MUC859 Psicoacústica

Voz humana y Habla (Speech)

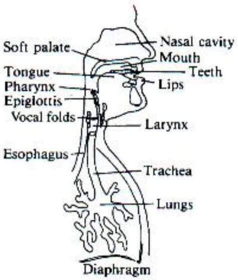
Rodrigo F. Cádiz, Ph.D.

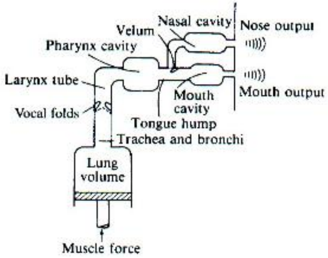
[1]Centro de Investigación en Tecnologías de Audio
Pontificia Universidad Católica de Chile

Apuntes

Voz humana

Voice Production





Voz humana

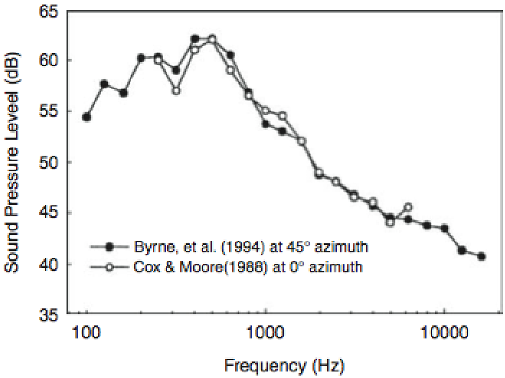
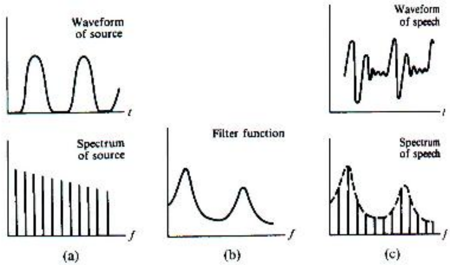


Figure 14.12 Long-term average speech spectra for speech presented at an overall level of 70 dB SPL. *Filled symbols*: composite of male and female speech samples from 12 languages at 45° azimuth (based on Byrne et al. (1994)). *Open symbols*: composite of male and female speech in English at 0° azimuth (based on Cox and Moore, 1988).

Voz humana

Formants

FIGURE 15.9
The effect of formants on sound: (a) waveform and spectrum of source sound; (b) filter function showing two formants (resonances); (c) waveform and spectrum of transmitted sound. t = time; f = frequency.



Voz humana

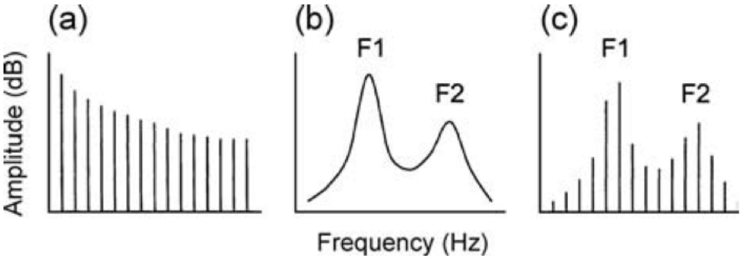


Figure 14.2 The source-filter theory (acoustic theory) of speech production: Idealized spectra showing that when the glottal source spectrum (a) is passed through the vocal tract filters (b) the resulting (output) spectrum (c) represents characteristics of the vocal tract. F1 and F2 indicate the first two formants.

Voz humana

Formants

- Formants vary by changing the position of the articulators (lips, tongue body, tongue tip, lower jaw, velum, pharyngeal sidewalls, and larynx)
- The two lowest formants can be changed in excess of two octaves.
  - They determine the identity of most vowels
- Higher formants cannot be varied as much
  - Do not contribute much to vowel quality
  - Contribute to personal voice timbre

Voz humana

Vowels

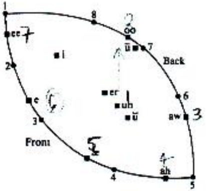


FIG. 15.8
Approximate tongue positions for articulating the vowels listed in Table 15.1. Number 1–8 are the eight cardinal vowels, which serve as a standard of comparison between languages.

TABLE 15.1 The vowels of Great American English

Pure vowels						Diphthongs		
ee	heat	/i/ 7	3 aw	call	/ɔ/	3 ou	tone	/oʊ/
i	hit	/ɪ/	ů	put	/ʊ/	ei	take	/eɪ/
e	head	/ɛ/	2 oo	cool	/u/	ai	might	/aɪ/
ae	had	/æ/ 5	1 ŭ	ton	/ʌ/	au	shout	/aʊ/
uh	the	/ə/	er	bird	/ɝ/	oi	toil	/ɔɪ/
ah	father	/ɑ/ 4				ju	fuse	/ju/

Voz humana

Vowels spectra

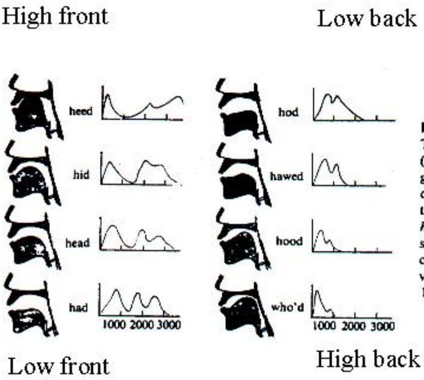


FIG. 15.16
The positions of the vocal organs (based on data from X-ray photographs of the author) and the spectra of the vowel sounds in the middle of the words *heed*, *hid*, *head*, *had*, *hod*, *hawed*, *hood*, *who'd*. Compare the sounds *hod*, *heed*, and *who'd* with the corresponding two-tube models of the vocal tract in Figs. 15.11, 15.12, and 15.13. (From P. Ladefoged 1962).

Voz humana

Vowels
Vocal Tract

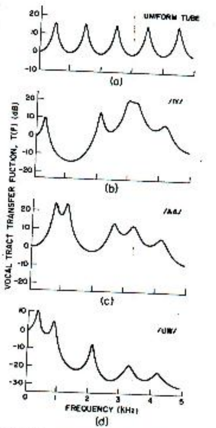
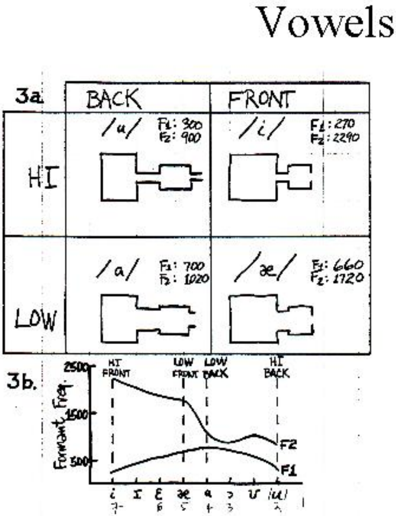


FIG. 9. The magnitude of the vocal tract transfer function is plotted for an ideal uniform vocal tract, and for the vowels [i], [a], and [u].

Dennis H. Klatt: Software for a formant synthesizer 979

Voz humana


Vowels
3a
BACK
FRONT
HI
LOW
/u/
F1: 300
F2: 900
/i/
F1: 270
F2: 2290
/a/
F1: 700
F2: 1020
/æ/
F1: 660
F2: 1720
3b.
Formant Freq.
2500
1500
500
HI FRONT
LOW FRONT
LOW BACK
HI BACK
F2
F1

Voz humana

Vowels F1, F2

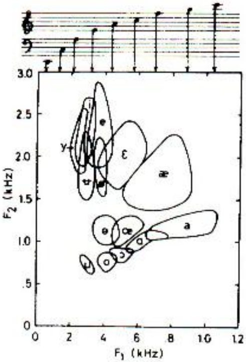
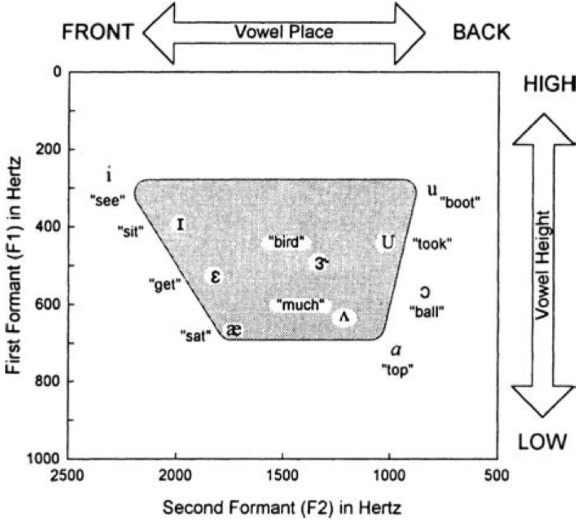


Fig. 1. Ranges of the two lowest formant frequencies for different vowels represented by their symbols in the International Phonetic Alphabet (IPA). Above, the scale of the first formant frequency is translated into musical notation.

Voz humana


FRONT
Vowel Place
BACK
HIGH
Vowel Height
LOW
0
200
400
600
800
1000
First Formant (F1) in Hertz
2500
2000
1500
1000
500
Second Formant (F2) in Hertz
i
"see"
I
"sit"
ɛ
"get"
æ
"sat"
"bird"
ɝ
"much"
ʌ
u
"boot"
U
"took"
ɔ
"ball"
ɑ
"top"

Voz humana

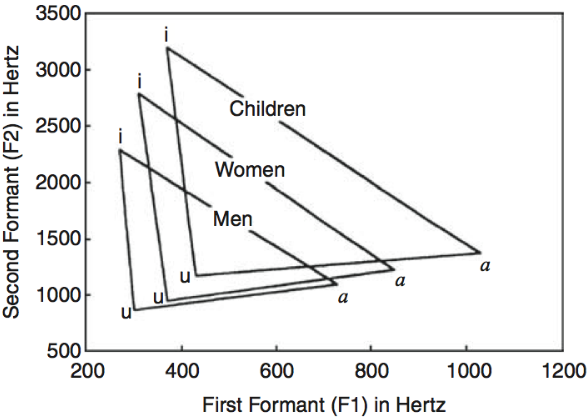


Figure 14.5 Average formant frequencies increase going from men to women to children, as illustrated here by the shifts in the average values of F1 and F2 for the vowels /i/, /a/, and /u/. *Source*: Based on data by Peterson and Barney (1952).

Voz humana

Vowels F1, F2, F3

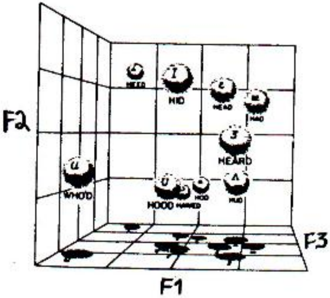


Fig. 2a. Three-dimensional spatial representation for 10 vowel phonemes (based on the data of Peterson and Barney).

Voz humana

Consonants

ABLE 15.2 The classification of English consonants

	Manner of articulation						
	Plosive		Fricative		Nasal	Semivowel	Liquids
ace of articulation	Unvoiced	Voiced	Unvoiced	Voiced			
bial (lips)	p	b			m	w	
biodental (lips and teeth)			f	v			
ental (teeth)			th /θ/ (thin)	th /ð/ (then)			
lveolar (gums)	t	d	s	z	n	y /j/	l, r
latal (palate)			sh /ʃ/	zh /ʒ/			
lar (soft palate)	k	g			ng /ŋ/		
ottal (glottis)			h				

etic symbols are given where they differ from the English letter.

Voz humana

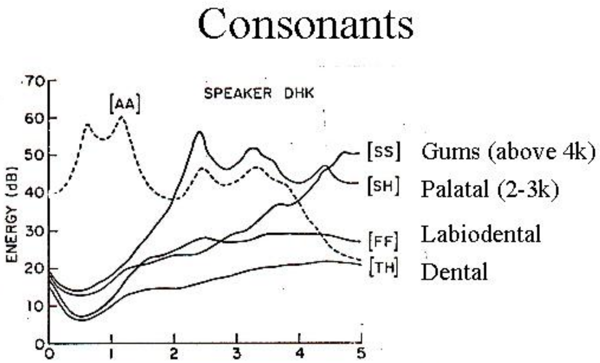


Figure 9-7. Average linear prediction spectra of the fricatives [FF], [TH], [SS], [SH]. A typical spectrum of the vowel [AA] is shown as the dashed curve to indicate the relative intensities of the fricatives.

Voz humana

Consonants

Lips

Gums

Velar

FIG. 15.17
Profiles of the vocal tract showing place of articulation of the stop or plosive consonants.

Voz humana

Consonants - Formant transitions

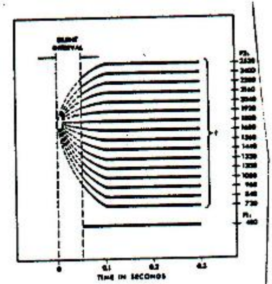


FIG. 16.7
Second-formant transitions perceived as the same plosive consonant "t." (After Delattre, Liberman, and Cooper, 1955).

Voz humana

Consonants - Formant transitions

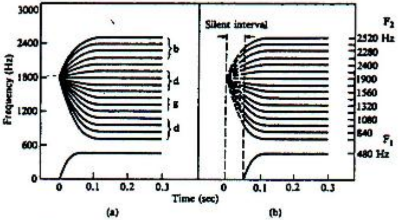
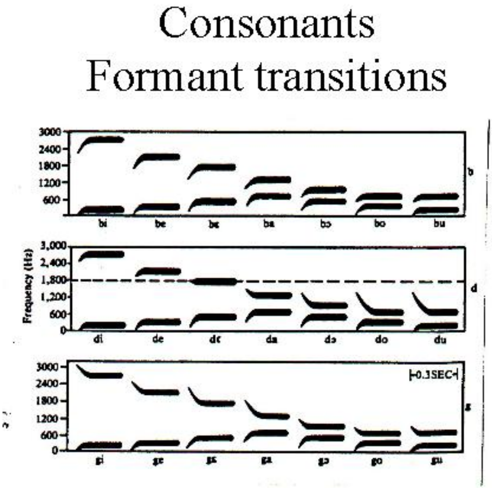


FIG. 16.9
(a) Second-formant transitions that start at the /d/ locus. (b) Comparable transitions that merely "point" at it, as indicated by the dotted lines. Those of (a) produce syllables beginning with /b/, /d/, or /g/, depending on the frequency level of the formant; those of (b) produce only syllables beginning with /d/. (From Delattre, Liberman, and Cooper, 1955).

Voz humana


Consonants
Formant transitions
3000
2400
1800
1200
600
bi
be
ba
b
Frequency (Hz)
3,000
2,400
1,800
1,200
600
0
di
de
da
do
du
d
0.3SEC
go
g

Voz humana

Formant transitions

A. M. Liberman and M. Studdert-Kennedy: Phonetic Perception

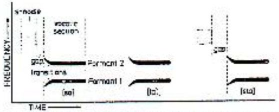


Fig. 3. Schematic spectrograms illustrating the importance of silence for the perception of a stop consonant: [sa] becomes [ta] when the noise is removed, or [sta] when a silent interval of appropriate length is introduced between the noise and the rest of the syllable

Phonetic Interpretation of the Sounds of Speech

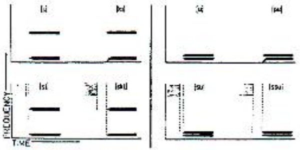


Fig. 4. Spectrographic patterns that illustrate how two very different acoustic cues — a transition of the first formant (top half) and an appropriate interval of silence (bottom half) — are phonetically equivalent in the perception of stop consonants

Voz humana

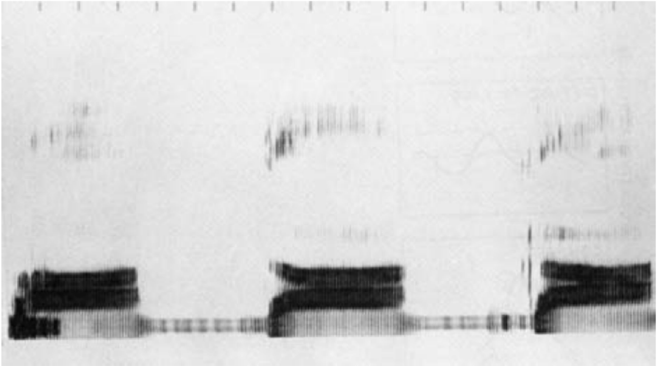


Figure 14.6 Spectrograms of /ba/, /da/, and / ga/ (left to right). Note second formant transitions.

Voz humana

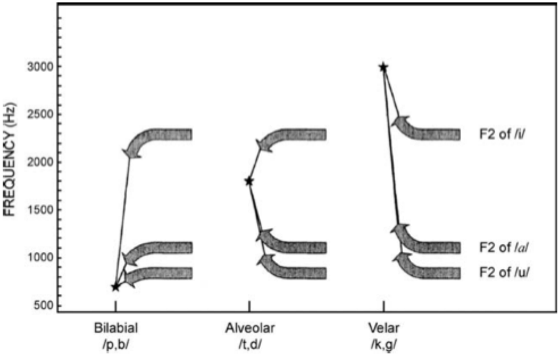


Figure 14.7 Artist's conceptualization of the second formant transition locus principle for stop consonant place of articulation. Stars indicate the locus (target frequency) toward which the second formant transitions point for bilabials, alveolars, and velars.

Voz humana

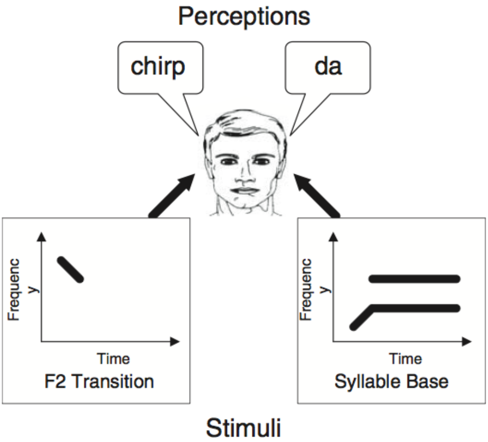


Figure 14.11 Idealized illustration of duplex perception. Presenting a syllable base without the second formant transition to one ear and just the second formant transition to other ear, causes the perception of both a speech sound (syllable /da/) in one ear and a nonspeech chirp-like sound in the other ear.

Voz humana

Consonants Clustering

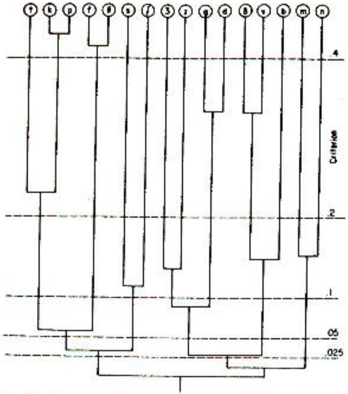


Fig. 2b. Hierarchical clustering representation for 16 consonant phonemes (based, again, on the pooled data from Miller and Niceley's six "flat" conditions).

Voz humana

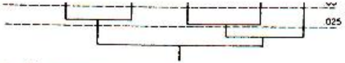


Fig. 2b. Hierarchical clustering representation for 16 consonant phonemes (based, again, on the pooled data from Miller and Niceley's six "flat" conditions).

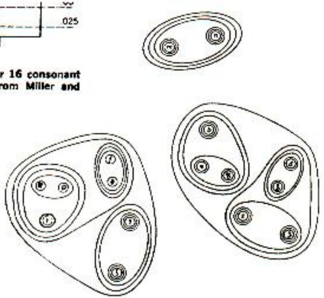


Fig. 2c. Combined spatial and hierarchical representation (in which the hierarchical clusters of Fig. 4.5 are embedded into the spatial configuration of Fig. 4.1).

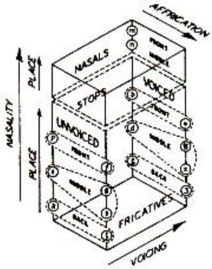


Fig. 2d. Representation of the 16 consonants in terms of five distinctive features.

Voz humana

Singing

- Formant frequencies are modified
  - The two lowest may be modified as much as to fall in the category of another vowel
- Pitch range changes dramatically
  - Speech fundamentals
    - 110-200Hz up to 350Hz
  - Singing (highest pitches)
    - Soprano 1400Hz (C6)
    - Alto 700 Hz (F5)
    - Tenor 523 Hz (C5)
    - Baritone 390 Hz (G4)
    - Bass 350Hz (F4)

Voz humana

Spectral differences between speech and singing

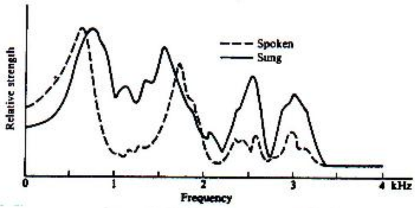


FIG. 17.3
Spectra of vowel sound |ae| as spoken and sung by a professional singer.

Voz humana

High -low Larynx

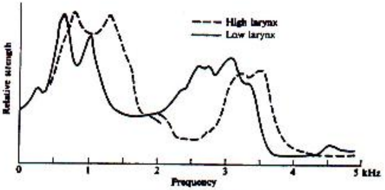


FIG. 17.6
Spectrum of vowel sound /ɑ/ sung with a high and a low larynx.

Voz humana

Formants difference
Speech and singing

- Singing
  - Wide pharynx
  - Lower larynx
- Less variability in timbre important, legato

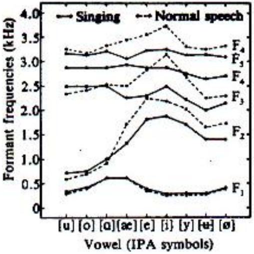


FIG. 17.4
Formant frequencies of long Swedish vowels in normal male speech (dashed lines) and in professional male singing (solid lines). (From Sundberg, 1974).

Voz humana

Singing

- Fundamental appears higher than first formant
- Singers move the frequency of first formant to be close to the fundamental
  - Usually by changing the jaw opening
- Amplitude of fundamental is therefore increased

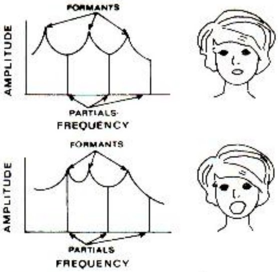
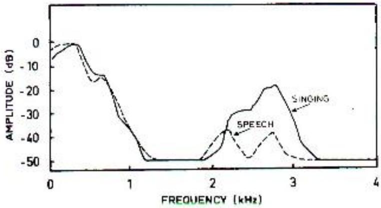


Fig. 2. Schematical illustration of the formant strategy in female singing at high pitches. In the upper case the singer has a small jaw opening. The first formant appears at a frequency far below the frequency of the lowest partial of the vowel spectrum. The result is a low amplitude of that partial. In the lower case the jaw opening is widened so that the first formant matches the frequency of the fundamental. The result is a considerable gain in amplitude of that partial (reprinted from Sundberg, 1977b).

Voz humana

Altos, Tenors, Baritones, Basses



- Partials in the frequency region 2.5-3kHz are higher in singing that in speech
  - "Singers formant"
  - Clustering of third, fourth, and fifth formant frequencies

Voz humana

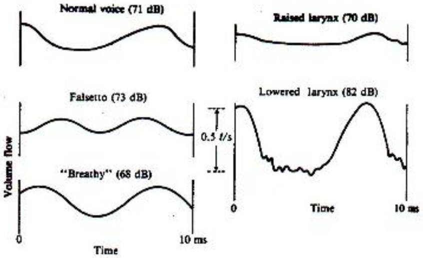


FIG. 17.15
Waveforms of glottal air flow during various modes of singing. (After Sundberg, 1978).

Voz humana

Register

- Female Chest register
  - Large et.al. (1970)
    - Chest tones consume more air than middle register and its more efficient
  - Large (1974)
    - Chest register vowels possessed strongre higher partials than mid register vowels.
    - Differences in registers may be accounted for by differences in vocal fold vibrations
  - Sundberg (1977)
    - Found formant frequency differences between registers.

Voz humana

Chest and Head

- Two sets of muscles involved
  - Cricothyroids
  - Thyroartyneoids
- In head register (falsetto) more air is consumed, less efficient

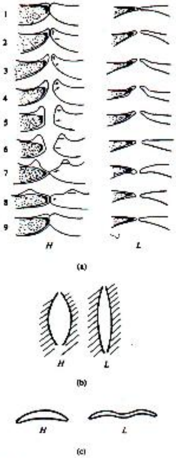


FIG. 17.14
Schematic diagrams of vocal cords vibrating: (a) side view (from Titze, 1973); (b) top view; (c) edge view. In each diagram H denotes the heavy mechanism (chest voice) and L the light mechanism (head voice).

Voz humana

Strengths of harmonics in different registers

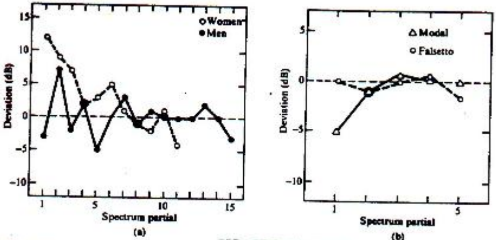


FIG. 17.18
(a) Relative strengths of harmonics in male and female voices. (b) Relative strengths of harmonics in a male voice in the modal and falsetto registers. In both cases the vertical axis shows the deviation from the overall decrease of 12 dB/octave that characterizes voice source. (From Sundberg, 1987).

Voz humana

Choral singing

- Singing solo, singers tend to produce more energy in the range 2-4kHz

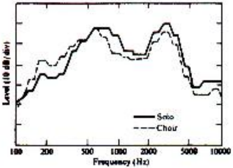


FIG. 17.19
Average spectrum envelopes for a male singer who sang a phrase as a solo singer and as a choral singer. In the latter case his lowest partials are somewhat stronger and his singer's formant is slightly weaker. (From Rossing et al., 1986).

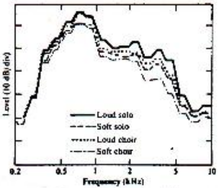


FIG. 17.20
Average spectrum envelopes for a female singer who sang the same phrase at two different sound levels as a solo singer and as a choral singer. Choral singing gives slightly weaker high partials. (From Rossing et al., 1987).

Voz humana

Amplitude - Audibility

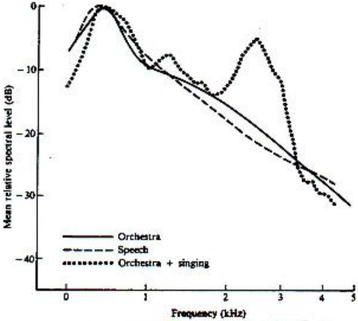


FIG. 17.7
Idealized average spectra of normal speech and orchestra music. The dotted curve shows the average spectrum of Jussi Björling singing with a loud orchestra accompaniment. (From Sundberg, 1977a).

- Vowels will tend to be masked if first formant/fundamental is below 500Hz (close to B4)
- Exception: [ɑ, a, æ]

Voz humana

Vowel Intelligibility

- Morozov (1965)
  - Based on fundamental frequency
    - Dropped below 80% above E4 (330Hz) for males, B4 (495Hz) for females
    - For C5 (523Hz), males, dropped to 50%
    - For C6 (1046Hz), females, dropped to 50%
  - At highest pitches in female singing, all vowels tend to be perceived as [a]
    - Maximum jaw opening
- Two factors that may explain
  - Deviation of formant frequency patterns by singers
  - At high pitches, few partials are present with information to correctly perceive vowel quality
- Other conclusions from other studes
  - Raised larynx (shorter vocal tract) produced more intelligible vowels

Voz humana

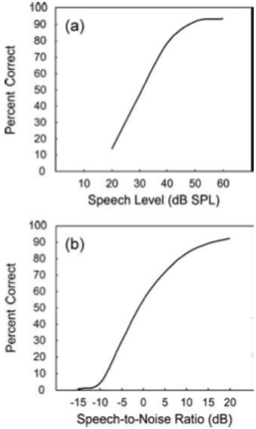


Figure 14.13 Speech recognition performance for single-syllable words improves with increasing (a) speech level and (b) speech-to-noise ratio. *Source*: Based on Gelfand (1998), used with permission.

Voz humana

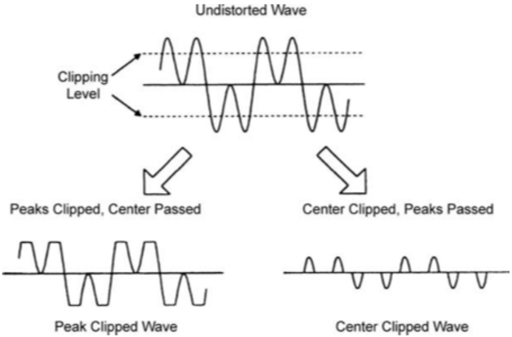


Figure 14.15 Effects of peak-clipping and center-clipping on the waveform. "Clipping level" indicates the amplitude above (or below) which clipping occurs.

Voz humana

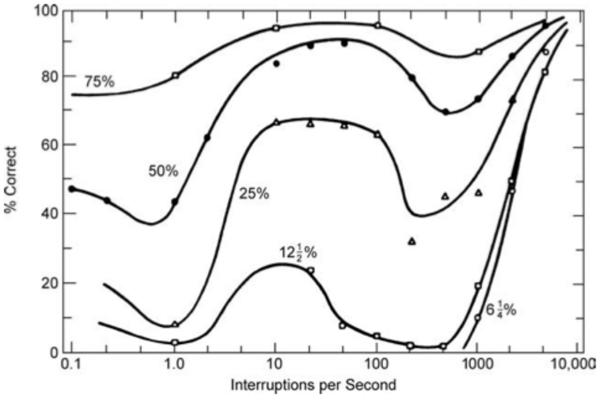


Figure 14.16 Discrimination as a function of interruption rate with speech–time fraction as the parameter. *Source:* Adapted from Miller and Licklider (1950), with permission of *J. Acoust. Soc. Am.*

Voz humana

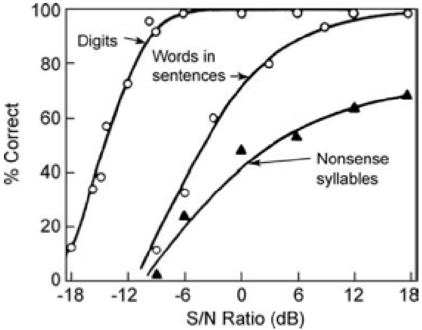


Figure 14.17 Psychometric functions showing the effects of test materials. *Source*: From Miller, Heise, and Lichten (1951), with permission of *J. Exp. Psychol.*

Voz humana

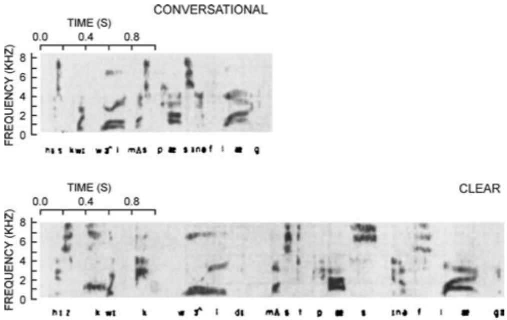


Figure 14.20 Spectrographic examples of conversational speech (above) and clear speech (below) produced by the same talker. *Source*: From Picheny, Durlach, and Braida (1986), with permission of the American Speech-Language-Hearing Association.