IT8006 / PRINCIPLES OF SPEECH PROCESSING

UNIT I SPEECH SIGNAL CHARACTERISTICS & ANALYSIS

- **UNIT I SPEECH SIGNAL CHARACTERISTICS & ANALYSIS 11**

  Speech production process - speech sounds and features- - Phonetic Representation of Speech -- representing= speech in time and frequency domains - Short-Time Analysis of Speech – Short Time Energy and Zero-Crossing Rate - Short-Time Autocorrelation Function - Short-Time Fourier Transform (STFT) - Speech Spectrum - Cepstrum - Mel-Frequency Cepstrum Coefficients - Hearing and Auditory Perception - Perception of Loudness - Critical Bands - Pitch Perception

- **UNIT II SPEECH COMPRESSION 12**

  Sampling and Quantization of Speech (PCM) - Adaptive differential PCM - Delta Modulation - Vector Quantization- Linear predictive coding (LPC) - Code excited Linear predictive Coding (CELP)

- **UNIT III SPEECH RECOGNITION 12**

  LPC for speech recognition- Hidden Markov Model

- **UNIT IV SPEAKER RECOGNITION 5**

  Acoustic parameters for speaker verification-Feature space for speaker recognition-similarity measures- Text dependent speaker verification-Text independent speaker verification techniques
- **UNIT V SPEAKER RECOGNITION AND TEXT TO SPEECH SYNTHESIS 5**

  Text to speech synthesis(TTS)- Concatenative and waveform synthesis methods, sub-word units for TTS, intelligibility and naturalness-role of prosody

- TEXT BOOKS:
  - 1. L. R. Rabiner and R. W. Schafer, Introduction to Digital Signal Processing, Foundations and Trendsin Signal Processing Vol. 1, Nos. 1–2 (2007) 1–194
  - 2. Ben Gold and Nelson Morgan —Speech and Audio signal processing- processing and perception of speech and music‖, John Wiley and sons 2006
- REFERENCES
  - 1. Lawrence Rabiner, Biiing and– Hwang Juang and B.Yegnanarayana —Fundamentals of Speech Recognition‖, Pearson Education, 2009
  - 2. Claudio Becchetti and Lucio Prina Ricotti, —Speech Recognition‖, John Wiley and Sons, 1999
  - 3. Donglos O shanhnessy —Speech Communication: Human and Machine —, 2nd

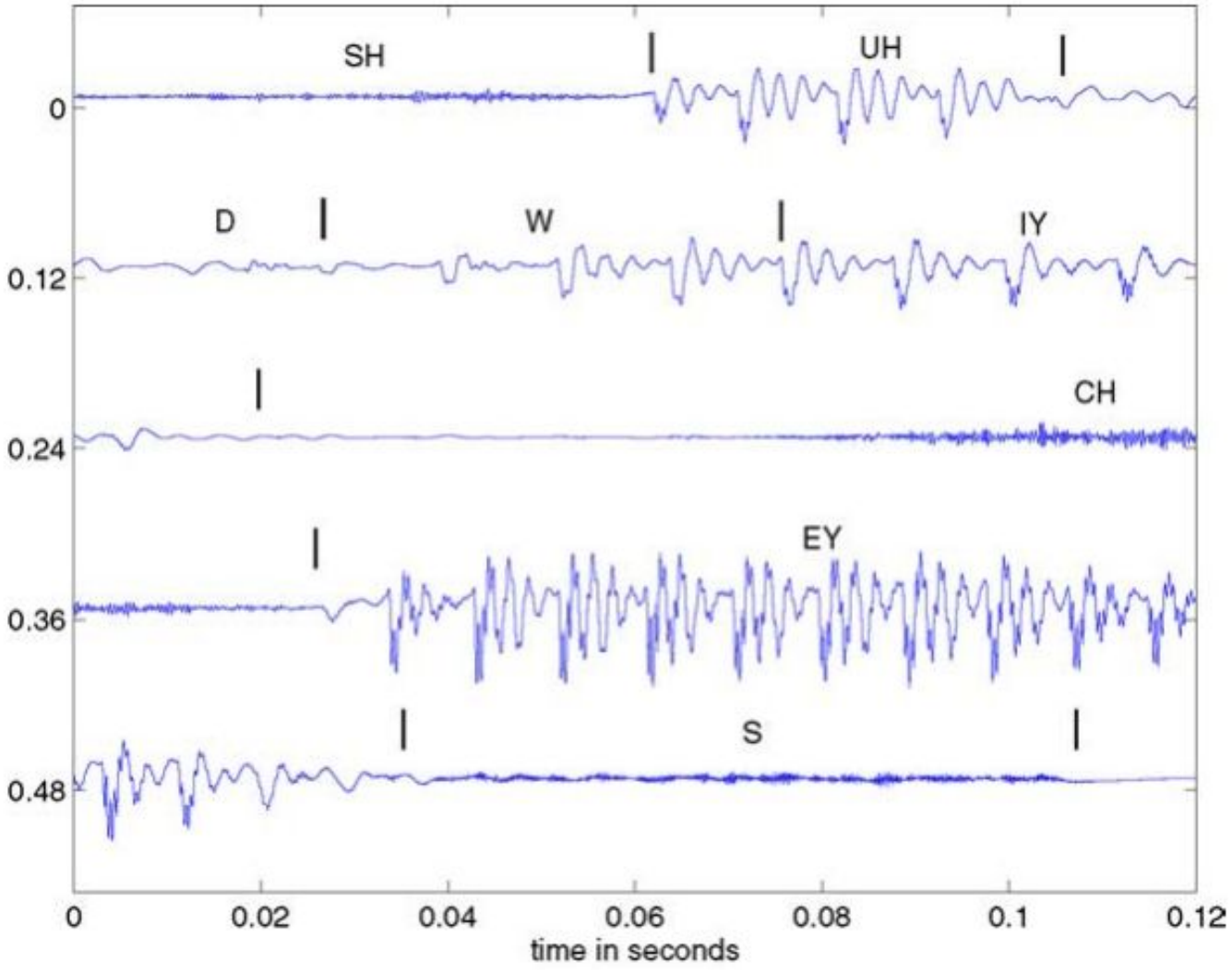
UNIT I

SPEECH SIGNAL CHARACTERISTICS & ANALYSIS

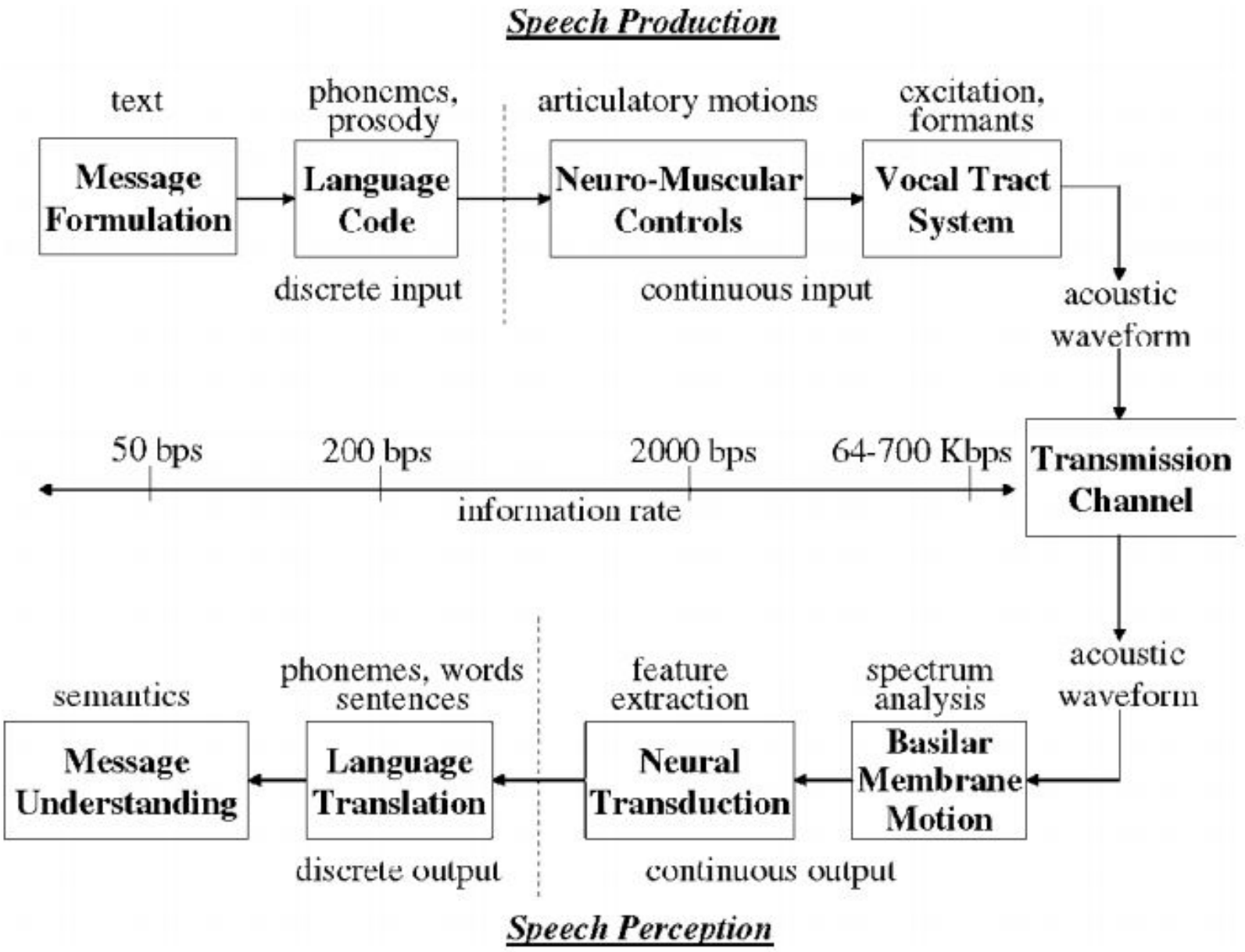
Speech signals

- Speech signals can be converted to an electrical waveform by a microphone, further manipulated by both analog and digital signal processing, and then converted back to acoustic form by a loudspeaker, a telephone handset or headphone, as desired.
- This form of speech processing is, of course, the basis for Bell's telephone invention as well as today's multitude of devices for recording, transmitting, and manipulating speech and audio signals.

A speech waveform with phonetic labels for the text message “Should we chase.”



The Speech Chain



- **language code generator**
  - to “speak” the message, the talker implicitly converts the text into a symbolic representation of the sequence of sounds corresponding to the spoken version of the text.
  - Phonetics - describe the basic sounds of a spoken version of the message
  - Prosody - the speed and emphasis in which the sounds are intended to be produced.
- phonetic symbols – developed using ARPAbet
  - Ex: “***Should we Chase”*** □ [SH UH D — W IY — CH EY S]
- **neuro-muscular controls**
  - the set of control signals that direct the neuro-muscular system to move the speech articulators, namely the tongue, lips, teeth, jaw and velum, in a manner that is consistent with the sounds

- **vocal tract system** - physically creates the necessary sound sources and the appropriate vocal tract shapes over time so as to create an acoustic waveform
- **rate of information flow**
  - Assume that there are about $32 = 2^5$ symbols
  - rate of speaking for most people is about 10 symbols per second
  - Thus the base information rate of the text message is 50 bps
  - During conversion to phonemes and prosody the information rate is estimated to increase by a factor of 4 to about 200 bps
  - After first two stages in the speech production part of the speech chain, the representation becomes continuous - estimate the spectral bandwidth of these control signals and appropriately sample and quantize these signals to obtain equivalent digital signals for which the data rate could be estimated ~

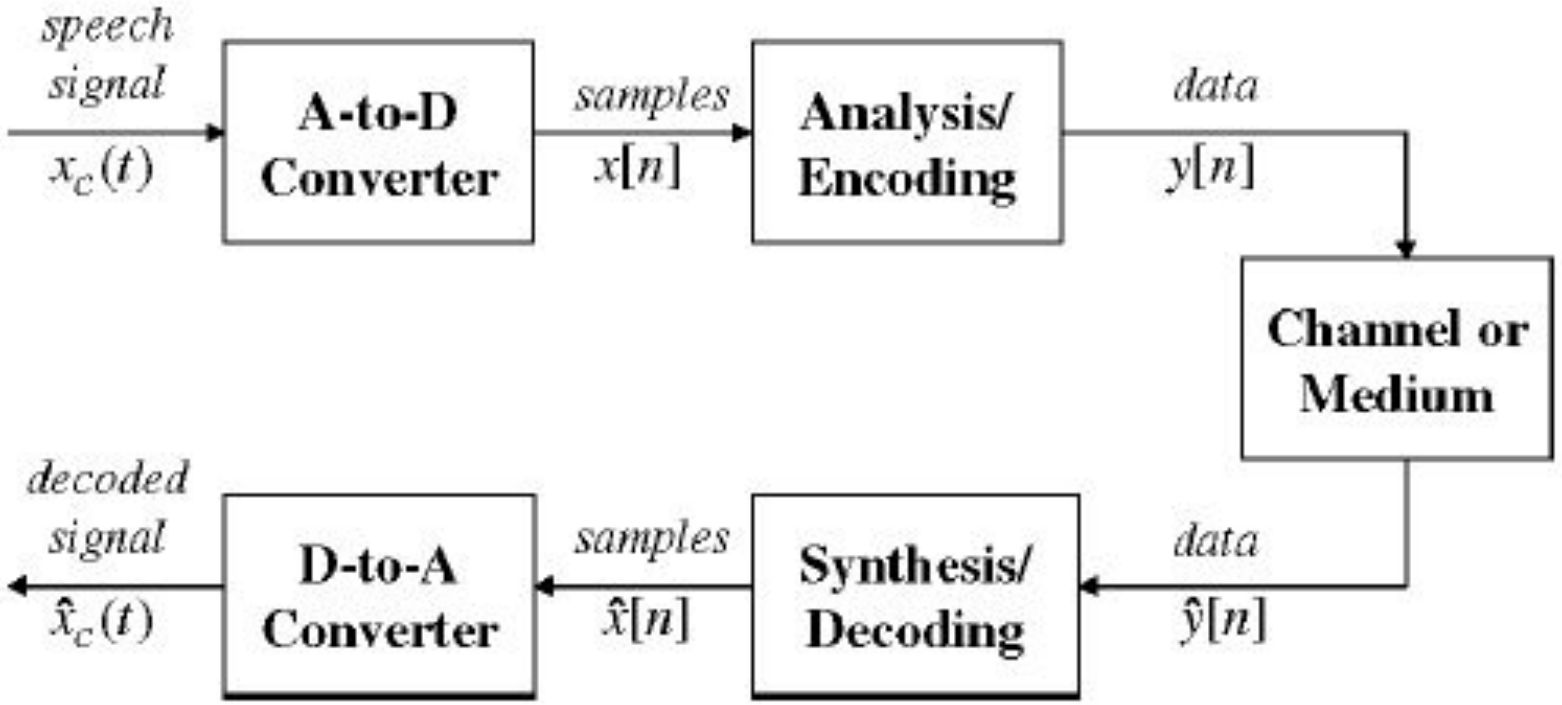
- The “telephone quality” requires that a bandwidth of 0–4 kHz be preserved, implying a sampling rate of 8000 samples/s. Each sample can be quantized with 8 bits on a log scale, resulting in a bit rate of 64,000 bps - highly intelligible but to most listeners, it will sound different from the original speech signal uttered by the talker.
- “CD quality” using a sampling rate of 44,100 samples/s with 16 bit samples, or a data rate of 705,600 bps - indistinguishable from the original speech signal
- **As we move from text to sampled speech waveform, the data rate can increase by a factor of 10,000.** Part of this extra information represents characteristics of the talker such as emotional state, speech mannerisms, accent, etc., but much of it is due to the inefficiency.
- **central theme of digital speech processing is to obtain**

speech perception model

- Effective conversion of the acoustic waveform to a spectral representation.
  - done within the inner ear by the basilar membrane, which acts as a non-uniform spectrum analyzer by spatially separating the spectral components of the incoming speech signal and thereby analyzing them by what amounts to a non-uniform filter bank.
- neural transduction of the spectral features into a set of sound features that can be decoded and processed by the brain.
- conversion of the sound features into the set of phonemes, words, and sentences associated with the in-coming message by a language translation process in the human brain.
- conversion of the phonemes, words and

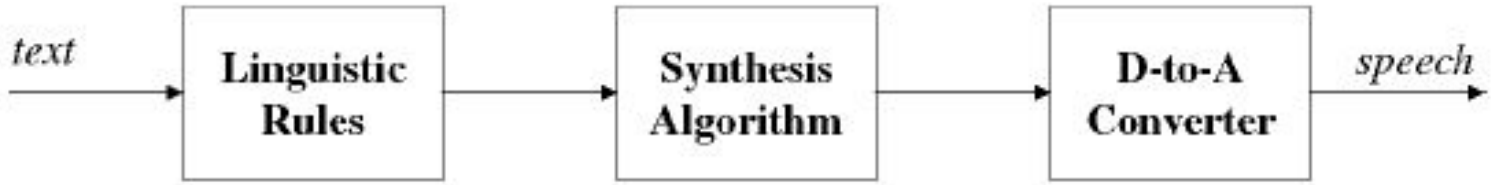
Applications of Speech Signal Processing

Speech Coding



- Applications of Speech Coders:
  - Narrowband and broadband wired telephony,
  - cellular communications,
  - voice over internet protocol (VoIP) (which utilizes the internet as a real-time communications medium),
  - secure voice for privacy and encryption (for national security applications),
  - extremely narrowband communications channels (such as battlefield applications using high frequency (HF) radio), and
  - for storage of speech for telephone answering machines,
  - interactive voice response (IVR) systems, and
  - Prerecorded messages.

Text-to-Speech Synthesis



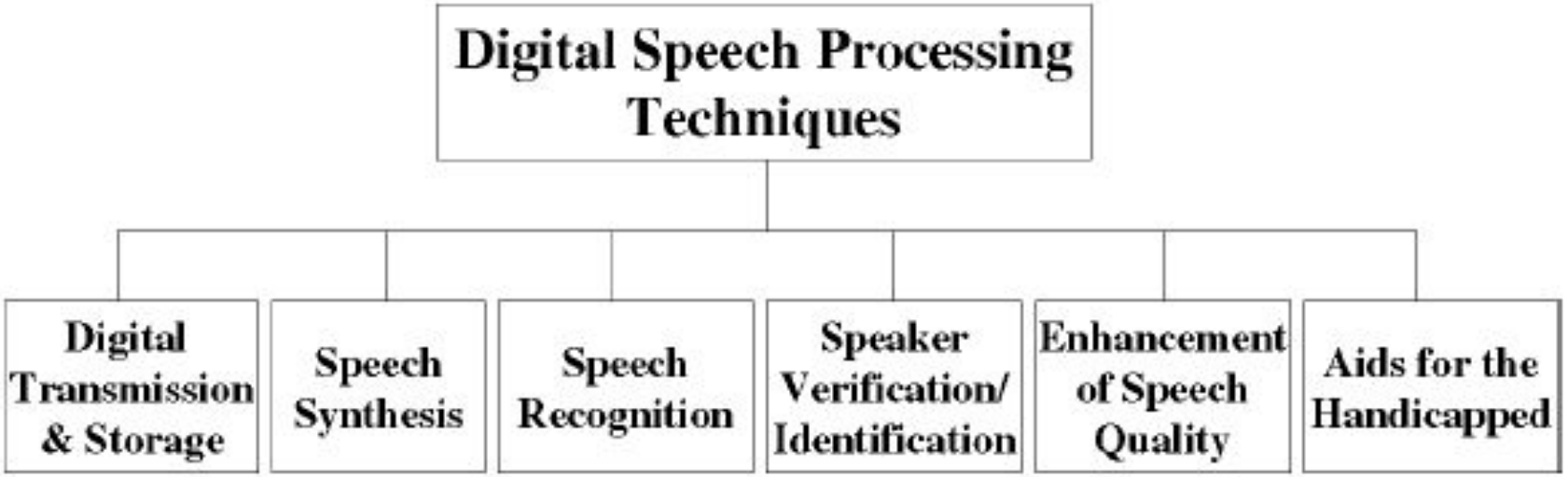
- The conversion from text to sounds involves a set of linguistic rules that must determine the appropriate set of sounds (perhaps including things like emphasis, pauses, rates of speaking, etc.) so that the resulting synthetic speech will express the words and intent of the text message in what passes for a natural voice that can be decoded accurately by human speech perception.
- the linguistic rules must determine how to pronounce acronyms, how to pronounce ambiguous words like read, bass, object, how to pronounce abbreviations like St. (street or Saint),

- the role of the synthesis algorithm is to create the appropriate sound sequence to represent the text message in the form of speech.
- the synthesis algorithm must simulate the action of the vocal tract system in creating the sounds of speech.
- “unit selection and concatenation.” Algorithm: the computer stores multiple versions of each of the basic units of speech, and then decides which sequence of speech units sounds best for the particular text message that is being produced.
- Appln:
  - read email messages over a telephone,
  - provide voice output from GPS systems in automobiles,
  - provide the voices for talking agents for completion of transactions over the internet, handle call center help

Speech Recognition and Other Pattern Matching Problems

- pattern matching problems in speech processing include the following:
  - speech recognition, where the object is to extract the message from the speech signal;
  - speaker recognition, where the goal is to identify who is speaking;
  - speaker verification, where the goal is to verify a speaker's claimed identity from analysis of their speech signal
  - word spotting, which involves monitoring a speech signal for the occurrence of specified words or phrases;

- automatic language translation - to convert spoken words in one language to spoken words in another language so as to facilitate natural language voice dialogues between people speaking different languages.



Representing Speech in time and frequency domains

The simplest and straightforward method is via the state of the speech production source – the vocal cords.

Three state representation

- Silence (S) – no speech is produced
- Unvoiced (U) – vocal cords are not vibrating, so the resulting speech is aperiodic or random in nature
- Voiced (V) – vocal cords are tensed and therefore vibrate periodically when air flows from the lungs.

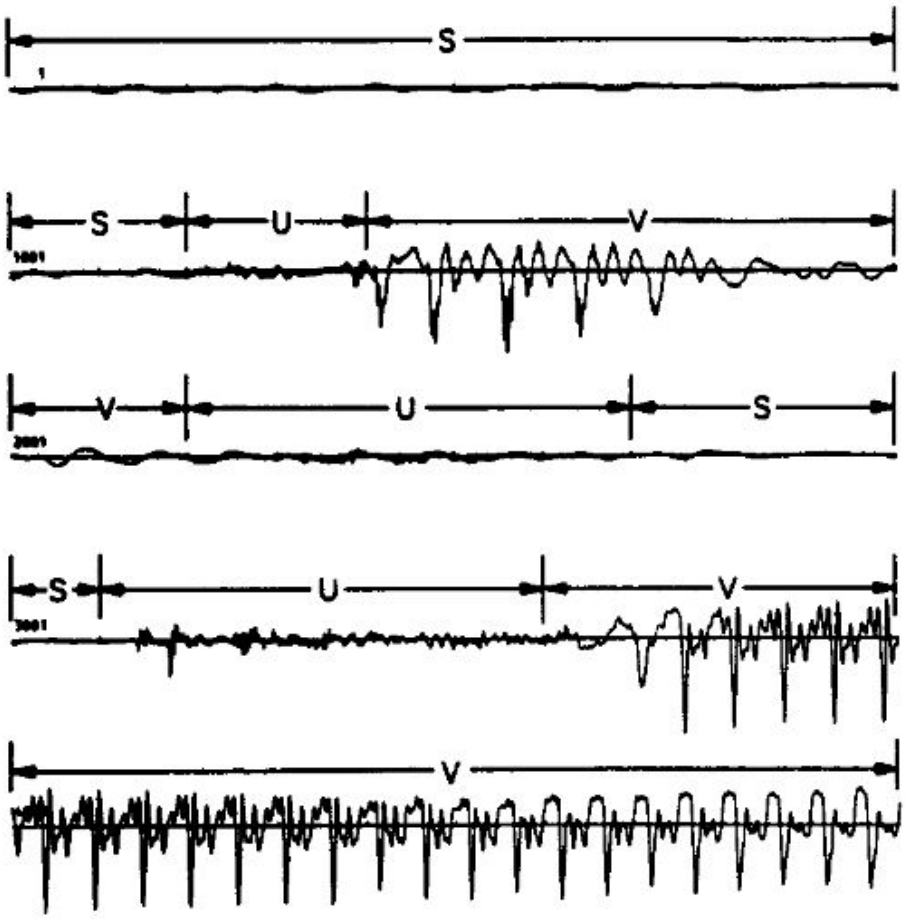


Figure 2.7 Waveform plot of the beginning of the utterance "It's time."

Spectral representation:

Spectrogram – three dimensional representation of the speech intensity, in different frequency bands, over time is portrayed.

- Wideband Spectrogram – spectral analysis on 15-msec sections of waveform using a broad analysis filter – spectral envelope of individual periods of the speech during voiced sections are resolved.
- Narrowband spectrogram – spectral analysis on 50-msec sections using a narrow analysis filter – individual spectral harmonics corresponding to the pitch of the speech waveform during voiced regions are resolved.

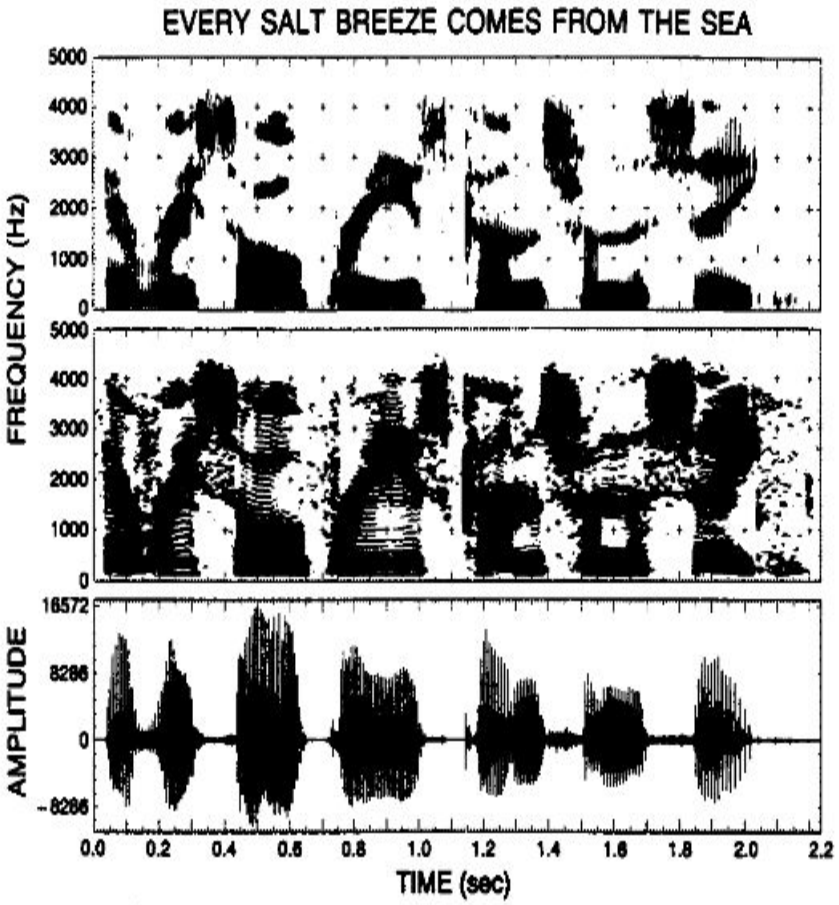


Figure 2.8 Wideband and narrowband spectrograms and speech amplitude for the utterance "Every salt breeze comes from the sea."

Parameterization of the spectral activity based on the model of speech production.

Because human vocal tract is essetially a tube, of varying cross-sectional area that is excited either at one end or at a point along the tube, acoustic theory tells that the transfer function of energy from the excitation source to the output can be described in terms of the natural frequencies or resonances of the tube. Such resonances are called ***formants*** for speech.

Disadv: difficulty of reliably estimating the formant frequencies for low-level voiced sound, and difficulty of defining the formants for unvoiced or silence regions.

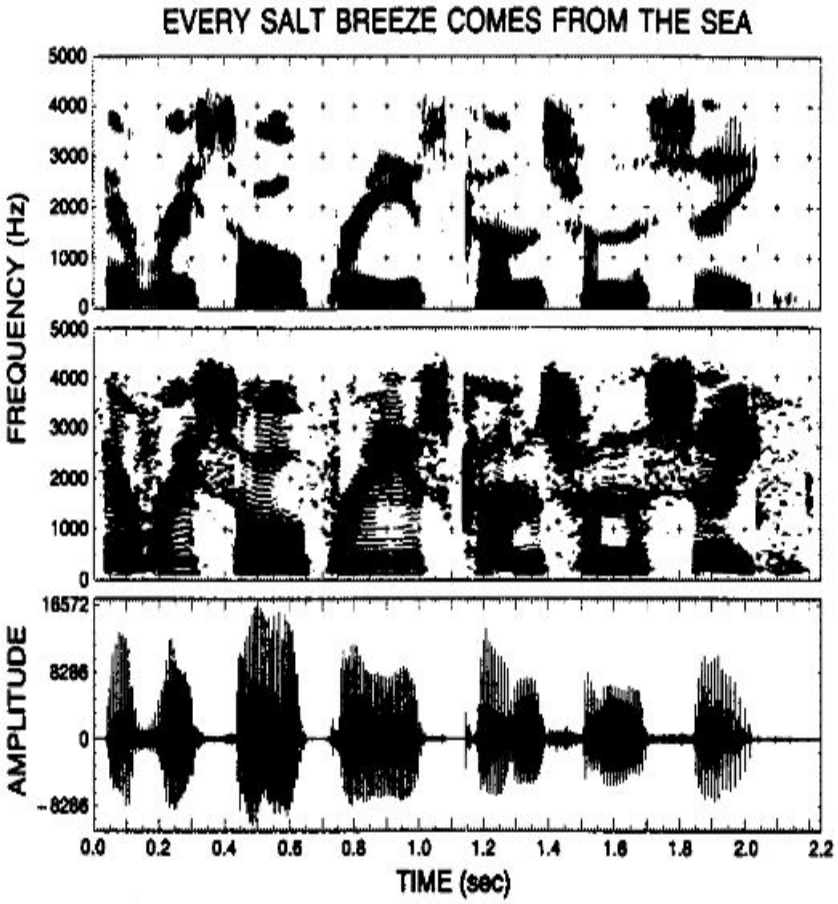


Figure 2.8 Wideband and narrowband spectrograms and speech amplitude for the utterance "Every salt breeze comes from the sea."

Speech sounds and Features

Phonemes – number of linguistically distinct speech sounds in a language.

- 18 vowels or vowel combinations
- 21 standard consonants
- 4 syllabic sounds
- 4 vowel-like consonants
- 1 glottal stop

TABLE 2.1. A condensed list of phonetic symbols for American English.

Phoneme	ARPABET	Example	Phoneme	ARPABET	Example
/i/	IY	beat	/ŋ/	NX	sing
/ɪ/	IH	bit	/p/	P	pet
/e/ (eʸ)	EY	bait	/t/	T	ten
/ɛ/	EH	bet	/k/	K	kit
/æ/	AE	bat	/b/	B	bet
/a/	AA	Bob	/d/	D	debt
/ʌ/	AH	but	/g/	H	get
/ɔ/	AO	bought	/h/	HH	hat
/o/ (oʷ)	OW	boat	/f/	F	fat
/ʊ/	UH	book	/θ/	TH	thing
/u/	UW	boot	/s/	S	sat
/ə/	AX	about	/ʃ/ (sh)	SH	shut
/ɨ/	IX	roses	/v/	V	vat
/ɝ/	ER	bird	/ð/	DH	that
/ɚ/	AXR	butter	/z/	Z	zoo
/aʷ/	AW	down	/ʒ/ (zh)	ZH	azure
/aʸ/	AY	buy	/č/ (tsh)	CH	church
/ɔʸ/	OY	boy	/ǰ/ (dzh, j)	JH	judge
/y/	Y	you	/ʍ/	WH	which
/w/	W	wit	/l̩/	EL	battle
/ɾ/	R	rent	/m̩/	EM	bottom
/l/	L	let	/n̩/	EN	button
/m/	M	met	/ɾ̃/	DX	batter
/n/	N	net	/ʔ/	Q	(glottal stop)

Vowels: most practical speech-recognition systems rely heavily on vowel recognition

Th_y n_t_d s_gn_f_c_nt _mpr_v_m_nts i_ th_ c_mp_ny's _m_g_, s_p_rv_s__n,
th__r w_rk_ng c_nd_t__ns, b_n_f_ts _nd _pp_rt_n_t__s f_r gr_wth.

A__i_u_e_ _o_a___a___a_e_ e__e__ia___ __e _a_e, _i__ __e __o_e_ o_
o__u_a_io_a_ e___o_ee_ __i_____ _e__ea_i__.

- Vowels are produced by exciting an essentially fixed vocal tract shape with quasi-periodic pulses of air caused by the vibration of the vocal cords.
- Ling in duration and spectrally well defined.
- Front vowel sound – high-frequency resonance
- Mid vowels – balance of energy over a broad frequency range
- Back vowels – predominance of

TABLE 2.2. Formant frequencies for typical vowels.

ARPABET Symbol for Vowel	IPA Symbol	Typical Word	F_1	F_2	F_3
IY	/i/	beet	270	2290	3010
IH	/ɪ/	bit	390	1990	2550
EH	/ɛ/	bet	530	1840	2480
AE	/æ/	bat	660	1720	2410
AH	/ʌ/	but	520	1190	2390
AA	/ɑ/	hot	730	1090	2440
AO	/ɔ/	bought	570	840	2410
UH	/ʊ/	foot	440	1020	2240
UW	/u/	boot	300	870	2240
ER	/ɝ/	bird	490	1350	1690

Diphthongs: gliding monosyllabic speech sound that starts at or near the articulatory position for one vowel and movers to or toward the position for another.

Six diphthongs in American English:

- /a^{y}/ (as in b<u>uy</u>)
- /a^{w}/ (as in d<u>ow</u>n)
- /e^{y}/ (as in b<u>ai</u>t)
- /o^{y}/ (as in b<u>oy</u>)
- /o/ (as in b<u>oa</u>t)
- /ju/ (as in y<u>ou</u>)

Semivowels: vowel-like nature

- Gliding transition in vocal tract are function between adjacent phonemes.
- /w/, /l/, /r/ and /y/

Nasal Consonants: glottal excitation and vocal tract totally constricted at some point along the oral passageway.

- **/m/** - as in sa**m**e - constriction at the lips
- **/n/** - as in **n**ext - constriction just behind the teeth
- **/ng/** - as in worki**ng** – constriction is just forward of the velum itself.

Unvoiced Fricatives: exciting the vocal tract by a steady air flow, which becomes turbulent in the region of a constriction in the vocal tract.

- /f/ - constriction near the lips
- /s/ - middle of the oral tract
- /sh/ - back of the oral tract

Voiced Fricatives: place of construction is same as unvoiced fricatives except that two excitation sources are involved in their production.

Ex: /v/, /th/, /z/, /zh/

Voiced and Unvoiced stops:

Voice stop consonants – transient, noncontinuant sounds produced by building up pressure behind a total constriction somewhere in the oral tract and then suddenly releasing the pressure.

- /b/ - constriction at the lips
- /d/ - constriction at the back of the teeth
- /g/ - near the velum

Unvoiced stop consonants – similar to voiced counterpart except that vocal cords do not vibrate

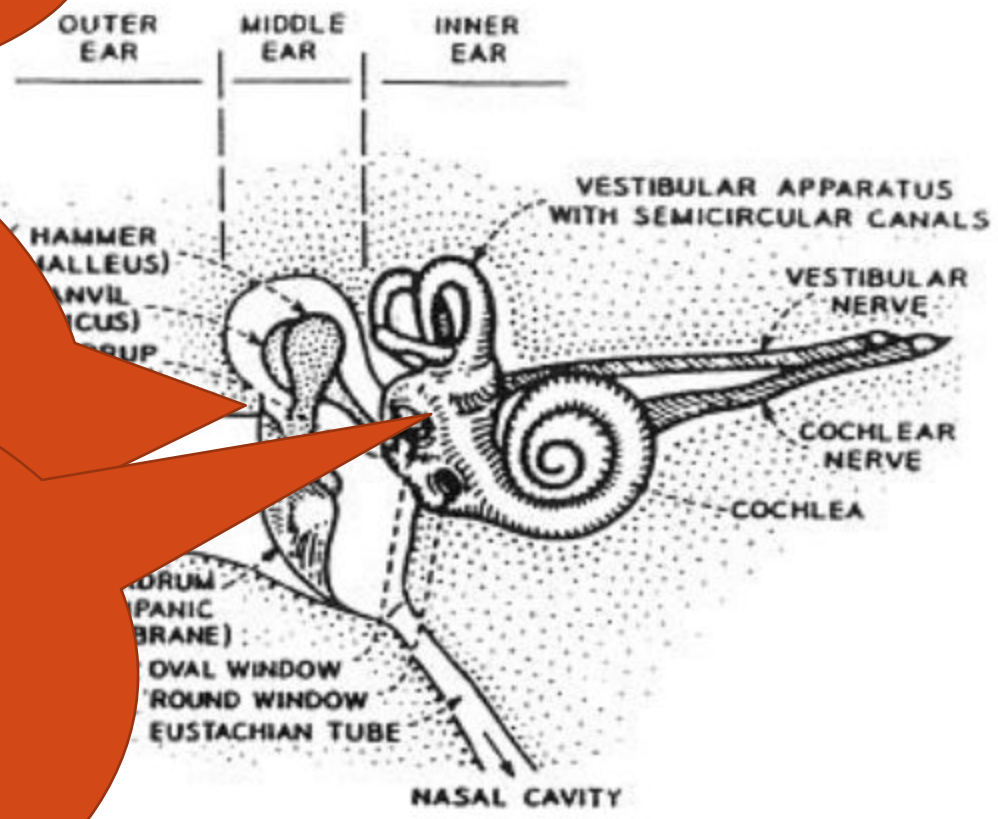
- /p/ , /t/, /k/

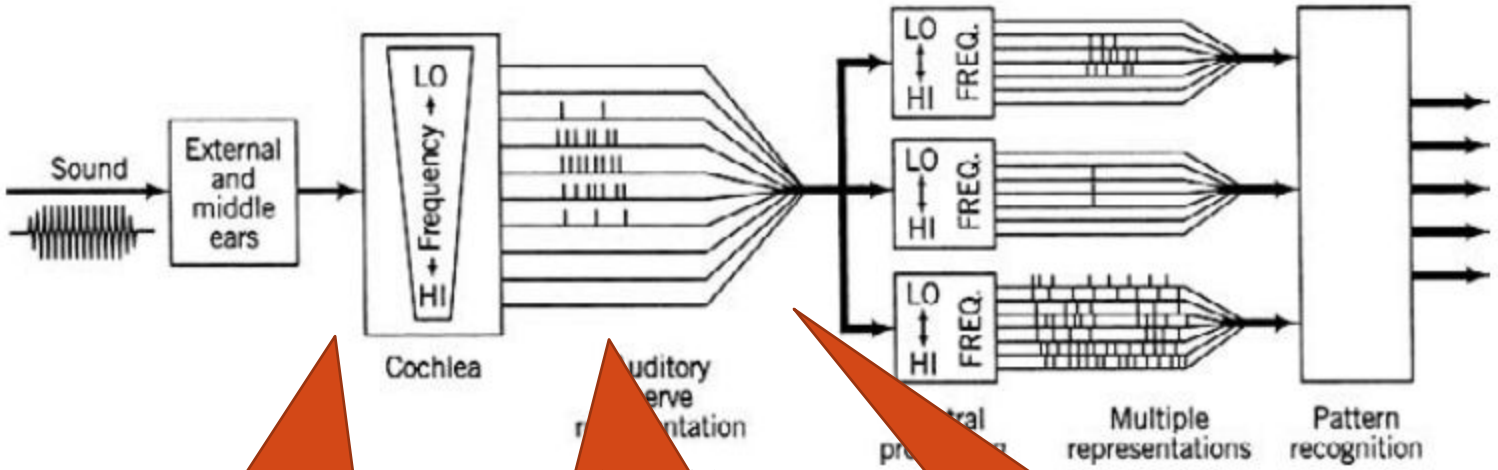
Hearing and Auditory Perception

outer ear consisting of the pinna, which gathers sound and conducts it through the external canal to the middle ear

middle ear beginning at the tympanic membrane, or eardrum, and including three small bones, the *malleus* (also called the

inner ear, which consists of the cochlea and the set of neural connections to the auditory nerve, which conducts the neural signals to the brain



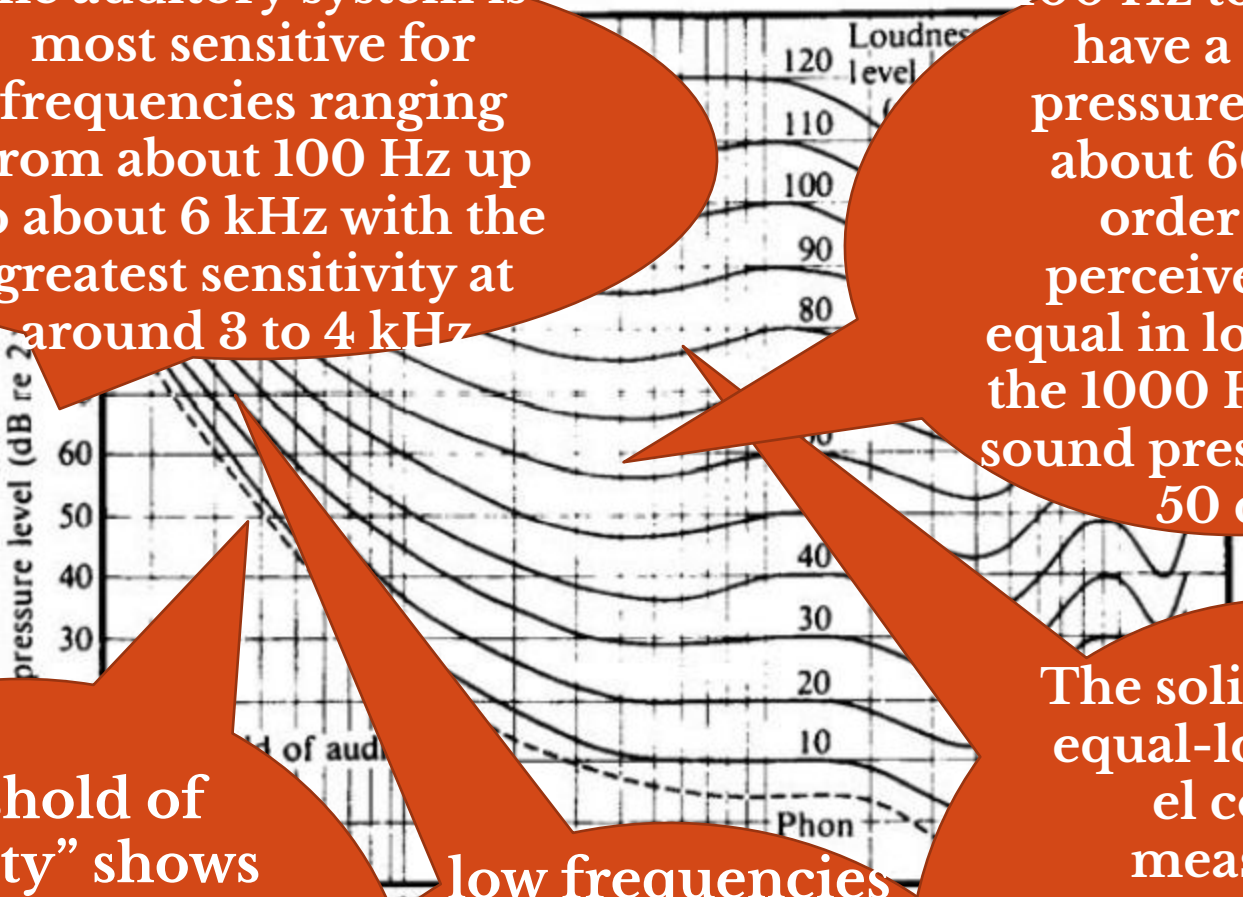


ear drum and *bor*
structures con
the sound wa
mechani
vibrations w
are transferr
the *basilar*
membrane insid
the *cochlea*.

basilar membrane vibrates in a frequency-selecti ve manner along its extent and thereby performs a rough (non-uniform) spectral analysis of the sound

the basilar membrane are a set of inner hair cells that serve to convert motion along the basilar membrane to neural activity. This produces an auditory nerve representation in both time and frequency.

- **Perception of Loudness**
  - Loudness is a perceptual quality that is related to the physical property of sound pressure level.
  - Loudness is quantified by relating the actual sound pressure level of a pure tone (in dB relative to a standard reference level) to the perceived loudness of the same tone (in a unit called phons) over the range of human hearing (20 Hz–20 kHz).



the auditory system is most sensitive for frequencies ranging from about 100 Hz up to about 6 kHz with the greatest sensitivity at around 3 to 4 kHz

Hz tone must have a sound pressure level of about 60 dB in order to be perceived to be equal in loudness to the 1000 Hz tone of sound pressure level 50 dB.

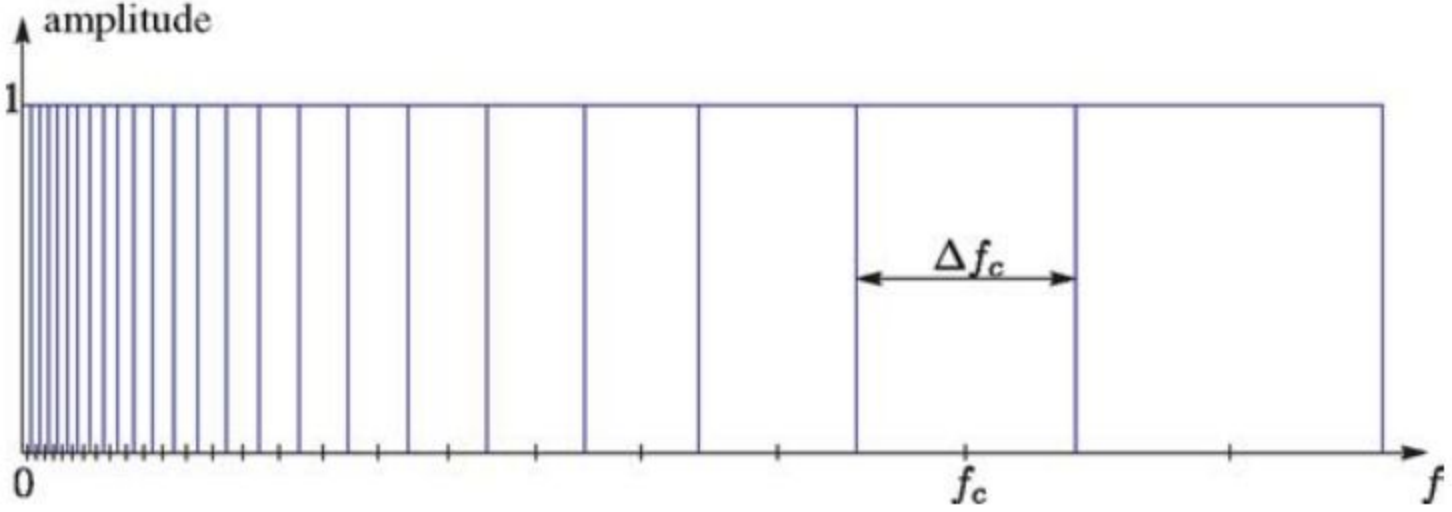
"threshold of audibility" shows the sound pressure level that is required for a sound of a given frequency to be just audible

low frequencies must be significantly more intense than frequencies in the mid-range in order that they be perceived at all

The solid curves are equal-loudness-level contours measured by comparing sounds at various frequencies with a pure tone of frequency 1000 Hz and known sound pressure level.

Critical Bands

- Non-uniform frequency analysis – by the basilar membrane - equivalent to that of a set of bandpass filters whose frequency responses



- In reality, the bandpass filters are not ideal, but their frequency responses overlap significantly since points on the basilar membrane cannot vibrate independently of each other

- the effective bandwidths are constant at about 100 Hz for center frequencies below 500 Hz, and with a relative bandwidth of about 20% of the center frequency above 500 Hz.
- equatic nts over the auc $\Delta f_c = 25 + 75[1 + 1.4(f_c/1000)^2]^{0.69}$

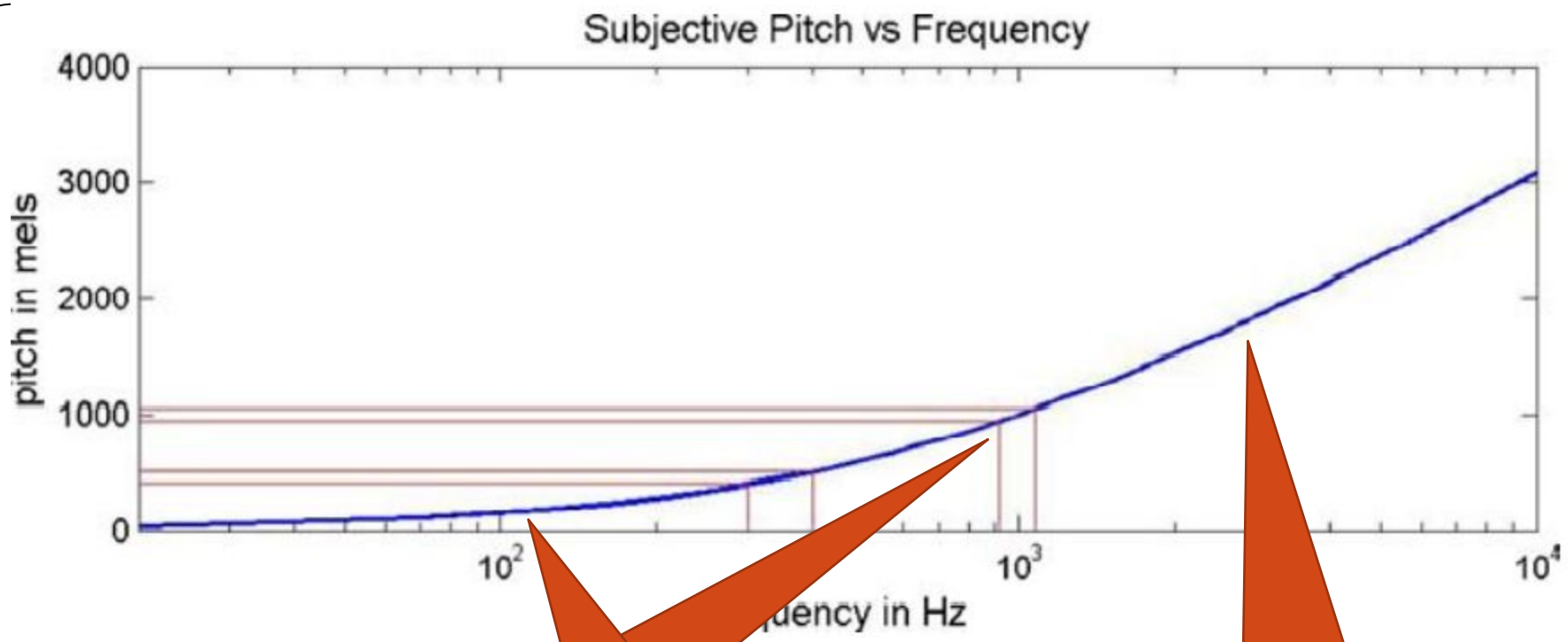
  Δf_c

  - - critical bandwidth associated with center frequency
- Approximately 25 critical band filters span the range from 0 to 20 kHz.

Pitch Perception

- Most musical sounds as well as voiced speech sounds have a periodic structure when viewed over short time intervals, and such sounds are perceived by the auditory system as having a quality known as pitch.
- The relationship between pitch (measured on a nonlinear frequency scale called the mel-scale) of a pure tone is approximated by the equation

$$\text{Pitch in mels} = 1127 \log_e(1 + f/700)$$



a frequency of 1000 Hz corresponds to a pitch of 1000 mels

Below 1000 Hz, the relationship between pitch and frequency is nearly proportional

For higher frequencies, however, the relationship is nonlinear.

- Independently of the center frequency of the band, one critical bandwidth corresponds to about 100 mels on the pitch scale.