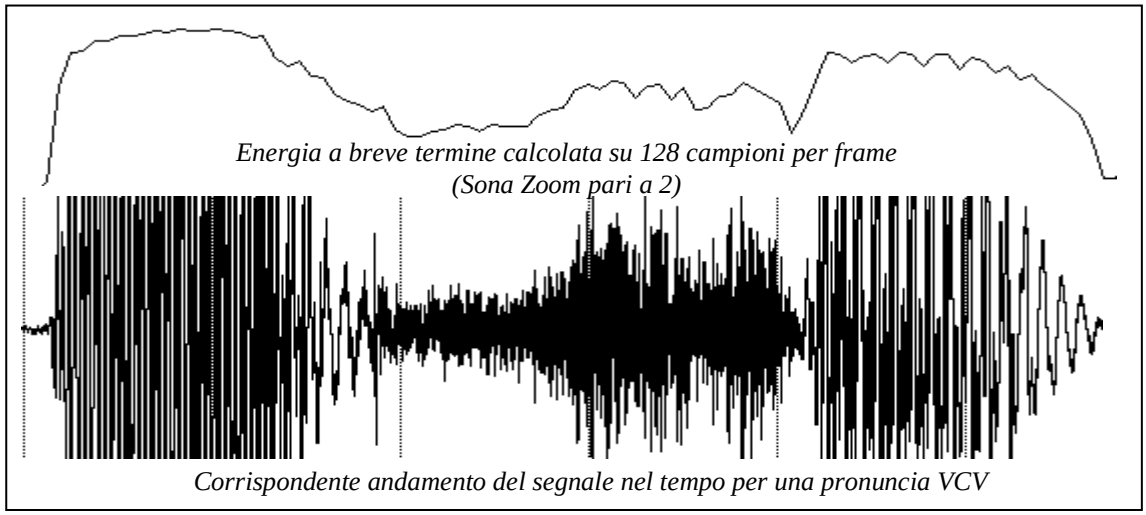


Fig. 3.2 Esempio di visualizzazione di energia a breve termine con il programma UNICE; sono visualizzati sullo stesso asse temporale circa 550 ms di segnale.

Per una corretta analisi è importante chiedersi:

- quanto dovrebbe essere **l'ampiezza della finestra** o **frame** N e su cosa influisce questo numero;
- se sia il caso di calcolare Q_n per tutti gli n, o se si può saltare il calcolo di alcuni elementi della sequenza e ripeterlo solo per dei multipli interi di n: in pratica, quanto dovrebbe essere il **fattore di sovrapposizione** tra frame adiacenti.

La risposta a queste domande dipende evidentemente dal tipo di analisi che si vuole effettuare. Per chiarire questo concetto consideriamo l'esempio del calcolo della FFT. Le FFT narrow-band e wide-band, costituiscono un esempio di trasformazione T[·] e si differenziano per il fatto che la prima ha una finestra di analisi di ampiezza maggiore della seconda. In figura 3.3 sono mostrati tre esempi di FFT (a 128, a 256 e a 512 campioni), per uno stesso segnale di voce campionata a 16 kHz (rappresentante una vocale). Alle tre FFT corrispondono, rispettivamente, una finestra temporale di analisi di 8, 16 e 32 ms. Potremmo affermare che: la prima è una wide-band, la terza è una narrow-band mentre la seconda è una via di mezzo tra le altre due. La figura 3.3 è molto esplicativa: risulta evidente che, più la finestra è ampia più la risoluzione in frequenza aumenta, anche se, ovviamente, ne risente la velocità di calcolo (il numero di punti della FFT è più grande). Lo svantaggio maggiore, in ogni modo, è quello di una diminuzione di risoluzione temporale. Il fattore di sovrapposizione può essere usato per raggiungere dei buoni compromessi tra le due esigenze.

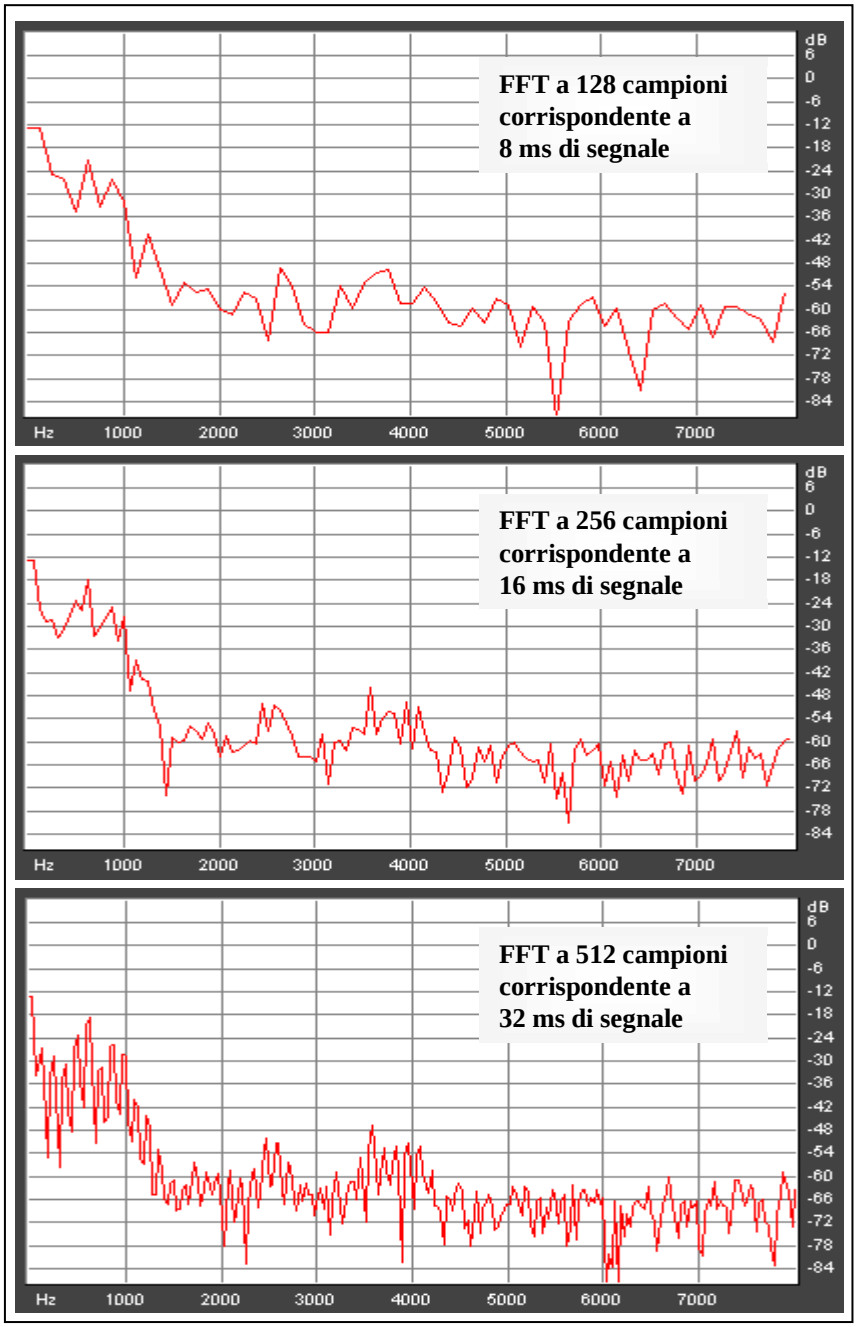


Fig. 3.3 Spettri FFT a 128, 256 e 512 campioni per un segmento di segnale vocale campionato a 16 kHz.

Nel caso delle analisi riguardanti il segnale vocale, bisogna considerare che, mediamente, per la voce di un uomo il periodo di pitch è di 8 ms mentre per una donna è di 4.4 ms (vedi tabella 1.4), e che la lunghezza di un fonema è in media di 150 ms. Perciò, occorre prestare molta attenzione nello scegliere l'ampiezza N della finestra di analisi in funzione del caso oggetto di studio: per esempio, se si è interessati alle caratteristiche prosodiche delle parole servirà sicuramente una finestra più ampia che per analizzare i singoli fonemi, per i quali servirà, a loro volta, una finestra più ampia che per l'analisi delle zone di transizione o di brusche variazioni nel segnale e così via. Per finire, si affermerà che un giusto

compromesso tra tutte le esigenze è, come al solito, la soluzione ottimale; per arrivare a questo, tuttavia, si devono conoscere a fondo tutti i vantaggi e gli svantaggi delle scelte che ci si presentano.

UNICE gestisce la short time analysis suddividendo il segnale nel tempo in frame la cui lunghezza N varia in funzione della frequenza di campionamento f_c (in kHz) secondo la semplice relazione:

$$N = \frac{128 \cdot f_c}{10} \tag{3.3}$$

La durata di ciascun frame, uguale a N/f_c, è, invece, fissa e pari a 12.8 ms (comprendendo quindi 128 campioni per una frequenza di campionamento di 10kHz).

3.2.3 L'analisi in frequenza con UNICE

Per l'analisi in frequenza Unice mette a disposizione sia uno spettrogramma a tutto schermo del tipo di quello mostrato in figura 1.17 che un'altra finestra di analisi più piccola, a sua volta suddivisa in due semifinestre, dove sono visualizzati gli spettri e/o il segnale nel tempo, frame per frame, come mostrato in figura 3.4. Ricordiamo che lo spettrogramma è un diagramma tridimensionale: tempo (o meglio frame essendo la dimensione temporale quantizzata in pacchetti) sulle ascisse, frequenza sulle ordinate e ampiezza, visualizzata tramite una tonalità di grigio, più scura se l'ampiezza è più alta (Oppenheim, Schafer, 1975).

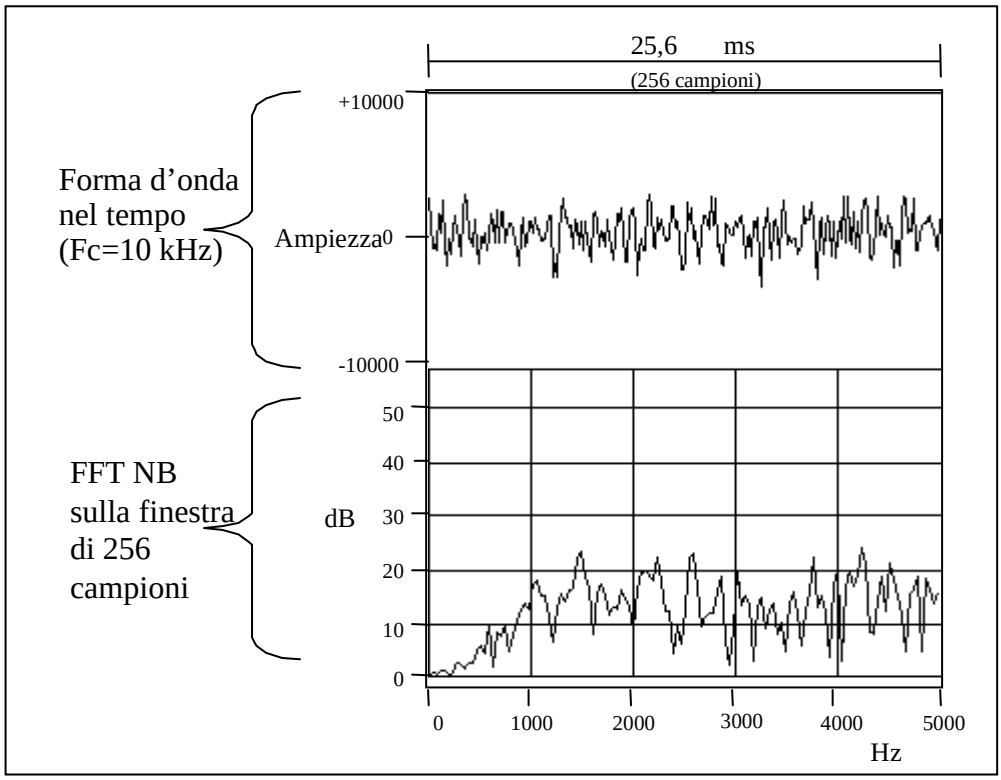


Fig. 3.4 Esempio di segnale + spettro visualizzato da UNICE relativamente ad un solo frame di analisi.

Le modalità di calcolo della FFT usate dal programma per questo tipo di analisi sono le seguenti:

1. FFT a banda stretta, realizzata a partire da 256 campioni nel tempo, precedentemente finestrati con finestra di Hamming, che restituisce 128 campioni in frequenza (e non 256, per effetto della simmetria che presenta lo spettro di un segnale campionato). Nel caso di un segnale campionato a 10Khz si ha una risoluzione in frequenza di 39.0625Hz (vedi formula 1.7).
2. FFT a banda larga, realizzata a partire da 60 campioni nel tempo precedentemente finestrati con finestra di Hamming, che restituisce ancora 128 campioni in frequenza. In effetti, teoricamente dovrebbero essere 30, ma, usando l'artificio di considerare 98 campioni nulli seguiti dai 60 campioni di cui si vuole la FFT a banda larga (FFT WB) e poi ancora da 98 campioni nulli, e calcolando su questi 256 campioni totali una FFT a banda stretta (FFT NB), si ottengono di fatto 128 campioni in frequenza, come mostrato in figura 3.5. Questa tecnica, per cui si aggiungono dei campioni nulli, è chiamata "zero padding" e non consente di aumentare la risoluzione in frequenza ma solo di migliorare la visualizzazione della FFT. La risoluzione effettiva in frequenza sarà di 188Hz (formula 1.7 con N pari a 60) per una frequenza di campionamento di 10kHz. Per maggiori dettagli sullo zero padding rimandiamo a Oppenheim e Schafer (1975).

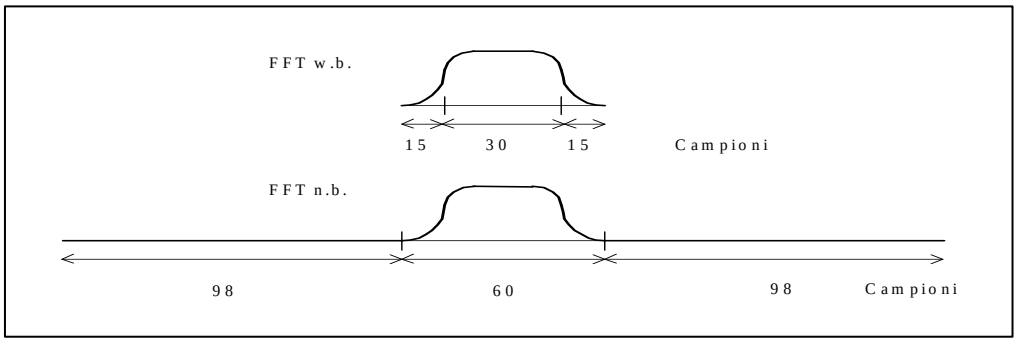


Fig. 3.5 FFT wide band e narrow band con *zero-padding*.

Quindi la FFT NB consente un'analisi più accurata in frequenza (essendo maggiore la risoluzione che offre) rispetto alla FFT WB, tuttavia, quando le caratteristiche del segnale subiscono variazioni repentine (in conseguenza di altrettanto rapide variazioni del tratto vocale), la FFT WB permette di isolare più selettivamente la zona di analisi, in virtù dei soli 60 campioni nel tempo di cui fa uso per il calcolo.

Nell'analisi in frequenza oltre alla FFT è disponibile anche l'LPC, con cui vengono calcolati i coefficienti di predizione su 256 campioni precedentemente finestrati con finestra di Hamming.

In tutti i casi è prevista un'*enfatizzazione* dello spettro con coefficiente pari a 0.95 tesa ad esaltare l'importanza del contenuto energetico in alta frequenza (vedere par. 1.4.4 per i dettagli).

Né la durata, né la posizione dell'inizio o della fine di un frame sono modificabili. Per ovviare a questa limitazione, che non consentirebbe di analizzare sequenze di campioni appartenenti a frame diversi, ma adiacenti, si può far uso delle opzioni offerte dal *Sona Zoom*. Si può impostare, infatti, il parametro Sona Zoom in una scala di valori tra 1 e 8: la dimensione della finestra tramite la quale viene condotta l'analisi visiva è 12.8/SZ, fino perciò ad un minimo di 1.6 ms.

La scala temporale di visualizzazione dipende dal fattore di Sona Zoom. Se esso vale 1, per ogni frame viene visualizzata una sola FFT: l'analisi in frequenza viene quindi ripetuta ogni 12.8 ms. Quando Sona Zoom è impostato ad *i*, per ogni frame vengono visualizzate *i* FFT. Chiaramente se si è impostato il

parametro Sona Zoom pari a 1 allora lo spettro narrow mostrato in una semifinestra è calcolato esattamente sulla porzione di segnale visualizzata nell'altra semifinestra. Se ci si sposta di un frame a destra o a sinistra in Sona Zoom pari a SZ (per SZ=1, 2,...8), lo spettro narrow mostrato nella finestra di visualizzazione dettagliata è calcolato in una finestra di 25.6ms che si sovrappone alla precedente per un fattore pari a

$$S = \frac{(2^{\cdot}SZ - 1)}{(2^{\cdot}SZ)} \qquad (3.4)$$

In tal modo si può impostare il passo di spostamento della finestra di calcolo degli spettrogrammi tra otto valori diversi.

A conclusione delle caratteristiche di UNICE riguardo all'analisi in frequenza pensiamo sia interessante far vedere analogie e differenze tra i tre tipi di analisi FFT NB, FFT WB e LPC sia sulla base degli spettrogrammi (per l'analisi complessiva di una pronuncia VCV) sia sulla base degli spettri (per l'analisi di un singolo frame posto al centro di una vocale).

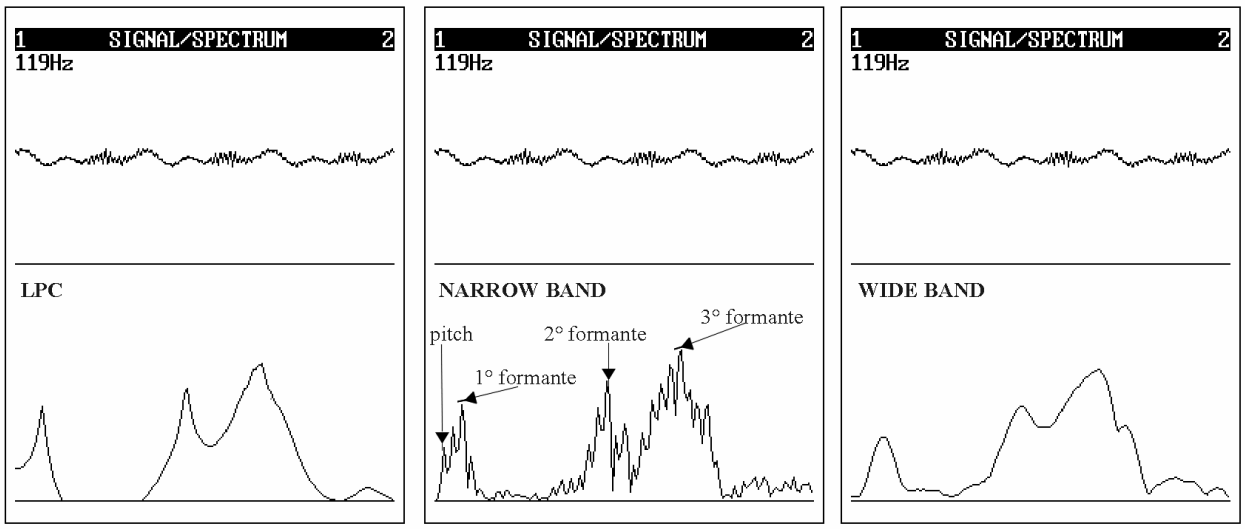


Fig. 3.6 Forma d'onda e corrispondenti spettri LPC, FFT NB e FFT WB relativi al frame centrale della prima vocale /i/ della pronuncia /ιτΣι/. In alto a sinistra è indicata la frequenza di pitch calcolata dall'algoritmo automatico di UNICE.

A questo scopo, si osservino attentamente i grafici di figura 3.6 dove sono mostrati gli spettri calcolati al centro di una vocale. Si supponga di voler ricercare i valori delle formanti. Ci si accorge facilmente che nello spettro WB (a destra) la risoluzione in frequenza minore comporta anche una minore accuratezza nella ricerca del valore esatto delle formanti, tanto è che due picchi "vicini" non vengono distinti (correndo il rischio di perdere qualche formante rispetto al NB, visualizzato al centro); di contro, con il WB si vede meglio l'inviluppo o la forma dello spettro e quindi questo risulta più indicato per osservare più in generale picchi e larghezze di banda. La stessa cosa può dirsi per lo spettro LPC (a sinistra) e di come esso "assomigli" al WB. La differenza sostanziale è che il primo individua meglio i picchi delle formanti, ma peggio, senza dubbio, la forma dello spettro e quindi le larghezze di banda.

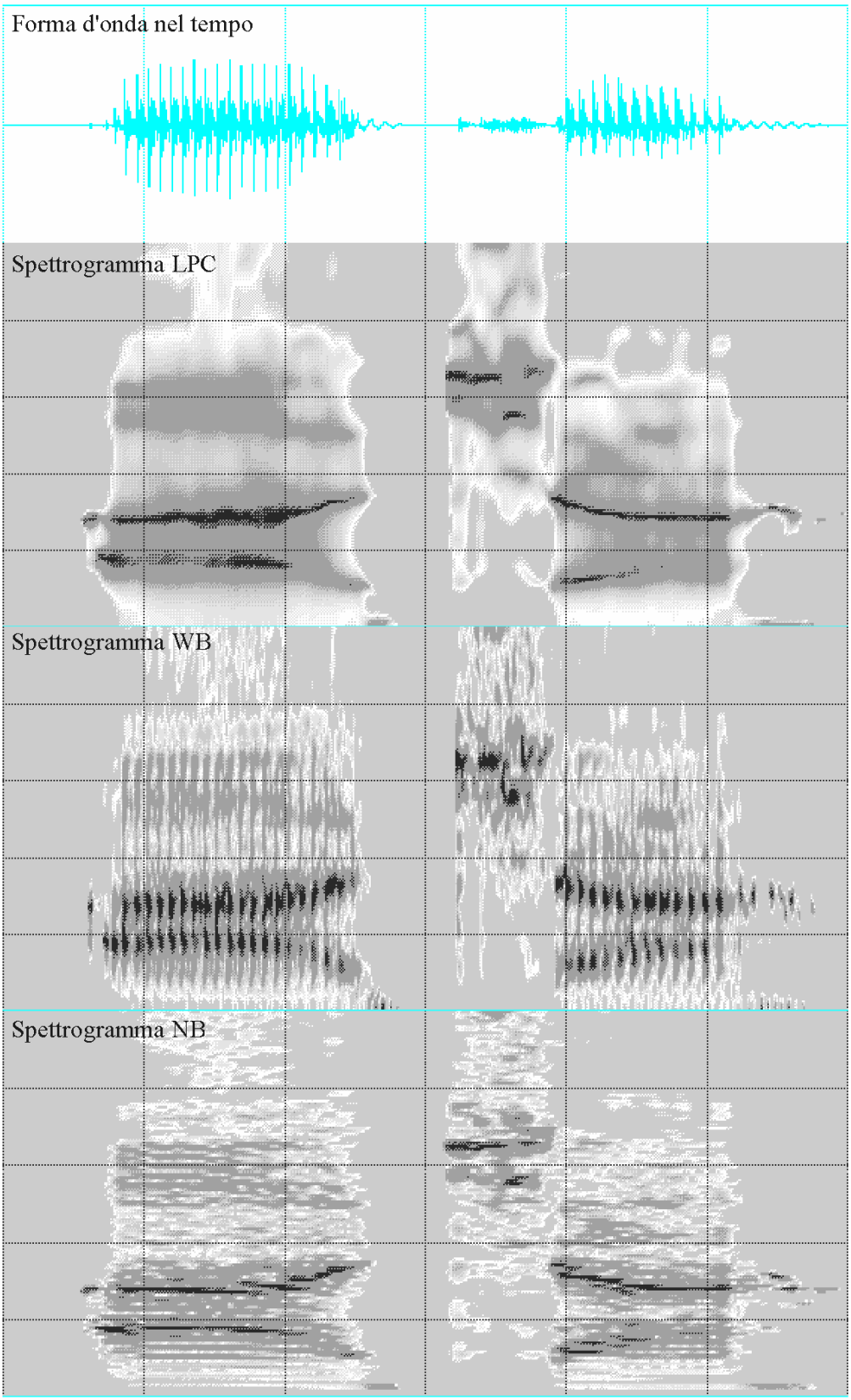


Fig. 3.7 Forma d'onda e corrispondenti spettrogrammi LPC, FFT NB e WB calcolati da UNICE sulla pronuncia della parola /ατΣα/. Sona Zoom è stato impostato a 2: l'intervallo temporale visualizzato corrisponde a circa 600 ms

3.3 GLI ALTRI SOFTWARE UTILIZZATI

Oltre ad UNICE sono stati utilizzati molti altri software per la stesura della tesi. Ne daremo ora una panoramica, necessariamente sintetica e schematica per non appesantire troppo la trattazione, con particolare riguardo alle caratteristiche e potenzialità utilizzate nel presente lavoro. Si spera comunque che queste brevi note siano utili per eventuali sviluppi futuri e che possano dare una idea più chiara di come si sia proceduto nella stesura del presente lavoro.

- **Compilatore C:** è stato utilizzato il Turbo C++ della Borland, versione 3.0 (1992). Anche se si tratta di una versione ormai superata, si è rivelato di estrema praticità e di facile utilizzo. Non è stato necessario usare versioni più recenti e con maggiori funzionalità in quanto si è sempre rimasti nello standard dell'ANSI C. Ciò principalmente per due motivi, sia perchè non sono mai servite funzioni non appartenenti a questo standard e sia per favorire la comprensione dei listati da parte di chiunque volesse utilizzare i programmi. Inoltre in questo modo è anche possibile la portabilità su compilatori che non siano della Borland.
- **Software di statistica:** la scelta è ricaduta su **Statgraphics plus 2.1**, il quale risulta decisamente completo e potente per gli scopi necessari e inoltre permette la facile importazione di dati da fogli di lavoro Microsoft Excel. Grazie ad un piccolo accorgimento è inoltre possibile esportare in un documento Word i risultati completi dell'analisi statistica condotta, la quale è parte fondamentale della tesi ed è riportata in appendice E. Gli strumenti di tale programma utilizzati sono: analisi della varianza mono e multivariata e test di correlazione di Spearman (vedi successivo Paragrafo 3.4).
- **Scrittura e foglio di calcolo:** a questo scopo è stato utilizzato il pacchetto Microsoft Office 2000. Questo pacchetto, comune e diffuso oramai in tutto il mondo, rappresenta un vero e proprio ambiente di lavoro integrato sotto il sistema operativo Windows. Si compone di più programmi che permettono varie funzionalità. Per l'utilizzo che se ne è fatto durante lo svolgimento della tesi, due sono gli applicativi che sono stati fondamentali: Excel e Word. Il primo è stato utilizzato per la stesura delle tre appendici A, B e C e per la maggior parte delle tabelle presenti nella tesi. Oltre alla normale formattazione di una tabella, per altro resa molto veloce dalla possibilità di automatizzare molte procedure, il programma è in grado di definire elaborazioni matematiche tramite formule che collegano tra loro le caselle delle tabelle stesse. Queste potenzialità sono state usate per le elaborazioni statistiche più semplici come medie e deviazioni standard senza bisogno di utilizzare il programma di statistica. Il programma Word è stato utilizzato per l'intera stesura della tesi, permettendo di inglobare nel testo grafici, tabelle, disegni, formule, immagini ecc. con relativa semplicità e con una impostazione della grafica tale da rendere il più chiara ed immediata possibile la lettura e la comprensione del lavoro svolto.
- **Grafica:** due sono stati i programmi utilizzati. Il primo, **Paint Phop Pro**, installato sullo stesso computer del Laboratorio Voce, ha permesso la "cattura" dallo schermo di parte delle immagini presenti nel testo e il suo salvataggio in uno dei tanti formati di file di grafica esistenti (bmp, pcx, jpg ...). Inoltre, grazie ad esso, è stato possibile visualizzare vicine due o più immagini catturate da UNICE, il quale non permette la visualizzazione multipla su più finestre contemporaneamente di diversi file di segnale. Per le elaborazioni delle immagini (anche di quelle catturate con Paint Shop Pro) si è preferito utilizzare il pacchetto **Corel Draw 6.0** della canadese Corel, decisamente più

completo e funzionale. Si compone di varie unità, tra cui ricordiamo le tre che sono state utilizzate e che sono: **Photo Paint**, particolarmente adatto al fotoritocco e al trattamento delle immagini scannerizzate da libri; **Corel Capture**, per la cattura di immagini dallo schermo del computer; infine **Corel Draw**, programma di grafica vettoriale che ha permesso l'inserimento e la modifica di testo e ogni tipo di simbolo grafico nelle immagini presenti sulla tesi. Grazie al supporto delle immagini così trattate e modificate è stata resa più chiara e semplice la comprensione del lavoro svolto.

- **Programma audio:** oltre al già ampiamente descritto UNICE, per l'analisi dei segnali audio, soprattutto in sede di sintesi (vedi Capitolo 5), è stato di grande aiuto il programma **Sound Forge 4.5** della Sonic Foundry. Si ricordi a tal proposito che UNICE permette l'ascolto in cuffia del singolo frame, ma che se si volesse ascoltare mezzo frame o addirittura il singolo campione, non sarebbe possibile. Importando invece i file .sig in Sound Forge è possibile ascoltare ogni porzione di segnale che si vuole, senza il limitante vincolo che tale porzione sia multiplo intero del frame. Grazie a ciò è stato possibile effettuare le segmentazioni delle pronunce nella maniera più precisa possibile, limitatamente ai casi più difficili. Per il suo uso in sede di sintesi si veda il già citato Capitolo 5 della presente tesi, in cui sono esposte più dettagliatamente le funzionalità utilizzate a tal scopo.
- **Programma di sintesi:** ultimo, ma non per importanza, il programma **HLsyn 2.2** della Sensimetrics Corporation, progettato e sviluppato da vere autorità nel campo dell'acustica, tra cui ad esempio Kenneth Stevens. Tale software permette la sintesi del segnale vocale tramite l'utilizzo di un limitato set di parametri che schematizzano il comportamento dell'apparato fonatorio umano durante la produzione del parlato. Inserendo gli istanti temporali corrispondenti alle durate medie dei fonemi misurate in sede sperimentale sia per le pronunce delle affricate singole che geminate, e modificando opportunamente i parametri del sintetizzatore si è riusciti a riprodurre sinteticamente i fonemi che sono stati oggetto di studio in questa tesi. Questo lavoro di sintesi è finalizzato ad un successivo esperimento di analisi percettiva. Anche qui si rimanda al Capitolo 5 per maggiori delucidazioni a riguardo.

3.4 GLI STRUMENTI STATISTICI PER L'ANALISI DEI DATI

Uno dei maggiori problemi associati alle misure e valutazioni di qualsiasi aspetto del comportamento umano è la sua intrinseca variabilità. **Variabilità**, semplicemente, significa che valori ottenuti dalla misura di un parametro non saranno le stesse in differenti circostanze, rendendo impossibile la decisione di quale sia quello "giusto". Si può in ogni modo pensare che l'uomo, pur producendo una certa variabilità, riporti la sua intenzione a prodursi in atti identici di comportamento. Questa "intenzione", potrebbe considerarsi un'astrazione che non contenga variabilità. Si rendono allora necessari dei metodi automatici che muovano dai dati variabili misurati verso "invarianti" astrazioni. Di questo aspetto così importante per lo studio condotto in questa tesi, si occupa la statistica alla quale si è voluto dedicare questo intero paragrafo.

3.4.1 Media aritmetica e deviazione standard

La più semplice statistica che si che si può estrarre da *n* dati raccolti è la media aritmetica, così definita:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{3.5}$$

dove *n* è il numero di campioni mentre x_i rappresenta il valore dell' i-esimo campione.

L'attendibilità della media aritmetica quale valore rappresentativo di un insieme di misure di un parametro dipende dal numero di campioni misurati e dal *range* di variabilità di ciascuno. Un'indicazione sul *range* della maggior parte dei valori può essere data dalla deviazione standard calcolata nel modo seguente:

$$StD = \sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1}} \tag{3.6}$$

A proposito di questa formula bisogna fare alcune precisazioni. Considerando che le misure vengono eseguite su un campione finito della popolazione, e che quest'ultima segue generalmente una distribuzione continua di probabilità per il parametro in esame, si ha che la deviazione standard, come statistica di campione, approssima in maniera più o meno precisa la radice della varianza (incognita) della popolazione (StD $\approx \sigma$). Quando si ha a che fare con un numero piccolo di campioni (in genere <30) la formula (3.6) costituisce una buona approssimazione; mentre, se il numero di campioni è grande (in genere >30), si usa normalmente la stessa formula con al denominatore n al posto di (n-1). Nel presente lavoro si è sempre usata l'approssimazione per piccoli campioni considerando che i dati da mediare

raramente hanno superato il numero di qualche decina di unità e che in pratica, al crescere di n, non sussiste alcuna differenza tra le due definizioni (Dillon e Goldstein, 1984; M.Spiegel, 1988).

3.4.2 Il test di analisi della varianza: l'ANOVA

Introduzione

L'analisi della varianza è la metodologia statistica usata per individuare e quantificare l'eventuale influenza delle tipologie prese in considerazione (sesso, pronuncia singola o geminata, vocali, consonanti) sulle misure rilevate dei diversi parametri scelti per l'analisi delle consonanti affricate.

Nel presente sottoparagrafo illustreremo tale metodologia basandoci su "Introduzione alla statistica" di T.H. Wonnacot, R.J. Wonnacot (1972), utilizzando alcune esemplificazioni classiche per questo tipo di trattazione e cercando di utilizzare, quando possibile, dei concetti intuitivi anziché lunghe dimostrazioni.

Analisi della varianza a un fattore

La significatività dei risultati di un'indagine può essere compresa mediante il seguente esempio: vogliamo confrontare tra loro tre macchine (A, B e C), le quali, essendo azionate da uomini e a causa di altre ragioni inesplicabili, danno luogo ad un prodotto orario soggetto a fluttuazioni casuali. Nella speranza di "mediare" e quindi di ridurre gli effetti di tali fluttuazioni, si effettua un campione casuale di 5 ore per ciascuna macchina, i cui risultati sono raccolti nella Tabella 3.3, insieme alle relative medie.

Macchine o numero del campione	*Campione della macchina i*					$\overline{X}_i$
$i = 1$	48,4	49,7	48,7	48,5	47,7	48,6
$= 2$	56,1	56,3	56,9	57,6	55,1	56,4
$= 3$	52,1	51,1	51,6	52,1	51,1	51,6

$$Media delle \overline{X} = \overline{\overline{X}} = 52,2$$

Tab. 3.3 Campioni dei prodotti da 3 macchine

La prima domanda che ci poniamo è: "Le macchine sono realmente differenti?". In altre parole, si vuole stabilire se le medie campionarie $\overline{X}_i$ nella Tabella 3.3 differiscono tra loro a causa della differenza nelle medie μ_i delle popolazioni da cui provengono (μ_i rappresenta la produzione media per tutto il periodo di vita della macchina *i*) oppure se queste differenze tra le $\overline{X}_i$ possono essere ragionevolmente attribuite solamente alle fluttuazioni casuali.

A scopo illustrativo, si supponga che siano stati effettuati tre esperimenti campionari su una macchina, i cui risultati sono raccolti nella Tabella 3.4. Come previsto, le fluttuazioni statistiche campionarie causano piccole differenze nelle medie dei campioni anche se le μ sono identiche.

N. del campione	*Valori campionari*					$\overline{X}_i$
i = 1	51,7	53,0	52,0	51,8	51,0	51,9
= 2	52,1	52,3	52,9	53,6	51,1	52,4
= 3	52,8	51,8	52,3	52,8	51,8	52,3

$\overline{\overline{X}} = 52,2$

Tab. 3.4 Tre campioni del prodotto di una macchina

Ne segue che la domanda può essere posta in altri termini: "Le differenze tra le $\overline{X}$ della Tabella 3.3 sono dello stesso ordine di grandezza di quelle della Tabella 3.4 (e così attribuibili alle fluttuazioni casuali), o risultano sufficientemente grandi da indicare una differenza effettiva tra le medie delle corrispondenti popolazioni?". In prima approssimazione, questa seconda spiegazione sembra la più plausibile, ma è chiaro che occorre sviluppare un test formale che fornisca elementi per rispondere con maggior rigore.

L'ipotesi di "nessuna differenza" tra le medie delle popolazioni diviene l'ipotesi nulla:

$$H_0 : \mu_1 = \mu_2 = \mu_3 \qquad (3.7)$$

L'ipotesi alternativa è che qualcuna delle μ (ma *non* necessariamente tutte) siano realmente differenti.

$$H_1 : \mu_i \neq \mu_j \qquad \text{per qualche } i \text{ e } j \qquad (3.8)$$

Per sviluppare un test plausibile di questa ipotesi, dobbiamo trovare in primo luogo una misura numerica del grado in cui le medie campionarie differiscono. A tal fine, consideriamo le tre medie campionarie nell'ultima colonna della Tabella 3.3 e ne calcoliamo la varianza; occorre sottolineare, in proposito, che stiamo calcolando la varianza delle medie campionarie e non la varianza di tutti i valori della tabella.

Avremo pertanto:

$$s^2_{\overline{X}} = \frac{1}{(r-1)} \sum_{i=1}^{r} (\overline{X}_i - \overline{\overline{X}})^2$$

$$= ½ [(48,6 - 52,2)^2 + (56,4 - 52,2)^2 + (51,6 - 52,2)^2] = 15,5 \qquad (3.9)$$

in cui r = numero delle righe (cioè numero delle medie campionarie) e

$$\overline{\overline{X}} = \textbf{media} \text{ delle } \overline{X} = \frac{1}{r} \sum_{i=1}^{r} \overline{X}_i = 52,2 \qquad (3.10)$$

Tuttavia $s^2_{\overline{X}}$ non esaurisce la questione, poiché, se consideriamo ad esempio i dati della seguente Tabella 3.5, è facile osservare che essi, pur presentando un $s^2_{\overline{X}}$ uguale a quello della Tabella 3.3, si riferiscono a macchine con maggiore variabilità, che producono grandi fluttuazioni casuali nell'ambito di ciascuna riga.

Macchine	*Prodotto campionario della macchina i*					$\overline{X}_i$
i = 1	54,6	45,7	56,7	37,7	48,3	48,6
= 2	53,4	57,5	54,3	52,3	64,5	56,4
= 3	56,7	44,7	50,6	56,5	49,5	51,6
						$\overline{\overline{X}} = 52,2$

Tab. 3.5 Campioni della produzione di 3 macchine diverse

Le implicazioni di tale fatto sono rappresentate nella Figura 3.10. In particolare, nella Figura 3.10 a) le macchine presentano una variabilità tale che tutte le produzioni campionarie potrebbero essere state ottenute da macchine della stessa popolazione, cioè le differenze nelle medie campionarie possono essere spiegate dal caso. D'altra parte le (stesse) differenze delle medie campionarie possono difficilmente essere spiegate dal caso nella Figura 3.10 b), poiché in quest'ultimo esempio le macchine *non* presentano una variabilità accentuata.

Abbiamo ora degli elementi per poter operare i confronti. Per quanto riguarda il caso rappresentato nella Figura 3.10 b), concludiamo che i valori delle μ sono diversi e rifiutiamo H_0 poiché la varianza delle medie campionarie $s^2_{\overline{X}}$ è grande *relativamente alla* varianza casuale.

Occorre tuttavia predisporre un indice per misurare la variazione dovuta al caso. Intuitivamente, ci sembra che essa possa interpretarsi come dispersione (o varianza) dei valori osservati *entro* ciascun campione, e quindi calcoliamo senz'altro la varianza entro il primo campione nella Tabella 3.3

$$s_1^2 = \frac{1}{(n-1)} \sum_{j=1}^{n} (X_{1j} - \overline{X}_1)^2 = \frac{(48,4 - 48,6)^2 + ...}{4} = 0,52 \qquad (3.11)$$

in cui X_{1j} è il j-mo valore osservato nel primo campione.

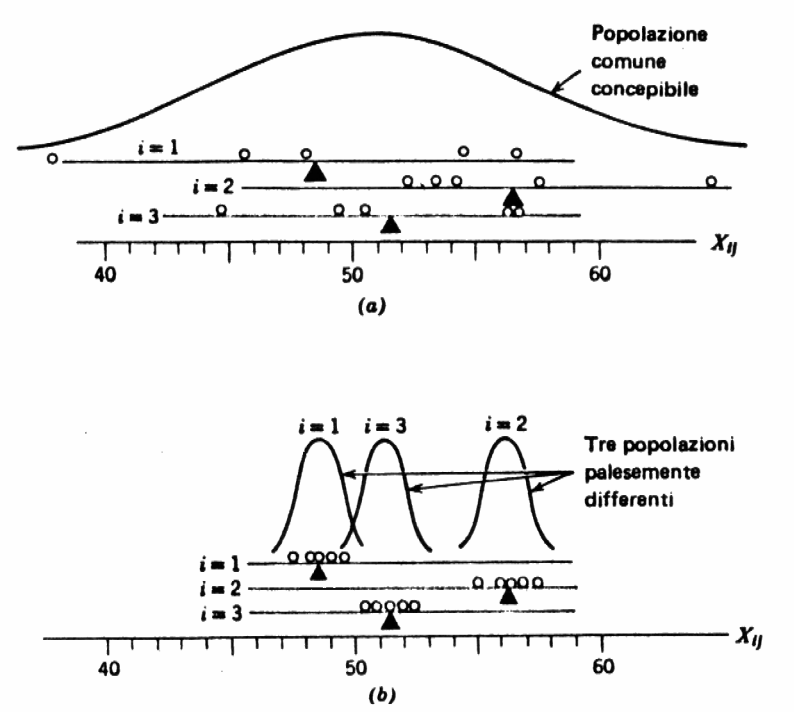


Fig. 3.10 (a) Grafico della Tabella 3.5; (b) Grafico della Tabella 3.3

Allo stesso modo calcoliamo la varianza della fluttuazione casuale entro il secondo (s_2^2) e il terzo campione (s_3^2). La media aritmetica semplice di queste varianze, che vale

$$s_p^2 = \frac{1}{r}\sum_{i=1}^{r} s_i^2 = \frac{0{,}52 + 0{,}87 + 0{,}25}{3} = 0{,}547 \qquad (3.12)$$

si assume come misura della fluttuazione casuale, e viene chiamata "varianza comune". Si noti che da ciascuno degli *r* campioni otteniamo una varianza campionaria con (n - 1) gradi di libertà, cosicché la varianza comune s_p^2 ha *r(n* - 1) gradi di libertà.

A questo punto possiamo porci la questione fondamentale consistente nel decidere se $s_{\bar{X}}^2$ è grande relativamente a s_p^2. L'esame del rapporto

$$F = \frac{n s_{\bar{X}}^2}{s_p^2} \qquad (3.13)$$

chiamato rapporto delle varianze, ci aiuta a risolvere la questione. Si noti che si è introdotto *n* nel numeratore in modo che, se H_0 è vera, il rapporto avrà, in media, un valore vicino ad 1; questo dipende dalla relazione che esiste tra la varianza delle medie campionarie e quella della popolazione. Accadrà peraltro che, a causa delle fluttuazioni statistiche, il rapporto stesso risulterà qualche volta superiore e qualche volta inferiore all'unità.

Se H_0 non è vera (e i valori di μ non sono gli stessi), allora $n\,s_{\bar{X}}^2$ sarà relativamente grande in confronto ad s_p^2 e il valore di *F* nella (3.13) risulterà più grande di 1. Formalmente si rifiuta l'ipotesi H_0 se il valore calcolato di *F* risulta significativamente maggiore di 1.

Il test formale di H_0, come del resto qualsiasi altro test, richiede la conoscenza della distribuzione della statistica osservata se H_0 è vera. Tale statistica, che si indica in questo caso con il simbolo *F*, ha una distribuzione che, nel caso particolare sopra esaminato, assume la forma della curva rappresentata nella

Figura 3.11, nella quale abbiamo anche indicato il valore critico $F_{.05}$ che lascia a destra il 5% della distribuzione. Pertanto, se H_0 è vera, vi è solamente una probabilità del 5 % che si possa osservare un valore di F superiore a 3,89; se si ottiene un valore superiore a 3,89 si rifiuta di conseguenza H_0. Natura1mente è anche possibile essere molto sfortunati ed osservare un valore di F superiore a 3,89 pur essendo H_0 vera, preferiamo tuttavia assumere H_0 come falsa.

Per illustrare questo procedimento, consideriamo le tre serie di risultati campionari nelle Tabelle 3.3, 3.4 e 3.5 e in ciascun caso ci chiediamo se le differenze che abbiamo rilevato per la produzione delle macchine siano statisticamente significative. In altre parole, in ciascun caso vogliamo provare H_0 : $\mu_1=\mu_2=\mu_3$ contro l'ipotesi alternativa che non siano uguali.

Per i dati della Tabella 3.4 una valutazione della (3.13) è:

$$F = \frac{ns_{\bar{X}}^2}{s_p^2} = \frac{0,35}{0,547} = 0,64 \tag{3.14}$$

Poiché il risultato è inferiore al valore critico di $F_{.05}$ = 3,89 concludiamo che le differenze osservate tra le medie possono essere spiegate ragionevolmente solo da variazioni casuali. La cosa non sorprende perché i tre campioni della Tabella 3.4 sono stati ottenuti dalla stessa macchina.

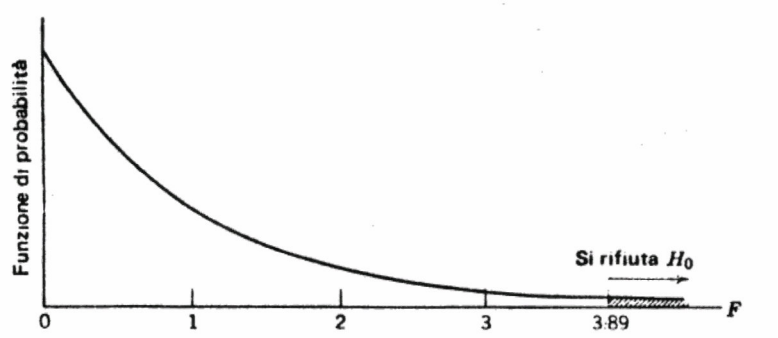


Fig. 3.11 Distribuzione di F quando H0 è vera (con 2,12 gradi di libertà).

Per i dati della Tabella 3.5 il rapporto F è

$$F = \frac{77,4}{35,7} = 2,17 \tag{3.15}$$

in questo caso, la differenza fra le medie campionarie (cioè il numeratore) è molto più grande; ma la stessa cosa accade anche per la variazione casuale (il che si riflette nel denominatore). Anche questa volta il valore di F risulta inferiore al valore critico 3,89.

Infine per i dati della Tabella 3.3, il rapporto F è pari a

$$F = \frac{77,4}{0,547} = 141 \tag{3.16}$$

In quest'ultimo caso, la differenza tra le medie campionarie è molto grande se confrontata con la variazione casuale, il che dà luogo ad un rapporto *F* che eccede di gran lunga il valore critico 3,89, e quindi l'ipotesi H_0 viene rifiutata.

Questi tre test confermano le conclusioni intuitive già sviluppate in precedenza. La Tabella 3.3 fornisce l'unico caso nel quale concludiamo che le popolazioni hanno medie diverse.

La distribuzione di F

Poiché questa distribuzione è importante, è bene esaminarla dettagliatamente. La distribuzione di *F* mostrata nella Figura 3.11 non è che una delle tante possibili, dato che ne esistono diverse in dipendenza dei gradi di libertà *(r - 1)* del numeratore e dei gradi di libertà *r(n - 1)* del denominatore. In questa sede possiamo vederne solo intuitivamente il perché. In effetti, maggiori sono i gradi di libertà nel calcolo del numeratore e del denominatore, più queste due stime di varianze risulteranno vicine al loro valore esatto: di conseguenza, il loro rapporto risulterà più vicino all'unità, come può desumersi dalla Figura 3.12.

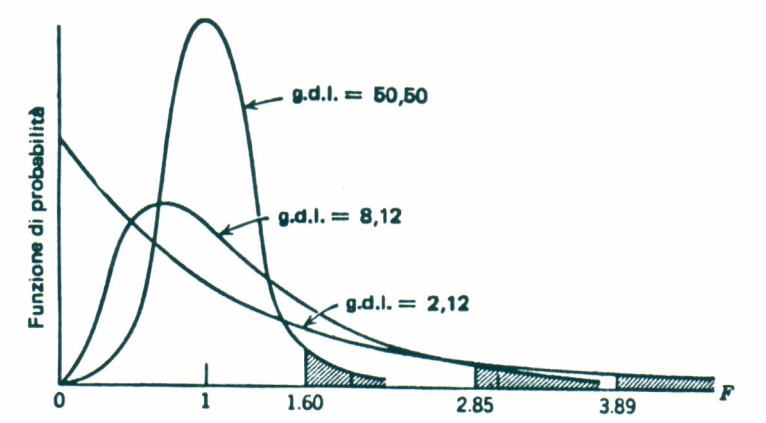


Fig. 3.12 Distribuzione di F con diversi gradi di libertà al numeratore e al denominatore.

Si noti come il punto critico (per il rifiuto di H0) si sposti verso 1 quando aumentano i gradi di libertà.

Si potrebbe compilare tutto un insieme di tabelle di F, ciascuna corrispondente ad una diversa combinazione di gradi di libertà. In pratica, però, ciò non appare necessario dato che in genere si richiedono solamente i valori critici al 5% e all'1%. Come risultato di un test di anova, a volte, al posto di F viene fornito p, la probabilità di osservare un valore di F maggiore di quello effettivamente calcolato.

La tabella ANOVA

In questa sezione è sintetizzato il modo con cui vengono effettuati i calcoli di cui si è appena parlato. Il modello è riassunto nella Tabella 3.6 e nella colonna 2 viene assunta l'ipotesi che tutti i campioni siano estratti da popolazioni normali con la stessa varianza σ^2, ma, ovviamente, con medie che possono essere o no uguali (sono proprio le possibili differenze fra le medie che dobbiamo esaminare).

I calcoli che ne risultano vengono esposti nella Tabella 3.7, chiamata tabella ANOVA (abbreviazione per ANalysis Of VAriance). Nella prima riga sono raccolti i calcoli per il numeratore di F, nella seconda riga le elaborazioni per il denominatore; nella parte (b) di questa stessa tabella sono riportati i valori per l'esempio specifico delle tre macchine della Tabella 3.3.

(1) *Popolazione*	(2) *Distribuzione ipotizzata*	(3) *Valori campionari osservati*
1	$N(\mu_1,\sigma^2)$	X_{1j} $(j=1 \ldots n)$
2	$N(\mu_2,\sigma^2)$	X_{2j} $(j=1 \ldots n)$
3	$N(\mu_3,\sigma^2)$	X_{3j} $(j=1 \ldots n)$
.		
.		
.		
In generale:		
i	$N(\mu_i,\sigma^2)$	X_{ij} $(j=1 \ldots n)$

Tab. 3.6 Sommario delle ipotesi

a) Tabella ANOVA in generale

(1) *Fonte di variazione*	(2) *Devianza: somma dei quadrati*	(3) *Gradi di libertà*	(4) *Varianza (MSS)*	(5) *F(rapporto)*
Tra le righe "spiegata" dalle differenze tra le $\overline{X}_i$	$n\sum_{i=1}^{r}(\overline{X}_i - \overline{\overline{X}})^2 = SS_r$	(r-1)	$MSS_r = SS_r/(r-1)$ $= ns^2_{\overline{X}}$	$\frac{\text{varianza spiegata}}{\text{varianza non spiegata}} = F$
Entro le righe; variazione residua, casuale "non spiegata"	$\sum_{i=1}^{r}\sum_{j=1}^{n}(X_{ij} - \overline{X}_i)^2 = SS_u$	r (n-1)	$MSS_u = SS_u/r(n-1)$ $= s_p^2$	
Totale	$\sum_i\sum_j(X_{ij} - \overline{\overline{X}})^2$	(nr - 1)		

b)Tabella ANOVA, per i valori della Tabella 3.3

(1) *Fonte di variazione*	(2) *Devianza*	(3) *Gradi di libertà*	(4) *Varianza*	(5) *F(rapporto)*
Tra le macchine; "spiegata"	154,8	2	77,4	77,4 / 0,547 = 141
Entro le macchine; "non spiegata"	6,56	12	0,547	
Totale	161	14		

Tab. 3.7 a) tabella ANOVA in generale; b) tabella ANOVA per i valori della Tabella 3.3

La tabella ANOVA ci fornisce inoltre due utili controlli intermedi per i nostri calcoli. Il primo riguarda i gradi di libertà della colonna 3. L'altro è relativo alla somma dei quadrati nella colonna 2, poiché la somma dei quadrati *tra* le righe aggiunta alla somma dei quadrati *entro* le righe deve dare come risultato la somma totale dei quadrati. In definitiva:

$$\sum_i \sum_j (X_{ij} - \overline{\overline{X}})^2 = n\sum_i \overline{X}_i - \overline{\overline{X}})^2 + \sum_i \sum_j (X_{ij} - \overline{X}_i)^2 \qquad (3.17)$$

In altre parole la variazione totale è uguale alla somma della variazione spiegata con la variazione non spiegata.

Quando ogni variazione (devianza) viene divisa per i corrispondenti gradi di libertà si ha la varianza. La varianza tra le righe è "spiegata" dal fatto che le righe possono provenire da diverse popolazioni (per esempio, macchine che si comportano in modo diverso). La varianza entro le righe è "non spiegata" poiché dipende dalle variazioni casuali che assumono i valori, variazioni che non possono essere spiegate sistematicamente (dalle differenze nelle macchine). Perciò qualche volta ci si riferisce ad F come ad un rapporto tra varianze.

$$F = \frac{Varianza \quad spiegata}{Varianza \quad non \quad spiegata} \qquad (3.18)$$

Le considerazioni precedenti ci suggeriscono un mezzo possibile per rafforzare il test F. Si supponga che le tre macchine dell'esempio siano sensibili alle differenze di temperatura. Allora si può introdurre esplicitamente la temperatura nella nostra analisi. Se parte delle variazioni non spiegate possono essere ora spiegate dalla temperatura, il denominatore della (3.13) si ridurrà, dando luogo ad un valore di F più grande del precedente, il che ci metterà a disposizione un test più potente per le macchine (cioè saremo in una posizione più forte per rifiutare H0). Ne segue che l'introduzione di altre spiegazioni della varianza ci permetterà di determinare se una specifica causa (quella delle diverse macchine) è importante o meno. Ciò ci conduce all'esame dell'argomento "ANOVA a due fattori".

Analisi della varianza a due fattori

Riferendoci sempre all'esempio delle macchine, vediamo come si possa tenere conto del fatto che parte della varianza comune è dovuta al fattore umano.

Si supponga che le produzioni campionarie nella Tabella 3.5 siano state ottenute da cinque diversi operatori e che ogni operatore produca uno dei valori campionari su ciascuna macchina. In tali condizioni, conviene raggruppare i dati precedenti mediante una classificazione a due caratteri (a seconda della macchina *e* dell'operatore) ed ottenere la Tabella 3.8.

Operatore / *Macchine*	j =1	2	3	4	5	*Media della macchina* $\overline{X}_{i.}$
i = 1	56,7	45,7	48,3	54,6	37,7	48,6
2	64,5	53,4	54,3	57,5	52,3	56,4
3	56,7	50,6	49,5	56,5	44,7	51,6
Media dell'operatore $X_{j.}$	59,3	49,9	50,7	56,2	44,9	$\overline{\overline{X}} = 52{,}2$

Tab. 3.8 Campioni della produzione (X_{ij}) di tre diverse macchine (come nella Tabella 3.5 ma ordinate secondo l'operatore)

E' necessario a questo punto complicare la notazione poiché ci interessa sia la media di ciascun operatore (X_j , media di ciascuna colonna) sia la media di ciascuna macchina (X_i , media di ciascuna riga)[3].

Ora il quadro è più chiaro: alcuni operatori sono efficienti (il primo e il quarto), mentre altri non lo sono. Le macchine dopo tutto non presentano una notevole variabilità poiché si osserva soltanto una grande differenza nell'efficienza degli operatori. Pertanto, se potessimo tenere conto di quest'ultima circostanza, riusciremmo a ridurre la nostra varianza non spiegata (o casuale) al denominatore della (13.18). E poiché il numeratore rimarrà invariato, il rapporto F risulterà di conseguenza così grande da consentirci, forse, di rifiutare l'ipotesi H0. In tale caso, apparirebbe chiaramente che un'altra influenza (differenza negli operatori) sarebbe responsabile della maggior parte delle difficoltà della nostra analisi della varianza della sezione precedente; superando questa difficoltà speriamo di ottenere un test molto più potente per le nostre macchine.

L'analisi appare come un'estensione dell'analisi della varianza (ANOVA) ad un fattore, ed è sintetizzata nella Tabella 3.9.

Naturalmente in questa tabella, la lettera minuscola c rappresenta il numero delle colonne nella Tabella 3.8 e sostituisce n nella Tabella 3.5, mentre, come nel caso precedente, le diverse componenti delle variazioni della seconda colonna hanno per somma la variazione totale in fondo a questa colonna, cioè

$$\sum_{i=1}^{r}\sum_{j=1}^{c}(X_{ij}-\overline{\overline{X}})^2 = c\sum_{i=1}^{r}(\overline{X}_i-\overline{\overline{X}})^2 + r\sum_{j=1}^{c}(\overline{X}_{.j}-\overline{\overline{X}})^2 + \sum_{i=1}^{r}\sum_{j=1}^{c}(X_{ij}-\overline{X}_{i.}-\overline{X}_{.j}+\overline{\overline{X}})^2 \qquad (3.19)$$

Questa formula ci dice che la variazione totale è pari alla variazione delle macchine (righe) sommata alla variazione dell'operatore (colonna) e alla variazione casuale

[3] Il punto indica l'indice rispetto al quale si effettua la sommatoria. Per esempio, il punto sostituisce j in $\overline{X}_{i.} = \frac{1}{n}\sum_{j} X_{ij}$.

(1) *Fonte delle variazioni*	(2) *Devianza; Somma dei quadrati (SS)*	(3) *Gradi di libertà*	(4) *Varianza (MSS)*	(5) *F*
Tra le righe: **Spiegata** dalle differenze tra le macchine; cioè differenze tra le $\overline{X}_{i.}$	$SS_r = c\sum_{i=1}^{r}(\overline{X}_{i.} - \overline{\overline{X}})^2$	r - 1	$MSS_r = \frac{SS_r}{r-1} = cs^2_{\overline{X}_{i.}}$	$\frac{MSS_r}{MSS_u}$
Tra le colonne: **Spiegata** dalle differenze tra gli operatori. Cioè differenze nelle $\overline{X}_{.j}$	$SS_c = r\sum_{j=1}^{c}(\overline{X}_{.j} - \overline{\overline{X}})^2$	c - 1	$MSS_c = \frac{SS_c}{c-1} = rs^2_{\overline{X}_{.j}}$	$\frac{MSS_c}{MSS_u}$
Non spiegata: cioè residuo risultante da fluttuazioni casuali.	$SS_u = \sum_{i=1}^{r}\sum_{j=1}^{c}(X_{ij} - \overline{X}_{i.} - \overline{X}_{.j} + \overline{\overline{X}})^2$	(r - 1) (c - 1)	$MSS_u = \frac{SS_u}{(r-1)(c-1)} = s_p^2$	
Totale	$SS = \sum_{i=1}^{r}\sum_{j=1}^{c}(X_{ij} - \overline{\overline{X}})^2$	rc - 1		

Tab. 3.9 ANOVA a due fattori

Notiamo che la variazione dovuta all'operatore è definita analogamente a quella dovuta alla macchina, con l'unica differenza che, in questo caso, la variazione dovuta all'operatore è data dalla variazione registrata dalle medie per *colonna*. La (3.19) viene stabilita mediante una complessa serie di passaggi, simili a quelli necessari per stabilire la (3.17) nel caso semplice.

Prova delle ipotesi

Avendo scisso nella (3.19) la variazione totale in componenti, possiamo ora verificare se si è prodotta una differenza significativa fra le macchine o fra gli operatori, tenendo conto, in ambedue i test dell'influenza estranea dell'altro fattore.

Iniziamo col verificare l'ipotesi della differenza fra le macchine, costruendo il rapporto

$$F = \frac{Mss_r}{Mss_u} = \frac{Varianza\ \ spiegata\ \ delle\ \ macchine}{Varianza\ \ non\ \ spiegata} \qquad (3.20)$$

il quale, se H0 è vera, ha una distribuzione F. Così se il valore di F osservato, calcolato nella (3.20), supera il valore critico di F possiamo rifiutare l'ipotesi nulla, concludendo che c'è una differenza tra le medie per righe della popolazione. I calcoli sono sviluppati nella Tabella 3.10.

Dalla Tabella 3.10 si ottiene che la (3.20) è pari a:

$$F = \frac{77,4}{5,9} = 13,1 \qquad (3.21)$$

(1) *Fonte di variazione*	(2) *Devianza (SS)*	(3) *Gradi di libertà*	(4) *Varianza (MSS)*	(5) *F*	(6) *F critico*
Tra le macchine	154,8	2	77,4	13,1	4,46
Tra gli operatori	381,6	4	95,4	16,2	3,84
Residuo	47,3	8	5,9		
Totale	583,7	14			

Tab. 3.10 ANOVA a due criteri. (Per i dati si veda Tab.3.8)

Poiché il valore ottenuto supera il valore critico[4] di F (4,46), rifiutiamo l'ipotesi nulla che le macchine siano simili.

Se confrontiamo il risultato ora ottenuto con il test F nella (3.15), in cui non eravamo in grado di rifiutare l'ipotesi nulla, osserviamo che mentre il numeratore rimane invariato, la variazione casuale nel denominatore è molto più piccola, poiché si è tenuto conto degli effetti delle differenze tra gli operatori. Ciò ci ha dato una grande "potenza"[5] in senso statistico, che ci ha permesso il rifiuto dell'ipotesi nulla.

Allo stesso modo potremmo sottoporre a test l'ipotesi nulla che gli operatori lavorino nella stessa maniera. Ancora una volta F e' il rapporto tra una varianza spiegata e una non spiegata, ma questa volta, naturalmente, il numeratore è la varianza stimata attraverso le differenze tra le colonne.

$$F = \frac{Varianza\ \ spiegata\ \ dagli\ \ operatori}{Varianza\ \ non\ \ spiegata} = \frac{Mss_r}{Mss_u} = \frac{95,4}{5,9} = 16,2 \qquad (3.22)$$

In questo caso abbiamo isolato l'azione dovuta alle macchine, perciò abbiamo ottenuto un test più potente per confrontare l'azione degli operatori. Poiché il valore osservato di F è pari a 16,2 ed è superiore al valore critico[6] di F (3,84), rifiutiamo l'ipotesi nulla concludendo che gli operatori in realtà lavorano in modo diverso.

[4] 2 e 8 gradi di libertà, e livello di significatività del 5 %.

[5] A rigor di termini, abbiamo un test più potente poiché abbiamo ridotto la varianza non spiegata; ciò facendo abbiamo guadagnato più di quello che avevamo perso riducendo i gradi di libertà al denominatore di 4.

[6] Diverso dal test precedente poiché ora i gradi di libertà sono 4 e 8.

C'è un argomento che può essere ulteriormente chiarito. Nel test a un fattore, abbiamo calcolato la varianza non spiegata ricercando la variabilità degli *n* valori osservati entro un campione, cioè entro l'intera riga nella Tabella 3.5. In un test a due criteri di classificazione (Tabella 3.8), però, avendo scisso le osservazioni per colonna e per riga, siamo rimasti con una sola osservazione per ciascuna casella: ad esempio, c'è una sola osservazione (57,5) del prodotto ottenuto dall'operatore 4 sulla macchina 2. Non possiamo allora calcolare la variazione entro tale casella. Cosa faremo? Ci chiediamo: "Se non ci sono errori casuali, come potremmo prevedere la produzione dell'operatore 4 sulla macchina 2?" Notiamo incidentalmente che questa è una macchina migliore della media ($\overline{X}_{2.}$ = 56,4) e con un operatore relativamente efficiente ($\overline{X}_{.4}$ = 56,2) e quindi, in ogni caso, dovremmo prevedere un prodotto superiore alla media. Questa osservazione può essere facilmente usata per prevedere $\hat{X}_{2,4}$. In effetti, se stimiamo in ciascuna casella l'elemento casuale come differenza tra il nostro valore osservato (X_{ij}) e il corrispondente valore stimato $\hat{X}_{ij}$, otterremo un insieme d'elementi casuali la cui somma dei quadrati sarà esattamente la variazione non spiegata SS_u (l'ultimo termine nell'equazione (3.19) che appare anche nella colonna 2 della Tabella 3.9); dividendo per i gradi di libertà si otterrà la varianza non spiegata usata nel denominatore di ambedue i test condotti sull'ultimo esempio considerato.

In dettaglio, il valore previsto $\hat{X}_{ij}$ è così definito:

$$\hat{X}_{ij} = \overline{\overline{X}} + \text{correzione dovuta al comportamento della macchina} +$$
$$+ \text{correzione dovuta al comportamento dell'operatore} =$$
$$= \overline{\overline{X}} + (\overline{X}_{i\cdot} - \overline{\overline{X}}) + (\overline{X}_{\cdot j} - \overline{\overline{X}}) \qquad (3.23)$$

Nel nostro esempio

$$\hat{X}_{2,4} = 52,2 + (56,4 - 52,2) + (56,2 - 52,2) = 52,2 + 4,2 + 4,0 = 60,4$$

Così, la previsione del comportamento dell'operatore 4 sulla macchina 2 si calcola correggendo il comportamento medio (52,2) con il grado in cui la macchina è superiore alla media (4,2) e il grado in cui lo è l'operatore (4,0). Semplificando i valori $\overline{\overline{X}}$ nella (3.23):

$$\hat{X}_{ij} = \overline{X}_{i\cdot} + \overline{X}_{\cdot j} - \overline{\overline{X}} \qquad (3.24)$$

e l'elemento casuale, che è la differenza tra il valore teorico e quello osservato, diviene:

$$X_{ij} - \hat{X}_{ij} = X_{ij} - \overline{X}_{i\cdot} - \overline{X}_{\cdot j} + \overline{\overline{X}} \qquad (3.25)$$

Notiamo che questo elemento casuale è il prodotto non spiegato dopo aver introdotto le correzioni per la macchina *i* e l'operatore *j*.

Nel nostro esempio:

$$X_{2,4} - \hat{X}_{2,4} = 57,5 - 60,4 = -2,9 \qquad (3.26)$$

Così questo prodotto osservato è di 2,9 al di sotto del previsto, e deve rimanere non spiegato (risultato delle influenze casuali). La variazione non spiegata (SS_u) viene ad essere uguale alla somma dei prodotti di tutti gli elementi casuali come nella (3.25).

Cenni sull'analisi della varianza a più fattori e sul problema dell'interazione

Abbiamo dunque visto quali sono le differenze tra l'analisi della varianza ad un fattore e quella a due fattori. E' facile, a questo punto, immaginare che, in presenza di più fattori di variabilità dei dati, potranno complicarsi in maniera notevole le formule ma il principio dell'analisi della varianza rimarra lo stesso.

Essendo queste pagine semplicemente a supporto di un lavoro di analisi di dati sperimentali, non ci addentreremo nell'analisi della varianza multifattoriale, che pure abbiamo usato in maniera sistematica nelle analisi che verranno a breve descritte, rimandando ai testi citati in bibliografia per maggiori dettagli. Riteniamo comunque di aver fornito gli elementi essenziali alla comprensione di quanto verrà detto di seguito. Questo stesso discorso è valido anche per il concetto di interazione di cui daremo solo un breve cenno.Sottolineiamo a proposito la difficoltà nel trovare una trattazione esauriente ed approfondita riguardo all'interazione, anche in testi considerati capisaldi della letteratura sull'analisi statistica.

Nel calcolare la produzione prevista $\hat{X}_{ij}$, nell'ultimo esempio fatto, abbiamo supposto che non ci sia interazione tra i due fattori cosa che invece avverrebbe, ad esempio se alcuni operatori lavorassero bene con alcune macchine e non con altre.

La presenza dell'interazione richiede un modello più complesso ed osservazioni in numero maggiore. Se per ogni combinazione dei due fattori sono disponibili n osservazioni, queste ultime possono essere considerate come un campione casuale estratto da una popolazione caratterizzata dai livelli i e j dei fattori ed aventi media μij. Anche in questo caso, il valore della singola osservazione Xijk può essere scomposta come la parte dovuta al primo fattore (la macchina), la parte dovuta al secondo fattore (l'operatore) e la parte dovuta alle fluttuazioni casuali.

Gli effetti che concernono i livelli del singolo fattore sono chiamati effetti principali. Se l'effetto del livello i del primo fattore sul valore atteso di Xijk è costante al variare del livello j del secondo fattore, gli effetti dei due fattori sono additivi. Altrimenti tra i due fattori c'è interazione: l'effetto dei fattori non è la somma dell'effetto del primo fattore e del secondo fattore ed esiste un ulteriore fattore correttivo. Per maggiori dettagli sull'interazione rimandiamo al testo di Cicchitelli (1984).

3.4.3 Misura della correlazione: il test di Spearman

Un problema che spesso si presenta quando si affronta l'analisi di dati sperimentali, è quello di capire se tra due serie di dati relativi a due parametri di un certo evento vi sia correlazione; si vuole capire, cioè, se esiste una relazione diretta o inversa tra i parametri. Ha interesse, inoltre, quantificare il grado di correlazione.

Vi sono diversi test statistici che permettono di dare una risposta alle domande appena formulate; uno di questi è il **test di correlazione di Spearman**.

In questo test il grado di correlazione è indicato dal coefficiente $\mathbf{r_s}$ (Spearman Rank Correlation Coefficient). Il valore di questo coefficiente è sempre compreso tra –1 e +1 ed il grado di correlazione massimo corrisponde a 1 (in modulo) mentre il grado di correlazione minimo corrisponde al valore 0. Un valore positivo del coefficiente indica, inoltre, che, in media, all'aumentare di una grandezza aumenta anche l'altra, mentre un coefficiente negativo è indice di un comportamento esattamente opposto.

I passi da seguire per il calcolo del coefficiente $\mathbf{r_s}$ sono i seguenti:

1. Mantenendo in due vettori distinti (di *n* dati ciascuno) le due serie di dati, per ognuna si calcola un vettore di ranghi secondo le seguenti regole (si veda l'esempio di tabella 3.11):
   a) ha rango ***1*** l'elemento del vettore con il valore più basso, …, rango ***n*** l'elemento del vettore con valore più alto;
   b) se ***k*** valori sono coincidenti essi hanno stesso rango pari alla media aritmetica dei ranghi che avrebbero avuto se fossero stati diversi ma comunque adiacenti rispetto all'ordinamento.
2. Per ogni riga i dei vettori, si calcola la quantità d_i sottraendo al rango relativo al dato della i-esima riga del primo vettore quello relativo al dato della i-esima riga del secondo vettore (tab. 3.11).
3. Si calcola il coefficiente di correlazione secondo la formula

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n^3 - n} \qquad (3.27)$$

Si noti che questo test, diversamente da altri test di correlazione, non si basa direttamente sui valori dei dati ma ne sfrutta l'ordine.

Infine è importante ricordare che, accanto al valore di $\mathbf{r_s}$, il test fornisce anche il valore p (come per l'ANOVA) che indica se il valore del coefficiente di correlazione trovato è statisticamente significativo.

parametro X		parametro Y			
score	rank	score	rank	d_i	
31	3	79	7	-4	
40	9.5	92	10	-0.5	
26	1	74	3	-2	
33	4.5	78	5.5	-1	
39	8	82	8	0	
40	9.5	86	9	0.5	
37	7	77	4	3	***rs = +0.66***
33	4.5	78	5.5	-1	
35	6	72	1	5	
30	2	73	2	0	

Tab. 3.11 Esempio di calcolo dei vettori dei ranghi e differenze i-esime tra i ranghi della riga i, allo scopo finale del calcolo di r_s.

3.4.4 Criteri di classificazione

Nell'analisi statistica di dati sperimentali, dopo aver investigato su quali tra i parametri presi in considerazione influenzino significativamente il fenomeno che si sta studiando, ci si può porre la domanda se sia possibile classificare i dati rispetto al fenomeno stesso dal valore di qualcuno dei parametri e con quale precisione. In altre parole può avere interesse la misura della separabilità dei dati in due o più gruppi rispetto al fenomeno. Considereremo il solo caso di classificazione in due gruppi. Facciamo un semplice esempio per chiarire quanto appena detto. Supponiamo di misurare l'altezza di un certo numero di persone, uomini e donne. Supponiamo quindi di trovare, ad esempio applicando ai dati un test di ANOVA, che l'altezza di un individuo è significativamente dipendente dal sesso. A questo punto ci si può chiedere se sia possibile e con quale precisione, dedurre il sesso di una persona conoscendo la sua altezza.

Il criterio di classificazione preso in considerazione nel presente lavoro è il criterio MLC (Maximum Likelihood Criterion o Criterio di massima probabilità).

Si suppone che le misure dei parametri di un certo insieme omogeneo, siano statisticamente descrivibili tramite una gaussiana, con un valore medio m e una varianza σ^2. Riportiamo di seguito l'espressione d.d.p della gaussiana :

$$p(x) = \frac{1}{\sigma \cdot \sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}} \qquad (3.28)$$

Questa ipotesi è molto ragionevole ogniqualvolta si tenti di descrivere un qualsiasi fenomeno naturale e non rappresenta perciò in alcun modo una limitazione. Il criterio MLC prevede, come misura della separabilità, il calcolo della percentuale di errori commessi operando una classificazione a posteriori di ciascuno dei dati nei due gruppi, secondo un criterio di massima verosimiglianza.

Si procede come segue:

1. Si dividono i dati in due gruppi a seconda dell'aspetto su cui si vuole basare la classificazione (nel nostro esempio, il sesso).
2. Si calcolano *m* e σ delle gaussiane relative ai due gruppi (nel nostro esempio uomini e donne).
3. Si classifica ciascuno dei dati come appartenente ad un gruppo o all'altro a seconda di quale delle funzioni gaussiane relative ai due gruppi sia maggiore, quando la si valuti in quel punto.
4. Si calcola il numero di errori commessi sfruttando il fatto che si conosce già il gruppo di appartenenza di ogni dato in esame.

Il procedimento è a posteriori proprio per questa ultima ragione: si conosce già il gruppo di appartenenza di ogni dato in esame; inoltre i parametri delle due gaussiane sono calcolati proprio tramite i dati che si vanno a classificare.

La figura 4.9 descrive graficamente il procedimento sopra esposto. Come si vede tale tecnica porta all'individuazione di una frontiera tra i due gruppi. Tutti i dati per cui la misurazione del parametro X dia un valore superiore alla frontiera segnata in figura saranno classificate come appartenenti al gruppo 2, le altre al gruppo 1.

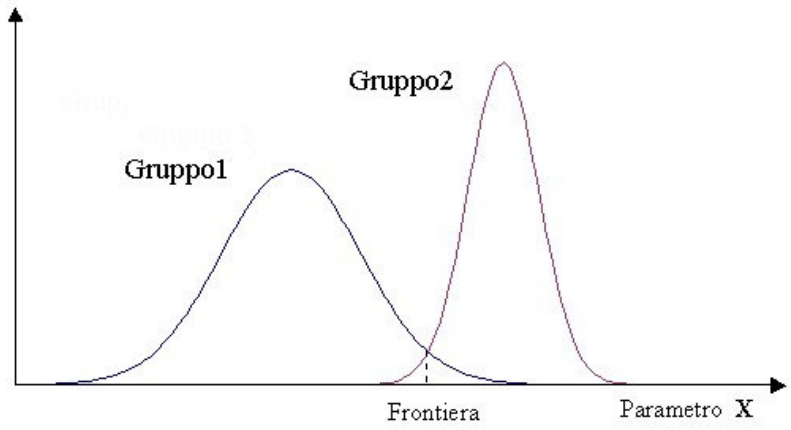


Fig. 3.13 Criterio MLC.

Il criterio MLC è già stato utilizzato per la classificazione in diversi lavori tra quelli del progetto GEMMA. Il test è stato implementato con un programma scritto in Pascal. Si rimanda per maggiori dettagli a A. Vannucci, 1993 e R. Rossetti, 1993.

CAPITOLO 4

L'ANALISI ACUSTICA DELLE CONSONANTI AFFRICATE: METODOLOGIA E RISULTATI

INTRODUZIONE

Nei primi tre capitoli sono state descritte, in maniera molto sintetica, le tecniche utilizzate per studiare il segnale vocale, sia da un punto di vista strettamente teorico che da uno più pratico ed applicativo.

Naturalmente non è stato possibile descrivere dettagliatamente quali sono le metodologie di studio, gli strumenti matematici ed i principi fisici grazie ai quali si riesce ad indagare sull'acustica del segnale vocale per ovvi motivi di spazio. Si è cercato di dare una serie di nozioni di base necessarie alla comprensione del presente lavoro, che poi il lettore può ampliare grazie anche ai testi citati in bibliografia.

Nel presente capitolo verranno esposte le metodologie applicate allo studio delle consonanti affricate italiane e verranno esposti i risultati di tale lavoro. La grande mole di dati misurati in sede sperimentale, insieme alle rispettive medie e deviazioni standard, è stata raccolta nelle appendici A, B e C, dedicate all'analisi nel dominio del tempo, dell'energia e della frequenza. Tali appendici sono parte integrante di tale tesi, che tra l'altro si prefiggeva come scopo anche quello di creare un database utilizzabile per futuri lavori.

Infine nelle appendici D ed E sono raccolti i risultati completi dell'analisi statistica condotta sui dati e i listati dei programmi C utilizzati.

4.1 I PARAMETRI SCELTI PER L'ANALISI ED I CRITERI DI MISURA

La scelta dei parametri temporali, energetici e frequenziali da misurare è stata fatta in base anche ai lavori precedenti, in quanto, come già detto, la presente tesi si pone nell'ambito del più ampio Progetto GEMMA sulla geminazione delle consonanti italiane. Ovviamente si sono adattati i parametri misurati nei precedenti lavori alle particolarità delle consonanti affricate italiane.

4.1.1 Le misure nel dominio del tempo

Ricordando che il database utilizzato è composto di pronunce del tipo VCV (vocale-consonante-vocale, pronuncia singola) e VCCV (pronuncia geminata), si è deciso di misurare le durate dei seguenti segmenti di pronuncia:

- durata della prima vocale, indicata con **V1d**
- durata della fase occlusiva della consonante, indicata con **C1d**
- durata della fase fricativa della consonante, indicata con **C2d**
- durata della seconda vocale, indicata con **V2d**
- durata della pronuncia completa, indicata con **Utd** (utterance duration)

Si fa notare che la divisione della consonante in due non è stata effettuata nei precedenti lavori, mentre qui si è resa necessaria a causa della particolarità delle consonanti affricate di presentare due diverse fasi, la prima occlusiva e la seconda fricativa (vedi Paragrafo 2.2).

Per misurare le durate dei singoli fonemi, si è dovuto scegliere come comportarsi rispetto alle zone di transizione. Considerato che ciò che interessa in questa sede è il confronto tra le durate dei fonemi, si è deciso di non considerare le zone di transizione e inglobare le loro durate in parte sulla vocale ed in parte sulla consonante. In effetti, dato che la media tra le durate di tutti i fonemi della base dati è di 158 ms [1], le zone di transizione rappresentano appena il 5÷10% di un fonema. L'importante quindi è che il criterio adottato per operare la separazione tra vocale e consonante, sia uniforme, così che i risultati finali non risentano di questa approssimazione.

Ricordando che UNICE permette l'individuazione di determinati istanti di tempo tramite l'inserimento di marker posizionabili semplicemente con un click del mouse, la separazione tra i fonemi si è ridotta semplicemente all'inserimento dei suddetti marker direttamente sulla forma d'onda nel tempo. Sono stati di conseguenza individuati i seguenti campioni nel tempo:

1. Campione di attacco della prima vocale (V1 onset)
2. Campione di attacco della consonante (C1 onset) o di fine della prima vocale (V1 offset)
3. Campione di attacco della fase fricativa della consonante (C2 onset)
4. Campione di attacco della seconda vocale (V2 onset) o di fine della consonante (C2 offset)
5. Campione di fine della seconda vocale (V2 offset)

[1] Questo risultato è in accordo con i dati riportati in letteratura per un parlato a velocità normale (6÷8 fonemi per secondo). Infatti, per le pronunce analizzate, risulta un "ritmo fonetico" medio di 6.33 fonemi per secondo.

La seguente figura illustra chiaramente il posizionamento dei suddetti marker.

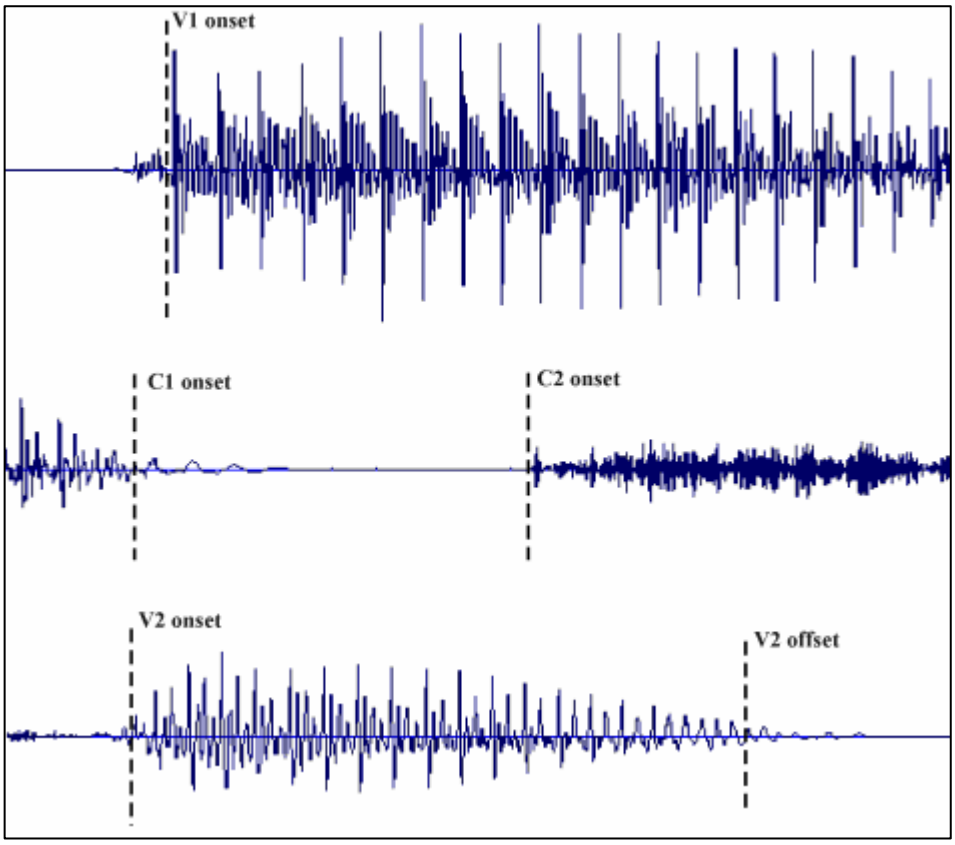


Fig. 4.1 Esempio di segmentazione per una pronuncia singola aτΣa, di un parlatore maschile. Una riga intera corrisponde a circa 200 ms di segnale.

Una volta messi i marker, le loro posizioni vengono memorizzate automaticamente nel file .key che contiene appunto le informazioni relative alla segmentazione di ciascuna pronuncia. Il calcolo successivo delle durate dei fonemi è stato poi effettuato in modo automatico grazie al programma DURATE.C, il cui listato completo si trova nell'appendice D.

Ci sono ora da fare alcune precisazioni riguardo i criteri adottati per il posizionamento dei marker. Innanzi tutto il marker iniziale, ossia V1 onset, non è sempre stato messo immediatamente all'inizio della prima vocale. Infatti, soprattutto per la vocale [a], all'inizio è spesso presente un segnale che presenta delle caratteristiche molto diverse dal resto della vocale, soprattutto per il suo periodo di pitch. Ciò è imputabile al "colpo di glottide" iniziale (che si presenta come un suono sonoro ma molto "sporco"). In questi casi il colpo di glottide è stato escluso dalla vocale. C'è da dire che alcune volte, soprattutto se l'inizio della pronuncia era abbastanza graduale, ascoltando la pronuncia stessa senza il colpo di glottide, questa appariva molto innaturale. Si è deciso, allora, di mantenere in questi rari casi parte dell'attacco iniziale.

Nella scelta di C1 onset le difficoltà maggiori si sono avute con le due consonanti sonore, ossia [δZ,δζ]. Per le consonanti sorde il segnale scende a zero (o quasi) molto rapidamente, e comunque le oscillazioni residue, non facenti parte della vocale, sono sempre distinguibili molto nettamente dalla fine della vocale stessa. Invece per le consonanti sonore, dato che c'è oscillazione delle corde vocali durante la pronuncia di tutto il fonema consonantico, si sono incontrate delle difficoltà maggiori, risolte anche grazie

all'aiuto degli spettri, soprattutto il Narrow Band. Infatti su esso si sono facilmente individuate le formanti caratteristiche della vocale, contro il comportamento in frequenza tipico delle consonanti affricate, caratterizzate dalla sola frequenza di pitch e, al più, da una "formante" di frequenza doppia del pitch.

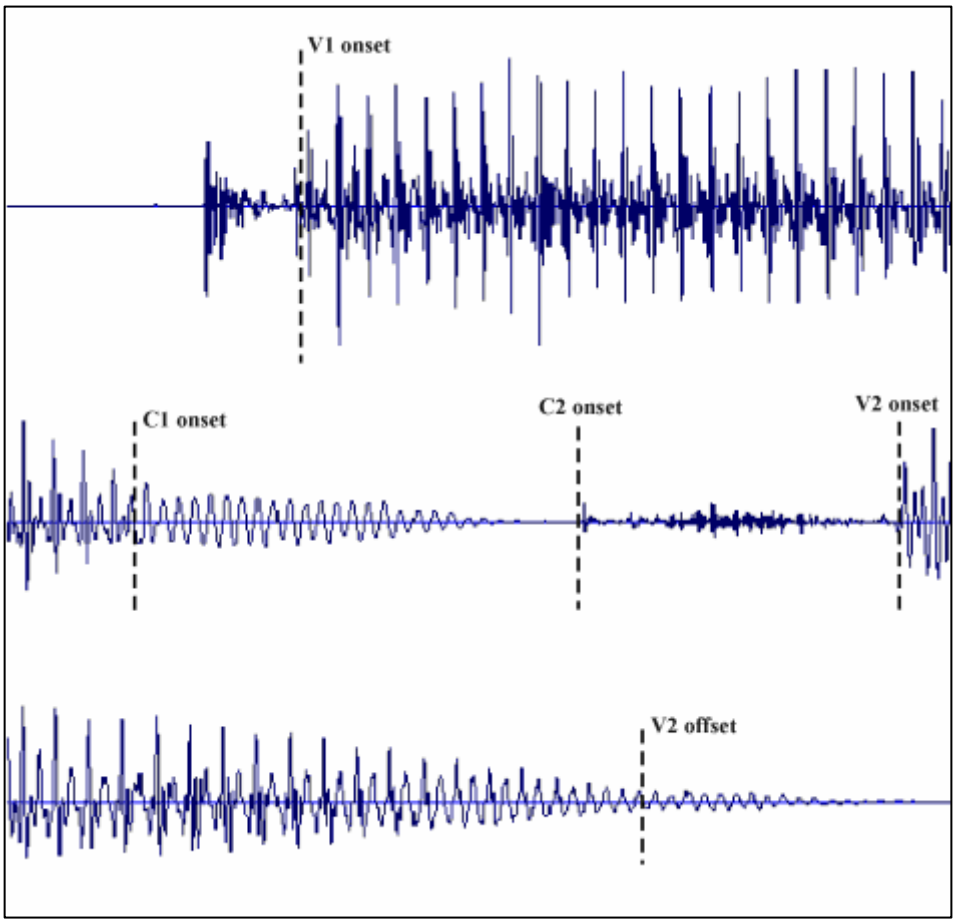


Fig. 4.2 Esempio di segmentazione per una pronuncia singola aδZa di un parlatore femminile. Una riga intera corrisponde a circa 200 ms di segnale.

L'istante di transizione tra la fase occlusiva e la fase fricativa della consonante (ossia tra C1 e C2) è stato quello di più semplice individuazione, in quanto la seconda parte della consonante è caratterizzata da un contenuto in alta frequenza molto evidente. Ciò si nota sia sullo spettro, sia sullo spettrogramma, sia sul segnale nel tempo, che in quasi tutti i casi presenta un forte burst di energia proprio in corrispondenza della transizione, dovuto all'improvviso rilascio dell'occlusione da parte della lingua (vedi Paragrafo 2.2).

Anche la transizione tra C2 e V2 non ha dato grossi problemi nella sua individuazione. Infatti la ricomparsa dell'andamento nel tempo del segnale caratteristico della vocale insieme alle sue proprietà spettrali (vedi andamento delle formanti) è stata quasi sempre abbastanza netta.

Difficoltà ben maggiori ci sono state nell'individuazione del campione di fine pronuncia, ossia V2 offset. Ciò risiede principalmente nel lento decadimento della vocale conclusiva dovuto all'intonazione discendente di fine parola. L'istante di fine pronuncia si è posto generalmente dove il periodo non aveva più la forma tipica della vocale stazionaria e le formanti dalla seconda in poi scomparivano. Accadeva però, relativamente di frequente che per diversi periodi l'ampiezza del segnale tendeva lentamente a zero, senza tuttavia mostrare, da un certo punto in poi, le caratteristiche tipiche di una vocale, neanche sullo

spettrogramma. Ciò è accaduto soprattutto per le vocali [i] e [u]. Si è deciso in questi casi di collocare il campione V2offset nel punto in cui l'ampiezza si abbassava di una certa percentuale (85÷90%) sotto il picco massimo. Si è provveduto poi ad ascoltare l'intera pronuncia fino all'istante scelto come finale, valutando la naturalezza della fine della parola. In base a questo il marker è stato poi spostato a destra o a sinistra per gli aggiustamenti "fini".

Nella Figura 4.2 sono illustrati alcuni dei problemi esposti. Si notino in particolare il colpo di glottide, la gradualità nel passaggio tra V1 e C1 e il lento decadimento alla fine della pronuncia.

Concludiamo questo paragrafo dicendo che, per eliminare le incertezze e rendere le misure il più possibile coerenti tra loro, a distanza di tempo, si sono effettuate nuovamente alcune segmentazioni, per poi confrontarle con quelle precedenti: il risultato è stato che in oltre il 90% dei casi gli istanti presi erano praticamente coincidenti, mentre nel restante 10% le differenze restavano modeste (al massimo uno o due periodi di pitch di variazione), a conferma della bontà delle misure effettuate.

4.1.2 Le misure nel dominio della frequenza

Le misura standard nel progetto GEMMA sono relative alla frequenza di pitch e alle prime tre formanti, con le relative ampiezze. Vediamo ora più in dettaglio quali parametri sono misurati in ogni frame di analisi e le modalità di misura che sono state seguite. Sono state misurate le seguenti grandezze per ogni frame di interesse:

1. F0, A0, F1, A1, F2, A2, F3, A3 nel frame centrale di V1 (V1 center)
2. F0, A0, F1, A1, F2, A2, F3, A3 nel frame finale di V1 (V1 offset)
3. F0, A0, F1, A1, F2, A2, F3, A3 nel frame di transizione tra V1 e C1 (V1 offset 2 C)
4. F0, A0 nel frame iniziale di C1 (solo consonanti sonore) (C onset)
5. F0, A0 nel frame centrale di C1 (solo consonanti sonore) (C1 center)
6. F0, A0 nel frame centrale di C2 (solo consonanti sonore) (C2 center)
7. F0, A0 nel frame finale di C2 (solo consonanti sonore) (C offset)
8. F0, A0, F1, A1, F2, A2, F3, A3 nel frame iniziale di V2 (V2 onset)
9. F0, A0, F1, A1, F2, A2, F3, A3 nel frame centrale di V2 (V2 center)

La Figura 4.3 illustra schematicamente le posizioni dei frame, e le rispettive grandezze misurate, all'interno della pronuncia.

Si noti in particolare nella figura la sovrapposizione per metà dei tre frame posti tra V1 e C1. In questo modo vengono coperti 51,2 ms di pronuncia (512 campioni di segnale)[2]. Anche tra C2 e V2 viene coperto lo stesso intervallo temporale, però con solo due frame di misura (vedi Figura 4.3). Si è utilizzato di un frame in più nella prima transizione in quanto si è visto che quella zona poteva essere di particolare interesse per lo studio del fenomeno della geminazione.

[2] Si ricordi che un frame di UNICE, alla frequenza di campionamento di 10 kHz, è composto da 256 campioni e che, con le impostazioni di *sona zoom* utilizzate nel presente lavoro, c'è un fattore di sovrapposizione del 50% tra frame adiacenti.

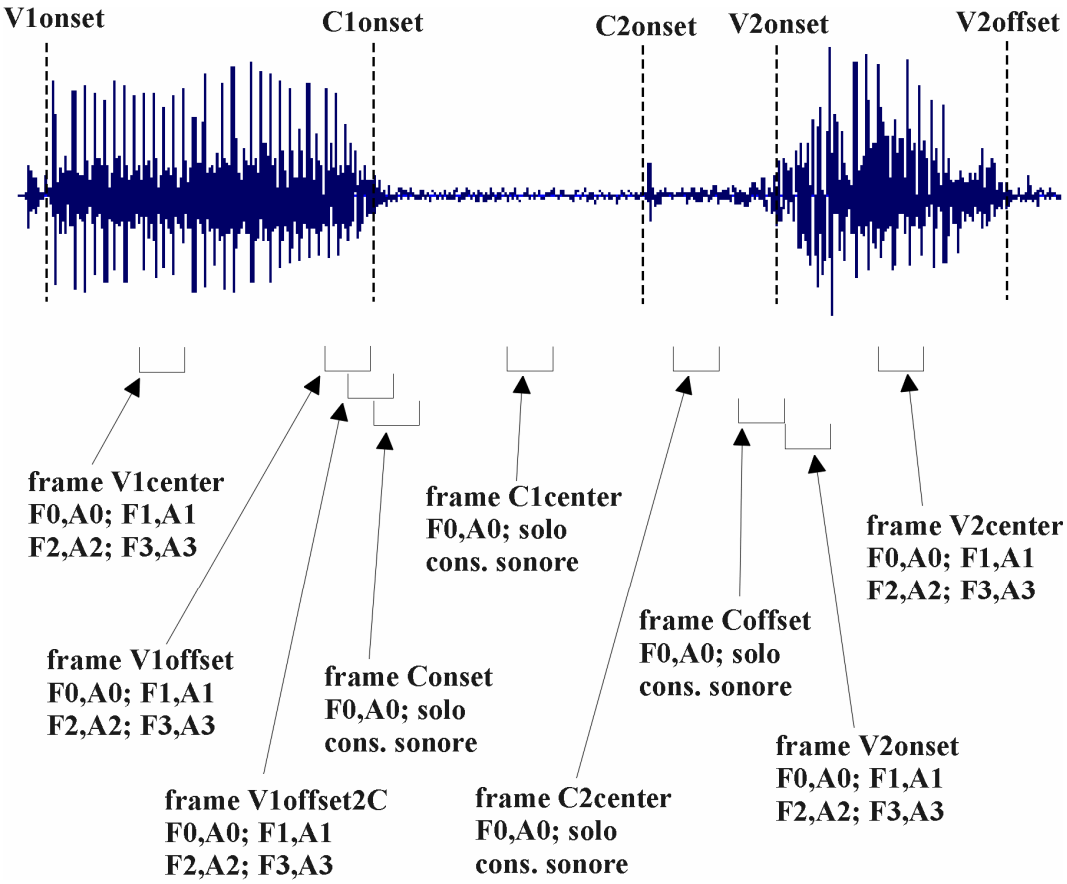


Fig. 4.3 Descrizione grafica dei punti (frame) della pronuncia dove si sono eseguite le misure in frequenza e dei corrispondenti parametri calcolati. Si noti la sovrapposizione tra frame adiacenti nella zona di transizione tra V1 e C1.

Illustriamo ora come sono stati individuati i frame indicati sopra. Per tutti i frame centrali, sia delle vocali che delle consonanti, non è stato preso esattamente il centro del fonema, bensì si è scelto un punto in cui il segnale appariva stazionario all'interno del frame di misura. Il frame V1offset2C è stato scelto in modo che contenesse almeno metà vocale, visto che vi si misurano proprio i parametri caratteristici di una vocale. I due frame adiacenti (V1offset e Conset) sono stati presi di conseguenza. Infine si è fatto in modo che nel frame Coffset ci fossero almeno i ¾ della consonante. Anche qui il frame adiacente V2onset è stato scelto di conseguenza.

Si fa notare che le misure sono state effettuate dopo la segmentazione nel tempo, di conseguenza il posizionamento dei frame di misura è risultato sempre univoco, in quanto si sono prese come riferimento le posizioni dei marker.

Vediamo ora i criteri utilizzati per effettuare le misure in frequenza:

- Per l'individuazione del pitch si è fatto uso dell'algoritmo di calcolo automatico di cui dispone UNICE. A volte (soprattutto per la [i] e la [u]) si notavano forti discontinuità nell'andamento del pitch, con salti tra un frame e l'altro dell'ordine delle decine di Hz, segno evidente che in quei casi il suddetto algoritmo non riusciva a calcolarlo correttamente. Si è allora ricorso al metodo più sicuro, individuando il periodo sulla forma d'onda nel tempo e calcolando il pitch come inverso del suddetto periodo. A volte, non è stato possibile trovare in due periodi adiacenti due picchi puliti che lo individuassero in maniera esatta, pur essendo evidente che il periodo terminava. In questi casi, si

è calcolata la distanza temporale tra picchi distanti tra loro più di un periodo e poi si è diviso per il numero dei periodi presi in considerazione (una sorta di periodo medio a breve termine tra due o tre adiacenti).

- L'ampiezza del periodo è stata misurata sulla prima armonica dello spettro NB. Si fa notare che la frequenza di pitch non è stata misurata direttamente sullo spettro in quanto UNICE, come già precisato, dispone du una risoluzione in frequenza di poco inferiore a 40 Hz, assolutamente troppo grande per l'armonica fondamentale ma sufficiente per le armoniche superiori.
- Per il calcolo delle prime due frequenze formanti e delle loro ampiezze ci si è serviti contemporaneamente delle informazioni derivanti dallo spettrogramma (in modalità WB) e dallo spettro NB[3]. Il primo era utile per visualizzare con un solo colpo d'occhio l'andamento delle formanti durante tutta la pronuncia (migliore risoluzione temporale), mentre il secondo era indispensabile per calcolare i picchi delle formanti con precisione (migliore risoluzione in frequenza). La F1 e la F2 sono state spessissimo individuate con una probabilità di sbagliare minima. Nei casi di [a] e [u] le due formanti sono vicine e questo ha richiesto maggiore attenzione, soprattutto nel caso della [a]. In questi casi ci si è aiutati con lo spettro LPC e con l'andamento delle due formanti nei frame adiacenti e addirittura in tutta la pronuncia.
- Il calcolo della F3 è stato più difficoltoso sia perché a quelle frequenze la variabilità è più alta, sia perché, altrettanto spesso, diversi erano i picchi di intensità confrontabile. A parte questo, valgono le stesse considerazioni fatte in precedenza per F1 e F2.

Anche per le misure in frequenza, come per le segmentazioni nel tempo, si sono misurati una seconda volta tutti i parametri di buona parte delle prime pronunce anallizzate, sia per verificare che non si fossero commessi errori, sia per ricontrollare i valori delle grandezze dopo aver acquisito una buona esperienza. Nella quasi totalità dei casi non si sono corretti i valori in quanto ritenuti esatti.

Concludiamo dicendo che le misure complessivamente eseguite in frequenza sono state:
(216 pron. sorde x 40 parametri pron. sorde) + (216 pron. sonore x 48 parametri pron. sonore) = 19008
tutte eseguite manualmente. Per far ciò sono state necessarie diverse settimane di lavoro. Per quanto onerosa, la scelta di misurare manualmente tutte le formanti, senza ricorrere ad algoritmi automatici, è stata necessaria. Infatti, già in lavori precedenti a questo, era stata riscontrata la bassa affidabilità di algoritmi automatici, dovuta al fatto che sono veramente molti i parametri che influenzano la scelta di un picco anziché di un altro come formante, non ultimo l'andamento delle formanti in tutta la pronuncia.

[3] Lo spettro LPC si è rivelato invece di grande aiuto all'inizio dell'analisi in frequenza per imparare a discernere tra tutti i picchi del NB quali erano le formanti vere. In fase di misura vera e propria però, di solito lo spettro NB dava informazioni più precise (non bisogna dimenticare che la tecnica LPC studia un'approssimazione, nel senso dei minimi quadrati, del segnale).

4.1.3 Le misure nel dominio energetico

Il progetto GEMMA prevede un set di parametri energetici standard che sono adottati anche nella presente tesi, con alcuni adattamenti alle particolarità della classe delle consonanti affricate. I parametri misurati sono i seguenti:

1. *Energia totale della prima vocale*, EtotV1, data dalla semplice formula

$$E_{totV1} = \sum_{i=t_1}^{t_2} x^2(i) \qquad (4.1)$$

   dove x(i) è l'iesimo campione del segnale e t1 e t2 sono gli istanti di V1 onset e di V1 offset.
2. *Potenza media della prima vocale*, PmV1, data da

$$P_{mV1} = \frac{E_{totV1}}{t_2 - t_1} \qquad (4.2)$$

3. *Energia della fase occlusiva della consonante* EtotC1, data ancora dalla (4.1) in cui, però, t1 e t2 corrispondono rispettivamente agli istanti C1 onset e C1 offset.
4. *Potenza media della fase occlusiva della consonante* PmC1, calcolata tramite una formula analoga alla (4.2) dove a numeratore figura l'energia di C1
5. .*Energia della fase fricativa della consonante* EtotC2, data dalla (4.1) in cui, t1 e t2 corrispondono rispettivamente agli istanti C2 onset e C2 offset.
6. *Potenza media della fase fricativa della consonante* PmC2, calcolata tramite una formula analoga alla (4.2) dove a numeratore figura l'energia di C2
7. *Energia totale della consonante* EtotC, data dalla somma di quelle calcolate ai punti 3) e 5).
8. *Potenza media della consonante* PmC, data dal rapporto tra EtotC e la durata di tutta la consonante.
9. *Energia istantanea al centro di V1* EiV1cent, data dalla (4.1), ma calcolata in una finestra temporale di 256 campioni posizionata al centro[4] di V1.
10. *Energia istantanea alla transizione V1-C1*, EiV1-C1; la finestra temporale di 256 campioni è centrata questa volta sul campione corrispondente a V1 offset.
11. *Energia istantanea al centro di C1* EiC1cent; la finestra temporale di 256 campioni è posizionata al centro di C1.
12. *Energia istantanea alla transizione C1-C*2 EiC1-C2; la finestra temporale di 256 campioni è centrata sul campione corrispondente a C1 offset.
13. *Energia istantanea al centro di C2* EiC2cent; la finestra temporale di 256 campioni è posizionata al centro di C2.
14. *Energia istantanea alla fine di C2*, EiC2offset; la finestra di 256 campioni è posizionata in modo che l'ultimo campione della finestra temporale sia quello corrispondente a V2 onset.

[4] Diversamente da quanto detto relativamente alla scelta dei frame centrali nell'analisi in frequenza, in questo caso con "centro" si intende proprio che i 256 campioni sono presi a metà del fonema.

Tutti i parametri sono espressi in dB. La Figura 4.4 riassume graficamente i punti della pronuncia dove sono stati valutati i parametri energetici.

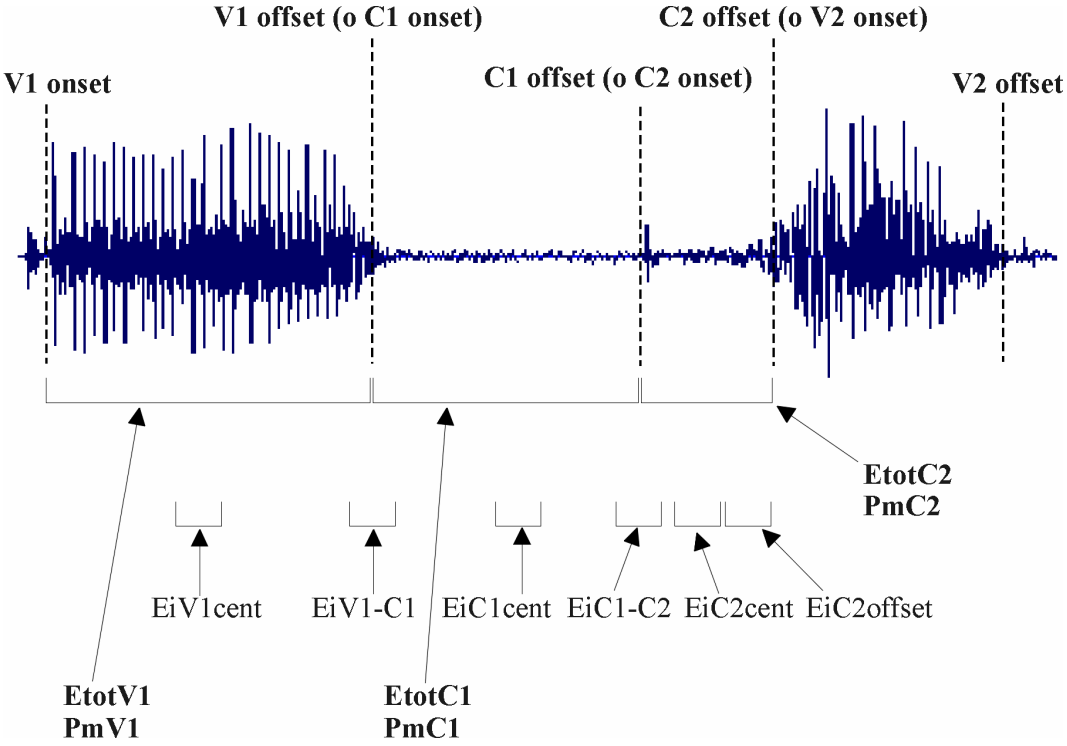


Fig. 4.4 Descrizione grafica dei punti (frame) della pronuncia dove si sono eseguite le misure energetiche e dei corrispondenti parametri calcolati.

I parametri energetici, diversamente da quelli frequenza, sono stati calcolati in maniera automatica con il programma *Energie.C*, scritto appositamente in C. La prima versione del programma è stata scritta da Giovanardi, 1998, per l'analisi delle consonanti fricative. Partendo da questa base sono state effettuate le modifiche necessarie per adattare il programma all'analisi delle consonanti affricate. Il programma *Energie.C* calcola i vari parametri energetici, sia dal domino temporale sia da quello frequenziale, sfruttando nel primo caso le informazioni contenute nei file .sig e .key (in particolare, la segmentazione tramite campioni) e nel secondo caso quelle contenute nei file .fft e .key (la segmentazione tramite frame). Il listato completo del codice è riportato in appendice E.

Concludiamo con una nota di terminologia: tutti i parametri sono stati calcolati sfruttando il concetto di energia a breve termine, esposto nel paragrafo 3.2.2; tuttavia, alcuni parametri misurati all'interno di uno o due frame al massimo sono stati chiamati "istantanei", per distinguerli da quelli "totali" riferiti a tutto un fonema[5].

[5] Per chiarire la terminologia adottata, si fa notare che in genere per istantanei (nell'analisi della voce) si intendono parametri misurati in intervalli di tempo di circa 1 ms (cioè molto più piccoli della durata di un fonema), mentre si dà agli altri il nome di parametri a breve termine. Per l'analisi energetica in esame sarebbero tutti parametri a breve termine, ma l'uso della parola istantanea servirà a distinguere meglio i due gruppi.

4.2 RISULTATI DELL'ANALISI STATISTICA

In questo paragrafo verrà illustrato come i vari test statistici siano stati applicati ai dati misurati sperimentalmente e a quali risultati si è giunti. Ciò verrà fatto per i tre diversi tipi di analisi svolte (ossia nel dominio del tempo, della frequenza e in quello energetico). Salvo diversa indicazione, ogni volta che si menzionerà un test ANOVA multivariato si intenderà che esso è stato effettuato sui tutti i quattro parametri studiati, ossia **sesso** (uomini, donne), **tipo** (singola, geminata), **vocale** (a, i, u), **consonante** (τΣ,δZ,τσ,δζ). Inoltre in alcuni casi, sempre specificati, lo studio sul parametro consonante è stato sostituito da quello su **sorda-sonora**, per valutare se la dipendenza da questo fattore fosse data dalla sonorità o meno della consonante (ricordiamo che τΣ e τσ sono sorde mentre δZ e δζ sono sonore).

Verranno anche illustrati i test di classificazione che si sono adoperati e a quali risultati hanno portato.

4.2.1 Elaborazioni statistiche e risultati dell'analisi nel dominio del tempo

Scopo di questa analisi è:

1. studiare quali fattori influenzano le durate dei fonemi
2. individuare, se vi sono, quali relazioni esistono tra le durate dei vari fonemi all'interno della pronuncia
3. individuare un possibile criterio automatico di distinzione tra una pronuncia singola e la sua corrispondente geminata e valutarne il grado di precisione

Medie e deviazioni standard

Per avere delle prime indicazioni, si sono calcolate le medie e deviazioni standard dei dati raccolti sperimentalmente.

Sono state calcolate le medie e deviazioni standard di tutte le durate dei fonemi presi in considerazione rispetto alle tre ripetizioni di uno stesso parlatore, poi rispetto alle ripetizioni di parlatori dello stesso sesso, poi rispetto alle ripetizioni di tutti i parlatori (indifferentemente dal sesso di appartenenza) e infine statistiche globali rispetto ad una consonante, ad una vocale e infine indifferentemente dalla vocale e consonante. Visti i particolari scopi del presente lavoro si sono lasciate separate le pronunce singole da quelle geminate, in modo da poter sempre fare un confronto relativamente a questo aspetto. Tutti i dati raccolti nel dominio del tempo e le varie tipologie di medie e deviazioni standard effettuate sono raccolte (anche per non appesantire troppo la trattazione) nelle trentaquattro tabelle che compongono l'Appendice A. Tale appendice, come del resto tutte le altre, è parte integrante e fondamentale della presente tesi che, tra l'altro, si proponeva come scopo anche quello di raccogliere e classificare dati che possano essere utili anche a chi, in futuro, vorrà occuparsi di argomenti correlati.

Riportiamo, per comodità, la tabella delle medie "globali" dell'ultima pagina dell'Appendice, che ci sarà utile per delle prime considerazioni.

	V1d	C1d	C2d	V2d	Utd
Singole	149.5	81.8	95.3	128.4	455.0
(StD)	*33.3*	*25.0*	*40.5*	*27.1*	*41.6*
Geminate	111.4	133.3	121.5	125.3	491.6
(StD)	*22.5*	*33.0*	*47.4*	*24.1*	*49.0*

Tab. 4.1 Medie e deviazioni standard (StD) rispetto a tutti i parlatori ,le ripetizioni ,le vocali e le consonanti per il gruppo delle singole (216 pronunce) e per quello delle geminate (216 pronunce). Tutte le misure di durata sono in ms.

Dalle prime osservazioni intuitive sui dati della Tabella 4.4 si possono dedurre i seguenti punti:

- La durata di V1d diminuisce passando dalla pronuncia singola a quella geminata
- Aumentano sia la durata della fase occlusiva della consonante (C1d) sia la durata della fase fricatitiva (C2d). Aumenterà anche, di conseguenza, la durata totale della consonante (data da C1d+C2d)
- Non sembrano esserci variazioni particolarmente evidenti nella durata della seconda vocale
- La durata totale della pronuncia geminata è maggiore della singola, anche se in parte compensata dalla diminuzione di V1d.

Tali affermazioni verranno giustificate e legittimate dall'analisi statistica che è stata effettuata sui dati. Inoltre, grazie a queste prime valutazioni "ad occhio", si è ritenuto opportuno indagare ulteriormente sui seguenti parametri, in base all'obiettivo del presente lavoro che è (anche) quello di fornire un metodo automatico per la distinzione tra pronunce singole e geminate:

- Rapporto tra durata della consonante totale e durata totale della pronuncia (Cd/Utd)
- Rapporto tra durata della consonante totale e durata della prima vocale (Cd/V1d)
- Rapporto tra durata della fase occlusiva della consonante e durata della prima vocale (C1d/V1d)
- Rapporto tra durata della fase fricativa della consonante e durata della prima vocale (C2d/V1d)

Va fatta una precisazione sul modo di eseguire i calcoli: le medie dei vari rapporti appena elencati sono state effettuate prima calcolando tutti i rapporti e poi effettuandone la media; invertendo le operazioni si sarebbe trovato il rapporto tra le medie, più semplice, ma che in questo caso sarebbe stato meno significativo. C'è da aggiungere che per questi parametri verranno evidenziate in particolare le caratteristiche rispetto al fenomeno della geminazione, in quanto sono stati calcolati appositamente e, come vedremo, hanno portato a risultati, sotto questo punto di vista, particolarmente interessanti.

Analisi della varianza

L'indagine sulla significatività delle varie medie calcolate è stata effettuata con il test ANOVA, il quale è ampiamente illustrato nel Capitolo 3. Riportiamo i risultati principali di tale analisi, rimandando il lettore all'Appendice E per i risultati completi. Per illustrare ciò si è deciso, per una maggiore chiarezza e comprensione, di elencare le varie grandezze esaminate, i parametri presi in considerazione per ognuna ed i risultati ottenuti. Per ogni grandezza sono indicati i fattori da cui c'è una forte dipendenza, il livello di significatività (indicato con p) e le medie ottenute (tra parentesi). Come ultima considerazione si ricorda che ogni volta che si è presentata una dipendenza dalla consonante si è eseguita una ulteriore analisi "sorda-sonora" per verificare se fosse questa caratteristica a dare la suddetta dipendenza.

1. **V1d:** presenta una forte dipendenza da: **tipo**, per il quale p=0.0000 (singole=149.5, geminate=111.4); **vocale**, con p=0.0000 (A=136, I=121.2, U=134.2); **consonante**, con p=0.0000 (τΣ=130.7, δZ=144.7, τσ=110.9, δζ=135.5); analisi sorda-sonora, p=0.0000 (sorda=121, sonora=140). V1d è allora indipendente dal sesso, mentre dipende dalla vocale (minore per la [i] che non per la [a] e la [u]), dipende dalla sonorità della consonante con cui è coarticolata (più breve se la consonante è sorda) e dipende, questione che ci interessa maggiormente, dal tipo, risultando più lunga per le pronunce singole.
2. **C1d:** presenta una forte dipendenza da: **tipo**, con un livello di significatività p=0.0000 (singole=81.8, geminate=133.3); **consonante**, con p=0.0000 (τΣ=102.5, δZ=124.3, τσ=95, δζ=108.4); analisi sorda-sonora, p=0.0000 (sorda=99, sonora=116). Anche qui la differenza delle medie tra singole e geminate è statisticamente significativa. Inoltre c'è dipendenza anche dalla consonante (cosa facilmente intuibile visto che stiamo parlando della durata della prima parte della consonante).
3. **C2d:** presenta una forte dipendenza da: **tipo**, con p=0.0000 (singole=95.3, geminate=121.5); **vocale**, con p=0.0000 (A=102.2, I=117.4, U=105.6); **consonante**, con p=0.0000 (τΣ=122.9, δZ=57.1, τσ=156.1, δζ=97.5); analisi sorda-sonora, p=0.0000 (sorda=139, sonora=77). Come per C1d la differenza delle medie tra singole e geminate è significativa, con una durata per le geminate maggiore. E' ovvia la differenza per consonante, con una durata maggiore per le due consonanti sorde, mentre si nota una dipendenza dalla vocale con cui la consonante è coarticolata (anche se le medie non sono poi così differenti).
4. **Cd:** visti i risultati ottenuti per C1d e C2d, non possiamo altro che aspettarci una spiccata dipendenza da **tipo**, con p=0.0000 (singole=177.1, geminate=254.8); **vocale**, con p=0.0001 (A=213.5, I=225.6, U=208.7); **consonante**, con p=0.0000 (τΣ=225.3, δZ=181.4, τσ=251.1, δζ=205.9); analisi sorda-sonora, p=0.0000 (sorda=238, sonora=194). La dipendenza dalla vocale di coarticolazione c'è, anche se la differenza tra le medie non è molto grande (circa 8% di differenza tra [i] ed [u]).
5. **V2d:** si è osservata una forte dipendenza da **sesso**, con un valore per p=0.0000 (uomini=121.4, donne=132.4); **vocale**, con p=0.0206 (A=123.9, I=125.9, U=130.8); **consonante**, con p=0.0000 (τΣ=115.3, δZ=137.1, τσ=115.1, δζ=140); analisi sorda-sonora, p=0.0000 (sorda=115, sonora=139). Come osservato in precedenza non c'è dipendenza della durata della seconda consonante dalla geminazione. La particolarità di V2d è che, pur dipendendo dalla vocale, la differenza tra le tre medie è estremamente contenuta (ciò è giustificato anche dal valore di p al limite della significatività). Anche la dipendenza dal sesso è caratterizzata da una differenza tra uomini e donne piccola.
6. **Utd:** presenta una forte dipendenza da: **sesso**, con p=0.0016 (uomini=466.6, donne=480); **tipo**, con p=0.0000 (singole=455, geminate=491.6). Per quanto riguarda la geminazione possiamo dire che l'aumento di Cd si riflette in un aumento di Utd, anche se in parte "attenuato" dalla parziale compensazione dovuta alla diminuzione di V1d.

	V1d	C1d	C2d	V2d	Utd		V1d	C1d	C2d	V2d	Utd		V1d	C1d	C2d	V2d	Utd
	160.0	73.1	100.9	112.3	446.3		137.4	64.0	122.4	104.6	428.4		163.6	66.0	103.7	131.7	465.0
(StD)	*27.6*	*34.7*	*20.5*	*19.6*	*43.8*	*(StD)*	*20.8*	*29.2*	*16.2*	*17.9*	*29.8*	*(StD)*	*27.4*	*37.9*	*24.0*	*23.7*	*32.0*
	113.2	137.8	128.7	107.5	487.2		99.3	122.8	158.4	110.7	491.3		110.9	151.1	123.0	125.0	509.9
(StD)	*19.2*	*13.9*	*28.1*	*12.2*	*29.3*	*(StD)*	*17.9*	*20.4*	*26.1*	*21.0*	*37.5*	*(StD)*	*25.4*	*39.4*	*24.7*	*22.4*	*51.7*
	169.0	92.0	49.1	142.3	452.3		166.7	95.9	52.6	141.6	456.8		173.5	85.7	44.1	146.1	449.5
(StD)	*20.6*	*18.9*	*13.6*	*26.1*	*47.4*	*(StD)*	*28.3*	*17.5*	*15.7*	*30.6*	*53.4*	*(StD)*	*32.1*	*21.1*	*16.5*	*26.5*	*45.0*
	127.3	156.1	61.5	125.9	470.9		111.7	162.1	74.1	129.4	477.3		120.2	154.0	61.3	137.3	472.8
(StD)	*16.0*	*17.7*	*11.0*	*15.9*	*42.2*	*(StD)*	*21.3*	*28.2*	*25.5*	*30.6*	*56.6*	*(StD)*	*21.6*	*21.3*	*20.8*	*29.9*	*67.7*
	121.3	89.6	129.8	109.9	450.6		106.7	84.4	149.6	109.7	450.4		133.2	73.3	140.7	115.3	462.5
(StD)	*23.3*	*11.0*	*34.0*	*23.1*	*37.0*	*(StD)*	*25.9*	*20.2*	*31.3*	*18.1*	*32.2*	*(StD)*	*30.6*	*26.9*	*22.4*	*16.3*	*41.1*
	106.0	112.2	167.0	117.4	502.6		94.5	114.0	171.0	123.2	502.7		103.8	96.3	178.8	115.1	493.9
(StD)	*18.7*	*18.8*	*22.0*	*20.6*	*43.5*	*(StD)*	*17.9*	*31.4*	*34.7*	*22.8*	*48.0*	*(StD)*	*21.9*	*20.4*	*19.4*	*15.8*	*40.4*
	163.4	89.9	78.6	139.7	471.7		148.4	85.9	90.9	148.1	473.4		150.8	81.6	80.9	139.7	453.0
(StD)	*24.7*	*13.5*	*19.3*	*18.9*	*42.9*	*(StD)*	*37.5*	*16.5*	*21.6*	*20.7*	*35.7*	*(StD)*	*23.7*	*18.8*	*18.1*	*23.8*	*44.8*
	127.8	139.8	102.3	136.3	506.2		104.7	136.5	120.2	139.7	501.1		117.7	116.8	112.3	136.4	483.1
(StD)	*24.5*	*35.3*	*23.0*	*29.0*	*57.4*	*(StD)*	*23.9*	*36.4*	*38.1*	*19.0*	*53.0*	*(StD)*	*17.1*	*26.9*	*29.4*	*20.0*	*43.1*

Tab. 4.2 Riepilogo delle misure di durata (in ms): medie (e deviazioni standard) rispetto a tutti i parlatori e a tutte le ripetizioni, eseguite per gruppi appartenenti alla stessa vocale e alla stessa consonante, tenendo separate le singole dalle geminate.

Per quanto riguarda gli altri quattro parametri scelti (e calcolati) per l'analisi temporale si sono ottenuti i seguenti risultati:

7. **Cd/Utd:** si osserva una forte dipendenza da: **tipo**, con p=0.0000 (singole=0.390, geminate=0.519); **vocale**, con p=0.0000 (A=0.449, I=0.476, U=0.438); **consonante**, con p=0.0000 ($\tau\Sigma$=0.475, δZ=0.391, $\tau\sigma$=0.525, $\delta\zeta$=0.426); analisi sorda-sonora, p=0.0000 (sorda=0.500, sonora=0.409).
8. **Cd/V1d:** forte dipendenza da: **tipo**, con p=0.0000 (singole=1.305, geminate=2.416); **vocale**, con p=0.0000 (A=1.734, I=2.105, U=1.742); **consonante**, con p=0.0000 ($\tau\Sigma$=1.942, δZ=1.380, $\tau\sigma$=2.445, $\delta\zeta$=1.674); analisi sorda-sonora, p=0.0000 (sorda=2.194, sonora=1.527).
9. **C1d/V1d:** presenta una forte dipendenza da: **tipo**, con p=0.0000 (singole=0.592, geminate=1.249); **vocale**, con p=0.0017 (A=0.895, I=1.002, U=0.864).
10. **C2d/V1d:** forte dipendenza da: **tipo**, con p=0.0000 (singole=0.713, geminate=1.167); **vocale**, con p=0.0000 (A=0.839, I=1.103, U=0.878); **consonante**, con p=0.0000 ($\tau\Sigma$=1.027, δZ=0.428, $\tau\sigma$=1.521, $\delta\zeta$=0.783); analisi sorda-sonora, p=0.0000 (sorda=1.274, sonora=0.605).

Le dipendenze da alcuni fattori piuttosto che da altri risultano facilmente spiegabili se si vanno ad osservare le grandezze da cui derivano questi quattro rapporti. Ad esempio, pensando proprio al fenomeno della geminazione, dato che V1d, C1d, C2d, Cd e Utd dipendono da tipo, anche le grandezze da esse derivate dovranno presentare un comportamento analogo, soprattutto se, come nel nostro caso, si

considerano nei rapporti a numeratore le grandezze che crescono con la geminazione e a denominatore quelle che invece decrescono (ad eccezione di Cd/Utd). Questa è una spiegazione del perché, come vedremo dettagliatamente in seguito, queste grandezze sono migliori dal punto di vista del riconoscimento automatico di una pronuncia singola da una geminata basandosi su una analisi nel dominio del tempo.

Nella Tabella 4.2 sono riportate le medie e le deviazioni standard (StD) rispetto a tutti i parlatori e alle ripetizioni che possono essere utili per avere, nella maggior parte dei casi, un immediato riscontro dei risultati dei test di anova appena descritti.

Vista la significatività delle grandezze temporali esaminate rispetto al fenomeno della geminazione, se ne calcoleranno ed esporranno ora le differenze (assolute e percentuali).

$$\Delta V1d = V1d_{gem} - V1d_{\sin} = -38.1$$
$$\Delta V1d\% = \frac{\Delta V1d}{V1d_{\sin}} * 100 = -25.5\% \quad (4.3)$$

$$\Delta C1d = C1d_{gem} - C1d_{\sin} = 51.5$$
$$\Delta C1d\% = \frac{\Delta C1d}{C1d_{\sin}} * 100 = 63\% \quad (4.4)$$

$$\Delta C2d = C2d_{gem} - C2d_{\sin} = 26.2$$
$$\Delta C2d\% = \frac{\Delta C2d}{C2d_{\sin}} * 100 = 27.5\% \quad (4.5)$$

$$\Delta Utd = Utd_{gem} - Utd_{\sin} = 36.6$$
$$\Delta Utd\% = \frac{\Delta Utd}{Utd_{\sin}} * 100 = 8\% \quad (4.6)$$

Test di correlazione sulle durate dei fonemi

Nell'analisi sulle durate dei fonemi appena conclusa si è trovato che la durata di V1 si comporta in maniera inversa rispetto a C1 e C2 passando da una pronuncia singola a una geminata. Infatti V1 tende ad accorciarsi e C1 e C2 ad allungarsi (vedi anche le differenze riportate nelle formule 4.3, 4.4 e 4.5). E' allora naturale chiedersi quale correlazione esista tra le durate dei suddetti fonemi e se questa sia imputabile alla geminazione. Per fare ciò si è ricorso, lo ricordiamo, al test di correlazione di Spearman, la cui spiegazione di come venga calcolato e di come vada interpretato è riportata nel Paragrafo 3.4.3.

Si sono calcolati i coefficienti $\mathbf{r_s}$ prima solo per le pronunce singole, poi per quelle geminate e infine per tutte le pronunce insieme, per poter valutare se le correlazioni tra le durate dei fonemi siano o no imputabili alla geminazione. I valori di $\mathbf{r_s}$ e la loro significatività sono riportati in matrici di correlazione che presentano la caratteristica di avere la diagonale principale unitaria (correlazione di una grandezza con sé stessa =1) e di essere simmetriche (per questo sono riportati i valori solo sulla diagonale inferiore).

Vediamo ora i risultati separando le pronunce singole da quelle geminate.

Pronunce singole

	V1d	C1d	C2d	V2d
V1d	**1**			
C1d	**-0.2708**	**1**		
C2d	**-0.4321**	**-0.1999**	**1**	
V2d	**0.4956**	0.0292	**-0.4980**	**1**

Pronunce geminate

	V1d	C1d	C2d	V2d
V1d	**1**			
C1d	0.0369	**1**		
C2d	**-0.3207**	**-0.5095**	**1**	
V2d	**0.5445**	0.0804	**-0.2382**	**1**

Tab. 4.3 Matrici di correlazione dei coefficienti r_s. Ogni elemento delle matrici rappresenta il coefficiente di correlazione r_s tra la variabile riga e la variabile colonna. Sono presi in considerazione i valori di durata dei fonemi di tutte le pronunce considerando i gruppi singole e geminate separatamente. I valori in grassetto sono quelli statisticamente significativi ($p<0.05$).

Commentiamo brevemente i risultati del test:

1. C'è una debole correlazione negativa tra C1d e V1d per le pronunce singole, mentre non esiste per le geminate
2. C'è correlazione negativa anche tra C2d e V1d, in questo caso un po' più forte che tra C1d e V2d
3. Esiste una correlazione positiva tra le durate delle due vocali
4. C'è correlazione negativa tra C2d e V2d

Guardiamo ora la tabella delle correlazioni per tutte le pronunce:

Tutte le Pronunce

	V1d	C1d	C2d	V2d
V1d	**1**			
C1d	**-0.4711**	**1**		
C2d	**-0.4739**	-0.0475	**1**	
V2d	**0.4666**	-0.0148	**-0.3623**	**1**

Tab. 4.5 Matrice di correlazione (secondo il coefficiente r_s) tra i valori di durata dei fonemi di tutte le pronunce (singole e geminate assieme). I valori in grassetto sono quelli statisticamente significativi ($p<0.05$).

Dai risultati del test di Spearman si vede che:

5. Esiste una correlazione negativa tra C1d e V1d abbastanza forte
6. C'è una correlazione negativa anche tra C2d e V1d

7. Esiste ancora la correlazione positiva tra V1d e V2d
8. C'è correlazione negativa tra C2d e V2d

La correlazione negativa tra C1d e V1d è allora causata dalla geminazione, in quanto, come già osservato, tale situazione non si presenta considerando le pronunce singole separate da quelle geminate. Questo cambiamento di correlazione così netto non si osserva invece tra C2d e V2d. Infatti, andando a rivedere i risultati dell'analisi della varianza per i due rapporti C1d/V1d e C2d/V2d si possono fare le seguenti considerazioni: entrambi presentano una forte significatività rispetto a **tipo** (p=0.0000), mentre però la differenza tra le medie è di 0.454 (singole=0.713, geminate=1.167) per C2d/V1d, questa sale al valore 0.657 (singole=0.592, geminate=1.249) per C1d/V1d.In base a queste considerazioni si può trarre la conclusione che nella geminazione si tende ad allungare più C1d che non C2d rispetto alla durata della prima vocale. Questo si riflette, come vedremo tra poco, in migliori risultati riguardo alla classificazione automatica delle pronunce usando la grandezza C1d/V1d piuttosto che C2d/V1d.

La correlazione positiva tra V1d e V2d può essere probabilmente imputabile alla struttura ritmica del parlato che produce delle compensazioni tra le lunghezze dei fonemi (ossia se si tende a parlare più lentamente si allungano mediamente le durate di tutti i fonemi).

Per quanto riguarda C2d e V2d possiamo dire che la correlazione esistente non dipende dalla geminazione in quanto i valori nelle tre tabelle sono confrontabili e comunque non molto alti.

Classificazione delle pronunce sulla base delle durate dei fonemi

Come ultimo passo dell'analisi vogliamo vedere se sia possibile classificare efficacemente il tipo delle pronunce sulla base dei parametri di durata che sono risultati significativi per la geminazione. Abbiamo utilizzato a tal proposito il Maximum Likelihood Criterion, già introdotto nel paragrafo 3.4.4. La classificazione è stata fatta su tutte le pronunce, poi dividendo uomini e donne, poi dividendo le consonanti e infine dividendo le vocali. I parametri sulla base dei quali sono state effettuate le classificazioni sono nell'ordine Cd/Utd, Cd/V1d, C1d/V1d, C2d/V1d, V1d, C1d, C2d, Cd. Pur essendo significativo anche Utd, non è stato incluso nella Tabella 4.6 in quanto ha dato risultati pessimi per tutte le classificazioni.

Analizzando la tabella si possono fare le seguenti osservazioni:

1. Considerando le pronunce tutte insieme la percentuale di errori è insoddisfacente, raggiungendo un valore minimo pari al 13.89% considerando il rapporto C1d/V1d.
2. Conducendo l'analisi sulle sole pronunce maschili le cose migliorano un po' arrivando a compiere il 9.72% di errori sempre considerando C1d/V1d.
3. Per le donne ci sono problemi ancora maggiori dato che non si scende sotto il 16.67% di errori.
4. Nella divisione per consonanti si nota un fatto molto particolare: risultati in assoluto ottimi si ottengono per le due consonanti alveopalatali [τΣ,δZ], con una percentuale di errori pari rispettivamente a 4.63% (grandezza Cd) e 0% (grandezza C1d/V1d); decisamente peggio va per le due consonanti dentali [ts, dz]. Per la prima non si fa meglio del 17.59% di errori mentre un po' meglio si comporta la [dz] con il 10.19%. Per entrambe il parametro migliore è Cd.
5. Nell'analisi eseguita rispetto a ciascuna vocale i risultati migliori sono i seguenti: [a], 10.42% di errori nella grandezza C1d/V1d; [i], 15.28% nella grandezza C1d/V1d; [u], 12.50% anche qui in C1d/V1d.

CRITERIO MLC												
	Cd/Utd			Cd/V1d			C1d/V1d			C2d/V1d		
	EPP	Errori	Err. %	EPP	Errori	Err. %	EPP	Errori	Err. %	EPP	Errori	Err. %
Tutte	0.43	86	19.91	1.6	72	16.67	0.76	60	13.89	0.78	131	30.32
Uomini	0.42	40	18.52	1.68	30	13.89	0.69	21	9.72	0.78	66	30.56
Donne	0.43	45	20.83	1.55	40	18.52	0.85	36	16.67	0.78	65	30.09
♦♠	0.44	15	13.89	1.6	9	8.33	0.65	9	8.33	0.87	20	18.52
♋	0.37	6	5.56	1.26	3	2.78	0.91	0	0.00	0.36	17	15.74
♦♦	0.49	31	28.70	1.9	26	24.07	0.77	29	26.85	1.41	28	25.93
♒	0.42	15	13.89	1.45	14	12.96	0.76	13	12.04	0.75	18	16.67
a	0.43	23	15.97	1.29	19	13.19	0.74	15	10.42	0.78	38	26.39
i	0.48	32	22.22	1.75	25	17.36	0.92	22	15.28	1.13	41	28.47
u	0.44	27	18.75	1.57	21	14.58	0.68	18	12.50	0.87	41	28.47
	V1d			C1d			C2d			Cd		
	EPP	Errori	Err. %	EPP	Errori	Err. %	EPP	Errori	Err. %	EPP	Errori	Err. %
Tutte	136	101	23.38	102	73	16.90	129	157	36.34	208	76	17.59
Uomini	135	47	21.76	93	26	12.04	146	82	37.96	215	36	16.67
Donne	135	54	25.00	112	41	18.98	129	67	31.02	208	37	17.13
♦♠	137	20	18.52	95	10	9.26	139	30	27.78	224	5	4.63
♋	136	15	13.89	125	5	4.63	67	34	31.48	166	8	7.41
♦♦	124	34	31.48	91	27	25.00	157	28	25.93	244	19	17.59
♒	153	28	25.93	100	17	15.74	108	28	25.93	201	11	10.19
a	145	32	22.22	104	17	11.81	107	53	36.81	198	23	15.97
i	122	33	22.92	102	24	16.67	139	48	33.33	235	28	19.44
u	135	27	18.75	94	24	16.67	129	54	37.50	207	24	16.67

Tab. 4.6 Criteri MLC per la classificazione del tipo, condotti sulla base di tutti i parametri di durata statisticamente significativi. EPP rappresenta il punto di equiprobabilità o di separazione delle due gaussiane. Le unità di misura di EPP sono coerenti con quelle delle grandezze cui è riferito (ms per quelle "assolute", numeri puri per i rapporti).

In conclusione si fa notare che la grandezza che mediamente fornisce i migliori risultati è C1d/V1d, come si poteva anche supporre guardando il livello di significatività nell'analisi della varianza rispetto a **tipo** e i valori medi che essa assume nelle pronunce singole e geminate (valori medi abbastanza diversi tra loro in rapporto alle deviazioni standard delle misure eseguite).

4.2.2 Elaborazioni statistiche e risultati dell'analisi in frequenza

Scopo dell'analisi in frequenza è:

1. indagare quali fattori influenzano le grandezze in frequenza, con particolare attenzione al fenomeno della geminazione
2. verificare se c'è una dipendenza del pitch e delle formanti dal punto della parola in cui vengono pronunciate
3. valutare se sia possibile fornire un metodo di riconoscimento della pronuncia singola o geminata sulla base delle grandezze in frequenza

Medie e deviazioni standard

Sono state calcolate le medie e deviazioni standard di tutti i parametri frequenziali presi in considerazione rispetto alle tre ripetizioni di uno stesso parlatore, poi rispetto alle ripetizioni di parlatori dello stesso sesso, poi rispetto alle ripetizioni di tutti i parlatori (indifferentemente dal sesso di appartenenza) e infine statistiche globali rispetto ad una consonante, ad una vocale e infine indifferentemente dalla vocale e consonante. Visti i particolari scopi del presente lavoro si sono lasciate separate le pronunce singole da quelle geminate, in modo da poter sempre fare un confronto relativamente a questo aspetto. Tutti i dati raccolti nel dominio della frequenza e le varie tipologie di medie e deviazioni standard effettuate sono raccolte (anche per non appesantire troppo la trattazione) nelle quaranta tabelle che compongono l'Appendice C.

Analisi della varianza

La lettura dei dati si presenta molto più complessa e meno immediata di quanto è stato per l'analisi temporale. Passiamo quindi direttamente ai risultati principali dell'analisi statistica in frequenza rimandando all'Appendice E per tutti i dettagli del caso. Il tipo di analisi effettuata è stato un ANOVA multivariato su ognuno degli 8 parametri (F0, A0, F1, A1, F2, A2, F3, A3) in ciascuno dei frame di analisi.

E' stata dapprima eseguita una analisi statistica senza dividere gli uomini dalle donne per valutare come il sesso influisse sulle grandezze. E' stato necessario eseguire (5 x 8) + (4 x 2) = 48 analisi multivariate. Si è ottenuto che tutte le grandezze sono influenzate dal sesso tranne l'ampiezza della seconda formante A2 in tutti i frame considerati. Tutte le frequenze delle formanti sono risultate più alte nelle donne che negli uomini, come era facilmente prevedibile da considerazioni di tipo acustico-fisiologiche.

A questo punto, vista la forte dipendenza dei parametri frequenziali dal sesso, si sono eseguite due nuove analisi complete (ciascuna composta da 48 analisi multivariate), la prima solo per le donne, la seconda solo per gli uomini. Ricordiamo infine che in ogni caso in cui una grandezza risultava influenzata dalla consonante, è stata eseguita una ulteriore analisi di tipo sorda-sonora (vedi introduzione del presente Paragrafo 4.2). Analizziamo i risultati:

- **Donne:** è emersa una forte dipendenza delle formanti dalla vocale analizzata, e questo è ovvio considerando che è proprio la frequenza delle formanti a caratterizzare una vocale. Soltanto il pitch

non dipende dalla particolare vocale in esame. Inoltre la frequenza fondamentale non dipende in nessun frame di analisi neanche da **tipo** e da **consonante**. Rimanendo focalizzati sulla dipendenza dalla geminazione, nessuna frequenza delle formanti dipende da tipo. Soltanto le ampiezze A2 ed A3 in V1 offset e A2 in V1 offset 2 C dipendono da tipo, risultando più alte (2-3 dB) nelle pronunce geminate. Per quanto riguarda il parametro **consonante** si è visto che la sua influenza sulle formanti è limitata ai frame adiacenti alla consonante stessa (V1 offset 2 C e V2 onset), in particolare dalla caratteristica della consonante di essere sorda o sonora.

- **Uomini:** anche qui si nota ovviamente una forte dipendenza delle formanti dalla vocale considerata. Contrariamente alle donne anche la frequenza fondamentale dipende dalla vocale nei primi frame di analisi della pronuncia, ossia in V1 center, V1 offset, V1offset 2 C e C1 onset. Più precisamente il pitch è più basso per la [a], intermedio per la [i] e più alto per la [u]. Per quanto riguarda il comportamento rispetto alla geminazione, ora sono molti i parametri che dipendono da questo aspetto. Innanzitutto il pitch nei primi quattro frame di analisi (gli stessi elencati sopra) è più alto per le geminate rispetto alle singole (fino a +15 Hz in V1 offset e V1 offset 2 C, corrispondenti ad un aumento di circa l'11%). Anche ora, sempre nei confronti del fattore **tipo**, si notano delle differenze significative nelle ampiezze delle formanti (A1, A2 ed A3), limitatamente alla prima vocale, con un incremento di ampiezza nelle geminate che arriva fino a 4 dB. Infine, riguardo al fattore **consonante**, le variazioni più interessanti riguardano i tre frame vicini alla consonante per cui si notano delle variazioni significative nelle due formanti F1 e F2. Queste variazioni si possono interpretare come dovute alla preparazione all'occlusione della consonante tra V1 e C e alla fase di rilascio dell'occlusione stessa tra C e V2. In particolare sembra che F1 sia maggiormente influenzata dalla caratteristica sorda-sonora, mentre le variazioni di F2, percentualmente più piccole di quelle di F1, non sembrano così legate a questo fattore.

Una ulteriore analisi in frequenza è stata effettuata considerando il frame di analisi come parametro. Si è ritenuto necessario fare ciò in quanto, durante la fase di misura dei dati, si è notato che le formanti si spostavano cambiando il frame di analisi. In pratica si è effettuata una analisi mirata a vedere se le variazioni delle frequenze formanti durante la pronuncia siano statisticamente significative rispetto al fattore tempo. Per fare ciò è stato necessario riordinare tutti i dati in frequenza ed effettuare le analisi della varianza assumendo il frame di misura come parametro. E' stato necessario dividere uomini e donne e le diverse vocali in quanto, in una fase preliminare, si è visto che esistevano forti interazioni tra **frame**, **vocale** e **sesso**. Non è stata fatta l'analisi considerando anche la distinzione su **tipo** in quanto si è già visto che influenza solo la frequenza di pitch (al massimo di 15 Hz) in pochi frame e solo negli uomini e le ampiezze delle formanti al più di 2-4 dB.

Il risultato di tale analisi è che per tutte le grandezze in frequenza considerate (F0, A0, F1, A1, F2, A2, F3 A3) il parametro frame è risultato statisticamente significativo, eccezion fatta per la F3 nella [u] delle pronunce femminili.

Nella Tabella 4.7 sono riportate le medie di tale analisi mentre nelle Figure 4.5 - 4.12 sono graficati tali andamenti per una più chiara ed immediata comprensione dell'andamento delle formanti.

DONNE								
A								
Frame	**F0**	**A0**	**F1**	**A1**	**F2**	**A2**	**F3**	**A3**
V1 center	186.1	11.5	1062.8	47.1	1636.9	41.8	2754.8	32.1
V1 offset	181.4	11.8	907.2	39.4	1731.2	36.1	2841.0	27.7
V1 offset 2 C	178.3	10.9	786.2	32.9	1777.3	31.3	2916.8	23.5
C1 onset	157.4	6.4						
C1 center	146.8	5.1						
C2 center	137.6	6.5						
C2 offset	155.4	6.9						
V2 onset	159.1	7.4	677.6	33.7	1728.0	30.2	3024.1	23.2
V2 center	151.0	6.3	939.5	35.9	1612.4	32.3	3034.4	22.4
I								
Frame	**F0**	**A0**	**F1**	**A1**	**F2**	**A2**	**F3**	**A3**
V1 center	200.6	18.6	400.3	28.8	2792.0	29.1	3566.2	30.8
V1 offset	191.9	15.7	379.2	25.8	2768.7	24.0	3489.5	26.8
V1 offset 2 C	185.1	12.4	364.5	22.6	2743.2	20.0	3444.0	22.5
C1 onset	170.4	8.8						
C1 center	153.2	6.4						
C2 center	130.8	8.7						
C2 offset	145.4	7.2						
V2 onset	158.2	7.3	314.7	28.7	2494.6	24.9	3145.6	25.3
V2 center	155.5	7.2	313.1	32.3	2632.5	24.8	3202.5	25.2
U								
Frame	**F0**	**A0**	**F1**	**A1**	**F2**	**A2**	**F3**	**A3**
V1 center	202.4	17.4	403.5	34.8	756.6	29.3	2861.1	9.3
V1 offset	194.2	15.8	385.1	31.8	989.0	24.4	2816.5	13.0
V1 offset 2 C	189.5	13.8	373.8	28.2	1048.6	19.7	2800.8	13.4
C1 onset	175.3	8.8						
C1 center	151.7	6.5						
C2 center	141.7	7.5						
C2 offset	143.5	7.8						
V2 onset	162.1	7.8	326.6	29.1	1285.8	21.1	2814.0	14.7
V2 center	157.5	7.2	327.7	31.8	901.2	24.7	2851.1	11.3
UOMINI								
A								
Frame	**F0**	**A0**	**F1**	**A1**	**F2**	**A2**	**F3**	**A3**
V1 center	119.3	10.8	849.3	43.1	1352.3	41.9	2513.1	32.8
V1 offset	117.1	12.1	710.9	36.5	1449.3	37.2	2490.2	29.3
V1 offset 2 C	115.8	12.7	618.4	32.1	1494.2	31.5	2494.0	26.2
C1 onset	109.3	13.1						
C1 center	102.4	11.4						
C2 center	101.2	8.7						
C2 offset	107.4	11.5						
V2 onset	111.1	12.7	534.7	33.3	1525.2	32.3	2433.2	26.7
V2 center	106.6	12.1	672.3	34.6	1447.7	34.3	2419.6	26.8
I								
Frame	**F0**	**A0**	**F1**	**A1**	**F2**	**A2**	**F3**	**A3**
V1 center	134.1	11.1	284.4	25.9	2285.5	28.4	3268.4	35.1
V1 offset	128.3	12.4	286.0	23.3	2282.9	25.5	3239.0	31.9
V1 offset 2 C	124.9	12.9	296.8	19.6	2263.9	21.6	3179.9	26.2
C1 onset	119.8	13.2						
C1 center	105.8	11.3						
C2 center	105.0	10.7						
C2 offset	107.7	12.1						
V2 onset	110.2	13.0	307.1	21.7	2154.0	25.7	3069.3	24.6
V2 center	107.0	13.5	300.6	24.4	2214.7	26.8	2981.4	26.1
U								
Frame	**F0**	**A0**	**F1**	**A1**	**F2**	**A2**	**F3**	**A3**
V1 center	144.2	9.0	304.4	31.4	684.9	27.7	2405.6	14.2
V1 offset	134.8	11.4	309.3	26.8	911.5	22.0	2252.8	13.8
V1 offset 2 C	130.0	12.4	303.3	23.8	1004.8	18.3	2195.1	14.0
C1 onset	124.0	14.1						
C1 center	108.3	12.6						
C2 center	104.7	11.4						
C2 offset	109.6	13.0						
V2 onset	117.3	13.5	323.9	24.2	1213.2	22.8	2233.1	14.2
V2 center	112.9	13.4	314.7	26.5	920.2	20.1	2184.3	18.6

Tabella 4.7 Andamento delle formanti all'interno della pronuncia. Le medie sono eseguite dividendo i dati per sesso e vocale. Tutte le frequenze sono in Hz e le ampiezze in dB.

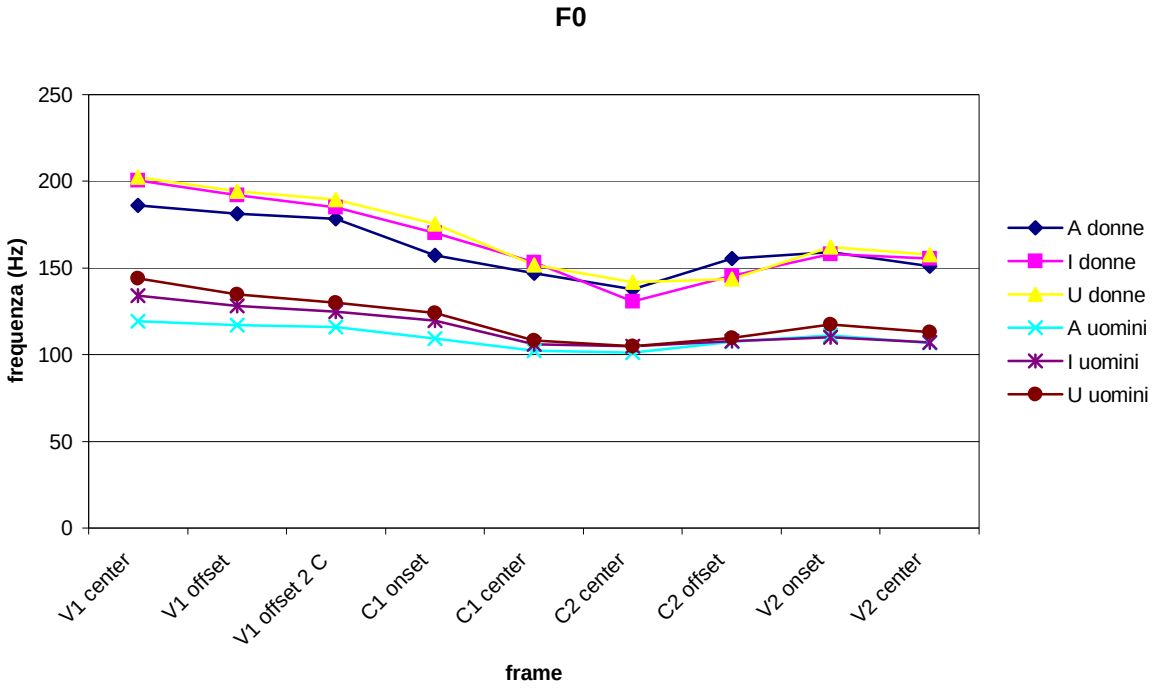


Figura 4.5 Andamento della frequenza fondamentale nei vari frame di analisi.

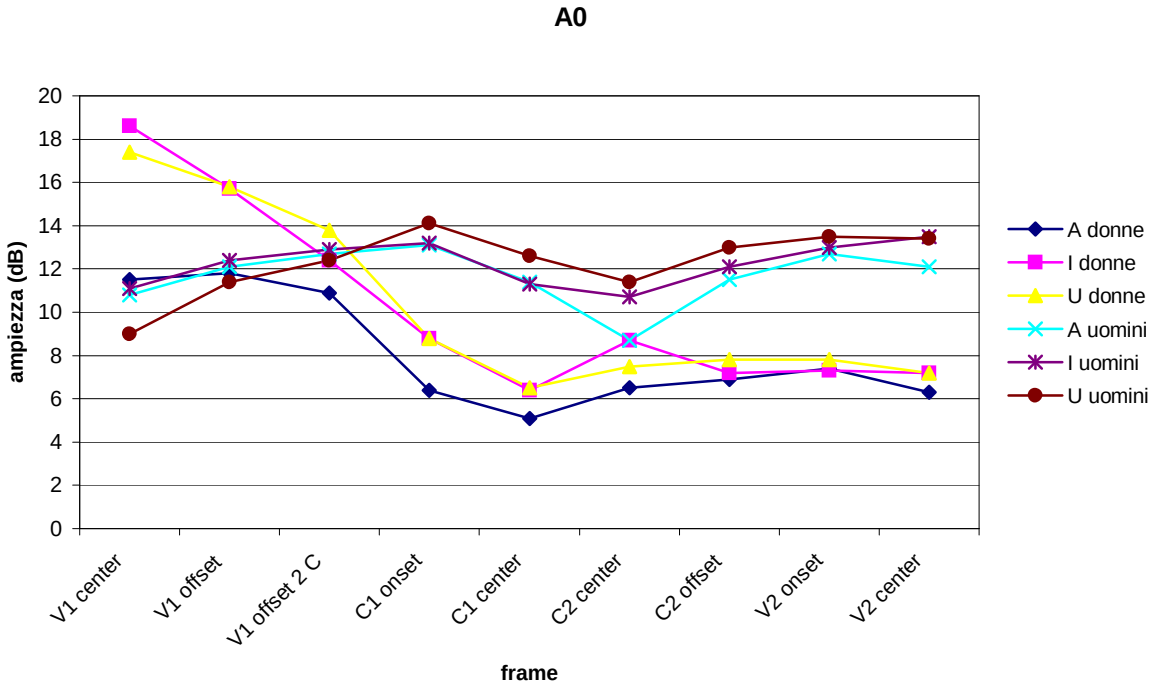


Figura 4.6 Andamento dell'ampiezza della frequenza fondamentale nei vari frame di analisi.

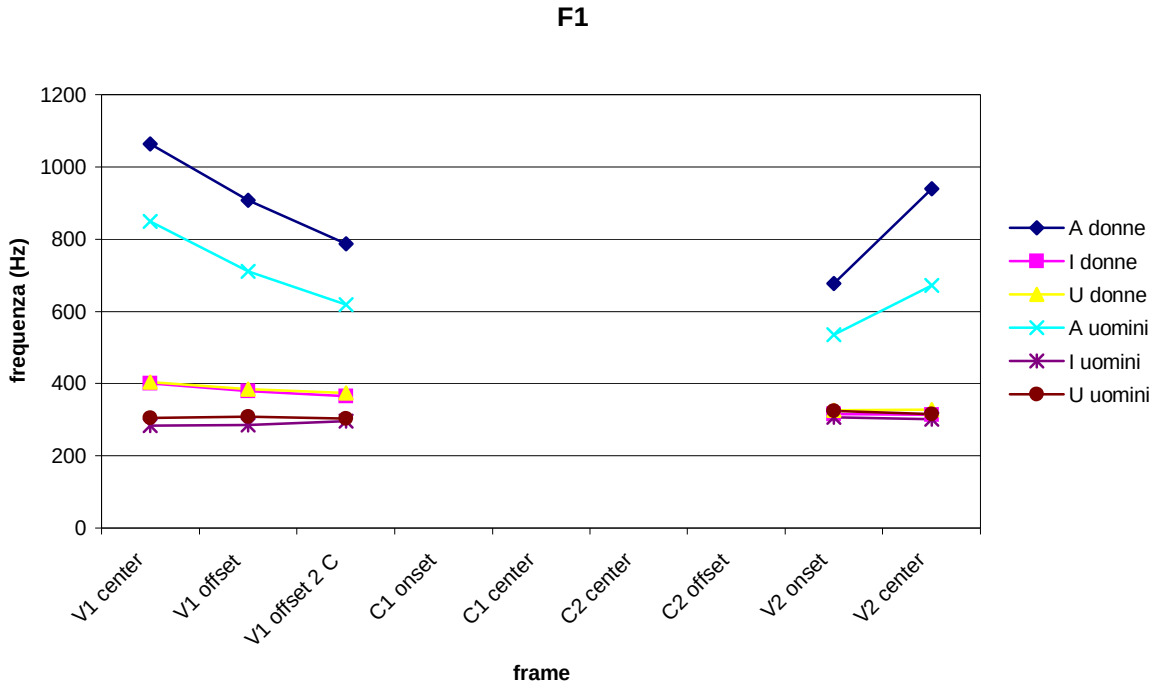


Figura 4.7 Andamento della prima formante nei vari frame di analisi.

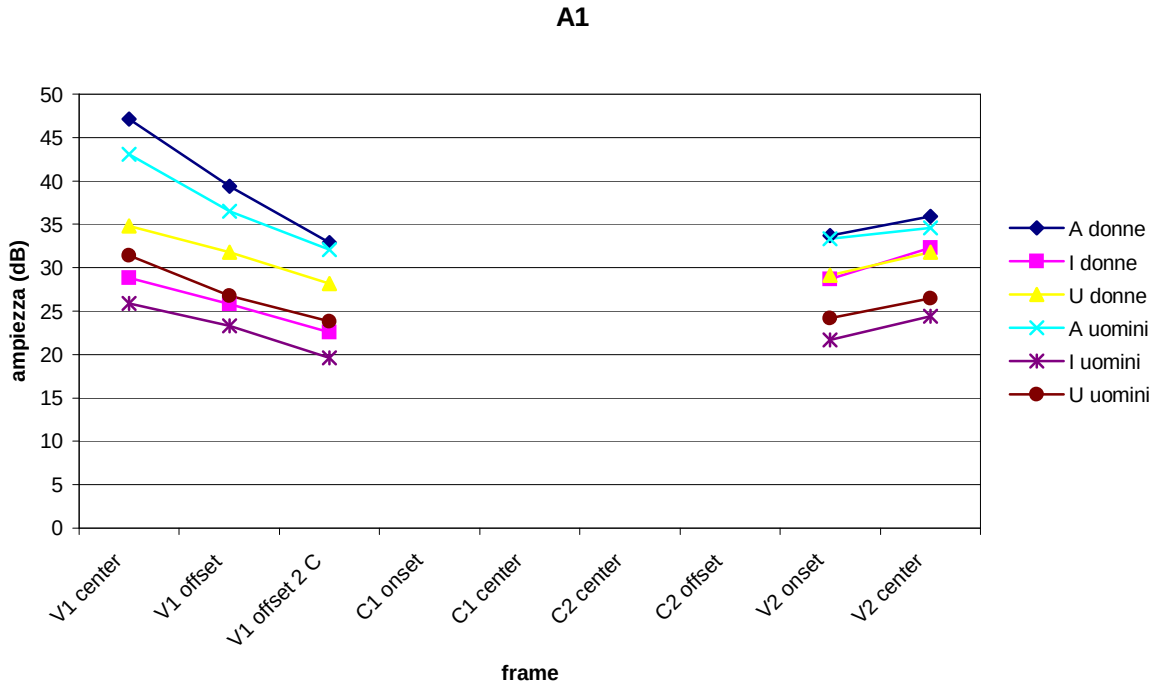


Figura 4.8 andamento dell'ampiezza della prima formante nei vari frame di analisi.

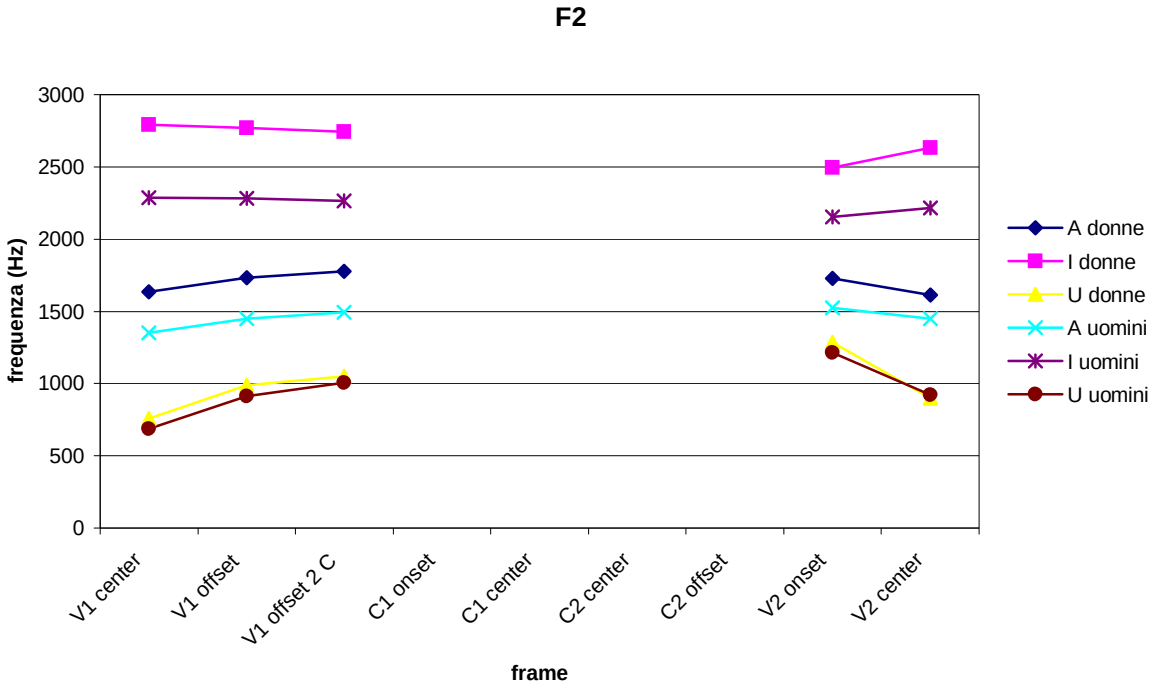


Figura 4.9 Andamento della seconda formante nei vari frame di analisi.

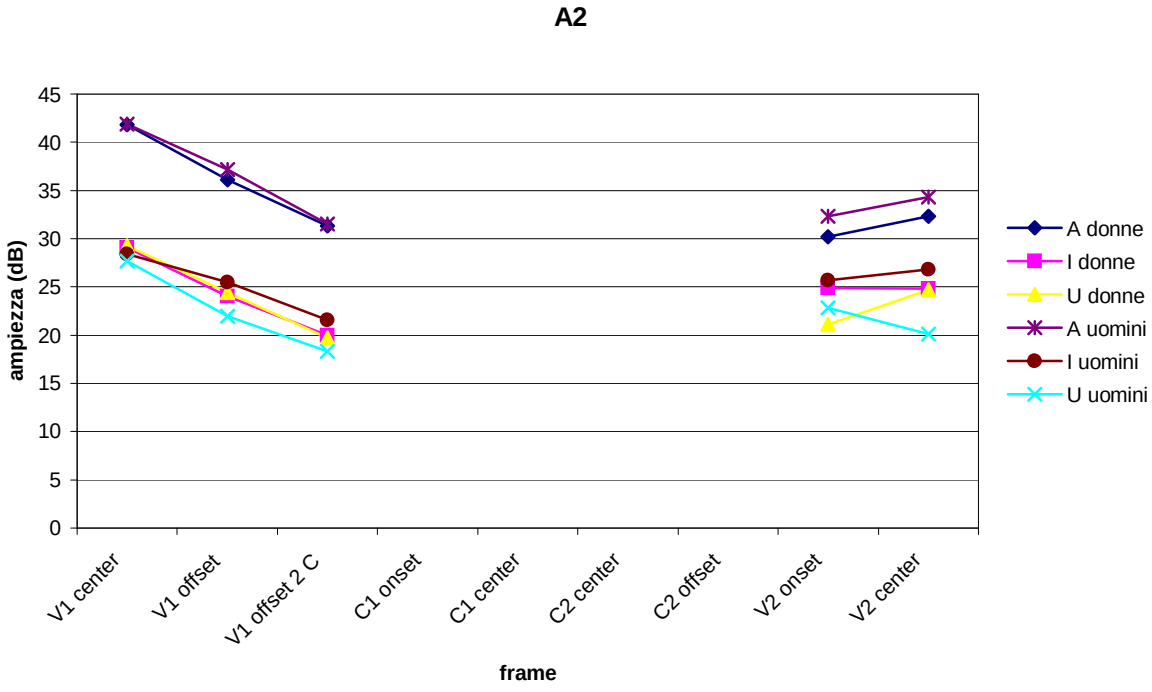


Figura 4.10 Andamento dell'ampiezza della seconda formante nei vari frame di analisi.

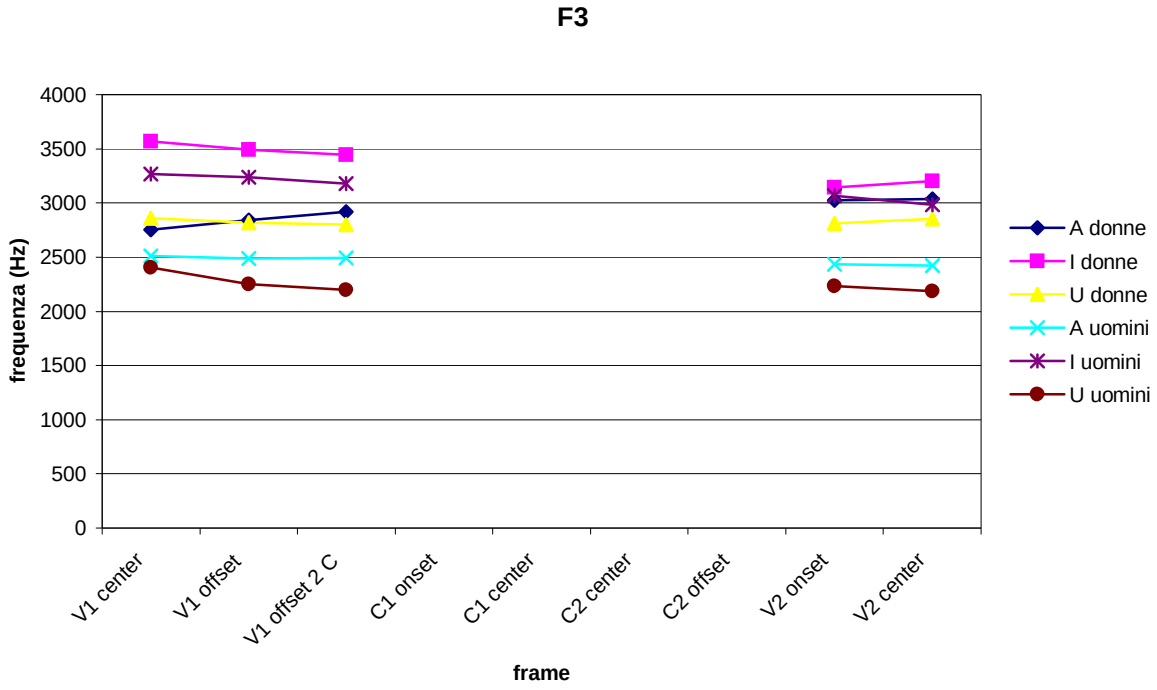


Figura 4.11 Andamento della terza formante nei vari frame di analisi.

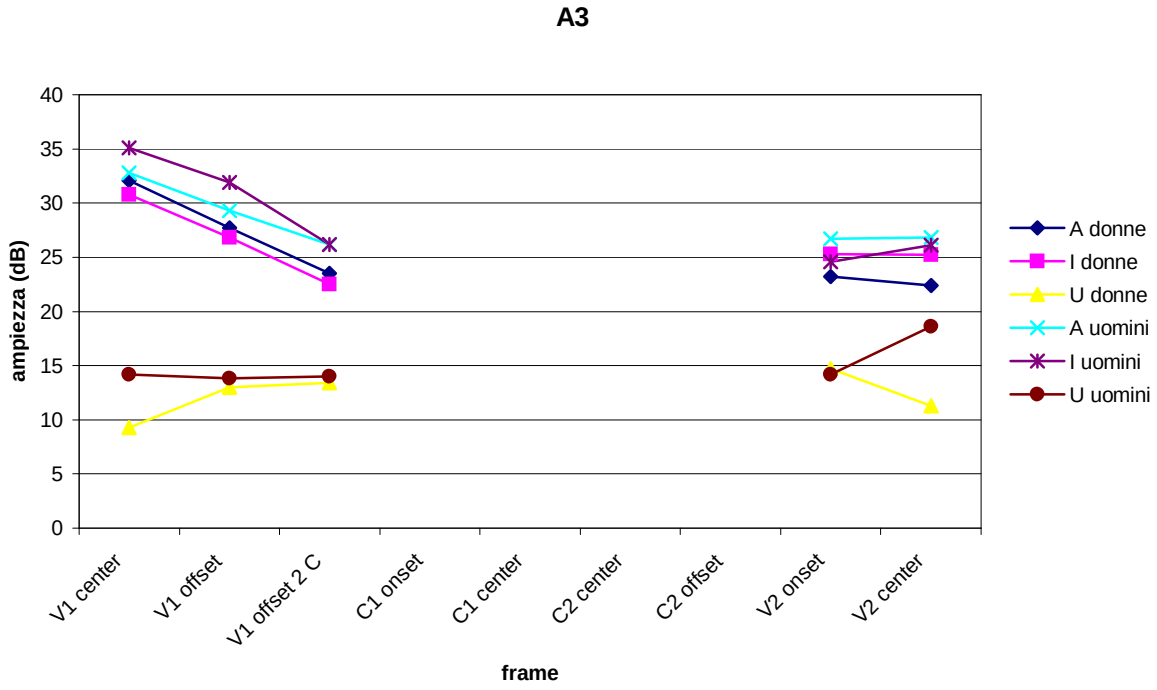


Figura 4.12 Andamento dell'ampiezza della terza formante nei vari frame di analisi.

Classificazione delle pronunce sulla base delle grandezze in frequenza

Tentiamo anche adesso di classificare il tipo delle pronunce sfruttando le grandezze frequenziali che sono risultate statisticamente significative per la geminazione. Abbiamo utilizzato il Maximum Likelihood Criterion (vedi Paragrafo 3.4.4). Tra tutte le grandezze frequenziali significative rispetto al parametro **tipo**, ne sono state selezionate quattro, ossia quelle che, guardando i valori medi e le deviazioni standard delle misure effettuate, si è pensato potessero fornire i risultati migliori. Tali grandezze sono le ampiezze del pitch e delle tre formanti (A0, A1, A2, A3) prese nel frame V1 offset. Proprio in questo frame le suddette grandezze subiscono le variazioni più vistose a causa della geminazione.

La classificazione è stata fatta dividendo le pronunce per sesso e per vocale, in quanto questi due fattori influenzano molto i valori delle grandezze in frequenza. Nella Tabella 4.8 sono riportati i risultati di tale classificazione.

CRITERIO MLC												
	V1 offset A0			V1 offset A1			V1 offset A2			V1 offset A3		
	EPP	Errori	Err. %	EPP	Errori	Err. %	EPP	Errori	Err. %	EPP	Errori	Err. %
Uomini [a]	13.5	25	34.72	33.5	21	29.17	41.5	26	36.11	29.5	29	40.28
Uomini [i]	12.5	25	34.72	25.5	20	27.78	26.5	25	34.72	32.5	22	30.56
Uomini [u]	12.5	29	40.28	31.5	27	37.50	21.5	19	26.39	13.5	22	30.56
Donne [a]	15.5	30	41.67	41.5	25	34.72	37.5	31	43.06	27.5	23	31.94
Donne [i]	14.5	30	41.67	21.5	31	43.06	24.5	25	34.72	26.5	29	40.28
Donne [u]	9.5	24	33.33	26.5	29	40.28	24.5	24	33.33	15.5	26	36.11

Tabella 4.8 Criterio MLC per la classificazione del tipo, condotto sulla base dei parametri frequenziali. E.P.P. rappresenta il punto di equiprobabilità o di separazione delle due gaussiane. E.P.P. è espresso in dB.

Come si può notare i risultati di tale classificazione sono pessimi, non scendendo mai sotto il 26% di errori e avvicinandosi spesso al valore del 50% che si otterrebbe in media con una classificazione completamente casuale.

4.2.3 Elaborazioni statistiche e risultati dell'analisi nel dominio energetico

Scopo dell'analisi energetica è:

1. indagare quali fattori influenzano le grandezze energetiche, con particolare attenzione al fenomeno della geminazione
2. individuare, se vi sono, quali relazioni esistono tra tali parametri energetici
3. fornire, se possibile, un criterio automatico di distinzione tra una pronuncia singola e la sua corrispondente geminata e valutarne il grado di precisione

Medie e deviazioni standard

Anche in questo caso, come per l'analisi nel tempo e in frequenza, le medie sono state calcolate rispetto alle ripetizioni, ai parlatori, al sesso ed infine sulla totalità dei dati. La raccolta di tutti i dati elaborati si trova nelle venticinque tabelle dell'appendice B. Nella Tabella 4.9 sono riportate le medie totali (con le rispettive deviazioni standard) dei quattordici parametri energetici misurati.

	EtotV1	**PmV1**	**EtotC1**	**PmC1**	**EtotC2**	**PmC2**	**EtotC**	**PmC**
Singole	93.7	62.1	75.9	47.1	78.4	49.1	81.4	59.1
(StD)	*6.0*	*5.5*	*5.9*	*6.0*	*4.8*	*5.3*	*4.1*	*4.5*
Geminate	94.1	63.7	76.9	45.8	80.9	50.5	83.4	59.4
(StD)	*5.7*	*5.2*	*6.0*	*5.9*	*5.1*	*5.5*	*4.5*	*4.7*
	EiV1cent	**EiV1-C1**	**EiC1cent**	**EiC1-C2**	**EiC2cent**	**EiC2offs**		
Singole	86.8	79.4	67.3	68.6	73.1	72.5		
(StD)	*5.9*	*5.1*	*9.0*	*6.6*	*6.2*	*5.1*		
Geminate	88.8	80.9	63.2	68.2	74.3	74.5		
(StD)	*5.3*	*4.7*	*9.5*	*6.2*	*6.6*	*5.5*		

Tabella 4.9 Medie e deviazioni standard (StD) rispetto a tutti i parlatori, le ripetizioni, le vocali e le consonanti per il gruppo delle singole (216 pronunce) e per quello delle geminate (216 pronunce). Tutte le misure sono in dB.

Analisi della varianza

Visto che la caratteristica sordità-sonorità delle consonanti prese in esame influenza in maniera molto evidente i risultati dell'analisi energetica, sono state condotte due analisi della varianza multifattoriali (i cui dettagli sono riportati in Appendice E): la prima sulle due consonanti sorde [τΣ,τσ], la seconda sulle due sonore [δZ,δζ]. In entrambi i casi, come in tutte le analisi precedentemente effettuate, sono stati assunti come fattori di variabilità **sesso**, **tipo**, **vocale** e **consonante**.

Esponiamo ora i risultati relativamente alle due classi di consonanti considerate riguardo al fattore **tipo** (ossia alla geminazione) riservandoci di illustrare i risultati per gli altri tre fattori in seguito.

1. **Consonanti sorde.**
    - **V1**, l'energia totale (Etot) non risulta variare tra pronunce singole e geminate mentre la potenza media (Pm) è maggiore nelle geminate. Questo può essere spiegato con la minore durata di V1 nelle pronunce geminate (la stessa energia concentrata in un tempo minore ci dà ovviamente una maggiore potenza)

- **C1**, L'energia totale non cambia tra pronunce singole e geminate, di conseguenza la potenza media nelle singole è maggiore in quanto la durata di C1 è minore.
- **C2**, In questo caso sia l'energia totale che la potenza media sono significativamente maggiori nelle pronunce geminate.
- **C**, si osserva una energia totale maggiore nelle geminate mentre non ci sono variazioni significative per la potenza media. Questo si può spiegare in maniera analoga a quanto fatto per V1, solo che ora è la potenza media ad essere uguale nei due casi. Essendo però la durata della consonante geminata maggiore della singola si ottiene una energia più alta nella pronuncia geminata.
- Per le **energie istantanee** si osservano dei valori maggiori nelle pronunce geminate per EiV1cent, EiV1-C1, EiC2cent e EiC2offs mentre per EiC1cent il valore più alto si ha per le pronunce singole. Il burst di energia tra C1 e C2 (ossia EiC1-C2) non risulta essere influenzato dalla geminazione.

2. **Consonanti sonore.**
   - **V1**, presenta lo stesso comportamento che per le sorde, ossia un valore maggiore per Pm nelle geminate e variazioni statisticamente non significative per Etot.
   - **C1**, si nota una maggiore energia totale nelle geminate mentre la potenza media non cambia. Anche qui la spiegazione risiede nelle diverse durate di C1 tra singole e geminate.
   - **C2**, comportamento uguale a C1.
   - **C**, i risultati per la consonante considerata nella sua interezza sono diretta conseguenza di ciò che accade per C1 e C2. Infatti l'energia totale è maggiore nelle pronunce geminate mentre la potenza media non varia significativamente.
   - Le **energie istantanee** hanno un valore maggiore in EiV1cent e E1C2offs, mentre non ci sono variazioni statisticamente significative tra singole e geminate negli altri quattro parametri.

Per avere una idea d'insieme di ciò che accade si riassumono nella Tabella 4.10 i risultati sopra esposti, indicando con G o S se la grandezza in esame risulta significativamente più grande per le pronunce geminate o singole rispettivamente. Si è svolta anche una analisi complessiva (senza la distinzione tra sorde e sonore) per fornire un quadro generale del comportamento delle consonanti affricate.

Come si può notare da una visione d'insieme dei risultati esposti c'è la tendenza da parte del parlatore ad enfatizzare la pronuncia geminata dando una maggiore potenza soprattutto alla prima vocale (fenomeno osservabile in tutti e tre i tipi di analisi svolte). Nella consonante, considerata nella sua interezza, non si osserva, invece, un aumento di potenza, che si traduce però in un aumento di energia a causa della maggior durata della consonante stessa (vedi i valori EtotC e PmC).

Consideriamo adesso l'analisi statistica svolta rispetto agli altri tre parametri, ossia **sesso**, **vocale** e **consonante**:

- **Sesso.** Consonanti sorde: tutte le grandezze sono influenzate dal sesso ad eccezione di EtotC, PmC, EiV1-C1 ed EiC2cent. In tutte le grandezze si notano dei valori per le donne maggiori che per gli uomini, ad eccezione di EtotC2, PmC2 ed EiC2cent. Le variazioni tra uomini e donne, dove statisticamente significative, vanno da uno a quattro dB. Consonanti sonore: qui è solo l'energia istantanea tra V1 e C1 (EiV1-C1) a non dipendere dal sesso. In questo caso sono solo le grandezze

relative a V1 a presentare valori maggiori per le donne. Per tutte le altre (ossia quelle misurate sulla consonante) si ottengono valori maggiori per gli uomini.

- **Vocale**. Consonanti sorde: le grandezze relative alla vocale risultano chiaramente influenzate dal parametro vocale, risultando nettamente maggiori per la [a] rispetto alla [i] e alla [u], fino ad arrivare ad una differenza di 10 dB (in EtotV1). Anche potenza ed energia della consonante risultano essere influenzate dalla vocale con cui sono coarticolate. In questo caso però i valori maggiori si osservano nelle pronunce con la [u]. Queste variazioni sono relativamente limitate dato che non superano mai i due dB Consonanti sonore: in questo caso **vocale** influenza solo le grandezze legate direttamente alla vocale di coarticolazione, ossia EtotV1, PmV1, EiV1cent, EiV1-C1 ed EiC2offs[6]. Anche in questo caso i valori relativi alla [a] sono i più grandi, quelli della [u] intermedi e quelli della [i] i più piccoli.
- **Consonante**. in questo caso si ritiene opportuno analizzare i dati relativi a tutte le consonanti insieme per poter valutare meglio le differenze energetiche tra consonanti sorde e sonore. L'unica grandezza relativa alla vocale che dipende dalla consonante di coarticolazione è EtotV1, che risulta maggiore nelle due consonanti sonore rispetto a quelle sorde. Ovviamente tutte le grandezze misurate sulla consonante dipendono dalla consonante stessa. In particolare in EtotC1, PmC1, EiV1-C1, EiC1cent, EiC1-C2, EiC2offs la dipendenza è strettamente legata alla caratteristica di sonorità della consonante. Infatti tutti i valori relativi alle consonanti sorde sono minori di quelli relativi alle consonanti sonore. Ciò è facilmente spiegabile se consideriamo che nelle consonanti sonore, durante la produzione del parlato si aggiunge la vibrazione delle corde vocali che fornisce potenza (ed energia) al segnale vocale.

CONSONANTI SORDE [♦♦♦]							
EtotV1	PmV1	EtotC1	PmC1	EtotC2	PmC2	EtotC	PmC
-	G	-	S	G	G	G	-
	EiV1cent	EiV1-C1	EiC1cent	EiC1-C2	EiC2cent	EiC2offs	
	G	G	S	-	G	G	
CONSONANTI SONORE [♦♦♦]							
EtotV1	PmV1	EtotC1	PmC1	EtotC2	PmC2	EtotC	PmC
-	G	G	-	G	-	G	-
	EiV1cent	EiV1-C1	EiC1cent	EiC1-C2	EiC2cent	EiC2offs	
	G	-	-	-	-	G	
TUTTE LE CONSONANTI							
EtotV1	PmV1	EtotC1	PmC1	EtotC2	PmC2	EtotC	PmC
-	G	-	S	G	G	G	-
	EiV1cent	EiV1-C1	EiC1cent	EiC1-C2	EiC2cent	EiC2offs	
	G	G	S	-	G	G	

Tabella 4.10 Risultati sintetici dell'analisi energetica condotta sul parametro tipo. Nelle celle è riportato G o S se la grandezza misurata è significativamente maggiore per le pronunce geminate o singole rispettivamente.

[6] Ricordiamo che dopo la consonante c'è di nuovo una vocale e che quindi è naturale che questa influenzi le grandezze caratteristiche della consonante stessa.

Test di correlazione tra i parametri energetici

Anche per l'analisi nel dominio energetico è stato eseguito un test di Spearman per mettere in evidenza eventuali correlazioni tra le grandezze misurate. Come per le durate dei fonemi in questa sede interessa particolarmente il comportamento rispetto alla geminazione. Sono stati allora calcolati i coefficienti $\mathbf{r_s}$ prima solo per le pronunce singole, poi per quelle geminate e infine per tutte le pronunce insieme, per poter valutare se le correlazioni tra le grandezze energetiche siano o no imputabili alla geminazione. Contrariamente a quanto fatto per l'analisi temporale ci limitiamo a riportare solo i risultati principali e statisticamente significativi, senza mostrare le matrici di correlazione complete. Ovviamente non si discuterà di risultati ovvi, come ad esempio la forte correlazione esistente tra l'energia totale di C1 (EtotC1) e l'energia istantanea al centro della stessa parte di consonante (EiC1cent) oppure tra energia e potenza dello stesso fonema.

Per le pronunce singole si osservano le seguenti correlazioni:

- EtotV1 con PmC1, $\mathbf{r_s} = 0.4553$
- PmV1 con PmC1, $\mathbf{r_s} = 0.4249$
- EiV1cent con PmC1, $\mathbf{r_s} = 0.4301$

Per le pronunce geminate non si osserva nessuna correlazione tra i parametri energetici mentre considerando le pronunce tutte insieme si ottiene la correlazione seguente:

- EtotV1 con PmC1, $\mathbf{r_s} = 0.4062$

Sulla base dei risultati appena esposti possiamo concludere che la correlazione tra EtotV1 e PmC1 non è causata dalla geminazione, in quanto già presente nelle sole pronunce singole. Inoltre non ci sentiamo di trarre delle conclusioni perentorie riguardo ai risultati ottenuti per le pronunce singole in quanto i valori di correlazione non sono così alti da far pensare a dei forti legami tra le suddette grandezze.

Classificazione delle pronunce

Tentiamo anche ora di classificare efficacemente il tipo delle pronunce sulla base dei parametri energetici che sono risultati significativi per la geminazione. E' stato utilizzato come al solito il MLC. La classificazione è stata fatta su tutte le pronunce, poi dividendo uomini e donne, poi dividendo le consonanti e infine dividendo le vocali. Per la classificazione sono stati presi in esame le grandezze che, oltre ad essere significative per la geminazione, presentassero uno scostamento tra i valori medi superiore ad almeno 2 dB per garantire il minor numero di errori possibile. La scelta è allora ricaduta su EtotC2, EtotC, EiV1cent, EiC1cent ed EiC2offs. Nella Tabella 4.11 sono esposti i risultati di tale classificazione. Come si può vedere la classificazione sui parametri energetici porta a risultati decisamente pessimi. A parte una percentuale di errore del 22.22% nelle pronunce con la [$\tau\Sigma$] sull'energia istantanea al centro di C1, le altre sono tutte abbondantemente sopra il 30% (la media degli errori è di circa il 38%), arrivando in vari casi vicino a quel 50% che si otterrebbe con una classificazione del tutto casuale.

CRITERIO MLC															
	EtotC2			EtotC			EiV1cent			EiC1cent			EiC2offs		
	EPP	Errori	Err. %	EPP	Errori	Err. %	EPP	Errori	Err. %	EPP	Errori	Err. %	EPP	Errori	Err. %
Tutte	79.5	170	39.35	84.5	168	38.89	84.5	182	42.13	61.5	174	40.28	75.5	170	39.35
Uomini	79.5	87	40.28	86.5	88	40.74	83.5	85	39.35	61.5	84	38.89	75.5	89	41.20
Donne	80.5	83	38.43	84.5	75	34.72	86.5	91	42.13	57.5	88	40.74	75.5	81	37.50
♦♠	86.5	38	35.19	86.5	38	35.19	82.5	48	44.44	59.5	24	22.22	75.5	35	32.41
♋	79.5	41	37.96	86.5	40	37.04	90.5	48	44.44	55.5	52	48.15	79.5	44	40.74
♦♦	76.5	42	38.89	79.5	40	37.04	84.5	40	37.04	54.5	41	37.96	69.5	41	37.96
♋	78.5	37	34.26	82.5	40	37.04	82.5	37	34.26	68.5	49	45.37	76.5	38	35.19
a	81.5	57	39.58	84.5	53	36.81	89.5	56	38.89	55.5	57	39.58	78.5	59	40.97
i	86.5	63	43.75	86.5	62	43.06	82.5	49	34.03	55.5	60	41.67	75.5	57	39.58
u	79.5	44	30.56	83.5	49	34.03	85.5	54	37.50	61.5	51	35.42	75.5	50	34.72

Tabella 4.11 Criterio MLC per la classificazione del tipo, condotto sulla base dei parametri energetici. E.P.P. rappresenta il punto di equiprobabilità o di separazione delle due gaussiane. E.P.P. è espresso in dB.

Considerazioni complessive sulla classificazione delle pronunce

Analizzando i risultati delle classificazioni nei tre domini in cui sono state studiate le pronunce, si osserva che i migliori risultati sono forniti dalle grandezze misurate nel tempo. Si era inoltre pensato che i risultati in frequenza e nel dominio dell'energia potessero essere utilizzati in qualche maniera per migliorare i già discreti (e in alcuni casi particolari ottimi) risultati di classificazione basati sulle durate dei fonemi. Appare ora ovvio, alla luce dei pessimi risultati raggiunti, che ciò non è possibile e che ci si deve accontentare delle percentuali di errore della classificazione nel dominio del tempo.

CAPITOLO 5

SINTESI DELLE CONSONANTI AFFRICATE

INTRODUZIONE

Nel precedente capitolo è stato illustrato come si è proceduto all'analisi dei dati misurati e a quali risultati ciò ha portato. Si è visto quali parametri influiscono sulle grandezze che caratterizzano i fonemi e si è cercato, ove possibile, di fornire una interpretazione acustico-fisica ai risultati trovati.

Cerchiamo adesso di mettere in pratica ciò che è stato studiato ed analizzato finora realizzando delle pronunce sintetiche di consonanti affricate coarticolate con vocali. Verranno dapprima esposti i fondamenti della sintesi del segnale vocale, verrà poi descritto il funzionamento del sintetizzatore utilizzato, che è l'HLsyn della Sensimetrics, si spiegherà in particolare come le consonanti affricate sono state sintetizzate e verranno infine fatte delle considerazioni sui risultati raggiunti.

5.1 FONDAMENTI DI SINTESI DEL SEGNALE VOCALE

Verranno ora esposte le nozioni base per la comprensione del lavoro svolto, ossia i fondamenti della sintesi, i diversi modi di effettuarla e i diversi tipi di sintetizzatori disponibili. Infine si illustreranno le applicazioni pratiche della sintesi del parlato.

5.1.1 Metodi di sintesi

Volendo fare una classificazione sui sistemi attualmente in uso per la generazione sintetica della voce, si può dire che appartengono a due categorie di base: **sintesi da analisi** e **sintesi da testo**. Nel prosieguo

ci si soffermerà soprattutto su questa seconda metodologia di sintesi dato che questo è l'approccio con cui è stato realizzato il progetto del sintetizzatore HLsyn.

Sintesi da analisi

Nota anche come **sintesi per concatenazione di unità acustiche** effettua la conversione di testo in voce basandosi su unità di base costituite da tratti di voce preregistrati. I parametri di controllo provvedono alla concatenazione di tali tratti.

Con l'uso di questi sistemi si procede alla generazione del parlato solo dopo aver eseguito una analisi accurata di una voce già registrata. Si tratta in sostanza di una riproduzione di un segnale vocale già esistente, una sorta di "copia e incolla" realizzata principalmente sfruttando le capacità di memorizzazione degli attuali apparati di elaborazione. Le parole, ove possibile, e i segmenti più piccoli vengono fusi tra loro considerando anche i caratteri prosodici della frase, con particolare riferimento all'intonazione, per cercare di ottemperare al requisito qualitativo della naturalezza.

Un problema che sorge con l'uso di questo sistema di sintesi è dato dalla difficoltà di desumere regole semplici dove invece sono molto articolate. In base a ciò si è arrivati alla definizione di una unità detta *difono*, che include la porzione temporale che intercorre tra le parti stabili di un fonema e del successivo. Questa unità contiene parte delle caratteristiche di transizione tra un fonema e l'altro. In questo modo si tiene conto del fatto che, nella maggioranza dei casi, l'assetto articolatorio del tratto vocale caratteristico di un fonema influenza la realizzazione acustica del fonema seguente. A tal proposito, per coprire un maggior numero di combinazioni, si stanno sviluppando algoritmi di regole più complesse che definiscono con maggior precisione la concatenazione tra fonemi adiacenti o che coinvolgono unità acustiche più estese, come *trifoni*, *quadrifoni* o addirittura parti intere di frasi.

Un altro problema è dato dalla dimensione del **corpus** delle unità. Con questa dizione si intende la collezione delle parole o delle unità multifonetiche immagazzinate nella memoria dell'elaboratore e che risultano così disponibili all'impiego. La dimensione di questo vocabolario è finita, così rimangono esclusi neologismi, nomi propri, citazioni in altre lingue che costringono ad un continuo aggiornamento. Nella lingua italiana è stato individuato un corpus di difoni che può arrivare anche alle migliaia di unità, a seconda delle specifiche richieste dall'utente. Gli sviluppi futuri sono rivolti al superamento di queste limitazioni sia per motivi di praticità che per risparmio di risorse. Ci si rivolge così ai sistemi di sintesi da testo (il vero e proprio *text-to-speech*) che verranno ora descritti.

Sintesi da testo

Nota anche come **sintesi per regole** effettua una conversione da testo in voce tramite l'implementazione di un sistema *text-to-speech*, realizzato spesso via software. Gli algoritmi di tale software devono poter riprodurre i meccanismi e le regole di generazione del parlato umano per ottenere un risultato corretto sia dal punto di vista acustico-articolatorio che logico-grammaticale. La natura del parlato che si ottiene con queste tecniche è completamente sintetica. In questo caso le unità fonetiche sono i *fonemi* mentre il compito di pilotare il sintetizzatore è delegato a quei parametri caratteristici, come le frequenze formanti o la posizione degli organi articolatori, che individuano l'applicazione delle regole ed i tanto importanti meccanismi di transizione fra coppie di fonemi adiacenti.

Questo tipo di sintesi viene definito anche sintesi per regole in quanto un gran lavoro di ricerca e di studio è stato, ed è tuttora, dedicato alla conversione in algoritmi delle complesse regole che gestiscono

una corretta trasformazione dell'informazione da un ambito astratto ad un ambito reale-vocale. A tal riguardo sono importanti gli studi compiuti sui *compilatori di regole*, che sembrano essere in grado di gestire via software i fonemi di un linguaggio a prescindere dalla lingua a cui appartengono. I vantaggi sono ovviamente legati alla versatilità di impiego. I primi a dedicarsi a questo tipo di implementazione, con l'uso di un apposito linguaggio e ottenendo buoni risultati, furono Carlson e Granström negli anni 1975-1976. Purtroppo sorgono dei problemi quando alcune regole modificano dei segmenti fonetici senza che siano intervenuti dei simboli sintattici o fonologici. In proposito studi più recenti sono rivolti alla modellizzazione delle strutture linguistiche con una fonologia che tenga conto anche di queste regole. Per ovviare a questi problemi nasce così la *fonologia tridimensionale* (Clements e Halle).

Ricerche per l'adattamento di queste tecniche di compilazione di regole a macchine per la sintesi da testo in tempi relativamente recenti sono state eseguite nel 1984 da Klatt ed Aoki. La Figura 5.1 mostra il passaggio fra la rappresentazione linguistica astratta di una frase e la sua rappresentazione in termini di forma d'onda. Questo processo di generazione di una frase si sviluppa con la successione di passaggi intermedi tendenti a soddisfare tutte le regole semantiche, sintattiche e lessicali nonché le regole per il corretto funzionamento acustico-articolatorio.

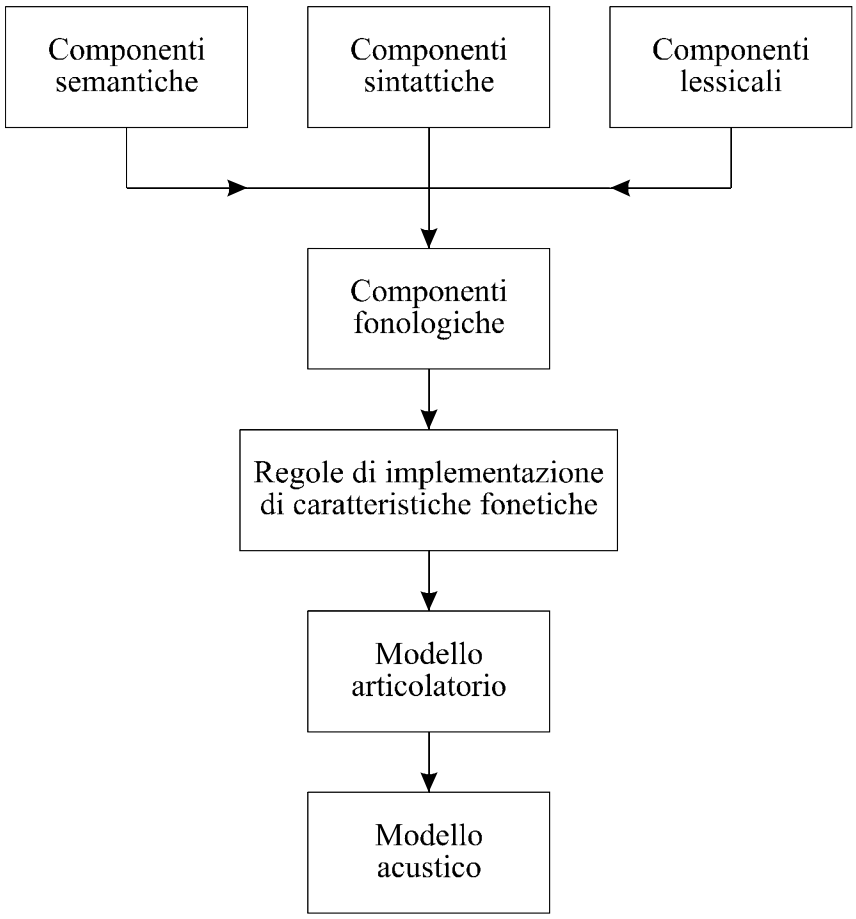


Figura 5.1 Passaggio fra la rappresentazione linguistica astratta di una frase e la sua rappresentazione in termini di forma d'onda.

5.1.2 Modelli per la generazione di voce sintetica

Tra i metodi per la generazione della voce sintetica, opportuni per effettuare la sintesi del segnale vocale, illustriamo il **modello a formanti** e il **modello articolatorio** (di cui fa parte il sintetizzatore HLsyn, oggetto del presente capitolo).

I sintetizzatori che utilizzano un approccio che privilegia la modellizzazione fisica del condotto vocale, rappresentandone la geometria e le proprietà, sono detti **sintetizzatori articolatori**. Il progetto di questi sistemi prevede l'implementazione software o hardware (preferibile la prima per motivi di flessibilità) di una rappresentazione delle sezioni del tratto vocale, con parametri variabili nel tempo in grado di modellare opportunamente le sorgenti e le interazioni fra le varie parti. Alla base di questi sistemi c'è una descrizione del processo di produzione della voce che tiene conto, con grande precisione, del funzionamento dell'apparato fonatorio umano, descritto nel Paragrafo 1.1 del presente lavoro. La modellizzazione che ne segue rispecchia quindi la descrizione data.

I sintetizzatori che privilegiano alle caratteristiche fisiche del tratto vocale quelle di trasmissione del condotto vocale solo dal punto di vista ingresso-uscita del sistema "condotto vocale" sono classificati come **sintetizzatori per formanti**. Una base comune per questi modelli è considerare il segnale vocale come un segnale di uscita di un sistema segmentato a filtri e sorgenti di suoni e rumori, tutti variabili nel tempo. Per poter modellare i parametri (come le frequenze formanti e le relative ampiezze e larghezze di banda) responsabili della generazione della voce, si utilizzano diversi circuiti risonatori in serie e/o parallelo. La rappresentazione usata per la sorgente sonora prevede la generazione di una forma d'onda tale da essere in grado di seguire, con una modellizzazione a tubi stazionari, le variazioni della velocità del flusso d'aria attraverso gli organi fonatori. Per maggiori dettagli sui sintetizzatori per formanti si veda (Scarlino, 1993).

5.1.3 Prospettive ed applicazioni future

I possibili campi di applicazione di un sintetizzatore vocale sono molteplici. Possono far parte di dispositivi atti a sostituire l'operatore umano lì dove prima era indispensabile la voce e quindi nei servizi di telecomunicazione in senso lato. Si può generare la voce dalla lettura di dati, o in generale di altre informazioni immagazzinate in file testo o in altra modalità sempre di tipo numerico, come ad esempio la trasduzione diretta della lettura di bollettini di informazione (meteo, borsa valori, viabilità ecc.) su richiesta dell'utente. Molteplici possono essere le applicazioni per agevolare la vita ai portatori di gravi handicap, come ad esempio la cecità. Sono inoltre possibili sistemi di traduzione simultanea in lingue diverse, trasformando il segnale vocale di un interlocutore in formato numerico tramite un riconoscitore del parlato, lasciando poi il compito ad un sistema automatico di tradurre le informazioni in un'altra lingua e infine riconvertendo i dati in formato vocale tramite un sintetizzatore del parlato, come illustrato dalla Figura 5.2.

Figura 5.2 Schema di un possibile traduttore simultaneo che non necessita dell'intervento diretto dell'uomo.

5.2 IL SINTETIZZATORE HLSYN

In questo paragrafo verranno descritte le principali caratteristiche e funzionalità del sintetizzatore articolatorio HLsyn. Si daranno soltanto le informazioni necessarie a comprendere il lavoro svolto per ovvi motivi di spazio, rimandando al manuale per una descrizione più approfondita e completa del del sintetizzatore.

5.2.1 Caratteristiche generali e parametri di controllo

Il sintetizzatore articolatorio HLsyn si basa sul precedente sintetizzatore per formanti KLsyn (Scarlino, 1993). In pratica si può dire che il sintetizzatore HL utilizza il precedente KL tramite delle relazioni matematiche che convertono i valori dei parametri impostati nell'HL nei valori del KL. Tale approccio è basato sull'osservazione che esistono dei legami e dei vincoli tra gli oltre quaranta parametri di controllo (formanti, loro ampiezze e larghezze di banda, ampiezze delle eccitazioni fricative e sonore ecc.), del sintetizzatore KLsyn. Questi vincoli esistono perché il processo fisico della produzione del parlato impone dei limiti sulle combinazioni dei parametri di sintesi che ci possono essere in ogni particolare istante della fonazione e in come questi parametri possono variare nel tempo. In accordo a questi limiti, è stato proposto un insieme di 10 (poi ampliato a 13) parametri ad un più alto livello (HL, higer level) di quelli del sintetizzatore per formanti KL. Questi parametri HL sono legati più direttamente allo stato e ai movimenti del tratto vocale di quanto non lo fossero i parametri del KLsyn. Un insieme di relazioni, implementate nell'HLsyn, trasforma i parametri HL in parametri KL che si occupano di controllare il sintetizzatore KLsyn88. Oltre a questi 13 parametri che possono essere variati a proprio piacimento (sempre entro i limiti previsti) durante la pronuncia, ce ne sono altri 24 che possono essere impostati dall'utente ma che restano costanti per tutta la durata della pronuncia sintetizzata e alcune altre decine invisibili all'utente e che non possono essere modificate.

Analizziamo ora quali sono i parametri di controllo e come essi sono legati alle caratteristiche che l'apparato vocale assume durante la fonazione. In Tabella 5.1 sono illustrati i parametri di controllo con una loro breve descrizione mentre in Figura 5.3 si può vedere come essi agiscono sulle caratteristiche dell'apparato fonatorio umano.

I primi cinque parametri del sintetizzatore HLsyn sono molto simili (e in alcuni casi uguali) ai parametri del KLsyn. Questi sono la frequenza fondamentale **f0** e le quattro frequenze formanti **f1**, **f2**, **f3** e **f4** che specificano le frequenze naturali del tratto vocale assumendo che non ci siano accoppiamenti acustici con la trachea o con la cavità nasale e che non ci siano costrizioni localizzate causate dalla punta della lingua e dalle labbra. Le frequenze formanti specificano come la forma del tratto vocale cambia durante la produzione del parlato (si pensi, ad esempio, alle differenti forme che assume la bocca pronunciando una [a] o una [u] e a come si ripercuotono sulla posizione ed ampiezza delle formanti). Se ci sono accoppiamenti con la trachea o con il naso o se c'è una costrizione localizzata (come specificato dai parametri **an**, **ag**, **al** e **ab**) le relazioni di mappatura modificano i parametri del sintetizzatore KLsyn. I parametri **f1**, **f2**, **f3** e **f4** descrivono gli aspetti del tratto vocale che sono determinati dalla posizione del corpo della lingua, dalla posizione della mascella, dalla forma della faringe e dall'eventuale arrotondamento delle labbra.

Parametro	Descrizione
f1, f2, f3, f4	Prime quattro frequenze naturali del tratto vocale. Queste sono le frequenze naturali quando la faringe è chiusa, non c'è accoppiamento acustico con la trachea e non ci sono occlusioni, anche parziali, davanti al tratto vocale formate dalla lingua o dalle labbra..
f0	Frequenza fondamentale di vibrazione delle corde vocali. E' data un decimi di Hz.
ag	Area dell'apertura della glottide. Il range di variazione normale è tra 0 e 40 mm^2. Il valore medio per suoni sonori è di circa 3 - 5 mm^2.
al	Area trasversale della costrizione formata dalle labbra durante la produzione delle consonanti. Il range di variazione è tra 0 e 100 mm^2. Il valore 100 mm^2 corrisponde alla configurazione senza costrizione.
ab	Area trasversale della costrizione formata dalla lingua durante la produzione delle consonanti. Il range di variazione è tra 0 e 100 mm^2. Il valore 100 mm^2 corrisponde alla configurazione senza costrizione
an	Area trasversale della costrizione del velo faringeo. Il range di variazione è tra 0 e 100 mm^2.
ue	Rapidità di aumento del volume del tratto vocale durante l'intervallo di occlusione di una consonante occlusiva sonora. Un valore positivo di ue corrisponde ad una espansione della cavità dietro al punto di occlusione, un valore negativo ad una contrazione. L'integrale di ue calcolato sull'intervallo di costrizione è l'aumento o la diminuzione totale del volume.
ps	Pressione subglottale. Permette di aumentare o diminuire l'intensità del segnale prodotto. L'unità di misura è in cm di H_2O.
dc	Variazione percentuale dell'elasticità delle pareti dell'apparato fonatorio durante la pronuncia.
ap	Area dell'interstizio glottale posteriore che persiste attraverso un ciclo glottale. L'unità di misura è mm^2.

Tabella 5.1 Elenco completo dei parametri di controllo del sintetizzatore HLsyn. Gli ultimi 3 (ps, dc e ap) sono stati introdotti sulla attuale versione del sintetizzatore (Versione 2.2).

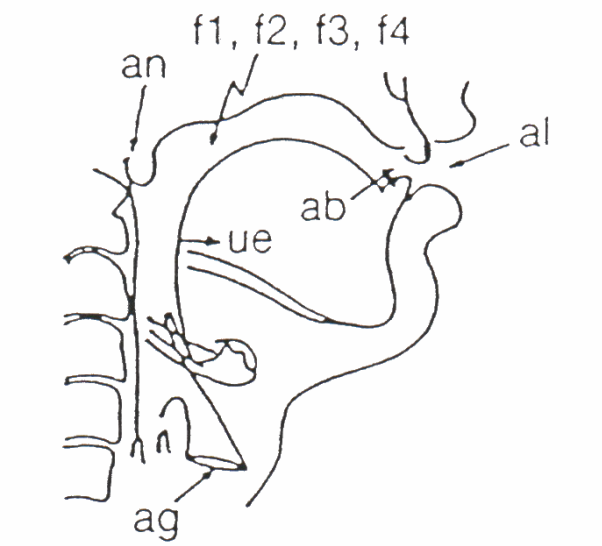


Figura 5.3 Legame tra i parametri del sintetizzatore e le caratteristiche del tratto vocale.

I parametri HL includono le aree di quattro costrizioni che si possono avere nella bocca e che sono:

- **an**, sezione di apertura della cavità nasale, data dal maggiore o minore abbassamento del velo palatino
- **ag**, area media dell'apertura della glottide
- **al**, area della costrizione formata dalle labbra
- **ab**, area della costrizione formata dalla punta della lingua

C'è da dire che **an** interviene solo per le consonanti nasali o, più in generale, quando c'è una nasalizzazione di qualche fonema, mentre **al** e **ab** intervengono solo durante la produzione delle consonanti.

Nella produzione di consonanti occlusive sonore si ha il passaggio di aria attraverso le corde vocali (per la produzione della sonorità) che però non può fuoriuscire all'esterno fino al momento del rilascio a causa dell'occlusione formata per produrre la consonante stessa. Si ha allora all'interno della bocca un aumento del volume compreso tra le corde vocali e il punto di occlusione. Di ciò tiene conto il parametro **ue**, che rappresenta la rapidità con cui questo volume varia e può essere sia positivo (per permettere la vibrazione delle corde vocali durante le consonanti occlusive) che negativo. Il suo integrale rappresenta ovviamente l'aumento o la diminuzione totale del volume all'interno della bocca.

Gli ultimi tre parametri, introdotti su questa ultima versione del sintetizzatore HL, sono **ps**, **dc** e **ap**. Il primo, **ps**, rappresenta la pressione sub-glottale e permette di variare l'intensità della sorgente sonora. Si può utilizzare, per esempio, per aumentare l'ampiezza di una vocale per le sillabe accentate. Per quanto riguarda **dc** c'è da fare una premessa: è stato dimostrato che la tensione delle pareti del tratto vocale, quando sottoposto ad una forza periodica, come ad esempio l'eccitazione dovuta alla vibrazione delle corde vocali, può variare significativamente durante una pronuncia (Svirsky et al., 1997). Il parametro **dc** (delta compliance) tiene conto di ciò, rappresentando la variazione percentuale che l'elasticità delle pareti dell'apparato fonatorio subisce durante la pronuncia. Infine **ap** rappresenta l'area dell'interstizio glottale posteriore che persiste durante un ciclo glottale. Grazie ad esso ora si può, per esempio, avere un miglior controllo del flusso d'aria per sintetizzare fricative sonore e si possono sintetizzare occlusive sonore aspirate.

Parametro	Descrizione	Val. Default	Parametro	Descrizione	Val. Default
TLm	tilt	5 dB	**Cwm**	elasticità pareti tratto vocale	0.001 cm^5/dina
OQm	quoziente di apertura	50%	**Rw**	resistenza pareti tratto vocale	10 dina*s*cm^-5
B1m	largh. di banda 1° formante	80 Hz	**Cgm**	elasticità corde vocali	8E-6 cm^5/dina
B2m	largh. di banda 2° formante	90 Hz	**Lg**	lungh. orizzontale glottide	1 cm
B3m	largh. di banda 3° formante	150 Hz	**LabialAB**	guadagno per il filtro parallelo	55 dB
B4m	largh. di banda 4° formante	350 Hz	**PalVelarA2f**	A2F per fricaz. palatovelare	55 dB
B5m	largh. di banda 5° formante	500 Hz	**PalVelarA3f**	A3F per fricaz. palatovelare	60 dB
B2f	largh. di banda per F2 in parall.	250 Hz	**PalVelarA5f**	A5F per fricaz. palatovelare	50 dB
B3f	largh. di banda per F3 in parall.	320 Hz	**RetroflexA3f**	A3F per fricaz. retroflessa	50 dB
B4f	largh. di banda per F4 in parall.	350 Hz	**LateralA3f**	A3F per fricaz. laterale	40 dB
B5f	largh. di banda per F5 in parall.	500 Hz	**F5**	quinta formante	4500 Hz
Psm	pressione subglottale	8 cm H_2O	**F6**	sesta formante	4990 Hz

Tabella 5.2 Elenco dei parametri caratteristici del singolo parlatore

Come già detto oltre a questi 13 parametri che variano durante la pronuncia ce ne sono altri 24 che possono essere impostati dall'utente ma che si mantengono costanti per tutta la durata della parola

sintetizzata (si può pensare ad esse come delle grandezze caratteristiche di ciascun parlatore). Questi sono elencati in Tabella 5.2 con una brevissima descrizione. Si fa presente che per la sintesi delle pronunce di questa tesi sono stati utilizzati i valori di default, validi per un generico parlatore maschile (per maggiori dettagli si veda la documentazione del sintetizzatore).

Analizziamo ora un po' più in dettaglio come i parametri di controllo appena descritti possono essere utilizzati nella sintesi di una pronuncia. Verranno descritti solo gli aspetti principali per ovvi motivi di spazio, lasciando al lettore interessato lo studio del manuale del sintetizzatore.

Costrizioni del tratto vocale e ampiezza delle sorgenti

Le proprietà della sorgente sono determinate dai flussi e dalle cadute di pressione attraverso le costrizioni nella glottide e nelle vie superiori. Ci sono tre tipi di orifizi che possono influenzare i flussi e le pressioni:

1. l'area trasversale dell'apertura del velo faringeo
2. l'area trasversale dell'apertura della glottide
3. la minima area trasversale presente nel tratto vocale sopra la laringe

Il primo di questi è dato semplicemente dal parametro **an** ed è diverso da zero solo per le pronunce nasalizzate (limitate a [m, n, ŋ] nell'italiano ma molto frequenti nella lingua inglese). Il secondo è dato dal parametro **ag**, escluso il caso in cui la pressione aumenti nel tratto sopra la glottide. In questo caso viene imposta sulla superficie delle corde vocali un aumento di pressione che può portare ad un aumento dell'area di apertura della glottide. In questo caso il sintetizzatore utilizza, per calcolare i flussi e le pressioni, un parametro modificato chiamato **agx**, che ottiene in base a calcoli ed algoritmi implementati sul software stesso. Il terzo tipo di strettoia che si può avere nel tratto vocale può essere formata con le labbra, con la punta della lingua o con il corpo della lingua. Se la costrizione è formata dalle labbra o dalla punta della lingua, l'area della sezione così formata è data rispettivamente da **al** o **ab**. Quando invece è l'intera lingua a formare il restringimento alzandosi verso il palato, la lunghezza della costrizione è maggiore rispetto alle due precedenti. Ciò provoca un effetto globale sulla forma del tratto vocale. In questo caso la sezione del restringimento non è data da un semplice parametro del sintetizzatore ma viene calcolata in base ad altre grandezze, soprattutto la prima formante. L'innalzamento della lingua provoca infatti un abbassamento della frequenza di f1. Quando allora si è di fronte a una occlusione formata da tutto il corpo della lingua (come avviene ad esempio nella pronuncia della [τΣ]) si deve modificare la grandezza **f1** per sintetizzare correttamente tale fenomeno.

Filtraggio delle sorgenti per la produzione di consonanti sonore e vocali

Per le vocali non nasalizzate (**an**=0) la funzione di trasferimento tra velocità del flusso d'aria nella glottide e velocità sulle labbra è una funzione a tutti poli. Assumendo che, durante un ciclo di vibrazione delle corde vocali, non ci siano cambiamenti significativi nella frequenza o nella larghezza di banda delle formanti, la sintesi di una vocale si può ottenere con la sorgente glottale standard (controllata, lo ricordiamo, dal parametro **ag** compreso tra 3 e 5 mm^2) filtrata da una cascata di cinque frequenze formanti. Le quattro frequenze formanti **f1**, **f2**, **f3** e **f4** possono essere variate a piacere durante la pronuncia mentre la quinta va impostata come costante per ogni parlatore. Tali formanti dovranno essere quelle caratteristiche della vocale che si sta sintetizzando, potendo subire delle variazioni in base alle caratteristiche del singolo parlatore (ad esempio se si sta sintetizzando una voce maschile o femminile). In

questa versione del sintetizzatore le larghezze di banda nominali delle diverse formanti sono fissate per tutta la pronuncia e i valori di default sono quelli in Tabella 5.2. Queste sono le larghezze di banda utilizzate quando la sorgente glottale è impostata per la produzione di suoni sonori (tipicamente ag=4 mm^2). Le effettive larghezze di banda dipendono dalla vocale (ossia dalla frequenza delle formanti e da quanto esse sono vicine l'una all'altra) e dalla lunghezza del tratto vocale del parlatore. Attualmente tali variazioni non sono incluse nelle relazioni di mappatura del software e la larghezza di banda delle formanti è un parametro fisso.

Filtraggio delle sorgenti di rumore (sorgenti fricative)

Dalle costrizioni che si possono avere nell'apparato fonatorio (labbra, punta o corpo della lingua) si può ottenere, per ogni istante, quella che ci dà la più piccola sezione di passaggio dell'aria. Si può pensare che il flusso d'aria che attraversa l'apparato boccale sia controllato da tale sezione minima e che la turbolenza dell'aria sia generata nelle vicinanze di tale costrizione. Il rumore così prodotto attraversa un insieme di filtri in parallelo che hanno il compito di modellizzare il comportamento dell'apparato fonatorio umano. Dato che le quattro frequenze formanti sono conseguenza della forma del tratto vocale, può essere possibile dedurre la posizione della costrizione da queste frequenze. La posizione e forma della costrizione determina quali formanti sono eccitate dal rumore di fricazione.

5.2.2 Il software del sintetizzatore

L'unità completa che contiene tutte le informazioni di un file sintetizzato è l'HL Document (file con estensione .hld). E' un file binario composto da sette gruppi di dati. Ogni gruppo può anche essere esportato separatamente in un file a sé stante con le seguenti estensioni:

1. file di descrizione del documento (.hli)
2. file di descrizione HL Speaker (.hls)
3. file di descrizione KL Speaker (.kls)
4. file con i parametri HL (.hl)
5. file con i valori di pressione dei flussi (.pf)
6. file con i parametri KL (.kl)
7. file in formato wave (.wav)

Anche un file nel formato del sintetizzatore KLsyn (.kld) può essere aperto e modificato con il programma HLsyn. La sintesi effettuata in questo modo corrisponde ad usare un sintetizzatore a formanti cascata-parallelo (Scarlino, 1993). Si può anche salvare un file di sintesi nel formato KL. In questo caso il file salvato (.kld) contiene quattro gruppi di dati, analogamente al formato .hld, e che contengono le seguenti informazioni:

1. file di descrizione del documento (.hli)
2. file di descrizione KL Speaker (.kls)
3. file con i parametri KL (.kl)
4. file in formato wave (.wav)

Tutte le operazioni sui file appena descritte si possono eseguire dal menù 'file' dell'interfaccia grafica del sintetizzatore. E' anche possibile importare file in formato wave per visualizzare forma d'onda, spettrogramma ecc. per poter fare dei confronti con le pronunce sintetizzate.

Il programma è in grado di visualizzare due tipi di finestre: finestre di testo e finestre grafiche. Le tre finestre di testo disponibili permettono di visualizzare, modificare e salvare i parametri HL e KL e di vedere i valori delle pressioni dei flussi (PF Values). Le quattro finestre grafiche permettono di visualizzare l'andamento dei parametri HL, KL, dei flussi PF e dello spettrogramma della pronuncia.

Il programma HLsyn implementa il metodo dei *punti di controllo* (control points) per l'inserimento dei valori dei parametri. Grazie a questo metodo si devono inserire i valori solo in corrispondenza di istanti di tempo scelti dall'utente. Il programma provvederà poi automaticamente a ricostruire con una interpolazione lineare i valori dei parametri tra due istanti precedentemente fissati. I punti di controllo possono essere fissati nelle finestre dei parametri HL e KL. La Figura 5.4 mostra appunto la finestra dei parametri HL. La prima colonna a sinistra contiene gli istanti temporali in msec, anche essi inseriti dall'utente secondo necessità. I caratteri più scuri indicano i valori fissati dall'utente mentre quelli più chiari sono i valori ricavati per interpolazione lineare dal programma stesso.

acia19 - HL Parameters

	ag	al	ab	an	ue	f0	f1	f2	f3	f4	ps	dc	ap
0.0	0.0	100.0	100.0	0.0	0.0	1050	740.0	1300	2500	3500	8.000	0.0	0.0
30.00	0.0	100.0	100.0	0.0	0.0	1050	740.0	1300	2500	3500	8.000	0.0	0.0
35.00	4.000	100.0	100.0	0.0	0.0	1051	740.0	1300	2500	3500	8.000	0.0	0.0
100.0	4.000	100.0	100.0	0.0	0.0	1070	740.0	1300	2500	3500	8.000	0.0	0.0
185.0	4.425	100.0	100.0	0.0	0.0	1035	740.0	1300	2500	3500	8.000	0.0	0.0
200.0	4.500	100.0	65.38	0.0	0.0	1028	687.5	1300	2500	3500	8.000	0.0	0.0
220.0	21.58	100.0	19.23	0.0	0.0	1020	617.5	1300	2500	3500	8.000	0.0	0.0
224.0	25.00	100.0	10.00	0.0	0.0	1020	603.5	1300	2500	3500	8.000	0.0	0.0
225.0	0.0	100.0	10.00	0.0	0.0	1019	600.0	1300	2500	3500	8.000	0.0	0.0
275.0	0.0	100.0	10.00	0.0	0.0	1014	600.0	1845	2318	3318	8.000	0.0	0.0
280.0	20.00	100.0	70.00	0.0	0.0	1013	600.0	1900	2300	3300	11.00	0.0	0.0

Figura 5.4 Finestra dei parametri HL di una parte di pronuncia. Si ricorda che il parametro f0 è in decimi di Hz mentre le frequenze formanti sono inHz.

In Figura 5.5 è mostrato un esempio di finestra grafica. Sono rappresentati i valori dei parametri HL della stessa pronuncia di Figura 5.4. E' sufficiente selezionare con il mouse un punto di una curva di interesse per vederne visualizzati i valori di ascissa (tempo) e di ordinata (valore del parametro).

Altre utili funzionalità del software di controllo riguardano gli spettrogrammi e gli spettri delle pronunce. Si possono visualizzare in finestre grafiche la forma d'onda del segnale sintetizzato, il suo spettrogramma e il suo spettro. In Figura 5.6 ne è illustrato un esempio. Le quattro opzioni di calcolo e di visualizzazione possibili sono tutte attivabili cliccando con il tasto destro del mouse sulla finestra di interesse e selezionando una delle opzioni possibili dal menù che si apre. Tali opzioni sono:

- Pre-Emphasis: può essere abilitato o disabilitato il filtro di pre-enfasi nella visualizzazione dello spettro

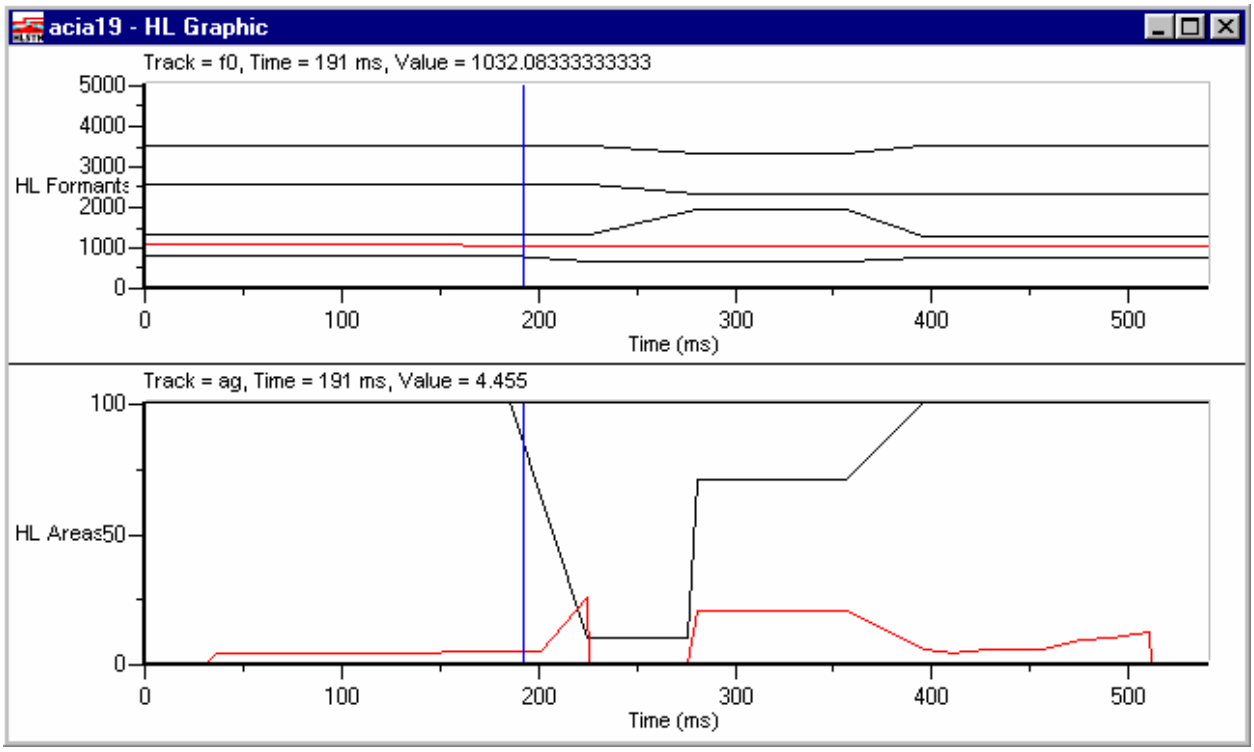


Figura 5.5 Finestra grafica dei parametri HL di una pronuncia VCV (vocale-consonante-vocale). Nel riquadro in alto sono graficati gli andamenti del pitch f0 e delle formanti, in basso le aree delle varie sezioni (ag, al,...). I valori (tempo e ampiezza) delle grandezze si possono leggere cliccando con il mouse sulla curva di interesse.

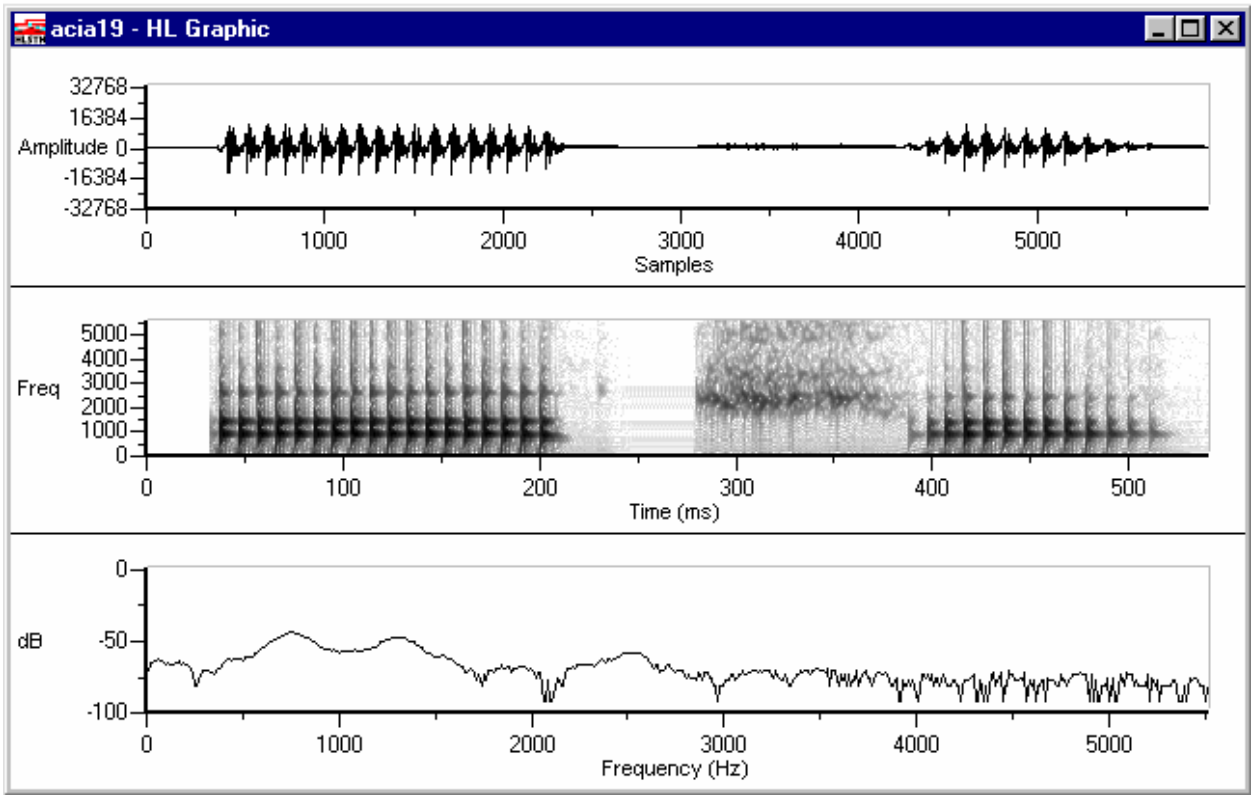


Figura 5.6 Finestra grafica della forma d'onda, dello spettrogramma e dello spettro (rispettivamente dall'alto in basso) di una pronuncia VCV.

- Window Size: si può impostare la dimensione (in numero di campioni) della finestra di Hamming per il calcolo dello spettro. Impostandolo a 64 campioni si ottiene uno spettro wide band mentre con una finestra di 512 si ha uno spettro narrow band
- Spectrum size: permette di scegliere il numero di campioni per il calcolo della FFT
- dB range: permette di aggiustare il livello di luminosità e contrasto dello spettrogramma per una visualizzazione ottimale

Tutti i valori caratteristici del singolo parlatore (elencati in Tabella 5.2) possono essere visualizzati e modificati aprendo l'apposita finestra con il comando 'KL Speaker' nel menù 'View'. Per impostare tutti i parametri di default del parlatore maschile o femminile è sufficiente selezionare il comando 'Generic Male Speaker' o 'Generic Female Speaker' dal menù 'Edit'.

Il software del sintetizzatore permette anche di selezionare la frequenza di campionamento e il numero di campioni per frame di analisi della pronuncia sintetizzata. Tali grandezze si possono modificare aprendo la finestra 'Document Info' nel menù 'View'. I valori usuali sono fc=10000 Hz con 50 campioni per frame o fc=11025 con 55 campioni per frame (sufficienti per l'analisi di un segnale vocale).

Una ultima considerazione riguarda la modalità di inserimento dei valori nelle finestre dei parametri. Purtroppo su questa versione non sono disponibili le familiari operazioni di 'taglia', 'copia' e 'incolla'. Ciò ha costretto all'inserimento manuale di tutti i valori dei parametri, anche se questi erano già disponibili da un altro file sintetizzato o, come spesso è accaduto visto il metodo di lavoro seguito, da un foglio di lavoro Excel.

5.2.3 Un aiuto alla sintesi: il software audio

Come è già stato detto nel Paragrafo 3.3 di grande aiuto alla sintesi delle consonanti affricate è stato il software **Sound Forge 4.5** della Sonic Foundry. Tale software permette, tra l'altro, la visualizzazione della forma d'onda del segnale audio con diversi fattori di zoom, l'ascolto totale e parziale del segnale (con una risoluzione temporale che arriva al singolo campione), l'analisi dello spettro del segnale ecc.

Particolarmente apprezzate sono state le possibilità di intervento sul segnale tramite varie funzioni di elaborazione. Molto utile si è rivelata la possibilità di effettuare semplicemente delle operazioni di "copia e incolla" su diverse porzioni di segnale e che hanno permesso di valutare i cambiamenti all'ascolto se ad una parte di pronuncia sintetica se ne sostituiva una naturale (ossia estratta dalle pronunce della base di dati) per capire come si doveva intervenire per modificare e migliorare il file sintetizzato. Un'altra delle possibilità offerte era quella di poter equalizzare e variare il volume di singoli tratti temporali del segnale, sempre per poter apprezzare come queste modifiche potessero influire sulla naturalezza della pronuncia sintetizzata.

La metodologia di lavoro seguita è stata in genere la seguente: da una base iniziale di una pronuncia sintetica si è provveduto all'esportazione del file nel noto formato audio .wav. Tale file è stato aperto con il Sound Forge e, dopo una ispezione visiva dell'andamento della forma d'onda, sono state effettuate le modifiche grazie alle funzionalità prima descritte. Si è poi provveduto all'ascolto in cuffia del segnale così ottenuto e, se il risultato era giudicato soddisfacente, si è tentato di ottenere un segnale dalle stesse caratteristiche tramite il sintetizzatore HLsyn modificando opportunamente i suoi parametri di controllo. Questo modo di operare ha portato, grazie ad una serie di affinamenti successivi, a quelle che sono le pronunce sintetizzate nel presente lavoro.

5.3 LA SINTESI DELLE CONSONANTI AFFRICATE

Le consonanti scelte per la sintesi sono [τΣ] e [τσ], la prima alveopalatale, la seconda dentale. Tali consonanti sono state sintetizzate coarticolate con la vocale [a], sia nella loro versione singola che in quella geminata. I dati sperimentali (vedi Capitolo 4) hanno messo in evidenza che il maggior carattere distintivo tra una pronuncia singola e una geminata è la durata dei fonemi. In base a ciò sono state cambiate soltanto le durate dei fonemi tra la pronuncia singola e la corrispondente geminata, lasciando inalterate le caratteristiche spettrali (prime fra tutti le frequenze formanti). Dato che sono state sintetizzate voci maschili sono state prese come riferimento le durate medie dei fonemi dei tre parlatori uomini. Tali valori sono riportati in Tabella 5.3.

	V1d	C1d	C2d	V2d	Utd
♋◆●♋	173	73	96	114	456
♋◆◆●♋	115	138	124	101	478
♋◆• ♋	113	91	146	105	455
♋◆◆• ♋	104	114	160	118	496

Tabella 5.3 Durate medie dei fonemi delle pronunce maschili sintetizzate. Tutti i valori sono in msec.

Per il calcolo degli istanti temporali è stata automatizzata una procedura tramite un foglio di calcolo Excel. E' sufficiente inserire nel foglio le durate dei fonemi della pronuncia e automaticamente si ottengono gli istanti temporali da inserire nel sintetizzatore HLsyn insieme ai corrispondenti valori dei parametri.

5.3.1 Sintesi della vocale [a]

Per sintetizzare la vocale [a] sono stati considerati innanzitutto i valori medi del pitch e delle formanti misurati in sede sperimentale. Tali valori sono stati inseriti come parametri di controllo nel software del sintetizzatore. L'elenco completo delle grandezze di controllo con i valori ad essi assegnati sono riportati in Tabella 5.4.

Parametro	**ag**	**al**	**ab**	**an**	**ue**	**f0**	**f1**	**f2**	**f3**	**f4**	**ps**	**dc**	**ap**
Valore	4	100	100	0	0	1070	750	1300	2500	3500	8	0	0

Tabella 5.4 Elenco completo dei parametri e valori loro assegnat per la sintesi della vocale [a].

C'è da dire che tali valori non corrispondono esattamente alle medie delle grandezze misurate in quanto si è visto (o meglio ascoltato) che la pronuncia sintetizzata non era naturale. In particolare è stato abbassato il pitch (come si vede a 107[1] Hz) in quanto, con il valore medio sperimentale, la voce sembrava più femminile che maschile. **ag** è stato impostato a 4, un valore tipico per la produzione di suoni sonori,

[1] Ricordiamo che l'unità di misura di f0 è decimi di Hertz, quindi scrivere 1070 equivale a 107 Hz.

mentre, ovviamente, non ci sono costrizioni nell'apparato fonatorio (**al**=100, **ab**=100) ad esclusione di **an** in quanto non è nostra intenzione produrre una vocale nasalizzata. Anche **ue** è stato posto pari a zero in quanto interviene solo nella produzione di fonemi occlusivi sonori. Per gli ultimi tre parametri (**ps**, **dc** e **ap**) sono stati utilizzati i valori di default.

5.3.2 Sintesi della pronuncia ατΣα

Illustriamo ora quali sono le considerazioni che hanno portato alla scelta di determinati valori dei parametri di controllo nella sintesi della pronuncia ατΣα. Innanzitutto facciamo alcune considerazioni su come si modifica l'apparato fonatorio nella produzione della consonante. Si parte dalla posizione della [a] in cui non ci sono costrizioni al passaggio dell'aria. Progressivamente la lingua si alza verso il palato per produrre l'occlusione (sorda) della prima parte della consonante (indicata nel corso della tesi con C1). Successivamente, passando alla fase fricativa della consonante, avviene il rilascio dell'occlusione da parte della lingua, spostandosi solo di quel tanto che basta per lasciar passare l'aria e causarne la frizione costringendola nello stretto passaggio del contoide fricativo [Σ]. Infine c'è l'abbassamento della lingua che causa la fine della fricazione e il passaggio alla seconda vocale [a]. In Tabella 5.5 sono illustrati tutti i valori dei parametri utilizzati nella pronuncia ατΣα e nella corrispondente geminata. Tra le due cambiano soltanto gli istanti temporali, indicati sulle due colonne più a sinistra.

t. sing.	t. gem.	ag	al	ab	an	ue	f0	f1	f2	f3	f4	ps	dc	ap
0	0	0	100	100	0	0	1070	750	1300	2500	3500	8	0	0
30	30	0												
35	35	4												
100	78	4					1070							
185	125		100	100				750	1300					
200	140	4.5												
220	160						1040							
224	164	25	30	10										
225	165	0						600	1800	2500	3500			
275	285	0		10								8		
280	290	20		90						2400	3300			
295	305											20		
355	395	20	30	90				600	1800	2400	3300			
375	415	5												
395	435	5	100	100			1020	740	1250	2500	3500	7		
407	443	4										8		
416	451	4.5												
426	459	5												
449	479	5.5												
453	483	6												
467	495	9												
480	505	10												
500	525	12												
501	526	0												
530	555	0	100	100	0	0	1000	740	1250	2500	3500	8	0	0

Tabella 5.5 Sintesi completa della pronuncia ατΣα nella sua versione singola e geminata. Gli istanti temporali nelle prime due colonne sono in msec. I 13 parametri di sintesi sono espressi ognuno nella sua unità di misura. Sono visualizzati solo i valori inseriti dall'utente, nelle celle vuote il valore viene calcolato automaticamente per interpolazione lineare dal software del sintetizzatore.

Giustifichiamo ora i valori assegnati ai parametri.

- **ag**: dal valore 4, caratteristico di una pronuncia sonora e utilizzato per la vocale [a], si passa a zero in corrispondenza alla fase occlusiva della consonante. Successivamente, per la seconda parte della consonante (ossia quella fricativa) si è portato a 20 in modo da permettere il passaggio d'aria attraverso la glottide senza la produzione di alcuna sonorità.[2] Per i valori assegnati durante la seconda vocale c'è da dire che si doveva ottenere una riduzione dell'ampiezza del segnale andando verso la fine della pronuncia. Ciò è stato ottenuto aumentando progressivamente la sezione della glottide, riducendo così l'ampiezza e la forza delle vibrazioni delle corde vocali.
- **al** e **ab**: si è tentato di dare a questi due parametri un andamento che rispecchiasse il più possibile l'effettivo movimento della punta della lingua e delle labbra. Sono stati poi necessari degli aggiustamenti per riuscire ad avere una pronuncia il più naturale possibile. Ciò ha portato a dei valori per **al** e **ab** che forse non rispecchiano in maniera assolutamente fedele quello che accade all'apparato fonatorio umano durante la produzione di una pronuncia $\alpha\tau\Sigma\alpha$ ma che sicuramente hanno permesso di ottenere un suono molto più naturale e realistico.
- **an** e **ue**: tali grandezze non sono intervenute in quanto la pronuncia non doveva essere nasalizzata e non esiste la sonorità durante la fase occlusiva della consonante.

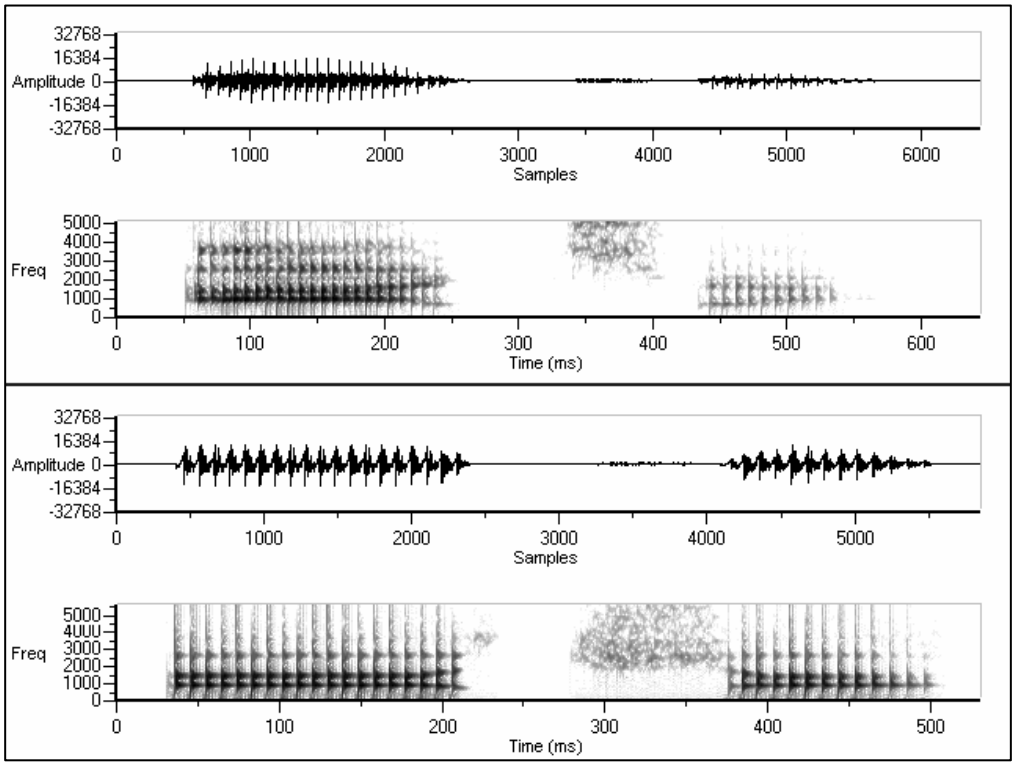


Figura 5.7 Forme d'onda e spettrogrammi di una pronuncia $\alpha\tau\Sigma\alpha$ registrata (sopra) e di quella sintetizzata (sotto).

[2] Si ricorda che la consonante $\tau\Sigma$ è sorda, ossia non c'è vibrazione delle corde vocali durante la sua produzione.

- **f0**: alla frequenza di pitch è stato dato un andamento decrescente in modo da simulare una parola con accento sulla prima sillaba (caratteristica delle pronunce della base di dati analizzate).
- **f1**, **f2**, **f3** e **f4**: alle quattro frequenze formanti è stato dato l'andamento caratteristico osservato durante l'analisi dei dati. In particolare si notino l'abbassamento di **f1** e l'innalzamento di **f2** ai lati della consonante. Per quanto riguarda i valori assegnati durante la fase fricativa della consonante c'è da dire che non si disponeva né dei dati sperimentali (in quanto non misurati), né di valori trovati in bibliografia. Si è allora proceduto ad un affinamento successivo basato su prove di ascolto, aiutandosi anche con gli spettrogrammi delle pronunce della base di dati e di quelle sintetizzate.
- **ps**: si è data una pressione maggiore in corrispondenza della fase fricativa della consonante in quanto si è visto che con il valore di default pari a 8 il volume della consonante era troppo basso rispetto alla vocale. E' stato poi fatto scendere fino a 7 alla fine della consonante per avere un andamento della forma d'onda più simile a quello che è nella realtà.
- **dc** e **ap**: sono stati lasciati i valori di default per tutta la durata della pronuncia.

In Figura 5.7 sono riportati la forma d'onda e lo spettrogramma di una pronuncia registrata e della pronuncia sintetizzata.

5.3.3 Sintesi della pronuncia atsa

Anche per la pronuncia atsa (e la sua geminata) le varie fasi della fonazione sono molto simili a quelle di αtΣα. In questo caso però ad alzarsi per provocare l'occlusione è più la punta della lingua che non tutto il suo corpo. Il punto di occlusione si porta più in avanti coinvolgendo la punta della lingua con i denti. Anche al momento del rilascio dell'occlusione, la fase fricativa è caratterizzata da un contenuto spettrale più spostato verso le alte frequenze, come si può anche vedere dagli spettrogrammi delle pronunce della base di dati. In Tabella 5.6 sono illustrati tutti i valori dei parametri utilizzati nella pronuncia atsa e nella corrispondente geminata.

Per giustificare le scelte fatte valgono considerazioni simili a quelle fatte per la sintesi della pronuncia αtΣα. Possiamo solo far notare le differenze principali che riguardano:

- il parametro **al**, che in questo caso non varia durante la pronuncia in quanto la chiusura delle labbra è minore;
- i valori delle formanti durante la fase fricativa, in quanto il rumore viene prodotto dal rilascio di una occlusione formata dalla lingua a contatto con i denti e non con il palato. In particolare **f3** e **f4** hanno un valore maggiore in quanto il contenuto frequenziale di [ts] è spostato più in alto rispetto a [tΣ];
- non si è reso necessario aumentare **ps** in corrispondenza della fase fricativa della consonante in quanto l'ampiezza del segnale si è ritenuta più che sufficiente. E' stato addirittura leggermente ridotto così da avere un migliore equilibrio dell'intera pronuncia.

In Figura 5.8 sono riportati la forma d'onda e lo spettrogramma di una pronuncia registrata e della pronuncia sintetizzata.

atsa - attsa														
t. sing.	**t. gem.**	**ag**	**al**	**ab**	**an**	**ue**	**f0**	**f1**	**f2**	**f3**	**f4**	**ps**	**dc**	**ap**
0	0	0	100	100	0	0	1070	750	1300	2500	3500	8	0	0
30	30	0												
35	35	4												
78	74	4					1070							
125	115			100										
140	130	4.5						750						
160	150						1040							
164	154	25		10										
165	155	0						600	1300	2500		8		
235	250	0		10										
240	255	20		35				400	1600	2670		7		
365	395	20		35				400	1600	2670		7		
385	415	5												
405	435	5		100			1020	740	1250	2500		8		
414	447	4												
423	456	4.5												
431	466	5												
452	489	5.5												
456	493	6												
469	507	9												
480	520	10												
500	540	12												
501	541	0												
530	570	0	100	100	0	0	1000	740	1250	2500	3500	8	0	0

Tabella 5.6 Sintesi completa della pronuncia atsa nella sua versione singola e geminata. Gli istanti temporali nelle prime due colonne sono in msec. I 13 parametri di sintesi sono espressi ognuno nella sua unità di misura. Sono visualizzati solo i valori inseriti dall'utente, nelle celle vuote il valore viene calcolato automaticamente per interpolazione lineare dal software del sintetizzatore.

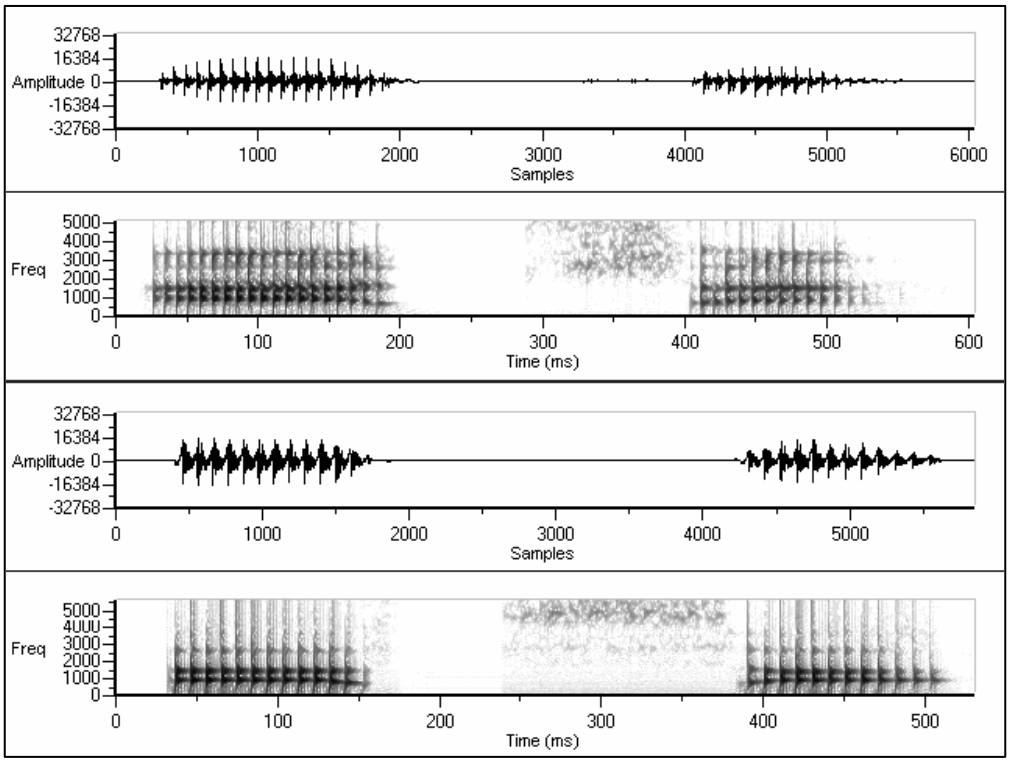


Figura 5.8 Forme d'onda e spettrogrammi di una pronuncia atsa registrata (sopra) e di quella sintetizzata (sotto).

CAPITOLO 6

CONFRONTI E CONCLUSIONI

INTRODUZIONE

Nel capitolo quattro sono stati descritti in maniera dettagliata i risultati dell'analisi acustica condotta sulle consonanti affricate italiane, con particolare attenzione al fenomeno della geminazione. In questo capitolo finale si riprenderanno i risultati più importanti di questo lavoro per poi confrontarli con quelli degli studi svolti per altre classi di consonanti nell'ambito del progetto GEMMA. Inoltre si confronteranno i risultati ottenuti anche con quelli di studi sulla geminazione in lingue diverse dall'Italiano. Considerazioni verranno fatte anche sulla sintesi delle pronunce affricate con una prima valutazione qualitativa dei risultati ottenuti. Infine saranno dati alcuni spunti per ulteriori ricerche.

6.1 RIEPILOGO DEI RISULTATI DELL'ANALISI SULLA GEMINAZIONE DELLE CONSONANTI AFFRICATE

Riassumiamo brevemente i risultati delle misure sperimentali e delle analisi condotte (riportati per intero nel Capitolo 4) sulla geminazione delle consonanti affricate.

1. **Risultati nel dominio del tempo**: le durate dei fonemi che presentano differenze statisticamente significative sono quelle della prima vocale e della consonante (sia della sua fase occlusiva che di quella fricativa). Si ha infatti una diminuzione di V1d e un aumento di C1d e C2d passando dalla pronuncia singola a quella geminata. Le durate della prima vocale e della consonante sono allora legate in maniera inversa, con un coefficiente di correlazione di Spearman r_s=-0.7. Considerando invece separatamente le due fasi della consonante, si ottiene per la correlazione con la durata della vocale un valore pari a -0.47 (lo stesso sia tra V1d e C1d che tra V1d e C2d). Anche la durata dell'intera pronuncia è un parametro che si è rivelato dipendere in maniera statisticamente significativa dalla geminazione. Questa dipendenza è meno forte che non per i precedenti

parametri. I risultati appena riepilogati hanno portato all'ipotesi che vi sia un effetto di compensazione tra le durate dei fonemi che, però, non appare completo.

2. **Risultati nel dominio della frequenza**: non ci sono differenze statisticamente significative tra i valori misurati ad eccezione del pitch in due specifici frame. F0 è di 14 Hz e 12 Hz più alto nella forma geminata (+9% e +8%) nei frame V1 offset e V1 offset to C rispettivamente. Non si sono osservate variazioni nella frequenza delle formanti F1, F2 e F3 mentre le loro ampiezze A1, A2 e A3 sono significativamente più alte (di circa 1-3 dB) nelle pronunce geminate. Ciò si è osservato nei frame V1 center, V1 offset, V1 offset to C e V2 onset.
3. **Risultati nel dominio energetico**: Senza scendere troppo in dettagli si può dire che, guardando la Tabella 4.10, c'è la tendenza a pronunciare con maggiore enfasi la parola geminata, fatto confermato dall'ampiezza delle formanti (vedi punto precedente). Dai risultati del test di correlazione di Spearman non ci sentiamo comunque di trarre delle conclusioni perentorie riguardo ai risultati ottenuti in quanto i valori di correlazione non sono così alti da far pensare a dei forti legami tra le grandezze energetiche e il fenomeno della geminazione.

Per quanto riguarda i risultati della classificazione delle pronunce, che verranno esposti più in dettaglio nel prossimo paragrafo insieme ai confronti con gli altri lavori, esiti appena soddisfacenti si ottengono solo con le durate dei fonemi. Assolutamente pessime sono le classificazioni basate sui dati frequenziali ed energetici, come già messo in evidenza nel Capitolo 4.

6.2 CONFRONTO TRA GLI EFFETTI DELLA GEMINAZIONE NELLE DIVERSE CLASSI DELLE CONSONANTI ITALIANE

Innanzitutto, prima di procedere con i confronti tra i risultati del presente studio e quelli precedenti, è d'obbligo una premessa. Le consonanti affricate presentano delle caratteristiche distintive molto particolari. Come infatti già precisato in precedenza, e come si può vedere anche dalle specifiche grandezze scelte per l'analisi, è stato necessario dividere in due parti la consonante: la prima, indicata con C1, che rappresenta la fase occlusiva; la seconda, indicata con C2, che rappresenta la fase fricativa. Tale necessità non si era manifestata nei precedenti lavori sulla geminazione delle consonanti italiane. Di conseguenza non sempre sarà possibile fare un confronto e trovare un riscontro diretto con i risultati ottenuti per le altre pronunce.

Fatta questa fondamentale osservazione, procederemo con i confronti facendo riferimento ai lavori sulle consonanti occlusive [p, b, t, d, c, g] (A. Vannucci 1993; R. Rossetti, 1993), sulle consonati liquide [l, r] (F. Argiolas, 1995; F. Macrì 1995), sulle consonanti fricative [f, v, s, z, Σ] (M.Giovanardi, 1998) e sulle consonanti nasali [m, n] (M. Mattei, 1999).

Una prima osservazione riguarda la variazione delle durate dei fonemi nella geminazione. In tutti gli studi precedenti è stato osservato una diminuzione della durata della prima vocale e un aumento della durata della consonante passando dalle pronunce singole a quelle geminate. In Tabella 6.1 sono riportate le durate dei fonemi misurate per le altre classi di consonanti e i rapporti Cd/V1d mentre in Tabella 6.2 si possono vedere le durate delle pronunce con consonanti affricate e i valori dei rapporti C1d/V1d,

C2d/V1d e Cd/V1d. In questa ultima Tabella è riportata anche la durata totale della consonante così da permettere un confronto diretto.

	OCCLUSIVE			LIQUIDE			FRICATIVE			NASALI		
	V1d	Cd	Cd/V1d	V1d	Cd	Cd/V1d	V1d	Cd	Cd/V1d	V1d	Cd	Cd/V1d
Singole	168	91	0.57	171	61	0.37	176	135	0.8	184	91	0.51
Geminate	125	182	1.56	122	174	1.52	127	233	1.97	125	212	1.78

Tabella 6.1 Durate dei fonemi V1d e Cd e rapporto Cd/V1d delle pronunce con consonanti occlusive, liquide, fricative e nasali. V1d e Cd sono in msec, Cd/V1d è adimensionale.

	AFFRICATE						
	V1d	C1d	C2d	Cd	C1d/V1d	C2d/V1d	Cd/V1d
Singole	150	82	95	177	0.55	0.63	1.18
Geminate	111	133	122	255	1.20	1.10	2.30

Tabella 6.2 Durate dei fonemi per le pronunce affricate. Le grandezze sono in msec ad esclusione dei rapporti che sono adimensionali.

Come prima considerazione si può notare nelle affricate una minore "separazione" dei valori medi tra le pronunce singole e le geminate. Si hanno infatti le seguenti differenze:

- occlusive: ΔV1d = 168-125 = 43 ms (-26% gem.); ΔCd = 182-91 = 91 ms; (+100% gem.);
- liquide: ΔV1d = 171-122 = 49 ms (-29% gem.); ΔCd = 174-61 = 113 ms (+185% gem.);
- fricative: ΔV1d = 176-127 = 49 ms (-28% gem.); ΔCd = 233-135 = 98 ms (+73% gem.);
- nasali: ΔV1d = 184-125 = 59 ms (-32% gem.); ΔCd = 212-91 = 121 ms (+133% gem.);
- affricate: ΔV1d = 150-111 = 39 ms (-26% gem.); ΔCd = 255-177 = 78 ms (+44% gem.);

Come si vede le differenze di durata di V1 e C tra pronunce singole e geminate per le consonanti affricate sono le più piccole (sia in valore assoluto che in percentuale). Ciò è una giustificazione dei risultati ottenuti nella classificazione delle pronunce. Infatti sono state ottenute le seguenti percentuali di errore (non sono incluse le liquide poiché questo dato non è disponibile):

- occlusive: 4% su Cd e 8% su Cd/V1d
- fricative: 12% sia su Cd che su Cd/V1d
- nasali: 0.47% sia su Cd che su Cd/V1d
- affricate: 17.6% su Cd e 16.7% su Cd/V1d

Andando invece a considerare il rapporto C1d/V1d (caratteristico delle consonanti affricate) la percentuale di errore scende al 13.9%. Inoltre c'è da dire che separando le diverse vocali e consonanti si ottengono dei risultati decisamente migliori, come ad esempio lo 0% di errori per la consonante [δZ], calcolato sul parametro C1d/V1d.

Per avere un'idea più immediata della maggiore difficoltà di classificazione delle affricate rispetto, ad esempio, alle nasali, riportiamo nelle Figure 6.1 e 6.2 i grafici a dispersione nel piano bidimensionale V1d-Cd per entrambe le classi di consonanti. Si nota che la separazione tra singole e geminate è molto più netta nelle nasali che non nelle affricate.

In Figura 6.3 è inoltre riportato per completezza il diagramma a dispersione tra V1d e C1d, poiché la classificazione su C1d/V1d è quella che ha dato i migliori risultati.

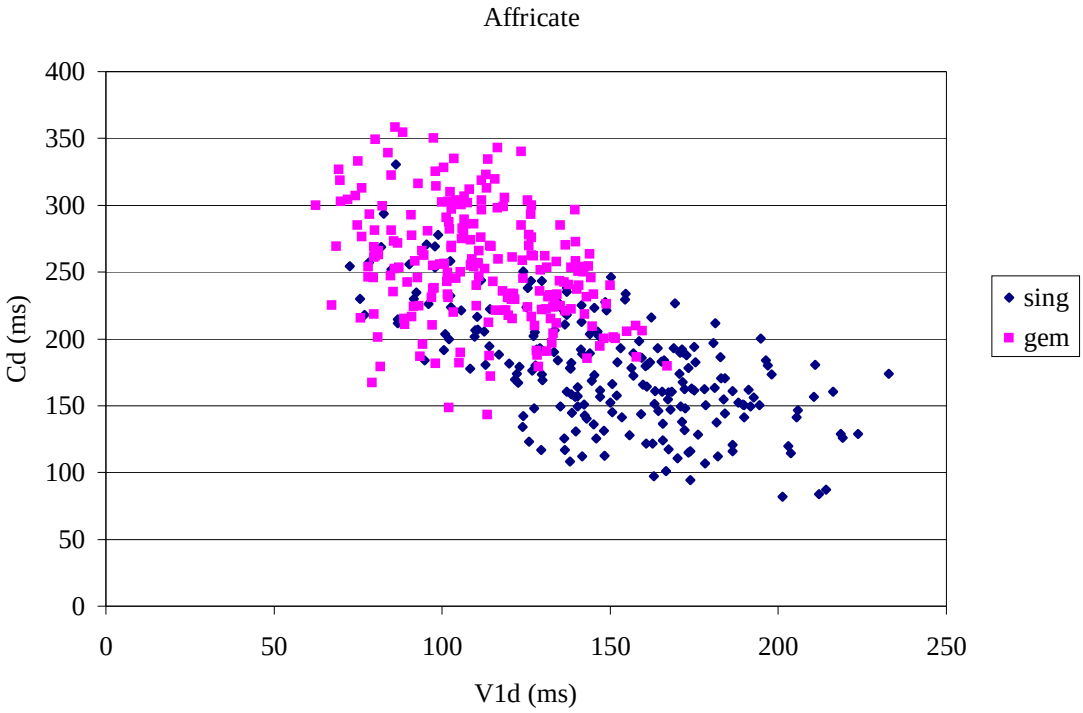


Figura 6.1 Diagramma a dispersione tra V1d e Cd per le consonanti affricate (216 singole e 216 geminate).

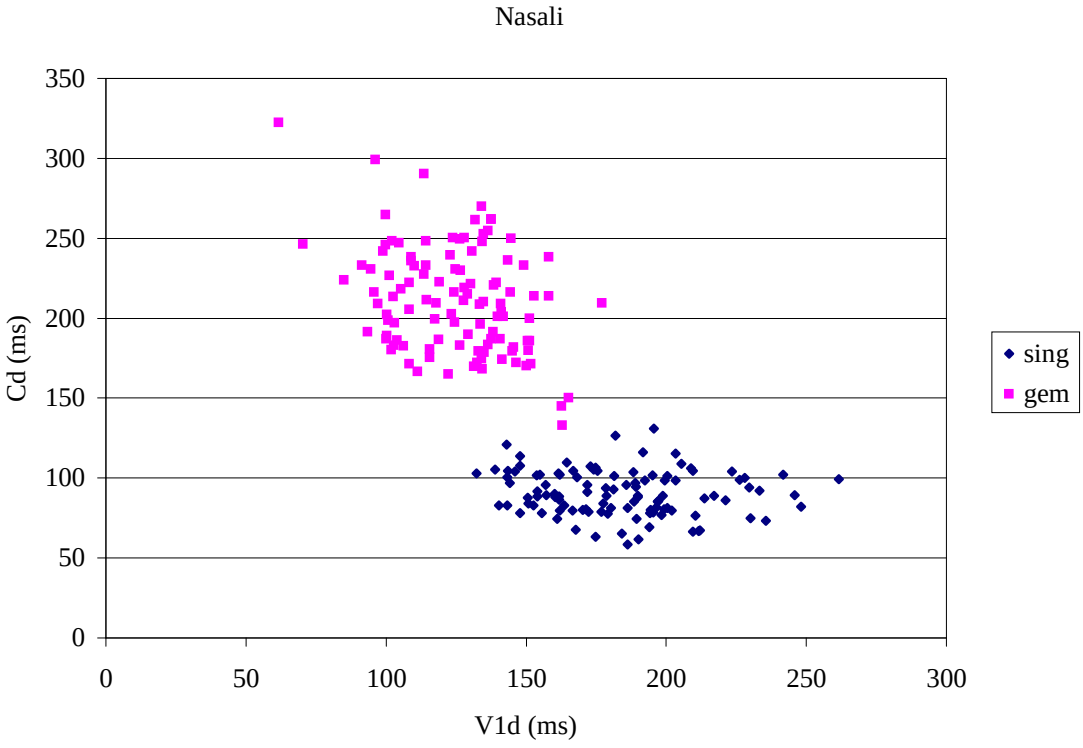


Figura 6.2 Diagramma a dispersione tra V1d e Cd per le consonanti nasali (108 singole e 108 geminate).

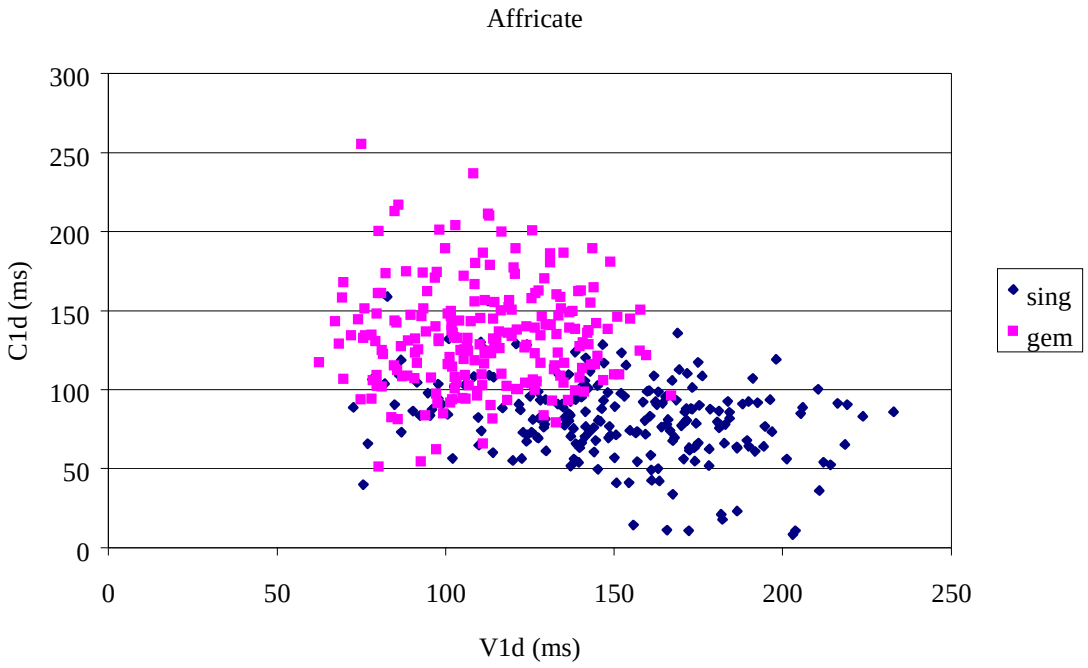


Figura 6.3 Diagramma a dispersione tra V1d e C1d per le consonanti affricate (216 singole e 216 geminate).

Dall'osservazione della tabella 6.1 si trae un'altra interessante considerazione: il rapporto medio Cd/V1d è per le quattro classi di consonanti inferiore ad 1 nel caso delle singole e superiore ad 1 nel caso delle geminate. Ciò non avviene per le consonanti affricate (Cd/V1d=1.18 per le singole e Cd/V1d=2.30 per le geminate). Questo è dovuto soprattutto alla maggiore durata della consonante. L'ipotesi fatta nei precedenti lavori che l'intenzione astratta che il parlatore ha nel produrre una geminata si traduca nella realizzazione di un fonema consonantico che sia almeno più lungo di quello che lo precede, ora non è più valida.

Un ulteriore riscontro a quanto detto può essere ottenuto dalla seguente considerazione: una classificazione su 1512 pronunce appartenenti alle sopraindicate classi di consonanti (occlusive, liquide, fricative e nasali) sulla base di Cd/V1d ha portato ad un minimo di errori commessi pari al 7.2% in corrispondenza ad un valore del suddetto rapporto pari a 1.03 (molto prossimo quindi all'unità). Per le consonanti affricate il punto di equiprobabilità per il rapporto Cd/V1d che minimizza il numero di errori vale invece 1.6. Considerando invece separatamente le due parti della consonante e analizzando i rapporti C1d/V1d e C2d/V1d, si vede che i punti di equiprobabilità valgono circa 0.77. Ciò avvalora l'ipotesi che le affricate vadano considerate e classificate in maniera a sé stante rispetto agli altri gruppi di consonanti italiane.

Concludiamo con un'ultima osservazione. Come ci si poteva attendere da quanto detto finora, osservando le matrici di correlazione, costruite per le occlusive (A. Vannucci, 1993; R. Rossetti, 1993), le fricative (M. Giovanardi, 1998) e le nasali (M. Mattei, 1999), si vede che i risultati sono perfettamente analoghi in tutte e tre le classi di consonanti. In particolare il valore del coefficiente di correlazione tra Cd e V1d oscilla tra -0,71 e -0,78. Per le affricate tale valore è -0.7 (in linea con quelli precedenti) mentre è

minore se la consonante viene considerata divisa nelle sue due parti (r_s = -0.47 sia tra V1d e C1d che tra V1d e C2d).

6.3 CONFRONTO TRA GLI EFFETTI DELLA GEMINAZIONE NELL'ITALIANO E IN ALTRE LINGUE.

Come è stato detto in precedenza, il fenomeno della geminazione è caratteristico della lingua italiana. Tuttavia la geminazione risulta un argomento di particolare interesse anche per studiosi non italiani. Un motivo di ciò può essere individuato nel fatto che la geminazione è forse l'unico carattere distintivo legato soprattutto ad aspetti prosodici. Sta di fatto che sono molti gli studi condotti sul fenomeno nelle più disparate lingue e dialetti e da molti punti di vista.

Ad esempio citiamo lo studio condotto da Shrotriya et al. (1995), sulle consonanti occlusive dell'Hindi. Anche in questo lavoro è stato trovato un significativo allungamento della consonante nelle pronunce geminate. Citiamo, inoltre, altri studi sulla geminazione in lingue diverse dall'Italiano: quello di Blumstein et al. (1998), e quello di Rochet e Rochet (1995).

E' inoltre doveroso in questa sede ricordare che si è tenuto recentemente (Agosto 1999) un simposio sulla geminazione nelle lingue presso l'International Conference of Phonetic Sciences a San Francisco. Gli articoli presentati alla conferenza, si riferivano a tre dialetti indonesiani (Cohn et al., 1999), al Pattani Malay (Abramson, 1999), al Malayalam (Local e Simpson, 1999), al Greco cipriota (Arvaniti, 1999) e al Berbero (Louali e Maddieson, 1999).

Molti dei risultati presentati negli articoli appena citati sono in accordo con quelli ottenuti per l'Italiano; in particolare si è trovato che, sia per i dialetti indonesiani,sia per il Greco cipriota, la durata è il principale correlato acustico per la classificazione delle pronunce singole e geminate. Lo studio condotto sul Pattani Malay (Abramson, 1999) focalizza la propria attenzione sulle variazioni di F0 in relazione a pronunce che presentano la geminazione della consonante iniziale (fenomeno tra l'altro inesistente nell'italiano). Il risultato di questo studio indica che c'è una variazione significativa della F0 in dipendenza della geminazione ma non per tutte le classi di consonanti. In particolare le affricate non sono state studiate in quanto si è visto, da un test preliminare, che la percentuale di errori commessi nel riconoscimento di pronunce singole o geminate era la più alta di tutte le classi di consonanti. Lo studio sul Malayalam (Local e Simpson, 1999) si discosta leggermente dai risultati degli altri studi contraddicendo l'affermazione che la durata è il principale correlato della geminazione. In particolare per il Malayalam sono risultati significativi aspetti legati sia al tempo che alla frequenza. Infine, lo studio sul Berbero si è interessato del problema della classificazione delle occlusive geminate anche quando, in alcuni dialetti, non esistono più le corrispondenti singole che nei secoli sono diventate aspirate. I risultati di questo studio indicano che è appropriato considerare queste consonanti ancora come geminate e che esse sono effettivamente caratterizzate da una durata dell'occlusione superiore a quella delle occlusive singole che ancora esistono in altri dialetti berberi.

6.4 CONSIDERAZIONI SULLE PRONUNCE SINTETIZZATE

Nel presente lavoro si sono sintetizzate per la prima volta pronunce singole e geminate di consonanti affricate italiane con il sintetizzatore articolatorio HLsyn. Anche se non è stata condotta una analisi percettiva sui risultati ottenuti, si può affermare che è sufficiente cambiare le durate dei fonemi, senza modificare in alcun modo i parametri spettrali, per ottenere una pronuncia geminata dalla corrispondente singola. Le pronunce così sintetizzate sono chiaramente riconoscibili come singole o geminate variando in maniera opportuna soltanto i parametri temporali. Ciò avvalora l'ipotesi che i principali tratti distintivi tra una pronuncia singola e una geminata vadano ricercati nelle durate dei fonemi. Tale ipotesi potrebbe essere definitivamente confermata conducendo un esperimento di analisi percettiva che permetterebbe, tra l'altro, di stabilire come l'orecchio umano coglie le variazioni temporali tra una pronuncia e l'altra. In particolare, per le consonanti affricate, potrebbe essere interessante studiare cosa succede non solo variando la durata di tutta la consonante ma cambiando i rapporti tra le durate delle due fasi (occlusiva e fricativa) di cui è composta la consonante stessa.

Un'ultima considerazione riguardo alle due particolari pronunce sintetizzate. Mentre per la ατΣα la separazione tra le durate dei fonemi della pronuncia singola e di quella geminata è molto netta e porta ad una facile distinzione delle due, ciò non può dirsi per la pronuncia atsa (ricordiamo che le durate medie dei fonemi utilizzate per la sintesi sono quelle ottenute dalle corrispondenti pronunce registrate). Infatti per quest'ultima, come si può vedere dalle durate dei fonemi, la separazione tra pronuncia singola e geminata non è così netta, portando ad interpretare per geminata la pronuncia singola. Probabilmente ciò è dovuto al fatto che in effetti, anche quando si deve pronunciare una consonante [ts] singola, si tende a dare una certa enfasi al fonema stesso che tende ad allungarne la durata. Una risposta a questo quesito potrebbe venire da un appropriato esperimento percettivo finalizzato allo studio della geminazione del fonema consonantico [ts].

6.5 CONCLUSIONI

In base a quanto emerso dal presente studio sulla geminazione delle consonanti affricate italiane, dal confronto con le altri classi di consonanti e con gli studi su altre lingue possiamo riassumere brevemente i risultati principali come segue:

- la classificazione delle affricate basata su parametri temporali risulta più difficile che per le altre classi di consonanti;
- al contrario di tutti gli altri lavori sulla geminazione delle consonanti italiane, non si ha il valore distintivo del rapporto Cd/V1d molto prossimo all'unità;
- la dipendenza della geminazione da parametri di durata è ricorrente in tutte le lingue citate nel Paragrafo 6.3;
- la sintesi delle consonanti affricate ha messo in risalto i parametri temporali come i principali correlati acustici alla geminazione.

In conclusione desidero ringraziare la Professoressa Di Benedetto per l'aiuto datomi nella stesura del presente lavoro e dell'articolo "Acoustic analysis of singleton and geminate affricates in Italian" in corso di pubblicazione sul journal "The European Student Journal of Language and Speech" e per la sua disponibilità ad ascoltare e risolvere i vari problemi incontrati. Un grazie anche a Marco Mattei che è stato di fondamentale aiuto, tra l'altro, negli affinamenti delle pronunce sintetizzate.

6.6 SPUNTI PER RICERCHE FUTURE

Eventuali ricerche future potrebbero orientarsi sui seguenti punti:

- condurre sulle affricate un esperimento percettivo per indagare sui valori ottimi dei rapporti Cd/V1d e C1d/V1d che discriminano le singole dalle geminate;
- analizzare i tratti distintivi dei rapporti Cd/V1d e C1d/V1d sulla geminazione in funzione dello *speaking rate*;
- analizzare le correlazioni tra le durate dei fonemi in parole intere più lunghe dei semplici bisillabi o addirittura all'interno di frasi complete;
- studiare come gli elementi prosodici influenzino le caratteristiche (temporali e spettrali) del segnale vocale;
- sfruttare tutti i dati raccolti nell'ambito del progetto GEMMA (ormai disponibili per tutte le classi di consonanti italiane) per progettare e implementare un sistema di riconoscimento o un sintetizzatore vocale per l'Italiano per scopi generali.

BIBLIOGRAFIA

Arthur S. Abramson, "Fundamental frequency as a cue to word-initial consonant length: Pattani Malay", ICPhS99 San Francisco pp 591-594, 1999.

Francesca. Argiolas, "Analisi acustica e percettiva delle consonanti liquide [l, r] in italiano", Tesi Univ. di Roma "La Sapienza", 1995.

Francesca Argiolas, Federico Macrì, M.G. Di Benedetto, "Acoustic analysis of Italian [r] and [l]", Journal of the Acoustical Society of America 97, no. 5, pt.2, pp.3418, 1995.

Amalia Arvaniti, "Effects of speaking rate on the timing of single and geminate sonorants", ICPhS99 San Francisco pp 599-602, 1999.

M. Bertinetto, E. Vivalda, "Recherches sur les oppositions des quantité en Italien", Journal of Italian Linguistics, No. 3, 1991, pp. 97-119.

Blumstein S.E., Pickett E., Burton M., "Effects of speaking rate on Singleton/Geminate consonant contrast in Italian", unpublished manuscript, 1998.

Brozovic D., "Sull'inventario dei fonemi serbocroati e i loro tratti distintivi", in "WSI", XII, pp161.172, 1967.

L. Canepari, "Introduzione alla fonetica", Einaudi, 1979.

L. Canepari, "Manuale di pronuncia italiana", Ed. Zanichelli, 1992.

Abigail C. Cohn, William H. Ham, Robert J. Podesva, "The phonetic realization of singleton-geminate contrasts in three languages of Indonesia", ICPhS99 San Francisco pp 587-590, 1999.

Giuseppe Cicchitelli, "Probabilità e statistica" Maggioli editore, 1984.

R. Carlson, B. Granström, "A phoneticaly oriented programming language for rule description of speech", Speech communication, vol. 2 pp. 245-253, 1975.

R. Carlson, B. Granström, "A text-to-speech system based entirely on rules", I.C.A.S.S.P., 1976.

R. Carlson, B. Granström, "A multi-language text-to-speech module", I.C.A.S.S.P., 1982.

G. N. Clements, "The geometry of phonological features", Phonology 2, 1985.

Clifford, "Microphones (3rd edition)", Tab books inc., 1986.

W. R. Dillon, M. Goldstein, "Multivariate analysis", J. Wiley & Sons, 1984.

Di Pietro R.J., "Phonemics, Generative grammar and the italian sibilants", in "SL", XXI pp. 96,106, 1967.

Esposito A., Di Benedetto M.G., "Acoustic and Perceptual Study of Gemination in Italian Stops", Journal of the Acoustical Society of America, 1999.

G. Fant, "Acoustic theory of speech production", Mounton and Company, Gravenhage, 1960.

Giovanni Flammia, "Classificazione statistica e neurale su base percettiva del riconoscimento delle vocali italiane", Tesi Univ. di Roma "La Sapienza", 1988.

J. L. Flanagan, A. Rosemberg, "Effecct of glottal pulse shape on the quality of natural vowels", J.A.S.A. 53, 1971.

J. Flanagan, L. R. Rabiner, "Speech synthesis", Stroudsburg, 1973.

Luisa Franchina, Piero Marietti, "Sistemi elettronici a banda frazionale stretta", Masson Editore, p. 239, 1994.

Fujimura O. e Lindqvist G, "Sweep-tone measurements of vocal tract characteristics", Journal of the Acoustical Society of America, Vol 49, No. 2, pp 541-558, 1971.

Giovanardi M., "Analisi Acustica e Sintesi delle consonanti fricative singole e geminate in Italiano", Tesi Univ. di Roma "La Sapienza", 1998.

Giovanardi M.," Acoustic analysis of singleton and geminate fricatives in Italian" European student journal of language and speech, 1998.

M. Halle, J. R. Vergnaud, "Three-dimensional phonology", J. Ling. Res. 1, 1980.

HLSYN, manuale di riferimento, 1997.

Al Kelley, Ira Pohl, "Didattica e programmazione C" Addison-Wesley, 1996.

B. W. Kernighan, D. M. Ritchie, "Linguaggio C", Jackson, 1990.

Kewley-Port D. and Watson C.S., "Formant-frequency discrimination for isolated English vowels", Journal of the Acoustical Society of America. Vol 95,No. 1, pp 485-496, 1994.

D.H. Klatt, C. Aoki, "Syntesis by rule of Japanese", J.A.S.A., suppl. 1 76, 1984.

KLSYN88, manuale di riferimento, L. Godstein, S. Levy, 1987.

John Local and Adrian P. Simpson, "Phonetic implementation of geminates in Malayalam nouns", ICPhS99 San Francisco pp 592-595, 1999.

Naima Louali and Ian Maddieson, "Phonological contrast and phonetic realization: the case of Berber stops", ICPhS99 San Francisco pp 603-606, 1999.

Pierpaolo Luzzato Fegiz, "Appunti di Statistica Metodoligica", Kappa Librerie editrice, 1965-66.

F. Macrì, "Raddoppiamento nelle liquide [l], [r]: acustica e percezione", Tesi Univ. di Roma "La Sapienza", 1995.

Shinji Maeda, "Acoustics of vowel nasalization and articoulatory shifts in french nasal vowels", in "Phonetic and Phonology" Volune 5 "Nasals,Nasalization, and the Velum", ACADEMIC PRESS INC. 1993.

B. Malmberg, "Manuale di fonetica generale", Ed. Il Mulino, Bologna, 1977.

P. Mandarini, "Comunicazioni elettriche", Ed. Ingegneria 2000, 1990.

Angelo Marchese, "Pratiche comunicative", Principato Editore, 1979.

Z. Muljacic, "Fonologia della lingua italiana", Ed. Il Mulino, Bologna, 1972.

V. Oppenheim, R. W. Schafer, "Digital signal processing", Prentice Hall, 1975.

OROS AU21 CARD, manuali di riferimento, OROS, 1991.

Athanasios Papoulis, "Probabilità, variabili aleatorie e Processi stocastici", Boringhieri editore, 1973.

R. Rabiner, R. W. Schafer, "Digital processing of speech signals", Prentice Hall, 1978.

Rochet,L.B., and Rochet,A.P., "The perception of the single-geminate consonant contrast by native speakers of Italian and Angliphones" in proceedings of ICPhS95, edited by K.Elenius and P.Branderud,Vol 3 (Arne Strombergs, Stockholm) pp. 616-619, 1995.

R. Rossetti "Gemination of Italian stops", Journal of the Acoustical Society of America, 95, 2pSP25, pp.2874, 1994.

R. Rossetti, "Caratteristiche acustiche del fenomeno di geminazione nelle consonanti occlusive Italiane: applicazione all'adattamento automatico di pronunce straniere", Tesi Univ. di Roma "La Sapienza", 1993.

Saltarelli M., "A phonology of italian generative grammar", The Hague-Paris, 1970.

Shrotriya N., Siva Sarma A.S., Verma R., Agrawal S.S., "Acoustic and perceptual characteristics of geminate Hindi stop consonants", in Proceedings of ICPhS95, edited by K.Elenius and P.Branderud, 4, (Arne Strombergs Grafiska Stockolm), pp.132-135, 1995.

M. Spiegel, "Statistica", sec. Ed., McGraw-Hill, 1988.

Statgraphics Plus User Manual - Statistical graphics corp.1996.

Kenneth N. Stevens, Gunnar Fant, Sarah Hawkins, "Some acoustical and perceptual correlates of nasal vowels", 1987.

Kenneth N. Stevens, "Acoustic phonetics", 1998.

H. W.Strube, "Determination of the istant of glottal closure from the speech wav.", J.A.S.A., vol. 56, n. 5, November 1974.

M. Svirsky, K. N. Stevens, M. L. Matthies, J. Manzella, J. S. Perkell, R. Wilhems, "Tongue surface displacement during bilabial stops", Journal of the Acoustical Society of America, 102, pp. 562-571.

Turbo Pascal 6.0 Manuale di riferimento", Borland, 1992.

A. Vannucci, "Correlati acustici di tratti distintivi: applicazione alla caratterizzazione del punto di articolazione delle consonanti occlusive dell'italiano e loro riconoscimento automatico", Tesi Univ. di Roma "La Sapienza", 1993.

Vecsys, "The Unice User Manual", Vecsys - Chemin du Chene rond - 91570 Bièvres, France, 1989.

T.H. Wonnacott, R.J Wonnacott, "Introduzione alla statistica" Franco Angeli editore, 1972.