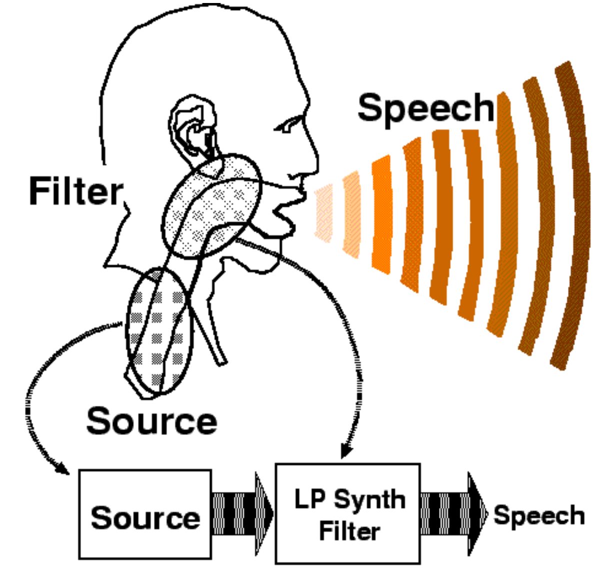


LP Synth Filter: Linear Prediction Synthesis Filter

Infocommunication
Speech Processing

Dr. Mohammed Salah Al-Radhi

Dr. Tamás Gábor Csapó

malradhi@tmit.bme.hu

Copyright

SPEECH PROCESSING, SPEECH TECHNOLOGY

Speech

- the most natural form of human-human communications
- related to language; linguistics is a branch of social science
- related to human physiological capability; physiology is a branch of medical science
- also related to sound and acoustics, a branch of physical science
- one of the most intriguing signals that humans work with every day

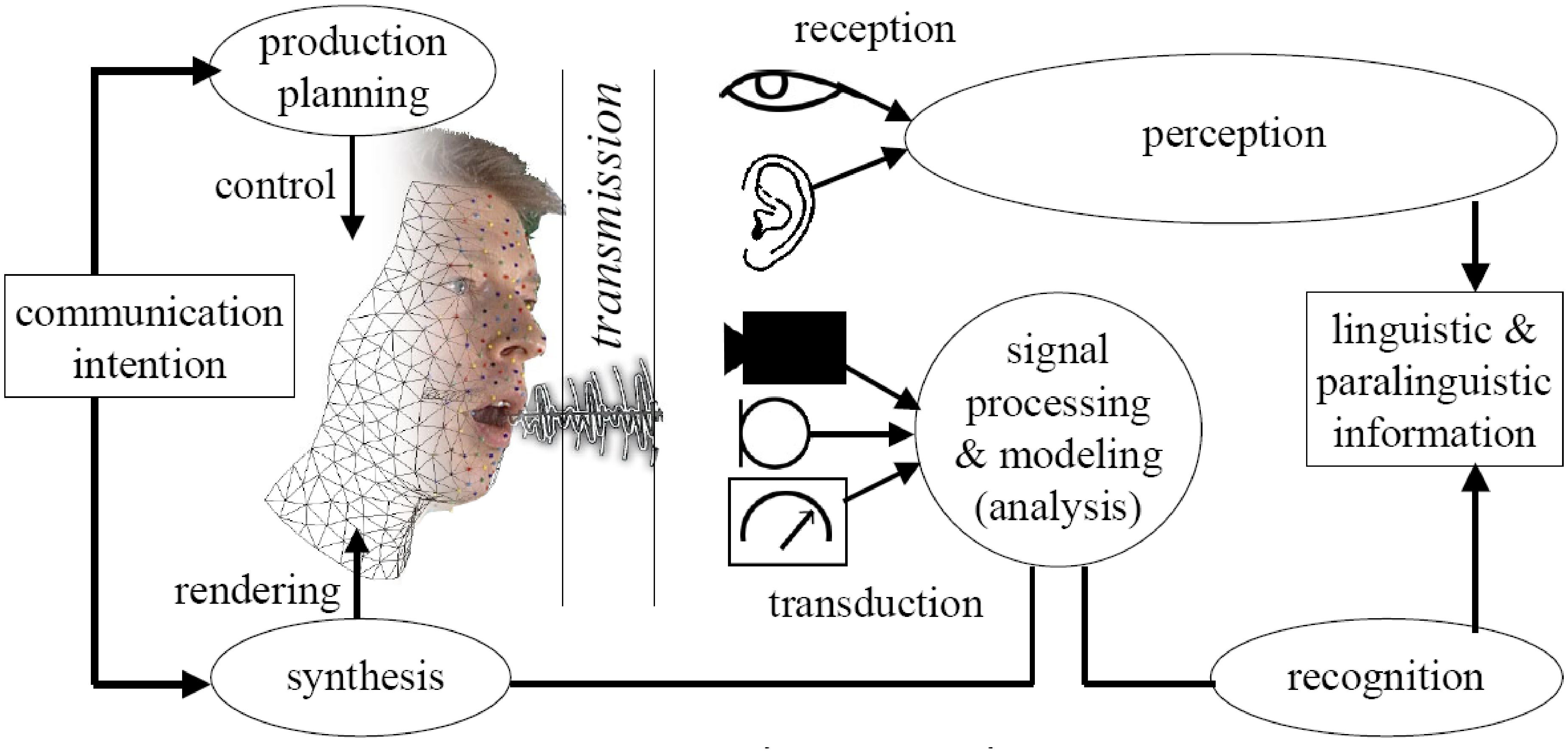
Source: Rabiner (2015) http://www.ece.ucsb.edu/Faculty/Rabiner/ece259/speech%20course.html

Speech processing

- Purposes:
  - to understand speech as a means of communication
  - to represent speech for transmission and reproduction
  - to analyze speech for automatic recognition and extraction of information
  - to discover some physiological characteristics of the talker

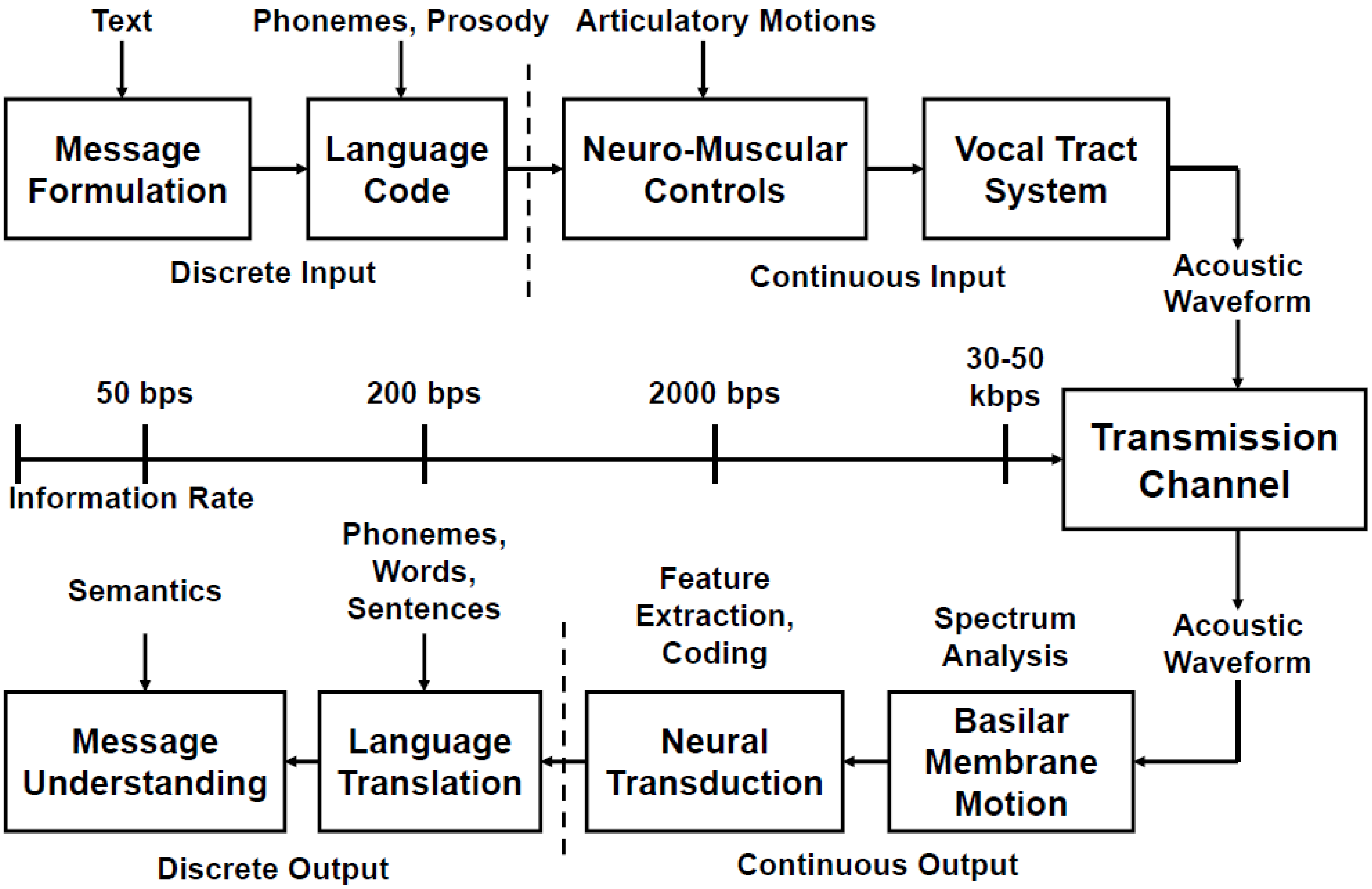
Source: Rabiner (2015) http://www.ece.ucsb.edu/Faculty/Rabiner/ece259/speech%20course.html

Speech technology

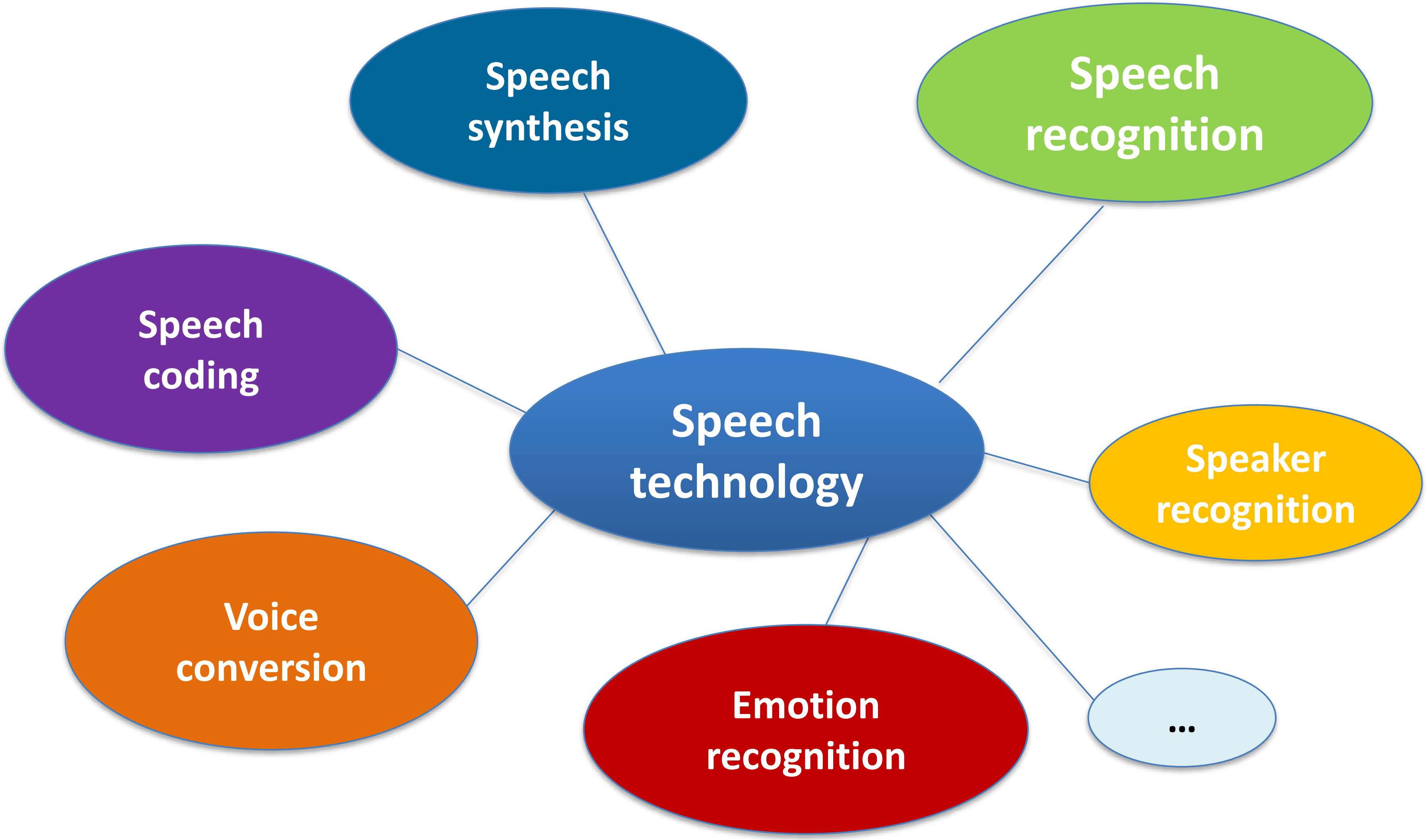


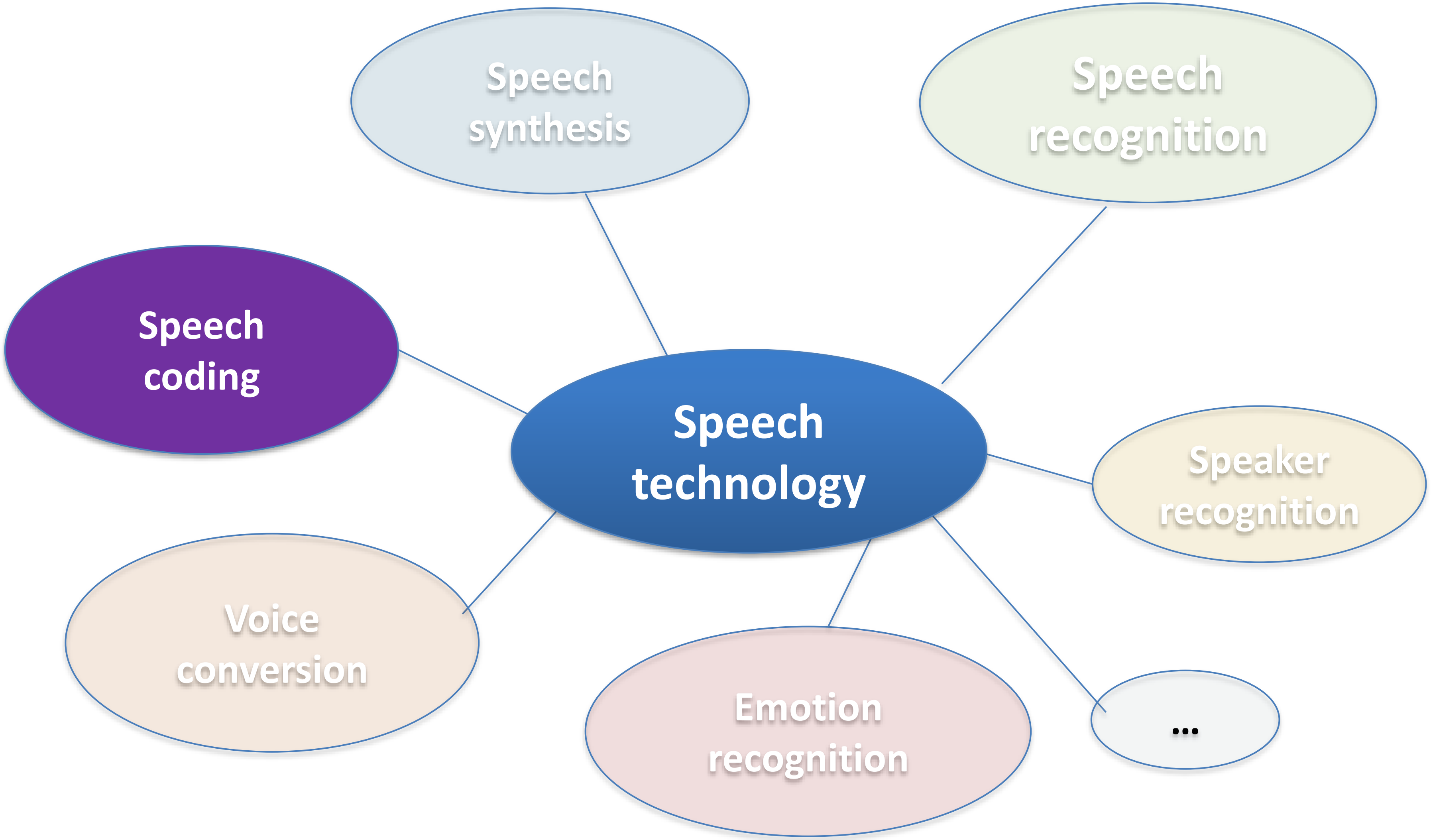
Source: Fagel (2007)

The Speech Chain



Source: Rabiner (2015) http://www.ece.ucsb.edu/Faculty/Rabiner/ece259/speech%20course.html


Speech synthesis
Speech recognition
Speech coding
Speech technology
Speaker recognition
Voice conversion
Emotion recognition
...



Speech technology

Speech synthesis

Speech recognition

Speech coding

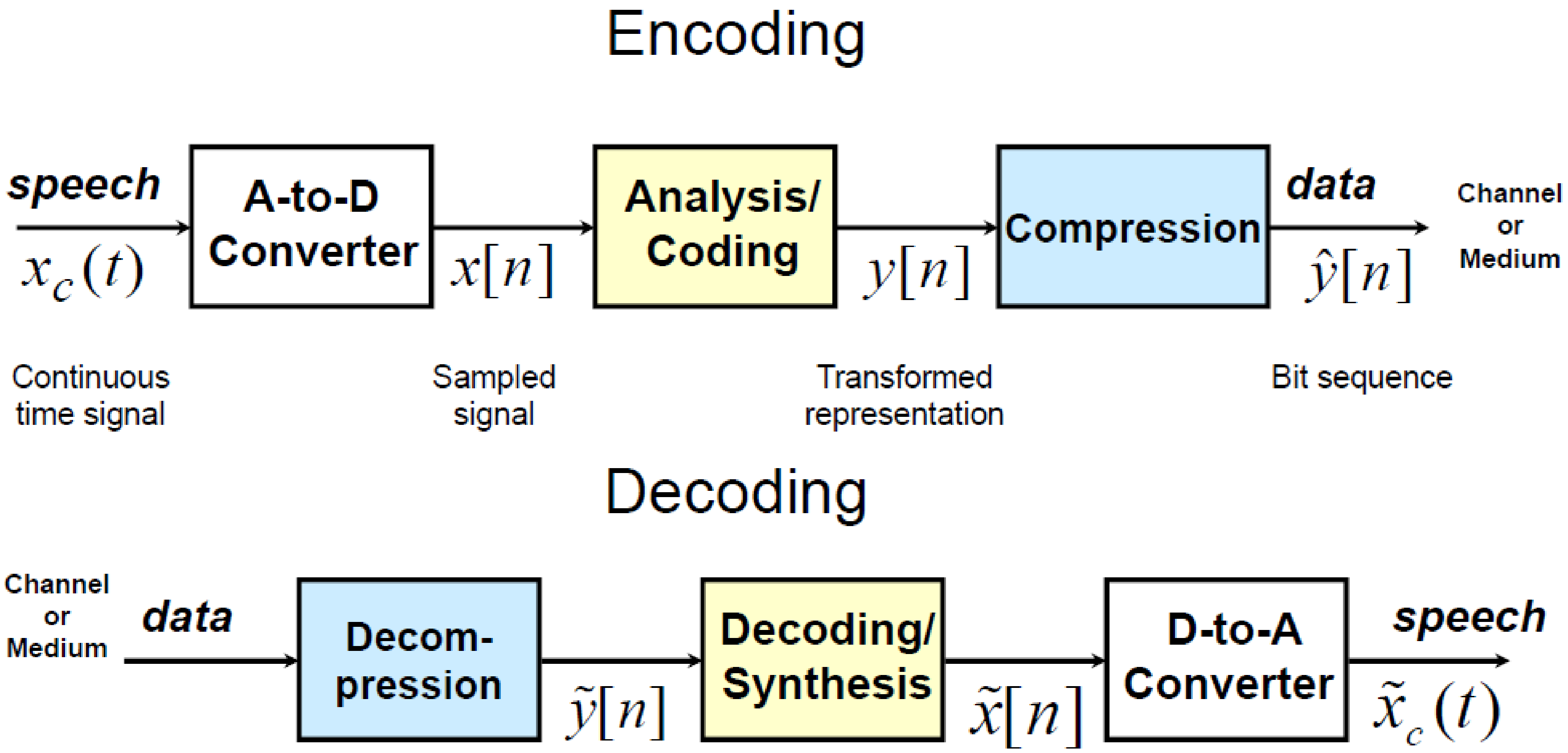
Speaker recognition

Voice conversion

Emotion recognition

...

Speech coding



Source: Rabiner (2015) http://www.ece.ucsb.edu/Faculty/Rabiner/ece259/speech%20course.html

Speech coding

- ***Speech Coding*** is the process of transforming a speech signal into a representation for efficient transmission and storage of speech
  - narrowband and broadband wired telephony
  - cellular communications (e.g. GSM, UMTS)
  - Voice over IP (VoIP) to utilize the Internet as a real-time communications medium
  - extremely narrowband communications channels, e.g., battlefield applications using HF radio
  - storage of speech for telephone answering machines, IVR systems, prerecorded messages

Source: Rabiner (2015) http://www.ece.ucsb.edu/Faculty/Rabiner/ece259/speech%20course.html

Information Rate of Speech

- from a Shannon view of information
  - message content/information--2**6 symbols (phonemes) in the language; 10 symbols/sec for normal speaking rate => **60 bps** is the equivalent information rate for speech
- from a communications point of view
  - speech bandwidth is between 4 (telephone quality) and 8 kHz (wideband hi-fi speech)—need to sample speech at between 8 and 16 kHz, and need about 8 (log encoded) bits per sample for high quality encoding => 8000x8=64000 bps (telephone) to 16000x8=128000 bps (wideband)

Source: Rabiner (2015) http://www.ece.ucsb.edu/Faculty/Rabiner/ece259/speech%20course.html

Information Rate of Speech

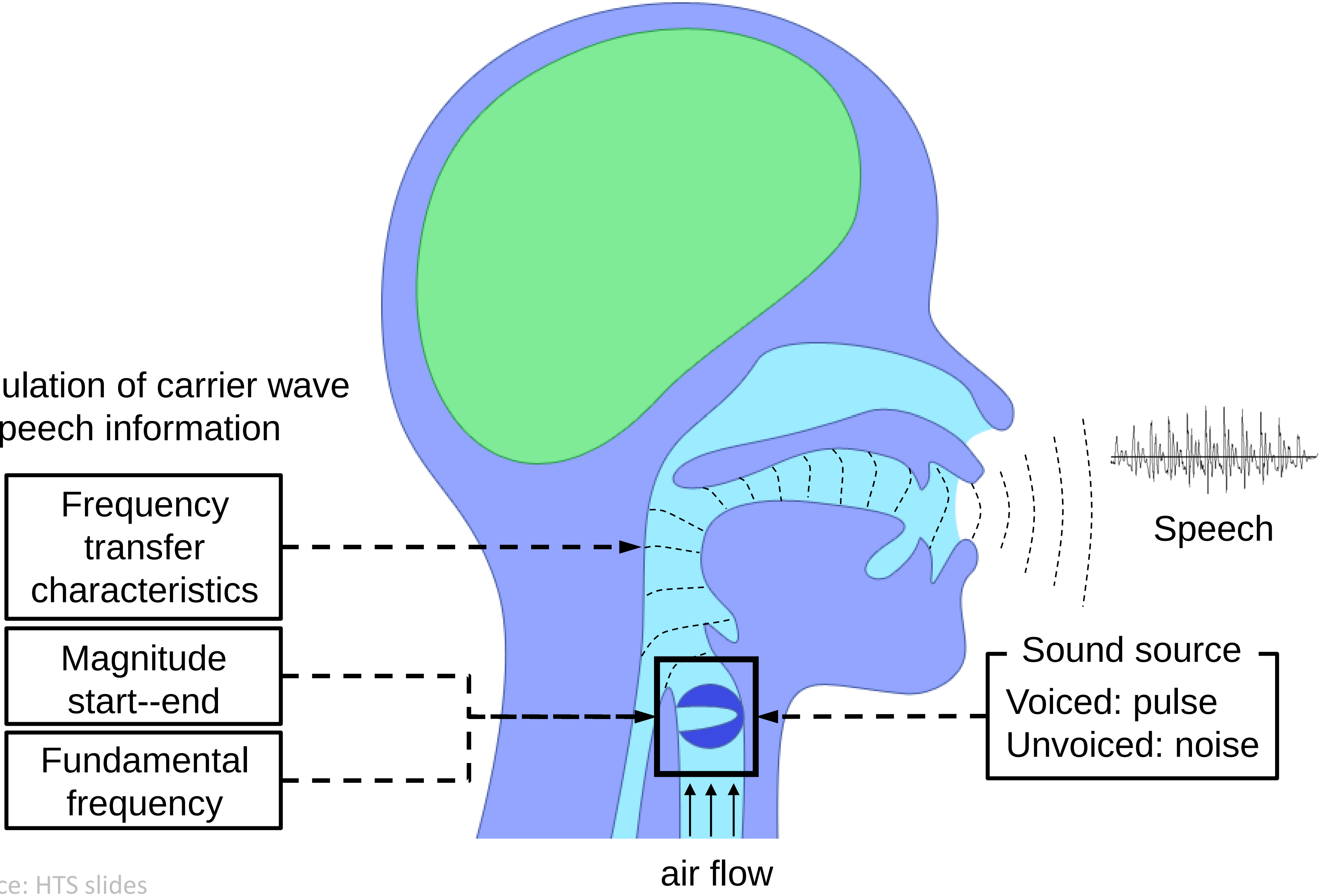
- from a Shannon view of information
  - message content/information--2**6 symbols (phonemes) in the language; 10 symbols/sec for normal speaking rate => **60 bps** is the equiv
- from a
  - spee Hz (wideband hi-fi speech)—need to sample speech at between 8 and 16 kHz, and need about 8 (log encoded) bits per sample for high quality encoding => 8000x8=64000 bps (telephone) to 16000x8=128000 bps (wideband)

1000-2000 times change in information rate from discrete message symbols to waveform encoding => can we achieve this three orders of magnitude reduction in information rate on real speech waveforms?

Source: Rabiner (2015) http://www.ece.ucsb.edu/Faculty/Rabiner/ece259/speech%20course.html

Speech production mechanism

Modulation of carrier wave
by speech information



Frequency transfer characteristics

Magnitude start--end

Fundamental frequency

Speech

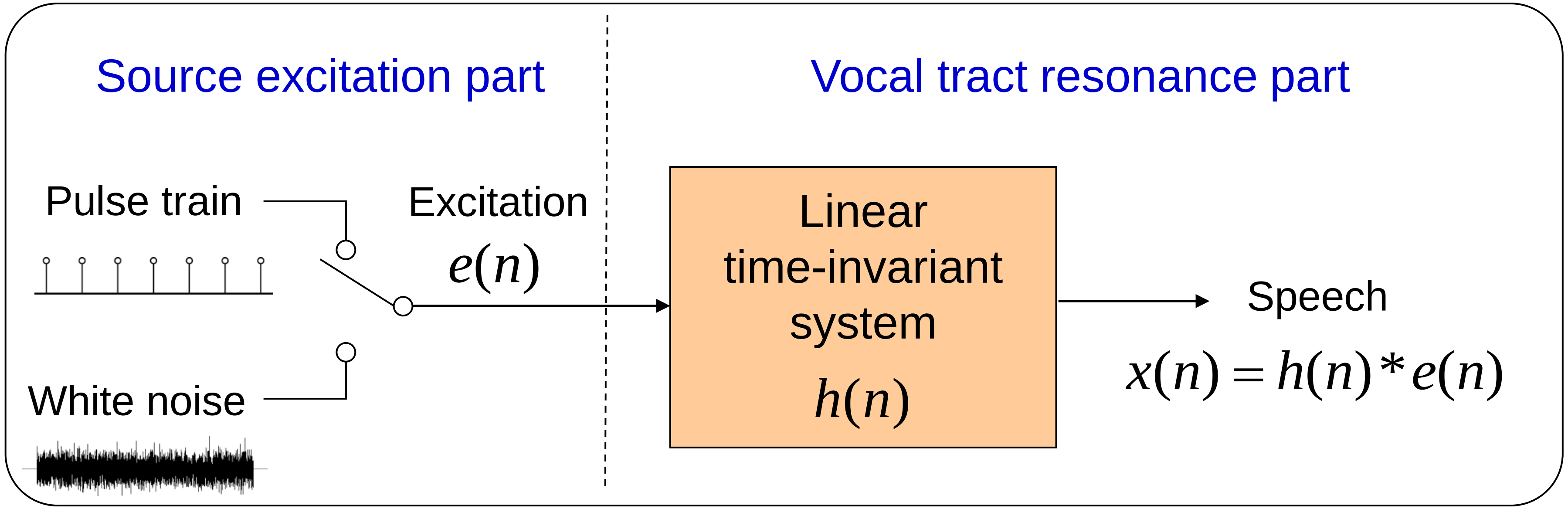
Sound source

Voiced: pulse
Unvoiced: noise

air flow

Source: HTS slides

Source-filter model



Source excitation part

Vocal tract resonance part

Pulse train

White noise

Excitation

$e(n)$

Linear time-invariant system

$h(n)$

Speech

$x(n) = h(n) * e(n)$

$$x(n) = h(n) * e(n)$$

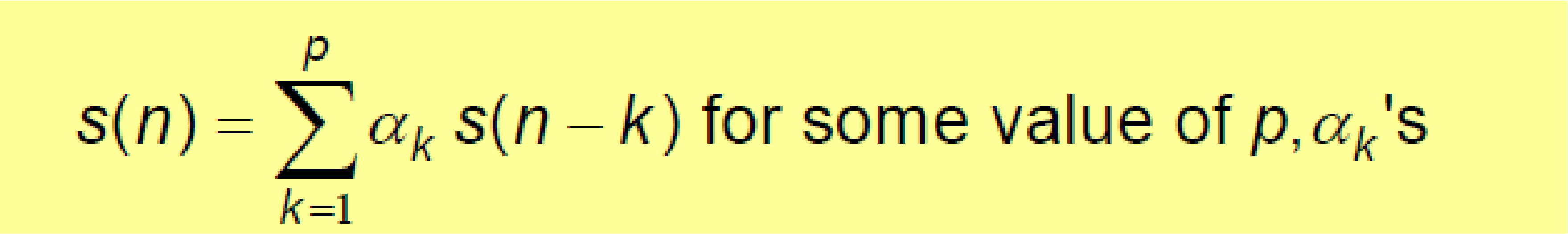
$$\Big\downarrow \ \text{Fourier transform}$$

$$X(e^{j\omega}) = H(e^{j\omega}) E(e^{j\omega})$$

Source: HTS slides

Linear Predictive Coding (LPC)

- LPC methods provide extremely accurate estimates of speech parameters, and does it extremely efficiently
- Basic idea of Linear Prediction: current speech sample can be closely approximated as a linear combination of past samples, i.e.,



$$s(n) = \sum_{k=1}^{p} \alpha_k \, s(n-k) \text{ for some value of } p, \alpha_k\text{'s}$$

Source: Rabiner (2015) http://www.ece.ucsb.edu/Faculty/Rabiner/ece259/speech%20course.html

LPC methods /1

- for periodic signals with period *Np*, it is obvious that

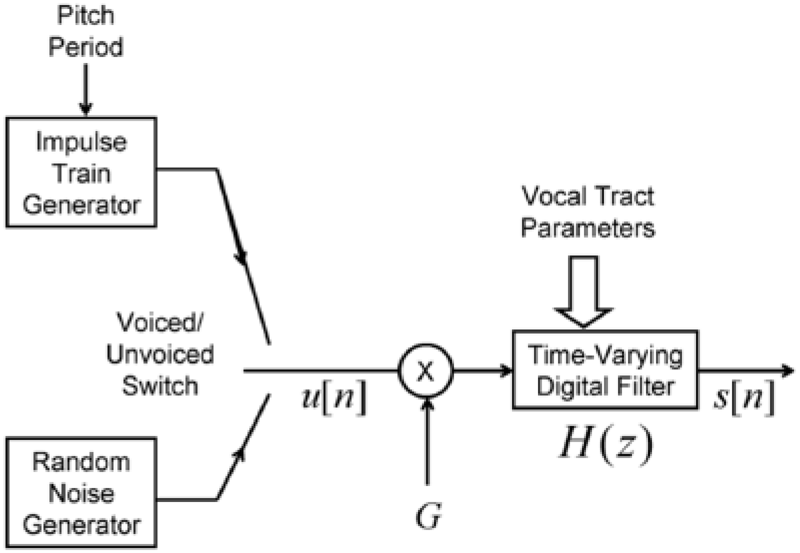
$$s(n) \approx s(n - N_p)$$

- but that is not what LP is doing; it is estimating *s(n)* from the *p (p << Np)* most recent values of *s(n)* by linearly predicting its value
- for LP, the predictor coefficients (the *αk* 's) are determined (computed) by ***minimizing the sum of squared differences*** (over a finite interval) ***between the actual speech samples and the linearly predicted ones***

Source: Rabiner (2015) http://www.ece.ucsb.edu/Faculty/Rabiner/ece259/speech%20course.html

LPC methods /2

- LP is based on speech production and synthesis models
  - speech can be modeled as the output of a linear, time-varying system, excited by either quasi-periodic pulses or noise;
  - assume that the model parameters remain constant over speech analysis interval

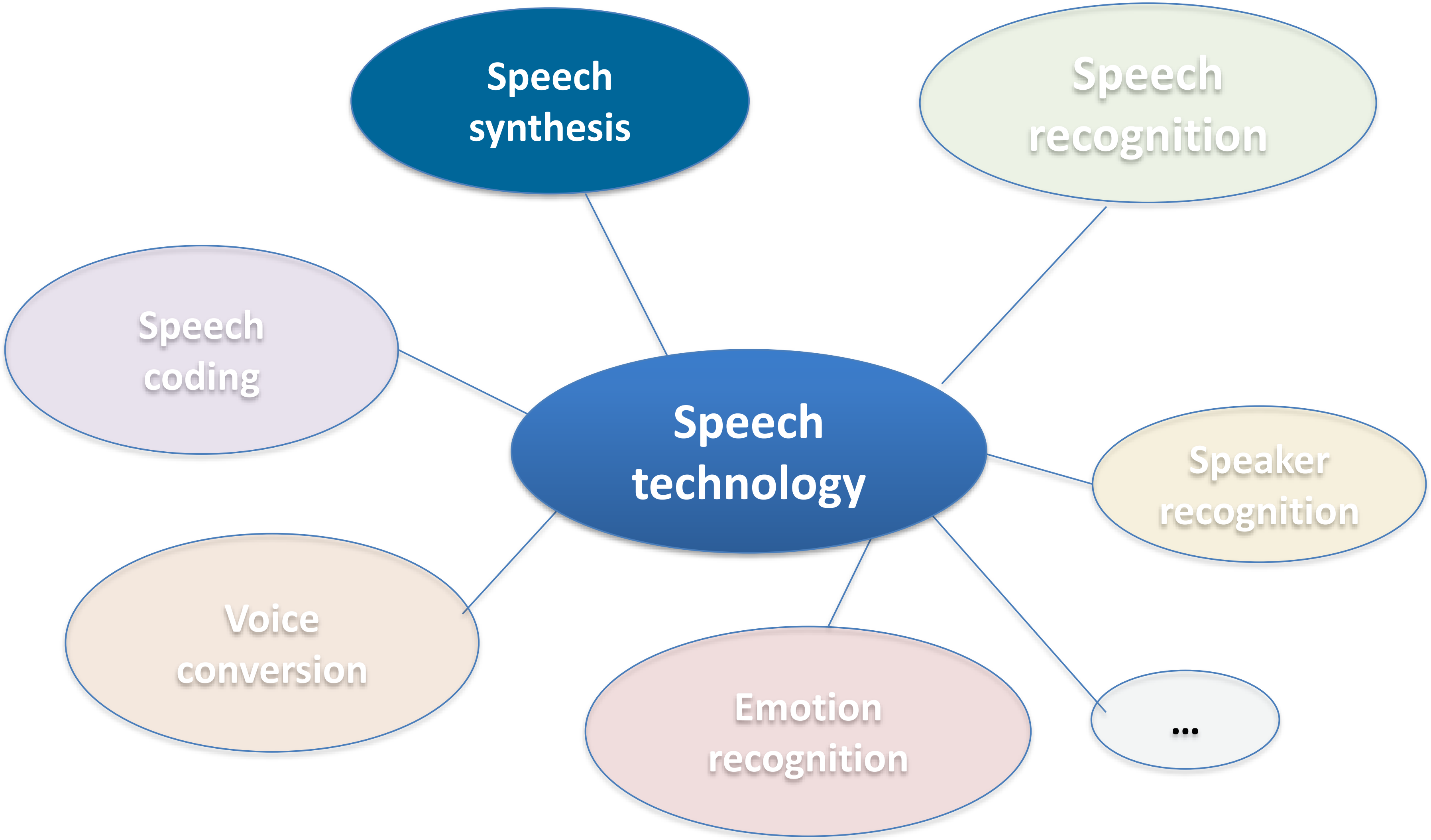


Source: Rabiner (2015) http://www.ece.ucsb.edu/Faculty/Rabiner/ece259/speech%20course.html

LPC examples

- Waveform coding
  - – Original (64 kbps)
  - – ADPCM (32 kbps)
- Linear Predictive Coding
  - – CELP (4800 bps)
  - – LPC-10 (2400 bps)

Source: http://www.data-compression.com/speech.html

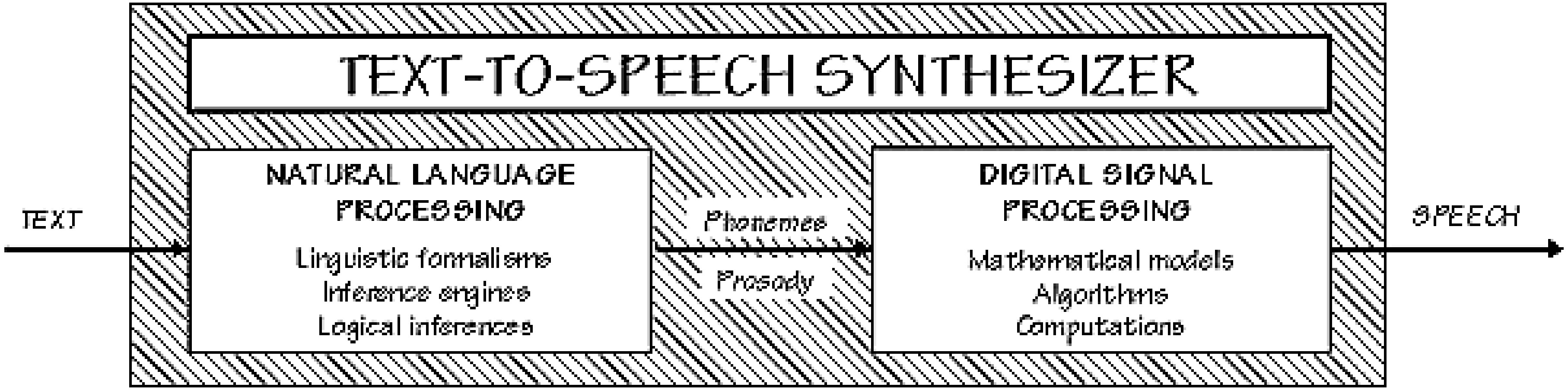

Speech synthesis
Speech recognition
Speech coding
Speech technology
Speaker recognition
Voice conversion
Emotion recognition
...

Text-to-speech synthesis

- ***Synthesis of Speech*** is the process of generating a speech signal using computational means for effective human-machine interactions
  - machine reading of text or email messages
  - telematics feedback in automobiles
  - talking agents for automatic transactions
  - announcement machines that provide information such as stock quotes, airlines schedules, weather reports, etc.
  - screen reader for the blind
  - speech communication help for the speaking impaired

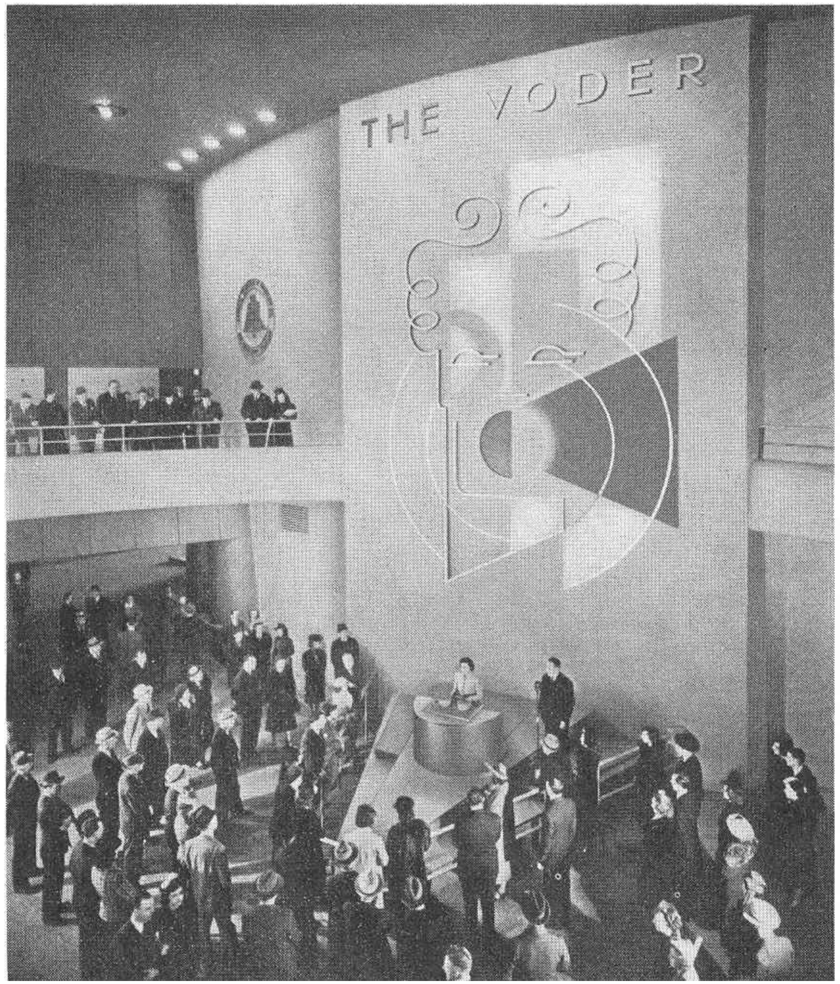
Source: Rabiner (2015) http://www.ece.ucsb.edu/Faculty/Rabiner/ece259/speech%20course.html

Text-to-speech (TTS)



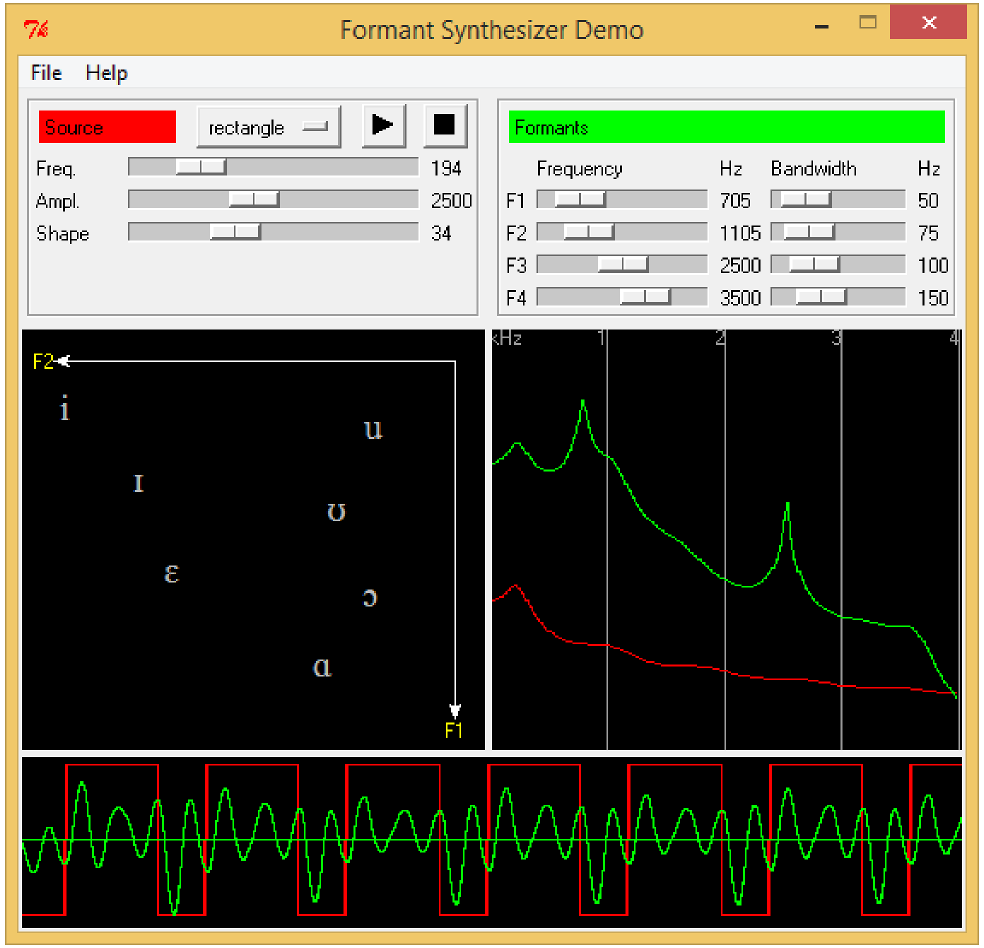
Speech synthesis - history

- 1939, „Voder” electromechanical system
- https://www.youtube.com/watch?v=0rAyrmm7vv0



Formant synthesis

- http://www.speech.kth.se/wavesurfer/formant/



Formant Synthesizer Demo

File Help

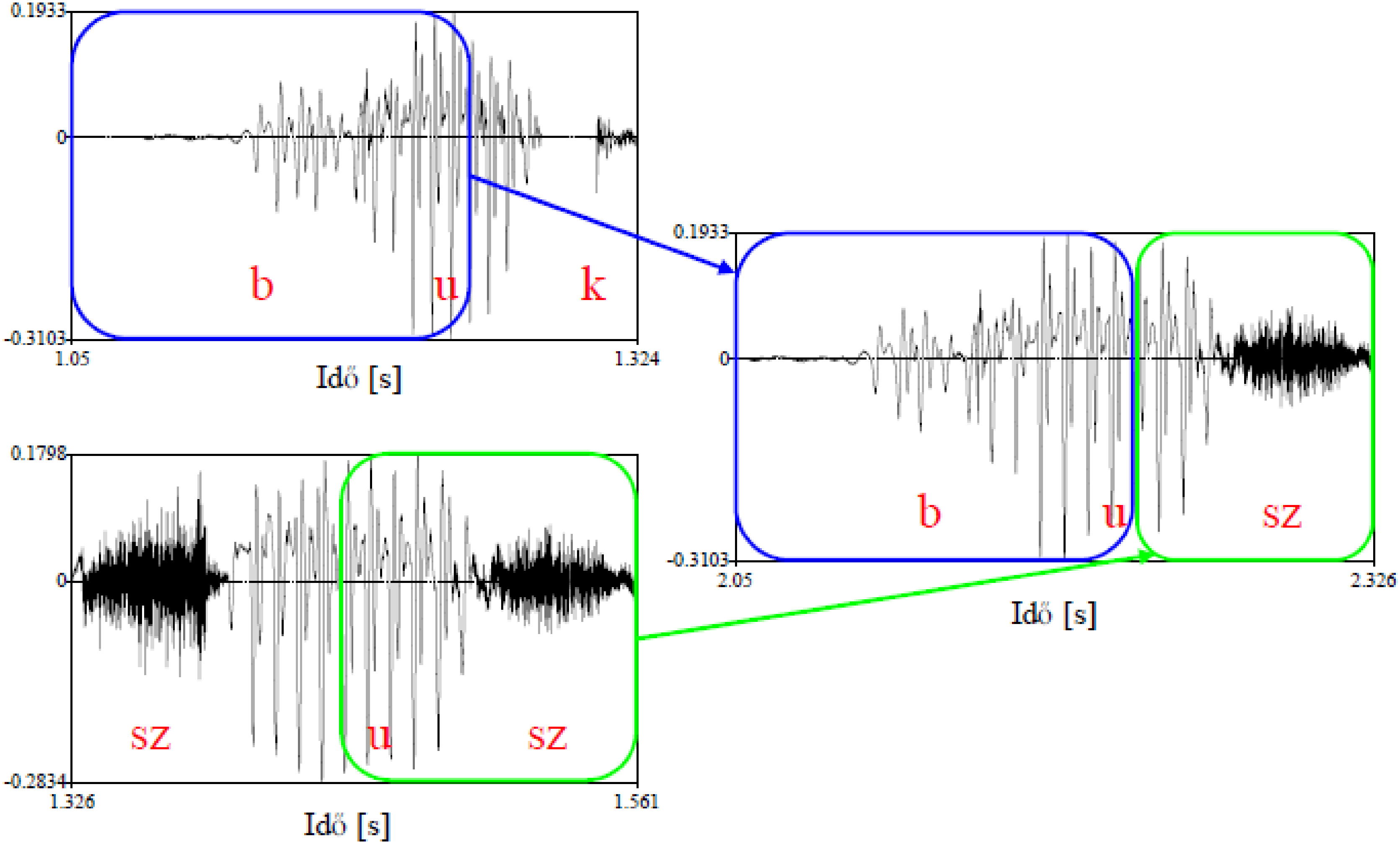
Source rectangle

Freq.	194
Ampl.	2500
Shape	34

Formants

	Frequency	Hz	Bandwidth	Hz
F1		705		50
F2		1105		75
F3		2500		100
F4		3500		150

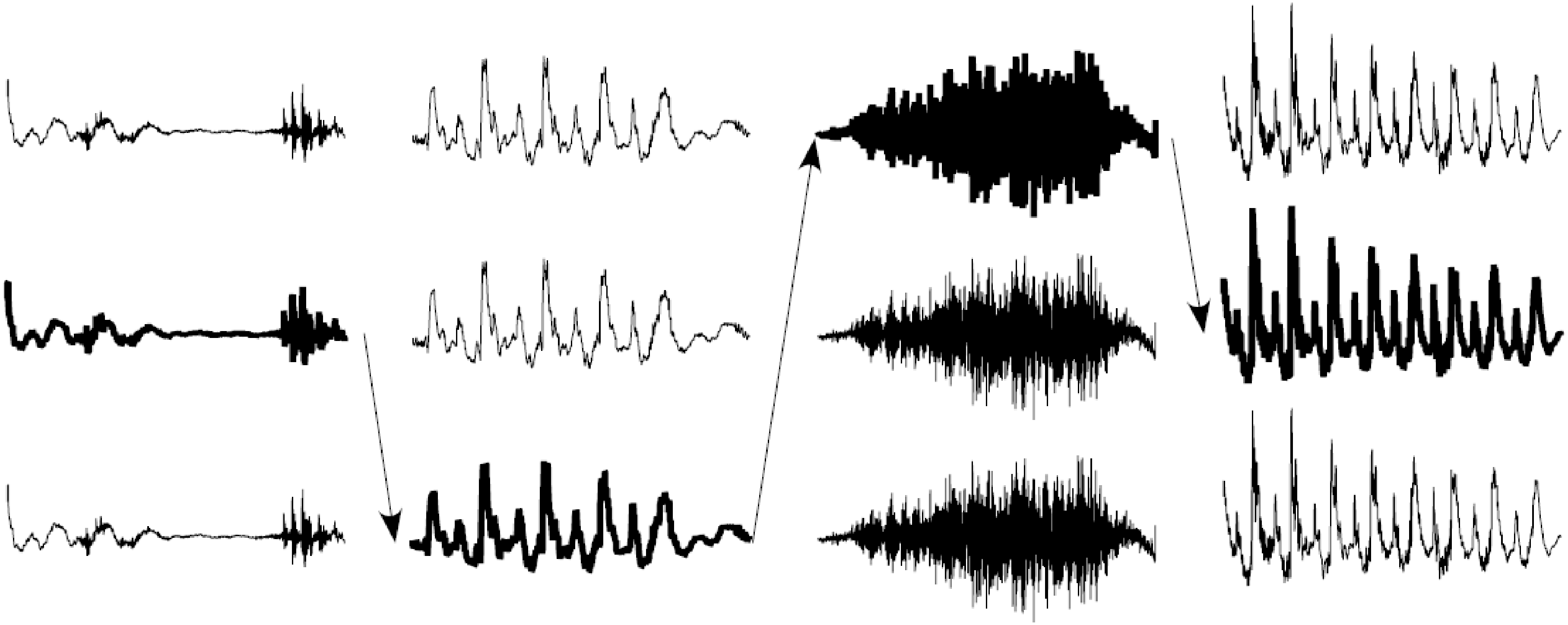
Diphone concatenation



Unit selection /1

Source: King (2015) http://www.research.ed.ac.uk/portal/files/21345037/Simon_King_ICPhS2015_keynote_for_publication.pdf

Unit selection /2



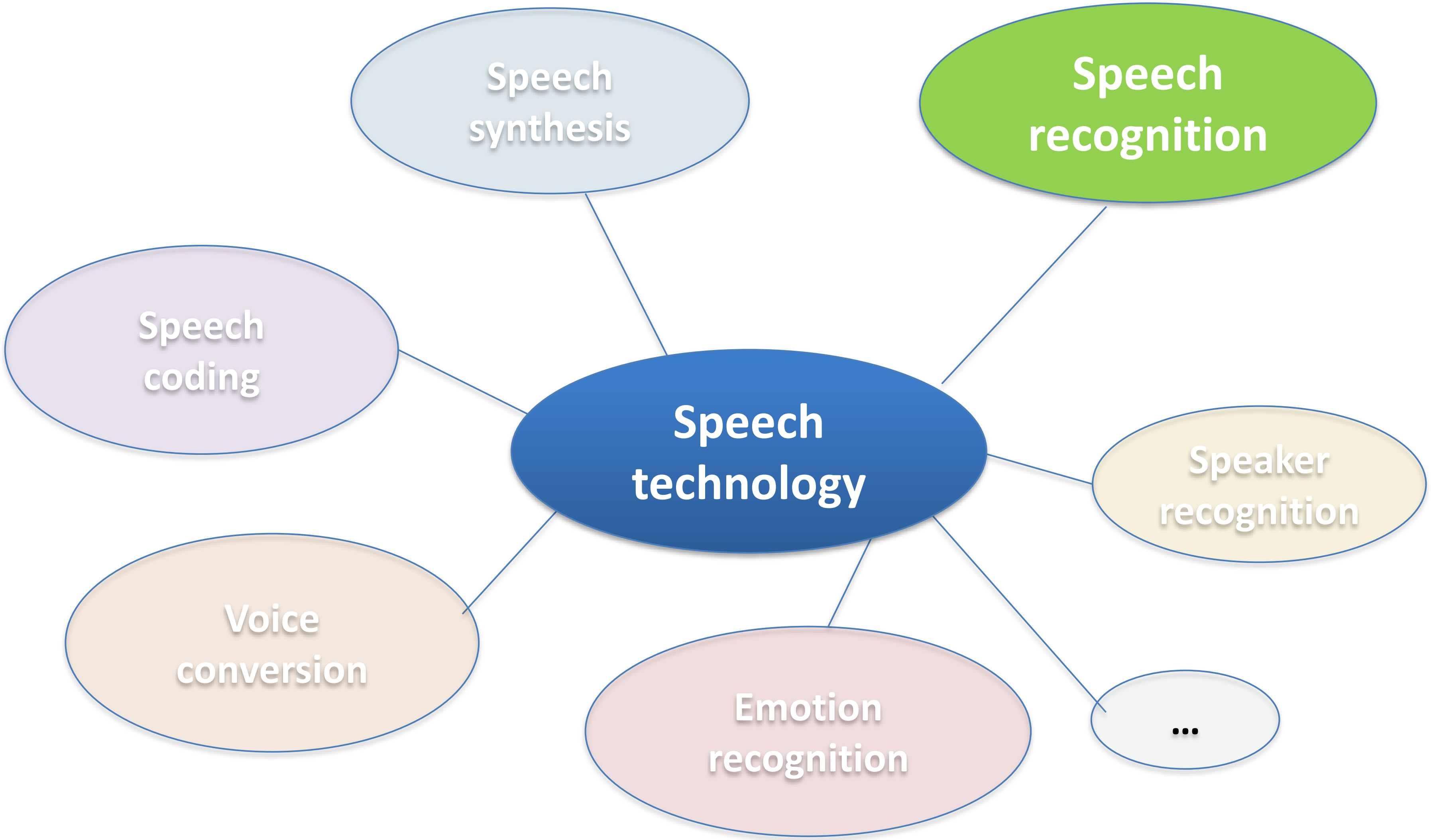
Speech synthesis samples

- Formant synthesis (’70s)

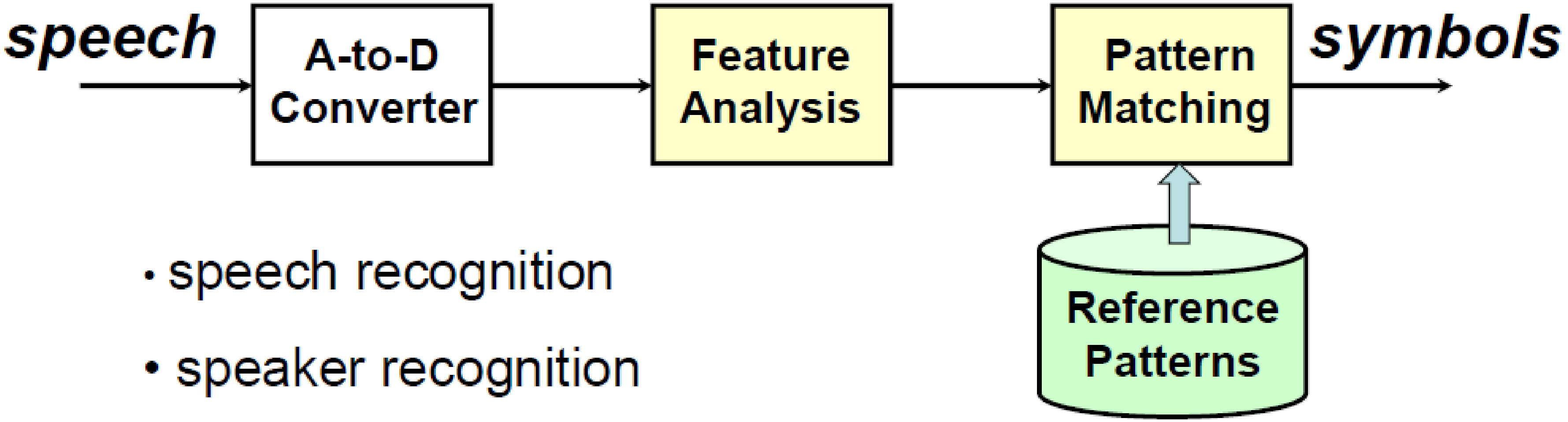
- Diphone concatenation (’80s)

- Unit selection (’90s)

- Statistical speech synthesis (2005-)


Speech synthesis
Speech recognition
Speech coding
Speech technology
Speaker recognition
Voice conversion
Emotion recognition
...

Pattern Matching Problems



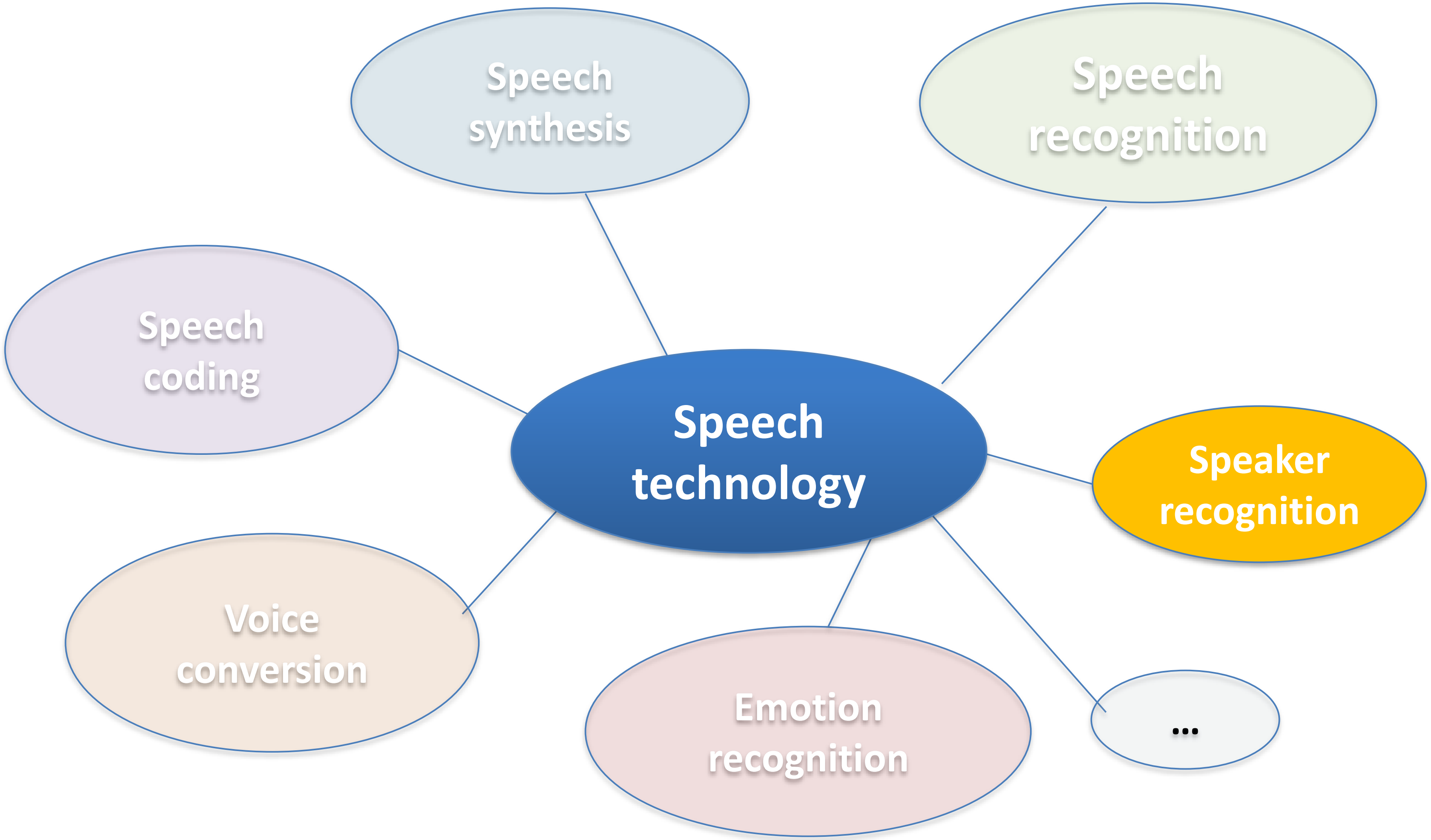
- speech recognition
- speaker recognition
- speaker verification
- word spotting
- automatic indexing of speech recordings

Source: Rabiner (2015) http://www.ece.ucsb.edu/Faculty/Rabiner/ece259/speech%20course.html

Automatic Speech Recognition

- ***Recognition of Speech*** is the process of extracting usable linguistic information from a speech signal in support of human-machine communication by voice
  - command and control (C&C) applications, e.g., simple commands for spreadsheets, presentation graphics, appliances
  - voice dictation to create letters, memos, and other documents
  - natural language voice dialogues with machines to enable Help desks, Call Centers
  - voice dialing for cellphones and from PDA's and other small devices

Source: Rabiner (2015) http://www.ece.ucsb.edu/Faculty/Rabiner/ece259/speech%20course.html



Speech technology

Speech synthesis

Speech recognition

Speech coding

Speaker recognition

Voice conversion

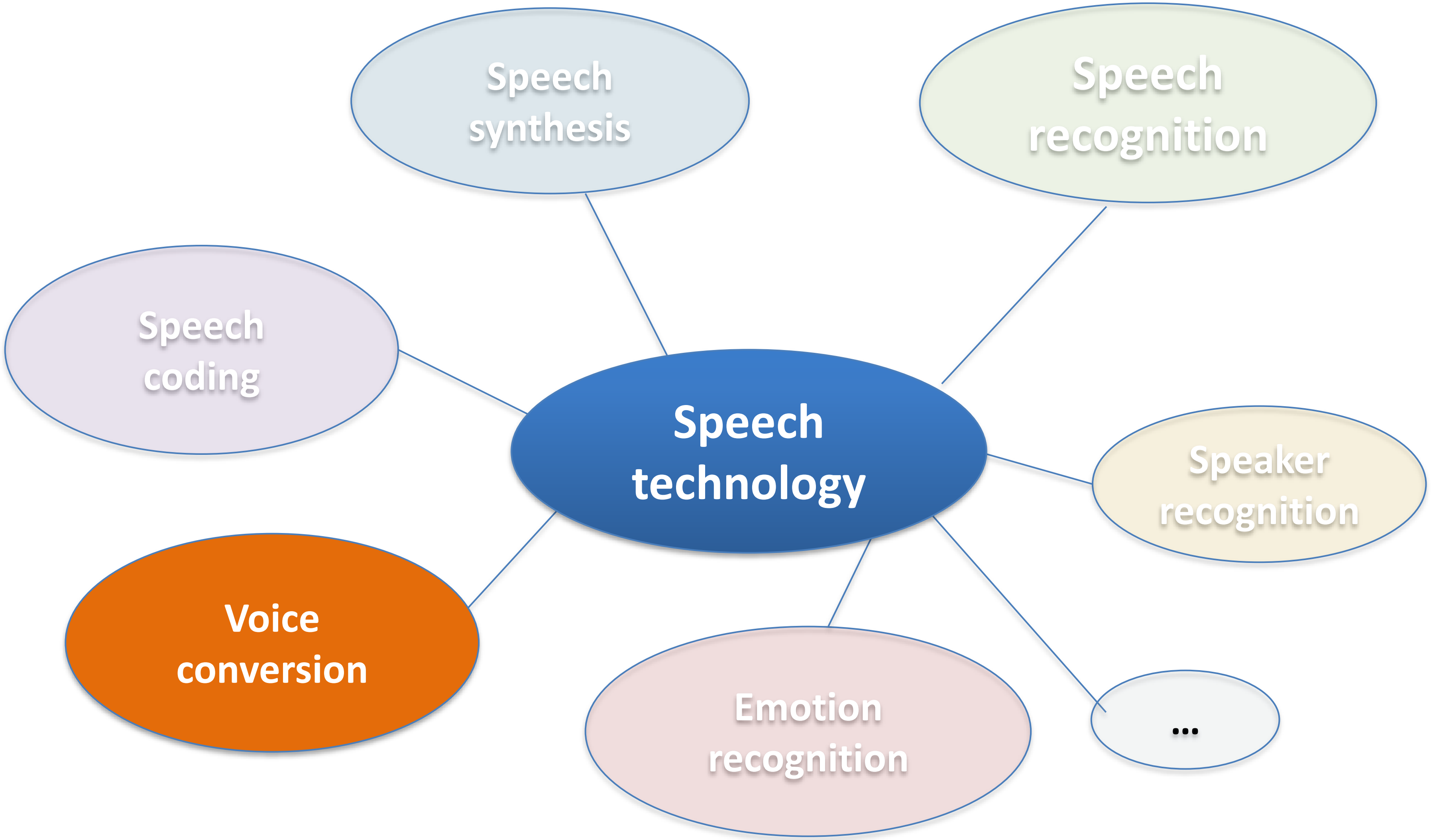
Emotion recognition

...

Speaker verification, recognition

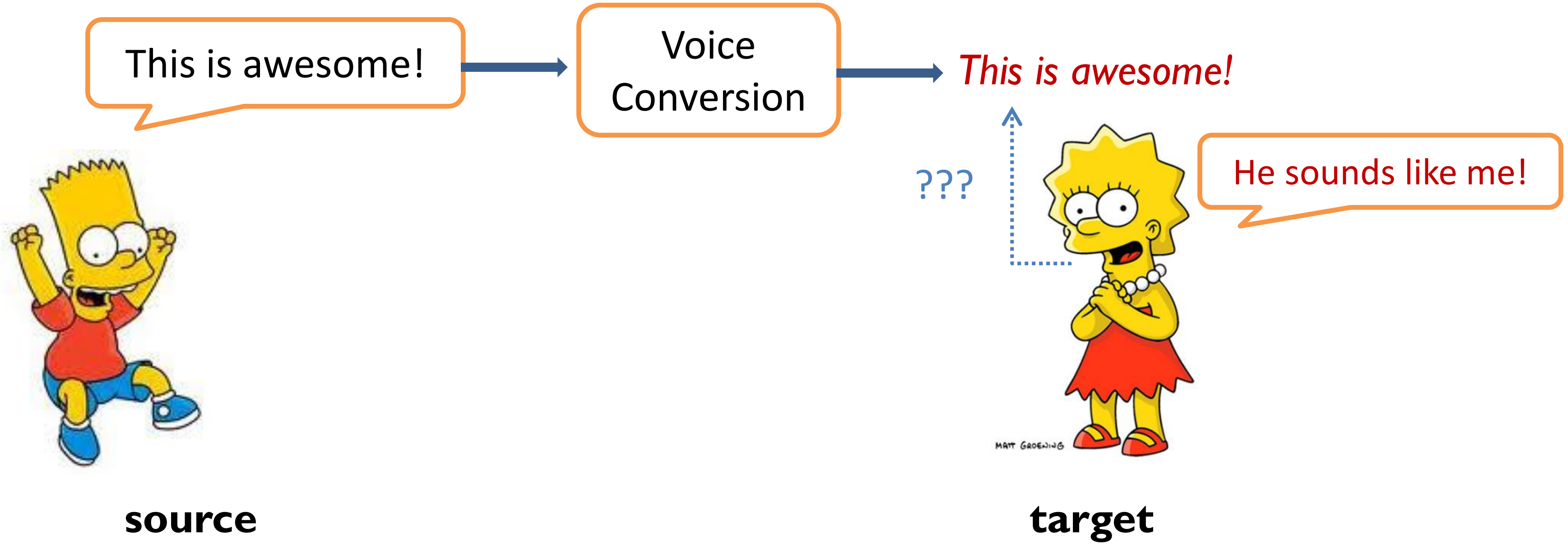
- ***Speaker Verification***
  - secure access to premises, information, virtual spaces
- ***Speaker Recognition***
  - legal and forensic purposes - national security; also for personalized services

Source: Rabiner (2015) http://www.ece.ucsb.edu/Faculty/Rabiner/ece259/speech%20course.html


Speech synthesis
Speech recognition
Speech coding
Speech technology
Speaker recognition
Voice conversion
Emotion recognition
...

Voice conversion

- Transform the speech of a (source) speaker so that it sounds like the speech of a different (target) speaker.



This is awesome!

Voice Conversion

This is awesome!

???

He sounds like me!

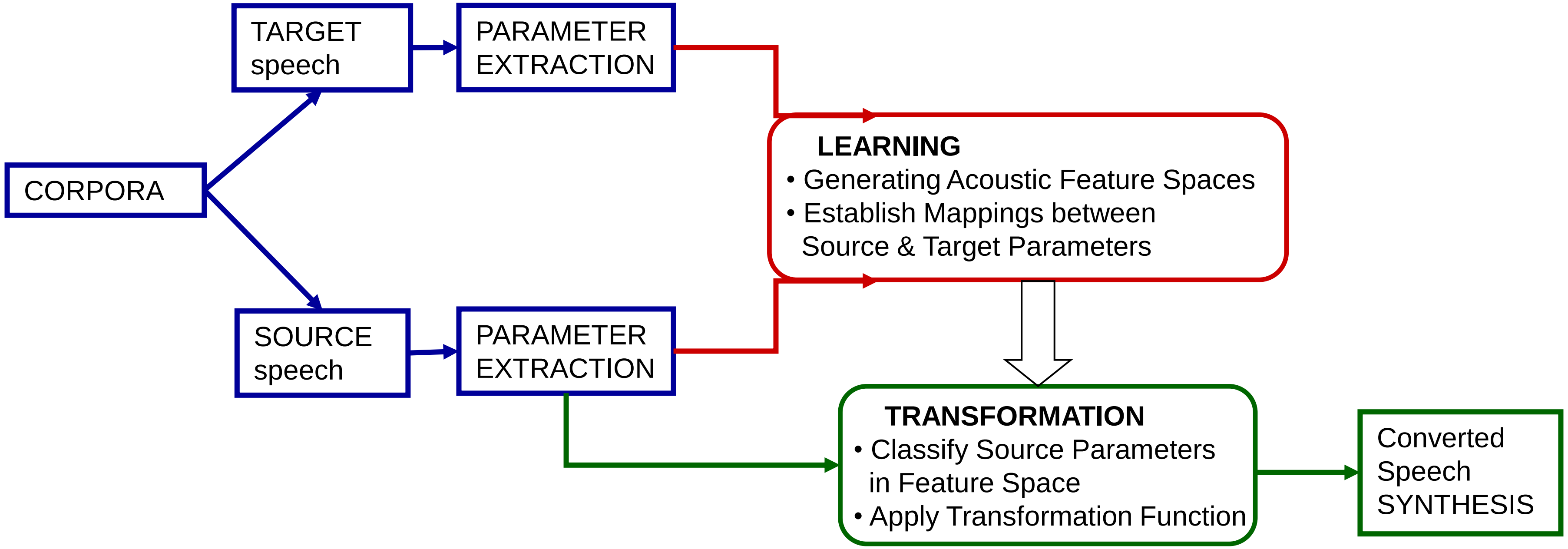
source

target

Source: http://www.csd.uoc.gr/~hy578/2012/Godoy-VC/GodoyLecture_Dec11.ppt

Stages of Voice Conversion

1) Analysis, 2) Learning, 3) Transformation


TARGET speech
PARAMETER EXTRACTION
CORPORA
SOURCE speech
PARAMETER EXTRACTION
LEARNING
• Generating Acoustic Feature Spaces
• Establish Mappings between Source & Target Parameters
TRANSFORMATION
• Classify Source Parameters in Feature Space
• Apply Transformation Function
Converted Speech SYNTHESIS

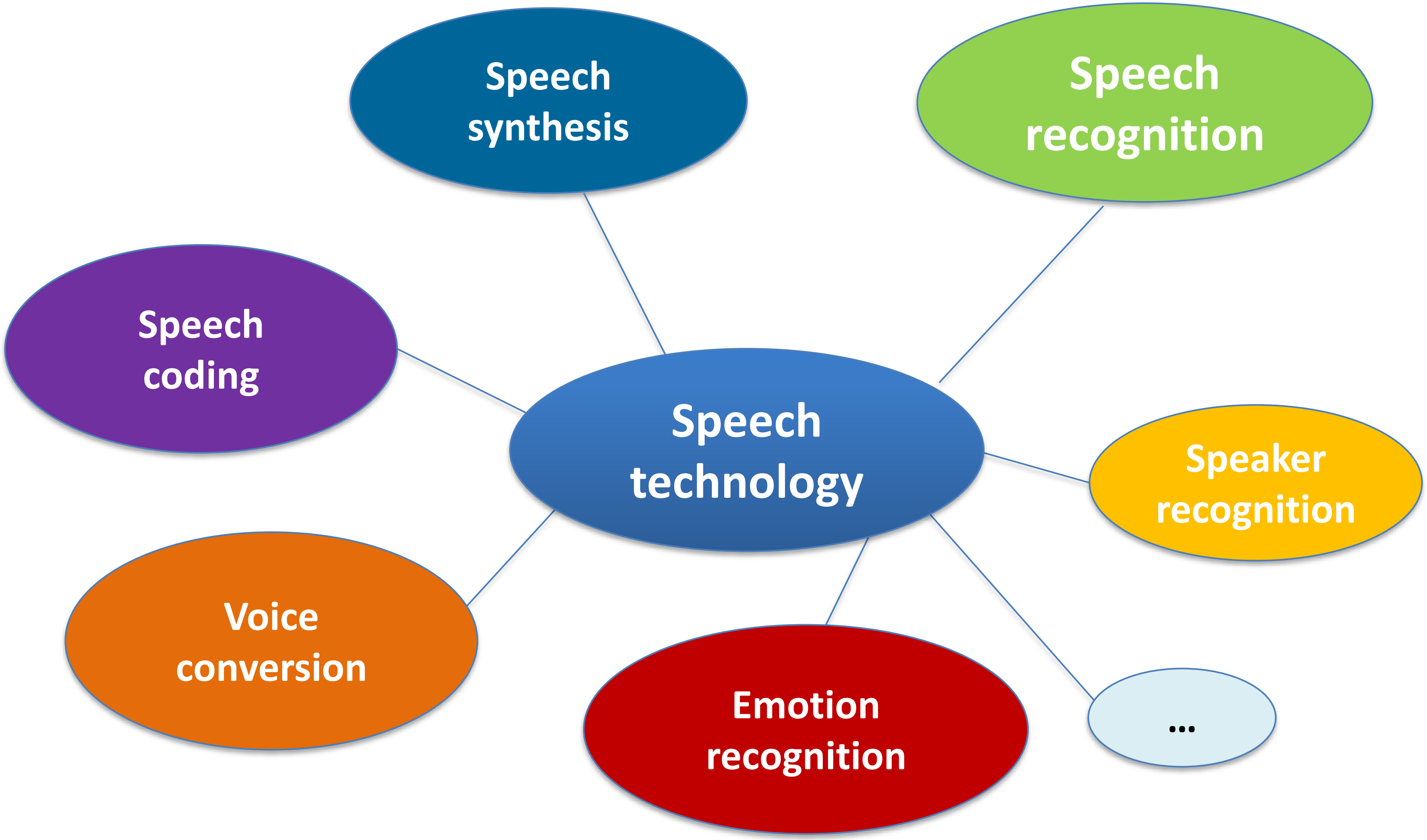
- Key Parameter: the spectral envelope (relation to timbre)

Source: http://www.csd.uoc.gr/~hy578/2012/Godoy-VC/GodoyLecture_Dec11.ppt

Voice conversion examples

	Source	Target	GMM	DFWA	DFWE
slt → clb (FF)	🔊	🔊	🔊	🔊	🔊
bdl → clb (MF)	🔊	🔊	🔊 🔊	🔊 🔊	🔊

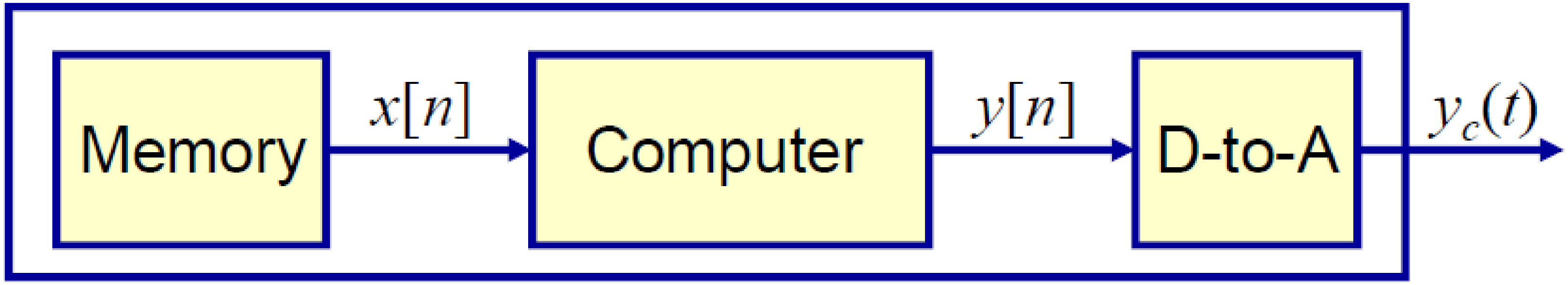
Source: http://www.csd.uoc.gr/~hy578/2012/Godoy-VC/GodoyLecture_Dec11.ppt


Speech synthesis
Speech recognition
Speech coding
Speech technology
Speaker recognition
Voice conversion
Emotion recognition
...

PERCEPTUAL CODING OF AUDIO SIGNALS

Apple iPod

- stores music in MP3, AAC, MP4, wma, wav, … audio formats
- compression of 11-to-1 for 128 kbps MP3
- can store order of 20,000 songs with 30 GB disk
- can use flash memory to eliminate all moving memory access
- can load songs from iTunes store – more than 1.5 billion downloads
- tens of millions sold



Source: Rabiner (2015) http://www.ece.ucsb.edu/Faculty/Rabiner/ece259/speech%20course.html

Compression

- High data rates, such as CD audio (4.32 Mb/s), are incompatible with internet & wireless applications.
- Audio data must somehow be compressed to a smaller size (less bits), while not affecting signal quality (minimizing quantization noise).
- **Perceptual Audio Encoding** is the encoding of audio signals, incorporating psychoacoustic knowledge of the auditory system, in order to reduce the amount of bits necessary to faithfully reproduce the signal.
    - MPEG-1 Layer III (aka mp3)
    - MPEG-2 Advanced Audio Coding (AAC)

Source: http://web.mit.edu/hst.723/www/Labs/Fundamentals%20of%20Perceptual%20Audio%20Coding.ppt

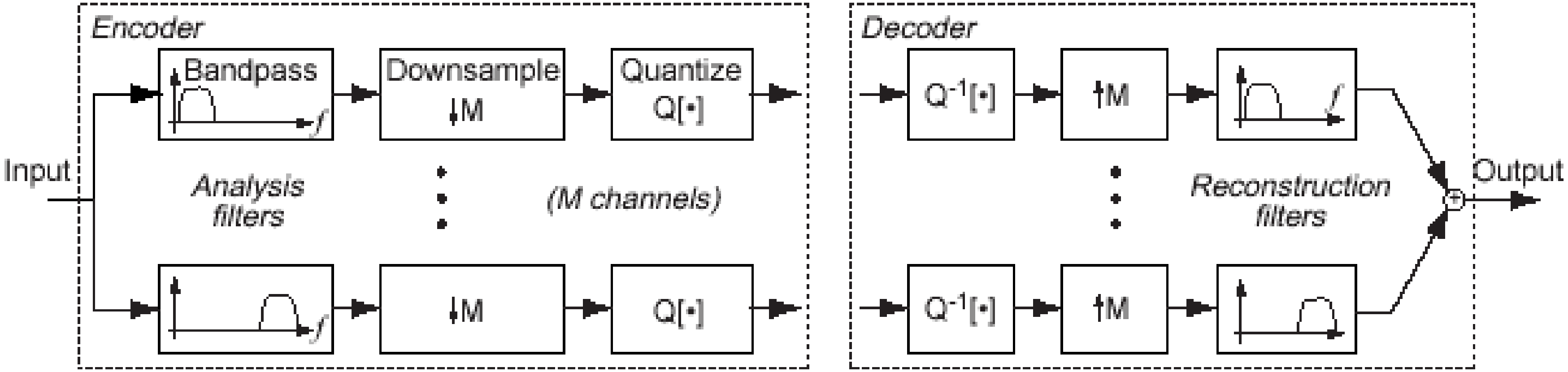
Perceptual coding

- Goal: compress audio (e.g. music) without quality loss

- Use properties of hearing
  - Critical bands
  - Hearing limitations
  - Masking
    - Time domain
    - Frequency comain

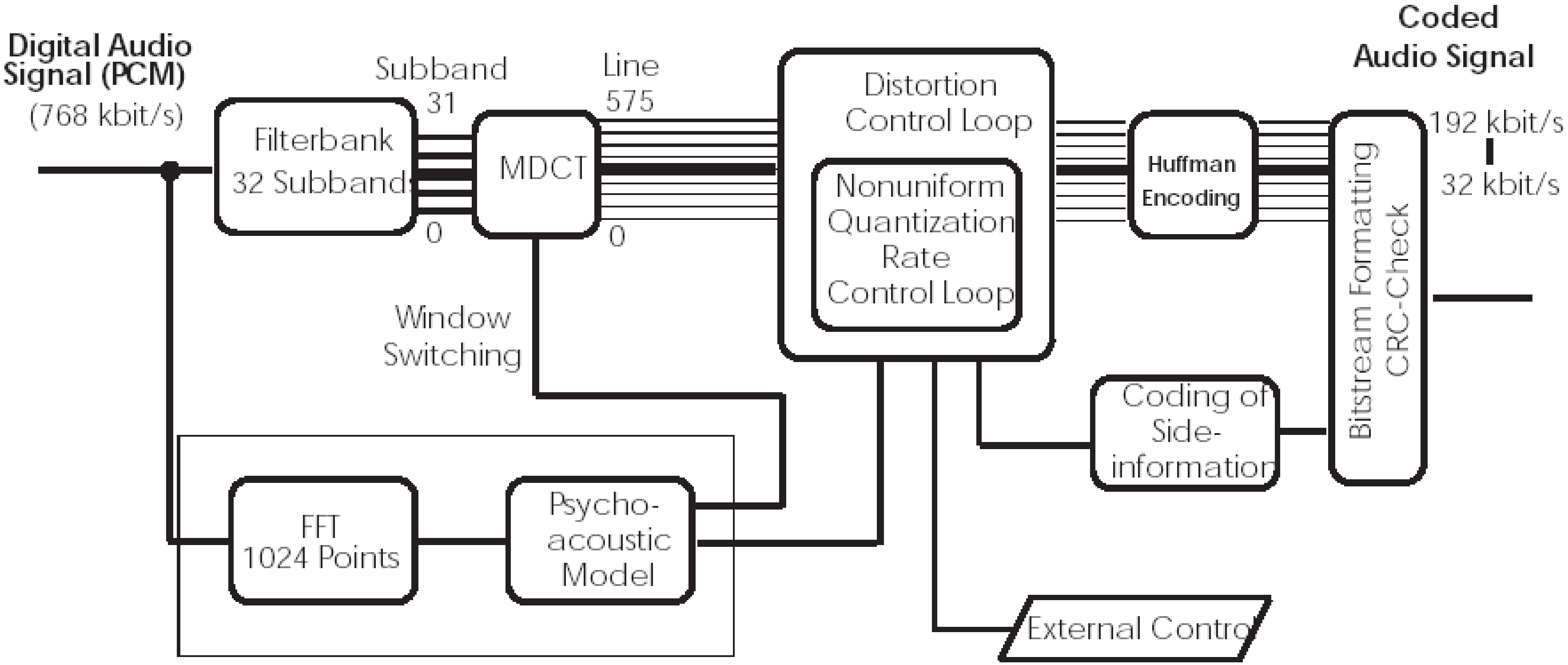
Subband coding

- Analysis filter bank, M bandpass filters
- Quantize separately in different bands
  - quantization noise stay within band; gets masked

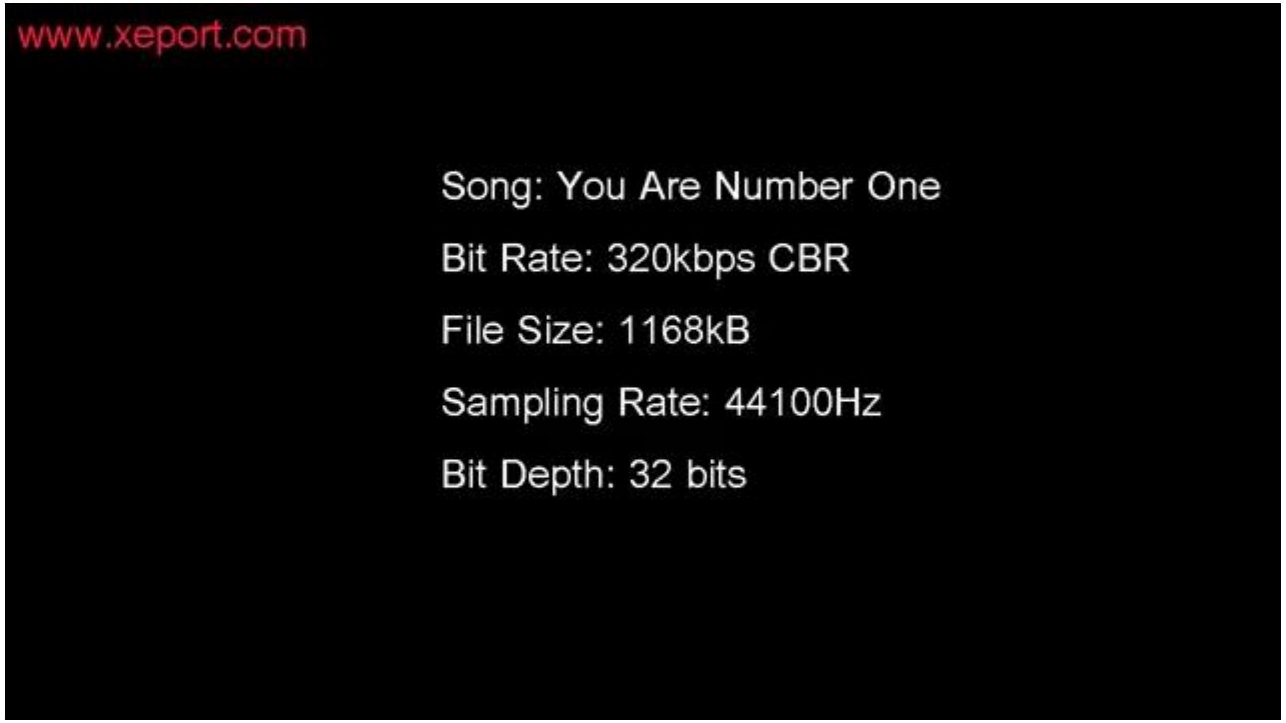


Source: http://www.aisv.it/AISVScuolaEstiva2008/materials/N.Orio/Compressione-MP3.ppt

MP3


Digital Audio Signal (PCM)
(768 kbit/s)
Filterbank
32 Subbands
Subband
31
0
MDCT
Line
575
0
Window
Switching
FFT
1024 Points
Psycho-
acoustic
Model
Distortion
Control Loop
Nonuniform
Quantization
Rate
Control Loop
Huffman
Encoding
Coding of
Side-
information
External Control
Bitstream Formatting
CRC-Check
Coded
Audio Signal
192 kbit/s
32 kbit/s

MP3 Bit Rate vs. Audio Quality



www.xeport.com

Song: You Are Number One

Bit Rate: 320kbps CBR

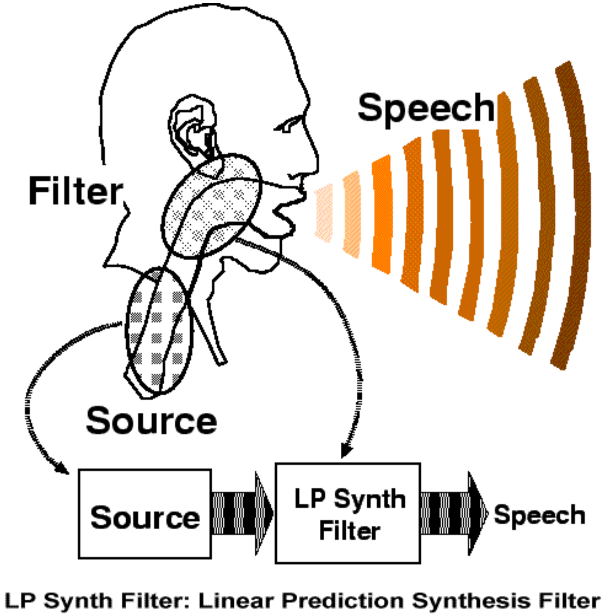
File Size: 1168kB

Sampling Rate: 44100Hz

Bit Depth: 32 bits

video

Source: https://www.youtube.com/watch?v=0O8MclLD1d0



The END

Infocommunication Speech Processing

Dr. Mohammed Salah Al-Radhi

Dr. Tamás Gábor Csapó