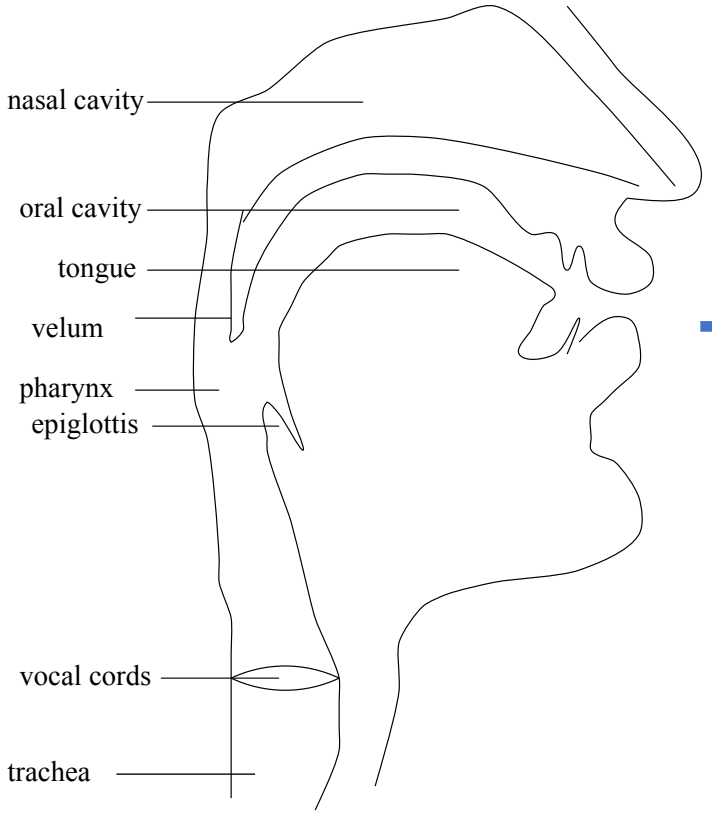
Automatic Speech Processing
EE-554

General Information

- Course, Textbook suggestions and lab notes available on moodle
- Practice questions
  - Purpose is to test the understanding of the main concepts
  - Feedback through hand correction
- Exam: written, without notes
- Contact: {mathew,petr.motlicek}@idiap.ch

What is speech?



Most common mode of communication

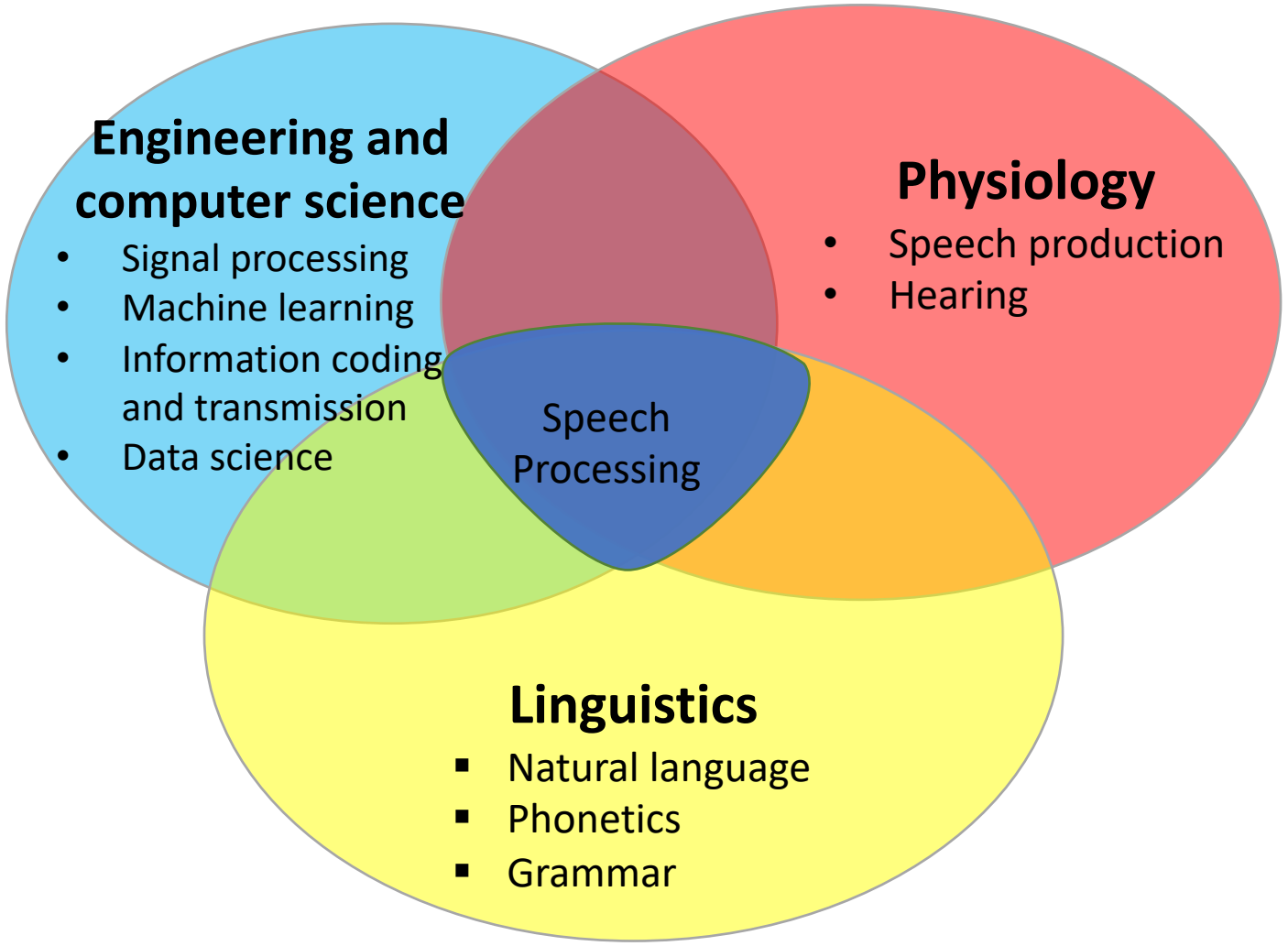
Information in speech

- A few day-to-day examples
  - “You sound happy.”
  - “Oh, I did not recognize your voice.”
  - “Can you say that again in English?”
  - “Did you hear what I said?”
  - “Oh, I can not hear clearly on the mobile, could you please call me on the landline?”
  - “He sounds like a French guy.”
  - “Oh that is a female voice.”
  - “It seems there is an old guy on the phone.”

Why speech processing? (1)

Emulate the ability of humans to speak and listen with machines

Why speech processing? (2)



Multi- and Inter-disciplinary

Why speech processing? (3)

"If I could make a current of electricity vary in intensity precisely as the air varies in density during the production of a speech sound, I should be able to transmit speech telegraphically."
Alexander Graham Bell

Opportunities to innovate and create

A few milestones related to speech transmission and storage

Milestone	Year
Commercial Telegraph	1844
Transatlantic Telegraph	1858
Telephone	1876
Phonograph	1877
Flat Disc	1887
Radio	Early 20th Century
Radio Program	1906
Transcontinental Telephone	1915
Transatlantic Radiotelephone	1927
Pulse Code Modulation (PCM)	1938
Speech and Hearing Research	1939

Milestone	Year
Transistor	1947
Information Theory	1948
$33\frac{1}{3}$ LP	1950
Transatlantic Telephonecable	1956
Stereo LP	1958
Commercial Digital Computers	1958
Communication Satellites	1962
Digital Telephone Transmission (using PCM)	1962
Integrated chips	1971
Singlechip Digital Signal Processing	1981

Milestone	Year
Compact Disc	1982
Cellular telephone	1984
Groupe Spécial Mobile (GSM)	1989
Digital Versatile Disk (DVD)	1995-1996
Voice over Internet Protocol (VoIP)	around 1996
3rd Generation (3G)	2001
In US, Western Union discontinued Telegraph	2006

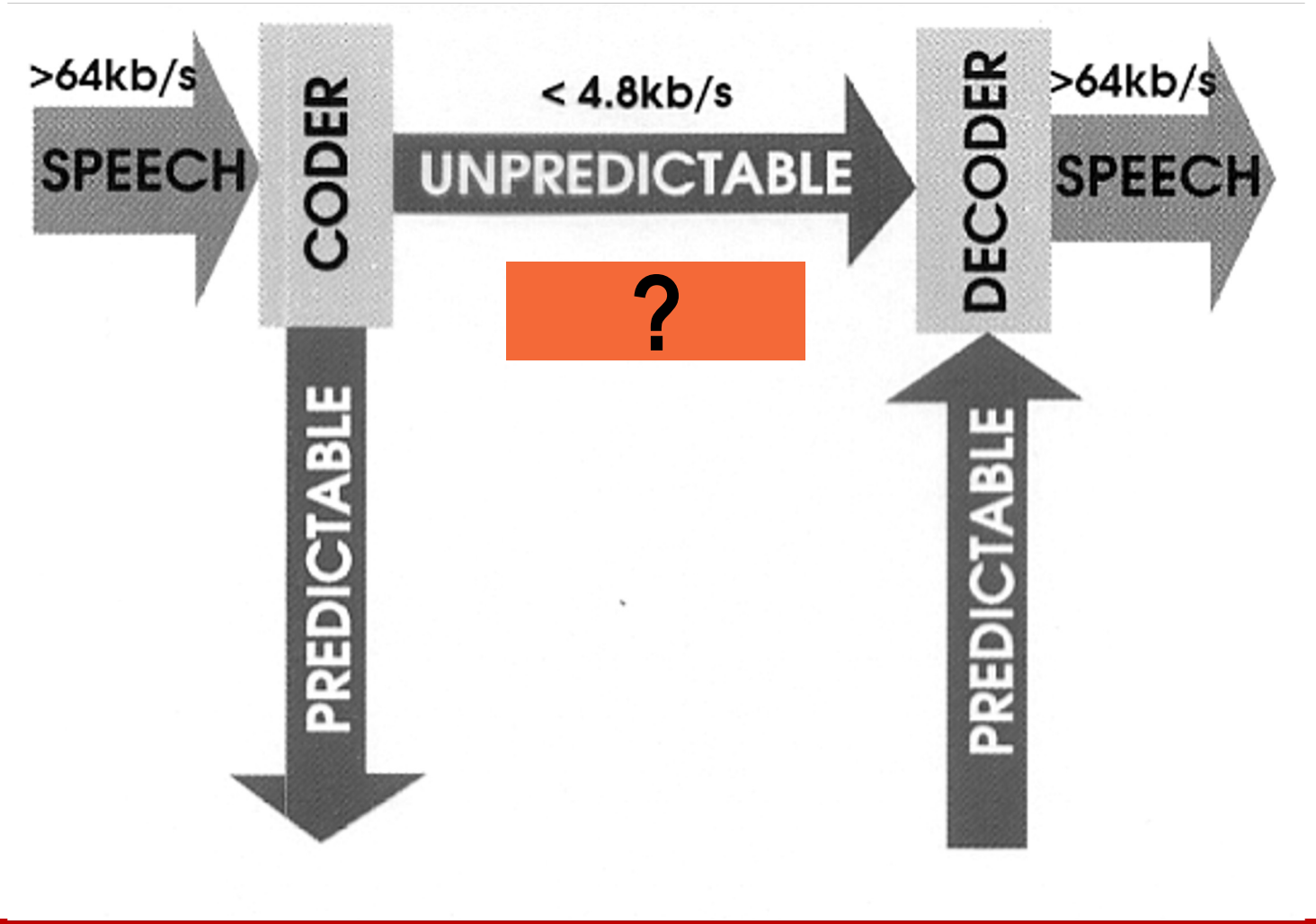
Course overview

- Introduction: speech processing, language engineering applications, speech science
- Fundamentals of speech signal processing: analysis and spectral properties, linear predictive coding
- Basic tools in statistical pattern recognition, Markov models, dynamic programming, hidden Markov models
- Automatic speech recognition
- Text-to-speech synthesis
- Speaker identification and verification

Lab exercises

- Lab. No. 1: Speech signal and its analysis, pitch, etc
- Lab. No. 2: Linear Predictive (LP) analysis of speech (source-system decomposition)
- Lab. No. 3: Statistical pattern recognition, clustering, EM algorithm, etc
- Lab. No. 4: Markov models and hidden Markov models

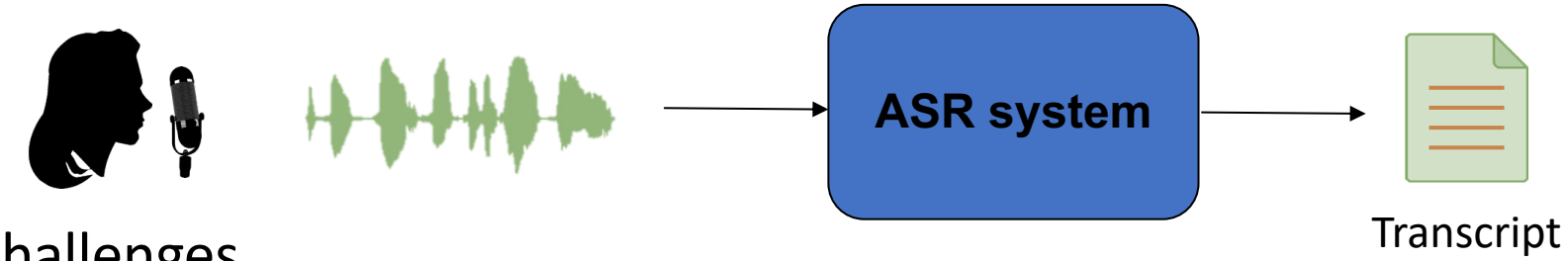
Speech Coding



Speech Coding (cont.)

- 16000 bps (adaptive pulse code modulation) 
- 13000 bps (multi-pulse LPC) 
- 8000 bps (code-excited LPC) 
- 2400 bps (LPC) 

Automatic Speech Recognition (ASR)



Challenges

- Reading vs. Conversation, Speaker-dependent vs. Speaker-independent
- Multilingual speech recognition (many languages are under-resourced)
- Noise-robust speech recognition
  - Distant speech recognition (e.g., on home devices Amazon Alexa, Google home, Apple HomePod)

Clean Reverberant Reverberant + noise

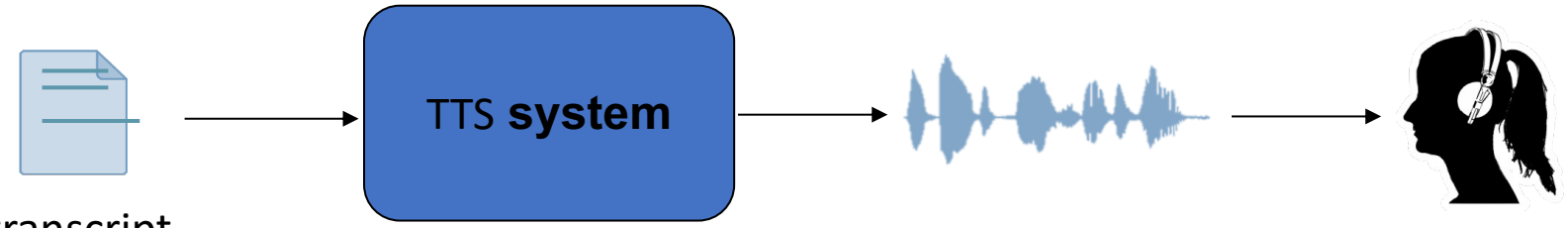
- Recognition of *atypical* speech, e.g. accented speech, children speech, impaired speech

Unimpaired speech Impaired speech (Hypernasal)

Automatic Speech Recognition (cont.)

- Office/desktop:
  - Voice control of PC/Workstations, of programs, dictation systems
- Manufacturing/Business:
  - Aid in manufacturing process, quality control, stock control and management
- Medical/Legal:
  - Creation of medical/legal reports, briefs, diagnostics...
- Others:
  - Games, aid to handicapped, interactive kiosk information systems

Text-to-Speech Synthesis (TTS)

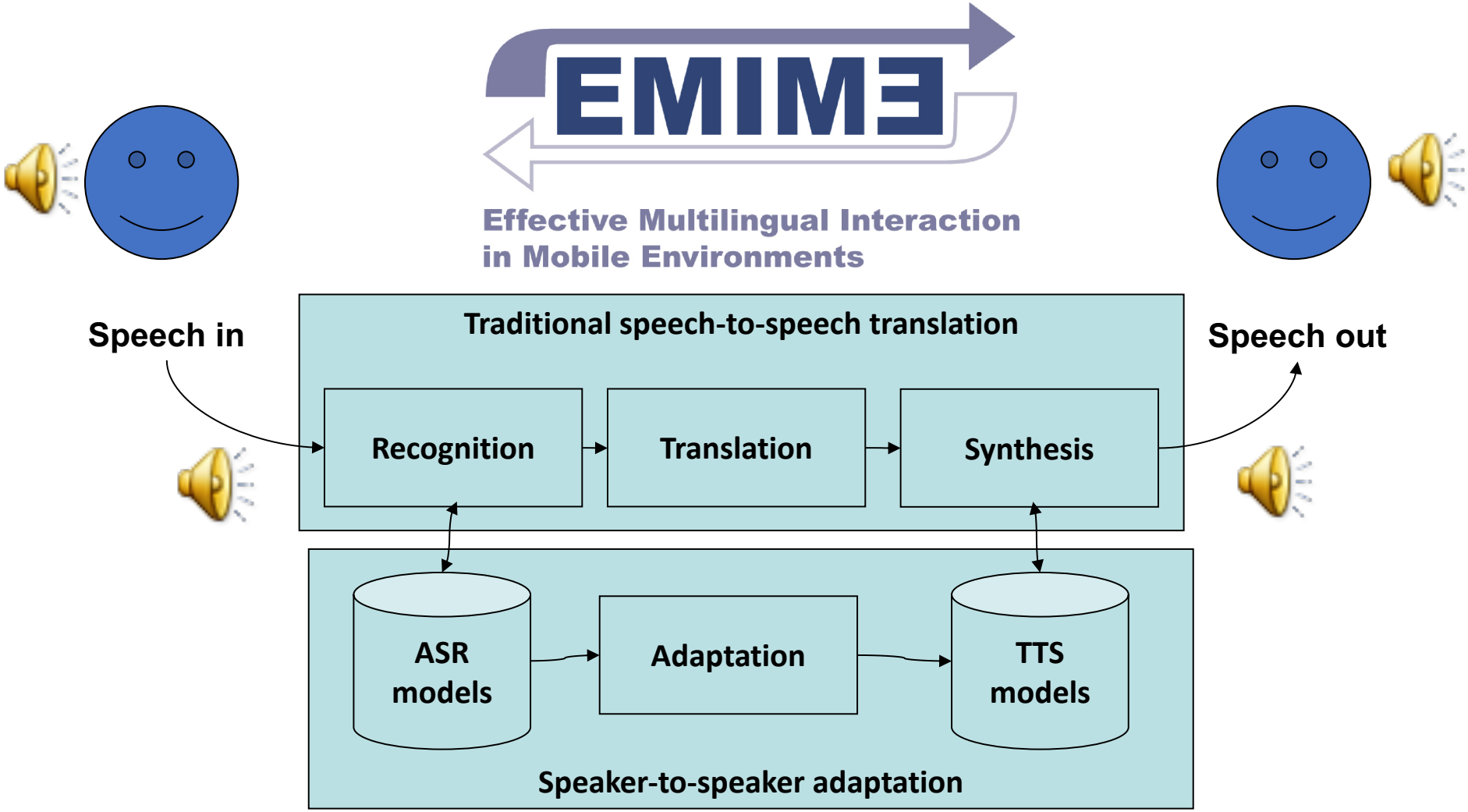


Text transcript

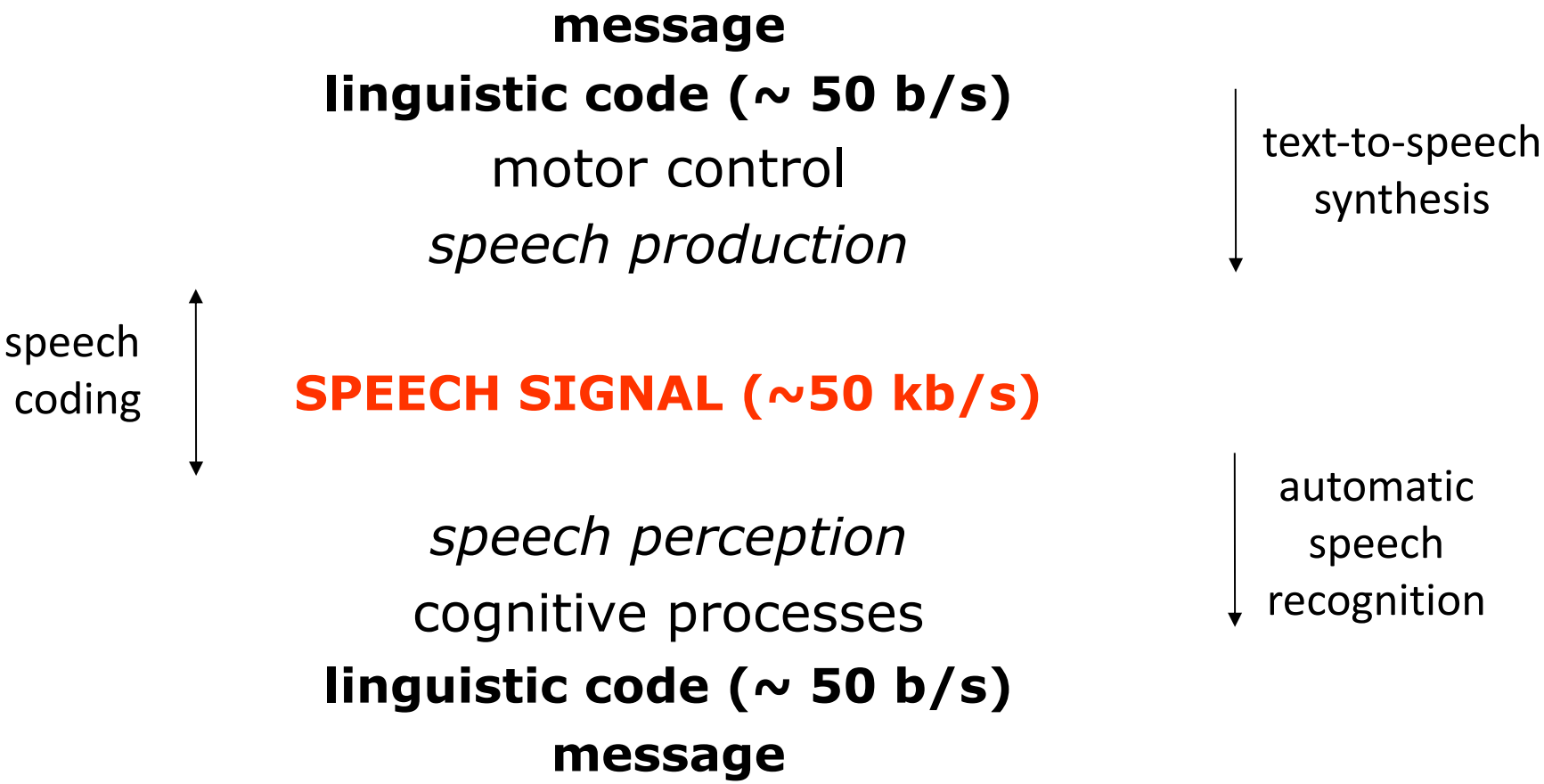
- Challenges
  - Fast adaptation to new speaker
  - Multilingual speech synthesis
  - Affective/Expressive speech synthesis
  - “Objective” evaluation
- End-use
  - Announcement systems
  - Dialog systems
  - Assistive systems for visually impaired and speech impaired persons
  - Voice banking

Stephen Hawkins, Cambridge Univ.

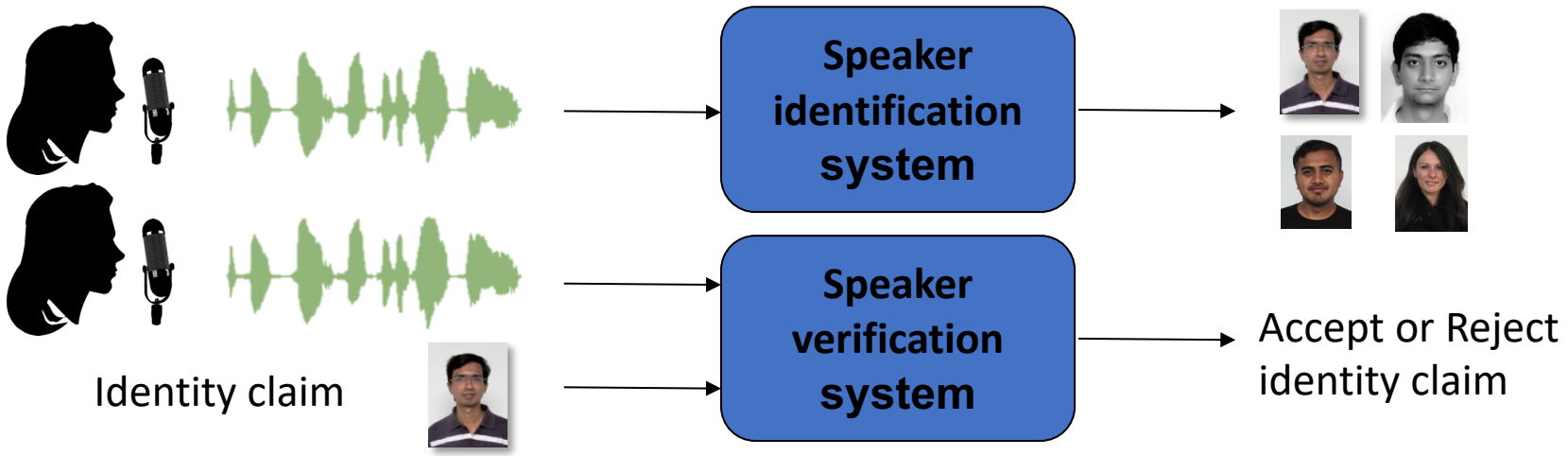
Speech-to-Speech Translation



Human Speech Communication

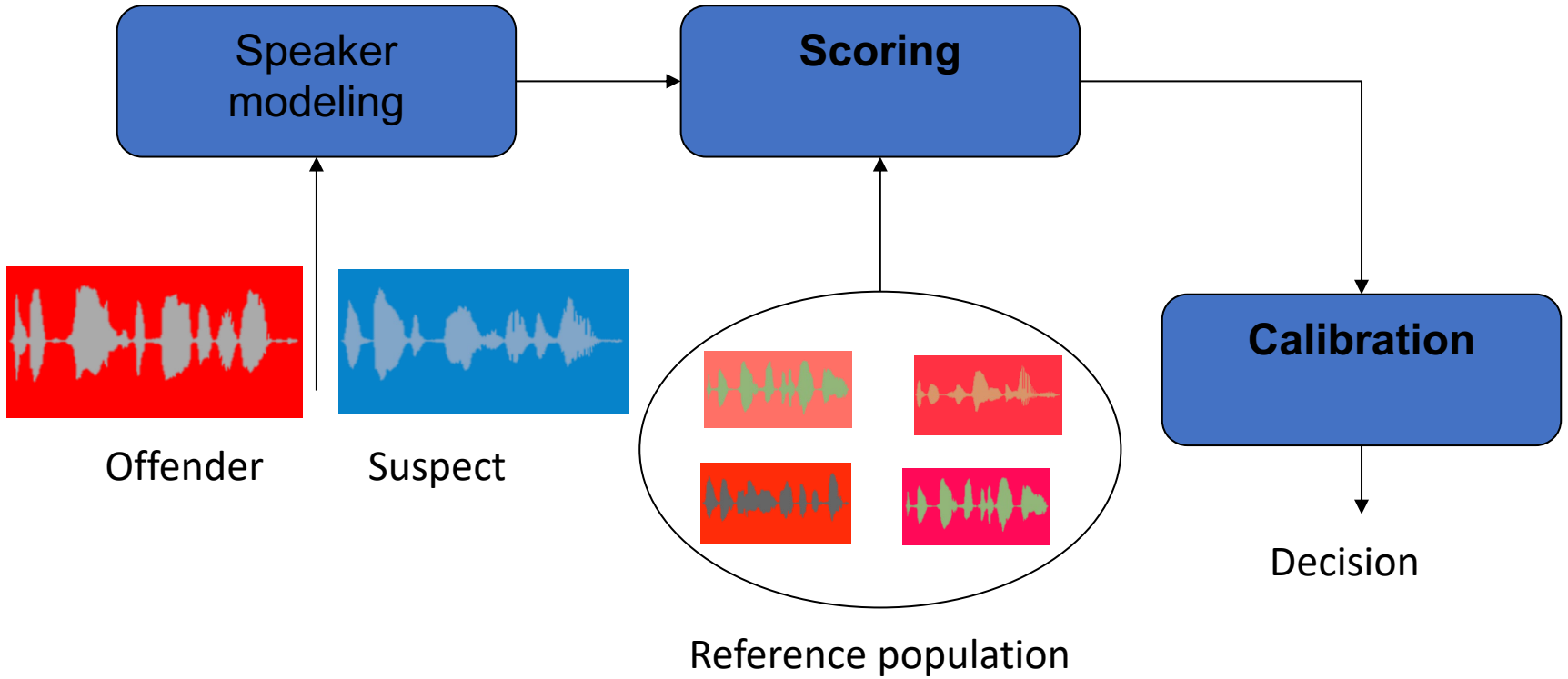


Automatic Speaker Recognition



- Challenges
  - Robustness to environment and channel variation
  - Language independence
  - Legally admissible forensic evidence
- End use
  - Secured access
  - Forensics

Forensic speaker recognition



Speech Assessment

Broad category of problems pertaining to speech communication with human in the loop

- Speech coding and transmission system assessment (ITU standards)

Change of codec and/or transmission channel *Average Listener*

- How intelligible is the transmitted speech?
- How good or bad is the transmitted speech quality?
- Quality of service?

- Language learning

Learner *Language expert*

- Degree of nativeness
- Fluency
- Pronunciation errors

- Clinical domain

Clinical expert

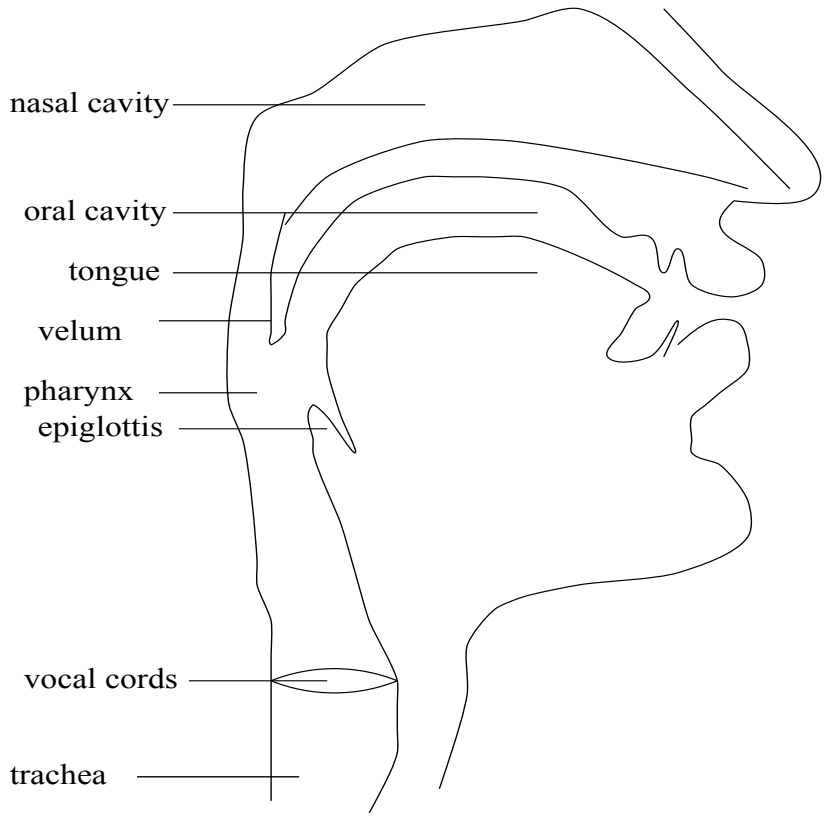
- Control versus Pathological
- Type of speech pathology

Speech Assessment (cont.)

- Drawback of human listener based assessment
  - Costly in terms of both time and money
  - Different expertise needed for different tasks
  - Difficult to reproduce

    Solution: Automatic prediction of human assessment
- Challenges
  - Developing models that are sensitive to the specific variabilities with limited data and knowledge
  - Exploiting existing speech technology components
  - Handling variabilities in human assessment

Speech Production

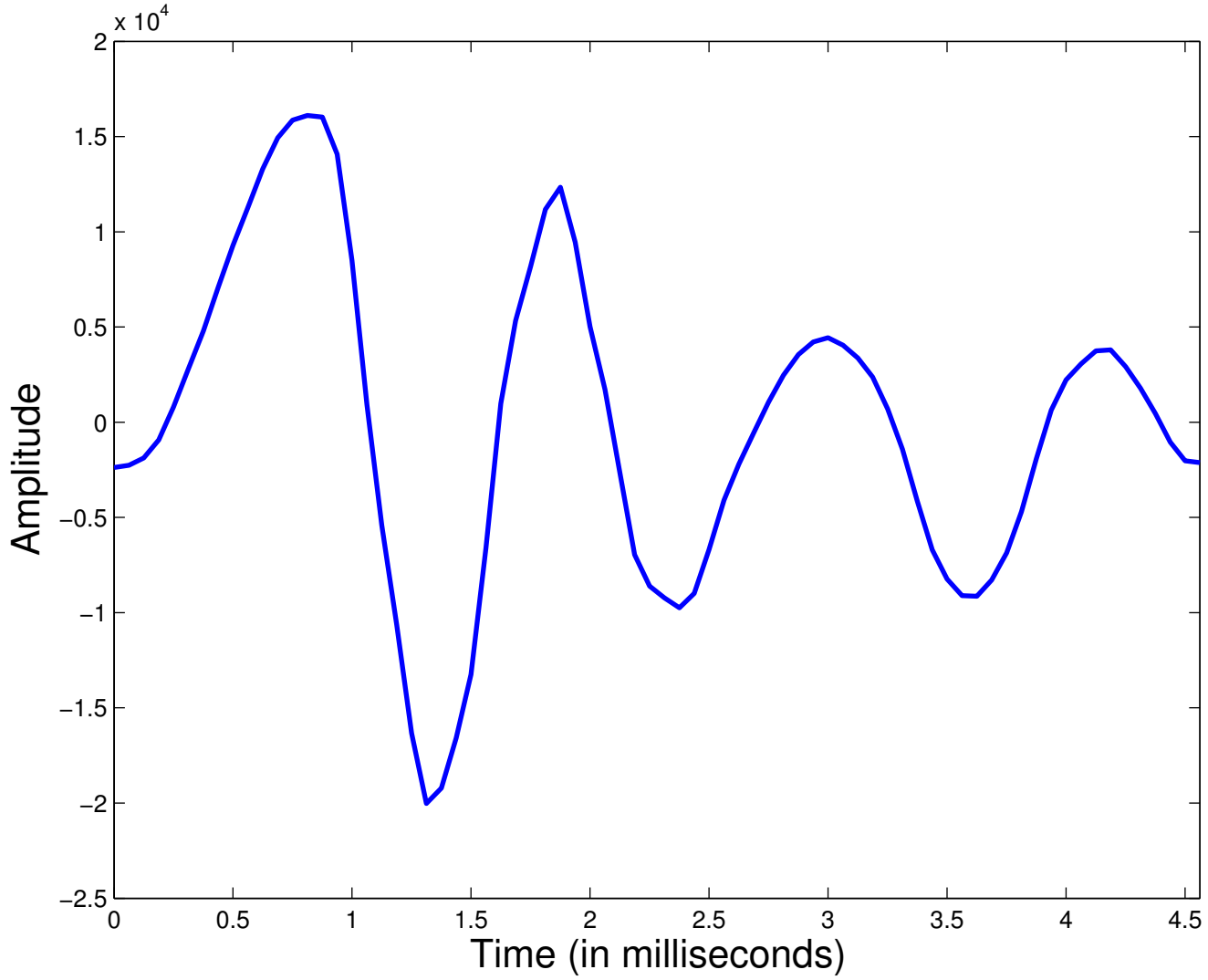


Excitation: vibration of vocal cords

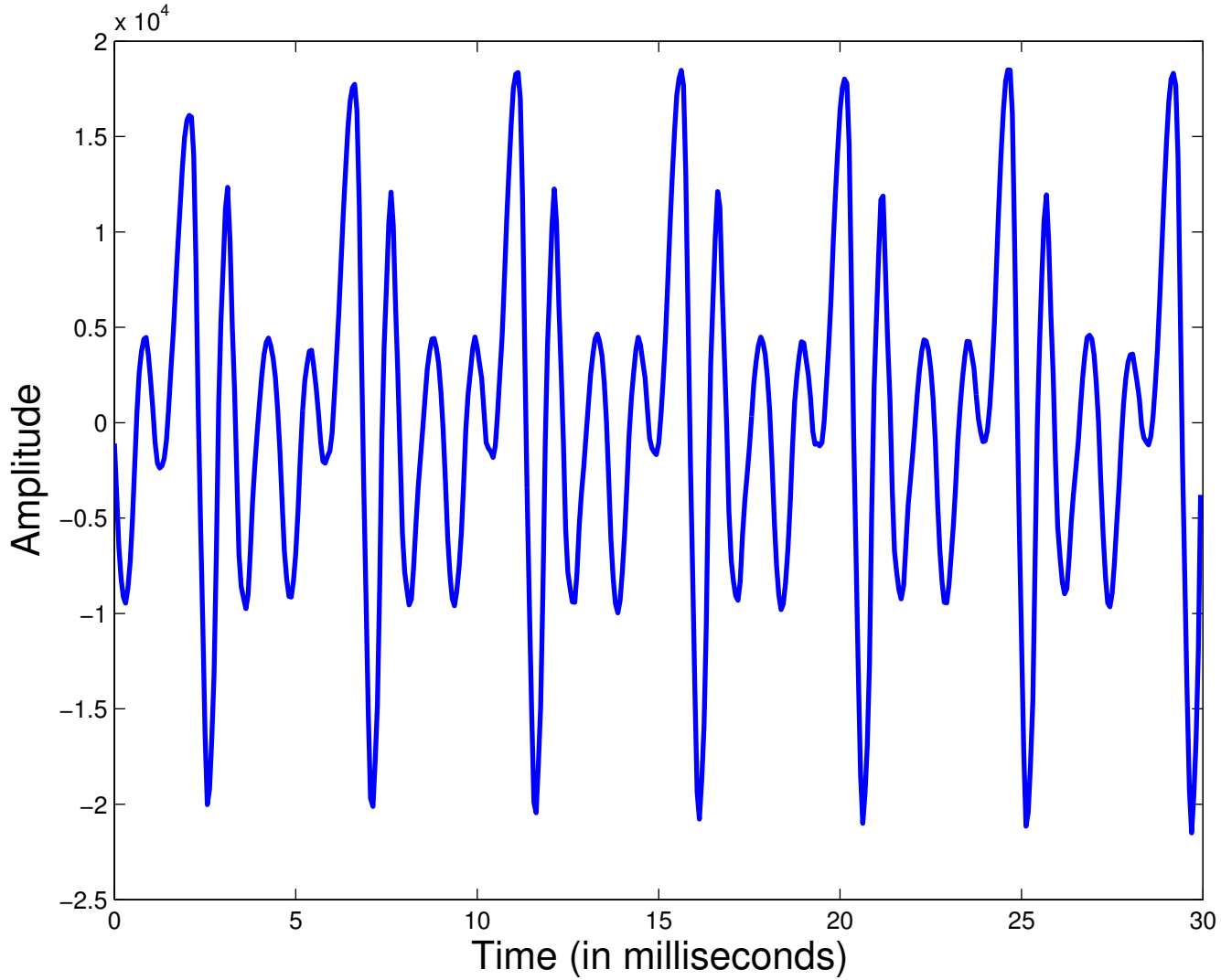
System: Vocal tract shape and sometimes nasal cavity

Response: Speech

One excitation



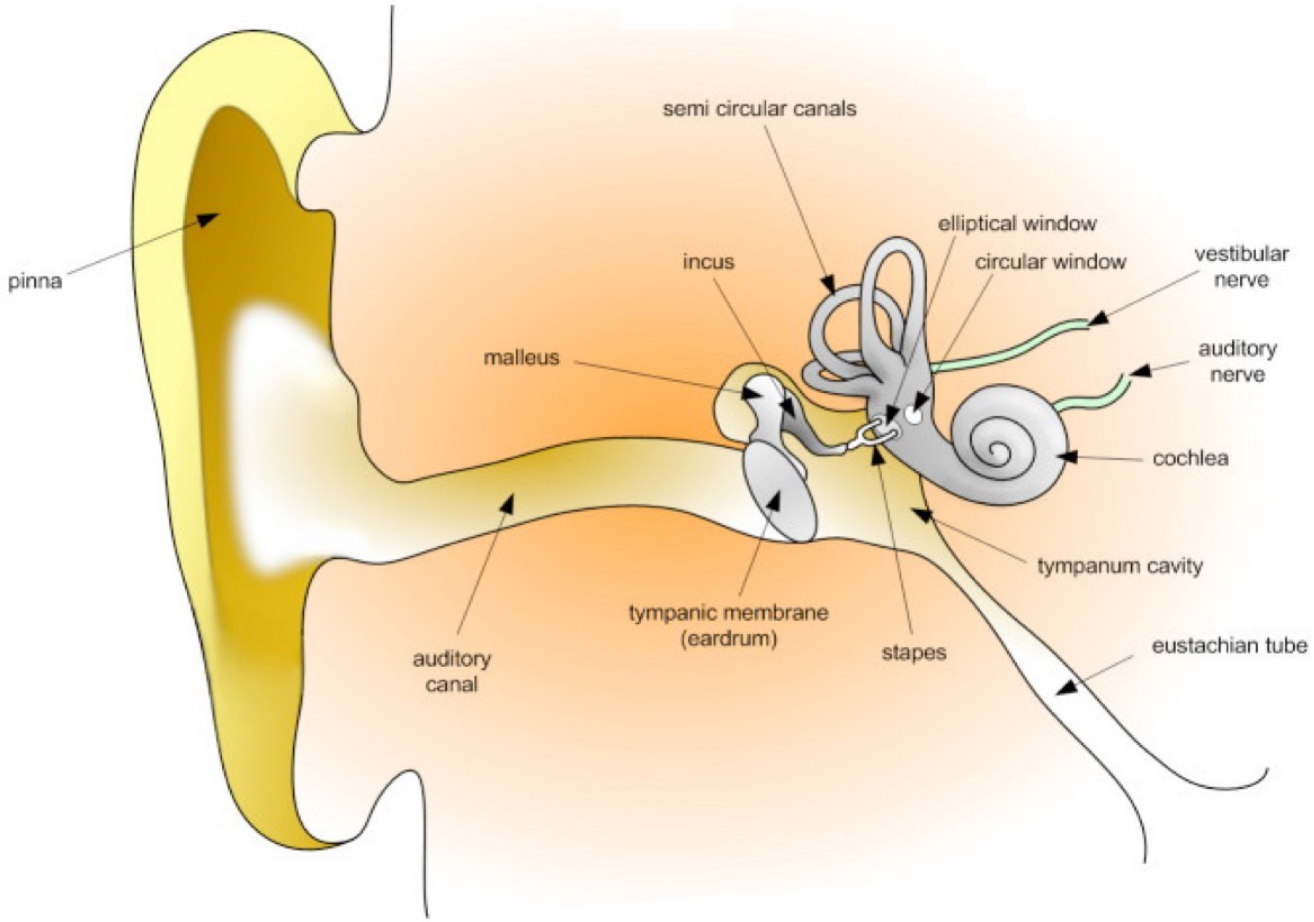
Six excitations



Speech Production (cont.)

- Rate of periodic vibration of vocal cords differs from person to person due to differences in vocal fold size
  children > female > male
- Vocal cords under tension tend to vibrate faster
- For certain sound productions there is no periodic vibration of vocal cords
- Length of vocal tract differs from person to person
  male > female > children
- Shape of vocal tract changes for different sounds

Human Hearing

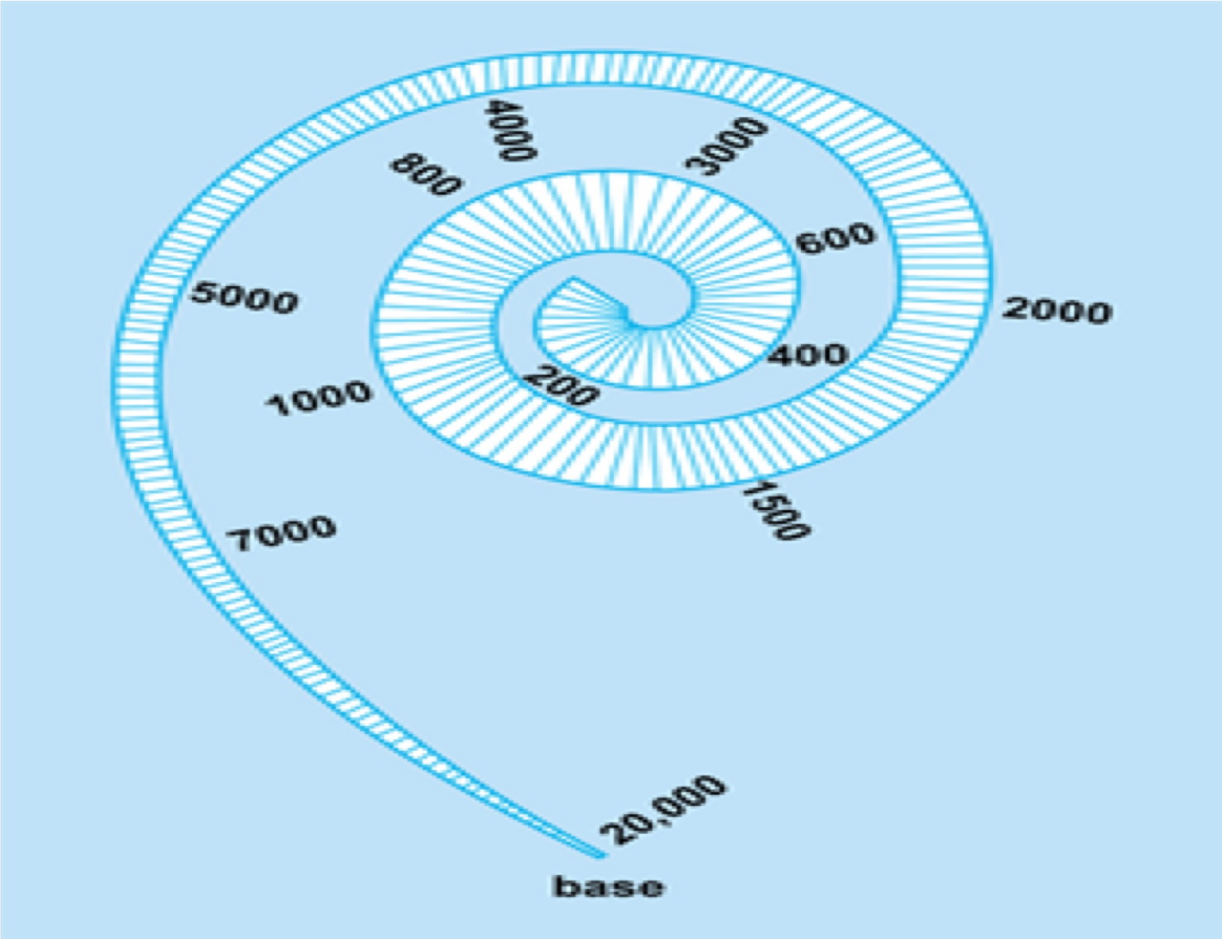


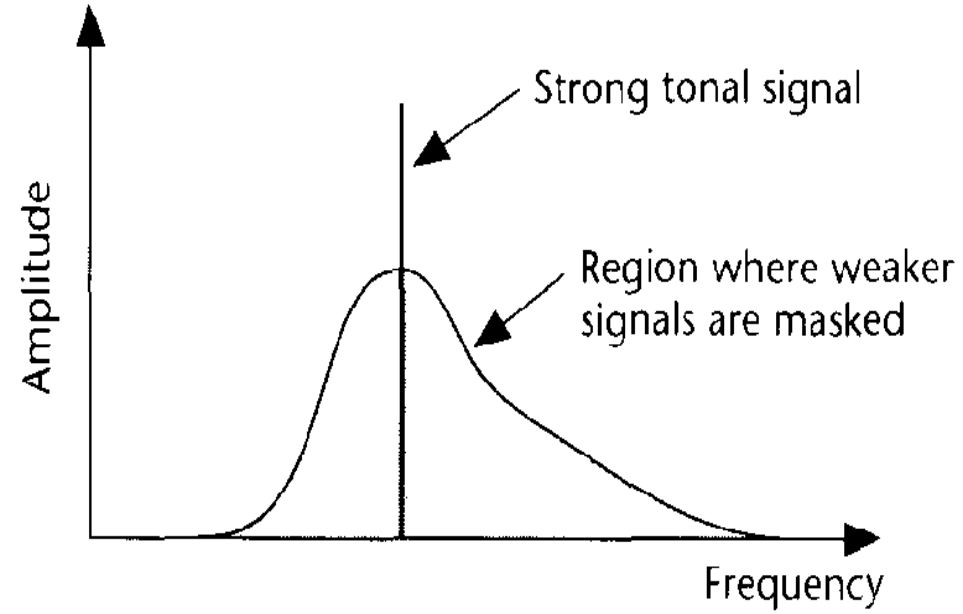
- Ear transforms sound's acoustic form of energy into nerve impulses
- Loudest sound that can be heard without feeling pain is about 120 dB
- Ear is not equally sensitive to all frequencies (highest sensitivity 1 kHz - 5 kHz)

Human Hearing: Sound pressure and loudness

Sound	Intensity N/m^2	Intensity level dB
Threshold of hearing	10^{-12}	0
Rustling Leaves	10^{-11}	10
Whisper	10^{-10}	20
Speech	10^{-6}	60
Orchestra	$6.3 \cdot 10^{-3}$	98
Walkman at Maximum Level	10^{-2}	100
Threshold of Pain	10^{1}	130
Instant Perforation of Eardrum	10^{4}	160

Human Hearing: Basilar membrane and critical bands





2000 Hz tone
CB ~ 280 Hz

Masked by broadband noise

Noise bandwidth 1000 Hz

Noise bandwidth 250 Hz

Noise bandwidth 10 Hz

Atomic unit of sound: Phoneme

- Phoneme: smallest sound unit that distinguishes meaning of words, e.g., *bed (/b/ /eh/ /d/) , bad (/b/ /ae/ /d/), bud (/b/ /ah/ /d/)*
- Phone: Several recognizably different sounds of phonemes
- Number of phonemes or phones in each language is different
  - English has about 42 phonemes
- Phonemes can be grouped as,
  - Vowels (no constriction in vocal tract) e.g., /aa/, /ey/, /iy/, /ow/, /uw/
  - Consonants
    e.g., /p/, /b/, /m/, /v/, /k/, /t/, /f/, /s/

Other groupings: Articulatory-based

- Manner of articulation
  e.g., Plosive: /p/, Fricative: /s/, Approximant: /er/
- Place of articulation
  e.g., Labial: /p/, Dental: /th/, Alevolar: /d/, Labio-dental: /v/,
- Palatal-Alevolar: /ch/, Lateral /l/
- Voicing
  - Voiced sounds (periodic vibration of vocal cords, has nonzero fundamental frequency) e.g., /aa/, /b/, /m/, /g/, /d/, /v/
  - Unvoiced sounds (no fundamental frequency) e.g., /p/, /k/, /t/, /f/, /s/
  - Aspirant sounds (mix of both) e.g., /hh/

Other groupings: Articulatory-based (cont.)

- Nasality
  e.g., /m/, /n/
- Roundedness
  e.g., /uw/
- Frontedness (only for vowels)
  e.g., Front: /ih/, Mid: /ah/, Back: /ao/
- Height of tongue (only for vowels)
  e.g., High: /ey/, Mid: /er/, Low: /ae/

See also International Phonetic Alphabet

Thank you for your attention!