Speech Signal Analysis

Dr. Mathew Magimai Doss

September 21, 2022

Outline

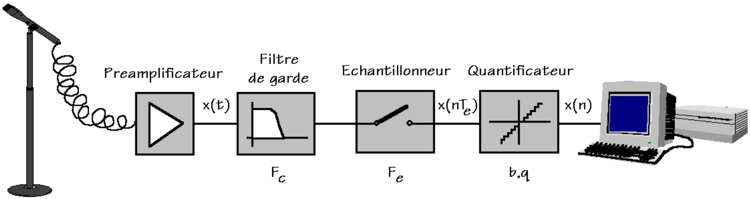
Speech signal acquisition

Time domain analysis

Frequency domain analysis

Outline

Speech Acquisition



- Convert acoustic signal to electrical analog signal
- Low pass filter the electrical analog signal with a cut-off frequency F_c
- Discretization of the low pass filtered electrical signal across time (Sampling)
- Discretization of amplitude of each sample (Quantization)

Sampling

- Apply Nyquist-Shannon's sampling theorem
- Low pass filter the electrical signal with a cut-off frequency F_c
  - Telephone speech: 4000 Hz (dictated by analog speech transmission bandwidth)
  - Microphone speech: 8000 Hz
  - CD quality speech: 22050 Hz (covers the entire auditory frequency range)
- Sample the low pass filtered electrical signal at frequency F_s (sampling frequency)

$$F_s = 2 \times F_c$$

  - Telephone speech: 8000 Hz
  - Microphone speech: 16000 Hz
  - CD quality speech: 44100 Hz

Guarantees reconstruction of the analog signal without aliasing

Quantization

- Sampling yields discrete signal, i.e. time is discrete but amplitude is continuous
- Quantize the amplitude of each sample to get the digital signal
- How many bits for quantization 8 bits or 16 bits?
  - 8 bits: minimum amplitude is 1 and maximum is 255

$$\textit{largest sound} \quad = 20 \cdot \log 10\left(\frac{255}{1}\right) \quad = 48 \; dB$$

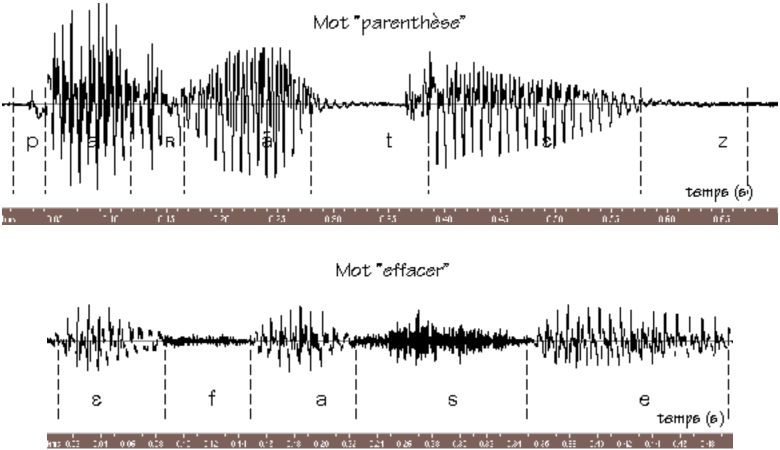
  - 16 bits: minimum amplitdue is 1 and maximum is 65535

$$\textit{largest sound} \quad = 20 \cdot \log 10\left(\frac{65535}{1}\right) \quad = 96 \; dB$$

- Digital speech signal can be stored and read as an array of short integers (2 bytes)
- Bitrate (bits per second) = $F_s \times$ (# of bits per sample)
  - Telephone speech: 8000 $\times$ 16 bps (2 bytes per sample) or 8000 $\times$ 8 bps (1 byte per sample through dynamic range companding, see μ-law algorithm, A-law algorithm)
  - Microphone speech: 16000 $\times$ 16 bps
  - CD quality speech: 44100 $\times$ 16 bps

Outline

Time domain speech signal



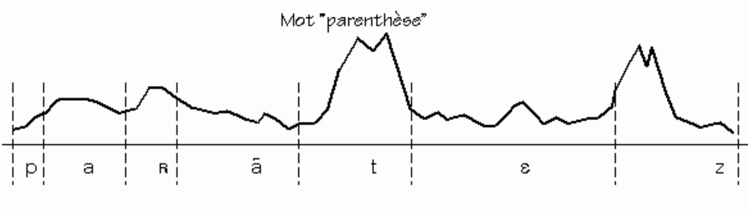
Some statistical measures

Mean: $\mu_s = \lim_{N->\infty} \frac{1}{2N+1} \sum_{n=-N}^{N} s(n)$

- DC component can removed by subtracting each sample by mean of the signal
- speech signal has zero mean

Variance: $\sigma_s^2 = \lim_{N->\infty} \frac{1}{2N+1} \sum_{n=-N}^{N} s^2(n)$

Signal energy at time n: $E(n) = \sum_{k=-N}^{N} s^2(n+k)$



Zero crossing rate differs for vowels and consonants (in particular fricatives)

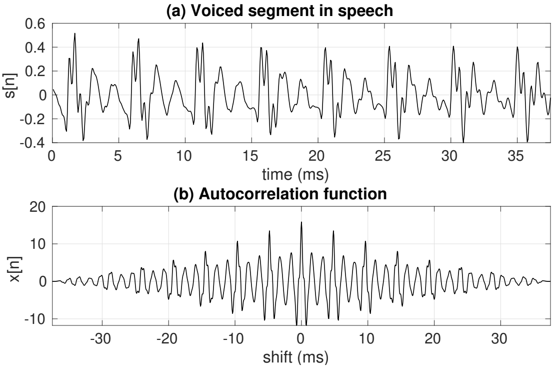
Autocorrelation (1)

Autocorrelation Function:

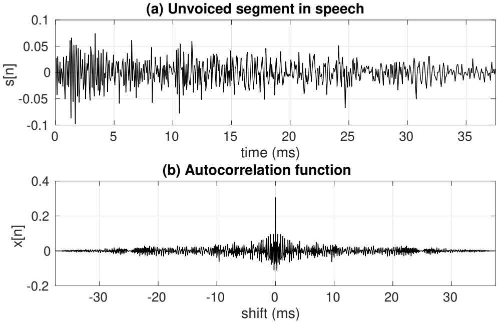
$$R_x(k) = \lim_{N->\infty} \frac{1}{2N+1} \sum_{n=-N}^{N} s(n) \cdot s(n+k)$$

Autocorrelation Coefficient:

$$RC_x(k) = R_x(k)/R_x(0)$$



Autocorrelation (2)

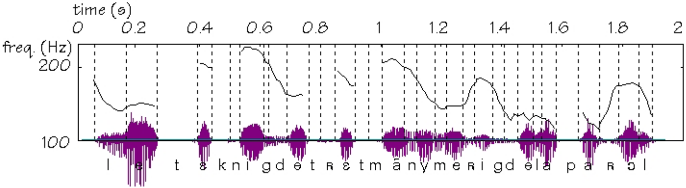


- Second order property of the signal (relation between samples)
- Autocorrelation "signal" is periodic if the signal is periodic (property used for pitch frequency estimation)
- Peak of the autocorrelation signal occurs at time 0, i.e. $R_x(0)$ (measures energy of the signal)
- Threshold the second peak $RC_x(k)$ to detect if the speech signal is voiced or unvoiced

Pitch frequency (F_0)

Fundamental (pitch) frequency: acoustic correlate of rate of periodic vibration of vocal cords.

- Between 70 and 250 Hz for men
- Between 150 and 400 Hz for women
- Between 200 and 600 Hz for children



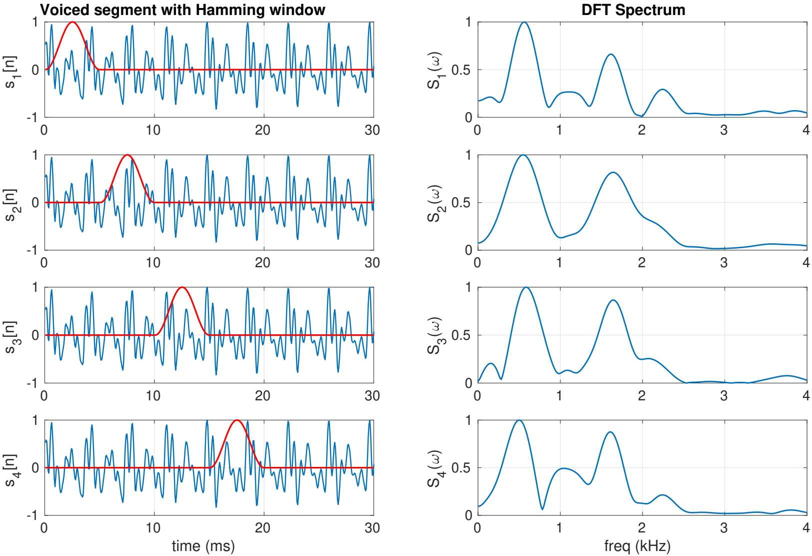
Pitch frequency evolution for sentence "Les techniques de traitement numérique de la parole"; frequency in log scale.

Outline

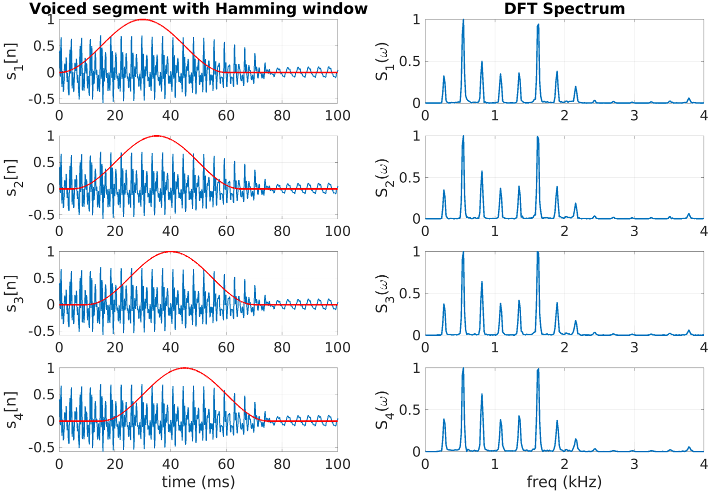
Frequency domain processing

- Human speech perception studies shows that human is able to distinguish between sounds mainly using frequency content of the signal
- Time domain information can be affected during transmission, e.g. time delay (shift), change of amplitude of signal (scaling).
- **Power spectrum**
  Fourier transform of the autocorrelation function:
  $S_x(\theta) = \sum_{k=-\infty}^{\infty} R_x(k) \cdot e^{-jk\theta}; \quad \theta = \omega . T_s$
  ω - Frequency, k - Time
- Difficulty: Speech signal is inherently "nonstationary"
  Vocal fold vibration and shape of the vocal tract keeps changing over the time, so the spectral (frequency) properties
- Solution: Short-term spectral processing with quasi-stationary assumption
  $S_x(\theta) = \sum_{k=-N}^{k=+N} R_x(k) \cdot w(k) \cdot e^{-jk\theta}$, $w(.)$ denotes a window function (typically Hamming or Hanning)

Short-term spectral processing (1)

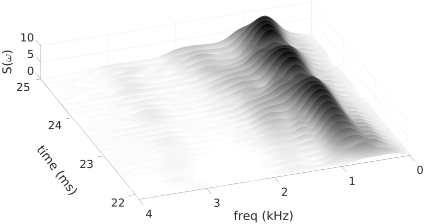


Short-term spectral processing (2)

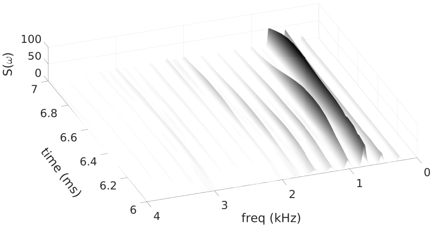


Time-frequency dilemma

similar to uncertainty principle in quantum mechanics



Wideband spectrogram (short analysis window)



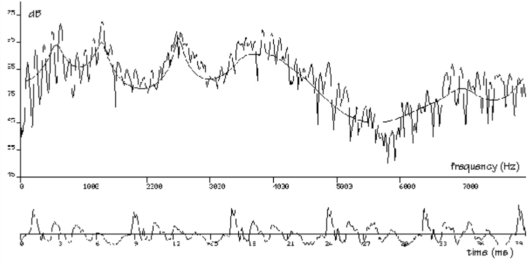
Narrowband spectrogram (long analysis window)

Analysis window hyper-parameters

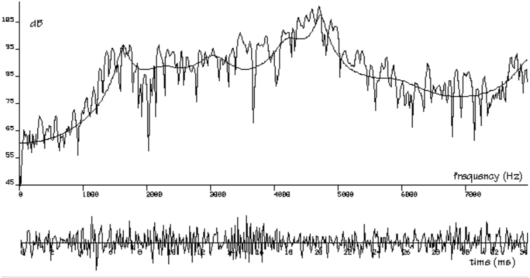
- Thumb rule for window size: should cover at least 2-3 pitch periods
  - Assuming 80 Hz (12.5 ms pitch period) or 100 Hz (10 ms pitch period) as the minimum pitch frequency
  - Window size is typically between 20-40 ms
- in the spectrum both source (vocal fold related) information and system (vocal tract related) information can be observed
- Enables speech signal to be decomposed into source component and system component (analysis) and then put them back together (synthesis)
- Window shift is typically 10 ms
- Number of frames:
  $\frac{(\text{length of signal} - \text{window size})}{\text{window shift}} + 1$
  all quantities in terms of # of samples

Power spectrum density (example)

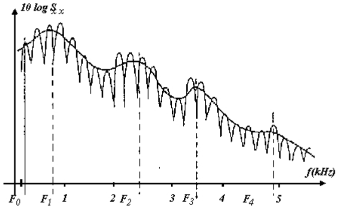
Voiced sound (“a” of “baluchon”):



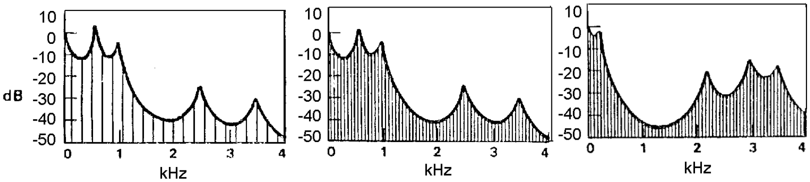
Unvoiced sound (“ch” of “baluchon”):



Spectrum of voiced sounds



F_0 - Fundamental Frequency, F_1 - First Formant, F_2 - Second formant, F_3 - Third formant, F_4 - Fourth formant

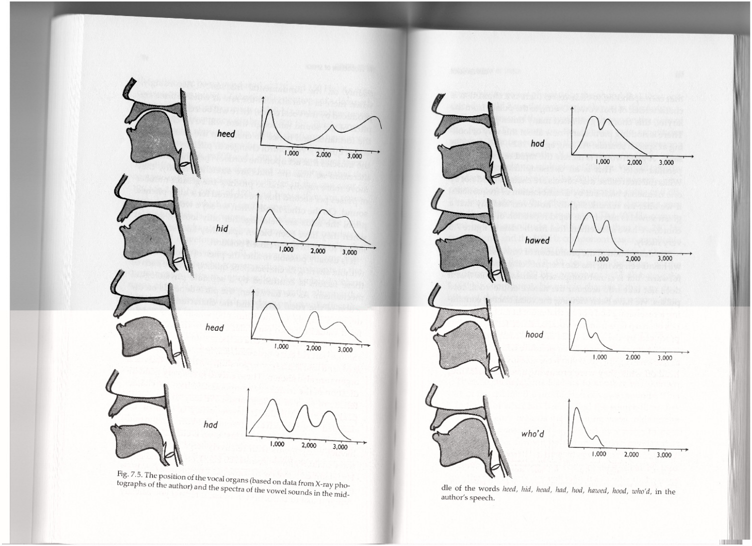


(a) RELATIVELY HIGH PITCH

(b) SAME VOWEL AS (a) WITH LOWER PITCH

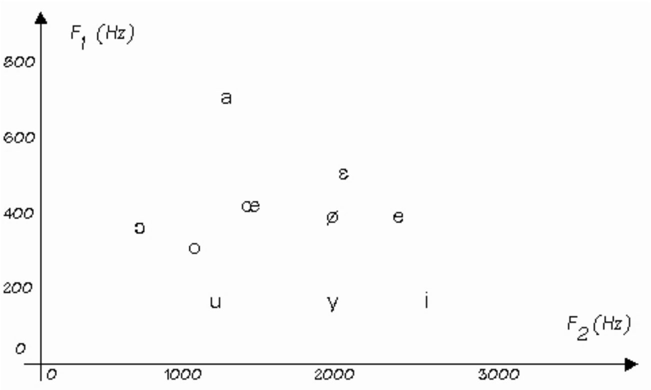
(c) DIFFERENT VOWEL WITH SAME PITCH AS (b)

Spectral envelop of different vowels

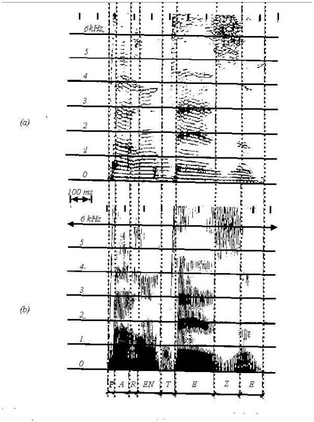


(Courtesy: Elements of Acoustic Phonetics by Peter Ladefoged)

Formants and Vowels (French)



Spectrogram



word "parenthèse": narrow band spectrogram (top) and wide band spectrogram (bottom)

- 3D: time, frequency, energy
- Narrowband spectrogram
  - long analysis window (60-100 ms)
  - time resolution is low and frequency resolution is high
  - Can observe well F_0 not Formants
- Wideband spectrogram
  - short analysis window (20-40 ms)
  - time resolution is high and frequency resolution is low
  - Can observe well Formants and F_0 (vertical strips)

Thank you for your attention!

Dr. Mathew Magimai Doss

Idiap Research Institute, Martigny, Switzerland