Human Computer Interaction

9. Audio and Speech

Course no. ILE5013

National Chiao Tung Univ, Taiwan

By: I-Chen Lin, Assistant Professor

Objectives

- Audio and 3D sound
- User interfaces with speech recognition or synthesis

Ref:
• D.A. Bowman, E. Kruijff, J.J. LaViola, I. Poupyrev, 3D User Interfaces: Theory and Practice, Addison Wesley Professional, 2005.
• Course notes, “3D User Interfaces”, CS, Columbia Univ..
• Course notes, “Introduction to Virtual Reality”, EPFL.
• Course notes, “Introduction to the Design, Prototyping, & Evaluation of Human-Computer Interfaces ”, CS, UC Berkeley.

Audio in UI

- Audio is another channel of communication between user and environment.
- Useful in designing systems where a user monitors several communication channels at once
- display of spatial information (particularly space beyond field of view)
- Simple stereo sounds are not convincing or short of full spatial information.

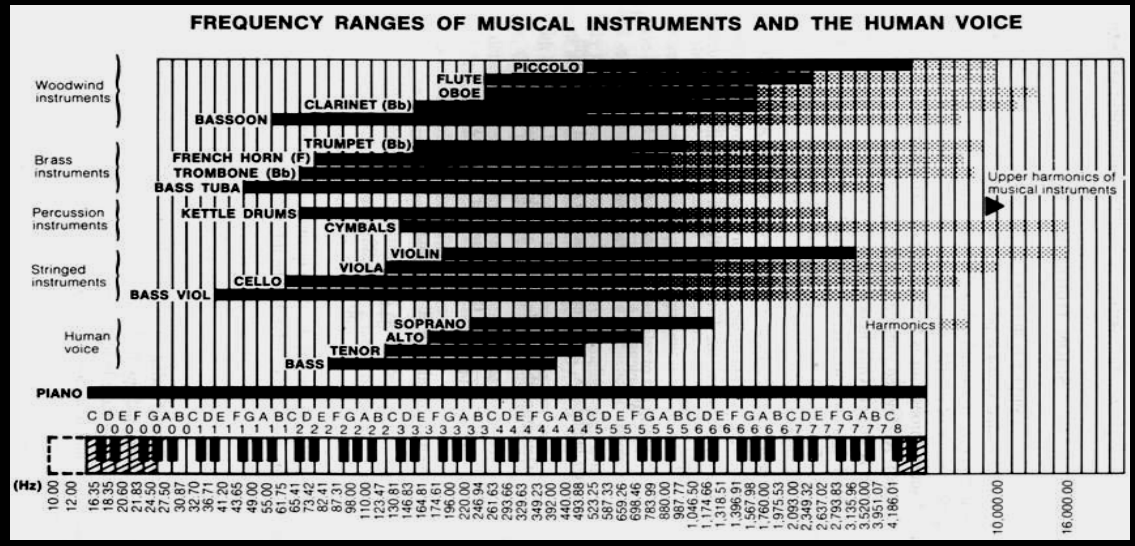
⇨ 3D sound

The Basics of Acoustics

- The sound source: Object that emits sound waves.
- The acoustic environment: In the medium sound waves are absorbed, reflected, refracted and diffracted in different ways.
  - depending on their frequency, and material and geometry of the environment.
- The listener: Sound receiving object, typically a 'pair of ears'.
  - extract information about the sound sources and the environment (spatial, motion info., etc.)
  - Audio signals provide a higher degree of temporal resolution than visual display.

Audio Perception

- Amplitude → loudness
  - ~10^{13} amplitude range.
- Frequency → pitch
  - 20Hz ~ 20KHz
- Waveshape → Timbre



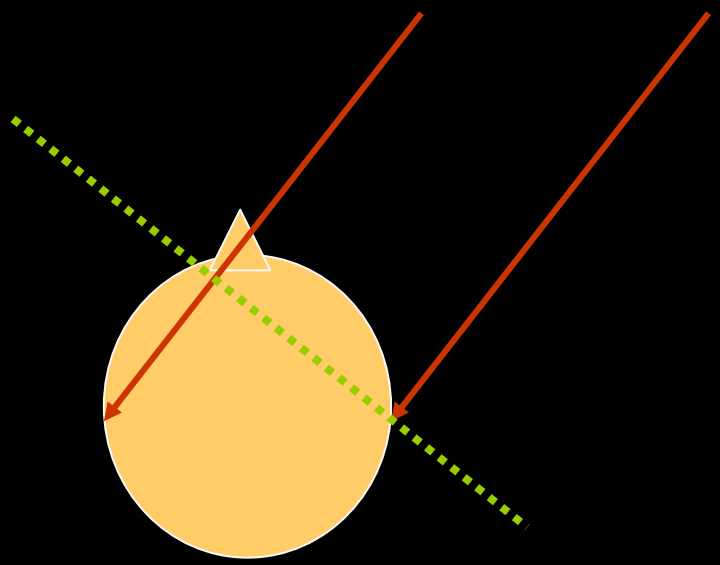
J. Eargle, Handbook of Recording Engineering

Spatial Cue: Binaural

- Binaural = hear using two ears
  - From the two signals we extract information about the location of sound sources.
- Difference in arrival time at ears
  - Interaural time difference (ITD) → azimuth
- Difference in intensity at ears
  - Interaural intensity difference (IID) → azimuth

Interaural Time Difference (ITD)

- Sound travels at a speed c around 343 m/s.
- Consider a sound wave from the right, the sound arrives at the right ear before the left.



Interaural Intensity Difference (IID)

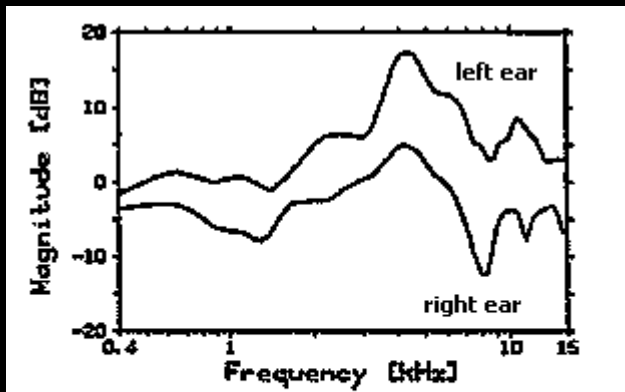
- Incident sound waves are diffracted by the head.
- IID is highly frequency dependent.
  - At low frequencies: The wavelength of the sound is long relative to the head diameter; there is hardly any difference in sound pressure at the two ears.
  - At high frequencies: The wavelength is short; there may well be a 20-dB or greater difference.

Spatial Cue: Binaural (cont.)

- Midline ambiguities
  - How to account for ability to hear sources as being front, behind, above, below?
- ITD and IID are not sufficient to precise location.
- The two pinnae act as variable filter that affects every sound that passes through them.
  - By this information the brain knows how to figure out the exact location of a sound in space.
  - In general the higher the frequency of a sound, the shorter its wavelength, and the better it can be localized

Binaural Cues

- Head-Related Transfer Function (HRTF)
  - Spatial filter characterizing interaction of sound waves with torso, shoulders, head, and pinna, based on source azimuth, elevation, and range



HRTFs of the left and right ears for a sound source at 30-degrees left of center.

source: E. MacPherson, A Computer Model of Binaural Localization for Stereo Imaging Measurement, JAES, September 1991.

Binaural Cues

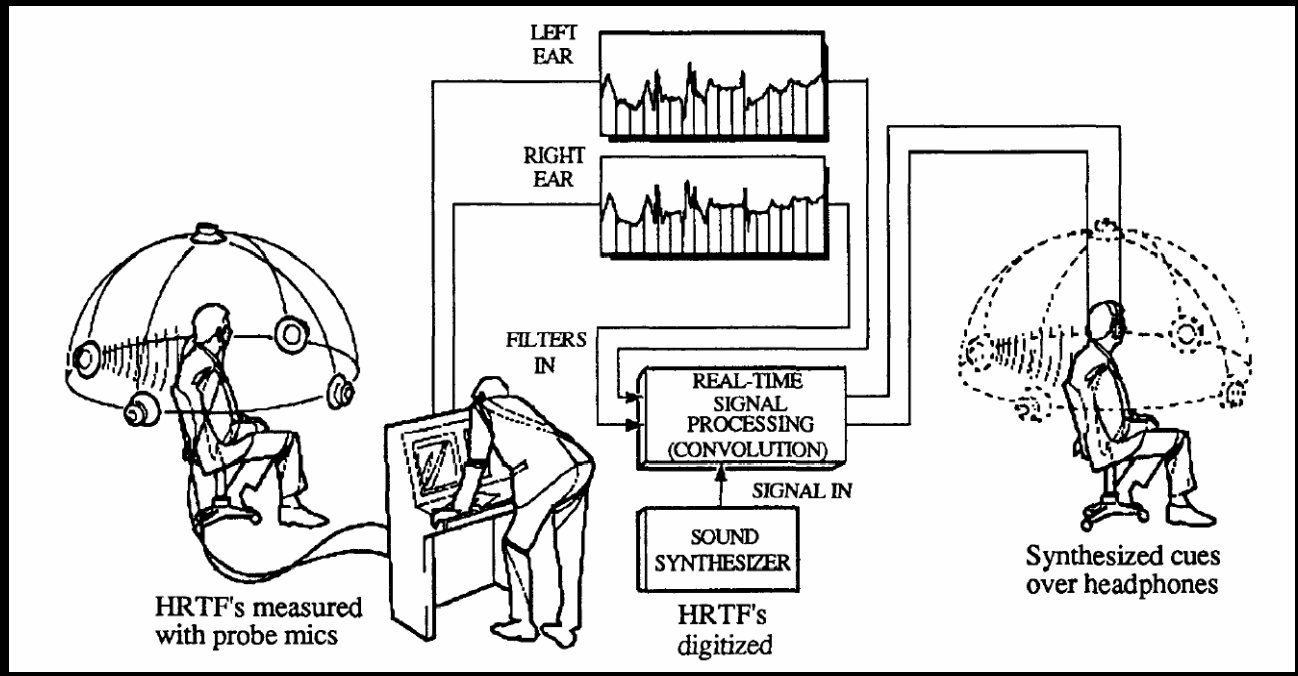
- Head-Related Transfer Function (HRTF)
  - Computed from recordings of localized sources with in-ear probe microphones

http://www.lce.hut.fi/~kar/pinna.jpg

interface.cipic.ucdavis.edu

Audio Spatialization

- Use HRTFs to process audio
  - Realism
  - But, doesn't account for environment



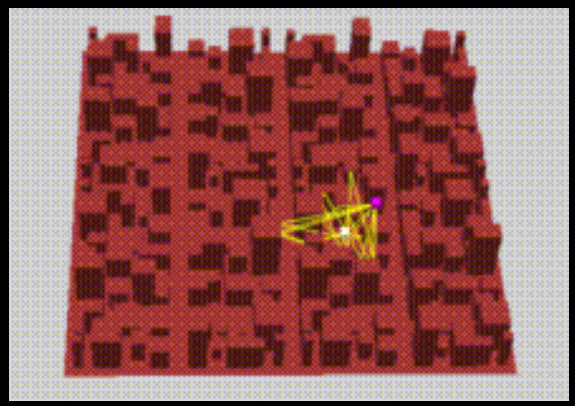
NASA Ames

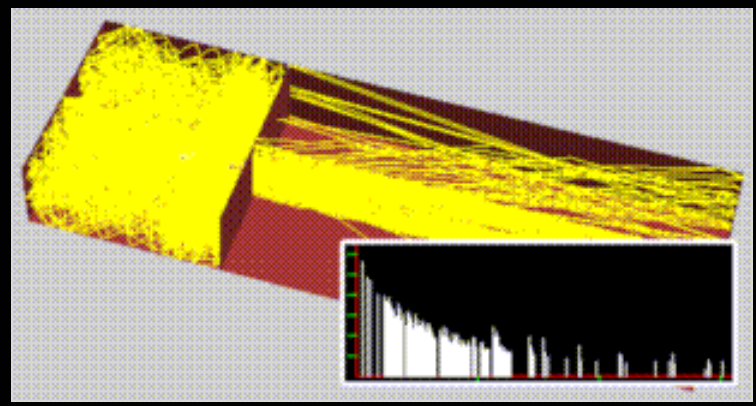
HRTF-Based Systems

- Able to produce elevation and range effects as well as azimuth effects.
- In practice, because of person-to-person differences and computational limitations, it is much easier to control azimuth than elevation or range.
- Becoming the standard for advanced 3-D audio interfaces

Sound Rendering

- Generate sound field by modeling interaction of sources with environment
- Usually defined to include potential application of HRTFs





http://www.cs.princeton.edu/~funk/rsas.html

Audio in 3D UIs

- Location cue
  - Accomplished through spatialization
- Realistic sound effects
- Sensory substitution
  - For example, to indicate physical contact
- .....

Speech in User Interfaces

- Speech recognition
- Speech synthesis
- Applications with speech techniques

Speech Recognition

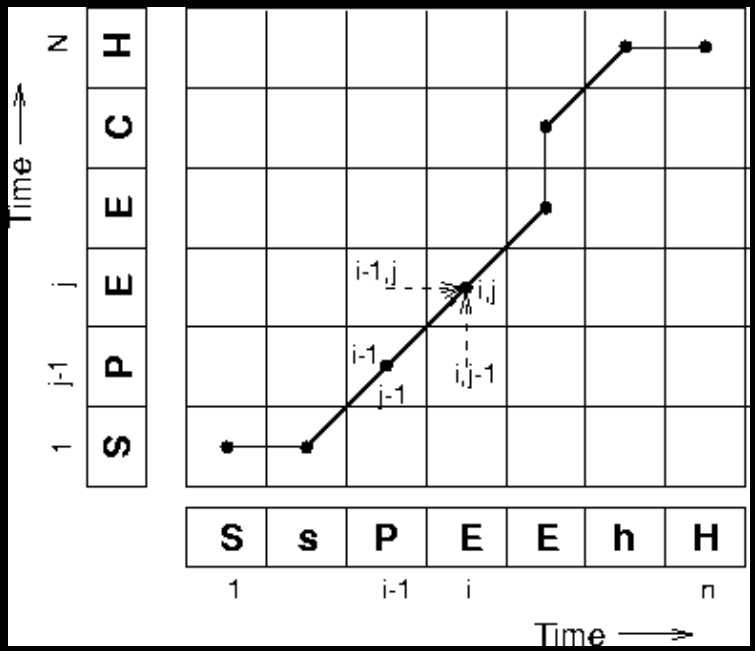
- Converting acoustic signal → a set of words.
  - Recognized words can be final results for commands and control, data entry, etc.
  - For further linguistic processing in order to achieve speech understanding.
- One of the ultimate goal of HCI research
  - very natural for communication
- More recently, the focus of the research shifted more towards multi–modal interfaces
  - e.g. combination of speech with gestures or keyboard.

Speech Recognition (cont.)

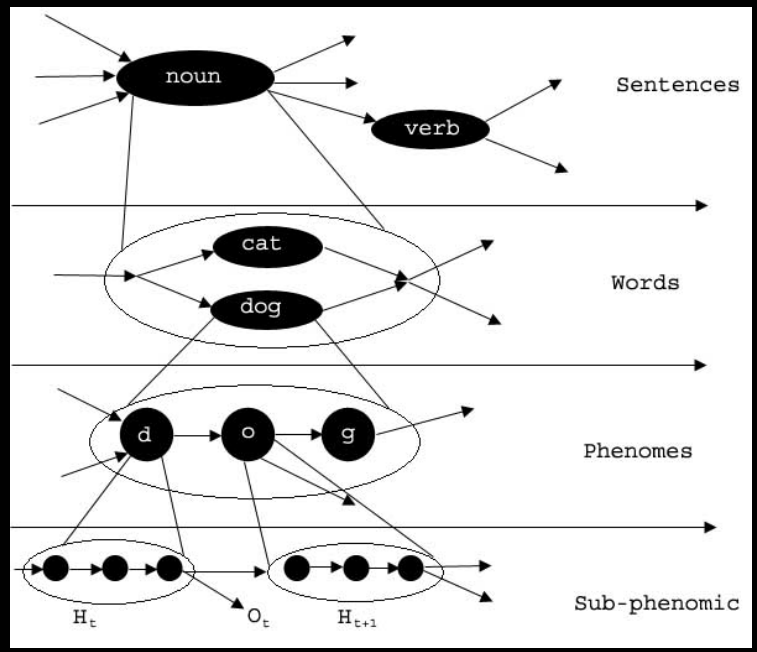
- Difficulties of speech recognition
  - Many sources of variability associated with signal
  - Phonemes are highly dependent on context
  - variability can result from changes in environment
- Either whole segments (words) are directly recognized (global method), or an intermediate phonetic labeling is used before lexical search

Speech Recognition (cont.)

- Template matching by dynamic time warping (DTW).
- Hidden Markov Model (HMM) for speech recognition.



www-lehre.inf.uos.de/~cmeyer/gmd/Do-Dateien/image002.gif



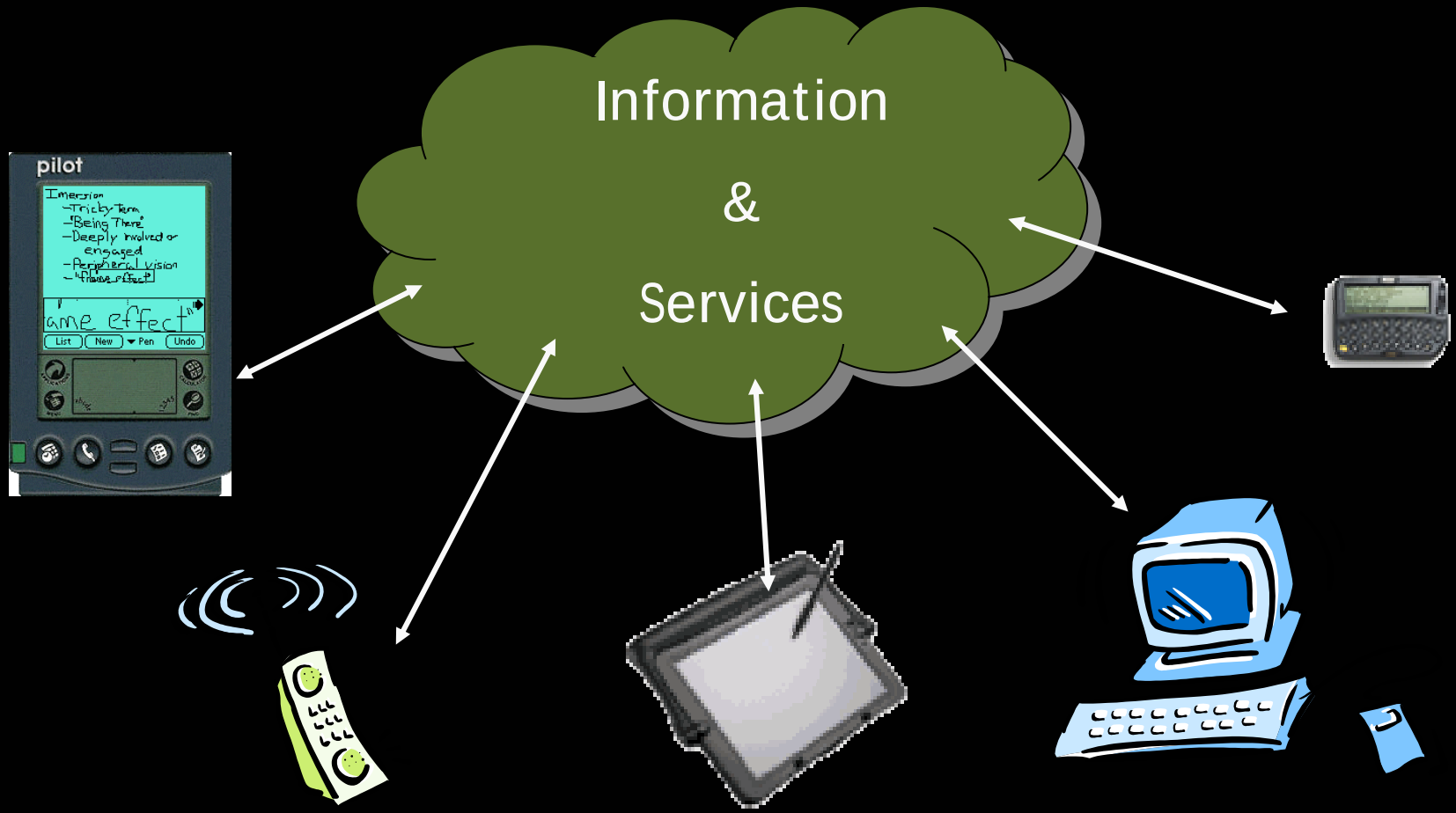
www.cs.wustl.edu/~sg/CS527_SP01/hmm3.jpg

Speech Synthesis

- Speech : a natural method of communication
- Especially helpful for disabled
- Text-to-Speech (TTS)
  - Concatenative approach: speech signal can be generated by concatenation of speech segments
  - Co-articulation effects.
  - Hardest part of speech synthesis: adding prosodic features of voice dynamics such as pitch, timing and amplitude.

Motivation for Speech UIs

- Pervasive information access



UIs in the Pervasive Computing

- Future computing devices won't have the same UI as current PCs
- Wide range of devices
  - Small or embedded in environment
- Smaller devices -> difficult I/O
- Freedom for other body parts
- Natural communication

Why are Speech UIs Hard to Get Right?

- Speech recognition far from perfect
- Speech UIs have no visible state
  - can't see what you have done before or what affect your commands have had
- Speech UIs are hard to learn
  - how do you explore the interface? how do you find out what you can say?

Recognition Problems

- Poor recognition
  - humans < 1% error rate on dictation
  - top recognition systems get 5-10% error rates
- Background noise
  - even worse recognition rates (20-40% error)
- Slow for higher accuracy
  - Segmentation
    - silly versus sill lea
  - Spelling
  - mail vs. male -> need to understand language

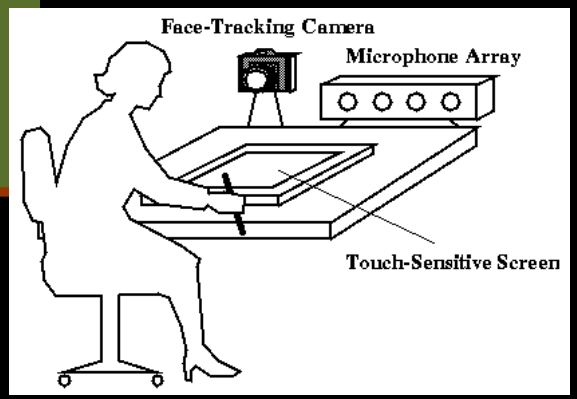
Speech UI Problems

- Speech UI no-nos
  - no feedback
    - certain commands only work when in specific states
  - deep hierarchies (aka voice mail hell)
- Verbose feedback wastes time/patience
  - only confirm consequential things
  - use meaningful, short cues
- Interruption
  - half-duplex communication (i.e., no barge-in support)
- Too much speech on the part of customer is tiring

- Speech takes up space in working memory
  - can cause problems when problem solving

Speech UIs

- Multimodal UIs address some of the problems with pure speech UIs.
  - help disambiguate
  - help w/ correction



www.cs.cmu.edu/afs/cs.cmu.edu/user/tue/www/papers/slt95/paper.html

www.research.att.com/~johnston/

www.miralab.unige.ch/

Resources

- Resources of speech recognition and synthesis
  - Microsoft Speech API 5.1
    - http://www.microsoft.com/speech/download/sdk51/
  - Sphinx, CMU
    - http://www.speech.cs.cmu.edu/#software
  - Viavoice, IBM
    - http://www-306.ibm.com/software/voice/viavoice/