INTERAZIONE MULTIMODALE
Errori nei sistemi multimodali

Si verifica un errore quando l'utente non raggiunge l'obiettivo desiderato e nessuna coincidenza può esserne responsabile. La robustezza agli errori influenza significativamente l'accettazione di un sistema da parte dell'utente («user acceptance»), quindi è necessario tenere conto della robustezza agli errori. Generalmente, gli errori non possono essere mai evitati completamente, quindi sono molto importanti le procedure di gestione degli errori passiva (*a-priori error avoidance*) e attiva (*a-posteriori error avoidance*).

Le risposte indesiderate del sistema dovute a un funzionamento difettoso e a errori interni al sistema devono essere evitate il più possibile. La risoluzione degli errori deve essere efficiente, trasparente e robusta.

Il seguente scenario mostra una situazione soggetta a errori: un utente alla guida di un veicolo vuole cambiare la stazione radio utilizzando il comando vocale «ascolta la radio alfa»; il sistema di riconoscimento interpreta in modo errato il comando con «spegni la radio» e la radio viene spenta.

INTERAZIONE MULTIMODALE
Errori nei sistemi multimodali

Gli errori possono essere classificati in *user specific errors* e *system specific errors*.

User Specific Errors

L'utente che interagisce con il sistema può essere una fonte di errore. Una possibile classificazione è la seguente:

- Errori associati allo *skill-based level* (esempio: scivolamento del dito da un pulsante).
- Errori associati al *rule-based level* (esempio: utilizzo di un comando vocale valido non permesso nella situazione corrente ma ammesso in un'altra situazione).
- Errori associati al *knowledge-based level* (esempio: utilizzo di un comando vocale che non è noto al sistema).

Per quanto riguarda gli errori associati allo *skill-based level*, tale livello comprende azioni di routine fluide, automatizzate e altamente integrate che avvengono senza attenzione o controllo. Le funzioni umane sono governate da schemi memorizzati di istruzioni pre-programmate rappresentate come strutture in un dominio spazio-temporale. Gli errori a questo livello sono legati alla variabilità intrinseca della forza, dello spazio o della coordinazione temporale.

INTERAZIONE MULTIMODALE
Errori nei sistemi multimodali

Sporadicamente, l'utente controlla se l'azione avviata si svolge come previsto e se il piano per raggiungere l'obiettivo è ancora adeguato. I pattern di errore associati allo *skill-based level* sono errori di esecuzione o di memoria che derivano da disattenzione o eccessiva attenzione dell'utente.

Per quanto riguarda gli errori associati al *rule-based level*, l'utente viola regole primarie memorizzate. Gli errori sono tipicamente associati al richiamo errato di procedure o all'errata classificazione delle situazioni che porta all'applicazione della regola sbagliata.

Per quanto riguarda gli errori associati al *knowledge-based level* l'utente applica le conoscenze e i processi analitici memorizzati in situazioni nuove, in cui le azioni devono essere pianificate on-line. Gli errori a questo livello derivano da limitazioni delle risorse (razionalità limitata) e da una conoscenza incompleta o errata.

INTERAZIONE MULTIMODALE
Errori nei sistemi multimodali

System Specific Errors
Il sistema può essere una fonte di errore. Una possibile classificazione è la seguente:

- Errori associati al *recognition level* (esempi: interpretazione errata, riconoscimento errato di un input corretto dell'utente, errata attivazione intrinseca al sistema di un riconoscitore vocale in una conversazione).

- Errori associati al *processing level*: problemi di temporizzazione o risultati di riconoscimento discordanti dei differenti sistemi di riconoscimento monomodali (esempio: il risultato del riconoscimento del parlato differisce da quello relativo al riconoscimento dei gesti) possono causare errori associati a tale livello.

- Errori associati al *technical level*: il malfunzionamento dei componenti del sistema può portare a errori associati a tale livello.

INTERAZIONE MULTIMODALE
Errori nei sistemi multimodali

Error Avoidance

Alcune linee guida per cercare di evitare gli errori nelle interfacce multimodali sono:

- *Cercare la coerenza*: contesti simili dovrebbero essere caratterizzati da sequenze simili di azione, così come da una terminologia identica, ad esempio riguardo a menu o richieste. La coerenza per le interfacce multimodali è caratterizzata da due aspetti principali. In primo luogo, all'interno di ogni singola situazione tutti i comandi devono essere gli stessi per tutte le modalità. In secondo luogo, all'interno di ogni singola modalità tutti i comandi simili in situazioni diverse devono essere uguali. Ad esempio, il comando per tornare al menu principale dovrebbe essere lo stesso in tutti i sottomenu e tutti i sottomenu devono essere accessibili da tutte le modalità con lo stesso comando (esempio: il nome del menu sul pulsante è identico al comando vocale associato).

INTERAZIONE MULTIMODALE
Errori nei sistemi multimodali

Error Avoidance

Alcune linee guida per cercare di evitare gli errori nelle interfacce multimodali sono:

- *Fornire un feedback informativo*: ad ogni fase dell'interazione deve essere fornito un feedback del sistema. Questo feedback dovrebbe essere modesto per le azioni frequenti e secondarie e rilevante per azioni poco frequenti o importanti. Le interfacce multimodali hanno la possibilità di sfruttare il vantaggio associato alle differenti modalità di output. Il feedback dovrebbe essere fornito nella modalità corrispondente alla modalità di input utilizzata (esempio: ad un comando vocale ci si aspetta un feedback acustico del sistema).

- *Ridurre il carico di memoria a breve termine*: l'elaborazione delle informazioni da parte dell'uomo è limitata nella memoria a breve termine. Ciò richiede un sistema di dialogo semplice. I sistemi multimodali dovrebbero utilizzare le modalità in modo da ridurre il carico di memoria dell'utente. Ad esempio, la selezione di un oggetto è facile puntando su di esso, ma è difficile con l'utilizzo dei comandi vocali.

INTERAZIONE MULTIMODALE
Errori nei sistemi multimodali

Error Avoidance

Alcune linee guida per cercare di evitare gli errori nelle interfacce multimodali sono:

- *Sincronizzare modalità multiple*: l'interazione attraverso il parlato dipende fortemente dal tempo; l'interazione visiva è spaziale. Occorre quindi una sincronizzazione di queste modalità di input (esempio: selezionare oggetti puntando il dito su di essi e avviare la selezione mediante l'utilizzo del parlato («Seleziona questo oggetto»)).

Tali linee guida sono dedotte dalle regole di progettazione delle interfacce utente. Tali regole derivano dall'esperienza e si possono applicare in molti sistemi interattivi. Esse non portano direttamente ad evitare gli errori, ma semplificano l'interazione dell'utente con il sistema.

Esistono anche molti altri aspetti progettuali e linee guida per evitare errori. Rispetto alle interfacce monomodali, le interfacce multimodali possono migliorare la prevenzione degli errori, consentendo all'utente di scegliere liberamente quale modalità di input utilizzare. In tal modo, l'utente è in grado di selezionare la modalità di input più comoda ed efficiente

per raggiungere il suo obiettivo. Inoltre, se l'interazione avviene tramite l'utilizzo di più di un canale, gli errori tipici di una singola modalità possono essere compensati combinando tutti i dati in ingresso. Quindi, fornire più di una modalità di input aumenta la robustezza del sistema e aiuta a prevenire gli errori.

Error Resolution

In caso di errori (errori del sistema o dell'utente), il sistema cerca di risolvere i problemi iniziando dialoghi con l'utente. Le strategie di risoluzione degli errori si differenziano in strategie di dialogo *single-step* e *multi-level*. Nel contesto di una strategia *single-step*, viene generata una richiesta dal sistema, alla quale l'utente può rispondere con un input singolo. Nel contesto di una strategia *multi-level* viene avviata una complessa procedura di dialogo, in cui l'utente viene guidato passo dopo passo nel processo di risoluzione dell'errore. In particolare, il secondo approccio offre la possibilità di adattamento all'utente corrente e alla situazione ambientale momentanea.

INTERAZIONE MULTIMODALE
Errori nei sistemi multimodali

Ad esempio, possono essere differenziate le seguenti strategie di gestione dell'errore:

- *Warning*
- *Richiesta di ripetizione dell'ultimo input*
- *Richiesta di cambiare la modalità di input*
- *Offerta di modalità di input alternative*

Queste strategie si differenziano per caratteristiche come l'inizializzazione dell'avviso di errore, la rilevanza del contesto, le caratteristiche individuali dell'utente, la complessità della strategia di errore e l'inclusione dell'utente. La scelta della strategia di dialogo da utilizzare dipende principalmente da parametri del contesto e dallo stato attuale del sistema (esempi: la modalità di input eventualmente scelta, lo stato dell'applicazione).

In accordo a quanto descritto in precedenza sulla fusione multimodale, la componente di gestione dell'errore si colloca nell'*integration level* di un'architettura multimodale.

INTERAZIONE MULTIMODALE
Errori nei sistemi multimodali

Il processo di gestione dell'errore è composto da quattro fasi:

- estrazione delle features dell'errore;
- analisi dell'errore;
- classificazione dell'errore;
- risoluzione dell'errore.

In primo luogo, un modulo estrae continuamente alcune features dal flusso di messaggi in arrivo e verifica un potenziale errore. Poi, si possono classificare i singoli pattern di errore. In questa fase del processo di gestione degli errori, vengono determinati i tipi di errore risultanti. Successivamente, per la selezione di una strategia di dialogo dedicata, vengono analizzati i parametri del contesto attuale e i tipi di errore. Valutando i risultati di tale analisi viene scelta la strategia con la maggiore plausibilità; essa aiuta a risolvere il problema in modo confortevole per l'utente.

RICONOSCIMENTO DELLE EMOZIONI

Emozioni associate al parlato e alle espressioni facciali

Anche se l'utilizzo di tasti può ormai essere sostituito da forme di comunicazione più naturali, come il parlare e il gesticolare, la comunicazione uomo-macchina continua a sembrare in qualche modo impersonale, insensibile e meccanica.

Se si confronta la comunicazione uomo-macchina con la comunicazione umano-umano, ci si può accorgere che a volte manca l'informazione supplementare percepita dall'uomo sullo stato emotivo dell'interlocutore.

Queste informazioni emotive influenzano fortemente le informazioni esplicite, come riconosciuto dagli odierni sistemi di comunicazione uomo-macchina, e, con una comunicazione sempre più naturale, ci si potrà aspettare il loro rispetto. Nella progettazione delle interfacce uomo-macchina l'inclusione di questo canale implicito sembra quindi obbligatoria.

RICONOSCIMENTO DELLE EMOZIONI

Emozioni associate al parlato e alle espressioni facciali

Alcuni dei campi pionieri di applicazione del riconoscimento delle emozioni sono stati:

- Call center: un cliente infastidito che attualmente viene gestito da un robot viene trasferito ad un operatore umano.
- Software ludici che rilevano le bugie, il livello di stress o il livello di benessere di un interlocutore telefonico.

Oltre a questi, campi di applicazione più generali sono la migliore comprensione dell'intenzione dell'utente, l'adattamento emotivo nella comunicazione (esempio: adattamento dei parametri acustici per la sintesi vocale se l'utente sembra triste), l'osservazione del comportamento (esempio: rilevare se un passeggero di un aereo sembra aggressivo), la misurazione oggettiva delle emozioni (esempio: utilizzo di tale informazione come linea guida per i medici), la trasmissione di emozioni (esempio: invio di immagini di risate o pianti all'interno di e-mail testuali), il recupero di contenuti multimediali legati alle emozioni (esempio: individuazione dei momenti salienti in un evento sportivo) e i prodotti multimediali sensibili alle emozioni (esempio: mirino tremante nei videogiochi se il giocatore sembra nervoso).

RICONOSCIMENTO DELLE EMOZIONI

Emozioni associate al parlato e alle espressioni facciali

L’emozione umana è osservabile mediante modalità diverse.

I primi tentativi di riconoscimento automatico hanno applicato misure invasive, ad esempio della conduttività cutanea, della frequenza cardiaca o della temperatura. Sebbene lo sfruttamento di queste fonti di informazione fornisce una stima affidabile dell’emozione esistente, spesso viene percepito come scomodo e innaturale, in quanto l’utente deve essere collegato o almeno deve stare a contatto con un sensore.

Alcuni studi affermano che gli umani comunicano per il 55% visivamente (attraverso il linguaggio del corpo), per il 38% attraverso il tono della voce e per il 7% attraverso il parlato. Tenendo conto di ciò, l’approccio più promettente sembra chiaramente essere una combinazione di queste fonti. Tuttavia, in alcuni sistemi e situazioni può essere disponibile solo una di esse.

RICONOSCIMENTO DELLE EMOZIONI

Emozioni associate al parlato e alle espressioni facciali

È interessante notare che, contrariamente alla maggior parte delle altre modalità, il parlato consente all'utente di controllare la quantità di emozioni mostrate. Il riconoscimento delle emozioni basato sul parlato fornisce risultati accettabili. Tuttavia, le informazioni visive contribuiscono ad ottenere una stima più robusta.

Dal punto di vista economico, il microfono come sensore è oggi un hardware standard in molti sistemi uomo-macchina, e anche le camere sono sempre più diffuse, come nei telefoni cellulari della generazione attuale. In tal senso, sembra ragionevole adottare una strategia di riconoscimento delle emozioni basata sulla fusione di informazioni acustiche, linguistiche e visive.

RICONOSCIMENTO DELLE EMOZIONI

Emozioni associate al parlato e alle espressioni facciali

Per il riconoscimento delle emozioni occorre un modello associato alle emozioni.

Esistono due approcci principali riguardo alla modellizzazione delle emozioni.

Nel primo approccio l'emozione viene rappresentata mediante due o tre assi ortogonali:

- *arousal* o *activation* (attivazione, stimolo), rispettando la prontezza nell'intraprendere un'azione;
- *valence* o *evaluation* (valutazione), considerando un'attitudine positiva o negativa;
- *control* o *power* (controllo), analizzando il dominio o la sottomissione del parlante.

Tale approccio offre una buona base per la sintesi delle emozioni, ma esso è troppo complesso per scenari associati ad applicazioni pratiche.

Il secondo approccio (quello più diffuso) si basa sulla classificazione delle emozioni tramite un insieme limitato di etichette associate alle emozioni (si ricordi quanto descritto in precedenza per le emozioni associate alle espressioni facciali).

RICONOSCIMENTO DELLE EMOZIONI

Emozioni associate al parlato e alle espressioni facciali

Per il riconoscimento delle emozioni occorre un modello associato alle emozioni.

Per addestrare e testare i sistemi di riconoscimento, è necessario disporre di database relativi a campioni delle emozioni. Tali database dovrebbero fornire un comportamento emozionale spontaneo e realistico.
La qualità dei campioni deve garantire una qualità audio e video da laboratorio, ma per l'analisi della robustezza al rumore possono essere richiesti anche campioni con condizioni di rumore di fondo note. Un database deve inoltre essere composto da un numero elevato di campioni equamente distribuiti per ogni emozione; sono necessari campioni associati alla singola persona e a molte persone diverse. Queste persone dovrebbero fornire un modello chiaro che tenga conto dei generi, delle fasce d'età e dei contesti etici. Per quanto riguarda la variabilità, le frasi pronunciate dovrebbero avere contenuti, lunghezze o addirittura lingue diverse. Per questo motivo, l'assegnazione univoca dei campioni raccolti a classi di emozioni è particolarmente difficile.

RICONOSCIMENTO DELLE EMOZIONI

Emozioni associate al parlato e alle espressioni facciali

Inoltre, i test di percezione condotti su persone umane sono molto utili: può essere difficile valutare con certezza le proprie emozioni.

Un database dovrebbe essere reso disponibile, il che però potrebbe rappresentare un problema considerando la mancanza di consenso sulle classi di emozioni utilizzate e sulla privacy delle persone coinvolte nel test.

Esistono diversi metodi per creare un database, con punti di forza differenti. I metodi principali sono la recitazione o l'evocazione di emozioni in scenari di test, l'osservazione nascosta o consapevole a lungo termine e l'uso di clip di contenuti pubblici. Tuttavia molti database utilizzano emozioni recitate che consentono di soddisfare i requisiti eccetto la spontaneità, poiché si dubita che le emozioni recitate siano in grado di rappresentare le vere caratteristiche delle emozioni stesse.

Esempi di database associati a espressioni facciali sono stati già descritti in precedenza. Esempi di database associati al parlato sono: Danish Emotional Speech Database (CEICES), Berlin Emotional Speech Database (EMO-DB) e AIBO Emotional Speech Corpus (AEC).

RICONOSCIMENTO DELLE EMOZIONI
Emozioni associate al parlato e alle espressioni facciali

Informazione Acustica
Gli obiettivi predominanti, oltre all'alta affidabilità, sono l'indipendenza dal parlante, l'indipendenza dalla lingua parlata, l'indipendenza dal contenuto e l'indipendenza dal rumore di fondo.

Oltre al modello delle emozioni e alle dimensioni e alla qualità del database, i seguenti aspetti influenzano fortemente la qualità di un sistema di riconoscimento:

- Acquisizione del segnale
- Pre-elaborazione
- Selezione delle features
- Metodo di classificazione
- Integrazione nell'interfaccia e nel contesto di applicazione

RICONOSCIMENTO DELLE EMOZIONI

Emozioni associate al parlato e alle espressioni facciali

Informazione Acustica

Per stimare l'emozione di un utente in base alle informazioni acustiche è necessario selezionare con attenzione le features adatte. Queste devono contenere informazioni sull'emozione trasmessa, ma devono anche adattarsi al modello scelto mediante algoritmi di classificazione.

Dopo aver individuato le features, potrebbe essere necessario ridurre la loro dimensione. Per tale scopo si può utilizzare la Principal Component Analysis (PCA).

La scelta del metodo di classificazione è influenzata da diversi fattori. Sono da considerare lo scopo di ottenere elevati tassi di riconoscimento e un'elevata efficienza ma occorre anche tenere conto di aspetti economici. Inoltre è opportuno provare a garantire un'agevole integrazione nel framework dell'applicazione considerata.

RICONOSCIMENTO DELLE EMOZIONI

Emozioni associate al parlato e alle espressioni facciali

Informazione Linguistica

Anche l'interpretazione del contenuto del parlato ha un ruolo importante nel riconoscimento delle emozioni. In alcuni studi di psicologia si sostiene l'esistenza di una connessione tra i vocaboli utilizzati dal parlante e l'emozione associata.

Poiché l'espressione delle emozioni da parte del parlante consiste nell'uso di alcune frasi che potrebbero essere combinate con affermazioni significative nel contesto del dialogo, è necessario un approccio con capacità di individuare le informazioni rilevanti dal punto di vista emotivo.

Si consideri come esempio la frase «Potrebbe per favore dirmi molto di più su questo fantastico campo di ricerca». La percentuale di parole emotive dipende chiaramente dal contesto applicativo e dalla personalità del parlante, ma nella maggior parte dei casi sarà molto bassa. Ci si può quindi chiedere se le informazioni linguistiche possano essere sufficienti per il riconoscimento delle emozioni. Tuttavia, la loro integrazione ha dimostrato un chiaro aumento delle prestazioni dei sistemi di riconoscimento.

RICONOSCIMENTO DELLE EMOZIONI

Emozioni associate al parlato e alle espressioni facciali

Informazione Visiva

Per uno spettatore umano l'aspetto visivo di una persona fornisce molte informazioni sul suo stato emotivo. Si possono quindi identificare diverse fonti, come la postura del corpo (eretta, dinoccolata), i gesti delle mani (agitarsi, braccia conserte), i gesti della testa (annuire, inclinarsi) e, soprattutto in una conversazione diretta e ravvicinata, la varietà di espressioni facciali (gioia, sorpresa, rabbia, tristezza, …).

Fusione delle informazioni

Per quanto riguarda la fusione delle informazioni acustiche, linguistiche e visive, possono essere adottate le tecniche di fusione multimodale precedentemente descritte. Uno dei problemi è la sincronizzazione del video e dei flussi audio (acustici e linguistici). Soprattutto se l'audio è classificato a livello globale di parola o di enunciato, può diventare difficile trovare i segmenti video che corrispondono a queste unità.

Riferimenti Bibliografici

[1] Kraiss, K. -F. (2006). Advanced Man-Machine Interaction: Fundamentals and Implementation. Springer-Verlag Berlin Heidelberg. ISBN-10: 3-540-30618-8

[2] Paramartha Dutta, Asit Barman (2020). Human Emotion Recognition from Face Images. Springer Singapore. ISBN: 978-981-15-3883-4