UNIVERSITÀ
DEGLI STUDI DI TRIESTE

dipartimento
di ingegneria
e architettura

La voce

A.Carini – Elettronica per l'audio e l'acustica

L’apparato vocale umano

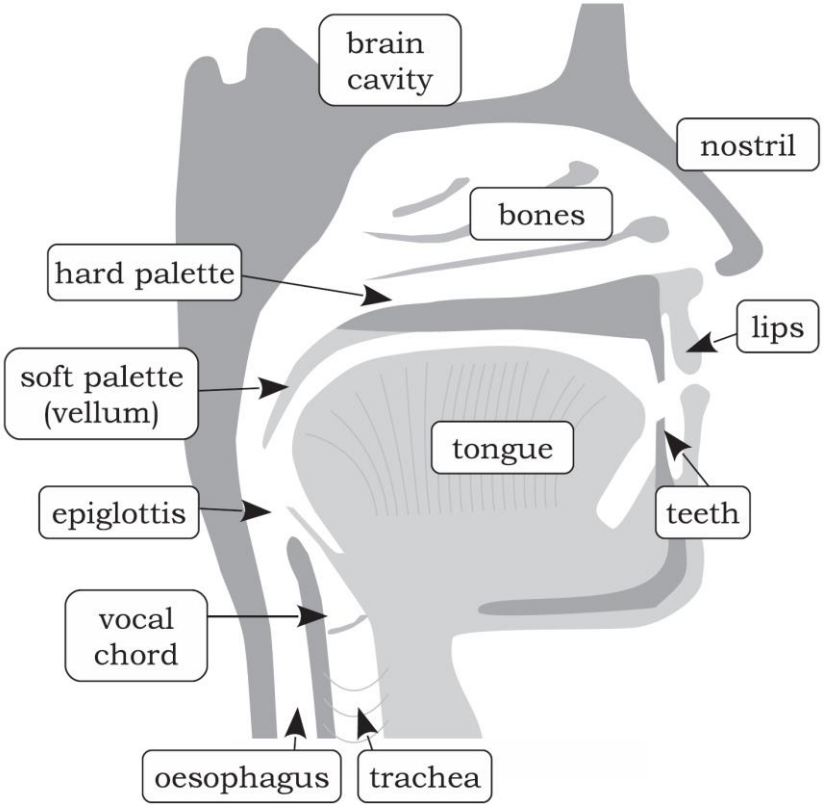
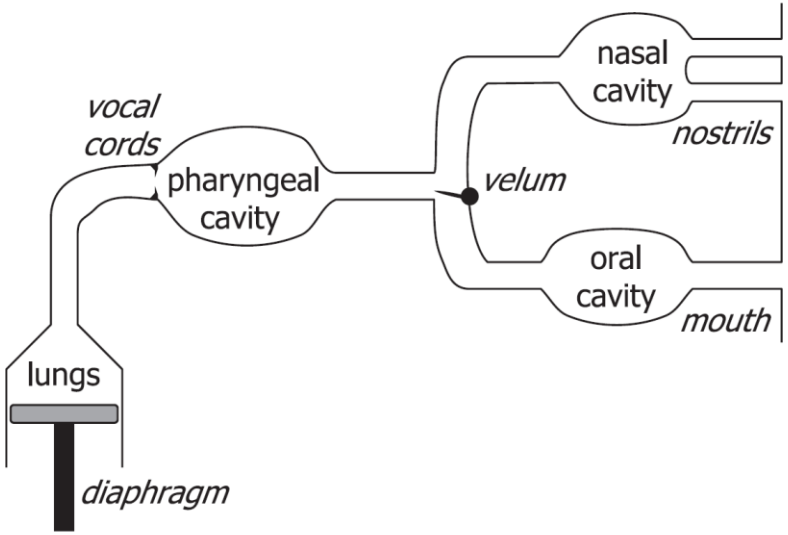


Diagramma funzionale dell'apparato vocale



Articolazioni e generazione della voce

- La potenza dei polmoni influenza il volume del suono.
- Se la glottide si chiude per un breve periodo avremo un *glottal stop,* es. /t/.
- La tensione delle corde vocali determina il *pitch*. I suoni possono essere *vocalizzati* o *non vocalizzati.*
- Un parte dell'aria può essere deviata dal *velum* verso il naso producendo suoni nasali, es. mamma.
- Se l'aria attraversa la bocca con la lingua abbassata e con l'apertura e chiusura della mandibola avremo una vocale (/a/, …) . Se la mandibola non viene chiusa avremo un *glide*, /w/ di «won».
- Se l'aria viene forzata ai lati della lingua che tocca palato o denti: /l/ e /th/ come in «la» o «determinare».
- Le *plosive* sono dei corti stop seguiti da un rilascio esplosivo, es. /d/ di dog.

Articolazioni della voce

Many phonemes are defined by their place, or method of articulation within the vocal tract. Here is a list of some of the more common terms used for vowels and consonants:

- *affricative* – a turbulent airflow fricative following an initial stop. E.g. /ch/ in ‘chip’.
- *diphthong* – a two-part sound consisting of a vowel followed by a glide. E.g. /i/ and /n/ in ‘fine’.
- *fricative* – a very turbulent airflow due to a near closure of the vocal tract. E.g. /sh/ in ‘ship’.
- *glide* – a vowel-like consonant spoken with almost unconstricted vocal tract. E.g. /y/ in ‘yacht’ (indeed many academics class /y/ as a vowel).
- *nasal* – a consonant spoken with the vellum lowered, so sound comes through the nasal cavity. E.g. /m/ in ‘man’.
- *plosive* – an explosive release of air caused by the rapid removal of a vocal tract blockage or *stop*. E.g. /p/ in ‘pop’.

Most of the consonant sounds can be either voiced or unvoiced, depending upon whether the glottis is resonating. For example /c/ in ‘cap’ is unvoiced whereas /g/ in ‘gap’ is voiced. In true whispers, all sounds, whether lexical consonant or lexical vowel, are unvoiced.

Caratteristiche della voce

- Generazione, propagazione e ricezione (udito) della voce sono ben adattati alle caratteristiche fisiche della voce stessa.
- Ci siamo adattati al mezzo fisico e agli strumenti in nostro possesso.
- Se le frequenze principali della voce fossero state di 30 kHz invece che 300 Hz, la voce si sarebbe propagata per una distanza 50 volte inferiore.
- A 300 kHz la voce sarebbe arrivata alle orecchie distorta dalla dispersione dell'aria.
- A 30 Hz saremmo stati costretti a parlare più lentamente.

- La voce può essere vista come una sequenza di elementi *vocalizzati, non vocalizzati, plosivi,* e di *silenzio.*

Caratteristiche della voce

- I **suoni vocalizzati** sono caratterizzati dalla frequenza fondamentale, il *pitch.*
  - Varia tra i 50 e 250 Hz negli uomini, tra i 120 e i 500 Hz nelle donne.
  - L'inviluppo delle armoniche crea almeno due massimi: le *formanti*
  - Voci maschili hanno fino a 3 formanti, quelle femminili fino a 5.
  - Le prime 2 formanti determinano la vocale, quelle superiori consentono di riconoscere chi sta parlando.
- I segmenti **non vocalizzati** sono dei tratti dall'aspetto di puro rumore, caratterizzati principalmente dalla forma dell'inviluppo spettrale.
- Le **plosive** contengono cambiamenti transienti molto veloci nello spettro. Sono segnali con un ampio spettro che cambia rapidamente.
- I **periodi di silenzio** sono usati per separare le parole ma anche fonemi e sono parti integranti della voce.

Formanti

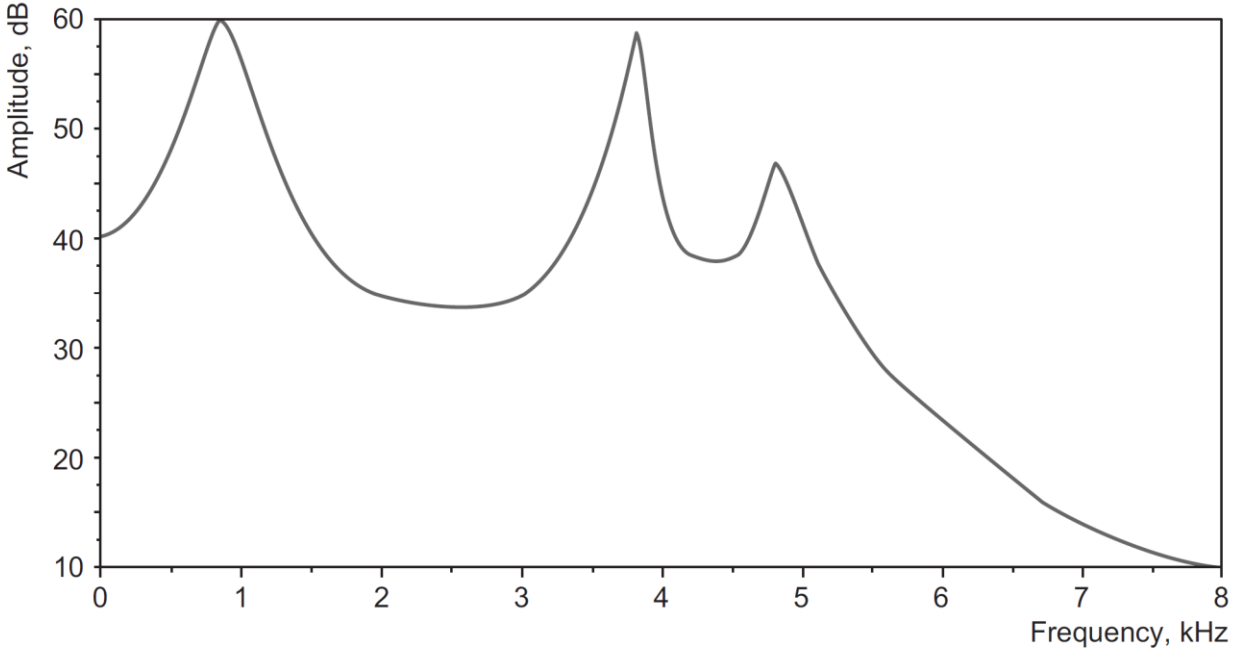
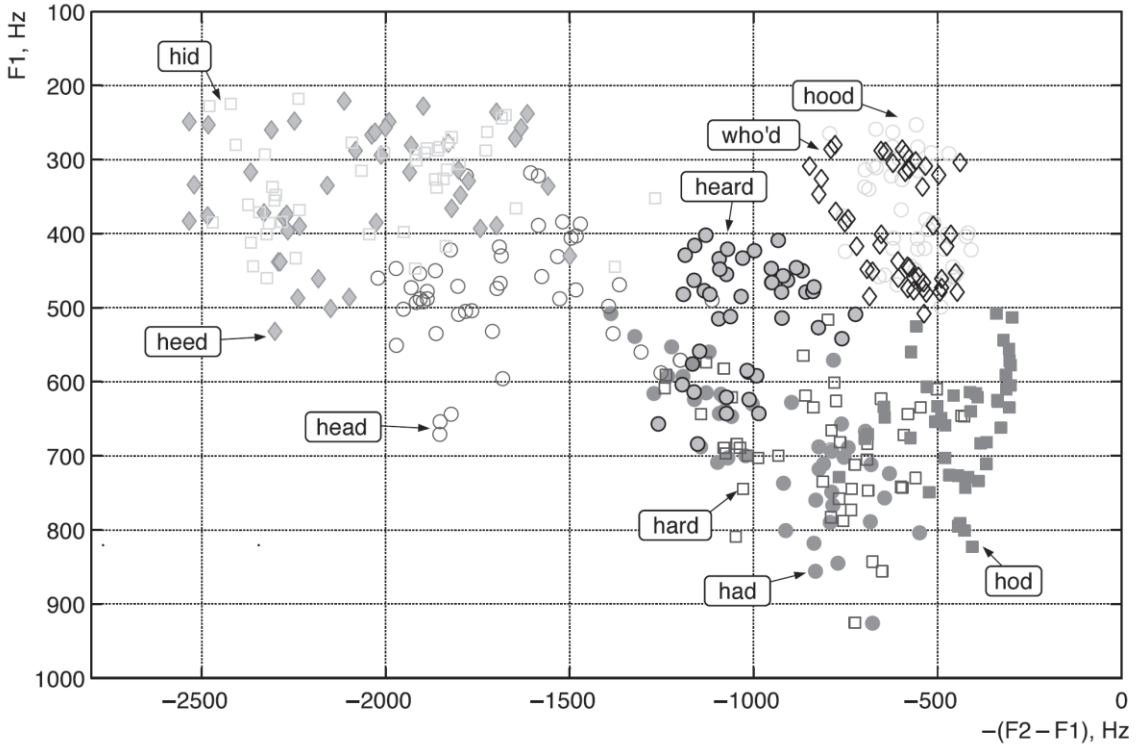


Figure 3.3 The spectral envelope of a 20 ms recording of voiced speech, showing three distinct formant peaks.

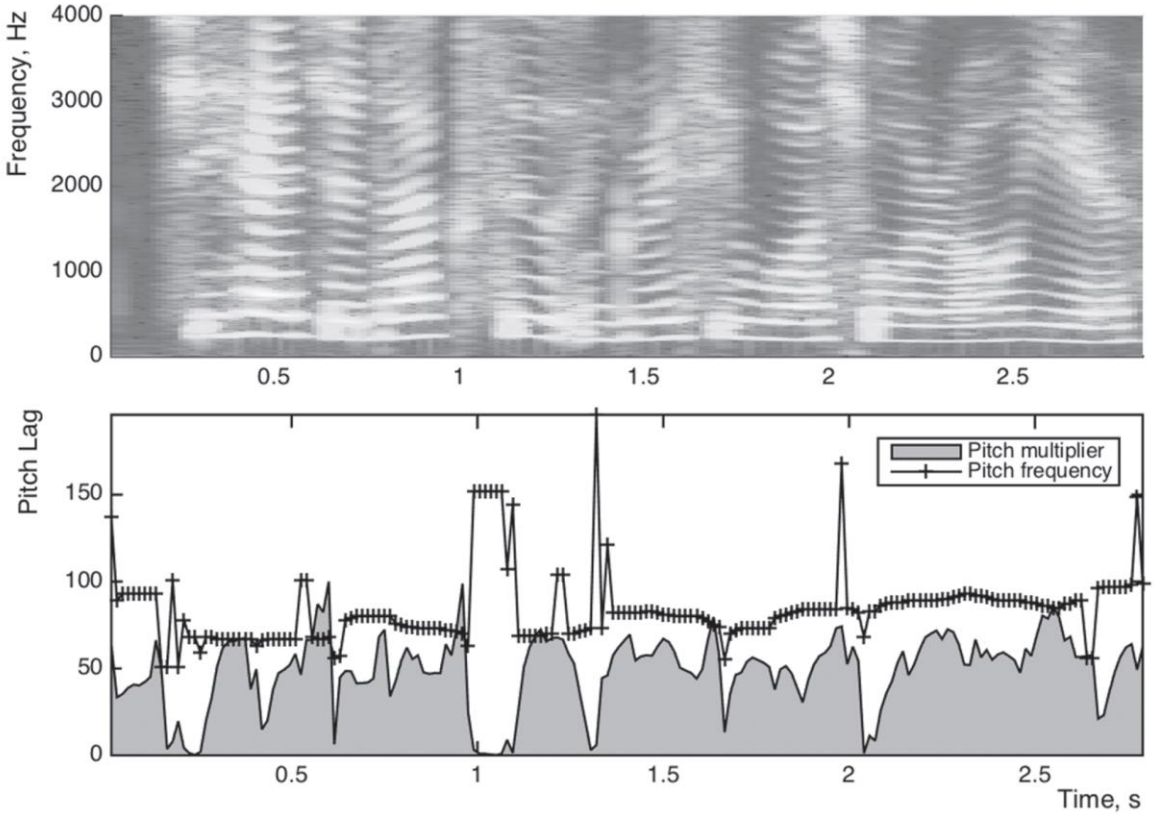
Formanti


F1, Hz
100
200
300
400
500
600
700
800
900
1000
hid
hood
who'd
heard
heed
head
hard
had
hod
−2500
−2000
−1500
−1000
−500
0
−(F2 − F1), Hz

Formanti

- La frequenza delle formanti varia nel tempo con i movimenti della bocca.
- Vengono in genere indicate con F1, F2, F3, …, mentre il pitch con f0.
- Solo le prime tre formanti contribuiscono all'intellegibilità della voce.
- F1 contiene la maggior parte dell'energia, ma sono F2 e F3 che contribuiscono di più all'intellegibilità.
- Il pitch f0 contribuisce molto poco all'intellegibilità delle lingue occidentali, diverso il discorso in quelle orientali.

Formanti


Frequency, Hz
4000
3000
2000
1000
0
0.5
1
1.5
2
2.5
Pitch Lag
150
100
50
0
Pitch multiplier
Pitch frequency
Time, s

Volume della voce

Table 3.1 Average amplitude of speech in several environments, from [17].[a]

Location	Noise level (dB_{SPL})	Speech level (dB_{SPL})
School	50	71
Home (outside, urban)	61	65
Home (outside, suburban)	48	55
Home (inside, urban)	48	57
Home (inside, suburban)	41	55
Department store	54	58
On a train	74	66
In an aircraft	79	68

[a] These data were originally published in *The Handbook of Hearing and the Effects of Noise*, K. Kryter, Chapter 1, Copyright Elsevier (Academic Press) 1994.

- Range: uomini 52-90 dB SPL, donne 50-82 dB, bambini 53-83 dB.
- Range dinamico di una conversazione 30 dB.

Volume della voce

Table 3.2 Average amplitude of phonemes by class, also showing amplitude range within each class, measured with respect to the quietest phoneme in English, the voiceless fricative /th/ in 'thought'.

Phoneme class	Example	Amplitude (range), dB
Vowel	c*a*rd	26.0 (4.9)
Glide	*l*uck	21.6 (3.2)
Nasal	*n*ight	17.1 (3.0)
Affricative	*j*ack	14.9 (2.6)
Voiced fricative	a*z*ure	11.5 (2.2)
Voiceless fricative	*sh*ip	10.0 (10.0)
Voiced plosive	*b*ap	9.6 (3.3)
Voiceless plosive	*k*ick	9.5 (3.3)

Distribuzione in frequenza e energia della voce

- La distribuzione in frequenza della voce rispecchia la sensibilità dell'orecchio.
- La maggior parte delle frequenze cadono lì dove l'orecchio è più sensibile.
- C'è però un forte disadattamento tra frequenze con maggiore energia e quelle più importanti per l'intellegibilità.
- Gran parte dell'energia cade alle basse freq., attorno ai 500Hz negli uomini, 800 Hz nelle donne, ma queste freq. non sono essenziali per l'intellegibilità.
- Le frequenze più importanti per l'intellegibilità sono superiori a 1kHz: F2, F3

Distribuzione in frequenza e energia della voce

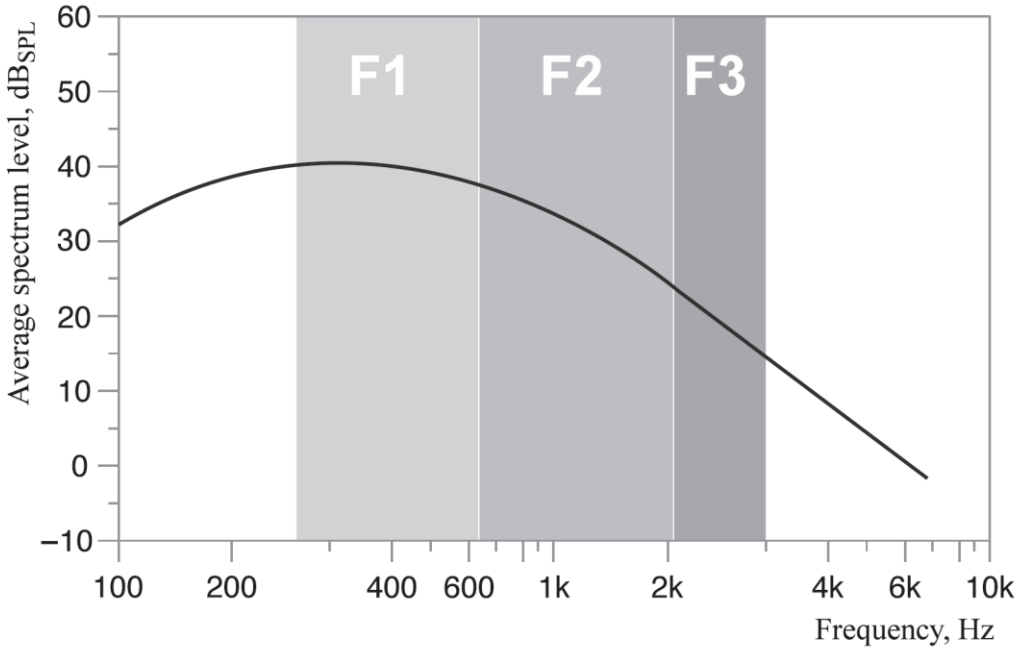


Figure 3.6 Long-time averaged speech power distribution plotted against frequency, with the approximate regions of the first three formants identified through vertical grey bands.

Distribuzione temporale

- Esiste un limite intrinseco nella velocità di articolazione di fonemi e sillabe.
- Prove sperimentali suggeriscono che per la voce normale, la velocità di articolazione è indipendente dalla rapidità del parlato.
- Quando parliamo più rapidamente, usiamo lo stesso tempo per ogni sillaba, ma riduciamo lo spazio tra sillabe e parole.
- Parlando lentamente, la durata dei fonemi è la stessa ma aumentano gli spazi.
- La massima velocità con cui si possono muovere i muscoli è indipendente dalla lingua, dal contesto, ed è abbastanza costante tra parlatori diversi.
- Il limite di questa velocità fa sì che possiamo considerare la voce **stazionaria su periodi di 20 ms**.

Intellegibilità e qualità della voce

- Importante distinguere tra le due.
- La **qualità** è una misura della fedeltà o correttezza della voce. Include quanto la voce assomiglia a un target, quanto è piacevole. E' un attributo soggettivo ma può essere approssimato con misure oggettive.
- L'**intellegibilità** è una misura di quanto la voce sia comprensibile. Si concentra sul contenuto di informazione, si riferisce al quantitativo di informazione che viene correttamente trasmesso.

Mean opinion score

Table 3.3 MOS rating scale for objective assessment of speech quality.

Score	Description	Impairment
5	Excellent	Imperceptible
4	Good	Perceptible but not annoying
3	Fair	Slightly annoying
2	Poor	Annoying
1	Bad	Very annoying

- Standardizzato dall'ITU nella raccomandazione P.800

Mean squared error - MSE

$$E = \frac{1}{N} \sum_{i=0}^{N-1} \{s[i] - p[i]\}^2.$$

```
mse=mean((s-p).^2)
```

MSE segmentale:

$$E(j) = \frac{1}{N} \sum_{i=jN}^{(j+1)N-1} \{s[i] - p[i]\}^2.$$

```
mse(j)=mean(s((j-1)*N+1:j*N)-p((j-1)*N+1:j*N).^2);
```

Signal to noise ratio - SNR

$$SNR = 10 \log_{10} \left[\frac{\sum_{i=0}^{N-1} s^2(n)}{\sum_{i=0}^{N-1} n^2(n)} \right]$$

```
snr = 10*log10( sum(s.^2)/sum(n.^2) )
```

$$SEGSNR(j) = 10 \log_{10} \left[\frac{\sum_{i=jN}^{jN-1} s^2(n)}{\sum_{i=jN}^{jN-1} n^2(n)} \right]$$

```
Ssnr(j) = 10*log10( sum(s((j-1)*N+1:j*N).^2)/
                                    sum(n((j-1)*N+1:j*N).^2) )
```

Signal to noise ratio - SNR

Con voice activity detector:

$$\Xi = 10\log_{10}\frac{\frac{1}{N_{sig}}\sum_{\substack{signal\\frames}}\sum_{k=0}^{K-1}|x^{(n)}(kT)|^2}{\frac{1}{N_{noise}}\sum_{\substack{signal\\frames}}\sum_{k=0}^{K-1}|x^{(n)}(kT)|^2}$$

Segmentale:

$$\Xi^{(n)} = \frac{\sum_{k=0}^{K-1}|X_k^{(n)}|^2}{\sigma^2}$$

Per singola frequenza:

$$\xi_k = \frac{|X_k|^2}{\sigma_k^2}$$

Spectral distortion: Log spectral distance

$$LSD = \sqrt{\frac{1}{2\pi}\int_{-\pi}^{+\pi}\left(10\ \log_{10}\frac{\left|S(e^{j\omega})\right|^2}{\left|P(e^{j\omega})\right|^2}\right)^2 d\omega}$$

(valore quadratico medio della differenza in dB tra i due spettri)

Perceptual audio quality assessment

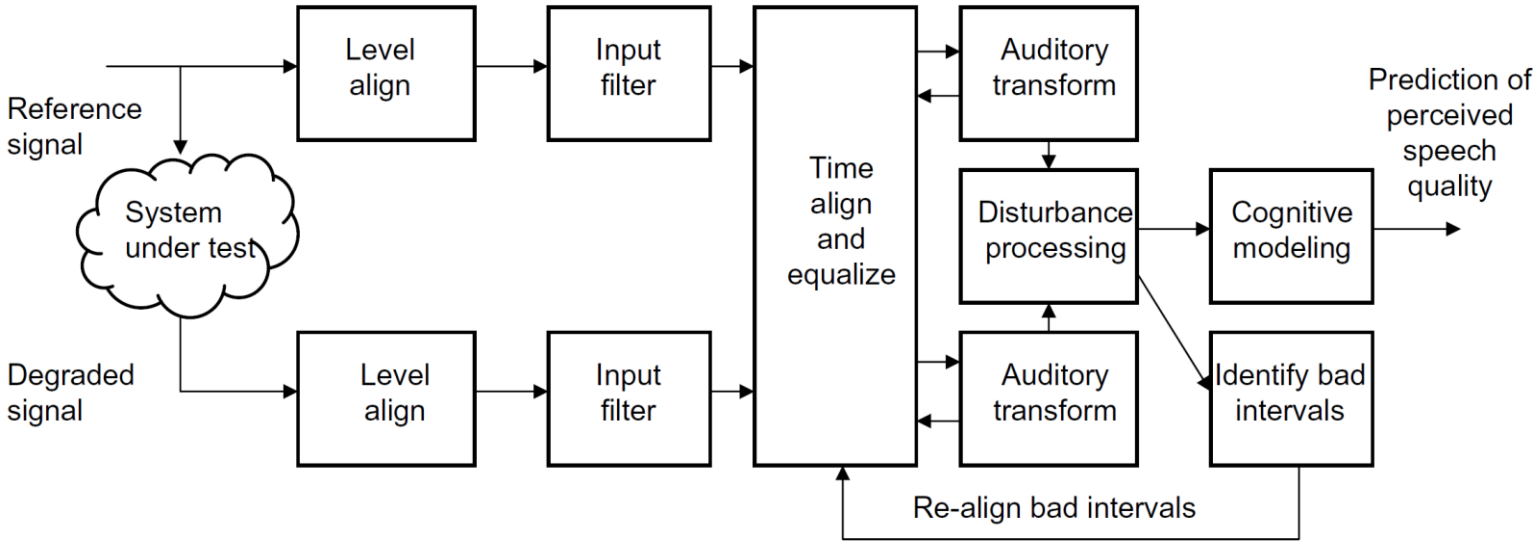


Figure 2.18 Block diagram of perceptual audio sound quality assessment algorithm

Intellegibilità e contesto

- L'intellegibilità viene aiutata dal contesto, dalla riduzione del vocabolario, da ripetizioni, ridondanze o parole più lunghe del necessario (alpha bravo foxtrot).

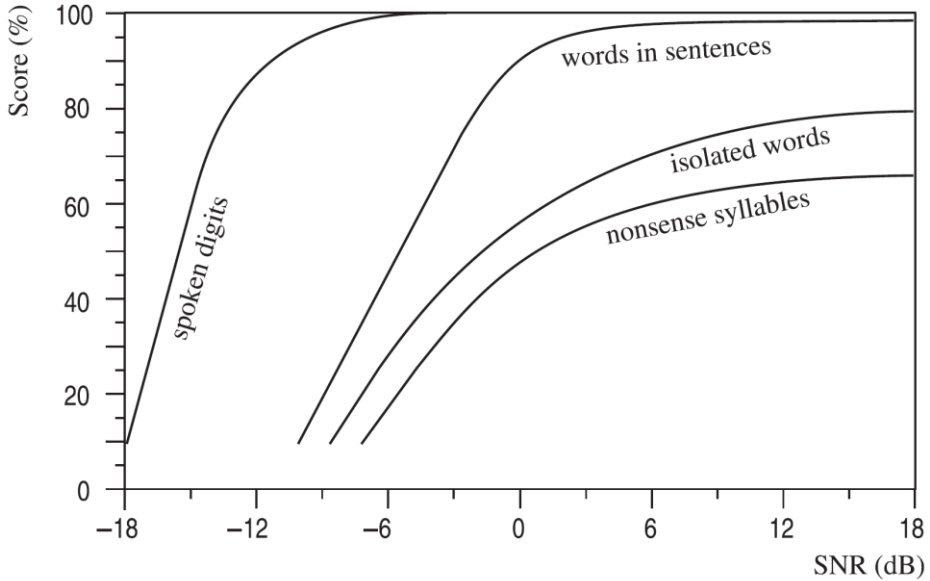


Figure 3.9 Effect of contextual information on the intelligibility of speech.

Effetto della dimensione del vocabolario

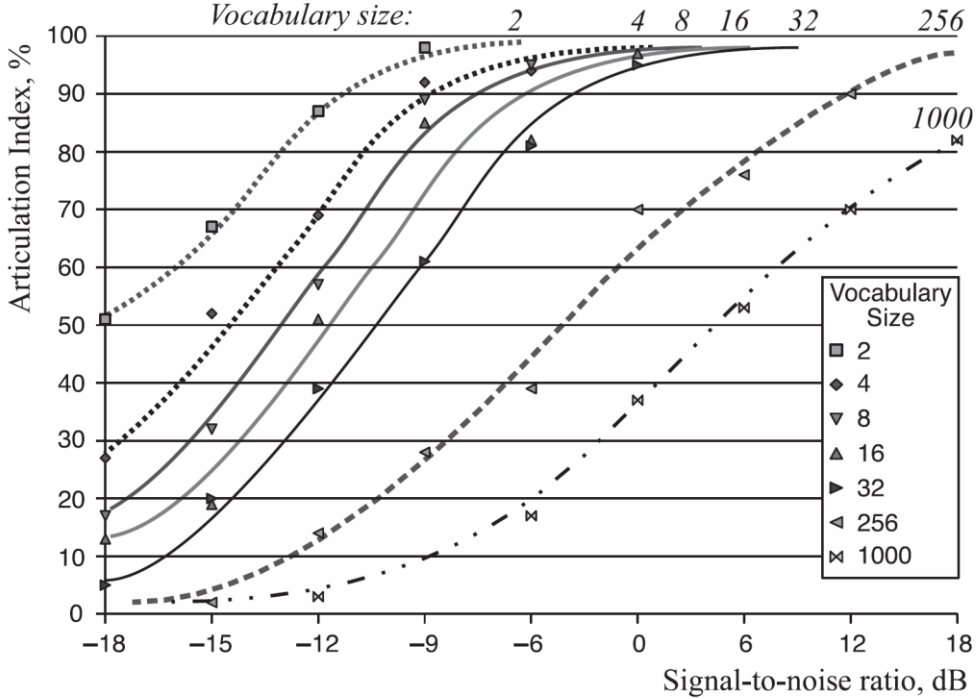


Figure 3.10 Effect of spoken vocabulary size on the intelligibility of speech.

Caratteristiche spaziali della voce

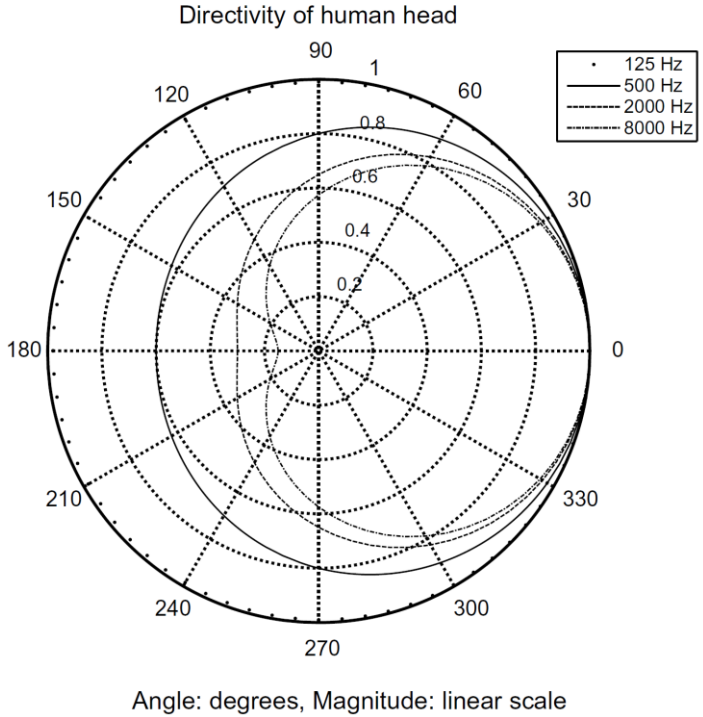


Figure 2.6 Directivity of a human head for various frequencies

Vedere:

- Ian Vince McLoughlin, “Speech and Audio Processing”- Cambridge Univesity Press (2016)
  - Cap. 3.1, 3.2
- Ivan Tashev “Sound Capture and Processing”, John Wiley & Sons, 2009
  - Cap. 2.2 (introduzione)
- Ian Vince McLoughlin, “Speech and Audio Processing”- Cambridge Univesity Press (2016)
  - Cap. 3.4.1 - 3.4.2.3
  - Cap. 3.4.3 – 3.4.4
- Ivan Tashev “Sound Capture and Processing”, John Wiley & Sons, 2009
  - Cap. 2.2.4