LA LETTURA LABIALE: DATI SPERIMENTALI E PROBLEMI TEORICI

E. Magno Caldognetto, K. Vagges, P. Cosi, F. E. Ferrero
Centro di Studio per le Ricerche di Fonetica - C.N.R. - Padova

RIASSUNTO

Le attuali ricerche sulla lettura labiale, partendo dalle tradizionali quantificazioni dell'intelligibilità dei movimenti articolatori visibili, si concentrano sull'indagine delle caratteristiche spaziali e temporali dei movimenti di labbra e mandibola per individuare gli indici visibili significativi per l'identificazione. L'analisi confrontativa dei segnali articolatori e acustici permette successivamente di studiare l'integrazione dei due flussi di informazione nella percezione bimodale, fenomeno fondamentale nella comunicazione linguistica orale faccia-a-faccia non solo per soggetti patologici (p. es. ipoacusici), ma anche per soggetti normali in situazioni di degradazione di segnale acustico. Sulla base di tali conoscenze sarà possibile mettere a punto, in particolare per ipoacusici, dispositivi per la comunicazione multimodale e per la riabilitazione.

PAROLE CHIAVE

Lettura labiale, Percezione bimodale, Intelligibilità visiva, Intelligibilità uditiva, Visemi.

1. La lettura labiale costituisce per i soggetti ipoacusici una importante, se non unica, via di accesso all'identificazione del messaggio verbale prodotto dai loro interlocutori in una interazione faccia-a-faccia. È quindi importante conoscere caratteristiche e limiti di questo fenomeno percettivo per rendere conto del sinergismo uditivo-visivo che ha luogo nella comunicazione bimodale anche per i soggetti normali e per sfruttarlo a scopo riabilitativo.

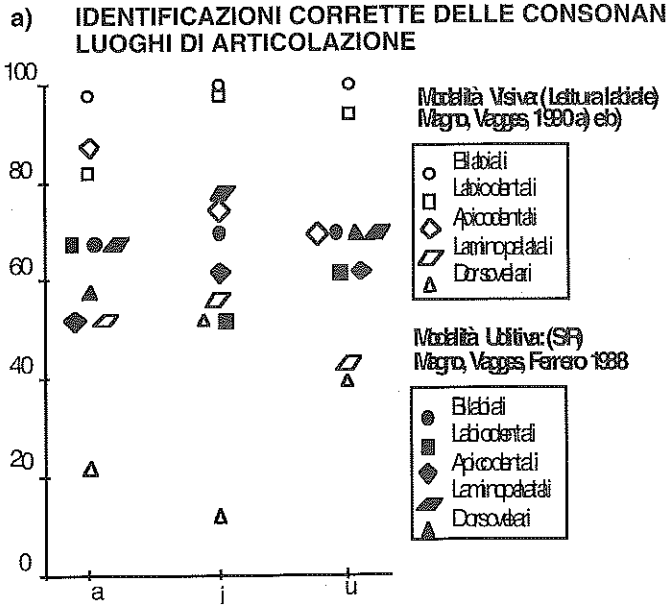
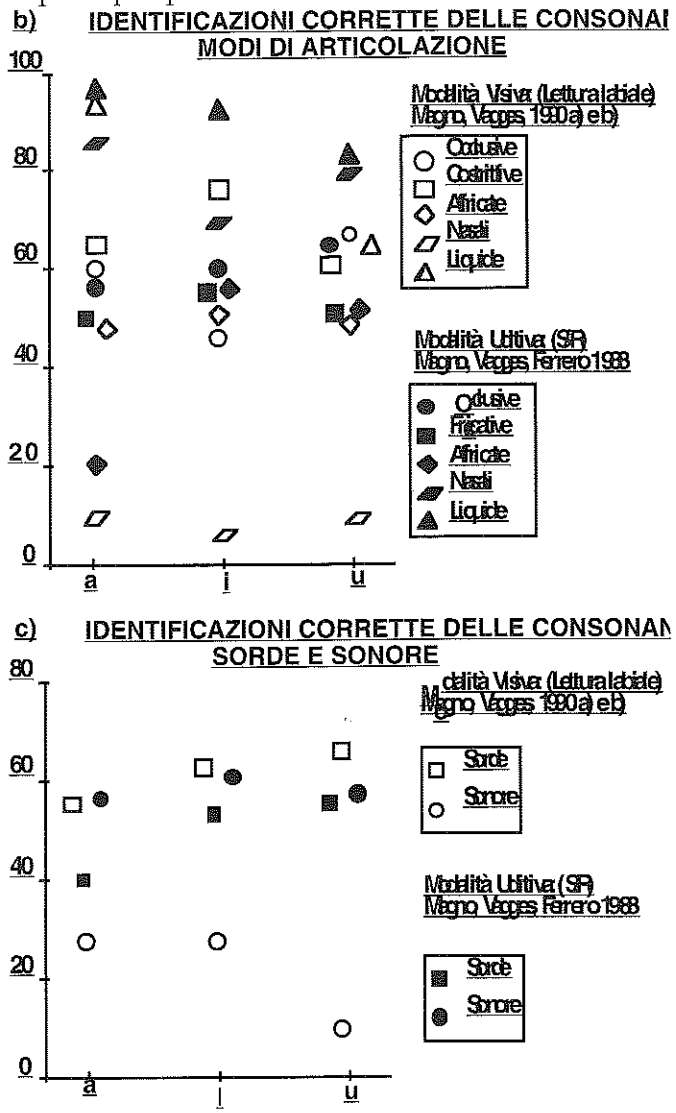


Fig. 1a. Risultati di test di intelligibilità visiva e uditiva per le consonanti dell'italiano relativi ai luoghi di articolazione.

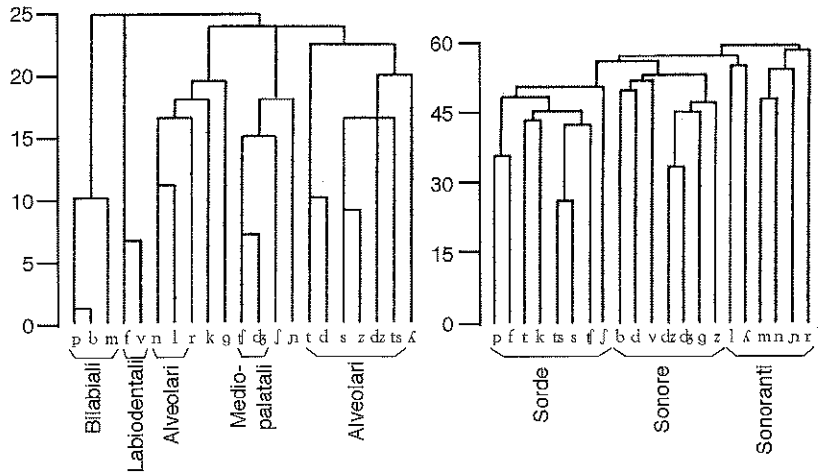
2. Il primo stadio delle ricerche sulla lettura labiale è costituito dai test di intelligibilità eseguiti con stimoli naturali, sebbene in questo caso il temine "lettura labiale" sia riduttivo perché vengono sfruttati per il riconoscimento delle unità fonologiche anche gli spostamenti della mandibola, la visibilità dei denti e/o della punta della lingua e possibili concomitanti facciali dell'articolazione, quali il rigonfiamento delle guance [1] [2]. I grafici di Fig. 1a,b,c riportano i risultati di un test di intelligibilità visiva relativo alle 21 consonanti dell'italiano in strutture CV con V=/a,i,u/ [3] [4]. Secondo le aspettative l'intelligibilità globale dei movimenti visibili è bassa (28% di identificazione corretta); la trasmis-

sione di informazione fonologica è elevata solo per bilabiali [p,b] e labiodentali [f,v], e i riconoscimenti corretti diminuiscono gradualmente passando dai luoghi più anteriori a quelli più posteriori.



Figg. 1b e 1c. Risultati di test di intelligibilità visiva e uditiva per le consonanti dell'italiano relativi b) ai modi di articolazione e c) all'opposizione di sonorità.

Per quanto riguarda i modi, viene penalizzato il riconoscimento delle consonanti nasali e di tutte le consonanti sonore poiché né i movimenti del velo né quelli delle corde vocali sono visibili. Queste tendenze non variano per effetto della coarticolazione con varie vocali, anche se la percentuale media dei riconoscimenti corretti risulta migliore per [a] con il 31% rispetto al 28% per [i] e al 25% per [u]. Il confronto con i risultati di un test di intelligibilità uditiva relativo alle stesse consonanti negli stessi contesti in condizioni di mascheramento del segnale con rumore a varia intensità [5] [6] e [7] riportati sempre in Fig. 1a, b, c, dimostra che nell'uditivo viene favorita l'identificazione delle consonanti nasali, delle liquide (laterali e vibrante) e delle consonanti sonore, cioè delle categorie peggio riconosciute visivamente. L'analisi degli errori di identificazione visiva, cioè delle confusioni tra consonanti documentate dagli stessi test, evidenzia (Fig. 2a) raggruppamenti di consonanti giudicate simili tra loro [3], [4]. Ciascuno di tali raggruppamenti individua un *visema*, cioè una categoria percettiva visiva cui corrisponde una classe di movimenti articolatori che trasmettono uno stesso significato: per questo le consonanti che costituiscono un visema vengono dette *omofene*.



Figg. 2a e 2b. Analisi in cluster delle confusioni a) visive e b) uditive tra le consonanti dell'italiano.

Come risulta dal dendrogramma di Fig. 2a, i visemi tendono a corrispondere a raggruppamenti definiti in base al luogo di articolazione.

L'analisi parallela delle confusioni tra le consonanti in condizione di mascheramento del segnale acustico, riportata nel cluster di Fig. 2b, individua tre aggregazioni principali, corrispondenti agli insiemi delle consonanti sorde, sonore e sonoranti, cioè a gruppi di consonanti individuate da rilevanti differenze tipologiche spettrali, quali presenza di segnale aperiodico, periodico o di struttura formantica, che implicano modi di articolazione diversi e presenza o assenza di vibrazioni delle corde vocali [6], [7].

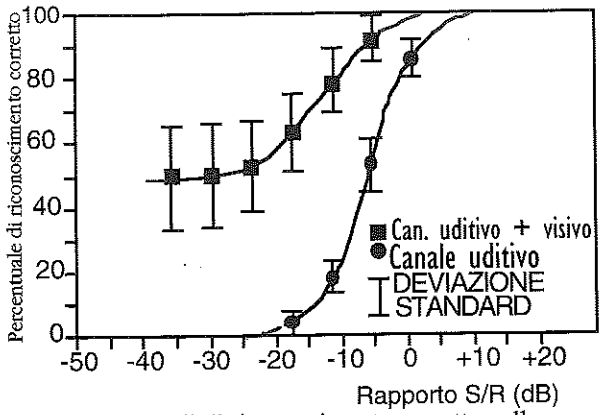


Fig. 3. Percentuali di riconoscimento corretto nella trasmissione di informazione acustica degradata e di trasmissione associata uditivo-visiva secondo Erber 1969.

Sulla base di questi risultati si è parlato di specializzazione del canale visivo per la trasmissione di informazione relativa ai luoghi di articolazione e del canale uditivo per i modi.

Tale complementarità è alla base del cosiddetto *sinergismo bimodale uditivo-visivo*, cioè ai miglioramenti dell'intelligibilità (cfr. Fig. 3) quando ad un segnale verbale degradato viene aggiunta la corrispondente, coerente, informazione visiva [8] e [9]. Poiché l'integrazione visivo-uditiva non ha effetti additivi ma moltiplicativi [10], la trasmissione bimodale del parlato, che è la condizione naturale dell'interazione comunicativa faccia-a-faccia, è anche la condizione più favorevole per i soggetti ipoacusici o portatori di impianto cocleare.

3. Il secondo livello delle ricerche sulla lettura labiale è costituito dall'analisi delle caratteristiche spaziali e temporali dei movimenti articolatori visibili al fine di isolare i possibili indici rilevanti per il riconoscimento. A questo scopo, per ovviare al problema della complessità della forma di labbra e mandibola, viene analizzata di solito la cinematica di punti rappresentativi di tali organi (per metodi diversi di rilevamento [11], [12], [13], [14] e [15]).

Il sistema ELITE da noi utilizzato esegue un'analisi automatica tridimensionale degli spostamenti di markers catarifrangenti incollati sui punti centrali del bordo del labbro superiore (LS) e inferiore (LI) sugli angoli delle labbra e sul centro del mento (Fig. 4), mentre altri markers costituiscono punti di riferimento per l'elaborazione dei dati, e permette l'acquisizione contemporanea del segnale verbale coprodotto [16], [17]. [18].

Partendo dai movimenti di LS, LI e M il dispositivo permette anche di calcolare una serie di parametri fonetici correlati a tratti distintivi fonologici:

- l'altezza dell'apertura mandibolare (AM), correlata al tratto alto/basso o chiuso/aperto;
- l'altezza dell'apertura labiale (AL), correlata anch'essa al tratto alto/basso;
- la larghezza dell'apertura labiale (LL), correlata al tratto arrotondato/appiattito;
- la protrusione del LS (PS) e LI (PI), correlata al tratto protruso/retratto (per la spiegazione dettagliata e discussione [16], [17] e [18]).

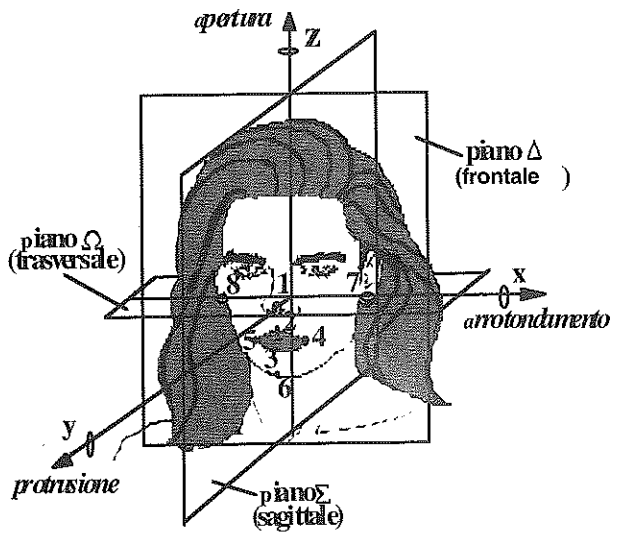


Fig. 4. Disposizione dei markers nel sistema ELITE per l'analisi tridimensionale dei movimenti labiali e mandibolari

Come hanno dimostrato le analisi statistiche della varianza, le tre dimensioni più significative sono risultate AM, PI e LL: AM è il parametro che permette di suddividere le vocali toniche dell'italiano secondo 4 gradi di apertura, PI individua anch'essa 4 gruppi, caratterizzati da 2 gradi di protrusione e 2 gradi di ritrazione, mentre LL divide le vocali in due gruppi, quello delle vocali arrotondate e quello delle vocali appiattite.

I cluster riportati in Fig. 5 illustrano questi risultati e contemporaneamente evidenziano, sulla base dei dati articolatori, i possibili visemi vocalici.

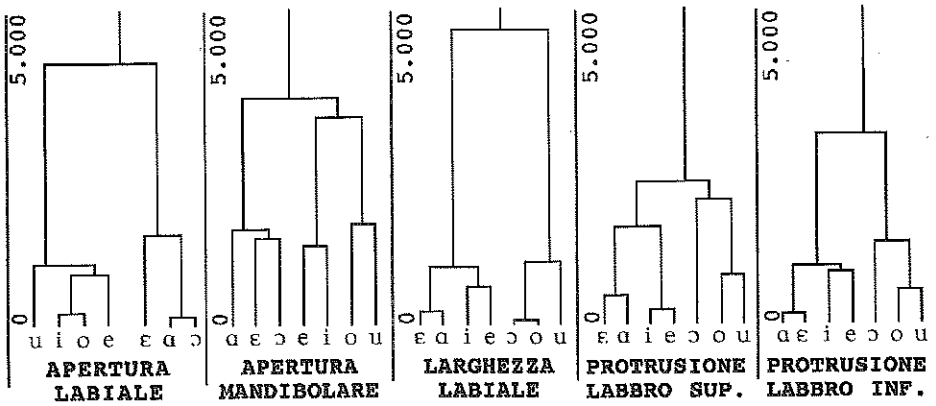


Fig. 5. Cluster gerarchici per le vocali toniche dell'italiano secondo cinque parametri articolatori (vedi testo)

In Fig. 6 viene proposta la caratterizzazione bidimensionale delle vocali toniche dell'italiano secondo questi tre parametri più rilevanti (per sistemi vocalici di lingue diverse: [19], [11], [12], [20]).

Anche per le consonanti /p, b, f, t, d, s, ʃ/, prodotte in contesto simmetrico [a], sono state elaborate le configurazioni visive tridimensionali (Fig. 7).

Si devono distinguere i casi in cui le labbra sono articolatori primari, come per le occlusive bilabiali /p/ e /b/ o la costrittiva labiodentale /f/, dalle consonanti /t, d, s, ʃ/in cui l'articolatore primario è la lingua e i movimenti delle labbra dipendono quindi dal movimento della mandibola, che coopera con la lingua alla realizzazione delle costrizioni del cavo orale, oltre che dalle caratteristiche delle vocali contestuali (per le relazioni tra movimenti della lingua nella realizzazione di /t/ e andamento dell'apertura labiale [21]). A parità di contesto vocalico, le configurazioni labiali delle consonanti da noi analizzate si differenziano principalmente per gradi di AL: si passa infatti da valori negativi di apertura labiale dovuti alla compressione labiale per l'occlusione di /p/ e /b/, a valori minimi di apertura per la costrizione di /f/ e a valori che aumentano via

via che il luogo di articolazione della consonante si posteriorizza. Lungo la dimensione LL tutte le consonanti mostrano una tendenza all'appiattimento, mentre è interessante notare che la costrittiva mediopalatale /'/ risulta protrusa e può quindi considerarsi labializzata.

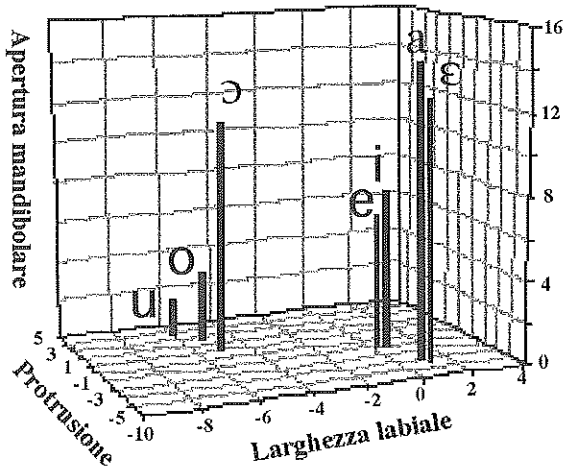


Fig. 6. Configurazione tridimensionale delle caratteristiche visive delle vocali toniche dell'italiano.

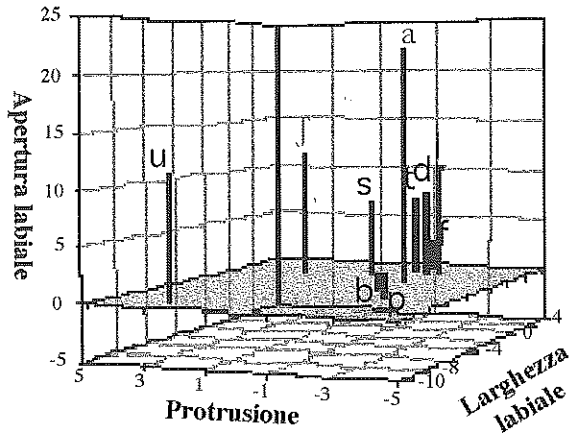


Fig. 7. Configurazione tridimensionale delle caratteristiche visive delle consonanti /p, b, f, t, d, s, ʃ/ .

Tanto i movimenti quanto le configurazioni labiali per le consonanti subiscono importanti modifiche al variare delle vocali contestuali, come viene esemplificato per /p/ in Fig. 8 a e b in cui vengono presentate le variazioni per /p/ in contesti simmetrici 'VCV in cui V = /a, i, u/.

Un altro fenomeno rilevante per la lettura labiale è quello dell'estensione della coarticolazione anticipatoria del parametro di arrotondamento labiale. Poiché è stato dimostrato che il movimento per l'arrotondamento di una vocale /u/ inizia in corrispondenza del primo elemento di un nesso triconsonantico come [istrù] [22], è evidente che il riconoscimento visivo delle 3 consonanti costituenti il nesso potrebbe risultare difficile, data la sovrapposizione dell'arrotondamento alle specifiche configurazioni illustrate precedentemente.

4. Il terzo livello delle ricerche sulla lettura labiale è costituito dallo studio delle regole dell'integrazione percettiva dei segnali visivi e uditivi per l'elaborazione di una teoria della percezione bimodale [2], [23], [24], [25] e [26]. In condizioni naturali, tra stimoli visivi e uditivi vi è coerenza nel senso che vi sono rapporti causali tra gesti articolatori, forme del condotto vocale e struttura del segnale acustico.

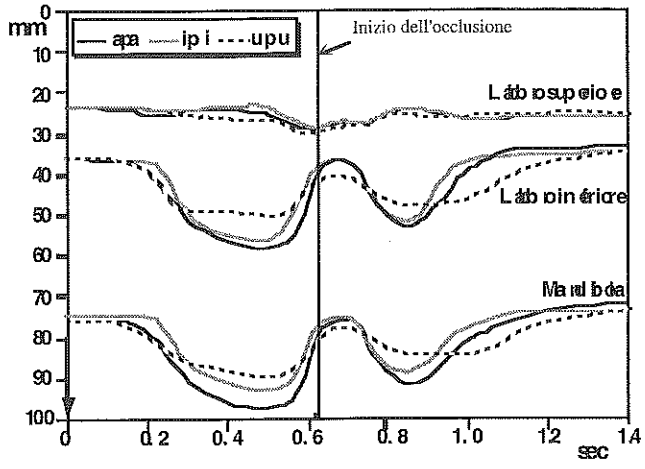


Fig. 8a. Modificazioni dei movimenti verticali di labbra e mandibola per /p/, indotte dalle vocali contestuali.

Per esempio, a gradi diversi di apertura labiale corrispondono, nella produzione delle vocali, posizioni frequenziali diverse della prima formante, a gradi di arrotondamento labiale graduali abbassamenti della seconda formante, a rapidi gesti di abbassamento o di innalzamento della mandibola rapide variazioni spettrali del segnale acustico. In questi casi di coerenza, qualora il segnale acustico venga danneggiato o ridotto, la lettura labiale può sostituire la percezione uditiva. Le ricerche sperimentali hanno anche

dimostrato che il segnale acustico e quello ottico possono però essere non isomorfici, come nel caso di porzioni iniziali o finali di gesti labiali e mandibolari all'inizio o alla fine di un enunciato ai quali non corrisponde produzione di segnale acustico oppure nel caso dei movimenti labiali che hanno luogo durante la fase di occlusione di consonanti quali /t/ o /d/ (Fig. 9).

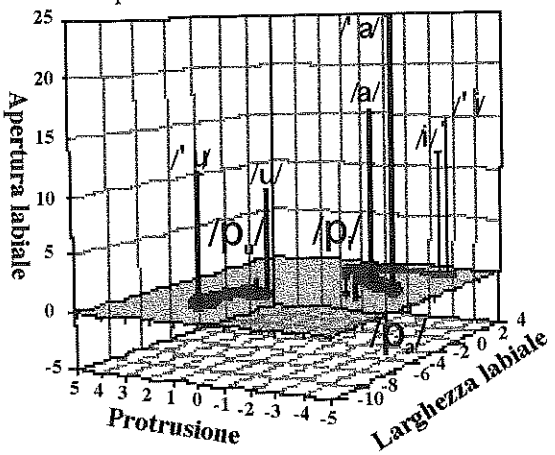


Fig. 8b. Modificazioni delle configurazioni visive tridimensionali per /p/, indotte dalle vocali contestuali.

Per questo, oltre all' integrazione di movimenti visibili e udibili, si deve tener conto anche di movimenti visibili e non udibili e di movimenti non visibili (come quelli del velo e delle corde vocali), ma udibili, cioè di casi in cui tra segnale uditivo e visivo si instaura una relazione complementare. Che l'integrazione tra questi segnali sia un'operazione percettiva complessa (senza tener conto qui delle variazioni imposte da fattori di angolazione di visibilità della faccia, di illuminazione, di visualizzazione della totalità o di una parte della faccia: [27], [26]) è dimostrato dall' "effetto McGurk" cioè dalle illusioni da fusione uditivo visiva che hanno luogo quando ad un soggetto vengono trasmessi uno stimolo visivo e uno uditivo non coerenti tra loro [28], [29], [2], [23], [24], [25]. Come risulta dallo schema seguente, l'effetto percettivo varia, a seconda delle

Stimolo visivo	stimolo uditivo	percetto
[ta]	[ma]	[na]
[ma]	[ta]	[pa]
[ga]	[ba]	[da]
[ba]	[ga]	[bga]

caratteritiche dei due stimoli trasmessi, da una vera e propria fusione a delle combinazioni tra i due.

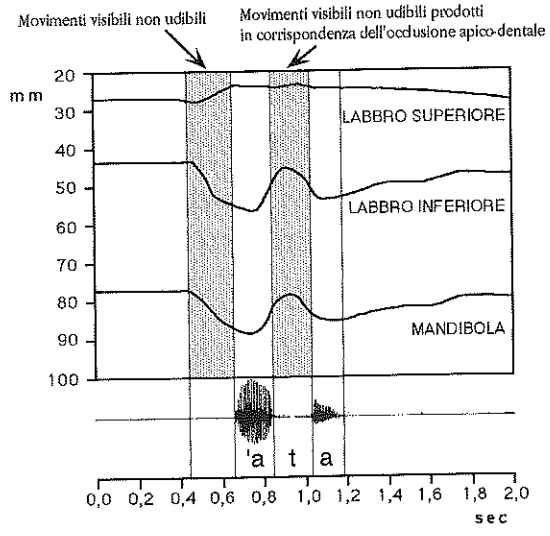


Fig. 9. Andamenti dei movimenti labiali e mandibolari per /t/ e segnale acustico corrispondente (vedi testo)

Nei primi due casi la fusione sembra aver luogo secondo la regola VPAM (Vision: Place; Audition: Mode) che prevede che lo stimolo visivo determini le caratteristiche di luogo del percetto finale, mentre lo stimolo uditivo vi contribuisce con le caratteristiche di modo, parallelamente a quanto riscontrato nei test di intelligibilità. Nel terzo e nel quarto caso il risultato percettivo è invece una combinazione più complessa che dipende dal grado di confondibilità o di ambiguità di uno dei due segnali, per esempio quando la consonante trasmessa visivamente è [-labiale], come [ga], e quindi risulta potenzialmente ambigua relativamente al luogo di articolazione [23], [24], [25].

Proprio perché i processi sottesi dall'integrazione bimodale sono più complessi di quanto previsto dallo schema VPAM, Summerfield [23], [30] ha proposto altre quattro ipotesi, diversificate dal tipo di

rappresentazione dell'informazione dvisiva e uditiva: integrazione di stime visive e uditive della funzione di filtro e di trasferimento del condotto vocale; di spettri acustici e di immagini visive fronto-facciali; di configurazioni tridimensionali del condotto vocale e infine di informazioni amodali la cui struttura profonda è la dinamica articolatoria. Non deve stupiure che il problema della rappresentazione sia fondamentale in una teoria della percezione bimodale: anzi esso è considerato il banco di prova per le teorie percettive uditive che si fondano sull'analisi della stimolazione prossimale, cioè le onde acustiche, perché devono confrontarsi con l'estrazione di informazione visiva direttamente relata all'evento distale, cioè all'articolazione [30], [31], [32], [33], [34].

Un interrogativo ancora più rilevante, e tuttora lontano dalla soluzione, riguarda l'esistenza di un sistema specializzato, o modulo, per la percezione bimodale. A questo proposito si oppongono le teorie computazionali dell'elaborazione a più stadi dell'informazione, quali il "Fuzzy Logic Model of Perception" di Massaro [24], [25] che prevede tre operazioni successive di valutazione delle caratteristiche, di integrazione delle caratteristiche e di decisione tra i vari prototipi che rappresentano le unità percettive di una lingua, e i modelli che prevedono che si percepiscanonon gli indici acustici isolati, ma la struttura dinamica profonda [32], [33] e che questa percezione sia da attribuire ad un modulo, senza la mediazione della rappresentazione cognitiva dello stimolo prossimale [30] e [31].

BIBLIOGRAFIA

[1] Erber N.P. (1972), Auditory, visual, and auditory-visual recognition of consonants by children with normal and impairment hearing, Journal of Speech and Hearing Research 15, 413-422.

[2] Summerfield A.Q. (1983), Audio-visual speech perception, lipreading and artificial stimulation, in M.E.Lutman & M.P.Haggard (Eds.), Hearing sciences and hearing disorders, Academic Press, London, 131-182.

[3] Magno Caldognetto E. & Vagges K. (1990), Il riconoscimento delle consonanti in un test di lettura labiale, Atti del Congresso Nazionale della Società Italiana di Acustica, l'Aquila, 94-99.

[4] Magno Caldognetto E. & Vagges K. (1990), Il riconoscimento visivo dei movimenti articolatori da parte di soggetti normali e ipoacusici. In Scritti in onore di Lucio Croatto, Padova, 1990, 153-166.

[5] Magno Caldognetto E., Vagges K., Ferrero F.E. (1983), Un test di confusione fra le consonanti dell'italiano: primi risultati, Atti del Seminario "La percezione del linguaggio" (Firenze, 17-20 dicembre 1980), Accademia della Crusca 123-179.

[6] Magno Caldognetto E., Ferrero F.E., Vagges K (1982), Intelligibilità delle consonanti dell'italiano in condizioni di mascheramento (S/R), di filtraggio passa-alto (PA) e passa-basso (PB), Bollettino Italiano di Audiologia e Foniatria, vol. 5, 163-172.

[7] [7] Magno Caldognetto E., Vagges K., Ferrero F.E. (1988), Intelligibilità e confusioni consonantiche in italiano, Rivista Italiana di Acustica, Vol. 12, 121-134.

[8] [8] Erber N.P. (1975), Auditory-visual perception of speech, Journal of Speech and Hearing Disorders 40: 481-492.

[9] [9] Mohamadi T. & Benoît C. (1992), Apport de la vision du locuteur à l'intelligibilité de la parole bruitée en français, Bulletin de la Communication Parlée 2, 31-41.

[10] Cohen M.M. & Massaro D.W. (1995), Perceiving visual and auditory information in consonant-vowel and vowel syllables. In C. Sorin et al. (Eds.), Levels in speech communication: Relations and interactions, Elsevier Science: 25-37.

[11] Linker W. (1982) Articulatory and acoustic correlates of labial activity in vowels: A cross-linguistic survey, UCLA, Working Papers in Phonetics 56, 1-134.

[12] Abry C. & Boe L.J. (1986), "Laws" for lips, Speech Communication 5, 97-104.

[13] Kelso J.A.S., Saltzman E.L., Tuller B. (1986), The dynamical perspective on speech production: Data and theory, Journal of Phonetics 14, 29-59.

[14] Gracco V.L. (1992), Analysis of speech movements: Practical considerations and clinical applications, Haskins Labs Status Report on Speech Research, SR-109/110, 45-58.

[15] Benoît C., Lallouache T., Mohamadi T., Abry C. (1992), A set of French visemes for visual speech synthesis, in G.Bally, C. Benoît, T.S. Sawallis (Eds.), Talking machines: theories, models, and designs, Elsevier Science Publ. 1992, 485-504.

[16] Magno Caldognetto E., Vagges K., Pedotti A., Ferrigno G. (1992), Parametri articolatori labiali e mandibolari nelle vocali cardinali dell'italiano, Atti delle III Giornate di Studio del Gruppo di Fonetica Sperimentale, Padova, 75-85.

[17] Magno Caldognetto E. & Vagges K. (1994), Caratteristiche articolatorie visibili delle vocali toniche e atone dell'italiano, Atti del XX Convegno Nazionale dell' A.I.A., Lecce, 479-484.

[18] Magno Caldognetto E., Vagges K., Zmarich C. (1995), Visible articulatory characteristics of the Italian stressed and unstressed vowels, Proc. of the XIII Int. Congr. of Phonetic Sciences, Stockholm, Vol. I, 366-369.

[19] Fromkin V. (1964), Lip positions in American English vowels, Language and Speech 7, 217-225.

[20] Zerling J.P. (1991), Labialité vocalique: Etude comparèe des types, degrés et strategies articulatoires de plusieurs langues, Proc. 12th International Congress of Phonetic Sciences, 1991, vol. 3, 46-49.

[21] Magno Caldognetto E., Vagges K., Zmarich C., Gelsomini F. (1995), L'analisi multiparametrica della cinematica articolatoria per la sintesi articolatoria del parlato, in Atti del XXIII Convegno Nazionale A.I.A., in corso di stampa.

[22] Magno Caldognetto E., Vagges K., Ferrero F.E., Busà G. (1992), Lip rounding coarticulation in Italian, Proc. of International Conference on Spoken Language Processing, Banff, vol. I, 61-64.

[23] Summerfield A.Q. (1987), Some preliminaries to a comprehensive account of audio-visual speech perception, in B.Dodd & R. Campbell (Eds.), Hearing by eye: The psychology of lip-reading, Lawrence Erlbaum Ass. Publ., Hillsdale, N.J., 3-51.

[24] Massaro D.W. (1987), Speech perception by hear and eye, in B.Dodd & R.Campbell (Eds.), Hearing by eye: The psychology of lip-reading, Lawrence Erlbaum Ass. Publ., Hillsdale (N.J.), 53-83.

[25] Massaro D.W. & Cohen M.M. (1992), Speech by eye. In G.Bally, C. Benoît, T.S. Sawallis (Eds.), Talking machines: theories, models, and designs, Elsevier Science Publ., 479-484.

[26] Cathiard M.A. (1988-1989), La perception visuelle de la parole: aperçu de l'état des connaissances, Bulletin de l'Institut de Phonétique de Grenoble, vol. 17-18, 109193.

[27] Summerfield Q., MacLeod A., Mc Grath M., Brooke M. (1989), Lips, teeth and the benefits of lipreading. In A.W.Young & H.D.Ellis (Eds.) Handbook of Research on Face Processing, Elsevier Science Publ., North Holland, 223-233.

[28] McGurk H. & Mac Donald J.W. (1976), Hearing lips and seeing voices, Nature 264, 746-748.

[29] Mac Donald J.W. & McGurk H. (1988), Visual influences on speech perception processes, Perception and Psychophysics, 24, 253-257.

[30] Summerfield A.Q. (1991), Visual perception of phonetic gestures. In I.G. Mattingly & M. Studdert-Kennedy (Eds.), Modularity and the motor theory of speech perception, L. Erlbaum Ass. Publ., Hillsdale (N.J.), 1991, 117-137.

[31] Liberman A.M. & Mattingly I.G. (1985), The motor theory of speech perception revised, Cognition21, 1-36.

[32] Fowler C.A. (1986), An event approach to the study of speech perception from a direct- realist perspective, Journal of phonetics 14, 3-28.

[33] Fowler C.A. & Rosenblum L.D. (1991), The perception of phonetic gestures. In I.G. Mattungly & M. Studdert-Kennedy (Eds.), Modularity and the Motor Theory of Speech Perception, Lawrence Erlbaum Ass., Hillsdale, N.J., 33-59.

[34] Magno Caldognetto E. (1990), Aspetti percettivi della coarticolazione, Rivista Italiana di acustica 14, 53-68.