Progettazione e realizzazione di un algoritmo di Voice Activity Detection

Giacomo Camposampiero, matricola 1187180

13 giugno 2021

1 Introduzione

Gli algoritmi di *Voice Activity Detection* (VAD) sono algoritmi sviluppati allo scopo di rilevare la presenza o l'assenza di voce umana all'interno di segnali audio. Questi algoritmi trovano applicazione in una vasta gamma di sistemi per l'elaborazione del suono e per la comunicazione audio real-time. In questi ultimi in particolare i sistemi VAD si rivelano molto efficaci nella riduzione dell'informazione media trasmessa tra utenti, essendo i silenzi una componente non marginale nella maggior parte delle conversazioni umane.

La letteratura scientifica è al giorno d'oggi ricca di metodi, proposti dalla comunità scientifica nel corso degli ultimi anni, per l'implementazione di un Voice Activity Detector. In questo lavoro ne sono stati selezioni alcuni allo scopo di implementare un classificatore per l'identificazione del parlato in un segnale audio digitale mono, codificato mediante una modulazione ad impulsi codificati (PCM, dall'inglese *Pulse-Code Modulation*) e pacchettizzati dal trasmettitore in pacchetti di 160 campioni audio. La classificazione è eseguita con lo scopo di decidere quali pacchetti trasmettere (perché contenenti voce, a cui si farà di seguito riferimento con la sigla *ACTIVE*) e quali invece scartare perché privi di contenuto vocale (*INACTIVE*). Gli obiettivi del classificatore comprendono la massimizzazione della compressione del segnale, la minimizzazione del clipping nel segnale vocale e il mantenimento del ritardo sotto una soglia massima di 50ms.

2 Descrizione del metodo

Come già accennato in precedenza, l'approccio alla risoluzione della consegna che si è voluto adottare è stata la creazione di un *ensemble* di classificatori. Essendo i metodi utilizzati poco elaborati, raggiungere elevate prestazioni mediante l'uso di uno solo di essi è stato infatti verificato essere molto meno efficace rispetto al loro utilizzo parallelo. I metodi che si è deciso di includere sfruttano due tipi di analisi del segnale:

- **analisi nel dominio del tempo**, che deriva la presenza di voce dalle ampiezze del suono
- **analisi nel dominio della frequenza**, che deriva la presenza di voce a partire dallo studio in frequenza del segnale

In Figura (1) è riportato il diagramma di flusso del metodo proposto. Il metodo proposto è stato implementato in Matlab. Ognuno dei singoli approcci inclusi nell'ensemble viene trattato più nel dettaglio nelle sezioni successive.

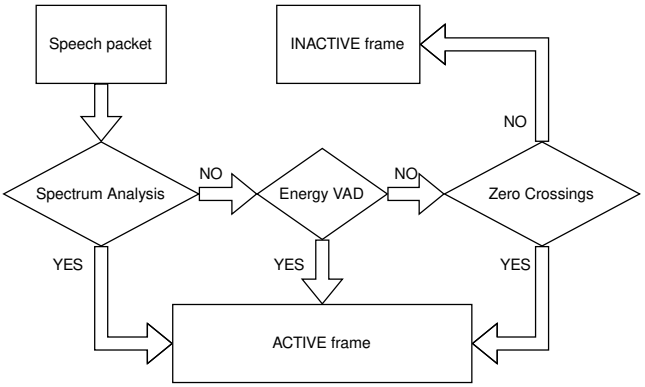


Figura 1: Flow chart dell'implementazione di VAD proposta.

2.1 Linear energy-based detector

Il *Linear Energy-based detector* è stato il primo tipo di classificatore implementato in questo progetto. Si tratta di un classificare basato sull'analisi nel dominio del tempo, che classifica un frame in base alla sua energia. L'energia di un frame può essere calcolata come

$$E_{frame} = \frac{1}{n} \sum_{i=1}^{n} x^2(i)$$

dove n è il numero di campioni del frame e $x(i)$ il campione i-esimo del frame. La decisione sul tipo di frame viene effettuata secondo la regola

```
IF (E_frame > k·E_s) ACTIVE
ELSE INACTIVE
```

dove E_s equivale all'energia media dei frame INACTIVE e k è un iperparametro del metodo di classificazione.

Una stima iniziale dell'energia media dei frame INACTIVE è fatta lavorando nella verosimile ipotesi che i primi 10 frame (corrispondenti a 200ms) della trasmissione non abbiano contenuto vocale. Il valore di questa soglia viene poi aggiornato ad ogni nuovo frame classificato come INACTIVE secondo la relazione

$$E_s = E_s \cdot (1 - p) + E_{frame} \cdot p \tag{1}$$

dove E_{frame} è l'energia del frame appena classificato e p un dato valore di probabilità che regola la velocità di aggiornamento del valore di soglia.

Il principale vantaggio di questo algoritmo risiede nella sua facilità di implementazione. Questa tecnica risulta tuttavia essere inefficiente nel caso di rumore di fondo altamente variabile o di SNR (*Signal to Noise Ratio*) basso.

2.2 Zero crossings detector

Un secondo metodo di classificazione implementato al fine di migliorare le prestazioni del precedente, soprattutto nel caso di segnali a bassa energia, è basato sul numero di *zero crossing*. Tale valore è pari al numero di volte in cui il segnale interseca l'asse delle ascisse in un dato intervallo temporale. Il numero di crossing di un frame ACTIVE è contenuto in un range fisso **[Rabiner1975AnAF]**, nel caso di un frammento ACTIVE di 10ms ad esempio pari a $[5, 15]$. Nel caso di frame INACTIVE, al contrario, il numero di zero crossings è totalmente casuale. Per sua stessa definizione, anche questo metodo esegue un'analisi del segnale nel dominio del tempo.

Questo ci permette quindi di definire una regola di decisione completamente indipendente dall'energia del frame in analisi e che, di conseguenza, è in grado di identificare frammenti di parlato a bassa energia con maggior facilità. La regola sulla base di cui è definito il metodo equivale a

```
IF (N_zc ∈ R ) ACTIVE
ELSE INACTIVE
```

dove N_{zc} equivale al numero di zero crossing in un dato pacchetto e R è l'intervallo in cui indicativamente dovrebbe ricadere lo stesso per essere classificato come voce.

Anche in questo caso uno dei principali vantaggi dell'approccio consiste nella facilità di implementazione. Di contro, l'approccio risulta essere in alcuni casi impreciso a causa della forte casualità nel numero di crossing di segnali diversi dalla voce umana.

2.3 Linear Sub-Band Power Detector

L'ultimo approccio implementato è basato invece su di un'analisi in frequenza della potenza del segnale in ingresso. In particolare, i frame che comprendono un contenuto vocale sono caratterizzati da potenze nella banda $1Hz$-$1kHz$ più grandi rispetto a frame che contengono invece silenzi o rumori di fondo. Al fine di ottenere lo spettrogramma e le relative potenze del segnale è stato utilizzato il metodo `pspectrum()` fornito da Matlab.

La decisione sul tipo di frame viene quindi effettuata secondo la regola

```
IF (P_frame > k·P_s) ACTIVE
ELSE INACTIVE
```

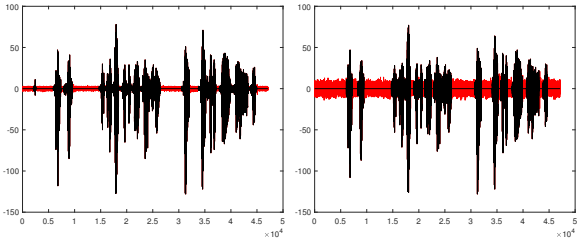
dove P_s equivale alla potenza media dei frame INACTIVE e k è un iperparametro del metodo di classificazione. Come per il classificatore tratta nella Sezione 2.1, una stima iniziale della potenza media dei frame INACTIVE è fatta lavorando nella verosimile ipotesi che i primi 10 frame della trasmissione non abbiano contenuto vocale. Il valore di questa soglia viene anche in questo caso dinamicamente aggiornato secondo una relazione simile a (1).

2.4 Finestra scorrevole

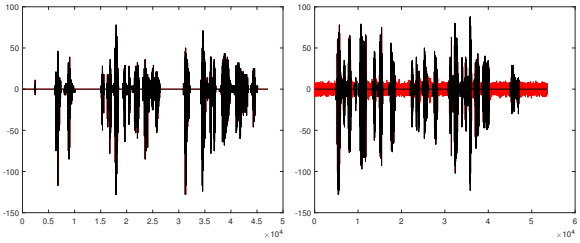
Al fine di migliorare le prestazioni dell'algoritmo, ogni frame viene classificato non solo sulla base dei campioni a lui appartenenti. É stata infatti implementata una di finestra di classificazione scorrevole, per cui la classificazione di ogni pacchetto è eseguita sulla base dei suoi campioni e di quelli del frame precedente e successivo. Questo ha permesso di ridurre sensibilmente il clipping sulla voce e rendere più fluido il risultato finale.

3 Risultati

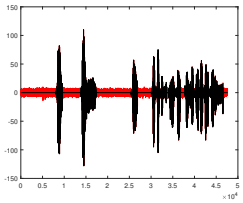
Sono di seguito riportati in Figura (2) i risultati della simulazione per i 5 file di test forniti dalla consegna. Nei grafici riportati la componente di colore nero corrisponde alle porzioni di segnale appartenenti a pacchetti classificati come ACTIVE, mentre la componente di colore rosso corrisponde alle porzioni classificate come INACTIVE.



(a) inputaudio1.data (b) inputaudio2.data



(c) inputaudio3.data (d) inputaudio4.data



(e) inputaudio5.data

Figura 2: Risultati dell'applicazione dell'algoritmo implementato alle tracce di test.

4 Conclusioni

Il metodo proposto per l'implementazione di un VAD sembra, almeno per le tracce audio fornite come esempio, eseguire una discriminazione efficace tra pacchetti ACTIVE e INACTIVE. Tuttavia, il metodo proposto rappresenta di fatto un'implementazione di fatto basilare e potrebbe, in presenza di rumori di fondo fortemente variabili o eterogenei o nel caso di SNR estremamente ridotto, risultare meno efficace nella classificazione dei pacchetti.