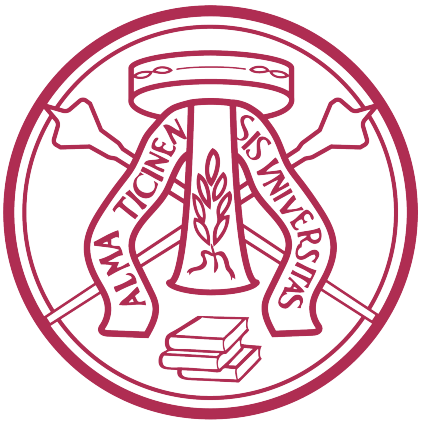


UNIVERSITÀ DI PAVIA

FACOLTÀ DI INGEGNERIA
DIPARTIMENTO DI INGEGNERIA INDUSTRIALE E DELL'INFORMAZIONE

CORSO DI LAUREA MAGISTRALE IN BIOINGEGNERIA

TESI DI LAUREA

Analisi della Voce mediante Deep Learning: Riconoscimento delle emozioni come supporto alla diagnosi del disturbo dello spettro autistico

Relatore:
Prof. Pietro Savazzi

Candidata:
Martina Iozzi

Correlatori:
Prof.ssa Natascia Brondino
Ing. Mauro Marchese

A.A. 2024/2025

Indice

Elenco degli Acronimi

SER Speech Emotion Recognition

ASD Autism Spectrum Disorder

RAVDESS Ryerson Audio-Visual Database of Emotional Speech and Song

EMO-DB Berlin Emotional Database

HCI Human Computer Interaction

ASR Automatic Speech Recognition

ADOS-2 Autism Diagnostic Observation Schedule

ADI-R Autism Diagnostic Interview-Revised

CSS Calibrated Severity Score

RMS Root Mean Square

LTI Lineare Tempo Invariante

ZCR Zero Crossing Rate

MFCC Mel-Frequency Cepstral Coefficients

DCT Trasformata Discreta del Coseno

LLD Low-Level Descriptors

SVM Support Vector Machine

HMM Hidden Markov Model

LFPC Log Frequency Power Coefficients

CNN Convolutional Neural Network

LSTM Long Short Term Memory

LOSO Leave-One-Speaker-Out

NCV Nested Cross Validation

CHAT Codes for the Human Analysis of Transcripts

VAD Voice Activity Detection

eGeMAPS Geneva Minimalistic Acoustic Parameter Set

1D-CLDNN Convolutional LSTM Deep Neural Network

LFLB Local Feature Learning Blocks

ReLU Rectified Linear Unit

GAP Global Average Pooling

BN Batch Normalization

AWGN Additive White Gausian Noise

SNR Signal-to-Noise Ratio

VGG (Visual Geometry Group)

STFT Short-Time Fourier Transform

ROC Receiver Operating Characteristic

AUC Area Under the Curve

Abstract

Lo Speech Emotion Recognition (SER) consiste nello sviluppo di modelli computazionali volti a decodificare lo stato emotivo di un soggetto mediante la sola analisi della microstruttura acustica del parlato, escludendo il contenuto semantico del discorso. Nonostante l'impiego di architetture di Deep Learning abbia migliorato la capacità di estrarre pattern prosodici complessi, l'analisi della letteratura evidenzia spesso criticità legate alle metodologie di validazione. Un errore ricorrente risiede nell'applicazione della Data Augmentation prima della separazione tra i set di addestramento e di test. Tale incoerenza procedurale determina inevitabilmente il problema del Data Leakage, inducendo il modello a memorizzare l'impronta biometrica del parlante o le variabili ambientali a discapito dei tratti emotivi. Ne consegue una sistematica sovrastima delle prestazioni che, di fatto, invalida la reale capacità di generalizzazione del sistema. Il presente lavoro di tesi si propone di sviluppare una pipeline di classificazione robusta, ponendo particolare attenzione al rigore metodologico nella fase di addestramento, per poi verificare l'applicabilità di queste tecniche in ambito biomedico. Nella prima fase del progetto, lo studio si è focalizzato sul riconoscimento delle emozioni utilizzando i dataset pubblici RAVDESS ed EMO-DB. L'elaborazione si è basata sull'estrazione ed il confronto di molteplici set di feature acustiche (cepstrali, prosodiche e fisiche), al fine di isolare i biomarcatori emotivi maggiormente generalizzabili. Lo spazio delle feature, così definito, è stato poi elaborato tramite un'architettura 1D-CLDNN integrata con un modulo di Self-Attention. Per garantire la robustezza dei risultati e minimizzare il rischio di overfitting, si è utilizzato un protocollo di Nested Cross-Validation a 10 fold. All'interno di questa architettura, ogni singolo processo di ottimizzazione è stato eseguito esclusivamente sul training set,

mantenendo il test set completamente isolato. Nella seconda fase, l'approccio metodologico è stato utilizzato per la ricerca di biomarcatori vocali nel contesto del Disturbo dello Spettro Autistico (ASD). Lo studio ha previsto la validazione dei modelli non solo sul corpus pubblico ASDBank, ma anche su un secondo dataset sperimentale. Date le limitazioni dovute alle ridotte dimensioni dei campioni e alla complessità dei dati, si è scelto di adottare una strategia di Transfer Learning. Nello specifico, si è utilizzato il modello pre-addestrato VGGish per estrarre rappresentazioni numeriche (embedding) del segnale audio. Questo ha permesso di confrontare i profili vocali tra il gruppo clinico e quello di controllo, identificando le differenze nella prosodia e nel ritmo tipiche dell'autismo. L'analisi dei dati evidenzia come l'impiego di protocolli di validazione rigorosi porti a risultati più contenuti rispetto alle stime spesso ottimistiche presenti in letteratura, ma maggiormente rappresentativi dell'effettiva sovrapposizione acustica del parlato. Nonostante la difficoltà dei modelli nel separare in modo netto gli stati affettivi, il lavoro di tesi definisce un approccio metodologico riproducibile e rappresenta un punto di partenza concreto per valutare le reali potenzialità dell'analisi vocale in ambito clinico.

1. Introduzione

Il presente capitolo fornisce le nozioni necessarie per comprendere il contesto teorico e applicativo in cui si inserisce il lavoro di tesi. La trattazione descrive l'evoluzione dell'interazione uomo-macchina, dai modelli computazionali classici fino alla nascita dell'Affective Computing, focalizzandosi sulla traduzione degli stati affettivi in variabili computazionali. Viene poi introdotto lo Speech Emotion Recognition (SER), ambito in cui il segnale vocale è definito come vettore complesso di informazioni paralinguistiche. In questa sezione sono inoltre approfondite le criticità tecniche relative all'estrazione di feature e le complessità legate all'elaborazione del segnale audio in contesti reali. La sezione conclusiva delinea l'attuale profilo clinico del Disturbo dello Spettro Autistico (ASD) esaminando i limiti strutturali che caratterizzano gli attuali protocolli diagnostici. Tale disamina risulta propedeutica all'individuazione della voce quale potenziale biomarcatore quantitativo, oggettivo e riproducibile.

1.1 Evoluzione dell'interazione uomo-macchina

L'attuale architettura delle interfacce uomo-macchina (Human-Computer Interaction, HCI) risente di quello che Rosalind Picard definisce un "bias razionalista" [1], ovvero una visione che identifica l'intelligenza solo con la capacità di eseguire operazioni matematiche. Fin dalle sue origini, l'informatica si è sviluppata con l'obiettivo di automatizzare processi matematici deterministici, valutando l'efficienza dei sistemi unicamente in termini di precisione e velocità d'esecuzione. All'interno di questo rigido paradigma, la natura non lineare delle emozioni è risultata del tutto refrattaria a una codifica matematica, venendo a lungo trattata come un semplice rumore stocastico o un'interferenza capace di compromettere la stabilità del siste-

ma. Questo retaggio ha condizionato per molto tempo lo sviluppo dell'HCI, limitandolo a quello che Cowie et al. definiscono canale esplicito [2]. Le interfacce standard supportano infatti una comunicazione focalizzata sulla trasmissione di messaggi semantici, in cui il successo dell'interazione dipende dalla correttezza sintattica dei comandi immessi dall'utente. Al contrario, la comunicazione umana è per sua natura multimodale e necessita dell'integrazione del canale implicito, nel quale i segnali paralinguistici e prosodici veicolano il contesto affettivo e definiscono le modalità di interpretazione del messaggio [2]. L'esclusione di questa dimensione informativa preclude alla macchina una reale comprensione del contesto e genera quei sistemi che Picard definisce «ottusi» (dummy) [1]. Tali macchine, pur vantando un'enorme potenza computazionale, non possiedono i criteri per dare priorità alle informazioni in base allo stato emotivo dell'individuo e falliscono sistematicamente nella gestione degli imprevisti tipici degli ambienti reali. Per superare questo limite strutturale, nel 1997 Picard introduce l'Affective Computing, definito come l'informatica che «si riferisce a, nasce da o influenza deliberatamente le emozioni» [1]. Lo scopo non risiede nella mimesi antropomorfica, quanto nell'integrazione dello stato affettivo dell'utente come variabile fondamentale per ottimizzare le prestazioni del sistema. L'informatica affettiva è finalizzata dunque a chiudere un ciclo di retroazione continuo, in cui il sistema acquisisce i segnali del canale implicito, li decodifica e adatta dinamicamente la propria risposta. Questo passaggio risulta essenziale per trasformare il computer da semplice calcolatore a sistema adattivo, capace di gestire con efficacia la complessità delle interazioni umane.

1.2 L'emozione come variabile computazionale

Il passaggio verso sistemi adattivi trova un solido fondamento scientifico nelle ricerche di Antonio Damasio [3], le quali evidenziano il ruolo della componente affettiva non come una semplice opzione progettuale, ma come una necessità funzionale per l'efficacia dei processi decisionali e la gestione dell'incertezza. Attraverso lo studio di soggetti con lesioni alla corteccia orbito-frontale (come nel caso clinico di "Elliot"), Damasio ha osservato come l'incapacità di processare gli stimoli emozionali comprometta sistematicamente la facoltà di scelta, pur mantenendo inalterate la memoria e le funzioni cognitive superiori[3]. In ambito ingegneristico tale deficit è assimilabile a un sistema incapace di gerarchizzare le informazioni. In assenza di un feedback affettivo, infatti, ad ogni opzione disponibile viene attribuito lo stesso peso computazionale, rendendo impossibile la selezione di un output tra le alternative possibili. Questo meccanismo trova una spiegazione nell'ipotesi dei marcatori somatici [3], secondo cui le emozioni agiscono come filtri euristici di pre-processing. Esse restringono lo spazio delle soluzioni possibili prima dell'analisi razionale, prevenendo così la paralisi computazionale in scenari complessi. Per tradurre tali architetture in parametri elaborabili da software, la letteratura propone due strategie principali di quantificazione quali la modellazione categorica e quella dimensionale.

1.2.1 Il modello categorico di Ekman

Il modello categorico proposto da Paul Ekman definisce le emozioni come meccanismi di adattamento intrinseci finalizzati alla gestione dei cosiddetti "fundamental life-tasks", ovvero scenari critici per la sopravvivenza che richiedono risposte biologiche immediate [4]. Il sistema è regolato dall'automatic appraisal, un processo di valutazione istantanea che, prima ancora

che intervenga la consapevolezza del soggetto, attiva mutamenti somatici involontari quali, ad esempio, il battito cardiaco o il tono della voce. La natura riflessa di queste reazioni, caratterizzate da un'insorgenza rapida e da una durata temporale limitata, rende l'emozione un segnale spontaneo difficile da simulare e oggettivamente misurabile attraverso parametri fisici e comportamentali. L'architettura teorica identifica sei emozioni primarie (rabbia, paura, tristezza, gioia, disgusto e sorpresa), la cui struttura interna è definita dal rapporto tra temi e varianti. Il tema rappresenta il nucleo fisiologico universale, ereditato biologicamente, che determina gli elementi comuni a ogni individuo per una specifica categoria emotiva. Al contrario, le varianti costituiscono le declinazioni specifiche di tale nucleo. Esse sono modellate dall'esperienza del singolo e dalle display rules, ovvero le norme socioculturali che stabiliscono le modalità di espressione pubblica in base al contesto. Per gestire la varietà delle risposte osservabili, Ekman introduce il concetto di famiglie emotive, secondo cui ogni emozione raggruppa un insieme di stati correlati che condividono la medesima identità biologica, pur variando per intensità e sfumature. Dal punto di vista computazionale, questa tassonomia permette di inquadrare il riconoscimento delle emozioni come un problema di classificazione multiclasse, fondamentale per la definizione delle ground truth nei dataset supervisionati. Tuttavia, la rigidità delle categorie costituisce un limite tecnico significativo nella rappresentazione della comunicazione spontanea, dove la fluidità delle transizioni e la sovrapposizione tra stati diversi rendono i confini tra le classi meno netti rispetto ai criteri previsti dal modello originale.

Modello dimensionale: il Circumplex Model of Affect di Russel

In alternativa alla rigidità dei sistemi categorici, James A. Russell ha introdotto il Circumplex Model of Affect [5], interpretando l'esperienza emotiva come un sistema spaziale continuo. In questa topologia, le diverse emozioni

sono mappate lungo il perimetro di un cerchio definito dall'intersezione di due assi ortogonali: la Valenza (V), che esprime la polarità edonica dello stimolo (piacevolezza o spiacevolezza), e l'Attivazione (A), che ne misura l'intensità neurofisiologica (Figura:1.1). Tale approccio introduce il concetto fondamentale di fuzziness, definendo le classi emotive come fuzzy sets (insiemi sfumati) caratterizzati da confini non netti e ampie regioni di sovrapposizione lungo il continuum spaziale. Questo paradigma trasforma il riconoscimento dell'emozione da un compito di classificazione discreta a un problema di posizionamento vettoriale in coordinate (V, A), in modo da rappresentare con maggiore fedeltà la fluidità intrinseca del parlato naturale. Tuttavia, la continuità dello spazio introduce una criticità legata alla prossimità di stati psicologicamente diversi ma acusticamente simili. Un caso rappresentativo è costituito dal quadrante della valenza negativa ad alta attivazione. In questa regione rabbia e paura occupano posizioni limitrofe (Figura 1.1) e condividono tratti acustici spesso sovrapponibili, come l'incremento dell'intensità e della frequenza fondamentale. Per ovviare a tale ambiguità, molti modelli computazionali integrano una terza dimensione, definita dominanza, la quale permette di distinguere gli stati in base al senso di controllo percepito dal soggetto. In quest'ottica, un eventuale errore di classificazione tra emozioni spazialmente vicine non costituisce un fallimento algoritmico, quanto piuttosto una validazione della continuità intrinseca dello spazio emotivo descritto dal modello di Russell.

Figura 1.1: Figura 1:Coordinate di scaling circolare per 28 termini affettivi [5]

1.3 Speech Emotion Recognition

Lo Speech Emotion Recognition (SER) afferisce al dominio dell'Affective Computing ed è finalizzato alla decodifica dello stato emotivo attraverso l'analisi della microstruttura acustica del segnale vocale [6]. A differenza dei sistemi di Automatic Speech Recognition (ASR), focalizzati sulla trascrizione semantica, il SER indaga la sfera paralinguistica interpretando la voce come un vettore biologico complesso [7]. La criticità primaria di questo compito risiede nella natura non stazionaria del segnale, dove l'informazione emotiva risulta sovrapposta a variabili quali l'identità del parlatore, la fonetica linguistica e il disturbo ambientale. L'efficacia algoritmica è pertanto subordinata alla capacità di disaccoppiare questi elementi, isolando pattern acustici che garantiscano robustezza a fronte della variabilità intersoggettiva e dei vincoli del contesto in cui viene utilizzato.

1.3.1 Ambiti applicativi del SER

La versatilità del riconoscimento emotivo lo rende adatto a una vasta gamma di domini in cui lo stato affettivo funge da variabile di controllo per l'ottimizzazione dell'interazione tra utente e sistema. L'ambito di maggiore rilievo per il presente lavoro è quello clinico, dove l'analisi paralinguistica si è consolidata come una metodologia di indagine non invasiva capace di estrarre biomarcatori vocali oggettivi. Tali parametri risultano determinanti per la diagnosi e il monitoraggio di disturbi psichiatrici e del neurosviluppo, permettendo di superare i limiti della soggettività intrinseca nelle valutazioni osservative tradizionali [8]. In particolare, l'identificazione di alterazioni micro-acustiche nella prosodia consente di quantificare stati quali ansia o depressione e fornisce indicatori precoci essenziali per la medicina personalizzata [6]. Oltre alla sfera medicale, l'efficacia del SER trova riscontro in contesti legati alla sicurezza del soggetto. Nel settore automobilistico, l'integrazione di moduli affettivi permette lo sviluppo di tecnologie volte alla prevenzione di incidenti causati da stress o affaticamento del conducente. Analogamente, nei protocolli di pubblica sicurezza, il rilevamento automatico di segnali di panico supporta i processi di triage nelle centrali operative, ottimizzando la distribuzione delle risorse in base alla gravità rilevata nel tono del chiamante [7]. Tali evidenze confermano la transizione verso sistemi adattivi, in grado di variare la risposta tecnologica in funzione del feedback paralinguistico dell'utente.

1.3.2 Criticità strutturali

Nonostante il miglioramento degli algoritmi, l'efficacia dei sistemi SER è limitata da alcune criticità strutturali che ne compromettono l'affidabilità in contesti reali e non controllati. Una delle problematiche principali riguarda la discrepanza tra le emozioni simulate in laboratorio e quelle

prodotte spontaneamente nella realtà quotidiana. La letteratura evidenzia come i database basati su attori presentino spesso manifestazioni emotive esasperate e prototipiche ,che falliscono nel modellare la sottile ambiguità degli stati affettivi reali, dove i tratti paralinguistici risultano meno marcati e più complessi da discriminare [6]. Dal punto di vista dell'elaborazione del segnale, la robustezza dei modelli deve confrontarsi con l'eterogeneità della firma vocale. Poiché ogni individuo possiede caratteristiche uniche influenzate da variabili biometriche (età, genere, anatomia), risulta complesso isolare pattern emotivi universali che siano realmente indipendenti dall' interlocutore (speaker-independent). Alla variabilità intersoggettiva si sovrappone la problematica delle interferenze ambientali, presente soprattutto nelle configurazioni a microfono aperto (open microphone setting). In questo contesto, il sistema deve riuscire a separare con precisione le reali variazioni emotive dal rumore paralinguistico irrilevante o dal disturbo stocastico ambientale. Senza architetture robuste e tecniche di preelaborazione avanzate, il sistema rischia infatti di interpretare il rumore ambientale come un segnale emotivo, compromettendo l'affidabilità dell'intera analisi [7].

1.4 L'Autismo come scenario applicativo

Nel presente lavoro si indaga l'applicabilità dei sistemi SER come supporto alla diagnosi di ASD. L'obiettivo è verificare se l'analisi automatica della voce fornisca parametri oggettivi per affiancare la diagnosi tradizionale, riducendo l'incertezza delle valutazioni osservative. La trattazione delinea le evidenze cliniche necessarie a definire il contesto della successiva indagine tecnica.

1.4.1 Il Disturbo dello Spettri Autistico (ASD)

Il Disturbo dello Spettro Autistico (ASD) è una condizione del neurosviluppo a esordio precoce, caratterizzata da deficit persistenti nella comunicazione sociale e da pattern di comportamento o interessi ristretti e ripetitivi [9]. Negli ultimi anni l'approccio diagnostico ha subìto una profonda evoluzione, passando da una suddivisione in categorie cliniche distinte a una visione unitaria e multidimensionale della patologia. Tale transizione, formalizzata nel DSM-5, ha comportato l'integrazione di quadri clinici precedentemente separati, quali il disturbo autistico, la sindrome di Asperger e il disturbo pervasivo dello sviluppo, in un'unica macrocategoria definita appunto spettro autistico. Questa impostazione risponde alla necessità di evidenziare come la sintomatologia si manifesti in modo eterogeneo in funzione della gravità e dello stadio evolutivo del soggetto. Tuttavia, la letteratura evidenzia limiti legati alla rigidità dei criteri che impongono compromissioni in entrambi i domini della diade sintomatologica. Come rilevato da Hodges et al. (2020), questa impostazione rischia di escludere i profili ad alto funzionamento, i quali possiedono capacità di adattamento che rendono i deficit meno evidenti durante le osservazioni standard [10].

1.4.2 Epidemiologia

Gli studi epidemiologici internazionali segnalano un incremento generalizzato della prevalenza di ASD. Tale fenomeno è riconducibile a una migliore formazione medica, all'affinamento dei criteri diagnostici e alla crescente consapevolezza sociale favorita dal contesto socio-economico. Le stime variano sensibilmente su base geografica e in funzione delle metodologie di ricerca adottate, passando da un caso ogni 54 bambini negli Stati Uniti a valori di uno su 86 in Gran Bretagna e uno su 160 in Danimarca e Svezia. Secondo i dati dell'Osservatorio coordinato dall'Istituto Superiore di Sanità

e dal Ministero della Salute, in Italia circa un bambino su 77 nella fascia d'età tra i 7 e i 9 anni presenta il disturbo. La rilevazione, ottenuta applicando i protocolli di screening del progetto europeo ASDEU, evidenzia infine una netta sproporzione di genere con una presenza maschile 4.4 volte superiore rispetto a quella femminile. [11]

1.4.3 Metodologie di valutazione clinica e limiti dell'osservazione

Ad oggi la diagnosi dell'autismo si basa principalmente sull' integrazione tra l'osservazione comportamentale e la storia clinica. La comunità scientifica internazionale identifica come "gold standard" l'uso combinato di ADOS-2 (Autism Diagnostic Observation Schedule) e ADI-R(Autism Diagnostic Interview-Revised) [12]. L'ADOS-2 consiste in una valutazione semistrutturata che, attraverso attività ludiche o compiti d'interazione standardizzati, elicita risposte comunicative spontanee nel paziente. Lo strumento si articola in cinque moduli distinti, selezionati dal clinico in base all'età e alle capacità linguistiche, per esaminare funzioni che spaziano dall'assenza di linguaggio fino alla fluidità verbale. Questa modularità garantisce un'analisi mirata di competenze quali l'attenzione condivisa e la reciprocità sociale. I dati raccolti durante la sessione vengono poi normalizzati tramite il Calibrated Severity Score (CSS) ovvero un indice standardizzato su una scala da 1 a 10 che permette di quantificare la gravità della sintomatologia indipendentemente dal modulo utilizzato, favorendo un monitoraggio costante nel tempo [12]. Il quadro diagnostico viene completato dall' integrazione dell'ADI-R, un' intervista strutturata rivolta ai caregiver necessaria per ricostruire l'anamnesi evolutiva completa. Tale strumento permette di rilevare tratti atipici presenti sin dalle prime fasi dello sviluppo, verificando la conformità ai criteri definiti dal DSM-5. Nonostante la sua diffusione, questa metodologia presenta delle criticità legate alla natura stessa dell'osservazione [13]. Il contesto ambulatoriale

rappresenta un ambiente artificiale che può indurre i soggetti con maggiori capacità adattive a mascherare i propri deficit attraverso il social camouflaging, compromettendo la validità ecologica della valutazione. A questo si aggiungono la soggettività intrinseca del giudizio clinico e l'onerosità dei tempi di somministrazione ed elaborazione dei risultati, i quali possono rallentare l'intero percorso diagnostico. In questo scenario, la necessità di disporre di indicatori oggettivi e svincolati dalle strategie adattive del paziente sposta l'indagine verso le proprietà acustiche del parlato. Attraverso l'impiego di sistemi SER, la funzionalità di processi bio-meccanici quali la dinamica respiratoria e il controllo laringeo [14] si traduce in metriche quantitative stabili e riproducibili. Questa metodologia fornisce uno strumento non invasivo, rapido ed economico per integrare la valutazione clinica e mitigare l'influenza del controllo volontario sulla validità della diagnosi.

2. Stato dell'arte

L'efficacia dei sistemi di Speech Emotion Recognition (SER) dipende dalla capacità di tradurre la complessità del parlato in parametri fisici, misurabili e riproducibili. Partendo dall'analisi dei meccanismi biologici della fonazione, il presente capitolo ripercorre l'evoluzione tecnologica del settore, dai modelli di apprendimento automatico tradizionale fino alle architetture neurali profonde e al paradigma del Transfer Learning. L'analisi della letteratura qui proposta integra una revisione delle principali criticità metodologiche con l'obiettivo di evidenziare l'importanza di protocolli di validazione rigorosi nella prevenzione del Data Leakage. La trattazione si focalizza infine sull'ambito clinico del Disturbo dello Spettro Autistico, dove la prosodia viene indagata quale potenziale biomarcatore diagnostico oggettivo e quantitativo.

2.1 La fonazione e il modello Sorgente-Filtro

La comprensione della voce umana come veicolo di informazioni affettive richiede un'analisi che trascenda la natura puramente acustica del segnale, indagando i processi biomeccanici sottostanti. La produzione del parlato è l'esito di una coordinazione pneumo-fonica, regolata dal sistema nervoso centrale, e basata su tre stadi fisiologici: i polmoni (sistema sottoglottideo), la laringe, e il tratto vocale (sistema sopraglottideo). Come illustrato in Figura 2.1, i polmoni fungono da sorgente energetica, fornendo un flusso d'aria continuo che viene modulato a livello laringeo. Qui, in accordo con la teoria mioelastica-aerodinamica, la pressione dell'aria vince la resistenza elastica delle pliche vocali, innescando un'oscillazione che trasforma il flusso costante in una sorgente di eccitazione. Tale segnale può assumere tre differenti configurazioni acustiche:

- Vibrazioni glottali periodiche: una serie di impulsi quasi-periodici la cui frequenza di ripetizione definisce la frequenza fondamentale F_0 (pitch). Esso rappresenta il meccanismo alla base dei suoni sonorizzati come le vocali.

- Rumore: fonte di eccitazione aperiodica generata dal passaggio dell'aria attraverso una costrizione del condotto vocale, tipica dei fonemi fricativi $(es./s/, /f/)$.

- burst transitori: brevi impulsi di eccitazione generati dal rilascio istantaneo della pressione intraorale accumulata durante un'occlusione del condotto vocale. Sono caratteristici dei fonemi occlusivi $(es./p/, /t/)$.

Indipendentemente dalla sua natura, il segnale di eccitazione attraversa il tratto vocale sopraglottideo che agisce come un sistema di risonanza a geometria variabile. Questo condotto filtra il suono originario e ne modella le caratteristiche timbriche finali attraverso il continuo mutamento della propria conformazione spaziale[15].

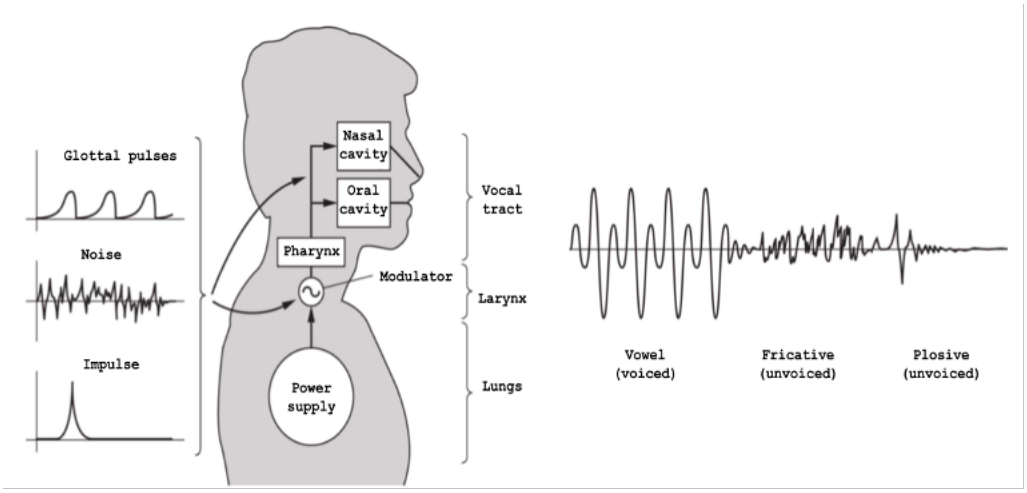


Figura 2.1: Rappresentazione schematica del meccanismo di produzione del parlato. A sinistra sono evidenziate le tre forme d'onda dell'eccitazione laringea, mentre a destra il segnale vocale irradiato risultante[15]

2.1.1 Il modello matematico Sorgente-Filtro

Tale architettura biologica trova la sua sintesi teorica nel paradigma sorgente-filtro, formalizzato da Gunnar Fant (1960). Secondo questo modello, il segnale vocale viene interpretato come la risposta di un sistema di filtri (il tratto vocale) a una o più sorgenti di eccitazione laringea [16]. Assumendo la mutua indipendenza tra i due stadi, il segnale vocale p(t) nel dominio del tempo è modellabile come la convoluzione tra la sorgente $s(t)$, la risposta all'impulso del tratto vocale $v(t)$ e l'effetto di irradiazione labiale $r(t)$.

$$p(t) = s(t) * v(t) * r(t) \tag{2.1}$$

Nel dominio di Laplace, l'operazione di convoluzione si semplifica in un prodotto algebrico tra le rispettive funzioni di trasferimento. Di conseguenza, l'andamento spettrale del segnale vocale può essere ricavato analiticamente come segue:

$$P(s) = S(s) \cdot V(s) \cdot R(s) \tag{2.2}$$

Sebbene la fonazione reale comporti un inevitabile accoppiamento acustico tra la laringe e il tratto vocale, l'ipotesi di indipendenza lineare di Fant rimane un'approssimazione necessaria per l'elaborazione del segnale. Passando infatti dal modello fisico all'implementazione algoritmica, bisogna considerare che i classificatori operano esclusivamente su dati campionati. Il problema va quindi trasposto nel dominio del tempo discreto, formalizzando il segnale digitalizzato nel seguente modo [15]:

$$s(n) = e[n] * v[n] \tag{2.3}$$

dove $e[n]$ quantifica l'eccitazione laringea discreta e $v[n]$ rappresenta la

risposta all'impulso del filtro digitale. Applicando la Trasformata Z, la relazione assume la forma compatta:

$$S(z) = E(z) \cdot V(z) \tag{2.4}$$

Questa scomposizione è fondamentale nell'ambito del SER perché consente di isolare i correlati acustici dell'emozione e delle patologie vocali. Nello specifico, l'attivazione neurofisiologica (arousal) altera primariamente le dinamiche pressorie della laringe andando a modificare lo spettro della sorgente $E(z)$. Al contrario, la valenza affettiva (valence) e gli eventuali deficit di controllo motorio agiscono sull'assetto articolatorio e si manifestano come una variazione della distribuzione energetica spettrale che altera il timbro e la qualità vocale modellati dalla funzione di trasferimento $V(z)$ [17].

2.2 Caratterizzazione delle feature acustiche

La conversione del segnale vocale in una variabile misurabile richiede la risoluzione della sua intrinseca non stazionarietà. Poiché le configurazioni del tratto vocale e della laringe mutano rapidamente durante l'eloquio, le proprietà statistiche del segnale sono considerate costanti solo su intervalli temporali ridotti. L'elaborazione si basa dunque sul paradigma della Short-Time Analysis, che prevede una segmentazione in finestre temporali (frame) di durata compresa tra 10 e 40 ms. Al fine di minimizzare le distorsioni causate dal troncamento netto del segnale, a ogni frame viene applicata una funzione di finestratura, tipicamente quella di Hamming. Tale preelaborazione è propedeutica all'estrazione di feature acustiche capaci di mappare sia la biomeccanica della sorgente $S(z)$ che la configurazione del tratto vocale $V(z)$ [18].

2.2.1 Caratteristiche della sorgente

La caratterizzazione della sorgente si focalizza sulla componente laringea, la cui oscillazione è determinata dall'interazione tra la pressione dell'aria (sottoglottidea) e la tensione delle pliche vocali [18]. Il parametro prosodico primario è la frequenza fondamentale F_0, definita come l'inverso del periodo glottale T_0 [19]:

$$F_0 = \frac{1}{T_0} \tag{2.5}$$

In ambito acustico, la F_0 rappresenta il principale indicatore dell'attivazione fisiologica (arousal), in quanto un incremento della spinta aerodinamica e della rigidità tissutale inducono un innalzamento del tono percepito [17]. Sebbene la F_0 tenda a stabilizzarsi in età adulta in un range compreso tra 60 e 400 Hz,le sue alterazioni sono clinicamente rilevanti. Ad esempio, nel Disturbo dello Spettro Autistico (ASD), la letteratura evidenzia pattern di monotonia tonale o variabilità atipiche che deviano significativamente dai profili neurotipici. [14]. Oltre al tono, la stabilità della sorgente viene analizzata tramite i parametri di micro-perturbazione. Essi includono il Jitter, che valuta l'instabilità temporale della F_0 tra cicli glottali consecutivi, e lo Shimmer, che invece mappa le fluttuazioni involontarie dell'ampiezza dell'onda [19]. Parallelamente, l'intensità dell'eccitazione laringea viene quantificata tramite l'energia Root Mean Square (RMS), la quale fornisce la potenza quadratica media del segnale all'interno della finestra di analisi [18]

$$RMS(i) = \sqrt{\frac{1}{N}\sum_{n=0}^{N-1} x(n)^2} \tag{2.6}$$

Questo parametro è un indicatore dello sforzo fonatorio e dell'intensità sonora. Emozioni ad alta intensità, come rabbia o gioia, inducono una crescita sistematica dei valori energetici, mentre stati a bassa attivazione, come la tristezza, presentano una riduzione dell'RMS e una minore variabilità [17].

2.2.2 Caratteristiche del tratto vocale

Il tratto vocale agisce come un modulatore spettrale dinamico il cui compito primario è la trasformazione del suono grezzo generato dalla laringe in un segnale acustico strutturato e intellegibile. Attraverso meccanismi di risonanza selettiva, tale sistema accentua specifiche bande di frequenza, conferendo un'impronta timbrica distintiva a una sorgente che, altrimenti, risulterebbe acusticamente indifferenziata. La risposta in frequenza del tratto vocale dipende strettamente dalla sua configurazione spaziale, soggetta a variazioni continue indotte dall'attività degli organi articolatori (lingua, labbra, velo palatino). Sebbene la morfologia cambi tra i diversi soggetti, in letteratura si assume convenzionalmente una lunghezza media del condotto pari a 17.5 cm per un uomo adulto[16]. In regime di quasi-stazionarietà, il comportamento del tratto vocale può essere descritto matematicamente come un filtro lineare tempo-invariante (LTI). La modellazione analitica prevalente rappresenta tale sistema attraverso una funzione di trasferimento a soli poli (all-pole model), la cui formulazione nel dominio Z è :

$$H(z) = \frac{G}{1 - \sum_{k=1}^{P} a_k z^{-k}} \tag{2.7}$$

In tale architettura, i poli complessi coniugati c_k del sistema, situati all'interno di un cerchio unitario ($|c_k| < 1$), identificano le frequenze naturali di vibrazione del condotto, denominate formanti. L'insieme delle prime tre componenti (F_1, F_2, F_3) definisce il cosiddetto "F-pattern", ovvero un set di parametri necessario per determinare sia l'identità dei fonemi che il profilo timbrico del parlatore [18]. La stretta correlazione tra dimensione fisica e risposta acustica risulta evidente analizzando la produzione della vocale neutra schwa [ə]. In questa specifica configurazione, il tratto vocale è approssimabile a un tubo cilindrico uniforme, aperto all'estremità labiale

e chiuso in quella glottale. Assumendo un'impedenza acustica trascurabile all'apertura orale, si generano onde stazionarie la cui frequenza fondamentale F_n si ottiene quando la lunghezza del condotto equivale esattamente a un quarto della lunghezza d'onda sonora:

$$F_n = \frac{(2n+1)c}{4L} \text{ con n=0,1,2...} \tag{2.8}$$

Con una velocità di propagazione del suono $c = 350m/s$ e una lunghezza L=17.5 cm, le prime tre formanti risultano pari a 500 Hz (F_1), 1500 Hz (F_2) e 2500 Hz (F_3). Qualsiasi alterazione della conformazione del tratto, sia essa dovuta a variabili biologiche o a pattern articolatori atipici indotti da stati emotivi o condizioni patologiche, determina una traslazione di queste frequenze, modificando radicalmente il profilo spettrale del segnale irradiato [20].

2.2.3 Caratteristiche combinate sorgente-filtro

Sebbene il modello sorgente-filtro definisca teoricamente l'indipendenza tra la sorgente sonora e il tratto vocale [16], l'analisi del segnale campionato $s[n]$ rivela come questi due contributi siano intrinsecamente legati da un'operazione di convoluzione. Tale complessità strutturale rende necessaria l'adozione di descrittori acustici "combinati", capaci di mappare l'interazione globale del sistema fonatorio. Un indicatore fondamentale in questo ambito è lo Zero-Crossing Rate (ZCR). Esso quantifica il tasso di inversione della polarità del segnale all'interno del frame di lunghezza W_L [21]

$$Z(i) = \frac{1}{2W_L} \sum_{n=1}^{W_L} |\text{sgn}[x_i(n)] - \text{sgn}[x_i(n-1)]| \tag{2.9}$$

Dal punto di vista dell'elaborazione, esso funge da stimatore della densità spettrale, permettendo di discriminare tra segmenti sonorizzati (voiced),

caratterizzati da una forte periodicità e valori di ZCR ridotti, e componenti aperiodiche (unvoiced). Biomeccanicamente, una variazione dello ZCR riflette il passaggio tra diversi regimi di eccitazione. In stati di forte attivazione emozionale, quali l'ansia o la paura, la tensione della laringe e l'instabilità del flusso d'aria generano componenti turbolente ad alta frequenza. Esse alterano la forma d'onda acustica provocando un incremento sistematico dei passaggi per lo zero rispetto ai regimi laminari e periodici tipici della voce neutra. Mentre lo Zero Crossing Rate (ZCR) è definito nel dominio nel tempo, le feature spettrali di forma quantificano la distribuzione dell'energia lungo l'asse delle frequenze. In questo contesto, lo Spectral Centroid rappresenta il baricentro dello spettro delle ampiezze [21] calcolato come media pesata delle frequenze $f(k)$ sui rispettivi moduli dell'ampiezza $|X(k)|$

$$C_i = \frac{\sum_{k=1}^{W_L} f(k) \cdot |X_i(k)|}{\sum_{k=1}^{W_L} |X_i(k)|} \tag{2.10}$$

Questo parametro mappa la percezione della "brillantezza" sonora. Dal punto di vista biologico, un centroide elevato indica uno spostamento energetico verso le frequenze acute, causato da una chiusura rapida e netta della sorgente sonora, tipica della rabbia. Al contrario, una chiusura lenta e incompleta delle pliche vocali agisce come un filtro passa-basso naturale che smorza le alte frequenze, producendo lo spettro caratteristico della tristezza [17]. In stretta correlazione con il centroide, vi è lo Spectral Rolloff che misura la frequenza m al di sotto della quale è contenuta una frazione specifica (convenzionalmente il 90%) della potenza totale del segnale: [21]:

$$\sum_{k=1}^{m} |X_i(k)| = C \cdot \sum_{k=1}^{W_L} |X_i(k)| \tag{2.11}$$

Nel Disturbo dello Spettro Autistico (ASD), valori anomali di questi parametri possono segnalare anomalie nel controllo motorio fine, le quali si manifestano attraverso voci ipertese (con rolloff e centroide elevati) o ipofoniche. In questo scenario, le feature di riferimento per il SER rimangono i Mel-Frequency Cepstral Coefficients (MFCC). L'obiettivo principale di questo algoritmo è isolare le caratteristiche del tratto vocale (il filtro), eliminando le interferenze legate alla vibrazione delle pliche vocali (la sorgente) [20]. Il calcolo emula la risposta del sistema uditivo umano e si articola in cinque passaggi matematici fondamentali [22]:

- Pre-enfasi: Il segnale audio viene filtrato per enfatizzare le alte frequenze, simulando il guadagno naturale dell'orecchio umano per tali onde. Si applica un filtro del primo ordine secondo la relazione:

$$S(n) = Data(n) - 0.95 \cdot Data(n-1) \tag{2.12}$$

  dove Data(n) rappresenta il campione originale e S(n) il segnale filtrato.

- Finestratura: Il segnale vocale viene diviso in brevi segmenti temporali (frame) a cui si applica una finestra (tipicamente di Hamming) per minimizzare la dispersione di energia ai bordi

$$W(n) = 0.54 - 0.46\cos\left(\frac{2n\pi}{N-1}\right), \quad 0 \leq n \leq N-1 \tag{2.13}$$

  , dove N indica la lunghezza in campioni del frame.

- Trasformata di Fourier e Banco di Filtri Mel: Lo spettro di ogni frame viene calcolato tramite la Fast Fourier Transform (FFT). Lo spettro di potenza risultante viene processato da un banco di M filtri triangolari sovrapposti $B_m[k]$, le cui frequenze centrali sono spaziate secondo la scala non-lineare Mel. Essa è concepita per modellare la

percezione non lineare dell'apparato uditivo umano, il quale manifesta una sensibilità molto elevata ai cambiamenti nelle basse frequenze (sotto i 1000 Hz) e una risoluzione progressivamente ridotta per i toni più acuti. La mappatura dalla frequenza fisica f (espressa in Hz) alla scala psicofisica Mel è determinata dalla seguente relazione logaritmica:

$$M(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \tag{2.14}$$

La risposta in ampiezza del singolo filtro triangolare è definita come:

$$B_m(k) = \begin{cases} \frac{k - f_{m-1}}{f_m - f_{m-1}} & f_{m-1} \leq k \leq f_m \\ \frac{f_{m+1} - k}{f_{m+1} - f_m} & f_m \leq k \leq f_{m+1} \\ 0 & \text{altrove} \end{cases} \tag{2.15}$$

- Calcolo dell'Energia Logaritmica: Per ogni filtro m, si calcola l'energia logaritmica del segnale nella specifica banda di frequenza:

$$Y(m) = \log\left(\sum_{k=f_{m-1}}^{f_{m+1}} |X[k]|^2 B_m[k]\right) \tag{2.16}$$

In questa espressione, $X[k]$ rappresenta lo spettro del frame e $B_m[k]$ la risposta del filtro triangolare m-esimo. L'applicazione del logaritmo trasforma il legame tra sorgente sonora e tratto vocale da moltiplicativo a additivo, semplificando la separazione dei due contributi e facilitando l'isolamento del filtro acustico.

- Trasformata Coseno Discreta (DCT): Quest'ultimo passaggio decorrela le energie logaritmiche per ottenere i coefficienti cepstrali effettivi. La formulazione moderna utilizzata nella letteratura algoritmica è:

$$C_x(n) = \frac{1}{M} \sum_{m=1}^{M} Y(m) \cos\left(n\pi \frac{m - 0.5}{M}\right) \tag{2.17}$$

dove $C_x(n)$ è l'n-esimo coefficiente MFCC generato a partire dalle M energie logaritmiche $Y(m)$.

Poiché i coefficienti MFCC forniscono una rappresentazione esclusivamente statica del parlato, una loro variazione è l'evidenza matematica di una modifica nella posizione degli organi articolatori indotta da variabili affettive o cliniche. L'analisi viene dunque estesa alle componenti dinamiche Delta (Δ) e Delta-Delta ($\Delta - \Delta$) le quali misurano rispettivamente la velocità e l'accelerazione dei cambiamenti timbrici tra i segmenti consecutivi, permettendo di catturare l'evoluzione temporale del segnale e di compensare le eventuali distorsioni tempo-invarianti di canale [18]. La selezione specifica dei parametri estratti verrà approfondita nel Capitolo 3, dove saranno descritti anche la metodologia sperimentale e i modelli di apprendimento automatico adottati per l'analisi dei dati.

2.3 Evoluzione tecnologica del SER

L'evoluzione della Speech Emotion Recognition (SER) è caratterizzata dal superamento dei descrittori euristici estratti manualmente a favore di architetture capaci di apprendere autonomamente rappresentazioni gerarchiche del segnale. Tale progresso ha trasformato l'analisi acustica da una sintesi di statistiche aggregate a una modellazione dinamica dell'evoluzione temporale e spettrale della voce. Le attuali soluzioni integrano diverse metodologie per mitigare la variabilità intersoggettiva tramite l'estrazione di feature astratte e indipendenti dalle caratteristiche fisiche del soggetto.

2.3.1 Tecniche di Machine Learning

L'approccio tradizionale si basa sul paradigma del Machine Learning, il quale prevede una netta separazione tra la fase di estrazione delle caratteristiche acustiche (feature engineering) e la successiva fase di inferenza

statistica. Tale metodologia trasforma la natura non stazionaria del segnale vocale in una rappresentazione vettoriale statica di lunghezza fissa, permettendo al sistema di confrontare audio di diversa durata all'interno di uno spazio decisionale omogeneo. La fase di caratterizzazione del segnale prevede l'estrazione dei Low-Level Descriptors (LLD), ovvero parametri che catturano le proprietà fisiche della voce su brevi intervalli temporali. Per ottenere una firma acustica globale dell'intero enunciato, a questi descrittori vengono applicati dei funzionali statistici quali media, varianza e skewness, che condensano le informazioni prosodiche e spettrali in vettori sintetici [23]. L'efficacia di questa rappresentazione risulta strettamente legata all'impiego delle Support Vector Machines (SVM), algoritmi finalizzati alla determinazione dell'iperpiano di separazione ottimale tra le diverse classi affettive. A differenza dei classificatori basati sulla minimizzazione dell'errore medio, le SVM operano infatti secondo il principio della massimizzazione del margine di separazione per conferire al sistema una maggiore robustezza rispetto alle variabilità dei dati in input. Inoltre, data la natura spesso non linearmente separabile del segnale emotivo, l'impiego di funzioni kernel consente di proiettare le caratteristiche acustiche in spazi a dimensionalità superiore. Tale operazione permette di tracciare confini decisionali netti tra stati affettivi altrimenti sovrapposti, rendendo computazionalmente trattabili anche i pattern più complessi. Tuttavia, questo approccio manifesta un limite strutturale dovuto proprio all'aggregazione statistica dei dati, la quale determina una perdita di risoluzione temporale che impedisce di modellare le micro-variazioni del parlato. Per integrare tale dimensione la letteratura ha introdotto i Modelli di Markov Nascosti (HMM), i quali interpretano l'emozione come una sequenza dinamica di stati di energia spettrale [24]. In questa configurazione la classificazione non dipende esclusivamente dall'ampiezza dei parametri acustici ma dalla probabilità di transizione tra le diverse configurazioni nel tempo. Per

ottimizzare tale analisi i parametri tradizionali vengono spesso sostituiti dai Log Frequency Power Coefficients (LFPC), capaci di mappare con precisione le fluttuazioni della frequenza fondamentale e la velocità di eloquio. Nonostante la validità di questi approcci, i singoli classificatori convenzionali mostrano limiti significativi in scenari reali non controllati [6]. La difficoltà nel distinguere emozioni acusticamente simili e la scarsa resistenza al rumore ambientale hanno spinto la ricerca verso l'adozione di architetture ibride e tecniche di *ensemble*. Queste soluzioni integrano diversi modelli per garantire una maggiore capacità di generalizzazione e prestazioni più stabili in contesti operativi complessi [6].

2.3.2 Tecniche di Deep Learning

L'adozione dei paradigmi di Deep Learning permette l'apprendimento autonomo di gerarchie di rappresentazioni complesse a partire da dati strutturati. A differenza dei modelli convenzionali le architetture profonde elaborano direttamente le matrici di feature acustiche per preservare l'integrità informativa del segnale ed evitare le approssimazioni derivanti dalla rappresentazione statistica. In questo contesto, le Reti Neurali Convoluzionali (CNN) analizzano la matrice delle feature come una griglia bidimensionale dove ogni riga rappresenta un descrittore acustico e ogni colonna un istante temporale. Per mezzo di filtri matematici, denominati kernel, queste architetture rilevano pattern locali salienti attraverso l'identificazione di variazioni di intensità lungo gli assi temporale e frequenziale [25]. La successione di operazioni di convoluzione e layer di pooling costruisce una gerarchia di descrittori progressivamente più astratti, garantendo l'identificazione dei tratti emotivi indipendentemente dalla loro esatta collocazione temporale o dalla tonalità della voce. Tale proprietà rende il modello resiliente alle variabilità individuali dei parlanti oltre che ai disturbi ambientali. Mentre le CNN si focalizzano sulla forma dei pattern acustici l'impiego

delle reti Long Short-Term Memory (LSTM) risulta fondamentale per la modellazione della dimensione sequenziale e della prosodia. Queste architetture si distinguono per un'organizzazione a celle regolata da meccanismi di gating che gestiscono attivamente il flusso informativo lungo l'intero enunciato. Nello specifico, l'interazione tra i gate di input, forget e output permette la selezione delle informazioni da conservare nello stato della cella o da scartare se ritenute irrilevanti per il riconoscimento finale [26]. Questo meccanismo assicura la persistenza delle dipendenze a lungo termine e garantisce che la rete possa correlare con precisione eventi acustici distanti tra loro per mantenere l'integrità del segnale anche attraverso sequenze audio di durata prolungata. Tuttavia, l'esigenza di gestire enunciati complessi ha promosso l'adozione di meccanismi di Self-Attention [27]. Tale tecnica permette di superare l'uniformità di trattamento del segnale tramite l'attribuzione di pesi differenti ai diversi istanti temporali in funzione della loro rilevanza informativa. L'analisi selettiva dei segmenti che presentano i marker acustici più significativi, quali picchi energetici o modulazioni enfatiche, mitiga i limiti di saturazione delle LSTM. Tale approccio garantisce che la rete si focalizzi esclusivamente sui tratti fonatori realmente distintivi in modo da ottimizzare la separabilità tra le diverse classi emotive.

2.3.3 Tecniche di Transfer Learning

L'introduzione del Transfer Learning si configura come un'istanza metodologica necessaria per ovviare alla scarsità di dati etichettati, superando i limiti dell'addestramento condotto esclusivamente ex novo. Tale paradigma, consiste nella trasposizione di modelli ottimizzati su domini massivi (classificazione audio su larga scala) verso obiettivi più specifici (classificazione emozioni) mediante il Representation Learning [28]. In questo processo, il segnale audio viene convertito in embedding semantici ovvero vettori di dati che sintetizzano le proprietà acustiche e prosodiche del parlato in una

rappresentazione matematica stabile e altamente discriminativa. In accordo con la rassegna di Zhuang et al. (2020) [29], l'impiego di parametri stabilizzati mediante il pre-addestramento su dataset estesi, come l'ontologia AudioSet di Google, garantisce una superiore capacità di generalizzazione e protegge la struttura dal rischio di overfitting. Nel panorama attuale la ricerca include sia modelli operanti sulla forma d'onda grezza sia architetture convoluzionali basate sull'analisi tempo-frequenza come YAMNet e VGGish. La solidità di questi modelli consente di limitare l'apprendimento ai soli strati terminali della rete, assicurando un'analisi affidabile anche per segnali acustici complessi. In virtù di tali considerazioni, in questo lavoro di tesi si integra l'architettura VGGish per la produzione di descrittori audio ad alto livello, le cui specifiche tecniche verranno dettagliate nel Capitolo 3.

2.4 Limiti del riconoscimento emotivo

L'affidabilità dei sistemi SER è strettamente vincolata a una serie di criticità che spaziano dal rigore dei protocolli di validazione alla natura stessa del segnale acustico. Un'analisi coordinata di tali fattori permette di distinguere tra le distorsioni indotte da una progettazione sperimentale fallace e i limiti biologici insiti nella percezione umana delle emozioni.

2.4.1 Integrità del protocollo di validazione e il problema del Data Leakage

Un'analisi critica della letteratura recente evidenzia alcune criticità metodologiche riguardanti l'affidabilità dei protocolli di validazione adottati nei sistemi SER. Nonostante la diffusione di reti di Deep Learning che dichiarano tassi di accuratezza superiori al 95%, alcune rassegne sistematiche, come quella condotta da Wani et al. (2021), sollevano forti dubbi sulla validità

di tali valori, definendoli il risultato di una sovrastima ottimistica legata a protocolli di validazione deboli [30]. Il problema centrale risiede nello speaker bias, ovvero una distorsione che emerge quando la partizione dei dati non garantisce l'indipendenza dei soggetti tra le fasi di addestramento e di test. In questo contesto, il sistema attua un riconoscimento biometrico latente. La rete neurale correla l'identità vocale specifica a etichette emotive predefinite anziché estrapolare pattern affettivi universali. Tale fenomeno, definito Data Leakage, compromette drasticamente la capacità di generalizzazione del modello, rendendolo inefficace su soggetti estranei al dataset di origine. Per assicurare dunque l'affidabilità dei risultati, la letteratura identifica come protocolli di validazione più rigorosi il Leave-One-Speaker-Out (LOSO) e la Nested Cross-Validation (NCV). Mentre il LOSO garantisce la speaker independence obbligando il sistema a confrontarsi con voci mai udite durante l'addestramento [31], la validazione nidificata assicura l'indipendenza del processo di model selection. Attraverso la separazione tra i processi di tuning e valutazione, la NCV impedisce che la scelta degli iperparametri risenta dei dati di test. Questa procedura mitiga il rischio di produrre stime iper-ottimistiche che possono gonfiare artificialmente i risultati di oltre il 10% [32]. L'integrazione di tali criteri conduce ad un necessario ridimensionamento delle performance dichiarate. Come confermato dalla rassegna di Wani et al. (2021), le accuratezze reali si attestano tipicamente in un range compreso tra il 55% e il 72%. Tale intervallo,sebbene appaia numericamente inferiore, riflette la robustezza del sistema in scenari reali e la sua effettiva utilità in ambito clinico o ingegneristico[30].

2.4.2 Ambiguità acustica e soggettività della Ground Truth

Oltre ai limiti procedurali, i sistemi SER devono confrontarsi con la natura intrinsecamente soggettiva dell'emozione umana, che rende estremamente

complessa la definizione di una "verità assoluta" (*ground truth*) del segnale. A differenza di altri domini del *pattern recognition*, le classi emotive sono soggette alla varianza interpretativa degli esperti e alla qualità della recitazione nei dataset simulati. Come evidenziato da Akçay e Oğuz (2020) ,la letteratura rileva una marcata asimmetria acustica tra le diverse categorie emozionali [6].Nello specifico, le emozioni ad alta attivazione, come la rabbia, presentano biomarker energetici netti e facilmente discriminabili. Al contrario, gli stati a bassa attivazione o a valenza sfumata mostrano ampie aree di sovrapposizione spettrale che impediscono una netta distinzione matematica del segnale. Tale ambiguità non costituisce un limite esclusivo del calcolo neurale ma riflette un tetto prestazionale biologico osservabile anche nei test percettivi condotti su ascoltatori umani. In assenza di informazioni visive o di un contesto semantico, la capacità di discernere correttamente uno stato emotivo dal solo audio risulta sensibilmente inferiore rispetto ai risultati dichiarati in studi privi di rigore metodologico. In questa prospettiva, la sfida della ricerca contemporanea non risiede più nel raggiungimento della migliore accuratezza, quanto piuttosto nello sviluppo di sistemi capaci di mantenere la stabilità predittiva nonostante l'ambiguità del segnale e la scarsità di campioni etichettati.

2.4.3 Paradigmi di Machine Learning e Deep Learning nel riconoscimento dell'ASD

L'analisi computazionale dei disturbi dello spettro autistico (ASD) ha registrato negli ultimi anni una transizione dai modelli statistici tradizionali verso le architetture neurali profonde, nel tentativo di gestire la variabilità intrinseca dei segnali biologici. Inizialmente la ricerca si è focalizzata sull'estrazione manuale di parametri acustici lineari, quali la frequenza fondamentale (f_0) e l'intensità energetica, elaborati mediante algoritmi di

apprendimento supervisionato come le Support Vector Machine (SVM) o le Random Forest [7]. La meta-analisi di Fusaroli et al. (2017) ha tuttavia ridimensionato l'efficacia di tali modelli, evidenziando una netta discrepanza tra approcci univariati e multivariati . Se l'analisi dei singoli parametri mostrava capacità discriminatorie modeste (61-64%), le prestazioni superiori dichiarate dai modelli multivariati risultavano spesso inficiate da sistematici errori metodologici [14]. L'impiego di coorti ridotte e registrazioni in ambienti asettici ha infatti favorito l'insorgenza di fenomeni di overfitting, portando i classificatori a memorizzare pattern specifici dei dati di addestramento. Per superare l'inefficacia delle singole feature isolate, la ricerca si è evoluta verso reti profonde capaci di processare combinazioni complesse di parametri acustici. Un esempio è rappresentato dal lavoro di Eni et al. (2020), i quali hanno utilizzato 60 caratteristiche prosodiche, acustiche e conversazionali come input di reti neurali convoluzionali CNN per quantificare la severità dell'autismo. Attraverso una procedura di cross-validation bilanciata ripetuta 50 volte, il modello ha raggiunto una correlazione di 0.72 con le valutazioni cliniche e un errore quadratico medio (RMSE) di 4.65 [33]. Tale risultato dimostra la capacità delle reti neurali di estrarre informazioni rilevanti da set di feature eterogenei e complessi. Nonostante la superiorità delle architetture profonde l'addestramento di queste reti resta subordinato alla disponibilità di grandi set di dati etichettati. Per questo motivo lo stato dell'arte attuale predilige il ricorso al Transfer Learning, un paradigma in grado di apprendere rappresentazioni latenti direttamente dalla forma d'onda del segnale (raw audio)[29]. L'efficacia di queste tecniche emerge nel lavoro di Chi et al. (2022) condotto su registrazioni domestiche di bambini con autismo [34]. Gli autori riportano un'accuratezza del 76.9% con l'impiego di wav2vec 2.0 mentre l'applicazione di modelli CNN agli spettrogrammi raggiunge il 79.3%. Tali risultati dimostrano la robustezza dei modelli pre-addestrati a fronte delle distorsioni acustiche

tipiche dei contesti reali. Il contributo di Narain et al. (2019) valida l'uso della rete VGGish per l'estrazione di embedding a 128 dimensioni da vocalizzazioni naturali di soggetti con autismo. Un classificatore LSTM raggiunge un'accuratezza del 70.3% nella classificazione della risata e del 69% per gli stati affettivi negativi. Lo studio impiega una procedura di verifica su dati indipendenti per valutare la capacità di generalizzazione del modello e contrastare l'overfitting. Queste evidenze confermano la solidità del Transfer Learning come base tecnica per lo screening digitale.

3. Materiali e Metodi

In questo capitolo sono illustrate le varie fasi del progetto, a partire dal reperimento dei dataset audio per l'ambito emotivo e clinico fino alla classificazione tramite reti di Deep Learning. Il testo segue la successione delle attività svolte, descrivendo le procedure di pre-elaborazione e i criteri di validazione adottati per l'analisi dei dati.

3.1 Dataset

La prima fase della ricerca ha riguardato la selezione di quattro dataset, differenziati tra loro per lingua e ambito di analisi. Per il riconoscimento emotivo sono stati impiegati il Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) e il Berlin Emotional Database (EMO-DB), mentre per lo studio dello spettro autistico (ASD) è stato scelto il corpus ASDBank e un ulteriore dataset sperimentale volto a validare la capacità di generalizzazione dei modelli in un contesto clinico reale.

3.1.1 Dataset RAVDESS

Il Ryerson Audio-Visual Database of Emotional Speech and Song, noto con l'acronimo RAVDESS, è un corpus multimodale progettato per lo studio dell'espressione emotiva attraverso i canali audio e video. Sebbene il database originale comprenda anche una sezione dedicata al canto, la presente ricerca si focalizza esclusivamente sulla componente audio del parlato (speech). Il dataset comprende 1440 registrazioni vocali prodotte da un campione di 24 attori professionisti (12 uomini e 12 donne), selezionati per la loro capacità di recitare con un accento nordamericano neutro. Al fine di neutralizzare il bias lessicale e garantire che l'informazione emotiva

derivi unicamente dalla prosodia, gli attori hanno pronunciato due frasi prive di connotazione semantica: *"Kids are talking by the door"* e *"Dogs are sitting by the door"*. Le emozioni sono suddivise in otto categorie: stato neutrale, calma, felicità, tristezza, rabbia, paura, disgusto e sorpresa. Ad esclusione della condizione neutrale, ogni espressione è stata acquisita con due differenti gradi di intensità (normale e forte), fornendo la variabilità necessaria per l'addestramento dei modelli. Sotto il profilo tecnico, i campioni sono stati registrati in ambienti controllati utilizzando microfoni a condensatore di alta qualità e sono distribuiti in formato `.wav` con una frequenza di campionamento di 48 kHz e una risoluzione di 16 bit. Un aspetto rilevante della post-produzione riguarda la normalizzazione dei picchi a -3 dBFS, procedura volta a preservare la naturale varianza della loudness fisiologica tra le diverse emozioni. La gestione computazionale del materiale è facilitata dalla nomenclatura dei file, basata su una sequenza di sette identificatori numerici:

- Modalità (03: audio-only);
- Canale (01: speech);
- Emozione (01: neutra, 02: calma, 03: felicità, 04: tristezza, 05: rabbia, 06: paura, 07: disgusto, 08: sorpresa);
- Intensità (01: normale, 02: forte);
- Ripetizione della prova (01: prima, 02: seconda);
- Attore (01-24: i numeri dispari indicano soggetti maschili, i pari femminili).

Ad esempio, un file etichettato come `03-01-05-01-02-01-12.wav` identifica una traccia audio di parlato relativa alla rabbia, con intensità normale, prodotta da un'attrice (soggetto 12). La validità del corpus è confermata dai

test di ascolto condotti per la validazione percettiva. Per la sola componente vocale, l'accuratezza media di identificazione umana si attesta al 62%, a fronte dell'81% ottenuto nelle prove audiovisive. Tale divario conferma come la privazione del supporto mimico-facciale renda la decodifica prosodica un compito intrinsecamente complesso. L'analisi evidenzia inoltre una marcata variabilità nel riconoscimento delle diverse categorie, mentre la rabbia viene identificata con tassi superiori al 90%, stati come la felicità (44%) e il disgusto (54%) registrano cali drastici, stabilendo un benchmark di riferimento per la valutazione dei sistemi automatici [35]

3.1.2 Dataset EMO-DB

Il Berlin Emotional Database (EMO-DB), sviluppato presso l'Istituto di Scienze della Comunicazione della *Technische Universität Berlin*, costituisce uno dei corpus di riferimento più autorevoli per l'analisi dei parametri acustici in lingua tedesca. Il dataset integra 535 file audio prodotti da 10 attori professionisti (5 uomini e 5 donne) di età compresa tra 21 e 35 anni, impegnati nella recitazione di 10 frasi di uso quotidiano, quali *"Der Lappen liegt auf dem Eisschrank"* o *"Das will sie am Mittwoch abgeben"*. Questa neutralità semantica è fondamentale per garantire che le variazioni acustiche dipendano esclusivamente dallo stato affettivo e non dal significato delle parole. Per indurre risposte fisiologiche autentiche, gli interpreti hanno utilizzato il Metodo Stanislavskij, rievocando esperienze reali legate a sette categorie emotive: rabbia (*Ärger*), noia (*Langeweile*), ansia (*Angst*), felicità (*Freude*), tristezza (*Trauer*), disgusto (*Ekel*) e uno stato neutro. Le registrazioni originali sono state effettuate nella camera anecoica dell'università, un ambiente privo di eco che ha permesso di isolare il segnale da ogni rumore esterno. In questa sede, gli attori potevano muoversi liberamente a una distanza nominale di circa 30 cm dal microfono. Questa mobilità ha però reso variabile la distanza effettiva dal trasduttore, causando instabilità

nell'energia sonora del segnale. Per rimediare a questa criticità e prevenire distorsioni nei picchi di rabbia o la perdita dei sussurri legati alla tristezza, i tecnici hanno regolato manualmente il guadagno durante le sessioni. Questo intervento ha garantito la nitidezza dell'audio, ma ha inevitabilmente alterato il rapporto di intensità naturale tra le diverse emozioni. Per quanto riguarda le specifiche tecniche, il campionamento iniziale a 48 kHz è stato ridotto a 16 kHz con una risoluzione a 16 bit. La gestione dei file è garantita da una nomenclatura sistematica a sette posizioni (es. 03a01Fa.wav) che identifica univocamente l'identità del parlatore, il testo dell'enunciato, l'emozione target e la variante della registrazione. La validità scientifica dell'intero corpus è infine corroborata da un protocollo di validazione percettiva su 20 ascoltatori, il quale ha permesso di selezionare esclusivamente le registrazioni capaci di superare una soglia di riconoscimento umano dell'80% e un indice di naturalezza superiore del 60% . Questa rigorosa procedura di filtraggio garantisce che il database finale sia composto solo da espressioni emotive acusticamente nitide e prototipiche, fornendo così un dataset validato per testare l'accuratezza dei sistemi di classificazione automatica.[36].

3.1.3 Dataset ASDBank

Ai fini dell'applicazione clinica oggetto della presente tesi, la ricerca si è basata sull'analisi del dataset ASDBank Dutch Asymmetries Corpus , con particolare riferimento al sottocorpus SK. Questa risorsa, sviluppata tra il 2007 e il 2012 presso l'Università di Groningen nell'ambito del progetto *"Asymmetries in Grammar"* [37], analizza le asimmetrie tra la produzione e la comprensione linguistica in età evolutiva. La coorte sperimentale comprende 46 bambini con Disturbo dello Spettro Autistico (ASD) e 38 con sviluppo tipico (TD), caratterizzati da un'età media di 9 anni (range 6-12) e una prevalenza maschile rispettivamente dell'87% e del 66%.L'affidabilità

clinica del campione è garantita da un rigoroso protocollo di screening basato sui criteri del DSM-IV-TR e confermato tramite l'impiego dei protocolli ADI-R e ADOS-2. In particolare, dei 51 candidati iniziali (10 casi di autismo, 34 PDD-NOS e 7 Asperger), 29 hanno soddisfatto i criteri di entrambi i test, mentre 14 hanno risposto positivamente solo all'ADI-Re 5 al solo ADOS-2. L'esclusione di tre partecipanti non conformi a tali soglie e di ulteriori due soggetti per anomalie tecniche durante la registrazione ha fissato il campione ASD definitivo a 46 unità. Il gruppo di controllo (TD) è invece composto da soggetti privi di anamnesi psichiatrica o neurologica, i cui punteggi nei test sono risultati ampiamente al di sotto delle soglie di rilevanza clinica. Per quanto concerne la fase di acquisizione, i segnali audio sono stati generati attraverso un compito di narrazione strutturata (structured storytelling). Nello specifico, sono stati impiegati quattro libri di fiabe (protagonisti: ballerina, pirata, principessa e indiano), ciascuno costituito da sei tavole illustrate progettate per sollecitare il topic shift e l'uso di pronomi. La struttura narrativa seguiva uno schema fisso: le prime due tavole presentavano il protagonista, la terza introduceva un secondo personaggio per variare l'argomento, mentre le sequenze successive servivano a verificare se il bambino riuscisse a mantenere il filo del discorso sullo stesso soggetto o a riprendere il protagonista iniziale.

Figura 3.1: Illustrazione dei quattro libri di fiabe mostrati ai bambini durante la sessione di narrazione strutturata. Ciascun libro era composto da sei immagini e un personaggio differente.

Al fine di massimizzare la comunicazione, la configurazione sperimentale ha previsto il coinvolgimento di due operatori con ruoli distinti. Mentre il primo assistente si posizionava accanto al partecipante per guidare la narrazione, il secondo si collocava a debita distanza in modo da simulare la condizione di "ascoltatore cieco". Questa disposizione obbligava il bambino a descrivere la storia in modo esplicito, rendendo il racconto comprensibile anche a chi non poteva osservare le immagini. Le sessioni, condotte presso l'Eye Lab della Facoltà di Lettere di Groningen, hanno prodotto un corpus audio con una durata individuale compresa tra i 7 e i 15 minuti, per un totale di 6.900 secondi di parlato per il gruppo ASD e 5.700 secondi per il gruppo TD. Dal punto di vista tecnico, i segnali sono stati digitalizzati in formato WMA tramite dispositivi Olympus, adottando frequenze di campionamento differenziate: 44.1 kHz per il gruppo ASDe 24 kHz per il gruppo TD. Il post-processing ha previsto una fase di trascrizione ortografica manuale in formato Word, seguita da una revisione sistematica volta a garantire la perfetta coerenza tra segnale audio e testo. Infine, i dati sono stati convertiti e codificati secondo gli standard del sistema CHAT (Codes for the Human Analysis of Transcripts) per l'integrazione nel database TalkBank.

3.1.4 Dataset Sperimentale

Il dataset sperimentale è stato acquisito presso l'ASST di Pavia e coinvolge un campione di 61 partecipanti adulti, suddivisi in 40 individui con diagnosi di autismo ad alto funzionamento e 21 soggetti neurotipici. La distribuzione anagrafica del campione copre una fascia d'età compresa tra i 20 e i 50 anni, con una distribuzione di frequenza maggiormente concentrata nella fascia giovane-adulta (20-30 anni). La procedura di valutazione consiste nell'applicazione del paradigma Short Story Task, basato sulla lettura del racconto "La fine di qualcosa" di Ernest Hemingway. Il protocollo prevede la somministrazione di sedici domande poste in ordine di complessità crescente. Nello specifico, i primi due quesiti accertano la familiarità con l'opera e la capacità di sintesi mentre quelli seguenti approfondiscono l'attitudine del soggetto a interpretare correttamente i pensieri, le emozioni e le intenzioni dei personaggi. Al fine di escludere variabili legate al carico cognitivo o a potenziali deficit della memoria, ad ogni partecipante è stato permesso di consultare liberamente sia il testo narrativo sia l'elenco dei quesiti per l'intera durata della sessione. Le interviste sono state acquisite tramite registrazioni audio mp3 a 48 kHz, con una durata variabile tra i 10 e i 25 minuti in funzione della ricchezza dell'esposizione verbale fornita.

3.2 Pipeline per lo Speech Emotion Recognition

L'intera architettura per il riconoscimento emotivo è stata implementata in ambiente *MATLAB*. L'utilizzo dei relativi *Audio* e *Deep Learning Toolbox* ha permesso una gestione integrata e standardizzata dei dataset RAVDESS ed EMO-DB, garantendo il controllo diretto su ogni fase di elaborazione del segnale e di addestramento del classificatore. La Figura 3.2 illustra la pipeline metodologica seguita.

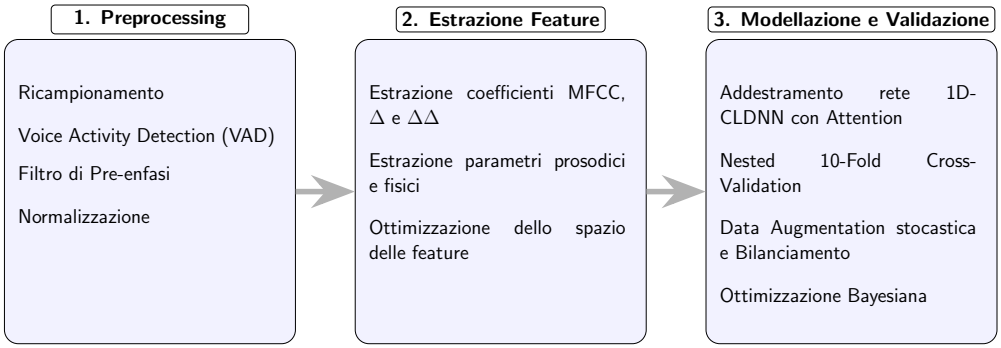


Figura 3.2: Pipeline per il riconoscimento dello stato emotivo implementata in MATLAB.

3.2.1 Pre-processing

La fase di Pre-processing è stata eseguita in modo sequenziale su tutti i segnali audio e si compone di quattro step principali: Ricampionamento, Voice Activity Detection (VAD), Pre-enfasi e Normalizzazione.

Ricampionamento

La fase iniziale di pre-elaborazione ha riguardato l'uniformazione della frequenza di campionamento dei segnali audio a 16 kHz. Tale operazione è stata applicata esclusivamente al corpus RAVDESS, in quanto il database EMO-DB viene fornito dagli autori già conforme a questo standard. La scelta rispetta i protocolli dello speech processing e trova riscontro nella natura acustica del parlato umano, le cui componenti fonetico-prosodiche più rilevanti risiedono nella banda di frequenze comprese tra 50 Hz e 7500 Hz. Dal punto di vista analitico, l'affidabilità della discretizzazione è garantita dal teorema di Nyquist-Shannon, secondo cui la frequenza di campionamento f_s deve essere pari almeno al doppio della massima frequenza f_{max} contenuta nel segnale originale.

$$f_s \geq 2 \cdot f_{max} \tag{3.1}$$

Il limite di Nyquist a 8 kHz assicura dunque la conservazione dei tratti paralinguistici e la rimozione delle componenti spettrali elevate prive di valore informativo [18].

Voice Activity Detection (VAD)

L'isolamento dei segmenti vocali rappresenta una fase determinante per garantire l'estrazione di feature affidabili. Sebbene i dataset EMO-DB e RAVDESS derivino da ambienti controllati, gli enunciati presentano sistematicamente intervalli di silenzio iniziali e finali che rischiano di alterare i parametri statistici, introducendo bias indipendenti dalla componente emotiva. Per ovviare a questa criticità si è utilizzato un algoritmo di Voice Activity Detection (VAD) basato sulla funzione *detectSpeech* di MATLAB, la quale implementa la metodologia proposta da Giannakopoulos [38]. Il processo ha inizio con la conversione del segnale audio in uno spettrogramma attraverso operazioni di finestratura e sovrapposizione. Per ogni fotogramma l'algoritmo calcola l'energia a breve termine e lo spread spettrale per misurare la distribuzione della potenza sonora tra le diverse frequenze. Tale analisi permette di discriminare il parlato dal rumore di fondo poiché la voce concentra l'energia in bande specifiche, diversamente dalle componenti di disturbo che presentano una distribuzione spettrale uniforme. Il sistema genera quindi gli istogrammi dei due parametri per ricavare le soglie adattive T basate sui primi due massimi locali M_1 e M_2 secondo la relazione matematica

$$T = \frac{W \cdot M_1 + M_2}{W + 1} \tag{3.2}$$

nella quale il peso W assume un valore pari a 5. Prima della classificazione definitiva, i valori di energia e spread vengono regolarizzati nel tempo attraverso successivi filtri mediani mobili a cinque elementi. Un fotogramma

viene classificato come attività vocale solo se entrambi i parametri superano contemporaneamente le rispettive soglie. Le regioni identificate vengono infine accorpate se la distanza temporale che le separa è inferiore a una *MergeDistance* di 50 ms. Questa operazione garantisce la continuità degli enunciati ed evita che le micro pause fisiologiche frammentino il segnale. La maschera risultante definisce i punti di taglio per il *trimming* isolando l'intervallo di attività rilevata tra il primo e l'ultimo indice utile. Qualora la dinamica vocale risulti troppo debole per generare una maschera valida, il sistema attiva un controllo di sicurezza che mantiene il file originale integro per prevenire la perdita di campioni informativi durante l'elaborazione automatica.

Pre-enfasi

Ottenuto il solo segnale vocale utile tramite il VAD, si procede all'applicazione di un filtro di pre-enfasi. Nella voce umana, l'energia tende a concentrarsi nelle basse frequenze, creando una pendenza spettrale che può oscurare le componenti medie e alte, fondamentali per distinguere i tratti emotivi. L'impiego di un filtro passa alto permette di attenuare questa pendenza, restituendo dunque uno spettro più equilibrato. L'operazione viene eseguita mediante la funzione filter di Matlab che applica al segnale $s(n)$ la seguente equazione alle differenze:

$$y(n) = s(n) - \alpha s(n-1) \tag{3.3}$$

dove α rappresenta il coefficiente di pre-enfasi, il cui valore è generalmente compreso nell'intervallo $0.9 \leq \alpha \leq 1$ [18]. In questa ricerca, il coefficiente α è stato impostato a 0.97, scelta che permette di compensare adeguatamente la pendenza spettrale naturale, mettendo in risalto le componenti armoniche e le formanti [18]. La risposta in frequenza del filtro risultante è

mostrata in Figura 3.3

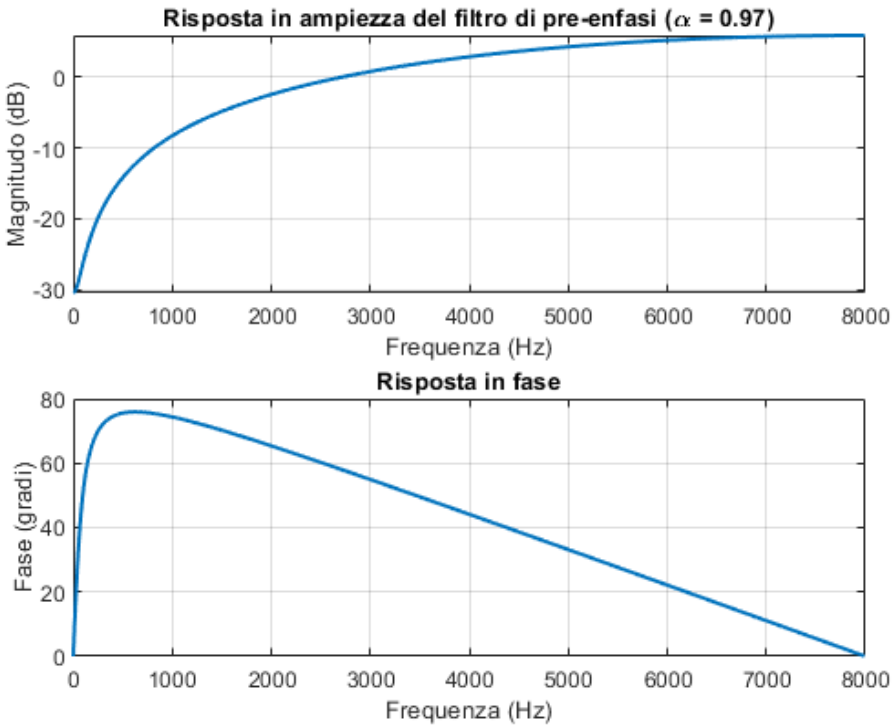


Figura 3.3: Filtro preenfasi con $\alpha = 0.97$

Normalizzazione

A completamento della fase di pre-elaborazione, il segnale viene sottoposto a una procedura di normalizzazione del picco (*Peak Normalization*). Questa operazione è fondamentale per uniformare i livelli di ampiezza e garantire che le variazioni di volume non introducano bias nell'estrazione delle feature acustiche [20]. Il segnale $y_{norm}(n)$ viene calcolato come:

$$y_{norm}(n) = \frac{y(n)}{\max(|y(n)|) + \varepsilon} \tag{3.4}$$

dove il termine $\varepsilon = 10^{-6}$ garantisce la stabilità numerica del sistema prevenendo divisioni per zero. Tale passaggio confina il segnale nell'intervallo standard $[-1, 1]$. In Figura 3.4 viene illustrata l'evoluzione temporale della

forma d'onda di un campione audio del dataset EMO-DB durante le diverse fasi della catena di pre-elaborazione implementata

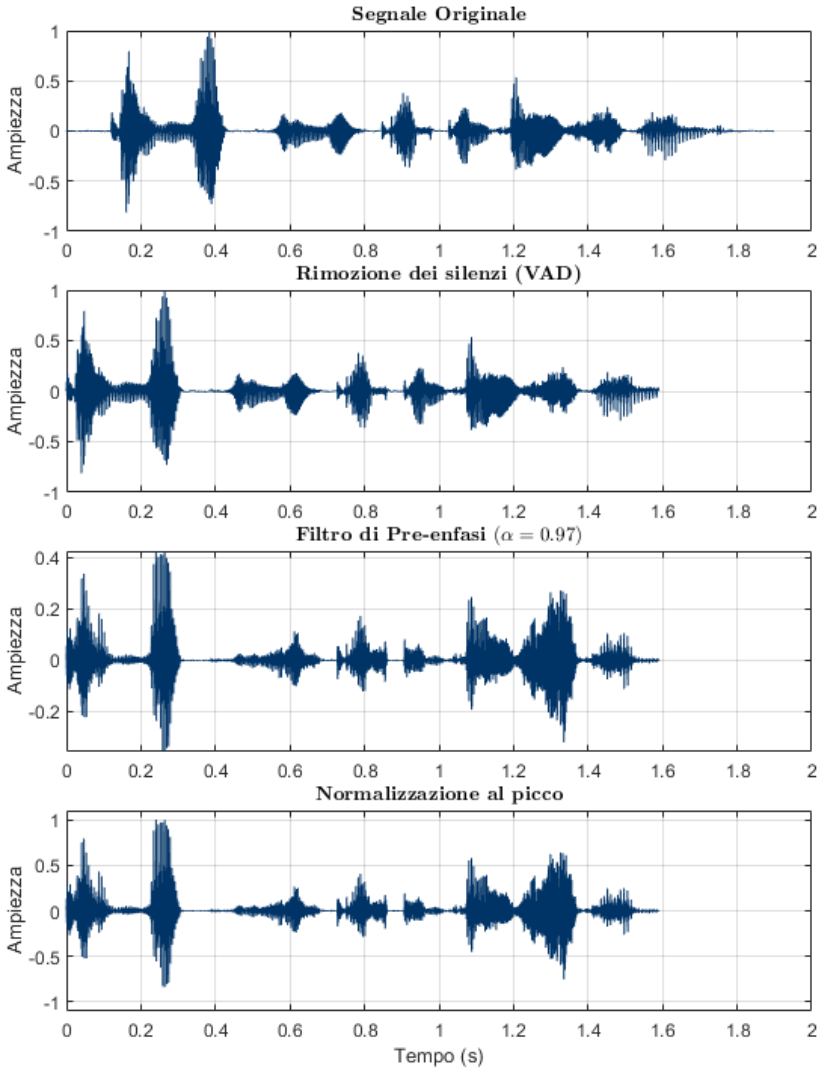


Figura 3.4: Esempio preprocessing file audio EMODB

3.2.2 Estrazione delle feature

L'estrazione delle feature mira a ridurre la dimensionalità del segnale audio, mantenendo esclusivamente le informazioni paralinguistiche ed emotive rilevanti per la classificazione. Tale processo avviene convertendo la forma d'onda tempo-variante in una serie di vettori di caratteristiche che ne descrivono le proprietà spettrali (timbro) e temporali (prosodia).

Su questa base, la metodologia proposta effettua un confronto tra diverse configurazioni di descrittori, così da identificare il set di feature più robusto per il riconoscimento in scenari cross-dataset.

Framing

Per poter estrarre correttamente questi parametri risulta necessario gestire la natura intrinsecamente non stazionaria del parlato, le cui proprietà statistiche evolvono rapidamente in funzione della dinamica articolatoria. Il ricorso alla tecnica del framing, già presentata nel Capitolo 2, consente dunque di isolare brevi intervalli entro i quali sia possibile assumere una condizione di quasi stazionarietà [18]. Nel presente studio è stata adottata una durata dei frame di 16 ms, che alla frequenza di campionamento di 16 kHz corrisponde a una lunghezza di 256 campioni Questa scelta garantisce un equilibrio ottimale tra la risoluzione temporale, necessaria per tracciare le rapide transizioni fonetiche, e la risoluzione frequenziale, indispensabile per una stima stabile dell'inviluppo spettrale. Al fine di mitigare le distorsioni spettrali ai bordi dei segmenti, ad ogni intervallo viene applicata la finestra di Hamming. Tale funzione di pesatura riduce la dispersione energetica e facilita l'isolamento delle formanti vocali, rendendo più nitida la firma acustica del segnale. Tuttavia, essa attenua forzatamente l'ampiezza alle estremità di ogni frame. Per questa ragione, la procedura include un overlap del 50%, pari a 128 campioni. Questa sovrapposizione compensa l'attenuazione ai margini e assicura la fluidità delle traiettorie acustiche, preservando le micro-variazioni prosodiche che risultano determinanti per il riconoscimento dello stato emotivo.

Implementazione e configurazione del set di feature baseline

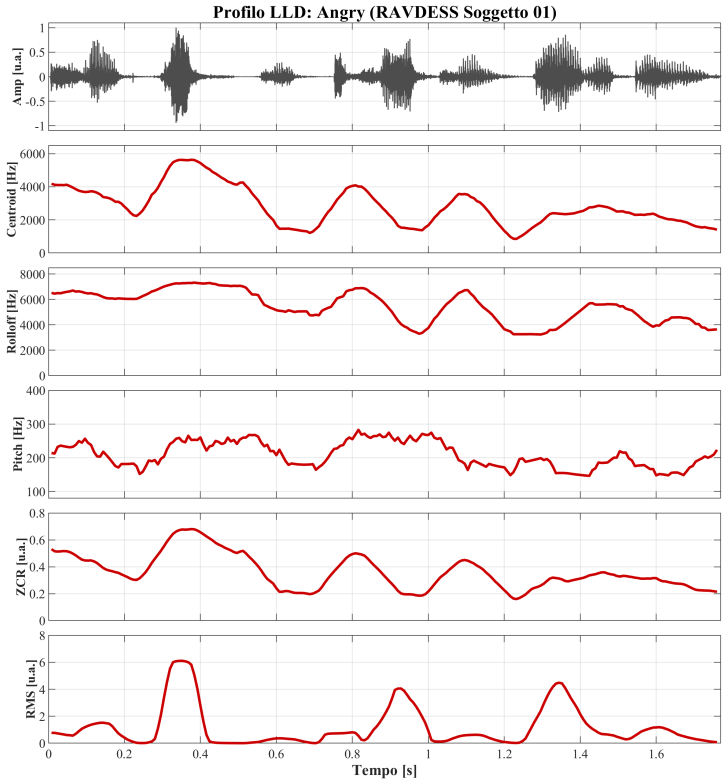
Una volta conclusa la segmentazione del segnale, l'analisi si focalizza sulla definizione di un vettore di caratteristiche capace di restituire una

descrizione numerica coerente dell'impronta emotiva. La configurazione adottata prevede un set di 65 descrittori acustici strutturato per bilanciare la rappresentazione dello spettro vocale con la dinamica temporale della voce. Questa strategia non si basa su una selezione arbitraria ma integra i coefficienti cepstrali con i parametri fisici previsti dal protocollo eGeMAPS [19]. Nello specifico, il set è costituito da 60 parametri Mel-Frequency Cepstral Coefficients (MFCC), ripartiti in un gruppo di 20 coefficienti statici ai quali vengono affiancate le rispettive derivate del primo e del secondo ordine ($\Delta - \Delta\Delta$). La scelta di limitare i coefficienti statici a venti unità permette di isolare l'inviluppo spettrale legato alle prime quattro formanti vocali. In questo modo la procedura trascura i dettagli della sorgente laringea che potrebbero introdurre variabilità non desiderata tra i diversi soggetti [18]. L'integrazione delle componenti dinamiche cattura invece la velocità e l'accelerazione dei cambiamenti timbrici e riflette lo stato di attivazione fisiologica del soggetto[18]. A completamento della firma acustica, sono stati selezionati cinque Low-Level Descriptors (LLD) estratti in conformità con lo standard eGeMAPS [19] . Il vettore include Pitch, energia (RMS), ZCR, Spectral Centroid e Spectral Rolloff. Mentre la frequenza fondamentale e l'energia descrivono lo sforzo vocale e la tensione laringea, i descrittori spettrali permettono di quantificare la brillantezza del timbro, un parametro fondamentale per distinguere stati emotivi a valenza contrastante . Le restanti caratteristiche del protocollo sono state escluse poiché ridondanti rispetto alle informazioni già codificate nel dominio cepstrale. Contestualmente, si è proceduto all'omissione di Jitter e Shimmer in quanto sono eccessivamente sensibili alla qualità dei microfoni e alle asimmetrie fonetiche tra i corpus RAVDESS ed EMO-DB [39].

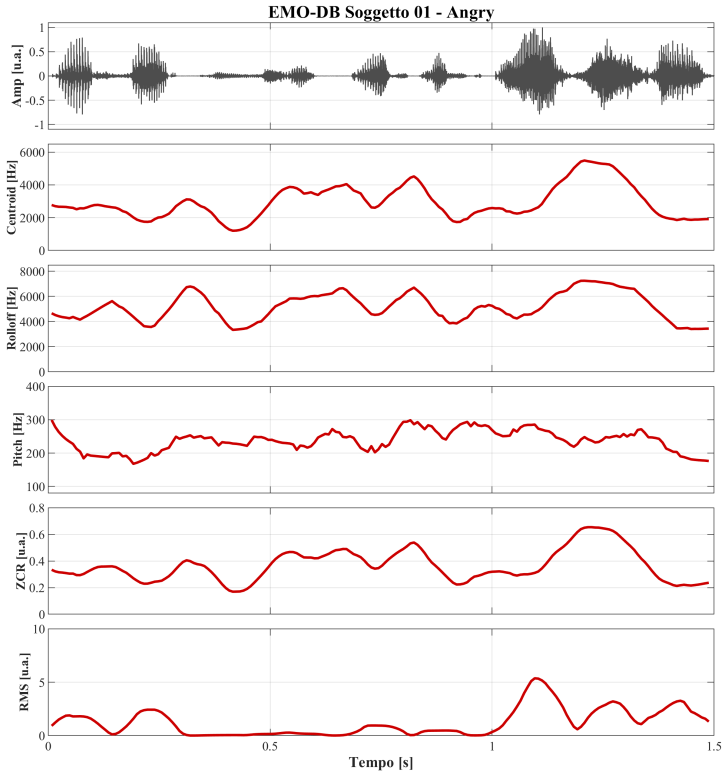
Al fine di validare la coerenza del processo di estrazione, nelle figure seguenti vengono analizzate le rappresentazioni tempo-frequenza e i profili LLD (Low-Level Descriptors) scelti come base per l'addestramento della

rete. La Figura 3.5 mette a confronto i descrittori estratti da un soggetto di lingua inglese appartenente al dataset RAVDESS e uno di lingua tedesca proveniente dal corpus EMO-DB, evidenziando una sostanziale analogia nelle componenti fisiche del segnale. Nonostante le divergenze fonetiche i grafici mostrano un'energia con picchi compresi tra 5 e 6 unità arbitrarie e una frequenza fondamentale (F_0) stabilizzata nella banda tra 200 e 300 Hz per entrambi i campioni. Inoltre, i valori elevati di Spectral Centroid e Spectral Rolloff indicano la presenza del timbro aspro tipico degli stati di rabbia. La coerenza di tali evidenze motiva l'adozione di modelli di Deep Learning per l'identificazione di pattern trasversali necessari a una efficace generalizzazione tra dataset differenti.

Per quanto concerne i parametri cepstrali la Figura 3.6 riporta gli spettrogrammi e le relative heatmap estratti da un soggetto del dataset EMO-DB.Le matrici dimostrano l'efficacia del vettore nel separare le informazioni statiche da quelle dinamiche. La metà superiore di ogni heatmap (i 20 coefficienti statici) presenta zone cromatiche ampie che mappano la postura base del tratto vocale. La metà inferiore (le derivate $\Delta - \Delta\Delta$) appare invece densa e frastagliata, poiché traccia i rapidi cambiamenti articolatori e le accelerazioni della voce. La combinazione di queste due componenti assicura un'impronta acustica completa.



(a) Profilo LLD: Angry (RAVDESS)



(b) Profilo LLD: Angry (EMO-DB)

Figura 3.5: Confronto cross-corpus della firma acustica per la classe emozionale "Anger".

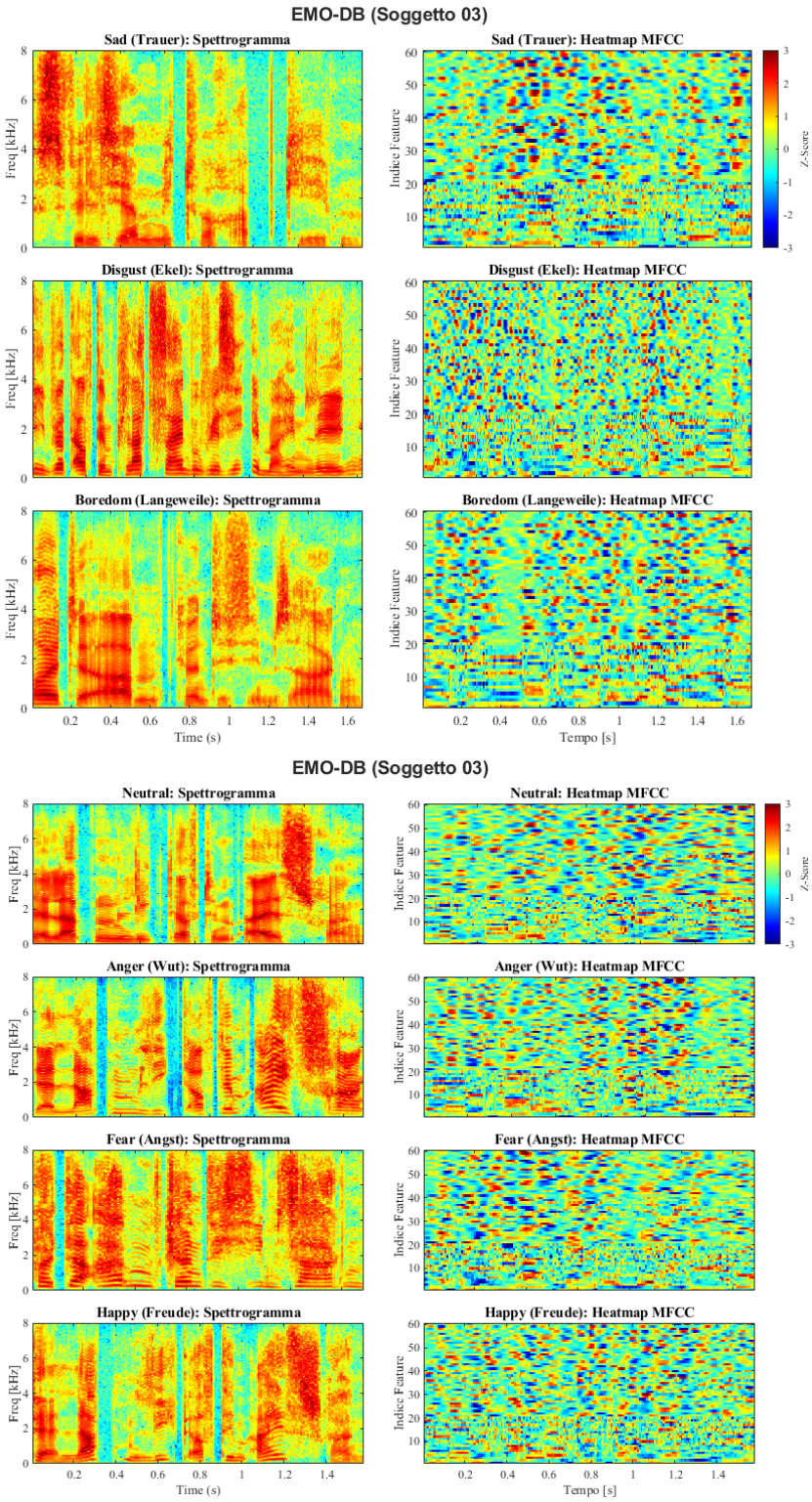


Figura 3.6: Rappresentazione tempo-frequenza per il dataset EMO-DB. Per ogni classe è riportato lo spettrogramma (sinistra) e la relativa heatmap dei 60 coefficienti MFCC, comprensivi delle derivate Δ e $\Delta\Delta$ (destra).

La medesima procedura di estrazione è stata infine applicata in modo sistematico e automatizzato su tutti i file audio. Questa elaborazione ha generato il set di dati strutturato e omogeneo che costituirà l'input per i modelli di Deep Learning descritti nei paragrafi successivi.

Selezione delle Feature tramite Diagnostic Feature Designer

Oltre al set di 65 feature estratte sulla base dei principali descrittori acustici consolidati in letteratura, si è scelto di procedere con una fase di ottimizzazione volta all'individuazione dei parametri dotati della massima capacità discriminante per i dataset RAVDESS ed EMO-DB. A tale scopo è stato impiegato il Diagnostic Feature Designer di MATLAB [40], un applicativo progettato per l'analisi massiva di segnali e l'estrazione automatizzata di descrittori matematici. La sua funzione primaria è l'identificazione di "indicatori di condizione" (condition indicators) capaci di distinguere tra stati diversi di un sistema (in questo caso, le diverse attivazioni emotive) partendo dall'analisi di grandi volumi di dati (ensemble data). Il procedimento adottato segue fedelmente la pipeline documentata da MathWorks, la quale integra sequenzialmente l'importazione dei segnali, il pre-processing e la generazione delle feature. Tale flusso operativo è sintetizzato nello schema a blocchi riportato di seguito (Figura: 3.7):

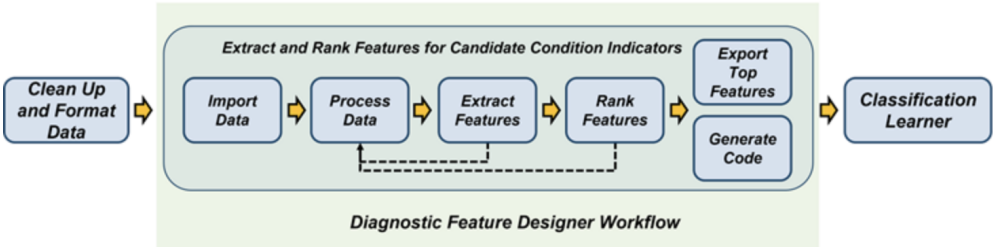


Figura 3.7: Workflow Diagnostc Feature Designer [40]

I dataset RAVDESS ed EMO-DB sono stati valutati separatamente. I segnali, precedentemente sottoposti alla fase di pre-elaborazione, sono stati

importati nell'applicativo mantenendo la segmentazione in frame di 16 ms con un overlap del 50%. Inizialmente sono state estratte le feature temporali tra cui media, deviazione standard, skewness e kurtosis per mappare l'andamento statistico della forma d'onda. Successivamente, per l'analisi nel dominio delle frequenze, è stata calcolata la densità spettrale di potenza tramite il metodo di Welch. Questa tecnica risulta preferibile rispetto al calcolo spettrale classico poiché, attraverso la media di segmenti sovrapposti, fornisce una stima della distribuzione energetica molto più stabile e resiliente al rumore stocastico. Da questa elaborazione sono state derivate le feature spettrali, comprese le ampiezze e le frequenze dei primi cinque picchi. L'efficacia di questi parametri è stata validata tramite un algoritmo di ranking supervisionato basato sul test One-way ANOVA . Il metodo assegna a ogni caratteristica la statistica F, una metrica che esprime il rapporto tra la varianza tra le diverse classi emotive e quella registrata all'interno delle classi stesse. Un punteggio F elevato identifica una feature capace di separare nettamente le etichette emotive, minimizzando le zone di ambiguità acustica. I risultati dei due ranking sono riportati nella Tabella: 3.1

Tabella 3.1: Ranking delle feature estratte per EMO-DB e RAVDESS in base all'F-Score.

Rank	**EMO-DB (F-Score)**	**RAVDESS (F-Score)**
1	SNR (744.47)	Std (695.06)
2	PeakFreq1 (694.37)	RMS (694.81)
3	PeakValue (648.77)	PeakValue (559.79)
4	Std (577.36)	PeakFreq1 (425.24)
5	SINAD (575.79)	PeakAmp3 (383.65)
6	RMS (574.72)	PeakAmp2 (382.82)
7	Skewness (486.42)	PeakAmp1 (380.42)
8	PeakFreq2 (456.09)	BandPower (372.62)
9	PeakFreq3 (359.10)	Skewness (352.85)
10	CrestFactor (287.72)	PeakAmp4 (335.82)
11	PeakAmp3 (244.44)	PeakAmp5 (284.97)
12	PeakAmp2 (220.31)	CrestFactor (200.44)
13	PeakAmp4 (219.58)	PeakFreq2 (195.63)
14	ShapeFactor (198.98)	PeakFreq3 (180.32)
15	BandPower (181.99)	ImpulseFactor (144.16)
...	...	...
Ultimi	Mean (4.67)	SNR (0.00)
Ultimi	ClearanceFactor (1.66)	SINAD (0.00)

Al fine di massimizzare la capacità di generalizzazione del sistema e mitigare i bias intrinseci dei singoli corpus,la selezione delle caratteristiche si è basata sulla stabilità dei parametri comuni a entrambi i dataset. Tale approccio ha portato alla definizione di una nuova configurazione, sviluppata a partire dai 20 MFCC e dalle relative derivate del primo ordine Δ. A questi 40 descrittori base sono stati integrati otto parametri aggiuntivi per garantire una mappatura fisica completa. Nel dominio del tempo, la Deviazione Standard (Std), la RMS e il PeakValue quantificano l'intensità e l'inviluppo energetico, mentre la Skewness modella l'asimmetria delle ampiezze. Il Crest Factor, definito dal rapporto tra valore di picco e valore efficace (RMS), misura l'impulsività del segnale. Tale descrittore permette di isolare i transienti acustici tipici di emozioni ad alta attivazione, come rabbia o sorpresa, in cui i picchi di ampiezza si discostano significativamente dall'energia media della finestra temporale. L'analisi spettrale è completata dalla BandPower e dalle prime due frequenze di picco (PeakFreq1 e PeakFreq2), che descrivono le risonanze laringee con una precisione superiore rispetto alla sola individuazione della frequenza dominante. La sintesi delle due configurazioni che verranno utilizzate per l'addestramento della rete è riportata nella Tabella 3.2 .

Tabella 3.2: Sintesi delle configurazioni e composizione dei set di feature utilizzati per l'addestramento.

Configurazione	**N. Feature**	**Elenco Feature**
Set Baseline	65	20 MFCC, 20 Δ, 20 ΔΔ, Pitch, Spectral Centroid, Spectral Rolloff, ZCR, RMS
Set ANOVA-8	48	20 MFCC, 20 Δ, Std, RMS, PeakValue, PeakFreq1, Skewness, PeakFreq2, Crest Factor, BandPower

3.2.3 Descrizione della Rete: 1D-CLDNN con Self-Attention

L'architettura implementata segue il paradigma 1D-CLDNN (Convolutional LSTM Deep Neural Network) [41], integrato con un meccanismo di Self-Attention [27]. La rete è strutturata gerarchicamente per estrarre pattern locali, modellare le dipendenze temporali e pesare la rilevanza informativa dei diversi segmenti del segnale vocale. La scelta di questa specifica topologia risponde alla necessità di elaborare dati in serie temporale complessi, dove l'informazione affettiva è codificata sia in micro-variazioni spettrali sia nell'evoluzione prosodica globale.

Il front-end dell'architettura analizza la matrice di input attraverso cinque stadi denominati Local Feature Learning Blocks (LFLB).Ciascun blocco impiega layer di convoluzione monodimensionale (1D-CNN) per intercettare le correlazioni multidimensionali tra coefficienti MFCC variazioni di pitch e parametri statistici. La configurazione dei filtri segue una logica di astrazione progressiva. Nei primi tre blocchi, l'adozione di kernel di dimensione 5 e un numero elevato di filtri (256 e 128) permette di mappare l'eterogeneità del segnale su contesti temporali estesi. Al contrario, negli stadi finali, la contrazione a 64 filtri e l'uso di kernel di dimensione 3 forzano il modello verso una sintesi informativa. Questo processo è essenziale per eliminare le ridondanze acustiche e isolare esclusivamente i marker emotivi più discriminanti [22].

L'efficacia di questa struttura profonda è strettamente legata alla stabilità del processo di addestramento, garantita da una precisa sequenza di operazioni di regolarizzazione e attivazione. Poiché la rete apprende tramite la retropropagazione dell'errore (backpropagation), il gradiente (ovvero il segnale di correzione) deve fluire dall'output verso l'input per aggiornare i pesi. Nelle architetture stratificate, tuttavia, questo segnale tende a ridursi esponenzialmente man mano che risale i layer, portando al fenomeno del

gradiente evanescente (vanishing gradient) che rischia di bloccare l'apprendimento dei primi blocchi convoluzionali [42]. Per ovviare a tale criticità, ogni convoluzione è seguita dalla Batch Normalization (BN), uno strato che normalizza le attivazioni di ogni mini-batch affinché mantengano media nulla e varianza unitaria. Analiticamente, la BN impedisce ai segnali di ricadere nelle zone di saturazione delle funzioni di attivazione, assicurando che il gradiente conservi un'intensità sufficiente per propagarsi lungo tutta la profondità della rete, accelerando così la convergenza [43] Subito dopo la stabilizzazione interviene il layer ReLU (Rectified Linear Unit). La sua funzione di attivazione,

$$f(x) = \max(0, x) \tag{3.5}$$

introduce la non-linearità necessaria per interpretare segnali complessi come la voce umana. La ReLU agisce come un filtro decisionale che azzera i contributi negativi, interpretati come rumore, e propaga solo le attivazioni positive, facilitando ulteriormente il flusso dei gradienti. Il blocco LFLB si conclude con il Max-Pooling 1D, un'operazione di sottocampionatura che seleziona il valore massimo in una finestra di scorrimento. Oltre a ridurre la dimensionalità dei dati e il carico computazionale, il pooling conferisce alla rete l'invarianza alle traslazioni. In termini pratici, ciò assicura che un evento acustico importante,come un improvviso aumento di energia tipico della rabbia, venga intercettato correttamente dal sistema a prescindere dalla sua esatta posizione temporale nel frame, aumentando la robustezza complessiva del classificatore.

L'output spaziale così generato costituisce l'input per lo stadio LSTM, deputato alla modellazione delle dinamiche temporali a lungo termine. Essa gestisce il flusso informativo tramite una cella di memoria (C_t) e tre gate logici che ne determinano matematicamente lo stato [44]. Il primo

stadio è il Forget Gate (f_t) che analizza lo stato precedente h_{t-1} e l'input attuale x_t per decidere quali informazioni della vecchia memoria debbano essere rimosse:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{3.6}$$

Successivamente l'Input Gate (i_t) seleziona i nuovi dati da integrare nello stato della cella ((C_t) mentre un layer tanh genera un vettore di valori candidati $\tilde{C}_t$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{3.7}$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{3.8}$$

Lo Stato della cella (C_t) viene aggiornato combinando la memoria residua e il nuovo contributo

$$C_t = f_t \circ C_{t-1} + i_t \circ \tilde{C}t \tag{3.9}$$

Infine l' output Gate (o_t) filtra l'informazione da trasmettere come stato nascosto (h_t) ovvero la rappresentazione temporale che verrà trasmessa ai layer successivi:

$$o_t = \sigma(W_o \cdot [ht-1, x_t] + b_o) \tag{3.10}$$

$$h_t = o_t \circ \tanh(C_t) \tag{3.11}$$

A valle della componente ricorrente, l'integrazione di un modulo di Multi-head Self-Attention a due teste permette di pesare selettivamente la salienza di ogni frame [27]. Il layer proietta l'input negli spazi vettoriali di Query (Q), Key (K) e Value (V), calcolando la rilevanza tramite la seguente formula:

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3.12}$$

Questo meccanismo forza la rete a focalizzarsi sui segmenti acustici emotivamente rilevanti, attenuando l'impatto di porzioni non informative. Per

contrastare l'overfitting, è stato inserito un layer di Dropout, che disattiva casualmente una frazione di neuroni durante l'addestramento e forza l'apprendimento di rappresentazioni più robuste. Segue poi il Global Average Pooling 1D (GAP), che sintetizza la sequenza calcolandone la media globale in modo da ridurre drasticamente il numero di parametri. La classificazione definitiva è demandata ad un layer Fully Connected con attivazione Softmax, che trasforma i dati estratti in una distribuzione di probabilità sulle classi emotive. Infine, l'addestramento specifico sul dataset EMO-DB ha richiesto l'implementazione della tecnica dei Class Weights. Questo approccio riequilibra la funzione di costo e compensa il naturale sbilanciamento statistico dei campioni, con l'obiettivo di assicurare prestazioni eque su tutte le categorie emotive. Di seguito sono riportati lo schema a blocchi (Figura: 3.8) e la Tabella 3.3 riassuntiva delle specifiche tecniche della rete implementata.

Tabella 3.3: Specifiche tecniche e parametri dell'architettura 1D-CLDNN con Attention.

Layer	**Configurazione e Parametri**
LFLB 1	Conv1D: 256 filtri, kernel 5, strides 1 BatchNormalization, ReLU MaxPool1D: size 5, strides 2
LFLB 2	Conv1D: 128 filtri, kernel 5, strides 1 BatchNormalization, ReLU MaxPool1D: size 5, strides 2
LFLB 3	Conv1D: 128 filtri, kernel 5, strides 1 BatchNormalization, ReLU MaxPool1D: size 5, strides 2
LFLB 4	Conv1D: 64 filtri, kernel 3, strides 1 BatchNormalization, ReLU MaxPool1D: size 5, strides 2
LFLB 5	Conv1D: 64 filtri, kernel 3, strides 1 BatchNormalization, ReLU MaxPool1D: size 3, strides 2
LSTM	Unità ricorrenti: 64
Attention	Multi-head Self-Attention: 2 teste
Dropout	Rate: 0.3 – 0.5
GAP 1D	Global Average Pooling
Output	Dense Layer: 8 (RAVDESS) / 7 (EMO-DB) Attivazione: Softmax

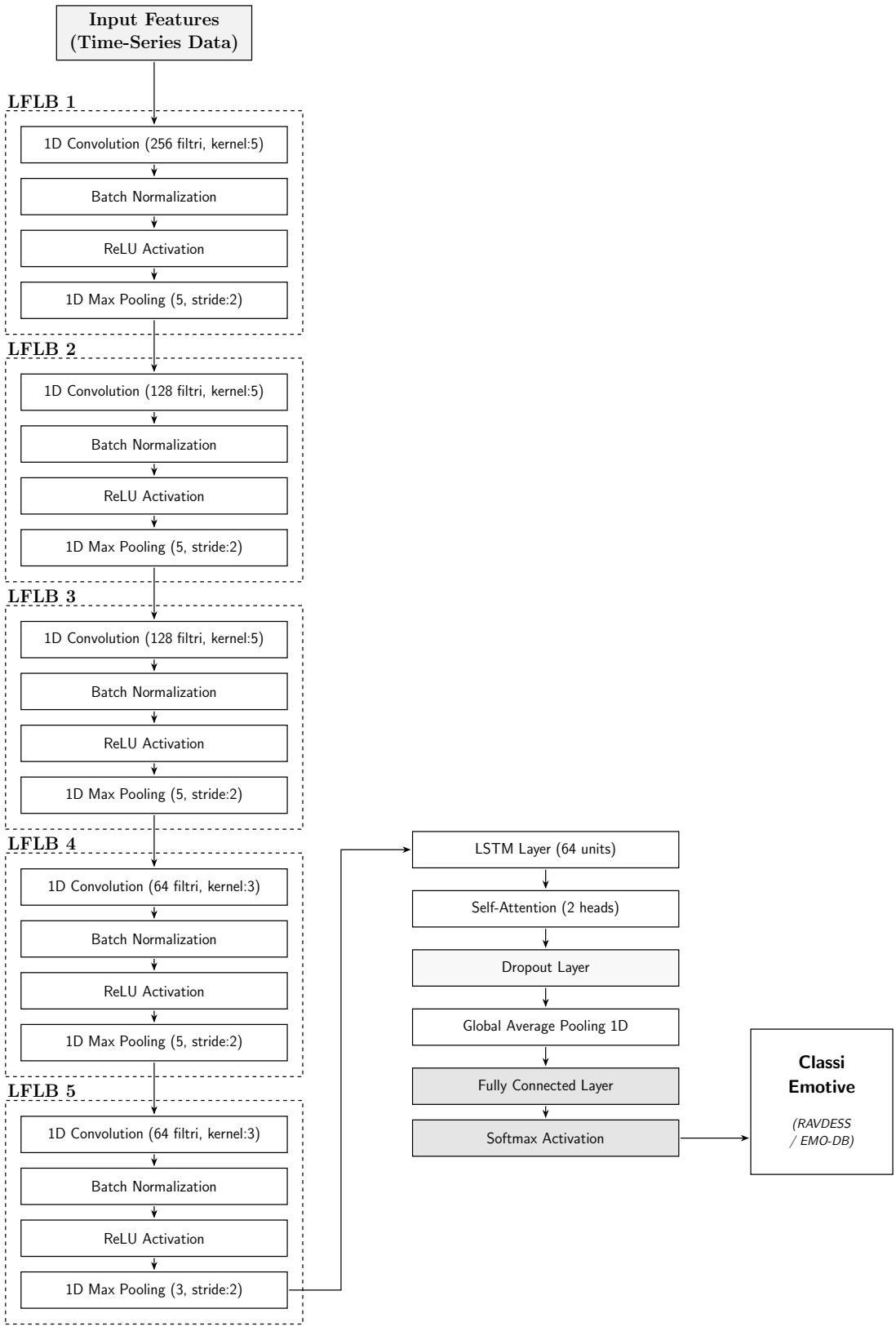


Figura 3.8: Schema a blocchi dell'architettura 1D-CLDNN con modulo di Self-Attention.

3.2.4 Metodologia Sperimentale e Validazione

L'efficacia di un sistema di *Speech Emotion Recognition* (SER) non dipende esclusivamente dalla struttura del classificatore, ma anche dalla robustezza della pipeline di addestramento. In questa sezione vengono descritte le strategie di *Data Augmentation*, i metodi di bilanciamento delle classi, l'ottimizzazione bayesiana e il protocollo di validazione implementati per garantire l'affidabilità del modello sui dataset di riferimento.

Data Augmentation Stocastica

La ridotta estensione campionaria tipica del RAVDESS e EMO-DB ha suggerito l'adozione di protocolli di regolarizzazione volti a contrastare il rischio di overfitting e a consolidare la capacità di generalizzazione della rete. In questa prospettiva si è strutturata una pipeline di Data Augmentation stocastica mediante la funzione MATLAB *audioDataAugmenter*, la quale interviene direttamente sulle forme d'onda grezze. L'espansione del set di addestramento avviene attraverso la generazione di tre varianti sintetiche per ogni traccia audio. Grazie alla configurazione a eventi indipendenti, le trasformazioni acustiche non seguono uno schema deterministico. Al contrario, ogni replica ha una probabilità dell'80% di subire una traslazione della frequenza fondamentale (pitch shifting) e una distinta probabilità dell'80% di perturbazione tramite l'addizione di rumore bianco gaussiano (AWGN). L'algoritmo genera dunque 3 tipologie di cloni:campioni alterati esclusivamente nel pitch, campioni affetti solo da rumore additivo e varianti che presentano la sovrapposizione simultanea di entrambe le distorsioni (Figure 3.9 e 3.10). Tale eterogeneità forza l'architettura 1D-CLDNN ad apprendere rappresentazioni acustiche robuste, indipendenti dal tipo di interferenza presente [22]. Dal punto di vista bioacustico, la manipolazione del segnale è stata circoscritta a parametri che non corrompono l'integrità

prosodica dell'enunciato. Il pitch shifting è contenuto entro l'intervallo di ± 2 semitoni [45]; variazioni superiori indurrebbero distorsioni formantiche tali da invalidare l'identità del parlatore o snaturare l'emozione espressa. Analogamente, il rapporto segnale rumore SNR è stato definito tra i 15 e i 30 dB. Il limite inferiore di 15 dB simula un'interferenza ambientale udibile ma non distruttiva, fondamentale per preservare le componenti spettrali delle emozioni a bassa attivazione, come la tristezza. Infine, si è scelto di escludere trasformazioni quali il time stretching e il volume adjustment, in quanto la velocità del parlato e l'intensità sonora non rappresentano variabili di disturbo, bensì sono i descrittori primari dell'arousal e della valenza emotiva. In accordo con i modelli di Scherer [46], la modifica artificiale di questi parametri altererebbe la natura stessa dell'emozione, privando lo stadio LSTM dei pattern sequenziali necessari per una corretta classificazione.

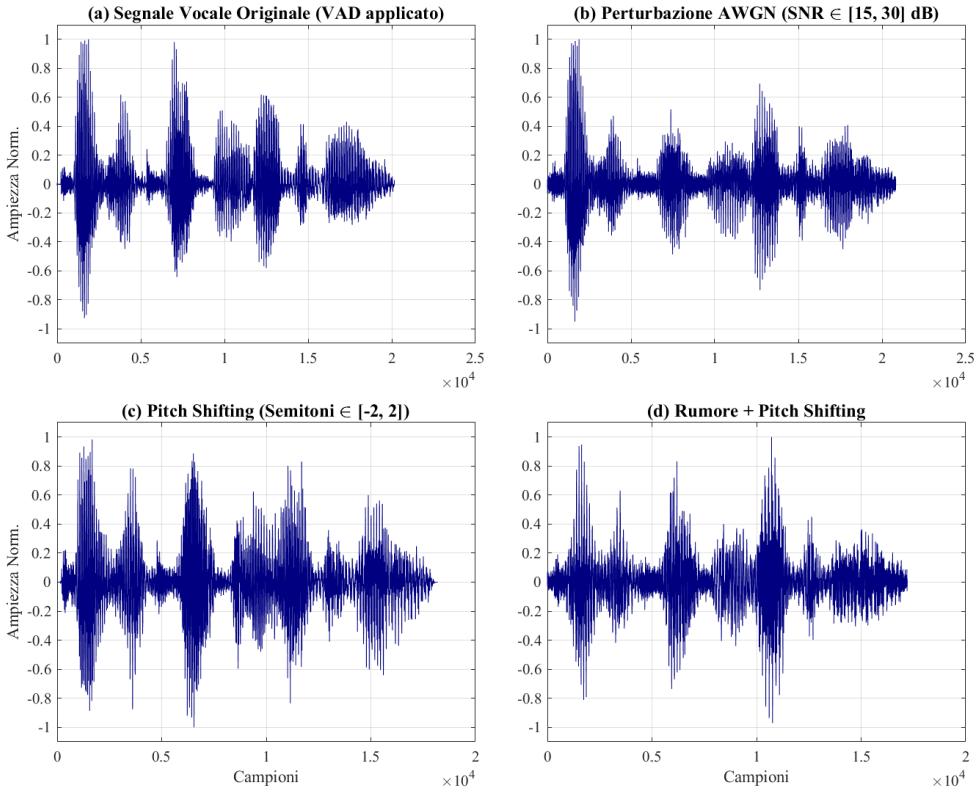


Figura 3.9: Esempio di data augmentation stocastica applicata a un segnale del dataset RAVDESS. Il layout mostra l'integrità dell'onda sonora originale (a) e l'impatto delle tre tipologie di distorsione acustica (b, c, d).

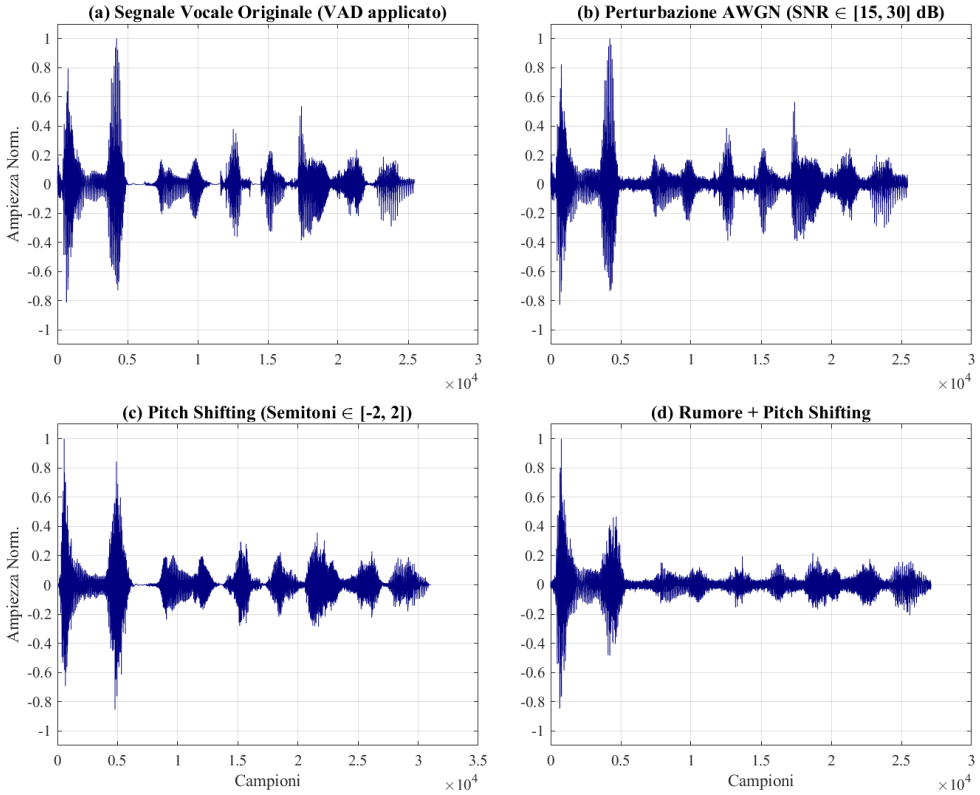


Figura 3.10: Visualizzazione dei processi di perturbazione su un enunciato del dataset EMO-DB. Il grafico evidenzia la preservazione dei pattern energetici della voce a seguito dell'introduzione di rumore AWGN e della traslazione del *pitch*.

Bilanciamento delle classi

La distribuzione non uniforme dei campioni tra le categorie emotive può indurre il classificatore a prevedere più spesso le classi maggioritarie. Come evidenziato nelle descrizioni dei dataset, sia il corpus RAVDESS che l'EMO-DB presentano squilibri strutturali che hanno richiesto l'adozione di strategie di bilanciamento differenti. Il dataset RAVDESS ha una struttura bilanciata per quasi tutte le classi, in cui ogni emozione è rappresentata da due livelli di intensità acustica (Normal e Strong). Tuttavia, la classe Neutral costituisce un'eccezione poiché, per sua natura, non prevede una variante ad alta intensità. Come si evince dalla Figura 3.11, tale peculiarità determina una numerosità del neutro esattamente dimezzata rispetto alle altre emozioni.

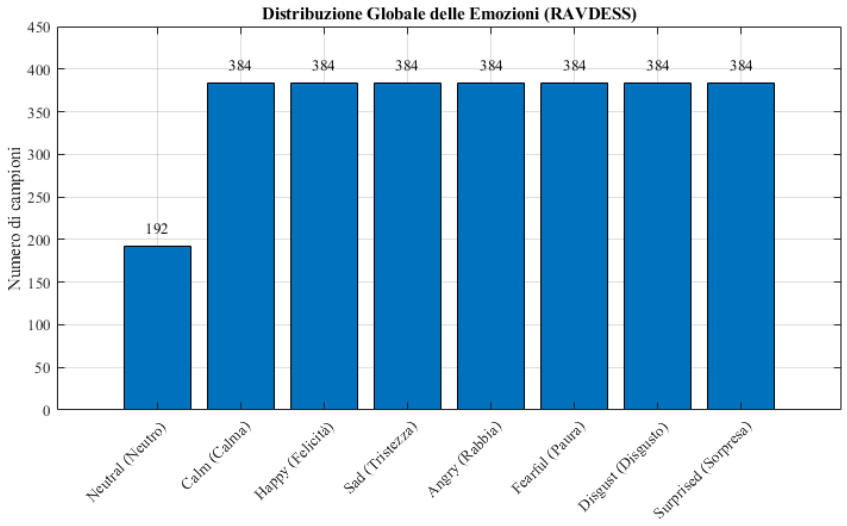


Figura 3.11: Distribuzione delle emozioni su dataset RAVDESS

Per riportare il dataset in equilibrio senza dover scartare file audio dalle classi maggioritarie, si è scelto di intervenire sulla classe minoritaria attraverso un'operazione di oversampling. Il procedimento adottato non si limita a una duplicazione deterministica dei file. Infatti, per ogni campione neutro, oltre alla generazione dei tre cloni stocastici standard previsti dalla *Data Augmentation*, è stata prodotta una seconda serie di varianti caratterizzate da un'ulteriore aggiunta di rumore bianco gaussiano AWGN con un fattore di disturbo pari a 0.01. Questa procedura genera dati sintetici che, occupando posizioni distinte nello spazio delle feature, permettono di pareggiare la cardinalità della classe Neutral e forzano il modello a identificare tratti vocali robusti anziché memorizzare i singoli campioni.

Per quanto riguarda il dataset EMO-DB,al contrario, presenta uno sbilanciamento eterogeneo che interessa l'intero spettro delle classi emotive Fig.3.12. In questo contesto, il ricorso a tecniche di oversampling risulta inadeguato poiché la replicazione di campioni per un numero elevato di classi minoritarie introdurrebbe una ridondanza tale da distorcere la distribuzione originale delle feature acustiche.

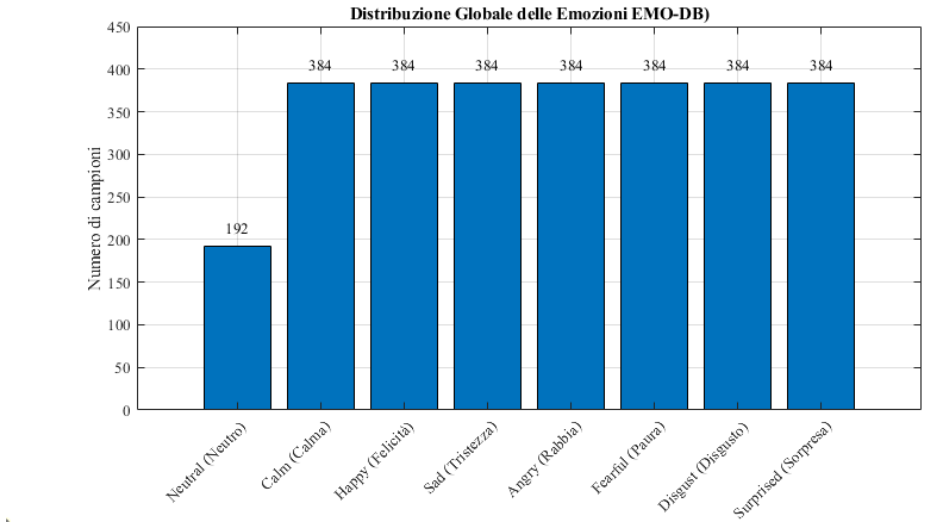


Figura 3.12: Distribuzione delle emozioni nel dataset EMO-DB

Si è pertanto adottata una compensazione a livello algoritmico mediante l'integrazione di una *Class-Balanced Loss* nel processo di ottimizzazione [47].Questa metodologia si basa sul presupposto teorico che il beneficio informativo marginale derivante dall'aggiunta di nuovi dati diminuisca esponenzialmente all'aumentare della numerosità della classe. Tale fenomeno viene formalizzato attraverso il concetto di numero effettivo di campioni (E_n), definito dalla relazione matematica:

$$E_n = \frac{1-\beta^n}{1-\beta} \tag{3.13}$$

dove n rappresenta il numero di campioni reali della classe, mentre l'iperparametro $\beta \in [0,1)$ definisce la probabilità che un nuovo campione condivida lo stesso spazio informativo di quelli già analizzati. Quando β tende all'unità, il valore di E_n satura rapidamente, indicando che l'aggiunta di ulteriori istanze non apporta nuova conoscenza. Al contrario, per le classi minoritarie d, il E_n si mantiene prossimo al numero reale di campioni n . Sulla base di questa logica, l'algoritmo assegna a ogni categoria un peso

W_c inversamente proporzionale al suo reale volume informativo:

$$W_c = \frac{1}{E_n} \tag{3.14}$$

Dal punto di vista dell'implementazione, questo coefficiente agisce direttamente sulla funzione di costo come un fattore moltiplicativo durante la fase di retropropagazione dell'errore. In caso di errata classificazione di un'emozione rara, il peso elevato amplifica il gradiente, costringendo la rete ad aggiornare i parametri in modo più incisivo. Nel presente lavoro di tesi il parametro β non è stato assegnato empiricamente ma è stato inserito nello spazio di ricerca dell'ottimizzazione bayesiana. Tale scelta ha consentito di individuare il punto di saturazione ottimale per l'architettura proposta, garantendo un apprendimento bilanciato.

Ottimizzazione Bayesiana degli Iperparametri

L'addestramento di reti neurali profonde, come la 1D-CLDNN, richiede la calibrazione di numerosi iperparametri (es. tasso di apprendimento, livello di regolarizzazione, dropout). Poiché non esiste una formula matematica capace di calcolare a priori la configurazione perfetta, il processo si traduce nella minimizzazione di una funzione di errore. In termini analitici, l'obiettivo consiste nell'individuare il set di parametri x^* all'interno dello spazio di ricerca ammissibile $\mathscr{X}$ in grado di minimizzare l'errore di validazione

$$x^* = \underset{x \in \mathscr{X}}{\operatorname{argmin}} f(x) \tag{3.15}$$

Al fine di superare i limiti legati a un'esplorazione puramente empirica, nel presente lavoro si è implementata l'ottimizzazione bayesiana [48].La procedura si basa sulla costruzione di un modello surrogato, ovvero una funzione statistica che approssima l'andamento dell'errore reale. Tale

rappresentazione è definita analiticamente da un Processo Gaussiano [49], il quale assegna a ogni punto dello spazio di ricerca una distribuzione di probabilità:

$$f(x) \sim \mathscr{GP}(\mu(x), k(x, x\prime))$$

dove $\mu(x)$ è la stima della prestazione attesa (media) per una specifica configurazione mentre $k(x, x\prime)$ è la funzione di covarianza che quantifica l'incertezza associata alle aree dello spazio di ricerca non ancora campionate. La distribuzione probabilistica così ottenuta permette di determinare, ad ogni passaggio, la successiva combinazione di iperparametri da sottoporre a test. Il processo decisionale è controllato da una funzione ausiliaria, denominata Funzione di Acquisizione, che nel framework implementato coincide con l'Expected Improvement. Essa consiste nel quantificare il miglioramento atteso rispetto alla prestazione ottimale registrata fino a quel momento, indicata con $(f(x^+))$:

$$a_{EI}(x) = \mathbb{E}\left[\max(0, f(x^+) - f(x))\right] \tag{3.16}$$

Attraverso la massimizzazione di a a_{EI}, l'algoritmo bilancia in modo automatico lo sfruttamento delle zone con errore basso già noto (exploitation) e l'esplorazione delle aree dove l'incertezza è ancora elevata (exploration). Tale meccanismo garantisce una convergenza rapida verso la configurazione ideale della rete neurale. L'implementazione dell'ottimizzazione bayesiana è stata adattata alle specifiche proprietà strutturali dei due dataset analizzati, al fine di bilanciare l'accuratezza della ricerca con la sostenibilità dei tempi di calcolo. Nel caso del corpus RAVDESS, l'addestramento sull'intero set di dati espanso avrebbe richiesto tempi di calcolo eccessivi, rendendo l'ottimizzazione bayesiana poco pratica. Per ovviare a questo problema, si è scelto di far lavorare l'algoritmo su una versione ridotta, limitata ai campioni audio originali e al bilanciamento della classe Neutral. Questa

configurazione più leggera ha permesso di esplorare rapidamente lo spazio dei parametri, individuando i valori ideali per il learning rate, la regolarizzazione L2, la dimensione del mini-batch e il fattore di dropout. In questa fase, l'obiettivo della funzione è stato impostato sulla minimizzazione dell'errore di accuratezza ($1 - \text{Accuracy}$) , poiché il bilanciamento preventivo delle classi rende tale metrica statisticamente robusta. Una volta trovata la combinazione ottimale, questi parametri sono stati estratti e utilizzati per l'addestramento finale sull'intero dataset. Le dimensioni contenute del corpus EMO-DB hanno consentito di estendere l'ottimizzazione bayesiana all'intero set di dati, integrando nello spazio di ricerca l'iperparametro β per una gestione dinamica e automatizzata della Class-Balanced Loss. A differenza del RAVDESS la funzione obiettivo è stata orientata alla minimizzazione dell' F1-score ($1 - \text{F}_1$) in modo da bilanciare precisione e richiamo e garantire una valutazione imparziale anche per le categorie emotive clinicamente meno rappresentate. La Tabella 3.4 seguente riassume lo spazio di ricerca definito per le variabili oggetto di ottimizzazione:

Tabella 3.4: Spazio di ricerca definito per l'ottimizzazione bayesiana degli iperparametri.

Iperparametro	**Range / Valori**	**Dataset**
Learning Rate	$[10^{-5}, 10^{-2}]$	Entrambi
Regolarizzazione L_2	$[10^{-6}, 10^{-2}]$	Entrambi
Dropout Rate	$[0.2, 0.5]$	Entrambi
Mini-batch size	$\{16, 32, 64\}$	RAVDESS
Coefficiente β (CB Loss)	$[0.99, 0.9999]$	EMO-DB

Protocollo di Addestramento

Il protocollo di addestramento definisce i parametri di controllo necessari a regolare la dinamica di apprendimento del modello. Mentre gli iperparametri strutturali sono determinati tramite procedura bayesiana, i parametri di gestione del training sono impostati su base empirica per garantire la

stabilità del processo. L'algoritmo selezionato per l'aggiornamento dei pesi è Adam poiché integra tassi di apprendimento adattivi calcolati individualmente per ogni parametro della rete [50]. A differenza degli altri metodi, esso analizza le variazioni dei gradienti attraverso due grandezze statistiche che ne guidano il comportamento. Il primo termine m_t rappresenta la media mobile dei gradienti passati e serve a tracciare la direzione principale della discesa. Il secondo termine v_t memorizza la media dei gradienti al quadrato, misurando di fatto l'instabilità e la varianza del segnale analizzato. Le relazioni matematiche applicate sono

$$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t \tag{3.17}$$

$$v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2 \tag{3.18}$$

In queste espressioni i coefficienti β_1 e β_2 rappresentano i tassi di decadimento esponenziale delle medie mobili. Il sistema utilizza tali informazioni per calibrare automaticamente l'intensità delle correzioni e riduce il passo di apprendimento per i parametri caratterizzati da gradienti instabili garantendo una convergenza robusta anche con segnali audio complessi. Al fine di incrementare la precisione nelle fasi conclusive del processo è stato introdotto un LearnRateSchedule di tipo piecewise che dimezza il passo ogni 20 epoche. Tale progressiva riduzione impedisce ai pesi della rete di oscillare in prossimità del minimo della funzione di errore e favorisce un assestamento definitivo della soluzione. Parallelamente il rischio di overfitting viene gestito mediante un protocollo di Early Stopping con una pazienza di 25 epoche [51]. Questo meccanismo arresta l'addestramento appena l'errore di validazione smette di scendere obbligando il modello ad apprendere pattern emotivi generali anziché memorizzare le specificità dei singoli campioni. Le specifiche tecniche adottate per i due dataset sono

riassunte nelle tabelle comparative seguenti

Tabella 3.5: Training Options RAVDESS

Parametro	**Valore**
Ottimizzatore	Adam
Obiettivo Bayes	$1 - \text{Accuracy}$
Max Epochs	120
Pazienza	25 Epoche
LR Schedule	Piecewise 20 ep
Drop Factor	0.5
Shuffle	every epoch

Tabella 3.6: Training Options EMODB

Parametro	**Valore**
Ottimizzatore	Adam
Obiettivo Bayes	$1 - F_1$
Max Epochs	100
Pazienza	25 Epoche
LR Schedule	Piecewise 20 ep
Drop Factor	0.5
Shuffle	every epoch

Nested Cross Validation 10 fold

La valutazione della rete 1D-CLDNN segue un protocollo di Nested Cross-Validation a 10 fold come illustrato nello schema di Figura 3.13. Tale configurazione permette di separare la fase di ottimizzazione bayesiana dalla valutazione finale, garantendo che le metriche riportate riflettano l'effettiva capacità di generalizzazione del sistema. La procedura si articola in due livelli funzionali:

- Ciclo Esterno: La partizione del dataset originale è stata eseguita tramite la funzione cvpartition in ambiente MATLAB. Il comando esegue un campionamento stratificato rispetto alla distribuzione delle classi emotive (y) e genera 10 fold mutuamente esclusivi. delle 10 iterazioni nove fold costituiscono il set di addestramento mentre il decimo fold rimane isolato come Test Set indipendente. Al fine di garantire una valutazione imparziale e rappresentativa di scenari reali tale partizione non subisce alcuna operazione di data augmentation o manipolazione sintetica. La validazione finale del modello avviene pertanto esclusivamente sui campioni audio originali.

- Ciclo Interno: Per ogni iterazione del ciclo esterno, i dati di addestramento sono stati elaborati secondo una sequenza a due stadi. Inizialmente, è stata definita una partizione di tipo hold-out (80/20) dedicata alla ricerca della configurazione ottimale degli iperparametri mediante ottimizzazione bayesiana. In questa fase non è stata applicata alcuna tecnica di data augmentation, limitando l'intervento sul dataset al solo bilanciamento della classe Neutral per stabilizzare il processo di convergenza.Una volta individuata la combinazione di parametri ideale, un secondo frazionamento (90/10) integra l'intera procedura di augmentation e bilanciamento per l'addestramento definitivo del modello.

Il classificatore, così configurato e addestrato, è stato infine valutato sul Test Set estratto dal ciclo esterno. La media dei risultati ottenuti sui 10 fold fornisce la stima delle prestazioni della rete 1D-CLDNN.

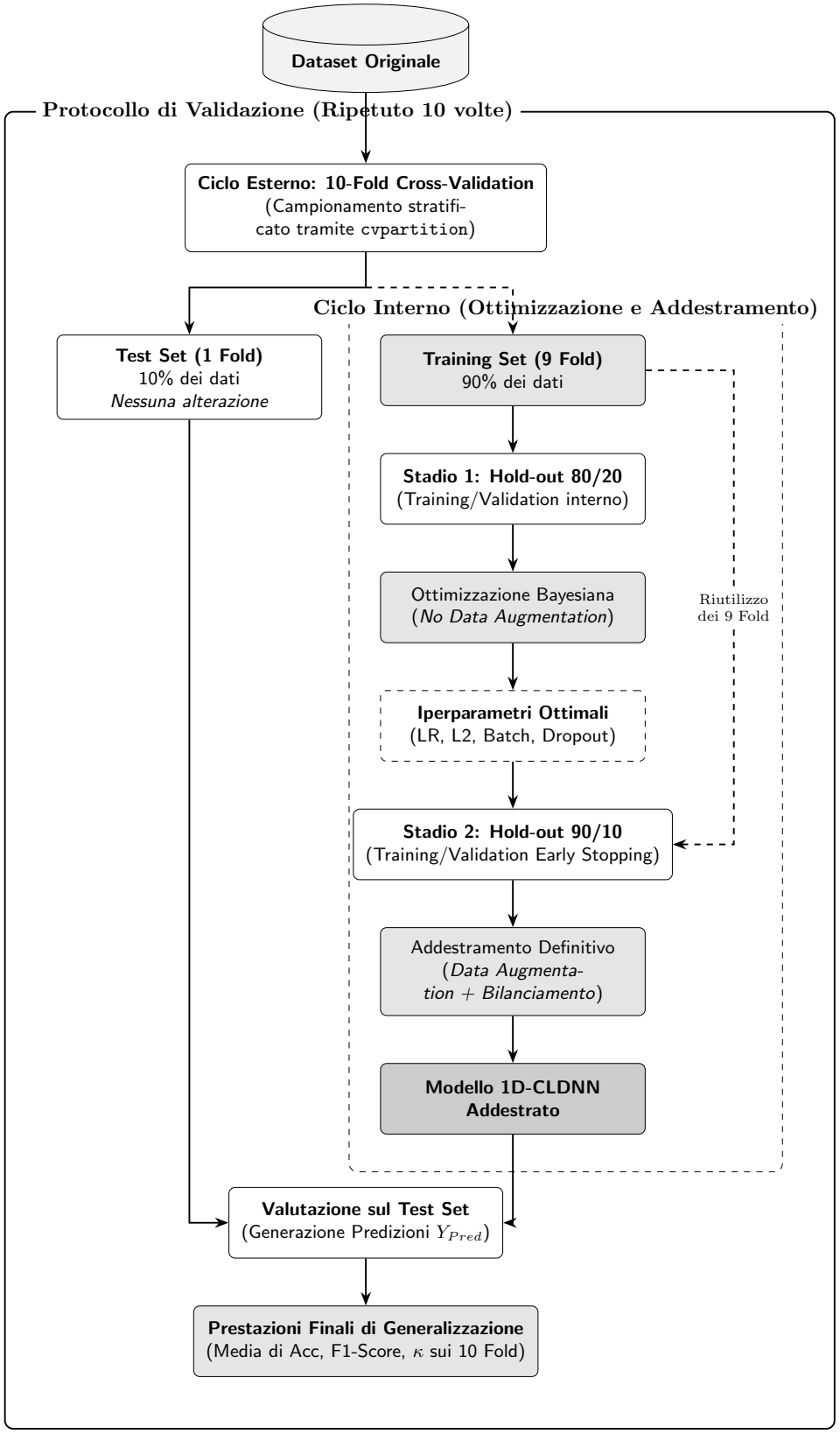


Figura 3.13: Schema a Blocchi del Nested Cross Validation a 10 fold

Metriche di Valutazione

L'analisi quantitativa del sistema di classificazione si basa sul confronto diretto tra il vettore delle classi reali (Y_{True}) e quello delle predizioni generate dal modello (Y_{Pred}) [52]. Tale processo permette di determinare, per ogni classe emotiva $i \in 1, ..., C$, i quattro parametri fondamentali per la valutazione statistica. Nello specifico vengono identificati i Veri Positivi (TP) e i Veri Negativi (TN) come misure di corretta classificazione ed esclusione mentre i Falsi Positivi (FP) e i Falsi Negativi (FN) mappano rispettivamente gli errori di tipo I e II. Sulla base di tali conteggi elementari viene determinata l'Accuratezza Globale (Acc) espressa dal rapporto tra la somma delle istanze correttamente identificate e il numero totale di campioni N:

$$Acc = \frac{TP + TN}{N} \tag{3.19}$$

In scenari caratterizzati da sbilanciamento delle classi l'accuratezza globale può tuttavia generare stime ottimistiche che non riflettono l'effettiva efficacia del modello sulle categorie meno rappresentate. Per minimizzare tale distorsione l'analisi integra metriche più selettive quali la Precisione e la Recall. La Precisione (P) quantifica l'affidabilità delle predizioni positive ed esprime la proporzione di istanze correttamente classificate rispetto al totale delle assegnazioni effettuate per una determinata categoria. Elevati valori di precisione indicano una spiccata capacità del sistema nel limitare l'occorrenza di Falsi Positivi.

$$P = \frac{TP}{TP + FP} \tag{3.20}$$

Parallelamente la Recall (R) definita anche come sensibilità misura la completezza del classificatore nell'identificare i campioni appartenenti a una specifica classe. Un'alta recall attesta l'efficacia del modello nel ridurre

i Falsi Negativi garantendo che la maggior parte dei target reali venga correttamente riconosciuta.

$$R = \frac{TP}{TP + FN} \tag{3.21}$$

La sintesi di queste due grandezze avviene tramite l'F1-Score calcolato come media armonica tra la precisione e la recall. Questo indice fornisce una valutazione equilibrata della capacità predittiva del modello poiché penalizza i casi in cui una delle due componenti risulti significativamente inferiore all'altra.

$$F_1 score = 2 \cdot \frac{P \cdot R}{P + R} \tag{3.22}$$

Per discriminare l'apprendimento dei pattern acustici da una convergenza puramente aleatoria, l'analisi integra il Coefficiente Kappa di Cohen (κ). Tale metrica normalizza l'accuratezza osservata (p_o) rispetto alla probabilità di concordanza casuale (p_e), stimata sulla base delle distribuzioni marginali della matrice di confusione:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{3.23}$$

Questo parametro fornisce una misura della robustezza del classificatore al netto delle fluttuazioni stocastiche. Per l'interpretazione di tale indice, lo standard accademico fa riferimento alla tassonomia introdotta da Landis e Koch [53], la quale classifica il livello di concordanza statistica in diverse fasce di affidabilità. Ad esempio, valori di κ compresi tra 0.61 e 0.80 denotano un accordo "sostanziale", fornendo una prova che il sistema ha effettivamente appreso pattern discriminativi senza affidarsi al puro caso, mentre valori oltre lo 0.81 indicano una classificazione quasi perfetta. Infine, la Matrice di Confusione viene impiegata per esaminare la distribuzione delle predizioni rispetto ai dati reali. Essa è fondamentale per identificare

le intersezioni critiche tra classi emotive caratterizzate da profili acustici sovrapponibili fornendo una chiave di lettura qualitativa dei limiti del classificatore.

3.3 Pipeline per la diagnosi del Disturbo dello Spettro Austico (ASD)

Lo studio del Riconoscimento delle Emozioni (SER) su dataset attoriali ha fornito le basi metodologiche per l'estrazione e la classificazione di pattern acustici complessi. La seconda fase di questo lavoro di tesi ha previsto l'applicazione di paradigmi di apprendimento profondo in ambito clinico al fine di supportare la diagnosi del Disturbo dello Spettro Autistico ASD. In questo nuovo contesto l'algoritmo non deve più interpretare un'emozione simulata volontariamente ma identificare specifici biomarcatori vocali involontari derivanti dalle alterazioni prosodiche tipiche della patologia. A tal fine sono stati impiegati due distinti corpus di dati ovvero il dataset ASDBank e un secondo dataset sperimentale le cui specifiche tecniche sono state descritte nel paragrafo 3.1. Essi sono stati elaborati e valutati separatamente per dimostrare l'indipendenza predittiva del modello dal singolo ambiente di acquisizione. Nelle sezioni successive verrà illustrata nel dettaglio l'intera pipeline implementata. Inizialmente verranno descritte le fasi di preprocessing necessarie per la pulizia e la standardizzazione dei segnali audio, successivamente si analizzerà l'impiego della rete VGGish per l'estrazione delle feature e infine verrà definita l'architettura neurale adottata per la classificazione.

3.3.1 Preprocessing dei segnali audio

Per garantire l'integrità della pipeline analitica i file audio sono stati sottoposti a una fase di editing manuale tramite il software Audacity [54]

finalizzata all'isolamento delle componenti fonatorie di interesse e alla rimozione degli artefatti. In primo luogo è stato eseguito il ritaglio dei segmenti non pertinenti. Attraverso l'ispezione visiva della forma d'onda e l'ausilio della funzione di zoom sono stati eliminati i silenzi iniziali, i rumori ambientali estemporanei e le voci riconducibili a figure adulte quali clinici o genitori. Questa selezione permette di circoscrivere l'analisi alle sole unità fonatorie del paziente impedendo alla rete di correlare la diagnosi a tratti timbrici estranei alla patologia sotto osservazione. Successivamente il segnale è stato trattato per mitigare il rumore mediante lo strumento di Noise Reduction nativo di Audacity basato sull'analisi di Fourier. La procedura richiede la selezione preliminare di un segmento di puro rumore della durata minima di 0.05 s per i file campionati a 44.1 kHz. Il software elabora questo frammento applicando una Trasformata di Fourier strutturata su una finestra di Hann da 2048 campioni. Questa operazione matematica suddivide il rumore in 1025 bande di frequenza e ne calcola l'energia media generando così una precisa impronta spettrale del disturbo che verrà poi sottratta dall'intera registrazione . La ricostruzione dell'audio ripulito nel dominio temporale avviene attraverso una trasformata inversa applicando un processo di time smoothing per prevenire discontinuità o distorsioni. L'applicazione del filtro sull'intera registrazione è stata ottimizzata attraverso l'utilizzo della funzione anteprima e la conseguente calibrazione di tre parametri operativi [55]:

- Riduzione del rumore (6 dB): definisce l'entità dell'attenuazione applicata alle frequenze di disturbo per preservare l'integrità delle formanti vocali.

- Sensibilità (6): stabilisce la soglia di energia per la classificazione degli eventi acustici, evitando la rimozione accidentale di componenti deboli del parlato.

- Smorzamento di frequenza: settato a 6 bande per rendere omogeneo il passaggio tra le componenti filtrate e quelle conservate

Per quanto riguarda il dataset sperimentale, l'analisi ha evidenziato la presenza di estesi intervalli di inattività fonatoria i quali avrebbero potuto compromettere l'efficacia della classificazione e aumentare inutilmente il carico computazionale della pipeline. Per ovviare a tale criticità è stata applicata la funzione Truncate Silence di Audacity. Tale operazione identifica i segmenti con energia inferiore alla soglia di -35 dB e, qualora la durata del silenzio superi gli 0.5 s, ne effettua la contrazione automatica a un valore fisso di 0.1 s. Questa compressione elimina i vuoti acustici inutili senza alterare il ritmo naturale del parlato, fornendo alla rete neurale input più densi di informazione utile. In Figura 3.14 è riportato il risultato delle Prepocessing

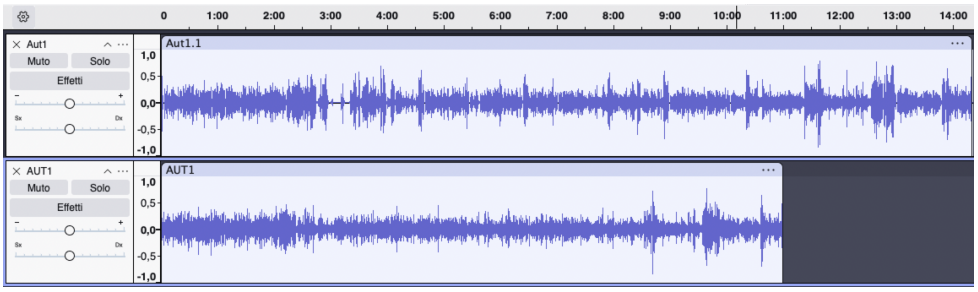


Figura 3.14: Confronto tra il segnale audio grezzo (in alto) e il risultato finale della pipeline di preprocessing (in basso).

3.3.2 Estrazione delle Feature: Il paradigma VGGish

La scelta di discostarsi dalla pipeline impiegata per lo Speech Emotion Recognition (SER) è dettata dalla necessità di garantire una maggiore robustezza statistica a fronte della ridotta numerosità dei campioni clinici, costituiti dagli 84 soggetti dell'ASDBank e dai 61 del corpus sperimentale. L'addestramento ex-novo di architetture profonde come la 1D-CLDNN su volumi di dati così ridotti comporterebbe un alto rischio di overfitting. Si

è pertanto adottato un paradigma di Transfer Learning basato sul modello VGGish, un estrattore di caratteristiche acustiche rilasciato da Google nel 2017. Questa rete è stata pre-addestrata su Audio Set, un'ontologia che comprende circa 1.8 milioni di segmenti audio estratti da YouTube e annotati manualmente in 632 classi sonore [56]. L'intuizione alla base di VGGish risiede nel trattare il segnale audio non come una serie temporale unidimensionale, ma come una rappresentazione tempo-frequenza bidimensionale. Ispirandosi all'architettura VGG (Visual Geometry Group) impiegata nella visione artificiale, il sistema è progettato per identificare pattern acustici complessi all'interno degli spettrogrammi log-mel, trattandoli come se fossero caratteristiche visive di un'immagine [57]. Il processo di estrazione, implementato in ambiente MATLAB tramite il tool vggish prevede

- Pre-processing: il segnale audio viene ricampionato a 16 kHz e convertito in mono.
- Framing: l'audio viene suddiviso in "patch" non sovrapposte della durata di 960 ms.
- Trasformazione Log-Mel: per ogni patch viene calcolata una STFT con finestre di 25 ms e passo di 10 ms. I risultati vengono poi integrati in 64 bande Mel e sottoposti a trasformazione logaritmica per emulare la percezione uditiva umana.

La rappresentazione bidimensionale così ottenuta, avente dimensioni pari a 96×64 bin [57], costituisce l'input di un'architettura convoluzionale profonda articolata in undici strati principali. Il modello è strutturato in quattro blocchi convoluzionali consecutivi che impiegano filtri di piccole dimensioni (3×3) con stride unitario. All'interno di ciascun blocco, l'integrazione sistematica di strati di Batch Normalization garantisce la stabilità del gradiente durante l'apprendimento, mentre le funzioni di attivazione ReLU introducono la non-linearità necessaria alla discriminazione

di pattern. A valle di ogni stadio convoluzionale, viene inserito un layer di Max-Pooling (2×2, stride 2) che effettua una riduzione della dimensionalità spaziale, rendendo la rete invariante rispetto alle micro-traslazioni temporali e frequenziali tipiche del segnale vocale. La sezione terminale della rete è caratterizzata da due livelli fully connected da 4096 unità che convergono in uno strato denominato Embedding Layer finale composto da 128 neuroni. In questa configurazione, l'utilizzo della funzione Matlab *vggishEmbeddings* non restituisce una classe discreta, ma un embedding numerico ovvero un vettore denso a 128 dimensioni che proietta il segnale in uno spazio latente dove la distanza matematica tra i punti riflette fedelmente la somiglianza semantica e prosodica del parlato.

3.3.3 Architettura del Classificatore e Strategia di Bilanciamento

L'architettura proposta elabora le caratteristiche estratte tramite VGGish configurando un blocco di classificazione dedicato. Il vettore di input 128 dimensioni viene elaborato attraverso uno strato Fully Connected con ampiezza ottimizzata tra 32 e 256 unità, seguito da stadi di Batch Normalization e attivazione ReLU. L'inserimento di un livello di Dropout (con rate compreso tra 0.3 e 0.7) agisce come regolarizzatore stocastico per impedire che la rete si focalizzi su artefatti acustici o tratti biometrici specifici dei singoli soggetti presenti nel training set. L'assegnazione della probabilità finale è affidata a uno strato softmax combinato con un livello decisionale strutturato sui principi del Cost-Sensitive Learning [58]. Questa strategia metodologica risulta fondamentale per correggere lo sbilanciamento numerico intrinseco ai dataset clinici. Il sistema compensa la disparità tra i gruppi calcolando un peso W_c per ciascuna classe c in modo inversamente

proporzionale alla sua frequenza secondo la formula

$$W_c = \frac{N}{k \cdot n_c} \tag{3.24}$$

dove N rappresenta il numero totale di segmenti nel training set, K il numero di classi e n_c la numerosità della classe specifica. L'integrazione di tali coefficienti all'interno della funzione di perdita (Loss Function) costringe il modello a penalizzare severamente gli errori commessi sui soggetti minoritari garantendo una valutazione oggettiva della capacità discriminativa. In linea con la metodologia precedentemente illustrata la configurazione ottimale della rete viene individuata tramite ottimizzazione bayesiana. L' algoritmo valuta quindici diverse combinazioni testandone l'efficacia su un set di validazione interno corrispondente al 20% . Lo spazio di ricerca assegnato a ciascun iperparametro è dettagliato nella tabella successiva.

Tabella 3.7: Iperparametri e intervalli di ricerca per l'ottimizzazione bayesiana.

Iperparametro	**Range**
Learning Rate (LR)	$10^{-4} - 10^{-2}$
L2 Regularization	$10^{-4} - 0.1$
Dropout Rate	$0.3 - 0.7$
FC Units	$32 - 256$
Batch Size	$32 - 128$

Una volta determinati gli iperparametri il modello definitivo viene addestrato impiegando l'ottimizzatore Adam. Le dinamiche di training integrano i meccanismi di decadimento del tasso di apprendimento e di arresto anticipato già argomentati nei capitoli precedenti applicati su un set di validazione dedicato . I valori esatti adottati per questa fase finale sono riassunti nella Tabella 3.8 .

Tabella 3.8: Impostazioni di addestramento del modello finale (Training Options).

Parametro	**Valore / Impostazione**
Ottimizzatore	Adam
Learning Rate Iniziale	Ottimizzazione Bayesiana
L2 Regularization	Ottimizzazione Bayesiana
Mini Batch Size	Ottimizzazione Bayesiana
Max Epochs	60
Pazienza (Early Stopping)	8 Epoche
Learn Rate Schedule	Piecewise (Decadimento ogni 15 epoche)
Learn Rate Drop Factor	0.5
Shuffle	Ogni epoca

3.3.4 Patient-Wise e Protocollo di Validazione

L'integrazione del modello VGGish introduce criticità legate alla gestione della risoluzione temporale del segnale. Poiché ogni registrazione audio viene decomposta in una moltitudine di segmenti da 960 ms, il dataset risultante è comprende migliaia di vettori di embedding che conservano l'impronta biometrica del soggetto d'origine. Un partizionamento casuale tra i set di addestramento e di test a livello di singoli segmenti indurrebbe un fenomeno di Data Leakage, portando la rete neurale a riconoscere l'identità vocale dei pazienti piuttosto che i biomarcatori prosodici associati alla patologia.Per mitigare tale distorsione sistematica è stata implementata una suddivisione Patient-Wise. Attraverso l'impiego degli identificativi univoci associati ai file audio, l'insieme di embedding appartenente a un singolo paziente viene vincolato a un unico fold di test. Questa procedura garantisce che il classificatore agisca esclusivamente su voci mai udite in precedenza, simulando un contesto diagnostico reale e assicurando che le prestazioni rilevate dipendano esclusivamente da caratteristiche generalizzabili. Tale logica è stata applicata all'interno di uno schema di Nested Cross-Validation a 5 fold. Sebbene la procedura segua la medesima impostazione adottata per il SER, il numero di partizioni è stato ridotto da 10 a 5

per evitare che i set di test risultassero numericamente troppo esigui. Con una coorte di 84 pazienti per l'ASDBank e 61 per il corpus supplementare, un frazionamento in 10 parti avrebbe infatti generato fold composti da soli 6-8 soggetti, rendendo le metriche di accuratezza eccessivamente sensibili alla variabilità dei singoli profili o a rumori acustici localizzati. L'impiego di 5 fold assicura invece una base campionaria del 20% per ogni iterazione, permettendo una valutazione più consistente della sensibilità e della specificità diagnostica. Il protocollo si articola in un ciclo interno dedicato alla ricerca degli iperparametri e un ciclo esterno per la verifica delle prestazioni. Questo garantisce l'imparzialità del classificatore e l'indipendenza assoluta dei soggetti tra le diverse fasi sperimentali.

3.3.5 Aggregazione dei Risultati e Metriche di Validazione Clinica

Per convertire le predizioni dei singoli segmenti in un risultato univoco per ogni paziente, in accordo con la logica patient-wise, il sistema adotta una strategia di Soft Voting. Tale procedura non si basa sulla frequenza delle classi predette, bensì calcola la media aritmetica delle probabilità restituite dal modello per ciascuna classe c su tutti gli M segmenti che compongono la registrazione del soggetto[59]:

$$P(c) = \frac{1}{M} \sum_{j=1}^{M} P_{j,c} \tag{3.25}$$

Il superamento della soglia decisionale convenzionalmente pari a 0.5 determina l'identificazione finale del paziente come appartenente alla classe ASD, mentre un valore inferiore comporta la classificazione come controllo. Questa metodologia garantisce una affidabilità superiore rispetto alla votazione di maggioranza, poiché la media probabilistica, conformemente

alla *Sum Rule* [59], attenua l'influenza di eventuali outlier o tratti audio privi di valore discriminante, come silenzi e interferenze ambientali. Al fine di caratterizzare il potenziale diagnostico dell'architettura oltre i parametri già impiegati per il SER, l'analisi viene estesa allo studio della curva ROC (Receiver Operating Characteristic) e del relativo indice AUC (Area Under the Curve) [52]. La curva ROC illustra la capacità discriminativa del classificatore al variare della soglia decisionale mettendo in relazione la sensibilità (True Positive Rate) con il tasso di falsi positivi (False Positive Rate). L'indice AUC sintetizza l'efficacia del modello in un valore compreso tra 0.5 e 1 che esprime la probabilità di distinguere correttamente un paziente ASD rispetto a un controllo sano [52]. Un punteggio prossimo all'unità conferma che le distribuzioni dei due gruppi sono ben separate e documenta la capacità della rete neurale di identificare i profili clinici con precisione indipendentemente dalla soglia di attivazione scelta.

4. Risultati

Il presente capitolo descrive i risultati sperimentali ottenuti nelle due fasi del progetto. La prima sezione è dedicata al riconoscimento delle emozioni tramite l'architettura 1D-CLDNN sui dataset pubblici RAVDESS ed EMO-DB. Per entrambi i corpus, verranno confrontate le prestazioni del modello addestrato con il set acustico di base (65 feature) e con il set ottimizzato (48 feature). La seconda sezione del capitolo illustra invece l'applicazione clinica del sistema. In quest'ambito, l'architettura VGGish viene valutata nel compito di classificazione del Disturbo dello Spettro Autistico ASD , riportando gli esiti dei test condotti sia sul dataset ASDBank sia su sul dataset sperimentale.

4.1 Risultati sul dataset EMO-DB

L'analisi dei risultati ottenuti sul dataset EMO-DB permette di valutare concretamente come la dimensionalità dello spazio delle feature influenzi la capacità di generalizzazione del modello.

4.1.1 Analisi della configurazione a 65 Feature

L'elaborazione condotta sul set di 65 feature ha restituito un'accuratezza globale del 71.21%, associata a un F1-Score di 0.7044 e a un coefficiente Cohen's Kappa di 0.6594, parametro che indica un accordo sostanziale tra le etichette reali e le predizioni del sistema. Come riportato nella Tabella 4.1, il modello risponde in modo differenziato alla natura spettrale delle emozioni. La Tristezza (Sadness) emerge come la categoria meglio identificata, raggiungendo un F1-Score di 0.8148 e una Recall di 0.8871, a dimostrazione di quanto i descrittori acustici siano efficaci nel mappare

i profili a bassa energia. Risultati altrettanto validi si riscontrano per il Disgusto (Disgust) e la Rabbia (Angry), i cui F1-Score si attestano rispettivamente a 0.78 e 0.7795. Le classi caratterizzate da una ridotta attivazione energetica, come il Neutro (Neutral) e la Noia (Boredom), hanno riportato prestazioni bilanciate con F1-Score di 0.7561 e 0.7248. Al contrario, le criticità maggiori si concentrano sulla Paura (F1=0.5649) e sulla Felicità (F1=0.5109) che risulta essere l'emozione più complessa da classificare.

Tabella 4.1: Metriche dettagliate per classe (65 feature)

Classe	**Precision**	**Recall**	**F1-Score**
Angry	0,7795	0,7795	0,7795
Boredom	0,7941	0,6667	0,7248
Disgust	0,7222	0,8478	0,7800
Fearful	0,5968	0,5362	0,5649
Happy	0,5303	0,4930	0,5109
Neutral	0,7294	0,7848	0,7561
Sadness	0,7534	0,8871	0,8148

L'analisi della matrice di confusione evidenzia come la discriminazione delle classi sia influenzata dalla distribuzione dei livelli di attivazione e dalla valenza del segnale (Figura 4.1). Nel dominio dell'alto arousal si riscontra una significativa sovrapposizione tra Felicità e Rabbia (28.2% e 12.6%), dovuta alla similitudine dei profili energetici. La Felicità manifesta inoltre un'ulteriore ambiguità verso la Paura (16.9%). Nelle categorie a bassa attivazione, la classe Noia presenta un'incidenza di errore distribuita equamente tra la Tristezza (13.6%) e lo stato Neutro (13.6%). La Paura si conferma invece l'emozione con la maggiore instabilità decisionale, con errori che si ripartiscono tra le classi ad alta attivazione quali Felicità (14.5%) e Rabbia (10.1%), e lo stato Neutro (11.6%). Tale dispersione indica che, in assenza di marker spettrali univoci , il classificatore non definisce un confine decisionale stabile tra i diversi livelli di attivazione dello spazio delle feature.

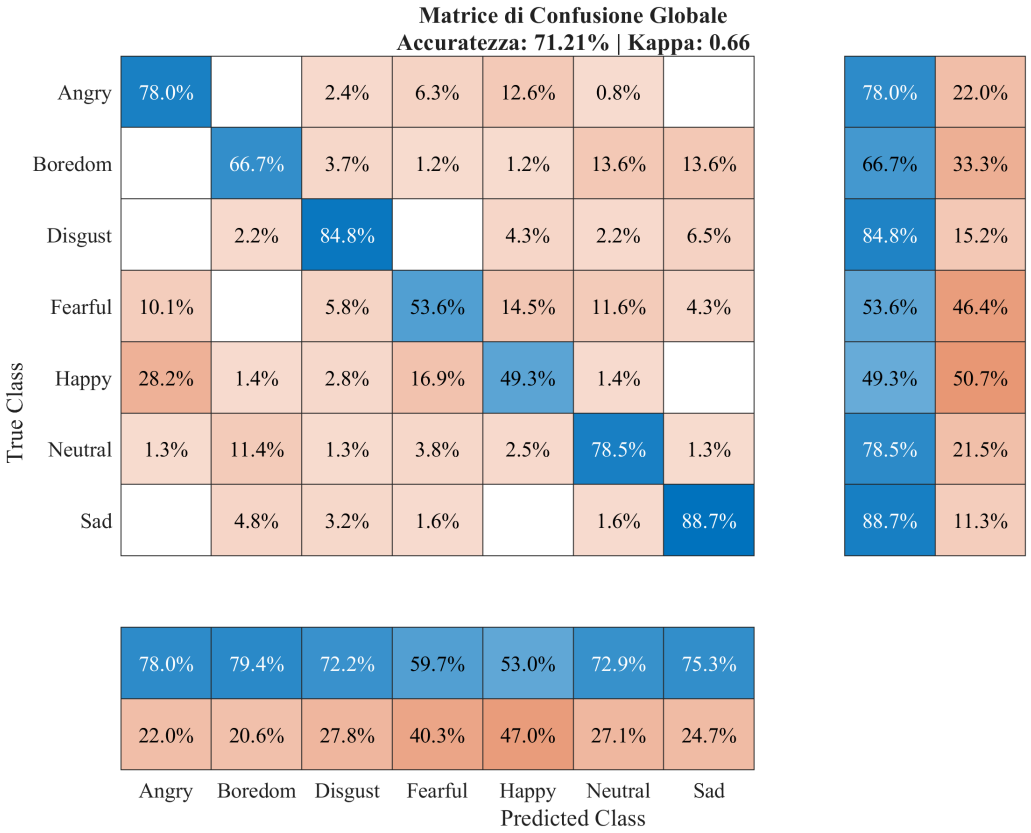


Figura 4.1: Matrice di confusione globale del Dataset EMO-DB (65 feature).

Al fine di verificare la reale robustezza del classificatore rispetto alle variabili biometriche degli attori è stata effettuata un'analisi comparativa basata sul genere. I risultati evidenziano una discrepanza prestazionale tra i due gruppi in quanto l'accuratezza del 72.85% ottenuta sui campioni femminili subisce una flessione attestandosi al 69.10% per i soggetti maschili. L'analisi della Figura 4.2 e della Tabella 4.6 mostra come lo sbilanciamento non sia uniforme tra le categorie emotive, ma si concentri su specifici profili acustici. Il modello evidenzia una risoluzione nettamente superiore nel discriminare il Disgusto e la Tristezza nelle voci femminili, con valori di F1-Score che superano la soglia dello 0.85. Al contrario, per i soggetti maschili, tali performance subiscono una contrazione significativa, attestandosi rispettivamente a 0.5217 e 0.7368. Il divario più critico si registra tuttavia nella classe Felicità, dove il rendimento maschile si ferma a un esiguo 0.3922 rispetto allo 0.5814 del campione femminile.

Tabella 4.2: Confronto delle metriche per genere (65 feature)

Campione Maschile

Classe	Prec.	Rec.	F1
Angry	0,7969	0,8500	0,8226
Boredom	0,7586	0,6286	0,6875
Disgust	0,5000	0,5455	0,5217
Fearful	0,5938	0,5278	0,5588
Happy	0,4167	0,3704	0,3922
Neutral	0,8000	0,8205	0,8101
Sadness	0,6563	0,8400	0,7368

Campione Femminile

Classe	Prec.	Rec.	F1
Angry	0,7619	0,7164	0,7385
Boredom	0,8205	0,6957	0,7529
Disgust	0,7857	0,9429	0,8571
Fearful	0,6000	0,5455	0,5714
Happy	0,5952	0,5682	0,5814
Neutral	0,6667	0,7500	0,7059
Sadness	0,8293	0,9189	0,8718

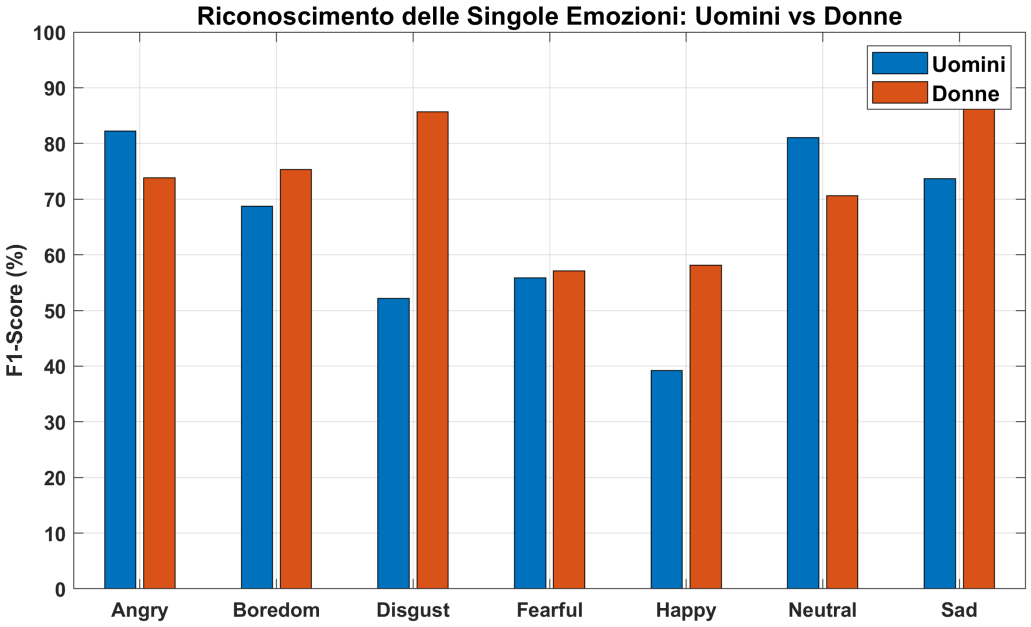


Figura 4.2: Confronto delle prestazioni per genere (65 feature).

L'origine di tale asimmetria trova riscontro nei pattern di errore evidenziati dalle matrici di confusione. Nel campione maschile (Figura: 4.3) si osserva una marcata difficoltà nel distinguere la valenza dei segnali ad alto arousal. La Felicità confluisce infatti spesso nella Paura (29.6%) e nella Rabbia (25.9%), un dato che ripropone ed accentua le criticità già emerse nell'analisi globale. Questa instabilità energetica interessa anche la Paura, le cui predizioni risultano disperse tra Felicità (16.7%) e Rabbia (13.9%). Inoltre, anche negli stati a bassa attivazione si rileva una sovrapposizione significativa, poiché la Noia converge verso la Tristezza (22.9%). Questi

elementi confermano la scarsa specificità dei descrittori spettrali maschili in presenza di un set a 65 feature.

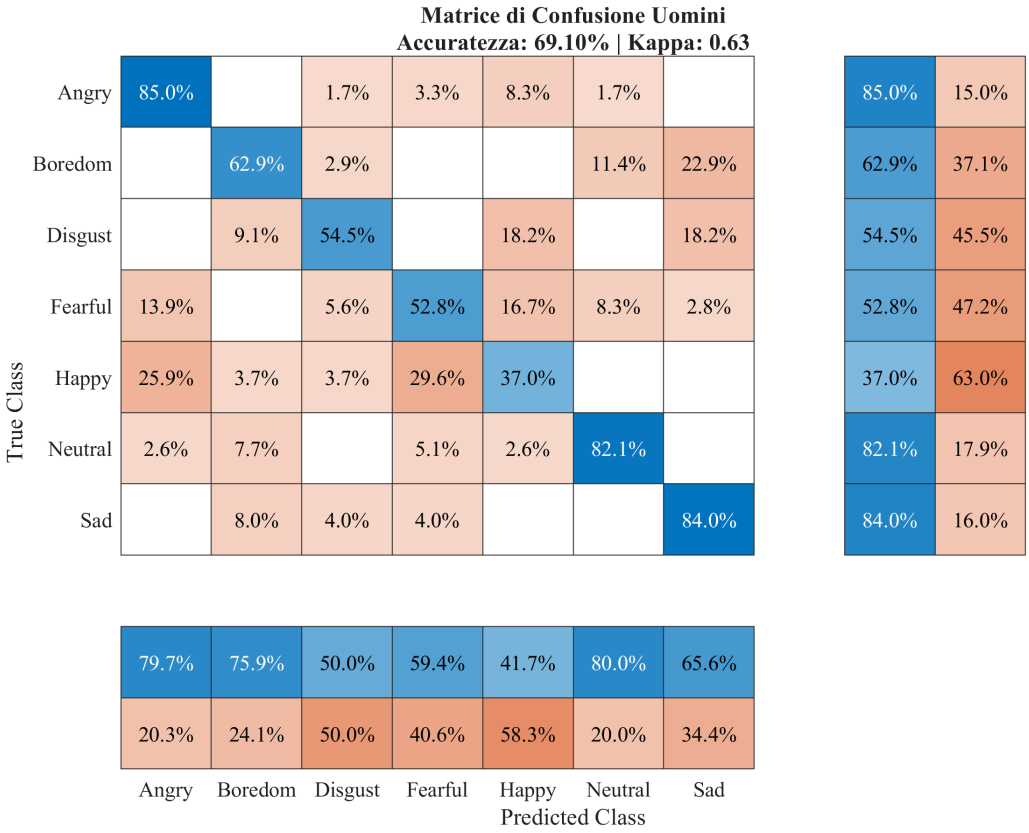


Figura 4.3: Matrice di confusione Uomini del Dataset EMO-DB (65 feature)

Al contrario, la fonazione femminile garantisce una separabilità dei cluster molto più netta (Figura: 4.4). La robustezza di questo gruppo emerge dalle prestazioni su Disgusto e Tristezza, che raggiungono livelli di Recall pari al 94.3% e al 91.9%. Questi valori superano ampiamente i risultati maschili ottenuti nelle medesime categorie. Sebbene persistano alcune interferenze fisiologiche, quali l'erronea classificazione della Noia come stato Neutro (15.2%) o della Paura come Felicità (12.1%), il sistema definisce i confini decisionali in modo più solido. In conclusione, mentre il parlato femminile agevola l'estrazione di biomarcatori emotivi distintivi, la varianza acustica maschile introduce un rumore biometrico che ostacola la generalizzazione di pattern universali.

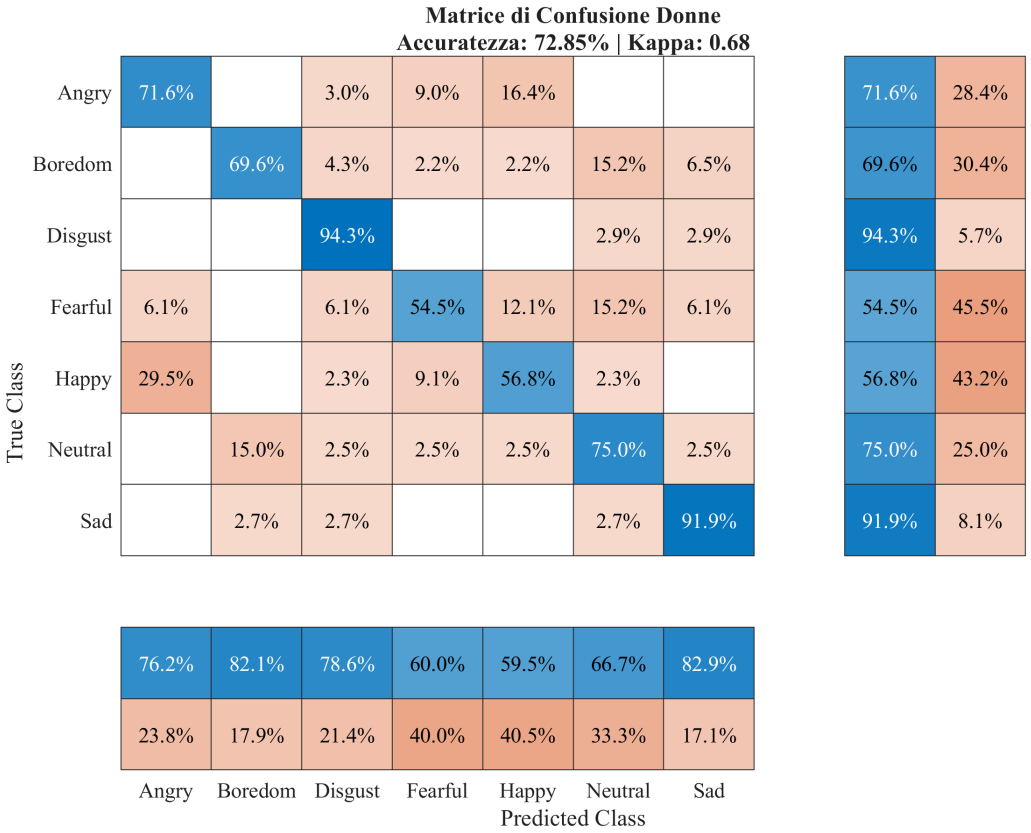


Figura 4.4: Matrice di confusione Donne del Dataset EMO-DB (65 feature)

4.1.2 Analisi della configurazione a 48 Feature

L'integrazione del set ridotto a 48 parametri determina un incremento della stabilità complessiva del sistema. L'accuratezza globale si attesta al 71.78% con un F1-Score di 0.7132 e un coefficiente Kappa di 0.6655. Dall'analisi della Tabella 4.3 si nota come la contrazione dello spazio delle feature agisca in modo positivo sulle classi che risultavano più critiche nella configurazione precedente. In particolare si registra un miglioramento dell'F1-Score per la Paura $(+0.11)$ e per la Felicità $(+0.03)$, i cui valori passano rispettivamente a 0.6721 e 0.5455.

Tabella 4.3: Metriche dettagliate per classe (48 feature)

Classe	Precision	Recall	F1-Score
Rabbia	0,7760	0,7638	0,7698
Noia	0,7250	0,7160	0,7205
Disgusto	0,7674	0,7174	0,7416
Paura	0,7736	0,5942	0,6721
Felicità	0,5417	0,5493	0,5455
Neutro	0,6778	0,7722	0,7219
Tristezza	0,7639	0,8871	0,8209

Osservando la matrice di confusione (Figura 4.5) è possibile quantificare la ridistribuzione della sensibilità tra i diversi cluster emotivi. Il sistema ottiene una maggiore precisione nel distinguere gli stati di Paura $(+5.8\%)$, Felicità $(+5.6\%)$ e Noia $(+4.9\%)$. Tuttavia, questo progresso comporta una flessione nelle categorie che prima apparivano più solide. Il Disgusto subisce la contrazione più evidente della Recall con una perdita del 13.1%, mentre la Rabbia e il Neutro registrano cali più contenuti, rispettivamente dell'1.6% e dell'1.3%. La Tristezza rimane invece la categoria più stabile con una Recall invariata dell'88.71%.

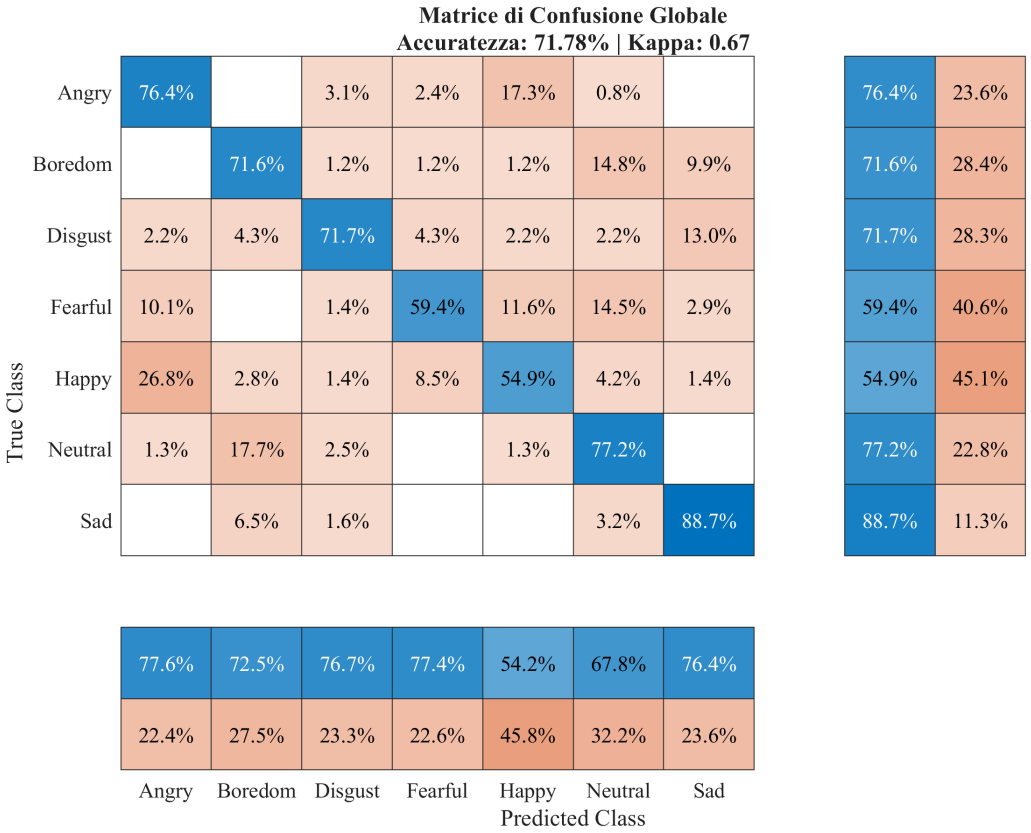


Figura 4.5: Matrice di confusione globale del Dataset EMO-DB (48 feature).

L'aspetto più rilevante dell'ottimizzazione riguarda l'efficace mitigazione del bias biometrico. A differenza della configurazione precedente, il sistema mostra una sostanziale uniformità tra i generi, con l'accuratezza maschile (72.53%) che supera leggermente quella femminile (71.19%). Tale convergenza dimostra che la selezione delle feature abbia rimosso parte dei tratti puramente fisiologici del parlato, permettendo al modello di focalizzarsi su biomarcatori emotivi più universali. Come illustrato nella Figura 4.6, la stabilità complessiva deriva da prestazioni eterogenee tra le diverse categorie emotive. I dati della Tabella 4.4 mostrano che la rete raggiunge la massima risoluzione nella Tristezza femminile, dove la Recall tocca il 97.30%, mentre nelle voci maschili presenta una solidità superiore per le classi Rabbia (86.67%) e Neutro (84.62%).

Tabella 4.4: Confronto delle metriche per genere (48 feature)

Campione Maschile

Classe	Prec.	Rec.	F1
Angry	0,8387	0,8667	0,8525
Boredom	0,7576	0,7143	0,7353
Disgust	0,6364	0,6364	0,6364
Fearful	0,7419	0,6389	0,6866
Happy	0,5263	0,3704	0,4348
Neutral	0,6735	0,8462	0,7500
Sadness	0,6786	0,7600	0,7170

Campione Femminile

Classe	Prec.	Rec.	F1
Angry	0,7143	0,6716	0,6923
Boredom	0,7021	0,7174	0,7097
Disgust	0,8125	0,7429	0,7761
Fearful	0,8182	0,5455	0,6545
Happy	0,5472	0,6591	0,5979
Neutral	0,6829	0,7000	0,6914
Sadness	0,8182	0,9730	0,8889

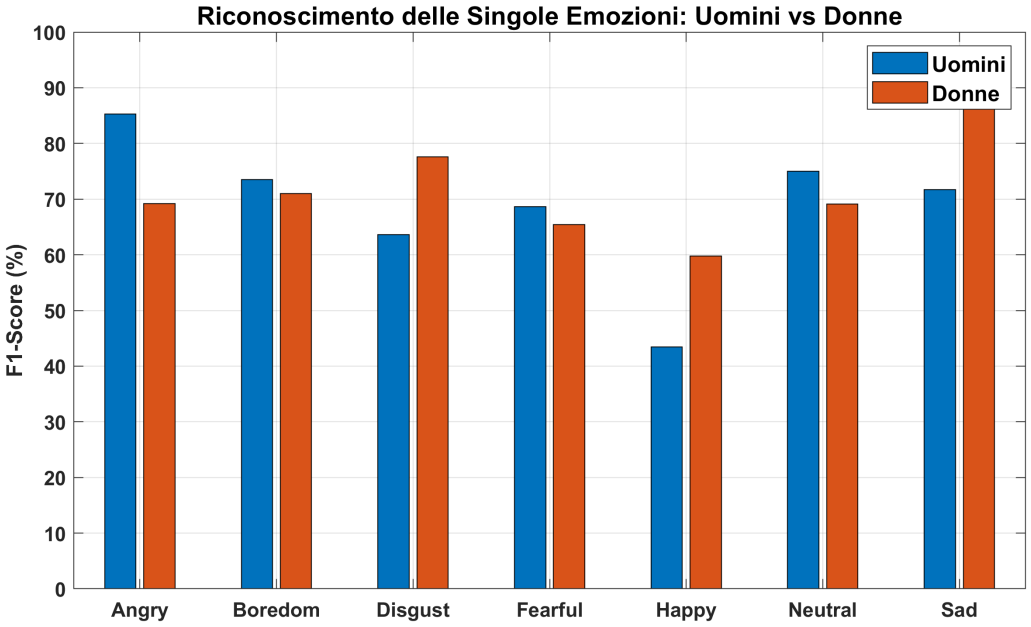


Figura 4.6: Confronto delle prestazioni per genere sul Dataset EMO-DB (48 feature).

Persiste tuttavia la criticità nella rilevazione della Felicità maschile (37.04%) che viene confusa con Paura (29.6%) e Rabbia (25.9%), a causa della difficoltà del sistema nel discriminarne correttamente la valenza. (Figura 4.7). Al contrario, la Felicità femminile è più distinguibile (65.9%) grazie a variazioni di frequenza fondamentale più ampie. La Paura mostra invece errori divergenti. Nelle donne essa viene confusa con la felicità (16.7%) per via della brillantezza vocale, mentre negli uomini tende verso lo stato Neutro (15.3%), indicando una fonazione maschile più attenuata. Rispetto alle 65 feature di partenza, il set ottimizzato garantisce un miglior compromesso

tra efficienza e stabilità, rendendo il sistema più robusto a prescindere dalle caratteristiche biometriche individuali.

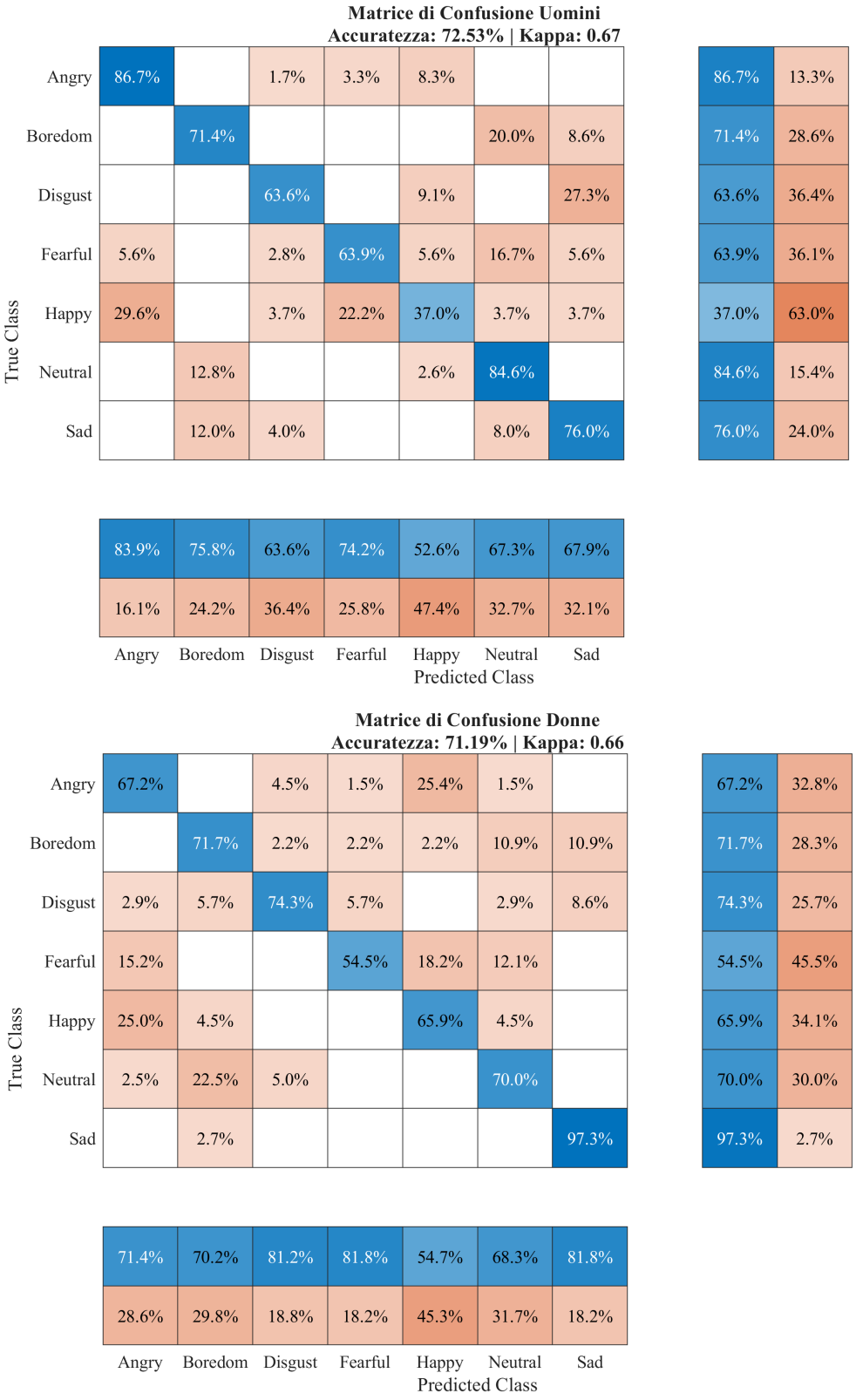


Figura 4.7: Matrici di confusione per genere del Dataset EMO-DB (48 feature)

4.2 Risultati Dataset RAVDESS

4.2.1 Analisi della configurazione a 65 Feature

L'analisi del dataset RAVDESS con la configurazione a 65 feature raggiunge un'accuratezza globale del 69.38%, un F1-Score di 0.6824 e un coefficiente Kappa di 0.6479. Questi risultati superano la soglia della validazione umana del corpus, mediamente attestata al 62%, confermando che l'estrazione automatica delle feature identifica pattern acustici oggettivi non sempre rilevabili mediante la sola analisi percettiva. I risultati ottenuti per singola categoria emotiva (Tabella 4.5) mostrano le prestazioni più elevate per le classi Sorpresa (F1: 0.792), Disgusto (F1: 0.789) e Rabbia (F1: 0.767) favorite da profili energetici marcati.

Classe	Prec.	Rec.	F1
Angry	0,790	0,745	0,767
Calm	0,664	0,823	0,735
Disgust	0,725	0,865	0,789
Fearful	0,688	0,562	0,619
Happy	0,602	0,661	0,630
Neutral	0,682	0,469	0,556
Sad	0,577	0,568	0,572
Surprised	0,846	0,745	0,792

Tabella 4.5: Metriche globali per classe (RAVDESS, 65 feature).

Sebbene la classe Calma registri la sensibilità più alta dell'intero sistema (82.3%), la sua stabilità acustica incide sulla corretta identificazione delle altre emozioni a basso arousal. Come si evince dalla matrice di confusione in Figura 4.8, lo stato Neutro, presenta una Recall limitata al 46.9% a causa della sistematica sovrapposizione con la Calma (32.3%). Un comportamento simile interessa la Tristezza (Recall: 56.8%) che presenta errori equamente distribuiti tra la Calma e la Felicità (10.4%). Tale distribuzione

mostra una difficoltà strutturale nel decodificare la valenza emotiva in assenza di variazioni energetiche marcate. Infine, nonostante l'elevato arousal, la Paura registra una Recall del 56.2% a causa di una parziale sovrapposizione acustica con la Sorpresa (14.6%) e la Tristezza (12.5%).

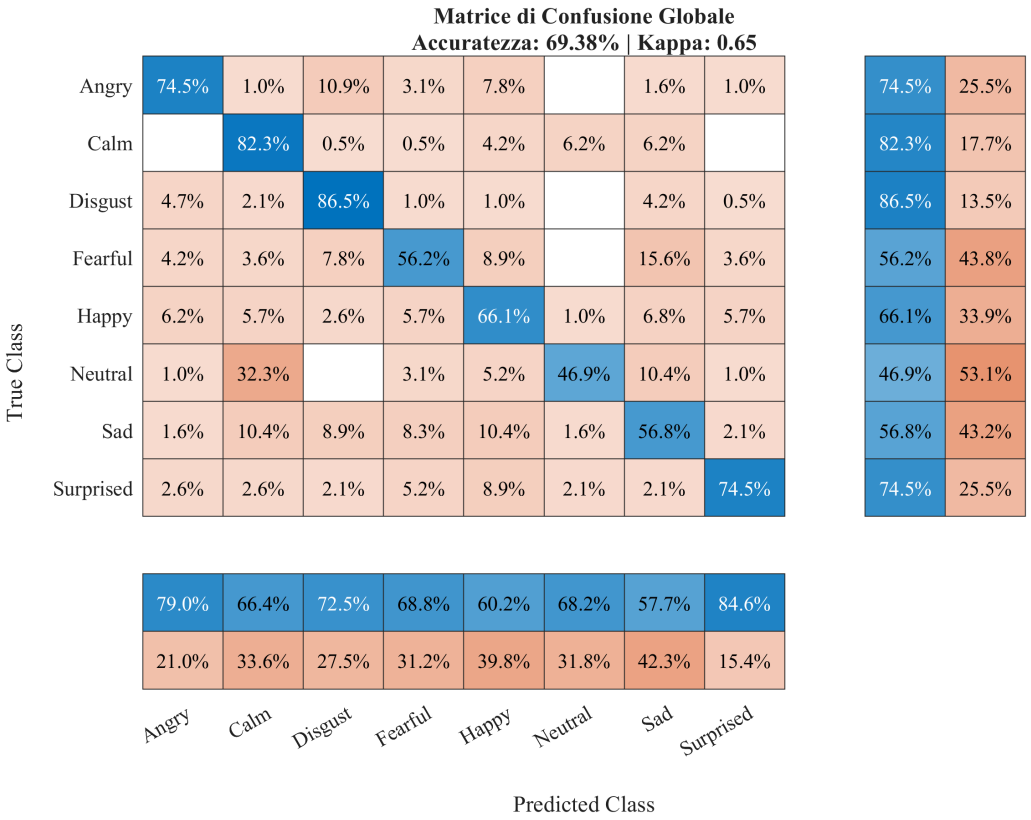


Figura 4.8: Matrice di confusione globale RAVDESS su 65 feature.

L'integrazione dei metadati ha permesso di isolare l'impatto del genere e dell'intensità vocale sulla capacità discriminante del modello. L'analisi per genere rileva un'accuratezza del 72.22% per il campione femminile e del 66.53% per quello maschile. Lo scarto di 5.69 punti percentuali è riconducibile alla diversa risoluzione acustica di specifiche classi emotive (Figura 4.10). In accordo con le tendenze osservate nella matrice di confusione globale, la Calma si conferma l'emozione con la sensibilità più elevata in entrambi i campioni con 80.2% per gli uomini e 84.4% per le donne. Rabbia e Disgusto si rivelano classi particolarmente robuste. Quest'ultimo,

in particolare, mantiene Recall elevate sia negli uomini (90.62%) che nelle donne (82.3%), validando l'efficacia delle 65 feature nel catturare marker di costrizione laringea indipendentemente dalla frequenza fondamentale (f_0). Diverso è il caso della Sorpresa, la cui accuratezza molto alta nel campione femminile (F1= 0.862) decade sensibilmente in quello maschile (F1= 0.722). Tale discrepanza suggerisce che le attrici adottino modulazioni del pitch e delle formanti acusticamente più discriminanti, facilitando il compito del classificatore.

Tabella 4.6: Confronto delle metriche per genere (65 feature)

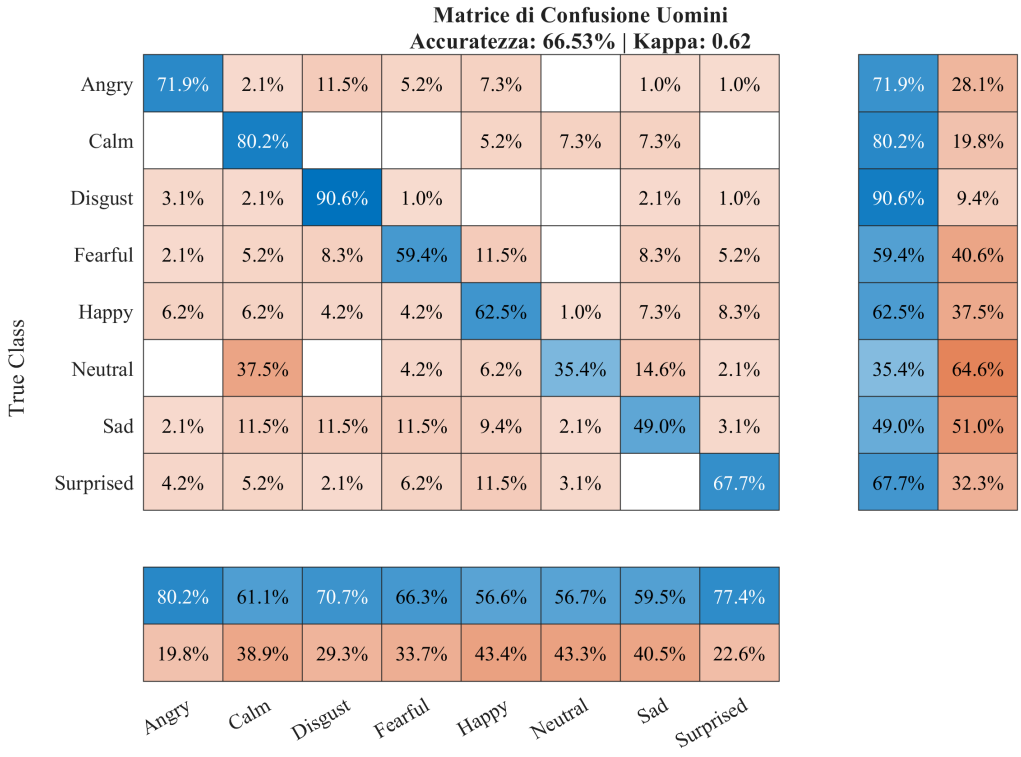
Campione Maschile

Classe	Prec.	Rec.	F1
Angry	0,802	0,719	0,758
Calm	0,611	0,802	0,694
Disgust	0,707	0,906	0,795
Fearful	0,663	0,594	0,626
Happy	0,566	0,625	0,594
Neutral	0,567	0,354	0,436
Sad	0,595	0,490	0,537
Surprised	0,774	0,677	0,722

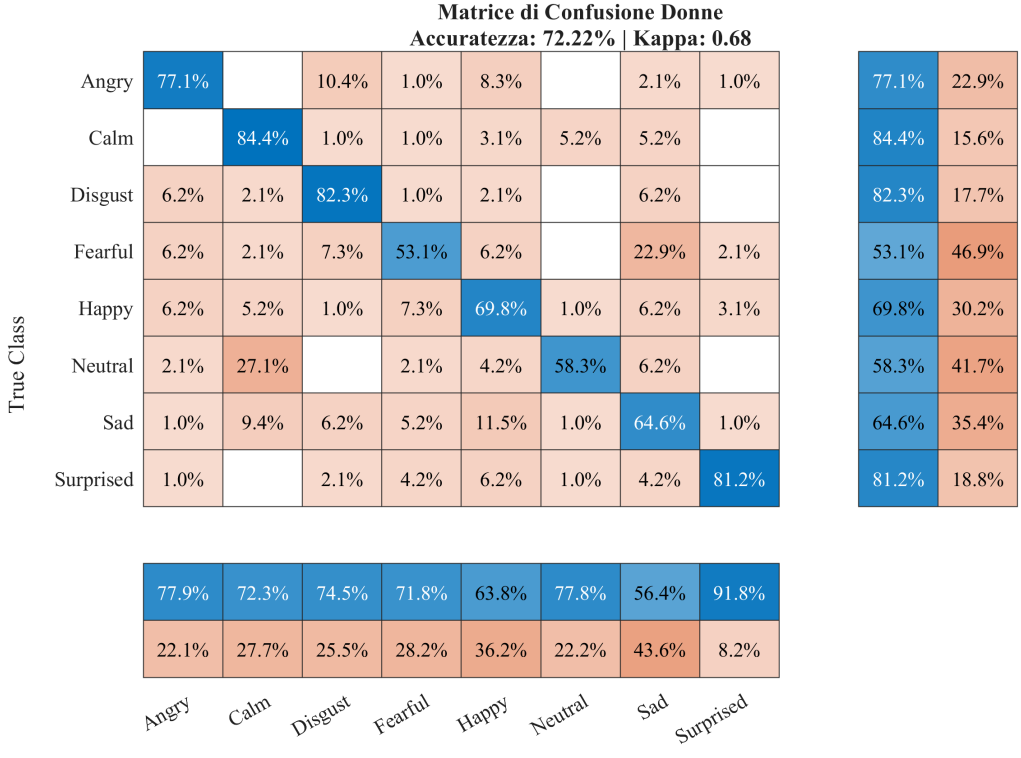
Campione Femminile

Classe	Prec.	Rec.	F1
Angry	0,779	0,771	0,775
Calm	0,723	0,844	0,779
Disgust	0,745	0,823	0,782
Fearful	0,718	0,531	0,611
Happy	0,638	0,698	0,667
Neutral	0,778	0,583	0,667
Sad	0,564	0,646	0,602
Surprised	0,918	0,813	0,862

La criticità maggiore del campione maschile riguarda gli stati a bassa attivazione. Dalla matrice di confusione riportata in Figura a si nota che il Neutro registra una Recall del 35.4%. Nello specifico la percentuale di campioni erroneamente attribuiti alla classe Calma (37.5%) risulta superiore alla quota di quelli correttamente identificati. Tale fenomeno risulta attenuato nel campione femminile, dove la Recall si attesta al 58.3%(Figura b). Un andamento analogo interessa la Tristezza (Recall del 48.96%), che viene frequentemente confusa con la Calma stessa e la Felicità. Questo dato conferma come la voce maschile, risultando più piatta e meno modulata, renda queste emozioni acusticamente troppo simili tra loro per essere separate dal classificatore.



(a) Matrice di Confusione (Uomini).



(b) Matrice di Confusione (Donne).

Figura 4.9: Confronto delle matrici di confusione tra i generi (65 feature).

Le differenze prestazionali tra i due sessi, per ogni singola emozione, sono riassunte visivamente in Figura 4.10

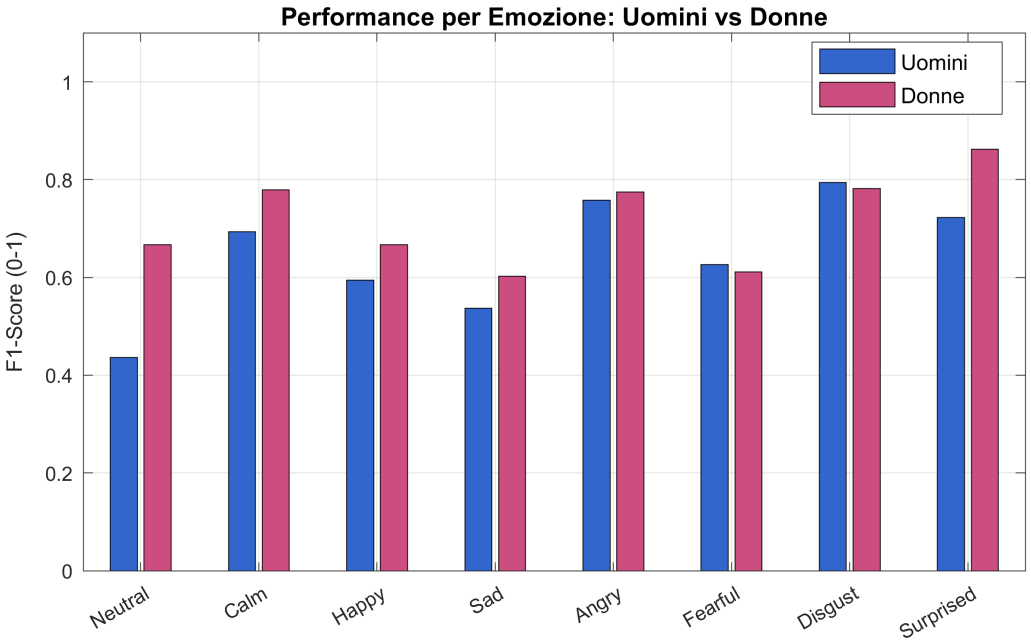


Figura 4.10: Confronto delle performance (F1-Score) tra i generi.

L'analisi sull'intensità vocale dimostra quanto l'energia del segnale incida sulla classificazione. I risultati globali mostrano uno scarto netto tra la modalità Normal ferma al 65.76% e quella Strong che raggiunge il 73.51%.

Tabella 4.7: Confronto delle metriche per intensità (65 feature)

Normal Intensity

Classe	Prec.	Rec.	F1
Angry	0,875	0,656	0,750
Calm	0,487	0,781	0,600
Disgust	0,706	0,875	0,781
Fearful	0,732	0,427	0,539
Happy	0,620	0,698	0,657
Neutral	0,726	0,469	0,570
Sad	0,523	0,604	0,560
Surprised	0,837	0,750	0,791

Strong Intensity

Classe	Prec.	Rec.	F1
Angry	0,734	0,833	0,780
Calm	0,988	0,865	0,922
Disgust	0,745	0,854	0,796
Fearful	0,663	0,698	0,680
Happy	0,583	0,625	0,603
Neutral	0,000	0,000	0,000
Sad	0,654	0,531	0,586
Surprised	0,855	0,740	0,793

In modalità Normal Intensity, la bassa energia limita la separabilità degli stati emotivi (Figura 4.11). Sebbene la Calma registri una Recall del 78.1%,

la sua Precision decade al 48.7% a causa di sistematici falsi positivi derivanti dal Neutro e dalla Tristezza. Anche la Paura risulta fortemente penalizzata (Recall del 42.7%), a dimostrazione della difficoltà di isolare i tratti della tensione vocale in assenza di un volume sonoro sostenuto.

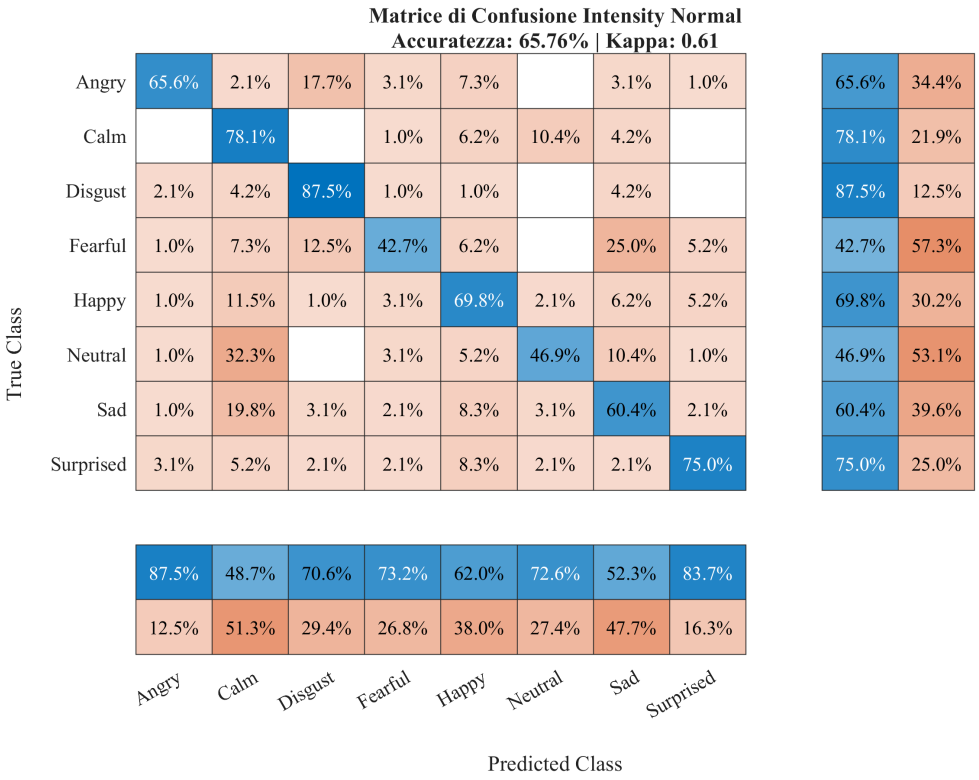


Figura 4.11: Risultati della classificazione per la modalità Normal Intensity

Al contrario, la modalità Strong Intensity (Figura 4.12) mostra prestazioni molto più stabili. La Calma raggiunge un F1-Score di 0.922, anche grazie all'assenza della classe neutra che rende la sua firma spettrale meno soggetta a confusioni. La Rabbia trae vantaggio dall'incremento energetico e raggiunge una Recall dell'83.3%, poiché l'enfasi vocale produce una firma acustica estremamente marcata e facilmente discriminabile dal classificatore. In questo scenario la Tristezza registra la Recall minima del 53.1% disperdendo gli errori verso Disgusto e Paura (14.6%) e Felicità (12.5%). Tale frammentazione indica che l'incremento del volume altera i

tratti acustici tipici della bassa attivazione, rendendo il segnale assimilabile a manifestazioni di tensione o elevata eccitazione.

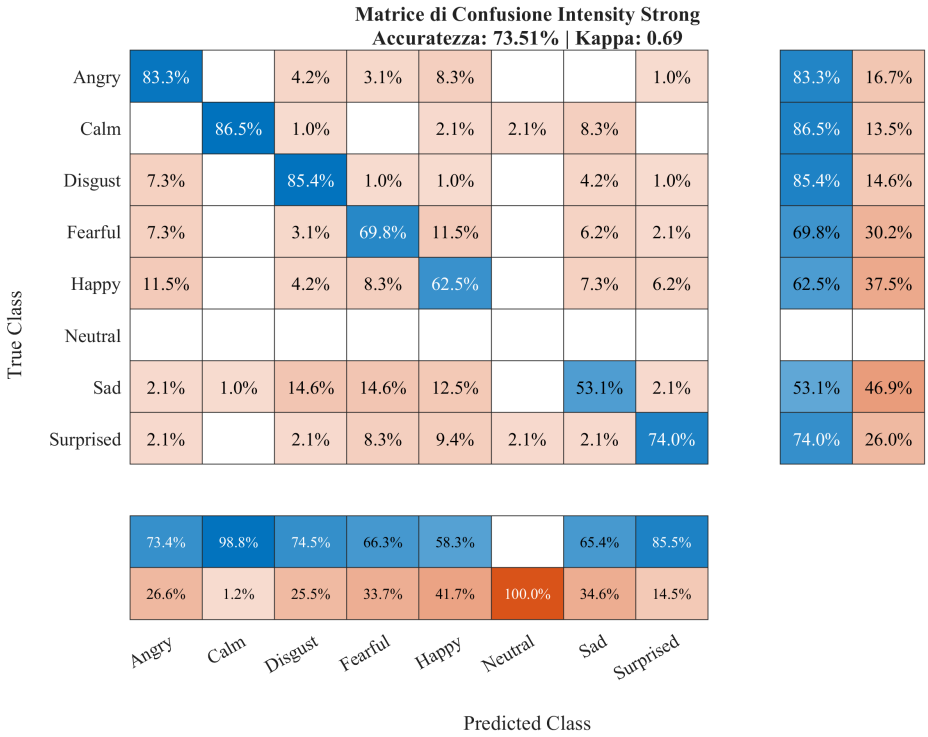


Figura 4.12: Risultati della classificazione per la modalità Strong Intensity

Il confronto tra le singole classi è sintetizzato nella Figura 4.13.

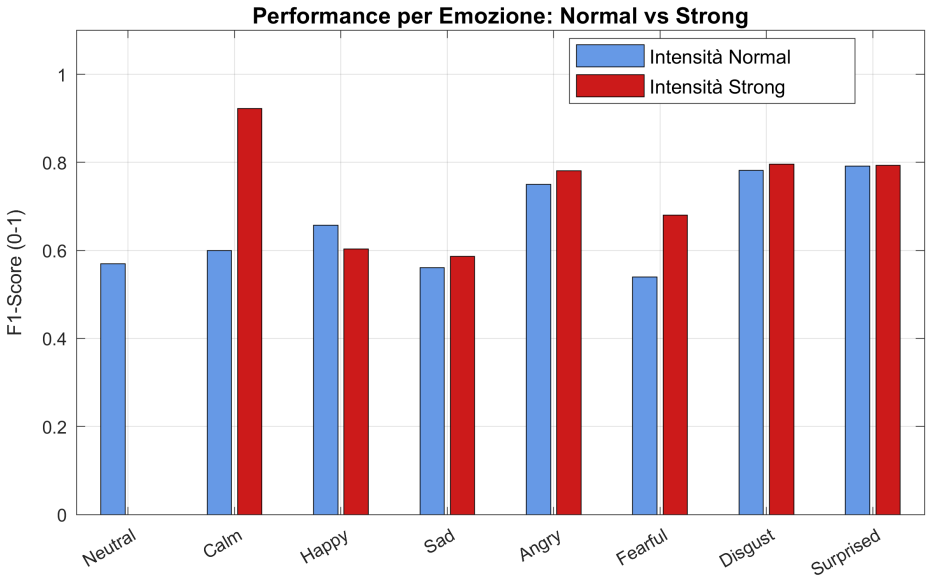


Figura 4.13: Confronto delle performance (F1-Score) tra le diverse intensità vocali

4.2.2 Analisi della configurazione a 48 feature

La configurazione a 48 feature registra un'accuratezza globale del 69.79% (Figura: 4.14). Rispetto al modello a 65 parametri, si osserva un lieve miglioramento delle performance che valida l'efficacia della riduzione dimensionale. L'eliminazione di 17 feature ha infatti ha ottimizzato lo spazio decisionale, eliminando il rumore senza compromettere il contenuto informativo. A livello globale, la Rabbia (F1: 0.793) e il Disgusto (F1: 0.783) si confermano le classi strutturalmente più solide (Figura 4.8). Permangono tuttavia le criticità sulle emozioni a bassa energia quali lo stato Neutro che ha una Recall del 52.08%, a causa della persistente sovrapposizione con la categoria Calma. Tale dato indica che l'ambiguità tra queste due classi rappresenta un limite strutturale legato alla natura dei segnali acustici del dataset piuttosto che alla complessità del modello di classificazione.

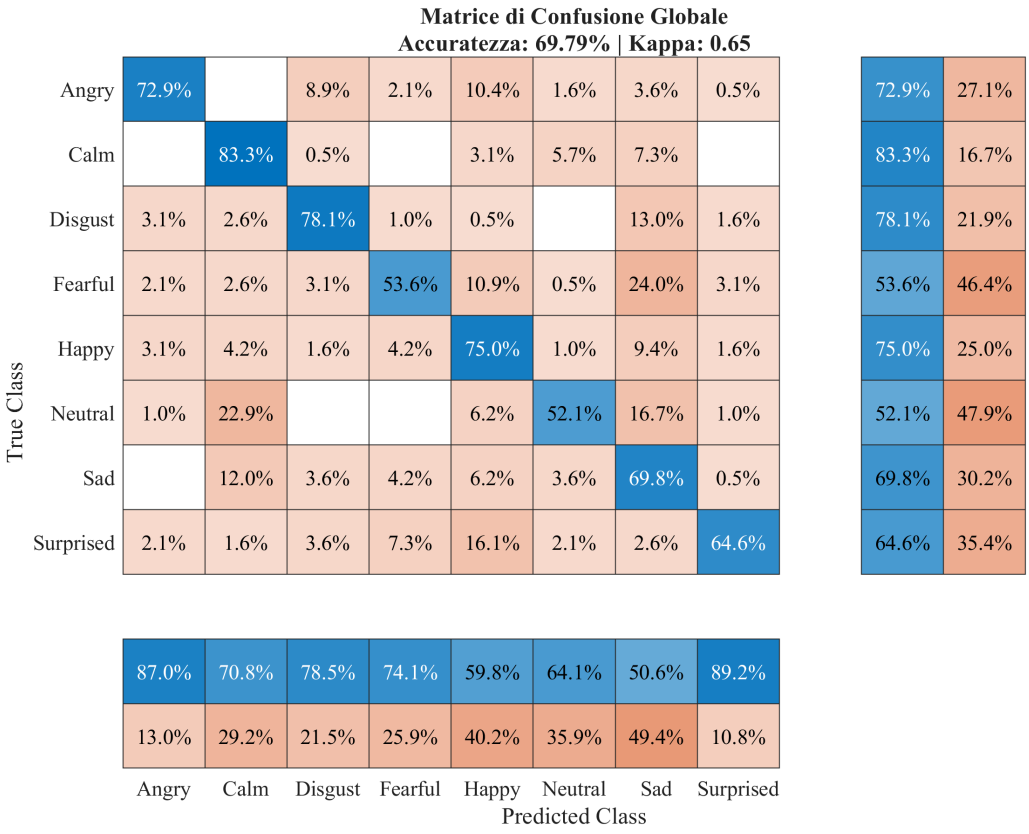


Figura 4.14: Matrice di confusione globale RAVDESS (48 feature)

Tabella 4.8: Metriche globali della classificazione con 48 feature

Classe	Prec.	Rec.	F1
Angry	0,870	0,729	0,793
Calm	0,708	0,833	0,766
Disgust	0,785	0,781	0,783
Fearful	0,741	0,536	0,622
Happy	0,598	0,750	0,665
Neutral	0,641	0,521	0,575
Sad	0,506	0,698	0,586
Surprised	0,892	0,646	0,749

La scomposizione per genere conferma e accentua le asimmetrie rilevate nella configurazione precedente. Il campione femminile raggiunge un'accuratezza del 73.19% (in crescita rispetto al precedente 72.22%), mentre quello maschile rimane stabile intorno al 66%. L'invarianza di tale divariom nonostante la contrazione dell matrice delle feature, suggerisce che la discrepanza derivi da fattori acustici intrinseci del corpus. Nel campione femminile (Figura 4.15), i punteggi F1 indicano una maggiore solidità per le categorie Calma, Sorpresa e Rabbia con valori rispettivamente di 0.825, 0.82 e 0.804. La Sorpresa, in particolare, mostra una Precision del 92.2% a testimonianza di una elevata capacità discriminativa del sistema. Per la Calma si osserva una Recall dell'88.5% a fronte di una Precision di 0.772 confermando la tendenza del modello a includere campioni appartenenti allo stato Neutro il quale si ferma a un punteggio F1 di 0.719. Per il campione maschile (Figura 4.16) permangono le difficoltà nell'identificazione degli stati a bassa attivazione fisiologica. Lo stato Neutro registra il valore F1 minimo di 0.424 con una Recall del 37.5%. Al contrario, il Disgusto e la Rabbia si confermano i cluster più stabili per le voci maschili con punteggi F1 rispettivamente di 0.786 e 0.781. In conclusione, la configurazione a 48 feature, pur ottimizzando l'identificazione emotiva nei segnali del campione femminile, non risulta sufficiente a mitigare le sovrapposizioni nelle

frequenze gravi del parlato maschile le quali si confermano il principale fattore limitante della sperimentazione. Il quadro complessivo dei risultati descritti troverà una sintesi visiva nel confronto per singola categoria proposto successivamente in Figura 4.17

Tabella 4.9: Metriche per il campione femminile (48 feature).

Classe	Prec.	Rec.	F1
Angry	0,867	0,750	0,804
Calm	0,772	0,885	0,825
Disgust	0,825	0,739	0,779
Fearful	0,796	0,489	0,606
Happy	0,617	0,822	0,705
Neutral	0,780	0,666	0,718
Sad	0,514	0,729	0,603
Surprised	0,922	0,739	0,820

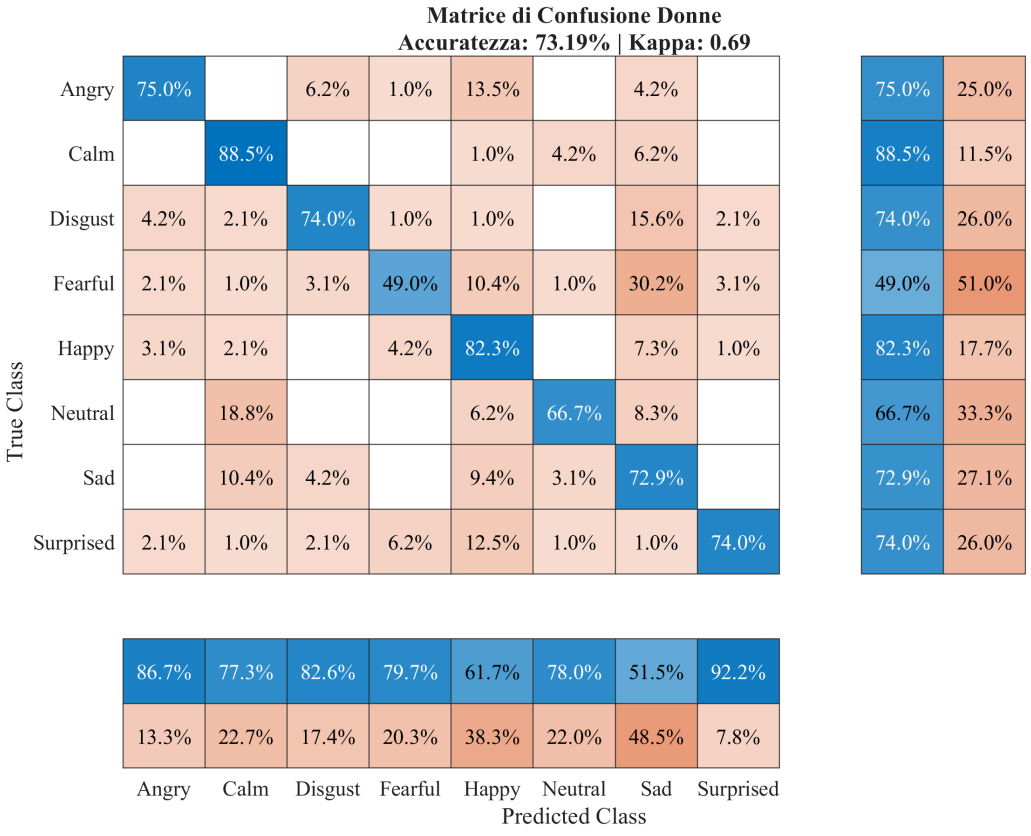


Figura 4.15: Analisi delle performance per il campione femminile (48 feature)

Tabella 4.10: Metriche per il campione maschile (48 feature).

Classe	Prec.	Rec.	F1
Angry	0,872	0,708	0,782
Calm	0,647	0,781	0,708
Disgust	0,752	0,823	0,786
Fearful	0,700	0,583	0,636
Happy	0,575	0,677	0,622
Neutral	0,486	0,375	0,424
Sad	0,496	0,667	0,569
Surprised	0,855	0,552	0,671

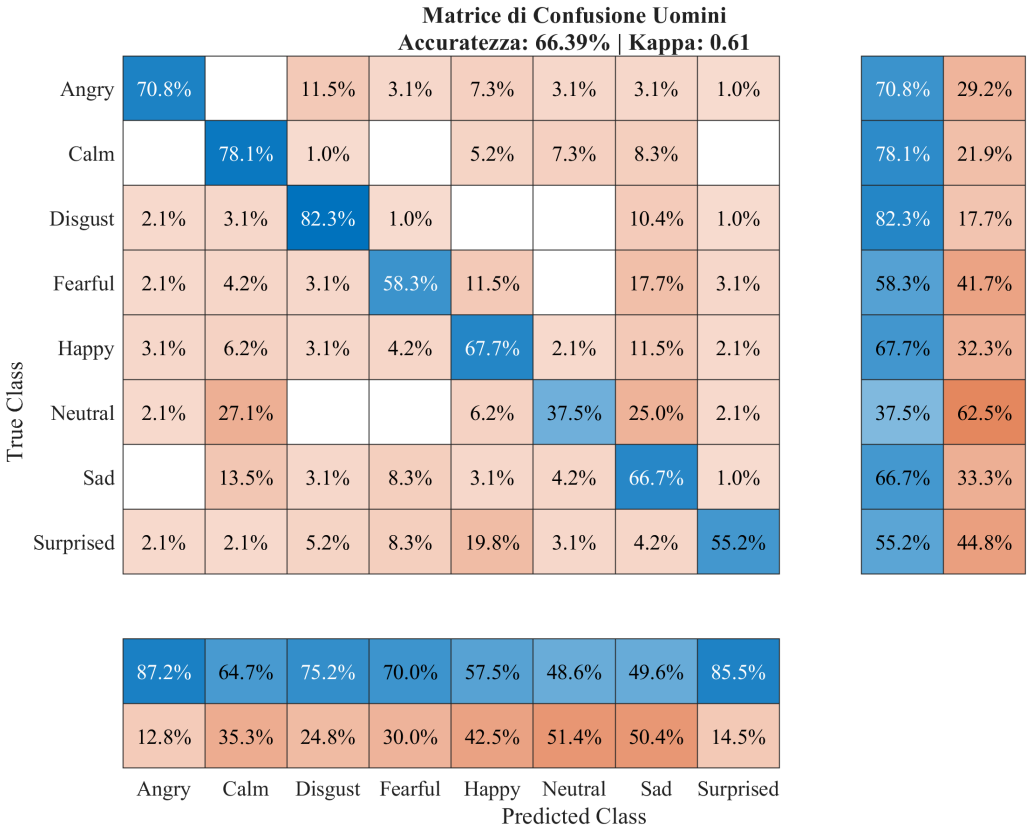


Figura 4.16: Analisi delle prestazioni per il campione maschile (48 feature).

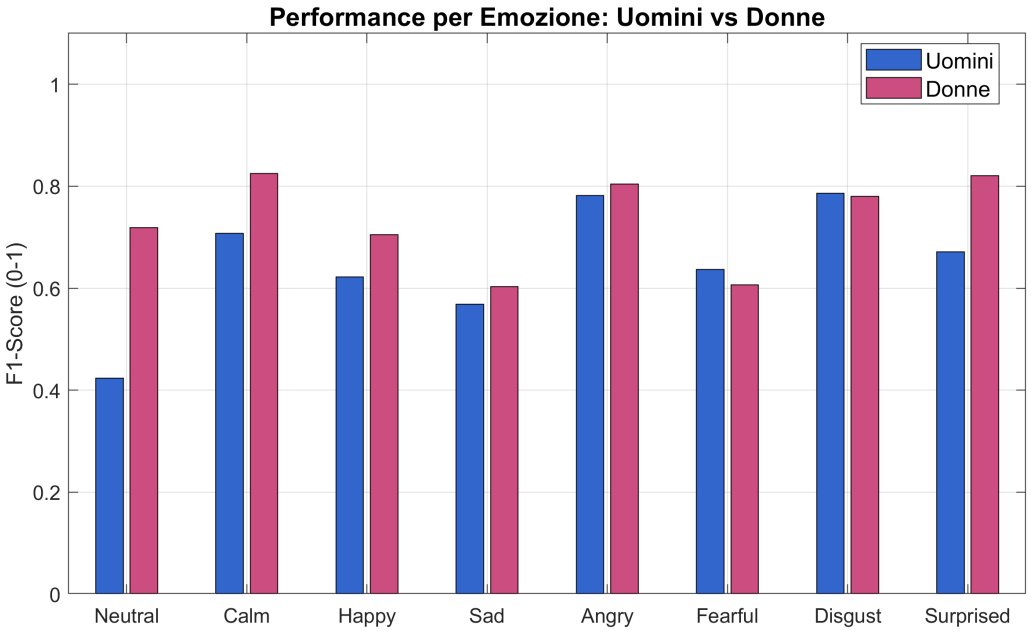


Figura 4.17: Confronto delle performance tra i generi (48 feature).

L'impiego del set ridotto a 48 parametri ha prodotto risultati differenti tra i due livelli di intensità rispetto alla configurazione a 65 feature. Il campione ad intensità normale ha registrato un decremento dell'accuratezza dell'1.70% attestandosi al 64.06%, mentre il gruppo ad alta intensità ha ottenuto un incremento dell'1.34%, raggiungendo il 76.34%. Tale andamento indica che la selezione delle feature ha migliorato la robustezza del modello sui segnali più energici pur comportando una lieve perdita di precisione nei contesti a bassa attivazione. Nell'analisi dell'intensità normale (Figura: 4.18) le categorie di Rabbia e Sorpresa mantengono una buona stabilità con punteggi F1 di 0.793 e 0.711. La Tristezza registra invece l'F1-score minimo (0.498) a causa di un elevato numero di falsi positivi riscontrabili nella matrice di confusione. Nel sottoinsieme ad alta energia (Figura:4.19), la maggiore enfasi vocale favorisce una distinzione più netta tra i profili, con risultati particolarmente solidi per le classi Disgusto (0.826) Rabbia (0.793) e Sorpresa (0.840). In questo contesto, la Calma raggiunge un F1-score di 0.946 grazie alla già citata assenza strutturale dello stato Neutro. Tale dinamica trova riscontro nella Figura 4.20 dove l'aumento

dell'ampiezza sonora agevola la separazione dei cluster emotivi sull'intero spettro analizzato.

Tabella 4.11: Metriche per il campione a intensità Normale (48 feature).

Classe	Prec.	Rec.	F1
Neutral	0,694	0,521	0,595
Calm	0,529	0,750	0,621
Happy	0,595	0,688	0,638
Sad	0,405	0,646	0,498
Angry	0,918	0,698	0,793
Fearful	0,740	0,562	0,639
Disgust	0,861	0,646	0,738
Surprised	0,735	0,688	0,711

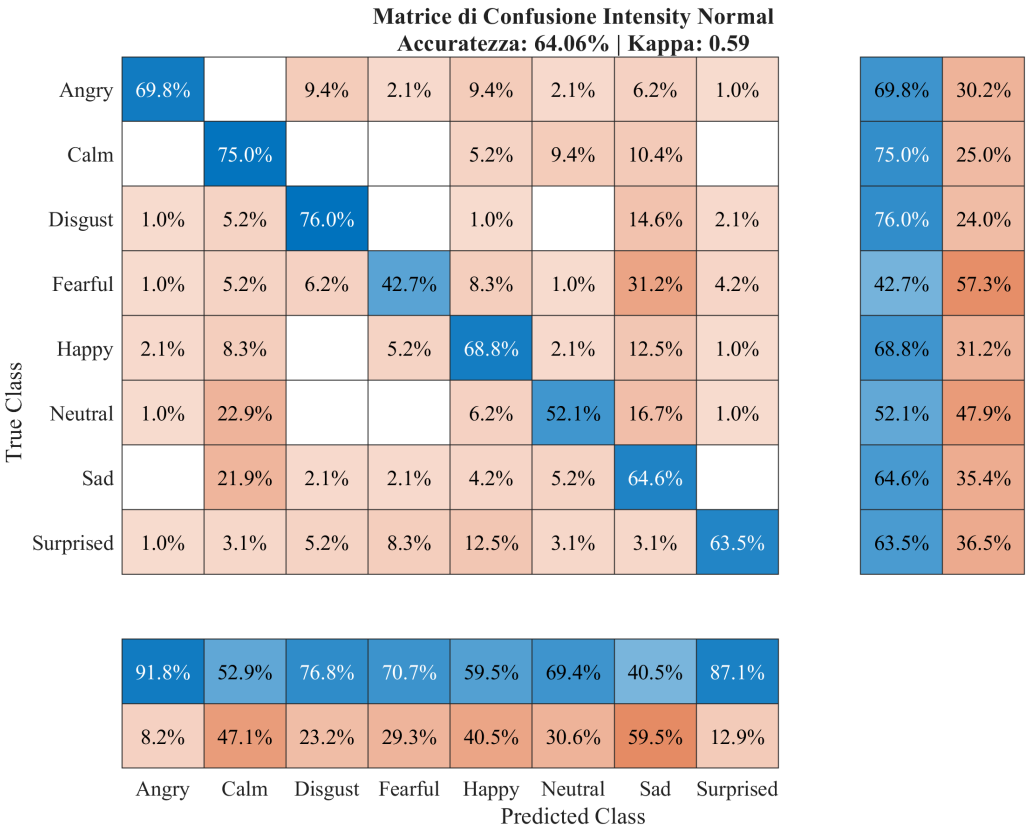


Figura 4.18: Analisi delle prestazioni sul RAVDESS l'intensità Normal (48 feature).

Tabella 4.12: Metriche per il campione a intensità Forte (48 feature).

Classe	Prec.	Rec.	F1
Neutral	0,000	0,000	0,000
Calm	0,978	0,917	0,946
Happy	0,600	0,812	0,690
Sad	0,643	0,750	0,692
Angry	0,830	0,760	0,793
Fearful	0,765	0,646	0,701
Disgust	0,864	0,792	0,826
Surprised	0,808	0,875	0,840

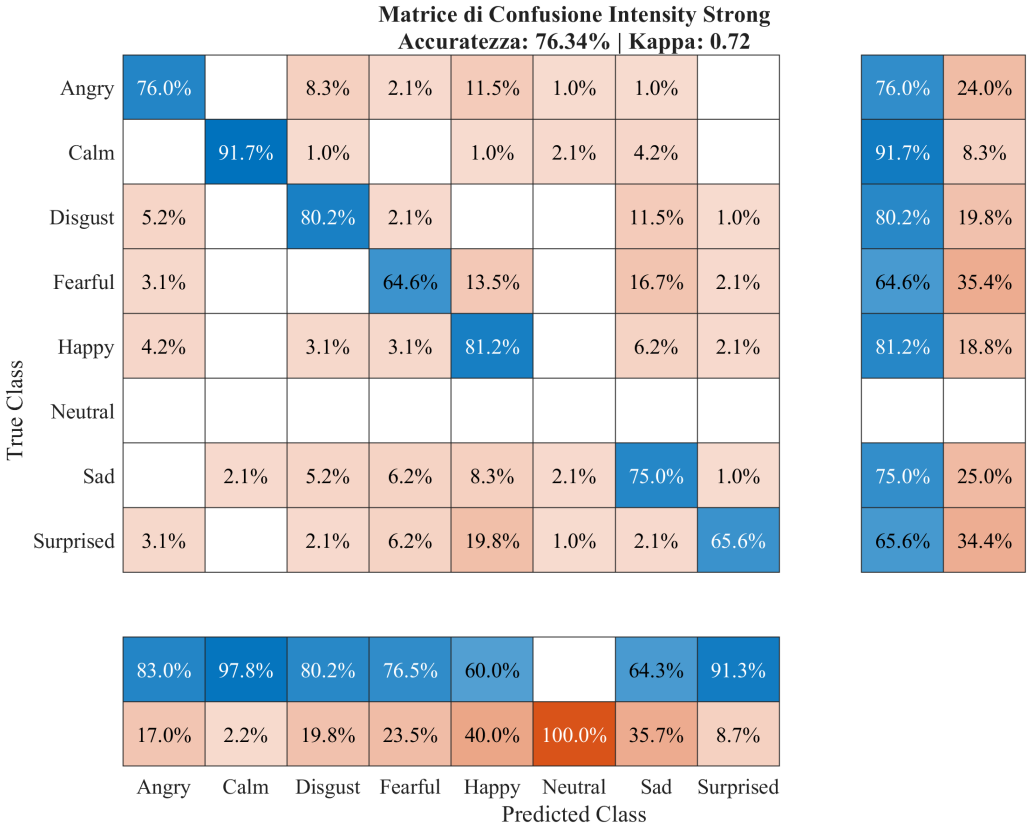


Figura 4.19: Analisi delle prestazioni sul RAVDESS l'intensità Strong 48 feature).

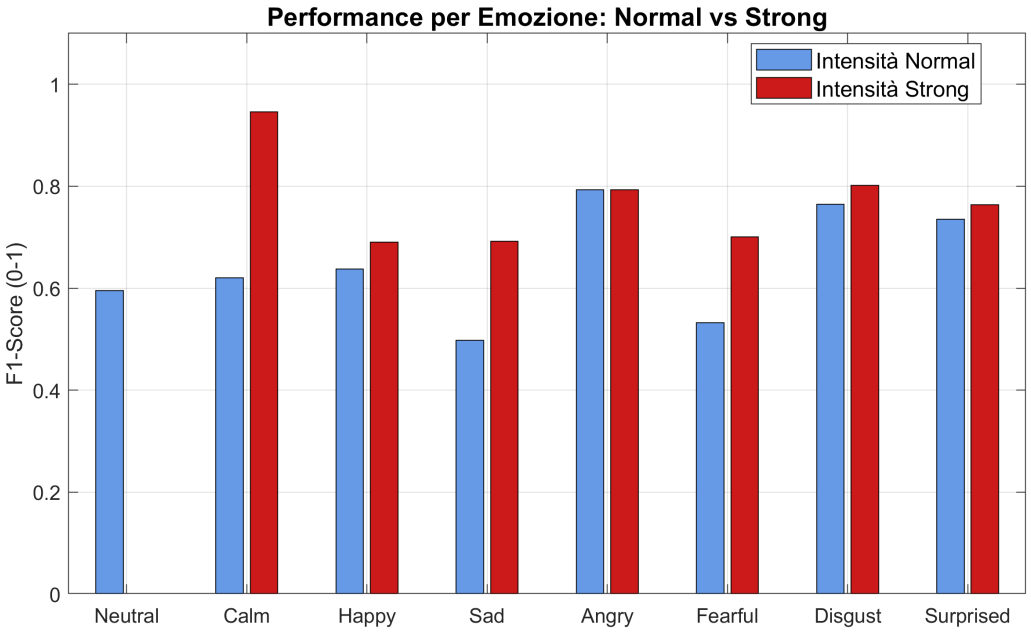


Figura 4.20: Confronto delle performance tra intensità Normale e Forte (48 feature).

4.3 Risultati Dataset clinici

L'efficacia dell'architettura VGGish nella caratterizzazione del segnale vocale clinico è stata approfondita confrontando le prestazioni del modello su popolazioni eterogenee per età e profilo funzionale, al fine di validare la capacità di generalizzazione dei parametri estratti. L'analisi si è focalizzata inizialmente sul dataset pediatrico ASDBank, composto da 84 soggetti (46 ASD e 38 TD), i cui risultati sono sintetizzati nella Tabella 4.13.

Tabella 4.13: Prestazioni globali del modello VGGish sul dataset pediatrico ASDBank.

Metrica	**Valore**
Accuracy	78,6%
Precision	82%
Recall	78%
AUC	(0,87)

I dati riportati evidenziano un'accuratezza del 78.6% e un'AUC di 0.87, confermando una solida capacità discriminativa del classificatore. Il bilan-

ciamento tra Precision (82%) e Recall (78%) indica un sistema capace di identificare efficacemente i tratti patologici della voce infantile, limitando sensibilmente l'incidenza dei falsi positivi sulla classe di controllo. A supporto di queste metriche, la Figura 4.21 illustra la curva ROC e la relativa matrice di confusione. Quest'ultima evidenzia la corretta individuazione di 36 soggetti ASD su 46 e 30 soggetti TD su 38, con una distribuzione degli errori (10 falsi negativi e 8 falsi positivi) che riflette la stabilità analitica del modello su entrambi i gruppi. L'andamento della curva ROC, caratterizzato da una pendenza elevata già per bassi tassi di falsi positivi, valida ulteriormente l'utilizzo degli embedding audio come biomarcatori digitali affidabili per l'identificazione dello spettro autistico in età pediatrica.

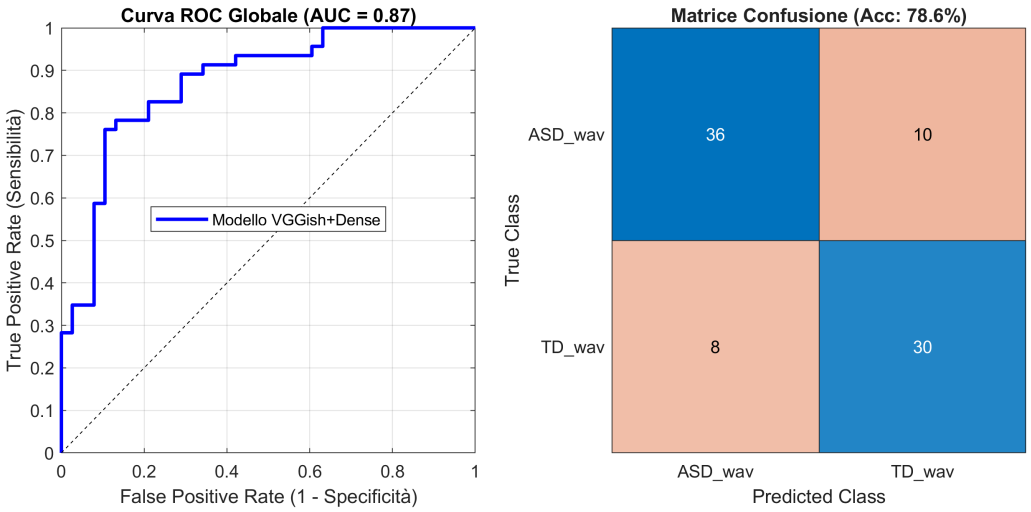


Figura 4.21: Risultati Dataset ASDBank

L'analisi delle prestazioni sulla popolazione composta da adulti ad alto funzionamento documenta un task diagnostico sensibilmente più complesso rispetto a quello pediatrico. Le metriche registrate per questo secondo gruppo sperimentale sono riportate nella Tabella 4.14. I valori ottenuti mostrano un'accuratezza del 70.49% e un'AUC di 0.6976. Il divario tra Precision (78.95%) e Recall (75%) indica che il sistema identifica con discreta affidabilità i casi positivi, pur non riuscendo a riconoscere il 25%

Tabella 4.14: Parametri prestazionali del modello VGGish sul dataset adulti ad alto funzionamento

Metrica	Valore
Accuracy	70,49%
Precision	78,95%
Recall	75,00%
F1-Score	76,92%
AUC	0,6976

dei soggetti appartenenti alla classe AUT. Questo rendimento riflette una minore separabilità tra le classi nello spazio delle feature, dove le tracce acustiche dei due gruppi mostrano una sovrapposizione superiore rispetto al campione pediatrico. Per analizzare puntualmente la distribuzione delle predizioni, nella Figura 4.22 sono riportate la matrice di confusione e la curva ROC.

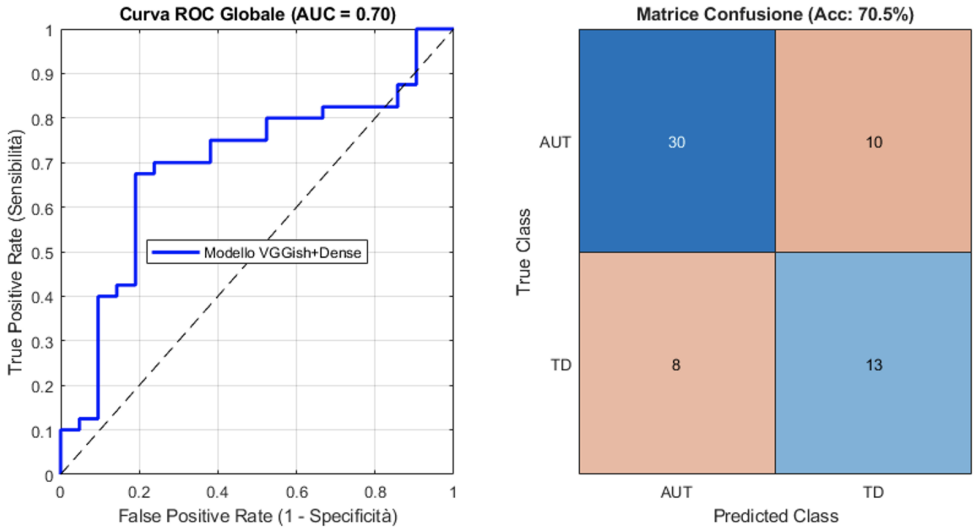


Figura 4.22: Risultati Dataset Sperimentale

Dalla matrice di confusione emerge la corretta classificazione di 30 soggetti AUT su 40 e di 13 soggetti TD su 21. La rilevazione di 8 falsi positivi evidenzia che una parte della popolazione neurotipica adulta presenta caratteristiche vocali che il modello non discrimina nettamente dai tratti clinici. Tale risultato riflette verosimilmente l'impatto del masking e delle strategie

compensative sociali che, in età adulta, tendono ad attenuare le divergenze prosodiche tipiche dell'infanzia. In conclusione, sebbene la capacità di adattamento fonetico e sociale dei soggetti renda il confine decisionale meno netto rispetto al target pediatrico, il valore dell'F1-Score (76.92%) conferma che gli embedding VGGish conservano una valenza informativa utile per l'identificazione di biomarcatori vocali anche in presenza di atipicità estremamente sfumate.

5. Conclusioni

Il presente studio ha permesso di validare l' impiego di architetture di Deep Learning per l'elaborazione del segnale vocale, con l'obiettivo di individuare biomarcatori oggettivi in ambito sia affettivo che clinico. Nella fase dedicata al riconoscimento delle emozioni (SER) è stato valutato l'impatto della selezione dei parametri acustici confrontando un set di 65 feature con uno ottimizzato di 48 elementi. L'integrazione del set ridotto ha prodotto incrementi marginali dell'accuratezza globale (71.78% per EMO-DB e 69.79% per RAVDESS) dimostrando come l'esclusione di 17 parametri ridondanti migliori l'efficienza del sistema senza sacrificarne la capacità predittiva. L'efficacia di questa selezione è risultata tuttavia legata alla natura specifica dei dati. Nel dataset tedesco EMO-DB la rimozione delle feature superflue ha consentito la minimizzazione del bias biometrico grazie all'allineamento dei risultati tra i generi. Al contrario, nel dataset RAVDESS lo scarto tra i sessi è rimasto costante a testimonianza del fatto che le divergenze dipendono da variabili acustiche intrinseche degli attori piuttosto che dalla dimensionalità delle feature. Questa analisi attesta che la scelta delle feature acustiche rimane una sfida ancora aperta data l'impossibilità di definire un set universale capace di operare con la medesima efficacia tra differenti domini linguistici e contesti di acquisizione. La solidità dei modelli trova conferma nel superamento della soglia di riconoscimento umana stimata intorno al 60% per la sola componente audio. Il divario rispetto ai risultati superiori al 95% riportati in parte della letteratura è giustificato dall' applicazione del protocollo Nested Cross Validation a 10 fold. Tale metodo ha impedito alle reti di memorizzare le identità vocali degli attori garantendo una valutazione precisa della reale capacità di generalizzazione del sistema. In ambito clinico, l'architettura VGGish

ha superato i limiti dei descrittori convenzionali, spesso inaffidabili a causa della sensibilità al rumore e del rischio di overfitting su campioni limitati. L'uso del transfer learning ha favorito invece l' estrazione di rappresentazioni profonde con il conseguente raggiungimento di una AUC di 0.87 nella popolazione pediatrica e di 0.70 negli adulti ad alto funzionamento. Tale scarto riflette l'impatto del social camouflaging poiché in età adulta i meccanismi di compensazione rendono le atipicità vocali acusticamente sovrapponibili ai profili neurotipici. Sebbene i risultati ottenuti siano promettenti, l'analisi della sola componente vocale risulta parziale per un supporto diagnostico pienamente efficace. Una reale evoluzione dello screening richiede l'integrazione di sistemi capaci di rilevare anche il contesto semantico e il significato del discorso. L'unione tra la microstruttura del segnale e l'analisi automatizzata del linguaggio favorirà una valutazione oggettiva delle difficoltà comunicative e fornirà strumenti più precisi per superare i meccanismi di compensazione dei pazienti.

Bibliografia

[1] Rosalind Wright Picard. *Affective Computings*. MIT Press, 1997.

[2] R. Cowie et al. «Emotion recognition in human-computer interaction». In: *IEEE Signal Processing Magazine* 18.1 (2001), pp. 32–80.

[3] Antonio Damasio. *Descartes' Error: Emotion, Reason, and the Human Brain.* New York: Avon Books, 1994. ISBN: 0-380-72647-5.

[4] Paul Ekman. «An argument for basic emotions». In: *Cognition and Emotion* 6.3-4 (1992), pp. 169–200. DOI: 10.1080/02699939208411068.

[5] James A. Russell. «A circumplex model of affect». In: *Journal of Personality and Social Psychology* 39.6 (1980), pp. 1161–1178. DOI: 10.1037/h0077714.

[6] Mehmet Berkehan Akçay e Kaya Oğuz. «Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers». In: *Speech Commun.* 116 (2020), pp. 56–76. DOI: 10.1016/j.specom.2019.12.001.

[7] Björn W. Schuller e Anton Batliner. *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing*. Chichester, UK: John Wiley & Sons, Ltd, 2014. ISBN: 978-1-119-97136-8.

[8] Nicholas Cummins et al. «A review of depression and suicide risk assessment using speech analysis». In: *Speech Commun.* 71 (2015), pp. 10–49. DOI: 10.1016/j.specom.2015.03.004.

[9] American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders (DSM-5-TR)*. 5th, Text Revision. Washington, DC: American Psychiatric Association, 2022. ISBN: 978-0-89042-576-3. DOI: 10.1176/appi.books.9780890425763.

[10] Holly Hodges, Catherine Fealko e Neelkamal Soares. «Autism spectrum disorder: causes, diagnosis, and treatments». In: *Children* 7.2 (2020), p. 14. DOI: 10.3390/children7020014.

[11] Ministero della Salute. *Autismo*. Consultato il: 2026-04-09. 2023. URL: https://www.salute.gov.it/new/it/tema/salute-mentale/autismo/.

[12] Catherine Lord et al. «Autism diagnostic observation schedule: ADOS-2». In: *Western Psychological Services* (2012).

[13] Catherine Lord et al. «The Lancet Commission on the future of care and clinical research in autism». In: *The Lancet* 399.10321 (2022), pp. 271–334. DOI: 10.1016/S0140-6736(21)01541-5.

[14] Riccardo Fusaroli et al. «Is voice a marker for Autism spectrum disorder? A systematic review and meta-analysis». In: *Autism Research* 10.3 (2017), pp. 384–407. DOI: 10.1002/aur.1678.

[15] Sudarsana Reddy Kadiri, Paavo Alku e B. Yegnanarayana. «Extraction and Utilization of Excitation Information of Speech: A Review». In: *Proceedings of the IEEE* 109.11 (2021), pp. 1814–1846. DOI: 10.1109/JPROC.2021.3106857.

[16] Gunnar Fant. *Acoustic Theory of Speech Production: With Calculations Based on X-Ray Studies of Russian Articulations*. A cura di Roman Jakobson e C. H. van Schooneveld. Second printing. Vol. 2. Description and Analysis of Contemporary Standard Russian. The Hague, Paris: Mouton, 1970.

[17] Rainer Banse e Klaus R. Scherer. «Acoustic Profiles in Vocal Emotion Expression». In: *Journal of Personality and Social Psychology* 70.3 (1996), pp. 614–636. DOI: 10.1037/0022-3514.70.3.

[18] Lawrence Rabiner e Ronald Schafer. *Theory and applications of digital speech processing*. Prentice Hall Press, 2010.

[19] Florian Eyben et al. «The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing». In: *IEEE Transactions on Affective Computing* 7.2 (2016), pp. 190–202. DOI: 10.1109/TAFFC.2015.2457417.

[20] Thomas F. Quatieri. *Discrete-Time Speech Signal Processing: Principles and Practice*. Prentice Hall Signal Processing Series. Upper Saddle River, NJ: Prentice Hall PTR, 2002. ISBN: 0-13-242942-X.

[21] Theodoros Giannakopoulos e Aggelos Pikrakis. *Introduction to Audio Analysis: A MATLAB Approach*. Oxford, UK: Academic Press, 2014. ISBN: 978-0-12-405865-1.

[22] Shing-Tai Pan e Han-Jui Wu. «Performance Improvement of Speech Emotion Recognition Systems by Combining 1D CNN and LSTM with Data Augmentation». In: *Electronics* 12.11 (2023). ISSN: 2079-9292. DOI: 10.3390/electronics12112436. URL: https://www.mdpi.com/2079-9292/12/11/2436.

[23] Björn Schuller, Stefan Steidl e Anton Batliner. «The INTERSPEECH 2009 Emotion Challenge». In: *Proceedings of the 10th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. ISCA. Brighton, UK, 2009, pp. 312–315.

[24] Tin Lay Nwe, Say Wei Foo e Liyanage C. De Silva. «Speech emotion recognition using Hidden Markov Models». In: *Speech Communication* 41.4 (2003), pp. 603–623. DOI: 10.1016/S0167-6393(03)00099-2.

[25] Ossama Abdel-Hamid et al. «Convolutional Neural Networks for Speech Recognition». In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 22.10 (2014), pp. 1533–1545. DOI: 10.1109/TASLP.2014.2339736.

[26] Alex Graves, Abdel-rahman Mohamed e Geoffrey E. Hinton. «Speech recognition with deep recurrent neural networks». In: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (2013), pp. 6645–6649. URL: https://api.semanticscholar.org/CorpusID:206741496.

[27] Ashish Vaswani et al. «Attention is All you Need». In: *Advances in Neural Information Processing Systems*. A cura di I. Guyon et al. Vol. 30. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf. Curran Associates, Inc., 2017.

[28] Sinno Jialin Pan e Qiang Yang. «A Survey on Transfer Learning». In: *IEEE Transactions on Knowledge and Data Engineering* 22.10 (2010), pp. 1345–1359. DOI: 10.1109/TKDE.2009.191.

[29] Fuzhen Zhuang et al. «A Comprehensive Survey on Transfer Learning». In: *Proceedings of the IEEE* 109.1 (2021), pp. 43–76. DOI: 10.1109/JPROC.2020.3004555.

[30] Taiba Majid Wani et al. «A Comprehensive Review of Speech Emotion Recognition Systems». In: *IEEE Access* 9 (2021), pp. 47795–47814. DOI: 10.1109/ACCESS.2021.3068045.

[31] Ingo Siegert et al. «Investigation of Speaker Group-Dependent Modelling for Recognition of Affective States from Speech». In: *Cognitive Computation* 6.4 (2014), pp. 892–913. DOI: 10.1007/s12559-014-9296-6.

[32] Sudhir Varma e Richard Simon. «Bias in error estimation when using cross-validation for model selection». In: *BMC Bioinformatics* 7.1 (2006), p. 91. DOI: 10.1186/1471-2105-7-91.

[33] Marina Eni et al. «Estimating Autism Severity in Young Children From Speech Signals Using a Deep Neural Network». In: *IEEE Access* 8 (2020), pp. 139489–139500. DOI: 10.1109/ACCESS.2020.3012532.

[34] N. A. Chi et al. «Classifying Autism From Crowdsourced Semistructured Speech Recordings: Machine Learning Model Comparison Study». In: *JMIR Pediatrics and Parenting* 5.2 (2022), e35406. DOI: 10.2196/35406. URL: https://doi.org/10.2196/35406.

[35] Steven R. Livingstone e Frank A. Russo. «The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English». In: *PLoS ONE* 13.5 (2018), e0196391. DOI: 10.1371/journal.pone.0196391.

[36] Felix Burkhardt et al. «A Database of German Emotional Speech». In: *Proceedings of Interspeech 2005*. Lisbon, Portugal, 2005, pp. 1517–1520. DOI: 10.21437/Interspeech.2005-446.

[37] P. Hendriks. *Asymmetries in grammar: NWO/Vici-project*. [Consultato il 26 marzo 2026]. 2013. URL: http://www.let.rug.nl/~hendriks/asymmetries/.

[38] Theodoros Giannakopoulos. «A Method for Silence Removal and Segmentation of Speech Signals, Implemented in MATLAB». Tesi di dott. Athens, Greece: University of Athens, 2009.

[39] Björn Schuller et al. «Cross-corpus acoustic emotion recognition: Variances and benefits». In: *Proceedings of INTERSPEECH 2010*. Makuhari, Japan, 2010, pp. 454–457.

[40] The MathWorks, Inc. *Diagnostic Feature Designer*. MATLAB Documentation. 2024. URL: https://www.mathworks.com/help/stats/diagnostic-feature-designer-app.html (visitato il giorno 30/03/2026).

[41] Tara N. Sainath et al. «Convolutional, Long Short-Term Memory, fully connected Deep Neural Networks». In: *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2015, pp. 4580–4584. DOI: 10.1109/ICASSP.2015.7178838.

[42] Sepp Hochreiter e Jürgen Schmidhuber. «Long Short-Term Memory». In: *Neural Computation* 9.8 (1997), pp. 1735–1780. DOI: 10.1162/neco.1997.9.8.1735. URL: https://doi.org/10.1162/neco.1997.9.8.1735.

[43] Sergey Ioffe e Christian Szegedy. «Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift». In: *Proceedings of the 32nd International Conference on Machine Learning*. A cura di Francis Bach e David Blei. Vol. 37. Proceedings of Machine Learning Research. Lille, France: PMLR, lug. 2015, pp. 448–456. URL: https://proceedings.mlr.press/v37/ioffe15.html.

[44] Zachary C Lipton, John Berkowitz e Charles Elkan. «A critical review of recurrent neural networks for sequence learning». In: *arXiv preprint arXiv:1506.00019* (2015).

[45] Justin Salamon e Juan Pablo Bello. «Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification». In: *IEEE Signal Processing Letters* 24.3 (2017), pp. 279–283. DOI: 10.1109/LSP.2017.2657381.

[46] Klaus R. Scherer. «Vocal communication of emotion: a review of research paradigms». In: *Speech Commun.* 40.1–2 (apr. 2003), pp. 227–

256. ISSN: 0167-6393. DOI: 10.1016/S0167-6393(02)00084-5. URL: https://doi.org/10.1016/S0167-6393(02)00084-5.

[47] Yin Cui et al. «Class-balanced loss based on effective number of samples». In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 9268–9277.

[48] Jasper Snoek, Hugo Larochelle e Ryan P Adams. «Practical bayesian optimization of machine learning algorithms». In: *Advances in neural information processing systems* 25 (2012).

[49] C.E. Rasmussen e C.K.I. Williams. *Gaussian Processes for Machine Learning*. Adaptative computation and machine learning series. University Press Group Limited, 2006. ISBN: 9780262182539. URL: https://books.google.it/books?id=vWtwQgAACAAJ.

[50] Diederik P Kingma e Jimmy Ba. «Adam: A method for stochastic optimization». In: *arXiv preprint arXiv:1412.6980* (2014).

[51] Lutz Prechelt. «Early Stopping-But When?» In: *Neural Networks*. 1996. URL: https://api.semanticscholar.org/CorpusID:14049040.

[52] Alaa Tharwat. «Classification assessment methods». In: *Applied Computing and Informatics* 17.1 (lug. 2020), pp. 168–192. ISSN: 2634-1964. DOI: 10.1016/j.aci.2018.08.003. eprint: https://www.emerald.com/aci/article-pdf/17/1/168/38503/j_aci_2018_08_003.pdf. URL: https://doi.org/10.1016/j.aci.2018.08.003.

[53] J. R. Landis e G. G. Koch. «The measurement of observer agreement for categorical data». In: *Biometrics* 33.1 (1977), pp. 159–174.

[54] *«Audacity ®»*. LOnline]. Available: Aurlhttps://www.audacityteam.org/. 2022.

[55] Audacity Team. *Audacity Manual: Noise Reduction*. Ultima modifica: 04-12-2025. Disponibile online: `https://manual.audacityteam.org/man/noise_reduction.html`. Audacity Team. 2025.

[56] Jort F. Gemmeke et al. «Audio Set: An ontology and human-labeled dataset for audio events». In: *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2017), pp. 776–780. URL: `https://api.semanticscholar.org/CorpusID:21519176`.

[57] Shawn Hershey et al. «CNN architectures for large-scale audio classification». In: *2017 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE. 2017, pp. 131–135.

[58] Haibo He e Edwardo A Garcia. «Learning from imbalanced data». In: *IEEE Transactions on Knowledge and Data Engineering* 21.9 (2009), pp. 1263–1284.

[59] J. Kittler et al. «On combining classifiers». In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20.3 (1998), pp. 226–239.

Ringraziamenti

Desidero rivolgere un sincero ringraziamento al mio relatore, il Prof. Pietro Savazzi, per avermi guidata in questo progetto di tesi. La sua disponibilità e il suo costante incoraggiamento sono stati essenziali per portare a termine questo percorso.
Un ringaziamento va anche ai miei corellatori il Dott. Mauro Marchese e la Prof.ssa Natascia Brondino per i consigli forniti durante le varie fasi del progetto.