**Università Politecnica delle Marche**

Facoltà di Ingegneria

Dipartimento di Ingegneria dell'Informazione

Corso di Laurea Magistrale in Ingegneria Informatica e dell'Automazione

---


**Tesi di Laurea**

# Metodi automatici a supporto della stima dell'intelligibilità nell'ambito della riabilitazione vocale dei pazienti affetti da malattie neurologiche

# Automated methods to support clinicians in intelligibility estimation during the vocal rehabilitation of patients with neurological disorders

Relatore

*Prof.* Stefano Squartini

Correlatore

*Dott.ssa* Lucia Migliorelli

Candidato

Lorenzo Scoppolini Massini

---

Sessione di Laurea Luglio 2021
Anno Accademico 2020/2021

# Indice

# Elenco delle figure

# Sommario

La sclerosi laterale amiotrofica, o SLA, è una malattia neurodegenerativa progressiva del motoneurone, che colpisce selettivamente sia i motoneuroni centrali sia quelli periferici. I primi riferimenti certi della malattia risalgono al 1824 da parte di Charles Bell, tuttavia la patologia fu battezzata SLA solo nel 1874. La SLA rimase per diversi decenni una malattia poco conosciuta finché, nel 1939, colpì il famoso giocatore di baseball Lou Gehrig [1], catturando l'attenzione pubblica.

Il tasso annuale di incidenza nei paesi industrializzati viene stimato pari a 2 casi ogni 100000 abitanti, solo in Italia vivono 3600 pazienti affetti da questa patologia. Ad oggi rimangono ancora molti aspetti da chiarire sui fattori scatenanti, in particolare sono state distinte due forme principali: quella di origine genetica e quella sporadica. Sulla prima, che riguarda il 5-10% dei casi sono state fatte recenti scoperte di rilievo che aiuteranno lo sviluppo di trattamenti efficaci, sulla seconda si sono avanzate diverse ipotesi, ma ancora rimangono molte incertezze. In assenza di cure specifiche, il decorso della SLA va normalmente dai 3 ai 5 anni: durante questo periodo i pazienti vengono sottoposti a sessioni di fisioterapia periodiche che permettono ai medici specialisti di rilevare lo stato di avanzamento della malattia. I dati raccolti in questo modo, oltre a chiarire al paziente la sua condizione clinica, sono cruciali per la ricerca, in particolare per valutare l'efficacia dei trattamenti che vengono continuamente sperimentati. In questo senso sono state proposte diverse scale di valutazione che, attraverso l'esecuzione di esercizi specifici, permettono di "quantificare" al meglio la gravità della malattia.

Purtroppo i protocolli formali definiti in queste scale da soli non sono sufficienti a garantire un'elevata oggettivizzazione del dati raccolti. Infatti, molti degli esercizi previsti sono valutati direttamente dal clinico senza l'ausilio di alcuno strumento. Inevitabilmente la raccolta di questi parametri rimane ancora affetta da errori individuali e difficilmente ripetibile, rendendo difficoltoso il confronto tra pazienti seguiti da strutture diverse.

Oggetto della presente tesi è l'approfondimento di una di queste scale valutative: il Profilo Robertson. In particolare di quest'ultimo vengono considerati gli esercizi vocali

[1] per questa ragione la patologia è anche conosciuta come malattia di Lou Gehrig

che rimangono tra i più complessi da misurare quantitativamente e per i quali sono presentate tre soluzioni software. L'obiettivo di questi sistemi è supportare il clinico durate la valutazione, permettendo la raccolta di misure quantitative e ripetibili. La tesi si struttura in 5 capitoli: nel primo viene trattata in modo più dettagliato la malattia e viene presentato il Profilo Robertson. Nel successivo vengono formalizzati alcuni esercizi della scala in modo da rintracciare in letteratura tecniche ed approcci utili alla loro valutazione oggettiva. Nel terzo capitolo sono presentate due soluzioni originali per affrontare una particolare tipologia di esercizio e nel capitolo seguente sono mostrati e discussi i risultati raggiunti. Infine, nell'ultimo capitolo, vengono tratte le conclusioni sull'efficacia e l'applicabilità dei metodi affrontati.

# Capitolo 1

# Introduzione

Inizialmente, in questo capitolo verrà approfondita la patologia SLA dal punto di vista clinico, poi sarà presentata la scala Robertson che successivamente verrà presa come riferimento per individuare delle criticità tecniche adatte per essere affrontate tramite algoritmi appositamente studiati.

## 1.1 La Sclerosi Laterale Amiotrofica

La SLA è una patologia che inibisce sia i motoneuroni centrali, cioè quelli che collegano il cervello al midollo spinale, sia i motoneuroni periferici, che collegano i neuroni motori superiori dal midollo spinale a tutti muscoli del corpo [13].

Questi neuroni comunicano attraverso l'invio di impulsi elettrici che muovendosi da un neurone all'altro raggiungono, alla fine, il bersaglio desiderato, ossia i muscoli. Nella SLA, questo sistema di comunicazione neuronale si degrada. I motoneuroni non sono in grado di portare le informazioni dal cervello e dal midollo spinale al muscolo, che come risultato diventa quindi inattivo. Se un muscolo è inattivo nel tempo la sua massa inizia a diminuire, ossia si atrofizza. Per questa ragione tra i sintomi più diffusi della SLA c'è il deperimento muscolare. [21]

L'indebolimento generale dei muscoli comporta la disfunzione di molte attività fondamentali come: camminare, impugnare oggetti, parlare, deglutire, respirare.

Le origini della malattia rimangono ad oggi ancora incerte, tuttavia l'ipotesi attuale è che sia una combinazione di fattori genetici ed ambientali. Studi condotti negli anni passati hanno evidenziato che in alcune famiglie la malattia è più ricorrente rispetto alla media e questo ha dato credito all'idea che la patologia avesse alla base una mutazione genetica ereditaria [5]. Studi paralleli hanno mostrato che in alcuni soggetti la SLA dipende da mutazioni genetiche sviluppate nel corso della vita [2]. Per via di questa duplice natura, la patologia viene distinta in due forme: quella genetica o familiare, e quella sporadica o acquisita.

Sfortunatamente, ad oggi, non sono state trovate cure capaci di far guarire dalla malattia. Tuttavia sono stati sviluppati alcuni trattamenti che permettono di rallentarne il decorso. In questo senso l'unico farmaco approvato dalla *Food and Drug Administration* (FDA) per affrontare la patologia è il riluzolo che è in grado di offrire ai motoneuroni una parziale difesa alla degenerazione. Inoltre, i soggetti affetti da SLA sono sottoposti periodicamente a sessioni di fisioterapia che hanno l'obiettivo di monitorare il loro stato di salute e di stimare la risposta ai trattamenti sopra citati. Allo scopo, in queste sessioni vengono impiegate delle scale di valutazione che permettono di guidare e uniformare la prognosi, una delle più diffuse è il Profilo Robertson.

## 1.2 Il Profilo Robertson

La scala di Robertson [3] è un protocollo pensato per valutare lo stato dei pazienti affetti da malattie neurologiche. Ad oggi, in questo ambito, risulta l'unica scala validata scientificamente in Italia. Essa prevede la stima di 71 indici che combinati permettono di determinare il livello di sviluppo di queste patologie. Per la stima di ciascun indicatore si richiede al paziente di svolgere un esercizio che viene valutato quantitativamente o qualitativamente con un punteggio da 1 a 4. Lo svolgimento dell'intera scala richiede circa un'ora, durante la quale vengono analizzati i seguenti aspetti:

- Respirazione
- Fonazione
- Muscolatura Bucco-Facciale
- Diadococinesi
- Riflessi
- Articolazione
- Intelligibilità
- Prosodia

La scala include molti esercizi che richiedono al paziente di pronunciare parole o fonemi: più precisamente questi ricadono nell'ambito della respirazione, fonazione, diadococinesi e intelligibilità. Attraverso di essi, infatti, è possibile stimolare il movimento di molti muscoli facciali oltre che della lingua, corde vocali e polmoni, che con l'avanzare delle patologie tendono ad indebolirsi gradualmente.

Come anticipato, durante l'esecuzione degli esercizi, il clinico spesso non può contare su alcun supporto strumentale, e per quanto concerne gli esercizi che coinvolgono la voce, la valutazione diretta e soggettiva è ancora più frequente. Per quest'ultimi,

eventualmente, il clinico può disporre solamente di software generici, come il Praat, che risultano di difficile utilizzo, e che forniscono solo indicazioni parziali.

Nei capitoli successivi si approfondiscono due tipologie di esercizi della scala per discutere dei possibili approcci capaci di restituire ai medici specializzati dati completi e di qualità, che non necessitano quindi di alcuna interpretazione o integrazione spesso fonti di errore.

# Capitolo 2

# Stato dell'arte

Nel seguente capitolo verranno trattati gli esercizi che concorrono alla valutazione pneufonicoarticolatoria. Un focus particolare sarà riservato a quelli relativi alla diadococinesi e all'intelligibilità che saranno quindi affrontati anche a livello tecnico attraverso alcuni articoli disponibili in letteratura.

## 2.1 I compiti vocali nel Profilo di Robertson

Gli esercizi di respirazione, fonazione, diadococinesi e intelligibilità, insieme, permettono di ottenere una valutazione pneuofonicoarticolatoria. Ad esempio, per la respirazione è chiesto al paziente di emettere una "S" prolungata, in questo modo misurando il tempo di esecuzione e l'intensità sonora è possibile stimare la capacità di ventilazione polmonare. La fonazione comprende tutti gli esercizi che agiscono a livello laringeo, in particolare richiedono di riprodurre vocalizzi per identificare parametri come la frequenza fondamentale e la massima pressione sonora. Alla diadococinesi fanno capo tutte le attività che richiedono azioni rapide e ripetute [1], quindi comprende sia esercizi che richiedono di muovere la muscolatura facciale sia che richiedono di riprodurre continuamente brevi sequenze di sillabe. L'intelligibilità infine viene valutata chiedendo al paziente di leggere alcuni brani appositamente scelti per mettere alla luce eventuali anomalie.

Tra le categorie della scala Robertson che contribuiscono alla valutazione pneuofonicoarticolatoria, nelle seguenti sottosezioni si approfondiscono le due che risultano meno supportate da strumentazione, cioè: l'intelligibilità e la diadococinesi.

[1] https://www.corriere.it/salute/dizionario/diadococinesi

### 2.1.1 Intelligibilità

Come sottolineato da Larry E. Smith et al. in [19], per fissare il concetto di intelligibilità occorre distinguere tre concetti:

- intelligibilità, che dipende dal numero di parole riconoscibili
- comprensibilità, che dipende dal numero di parole di cui si coglie il significato
- interpretabilità, che dipende dalla chiarezza del messaggio inteso dall'oratore

Seguendo questa caratterizzazione, risulta che l'intelligibilità è un parametro fondamentale per valutare lo stadio di avanzamento della SLA, infatti questa malattia tende nel tempo ad indebolire, fino a paralizzare completamente, i muscoli del paziente, e quindi anche quelli adibiti alla parola. Nella scala Robertson questa categoria si compone di esercizi che richiedono la lettura di alcune frasi allo scopo di valutare la capacità del paziente di pronunciare correttamente le parole.

### 2.1.2 Diadococinesi

La diadococinesi serve a valutare la velocità, la precisione, la coordinazione e l'estensione dei movimenti. Nell'ambito della valutazione dei pazienti affetti da SLA, e più precisamente prendendo in considerazione le prove relative all'articolazione del linguaggio, è pratica comune richiedere la pronuncia ripetuta di alcuni fonemi quali: "Pa", "Pa-Ta-Ka" e "Ui". In questi esercizi l'obiettivo e quello di ripetere il maggior numero di volte questi versi cercando di scandire quanto più possibile ogni pronuncia.

Lo scopo di questi esercizi è quello di esaminare analiticamente le eventuali anomalie del linguaggio che caratterizzano il paziente. Infatti quando si valuta l'intelligibilità, dalla lettura del brano possono emergere alcuni difetti di pronuncia, senza però distinguere esattamente le ragioni di ogni criticità. Al contrario, durante lo svolgimento di queste attività è possibile localizzare l'origine che determina ogni problema di pronuncia. In particolare le quattro sillabe "Pa", "Ta", "Ka" e "Ui" per essere pronunciate correttamente richiedono rispettivamente un'azione bilabiale, dentale, velare e vocale. In questo modo è possibile, ad esempio, riconoscere una criticità in corrispondenza delle labbra nel caso in cui si riscontri una sistematica difficoltà durante le ripetizioni di "Pa".

## 2.2 Metodi automatici per la valutazione di intelligibilità e diadococinesi

In questa sezione verranno esposte alcune tecniche, disponibili in letteratura, che hanno affrontato il tema dell'intelligibilità. Per quanto riguarda la diadococinesi applicata alla valutazione dell'articolazione del linguaggio non ci sono riferimenti specifici, quindi verrà trattato un articolo che affronta il problema più generale del *keyword spotting* dal quale verranno presi degli spunti per implementare una soluzione più specifica.

Per trattare l'intelligibilità bisogna distinguere l'accezione intesa lato clinico e lato ingegneristico. Seguendo la prima, quando si valuta l'intelligibilità si tengono in considerazione diversi aspetti tra cui: la frequenza di errori di pronuncia, le pause tra una parola e la successiva, la capacità espressiva e l'intensità sonora. Questi vengono aggregati assieme per produrre un giudizio che nel Profilo di Robertson viene espresso tramite un numero da 1 a 4. Secondo la seconda, come su scritto, si intende il rapporto tra le parole comprensibili su tutte le parole dette, ovvero il *Word Recognition Rate*. Nonostante questa seconda definizione sia più puntuale, rimane aperta la questione di stabilire quando una parola si possa considerare effettivamente riconoscibile dall'ascoltatore. Per risolvere questo secondo problema i principali *dataset* specializzati, [7] dopo aver catturato delle registrazioni dai pazienti, arruolano un gruppo di uditori. Ogni audio contiene una sola parola, gli uditori, autonomamente, possono ascoltare ogni traccia quante volte preferiscono prima di annotare la parola riconosciuta. Alla fine, facendo una media fra tutte le annotazione giuste e sbagliate, si ricava la percentuale di intelligibilità del soggetto.

Le soluzioni proposte in letteratura per la stima automatica dell'intelligibilità possono essere suddivise in due categorie principali:

1. basate su sistemi di *Automatic Speech Recognition* (ASR) [4]
2. basati sulla raccolta di *feature* [17]

I sistemi basati su ASR fanno affidamento su algoritmi di Intelligenza Artificiale progettati per riconoscere le parole registrate a partire da un file audio. Alla stregua di un uditore, questi algoritmi possono ricavare una stima dell'intelligibilità comparando le parole riconosciute con quelle effettivamente dette nella traccia. Il problema di questo approccio risiede nella sua complessità: infatti, i migliori algoritmi di ASR attualmente disponibili sono molto complessi e richiedono una elevata capacità computazionale, oltre che un *dataset* molto vasto per la fase di allenamento.

Il secondo approccio rappresenta ogni espressione attraverso un vettore definito in uno spazio multidimensionale detto *acustic space*. Il vettore viene quindi inviato ad

un predittore, come ad esempio un *support vector predictor*, che converte il vettore in un indice di intelligibilità.

Quest'ultima tipologia di algoritmi, per funzionare, può richiedere un segnale di riferimento che viene confrontato con il segnale da valutare, oppure può regredire direttamente l'intelligibilità [8]. In quest'ultimo caso si garantisce una maggiore applicabilità del sistema in contesti reali, tuttavia, molte di queste soluzioni, non seguono rigorose strategie di validazione.

### 2.2.1 La stima dell'intelligibilità

L'approccio approfondito in questa tesi è quello proposto da Janbakhshi et al. [14]. La soluzione in questione è basata sull'estrazione di *feature* senza far uso di alcun sistema ASR, ne di un predittore. L'idea di base consiste nello stimare l'intelligibilità di un segnale di test basandosi su un segnale di riferimento, ed in particolare valutandone la differenza.

Per quantificare la distanza tra i due segnali e tradurla in una misura dell'intelligibilità vengono impiegati due indici che hanno già dato ottimi risultati in contesti simili: lo *short-time objective intelligibility* (STOI) e lo *extended short-time objective intelligibility* (ESTOI) [App. A]. Questi due indici sono usati per stimare l'intelligibilità comparando un segnale di test rumoroso con un segnale di riferimento nitido allineati temporalmente. Per estendere l'uso di questi due indici nel contesto in esame, è possibile considerare il segnale del paziente come quello rumoroso, mentre il segnale del soggetto sano come quello nitido. Inoltre, poiché i due indici devono essere calcolati a partire da due segnali sincronizzati e di pari lunghezza, bisognerà introdurre una fase iniziale di allineamento attraverso il *Dynamic Time Warping* [App. B]. Per via di queste variazioni, gli autori dell'Articolo [14] hanno deciso di rinominare rispettivamente i due indici come: P-STOI e P-ESTOI.

In Figura 2.1 è possibile osservare l'intero flusso che segue l'algoritmo per estrarre il P-STOI e il P-ESTOI. Come si può notare, il segnale di riferimento non appare come un input, infatti, a differenza del contesto applicativo originale, in questo caso non si ha nessun segnale nitido che corrisponde a quello del paziente. Quindi il segnale di riferimento viene ottenuto a partire da un gruppo di registrazioni raccolte da soggetti sani. Queste registrazioni, nelle quali viene ripetuta la stessa frase o parola richiesta poi al paziente, vengono fuse per mezzo di una media aritmetica dopo essere prima allineate tra loro attraverso il DTW.

A tutti i segnali coinvolti viene applicata una fase iniziale di preprocessamento nella quale vengono portati nel dominio della frequenza attraverso una *Fast Fourier Transform* (FFT). Inoltre vengono esclusi i *frame* a basso contenuto energetico usando

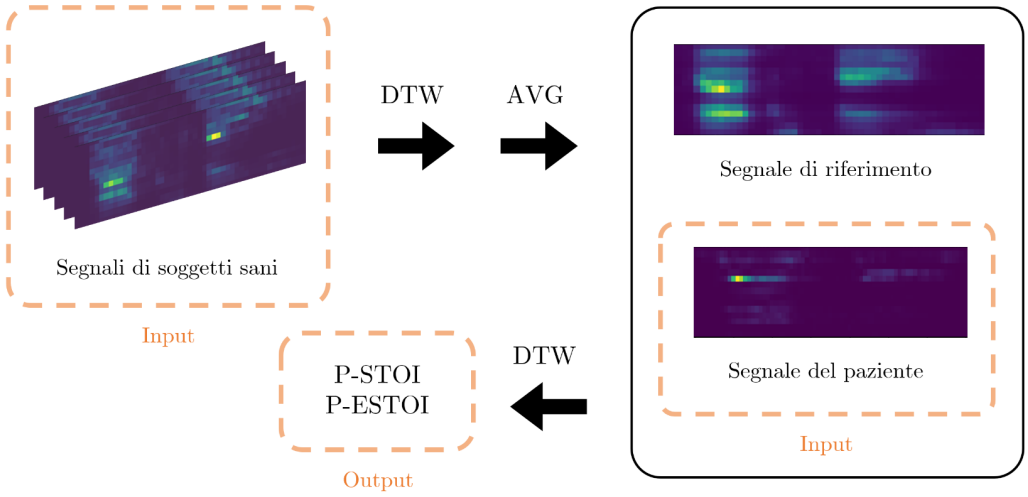



Figura 2.1: Flusso di esecuzione per la stima dell'intelligibilità

un sistema di *Voice Activity Detection* (VAD). Infine viene ridotta la dimensionalità del segnale aggregando l'energia sulle singole frequenze in 15 bande, riportate nella Tabella 2.1 attraverso una *one-third octave band analysis*.

Una volta processati tutti i segnali in input, come su scritto, viene prodotto il segnale di riferimento. Allo scopo, viene selezionato un segnale tra quelli registrati ai soggetti sani, al quale, tramite DTW vengono allineati i rimanenti [2]. Al termine dell'allineamento tutti i segnali raggiungono la stessa lunghezza e sarà quindi possibile unirli attraverso una media aritmetica. Il segnale di riferimento così ottenuto, a livello astratto, rappresenta la parola o la frase registrata, con le caratteristiche generali che distinguono la sua completa intelligibilità. Infatti, raccogliendo le registrazioni da un campione sufficientemente eterogeneo, grazie alla media, è possibile attenuare tutte le caratteristiche peculiari del linguaggio di ogni soggetto, mantenendo solo gli aspetti in comune che caratterizzano la pronuncia corretta dell'espressione. La cardinalità del gruppo di soggetti sani utile affinché questa fase si svolga in maniera ottimale non deve essere necessariamente molto elevata. Ad esempio, usando l'*Universal Access database* [7] si hanno a disposizione gli audio di 15 soggetti sani: 4 femmine e 11 maschi. Una volta generato il segnale di riferimento, sarà possibile procedere al calcolo del P-STOI e del P-ESTOI seguendo la procedura descritta nell'Appendice A, dopo aver allineato il segnale del paziente con quello di riferimento.

[2]Nelle prove effettuate dagli autori e nell'ambito della tesi, la scelta del segnale tra tutti quelli disponibili, non ha mostrato un impatto significativo sul risultato finale

| **Bande di aggregazione della 1/3 octave band analysis** | | | |
|---|---|---|---|
| | Limite inferiore | Limite superiore | Ampiezza |
| 01 | 133.63 | 168.37 | 34.73 |
| 02 | 168.37 | 212.13 | 43.76 |
| 03 | 212.13 | 267.27 | 55.14 |
| 04 | 267.27 | 336.74 | 69.47 |
| 05 | 336.74 | 424.26 | 87.53 |
| 06 | 424.26 | 534.54 | 110.28 |
| 07 | 534.54 | 673.48 | 138.94 |
| 08 | 673.48 | 848.53 | 175.05 |
| 09 | 848.53 | 1069.08 | 220.55 |
| 10 | 1069.08 | 1346.95 | 277.88 |
| 11 | 1346.95 | 1697.06 | 350.10 |
| 12 | 1697.06 | 2138.16 | 441.10 |
| 13 | 2138.16 | 2693.91 | 555.75 |
| 14 | 2693.91 | 3394.11 | 700.20 |
| 15 | 3394.11 | 4276.31 | 882.20 |

Tabella 2.1: Bande di aggregazione dell'energia in frequenza

### 2.2.2 Il keyword spotting

Come anticipato, in letteratura non è stato trovato alcun riferimento specifico per il conteggio di parole, o più in generale fonemi, utile nell'ambito della diadococinesi. Per questo motivo è stata approfondita la metodologia del *keyword spotting*, ovvero una tecnica che permette di individuare delle parole chiave all'interno di una traccia audio. Questo tipo di problema ha assunto particolare importanza in questi ultimi anni grazie alla diffusione degli assistenti vocali. Infatti, tali soluzioni sono utilizzate per attivare gli assistenti virtuali quando viene rilevata una certa parola specifica come: "Hey Siri", oppure "Ok Google" [11]. Questa tipologia di algoritmi, rispetto ai più complessi ASR, non è capace di distinguere tutte le parole del dizionario, ma solo un numero molto ristretto, d'altra parte però è più efficiente e non richiede imponenti *dataset* di allenamento.

Questo elaborato di tesi prende spunto dall'Articolo di Zhang et al. [22]. L'approccio spiegato nell'articolo viene chiamato *Deep Tempalte Matching* (DTM). Il DTM sfrutta un modello *end-to-end* basato sul *deep learning*. L'architettura, mostrata in Figura 2.2, si sviluppa in tre componenti principali:

1. l'estrattore di *feature*
2. il meccanismo di *template matching*
3. il classificatore binario

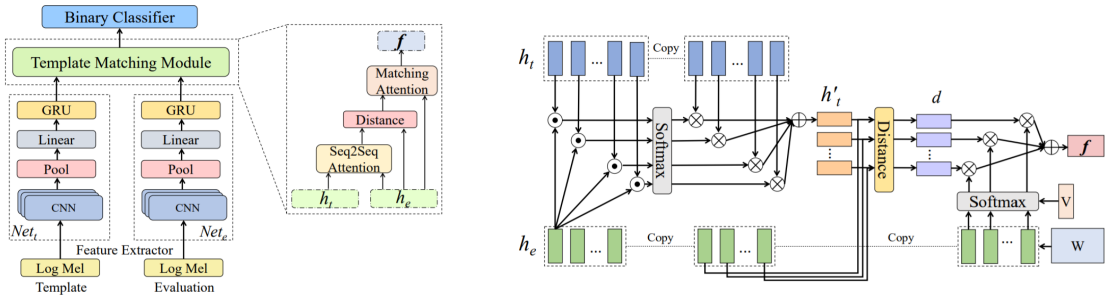



Figura 2.2: Immagini tratte dall'Articolo [22]; (a) rappresenta l'architettura generale del sistema di *keyword spotting*; (b) rappresenta più dettagliatamente la parte del sistema adibita al *template matching*

Il sistema funziona confrontando tra loro una traccia audio di riferimento, detta *template*, con quella da valutare, chiamata *evaluation*. Entrambi i segnali vengono processati sotto forma di *Log Mel Spectrogram* e passati all'estrattore di *feature*. Quest'ultimo è implementato attraverso una rete Siamese [1], ovvero una rete che viene eseguita su due input distinti ai quali vengono fatti attraversare gli stessi *layer*. In questo modo, la rete restituisce due uscite che rappresentano i *feature vector* relativi ai due ingressi. In questo caso la rete Siamese si compone di una *Convolutional Recurrent Neural Network* (CRNN) che include tre *layer* convoluzionali e un *Gated Recurrent Unit* (GRU).

I due *feature vector* ottenuti, denotati $h_t$ e $h_e$ e ricavati rispettivamente dal segnale *template* e dal segnale *evaluation*, vengono introdotti nel sistema di *template matching*. Questo modulo allinea i due segnali sfruttando una *Seq2Seq* e successivamente estrae un vettore $d$ che contiene la distanza *frame* per *frame*. Il vettore $d$ viene quindi trasformato in un vettore $f$ a dimensione fissa attraverso un meccanismo chiamato *matching attention*.

Una volta ottenuto il vettore $f$, attraverso il classificatore binario costituito da un solo *layer fully-connected*, vengono riconosciuti tutti i segnali di *evaluation* che corrispondono effettivamente alla *keyword* scelta.

# Capitolo 3

# Protocollo sperimentale e metodi

In questo capitolo, dopo una breve esposizione dei *dataset* utilizzati, verranno mostrate le soluzioni sviluppate nell'ambito della tesi. Per quanto riguarda l'intelligibilità non sarà presentato alcun approccio, in quanto si è seguito fedelmente il sistema riportato nella Sottosezione 2.2.1, mentre per la diadococinesi applicata all'articolazione del linguaggio verranno esposti due sistemi originali distinti.

## 3.1 Protocollo sperimentale

In questa sezione verranno descritti i *dataset* sui quali sono stati messi alla prova gli algoritmi sviluppati.

### 3.1.1 Dataset per la stima dell'intelligibilità

Per valutare le prestazioni dell'algoritmo volto alla stima dell'intelligibilità è stato utilizzato il *dataset Universal Access database*. Questa raccolta contiene 765 parole isolate per ciascuno dei 15 soggetti affetti da disartria coinvolti: 11 maschi e 4 femmine. Le parole comprendono 100 termini comuni, 300 termini non comuni, numeri, lettere dell' *International Radio Alphabbet* e termini di natura informatica. L'intelligibilità dei soggetti coinvolti è molto variabile, in alcuni casi è pari al 2% in altri supera il 90%. Inoltre, nel *dataset* sono anche contenute le stesse registrazioni ottenute a partire da un gruppo di 15 persone sane con un intelligibilità di riferimento pari al 100%.

Nelle prove svolte, seguendo quanto fatto nell'Articolo [12], sono stati considerati 10 soggetti affetti da disartria e 13 di controllo.

### 3.1.2 Dataset per la valutazione della diadococinesi

Per provare, e nel caso di uno dei due approcci allenare, i due algoritmi sviluppati per contare le ripetizioni nell'ambito degli esercizi di diadococinesi, è stato appositamente raccolto un *dataset*. Questo comprende le registrazioni di 12 soggetti sani per ognuno dei quali sono stati ottenuti 6 audio: uno di prova e uno di controllo per i tre esercizi "Pa", "Pa-Ta-Ka" e "Ui". Compatibilmente con la situazione pandemica sono stati coinvolti due pazienti affetti da SLA bulbare attualmente in cura presso gli Ospedali Riuniti di Ancona. Ogni traccia audio contiene una sequenza di ripetizioni di un certo esercizio per una durata di circa 30 secondi. Le registrazioni dei soggetti sani sono state raccolte seguendo un protocollo rigido che prevedeva l'impiego di una scheda audio Focusrite Scarlett 2i2 ed un microfono a canale singolo usati in un ambiente silenzioso. Per quanto riguarda i due pazienti, è stato utilizzato il microfono di un auricolare collegato ad uno smartphone in un contesto a basso rumore di sottofondo.

Inoltre per verificare la robustezza degli algoritmi sono state registrate tre ripetizioni di "Ui" di un soggetto sano in contesti critici. La prima registrazione è stata fatta in un ambiente silenzioso attraverso un microfono di un auricolare collegato ad un computer portatile. Per la seconda è stato utilizzato lo stesso apparato di registrazione, ma con una televisione accesa in sottofondo. L'ultima prova si è svolta in maniera analoga alla seconda, eccetto per l'aggiunta di brevi rumori intensi emessi in istanti casuali.

## 3.2 Metodologie per la valutazione della diadococinesi

In questa sezione verranno mostrati due metodi che affrontano il problema del conteggio di fonemi nell'ambito della diadoconinesi. I due sistemi prendono spunto dai due articoli discussi nel capitolo precedente, in particolare con il primo hanno in comune la fase di preprocessamento, mentre con il secondo uno dei due approcci condivide l'impiego di una rete Siamese. Per quanto riguarda l'intelligibilità, vista la ricchezza di soluzioni trovate in letteratura, e le ottime prestazioni raggiunte da alcune di queste, non verranno introdotti nuovi approcci.

### 3.2.1 Metodo basato su Dynamic Time Warping

La prima soluzione sviluppata, rappresentata in Figura 3.4, si affida unicamente al DTW. In particolare, gli input dell'algoritmo sono due: da una parte il segnale raccolto durante l'esecuzione dell'esercizio al paziente $p$, dall'altra un insieme di ripetizioni sin-

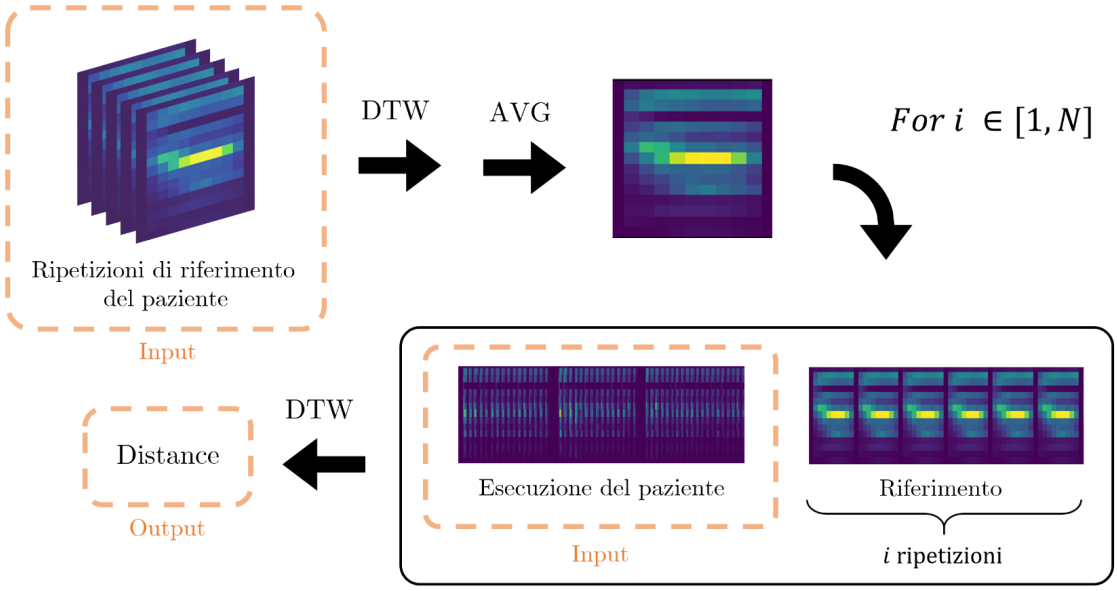



Figura 3.1: Flusso di esecuzione per il conteggio di versi

gole appartenenti allo stesso soggetto [1]. Tutti i segnali coinvolti vengono preprocessati come descritto nella Sottosezione 2.2.1 ottenendo così, per ogni *frame*, la distribuzione energetica del segnale divisa in 15 bande.

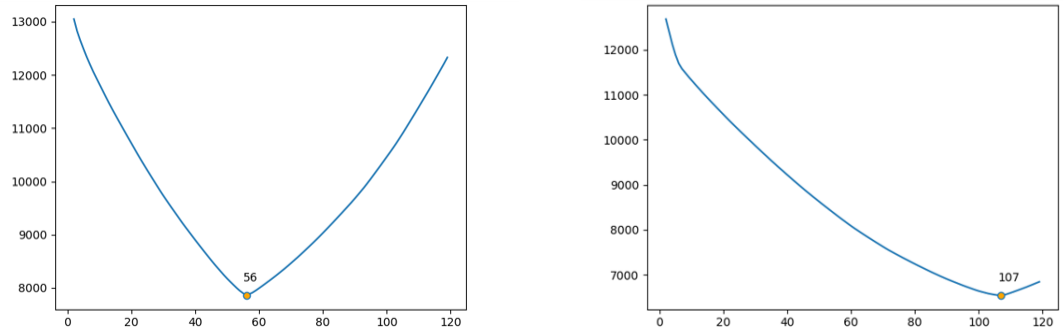
Seguendo la stessa procedura introdotta per l'intelligibilità, viene prodotto un segnale di riferimento iniziale $r_0$ a partire dalle singole ripetizioni. Una volta generato $r_0$ inizia una fase iterativa durante la quale, ad ogni ciclo, vengono svolti i seguenti passaggi:

1. viene prodotto un nuovo segnale di riferimento $r_i$
2. vengono allineati tra loro il segnale di riferimento $r_i$ e quello registrato durante l'esecuzione dell'esercizio
3. viene calcolata la *DTW distance*

Il nuovo segnale di riferimento $r_i$, prodotto ed utilizzato ad ogni iterazione, viene creato concatenando $r_{i-1}$ con $r_0$. In questo modo, si creano dei segnali periodici, che contengono un numero di iterazioni che aumenta sempre di un'unità. Successivamente si allineano tra loro $r_i$ e $p$ e viene estratta la *DTW distance* che viene salvata in memoria.

In Figura 3.2 sono riportate, a titolo di esempio, due sequenze di distanze ottenute al termine di due istanze dell'algoritmo. Come si può osservare, in entrambi i casi la distanza mostra un andamento decrescente fino a raggiungere un punto di minimo,

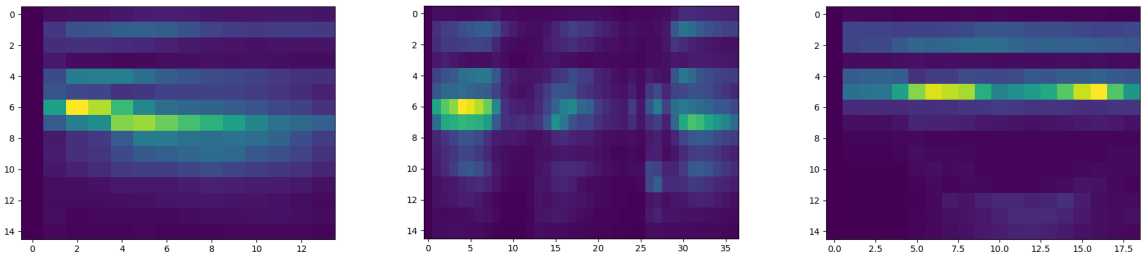
[1] Vengono raccolte ed utilizzate tra le 5 e le 10 ripetizioni per soggetto




(a) (b)

Figura 3.2: (a) ripetizioni "Pa-Ta-Ka" soggetto 4, corrette: 56, stimate: 56; (b) ripetizioni "Ui" soggetto 4, corrette: 108, stimate: 107

passato il quale inizia ad aumentare. I punti di minimo, messi in evidenza, permettono di stimare il numero di ripetizioni effettivamente pronunciate durante lo svolgimento dei due esercizi. Questa scelta è motivata dal fatto che generalmente la coppia di segnali allineati che raggiunge la minore distanza è quella in cui il numero di ripetizioni corrisponde. Infatti quando ciò accade, il DTW può far coincidere esattamente ogni ripetizione di $r_i$ ad una su $p$.



(a) (b) (c)

Figura 3.3: (a) segnale di riferimento $r_0$ della ripetizione "Pa"; (b) segnale di riferimento $r_0$ della ripetizione "Pa-Ta-Ka"; (c) segnale di riferimento $r_0$ della ripetizione "Ui"

In Figura 3.3 sono mostrati i segnali di riferimento $r_0$ dei tre versi: "Pa", "Pa-Ta-Ka" e "Ui" che sono stati presi in esame. Come si può notare, mentre i primi due hanno una distribuzione energetica per *frame* piuttosto uniforme, il segnale relativo a "Ui" mostra una forma ad arco più caratteristica nelle bande superiori. Questa componente del segnale, essendo la più distintiva, è la più utile ai fini del conteggio. Tuttavia, il contenuto energetico in questa fascia risulta più ridotto di quello presente tra la prima e la sesta banda, e questo impedisce al DTW di dargli la giusta rilevanza. Per migliorare le prestazioni del sistema si è dunque deciso di sviluppare due varianti dell'algoritmo: la prima $V_1$ ottimizzata per il conteggio di "Pa" e "Pa-Ta-Ka", mentre la seconda $V_2$ specializzata per "Ui".

La prima differenza tra le due versioni è il *range* di bande effettivamente considerato: mentre $V_1$ considera tutte le 15 bande, $V_2$ utilizza solo quelle comprese tra la settima e la quindicesima, cioè dove ricade l'arco di frequenze ad alta energia. La seconda differenza è sempre relativa alla fase di preprocessamento, più precisamente il VAD, infatti nel caso di $V_2$ questo è più selettivo. Il motivo di questa scelta è sempre dovuto alle caratteristiche dei segnali, infatti il principale aspetto distintivo di "Pa" e "Pa-Ta-Ka" è l'alternanza di *frame* ad alto e basso contenuto energetico. In questo caso, quindi, rimuovendo con il VAD tutti i *frame* di silenzio, si ottiene una sequenza continua ad alto livello energetico che non permette di distinguere le singole ripetizioni. L'ultima differenza tra le due versioni è la funzione di dissimilarità locale $f$ impiegata nel DTW ed introdotta nell'Appendice B. Questa funzione permette l'allineamento dei singoli *frame* e quindi ha un impatto diretto anche nella *DTW distance*. Nello specifico sono state utilizzate queste due funzioni:

$$f_1 = ||x - y||_2 \cdot \left(1 - \frac{(x - \overline{x}) \cdot (y - \overline{y})}{||(x - \overline{x})||_2 ||(y - \overline{y})||_2}\right) \tag{3.1}$$

$$f_2 = ||x - y||_2 \cdot \left(1 + \frac{x \cdot y}{||x||_2 ||y||_2}\right) \tag{3.2}$$

L'Equazione 3.1 che moltiplica tra loro l'*euclidean distance* con la *cosine distance*, è stata usata in $V_1$, mentre l'Equazione 3.1 che moltiplica l'*euclidean distance* con la *correlation distance* è stata impiegata in $V_2$. Il motivo della definizione di queste due distanze, rispetto alla più semplice e diffusa *euclidean distance*, è che dipendendo dalla *cosine distance* piuttosto che dalla *correlation distance*, le due misure risultano meno sensibili a differenze di fattori di scala. In questo modo per il DTW è più facile allineare tra loro due ripetizioni anche se emesse con intensità diverse.

### 3.2.2 Metodo basato su rete Siamese

La seconda soluzione sviluppata come fatto in [22] introduce l'Intelligenza Artificiale tramite l'impiego di una rete Siamese. Anche in questo caso la fase di preprocessamento è analoga a quella descritta nella Sottosezione 2.2.1, tuttavia si è fatta una diversa aggregazione dell'energia in frequenza. Le bande usate in questo approccio sono state ristrette ed aumentate, in modo da ottenere una maggiore definizione. Infatti, rispetto alla *one-third octave band analysis* dove il centro di ogni banda $b$ è calcolato come:

$$b_{center}(k) = 2^{\frac{k}{3}} \cdot f_{min} \tag{3.3}$$

la formula utilizzata in questo caso è:

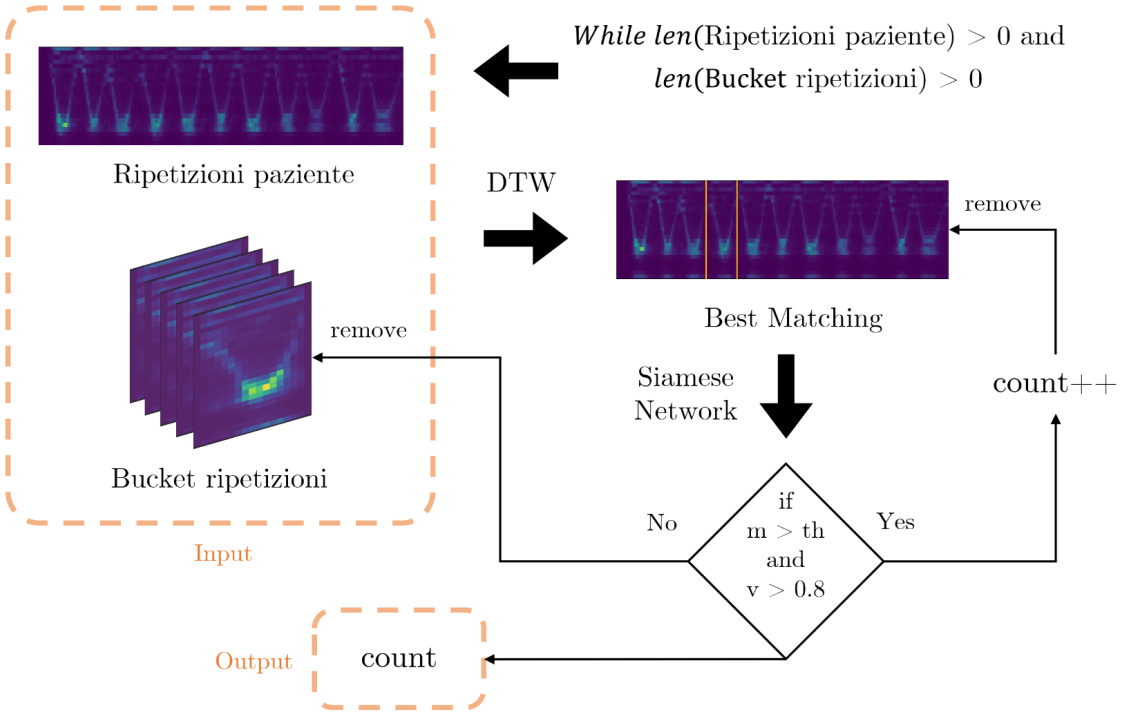



Figura 3.4: Flusso di esecuzione per il conteggio di versi con approccio basato su *deep learning*

$$b_{center}(k) = 2^{\frac{k}{9}} \cdot f_{min} \tag{3.4}$$

Per non ridurre eccessivamente il *range* di frequenze coperto, si sono definite 26 bande.

Il flusso di esecuzione dell'algoritmo è stato rappresentato in Figura 3.4. Anche in questo caso l'algoritmo è iterativo, tuttavia il numero di cicli non è fissato, ma dipende da una condizione di interruzione. In particolare ad ogni iterazione vengono svolte le seguenti operazioni:

1. Viene controllata la condizione di interruzione
2. Ogni segnale $b_i$ presente nel *bucket* ripetizioni, viene allineato attraverso il DTW *open-ended* con il segnale relativo alle ripetizioni del paziente $p$
3. Ogni coppia di segnali allineati nella fase precedente se ha un contenuto energetico medio maggiore di una certa soglia $th$ viene valutata dalla rete Siamese: se l'output della rete è maggiore di 0.8, allora la porzione di segnale individuata all'interno di $p$ viene rimossa e viene contata una nuova ripetizione, altrimenti viene rimosso il segnale $b_i$

La condizione di interruzione controlla due aspetti: da una parte che il *bucket* ripetizioni non sia vuoto, dall'altra che la lunghezza in *frame* del segnale $p$ non sia

scesa sotto una certa soglia. In caso negativo, l'algoritmo restituisce il numero di ripetizioni fino a quel momento individuate e termina.

In fase di inizializzazione, all'interno del *bucket* ripetizioni vengono conservati dei segnali $b_i$ ciascuno dei quali contiene un singola ripetizione. A differenza dell'algoritmo descritto nella sottosezione precedente, in questo caso non è necessario che le ripetizioni appartengano allo stesso soggetto di cui viene valutata l'esecuzione dell'esercizio. Nella seconda fase, ogni segnale $b_i$ rimasto viene allineato al segnale $p$ attraverso il DTW *open-ended*. Questo algoritmo, diversamente al DTW tradizionale, rimuove la condizione per la quale i due segnali in ingresso devono raggiungere la stessa lunghezza finale. In questo modo ogni $b_i$, contenendo una sola ripetizione, può adattarsi ed allinearsi ad una singola ripetizione rimasta in $p$.

**Composizione dei layer della rete neurale**

| Layer | Input size | Hyperparameters | Output size |
|---|---|---|---|
| conv2D 1 | $1 \times T \times 26$ | $5 \times 5, (2,2), 16$ | $16 \times T \times 26$ |
| conv2D 2 | $16 \times T \times 26$ | $5 \times 5, (2,2), 32$ | $32 \times T \times 26$ |
| conv2D 3 | $32 \times T \times 26$ | $5 \times 5, (2,2), 64$ | $64 \times T \times 26$ |
| conv2D 4 | $64 \times T \times 26$ | $5 \times 5, (2,2), 32$ | $32 \times T \times 26$ |
| conv2D 5 | $32 \times T \times 26$ | $5 \times 5, (2,2), 16$ | $16 \times T \times 26$ |
| conv2D 6 | $16 \times T \times 26$ | $5 \times 5, (2,2), 1$ | $1 \times T \times 26$ |
| $\max_p ool$ | $1 \times T \times 26$ | $2 \times 2, (2,1)$ | $1 \times T \times 26$ |
| cosine distance | $1 \times T \times 26$ | // | $1 \times 26$ |
| fc 1 | $1 \times 26$ | $26 \times 52 \times 2$ | $1 \times 52$ |
| fc 2 | $1 \times 52$ | $52 \times 26 \times 2$ | $1 \times 26$ |
| fc 3 | $1 \times 26$ | $26 \times 1 \times 2$ | $1 \times 1$ |

Tabella 3.1: Architettura della rete neurale

Una volta che, attraverso il DTW *open-ended*, si è trovata una possibile ripetizione all'interno di $p$, la parte di $p$ individuata insieme al segnale $b_i$ coinvolto nell'allineamento vengono passati in ingresso alla rete Siamese dopo essere normalizzati individualmente. La rete Siamese, mostrata in Figura 3.5 e riportata in Tabella 3.1, riceve quindi coppie di ingressi di dimensione $T \times 26$, dove $T$ si riferisce al numero di *frame* che a seguito dell'allineamento può variare, mentre 26 sono le bande dei segnali dopo il preprocessamento. Dovendo gestire ingressi di dimensione variabile, la rete è stata realizzata esclusivamente da *layer* convoluzionali. Al termine della rete Siamese, viene calcolata la *cosine distance* tra le due mappe di attivazione restituite in uscita dall'ultimo strato convoluzionale per produrre un vettore di dimensione fissa pari a 26. Il vettore così ottenuto viene passato in ingresso al classificatore binario composto da tre *layer fully-connected* che permette di stabilire se la porzione di $p$ individuata nella fase precedente contiene effettivamente una ripetizione. In base all'output della

rete neurale, come su scritto, si rimuove il segnale $b_i$ dal *bucket* ripetizioni, oppure si sottrae al segnale $p$ la ripetizione individuata.

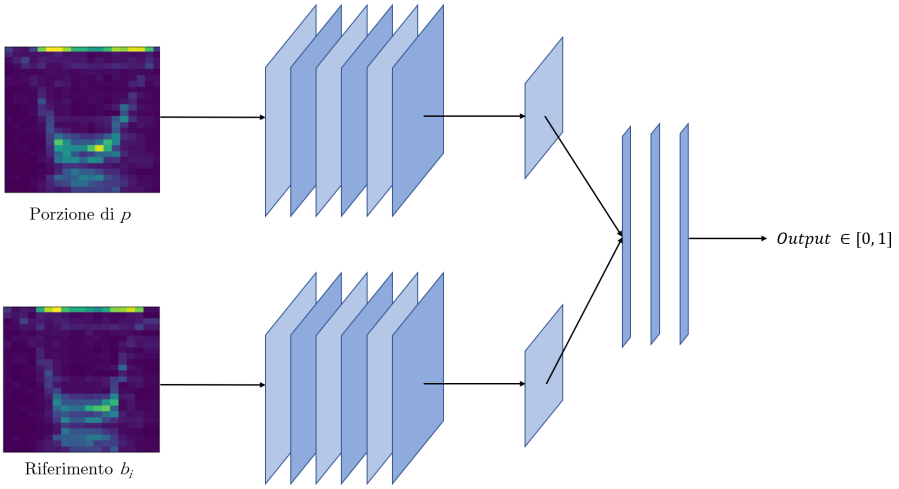
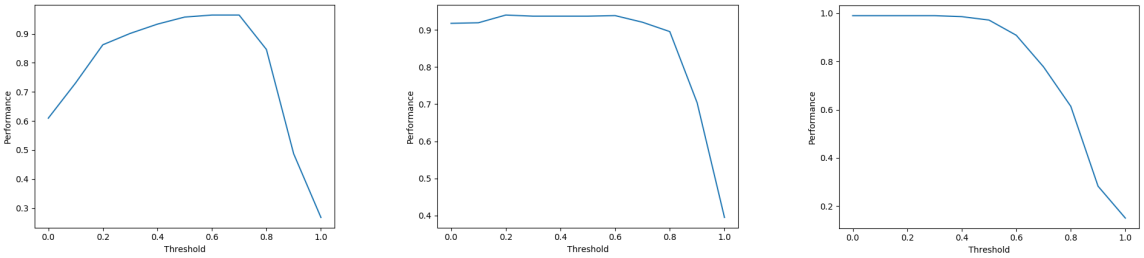



Figura 3.5: Schema della rete neurale




(a) (b) (c)

Figura 3.6: (a) andamento delle prestazioni in funzione della soglia sull'energia media per l'esercizio "Pa"; (b) andamento delle prestazioni in funzione della soglia sull'energia media per l'esercizio "Pa-Ta-Ka"; (c) andamento delle prestazioni in funzione della soglia sull'energia media per l'esercizio "Ui"

Al fine di migliorare le prestazioni, anche per questo approccio sono state fatte delle ottimizzazioni basate sulle diverse caratteristiche dei tre segnali "Pa", "Pa-Ta-Ka" e "Ui". Più in particolare, in fase di preprocessamento cambia il *range* di frequenze considerate: per gli esercizi "Pa" e "Pa-Ta-Ka" la frequenza inferiore $f_{min}$ della prima banda è 133.63 Hz, mentre per "Ui" è 989.83 Hz. Inoltre, il valore della soglia $th$ utilizzata per identificare ed escludere le corrispondenze trovate dal DTW *open-ended* con *frame* a ridotto contenuto energetico varia nei i tre esercizi. Più precisamente, come si può notare nei grafici in Figura 3.6 dove viene mostrata la relazioni tra le prestazioni in funzione di $th$, la soglia fissata ha un impatto differente in base alla tipologia di ripetizione. Per l'esercizio "Pa" in particolare si osserva un netto incremento di prestazioni grazie all'adozione di $th$ pari a 0.6: attraverso questa soglia le prestazioni

aumentano dal 61% al 96%. Per gli altri due tipi di ripetizione, invece, la differenza è meno marcata: per l'esercizio "Pa-Ta-Ka" si è fissata una soglia pari a 0.2 con un miglioramento del 4%, mentre per l'esercizio "Ui" non si ha nessun beneficio, quindi la *th* è fissata pari a 0. La ragione di questa differenza è dovuta alle caratteristiche del segnale "Pa", infatti questo, essendo il più semplice, è anche il più simile al rumore specialmente dopo la fase di normalizzazione. Quindi escludendo le corrispondenze trovate in $p$ a basso contenuto energetico si evitano molti falsi positivi. Infine, come fatto per l'approccio precedente, sono state utilizzate soglie diverse per il sistema VAD, anche in questo caso, per l'esercizio "Ui" si è applicato filtraggio più selettivo.

Riguardo la fase di allenamento del modello sono stati utilizzati i dati descritti nella Sottosezione 3.1.2. A partire da questi, per ogni soggetto è stata presa la registrazione di controllo dalla quale sono state isolate manualmente le singole ripetizioni al fine di creare tre *dataset* distinti: uno per tipologia di esercizio. Ogni *dataset* si compone di circa 1200 elementi equamente distribuiti tra ripetizioni isolate e rumore. Sfruttando il fatto che la rete Siamese lavora su due input alla volta si è comunque potuto fornite al modello un numero sufficiente di combinazioni di coppie diverse per ottenere ottimi risultati senza andare incontro ad *overfitting*. Durante l'allenamento come *loss function* è stata utilizzata la *binary cross entropy loss* definita come:

$$loss = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot log(p(y_i)) + (1 - y_i) \cdot log(1 - p(y_i)) \tag{3.5}$$

dove $y$ rappresenta la classe corretta relativa ad una certa coppia: $y = 1$ se i due segnali in ingresso contengono entrambi una ripetizione, mentre $y = 0$ altrimenti, $p(y_i)$ rappresenta l'output della rete, ed $N$ indica la dimensione del *batch*. Gli allenamenti sono stati svolti in 70 epoche ognuna composta da 3000 *batch* di dimensione pari ad 8. Infine, come ottimizzatore è stato scelto lo *stochastic gradient descent* (SGD) con *momentum* pari a 0.5, mentre il *learning rate* è stato definito con la seguente formula:

$$lr = 0.99^k \cdot 10^{-1} \tag{3.6}$$

dove $k$ indica l'epoca corrente.

# Capitolo 4

# Risultati e discussione

In questo capitolo verranno presentati i risultati ottenuti dai tre algoritmi implementati: quello sviluppato seguendo l'Articolo [12] e i due originali progettati per contare le ripetizioni durante gli esercizi di diadococinesi volti alla valutazione dell'articolazione del linguaggio.

## 4.1 Intelligibilità

Nei test effettuati, i due indici P-STOI e P-ESTOI hanno mostrato prestazioni in linea con quanto affermato nell'articolo di riferimento [14]. In particolare per i due indici viene riportata una correlazione di Pearson $R$ pari a $0.90 \pm 0.004$ per il P-STOI e pari a $0.95 \pm 0.004$ per il P-ESTOI, mentre quella raggiunta nelle prove è rappresentata nel grafico in Figura 4.1.

In particolare, in questo diagramma sono riportati gli andamenti della correlazione tra i due indici rispetto all'intelligibilità in funzione del numero di tracce audio analizzate. Infatti nel *dataset Universal Access database*, ogni registrazione include una sola parola, quindi per aumentare la correlazione è necessario calcolare gli indici su più tracce, e successivamente ricavare la media aritmetica. In questo modo, come si può osservare, per entrambe le misure aumenta la correlazione fino a raggiungere valori simili a quelli sopra specificati. Più precisamente si può notare come siano sufficienti circa 20 parole affinché la correlazione raggiunga il massimo e si stabilizzi.

Nei due grafici riportati in Figura 4.2 si può osservare la relazione tra l'intelligibilità e i due indici ottenuta a partire da dieci pazienti affetti da disartria. Com'è possibile notare, i dieci soggetti sono affetti da forme della malattia più o meno grave, alcuni infatti mostrano un'intelligibilità elevata, superiore al 90%, mentre altri riportano un'intelligibilità ridotta, inferiore al 10%. I dati esposti sono stati ricavati mediando i risultati ottenuti in 30 prove svolte impiegando di volta in volta 20 parole diverse.

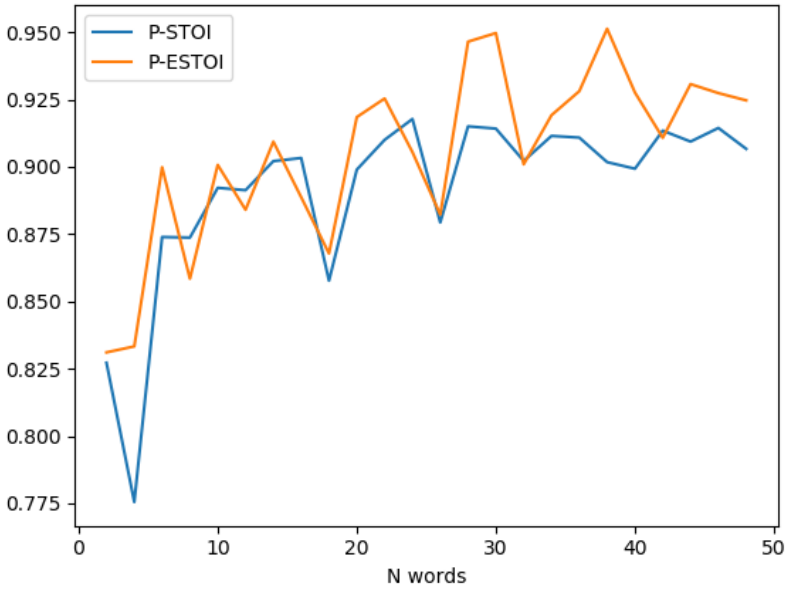



Figura 4.1: Andamento della correlazione tra P-STOI e P-ESTOI al variare del numero di parole considerate

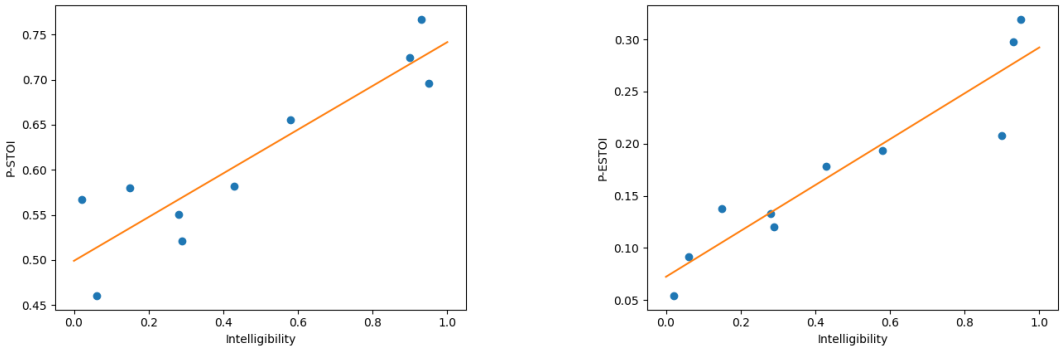
## 4.2 Diadococinesi

Nella Tabella 4.2 sono mostrati i risultati raggiunti nei tre esercizi dall'algoritmo basato su DTW, mentre per quanto riguarda l'approccio basato su rete Siamese i risultati ottenuti sono riportati in Tabella 4.3.

In entrambi i casi le misure sono state ottenute a partire dalla raccolta di registrazioni descritta nella Sottosezione 3.1.2. Inoltre le prestazioni riportate nelle due tabelle sono state calcolate come segue:

$$performance = 1 - \frac{error}{correct} \tag{4.1}$$

Confrontando i due metodi si può notare che, nonostante entrambi raggiungono ottime prestazioni, il primo commette un errore inferiore in tutti gli esercizi. Tuttavia vanno fatte alcune considerazioni. Innanzitutto le ottimizzazioni basate sulle caratteristiche del segnale fatte nei due algoritmi, riassunte in Tabella 4.4, fanno perdere ad entrambe le soluzioni un certo grado di generalità. Inoltre, l'approccio basato su rete Siamese presenta due vantaggi: per prima cosa per generare il segnale di riferimento non richiede le ripetizioni già isolate del paziente di cui si vogliono contare quelle raggiunte durante un esercizio, in secondo luogo, come riportato in Tabella 4.1, tale approccio mostra una maggiore robustezza al rumore. In questo senso sono state svolte tre prove contando le ripetizioni di "Ui" pronunciate da un soggetto sano. Nel




(a) (b)

Figura 4.2: (a) relazione tra intelligibilità e P-STOI; (b) relazione tra intelligibilità e P-ESTOI

primo caso è stato utilizzato il microfono di un auricolare collegato ad un computer in un ambiente silenzioso. Nel secondo caso è stato utilizzato lo stesso apparato di registrazione ma con una televisione accesa in sottofondo. Infine, nell'ultima prova si è aggiunto in momenti casuali un rumore intenso. Come si può notare le prestazioni in entrambi i casi subiscono una flessione, tuttavia il secondo approccio mostra una migliore risposta. Inoltre va considerato che le reti neurali impiegate sono state allenate esclusivamente su segnali raccolti in ambiente silenzioso, quindi, estendendo opportunamente il *dataset* le prestazioni potrebbero migliorare anche in questi contesti.

| | Contesto | Ui(Pred/True) | Performance(%) |
|---|---|---|---|
| Algoritmo 1 | mic | 15/21 | 57.14 |
| | mic + tv | 3/15 | 20.00 |
| | mic + tv + noise | 3/21 | 14.28 |
| Algoritmo 2 | mic | 21/21 | 100.00 |
| | mic + tv | 13/15 | 86.66 |
| | mic + tv + noise | 16/21 | 76.19 |

Tabella 4.1: Risultati ottenuti durante il conteggio di ripetizioni per l'esercizio "Ui" in contesti difficili dai due approcci: Algoritmo 1 si riferisce al sistema basato su DTW, mentre Algoritmo 2 si riferisce a quello basato su rete Siamese

| **Risultati algoritmo per la valutazione della diadococinesi basato su DTW** | | | | | | |
|---|---|---|---|---|---|---|
| | Sesso | Età | Pa (Pred./True) | Pa-Ta-Ka (Pred./True) | Ui (Pred./True) | Performance (%) |
| S01 | F | 24 | 65/65 | 34/34 | 61/62 | 99.37 |
| S02 | F | 32 | 114/114 | 61/60 | 90/92 | 98.87 |
| S03 | F | 28 | 75/75 | 48/47 | 48/48 | 99.41 |
| S04 | M | 24 | 112/112 | 56/56 | 107/108 | 99.63 |
| S05 | M | 61 | 186/186 | 47/46 | 72/72 | 99.67 |
| S06 | F | 24 | 56/56 | 32/32 | 61/61 | 100.00 |
| S07 | M | 24 | 154/155 | 60/62 | 87/88 | 98.68 |
| S08 | F | 61 | 120/120 | 54/53 | 94/95 | 99.25 |
| S09 | M | 39 | 139/140 | 51/50 | 80/80 | 99.25 |
| S10 | F | 57 | 56/56 | 49/49 | 55/55 | 100.00 |
| S11 | M | 24 | 160/159 | 63/62 | 81/81 | 99.30 |
| S12 | M | 30 | 127/125 | 46/44 | 94/94 | 98.50 |
| P01 | F | // | 73/74 | 44/43 | 48/48 | 98.78 |
| P02 | F | // | // | 34/34 | // | 100.00 |
| Performance (%) | | | 99.60 | 98.37 | 99.38 | |

Tabella 4.2: Risultati ottenuti nei tre esercizi "Pa", "Pa-Ta-Ka" e "Ui" dall'algoritmo basato su DTW

| **Risultati algoritmo per la valutazione della diadococinesi basato su rete Siamese** | | | | | | |
|---|---|---|---|---|---|---|
| | Sesso | Età | Pa (Pred./True) | Pa-Ta-Ka (Pred./True) | Ui (Pred./True) | Performance (%) |
| S01 | F | 24 | 68/65 | 41/34 | 62/62 | 93.78 |
| S02 | F | 32 | 116/114 | 59/60 | 92/92 | 98.87 |
| S03 | F | 28 | 72/75 | 46/47 | 47/48 | 97.05 |
| S04 | M | 24 | 113/112 | 56/56 | 108/108 | 99.63 |
| S05 | M | 61 | 169/186 | 50/46 | 72/72 | 93.09 |
| S06 | F | 24 | 61/56 | 33/32 | 61/61 | 95.97 |
| S07 | M | 24 | 159/155 | 65/62 | 84/88 | 96.39 |
| S08 | F | 61 | 120/120 | 54/53 | 95/95 | 99.62 |
| S09 | M | 39 | 139/140 | 51/50 | 80/80 | 99.25 |
| S10 | F | 57 | 64/56 | 51/49 | 54/55 | 93.12 |
| S11 | M | 24 | 156/159 | 55/62 | 78/81 | 95.69 |
| S12 | M | 30 | 124/125 | 40/44 | 94/94 | 98.09 |
| P01 | F | // | 69/74 | 45/43 | 49/48 | 95.15 |
| P02 | F | // | // | 40/34 | // | 79.41 |
| Performance (%) | | | 96.45 | 94.04 | 98.98 | |

Tabella 4.3: Risultati ottenuti nei tre esercizi "Pa", "Pa-Ta-Ka" e "Ui" dall'algoritmo basato su rete Siamese

| **Algoritmo 1** | **Algoritmo 2** |
|---|---|
| **Preprocessing** | |
| 1.1 Per gli esercizi "Pa" e "Pa-Ta-Ka" vengono sfruttate 15 bande della *one-third octave band analysis*, mentre per "Ui" viene considerato solo il *range* dalla settima alla quindicesima banda. | 2.1 Per tutti gli esercizi vengono sfruttate 26 bande definite dall'Equazione 3.4. Per le ripetizioni di "Pa" e "Pa-Ta-Ka" la frequenza $f_{min}$ viene fissata a 133.63 Hz, mentre per "Ui" viene fissata a 989.83 Hz. |
| 1.2 Per gli esercizi "Pa" e "Pa-Ta-Ka" viene applicato ai segnali un VAD meno selettivo, cioè che rimuove meno *frame* rispetto a quello usato per "Ui". | 2.2 Per gli esercizi "Pa" e "Pa-Ta-Ka" ai segnali viene applicato un VAD meno selettivo, cioè che rimuove meno *frame* rispetto a quello usato per "Ui". |
| **Dynamic time warping** | |
| 1.3 Per gli esercizi "Pa" e "Pa-Ta-Ka" viene impiegata la funzione di dissimilarità riportata nell'Equazione 3.1, mentre per l'esercizio "Ui" viene adottata quella riportata nell'Equazione 3.2. | 2.3 Per tutti gli esercizi come funzione di dissimilarità viene impiegata la *cosine distance*. |
| **Model** | |
| 1.4 Non viene impiegato alcun modello. | 2.4 Ogni esercizio richiede un modello allenato su un *dataset* specifico, tuttavia l'architettura della rete rimane invariata. |
| **Execution** | |
| 1.5 Non vengono introdotte altre differenze in fase di esecuzione. | 2.5 La soglia *th* utilizzata per escludere le corrispondenze trovate dal DTW *open-ended* a basso contenuto energetico varia da esercizio ad esercizio: per "Pa" è 0.6, per "Pa-Ta-Ka" è 0.2 e per "Ui" è 0.0. |

Tabella 4.4: Ottimizzazioni specifiche per esercizio per i due approcci trattati: Algoritmo 1 corrisponde a quello basato su DTW, mentre Algoritmo 2 a quello basato su rete Siamese

# Capitolo 5

# Conclusioni

Nei capitoli di questa tesi sono state introdotte ed affrontate alcune problematiche relative al Profilo Robertson impiegato per valutare lo stato clinico dei pazienti affetti da SLA, o disartria più in generale. Dopo una breve descrizione della malattia e della scala si sono approfondite le attività relative alla valutazione pneuofonicoarticolatoria, e più precisamente quelle legate all'intelligibilità e alla diadococinesi. Nel secondo capitolo si sono dunque formalizzati gli esercizi richiesti ai pazienti per cercare dei riferimenti in letteratura che sono poi stati presentati. Nel terzo capitolo sono descritti i *dataset* utilizzati e sono introdotti due approcci originali per supportare la valutazione del clinico durante le attività di diadococinesi relative alla valutazione dell'articolazione del linguaggio. Infine nel quarto capitolo sono stati mostrati tutti i risultati ottenuti, discutendo anche un breve confronto fra le due soluzioni sopra citate.

I risultati raggiunti permettono di affermare che la tecnologia odierna è matura per aiutare i clinici durante le loro visite, oppure per realizzarle a distanza. Inoltre gli approcci mostrati con ricerche più approfondite potranno essere ulteriormente affinati: nel caso del sistema basato su rete Siamese, ad esempio, sarebbe sufficiente aumentare il *dataset* utilizzato, magari inserendo delle ripetizioni raccolte con rumore di sottofondo, per migliorare le prestazioni in contesti meno controllati.

In conclusione, dalla tesi emerge che la valutazione soggettiva può essere affrontata con successo da soluzioni software appositamente studiate. Queste, oltre a migliorare la prognosi rendendola quantitativa e ripetibile, permettono ai pazienti affetti da queste patologie di svolgere le visite da remoto ed ai clinici di consultare resoconti generati automaticamente senza la necessità di assistere in diretta all'esecuzione.

# Appendice A

# STOI ed ESTOI

Gli indici di STOI ed ESTOI definiti rispettivamente in [20], [9] possono essere ottenuti da un segnale rumoroso ed uno nitido allineati temporalmente. Il calcolo avviene a partire dalla *one-third octave band analysis* che viene applicata agli spettri dei due segnali ottenuti tramite la *Fast Fourier Transform* (FFT). La *one-third octave band analysis* permette di aggregare l'energia delle singole frequenze in $J$ bande, riducendo la dimensionalità del segnale. In questo modo, dopo la sua applicazione, si ottengono due segnali $H_j(i)$ e $P_j(i)$ , rispettivamente il segnale nitido e quello degradato, entrambi con dimensione $J \times T$, con $T$ pari al numero di *frame* restituiti dalla FFT. Successivamente viene ricavato $P'_j(i)$ scalando $P_j(i)$ di un fattore $\alpha$ in modo da avere un contenuto energetico pari a $H_j(i)$, e applicando il *clipping* in modo da limitare inferiormente il *signal-to-distortion ratio* (SDR).

Per ricavare lo STOI denotato da $d^S$, è necessario determinare inizialmente una misura di intelligibilità preliminare $d_j^S(t)$ definita su un intervallo di $I$ *frame* consecutivi su ogni banda $j$. Questa grandezza intermedia è definita come segue:

$$d_j^S(t) = \frac{\sum_{i=t}^{t+I-1} (H_j(i) - \overline{H_j(i)})(P'_j(i) - \overline{P'_j(i)})}{\sqrt{\sum_{i=t}^{t+I-1} (H_j(i) - \overline{H_j(i)})^2 \sum_{i=t}^{t+I-1} (P'_j(i) - \overline{P'_j(i)})^2}} \tag{A.1}$$

con $i \in \{t, t+1, ..., t+I-1\}$ e dove

$$\begin{aligned} \overline{H_j(i)} &= \frac{1}{I} \sum_{i=t}^{t+I-1} H_j(i) \\ \overline{P'_j(i)} &= \frac{1}{I} \sum_{i=t}^{t+I-1} P'_j(i) \end{aligned} \tag{A.2}$$

Una volta calcolati tutti gli indici $d_j^S(t)$, attraverso una media tra tutti gli intervalli e tutte le bande, sarà possibile stabilire l'intelligibilità:

$$d^S = \frac{1}{(T-I+1)J} \sum_{j,t} d_j^S(t) \tag{A.3}$$

Per quanto riguarda l'indice ESTOI, entrambi i segnali $H_j(i)$ e $P_j(i)$ sono normalizzati rispetto alla media e alla varianza ed indicati rispettivamente con $\tilde{H}_j(i)$ e $\tilde{P}_j(i)$. Analogamente al caso precedente, l'intelligibilità complessiva $d^E$ viene calcolata a partire dalla media di misure intermedie $d^E(t)$.

$$d^E = \frac{1}{(T-I+1)} \sum_{t} d^E(t) \tag{A.4}$$

Come si può notare, in questo caso la media non viene eseguita anche tra le bande, infatti a differenza dello STOI, l'ESTOI non assume il contributo delle singole bande mutualmente indipendente. Per questa ragione la formula per il calcolo della misura intermedia $d^E(t)$ diventa:

$$d^E(t) = \frac{1}{I} \sum_{i=t}^{t+I-1} \frac{\sum_{j=1}^{J} (\tilde{H}_j(i) - \overline{\tilde{H}_j(i)})(\tilde{P}_j(i) - \overline{\tilde{P}_j(i)})}{\sqrt{\sum_{j=1}^{J} (\tilde{H}_j(i) - \overline{\tilde{H}_j(i)})^2 \sum_{j=1}^{J} (\tilde{P}_j(i) - \overline{\tilde{P}_j(i)})^2}} \tag{A.5}$$

con $i \in \{t, t+1, ..., t+I-1\}$ e dove

$$\begin{aligned} \overline{\tilde{H}_j(i)} &= \frac{1}{J} \sum_{j=1}^{J} \tilde{H}_j(i) \\ \overline{\tilde{P}_j(i)} &= \frac{1}{J} \sum_{j=1}^{J} \tilde{P}_j(i) \end{aligned} \tag{A.6}$$

# Appendice B

# Dynamic Time Warping

Il Dynamic Time Warping (DTW) [6] è un algoritmo molto utilizzato nell'ambito delle serie temporali. Esso permette di adattare tra loro due segnali: il primo detto *query*, $X = (x_1, ..., x_n)$ ed il secondo detto *reference*, $Y = (y_1, ..., y_m)$ in modo da minimizzare la loro differenza [Fig. B.1].

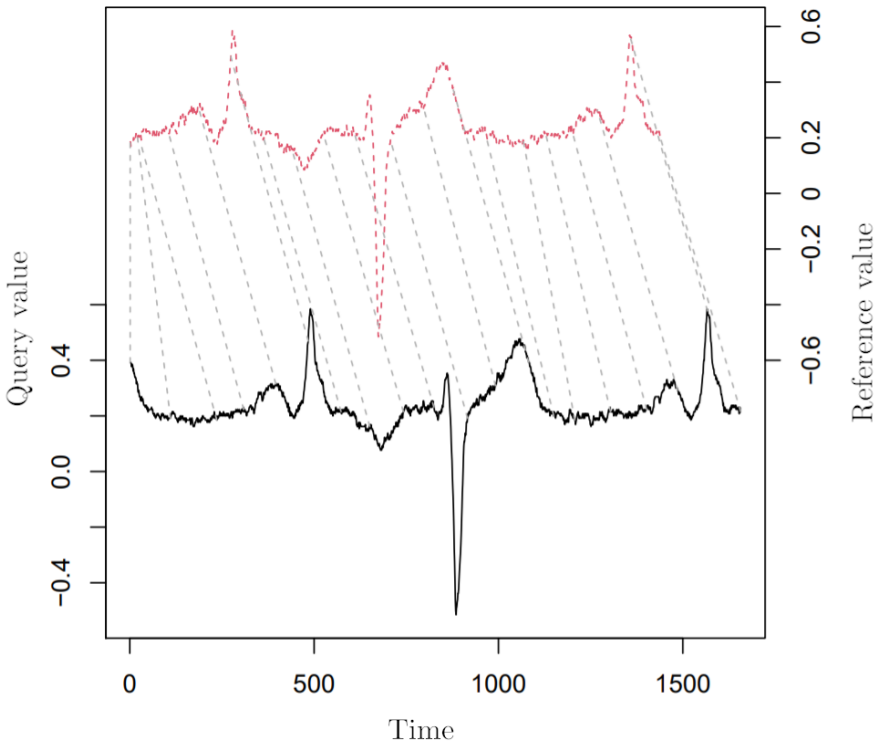



Figura B.1: Allineamento di due serie temporali: il segnale *query* in nero e il segnale *reference* in rosso tratteggiato

L'algoritmo agisce attraverso un solo grado di libertà: il tempo. Infatti l'unico modo per aumentare la similarità tra i due segnali è ripetere i valori della *query* e della *reference* in modo da ottenere due serie di pari lunghezza. Per trovare la migliore

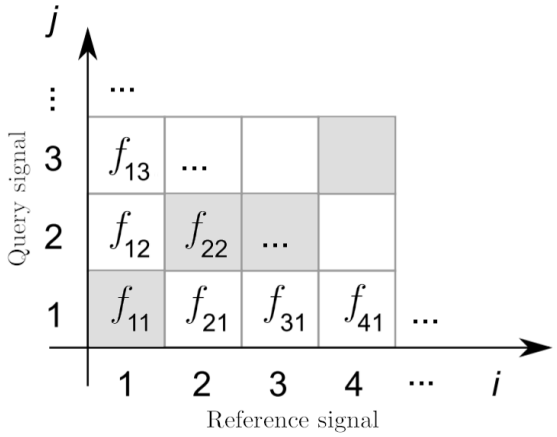



Figura B.2: *Cross-distance matrix* con *warping path* rappresentato in grigio

sequenza di valori ripetuti, l'algoritmo ricava la *cross-distance matrix* $C$ a partire dalle due serie. La matrice ha dimensione $(n \times m)$ ed è definita come:

$$c(i, j) = f(x_i, y_j) \tag{B.1}$$

dove $f$ è una funzione di dissimilarità locale. [1]

Una volta ottenuta la *cross-distance matrix*, come mostrato in Figura B.2, viene cercato il migliore *warping path*. Il *warping path* $\phi(k), k = 1, ..., T$, è definito come:

$$\begin{aligned} \phi(k) = (\phi_x(k), \phi_y(k)) \, con \\ \phi_x(k) \in \{1...N\}, \\ \phi_y(k) \in \{1...M\} \end{aligned} \tag{B.2}$$

dove $\phi_x$ e $\phi_y$ sono le *warping function* che mappano gli indici temporali di $X$ e $Y$ rispettivamente sulla serie temporale allineata. La ricerca del *warping path* ottimo si basa sulla minimizzazione della distorsione media complessiva tra i due segnali $d_\phi$ data da:

$$d_\phi(X, Y) = \sum_{k=1}^{T} c(\phi_x(k), \phi_y(k)) w_\phi(k) / p_\phi \tag{B.3}$$

dove $w_\phi(k)$ è un coefficiente pesato definito lungo il *warping path*, mentre $p_\phi$ è il fattore di normalizzazione corrispondente per rendere comparabili le distanze ottenute su *path* differenti.

Quindi, data la *cross-distance matrix* ottenuta a partire dalle due serie temporali, è possibile trovare il miglior *warping path* cercando quello che minimizza $d_\phi(X, Y)$,

[1] Solitamente viene impiegata la *euclidean distance*

inoltre sarà possibile definire la distanza $D$, detta $DTW$ *distance*, tra i due segnali come:

$$D(X, Y) = \min_{\phi} d_{\phi}(X, Y) \tag{B.4}$$

La soluzione dell'Equazione B.4 generalmente viene risolta considerando dei vincoli aggiuntivi tra i quali i più comuni sono quelli di monoticità che impediscono cicli di ripetizioni:

$$\begin{aligned} \phi_x(k+1) &\geqslant \phi_x(k) \\ \phi_y(k+1) &\geqslant \phi_y(k) \end{aligned} \tag{B.5}$$

Quindi l'algoritmo $DTW$ è un utile strumento che permette di allineare tra loro due segnali, e che fornisce una misura di distanza tra essi. Per questa ragione è particolarmente diffuso anche in analisi audio per riconoscere *pattern* indipendentemente dal loro sfasamento temporale.

# Bibliografia


[1] Luca Bertinetto, Jack Valmadre, João F. Henriques, Andrea Vedaldi, and Philip H. S. Torr. Fully-convolutional siamese networks for object tracking. *CoRR*, abs/1606.09549, 2016.

[2] Virginia Bozzoni. Amyotrophic lateral sclerosis and environmental factors. 2016.

[3] Anna Cantagallo, Fabio La Porta, and L. Abenante. Dysarthria assessment; robertson profile and self-assessment questionnaire. *Acta Phon Lat*, 28:246–261, 2006.

[4] Jean-Pierre Martens Catherine Middag, Gwen Van Nuffelen and Marc De Bodt. Objective intelligibility assessment of pathological speakers. pages 1745–1748, 2008.

[5] Mehdi Ghasemi and Jr. Brown, Robert H. Genetics of amyotrophic lateral sclerosis. *Cold Spring Harbor Perspectives in Medicine*, 8:a024125, 2017.

[6] Toni Giorgino. Computing and visualizing dynamic time warping alignments inr: Thedtwpackage. *Journal of Statistical Software*, 31, 2009.

[7] Adrienne Perlman Jon Gunderson Thomas Huang Kenneth Watkin Simone Frame Heejin Kim, Mark Hasegawa-Johnson. Dysarthric speech database for universal access research. pages 1741–1744, 2008.

[8] Parvaneh Janbakhshi, Ina Kodrasi, and Herve Bourlard. Automatic pathological speech intelligibility assessment exploiting subspace-based analyses. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:1717–1728, 2020.

[9] Jesper Jensen and Cees H. Taal. An algorithm for predicting the intelligibility of speech masked by modulated noise maskers. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24:2009–2022, 2016.

[10] Iván López-Espejo, Z. Tan, and J. Jensen. Keyword spotting for hearing assistive devices robust to external speakers. 2019.

[11] Ivan Lopez-Espejo, Zheng-Hua Tan, and Jesper Jensen. Keyword spotting for hearing assistive devices robust to external speakers. pages 3223–3227, 2019.

[12] David Martinez, P. Green, and H. Christensen. Dysarthria intelligibility assessment in a factor analysis total variability space. 2013.

[13] P. Masrori and P. Van Damme. Amyotrophic lateral sclerosis: a clinical review. *European Journal of Neurology*, 27:1918–1929, 2020.

[14] Herv´e Bourlard Parvaneh Janbakhshi, Ina Kodrasi. Pathological speech intelligibility assessment based on the short-time objective intelligibility measure. pages 6405–6409, 2019.

[15] Emanuele Principi, Stefano Squartini, Erik Cambria, and Francesco Piazza. Acoustic template-matching for automatic emergency state detection: An elm based algorithm. *Neurocomputing*, 149:426–434, 2015.

[16] Lawrence Rabiner and Biing-Hwang Juang. *Fundamentals of Speech Recognition*. Prentice-Hall, Inc., 1993.

[17] Fadhilah Rosdi, Mumtaz Begum Mustafa, Siti Salwah Salim, and Nor Azan Mat Zin. Automatic speech intelligibility detection for speakers with speech impairments: The identification of significant speech features. *Sains Malaysiana*, 48:2737–2747, Dec 2019.

[18] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26:43–49, 1978.

[19] Larry E. Smith and Cecil L. Nelson. International intelligibility of english: directions and resources. *World Englishes*, 4:333–342, 1985.

[20] Cees H. Taal, Richard C. Hendriks, Richard Heusdens, and Jesper Jensen. A short-time objective intelligibility measure for time-frequency weighted noisy speech. pages 4214–4217, 2010.

[21] Sara Zarei, Karen Carr, Luz Reiley, Kelvin Diaz, Orleiquis Guerra, PabloFernandez Altamirano, Wilfredo Pagani, Daud Lodin, Gloria Orozco, and Angel Chinea. A comprehensive review of amyotrophic lateral sclerosis. *Surgical Neurology International*, 6:171, 2015.

[22] Peng Zhang and Xueliang Zhang. Deep template matching for small-footprint and configurable keyword spotting. 2020.