IV. L'analisi fonetica strumentale [a cura di A. Romano; tratto da: N. MINISSI, M. RIVOIRA, A. ROMANO, in prep., "Manuale di Fonetica" Alessandria: Dell'Orso]

Nonostante gli sviluppi e la diffusione di tecniche di rappresentazione alternative, rispetto a quelle descritte nella prima edizione di questo manuale, lo **spettrogramma** resta ancora oggi il principale riferimento documentario per l'interpretazione acustica e articolatoria del fatto fonetico (linguistico o no). Per quanto esso si possa ancora ottenere con uno dei modelli analogici di *Sonagraph* o con strumenti a controllo informatico, come il *CSL* (*Computerized Speech Lab*), oppure con diversi software, come *Multispeech*, commercializzati successivamente dalla Kay Elemetrics[1], la possibilità di ottenere una rappresentazione di tipo spettrografico per mezzo di elaborazioni numeriche su segnali digitali è implementata in numerosi altri prodotti informatici di diversa destinazione e di varia qualità.

Dove non diversamente specificato, i grafici che qui si presentano a titolo d'esempio sono ottenuti mediante PRAAT, un applicativo multi-piattaforma (PC, Unix, Macintosh) sviluppato (dal 1992), aggiornato costantemente e messo a disposizione gratuitamente da Paul Boersma e David Wenink del Laboratorio di Fonetica dell'Università di Amsterdam (l'ultimo aggiornamento disponibile è scaricabile dal sito www.praat.org).

IV.1. Spettrogrammi a banda stretta e a banda larga

In tutte le rappresentazioni, comunque, gli spettrogrammi (detti sonagrammi o sonogrammi in altre tradizioni) di una registrazione sonora di parlato conservano le stesse caratteristiche grafiche, presentando la distribuzione dell'energia spettrale, alle varie frequenze (sull'asse delle ordinate, in cicli al secondo o *Hertz*, Hz) e nel corso del tempo (sull'asse delle ascisse, in millesimi di secondo o *millisecondi*, ms), con valori locali rappresentati da variazioni di livello cromatico (generalmente in scala di grigi).

Sebbene sia in genere possibile ottenere spettrogrammi del tipo di quelli un tempo definiti a *banda stretta* (v. nota 1), gli spettrogrammi più diffusi sono oggi quelli di tipo a *banda larga*, che permettono una più agevole lettura dei valori formantici, relativi cioè all'evoluzione temporale delle formanti (v. Fig. 1).

[1] Questi prodotti rappresentano la realizzazione commerciale del procedimento spettrografico *Visible Speech* messo a punto presso i laboratori Bell Telephone. Una prima presentazione del procedimento, illustrato esaurientemente da Potter, Kopp & Green (1947) e dalle altre fonti fondamentali citate nella prima parte, è in un numero di *Science* del 1945 e viene poi descritto più dettagliatamente, nel 1946, in una monografia della serie "Bell Telephone System Monograph" e nel numero 17 della rivista *JASA* (*Journal of the Acoustic Society of America*). Il Sonagraph registrava originariamente una produzione sonora (segnale) di durata massima di 2,4 s. Le successive riproduzioni di questa registrazione, per circa 5 minuti (a una velocità 3,33 volte quella di registrazione), analizzate da un banco di filtri analogici, permettevano di ottenere una rappresentazione grafica delle caratteristiche acustiche del prodotto di ciascun filtraggio. A ogni ripetizione, le caratteristiche energetiche di una porzione dello spettro del segnale, evidenziata da un filtro passabanda, restavano impresse su un foglio di carta termo-sensibile (avvolta su un tamburo rotante in sincronia con il disco magnetico contenente il segnale) per mezzo di un ago termico che, alla fine del ciclo, scatta di un gradino verso l'alto (con un'escursione verticale di 4 pollici, cioè 101, 5 mm, che andava successivamente graduata secondo un'opportuna scala lineare, ad es. 6 o 8 kHz). La distribuzione dell'energia spettrale alle varie bande (sull'asse delle ordinate) era così tracciata progressivamente (col tempo graduato in ms sull'asse delle ascisse) in base alla variazione dell'oscuramento subito dalla carta, il grado di annerimento essendo proporzionale all'intensità sonora. Il filtro poteva avere (a scelta dello sperimentatore) una larghezza di banda di 300 (*banda larga*) o di 45 (*banda stretta*) Hz: le componenti che mostravano il maggiore annerimento (e quindi l'evoluzione temporale dei massimi locali di energia) potendo essere rispettivamente le formanti (a banda larga) o le armoniche (a banda stretta).

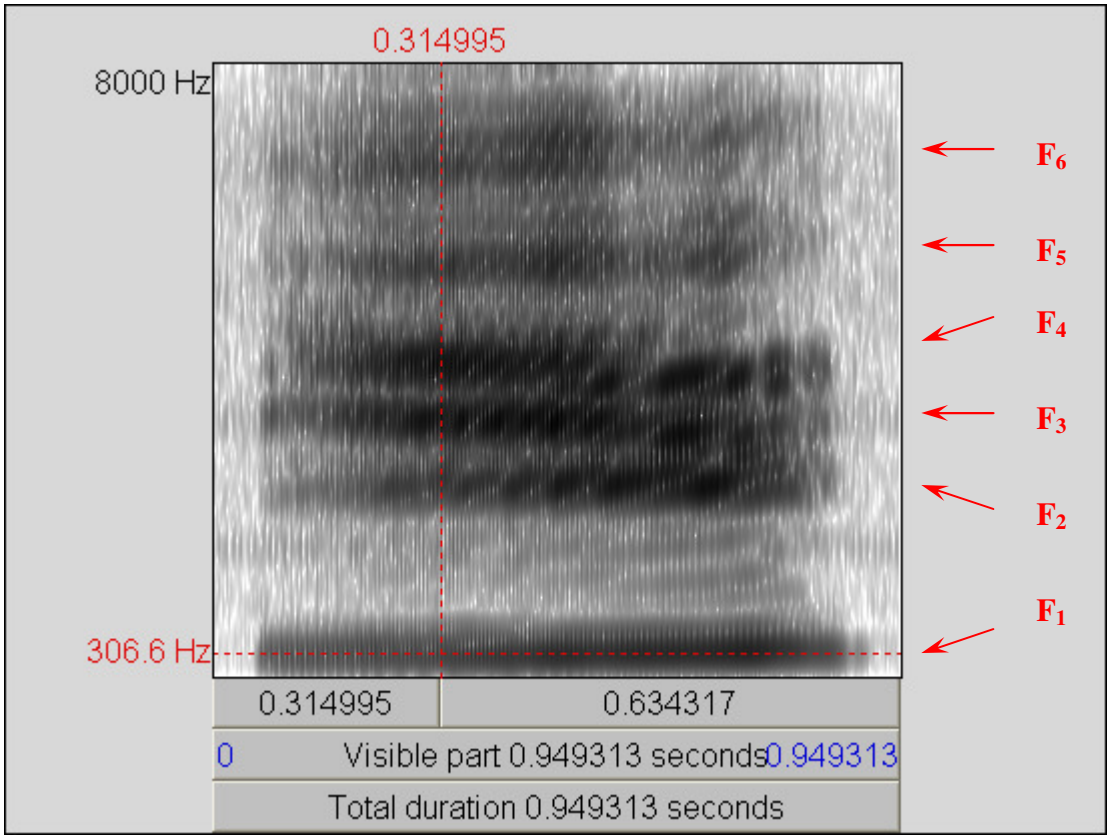

Figura 1. Esempio di spettrogramma del tipo a banda larga per la sillaba [ji:˥] pronunciata da un locutore di cinese mandarino in realizzazione della parola 遗 *yi*[2] 'perdere'. Oltre a un generale progressivo aumento di energia nel graduale ma rapido passaggio da [j] a [i:] (un confine convenzionale potrebbe essere posto in corrispondenza del demarcatore verticale qui indicato posizionando il cursore, nella finestra interattiva di PRAAT, a circa 315 ms dall'inizio), si distinguono chiaramente le prime sei formanti (in questo caso praticamente orizzontali, data la particolare stabilità timbrica del suono vocalico); la prima formante è quella qui sovrastata dalla linea tratteggiata orizzontale posizionata in questo caso a circa 300 Hz).

Le **formanti** (cioè le maggiori componenti del timbro di un suono, indicate con F_1, F_2 etc.) dipendono quasi esclusivamente dalla conformazione istantanea delle cavità epilaringee e quindi sono notevolmente indipendenti dalle **armoniche**, le quali dipendono dalle caratteristiche del suono prodotto dal movimento delle pliche vocali. La differenza tra formanti e armoniche si può apprezzare negli spettrogrammi effettuati a banda stretta nei quali le formanti appaiono talvolta segmentate in tante striature parallele (v. Fig. 2) la cui evoluzione temporale può essere contenuta, in condizioni di minor variazione della velocità di vibrazione delle pliche vocali (misurata dall'andamento della prima armonica f_1 che, coincidendo con l'armonica fondamentale, s'indica anche con f_0 ed è di solito detta **frequenza fondamentale**)[2] o, al contrario, molto più dinamica in presenza di rapide variazioni di questa[3]. Essendo l'altezza delle armoniche determinata da multipli di f_0, anche la distanza che separa un'armonica dalla successiva corrisponde alla frequenza fondamentale che, in tal modo, può essere facilmente misurata dividendo il valore di frequenza di un'armonica per il numero che la individua convenzionalmente in successione (e tenendo conto che la seconda armonica è già f_2; ad es. si divide per 9 il valore della nona armonica, che è appunto 9 volte f_0).

[2] Si noti che in alcuni casi (in voci gravi o con forte componente laringale), la presenza di più armoniche con energia considerevole in bassa frequenza definisce una F_0 (v. esempio in Fig. 2).

[3] Si noti ancora che la distanza tra le striature verticali corrisponde al **periodo fondamentale**, T_0, corrispondente alla durata di un ciclo di vibrazione delle pliche vocali (periodo della glottide) e pari al reciproco di f_0 ($T_0 = 1 / f_0$). In molti casi, potendo stimare T_0, è facilmente deducibile f_0 (= $1 / T_0$). Ad es. misurando un valore di T_0 pari a circa 10 ms, si ha una f_0 di circa 100 Hz.

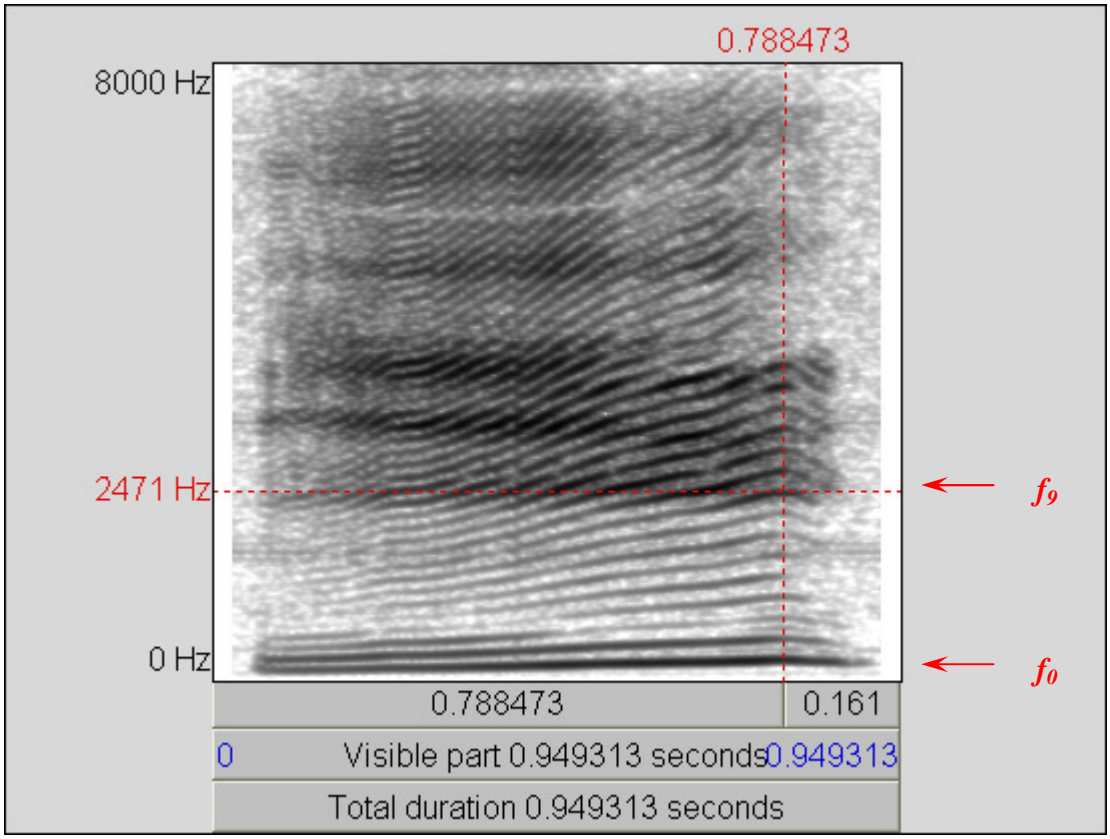

Figura 2. Esempio di spettrogramma del tipo a banda stretta per la stessa produzione analizzata in Fig. 1. In questo caso, per quanto ancora parzialmente visibili le sei formanti evidenziate in Fig. 1, le striature globalmente ascendenti rendono evidenti le variazioni imposte alle armoniche da un tono ascendente. Pur essendo visibile (e misurabile, posizionando su di essa il cursore nella finestra interattiva di PRAAT) la frequenza fondamentale, f_0, una misura di questa può essere ottenuta leggendo il valore indicato a sinistra del grafico in corrispondenza della linea orizzontale definita dalla posizione del cursore che in questo caso individua ad esempio la nona armonica, f_9 = 2471 Hz (a 788 ms dall'inizio) e dividendolo per 9 (circa 274 Hz; cfr. Fig. 4).

IV.2. Altre curve

In molti di questi, alle possibilità di rappresentazione spettrografica sono aggiunte quelle di rappresentazione congiunta con altre curve, come l'**oscillogramma** (andamento dell'ampiezza istantanea della forma d'onda o *amplitude waveform*, con scala su un numero di livelli convenzionali, v. dopo), di solito affiancato verticalmente in alto (v. Fig. 3) oppure del **profilo melodico** (andamento della frequenza fondamentale, curva di f_0 o *pitch*, con scala in Hz), di solito sovrapposto (v. Fig. 4) o dell'**intensità** sonora (curva dell'energia, cioè del volume istantaneo, *loudness* o *intensity* in PRAAT, con scala virtuale in *decibel*, dB, basata sui livelli numerici dell'ampiezza, v. nota 3), anche questo di solito sovrapposto (v. Fig. 5).

In seguito alla diffusione delle tecniche *LPC* (v. §5), si è diffusa anche la possibilità di ricorrere a una stima robusta delle variazioni formantiche nel corso di una stessa produzione. Il grafico che le rappresenta si chiama **tracciato formantico** (*formant track*) ed è di solito anch'esso sovrapposto allo spettrogramma (v. Fig. 6).

Per una migliore analisi della composizione spettrale di un suono, inoltre, è in genere possibile ricorrere a un altro tipo di rappresentazione: la **sezione spettrale** (generalmente indicata come **spettro** d'ampiezza, o *spectral slice* in PRAAT). La risoluzione dell'analisi spettrale dipende però dalla lunghezza della finestra selezionata (v. Fig. 7) e permette in genere d'isolare armoniche e formanti dei suoni periodici, ma anche di stimare la pendenza (*spectral tilt*), la densità o il centro di gravità che caratterizzano la composizione spettrale di suoni non periodici (rumori di frizione, v. §10).

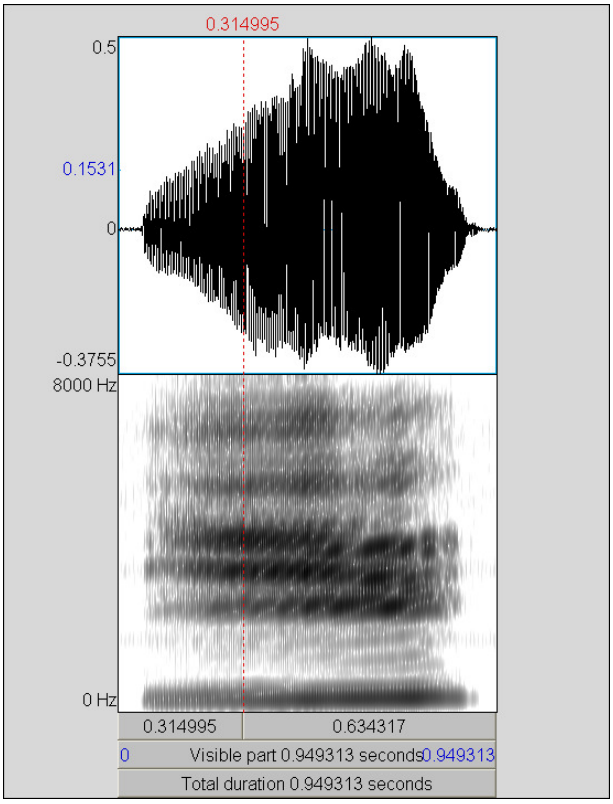

Figura 3. Esempio di associazione tra oscillogramma (sopra) e spettrogramma (sotto) per la stessa produzione analizzata in Fig. 1.

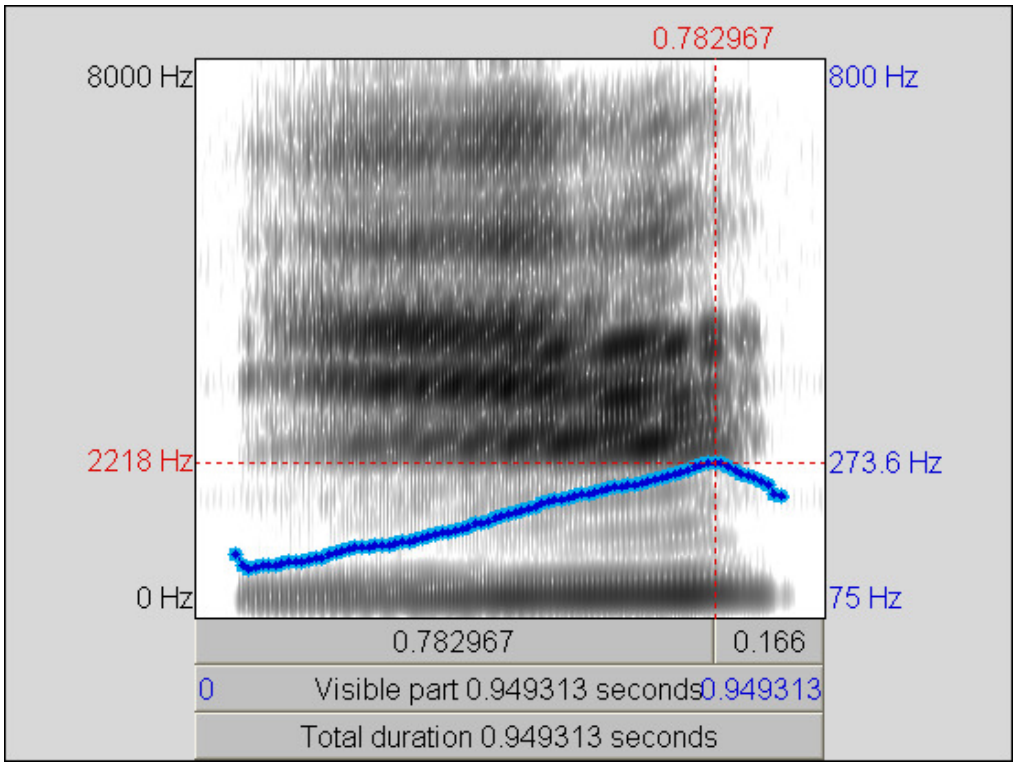

Figura 4. Esempio di profilo melodico sovrapposto allo spettrogramma della stessa produzione analizzata in Fig. 1. Il picco massimo segnalato a circa 274 Hz corrisponde al valore calcolato in Fig. 2 in base alla misura della frequenza della nona armonica (si noti che il minimo iniziale di f_0 è pari in quest'esempio a 137 Hz, esattamente coincidente con la metà del valore massimo: il locutore ha realizzato quindi in questa sillaba un *glissando* ascendente di un'ottava).

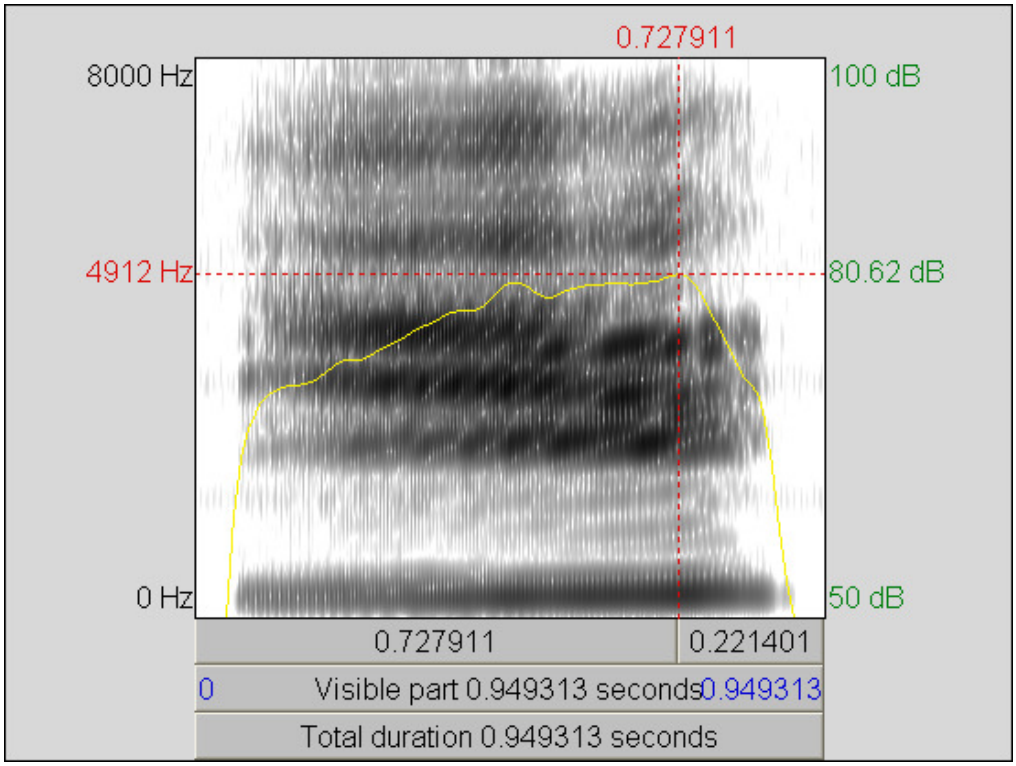

Figura 5. Esempio di curva dell'intensità sovrapposta allo spettrogramma della stessa produzione analizzata in Fig. 1. Il picco massimo segnalato a circa 80 dB indica un valore di riferimento virtuale (se non si esegue un'adeguata *calibrazione*, non si tratta di dB_{SPL} ovviamente)[4].

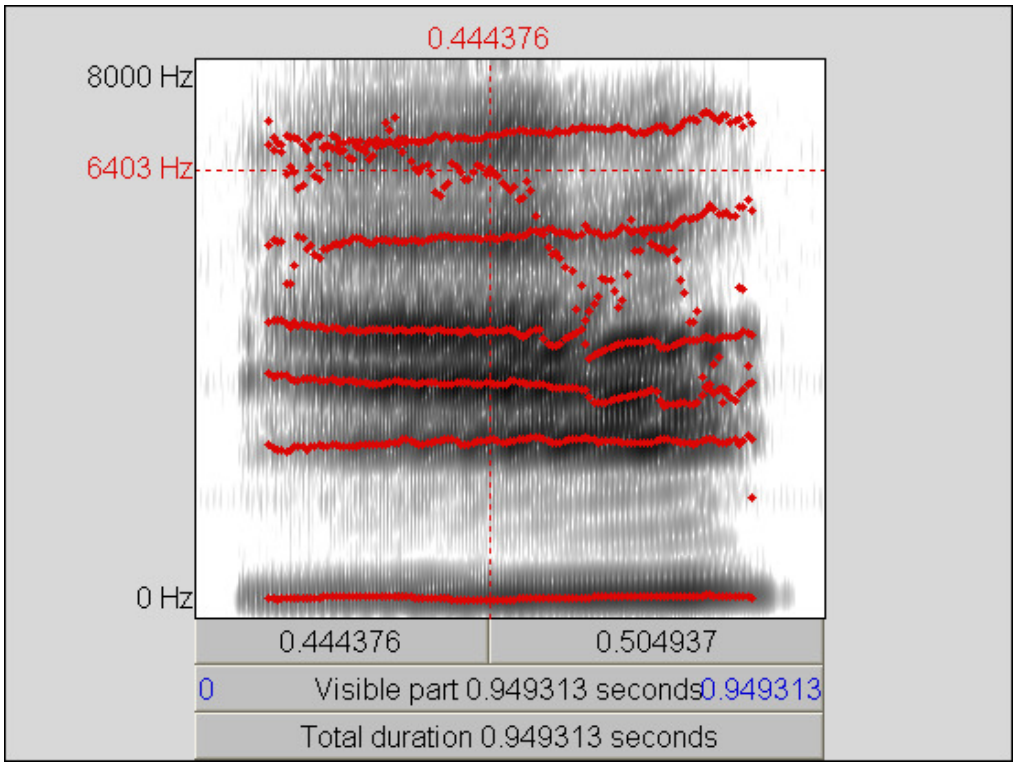

Figura 6. Esempio di tracciato formantico sovrapposto allo spettrogramma della stessa produzione analizzata in Fig. 1. Si evidenziano in tal modo le sei formanti già individuate (il cursore indica però una formante stimata erroneamente).

[4] I dB_{SPL} (*SPL = Sound Pressure Level*) sono una misura d'intensità basata su una media dell'accumulo di pressione sonora nel tempo in riferimento alla minima pressione udibile dall'orecchio umano (0,00002 Pascal (Pa)). Il locutore ha prodotto la parola scandendola con un assetto energetico di conversazione ordinaria, quindi intorno ai 50 dB_{SPL}; 80 dB_{SPL} corrispondo solitamente a rumori molto intensi e duraturi come quello di un martello pneumatico.

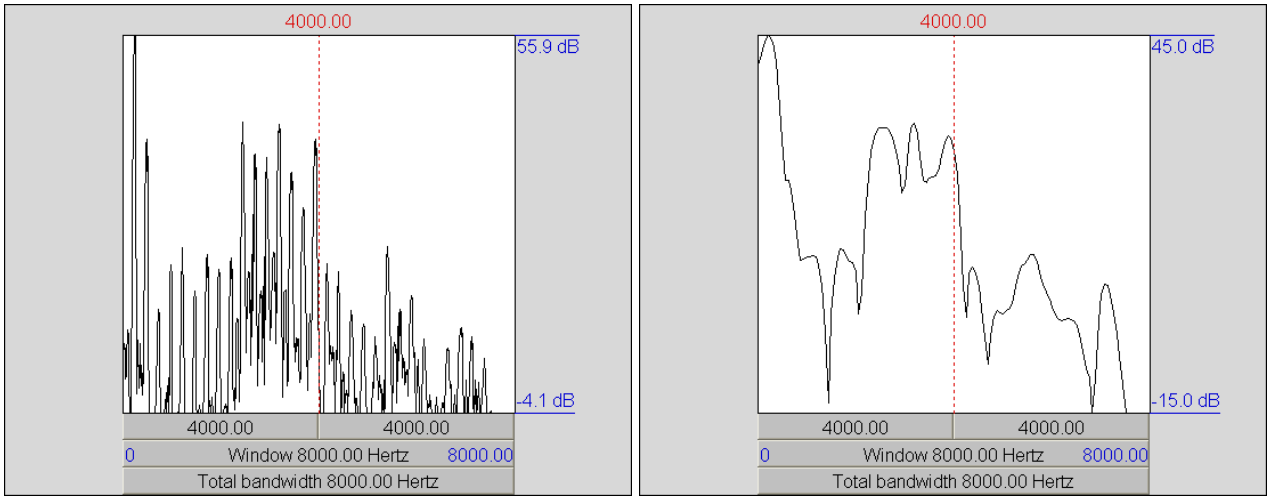

Figura 7. Esempi di sezioni spettrali per il suono vocalico [i] della produzione analizzata in Fig. 1. Lo spettro di sinistra rappresenta il risultato di un'analisi condotta in una finestra di circa 100 ms e si presenta dettagliato al punto da rendere riconoscibili quasi tutte le armoniche: le formanti emergono grossolanamente come raggruppamenti di armoniche. Lo spettro di destra è invece ottenuto a partire da una finestra di segnale di 10 ms e si presenta quindi decisamente più approssimativo, ma permette d'individuare ancora le principali formanti.

IV.3. I segnali digitali

Rispetto alle tradizionali analisi analogiche, l'analisi di segnali digitali ha introdotto diversi nuovi parametri descrittivi delle loro caratteristiche e determinanti per una loro corretta rappresentazione.

I segnali digitali sono infatti discreti nel tempo e si ottengono effettuando una *discretizzazione* cioè un *campionamento* di un segnale in origine continuo (analogico, appunto).

Un **campionamento** a 16 kHz, ad esempio, fa sì che vengano prelevati 16 campioni ogni millisecondo (*ms*) di segnale analogico (un parametro importante è appunto la *frequenza di campionamento*, in quest'esempio F_c=16 kHz).

Il segnale così definito è discreto nel tempo (una discretizzazione viene introdotta anche in ampiezza mediante la cosiddetta *quantizzazione*) e presenta una certa perdita di qualità rispetto al segnale analogico (questo deterioramento si può evidentemente contenere, aumentando la F_c). Anche se si tratta di un fatto abbastanza intuitivo, esiste un teorema (il teorema di Nyquist) che stabilisce che, campionando a una data F_c, si perde la possibilità di descrivere tutte quelle oscillazioni con frequenza superiore a $F_c/2$. Questo è il motivo per cui spettri e spettrogrammi di segnali acquisiti con F_c=16 kHz sono rappresentati su una scala di frequenza fino a 8 kHz.

Riguardo alla **quantizzazione** dell'ampiezza, diciamo solo che questa contribuisce a far rappresentare la variazione continua del segnale come se si trattasse di una curva a gradini, con salti da uno all'altro di *n* valori prestabiliti[5]. Una quantizzazione a 16 *bit* nella digitalizzazione (N=16) determina una rappresentazione dell'ampiezza istantanea del segnale su 65536 livelli ($n=2^N$) generalmente suddivisi tra valori negativi (da –32768 a –1), 0 e i valori positivi (da +1 a +32767) [6].

Alla base dell'**analisi spettrografica digitale** è la cosiddetta ***FFT*** (*Fast Fourier Transform*), un'evoluzione della ***DFT*** (*Discrete Fourier Transform*), disponibile ormai in tutti i prodotti informatici che permettono di eseguire un'analisi acustica elementare sottoforma di rappresentazione in frequenza (basata sulla trasformata di Fourier, appunto).

[5] La convenzionalità di questi gradini rende impossibile risalire ai valori assoluti che in un segnale analogico erano invece associati a variazioni di variabili elettriche, come ad es. la tensione istantanea (misurata in μV), con valori dipendenti dalle modalità di trasduzione della pressione acustica captata dalla membrana del microfono.

[6] Fino a qualche anno fa, molti prodotti commerciali usavano *per default* una quantizzazione a 16 *bit*. Oggi sono invece disponibili dispositivi di acquisizione con quantizzazione a 20, 32 o 48 *bit*.

Queste analisi sfruttano un algoritmo, basato sul teorema di Fourier, che estrae in modo rapido (ma approssimativo) una stima dell'ampiezza corrispondente a ciascuna componente armonica del suono analizzato. L'analisi di Fourier necessita però di un suono sufficientemente lungo e stabile – tanto da poter essere considerato un suono periodico di durata teoricamente infinita – per poterne effettuare la decomposizione in onde sinusoidali elementari. L'algoritmo *FFT*, applicato su finestre di durata limitata in cui il suono presenta caratteristiche più o meno costanti e pseudo-periodiche, produce quindi necessariamente una stima approssimativa della reale composizione armonica del suono[7]. Sono spettri *FFT*, ad esempio, quelli proposti in Fig. 7 (ma sono basati su analisi *FFT* anche gli spettrogrammi delle figure precedenti).

IV.4. L'analisi predittiva lineare LPC

Esiste inoltre anche un altro approccio che consente anch'esso di determinare le proprietà timbriche dei suoni sulla base di stime numeriche.

Quest'approccio, affermatosi nell'analisi dei segnali digitali (cfr. Markel & Gray 1976), è noto come ***LPC*** (*Linear Predictive Coding*) ed è stato messo a punto nell'ambito di tentativi miranti a realizzare una codifica del segnale vocale per applicazioni nel campo delle telecomunicazioni.

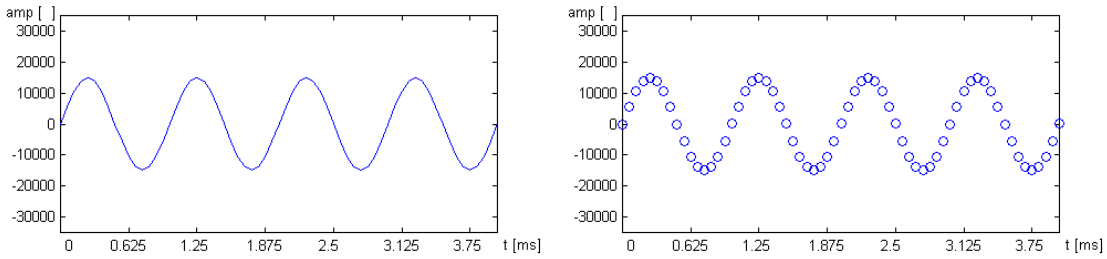

Figura 8. A sinistra, segnale analogico (quattro cicli di un tono puro a frequenza 1000 Hz e ampiezza convenzionale). A destra, stesso suono campionato a 16 kHz (vengono prelevati 16 campioni ogni ms di segnale; in questo caso, essendo raffigurati 4 ms, corrispondenti a 4 cicli, sono presenti 64 campioni) [grafici ottenuti con opportuni *script* di Matlab©].

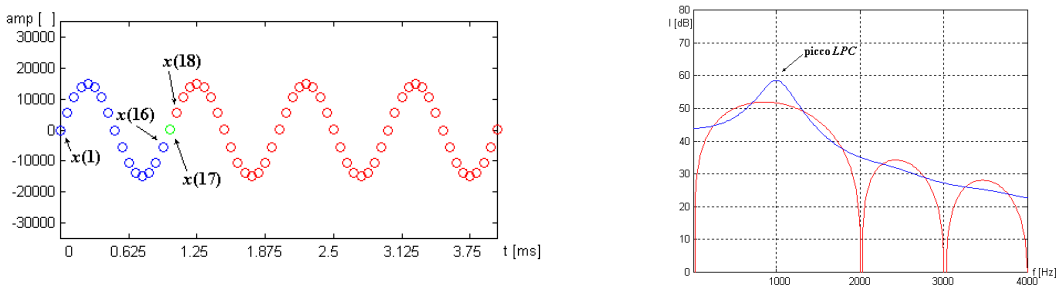

Figura 9. A sinistra, campioni del tono puro di fig. 8 coinvolti nella predizione *LPC* del 17° campione (evidenziati in blu e verde). A destra, profilo del(la funzione di trasferimento) del filtro *LPC* e curva dello spettro *FFT* ottenuti sullo stesso numero di punti (i 16 campioni usati per la predizione). Si noti come in questo caso la stima *LPC* (il picco) dia migliori risultati, determinando un profilo che individua la principale componente frequenziale [grafici ottenuti con Matlab©].

[7] La stima può avvenire su finestre di segnale più o meno ampie con effetti simili a quelli della tradizionale distinzione tra un'analisi a *banda larga* (condotta a partire da filtri che operano una scansione dello spettro, isolando di volta in volta bande di 300 Hz) e un'analisi a *banda stretta* (condotta a partire da filtri che operano per bande di 45 Hz). Con le tecniche odierne, le analisi *FFT* che meglio approssimano quella tradizionale a *banda larga* possono essere quelle condotte su finestre di meno di 8 ms (5 ms nell'esempio di Fig. 1), mentre per ottenere un'analisi simile a quella a *banda stretta* occorrono finestre di almeno 30 ms (come nell'esempio di Fig. 2). Per segnali vocali digitali con F_c=16 kHz questi limiti si traducono in analisi *FFT* rispettivamente su 128 e su 512 campioni (punti). L'analisi spettrografica proposta da PRAAT è basata *per default* su finestre di 5 ms (è quindi un'analisi a *banda larga* utile per la lettura dei movimenti formantici) ma è possibile modificare questo parametro nella finestra interattiva "Spectrogram settings" disponibile nella voce di menu "Spectrum".

Il principio è semplice: in una porzione di segnale digitale con caratteristiche periodiche costanti (come quello nel grafico a destra della figura 8), in una sequenza temporale di campioni prelevati sulla forma d'onda, ogni campione può essere descritto in base a una valutazione ponderata dell'andamento descritto dai campioni precedenti (il suo valore può essere cioè "previsto" in base alle caratteristiche di un certo numero di valori degli ultimi campioni osservati). I coefficienti attribuiti a ognuno di questi campioni definiscono una funzione di trasferimento (in uno spazio numerico opportuno) che segue in buona misura l'inviluppo dello spettro di una trasformata di Fourier (v. figura 2)[8].

In figura 9 sono evidenziati i primi 17 campioni di 4 cicli di un tono puro a 1000 Hz (un campione dato, il 17°, e i 16 campioni precedenti). La stima del valore di ampiezza del 17° campione ad esempio, che assumiamo ignota, può essere ottenuta mediante un algoritmo che "valuta" l'andamento dei 16 campioni precedenti (*LPC* di ordine 16, detto anche a 16 poli).

Definendo $x(n)$ questo segnale (con n variabile temporale), le ampiezze dei primi campioni nell'esempio prescelto sono rispettivamente:

$x(1) = 0$
$x(2) = 5740$
...
$x(15) = -10607$
$x(16) = -5740$

In generale, sulla base degli N campioni precedenti, si può rappresentare il valore predetto del generico campione n come:

$$\bar{x}(n) = a_N * x(n-1) + a_{N-1} * x(n-2) + ... + a_2 * x(n-N+1) + a_1 * x(n-N)$$

Un'analisi *LPC* di ordine 16, effettuata con un opportuno algoritmo sui primi 16 campioni, produce come risultato i seguenti coefficienti:

$a_{16} = -1.7522$
$a_{15} = 0.8984$
$a_{14} = 0$
... (coefficienti nulli per tutti gli altri campioni)
$a_3 = 0$
$a_2 = 0.0922$
$a_1 = -0.0688$

Con questi valori, diviene quindi possibile predire un campione, conoscendone i 16 precedenti. Ad es.:

$$\bar{x}(17) = a_{16} * x(16) + a_{15} * x(15) + ... + a_2 * x(2) + a_1 * x(1)$$
$$= 1.7522*5740 - 0.8984*10607 + ... - 0.0922*5740 - 0.0688*0 = -0.9288 \text{ (circa 0)}$$

e infatti, data la periodicità, sappiamo essere $x(17) = 0$, oppure:

$$\bar{x}(18) = a_{16} * x(17) + a_{15} * x(16) + ... + a_2 * x(3) + a_1 * x(2)$$
$$= -1.7522*0 + 0.8984*5740 + ... + 0.0922*10607 - 0.0688*5740 = 5739.9$$

e infatti sappiamo essere $x(18) = 5740$.

[8] Il dosaggio in ampiezza dei campioni permette a questi coefficienti di definire infatti una misura delle diverse periodicità presenti nel segnale.

La stima della periodicità del segnale studiato, permette di definire – mediante opportuni calcoli – un **profilo *LPC*** come quello riportato nel grafico a destra della stessa figura 2 (che individua chiaramente, nonostante il basso numero di punti della valutazione, la frequenza del tono: nel grafico il picco è, infatti, a 1000 Hz).

Come si diceva, quindi, usando questi coefficienti è possibile definire un filtro le cui proprietà di selezione (funzione di trasferimento) possono essere riprodotte mediante una curva il cui profilo riproduce all'incirca l'inviluppo spettrale del suono filtrato (v. §5). Il profilo della funzione di trasferimento *LPC*, calcolato per un qualsiasi segnale (pseudo-)periodico, si presenta, infatti, di solito, come una curva con un certo numero di rilievi (picchi) situati in prossimità delle formanti del suono campionato[9].

Un'applicazione dell'analisi *LPC* (v. Fig. 10) alle stesse porzioni sonore in cui sono calcolati gli spettri di Fig. 7, permette di apprezzare la relativa stabilità delle stime formantiche basate su queste curve rispetto a quelle che si potrebbero fare su uno spettro. La maggior facilità con cui è possibile individuare in modo procedurale i massimi di curve come queste è anche all'origine della diffusione di grafici come quello del tracciato formantico (v. Fig. 6).

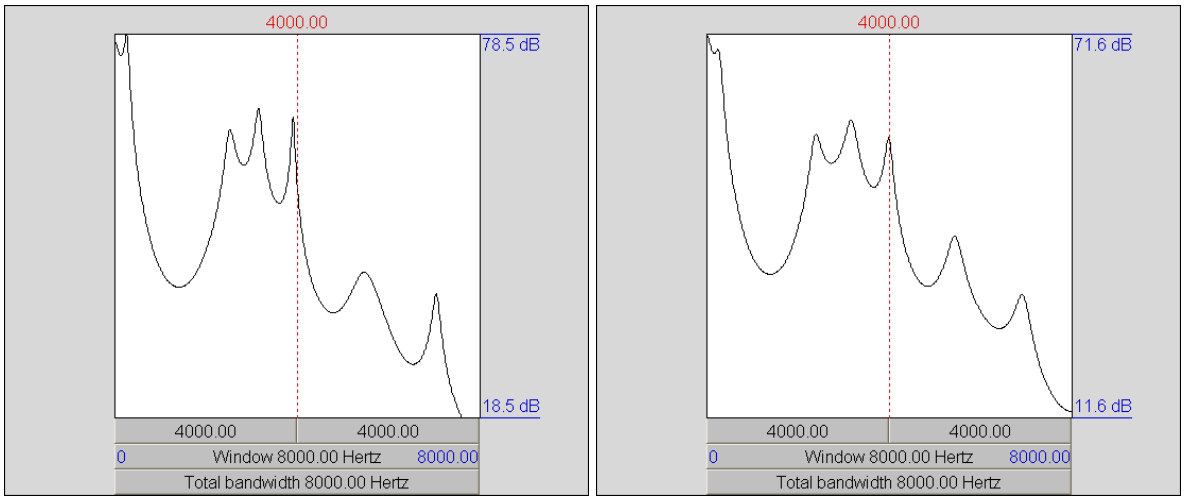

Figura 10. Esempi di analisi *LPC* del suono vocalico [i] della produzione analizzata in Fig. 1. A sinistra un profilo ottenuto con un'analisi condotta in una finestra di segnale di circa 100 ms e a destra uno ottenuto a partire da una finestra di 10 ms: le formanti emergono in entrambi i casi con definizione simile.

Il ricorso alla predizione lineare, e alle tecniche di sintesi vocale basate su questa, ha rafforzato la visione del *continuum* sonoro come risultato dell'interazione tra un **filtro** (associato alla funzione selettiva delle frequenze dipendente dalla forma del condotto vocale), la cui funzione di trasferimento risulta modellizzabile per mezzo di un'adeguata analisi *LPC*, e una **sorgente** impulsiva o rumorosa (effettivamente presenti nel caso di suoni sonori e/o fricativi, rispettivamente come vibrazioni glottidali o come turbolenze originate nel tratto vocale)[10].

Inoltre, come si vedrà nel paragrafo seguente, ha permesso di ottenere stime più affidabili dei timbri nel caso di voci femminili (per le quali l'analisi di Fourier si presenta di solito più difficoltosa).

---

[9] Solitamente, occorre almeno un ordine *N* (*prediction order* in PRAAT, si dice allora *LPC* con ordine di predizione *N* oppure a *N* poli) per ottenere *N*/2 picchi nel profilo *LPC* (sono infatti necessari due poli per ciascuno di questi). Nelle regioni di pseudo-periodicità di un segnale vocale si trova una formante circa ogni kHz. Per un segnale campionato a 16 kHz, il cui spettro si estende fino a 8 kHz, possono essere quindi visibili circa 8 formanti. L'ordine *LPC* più indicato sarebbe quindi *N*=16, adatto ad approssimare meglio uno spettro con 8 formanti. Tenendo conto però della frequente presenza di un'importante concentrazione energetica anche in prossimità di F_0, si consiglia talvolta di tenere il numero di poli dell'analisi *LPC* pari almeno a F_c/1000+2 (in presenza di nasalità, si rivela utile a volte aumentare ancora l'ordine dell'analisi, scegliendo ad esempio *N*=20, per F_c=16 kHz).

[10] Si tratta di un modello (v. §5) proposto da Fant (1960) e usato nei primi sistemi di sintesi vocale di tipo *Vocoder*. Ancora recentemente il modello è stato usato da Klatt & Klatt (1990).

IV.5. La teoria Sorgente-Filtro

Nell'ambito di questa teoria, la cui prima efficace formalizzazione risale a Fant (1960), si fa l'ipotesi che, nella produzione linguistica dei suoni, il contributo dato dalle pliche vocali in vibrazione (all'interno della laringe) sia separabile, in termini di generazione fisiologica e di proprietà acustiche di un suono originario impulsivo periodico, dal contributo dato dalle cavità superiori (faringe, cavità orale e cavità nasali), caratterizzate da effetti successivi di filtraggio e amplificazione delle caratteristiche acustiche di questo suono originario.

Alla **laringe**, organo in cui si genera la voce (fonazione), corrisponderebbe quindi il ruolo di generatore di un'eccitazione variabile (***sorgente***), costituita da una **vibrazione periodica**, delle **cavità superiori**, modellizzabili come casse di risonanza a geometria variabile, che – proprio grazie a questa variabilità – renderebbero possibile un'azione selettiva dinamica (***filtro***) tale da consentire una **modifica timbrica della voce** e quindi l'articolazione di suoni qualitativamente distinti.

Le caratteristiche dell'eccitazione sono descritte come quelle di un treno d'impulsi la cui forma d'onda elementare può essere osservata, nella descrizione del meccanismo di vibrazione delle corde vocali, mediante l'ausilio di rappresentazioni glottografiche (ottenute con speciali apparecchi).

Sul piano dell'analisi armonica (decomposizione di Fourier), le caratteristiche spettrali di questo suono periodico sono naturalmente quelle di uno spettro armonico (a righe) con energia decrescente dalla prima armonica alle armoniche più alte. La pendenza dell'inviluppo di queste righe viene idealmente descritta come un andamento decrescente a –12 dB/ottava[11].

Le caratteristiche di trasmissione del filtro sono descritte solitamente da un grafico continuo (il profilo della funzione di trasferimento) che rappresenta le regioni di frequenza in cui le cavità epilaringee operano una selezione delle componenti della sorgente acustica, che vengono filtrate in base alle frequenze di risonanza di queste cavità (notare che questa curva è mediamente piatta, con dei rilievi che rinforzano alcune bande di frequenza e delle depressioni che ne attenuano altre).

Dalla sovrapposizione delle caratteristiche energetiche dei contributi di sorgente e filtro si ottiene la composizione frequenziale del suono in uscita, irradiato in corrispondenza delle labbra (uno spettro con armoniche ponderate). La figura 11 presenta tre grafici che permettono di apprezzare i principi di base di questo modello e a sua utilità finale. Dalla semplice osservazione dello spettro di un suono (che si presenta di solito come quello nel grafico più a destra), visto come risultato dell'interazione tra quello della sorgente glottidale (grafico più a sinistra) e le caratteristiche filtranti del tratto vocale, configurato per determinare il suo timbro particolare (grafico al centro), è infatti possibile ottenere le principali informazioni sulle caratteristiche di fonazione e articolazione del locutore al momento della sua produzione. Nell'esempio in figura si osservano delle righe equispaziate a una distanza frequenziale che determina la frequenza fondamentale del suono (anche visibile dalla posizione della prima armonica) e quindi la frequenza laringea con cui è stato ipoteticamente prodotto il suono (f_0 = 100 Hz).

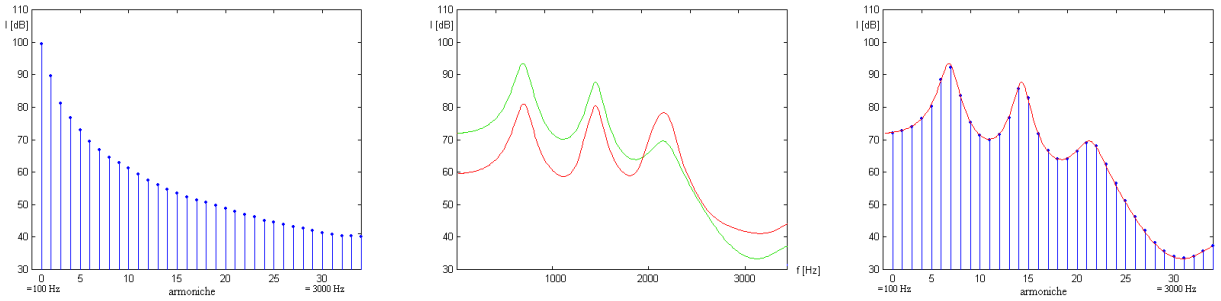

Figura 11. Da sinistra a destra: (1) rappresentazione armonica di un segnale di sorgente ideale (spettro glottidale); (2) funzione di trasferimento delle cavità epilaringee (linea scura - la linea più chiara rappresenta invece la stessa curva, tenendo conto però della pendenza complessiva introdotta dal segnale glottidale e dalla caratteristica di irraggiamento); (3) spettro del segnale risultante immesso nell'ambiente di propagazione (canale) [grafici ottenuti con Matlab©].

[11] Un'*ottava* è un intervallo di frequenze il cui estremo superiore è pari al doppio di quello inferiore. È un'ottava, ad es., l'intervallo tra 100 e 200 Hz. Per fare un altro esempio, è un'ottava anche la distanza frequenziale tra 160 e 320 Hz.

Seguendo inoltre il profilo che descrivono i picchi delle armoniche, si può ricostruire la curva caratterizzante del filtro, individuandone regioni di attenuazione armonica in contrapposizione a regioni di rafforzamento (formanti e antiformanti). Negli esempi fittizi di figura 12, grazie alla presenza di un inviluppo di raccordo tra le sommità delle varie armoniche, si individuano facilmente i picchi delle tre formanti, F_1, F_2 e F_3, compresi tra 600 e 700 Hz, tra 1400 e 1500 Hz, e tra 2100 e 2200 Hz[12].

Rispetto a questi esempi di lettura agevolata (dalla presenza dell'inviluppo spettrale), le uniche differenze che presenterebbero i casi reali (idealizzati anche questi nei grafici della parte destra di Fig. 12) sarebbero legate all'assenza di qualsiasi indicazione sul profilo originario della funzione di trasferimento del tratto vocale (che sarebbe quindi deducibile ricostruendo un inviluppo immaginario) e una minore chiarezza nella separazione delle armoniche indotta dalle limitazioni degli strumenti che eseguono l'analisi di Fourier (v. anche Fig. 7).

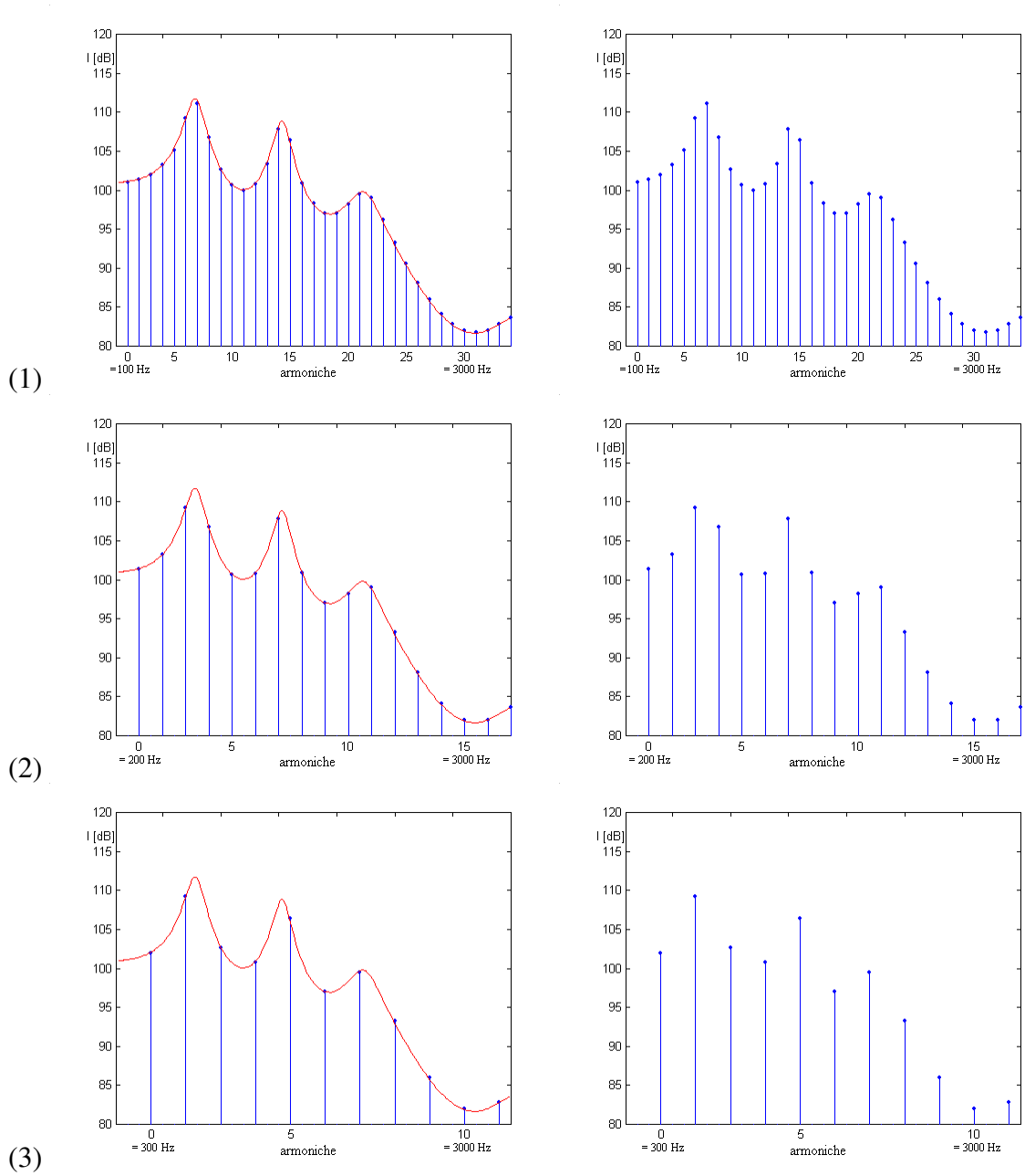

Figura 12. Spettri ideali con (a sinistra) e senza (a destra) l'indicazione del profilo della funzione di trasferimento del filtro vocale. Dall'alto in basso: (1) voce ideale a bassa f_0 (*M*); (2) voce ideale a media f_0 (*F*); (3) voce ideale ad alta f_0 (*B*) [grafici ottenuti con Matlab©].

[12] Più esattamente i picchi delle tre formanti si trovano rispettivamente a 680 Hz, 1430 Hz e 2120 Hz.

Questi grafici (Fig. 12) permettono altresì di valutare idealmente le conseguenze di una diversa frequenza di sorgente sulla stima delle regioni di rafforzamento armonico, le formanti appunto, individuali sullo spettro del suono.

Sappiamo infatti che passando da voci maschili (*M*) a voci femminili (*F*) e a voci infantili (*B*) la frequenza fondamentale media tende a crescere progressivamente.

Assumendo tre diversi valori di riferimento per f_0, 100, 200 e 300 Hz, e considerandoli arbitrariamente come rappresentativi di questi tre tipi di voce, vediamo cosa dobbiamo aspettarci di vedere cambiare sullo spettro di un suono. Assumiamo ancora, ipoteticamente che il tratto vocale stimolato da queste tre frequenze sia esattamente identico e assuma esattamente la stessa configurazione (e abbia quindi la stessa funzione di trasferimento, v. Fig. 12, grafici a sinistra). Gli spettri che idealmente possiamo aspettarci di ottenere dall'applicazione dell'analisi *FFT* sono quelli dei tre grafici a destra. È facile constatare il tipo d'errore cui saremmo indotti da questi grafici. Se cioè provassimo a misurare (a stimare) la posizione frequenziale delle tre formanti (F_1, F_2 e F_3) riconoscibili su questi spettri (che dai grafici in alto sappiamo essere esattamente le stesse), diremmo rispettivamente che, nel caso della prima 'voce' (*M*, con f_0 = 100 Hz), le formanti sono a 700 Hz, 1400 Hz e 2100 Hz; nel caso della seconda 'voce' (*F*, con f_0 = 200 Hz), sono a 600 Hz, 1400 Hz e 2100 Hz; nel caso della terza 'voce' infine (*B*, con f_0 = 300 Hz), sono a 600 Hz, 1500 Hz e 2100 Hz.

Sappiamo che tutti e tre gli spettri ci fanno commettere degli errori rispetto ai valori reali delle tre formanti. È però abbastanza evidente che commettiamo più errori, e di maggiore gravità, nel caso della terza voce. Infatti, se per le prime due voci (*M* e *F*) commettiamo lo stesso errore per F_2 e F_3 (sbagliando rispettivamente di 30 e 20 Hz), per la seconda voce (*F*) l'errore commesso su F_1 è di 80 Hz in difetto (rispetto ai 20 Hz in eccesso di *M*), mentre per la terza voce (*B*) che presenta lo stesso errore della seconda (*F*) su F_1, è presente anche un errore di 70 Hz su F_2[13].

Questa è la principale ragione per cui le misure *FFT* sono più affidabili quando eseguite su voci gravi; ed è questo il motivo per cui le voci maschili sono state le più studiate nei primi decenni di diffusione di questi strumenti. L'avvento dell'analisi *LPC* ha controbilanciato quest'interesse, rendendo affidabili le analisi formantiche anche per voci acute (femminili o infantili).

IV.6. Il timbro delle vocali

Come discusso nella parte I e nelle sezioni precedenti (nonché verificato negli esempi illustrati sopra), la disposizione delle formanti (soprattutto le prime) determina il timbro dei suoni vocalici[14].

Anche se è la struttura spettrale nella sua interezza che dà informazioni sull'esatta qualità vocalica, per semplicità si considerano solitamente le prime tre formanti (a volte, come in italiano, per via dell'estrema riduzione nel numero di elementi del sistema, si assume per comodità che bastino anche solo le prime due) riconducendo la valutazione sulle caratteristiche timbriche a un'operazione di misura di pochi parametri[15].

Al di là dei valori effettivamente misurabili di ciascuna di queste formanti dello spettro d'energia dei singoli suoni vocalici (v. esempi nelle Figg. 13-15), possiamo osservare come si dispone in generale l'energia di un vocoide in funzione del suo punto d'articolazione, del suo grado

[13] Inoltre, se nel caso di una serie di armoniche più fitte, seguendo l'inviluppo immaginario, con un po' d'esperienza possiamo dedurre una posizione più probabile delle formanti rispetto ai picchi d'armonica e scegliere di non ritenere come migliore stima della frequenza di formante la posizione di uno di questi, ma una posizione intermedia, questa strategia non è altrettanto applicabile nel caso di uno spettro con armoniche distanti.

[14] Le formanti più alte (F_4, F_5, F_6...) sono in genere più variabili da persona a persona e identificano maggiormente proprio le caratteristiche idiosincratiche del parlante.

[15] In realtà, oltre alla posizione media delle formanti, sarebbero rilevanti anche la larghezza di banda di ciascuna di queste e la presenza di contributi energetici spuri (attribuibili a nasalità, laringalità o altre proprietà dell'assetto fono-articolatorio generale). Di notevole importanza nella caratterizzazione fonetica è la presenza più o meno sistematica di relative instabilità articolatorie che si possono manifestare con derive timbriche.

d'apertura e di altri parametri come il grado di arrotondamento delle labbra, di retroflessione della lingua, d'innalzamento del velo o di arretramento della radice della lingua.

Per una voce maschile, l'energia di un suono di tipo [i] presenta formanti intorno a 300 Hz e nella banda 2500-3500 Hz ed è proprio quest'ultima concentrazione che differenzia fortemente il vocoide [i] dagli altri suoni vocalici (nell'esempio di Fig. 13 si ha all'incirca: F_1 = 260, F_2 = 2520 e F_3 = 3330 Hz).

Nel caso di un suono di tipo [a] le componenti energetiche più importanti si distribuiscono di solito in una banda che va da 800 a 2400 Hz (con particolare concentrazione nell'intervallo 800-1500 Hz), proprio laddove, al contrario, [i] e [u] presentano contributi poco significativi (nell'esempio di Fig. 14 si ha all'incirca: F_1 = 920, F_2 = 1360 e F_3 = 2650 Hz).

Buon parte dell'energia di un suono di tipo [u], infine, resta prevalentemente concentrata sotto i 1000 Hz (nell'esempio di Fig. 15 si ha all'incirca: F_1 = 360, F_2 = 860 e F_3 = 2680 Hz).

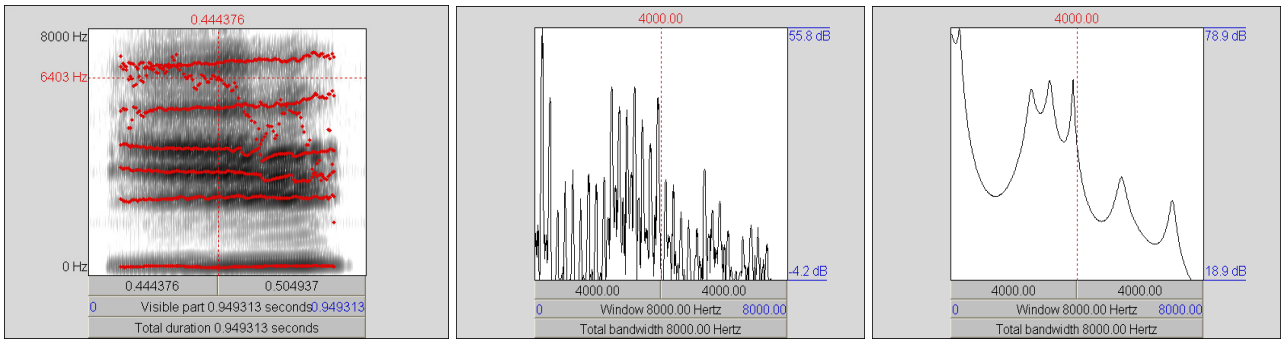

Figura 13. A sinistra, spettrogramma per la sillaba [ji:˩] pronunciata da un locutore di cinese mandarino in realizzazione della parola 遗 *yi*^{2} 'perdere' (v. Fig. 1). Al centro, sezione spettrale di [i:]. A destra, profilo *LPC* corrispondente.

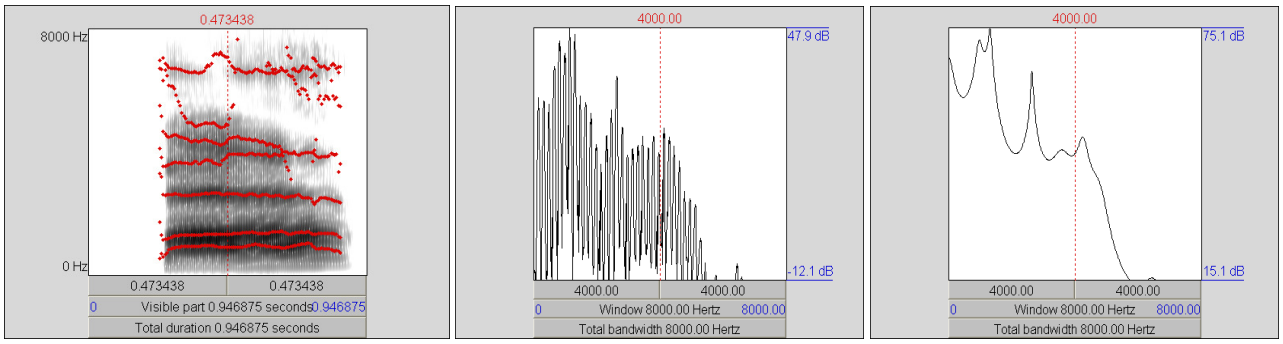

Figura 14. A sinistra, spettrogramma per la sillaba [b̥a:˩] pronunciata da un locutore di cinese mandarino in realizzazione della parola 拔 *ba*^{2} 'estrarre'. Al centro, sezione spettrale di [a:]. A destra, profilo *LPC* corrispondente.

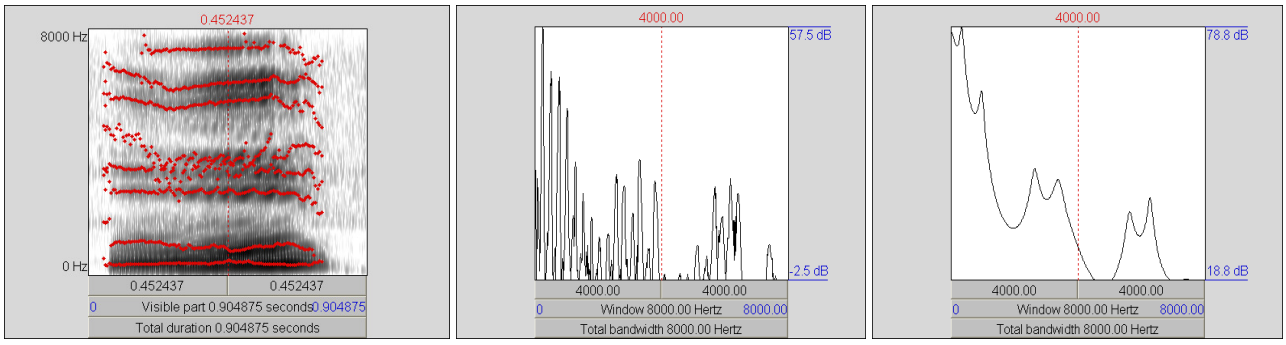

Figura 15. A sinistra, spettrogramma per la sillaba [wu:˩] pronunciata da un locutore di cinese mandarino in realizzazione della parola 无 *wu*^{2} 'senza'. Al centro, sezione spettrale di [u:]. A destra, profilo *LPC* corrispondente.

Con queste premesse, si può concludere sommariamente affermando che a una distinta disposizione degli organi articolatori corrisponda una diversa configurazione spettrale. A ogni diversa conformazione delle cavità epilaringee una distinta selezione viene operata sulle componenti spettrali, permettendo in un gran numero di casi di ritrovare indici acustici delle proprietà articolatorie di ciascun suono prodotto con un'articolazione stabile (mantenuta un certo tempo). Come osservato, infatti, in funzione della conformazione assunta dal tratto vocale, si determina un effetto filtrante che porta alla selezione di bande in cui le componenti vengono enfatizzate e bande in cui invece le componenti vengono attenuate.

Questo resta maggiormente vero per i suoni vocalici (almeno quelli articolati mantenendo un timbro statico per un certo tempo) che presentano una struttura spettrale più connotata e più facilmente leggibile. L'assenza di particolari ostruzioni nell'articolazione di questi suoni fa sì che l'energia glottidale venga irradiata con una minore attenuazione e con una maggiore conservazione della sua struttura armonica.

Come descritto nella I parte, gli effetti che l'assunzione di distinte conformazioni delle cavità supralaringali provocano sullo spettro finale del suono prodotto sono in parte prevedibili mediante una rappresentazione semplificata di queste come dei tubi acustici e facendo ricorso a un modello acustico studiato secondo le teorie della risonanza e della perturbazione[16].

La ricostruzione del sistema timbrico di una lingua su un piano acustico è quindi affidato a uno o più diagrammi cartesiani (con scelte diverse, a seconda degli autori, sulle variabili da rappresentare) di solito miranti a ritrovare la disposizione dei timbri del trapezio vocalico nelle aree di esistenza di un gran numero di parlanti (maschi e femmine)[17].

Un diagramma che si presta bene a queste operazioni (e che sia facile da ottenere) è il **diagramma F_2-F_1** che, per i tre vocoidi delle figure 13-15, avrebbe le caratteristiche di quello in Fig. 16. In molti casi, tuttavia, volendo tenere conto almeno degli effetti della labialità nelle opposizioni vocaliche, si fa riferimento anche a un diagramma F_2-F_3 (v. dopo)[18].

Da Delattre (1948) e Delattre *et alii* (1952) a oggi, sono state numerose le pubblicazioni di autori che si sono appoggiati su questi diagrammi per discutere dei diversi sistemi vocalici delle lingue o per approfondire caratteristiche specifiche.

Nonostante le diverse modalità di variazione che si può avere, anche tra parlanti della stessa varietà linguistica, in base alle proprie caratteristiche anatomo-fisiologiche dell'apparato di produzione e in base al proprio profilo socio-geo-linguistico, è evidente che diagrammi di questo tipo possono essere sfruttati per rendere conto, caso per caso, delle diverse opposizioni che ogni lingua si ritrova ad affidare a diverse selezioni di timbri.

---

[16] Per una descrizione accurata di queste teorie si rimanda a contributi più specializzati (v., tra gli altri, *Calliope* 1989 e Kent & Read 1992). L'assunzione di un modello acustico che stabilisce un'analogia tra il condotto vocale e un tubo acustico uniforme chiuso a un'estremità risale al XIX sec., ma si deve a Fant (1960) un'esposizione esaustiva dell'argomento. Sono note le conseguenze sulle frequenze di risonanza di un tubo con queste caratteristiche delle deformazioni imposte al tubo in punti diversi. Sono noti anche i cambiamenti di affiliazione tra le frequenze delle risonanze e le cavità che si formano per via di restringimenti locali in certe posizioni. Alcune valutazioni di questo genere sono proposte anche da Ladefoged (1996[2]) per spiegare alcune dinamiche formantiche. Si veda però il §7.

[17] Generalmente si osservano valori significativamente più elevati di F_1 e F_2 in voci femminili. Numerose tecniche sono state proposte per normalizzare le differenze specifiche legate alle dimensioni degli organi di produzione dei suoni vocali (in particolar modo nel tentativo di ridurre l'effetto di una diversa lunghezza del canale epilaringeo, v. Wakita 1977, e per neutralizzare le conseguenze di una diversa f_0, v. anche §5). Le variabili normalizzate usate più comunemente si riferiscono tuttavia a scale definite su base percettiva come quella in *ERB* (*Equivalent Rectangular Bandwidth*) o quella in *bark* (ssssss) o a complesse riflessioni sull'interazione di variabili diverse, come già illustrato in Fant (1975) e Disner (1980) e recentemente ridiscusso in lavori come quello di Halberstam & Raphael (2004).

[18] L'interazione tra F_2 e F_3 ha indotto alcuni studiosi a valutare gli effetti che F_3 ha nella percezione di F_2 (determinando anche numericamene, con opportune ponderazioni, valori di F_2 *efficaci* che tengano conto anche di questi contributi). Inoltre, l'osservazione della disposizione delle principali formanti nella definizione dei timbri vocalici più comuni (Schwartz *et alii* 1997) ha permesso di stabilire l'influenza che strategie di convergenza/divergenza tra formanti possono avere nella determinazione dei sistemi vocalici nelle diverse lingue.

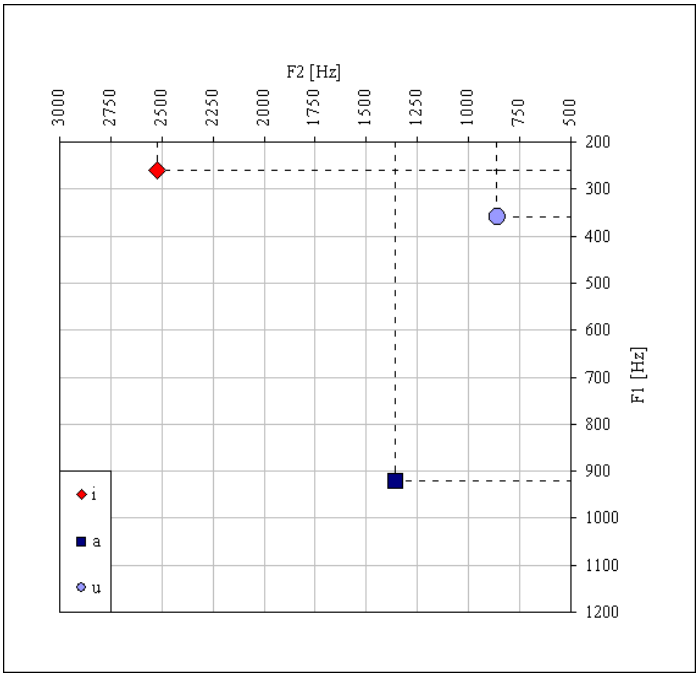

Figura 16. Diagramma F_2-F_1 per le tre realizzazioni vocaliche analizzate nelle Figg. 13-15. Oltre alla tipica inversione delle scale, si noti la disposizione di F_2 sull'asse delle ascisse e di F_1 sull'asse delle ordinate.

In Fig. 17 si propone una rappresentazione delle misure delle prime tre formanti eseguite sulle 28 vocali del trapezio vocalico dell'Associazione Fonetica Internazionale corrispondenti a quelle disponibili nell'illustrazione dei suoni vocalici sul sito interattivo del *LFSAG* (http://www.lfsag.unito.it).

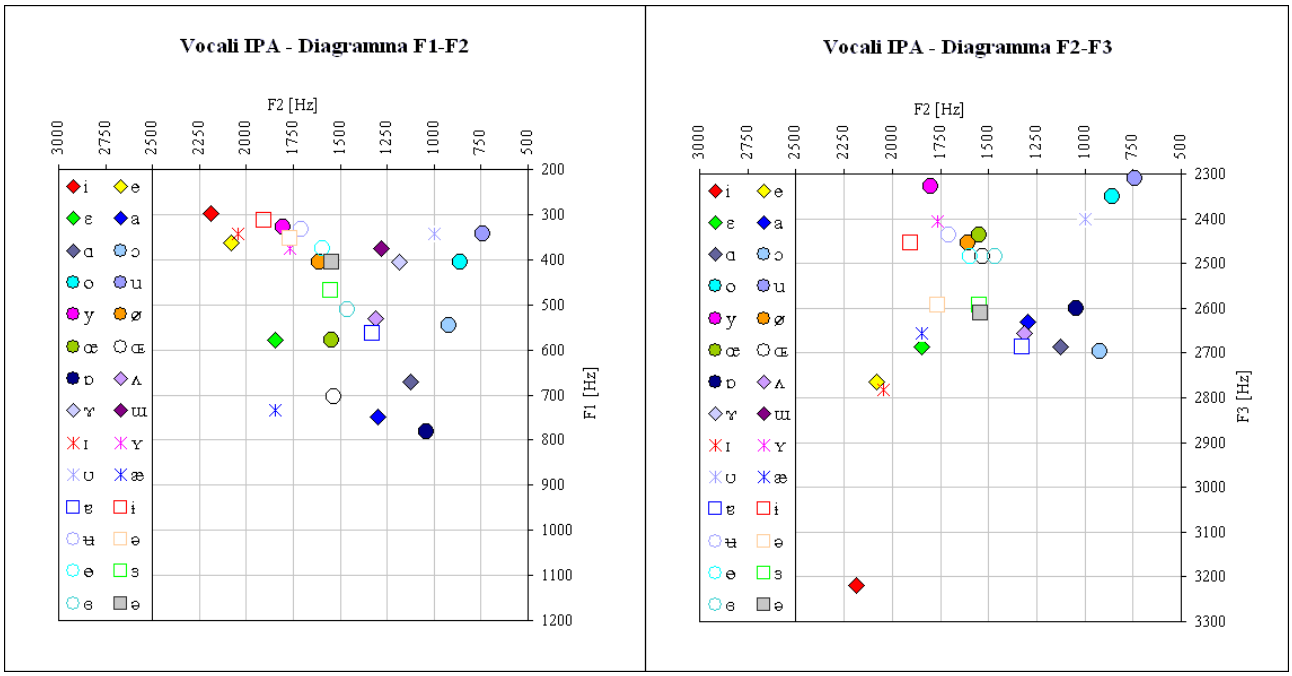

Figura 17. Diagrammi F_2-F_1 e F_2-F_3 di una realizzazione delle 28 vocali *IPA*. I rombi sono riservati alle vocali non arrotondate (i circoli invece alle vocali arrotondate), i quadrati alle vocali centrali e gli asterischi alle vocali abbassate, alzate o centralizzate (quasi-chiuse o quasi-aperte).

In Fig. 18 si propone invece un diagramma F_2-F_1 per un insieme di misure eseguite sulle produzioni di un dicitore professionale d'italiano. Nel grafico sono riportate le prime due formanti di dieci realizzazioni di ciascuna delle 7 vocali dell'italiano in posizione accentata, misurate nella regione di massima stabilità del timbro. Come si può osservare le diverse misure definiscono delle aree di dispersione distinte per ciascuna vocale. Per questo motivo si parla di diagrammi di dispersione sul piano F_2-F_1 (e di aree di esistenza dei fonemi vocalici, delimitate di solito in ellissi dette di "equi-probabilità")[19].

Il riferimento più classico per il sistema vocalico dell'italiano è Ferrero (1968), al quale sono seguiti numerosi contributi specifici relativi a varietà geolettali o sociolettali (v. già Ferrero, Genre , Boë & Contini1979)[20]. Una discussione sulle problematiche relative alla lettura di questi grafici (e in generale alle modalità di misurazione delle formanti) è in Ferrero (1996).

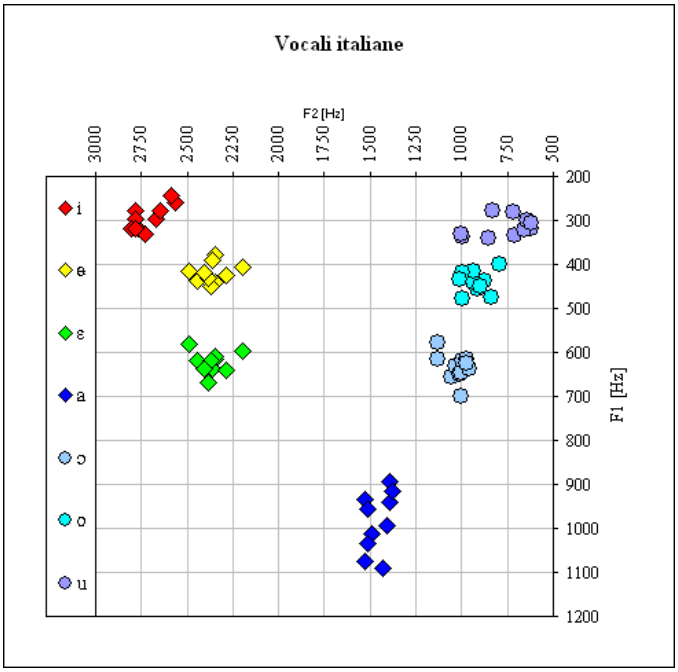

Figura 18. Diagrammi di dispersione sul piano F_2-F_1 di dieci realizzazioni da parte di un locutore professionale di ciascuna delle 7 vocali possibili in italiano in posizione accentata (v. Fig. 17).

[19] L'analisi statistica applicata alle misure di F_1 e F_2 di una stessa vocale tiene conto della relativa dipendenza statistica tra queste due variabili e porta quindi al tracciamento di ellissi che possono risultare più o meno inclinate sul piano F_2-F_1 e sono convenzionalmente individuate per mezzo di un centroide (la cui posizione dipende dalla media dei valori misurati per ciascuna formante) e da due semi-assi (le cui dimensioni dipendono da una soglia di equi-probabilità stabilità a priori e che sono proporzionali alla varianza congiunta stimabile proprio a partire dalla disposizione dei dati misurati)

[20] Una rassegna di studi sperimentali sui sistemi vocalici diffusi in Italia, aggiornata al 2002, è quella di Calamai (2003).

IV.7. La lettura degli spettrogrammi

Mentre il concetto di formante è abbastanza preciso e in casi ideali non pone problemi, manca una precisa definizione operazionale che ne permetta un riconoscimento automatico nel caso più generale (che autorizzi ad esempio l'esclusione di quella evidenziata in Fig. 6).

Si entra qui nei problemi dell'interpretazione dello spettrogramma, problemi che sono complessi per il fatto – messo ben in evidenza sin da Fant (1960) – che una stessa variante articolatoria, da un individuo all'altro e da un'esecuzione all'altra dello stesso parlante, può determinare manifestazioni spettrografiche diverse e che, reciprocamente, a un'indicazione spettrografica simile possono corrispondere strategie articolatorie diverse.

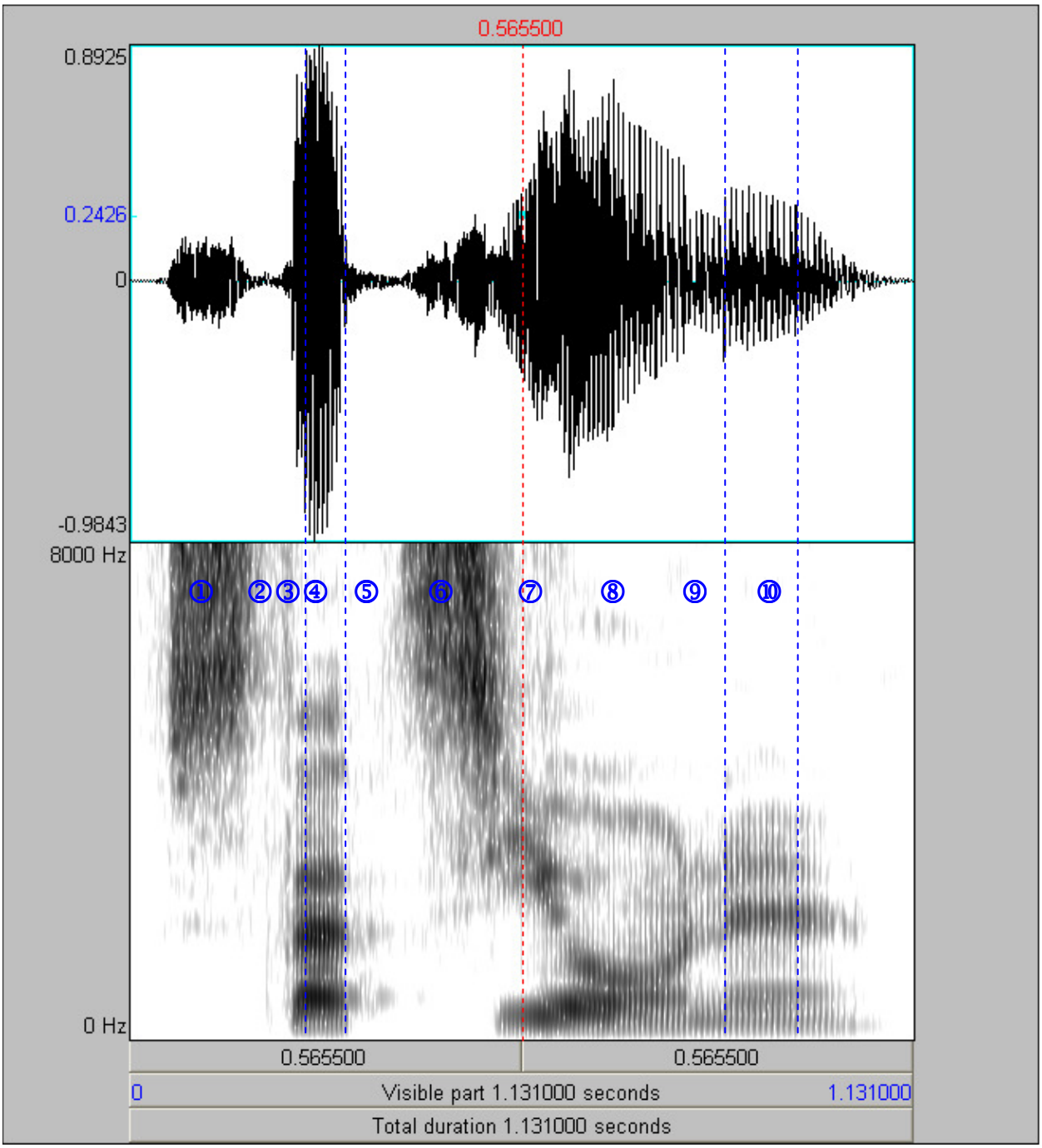

Figura 19. Oscillogramma (in alto) e spettrogramma (in basso) di una realizzazione della parola *stazione* pronunciata da un parlante italiano (dicitore professionale). Si possono distinguere 10 fasi temporali distinte (indicate dalle cifre cerchiate) associate a diversi momenti articolatori dei vari segmenti (v. testo).

L'interpretazione inizia con la segmentazione del tracciato, cioè con l'individuare le caratteristiche delimitative dei segmenti e con la classificazione degli eventi presenti in ciascun segmento[21].

Per riuscire a trarre dagli spettrogrammi il maggior numero d'informazioni relative al messaggio sottoposto ad analisi (quando è noto il codice usato dal locutore che lo ha prodotto) sono tuttavia necessarie alcune indicazioni di base. Ci proponiamo di discuterne qui alcune a partire dall'osservazione di un solo esempio relativo a una realizzazione della parola 'stazione' da parte di un parlante italofono senza particolari marche geo- o socio-fonetiche (v. Fig. 19)[22].

Le principali osservazioni che possiamo fare su questi grafici sono relativi alla particolare successione di segmenti che riusciamo a individuare nel corso dello sviluppo temporale della produzione (lungo l'asse delle ascisse), nonostante lo sviluppo continuo e gl'importanti fenomeni di coarticolazione.

Osserviamo sin dall'inizio, grazie alla bassa densità di righe verticali delle regioni di sonorità vocale, che si tratta di una voce maschile[23].

Al di là di questo dato, facendo finta di non conoscere il contenuto sonoro (e d'ignorare quindi la rappresentazione fonetica simbolica) di questa produzione, si possono distinguere i segmenti acustici che si candidano a rappresentare i suoni dell'ideale catena sonora che la contraddistingue.

Notiamo da sinistra verso destra, una macchia informe concentrata prevalentemente sulle alte frequenze (tipica di un rumore di frizione) ①, seguita da una pausa di silenzio ② e da una barra verticale piuttosto stretta ③ (distinta dalle striature verticali che segnalano solitamente la sonorità). Segue poi – appunto – una regione di sonorità ④, caratterizzata da un'ampia striatura orizzontale tipica delle strutture formantiche vocaliche, interrotto nuovamente da una pausa silente (con un'eco parassita) ⑤.

Si ripresenta poi un rumore con caratteristiche frequenziali simili a quelle del primo, ma con una distinte dinamiche frequenziali ed energetica ⑥ (questa fase inizia con una barra di esplosione leggermente accennata). A questo fa seguito una lunga regione di sonorità caratterizzata da un'energia variabile, ma soprattutto da notevoli evoluzioni formantiche.

Dapprincipio, con un graduale aumento d'intensità, si presenta un *pattern* formantico con una certa convergenza in una stessa regione delle formanti più alte (le quali sembrano svilupparsi direttamente dal rumore che segnala il segmento precedente) ⑦. In concomitanza con un aumento d'energia, si ha poi il passaggio a una nuova configurazione abbastanza stabile per un certo tempo ⑧. Questa è a sua volta interrotta da un improvviso calo d'energia associato a una leggera perturbazione del tracciato di evoluzione delle singole formanti (le quali, oltre che essere più deboli, in alcuni tratti sembrano soggette a importanti oscillazioni sulla larghezza di banda) ⑨. L'energia poi riaumenta nel segmento finale che riacquista stabilità nel *pattern* formantico a un'energia comunque relativamente ridotta ⑩.

---

[21] Interessanti contributi in merito a questo delicato tema sono in Abry *et alii* (1985) e Salza (1991) i quali propongono di basarsi su un insieme specifico d'indici temporali per una delimitazione convenzionale dei segmenti.

[22] Sono numerosi nella letteratura internazionale i manuali che si propongono come guide per la lettura degli spettrogrammi (v., tra gli altri, Ferrero *et alii* 1979 e, in ambito internazionale, Ladefoged 1996). Si tratta in alcuni casi di prontuari d'uso (come quello di Painter 1979) che offrono una rassegna d'esempi di realizzazione dei suoni di una determinata lingua studiati spettrograficamente e discussi in termini di tratti caratterizzanti sul piano articolatorio e acustico. Un manuale utile a questo scopo, rivolto all'illustrazione dei suoni dell'italiano, è quello di Giannini & Pettorino (1992) che, oltre a discutere la lettura di numerosi esempi (di laboratorio), propone anche un certo numero di esercizi (con soluzione). In Albano Leoni & Maturi (1995) si trova, invece, un interessante quadro di variazione che espone alcune caratteristiche più critiche della lettura degli spettrogrammi quando relativi a un parlato meno controllato (più spontaneo e, in genere, ipoarticolato).

[23] Possiamo valutare il numero approssimativo di cicli di vibrazione delle corde vocali nell'unità di tempo contando le righe verticali: ad esempio ne rileviamo 8 in 50 ms nel caso della prima vocale (nella finestra delimitata dai primi due demarcatori verticali; in 1 s ce ne saranno 20 volte tante, quindi 160 al secondo) e 12 in 100 ms nel caso dell'ultima vocale (nella finestra delimitata dagli ultimi due demarcatori verticali; in 1 s ce ne saranno 10 volte tante, quindi 120 al secondo). Queste misure indicano una variazione di frequenza fondamentale (f_0) da circa 160 Hz a circa 120 Hz, dato tipico di voci maschili (in enunciazioni di tipo dichiarativo).

Ripercorrendo ora queste distinte regioni, proviamo a formulare delle ipotesi sugl'indici rilevati per i singoli segmenti, osservando continuità e discontinuità della barra di sonorità (che corre nella parte più bassa dello spettrogramma lungo l'asse del tempo, marcando come sonori i segmenti ④, ⑦, ⑧, ⑨ e ⑩).

Sempre simulando d'ignorare la parola pronunciata, notiamo che il primo segmento ① si candida a rappresentare una fricativa sorda che, date le sue proprietà spettrali (maggiore concentrazione nella banda 5÷8 kHz, v. §9), non può che corrispondere a un suono alveodentale [s]. Segue la pausa silente ② che contraddistingue la fase di tenuta di un'occlusiva sorda; solo dagl'indici (di durata e di estensione frequenziale) dell'esplosione ③ possiamo dedurre che si tratta anche in questo caso di un suono alveodentale [t]. Della vocale seguente ④ possiamo dire che deve trattarsi di una vocale aperta (F_1 è pari all'incirca a 600 Hz o più) e piuttosto centrale o anteriore (F_2 è superiore a 1600 Hz): il fonema vocalico italiano che più spesso presenta realizzazioni con queste caratteristiche è /a/.

Della fase seguente ⑤ possiamo solo dire che rappresenta di nuovo la manifestazione della tenuta di un segmento occlusivo puro o affricato (sordo). Data la presenza di una fase di frizione, preceduta da una leggera barra di esplosione, ⑥ con le stesse proprietà del primo segmento, concludiamo che si tratta qui di un'affricata dentale [t͡s] (e che, data la sua lunghezza complessiva, si tratti piuttosto di una realizzazione lunga, solitamente considerata di tipo [t:s]).

Il caratteristico schema formantico del segmento ⑦ (con F_1 sotto i 400 Hz e F_2 che corre su valori pari circa a 2300 Hz), unitamente al suo andamento energetico gradualmente crescente, ci fa riconoscere una realizzazione approssimante palatale [j] (una sorta di /i/ consonantica).

Nel segmento ⑧ è facilmente riconoscibile una vocale (che possiamo anche assumere accentata, data la durata della sua fase di maggiore stabilità) che, con i valori approssimativi di 400 Hz per F_1 e 800 Hz per F_2, è sicuramente alta e posteriore (possiamo pensare a una realizzazione di /o/).

L'ultimo segmento ⑩ è ancora una vocale (debole, ma abbastanza lunga) con F_1 intorno a 500 Hz e F_2 intorno ai 2000 Hz: anche per via della sua intensità e della sua posizione, la possiamo riconoscere come una realizzazione di /e/ finale non accentata dell'italiano.

Il segmento tra questi ultimi due, il ⑨, si presenta con un netto calo energetico rispetto a questi: anche per la sua durata e per le caratteristiche formanti 'allargate' (di cui la prima sotto i 400 Hz) si configura tipicamente come una consonante nasale. Per le deviazioni formantiche causate nei segmenti adiacenti, in italiano non può che essere una delle rese possibili di /n/.

Abbiamo così ricomposte una sequenza di tipo [statsjone] che possiamo migliorare, tenendo conto delle durate dei segmenti individuati, trascrivendo [staˈt͡sjo:ne] (che rappresenta bene una resa non connotata dell'italiano *stazione*, cui – a seconda degli autori – possono corrispondere le forme /staˈt͡sjone/, /statsˈtsjone/o /statˈtsjone/).

Le caratteristiche classificatorie su cui basare le considerazioni che portano al riconoscimento dei suoni presenti in una produzione linguistica analizzata su base spettrografica sono state codificate in diverso modo nel corso degli anni. La più celebre tipologia di tratti distintivi basata su considerazioni acustiche è quella proposta da Jakobson, Fant & Halle (1952) e illustrata in Ferrero *et alii* (1979). D'altra parte, come discusso nella I parte, diversi approcci di analisi fonologica si sono appoggiati sull'osservazione sperimentale e hanno portato a metodi di parametrizzazione (e di rappresentazione) che, pur limitandosi di solito a poche applicazioni, sono ora diffusamente accolti nei principali trattati di fonetica sperimentale (si veda ad es. Laver 1994).

Si propone in Fig. 20 un esempio di parametrizzazione articolatoria dell'esempio analizzato in Fig. 19 che mette in evidenza ampie regioni di sovrapposizione (e tracimazione) dei gesti articolatori necessari per il raggiungimento di determinati obiettivi (bersagli acustici) necessari per la realizzazione del "segmento"[24].

[24] Si noti che un *tool* di sintesi articolatoria basato sul controllo temporale di una diversa selezione di parametri articolatori (v. Boersma 1998) è disponibile anche in PRAAT (*Artword*).

L'attività (attivazione e/o disattivazione) di un "organo mobile" nel corso del tempo è ovviamente continua ma può essere valutata in termini binari stabilendo delle soglie convenzionali (presenza, posizione o valore sopra la soglia, vs. assenza, posizione o valore sotto la soglia).

Un flusso d'aria polmonare è presente ad esempio nella realizzazione delle fasi ①, ③, ④, ⑥, ⑦, ⑧, ⑨ e ⑩, ma non in ② e ⑤, caratterizzate da interruzione di questo nella fase di tenuta (occlusione) di [t] e [t͡s]. In particolare si nota, dalla linea relativa all'attività delle pliche vocali, che solo nelle fasi ④, ⑦, ⑧, ⑨ e ⑩ questo flusso è impiegato per la produzione di energia vocale (sonorità) dato che nelle fasi ①, ③ e ⑥ è invece sfruttato per la produzione di rumore di costrizione, rilascio o esplosione. In tal modo i segmenti [s], [t] e [t͡s] si caratterizzano come non sonori (sordi).

La parte anteriore della lingua (la corona, il cui abbassamento è controllato in modo prevalente dal muscolo genioglosso) è sicuramente sollevata (oltre una certa soglia, al confine tra suoni vocalici e consonantici) nelle fasi ①, ②, ③, ⑤, ⑥, ⑦ e ⑨, determinando occlusioni, costrizioni o approssimazioni, mentre non lo è nelle fasi (④, ⑧ e ⑩) in cui si determina un'apertura sufficiente alla produzione di suoni vocalici. L'arrotondamento labiale (determinato principalmente dal controllo del muscolo orbicolare) è presente solo in una di queste: la ⑧, relativa alla produzione di [o] che richiede anche un innalzamento e un arretramento del dorso della lingua (regolati dal muscolo stiloglosso, attivo in questa stessa fase). Al contrario, un abbassamento della regione posteriore della lingua (regolato dal muscolo antagonista ioglosso) è invece necessario per gli altri due suoni vocalici (④ e ⑩).

L'abbassamento del velo palatino (controllato dal levatore e dal tensore palatini) è necessario a un certo punto per consentire al flusso di mettere in risonanza le cavità nasali (⑨) per la realizzazione di [n]. Com'è stato messo in evidenza nel grafico, quest'abbassamento incomincia progressivamente prima del previsto e termina con un certo ritardo, determinando parziale nasalizzazione dei segmenti vocalici adiacenti ([o] e [e])[25].

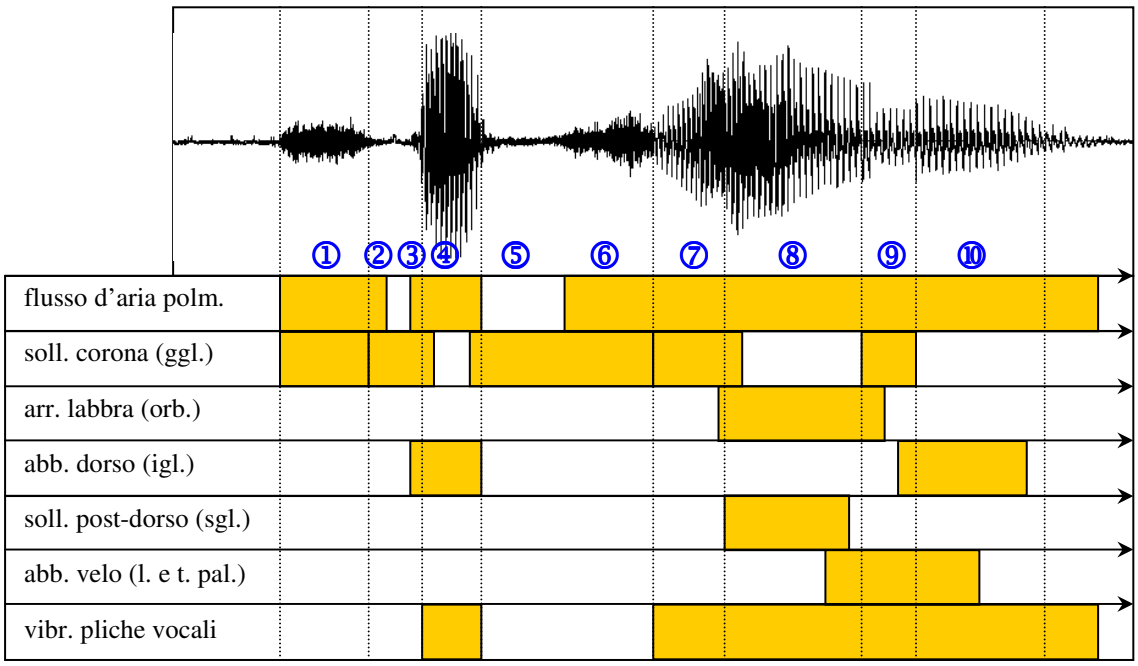

Figura 20. Oscillogramma con i segmenti delimitati per l'esempio di Fig. 19 (in alto) in associazione con gli eventi articolatori verosimilmente responsabili degl'indici acustici osservabili nei grafici dell'analisi spettrografica. L'attività di un dato articolatore parametrizzato sull'asse temporale è valutata in termini binari (presenza, in bande grigie, vs. assenza; v. testo; cfr. III parte, §III.II.2).

[25] Nel caso di [o], inoltre, dato che quest'abbassamento – in condizioni di sollevamento del dorso – potrebbe portare a un'eccessiva costrizione nella regione dorso-velare, al progressivo sollevamento della corona (necessario per articolare [n] al livello orale) si accompagna una distensione anticipata dello stiloglosso la cui contrazione sarebbe altrimenti in competizione con queste attività.

IV.8. Indici spettrografici dell'attività articolatoria

Tutte le dinamiche fin qui osservate possono essere sommariamente schematizzate riferendosi a un insieme di categorie articolatorie (*tratti*) più tradizionali nell'ambito della linguistica e adattato già nella prima edizione di questo manuale, distinguendo tra quei tratti relativi alla sorgente e quelli relativi al filtro (cavità di risonanza). A ciascun tratto sono poi associati degl'*indici* spettrografici[26].

Tratti relativi alla sorgente:
1. periodica (o quasi periodica; sonorità di vario tipo: anche voce mormorata, cricchiata etc.);
2. aperiodica (a eccitazioni multiple non periodiche; rumore);
3. a carattere impulsivo, intermedio o transitorio (esplosioni, fasi d'innesco e/o disattivazione).

Tratti relativi al risonatore:
4. occlusione;
5. costrizione (fricatività);
6. lateralità;
7. nasalità;
8. vocalità (apertura, sonorità vocalica);
9. a carattere intermedio o transitorio (transizioni + fasi e/o movimenti approssimanti[27]).

Un tratto come quello del tipo 9 è necessario anche perché non di rado le caratteristiche di passaggio da un segmento all'altro non sono immediate e può essere necessario distinguere transizioni più o meno lunghe (v. §9).

Così, nell'esempio delle Figg. 19 e 20, si assume che la lunga transizione tra [t͡s] e [o] (fase ⑦) corrisponda a un suono di tipo [j] (anche per via dell'importante variazione timbrica associata) mentre la più breve transizione nel passaggio da [o] a [n] (pur consistente) si attribuisce ancora al segmento vocalico per via del mantenimento di condizioni energetiche di tipo vocalico (vibrazione delle pliche vocali irradiata dal tratto vocale aperto) prima della brusca diminuzione d'energia causata dalla repentina metastasi dell'articolazione occlusiva orale della nasale (e dall'attivazione dei risonatori nasali attraverso i quali avviene l'irraggiamento). Questo corrisponde tra l'altro anche all'esito della valutazione tradizionalmente condotta su basi (orto-)grafiche (e talvolta storiche) di natura spesso percettiva.

I problemi di classificazione che nascono da queste osservazioni dipendono talvolta da convenzioni di segmentazione, ma tengono conto in genere di considerazioni legate a valutazioni di tipo macro-categoriale (distinguendo ad es. il tratto 8 da tutti gli altri per via delle caratteristiche energetiche delle risonanze, associate a condizioni di apertura, o per la presenza/assenza di rumore, prima ancora che in base al tipo di rumore che caratterizza ad es. i tratti 4 e 5). Altri tratti che sono stati proposti non compaiono in quest'elenco perché le loro caratteristiche possono essere date dall'alternanza o dalla sovrapposizione di più di uno degli altri. Ad es. il tratto che caratterizza i suoni vibranti può essere considerato ridondante in quanto i vibrati (monovibranti) possono essere in genere descritti usando il tratto 4, mentre le polivibranti sono in genere date dall'alternanza di fasi associate a condizioni di tipo 4 e 8.

[26] Almeno da Delattre (1970), "des indices acoustiques aux traits pertinents", si parla d'indici, in senso lato, per riferirsi a qualsiasi indicazione fonologica (linguistica) desumibile dall'osservazione di una variabile acustica. Per "indice" (*cue*) s'intendeva originariamente la deviazione positiva o negativa (o nulla) che subisce sull'asse verticale una formante di un segmento per effetto della configurazione assunta nel segmento contiguo. Gl'indici di deviazione formantica in una vocale sono, ad esempio, secondo Delattre (1962), la "chiave della percezione delle consonanti adiacenti" (spesso povere di caratterizzazione acustica distintiva).

[27] Questi fenomeni, un tempo classificati come *glide* o come "costrittive larghe", hanno progressivamente assunto (partendo dalla definizione di *approssimanti* introdotta da Ladefoged 1964: 25) una definizione che ne enfatizza l'aspetto dinamico di approssimazione da parte di un organo mobile nei confronti di un organo fisso (o tra due organi mobili). Ladefoged (1975: 277), rielaborando la sua originaria definizione, descrive le approssimanti come consonanti per la cui produzione si verifica un 'approach of one articulator towards another but without the vocal tract being narrowed to such an extent that a turbulent airstream is produced'.

Gl'indici spettrografici che determinano l'ipotesi della presenza di uno dei tratti succitati si possono riassumere nel seguente elenco:

1. striatura verticale variabile (dipendente da f_0) e "colorata"[28] (energia a bande formantiche) associata a un evidente annerimento orizzontale in bassa fequenza (barra di sonorità, F_0);
2. striatura irregolare, debole o assente;
3. striatura verticale impulsiva (singola) e localizzata (*spike*);
4. estesa regione bianca (o comunque chiara) a tutte le frequenze (tranne l'eventuale barra di sonorità, descritta in 1.) seguita da uno o più degli eventi descritti in 3., 5. e 8.;
5. estese regioni di annerimento irregolare (con parziale striatura) soprattutto alle alte frequenze (e, nelle sonore, anche in presenza della barra di sonorità, descritta in 1.) che determinano un rumore "colorato" (*noise*, per le fricative, oppure *burst*, per le occlusive aspirate e/o affricate);
6. indici di tipo 1. e 8. ma con energie ridotte (intorno ai –20dB) e distinte localizzazioni formantiche con bande in genere allargate e talvolta degeneranti in caratteristiche di tipo 5. (possibilità di antirisonanze, sottili linee bianche tra le formanti, e di formanti additive alle alte frequenze);
7. indici di tipo 1. e 8. ma con energie ridotte (intorno ai –10dB) e distinte localizzazioni formantiche con bande in genere allargate e/o separate da antirisonanze (*formant splitting*) e arricchite da formanti additive a diverse frequenze (*nasal murmur*, F_{1N} = 400-500 Hz, F_{2N} = 1000-1200 Hz etc.) oppure indici di tipo 1. e struttura formantica arricchita (10-11 formanti anziché 6-8) con contributi ben visibili (rispetto ai suoni orali corrispondenti) di F_{1N}, F_{2N}, F_{3N} etc.;
8. strutture formantiche ben visibili, in genere in associazione con tratti di tipo 1. o 2.;
9. strutture formantiche piuttosto ben visibili ma con progressivo indebolimento e/o intensificazione associato a decisi movimenti formantici ben delineati rispetto alle strutture formantiche dei segmenti vocalici (o consonantici) precedenti e/o seguenti.

La necessità di distinguere all'interno di quest'ultimo tipo d'indici tra transizioni, da un lato, e fasi e/o movimenti approssimanti, dall'altro rileva di processi distinti: nel primo caso si tratta, infatti, di transizioni acustiche associate a una modificazione fisiologica degli pneumatocelomi nel passaggio da una configurazione all'altra, mentre nel secondo caso si tratta di movimenti volontari (additivi) dotati di valore acustico e salienza percettiva tale da determinare rappresentazioni linguistiche distinte.

La direzione dei movimenti del primo tipo è il riflesso del cambiamento del punto d'articolazione da un suono all'altro e, com'è discusso sin da Öhman (1966), dipende da condizioni specifiche di co-articolazione.

Al variare del suono vocalico contiguo, per una stessa consonante, le deviazioni di una stessa formate convergono – procedendo con inclinazione differente – verso un punto fisso caratteristico della consonante (*locus*); in particolare i *loci* di F_2 e F_3 suggeriscono il punto d'articolazione di questa (v. §9)[29].

A titolo d'esempio, in fig. 21 si riportano le realizzazioni di tre sillabe *CV* (consonante-vocale) formate dalla stessa consonante occlusiva bilabiale sonora /b/ e dalle tre vocali /i/, /a/ e /u/. Si nota come, al variare della vocale, anche la struttura formantica della consonante (per quanto debole) risulti leggermente diversa (per gli effetti anticipatori della co-articolazione). I tratti più caratteristici del segmento consonantico risultano di tipo 4 (che include indici di tipo 1 – barra di sonorità –, per tutta la sua durata, e di tipo 3 – esplosione o *spike* – alla fine, subito prima dell'inizio del segmento seguente). Il segmento vocalico seguente si caratterizza invece, nei tre casi, come di tipo 8, con transizioni formantiche che presentano deviazioni simili (dal basso verso l'alto).

[28] Si definisce "colorato" uno spettro non piatto, ma con addensamenti energetici a bande ben precise (in opposizione a uno spettro piatto che, nel caso di rumore – cioè di segnale aperiodico –, si definisce "bianco"). Si noti che a uno spettro piatto corrisponde una vista spettrografica con annerimento uniforme.

[29] In generale, per le consonanti con punto d'articolazione orale, il *locus* di F_1 (L_1) è inferiore a 500 Hz, mentre per quelle con punto d'articolazione faringale $L_1 > 500$ Hz.

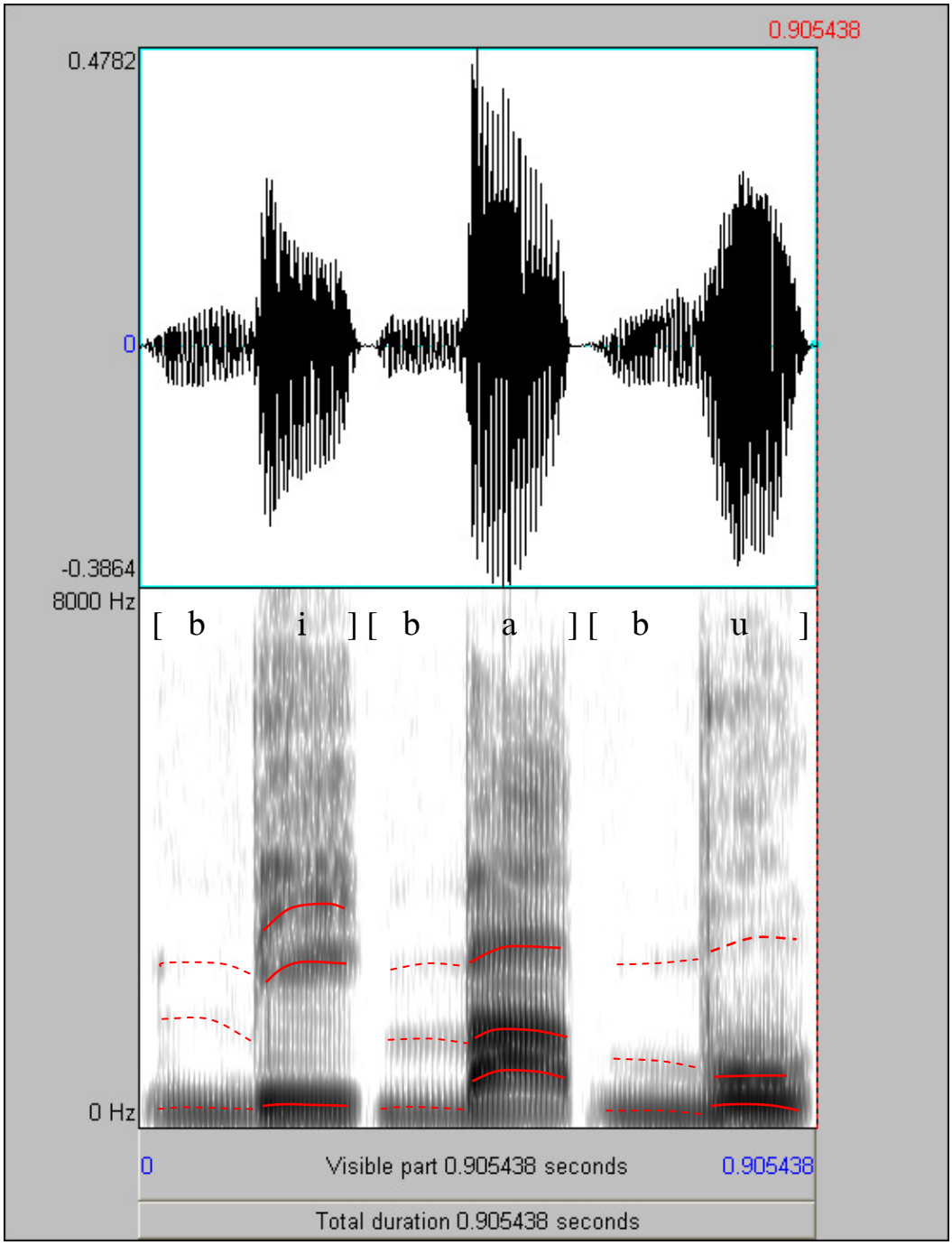

Figura 21. Oscillogramma e spettrogramma di una realizzazione delle tre sillabe [bi], [ba] e [bu] pronunciate da un parlante italiano (dicitore professionale). Oltre al distinto timbro delle tre vocali (qui evidenziato dall'andamento delle prime tre formanti), si notino le diverse strutture formantiche visibili nelle tre realizzazioni di /b/ e le transizioni formantiche nella fase di metastasi delle tre vocali (v. testo).

Le deviazioni formantiche nei segmenti vocalici suggeriscono, con il loro andamento, il luogo ma anche il modo d'articolazione della consonante precedente (oppure seguente, al confine tra vocale e consonante, nel caso di nessi *VC*): in genere essi sono tanto più lenti quanto maggiormente sono presenti le caratteristiche 1 e 8). Tra gl'indici che contribuiscono alla caratterizzazione del luogo d'articolazione, oltre alle deviazioni di F_2 (v. §9), ricordiamo anche una certa interazione tra F_2 e F_3 (con F_2 che converge su F_3 nelle articolazioni prepalatali o si spostano insieme al massimo livello nelle articolazioni mediopalatali), l'effetto generale della protrusione labiale (che abbassa tutte le formanti) o la deviazione delle formanti verso regioni di particolare intensità di rumore ("formanti di rumore", v. §10).

IV.9. La teoria del loci *acustici*

Se la teoria della risonanza (e della perturbazione) acustica, ci permette di prevedere (o riconoscere *a posteriori* le ragioni di) una diversa disposizione di formanti nella fase di maggiore stabilità di una vocale, sin dai primordi dell'analisi spettrografica (Delattre *et alii* 1955; Delattre 1961) era apparso chiaramente come, osservando le deviazioni delle formanti di una vocale, era anche possibile definire le conseguenze acustiche di diversi punti d'articolazione assunti nel corso della produzione dei suoni contigui.

Anche in questo caso, se da un lato l'assunzione di configurazioni articolatorie successive causa riaggiustamenti in tutta la struttura formantica del suono risultante, portatrici dei maggiori indici dei cambiamenti articolatori sono soprattutto le transizioni della seconda formante (F_2).

In generale, dall'osservazione delle formanti e delle loro diverse deviazioni in contesti *CVC*, in funzione del timbro della vocale, è possibile definire delle regioni frequenziali immaginarie di provenienza o di destinazione (prima o dopo la vocale).

In particolare, nel caso di F_2, il *locus* acustico di questa regione – indicato con L_2 –, in seguito a numerose osservazioni e classificazioni, si è confermato come un valido indicatore del luogo d'articolazione della consonante (soprattutto per quelle ostruenti, che presentano pochi altri indici e per di più, spesso, anche poco affidabili)[30].

I metodi per la sua determinazione sono essenzialmente due: uno manuale e uno algebrico.

Il metodo manuale (di cui si presenta una schematizzazione in figura 22 per le tre consonanti [b], [d] e [g]), noto anche come **metodo dell'intersezione**, si applica a partire dall'osservazione di F_2 almeno nei tre contesti forniti dalle vocali agli estremi dello spazio vocalico (generalmente, se presenti nel sistema studiato, [i], [a] e [u]).

La direzione di deviazione di F_2 viene annotata caso per caso e riprodotta come un insieme di rette convergenti in un unico grafico cumulativo. Il punto (o più spesso l'area) d'intersezione delle tre rette definisce una stima di L_2.

Questo procedimento, illustrato nell'ultima serie di grafici in basso in figura 22, permette ad esempio di determinare, in un caso ideale (basato però su misurazioni reali relative a un italiano di laboratorio), i tre *loci* corrispondenti ai punti d'articolazione bilabiale, alveodentale e velare.

Se per i primi due punti si ottengono valori univoci (in accordo con quelli ottenuti nella maggior parte degli studi condotti sull'italiano) e cioè, rispettivamente, 650 Hz e 1600 Hz, nel caso delle velari è noto (cfr. Giannini & Pettorino 1992: 199) che le aree di convergenza dei prolungamenti delle rette di deviazione sono rese diverse dalla presenza/assenza della labialità nell'articolazione della vocale.

Nel caso di vocali non protruse il *locus* è di solito piuttosto alto (2600 Hz nel nostro esempio); in presenza di protrusione sembra invece dominare la labialità, che riporta molto più in basso l'area di convergenza.

L'altro metodo, noto come **metodo dell'equazione dei *loci***, sembra esser più robusto e mira a una determinazione algebrica di L_2, risultante dalla stima preliminare di altri due parametri: il parametro angolare e l'intercetta della retta di regressione descritta dalla dispersione dei valori misurabili di F_2 nelle fasi di transizione (da e verso la consonante, entrambe definibili *onset* in base a un'adeguata direzione d'osservazione) in funzione dei valori raggiunti nella fase di massima stabilità della vocale (*offset*).

L'equazione dei *loci* è una retta $y = mx + q$ definita da un parametro angolare m e da un'intercetta q, i cui valori risultano empiricamente sulla base delle coppie ($y = F2_{offset}$, $x = F2_{onset}$).

[30] Naturalmente, la presenza di una struttura formantica visibile nel segmento consonantico darebbe già di per sé indicazioni di massima per la determinazione del luogo d'articolazione (v. es. in Fig. 21). Similmente, informazioni di luogo possono essere anche desunte dal tipo, dalla concentrazione e dall'energia delle tracce di rumore eventualmente presenti (v. §10). Il metodo dei *loci* si rivela invece molto utile innanzitutto nel caso di consonanti sorde, per via del fatto che, ovviamente, nessuna formante può apparire nel loro caso, ma secondariamente anche nel caso di molte consonanti sonore che non presentino abbastanza energia armonica tale da permetterne la classificazione.

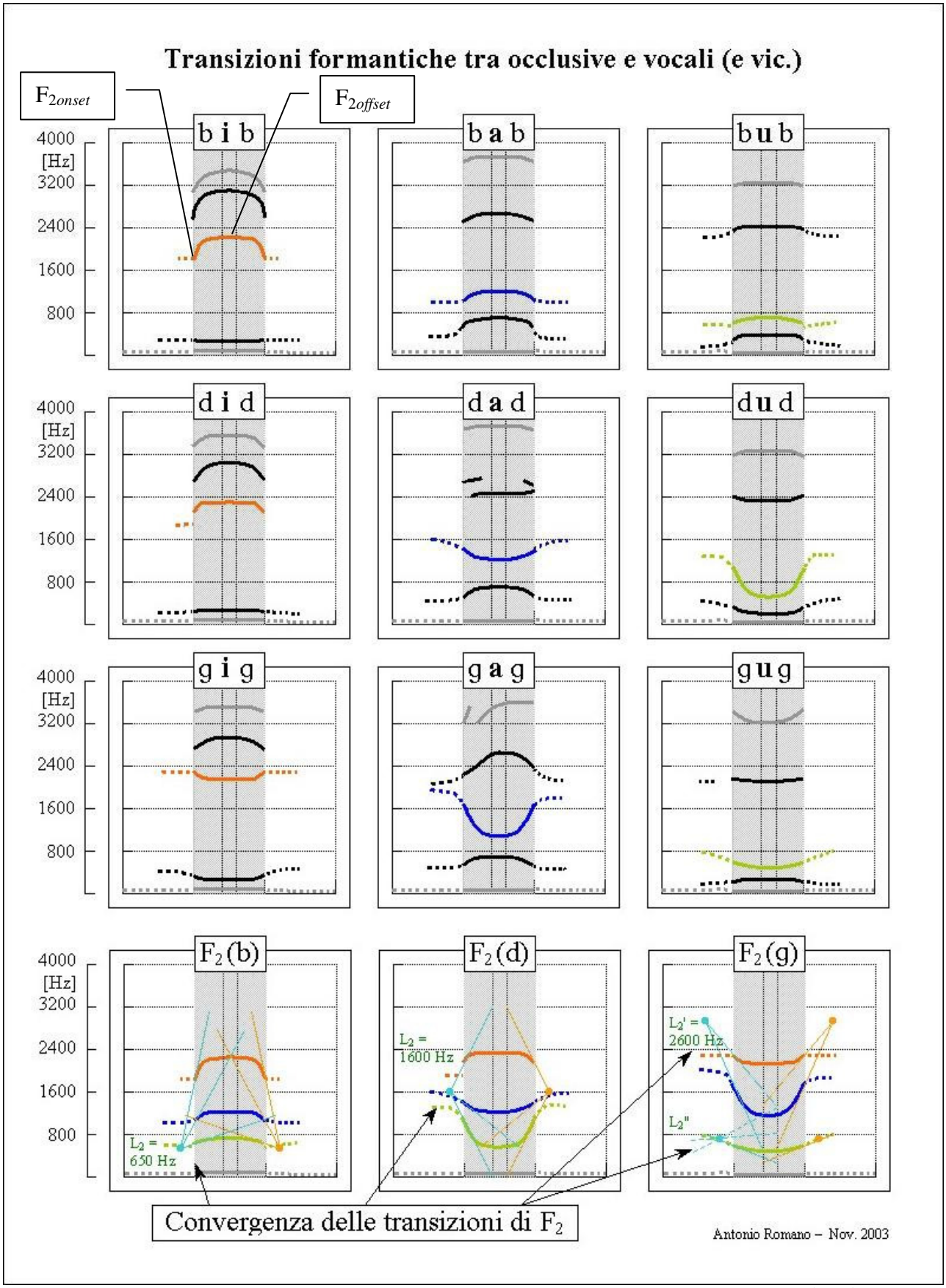

Figura 22. Schema raffigurante l'applicazione del metodo manuale di determinazione del *locus* L_2 di [b], [d] e [g]. Per ogni consonante si riportano in un unico grafico cumulativo (v. grafici in basso) le rette di deviazione di F_2 (prima e dopo le vocali [i], [a] e [u]). Il punto d'intersezione delle tre rette definisce una stima di L_2.

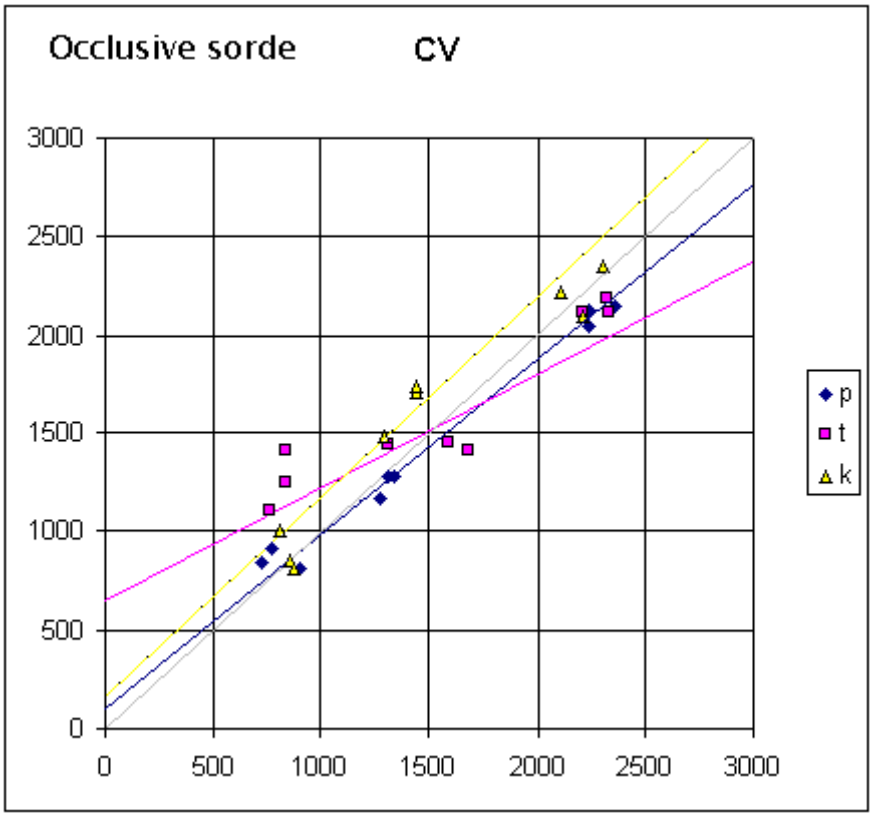

Figura 23. Rette di regressione per la determinazione di L_2 (equazione dei *loci*) per i suoni occlusivi [p], [t] e [k] dell'italiano.

Misurando F_{2onset} e $F_{2offset}$ (come indicato nel primo grafico in alto a sinistra di Fig. 22) si determinano coppie di valori che possono essere disposte in grafici come quello di Fig. 23. In questo in particolare sono state riportate le misure eseguite sulle vocali presenti in tre ripetizioni di logatomi *CVC* (*pip*, *tit*, *kik*, *pap*, *tat*, *kak*, *pup*, *tut*, *kuk*).

In questo i valori di *m* e *q* (v. anche Fig. 24) sono quelli delle **rette di regressione lineare** che ottimizzano la presa in carico dei punti delle dispersioni di valori che si determinano da queste coppie di misure.

La pendenza di queste rette è fissata dal parametro angolare *m*, che – come si può osservare sul grafico –, assume valori raramente valori > 1 (cioè con angoli > 45°). I valori più bassi si ottengono per le alveolari, mentre le occlusive per le quali più si approssima a 1 sono di solito le velari (a meno che queste non siano soggette a palatalizzazione, v. dopo).

Nel caso delle coppie di valori riportate in Fig. 23 sono stati ricavati i seguenti valori di *m* e *q* (v. anche figura 24):

	m	*q*
p	0,85	151
t	0,60	691
k	0,96	162

Un confronto tra i valori di parametro angolare *m* (corrispondente alla pendenza della retta di regressione) nell'equazione dei *loci* per articolazioni bilabiali, alveodentali e velari è presentato nella seguente tabella dove, insieme ai dati qui presentati per l'italiano, sono riassunti i risultati ottenuti in vari studi per altre lingue (cfr. Blumstein & Stevens 1979, Krull 1989, Sussman *et alii* 1993, Celdrán & Villalba 1995, Eek & Meister 1995)[31].

m	Luogo d'articolazione					
	bilabiali (VC/ CV)		alveodentali (VC/ CV)		velari (VC/ CV)	
Italiano (sorde)	**0.93**	**0.86**	**0.77**	**0.61**	**1.09**	**0.96**
Italiano (sonore)	**0.89**	**0.85**	**0.76**	**0.58**	**1.12**	**0.96**
Italiano (media)	***0.88***		***0.68***		***1.04***	
Spagnolo	0.83		0.58*		1	
Inglese	0.87		0.43		0.66**	
Estone	0.64	0.81	0.22	0.61	0.74	0.89
Thai	0.70		0.30***		0.95	
Arabo	0.77		0.25***		0.92	
Urdu	0.81		0.50		0.97	
MEDIA	*0.78*		*0.41*		*0.88*	

* Secondo Celdrán & Villalba (1995: 33), questo valore è più alto di quello dell'inglese perché il luogo d'articolazione di queste occlusive sarebbe dentale in Spagnolo e alveolare in Inglese (come suggerito anche dai valori più alti dell'Urdu che nel suo sistema ha suoni dentali in contrasto con suoni alveolari). Nel nostro caso si tratterebbe di articolazioni ancor più dentali.

** Questo valore è più basso forse a causa di una leggera palatalizzazione delle velari davanti a vocali di massima apertura (se sono escluse le posteriori).

*** Questi valori sembrerebbero indicare delle articolazioni praticamente postalveolari (sarebbe interessante verificare se le retroflesse hanno valori ancora più bassi).

I valori ottenuti per l'italiano, pur differenziandosi da quelli medi dello spagnolo, sono in linea con quelli rappresentati da singoli locutori considerati da Celdrán & Villalba (1995: 31).

Per valutare il loro uso in termini discriminatori occorre definire il loro intervallo di variazione e l'eventuale sovrapposizione di valori (v. dopo)[32].

[31] Riportiamo separatamente i valori ottenuti da Barillari *et alii* (1995), per voci maschili e femminili, probabilmente su una scala di rappresentazione invertita (F_2*offset* vs. F_2*onset*, anziché F_2*onset* vs. F_2*offset*). Questo fa sì che i valori non siano confrontabili con quelli degli altri studi.

	Luogo d'articolazione					
	bilabiali		alveodentali		velari	
	m	*q*	*m*	*q*	*m*	*q*
M	0.65	265	0.43	896	0.28	1643
F	0.66	307	0.44	1203	0.32	1793

I valori dei *loci* ottenuti nello studio di Barillari *et alii* (1995) (con i due metodi del punto di convergenza – *intersezione* – e della coincidenza dei valori di y e x nell'equazione dei *loci*, v. dopo), sono invece i seguenti:

	Locus di F_2 [Hz]					
	bilabiali		alveodentali		velari	
	intersezione	*equazione*	*intersezione*	*equazione*	*intersezione*	*equazione*
M	800	757	1650	1572	2280	2281
F	920	902	2100	2148	2760	2636

[32] In *VC* c'è una leggera sovrapposizione tra bilabiali e dentali (nel materiale analizzato la dentale implosiva è in effetti labializzata, cioè "suona" un po' come [pˀ/bˀ]). In *CV* gl'intervalli di variazione sono ben separati a eccezione di una sospetta corrispondenza tra il minimo delle velari e il massimo delle bilabiali (che potrebbero però essere discriminate da un decimale in più 0.889 vs. 0.888).

m	Luogo d'articolazione					
	bilabiali (VC/ CV)		alveodentali (VC/ CV)		velari (VC/ CV)	
minimo	0.83	0.79	0.71	0.45	1.04	0.89
massimo	0.97	0.89	0.84	0.63	1.14	1.01

Un altro uso interessante che si può fare dell'equazione dei *loci* è quello in cui si valuta il valore dell'intersezione tra la retta di regressione e la retta a parametro angolare 1 cioè quella retta in cui i valori di *offset* e di *onset* coincidono. Quest'assunzione equivale a porre $y = x$ nell'equazione e a determinare quel valore di x che soddisfa la condizione. Questo valore dà un'indicazione sul *locus* formantico delle transizioni descritte dalla retta.

Confrontando la media dei valori ottenuti per le occlusive sorde e sonore, distinguendo tra VC e CV, si sono ottenuti i seguenti valori:

Loci [Hz]	VC	CV	*medio*
L_2 (bilabiale)	585	551	*568*
L_2 (dentale)	1834	1890	*1862*
L_2 (velare)	2540	2442	*2491*

Questi valori corrispondono, con buona approssimazione, ai valori prototipici riportati solitamente in letteratura (v. Delattre *et alii* 1952, Giannini & Pettorino 1992) e cioè rispettivamente: 700 Hz, 1800 Hz, 3000 Hz.

I corrispondenti valori, ricavati manualmente, determinando i punti di intersezione dei prolungamenti delle transizioni hanno invece portato alle seguenti tabelle:

Occlusive sorde

Loci [Hz]	VC	CV	*media*
L_2 (bilabiale)	1000	550	*775*
L_2 (dentale)	1700	1400	*1550*
L_2 (velare)	2800	2300	*2550*

Occlusive sonore

Loci [Hz]	VC	CV	*media*
L_2 (bilabiale)	900	380	*640*
L_2 (dentale)	1850	1900	*1875*
L_2 (velare)	3050	2300	*2675*

I *loci* medi così determinati sono quindi nello stesso ordine di grandezza di quelli stimati col ricorso all'equazione dei *loci* riassunti nella seguente tabella:

Loci medi [Hz]	*intersezione*	*equazione*
L_2 (bilabiale)	*672*	*568*
L_2 (dentale)	*1706*	*1862*
L_2 (velare)	*2521*	*2491*

Tornando al tema dello sfruttamento di queste informazioni per un'analisi discriminante facciamo riferimento a Sussman *et alii* (1993) e ricorriamo a un altro metodo di presentazione dei dati.

Una discriminazione perfetta non sembra infatti possibile sulla base di un solo parametro, mentre invece, tenendo conto di entrambi i parametri *m* e *q* si ottiene una distribuzione di valori come quella nel diagramma in figura 24 che, ad esempio, consente di individuare delle aree distinte per i tre luoghi bilabiale, alveodentale e velare.

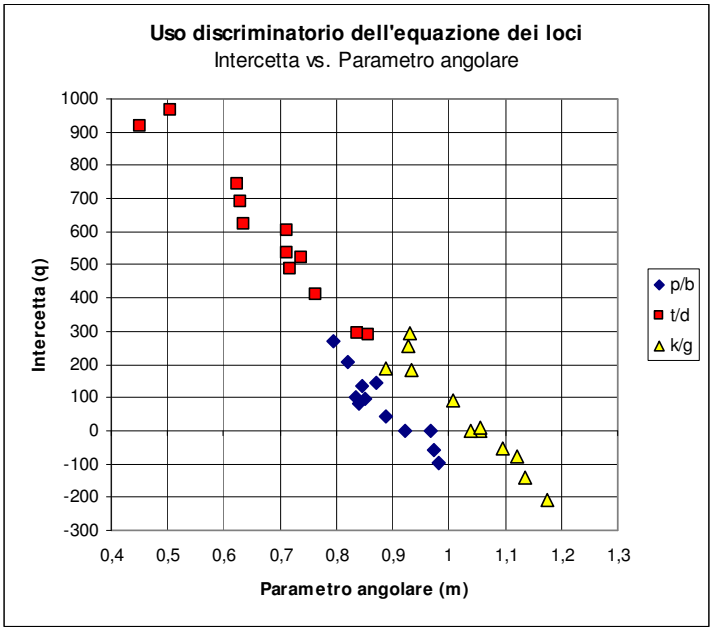

Figura 24. Aree di dispersione definite dai parametri *m* e *q* per i tre luoghi d'articolazione consonantici bilabiale, alveodentale e velare (consonanti sorde e sonore).

Così come per i suoni vocalici, le variabili acustiche F_1 e F_2 (e F_3) suggeriscono una rappresentazione grafica che consenta di riprodurre la disposizione dei diversi timbri nello spazio articolatorio, nel caso dei suoni occlusivi sono questi due parametri (*m* e *q*), unitamente agl'indici spettrali dell'esplosione, a garantire la separazione acustica degli effetti di distinte articolazioni (cfr. Romano *et alii* 2005).

IV.10. L'analisi dei suoni costrittivi

Anche per l'analisi dei suoni costrittivi, quelli cioè che sono costrittivi sul piano articolatorio e si presentano come dei rumori di frizione su quello acustico, lo strumento d'osservazione principale può essere costituito dallo spettrogramma.

Un esempio di confronto tra varie possibili configurazioni acustiche è proposto in figura 25 (tratta da De Sio & Romano 2003)[33].

Nel caso di suoni particolarmente stabili, è però invalso anche un frequente riferimento alla sezione spettrale (*FFT* o *LPC*), il cui profilo si rivela particolarmente utile per apprezzare meglio le associazioni tra frequenze ed energia (si vedano i numerosi lavori citati in bibliografia).

In Italia i contributi più recenti allo studio acustico delle fricative sono quelli di De Sio & Romano (2003) e di Sorianello (2003, 2004).

Rifacendosi ai principali manuali di fonetica acustica, Sorianello (2004) riassume le caratteristiche spettrografiche dei suoni costrittivi descrivendoli come dei segnali aperiodici piuttosto intensi, la cui distribuzione energetica lungo l'asse frequenziale è variamente distribuita a seconda di vari fattori.

Tale distribuzione dà luogo a una configurazione spettrale di rumore condizionata in realtà da molteplici fattori, tra cui il punto di articolazione del segmento, il grado di chiusura diaframmatica che si stabilisce tra gli organi articolatori, il livello di pressione raggiunto dall'aria che fuoriesce e, infine, la forma della cavità di risonanza (pneumatoceloma).

Oltre che essere funzione del punto di articolazione, quest'ultima può essere infatti fortemente condizionata dalla superficie del contatto (quindi dalle regioni articolatorie realmente coinvolte, si veda Ladefoged & Maddieson 1996: 160), dal grado di sulcalizzazione della lingua, nonché dalla concomitanza di diversi gradi di labializzazione e velarizzazione.

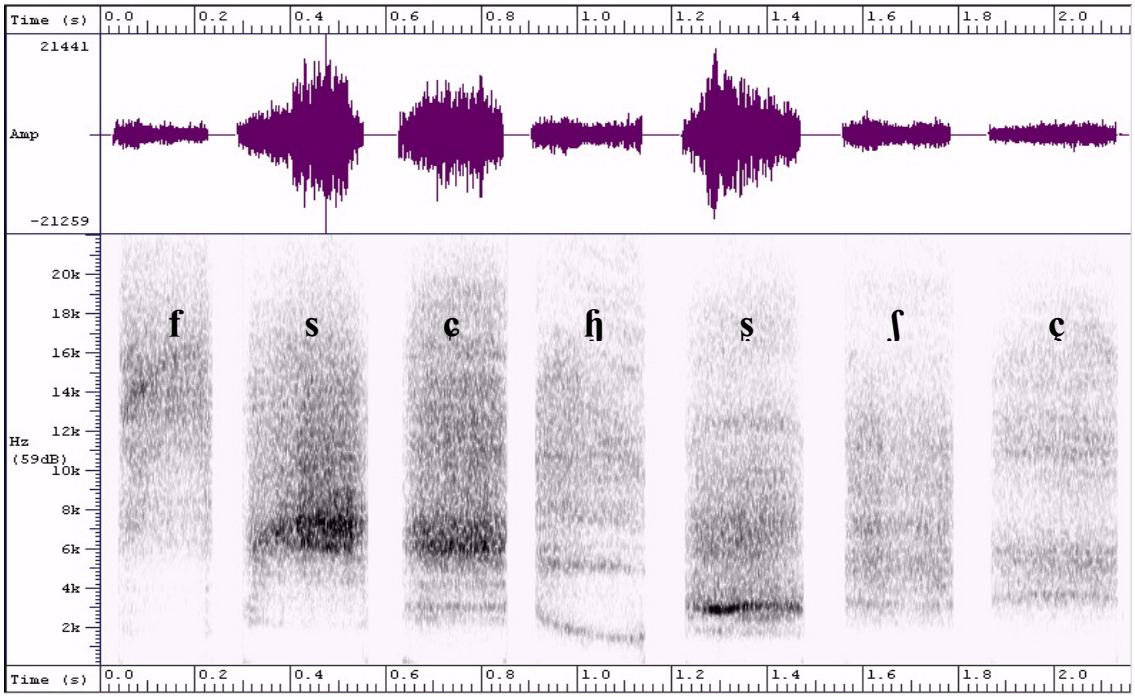

Figura 25. Oscillogramma e spettrogramma per le 7 distinte fricative possibili in svedese avulse dal loro contesto originario (i)C(i). Si noti la diversa distribuzione dell'energia e la possibile classificazione in due gruppi "energetici" (in particolare la seconda, la terza e la quinta si presentano nettamente più forti delle altre quattro). Si noti inoltre la variabilità temporale della concentrazione di energia per la prima, la seconda e la quarta [riprodotta da De Sio & Romano 2003].

[33] Lo spettrogramma è ottenuto con il programma *WASP/SFS*. Si noti che in questo caso, trattandosi di uno studio in cui non si esclude *a priori* la possibilità che questi suoni presentino componenti interessanti anche a frequenza più alta di 8 kHz, la frequenza di campionamento adottata è stata di 44,1 kHz (che garantisce la stessa qualità dei CD musicali).

Sulla base di questi fattori, lo spettro di una fricativa si manifesta di solito come quello di un rumore 'colorato' dalle caratteristiche spesso dinamicamente variabili nel corso dell'articolazione anche in funzione del contesto segmentale[34].

Come ci ricorda Sorianello (2003), riguardo all'italiano i rilievi sperimentali disponibili nella letteratura specialistica sull'argomento, pur nella loro varietà, sono abbastanza omogenei. La labiodentale /f/ mostra una frizione di debole intensità che inizia da 1500-2000 Hz fino a ricoprire le frequenze più alte. Il rumore di /s/, che è invece molto più intenso, si estende da 4000-5000 Hz in su. Una forte energia concentrata tra 2000 e 4000 Hz caratterizza infine /ʃ/[35].

Anche autori che ricorrono alla valutazione congiunta di *FFT* e *LPC*, per l'analisi acustica delle fricative in generale, concordano con queste descrizioni. Kent & Read (1992: 125) dànno ad esempio per lo spettro di [s] un massimo a 5000 Hz e per quello di [ʃ] uno a 2500 Hz, mentre Ladefoged (2003: 156), che dà l'indicazione dei centroidi degli spettri, riporta 7,1 kHz per [s] e 5,1 kHz per [ʃ]).

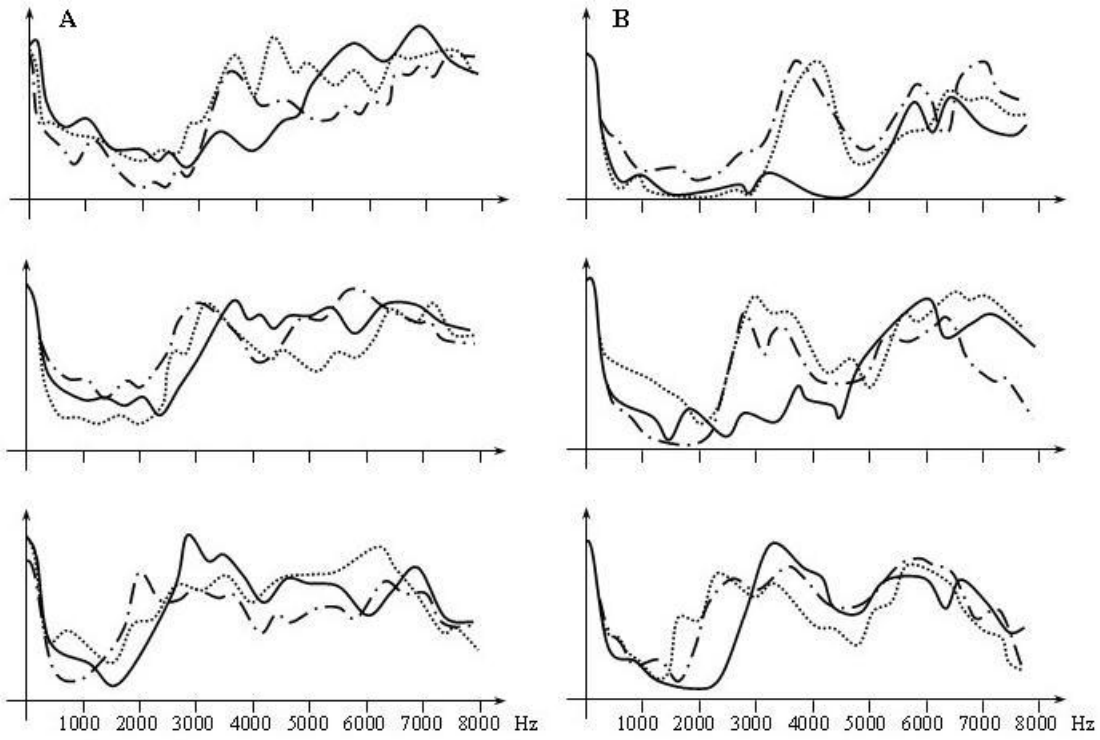

Fig. 26. Spettri delle fricative svedesi /s/, /ɕ/, /ʂ/ prima di /iː/ (linea continua), /yː/ (linea tratteggiata), /uː/ (linea tratto - punto) prodotte da due locutori (A, prima colonna, e B, seconda colonna) [adattati da Lindblad 1980].

[34] Per la definizione di 'formanti di rumore', già ripresa nel lavoro sulle fricative svedesi da De Sio & Romano (2003), ci possiamo rifare ai lavori di Badin (1991) e di Shadle *et alii* (1991) che discutono gli spettri medi di potenza per diversi luoghi di costrizione. Le tecniche d'analisi spettrale a breve termine sono invece esemplificate in diversi manuali cui rinviamo per maggiori dettagli. Kent & Read (1992: 125), ad esempio, ricorrono a *FFT* e *LPC*. La tecnica dell'avanzamento delle finestre d'analisi è ottimamente illustrata da Ladefoged (2003: 155-156) che presenta un'analisi congiunta *FFT* e *LPC* con l'indicazione dei centroidi degli spettri.

[35] In uno studio condotto sulle caratteristiche acustiche delle fricative dell'italiano, Vagges *et alii* (1978) prendono in esame un corpus di parole isolate lette da dieci giovani locutori fiorentini. Tra i parametri acustici analizzati troviamo il limite frequenziale inferiore del rumore presente nello spettro, valutato considerando il solo contesto intervocalico e limitatamente alla posizione lessicale accentata. I valori schematici riportati per questo parametro indicano 2200 Hz per /f/, 4300 Hz per /s/ (ma 5080 Hz per /z/) e 1925 Hz per /ʃ/. Presentando un caso dinamico di labializzazione e facendone osservare gli effetti, Giannini & Pettorino (1992) rilevano come nei suoni fricativi le concentrazioni di rumore decrescano in frequenza con l'aumentare della lunghezza della cavità anteriore al punto di costrizione del suono. Per le tre costrittive sorde dell'italiano la concentrazione più importante si colloca a circa 12 kHz per /f/, a 7 kHz per /s/ e a 3 kHz per /ʃ/. I dati di Sorianello (2003: 35) dànno infine, per tre diversi locutori fiorentini, valori di massima concentrazione energetica tra 6643 e 8427 Hz per /f/, tra 4749 e 6212 Hz per /s/, tra 3062 e 4933 Hz per /ʃ/.

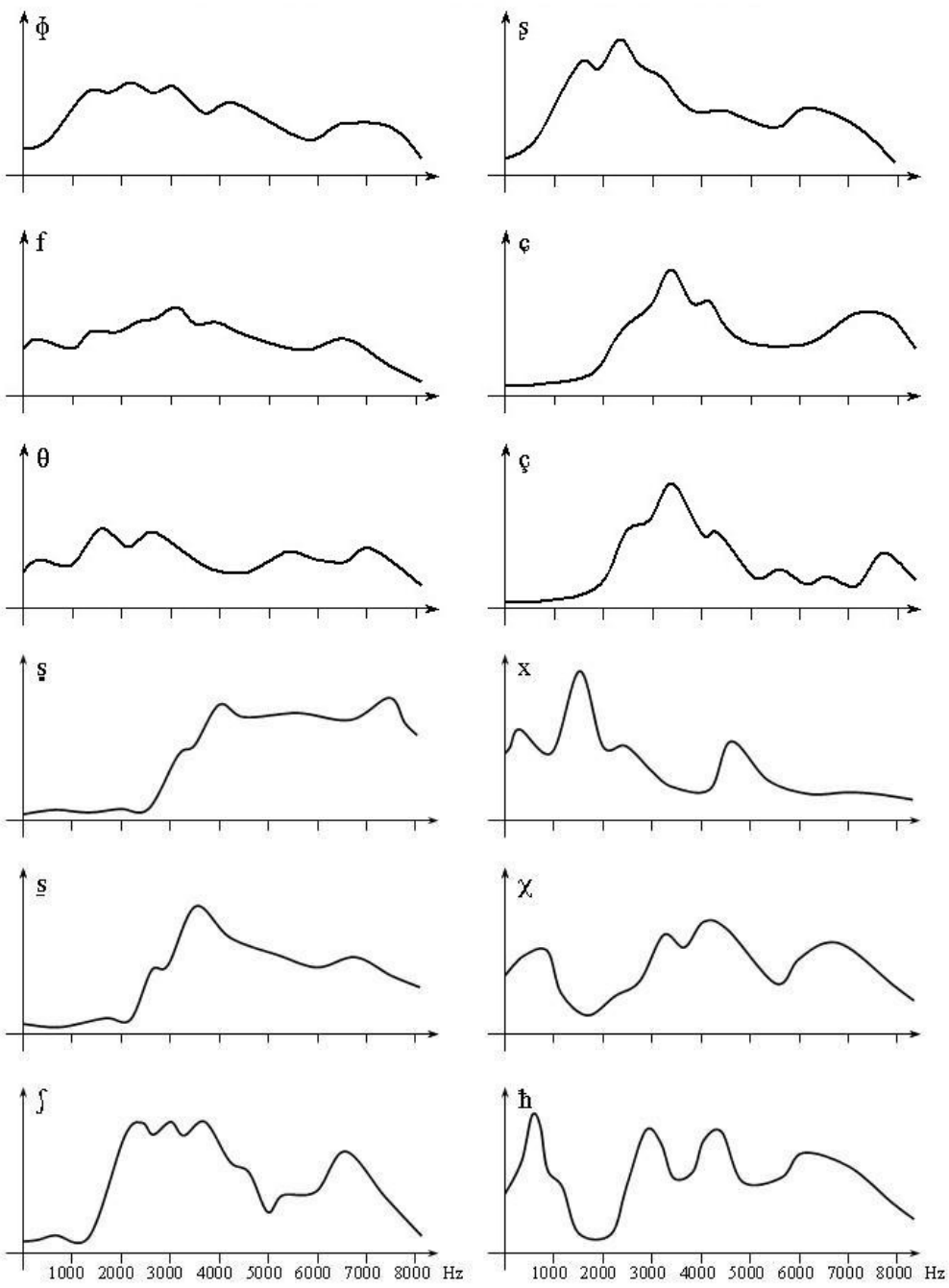

Fig. 27. Profili spettrali delle fricative [ɸ, f, θ, s̪, s̲, ʃ, ʂ, ɕ, ç, x, ħ] [selezionati e adattati da Jassem 1968].

Se però in questi lavori, che si concentrano su una selezione ridotta e semplificata di suoni fricativi, ritroviamo una buona suddivisione dello spazio spettrale, le cose non si presentano altrettanto semplici quando si voglia definire un quadro completo tenendo conto dei vari fattori di cui sopra e delle loro conseguenze sul piano acustico (oltre a Shadle *et alii* 1991, sulla necessità di un approccio fonetico si veda ad es. anche Evers *et alii* 1998).

Shadle *et alii* (1991) discutono gli spettri medi di potenza di una selezione di tre luoghi di costrizione per due locutori che ne distinguono volontariamente 7: per [s] i grafici indicano un aumento di energia a partire da 5,5-6 kHz; per [ʃ] l'aumento di energia si presenta a partire da 2,5 kHz ma decresce dopo un ultimo massimo ancora significativo a 5-6,5 kHz; per [ç] i grafici indicano invece un primo massimo a 3,5-4 kHz e una concentrazione che si mantiene alta fino a ca. 6 kHz. Queste ricerche confermano tuttavia una certa dipendenza dal locutore, quella stessa che era stata messa in evidenza da Lindblad (1980; v. Fig. 26; cfr. anche Faber 1991).

Dati ancora più completi sono quelli pubblicati da Badin (1991) che, nonostante la variabilità riscontrata, riassumono infine: per le labiodentali, spettri medi sostanzialmente piatti (con massimi totalmente differenti per i due locutori analizzati: 2000 e 7500 Hz); per le (inter-)dentali, profili tendenzialmente piatti-ascendenti (fino a 10 kHz) con deboli massimi variamente distribuiti; per le alveolari, spettri con poca energia alle basse frequenze e un salto energetico tra i 5 e i 7 kHz; per le post-alveolari, profili con diverse concentrazioni tra i 2 e i 5 kHz e energia decrescente alle alte frequenze; per le palatali, massimi maggiormente concentrati intorno a 3 e (in misura incostante) 5 kHz; per le velari, profili discendenti con picchi energetici tra i 1,2-1,5 kHz, uno tra 3,5 e 5 kHz e vari altri d'importanza relativamente limitata.

Tra le numerose altre fonti che hanno proposto delle schematizzazioni 'statiche' citiamo in particolare Jassem (1962) e Lindblad (1980) (i cui dati sono riprodotti in Ladefoged & Maddieson 1996) e Jassem (1968). Riportiamo in Figura 27 alcuni spettri di fricative selezionati e adattati da Jassem (1968) i quali permettono di avere un insieme di riferimento.

Vediamo ora il quadro proposto da De Sio & Romano (2003). Nonostante uno degli obiettivi degli autori fosse quello di mostrare una certa variabilità dinamica di questi suoni, presentando i risultati di un monitoraggio temporale degl'indici di frizione, De Sio & Romano (2003: 326) propongono un grafico riassuntivo sulle proprietà spettrali di 7 fricative in tre diversi contesti vocalici (V_1CV_2, con V_1=V_2). I diagrammi riportati in quel caso sono qui riprodotti (v. Fig. 28) appiattendo la dimensione di variazione temporale e sottolineando la presenza di configurazioni radicalmente diverse per alcuni fonemi (in virtù di una variazione che, oltre ad essere indotta dalla co-articolazione, sembra soprattutto attribuibile a una diversa selezione d'allofoni)[36].

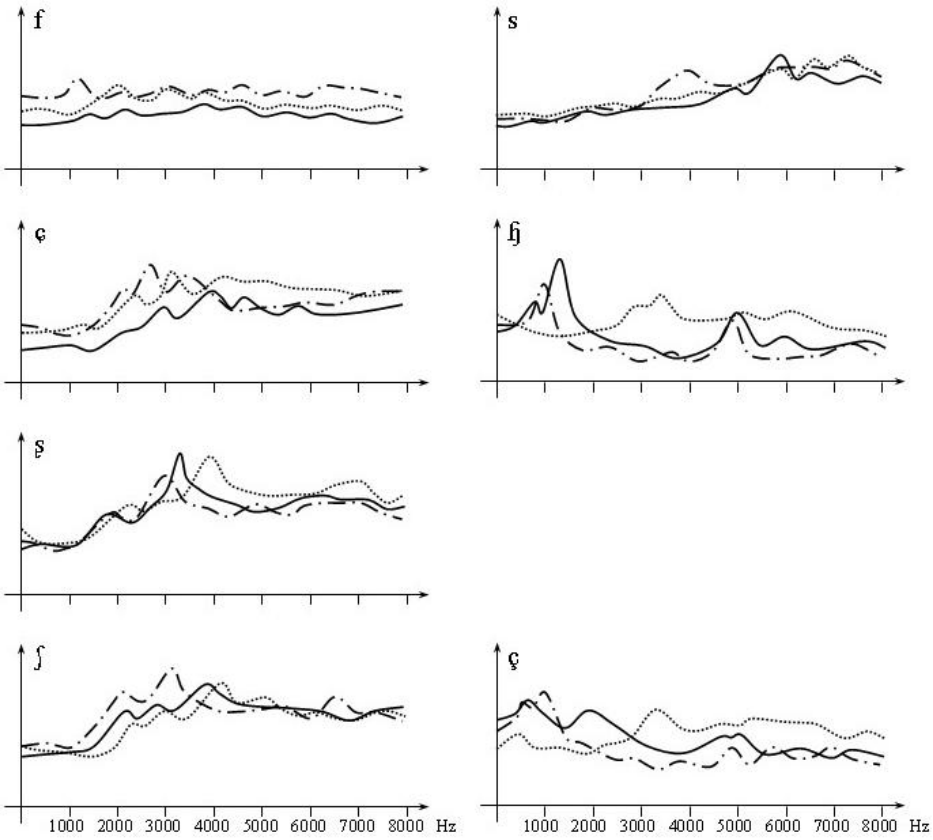

Fig. 28. Profili spettrali delle fricative [f, s, ʃ, ʂ, ɕ, ç, ʃj] in contesto di [a] (linea continua), [i] (linea tratteggiata), [u] (linea tratto - punto) prodotte da una locutrice svedese [adattati da De Sio & Romano 2003].

[36] Oltre a sottolineare ancora una volta l'incidenza di idiosincrasie individuali, questo studio ha permesso di rilevare un fattore decisivo di variabilità indotto dalla variazione contestuale. Le notevoli differenze tra spettri di suoni fricativi classificati 'emicamente' come realizzazioni dello stesso fonema sottolineano importanti oscillazioni nella pronuncia che portano all'accavallamento (e a volte allo scavalcamento) di *pattern* prossimi.

I grafici confermano un andamento piuttosto piatto dell'energia per [f] (con un particolare aumento generale d'energia nel caso di [u]) e un andamento ascendente con progressiva colorazione di [s] a 4-6 kHz. Tralasciando le considerevoli differenze causate, evidentemente, dalle varianti combinatorie realizzate nei casi di [ç] e /ʃj/ (oscillanti tra tassofoni con dominanza di palatalità o di labialità-velarità), vengono messe in rilievo le formanti di rumore caratteristiche di [ʃ], [s̟] e [ç].

Relativamente stabili e costanti, tre componenti equidistanti che si presentano a circa 2, 3 e 4 kHz per [ʃ], si riducono essenzialmente a due per [s̟] (una tra i 1800 e i 2200 Hz e l'altra più mobile e nettamente più energetica tra i 3 e i 4 kHz), mentre diventano più variabili per [ç] che, oltre a presentare una prima formante più precaria a 1,8-2 kHz, intensifica i contributi a frequenza più alta con una formante a 2,8-3 kHz e un'altra a 3,5-4 kHz (che ricordano la F_3 e la F_4 di [i]).