Interaction Design and Architecture(s) Journal - IxD&A, N. 5-6, 2009, pp. 107-110

Challenging emotivity: the voice

D. Conflitti, R. Santoboni
C. Giovannella
ScuolaIAD & ISIM_lab
University of Rome Tor Vergata
via della ricerca scientifica, 1
I-00133 Rome
info@mifav.uniroma2.it

A. Paoloni
Fondazione Ugo Bordoni (FUB)
University of Rome Tor Vergata
via Baldassarre Castiglioni, 59
I-00142 Rome
pao@fub.it

ABSTRACT

We present a new tool and test procedure - based on the use of the finite state emotional model described by the Plutchick "flower" - designed to measure the ability of the human to recognize the emotion carried by the voice when the italian language is used. Using the no-sense sentences contained in the corpus EMOVO we have been able to that in natural conditions the human ability to recognize the emotional states using only the vocal channel are quite poor (30% in average). The recognition rate depends on many factors, among them the sex and the peculiarities of the speaker, the specific emotion, the specific contest.

Categories and Subject description

H.1.2 [Human factors]; H.5 INFORMATION INTERFACES AND PRESENTATION; H.5.1 **[User Interfaces]**.

General Terms

Human Factors, Measurements

Keywords

Emotivity, affective computing, emotional colored voice corpora, multimodal interaction, human factors evaluation

1. Introduzione

L'interazione naturale e la progettazione per l'esperienza non possono prescindere dal livello emotivo dell'interazione [1]. Non a caso l'interesse per "l'affective computing" [2] è da anni in costante crescita e, di recente, ha valicato il confine rappresentato dalla sfera emotiva del singolo per arivare ad indagare gli aspetti emotivi originati dall'interazione sociale [3]. Nonostante esistano da moltissimi anni metodi consolidati per la rilevazione dell'emotività (basti pensare a tutte le grandezze fisiche che compongono il bouquet del biofeedback), non è da tantissimo che l'attenzione generale si è spostata verso una rilevazione non invasiva delle tracce emotive per mezzo di webcam, microfoni, dell'analisi delle relazioni e dei testi. Questo è stato reso possibile: a) dal recente sviluppo delle infrastrutture di rete su cui insistono un numero sempre crescente di applicazioni di "social networking"; b) da una penetrazione di apparati sufficientemente "smart", ormai numericamente equivalente al numero di possessori di computer e/o cellulari.

Nonostante gli enormi progressi tecnologici e algoritmici la rilevazione non invasiva del livello emotivo dell'interazione risulta decisamente complessa, poiché non banali sono le modalità attraverso le quali gli umani percepiscono le emozioni. L'uomo, infatti, identifica lo stato emotivo di una persona, a partire dall'analisi combinata di indicatori ricavati dai segnali provenienti da un insieme di canali percettivi paralleli, attraverso tecniche di vera e propria "data fusion" e, quando possibile, di analisi comparata. E nonostante la raffinatezza di tali metodi, sovente, non si riesce ad identificare in maniera corretta l'emozione che si intendeva trasmettere!

Il lavoro di questo articolo mira proprio a mostrare come l'umano, quando viene privato di alcuni dei canali informativi, sovente, stenta a riconoscere lo stato emotivo di colui con cui è interazione. Un limite, questo, che giustifica anche le attuali difficoltà della macchina nel procedere al riconoscimento, e nella sintesi, di stati emotivi. L'indagine qui riportata si è concentrata sul canale audio e più specificamente sulle emozioni veicolate della voce.

Come ben noto molte sono le caratteristiche della voce che si ritiene possano veicolare il portato emotivo di una persona [4,5,6]: velocità del parlato, l'intonazione media (frequenza media) e la sua variazione, l'ampiezza delle frequenze utilizzate, l'intensità del suono, il tono della voce, l'articolazione, ecc.... Gli attori, per le finalità della propria arte, si esercitano con abnegazione sul controllo di tutti questi parametri per riuscire a trasferire al meglio le emozioni di una piece teatrale, specie quando, come nel caso delle trasmissioni radiofoniche, hanno a disposizione soltanto il canale vocale. Situazioni non dissimili si prospettano a tutti

noi quando dobbiamo comunicare attraverso un telefono, o via Skype.

Per poter valutare le capacità di riconoscimento degli stati emotivi da parte dell'uomo, o di un sistema automatico, sarebbe opportuno disporre di corpora di registrazioni prese nel corso di conversazioni naturali; ma ottenere database audio di tale natura risulta estremamente arduo, sia per la difficoltà di ottenere registrazioni significative, sia per la difficoltà di ottenere frammenti di discorsi in cui sia chiaramente predominante un solo stato emotivo. Non resta dunque che ricorrere all'ausilio di attori, come ha fatto la Fondazione Ugo Bordoni (FUB) di Roma che, di recente, ha realizzato EMOVO, un corpus vocale composto da frasi emotivamente colorate a contenuto emotivo controllato.

In questo lavoro è stato utilizzato EMOVO in una condizione sperimentale molto simile a quella della comunicazione reale monocanale in modo da poter valutare la capacità dell'uomo sia di trasmettere che di percepire un contenuto emotivo, avendo sempre ben chiaro i limiti posti allo studio dall'uso di un corpus "artificiale", ovvero realizzato da attori.

2. EMOVO e la sua validazione

EMOVO è un database che contiene 588 record: 14 frasi in lingua italiana ciascuna colorata con 7 differenti stati emotivi da 6 attori professionisti. Le colorazioni emotive utilizzate sono state: disgusto (inteso come fisico più che morale), gioia, paura, rabbia (nella accezione definita da Klaus Scherer come "calda"), sorpresa (utilizzata per tutta la durata della frase), tristezza; a queste è stata aggiunta, per comparazione, la colorazione neutra. per quel che concerne gli attori coinvolti si è trattato di 3 maschi - 30 anni palermitano M1, 27 anni romano M2, 30 anni padovano M3 - e 3 femmine - 28 anni gaetana F1, 23 anni foggiana F2, 25 anni romana F3. Il corpus è stato registrato presso la FUB in condizioni di quiete acustica. Sono stati impiegati due microfoni professionali marca SHURE modello SM58LC ed un registratore digitale MARANTZ modello PMD670. Le registrazioni sono state eseguite con una frequenza di campionamento di 48 kHz, 16 bit stereo, formato wav.

TAB. 1 - Percentuali di riconoscimento della colorazione emotiva: mediante test di discriminazione (FUB); test ISIM_Plutchick: somma su tutte le tonalità di uno stesso colore (intero petalo) (A), solo tonalità espressa dall'attore (porzione di petalo) (B); rapporto A/FUB=C, rapporto B/FUB=D

Actor	FUB	A	B	C	D
M1	82%	32%	13%	39%	16%
M2	79%	27%	17%	34%	21%
M3	77%	21%	6%	27%	8%
F1	77%	32%	16%	42%	21%
F2	84%	36%	20%	43%	24%
F3	84%	32%	17%	38%	20%
Average	81%	30%	15%	37%	17%

Gli attori hanno avuto la possibilità di muoversi liberamente e ciò ha ovviamente inficiato i dati relativi all'intensità assoluta del segnale, dipendente dalla lontananza della bocca dal microfono. Inoltre nei casi di emozioni che generano livelli di energia elevati, come la rabbia, si è dovuto provvedere all'abbassamento manuale del volume di registrazione. Si è consentito agli attori di avere tutto il tempo necessario per prepararsi all'autoinduzione dell'emozione, concedendo loro le pause richieste e consentendo la ripetizione delle frasi la cui esecuzione non li convinceva.

La validazione del corpus è stata effettuata mediante test di discriminazione, utilizzando le due frasi non-sense contenute nel corpus ("La casa forte vuole col pane", "Il gatto sta scorrendo nella pera") e richiedendo a 24 soggetti di indicare - tra una coppia di stati - quale fosse quello più vicino ad un determinato stato emotivo (ai soggetti è stato chiesto di indicare le proprie scelte per ciascuna delle possibili combinazioni di coppie di stati). In base a questa strategia di validazione (comparazione tra due stati) la percentuale di riconoscimento è risultata essere dell'81% (84% per F2 e F3, 82% per M1, 79% per M2 e 77% per F1 e M3), tale da lasciar ben sperare sulla capacità dell'uomo di trasmittere-identificare un determinato stato emotivo comunicando esclusivamente tramite il canale audio.

3. EMOVO in condizioni di "comunicazione naturale"

Quando si ascoltano trasmissioni radiofoniche, o nelle interazioni naturali vocali (es. skype vocale), non si procede all'identificazione dell'emozione per confronto tra un numero limitato di stati ma operando una selezione a partire dall'intera tavolozza delle possibili colorazioni vocali, aiutati, in questo, dal contesto della conversazione e dal senso della singola frase.

Per poter effettuare una validazione del corpus e avviare uno studio dell'identificazione dello stato emotivo trasmesso dalla voce in una situazione prossima al reale, abbiamo deciso di disegnare un esperimento, ISIM_Plutchick test, in cui al soggetto viene presentata una frase non-sense e le/gli viene data la possibilità di indicare l'emozione percepita all'interno dell'intera gamma dei colori di cui si compone il "fiore" di Plutchick [7]. In un certo senso, il test di riconoscimento viene condotto in una situazione che si potrebbe addirittura definire di "iperrealtà vocale" per l'eliminazione degli appigli che possono derivare dal senso e dalla contestualizzazione del discorso. E' bene sottolineare che nel fiore di Plutchick, rispetto al modello di Eckman [8], sono presenti (con riferimento alla porzione centrale del petalo) due emozioni aggiuntive, fiducia e attesa che nel test, a tutti gli effetti, operano da distrattori.

Per questo, i risultati che abbiamo ottenuto, e che descriveremo in seguito, devono essere considerati come un limite inferiore alla riconoscibilità dello stato emotivo, mentre quelli ottenuti dalla FUB (validazione per differenza) potrebbero essere interpretati come vicini al limite superiore della riconoscibilità dell'emozione.

Per realizzare l'esperimento abbiamo sviluppato in Processing due semplici applicativi (per ora non ancora integrati) da utilizzare rispettivamente per: a) l'ascolto dei

brani vocali e la registrazione della colorazioni emotive percepite dal soggetto; b) la visualizzazione dei dati.

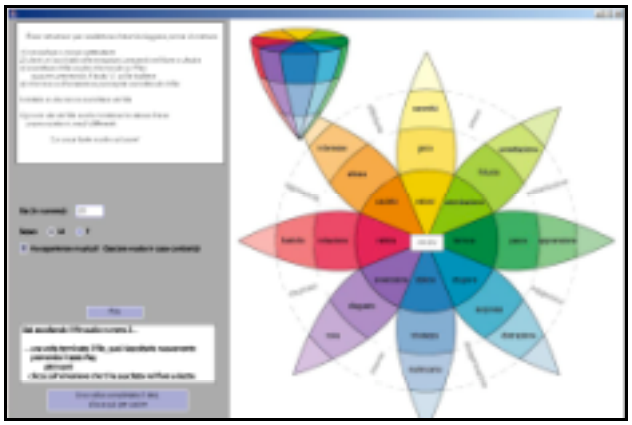

Fig. 1 - Organizzazione dello schermata utilizzata per proporre ai soggetti il test su EMOVO

Il protocollo dell'esperimento prevede cha al soggetto sia mostrata la schermata di fig. 1. In alto a sinistra sono riportate le istruzioni che dovranno essere seguite dal soggetto; si richiede: a) di compilare, dapprima, i campi relativi all'anagrafica e al possesso di eventuali abilità musicali (per ora non differenziate); b) di effettuare, poi, un'esplorazione visiva del "fiore" delle emozioni, al quale è stato aggiunto al centro un rettangolo bianco con la scritta neutro/a; c) quindi, una volta che ci si sente pronti e rilassati, di cominciare l'ascolto dei brani vocali, a partire dal primo, e di indicare, alla fine dell'ascolto di ogni brano, l'emozione percepita facendo click sulla corrispondente area della rappresentazione di Plutchick.

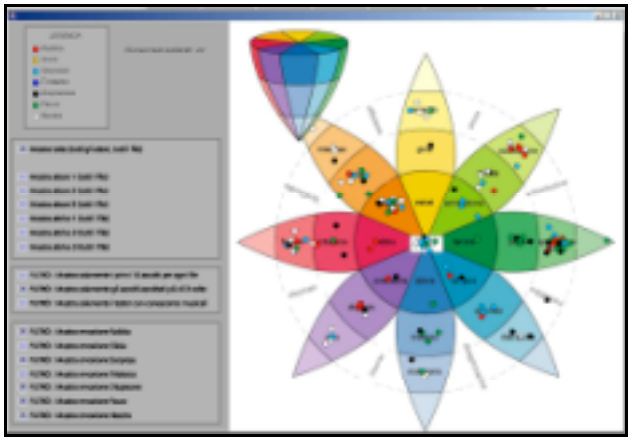

Fig. 2 - Schermata dell'applicativo utilizzato per l'analisi dei dati: sulla sinistra legenda e lista dei filtri disponibili; sulla destra fiore delle emozioni di Plutchick su quale viene mostrata la distribuzione delle risposte ottenute.

L'avanzamento al brano successivo viene determinato dal click che indica l'avvenuta segnalazone dell'emozione percepita. Al soggetto è stata lasciata la possibilità di riascoltare tutte le volte che reputasse necessario le singole frasi in modo che potesse arrivare a indicare con la massima sicurezza l'emozione percepita. Ai soggetti, 22 in tutto (15 maschi e 7 femmine comprese tra 16 e 53 anni), è stata fatta ascoltare la frase "La casa forte vuole col pane" pronunciata da tutti gli attori e con tutte le colorazioni emotive, per un totale di 42 ascolti. L'ordine di ascolto, pseudo-casuale, è stato identico per tutti i soggetti.

Un secondo applicativo, fig. 2, ci ha consentito poi di esaminare i risultati ottenuti filtrandoli, a secondo del bisogno, per attore, emozione, per finestra temporale di ascolto (in questo test solo primi dieci), per numero di ascolto dello stesso brano (in questo test uguale o maggiore a 3), per abilità musicale dei soggetti sottoposti a test.

4. Analisi dei dati

In tabella 1 sono riportate le percentuali di riconoscimento medio riscontrate per ciascun attore e per il corpus nel suo complesso. Salta subito agli occhi che, come ci si attendeva, nelle nostre condizioni sperimentali il riconoscimento medio della colorazione emotiva scende drasticamente e si attesta al 30% se si considerano tutte le sfumature di un petalo del fiore di Plutchick e addirittura al 15% se si considera la sola tonalità di interesse (ovvero utilizzata dall'attore). Per questi due casi, dunque, la riconoscibilità, scende rispettivamente al 37% e 17% di quanto riscontrato nella validazione effettuata dalla FUB. Per quel che concerne gli attori nel nostro caso si riscontra per le voci maschili lo stesso trend riscontrato in FUB con percentuali di riconoscimento tali che M1 > M2 > M3, ma un sostanziale decremento delle percentuali di riconoscimento per M3. La migliore prestazione è di nuovo associabile a F2, ancora molto elevati i risultati di F3 (comparabili a M1) ma altrettanto elevati sono i riscontri per F1 con un netto miglioramento rispetto all'analisi FUB. La variabilità del riconoscimento sui singoli attori è tale che si va dal 36% al 21%. Il confronto tra i sessi favorisce i parlatori di genere femminile che si attestano al 33% mentre per i parlatori di genere maschile non si va oltre il 27%.

In generale questi primi riscontri ci dicono della difficoltà di riconoscere le colorazioni emotive nel vocale in assenza di termini di paragone e di contesto, e che un 15% circa della riconoscibilità del colore è da associarsi alle caratteristiche peculiari del parlatore.

A queste ultime è dedicata la tabella 2 nella quale è riportata la matrice di riconoscibilità attore-colorazione emotiva. Anche qui saltano subito agli occhi dei trend e delle differenze che ancor di più tendono a sottolineare la rilevanza delle caratteristiche del parlatore ai fini del riconoscimento delle emozioni trasmesse dal vocale, ma anche le specificità delle singole colorazioni.

Colorazioni molto forti, come quelle relative ai petali della rabbia e della paura, vengono più facilmente riconosciute. La colorazione emotiva del petalo della rabbia, fatta eccezione per M3, sembrerebbe più facilmente trasmissibile da una voce maschile. A ruota seguono gli opposti gioia-tristezza che sembrano decisamente appannaggio del genere femminile. Molto altalenante e decisamente meno efficace la trasmissione di colorazioni meno marcate come il disgusto e la sorpresa. Sulla stessa percentuale di queste ultime anche la capacità di riconoscere la tonalità neutra.

Se dalle generalità si passa alle specificità del singolo parlatore si può osservare che, limitatamente a questo test: F3 è quella che riesce ad esprimere con più costanza le colorazioni emotive, eccezion fatta per il disgusto; F2 riesce particolarmente bene nella trasmissione delle emozioni marcate, quali paura, rabbia e gioia; F1 eccelle nella trasmissione della tristezza e del disgusto mentre ha difficoltà a colorare la sorpresa e la paura; M3 risulta essere un parlatore non particolarmente espressivo ad eccezion fatta per la trasmissione del disgusto; M2 un po' come F2, ma in maniera meno efficace, riesce a trasmettere colorazioni emotive marcate (vedere in particolare la rabbia) e nel disgusto; anche M1 riesce a trasmette emozioni marcate, incontra difficoltà nella trasmissione della gioia ma eccelle nella trasmissione della tristezza e della sorpresa. Per quel che riguarda l'attore M3 è lecito chiedersi se la sua provenienza geografica non possa aver inficiato il riconoscimento dei colori vocali da parte di un pubblico di uditori principalmente di area geografica Centro-Italia. In ogni caso, si ritiene necessaria un'indagine più accurata.

Quelli appena discussi, pur essendo dati molto preliminari, mettono già in evidenza come la colorazione vocale non possa essere ricostruita facilmente a tavolino sulla base delle caratteristiche di un solo parlatore (infatti anche il miglior candidato F3 sembra carente nell'espressione del disgusto) e come, invece, i modelli per l'implementazione di parlatori sintetici dovranno far ricorso a un'accurata osservazione sul campo per poter estrarre le caratteristiche che permettono di ricostruire al meglio tutti i colori della tavolozza, al fine di renderli percepibili.

TAB. 2 - matrice di riconoscibilità attore-colorazion emotiva

Actor	rabbia	disgusto	tristezza	sorpresa	paura	gioia	neutro
M1	54%-23%	5%-5%	36%-14%	41%-23%	54%-23%	5%-5%	27%
M2	68%-59%	23%-14%	9%-0%	9%-0%	45%-23%	23%-9%	14%
M3	5%-0%	50%-14%	18%-0%	9%-0%	14%-5%	18%-0%	27%
F1	45%-14%	38%-28%	57%-33%	9%-0%	9%-0%	45%-18%	14%
F2	45%-14%	5%-0%	19%-5%	24%-19%	82%-41%	54%-45%	18%
F3	38%-24%	5%-5%	41%-18%	32%-14%	38%-5%	36%-23%	41%
Average	35%-22%	21%-11%	30%-12%	21%-9%	40%-16%	30%-17%	23%

5. Prospettive

Quanto presentato e discusso nel precedente paragrafo è solo la punta dell'iceberg di un lavoro di analisi, in progress, che nel caso della lingua italiana non era mai stato fatto e che sta coinvolgendo anche lo studio delle correlazioni tra colorazioni emotive e relazione parlatore-uditore (e che, eventualmente, si potrà allargare ad elementi di contestualizzazione sociale). Si tratta di parametri che devono essere tutti presi in considerazione sia ai fini di una migliore riconoscibilità di tutte le sfumature di colore che per generare migliori personalizzazioni, in condizioni di interazione naturale, ovvero realistica.

Va da sé che tra gli obiettivi futuri vi sono a) il potenziamento dei tool utilizzati nell'esperimento qui descritto allo scopo di poter eseguire una più accurata analisi dell'influenza delle qualità sia del parlatore che dell'ascoltatore nel riconoscimento della coloritura emotiva; b) la correlazione tra risultati psicoacustici e le caratteristiche fisiche misurabili della voce degli attori.

I risultati sul riconoscimento medio del colore vocale - che abbiamo visto non superare il 30% - prospettano anche la necessità di affrontare un accurato studio delle cross-interazioni modali (vocale, immagini, gesti) che come noto influenzano il riconoscimento degli stati emotivi [4,5].

Prospettive e sviluppi di lavoro molto densi, dunque, che fanno intravvedere, a loro volta, moltissime applicazioni in tutti i settori del riconoscimento e della sintesi vocale associati all'affective computing ma anche in ambiti, forse meno ampi ma non meno importanti, come quelli della diversabilità/difficoltà espressiva e delle scuole di recitazione.

6. Riferimenti bibio-webgrafici

[1] see for example the humaine portal: http://emotion-research.net/ and link and references therein

[2] Picard W.J., Affective Computing, MIT Press, 1997; http://affect.media.mit.edu/publications.php

[3] see the proceedings of the workshop on "Measuring Affect in HCI: Going Beyond the Individual" http://www.emotion-in-hci.net/chi08/proceedings.html

[4] Scherer, K. R.; Johnstone, T.; Klasmeyer, G., Handbook of the Affective Sciences, Oxford University Press, New York and Oxford, p.433–456 (2003)

[5] Juslin, P. N.; Scherer, K. R., The New Handbook of Methods in Nonverbal Behavior Research, Oxford University Press, Oxford, UK, p.65–135 (2005)

[6] See "Verbal and Nonverbal Communication Behaviours", ed. by A. Esposito et al., Springer LNA5 4775 (2007)

[7] Plutchik R., EMOTION: A Psychoevolutionary Synthesis, NEW YORK: Harper & Row, 1980

[8] Ekman, P. 1992. An Argument for Basic Emotions. Cognition and Emotion 6(3/4): 169-200