UNIVERSITÀ
DEGLI STUDI
DI PADOVA

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

CORSO DI LAUREA IN INGEGNERIA BIOMEDICA

"RILEVAMENTO PRECOCE DELLE MALATTIE DELLA VOCE: UN APPROCCIO BASATO SULL'ANALISI DEL SEGNALE VOCALE"

Relatore: Prof. Sarah Tonello

Laureanda: Maria Chiara Rigoni

ANNO ACCADEMICO 2022 – 2023

Data di laurea 21 Novembre 2023

Indice

Abstract 4
Introduzione 5
Capitolo 1: L'origine della voce 6
1.1 Anatomia dell'apparato fonatorio 6
1.2 Fisiologia: come si genera la voce 9
Capitolo 2: Caratteristiche del segnale voce 11
2.1 La voce come segnale biomedico 12
Capitolo 3: Patologie individuabili dall'analisi della voce 16
3.1 I biomarcatori vocali 17
Capitolo 4: Tipologie di sensori per l'acquisizione e l'analisi del segnale vocale 20
4.1 Microfoni 20
4.2 Sensori acustici vibrazionali morbidi 21
4.2.1 Sensori acustici vibrazionali morbidi resistivi 22
4.2.2 Sensori acustici vibrazionali morbidi capacitivi 23
4.2.3 Sensori acustici vibrazionali morbidi triboelettrici 24
4.2.4 Sensori acustici vibrazionali morbidi piezoelettrici 24
4.2.5 Altri sensori acustici vibrazionali morbidi 25
Capitolo 5: Principali tecniche di signal processing 26
5.1 Estrazione di parametri rilevanti 26
5.1.1 Analisi acustica 28
5.1.2 Analisi temporale: perturbazioni e fluttuazioni 28
5.1.3 Analisi spettrale 30
5.1.4 Analisi della complessità 31
5.1.5 Rappresentazione tridimensionale 31
5.2 Valutazione dei parametri 33
Capitolo 6: Un esempio da letteratura: un dosimetro per la valutazione e il monitoraggio dei disturbi vocali 34
Conclusioni e sviluppi futuri 38
Bibliografia 39

Abstract

Il presente elaborato è incentrato sull'analisi vocale come strumento diagnostico per la rilevazione precoce di patologie vocali. In una fase iniziale si indaga l'origine anatomica e fisiologica della voce, elemento cruciale per comprendere le caratteristiche del segnale vocale. Quest'ultimo viene successivamente trattato come un segnale biomedico dal quale ricavare importanti informazioni che riflettono lo stato di salute della persona. Inoltre vengono esaminate in dettaglio le patologie che possono essere individuate mediante l'analisi del segnale vocale, tra cui malattie proprie dell'apparato fonatorio e non, come ad esempio il morbo di Parkinson.

Vengono poi approfondite le principali tipologie di sensori necessari per acquisire il segnale vocale e le tecniche di signal processing, fondamentali per estrarre parametri significativi dai segnali allo scopo di diagnosticare e trattare le patologie tempestivamente.

Infine l'elaborato si conclude con l'analisi di un dosimetro vocale presente in letteratura: esso rileva e registra dati vocali in maniera non invasiva e per lunghi periodi di tempo, consentendo agli esperti di valutare l'andamento delle patologie vocali e di personalizzarne il trattamento.

Introduzione

La voce è una componente fondamentale della comunicazione umana, che consente all'uomo di esprimersi e di creare connessioni con gli altri. Essa è il risultato della complessa cooperazione tra organi e muscoli che, svolgendo la loro funzione e muovendosi sinergicamente, producono il suono. Oltre ad essere un importante mezzo di comunicazione, la voce è uno strumento che custodisce informazioni preziose riguardanti la salute fisica e mentale di una persona: infatti viene influenzata da fattori quali, per esempio, patologie, tensione e stress. Negli ultimi anni l'analisi vocale si è rivelata una risorsa preziosa per la diagnosi e il monitoraggio delle malattie vocali, le quali pur essendo di natura molto diversa si traducono in alterazioni e disturbi della voce. Proprio per tale motivo diviene essenziale comprendere le caratteristiche del segnale vocale, sottolineando che una rilevazione precoce e accurata delle patologie vocali consente di consente di effettuare un trattamento tempestivo e di prevenire complicazioni a lungo termine. Inoltre, l'analisi vocale, rispetto a procedure di individuazione delle patologie vocali già esistenti, si rivela uno strumento semplice, economico e non invasivo.

Nel contesto di questa tesi verranno esaminate le complesse origini anatomiche e fisiologiche della voce umana, verranno studiate le caratteristiche del segnale vocale e le principali patologie che possono essere individuate tramite la sua analisi. Successivamente si discuteranno alcune tipologie di sensori impiegate per raccogliere dati vocali ed alcune tecniche di signal processing, imprescindibili per l'estrapolazione dei parametri di interesse.

Infine, la ricerca terminerà con la presentazione di un prototipo di dosimetro per la valutazione e il monitoraggio dei disturbi vocali, che accomuna diverse tecnologie esistenti in maniera originale e innovativa, costituendo un'importante risorsa per medici e ricercatori che lavorano in questo settore.

Capitolo 1: L'origine della voce

La voce è il suono prodotto dalle corde vocali di un essere umano quando l'aria espirata passa attraverso di esse. È un tratto distintivo della nostra specie, nonché una parte essenziale del linguaggio e della comunicazione che permette l'espressione di pensieri, emozioni e idee attraverso il parlato e il canto.

1.1 Anatomia dell'apparato fonatorio

Dal punto di vista anatomico, la voce si origina dai polmoni, dove l'aria viene inspirata ed espirata, grazie a sua volta all'azione del diaframma, per passare successivamente attraverso le corde vocali presenti nella laringe e diventare un suono.

La laringe ricopre un ruolo di grande importanza in quanto è la parte più sensibile ed espressiva nel meccanismo di fonazione [1]. Essa è un organo che consiste in un condotto tubulare situato nella parte anteriore del collo e collega la faringe alla trachea, costituendo l'ultimo tratto delle vie aeree superiori. Può essere pensata come una cavità formata da strati di mucosa avvolti su uno scheletro di cartilagine e muscolo [2]. Lo spazio presente tra le corde vocali è denominato *glottide*: si individuano una regione superiore ad essa, o sopraglottica, ed una inferiore, o sottoglottica [1].

La laringe possiede una struttura rigida dotata di cartilagini che tramite legamenti e muscoli si collegano all'osso ioide, alla base del cranio e alla trachea. Vi sono quattro cartilagini laringee principali: tiroidea, cricoide, aritenoide accoppiata ed epiglottide. Le prime tre sono di natura ialina e forniscono un supporto rigido per le componenti mobili della laringe, mentre l'epiglottide, assieme ad altre tre cartilagini laringee minori, conferisce flessibilità essendo costituita da fibrocartilagine elastica. Le cartilagini sono unite tra loro mediante membrane e legamenti fibroelastici e sorgono al di sotto dell'osso ioide. Quest'ultimo è caratterizzato da una forma a ferro di cavallo e serve a sospendere ed ancorare la laringe durante i movimenti che si compiono durante la deglutizione e la fonazione [2, 3].

Strutture di tessuto connettivo danno origine a membrane e legamenti all'interno della laringe. I principali legamenti di questo organo sono i legamenti vocali, ovvero dei legamenti accoppiati che si estendono nella zona immediatamente sottostante all'epiglottide, collegati anteriormente alla cartilagine tiroidea e posteriormente al processo vocale dell'aritenoide. Tali legamenti rappresentano la base delle corde vocali. Lo spazio che si crea in corrispondenza dell'inserimento del legamento sulla parte anteriore della laringe è denominato *commessura anteriore* ed è un punto di riferimento nell'imaging tumorale; allo stesso modo, lo spazio che

si crea sulla parte posteriore è definito *commessura posteriore*, che risulta più ampio durante la respirazione tranquilla e il riposo. In realtà i legamenti vocali sono i margini più spessi di tessuto connettivo appartenente al cono elastico, un complesso composto da membrane elastiche collocate nella parte inferiore della laringe e alla cartilagine cricoide. Un'ultima categoria di legamenti laringei, è data dai legamenti ventricolari, che sono situati al di sopra dei legamenti vocali e vi scorrono parallelamente. Essi sono rivestiti dalle false corde vocali, ossia delle pieghe di una membrana mucosa. Quest'ultime si estendono tra l'aritenoide e la cartilagine tiroidea e si trovano al di sopra delle corde vocali vere così da proteggerle. Le corde vocali vere vengono lubrificate mediante muco che viene secreto dalle saccule laringee [2, 3].

Infine, una membrana formata da una lamina fibroelastica collega il margine superiore dell'osso ioide e la cartilagine tiroide: essa prende il nome di membrana tiroioidea (Figura 1.1) [2, 3].

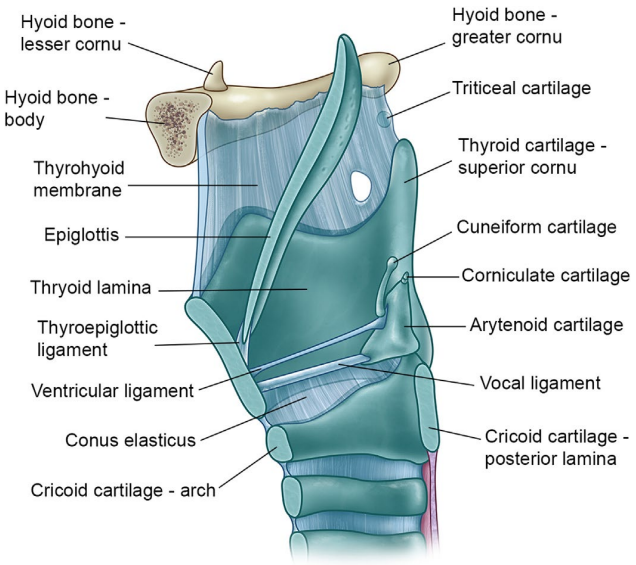

Figura 1.1: Visione trasversale mediana sagittale della laringe che mostra cartilagini, legamenti e membrane.

Per quanto concerne la muscolatura laringea, essa si divide in due categorie: i muscoli intrinseci e i muscoli estrinseci. I primi sono interamente collocati all'interno della laringe, facilitano lo scorrimento delle cartilagini laringee l'una contro l'altra e originano il movimento delle corde vocali; i secondi, invece, si estendono a partire dal collo, sono attaccati all'osso ioide e svolgono le funzioni di alzare, abbassare, nonché stabilizzare la laringe. Tra i muscoli estrinseci si trova il muscolo cricotiroideo o muscolo vocale che, assieme ai legamenti vocali, forma le corde vocali vere. Muscoli, cartilagini e legamenti della

laringe sono ricoperti da una mucosa costituita da un epitelio ciliare colonnare pseudo-stratificato (Figura 1.2) [2, 3].

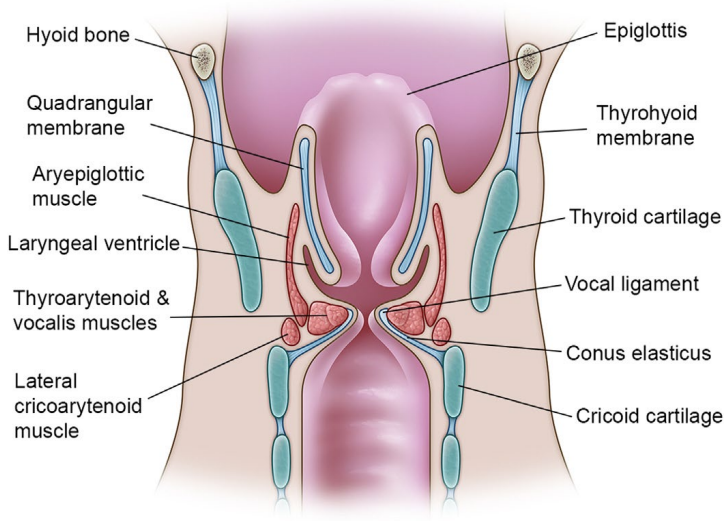

Figura 1.2: Visione coronale trasversale della laringe raffigurante le relazioni tra cartilagini, muscoli e legamenti con le superfici mucose della laringe.

La laringe è innervata dai nervi laringei superiori e dai nervi laringei ricorrenti. I primi si originano dal nervo vago e si dividono in nervi laringei interni ed esterni in base alla loro collocazione rispetto all'osso ioide. I nervi laringei interni conferiscono innervazione sensoriale, mentre quelli esterni forniscono innervazione motoria ai muscoli faringeo inferiore e cricotiroideo. I nervi laringei ricorrenti, invece, sono adibiti all'innervazione sensoriale della parte di laringe inferiore alla glottide ed ai muscoli intrinseci della laringe, escluso il muscolo cricotiroideo [3].

Le corde vocali sono in grado di produrre il suono grazie alla vibrazione, l'una contro l'altra, di due loro estremità libere. La larghezza e le dimensioni dello spazio tra le corde vocali, ossia della glottide, variano in base alla respirazione e alla parola. Le corde vocali sono costituite da diverse parti, ma principalmente si individuano: uno strato muscolare, comprendente le fibre del muscolo tiroaritenoide, che a sua volta forma una parte del legamento vocale; una mucosa propria delle corde vocali, coperta da cellule epiteliali squamose stratificate, sotto le quali si trova uno strato di epitelio colonnare e quindi la lamina propria. Quest'ultima è composta da proteine extracellulari generate dai fibroblasti, che conferiscono la flessibilità e la viscosità necessarie per la vibrazione e il movimento delle corde vocali [3].

Infine, la laringe è vascolarizzata ad opera delle arterie tiroidea superiore e tiroidea inferiore, e ad essa convergono numerosi vasi linfatici [2].

Oltre alla laringe, anche il tratto vocale sopraglottico riveste un ruolo fondamentale nell'anatomia della voce. Infatti la laringe sopraglottica, le labbra, la lingua, il palato, la faringe e i seni nasali consentono al suono che si origina dalle corde vocali di essere amplificato e modellato [1].

1.2 Fisiologia: come si genera la voce

La fisiologia della voce è il meccanismo attraverso il quale la voce viene prodotta dall'uomo e si rivela parecchio complesso. La laringe svolge principalmente quattro funzioni in questo processo: la fonazione, in quanto genera il suono e la voce; la respirazione, consentendo il passaggio dell'aria nei polmoni; si eleva facilitando la deglutizione; infine protegge i polmoni dall'ingresso di oggetti estranei [3].

La fonazione è il risultato della complessa interazione tra la generazione di forza per consentire il passaggio del flusso d'aria nei polmoni, la modifica dello stesso ad opera della laringe e la risonanza delle cavità faringea, orale e nasale, che definisce in modo unico la voce di una persona. Grazie all'azione coordinata di diaframma, muscoli addominali e del torace, e della gabbia toracica, si genera un flusso d'aria che viene modulato dalle corde vocali, le quali vengono spostate al centro della laringe dalla muscolatura laringea, dai nervi e dalle cartilagini [3].

Durante la fonazione si verifica il ciclo glottico, che si articola in quattro fasi: chiusa, di apertura, aperta e di chiusura. Nella fase chiusa, l'aria viene trattenuta al di sotto delle corde vocali, che sono chiuse, e la pressione dell'aria inizia ad aumentare. Nella fase di apertura, la pressione supera una certa soglia e le corde vocali iniziano ad aprirsi: l'aria inizia a fluire attraverso la glottide e si origina il suono. Successivamente, nella fase aperta, le corde vocali sono completamente aperte e l'aria fluisce attraverso di loro: si produce così un vero e proprio suono, ossia la voce. Infine, con la fase di chiusura, la pressione subglottica diminuisce, a causa della precedente fuoriuscita d'aria, e le corde vocali iniziano a chiudersi nuovamente. A questo punto il ciclo viene ripetuto [3].

La comprensione della fisiologia della fonazione è stata possibile grazie a due teorie. La prima è la teoria mioelastica-aerodinamica: l'originale, risalente al 1958, stabilisce che l'aria, passando attraverso il lume stretto delle corde vocali, aumenti la sua velocità secondo il principio di continuità, causando una diminuzione di pressione secondo l'equazione di Bernoulli. Questa pressione più bassa, combinata con l'elasticità della copertura delle corde vocali, provoca una chiusura delle stesse, a partire dalla porzione inferiore. Tuttavia, ulteriori ricerche hanno dimostrato che durante la fonazione, la pressione subglottica supera la

pressione sopraglottica, spingendo le corde vocali inferiori a separarsi. Le differenze di fase verticale tra le parti inferiori e superiori delle corde creano un'onda che si muove verso l'alto lungo la loro superficie mediale. L'elasticità delle corde vocali provoca poi la loro chiusura inferiore, riducendo la pressione intraglottica e facendo chiudere le corde vocali. Questo ciclo si ripete grazie alla loro elasticità. La seconda teoria che spiega la fonazione, afferma che, essendo il corpo delle corde vocali costituito prevalentemente da muscolo tiroaritenoideo e la loro copertura dall'epitelio e dalla lamina propria superficiale, le due componenti possiedono proprietà biomeccaniche distinte, che influenzano la vibrazione. In realtà la vibrazione di una singola struttura è il risultato della vibrazione di strati multipli con differenze di fase, dovute a un corpo rigido e a una copertura flessibile [3].

Per una fonazione efficace, occorrono cinque requisiti essenziali: supporto respiratorio e forza espiratoria sufficienti, posizionamento e allineamento corretti delle corde vocali, elasticità delle corde vocali per agevolare la vibrazione, una forma idonea delle corde vocali e la capacità di controllare le loro tensione e lunghezza [3].

Il suono prodotto dalla glottide viene modellato mentre l'aria scorre attraverso le varie strutture della laringe. Le caratteristiche proprie della voce del singolo individuo sono influenzate da torace, faringe e cavità orale e nasale, la cui forma influisce sulla filtrazione e amplificazione del suono, che può essere a sua volta modificato dal movimento di palato, laringe, faringe, lingua e mandibola [3].

Capitolo 2: Caratteristiche del segnale voce

Oltre ad essere un importante mezzo di comunicazione, la voce costituisce un prezioso segnale che porta con sé molte informazioni e caratteristiche distintive. Quest'ultime rendono ogni voce unica e consentono di riconoscere chi sta parlando. In aggiunta, la voce dà informazioni sulla salute generale e il benessere dell'individuo, oltre che sulla struttura e sulla funzione di alcune parti del corpo [5].

Nel segnale vocale si verificano perturbazioni derivanti dai rumori di fondo del corpo umano, che ne fanno variare frequenza, ampiezza e forma d'onda. È possibile pensare che la voce venga prodotta da un sistema di oscillatori biomeccanici, neurali ed acustici: il segnale portante è prodotto dalle corde vocali, che sono gli oscillatori principali; mentre ad esempio la laringe, la respirazione e il battito cardiaco rappresentano gli oscillatori secondari (Figura 2.1). Tale sistema di oscillatori fornisce informazioni riguardanti genetica, sviluppo, età, malattie, lingua, cultura, alimentazione e assunzione di farmaci, e risposta all'ambiente del corpo umano (Figura 2.1) [5].

Le oscillazioni sono tante, così come le informazioni che queste apportano, per cui la loro analisi è particolarmente complessa [5].

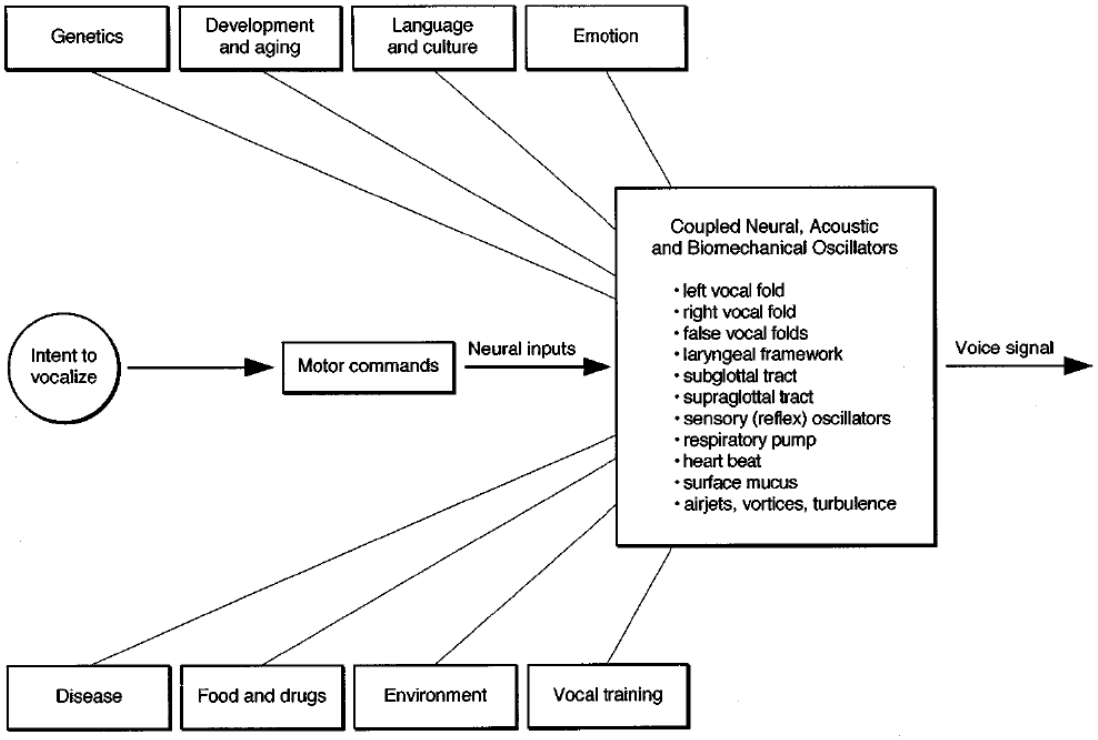

Figura 2.1: Oscillatori biologici coinvolti nella produzione vocale e fattori che possono influenzarli.

Vi sono alcune caratteristiche che consentono di descrivere il segnale vocale, tra cui perturbazione, fluttuazione, variabilità, tremore e vibrato vocale. La *perturbazione* è un cambiamento temporaneo del comportamento previsto dal segnale, che nel caso della voce è sempre presente ed è spesso dovuto ai rumori fisiologici di fondo del corpo umano [5, 6]. Una deviazione significativa dal modello del segnale vocale costituisce una *fluttuazione*, che si distingue dalla perturbazione perché, a differenza di quest'ultima, non consente di riportare il sistema alla stabilità. La *variabilità* è definita come la quantità di variazione determinata da una misura statistica nel segnale vocale: essa apporta un cambiamento, positivo o negativo, nel risultato finale del processo vocale. Il *tremore* è una fluttuazione a bassa frequenza nell'ampiezza o nella frequenza della voce, spesso ricondotta a cause neurologiche. Se da un lato essa può contribuire a rendere la voce piacevole dall'altro un tremore eccessivo è considerato patologico. Infine, il *vibrato vocale*, la cui origine non è ancora completamente nota, è ipotizzato come un tremore fisiologico della muscolatura laringea [5].

2.1 La voce come segnale biomedico

Un segnale è una funzione del tempo che rappresenta il variare di una grandezza fisica e che possiede specifiche caratteristiche nei domini di tempo e frequenza. Il segnale vocale, in base alla propria natura, viene classificato in tre tipologie: tipo 1, tipo 2 e tipo 3. I segnali di tipo 1 sono quasi periodici e risulta vantaggioso analizzarne le perturbazioni; quelli di tipo 2 contengono intermittenze, frequenze subarmoniche o frequenze di modulazione e richiedono un'analisi visiva del contenuto in frequenza mediante spettrogrammi; quelli di tipo 3 sono di natura stocastica e vengono esaminati nel dominio del tempo. Ad esempio, il segnale derivante dalla pronuncia di una vocale sostenuta è quasi periodico, dunque in condizioni normali è di tipo 1 (Figura 2.2) [5].

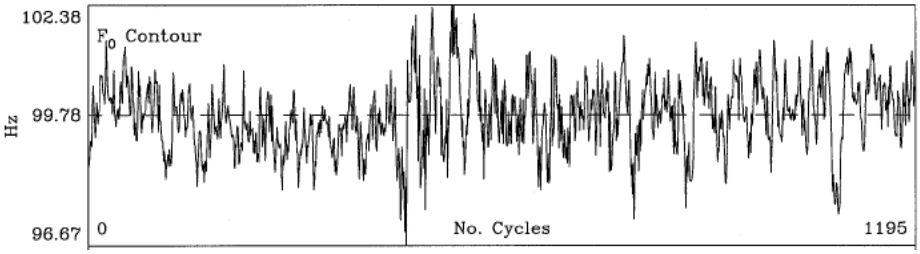

Figura 2.2: Variazione della frequenza fondamentale F_0 durante la fonazione continua di una vocale per un intervallo di 12 secondi circa. La F_0 media è di 99.78 Hz.

Dunque essendo la voce un segnale, essa consente di estrarre utili parametri nel tempo e nella frequenza. Nel dominio della frequenza si ricavano parametri come la frequenza fondamentale, le armoniche e lo spettro in frequenza. La *frequenza fondamentale* F_0 è un parametro acustico fondamentale, nonché il più importante, ed è definito come il numero totale di cicli per secondo eseguiti dalle corde vocali [5]. Le *armoniche* sono multipli interi della frequenza fondamentale, che nel caso del segnale vocale sono dovute alle vibrazioni delle corde vocali. Dal contenuto in frequenza del segnale è possibile estrarre lo spettro, ossia il contenuto in frequenza di un segnale: quest'ultimo viene spesso rappresentato mediante uno spettrogramma, un grafico avente la frequenza sull'asse delle ordinate, il tempo sull'asse delle ascisse e diversi colori o tonalità che indicano l'intensità delle componenti in frequenza. Nella *Figura 2.3* è possibile osservare lo spettrogramma di un paziente con disfonia iperfunzionale infantile: la frequenza fondamentale è all'incirca di 300 Hz e il segnale è di tipo 2 [5].

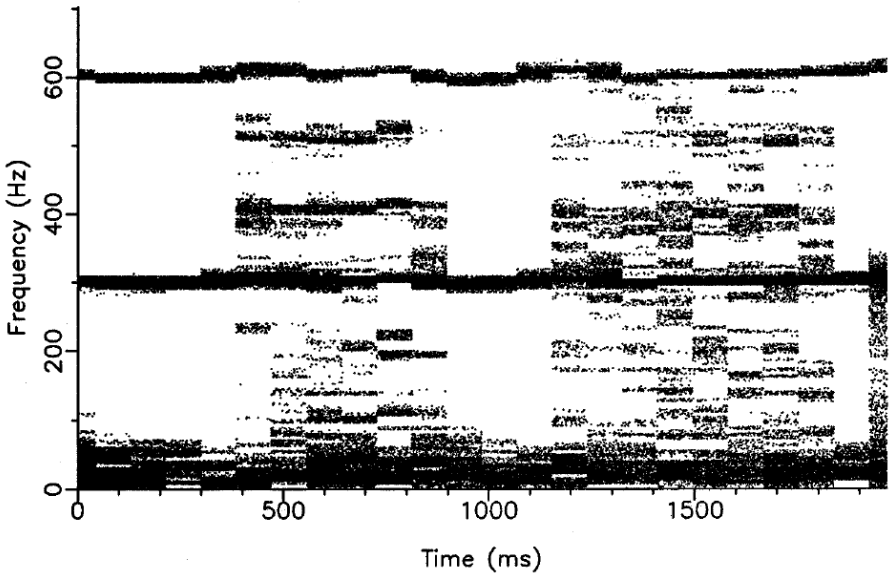

Figura 2.3: Spettrogramma di un paziente don disfonia iperfunzionale infantile.

Nel dominio del tempo si ottengono informazioni riguardanti il periodo fondamentale, il tempo di fonazione, lo jitter e lo shimmer e il rapporto armoniche-rumore. Il *periodo fondamentale* è l'inverso della frequenza fondamentale. Per verificare la salute delle corde vocali si indaga il tempo di fonazione massimo in secondi, definito come la durata massima in cui un individuo può sostenere un suono vocalico. Tale indicatore esprime l'efficienza glottica, ossia la capacità di addurre le corde vocali per tutta la loro lunghezza in maniera efficace, che è importante sia per la fonazione normale e sana, sia per proteggere il passaggio dell'aria da particelle estranee. Per svolgere adeguatamente questa funzione è necessario

fornire una specifica tensione alle corde vocali, che durante la fonazione vibrano tra le cento volte al secondo per gli uomini adulti e le duecento volte al secondo per le donne adulte [6].
Per descrivere le caratteristiche vocali si rivela altresì utile lo *shimmer*, che è una perturbazione dell'ampiezza da un ciclo all'altro delle corde vocali, percepita come un suono crepitante o ronzante; in alcuni casi in cui l'effetto è più marcato può diventare molto sgradevole. Il *jitter* indica una variazione, o un'irregolarità, nella frequenza fondamentale della voce da ciclo a ciclo delle corde vocali. Questi due parametri forniscono informazioni importanti sulla salute vocale: nei pazienti con patologie alle corde vocali il *jitter* è più elevato; mentre lo *shimmer* è alterato in pazienti con ridotta resistenza glottica, lesioni alle corde vocali e affaticamento della voce [5, 6].
Il jitter viene misurato tramite altri sotto-parametri più specifici. Il jitter assoluto è la variazione della frequenza F_0 ad ogni ciclo; il jitter relativo è la differenza assoluta media tra i periodi consecutivi, divisa per il periodo medio [7]; il jitter RAP è la perturbazione media relativa, calcolata come la differenza assoluta media tra un periodo e la media di esso e dei suoi due vicini, divisa per il periodo medio [7]; il jitter PPQ5 è definito come il quoziente di perturbazione del periodo a cinque punti ed è calcolato come la differenza assoluta media tra un periodo e la media di esso e dei suoi quattro elementi più vicini, divisa per il periodo medio [7]. Il jitter relativo, RAP e PPQ5 sono espressi in forma percentuale. Infine il jitter DDP è la differenza assoluta media tra i periodi consecutivi, divisa per il periodo medio. Il suo valore è tre volte quello del jitter RAP [6, 7].
Dall'analisi di una vocale pronunciata in maniera sostenuta per un intervallo di tempo sufficientemente lungo è possibile misurare lo shimmer. Anche in questo caso si utilizzano dei sotto-parametri specifici che agevolano la misura dello stesso. Vi è dunque lo shimmer locale, ossia la differenza assoluta media tra le ampiezze dei periodi consecutivi, divisa per l'ampiezza media [7]. Lo shimmer in dB esprime la variabilità dell'ampiezza da picco a picco [7]; lo shimmer APQ3 è il quoziente di perturbazione dell'ampiezza a tre punti, la differenza assoluta media tra l'ampiezza di un periodo e la media delle ampiezze dei suoi vicini, divisa per l'ampiezza media [7]. Lo shimmer APQ5 è il quoziente di perturbazione dell'ampiezza a cinque punti, ovvero la differenza assoluta media tra l'ampiezza di un periodo e la media delle ampiezze dei suoi quattro vicini più vicini, divisa per l'ampiezza media. Lo shimmer DDA è la differenza assoluta media tra le differenze consecutive, tra le ampiezze dei periodi consecutivi [7]. Shimmer locale, APQ3 ed APQ5 sono in forma percentuale [6, 7].

Infine, molto utilizzato è l'*HNR*, ossia il rapporto armoniche-rumore. Esso fornisce la periodicità generale del segnale vocale calcolando il rapporto tra le componenti periodiche e quelle aperiodiche; è calcolabile solo per segnali periodici, come una vocale sostenuta [6].

Capitolo 3: Patologie individuabili dall'analisi della voce

Come si è potuto vedere nel *Capitolo 2*, la voce consente di ricavare diversi parametri acustici, utili per la sua analisi. In campo medico, grazie ad essa, si possono riconoscere nel paziente patologie che riguardano l'apparato fonatorio oppure patologie neurologiche, funzionali, strutturali e psicogene che si riflettono in disturbi vocali. L'individuazione precoce delle malattie dell'apparato vocale, in particolare, è molto importante in quanto esse necessitano di un trattamento urgente [6, 8].

Una delle tecniche attualmente adottate per riconoscere la presenza di anomalie alle corde vocali è la laringoscopia diretta, un metodo costoso, invasivo e scomodo. Esso prevede la visualizzazione delle corde vocali mediante una telecamera apposita che viene posta nella gola della persona, motivo per cui spesso è richiesto l'utilizzo di un anestetico locale. In alternativa vi è la laringoscopia indiretta, che svolge lo stesso test di quella diretta, ma mediante uno specchio: è sicuramente un metodo molto meno invasivo, ma anch'esso necessita di piccole quantità di sedativi locali e l'attrezzatura è abbastanza costosa. Vi è poi l'elettroglottografia (EGG) che rileva le vibrazioni delle corde vocali mediante elettrodi posti sulla superficie del collo ed è spesso associata ad un'analisi acustica [8, 9].

Tuttavia, le tecniche appena citate sono sempre precedute da uno screening effettuato dal medico tramite mezzi non strumentali, che si rivela parecchio difficoltoso e poco affidabile a causa della somiglianza negli effetti uditivi di patologie differenti e del background e dell'esperienza del medico che svolge la visita [8]. Proprio per questo motivo si sta incoraggiando sempre più l'impiego di misurazioni oggettive della voce, anche e soprattutto, nello screening. L'analisi vocale si presta bene a tale scopo, essendo un metodo semplice, efficiente, economico e non invasivo: non prevede attrezzature particolari, quindi rende lo screening più accessibile; in aggiunta, dall'indagine di segnali audio semplici è possibile ottenere immediatamente risultati precisi e accurati, che consentono una valutazione sia iniziale, sia complementare per la diagnosi di alcune malattie vocali [6, 8, 9].

Tra le principali malattie riconoscibili dalla voce vi sono noduli, edema e paralisi alle corde vocali. Quando esse vibrano, nelle zone di impatto dove l'attrito è maggiore si possono formare lesioni benigne che possono portare alla formazione di noduli. Questi sono delle dimensioni della testa di uno spillo, possono essere a gruppi e impediscono la completa chiusura delle corde vocali, motivo per cui la loro presenza si manifesta in disfonie, cambiamenti di voce e raucedine; nei casi più estremi si può arrivare altresì alla totale afonia. Per compensare i sintomi la persona tende ad aumentare la tensione muscolare e, di

conseguenza, le collisioni tra le corde vocali. I noduli sono molto comuni nei soggetti che utilizzano in maniera intensa la voce, ad esempio insegnanti, cantanti e bambini che urlano spesso [10].

L'edema delle corde vocali provoca un aumento del volume delle corde vocali, che altera la loro elasticità e, successivamente, il timbro della voce della persona. Tale patologia deriva da un abuso della voce e, in alcuni casi, il soggetto può essere affetto anche da noduli; cause più esterne si riconducono al consumo di tabacco o droghe, dalla tosse eccessiva, dalle mestruazioni, dalla menopausa o dalla gravidanza [10].

Il sistema nervoso controlla i muscoli della laringe, che a loro volta regolano tensione e posizione delle corde vocali. Quando la muscolatura laringea non è più in grado di svolgere il suo compito, è presente una paralisi alle corde vocali, un disturbo che ha origine neurologica periferica. Solitamente questa patologia colpisce un'unica corda vocale, il che comporta vibrazioni a frequenze differenti per ciascuna corda: il paziente perde la capacita di amplificazione vocale, dunque non può parlare ad alta voce, la quale assume un suono bimodale [10].

Ulteriori patologie sono individuabili grazie ai biomarcatori vocali e verranno discusse nel seguente paragrafo.

3.1 I biomarcatori vocali

Un biomarcatore è un fattore misurato e valutato oggettivamente che rappresenta un processo biologico o patogeno, o una risposta farmacologica a un intervento terapeutico [11]. In particolare un biomarcatore vocale è un'impronta, una caratteristica, o una combinazione di caratteristiche del segnale vocale. Esso può essere utilizzato per diagnosticare una condizione o valutare e monitorare una malattia o lo sviluppo di farmaci. È importante che il biomarcatore vocale possieda tutte le proprietà di un biomarcatore tradizionale, convalidate analiticamente e qualificate tramite una valutazione delle prove [11].

L'impiego di biomarcatori si sta rivelando particolarmente utile in diversi ambiti patologici, tra i quali i principali includono alcune malattie neurodegenerative, malattie infiammatorie, stress post-traumatici, malattie mentali o disturbi correlati, malattie dell'apparato circolatorio e malattie dell'apparato respiratorio.

Nel contesto della diagnosi di malattie neurodegenerative, al giorno d'oggi si stanno svolgendo studi sui biomarcatori vocali per l'individuazione, ad esempio, di morbo di Parkinson, malattia di Alzheimer, declino cognitivo lieve e sclerosi multipla. Ad esempio nel morbo di Parkinson sono frequentemente presenti numerosi disturbi vocali, generati da

anomalie nella fonazione e nell'articolazione, che a loro volta provocano variazioni di tono, diminuzione dell'energia nelle parti superiori dello spettro armonico e articolazione imprecisa di vocali e consonanti, con una conseguente diminuzione dell'intelligibilità. Si prevede che tali variazioni della voce, dunque, possano essere impiegate come biomarcatori vocali diagnostici precoci, o come indicatori della progressione della malattia, in modo da poter integrare in futuro l'esame manuale per valutare i sintomi del morbo o monitorare l'efficacia del trattamento. Tuttavia, spesso sia il paziente che il medico trascurano, almeno in una fase iniziale, i cambiamenti vocali, che si verificano nel 78% dei pazienti che soffrono di Parkinson [11].

Le persone affette da malattia di Alzheimer o da declino cognitivo lieve, invece, dialogano in maniera incerta, parlano con un ritmo discorsivo lento e spesso incontrano difficoltà a trovare le parole; si ripetono frequentemente, utilizzano termini indefiniti e via via perdono le capacità semantiche in generale. In aggiunta spesso il discorso è poco coerente, molto semplice, con dettagli irrilevanti. Tutte queste caratteristiche potenzialmente possono diventare biomarcatori vocali per la diagnosi precoce di condizioni patologiche associate alla demenza: basti pensare che alterazioni nella voce e nel linguaggio possono essere percepite anni prima dell'insorgenza della malattia di Alzheimer o del declino cognitivo lieve [11].

Nella sclerosi multipla, che provoca disturbi vocali e disartria, al fine di monitorare il suo decorso vengono controllati periodicamente articolazione, respirazione e prosodia: ciò serve altresì ad individuare il momento più adeguato per avviare trattamenti come la stimolazione cerebrale [11].

Al di fuori delle malattie neurodegenerative l'impiego di biomarcatori può risultare potenzialmente utile anche per la diagnosi di malattie infiammatorie croniche, come l'artrite reumatoide. In questa patologia, infatti, si riscontrano cambiamenti patologici nelle corde vocali man mano che la malattia avanza, perciò il controllo delle caratteristiche della qualità della voce consente il monitoraggio delle condizioni del paziente [11].

Anche lo stress influisce sulla fonazione e la risposta da parte dell'organismo varia molto da individuo a individuo, spesso si traduce in cambiamenti della frequenza fondamentale F_0 del segnale vocale, fino a giungere a frequenze di 700 Hz nelle voci femminili. Inoltre all'aumentare dello stress si innalza il tono vocale, quindi predominano le frequenze più alte e aumentano energia del segnale e variazioni di frequenza [4]. Le alterazioni della voce sono maggiormente ricorrenti in persone con alti livelli di cortisolo, tipici della depressione. Le variazioni vocali vengono dunque utilizzate anche per diagnosticare quest'ultima, oppure per stimarne la gravità. Tramite uno smartphone si sono svolti dei test per valutare la voce in

soggetti affetti da depressione e per verificare l'efficacia dei trattamenti, o ancora per attribuire un punteggio di salute mentale per il disturbo da stress post-traumatico. Infine si è notata una correlazione positiva tra i livelli di stress e la durata dell'interazione verbale [11].
Inoltre si riscontra che le malattie mentali influenzano gli aspetti linguistici della voce. Ad esempio il discorso è incoerente, disorganizzato, vi sono digressioni e associazioni insensate nel caso della schizofrenia; le persone con disturbo bipolare e di personalità istrionica tendono invece a dilungarsi molto e a sottolineare dettagli superflui. Un altro aspetto su cui si sta ponendo sempre più attenzione è il riconoscimento delle emozioni mediante la voce, che è utile per il monitoraggio del paziente durante le visite e per raccogliere informazioni della vita reale in studi clinici o epidemiologici [11].
Nell'ambito di ricerca delle malattie dell'apparato circolatorio, un team della Mayo Clinic ha indagato la correlazione tra malattie alle arterie coronarie e voce. Si è anche dimostrato che le persone affette da diabete di tipo 2 con scarsa regolazione glicemica o neuropatia sono caratterizzate da maggior affaticamento vocale, debolezza della voce e raucedine [11].
In conclusione, per quanto concerne le patologie respiratorie, la recente pandemia da COVID-19 ha stimolato l'attività di ricerca ad utilizzare i suoni respiratori come biomarcatori. Il COVID-19 è una condizione respiratoria che influenza la respirazione e la voce, causando, tra gli altri sintomi, tosse secca, mal di gola, voce eccessivamente affannosa e schemi di respirazione tipici [11]. Questi sintomi consentono di rendere le voci dei pazienti malati riconoscibili, tanto che si sono studiati algoritmi per prevedere la presenza di un'infezione da SARS-COV-2 o, alternativamente, per valutare la gravità della malattia. Grazie all'Università di Cambridge e agli scienziati del MIT si stanno ottenendo risultati promettenti e in alcuni progetti si sta cercando di analizzare la tosse per di sviluppare un sistema di valutazione del rischio di infezione da COVID-19 basato su robot [11].

Capitolo 4: Tipologie di sensori per l'acquisizione e l'analisi del segnale vocale

I sensori ricoprono una funzione essenziale nel rilevare un segnale vocale affidabile e preciso e, di conseguenza, dei parametri accurati, dai quali è possibile individuare la presenza o meno di patologie legate alla voce. Il progresso tecnologico ha portato alla creazione di sensori di acquisizione del segnale vocale di vario tipo: essi spaziano dai comuni microfoni, a quelli integrati negli smartphone, fino ai sensori acustici vibrazionali, che trasformano le vibrazioni acustiche in un altro segnale [12]. In questo modo il monitoraggio della propria salute avviene semplicemente parlando al telefono, oppure senza rendersi conto di un minuscolo sensore che sorge sulla superficie della propria cute. Tali dispositivi offrono il vantaggio di essere non invasivi e facili da reperire, incentivando così il controllo del proprio stato di salute, che si rivela più accessibile delle comuni procedure di individuazione delle patologie vocali.

4.1 Microfoni

Attualmente molti dispositivi moderni sono capaci di misurare un segnale vocale in tempo reale e di analizzarlo. Infatti, come detto in precedenza, è possibile utilizzare direttamente il microfono del proprio smartphone per effettuare misure riguardanti la voce: esistono già applicazioni che adempiono a questo compito [13]. In un moderno smartphone sono contenuti più microfoni posizionati in parti differenti del dispositivo, che cooperano per cancellare il rumore e registrare suoni. Ad esempio in *Figura 4.1* viene indicata la collocazione dei quattro microfoni integrati in un iPhone: uno nella parte anteriore, uno nella parte posteriore e due nella parte inferiore. La presenza di più microfoni consente di recepire il suono anche se l'utente cambia posizione dello smartphone mentre parla, offrendo quindi maggiori sensibilità ed accuratezza [13].

Figura 4.1: Posizione dei microfoni integrati in un iPhone.

In generale la geometria e la dimensione del microfono influenzano la sua sensibilità direzionale: essa si manifesta quando le dimensioni del dispositivo si avvicinano o superano un quarto della lunghezza d'onda del suono percepito; al contrario se il microfono è di dimensioni infinitesime, la direzione del suono non ne determina in alcun modo la sensibilità. Si individuano due classi di microfoni: quelli a campo libero, che quando sono rivolti verso la sorgente di onde piane forniscono una risposta più piatta, ossia indipendente dalla frequenza; quelli a campo diffuso, che sono progettati per ottenere una risposta in frequenza piatta quando il suono proviene da tutte le direzioni contemporaneamente [13].

4.2 Sensori acustici vibrazionali morbidi

I sensori morbidi sono una tipologia di sensori realizzati in materiali sottili ed elastici, che conferiscono grande flessibilità e sono in grado di adattarsi al meglio alla superficie del corpo umano, sulla quale vengono posizionati. Essi sfruttano nanomateriali che forniscono altresì elevati livelli di precisione e accuratezza, data la loro capacità di rilevare stimoli esterni impercettibili. Alla categoria dei sensori morbidi appartengono i sensori acustici vibrazionali, che convertono onde acustiche analogiche appartenenti ad un vasto range di frequenze in dati interpretabili e analizzabili [12].

Le prestazioni dei sensori vibrazionali acustici dipendono da: sensibilità, linearità, gamma dinamica, larghezza di banda, piattezza della risposta in frequenza e frequenza di risonanza. La sensibilità dei sensori acustici vibrazionali morbidi riflette la loro capacità di convertire

onde acustiche in segnali elettrici ed è misurata come il rapporto tra la variazione del segnale elettrico e la variazione di pressione o accelerazione. La linearità dei sensori è importante per una facile elaborazione dei dati e viene spesso valutata utilizzando il coefficiente di determinazione (R^2). Il rapporto segnale-rumore (SNR) è essenziale per valutare la qualità dei sensori nelle applicazioni pratiche e si misura in decibel (dB): un alto SNR significa maggior precisione delle misurazioni e affidabilità del sensore. La gamma dinamica dei sensori acustici vibrazionali morbidi rappresenta l'intervallo tra le intensità di vibrazione più alte e più basse che possono essere rilevate dai sensori. La larghezza di banda, d'altra parte, è l'intervallo tra le frequenze più alte e più basse che i sensori misurano senza una significativa perdita di sensibilità. La piattezza della risposta in frequenza indica il mantenimento uniforme, da parte dei sensori, della sensibilità in tutto il range di frequenze in ingresso. La frequenza di risonanza, invece, è quella in cui i sensori raggiungono la massima sensibilità: se essa si trova all'interno della larghezza di banda può causare alterazioni della sensibilità vicino alla risonanza [12].
I sensori acustici vibrazionali basati su diversi meccanismi di trasduzione, in base ai quali vengono solitamente classificati.

4.2.1 Sensori acustici vibrazionali morbidi resistivi

I sensori resistivi sfruttano l'effetto resistivo per cui cambiamenti nelle aree di contatto tra gli strati conduttivi generano variazioni nella resistenza. In base all'aumento o alla diminuzione della contiguità tra i riempitivi conduttivi la resistenza aumenta o diminuisce rispettivamente. L'entità e la frequenza dei cambiamenti rappresentano ampiezza e frequenza della vibrazione individuata. Questi sensori sono semplici da fabbricare, economici, facilmente integrabili in altri dispositivi e permettono di leggere facilmente il segnale. D'altra parte, però, forniscono un output non lineare, necessitano di un'alimentazione esterna, possiedono una risposta in frequenza non uniforme e una larghezza di banda limitata [12].
Alcuni esempi di materiali utilizzati per la realizzazione dei sensori resistivi sono: l'ossido di grafene ridotto (rGO), il grafene e le nanoparticelle metalliche sottoforma di film in cui sono presenti delle micro-crepe; il grafene indotto con laser (LIG), prodotto a partire dal poliimmide; il grafene verticale (VGr) ad alta densità di crepe anch'esso sottoforma di film. Inoltre per la produzione di sensori acustici vibrazionali morbidi sono stati impiegate strutture particolari e innovative, come i nanofili d'oro orientati verticalmente (AuNWs) con micro-crepe in grado di rilevare con notevole sensibilità le vibrazioni acustiche (Figura 4.2) [12].

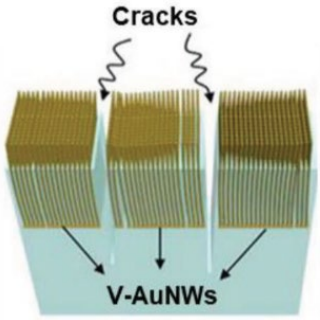

Figura 4.2: Schema del sensore in VAuNW con micro-crepe.

4.2.2 Sensori acustici vibrazionali morbidi capacitivi

I sensori acustici vibrazionali capacitivi misurano le variazioni di pressione sonora mediante cambiamenti nella capacità, causati dalla deformazione della loro struttura: il diaframma. Questo costituisce il meccanismo di trasduzione che prende il nome di *effetto capacitivo*, per il quale i cambiamenti rilevati dal sensore vengono convertiti in una corrente elettrica, grazie ad una tensione di polarizzazione fissa. I vantaggi di questa tipologia di sensori sono offrire un output lineare, durare nel tempo, fornire una risposta in frequenza uniforme e presentare un'ampia larghezza di banda; tuttavia hanno bassa sensibilità e necessitano di un amplificatore operazionale che potenzi il segnale in uscita [12].

La struttura a diaframma dei sensori capacitivi morbidi può essere realizzata introducendo fori di ventilazione al centro (Figura 4.3), come nel caso del sensore *MEMS* (Micro Electro-Mechanical Systems) il quale ha dimostrato buone prestazioni di variazioni nella pressione acustica. Altri sensori impiegano diaframmi ottenuti tramite fotolitografia oppure caricati elettricamente: i primi conferiscono elevate prestazioni e una risposta in frequenza ampia, i secondi una buona sensibilità e non hanno bisogno di alcuna fonte di alimentazione [12].

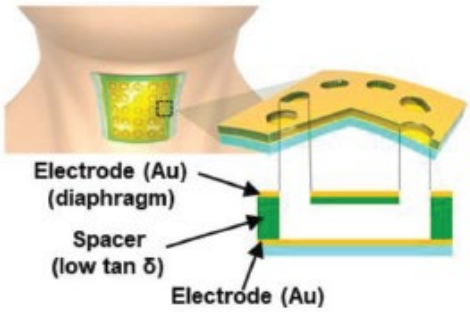

Figura 4.3: Schema di un sensore vibrazionale acustico capacitivo con diaframma con fori.

Infine vi sono anche sensori acustici vibrazionali capacitivi morbidi in gel, come idrogel e ionogel: a differenza degli altri sensori capacitivi, la loro capacità è data dalla formazione di

due strati elettrici quando un elettrodo entra in contatto con il gel sottoposto ad una tensione. Hanno dimostrato buone prestazioni nel rilevare il suono se immersi in acqua [12].

4.2.3 Sensori acustici vibrazionali morbidi triboelettrici

I sensori acustici vibrazionali triboelettrici sfruttano la triboelettrificazione e l'induzione elettrostatica per convertire le pressioni sonore in tensioni a circuito aperto o correnti di corto circuito. La struttura più semplice di questa tipologia di sensori è data da un sottile strato dielettrico e da due elettrodi, superiore e inferiore. Dunque l'effetto triboelettrico è basato sulla generazione di cariche elettriche durante i contatti tra gli elettrodi e lo strato dielettrico, infatti i sensori triboelettrici danno vita a correnti di corto circuito alternate attraverso contatti e separazioni ripetuti tra l'elettrodo e lo strato dielettrico. Essi possono autoalimentarsi, sono di facile fabbricazione e hanno bassi costi di produzione, però non hanno un output non lineare, la risposta in frequenza non è uniforme, hanno alta impedenza e sono molto suscettibili all'ambiente in cui si trovano [12].

Esistono diversi sensori acustici vibrazionali morbidi triboelettrici: alcuni sono costituiti da particelle di politetrafluoroetilene (PTFE) che vibrano alla medesima frequenza del suono misurato (Figura 4.4); altri sono in polimeri come il politetrafluoroetilene (PTFE) o il polietilene ad alta densità (HDPE) sottoforma di nanofili con elevata densità di carica; altri ancora sono realizzati in con nano-fibre di polimeri ottenute mediante elettrofilatura e sono sensibili alle pressioni sonore [12].

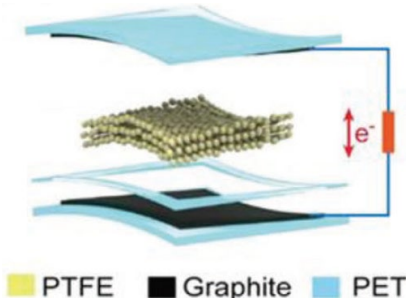

Figura 4.4: Schema di un sensore acustico basato su microparticelle di PTFE.

4.2.4 Sensori acustici vibrazionali morbidi piezoelettrici

I sensori acustici vibrazionali piezoelettrici sono composti da due elettrodi tra i quali è inserito un materiale piezoelettrico in grado di rispondere alle sollecitazioni meccaniche applicate deformandosi e inducendo al suo interno momenti dipolari e separazioni di carica. Si genera quindi una corrente elettrica dovuta alla differenza di potenziale tra i due elettrodi, che è

correlata all'intensità delle vibrazioni percepite: una volta che la pressione viene rimossa, il sensore riassume la sua forma originale e produce una corrente in cortocircuito alternata. Questi sensori godono di un'auto-alimentazione, di un'ampia larghezza di banda, di una grande sensibilità e di una struttura semplice. Tuttavia, presentano alcuni svantaggi come ad esempio una risposta in frequenza non uniforme, un'impedenza elevata e perdite di cariche elettriche [12].

Sono stati sviluppati sensori piezoelettrici morbidi per rilevare suoni e vibrazioni, utilizzando tecniche come l'elettrofilatura e materiali come le nanofibre di polivinilidenfluoruro (PVDF) e il solfuro di zinco (Figura 4.5). Altri sensori, invece, sono stati creati in materiali biodegradabili, come ad esempio la chitina, oppure fili compositi piezoelettrici termicamente disegnati per aumentare la sensibilità, nonostante la risposta in frequenza non uniforme. Infine, alcuni sensori consentono di operare una selezione in frequenza grazie alla geometria della struttura dei materiali piezoelettrici [12].

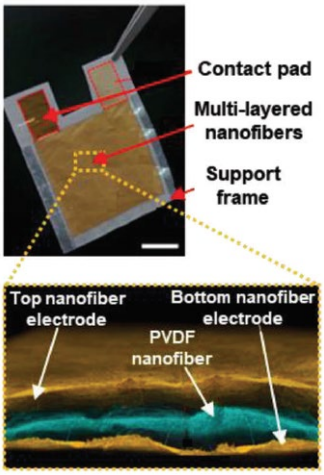

Figura 4.5: Sensore acustico multistrato basato su nano-fibre.

4.2.5 Altri sensori acustici vibrazionali morbidi

Di recente si sono combinati i meccanismi di trasduzione triboelettrica e piezoelettrica per ottenere un sensore acustico vibrazionale morbido allo scopo di migliorarne le prestazioni: un esempio è il sensore piezo-trio ibrido contenente film di alluminio e film di PVDF polarizzati che conferiscono al dispositivo una sensibilità lineare. In aggiunta, alcuni sensori sfruttano la struttura nanoporosa e la fibra cava di materiali come il PVDF; altri si basano su effetti elettromagnetici e altri ancora su effetti ottici. Quest'ultimi sensori inviano la luce attraverso un cavo a fibra ottica per illuminare un diaframma riflettente, il quale si muove modificando l'intensità della luce, che viene convertita in un segnale digitale [12].

Capitolo 5: Principali tecniche di signal processing

La crescente necessità di migliorare la diagnosi di malattie legate alla voce ha dato origine a un campo emergente chiamato *Analisi Automatica delle Condizioni Vocali* (AVCA). Esso si propone di esaminare, classificare e quantificare la presenza di una patologia vocale in un paziente, sfruttando sistemi automatici che forniscono misurazioni obiettive. La prima distinzione eseguita in questo ambito è tra voce normofonica e voce patologica. Una voce normofonica è caratterizzata da una qualità piacevole e possiede un'altezza e un volume adeguati che variano in base al significato del discorso. Contrariamente, una voce anormale è affetta da rumori, interruzioni inappropriate, perturbazioni e atonalità, che si traducono in afonie, mutismo e disfonie. I sistemi AVCA distinguono solamente tra voci normofoniche e disfoniche in quanto nel mutismo l'audio è totalmente assente e l'afonia è rilevabile senza dover condurre uno studio dettagliato [14].

I sistemi AVCA sono solitamente costituiti da una struttura base comune che consente l'estrazione di parametri, sottoforma di caratteristiche acustiche, che vengono successivamente analizzati dallo stesso (Figura 5.1) [14].

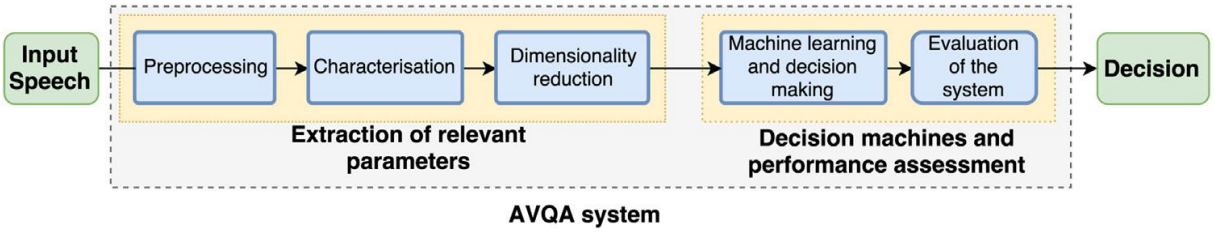

Figura 5.1: Schematizzazione della struttura di un comune sistema AVCA.

5.1 Estrazione di parametri rilevanti

Il discorso che il sistema riceve in input influenza la capacità di individuare la presenza o meno di malattie vocali, motivo per cui deve essere rilevato in maniera accurata. Esso può consistere in una vocale sostenuta oppure in un discorso continuo, contraddistinto da variazioni di intensità e pause. Al fine di poter confrontare i parametri vocali del paziente con dei parametri standard di una voce normofonica o patologica, viene preliminarmente creato un dataset contenente registrazioni di diversi interlocutori, che verrà poi impiegato per addestrare e testare il sistema AVCA. È necessario che durante l'acquisizione dei dati vengano seguite determinate linee guida, così da conferire oggettività, precisione e affidabilità elevate alle misurazioni effettuate. Ad esempio si richiede l'utilizzo di microfoni con sensibilità minima di -60 dB, distanze costanti tra bocca e microfono inferiori ai 10 cm, frequenze di

campionamento tra i 20 e i 100 kHz e stanze insonorizzate con un rumore ambientale inferiore a 50 dB. Inoltre il corpus registrato deve essere molto ampio in modo tale da contenere molte casistiche all'interno della classe e da essere equilibrato in termini di attributi come età e sesso. Esistono dataset pubblici, ad esempio il *Massachussets Ear and Eye Infirmary* (MEEI) e il *Saarbrücken Voice Database* (SVD), oppure esistono dataset privati come il corpus dell'*Ospedale Prìncipe de Asturias* (HUPA) e l'*Arabic Voice Pathology Dataset* (AVD). Infine, un altro corpus importante è stato realizzato negli ospedali di Marsiglia e Aix-en-Provence in Francia: in esso si trovano registrazioni del livello di pressione sonora (SPL), del flusso d'aria orale e della pressione dell'aria subglottica di varie categorie di interlocutori [14].

Una volta che il segnale vocale in input viene acquisito, mediante i sensori che si sono discussi nel *Capitolo 4*, viene sottoposto ad una preelaborazione. Solitamente si esegue una short-time analysis, che prevede la suddivisione del segnale in frame, ossia segmenti di uguale lunghezza, che permettono di trattarlo come un segnale stazionario o quasi stazionario. Successivamente avviene l'operazione di windowing: i frame vengono finestrati tramite il prodotto con una funzione di finestra allo scopo di attenuare inizio e fine dei diversi frame e di migliorare le proprietà spettrali del segnale. La scelta della finestra dovrebbe essere fatta in modo che essa possieda una risposta in frequenza con banda stretta nel lobo principale, per enfatizzare la frequenza principale del segnale, e un'ampia attenuazione nei lobi laterali, per ridurre le frequenze indesiderate. La lunghezza della finestra varia in base all'applicazione: di solito la durata è compresa tra i 10 e i 55 ms, ma in alcuni casi può essere anche di 800 ms. Tra le più comuni vi sono le finestre triangolari e le finestre di Hanning. È importante notare come nella preelaborazione del segnale vocale non sia adatto ridurre il rumore di fondo del segnale in quanto è il rumore stesso a consentire l'individuazione di condizioni patologiche perché esso deriva da fenomeni come irregolarità durante la vibrazione delle corde vocali o turbolenza dovuta alla loro chiusura incompleta [14].

Vi è poi la fase di caratterizzazione, che ha l'obbiettivo di creare un vettore contenente i parametri vocali da esaminare, che possono essere estratti in due modi: a breve termine, generando un numero di vettori pari al numero di frame; a lungo termine, fornendo un unico vettore per ogni registrazione audio. È molto difficile trovare attributi che descrivano in maniera efficace la presenza di disturbi vocali, per cui si ricercano molti parametri che siano poco correlati tra loro. Diverse sono le tecniche adottate per caratterizzare un segnale vocale, ma tra le più efficaci ci sono: l'analisi acustica, l'analisi temporale, l'analisi spettrale; l'analisi della complessità e la rappresentazione tridimensionale. Di seguito si farà riferimento ad un

esempio di due registrazioni audio di due voci: una normofonica e una disfonica. Tali audio sono registrazioni della vocale /a/, provengono dal dataset HUPA e sono state campionate con una frequenza di 50 kHz e con una risoluzione di 16 bit. La voce normofonica è di una donna di 36 anni ed ha una durata di 2.98 secondi, mentre la voce disfonica è di un uomo di 36 anni affetto da tumore ed ha una durata di 2.45 secondi. È stata eseguita una short-time analysis utilizzando finestre di Hanning della lunghezza che varia tra 40 e 180 ms [14, 15]

5.1.1 Analisi acustica

L'analisi acustica prevede la misurazione della funzione vocale, che si svolge attraverso la quantificazione dell'energia contenuta nel segnale. Tuttavia, siccome quest'ultima dipende dalla distanza tra la bocca e il microfono, spesso è preferibile indagare il livello di pressione sonora (SPL) [14].

Inoltre nell'analisi acustica vengono individuati rumore additivo e di modulazione attraverso tecniche basate sul monitoraggio della frequenza fondamentale F_0 e sulla derivazione di statistiche di basso ordine per rilevare anomalie rispetto ai normali modelli di vibrazione delle corde vocali. Tuttavia non esistono sistemi automatici capaci di correlare SPL e F_0 con le peculiarità del contesto e della persona che sta parlando [14].

5.1.2 Analisi temporale: perturbazioni e fluttuazioni

L'analisi temporale consente di individuare perturbazioni e fluttuazioni all'interno del segnale vocale. Le perturbazioni sono molto indicative per l'analisi dell'aperiodicità vocale conseguente a rumore di modulazione o additivo, causati da patologie. Infatti, quando viene pronunciata una vocale sostenuta si ottiene un segnale vocale periodico, che subisce modifiche anomale nel caso in cui siano presenti disturbi vocali. In particolare i parametri solitamente utilizzati per valutare le perturbazioni sono lo jitter e lo shimmer: in *Figura 5.2* si può osservare che il jitter percentuale della voce patologica presenta valori molto più alti rispetto a quello della voce normofonica e ciò è dovuto alla mancanza di stabilità nell'intonazione della persona; lo stesso fatto si verifica per lo shimmer (Figura 5.3) [14, 15].

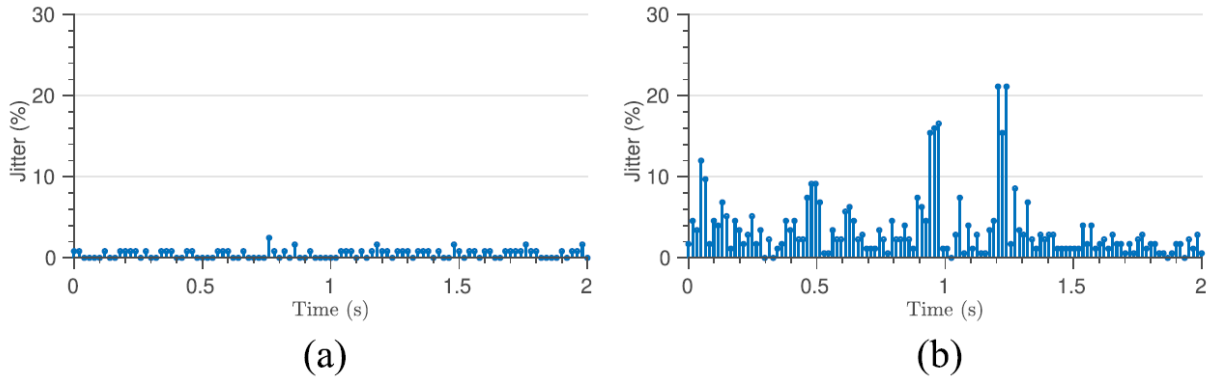

Figura 5.2: Jitter (%) di una voce normofonica (a) e di una voce patologica (b).

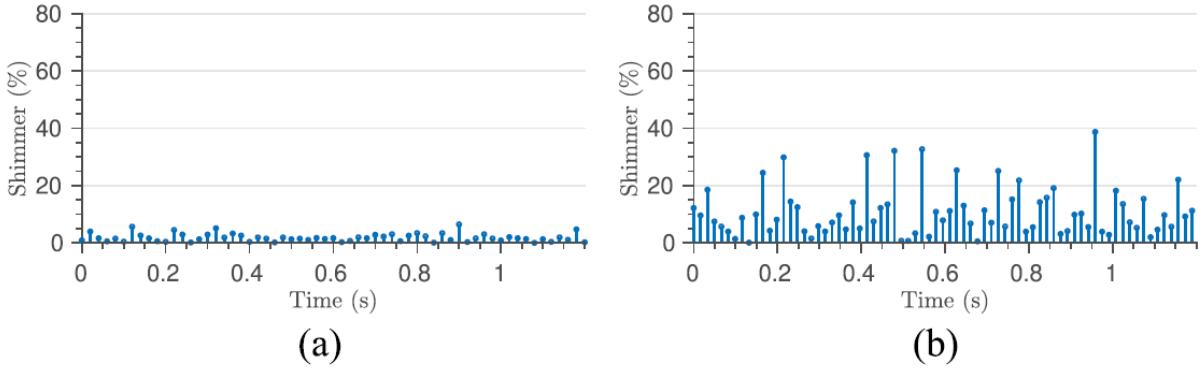

Figura 5.3: Shimmer (%) di una voce normofonica (a) e di una voce patologica (b).

L'analisi delle fluttuazioni, invece, consente lo studio di gravi disturbi nei meccanismi di movimento delle corde vocali, come ad esempio il tremore che si riflette in fluttuazioni a bassa frequenza nell'ampiezza oppure nella frequenza [14].

Altre caratteristiche a cui viene data importanza nell'individuazione di rumore nel segnale vocale sono il Rapporto Segnale-Rumore (SNR) e il rapporto Armoniche-Rumore (HNR). L'SNR espresso in dB è definito come il rapporto logaritmico tra la potenza del segnale ($P_{segnale}$) e la potenza del rumore (P_{rumore}) [15]:

$$SNR = 10log\left(\frac{P_{segnale}}{P_{rumore}}\right)$$

L'HNR, è il rapporto logaritmico tra l'energia delle armoniche (ι_{En}) e l'energia associata al rumore (γ_{En}) [15] :

$$HNR = 20\log\left(\frac{\iota_{En}}{\gamma_{En}}\right)$$

5.1.3 Analisi spettrale

Per misurare il rumore additivo viene considerato anche lo spettro del segnale, che permette di caratterizzare la qualità della voce, infatti molti dei suoi parametri hanno dimostrato una correlazione con la valutazione percettiva della disfonia. Mediante questa tecnica si possono analizzare indistintamente sia vocali sostenute che discorso continuo e sono stati sviluppati anche diversi approcci che prevedono l'impiego di banchi di filtri al fine di suddividere i segnali vocali in tante sotto-bande per consentire una maggiore precisione nella diagnosi di patologie vocali [14].

Per quanto riguarda l'HNR, ad esempio, spesso si usa una sua variante più robusta, ossia il Rapporto Capstrale Armoniche-Rumore (CHNR), che è dato dalla differenza tra lo spettro di rumore e lo spettro della voce e impiega uno strumento matematico che serve per analizzare i segnali acustici: la trasformata cepstrale. Per ricavare il CHNR, quindi, per ciascun frame: prima si calcola la trasformata cepstrale del segnale, vengono identificate le armoniche periodiche nel dominio cepstrale, le quali vengono in seguito rimosse per ottenere l'energia equivalente al rumore del segnale; infine si calcola l'energia totale nel dominio cepstrale e la differenza tra l'energia del rumore e l'energia totale fornisce l'HNR. In *Figura 5.4* si nota come l'ampiezza dello spettro di rumore della voce disfonica sia dello stesso ordine di grandezza dello spettro della voce disfonica, mentre lo spettro di rumore della voce normofonica è inferiore allo spettro della voce normofonica [14, 15].

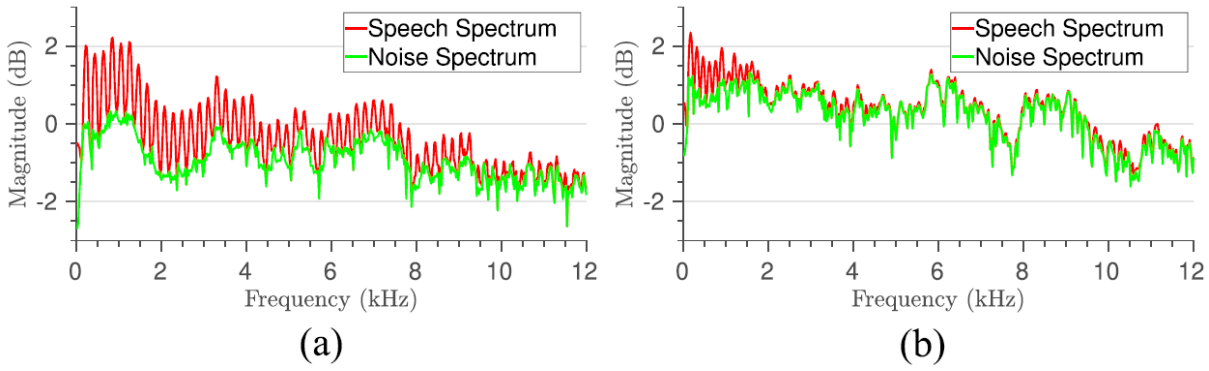

Figura 5.4: Spettro delle ampiezze e spettro del rumore di una voce normofonica (a) e di una voce disfonica (b) nel dominio cepstrale.

Infine, un altro parametro importante è l'Energia del Rumore Normalizzata (NNE). La NNE è la misura del livello di rumore additivo per ogni frame del segnale in analisi ed è espressa in funzione dell'energia del rumore (γ_{En}) e dell'energia totale del segnale (ι_{En}) [15]:

$$NNE = 20\log\left(\frac{\gamma_{En}}{\iota_{En}}\right)$$

Tale parametro si misura calcolando la frequenza fondamentale e lo spettro logaritmico del segnale, poi viene determinata l'energia associata al rumore direttamente dallo spettro e viene stimata l'energia dei picchi armonici; infine la NNE viene calcolata come la differenza tra l'energia spettrale totale e l'energia del rumore. In *Figura 5.5* si può vedere che la componente rumorosa nello spettro della voce disfonica è simile allo spettro vocale, mentre nella voce normofonica lo spettro del rumore è maggiore [15].

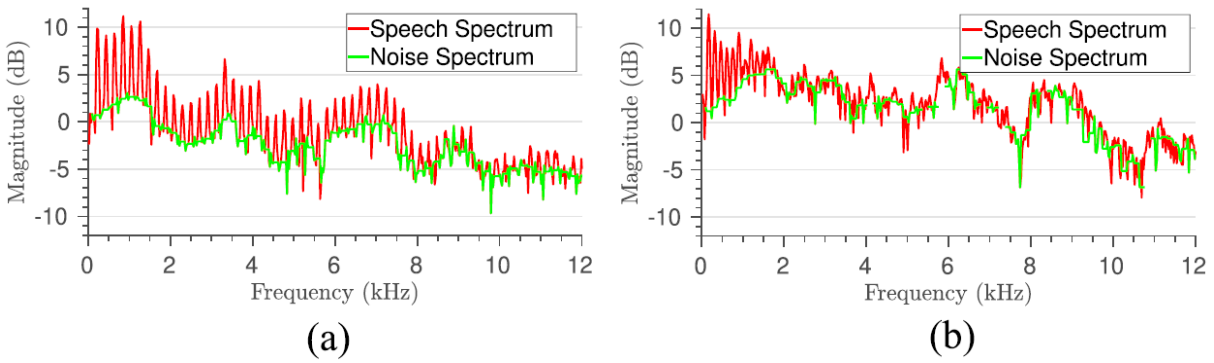

Figura 5.5: Spettro delle ampiezze e spettro del rumore di una voce normofonica (a) e di una voce disfonica (b) per il calcolo della NNE.

5.1.4 Analisi della complessità

Nei sistemi biologici la complessità si manifesta sottoforma di almeno uno dei seguenti attributi: non linearità, non stazionarietà, irreversibilità temporale, variabilità multiscala. In tutti i processi fisiologici si verifica la non linearità, dunque anche la voce assume questo comportamento che è dovuto ad esempio a collisioni tra le corde vocali, relazioni pressione-flusso non lineari nella glottide, asimmetrie tra le vibrazioni tra le corde vocali destra e sinistra. Uno dei metodi più utilizzati per indagare la complessità di un sistema è l'Analisi delle Dinamiche Non Lineari (NDA) oppure l'analisi dell'entropia, che sono state largamente introdotte nei sistemi AVCA. In aggiunta sono state impiegate decomposizioni tempo-frequenza per esplorare le proprietà frattali del linguaggio o per caratterizzare la complessità di ogni sotto-banda decomposta [15].

5.1.5 Rappresentazione tridimensionale

Ampiamente diffuse sono le rappresentazioni multidimensionali del segnale vocale, che si basano su tecniche di elaborazione delle immagini o strumenti matriciali per l'estrazione di parametri utili. Ad esempio gli Spettri di Modulazione (MS) sono rappresentazioni spettrali tridimensionali in cui sull'asse delle ascisse si trova la frequenza di modulazione, su quella

delle ordinate la frequenza del linguaggio parlato e sull'asse z l'energia del segnale: il loro vantaggio è dunque poter osservare contemporaneamente più caratteristiche vocali. Per realizzare un MS il segnale in input viene inizialmente filtrato tramite banchi di filtri che sfruttano la *short-time Fourier transform* (sTFT), poi una seconda sTFT esegue la modulazione in frequenza delle sotto-bande appena ottenute e si ottiene in questo modo una matrice E [M×A], costituita da A bande acustiche e da M bande di modulazione. Un generico punto nello spazio tridimensionale viene individuato nel seguente modo: $E(f^a, f^m)$, con f^a frequenza centrale della banda acustica e f^m frequenza centrale della banda di modulazione (Figura 5.6) [14, 15].

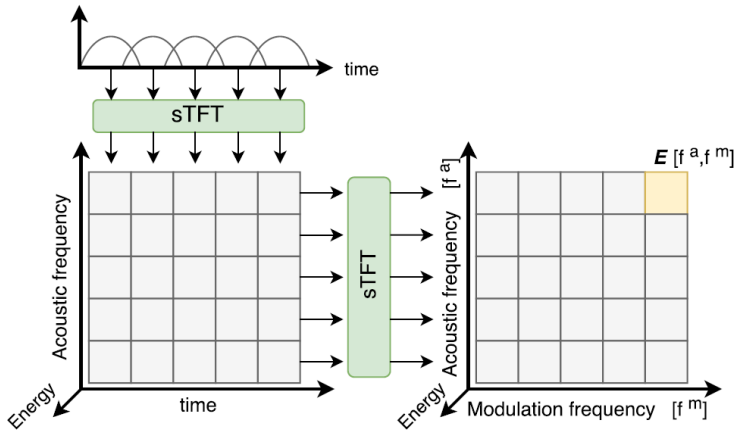

Figura 5.6: Passaggi per ottenere uno spettro di modulazione.

La matrice E può essere rappresentata anche come immagine bidimensionale in cui il colore varia in base all'energia del segnale ed è stato riscontrato che gli MS delle voci normofoniche sono più omogenei rispetto a quelli delle voci patologiche, come si può vedere in *Figura 5.7*. Ciò è dovuto ad una distribuzione più diffusa di energia nel caso disfonico, che a sua volta indica la presenza di rumore additivo e di modulazione [15].

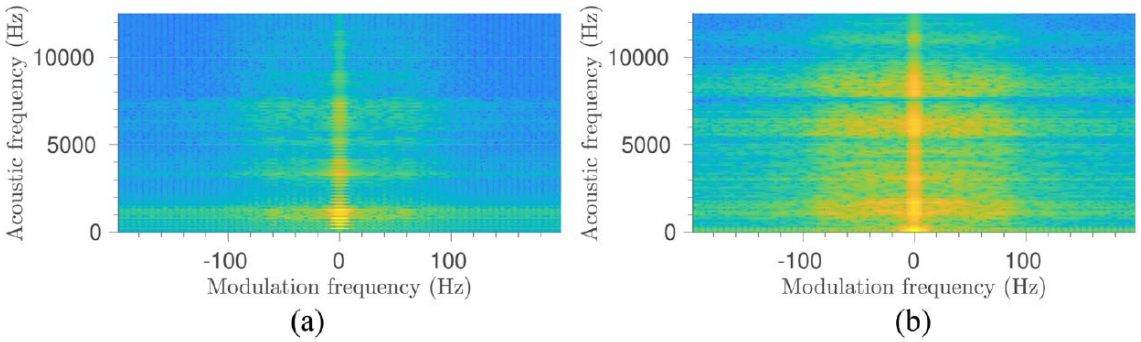

Figura 5.7: Modulo dello spettro di modulazione per una voce normofonica (a) e per una voce disfonica (a)

5.2 Valutazione dei parametri

Allo scopo di diminuire il numero di parametri ridondanti o irrilevanti che potrebbero influenzare il risultato dell'analisi vocale, si opera infine una riduzione del numero di caratteristiche considerate. Due sono le principali tecniche per realizzare questa riduzione: l'estrazione delle caratteristiche, che trasforma lo spazio in input in un insieme più compatto e quindi genera un nuovo spazio; la selezione delle caratteristiche, che non prevede alcuna trasformazione, bensì una scelta dei parametri più pertinenti. Nei sistemi AVCA si utilizzano inoltre tecniche di machine learning per automatizzare i processi. Il meccanismo con cui il machine learning impara dai dati si suddivide in apprendimento supervisionato e non supervisionato: nei sistemi AVCA viene impiegato il primo, che consente di mappare un insieme di osservazioni vocali con le relative etichette per riconoscere le patologie vocali [14].

Capitolo 6: Un esempio da letteratura: un dosimetro per la valutazione e il monitoraggio dei disturbi vocali

Le attuali tecniche di analisi per l'individuazione di patologie della voce, come si è già visto, prevedono dei test che vengono svolti in strutture specifiche mediante attrezzature specializzate. I risultati forniti rispecchiano lo stato di salute del paziente in un determinato momento e in determinate condizioni, tuttavia solo in casi specifici a scopo di ricerca vengono effettuate misurazioni continue per lunghi periodi di tempo, procedura che offre risultati più informativi e affidabili, oltre che una diagnosi e un trattamento personalizzati [16]. Proprio per questi motivi è stato creato un prototipo di dispositivo per l'analisi vocale descritto nell'articolo *Introducing a New Dosimeter for the Assessment and Monitoring of Vocal Risk Situations and Voice Disorders* realizzato da Marcos Llorente-Ortega, Adam Podhorski, e Secundino Fernandez. In un primo momento si sono indagati i diversi dispositivi di dosimetria vocale già sviluppati: è emerso che si suddividono in dispositivi di monitoraggio, che registrano i dati, e dispositivi di risposta, che rispondono a stimoli acustici o vibratori quando alcuni valori dei parametri escono da range prestabiliti. Alcuni di questi dispositivi sfruttano un microfono, mentre altri un accelerometro per registrare i dati. In aggiunta solo una piccola parte di essi può registrare tutti i parametri di interesse e spesso sono costosi e difficilmente reperibili [16].

Il dispositivo ideato da Llorente-Ortega, Podhorski e Fernandez è costituito da due componenti: un accelerometro in grado di convertire le vibrazioni vocali in un segnale elettrico, che sorge a contatto con la pelle su un sopporto realizzato con una stampante 3D e posizionato nella regione sovrasternale del paziente; un microcontrollore che esegue analisi ed acquisizione del segnale. Il modello di accelerometro utilizzato è il *BU-27135-0000* della *Knowles Electronics* in quanto è di dimensioni ridotte e possiede una risposta in frequenza unidirezionale e piatta tra 20 e 2000 Hz. Il microcontrollore ha un consumo energetico molto basso ed è il modello *MSP430FR5994* della *Texas Instruments*. In esso è incluso un coprocessore matematico a basso consumo denominato *LEA* (Low Energy Accelerator) che può eseguire la Fast Fourier Transform per convertire il segnale nel dominio della frequenza ed un filtraggio lineare. La sua fonte di alimentazione è una batteria da 5V con capacità di 2000 mAh e i dati elaborati vengono memorizzati in una scheda micro SD. Il dispositivo lavora ad una frequenza di campionamento di 2048 Hz e calcola la frequenza fondamentale ogni 250 ms, mentre l'ampiezza e l'attività fonatoria ogni 31.25 ms. Date le componenti a

basso consumo energetico e di modeste dimensioni, l'attrezzatura sviluppata per l'analisi vocale è indossabile [16].

Il dispositivo è stato testato su una donna di 46 anni (Soggetto 1) e un uomo di 26 anni (Soggetto 2), per una durata di quattordici e undici giorni rispettivamente, durante i quali quotidianamente i dati venivano scaricati e la batteria ricaricata. I parametri analizzati sono la frequenza fondamentale , l'ampiezza del segnale e il tempo di fonazione: è stato richiesto ai due soggetti di utilizzare normalmente la loro voce per tutta la durata del test, tranne che per un giorno in cui dovevano parlare con maggiore intensità o frequenza rispetto al solito [16].

Il Soggetto 1 ha parlato quotidianamente tra le 0.24 e le 1.03 ore su una durata di registrazione compresa tra le 4.38 e le 7.17 ore, calcolate nei quattordici giorni. Nel settimo giorno il Soggetto 1 ha utilizzato la voce in modo insolito (Figura 6.1 A). La frequenza fondamentale (F_0) è di 23.39 ± 55.62 Hz, mentre l'intensità relativa è di 42.50 ± 2.26 m/s². Rimuovendo il settimo giorno, la F_0 è di 220.53 ± 23.78 Hz e l'intensità relativa è di 42.17 ± 1.87 m/s²: in questo giorno la F_0 è di 320.81 ± 122.43 Hz, e l'intensità relativa è di 45,17 ± 3,12 m/s² (Figura 6.2 A e C) [16].

Il Soggetto 2, invece, ha effettuato undici registrazioni con una durata quotidiana compresa tra le 2.76 e le 6.39 ore e un tempo di fonazione tra le 0.1 e le 0.73 ore. Il nono giorno il soggetto ha seguito uno schema inusuale (Figura 6.1 B). La frequenza fondamentale (F_0) è di 108.10 ± 17.75 Hz, mentre l'intensità relativa è di 43.68 ± 2.26 m/s². Rimuovendo il nono giorno, la F_0 è di 105.76 ± 11.36 Hz e l'intensità relativa è di 43.70 ± 2.30 m/s²: in questo giorno la F_0 è di 148.91 ± 42.72 Hz, e l'intensità relativa è di 43.34 ± 1.41 m/s² (Figura 6.2 B e D) [16].

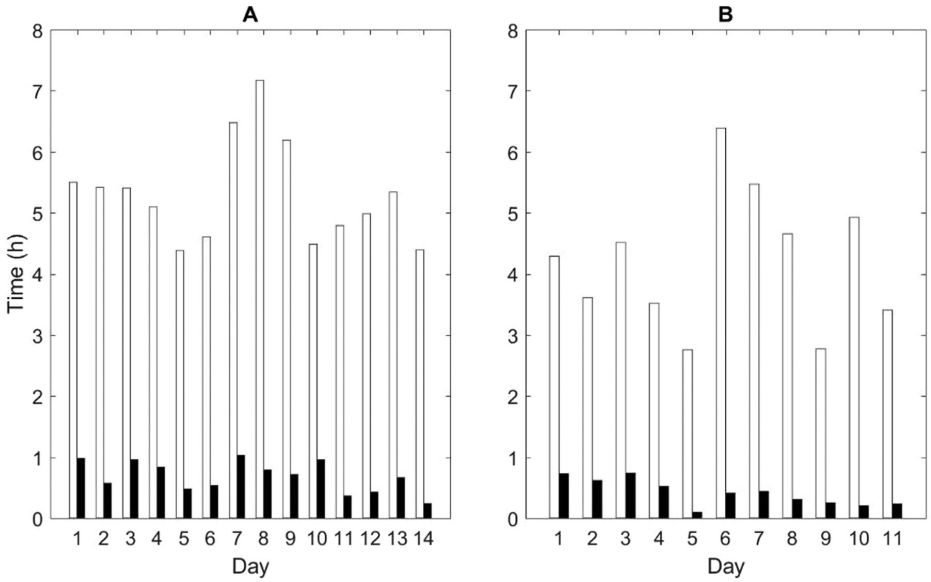

Figura 6.1: Confronto tra il tempo di registrazione (bianco) e il tempo di fonazione (nero) per il Soggetto 1 (A) e il Soggetto 2 (B).

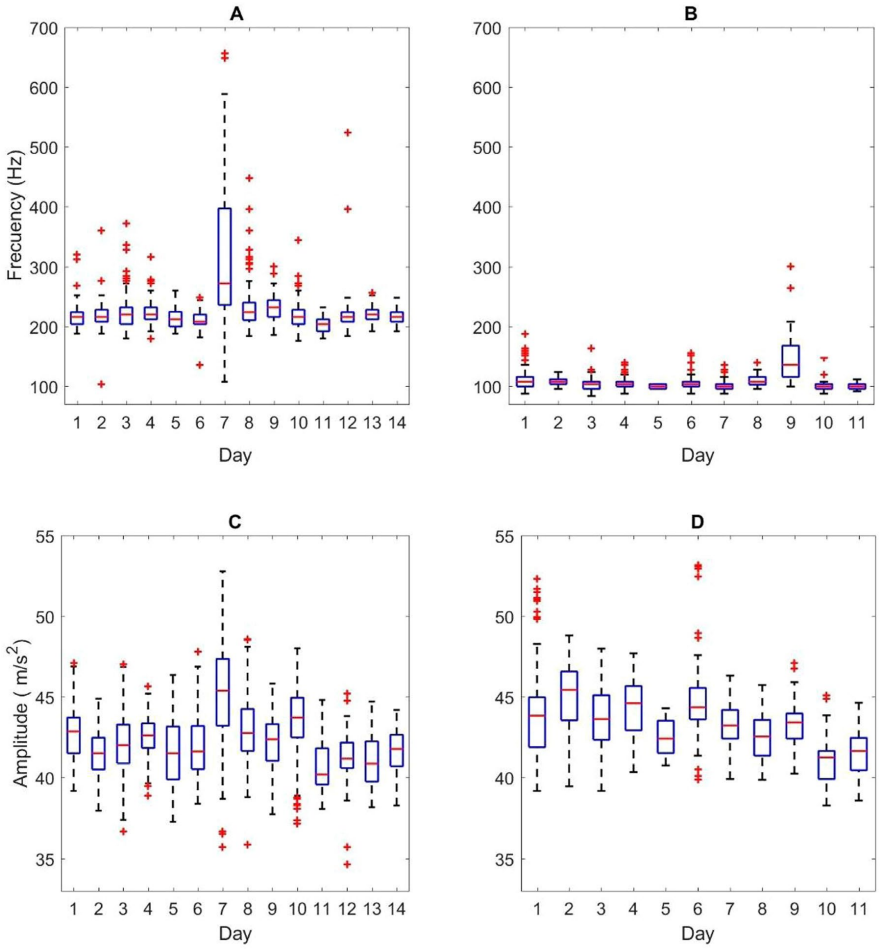

Figura 6.2: Confronto di F_0 (A e B) e ampiezza (C e D) nei diversi giorni di registrazione del Soggetto 1 (A e C) e del soggetto 2 (B e D).

Si può notare, dunque, che il prototipo ideato, oltre ad essere indossabile ed ad adattarsi perfettamente al paziente, offre numerosi altri vantaggi tra cui oggettivare e quantificare il tempo di fonazione, consentendo di approfondire lo studio dell'uso della voce per lunghi periodi. Inoltre il sistema rispetta a pieno la privacy della persona, dal momento che il sensore è costituito da un accelerometro che rileva vibrazioni e non da un microfono che registra anche il contenuto del discorso. Infine il segnale viene elaborato ed archiviato immediatamente [16].

L'invenzione di un dosimetro come quello appena descritto rappresenta una vera e propria rivoluzione in ambito clinico visto che permette di mantenere monitorato il proprio segnale vocale, e quindi il proprio stato di salute, quotidianamente e in maniera continua. Ciò è di grande aiuto soprattutto nei casi di soggetti malati, ma anche in casi di soggetti che hanno subito interventi o che potenzialmente possono sviluppare più facilmente patologie che si riflettono sul segnale vocale. Infatti è bene ricordare che la progettazione di dispositivi biomedici ha come obbiettivo il miglioramento della salute e della qualità di vita della persona affetta da malattie, cercando allo stesso tempo di consentirle di condurre una vita quanto più normale possibile, ed è proprio ciò che questo dosimetro promette di fare.

Inoltre anche il sistema ospedaliero può trarre vantaggio da una tecnologia simile in quanto grazie ad essa non vi è più la necessità di un esperto che svolga l'esame vocale presso l'ospedale, come è invece previsto dalle attuali tecniche di analisi vocale. Questo rappresenta un risparmio di tempo perché il paziente si recherà in ospedale solo per ritirare e consegnare il dispositivo, un risparmio di personale ospedaliero poiché il dosimetro rileva autonomamente i parametri vocali che il medico dovrà solo analizzare, e un risparmio di denaro dato che le componenti del dispositivo sono relativamente economiche, caratteristica che lo rende altresì più accessibile a tutti.

Conclusioni e sviluppi futuri

Come si è potuto notare, l'analisi vocale si dimostra promettente ed efficace nell'individuazione di disturbi vocali, tuttavia non esistono ancora dei dispositivi che non siano sperimentali utilizzabili per lo svolgimento della stessa. Si spera che in futuro le tecnologie non invasive per l'analisi vocale diventino sempre più accessibili e diffuse, soprattutto per le categorie di persone che possono essere maggiormente affette da patologie vocali, tra cui insegnanti, cantanti e attori [16].

Si prevede altresì che l'analisi audio verrà superata dall'analisi video, che include anche le immagini. Lo studio di quest'ultime, infatti, se integrato al segnale vocale, consentirà una migliore caratterizzazione dei parametri dei pazienti riuscendo ad individuare anche emozioni e sfaccettature legate alla salute attraverso il riconoscimento facciale [11].

Inoltre con l'aumento delle capacità di trasferimento dei dati, grazie alle reti 5G e alle sue versioni future, e con l'incremento costante della porzione di popolazione in possesso di uno smartphone, si renderà più agevole la raccolta e l'elaborazione di molti campioni vocali sia in formato grezzo che in alta definizione [11].

Per quanto riguarda la ricerca, verrà sicuramente considerato ed incluso un maggior numero di parametri vocali testati sia in studi sperimentali che nella quotidianità delle persone [11]: quindi aumenterà sia il numero di dati presenti nei dataset già esistenti, sia la tipologia di dati considerati.

Infine, dal punto di vista dell'assistenza sanitaria, l'analisi vocale potrà essere svolta direttamente nei centri di chiamata sanitaria dove consentirà consulenze potenziate, una più accurata autenticazione dei chiamanti e l'analisi in tempo reale della salute della persona. Tali tecnologie in futuro saranno sempre più integrate nello sviluppo di medici virtuali e cliniche virtuali o digitali [11].

Bibliografia

[1] *Clinical Anatomy and Physiology of the Voice,* Robert T. Sataloff, MD, DMA, Yolanda D. Heman-Ackah, MD, Mary J. Hawkshaw, BSN, RN, CORLN

[2] *Anatomy of the Larynx and Cervical Trachea,* Kassie, L. McCullagh, MD, Rupali N. Shah, MDb, Benjamin Y. Huang, MD, MPHa

[3] *Anatomy and Physiology of Phonation,* Candace Hrelec, Emily Zhang

[4] *Intonation processing for speech technology,* Grażyna Demenko

[5] *Workshop on Acoustic Voice Analysis: Summary Statement,* National Center for Voice and Speech, Titze IR

[6] *Acoustic Analysis of Normal Voice Patterns in Pakistani Adults,* Saira Ambreen, Nazia Bashir, Sharmeen Aslam Tarar, and Rukhsana Kausar

[7] *Vocal acoustic analysis—jitter, shimmer and HNR parameters,* Teixeira JP, Oliveira C, Lopes C.

[8] *Automated voice pathology discrimination from audio recordings benefits from phonetic analysis of continuous speech,* Mark Huckvale, Zhuoya Liu, Catinca Buciulea

[9] *LPC Spectrum First Peak Analysis for Voice Pathology Detection,* Hugo Tito Cordeiroa, José Manuel Fonsecaa, Carlos Meneses Ribeirob

[10] *Continuous Speech Classification Systems for Voice Pathologies Identification,* Hugo Cordeiro, Carlos Meneses, and José Fonseca

[11] *Voice for Health: The Use of Vocal Biomarkers from Research to Clinical Practice,* Guy Fagherazzi, Aurélie Fischer, Muhannad Ismael, Vladimir Despotovic

[12] *Emerging Trends in Soft Electronics: Integrating Machine Intelligence with Soft Acoustic/Vibration Sensors,* Jeng-Hun Lee, Kang Hyuk Cho, and Kilwon Cho

[13] *Smartphone-based sound level measurement apps: Evaluation of directional response,* Metod Celestina, Chucri A. Kardous, Andrej Trost

[14] *On the design of automatic voice condition analysis systems. Part I:Review of concepts and an insight to the state of the art,* J.A. Gómez-García, L. Moro-Velázquez, J.I. Godino-Llorente

[15] *On the design of automatic voice condition analysis systems. Part III: review of acoustic modelling strategies*, J.A. Gómez-García, L. Moro-Velázquez, J.D. Arias-Londoño, J.I. Godino-Llorente

[16] *Introducing a New Dosimeter for the Assessment and Monitoring of Vocal Risk Situations and Voice Disorders, Marcos Llorente-Ortega*, Adam Podhorski, and Secundino Fernandez