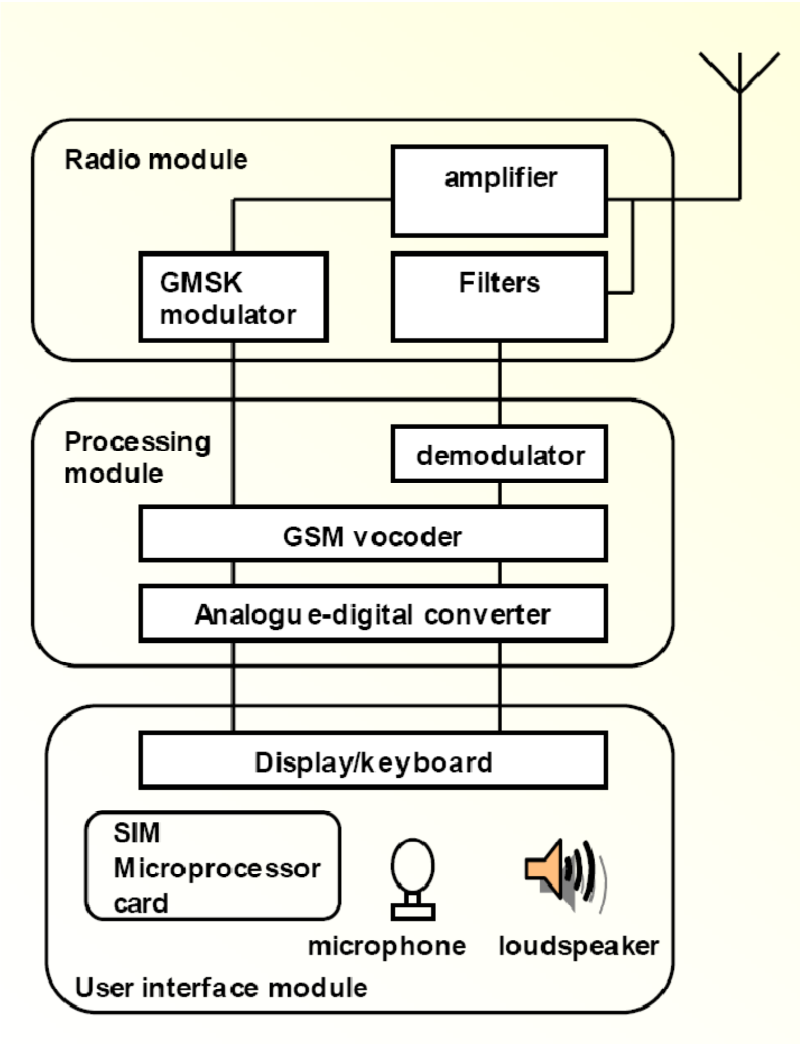
Principles of Source coding
(applied to GSM & UMTS)

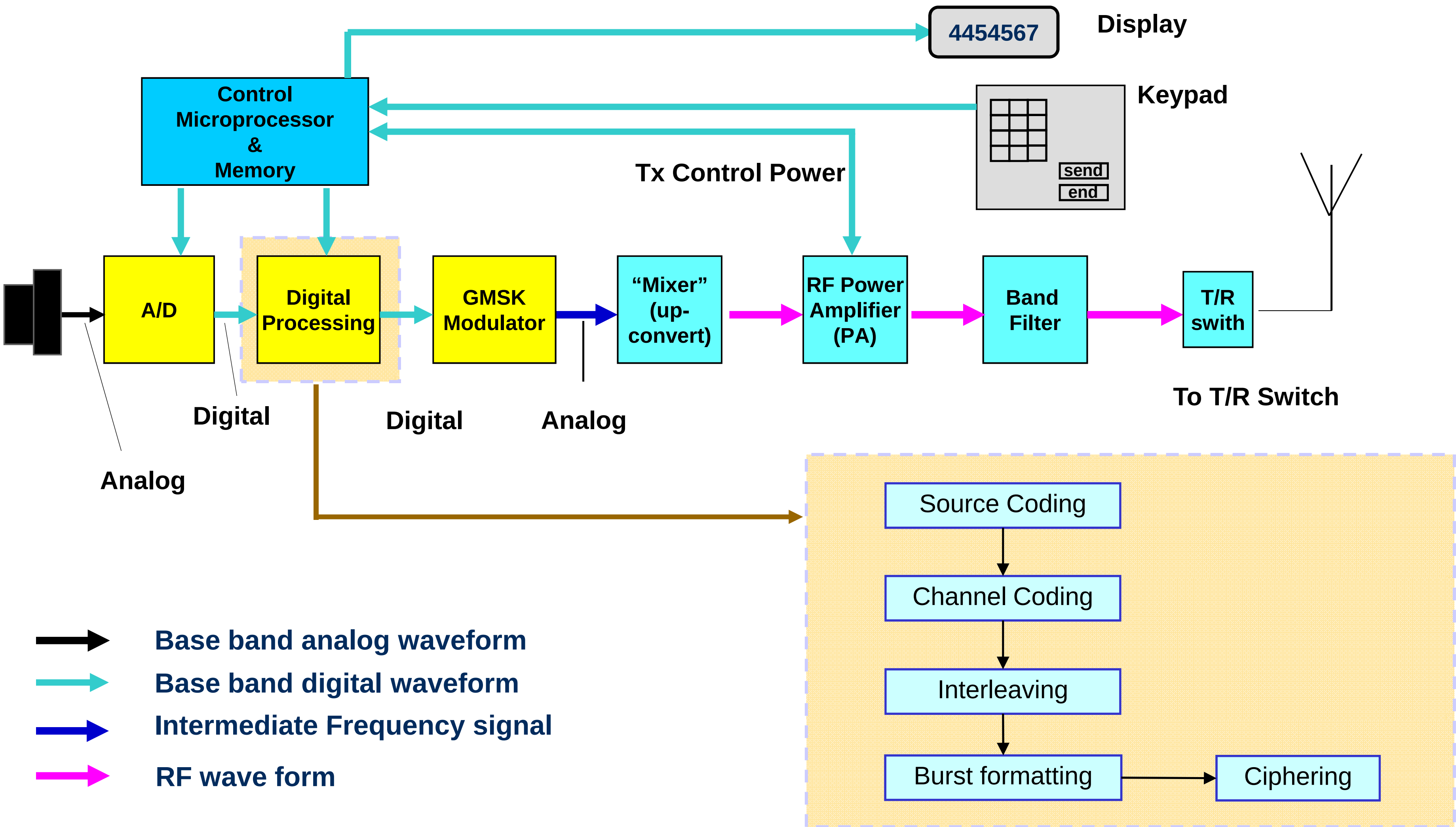
Dr. Hicham Aroudaki

Damascus, 18th March 2010

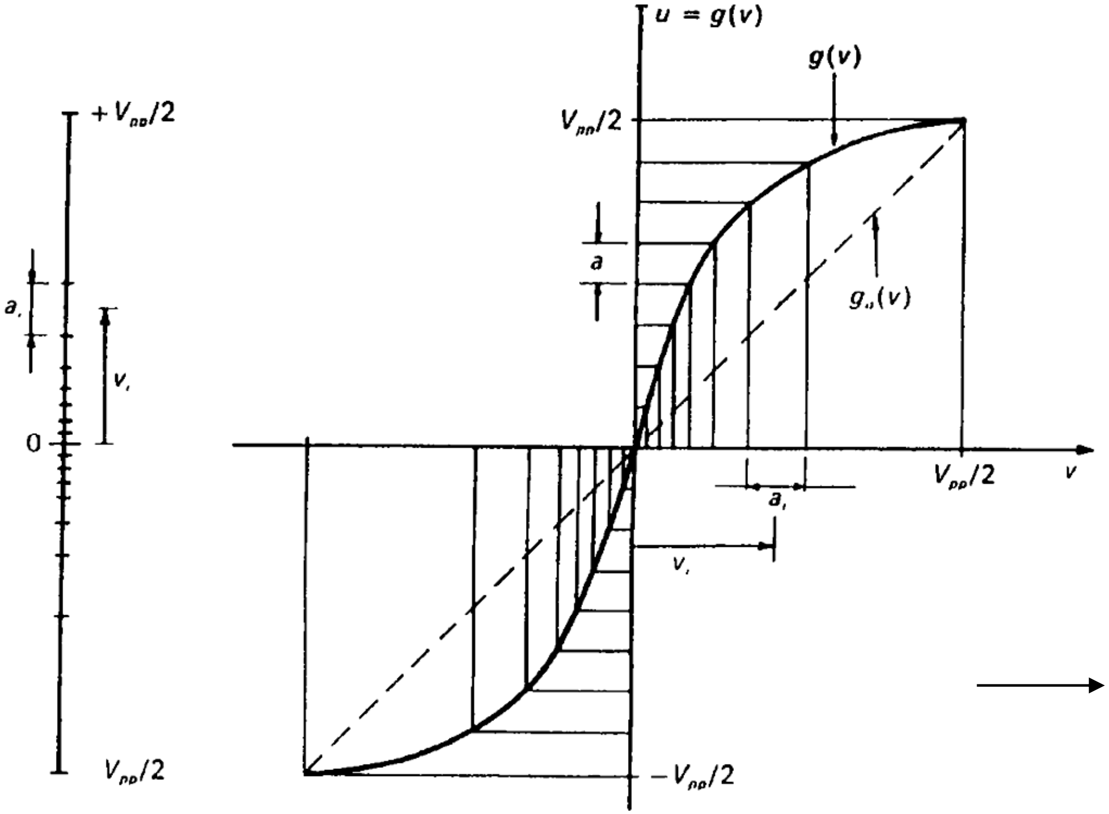
Behind the facade



Modular block diagram (Transmitter)



GSM Quantisation



- Speech is logarithmically quantized.
- 13 quantisation levels.
- (8000 samples / second) * 13 = 104 kbit/s

Source Coding

- **Definition**
  - Reduce the number of bits in order to save on transmission time or storage space

Types of Voice Codecs

- **Wave Form codecs:**
  - Just sampling and coding without thought of speech generation
  - High-quality and not complex
  - Large amount of bandwidth
  - PCM (Pulse Code Modulation) G.711, ISDN: 8b x 8k/s = 64 kbit/s, CD: 16b x 44 k/s x 2ch = 1.408 Mbit/s
  - ADPCM (Adaptive Differential PCM) G.726/27, 40 / 32 / 24 / 16 kbit/s
  - CVSD (Continuously Variable Slope Delta)

Types of Voice Codecs

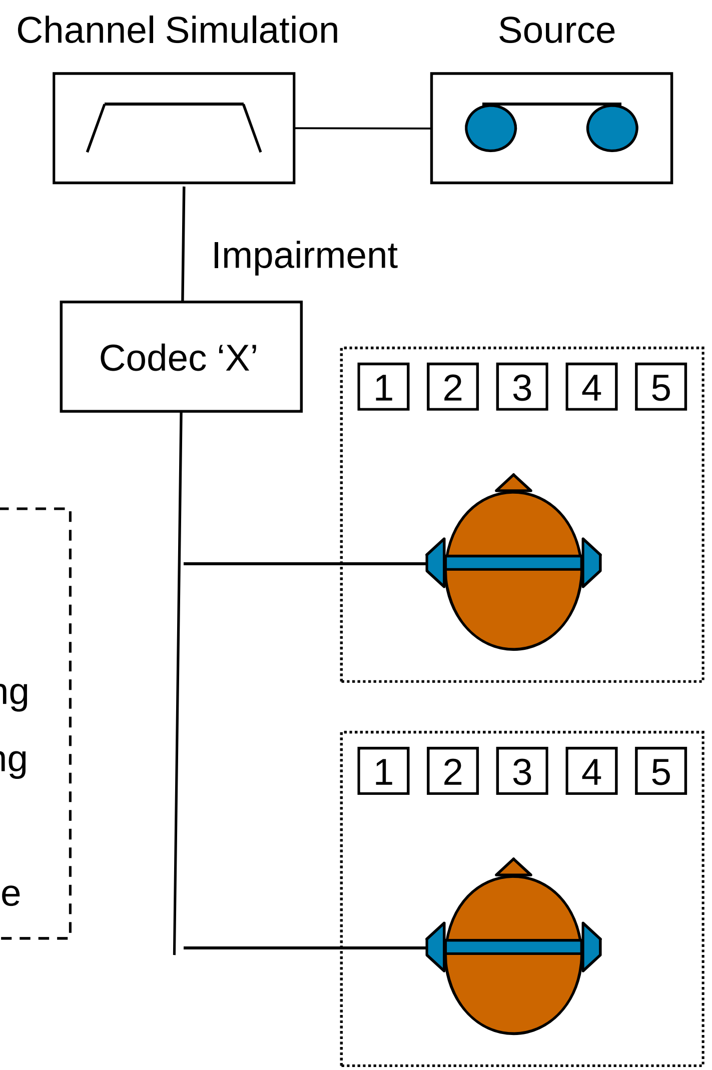
- **Source codecs** (vocoders, synthetic voice):
  - **Encoding**: Match the incoming signal to a math model of speech generation
    - Linear-predictive filter model of the vocal tract
      - Parameters: voiced/unvoiced flag for the excitation
    - Parameters of the filter is sent! (Not the sampled signal)
  - **Decoding** : Apply the parameters to the same filter model (i.e. like to speech synthesis)
  - Low bit rate, but sounds synthetic
    - Higher bit rate does not improve much
  - Codecs: Linear Predictive Coding (LPC)

Types of Voice Codecs

- **Hybrid codecs:**
  - Attempt to provide the best of both
  - Perform a degree of waveform matching
  - Utilize the sound production model
  - Quite good quality at low bit rate
  - Time Domain Analysis-by-Synthesis(AbS) codecs:
  - The most commonly used Not a simple two-state, voiced/unvoiced
  - Different excitation signals are attempted
  - Closest to the original waveform is selected
  - CELP (Code book Excited Linear Prediction)
  - ACELP (Algebraic CELP)
  - RPE-LTP (Regular Pulse Excitation - Long-Term Prediction)
  - VSELP (Vector-Sum Excited Linear Prediction

Mean Opinion Score

“Nowadays, a chicken leg is a rare dish”



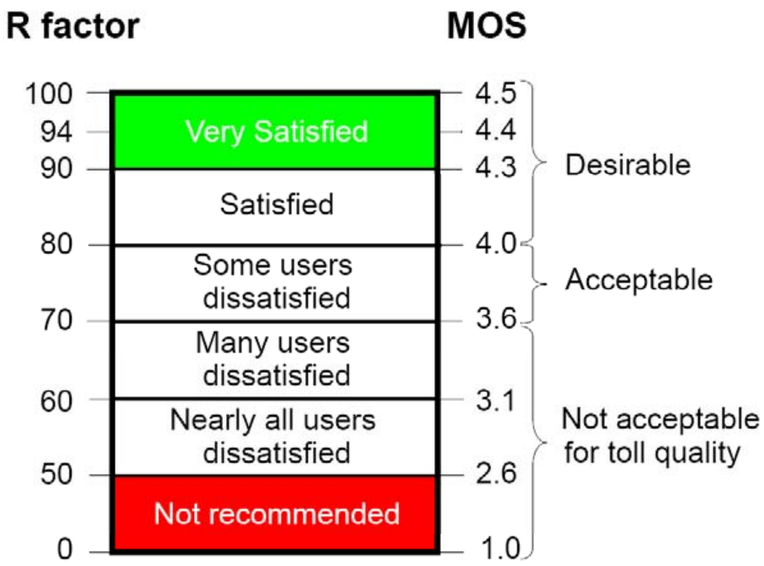
Rating	**Speech Quality**	**Level of Distortion**
5	Excellent	Imperceptible
4	Good	Just perceptible but not annoying
3	Fair	Perceptible and slightly annoying
2	Poor	Annoying but not objectionable
1	Unsatisfactory	Very annoying and objectionable

MOS: Mean Opinion Square

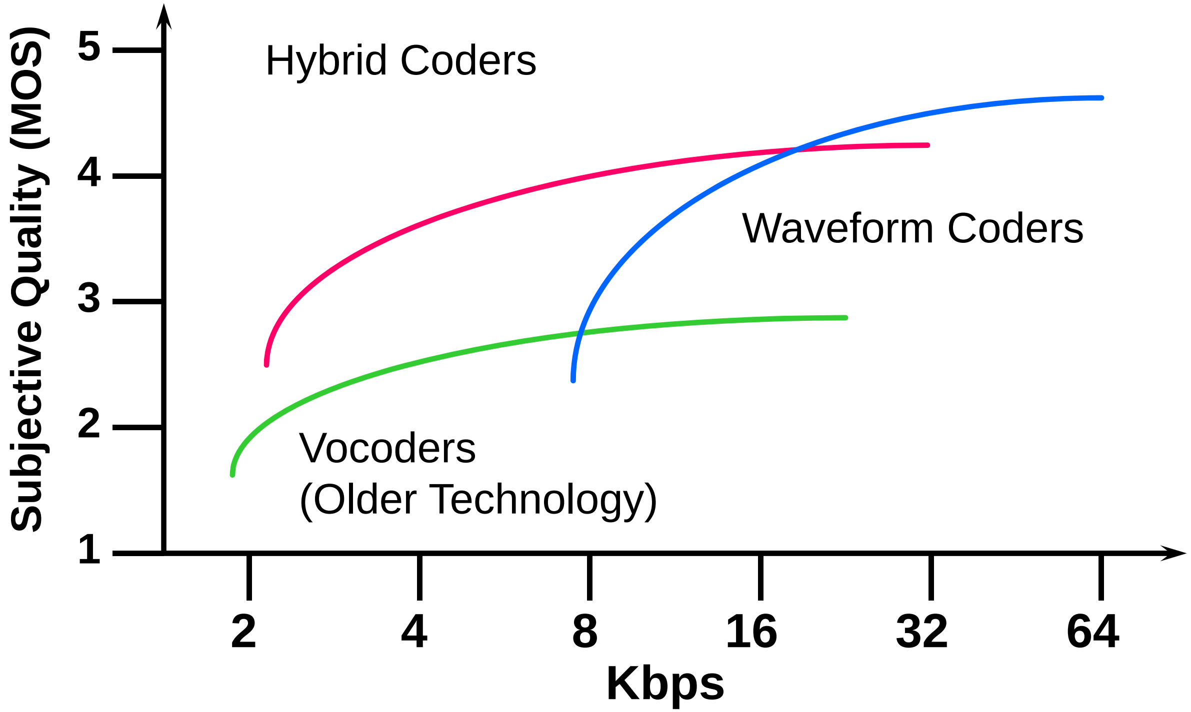
R-factor: in-service voice quality measurement based on observed traffic flow for every phone call

Mean Opinion Square: To arrive at an MOS score, a tester assembles a panel of “expert listeners” who rate the quality of speech samples that have been processed by the system under test.

- Ideally, a panel would consist of a mix of male and female listeners of various ages
- The samples should reflect a range of typical voice conversations
- The panel rates the quality of the system output from 1 to 5, with 1 indicating the worse and 5 the best
- The scores of the panelists are then averaged

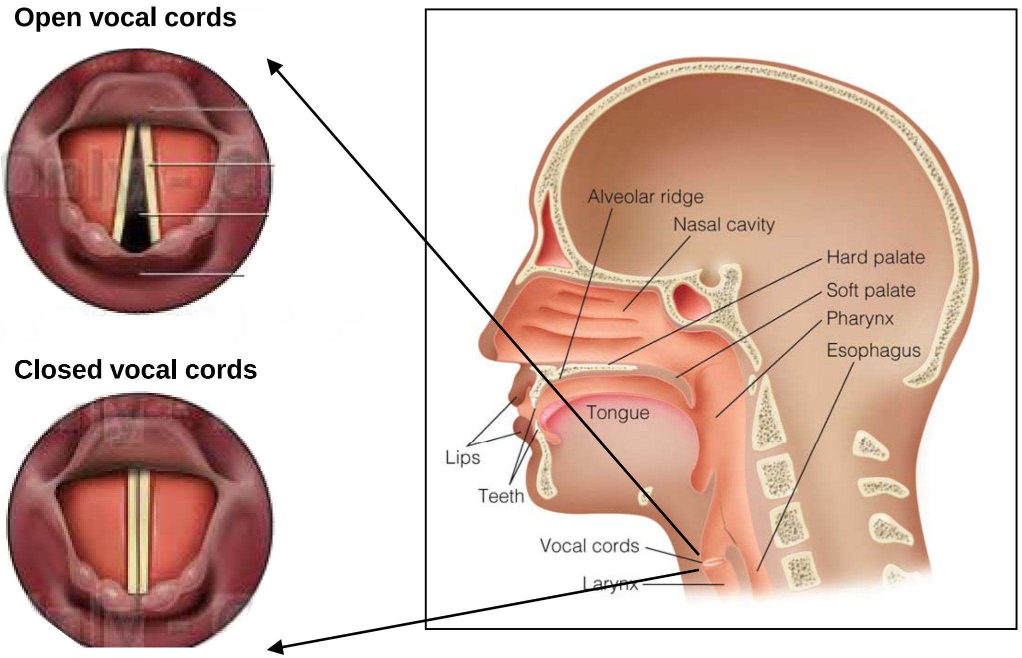


Mean Opinion Scores



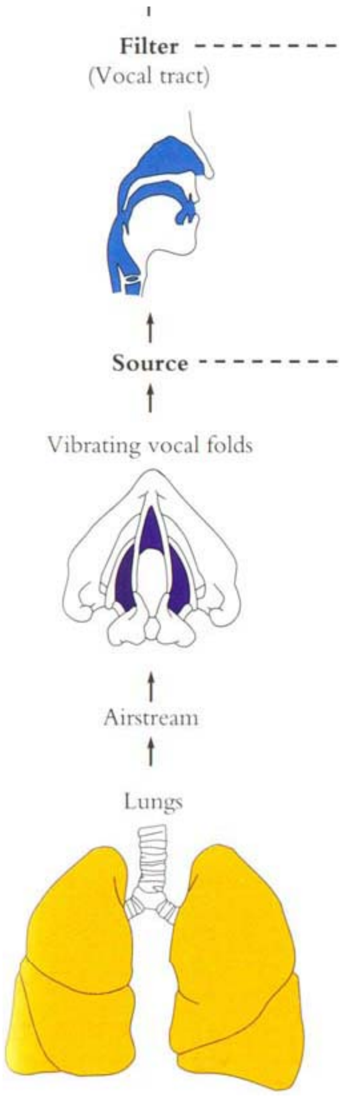
Score	Quality	Description of Impairment
5	Excellent	Imperceptible
4	Good	Just Perceptible, not Annoying
3	Fair	Perceptible and Slightly Annoying
2	Poor	Annoying but not Objectionable
1	Bad	Very Annoying and Objectionable

Voice production



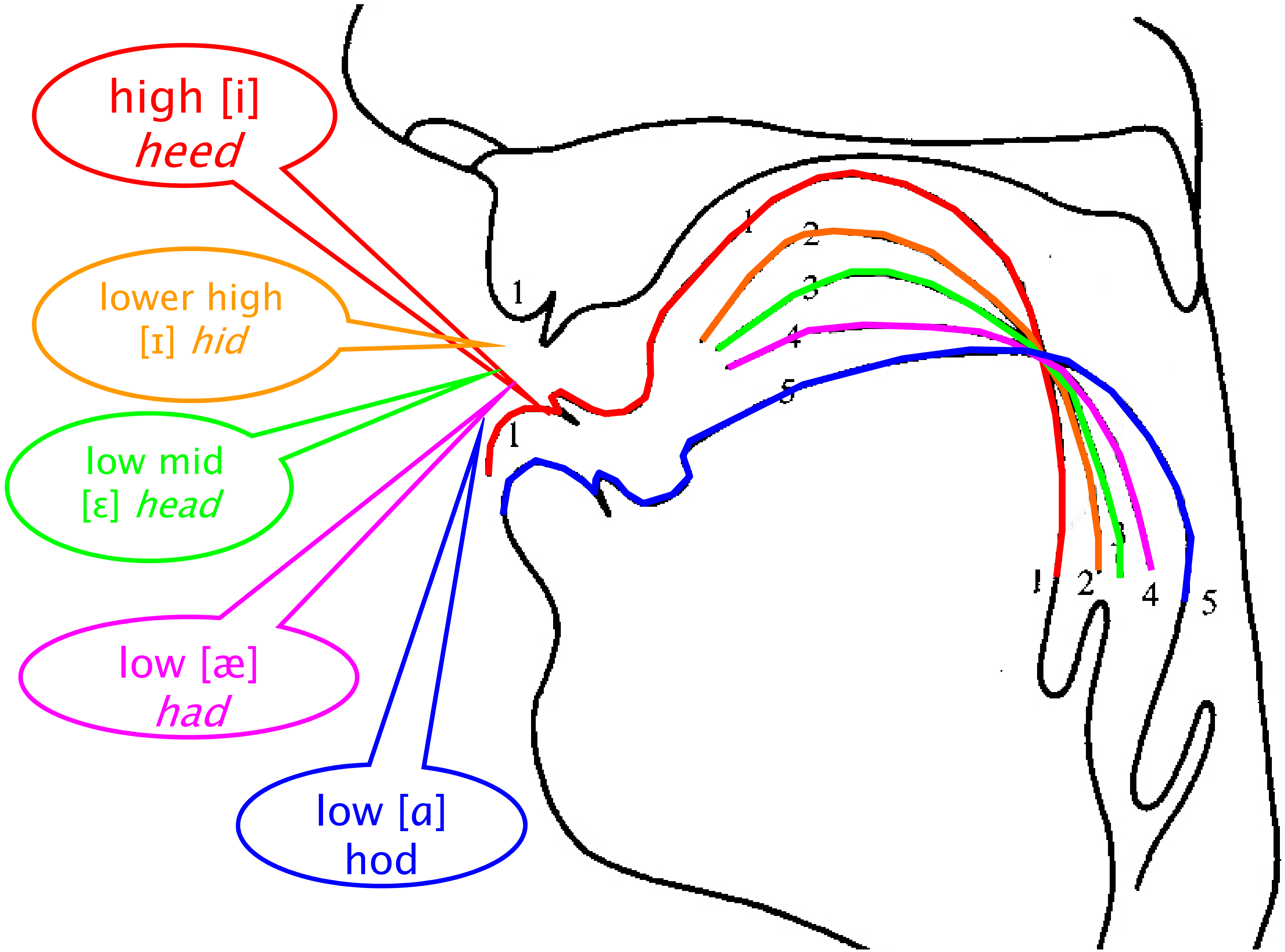
- Speech is produced by a cooperation of lungs, glottis (with vocal cords) and the vocal tract (guttural cavity, oral cavity, nasal cavity).
- The vocal tract is excited with pressure pulses of airflow (product of the vocal cords) with period of opening and closing phase about 10 ms.
- The larynx and the vocal cords produce periodic or noisy signal for voiced or unvoiced phonemes.

Voice production



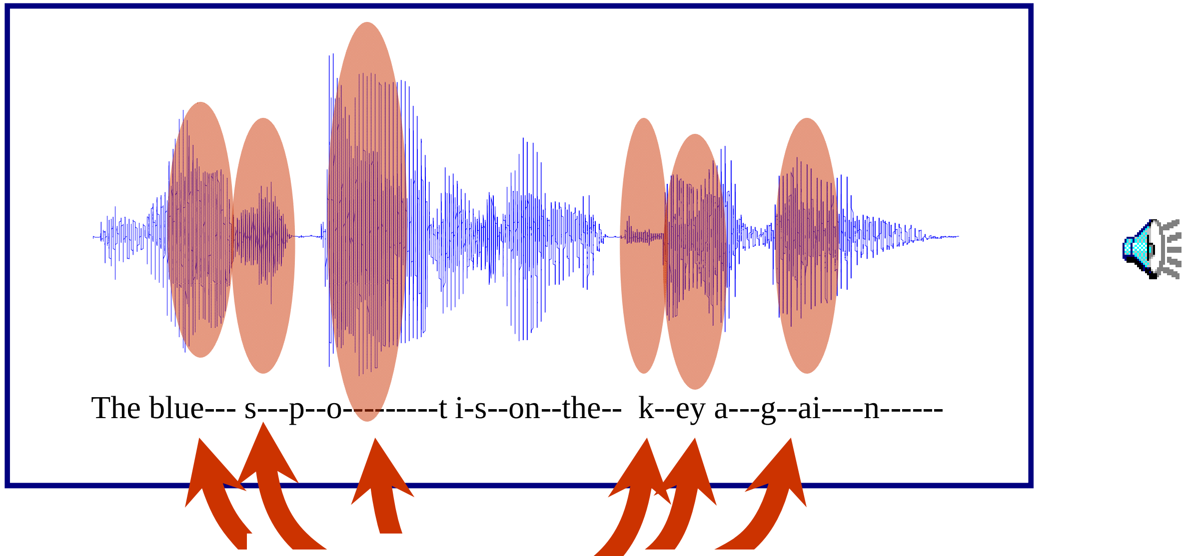
- Mouth and nasal cavities act as resonators with characteristic resonance frequencies, called formant frequencies.
- Parameters of the cavity (its transverse section in single cuts) are varied during the speech.
- Since the mouth cavity can be greatly changed, we are able to pronounce many different sounds.
- For voiced sounds pitch impulses (generated by the vocal cords) stimulate the air in the mouth and for certain sounds (nasal) also stimulate the nasal cavity.
- In the case of unvoiced sounds, the excitation of the vocal tract is more noise-like.

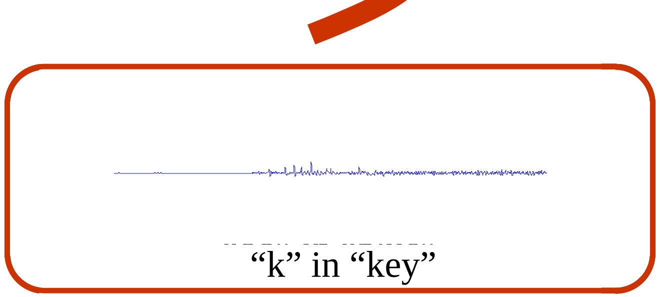
Articulation



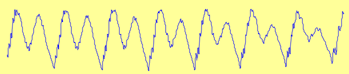
Some properties of speech

The blue spot is on the key again

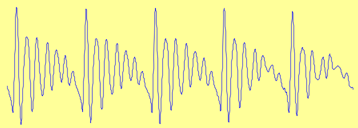




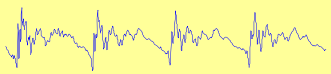
Some Properties of Speech
Vowels (voiced sounds)



“oo” in “blue”

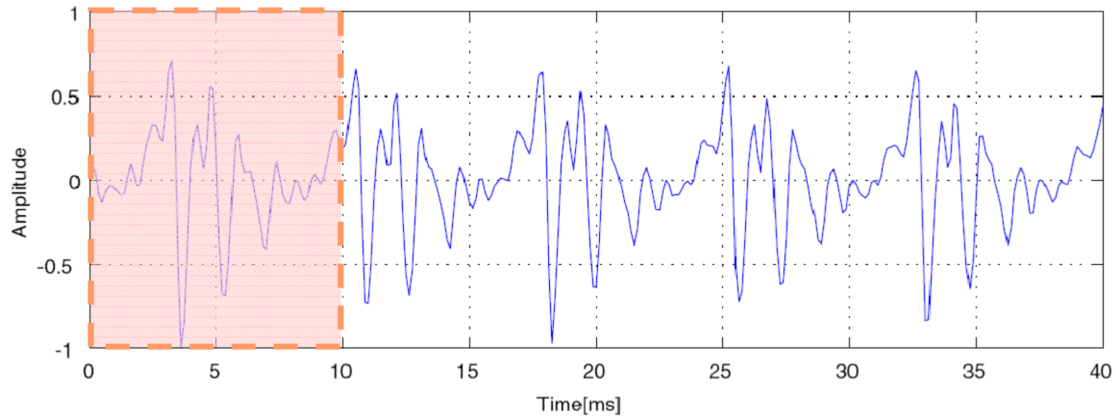


“o” in “spot”



“e” in “again”

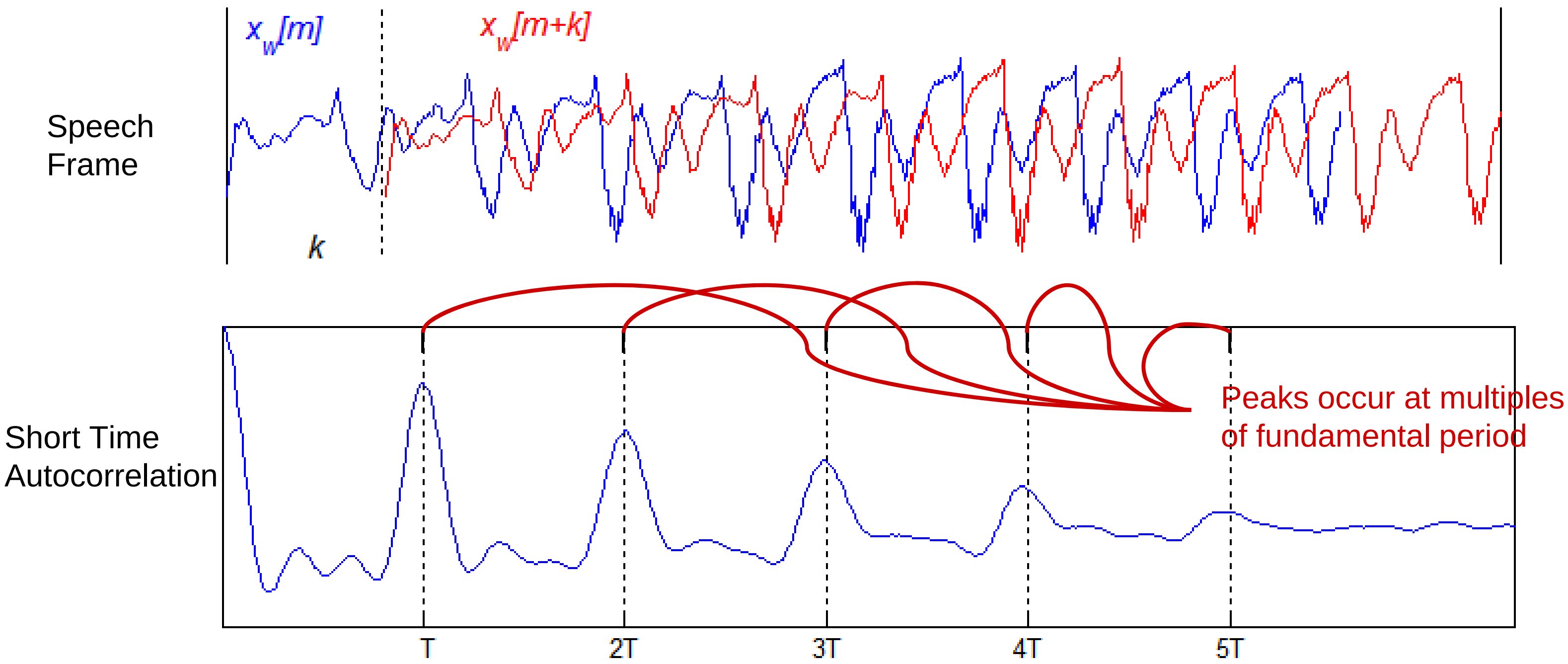
- Created due vocal cords vibrations at frequencies between 50 Hz & 1 kHz.
- Vibrations produce a quasi-periodic pressure wave which excites the vocal tract.
- Frequency of the pressure signal is the pitch frequency or fundamental frequency (F0).
- Voiced waveforms are quasi-periodic.



Autocorrelation of Voice Frames

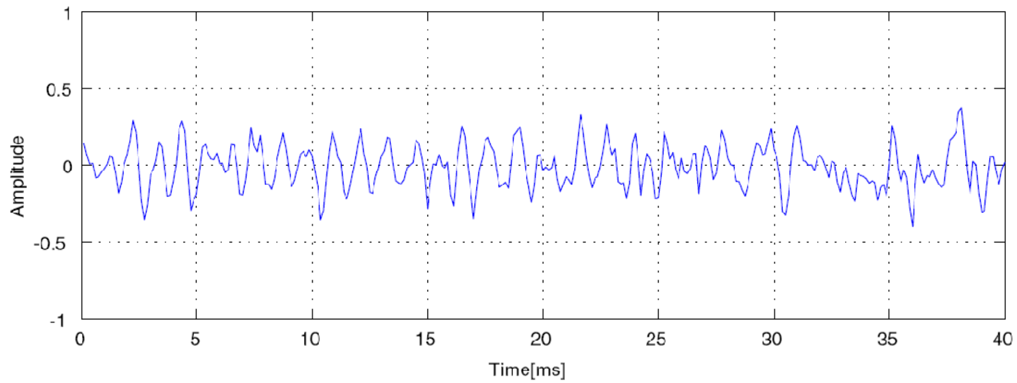
- **If $x_w[n]$ is a speech frame, then autocorrelation $A(k)$ is**

$$A(k) = \sum_m x_w[m] \cdot x_w[m+k]$$



Some properties of speech
Consonants (unvoiced sounds)

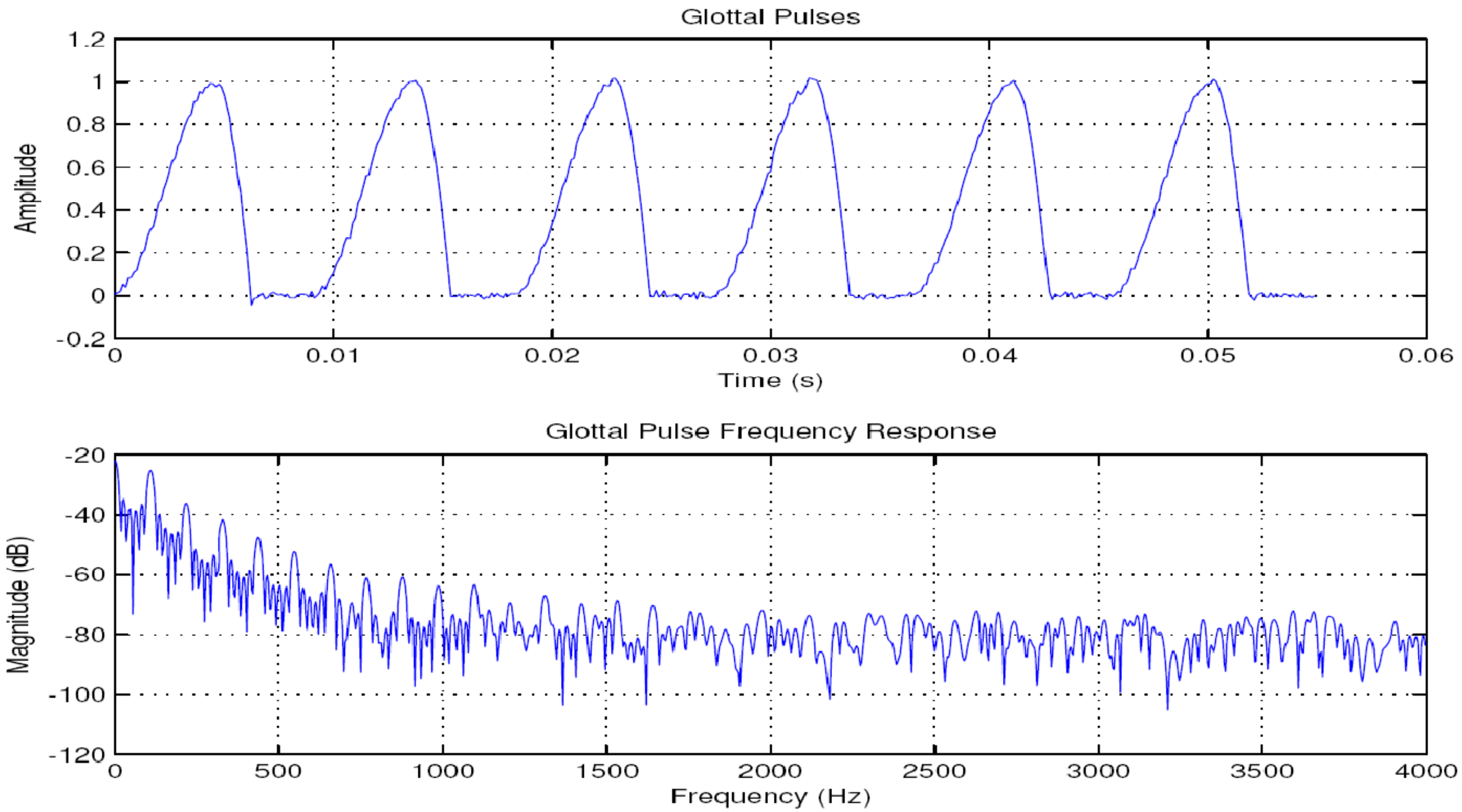
- Generated by forming a constriction at some point in the vocal tract, such as the teeth or lips, and forcing air through the constriction to produce turbulence.
- This is regarded as a broad-spectrum noise source to excite the vocal tract.



Defining terms

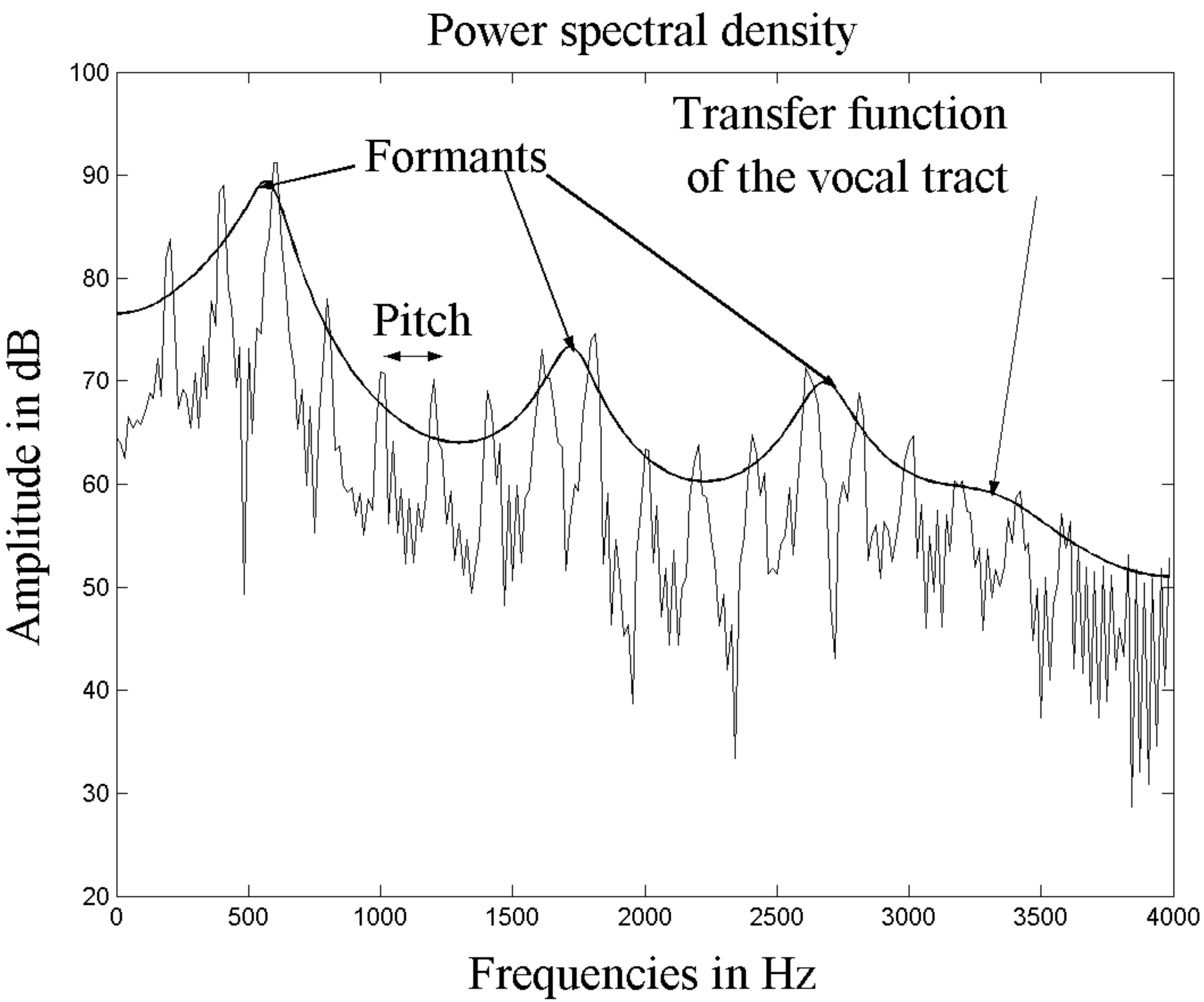
- **Fundamental frequency** (F0) is an **acoustic term** referring to the signal itself: how many pulses per second does the signal contain.
  - In the case of speech signal, each pulse is produced by a signal vibration of the vocal folds.
  - The frequency of these pulses is measured in Hertz (Hz).
- **Pitch** is a **perceptual term**.
  - What is the hearer's perception of this signal: is it heard as high in pitch or low in pitch, the same pitch as the previous portion of the signal, or different?
  - Pitch can be a property of speech or non-speech signals. I.e., music, high-pitched scream, bird-call.
- **Tone** is **a linguistic term**.
  - Tone refers to a phonological category that distinguishes two words or utterances.
  - Tone is a term relevant for language, and only for languages in which pitch plays a linguistic role (convey meaning).

Typical impulse sequence of a voiced sound



- Typical impulse sequence (sound pressure function) produced by the vocal cords for a voiced sound (glottal pulses).
- It is the part of the voice signal that defines the speech melody.

Speech Spectrum for a Voiced Sound



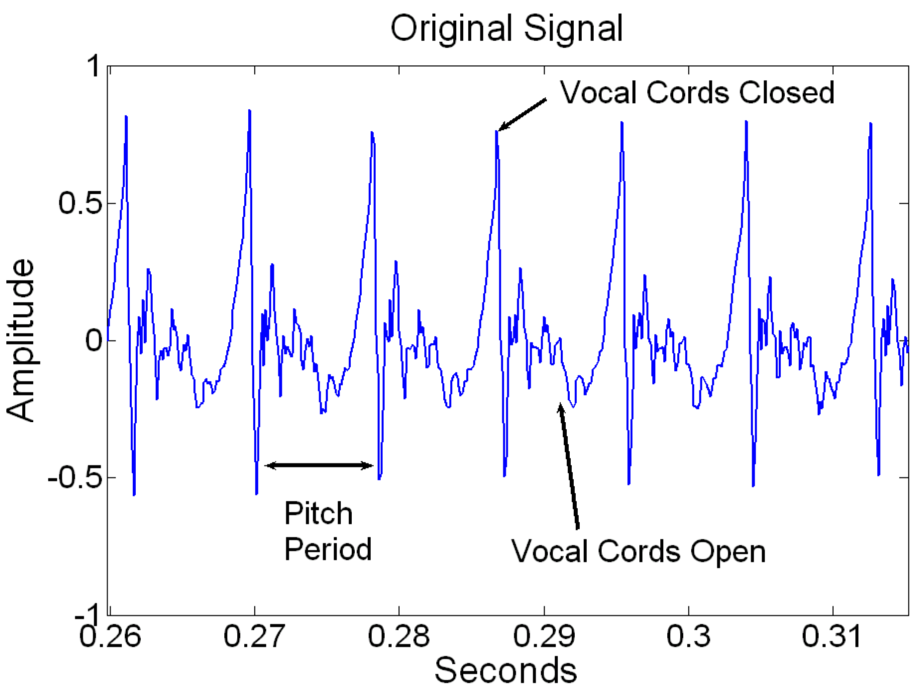
Formants

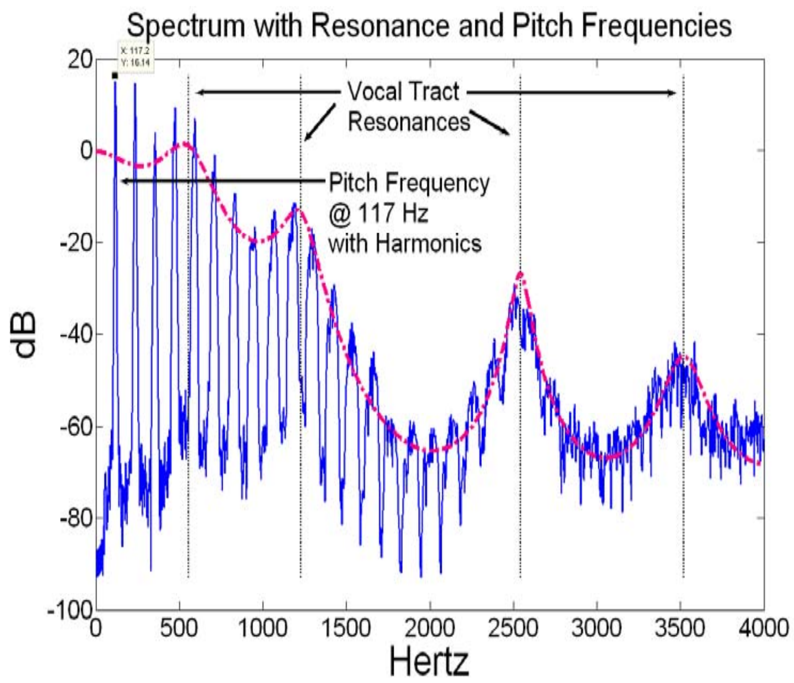
- Resonance frequencies of the vocal tract.
- Shapes and filters the sound of vocal cords.

Fine (pitch) harmonic structure:

- Attributed to the vibrating vocal cords (narrow peaks).

Speech Spectrum for a Voiced Sound

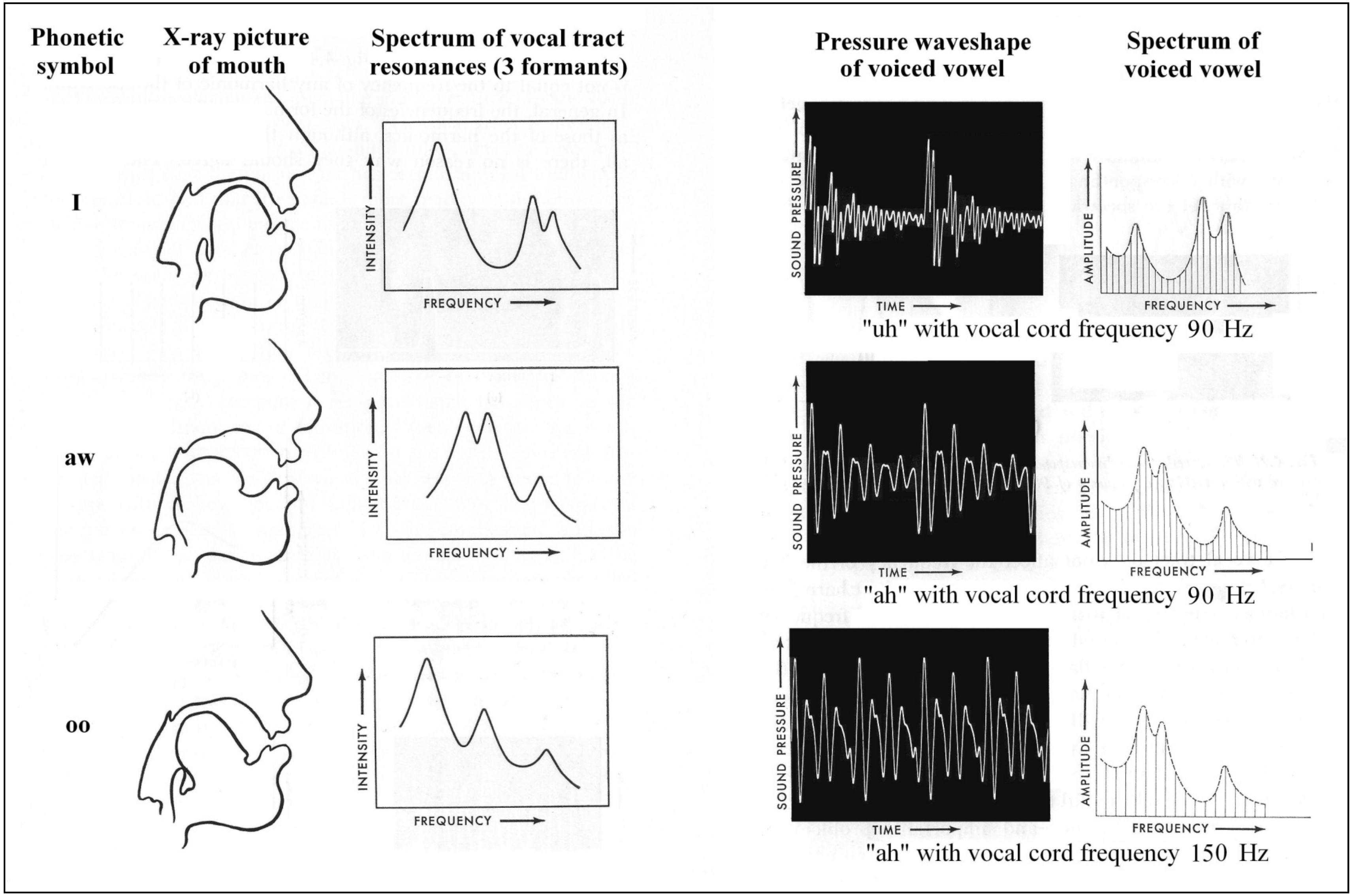




- Fine (pitch) harmonic structure:
  - reflects the quasi-periodicity of speech
  - attributed to the vibrating vocal cords (narrow peaks).
- Formant structure (envelope peaks):
  - due to the interaction of the source and the vocal tract (resonances of the vocal tract).
  - Short term correlation in the time domain.

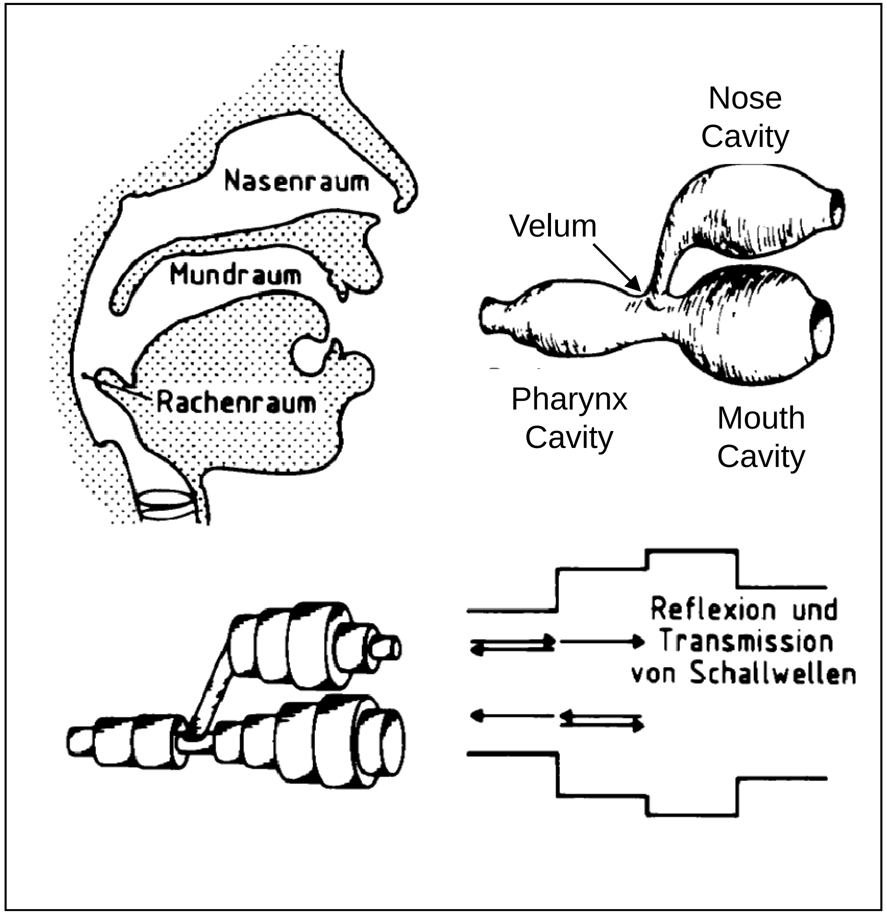
- 1st formant 150-850 Hz
- 2nd formant 500-2500 Hz
- 3rd formant 1500-3500 Hz
- 4th formant 2500-4800 Hz

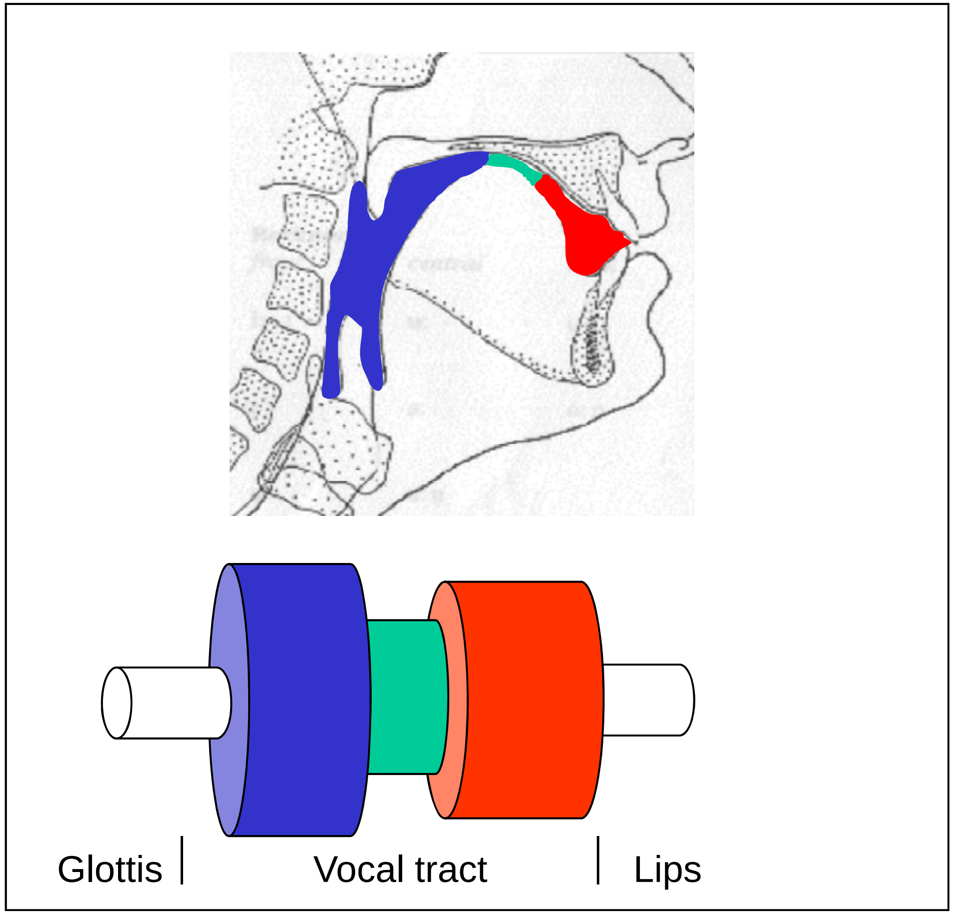
Phonetic speech mouth, formants, wave, and spectrum



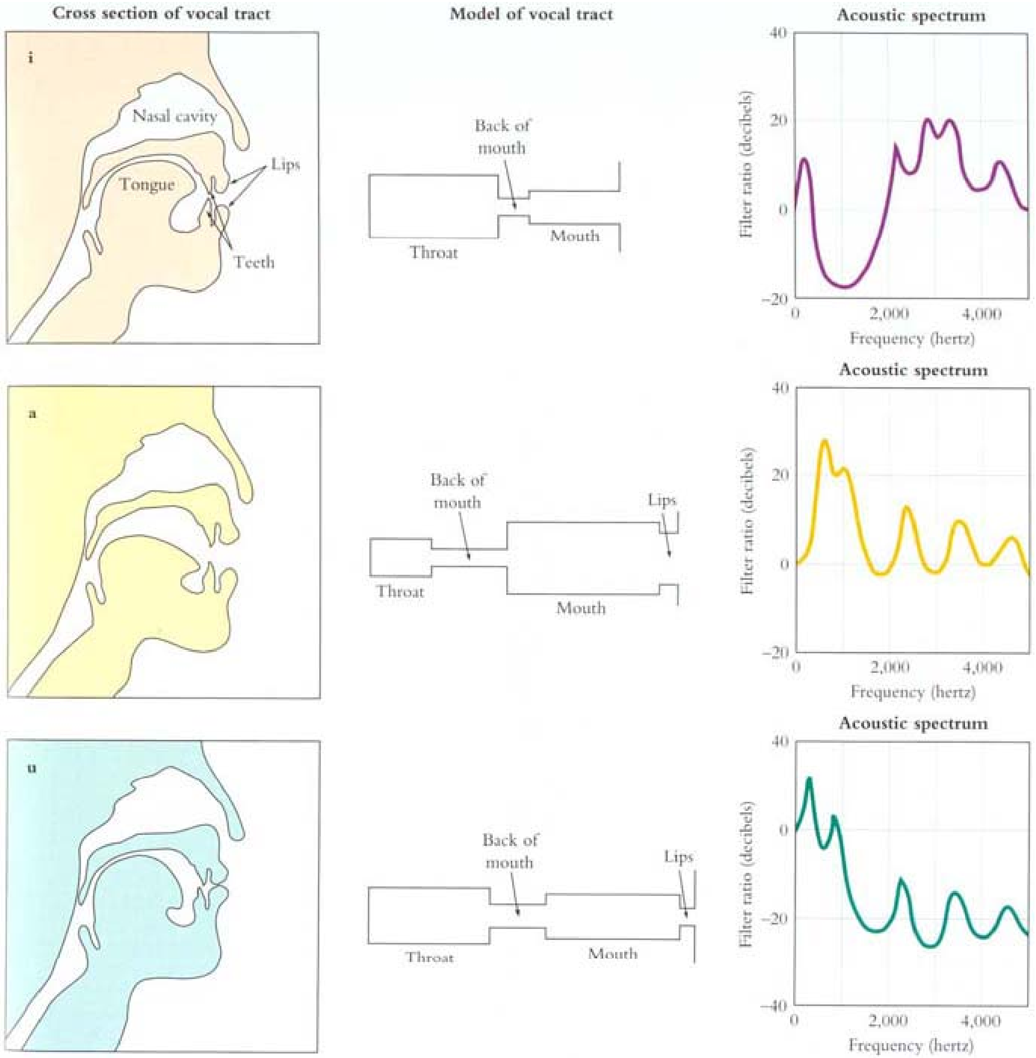
Vocal tract model

The vocal tract can be modeled as a hard-walled lossless tube resonator consisting of N tubes with different cross-sectional areas





Vocal tract formants



$$F_n = \frac{(2n - 1)\,c}{4L}$$

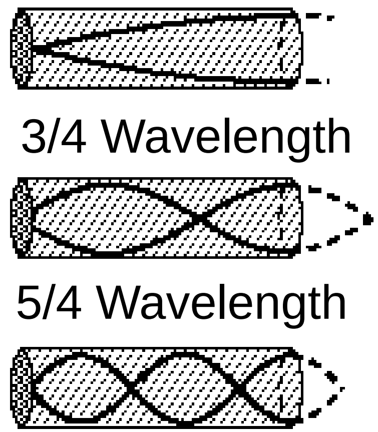
F_n = nth formant freq. [Hz]

L = length of the tube [m]

1/4 Wavelength

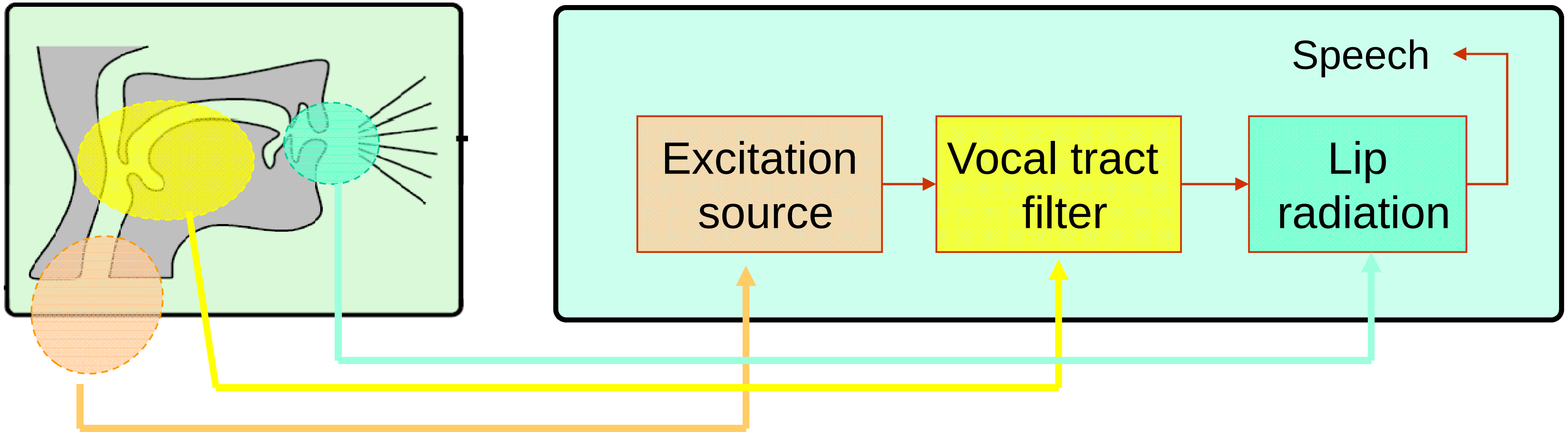
3/4 Wavelength

5/4 Wavelength



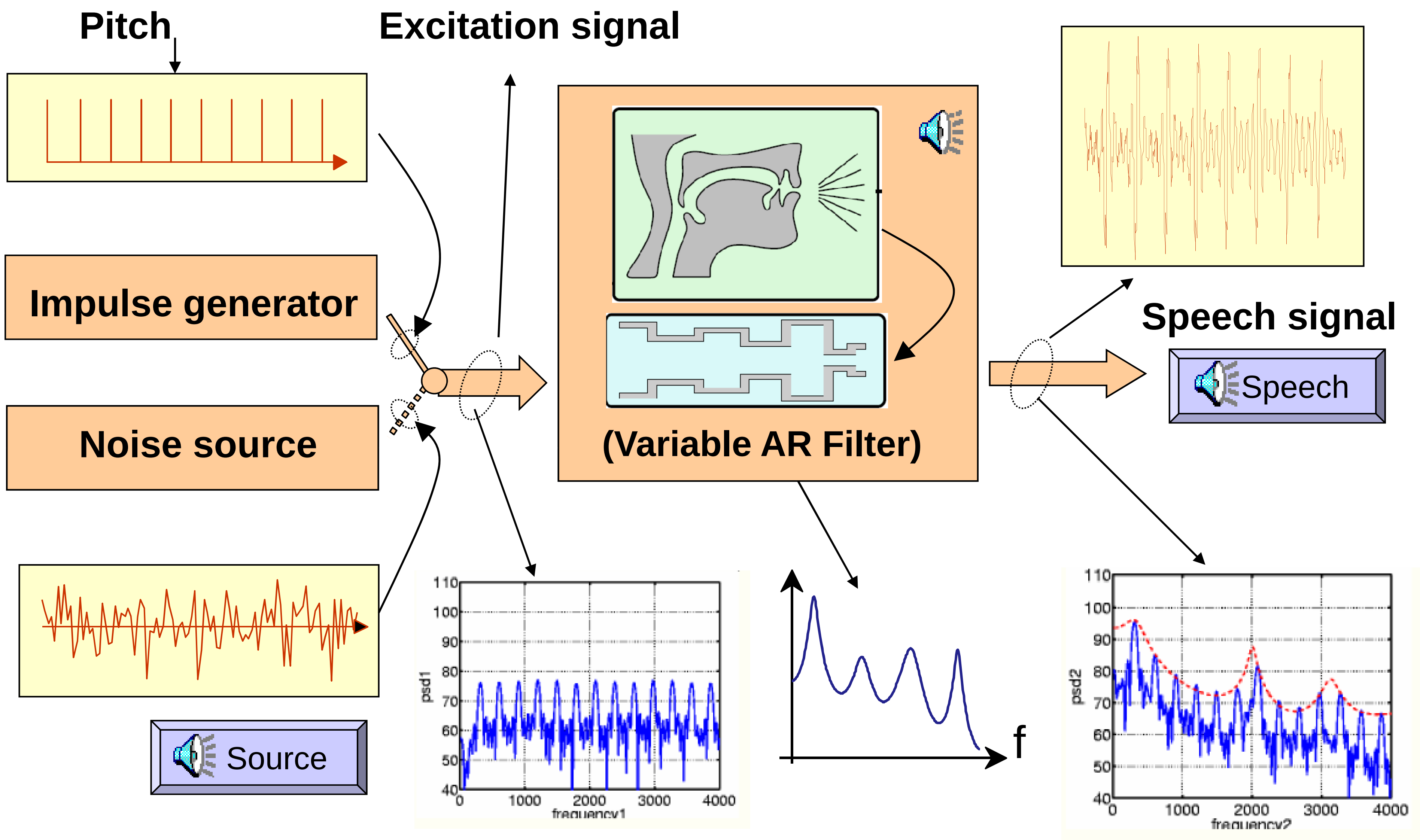
First 3 resonances of tube with 1 closed end

Source-filter theory of human speech production

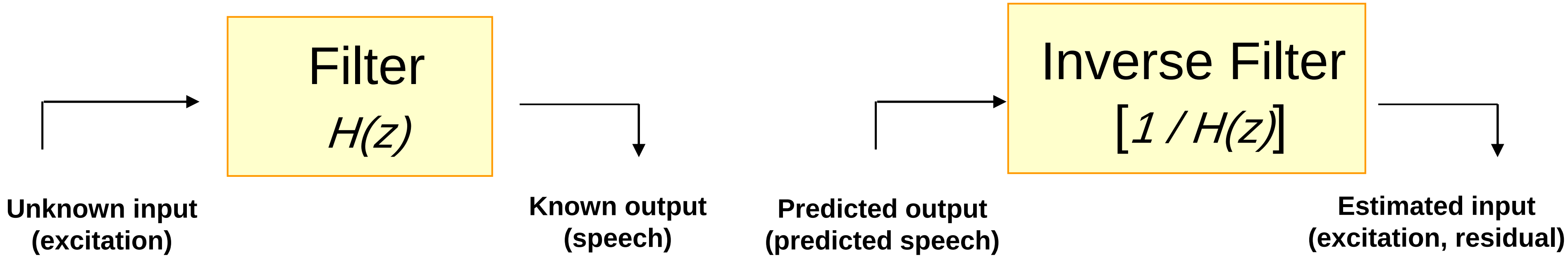


- Voice production mechanism can be modeled as a series connection of an excitation source and a filter system.
- Source and filter are considered independent of each other.
- Glottal pulses correspond to source and vocal tract corresponds to the filter
- In the case of voiced speech sounds, the excitation is modeled by an impulse train corresponding to the glottal pulses.
- In the case of unvoiced sounds, the excitation is modeled by a noise-like signal.
- The vocal tract functions as a phone dependent filter.
- The theory provides a theoretical background for the inverse filtering technique.

Source-Filter model of speech production

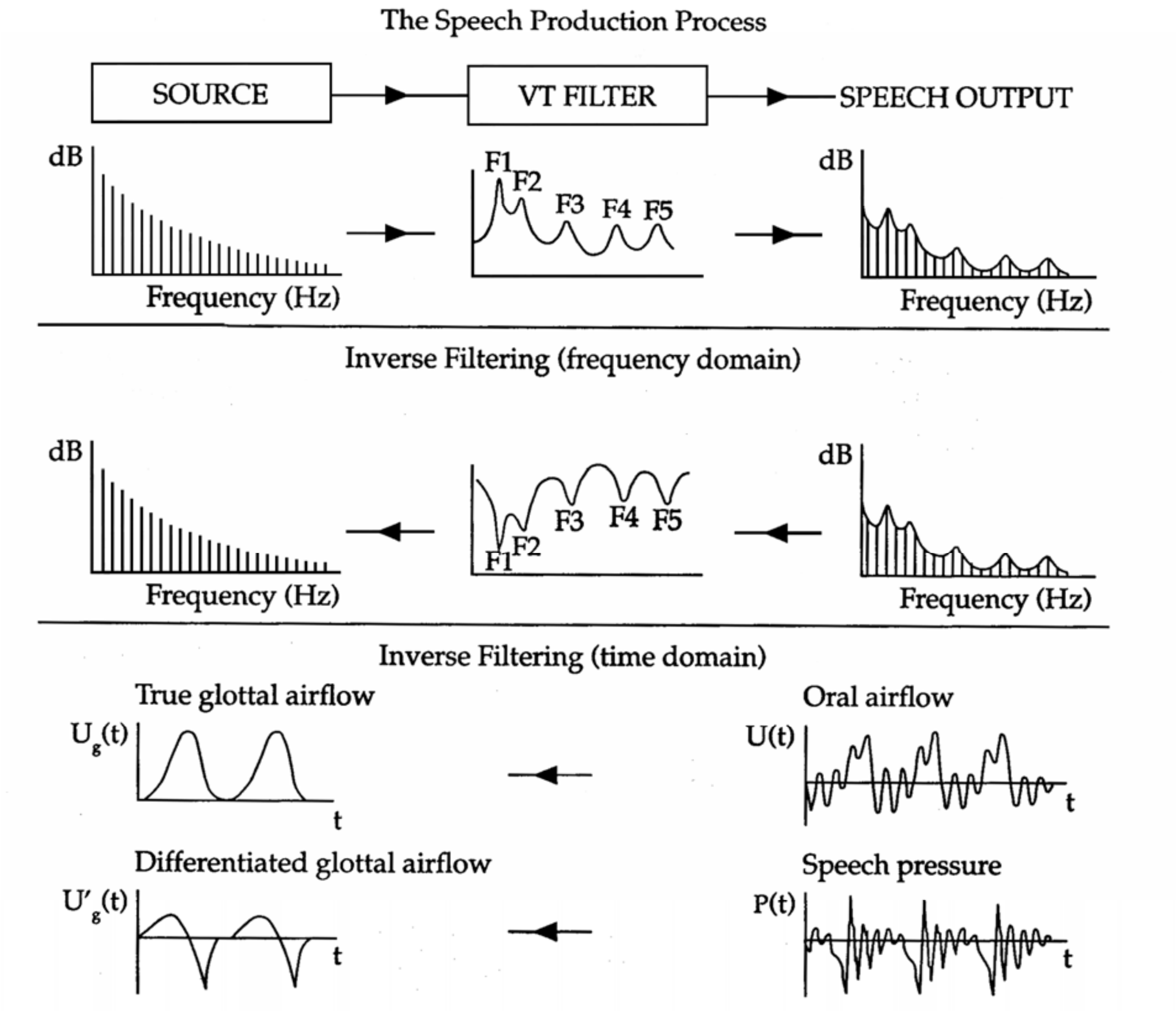


Source-filter theory of human speech production



- The source-filter theory of speech production provides theoretical background for the inverse filtering technique.
- If the transfer function of the vocal tract filter is known, an inverse filter can be constructed.
- Inverse filtering basically involves extracting two signals, the volume velocity waveform at the glottis, and the effect of the vocal tract filter, from a single source signal. It simulates the inverse characteristics of the vocal tract
- In principle, the glottal excitation signal can then be reconstructed by feeding the speech signal through the inverse of the vocal tract filter.
- The result of inverse filtering has to be regarded as an estimate of the glottal flow. The actual volume velocity waveform at the glottis is not known exactly.
- Automatic methods build a vocal tract model and automatically find filter parameters, often by means of LPC analysis.

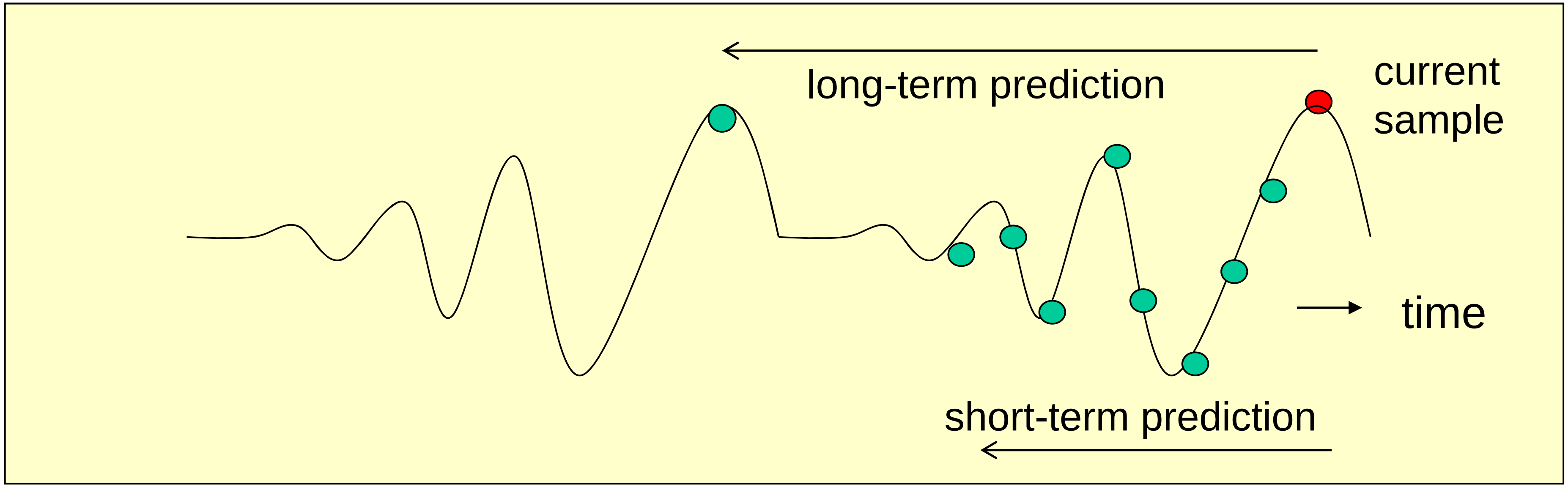
Inverse filtering



Filtering

Inverse Filtering

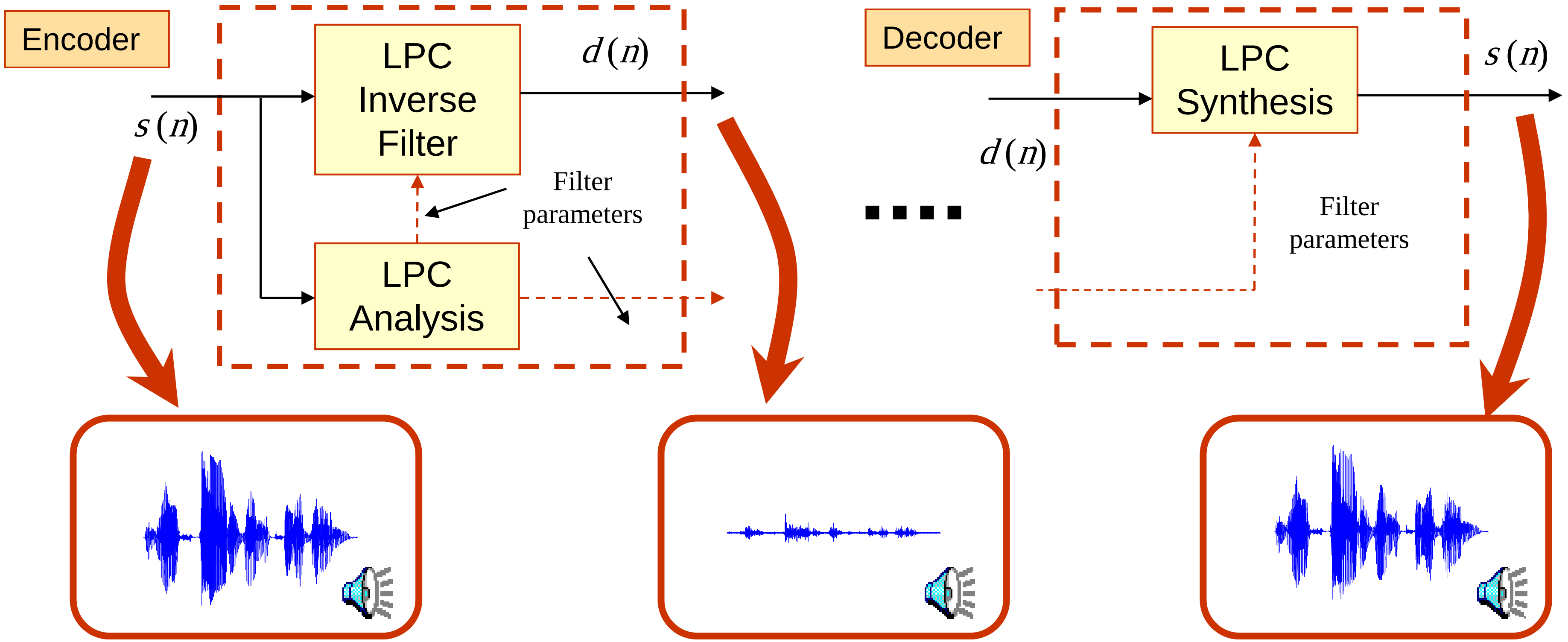
Predictive Coding – Basic principle



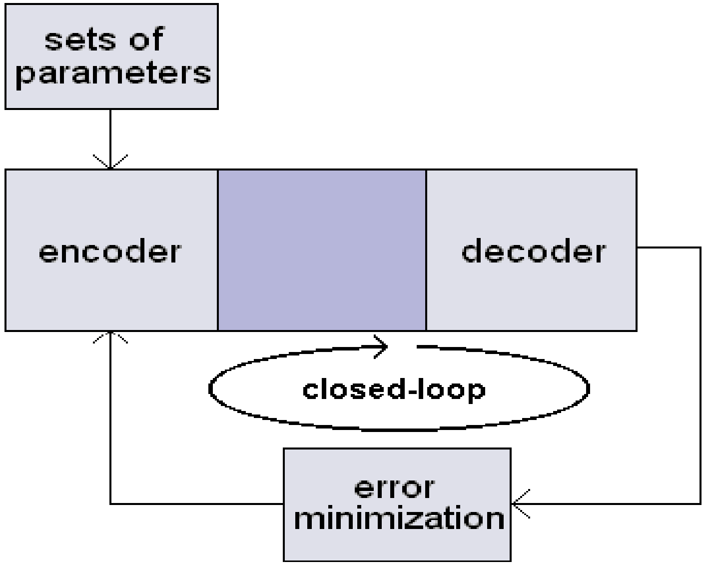
- Because of resonance properties of the vocal tract there is high correlation between successive samples of the speech signal (redundancy).
- Predictive Coding: Remove redundancy between successive pixels and only encode residual between actual and predicted.
- Residue usually has much smaller dynamic range, allowing fewer quantization levels for the same mean square error (= compression).
- short-term (resonance of vocal tract)
- long-term (periodicity of voiced speech (vocal cord vibration))

Linear Prediction

- Speech is segmented into frames with a length of 20 ms each, each frame contains 160 samples.
- Due to the vocal tract resonances, there is a correlation between subsequent samples.
- Each sample can be represented by a linear combination of past samples.
- With mathematical analysis, vocal tract emulating parameters can be found.
- LPC analysis: extract the parameters, LPC (inverse) filtering: find the excitation signal.

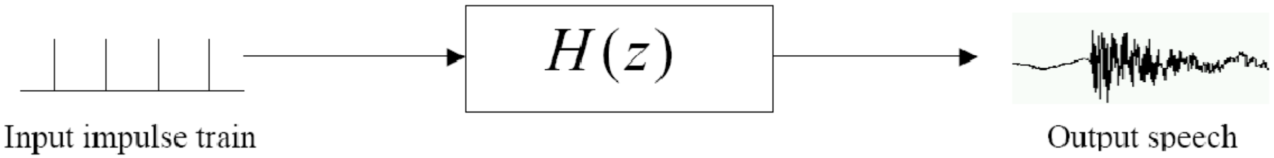


Analysis-by-synthesis (AS)



- Synthesize the input speech
- Compare to the original
- Try all sets of parameters
- The set that minimizes the error is used
- This is called closed-loop quantization

Linear prediction Coding



To predict H(z) from output speech signal, we use linear prediction analysis or auto-regressive(AR) modeling.

- Basic Idea :

$$H(z) = \frac{X(z)}{E(z)} = \frac{1}{1 - \sum_{k=1}^{p} a_k z^{-k}} = \frac{1}{A(z)}$$

  - where *E(z), X(z),* and *H(z)* are Z-transform of e(n), x(n), and h(n). e(n) is input impulse train, x(n) is output speech signal, h(n) is vocal tract filter.
  - *A(z)* is called inverse filter

Linear prediction Coding

We have following relationship between input and output by inverse z-transform.

$$x(n) = \sum_{k=1}^{p} a_k x(n-k) + e(n)$$

- How to get coefficients of *H(z)*

We assume

- e(n) is impulse train and most of it is zero value
- $\tilde{x}(n) = \sum_{k=1}^{p} a_k x(n-k)$ is predictive output of *H(z)*

then,

$$e(n) = x(n) - \tilde{x}(n) = x(n) - \sum_{k=1}^{p} a_k x(n-k)$$

and, we define short term prediction error

$$E_m = \sum_n e_m^2(n) = \sum_n \{x_m(n) - \tilde{x}_m(n)\}^2$$

where, $x_m(n) = x(m+n)$ means the segment of *x(n)* in the vicinity of *m, and b*ecause we assumed e(n) is zero for most part of it, we choose a_k such that make E_m has minimum value.

Linear prediction Coding

- Autocorrelation Method

We define predictive error as

$$E_m = \sum_{n=0}^{N+p-1} e_m^2(n)$$

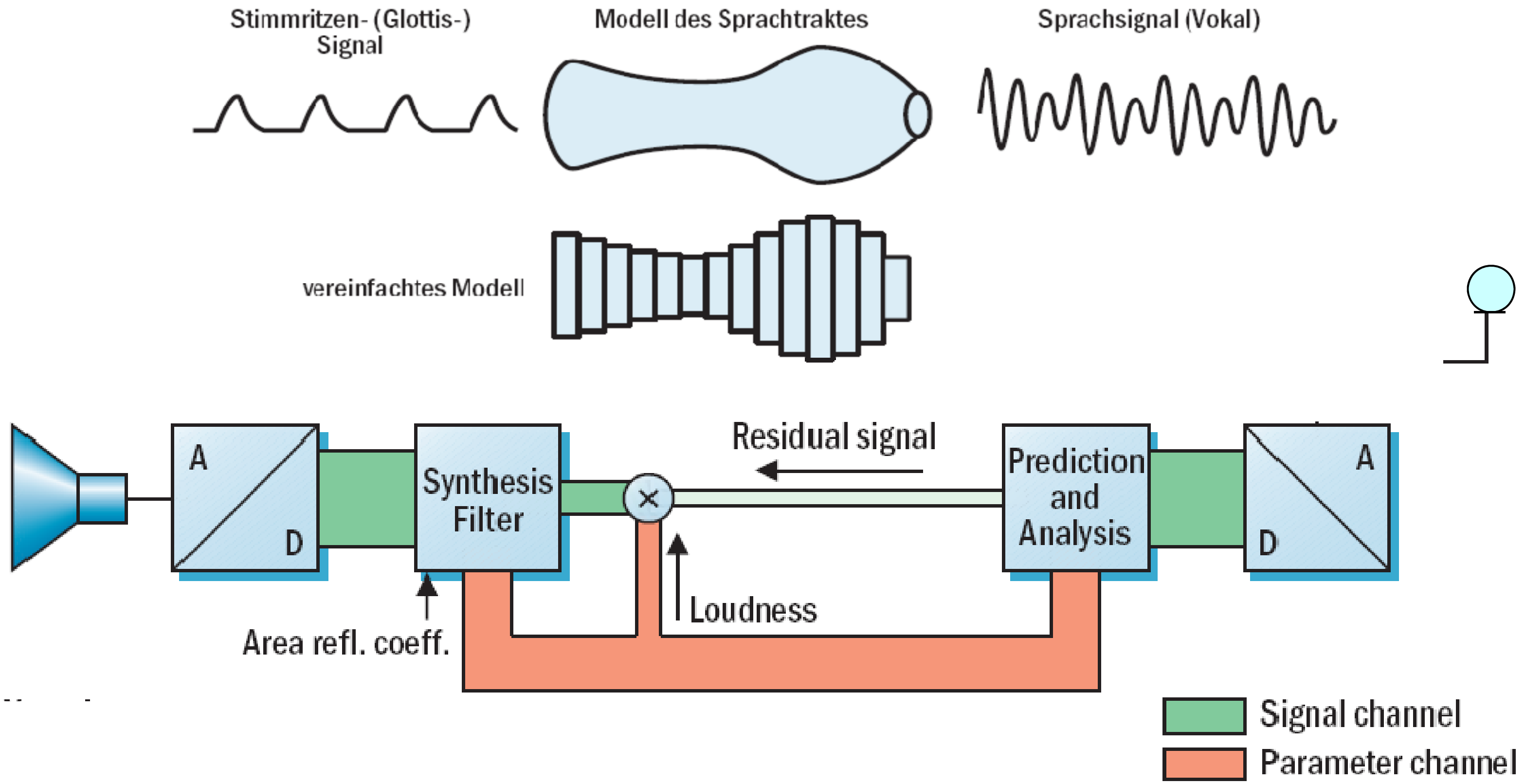
by minimizing this prediction error we get following equation

$$\begin{bmatrix} R_m(0) & R_m(1) & \cdots & R_m(p-1) \\ R_m(1) & R_m(0) & \cdots & R_m(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ R_m(p-1) & R_m(p-2) & \cdots & R_m(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} = \begin{bmatrix} R_m(1) \\ R_m(2) \\ \vdots \\ R_m(p) \end{bmatrix}$$

where,

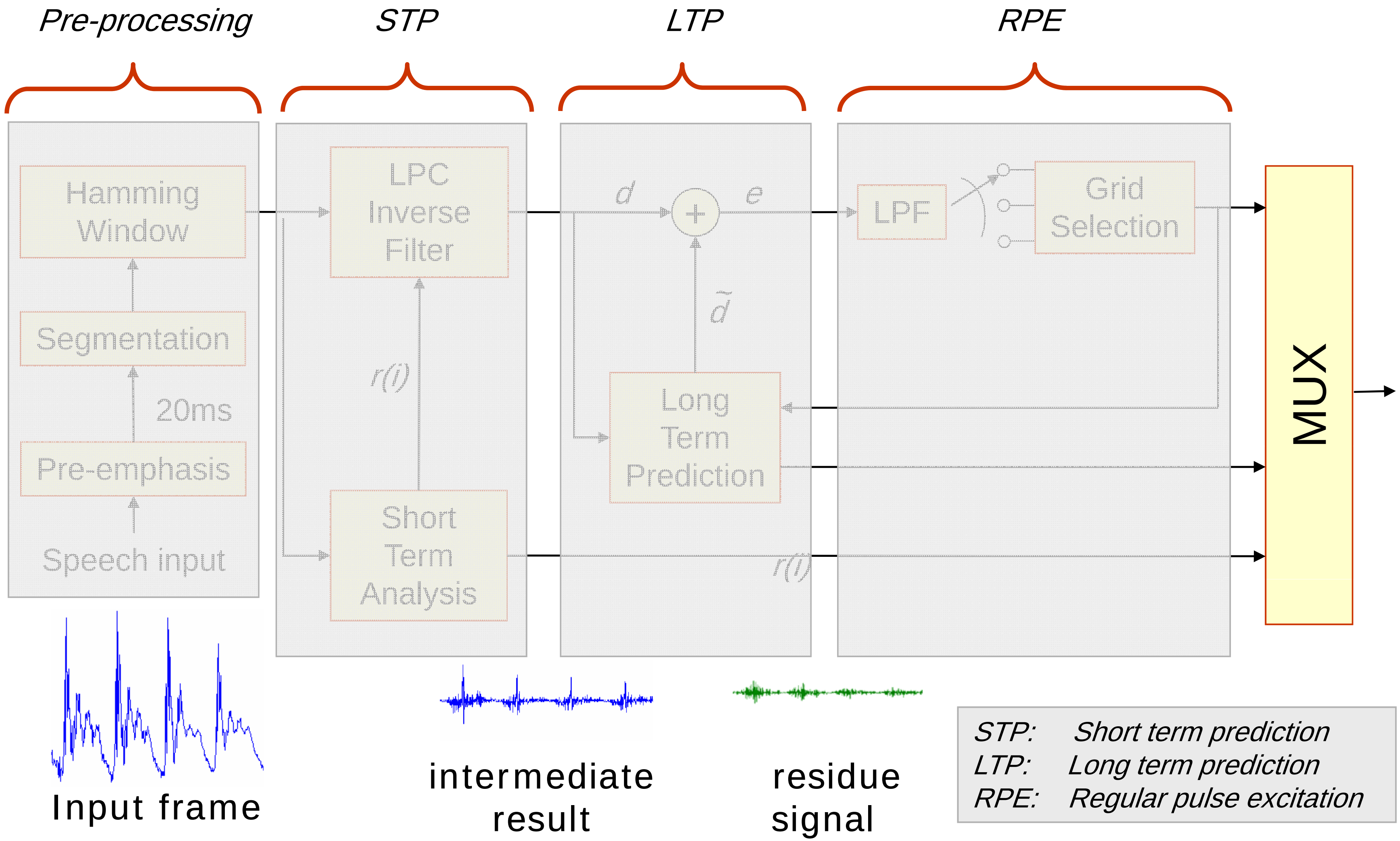
$$R_m(k) = \sum_{n=0}^{N-1-k} x_m(n) x_m(n+k)$$

Speech Coding – Linear Prediction



- LPC vocoders extract salient features of speech (formants) directly from the waveform, rather than transforming the signal to the frequency domain.
- LPC Features:
  - uses a time-varying model of vocal tract sound generated from a given excitation
  - transmits only a set of parameters modeling shape and excitation of the vocal tract, not actual signals.

GSM - Encoder

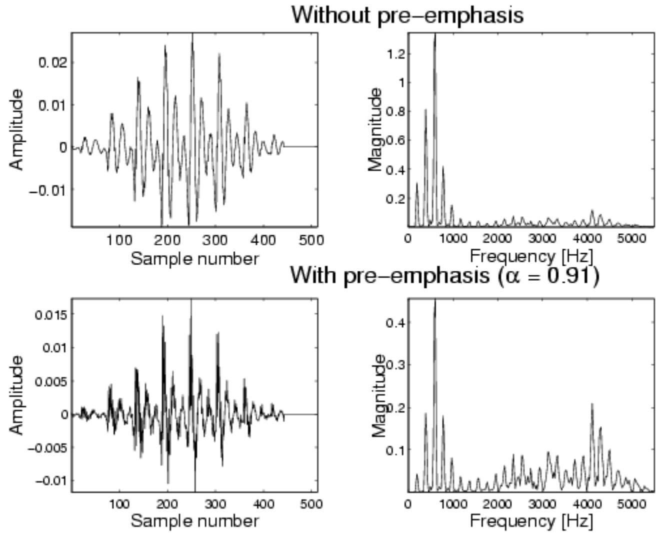


STP: Short term prediction
LTP: Long term prediction
RPE: Regular pulse excitation

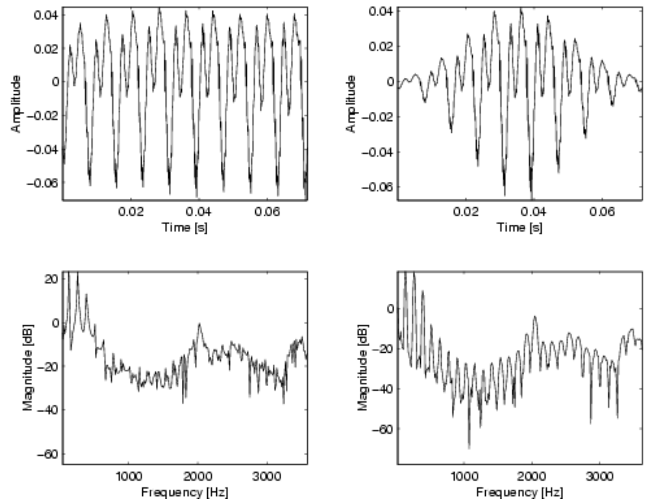
General Steps Before Feature Extraction

- **Pre-emphasis filtering :**
  - The natural attenuation that arises from voice source is about -12 dB/octave. Pre-emphasis makes higher frequencies of voiced sounds more apparent.

Usually: $H(z) = 1 - \alpha z^{1}$, with $\alpha \approx 1$



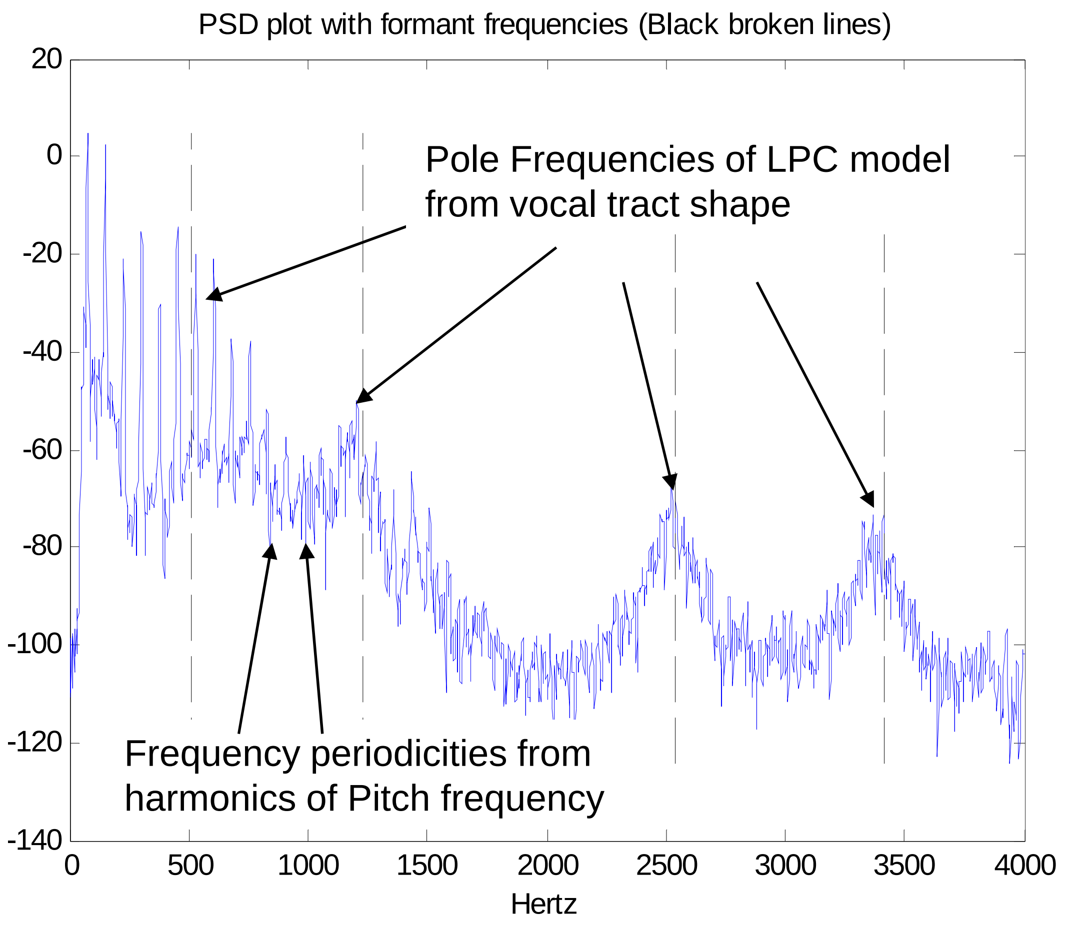
- **Windowing :**
  - Discrete Fourier transform (DFT) assumes that the signal is periodic. Windowing reduces the effect of the spectral artefacts (spectral leakage/smearing) that arise from discontinuities at the frame endpoints.
    Typically Hamming window is used.



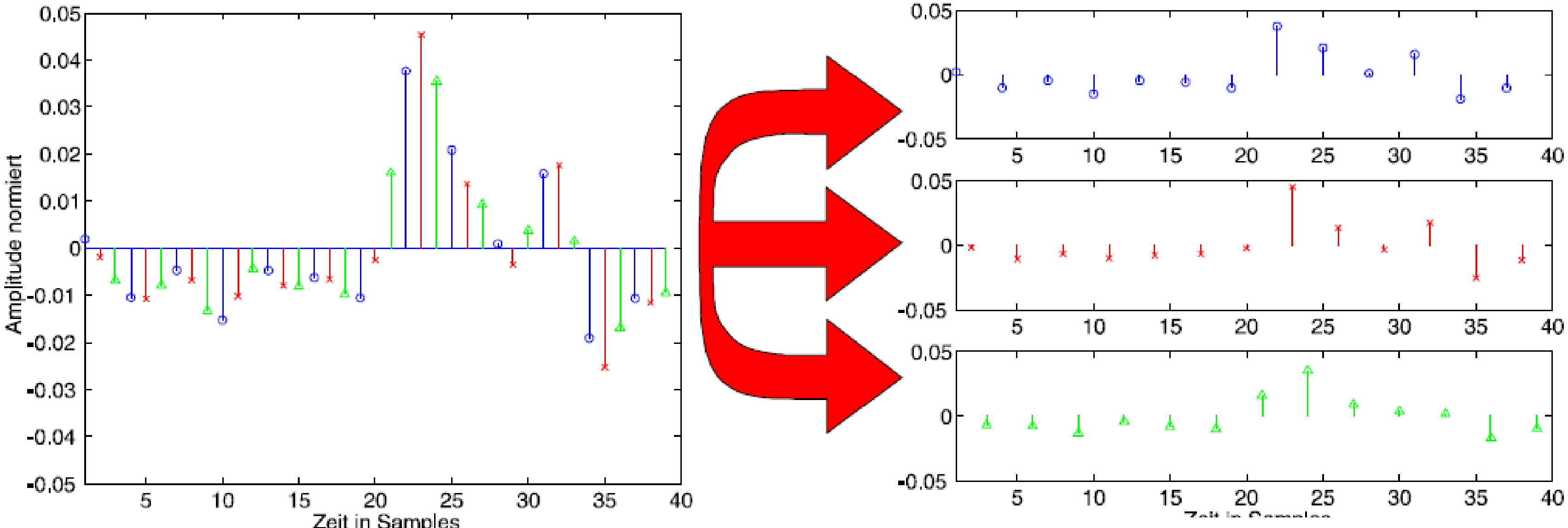
Short Term & Long term prediction

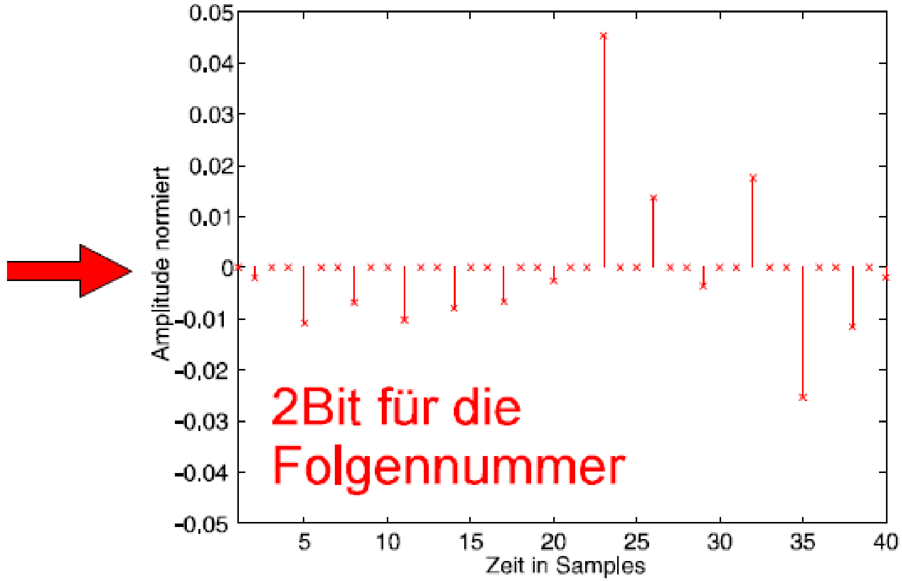
- Short Time Correlation refers to the predictability of the signal on a sample by sample basis.
- Primarily associated with the resonances of the vocal tract.
- Occurs within one pitch period.
- Short term predictor removes the short-term correlation and results in a glottal excitation signal

- Long term correlation refers to the predictability of the signals based on samples which do not immediately precede the current one.
- Primarily associated with the quasi-periodical nature of voice production (periodicity of vocal cords vibration).
- LTP (Long time prediction): try to reduce redundancy in speech signals by finding the basic periodicity or pitch that causes a waveform that more or less repeats
- Occurs across consecutive pitch periods
- Long-term/pitch prediction removes the correlation across consecutive periods.



Regular Pulse Excitation





- The residue is down sampled by a factor of 3.
- The series with the with the highest energy is selected, quantized, and transmitted.
- Extract 6-bit overall gain
- Encode remainder with 3 bits/sample

GSM Speech Coding

Regular pulse excited - Long Term Prediction-Linear predictive Coder

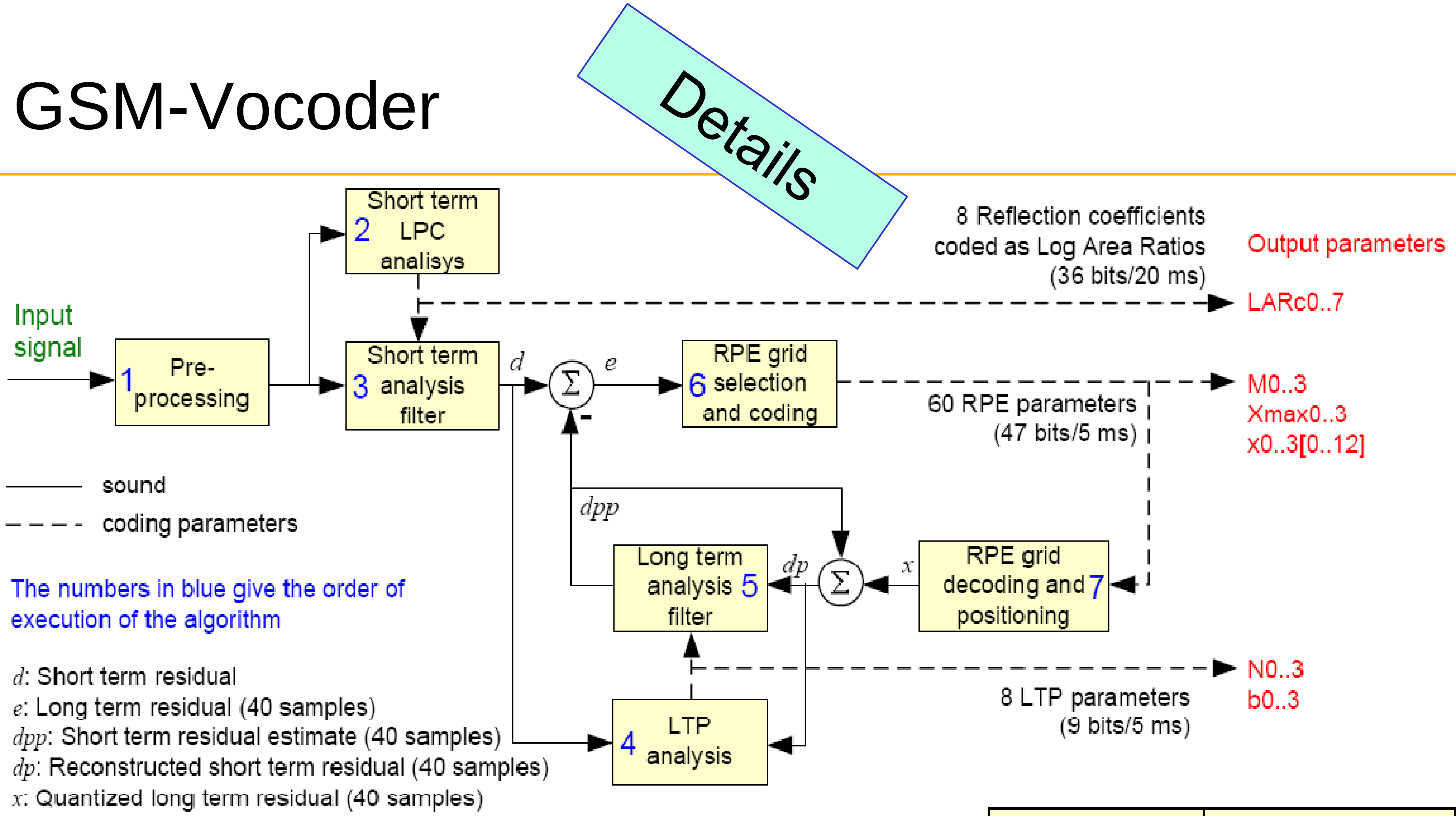
Input:

- A/D converted signal
- 8kHz sampling frequency
- Coded @ 13bit/sample
  = 104 kbps
- Frame of 20 ms duration
  = 2080 bits
- Divided by 13 bit/sample
  =160 samples

RPE-LTP Speech Encoder

	Bits per 20 ms
Linear Prediction Coding (LPC) filter	36
Long Term Prediction (LTP) filter	9
Excitation Signal	188
Total:	**260 bits/20 ms = 13 kbps**

GSM-Vocoder



		bits per 5 ms	Bits per 20 ms
Linear Prediction Coding (LPC) filter	▪ 8 parameters		36
Long Term Prediction (LTP) filter	▪ Nr (delay) ▪ br (gain)	2 7	28 8
Excitation Signal	▪ Sub-sampling phase ▪ Maximum amplitude ▪ 13 samples	2 6 39	8 24 156

Speech coding related delay

- Dividing the speech into 20 ms frames (160 samples each) leads to a signal delay of 20 ms.
- In addition there is additional delay due to mathematical calculations.
- In total, the overall delay is approximately 90 ms.
- In duplex communications, the delay starts to be noticebale and affects the communication quality when it reaches 300 to 500 ms.
- Therefore, 90 ms are tolearble.

Other voice coding standards

- **Half rate (HR)**
  - 6.5 kb/s used to increase capacity since it only needs a half rate TCH.
  - Speech quality is considerably less.

- **Enhanced Full Rate (EFR)**
  - 12.2 kb/s but further protected by CRC in a resulting 13 kb/s codec.
  - Better speech quality.

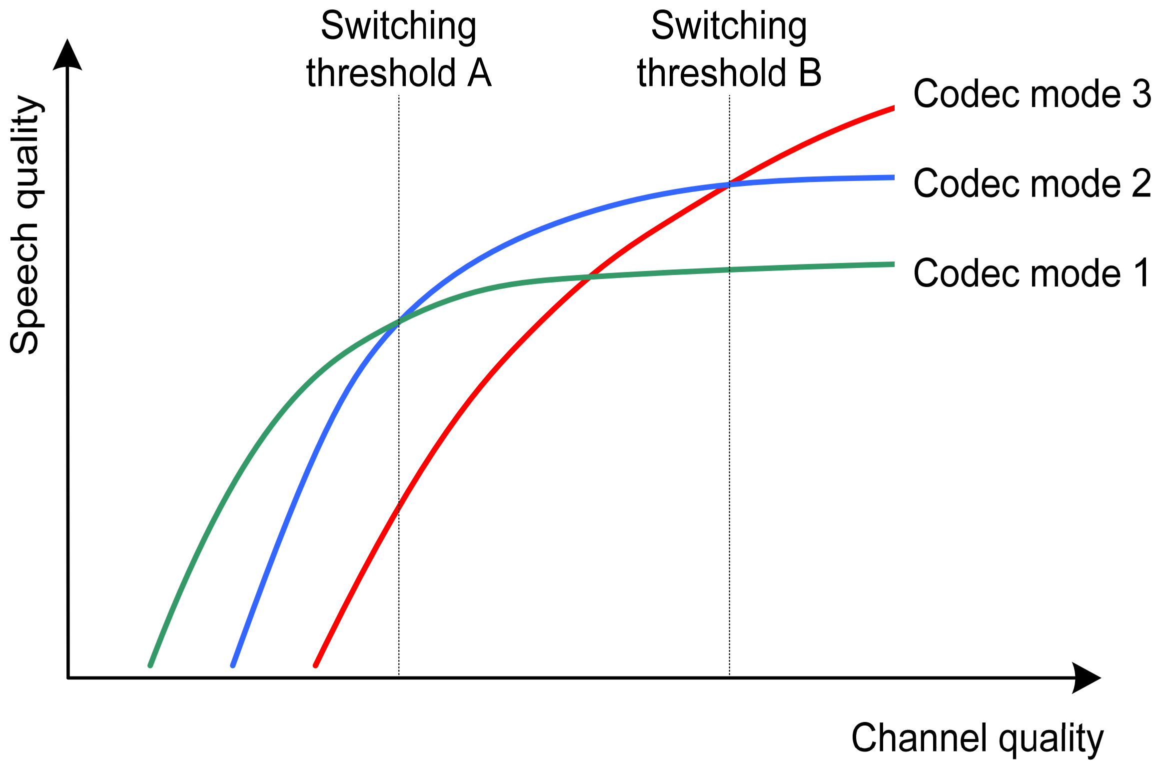
- **Adaptive Multi Rate (AMR)**
  - The codec adapts to the radio conditions.
  - High BER will use a low rate codec that is better protected.

Card games are fun to play

GSM Full Rate	
GSM Half Rate	
GSM EFR	

Adaptive Multi-Rate codec (1/2)

- The philosophy behind AMR is to lower the codec rate as the interference increases and thus enabling more error correction to be applied.
- The AMR codec is also used to harmonize the codec standards amongst different cellular systems.



Adaptive Multi-Rate codec (2/2)

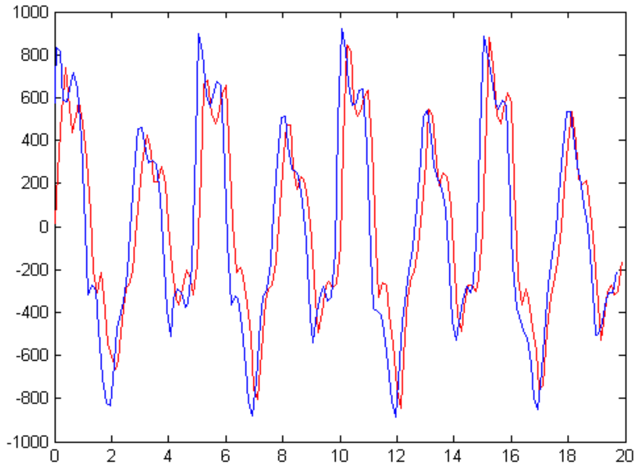
Narrowband AMR (GERAN Rel'98)

- Audio bandwidth equal to 300-3400 Hz.
- Speech quality lower than the quality of wireline communications.
- Most spectrum efficient codec among the speech codecs of GSM.

Wideband AMR (GERAN Rel'5)

- Audio bandwidth extended to 50-7000 Hz.
- The extension improves intelligibility and naturalness of speech.
- The quality of the highest codec modes exceeds the quality of 64 kbit/s PCM speech.
- High quality means more bits and reduced network capacity.

CELP (Code Excited Linear prediction)



• Assume that previously synthesized frames are highly correlated

• Correlate with different lags to obtain optimum delay

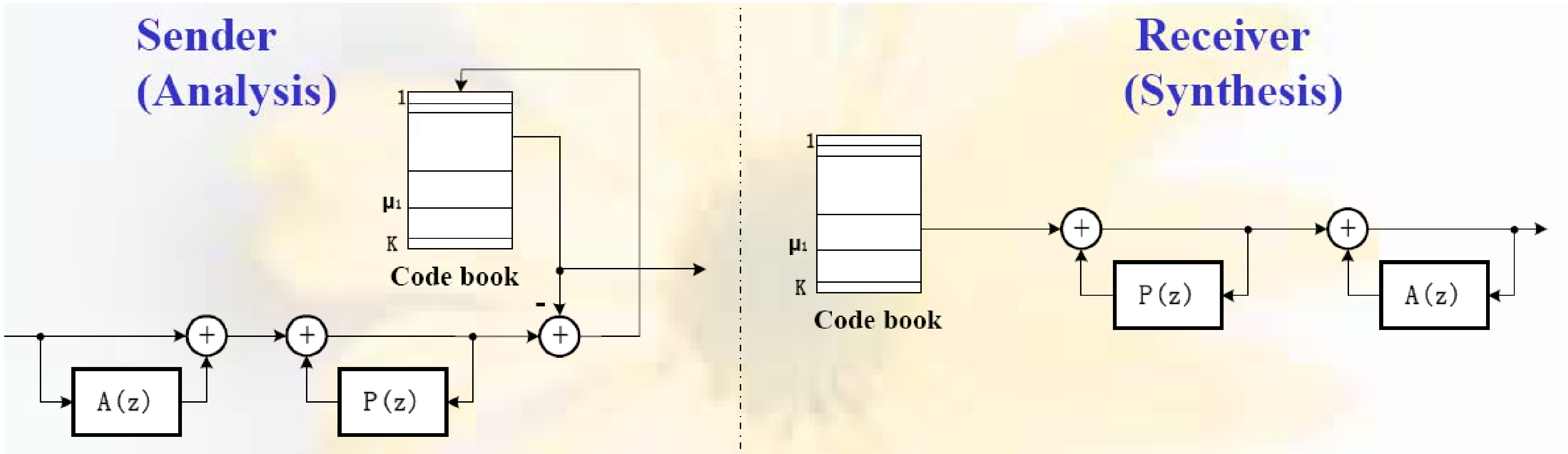
• Scale to fit

• Transmit delays and gains

....using only the adaptive excitation, the synthesis is still decent....

CELP (Code Excited Linear prediction)

- Coder and decoder have a predetermined *code book of excitation signals.*
- *Index of the code book where the best match was found is transmitted.*
- The receiver uses this *index to pick the correct excitation signal for its* synthesizer filter



The listening room

The listening room....

Take the opportunity to listen and compare....

The original (128 kbit/s)

Different codecs:

Normal precision (12,25 kbit/s)

Low precision (10 kbit/s)

An unstable formant filter has this effect....

This is the result of removing the algebraic codebook....

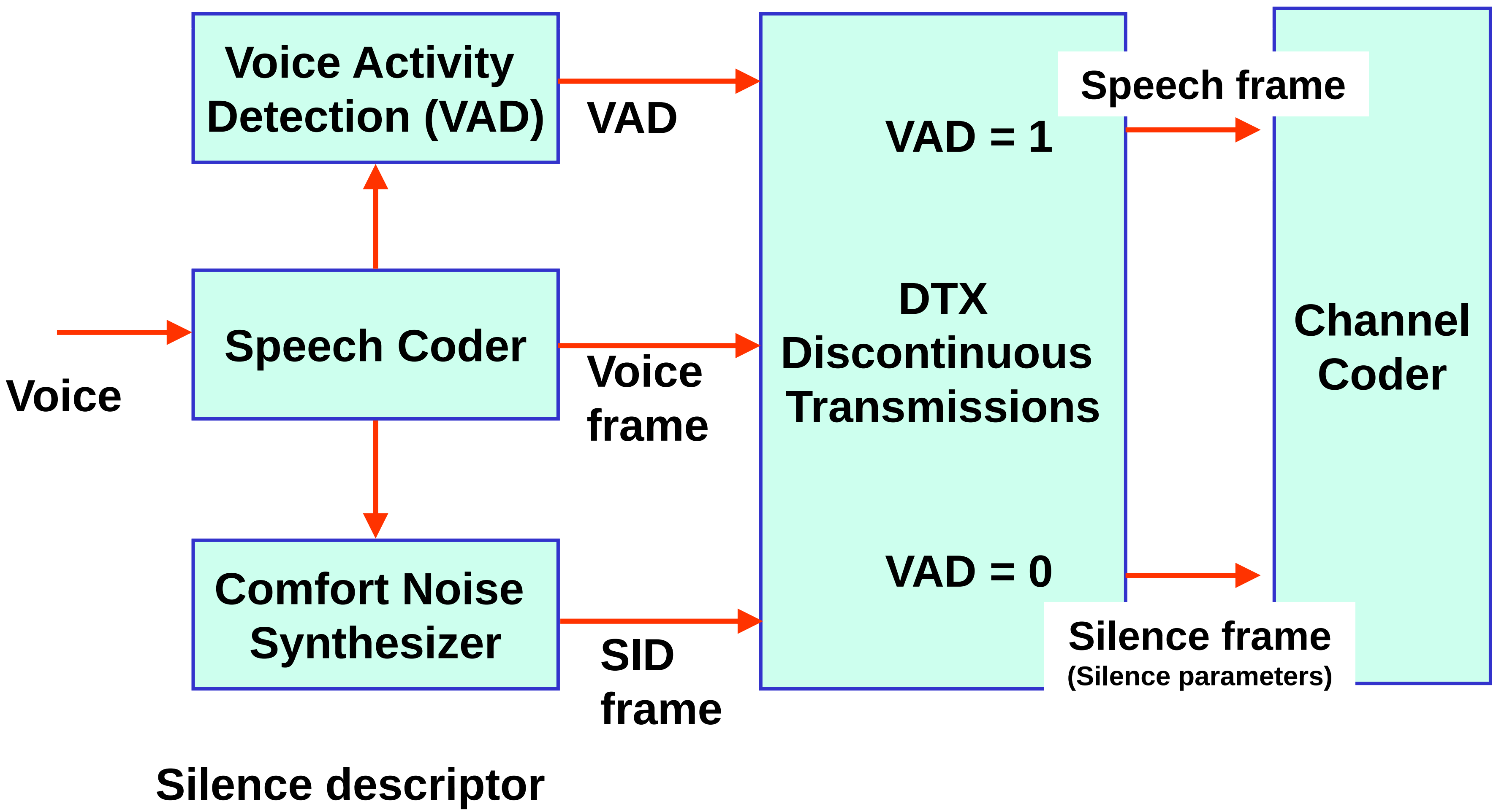
....and if we remove the adaptive codebook....

use white noise excitaion instead of the codebooks....

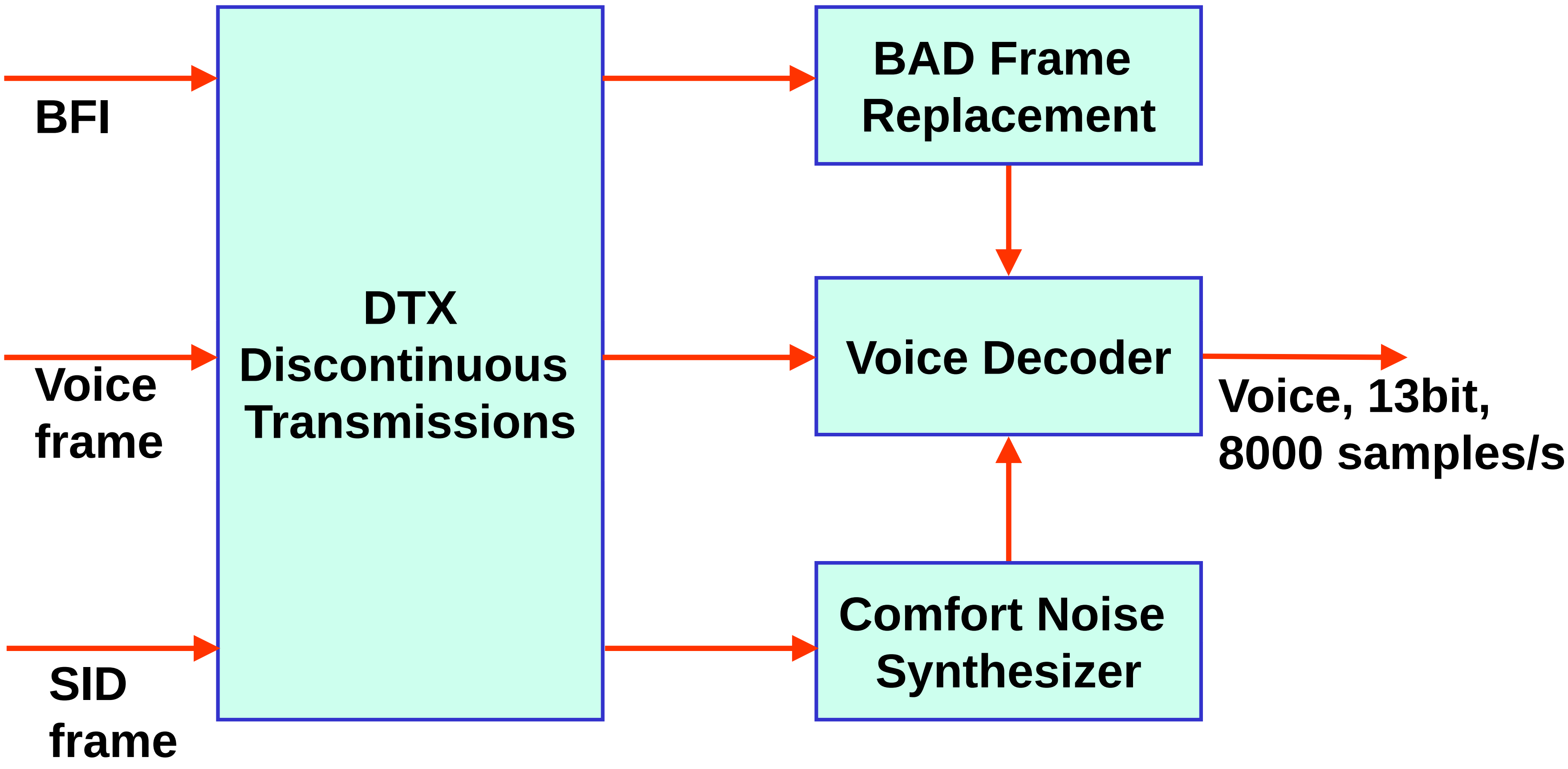
Discontinuous transmission

- **Speech coder implements Voice Activity Detection (VAD)**
  - Voice activity: idle for about 40% of the time.
- **When IDLE, do not transmit**
  - Reduced battery consumption.
  - Reduced interference.
- **Receiver side: silence is disturbing!**
  - Missing received frames replaced with “comfort noise”.
  - Comfort noise spectral density evaluated by TX decoder.
  - And periodically (480ms) transmitted in special frame (SID= Silence Descriptor).

Speech functions at the transmitter



Speech functions at the receiver



Bad Frame Indication (BFI)
Silence Descriptor (SID)

Discontinuous transmission

DTX coder

- The coder will detect if the sample is voice or "silence". When detected the Voice activity detection will set a VAD bit to 1.
- The voice coder will output a voice frame of 260 bits or a SID frame of 35 bits.
- Depending on the VAD bit the DTX will output voice frames or SID frames.
- The DTX will pass the voice frames or *in band encoded* SID frames to the channel coder.

DTX encoder

- Correct voice frames are passed directly to the voice decoder.
- Correct SID frames are passed to the comfort noise synthesizer.
- If the Bad Frame Indicator (BFI) is set then
  - an incorrect voice frame is replaced by the previous
  - an incorrect SID frame is replaced by the last valid SID frame or last valid speech frame (that probably contains noise).

Summary – Source Coding

- **Source coding: compression of the voice signal.**
- **Vocoding: Extraction of the vocal tract parameters.**
- **Vocal tract parameters & residual signal are sent instead of the signal itself. which are sent to the receiver.**
- **Signal is regenerated at the receiver.**