RICONOSCIMENTO DI EMOZIONI NEL PARLATO PER MEZZO DI PARAMETRI PROSODICI

Roberto Gretter, Dino Seppi
ITC-irst, Trento, Italia
gretter@itc.it, *seppi@itc.it*

1. SOMMARIO

Il presente lavoro si propone di descrivere la realizzazione di un sistema automatico per il riconoscimento di emozioni nel parlato. La peculiarità del progetto consiste nell'aver utilizzato *database* di parlato spontaneo, in particolare registrazioni di interazioni vocali uomo-macchina. L'intero progetto può essere suddiviso in tre differenti fasi: la raccolta e l'etichettatura dei dati, l'estrazione di parametri prosodici dalle registrazioni audio e la classificazione emozionale di ciascuna frase.

Sono stati considerati due *database*. Il primo, in italiano, consiste in registrazioni di utenti di *call-center* automatici. Questi dati sono stati selezionati ed etichettati da annotatori professionisti in base alle emozioni espresse da ciascun parlatore. Il secondo, in tedesco, consiste invece in registrazioni raccolte durante la messa punto di un sistema di dialogo automatico. Questa seconda raccolta contiene un numero più significativo di frasi non emotivamente neutre ed è stata utilizzata per un raffronto con i dati italiani. La scelta di fare uso di parlato spontaneo comporta inevitabilmente alcuni problemi: la presenza preponderante di frasi emotivamente neutre che rende i *database* molto sbilanciati, la disomogenea distribuzione delle emozioni rilevate e i bassi livelli di consenso registrati tra gli annotatori.

Ulteriori problematiche derivano dall'estrazione automatica di parametri significativi: a tutt'oggi, infatti, le tecniche utilizzate in letteratura non consentono di fare affidamento su un insieme limitato di *features* robuste. L'approccio adottato prevede quindi il calcolo di un gran numero di parametri, anche molto correlati tra loro, per il successivo addestramento di un classificatore automatico. Tali parametri, comunemente utilizzati in letteratura, derivano da funzioni del segnale audio come energia, frequenza fondamentale, durata ed eventuale presenza di pause tra parole e sono stati calcolati a livello di parola. La segmentazione del segnale vocale è stata ottenuta per mezzo di un segmentatore automatico.

Date le potenziali ambiguità delle annotazioni e le lacune nelle informazioni codificate dai parametri acustici, il compito svolto dalla classificazione diviene difficile se non critico. Per questo motivo abbiamo testato due diversi tipi di classificatori: le reti neurali e gli alberi binari di classificazione; per entrambi forniamo i risultati nelle configurazioni che si sono rivelate più robuste. Nonostante entrambi i metodi proposti non si comportino particolarmente bene se applicati a dati sparsi e molto sbilanciati, siamo riusciti a ottenere, per entrambi i *database* utilizzati, risultati equiparabili e prestazioni più che soddisfacenti.

In conclusione restano ancora irrisolti numerosi problemi di carattere tecnico e concettuale, tra cui la significatività dei parametri prosodico-acustici utilizzati e l'affidabilità dell'etichettatura manuale dei dati. Quindi sforzi futuri per il miglioramento dei risultati andranno direzionati soprattutto verso un'attenta e più approfondita analisi dei parametri, che andranno modificati, selezionati e affiancati da altri, non necessariamente solo di natura prosodica.

2. DESCRIZIONE DEI DATI

La prima parte di questo studio è consistita nella raccolta e nell'annotazione dei dati, e nella scelta e nello studio di un *database* integrativo, già etichettato, per un confronto più significativo nella successiva fase sperimentale. In particolare abbiamo utilizzato due *database* di parlato spontaneo, il primo in lingua italiana, il secondo in lingua tedesca.

2.1 Il database *Targhe*

Il *database* italiano è stato assemblato in ITC-irst. Esso deriva sostanzialmente da altri due *database* precedentemente raccolti: *Car-Plates e Tal-Trains*. Entrambe queste raccolte consistono di parlato spontaneo uomo-macchina. Car-Plates è una raccolta di dialoghi tra utente e un sistema di dialogo automatico che permette di pagare tasse automobilistiche via telefono; Tal-Trains (Cattoni, 2002) contiene conversazioni telefoniche tra utente e un sistema prototipo di dialogo per la consultazione di orari dei treni. La peculiarità di entrambe le raccolte risiede nel fatto che i dialoghi sono condotti da parlatori comuni e il contenuto emotivo è assolutamente spontaneo.

Da queste due raccolte si è provveduto ad estrarre, in modo semi-automatico, le frasi che potessero contenere effettivamente stati emotivi non neutrali, andando a selezionare frasi con interruzioni o con un numero ridotto di parole. Successivamente, questi dati sono stati segmentati in modo automatico utilizzando il riconoscitore automatico del parlato sviluppato in ITC-irst (Angelini, 1993); in seguito sia le trascrizioni che le segmentazioni sono state controllate e corrette manualmente. Complessivamente il *database* italiano utilizzato, denominato **Targhe**, consiste di circa 15 ore di parlato spontaneo, di qualità telefonica, suddiviso in 9444 frasi. I parlatori sono adulti, sia maschi che femmine.

Successivamente, si è provveduto a etichettare i dati da un punto di vista emotivo, avvalendosi di due annotatori esperti. L'etichettatura emotiva è stata apposta a livello di frase senza utilizzare un insieme prestabilito di etichette emotive a cui attingere. Complessivamente sono stati individuati 9 diversi tipi di stati emotivi nei parlatori (inclusa la classe NEUTRO). In Tabella 1 è rappresenta la statistica delle classi utilizzate dagli annotatori. La maggior parte delle frasi (95%) si è rivelata non contenere alcuno stato emotivo. E' quindi stato chiesto ai due annotatori di riconsiderare la parte del *database* inerente le frasi non neutre e cercare di ottenere un consenso sull'etichetta finale.

Targhe – 9444 frasi	
Emozione	Frequenza
NEUTRO	95.2%
NOIA	2.5%
RABBIA	1.9%
PREOCCUPAZIONE	0.2%
IRONIA	0.1%
FELICITÀ	0.1%
SCORAMENTO	0.02%
SORPRESA	0.01%
DISPIACERE	0.01%

Sympafly – 5283 frasi	
Emozione	Frequenza
NEUTRO	83.5%
ALTERAZIONE	10.9%
PERPLESSITÀ	3.0%
IRONIA	1.1%
SCORAMENTO	1.0%
FELICITÀ	0.2%
PANICO	0.1%
SORPRESA	0.1%
RABBIA	0.06%

Tabella 1: Statistica delle classi emotive in entrambi i *database* considerati.

Al fine di valutare l'affidabilità dell'etichettatura finale è stato calcolato il consenso tra i due annotatori utilizzando un campione di 99 frasi selezionate in modo da avere una distribuzione delle classi quanto più uniforme possibile. A tal scopo, dato che alcune

emozioni hanno frequenze molto basse, ne sono state selezionate non più di 20 per tipo. La matrice di confusione, esposta in Tabella 2, mostra come sia effettivamente necessario introdurre delle *sovra-classi* che possano in qualche modo raggruppare le etichette originali, cercando in tal modo di evitare la scarsità e lo sbilanciamento dei dati. Abbiamo quindi optato per scelta di tre sovra-classi principali a seconda della valenza emotiva espressa da ciascuna emozione: NEUTRO, POSITIVO e NEGATIVO (Per la classificazione le classi POSITIVO e NEGATIVO confluiranno a loro volta nella sovra-classe NON-NEUTRO). Le nuove classi POSITIVO e NEGATIVO raggruppano rispettivamente le classi originali SORPRESA e FELICITÀ da una parte e NOIA, RABBIA, PREOCCUPAZIONE, IRONIA, SCORAMENTO e DISPIACERE dall'altra. Il consenso può quindi essere condotto calcolando il rapporto del numero di corrispondenze rispetto al numero di combinazioni. Se si considera il raggruppamento proposto si passa da un consenso del 62.6% a uno del 83.8%.

	NE.	NO.	IR.	FE.	SO.	PR.	SC.	RA.	DI.
NEUTRO	**36**	5	0	0	1	1	1	3	1
NOIA	2	**10**	0	0	0	2	1	8	0
IRONIA	0	1	**5**	1	0	0	0	1	0
FELICITA'	1	0	0	**1**	1	0	0	0	0
SORPRESA	0	0	0	0	**0**	0	0	0	0
PREOCCUPAZ.	0	2	0	0	0	**6**	3	0	0
SCORAMENTO	0	0	0	0	0	1	**0**	0	0
RABBIA	0	1	0	0	0	0	0	**4**	0
DISPIACERE	0	0	0	0	0	0	0	0	**0**

Tabella 2: Matrice di confusione dell'etichettatura di classi di emozioni da parte dei due annotatori.

2.2 Il database *Sympafly*

La seconda raccolta di dati audio, **Sympafly** (Batliner *et al.*, 2004a), è stata fornita dall'Università di Erlangen-Nuernberg e dalla ditta Sympalog. Essa consiste di circa 5 ore di conversazione uomo-macchina. La qualità delle registrazioni è pure telefonica, in quanto si tratta di dialoghi di utenti con un *call-center* per la prenotazione di voli aerei. Gli utenti sono sia gli sviluppatori stessi del sistema, sia utenti veri, adulti, sia maschi che femmine.

Poiché le registrazioni hanno avuto luogo mentre il sistema era in fase di realizzazione e collaudo, la parte emotiva è risultata essere relativamente abbondante. Dalla Tabella 1 si evince come Sympafly consti del 13% di frasi emotivamente non neutre.

Le registrazioni di questo corpus sono state fornite assieme alle trascrizioni e alle segmentazioni. Inoltre sono state anche messe a disposizione la frequenza fondamentale delle registrazioni e la segmentazione in sillabe.

2.3 Considerazioni

Per quanto riguarda l'etichettatura emotiva, il *set* di classi utilizzato per annotare il *database* tedesco non corrisponde, se non in minima parte, all'insieme adoperato dai nostri annotatori. Inoltre non è detto che a etichette uguali corrispondano effettivamente gli stessi stati emotivi. Per questo motivo non è stato possibile eseguire delle prove utilizzando allo stesso tempo entrambe le raccolte a disposizione: gli esperimenti sono stati portati avanti comparativamente, cercando di confrontare per quanto possibile tecniche e approcci su entrambi i *corpora*.

Sia l'utilizzo che la creazione di una raccolta di dati emotivamente ricchi ha posto diversi problemi. Dal punto di vista dell'annotatore, la difficoltà principale è stata la scelta delle etichette e il loro numero. Come già accennato, abbiamo cercato di aggirare questo

problema in due modi: dapprincipio la scelta delle etichette è stata demandata agli annotatori, successivamente si è cercato di trovare un accordo tra gli stessi, quindi abbiamo raggruppato le varie classi a seconda della valenza emotiva. Dato l'esiguo numero di manifestazioni positive, abbiamo infine optato per due sole sovra-classi principali: NEUTRO e NON-NEUTRO. In questa maniera ci siamo inizialmente preposti di restringere lo studio all'identificazione di uno stato d'animo non neutrale, sia esso a valenza positiva o negativa, che potesse essere indice di "anomalie" nella comunicazione uomo-macchina.

Questo approccio aggira il problema di assegnare definizioni emotive (Cowie, 2001) più o meno condivisibili secondo le discipline più diverse (filosofiche, biologiche, psicologiche), e si basa maggiormente sulla discrezione degli annotatori nel determinare alterazioni nella voce dei parlatori. Il nostro obiettivo è infatti costruire un sistema che sia in grado replicare la capacità umana – dei nostri annotatori per la fattispecie – di individuare problemi nella comunicazione. In questa maniera il nostro lavoro rimane naturalmente limitato alle capacità interpretative degli annotatori ed espressive degli individui registrati.

Database	ore	# frasi	neutre	Lingua
Targhe	15:16	9444	95%	Italiano
Sympafly	5:18	5283	87%	Tedesco

Tabella 3: Caratteristiche dei *database* utilizzati.

Il secondo fondamentale problema nel quale ci siamo imbattuti in questa prima fase riguarda la carenza di alcune classi e la loro distribuzione tutt'altro che uniforme. Questa caratteristica era prevedibile in quanto i dati raccolti sono di parlato spontaneo, e quindi la presenza di emozioni è limitata da molti fattori: dall'abilità dei parlatori a esprimere (o nascondere) il loro stato d'animo, dalla durata del rapporto uomo-macchina, dalla pazienza e dalla disposizione che i parlatori mostrano, in generale, verso sistemi di dialogo automatici. Anche per questi motivi la scelta del raggruppamento delle etichette in due sovra-classi principali ha reso possibile la successiva costruzione di un sistema limitato ma molto più affidabile.

Da ultimo va considerato che ogni frase può essere intrinsecamente ambigua dal punto di vista emotivo (Cowie, 2001). Questo aspetto dipende anche dall'unità linguistica fondamentale nella quale si manifesta una certa emozione, la quale a sua volta varia notevolmente a seconda della forza con cui la si vuol esprimere. Può infatti succedere che un'intera frase venga percepita come caratterizzata da una stessa emozione, oppure una sola parola all'interno di un discorso. Infine nella stessa frase possono coesistere stati emotivi differenti. Va da sé che in questi casi la scelta degli annotatori diviene ancora più difficile e contraddittoria, mentre l'uso di sovra-classi, pur generalizzando, media gli errori di valutazione.

3. ESTRAZIONE DEI PARAMETRI

I parametri acustici utilizzati per la classificazione automatica delle emozioni hanno natura prettamente acustica e in particolare prosodica. Più in particolare essi sono stati estratti a partire da elaborazioni del segnale audio come l'energia, il *pitch* e la durata calcolati su ogni singola parola della frase. Per una descrizione particolareggiata dei parametri prosodici utilizzati si consulti Kiessling (2001). Per alcune *features* estratte sono stati considerati anche orizzonti temporali più lunghi della parola stessa (contesto=0), come l'insieme della parola analizzata e la precedente o seguente (contesto=1), oppure due parole

seguenti o due precedenti (contesto=2). L'utilizzo di contesti diversi da zero permette di considerare anche variazioni dei parametri analizzati e non solo i loro valori assoluti.

3.1 Parametri derivanti dall'energia

L'energia del segnale vocale è stata estratta frase per frase. Dopo la segmentazione in parole, sono state computate le *features* energetiche per ogni parola. In particolare sono stati identificati i massimi e minimi dell'energia, le relative posizioni normalizzate sulla lunghezza dalla parola in considerazione, i parametri della regressione lineare dell'energia come il coefficiente angolare e l'errore, e infine l'energia media della parola.

Parametri	Contesto				
	-2	**-1**	**0**	**+1**	**+2**
Pitch: valor medio di frase			-		
Pitch: valor medio		●	●	●	
Pitch: Massimo		●	●	●	
Pitch: posizione normalizzata del massimo		●	●	●	
Pitch: minimo		●	●	●	
Pitch: posizione normalizzata del minimo		●	●	●	
Pitch: valore di *on-set*			●	●	
Pitch: posizione normalizzata di *on-set*			●	●	
Pitch: valore di *off-set*		●	●		
Pitch: posizione normalizzata di *off-set*		●	●		
Pitch: regressione e relativo errore	●●	●●	●●	●●	●●
Energia: valor medio	●	●	●	●	●
Energia: valor assoluto	●	●	●	●	●
Energia: Massimo		●	●	●	
Energia: posizione normalizzata del massimo		●	●	●	
Energia: regressione e relativo errore	●●	●●	●●	●●	●●
Energia: *Tau*			-		
Energia: valor medio normalizzato	●	●	●	●	●
Durata: valor assoluto	●	●	●	●	●
Durata: *Tau*			-		
Durata: valore normalizzato	●	●	●	●	●
Durata: normalizzazione sulle sillabe	●	●	●	●	●
Pausa	●	●		●	●

Tabella 4: Contesto dei parametri estratti dal *pitch*, dall'energia e dalle durate. I tre parametri contrassegnati con un trattino sono calcolati su tutta la frase. Non è indicato il contesto -½ (5 *features*).

Al fine di rendere i parametri derivanti dall'energia più affidabili e meno influenzati da caratteristiche locali del segnale audio (per esempio la dipendenza dal parlatore), si è anche proceduto a diverse normalizzazioni: il segnale energia è stato infatti normalizzato rispetto all'energia media della frase della quale faceva parte la parola sotto analisi e rispetto all'energia della parola stessa in tutto il *database* (si veda più avanti *Tau*). Per quanto riguarda l'energia, sono state raccolte complessivamente 33 *features* a livello di parola, considerando anche contesti diversi da zero.

3.2 Parametri derivati dalla frequenza fondamentale

La frequenza fondamentale, o *pitch*, è stata calcolata in un primo tempo utilizzando metodi differenti. A differenza del calcolo dell'energia del segnale, l'estrazione del *pitch* presenta infatti non pochi problemi ancora irrisolti. Tutti gli algoritmi disponibili presentano, se pur in quantità diverse, degli errori di stima e purtroppo anche un solo valore sbagliato potrebbe invalidare molti dei parametri che derivano dal *pitch* stesso e che ne

dovrebbero riassumere le caratteristiche salienti. Un errore di raddoppiamento a fine parola, peraltro non così raro, comprometterebbe per esempio la posizione e il valore del massimo della frequenza fondamentale, e i coefficienti della retta di regressione.

Complessivamente abbiamo provato tre diverse curve di *pitch*, due ottenute grazie a due diversi algoritmi di estrazione, la terza fornita assieme al *database* Sympafly. Per quanto riguarda i due algoritmi implementati, essi utilizzano due delle più usate tecniche di estrazione, vale a dire l'auto-correlazione pesata (*weighted autocorrelation*, WAC) e la cross-correlazione normalizzata (*normalized crosscorrelation*, NCC). Entrambe le tecniche sono dettagliatamente descritte nei rispettivi riferimenti originali (Shimamura & Kobayashi, 2001; Medan *et al*., 1991). Il secondo di questi due algoritmi è stato successivamente perfezionato con l'ausilio di tecniche di programmazione dinamica, al fine di minimizzare soprattutto i salti d'armonica. Per quanto riguarda la terza curva di *pitch* (UERLN), essa può essere descritta come la sostanziale concatenazione di un modulo ad auto-correlazione seguito da un filtro non lineare. Purtroppo la realizzazione di un estrattore di *pitch* comprende anche e soprattutto la regolazione di soglie e parametri. Questo difficile aspetto è indispensabile e spesso determinante ai fini dell'ottimizzazione dell'estrattore stesso e frequentemente richiede ripetute tarature.

Per i due algoritmi implementati sono state calcolate le prestazioni (Rabiner, 1976). Queste si possono misurare principalmente tramite due grandezze: errori di vocalizzazione ed errori di precisione. Per quanto riguarda questi ultimi ci siamo focalizzati in particolare su quelli di raddoppiamento o dimezzamento del *pitch*, genericamente individuati da differenze tra stima e riferimento maggiori di 10 ms. Tutte le prove sono state condotte sul *database* Keele (Plante, 1995) per il quale viene fornito anche il *pitch* di riferimento.

PDA	V|V	V|UV	UV|V	UV|UV	Errori >10ms
NCC	45%	3%	6%	46%	5%
WAC	42%	5%	8%	45%	7%

Tabella 5: Prestazioni dei due algoritmi di estrazione del *pitch* (PDA) implementati. V significa vocalizzato, UV non vocalizzato. Nell'ultima colonna sono indicati le percentuali di *frame* V|V per i quali la differenza con il riferimento supera 10 ms. (di solito errori di raddoppiamento o dimezzamento). V|UV indica la percentuale di *frame* riconosciuti come vocalizzati (V) quando nel riferimento risultano invece non vocalizzati (UV).

Una volta stimata la curva del *pitch*, in analogia a quanto fatto per l'energia, sono stati calcolati i parametri che ne derivano. Oltre ai parametri descritti per l'energia, sono state introdotte altre quattro *features*: *on-set* e *off-set* e loro posizione relativa. Con *on-set* si intende il valore del *pitch* nel primo *frame* vocalizzato, mentre con *off-set* si indica il valore del *pitch* dell'ultimo *frame* vocalizzato della parola.

Complessivamente dal *pitch* sono stati estratti 35 parametri a livello di parola.

3.3 Parametri derivanti dalle durate

Infine sono stati aggiunti dei parametri legati alle durate delle parole e alle pause tra le parole stesse. Come per quanto fatto per l'energia, le durate sono state normalizzate seguendo diversi criteri: secondo il numero di sillabe contenute nella parola, secondo la durata media delle parole all'interno della frase, e infine secondo la durata media della stessa parola all'interno di tutto il *database*.

Dalla durata è stato possibile raccogliere un totale di 20 parametri a livello di parola. L'elenco completo delle *features* utilizzate è illustrato in Tabella 4. Alle *features* di

pertinenza delle singole parole descritte precedentemente vanno aggiunti tre parametri calcolati a livello di frase, per un totale di 91 *features* prosodiche. Si tratta della media del *pitch* su tutta la frase, e dei *Tau* (Buckow, 2003, pag. 58) di energia e durata, usati anche per le successive normalizzazioni. *Tau* è un fattore di scala che indica quanto la parola i-esima di una data frase sia pronunciata più forte o più a lungo rispetto alla media calcolata su tutto il *database*.

3.4 Selezione dei parametri

Le *features* raccolte derivano, come precedentemente spiegato, dal segnale delle singole parole. Questa scelta è stata dettata, oltre che dalla volontà di attenersi ai metodi finora utilizzati in letteratura, anche dalla necessità di poter codificare variazioni prosodiche non necessariamente legate alla struttura della frase, bensì anche a singoli gruppi di parole.

In tal modo si viene a verificare un disallineamento tra l'etichettatura, effettuata a livello di frase, e i parametri, estratti a livello di parola. Esiste quindi la necessità di una successiva fase intermedia di selezione, a monte della classificazione. Tale selezione si propone di associare un eguale numero di *features* alle rispettive etichette emotive. Dato il numero variabile di parole all'interno di ciascuna frase abbiamo deciso di estrarre e conservare, per ogni frase solo il parametro che presentasse il valore massimo, quello minimo e la media di tutti i valori. Cosi facendo abbiamo ottenuto complessivamente 273 *features* per ogni frase.

Un'altra soluzione, che però non è stata testata, sarebbe consistita nell'associare la medesima etichetta emotiva di frase a tutte le parole della frase stessa ed effettuare la classificazione a livello di parola piuttosto che di frase. Quindi sarebbe stato possibile determinare la classe delle frasi partendo dalle etichette delle parole utilizzando metodi più o meno complessi. Ciò pone di fronte alla difficoltà di scegliere un'euristica adeguata e all'approssimazione di assegnare la stessa etichetta emotiva a tutte le parole della frase.

4. CLASSIFICATORI UTILIZZATI

Al fine di costruire un sistema di riconoscimento automatico delle emozioni si è deciso di utilizzare classificatori statistici. In particolare ne abbiamo scelti due dalle caratteristiche molto differenti: gli alberi binari di classificazione (CART), descritti in Breiman *et al*. (1984) e le reti neurali a singolo strato (NN) descritte in Zell (1994).

I CART sono strumenti molto efficienti e consentono l'addestramento in tempi molto rapidi. Inoltre essi consentono di classificare anche classi con *features* non linearmente separabili. D'altra parte gli alberi sono strumenti delicati e poco stabili in quanto piccole differenze nei dati si ripercuotono in maniera a volte sorprendente nei risultati finali. Un altro inconveniente è il fatto che i CART siano strumenti sub-ottimi nel senso che vengono costruiti cercando di massimizzare localmente una funzione obiettivo. I risultati ottenuti con questi classificatori possono comunque essere considerati affidabili in quanto le prestazioni sono state mediate su 10 alberi differenti (si veda più avanti).

L'altro classificatore, NN, è uno strumento molto più sofisticato, nonostante ci si sia limitati a considerare esclusivamente NN a singolo strato, ovvero "quasi" lineari. Le reti hanno l'indubbio vantaggio di minimizzare l'errore di classificazione e di gestire classi sbilanciate. Purtroppo però esse sono lente da addestrare e nel caso considerato (1 *layer*) non gestiscono nella maniera ottimale classi non linearmente separabili.

Data l'esigua quantità di frasi non neutre, abbiamo dovuto ricorrere alla tecnica nota come k-*cross-validation*: abbiamo diviso i dati in k=10 parti, e abbiamo condotto 10 differenti addestramenti utilizzando di volta in volta 9 parti per addestrare il classificatore e

testando sulla parte restante, ruotando su quest'ultima. Inoltre, esclusivamente per quanto riguarda i CART, abbiamo dovuto provvedere al bilanciamento delle classi prima dell'addestramento stesso. Il bilanciamento è stato ottenuto sovracampionando i dati NON-NEUTRO, quelli cioè meno numerosi, fino ad ottenere per entrambe le classi la medesima cardinalità per quanto riguarda il *training set*.

Le prestazioni dei classificatori utilizzati sono state misurate utilizzando le misure *Recognition Rate*, RR, e il *CLass-wise recognition rate*, CL. RR rappresenta il rapporto tra il numero di elementi riconosciuti come corretti e il numero totale di elementi presenti nel *test set*, mentre CL rappresenta la media dei RR calcolati separatamente per le due classi. In generale CL è diverso da RR; i due valori coincidono quando il *test set* è bilanciato. Buoni risultati di classificazione richiedono che entrambe queste misure, oltre ad essere quanto più elevate possibili, non si discostino troppo tra loro.

5. ESPERIMENTI

Sono stati condotti numerosi esperimenti con lo scopo di studiare il comportamento di entrambi gli strumenti di classificazione utilizzati su di entrambi i *database*. Come già detto, considerati i problemi di scarsità di dati e della loro affidabilità abbiamo condotto questi esperimenti utilizzando prevalentemente le due classi NEUTRO e NON-NEUTRO. In Tabella 6 sono rappresentati i risultati finali. Dall'analisi degli esiti di classificazione si nota come, a parità di altri fattori, i risultati ottenuti con il *database* in lingua tedesca siano evidentemente migliori. Questo fenomeno può essere spiegato col fatto che, come precedentemente affermato, questa raccolta di dati presenta una maggior ricchezza di frasi emotivamente significative. Questa stessa caratteristica si riflette anche sulle diverse prestazioni dei due classificatori: gli alberi sembrano essere penalizzati dal bilanciamento artificiale dei dati che, per il *database* Targhe è molto più consistente. D'altro canto non si può escludere che il *set* di parametri utilizzati possa essere più efficiente per una lingua piuttosto che per l'altra a causa di un differente utilizzo dell'accento frasale, oppure alla diversa qualità delle annotazioni (Buckow, 2003).

Classificatore	CART		NN	
Database	Targhe	Sympafly	Targhe	Sympafly
RR	73.2%	73.9%	74.2%	73.5%
CL	70.7%	72.1%	69.4%	74.1%

Tabella 6: Risultati di classificazione di due classi (NEUTRO vs. NON-NEUTRO) utilizzando entrambi i *database* ed entrambi i classificatori. Per entrambi i *database* sono stati utilizzati i PDA migliori (NCC per Targhe e UERLN per Sympafly).

Nonostante siano strumenti lineari, le reti neurali ottengono risultati migliori dei CART. Questo aspetto è sorprendente, e fa pensare che strumenti non lineari, ma più stabili degli alberi considerati, possano probabilmente raggiungere prestazioni ancora migliori.

Successivamente abbiamo anche effettuato dei tentativi di classificazione cercando di studiare l'efficacia di singoli parametri o gruppi di parametri. Ci siamo focalizzati in particolare sull'efficienza di *features* legate alla frequenza fondamentale del segnale vocale e sull'utilità della precisione degli algoritmi di estrazione (PDA). I risultati offrono interpretazioni interessanti: in primo luogo si vede, come del resto era prevedibile, che le caratteristiche del PDA utilizzato influenzano la classificazione finale se si utilizzano solo i parametri relativi al *pitch*. Analizzando la Tabella 5 e la Tabella 7 e ricordando come le prestazioni del PDA UERLN siano paragonabili al PDA NCC, si nota un degrado dei

risultati ottenuti con il PDA meno performante. D'altra parte le differenze di riconoscimento scompaiono quasi completamente dopo l'introduzione di tutte le altre *features* a disposizione. Almeno per quanto riguarda le reti neurali l'utilizzo di PDA efficienti diviene praticamente irrilevante.

Database	Sympafly					
Classificatore	CART			NN		
PDA	UERLN	NCC	WAC	UERLN	NCC	WAC
Solo parametri derivanti dal *Pitch*						
RR	68.8%	69.6%	64.4%	77.5%	77.0%	74.6%
CL	65.1%	66.4%	64.1%	68.6%	68.6%	58.6%
Tutti i parametri						
RR	73.9%	73.4%	73.4%	73.5%	72.5%	73.1%
CL	72.1%	71.7%	70.7%	74.1%	74.6%	72.3%

Tabella 7: Risultati di classificazione utilizzando due classi e due differenti insiemi di parametri: il primo blocco utilizza solo *features* derivanti dal *pitch*, il secondo tutte le *features*. Ogni risultato è stato ripetuto utilizzando tre differenti estrattori di *pitch*.

Infine abbiamo anche provato a utilizzare per la classificazione l'etichettatura originale, selezionando esclusivamente le tre classi più frequenti, vale a dire NEUTRO, RABBIA e NOIA, e classificando ognuna di esse contro le altre. Questi esperimenti sono stati condotti esclusivamente sul *database* Targhe utilizzando i CART. In Tabella 8 si nota come alcune classi (RABBIA) siano più discriminabili rispetto alle altre. Questo fatto può suggerire che l'utilizzo delle due sovra-classi NEUTRO e NON NEUTRO possa in futuro venir sostituita dalle classi POSITIVO e NEGATIVO, magari ottenendo risultati migliori senza penalizzare l'informazione raccolta.

Classi	RABBIA *vs*. altre	NOIA *vs*. altre	NEUTRO *vs*. altre
RR	66.2%	49.6%	53.0%
CL	67.6%	53.2%	56.6%

Tabella 8: Classificazione di una classe emotiva contro le altre due. Il *database* utilizzato è Targhe, mentre per classificatore sono stati considerati solo gli alberi.

6. CONCLUSIONI

Il *set* di parametri prosodici utilizzato, precedentemente sviluppato con scopi differenti come l'individuazione di interruzioni prosodiche o dell'accento di frase si è rivelato essere un buon punto di partenza per la realizzazione di un sistema per l'individuazione di emozioni nel parlato spontaneo. In particolare ci siamo limitati alla realizzazione di un sistema funzionante ma limitato alle sovra categorie di parlato NEUTRO e NON-NEUTRO, ottenendo risultati comparabili a quelli di sistemi focalizzati su *task* simili. Sono stati condotti esperimenti con *database* con caratteristiche simili ma linguisticamente differenti e con diversi classificatori. Ciò nonostante le prestazioni riscontrate si sono rivelate uniformi e coerenti. Più in particolare è stata studiata l'influenza dei parametri legati al *pitch*, riscontrando che altri parametri prosodici mascherano la precisione dei singoli PDA, minimizzando le prestazioni finali.

In futuro si pensa di focalizzare lo studio su *database* più ricchi e più bilanciati, come ad esempio il *corpus* CEICES (Batliner *et al*., 2004b). Si cercherà inoltre di classificare

direttamente le classi utilizzate per l'etichettatura senza ricorrere a mappature in sovra-classi. Si cercherà naturalmente anche di perfezionare l'utilizzo degli strumenti di classificazione tramite l'adozione di multiclassificatori.

RINGRAZIAMENTI

Questo lavoro è stato parzialmente finanziato dal progetto europeo PF-STAR (IST–2001–37599). Si desidera inoltre ringraziare Silvia Rocchi per la preparazione del *database*.

7. BIBLIOGRAFIA

Angelini, B.; Brugnara, F.; Falavigna, D.; Giuliani, D.; Gretter, R.; Omologo, M., 1993. Automatic Segmentation and Labeling of English and Italian Speech Database. In *Proceedings of Eurospeech*, Germany.

Batliner, A.; Hacker, C.; Steidl, S.; Noeth, E.; Haas, J., 2004a. User states, user strategies, and system performance: How to match the one with the other. In *Proceedings of the International Speech Communication Association Tutorial and Research Workshop on Error Handling in Spoken Dialogue Systems*, Chateau d'Oex, France, 5-10.

Batliner, A.; Hacker, C.; Steidl, S.; Noeth, E.; D'Arcy, S.; Russel, M.; Wong, M., 2004b. You stupid tin box – children interacting with the AIBO robot: A cross-linguistic emotional speech corpus. In *Proceedings of the IV International Conference on Language Resources and Evaluation*, Lisbon, 171-174.

Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C. J., 1984. *Classification and Regression Trees*. Pacific Grove, CA, USA: Wadsworth and Brooks.

Buckow, J., 2003. *Multilingual Prosody in Automatic Speech Understanding*. Berlin: Logos.

Cattoni, R.; Danieli, M.; Sandrini, V.; Soria, C., 2002. ADAM: the SI-TAL Corpus of Annotated Dialogues. In *Proceedings of the III International Conference on Language Resources and Evaluation*, Las Palmas, Spain.

Cowie, R.; Douglas-Cowie, E.; Tsapatsoulis, N.; Votsis, G.; Kollias, S.; Fellenz, W.; Taylor J., 2001. Emotion Recognition in human-computer interaction. *Institute of Electrical and Electronics Engineers Signal Processing Magazine*, 18, 32-80.

Kiessling, A., 1997. *Extraktion und Klassifikation prosodischer Merkmale in der automatischen Sprachverarbeitung*. Aachen, Germany: Shaker Verlag.

Medan, Y.; Yair, E.; Chazan, D., 1991. Super resolution pitch determination of speech signals. *Institute of Electrical and Electronics Engineers Transaction of Signal Processing*, 39, 40-48.

Plante, F.; Ainsworth, W. A.; Meyer, G., 1995. A pitch extraction reference database. In *Proceedings of European Conference on Speech Communication and Technology*, 837-840.

Rabiner, L. R.; Cheng, M. J.; Rosenberg, A. E.; McGonegal, C. A., 1976. A comparative performance study of several pitch detection algorithms. *Institute of Electrical and Electronics Engineers Transaction on Acoustic, Speech and Signal Processing*, 24, 399-417.

Shimamura, T.; Kobayashi, H., 2001. Weighted Autocorrelation for Pitch Extraction of Noisy Speech. *Institute of Electrical and Electronics Engineers Transaction on Speech and Audio Processing*, 9, 727-730.

Zell, A., 1994. *Simulation Neuronaler Netze*. Addison-Wesley.