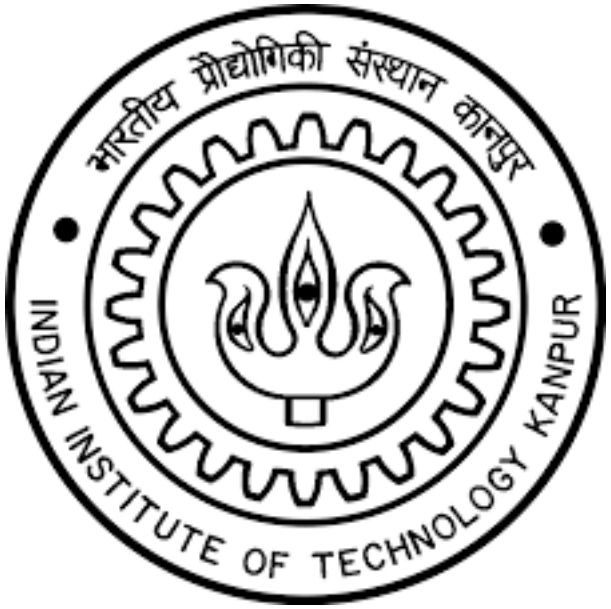


Indian Institute of Technology
Kanpur

In Collaboration
with

National Program on Technology Enhanced Learning (NPTEL)

Presents

Course Title:

Basic Cognitive Processes

By: Dr. Ark Verma,
Assistant Professor of Psychology,
Department of Humanities & Social Sciences,
IIT Kanpur

Lecture 23: Auditory Perception

Hearing: The Preliminaries

- Hearing, too begins with transduction.
    - sound waves are collected by our ears and converted into neural impulses, which are sent to the brain where they are integrated with past experience and interpreted as the sounds we experience.
- the human ear is sensitive to a wide range of sounds, ranging from the faint click of a clock to the roar of a rock band.
- but the human ear is particularly sensitive to the sounds in the same frequency range as the human voice.

- the Ear: detects sound waves.
    - vibrating objects (such as the human vocal chords or guitar strings) cause air molecules to bump into each other and produce sound waves, which travel from they source as peaks and valleys much like the ripples that expand outward when a stone is tossed into a pond.
    - sound waves are carried within medium such as air, water or metal, & it is the changes in pressure associated with these mediums that the ear detects.

- ***Physical Characteristics of Sound***
  - we detect both the *wavelength* & the *amplitude* of sound waves.
    - the wavelength of the sound wave (known as frequency) is measured in terms of the number of waves that arrive per second and determines our perception of pitch, i.e. the perceived frequency of the sound.
    - longer sound waves have lower frequency & produce a lower pitch whereas shorter sound waves have higher frequency & higher pitch.
    - the amplitude, or height of the sound wave, determines how much energy it contains and is perceived as loudness (the degree of sound volume).

- larger waves are perceived as louder.

- loudness is measured using the unit of relative loudness known as ***decibel***.

  - zero decibels represent the absolute threshold for human hearing, below which we cannot hear a sound. each increase in 10 decibels represents a ten - fold increase in the loudness of the sound.

  - the sound of a typical conversation (about 60 decibels) is 1,000 times louder that the sound of a whisper (30 decibels).

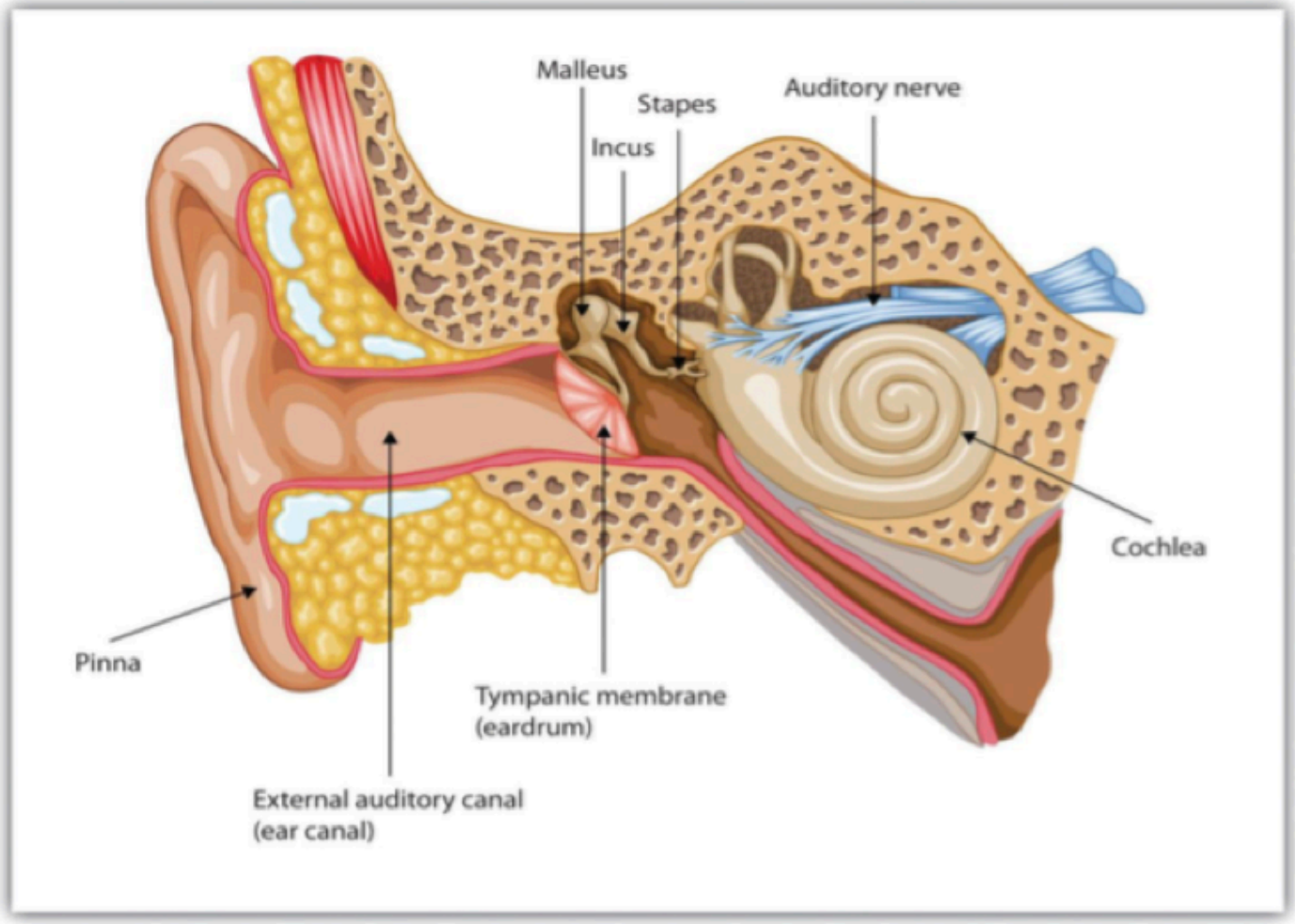
- ***The Structure of the Ear***:
  - audition begins in the ***pinna***, the external & visible part of the ear, which is shaped like a funnel to draw in sound waves & guide them into a auditory canal,
  - at the end of the canal, the sound waves strike the tightly stretched, highly sensitive membrane known as the ***tympanic membrane (or eardrum)***, which vibrates with the waves.
  - the resulting vibrations are relayed into the middle ear through three tiny bones, known as the ***ossicles - the hammer*** (malleus), ***the anvil*** (incus) and ***the stirrup*** (stapes) - to the ***cochlea***, a sail shaped liquid filled tube in the inner ear.

- the vibrations cause the oval window, the membrane covering the opening of the cochlea, to vibrate, disturbing the fluid inside the cochlea.

- the movements of the fluid in the cochlea bend the hair cells of the inner ear. the movement of the hair cells trigger nerve impulses in the attached neurons, which are sent to the auditory nerve and then to the auditory cortex in the brain.

- the cochlea contains about 16,000 hair cells, each of which hold a bundle of fibres known as *cilia* on its tip.

- the cilia are so sensitive that they can detect a movement that pushes them the width of a single atom or shifting the Eiffel Tower by half an inch (Corey et al., 2004).



Sound waves enter the outer ear and are transmitted through the auditory canal to the eardrum. The resulting vibrations are moved by the three small ossicles into the cochlea, where they are detected by hair cells and sent to the auditory nerve.

Image: Stangor (2010). Introduction to Psychology. *Flat World Knowledge*. Creative Commons license. (p/ 203).

- the loudness of the sound is directly determined by the number of hair cells that are vibrating.

- two different mechanisms are used to detect pitch.

  - ***the frequency theory*** of hearing proposes that *whatever the pitch of a sound wave, nerve impulses of a corresponding frequency will be sent to the auditory nerve.* for e.g. a tone measuring 600 Hz will be transducer into 600 nerve impulses a second.
    - but for high pitched sounds this theory can't explain, because neurons won't be able to fire fast enough for higher frequencies.

  - a solution could be that to reach the necessary speed, the neurons work together in a sort of volley system in which different neurons fire in sequence, allowing us to detect sounds up to 4000 Hz.

- the place theory of hearing proposes that different areas of the cochlea respond to different frequencies.
    - higher tones excite areas closest to the opening of the cochlea (near the oval window). whereas lower tones excite areas near the narrow tip of the cochlea, at the opposite end.
- pitch is therefore determined in part by the area of the cochlea firing the most frequently.

- that the ears are placed on either side of the head enables us to benefit from stereophonic, or three dimensional hearing.

- if a sound occurs on your left side, the left ear will receive the sound slightly sooner than the right ear and the sound will receive will be more intense, allowing you to quickly determine the location of the sound.

    - although the distance between the two ears is barely 6 inches & sound waves travel at 750 miles an hour; the time & intensity differences are easily detected (Middlebrooks & Green, 1991).

- when a sound is equidistant from both ears (such as when it is directly in front or back, beneath or overhead; we have more difficulty pinpointing its exact location & we may maneuver to facilitate localization.

Speech Perception

- The most important class of stimuli that we perceive via auditory perception is speech stimuli.
- In that reference, speech perception deserves a more important mention.
- During speech perception, the auditory system needs to analyze the sound vibrations generated by someone's conversation.

- ***Characteristics of Speech Perception*** (Matlin, 2008)
  - When describing these speech sounds, psychologists & linguists use the term ***phoneme*.**
    - a phoneme refers to the basic unit of spoken language, which includes basic sounds as *a, k, th*. The English language uses about 45 phonemes, including both consonants & vowels.
  - Listeners can impose boundaries between words, even when these words are not separated by silence. #speech segmentation

- Phoneme pronunciation varies tremendously. #phoneme variation

- Context allows listeners to fill in missing sounds. #role of context

- Visual cues from the speaker's mouth help us interpret ambiguous sounds. #multi – modal perception

Theories of Speech Perception

- The ***special mechanism approach*** proposes that speech perception is accomplished by a naturally selected module (Fodor, 1983).
    - this speech perception module monitors incoming acoustic stimulation and reacts strongly when the signals contains the characteristic complex patterns that make up speech.
    - when the speech module recognized an incoming stimulus as speech, it preempts other auditory processing systems, preventing their output from entering consciousness.

- So, while the non - speech sounds are analyzed according to the basic properties of frequency, amplitude, and timbre, and while we are able to perceive those characteristics of non-speech sounds accurately, when the speech module latches onto an acoustic stimulus; it prevents the kind of spectral analysis that general auditory processing mechanisms generally carry out for non - speech auditory stimuli.

- the preemption of normal auditory perceptual processes for speech stimuli can lead to *duplex perception* under special, controlled lab conditions (Liberman & Mattingly, 1989).

  - to create their experimental stimuli, researchers constructed artificial speech stimuli that sounded like /da/ or /ga/ depending upon whether the second formant transition decreased (/da/) in frequency over time or increased (/ga/).
  - next, they edited their stimuli to create separate signals for the transition and the rest of the syllable, which they called the *base*.
  - they played the two parts of the stimulus over headphones, with the transition going in one ear & the base going in one ear.

- the question was, how would people perceived the stimulus?
    - it turned out that people perceived two different things at the same time. at the ear that the transition was played into, people perceived a high - pitched chirp or whistle. But at the same they perceived the original syllable, just as if the entire, intact stimulus had been presented.
- Liberman & colleagues, argued that simultaneously perceiving the transition in two ways - as a chirp & as a phoneme - reflected the simultaneous operation of the speech module and general purpose auditory processing mechanisms.

- duplex perception happened because the auditory system could not treat the transition and base as coming from the same source (as they were played in two different ears).

- as the auditory system recognised two different sources, it had to do something with the transition that it would not normally do., i.e. it had to analyse it for the frequencies it contained and the result was hearing it as a chirp.

- but simultaneously, speech processing module recognised a familiar pattern of transitions and formants & as a result the auditory system reflexively integrated the transition & the base and led to the experience of hearing a unified syllable.

- ***the motor theory of speech perception*** :

  - that gestures, rather sound, represent the fundamental unit of mental representation in speech (Liberman & Whalen, 2000; Fowler, 2008).

    - i.e. when we speak, we attempt to move your articulators to particular places in specific ways. each of these movement constitutes a gesture.

- the motor part of speech production system takes the sequence of words we want to say & comes up with a *gestural score,* that tells our articulators how to move.

- acc. to the theory, if you can figure out what gestures created a speech signal, you can figure out what the gestural plan was, which takes you back to the sequence of syllables or words that went into the gestural plan in the first place.

- So, by knowing what the gestures are, you can tell what was the set of words that produced that set of gestures.
    - For e.g. the “core” part of the gesture to produce either “di” or “du” sounds is tapping the tip of your tongue against the back of your teeth (or your alveolar ridge).
- Other parts of the gesture, like lip position are affected by coarticulation, but the core component of the gesture is the same regardless of the phonological context.

- Thus, rather than trying to map acoustic signals directly to phonemes, Alvin Liberman & his colleagues proposed that we map acoustic signals to gestures that produced them, as there is a closer relationship between gestures and phonemes than there is between acoustic signals & phonemes.

- In their words, "The relation between perception & articulation will be considerably simple than the relation between perception and the acoustic stimulus."

- Further, "perceived similarities and differences will correspond more closely to the articulatory than the acoustic similarities among the sounds."

- So, differences between two acoustic signals will not cause you to perceive two different phonemes as long as the gestures that created those two different acoustic signals are the same.

- Another aspect of the motor theory proposes, ***categorical perception*** is another product of the speech perception module.

- categorical perception happens when a wide variety of physically distinct stimuli are perceived as belonging to one of a fixed set of categories.

  - for example: every vocal tract is different from every other vocal tract & as a result the sound waves that come out of your mouth when you say *pink* are very different that the sound waves that come out of my mouth when I say *pink,* and so on.

- nonetheless, your phonological perception is blind to the physical differences and perceives all of those signals as containing an instance of the category /p/.

- Further, it may be noted that all of our voices have different qualities than each other, but we categorize the speech sounds from each of us, in much the same way. This is because, all of those different noises map to the same set of 40 phonemes (in English).

- In addition, although the acoustic properties of speech stimuli can vary across a wide range, our perception does not change in little bitty steps with each little bitty change in the acoustic signal.

- We are insensitive to some kinds of variation in the speech signal, but if the speech signal changes enough , we perceive that change as the difference between one phoneme and another (Liberman et al., 1957).

- An example:
  - the difference between /b/ & /p/ is that the /b/ is voiced while the /p/ is not.
  - other than voicing the two phonemes are essentially identical; in that they are both *labial plosives*, meaning that we make these sounds by closing our lips & allowing air pressure to build up behind our lip dam and then releasing the pressure suddenly, creating a burst of air that rushes out of the mouth.
  - the difference between the two phonemes has to do with the timing of the burst and the vocal fold vibrations that create voicing.
    - for the /b/ sound, the vocal folds begin vibrating while your lips are closed or just after; but for the /p/ sound, there is a delay between the burst and the point in time when the vocal folds begin to vibrate. This gap is the *voice onset time*.

- the VOT is a variable that can take any value whatsoever, so it is called a continuous variable. but even though VOT can vary continuously in this way, we do not perceive much of that variation.
    - for e.g. we can not greatly hear the difference between a bot of 2ms and 7ms or between 7 ms & 15ms.
- instead we map a range of VOTs on the same percept.
    - Those different acoustic signals are called *allophones* - different signals that are perceived as being the same phoneme.
    - so the experience with a range of short VOTs is as /b/ & long VOTs is as /p/; the difference point being 20ms.

References

- Traxler, M. J. (2013). Introduction to Psycholinguistics: Understanding Language Science. *Wiley – Blackwell*.