EE6424 Digital Audio Signal Processing
Part 2
Lecture 5:
Short-Time Analysis of Speech

Outline of lecture

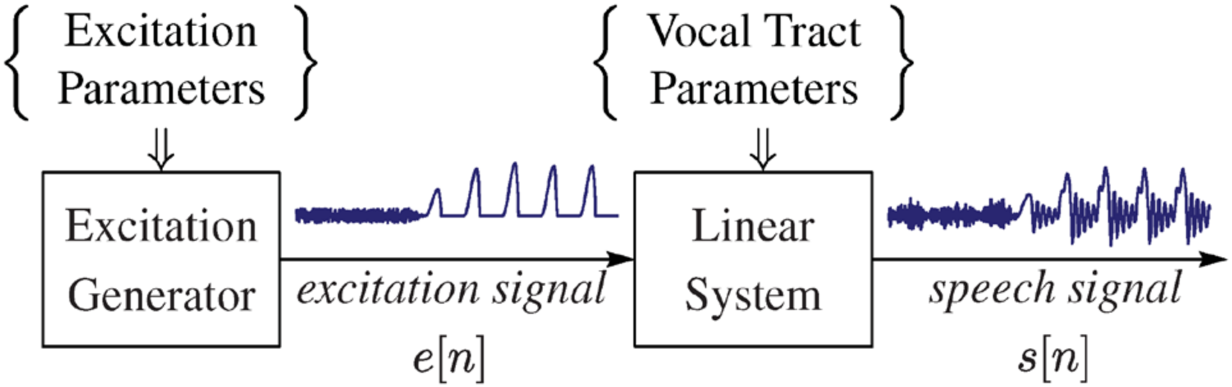
- Short-time analysis
  - General form
  - short-time energy
  - Short-time Fourier transform (STFT)
  - Speech spectrogram
- Applications
  - Speech activity detection (SAD)
  - Pitch period estimation

EE6424 Part 2: Lecture 5.1

SHORT-TIME ANALYSIS

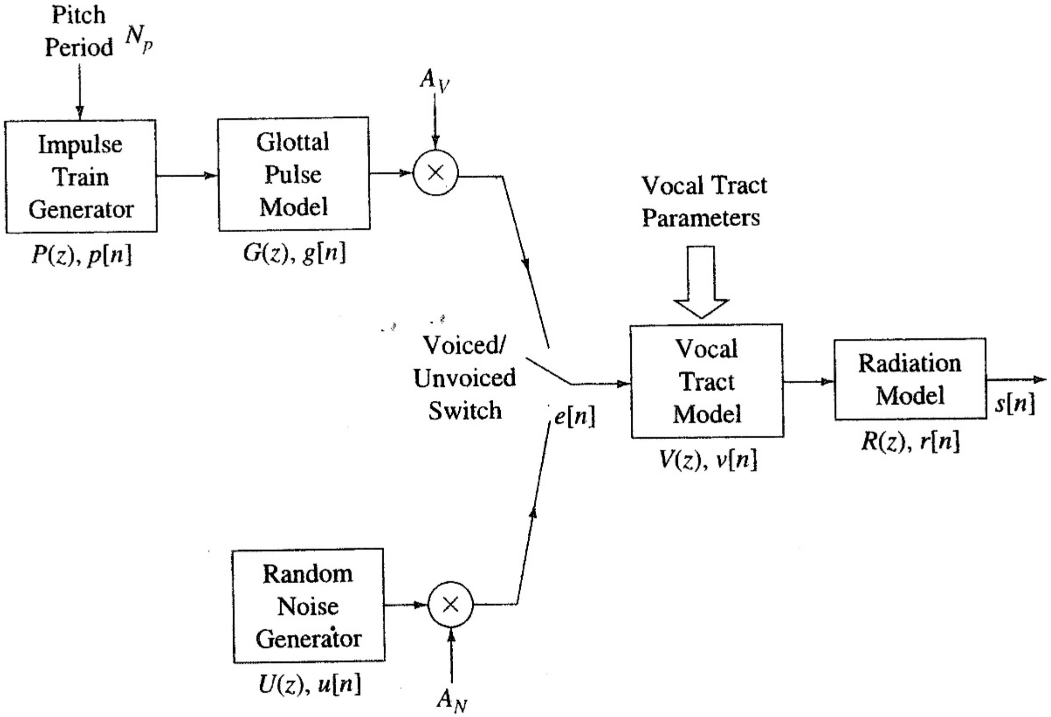
General form

In Lecture 2, a simplified source-filter model was given for speech production:



General form

A more general discrete-time model of speech production showing explicit sources for voiced and unvoiced speech sounds is given below:

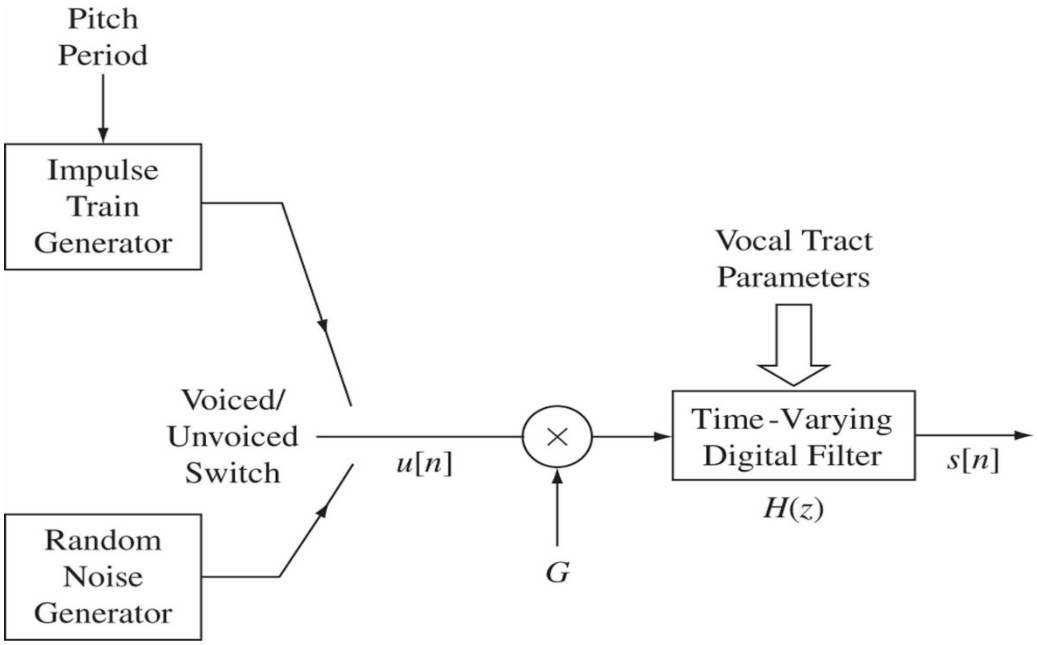


General form

where N_p is the (time-varying) pitch period (in samples) of the impulse train, for regions of voiced speech, $g(n)$ is the glottal pulse model, A_V is the (time-varying) amplitude of voiced excitation, A_N is the (time-varying) amplitude of unvoiced excitation, $\mathrm{v}(n)$ is the (time-varying) vocal tract model, and $r(n)$ is the radiation model (for the lip, fixed over time), and $u(n)$ is a random signal for unvoiced speech.

Note that in linear prediction analysis, the above general discrete-time model can be simplified as in the block diagram given next

General form



where the impulse train generator produces the impulse train for voiced speech; the random noise generator produces unvoiced excitation; the time-varying digital filter represents the vocal track resonances, the effects of radiation at the lips and, in the case of voiced speech, the effects of the glottal pulse shape. The vocal track parameters are coefficients of the time-varying digital filter.

General form

To capture the time-varying nature of speech signals, short-time analysis is necessary.

- The long-time autocorrelation function of a stationary random signal or a periodic signal (finite-power signal) is defined as

$$\phi(k) = \lim_{L \to \infty} \frac{1}{2L+1} \sum_{m=-L}^{L} x(m)x(m+k)$$

- On the other hand, the short-time autocorrelation function is defined as

$$R_{\hat{n}}(k) = \sum_{m=0}^{L-|k|-1} x(\hat{n}+m)w'(m)x(\hat{n}+m+k)w'(m+k)\,,\ \text{for } -(L-1) \le k \le L-1$$

where $w'(m)$ is a causal window of length *L*. The quantity $\hat{n}$ determines the shift of the window, and is therefore the analysis time.

General form

- Over the interval on the order of **10 to 40 ms**, it is safe to assume that the properties of the speech waveform remain relatively constant. This leads to the principle of **short-time analysis** where speech is processed in **blocks** (also called **frames**).

- A **general form of short-time analysis** is given by

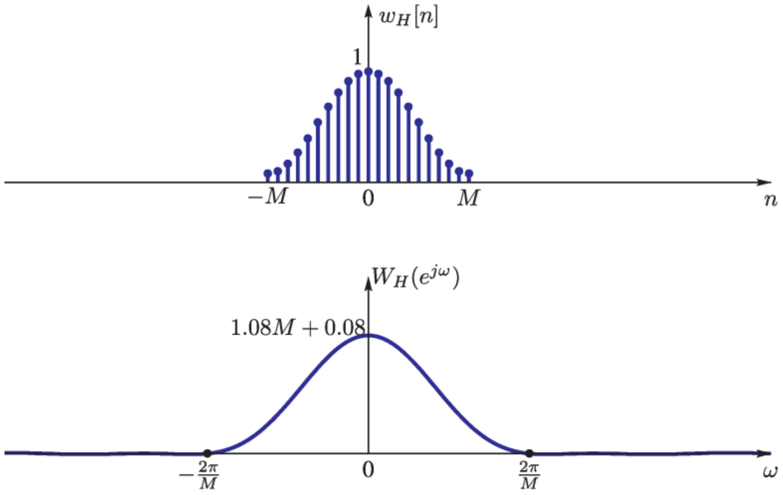
$$X_{\hat{n}} = \sum_{m=-M}^{M} T\{x(m)w(\hat{n} - m)\}$$

  - $X_{\hat{n}}$ is the short-time analysis parameters (or vector of parameters) at analysis time $\hat{n}$.
  - $T\{\cdot\}$ is the operator that defines the nature of short-time analysis function.
  - $x_{\hat{n}}(m) = x(m)w(\hat{n} - m)$ is the windowed frame of length $2M + 1$
  - $w(m)$ denotes a finite duration data window of length $2M + 1$

Window function

- An example of finite duration window is the **Hamming window**

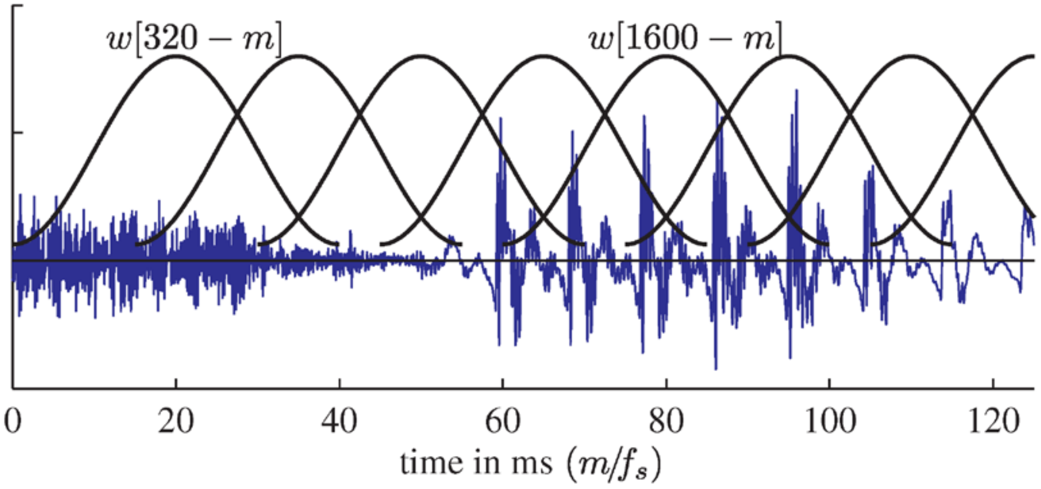
$$w_{\mathrm{H}}(m) = \begin{cases} 0.54 + 0.46\cos(\pi m/M) & -M \le m \le M \\ 0 & \text{otherwise} \end{cases}$$



- A causal window could be obtained by shifting to the right by M samples.

- A $L = (2M + 1)$ samples Hamming window has a **main lobe** with a **bandwidth** of $4\pi/M$.

- Figure below shows the operation of short-time analysis, with a data window of duration **40 ms** and shifted by **20 ms** (this corresponds to **320** samples at 16 kHz sampling rate).



Short-time energy

- The short-time energy is defined as

$$E_{\hat{n}} = \sum_{m=-M}^{M} [x(m)w(\hat{n}-m)]^2 = \sum_{m=-M}^{M} x^2(m)w^2(\hat{n}-m)$$

- Compared to the general form, the short-time analysis function $T\{\cdot\}$ here is simply a **squaring operator**.

- Let $h(n) = w^2(n)$, the short-time analysis operator could be expressed as a **convolution** or **linear filtering** operation:

$$E_{\hat{n}} = x^2(\hat{n}) * h(\hat{n})$$

- The short-time energy is the output of a **lowpass filter** whose frequency response is determined by the window function used.

- For the case of Hamming window, $2M + 1 = 401$ and a sampling rate of $F_s = 16$ kHz, the **low-pass** cut-off frequency is

$$\omega_{\mathrm{LP}} = \frac{2\pi}{M} = \frac{2\pi}{200}\,\mathrm{rad/s}$$

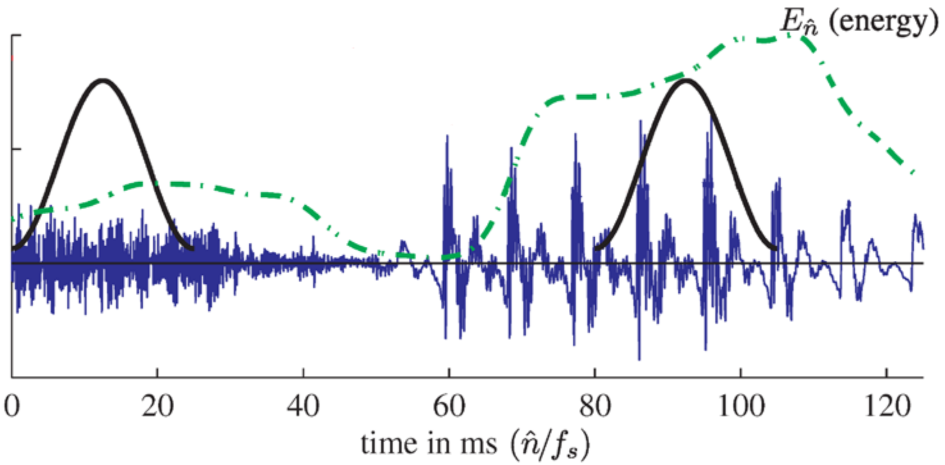
$$F_{\mathrm{LP}} = \frac{\omega_{\mathrm{LP}}}{2\pi} F_s = \frac{16000}{200} = 80\ \mathrm{Hz}$$

- Notice that the actual bandwidth of $h(n)$ is slightly increased due to the squaring operator.

- The short-time energy function $E_{\hat{n}}$ is **slowly varying** compared to the time variations of the speech signal, and therefore can be computed at a much **lower rate** than that of the original speech signal, $2F_{\mathrm{LP}} \leq F \leq F_s$.

- This reduction in the sampling rate is accomplished by moving the window position $\hat{n}$ in shift of more than one sample.

- The **short-time energy** $E_{\hat{n}}$ is an indication of the amplitude of the signal in the **interval** around analysis time $\hat{n}$.
- **Unvoiced** regions have lower short-time energy than the **voiced** region.
- There is a small shift of the short-time energy curve relative to events in the time domain waveform.



- The **time delay** of $(L-1)/2 = M$ samples is added to make the analysis window **causal**.

Short-time Fourier transform (STFT)

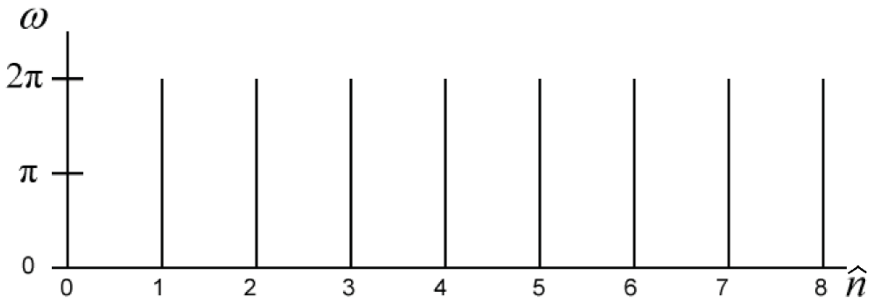
- The **short-time Fourier transform** (STFT) is defined as

$$X_{\hat{n}}\left(e^{j\omega}\right) = \sum_{m=-M}^{M} x(m)w(\hat{n}-m)e^{-j\omega m}$$

- Compared to the general form of short-time analysis, the STFT is obtained by letting the **analysis function** $T\{\cdot\}$ be the **discrete-time Fourier transform** (DTFT):

$$X\left(e^{j\omega}\right) = \sum_{n=-\infty}^{\infty} x(n)e^{-j\omega n}$$

- The STFT $X_{\hat{n}}(e^{j\omega})$ is a two-dimensional function. It is a function of two variables:
  - $\hat{n}$ the **discrete time index** denoting the window position
  - ω representing the **analysis frequency**
- STFT can be seen as a sequence of DTFTs of windowed signal segments (or frames) at **analysis time** $\hat{n}$.



- In practice, the **DTFT** is replaced with Discrete Fourier transform (**DFT**), in which the continuous analysis frequency ω is replaced with a **discrete** set of N **frequencies** as in

$$X_{rR}(k) = \sum_{m=rR-L+1}^{rR} x(m)w(rR-m)e^{-j(2\pi k/N)m} \quad k = 0,1,\dots,N-1$$

  - N is the number of uniformly spaced frequencies across the interval $0 \le \omega \le 2\pi.$
  - $L = 2M + 1$ is the window length (in samples)
  - R is the temporal sampling period (shift between successive frames)
  - A windowed segment $x(m)w(rR-m)$ is non-zero over $(rR-L+1) \le m \le rR$, *r* = 1, 2, …

- The parameters R and N (shift and DFT length) are determined by the time width and frequency bandwidth of the window function $w(m)$ used.
  - $R \leq M/C$, where C is a constant that is dependent on the window frequency bandwidth. $C = 2$ for Hamming window, $C = 1$ for rectangular window.
  - $N \geq L$
- Considering that each frequency ω is an output of a **lowpass filter** with bandwidth determined by the window function used, the first constraint is related to sampling the STFT in time at a rate of **twice the bandwidth** in order to eliminate aliasing.

$$\omega_{\mathrm{LP}} = \frac{2\pi}{M}$$

$$F_{\mathrm{LP}} = \frac{\omega_{\mathrm{LP}}}{2\pi} F_{\mathrm{s}} = \frac{2\pi}{M} \times \frac{1}{2\pi} \times F_{\mathrm{s}} = \frac{F_{\mathrm{s}}}{M} \mathrm{Hz}$$

$$F \geq 2F_{\mathrm{LP}} = 2 \times \frac{F_{\mathrm{s}}}{M} \mathrm{Hz}$$

$$\frac{1}{R'} \geq 2 \times \frac{F_{\mathrm{s}}}{M} \mathrm{Hz}$$

$$R' \leq \frac{M}{2F_{\mathrm{s}}} \mathrm{second}$$

$$R = \frac{R'}{T_{\mathrm{s}}} \leq \frac{M}{2} \mathrm{sample}$$

- For applications in which some **modification** is to be performed on the STFT (e.g., linear or non-liner filtering) and then **re-synthesizing** the modified signal, it is essential that **no aliasing** should occur in sampling the STFT.

- **Under-sampled** representations are useful for short-time analysis, for examples, pitch and formant analysis, speech spectrogram, spectral estimation etc.

Speech spectrogram

- **Spectrograms** are **magnitude plots** of the STFT

$$S(t_r, f_k) = 20\log_{10}|X_{rR}(k)|$$

- The plot axes are labeled in terms of analog time and frequency through the relations (where T is the sampling period, N indicates the number of points in DFT)

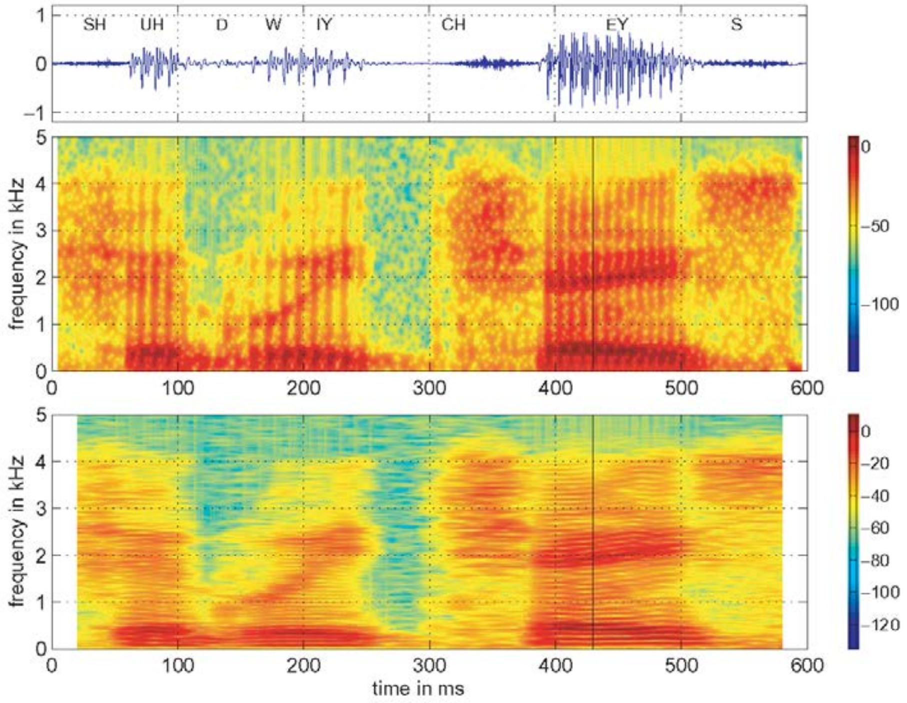
$$t_r = rRT$$

$$f_k = k/(NT)$$

- The shift R between successive frames is usually smaller than the window length L while N is much larger than the window length L to increase the resolution of the spectrogram (to make a more smooth looking plot).

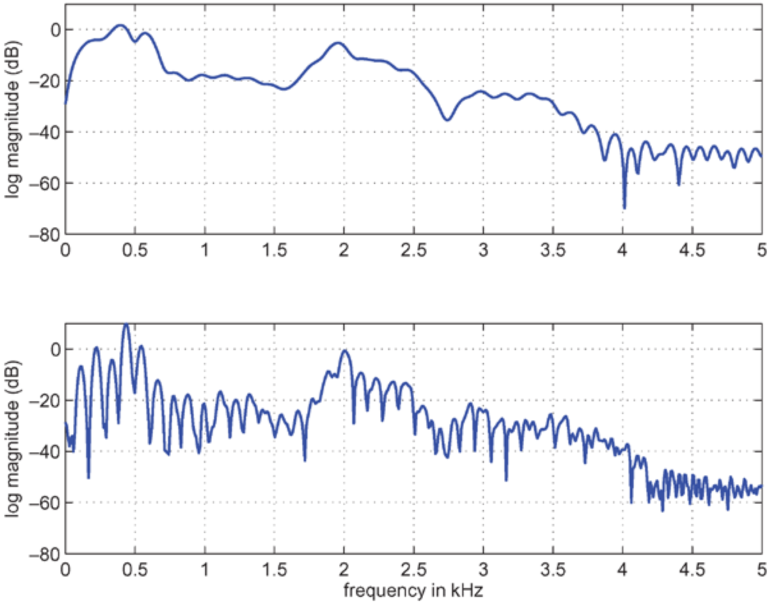
- The window length L has a major effect on the STFT and the spectrogram:
  - **Short** analysis window (on the order of a pitch period around 10 ms) results in good time resolution but poor frequency resolution.
    - Each individual pitch period is resolved in the time dimension as **vertically oriented striations**.
  - **Long** analysis windows (on the order of several pitch periods around 40 ms) results in good frequency resolution but poor time resolution.
    - The fine structure appears as **horizontally oriented striations** since the fundamental frequency and its harmonics are all resolved.
    - Several periods are included in the window. The spectrogram is not sensitive to rapid time variation.

- If the analysis window is short the spectrogram is called **wideband** spectrogram (i.e., poor frequency resolution). Conversely, when the window length is long, the spectrogram is call **narrowband** spectrogram.



- The upper plot is a **wideband** spectrogram computed with a window length of 10 ms.
- The lower plot is a **narrowband** spectrogram computed with a window length of 40 ms, while

- The formants could be clearly observed on both wideband (upper plot) and narrowband spectrogram (lower plot).



EE6424 Part 2: Lecture 5.2

APPLICATIONS

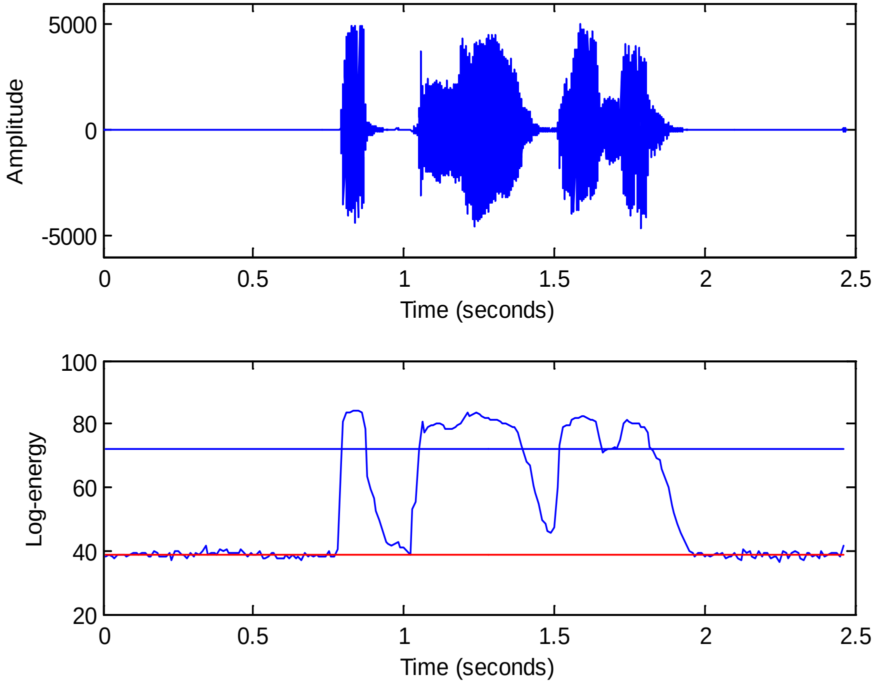
Speech activity detection

- Speech Activity Detection (SAD) or Voice Activity Detection (VAD) is the problem of determining the **presence** of speech in a given signal and locating the **region** of speech in a background noise (non-speech event).
- **Short-time log-energy** can be used as the basis for VAD, with the assumption that speech regions have higher energy than the background noise (SNR > 0).
- The log-energy $\log E_{\hat{n}}$ is computed for the entire recording:

$$\log E_{\hat{n}} = 10 \log_{10}\left(\sum_{m=-M}^{M} [x(m)w(\hat{n}-m)]^2 + \varepsilon\right)$$

- A rectangular (or hamming window) is commonly used for SAD and ε is a small constant to avoid log of zero.

- At 8 kHz sampling frequency, typical value for window length is $L = 2M + 1 = 161$ samples (20 ms) with 80 samples (10 ms) shift between adjacent frames.



- A simple VAD algorithm to classify individual frames of the signal could be devised by fitting a **bi-Gaussian model** onto the **log-energy contour**.

- Let $e = \log E_{\hat{n}}$ be the frame log-energy, a **bi-Gaussian** model is given by

$$p(e) = \sum_{i=1}^{2} w_i \times \mathcal{N}\left(e \mid \mu_i, \sigma_i^2\right) = \sum_{i=1}^{2} w_i \times \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{(e-\mu_i)^2}{2\sigma_i^2}\right)$$

- A **bi**-Gaussian model consists of **two** Gaussian components, where
  - μ_1 and μ_2 are the means
  - σ_1^2 and σ_2^2 are the variances
  - w_1 and w_2 are the weights (the percentage of frames in each of the two classes)
- Gaussian component with **lower mean value** corresponds to a **non-speech** event, while Gaussian component with **higher mean value** corresponds to a **speech** event.

The EM algorithm

- Training the bi-Gaussian model using **Expectation Maximization** (EM) algorithm

- **Initialization:**

Set $w_1 = w_2 = 0.5$ for equal weight. Take the top P percent (usually 0.1 to 0.2) of frames with highest and lowest log-energy to initialize the mean μ_i, variance σ_i^2:

$$\boxed{\begin{aligned} \mu_i &= \frac{1}{N_i} \times \sum_{e_{\hat{n}} \in \Omega_i} e_{\hat{n}} \\ \sigma_i^2 &= \frac{1}{N_i} \times \sum_{e_{\hat{n}} \in \Omega_i} \left(e_{\hat{n}} - \mu_i\right)^2 \end{aligned}}$$

- **E-step:**

Compute the membership of each frame to the two Gaussians based on its log-energy $e_{\hat{n}}$

$$\boxed{\lambda_i\left(e_{\hat{n}}\right)=\frac{\mathcal{N}\left(e_{\hat{n}} \mid \mu_i,\sigma_i^2\right)\cdot w_i}{\sum_{i=1}^{2}\mathcal{N}\left(e_{\hat{n}} \mid \mu_i,\sigma_i^2\right)\cdot w_i}}$$

- **M-step:**

Update the mean μ_i, variance σ_i^2, and weights w_i based on the membership information $\lambda_i(e_{\hat{n}})$ of each frame (N is the number of frames).

$$\mu_i=\frac{1}{n_i}\times\sum_{\forall\hat{n}}\lambda_i\left(e_{\hat{n}}\right)\cdot e_{\hat{n}} \qquad w_i=\frac{1}{N}\underbrace{\sum_{\forall\hat{n}}\lambda_i\left(e_{\hat{n}}\right)}_{n_i}$$

$$\sigma_i^2=\frac{1}{n_i}\times\sum_{\forall\hat{n}}\lambda_i\left(e_{\hat{n}}\right)\cdot\left(e_{\hat{n}}-\mu_i\right)^2$$

Likelihood computation

- To classify individual frames given the frame log-energy $e_{\hat{n}}$, we compute the **log likelihood** for both classes (speech and non-speech) and assign the frames to the class with the higher likelihood.

$$l_i\left(e_{\hat{n}}\right) = \ln \mathcal{N}\left(e_{\hat{n}} \mid \mu_i, \sigma_i^2\right) = \ln \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{\left(e_{\hat{n}} - \mu_i\right)^2}{2\sigma_i^2}\right)$$

$$= -\ln\left(\sqrt{2\pi}\sigma_i\right) + \ln\left(\exp\left(-\frac{\left(e_{\hat{n}} - \mu_i\right)^2}{2\sigma_i^2}\right)\right)$$

$$= -\ln\left(\sqrt{2\pi}\sigma_i\right) - \frac{\left(e_{\hat{n}} - \mu_i\right)^2}{2\sigma_i^2}$$

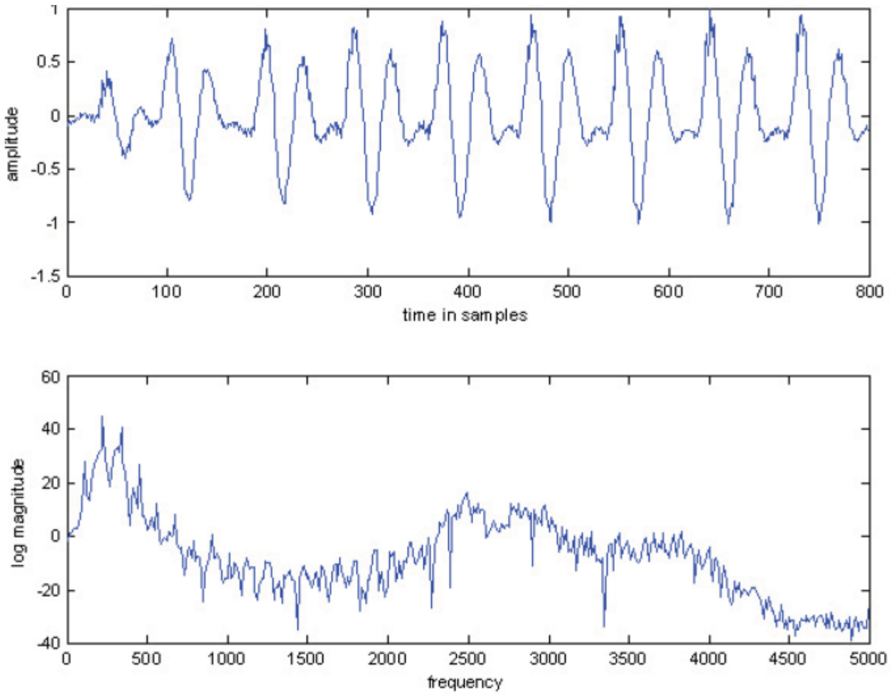
- Let class $i = 1$ to represent a speech event while class $i = 2$ to represent a non-speech event. A frame with log-energy $e_{\hat{n}}$ is taken a speech frame if and only if $l_1(e_{\hat{n}}) > l_2(e_{\hat{n}})$.

Pitch detection

- The goal of a pitch detector is to estimate the **pitch period** (or equivalently, the pitch **fundamental frequency**) of the speech waveform during the **voiced sections** of speech.

- Reliable and accurate estimation is difficult due to the following **challenges**:
  - The glottal source is not truly periodic, but **quasi-periodic** with a period that changes (usually slowly) over time, such as Jitter and shimmer of the glottal pulses.
  - The **vocal tract changes** over time. Hence, the frequency shaping and the speech signal itself (where the estimation is based on) changes from period to period.
  - Pitch period is **ill-defined** at the beginning and ending of voicing, during which the glottal excitation is **building up** and **breaking down**.

- Four general approaches for pitch detection:
  - Time-domain measurements
  - Frequency-domain measurements
  - Cepstral-domain measurements
  - LPC-based measurements



Pitch detection in the spectral domain

- The basic principle for pitch detection in the spectral domain:

  In a ***narrowband*** *spectrogram, the excitation for voiced speech is manifested in sharp peaks that occur at the integer multiples of the fundamental frequency.*

- Let $X_{\hat{n}}(e^{j\omega})$ be the STFT at time $\hat{n}$, we define the **harmonic product spectrum** as follows

$$P_{\hat{n}}(e^{j\omega}) = \prod_{r=1}^{K} \left|X_{\hat{n}}(e^{j\omega r})\right|^2$$

- $r \geq 1$ is a positive integer indicating the **compression scale**. The frequency scale is compressed by a r factor.

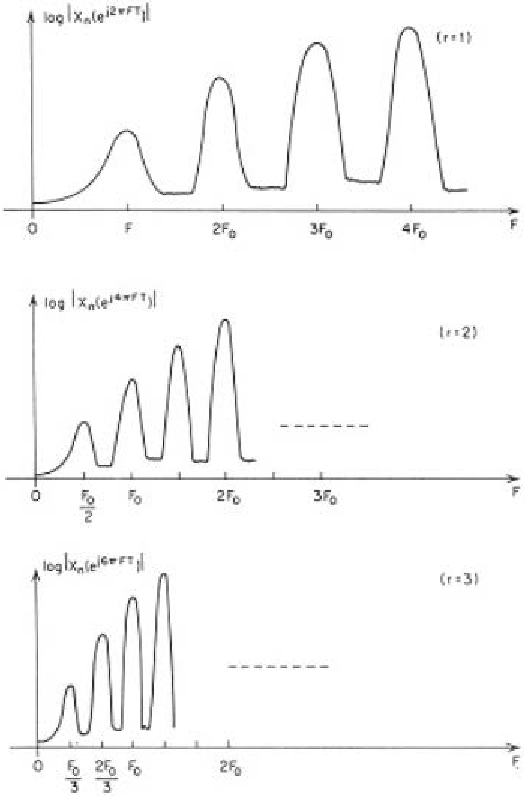
- Taking the logarithm of the harmonic product spectrum, we arrive at the **log harmonic product spectrum**:

$$\log P_{\hat{n}}\left(e^{j\omega}\right) = \log \prod_{r=1}^{K} \left|X_{\hat{n}}\left(e^{j\omega r}\right)\right|^{2} = 2 \sum_{r=1}^{K} \log\left|X_{\hat{n}}\left(e^{j\omega r}\right)\right|$$

- In the log domain, $\log P_{\hat{n}}\left(e^{j\omega}\right)$ is seen as the sum of K frequency compressed replicas of $\log\left|X_{\hat{n}}\left(e^{j\omega}\right)\right|$.

- For **voiced speech**, compressing the frequency scale by integer factors causes harmonics of the fundamental frequency to coincide at the fundamental frequency.
  - For the compression by a factor of r, the rth harmonic $r \times F_0$ will coincide with the fundamental frequency at F_0.

- The process of compressing the frequency scale is depicted below, showing the **log magnitude spectrum** without compression (i.e., $r = 1$) , and with a compression factor of 2 and 3.



- A compression factor of $r = 2$ causes the second harmonic to coincide with the fundamental.

- A compression factor of $r = 3$ causes the third harmonic to coincide with the fundamental.

- **Adding** up the **log magnitude spectra** with different scaling factors **reinforces** the **fundamental frequency** while random noises are **suppressed** in the resulting **log harmonic product spectrum** $\log P_n\left(e^{j\omega}\right)$.

- The contribution of noises to the spectrum $X_n\left(e^{j\omega}\right)$ has no **coherence structure** when viewed as a function of frequency. The noise components tend to add incoherently.

- For the same reason, **unvoiced speech** will not exhibit a peak in the log magnitude product spectrum.

- For the case of **missing fundamental** (as in telephone speech), the compressed harmonics would still add coherently at the fundamental frequency without the existence of the missing fundamental.

Summary

- A general discrete-time model of speech production showing explicit sources for voiced and unvoiced speech sounds

- Long-time autocorrelation function vs. short-time autocorrelation function

- The properties of the speech waveform remain relatively constant over the **time scale of phoneme**.

- This leads to the principle of **short-time analysis** where speech is processed in **blocks** (or **frames**), usually **10 to 40** ms.
  - **Short-time energy**, where the short-time analysis function $T\{\cdot\}$ is a square operator.
  - **Short-time Fourier transform** (STFT), where the short-time analysis function $T\{\cdot\}$ is the DTFT

- In practice, the **DTFT** is replaced with Discrete Fourier transform (**DFT**) in a STFT.

- **Spectrograms** are magnitude plots of the STFT.
  - A **wideband** spectrogram is produced with a short window (10 ms)
  - A **narrowband** spectrogram is produced with a long window (40 ms)

- The short-time analysis function is **slowly varying** compared to the time variations of the speech signal and and therefore can be computed at a much lower rate than that of the original speech signal, $2F_{\mathrm{LP}} \leq F \leq F_s$.

- The **lowpass filtering** characteristic is determined by the window function used.