POLITECNICO DI MILANO
Corso di Laurea MAGISTRALE in Ingegneria Informatica
Dipartimento di Elettronica, Informazione e Bioingegneria

Speech2Emotion: un'applicazione per l'estrazione dello stato emozionale del parlatore da audio e testo

Relatore: Prof. Licia Sbattella
Correlatore: Ing. Roberto Tedesco

Tesi di Laurea di:
Danilo Petrone, matricola 782806

Anno Accademico 2014-2015

Al coso guido...

Sommario

Il lavoro svolto ha visto la realizzazione di una applicazione che svolge la funzione di estrazione dello stato emozionale di un parlatore, partendo dall'analisi di un file audio registrato, contenente una espressione pronunciata, e la sua trascrizione testuale.

Tale applicazione è stata sviluppata facendo leva su meccanismi di *feature extraction*, al fine di poter sintetizzare modelli emozionali, mediante l'ausilio di tecniche di classificazione dei dati così raccolti.

La ricerca rappresenta il proseguo e l'estensione del lavoro di ricerca precedente, che ha visto lo sviluppo dell'applicazione PrEmA [26].

Il lavoro di ricerca, oltre a quello di realizzare un sistema funzionante, si è proposto di esplorare, valutare, confrontare e migliorare, le diverse soluzioni esistenti, raggiungendo risultati incoraggianti e tracciando le linee guida rispetto a prosegui futuri.

Ringraziamenti

Ringrazio la mia famiglia, inesauribile fonte di pazienza, incoraggiamento e sicurezza.
L'ingegnere Roberto Tedesco, per l'assistenza nello svolgimento di tutto il lavoro. I miei coinquilini ed amici, per avermi accompagnato durante questa bellissima odissea.
Me stesso, per averci creduto.

Indice

Capitolo 1

Introduzione

1.1 Motivazioni

Uno dei campi di conoscenza su cui la ricerca sta concentrando l'attenzione è sicuramente quello dell'interazione della macchina con l'uomo e della sua integrazione anche in quelle circostanze dove, fino a pochi anni fa, sembrava impensabile. In particolare lo svolgimento è stato dettato dalla necessità sempre crescente di esplorare un mondo apparentemente(?) soggettivo come può essere uno stato emozionale ed analizzarlo con un occhio scientifico.

1.2 Obiettivi

L'obiettivo che ci si è proposti di raggiungere, ha riguardato lo sviluppo di una applicazione che permettesse di estrarre lo stato emozionale di un parlatore, sulla base dell'analisi del contenuto informativo all'interno del file audio della registrazione di tale espressione.

Il sistema rappresenta il proseguo del lavoro che ha visto lo sviluppo del sistema PrEmA, descritto in [26], [20].
Il lavoro svolto, infatti, vede il suo modello concettuale condiviso con quello sopracitato, che guarda in primis, allo sfondo emotivo, come la combinazione di caratteristiche comuni che si riversano nell'espressività del singolo individuo, e che quindi seguono un pattern, in generale, un modello.
Nei capitoli successivi saranno presentati i modelli che la ricerca ha sintetiz-

zato rispetto alla natura di una emozione.
In particolare si è puntata la ricerca rispetto all'estrazione dello stato emotivo da file audio, funzionalità già implementata in PrEmA, facendo leva su meccanismi di classificazione di vettori di feature. Saranno proprio questi meccanismi argomento di indagine in questo lavoro, al fine di ottenere prestazioni globali migliori.
Da sottolineare è che il dataset utilizzato per lo sviluppo di PrEmA, sarà lo stesso analizzato in questo lavoro. Questa caratteristica permetterà di valutare e confrontare in seguito i risultati ottenuti, anche rispetto ad una applicazione che, come sarà chiaro in seguito, condivide un componete decisivo, come il dataset.
L'applicazione, inoltre, ha previsto l'analisi aggiuntiva del contenuto informativo riguardante la trascrizione testuale dell'espressione analizzata, con il fine di valutare il contributo di tale analisi all'efficacia del sistema.

In generale, data la natura sperimentale di tale ricerca, si è puntato anche ad analizzare diverse combinazioni dei componenti che avrebbero costituito il sistema nella sua configurazione ultima, al fine di garantirsi una buona base su cui proseguire la ricerca.

1.3 Struttura della tesi

La tesi è strutturata nel modo seguente:

Capitolo 2 : si mostra lo stato dell'arte rispetto alle soluzioni adottate per sistemi simili, o che implementano alcune funzioni di interesse per il nostro problema. Inoltre si pone in evidenza quanto la ricerca stia spingendo verso tale direzione e quanto i risultati raggiunti lascino ben sperare rispetto a miglioramenti significativi.

Capitolo 3 : si descrive la natura del problema analizzato. In particolare si mette in evidenza la posizione del problema tra la conoscenza di strumenti in grado di creare modelli legati a fenomeni, ed, invece, una conoscenza ancora troppo esile rispetto al legame tra i meccanismi di generazione di una emozione e la loro esternazione attraverso gli sva-

riati canali comunicativi, tra cui, naturalmente, il canale vocale.

Capitolo 4 : si illustra il modello dell'applicazione messo a punto. Nello specifico si analizzano tutti i metodi che, nei vari punti del sistema, meglio realizzano la funzione che è necessario implementare. Una volta confrontati tali metodi, si realizza il modello finale dell'applicazione.

Capitolo 5 : si descrive l'architettura dell'applicazione sviluppata, sottolineando i componenti messi a punto, che implementano le funzioni di interesse e le loro interazioni.

Capitolo 6 : si elencano i risultati ottenuti con le varie configurazioni di sistema, fornendo figure di merito per la stima dell'accuratezza e dei tempi di computazione.

Capitolo 7 : le conclusioni, in cui si riassumono gli scopi, le valutazioni dei risultati ottenuti e le prospettive future.

Capitolo 2

Stato dell'arte

2.1 Introduzione

Negli ultimi anni è diventato di interesse globale l'estrazione di informazioni collaterali rispetto al messaggio primario veicolato da una espressione verbale.
A tale fine, la ricerca ha puntato la sua attenzione verso la messa a punto di sistemi di tipo *machine learning, supervised*, che vedessero una preliminare raccolta di dati che permettesse, attraverso una fase di *training*, di identificare dei modelli che rappresentassero al meglio i dati a disposizione e che quindi permettessero di sintetizzare nuove informazioni utili.
Basandosi su una logica di *feature extraction*, i sistemi di questo tipo puntano all'estrazione di caratteristiche che poi saranno discriminatorie al momento della classificazione dello stato emozionale di un parlatore in una espressione.
Attualmente la ricerca punta l'attenzione su diversi aspetti della problematica, poichè esistono ancora punti poco chiari su quali siano le logiche che generano a monte una emozione, e quanto questi meccanismi siano influenzati da diversi parametri quali:

fattori ambientali : sfondo culturale, lingua ...

fattori individuali : il sesso, l'età, personalità...

Inoltre è da sottolineare che i dati relativi le prestazioni dei vari sistemi descritti, non sono direttamente confrontabili, proprio per i motivi sopracitati. Infatti tutti i sistemi condividono l'obiettivo, ma è come tentano di

raggiungere tale obiettivo che invece risulta essere diversificato. In particolare si vedranno quanto influenti saranno i dati utilizzati, la loro quantità, qualità, diversificazione; quanto determinanti siano le diverse tecniche di preprocessing, quindi di classificazione. Quindi tutte le figure di merito fornite dovranno essere contestualizzate, prorpio rispetto alle condizioni iniziali e i presupposti del sistema stesso da sviluppare.

Fatta questa premessa, analizziamo quelli che sono alcuni sistemi che hanno gettato luce su questa problematica e quali sono stati gli approcci al problema e le soluzioni quindi suggerite.

2.2 Analisi Audio e Testuale con Support Vector Machine

Di seguito si presenta il sistema messo a punto in [33]. L'approccio al problema è principalmente basato sul concetto di *feature extraction*: in particolare si fa riferimento ad un set di caratteristiche proprie del segnale audio registrato, *misurabili ed identificative* delle variazioni dell'espressività del parlatore e delle sue caratteristiche.
Il sistema parte, quindi, dall'estrazione di tali caratteristiche; ciò che si otterrà sarà un vettore di feature, che identifica il sample di registrazione analizzato.
Estendendo questo processo a tutti i file di registrazione, avremo a disposizione un dataset costituito dai sample di registrazione, ridotti a vettori di feature.

Si fa riferimento a caratteristiche acustiche, proprie del file audio:

- 4th-order Legendre parameters for the pitch contour;
- 4th-order Legendre parameters for the energy contour;
- 4th-order Legendre parameters for the formant one F1 contour;
- 4th-order Legendre parameters for the zero crossing rate ZCR contour;
- Maximum energy;
- Maximum smoothed energy;

- Minimum, median, and standard deviation of the pitch contour;
- Minimum, median, and standard deviation of the energy contour;
- Minimum, median, and standard deviation of the smoothed pitch contour;
- Minimum, median, and standard deviation of the smoothed energy contour;
- Ratio of the sample number of the upslope to that of the downslope for the pitch contour;
- Ratio of the sample number of upslope to that of the downslope for the energy contour;
- Pitch vibration.

Tali feature sono calcolate mediante una *short term analisys*, il che, prevede una stima spettrale di frame temporali del segnale audio in ingresso. In più sono considerati valori di minimo, massimo, mediana, deviazione standard delle stesse misurazioni sopracitate:

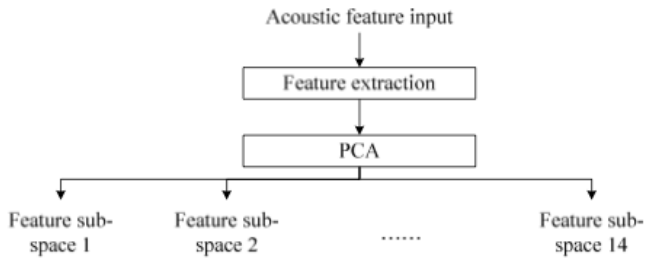


Figura 2.1: Modulo di analisi audio

In figura 2.1 vediamo come il set di feature abbia subito un'analisi del tipo Principal Component Analysis, al fine di valutare il potere informativo, quindi discriminatorio, di *subsets* ottenuti dal set iniziale di features.

A questa analisi è accostata una di tipo *semantica*, che prevede, invece, dapprima l'estrazione del contenuto testuale (trascrizione), attraverso uno

speech recognizer, basato su un modello HMM, avente in input i coefficienti MFCC come feature acustiche [33], come mostrato in figura 2.2.
In particolare, non si è interessati a tutte le parole, bensì solo a quelle cui è associato un valore di trigger (*emotion trigger*) che indichi un contenuto emozionale al loro interno.
Questo è possibile grazie ad un lavoro di *tagging manuale*, di un *corpus*, ossia di una raccolta strutturata di vocaboli.
In questo caso la struttura consiste nell'associare a ciascun vocabolo un contenuto emozionale, che sia da scegliere tra (Gioia, Paura, Rabbia, Tristezza, Sorpresa, Antipatia), basandosi sulla logica del corpus HowNet.
Questo sistema prende il nome di *keyword spotting system*.
Successivamente si procede con l'analisi emozionale delle parole facendo leva su una *sematic network*, ossia una rete che tiene conto delle relazioni tra le parole presenti e le loro implicazioni emozionali.
Questo permette di associare un *emotion modifier*, ossia un fattore associato a vocaboli che rafforzano o meno il significato emozionale di una *keyword*.
Si faccia riferimento all'esempio: "siamo molto felici": in questo caso alla parola chiave "felici", si associa la parola "molto" che ne rafforza il contenuto emozionale.
Dato che i dati in ingresso sono costituiti da espressioni proveninti da dialoghi di testi teatrali, si punta a misurare anche la propagazione dello stato emozionale del parlatore lungo l'evoluzione dell'eloquio.

Il sistema vede la classificazione delle feature acustiche mediante l'algoritmo SVM (vedi dopo), mentre la *semantic network statica* fornisce il contenuto emozionale proveniente dall'analisi testuale delle singole espressioni.
La classificazione finale prevede l'integrazione delle due stime come specificato in [33].

I risultati di tale sistema sono riposrtati nella tabella 2.3:
Si può notare come il sistema effettivamente sia in grado di distinguere tra le varie classi di emozioni e quindi di classificare correttamente i sample.
Ciò che resta evidente è la dipendenza rispetto ad un corpus di vocaboli etichettati, che prevede l'intervento diretto umano.
Resta comunque un ottimo esempio di sistema di estrazione dello stato emozionale, infatti sarà su questo modello che si andrà a puntare l'attenzione e la ricerca.

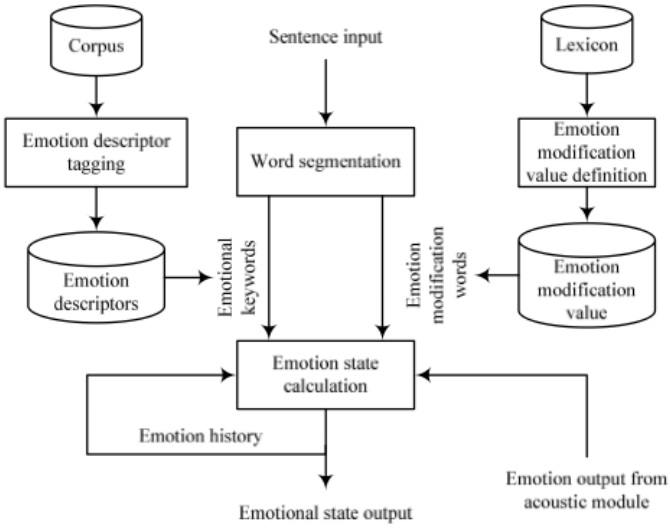


Figura 2.2: Modulo di analisi testuale

	Inside	Outside
Happiness	84.44%	66.67%
Sadness	82.98%	73.91%
Anger	79.66%	67.65%
Fear	78.24%	62.37%
Surprise	80.33%	69.52%
Disgust	76.51%	70.43%
Neutral	88.24%	76.84%
Average	81.49%	69.63%

Figura 2.3: Accuratezza media del sistema di riconoscimento: Audio+ Testo (Inside), solo testo (Outside)

2.3 Analisi Audio con Artificial Neural Network ANN

Accanto a questo tipo di sistema vediamo un altro tipo di approccio, che prevede la sola analisi del contenuto informativo all'interno del segnale audio [2].

In particolare, la fase iniziale prevede la riduzione dei file audio in ingresso a vettori di feature cosi costituiti:

- Pitch media
- Pitch varianza
- Intensity media
- Intensity varianza
- Jitter
- Shimmer
- Speech rate #1 numero di periodi unvoiced per frame
- Speech rate #2 1/lunghezza media periodi voiced per frame

Ciascuna feature è calcolata su *frame di 30 ms* con *10 ms di overlap*.
Infine ciascuna feature è stata normalizzata rispetto alle espressioni che esternano *neutralità emotiva*.
Questo vuole evidenziare come si cerchi di rendere il sistema indipendente dal parlatore, cercando di far risaltare le differenze con uno stato di neutralità, piuttosto che confrontarle *assolutamente* tra loro.

Quattro reti neurali -*Feed Forward Neural Network*- saranno utilizzate per classificare i campioni vocali, in quattro classi di emozioni. In figura **??** è mostrata una rappresentazione di uno schema di una ANN.
Ciascuna rete ha dodici neuroni di ingresso ed una uscita nell'intervallo $[0, 1]$.

Ogni rete è modellata per distinguere una singola emozione dalle altre tre utilizzate.
Il metodo funziona come segue: si considera un livello iniziale di neuroni, che nella fattispecie saranno dodici, pari al numero di feature.
In un sistema feed-forward i nodi sono detti *Processing Elements* e sono disposti in strati distinti e ogni strato riceve l'input dal livello precedente, e fornisce l'output allo strato successivo. Non vi è alcun feedback. Ciò significa che i segnali provenienti da uno strato non vengono trasmessi a un livello precedente.

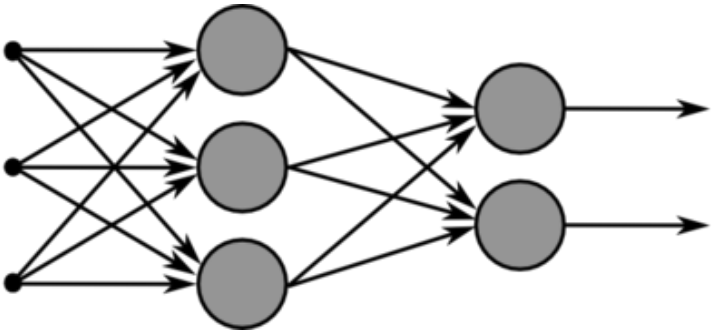


Figura 2.4: Multilayered Artificial Neural Network

Questo può essere indicato matematicamente come:

$$w_{ij} = 0 \quad if\ i = j$$
$$w_{ij} = 0 \quad if\ layer(i) \leq layer(j)$$

calcolo dei pesi tra i vari livelli:

$$\mathbf{y}_l = \rho_l(\mathbf{w}_l \mathbf{x}_l + \mathbf{b}_l)$$
$$\mathbf{y}_l = \rho_l(\mathbf{w}_l \rho_{l-1}(\mathbf{w}_{l-1} \mathbf{x}_{l-1} + \mathbf{b}_{l-1}) + \mathbf{b}_l)$$

Questo metodo può essere esteso per calcolare l'output di una rete con un numero arbitrario di strati. Si noti che aumentando il numero di strati, aumenta anche la complessità di calcolo.
Reti sufficientemente grandi possono rapidamente diventare troppo complesse per un'analisi matematica diretta.

Dopo che il processo di training è completo, la rete dovrebbe produrre un output di 1, o vicino ad 1 per ogni ingresso derivato da un campione vocale con l'emozione corrispondente per cui è stata modellata.
La funzione di trasferimento (*activation function*) utilizzata in tutti gli strati, di tutte le reti è la Log-Sigmoid.

Il metodo di training applicato è stato quello della *Levenberg-Marquardt backpropagation*.
Esso prevede l'aggiornamento dei pesi mediante la minimizzazione dell'*errore*

quadratico medio dei pesi, rispetto all'iterazione precedente, con un algoritmo del tipo *gradient descent*:

Initialize weights at random
repeat
for each sample in the training set
compute sample's output
compute quadratic error
for i = #levels down to 1
compute update for weights at level i
end
update all weights
end
until (all examples correctly classified
or max iterations reached)

L'architettura interna, quindi le dimensioni e il numero dei livelli nascosti della rete, è specifico per ogni emozione.
Ogni rete è modellata su un numero di epoch tra 100 e 200, con circa 2500 campioni in ingresso.

Nella tabella in 2.5 vengono riassunte le prestazioni del sistema così costituito.

Identified / Presented	Neutral	Happy	Sad	Angry
Neutral	0.95	0.22	0.21	0.22
Happy	0.15	0.55	0.23	0.43
Sad	0.24	0.10	0.98	0.18
Angry	0.25	0.29	0.40	0.49

Figura 2.5: Accuratezza media del sistema Artificial Neural Network con feature audio

Possiamo notare come siamo lontani da prestazioni accettabili, questo però non è sufficiente ad escludere questo tipo di approccio al problema, poichè soprattutto il metodo di classificazione risente della natura dei dati in ingresso.
Quindi è bene considerare anche solo di modificare gli stadi precedenti la classificazione e verificare i nuovi risultati.

2.4 Analisi Audio con CodeBook-HMM

In [29] viene presentata una ulteriore variante di tale sistema, che consiste nella configurazione in figura 2.6:

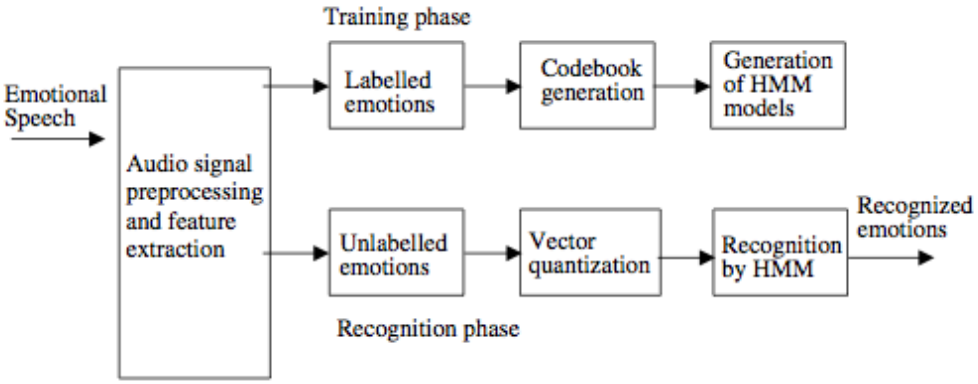


Figura 2.6: Schema generale sistema Codebook-HMM

Il segnale vocale viene campionato a 22,05 kHz e codificato con 16 bit PCM.
I campioni del segnale sono suddivisi in frame di 16 ms ciascuno con 9 ms overlap tra frame consecutivi. I valori tipici, minimi, di frequenza fondamentale (F_0) variano tra 100 e 200 Hz.
La dimensione della finestra è di 16 ms, si estende per circa due periodi di frequenza fondamentale minima (125 Hz), questa per poter garantire risultati accurati della stima dei varie feature.
Il numero totale di frame, N, per essere determinato, dipende dalla lunghezza dell'espressione analizzata.

Per ogni frame, si ottiene un vettore basato su coefficienti normalizzati LFPC - Low Frequency Power Coefficients -. Questi coefficienti, similmente a quelli MFCC, voglio riassumere l'energia contenuta in un segnale, vocale nella fattispecie, in dodici coefficienti, calcolati come l'energia all'interno di una delle dodici bande di un banco di filtri, la cui estensione in frequenza riguarda quella del segnale in analisi, quindi circa 4 Khz.
In particolare per determinare la larghezza delle singole bande si faccia riferimento a:

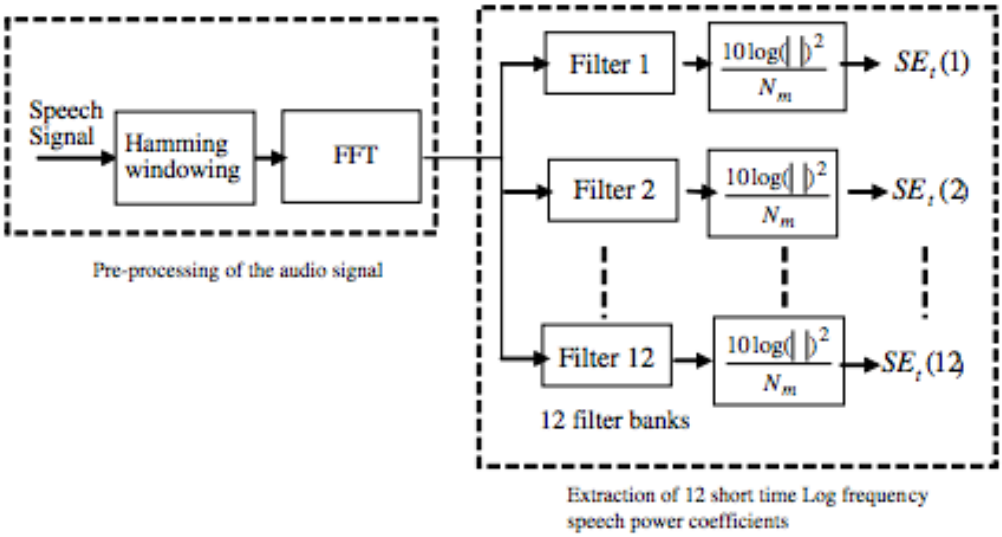


Figura 2.7: Modulo di estrazione dei coefficienti LFPC

$$b_1 = C \tag{2.1}$$

$$b_i = \alpha \cdot b_{i-1} \qquad con\ 0 \leq i \leq 12 \tag{2.2}$$

$$f_i = f_1 + \sum_{j=1}^{i-1} b_j \frac{(b_i - b_1)}{2} \tag{2.3}$$

con C la frequenza di banda centrale
f_1 la frequenza centrale della prima banda
$\alpha\ \epsilon\ [1, 1.4]$ fattore di crescita logaritmico, che determina una larghezza delle bande progressivamente crescente, in accordo con l'intervallo di frequenze scelto. In figura 2.7 sono rappresentati i blocchi che realizzano l'estrazione di tali coefficienti.
Per la sessione di training, un *Universal Codebook* è costruito utilizzando i feature vector di tutte le espressioni riservate per il training.
Nello specifico, a seguito della fase di estrazione delle feature, ogni frame del singolo sample di speech è rappresentato da un vettore. Tale vettore sarà costituito da 12 elementi, nel caso di siano costituiti dai coefficienti LFPC, MFCC e 16 elementi, nel caso si considerino i coefficienti LPCC come feature, rispettivamente.
Per comprimere ulteriormente i dati, per la presentazione alla fase finale del

sistema, una *quantizzazione vettoriale* è eseguita.
Tutti i coefficienti sono normalizzati prima della quantizzazione. Un codebook di dimensione 64 è costruito, utilizzando un ampio insieme di vettori che rappresentano le caratteristiche di tutti i sample di training.
La divisione in 64 cluster viene effettuata secondo l'algoritmo LGB (Linde-Buzo-Gray), generalizzazione dell'algoritmo di Lloyd.
Tutti i vettori che rientrano in un gruppo particolare sono codificati con il vettore che rappresenta il cluster.
La qualità del codebook può essere misurata dal parametro *Distortion*, che è la distanza media di un vettore di osservazione nei dati di training, dal suo corrispondente centroide nel codebook.
La distorsione può essere ridotta aumentando la dimensione codebook.
Nello speech recognition problem, utilizzando i coefficienti MFCC come feature, l'efficienza massima si ha per dimensioni del codebook di 32 o 64.
Un vocabolario più grande significa anche maggiore costo computazionale.
Da esperimenti, si è constatato che anche le prestazioni del sistema proposto non migliorano significativamente estendendo la dimensione codebook oltre 64.
Studi sperimentali [29] dimostrano che una *4-state discrete ergodic HMM* offre le migliori prestazioni rispetto alla struttura *left-right* (le espressioni evolvono nella direzione del tempo, idealmente da sinistra a destra, quindi la direzione di transizione tra gli stati è unica).
Le probabilità di transizione di stato e le probabilità dei simboli di uscita sono uniformemente inizializzati. Una HMM distinta è ottenuta per ciascuna emozione durante la fase di training.
Le probabilità dei simboli di uscita vengono smussate con una distribuzione uniforme per evitare la presenza di quantità di probabilità troppo piccole o probabilità nulle.
Il 60% degli eloqui per emozione di ogni parlatore sono stati usati per il training di ogni modello delle emozioni.
Dopo il training, sei modelli HMM vengono ottenuti per ogni parlatore, uno per ogni classe di emozione.
Le prove di testing, di riconoscimento, vengono effettuate sul restante 40% degli eloqui, con l'ausilio di un $forwardalgorithm$.
Il sistema proposto è *text independent*, ma *speaker dependent*, infatti frasi diverse sono utilizzate per ciascun parlatore e i modelli sono ottenuti per ciascuno di loro.

Quando un enunciato di test viene presentato al sistema, l'espressione viene valutata con l'algoritmo forward, passando attraverso il calcolo di uno score, scaturito dal confronto con i modelli delle emozioni.
Nella figura 2.8 è messa in evidenza l'accuratezza del sistema rispetto al numero di stati della HMM, mentre in figura 2.9 si fa riferimento all'accuratezza del sistema al variare dello speaker e del subset di feature scelto.

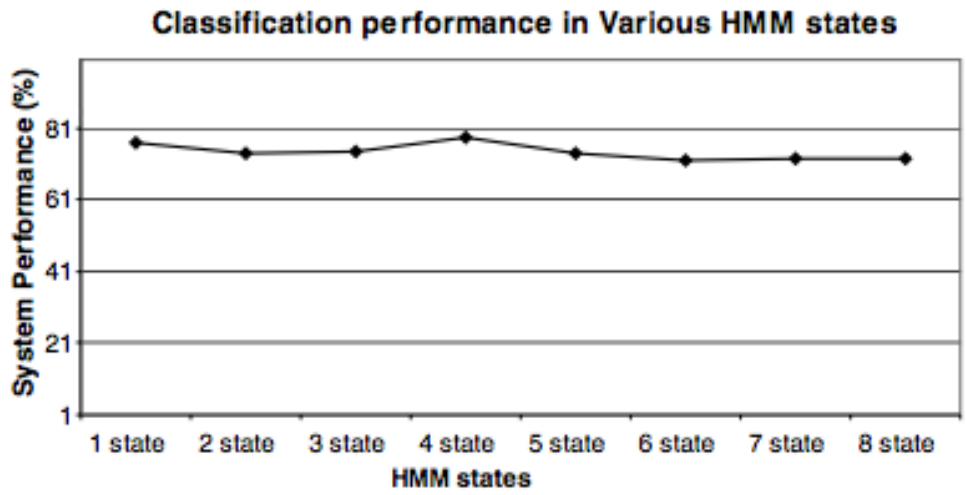


Figura 2.8: Accuratezza media al variare del numero di stati della HMM

Speaker	LPCC		MFCC	
	Burmese	Mandarin	Burmese	Mandarin
S1 (M)	54.2	79.2	58.3	66.7
S2 (M)	69.4	41.7	72.2	37.5
S3 (M)	52.8	54.2	63.9	58.3
S4 (F)	58.3	41.7	54.2	66.7
S5 (F)	62.5	50	66.7	58.3
S6 (F)	58.3	50	62.5	41.7
Average performance	59.3	52.8	63	54.9
Average performance	56.1		59	

Figura 2.9: Accuratezza media nel caso si usino LFPC o MFCC come feature nel caso del Brunense e Mandarino

Quello che emerge dalle prestazioni del sistema descritto, è che un approccio del genere non riesce a raggiungere risultati soddisfacenti. Questo,

partendo dal confronto con gli altri studi presentati, è parzialmente giustificabile con la tipologia e il numero esiguo di feature scelte e che sopratutto, come sarà chiaro più avanti, non tengano conto di quelle caratteristiche tipiche del segnale vocale, strettamente legate alla problematica in esame.

2.5 Analisi Audio con LDA - sistema PrEmA

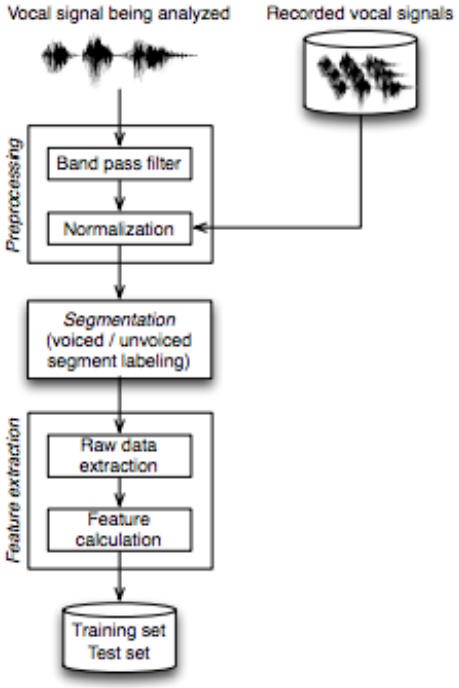


Figura 2.10: Schema blocchi funzionali sistema PrEmA per la costruzione del training/testing set

Un'altra configurazione proposta [26], che rappresenta la base del lavoro svolto, vede l'analisi di file audio, contenenti espressioni recitate da attori. Nella fattispecie, il dataset è costituito da 900 espressioni recitate da sei attori differenti, con tre stili di comunicazione differenti. Tali espressioni sono veicolo di stati emozionali che ricadono in 5 classi:

- Gioia
- Paura
- Rabbia

- Tristezza
- Neutralità

A meno della classe della neutralitá, il dataset descritto, è lo stesso analizzato in questo lavoro. Come mostrato in figura 2.10, ciascuna espressione ha visto dapprima una fase di preprocessing, in cui ciascun sample è stato sottoposto ad un filtraggio passabanda con intervallo $[100Hz \quad 6kHz]$ e normalizzazione dell'intensità ad un valore predefinito. Successivamente si sono estratti, per ciascun file audio, i segmenti *voiced*, ossia quelli in cui ci fosse effettivamente contenuto vocale, seguendo una logica di imposizione di soglie rispetto all'intensità di tali segmenti.
A seguito di tale preprocessing i file audio hanno visto l'estrazione delle seguenti feature:

Pitch (F0) :

- Average [Hz]
- Standard deviation [Hz]
- Maximum [Hz]
- Minimum [Hz]
- 25th quantile [Hz]
- 75th quantile [Hz]
- Median [Hz]

Intensity :

- Average [dB]
- Standard deviation [dB]
- Maximum [dB]
- Minimum [dB]
- Median [dB]

Time :

- Unvoiced frame ratio [%]
- Articulation break ratio [%]
- Articulation ratio [%]
- Speech ratio [%]

Voice quality :

- Jitter [%]
- Shimmer [%]
- HNR [dB]

A seguito dell'estrazione delle feature, i vettori ottenuti vedono l'applicazione di un classificatore basato su una *Linear Discriminant Analysis*.
Il metodo si basa sulla individuazione di combinazioni lineari tra le componenti (feature) dei dati analizzati, basandosi sulla massimizzazione del rapporto tra la *varianza inter-classe* -varianza di una componente rispetto alle altre classi- e la *varianza intra-classe* -varianza di una componente rispetto a alla classe di appartenenza-, in modo da massimizzare la separabilità tra le classi [26].
In figura 2.11 sono descritte le prestazioni esibite dal sistema.

	Predicted emotions				
	Joy	Neutral	Fear	Anger	Sadness
Joy	**63.81**	0.00	18.35	11.79	6.05
Neutral	3.47	**77.51**	2.14	1.79	15.09
Fear	33.75	0.00	**58.35**	6.65	1.25
Anger	10.24	1.16	8.16	**77.28**	3.16
Sadness	5.14	14.44	0.28	0.81	**79.33**

Figura 2.11: Prestazioni sistema PrEmA

Da sottolineare è che tale sistema vede, accanto all'analisi del contenuto emozionale, anche l'*analisi dello stile vocale* del parlatore, attraverso un'analisi analoga a quella descritta [26].

Capitolo 3

Impostazione del problema di ricerca

3.1 Caratteristiche della comunicazione verbale

Se pensiamo alla *voce umana*, come veicolo per lo scambio di informazioni, potremmo cadere nell'errore di far ricadere tale analisi vocale in un contenitore, quale può essere una lingua strutturata.
Essa risulta essere indipendente rispetto alla lingua, infatti la capacità dell'essere umano di produrre suoni mediante il tratto vocale, prescinde da qualsivoglia sovrastruttura esterna.
Questo rappresenta un punto di partenza rispetto allo studio che si sta svolgendo, infatti, guardando alle caratteristiche più primordiali della produzione di suoni da parte dell'essere umano, possiamo notare come in realtà esista un legame tra voce e linguaggio ed esso è proprio la **vocalità**, ossia un insieme di caratteristiche che sono *proprie dell'individuo* (e che accomunano il genere umano) e rappresentano l'esternazione di uno stato interiore.

3.2 Fonetica articolatoria: meccanismi di produzione dei segnali vocali

Ai fini della produzione fisica dei segnali vocali si descrivono brevemente i meccanismi coinvolti.

Tab. 1 **Modi di articolazione**

DENOMINAZIONE	DESCRIZIONE ARTICOLATORIA	AERODINAMICA	ACUSTICA
occlusive	chiusura diaframmatica per tutta la durata del fono con apertura improvvisa e completa alla fine	interruzione del flusso d'aria e brusca fuoriuscita all'apertura del diaframma	silenzio per tutta la durata della consonante e forte rumore alla fine
nasali	nella cavità orale si realizza un diaframma che segue la stessa meccanica delle occlusive, ma il passaggio rinovelare (tra velo pendulo e parete posteriore della faringe) resta aperto	l'aria passa liberamente attraverso le cavità nasali per tutta la durata del fono	mormorio nasale per tutta la durata della consonante
fricative	diaframma in posizione di stretta per tutta la durata del fono	l'aria passa in maniera continua in modo turbolento	fruscio per tutta la durata del fono
affricate	chiusura diaframmatica seguita da una posizione di stretta	interruzione del flusso d'aria e passaggio turbolento nella fase finale	silenzio seguito da fruscio
vibranti	alternanza di più chiusure e aperture diaframmatiche (polivibranti) oppure di una singola apertura e chiusura (monovibranti)	rapido alternarsi di impulsi d'aria e interruzioni del flusso	rapido alternarsi di silenzio e rumore
laterali	il diaframma è chiuso nella parte centrale del canale e aperto ai lati	l'aria passa liberamente ai lati	suono continuo di tipo vocalico, ma meno intenso

Figura 3.1: Modi di articolazione di foni

La combinazione di un determinato *modo di articolazione* e di un determinato *luogo di articolazione*, inteso come l'articolazione delle unità fonetiche - suoni linguistici atomici - che corrisponde ad una articolazione di punti del tratto vocale, come è descritto nelle figure 3.1 e 3.3, che permette alla pressione polmonare dell'aria di seguire percorsi diversi lungo tutto il tratto vocale e di dar luogo, di volta in volta, ad un diverso fono consonantico; questo a sua volta può essere *sonoro* o *sordo* secondo che, nella sua produzione, intervenga o no la vibrazione delle *pliche vocali*, coinvolte, invece, nella produzione di foni vocalici, dalla caratteristica peculiare di essere segnali armonici, come mostrato in figura 3.2.

Partendo da queste considerazioni, possiamo fare un passo ulteriore, ed esso vede l'inserimento di una lingua strutturata all'interno della comunicazione verbale, quindi un **linguaggio**, costituito da regole che permettono alla voce di essere lo strumento con cui l'uomo si rivolge all'esterno, costruendo

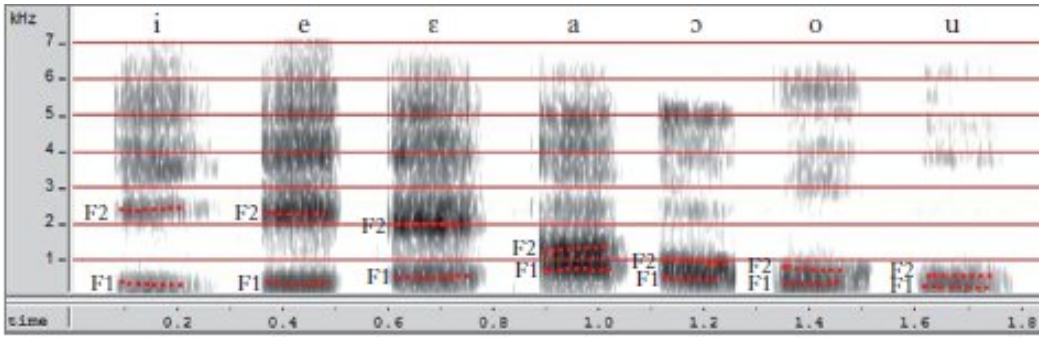


Figura 3.2: Rappresentazione dello spettro di segnali relativi al suono di vocali

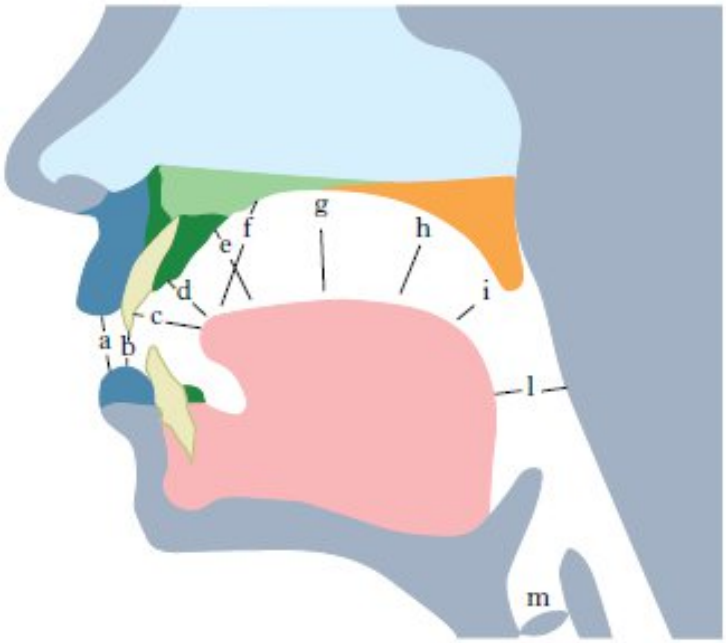


Figura 3.3: Luoghi di articolazione delle unità fonetiche: (a) bilabiale, (b) labiodentale, (c) dentale (o interdentale), (d) alveolare, (e) postalveolare, (f) retroflesso, (g) palatale, (h) velare, (i) uvulare, (l) faringale, (m) glottidale.

relazioni, esprimendo necessità, affermando l'individuo stesso.
Come già preannunciato, vocalità e linguaggio vedono le loro caratteristiche composte da diversi fattori quali:

fattori fonosimbolici : strettamente legati a livello microscopico da fonemi che costituiscono le parole

fattori soprasegmentali : legati alla struttura della frase, che vede l'ana-

lisi di come i singoli foni andranno ad essere concatenati per dar luce al messaggio nella sua interezza. Queste caratteristiche sono scelte del parlatore, e sarà su queste che si farà leva per la nostra ricerca.

fattori fonico-timbrici : caratteristiche proprie del parlatore, quali possono essere le caratteristiche fisiche del tratto vocale che ne determinano la unicità delle caratteristiche meramente qualitative della voce (*timbro vocale*).

La voce infine è **interprete** dei significati associati alla oggettività di una lingua, facendoli soggettivi, rispetto al parlatore.
Quindi possiamo concludere che la voce è la vocalità del linguaggio dell'interprete [20].

3.3 Stile vocale e prosodia

Lo stile vocale rappresenta la scelta del parlatore di una modalità di pronunciare una frase, quindi di veicolare il messaggio.
Esso è il punto d'incontro tra significante e significato e fa leva su meccanismi propri della lingua.
Questo compromette il legame isomorfico, ideale, tra significante e significato, ossia, nel naturale svolgimento di una conversazione, l'uso di forme metaforiche nell'espressività compromette una diretta associazione tra significante e significato, quindi per quanto sia possibile misurare caratteristiche prettamente fisiche come intensità, altezza, etc... resta ancora poco chiaro, ed è questo argomento di esplorazione, poichè è altresì vero che dovrà esistere una griglia referenziale, una sorta di mapping che determini il rapporto tra queste grandezze misurabili e come queste si traducano nella espressione di un sentimento, di una emozione.
Volendo riassumere e descrivere queste caratteristiche possiamo proseguire come segue:

- Pause e la loro distribuzione
- Distribuzione e tipologia di accenti
- Tempo

- Dinamiche
- Timbro
- Intonazione
- Elementi vocali non linguistici

3.3.1 La pausa

La pausa è in sè un elemento quasi dicotomico dell'eloquio, infatti essa rappresenta un silenzio, che però racchiude svariati tipi diversificati di informazioni.
Nella fattispecie, una pausa è in primis una necessità biologica, regolata dal respiro, in particolare nella fase di insiprazione, quindi involontaria, che però può essere in generale controllata, più o meno coscientemente, e che risponde ad un moto emotivo, che può, quindi, sospendere, enfatizzare, in generale, porre in evidenza una particolare espressione, all'interno di un discorso più ampio.
La quantità e la qualità di informazione fornita dall'analisi delle pause, quindi del loro peso all'interno di un eloquio, risiede nelle combinazioni delle caratteristiche proprie della pausa stessa, la loro *quantità e la distribuzione lungo l'espressione*.
Infine è necessario essere in grado di distinguere quali siano le cause che hanno determinato la presenza di una pausa, poichè data la grande varietà di possibilità, risulta evidente quanto sia semplice ambiguare il valore di tali fattori e attribuire informazioni errate ai dati rilevati [20].

3.3.2 L'accento

L'accento rappresenta per una lingua l'elemento principe che permette l'articolazione e la organizzazione (*accento lessicale*) dei foni; costituisce l'elemento di scansione ritmico in una sequenze di parole (*accento fraseologico*).
Tale elemento risulta fortemente dipendente dalla modalità di produzione, quindi dal singolo individuo e quindi si può indagare in questa direzione per l'analisi dello stile vocale [20].

3.3.3 Il tempo

Il tempo, nel caso dello stile vocale, rappresenta la velocità dell'eloquio, ossia quale quotaparte, della durata dell'intera espressione, sia di sole pause, o di parlato; quanto veloce sia l'articolazione delle sillabe, con il fine di distinguere tra gli stati psico-emotivi alla radice di tali fattori.
Più avanti verranno descritti, in funzione delle emozioni considerate, i valori qualitativi di tali fattori, quindi il loro potere di disambiguare tra le emozioni stesse [20].

3.3.4 Le dinamiche

Tale fattore è legato alla intensità dello stile vocale, quindi del suo andamento.
Essa è il frutto della pressione ipolaringea e della forza fonoespiatoria, che rispondono, ach'esse, ad un fine presupposto dal parlatore.
Possiamo immaginare quanto la voce in una circostanza, in generale poco usuale, quale può essere un pericolo imminente o una gioia improvvisa, possa essere un veicolo preponderante dello stato d'animo del parlatore, da cui dipenderà l'incisività nell'evidenza di tale espressione [20].

3.3.5 Il timbro

Rappresenta la caratteristica più legata al parlatore, esso infatti dipende dalla struttura fisica di tutto l'apparato fonatorio, atto alla produzione di suoni. Questo è di per sè uno dei punti nodali rispetto alla disambiguazione tra le emozioni sottese da una espressione, poichè è necessario dapprima rendere il sistema indipendente, o quanto meno robusto, rispetto al parlatore, in modo da estrarre i meccanismi di evoluzione dell'eloquio solo rispetto all'emozione espressa, piuttosto che al fatto che l'emozione sia espressa, ma da quel particolare individuo.
Il timbro rappresenta altresì proprio l'elemento in cui confluiscono tutti gli aspetti più personali, quindi anche le stesse emozioni; in sostanza, una voce a-timbrica, sarebbe una artificiale, sintetica, quindi priva di emozioni [20].

3.3.6 L'intonazione

L'intonazione rappresenta l'evoluzione melodica impressa dal parlatore, durante l'eloquio.
Essa fa riferimento alla *curva melodica*, intesa come la curva che rappresenta l'evoluzione della frequenza fondamentale F_0, lungo tutta la durata dell'espressione. In figura 3.4, viene rappresentato l'esempio di curva melodica della parola "Anna". In generale la variazione di intonazione è funzionale ad una codifica socialmente utile per intendere una espressione come una interrogazione, piuttosto che una affermazione od un ordine.
All'atto pratico, il parlatore si troverà a scandire lo stile della conversazione attaverso il principio della *focalizzazione*, che permetterà, attraverso variazioni della curva di intonazione, di mettere o meno in evidenza particolari informazioni. L'intonazione rappresenta sicuramente un fattore cardine nella discriminazione tra le varie emozioni [20].

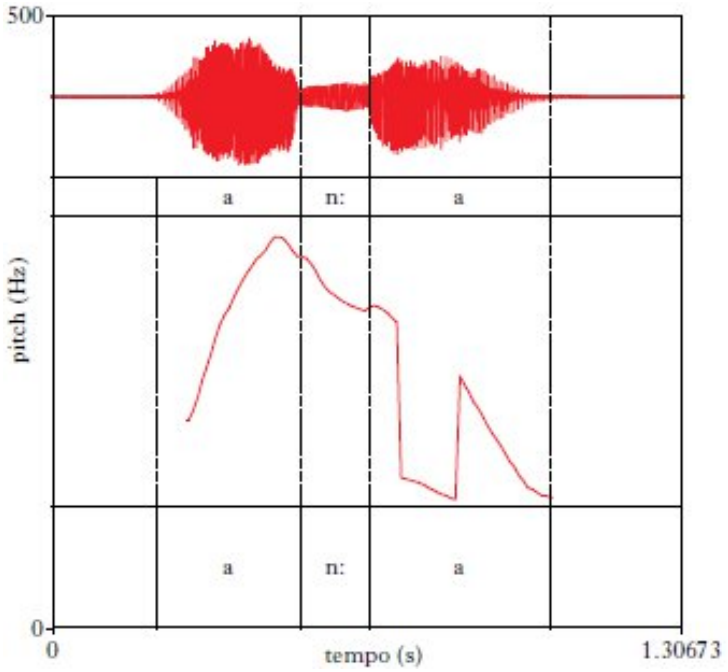


Figura 3.4: Esempio di curva melodica della parola Anna

3.3.7 Elementi vocali non linguistici

In questa categoria ricadono tutti quei suoni-rumori che il parlatore produce legati strettamente all'inconscio dello stesso e quindi di difficile identificazione e classificazione.
Questa considerazione però fa riflettere sulla qualità di informazione che questi fattori possono avere all'interno di una espressione e quanto può essere determinante studiarne il loro comportamento, anche a fronte del fatto che tali elementi siano strettamente personali, quindi di difficile generalizzazione.

3.4 Le emozioni e il linguaggio naturale

In questa parte del documento, andremo ad analizzare quali siano le caratteristiche, nell'espressività umana, che determinano, in via quanto più possibile inequivocabile, la definizione di una emozione.
Partiamo quindi da una definizione in psicologia, dove il paradigma *activation evaluation space,* rappresenta un semplice, ma efficace approccio nel considerare una emozione, come la *cognizione e l'elaborazione di una risposta ad uno stimolo*, che si riversa in una scelta (rapporto causa-effetto), che vedrà la sua risoluzione in una modalità *fight or flight* (combatti o scappa), che vede le sue origini in un aspetto meramente biologico, ossia una emozione come un pattern di azioni dettate da un principio primitivo di sopravvivenza rispondente alle naturali leggi evoultive, che si riversano quindi in una differenziazione dei parametri vitali, come la velocità del ritmo cardiaco, pressione sanguigna, diametro delle pupille, così per le espressioni facciali, o il tono di voce.
Possiamo quindi riassumere e schematizzare queste caratteristiche legate all'origine di una emozione come:

Componente neurofisiologica di attivazione : *arousal*, è una condizione temporanea del sistema nervoso, in risposta ad uno stimolo significativo e di intensità variabile, di un generale stato di eccitazione, caratterizzato da un maggiore stato attentivo-cognitivo di vigilanza e di pronta reazione agli stimoli esterni.

Componente cognitiva : mediante la quale l'individuo confronta la situazione-stimolo con i propri bisogni

Componente motoria : che mette in atto le disposizioni ad agire

Componente espressiva : mediante la quale l'individuo manifesta le prpoprie intenzioni, in funzione dell'interaizone sociale.

Componente soggettiva : che monitora il vissuto sperimentato dall'individuo.

Possiamo notare quindi come l'espressione di una emozione non si limiti solo a comunicare un'informazione, ma ha anche come obiettivo quello di promuovere ed evidenziare reazioni empatiche, infatti accanto a manifestazione di tipo lessicale (legata quindi ad un linguaggio), possiamo individuare attività complementari di natura acustica (urlo, inflessione vocale), gesturale ed espressiva; questo ci permette di distinguere tra [20]:

Emozioni indistinte : viste come il risultato del processo dinamico esistenziale, ossia sono tutte quelle sfuggenti espressioni dell'individuo che non vedono una chiara e definita natura

Emozioni discrete : quelle identificabili, definibili, discrete.

3.4.1 La fisiologia delle emozioni

Dal punto di vista della fisiologia nella produzione del discorso, il sistema nervoso simpatico è responsabile delle emozioni di Rabbia, Paura e Gioia.
Come conseguenza, la frequenza cardiaca e la pressione sanguigna aumentano, la bocca diventa secca e ci sono occasionali tremori muscolari.
L'eloquio è corrispondentemente forte, veloce ed enunciato con una forte energia ad alta frequenza.
D'altra parte, con l'eccitazione del sistema nervoso parasimpatico, come con la Tristezza, la frequenza cardiaca e la pressione arteriosa diminuiscono, mentre la salivazione aumenta, producendo un discorso è lento e con poca energia ad alta frequenza.
Gli effetti corrispondenti sul discorso di tali cambiamenti fisiologici sono mostrati, dall'energia globale del segnale prodotto, così come la sua distribuzione su tutto lo spettro di frequenza, così come la frequenza e durata delle pause all'interno del segnale vocale [5].

Nella tabella 3.1 si fa rifermineto a valori qualitativi delle caratteristiche legate allo stile vocale, che sono legate all'emozione sottesa dall'eloquio [29].

3.5 Modelli descrittivi delle emozioni

In questa sezione si vogliono descrivere quali siano i modelli su cui la ricerca ha fatto leva per poter sintetizzare un sistema che andasse oltre l'inserimento di una espressione all'interno di una classe di emzioni , e che potesse, invece, generare un modello di approssimazione dei meccanismi di origine dell'emozione stessa, quindi indipendente dal numero più ampio possibile di fattori che legano l'emozione alle caratteristiche del parlatore.

3.5.1 Modello di Plutchik

Con questo modello, Plutchik ha sintetizzato le caratteristiche delle emozioni primarie: Gioia, Accettazione, Paura, Sorpresa, Tristezza, Schifo, Collera, Aspettativa.
Queste vengono distinte anche in base ad un loro grado di intensità di espressività più o meno attenuata (tristezza, dolore, collera, noia, sorpresa..).
Grazie allo schema in figura possiamo notare come una combinazione di emozioni primarie, affiancata ad un livello di espressività, possono dar vita ad emozioni piùcomplesse.
Ad esempio:
Amore è il risultato di *Accettazione* e *Gioia*;
Spavento è invece la combinazione di *Paura* e *Sorpresa*.
In figura 3.5 vediamo una rappresentazione di come sia possibile combinare tra loro emozioni primarie e le loro diverse intensità, dando vita ad entità diverse, più complesse, quindi più precise [20].

3.5.2 Modello di Juslin

Tale modello nasce da un approccio più teorico legato all'analisi di emozioni primarie, comuni a tutte le culture, che hanno le loro radici nella famiglia dei primati e che si rivelano in pattern fisiologici che vanno a svilluparsi nell'età infantile.
Esse hanno modalità di espressione specifiche, sostanzialmente indipendenti

Emotion	Anger	Surprise	Joy	Fear	Disgust	Sadness
Pitch contour	Angular frequency curve, stressed syllables ascend frequently and rhythmically, irregular up and down inflection, level average pitch except for jumps of about a musical fourth or fifth on stressed syllables.	Sudden glide up to a high level within the stressed syllables, then falls to mid-level or lower level in last syllable.	Descending line, melody ascending frequently and at irregular intervals	Disintegration of pattern and great number of changes in direction of pitch.	Wide, downward terminal inflects.	Downward inflections.
Average pitch	Increased in mean	-	Increased in mean	Increase in mean F0	Very much lower	Below normal mean
Pitch range	Much wider	Wide range, median normal or higher	Much wider	Increase in range F0	Slightly wider	Slightly narrower
Intensity	Raised	-	Increased	Normal	Lower	Decreased
Rate	High rate	Tempo normal	Increased rate	Increased rate	Very much faster	Slightly slow, long pitch falls
Spectral	High midpoint for average spectrum for nonfricative portions	-	Increase in high frequency energy	Increase in high-frequency energy	-	Downward inflections
Voice Quality	Tense, breathy, heavy chest tone, blaring	Breathy	Tense , breathy, blaring tone	Tense, irregular voicing	Grumble chest tone	Lax, resonant

Tabella 3.1: Valori qualitativi delle caratteristiche legate allo stile vocale, rispetto alle emozioni sottese

Figura 3.5: Modello emozionale di Plutchik

dal canale espressivo adottato (non importa se facciale, gestuale, vocale o sonoro) e quindi più facilmente comunicabili e trasmissibili per mezzo di processi di encoding e di decoding ed universalmente riconoscibili.
Juslin pone anche l'accento sul fatto che i segnali non verbali come quello vocale hanno:

- *carattere continuo*, non discreto
- *probabilistico*, non univoco
- *iconico*, non arbitrario

Proprio per queste caratteristiche si distinguono dai segnali verbali.
Juslin giunge così a precisare le strategie che gli ascoltatori mettono in atto nella formulazione dei loro giudizi.
Questo autore precisa e riassume il codice delle quattro emozioni primarie.
A questo punto, ci si è chiesti quali fossero le unità atomiche delle emozioni, ossia degli archetipi naturali, che, così come con i colori primari, permettesse

di generare, via via, combinazioni piùcomplesse e che soprattutto, fosse possibile da implemetare in un sistema di riconoscimento automatico.

3.5.3 Modello di Darwin

L'idea centrale di organizzazione della prospettiva darwiniana è il concetto che le emozioni sono fenomeni evoluti selezionati con importanti funzioni di sopravvivenza che hanno risolto problemi che abbiamo affrontato come specie. Come tale, dovremmo vedere le stesse emozioni, più o meno, in tutti gli esseri umani.
Inoltre, poichè gli esseri umani condividono un passato evolutivo con altri mammiferi, dovremmo aspettarci di osservare somiglianze nelle emozioni di strettamente correlati specie [1]. Le idee di Darwin sono state enormemente influenti.
La sua eredità nello studio delle emozioni in psicologia e biologia consiste nel suo uso della teoria dell'evoluzione per la selezione naturale come un quadro di riferimento per la comprensione emotiva delle espressioni e, per estensione, delle emozioni stesse, e la sua insistenza che le espressioni emotive devono essere intese in termini di funzioni e, quindi, valore di sopravvivenza [25].

3.5.4 Modello Cognitive Prespective

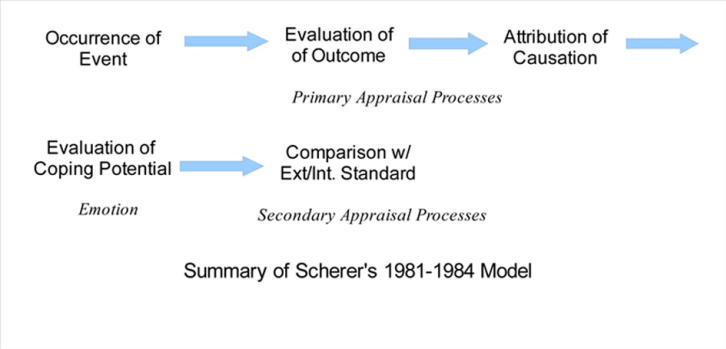


Figura 3.6: Schema Cognitive Perspective [27]

L'assunzione centrale della teoria Congnitive Perspective e la sua tradizione associata di ricerca, è che *pensiero ed emozione siano inseparabili.*

Più nel dettaglio, tutte le emozioni sono viste in questa prospettiva come dipendente su ciò che Magda Arnold [1] ha definito **appraisal**, il processo per cui gli eventi nell'ambiente sono giudicati come positivi o negativi per noi.
Ogni emozione è associata con un *pattern specifico e differente* appraisal.
Questi modelli forniscono il collegamento tra le particolari caratteristiche della persona o organismo, la storia del suo apprendimento, temperamento, personalità, stato fisiologico e particolari caratteristiche della situazione in cui la persona o organismo si ritrova.
Il concetto di appraisal, per molte teorie su modelli cognitivive-oriented moderni in materia di emozioni, vanno di pari passo con l'idea che le emozioni siano *tendenze di azione* [17].
In figura 3.6 si mettono in evidenza gli step successivi rispetto all'evoluzione di uno stato emotivo, partendo dal momento di occorenza di un evento, fino alla definizione dello stato emotivo stesso. Il processo di appraisal informa l'organismo di particolari caratteristiche dell'ambiente e porta uno stato di prontezza ad agire in funzione di quelle caratteristiche.

3.5.5 Modello Social Constracting Perspective

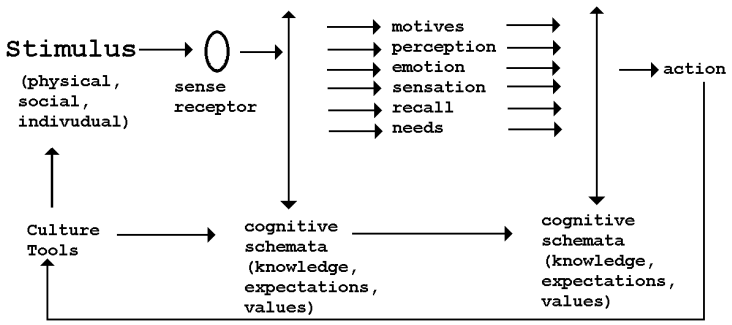


Figura 3.7: Schema Social Constructing Perspective

Rompendo le fila di coloro che vedono le emozioni come un fenomeno biologico, come adattamento evoluto, costruttivisti sociali ritengono che le emozioni sono prodotti culturali che devono il loro senso e coerenza alle regole sociali apprese.
Da quanto emerge in [4], le emozioni non sono solo i resti di un nostro passato

filogenetico, ne possono essere spiegate in termini strettamente fisiologici.
Piuttosto, sono costruzioni sociali, e possono essere pienamente comprese solo a livello di analisi sociali.
Anche i darwiniani più radicali [13], hanno riconosciuto il ruolo della cultura nella regolazione dell'esternazione emozionale, ma Averill e gli altri costruttivisti [4], suggeriscono qualcosa di molto più radicale.
Se si sta indagando sulla natura delle emozioni, basta guardare ciò che le emozioni realizzano socialmente.
Si scoprirà l'esistenza di sistemi di regole culturalmente specifiche che dettano come, quando, e da chi particolari emozioni devono essere vissute ed espresse.
Le differenze di genere e gruppo sociale che si riversano nell'espressione e l'esperienza delle emozioni non sono un caso.
Infatti, essi rivelano il modo in cui le emozioni sono costruite all'interno di una cultura e di come possano servire per particolari finalità sociali.
In 3.7 ci mette in evidenza il processo di creazione di una emozione. Si pone in evidenza il meccanismo di retroazione legato agli aspetti culturali di un individuo, quanto essi siano al tempo stesso, causa ed effetto di una particolare reazione ad uno stimolo.

Ciò che emerge dalla descrizione di tutti questi modelli è che delineare una descrizione prescisa e completa di quali siano tutti i meccanismi legati alla generazione di una emozione risulta difficile.
È evidente, però, il gran numero di fattori implicati, in particolare risulta comune a tutte le teorie, affidare una caratteristica primordiale al concetto di emozione, come risultato, reazione ad uno stimolo, che genererà una azione.
Restano da valutare le inferenze dell'ambiente sociale, della cultura, rispetto ai meccanismi emozionionali e come questi interferiscano nella loro esternazione.

3.6 Fonetica uditiva: la percezione del parlato naturale

In questa sezione ci si sofferma sull'analisi della percezione del linguaggio parlato; poichè se è importante capire l'origine di una emozione, è altrettan-

to importante determinare come questa venga estratta dal segnale vocale, dal nostro cervello, o se sia sufficiente il solo segnale vocale per questo scopo. È bene ricordare che il sistema che si vuole mettere a punto dovrà, sostanzialmente, funzionare da unità di rilevazione del segnale (quindi simula l'orecchio umano) e di elaborazione (il cervello), ossia in grado di distinguere cosa sia o meno funzionale (informazioni utili) a distinguere le emozioni tra loro.

Poichè sono molti i fattori nel sistema di percezione del linguaggio naturale, gli stimoli sintetici prodotti da un sistema di *speech synthesis*, spesso usati negli esperimenti di percezione uditiva, sicuramente non riescono a riprodurre tutte le variabili presenti in condizioni naturali.
Infatti i soggetti possono interpretare gli stimoli sintetici in modo completamente diverso, da come interpreterebbero gli stimoli naturali.
Un esperimento con stimoli del parlato sintetici mostra che un ascoltatore può percepire alcune variazioni linguisticamente, ma non prova che tale processo è rilevante per la percezione del parlato naturale.
Usualmente, comunque, si assume che, se gli stimoli sintetici sono sufficientemente vicini agli stimoli naturali, i risultati percettivi che si ottengono rappresentano un valido modello per la percezione del parlato, che include diversi stadi di analisi [20]:

- uditivo
- fonetico
- fonologico
- lessicale (delle parole)
- sintattico
- semantico

Questi dovrebbero essere visti come processi seriali, in cui il segnale relativo al parlato è trasformato, ad ogni stadio, in una rappresentazione più raffinata, che eventualmente termina nella definizione del messaggio linguistico.
Alcuni stadi, però, devono essere analizzati in parallelo attraverso un meccanismo di *retroazione*, per una correzione di incomprensioni che avvengono a basso livello (per esempio quello fonemico), usare, quindi, una conoscenza

più globale (per esempio della frase o del contesto) per permettere di ritardare la decisione, rispetto al messaggio finale, quando il solo segnale a basso livello non fornisce sufficienti informazioni.
Di fronte ad un'esperienza sonora come quella indotta dalla propria voce o da quella degli altri, una persona passa dal *racconto* all'*esperienza*.
Nell'esperienza di narrazione o di autonarrazione - esterna o interiore, rispettivamente - spicca lo stretto legame tra la dimensione emotiva e la dimensione prosodico espressiva come canale di attribuzione e di rilevamento del senso [20].
Questo rappresenta il campo di esplorazione di questo lavoro, e risulta, quindi, cruciale riuscire a far leva sulle sole componenti acustiche, come variabili discriminatorie, ben sapendo che in un contesto di quotidianeità, nelle comunicazioni interpersonali, è applicabile la *Regola di Mehrabian*:

$$Total\ Liking = 7\%\ Verbal\ Liking + 38\%\ Vocal\ Liking + 55\%\ Facial\ Liking$$

che mette in evidenza quanto sia determinante la componente facciale rispetto al potere informativo e discriminatorio di uno stato emozionale.

Si ricordano inoltre gli interessanti esperimenti di Abelin e Allwood [3] che hanno eseguito indagini statistiche su un gruppo di americani: questi hanno dovuto riconoscere l'emozione di un altro americano o di una persona giapponese usando soltanto le informazioni acustiche (le espressioni e le altre componenti corporee non erano osservabili).
Agli ascoltatori giapponesi è stato chiesto di decidere quali emozioni gli altri giapponesi o americani stavano cercando di trasmettere.
Sono stati raggiunti due risultati importanti: la percentuale di riconoscimento era elevata e c'era poca differenza fra le prestazioni di diversi soggetti nella rilevazione delle emozioni trasmesse da qualcuno che parla la stessa lingua e questo è normale per il giapponese ascoltato dai giapponesi così come per l'americano ascoltato dai soggetti americani; nel caso contrario (giapponesi che ascoltano l'americano e viceversa) si era ben lontani dal riconoscimento assoluto: il riconoscimento migliore è stato del 60 %.
Questo risultato potrebbe essere spiegato parzialmente dal fatto che gli oggetti ascoltati erano abbastanza artificiali (*frasi senza senso*, nonsense utterances), come è confermato dagli studi per la differenziazione fra le frasi in stile neutro e le frasi espressive.
Il primo risultato indica che l'obiettivo di costruire un sistema che può avere

competenze di riconoscimento con l'esattezza di un orecchio umano, è raggiungibile solo nella teoria.
Il secondo risultato indica che non si dovrebbe prevedere un riconoscimento perfetto e quindi confrontare le prestazioni della macchina con le prestazioni umane.
Il fatto che gli esseri umani non hanno dimostrato ottime prestazioni di riconoscimento dimostra una forte similarità delle componenti fisiologiche e acustiche.
Nelle situazioni reali, risolviamo tutte queste ambiguità usando il contesto, l'espressività, la mimica e altre modalità.
Effettivamente, alcuni esperimenti hanno indicato che la natura multi-modale dell'espressione può condurre a quello che viene chiamato *effetto di McGurk*: per le emozioni, una faccia che mostra l'emozione A e che parla con l'emozione B, è percepita come esprimere soltanto una delle due emozioni o, a volte, persino una terza non rappresentata.
Inoltre, contesti differenti possono condurre la gente a interpretare la stessa intonazione anche quando si esprimono emozioni differenti.
Questi risultati indicano che non chiederemo al nostro sistema di distinguere tutte le espressioni e le emozioni possibili ma soltanto quelle fondamentali che interessano le categorie studiate e maggiormente distintive.

Alla luce di queste considerazioni, si è proceduto con la scelta di un set di emozioni, che andranno a rappresentare le classi in cui far ricadere i file audio analizzati; in particolare esse saranno:

- Gioia
- Rabbia
- Tristezza
- Paura

Analizzando le caratteristiche dello speech, esso è costituito da due tipologie di informazioni:

Informazioni semantiche strettamente legata alle modalità di produzione di una espressione regolata dalla grammatica della lingua in questione, così come dalle sue regole fonetiche.

Informazioni paralinguistiche legate, appunto allo stato emozionale dell'agente.

Si è quindi, a questo punto, proceduto nel riassumere le carattiristiche su cui basare l'analisi dei file audio ai fini della classificazione degli stessi, come si vedrà nel capitolo successivo.

Capitolo 4

Progetto logico della soluzione del problema

4.1 Introduzione

In questa parte del documento si andranno a descrivere i componenti e le loro interazioni, dell'applicazione messa a punto.
Il modello fa rifereminto al più classico dei sistemi black box che vedono un segnale di ingresso, un blocco di elaborazione dei dati e un segnale di uscita. Infatti un obiettivo è stato quello di garantire una certa trasparenza, rispetto all'elaborazione dei dati, nei confronti dell'utente, garantendo, così, l'usabilità dello strumento. In figura 4.1 è rappresentato lo schema generale del sistema di classificazione messo a punto, volendo mettere in evidenza i macro componenti coinvolti.
Andando a descrivere il componente di elaborazione dei file audio,è possibile suddividerlo in due moduli, rappresentati dal *modulo di analisi testuale* e quello di *analisi dell'audio*.
Ciò che questi moduli andranno ad estrarre sono informazioni, dette feature [2].
Avendo ben chiaro l'obiettivo di dover discriminare tra le quattro classi di emozioni, è necessario capire quali siano le caratteristiche racchiuse all'interno del segnale vocale, così come quello testuale, che inequivocabilemente possano permettere di distinguere tra le diverse emozioni, facendo leva sulle diverse informazioni che posso trasportare.
Dato che durante la fase di progettazione si sono prese in considerazione

diverse soluzioni, le configurazioni del sistema si sono evolute in base alle caratteristiche dello stesso e ai risultati che man mano si ottenevano.

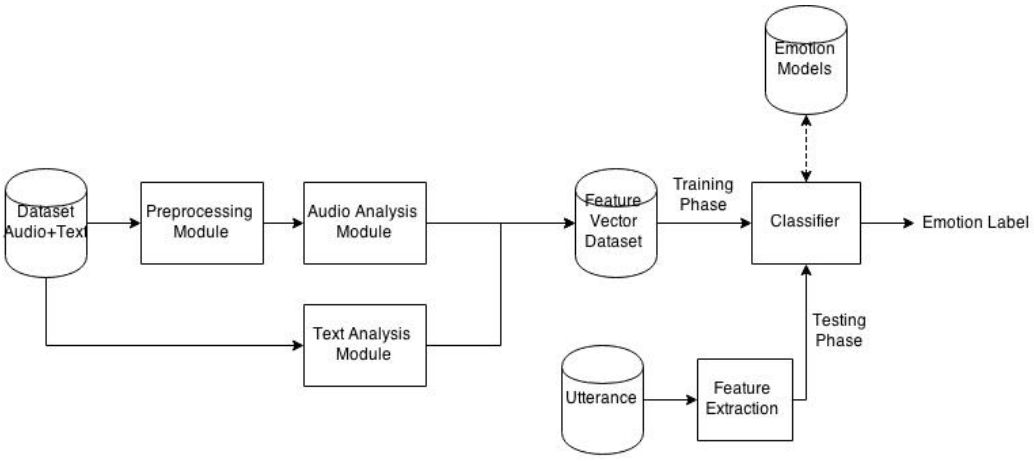


Figura 4.1: Schema generale di un sistema supervised di classificazione di dati

Dallo schema in figura 4.1 si evince come, una volta raccolti i dati, essi confluiranno in un classificatore che andrà, dapprima a sintetizzare un modello per ciascuna emozione e, in seconda battuta, ad etichettare le espressioni di test sottoposte al sistema, per la sua validazione.

4.2 Dataset

Durante la fase di ricerca, una volta chiara la natura del sistema da mettere a punto, si è dovuto tener conto dei dati a disposizione, delle loro caratteristiche, della loro organizzazzione.
Nella fattispecie l'insieme di dati è così costituito:

- 720 file audio: ciascun file audio rappresenta una singola espressione. Con questo si vuole intendere una o più frasi che, insieme, esprimono un pensiero che sottende un unico stato emotivo, inequivocabilmente distinguibile. Questa inequivocabilità è garantita dall'analisi manuale dei file audio, da parte di operatori.
- Ciascuna espressione è *recitata* da *6 attori maschi diversi*

- i file sono divisi in quattro gruppi, ciscuno corrispondente alle quattro emozioni -*Gioia, Paura, Tristezza, Rabbia*-.
    - Per ciascun gruppo, ci sono *dieci espressioni*, e ciascuna espressione è recitata con *tre interpretazioni diverse*, da ciascun attore per un *totale di 180 espressioni*, per classe di emozione [20].

I file audio sono stati registrati in uno studio professionale, ad una frequenza di campionamento di 44100 Hz e quantizzazione a 16 bit [20].

Dalla descrizione del dataset a disposizione, risultano chiare le limitazioni che il sistmea possiede intrinsecamente - si pensi al fatto che il sistema sarà modellato sulle sole voci maschili, all'interno di un certo intervallo di età-. Questo però non deve essere considerato un vero problema.
L'integrazione di altre tipologie di segnali audio, corrispondenti a parlatori di tipologie quanto più possibile diversificate è realizzabile, così come descritto nel Capitolo 2, con sistemi di *speaker recognition* e, in generale, con meccanismi di distinzione tra i parlatori sono già ad un stato di sviluppo avanzato, permettendo quindi di essere facilmente integrati in sistemi del nostro tipo.

4.3 Modulo di analisi dell'audio

In questa parte del documento, verranno illustrate le tecniche adottate per l'estrazione di informazioni di interesse dai file audio.
In particolare, saranno descritti i metodi, gli algoritmi, motivando la scelta degli stessi.

4.3.1 Pre-processing Audio

Una fase sicuramente determinate è stata rappresentata dal processing preliminare dei file audio.
In particolare, ciò a cui si è puntato è stato quello di rendere quanto più possibile, tutte le fasi successive, indipendenti (immuni), dalle caratteristiche preliminari del file audio stesso.
Banalmente, partendo dalle tecniche di registrazione, che ne implicano le

condizioni legate alla qualità audio, in particolare :

Ambiente di registrazione : rappresenta sicuramente un fattore determinante ai fini di una omogeneità e consistenza dei dati raccolti.
Infatti, le caratteristiche dell'ambiente di registrazione potranno influenzare i dati raccolti e quindi le informazioni estratte. è bene ricordare che tutto ciò che non fa riferimento al segnale di interesse, rappresenta un disturbo, che può rivelarsi determinante nelle prestazioni globali del sistema.

Tecniche di ripresa : rappresentate dai dispositivi implicati nella registrazioni, con relative tecniche di utilizzo, che implicano la scelta di parametri cruciali, ai fini di una corretta collezione di sample per costruire il dataset, quali, frequenza di campionamento, cos'ıcome la lunghezza della parola di quantizzazione in bit.

In figura 4.2 è riportata una descrizione del modulo di preprocessing di ciascun file audio, ponendo in evidenza che i file di ingresso agli stadi successivi sono suddivisi in due tipologie, rispettivamente, in accordo con la tipologia di feature da estrarre.

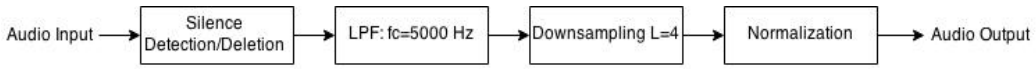


Figura 4.2: Preprocessing Module

Silence Detection/Delation

Tale componente ha il compito di eliminare i silenzi presenti all'inizio e alla fine di ciascun file audio, oltre che a silenzi, all'interno del file audio, che non siano pause.
Per avere un riferimento rispetto alla durata media di una pausa in un eloquio, si è fatto riferimento a [15], dove si nota come il valor medio di una pausa sia di circa $170\,ms$. Per la sua realizzazione si è considerato il segnale audio in ingresso, suddiviso in frame, se una quota successiva di frame, pari alla durata di una pausa -valore sopracitato- è al di sotto di una soglia di intensità imposta, allora sarà eliminata dal contenuto del file. L'output sarà costituito da un segnale audio, risultato della concatenazione dei singoli *chunk* estratti.

Low Pass Filter & Downsampling

Si è partiti dal presupposto, ragionevole, che i segnali analizzati fossero prettamente vocali; questo ha permesso di riddurre la frequenza di campionamento a $f_c = 11025$ (un quarto di quella di partenza), considerando che il segnale vocale ritrova il 90% della sua energia spettrale all'interno di una banda di circa 5kHz, come mostrato in figura 4.3, quindi, siamo più che all'interno dei limiti del teorema di campinamento.

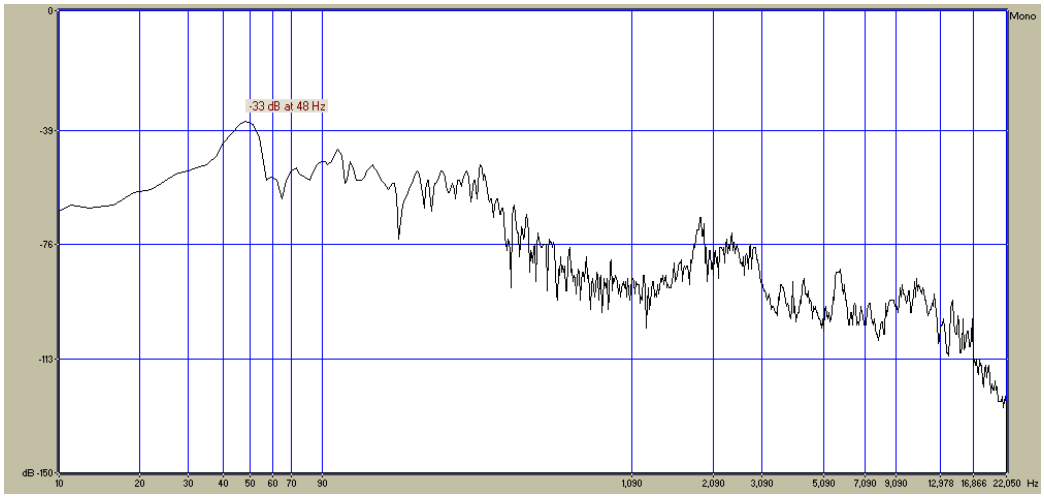


Figura 4.3: Esempio di spettro di un segnale vocale

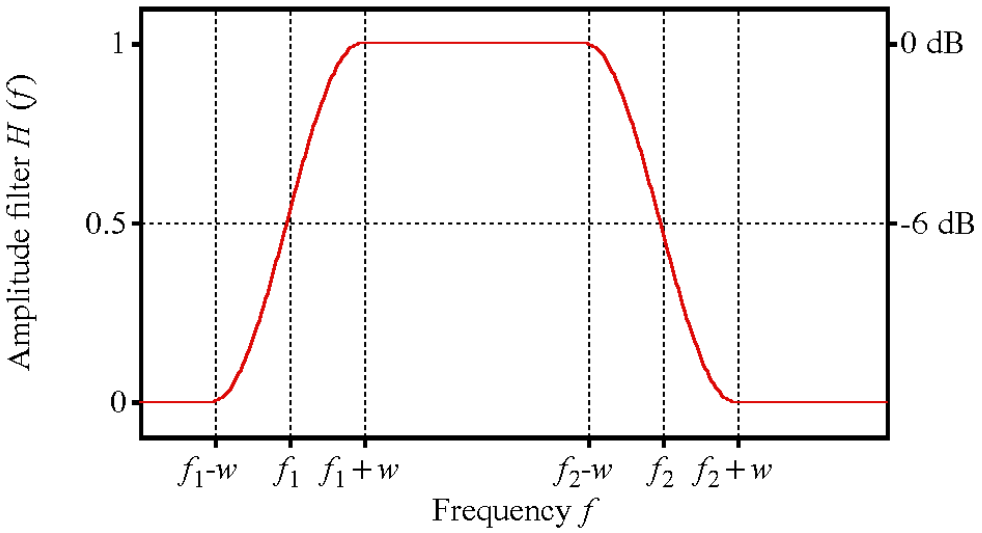


Figura 4.4: Esempio di filtro passabanda, con parametri $f_{1,2}$, ω settabili

Il filtro realizzato è un filtro passabasso di *Hann-like* - la forma in frequenza, è riconducibile ad una finestra di Hann-, con frequenza di taglio di a 5KHz e banda di transizione 200 Hz, così come descritto in figura 4.4.

Normalization

La normalizzazione è una operazione che ha il fine di far sì che l'analisi non sia interferita da errori di *offset* tra i campioni del dataset, ossia la differenza di intesità tra i sample potrebbe compromettere il confronto di alcune feature, dato che i valori di tali caratteristiche risuternno diversi, ma perchè diverso è il punto di partenza.
Per questo motivo, la normalizzazione dei file audio ha come riferimento la *pressione acustica* esercitata dal suono. Il valore di riferimento è quello legato alla *soglia minima di udibilità*: $p_0 = 20\mu Pa$. Si riporta il valore di intensità a 80 $dB\ SPL$.
La scelta di tale valore è il risultato di prove ripetute, con valori diversi, che però rispettassero lo standard di udibilità che si aggira intorno ai 70 $dB\ SPL$ [7].

In sostanza il nuovo segnale dopo la normalizzazione sarà:

$$\sqrt{1/(t_2 - t_1) \int_{t_1}^{t_2} x_2(t)dt} \tag{4.1}$$

Pause Detector

In questa fase i file audio vengono ridotti alle sole parti vocali, ossia tutte le pause nell'eloquio vengono eliminate, questo al fine della estrazione delle feature acustiche e i coefficienti MFCC, che vedono le loro informazioni legate alle sole parti vocali.
Così come evidenziato in [28] questo tipo di operazione, aumenta l'accuratezza nella stima di alcune variabili acustiche, quali le formanti, che, altrimenti, subirebbe una degradazione dovuta alla presenza di rumore, intrinseco all'interno delle pause. Per l'individuazione di queste ultime, è stato messo a punto un componente che facesse leva sulla divisione in frame del segnale in ingresso, e per ciascun frame, determinasse se il contenuto di tale frame fosse o meno vocale.
Dato che un segnale vocale è contraddistinto da una componente armonica - nell'articoazione delle vocali -, sarà possibile, con l'implementazione di un algoritmo di *pitch tracking*, individuare tali componenti. In figura 4.5 sono mostrati i blocchi funzionali del Pause Detector.
A questo punto, si è diviso il segnale in frame di 40 ms e applicato l'algoritmo di pitch tracking (vedi dopo).
A questo punto avremo, in output, i file audio delle espressioni, privati di pause e silenzi.
Così costituiti, essi saranno l'input del modulo di analisi delle variabili acustiche continue, e delle caratteristiche proprie del parlatore (MFCC coefficients).

Figura 4.5: Pause Dectector

è da sottolineare come comunque si sia puntato a considerare le pause come fattori determinanti, in primis, per la definizione di uno stile di espressione, che potesse essere associato alla emozione sottesa com edescritto nel

4.3.2 Audio Feature Extraction

In questa fase, le registrazioni sono state sottoposte al modulo di analisi dell'audio.
Una presmessa sulla scelta delle feature è necessaria: in particolare, è stato argomento di indagine, capire quali potessero essere, nel panorama delle caratteristiche estraibili da un file audio, quelle che potessero in qualche modo essere usate per disambiguare una emozione ripetto all'altra, definendole, nel concreto, in maniera univoca, attraverso un modello.

Frame-based Analysis

Si è distinto tra feature per le quali avesse più senso studiarne il comportamento evolutivo rispetto al tempo, ossia capire quanto la rapidità di variazione di determinati parametri potesse essere decisiva nella classificazione dei sample di registrazione, proprio per voler indagare, non solo sulle caratteristiche in via assoluta, ma anche indagare sui pattern evolutivi di tali parametri, proprio come descritto nel Capitolo 2, rispetto ai modelli teorici delle classi di emozioni.
Per questo fine, è necessaria l'applicazione di una analisi *frame-by-frame*, per poter determinare, con metodi alle differenze finite, le derivate di ordini successivi dei parametri di interesse.
In particolare, per le Speaker Model Feature e per le Continuous Acoustic Feature si è proceduto come segue [37]:

Windowing Pitch Analysis : Finestratura di *Hanning* di lunghezza 40 ms, 0.75 di overlap.

$$w(n) = 0.5 + 0.5 \cos\left(\frac{2\pi n}{N-1}\right)$$

Windowing MFCC Analysis : Finestratura di *Hamming* di lunghezza 25 ms, 0.5 di overlap.

$$w(n) = 0.54 + 0.46 \cos\left(\frac{2\pi n}{N-1}\right)$$

Partendo dalle considerazioni espresse in [11] [19], si è proceduto con questa configurazione poichè ha mostrato risultati migliori a seguito di una serie di test con:

- Finestra di Hanning di 20 ms con 50% di overlap
- Finestra di Hanning di 40 ms con 50% di overlap

Quindi i valori ottenuti, per ciascuna feature, per ciascun frame, *sono stati riassunti con un valore di deviazione standard e media*, rispetto al numero di frame nell'espressione analizzata.

Questa ultima fase è necessaria poichè è chiaro che, espressioni di lunghezza differente, avranno un numero di frame differenti e questo porterebbe ad avere feature vectors di lunghezza diversa, che avrebbe compromesso l'operazione successiva di classificazione. Per superare questa problematica, si è scelto, come si vedrà nelle sezioni successive, di uniformare la lunghezza dei feature vector, da un lato, utilizzando parametri statistici come media e deviazione standard, dall'altro, modificando l'approccio del classificatore, considerando come unità atomiche da classificare, i feature vector di ogni singolo frame, indipendentemente dalla espressione di appartenenza, purchè, naturalemente etichettati.

Al termine della fase di ricerca [6][11][33], si è giunti ad un *feature set* così costituito:

- **Speaker Model Feature**:
  12 MFCCoefficients
  12 delta MFCCoefficients
  12 delta-delta MFCCoefficients
  Per un totale di 72 features

- **Continuous Acoustic Feature**:
  Pitch, derivata Pitch

Intensità, derivata Intensità
Formanti I/II/III, con rispettive derivate
per un totale di 20 feature

- **Voice Quality Feature:**
  Jitter
  Shimmer
  Percentuale di Voice Breaks
  Harmonic Noise Ratio
  Quantile 25/50/75
  per un totale di 12 feature.

Pitch-Formant I/II/III-HNR:
il pitch rappresenta sicuramente una delle feature chiave al fine di disambiguare tra le varie classi di appartenenza delle espressioni analizzate.
L'algoritmo utilizizzato è quello implementato in PRAAT.
La scelta è stata dettata innanzitutto dalle caratteristiche di robustezza e precisione del metodo adottato, così come si può notare da quanto segue:[7]

> Our measurements of the places and the heights of the peaks in the lag domain are several orders of magnitude more accurate than those of the usual pitch-detection algorithms.

Così come specificato in precedenza il metodo appartiene a quella categoria detta frame based analysis, al fine di garantire risultati consistenti con l'algoritmo implementato, si è scelta una configurazione per la divisione in frame del segnale di ingresso, portandola a 40 ms di lunghezza, e su questi frame così ottenuti si applica una *short term analysis*, con finestre di 10 ms per la stima accurata del pitch.
Questo dipende dalla scelta del *Minimum Pitch*, ossia del valore minimo ammissibile come periodicità del segnale, che, nella fattispecie è stato settato a 75 Hz.
Questo poichè l'algoritmo necessita di almento tre periodi per poter stimare la periodicità del segnale con accuratezza, di qui la scelta di una lunghezza dei frame di 40 ms [7]; discorso analogo vale per l'HNR, che però necessita di sei periodi per essere stimato.

Il metodo prevede quindi l'estrazione del pitch attraverso l'analisi dell'autocorrelazione del segnale finestrato :

$$r(\tau) = \int a(t) \cdot a(t+\tau)dt \tag{4.2}$$

con $a(t) = (x(t_{mid} - 1/2T + t)\mu_x) \cdot w(t)$
dove t_{mid} è l'istante su cui è centrato x(t) μ_xè il valore medio del segnale x(t), $w(t)$ è invece la finestra scelta.
A questo proposito in [7] è possibile verificare, come la finestra di Hanning sia la più indicata ai fini dell'estrazione del pitch.
Di qui:

$$w(t) = 1/2 - 1/2(cos(2\pi t))/T$$

L'idea è quella di individuare i valori candidati della frequenza di pitch, attraverso una massimizzazione della funzione di autocorrelazione normalizzata rispetto all'autocorrelazione della finestra scelta:

$$(r_x(\tau) \approx (r_a(\tau))/(r_w(\tau))$$

Il calcolo della funzione di autocorrelazione passa attraverso la trasformazione del segnale in ingresso nel dominio di Fourier attraverso una FFT, per poi, naturalmente, tornare nel dominio del tempo, attraverso la trasformata inversa.
Su questa tecnica è basato il sistema di tracking del pitch, quindi della Formante I/II/III.
Infine definiamo un segnale:

$$(x(t) = h(t) + n(t)$$

dove $h(t) = h(t + T_0)$ è un segnale periodico $n(t)$ è rumore bianco.
Partendo da questo possiamo affermare che $h(t), n(t)$ siano incorrelate, quindi la correlazione massima la si ritrova in $t = 0$, quindi:

$$r_x(0) = r_H(0) + r_N(0) \tag{4.3}$$

$$r'(\tau_{max}) = \frac{r_H(0)}{r_x(0)} \tag{4.4}$$

$$1 - r'(\tau_{max}) = \frac{r_N(0)}{r_x(0)} \tag{4.5}$$

$$\tag{4.6}$$

dove

$$r'(\tau) = \frac{r(\tau)}{r(0)}$$

τ_{max} , sarà il lag che permette alla funzione di essere massima, nella fattispecie, deve essere $\tau_{max} = T_0$ Infine, quindi, possiamo definire l'Harmonic to Noise Ratio come:

$$HNR(dB) = 10 \cdot log_{10} \frac{r'(\tau_{max})}{1 - r'(\tau_{max})} \tag{4.7}$$

Intensità:
Così come evidenziato in [7]: I valori di ampiezza vengono dapprima elevati al quadrato, poi convoluti con una finestra di analisi Gaussiana (Kaiser-20; lobi secondari -190 dB).
La durata effettiva di questa finestra di analisi è $3.2/(minimum\_pitch)$, che garantisca che un segnale periodico viene analizzato come avente un ripple pitch-sincrono intensità non superiore alla nostra word di 4 byte floating-point precision (cioè $< 0,00001dB$).
Con minimum_pitch si intende la durata minima di un periodo, di un segnale, appunto, periodico, affinchè i segnali a con periodicità più brevi (pitch più alto), subiranno un deterioramento minore da parte della finestratura.

Parallelamente, i file audio pre processati, comprensivi delle pause, sono stati sottoposti all'analisi di caratteristiche globali, che riguardassero informazioni relative al comportamento del segnale sia nel tempo, che in frequenza.

Mean Period: è la stima del periodo mediata sul numero di frame, all'interno di una espressione, da cui si potrà stimare la frequenza fondamentale.

Standard Deviation of Period:
deviazione standard della durata del periodo, rispetto alla sua media calcolata su tutta l'espressione, divisa in frame.

Jitter (absolute):
Rappresenta la media tra le differenze assolute tra le frequenze individuate

tra periodi successivi, in particolare:

$$Jitter(absolute) = \frac{1}{N-1}\sum_{i=1}^{N-1}|T_i - T_{i+1}| \tag{4.8}$$

Shimmer (dB):
Rappresenta la media tra le differenze assolute tra le ampiezze individuate tra periodi successivi, in particolare:

$$Shimmer(dB) = \frac{1}{N-1}\sum_{i=1}^{N-1}|log_{10}\frac{A_{i+1}}{A_i}| \tag{4.9}$$

Quantile 25/ 50/ 75:
il quantile rappresenta, in statistica, una figura di merito rispetto alla distribuzione di una variabile aleatoria.
In particolare dato un valore $\alpha \epsilon [0,1], \alpha \epsilon \mathbf{R}$, il quantile è un valore q_α tale che la popolazione sia divisa in due parti proporzionali ad $\alpha e (1-\alpha)$.
Avendo a disposizione delle modalità ordinate, è stato possibile calcolare il 25-esimo quantile,così come la mediana e il 75-esimo quantile, che rispettivamente rappresentano i valori della frequenza di pitch, raggiunto il 25%, 50% e 75% dei valori di pitch estratti da ciascun frame.

Fraction of locally Unvoiced Frame:
rappresenta il rapporto de il numero di unvoiced frame, rispetto ai voiced frames.
Questo permette di avere informazioni rispetto all'entità delle pause in una espressione, in termini quantitativi.

Degree of Voice Brakes:
rappresenta il rapporto tra lunghezza totale del segmento analizzato e la somma delle durate degli unvoiced frame.
Infine i silenzi all'inizio e alla fine dell'espressione non sono considerati in tale calcolo.

HNR Harmonic to Noise Ratio (e NHR):
è una grandezza, misurata in dB, che rappresenta il rapporto tra l'energia del segnale sottesa dalla parte periodica del segnale (voiced), rispetto a quella non periodica (unvoiced).

$$10 \cdot log_{10} \frac{Periodic\ signal\ energy}{Noise\ signal\ energy} \tag{4.10}$$

Mel-Frequency Cepstral Coefficients MFCC:
questa feature è sicuramente tra le candidate ad essere determinanti per un corretto funzionamento del sistema, infatti esistono in letteratura svariate ricerche che vedono coinvolti questi coefficienti, che rappresentano la distribuzione dell'energia all'interno del spettro, soprattutto poichè esse sono strettamente legate al parlatore e modellano meglio, come l'orecchio umano ascolta e quindi rendono il sistema robusto rispetto al parlatore, migliorandone le prestazioni globali.

Un segnale audio è in continua evoluzione, ma per semplificare, senza che questo leda le prestazioni del sistema, ammettiamo che per un breve tempo il segnale audio non cambi molto (con non cambia, si intende, statisticamente, cioè statisticamente stazionario, ovviamente i campioni sono in continua evoluzione anche su scale temporali brevi).
Questo è il motivo per cui inquadriamo il segnale in frame di 25ms .
Se il frame è troppo breve, non avremo sufficienti campioni per ottenere una stima spettrale attendibile, se troppo, il segnale cambierà troppo in tutta la sua struttura.
Il passo successivo consiste nel calcolare la potenza spettrale di ogni frame.
Questo è motivato dalla coclea umana (un organo nell'orecchio) che vibra in punti diversi a seconda della frequenza dei suoni in ingresso.
A seconda della posizione nella coclea che vibra, il cervello riceverà informazioni su quali intervalli di frequenze sono presenti, come mostrato in figura 4.6.
La stima del periodogramma svolge un lavoro simile, identificando quali intervalli di frequenze sono presenti nel frame.

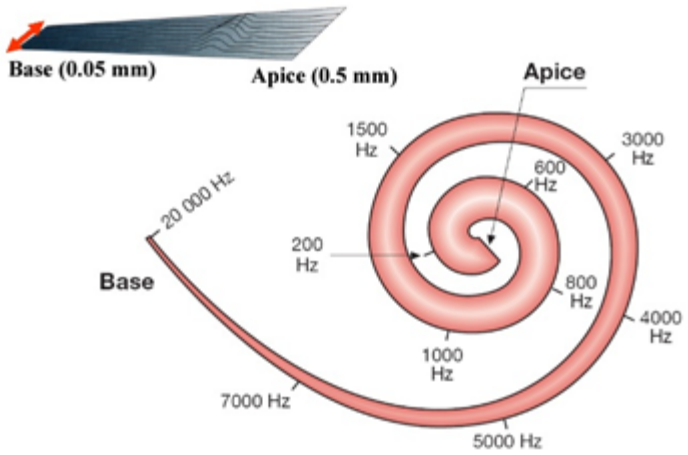


Figura 4.6: Illustrazione esemplificativa della risoluzione spettrale di una coclea, responsabile dell'estrazione del contenuto in frequenza del segnale in ingresso l'orecchio

La stima spettrale contiene ancora molte informazioni non richieste per l'analisi nella speech signal analysis.
In particolare, la coclea non può distinguere la differenza tra due frequenze

ravvicinate.
Questo effetto diventa più pronunciato, all'aumentare della frequenza.
Viene eseguita quindi dal nostro filtro a Mel: il primo filtro è a banda molto stretta e fornisce un'indicazione di quanta energia esiste vicino agli 0 Hertz. Poichè le frequenze diventano più elevate, i filtri diventano più ampi.
La scala Mel ci dice esattamente la distanza esistente tra i filtri del banco e la loro larghezza.
Una volta che abbiamo le energie dal banco di filtri, ne calcoliamo il logaritmo. Questo è motivato anche dal funzionamento dell'udito umano: infatti non percepiamo l'intensità sonora su una scala lineare.
Generalmente al raddoppio dell'intensità percepita di un suono, corrisponde un relativo aumento di un fattore 8 di energia in esso. Ciò significa che a grandi variazioni di energia possono non corrispondere grandi differenze in quello che si percepisce.
Questa operazione di compressione rende tali coefficienti più vicini all'informazione, che l'essere umano estrae naturalmente.
Il logaritmo ci permette di utilizzare la sottrazione del valor medio, che è una tecnica normalizzazione di canale.
Il passo finale è quello di calcolare la DCT delle energie in uscita dal banco filtro .
Ci sono due ragioni principali per cui questa operazione viene eseguita: la prima, poichè le bande dei singoli filtri sono sovrapposte, quindi esiste una certa correlazione tra loro e la DCT decorrela tali energie.
Ma si noti che solo 12 dei 26 coefficienti DCT sono conservati.
Questo perchè i coefficienti DCT di ordine via via superiore rappresentano cambiamenti veloci nelle energie del banco di filtri e questi abbasserebbero il livello di accuratezza nei sistemi in cui sono inseriti.
Vediamo come tali coefficienti vengono calcolati:
Il segnale in ingresso viene dapprima diviso in frame, di 25 ms (256 sample con Fs = 11025 Hz) con overlap del 50%, mediante l'utilizzo di una finestra di Hamming.
Per ogni frame, viene calcolata la FFT, quindi ciascun frame sarà l'input di un banco di filtri, con suddivisione in bande di Mel [32]:

$$M(f) = 1125ln(1 + f/700)$$

In 4.7 si mostra come il banco di filtri agisca sullo spettro del segnale in ingresso, rappresentanto dai singoli frame di cui sopra.

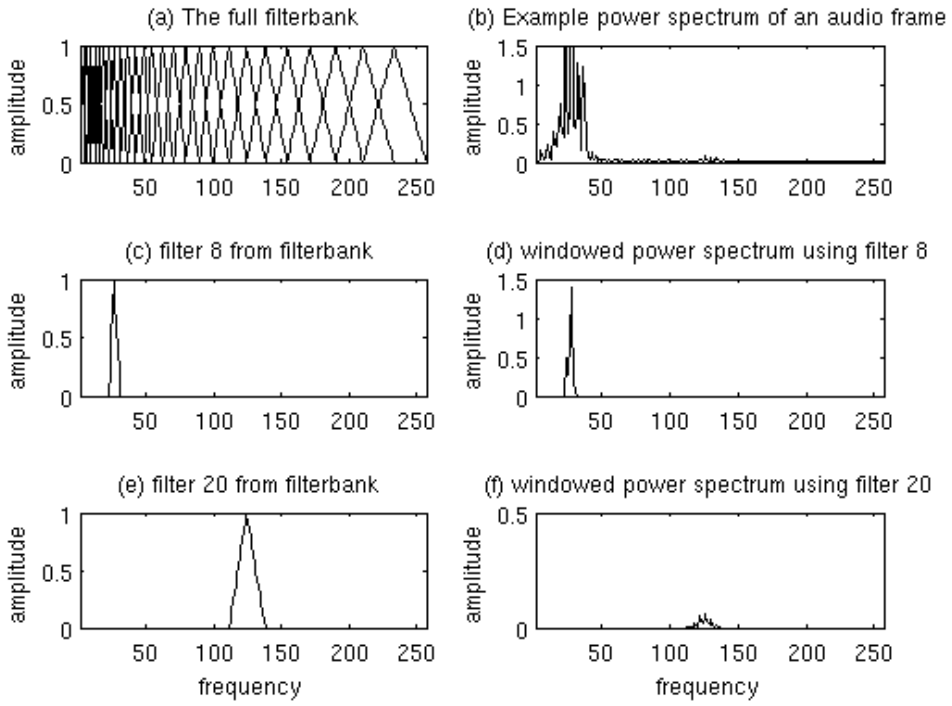


Figura 4.7: Divisione del segnale in in gresso attraverso un banco di filtri con divisione in bande di Mel

Quindi verrà applicata una trasformata DCT (Discrete Cosine Transform) del logaritmo dell'energia per banda, fornita dal banco di filtri, per ciscun frame.
Infine ciascun rispettivo coefficiente della DCT, facente riferimento ad ogni singolo frame, sarà mediato rispetto al numero di frame stesso e rappresenterà il contenuto energetico in quella particolare banda di Mel.
In figura 4.8 sono rappresentati i blocchi funzionali necessari all'estrazione dei coefficienti MFCC, con la relativa rappresentazione spettrale. Ciò che emerge è che a seguito della DCT, l'energia del segnale in ingresso è effettivamente ben rappresentata dai coefficienti MFCC.

Il passo ulteriore ha visto il calcolo dei coefficienti delta-MFCC e delta-delta-MFCC, che rappresentano la derivata prima e seconda, rispettivamente, dei coefficienti MFCC.
L'inclusione di tali figure di merito è stata dettata dai risultati che i sistemi di Speaker Recognition hanno mostrato successivamente alla loro implementazione[23].

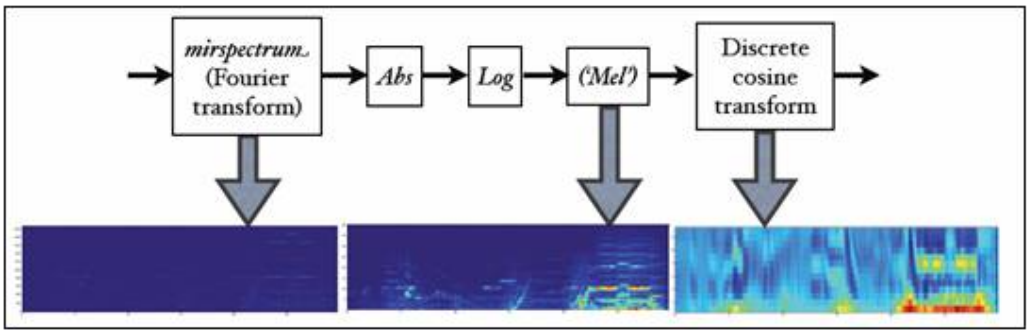


Figura 4.8: Schema a blocchi dell'estrazione dei coefficienti MFCC da un segnale vocale

4.4 Modulo Analisi Testuale

4.4.1 Introduzione

Il modulo di analisi testuale, risponde a quel ben noto paradigma della biometria, che vede nella combinazione di analisi degli stessi dati da più punti di vista, un punto di forza, in termini di robustezza e precisione del sistema messo a punto.
Nel caso in analisi, si è puntato ad associare a ciascun file audio, uno testuale, che rappresenta la sua trascrizione.
Partendo da questo, si è indagato su quali potessero essere i legami tra lo stato emozionale del parlatore, rispetto alla struttura semantica e grammaticale scelta dallo stesso.
In figura 4.9 sono rappresentati i blocchi funzionali necessari a sintetizzare le informazioni estratte dal testo, in feature vectors.

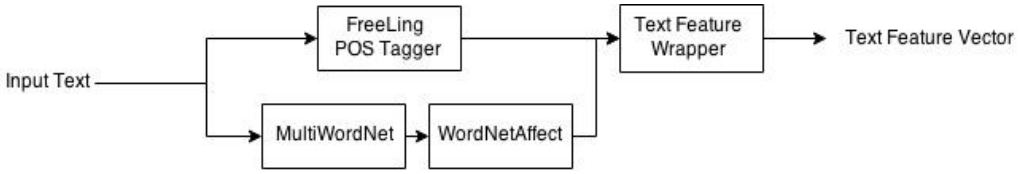


Figura 4.9: Modello del modulo di analisi testuale

4.4.2 Database lessicale: Freeling, Wordnet, MultiWordNet e WordNetAffect

Freeling

Freeling, rappresenta il primo passo verso l'analisi del contenuto emozionale all'interno della trascrizione relativa alle registrazioni. In particolare, si tratta di una libreria di funzioni, atte all'analisi del testo [24].
Per il nostro fine, è stato utilizzato come:

Tokenizer : avendo un testo in ingresso, analizza i singoli vocaboli, e ne fornisce il *lemma*- la citazione di una parola, ossia quella parola che per convenzione Ã¨ scelta per rappresentare tutte le forme di una flessione; per i verbi è rappresentata dal modo infinito, per gli aggettivi il maschile singolare ...

PoS Tagger : Rappresenta la funzione di etichettatura di cascun lemma, rispetto al ruolo grammaticale all'interno della frase. PoS, infatti, è proprio la *Position of Speech*. Questo è funzionale all'estrazione di informazioni successive, quali la collocazione all'interno del corpus WordNet (vedi dopo), del vocabolo in analisi.

Princeton WordNet

WordNet è una grande banca dati lessicale della lingua inglese.
Nomi, verbi, aggettivi e avverbi sono raggruppati in insiemi di *sinonimi cognitivi (synsets)*, ognuno esprimendo un concetto distinto.
I synsets sono collegati per mezzo di relazioni concettuali-semantiche e lessicali.
La rete risultante di parole e concetti legati da criteri, è navigabile attraverso il browser messo a punto dallo stesso team.
La relazione principale tra parole WordNet è la *sinonimia*, ad esempio come tra le parole *veicolo* e *automobile*.
Ciascuno dei 117000 synsets di WordNet è collegato ad altri synset mediante un piccolo numero di "relazioni concettuali".
Inoltre, un synset contiene una breve definizione ("gloss") e, in molti casi, uno o più brevi frasi che illustrano l'uso dei membri dei synset.
A morfologie di parole identiche, ma con più significati distinti, sono rappresentate in altrettanti synset distinti.
Così, ogni coppia *morfologia-significato* in WordNet è unica.
In figura 4.10 è rappresentato un esempio di relazioni tra synsets [21].

La relazione codificata più frequentemente tra synset è detta super-subordinata (chiamata anche *iperonimia, iponimia o relazione ISA*).
Così, WordNet afferma che la categoria Mobili comprende Letto, che a sua volta comprende Letto A Castello; al contrario, concetti come Letto A Castello costituiscono la categoria Mobili.
Tutte le gerarchie tra sostantivi riportano al nodo principale.
L'iponimia è transitiva: se una poltrona è un tipo di sedia, e se una sedia è una sorta di mobile, allora una poltrona è un tipo di mobile.
WordNet distingue tra i tipi (sostantivi comuni) e istanze (persone specifiche, i paesi e le entità geografiche).
Così, poltrona è un tipo di sedia, Barack Obama è un esempio di un presi-

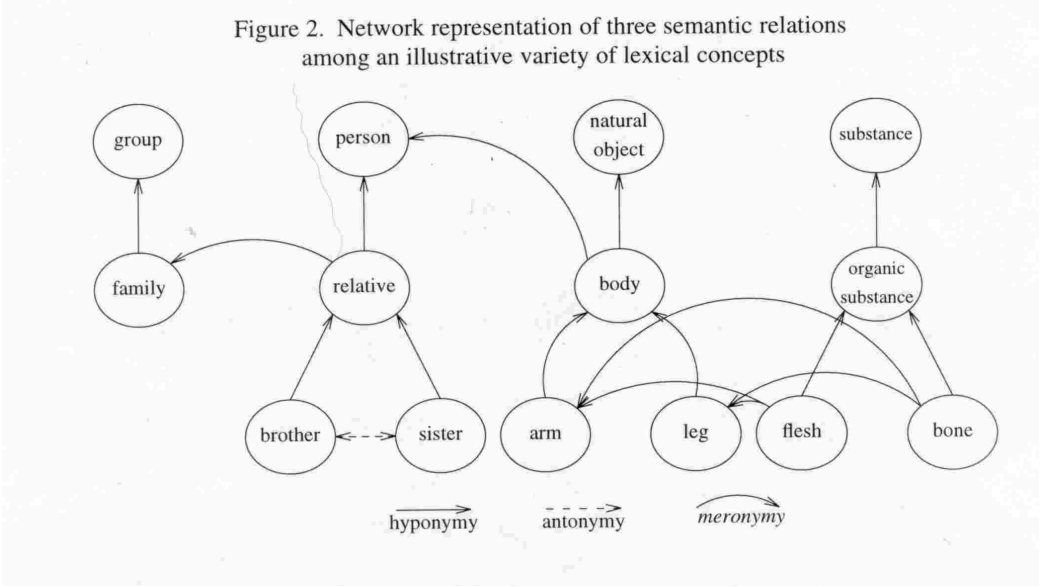


Figura 4.10: Esempio di relazioni tra synsets in WordNet

dente.
Le istanze sono sempre un nodo foglia (terminale) nelle loro gerarchie.

MultiWordNet

MultiWordNet è una banca dati lessicale multilingue in cui il WordNet per l'italiano è strettamente allineato con Princeton WordNet 1.6.
I synset italiani sono creati in corrispondenza dei synset Princeton WordNet, quando possibile, e le relazioni semantiche sono importate dai corrispondenti synsets inglesi; cioè, si assume che se ci sono due synsets in Princeton WordNet, ed esiste una relazione tra di loro, la stessa relazione vale tra i corrispondenti synset in italiano [14].
Vediamo più da vicino come sono stati realizzati tali allineamenti.

L'*Assign-procedure* sfrutta le informazioni sulle traduzione equivalenti contenute nel dizionario Collins per costruire synsets italiani in corrispondenza di synsets già esistenti nel Princeton Wordnet. Un algoritmo di mapping prende come input un senso della parola italiana, con tutte le informazioni correlate, e cerca di assegnare il senso ad un synset di WordNet inglese. L'algoritmo si basa sull'attivazione di una serie di regole, ciascuna di esse tenendo

conto di un particolare tipo di informazione, come, ad esempio, la presenza di un codice semantico nel senso italiano, e.g. "culinario" etichetta uno dei tre sensi della parola "pizza". Ogni regola contribuisce alla assegnazione di un senso ad un synset con un punteggio parziale. L'uscita dell'algoritmo è, o un'assegnazione del senso italiano per un certo synset inglese, quando il punteggio globale (dato dalla somma dei punteggi parziali fornite dalle singole regole) raggiunge una soglia fissa, o , quando il punteggio globale non raggiunge la soglia, viene scartato. L'insieme di assegnazioni prodotte, esaminando tutti i sensi delle parole italiane del dizionario costituisce un WordNet italiana creato automaticamente ed allineato con il Princeton WordNet. I dati della versione automatica vengono poi confrontati con dati acquisiti manualmente, con l'obiettivo di migliorare gradualmente il livello di precisione dell'algoritmo.

*Lexical-Gaps*identifica le lacune lessicali in modo semi-automatico. Per lacuna lessicale si intende, l'assenza di una corrispondenza diretta tra un vocabolo in inglese e lo stesso in italiano.
La procedura classifica le *Traduzione Equivalenti* in due gruppi principali: *espressioni idiomatiche* e *collocazioni ristrette* da un lato e *combinazioni libere di parole* (che implicano lacune) dall'altro. La conoscenza contenuta nei dizionari, la regolarità strutturale esposta da espressioni idiomatiche e le collocazioni ristrette, possono essere sfruttate per distinguere automaticamente tra loro con un certo grado di fiducia.

WordNet Affect

WordNet-Affect è un'estensione dei domini di WordNet , tra cui un sottoinsieme di synsets idonei a rappresentare concetti affettivi, correlati con parole affettive.
Similmente al metodo descritto per l'assegnazione di etichette di dominio, sono associate ad un numero di synsets di WordNet, una o più etichette affettive (*A-label).*
In particolare, i *concetti affettivi* che rappresentano lo stato emotivo, sono individuati da synsets, contrassegnati con l'etichetta di una emozione.

La risorsa è stata ampliata con una serie di ulteriori A-label (chiamate categorie emozionali), gerarchicamente organizzate, al fine di specializzare i synsets etichettati.

La struttura gerarchica del nuovo A-label è stata modellata sul rapporto *hyperonym di WordNet*.

Tutte le parole possono potenzialmente trasmettere significato affettivo. Ognuna di loro, anche quella più apparentemente neutra, può evocare esperienze piacevoli o dolorose.
Mentre alcune parole hanno significato emotivo rispetto alla storia individuale, per molte altre la potenza affettiva fa parte dell'immaginario collettivo (ad esempio, le parole "mamma", "fantasma", "guerra"...).

Pertanto, è interessante individuare un modo per misurare il significato affettivo di un termine generico.
A questo scopo,è stato studiato l'uso di parole in produzioni testuali, e in particolare le loro co-occorrenze con le parole in cui il significato affettivo è esplicito.
Dobbiamo distinguere tra le parole che direttamente fanno riferimento a stati emotivi (ad esempio i termini "paura", "allegro") e quelli con solo un riferimento indiretto che dipende dal contesto (ad esempio, le parole che indicano le possibili cause emotive, come "mostro", o risposte emotive, come un grido).
Chiamiamo le prime, *parole affettive dirette* e le ultime, *parole affettive indirette* [8].

Sono state organizzate le parole affettive dirette in WordNet-Affect.
Quindi, è stata sviluppata una funzione di selezione (denominata *Affective-weight*) basata su un meccanismo di similarità semantica acquisita automaticamente in modo *unsupervised*, da un grande corpus di testi (100 milioni di parole), al fine di individuare il lessico affettivo indiretto.

Applicato ad un concetto (ad esempio un WordNet synset) e una categoria emotiva, questa funzione restituisce un valore che rappresenta l'affinità semantica con quella emozione.
In questo modo è possibile assegnare un valore al concetto rispetto a ciascuna categoria emotiva, ed eventualmente selezionare l'emozione con il valore più alto.

Il sistema messo a punto, quindi, è in grado di mettere a fuoco in modo

A-Labels	Examples
EMOTION	*noun anger#1, verb fear#1*
MOOD	*noun animosisy#1, adjective amiable#1*
TRAIT	*noun aggressiveness#1, adjective competitive#1*
COGNITIVE STATE	*noun confusion#2, adjective dazed#2*
PHYSICAL STATE	*noun illness#1, adjective all in#1*
HEDONIC SIGNAL	*noun hurt#3, noun suffering#4*
EMOTION-ELICITING SITUATION	*noun awkwardness#3, adjective out of danger#1*
EMOTIONAL RESPONSE	*noun cold sweat#1, verb tremble#2*
BEHAVIOUR	*noun offense#1, adjective inhibited#1*
ATTITUDE	*noun intolerance#1, noun defensive#1*
SENSATION	*noun coldness#1, verb feel#3*

Figura 4.11: Elenco di etichette in WordNet Affect, A-Label, con relativi esempi

selettivo tipi positivi, negativi, ambigui o neutrali di emozioni.
Ad esempio, data la parola *difficoltà* come termine di ingresso, il sistema propone come emozioni correlate (A-Label): identificazione, negativo-preoccupato, ambiguo, apatia.
In figura 4.11 sono rappresentate le etichette messe a punto in WordNet Affect, A-Label, con relativi esempi di vocaboli associati. Si noti come, nel loro insieme, le etichette cerchino di rappresentare non solo informazioni strettamente legate all'emozione, quanto ai meccanismi, le cause e gli effetti, legati ad una emozione.

4.4.3 Text Feature Extraction

Nei paragrafi successivi si illustrano gli approcci più ampiamente utilizzati nel panorama della ricerca, al fine di estrarre informazioni da un testo scritto.

Keyword Spotting

L'approccio più intuitivo e probabilmente anche il più popolare a causa della sua accessibilità ed economia.

Il testo viene classificato in base alla presenza di parole che trasportino un significato, intrinsicamente legato ad una emozione.
Pensiamo alle parole come *felicità*. Questo tipo di metodo sicuramente risulta fallace da due punti di vista:
in primis, basterà una negazione all'interno dell'espressione per rendere inefficace l'analisi, infatti se pensiamo all'espressione: *"Oggi non è una bella giornata"*, il modello, non dando alcun peso alla presenza della negazione, sbaglierà la sua stima.
Un altro punto debole è rappresentato dalla dipendenza del sistema da parole chiave (*keyword*):
infatti l'espressione: *"Ho lasciato la mia ragazza"*, per quanto possa sottointendere uno stato d'animo ben evidente, esso non traspare dalla semplice analisi del testo, ossia non ci sono parole chiave, che possano inequivocabilmente permettere di classificare l'espressione in una emozione ben definita [31].

Lexical Affinity

Questo risulta essere una versione migliorata del metodo visto in precedenza, infatti l'idea è quella di assegnare ad una parola un certo grado di affinità con una emozione; pensiamo alla parola *Ospedale*:
potremmo pensare ad essa come veicolatrice di emozioni negative per buona parte delle espressioni in cui viene inserita; pensiamo altresì alla frase: *"Vado in ospedale per la nascita di mio nipote"*.
Risulta evidente come in questo caso, invece, il significato implichi una sensazione di positività.
Quindi in un ipotetico vettore dove ciascuna dimensione è una emozione, potremmo considerare una distribuzione, a somma unitaria, di quattro pesi, che vadano a descrivere il contenuto emozionale trasportato da ciascun vocabolo analizzato [31]:
$i.e. [w_g, w_t, w_r, w_p] \qquad con \ \sum w_i = 1;$
e.g. Incidente: $[0, 0.5, 0, 0.5]$;

Di qui si deduce che il modello preveda la dipendenza da un corpus che raccolga i vocaboli e le relative affinità implicate.
Questo però non sarebbe sufficiente, poichè in ogni caso il sistema resta sen-

sibile rispetto alla negazione (*Ïo evitato un incidente"*) oppure a significati secondari (*"L'ho incontrato per incidente"*).
Quindi questo suggerisce la necessità di rendere questo strumento *domain-independent*, ossia indipendente rispetto al dominio di appartenenza della parola analizzata, pensiamo alla parola *"fisica"*, essa può essere intesa come forza fisica, così come la materia scientifica.

Statistical Natural Language Processing

L'idea è quella di partire da un algoritmo di machine learning, partendo da un corpus di vocaboli etichettati emotivamente, che permetta, non solo, di estrarre le occorrenze individuate nel corpus, ma anche, similmente al lexical affinity, di affidare anche alle restanti parole, così come alla punteggiatura, una affinità emotiva proprio perchè affiancate a parole etichettate.
In generale è necessario avere a disposizione corpora molto grandi e inoltre il sistema sembra avere prestazioni accettabili solo in presenza di testi in input molto lunghi.
Risulta poco efficace con espressioni di breve durata.
Una tecnica utilizzata è proprio la LSA (Latent Semantic Analysis) [31].

I modelli appena descritti, rappresentano le logiche adottate per la manipolazione ed estrazioni di informazioni, provenienti da testi scritti. Per il sistema messo punto, l'approccio utilizzatoè quello della lexical affinity, poichè di immediata implementazione ed integrazione con il sistema messo a punto ed in linea con gli strumenti di ricerca adottati, come MultiWordNet e WordNet Affect. In più, come si evince da [9], si punterà all'estrazione di informazioni quanto più sensibili rispetto al contenuto emotivo che veicolano.

Il set di features scelto per le informazioni legate al testo scritto sono quindi state:

- Numero di parole per espressione
- Percentuale di verbi
- Percentuale di sostantivi
- Percentale di aggettivi
- Percentuale di avverbi

- Parole Negative
- Parole Positive

In paricolare, l'accezione di positivo e negativo, è rispetto al significato intrinseco della parola, che di per sè, può trasportare una informazione legata ad un senso di positività, quale può essere la parola "festa", così come negatività come "morte".
Questo getta luce su una problematica più sottile, quanto determinante e attuale.
Cercare di individuare il significato figurato in una espressione, che alla prima interpretazione può apparire completamente diverso; e soprattutto farlo automaticamente, attraverso l'uso di una macchina.

Infine, quello che si vuole sottolineare è che l'integrazione dell'analisi testuale, affiancante l'analisi audio, è ad uno stato di ricerca iniziale, che richiede un approfondimento rispetto alle tecniche di estrazione stesse delle informazioni di interesse, oltre che al rafforzamento di degli strumenti di analisi testuali, quali i corpora di vocaboli e le loro relazioni.

4.5 Ottimizzazione delle feature

Nel processo di scelta del set di features si è dapprima tenuto conto dei risultati ottenuti in altri studi che hanno affrontato il tema della classificazione di campioni di un dataset, rispetto all'estrazione dello stato emozionale dei parlatori.
In più si è tenuto conto anche di altre problematiche che potessero riguardare le caratteristiche del singolo speaker, ossia esplorare le interazioni tra l'espressione di una emozione e le caratteristiche legate più propriamente all'individuo che le esprime, questo per rendere il più possibile il sistema robusto rispetto al cambiamento del parlatore, che però esprime la stessa emozione.
Quindi si è dapprima partiti con un set di di 110 feature, così come elencato nella sezione dedicata al modello dei moduli di estrazione delle features.
Al fine di ridurre i costi computazionali dell'applicazione e, non di meno, per valutare il peso, il potere informativo, quindi discriminatorio, di una particolare feature o di un gruppo di esse, si è proceduto con alcune tecniche di riduzione dello spazio, e con la analisi di alcune figure di merito per valutare

questi aspetti.

4.5.1 Principal Component Analysis

L'idea intuitiva sta nel considerare un ellissoide di dimensione n che modelli un dataset.
E' ragionevole che la varianza attorno ad alcuni assi sia minima, quindi che trasportino un'informazione minore rispetto alle altre componenti.

Quindi, si procede proiettando ciascuna osservazione nel dataset, sulla prima componente principale, massimizzando la varianza di questa nuova variabile così ottenuta.
Procedendo, si otterrà un sistema di variabili, di pari numero rispetto al sistema iniziale, con la differenza però che la somma delle varianze delle prime componenti, dette appunto principali, racchiuderà la maggior parte dell'informzaione iniziale. Questo permette, quindi, di escludere dal feature set, tutte quelle componenti che non incrementano l'efficacia discriminatoria del sistema, ma ne compromettono l'efficienza. In termini più formali:
Considerata la matrice $X, n \times p$, possiamo considerare una trasformazione di tale dataset secondo la relazione:

$$t_{k(i)} = x_{(i)} \cdot w_{(k)} \tag{4.11}$$

dove:

$t_{(i)} = (t_1, ..., t_p)_{(i)}$ è il vettore di score

$w_{(k)} = (w_1, ..., w_p)_{(k)}$ è il vettore dei pesi o *loadings*

$x_{(i)}$ sono i vettori riga del dataset X

Quindi,

$$w_{(1)} = arg\,max_{\|\omega\|} \{\sum_i (t_1)^2_{(i)}\} = arg\,max_{\|\omega\|} \sum_i (x_{(i)} \cdot w)^2 \tag{4.12}$$

che in forma matriciale diventa:

$$w_{(1)} = argmax_{\|\omega\|=1} \|X\omega\|^2 = argmax_{\|\omega\|=1} \|\omega^T X^T X \omega^2\| \tag{4.13}$$

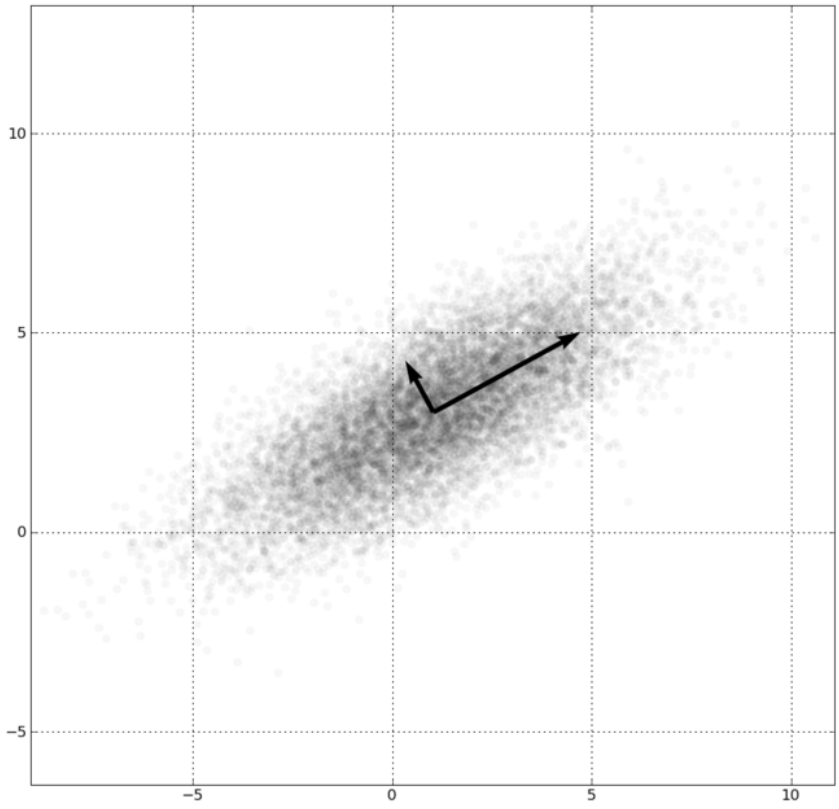


Figura 4.12: Elissoide di 2 dimensioni.
Si pone in evidenza la differenza tra le varianze, nelle due direzioni, rispettivamente. Quindi la componente a varianza minore può essere trascurata, approssimando l'informazione totale alla unica componente restante.

$$w_{(1)} = arg\,max\,\|\frac{(\omega^T X^T X\omega)}{(\omega^T\omega)\omega^2}\| \tag{4.14}$$

La quantità tra parentesi è detta quoziente di Rayleigh e, per matrici simmetriche come $X^T X$, si ha il massimo corrispondente al maggiore degli autovalori individuati, aventi w come corrispondente autovettore.
Di qui si può intuire che l'informazione contenuta nel generico $x_{(i)}$ (riga i-esima della matrice X) espressa nella sua prima componente nel nuovo spazio

è:

$$t_{1(i)} = x_{(i)} w_{(1)} \tag{4.15}$$

Le componenti successive si estraggono sottraendo le prime k-1 componenti dal dataset X:

$$\hat{X}_k = X - \sum_{s=1}^{k} X w_{(s)}^T w_{(s)} \tag{4.16}$$

e quindi:

$$w_{(k)} = arg\ max\ \{\| \frac{\omega^T \hat{X}_{k-1}^T \hat{X}_{k-1} \omega}{\omega^T \omega} \|^2\} \tag{4.17}$$

Quindi l'idea è quella di decorrelare i dati in maniera successiva, rispetto alla varianza delle singole variabili (feature).
In questo modo avremo un dataset rappresentato da un numero inferiore di dimensioni, ma con una perdita non determinante di informazione.
In 4.12 è rappresentato un semplice esempio che mette in evidenza il meccanismo e le ipotesi su cui si basa la tecnica PCA. Il punto debole di questo metodo risiede nel fatto che è sensibile alla tecnica con cui si sono normalizzati i dati, e non è definibile a priori quale sia il metodo di normalizzazione migliore.[wiki]

4.5.2 Analisi della Varianza (One-Way ANOVA)

Tale analisi si propone di individuare quali siano i singoli fattori (features) o gruppi di essi, che determinino la differenziazione tra raggruppamenti all'interno di una popolazione di dati.
Definiamo la variabilità delle medie intra-gruppo, come la variabilità del valor medio rispetto ad un fattore (analisi della varianza a un fattore) e definiamo la variabilità inter-gruppo come la variabilità del valor medio di tale fattore tra gruppi diversi.
Mentre la prima è considerata un errore casuale, la variabilità inter-gruppo è da attribuirsi alla differenza che esiste tra i raggruppamenti e viene definito effetto del trattamento.
A questo punto ci si propone di verificare due ipotesi mutuamente esclusive: $\{(H_1$: almeno una media è diversa $(H_0 : \mu_1 = \mu_2 = \cdots = \mu_c)$ con μ_i il valor medio diun fattore di ogni gruppo Per verificare tali ipotesi individuiamo alcune figure di merito:

$$SST(sommadeiquadratitotale) = SSA(sommadeiquadratideigruppi) + SSW(sommadeiquad$$

$$SSA = \sum_{j=1}^{c} (x_j - x)^2 n_j \tag{4.18}$$

variabilità inter-gruppo

$$SSW = \sum_{j=1}^{c} sum_{i=1}^{n_j} (x_{ij} - x_j)^2 \tag{4.19}$$

variabilità intra-gruppo

Dove:
c:numero di gruppi
n_j :campione j-esimo
n: numerosità complessiva
j : generico gruppo
$x_i j$: i-esima osservazione appartenete al j-esimo gruppo
(x_j) : media del j-esimo gruppo
(x) : media generale
A questo punto possiamo individuare:
$(c-1)$ gradi di libertà di SSW
$(n-c)$ gradi di libertà di SSA
Dividendo ciascuna somma di quadrati per i rispettivi gradi di libertà otteniamo MSA (media dei quadrati tra gruppi) MSW (media dei quadrati all'interno del gruppo) MST(media dei quadrati totale).
Definiamo infine come figura di merito F come:

$$F = \frac{SSA/n-c}{SSW/c-1} = \frac{MSA}{MSW} \tag{4.20}$$

Se l'ipotesi H_0 risulta vera $F \approx 1$, altrimenti $F > 1$ Guardando alla distribuzione di Snedecor/Fisher con $(c-1)$ gradi di libertà al numeratore e $(n-c)$ al denominatore, e fissato un valore di significatività α (normalmente al 5%), si rifiuta l'ipotesi H_0 se il valore di F rilevato supera un valore critico F_s.
Inoltre si considera un valore $p-value$ che indica la probabilità di osservare un valore di F uguale o superiore a quello osservato, nel caso l'ipotesi nulla sia vera, ossia se si osserva un $F \approx 1$.
Anche il per il $p-value$ si sceglie un valore di soglia sotto il quale si considera come probabile che si possa osservare un valore di F maggiore di 1 e che quindi l'ipotesi H_0 non sia vera.

Normalmente il valore del $p - value$ si fissa a 0,05.

4.5.3 Overfitting e Pre Clustering

Un approccio adottato per smussare il fenomeno dell'overfitting è stato l'utilizzo del metodo di k-means.
K-means è una tecnica di partizionamento.
La funzione Kmeans partiziona i dati in k cluster mutuamente esclusivi, e restituisce l'indice del cluster a cui si è assegnata ogni osservazione.
Kmeans tratta ogni osservazione nei dati come un oggetto, avente una posizione nello spazio.
Nella fattispecie la dimensione del nostro spazio sarà pari al numero di feature.
Esso trova una partizione in cui, gli oggetti all'interno di ciascun cluster, siano tra loro i più vicini possibile, e lontani da oggetti, in altri cluster possibili.
Ogni cluster nella partizione è definito dai suoi oggetti membro e dal suo centroide.
Il centroide per ogni cluster è il punto in cui la somma delle distanze da tutti gli oggetti del cluster è minima.
Kmeans calcola i centroidi dei cluster in modo diverso per ogni metrica di distanza, puntando a minimizzare la somma rispetto alla misura specificata.
Esso utilizza un algoritmo iterativo che minimizza la somma delle distanze da ogni oggetto al suo centroide cluster, su tutti i cluster.
Questo algoritmo muove gli oggetti tra i cluster fino a quando la somma non può essere diminuita ulteriormente.
Il risultato è un insieme di cluster che sono il più compatti e separati possibile.
Kmeans utilizza l'algoritmo k-means ++ per l'inizializzazione dei centroidi dei cluster e la Squared Euclidean Distance per determinare le distanze tra i gli oggetti ed i relativi cluster.[24]
La scelta di tale metrica è suggerita anche in [18], dove si pone in evidenza che non ci siano differenze in prestazioni tali da portare a cambiare questo parametro.
L'algoritmo appena descritto, vuole mettere in evidenza come si cerchi di ridimensionare, oltre che la dimesione dello spazio, anche la dimensione dei dati analizzati, predisponendoli ad una classificazione succesiva che godrà di

un dataset in ingresso robusto, rispetto al suo obiettivo.
Se un gruppo di dati fa riferimento ad uno stesso fenomeno o condivide una stessa caratteristica, è ragionevole pensare che, tali dati, saranno distribuiti più o meno uniformemente attorno ad un centroide (centro di massa).
Facendo leva su questa idea, è stato possibile individuare per ciascuna emozione, un cluster. In particolare, è stato funzionale, non tanto ai fini della classificazione, quanto ad un preprocessing dei dati; infatti è stato applicato tale algoritmo, su datasets, già etichettati, avendo come unico fine quello di garantire una robustezza successiva, nel modello messo a punto. Vediamo come:
si sono analizzate le distanze di ciascun sample dal centroide del gruppo di appartenenza che è stato funzionale all'esclusione, previa la scelta di una soglia, di alcuni dei sample, troppo distanti dal proprio centroide, e che quindi avrebbero potuto portare a errori di classificazione successivi.
Nella fattispecie si è scelta una figura di merito che quantificasse la dispersione dei sample nel dataset, rispetto alla classe di appartenza, quindi la distanza media dal centroide di appartenenza, rappresenta la nostra soluzione.
Per determinare quale fosse la distanza di soglia per escludere i sample dal nuovo dataset, si è proceduto con prove ripetute, con soglie diverse e si è misurata, quindi, l'accuratezza del sistema, a seguito di tali prove.

Le figure 4.13 4.14 4.15 4.16 4.17 4.18 rappresentano i risultati di tale analisi.

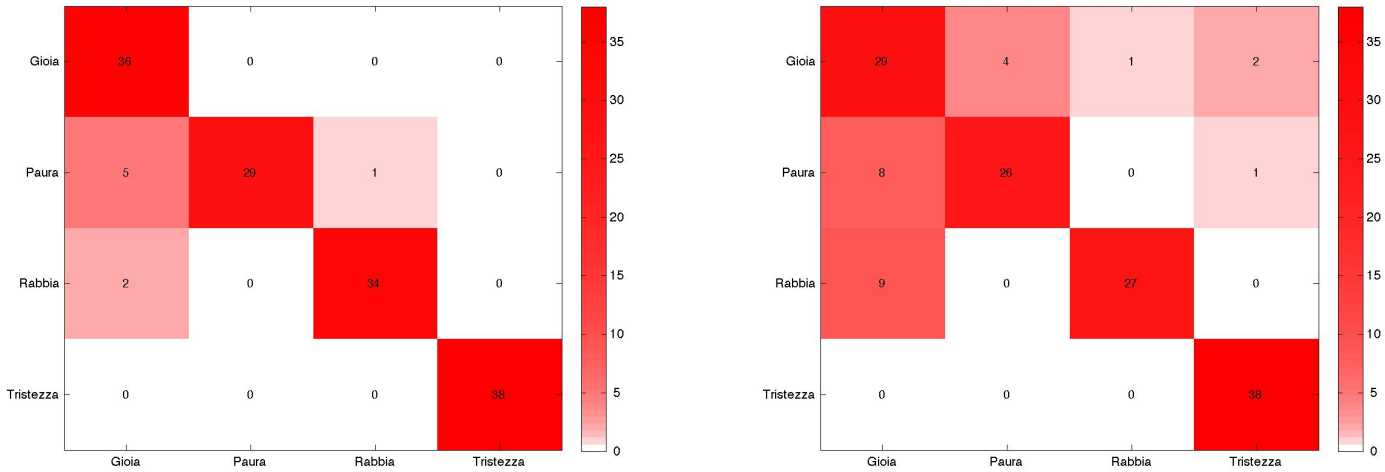


Figura 4.13: Classificazione con pre clustering, Threshold=avgDist+20% Confusion Matrix: (a) Best case (b) Worst case

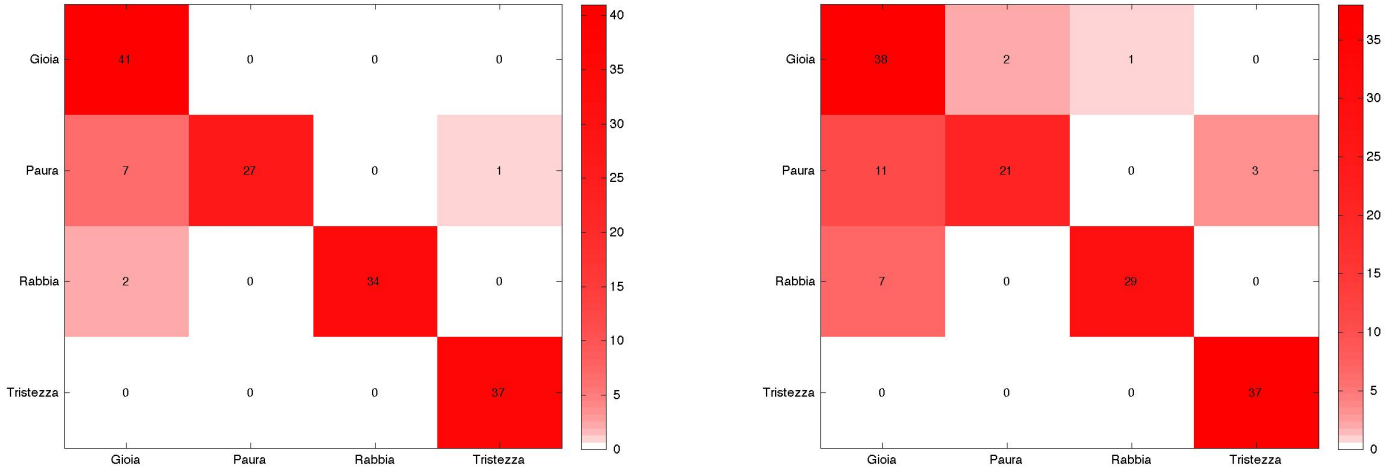


Figura 4.14: Classificazione con pre clustering, Threshold=avgDist+30% Confusion Matrix: (a) Best case (b) Worst case

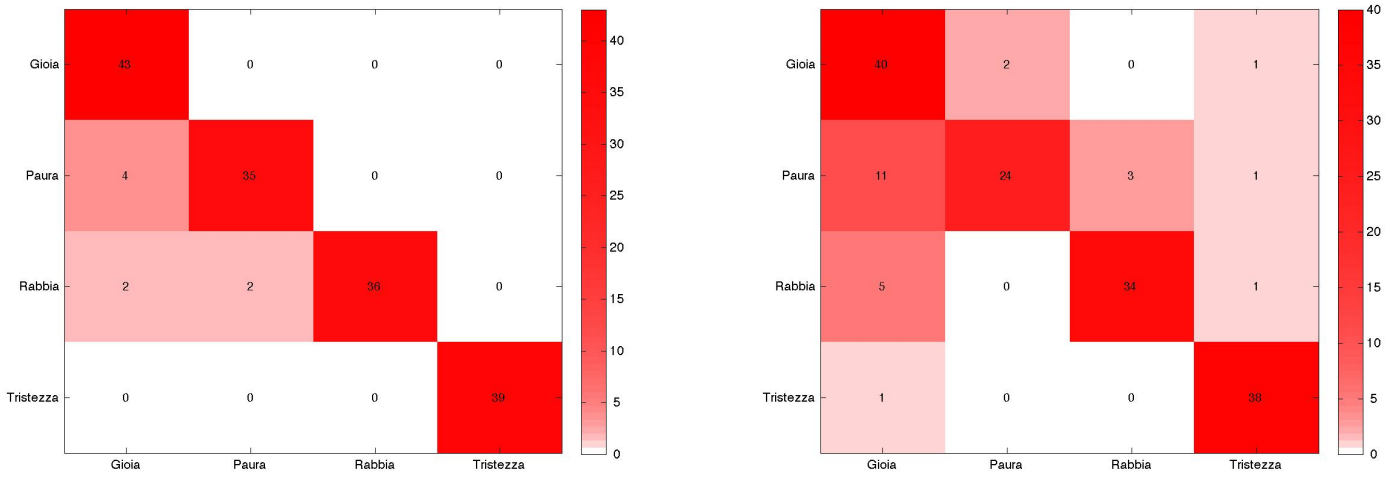


Figura 4.15: Classificazione con pre clustering, Threshold=avgDist+40% Confusion Matrix: (a) Best case (b) Worst case

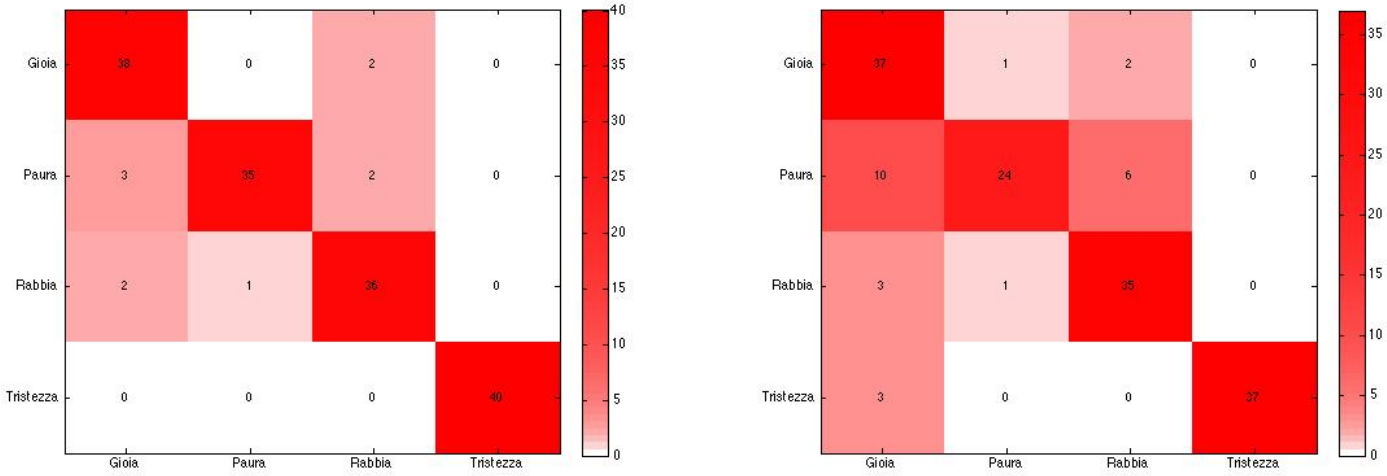


Figura 4.16: Classificazione con pre clustering, Threshold=avgDist+50% Confusion Matrix: (a) Best case (b) Worst case

Quindi tutti i sample la cui distanza è superiore alla soglia imposta, sono stati esclusi.

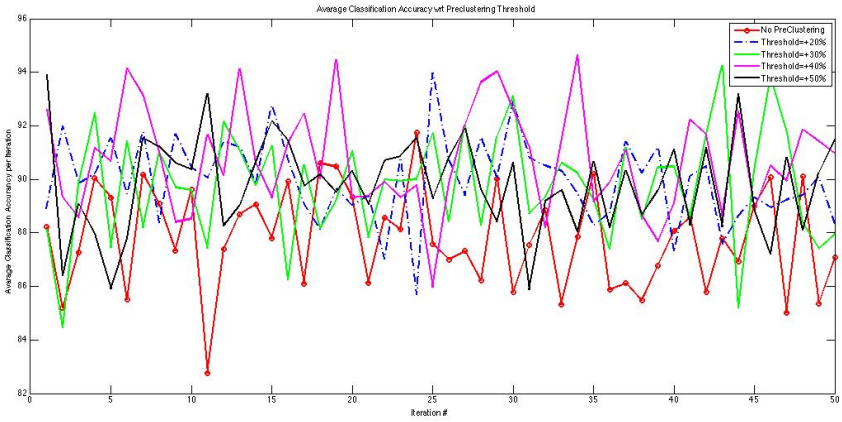


Figura 4.17: Confronto Prestazioni di Classificazione con Pre Clustering, al variare della soglia

Formalmente l'algoritmo si comporta come segue:

$x^{(i)} \epsilon \mathbf{R}^n$ siano i vettori delle osservazioni, di dimensione n, pari al numero di features estratte.

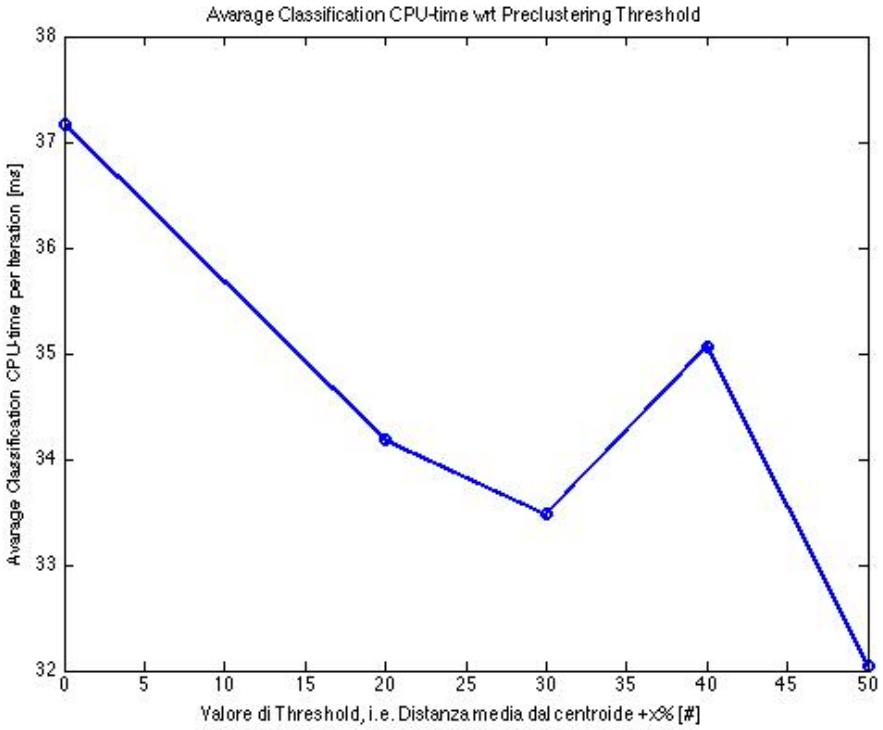


Figura 4.18: Confronto Tempi di Classificazione con Pre Clustering, al variare della soglia

$(\mu_1, \mu_2, \ldots \mu_k) \epsilon \mathbf{R}^n$ siano i k centroidi casualmente inizializzati $c^{(i)}$siano le etichette dell' i-esima osservazione Ripetere fino alla convergenza:

$\forall i$,

$$c^{(i)} := argmin_j \|x^{(i)} - \mu_j\|^2 \tag{4.21}$$

$\forall j$,

$$\mu_j := \frac{\sum_{i=1}^{m} 1\{c^{(i)} = j\} x^{(i)}}{\sum_{i=1}^{m} 1\{c^{(i)} = j\}} \tag{4.22}$$

In figura 4.19 sono rappresentate sei iterazioni dell'algoritmo Kmeans.

A questo punto avremo un feature set composto da 110 feature.
Il numero è considerevole, ma uno degli obiettivi risiede anche nell'esplorare quali siano le componenti rilevanti rispetto alla distinzione tra le emozioni. Quindi ad un feature set così composto è stato applicato l'algoritmo di Prin-

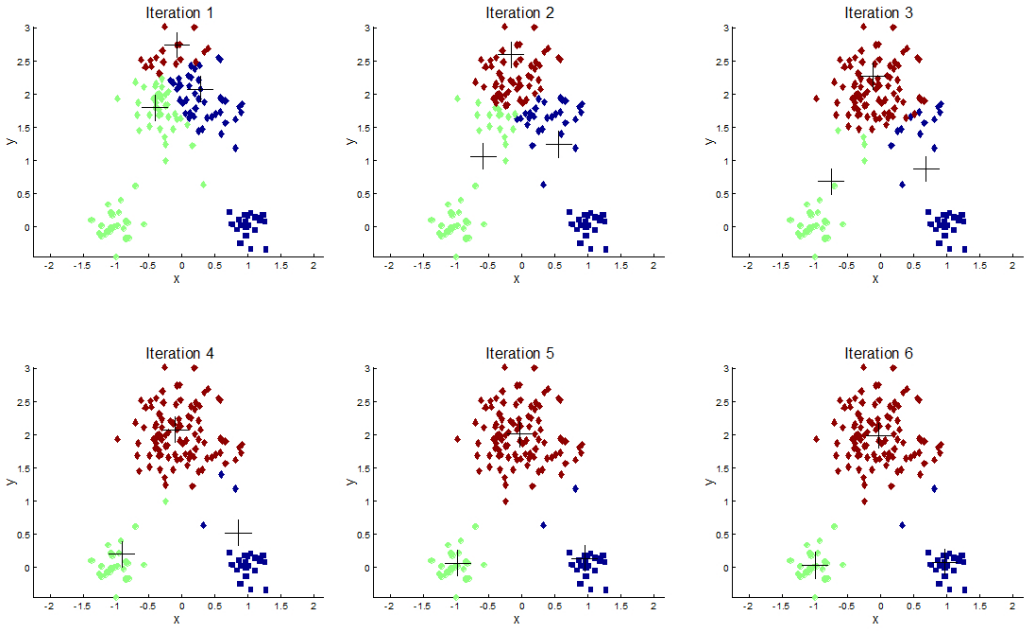


Figura 4.19: Esempio illustrativo di 6 iterazioni.
Si pongono in evidenza come ad ogni iterazione si aggiornino le posizioni dei centroidi -croci-, e i rispettivi dati appartenenti al cluster che rappresentano -gruppi di punti, di colore diverso-.

cipal Component Analysis, i cui risultati sono riportati in figura 4.20.

A questo punto si llustrano i risultati a seguito del processo di ottimizzazione del sistema. In figura 4.20 si mostra un riassunto delle informazioni ottenute dopo l'applicazione della PCA al feature set. Ciò che si evince è che l'algoritmo di PCA riesce a riassumere in uno spazio di 80 componenti, quindi ben 30 in meno rispetto a quelle di partenza, il 99.18% dell'informazione totale, riducendo la complessità del sistema.
Infine ciò che è necessario evidenziare è che il metodo PCA risulta molto sensibile ai dati analizzati e alla loro normalizzazione.
Nella fattispecie si è proceduto, con una normalizzazione z-score, in cui, ogni vettore di feature (ogni colonna della matrice dataset), sarà a *media nulla e a deviazione standard unitaria*, questo perchè tale normalizzazione è funzionale alla classificazione successiva.

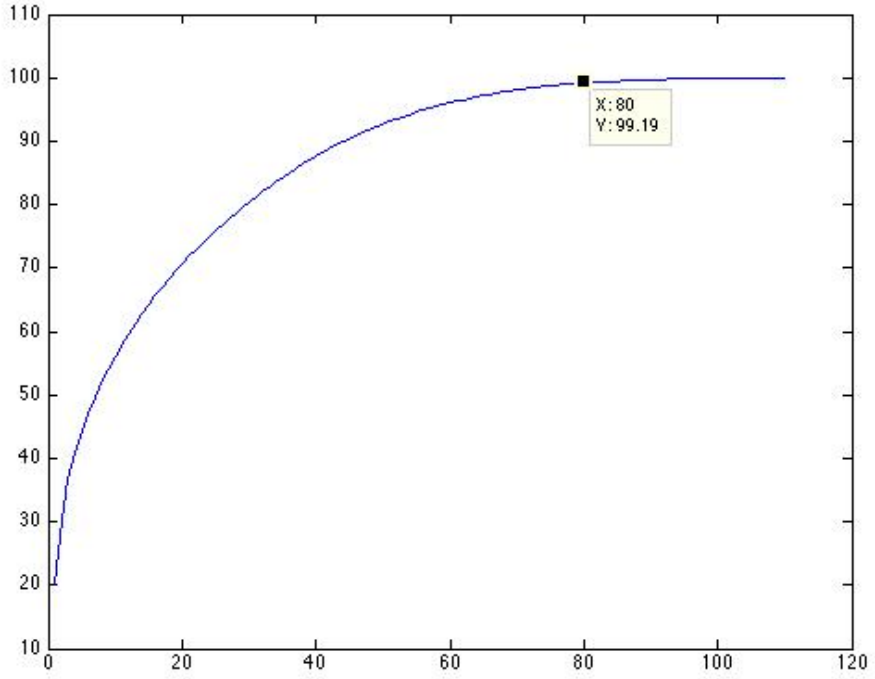


Figura 4.20: Risultato dell'applicazione della PCA, con X il numero di feature, Y la quotaparte cumulata di informazione, considerando il 100% come rappresentata dalla totalità del feature set.

Così come per il pre clustering, anche nel caso dell'ottimizzazione della dimensione dello spazio in cui operare, si è optato per prove successive, facendo variare il numero di componenti (feature) e quindi valutare le performance esibite dal sistema al variare di tale dimensione. Le figure 4.21 4.22 4.23 4.24 4.25 seguenti evidenziano tali risultati:

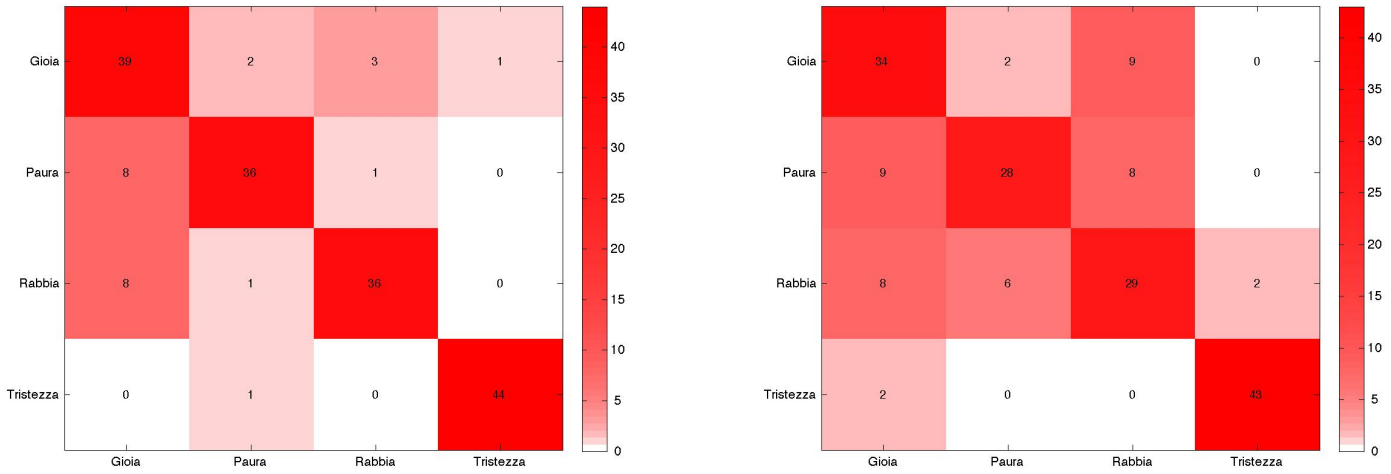


Figura 4.21: Classificazione con PCA, Dimensione=50 componenti Confusion Matrix: (a) Best case (b) Worst case

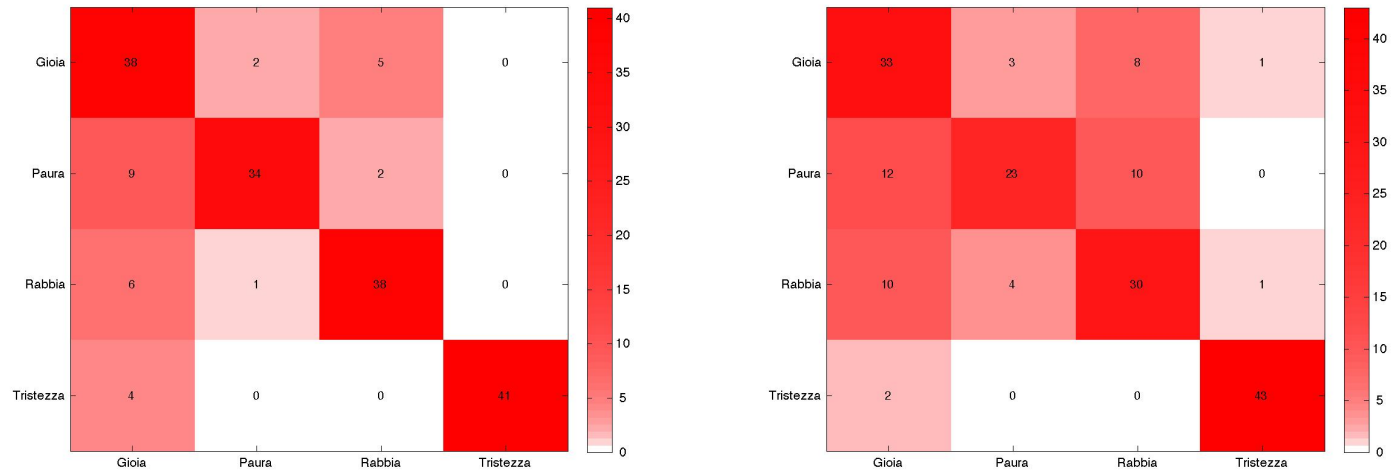


Figura 4.22: Classificazione con PCA, Dimensione=60 componenti Confusion Matrix: (a) Best case (b) Worst case

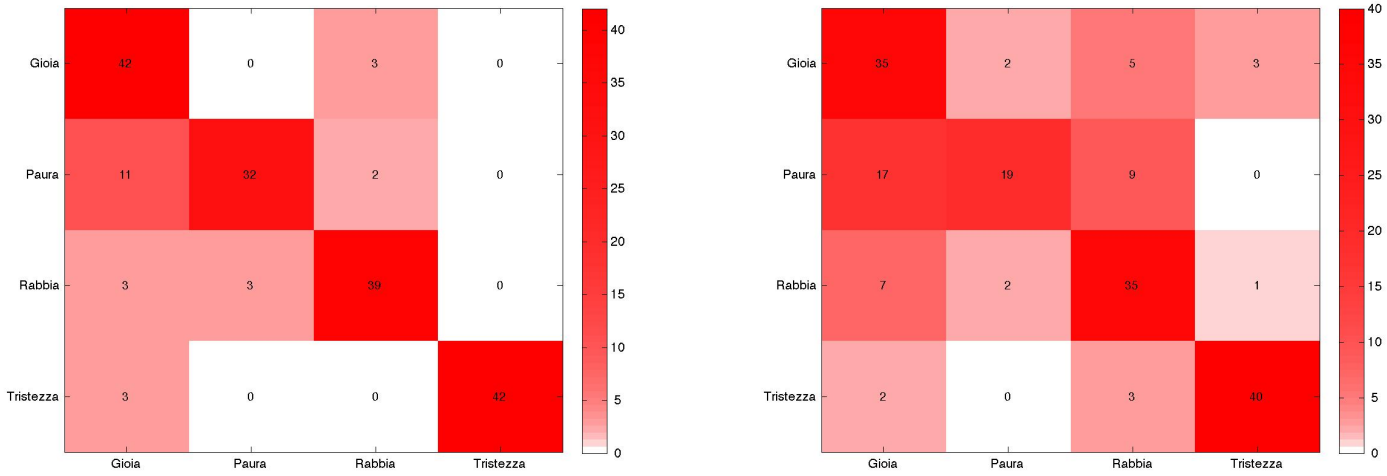


Figura 4.23: Classificazione con PCA, Dimensione=80 componenti Confusion Matrix: (a) Best case (b) Worst case

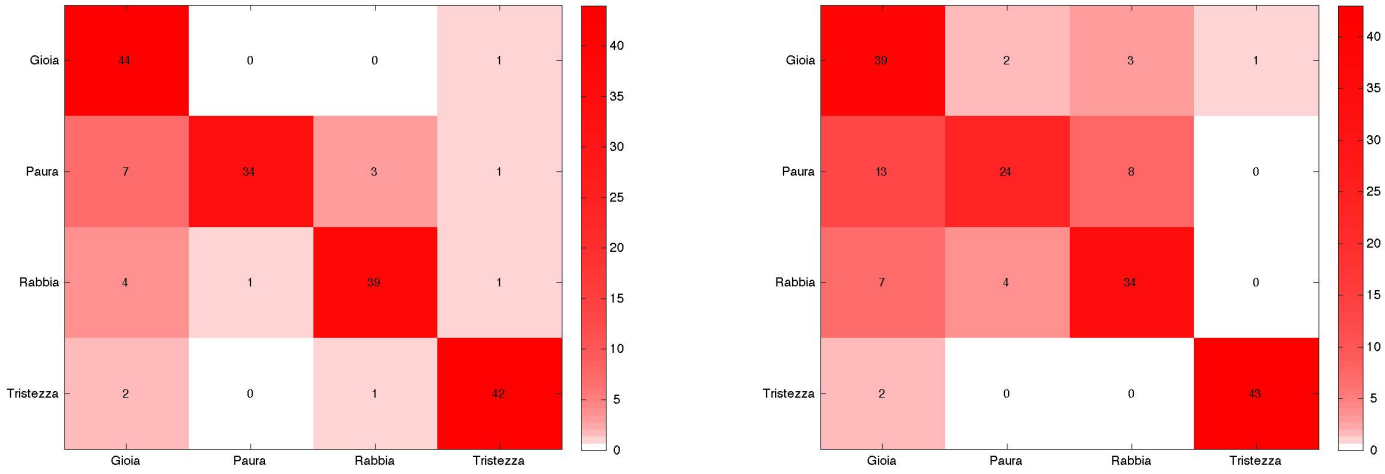


Figura 4.24: Classificazione con PCA, Dimensione=100 componenti Confusion Matrix: (a) Best case (b) Worst case

Parallelamente a questa analisi si è eseguito il test ANOVA che ha evidenziato quanto illustrato nelle figure 4.26 4.27:

il set di feature risulta essere abbastanza robusto, la maggior parte dei fattori (ossia le feature), ha la media, nei rispettivi gruppi, diversa, quindi è valida l'ipotesi per cui esse siano annoverabili feature per la modellazione delle varie emozioni.
In particolare è emerso che delle 110 feature, 34 siano, sostanzialmente, superflue.
In particolare il test rileva i coefficienti delta-delta MFCC come superflui.
Questo test è stato funzionale ad una esplorazione del potere informativo del set di feature messo a punto, e sarà un punto di riferimento nel ridimensionamento dello spazio, nel momento in cui si andrà ad applicare l'algoritmo di PCA.
A questo punto si è proceduto con l'implementazione dell'algoritmo di classificazione.
Al dataset così costituito è stato applicato un algoritmo di k-means, per un clustering preliminare dei dati, per eliminare dal dataset quei campioni, che rappresentano un errore, poichè non associabili ad nessuna emozione.

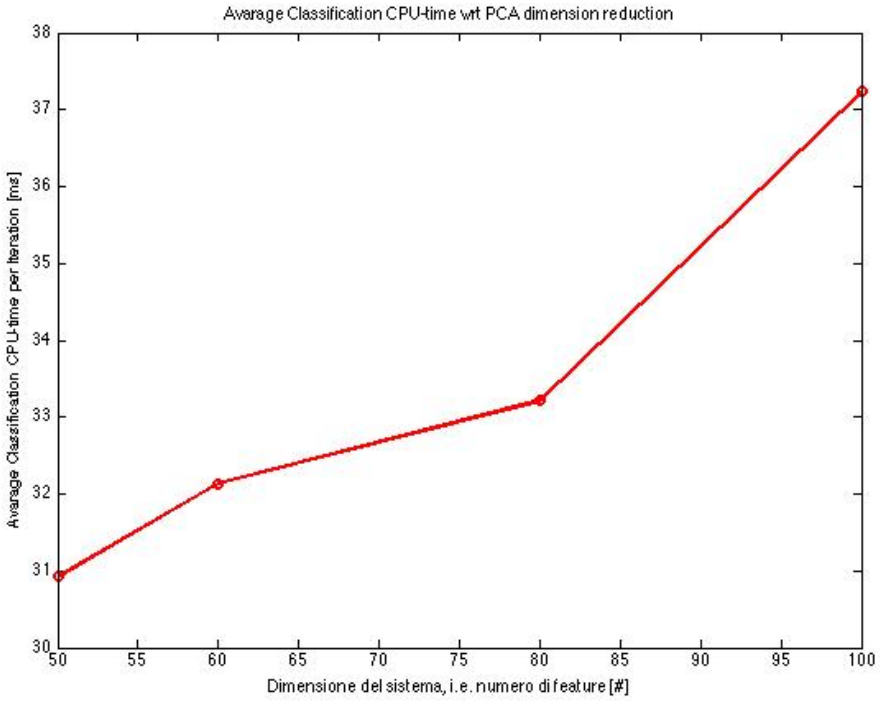


Figura 4.25: Confronto Tempi di Classificazione, al variare della dimensione del sistema

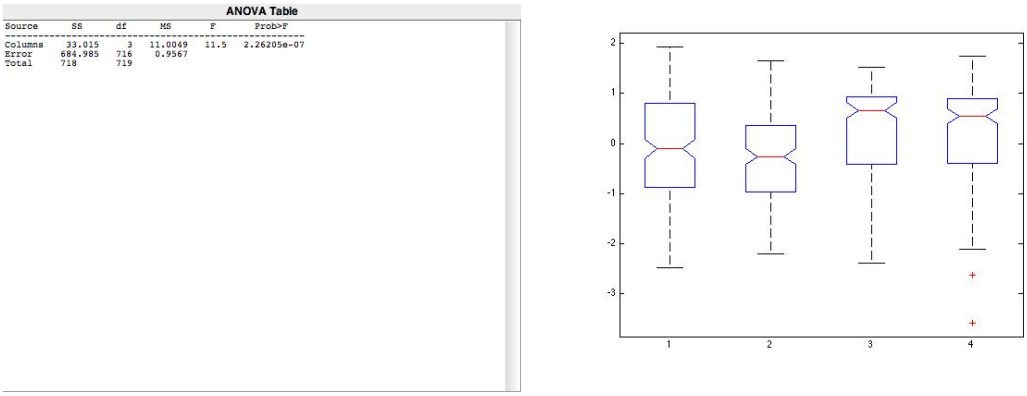


Figura 4.26: Esempio di feature che nega l'ipotesi H_0

4.6 Classificatori

Nella sezione seguente si andranno a descrivere i classificatori messi a punto per il sistema. Le scelte sono state principalmente dettante dai risultati ottenuti in altre ricerche, si vedano infatti [12] [22] oltre che frutto dell'adat-

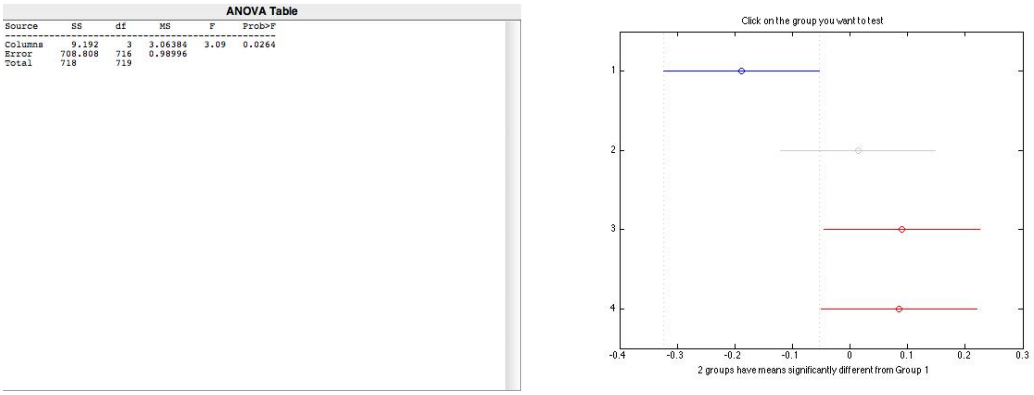


Figura 4.27: Esempio di feature che conferma l'ipotesi H_0

tamento rispetto alle risorse a disposizione. Per questi motivi si è proceduto come segue.

4.6.1 Gaussian mixture model (GMM)

In prima istanza è stato preso in considerazione questo approccio per la classificazione dei dati così raccolti. Il metodo è un classificatore di tipo *supervised*, ossia la classificazione sarà costituita da una fase di training, in cui si andrà ad estrarre il modello di interesse ed una di testing, in cui si validerà tale modello.
La scelta di tale metodo è stata dettata dagli studi precedentemente effettuati [12] che confermano quanto tale approccio sia funzionale nello Speech Processing, e, più specificatamente nello Speaker Recognition.
Infatti il modello si piega bene alle condizioni del nostro problema, ossia individuare caratteristiche acustiche e le loro evoluzioni, che possano accomunare un singolo stato emozionale.
Vediamo più nel particolare le caratteristiche di tale metodo:
Sia

$$p(\bar{x}|\lambda) = \sum_{i=1}^{M} p_i b_i(\bar{x}) i = 1, ..., M \tag{4.23}$$

una Gaussian Mixture Density, definita come una somma di M componenti, dove ciascuna è una una funzione di densità di probabilità $b_i(\bar{x})$ moltiplicata per il rispettivo peso p_i.
Infine $\bar{x}$ è definito come il vettore di valori random D-dimensionale.

Ciascuna densità $b_i(\bar{x})$ è così definita:

$$b_i(\bar{x}) = \frac{1}{2\pi^{D/2}|\Sigma_i|^{1/2}} \exp\{-1/2(\bar{x} - \bar{\mu_i})\Sigma_i^{-1}(\bar{x} - \bar{\mu_i})\} \tag{4.24}$$

dove $\bar{\mu_i}$ è il vettore della media e $\sum_i^{-1}$ è la matrice di covarianza.
L'unico vincolo è posto su $\sum_{i=1}^{M} p_i = 1$.
A questo punto potremo costruire il nostro modello e rappresentarlo come:

$$\lambda = \{p_i, \bar{\mu_i}, \Sigma_i\} \qquad i = 1, \ldots, M$$

Quindi, potremo definire quattro modelli λ_k, con $k = 1, \ldots, 4$, ciascuno per ogni emozione.
In base alla scelta della matrice di covarianza, possiamo definire un modello con una matrice di covarianza per ogni i-esima componente gaussiana (*nodal covariance*), oppure una singola matrice di covarianza per modello (g*rand covariance*) o, infine, una singola matrice di covarianza condivisa da tutti i modelli, nel nostro caso, quattro.
Infine la matrice di covarianza può essere *full* o *diagonal*.
Guardando al modello, potremmo pensare all'i-esimo vettore delle medie dell'i-esima classe (emozione) come al contenuto spettrale della i-esima classe e alla matrice delle covarianze come alle variazioni medie del contenuto spettrale.
Inoltre osservazioni sperimentali hanno permesso di verificare che la combinazioni lineare di *Gaussian Basis Function* rappresenta bene una larga classe di distribuzioni di dati diversificati tra loro.
Infatti potremmo vedere una GMM come un modello ibrido di una Vector Quantization e un modello Gaussiano mono-modale, dove il modello VQ vede la distribuzione di dati all'interno di template (ciascun vettore), mentre un modello gaussiano mono-modale prevede la distribuzione delle feature indicando una posizione (vettore media) e una ellissi, (la matrice di covarianza).
Una combinazione lineare di matrici di covarianza diagonali è già in grado di estrarre la correlazione tra i vettori delle osservazioni.
A questo punto, impostato il modello, potremo passare alla fase di training del sistema.
In questo caso si procederà con un algoritmo ben collaudato come quello della Maximum Likelihood Estimation per la stima dei parametri di ciascun modello GMM [12].

La GMM likelihood sarà quindi:

$$p(X|\lambda) = \prod_{t=1}^{T} p(\bar{x}_t|\lambda) \qquad t = 1, ..., T \tag{4.25}$$

con T indichiamo il numero di vettori di osservazioni nella fase di training. Tale funzione e non lineare nei suoi parametri, quindi per la massimizzazione della correlazione si passerà attraverso un algoritmo iterativo denominato Expectation Maximization che funziona partendo da un modello iniziale λ si stima un nuovo modello λ^{new} in modo che $p(X|\lambda^{new}) \geq p(X|\lambda)$.
Quindi si procederà fino ad una condizione di stop dell'algoritmo.
Le formule per la stima del nuovo modello sono:

Mixture Weights:

$$p_i^{new} = 1/T \sum_{t=1}^{T} p(i|\bar{x}_t, \lambda) \tag{4.26}$$

Means:

$$\bar{\mu_i}^{new} = \frac{\sum_{t=1}^{T} p(i|\bar{x}_t, \lambda)\bar{x}_t}{\sum_{t=1}^{T} p(i|\bar{x}_t, \lambda)} \tag{4.27}$$

Variances:

$$\sigma_i^{new\,2} = \frac{\sum_{t=1}^{T} p(i|\bar{x}_t, \lambda)x_t^2}{\sum_{t=1}^{T} p(i|\bar{x}_t, \lambda)} - \bar{\mu_i}^2 \tag{4.28}$$

dove σ_i^2, x_t, μ_i si riferiscono a elementi arbitrari dei vettori $\bar{\sigma_i}^2, \bar{x}_t, \bar{\mu}_i$ rispettivamente.
La probabilità a posteriori della i-esima classe è data da:

$$p_(i|\bar{x}_t, \lambda) = \frac{p_i b_i(\bar{x}_t)}{\sum_{k=1}^{M} p_k b_k(\bar{x}_t)} \tag{4.29}$$

I fattori critici da determinare preliminarmente sono l'ordine M del modello e il modello iniziale su cui applicare l'algoritmo di massimizzazione dell'aspettazione.

A questo punto avendo a disposizione i modelli, possiamo applicare in fase di testing, un processo di massimizzazione della probabilità a posteriori

che una sequenza osservata appartenga ad una determinata classe.
Si ha quindi:

$$\hat{S} = arg\,max\ Pr(\lambda_k|X) = arg\,max\ \frac{p(X|\lambda_k)Pr(\lambda_k)}{p(X)} \quad 1 \leq k \leq S \tag{4.30}$$

L'uguaglianza precedente è dovuta alla legge di Bayes.
Assumendo preliminarmente ogni emozione come ugualmente probabile rispetto alle altre, e notando che $p(X)$ è uguale per tutte le classi il problema diventa:

$$arg\,max\,p(X|\lambda_k) \quad 1 \leq k \leq S \tag{4.31}$$

Infine supponendo le osservazioni indipendenti tra loro:

$$\hat{S} = arg\,max \sum_{t=1}^{T} log\,p(\bar{x}_t|\lambda_k) \tag{4.32}$$

Punti critici dell'algoritmo
Come già citato in precedenza, tale approccio al problema vede tra i suoi punti critici in primis:

- Inizializzazione:
  Ciò che emerso da prove ripetute è che l'inizializzazione ossia il modello λ_0 da cui far partire l'algoritmo di massimizzazione dell'aspettazione non influisce sul risultato finale, così come non influisce sul numero di iterazioni per raggiungere lo stesso risultato.

- Ordine del modello:
  In letteratura non esiste un vero e proprio metodo per stabilire a priori l'ordine del modello, in generale si punta ad avere un modello abbastanza robusto, quindi con un numero sufficiente di componenti tale da modellare correttamente i dati e, altresì garantire un costo computazionale comprensibile [12] .

4.6.2 GMM-UBM Universal Background Model

Partendo da quanto visto in precedenza possiamo, a questo punto, pensare ad un approccio che parta dalla modellazione dei dati in un'unica GMM,

un modello ricavato da tutti i dati a disposizione per il training, un modello universale, di qui la dicitura Universal Background Model [30].
In più, partendo da tale modello λ_{UBM}, sarà possibile ricavare i modelli delle singole emozioni adattando, attraverso l'algoritmo Maximum A Posteriori (MAP), tale modello universale ai dati relativi alle singole emozioni.

Maximum A Posteriori algorithm (MAP)

L'algoritmo è simile a quello visto per la massimizzazione dell'aspettazione (EM) nella modellazione del modello universale.
In particolare consideriamo un insieme di vettori afferenti ad una delle classi all'interno del training set, ossia una delle emozioni: $X = \{x_1, x_2, \ldots, x_t, \ldots, x_T\}$.
A questo punto verifichiamo l'alliniamento tra tali vettori di training e le componenti delle mixture del modello universale, i.e. l'i-esima Gaussiana:

$$p_i = 1/T \sum_{t=1}^{T} p(i|\bar{x_t}, \lambda_{ubm}) \tag{4.33}$$

$$\mu_i = \frac{1}{p_i} \sum_{t=1}^{T} Pr(i|x_t, \lambda_{ubm}) x_t$$

$$\sigma_i^2 = \frac{1}{p_i} \sum_{t=1}^{T} Pr(i|x_t, \lambda_{ubm}) x_t^2$$

A questo punto potremo adattare i valori del vettore delle medie di ciascuna componente del modello universale:

$$\hat{w}_i = \alpha_i^w w_i + (1 - \alpha_i^w) w_i \tag{4.34}$$

$$\hat{\mu}_i = \alpha_i^m \mu_i + (1 - \alpha_i^m) \mu_i \tag{4.35}$$

$$\hat{\sigma}_i = \alpha_i^v \sigma_i + (1 - \alpha_i^v)(\sigma_i^2 + \mu_i^2) - \hat{\mu_i}^2 \tag{4.36}$$

$$\alpha_i^w = \alpha_i^m = \alpha_i^v = w_i\, T/(w_i\, T + r) \tag{4.37}$$

dove x_t rappresenta il t-esimo vettore di training appartanente ad una particolare classe di emozioni, r è detto *relevance factor*, normalmente compreso

tra 8 e 20.
Esso rappresenta un parametro che pesa il training set; in particolare se un traning set afferente ad una singola classe è poco popolato è plausibile che il modello ottenuto mediante MAP non si discosterà molto dal modello universale, quindi il relevance factor sarà vicino ai valori minimi; vale naturalmente il viceversa.
Infine, come si pone in evidenza in [30], il solo adattamento del vettore delle medie, affiancato dall'uso di un'unica matrice di covarianza, per ogni modello (global covariance matrix), si raggiunge un trade off ottimale tra prestazioni del modello e tempi di computazione.

4.6.3 Support Vector Machine (SVM)

L'idea di base di questo classificatore è che, avendo a disposizione dei dati in ingresso appartenenti a due classi differenti, essi saranno distinti come segue:
Sia

$$D = \{x_i, y_i\} \qquad i = 1, \cdots, l \qquad y_i \, \epsilon \, \{-1, +1\}$$

Con D ,il dataset a disposizione, x_i saranno i sample del dataset e y_i indichiamo l'etichetta di ciascun sample, se appartenente alla classe -1 o +1.
Quindi stiamo considerando un problema di natura binaria, più avanti sarà illustrato un metodo per estendere tale classificatore ad una etichettatura multiclasse.
Ciò che il metodo prevede, quindi, è di individuare un iperpiano H , tale che tutti i punti apparteneti ad una specifica classe siano disposti dallo stessa parte rispetto al piano e che la distanza tra l'iperpiano e le due classi sia massima.
Definiamo con M il margine, ossia la distanza, tra l'iperpiano individuato e i vattori x_i appartenenti alle due diverse classi. In figura 4.29 è raprresentato un generico iperpiano

Data l'equazione generica di un iperpiano

$$f(x) = \beta_0 + \beta^T x = 0, \quad \beta^T \quad x, \epsilon \mathbf{R}^d, \quad \beta_0 \epsilon \mathbf{R} \tag{4.38}$$

Definiamo la funzione obiettivo[16]:

$$min_{\beta,\beta_0} \quad \frac{1}{2}\|\beta^2\| \tag{4.39}$$

$$Subjected\, to \quad y_i(x_i^T \beta + \beta_0) \geq 1 \qquad coni = 1, \cdots, N \tag{4.40}$$

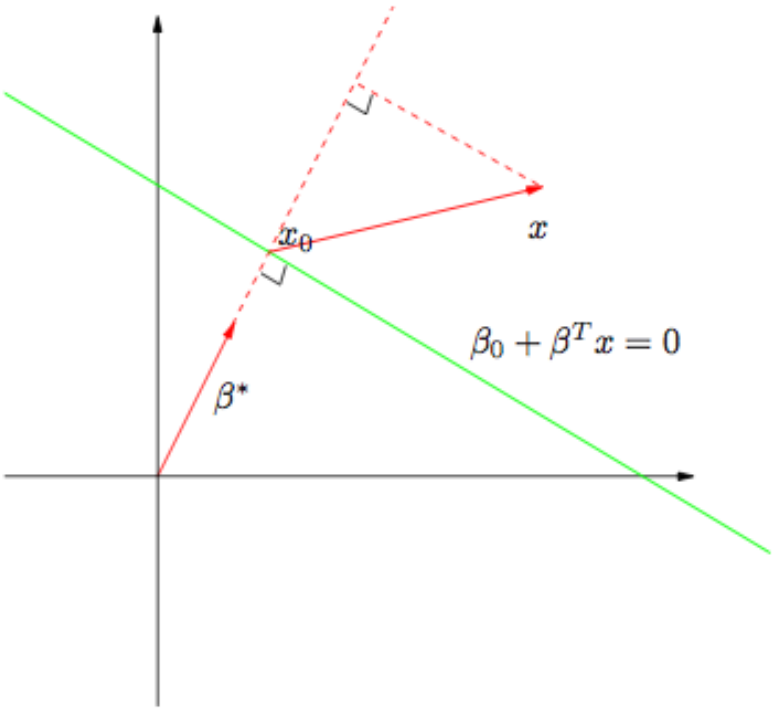


Figura 4.28: Definizione di un iperpiano generico

Nella formulazione raggiunta, il problema è nella forma di uno convesso (funzione obiettivo quadratica, vincoli lineari).
Possiamo quindi, riformularlo, con l'ausilio dei moltiplicatori di Lagrange [14], e quindi la funzione da minimizzare, rispetto a β ,β_0, diventa:

$$L_P = 1/2\beta^2 - \sum_{i=0}^{N} \alpha_i [y_i(x_i^T \beta + \beta_0) - 1] \tag{4.41}$$

Imponendo a zero le derivate:

$$\beta = \sum_{i=1}^{N} \alpha_i y_i x_i \tag{4.42}$$

$$0 = \sum_{i=1}^{N} \alpha_i y_i \tag{4.43}$$

Quindi sostituendo,

$$Maximize\, L_D = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,k=1}^{N} \alpha_i \alpha_k y_i y_k x_i^T x_k \tag{4.44}$$

$$Subjected\, to \sum_{i=1}^{l} \alpha_i y_i = 0 \qquad \alpha_i \geq 0 \tag{4.45}$$

$$\alpha_i [y_i (x_i^T \beta + \beta_0) - 1] = 0 \qquad \forall i \tag{4.46}$$

quindi, ciò che emerge è che se $\alpha_i > 0$, allora $y_i(x_i^T \beta + \beta_0) = 1$, quindi x_i, è un support vector, ossia è un vettore appartenente al margine di classe, o, per meglio dire, appartiene al contorno della classe che definisce, come mostrato in figura 4.30.
Viceversa, se $y_i(x_i^T \beta + \beta_0) > 1$ allora x_i non appartiene al piano di separazione, quindi $\alpha_i = 0$.
A questo punto il classificatore sarà della forma:

$$G(x) = sgn[f(x) = x^T \beta + \beta_0] \tag{4.47}$$

La trattazione fin ora proposta, fa riferimento al caso in cui le classi di dati siano separabili e che le superfici di separazione siano lineari.
Nel caso in esame, così come nella maggior parte delle applicazioni reali, è necessario far riferimento a superfici di separazione non lineari tra i dati e alla non perfetta separabilità delle classi di dati. Si procederà, quindi, con l'applicazione di una funzione che proietti le componenti di ogni sample del dataset su un nuovo spazio multidimensionale, e si andrà a considerare un *soft margin*, ossia, nell'individuare il miglior iperpiano che divide le due classi di dati, si ammette che alcuni sample siano classificati erroneamente, purchè entro certi vincoli imposti, come si evince dalla figura 4.29(b):

il nuovo problema diventa quindi:

$$min_{\beta,\beta_0} \quad \frac{1}{2}\|\beta\|^2 + C \sum_{i=1}^{N} \xi_i \tag{4.48}$$

$$Subjected\, to \quad \xi_i \geq 0, y_i(\phi(x_i^T)\beta + \beta_0) \geq 1 - \xi_i \qquad \forall i \tag{4.49}$$

(ξ_i rappresenta il soft margin, ossia il range di tolleranza entro cui un sample può essere classificato erroneamente, come mostrato in figura 4.29.
C è un parametro di costo, e nel caso di di perfetta separabilità, $C = \infty$ C

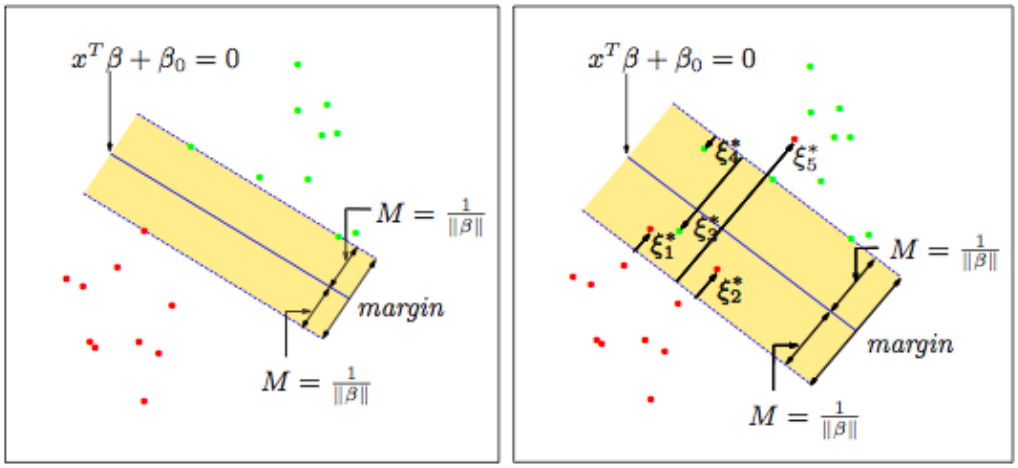


Figura 4.29: Definizione di margin, nel modello iniziale (a) e soft margin nel modello finale (b)

rappresenta un parametro settabile e l'approccio migliore per la sua scelta è l'applicazione di un cross folding che permetta di stimare quando le performance di classificazione siano migliori[10].

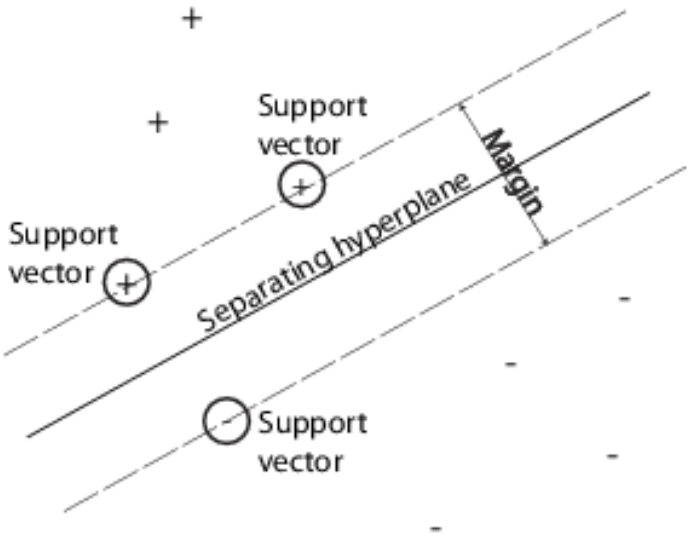


Figura 4.30: Definizione di Support Vector

A questo punto la trattazione è analoga a quella già illustrata, con la

differenza ulteriore che i vettori in ingresso saranno trasformati rispetto ad una funzione kernel, quindi l'equazione generica dell'iperpiano sarà:

$$f(x) = \beta \langle \phi(x^{(i)}), \phi(x^{(k)}) \rangle + \beta_0 = 0 \tag{4.50}$$

nella fattispecie sarà una funzione polinomiale del terzo grado nella forma:

$$K(x^{(i)}, x^{(k)}) = (\sum_{i=1}^{N} x^{(i)} x^{(k)} + c)^3,$$

con $x^{(i)}, x^{(k)}$ vettori nello spazio iniziale.
il problema risulta quindi nella sua formulzione finale come:

$$Maximize \quad L_D = \sum_{i=1}^{N} \alpha_i - 1/2 \sum_{i,k=1)}^{N} \alpha_i \alpha_k y_i y_k \langle \phi(x^{(i)}), \phi(x^{(k)}) \rangle \tag{4.51}$$

$$Subjected\, to: \quad \sum_{i=1}^{l} \alpha_i y_i = 0 \tag{4.52}$$

$$0 < \alpha_i < C \tag{4.53}$$

$$\alpha_i [y_i (x_i^T \beta + \beta_0) - 1] = 0 \qquad \forall i \tag{4.54}$$

KKT dual complementary conditions

L'algoritmo utilizzato per la risoluzione di questo problema di programmazione è il *Sequential Minimal Optimization* (SMO) [24].

Multiclass SVM

Così come preannunciato, il passo successivo è stato rappresentato dall'adattamento del classificatore binario SVM in un classificatore multiclasse.
A questo proposito, la soluzione è stata trovata in un approccio One vs One, ossia sono stati confrontati a coppie tutti i classificatori e con una logica di *major voting*, è stata decisa la classe di apparteneza del test sample in esame. Questo metodo è sicuramente il più preciso ma altresì computazionalmente dispendioso, infatti sarà necessario effettuare $n/2(n-1)$ classificazioni, per training sample nel dataset, dove con n indichiamo il numero di classi.

Capitolo 5

Architettura del sistema

5.1 Introduzione

Per la realizzazione del sistema, partendo dal modello descritto nel capitolo precedente, si è dapprima scelto l'ambiente di sviluppo, con il fine di avere un'applicazione che fosse usabile, e che fornisse sufficienti strumenti di analisi, anche durante la fase di sviluppo.
Inoltre è necessario sottolineare la natura dell'applicazione messa a punto, prettamente sperimentale, quindi, è stato preferito un ambiente di sviluppo già calcato dalla ricerca in questa direzione.
Infatti l'utilizzo di software quali Matlab, Praat, hanno permesso di sperimentare una grande varietà di combinazioni dei vari moduli di sistema, permettendo un confronto in tempo reale delle diverse configurazioni.
Inoltre non è da sottovalutare anche il supporto di cui queste applicazioni godono, quindi la possibilità di confrontare il prorpio lavoro con una grande comunità.
Infine si è scelto di sviluppare il framework dell'applicazione attraverso l'uso dell'IDE Eclipse, data la sua natura multipiattaforma e di usare Java come linguaggio di sviluppo, questo poichè tutti gli altri strumenti di sviluppo prevedono librerie di funzioni (API) che hanno permesso una facile interconnessione tra i moduli del sistema. Per questi motivi si è proceduto come segue.

5.2 Interfaccia Utente (UI)

L'idea di fondo di approccio all'utilizzo del software è molto semplice.
L'utente decide se effettuare un nuovo training del sistema, con un nuovo dataset, che fornirà al sistema come path alla cartella contenente i sample da analizzare, oppure effettuare un test su un dataset, con un modello precedentemente messo a punto.
Tale interfaccia è stata realizzata con la messa a punto di un semplice menù da riga di comando, una volta lanciata l'applicazione.

5.3 Modulo Analisi Audio

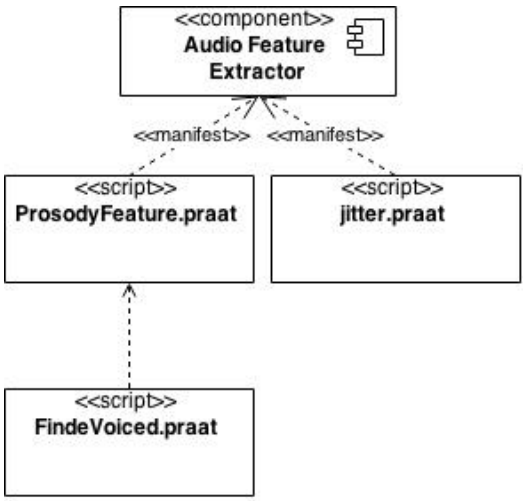


Figura 5.1: Dettaglio del Deployement Diagram del modulo di analisi audio

Si è partiti con l'acquisizione dei dati da analizzare, in forma di file audio codificati in wave a 44100 Hz, 16 bit.
Dapprima, si è partiti con l'analisi del file audio, quindi con il preprocessing, che ha visto la rimozione dei silenzi, iniziali e finali, poi la normalizzazione dell'intensità, un filtraggio passabasso con frequenza di taglio $f_{cut} = 5000$ Hz e sottocampionamento di un fattore 4.
Tale preprocessing è stato realizzato con l'ausilio di uno script messo a punto con PRAAT. Tale script viene lanciato attraverso un comando di shell, in

background.

A questo punto il file audio è stato ridotto alle uniche parti *voiced*, ossia si sono estratte le uniche parti di parlato, escludendo qualsiasi forma di pausa. Anche questa funzione è stata realizzata con PRAAT: in particolare si divide il file audio in frame di 40 ms, e se ne studia, attraverso un algoritmo di *pitch tracking* (capitolo 3), la componente armonica e, se presente, si etichetterà tale frame come voiced.
In questa fase dell'analisi, non si è interessati ad una stima accurata del pitch, quindi la soglia di decisione imposta per distinguere tra voiced/unvoiced frame, resta comunque alta, in modo da eliminare quella quotaparte di segnale che potrebbe rivelarsi degradante nelle analisi successive.

Il segnale contenente gli unici voiced frame, è stato sottoposto all'analisi delle Continuous Acoustic Feature e Speaker Feature.
Il primo, rispettivamente, messo a punto con PRAAT (capitolo 3) e il secondo con l'ausilio di Matlab e, in particolare, della libreria Voicebox, che riassume in un' unica funzione, tutte le singole operazioni, sfruttando funzioni native di Matlab.
In questo caso è stata utilizzata la libreria *matlabcontrol 4.0.1* per poter lanciare in background Matlab, e gli script ad esso associati.

Parallelamente, il file audio comprensivo delle pause è stato sottoposto all'analisi delle Voice Quality Feature, estratte attraverso un altro script di PRAAT, questo a causa della natura delle feature implicate in tale categoria, che prevedono proprio l'analisi delle pause e della loro natura.

5.4 Modulo di analisi testuale

Accanto al file audio, la sua trascrizione letterale, fornita come file di testo, sarà sottoposta, dapprima all'analisi del POS tagger. Tale componente ha lo scopo di fornire quali siano i ruoli grammaticali di ciascuna parola all'interno di una espressione. Tale sistema è stato messo a punto facendo leva su un componente già esistente per l'italiano, Freeling 3.0.
Quindi con un semplice script, si analizza l'espressione in forma testuale, ottenendo i *lemma* di ciascun vocabolo presente e il relativo POS.

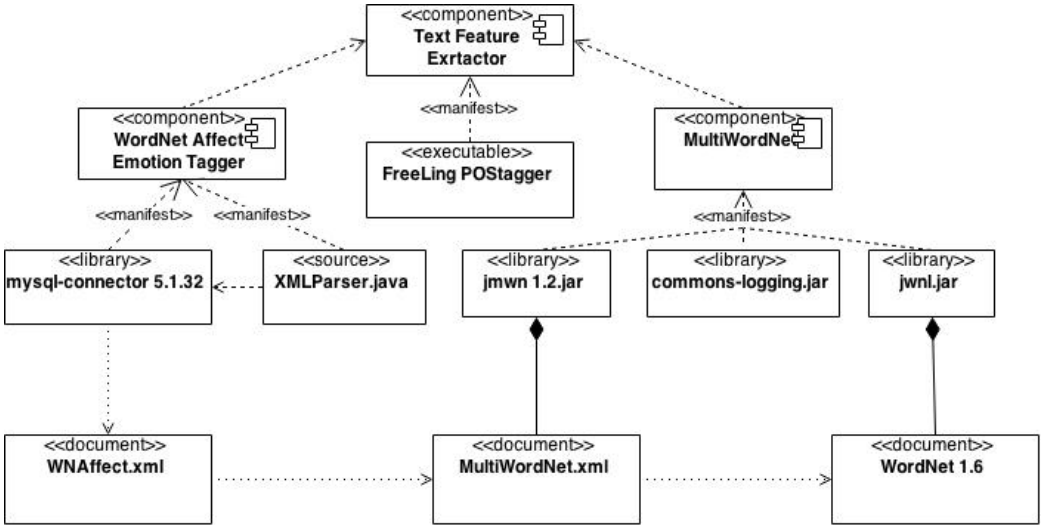


Figura 5.2: Dettaglio del Deployement Diagram del modulo di analisi testuale

A questo punto i singoli lemma così estratti, sono l'input della funzione di ricerca del contenuto emozionale di ciascun vocabolo.
Per la realizzazione di tale componente è stato necessario individuare una soluzione per superare il problema della lingua italiana, non supportata dalla libreria WordNet Affect, contenente i vocaboli etichettati emotivamente.
Quindi, attraverso l'utilizzo della libreria messa a disposizione da MultiWordNet, *MWN 1.5*, è stato possibile interfacciarsi con il corpus di WordNet Affect.
In particolare il vocabolo in analisi, sarà dapprima ricercato all'interno della base di dati fornita da MultiWordNet, che funzionerà da funzione di mappatura per il corpus WordNet 1.6 di Princeton, quindi sarà associata una stringa identificativa a tale vocabolo, del quale si cercherà, attraverso un XML parser, il contenuto emozionale, all'interno della base di dati WordNet Affect, visto che il sistema di identificazione dei vocaboli è lo stesso tra quest'ultimo e WordNet.
Avendo raccolto queste informazioni è stato possibile sintetizzare il vettore di feature testuali per ciascuna espressione.

5.5 Feature Wrapper

Tale componente ha come scopo, quello di raccogliere e sintetizzare in un'unica struttura (feature vector), le feature che rappresentano ciascuna espressione.
Successivamente, struttura i feature vector così ottenuti, in un dataset, compatibile con il modulo di classificazione messo a punto.

5.6 Classificatore SVM

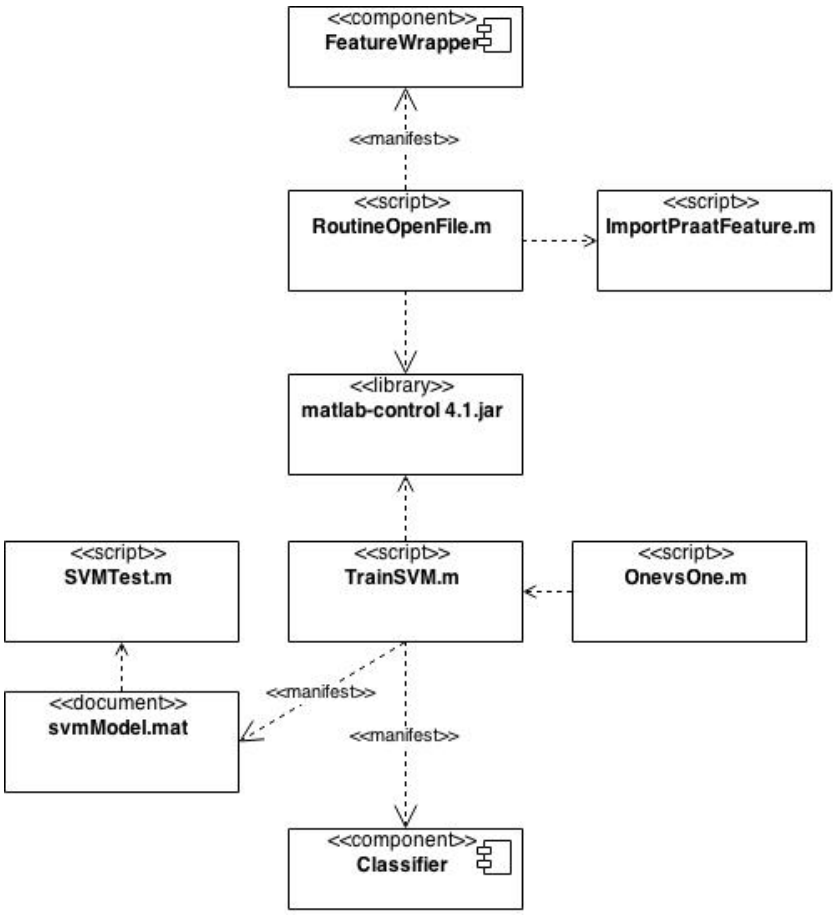


Figura 5.3: Dettaglio del Deployement Diagram del classificatore

Tale componente, è stato realizzato in Matlab.
In particolare, a seguito del compito svolto dal Feature Wrapper, il sistema

di classificazione ha il compito di stimare i quattro modelli emozionionali.
Il componente è stato realizzato sfruttando l'approccio One Vs One, in cui vengono dapprima stimati i quattro modelli emozionali, per poi eseguire il testing sul 20% del dataset, con un cross folding validation di cinquanta iterazioni (capitolo 3).
A questo punto implementando una tecnica di major voting, si individua quale sia la classe di appartenenza dei sample di test.
Per ciascun test effettuato si sono visualizzati i risutati relativi all'accuratezza e ai tempi di computazione di ciascuna iterazione del cross folding validation, fornendo le Confusion Matrix relative ai casi migliori e peggiori, che pongono in evidenza quanti siano i sample correttamente classificati (entry con stesso indice) e quali classificati erroneamente (entry con indice differente).
Le funzioni sono implementate partendo dalla libreria nativa di Matlab.

Capitolo 6

Realizzazioni sperimentali e valutazione

6.1 Introduzione

Nel sezione seguente si andranno ad illustrare le configurazioni messe a punto a seguito della fase di ricerca delle tecniche per la realizzazione dei componenti del sistema a vari livelli, rispetto al punto, nel sistema, di elaborazione dei dati.
In particolare le configurazioni di sistema che sono state messe a punto sono due:

- Audio Feature con UBM-GMM
- Sistema Multiclass-SVM, Text+Audio Feature

6.2 Audio Feature con UBM-GMM

La prima configurazione, quella preliminare, su cui si è puntata l'attenzione nella fase iniziale della ricerca delle soluzioni possibili e si è partiti da risultati già noti, per garantirsi una base di partenza[2][9].
In figura 6.1 è rappresentato il modello dei blocchi funzionali, che vedono i segnali audio in ingresso e, in uscita, gli stessi, ma etichettati. In più il sistema fornisce i modelli delle quattro emozioni, partendo dall'adattamento di quello

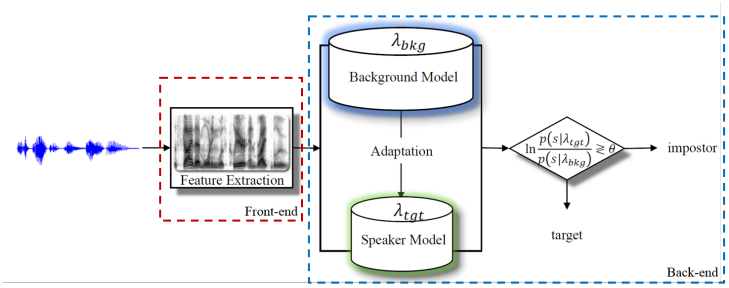


Figura 6.1: Schema sistema GMM-UBM

universale, risultato del training su tutto il dataset (vedi dopo). La scelta ha visto la considerazione di un fattore determinante per la progettazione del sistema: esso è rappresentato dalla *dipendenza dal tempo del sistema*, o, per meglio dire, il sistema si basa sull'analisi di espressioni di durata differente, quindi considerare feature legate al tempo avrebbero reso i feature vector di lunghezza diversa.
Per superare questa difficoltà si è deciso di modificare l'approccio al problema. In particolare si è proceduto con l'estrazione delle feature audio, basate su una *frame-based analysis*, quindi, ciscuna espressione è sintetizzata come una matrice $n \times m$ dove n è il numero di frame, ed m è il numero di feature estratte.
Questo mette in evidenza che, n è strettamente legato alla durata dell'espressione in analisi.
Il feature set è composto dalle sole *variabili acustiche continue*:

- Pitch (F_0) e derivata
- Formant I/II/III e derivate
- Harmonic To Noise Ratio (HNR)

Le tecniche di estrazione delle feature sono le stesse descritte nel capitolo 4, con la differenza che in questo caso, non si riassumono le feature in valori di media e varianza, piuttosto, si considera ciascun frame analizzato come una unità atomica da classificare.
Questo giustifica poichè il *sistema resta comunque indipendente dalla durata delle espressioni*, ció che conta sono i valori delle feature che caratterizzano i frame stessi nella fase di training - dove ciascun frame sarà già etichettato,

quindi il sistema estrae una correlazione tra frame appartenenti alla stessa classe emozionale -. A questo punto, i dati raccolti saranno riordinati dal feature wrapper, che preparerà il dataset per il classificatore GMM-UBM di cui (Capitolo 4).
In questo caso il modello GMM-UBM è stato implementato e configurato come segue: il modello universale (UBM) è stato sintetizzato partendo dal training di tutto il dataset, di un modello GMM a 256 core.
A questo punto si sono sintetizzati i modelli delle quattro emozioni, partendo dall'adattamento del modello universale ai training set delle singole emozioni, attraverso l'algoritmo MAP descritto precedentemente nel Capitolo 4.

I dati rilevati, sono sintetizzati nelle figure 6.2 (a) e (b):

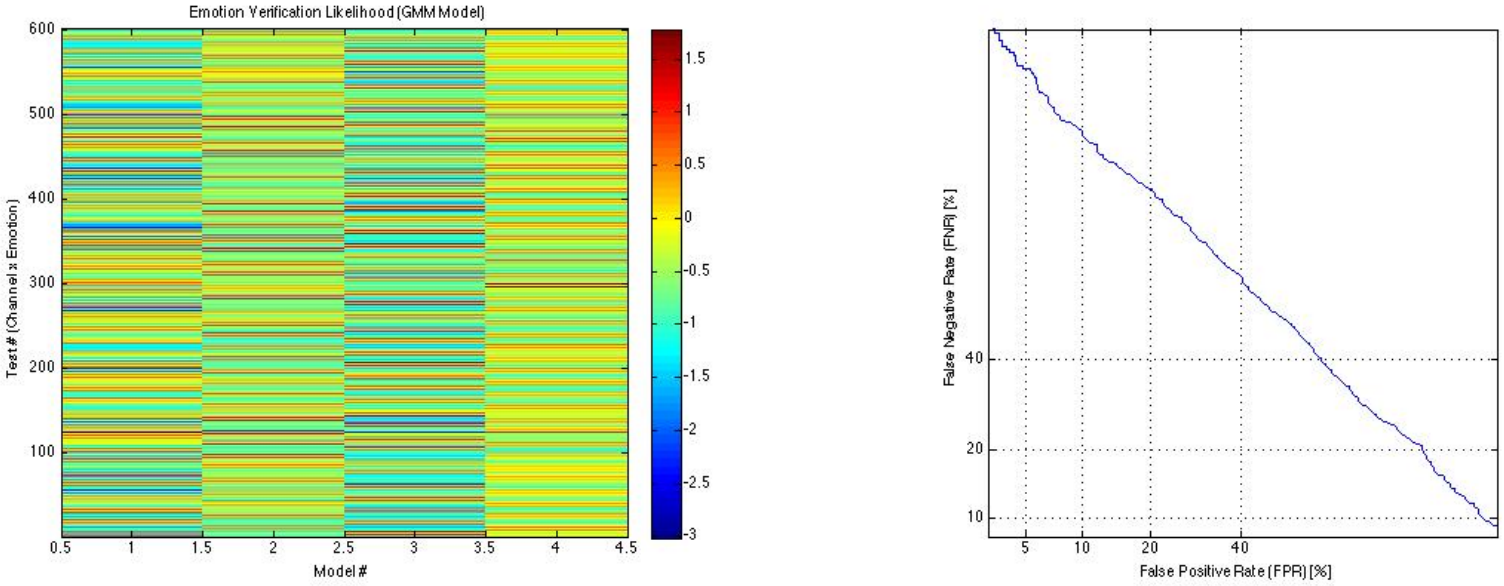


Figura 6.2: Performance GMM-UBM: Confusion Matrix (a), False Positive vs. False negative rate (b)

Initializing the GMM hyperparameters ...

Re-estimating the GMM hyperparameters for 1 components ...

EM iter#: 1 [llk = -34.02] [elaps = 0.29 s]

Re-estimating the GMM hyperparameters for 2 components ...

EM iter#: 1 [llk = -34.15] [elaps = 0.27 s]
EM iter#: 2 [llk = -33.72] [elaps = 0.26 s]

Re-estimating the GMM hyperparameters for 4 components ...

EM iter#: 1 [llk = -33.72] [elaps = 0.28 s]
EM iter#: 2 [llk = -32.79] [elaps = 0.31 s]
EM iter#: 3 [llk = -32.15] [elaps = 0.28 s]
EM iter#: 4 [llk = -31.86] [elaps = 0.28 s]

Re-estimating the GMM hyperparameters for 8 components ...

EM iter#: 1 [llk = -31.82] [elaps = 0.35 s]
EM iter#: 2 [llk = -31.16] [elaps = 0.34 s]
EM iter#: 3 [llk = -30.90] [elaps = 0.34 s]
EM iter#: 4 [llk = -30.79] [elaps = 0.33 s]

Re-estimating the GMM hyperparameters for 16 components ...

EM iter#: 1 [llk = -30.78] [elaps = 0.42 s]
EM iter#: 2 [llk = -30.40] [elaps = 0.45 s]
EM iter#: 3 [llk = -30.16] [elaps = 0.43 s]
EM iter#: 4 [llk = -30.05] [elaps = 0.43 s]

Re-estimating the GMM hyperparameters for 32 components ...

EM iter#: 1 [llk = -30.05] [elaps = 0.59 s]
EM iter#: 2 [llk = -29.75] [elaps = 0.58 s]
EM iter#: 3 [llk = -29.54] [elaps = 0.58 s]
EM iter#: 4 [llk = -29.33] [elaps = 0.58 s]

Re-estimating the GMM hyperparameters for 64 components ...

EM iter#: 1 [llk = -29.21] [elaps = 0.90 s]
EM iter#: 2 [llk = -28.89] [elaps = 0.87 s]
EM iter#: 3 [llk = -28.69] [elaps = 0.88 s]
EM iter#: 4 [llk = -28.54] [elaps = 0.89 s]
EM iter#: 5 [llk = -28.41] [elaps = 0.88 s]

EM iter#: 6 [llk = -28.31] [elaps = 0.89 s]

Re-estimating the GMM hyperparameters for 128 components ...

EM iter#: 1 [llk = -28.31] [elaps = 1.50 s]
EM iter#: 2 [llk = -28.04] [elaps = 1.48 s]
EM iter#: 3 [llk = -27.89] [elaps = 1.47 s]
EM iter#: 4 [llk = -27.77] [elaps = 1.48 s]
EM iter#: 5 [llk = -27.67] [elaps = 1.48 s]
EM iter#: 6 [llk = -27.59] [elaps = 1.48 s]

Re-estimating the GMM hyperparameters for 256 components ...

EM iter#: 1 [llk = -27.59] [elaps = 2.74 s]
EM iter#: 2 [llk = -27.31] [elaps = 2.72 s]
EM iter#: 3 [llk = -27.14] [elaps = 2.70 s]
EM iter#: 4 [llk = -27.00] [elaps = 2.70 s]
EM iter#: 5 [llk = -26.90] [elaps = 2.69 s]
EM iter#: 6 [llk = -26.82] [elaps = 2.67 s]
EM iter#: 7 [llk = -26.75] [elaps = 2.68 s]
EM iter#: 8 [llk = -26.70] [elaps = 2.73 s]
EM iter#: 9 [llk = -26.66] [elaps = 2.70 s]
EM iter#: 10 [llk = -26.62] [elaps = 2.73 s]

Come si evince dalle figure 6.2 *(a)* e *(b)*, il sistema risulta altamente impreciso, ossia non è in grado di discriminare tra le quattro classi che rappresentano le emozioni.
Inoltre è evidente come il sistema, necessiti di tempi di training molto lunghi, nonostante il dataset non sia enorme. In particolare, la confusion matrix mette in evidenza quanto l'errore sia diffuso lungo tutto il dataset, cosa che fa presuppore un approccio errato al problema.

6.3 Text+Audio Feature con Multiclass-SVM

Partendo da questi risultati, si è proceduto con la messa a punto di un nuovo modello per la raccolta e classificazione dei dati.

Il sistema ha visto le sue modifiche sia nel feature set, che nel modello di classificazione, infatti si sono prese in considerazione le feature descritte nel modello (Capitolo 3); A questo punto si puó procedere con la classificazione dei dati, attraverso un classificatore Support Vector Machine, multiclasse.
Il dataset è stato partizionato in proprorzioni di 75% per il Training e 25% per il Testing, per poi quindi applicare un 50 cross-fold validation, ossia si è ripetuta la fase di Training del sistema per 50 volte, in modo che le parti di dati utilizzati per la fase di Training e di Testing, mantenendo invariate le proporzioni, fossero casualmente scelte all'interno del dataset, questo, per evitare il fenomeno dell'overfitting e per stimare un' accuratezza media, su un numero ragionevole di iterazoni.
Nel prossimo paragrafo andremo ad illustrare le varie confgurazioni di sistema e saranno mostrati i risultati ottenuti in termini di accuratezza nella distinzione tra le quattro classi, questo per poter valutare quali siano i punti sensibili del sistema, quindi la direzione su cui insistere con la ricerca.

Configurazione 1

Dataset : completo (720 sample)

Feature set : completo (110 feature)

Numero emozioni : 4 (Gioia, Rabbia, Paura,Tristezza)

Training/Testing : 50 cross-folding, 75% Train, 25% Test

Pre clustering : no

PCA : no

Classificatore : SVM Multiclass

Overall Avarage Accurancy = 88.03%
Overall Avarage CPU-time for classification = 36.14 ms

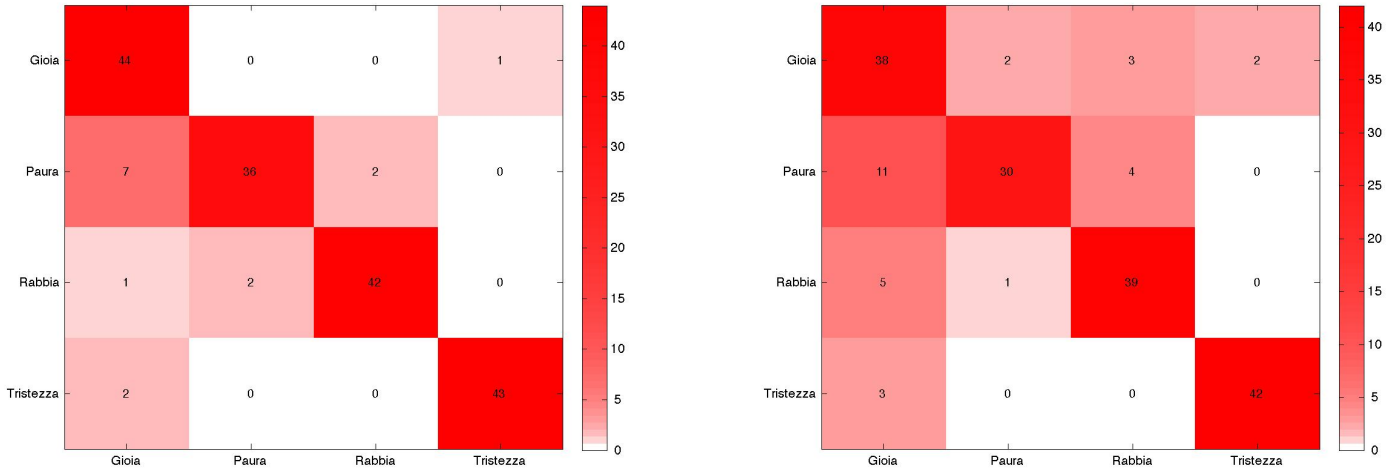


Figura 6.3: Confusion Matrix caso migliore (a), Confusion Matrix del caso peggiore (b) Configurazione 1

Configurazione 2

Dataset : completo (720 sample)

Feature set : 80

Numero emozioni : 4 (Gioia, Rabbia, Paura,Tristezza)

Training/Testing : 50 cross-folding, 75% Train, 25% Test

Pre clustering : no

PCA : si, si ridimensiona il sistema a 80 dimensioni

Classificatore : SVM Multiclass

Overall Avarage Accurancy = 81.38%
Overall Avarage CPU-time for classification = 33.21 ms

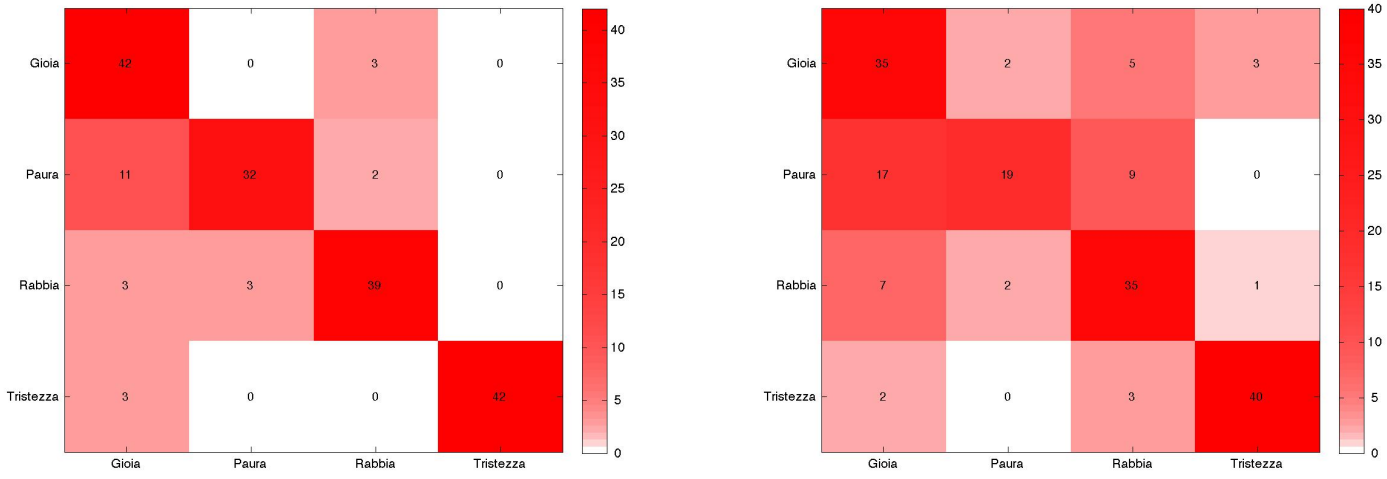


Figura 6.4: Confusion Matrix caso migliore (a), Confusion Matrix caso peggiore (b), Configurazione 2

Configurazione 3

Dataset : 648 sample, a seguito del pre clustering

Feature set : completo (110 feature)

Numero emozioni : 4 (Gioia, Rabbia, Paura,Tristezza)

Training/Testing : 50 cross-folding, 75% Train, 25% Test

Pre clustering : si, Threshold=Distanza media centroide+40%

PCA : no

Classificatore : SVM Multiclass

Overall Avarage Accurancy = 91.47%
Overall Avarage CPU-time for classification = 35.07 ms

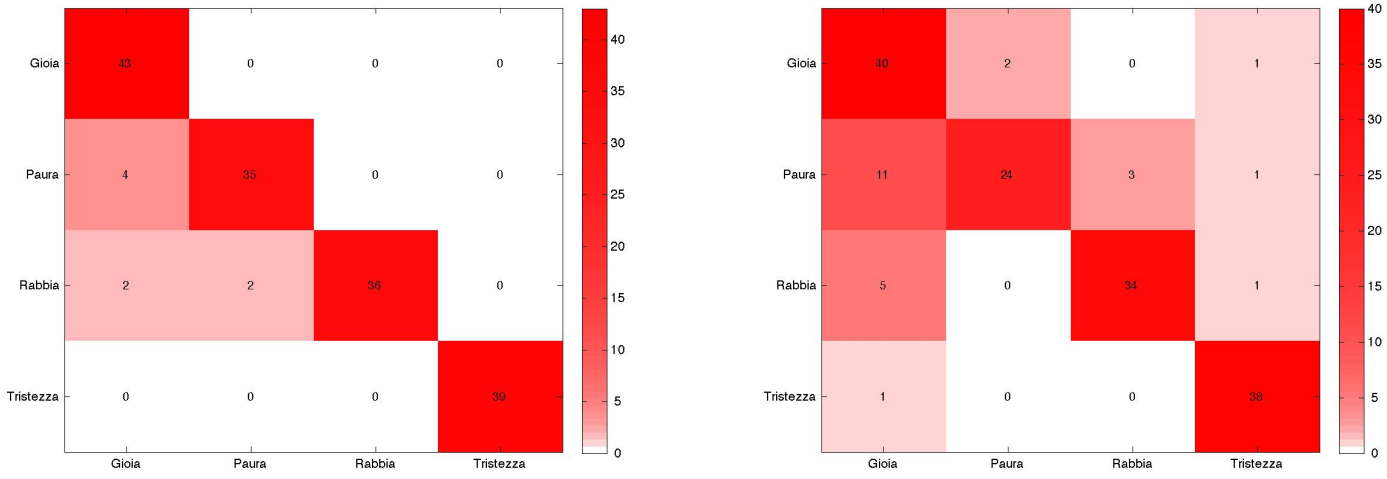


Figura 6.5: Confusion Matrix caso migliore (a), Confusion Matrix caso peggiore (b) Configurazione 3

Configurazione 4

Dataset : 583 sample, a seguito del pre clustering

Feature set : 80 feature

Numero emozioni : 4 (Gioia, Rabbia, Paura,Tristezza)

Training/Testing : 50 cross-folding, 75% Train, 25% Test

Pre clustering : si, Threshold=Distanza media centroide+40%
avarageDistance=distanza media dal centroide

PCA : si, si ridimensiona il sistema a 80 dimensioni

Classificatore : SVM Multiclass

Overall Avarage Accurancy = 84.60%
Overall Avarage CPU-time for classification = 31.93 ms

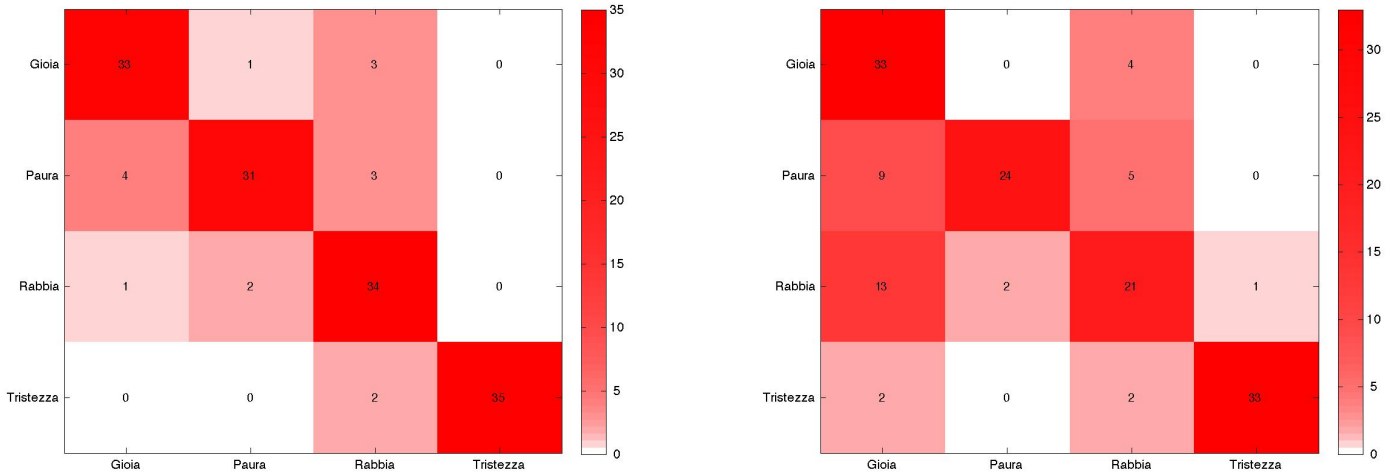


Figura 6.6: Confusion Matrix caso migliore (a), Confusion Matrix caso peggiore (b) Configurazione 4

A seguito dei risultati ottenuti, è stato possibile valutare, in via preliminare, quali fossero le prestazioni del sistema.
Ció che è emerso, a fronte di valori di accuratezza già incoraggianti, è che l'errore rispetto alla discriminazione delle quattro emozioni riguarda principalemente la classe legata alla Gioia, ossia il sistema risulta essere poco efficace nel distinguere tale classe rispetto alle altre.
Ció si evince dall'analisi delle confusion matrix.
Infatti possiamo notare dalle figure 6.3,6.4,6.5,6.6, come, nel caso delle altre tre classi di emozioni, il numero di sample classificati erroneamente sia prossimo a zero.
Come ulteriore test per la valutazione delle prestazioni del sistema si è considerata la configurazione che prevede un feature set costituito dalle sole feature audio, in modo da poter verificare quanto sia incidente il modulo di analisi testuale ai fini dell'accuratezza dell'intero sistema.

Configurazione 5

Dataset : 720 sample, a seguito del pre clustering

Feature set : 102 feature, solo audio

Numero emozioni : 4 (Gioia, Rabbia, Paura,Tristezza)

Training/Testing : 50 cross-folding, 75% Train, 25% Test

Pre clustering : no

PCA : no

Classificatore : SVM Multiclass

Overall Avarage Accurancy = 87.39%
Overall Avarage CPU-time for classification = 39 ms

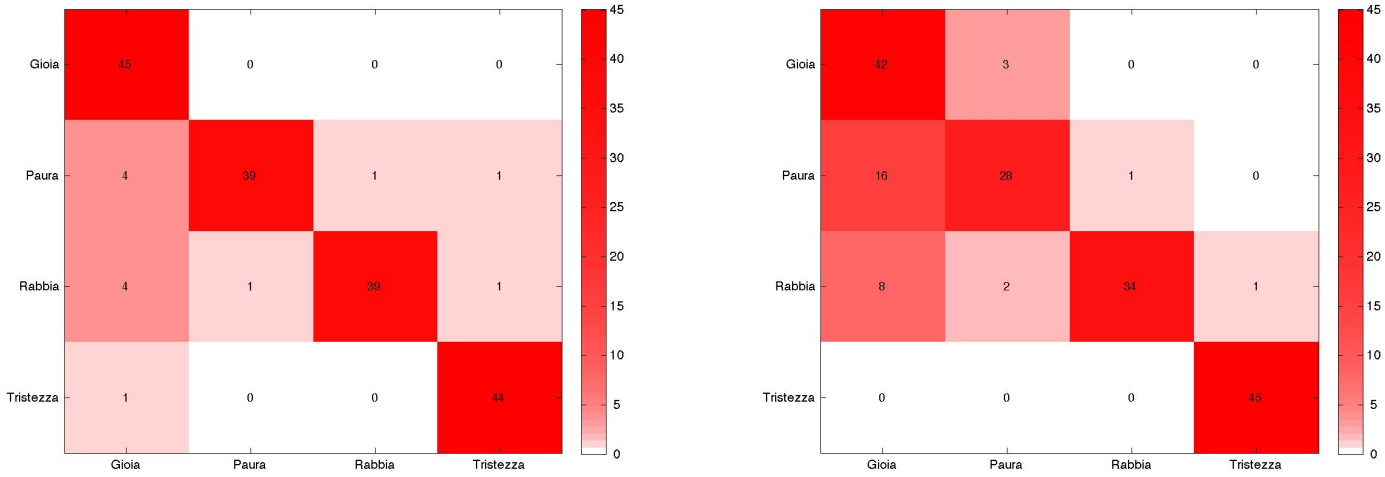
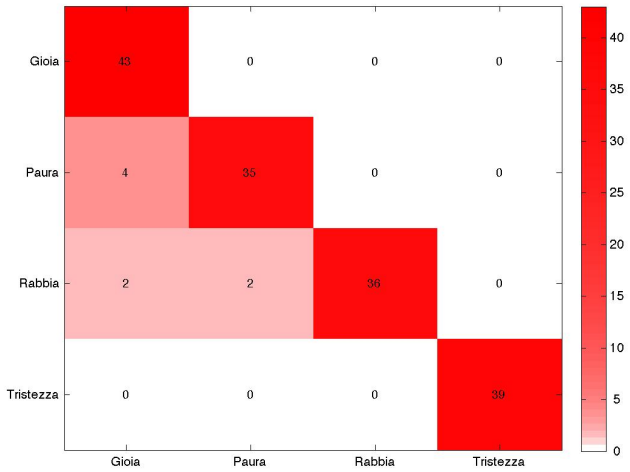


Figura 6.7: Confusion Matrix caso migliore (a), Confusion Matrix caso peggiore (b) Configurazione 5

Come si evince dalla figura 6.7 l'accuratezza mostrata dal sistema se pur inferiore, resta di gran lunga accettabile. Questo suggerisce che l'approccio adottato non incide in modo determinante sulle prestazioni del sistema, ma

comunque fa presupporre che una ricerca verso questa direzioni possa via via migliorare le prestazioni globali.

Si vogliono, infine, porre a confronto le prestazioni del sistema messo a punto, nella sua configurazione più performante, i.e. *configurazione 3*, in questo lavoro, rispetto al sistema messo a punto in [26], in particolare si prende in considerazione il solo modulo di analisi dello stato emozionale implementato in PrEmA. Questo rappresenta, rispetto al confronto con altri sistemi descritti, una figura di merito particolarmente esplicativa, poichè *entrambi i sistemi condividono lo stesso dataset*. Questo mostrerà quanto la scelta di un feature set e di tecniche di classificazione diverse, possa essere determinante rispetto alle prestazioni globali. La differenza rispetto ai dati analizzati é rappresentata dal numero di classi emozionali. In PrEmA si considera una quinta classe, rappresentata dalla *neutralitá*.



	Gioia	neutro	paura	rabbia	tristezza
gioia	63.81%	0.00%	**18.35%**	**11.79%**	6.05%
neutro	3.47%	77.51%	2.14%	1.79%	**15.09%**
paura	**33.75%**	0.00%	58.35%	6.65%	1.25%
rabbia	10.24%	1.16%	8.16%	77.28%	3.16%
tristezza	5.14%	14.44%	0.28%	0.81%	79.33%

Figura 6.8: Confusion Matrix sistema Speech2Emotion (a), Confusion Matrix sistema Preema [20] (b)

Ciò che si evince dal confronto tra i due sistemi é il significativo miglioramento, in termini di accuratezza, nella classificazione dei sample presenti nel dataset. La classe emozionale ulteriore considerata in PrEmA, viene ben modellata dal sistema, mentre, come si evince dalla figura 6.8 (b), i dati rispetto alle altre classi presentano percentuali di dati classificati erroneamente abbastanza alte. Il fatto che il dataset sia lo stesso, permette, effettivamen-

te, di concludere che si sta percorrendo la giusta direzione rispetto ad una risoluzione via via più accurata di una problematica del genere.

Capitolo 7

Conclusioni

7.1 Conclusioni

Al termine del lavoro svolto possiamo trarre alcune conclusioni in merito al sistema messo a punto.
Ciò che è emerso è che la tipologia del problema richiede sicuramente un incremento della ricerca, a prescindere dal sistema stesso, ossia è di evidente necessità esplorare quali siano i meccanismi che generano una emozione, quella fase in cui l'emozione è ancora un'idea, piú che una espressione verbale, in modo da poter via via affinare il modello che al meglio descrive le emozioni stesse, la loro manifestazione, quindi la loro natura.
Per quanto attiene il sistema possiamo analizzarne, in conclusione, le sue prestazioni e dare un significato ai valori ottenuti.
In particolare si andranno ad analizzare, seguendo il flusso di informazioni all'interno del sistema stesso, i comportamenti dei vari moduli.

In primis è da considerare il *dataset*, quindi i dati a disposizione, da cui poter estrarre le informazioni utili al training dei modelli emozionali.
La natura del dataset, costituita da espressioni diversificate nelle modalitá espressive, ma altresí limitato: ad esempio, sono assenti campioni di espressioni con voce femminile o di bambini.
Il numero in se è altrettanto limitativo, oltre che causa di overfitting.
Il modulo di analisi audio, invece, risulta essere efficiente ed efficace nel suo compito.
Infatti le prestazioni del sistema dipendono, principalmente, dalla qualitá

dell'informazione estratta.
Cruciale è risultato l'utilizzo di algoritmi precisi, quali quelli implementati in Praat e Matlab.

Un'altra considerazione da farsi è relativa alla scelta del feature set. A seguito dell'analisi attraverso la Principal Component Analysis e del test di Anova è evidente quanto il ridimensionamento sia relativo ad un numero consistente, ma non determinante rispetto alla scelta di un cambiamento consistente del feature set. Questo suggerisce che la scelta del feature set è consona alla problematica analizzata.

Per quanto riguarda il modulo di analsi testuale e la conseguente implementazione delle feature testuali, all'interno dei feature vector, possiamo concludere che l'uso delle trascrizioni delle espressioni non è ancora determinante per il riconoscimento di una emozione.

In particolare, è doveroso ricordare che tutto il sistema si è basato sull'analisi di espressioni in lingua italiana e questo è risultato un limite, rispetto all'utilizzo di corpora, piú consistenti e dettagliati, in altre lingue.
Rispetto al sistema in sè, alla sua implementazione, esso risulta essere un punto di partenza.

Il classificatore invece è stato sicuramente un punto critico e la sua implementazione con algoritmo SVM multiclasse è risultato un buon compromesso tra accuratezza e costi computazionali.
In generale si sono analizzate le prestazioni del classificatore in base alle diverse configurazioni preliminari, come la dimensione dello spazio delle feature, cosí come il preclustering dei dataset relativi alle singole emozioni.
Ció che è emerso è che il ridimensionamento dello spazio delle feature fornisce un contributo considerevole nell'abbattimento del costo computazionale, andando, peró, a compromettere sensibilmente anche l'accuratezza del sistema stesso.

Questo suggerisce che l'applicazione, anche avendo in ingresso un dataset in cui sia riassunto il 99.2% del contenuto informativo del dataset iniziale (a seguito della PCA), resta vincolato alla quota parte di informazione mancante, piú che per il contenuto informativo, che risulta trascurabile, per un

principio di overfitting.

Il preclustering, invece, nella configurazione 3, mostra come un ridimensionamento, intelligente, del dataset, predisponga i dati ad una classificazione ottimale, in termini di accuratezza e costo computazionale.

7.2 Lavori futuri

Al termine del lavoro svolto, è stato possibile valutare i punti critici a cui si è dovuto far fronte.
Ció che è emersa è la necessitá di corpora di vocaboli, quanto piú ampi possibile, al fine di poter, effettivamente ed efficacemente, implementare sistemi di analisi testuale per l'estrazione di informazione piú complesse, e ad alto livello.
Inoltre, un passo ulteriore è necessario verso una definizione via via piú specifica di un feature set idoneo ai modelli che ci si propone di sintetizzare, quindi spingere la ricerca verso la conoscenza dei meccanismi, biologici, psicologici e fisici, implicati nella generazione di una emozione.

Naturalmente, questo lavoro deve essere visto come la componente di un possibile sistema multi-biometrico, e che quindi rappresenta un punto di vista rispetto a sistemi piú generalizzati, che tengano conto di altre caratteristiche. Pensiamo, ad esempio, all'affiancamento di un'analisi delle espressioni facciali, legata a quelle vocali, rispetto alla discriminazione tra modelli emozionali.

Risulta evidente, infine, quanto un approccio di questo genere presupponga un'accuratezza, oltre che un margine di scalabilitá del sistema, molto piú estesa.

Bibliografia

[1] Newcastle, Northern Ireland, UK.

[2] Noam Amir. Classifying emotions in speech: a comparison of methods, 2001.

[3] Jens Allwood Asa Abelin. Cross linguistic interpretation of emotional prosody, 2000.

[4] James R. Averill. A constructivist view of emotion, 1980.

[5] Joe-Anne Bachorowski. Vocal expression and perception of emotion, 1999.

[6] Gerhard Rigoll Bjorn Schuller, Manfred Lang. Automatic emotion recognition by the speech signal. In *SCI 2002 - IIIS*, 2002.

[7] Paul Boersma. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound, 1993.

[8] Alessandro Valitutti Carlo Strapparava. Wordnet-affect: an affective extension of wordnet, 2004.

[9] Richard Sproat Cecilia Ovesdotter Alm, Dan Roth. Emotions from text: machine learning for text-based emotion prediction, 2005.

[10] Chih-Jen Lin Chih-Wei Hsu, Chih-Chung Chang. A practical guide to support vector classification, 2010.

[11] Kornel Laskowski Daniel Neiberg, Kjell Eleniusl. Emotion recognition in spontaneous speech using gmms. In *INTERSPEECH 2006-ICSLP*, Pittsburgh, Pennsylvania.

[12] Richard C. Rose Douglas A. Reynolds. Robust text independent speaker identification using gaussian mixture speaker model. In *IEEE transaction on speech and audio processing, Vol.3, no 1*, January.

[13] Paul Ekman. In *Nebraska Symposium on motivation, Vol. 19.*

[14] Luisa Bentivogli Emanuele Pianta and Christian Girardi. Multiwordnet: Developing and aligned multilingual database. in proceedings of the first international conference on global wordnet, 2002.

[15] Kristina Lundholm Fors. An investigation of intra-turn pauses in spontaneous speech. In *TMH QPSR Vol. 51*, 2011.

[16] Trevor Hastie Robert Tibshirani Jerome Friedman. The elements of statistical learning data: Mining, inference, and prediction, 2008.

[17] Nico H. Frijda. *The Emotions.* Cambridge University Press, 1986.

[18] Peter Grabusts. In *Proceedings of the 8th International Scientific and Practical Conference. Volume II.*

[19] James G. Lyons Kuldip K. Paliwal and Kamil K. Wocicki. Preference for 20-40 ms window duration in speech analysis, 2010.

[20] Carlo Rinaldi Luca Colombo. L'analisi dell'andamento prosodico e il riconoscimento di stati emotivi nel parlato: Progettazione e realizzazione del sistema integrato prema. Master's thesis, Politecnico di Milano, 2005/2006. Relatore: L. Sbattella, Correlatore: R. Tedesco.

[21] George A. Miller. Wordnet: A lexical database for english, 1995.

[22] Dr.B.Paramasivan M.JayaLakshmi, K.Maharajan. Instantaneous emotion detection system using vocalizations. In *IOSR Journal of Engineering (IOSRJEN)*, July.

[23] Y. Srinivas-J. Sirisha Devi N. Murali Krishna, P.V. Lakshmi. Emotion recognition using dynamic time warping technique for isolated words. In *IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 5, No 1*, September 2011.

[24] Lluis Padro and Evgeny Stanilovsky. Freeling 3.0: Towards wider multilinguality, 2012.

[25] Robert P.W. Duin Sang-Woon Kim. On using a pre-clustering technique to optimize lda-based classifiers for appearance-based face recognition, 2015.

[26] L. Sbattella, L. Colombo, C. Rinaldi, R. Tedesco, M. Matteucci, and A. Trivilini. Extracting emotions and communication styles from vocal signals. In *Proceedings of the International Conference on Physiological Computing Systems (PhyCS)*, Lisbon, Portugal, January 2014.

[27] K.R. Scherer. On the nature and function of emotion: a component process approach, 1984.

[28] Will Styler. Using praat for linguistic research, 2015.

[29] Liyanage C. De Silva Tin Lay Nwe, Say Wei Foo. Speech emotion recognition using hidden markov models. In *Elsevier Speech Communications Journal Vol. 41, Issue 4, pp. 603-623*, November 2003.

[30] Armin Kohlrausch Tobias May, Steven van de Par. Noise-robust speaker recognition combining missing data techniques and universal background modeling. In *IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 20, NO. 1*, January 2012.

[31] Ing. Ivar F.T. van Willigen. Reasoning about emotions, 2009.

[32] A. Acero X. Huang and H. Hon. Spoken language processing: A guide to theory, algorithm, and system development, 2001.

[33] Chung-Hsien Wu Ze-Jing Chuang. Multi-modal emotion recognition from speech and text. In *Computational Linguistics and Chinese Language Processing Vol. 9, No. 2*, August.