ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA

CAMPUS DI CESENA
DIPARTIMENTO DI INGEGNERIA DELL'ENERGIA ELETTRICA E DELL'INFORMAZIONE "*GUGLIELMO MARCONI*"

CORSO DI LAUREA IN INGEGNERIA BIOMEDICA

Sintesi vocale attraverso speech BCI invasive: nuove prospettive verso un parlato intelligibile

Elaborato in Strumentazione biomedica

Relatore:

Prof. Ing. Cristiano Cuppini

Presentata da:

Anna Zanuccoli

Sessione I
Anno Accademico 2019-2020

Indice

Introduzione

Ogni anno milioni di persone risultano affette da numerose patologie neurodegenerative o traumatiche come l'ictus, la sclerosi laterale amiotrofica (SLA), la Sindrome Locked-In (LIS) o le lesioni al midollo spinale. Spesso tali patologie comportano deficit molto invalidanti e lesioni permanenti delle vie nervose deputate al controllo dei muscoli coinvolti nell'esecuzione volontaria delle azioni, precludendo anche la possibilità di comunicare.
In particolare le persone affette da SLA soffrono di un disturbo neurologico cronico progressivo in cui i motoneuroni del cervello e del midollo spinale degenerano, riducendo la loro capacità di attivare il sistema muscolo scheletrico. Analogamente i soggetti colpiti da LIS vivono in uno stato di paralisi totale ma con cognizione e sensazione intatte.
Per queste persone, la perdita del linguaggio è un'ulteriore afflizione che peggiora la loro condizione di vita: rende molto difficile la comunicazione, e più in generale, può portare a un profondo isolamento sociale, fino alla depressione.

Per risolvere i problemi di comunicazione causati da queste tipologie di malattie negli ultimi trenta anni sono stati sviluppati numerosi strumenti che permettono il ripristino delle capacità comunicative.

Gli strumenti che permettono il ripristino della comunicazione sono le BCI (Brain Computer Interface), cioè delle interfacce che collegano l'attività celebrale ad un computer che ne registra e ne interpreta le variazioni.
Gli approcci che utilizzano le BCI sono molteplici e molto diversi tra loro, in particolare si possono distinguere per tecnica di rilevazione del segnale celebrale (invasiva o non invasiva), paradigma di decodifica del segnale (basato su varie caratteristiche del segnale) e metodo di comunicazione (indiretta o diretta).

Una caratteristica comune alla maggior parte di tali strumenti è che la comunicazione permessa dalle BCIs risulta solitamente essere molto lenta rispetto alla capacità comunicativa propria del linguaggio naturale; infatti spesso attraverso sistemi indiretti di produzione del linguaggio si riescono a riprodurre solo 5/6 parole al minuto, un numero nettamente inferiore rispetto a quelle di un discorso fluente.
Per questo motivo negli ultimi dieci anni la ricerca si è concentrata su altre possibili soluzioni in cui le tecniche di BCIs fossero in grado di controllare un sintetizzatore vocale in tempo reale al fine di ripristinare una comunicazione fluente decodificando l'attività neurale direttamente dalle aree del cervello deputate al controllo del parlato.

In questo contesto si inserisce l'obiettivo della tesi che consiste nel confrontare tre differenti sistemi di speech BCI in grado di riprodurre un discorso fluente e intelligibile.
I metodi di BCI confrontati nel presente elaborato sintetizzano il parlato attraverso la rilevazione invasiva dell'attività cerebrale misurata tramite elettrocorticografia (ECoG) da specifiche aree del cervello deputate al linguaggio.
Tutti i metodi descritti decodificano l'attività cerebrale attraverso delle reti neurali, cioè dei modelli matematici il cui funzionamento è ispirato alle connessioni neurali biologiche, e utilizzano l'attività cerebrale direttamente collegata alla produzione del linguaggio per controllare il sistema di speech BCI, attuando quindi una comunicazione diretta.

I tre approcci descritti si differenziano per la strategia di decodifica del segnale cerebrale, cioè per il tipo di caratteristiche del segnale che vengono decodificate, e per il tipo di reti neurali utilizzate per decodificarlo.

I primi due metodi, descritti rispettivamente nei capitoli 3 e 4, decodificano le caratteristiche acustiche del parlato a partire dall'attività cerebrale ma utilizzano due differenti rappresentazioni acustiche e due differenti tipologie di reti neurali.
Il sistema di speech BCI realizzato da Akbari et al (capitolo 3) [1] utilizza delle locally e fully connected neural networks (disposte secondo una particolare sequenza) per decodificare una rappresentazione acustica che è data dall'unione di più parametri utilizzati per la sintesi vocale (l'inviluppo spettrale, le frequenze fondamentali, la banda di aperiodicità e l'intonazione).
Il sistema di speech BCI realizzato da Angrick et al. (capitolo 4) [2] utilizza delle deep neural networks per decodificare una rappresentazione acustica basata sullo spettrogramma logaritmico in scala mel (logMel spectrogram).

Il terzo metodo che viene confrontato nel presente elaborato descrive il lavoro di Anumanchipalli et al. (capitolo 5) [3] in cui vengono utilizzate delle reti neurali ricorrenti per decodificare le caratteristiche articolatorie del parlato a partire dall'attività cerebrale. Le caratteristiche articolatorie del parlato descrivono come varia nel tempo la posizione dei principali organi del tratto vocale come la lingua, le labbra, la mandibola e la laringe. I sistemi BCIs che si basano su questa tecnica di decodifica predicono i movimenti degli articolatori del parlato a partire dall'attività celebrale, convertono la rappresentazione articolatoria in una rappresentazione spettrale acustica e a partire da quest'ultima ricostruiscono il discorso parlato attraverso un sintetizzatore vocale.

L'obiettivo della tesi è quello di confrontare tali differenti approcci di speech BCI che rilevano l'attività cerebrale in modo invasivo e utilizzano metodi di comunicazione diretta.
Ad oggi questi innovativi sistemi di speech BCI rappresentano la soluzione più promettente per risolvere problemi di comunicazione per persone affette da SLA e LIS poiché, a differenza dei sistemi di BCI che utilizzano la comunicazione indiretta e rilevano il segnale in modo non

invasivo, sono in grado di riprodurre un discorso intelligibile, fluente ed in tempo reale che permette di ripristinare una capacità comunicativa molto simile a quella del linguaggio naturale migliorando la qualità della vita dell'utente.

Capitolo 1

1.Brain Computer Interface e sintesi vocale

Con il termine Brain-Computer Interface (BCI, ovvero interfaccia cervello-computer) si definisce un dispositivo costituto da componenti hardware e software che permette all'utente di interagire e comunicare con l'ambiente esterno utilizzando solamente specifici segnali generati dall'attività cerebrale

Una speech BCI è un dispositivo che produce una forma di output vocale, come la selezione di parole / lettere o la generazione del suono del parlato, a partire dalla rilevazione dell'attività cerebrale da specifiche aree del cervello deputate al linguaggio.
Le tipologie di BCI si differenziano sulla base di alcune caratteristiche come la metodologia di registrazione del segnale celebrale, che può essere invasiva o non invasiva, alla modalità di comunicazione, indiretta o diretta, e alla strategia di decodifica del segnale a cui è associato un sistema di sintesi vocale.

Nel capitolo seguente viene inizialmente descritta l'organizzazione funzionale delle aree cerebrali coinvolte nella produzione del parlato e successivamente viene effettuata una classificazione delle BCIs concentrandosi in particolare sulla descrizione dell'ECoG che è la tecnica di registrazione utilizzata dai tre differenti approcci di speech BCI trattati dettagliatamente nella tesi.

1.1Organizzazione neuroanatomica-funzionale del linguaggio

La produzione linguistica e lessicale è un processo cerebrale complesso che coinvolge diverse aree del cervello. Nel caso di danni cerebrali e malattie neurodegenerative che provochino la perdita di tali funzioni l'approccio riabilitativo prevede il ripristino della comunicazione attraverso le speech BCI.

Per realizzare una speech BCI è fondamentale conoscere l'organizzazione neuroanatomica funzionale del linguaggio poiché permette di comprendere quali sono le aree cerebrali dalle quali l'interfaccia deve registrare l'attività neuronale. Inoltre, in base all'area cerebrale considerata e alle rappresentazioni neurali del parlato che essa codifica, variano anche le strategie di decodifica del sistema di speech BCI. Per questi motivi di seguito viene descritta l'organizzazione funzionale delle aree cerebrali coinvolte nella produzione del parlato.

1.1.1 Il modello a doppio flusso

L'elaborazione del parlato coinvolge un'ampia rete corticale che realizza due principali funzioni al fine di produrre il discorso: associare alla rappresentazione acustica del discorso una rappresentazione semantica e associare delle rappresentazioni articolatorie ai suoni del parlato. Al fine di realizzare entrambe le funzioni le informazioni vocali seguono due percorsi di elaborazione differenti: la via ventrale e la via dorsale. Questa descrizione del processo neurale correlato alla produzione linguistica viene chiamata "modello a doppio flusso".
Secondo tale modello le prime fasi dell'elaborazione del parlato si verificano bilateralmente nelle regioni uditive del giro temporale superiore (STG), in cui avviene l'analisi spettro temporale, e nel solco temporale superiore (STS), in cui sono contenute le rappresentazioni fonologiche del linguaggio. A partire da queste regioni il meccanismo di elaborazione del parlato si divide in due grandi flussi: il flusso ventrale, che coinvolge strutture nelle porzioni superiore e mediale del lobo temporale, responsabile dell'elaborazione dei segnali vocali per la comprensione del parlato; e il flusso dorsale, che coinvolge il lobo temporale posteriore, la giunzione parietale-temporale (Spt) e il lobo frontale posteriore, ed è responsabile della traduzione di rappresentazioni acustiche dei segnali vocali in rappresentazioni articolatorie.

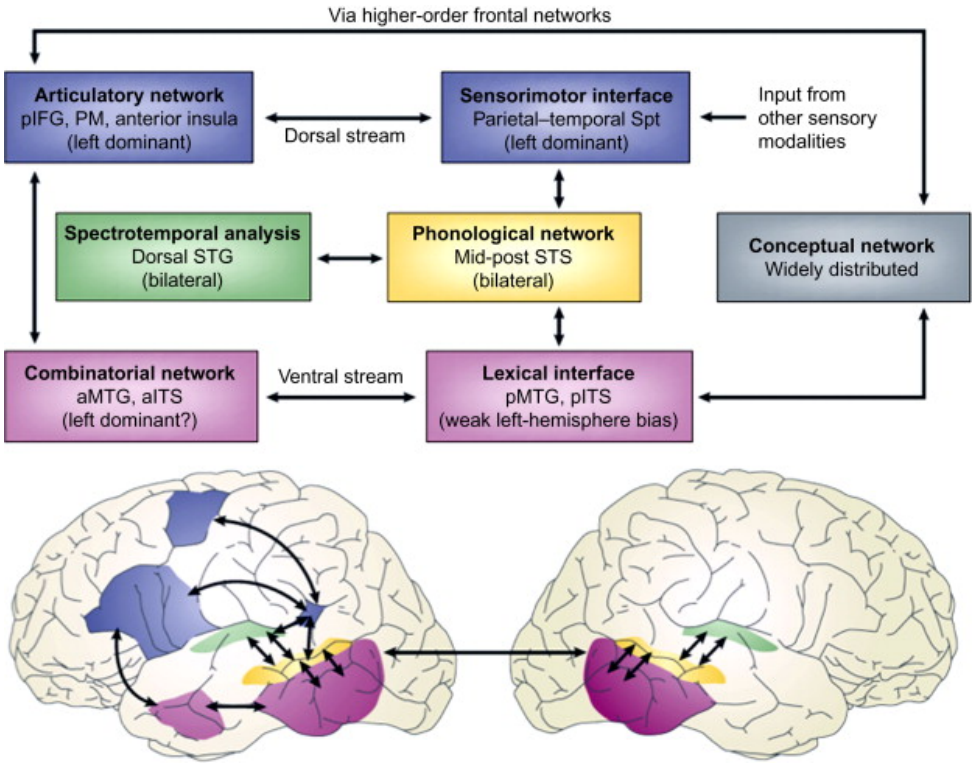

Figura 1.1: Modello a doppio flusso di elaborazione vocale. Regioni uditive STG (analisi spettrotemporale; verde) e STS (accesso fonologico / rappresentazione; giallo); un flusso ventrale del lobo temporale (accesso lessicale e processi combinatori; rosa), flusso dorsale (fortemente dominante a sinistra, blu) coinvolge strutture a livello della giunzione parietale-temporale (Spt) e del lobo

frontale. IFG, giro frontale inferiore; ITS, solco emporale inferiore, MTG, giro temporale medio, PM, premore, Spt, planum temporale; STG, giro temporale superiore; STS, solco temporale superiore.

1.1.2 Il flusso ventrale

Il flusso ventrale coinvolge le regioni del lobo temporale medio e inferiore ed è organizzato bilateralmente nei due emisferi. Il fatto che il flusso ventrale sia localizzato bilateralmente non implica una ridondanza del processo di elaborazione cerebrale poiché molti studi di neuroimaging hanno evidenziato l'esistenza di asimmetrie dal punto di vista computazionale nella via ventrale dell'emisfero destro e sinistro. La natura di queste differenze è attualmente in discussione ma la sua esistenza indica che l'elaborazione del parlato avvenga su percorsi paralleli nella mappatura dal suono al significato lessicale.

Il flusso ventrale collega la rappresentazione acustica del parlato alla rappresentazione fonetica attraverso STG e STS, e successivamente utilizza le informazioni fonologiche per accedere alla rappresentazione semantica delle parole.
Secondo il modello a doppio flusso le rappresentazioni concettuali-semantiche sono distribuite in tutta la corteccia; tuttavia esiste una regione cerebrale, implicata nella via ventrale, che funge da interfaccia computazionale tra rappresentazioni a livello fonologico e rappresentazioni concettuali distribuite. Questa interfaccia non è il sito per la memorizzazione di informazioni concettuali ma memorizza informazioni relative alla relazione tra informazione fonologica e informazione concettuale. La maggior parte dei ricercatori concorda sul fatto tale area sia contenuta all'interno del lobo temporale tuttavia sono presenti pareri contrastanti in letteratura poichè alcuni autori affermano che l'interfaccia sia costituita dalla parte anteriore [4], [5], [6] del lobo temporale mentre altri sostengono che si trovi nella parte posteriore [7], [8].

1.1.3 Il flusso dorsale

Il flusso dorsale, che coinvolge il lobo temporale posteriore e il lobo frontale posteriore, è responsabile dell'integrazione sensomotoria del parlato e associa i suoni del parlato alle rappresentazioni articolatorie.

La via dorsale proietta dalle cortecce uditive alla regione posteriore del lobo temporale, e poi al lobo frontale posteriore, inoltre la via dorsale proietta anche alle aree premotorie e alle aree motorie supplementari. La produzione vocale è il risultato di movimenti coordinati le cui rappresentazioni sono contenute nella parte ventrale della corteccia sensomotoria (vSMC), dove la rappresentazione degli organi articolatori del parlato (cioè labbra, mascella, lingua e laringe) è organizzato somatotopicamente.

Data questa ampia distribuzione delle aree coinvolte nella produzione del linguaggio, per realizzare una speech BCI, è necessario effettuare una scelta opportuna delle aree corticali da cui registrare e decodificare l'attività evocata durante la produzione vocale. Una possibilità, ad esempio, è quella di utilizzare segnali provenienti da aree uditive del flusso ventrale, come STG e STS, per decodificare le rappresentazioni spettrale-temporale del contenuto acustico del parlato. Al contrario si potrebbero utilizzare i segnali registrati nella vSMC per decodificare le caratteristiche articolatorie della produzione vocale.

Una volta determinata la regione cerebrale da cui si registra l'attività neurale correlata alla produzione linguistica è necessario definire le caratteristiche della speech BCI che si vuole realizzare. Di seguito vengono descritte le possibili tecniche di registrazione dell'attività cerebrale con cui può essere implementato un sistema di BCI.

1.2 Tecniche di registrazione dell'attività cerebrale

Le tecniche di registrazione dell'attività cerebrale si dividono in tecniche invasive e non invasive e soltanto alcune tra queste sono adatte per la realizzazione di una speech BCI.

Tra le tecniche di rilevazione non invasive troviamo i metodi che si basano sulla rilevazione di segnali metabolici per lo studio dell'attività celebrale. Le più utilizzate nell'ambito delle speech BCI sono la risonanza magnetica funzionale (fMRI) e la Near-Infrared Spectroscopy (NIRS), ossia la spettroscopia che usa la regione dello spettro del campo elettromagnetico vicino all'infrarosso. Queste tecniche rilevano l'attività cerebrale mediante misurazioni non invasive di variazioni del livello di ossigenazione cerebrale, le quali sono correlate ai cambiamenti dell'attività neuronale.
Tali metodi basati sulla rilevazione dell'attività metabolica possono fornire indicazioni dell'attività celebrale con una risoluzione spaziale molto elevata. Tuttavia i processi metabolici che fMRI e NIRS rilevano sono lenti in natura e impiegano diversi secondi per completarsi, per questo motivo la risoluzione temporale di tali metodi è dell'ordine del secondo. I processi vocali invece avvengono nell'ordine delle decine di millisecondi, il che rende tali tecniche di rilevazione non adeguati alla loro misura.

Le tecniche di rilevazione non invasiva che, invece, vengono principalmente utilizzate nell'ambito delle speech BCI sono quelle che sfruttano la registrazione di segnali elettrofisiologici per determinare l'attività cerebrale; tra queste troviamo l'elettroencefalografia (EEG) e la magnetoencefalografia (MEG). Tali metodologie sono più adeguate per rilevare la dinamica dei processi di produzione linguistica in quanto hanno una risoluzione temporale pari a 0,01 - 0,1 s.

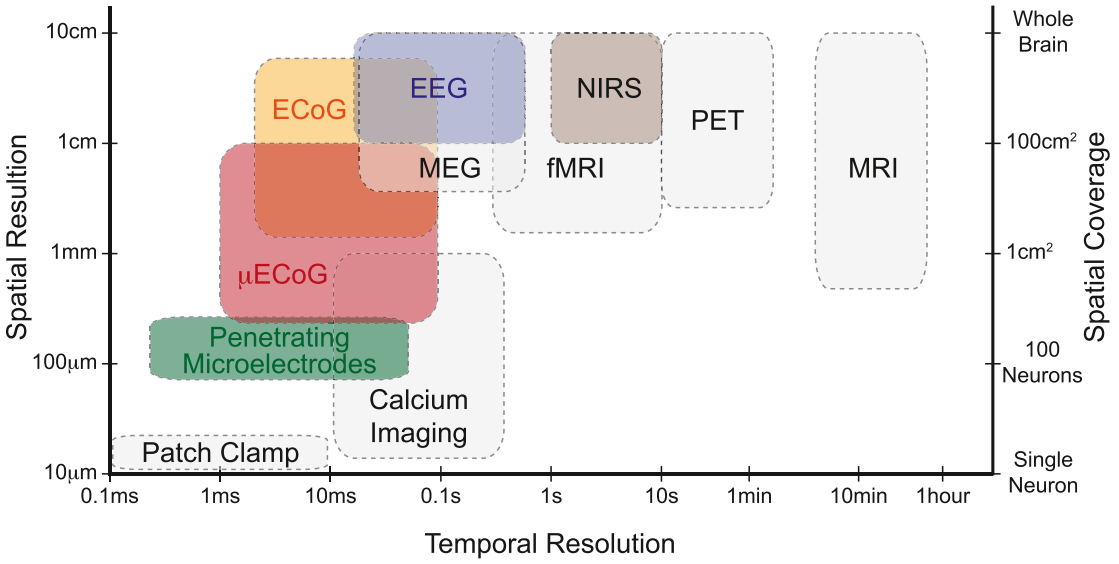

Figura 1.2: Risoluzione spaziale, temporale e copertura spaziale di varie modalità di monitoraggio dell'attività neuronale. Per ciascuna modalità mostrata, il limite inferiore del riquadro specifica la risoluzione spaziale indicata sull'asse sinistro, mentre il limite superiore specifica la copertura spaziale sull'asse destro. La larghezza di ogni riquadro indica il tipico intervallo raggiungibile di risoluzione temporale.

La MEG è una tecnica di registrazione non invasiva che prevede la rilevazione dei campi magnetici prodotti dall'attività elettromagnetica cerebrale utilizzando magnetometri posizionati intorno alla testa dell'utente. La magnetoencefalografia è caratterizzata da un'elevata risoluzione temporale e un'accettabile risoluzione spaziale, tuttavia presenta alcuni svantaggi tra cui l'elevata sensibilità agli artefatti da movimento dei muscoli facciali che possono determinare delle distorsioni nel segnale MEG. Un ulteriore svantaggio della magnetoencefalografia riguarda l'impossibilità di realizzare dispositivi portatili poichè la strumentazione per MEG non può essere spostata facilmente ed è ingombrante; per questi motivi la MEG è meno utilizzata dell'EEG nell'ambito delle speech BCI.

La tecnica di registrazione elettrofisiologica più diffusa è l'elettroencefalografia (EEG), la quale registra l'attività elettrica, derivante dalla conduzione volumetrica dell'attività neuronale sincronizzata di tutto il cervello, utilizzando elettrodi posizionati sul cuoio capelluto.

EEG si è dimostrata una tecnica adeguata a registrare in modo affidabile l'attività cerebrale per il controllo BCI poiché è caratterizzata da un'ottima risoluzione temporale e non è invasiva; tuttavia il fatto che gli elettrodi siano posizionati sul cuoio capelluto di un utente presenta alcuni svantaggi. In primo luogo questa tecnica è sensibile agli artefatti, come i movimenti oculari o i movimenti facciali residui, e la risoluzione spaziale e il rapporto segnale-rumore sono molto inferiori rispetto alle tecniche di registrazioni invasiva. Inoltre, quando si utilizza l'EEG gli elettrodi devono essere posizionati sulla testa del soggetto e ricalibrati ad ogni utilizzo, il che di solito comporta l'assistenza da parte di un ricercatore o di un assistente competente.

Nonostante la presenza di questi svantaggi l'EEG è ad oggi la tecnica di registrazione dell'attività cerebrale maggiormente utilizzata nell'ambito delle BCI grazie alla sua non invasività, alla buona risoluzione temporale e alla possibilità di progettare dispositivi portatili.

Negli ultimi anni la ricerca ha rivolto la propria attenzione verso le tecniche invasive di registrazione dell'attività cerebrale come metodo per la progettazione e realizzazione delle BCI. I principali metodi di registrazione invasiva dell'attività cerebrale sono i microelettrodi intracorticali e l'elettrocorticografia o elettroencefalografia intracranica (ECoG).

I primi approcci di registrazione invasiva prevedevano l'utilizzo di microelettrodi intracorticali che possono misurare i potenziali di azione extracellulare (EAP) e i potenziali di campo locale (LFP) da piccole popolazioni neuronali.
La registrazione attraverso microelettrodi intracranici presenta alcuni vantaggi poichè è caratterizzata da un'elevata risoluzione spazio-temporale, un elevato rapporto segnale rumore e robustezza nei confronti di artefatti oculari. L'utilizzo di microelettrodi intracorticali comporta tuttavia alcune grandi limitazioni. Innanzitutto questa tecnica consente di registrare l'attività cerebrale solamente da piccole regioni corticali, inoltre l'inserimento di elettrodi penetranti può causare danni al tessuto cerebrale e inevitabilmente generare la formazione di tessuto cicatriziale che a lungo termine compromette la stabilità del segnale. A causa di questi problemi di longevità ed estrema invasività l'impianto cronico di microelettrodi non viene quasi mai utilizzato.

La principale tecnica di rilevazione invasiva utilizzata nell'ambito delle speech BCI è l'elettrocorticografia o elettroencefalografia intracranica. L'ECoG registra l'attività elettrica cerebrale rilevando i potenziali postsinaptici sincronizzati attraverso il posizionamento di elettrodi sulla superficie corticale, sopra (epidurale) o sotto (subdurale) la dura madre. Rispetto all' EEG si trova quindi in una posizione più vicina alla sorgente del segnale ma contemporaneamente non è estremamente invasiva come l'impianto di microelettrodi intracorticali.

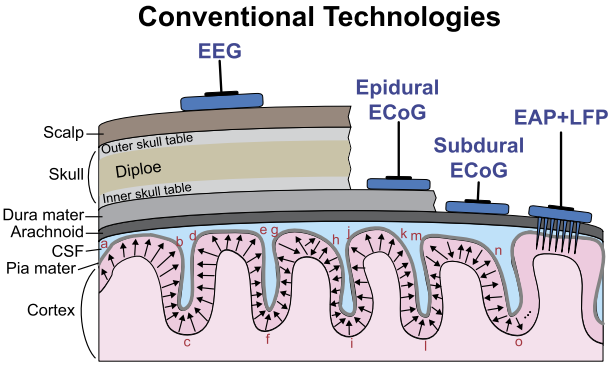

Figura 1.3: Metodi di elettrofisiologia convenzionali tra cui EEG, ECoG epidurale e subdurale e registrazione LFP con microelettrodi penetranti

L’impianto di elettrodi ECoG per registrare l’attività cerebrale, nella pratica medica, viene utilizzato, ad esempio, per pazienti con epilessia intrattabile al fine di localizzare la zona di insorgenza delle crisi, prima della resezione del tessuto cerebrale. A causa dell’invasività dell’ECoG, solo in rari casi i pazienti vengono sottoposti alla procedura di impianto; tuttavia, se i pazienti acconsento a partecipare ad attività di ricerca, durante il tempo per cui viene mantenuto l’impianto si possono studiare in maniera molto dettagliata alcuni meccanismi e funzioni cerebrali come l’abilità del linguaggio.

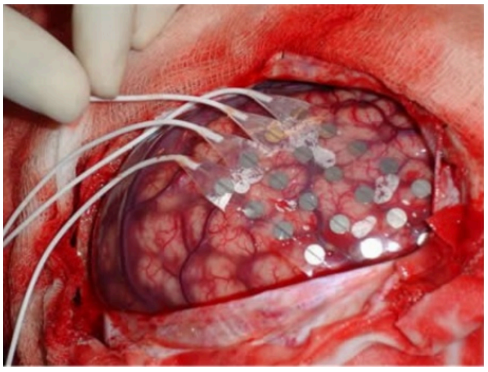

Figura 1.4: Impianto di elettrodi per ECoG

ECoG è una tecnica di registrazione che comporta molti vantaggi: offre un’ottima risoluzione spaziale, dell’ordine del millimetro, e spettrale (0–500 Hz); inoltre le sue prestazioni sono migliori rispetto a quelle dell’EEG anche in termini di ampiezza del segnale (50–100 mV) e rapporto segnale-rumore. Un ulteriore aspetto positivo di questa tecnica consiste nel fatto che i suoi elettrodi possono coprire e mappare ampie aree cerebrali (in genere corteccia frontale, temporale e parietale), il che è vantaggioso considerando l’estensione e il numero di aree cerebrali coinvolte nella produzione del linguaggio. Infine, l’ECoG ha una risoluzione temporale molto più elevata (millisecondi) rispetto alle tecniche di misurazione emodinamica, come fMRI e NIRS.

L’ECoG tuttavia presenta anche alcuni svantaggi e il più rilevante tra tutti è l’invasività degli elettrodi poiché il loro posizionamento sulla corteccia cerebrale richiede un intervento chirurgico. Un ulteriore problema dell’impianto di questo dispositivo è la conseguente formazione di tessuto cicatriziale che riduce l’efficacia del segnale che viene rilevato e aumenta il rischio per la salute del paziente.

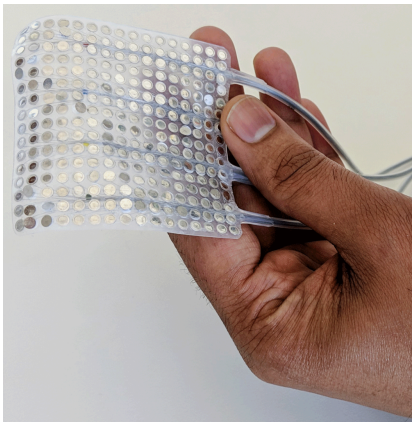

Figura 1.5: Elettrodo a griglia utilizzato nelle sperimentazioni della university of California, San Francisco. I conduttori sono realizzati in platino/iridio, questo fornisce la possibilità di localizzazione degli elettrodi attraverso risonanza magnetica. Gli elettrodi presentano un diametro che può variare tra i 2.5 e i 3.0 mm mentre lo spessore delle griglie varia da 0.5 mm a 0.8 mm

Gli studi effettuati su pazienti epilettici continuano a fornire importanti informazioni scientifiche per la futura realizzazione di sistemi di speech BCIs basati su ECoG, nonostante gli evidenti inconvenienti di questa tecnica.
Sulla base dell'esperienza accumulata dagli studi effettuati negli ultimi anni vi è una crescente fiducia nei potenziali benefici delle speech BCIs basate su ECoG per migliorare la qualità della vita nei pazienti con gravi disabilità comunicative. Per questi motivi nel presente elaborato ho deciso di confrontare dei nuovi approcci di speech BCI che rilevano l'attività cerebrale attraverso ECoG, e nei paragrafi successivi vengono descritte con maggiore dettaglio la generazione e le caratteristiche del segnale registrato attraverso questa tecnica.

1.3 Elettrocorticografia

1.3.1 Generazione del segnale

I neuroni corticali sono disposti con una determinata successione nella corteccia, andando a formare gli strati corticali, e tali neuroni sono classificati in due principali categorie: neuroni corticali piramidali e neuroni corticali non piramidali.

Le cellule piramidali costituiscono la parte preponderante della corteccia cerebrale e sono caratterizzate da una specifica disposizione: gli assoni delle cellule sono disposti perpendicolarmente alla corteccia mentre i dendriti sono disposti parallelamente tra loro, come mostrato in figura 1.6.
Ogni cellula è una sorgente di campo elettrico e se i neuroni piramidali vicini si attivano in modo sincrono si generano tanti campi elettrici che si sommano determinando un campo

elettrico complessivo; tale sommatoria dei singoli campi elettrici delle cellule è possibile grazie alla disposizione parallela dei dendriti.
Il campo elettrico complessivo che si genera dall'attivazione sincrona di più neuroni può essere rilevato tramite degli elettrodi andando a determinare il segnale ECoG.

Le cellule non piramidali sono caratterizzate da una forma ovale e il loro orientamento spaziale rispetto alla corteccia non è ordinato, alcune sono orientate orizzontalmente mentre altre verticalmente, pertanto non partecipano in modo significativo alla generazione del segnale elettrico.

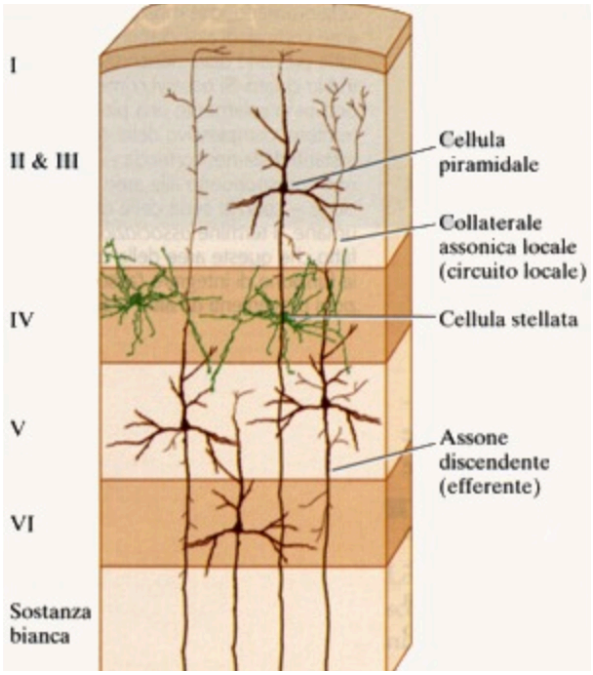

Figura 1.6: rappresentazione nei neuroni corticali all'interno degli strati della corteccia

Attraverso la trasmissione sinaptica i neuroni possono comunicare tra loro, gli input sinaptici tra due neuroni possono essere di due tipi: quelli che producono potenziali postsinaptici eccitatori (PPSE) e quelli che inducono potenziali postsinaptici inibitori (PPSI).
I primi provocano una depolarizzazione nella membrana postsinaptica del neurone, aumentando così la probabilità che venga generato un potenziale d'azione, cioè una breve e rapida variazione del potenziale di membrana che si genera solo se quest'ultimo subisce una variazione di polarizzazione superiore ad un determinato valore detto valore di soglia (circa -50mV). I secondi agiscono in modo contrario, iperpolarizzando la membrana del neurone e abbassando così la probabilità che quest'ultimo scarichi un potenziale d'azione.

Le attività elettriche descritte dai tracciati ECoG derivano però quasi totalmente da potenziali postsinaptici (eccitatori e inibitori) e non dai potenziali d'azione, sebbene quest'ultimi siano i

più ampi potenziali generati dai neuroni. Questo è dovuto alla loro breve durata, dell'ordine di 1ms, che non permette la generazione di un'attività elettrica sincronizzata.
Il potenziale post-sinaptico è quello che più contribuisce alla formazione del segnale misurabile sull'elettrodo poiché, nonostante la ridotta ampiezza del potenziale (circa 10 mV), le correnti sinaptiche hanno una durata maggiore (da 10 a 100 ms); questo permette di registrare il segnale ECoG come la somma dei singoli potenziali post-sinaptici sincronizzati di popolazioni di neuroni.

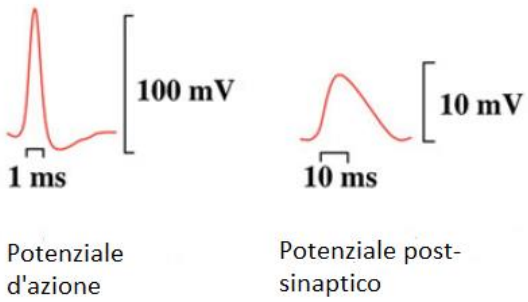

Figura 1.7: potenziale d'azione e potenziale post sinaptico.

1.3.2. Caratteristiche del segnale

L'analisi del segnale elettrocorticografico è fondamentale per lo sviluppo e il corretto funzionamento delle BCIs in quanto differenti componenti di segnale evidenziano processi neurologici differenti. L'analisi del tracciato ECoG mostra come l'attività cerebrale sia caratterizzata da fluttuazioni temporali scandite da fasi ritmiche, ossia da onde caratterizzate da diverso contenuto in frequenza, suddivisibile in specifici intervalli. Le onde che compongono il segnale sono riconoscibili, classificabili e riflettono l'attività sincrona di popolazioni di neuroni che concorrono alla realizzazione del medesimo processo cognitivo. Si possono distinguere cinque onde principali in base all'intervallo di frequenza.

- Il ritmo Delta (δ) caratterizzato da frequenze fino a 4Hz; tipicamente non è presente in condizioni fisiologiche ed è pertanto associato a stati patologici;
- Il ritmo Theta (θ) presenta frequenze nella banda 4 - 8Hz ed è presente durante stati di sonno profondo;
- Il ritmo Alpha (α) è caratterizzato da oscillazioni nella banda 8 - 12Hz; viene registrato ad occhi chiusi in un soggetto sveglio e viene tipicamente correlato a stati di rilassamento mentale;
- Il ritmo Beta (β) è un ritmo molto veloce, con frequenze tra 12 e 30Hz; è associato ad aree attive della corteccia e livelli di coscienza come l'attenzione e la concentrazione;

- Il ritmo Gamma (γ) ha oscillazioni con frequenze superiori a 30Hz fino a 500 Hz ed è legato a stati di elaborazione attiva delle informazioni da parte della corteccia e ad attività cognitive di alto livello.

Grazie alla sua invasività l'ECoG presenta delle caratteristiche vantaggiose rispetto ad EEG tra cui un'ampiezza di segnale maggiore (50-100 μV contro 10-20 μV) e una larghezza di banda più ampia (0–500 Hz contro 0–40 Hz).
La possibilità di accedere ad un intervallo di frequenze più ampio è molto importante poiché è stato dimostrato, attraverso numerosi studi ([9]-[15]), che i segnali a frequenze più elevate contengono informazioni sostanziali relative a compiti cognitivi, motori e linguistici e quindi possono fornire informazioni fondamentali (che diversamente non sono accessibili attraverso EEG) per il controllo delle BCIs.

In particolare, per quanto riguarda l'abilità del linguaggio, è stato dimostrato che le componenti di segnale ad alta frequenza (banda gamma∼70–500Hz) contengono informazioni fondamentali per determinare le rappresentazioni neurali del parlato.
Alcuni studi hanno dimostrato che la banda gamma è correlata alle proprietà acustiche spettro-temporali del parlato nel giro temporale superiore [11], alle caratteristiche fonetiche nel solco temporale superiore [12], e alle caratteristiche articolatorie nella corteccia sensorimotoria [13],[14],[15]. Tali evidenze dimostrano che molti aspetti del linguaggio sono codificati nelle frequenze appartenenti alla banda gamma, e più precisamente nella parte alta di tale intervallo (high-frequency gamma band), del segnale ECoG.

Alcuni aspetti del linguaggio vengono anche codificati alle basse frequenze; ad esempio le frequenze all'interno della banda theta tengono traccia dell'inviluppo acustico del parlato, sono correlate alla frequenza sillabica e possono discriminare le frasi parlate [16],[17],[18]. Inoltre, i ritmi theta hanno mostrato significativi cambiamenti di potenza nell'area di Broca e nelle aree temporali del linguaggio durante un compito di produzione linguistica e hanno mostrato interazioni con la banda ad alta frequenza, attraverso l'accoppiamento ampiezza-ampiezza e fase-ampiezza [19].

Pertanto ECoG, grazie alla sua caratteristica di poter registrare un'ampia larghezza di banda, rappresenta una promettente tecnica di registrazione per studiare e decodificare l'attività cerebrale associata alla produzione linguistica.

1.4 Metodi di comunicazione indiretta e diretta

La ricerca nell'ambito delle speech BCIs si pone l'obiettivo di creare un collegamento tra l'attività neurale e dispositivi esterni. Per realizzare questo collegamento si possono utilizzare due differenti metodi di comunicazione, uno indiretto l'altro diretto.

Il termine comunicazione vocale indiretta si riferisce a dispositivi BCI comunicativi che utilizzano l'attività neurale non direttamente correlata all'atto della produzione vocale come segnale di controllo per sistemi BCIs. Nelle BCIs che utilizzano questo meccanismo di comunicazione è necessario un passaggio intermedio per tradurre l'attività neurale che rappresenta informazioni non specifiche in un dominio vocale.

I principali sistemi di comunicazione indiretta assumono la forma di paradigmi di selezione di lettere e parole, in particolare negli ultimi trent'anni sono stati sviluppati metodi molto utilizzati che prevedono la selezione di lettere mediante il segnale EEG ([20-25]). In questi sistemi l'elettroencefalogramma registra dei potenziali elettrici evento correlati, cioè delle risposte cerebrali rilevate in corrispondenza di un evento specifico, che sono utilizzati per selezionare delle lettere e quindi per produrre delle parole. Alcune di queste tecniche utilizzano ad esempio il potenziale evento-correlato P300 (ERP) e i potenziali evocati visivi allo stato stazionario (SSVEP).

Un'onda P300 è un potenziale evento correlato che si verifica nel tracciato EEG ogni volta che l'utente rileva un evento raro o significativo tra una serie di altri eventi più frequenti. L'onda P300 può essere suscitata in modo affidabile con paradigmi relativamente semplici, tra cui lettere lampeggianti su una tastiera virtuale, e può essere sfruttata come segnale di comando per una BCI.
Il segnale P300 fu utilizzato per la prima volta come mezzo per selezionare lettere da una tastiera virtuale da Farwell e Donchin [22] che nel 1988 realizzarono il primo speller P300 cioè un sistema di speech BCI in cui il segnale di controllo dell'interfaccia è l'onda P300. In questo metodo l'utente guardava uno schermo in cui era raffigurata una matrice 6x6 contenente lettere e simboli, focalizzando l'attenzione sull'elemento desiderato mentre ogni 125 ms una riga o colonna della matrice veniva illuminata per 100 ms. L'illuminazione della riga o della colonna contenente il simbolo desiderato provocava una risposta evocata P300, permettendo così all'interfaccia di selezionare tale lettera.
Nel corso degli ultimi anni lo speller P300 ha subito modifiche e miglioramenti e al giorno d'oggi è stato dimostrato che questa tecnologia può essere utilizzata da persone affette da SLA raggiungendo un'accuratezza media del 95 % e la produzione di 6-12 caratteri al minuto [26].

Il potenziale evocato visivo allo stato stazionario (SSVEP) è una risposta cerebrale evocata che può essere rilevata nell'EEG in risposta a uno stimolo visivo che oscilla ad una frequenza fissa. In un'applicazione BCI, possono essere presentati diversi stimoli intermittenti, ciascuno con la

propria frequenza e / o fase; l'utente può quindi selezionare lo stimolo target semplicemente focalizzando la propria attenzione su di esso. Una delle realizzazioni più recenti di SSVEP-BCIs utilizza uno schermo in cui è rappresentata una matrice 5x8 con 40 stimoli individuali, inclusi caratteri e cifre, (Nakanishi et al., 2017 [27]), in cui ciascuno lo stimolo può essere selezionato in un solo passaggio.

È importante notare che, sebbene le frasi costruite dalla selezione ortografica delle singole lettere possano essere espresse ad alta voce usando dei sistemi di sintesi vocale, nessuno di questi sistemi decodifica un discorso parlato. Esistono due principali svantaggi di questo tipo di sistema di comunicazione indiretta. In primo luogo, sebbene spesso molto preciso, il tasso di selezione delle lettere può essere lento quanto una parola al minuto, limitando la capacità di un utente di conversare fluentemente in tempo reale. In secondo luogo, questi sistemi sono "generici" in quanto possono essere utilizzati per qualsiasi attività di selezione; pertanto ignorano delle informazioni neurologiche potenzialmente preziose nonché i vincoli relativi al linguaggio che possono migliorare le capacità di comunicazione. Infine per riuscire ad utilizzare correttamente questi sistemi l'utente necessita di essere allenato a selezionare le lettere con l'attenzione visiva e questo richiede spesso un elevato sforzo cognitivo che può affaticare il paziente poiché la comunicazione non avviene in modo naturale.

Studi recenti hanno cercato di affrontare questi problemi e di rendere la produzione vocale attraverso le BCIs più naturale e fluente grazie a metodi di comunicazione diretta.

La comunicazione vocale diretta si riferisce ai metodi BCI che utilizzano l'attività neurale correlata in modo innato all'atto della produzione vocale. Questi metodi utilizzano l'attività neurologica presente durante la produzione del linguaggio, o anche solo durante una fase in cui l'utente immagina di parlare, per controllare il sistema di BCI. Per i metodi diretti non è richiesta alcuna mappatura intermedia tra stati cognitivi e output vocale, perché per controllare il sistema di BCI viene utilizzata l'attività neurale direttamente correlata alla produzione vocale; ciò consente di aumentare considerevolmente la velocità di produzione delle parole per tendere ad una produzione vocale fluente in tempo reale. Inoltre la previsione diretta consente ai ricercatori di sfruttare i meccanismi neurali noti per la produzione del parlato e altre caratteristiche rilevanti del parlato negli algoritmi di decodifica e classificazione per la protesi del parlato.

I metodi di comunicazione diretta si differenziano sulla base delle caratteristiche del parlato che si intende decodificare a partire dall'attività neurale, alcuni ricostruiscono il discorso a partire dalle caratteristiche spettro-temporali mentre altri utilizzano le rappresentazioni articolatorie del parlato. Per questi motivi è necessario discutere le possibili strategie di decodifica del segnale ECoG e le caratteristiche del parlato che si vogliono estrarre da esso; per decodificare tali caratteristiche in modo diretto si utilizzano delle reti neurali, il cui funzionamento verrà discusso in seguito nel capitolo 2.

1.5 Strategie di decodifica

La sintesi del parlato può essere ottenuta attraverso diversi metodi, classificati in base al tipo di parametri che vengono decodificati dall'attività cerebrale. Come illustrato in figura 8, le strategie di decodifica sono tre, ciascuna corrispondente a una diversa rappresentazione del parlato: fonetica, acustica e articolatoria. Ognuna di queste rappresentazioni è codificata in modo più specifico in alcune aree cerebrali rispetto ad altre. Ad esempio il contenuto acustico del linguaggio è codificato più dettagliatamente nelle aree uditive temporali, mentre le caratteristiche articolatorie del parlato sono codificate in modo più specifico nella corteccia sensorimotoria. In base alla scelta della strategia di decodifica possono essere considerati diversi metodi di sintesi vocale.

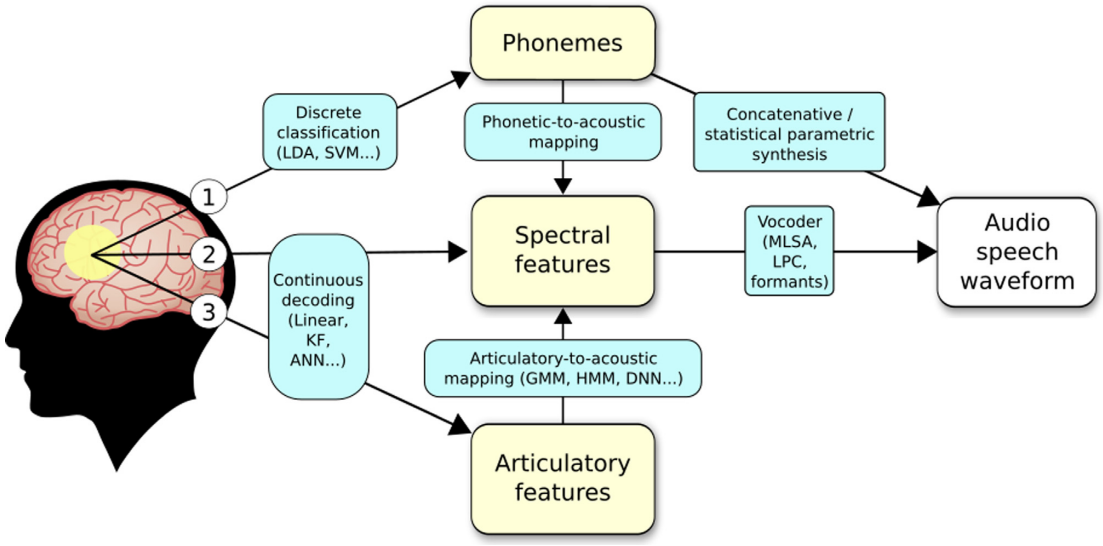

Figura 1.8: A partire dall'attività cerebrale possono essere decodificate tre rappresentazioni del parlato (fonetica, acustica o articolatoria), ognuna delle quali implica l'utilizzo di una specifica tecnica di sintesi vocale per l'implementazione di una speech BCI.

La prima categoria di sintesi vocale consiste nel concatenare singoli fonemi o singole parole. Un sistema BCI basato su tale sintesi consiste nel decodificare gli elementi vocali discreti a partire dall'attività cerebrale, ad esempio usando una classificazione discreta delle caratteristiche neuronali, per poi convertire la sequenza di fonemi decodificati in discorsi audio o rappresentazioni testuali.

La principale implementazione di tale metodo di decodifica è stata ottenuta da Herff e colleghi [28] che nel 2015 hanno realizzato l'interfaccia brain-to-text cioè un sistema di speech BCI che decodifica il discorso continuo attraverso delle rappresentazioni testuali di sequenze di singoli fonemi. Nello studio di Herff et al. l'attività cerebrale, misurata attraverso ECoG, è stata registrata contemporaneamente ed in modo sincronizzato alla produzione vocale di alcune parole. Successivamente, attraverso dei modelli matematici detti automatic speech recognition

(ASR), l'onda acustica prodotta durante il parlato è stata suddivisa in sequenze che sono state classificate in base ai fonemi associati al suono; poiché l'attività cerebrale era stata registrata in modo sincrono alla produzione vocale si è riusciti ad associare a ciascun fonema la corrispondente parte di segnale ECoG. Grazie agli ASR e ad un modello statistico probabilistico che determina la sequenza di fonemi e parole più adatta alla sequenza di attività neurale osservata, l'interfaccia brain-to-text è in grado di determinare una rappresentazione testuale dell'attività neuronale, andando a ricostruire le parole attraverso la concatenazione di singoli fonemi decodificati dall'attività cerebrale.

Una seconda categoria di sintesi vocale si basa sulle caratteristiche spettrali del parlato. Un sistema BCI basato su questo approccio converte i segnali cerebrali in una rappresentazione spettrale del discorso, attraverso delle reti neurali, e successivamente converte le caratteristiche acustiche prima decodificate in una forma d'onda del parlato utilizzando uno strumento che è in grado di codificare un segnale audio cioè un vocoder (voice encoder).

Le rappresentazioni acustiche del parlato utilizzate per decodificare l'attività cerebrale sono molteplici; in particolare un importante studio condotto in questo ambito è il lavoro di Guenther e colleghi [29].
Guenther et al. hanno implementato un sintetizzatore vocale in tempo reale che rileva l'attività cerebrale attraverso microelettrodi intracorticali e decodifica le rappresentazioni acustiche del parlato attraverso le formants cioè concentrazioni di energia nello spettro del parlato attorno a frequenze particolari. Questo studio è stato uno dei primi lavori che ha dimostrato il potenziale dell'utilizzo di tecniche di registrazione invasive per la realizzazione di una speech BCI; tuttavia Guenther et al. hanno realizzato un sistema che è in grado di riprodurre solamente la sintesi delle vocali e non delle consonanti poiché le formants sono adatte solo per la sintesi di vocali.

Recentemente altri studi hanno apportato dei miglioramenti implementando sistemi BCI per la sintesi vocale utilizzando differenti rappresentazioni del contenuto spettrale del parlato. Due dei più recenti ed innovativi esempi sono trattati nei capitoli 3 e 4 della tesi e sono stati implementati da Angrick [2], che ha utilizzato le caratteristiche spettrali dello spettrogramma logaritmico in scala mel (logMel spectrogram) (vedi capitolo 4), e Akbari [1] che ha utilizzato una rappresentazione acustica basata sui parametri di sintesi del discorso (vedi capitolo 3).

Infine, la terza categoria di sintesi vocale è la sintesi articolatoria in cui la strategia di decodifica dell'attività cerebrale si basa sulle caratteristiche articolatorie del parlato cioè le posizioni variabili nel tempo dei principali organi del tratto vocale quali: la lingua, le labbra, la mascella e la laringe. Un sistema BCI basato su tale sintesi consiste nel decodificare i movimenti degli articolatori del parlato dall'attività cerebrale, nel decodificare le rappresentazioni acustiche del parlato a partire dalle caratteristiche articolatorie precedentemente decodificate e infine nel

ricostruire la forma d'onda del parlato a partire dalla rappresentazione acustica. Un approccio a doppia decodifica come quello appena descritto è stato implementato da Anumanchipalli et al. [3] che sono riusciti a realizzare un sintetizzatore vocale in tempo reale che registra l'attività cerebrale attraverso ECoG. Questo approccio di sintesi vocale, che viene descritto dettagliatamente nel capitolo 5, si è rivelato molto promettente negli ultimi anni e il lavoro di Anumanchipalli et al. rappresenta una nuova frontiera nell' ambito delle speech BCI.

Capitolo 2

2. Deep learning e reti neurali

Nel capitolo precedente sono state descritte le principali caratteristiche di un sistema di speech BCI considerando la tecnica di registrazione del segnale cerebrale, il metodo di comunicazione e la strategia di decodifica utilizza dal sistema. Un ulteriore importante caratteristica di un sistema di speech BCI è il metodo di decodifica utilizzato per estrarre le rappresentazioni del parlato a partire dall'attività cerebrale, poiché in base al metodo di decodifica utilizzato cambiano radicalmente le prestazioni del sistema di sintesi vocale.
I tre approcci di speech BCI descritti nella tesi decodificano il parlato direttamente a partire dall'attività cerebrale attraverso delle reti neurali profonde, per questo motivo prima di analizzare nel dettaglio i tre approcci oggetto della tesi è necessario introdurre il concetto di deep learning e le reti neurali.

Nel seguente capitolo viene brevemente descritto il concetto di deep learning, in italiano apprendimento profondo (termine raramente utilizzato nella prassi), e vengono definite le reti neurali e il loro funzionamento. In particolare nella parte finale del capitolo vengono descritte con maggiore attenzione tre tipologie di reti neurali: locally connected e fully connected neural networks, convolutional neural networks e recurrent neural networks; tali reti neurali sono rispettivamente associate ai tre sistemi di speech BCI [1], [2], [3] che verranno trattati nei capitoli successivi.

2.1 Machine learning e deep learning

Negli ultimi decenni le tecniche di machine learning sono state applicate in molti settori determinando una grande influenza sulla nostra vita quotidiana, attraverso ad esempio sistemi di guida autonoma, visione artificiale e ricerche sul web efficienti [30-32]. Le tecniche di machine learning sono l'insieme delle tecniche di analisi di dati che hanno la capacità di apprendere dall'esperienza, cioè di migliorare di volta in volta le proprie prestazioni, attraverso l'utilizzo di metodi computazionali. Sebbene le tecniche di apprendimento automatico vengano applicate con successo in molti settori, le loro performance nell'elaborazione di segnali biologici, come ECoG, non sono soddisfacenti. Ispirandosi alle profonde strutture gerarchiche della percezione del parlato e della vista è stata realizzata all'inizio del 2000 una nuova sottocategoria dell'apprendimento automatizzato chiamata deep learning.

Gli algoritmi di deep learning si basano su tecniche di representation learning, ovvero algoritmi che consentono ad un computer di apprendere automaticamente, a partire dai dati in input "grezzi", le rappresentazioni necessarie per risolvere un problema di learning (come ad esempio un compito di classificazione o regressione).
Le tecniche di deep learning sono caratterizzate dalla presenza di più livelli di rappresentazione che elaborano i dati a partire dell'input "grezzo" fino ad ottenere rappresentazioni di livello più astratto.

In tutti gli algoritmi di deep learning in base al tipo di problema definito, ai dati a disposizione e a quello che si vuole ottenere si possono distinguere due tecniche di apprendimento: supervisionato e non supervisionato.
Nell'apprendimento supervisionato ad ogni elemento del data set che l'algoritmo utilizza per apprendere è associato una label (etichetta). L'algoritmo apprende rilevando le caratteristiche comuni di ogni label ed esegue il task. Questo tipo di apprendimento viene chiamato supervisionato perché all' algoritmo vengono fornite alcune soluzioni del task all'interno del data set, in questo modo la macchina ha a disposizione degli esempi pratici che consentono di supervisionare l'apprendimento.
Si sviluppa in questo modo un modello predittivo basato sia sui dati di ingresso che sulle risposte, tale modello verrà poi applicato a dati nuovi per generare previsioni. I problemi di learning che vengono risolti attraverso l'apprendimento supervisionato sono la classificazione e la regressione.
Nell'apprendimento non supervisionato il data set che l'algoritmo utilizza per apprendere non è etichettato, non sono quindi note le risposte al problema. L'algoritmo apprende tramite l'individuazione di strutture intrinseche dei dati o pattern nascosti. Una tecnica utilizzata in questo tipo di apprendimento è il clustering e in questo caso l'algoritmo apprende individuando gruppi (cluster) che hanno caratteristiche simili.

2.2 Reti neurali

Le reti neurali artificiali sono modelli matematici che implementano algoritmi di machine learning, la loro struttura è ispirata all'organizzazione delle reti neurali biologiche poiché sono composte da unità di calcolo elementari, che ricoprono un ruolo simile al neurone, interconnesse tra loro su più livelli.
Generalmente l'architettura di una rete neurale si articola in tre livelli:

- Input layer: costituisce il livello di ingresso dei dati ed è solitamente formato da linee di ingresso multiple.
- Hidden layer: riceve direttamente i dati dall' input layer e li elabora. A seconda del tipo di rete che si intende realizzare è possibile avere più hidden layers, in particolare una rete caratterizzata da molteplici hidden layers prende il nome di deep neural network.

- Output layer: riceve i dati dagli hidden layers e tali dati sono il risultato dell'apprendimento della rete stessa, non vengono determinati manualmente dall'operatore che programma la rete, ma sono ottenuti automaticamente attraverso gli algoritmi di deep learning e il loro addestramento.

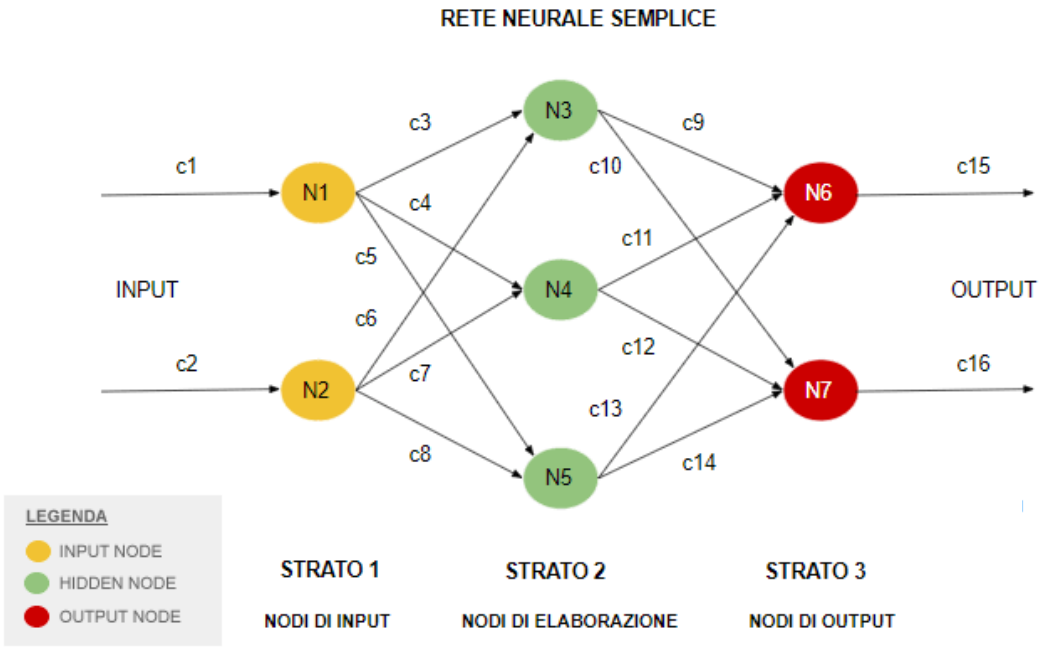

Figura 2.1: architettura di una rete neurale semplice costituita da tre livelli, input, hidden e output con due ingressi c1 e c2 e due uscite c15 e c16.

2.2.1 Unità funzionale: neurone artificiale

Le singole unità di elaborazione di una rete neurale sono dei neuroni artificiali cioè modelli matematici computazionali che hanno lo scopo di processare informazioni ed il cui funzionamento si ispira al comportamento dei neuroni biologici.
Un neurone biologico riceve in ingresso i segnali elettrici attraverso i suoi dendriti, integra le informazioni nel suo corpo centrale, detto soma, e genera un'uscita che viene trasmessa agli altri neuroni tramite una sinapsi. Allo stesso modo un neurone artificiale riceve i dati in input attraverso numerosi ingressi, ciascuno caratterizzato da un peso, ed integra le informazioni sommando il contributo degli ingressi moltiplicati per il relativo peso. Successivamente il neurone artificiale genera un'uscita, che dipende da un'opportuna funzione di attivazione, e invia l'output generato ad altri neuroni artificiali.

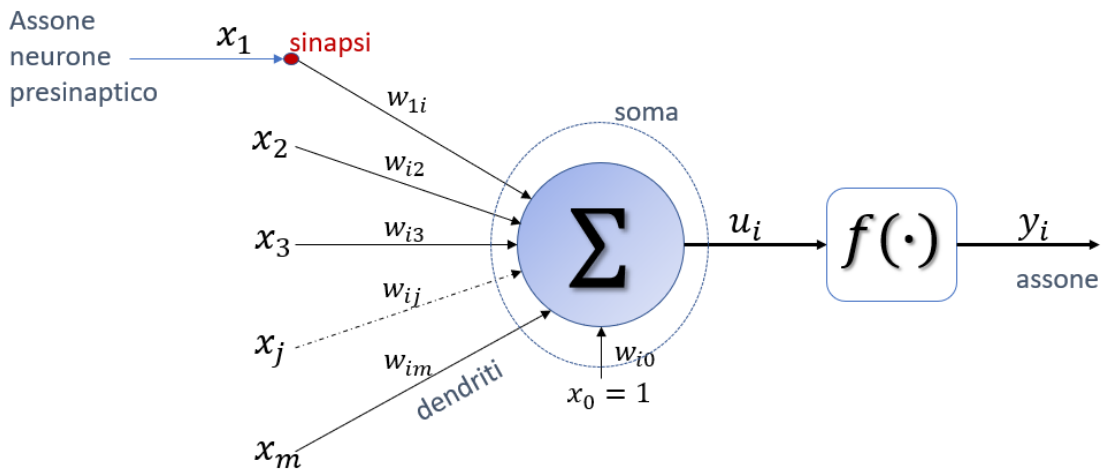

Figura 2.2: Struttura di un neurone artificiale

Come mostrato in figura 2.2 gli elementi presenti nel modello del neurone artificiale sono:

- $\boldsymbol{x_j}$: sono gli m ingressi che riceve il neurone post-sinaptico *i-esimo* dai neuroni presinaptici *j-esimi*.
- $\boldsymbol{w_{ij}}$: sono i pesi (weights) delle sinapsi. In base al loro valore, esse stabiliscono l'efficacia della connessione fra il neurone presinaptico *j-esimo* e il neurone post-sinaptico *i-esimo*.
- $\boldsymbol{w_{i0}}$: valore di bias. É un peso che si considera collegato ad un ulteriore ingresso fittizio $\boldsymbol{x_0}$ di valore 1. Serve ad impostare il punto di lavoro ottimale del neurone.
- $\boldsymbol{u_i}$: è il livello di eccitazione globale del neurone: $u_i = \sum_{j=1}^{m} w_{ij} x_j + w_{i0}$
- $\boldsymbol{f}(\cdot)$: funzione di attivazione; determina il comportamento di uscita: $\hat{y}_i = f(u_i)$

L'uscita del neurone sarà $\hat{y}_i$ cioè il valore predetto che è approssimazione del valore obiettivo y_i. L'apprendimento della rete consiste nell'aggiustamento dei pesi w_{ij} poiché cambiando i valori dei pesi si ottengono delle risposte diverse ad ogni input.

Nella progettazione di una rete neurale è determinante la scelta della funzione di attivazione. Esistono diverse tipologie di funzione di attivazione, come viene mostrato in figura 2.3, e generalmente sono tutte funzioni non lineari, che è una caratteristica necessaria per consentire la risoluzione di task con dati in input molto complessi. Inoltre, tutte le funzioni sono continue e differenziali poiché questa è una condizione necessaria per l'applicazione dell'algoritmo di back propagation cioè quell'algoritmo che consente la capacità di apprendimento della rete.

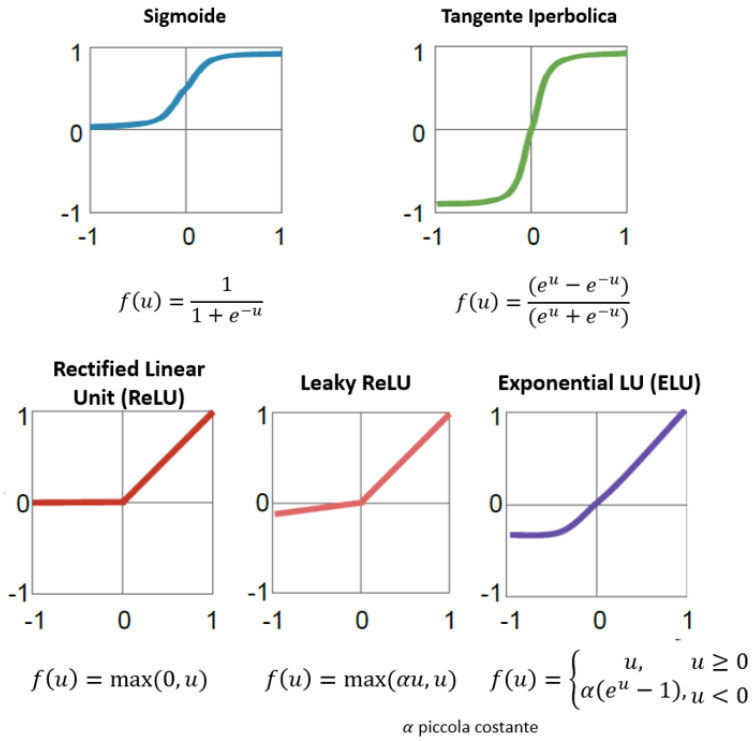

Figura 2.3: Principali funzioni di attivazione utilizzate nelle architetture di deep learning.

2.2.2 Tipologie di reti neurali

Le reti neurali sono composte da gruppi di neuroni artificiali organizzati in livelli, e come detto nel paragrafo precedente generalmente sono presenti: un livello di input, un livello di output, e uno o più livelli intermedi (hidden). Le reti che sono costituite da più hidden layer vengono chiamate deep neural network e da questo punto in poi faremo sempre riferimento a questa tipologia poiché è quella che viene utilizzata da tutti e tre gli approcci di speech BCI che verranno descritti nei capitoli successivi.

Le reti neurali vengono classificate, in base al modo in cui avviene il flusso di informazioni tra i layer, in due categorie: feedforward e feedback.

- Feedforward: nelle reti neurali feedforward il flusso di informazioni avviene in una sola direzione, le connessioni tra i nodi collegano neuroni di un livello con quelli del livello successivo e il flusso di dati procede sempre in avanti. Le reti neurali feedforward sono ampiamente utilizzate ed in particolare nei paragrafi successivi tratteremo le Locally Connected e le Fully Connected neural networks che sono delle Deep neural network caratterizzate da due differenti tipi di connessioni. Inoltre in un ulteriore paragrafo approfondiremo anche le Convolutional neural networks che sono una sottocategoria delle DNN e si contraddistinguono per l'utilizzo dell'operazione di convoluzione tra i layers.

- Feedback: nelle reti neurali ricorrenti sono previste connessioni in feedback che possono avvenire sia verso neuroni dello stesso livello, sia verso neuroni del livello precedente. La presenza di connessioni in retroazione complica notevolmente il flusso delle informazioni e di conseguenza anche l'addestramento della rete, poiché è necessario considerare il comportamento in più istanti temporali. Le reti neurali ricorrenti sono dotate di una capacità di memoria (di breve termine) che al tempo t rende disponibile l'informazione processata a $t-1$, $t-2$, ecc.

2.3 Locally connected and Fully connected neural networks

Le reti neurali localmente (locally) e completamente (fully) connesse sono delle deep neural networks e si differenziano per il modo in cui sono organizzate le connessioni tra due layer all'interno della rete.

Nelle locally connected neural networks (LCN) ogni neurone appartenente ad un layer riceve le connessioni soltanto da alcuni dei neuroni appartenenti al layer precedente. In particolare, come mostrato in figura 2.4 un neurone riceve le connessioni da gruppi di neuroni che sono spazialmente vicini nel layer precedente, per questo motivo si dice che i layer sono connessi localmente. Attraverso questo tipo di connessioni la complessità della rete si riduce ed è sufficiente una fase di traning minore (meno parametri da allenare).
È importante sottolineare che, a differenza delle reti neurali convoluzionali (vedi paragrafo 2.4) che sono anche esse localmente connesse, nelle LCN i pesi utilizzati tra i layers sono tutti diversi quindi non c'è la condivisione dei pesi.

Nelle fully connected neural networks (FCN) i neuroni appartenenti a due layer adiacenti sono completamente collegati tra loro, questo significa che ogni neurone del i-esimo layer è collegato a tutti i neuroni del j-esimo layer.
Ad oggi le FCN sono la tipologia di DNN maggiormente utilizzata tuttavia presentano alcuni svantaggi soprattutto durante la fase di addestramento che è molto complicata a causa dell'elevato numero di pesi sinaptici da addestrare.

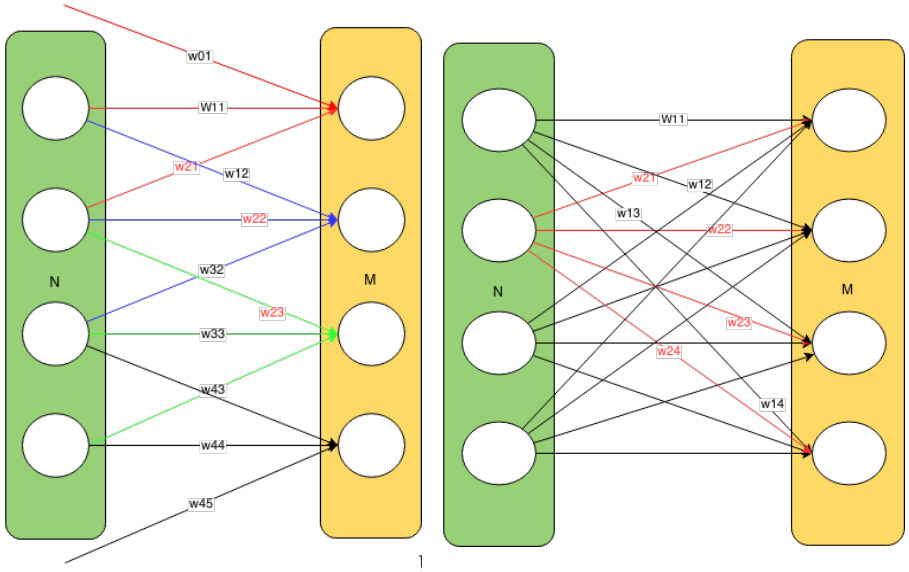

Figura 2.4: Differenza tra le connessioni, a sinistra troviamo una LCN mentre a destra troviamo una FCN.

2.4 Convolutional neural networks

Le reti neurali convoluzionali (ConvNet o Convolutional Neural Networks CNNs) sono una particolare tipologia di DNN specializzate nel processare dati 2D o 3D.
A differenza delle normali DNN in cui viene eseguita la moltiplicazione tra matrici (che costituiscono i dati della rete), nelle CNN viene effettuata l'operazione di convoluzione tra una matrice (solitamente multidimensionale) e un filtro detto kernel di dimensione minore della matrice. Durante l'operazione di convoluzione il filtro è fatto scorrere sulle diverse posizioni della matrice di input e per ogni posizione viene generato un valore di output eseguendo il prodotto scalare tra il filtro e la porzione dell'input coperta dal filtro in quel momento.

Le CNN si differenziano dalle normali DNN per due principali caratteristiche: processing dei dati a livello locale e condivisione dei pesi (weights).
La capacità di effettuare il processing dei dati a livello locale è dovuta al fatto che il kernel esegue la convoluzione a livello di gruppi più piccoli di dati in input rispetto alla matrice intera. Ogni neurone esegue quindi un'elaborazione locale e i neuroni di un livello sono connessi solo localmente ai neuroni del livello precedente; questo implica una forte riduzione del numero di connessioni che rende la rete più efficiente.
L'altra caratteristica che contraddistingue le CNN è la condivisione dei pesi. Il fatto che i pesi delle connessioni tra neuroni siano condivisi a gruppi implica una riduzione del numero di pesi. I pesi sono condivisi a livello di feature maps, ogni feature map è il risultato dell'operazione di filtraggio a livello locale dell'input; poiché i neuroni di una stessa feature map condividono i pesi, essi processano i dati del volume di input allo stesso modo.

La combinazione di processing locale e connessioni condivise consente ai neuroni di una stessa feature map di processare nello stesso modo porzioni diverse dell'input. Attraverso questa modalità di elaborazione gli strati convoluzionali della rete sono in grado di estrarre le features del segnale in input in modo efficiente sfruttando il fatto che in alcuni tipi di segnali, all'interno dello stesso volume di dati, regioni diverse contengono lo stesso tipo di informazioni. Per questo motivo la convoluzione è molto utile per estrarre features dalle immagini e dal discorso parlato, poiché in questi tipi di dati ci sono numerosi pattern che si ripetono.

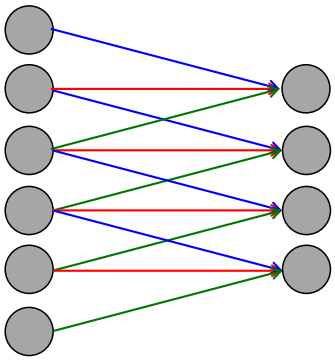

Figura 2.5: Connessioni localizzate e condivisione di pesi: ciascuno dei 4 neuroni a destra è connesso solo a 3 neuroni del livello precedente. I pesi sono condivisi (stesso colore stesso peso).

L'architettura di una CNN non si compone solamente di convolutional layer ma è costituita anche da layer che eseguono l'operazione di pooling. Il livello di pooling effettua un'aggregazione delle informazioni nel volume di input, generando feature maps di dimensione inferiore. L'obiettivo di questo layer è conferire invarianza rispetto a semplici trasformazioni dell'input mantenendo al tempo stesso le informazioni significative ai fini della discriminazione delle features.

Una rete neurale convoluzionale si compone quindi di un livello di input, un convolutional layer, e un pooling layer. Questa composizione di layers viene ripetuta più e più volte e solitamente la rete termina con un fully connected layer in cui tutti gli elementi in ingresso sono collegati con tutti i neuroni dello strato.

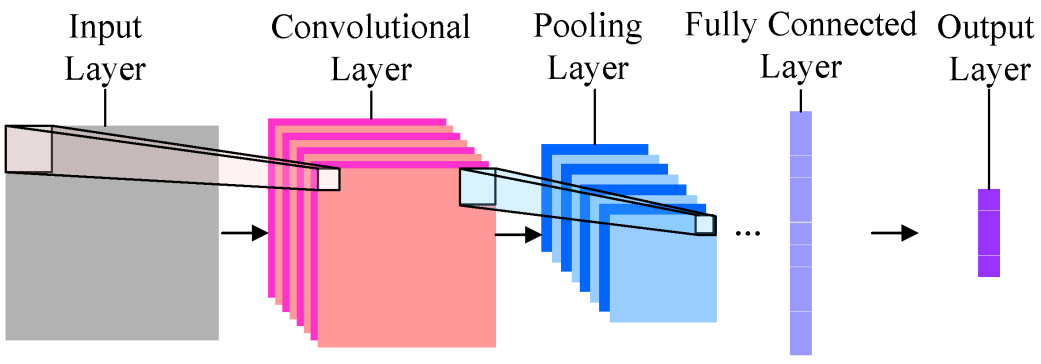

Figura 2.6: rappresentazione dei layers presenti all'interno di una CNN.

2.5 Recurrent neural networks

Le reti neurali ricorrenti sono caratterizzate da connessioni feedback tra i layers della rete e tali connessioni posso essere riferite ai livelli precedenti oppure possono essere connessioni all'interno dello stesso livello.

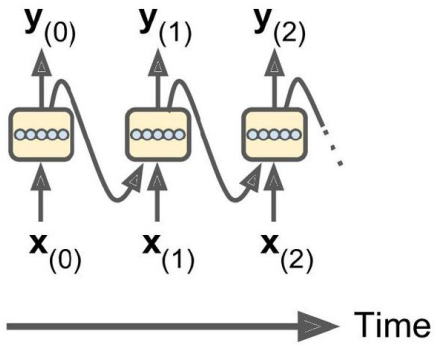

Figura 2.7: connessioni di una RNN

Nelle reti neurali ricorrenti, come mostrato nella figura 2.7, ad ogni istante di tempo t ogni nodo della rete riceve l'input x(t) ma anche l'output del nodo precedete all'istante t -1, y(t-1). Questa struttura consente alla rete di basare il meccanismo di apprendimento sulla "storia passata" ovvero su tutti gli elementi di una sequenza e sulla loro posizione reciproca.

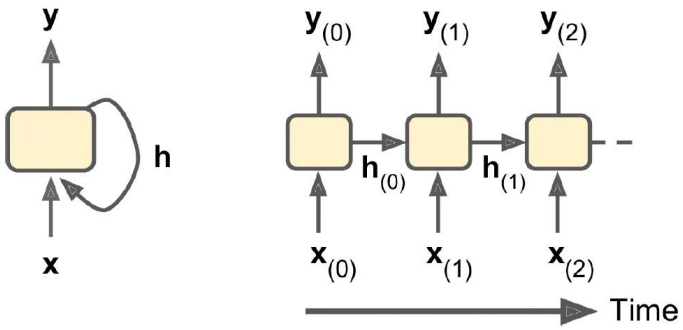

Figura 2.8: rappresentazione di una cella di una RNN

Una cella è una parte della rete ricorrente che conserva uno stato (o memoria) interno h(t) per ogni istante temporale.
Ogni cella è costituita da un numero prefissato di neuroni e può essere considerata come un layer della rete. Le celle così costituite hanno difficoltà a ricordare e ad utilizzare input che provengono da istanti di tempo lontani poichè la memoria dei primi input tende a svanire.

Per risolvere questo problema e facilitare la convergenza in applicazioni complesse, sono state proposte celle più evolute dotate di un effetto memoria a lungo termine. Una delle reti neurali che utilizza questo tipo di celle è la Long Short Term Memory network (LSTM) che è

caratterizzata da una capacità di memoria. Grazie a questa caratteristica negli ultimi anni le LSTM si sono dimostrate più indicate delle tradizionali RNN per la gestione di sequenze come per esempio audio, video e frasi in linguaggio naturale.
In particolare le LSTM sono state utilizzate anche da Anumanchipalli et al [3] nel loro studio "speech synthesis of neural decoding of spoken sentences" come viene spiegato nel capitolo 5.

Capitolo 3

Nei capitoli precedenti abbiamo descritto le caratteristiche delle differenti tipologie di BCI, il concetto deep learning e le reti neurali; tali concetti sono di fondamentale importanza per la comprensione delle caratteristiche dei tre differenti sistemi di sintesi vocale che analizzeremo nei prossimi capitoli.

I tre approcci di speech BCI oggetto della tesi si distinguono dai sistemi BCI precedentemente sviluppati poiché riescono a riprodurre un parlato fluente ed intelligibile. I risultati ottenuti da questi sistemi di sintesi vocale sono stati resi possibili grazie all'utilizzo dell'ECoG come tecnica di rilevazione dell'attività cerebrale e grazie ai progressi nell'ambito del deep learning come metodo di decodifica.
L'utilizzo dell'ECoG ha permesso di rilevare con più precisione l'attività e i processi cerebrali coinvolti nella produzione lessicale, in quanto questa tecnica di rilevazione possiede un'elevata risoluzione spaziale e temporale. Inoltre l'utilizzo di algoritmi di deep learning e i recenti progressi che sono stati raggiunti in questo campo hanno permesso di decodificare le rappresentazioni acustiche e articolatorie del parlato con una maggiore accuratezza permettendo una migliore ricostruzione del discorso.

Nei tre capitoli seguenti verranno descritti singolarmente i tre approcci di sintesi del parlato; la trattazione si concentrerà sulle differenti tecniche di decodifica utilizzate, sulle differenti architetture di reti neurali e sul diverso posizionamento degli elettrodi per ECoG sulle aree della corteccia. Infine per ogni metodo descritto verranno riportati i risultati ottenuti in termine di accuratezza della sintesi del parlato.

3. Sintesi del parlato attraverso la ricostruzione della rappresentazione acustica nella corteccia uditiva

3.1 Introduzione

Nel presente capitolo viene descritto il lavoro di Akbari e colleghi [1] che hanno sviluppato un sistema BCI in grado di produrre un parlato intelligibile attraverso la ricostruzione dello stimolo uditivo a partire dall'attività cerebrale codificata nella corteccia uditiva. Per realizzare questo sistema di sintesi vocale Akbari et al. hanno utilizzato una rappresentazione acustica basata sui parametri di uno speech vocoder (vedi paragrafo 3.4) e hanno implementato un sistema di decodifica costituito da una deep neural network a due stadi che utilizza delle locally connected e fully connected neural networks.

Il sistema uditivo umano è molto complesso e possiede la capacità di rilevare, separare e riconoscere le parole.
L'analisi dello stimolo uditivo inizia a livello della coclea e prosegue fino alla corteccia uditiva primaria, dove viene determinata una rappresentazione spettro-temporale dello stimolo; successivamente l'elaborazione fonetica e lessicale dello stimolo proseguono nei livelli superiori della corteccia.
La prima analisi dello stimolo uditivo, che produce una rappresentazione in tempo e in frequenza, fornisce una rappresentazione fedele delle proprietà spettro temporali della forma d'onda dello stimolo uditivo che l'utente ha ascoltato. Quindi, per poter riprodurre lo stimolo uditivo, è necessario riuscire a ricostruire tale rappresentazione spettro-temporale dello stimolo.

La tecnica di ricostruzione dello stimolo uditivo consiste nel rilevare l'attività neuronale dalla corteccia uditiva, in corrispondenza della presentazione di stimoli uditivi, e nel ricostruire la migliore approssimazione della rappresentazione acustica dello stimolo.
Tale tecnica è stata utilizzata da numerosi studi per determinare le rappresentazioni codificate nella corteccia uditiva [36-39], tuttavia negli ultimi anni la ricostruzione dello stimolo uditivo a partire dall'attività cerebrale è stata utilizzata anche per realizzare dei sistemi di speech BCI che fossero in grado di ripristinare il parlato per pazienti gravemente paralizzati [11].
Tra gli studi più recenti ed innovativi che utilizzano la ricostruzione dello stimolo acustico per realizzare un sistema di speech BCI che registra l'attività cerebrale dalla corteccia uditiva troviamo lo studio di Akbari e colleghi [1].

Il lavoro realizzato da Akbari et al. si distingue per la qualità e l'accuratezza del discorso sintetizzato poiché, sebbene altri studi in precedenza avevano dimostrato la fattibilità di realizzare un sistema di speech BCI utilizzando la ricostruzione dello stimolo uditivo, nessuno era riuscito a riprodurre un parlato intelligibile.
Grazie a recenti sviluppi tecnologici nell'ambito del deep learning e della sintesi vocale Akbari et al. sono riusciti ad implementare un sistema di speech BCI in grado di ricostruire lo stimolo uditivo e di riprodurlo realizzando un discorso fluente e intelligibile.

3.2 Registrazione dell'attività cerebrale

Akbari e colleghi hanno effettuato le registrazioni dell'attività neurale attraverso ECoG da cinque pazienti che indossavano elettrodi per elettrocorticografia poiché sottoposti ad un trattamento per l'epilessia.
Durante la fase di acquisizione gli elettrodi, posizionati su STG (giro temporale superiore) e su HG (giro di Heschl), registrano l'attività neurale mentre i pazienti ascoltano degli stimoli uditivi, parole, frasi e cifre, prodotti da alcuni oratori.

Per ricostruire lo stimolo uditivo sono state selezionate due bande di frequenza dell'attività neurale: banda a bassa frequenza (0-50 Hz) e banda ad alta frequenza (70-150 Hz).
La frequenza di acquisizione del segnale influenza le caratteristiche dello stimolo ricostruito, per questo motivo Akbari et al. hanno analizzato la qualità e l'accuratezza del parlato ricostruito in funzione della banda di frequenza utilizzata per decodificare le caratteristiche del segnale (vedi paragrafo 3.5).

Dopo aver registrato l'attività neurale attraverso ECoG, Akbari e colleghi hanno utilizzato due differenti modelli di regressione (vedi paragrafo 3.3), lineare e non lineare, per estrarre la rappresentazione acustica dello stimolo.

La strategia di decodifica del segnale, cioè il tipo di rappresentazione acustica che si intende decodificare, influenza molto il risultato della ricostruzione dello stimolo, per questo motivo Akbari et al. studiano la qualità del discorso sintetizzato decodificando due differenti rappresentazioni acustiche: lo spettrogramma uditivo e una rappresentazione costituita dall'insieme di parametri per uno speech vocoder (vedi paragrafo 3.4).

Akbari e colleghi hanno quindi implementato un sistema di speech BCI che sintetizza il parlato attraverso la ricostruzione dello stimolo uditivo a partire dalla registrazione dell'attività neuronale della corteccia uditiva e hanno studiato l'andamento dell'accuratezza del parlato prodotto in funzione di tre fattori:

- il modello di regressione (lineare e non lineare)
- la rappresentazione acustica decodificata (spettrogramma uditivo e parametri di uno speech vocoder)
- la banda di frequenza utilizzata per estrarre le caratteristiche del segnale (banda ad alte e/o basse frequenze)

Il sistema di sintesi vocale che ha prodotto il parlato di qualità ed intelligibilità più elevata è stato quello realizzato con modelli di regressione non lineare, cioè DNN, che decodificano una rappresentazione acustica data dall'insieme di parametri di uno speech vocoder, e che utilizza sia la banda ad alta frequenza che quella a bassa frequenza per estrarre le caratteristiche del segnale.

Nei paragrafi seguenti i tre aspetti caratterizzanti del sistema di speech BCI sono trattati nel dettaglio e vengono determinate le prestazioni del sistema misurando la qualità e l'accuratezza dello stimolo ricostruito attraverso il coefficiente di correlazione di Pearson, che valuta la correlazione lineare tra due variabili, e la misura ESTOI [40] (extended short time objective intelligibility) che misura la distorsione, nei pattern di modulazioni spettro-temporali, del segnale vocale.

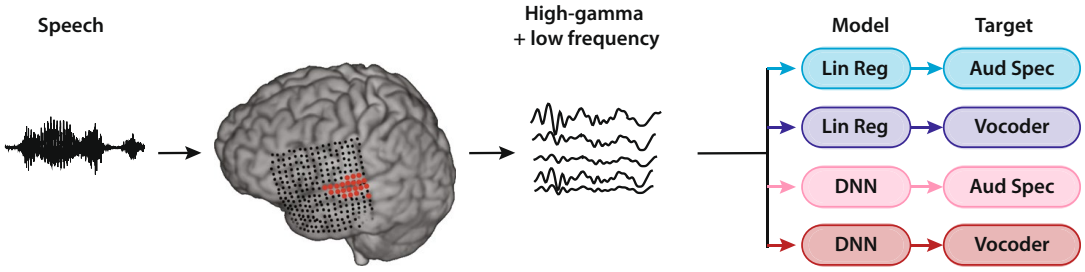

Figura 3.1 rappresentazione schematica del metodo di ricostruzione. Inizialmente i pazienti ascoltano delle frasi ripetute da alcuni oratori, viene registrata l'attività neurale nella corteccia uditiva. Vengono selezionate bande ad alta e bassa frequenza dell'attività neurale. Successivamente per decodificare il segnale vengono utilizzati due tipi di modelli di regressione e due tipi di rappresentazioni del parlato, risultanti in quattro combinazioni: regressione lineare allo spettrogramma uditivo, regressione lineare al vocoder, DNN allo spettrogramma uditivo e DNN al vocoder.

3.3 Modelli di regressione

Per poter effettuare la ricostruzione dello stimolo uditivo è necessario decodificare una rappresentazione acustica dello stimolo a partire dall'attività neurale, per fare questo Akbari e colleghi hanno utilizzato due modelli di regressione: lineare e non lineare (realizzato con una DNN).
Successivamente, per determinare quale modello di regressione fosse più efficiente, i ricercatori hanno confrontato le caratteristiche delle rappresentazioni acustiche ottenute tramite i due metodi di decodifica con la rappresentazione acustica originale dello stimolo.

Nel modello di regressione lineare l'algoritmo trova una relazione lineare tra la risposta neuronale evocata dalla presentazione dello stimolo uditivo e la rappresentazione acustica dello stimolo, questo viene implementato assegnando un filtro spazio-temporale a ciascun elettrodo e viene stimato minimizzando lo scarto quadratico medio tra la rappresentazione dello stimolo originale e quella ricostruita.

Il modello di regressione non lineare viene implementato attraverso una deep neural network con un'architettura a due livelli in cui vengono collegate tra loro delle locally connected neural networks e delle fully connected neural networks.
La struttura della rete neurale prevede 2 stadi: una fase di estrazione delle caratteristiche (feature extraction) e una fase di integrazione delle caratteristiche ricavate nello stadio precedente (feature summation).

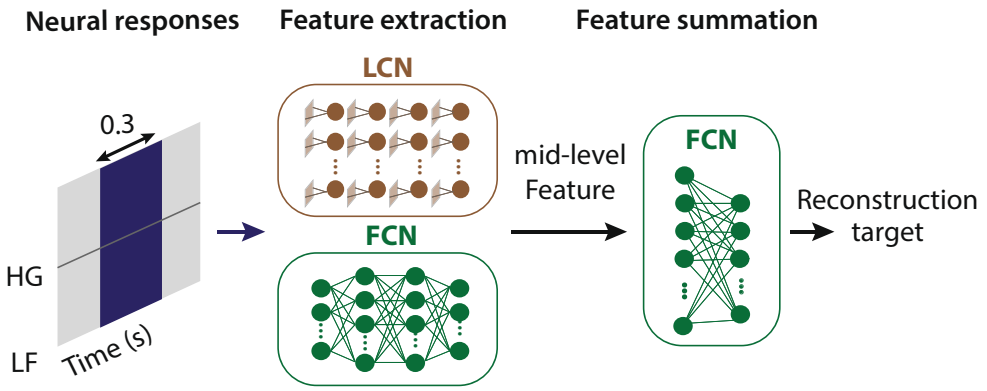

Figura 3.2: in figura è rappresentata l'architettura della rete neurale utilizzata per decodificare la rappresentazione acustica basata sui parametri dello speech vocoder. Il segnale cerebrale viene elaborato dal primo stadio della rete che ne ricava una rappresentazione intermedia, successivamente la rappresentazione intermedia viene elaborata dal secondo stadio che produce in output la rappresentazione acustica ricostruita.

Il primo stadio dell'architettura (feature extraction) è stato implementato con diverse combinazioni di reti neurali. Per determinare quale fosse la combinazione che forniva una migliore ricostruzione dello stimolo uditivo sono state utilizzate cinque combinazioni: FCN, LCN, CNN, FNC e CNN e FCN e LCN.
Le cinque differenti combinazioni sono state utilizzate per decodificare entrambe le rappresentazioni acustiche utilizzate per la ricostruzione dello stimolo uditivo (cioè lo spettrogramma uditivo e lo speech vocoder), e per ogni rappresentazione è stata determinata la struttura della rete neurale che permetteva di ricostruire al meglio lo stimolo.
Akbari e colleghi hanno determinato che il primo stadio della rete neurale che permette di decodificare al meglio la rappresentazione acustica ottenuta con lo spettrogramma uditivo è costituito da una FCN. Tuttavia l'utilizzo della stessa rete neurale per decodificare la seconda rappresentazione acustica, speech vocoder, non fornisce risultati altrettanto positivi poiché per decodificare questa rappresentazione, in cui vanno decodificati molti parametri contemporaneamente, serve un'architettura che possa gestire l'elevato numero di parametri e l'ampia variabilità statistica. Per questo motivo Akbari e colleghi come primo stadio della rete neurale per la decodifica della rappresentazione speech vocoder hanno utilizzato una rete costituita dalla combinazione di una LCN e una FCN.
Per il secondo stadio dell'architettura è stata usata solo una FCN per entrambe le rappresentazioni acustiche decodificate.

Analizzando la qualità e l'accuratezza del parlato sintetizzato a partire dai due modelli di regressione Akbari et al. hanno verificato che con l'utilizzo di un modello di regressione non lineare realizzato con un sistema di decodifica deep neaural network si possono ottenere delle ricostruzioni dello stimolo uditivo molto più simili allo stimolo originale. Questo risultato viene

evidenziato dal grafico in figura 3.3 che mostra come i metodi di decodifica che utilizzano delle DNN ottengano un valore di ESTOI più elevato rispetto ai metodi che utilizzano dei modelli di regressione lineare.
I risultati ottenuti da Akbari e colleghi sono in accordo con quelli ottenuti da altri studi [41, 42] che hanno dimostrato l'importanza dell'utilizzo dei modelli di regressione non lineare per la decodifica dell'attività neurale, pertanto lo studio di Akbari et al. ha confermato questi risultati mostrando i vantaggi dell'utilizzo delle reti neurali profonde nella ricerca per le neuro protesi del linguaggio.

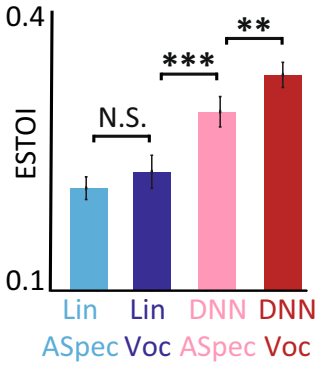

Figura 3.3: l grafico rappresenta il punteggio di intelligibilità oggettiva ottenuto dai differenti modelli. I risultati mostrano che l'utilizzo di un modello di regressione non lineare (DNN) permette di ricostruire un parlato più intelligibile e quindi di una qualità superiore

3.4 Rappresentazione acustica

Una caratteristica di un sistema di speech BCI che influenza fortemente la qualità e l'accuratezza del parlato ricostruito è il tipo di rappresentazione acustica che il sistema utilizza per ricostruire lo stimolo.
Akbari et al. hanno utilizzato due differenti rappresentazioni acustiche dello stimolo uditivo e hanno determinato quale delle due sia la più adeguata a ricostruire con maggiore precisione il parlato. Le rappresentazioni acustiche utilizzate sono lo spettrogramma uditivo e una rappresentazione acustica costituita dall'insieme di quattro parametri utilizzati per la sintesi vocale attraverso un vocoder (speech synthesis parameters o speech vocoder).

Lo spettrogramma è una rappresentazione grafica di un suono che viene rappresentato attraverso tre variabili: frequenza, tempo e intensità. Nello spettrogramma l'asse delle ascisse rappresenta il tempo, l'asse delle ordinate rappresenta la frequenza mentre l'intensità del segnale viene rappresentata attraverso una scala cromatica nel piano che può essere una scala di grigi oppure a colori.

Per ricostruire lo spettrogramma uditivo Akbari et al. hanno utilizzato un modello di calcolo elaborata da Yang e colleghi [43] che si basa su un modello matematico del funzionamento del sistema uditivo periferico. Attraverso questo modello Akbari e colleghi sono riusciti a determinare una rappresentazione spettro-temporale del segnale acustico su un asse tono topico delle frequenze.
Successivamente per ricostruire la forma d'onda (e quindi per sintetizzare il parlato) a partire dallo spettrogramma uditivo i ricercatori hanno utilizzato una procedura di ottimizzazione convessa iterativa [44].

La seconda rappresentazione acustica utilizzata nel presente studio è costituita dall' insieme di quei parametri acustici che un vocoder utilizza per sintetizzare il parlato. I parametri che devono essere stimati per ottenere questa rappresentazione sono quattro: l'inviluppo spettrale, la banda di aperiodicità, l'intonazione e la vocalizzazione.
Per ricostruire il parlato a partire da questa rappresentazione acustica Akbari e colleghi hanno utilizzato un algoritmo di speech synthesis, chiamato WORLD [45] che sintetizza il parlato proprio a partire da questi quattro parametri.

La scelta di utilizzare una rappresentazione acustica che si basi esattamente su quei parametri utilizzati da un vocoder per riprodurre il parlato è chiaramente la scelta più naturale e diretta per implementare un sistema di speech BCI, tuttavia questa strategia non era mai stata adottata in precedenza perché la stima di tutti questi parametri a partire dall'attività neurale richiede una decodifica accurata che è difficile da ottenere con le convenzionali tecniche di apprendimento automatico.
Quello che ha permesso ad Akbari et al. di decodificare tale rappresentazione in modo accurato è stato l'utilizzo di una rete neurale profonda che fosse in grado di gestire l'elevato numero di parametri della rappresentazione.

L'accuratezza con cui si riescono a decodificare i parametri utilizzati per la rappresentazione acustica speech vocoder viene valutata attraverso il coefficiente di correlazione di Pearson come mostrato in figura 3.4, in cui si evidenzia come i metodi di regressione non lineare riescano a ricostruire tutti i parametri con più precisione rispetto ai metodi di regressione lineare.

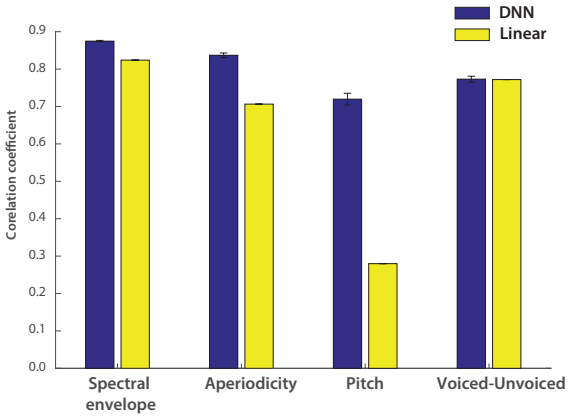

Figura 3.4: accuratezza della ricostruzione dei singoli parametri espressa in funzione del coefficiente di correlazione di Pearson

I risultati delle due differenti strategie di decodifica sono mostrati nel grafico in figura 3.3, in cui si evidenzia che l'utilizzo di una rappresentazione basata sui parametri per uno speech vocoder ricostruisce lo stimolo in maniera più accurata rispetto all'utilizzo di uno spettrogramma uditivo. Quindi, grazie all'utilizzo di una rappresentazione acustica basata sui parametri di uno speech vocoder, Akbari e colleghi sono riusciti ad ottenere una produzione vocale fluente e intelligibile, con una qualità e un'accuratezza superiore rispetto a quella ottenuta con lo spettrogramma uditivo.

3.5 Bande di frequenza

Recentemente molti studi hanno dimostrato che la banda a bassa ed alta frequenza del segnale di attività neurale codifica informazioni differenti e complementari di uno stesso stimolo [11-19].
Per determinare quale banda di frequenza del segnale di attività cerebrale permetta di ottenere la migliore ricostruzione dello stimolo uditivo, in termini di accuratezza e qualità della ricostruzione, Akbari e colleghi hanno utilizzato due differenti bande di frequenza per ricostruire lo stimolo: banda a bassa frequenza e banda ad alta frequenza.
L'evidenza sperimentale ha mostrato che le migliori prestazioni di accuratezza, misurate in termini di scala ESTOI, figura 3.5, sono state ottenute utilizzando una combinazione delle due bande di frequenza piuttosto che utilizzandone una soltanto. Il risultato ottenuto è quindi in accordo con studi precedenti che hanno dimostrato che le bande di frequenza codificano informazioni complementari del segnale cerebrale.

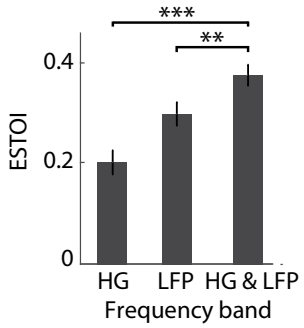

Figura 3.5: il grafico rappresenta il punteggio ESTOI ottenuto dai modelli in funzione della frequenza di acquisizione utilizzata per estrarre le caratteristiche del segnale. I risultati mostrano che utilizzando sia le alte che le basse frequenze si ricostruisce uno stimolo con punteggio di ESTOI più alto e quindi con un'intelligibilità maggiore.

3.6 Conclusioni

Akabari e colleghi hanno realizzato un sistema di sintesi vocale che riproduce il discorso parlato attraverso la ricostruzione della rappresentazione acustica basata sui parametri di uno speech vocoder.
In precedenza nessuno era riuscito ad ottenere questo risultato poiché, per ricostruire un parlato intelligibile attraverso questa rappresentazione, è necessario ricostruire la rappresentazione con un'elevata accuratezza altrimenti il vocoder, che utilizza la rappresentazione per ricostruire la produzione vocale, non riesce a ricostruire un parlato intelligibile se i parametri per la sintesi vocale sono poco accurati.
Akbari e colleghi sono riusciti a decodificare in modo accurato l'elevato numero di parametri grazie all'utilizzo di una tecnica di decofica basata sulle DNN che ha permesso di decodificare con molta accurateza tutti i parametri della rappresentazione e di conseguenza gli ha permesso di riprodurre un parlato intelligibile.

L'approccio di sintesi vocale implementato da Akbari e colleghi pone le basi per fututi sistemi di speech BCI che potranno ripristinare un parlato fluente ed intelligibile per pazienti affetti da sclerosi laterale amiotrofica e sindrone locked-in.

Capitolo 4

4. Sintesi vocale attraverso reti convoluzionali densamente connesse

4.1 Introduzione

Nel presente capitolo viene descritto il lavoro di Angrick e colleghi [2] che hanno realizzato un sistema di speech BCI che utilizza delle reti neurali convoluzionali densamente connesse (DenseNet [46] vedi paragrafo 4.3) per decodificare una rappresentazione acustica del parlato a partire dal segnale ECoG.

A differenza del lavoro realizzato da Akbari e colleghi, descritto nel capitolo precedente, in cui viene utilizzata una rappresentazione acustica speech vocoder, la rappresentazione acustica utilizzata da Angrick et al. è costituita da uno spettrogramma logaritmico in scala mel, cioè uno spettrogramma in cui l'asse delle frequenze è rappresentato secondo la scala Mel ossia una scala di percezione dell'intonazione vocale definita da Steven et al. [47].

Il sistema di sintesi vocale implementato da Angrick e colleghi decodifica una rappresentazione acustica del parlato attraverso delle reti neurali, e successivamente sintetizza il parlato a partire dalla rappresentazione acustica logMel spectrogram utilizzando un Wavenet vocoder [48, 49] condizionato secondo le stesse logMel features.

Questo sistema di speech BCI permette di ricostruire il parlato con elevata accuratezza ed intelligibilità e di seguito vengono descritte nel dettaglio le caratteristiche del sistema concentrandosi sulla rappresentazione acustica e sul tipo di reti neurali utilizzate.

4.2 Registrazione dell'attività cerebrale

Angrick e colleghi hanno registrato l'attività cerebrale attraverso ECoG da sei pazienti che indossavano gli elettrodi poiché dovevano effettuare un intervento al cervello da svegli per la resezione di un tumore cerebrale.
Durante la fase di acquisizione dati i partecipanti, che avevano elettrodi posizionati sulla corteccia ventrale sensorimotoria (vSMC), sulla corteccia premotoria (PM) oppure sul giro frontale inferiore (IFG), dovevano leggere a voce alta delle parole, che gli venivano mostrate su uno schermo, mentre venivano registrate contemporaneamente l'attività cerebrale e la voce del paziente.

La contemporanea registrazione dell'attività cerebrale e della voce del paziente costituiscono un passaggio fondamentale per la realizzazione del sistema di speech BCI poiché fornisco i dati necessari per la fase di addestramento delle reti neurali. Quando si vuole estrarre una rappresentazione acustica a partire dal segnale ECoG si utilizza un metodo di addestramento chiamato supervised learning in cui si allena la rete neurale a decodificare la rappresentazione confrontando il risultato della decodifica con la rappresentazione originale dello stimolo (ricavata dalla registrazione della voce del paziente) e modificando gradualmente i parametri della rete neurale fino a quando le due rappresentazioni non coincidono.

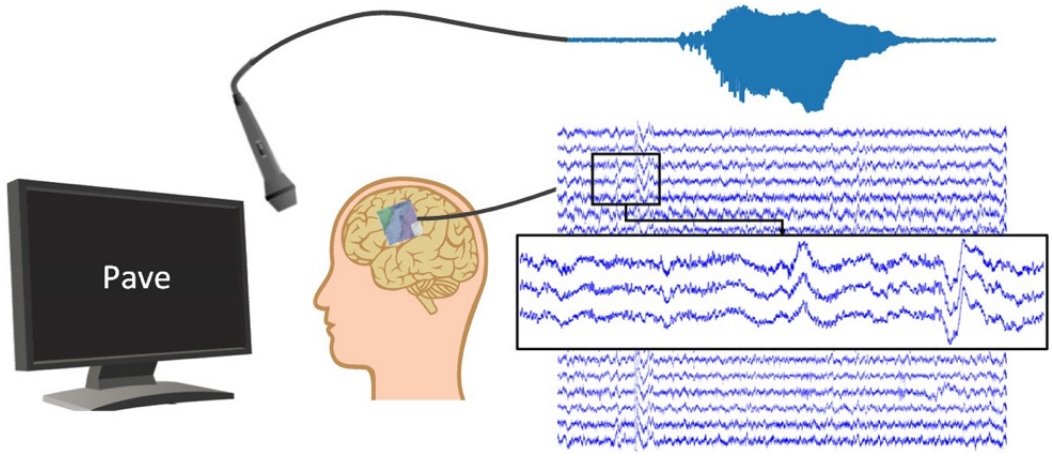

Figura 4.1: Illustrazione della fase di acquisizione: ai partecipanti viene chiesto di ripetere delle parole che vedono su uno schermo, durante la produzione vocale di queste parole vengono registrati contemporaneamente i dati acustici e l'attività cerebrale mediante ECoG.

Per estrarre le informazioni rilevanti dal segnale ECoG Angrick et colleghi hanno considerando la banda di frequenza gamma (70-170 Hz) poiché è stato dimostrato da numerosi studi [11-15] che l'attività neurale in questa banda di frequenze contiene informazioni legate alla produzione linguistica.
Successivamente i ricercatori hanno determinato la potenza del segnale ECoG, calcolando la media del segnale al quadrato e applicando una trasformazione logaritmica. Attraverso questo procedimento di data processing Angrick et al. hanno ricavato la potenza logaritmica del segnale nella banda di frequenze gamma e hanno utilizzato questa caratteristica del segnale ECoG per decodificare la rappresentazione acustica attraverso le reti neurali.

4.3 Rappresentazione acustica e approccio di decodifica

Dopo aver ottenuto le registrazioni dell'attività cerebrale del paziente e le registrazioni vocali, Angrick e colleghi hanno decodificato la rappresentazione acustica log Mel spectrogram a partire dal segnale ECoG.

Come detto in precedenza, per poter ricostruire la rappresentazione acustica vengono utilizzati dei modelli di regressione non lineare, costituiti da delle DNN, che necessitano di una fase di allenamento prima di poter decodificare lo spettrogramma acustico. Considerando che durante la fase di allenamento vengono confrontate le rappresentazioni acustiche decodificate e le rappresentazioni acustiche dello stimolo originale è prima necessario effettuare un'elaborazione della produzione vocale del paziente, che era stata registrata tramite un microfono, per decodificare la rappresentazione basata sullo spettrogramma logaritmico attraverso le reti neurali.
Per ottenere lo spettrogramma logaritmico in scala mel a partire dalla registrazione della voce del paziente i ricercatori hanno ricavato lo spettrogramma in ampiezza, non considerando le informazioni sulla fase, e successivamente hanno convertito lo spettrogramma in scala mel e hanno applicato una trasformazione logaritmica.

Dopo aver terminato la fase di elaborazione della produzione vocale del partecipante, necessaria per addestrare le reti, Angrick et al. hanno potuto procedere con la fase di decodifica del segnale ECoG. Per decodificare lo spettrogramma logaritmico in scala mel a partire dall'attività cerebrale Angrick e colleghi hanno utilizzato una particolare architettura di rete neurale chiamata DenseNet [46] costituita da delle reti convoluzionali densamente connesse.

L'architettura DenseNet è una rete neurale feed-forward multi-strato che utilizza dei collegamenti aggiuntivi tra i layer della struttura come mostrato in figura 4.2.
In questa rete neurale ogni layer riceve come input le feature maps (output) elaborate dai livelli precedenti e invia in output le proprie feature maps a tutti i layer successivi; questa struttura, in cui tutti i livelli sono collegati tra loro, è definita densamente connessa.
A differenza delle tradizionali reti neurali convoluzionali, in cui dati L layers della rete sono presenti L connessioni, nell'architettura DenseNet, dati L layers, troviamo $L(L+1)/2$ connessioni.

L'utilizzo di un'architettura di questo tipo permette, grazie alle connessioni aggiuntive tra i vari livelli, di riutilizzare delle informazioni nei vari strati della rete e quindi permette di apprendere modelli più compatti e più accurati; Angrick e colleghi hanno infatti adottato questo approccio di decodifica perché gli ha permesso di decodificare la rappresentazione in maniera più accurata rispetto all'utilizzo di una tradizionale CNN.

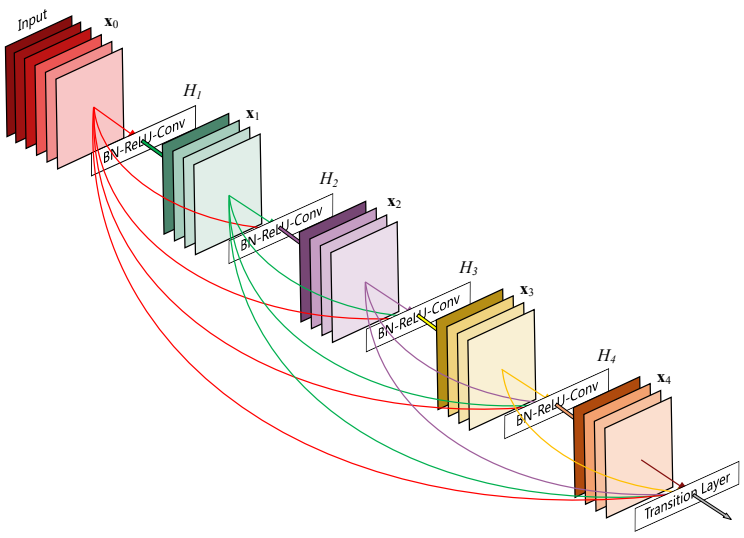

Figura 4.2: Architettura DenseNet

Dopo aver ricavato la rappresentazione logMel dello spectrogramma i ricercatori hanno utilizzato un sistema di sintesi vocale chiamato Wavenet vocoder [48, 49] per ricostruire la forma d'onda del parlato a partire dalla rappresentazione acustica.
La particolarità del vocoder utilizzato da Angrick e colleghi consiste nel poter essere condizionato sulle stesse caratteristiche acustiche della rappresentazione logMel spectrogram, in questo modo Angrick et al. hanno potuto sintetizzare il parlato in modo diretto a partire dalla rappresentazione acustica.

4.4 Conclusioni

Per valutare la qualità e l'intelligibilità del parlato sintetizzato Angrick e colleghi hanno utilizzato il coefficiente di correlazione di Pearson e la misura STOI, short time objective intelligibility.
Considerando i valori dei coefficienti di correlazione di Pearson mostrati in figura 4.3, possiamo dire che il discorso sintetizzato ha raggiunto dei valori di correlazione superiori rispetto al chance level per tutti e sei i partecipanti, tuttavia solo un partecipante è riuscito a raggiungere un valore elevato del coefficiente di correlazione di Pearson, pari a 0,69 mentre gli altri partecipanti hanno ottenuto valori minori.
Dal punto di vista dell'intelligibilità il sistema di speech BCI realizzato da Angrick et al. ha permesso di riprodurre un parlato che ha raggiunto dei valori STOI superiori al chance level per tutti i pazienti e il miglior risultato è stato ottenuto con il paziente cinque, per il quale la misura di STOI è arrivata a 0,53.

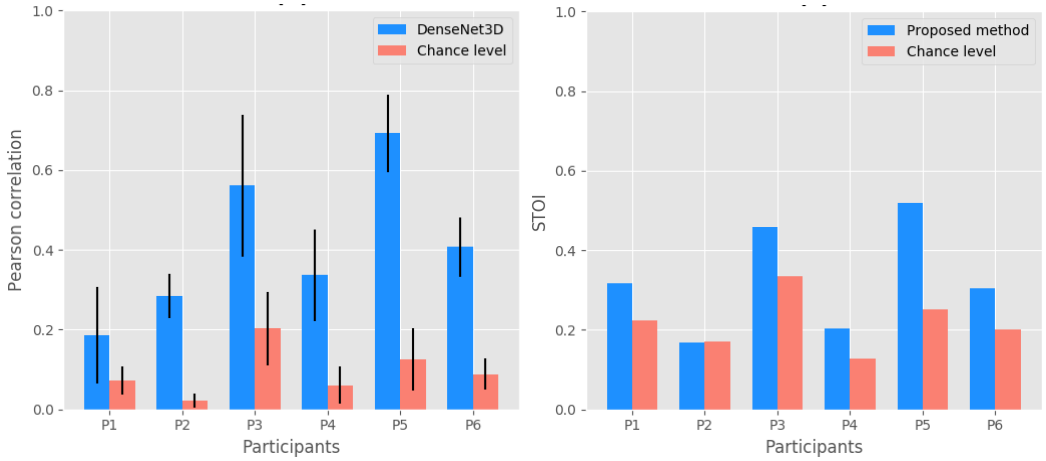

Figura 4.3: A sinistra è rappresentato il grafico che descrive i valori del coefficiente di correlazione di Pearson tra lo spettrogramma ricostruito e quello originale. A destra sono rappresentati i valori della misura STOI per ogni partecipante. In entrambi i grafici i risultati ottenuti da ciascun paziente sono confrontati con il chance level.

I risultati ottenuti da Angrick e colleghi dimostrano che è possibile sintetizzare un parlato intelligibile a partire dall'attività cerebrale registrata attraverso ECoG.
La particolarità del sistema di speech BCI implementato da Angrick et al. è di aver utilizzato una struttura di reti neurali DenseNet che gli ha permesso di ricavare con una discreta accuratezza la rappresentazione del parlato; inoltre Angrick et al. sono riusciti a riprodurre un parlato intelligibile grazie all'utilizzo di un Wavenet vocoder condizionato sulla stessa rappresentazione acustica del segnale decodificato, questo ha permesso di ricostruire il parlato con un livello di intelligibilità elevato.

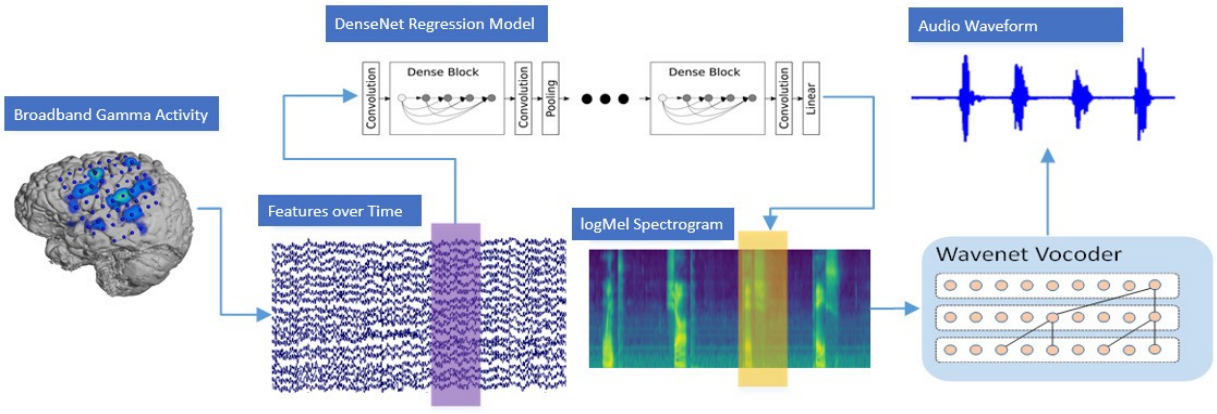

Figura 4.4: la figura rappresenta il flusso di lavoro utilizzato da Angrick et al. per sintetizzare il parlato a partire dall'attività cerebrale. Inizialmente viene registrata attività cerebrale attraverso ECoG e viene utilizzata la banda di frequenza gamma del segnale per estrarre lo spettrogramma logaritmico in scala mel attraverso delle reti neurali DenseNet. Successivamente tramite un Wavenet vocoder si ricostruisce il parlato a partire dalla rappresentazione acustica ricavata nella fase di decodifica

L’approccio di sintesi vocale realizzato da Angrick e colleghi pone le basi per il futuro delle speech BCI che sintetizzano il parlato in maniera diretta a partire dall’attività cerebrale registrata attraverso elettrocorticografia.
La caratteristica più rilevante del sistema di sintesi vocale realizzato da Angrick et al. è il sistema di decodifica dell’attività cerebrale poiché quest’ultimo è in grado di ricostruire accuratamente la rappresentazione acustica del parlato avendo a disposizione una piccola quantità di dati, circa 300 parole. Generalmente i tradizionali sistemi di speech BCI riescono a sintetizzare il parlato se vengono allenati per diverse ore con una grande quantità di dati a disposizione, tuttavia spesso, con i pazienti che hanno problemi di linguaggio, non è possibile avere a disposizione una grande quantità di dati e di conseguenza i tradizionali sistemi di decodifica non riescono a ricostruire un parlato intelligibile.
Il sistema di sintesi vocale implementato da Angrick e colleghi risolve questo problema grazie all’utilizzo di reti neurali DenseNet e pone le basi per la realizzazione di una neuroprotesi per il linguaggio per pazienti con problemi di comunicazione.

Capitolo 5

5. Sintesi vocale a partire da una rappresentazione articolatoria del parlato

5.1 Introduzione

Quando parliamo produciamo normalmente cento o centocinquanta parole al minuto e nella produzione vocale vengono coinvolti quasi 100 muscoli che si attivano per muovere gli organi del tratto vocale cioè la lingua, le labbra, la mandibola e la laringe.
Studi recenti hanno dimostrato che in alcune aree cerebrali viene codificata una rappresentazione articolatoria del parlato [13], [50], [51] cioè una rappresentazione che descrive come variano nel tempo le posizioni degli organi del tratto vocale. Grazie alle evidenze scientifiche mostrate da questi studi, recentemente si è pensato di poter realizzare un sistema di speech BCI che ricostruisca il parlato a partire da una rappresentazione articolatoria.

In questo contesto si inserisce il terzo ed ultimo approccio di speech BCI trattato nella tesi che descrive lo studio di Anumanchipalli e colleghi. Questi ricercatori sono stati i primi ad implementare un sistema di sintesi vocale che rileva l'attività cerebrale attraverso ECoG e che sintetizza il parlato utilizzando una rappresentazione articolatoria.

Anumanchipalli e colleghi hanno realizzato un sistema di speech BCI che sintetizza un parlato fluente e intelligibile, questo risultato è stato raggiunto grazie all'utilizzo di un sistema di decodifica a due stadi, implementato attraverso delle reti neurali ricorrenti.

Nel capitolo seguente analizzeremo nel dettaglio le componenti del sistema di speech BCI realizzato da Anumanchipalli et al. concentrando la nostra attenzione sulla fase di acquisizione dati e di allenamento delle reti neurali e sul metodo di decodifica dell'attività cerebrale.

5.2 Acquisizione dati

Anumanchipalli e colleghi hanno registrato l'attività cerebrale attraverso ECoG da cinque pazienti che indossavano gli elettrodi per elettrocorticografia poiché sottoposti ad un trattamento clinico per l'epilessia. Durante l'esperimento gli elettrodi rilevavano l'attività neurale da diverse aree cerebrali: la corteccia ventrale sensorimotoria (vSMC), il giro temporale superiore (STG) e il giro frontale inferiore (IFG).

La fase di acquisizione dati prevedeva la contemporanea registrazione della voce del paziente, mentre pronunciava determinate parole, e la registrazione dell'attività cerebrale.
Per simulare il funzionamento dell'approccio di decodifica per pazienti con sindromi neurodegenerative, per i quali, a causa dei problemi di comunicazione, è difficile acquisire molti dati, Anumanchipalli e colleghi hanno utilizzato una breve quantità di dati facendo leggere ad ogni partecipante circa 400 parole. Inoltre, per testare il funzionamento del sistema di sintesi vocale nella condizione in cui il paziente non abbia la possibilità di vocalizzare, come nel caso di afasia, Anumanchipalli et al. hanno effettuato un'acquisizione dati in cui un partecipante leggeva una sequenza di parole alternando una lettura ad alta voce e una lettura in cui veniva mimato il movimento del parlato senza vocalizzare.

Per decodificare le informazioni rilevanti del segnale ECoG i ricercatori hanno utilizzato sia la banda ad alte frequenze (70-200 Hz) che la banda a basse frequenze (1-30 Hz) dell'attività cerebrale, poiché numerosi studi hanno dimostrato che entrambe le bande di frequenza contengono informazioni determinanti per studiare i meccanismi del linguaggio [52-54].

5.3 Approccio di decodifica

Il sistema di speech BCI realizzato da Anumanchipalli et al. permette di sintetizzare il parlato attraverso l'utilizzo di una strategia di decodifica a due stadi implementata mediante delle reti neurali ricorrenti. Come mostrato in figura 5.1 il primo stadio decodifica una rappresentazione articolatoria del parlato a partire dall'attività neurale e successivamente, a partire dalla rappresentazione appena decodificata, il secondo stadio determina una rappresentazione acustica del parlato che viene poi utilizzata per la sintesi vocale.

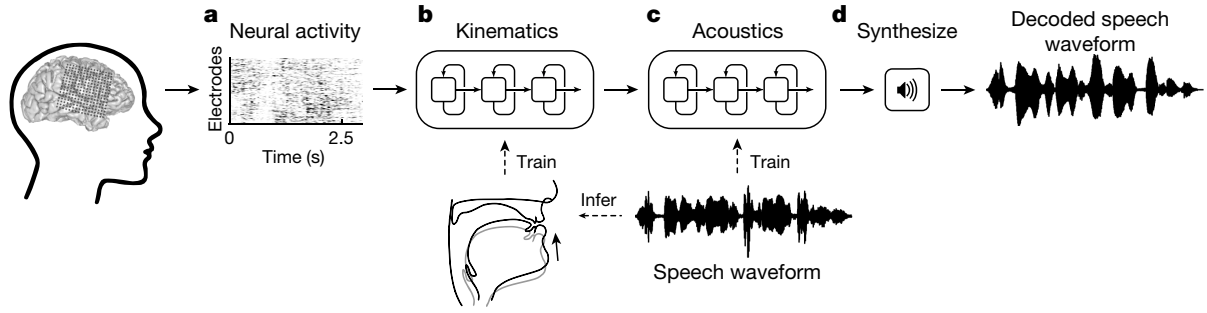

Figura 5.1: la figura mostra il procedimento di sintesi vocale a partire dall'attività neurale realizzato da Anumanchipalli et al. Inizialmente viene registrata l'attività cerebrale attraverso ECoG, successivamente avviene la fase di decodifica a due stadi: attività neurale-rappresentazione articolatoria e rappresentazione articolatoria-rappresentazione acustica ed infine, a partire dalla rappresentazione acustica del segnale, viene sintetizzato il parlato.

Per implementare questo sistema di decodifica è necessario allenare le reti neurali a decodificare le due rappresentazioni e per fare questo è necessario avere a disposizione i dati per allenare le reti neurali, cioè la rappresentazione articolatoria e la rappresentazione acustica del parlato originale.
Anumanchipalli et al. hanno ricavato la rappresentazione acustica del parlato a partire dalla registrazione della voce del paziente, mentre per ottenere la rappresentazione articolatoria i ricercatori hanno dovuto utilizzare un ulteriore sistema di decodifica in quanto non disponevano di una registrazione delle traiettorie degli organi articolatori del parlato.
Quindi per realizzare il sistema di speech BCI Anumanchipalli e colleghi hanno inizialmente dovuto determinare le traiettorie cinematiche degli organi del tratto vocale, necessarie per allenare il primo stadio del sistema di decodifica e successivamente hanno potuto implementare il sistema di decodifica che permette di sintetizzare il parlato.

Di seguito vengono analizzati in due paragrafi distinti gli approcci con cui Anumanchipalli et al. hanno determinato le traiettorie degli organi del tratto vocale e come hanno realizzato il sistema di decodifica del parlato.

5.3.1 Deduzione delle traiettorie degli articolatori del tratto vocale

Generalmente per ottenere la rappresentazione articolatoria del parlato si registrano le traiettorie degli organi del tratto vocale attraverso uno strumento chiamato EMA, electromagnetic midsagittal articulography.
Per registrare le traiettorie attraverso EMA vengono posizionati dei piccoli sensori sugli organi del tratto vocale, solitamente labbra e lingua, e viene inviato un campo magnetico alla testa del paziente.
Durante la produzione linguistica il paziente muove gli organi del tratto vocale e lo strumento registra la variazione della posizione dei sensori all'interno del campo magnetico permettendo così di ricostruire le traiettorie cinematiche degli organi articolatori del parlato.

Anumanchipalli e colleghi, tuttavia, non hanno potuto utilizzare la tecnica EMA per registrare le traiettorie degli articolatori del parlato poiché i campi magnetici che questa tecnica utilizza per registrare i movimenti possono interrompere il funzionamento degli elettrodi per ECoG.
A causa dell'incompatibilità tra la registrazione dell'attività cerebrale mediante ECoG e la registrazione delle traiettorie cinematiche attraverso EMA, Anumanchipalli e colleghi hanno dovuto utilizzare un metodo differente per ricavare la rappresentazione articolatoria del parlato.

Per dedurre le traiettorie cinematiche, al fine di allenare il sistema di decodifica che estrae una rappresentazione articolatoria del parlato a partire dal segnale ECoG, Anumanchipalli e colleghi hanno utilizzato un modello matematico statistico implementato attraverso delle RNN.

L'approccio statistico utilizzato da Anumanchipalli et al. ricostruisce le traiettorie cinematiche degli organi del tratto vocale a partire dalla rappresentazione acustica e fonologica della voce del paziente.
Per implementare questo modello matematico i ricercatori hanno utilizzato delle recurrent neural networks che realizzano un'architettura encoder-decoder. In tale architettura la funzione dell'encoder consiste nel costruire una rappresentazione articolatoria del parlato a partire dalla rappresentazione acustica attraverso un algoritmo di acoustic-to-articulatory inversion (AAI) implementato da Chartier e colleghi [55]. Successivamente il decoder, per verificare che la rappresentazione articolatoria codificata sia adatta per sintetizzare il parlato, ricostruisce il segnale acustico originale partendo dalla rappresentazione articolatoria precedentemente codificata.

Le reti neurali che realizzano questo sistema sono state allenate con dei dati speaker independent provenienti da un database [56] e, a seguito della fase di allenamento, le traiettorie cinematiche codificate dall'encoder a partire dalla rappresentazione acustica della voce del paziente sono state utilizzate come rappresentazione articolatoria per allenare il primo stadio del sistema di decodifica a partire dall'attività neurale.

5.3.2 Decodifica dall'attività neurale

Dopo aver determinato come vengono dedotte le traiettorie cinematiche del tratto vocale, necessarie per allenare le reti neurali che decodificano la rappresentazione articolatoria del parlato, possiamo descrivere l'approccio di decodifica realizzato da Anumanchipalli e colleghi. Il sistema di decodifica implementato in questo studio è costituito da una rete neurale ricorrente a due stadi entrambi realizzati con delle Long Short Term Memory network (LSTM). Il primo stadio della rete decodifica una rappresentazione articolatoria del parlato, costituita dalle traiettorie degli organi del tratto vocale, a partire dall'attività cerebrale mentre il secondo stadio decodifica una rappresentazione acustica del parlato a partire dalla rappresentazione articolatoria precedentemente decodificata.
La rappresentazione acustica che Anumanchipalli e colleghi hanno scelto di decodificare si basa sui Mel frequency cepstral coefficients (MFCCs), cioè dei coefficienti che descrivono l'inviluppo dello spettro di potenza del segnale acustico e che costituiscono una rappresentazione acustica molto utilizzata nell'ambito della sintesi vocale [57, 58].

Per valutare le prestazioni del sistema di decodifica e, di conseguenza, la qualità del parlato sintetizzato Anumanchipalli et al. hanno utilizzato una misura oggettiva chiamata mel cepstral distortion (MCD). Questa misura è data dalla differenza di distorsione dello spettro del parlato sintetizzato rispetto allo spettro del parlato originale; minore è il valore di MCD migliore è la prestazione del sistema di sintesi vocale poiché un basso valore di MCD indica che lo spettro ricostruito è poco distorto rispetto a quello originale.

Nella figura 5.2 vengono mostrati i valori di MCD per i vari partecipanti, considerando il valore ottenuto dal parlato sintetizzato dalle rappresentazioni articolatorie originali (reference), il valore ottenuto per il parlato sintetizzato a partire dalle rappresentazioni articolatorie decodificate dall'attività neurale (decoded) e il valore ottenuto da un decodificatore casuale.
I risultati mostrano che l'approccio di decodifica implementato da Anumanchipalli e colleghi permette di ricostruire accuratamente lo spettro del parlato poiché i valori di MCD decoded sono molto vicini ai valori MCD reference.

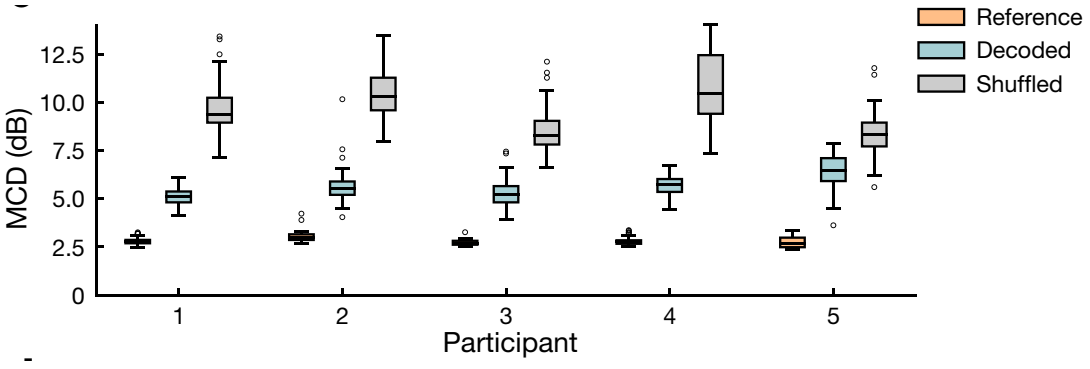

Figura 5.2 Il grafico mostra la distorsione spettrale, misurata attraverso MCD, del parlato prodotto dai 5 partecipanti. I valori reference di MCD si riferiscono al parlato sintetizzato a partire dalle rappresentazioni articolatorie originali, ricavate con il metodo statistico matematico.

Anumanchipalli e colleghi hanno valutato le prestazioni del sistema di decodifica anche in funzione dell'area cerebrale dalla quale veniva registrata l'attività neurale (vSMC, IFG, STG) poiché, come detto nel capitolo uno, in aree diverse del cervello vengono codificate rappresentazioni differenti del linguaggio e questo può influenzare l'accuratezza con cui una rappresentazione viene decodificata.
Anumanchipalli et al. hanno valutato le prestazioni del sistema di decodifica calcolando la differenza del valore di MCD tra un discorso sintetizzato utilizzando l'attività cerebrale proveniente da tutte tre le aree e un discorso sintetizzato escludendo l'attività cerebrale di una delle tre aree.
I risultati di questa analisi sono mostrati in figura 5.3 dalla quale si evidenzia che la peggiore ricostruzione del parlato viene ottenuta escludendo l'attività neurale proveniente dalla vSMC. Questo risultato è in accordo con le affermazioni effettuate da studi recenti che hanno dimostrato che vSMC è la principale regione cerebrale nella quale viene codificata una rappresentazione articolatoria del parlato [13], [50], [51].

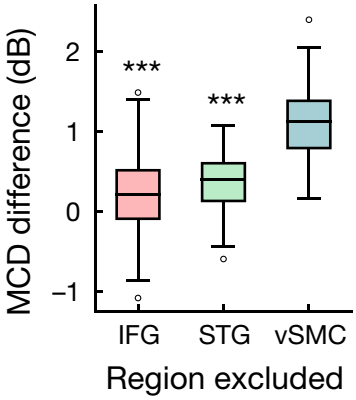

Figura 5.3: il grafico mostra la differenza dei valori di MCD tra un decoder allenato con segnali cerebrali provenienti da tutte le regioni e un decoder allenato escludendo una delle tre aree cerebrali. La maggiore differenza di MCD si ottiene escludendo i segnali cerebrali provenienti da vSMC questo implica che le prestazioni del decoder sono peggiori quando non viene considerata l'attività neurale proveniente da quest'area.

La chiave del sistema di sintesi vocale realizzato Anumanchipalli e colleghi è la rappresentazione articolatoria intermedia utilizzata nella tecnica di decodifica del parlato. Come mostrato in figura 5.4 attraverso l'utilizzo di questa rappresentazione viene ricostruito uno spettro del parlato meno distorto rispetto a quello ricostruito direttamente a partire dalla rappresentazione acustica del parlato. Questi risultati indicano che l'utilizzo di una rappresentazione articolatoria intermedia migliora le prestazioni del sistema di decodifica.

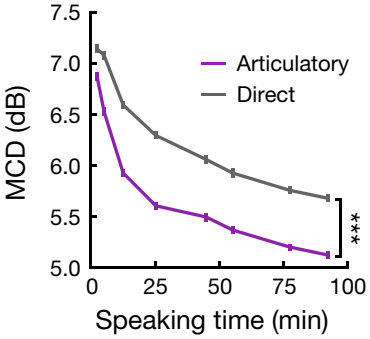

Figura 5.4: il grafico mostra i valori di MCD ottenuti da un decoder che utilizza una rappresentazione articolatoria intermedia (linea viola) e un decoder che decodifica direttamente una rappresentazione acustica a partire dal segnale ECoG. I valori di MCD più bassi sono ottenuti dal decoder che utilizza una rappresentazione articolatoria intermedia ed è quindi quest'ultimo che ricostruisce il parlato più accuratamente.

5.4 Conclusioni

Anumanchipalli e colleghi hanno realizzato un sistema di speech BCI capace di riprodurre un parlato fluente ed intelligibile a partire dall'attività cerebrale registrata mediante ECoG.
Il sistema di sintesi vocale implementato in questo studio è innovativo e si distingue dai precedenti sistemi poiché è il primo che riesce a ricostruire un discorso di elevata qualità a partire dall'attività cerebrale utilizzando una rappresentazione articolatoria del parlato. Il sistema a doppia decodifica, che utilizza la rappresentazione articolatoria del parlato come rappresentazione intermedia, è la parte più importante del sistema realizzato da Anumanchipalli et al. poiché l'utilizzo di tale rappresentazione ha permesso di migliorare le prestazioni del sistema di decodifica (vedi figura 5.4).

Considerando che gli ottimi risultati ottenuti dal sistema di decodifica dipendono dall'utilizzo della rappresentazione articolatoria del parlato, Anumanchipalli et al. hanno analizzato più approfonditamente queste rappresentazioni e hanno osservato che, quando diversi partecipanti ripetono le stesse parole, le traiettorie degli organi del tratto vocale eseguite dai diversi pazienti sono molto simili. La somiglianza delle traiettorie degli articolatori del parlato dei diversi partecipanti riferite alle stesse frasi è stata valutata attraverso il coefficiente di correlazione di Pearson, e come mostrato in figura 5.5 questo coefficiente è risultato circa uguale a 0.8, indicando un elevato grado di somiglianza tra le rappresentazioni articolatorie.

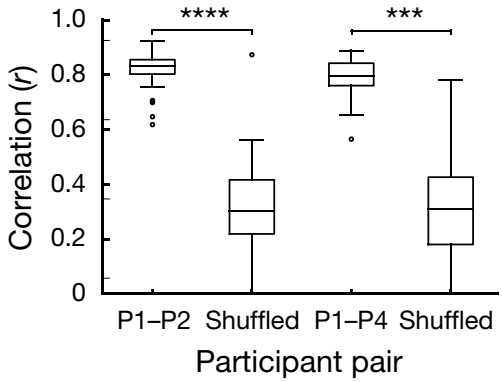

Figura 5.5: il grafico rappresenta il valore del coefficiente di correlazione di Pearson tra le traiettorie degli organi articolatori del parlato del partecipante 1 e 2 e del partecipante 1 e 4. I valori elevati di r indicano che le traiettorie degli organi del tratto vocale dei partecipanti sono molto simili.

La somiglianza delle traiettorie cinematiche tra i diversi partecipanti suggerisce che probabilmente esiste una rappresentazione articolatoria del parlato condivisa.
L'esistenza di una rappresentazione articolatoria condivisa tra i partecipanti potrebbe permettere di realizzare un sistema di speech BCI in cui: il primo stadio del sistema di decodifica ricostruisce la rappresentazione articolatoria del parlato a partire dall'attività neurale

specifica per ogni paziente, mentre il secondo stadio ricostruisce la rappresentazione acustica a partire dalla rappresentazione articolatoria condivisa, indipendente dallo specifico partecipante. Il vantaggio di utilizzare la rappresentazione articolatoria condivisa nel consiste nel fatto che il secondo stadio del sistema di decodifica può essere realizzato con un decodificatore che è indipendente dal partecipante e che quindi viene allenato con dei dati acustici registrati indipendentemente dalla voce dello specifico partecipante. In questo modo non sarebbe più necessario avere a disposizione la registrazione della voce del paziente per realizzare il sistema di speech BCI e quindi si potrebbe realizzare un sistema di sintesi vocale utilizzabile anche da persone che hanno perso capacità di parlare.

Il sistema di speech BCI realizzato da Anumanchipalli e colleghi è in grado di sintetizzare un parlato intelligibile a partire dall'attività neurale, tuttavia come detto in precedenza, per ricostruire il parlato il sistema di decodifica necessita di una fase di allenamento per la quale è necessario avere a disposizione la registrazione della voce del paziente. Spesso nei pazienti con problemi di comunicazione non è possibile ottenere la registrazione della voce e per questo motivo è difficile riuscire a realizzare un sistema di speech BCI che sintetizzi il parlato direttamente a partire dall'attività cerebrale.
Per cercare di risolvere questo problema Anumanchipalli e colleghi hanno testato il funzionamento del loro sistema di speech BCI nella condizione di parlato mimato cioè in una condizione in cui il partecipante non produce un parlato udibile ma muove gli organi del tratto vocale mimando la produzione linguistica.
Per simulare il comportamento del sistema nella condizione mimed speech Anumanchipalli et al hanno registrato l'attività neurale del partecipante 1 mentre leggeva determinate parole ad alta voce e mentre mimava i movimenti della produzione vocale silenziosamente.

I ricercatori hanno verificato che, anche se il sistema di decodifica non era stato allenato in modo specifico per sintetizzare il parlato mimato, gli spettrogrammi ricostruiti a partire dal discorso mimato mostrano caratteristiche spettrali simili a quelli ricostruiti a partire dal discorso udibile.
Le prestazioni del sistema di speech BCI in condizioni di parlato mimato sono state valutate calcolando MCD e il coefficiente di correlazione di Pearson tra lo spettro del parlato sintetizzato utilizzando il parlato udibile e il discorso mimato. I risultati, mostrati in figura 5.6, evidenziano che, sebbene le prestazioni del sistema di sintesi per il parlato mimato siano inferiori alle prestazioni ottenute a partire dal parlato udibile, è possibile decodificare importanti caratteristiche spettrali del discorso a partire da un parlato mimato, non udibile.

Anumanchipalli e colleghi hanno dimostrato che tale approccio di speech BCI potrebbe essere utilizzato anche da persone che hanno perso la capacità di parlare ma che siano ancora in grado di muovere gli organi articolatori del tratto vocale, come gli afasici o persone che hanno subito una laringectomia.

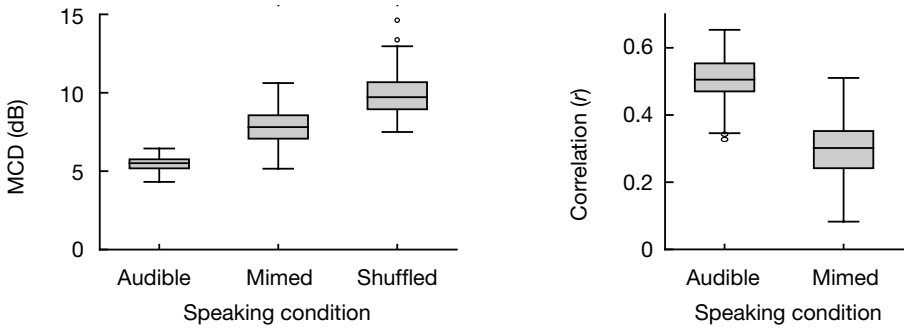

Figura 5.6: i grafici rappresentano i valori di MCD (a sinistra) e coefficiente di correlazione di Pearson (a destra) per il parlato decodificato a partire dall'attività neurale nella condizione di parlato udibile e mimato.

Il sistema di sintesi vocale realizzato da Anumanchipalli e colleghi è innovativo e mostra la fattibilità di poter ricostruire un parlato intelligibile e fluente a partire dall'attività neurale registrata attraverso ECoG. Inoltre, i risultati ottenuti in questo studio pongono importanti basi per futuri approcci di speech BCI utilizzabili da persone che hanno perso la capacità di parlare.

Conclusioni

Negli ultimi anni le protesi neurali per il ripristino del linguaggio sono diventate oggetto di ricerca di grande interesse poiché, grazie alle recenti scoperte relative ai processi cerebrali coinvolti nella produzione lessicale [4-8] e allo sviluppo degli algoritmi di deep learning [30-35], è nata la possibilità concreta di riprodurre un parlato fluente attraverso delle speech BCI.
L'obiettivo del presente elaborato era quello di descrivere le recenti innovazioni nell'ambito delle speech BCI che sintetizzano il parlato direttamente dall'attività neurale. In questo contesto ho scelto di descrivere gli approcci implementati da Akbari et al [1], Angrick et al [2] e Anumanchipalli et al [3] che per primi sembrano essere riusciti a ripristinare una comunicazione fluente e intelligibile.

L'elevata accuratezza e qualità del parlato sintetizzato dai ricercatori nei tre differenti approcci è resa possibile grazie all'utilizzo dell'ECoG come tecnica di registrazione dell'attività cerebrale. Studi precedenti [22-29] non erano riusciti a riprodurre un discorso fluente a causa dei limiti di accuratezza della tecnica di misurazione dell'attività cerebrale che utilizzavano, come EEG e MEG. L'elettrocorticografia permette di superare questi limiti poiché possiede la risoluzione spaziale e temporale necessaria per registrare dei processi cerebrali rapidi come la produzione linguistica; tuttavia l'utilizzo dell'ECoG comporta diversi svantaggi in quanto richiede un intervento chirurgico e l'impianto di elettrodi nel cervello.
I risultati ottenuti da Akbari, Angrick e Anumanchipalli dimostrano, tuttavia, l'importanza dell'utilizzo dell'ECoG come metodo di rilevazione dell'attività cerebrale; di conseguenza, in futuro, per poter applicare questi approcci di speech BCI alle persone affette da problemi di comunicazione, sarà necessario ridurre al minimo i rischi legati all'impianto degli elettrodi per elettrocorticografia.

I metodi descritti nel presente elaborato rappresentano il futuro delle applicazioni di speech BCI; tuttavia, passando ad analizzarne le caratteristiche implementative più nel dettaglio, essi presentano un grande svantaggio poiché per ricostruire il parlato utilizzano dei sistemi di decodifica che hanno inizialmente bisogno di essere allenati con la rappresentazione originale dello stimolo acustico e quindi con la registrazione della voce del paziente. Questo rappresenta un limite degli approcci di sintesi vocale poiché spesso, nella realtà di pazienti che hanno perso la capacità di parlare, i dati richiesti per allenare le reti non sono disponibili.
In questi termini l'approccio più limitato è quello implementato da Angrick e colleghi perché necessita obbligatoriamente della registrazione della voce del paziente e quindi non è realizzabile per coloro che hanno perso la capacità di comunicare. L'approccio di Angrick e colleghi potrebbe essere adattato a persone affette da SLA e LIS registrando la voce del paziente prima che questo arrivi allo stadio terminare della malattia, tuttavia questa soluzione non è adattabile a tutte le persone che hanno perso la capacità di parlare in quanto non potrebbe essere

utilizzata, ad esempio, da persone che perdono la capacità di parlare improvvisamente a causa di una lesione cerebrale.
Il metodo implementato da Anumanchipalli e colleghi risolve parzialmente questo problema attraverso il funzionamento mimed speech della BCI, in cui il sistema può essere utilizzato anche da coloro che non possono vocalizzare il parlato in modo udibile ma possono comunque muovere gli organi articolatori del parlato. Questa soluzione non è adattabile a tutte le persone che hanno perso la capacità di comunicare perché, ad esempio, chi è affetto da SLA e LIS non può muovere gli organi articolatori del tratto vocale; tuttavia l'approccio potrebbe essere utilizzato da persone afasiche o che persone hanno subito un intervento con asportazione della laringe.
L'unico metodo che risolve completamente il problema della registrazione della voce del paziente è quello implementato da Akbari e colleghi in quanto, basandosi sulla ricostruzione dello stimolo acustico a partire dalla corteccia uditiva, non necessita la produzione vocale da parte del paziente.

Ad oggi i tre approcci di speech BCI realizzati da Akbari, Angrick e Anumanchipalli non sono ancora stati testati su pazienti affetti da problemi di comunicazione poiché i metodi proposti sono ancora in fase sperimentale e comportano dei rischi legati all'impianti di elettrodi per ECoG. Tuttavia i risultati ottenuti da questi studi sono molto innovativi e pongono le basi per il futuro delle speech BCI, in cui tali dispositivi potrebbero ripristinare un parlato fluente ed intelligibile per le persone che hanno perso la capacità di parlare.

Bibliografia e sitografia

1. Akbari, H., Khalighinejad, B., Herrero, J. L., Mehta, A. D., and Mesgarani, N. (2019). Towards reconstructing intelligible speech from the human auditory cortex. *Sci. Rep.* 9:874. doi: 10.1038/s41598-018-37359-z
2. Angrick, M., Herff, C., Mugler, E., Tate, M. C., Slutzky, M. W., Krusienski, D. J., et al. (2019). Speech synthesis from ecog using densely connected 3d convolutional neural networks. *J. Neural Eng*. 16:036019. doi: 10.1088/1741-2552/ab0c59
3. Anumanchipalli, G. K., Chartier, J., and Chang, E. F. (2019). Speech synthesis from neural decoding of spoken sentences. *Nature* 568, 493–498. doi: 10.1038/s41586-019-1119-1
4. R. Vandenberghe, A.C. Nobre, C.J.Price The response of left temporal cortex to sentences J Cogn Neurosci, 14 (4) (2002), pp. 550-560
5. C. Humphries, T. Love, D. Swinney, *et al.* Response of anterior temporal cortex to syntactic and prosodic manipulations during sentence processing Hum Brain Mapp, 26 (2005), pp. 128-138
6. C. Humphries, J.R. Binder, D.A.Medler, *et al* .Syntactic and semantic modulation of neural activity during auditory sentence comprehension J Cogn Neurosci, 18 (4) (2006), pp. 665-679
7. G. Hickok, D. Poeppel Dorsal and ventral streams: A framework for understanding aspects of the functional anatomy of language
8. E.F. Lau, C. Phillips, D. Poeppel A cortical network for semantics: (de)constructing the N400 Nat Rev Neurosci, 9 (12) (2008), pp. 920-933
9. Axmacher, N., Henseler, M.M., Jensen, O., Weinreich, I., Elger, C.E., Fell, J., 2010a. Cross-frequency coupling supports multi-item working memory in the human hippocampus. Proceedings of the National Academy of Sciences of the United States of America 107, 3228–3233
10. Mormann, F., Fell, J., Axmacher, N., Weber, B., Lehnertz, K., Elger, C.E., Fernandez, G., 2005. Phase/amplitude reset and theta-gamma interaction in the human medial temporal lobe during a continuous word recognition memory task. Hippocampus 15, 890–900.

11. Pasley, Brian N., David, Stephen V., Mesgarani, Nima, Flinker, Adeen, Shamma, Shihab A., Crone, Nathan E., ... Chang, Edward F. (2012). Reconstructing speech from human auditory cortex Edited by Robert Zatorre. PLoS Biology, 10(1), e1001251. http://dx.doi.org/10.1371/journal.pbio.1001251.
12. Chang, Jochem W., Rieger, Keith Johnson, Berger, Mitchel S., Barbaro, Nicholas M., & Knight, Robert T. (2010). Categorical speech representation in human superior temporal gyrus. Nature Neuroscience, 13(11), 1428–1432. http://dx.doi.org/ 10.1038/nn.2641.
13. Bouchard, Kristofer E., Mesgarani, Nima, Johnson, Keith, & Chang, Edward F. (2013). Functional organization of human sensorimotor cortex for speech articulation. Nature, 495(7441), 327–332. http://dx.doi.org/10.1038/nature11911.
14. Cheung, C., Hamiton, L. S., Johnson, K., and Chang, E. F. (2016). The auditory representation of speech sounds in human motor cortex. *eLife* 5:12577. doi: 10.7554/eLife.12577
15. Conant, D. F., Bouchard, K. E., Leonard, M. K., and Chang, E. F. (2018). Human sensorimotor cortex control of directly measured vocal tract movements during vowel production. *J. Neurosci.* 38, 2955–2966. doi: 10.1523/JNEUROSCI.2382-17.2018
16. Luo, H., and Poeppel, D. (2007). Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. *Neuron* 54, 1001–1010. doi: 10.1016/j.neuron.2007.06.004
17. Ding, N., and Simon, J. Z. (2012). Emergence of neural encoding of auditory objects while listening to competing speakers. *Proc. Natl. Acad. Sci. U.S.A.* 109, 11854–11859. doi: 10.1073/pnas.1205381109
18. Giraud, A.-L., and Poeppel, D. (2012). Cortical oscillations and speech processing: emerging computational principles and operations. *Nat. Neurosci.* 15, 511–517. doi: 10.1038/nn.3063
19. Hermes, D., Miller, K. J., Vansteensel, M. J., Edwards, E., Ferrier, C. H., Bleichner, M. G., et al. (2014). Cortical theta wanes for language. *Neuroimage* 85, 738–748. doi: 10.1016/j.neuroimage.2013.07.029
20. Kubler A, Nijboer F, Mellinger J et al. (2005). Patients with ALS can use sensorimotor rhythms to operate a brain-computer interface. Neurology 64: 1775–1777.
21. Neuper C, M€uller GR, Staiger-S€alzer P et al. (2003). EEG- based communication—a new concept for rehabilitative support in patients with severe motor impairment. Rehabilitation (Stuttg) 42: 371–377.

22. Farwell, L., Donchin, E., 1988. Talking Off the Top of Your Head: toward a mental prosthesis utilizing event-related brain potentials. Electroencephalogr. Clin. Neurophysiol. 70, 510–523.
23. Kaufmann T, Schulz SM, Grunzinger C et al. (2011). Flashing characters with famous faces improves ERP-based brain- computer interface performance. J Neural Eng 8: 056016.
24. Sellers EW, Vaughan TM, Wolpaw JR (2010). A brain- computer interface for long-term independent home use. Amyotroph Lateral Scler 11: 449–455.
25. Wolpaw JR, Bedlack RS, Reda DJ et al. (2018). Independent home use of a brain-computer interface by people with amyotrophic lateral sclerosis. Neurology 91: e258–e267.
26. X. Wang, H. Guo and Z. Huang, "A brain-controlled speech generator based on the P300 speller," *2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, Shanghai, 2017, pp. 1-5, doi: 10.1109/CISP-BMEI.2017.8302263.
27. Nakanishi M, Wang Y, Chen X et al. (2017). Enhancing detec- tion of SSVEPs for a high-speed brain speller using task- related component analysis. IEEE Trans Biomed Eng 65: 104–112. https://doi.org/10.1109/TBME.2017.2694818.
28. Herff, C., Heger, D., de Pesters, A., Telaar, D., Brunner, P., Schalk, G., Schultz, T., 2015. Brain-to-text: decoding spoken phrases from phone representations in the brain. Front. Neurosci. 9, 1–11.
29. Guenther, F.H., Brumberg, J.S., Wright, E.J., Nieto-Castanon, A., Tourville, J.a., Panko, M., Law, R., Siebert, S.A., Bartels, J.L., Andreasen, D.S., Ehirim, P., Mao, H., Kennedy, P.R., 2009. A wireless brain-machine interface for real-time speech synthesis. PLoS One 4, e8218.
30. M.Chau,H.Chen,A machine learning approach to web page filtering using content and structure analysis, Decision Support Systems 44 (2) (2008) 482–494.
31. M. V. Martnez, I. D. Campo, J. Echanobe, and K. Basterretxea, “Driving Behavior Signals and Machine Learning: A Personalized Driver Assistance System,” in *2015 IEEE 18th International Conference on Intelligent Transportation Systems*, Sep. 2015, pp. 2933–2940.
32. G. Ghazaei, A. Alameer, P. Degenaar, G. Morgan, and K. Nazarpour, “Deep learning-based artificial vision for grasp classification in myoelec- tric hands,” J. Neural Eng. 14, 036025 (2017).

33. Yann LeCun, Yoshua Bengio and Geoffrey Hinton: Deep Learning (Nature, 2015)
34. Weibo Liu, Zidong Wang, Nianyin Zeng, Yurong Liu, Fuad E. Alsaadi: A survey of deep neural networks and their applications (Elsevier, 2017)
35. Convolutional Neural Networks for Speech Recognition IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 22, NO. 10, OCTOBER 2014
36. Bialek W, Rieke F, de Ruyter van Steveninck R. R, Warland D (1991) Reading a neural code. Science 252: 1854–1857.
37. Stanley G. B, Li F. F, Dan Y (1999) Reconstruction of natural scenes from ensemble responses in the lateral geniculate nucleus. J Neurosci 19: 8036–8042.
38. .Mesgarani N, David S. V, Fritz J. B, Shamma S. A (2009) Influence of context and behavior on stimulus reconstruction from neural activity in primary auditory cortex. J Neurophysiol 102: 3329–3339.
39. Ramirez A. D, Ahmadian Y, Schumacher J, Schneider D, Woolley S. M, et al. (2011) Incorporating naturalistic correlation structure improves spectrogram reconstruction from neuronal activity in the songbird auditory midbrain. J Neurosci 31: 3828–3842.
40. Jensen, J. & Taal, C. H. An Algorithm for Predicting the Intelligibility of Speech Masked by Modulated Noise Maskers. *IEEE/ACM Trans. Audio, Speech Lang. Process.* **24**, 2009–2022 (2016).
41. Yang, M. *et al*. Speech reconstruction from human auditory cortex with deep neural networks. In *Sixteenth Annual Conference of the International Speech Communication Association* (2015).
42. Hajinoroozi, M., Mao, Z., Jung, T.-P., Lin, C.-T. & Huang, Y. EEG-based prediction of driver's cognitive performance by deep convolutional neural network. *Signal Process. Image Commun.* **47**, 549–555 (2016).
43. Yang, X. & Shamma, S. A. W. K. Auditory representations of acoustic signals. *IEEE Trans. Inf. Theory* **38**, 824–839 (1992).
44. Chi, T., Ru, P. & Shamma, S. A. Multiresolution spectrotemporal analysis of complex sounds. *J Acoust Soc Am* **118**, 887–906 (2005)
45. Morise, M., Yokomori, F. & Ozawa, K. WORLD: a vocoder-based high-quality speech synthesis system for real-time applications.*IEICE Trans. Inf. Syst.* **99**, 1877–1884 (2016).
46. G. Huang, Z. Liu, L. Van Der Maaten and K. Q. Weinberger, "Densely Connected Convolutional Networks," *2017 IEEE Conference on Computer Vision and Pattern*

Recognition (CVPR), Honolulu, HI, 2017, pp. 2261-2269, doi: 10.1109/CVPR.2017.243.

47. Stevens S S, Volkmann J and Newman E B 1937 A scale for the measurement of the psychological magnitude pitch *J. Acoust. Soci. Am.* **8** 185–90
48. Van Den Oord A, Dieleman S, Zen H, Simonyan K, Vinyals O, Graves A, Kalchbrenner N, Senior A and Kavukcuoglu K 2016 Wavenet: a generative model for raw audio (arXiv:1609.03499)
49. Tamamori A, Hayashi T, Kobayashi K, Takeda K and Toda T 2017 Speaker-dependent wavenet vocoder *Proc. of Interspeech* pp 1118–22
50. F. Lotte, J.S. Brumberg, P. Brunner, A. Gunduz, A.L. Ritaccio, C. Guan, G.Schalk Electrocorticographic representations of segmental features in continuous speech Front. Hum. Neurosci., 9 (2015), p. 97
51. D. Carey, S. Krishnan, M.F.Callaghan, M.I. Sereno, F. Dick Functional and quantitative MRI mapping of somatomotor representations of human supralaryngeal vocal tract Cereb. Cortex, 27 (2017), pp. 265-278
52. Crone, N. E. et al. Electrocorticographic gamma activity during word production in spoken and sign language. *Neurology* **57**, 2045–2053 (2001).
53. Nourski, K. V. et al. Sound identification in human auditory cortex: differential contribution of local field potentials and high gamma power as revealed by direct intracranial recordings. *Brain Lang*. **148**, 37–50 (2015).
54. Pesaran, B. et al. Investigating large-scale brain dynamics using field potential recordings: analysis and interpretation. *Nat. Neurosci*. **21**, 903–919 (2018).
55. Chartier, J., Anumanchipalli, G. K., Johnson, K. & Chang, E. F. Encoding of articulatory kinematic trajectories in human speech sensorimotor cortex. *Neuron* **98**, 1042–1054 (2018).
56. Paul, B. D. & Baker, M. J. The design for the Wall Street Journal-based CSR corpus. In *Proc. Workshop on Speech and Natural Language* (Association for Computational Linguistics, 1992).
57. Kominek, J., Schultz, T. & Black, A. Synthesizer voice quality of new languages calibrated with mean mel cepstral distortion. In *Proc. The first workshop on Spoken Language Technologies for Under-resourced languages (SLTU-2008)* 63–68 (2008).
58. Bocquelet, F., Hueber, T., Girin, L., Savariaux, C. & Yvert, B. Real-time control of an articulatory-based speech synthesizer for brain computer interfaces. *PLOS Comput. Biol*. **12**, e1005119 (2016).

59. Ibayashi K, Kunii N, Matsuo T, Ishishita Y, Shimada S, Kawai K, Saito N (2018) Decoding speech with integrated hybrid signals recorded from the human ventral motor cortex. Front Neurosci 12:221.
60. C. Herff, L. Diener, M. Angrick, E. M. Mugler, M. C. Tate, M. Goldrick, D. Krusienski, M. W. Slutzky, and T. Schultz, "Generating natural, intelligible speech from brain activity in motor, premotor and inferior frontal cortices," Frontiers in Neuroscience, vol. 13, p. 1267, 2019.
61. Rabbani Q, Milsap G and Crone N E 2019 The potential for a speech brain–computer interface using chronic electrocorticography *Neurotherapeutics* **16** 1–22
62. Bocquelet F, Hueber T, Girin L, Chabardès S and Yvert B 2016 Key considerations in designing a speech brain– computer interface *J. Physiol.* **110** 392–401
63. Iljina, O. *et al*. Neurolinguistic and machine-learning perspectives on direct speech BCIs for restoration of naturalistic communication. *Brain-Computer Interfaces* **4**, 186–199 (2017).
64. Martin S, Millán J D R, Knight R T and Pasley B N 2016 The use of intracranial recordings to decode human language: challenges and opportunities *Brain Lang.*
65. Brumberg JS, Guenther FH. Development of speech prostheses: current status and recent advances. Expert Rev Med Devices. 2010; 7: 667–79.
66. S. Martin, I. Iturrate, J. Millan, R. Knight, and B. N. Pasley "Decoding Inner Speech Using Electrocorticography: Progress and Challenges Toward a Speech Prosthesis.," Front. Neurosci., vol. 12, pp. 422, 2018
67. http://bias.csr.unibo.it/maltoni/ml/ in particolare ho fatto riferimento alle dispense dalla 8 alla 10
68. https://www.sciencedirect.com/science/article/pii/S2405844018332067
69. https://www.sciencedirect.com/science/article/pii/B9780128197950000025
70. https://cs231n.github.io/convolutional-networks/
71. https://cs231n.github.io/neural-networks-1/
72. http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/