CAPITOLO 3

LA CONVERGENZA CROSS-MODALE AUDIO-VISIVA E LA SPECIFICITA' DEI PATTERN EMOTIVI

1. Introduzione

Come anticipato nei precedenti capitoli, la comunicazione degli eventi emotivi è un fenomeno complesso e multicomponenziale, che si esplica attraverso l'azione congiunta di una costellazione di segnali di diversa natura. All'espressione, così come al riconoscimento di un'emozione partecipano infatti numerose informazioni interconnesse: le parole, il tono di voce con cui vengono pronunciate, l'espressione del viso, la postura, un gesto, una risata, un sussulto, che di per sé possono essere ambigui, una volta combinati tra loro in un preciso pattern espressivo, acquistano un notevole valore comunicativo. Tale combinazione viene indicata come *integrazione* o *convergenza cross-modale*. L'integrazione è un fenomeno che si realizza nel momento in cui informazioni appartenenti a diverse modalità sensoriali vengono identificate e processate come appartenenti ad un singolo evento o ad una singola sorgente fisica. Generalmente, la percezione di questa unitarietà ha origine dalla vicinanza spaziale e dalla sincronia con cui le informazioni si presentano. Di seguito, viene inizialmente presentata una panoramica dell'attuale stato della ricerca sui processi integrativi, mentre nella seconda parte del capitolo il focus della trattazione verrà ristretto alle indagini che si sono occupate di comprendere i processi sottesi alla convergenza di informazioni cariche di valenza emotiva.

2. Il contributo degli studi comportamentali

2.1. L'integrazione multimodale come risposta alla complessità ambientale

Oggi si ritiene che i sistemi visivo, uditivo e somatosensoriale abbiano un'organizzazione di tipo gerarchico, tale per cui la stimolazione sensoriale produce una rappresentazione percettiva che passa attraverso una serie di stadi di processamento di complessità crescente. Tale organizzazione gerarchica sembra d'altro canto costituire un principio generale del funzionamento cerebrale.

Per completare il quadro, occorre inoltre soffermare l'attenzione sul fatto che normalmente i nostri organi di senso sono bersagliati contemporaneamente da una molteplicità di stimoli appartenenti a diverse modalità sensoriali. Ed infatti, corrispondentemente, numerosi recenti studi suggeriscono che, fin dalle prime fasi di elaborazione delle informazioni, si verifichino, sia a livello delle *aree di convergenza* come la corteccia parietale, temporale (Schroeder & Foxe, 2002) e frontale (Graziano et al., 1997) sia a livello delle *aree specifiche* di elaborazione sensoriale (ad esempio l'area visiva V1 o l'aera uditiva A1) fenomeni di convergenza cross-modale.

Ma qual è la funzione di tale processo di integrazione multisensoriale?

L'abilità nell'integrare stimoli ridondanti appartenenti a differenti modalità per formare un unico percetto costituisce una fondamentale componente alla base del comportamento e della cognizione guidati dai sensi. Essa ha una *funzione adattiva*, in quanto è finalizzata a migliorare l'elaborazione dello stimolo, in modo tale da produrre una risposta comportamentale più rapida e accurata. Ad esempio, le ricerche indicano che essa produce una migliore performance nei compiti di orientamento dell'attenzione e di riconoscimento (Schroeder et al., 2003).

2.2. I principi che regolano la percezione multimodale

L'ipotesi che l'integrazione multisensoriale abbia una funzione adattiva di facilitazione della risposta comportamentale è supportata dai risultati ottenuti da

quegli studi che hanno applicato il paradigma classicamente utilizzato per indagare la convergenza multisensoriale, il quale prevede un confronto tra le risposte agli stimoli unimodali con le risposte agli stimoli multimodali che derivano dalla loro combinazione. La letteratura indica che le risposte comportamentali a stimoli multimodali, se confrontate con quelle a stimoli di natura unimodale, sono più rapide in termini di tempi di risposta (TR) e più accurate (Welsch & Warren, 1986). Questo fenomeno viene detto *multisensory enhancement* (miglioramento multisensoriale). Ci si riferisce ad esso anche come *redundant target effect* (effetto di ridondanza del target), poiché l'effetto di miglioramento deriva proprio dal carattere di ridondanza, ripetitività e reciproca conferma degli stimoli. Tale processo, inoltre, è influenzato da alcuni fattori legati alle caratteristiche degli stimoli e alla modalità di presentazione degli stessi. L'azione di tali fattori è descritta da quelli che vengono considerati i tre principi che regolano l'integrazione. Secondo il *principio della vicinanza temporale*, la relazione temporale tra gli stimoli è un fattore critico ai fini della convergenza: gli stimoli separati da un intervallo temporale inferiore ai 100 ms sono quelli che hanno maggiore probabilità di elicitare un miglioramento della risposta. Oltre a ciò, secondo il *principio della vicinanza spaziale,* anche la prossimità spaziale tra gli stimoli ha un'importanza cruciale: quanto più gli stimoli sono ravvicinati e tanto maggiore sarà la possibilità che si verifichi il fenomeno di *multisensory enhancement*. Infine, secondo il principio dell'*inverse effectiveness effect* (effetto dell'efficacia inversa), la forza e l'efficacia dello stimolo unimodale sono inversamente correlate con il miglioramento della risposta multimodale. In altri termini, gli stimoli che di per sé sono poco efficaci producono i massimi livelli di *multisensory enhancement*, mentre gli stimoli che hanno una massima efficacia se presentati singolarmente producono uno scarso miglioramento della risposta se abbinati ad un altro stimolo.

2.3. La natura dell'integrazione: una questione aperta

Nonostante il fatto che l'integrazione multisensoriale sia una caratteristica fondamentale della percezione, tuttavia le nostre conoscenze circa il modo in cui un oggetto caratterizzato da componenti che fanno riferimento a differenti modalità sensoriali venga percepito come un oggetto unitario sono ancora incomplete e molte questioni rimangono aperte. In particolare, il tema di discussione maggiormente controverso riguarda la *natura dei processi implicati* nell'integrazione. Marks, Ben-Artzi e Lakatos (2003) offrono una panoramica degli studi che hanno cercato di dare una risposta a questo quesito, partendo dall'esame delle cosiddette corrispondenze cross-modali. Diversi studi suggeriscono che esistano, nella percezione, delle corrispondenze sensoriali. Tra di esse, quella maggiormente indagata è la sinestesia: nei fenomeni di sinestesia, la stimolazione di una certa modalità sensoriale comporta il coinvolgimento di una seconda modalità. Nella sinestesia audio-visiva gli stimoli acustici evocano risposte sia uditive sia visive, per cui ad esempio la persona può "vedere" colori e forme nei suoni. La sinestesia rivela evidenti corrispondenze cross-modali: ad esempio la brillantezza delle immagini visive aumenta all'aumentare della tonalità dello stimolo acustico. Anche i compiti di confronto e di giudizio sulla similarità cross-modale (Marks, 1989) rivelano le medesime corrispondenze: già in tenera età i bambini individuano delle corrispondenze del tipo brillantezza-intensità o luminosità-tonalità (Marks, 1978). Vi sarebbe poi una corrispondenza diretta tra congruenza e corrispondenza: la corrispondenza si instaura ad esempio tra suoni acuti e immagini molto luminose ma non tra suoni acuti e immagini poco luminose. Queste interazioni legate alla congruenza solitamente hanno natura bidirezionale.

L'interazione cross-modale legata alle corrispondenze cross-modali è stata rilevata anche nel caso dei compiti di discriminazione. I risultati mostrano che, quando si chiede di discriminare tra una luce debole e una brillante che compaiono contemporaneamente ad un suono acuto o grave, i tempi di reazione sono più rapidi

e la performance è più accurata quando vi è congruenza tra i due stimoli (luce brillante e suono acuto; luce debole e suono basso) (Marks, 1987).

In alcuni casi la corrispondenza, e di conseguenza l'integrazione, sembrano derivare da associazioni apprese tra gli stimoli, che potrebbero quindi avere una funzione di facilitazione nell'identificazione degli stimoli stessi. Un esempio di ciò è dato dalla corrispondenza tra colore caldo e temperatura elevata e tra colore freddo e bassa temperatura. L'ipotesi che tali associazioni siano apprese è suggerita dal fatto che esse non sono presenti nei bambini piccoli (Marks, 1987).

Tuttavia, in altri casi, all'origine di una corrispondenza (ad esempio quella tra tonalità del suono e brillantezza o tra intensità del suono e brillantezza) non può essere identificata un'influenza ambientale: queste associazioni infatti sono presenti già nella prima infanzia. In questo caso diventa più difficile individuare la loro natura e la loro funzione nella percezione. L'analisi della letteratura indica l'esistenza di tre diverse possibilità. La prima è che la corrispondenza cross-modale derivi dal fatto che, nei primi stadi di *processamento sensoriale*, si verifichi una qualche forma di dialogo tra l'elaborazione degli stimoli visivi e quella degli stimoli uditivi. Una seconda ipotesi invece attribuisce l'interazione cross-modale non al precoce processamento sensoriale ma ad un *processo decisionale più tardivo* (Marks, 2004). In questo caso la presenza di uno stimolo congruente avrebbe un effetto di facilitazione, abbassando così la soglia del criterio di risposta, senza andare in alcun modo a influenzare il processamento sensoriale dello stimolo. Infine, è plausibile ipotizzare che l'interazione cross-modale coinvolga sia processi di natura sensoriale sia processi di natura decisionale (Odgaard et al., 2003). Una recentissima ricerca (Colin, Radeau & Deltenre, 2005) offre a tal riguardo risultati assai interessanti. La ricerca, che ha sfruttato l'effetto McGurk[1], ha indagato l'*audiovisual speech* (integrazione delle componenti visive e uditive del parlato) modulando alcune variabili di natura sia

[1] L'effetto McGurk (McGurk & McDonald, 1976), indagato in relazione all'*audiovisual speech*, si verifica quando al soggetto vengono presentate due sillabe diverse, una in forma visiva (movimento delle labbra) e una in forma uditiva (parlato). In presenza quindi di una discordanza tra la componente uditiva e quella visiva, i soggetti, cui viene chiesto di riprodurre la sillaba percepita, combinano quanto hanno udito con quanto hanno visto. Ad esempio, le sillabe percepite visivamente "gaga" e le sillabe percepite uditivamente "baba" vengono integrate nelle sillabe percepite "dada".

sensoriale sia cognitiva. Nel corso di due esprimenti, sono state infatti manipolate due variabili sensoriali (l'intensità della voce e la grandezza del volto) e una variabile cognitiva (tipo di task, con risposta a scelta multipla o con risposta libera). I risultati indicano che l'integrazione dipende da entrambi gli ordini di fattori, percettivi e cognitivi.

3. L'apporto della neuropsicologia alla comprensione del processo di integrazione

La questione della natura dell'integrazione cross-sensoriale è stata ulteriormente approfondita affiancando ai dati comportamentali dati di natura neuropsicologica, che si pongono l'obiettivo di esplorarne i processi neurali sottostanti.

Gli studi che hanno utilizzato tecniche elettroencefalografiche e di *neuroimaging* hanno confermato l'esistenza del *redundant target effect,* già rilevato a livello comportamentale (Fort et al., 2002; Teder-Sälejärvi et al., 2002). Gli stimoli bimodali congruenti producono infatti, negli stadi di elaborazione sensoriale, risposte neurali più veloci e di maggiore intensità rispetto a quelle elicitate da stimoli unimodali o da stimoli bimodali incongruenti. L'insieme di tali dati suggerisce che, in presenza di stimoli congruenti, avvenga una qualche forma di facilitazione dovuta all'integrazione cross-modale. A questo proposito, il "*redundant target effect*" ha suggerito diverse interpretazioni circa i processi sottostanti implicati:

- i *race models* affermano che le due componenti vengono processate indipendentemente e che il tempo di risposta coincide con la fine del processamento di quella che, tra le due, termina in tempi più rapidi;
- gli *independent coactivation models* ipotizzano che le due componenti inducano attivazioni indipendenti che vengono sommate per elicitare la risposta;
- infine, gli *interactive coactivation models* affermano che il processamento di uno stimolo in una modalità influenza il processamento di uno stimolo in un'altra modalità, ipotizzando che l'integrazione possa avvenire a diversi livelli: di processamento sensoriale e/o cognitivo di selezione della risposta.

Recentemente alcuni studi neuropsicologici hanno fornito supporto a favore di quest'ultimo tipo di modello (Calvert et al., 1999; Calvert et al., 2000; Hadjikhani & Roland, 1998). In genere, vi è accordo tra i ricercatori sul fatto che l'integrazione abbia inizio ad uno stadio molto precoce del processo di elaborazione sensoriale degli stimoli (Giard & Peronnet, 1999). In un esperimento condotto da Giard & Peronnet, i soggetti furono sottoposti ad un compito di identificazione che utilizzava stimoli unimodali visivi e uditivi e stimoli bimodali congruenti risultanti dalla combinazione delle due componenti. Come previsto, l'identificazione degli stimoli bimodali fu più rapida ed accurata. Un'analisi spaziotemporale degli ERPs ha mostrato che già tra i 40 e i 200 ms dopo la presentazione dello stimolo si manifestano patterns multipli di integrazione cross-modale sia nelle aree corticali specifiche visiva e uditiva sia in aree non specifiche, come nella regione fronto-temporale destra. Gli effetti indotti da stimoli bimodali ridondanti sono stati interpretati come modulazione della risposta unimodale uditiva N1 e della risposta unimodale visiva N185 nelle rispettive cortecce sensoriali, nonché come nuova attività nella corteccia visiva e nelle aree fronto-temporali destre. Anche Teder-Sälejärvi (Teder-Sälejärvi et al., 2002) ha rilevato una prima deflessione che ha inizio attorno ai 130 ms e raggiunge il picco tra i 160 e i 170 ms nelle aree corticali occipito-temporali ventrali. Anche in questo caso, una simile interazione audio-visiva potrebbe essere interpretata come modulazione dell'onda visiva N1. Tale effetto infatti sembra rappresentare un'influenza dell'input uditivo sul processamento che ha luogo in un'area corticale prevalentemente visiva.

Relativamente alla presentazione di stimoli bimodali incongruenti, Fort (Fort et al., 2002) hanno trovato che in generale, come previsto, gli stimoli bimodali non ridondanti non producono un effetto di facilitazione a livello di dati comportamentali, dal momento che per portare a termine il compito di identificazione i soggetti devono processare in modo completo ogni componente dello stimolo. Invece, contrariamente a quanto previsto, sebbene essa sia di minore intensità e più tarda rispetto a quella registrata in presenza degli stimoli ridondanti, è stata rilevata una precoce attività cross-modale in risposta alla presentazione degli

stimoli bimodali non ridondanti, caratterizzata da attivazione sia nelle aree sensoriali specifiche sia nell' area fronto-temporale destra non specifica. Gli autori suggeriscono che l'ipotesi della coattivazione interattiva, che ha ricevuto supporto da recenti studi neuropsicologici, ben si adatta anche ai risultati di questo studio. Tali risultati, inoltre, sono in accordo con i principi neurali di integrazione multisensoriale che si applicano a livello dei singoli neuroni nel collicolo mammale superiore e nella corteccia polisensoriale: secondo tali principi, a questo livello la coincidenza spaziale e temporale degli stimoli è condizione sufficiente per innescare l'integrazione (Bushara et al., 2001; Stein & Wallace, 1996). Si può quindi ipotizzare che, a fronte di una precoce convergenza sensoriale sempre presente, solo in un secondo momento le modalità di elaborazione degli stimoli si differenzino sulla base della natura congruente o incongruente delle informazioni sensoriali. In altre parole, l'attivazione indica un'integrazione cross-modale non tanto nell'identificazione dello stimolo quanto nella sua mera ricezione (*detection*).

I meccanismi fisiologici dell'integrazione sono complessi e molteplici. L'insieme dei risultati dimostra infatti la *flessibilità* dei processi cross-modali, che presentano notevoli possibilità di adattamento in funzione delle caratteristiche dello stimolo. Essi infatti sono influenzati da fattori sia di natura endogena (ad esempio Fort et al. (2002) hanno rilevato un effetto del grado di *expertise* dei soggetti rispetto al task, per cui i soggetti a dominanza visiva e i soggetti a dominanza uditiva presentavano differenti patterns di attivazione) sia di natura esogena (condizioni sperimentali e tipo di compito). La *natura della stimolazione* ha sicuramente un effetto sulle modalità di integrazione (Callan et al., 2001; Calvert et al., 2001). Tale carattere flessibile dei processi di integrazione ha ancora una volta una funzione adattiva, in quanto risponde all'esigenza di produrre una risposta efficiente in presenza di condizioni ambientali variabili.

3.1. Circuiti neurali implicati nel decoding intersensoriale

Relativamente alle aree coinvolte nell'integrazione cross-modale, esistono due differenti scuole di pensiero: alcuni ritengono che ogni specifica combinazione di stimoli sensoriali (ad es. audio-visiva o audio-tattile) venga integrata in una precisa area "associativa" polimodale ad essa dedicata. Per quanto riguarda nello specifico l'integrazione audio-visiva, si ritiene che le aree dedicate siano la corteccia frontale destra inferiore, la corteccia temporale destra, il solco temporale superiore, il giro temporale superiore, l'insula e il lobo parietale, come evidenziato da studi PET (Bushara, Grafman & Hallett, 2001; Hadjikhani & Roland, 1998) e fMRI (Calvert et al., 2000; Downar et al., 2000).

Altri invece sostengono che le aree che processano gli stimoli unimodali processino anche gli stimoli multimodali, ipotizzando che i sensi abbiano accesso l'uno all'altro grazie ad aree di ritrasmissione (*relay*) subcorticali. Questa ipotesi è supportata dall'evidenza che la lesione delle presunte aree polimodali non preclude l'integrazione intersensoriale (Ettlingen & Wilson, 1990 per una rassegna). A tale proposito, con uno studio che ha utilizzato la risonanza magnetica funzionale (fMRI) per indagare il fenomeno del *lip-reading* (comprensione del linguaggio attraverso la lettura dei movimenti labiali), Olson (Olson, Gatenby & Gore, 2002) ha fornito supporto a questa seconda ipotesi dimostrando che le aree unimodali, utilizzando come aree subcorticali di ritrasmissione il claustrum e il putamen, elaborano stimoli appartenenti a diverse modalità sensoriali, mettendoli in comunicazione tra loro.

Recentemente Fort e Giard (2004) hanno suggerito una nuova prospettiva: essi hanno dimostrato che la convergenza ha inizio a livello delle aree sensoriali specifiche in uno stadio veramente molto precoce, intorno ai 40-50 ms dopo la presentazione dello stimolo. Rispetto a questi fenomeni precoci, studi condotti sulle scimmie suggeriscono l'esistenza di proiezioni dirette dalla corteccia uditiva primaria alla corteccia visiva primaria e viceversa (Falchier et al., 2002; Schroeder et al., 2001). Questo dato è difficilmente compatibile con l'ipotesi che siano presenti delle proiezioni dalle aree di convergenza polisensoriali verso le aree specifiche

(Calvert et al., 2001). Gli autori tuttavia ipotizzano che tali proiezioni siano coinvolte in stadi più tardivi di processamento, di natura cognitiva più che percettiva.

3.2. I neuroni multisensoriali

Infine, bisogna specificare che il fenomeno dell'integrazione cross-modale può essere indagato a diversi livelli. Un primo livello di analisi - quello che è stato esposto fino ad ora - pone l'attenzione su specifiche regioni che fungono da aree di convergenza o che partecipano al processo di integrazione. Tuttavia un'ulteriore analisi di tipo strettamente psicofisiologico può essere effettuata anche a livello di singoli neuroni detti *neuroni multimodali*, che hanno la particolarità di elaborare informazioni appartenenti a diverse modalità sensoriali. Esistono infatti neuroni bimodali e trimodali in grado di gestire informazioni sia uditive sia visive sia somatosensoriali.

Mentre della convergenza a livello di area cerebrale abbiamo oggi una discreta conoscenza, i meccanismi sottostanti alla convergenza a livello di singoli neuroni sono ancora poco noti. Ciò di cui siamo a conoscenza è l'esistenza di due tipologie di convergenza multimodale, che vengono differenziate sulla base del loro effetto. Il primo tipo di convergenza neuronale, che è stato studiato a livello del collicolo superiore e a livello della corteccia cerebrale, è detto *convergenza eccitatoria-eccitatoria*. Quando i neuroni multimodali che operano tale tipo di convergenza ricevono informazioni di diverso tipo, le integrano, provocando un miglioramento della risposta. Ad esempio, è possibile che un neurone risponda debolmente ad un certo stimolo uditivo e in modo più accentuato ad un certo stimolo visivo; nel caso in cui esso riceva simultaneamente i due stimoli, la sua risposta sarà nettamente più intensa. Questo tipo di risposta, come nel caso della convergenza a livello delle aree cerebrali, viene detto miglioramento multisensoriale (*multisensory enhancement*). Esso è influenzato da diversi fattori legati alle caratteristiche degli stimoli, alle modalità di presentazione degli stessi e alle caratteristiche del neurone che opera la convergenza e rispetta i principi della vicinanza temporale, della vicinanza spaziale e dell'*inverse*

effectiveness effect. Il secondo tipo di integrazione è denominato *convergenza eccitatoria-inibitoria*. Essa si verifica ad esempio nel caso di soppressione di una risposta in seguito alla presentazione di uno stimolo inatteso (per esempio uno stimolo visivo al posto di uno stimolo uditivo) durante un compito di attenzione selettiva. Si verifica in questo caso un fenomeno di inibizione della risposta. Infatti, a fronte dell'azione eccitatoria di una modalità, è presente un'azione inibitoria esercitata dall'altra modalità. Mentre nel caso della convergenza eccitatoria-eccitatoria l'influenza sulla risposta è molto accentuata, nel caso della convergenza eccitatoria-inibitoria si osserva invece semplicemente una lieve modulazione della risposta (Meredith, 2002).

4. La decodifica audio-visiva dei volti: riconoscere l'identità dal volto e dalla voce

La maggior parte delle ricerche che hanno indagato la cross-modalità hanno impiegato stimoli sensoriali molto semplici, rilevando la centralità delle fasi precoci di elaborazione percettiva. Sostanziali differenze sono state rilevate invece da coloro che hanno utilizzato stimoli che si collocano ad un più elevato livello informativo.
Particolarmente interessante è un recentissimo studio di Schweinberger (Schweinberger, in press) che ha dimostrato l'importanza dell'integrazione audiovisiva ai fini del riconoscimento dell'identità delle persone. Da precedenti studi era infatti emerso che sia il volto sia la voce costituiscono delle informazioni importanti quando dobbiamo giudicare il grado di familiarità, ma non era mai stata indagata la convergenza tra i due codici rispetto a tale compito. Un indizio della possibile presenza di fenomeni integrativi era stato fornito da una ricerca che, utilizzando la risonanza magnetica, ha dimostrato che la percezione di una voce familiare attiva la cosiddetta "area fusiforme del volto", che tipicamente viene appunto attivata dalla percezione dei volti (von Kriegsten et al., 2005). Nello studio di Schweinberg, ai soggetti veniva chiesto di giudicare se una frase standardizzata veniva pronunciata da una persona familiare o sconosciuta. Nella condizione unimodale veniva presentata solo la voce, mentre quella audiovisiva era caratterizzata dalla simultanea presentazione di un volto, familiare o sconosciuto,

congruente o incongruente. I risultati dimostrano che, in termini sia di accuratezza sia di tempi di risposta, la simultanea presentazione del volto produce sistematici costi (nella condizione di incongruenza) e benefici (nel caso della congruenza) nella valutazione delle voci familiari, mentre nel caso delle voci non note tali effetti non si verificano. Gli autori suppongono che ciò sia dovuto al fatto che, a seguito della presentazione delle informazioni audiovisive, viene operato un confronto queste e le rappresentazioni multimodali delle persone familiari che sono conservate nella memoria a lungo termine.

5. La convergenza di pattern emotivi

Se il riconoscimento del volto neutro costituisce un processo complesso, ancor più complessi sono i meccanismi che il nostro sistema cognitivo attua quando rileva che le informazioni multimodali sono cariche di significato emotivo.

Quando decodifichiamo un'emozione, utilizziamo molteplici fonti di informazione. Numerose ricerche si sono occupate di capire cosa avviene quando, allo scopo di riconoscere e comprendere un'emozione, l'individuo si trova a dover in qualche modo integrare tali informazioni appartenenti a diversi sistemi sensoriali. Gli studi che sono stati condotti allo scopo di comprendere questo particolare processo di integrazione cross-modale, hanno focalizzato l'attenzione in particolare sulla convergenza tra il *canale visivo e* quello *uditivo* che, come illustrato nei cap. 1 e 2 , hanno un ruolo fondamentale nel processo di decoding delle emozioni.

Sembra esistere una stretta interrelazione tra la decodifica della mimica facciale e la decodifica dell'espressione vocale delle emozioni. A tal proposito, van Lancker e Sidtis (1992) hanno trovato che alcuni pazienti con diagnosi di aprosodia presentavano anche un correlato deficit nel riconoscimento dei volti. Parallelamente, Scott et al. (1997) hanno osservato un'incapacità di decodifica delle componenti prosodiche in un paziente con difficoltà nel riconoscimento delle espressioni facciali. Questi dati non permettono tuttavia di stabilire con certezza che le informazioni facciali e prosodiche convergano in una rappresentazione amodale

comune: essi attestano soltanto l'esistenza di una semplice correlazione tra i due ordini di deficit. Tra l'altro, alcuni studi hanno indicato l'esistenza di asimmetrie tra riconoscimento della voce e riconoscimento del volto. Alcune emozioni sono infatti più facilmente riconoscibili sulla base del volto o viceversa: ad esempio la gioia viene facilmente riconosciuta sulla base dell'espressione facciale, ma spesso la voce della gioia viene confusa con l'espressione neutra (Vroomen et al., 1993). L'insieme di queste e simili ricerche ha fornito spunti di riflessione interessanti, che sono stati sviluppati da un filone sperimentale che si è posto l'obbiettivo specifico di esplorare il processamento di stimoli emotivi multimodali. Tale corpus di ricerche include sia studi di natura comportamentale sia studi di natura neuropsicologica.

6. L'apporto degli studi comportamentali

Prima di intraprendere un excursus sui risultati delle ricerche che hanno indagato la decodifica cross-modale delle emozioni, è necessario sottolineare che essa presenta delle caratteristiche qualitativamente diverse rispetto alla percezione cross-modale classicamente studiata, che ha impiegato come stimoli lampi di luce e semplici suoni inarticolati. Ciò che differenzia la decodifica cross-modale delle emozioni è la *complessità degli stimoli* implicati. Un parallelo può essere individuato negli studi che si sono occupati di indagare un caso particolare di processamento multimodale: lo *speech reading* o comprensione della lingua parlata attraverso il simultaneo processamento delle informazioni uditive e visive correlate. Normalmente, quando qualcuno ci parla, noi siamo impegnati sia ad ascoltare le sue parole sia a guardare il movimento delle sue labbra. La nostra comprensione è il risultato dell'integrazione tra questi due livelli di informazione. L'effetto McGurk (McGurk & McDonald, 1976) ha dimostrato che tale integrazione ha carattere automatico e obbligato (vedi cap. 2): quando ad un soggetto vengono presentate due diverse sillabe, l'una in forma visiva (movimento delle labbra) e l'altra in forma uditiva (linguaggio parlato) e gli viene chiesto di riferire la sillaba percepita, egli riporta un percetto derivante dalla combinazione delle due. Allo stesso modo, anche

le componenti mimiche e vocali dell'espressione emotiva costituiscono delle informazioni complesse.

Sono stati Beatrice de Gelder e il suo gruppo di ricerca a condurre buona parte degli studi empirici che si sono occupati specificamente di indagare la percezione cross-modale delle emozioni basata su informazioni di natura audio-visiva.

In primo luogo la de Gelder (de Gelder & Vroomen, 2000) si è posta lo scopo di determinare se, in una situazione bimodale in cui le informazioni sullo stato emotivo sono fornite sia attraverso il canale visivo sia attraverso quello vocale, entrambe le modalità contribuiscono al riconoscimento. A tal fine, ai soggetti sono stati mostrati stimoli costituiti da volti e frasi esprimenti tristezza o gioia, in condizione unimodale e bimodale. In una prima fase dell'esperimento, ai soggetti è stato semplicemente chiesto di indicare se la persona cui il volto e/o la voce si riferivano era triste o felice. Nelle due fasi successive è stata invece data loro istruzione di prestare attenzione solo all'espressione del volto o solo al tono di voce. I risultati indicano che, come avviene nei classici esperimenti sulla percezione bimodale, i tempi di latenza sono più veloci quando vengono somministrati due stimoli congruenti (volto e voce esprimenti la stessa emozione) rispetto a quando viene presentato un solo stimolo. Questo fatto indica che, per il sistema di processamento, l'integrazione delle informazioni visive ed uditive costituisce un meccanismo usuale ed efficace.

I tempi più lunghi si registrano invece nel caso di due stimoli incongruenti (volto e voce esprimenti due emozioni diverse), indicando che tale situazione rappresenta un condizione poco naturale e che quindi richiede un maggiore sforzo in termini di decodifica delle informazioni. Per quanto concerne la correttezza del riconoscimento, è stata osservata un'influenza del volto sulla voce e, viceversa, della voce sul volto, come precedentemente riscontrato dallo studio pionieristico di Massaro ed Egan (1996).

6.1. La funzione dell'integrazione delle informazioni emotive multimodali

La de Gelder (de Gelder, 2000) si è chiesta quale sia la funzione della convergenza cross-modale nel decoding delle emozioni. La decodifica simultanea di informazioni acustiche e visive rappresenta infatti un caso di ridondanza. Tre diverse ipotesi sono state formulate per spiegare tale fenomeno:

In primo luogo è possibile che la presenza di due diversi tipi di segnali sia utile quando la ricezione dei segnali provenienti da uno dei due sistemi è povera o assente, ad esempio in presenza di rumore o cecità. Questa ipotesi però non spiega perché, anche nel caso in cui entrambi i sistemi funzionino al meglio, l'organismo processi in modo completo tutte le informazioni disponibili.

Una seconda ipotesi è che l'organismo sia avvantaggiato dalla ridondanza perché i due sistemi sono complementari e che questa condizione gli permetta una maggiore efficienza nella risposta comportamentale. Viene assunto come prova di ciò il fatto che, come già accennato, alcune emozioni vengono meglio espresse dal sistema visivo ed altre da quello uditivo. Tuttavia, bisogna osservare che la convergenza avviene anche quando entrambi gli ordini di informazioni presentano scarsa ambiguità (de Gelder & Vroomen, 2000).

De Gelder (2000) avanza quindi una terza ipotesi, suggerendo che la ridondanza permetta una maggiore *efficienza nella risposta comportamentale* non perché i due sistemi siano complementari ma perché l'organismo già dai primissimi stadi di processamento integra gli stimoli e ciò gli consente di produrre una risposta molto più veloce rispetto al caso in cui processasse gli stimoli separatamente per poi integrare i percetti solo nella fase finale. A conferma di ciò, paragonando il decoding unimodale a quello multimodale, gli studi della de Gelder dimostrano che il processamento degli stimoli integrati precocemente avviene in modo più veloce ed efficiente.

6.2. L'integrazione come processo precoce ed automatico

Come già accennato, a livello di riconoscimento ed etichettamento delle emozioni presentate nella duplice modalità audiovisiva, esiste un reciproco effetto di influenzamento tra le due modalità sensoriali. Il fatto che tale *bias* cross-modale si verifichi anche quando viene esplicitamente richiesto di prestare attenzione ad un'unica modalità sensoriale (quella visiva o quella uditiva) (de Gelder & Vroomen, 2000) rinforza l'ipotesi che l'integrazione avvenga ad uno stadio di processamento molto *precoce* e in modo *automatico ed obbligato*. Ciò induce ad escludere l'ipotesi che il *bias* possa essere il frutto di una valutazione e di un giudizio consapevoli, attuati come risultato della presa di coscienza di un'incongruenza dopo che il processamento separato delle due fonti di informazione è terminato. Anzi, addirittura l'integrazione si verifica nonostante il fatto che i soggetti si dichiarino consapevoli dell'incongruenza. Evidentemente quindi questa impressione fenomenica di incongruenza si colloca ad un livello cosciente e molto differente da quello in cui avviene il processamento cross-modale. Si può quindi ritenere, secondo gli autori, che l'integrazione sia un *fenomeno percettivo*, analogamente a quanto rilevato nel caso dell' *audio-visual speech*. Tale processo percettivo precede ampiamente fenomeni come il riconoscimento e la comprensione delle emozioni, che sono centrati sul significato personale e sociale dell'emozione. A conferma delle proprie affermazioni, de Gelder e colleghi (de Gelder, Vroomen & Bertelson, 1998) hanno replicato gli esperimenti precedentemente condotti introducendo una variante: in alcuni casi, infatti, i volti venivano presentati invertiti. L'inversione del volto comporta una drastica diminuzione della possibilità di identificazione dell'identità e dell'espressione del volto. Ciò è dovuto al fatto che, come precedentemente esposto, l'identificazione del volto si differenzia dall'identificazione di altri tipi di oggetti, in quanto coinvolge la configurazione complessiva e non i singoli attributi. I risultati dello studio mostrano che il decoding dell'espressione facciale influenza il giudizio circa il tono della voce solo quando il volto è presentato dritto ma non quando è capovolto. Questo dato è interessante se

messo in relazione al fatto che il riconoscimento dell'emozione espressa dal volto diventa difficoltoso quando lo stimolo è capovolto. L'effetto cross-modale osservato quando lo stimolo è presentato dritto conferma invece l'ipotesi che tale processo sia un fenomeno percettivo automatico che non può essere ridotto ad un processo post-percettivo volontario di aggiustamento.

7. Il contributo della neuropsicologia

A livello neuropsicologico, diversi studi sono stati effettuati allo scopo di mettere in luce i processi cerebrali implicati nella decodifica multimodale delle emozioni e di spiegare con maggiore chiarezza i dati comportamentali disponibili. Come precedentemente esposto, gli studi behavioural mostrano che, quando vengono presentati simultaneamente due stimoli emotivi congruenti, l'uno vocale e l'altro facciale, la risposta è più accurata e i tempi sono più rapidi rispetto a quando viene presentato uno stimolo unimodale. Ciò suggerisce che l'organismo sfrutti le risorse multiple offerte dall'ambiente ai fini di produrre delle risposte comportamentali più rapide ed efficienti. Tuttavia, i dati disponibili non forniscono alcuna prova certa del fatto che l'integrazione avvenga ad uno stadio precoce del processamento. I tempi più brevi che si presentano in concomitanza con gli stimoli bimodali potrebbero essere spiegati da un *race model*, cioè da un modello secondo il quale i due stimoli vengono processati separatamente, e quello il cui processamento ha termine per primo determina la prestazione. Un'altra ipotesi possibile è che l'integrazione degli stimoli abbia luogo non appena essi si presentano e che il loro processamento congiunto sia il meccanismo che meglio potrebbe sfruttare la ridondanza della stimolazione, come supposto dagli *interactive coactivation models* (vedi cap. 3). Uno degli obiettivi che si pongono le ricerche neuropsicologiche è quindi quello di portare nuove conoscenze che possano disambiguare tale questione.

Un ulteriore principale argomento di discussione nella ricerca sulla convergenza cross-modale degli stimoli di natura emotiva riguarda le sue coordinate temporali. Di conseguenza, lo studio dei *potenziali evocati corticali (ERPs)* si è rivelato,

data la sua alta definizione temporale, particolarmente utile ed efficace (Rugg & Coles, 1997).

Proprio al fine di ampliare le conoscenze relative a tali questioni, pressoché tutti gli studi condotti hanno centrato l'attenzione sugli *stadi percettivi precoci* del processo di elaborazione degli stimoli emotivi. Infatti, gli indici ERP più frequentemente indagati, come la N1 e il MMN, segnalano processi di natura sensoriale.

7.1 La componente MMN (mismatch negativity) come indicatore indiretto della convergenza audio-visiva

I primi studi sull'argomento, effettuati alla fine degli anni '90, hanno utilizzato come indice la *MMN (mismatch negativity)*, un picco negativo che si presenta in concomitanza con stimolazioni di tipo uditivo quando, in una serie di stimoli ripetitivi, appare uno stimolo deviante (Näätänen, 1992). Tale ERP non è sotto controllo attentivo e segnala la ricezione di uno stimolo che tradisce le aspettative (Levänen & Sams, 1997). De Gelder e colleghi (de Gelder et al., 1999) hanno utilizzato la componente MMN al fine di indagare l'influenza dell'espressione facciale sul processamento delle componenti emozionali vocali. Ai soggetti sono state presentate coppie congruenti o incongruenti di stimoli uditivi (parole pronunciate in tono triste o arrabbiato) e stimoli visivi (espressioni facciali di tristezza e rabbia) con la consegna di prestare attenzione al volto e di ignorare la componente uditiva. I risultati indicano che quando, dopo una serie di stimolazioni congruenti, ne viene presentata una incongruente, appare una risposta cerebrale negativa precoce (latenza 178 ms) localizzata nelle aree anteriori della corteccia, in particolare in F3, Cz e soprattutto Fz. Lo stesso avviene quando, dopo una serie di stimoli congruenti, ne compare uno incongruente. I parametri della componente ERP evidenziata dalla de Gelder corrispondono a quelli della MMN, che ha infatti una latenza di 178 ms ed è principalmente localizzato in Fz. Gli autori ipotizzano che l'assenza di una lieve

positività identificabile come P3 o P3a indichi che il processo non avviene sotto controllo attentivo ma è obbligato.

I dati confermano ed estendono i precedenti risultati ottenuti negli studi comportamentali, pur non permettendo, data la tecnica impiegata, di localizzare con precisione le sedi in cui avviene l'integrazione cross-modale. Anche Surakka et al. (1998) hanno utilizzato la MMN per studiare l'integrazione tra stimolazioni visive e stimolazioni uditive, rilevando che gli stimoli visivi hanno un impatto sul processamento degli stimoli uditivi: Surakka ha infatti studiato l'effetto di immagini emotivamente connotate tratte dall'International Affective Picture System su stimoli uditivi (toni standard di 1000 Hz con probabilità pari a 0.85 e toni devianti di 1050 Hz con probabilità pari a 0.15), trovando che l'ampiezza dell'MMN era significativamente attenuata quando l'emozione legata alla figura era a basso arousal e positiva rispetto a quando era negativa o ad alto arousal. Per spiegare questo dato, gli autori suggeriscono che gli stimoli positivi a basso arousal segnalino la presenza di un ambiente non allarmante e non appetitivo, e che questo faccia sì che la tendenza a rilevare automaticamente cambiamenti inaspettati a livello delle stimolazioni uditive sia meno importante da un punto di vista adattivo e di conseguenza presenti un decremento.

7.1.1. Il ruolo dell'amigdala

Surakka inoltre propone un modello secondo il quale l'amigdala costituisce un importante elemento di mediazione nel contesto dell'elaborazione degli stimoli sensoriali. Gli stimoli positivi a basso arousal infatti determinerebbero una diminuzione dell'attività dell'amigdala, che a sua volta causerebbe una diminuzione dell'attività del meccanismo deputato a rilevare automaticamente i cambiamenti inattesi a livello della corteccia uditiva. Il fatto che non sia stato registrato un aumento dell'MMN in presenza di stimoli negativi ad alto arousal viene spiegato ipotizzando che già la detezione degli stimoli incongruenti fosse massimamente attivata. In sintesi quindi, gli autori suggeriscono che il processamento degli stimoli uditivi sia influenzato dall'amigdala e, attraverso quest'ultima, dagli stimoli

emozionali di tipo visivo. Studi precedenti in effetti hanno dimostrato che l'amigdala ha un ruolo di primo piano nel processamento di informazioni connotate emotivamente. Uno studio condotto da Dolan (Dolan et al., 2001) per mezzo della fMRI (risonanza magnetica funzionale) ha mostrato come l'amigdala sia coinvolta nell'integrazione di informazioni visive e uditive legate all'espressione della paura: l'attivazione dell'amigdala e del giro fusiforme aumentano quando un volto esprimente paura viene presentato congiuntamente ad un messaggio verbale pronunciato in tono impaurito. Inoltre, l'amigdala riceve proiezioni da tutte le principali aree corticali sensoriali e presenta importanti proiezioni verso le aree visive ed uditive. In base a tutto ciò, è stato suggerito che l'amigdala possa avere una funzione di modulazione sugli stadi relativamente precoci del processamento sensoriale (LeDoux, 1995). In particolare essa, oltre a partecipare all'elaborazione unimodale degli stimoli, è coinvolta nel processamento multimodale degli stimoli che hanno valenza affettiva, sia propriamente in termini emozionali, sia più in generale in termini di valenza edonica: è ritenuta essere una struttura associativa multimodale, perchè riceve afferenze sia dalle diverse aree sensoriali specifiche sia dalle aree polimodali della corteccia temporale (O'Doerty, Rolls & Kringelbach, 2004 per una rassegna).

7.1.2. Valenza edonica e integrazione cross-modale

Relativamente alla *valenza edonica*, la letteratura indica che, nel decoding delle espressioni facciali, le emozioni con valenza negativa vengono processate principalmente nell'emisfero destro, mentre quelle con valenza positiva sono elaborate prevalentemente nell'emisfero sinistro (Davidson & Irwin, 1999). Pourtois e colleghi (Pourtois et al., 2005) hanno voluto indagare con uno studio PET se tale effetto di lateralizzazione è rilevabile anche in presenza di una stimolazione bimodale. Essi hanno utilizzato stimoli visivi (espressioni facciali di gioia e paura), stimoli uditivi (una parola bisillabica pronunciata in tono felice o impaurito) e stimoli bimodali congruenti risultanti dalla combinazione di quelli unimodali. L'originalità dell'esperimento consiste nell'utilizzo di una consegna indiretta o "nascosta" (*covert*):

ai soggetti è stato chiesto di valutare il genere del soggetto che esprimeva l'emozione. Come emerge dai risultati, rispetto agli stimoli unimodali, quelli bimodali attivano maggiormente un'area di convergenza situata nella corteccia temporale sinistra. Tale effetto è descritto anche da Calvert (Calvert et al., 2001). L'attivazione nello specifico coinvolge il giro mediotemporale sinistro (MTG), già precedentemente indicato come area di convergenza multimodale (Mesulam, 1998) e il giro fusiforme sinistro, la cui attivazione era stata rilevata da uno studio fMRI (Dolan, 2001). Inoltre, le analisi condotte separatamente per le due emozioni rivelano la presenza di aree di convergenza supplementari, situate prevalentemente nell'emisfero sinistro per gli stimoli bimodali della gioia e nell'emisfero destro per gli stimoli bimodali della paura. Questo dato indica l'esistenza di sostrati neurali di processamento cross-modale differenziati sulla base della valenza edonica dello stimolo emotivamente connotato. Infine, confermando i dati già presenti in letteratura, i ricercatori hanno evidenziato un'attivazione dell'amigdala per gli stimoli unimodali facciali e per gli stimoli bimodali esprimenti paura. Complessivamente, lo studio condotto dimostra che la presentazione congiunta di stimoli emozionali appartenenti a diverse modalità sensoriali (visiva e uditiva) porta all'attivazione di aree di convergenza eteromodali e che tale processo, data la natura implicita della consegna, ha carattere obbligato.

7.2. Un altro indice indiretto: la componente N1

Come anticipato, la *N1*, componente ERP sensoriale legata al processamento degli stimoli uditivi, è stata utilizzata, al pari della MMN, al fine di indagare le coordinate temporali del fenomeno di integrazione cross-modale che si verifica quando vengono presentati simultaneamente stimoli emotivi visivi e uditivi. In uno studio ERP Pourtois e collaboratori (Pourtois et al., 2000) hanno presentato ai soggetti coppie congruenti e incongruenti di stimoli uditivi (frammenti di 4 sillabe pronunciate in tono triste o arrabbiato) e di stimoli visivi (espressioni facciali tristi o arrabbiate presentate normalmente o capovolte), con la consegna di prestare

attenzione ai volti ignorando la voce. I risultati rivelano che l'informazione visiva influenza il processamento dello stimolo uditivo già dopo 110 ms dopo la stimolazione. Tale influenza si manifesta come un aumento dell'ampiezza di N1, come già rilevato in precedenti studi che hanno utilizzato stimoli non di tipo emotivo (Giard & Peronnet, 1999). Inoltre, l'integrazione avviene solo in presenza di stimoli congruenti, confermando così quanto rilevato a livello comportamentale. Si può dunque ipotizzare che l'elaborazione degli stimoli uditivi venga facilitata dalla presentazione di uno stimolo visivo congruente in termini di contenuto emotivo. L'integrazione non avviene invece quando il volto è capovolto. Questo perchè la rotazione di 180° impedisce il normale processo di elaborazione del volto che, come accennato precedentemente, presenta un percorso di processamento specifico e dedicato.

La maggior parte degli studi ha indagato il processamento cross-modale delle emozioni prendendo in esame l'*ampiezza* delle componenti ERP implicate, rilevando un incremento o decremento delle componenti unimodali precoci, come il picco uditivo N1 o il picco visivo P1, che hanno luogo intorno ai 100 ms di latenza nelle aree sensoriali specifiche (Calvert, Brammer & Iversen, 2000; Giard & Peronnet, 1999; Raij, Uutela & Hari, 2000; Sams et al., 1991). Infatti, l'incremento dell'attività nella corteccia modalità-specifica è considerato un fondamentale correlato elettrofisiologico della cross-modalità (de Gelder, 2000; Driver & Spence, 2000). Ad esempio, è stata segnalata un'attivazione amplificata a livello della corteccia uditiva durante la lettura del labiale (Calvert et al., 1997), del giro fusiforme e dell'amigdala durante la percezione di stimoli emozionali bimodali (Dolan, Morris & de Gelder, 2001) e delle aree tattili durante una stimolazione visuo-tattile (Macaluso, Frith & Driver, 2000). Nel complesso, l'integrazione cross-modale è segnalata da un'amplificazione sia a livello delle aree specifiche sia a livello di quei network corticali, come la corteccia parietale posteriore e il giro temporale mediale, che hanno natura multimodale (Mesulam, 1998).

7.3. L'indice di integrazione multimodale P2b

Solo pochi studi neuropsicologici hanno invece studiato la convergenza audio-visiva a contenuto emotivo tenendo conto del fattore temporale, rappresentato dalla *latenza*.

Tra questi, Pourtois e colleghi (2002) hanno dimostrato empiricamente che l'elaborazione degli stimoli emotivi audio-visivi comporta anche precise implicazioni in termini di latenza. Essi si sono posti infatti l'obiettivo di verificare se la presentazione di uno stimolo facciale può influenzare anche la latenza, oltre che l'ampiezza, dei processi di natura uditiva, prendendo però come oggetto di osservazione un intervallo temporale relativamente meno precoce di quello indagato dagli studi precedentemente citati. In sintesi, le analisi effettuate sulla componente uditiva mostrano l'esistenza di un picco positivo intorno ai 240 ms con una topografia posteriore, che gli autori denominano P2b. Gli autori ritengono che la P2b rappresenti un *indice di integrazione* tra la componente uditiva e quella visiva. Tale picco segue le componenti modalità-specifiche uditive N1 e P2 e precede il complesso amodale N2-P3, che si sa essere deputato all'elaborazione cognitiva ad un più tardivo stadio decisionale. Gli stimoli bimodali congruenti elicitano una P2b più precoce rispetto agli stimoli incongruenti, suggerendo che il processamento uditivo, in presenza di informazioni incongruenti, sia ritardato. Questi risultati sono in accordo con i precedenti dati comportamentali, che dimostrano un accorciamento dei tempi di processamento in presenza di stimoli multimodali congruenti. Inoltre, essi sono in accordo con quelli recentemente ottenuti da studi basati sulla risonanza magnetica funzionale (Calvert, Campbell & Brammer, 2000) o sulla magnetoencefalografia (Raij et al., 2000). Un'analisi di localizzazione della fonte effettuata durante l'intervallo temporale corrispondente alla P2b ha messo in evidenza un'implicazione della corteccia cingolata anteriore, che è implicata nel processamento della congruenza/incongruenza tra stimoli (McLeold & McDonald, 2000). I risultati sono coerenti con un coinvolgimento della corteccia cingolata anteriore nell'integrazione audio-visiva interno ai 220 ms.

Nel complesso, lo studio ancora una volta dimostra che l'integrazione cross-modale degli stimoli emotivi audio-visivi avviene nel corso del processamento percettivo (intorno ai 220 ms nella zona posteriore) e non ad uno stadio decisionale più avanzato. Pourtois e colleghi suggeriscono che già intorno ai 100 ms, nella fase *percettiva*, avvenga un incremento in termini di ampiezza delle componenti modalità-specifiche (de Gelder et al., 1999; Giard & Peronnet, 1999; Pourtois et al., 2000), successivamente seguito dalla comparsa di altre componenti, come la P2b, sensibili al *contenuto* dello stimolo audio-visivo. Solo in un secondo momento le informazioni avrebbero accesso a stadi cognitivi più avanzati di natura *decisionale*.

8. Gli studi sui casi clinici

Un contributo fondamentale per la comprensione del decoding intersensoriale delle emozioni giunge dagli studi condotti su casi clinici, che ne mettono in risalto l'importante funzione adattiva: ad esempio, in un recente studio condotto su un paziente con grave deficit della localizzazione uditiva dovuto a lesione, Bolognini, Rasi e Ladavas (2005) hanno dimostrato che la contemporanea comparsa di uno stimolo visivo che viene presentato nella medesima posizione spaziale dello stimolo target uditivo migliora fortemente la localizzazione del suono.

Alcuni studi clinici si sono rivelati utili anche per chiarire le funzioni delle strutture corticali implicate nel processamento cross-modale. Tra di essi, uno studio condotto con la risonanza magnetica da Taylor e Brugger (2005) sul caso di un paziente affetto da sclerosi multipla e vittima di allucinazioni audio-visive ha portato gli autori ad ipotizzare che tali allucinazioni fossero legate ad un deficit nella regolazione dell'attività di integrazione cross-modale localizzata nel collicolo superiore e nel solco temporale superiore.

8.1. Il fenomeno del blindsight

Gli studi condotti su casi clinici si sono in particolar modo rivelati utili per indagare il ruolo della consapevolezza nell'integrazione intersensoriale delle

informazioni emotive di natura uditiva e visiva. A tal proposito, de Gelder e colleghi (de Gelder, Pourtois & Weiskrantz, 2002) hanno sottoposto ad uno dei classici esperimenti sul processamento cross-modale delle emozioni due pazienti che presentavano blindsight (emianopia) unilaterale. Essi, a causa di una lesione alla corteccia striata (V1), erano in grado di discriminare le espressioni del volto ma senza essere consapevoli di percepirle. Il processamento cosciente degli stimoli emotivi, che è di tipo cortico-corticale, coinvolge, oltre alla corteccia V1, la corteccia fusiforme e quella orbitofrontale, oltre a provocare un aumento dell'attivazione dell'amigdala destra. La percezione non consapevole coinvolge invece l'amigdala sinistra (Morris, Öhman & Dolan, 1998), il pulvinar e il collicolo superiore (Morris, Öhman & Dolan, 1999), che sono implicati in un circuito sottocorticale di elaborazione delle espressioni facciali. Queste strutture, nei due pazienti che hanno partecipato all'esperimento, erano intatte, permettendo così che venisse conservata l'elaborazione implicita degli stimoli emotivi. Finora il fenomeno del blindsight affettivo era stato studiato soltanto utilizzando come stimolo le espressioni facciali. De Gelder e colleghi hanno introdotto un secondo tipo di stimolo visivo, mostrando ai pazienti anche delle scene a contenuto emotivo. Nel presente esperimento, gli autori si sono chiesti se l'integrazione audiovisiva in soggetti con blindsight emotivo avvenga solo in presenza di abbinamenti naturali (volto della paura – voce della paura), o anche in presenza di abbinamenti semantici (immagine paurosa – voce della paura). Gli autori hanno ipotizzato che, se quest'ultimo caso si verifica, bisogna supporre che i circuiti sottocorticali compensino l'assenza di quelli corticali; nel caso in cui invece l'integrazione non abbia luogo, allora bisogna ipotizzare che la percezione cross-modale delle coppie immagine – voce richieda necessariamente l'intervento di circuiti di ordine superiore, deputati all'elaborazione delle proprietà semantiche che essi condividono. Propendendo per la seconda possibilità, gli autori hanno ipotizzato nello specifico che, nel caso delle coppie naturali, si verificasse un decremento dell'ampiezza di N1 in presenza di coppie incongruenti e che, nel caso delle coppie semantiche, tale decremento si verificasse solo quando gli stimoli erano presentati all'emisfero intatto e quindi processati consapevolmente. Essi quindi, con il loro

esperimento, si aspettavano che la presentazione di un'immagine all'emisfero danneggiato (assenza di consapevolezza) non potesse interferire con il processamento uditivo e che quindi non ci fosse integrazione, per il fatto che il circuito subcorticale in questo caso non sarebbe sufficiente. Dai risultati emerge che, in assenza di percezione consapevole (presentazione nel campo visivo danneggiato), la presentazione dello stimolo visivo influenza il processamento della voce solo nel caso in cui lo stimolo visivo sia costituito da un volto, come suggerito dall'analisi dell'indice ERP N1, che rivela un decremento nelle coppie incongruenti. L'effetto della percezione visiva su N1, indipendentemente dal lato della lesione, presenta una lateralizzazione, essendo maggiormente evidente nell'emisfero destro. Ciò è in linea con il dato che soprattutto tale emisfero è implicato nell'elaborazione delle componenti prosodiche (Ross, 2000). Per spiegare i risultati, viene ipotizzato che quando la corteccia visiva primaria è danneggiata, alcune strutture che ricevono afferenze dirette dalla retina, come il collicolo superiore e il pulvinar, possano compensare fino ad un certo punto la mancata attività di V1, e che tuttavia non possano compensare l'assenza di alcune proiezioni di feedback che mettono in collegamento V1 e aree corticali anteriori garantendo la percezione combinata audio-visiva (Lamme, 2001). La percezione congiunta di stimoli affettivi uditivi e visivi abbinati sulla base del contenuto semantico richiede l'intervento di circuiti corticali deputati all'elaborazione semantica che coinvolgono V1 così come aree corticali anteriori di alto livello. Questo suggerisce che invece, nel caso dell'abbinamento volto-voce, l'intervento dei circuiti corticali non sia del tutto cruciale ai fini dell'integrazione. Gli autori concludono che il riconoscimento delle emozioni a partire dal volto, o dalla voce o dall'integrazione tra i due può avvenire baypassando la coscienza e che ciò, molto probabilmente, è dovuto alla rilevanza che le emozioni rivestono da un punto di vista adattivo.

8.2. La prosopagnosia

Un risultato simile (de Gelder et al., 2000) è stato ottenuto indagando l'integrazione cross-modale di stimoli emotivi in una paziente con prosopagnosia

dovuta a lesione bilaterale dei lobi occipitali, del tutto incapace di riconoscere consapevolmente, sulla base del volto, l'identità e l'espressione emotiva. La paziente non presentava invece problemi nel riconoscere le emozioni espresse attraverso il tono di voce. La prosopagnosia consiste infatti nella compromissione della capacità di riconoscere i volti e di identificarne l'identità e l'espressione (Tranel, Damasio & Damasio, 1995). Tuttavia studi che hanno utilizzato metodi elettrofisiologici come la rilevazione della conduttanza cutanea (Tranel & Damasio, 1987) o come la registrazione dei potenziali evocati (Renault et al., 1989) hanno evidenziato come i pazienti affetti da prosopagnosia siano in grado di riconoscere in modo implicito e latente l'identità, a partire dall'osservazione del volto. Nessuno studio precedente aveva invece studiato l'esistenza di un riconoscimento latente dell'espressione del volto. La ricerca ha indagato il riconoscimento implicito dell'espressione emotiva andando a verificare se, presentando contemporaneamente un'espressione facciale e una parola pronunciata in tono emotivamente connotato, esistesse un'interferenza tra le due modalità sensoriali. Gli stimoli utilizzati a questo scopo esprimevano gioia o tristezza ed erano abbinati in coppie congruenti o incongruenti. In una prima fase, volta ad indagare l'effetto del tono di voce sul riconoscimento del volto, la paziente veniva invitata a ignorare la voce e di identificare il volto come felice o triste. A differenza di quanto rilevato sui soggetti normali, che presentavano un effetto di interazione tra volto e voce, il giudizio della paziente, nonostante la consegna, era interamente basato sul tono di voce. Un risultato molto diverso è stato ottenuto nella seconda fase dell'esperimento, che esplorava l'effetto del volto sul riconoscimento della voce e in cui la paziente veniva invitata a ignorare il volto e ad etichettare la voce come felice o triste. In questo caso è emerso un effetto cross-modale, tale per cui l'espressione del volto aveva un sistematico impatto sulla valutazione del tono di voce. Nel complesso, i risultati indicano l'esistenza di un riconoscimento implicito dell'espressione del volto. Inoltre, contribuiscono a validare l'ipotesi della de Gelder che il processo di integrazione cross-modale abbia carattere obbligato e che avvenga in una fase percettiva precoce: il fatto che la paziente non percepisca consapevolmente il volto esclude infatti che l'integrazione avvenga ad uno stadio

cognitivo decisionale. Diverse ipotesi sono state formulate per spiegare il riconoscimento latente: in primo luogo, esso potrebbe scaturire da una forma di rappresentazione degradata, impoverita e che quindi non ha la possibilità di essere concettualizzata a livello cosciente (Farah, O'Reilly & Vecera, 1993); inoltre, tale riconoscimento latente potrebbe indicare l'esistenza di due sistemi distinti di processamento del volto, l'uno ventrale dedicato alle rappresentazioni manifeste e l'altro dorsale dedicato a quelle latenti (Bauer, 1984). A questo proposito, è significativo il fatto che la paziente presentava una compromissione della via ventrale (occipitotemporale) a fronte della conservazione di quella dorsale. Infine, è stata ipotizzata l'esistenza di due diversi tipi di processamento, qualitativamente differenti e corrispondenti all'elaborazione implicita ed esplicita dei volti. Nella paziente sarebbe conservato solo il primo tipo di elaborazione, che tuttavia non prevede l'accesso alla consapevolezza. In effetti, recenti studi indicano che gran parte del processamento delle emozioni (LeDoux, 1996) e, nello specifico, delle espressioni facciali (Morris, Öhman & Dolan, 1998) avviene al di fuori della consapevolezza. Sarebbero necessarie ulteriori indagini per capire quale di queste ipotesi meglio rende conto del fenomeno osservato.

8.3. Sistemi multipli di decodifica delle emozioni

Recentemente, è stata proposta l'ipotesi che esistano diversi sistemi indipendenti di riconoscimento delle emozioni, che si differenziano sia per il tipo di modalità sensoriale implicata (visiva, uditiva o audio-visiva) sia - nel caso della modalità visiva - per la natura dello stimolo (dinamico o statico). Oggi si tende ad evidenziare il contributo di strutture bilaterali come l'amigdala, il giro cingolato e i gangli basali, oltre alla corteccia prefrontale nel processamento delle emozioni (Adolphs, 2002; Phillips et al., 2003). Diversi lavori hanno tuttavia messo in luce anche l'esistenza di una specializzazione emisferica (Borod, 1993; Tranel et al., 2002). Inoltre, un importante ruolo è ricoperto dalla corteccia somatosensoriale destra, che sembra essere fondamentale nella comprensione delle espressioni emotive facciali, perchè permette al soggetto di accedere alle qualità dell'espressione osservata "come

se" fosse la propria (Adolphs et al., 2003). Nei pazienti neurologici, il deficit nel processamento delle espressioni emotive coinvolge soprattutto specifiche categorie di emozioni, più frequentemente quelle negative, come paura, disgusto e tristezza. Proprio questa osservazione ha suggerito la possibilità che esistano diversi sistemi specializzati di processamento (Adolphs & Tranel, 2004).

Un interessante studio condotto da McDonald e Saunders (2005) su pazienti con severo danno cerebrale traumatico (*traumatic brain injury*, TBI) ha portato supporto a questa ipotesi. I danni cerebrali traumatici consistono in ampie lesioni delle aree frontali e temporali, oltre che delle strutture limbiche e di altre strutture ad esse associate. Possono comportare la disconnessione tra le strutture limbiche e le aree somatosensoriali, disconnessione che spesso è causa di deficit nel riconoscimento delle emozioni altrui (Green et al., 2004). In effetti, la maggior parte dei pazienti con TBI presenta evidenti difficoltà quando viene chiesto di decodificare le emozioni sulla base delle diverse modalità sensoriali (McDonald & Flanagan, 2004). In particolare, la lesione dei lobi frontali e parietali e delle strutture limbiche causa delle difficoltà nel riconoscimento delle espressioni sia facciali sia vocali delle emozioni (Adolphs, 2002). Bisogna poi osservare che spesso questi pazienti presentano deficit maggiori quando viene loro chiesto di riconoscere espressioni facciali statiche anziché dinamiche. Questo dato ha portato Adolphs (Adolphs et al., 2003) ad ipotizzare che esistano due differenti processi e che, in particolare, le espressioni statiche richiedano il contributo del sistema limbico e della corteccia prefrontale associata, a differenza delle espressioni dinamiche che sarebbero invece processate a livello della corteccia parietale.

McDonald e Saunders, sottoponendo i pazienti ad un compito di riconoscimento delle emozioni, hanno rilevato che essi presentavano una competenza deficitaria nel decoding delle emozioni, ma con alcune interessanti specificità. In particolare, gli stimoli facciali dinamici erano normalmente riconosciuti, a differenza degli stimoli facciali statici, degli stimoli emotivi di natura uditiva e, soprattutto degli stimoli audio-visivi, riconosciuti in modo altamente deficitario. Ciò supporterebbe l'ipotesi di sistemi di processamento indipendenti e

qualitativamente distinti. Nello specifico, gli autori suggeriscono che l'elaborazione degli stimoli visivi dinamici sia localizzata principalmente nelle aree parietali (Adolphs et al., 2003), non compromesse dalla lesione, e che invece quella degli stimoli facciali statici avvenga ad opera delle aree fronto-temporali danneggiate e delle strutture limbiche ad esse correlate. Rispetto agli stimoli uditivi, è possibile che il deficit sia dovuto al fatto che i pazienti elaborano il contenuto linguistico, a scapito dell'espressione emotiva. In effetti, tali pazienti normalmente tendono ad interpretare gli enunciati in modo molto letterale, tralasciando di effettuare delle inferenze (McDonald & Flanagan, 2004). Infine, per quanto riguarda gli stimoli audio-visivi, è possibile che, nonostante la compresenza di entrambe le modalità, i pazienti si focalizzino su una sola di esse e che non utilizzino le strategie di processamento normalmente impiegate nel riconoscimento delle informazioni emotive bimodali.