SAPIENZA UNIVERSITÀ DI ROMA
INGEGNERIA DELLE TELECOMUNICAZIONI

WORDIE: Riconoscitore universale di parole isolate in tempo reale (Word On the fly Recognition Device for Irksome Environments)

Relatore
Prof. M.-G. Di Benedetto

Candidato
Gianni Colasanti

Correlatore
Dott. Ing. Daniele Domenicali

Anno Accademico < 2007/2008 >

Ringraziamenti

Alla mia famiglia per aver creduto in me e per avermi dato la possibilità di intraprendere questo percorso

Alla prof. Maria Gabriella Di Benedetto, per la fiducia fin da subito dimostratami nell'aver accettato questo argomento di tesi

A Dome per avermi seguito durante lo svolgimento del lavoro con consigli e confronti che mi hanno aiutato ad intraprendere, ogni volta, le scelte più appropriate

A Giulia, stupenda compagna di viaggio, per il suo aiuto, per tutti i bei momenti passati insieme, per i suoi continui doppisensi inconsapevoli, per il suo supporto nei momenti di crisi, per tutte le risate, per le nottate passate sulle tabelle in latex etc.....
Ai ragazzi della casa, Daniele Luca e Vincenzo per aver condiviso con me questa fantastica esperienza

A tutti i componenti del laboratorio acts per avermi fatto passare otto mesi stupendi

A tutti i miei amici per il supporto da loro ricevuto

Indice

Capitolo 1

Introduzione

Il sistema di riconoscimento umano del linguaggio parlato è naturale, robusto ed efficiente. Questo sistema infatti riesce a funzionare correttamente anche in situazioni sfavorevoli, come quando c'è rumore di sottofondo o riverbero. Il sistema di riconoscimento umano compie, nel suo funzionamento, computazioni, filtraggi e adattamenti ai diversi parlanti con cui ha a che fare, riesce dunque a trasformare un segnale vocale in una successione di vocaboli, alla quale poi da un'interpretazione. I dettagli di come tutto questo succeda esula dalla nostra attuale conoscenza e nonostante le conferme teoriche ottenute circa le basi fisiologiche dell'ascolto, ci sono molti aspetti ancora da scoprire.
A dispetto di queste incomprensioni, la neurofisiologia del riconoscimento del parlato, negli ultimi 30 anni, ha ottenuto dei progressi nella creazione di metodi artificiali che emulino il comportamento umano. I Riconoscitori Automatici del Parlato (ASR) stanno cominciando a funzionare abbastanza bene per gli scopi del mercato di massa ed alcuni di essi vengono adoperati come sistemi di dettatura automatica. La progettazione e la costruzione di sistemi artificiali di questo tipo ha presentato difficoltà, a causa anche di problemi di complessità e robustezza: fattori come la variabilità del parlato da persona a persona, il rumore ambientale, la confondibilità delle parole, gli effetti di

coarticolazione, inficiano molto sul loro funzionamento e le prestazioni del sistema uditivo umano sono ancora lontane dall'essere raggiunte.
I moderni sistemi di dettatura automatica, nei quali l'utente parla in un microfono e le parole vengono scritte, dal riconoscitore, su una pagina elettronica, hanno prestazioni che si aggirano intorno al 90, ma necessitano di un addestramento preliminare e mantengono questi risultati solo per l'utente che li ha addestrati.
Un riconoscitore dovrebbe prescindere dal parlante con cui ha a che fare e questo implica che gli ASR di questo tipo abbiano un'architettura robusta a tali variazioni.
Attualmente il riconoscimento vocale è utilizzato in vari campi, sia civili che militari: ad esempio nei telefoni cellulari, negli ambienti di ufficio, per l'aiuto di persone con handicap, per il controllo del traffico aereo, per accessi sicuri tramite l'identificazione della voce, per lo speech-to-text, per corsi di lingue e traduzioni, per applicazioni di comando e controllo in ambito militare su elicotteri, aerei, carro armati, e su computer indossabili.
Proprio per l'ampio campo di applicazione, il sistema di riconoscimento vocale si trova a lavorare in ambienti molto eterogenei fra loro. In ambito militare possono presentarsi situazioni estreme di varia natura che possono influire in modo negativo nel corretto funzionamento del sistema: in particolare l'ambiente di lavoro potrebbe essere estremamente rumoroso, come la cabina di pilotaggio di un aereo, di un elicottero o di un carro armato.
Anche il cambio di intonazione della voce ha grossa influenza nel riconoscimento vocale: la voce del parlatore può essere influenzata da vari fattori come lo stress, il dover urlare o il dover bisbigliare. In ambito militare una situazione simile può capitare con frequenza, visto che un soldato può essere sottoposto a situazioni estreme.

In questa tesi verrà analizzata la possibilità di realizzare un sistema di riconoscimeto vocale per applicazioni militari. L' operazione di riconoscimento dovrà essere in tempo reale, indipendente dal parlatore, snella dal punto di vista computazionale e robusta al rumore ed allo stress.
Questo lavoro di tesi è organizzato in 7 capitoli che sono brevemente descritti di seguito:

Capitolo 1 è una breve introduzione al lavoro che segue;

Capitolo 2 introduce il concetto del riconoscimento vocale, presenta lo schema a blocchi di un Sistema automatico di riconoscimento vocale e anlizza brevemente gli effetti del rumore e dello stress nelle prestazioni del riconoscitore;

Capitolo 3 viene effettuata una analisi teorica delle caratteristiche utilizzate del riconoscitore implemetato in questa tesi;

Capitolo 4 é introdotto il concetto di endpoint detection per effettuare l' operazione di VAD (Voice activity detector);

Capitolo 5 è analizzato il concetto di classificazione polinomiale e come questo si adatta al riconoscimento vocale;

Capitolo 6 discute i vari blocchi che compongono il riconoscitore vocale realizzato in questa tesi.

Capitolo 7 vengono presentati i risultati e gli sviluppi futuri.

Capitolo 2

ASR (Automatic speech recognition)

2.1 Sistemi di Riconoscimeto Vocale

Lo scopo di un sistema di riconoscimento vocale è quello di riconoscere i fonemi o le parole generate da un determinato individuo. Tali sistemi vengono oggi utilizzati sia in ambito militare che civile. Possono, ad esempio, essere utilizzati nell'ambito della telefonia cellulare, per il controllo del traffico aereo, nei sistemi di sicurezza basati sull'identificazione della voce, per l'implementazione di dispositivi di tipo speech-to-text, per applicazioni di comando e controllo in ambito militare su elicotteri, aerei, carri armati e computer indossabili dal soldato.

In ambito militare tali dispositivi possono essere molto utili, in quanto il soldato, non dovendo controllare il sistema manualmente, può utilizzare le proprie mani per impugnare armi o pilotare un mezzo di trasporto.

L'eterogeneità caratterizzante il campo di applicazione dei sistemi di riconoscimento vocale si traduce in un fattore non trascurabile di complessità implementativa che incidere sul corretto funzionamento del sistema in contesti operativi che possono risultare molto diversi

tra loro. In particolare si pensi ad ambienti di lavoro particolarmente rumorosi o ad ambienti di lavoro che provocano un cambio di intonazione o in generale un fattore di deformazione nella voce del parlatore. Un esempio tipico di tali alterazioni rispetto al parlato nominale è fornito dalla condizione di stress in cui può versare il parlatore. Tali situazioni non sono infrequenti in ambito militare, visto che un soldato può trovarsi ad affrontare situazioni di estremo pericolo in grado di provocare un forte stress sia fisico che mentale.

2.2 Struttura di un sistema di riconoscimento vocale

La struttura generale di un ASR (Automatic Speech Recognition device) è mostrata in figura 2.1. L'acquisizione del segnale avviene attraverso il microfono, in grado di captare le onde sonore e di convertirle in un segnale elettrico. Tale segnale, dopo essere stato campionato e digitalizzato, passa attraverso un rivelatore di attività vocale (VAD - Voice Activity Detector), il quale stabilisce, in base ad un'opportuna procedura di analisi del segnale al suo ingresso, se è stata emessa una parola o un fonema, oppure se è stata acquisita una componente esclusivamente rumorosa. Segue il sistema di riconoscimento vocale vero e proprio, in cui il segnale viene opportunamente elaborato al fine di riconoscere lo specifico fonema o la parola emessa. Questo blocco è composto da un estrattore di caratteristiche, che mediante opportune elaborazioni fa corrispondere al segnale d'ingresso un vettore N-dimensionale. Tale vettore viene posto in ingresso al classificatore, che si basa su un vocabolario di riferimento costruito in una fase preliminare denominata addestramento, al fine di prendere una decisione conclusiva su quale sia l'espressione effettivamente pronunciata.

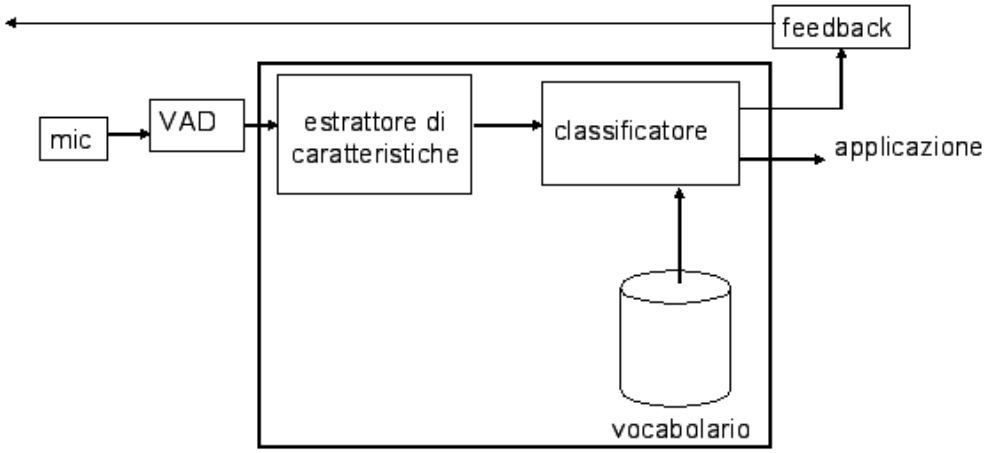

Figura 2.1: Struttura di un sistema di riconoscimento vocale

Risulta spesso presente anche un ramo di feedback che indica all'utilizzatore se la parola è stata riconosciuta o se il classificatore non è riuscito a lavorare nell'ambito dei parametri decisionali previsti in fase di progetto.
Di particolare importanza risulta, come precedentemente accennato, la fase preliminare di addestramento (*Training*) che serve a far sì che l'ASR distingua correttamente le varie parole. L'addestramento può essere effettuato utilizzando differenti metodologie, analizzate in dettaglio nelle sezioni successive.

2.3 Componenti di un ASR

2.3.1 Il microfono

Il primo elemento del sistema è il microfono, la sua funzione è di assoluta importanza in quanto da esso dipende la corretta acquisizione del segnale vocale e l'eventuale presenza di rumore esterno. In questa sezione sono descritti alcune tra le più diffuse tipologie di microfono

nell'ambito dei dispositivi ASR: i microfoni direzionali, a cancellazione di rumore, a conduzione ossea.

Microfono direzionale

Un microfono direzionale è un dispositivo la cui sensibilità varia fortemente con la direzione. I microfoni direzionali sono molto utili nel facilitare la comprensione della voce in ambienti pervasi da suoni interferenti, poiché riescono a sopprimere efficacemente i disturbi caratterizzati da direzioni diverse da quella verso cui è indirizzato il microfono.
L'angolo di reiezione indica le direzioni in cui la sensibilità del microfono risulta ridotta o nulla: esso definisce un raggio che parte dal microfono, si dirige lungo la direzione dove è puntato il microfono e si allarga fino ad un'ampiezza tale da toccare gli estremi dell'angolo di reiezione.

Microfono a cancellazione di rumore

Il microfono a cancellazione di rumore, che può essere anche direzionale, riesce a minimizzare il rumore ambientale (macchine, conversazioni), utilizzando ad esempio una coppia di microfoni, uno puntato sulla bocca del parlatore e l'altro sull'ambiente. I due segnali vengono poi sottratti al fine di ottenere un'acquisizione del parlato più pulita..
E' descritto di seguito un esempio commerciale di microfono a cancellazione, Andrea Active Noise Cancellation (Andrea ANC) [1].
Il microfono Andrea ANC è un microfono a cancellazione di rumore progettato per incrementare l'intelligibilità della voce riducendo il rumore nella banda tra 150 e 3500 Hz, dove risiede la parte più rilevante del segnale vocale. Il microfono presenta anche una risposta in frequenza piatta tra i 100 e i 15000 Hz.

Andrea ANC è composto da una coppia di microfoni omnidirezionali posizionati in modo da ottimizzare la cancellazione del rumore e raggiungere una risposta in frequenza piatta. Questi risultati dipendono in gran parte dai circuiti attivi caratterizzanti il dispositivo e dalla posizione assunta dalla coppia di microfoni omnidirezionali.

I microfoni direzionali acquisiscono la loro proprietà di direzionalità essendo sensibili al gradiente di pressione presente in due punti diversi dello spazio. I microfoni omnidirezionali, invece, misurano le variazioni di pressione di un'onda sonora in un volume definito d'aria, prescindendo quindi dalla direzionalità.

I microfoni direzionali hanno inoltre una risposta in frequenza non piatta, a differenza dei microfoni omnidirezionali che invece mostrano una risposta piatta. Poiché Andrea ANC è composto da una coppia di microfoni omnidirezionali, la risposta in frequenza di Andrea ANC è molto più piatta rispetto a quella di un tipico microfono direzionale.

Microfono a conduzione ossea

Il microfono a conduzione ossea è invece posto a diretto contatto con l'osso della mascella o con l'orecchio e capta le vibrazioni ossee dovute alla voce. Il rumore risulta praticamente trascurabile in questo caso. La voce può, però, e risultare distorta e priva di alcune delle caratteristiche originarie.

Non sono stati fatti per ora degli studi approfonditi di questi microfoni nell'ambito dei sistemi ASR, ma si pensa che potrebbero portare miglioramenti in quanto molto robusti al rumore. Bisogna però contare che tali microfoni esaltano le basse frequenze a discapito delle alte frequenze e potrebbero essere non molto robusti a fenomeni di vibrazione.

2.3.2 Rivelatore di attività vocale

Lo scopo di questo dispositivo è di separare l'effettiva parola pronunciata dal parlatore dal silenzio e dal rumore di fondo. Questo compito può non essere semplice a causa della presenza di sorgenti rumorose proprie dell'ambiente operativo; si consideri, inoltre, che non sempre il parlatore desidera interagire con l'ASR. Quest'ultima situazione può essere risolta prevedendo la pronuncia di una parola chiave prima del comando effettivo. Utilizzando questa strategia sorgono però almeno due problemi: il primo di natura energetica, in quanto il sistema deve permanere in uno stato di ascolto continuo, il secondo riguarda il conseguente aumento del numero di falsi allarmi e di errori di inserimento.

Un approccio alternativo è rappresentato dalla tecnica Push-To-Talk (PTT), che, a differenza del precedente approccio, preclude la possibilità che si verifichino attivazioni indesiderate. Il PTT prevede la pressione di un tasto che attiva l'ASR quando si deve pronunciare il comando. Un evidente lato negativo di questa tecnica, soprattutto in ambito militare, è che si deve utilizzare una mano per attivare il sistema. La posizione del pulsante risulta dunque di fondamentale importanza, ed è necessario prevedere una strategia che determini la conclusione della fase di ascolto.

Una tecnica alternativa è il Push-And-Hold-To-Talk (PHTT), in cui si mantiene premuto il pulsante finché la pronuncia del comando non risulta terminata. Si noti che questa soluzione tiene impegnata una mano durante tutta l'enunciazione del comando [2].

2.3.3 Il sistema di riconoscimeto vocale

Come si può osservare dalla figura 2.1, il sistema di riconoscimento vocale è composto da diversi sottosistemi. Il segnale in uscita dal VAD viene prima elaborato dall'estrattore di caratteristiche, la cui uscita è poi utilizzata dal classificatore. Queste operazioni sono molto diverse a seconda del contesto operativo considerato. Per questo motivo è bene descrivere alcune caratteristiche strutturali degli ASR.

2.3.4 Dipendenza del parlatore

Si possono avere ASR che funzionano solo per un parlatore o per più parlatori. Nel primo caso, il riconoscitore può essere definito *speaker-dependent* ed il modello vocale deve essere adattato alla voce del parlatore. Il tutto si traduce, nella fase di addestramento, nel far leggere al parlatore specifico un testo con voce e velocità naturali. Questi sono i sistemi che offrono i risultati migliori in termini di precisione. A valle di un opportuno periodo di addestramento, questa tipologia di ASR permette anche di correggere errori di interpretazione. Gli algoritmi alla base dei dispositivi *speaker-dependent* prevedono che venga tenuta traccia delle correzioni, per consentire di imparare dagli errori precedentemente commessi.

Nel secondo caso, il riconoscitore può essere definito *speaker-independent*. In questo situazione non si è legati ad un particolare timbro vocale e si ottiene in generale una precisione inferiore rispetto alla soluzione precedente. Tali sistemi raggiungono buoni risultati nel caso in cui le parole pronunciate facciano parte di una lista di espressioni a cardinalità ristretta. Questo tipo di riconoscitore trova usuale applicazione in ambito militare [2].

2.3.5 Il limite di parola

Il limite di parola o *word boundary* si riferisce al tempo intercorso tra una parola e la successiva. In base a questo paramentro è possibile distingure tra vari tipi di ASR:

- I sistemi *isolated-word* in cui le parole sono pronunciate con delle pause oscillanti tra i 100 e i 250 ms, affinché il sistema possa riconoscerle correttamente. Si utilizza una strategia del genere nei sistemi command and control, dove gli ordini da pronunciare sono tipicamente parole singole.

- I sistemi *connected-word* che hanno bisogno di un tempo inferiore a 100 ms, ma non tollerano cambiamenti di intonazione tra una parola e l'altra.

- I sistemi *continuos-speech* non richiedono, invece, alcuna interruzione tra le parole ed accettano quindi un parlato spontaneo e fluente. I diffusi sitemi speech-to-text fanno parte di questa categoria.

In generale un sistema *isolated-word* richiede meno risorse di un sistema *continuos-speech* [2].

2.3.6 Dimensione del vocabolario

La dimensione del vocabolario rappresenta il numero di parole che il riconoscitore è in grado di riconoscere correttamente. Gli ASR possono essere differenziati in base al numero di parole che sono in grado di riconoscere. Esistono infatti sistemi dal vocabolario ristretto, minore di 100 parole, grande, con un numero di parole fra 1000 e 5000 e illimitato. Un vocabolario dalle dimensioni ristrette permette mag-

giore velocità di riconoscimento, maggiore acuratezza e riduce spesso drasticamente il tempo di addestramento richiesto dal sistema.

Esistono diverse applicazioni reali dove un vocabolario a dimensioni ristrette è più adatto, si pensi per esempio alla chiamata vocale per telefonia cellulare all'interno di autoveicoli [3].

Un punto chiave nella progettazione di sistemi di tipo *isolated-word* risiede nella corretta scelta delle parole al fine di migliorare l'accuratezza del sistema. Devono in generale essere scelte parole che siano *distanti* tra loro, dove con distanza tra le parole si intende un parametro fortemente legato alle caratteristiche estratte e al tipo di classificazione utilizzata nel dispositivo.

2.3.7 Estrazione delle caratteristiche e classificazione

Non esiste una metodologia univoca per l'estrazione delle caratteristiche, l'approccio dipende considerevolmente dal classificatore utilizzato e dall'ambiente operativo.

La maggior parte dei riconoscitori moderni utilizzano classificatori basati sui modelli HMM (Hidden Markov Model), cioè modelli statistici che hanno il vantaggio di essere abbastanza accurati [4]. L' HMM utilizza metodi convenzionali di addestramento basati sulla massima verosimiglianza. Ogni parola all'interno del vocabolario è specificata in termini di componenti del fonema. In fase di riconoscimento una procedura di ricerca è utilizzata per determinare la sequenza di fonemi con la massima verosimiglianza. Tale ricerca tiene conto esclusivamente delle sequenze di fonemi che corrispondono alle parole contenute nel vocabolario.

Anche le reti neurali possono essere utilizzate per effettuare la classificazione, richiedono, però, elevata capacità di calcolo, e quindi di solito

sono utilizzate per riconoscitori con vocabolari limitati.

Esiste anche la classificazione di tipo polinomiale che si adatta bene a sistemi con vocabolari ridotti. Questa tipologia di classificazione presenta una struttura relativamente semplice che si adatta bene ai moderni dispositivi DSP.

Un approccio alternativo utilizza le regole della grammatica formale per il riconoscimento di parole isolate. In informatica una grammatica formale è una struttura astratta che descrive un linguaggio formale in modo preciso e cioè un sistema di regole che delineano matematicamente un sistema (di solito infinito) di lunghezze finite (stringhe) usando un alfabeto (di solito finito). Le grammatiche formali sono chiamate così per analogia con la grammatica propria dei linguaggi adottati dall'uomo. Uno dei principali problemi per i sistemi di riconoscimento di tipo isolated-word è proprio quello di costruire un modello basato sulla grammatica formale che rappresenti correttamente le caratteristiche della sorgente generatrice di simboli.

2.3.8 Addestramento

Si è visto nel precedente paragrafo che un ASR presenta al suo interno un vocabolario di riferimento contenente i modelli delle parole. Tali modelli sono sviluppati durante la fase di addestramento, in cui il sistema impara ad identificare le diverse parole. L'accuratezza del sistema dipende dal tipo di addestramento, che può essere effettuato in accordo a diverse strategie. l'addestramento più semplice consiste nel ripetere le parole (più volte se necessario, come nel caso di speaker-dependent) senza badare al possibile rumore che potrebbe essere presente in ambiente operativo. Nel caso in cui il sistema non sia molto robusto al rumore e l'ambiente di lavoro risulti molto rumoroso,

si possono ottenere prestazioni non soddisfacenti. Per ovviare a questo problema si può eseguire l'addestramento direttamente nell'ambiente di lavoro, ovvero utilizzare delle registrazioni del rumore dell'ambiente di lavoro da riprodurre durante l'addestramento. Esistono database di registrazioni relative a diversi rumori ambientali, come cabine di pilotaggio di jet, carri armati, mitragliatrici ecc.. Un database molto utilizzato è il NOISEX [5], liberamente scaricabile. Questo tipo di addestramento però comporta una forte dipendenza dall'ambiente di lavoro, ed inoltre, in caso di rumori le cui caratteristiche statistiche varino nel tempo, potrebbe rivelarsi non sufficientemente adatto.

2.3.9 Il feedback

Per fornire un riscontro immediato al parlatore e per assicurargli che il sistema sta effettivamente lavorando, si dovrebbe fornire un riscontro (feedback) per ogni comando ricevuto, anche se non riconosciuto. Un meccanismo di feedback permette anche un miglioramento delle prestazioni poiché il parlatore è in grado di sapere se il comando precedentemente fornito è stato riconosciuto o meno.
Il feedback può essere di tipo visivo (testuale e grafico), uditivo (verbale e non verbale), tattile o essere una risposta di sistema quale un cambio del canale radio o la focalizzazione di una fotocamera. Esistono tre tipologie di feedback: reattivo, strumentale, operazionale.

Feedback reattivo nel feedback reattivo il sistema compie un'azione fisica in seguito alla sollecitazione del parlatore: quando ad esempio si preme un bottone esso si muove e scatta. Solitamente non c'è feedback reattivo quando viene utilizzato un microfono e lo speaker non può sapere se è in funzione, se sta parlando abbastanza forte o se il segnale sta raggiungendo il riconoscitore vocale

in modo adeguato. In realtà i primi riconoscitori vocali di tipo isolated-word utilizzavano un beep per indicare l'operatività del sistema, ma a lungo andare diventava un accorgimento irritante per il parlatore. Non esiste attualmente una soluzione valida per il feedback reattivo per i sistemi di riconoscimento vocale.

Feedback strumentale il feedback strumentale indica all'utente quali parole sono state riconosciute. Può essere simultaneo (dopo ogni parola, cifra, lettera) o finale (dopo ogni comando, frase, stringa di cifre). Entrambe le soluzioni presentano dei limiti, poiché la prima può compromettere la corretta enunciazione del comando, mentre la seconda risulta meno affidabile nella rivelazione dell'errore.

Feedback operazionale il feedback operazionale riguarda la risposta del sistema ai comandi vocali, come cambiare visuale su un display visivo o cambiare una frequenza radio.

2.4 Valutazione delle prestazioni

Le performance di un riconoscitore vocale sono solitamente specificate in termini di accuratezza. Alcune misure di accuratezza includono il tasso di errore sperimentato sulla singola parola (SWER - Single Word Error Rate) e il tasso di successo del comando (CSR - Command Success Rate).
Per i primi sistemi la misura più adottata fu quella dell'accuratezza di parola, che indicava la percentuale di parole correttamente riconosciute in una frase di comando. Questa statistica fu applicata per sistemi di tipo isolated-word. Per sistemi connected-word e continuous-speech si è utilizzato all'inizio una misura di accuratezza per la frase e una

per l'intento. Poiché gli utenti sono in grado di comunicare una intera stringa di comando, l'accuratezza di frase misura la percentuale di frasi complete riconosciute dal sistema [6].

Per permettere la flessibilità dei comandi vocali per sistemi di tipo continuous-speech, la sintassi del vocabolario può essere progettata per permettere a più parole di accedere allo stesso comando. In questo caso la misura di accuratezza di intento, o la percentuale di azioni corrette eseguite dal sistema, è la più appropriata. Questo tipo di misura dà un punteggio che non penalizza l'utente per aver utilizzato una sintassi alternativa. Quello che interessa è sapere se il sistema è in grado di identificare le intenzioni del comando dell'utente, indipendentemente da quante parole effettive sono state identificate. Nel caso di comandi che non hanno forme alternative, l'accuratezza di intento e quella di frase sono identiche.

Si può parlare inoltre di individuazione della parola chiave per raggiungere un'accuratezza di intento robusta nel caso di applicazioni command and control in ambienti rumorosi.

La progettazione e l'implementazione di test per definire le prestazioni di un sistema di riconoscimento vocale richiedono che l'attenzione sia posta sui vari fattori che influenzano le prestazioni.

2.4.1 Fattori relativi alla modalità del linguaggio

La modalità di linguaggio ha un grande effetto sulla difficoltà di riconoscimento: le parole separate sono più semplici da riconoscere , le parole connesse devono essere invece pronunciate con attenzione poiché è difficile riconoscerne l'inizio e la fine in quanto influenzate dalle parole adiacenti. Nel caso di sistemi continuous-speech il riconoscimento è più difficile perché il suono dei segmenti tende a mischiarsi e lo stress

turba la chiarezza delle vocali. Infatti i sistemi continuos-speech attualmente in commercio non fanno riconoscimento sulla parola, bensì sui fonemi.

2.4.2 Fattori relativi al parlatore

Esistono importanti differenze nel modo di parlare dei diversi individui dovute essenzialmente a:

Età la qualità della voce di un adolescente o di un anziano è sicuramente diversa da quella di un signore di mezza età.

Sesso alcune caratteristiche del parlato come ad esempio i picchi e la lunghezza dei tratti vocali tendono ad essere strettamente correlati con il sesso della persona per gli adulti.

Dialetto la pronuncia di alcune parole dipende dal dialetto, la documentazione di dati relativi al dialetto può essere particolarmente rilevante nel caso di sistemi speaker-independent.

Stress questo fattore influenza molto il linguaggio parlato perciò verrà trattato in modo più esteso più avanti. Qui accenniamo solo che è necessario soprattutto in ambito militare ottenere campioni del linguaggio parlato in queste condizioni, in modo che sia possibile stimare il conseguente degrado prestazionale.

2.4.3 Fattori relativi all' applicazione

La progettazione di un buon vocabolario per il successo dell'applicazione è una operazione fondamentale. Vocabolari di ampiezza limitata richiedono accuratezza nella pianificazione. Le parole devono essere adattate al contesto e ben distinte tra loro per assicurare il riconoscimento con poche sostituzioni ad-hoc in caso di errori non accettabili.

Le performance sono di gran lunga incrementate se vengono imposti dei vincoli sintatticim rendendo il compito di riconoscimento molto più semplice, veloce ed affidabile.

2.4.4 Fattori ambientali

Il segnale parlato è spesso affetto da rumore ambientale, riverbero e fenomeni di trasmissione di canale. Questi fattori ambientali possono portare a comportamenti errati e attivazioni non desiderate del riconoscitore. L'utilizzo di microfoni senza fili può portare a errori di riconoscimento dovuti al cross-talk del canale di trasmissione, interferenze a RF ,etc.

2.5 Sviluppo di una procedura di test

Un processo sistematico per la progettazione sperimentale e il testing è descritto di seguito:

- Selezionare un progetto sperimentale che fornisca un modello dell'applicazione o dei dati rappresentativi.
- Selezionare speakers che rappresentino gli utenti o un qualche sottoinsieme.
- Selezionare un vocabolario di test che esemplifichi l'utilizzo dell'applicazione o che sia stato utilizzato da altri per test di riferimento.
- Effettuare il training del sistema o costruire schemi di riferimento da usare per riconoscitori di tipo speaker-dependent.
- caratterizzare l'ambiente di test per documentare fattori quali

rumore, limitazioni di comunicazione di canale o fattori relativi al compito da svolgere.

- Registrare il materiale di test per permettere la verifica della validità dei risultati e il loro eventuale riutilizzo.
- Riuscire ad ottenere risultati di test; le procedure sono delineate sia per sistemi isolated-word che connected-word.
- Trarre considerazioni pratiche per assicurare che l'equipaggiamento stia lavorando nella maniera appropriata e che i test siano condotti in modo deguato ai vari fattori in gioco.
- Trarre considerazioni statistiche per indicare la validità statistica dei dati sulle prestazioni.
- Documentare le condizioni di test e i dati sulle performance ottenuti per permettere la valutazione dei dati pubblicati.

La progettazione dei test e i risultati ottenuti considerando i vari fattori saranno preziosi per identificare i punti di forza e le debolezze dei sistemi ASR.

2.6 ASR in ambito rumoroso

Questa sezione si occupa dei sistemi di riconoscimento vocale (ASR – Automatic Speech Recognition) in presenza di pertubazioni di vario genere e natura, genericamete indicate con il termine rumore.
Normalmente, soprattutto in ambito civile, la presenza del rumore è spesso trascurabile poiché si lavora in ambienti silenziosi, come uno studio di lavoro, dove il rumore non è presente. Esistono, però, alcune applicazioni di riconoscimento vocale che sono impiegate in ambienti

rumorosi, si pensi ad un sistema che riconosca le ordinazioni in un fast-food, oppure alla cabina di pilotaggio di un aereo o di un carro-armato, dove il rumore può essere molto elevato, oltre i 100 dB. In questo caso i normali ASR di utilizzo commerciale potrebbero non funzionare e c'è bisogno di appositi strumenti e tecniche per migliorare le prestazioni di un sistema ASR in presenza di rumore.

Esistono alcuni microfoni progettati appositamente per eliminare il rumore additivo dal segnale vocale. Tali microfoni, detti a cancellazione di rumore, riescono a limitare molto l'effetto del rumore anche se non possono contrastare efficacemente il cosiddetto *effetto Lombard*, ossia la modifica dell'emissione del segnale vocale involontariamente indotta in un ambiente rumoroso. Il parlatore, sentendo del rumore, tende involontariamente ad alzare il volume della voce e a distorcerla. Tale effetto può essere molto evidente in ambienti molto rumorosi come una cabina di pilotaggio di un aereo o di un carro armato.

Nel seguito verranno analizzate le degradazioni che dipendono dal rumore e dall'effetto Lombard e le metodologie usate per diminuire gli effetti negtivi introdotti da questi fenomeni. Verranno descritte le tipologie di rumore più comuni negli ASR e le rispettive degradazioni introdotte.

2.6.1 Il rumore negli ASR

Il rumore ambientale influisce negativamente sul riconoscimento vocale, sia attraverso la degradazione del segnale, sia influenzando la pronuncia delle parole. Situazioni tipiche di rumore possono sussistere all'interno di una cabina di un veicolo come un aereo, un elicottero, o un carro armato. Altri ambienti rumorosi possono risultare i campi di battaglia, caratterizzati da colpi di arma da fuoco e voci e urla di

altre persone.

Come si può evincere da questi semplici esempi, il rumore può essere di natura molto varia, con caratteristiche statistiche e spettrali molto diverse caso per caso. Si prenda come esempio il rumore di un elicottero, che può essere modellato come un processo quasi-stazionario [7], le cui caratteristiche statistiche variano lentamente nel tempo, ovvero il rumore dovuto ad armi da fuoco, di natura impulsiva. L'intensità del rumore può essere piuttosto elevata: un elicottero Black-Hawk raggiunge in cabina di pilotaggio un intensità pari a 103-107 dBA, mentre un elicottero Apache 110 dBA. Altri elicotteri come il CH-47 Chinook o un elicottero da trasporto arrivano fino a 123 dBA.

Mentre la maggior parte dei riconoscitori vocali fornisce ottime prestazioni in caso di assenza di rumore, le loro prestazioni degradano quando vengono applicati a situazioni reali. Uno dei motivi principali per cui si verifica questa degradazione delle prestazioni è la differenza sostanziale che c'è tra l'addestramento e la fase operativa.

2.6.2 Compensazione del rumore

I principali sforzi nel combattere il rumore sono diretti nel ridurre le differenze che sono presenti tra l'addestramento e le reali condizioni operative.

Esistono studi basati sull' utilizzo dei filtri di Kalman e di Wiener, che riescono con successo ad aumentare l'SNR in presenza di rumore, ma ciò non aumenta necessariamente la qualità o l'intelligibilità del dialogo nel riconoscimento vocale. Sembra plausibile che, nell'ottimizzare l'SNR, lo spettro del segnale si alteri in qualche modo e così vengano introdotte delle distorsioni che non necessariamente migliorano la prestazione di riconoscimento.

Per mitigare gli effetti del rumore si applica una trasformazione del segnale in modo da diminuire il disallineamento tra l'addestramento e il condizioni operative. Tale trasformazione può riguardare sia l'addestramento, per riallinearlo all'ambiente operativo, che il segnale ricevuto durante il normale funzionamento, in modo da *avvicinarlo* alla fase di addestramento.
Le principali trasformazioni applicabili possono essere suddivise in tre distinte categorie:

- Caratteristiche robuste: si assume che il sistema sia indipendente dal rumore e che si usi la stessa configurazione di sistema sia in presenza di rumore che in assenza di rumore. In questo caso si cerca di isolare delle grandezze caratteristiche e delle misure di distanza che siano robuste rispetto al rumore;
- Miglioramento del segnale vocale: si trasforma il segnale corrotto dal rumore in un segnale più simile a quello utilizzato per l'addestramento;
- Compensazione del modello per il rumore: i modelli di segnale creati al termine della fase di addestramento sono trasformati in modo da adattarsi all'ambiente rumoroso presente in condizioni operative.

Caratteristiche robuste al rumore

L'insieme di parametri in grado di ottenere dei buoni risultati nel riconoscimento vocale in assenza di rumore può risultare inefficace quando il rumore risulta presente; i parametri scelti potrebbero risultaremolto sensibili ai disturbi introdotti, inducendo un grosso disallineamento tra l'addestramento e l'ambiente operativo e portando ad una grossa perdita di prestazione. Si è mostrato teoricamente che in presenza di

rumore bianco additivo la norma dei coefficienti cepstrali diminuisce al diminuire dell'SNR [8]. Ciò porta, per un sistema addestrato in assenza di rumore, a delle drastiche perdite di prestazioni [9].

Si è cercato quindi di caratterizzare gli effetti del rumore sulle caratteristiche del segnale vocale piuttosto che più che focalizzarsi sulla rimozione del rumore stesso. Lo scopo ultimo di questi studi è trovare delle caratteristiche del segnale quanto più possibile resistenti al rumore.

Uno dei vantaggi principali di queste tecniche riguarda l'assenza (totale o quasi) di assunzioni circa la natura del rumore. In generale non è richiesta alcuna stima esplicita dei parametri statistici del rumore. D'altra parte il non utilizzare appieno le caratteristiche di uno specifico tipo di rumore chiaramente porta ad un degrado prestazionale rispetto ad un sistema che ne faccia uso.

Di seguito analizzeremo alcuni tipi di caratteristiche resistenti al rumore.

Rappresentazione acustica Nel dominio della Trasformata Discreta di Fourier (DFT – Discrete Fourier Transform), la combinazione del segnale vocale e del rumore può essere considerata additiva e perciò abbastanza facile da elaborare matematicamente. Tuttavia, si è visto che in presenza di distorsioni le prestazioni dei riconoscitori vocali peggiorano notevolmente se le caratteristiche sono nel dominio DFT rispetto al dominio cepstrale.

Come già precedentemente accennato, il rumore bianco riduce la norma dei vettori cepstrali. Per compensare questa diminuzione, nella distanza euclidea cepstrale tra un vettore rumoroso di test ed il vettore di riferimento in assenza di rumore si inserisce un fattore di scala. Il

valore ottimo del fattore di scala è la proiezione ortogonale del vettore di test sul vettore di riferimento.
Tuttavia, non solo la norma dei vettori cepstrali è modificata dal rumore, ma anche altri parametri statistici dei coefficienti cepstrali ne risultano influenzati. All'aumentare del rumore bianco, il valor medio si modifica, la varianza si riduce e le distribuzioni tendono a scostarsi dalla distribuzione normale. Questo limita i miglioramenti introdotti dalla compensazione della norma.

Analisi del discriminante lineare L'analisi del discriminante lineare consiste nel trovare una trasformazione lineare dello spazio dei parametri che minimizzi la varianza all'interno di una classe e massimizzi la varianza tra le classi. L'analisi dei componenti principali è quindi applicata alla matrice di covarianza inter-classe per selezionare la direzione con la maggiore varianza. Un sottoinsieme di queste componenti principali è poi usato per formare i vettori di parametri trasformati. Un tipo di analisi molto frequente è denominata Discriminante Lineare di Fisher (FLD – Fisher Linear Discriminant) ed è di seguito descritta. La funzione del discriminante lineare di Fisher consiste nel separare gli insiemi dei vari vettori tramite una proiezione su un asse comune. L'asse è scelto in modo tale da discriminare nel modo migliore possibile le classi di vettori. Si supponga per semplicità di esposizione l'esistenza di due sole classi di vettori con stessa cardinalità pari a l. Si indichi il generico vettore j-esimo con il simbolo v_j^F dove F è la dimensione dello spazio delle caratteristiche. Indicando con v^1 e con v^2 rispettivamente i generici vettori associati alla classe 1 e alla classe 2 si definiscono gli insiemi delle due classi:

$$P_1 = \{v_j^1\} \qquad j = 1, ..., l \tag{2.1}$$

$$P_2 = \{v_j^2\} \qquad j = 1, ..., l \tag{2.2}$$

La fase di addestramento si conclude con la ricerca di un particolare $\bar{v}$ su cui proiettare gli insiemi P_1 e P_2, in modo da ottenere due regioni di punti $P_1^{'}$ e $P_2^{'}$ il più possibile raggruppate e disgiunte tra loro. Questo metodo ha il pregio di ridurre le dimensioni dei dati (i vettori delle caratteristiche) in ingresso preservandone però la discriminabilità. Più in generale si può descrivere il discriminante lineare di Fisher come quella trasformazione matematica che dallo spazio delle caratteristiche porta all'insieme dei numeri reali, preservando però la discriminabilità delle classi.

Misure di similarità Esistono molte tecniche per discriminare tra differenti distribuzioni di vettori di caratteristiche. Il classificatore a percettrone multistrato (MLP – Multi-Layer Perceptron) [10], ad esempio, offre due importanti vantaggi sugli altri metodi:

- non è assunta alcuna distribuzione di probabilità.
- si possono creare, tramite l'addestramento, delle mappature arbitrariamente complesse.

In un esperimento eseguito allo scopo di classificare singoli suoni vocali ottenuti da diversi parlatori in presenza di rumore, si è visto che un classificatore MLP che utilizzi coefficienti cepstrali ottiene un'accuratezza decisamente migliore a tutti i livelli di SNR rispetto ad un classificatore a massima verosimiglianza gaussiana multivariata e ad un classificatore kNN (k-nearest Neighbourhood). In aggiunta, il degrado delle prestazioni per il classificatore MLP avviene ad un livello di SNR più basso rispetto agli altri due classificatori.

Rimozione di variazioni lente Molti rumori additivi, così come le distorsioni della maggior parte dei canali, variano lentamente se comparati alle variazioni del segnale vocale. Filtrare le lente variazioni nei vettori delle caratteristiche può migliorare l'accuratezza del riconoscitore vocale significativamente. Il filtraggio può avvenire in domini differenti, come ad esempio nello spettro di potenza logaritmico oppure nel domino cepstrale.
Un tipo di elaborazione è detta RASTA (RelAtive SpecTrAl) e consiste nel sopprimere sfasamenti additivi costanti in ogni componente log spettrale dello spettro a breve termine del segnale. Questo metodo di analisi può essere applicato anche a parametri mel-cepstrali. Ogni banda di frequenza è filtrata da un filtro con un'alta attenuazione intorno alla frequenza zero. RASTA ha mostrato di poter ottenere buone prestazioni anche con variazioni del microfono dalla fase di addestramento a quella operativa. Il modo più semplice ed efficiente per rimuovere le variazioni lente è tramite il CMN (Cepstrum Mean Normalization) che rimuove il valor medio vettore dei coefficienti cepstrali. Si è visto che il metodo CMN ha prestazioni simili a RASTA, non degrada le prestazioni in caso di assenza di rumore e migliora l'accuratezza in caso di cambio di microfono.

Miglioramento del segnale vocale

Come passo di pre-elaborazione per il riconoscimento vocale, le tecniche di miglioramento del segnale vocale sono intese come recupero o della forma dell'onda oppure del vettore delle caratteristiche del segnale in assenza di rumore. Queste tecniche fanno uso differente di informazioni a priori sul segnale vocale e sul rumore. Il criterio utilizzato nelle tecniche di miglioramento del segnale vocale è basato o sulla distribuzione di probabilità del segnale vocale in assenza di rumore,

oppure sulla distorsione tra il segnale vocale in assenza di rumore e il segnale registrato in condizioni operative e usualmente non direttamente correlato alla funzione di riconoscimento vocale. La maggior parte delle tecniche presentate sono state sviluppate originariamente per miglioramenti della qualità del segnale vocale piuttosto che per il riconoscimento vocale. Tuttavia, possono anche essere utilizzate come passo di pre-elaborazione nel riconoscimento vocale. Le tecniche per il riconoscimento vocale possono essere divise in diverse categorie:

- Sottrazione spettrale;
- Mascheramento del rumore;
- Stima bayesiana;
- Modellizzazione parametrica spettrale.

Sottrazione spettrale Il metodo della sottrazione spettrale assume che il rumore e il segnale vocale siano incorrelati e additivi nel dominio temporale. In questo caso, lo spettro di potenza del segnale ricevuto è la somma dello spettro del rumore e dello spettro del segnale vocale in assenza di rumore. Il metodo assume anche che le caratteristiche statistiche del rumore varino lentamente nel tempo rispetto al segnale vocale, in modo tale che lo spettro del rumore stimato durante un periodo di silenzio possa essere utilizzato per sopprimere il rumore durante la produzione del segnale vocale.
Il metodo è semplice e efficiente per rumore additivo stazionario o quasi-stazionario a banda larga, ma purtroppo soffre di diversi problemi:

- Le prestazioni di un sistema a sottrazione di rumore si basano anche su un sistema di classificazione rumore/segnale vocale. Un'er-

rata classificazione può portare ad una stima errata del modello di rumore e quindi ad una degradazione della stima del segnale vocale;

- La sottrazione potrebbe portare a valori dello spettro di potenza negativi, che quindi sono posti a valori non negativi tramite un meccanismo a soglia. Questa operazione non lineare porta a del rumore residuo detto comunemente rumore musicale;

- Le tecniche di sottrazione dello spettro non possono essere utilizzate nel dominio spettrale logaritmico, perché il rumore, anche se incorrelato con il segnale vocale nel dominio temporale, diventa dipendente dal segnale.

Mascheramento del rumore Il mascheramento del rumore è un fenomeno psicologico di riduzione della percezione di un segnale vocale in presenza di rumore. In generale, le persone non riescono a rivelare uno stimolo acustico il cui livello sia più basso della soglia generata da altri stimoli. Quando si ascolta un segnale vocale in ambiente rumoroso, l'effetto del rumore può essere decrementato dal mascheramento. L'effetto di questo fenomeno può essere emulato per migliorare le prestazioni del segnale vocale in ambiente rumoroso. L'uscita di un banco di filtri è rimpiazzata da una maschera se l'uscita è sotto una certa soglia. L'operazione decrementa le fluttuazioni spurie che portano informazione per la maggior parte riguardante il rumore.
Ad alti livelli di rumore, il mascheramento non ha effetti positivi poiché in questo caso anche le bande di frequenza del segnale ad alta energia potrebbero avere energia più bassa del rumore. Il mascheramento assume che il rumore di fondo sia conosciuto con certezza e che alcuna

osservazione sul segnale vocale può essere estratta dalle osservazioni che sono sotto quel livello.

Stima Bayesiana Se si considera il segnale vocale, in particolare il vettore delle caratteristiche, come una grandezza aleatoria con una sua distribuzione, che viene osservata in presenza di rumore, allora si può applicare la stima bayesiana per ottenere una stima del vettore delle caratteristiche generato.

La stima bayesiana ha come caratteristica una funzione di costo che varia a seconda del criterio adottato. Per la stima del vettore si possono utilizzare quindi molte funzioni di costo. Le funzioni di costo più comuni sono la funzione di costo di errore quadratico che è utilizzata nello stimatore a minimo errore quadratico medio (MMS – Minimum Mean Square), e la funzione di costo uniforme che è utilizzata nella stimatore a massima probabilità a posteriori (MAP – Maximum A Posteriori probability).

Modellizzazione parametrica spettrale Un modello che è specifico per segnali vocali accetterà segnali con le proprietà del segnale vocale e rifiuterà il rumore: in pratica è un sistema robusto al rumore. Assumendo una certa struttura spettrale del segnale vocale, un modello AR è utilizzato per migliorare la qualità della rappresentazione del segnale vocale, i cui parametri sono ottenuti da una procedura iterativa implementante una stima MAP. La procedura stima il segnale vocale tramite filtro di Wiener e computa il modello AR da tale stima. Come miglioramento della procedura iterativa un classificatore HMM può essere utilizzato allo scopo di suddividere i campioni del segnale vocale in classi fonetiche e il classificatore termina anche la stima iterativa.

Compensazione del modello per il rumore

Piuttosto che tentare di derivare uno stimatore per il segnale vocale, si può utilizzare un approccio diametralmente opposto, come modificare il modello generato dall'addestramento in modo da adattarsi alla presenza del rumore.
Un HMM fornisce una base matematica per modellare variazioni temporali e spettrali nei segnali vocali. Nel contesto degli HMM esistono algoritmi ampiamente usati, aventi un carico computazionale sopportabile, per stimare le caratteristiche del segnale vocale e per eseguire il riconoscimento vocale. Riconoscitori vocali basati su HMM offrono anche la possibilità di sfruttare un determinato modello ottenuto dall'addestramento, e poi di cambiare i parametri del modello per compensare la presenza del rumore. In altre parole un HMM può utilizzare il segnale vocale sporcato dal rumore per adattare i parametri del modello, come ad esempio medie e varianze, per compensare la discrepanza che si presenta tra addestramento e reali condizioni operative.
Tali schemi adattativi sono potenzialmente in grado di lavorare in ambienti rumorosi che non sono stati riprodotti durante la fase di addestramento oppure che hanno caratteristiche tempo-varianti. Tecniche di compensazione del modello permettono l'ottimizzazione dei parametri del modello per fornire la migliore accuratezza possibile.

2.7 Effetto Lombard

La presenza del rumore oltre che a degradare la qualità del segnale influenza anche la produzione vocale. L'effetto Lombard può essere identificato con lo sforzo e l'alterazione vocale che lo speaker è indotto a compiere quando parla in presenza di rumore.

In genere il parlatore adatta l'intensità della sua voce a seconda di come essa viene percepita dal suo stesso udito. In presenza di rumore superiore a 50 dB(A) (dB(A) – misura in dB adattata alla risposta in frequenza dell'orecchio umano) un parlatore normoudente solitamente aumenta l'intensità della sua voce da 3 a 6 dB per ogni incremento di 10 dB del rumore mascherante il messaggio verbale.

Per costruire sistemi di riconoscimento vocale che abbiano delle buone prestazioni in ambienti rumorosi è necessario tenere in considerazione le differenze acustico-fonetiche che esistono tra il parlato normale e quello prodotto in presenza di rumore. Alcuni studi hanno dimostrato che l'effetto Lombard influenza la chiarezza del parlato, in particolare sono stati presi in considerazione alcuni parametri quali lo sforzo vocale, la durata delle parole, la dimensione del vocabolario. È stato mostrato inoltre che la chiarezza aumenta nel caso di parole lunghe e di un vocabolario piccolo, e che diminuisce quando il parlatore incrementa il suo sforzo vocale fino ad un livello che corrisponde ad un parlato urlato.

L'effetto Lombard modifica la frequenza, l'intensità, la durata ed altre componenti del suono. Dopo aver condotto analisi acustiche sui fonemi per circa 40 parametri, sono state riscontrate molte differenze a seconda che il parlatore sia di sesso maschile o femminile. I parametri selezionati corrispondono ad alcune caratteristiche rappresentative a cui sono sensibili gli attuali riconoscitori vocali: larghezza di banda, pendenza spettrale, energia, etc. Inoltre la variabilità dei fonemi è legata al contesto e ci sono importanti differenze tra i vari speaker che producono l'enunciazione in presenza di rumore.

Per compensare l'effetto Lombard sono state sviluppate varie tecniche: le più diffuse sono descritte in [14]. Un meccanismo di feedback audio della voce può mitigare l'effetto Lombard: infatti il parlatore, sentendo

la sua voce in cuffia non ha più bisogno di alzare l'intensità vocale [5], riducendo, o addirittura annullando, l'effetto Lombard.

2.8 Stress

In determinate applicazioni le operazioni di riconoscimento vocale sono spesso condotte in condizioni di stress fisico e mentale. Il rendimento di un ASR è fortemente influenzato in modo negativo dallo stress: è quindi necessario studiare questo fenomeno in modo approfondito.
Lo stress può essere indotto da un grosso carico di lavoro, o dalla mancanza di riposo, oppure da emozioni come paura, dolore, tensione psicologica e altre condizioni dovute allo stato di guerra (in ambito militare) ovvero da stati emotivi come rabbia o paura. È riconosciuto che queste condizioni alterano, anche significativamente, la produzione del linguaggio umano.
Essere sotto stress significa principalmente che esiste qualche forma di pressione applicata a chi parla, che sfocia nella perturbazione del processo di produzione della voce e dunque del segnale acustico. Il problema nasce quando il sistema, addestrato in un ambiente privo di stress, si trova ad operare in un ambiente con caratteristiche diverse e con condizioni dell'utente che sono variate rispetto a quelle con cui è stato effettuato l'addestramento.
Nelle applicazioni militari in particolare, l'ambiente di riferimento in cui è eseguito l'addestramento è normalmente la base militare, un ambiente confortevole e con poco stress quindi, ma durante le operazioni militari il parlatore, il soldato, opererà in ambienti molto differenti dalla base militare e sarà a sottoposto a stress di vario tipo come sforzo fisico, paura, fatica, necessità di non fare rumore, etc. . .
La causa di stress nel parlatore è indicata con il termine stressore. Il

processo di produzione vocale funziona *a strati* e gli stressori influiscono su questo strati modificando la produzione vocale. E' quindi possibile classificare gli stressori in base allo strato che influenzano. Esistono quattro livelli di stressori:

Ordine 0 sono gli stressori i cui effetti hanno una relazione diretta e fisica con la produzione vocale. Sono gli stressori più facili da comprendere.

Ordine 1 questi stressori causano modifiche fisiologiche nell'apparato di produzione vocale, alterando la traduzione dei comandi neuromuscolari nei movimenti dell'articolazione.

Ordine 2 sono stressori che condizionano la conversione del linguaggio in comandi neuromuscolari. Questo livello può essere descritto come percettivo, poiché coinvolge la percezione della necessità di cambiare la produzione vocale.

Ordine 3 gli stressori del terzo ordine sono quelli i cui effetti influenzano gli strati più alti del livello di produzione vocale. Uno stimolo esterno è soggetto all'interpretazione mentale e alla valutazione, di solito come minaccia (da qui il termine *stress*), ma anche altri stati emozionali come felicità e gioia hanno il loro effetto a questo strato.

2.8.1 Parametri del parlato sotto stress

I cambiamenti del parlato sotto condizioni di stress sono ancora oggi poco chiari. Le ricerche si sono concentrate principalmenti sui singoli parametri che sono di seguito descritti.

Tono

Gli studi hanno interessato gli assestamenti della forma del tono, l'analisi statistica del valor medio del tono, la varianza e la sua distribuzione. Le conclusioni sono le seguenti:

- I valori medi del tono possono essere utilizzati come indicatori di stili di dialogo come ad esempio un dialogo arrabbiato, tranquillo, veloce, chiaro, con effetto Lombard, interrogativo, urlato, quando sono comparati con condizioni neutre;
- I valori medi di tono per l'effetto Lombard, per urla, per domande o per arrabbiature sono estremamente differenti da tutti gli altri stili di dialogo;
- Il dialogo prodotto dall'effetto Lombard ha dei valori medi di tono molto associati ai toni di dialoghi veloci e chiari;
- La varianza del tono per i dialoghi forti e per le urla è completamente differente dalle varianze degli altri stili di dialogo;
- La varianza del tono per dialoghi chiari è simile alla varianza per l'effetto Lombard, ma entrambe sono molto differenti dalle varianze per tutti gli altri stili di dialogo considerati.

Durata

Solo da poco si è cominciati a studiarla, valutando statisticamente la durata di classi di fonemi. L'analisi della durata è condotta su parole intere oppure su classi di fonemi (vocali, consonanti, semivocali e dittonghi). Una lista parziale di conclusioni è riportata qua di seguito:

- La durata media di una parola è un indicatore significativo di

dialoghi lenti, chiari, arrabbiati, urlati, con effetto Lombard, o veloci, comparati con uno stile di dialogo neutro;

- La durata media di parola per stili di dialogo veloce e lento sono significativamente diversi da tutti gli altri stili;
- La durata media di una consonante chiara è significativamente differente da tutti gli altri stili, eccetto lo stile di dialogo lento;
- La varianza della durata per tutte le classi considerate (parole intere, vocali, consonanti, semivocali, dittonghi) incrementa sotto uno stile di dialogo lento;
- La varianza della durata per la maggior parte delle classi considerate decrementa per un stile di dialogo veloce;
- La durata della varianza incrementa significativamente per dialoghi arrabbiati;
- La varianza della durata per consonanti in stili di dialogo chiari è significativamente diversa da tutti gli altri stili.

Intensità

Alcune analisi sono state condotte sull'intensità di intere parole e sull'intensità di classi intere di fonemi. Test statistici sono stati eseguiti sul valore medio, sulla varianza, e sulla distribuzione. Una lista parziale di conclusioni è mostrata di seguito:

- Le intensità medie di parola sono indicatori significativi per dialoghi arrabbiati e urlati comparati a condizioni neutrali;
- Le intensità media di parola per dialoghi arrabbiati ed urlati sono significativamente diversi da tutti gli altri stili considerati;

- Le intensità media di vocali e dittonghi sono significativamente diverse da tutti gli altri stili considerati;
- Le intensità media di consonanti e semivocali non sono indicatori di nessuno stile di dialogo;
- La varianza dell'intensità di parola è una significativa indicatrice di dialoghi arrabbiati e urla;
- La varianza dell'intensità di parola per dialoghi arrabbiati ed urla è significativamente differente da tutti gli altri stili.

Capitolo 3

Analisi delle features

3.1 Introduzione

Una parte fondamentale di un sistema di riconoscimento vocale è quella che codifica le porzioni di segnale, provenienti dal microfono, attraverso una tra le possibili tecniche di analisi del segnale vocale orientate a questo scopo. Queste tecniche mettono in risalto le caratteristiche peculiari della porzione analizzata.

Il problema della codifica è fondamentale per un ASR, in quanto da essa dipende la possibilità di discriminare un'entità linguistica da un'altra. Una codifica breve e carica di informazione è la migliore possibile perché consente un buon riconoscimento con il minimo addestramento.

Nel proseguo del capitolo verrano desritti i sistemi di analisi utilizzati nella fase implementativa della tesi. Questo particolare ramo di analisi si occupa di estrarre informazione discriminante da una porzione di segnale, allo scopo di ottenerne una codifica utile per una post elaborazione da parte di un calcolatore. Si passa, dunque, da una successione di ampiezze rappresentanti variazioni di pressione dell'aria, ad un vettore di numeri che si riferisce a particolari caratteristiche di una zona del segnale.

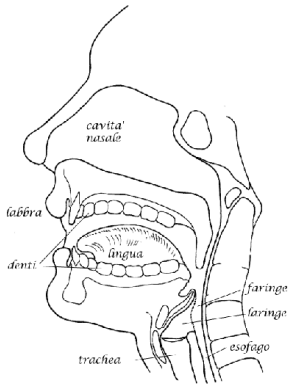

Figura 3.1: Schema apparato fonatorio

3.2 Modello generale tempo-discreto per la produzione della voce

Il sistema vocale umano è composto dal tratto vocale, che inizia all'apertura tra le corde vocali, o glottide, e termina alle labbra, e che a sua volta si compone di faringe e bocca o cavità orale. In media la lunghezza del cavo orale è di 17 cm [11]. L'area della sezione trasversale del tratto vocale, determinata dalla posizione della lingua, delle labbra, della mascella e del palato soffice, varia da zero a circa 20 cm^2, (Fig 3.1). Il tratto nasale comincia dal palato soffice e termina alle narici. Quando il palato soffice è abbassato, il tratto nasale è acusticamente accoppiato al tratto vocale per produrre i suoni nasali del parlato. Nello studio del processo di produzione del parlato è utile estrapolare importanti caratteristiche del sistema fisico in maniera tale da realizzare un modello matematico realistico.

La maggior parte dei foni che produciamo deriva dal flusso d'aria che fuoriesce dai polmoni durante l'espirazione (flusso polmonare egressivo, cioè diretto verso l'esterno). L'italiano usa solo questo tipo di flusso, altre lingue usano anche quello ingressivo, in cui il flusso d'aria è diretto verso l'interno. Il flusso d'aria polmonare egressivo sale

dai polmoni, ai bronchi, alla trachea, fino alla laringe; nella laringe si trova la glottide che contiene due membrane muscolari, le corde vocali, le quali possono essere accostate/vicine o aperte/lontane.

I suoni del linguaggio parlato possono essere classificati in classi distinte secondo il loro modo di eccitazione. I suoni sonori sono prodotti forzando l'aria attraverso la glottide con la tensione delle corde vocali posizionate in modo da vibrare con moto oscillatorio rilassato, producendo impulsi di aria quasi periodici che eccitano il tratto vocale. I suoni sordi (o fricativi) sono generati formando una costrizione allo stesso punto del tratto vocale e forzando l'aria attraverso tale costrizione ad una velocità sufficientemente elevata da produrre turbolenza, inoltre sono aperiodici.

Possiamo distinguere tra due categorie principali di suoni: le consonanti e le vocali. Nel percorso dalla glottide verso l'esterno il flusso d'aria polmonare può essere libero oppure più o meno ostruito. Nel caso in cui l'aria fluisca liberamente verso l'esterno otteniamo dei suoni vocalici; nel caso invece che vi sia una qualche ostruzione - parziale o totale - del flusso da parte degli articolatori, otteniamo dei suoni consonantici. Le lingue tendono a preferire parole caratterizzate da una sequenza di articolazioni di chiusura ed apertura, cioè di consonanti e vocali; il tipo di sillaba più comune nelle lingue è CV, cioè quello formato da una consonante (C) e da una vocale (V).

Vocali

Le vocali vengono prodotte modificando la forma del cavo orofaringeo, ma senza ostruire in maniera rilevante il corso del flusso d'aria proveniente dai polmoni, cosicché il suono prodotto dalla vibrazione delle corde vocali assume risonanze distinte, ma non viene trasformato in

rumore (come nel caso delle consonanti). Le vocali vengono prodotte con il dorso della lingua che avanza o arretra, e si alza ed abbassa, senza però creare una costrizione del flusso d'aria. Inoltre, il movimento della lingua può essere accompagnato da una concomitante protrusione ed arrotondamento delle labbra. Dunque, semplificando, le vocali possono essere posizionate in uno spazio tridimensionale, detto spazio vocalico, sulla base di tre parametri:

- posteriorità: chiamiamo anteriori le vocali prodotte con un avanzamento del dorso verso il palato, posteriori quelle prodotte con un arretramento del dorso verso il velo, centrali quelle prodotte senza avanzamento o arretramento.
- altezza: a seconda del grado di innalzamento della lingua (rispetto ad uno stadio di riposo che è più o meno centrale), possiamo distinguere tra vocali alte, medio-alte, medio-basse e basse.
- arrotondamento: le vocali prodotte con una protusione/ arrotondamento delle labbra si dicono arrotondate; le altre sono non arrotondate.

Consonanti

Possiamo classificare le consonanti sulla base di tre parametri:

- modo di articolazione, cioè in base al modo in cui gli articolatori producono la chiusura parziale o totale della cavità orofaringea;
- luogo/punto di articolazione, cioè in base a quali articolatori producono la chiusura parziale o totale della cavità orofaringea;
- sonorità, cioè in base al fatto che le corde vocali siano separate o accostate e in vibrazione durante la produzione del suono.

Occlusive

Un organo mobile (generalmente la lingua) tocca un organo fisso, ostruendo completamente il passaggio dell'aria: il suono è così la piccola esplosione che si ottiene rilasciando bruscamente l'ostruzione; ad esempio, per il suono t la lingua premuta contro i denti impedisce il passaggio dell'aria: appena la posizione si rilassa, l'aria esplode con un rumore caratteristico che è appunto quello della t.

Fricative

Gli organi articolatori sono avvicinati l'uno all'altro ma non si toccano: l'aria è così costretta a passare attraverso uno stretto canale producendo un fruscio, che è poi il suono specifico.

Affricate

Sono i suoni che combinano le caratteristiche degli occlusivi con quelle dei fricativi: nella z della parola italiana azione, ad esempio, l'ostruzione non viene liberata del tutto con un'esplosione: i due organi rimangono vicini e l'aria fruscia fra essi, come se si pronunciassero una t e una s a distanza molto ravvicinata.

Nasali

Se durante la produzione di un suono entrano in risonanza anche le cavità del naso (questo avviene se si abbassa il velo palatino e si lascia che l'aria entri nel naso) i suoni che derivano sono detti nasali.
Sia il tratto vocale che quello nasale possono essere visti come tubi con area della sezione trasversale non uniforme. Quando il suono, generato come discusso sopra, si propaga attraverso questi tubi lo spettro di frequenza viene sagomato dalla selettività in frequenza del tubo.

Questo effetto è molto simile a quello di risonanza che si osserva negli strumenti a fiato. Nel contesto della produzione vocale le frequenze di risonanza del tratto vocale sono chiamate frequenze formanti o semplicemente formanti. Le frequenze formanti dipendono dalla forma e dalle dimensioni del tratto vocale; ogni forma è caratterizzata da un set di frequenze formanti. I differenti suoni sono prodotti variando la forma del tratto vocale, perciò le proprietà spettrali del segnale vocale variano con il tempo al variare della forma del tratto vocale.

Le leggi su cui si basa la produzione dei suoni sono:

1. Variazione nel tempo della forma del tratto vocale
2. Perdite dovute alla viscosità delle pareti del tratto vocale e alla conduzione termica
3. Morbidezza delle pareti del tratto vocale
4. Radiazione del suono alle labbra
5. Accoppiamento nasale
6. Eccitazione del suono nel tratto vocale

Il tratto vocale impone le sue risonanze sulla sorgente di eccitazione per produrre i differenti suoni del parlato. Siamo interessati a costruire un modello tempo-discreto per rappresentare segnali del parlato campionati, cioè un sistema lineare la cui uscita ha proprietà simili a quelle del parlato se controllato da un set di parametri che sono in qualche modo correlati con il processo di produzione della voce. Il modello dunque avrà un'uscita simile a quella del modello fisico, ma la sua struttura interna non imita la produzione fisica della voce. Per produrre segnali simili a quelli del parlato l'eccitazione e le proprietà di risonanza del sistema lineare devono cambiare con il tempo. Per

molti suoni del parlato è ragionevole assumere che le proprietà generali dell'eccitazione e del tratto vocale rimangano fisse per un periodo di 10-20 msec. Perciò il modello è composto da un sistema lineare lentamente tempo-variante eccitato da un segnale la cui natura di base cambia da impulsi quasi periodici per suoni sonori a rumore bianco per suoni sordi.

Il tratto vocale è modellizzato come un tubo di sezione trasversale non uniforme e tempo variante. Per le frequenze corrispondenti a lunghezze d'onda comparabili con le dimensioni del tratto vocale (meno di circa 4000 Hz), è ragionevole assumere la propagazione di onde piane lungo gli assi del tubo. Un modello di soli poli è una buona rappresentazione degli effetti del tratto vocale per la maggior parte dei suoni del parlato; comunque la teoria acustica ci dice che le nasali e le fricative richiedono sia poli che zeri per quanto riguarda la funzione di trasferimento. In questi casi possiamo o includere gli zeri nella funzione di trasferimento o possiamo ragionevolmente includere più poli, approccio che nella maggior parte dei casi risulta preferibile.

La funzione di trasferimento può essere così rappresentata :

$$V(z) = \frac{G}{1 - \sum_{k=1}^{N} \alpha_k z^{-k}} \tag{3.1}$$

dove G e α_k dipendono dalla funzione d'area.

Da sommarsi alla risposta del tratto vocale per completare il modello c'è la rappresentazione del cambiamento della funzione di eccitazione e degli effetti della radiazione del suono attraverso le labbra. Se vogliamo ottenere un modello per la pressione al livello delle labbra, allora dobbiamo includere gli effetti della radiazione. La pressione è in relazione con la velocità di volume tramite un'operazione di filtraggio passa-alto. Infatti a basse frequenze possiamo considerare che la

pressione è approssimativamente la derivata della velocità del volume. Perciò per ottenere una rappresentazione tempo discreta di questa relazione dobbiamo utilizzare una tecnica di digitalizzazione che evita l'aliasing. Un'approssimazione ragionevole degli effetti della radiazione è data da :

$$R(z) = R_0(1 - z^{-1}) \tag{3.2}$$

Il carico della radiazione può essere messo in serie al modello del tratto vocale.

Per completare il modello dobbiamo discutere i modi in cui viene generato un input appropriato per il sistema composto dal tratto vocale e dalla radiazione. Ricordando che la maggior parte dei suoni del parlato possono essere classificati come sordi e sonori, possiamo dire in termini generali che la sorgente necessaria deve essere tale da poter produrre sia un'onda di impulsi quasi periodici o di rumore casuale. Questo segnale eccita un sistema lineare la cui risposta impulsiva $g(n)$ è quella glottale. Un guadagno di controllo, A_v, regola l'intensità dell'eccitazine vocale. La scelta della forma di $g(n)$ non è particolarmente critica grazie alle proprietà della trasformata di Fourier. È stato trovato che la naturale forma d'onda glottale può essere rimpiazzata da una forma d'onda del tipo

$$\begin{aligned} g(n) &= \frac{1}{2}[1 - cos(\pi n/N_1)] \qquad 0 \leq n \leq N_1 \\ &= cos(\pi(nN_1)/2N_2) \qquad N_1 \leq n \leq N_1 + N_2 \\ &= 0 \qquad\qquad\qquad\qquad altrove \end{aligned} \tag{3.3}$$

Il modello completo è mostrato nella figura 3.2.

Commutando tra i generatori di eccitazione sordi e sonori possiamo modellare il cambiamento nel modo di eccitazione.

Nel caso della predizione lineare è conveniente combinare i componenti dell'impulso glottale, della radiazione e del tratto vocale tutti insieme

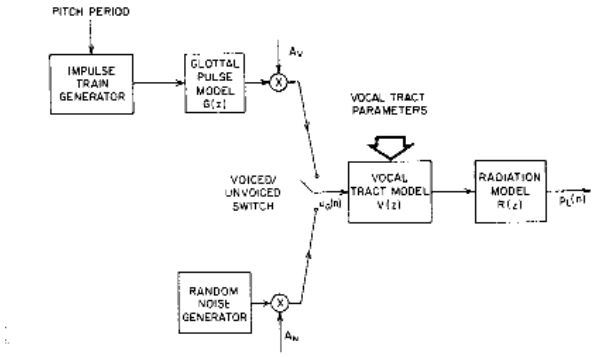

Figura 3.2: Modello generale tempo discreto per la produzione del parlato

e rappresentarli in una singola funzione di trasferimento :

$$H(z) = G(z)V(z)R(z) \tag{3.4}$$

composta da soli poli.

Ovviamente questo modello ha delle limitazioni, ma fortunatamente ciò non inficia seriamente la sua applicabilità. In primo luogo dobbiamo considerare la variazione temporale dei parametri. In questo modello possiamo assumere che essi siano costanti per intervalli di tempo tipicamente compresi fra 10-20 ms. La funzione di trasferimento $V(z)$ serve per definire la struttura di un modello i cui parametri variano lentamente nel tempo. Una seconda limitazione è la mancanza di zeri necessari in teoria per le nasali e le fricative. Questa è una limitazione severa per le nasali ma non per le fricative. Gli zeri possono essere inclusi nel modello se lo si desidera. Una terza limitazione è data dalla semplificazione nella dicotomia dell'eccitazione (voiced-unvoiced) che è inadeguata in quanto le fricative sono correlate con i picchi del flusso glottale. Infine una limitazione relativamente minore del modello richiede che gli impulsi glottali siano spaziati da un multiplo intero del periodo di campionamento, T.

3.3 Linear Predictive Coding

Una delle più efficienti tecniche di analisi per la voce è quella della linear prediction coding [11]. L'importanza di questo metodo risiede sia nell'abilità di fornire stime accurate dei parametri della voce sia nella sua relativa velocità di computazione. L'idea di base della predizione lineare è quella secondo cui un campione di voce può essere approssimato tramite una combinazione lineare dei campioni passati. Minimizzando l'errore quadratico tra un campione di voce ricevuto e quello calcolato tramite la linear prediction , su un intervallo finito, può essere determinato un set unico di coefficienti predetti.

La voce, come già detto in precenza, può essere modellata come l'uscita di un sistema lineare tempo variante eccitato da impulsi quasi periodici o da un rumore casuale. Il metodo della predizione lineare fornisce un approccio robusto e accurato per stimare i parametri che caratterizzano il sistema lineare tempo variante. Gli effetti spettrali derivanti dalla radiazione, dal tratto vocale, e dall'eccitazione glottale sono rappresentati da un filtro digitale tempo variante la cui funzione di trasferimento è la seguente:

$$H(z) = \frac{S(z)}{U(z)} = \frac{G}{1 - \sum_{k=1}^{p} a_k z^{-k}} \tag{3.5}$$

Questo modello è adatto per rappresentare suoni vocali non nasali in quanto sono presenti solo poli. Per rappresentare anche suoni nasali andrebbero inseriti oltre ai poli anche degli zeri. Nel caso in cui però l'ordine p è sufficientemente elevato anche il modello con soli poli può essere adatto a rapprentare la maggior parte dei suoni del linguaggio vocale. Il vantaggio di questo modello è che il guadagno G e i coefficienti del filtro a_k possono essere stimati in maniera computazionalmente

efficiente con il metodo della predizione lineare. I campioni della voce s(n) sono collegati con l'eccitazione u(n) dalla semplice equazione:

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) + Gu(n) \tag{3.6}$$

Un predittore lineare con coefficienti di predizione α_k, è definito come un sistema la cui uscita è :

$$\tilde{s}(n) = \sum_{k=1}^{p} \alpha_k s(n-k) \tag{3.7}$$

L'errore di predizione è definito come:

$$e(n) = s(n) - \tilde{s}(n) \tag{3.8}$$

Il problema principale dell'analisi con predizione lineare è quello di determinare un set di coefficienti direttamente dal segnale vocale per ottenere una buona stima delle proprietà spettrali del segnale attraverso l'uso dell'equazione:

$$H(z) = \frac{G}{A(z)} \tag{3.9}$$

dove con $A(z)$ si ha:

$$A(z) = 1 - \sum_{k=1}^{p} \alpha z^{-k} \tag{3.10}$$

A causa della natura tempo variante del segnale vocale la stima dei coefficienti deve essere effetuata su piccoli segmenti del segnale. Il principale approccio si basa sul trovare un set di coefficienti che minimizzino l'errore quadratico medio su un piccolo segmento dell'onda prodotta dal segnale vocale. Osserviamo che nel caso in cui a_k è uguale ad α_k allora $e(n) = Gu(n)$ cioè un treno di impulsi, quindi e(n) nel

caso di suoni sonori sarà piccolo la maggior parte delle volte. Inoltre l'utilizzo dell'errore quadratico medio per stimare i parametri del modello è giustificato dal fatto che questo approccio porta ad un set di equazioni lineari che possono essere efficientemente risolte per ottenere i parametri di predizione. Per stimare i coefficienti α_k, consideriamo delle selezioni $s_n(m)$ del segnale e minimizziamo l'errore quadratico medio dato da :

$$\begin{aligned} E_n &= \sum_m e_n^2(m) \\ &= \sum_m (s_n(m) - \tilde{s}_n(m))^2 \\ &= \sum_m (s_n(m) - \sum_{k=1}^{p} \alpha_k s_n(m-k))^2 \end{aligned} \tag{3.11}$$

dove $s_n(m) = s(m+n)$ cioè una selezione del parlato estratta nelle vicinanze del campione n. Imponendo:

$$\frac{\partial E_n}{\partial \alpha_i} = 0 \tag{3.12}$$

per $i = 1, 2, ..., p$ si ottiene:

$$\sum_m s_n(m-i)s_n(m) = \sum_{k=1}^{p} \hat{\alpha}_k \sum_m s_n(m-i)s_n(m-k) \tag{3.13}$$

per $1 \leq i \leq p$ dove $\hat{\alpha}_k$ sono i valori di α_k che minimizzano E_n. Definendo:

$$\phi_n(i,k) = \sum_m s_n(m-i)s_n(m-k) \tag{3.14}$$

possiamo esprimenere la 3.13 in maniera più compatta:

$$\sum_{k=1}^{p} \alpha_k \phi_n(i,k) = \phi_n(i,0) \qquad peri = 1, 2, ...p. \tag{3.15}$$

Si ottiene dunque un sistema di p equazioni in p incognite che risolto restituisce i valori cercati di α_k.

Al fine di implementare un buon sistema di predizione lineare è necessario risolvere le equazioni in una maniera efficiente. È perciò possibile utilizzare le particolari proprietà delle matrici dei coefficienti. Esistono vari metodi implementativi tra cui : il *metodo della covarianza*, il *metodo dell'autocorrelazione*, il *lattice method*, il *metodo del filtro inverso*, il *metodo della stima spettrale*, il *metodo della massima verosimiglianza*, il *metodo del prodotto interno*. Questi metodi si differenziano per efficienza computazionale e stabilità fisica e numerica. Nel caso del riconoscimento vocale il metodo più utilizzato è quello dell'autocorrelazione proprio in virtù della sua efficienza computazionale e della sua stabilità inerente.

Per quanto riguarda il guadagno G dell'equazione 3.9, esso può essere determinato uguagliando l'energia del segnale con l'energia dei campioni di predizione lineare. Tutto ciò in realtà è vero quando sono verificate determinate assunzioni riguardo il segnale di eccitazione per il sistema di predizione lineare.

È possibile mettere in relazione il guadagno costante G con il segnale di eccitazione e con l'errore di predizione. Il segnale di eccitazione $Gu(n)$ può essere espresso come:

$$Gu(n) = s(n) - \sum_{k=1}^{p} a_k s(n-k). \tag{3.16}$$

Dove l'errore di predizione lineare $e(n)$ è espresso da:

$$e(n) = s(n) - \sum_{k=1}^{p} \alpha_k s(n-k). \tag{3.17}$$

Nel caso in cui $a_k = \alpha_k$ i coefficienti reali di predizione e quelli del modello sono identici quindi:

$$e(n) = Gu(n) \tag{3.18}$$

cioè il segnale di ingresso è proporzionale al segnale di errore con costante di proporzionalità pari al guadagno G.
Essendo l'equazione solo approssimata(cioè valida solo nel caso in cui i reali coefficienti di predizione e quelli ideali sono identici) non è in generale possibile risolvere G in una maniera attendibile dal segnale di errore stesso.
Invece l'assunzione più ragionevole è che l'energia del segnale di ingresso sia uguale all'energia dell'input cioè:

$$G^2 \sum_{m=0}^{N-1} u^2(m) = \sum_{m=0}^{N-1} e^2(m) = E_n \tag{3.19}$$

A questo punto dobbiamo fare delle assunzioni su $u(n)$ per essere in grado di relazionare G alle quantità note. Esistono due casi di interesse per quanto riguarda l'eccitazione. Per i suoni sonori è ragionevole assumere che $u(n) = \delta(n)$ cioè che l'eccitazione sia un campione unitario per $n = 0$. Affinché questa assunzione sia valida è necessario che gli effetti della forma degli impulsi glottali utilizzati nell'eccitazione reale per i suoni sonori siano sommati alla funzione di trasferimento del tratto vocale, perciò entrambi questi effetti sono essenzialmente modellati da un predittore lineare tempo variante. Questo richiede che l'ordine del predittore, p, sia abbastanza elevato da tenere in conto sia gli effetti del tratto vocale che dell'impulso glottale. Per i suoni sordi è più ragionevole assumere che u(n) sia un processo di rumore bianco stazionario, a valor medio nullo e varianza unitaria.
Da tali assunzioni possiamo ora determinare la costante G utilizzando

la precedente equazione. Per i suoni sonori abbiamo come input $G\delta(n)$. Se chiamiamo l'uscita risultante per questo input particolare $h(n)$, troviamo la seguente relazione:

$$h(n) = \sum_{k=1}^{p} \alpha_k h(n-k) + G\delta(n) \tag{3.20}$$

È immediatamente mostrato come la funzione di autocorrelazione della funzione $h(n)$, definita come:

$$\tilde{R}(m) = \sum_{n=0}^{\inf} h(n)h(m+n) \tag{3.21}$$

soddisfi la relazione:

$$\tilde{R}(m) = \sum_{k=1}^{p} \alpha_k \tilde{R}(|m-k|) \qquad\qquad m = 1, 2, ..., p \tag{3.22}$$

e

$$\tilde{R}(0) = \sum_{k=1}^{p} \alpha_k \tilde{R}(k) + G^2 \tag{3.23}$$

da cui segue che:

$$\tilde{R}(m) = R_n(m) \qquad 1 \leq m \leq p. \tag{3.24}$$

Poiché le energie totali nel segnale ($R(0)$) e la risposta impulsiva ($\tilde{R}(0)$) devono essere uguali si ottiene:

$$G^2 = R_n(0) - \sum_{k=1}^{p} \alpha_k \tilde{R}(k) = E_n \tag{3.25}$$

È interessante notare come dall'equazione 3.22 e dalla necessità che l'energia della risposta impulsiva sia uguale all'energia del segnale, si ha che i primi $p+1$ coefficienti della funzione di autocorrelazione della

risposta impulsiva del modello devono essere uguali ai primi $p + 1$ coefficienti della funzione di autocorrelazione del segnale vocale.

Nel caso di suono sordo, le correlazioni sono definite come medie statistiche. Viene considerato che l'input è rumore bianco con media pari a zero e varianza unitaria:

$$E[u(n)u(n-m)] = \delta(m). \tag{3.26}$$

Se eccitiamo il sistema con un input random $Gu(n)$ e chiamiamo l'output $g(n)$ con:

$$g(n) = \sum_{k=1}^{p} \alpha_k g(n-k) + Gu(n) \tag{3.27}$$

Denotando con $\tilde{R}(m)$ la funzione di autocorrelazione di $g(n)$, allora:

$$\begin{aligned} \tilde{R}(m) &= E\{[g(n)g(n-m)]\} \\ &= \sum_{k=1}^{p} \alpha_k E\{[g(n-k)g(n-m)]\} + E\{[Gu(n)g(n-m)]\} \end{aligned} \tag{3.28}$$

poiché $E\{[u(n)g(n-m)]\} = 0$ per $m > 0$ perché $u(n)$ è incorrelata con il precedente segnale di u(n). Per m=0 abbiamo:

$$\begin{aligned} \tilde{R}(0) &= \sum_{k=1}^{p} \alpha_k \tilde{R}(k) + GE[u(n)g(n)] \\ &= \sum_{k=1}^{p} \alpha_k \tilde{R}(k) + G^2. \end{aligned} \tag{3.29}$$

poiché $E\{[u(n)g(n)]\} = E\{[u(n)(Gu(n) + \text{termini che precedono n})]\} =$

G. poiché l'energia nella risposta a Gu(n) deve essere uguale all'energia del segnale, abbiamo che:

$$G^2 = R_n(0) - \sum_{k=1}^{p} \alpha_k R_n(k). \tag{3.30}$$

come nel caso dell'impulso di eccitazione per il parlato.

È inoltre importante effettuare delle considerazioni sulla scelta del numero dei parametri predittori, p, e sulla scelta della lunghezza del frame N. La scelta di p dipende essenzialmente dalla frequenza di campionamento e non dipende dal metodo LPC che è stato utilizzato. Generalmente lo spettro del parlato può essere rappresentato con una densità media di 2 poli per KHz dovuti al contributo del tratto vocale, dunque in totale sono necessari Fs poli per rappresentare il contributo dello spettro vocale dove Fs è la frequenza di campionamento in KHz. Inoltre per rappresentare adeguatamente lo spettro della sorgente di eccitazione e il carico di radiazione sono necessari altri 3 o 4 poli. L'errore di predizione decresce al crescere di p, con p dell'ordine di 13-14 l'errore viene essenzialmente appiattito, solamente per uteriori incrementi di p ci saranno piccoli decrementi.

È interessante notare oltretutto che la radice quadratica media normalizzata della predizione di errore nel caso di rumore di fondo è significativamente più elevata che per il linguaggio parlato. La scelta della lunghezza N del frame è molto importante per l'implementazione di sistemi di analisi LPC. Ovviamente è vantaggioso scegliere N il più piccolo possibile poiché il livello di carico computazionale per tutti i metodi di risoluzione è proporzionale ad N. Nel caso del metodo dell'autocorrelazione N deve essere dell'ordine di alcuni periodi di picco per assicurare risultati affidabili.

Nel metodo dell'autocorrelazione viene utilizzata una finestra per pe-

sare il parlato dunque la durata del frame deve essere sufficientemente lunga in modo che gli effetti di smussamento della finestra non influenzino seriamente i risultati.
Perciò nel caso di un rate di 10 KHz sono state utilizzate analisi di durata di circa 100- 400 campioni per l'implementazione dell'LPC tramite il metodo dell'autocorrelazione con la maggior parte dei sistemi tendenti al valore massimo di campioni.

3.4 Coefficienti Cepstrali

È stato dimostrato che il parlato può essere rappresentato come l'uscita di un sistema lineare tempo variante le cui proprietà variano lentamente nel tempo. Questo ha portato al principio base dell'analisi vocale che asserisce che se si considerano piccoli segmenti del segnale vocale, ogni segmento può effettivamente essere modellato come generato eccitando un sistema tempo invariante da un treno di impulsi quasi teorico o un segnale random di rumore. Il problema dell'analisi del parlato risiede nello stimare i parametri del modello vocale e misurare la loro variazione con il tempo. L'eccitazione e la risposta impulsiva di un sistema lineare tempo invariante sono combinati attraverso una convoluzione, perciò il problema dell'analisi del parlato può essere visto come il problema di dover separare le componenti della convoluzione. Questo problema è spesso chiamato deconvoluzione. È necessario dunque introdurre il concetto del filtraggio omomorfo e dunque la teoria dei sistemi omomorfi per la convoluzione [11]. I sistemi omomorfi per la convoluzione obbediscono al principio generalizzato di sovrapposizione degli effetti che afferma che se un segnale di ingresso è composto dalla combinazione lineare di segnali elementari allora l'uscita è una combinazione lineare delle corrispondenti uscite. Un ri-

sultato immediato del principio di sovrapposizione degli effetti risiede nel fatto che l'uscita di un sistema lineare tempo invariante può essere espresso come una somma di convoluzione:

$$y(n) = \sum_{k=-\infty}^{\infty} h(n-k)x(k) = h(n) * x(n) \tag{3.31}$$

Per analogia con il principio di sovrapposizione per i sistemi lineari convenzionali si può definire una classe di sistemi che obbediscono al principio di sovrapposizione generalizzato dove l'addizione è sostituita con la convoluzione.

$$\begin{aligned} H[x(n)] &= H[x_1(n) * x_2(n)] \\ &= H[x_1(n)] * H[x_2(n)] \\ &= y_1(n) * y_2(n) \\ &= y(n) \end{aligned} \tag{3.32}$$

Questo tipo di sistemi sono nominati sistemi omomorfi per la convoluzione. Un filtro omomorfo è semplicemente un sistema omomorfo con la proprietà che un componente (quello desiderato) passa attraverso il sistema essenzialmente inalterato mentre il componente indesiderato è rimosso. Un importante aspetto della teoria dei sistemi omomorfi è che ogni sistema omomorfo può essere rappresentato come una cascata di tre sistemi omomorfi. Il primo sistema prende gli ingressi combinati tramite la convoluzione e li trasforma in una combinazione additiva delle corrispondenti uscite. Il secondo sistema è un sistema lineare convenzionale che obbedisce al principio di sovrapposizione degli effetti. Il terzo sistema è l'inverso del primo cioè trasforma i segnali legati da una combinazione additiva in segnali combinati tramite la convoluzione. L'importanza dell'esistenza di queste forme canoniche risiede nel fatto che la progettazione di questi sistemi si riduce al problema

di progettare un sistema lineare. Il sistema $D_*[]$ è chiamato sistema caratteristico per la deconvoluzione omomorfa ede è così espresso:

$$\begin{aligned} D_*[x(n)] &= D_*[x_1(n) * x_2(n)] \\ &= D_*[x_1(n)] + D_*[x_2(n)] \\ &= \hat{x}_1(n) + \hat{x}_2 \\ &= \hat{x}(n) \end{aligned} \tag{3.33}$$

In maniera analoga il sistema caratteristico inverso D_*^{-1} è definito da:

$$\begin{aligned} D_*^{-1}[\hat{y}(n)] &= D_*^- 1[\hat{y}_1(n) + \hat{y}_2(n)] \\ &= D_*^{-1}[\hat{y}_1(n)] * D_*^{-1}[\hat{y}_2(n)] \\ &= y_1(n) * y_2(n) \\ &= y(n) \end{aligned} \tag{3.34}$$

La rappresentazione matematica del sistema caratteristico dipende dal fatto che se l'input è una convoluzione:

$$x(n) = x_1(n) * x_2(n) \tag{3.35}$$

allora la trasformata z dell'input è il prodotto della corrispondente trasformata z:

$$X(z) = X_1(z)X_2(z) \tag{3.36}$$

.

Chiaramente la trasformata z dell'uscita del sistema caratteristico deve essere una combinazione di somme delle trasformate z. Il comportamento nel dominio della frequenza del sistema caratteristico per la convoluzione deve avere la proprietà che se il segnale è rappresentato da un prodotto di trasformate z degli input, l'uscita deve essere

una somma delle corrispondenti z-trasformate delle uscite. Questo approccio è basato sul fatto che il logaritmo di un prodotto può essere espresso come la somma dei logaritmi di termini individuali.

$$\begin{aligned} \hat{X}(z) &= log[X(z)] \\ &= log[X_1(z)X_2(z)] \\ &= log[X_1(z)] + log[X_2(z)] \end{aligned} \tag{3.37}$$

La trasformata z comunque è generalmente una quantità complessa. Una appropriata definizione del logaritmo complesso è:

$$\hat{X}(e^{j\omega}) = \log|X(e^{j\omega})| + j\arg[X(e^{j\omega})] \tag{3.38}$$

In questa equazione la parte reale non causa particolari difficoltà mentre si hanno problemi per quanto riguarda l'unicità nel definire la parte immaginaria che rappresenta semplicemente la fase dell'angolo della z-trasformata valutata sul cerchio unitario. Un approccio per affrontare il problema dell'unicità della fase dell'angolo è di richiedere che essa sia una funzione pari continua di ω. Date queste considerazioni la trasformata inversa del logaritmo complesso della trasformata di Fourier dell'ingresso è l'uscita del sistema caratteristico per la convoluzione:

$$\hat{x}(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \hat{X}(e^{j\omega})e^{j\omega n}d\omega \tag{3.39}$$

L'uscita del sistema caratteristico, $\hat{x}(n)$ è chiamata cepstrale complesso. Possiamo utilizzare il termine cepstrale per la quantità:

$$c(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \log|X(e^{j\omega})|e^{j\omega n}d\omega \tag{3.40}$$

È stato dunque definito il sistema caratteristico per la convoluzione omomorfa e perciò è stata definita la forma canonica per tutti i sistemi

omomorfi per la convoluzione. Tutti i sistemi di questa classe differiscono solamente nella parte lineare del sistema. La scelta della parte lineare dipende necessariamente dalle proprietà dei segnali di ingresso. Fondamentalmente le tecniche cepstrali sono state adattate all'analisi di dati che contengono echi o riverberi di una piccola onda fondamentale la cui forma deve essere conosciuta a priori. Consideriamo ad esempio un segnale con un semplice eco:

$$x(t) = s(t) + \alpha s(t - \tau). \tag{3.41}$$

La densità spettrale di un tale segnale è data da:

$$|X(f)|^2 = |S(f)|^2[1 + \alpha^2 + 2\alpha cos(2\pi f\tau)]. \tag{3.42}$$

Possiamo vedere dalla precedente equazione che la densità spettrale del segnale con un eco ha la forma di un inviluppo che modula una funzione periodica della frequenza (il contributo spettrale dell'eco). Prendendo il logaritmo dello spettro, questo prodotto viene convertito in una somma di due componenti:

$$C(f) = log|X(f)|^2 = log|S(f)|^2 + log[1 + \alpha^2 + 2\alpha cos(2\pi f\tau)]. \tag{3.43}$$

Dove $C(f)$ può essere vista come una forma d'onda che ha una componente additiva periodica la cui frequenza fondamentale è il ritardo dell'eco τ.

Nelle analisi convenzionali delle forme d'onda al variare del tempo , queste componenti periodiche si mostrano come linee o picchi nel corrispondente spettro di Fourier.

Perciò lo spettro del logaritmo dello spettro mostrerà dei picchi quando la forma d'onda nel tempo originale contiene un eco. Questa rappresentazione non riguarda nè il tempo nè le frequenze. Il dominio in cui ci troveremo viene chiamato da Borget come quefrency.

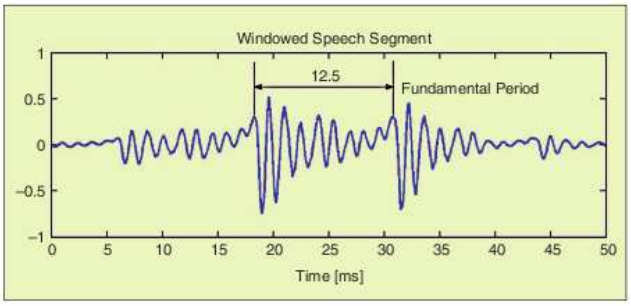

Figura 3.3: Segmento di segnale sonoro (una vocale) della duarata da 50 ms a cui è stata applicata una finestra di Hamming. Si può notare il periodo di picco di circa 12.5 ms

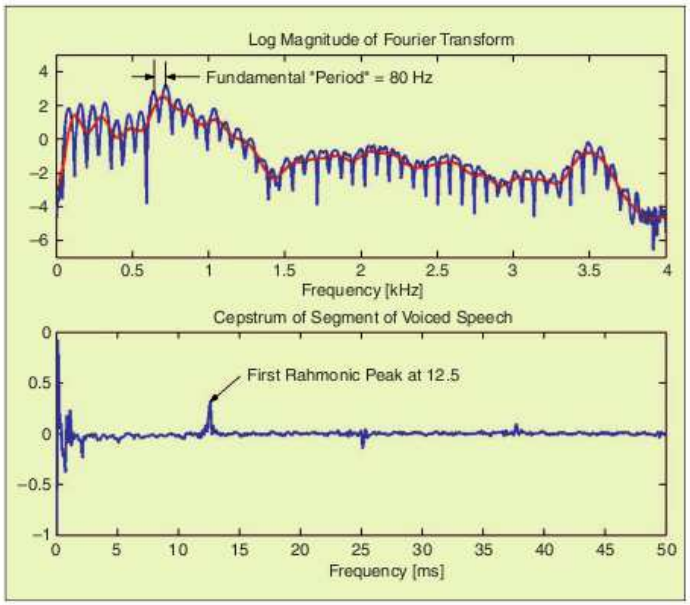

Figura 3.4: Logaritmo dello spettro e Cepstrali di un segmento di segnale sonoro (una vocale)

La potenza cepstrale è di solito utilizzata per determinare gli istanti di arrivo dell'onda fondamentale, i suoi echi e le loro relative: ampiezze; l'analisi dei cepstrali complessi serve a determinare la forma d'onda. I cepstrali complessi possono essere visti come la deconvoluzione di due segnali , una forma d'onda fondamentale e un treno di impulsi. Storicamente i cepstrali trovano le loro radici nel problema generale della deconvoluzione di due o più segnali. La potenza cepstrale fu dapprima descritta da Borget nel 1963 come una tecnica euristica per trovare gli istanti di arrivo degli eco in un segnale composto. Questo autore definì la potenza cepstrale di una funzione come lo spettro di potenza del logaritmo dello spettro di potenza del segnale. In pratica la potenza cepstrale ha effetto se le creste e il treno di impulsi la cui convoluzione compone i dati, occupano range di frequenza differenti. In verità la potenza cepstrale non esiste per la maggior parte dei segnali, è significativa solo nel caso di dati campionati. Consideriamo, dunque, la seguente definizione: la potenza cepstrale di una sequenza di dati è la radice della z-trasformata inversa della sequenza di dati. Poiché i cepstrali complessi contengono le informazioni sulla fase dei dati composti, possono essere utilizzati non solo per la rivelazione dell'eco ma anche per il riconoscimento delle increspature. Questo processo è anche conosciuto come deconvoluzione omomorfa o filtri omomorfi. Come accennato precedentemente il processo di produzione della voce consiste essenzialmente di un sistema lineare lentamente tempo variante, eccitato o da un treno di impulsi quasi periodico o da un rumore casuale. Pertanto è appropriato pensare ad un breve segmento di parlato come il risultato di una eccitazione di un sistema lineare tempo invariante da parte di un treno di impulsi periodici. Similmente un breve segmento di registrazione sorda può essere pensato come un sistema lineare tempo invariante eccitato da un rumore ca-

suale. Un breve segmento di segnale sonoro può quindi essere pensato come un segmento della forma d'onda:

$$s(n) = p(n) * g(n) * v(n) * r(n) = p(n) * h_v(n) = \sum_{r=-\inf}^{\inf} h_v(n - rN_p) \tag{3.44}$$

dove $p(n)$ è un treno di impulsi periodici con N_p campioni per periodo e $h_v(n)$ è la risposta impulsiva di un sistema lineare che combina gli effetti della forma dell'onda glottale $g(n)$ e della risposta impulsiva del tratto vocale $v(n)$ e la risposta impulsiva della radiazione $r(n)$. Alla stessa maniera un breve segmento di segnale sordo può essere visto come un segmento della forma d'onda:

$$s(n) = u(n) * v(n) * r(n) = u(n) * h_u(n) \tag{3.45}$$

dove $u(n)$ è un'eccitazione di un rumore casuale e $h_u(n)$ è la risposta impulsiva di un sistema che rappresenta gli effetti combinati del tratto vocale e della radiazione. Nel caso del segnale sonoro, la funzione di trasferimento del sistema lineare può essere espressa nella forma:

$$H_v(z) = G(z)V(z)R(z) \tag{3.46}$$

per il segnale sordo:

$$H_u(z) = V(z)R(z) \tag{3.47}$$

Nel caso di suoni non nasali il tratto vocale è modellizzato come un filtro di tutti poli su piccoli intervalli di tempo. La sorgente glottale è modellizzata come zeri nel dominio z ancora su piccoli intervalli temporali. Il segnale vocale è perciò modellizzato come la convoluzione di un treno di impulsi, della risposta impulsiva glottale e della risposta

impulsiva del tratto vocale. Questi tre segnali devono essere deconvoluti sfruttando le proprietà dei sistemi omomorfi precedentemente descritti.

Come già accennato sappiamo che i coefficienti della predizione lineare sono assunti come i coefficienti del denominatore della funzione di sistema che rappresenta il modello che combina gli effetti della risposta del tratto vocale, della forma d'onda glottale e della radiazione. Perciò dati i coefficienti di predizione lineare possiamo facilmente trovare la risposta in frequenza del modello per la produzione della voce. Valutando $H(z)$ per $z = e^{j\omega}$ abbiamo:

$$H(e^{j\omega}) = \frac{G}{1 - \sum_{k=1}^{p} a_k e^{-j\omega k}} = \frac{G}{A(e^{j\omega})} \tag{3.48}$$

Nel caso in cui il filtro di predizione lineare sia stabile (cosa garantita per l'analisi tramite autocorrelazione) è possibile derivare i coefficienti cepstrali effettuando delle operazioni ricorsive. poiché essi sono in effetti la trasformata di Fourier inversa della risposta impulsiva del modello di predizione lineare e quest'ultimo è un filtro con risposta impulsiva infinita , possiamo, in teoria, calcolare un numero infinito di coefficienti cepstrali. In realtà il numero di coefficienti cepstrali calcolati è solitamente comparabile al numero di coefficienti di predizione lineare.

3.5 Trasformata Affine

Introduciamo preliminarmente il concetto di Cepstral Mean Subtraction (CMS) [12], tecnica veloce ed efficiente, al fine di discutere il ruolo della trasformazione affine. Sappiamo che l'acquisizione di un segnale

vocale subisce distorsioni lineari dovute all'effetto di filtro del canale. Ciò si può semplicemente esprimere come:

$$T(z) = S(z)G(z) \tag{3.49}$$

dove $S(z)$ corrisponde al segnale originale pulito, $G(z)$ al canale e $T(z)$ al segnale vocale filtrato. Nel dominio logaritmico si ha:

$$logT(z) = logS(z) + logG(z) \tag{3.50}$$

Assumendo che lo spettro della parola pronunciata e del canale siano ben approssimati dal modello di predizione lineare costituito da soli poli, si può osservare che l'influenza del canale è costituita da una componente additiva ai coefficienti cepstrali del segnale privo di alterazioni $S(z)$. Assumendo inoltre che la media dei coefficienti cepstrali privi di disturbi sia zero, la stima dei cepstrali del canale è data solamente dalla media dei cepstrali del segnale vocale filtrato $T(z)$. Per compensare l'effetto del canale, la stima della media effettuata su di esso viene rimossa tramite la CMS. Nel riconoscimento vocale, la trasformazione affine, che si basa appunto sulla tecnica CMS, ha lo scopo di eliminare i mismatch dati dalle differenti condizioni di acquisizione del segnale in fase di training e in fase di testing [13]. I mismatch possono essere rappresentati come una trasformazione lineare nel dominio cepstrale:

$$y = Ax + b \tag{3.51}$$

dove x è il vettore dei coefficienti cepstrali relativi ad un frame della parola pronunciata in fase di test; A e b sono rispettivamente la matrice e il vettore che dobbiamo stimare per ogni enunciazione ed y è il vettore trasformato. Geometricamente, b rappresenta una traslazione e A rappresenta sia una scalatura che una rotazione. Quando A è una

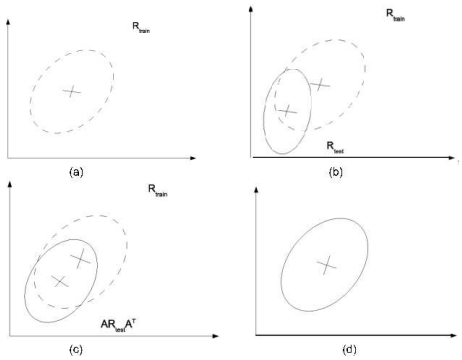

Figura 3.5: Visione geometrica della trasformazione affine

matrice diagonale, essa rappresenta una operazione di sola scalatura. La tecnica CMS fornisce una stima del vettore b ed assume che A sia una matrice identità.

Dalla figura 3.5 possiamo ottenere un'interpretazione geometrica della trasformazione affine. Nella fig. 3.5 (a) la linea tratteggiata è il contorno dei dati di training mentre in fig. 3.5(b) abbiamo il contorno dei dati di testing. A causa di differenti canali e livelli di rumore la media dei dati di test è traslata rispetto a quella dei dati di training e la distribuzione è ristretta e ruotata. Il mismatch potrebbe dunque causare un errore decisionale. Di seguito verrà presentato un algoritmo che in primo luogo trova dai dati di training la matrice di covarianza, R_{train}, la quale caratterizza l'intera distribuzione e in seguito la matrice di covarianza, R_{test}, dai dati di test e stima i parametri della matrice A necessaria per la trasformazione lineare. Dopo aver applicato la prima trasformazione, l'intera distribuzione dei dati di test è scalata e ruotata, $AR_{test}A^T$, per essere uguale ai dati di training ad eccezione della media, come si può notare in figura 3.5(c). In un passo successivo viene trovata la differenza fra le medie e i dati di test vengono traslati affinché assumano la stessa locazione di quelli di training, come mostrato in figura 3.5(d) dove risulta una completa sovrapposizione dei contor-

ni dei due insiemi. Nel caso di speech recognition vengono registrate tutte le parole appartenenti al vocabolario considerato, pronunciate da diversi parlatori e vengono utilizzati la matrice di covarianza R_{train} e il vettore della media m_{train} per rappresentare l'intera distribuzione dei dati di training per tutte le enunciazioni del training nel dominio cepstrale. La matrice di covarianza e il vettore delle medie sono così rappresentati:

$$R_{train} = \frac{1}{U}\sum_{i=1}^{U}\frac{1}{N_i}\sum_{j=1}^{N_i}(x_{i,j} - m_i)(x_{i,j} - m_i)^T \tag{3.52}$$

$$m_{train} = \frac{1}{U}\sum_{i=1} U m_i \tag{3.53}$$

dove $x_{i,j}$ è il j-esimo frame diverso dal silenzio dell'i-esima parola del training, U è il numero totale di parole del training, N_i e m_i sono rispettivamente il numero totale di frame diversi dal silenzio e il vettore media dell'i-esima parola di training. Nella fase di testing, verrà registrata un'unica parola alla volta e verificata. La matrice di covarianza per i dati di test è:

$$R_{test} = \frac{1}{N_f}\sum_{j=1}^{N_f}(y_j - m_{test})(y_j - m_{test})^T \tag{3.54}$$

dove y_J e m_{test} sono rispetivamente un frame diverso dal silenzio e il vettore media dei dati di test e N_f è il numero totale di frames diversi dal silenzio. Il criterio proposto per la stima dei parametri è quello di rendere R_{test} uguale ad R_{train} attraverso una rotazione, una scalatura e una traslazione dei dati di test. Per quanto riguarda la rotazione e la scalatura si fa riferimento alla seguente equazione:

$$R_{train} - AR_{test}A^T = 0 \tag{3.55}$$

dove A è stata definita in 3.51; R_{train} e R_{test} sono definite in 3.52 e 3.54. Risolvendo l'equazione precedente 3.55, abbiamo per la matrice A la forma:

$$A = R_{train}^{1/2} R_{test}^{-1/2} \tag{3.56}$$

dunque il termine di traslazione b si ottiene da:

$$b = m_{train} - m_{rs} = m_{train} - \frac{1}{N_f} \sum_{j=1}^{N_f} A x_j \tag{3.57}$$

dove m_{rs} è il vettore media dei frames ruotati e scalati; N_f è il numero totale dei frames diversi dal silenzio dell'enuncizione del test; x_j è il j-esimo vettore cepstrale del frame. Per verificare un comando pronunciato in fase di test confrontandolo con i modelli costruiti in fase di training, dapprima vengono calcolati R_{test}, A, b e in seguito viene applicata la trasformazione 3.51 per ridurre il mismatch.

Capitolo 4

Endpoint Detection

4.1 Introduzione

Come specificato in precedenza, il nostro sistema di riconoscimento è di tipo isolated-word perché rivolto ad applicazioni di comando e controllo. Questo tipo di riconoscimento si basa sul fatto che il segnale risulta costituito da una singola parola preceduta e seguita da silenzio o rumore di fondo. Si assume, dunque, che una volta pronunciata la parola, si possano separare i segmenti relativi al parlato dai restanti. Il processo di separazione della parola pronunciata dal rumore di fondo è denominato *endpoint detection*. L'accuratezza della scelta degli endpoints è importante per due ragioni:

1. Un riconoscimento attendibile della parola dipende in maniera critica dall'accuratezza della fase di endpoint detection.

2. Il peso a livello computazionale nell'elaborazione del parlato è minima nel caso di endponts determinati in modo accurato.

I problemi riguardanti l'endpoint detection crescono nel caso di transitori associati allo speaker e/o ai sistemi di trasmissione. Questo tipo di rumori di fondo complicano notevolmente il problema dell'endpoint detection. Per esempio spesso l'inizio e la fine di una parola sono oscurati

da rumori generati dal parlatore come ad esempio il respiro pesante. In molte applicazioni, come ad esempio quella presa in considerazione in questa tesi, il problema è inoltre complicato da un ambiente non stazionario dove possono presentarsi conversazioni concorrenti e rumori dovuti a movimenti che avvengono nell'ambiente circostante. Per minimizzare gli effetti dovuti all'ambiente mutevole, abbiamo utilizzato un microfono a cancellazione di rumore per registrare il segnale vocale. L'approccio dell'endpoint detection assume che la parola desiderata si presenti in un determinato intervallo di tempo. Esistono vari approcci per trovare gli endpoints delle parole: in maniera esplicita, implicita o ibrida [14]. Nell'approccio esplicito l'endpoint detection precede ed è indipendente dagli stadi relativi al riconoscimento e alla decisione. Gli endpoints della parola pronunciata vengono stimati da misure effettuate sull'input vocale e inviati allo stadio successivo del sistema. In un approccio puramente implicito gli endpoints sono determinati esclusivamente nelle fasi di riconoscimento e decisione del sistema. Un metodo implicito effettuerà il riconoscimento utilizzando tutti i possibili sets di endpoints. La tecnica ibrida incorpora idee provenienti da entrambi i precedenti metodi. In maniera simile al metodo esplicito, in quello ibrido una o più stime di ogni endpoint vengono ottenute dalle caratteristiche misurate dall'input pronunciato.

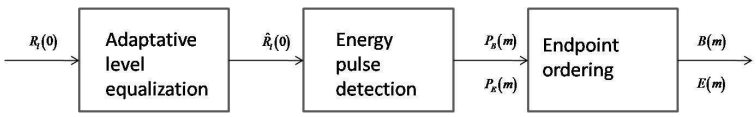

Figura 4.1: Schema a blocchi di un endpoint detector di tipo ibrido.

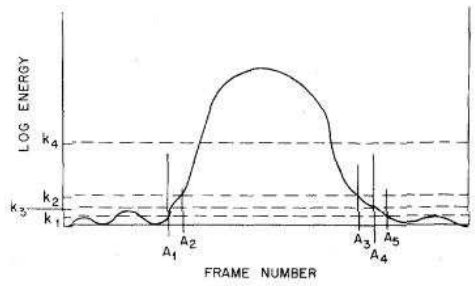

Figura 4.2: Esempio dell'uso di un sistema a soglie per determinare i limiti di un impulso.

4.2 Rilevazione degli impulsi

In questo lavoro di tesi è stata utilizzata una tecnica di tipo ibrido. Il primo stadio dell'endpoint detector ibrido, come è mostrato in figura 4.1, è l'equalizzatore adattativo di livello che normalizza l'array di energia al livello del rumore di fondo. L'array di energia possiede la proprietà seguente: durante il silenzio esso si aggira intorno al livello di 0 dB e durante il parlato possiede valori molto più elevati. Perciò le soglie del valore assoluto dell'energia possono essere definite per la rivelazione della presenza di segnali simili al parlato.

Basandosi sull'uscita dell'equalizzatore adattativo di livello $R_l(0)$, viene definito un set di quattro soglie k_1, k_2, k_3, k_4. Lo scopo di queste soglie è di definire la presenza di un impulso di energia (figura 4.2). Ci si basa sull'assunzione che la parola pronunciata contenga uno o più impulsi di energia e che l'unico problema sta nel trovare questi impulsi e nel determinare quali di questi appartengano alla parola.

Il secondo blocco dell' endpoint detector è il rivelatore d'impulsi(figura 4.1). La ricerca degli impulsi energetici procede da destra a sinistra. Vengono analizzati i valori di uscita dell'equalizzatore di livello $R_l(0)$ al variare di l e quando un valore supera la prima soglia k_1, viene

registrato il numero del frame corrispondente A_1. Se $R_l(0)$ supera la soglia più alta k_2 prima di ricadere sotto la soglia k_1 viene rilevato l'inizio dell'impulso di energia. Il punto di inizio è nominalmente A_1 a meno che il tempo di salita (da A_1 ad A_2) sia troppo lungo, in tal caso si sceglie A_2. Il frame che indica la fine dell'impulso energetico viene ricercato in modo simile utilizzando le soglie k_2 e k_3. Anche in questo caso se la durata fra A_3 e A_4 è troppo lunga viene utilizzato il frame A_3 come frame di fine dell'impulso energetico.
Possono essere effettuati altri due test per ogni impulso di energia rilevato. Viene misurato il picco di energia dell'impulso e se questo cade al di sotto della soglia k_4, l'impulso di energia viene scartato e non considerato come parte della parola. Inoltre viene misurata la durata dell'impulso che non deve essere minore di cinque frames.
In conclusione, all'uscita del rivelatore di impulsi energetici avremo una serie di punti di inizio $P_B(m)$ e di punti di fine $P_E(m)$ di ogni impulso energetico $m = 1, 2, ..., M$ con M pari al numero di impulsi rilevati nell'intervallo di registrazione. Se non sono stati trovati impulsi di energia viene chiesta nuovamente la registrazione. Dei controlli vengono effettuati anche sulla posizione degli impulsi di energia, se questi vengono rilevati agli estremi dell'intervallo di registrazione, come nel caso precedente viene richiesta una nuova registrazione.

4.3 Ordinamento degli impulsi

L'ultimo blocco costitutivo dell'endpoint detector è l'ordinatore di endpoint(figura 4.1). Il fine di questo blocco è quello di determinare dei possibili sets di coppie di endpoint di parola a partire dal set degli endpoints degli impulsi. L'ordinamento si basa sulle seguenti assunzioni:

1. Una parola singola i cui endpoints devono essere determinati è composta da uno o più impulsi di energia

2. Il frame che ha la massima energia in termini logaritmici deve appartenere alla parola pronunciata

3. Più è grande l'intervallo di pausa che separa due impulsi di energia, meno è probabile che appartengano ad una parola formata da impulsi multipli.

4. Impulsi di energia separati dall'impulso ad energia massima da un intervallo di pausa maggiore di 150 ms non fanno parte della stessa parola.

Basandosi su queste assunzioni, gli impulsi di energia vengo raggruppati in combinazioni di coppie di endpoints di parola e ordinati.
In particolare l'operazione di ordinamento si suddivide in tre parti principali. Nella prima parte si utlizza la durata NSEP(sec) per determinare gl'impulsi che molto probabilmente faranno parte della parola. Il risultato di questa fase è composto due parametri: LC ed RC, cioè il numero di impulsi a sinistra ed a destra dell'impulso contenente il frame ad energia massima, che con molta probabilità fanno parte della parola. Nella seconda fase vengono determinati la prima serie di endpoints della parola, calcolati facendo riferimento solo agli impulsi tenuti in conto dai parametri RC ed LC. Nella terza ed ultima fase vegono aggiunti altri enpoints, questa volta considerando anche gli impulsi non tenuti in conto da RC e da LC.
Per capire meglio il cocetto, di seguito sarà fatto un piccolo esempio per capire il metodo di ordinamento; l'esempio fa riferimento alla figura 4.3. In tale figura sono mostrati tre impulsi, P_1,P_2 e P_3, che sono distanziati tra di loro da X_1 e X_2 frames. Se X_1 e X_2 sono entrambi

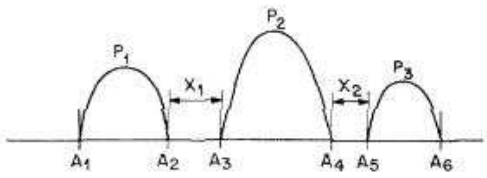

Figura 4.3: Esempio che illustra l'uso dei vincoli temporali usati per discriminare tra impulsi di energia adiacenti

minori di NSEP secondi, la prima coppia di endpoints è scelta come (A_1, A_6), la seconda coppia è scelata come (A_3, A_4) (assumendo che $X_1 > X_2$), la terza è scelta come (A_1, A_4) e la quarta (A_3, A_4). Se $X_1 > 150ms$ e $X_2 < 150ms$, le coppie di endpoints ordinate sono (A_3, A_6),(A_3, A_4) e (A_1, A_4). Infine se sia X_1 che X_2 sono più grandi di $150ms$, le coppie ordinate sono (A_3, A_4),(A_3, A_6) e (A_1, A_4).

Alla fine delle tre fasi otteniamo il risultato finale, cioè un lista di endpoints della parola ordinati in termini di probabilità decrescente.

Capitolo 5

Classificatore Polinomiale

5.1 Introduzione

Il riconoscimento automatico di oggetti (pattern) e la loro descrizione, classificazione e raggruppamento (clustering) sono argomenti importanti in un'ampia varietà di problemi ingegneristici. Il problema del pattern recognition si pone nella forma di classificazione o identificazione delle categorie di appartenenza dell'oggetto considerato. La progettazione di un sistema con capacità di pattern recognition richiede essenzialmente di affrontare i seguenti aspetti:

- Acquisizione e pre-elaborazione (e normalizzazione) dei dati.
- Rappresentazione e classificazione dei dati o pattern.
- Decisione e classificazione.

In genere un sistema di pattern recognition riceve in ingresso la descrizione di un oggetto, ovvero un insieme di misure che lo caratterizzano (features), e sulla base di esse decide a quale classe l'oggetto appartiene con più probabilità, restituendo un vettore discriminante d e l'etichetta $\hat{\mathrm{k}}$ della classe scelta. La decisione della classe di appartenenza ha un costo associato ad un errore di assegnazione di classe. L'obiettivo

è quello di minimizzare il costo di classificazione realizzando un buon sistema di riconoscimento.

5.2 Pattern Classification

In questa sezione verranno introdotti dei termini chiave per lo studio della pattern classification che serviranno a spiegare le nozioni fondamentali di tale processo. Consideriamo un set di possibili eventi:

$$\Omega = \{\omega_1, \quad \omega_2, \quad \dots \quad , \quad \omega_k, \quad \dots \quad , \omega_K\} \tag{5.1}$$

Dove ω_k sono le classi di Ω. Le K classi sono mutuamente esclusive e complete. Introduciamo adesso il concetto di target vector y, cioè un vettore K-dimensionale definito come la k-esima colonna della matrice identità di dimensione K. Nel caso di K=3, abbiamo quindi 3 target vector ognuno associato ad una particolare classe. Nel migliore dei casi essi formano un triangolo equilatero nello spazio tridimensionale come mostrato in figura 5.2. Lo spazio continuo K-dimensionale coperto dai K vettori di target, $y_1, \quad \dots \quad , \quad y_k$, è chiamato spazion decisionale D. L'obiettivo della classificazione è quello di effettuare una scelta sulla base di opportune osservazioni. Le osservazioni possono essere arbitrarie, possono avere origini differenti e sono chiamate misure. L'insieme di tutte le misure che si riferiscono alla stessa entità sono combinate nei relativi vettori:

$$v = [v_1 \quad v_2 \quad \dots \quad v_n]^T \tag{5.2}$$

La dimensione N di v, ovvero la sua cardinalità, è di grande importanza in quanto determina il peso computazionale della pattern calssification. Le misure sono delle semplici osservazioni potenzialmente utili

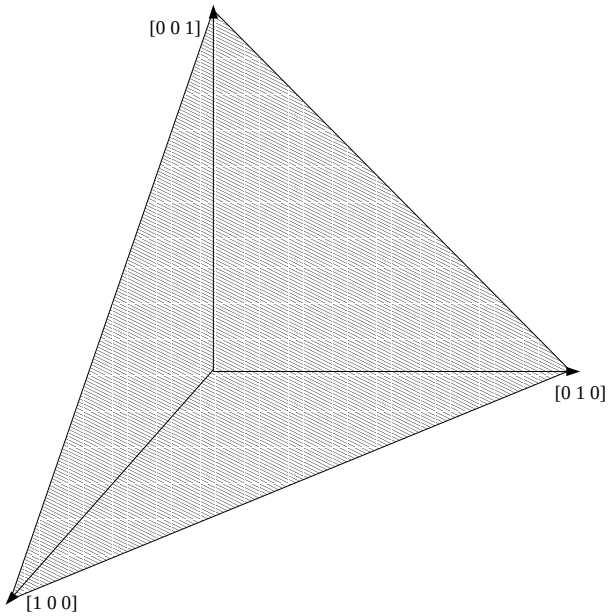

Figura 5.1: Spazio tridimensionale occupato dai tre vettori di target

al riconoscimeto. Le features sono costruite in maniera opportuna a partire dalle misure e quindi hanno maggiore capacità discriminativa. Le due variabili appena definite (v ed y) descrivono come il pattern si presenta e che cosa esso rappresenti. Esse sono i due aspetti di un singolo pattern:

$$Pattern = [v, y] \tag{5.3}$$

La pattern classification ha l'obiettivo di stabilire un mapping tra lo spazio delle misure V e lo spazio decisionale D (fig 5.2).

Questo mapping è implementato in due passi principali:

Approssimazione funzionale mappatura della misura v in un vet-

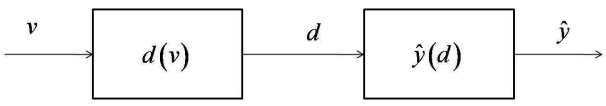

Figura 5.2: Mapping di v in d(v)

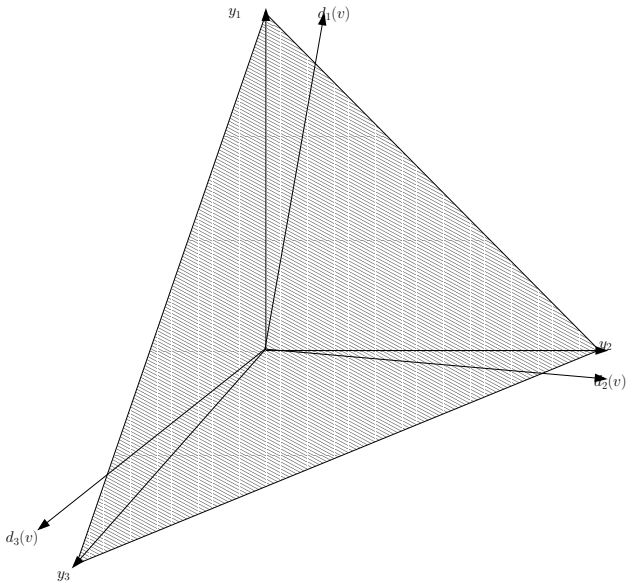

Figura 5.3: Mappatura di v in d effettuata in maniera corretta, i vettori $d(v)$ sono vicini ai vettori target y

tore d appartenente allo spazio decisonale D con lo scopo di rendere d il più vicino possibile al target vector y associato come mostrato in figura 5.3.

Classificazione a minima distanza Dato d si cerca il target vector $\hat{y}$ più vicino a d tra l'insieme dei target vector a disposizione.

In figura 5.4 una mappatura di v in d effettuata in maniera errata porta i vettori appartenenti allo spazio decisionale D ad allontanarsi dai vettori di target y e ad avvicinarsi tra loro creando così possibilità maggiori di errori decisionali.

5.3 Classificazione Polinomiale

5.3.1 Espansione Polinomiale

L'approccio più ovvio per costruire una funzione d(v) che approssimi un set di funzioni di base è quello di appoggiarsi al principio del-

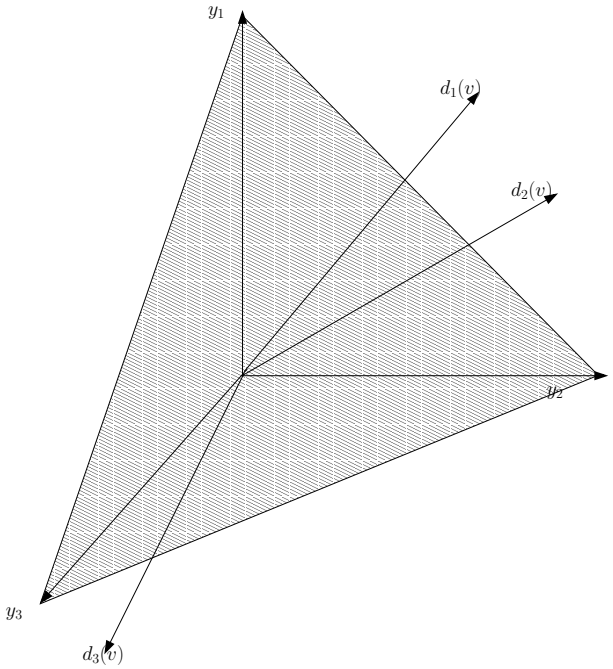

Figura 5.4: Vettori d_k appartenenti allo spazio decisionale D creati in maniera errata

l'approssimazione polinomiale. È noto che un polinomio di lunghezza sufficiente è adeguato a rappresentare ogni ragionevole funzione, basti far riferimento al teorema di Weierstrass e all'espansione in serie di Taylor.

Verrà ora introdotta la tecnica dell'approssimazione polinomiale prima nel caso di una funzione scalare d(v) con un vettore come argomento e verrà in un secondo momento generalizzato al caso di una funzione vettoriale d(v):

$$\begin{aligned} d(v) =& a_0 + a_1 v_1 + a_2 v_2 + \ldots + a_N v_N \\ &+ a_{N+1} v_1^2 + a_{N+2} v_1 v_2 + a_{N+3} v_1 v_3 + \ldots \\ &+ a_{\ldots} v_1^3 + a_{\ldots} v_1^2 v_2 + a_{\ldots} v_1^2 v_3 + \ldots \end{aligned} \tag{5.4}$$

Questo polinomio generale è formato da un termine costante a_0, seguito da N termini lineari $a_n v_n$ nella prima linea seguiti da $N(N+1)/2$ termini quadratici nella seconda linea, seguiti da un più grande nume-

ro di termini cubici nella terza linea e così via, fino all'arbitrario grado G dei termini polinomiali.
Il polinomio completo di grado G del vettore argomento v di dimensione N, possiede L termini polinomiali dove L è dato da:

$$L = \binom{N+G}{G} \tag{5.5}$$

Ogni termine polinomiale costituisce una delle L funzioni di base $f_l(v)$, $l = 1, ..., L$, di una espansione funzionale. Queste funzioni di base sono combinate in forma di somma pesata. I coefficienti a_l sono i pesi della combinazione lineare.
Il polinomio generale mostrato nell'equazione 5.4 può essere espresso in una forma più compatta introducendo un mapping v $\rightarrow$ x, generando un vettore L-dimensionale da un vettore variabile N-dimensionale:

$$\begin{aligned} x(v) = (1 \quad & v_1 \quad v_2 \quad \dots \quad v_N \quad v_1^2 \quad v_1v_2 \quad v_1v_3 \quad \dots \\ & v_1^3 \quad v_1^2v_2 \quad v_1^2v_3 \quad \dots \quad)^T \end{aligned} \tag{5.6}$$

La funzione x(v) determina il tipo del polinomio d(v) perciò x è chiamato struttura polinomiale. Introduciamo inoltre un coefficiente vettoriale L-dimensionale a e otteniamo la forma compatta per il polinomio scalare:

$$d(v) = a^T x(v) \tag{5.7}$$

È da notare che i polinomi di questo tipo non necessitano di essere completi. Nelle applicazioni pratiche, i polinomi di un certo grado G sono spesso costruiti in accordo con certe regole euristiche di costruzione che non implementano tutti i termini polinomiali di quel grado.

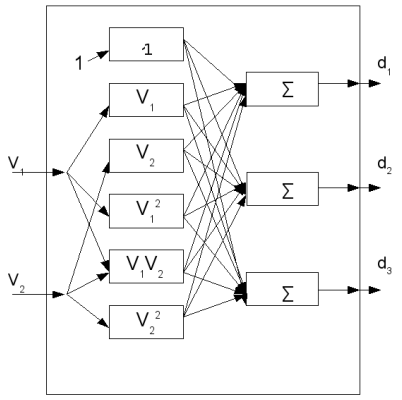

Figura 5.5: Rappresentazione del classificatore polinomiale con espansione completa con G=2 e N=2

5.3.2 Elementi base della classificazione polinomiale

Per il proposito della pattern classification è necessaria una funzione polinomiale vettoriale d(v) che consiste di K polinomi scalari:

$$d_k(v) = a_k^T x(v) \tag{5.8}$$

ognuna appartenente ad una delle K classi. Combiniamo gli a_k coefficienti vettoriali appartenti alla K-esima classe specifica in una matrice di coefficienti:

$$A = (a_1 \quad a_2 \quad ... \quad ... \quad a_k) \tag{5.9}$$

e si ottiene:

$$d(v) = A^T x(v) \tag{5.10}$$

Solamente la matrice dei coefficienti A verrà modificata durante la procedura di ottimizzazione. La struttura polinomiale è predeterminata e rimarrà invariata.

In figura 5.5 è rappresentato un semplice caso di classificatore con N=2 e G=2. Per valori maggiori di N e G, networks completi crescono rapidamente di dimensione e sono difficili da rappresentare.
Il nostro scopo è quello di costruire lo stimatore della minima media aritmetica quadratica d(v) per il vettore di target y che indica la vera appartenenza alla classe per il pattern [v,y] che deve essere classificato. In generale lo stimatore d(v) non combacia con il target y precisamente:

$$\Delta d(v) = d(v) - y = A^T x(v) - y \tag{5.11}$$

dove $\Delta d(v)$ è una variabile stocastica. Tutta l'adattabilità di questa struttura risiede nella matrice dei coefficienti A.
Esistono diversi metodi di training per il classificatore polinomiale basati o su metodi statistici o sul criterio della minimizzazione dell'errore quadratico medio. L'ottimizzazione basata sulla minimizzazione della media quadratica artimetica di A richiede che la varianza residua S^2 cioè il valore matematico atteso $E\left\{|\Delta d|^2\right\}$ della norma Euclidea quadratica del vettore di errore Δd, siano minima.

$$S^2 = E\left\{|d(v) - y|^2\right\} = min_d(V) \tag{5.12}$$

Il training del classificatore polinomiale si basa sul principio di imparare dagli esempi. In questo contesto ciò significa che la matrice dei coefficienti A dei polinomi deve essere adattata propriamente rispettando il criterio della minima media quadratica basandosi su un set dato v, y di campioni di apprendimento. L'ottimo della matrice dei coefficienti A è determinato da una matrice di equazioni lineari:

$$E\left\{xx^T\right\} A = E\left\{xy^T\right\} \tag{5.13}$$

composta delle due matrici dei momenti $E\left\{xx^T\right\}$ e $E\left\{xy^T\right\}$. Risolvere matrici di equazioni lineari non è un grosso problema, anche se le dimensioni sono grandi, dunque l'essenza di imparare dagli esempi di apprendimento significa stabilire le due matrici dei momenti basandosi su questi.

5.3.3 Applicazione della classificazione polinomiale al riconoscimento vocale

Focalizziamoci ora sul problema della classificazione nell'ambito del riconoscimento vocale per applicazioni militari. Sappiamo infatti che il dominio di applicazione influenza enormemente la scelta dei sensori, delle tecniche di pre-elaborazione e di normalizzazione dei dati, della rappresentazione degli stessi e del modello decisionale di classificazione.

Nei moderni sistemi di riconoscimento vocale sono richieste in particolare un'elevata accuratezza e una bassa complessità computazionale soprattutto per dispositivi embedded che non possiedono una grande quantità di memoria. I classificatori polinomiali hanno una struttura semplice che si adatta bene con i moderni DSP.

Per il classificatore polinimale del riconoscitore vocale implementato in questo lavoro di tesi, è stato utilizzato un nuovo metodo di training di tipo differenziale, che può facilmente trattare dataset molto ampi, utilizza poca memoria e può essere effettuato separandolo per classi.

La struttura di base del classificatore è mostrata in figura 5.6. I vettori delle caratteristiche, $x_1, x_2, ..., x_M$ sono gli input per la funzione discriminante,

$w^t p(x)$, la cui uscita viene mediata su tutte le M caratteristiche per produrre un punteggio. La funzione discriminante $w^t p(x)$ è composta

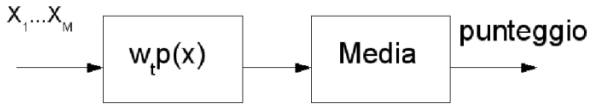

Figura 5.6: Struttura di base del classificatore polinomiale

da due parti: il termine w è il modello per la particolare classe, il termine $p(x)$ è il vettore su base polinomiale.
Quest'ultimo è un vettore composto dai monomi delle caratteristiche di ingresso fino al grado K. Per esempio per un vettore di feature bidimensionale:

$$x = [x_1 \quad x_2]^t \tag{5.14}$$

e K=2 abbiamo:

$$p(x) = [1 \quad x_1 \quad x_2 \quad x_1^2 \quad x_1x_2 \quad x_2^2]^t \tag{5.15}$$

Quindi la funzione discriminante di uscita è una combinazione lineare degli elementi della base polinimiale. Per ottenere una buona separazione fra le classi del sistema, viene utilizzato un criterio di errore di tipo MSE (mean-squared error o errore quadratico medio), che indica la discrepanza quadratica media fra i valori dei dati osservati ed i valori dei dati stimati. Per ogni classe del set di training, addestriamo la funzione discriminante $f_i(x) = w_i^t p(x)$ ad essere 1 per il vettore delle caratteristiche della classe considerata e 0 per le altre classi. Definiamo M_i come la matrice le cui righe sono composte dall'espansione polinomiale dei dati della classe i-esima:

$$M_i = [p(x_{i,1})^t] \tag{5.16}$$

dove N_i è il numero dei vettori di features per la classe i. Definiamo:

$$M = [M_1 \dots M_{N_{classes}}] \tag{5.17}$$

dove $N_{classes}$ è il numero delle classi. Il problema del training può essere così posto:

$$w_i^* = argmin_w ||Mw - o_i||_2 \tag{5.18}$$

Si può subito notare la somiglianza fra l'equazione 6.1 e la 5.12, dove Mw è l'insieme delle funzioni polinomiali vettoriali che si riferiscono alle classi in gioco e o_i è il target vector relativo alla i-esima classe composto da un numero di elementi pari al numero di righe della matrice M. In questo caso particolare ogni classe è rappresentata da N_i vettori di feature perciò il target vector è composto da N_i uno sulle righe corrispondenti all'i-esima classe e tutti zeri nelle restanti posizioni. Applicando il metodo delle equazioni normali arriviamo alla forma:

$$M^t M w_i = M^t o_i \tag{5.19}$$

definiamo con *1* il vettore di tutti uno e abbiamo:

$$\sum_{j=1}^{N_{classes}} M_j^t M_j w_i = M_i^t 1. \tag{5.20}$$

Se definiamo $R_j = M_j^t M_j$ e sostituiamo abbiamo:

$$(\sum_{j=1}^{N_{classes}} R_j) w_i = M_i^t 1 \tag{5.21}$$

Quest'ultima equazione è la base del metodo di training. Possiamo notare che ora il problema può essere separato, possiamo calcolare R_j

per ogni classe j e poi combinare il risultato finale in un'unica matrice $R_{all} = \sum_{j=1}^{N_{classes}} R_j$.
Per quanto rigurada il punteggio, considerando che il sistema è di tipo isolated word, il problema consiste nel calcolare i valori:

$$s_j = \frac{1}{M} \sum_{i=1}^{M} w_j^t p(x_i) \tag{5.22}$$

per ogni modello w_i delle classi del riconoscitore. La classe che ottiene il punteggio più elevato viene selezionata come la parola pronunciata. È possibile semplificare il calcolo definendo:

$$\bar{p} = \frac{1}{M} \sum_{i=1}^{M} p(x_i). \tag{5.23}$$

Notiamo che $s_j = w_j^t \bar{p}$. Perciò il punteggio per ogni modello diventa semplicemente un prodotto interno. Il numero di somme e prodotti per ottenere il punteggio è approssimativamente:

$$2N_{model}N_{frames} + 2N_{model}N_{words} \tag{5.24}$$

dove N_{model} è pari alla lunghezza di w. Un vincolo per questa equazione è dato da N_{words} che solitamente è un numero piccolo, la complessità computazionale in questa maniera infatti incrementa lentamente con l'ingrandirsi del vocabolario.
Questo tipo di calcolo per il punteggio utilizza il prodotto fra la funzione polinomiale discriminante e il modello di ogni parola j-esima calcolato dal processo di training, $w_j^t p(x)$, approssimando in questa maniera la probabilità a posteriori $p(j|x)$. Supponiamo di avere in ingresso dei vettori di feature $x_1, ..., x_N$, dapprima calcoliamo $p(x_1, ..., x_N|class_j)$, cioè la probabilità della sequenza dato il modello della j-esima parola, dunque abbreviamo questa scrittura ponendola come $p(x_1|class_j)$.

Una assunzione standard presente in letteratura per quanto riguarda il segnale vocale è di considerare i vettori di feature indipendenti quindi abbiamo:

$$p(x_1^N|class_j) = \prod_{i=1}^{N} p(x_i|class_j). \tag{5.25}$$

Utilizziamo ora la relazione:

$$p(x_i|class_j) = \frac{p(class_j|x_i)p(x_i)}{p(class_j)} \tag{5.26}$$

e otteniamo la funzione discriminante:

$$d'(x_1^N, class_j) = \prod_{i=1}^{N} \frac{p(class_j|x_i)}{p(class_j)} \tag{5.27}$$

Consideriamo il logaritmo della funzione discriminante:

$$log(d'(x_1^N, class_j)) = \sum_{i=1}^{N} log(\frac{p(class_j|x_i)}{p(class_j)}) \tag{5.28}$$

utilizzando le serie di Taylor, effettuiamo un'approssimazione lineare del $log(x)$ attorno ad $x = 1$ ponendolo uguale ad $x - 1$. Perciò otteniamo:

$$log(d'(x_1^N, class_j)) \approx \sum_{i=1}^{N} \frac{p(class_j|x_i)}{p(class_j)} - 1 \tag{5.29}$$

La funzione discriminante con le approssimazioni descritte è:

$$d'(x_1^N, class_j) = \frac{1}{N} \sum_{i=1}^{N} \frac{p(class_j|x_i)}{p(class_j)} \tag{5.30}$$

dove il termine -1 è stato trascurato poichè una costante di offset è ininfluente in una funzione razionale di probabilità. Nella formula precedente è stata effettuata una normalizzazione per il numero di frame

che assicurano che una soglia costante possa essere utilizzata per la verifica. L'approssimazione considerata è necessaria per la scalabilità di questa tecnica. Se sostituiamo al modello per ogni classe ,w_j, otteniamo:

$$d'(x_1^N, class_j) = \frac{1}{N}\sum_{i=1}^{N} p(x_i) w_j^t = w_j^t \bar{p} \tag{5.31}$$

dove:

$$\bar{p} = \frac{1}{N}\sum_{i=1}^{N} p(x_i). \tag{5.32}$$

Il punteggio per ogni modello di parola viene calcolato effettuando un prodotto interno, un'operazione di minima complessità.

Capitolo 6

Riconoscitore vocale: WORDIE

6.1 Training

La fase di training è essenziale per il corretto funzionamento del sistema ASR. Sappiamo infatti che il riconoscitore vocale si basa su un set stabilito di parole a lui note, chiamato vocabolario. Il riconoscimento si basa proprio su un confronto fra le caratteristiche estratte dalla parola pronunciata e il modello delle parole appartenenti al vocabolario. Questi modelli vengono sviluppati proprio durante la fase di addestramento momento in cui il sistema impara a riconoscere le diverse parole. Possiamo dunque affermare che dall'addestramento dipende l'accuratezza del sistema. In questo lavoro di tesi è stato implementato un nuovo metodo di training di tipo differenziale, che può facilmente trattare dataset molto ampi, utilizza poca memoria e può essere effettuato separandolo per classi. Il criterio seguito è quello della minimizzazione dell'errore quadratico medio. Per ogni classe del set di training, viene allenata la funzione discriminante $f_i(x) = w_i^t p(x)$ ad assumere il valore 1 per i vettori di feature corrispondenti alla classe della parola i considerata e 0 per i vettori di feature delle altre classi. Per costruire la matrice che conterrà i modelli relativi ad ogni parola del vocabolario definiamo M_i come una matrice le cui righe contengono

le espansioni polinomiali della classe di dati i.

$M_i =$

$$\begin{pmatrix} p(x_{i,1})^t \\ p(x_{i,2})^t \\ . \\ p(x_{i,Nj})^t \end{pmatrix}$$

dove N_i è il numero di vettori di feature per la classe i. Definiamo M=

$$\begin{pmatrix} M_1 \\ M_2 \\ \dots \\ \dots \\ M_{N_{classes}}] \end{pmatrix}$$

dove $N_{classes}$ è il numero delle classi. Il problema del training può essere posto come

$$w_i^* = argmin_w ||Mw - o_i||_2 \tag{6.1}$$

dove o_i è un vettore che consiste di N_i uni nelle righe in cui sono posizionati i dati relativi alla i-esima classe e tutti zeri nelle restanti posizioni.

Analizziamo ora nel dettaglio i passaggi che vengono eseguiti per completare il processo di addestramento del sistema.

L'intera fase di training è stata svolta tramite l'utilizzo del software Matlab. Una fase di inizializzazione è necessaria per impostare i valori di alcune costanti che verranno poi passate come parametri alle funzioni che compongono il sistema. Alcuni di questi valori rappresentano la durata dei frame, la larghezza minima di un impulso,la minima durata valida per una parola, il grado dell'espansione polinomiale, il numero delle parole che compongono il vocabolario, il numero di parlatori

del database, il numero di coefficienti LPC, il numero di coefficienti cepstrali, il numero di features estratte, etc. Per ogni comando vocale viene effettuta l'estrazione degli endpoints, cioè gli indici relativi ai frames in cui inizia e finisce la parola pronunciata, in modo da eliminare i silenzi e il rumore che precedono e seguono il parlato. Ogni parola di ogni parlatore che compone il database del sistema viene suddivisa in frame di 20 ms e per ogni frame viene calcolato il vettore delle energie normalizzato e viene centrato il profilo energetico del segnale intorno ai 0 dB. Dal confronto del vettore delle energie appena calcolato con opportune soglie sono calcolati due vettori: il primo, Pb, contiene gli indici relativi ai frame che corrispondono all'inizio di tutti gli impulsi contenuti nella parola pronunciata, il secondo, Pe, contiene gli indici dei frames relativi alla fine degli impulsi. L'algoritmo utilizzato per la rivelazione degli impulsi è descritto nell'articolo *An improved endpoint Detector for Isolated word recognition* [14]. L' assunzione di questo algoritmo è che la parola pronunciata contiene una sequenza di uno o più impulsi di energia, il solo problema sta nel determinare quali di questi appartiene alla parola suddetta. Dal file vocale selezionato attraverso Pb e Pe sono dapprima estratti i coefficienti di predizione lineare e in seguito i coefficienti cepstrali relativi ad ogni frame del segnale calcolati da questi attraverso opportune formule ricorsive e accumulati per riga all'interno di una matrice. Per ogni riga della matrice si calcola l'espansione polinomiale, di grado specificato al momento della chiamata, delle feature estratte dal segnale selezionato. Quello che si ottiene dunque è una matrice che ha per ogni riga l'espansione polinomiale delle feature del segnale vocale di partenza selezionato in base agli endpoints trovati. Il procedimento viene ripetuto per tutte le parole del vocabolario e per tutti i parlatori che compongono il database vocale. La matrice ottenuta quindi contiene i modelli relativi ad

ogni parola che compone il vocabolario del sistema di riconoscimento vocale. Il vocabolario di questo specifico riconoscitore è composto da 60 parole. Sono stati impiegati per questa fase 22 parlatori e i comandi sono stati suddivisi in sottomenù per garantire una buona distanza fra le classi.

6.2 Riconoscitore vocale

La struttura del riconoscitore vocale è mostrata in figura 6.1. Dall'osservazione dello schema risulta evidente la suddivisione del dispositivo in quattro stadi fondamentali:l'acquisizione del segnale, l'estrazione delle features, l'endpoint detection e la classificazione polinomiale. Per progettare il dispositivo e verificarne il funzionamento real-time è stato utilizzato il software Simulink che offre un ambiente simulativo specializzato per applicazioni caratterizzate da precisi vincoli di temporizzazione e operatività in tempo reale. È importante sottolineare il ruolo primario riservato alle operazioni di sincronizzazione, risultando fondamentale per una corretta risposta del riconoscitore che entrambe le fasi che precedono quella di classificazione terminino prima che essa inizi ad elaborare i dati presenti ai suoi ingressi. Per evitare che i dati elaborati dal classificatore polinomiale siano sovrascritti, è necessario fare in modo che le sue operazioni si concludano prima dell'inizio di una nuova iterazione. Nel contesto in esame tale possibilità di errore è limitata dalla struttura scelta per il riconoscitore che produce un risultato per ogni file vocale in ingresso nell'arco di una singola iterazione. Un ulteriore problema, non risolvibile a priori tramite opportuna scelta progettuale, si è presentato in termini di coerenza del tempo del clock tra i vari i rami in ingresso al classificatore polinomiale, condizione necessaria per ottenere un'elaborazione sensata delle informazioni

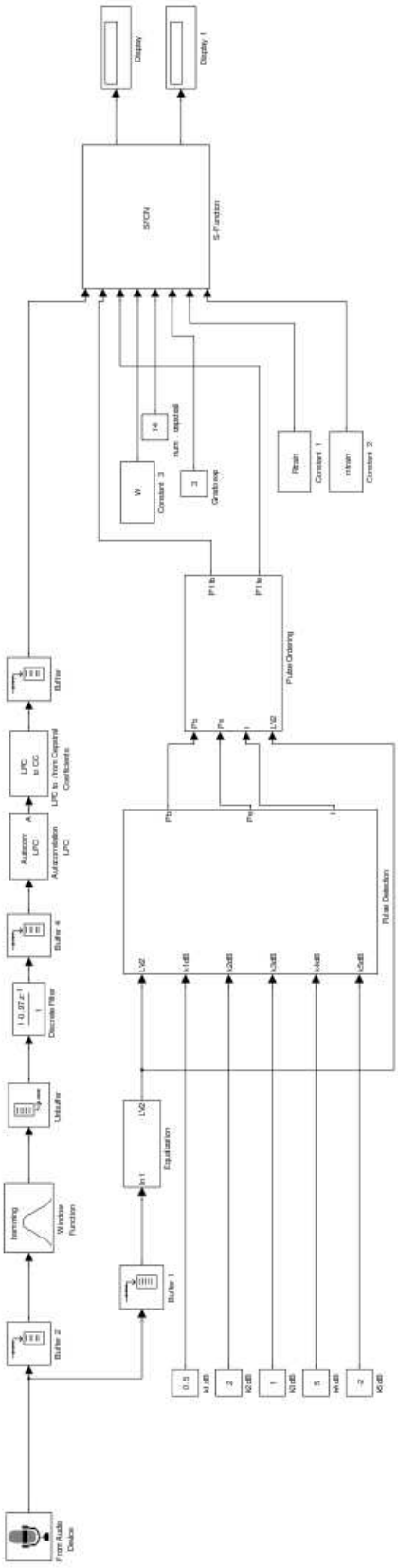

Figura 6.1: Schema a blocchi della fase di testing

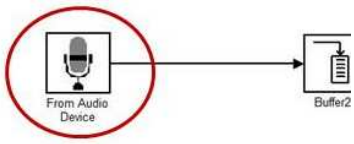

Figura 6.2: Microfono a cancellazione di rumore

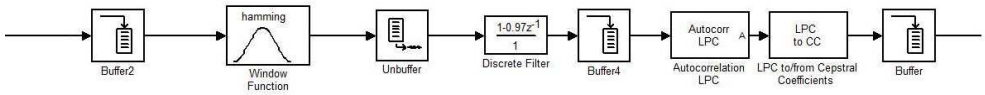

Figura 6.3: Estrazione delle features

contenute nel segnale in ingresso, data la stretta dipendenza fra i dati trattati in parallelo dal dispositivo.

6.2.1 Acquisizione del segnale

Il segnale vocale viene acquisito tramite un microfono a cancellazione di rumore, in grado di eliminare la presenza di rumore bianco (entro un range di potenza di rumore in ingresso ampio ma ovviamente limitato) nel segnale acquisito dal riconoscitore. Questo dispositivo è rappresentato dal primo blocco dello schema che è visibile nella figura 6.2. Il segnale viene campionato ad una frequenza di 8 kHz (sufficienti ad includere la parte significativa del parlato) e suddiviso in frames di 20 ms (in cui il segnale può in buona approssimazione essere considerato stazionario).

6.2.2 Estrazione delle features

La fase di estrazione delle features è rappresentata dal ramo superiore della figura 6.1 ed è mostrata in figura 6.3.
Si noti che ad ogni frame del segnale è applicata una finestra di Hamming, attualmente una delle maggiormente utilizzate nell'ambito del

trattamento iniziale del segnale per il riconoscimento vocale[[15]]. La finestra di Hamming è un caso specifico della finestra di Hanning. Una finestra di Hanning generalizzata è definita come segue:

$$w(n) = \frac{\alpha_w - (1 - \alpha_w)cos(2\pi n/(N_s - 1))}{\beta_w} \tag{6.2}$$

per $0 \leq n < N_s$ e $w(n) = 0$ altrove. Nell'intervallo [0,1], α_w è definita come una costante e N_s rappresenta il numero di campioni della finestra. Per implementare la finestra di Hamming, $\alpha_w = 0.54$. La costante normalizzata β_w è definita in modo che il valore efficace (RMS, *Root Mean Square*) della finestra risulti unitario. Definiamo β_w come segue:

$$\beta_w = \sqrt{\frac{1}{N_s} \sum_{n=0}^{N_s-1} w^2(n)} \tag{6.3}$$

Il blocco del sistema che applica al segnale la finestra così definita è mostrato nella figura 6.4.

Si noti che la finestra è stata applicata in modo che la potenza del segnale in ingresso e in uscita dal blocco che la implementa risulti approssimativamente invariata (*normalizzazione*). Questo tipo di normalizzazione è conveniente specialmente nelle implementazioni che utilizzano hardware a virgola fissa. Lo scopo principale della finestratura di Hamming è quello di pesare o privilegiare i campioni posizionati nella zona centrale della finestra. Tale operazione di pe-

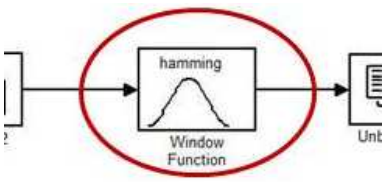

Figura 6.4: Finestra di Hamming utilizzata nel riconoscitore

satura, insieme ad un utilizzo opportuno della sovrapposizione delle finestre applicate al segnale di ingresso (*overlapping*), fornisce la base per l'ottenimento di una stima lentamente variante e che non perda di significatività nelle zone periferiche della finestra di Hamming considerata. È importante che la profondità del lobo principale nella risposta in frequenza sia più piccola possibile, altrimenti il processo finestrato potrebbe risultare inadeguato alla successiva analisi spettrale.

La durata di un frame, T_f, è definita come il tempo per il quale un set di parametri rimane valido. Il periodo di frame è utilizzato in maniera simile per indicare il periodo di tempo che intercorre tra il calcolo di un parametro e il successivo. Il rate di frame, un altro termine molto comune, è il numero di frame calcolati al secondo (Hz). La durata di un frame nei sistemi effettivamente implementati è tipicamente compresa fra 20 msec e 10 msec. I valori compresi in questo range rappresentano un compromesso accettabile tra il tasso di cambiamento dello spettro e la complessità del sistema. La durata di un frame è in ultimo dipendente dalla velocità del parlatore (tasso di cambiamento nella forma del tratto vocale). Alcuni suoni del parlato mostrano bruschi cambiamenti della transizione spettrale che possono risultare come picchi spettrali che traslano oltre gli 80 Hz/msec; di conseguenza non è normalmente utilizzata una durata di frame inferiore a 8 msec. La durata del frame e la durata della finestra sono normalmente settate in coppia: un durata della finestra di 30 msec è comunemente affiancata ad una durata di frame di 20 msec, mentre una durata della finestra di 20 msec è utilizzata insieme ad una durata di frame di 10 msec. In generale, poiché una durata di frame breve è utilizzata per catturare rapide dinamiche delle spettro, la durata della finestra dovrebbe in corrispondenza essere più breve, così che i dettagli dello spettro non siano eccessivamente smussati. Nell'implementazione di

questo specifico riconoscitore vocale si è optato per una finestra di durata pari a 20 msec. L'analisi basata sui frames è spesso chiamata analisi con overlapping, perché, per ogni nuovo frame, solo una frazione di dati del segnale cambia. La quantità dell'overlap controlla quanto velocemente possono cambiare i parametri da frame a frame. La percentuale di overlap è data da:

$$\%Overlap = \frac{(T_w - T_f)}{T_w} \times 100\% \tag{6.4}$$

dove T_w è la durata della finestra (in secondi) e T_f è la durata del frame. Se $T_w < T_f$ la percentuale di overlapping risulta nulla, quantità effettivamente adottata, per semplicità, nella soluzione implementata.

In seguito alla finestratura di Hamming ogni frame del segnale viene pre-enfatizzato attraverso un opportuno filtro FIR (*Finite Impulse Response*) mostrato in figura 6.5 , la cui risposta impusiva è data da:

$$H_{pre}(z) = \sum_{k=0}^{N_{pre}} a_{pre}(k) z^{-k}. \tag{6.5}$$

Normalmente viene utilizzato un filtro digitale ad un solo coefficiente, chiamato filtro di preenfasi[[15]]:

$$H_{pre}(z) = 1 + a_{pre} z^{-1}. \tag{6.6}$$

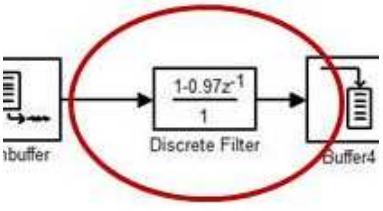

Figura 6.5: Filtro di preenfasi

Un range tipico di valori per a_{pre} è $[-1.0, -0.4]$. Valori vicini a -1.0 che possono essere efficientemente implementati tramite hardware a virgola fissa, come -1 o $-(1 - 1/16)$, sono i più comuni per il riconoscimento vocale. Il filtro di preenfasi ha lo scopo di incrementare il segnale di 20 dB per decade (un incremento di un ordine di grandezza in frequenza). Ci sono due spiegazioni principali per i vantaggi nell'utilizzo di questo filtro. La prima risiede nel fatto che le sezioni del segnale vocale presentano una pendenza spettrale negativa (attenuazione) che è approssimativamente di 20 dB per decade ed è dovuta a caratteristiche fisiologiche del sistema di produzione del linguaggio parlato. Il filtro di preenfasi serve a controbilanciare questa naturale attenuazione prima dell'analisi spettrale al fine di incrementare l'efficienza dell'analisi. Una spiegazione alternativa è che l'udito è più sensibile nelle regioni dello spettro superiori ad 1 kHz. Il filtro di preenfasi amplifica questa area dello spettro, coadiuvando l'algoritmo di analisi spettrale nel modellare gli aspetti più importanti dal punto di vista percettivo dello spettro del parlato. Si noti, inoltre, che questi filtri enfatizzano le frequenze oltre i 5 kHz, una regione nella quale il sistema uditivo diventa incrementalmente meno sensibile. Tuttavia, le frequenze sopra i 5 kHz sono naturalmente attenuate dal sistema di produzione del parlato e normalmente sono assegnate a pesi significativamente più piccoli in un tipico sistema di riconoscimento vocale.

Sono stati inoltre proposti algoritmi più sofisticati per l'ottenimento della preenfasi. Uno di questi è la preenfasi adattativa, nella quale la pendenza spettrale è automaticamente appiattita prima dell'analisi spettrale. Altri algoritmi che utilizzano filtri sagomati e che attenuano aree dello spettro risultano essere piuttosto rumorosi. Recentemente sono state implementate delle classificazioni parlato/rumore per gli

algoritmi basati su filtri adattativi. Nessuno degli approcci ora riportati ha comunque ancora fornito prestazioni ottimali nelle applicazioni riguardanti il riconoscimento vocale. Attualmente molti sistemi di riconoscimento vocale hanno eliminato lo stadio di preenfasi completamente e hanno compensato l'attenuazione spettrale come parte del modello statistico per il riconoscimento vocale.

Tra le tecniche utilizzate per l'implementazione del riconoscimento vocale la famiglia basata sui coefficienti lineari di predizione e sui coefficienti cepstrali è quella preminente per le sue prestazioni e la sua relativa semplicità realizzativa. Il metodo LPC/Cestrum modellizza un segnale che evolve nel tempo tramite un set ordinato di coefficienti rappresentanti l'inviluppo spettrale del segnale. Questo è costituito da una curva che passa vicino ai picchi dello spettro del segnale originale. Per ottenere la rappresentazione LPC/Cepstrum, il primo passo è quello di calcolare i coefficienti LPC. Questi sono coefficienti di un modello autoregressivo che minimizzano la differenza tra i valori di predizione lineare e i valori effettivi nella finestra temporale considerata. Una volta ottenuti i coefficienti LPC, è possibile, a partire dai loro valori, calcolare i coefficienti cepstrali. Le serie cepstrali rappresentano una progressiva approssimazione dell'inviluppo del segnale; come nel caso degli LPC, quanto più è elevato il numero di coefficienti cepstrali considerati, più l'inviluppo si avvicina allo spettro originale. Nel riconoscimento vocale la tecnica LPC/Cepstrum ha dimostrato di catturare le informazioni rilevanti contenute nelle serie originali. Questi coefficienti permettono di ottenere un rappresentazione molto sintetica dell'evoluzione del fenomeno nel tempo.

Nel caso del modello considerato in questo lavoro sono stati calcolati 11 coefficienti LPC e 12 coefficienti cepstrali ottenuti da questi tramite

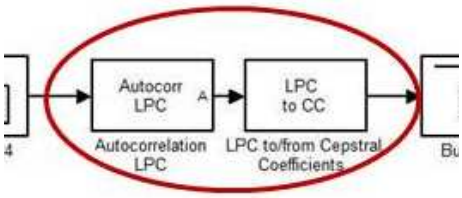

Figura 6.6: Coefficienti LPC/Cepstrum

il blocco di trasformazione LPCtoCC, fornito dal software Simulink, riportato nella figura 6.6.

Verranno in questa sezione validati i metodi di estrazione utilizzati nel simulatore del riconoscitore vocale facendo uso di *xKl*, un programma per l'analisi spettrale del segnale vocale in grado di fornire un'analisi estremamente accurata del segnale vocale di interesse. In particolare xKl usa per il calcolo di tale spettro un modello di rappresentazione dell'apparato di riproduzione vocale di tipo articolatorio, più sofisticato e accurato di quello utilizzato nell'ambito di Simulink. Ricordiamo infatti che per rappresentare lo spettro di uno specifico frame del segnale vocale attraverso i coefficienti LPC viene utilizzata la seguente funzione di trasferimento rappresentate il tratto vocale:

$$H(z) = \frac{S(z)}{U(z)} = \frac{G}{1 - \sum_{k=1}^{p} a_k z^{-k}} \tag{6.7}$$

dove con a_k vengono rappresentati i coefficienti LPC.

Nell' analisi che segue la durata di ogni frame analizzato è di 20 ms, ad ogni frame viene applicata una finestra di Hamming, e vengono calcolati 12 coefficienti LPC.

Nelle figure 6.7 e 6.8 sono rappresentati i confronti relativi ai primi due frame della parola presa in considerazione. Nella parte destra delle due figure è rappresentata l'approssimazione dello spettro mediante coefficienti LPC, mentre in quella di sinistra possiamo vedere i risultati ottenuti mediante xKl. Come è facile osservare, i coefficienti ottenuti

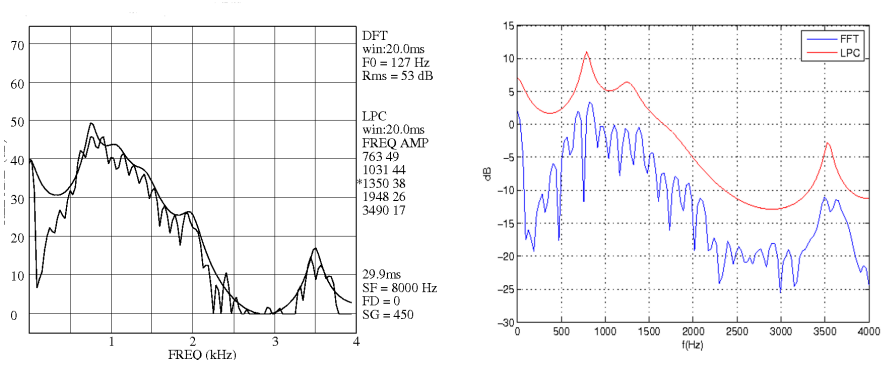

Figura 6.7: Confronto spettri relativi al primo frame

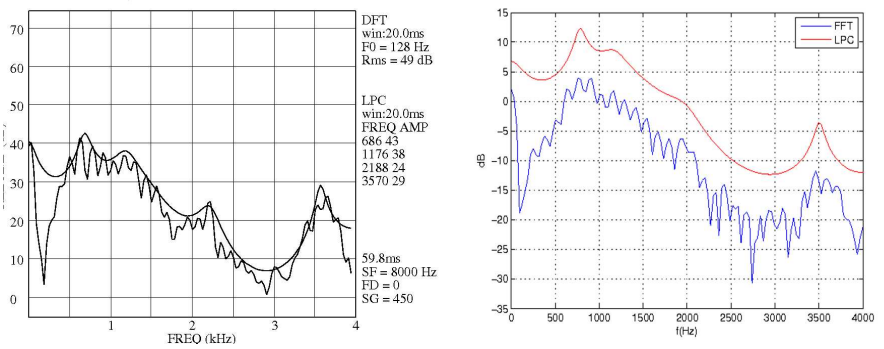

Figura 6.8: confronto spettri relativi al secondo frame

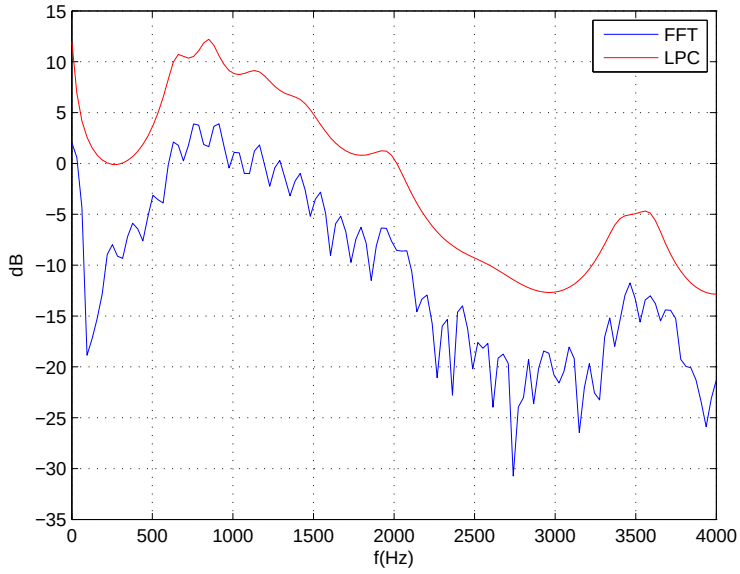

Figura 6.9: Approssimazione dello spettro relativa al secondo frame con 20 LPC

tramite il blocco Simulink sono coerenti in quanto approssimano in maniera sufficientemente accurata le frequenze formanti proprie dei frame del segnale vocale analizzato. L'accuratezza dell'approssimazione è migliore nei grafici ottenuti con xKl proprio perché, come è stato già sottolineato, il modello utilizzato da questo software è più raffinato. Poiché sappiamo che l'accuratezza del modello da noi utilizzato per rappresentare lo spettro del segnale a partire dai coefficienti LPC migliora all'aumentare di essi, viene mostrato nella figura 6.9 l'analisi ottenuta utilizzando un numero maggiore di coefficienti, ovvero pari a 20.

Allo stesso modo è possibile effettuare un'analisi dei coefficienti cepstrali presenti in uscita dal blocco LPCtoCC. Nella figura 6.10 è rappresentato l'andamento temporale di un frame del segnale vocalico di 30 msec rappresentante la vocale 'A'.

La scelta del segnale vocalico è stata determinata dalla sua intrinseca periodicità; possiamo infatti notare come il segnale si ripeta all'incirca

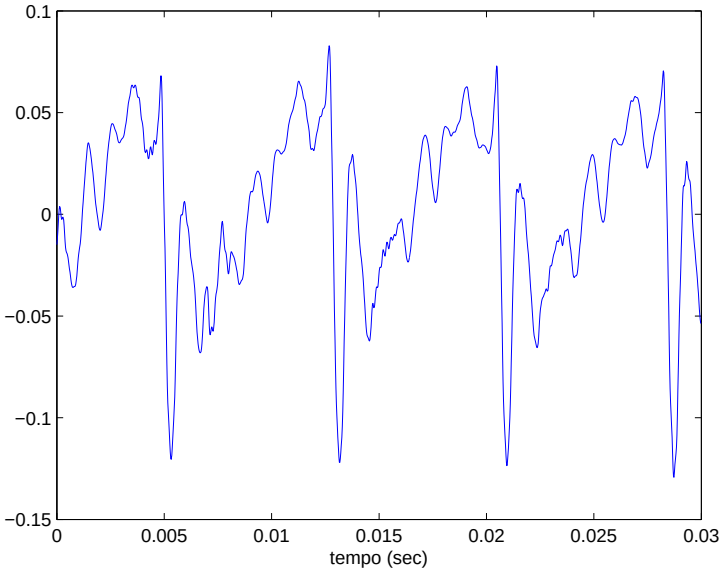

Figura 6.10: Andamento temporale di un frame del segnale vocalico

ogni 8 msec. I coefficienti cepstrali sono ottenuti da quelli LPC tramite la seguente formula:

$$c_m = -a_m + \frac{1}{m}\sum_{k=1}^{m-1}[-(m-k)\cdot a_k \cdot c_{(m-k)}] \qquad 1 \leq m \leq p \qquad (6.8)$$

$$c_m = \sum_{k=1}^{p}[\frac{-(m-k)}{m}\cdot a_k \cdot c_{(m-k)}] \qquad p < m < n \qquad (6.9)$$

dove con a_k vengono rappresentati i coefficienti LPC e con c_{m-k} i coefficienti cepstrali precedentemente calcolati. Poiché i coefficienti cepstrali rappresentano lo spettro del logaritmo dello spettro del segnale di partenza, è facile intuire che graficandoli si otterranno dei picchi proprio in corrispondenza del valore del periodo del segnale temporale iniziale. Nella figura 6.11 sono messi a confronto i valori dei coefficienti cepstrali ottenuti tramite la conversione degli LPC e quelli calcolati applicando la definizione teorica. Come si può osservare è presente un picco, in entrambi i casi, nell'intorno degli 8 msec, inoltre l'asse delle ascisse non è né nel dominio del tempo né in quello delle frequenze,

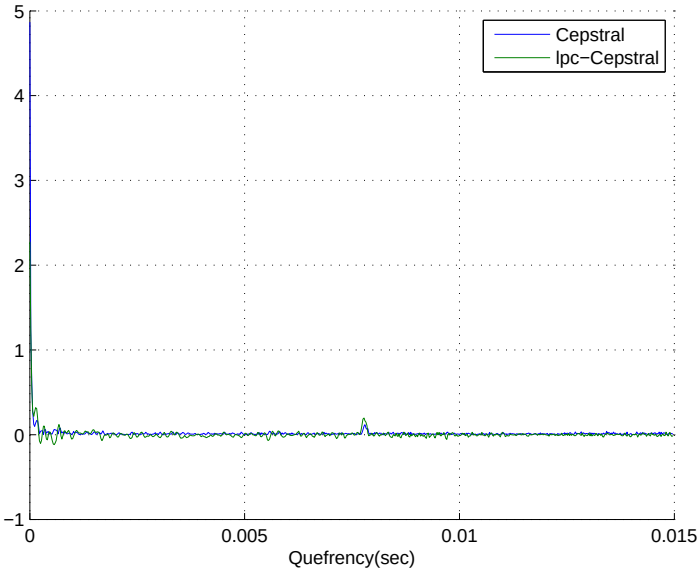

Figura 6.11: Confronto coefficienti cepstrali

ma in un nuovo dominio chiamato *Quefrency*. Nel modello Simulink utilizzato per il testing, è stato scelto di ricavare i coefficienti cepstrali dai coefficienti LPC per due motivi: per una questione di efficienza computazionale, in quanto con questo approccio ci si riduce a risolvere esclusivamente sistemi di equazioni lineari e si ottengono dei picchi più accentuati, come possiamo notare nella figura 6.12.

Una volta calcolati i coefficienti cepstrali relativi ad ogni frame del segnale, essi vengono raccolti in un buffer (figura 6.13) per ottenere in un'unica matrice tutte le features necessarie per le trasformazioni successive.

La matrice infatti diventerà uno degli ingressi del blocco S-function, necessario per poter includere un codice scritto in linguaggio $C/C++$ all'interno della struttura Simulink del riconoscitore. In questo blocco verranno implementate due fondamentali operazioni: la classificazione polinomiale, di cui parleremo tra poco, che produrrà il risultato finale del riconoscitore vocale, e la trasformazione affine. Quest'ultima è sta-

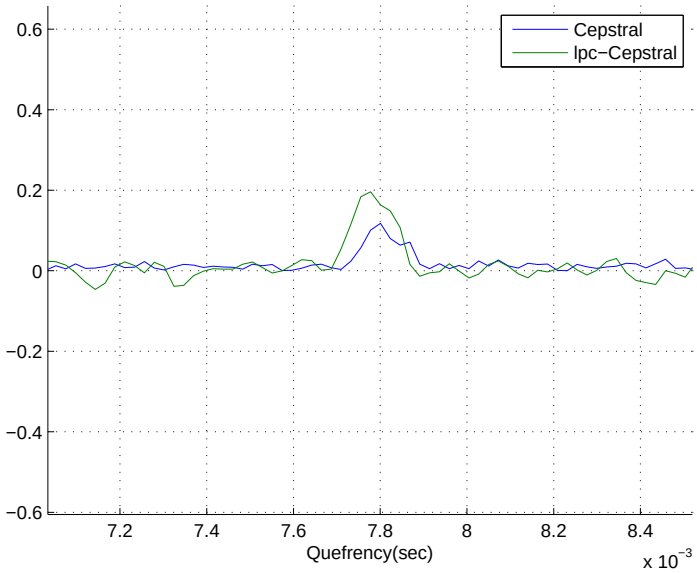

Figura 6.12: Confronto dei picchi dei coefficienti cepstrali calcolati tramite le due differenti metodologie

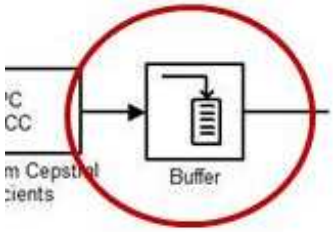

Figura 6.13: Buffer di raccolta di tutte le features relative al segnale

ta utilizzata per effettuare la compensazione del rumore e dei disturbi introdotti dal canale [16]. Ogni vettore di features x_i viene eleborato tramite la seguente trasformazione:

$$y_i = Ax_i + b_i, \tag{6.10}$$

in cui A è principalemente responsabile della distorsione dovuta al rumore e b della distorsione introdotta dal canale. Questo approccio è risultato di successo in molte situazioni. Gli effetti di questa operazione possono essere divisi in due parti: dapprima viene effettuata

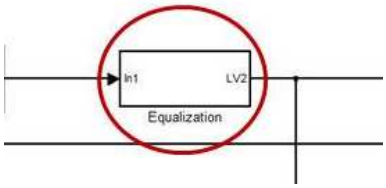

Figura 6.14: Blocco che calcola il vettore energetico normalizzato rappresentante il segnale vocale

una *cepstral mean subtraction* (CMS) sul vettore delle features e in un secondo momento l'input traslato viene normalizzato rispetto alla varianza dei dati di training al fine di compensare il rumore. È stato dimostrato che la mean subtraction incrementa significativamente le prestazioni del sistema per il quale il training è stato effettuato in condizioni di canale diverse da quelle di testing. E' stato però appurato che, nel caso di uguali condizioni di canale per il training ed il testing, si hanno notevoli perdite nell'accuratezza del riconoscitore. Nella consapevolezza di tale fenomeno, nel blocco S-function è stato implementato un metodo per poter sceglere se applicare o meno la trasformata affine alle features del segnale.

6.2.3 Endpoint Detection

In riferimento alla figura 6.1, possiamo vedere come la fase di endpoint detection venga svolta dal ramo inferiore dello schema. In base all'algoritmo descritto al capitolo 3 una volta suddiviso il segnale in frames di 20 ms, per ognuno di essi viene calcolato il profilo energetico normalizzato tramite il blocco Equalization mostrato in figura 6.14.
Una volta effettuata questa operazione, tramite il blocco Pulse Detection, mostrato in figura 6.15, viene effettuata la rivelazione degli impulsi scorrendo il vettore del profilo energetico da sinistra a destra e confrontandolo con un sistema di cinque soglie.

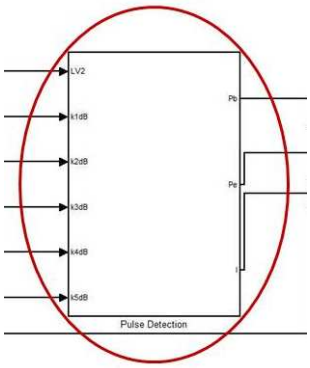

Figura 6.15: Blocco che effettua la rivelazione degli impulsi energetici presenti nel segnale vocale

L'uscita del blocco Pulse Detection è composta dai due vettori P_b e P_e, che tengono conto rispettivamente dei frame iniziali e finali degli impulsi energetici rilevati, e dal numero I degli impulsi trovati. Questi tre elementi, insieme al vettore del profilo energetico, sono gli ingressi del blocco Pulse Ordering(figura 6.16).
Lo scopo di quest'ultimo è di determinare un insieme di coppie di word endpoint ordinate in termini di probabilità decrescente. L'ordinamento si basa sulle seguenti assunzioni:

1. La parola da analizzare contiene uno o più impulsi di energia.
2. Il frame ad energia massima è sempre incluso all'interno della parola.
3. Più è grande lo stopgap tra due impulsi di energia, minore è la probabilità che essi facciano parte di una stessa parola.
4. Impulsi di energia separati da quello contenente il frame ad energia massima da uno stopgap più grande di 200 ms molto probabilmente non apparterranno alla parola.

In uscita dal blocco appena descritto dunque abbiamo due vettori $P11_b$ e $P11_e$ che contengono rispettivamente una lista di indici di

frame che indicano l'inizio e la fine della parola ordinati con probabilità decrescente.

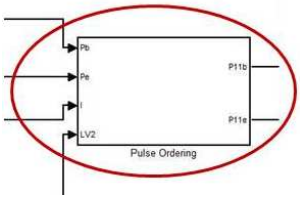

Figura 6.16: Blocco che rileva i limiti della parola e li ordina in termini di probabilità

6.2.4 Classificazione Polinomiale

Il classificatore polinomiale, mostrato in figura 6.17, è stato implementato tramite il blocco S-Function.
In ingresso a tale blocco abbiamo:

1. La matrice contenente sulle sue righe i coefficienti cepstrali relativi ai diversi frame del segnale di ingresso.

2. I due vettori $P11_b$ e $P11_e$ che contengono rispettivamente una lista di indici di frame che indicano l'inizio e la fine della parola ordinati con probabilità decrescente.

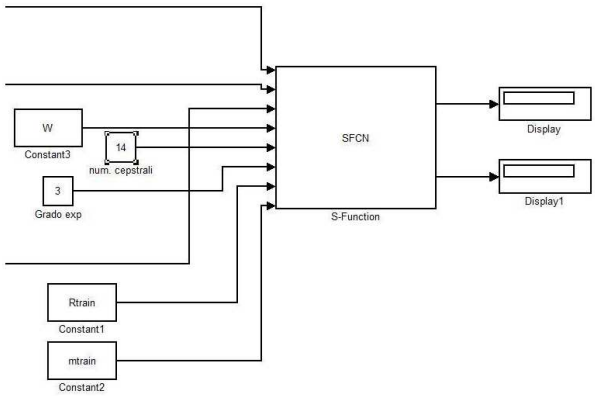

Figura 6.17: Blocco che effettua l'operazione di classificazione polinomiale

3. La matrice W, calcolata nella fase di training, che contiene sulle sue colonne i vettori rappresentativi di ogni classe del classificatore polinomiale.

4. Il numero di coefficienti cepstrali.

5. Il grado dell'espansione polinomiale.

6. R_train, la matrice di covarianza calcolata nella fase di training necessaria per la trasformata affine.

7. Il vettor medio b_train rappresentante la totale distribuzione dei dati di training.

La struttura di base del classificatore polinomiale è composta da diverse classi ognuna delle quali è associata ad una parola. Nella fase di training è stato calcolato un vettore numerico w per ogni parola rappresentante la particolare classe. Nella fase di testing per ogni classe si effettua un prodotto scalare del vettore w che la rappresenta, con le espansioni polinomiali dei vettori di feature estratte dai diversi frame che compongono la parola pronunciata. Al seguito di queste operazioni si ha un valore per ogni frame ed effettuando la media tra i valori calcolati, si ottiene un punteggio per la particolare classe. Dopo aver ripetuto l'operazione per tutte le classi esistenti, si decide per la parola corrispondente alla classe con punteggio più alto. L'uscita del blocco è infatti rappresentata dal valore del punteggio massimo e dall'indice della classe corrispondente.

Capitolo 7

Validazione e sviluppi futuri

In questo lavoro di tesi è stato progettato un sistema automatico di riconoscimento vocale (ASR) di tipo speaker independent rivolto ad un ambito militare. Abbiamo delineato in precedenza il vasto campo di applicazione di questi sistemi e sottolineato che, per una buona progettazione di un ASR, è assolutamente fondamentale definire bene il contesto operativo, in quanto fattore determinante per le prestazioni dell'intero sistema. In particolare l'ambito in cui lavora il riconoscitore vocale progettato è di difficile gestione in quanto le condizioni operative saranno spesso critiche a causa della presenza di forte rumore in ambienti quali la cabina di pilotaggio di un aereo o quella di un carro armato. Il rumore può anche causare un'alterazione della voce denominata effetto Lombard, che può anch'essa generare problemi nella corretta rivelazione delle parole. È opportuno inoltre considerare le condizioni psico-fisiche dell'utente dell'ASR. Le operazioni militari sono spesso condotte sotto condizioni di stress indotte da grossi carichi di lavoro, privazione del sonno, paura ed emozioni, confusione dovuta all'informazione conflittuale, tensione psicologica, dolore e altre tipiche condizioni incontrate in contesti militari. Queste condizioni influenzano le abilità fisiche e cognitive delle caratteristiche della voce umana.

I sistemi commerciali non sono ancora in grado di ottenere risultati apprezzabili in situazioni estreme. Il progresso in campo militare degli ASR è molto limitato a causa della mancanza di disponibilità di database di parole sotto stress. In particolare le condizioni in cui può trovarsi un soldato in un campo di battaglia non sono facilmente simulabili, e perciò è difficile collezionare sistematicamente dati per il loro utilizzo nella ricerca e nell'addestramento degli ASR. Un ulteriore limite è legato alla lingua, infatti database a disposizione di voci registrate in diverse condizioni ambientali e fisiche sono principalemente in lingua inglese. Questo comporta una grande difficoltà per l'addestramento del sistema, fase che sappiamo essere determinante per l'accuratezza del riconoscitore, soprattutto nel caso di un ASR speaker independent.

Come abbiamo visto nei precedenti capitoli, tramite l'utilizzo di software atti a tale scopo, quali Matlab e Simulink, sono stati implementati tutti gli elementi che costituiscono l'intero sistema quali l'endpoint detector, l'estrattore di caratteristiche e il classificatore. Successivamente è stato scelto un vocabolario composto da menù costituiti da un numero di parole massimo limitato a cinque comandi per poter suddividere lo spazio decisionale in maniera da fornire la maggior distanza possibile fra le varie classi. Il sistema implementato è stato addestrato da 22 parlatori utilizzando un microfono a cancellazione di rumore ed effettuando le registrazioni in condizioni nominali cioè prive di rumore e di stress.

La fase di test del riconoscitore vocale in condizioni operative nominali si è svolta utilizzando 15 parlatori esterni, cioè non addestratori del sistema, che in tempo reale hanno pronunciato più volte tutte le parole dei menù considerati al fine di calcolare una percentuale di riconoscimento. È stato validato un menù composto da 4 parole : *speaker*,

grado espansione polinomiale	3
n.coefficienti LPC	10
n.coefficienti cepstrali	11
n.parlatori training	22

Tabella 7.1: Tabella dei parametri utilizzati per la fase di training

parole	speaker	radio	monitor	conferenza
test1-diego	86,66667%	80%	100%	100%
test2-fabrizio	100%	93,3333%	100%	100%
test3-riccardo	100%	86,6667%	100%	100%
test4-gregorio	53,33333%	60%	86,66667%	100%
test5-raffaele	46,666667%	93,33333%	100%	100%
test6-germano	86,6667%	20%	93,33333%	100%
test7-paolo	86,66667%	93,33333%	86,66667%	73,333333%
test8-emilio	100%	86,6667%	86,6667%	100%
test9-nicola	93,33333%	53,33333%	93,33333%	100%
test10-luca	100%	20%	100%	100%
test11-alessandro	100%	100%	80%	100%
media	86,11%	72,78%	93,89%	93,89%

Tabella 7.2: Tabella dei risultati ottenuti per un menù composto da quattro parole

radio, *monitor*, *conferenza*. I coefficienti LPC estratti sono pari a 10 e i relativi coefficienti cepstrali sono 11, il grado di espansione polinomiale è pari a 3 e il training su cui viene fatto il confronto è composto da 22 parlatori. Nelle tabelle sottostanti possiamo osservare i parametri settati e i risultati ottenuti.

Notiamo che le percentuali di riconoscimento delle parole sono diverse per i vari parlatori, questo è dovuto alle differenze di pronuncia delle varie parole la cui causa predominante è il dialetto. È possibile comunque individuare in questo particolare menù una parola critica la cui percentuale media è pari a 72,78%: *radio*.

Questa parola ha la catteristica di essere breve ed iniziare per una

parole	monitor	conferenza	speaker	pilota	mappa
test1-diego	66%	100%	73,33333%	53,33333%	100%
test2-fabrizio	93,33333%	100%	60%	80%	86,66667%
test3-riccardo	93,33333%	100%	60%	93,33333%	93,33333%
test4-gregorio	66,66667%	100%	66,666667%	66,66667%	93,33333%
test5-raffaele	33,33333%	100%	86,66667%	100%	86,66667%
test6-germano	93,33333%	100%	80%	53%	93,33333%
test7-paolo	86,66667%	86,66667%	40,00%	80%	80%
test8-emilio	86,66667%	100%	100%	80%	100%
test9-nicola	60%	100%	93,33333%	10%	100%
test10-luca	100%	100%	93,33333%	73,333333%	86,66667%
test11-alessandro	60%	100%	86,66667%	93,33333%	100%
test12-lorenzo	100%	100%	93,33333%	100%	100%
media	78,28%	98,89%	77,78%	73,61%	93,33%

Tabella 7.3: Tabella dei risultati ottenuti per un menù composto da cinque parole

consonate vibrante di difficile pronuncia.

In base ai risultati ottenuti è stato ritenuto opportuno sostituire all'interno di questo menù la parola *radio* per creare un nuovo menù di 5 parole: *monitor*, *conferenza*, *speaker*, *pilota*, *mappa*. In questo caso sono stati calcolati 14 coefficienti LPC e 15 coefficienti cepstrali con un grado di espansione polinomiale pari a 3 e un training effettuato in condizioni nominali su 22 parlatori. In figura 7.3 possiamo osservare i risultati ottenuti.

Possiamo notare che sia in questo menù tabella (7.3) che nel precedente tabella (7.2) la parola *conferenza* ha ottenuto ottimi risultati con percentuali fino al 99%.

La parola *radio* è stata sostituita con la parola *mappa* per ottenere un menù strutturato in maniera più performante. È opportuno notare che sono state ottenute delle percentuali di riconoscimento più elevate rispetto al caso precedente considerando anche il fatto che i vettori

n. coefficienti LPC	10			
n. coefficienti cepstrali	11			
n. parlatori	16			
grado espansione	3			
uno	due	tre	quattro	cinque
87,5%	62,5%	100%	100%	75%

Tabella 7.4: Tabella dei risultati ottenuti per un menù composto da cinque parole a cui è stato sovrapposto il rumore del carro armato campionato

n. coefficienti LPC	10		
n. coefficienti cepstrali	11		
n. parlatori	16		
grado espansione	3		
speaker	radio	monitor	conferenza
75%	75%	62,5%	75%

Tabella 7.5: Tabella dei risultati ottenuti per un menù composto da quattro parole a cui è stato sovrapposto il rumore del carro armato campionato

relativi ad ogni parola sono più vicini essendo aumentata la cardinalità del menù.

Per quanto rigurda il rumore, è stato scelto di effettuare il training sovrapponendo ai comandi registrati in condizioni nominali il rumore del carro armato campionato. Il livello del rumore è stato scelto in maniera da rendere l'applicazione la più verosimile possibile. Rumori più elevati infatti porterebbero delle distorsioni della voce dell'utente dovute all'effetto Lombard che in questa fase non è stato considerato. La validazione del riconoscitore è stata eseguita utilizzando le registrazioni di 8 parlatori esterni e sovrapponendo, come per il training, lo stesso livello di rumore. Sono stati analizzati diversi menù di 4 o 5 parole.

Ricordiamo che il riconoscimento delle parole viene effettuato sulla

n. coefficienti LPC	10		
n. coefficienti cepstrali	11		
n. parlatori	16		
grado espansione	3		
ingaggia	rilascia	interrompi	ripeti
87,5%	62,5%	100%	37,5%

Tabella 7.6: Tabella dei risultati ottenuti per un menù composto da quattro parole a cui è stato sovrapposto il rumore del carro armato campionato

n. coefficienti LPC	10		
n. coefficienti cepstrali	11		
n. parlatori	16		
grado espansione	3		
affilia	fisio	sensori	esci
100%	37,5%	75%	50%

Tabella 7.7: Tabella dei risultati ottenuti per un menù composto da quattro parole a cui è stato sovrapposto il rumore del carro armato campionato

n. coefficienti LPC	10		
n. coefficienti cepstrali	11		
n. parlatori	16		
grado espansione	3		
luminosità	aumenta	diminuisci	attiva
100%	87,5%	75%	50%

Tabella 7.8: Tabella dei risultati ottenuti per un menù composto da quattro parole a cui è stato sovrapposto il rumore del carro armato campionato

n. coefficienti LPC	10		
n. coefficienti cepstrali	11		
n. parlatori	16		
grado espansione	3		
sei	sette	otto	nove
37,5%	100%	87,5%	87,5%

Tabella 7.9: Tabella dei risultati ottenuti per un menù composto da quattro parole a cui è stato sovrapposto il rumore del carro armato campionato

n. coefficienti LPC	10		
n. coefficienti cepstrali	11		
n. parlatori	16		
grado espansione	3		
affermativo	negativo	dentro	fuori
100%	62,5%	37,5%	62,5%

Tabella 7.10: Tabella dei risultati ottenuti per un menù composto da quattro parole a cui è stato sovrapposto il rumore del carro armato campionato

n. coefficienti LPC	10		
n. coefficienti cepstrali	11		
n. parlatori	16		
grado espansione	3		
allarmi	batterie	messaggi	mappa
75%	75%	62,5%	87,5%

Tabella 7.11: Tabella dei risultati ottenuti per un menù composto da quattro parole a cui è stato sovrapposto il rumore del carro armato campionato

n. coefficienti LPC	10		
n. coefficienti cepstrali	11		
n. parlatori	16		
grado espansione	3		
acceso	spento	precedente	disattivazione
75%	37,5%	100%	100%

Tabella 7.12: Tabella dei risultati ottenuti per un menù composto da quattro parole a cui è stato sovrapposto il rumore del carro armato campionato

n. coefficienti LPC	10		
n. coefficienti cepstrali	11		
n. parlatori	16		
grado espansione	3		
connetti	escludi	ritrasmissione	chiudi
63%	37,5%	75%	63%

Tabella 7.13: Tabella dei risultati ottenuti per un menù composto da quattro parole a cui è stato sovrapposto il rumore del carro armato campionato

n. coefficienti LPC	10		
n. coefficienti cepstrali	11		
n. parlatori	16		
grado espansione	3		
disconnetti	configurazione	chiamata	puntatore
37,5%	87,5%	100%	37,5%

Tabella 7.14: Tabella dei risultati ottenuti per un menù composto da quattro parole a cui è stato sovrapposto il rumore del carro armato campionato

base delle loro caratteristiche spettrali e su una approssimazione su base polinomiale. Anche da queste tabelle possiamo trovare conferma del fatto che le parole più lunghe hanno una percentuale maggiore di riconoscimento come ad esempio (7.6) *interrompi*, (7.10) affermativo, (7.12) *precedente* e *disattivazione*. Le parole invece che hanno una minore lunghezza ed iniziano per una consonante fricativa hanno bassa percentuale di riconoscimento perché questo particolare tipo di suono può essere confuso con rumore bianco. Questo è evidente nella parola *fisio*, fig. (7.7) e nella parola *sei*, fig. (7.9).

Dai test effettuati possiamo dunque concludere che nel caso nominale le percentuali di riconoscimento sono buone ad esclusione di alcune parole critiche. È quindi opportuno studiare le caratteristiche delle parole per poter eventualmente trovare dei sinonimi per i comandi critici e costruire dei menù più robusti possibile. I risultati ottenuti hanno validato la scelta del numero di feature estratte effettuata sulla base di un compromesso fra efficienza computazionale e buone prestazioni. Nel caso di rumore aggiunto l'attenzione si è concentrata sulla particolare applicazione del riconoscitore, nello specifico quindi per il carro armato. Per un livello di rumore tale da non creare effetto Lombard è stata riscontrata una buona percentuale di riconoscimento che ovviamente peggiora rispetto al caso nominale mettendo a confronto uno stesso menù, come possiamo notare dalle tabelle 7.2 e 7.4 ma si mantiene comunque sempre sopra al 63%.

È da sottolineare ancora che le prestazioni sono strettamente correlate al database vocale utilizzato per effettuare il training del sistema. I parlatori impiegati sono stati 22 per mancanza di disponibilità, si suppone dunque che le prestazioni possano essere di gran lunga migliorate senza effettuare alcuna modifica al riconoscitore ma solamente ampliando il database vocale. Come uteriore miglioramento del siste-

ma si potrebbero effettuare studi su diversi tipi di rumore al fine di renderlo robusto per vari ambienti. Uno studio futuro sarà sicuramente rivolto ad un'analisi dello stress poiché, come detto in precedenza, la distorsione che lo stress genera sulle parole pronunciate è di solito tale da mandare in confusione il riconoscitore vocale. Sarà quindi opportuno approfondire le principali tecniche di compensazione dello stress come ad esempio metodi basati su un particolare addestramento del sistema come il multi-stile. Questa tecnica richiede che i parlatori debbano produrre segnali vocali in condizioni di stress simulato e inserire questi stili nel processo di addestramento.
Tenendo conto delle caratteristiche del classificatore polinomiale è possibile considerare la possibilità di implementare una procedura automatizzata per la sostituzione di parole critiche, con i relativi sinonimi in un menù a cardinalità limitata.
Ua altra strategia è quella di selezionare i sottomenù più critici, effettuare un analisi spettrale più accurata delle parole che lo compongono e quindi ricavare delle feature ad hoc che rendono le parole considerate più immuni allo stress.
In conclusione, come era lecito attendersi, il riconoscitore può ancora essere migliorato ma già sono state riscontrate buone prestazioni considerando che il sistema è del tutto innovativo e che si differenzia nettamente dai software in commercio per il suo intento di lavorare in situazioni estreme.

Bibliografia

[1] http://www.andreaelectronics.com/.

[2] D. Tack L. Thompson. Voice controls and displays for the dismounted soldiers. *Defence Research and Development Canada*, October 2005.

[3] S. Sridharan D. Thambiratnam. Improving speech recognition accuracy for small vocabulary applications in adverse environments. *International Journal of Speech Technology 3*, 2000.

[4] W.M. Campbell C.C. Broun. Force xxi land warrior: A system approach to speech recognition. *IEEE*, 2001.

[5] www.speech.cs.cmu.edu/comp.speech/Section1/Data/noisex.html.

[6] G. Bristow. *Electronic Speech Recognition.* Collins Professional and Technical Books, 1986.

[7] L. Reed. *The Requirements and Applications of Speech Recognition Technology for Voice Activated Command and Control in the Tactical Military Environment.* Command and Control Research and Technology Symposium, 2004.

[8] B.H. Juang D. Mansour. A family of distorsion measures based upon projection operation for robust speech recognition. *IEEE Trans. Acoust. Speech Signal Process*, 1989.

[9] K.K. Paliwal. The short-time modified coherence representation and its application for noisy speech recognition. *IEEE Trans. Acoust. Speech Signal Process*, 1990.

[10] B.H. Juang D. Mansour. Neural net classifier for robust speech recognition under noisy environments. *Proc. IEEE Internat. Conf. Acoust*, 1989.

[11] R.W. Schafer L.R. Rabiner. *Digital Processing of Speech*. Prentice-Hall, 1978.

[12] RAVI P. RAMACHANDRAN RICHARD J. MAMMONE, XIAOYU ZHANG. Robust speaker recognition. *IEEE Signal Processing Magazine*, September 1997.

[13] S. Parthasarathyt Qi Li and Aaron E. Rosenbergt. A fast algorithm for stochastic matching with application to robust speaker verification. *IEEE Transactions on Communications*, 1997.

[14] A. E. Rosenberg J. G. Wilpon L. Lamel, L. R. Rabiner. An improved endpoint detection for isolated word recognition. *IEEE Signal Processing Magazine*, August 1981.

[15] J. Picone. Signal modeling techniques in speech recognition. 6, June 1993.

[16] K. T. Assalehb W. M. Campbella and C. C. Brouna. Audio sensors and low-complexity recognition technologies for soldier systems. *IEEE Transactions on Communications*.