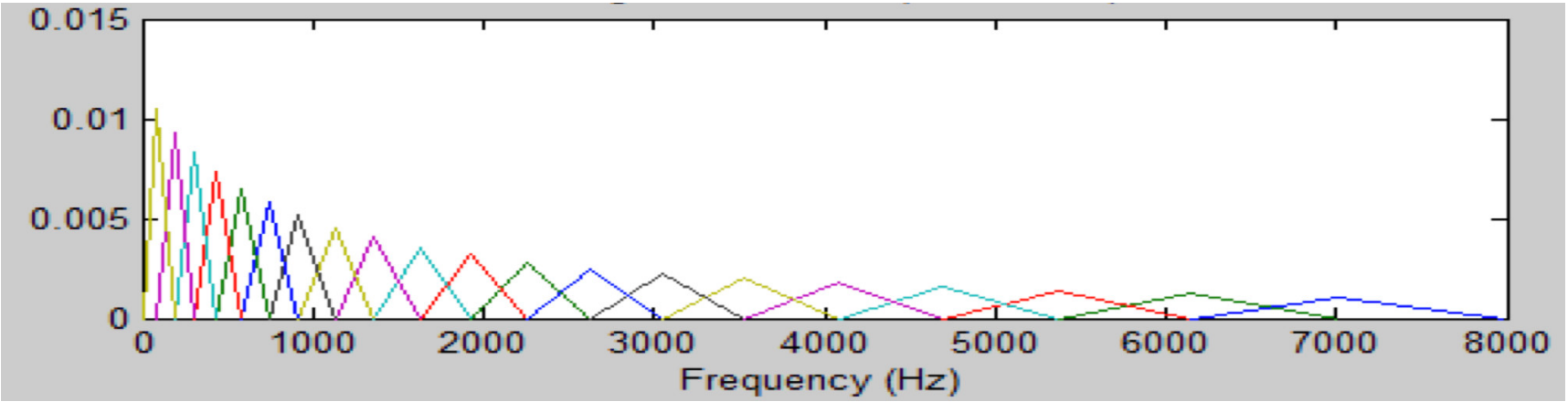
- - - Speech Processing - - -

- **Mel Frequency Cepstral Coefficients (MFCCs)**: it combines cepstrum with a specific scale called *Mel scale*, being used as coefficients which embed speech-related information for speech recognition, speaker identification, emotion recognition, and so on.
- The term *Mel* comes from *Melody*. The fundamental idea behind it is to mimic the way humans perceive sounds. The transformation from the linear scale (f) to the Mel Scale (m) is: $m = 1127 \cdot \ln\left(1 + \frac{f}{700}\right)$.
- MFCCs are determined as follows:
  - STEP 1: pre-processing, including pre-emphasis, mean removal, normalization, and Hamming windowing.
  - STEP 2: calculate the module of the DFT of the frame being considered, then apply the triangular filterbank shown ahead to filter the spectrum in the frequency domain based on multiplications **OR** filter the time-domain signal from the previous step by convolving it with the filters' impulse responses, according to the triangular filterbank shown ahead. The former option is more frequently used since the filters

adopted in the latter need high orders.



- Each triangular filter above has its area under the curve equals to the unit, in order to avoid unbalancing. Their frequencies responses are as follows:
  - 0, for $t < a_m$
  - $\frac{2(t-a_m)}{(b_m-a_m)(\frac{a_m+b_m}{2}-a_m)}$, for $a_m < t \leqslant \frac{a_m+b_m}{2}$
  - $\frac{2(b_m-t)}{(b_m-a_m)(b_m-\frac{a_m+b_m}{2})}$, for $\frac{a_m+b_m}{2} < t \leqslant b_m$
  - 0, for $t > b_m$

where a_m and b_m, in Hertz, are the initial and final frequencies of the m-th interval, as follows, approximately:

- - - Speech Processing - - -

m	a_m	b_m	m	a_m	b_m	m	a_m	b_m	m	a_m	b_m
1	0	200	2	100	300	3	200	400	4	300	500
5	400	600	6	500	700	7	600	800	8	700	900
9	800	1000	10	900	1148	11	1000	1318	12	1148	1514
13	1318	1737	14	1514	1995	15	1737	2291	16	1995	2630
17	2291	3020	18	2630	3467	19	3020	4000	20	3467	max.

- Proceeding, we have:
  - STEP 3: once the previous signal is filtered, separately, with each one of the 20 filters, the log energy of each one is calculated, resulting in 20 scalars, i.e., s_n, for $0 \leqslant n \leqslant 19$.
  - STEP 4: lastly, to attenuate the correlation among the coefficients, the Discrete Cosine Transform (DCT) is applied, as follows, producing the MFCCs:

$$\text{MFCC}_i = \sum_{n=0}^{M-1} s_n \cdot cos\left(\frac{i\pi(n+0.5)}{M}\right) \quad \text{, for} \quad i = 0, 1, 2, ..., M-1 \quad ,$$

    where M is the number of coefficients. Traditionally, $8 \leqslant M \leqslant 14$.

- Usually, velocity and acceleration coefficients, i.e., Δ and $\Delta\Delta$, are also computed, resulting in $M + (M-1) + (M-2)$ coefficients.

- - - Speech Processing - - -

- Thus, if, for instance, $\text{MFCC}[i] = \{f_0, f_1, f_2, f_3, ...\}$, then $\Delta = \{|f_1 - f_0|, |f_2 - f_1|, |f_3 - f_2|, ...\}$ and $\Delta\Delta = \{|f_2 - f_1| - |f_1 - f_0|, |f_3 - f_2| - |f_2 - f_1|, ...\}$.

- **Harmonic-to-noise Ratio (HNR)**: this feature corresponds to the ratio between the second harmonic amplitude and the difference between the first harmonic (fundamental frequency) amplitude and the second harmonic one, i.e., $\text{HNR} = \frac{r_x[t]}{r_x[0] - r_x[t]}$, where $r_x[0]$ is the autocorrelation central coefficient of the input signal $x[n]$, and $r_x[t]$ is the closest autocorrelation coefficient for which the amplitude corresponds to a peak.

- **Example**: letting $x[n] = \{1, 3, 1, 3, 1, 2, 1\}$, then its autocorrelation, i.e., $r_x[n]$ is $\{1, 5, 8, 11, 18, 16, 26, 16, 18, 11, 8, 5, 1\}$ for which the central coefficient is $r_x[0] = 26$. Its closest peak is $r_x[2] = 18$. Consequently, $\text{HNR} = \frac{r_x[t]}{r_x[0] - r_x[t]} = \frac{r_x[2]}{r_x[0] - r_x[2]} = \frac{18}{26 - 18} = 2.25$.

- - - Speech Processing - - -

- **Discrete Hilbert Transform**: it is a simple transformation which allows for the envelope to be extracted from a certain signal. Basically, from an input signal $s[n]$, that transformation produces $\tilde{s}[n]$ which corresponds to the same signal $s[n]$ with all its frequency components shifted 90 degrees. Particularly, the envelope of $s[n]$ is computed as

$$\sqrt{\left(s[n]\right)^2 + \left(\tilde{s}[n]\right)^2} \quad .$$

To obtain $\tilde{s}[n]$ from $s[n]$, we simply convolve $s[n]$ with $v[n]$, where:

$$v[n] = \begin{cases} 0 & \text{for}(n - \frac{M}{2}) = 0 \\ \frac{2sin^2(\frac{\pi}{2}(n-\frac{M}{2}))}{\pi(n-\frac{M}{2})} & \text{otherwise} \end{cases} \quad .$$

After filtering, in order to calculate the envelope, the exceeding samples of the filtered signal might be discarded.

- - - Speech Processing - - -

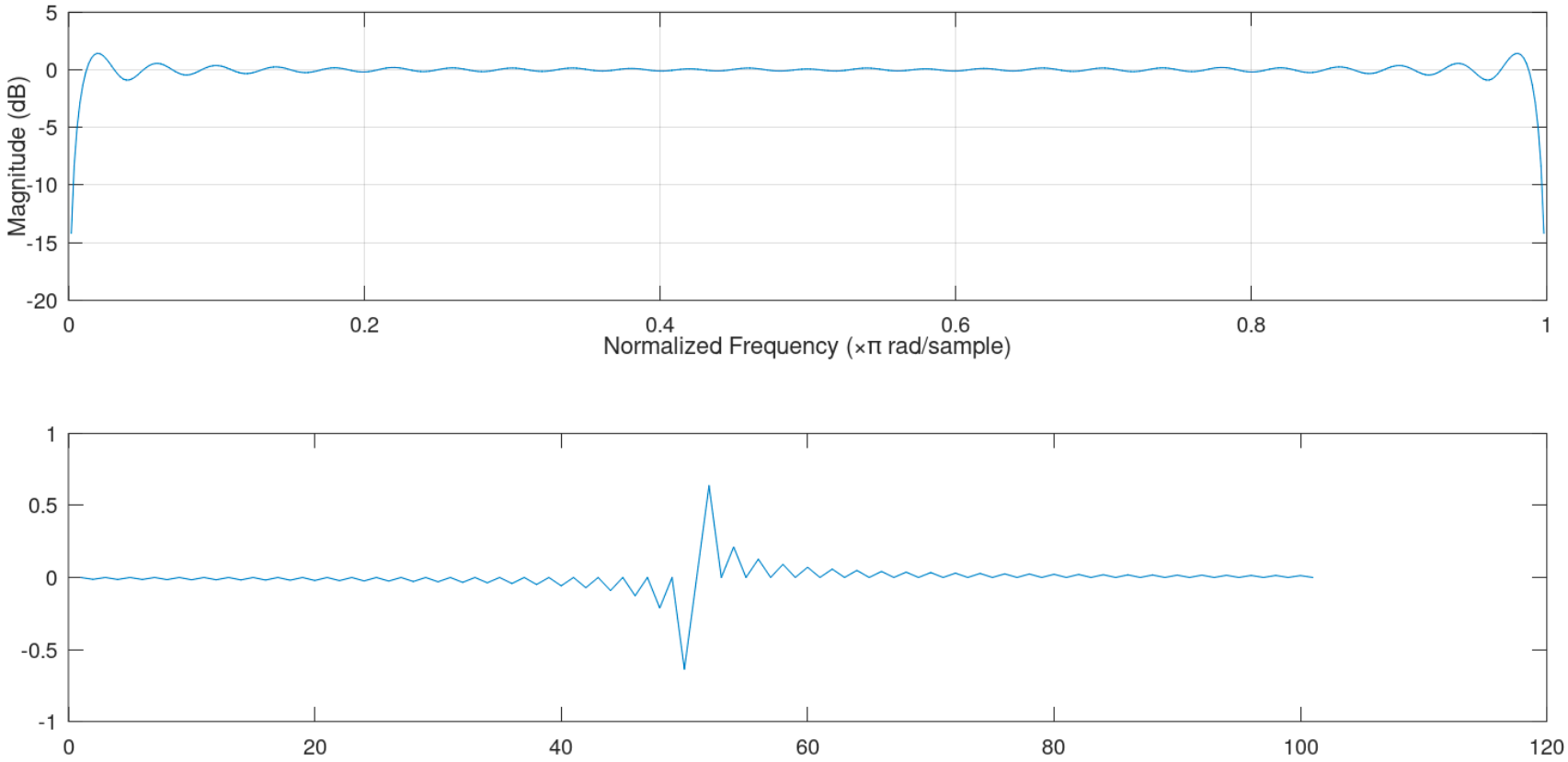


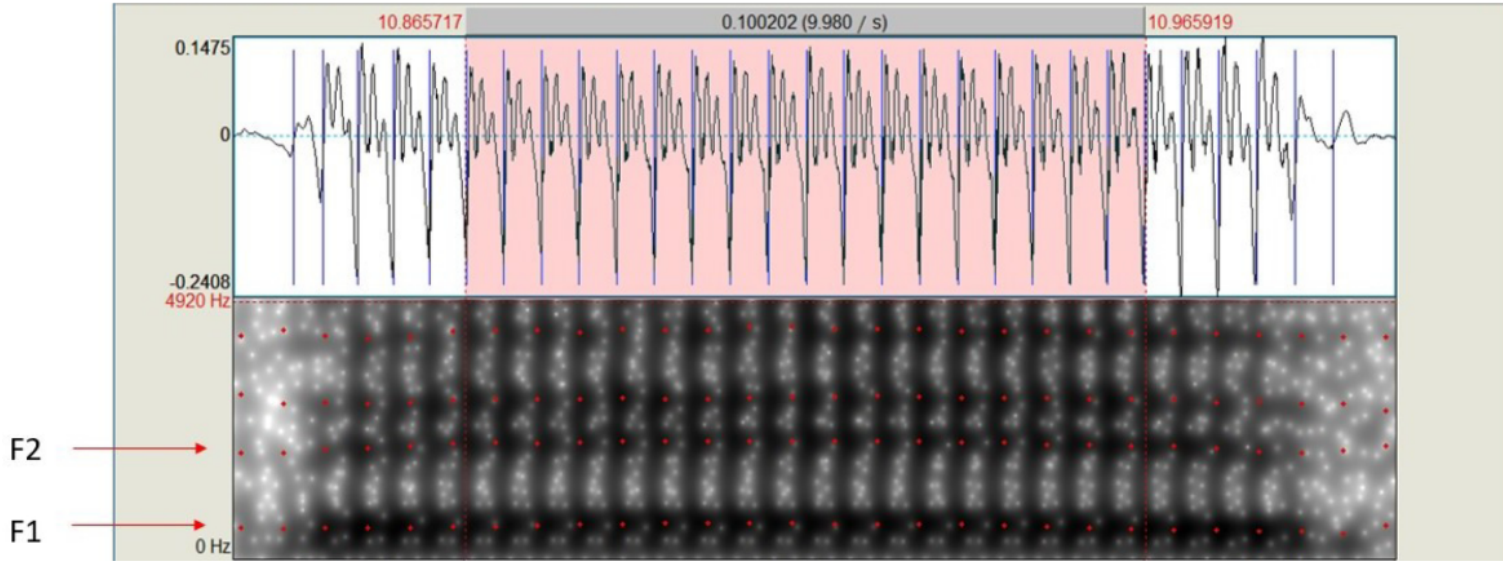
Figure: [above]: frequency response of the Hilbert Transform filter (observe that it is an ALL-PASS filter, i.e., it preserves all the frequencies, just acting to shift their phases); [below]: corresponding filter impulse response.

- - - Speech Processing - - -

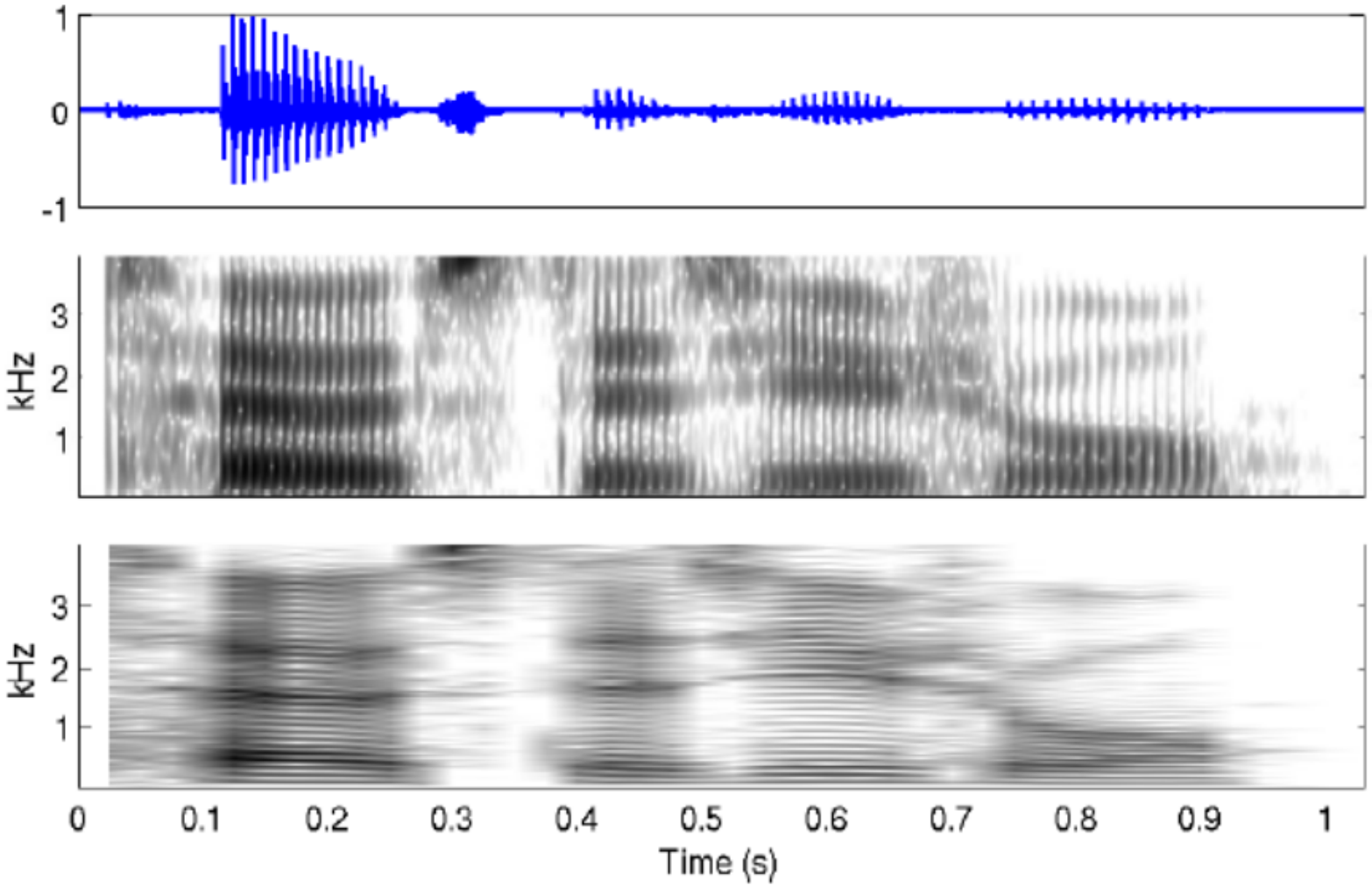
- **Teager Energy Operator (TEO)**: it consists of a transformation which converts an M-sample long speech signal $s[n]$ into an M-sample long energy signal, which is a function of both amplitude and frequency. TEO is inspired on the Newton's second law of mechanics, where a spring-mass system is described. According to TEO, the signal energy of an input discrete-time signal $x[n]$ is approximately given by $(\mathbb{E}(x[n]))_i \approx (x_i)^2 - (x_{i-1}x_{i+1})$, for all i.
- Example: find $(\mathbb{E}(x[n]))$ for $x[n] = \{1, 2, 3, 4, 5\}$.
- A detailed explanation on TEO, including an enhanced energy operator, is available in the paper "GUIDO, R.C. Enhancing Teager Energy Operator Based on a Novel and Appealing Concept: signal mass. *Journal of the Franklin Institute*, v.356, n.4, pp. 2346-2352, (2019)". Notably, as the signal amplitude $(A(x[n]))$ increases, $\mathbb{E}(x[n])$ also increases. The same holds true for the signal frequency $(\Omega(x[n]))$. Particularly, as shown in that paper, $\mathbb{E}(x[n]) \propto A^2\Omega^2$.

- - - Speech Processing - - -

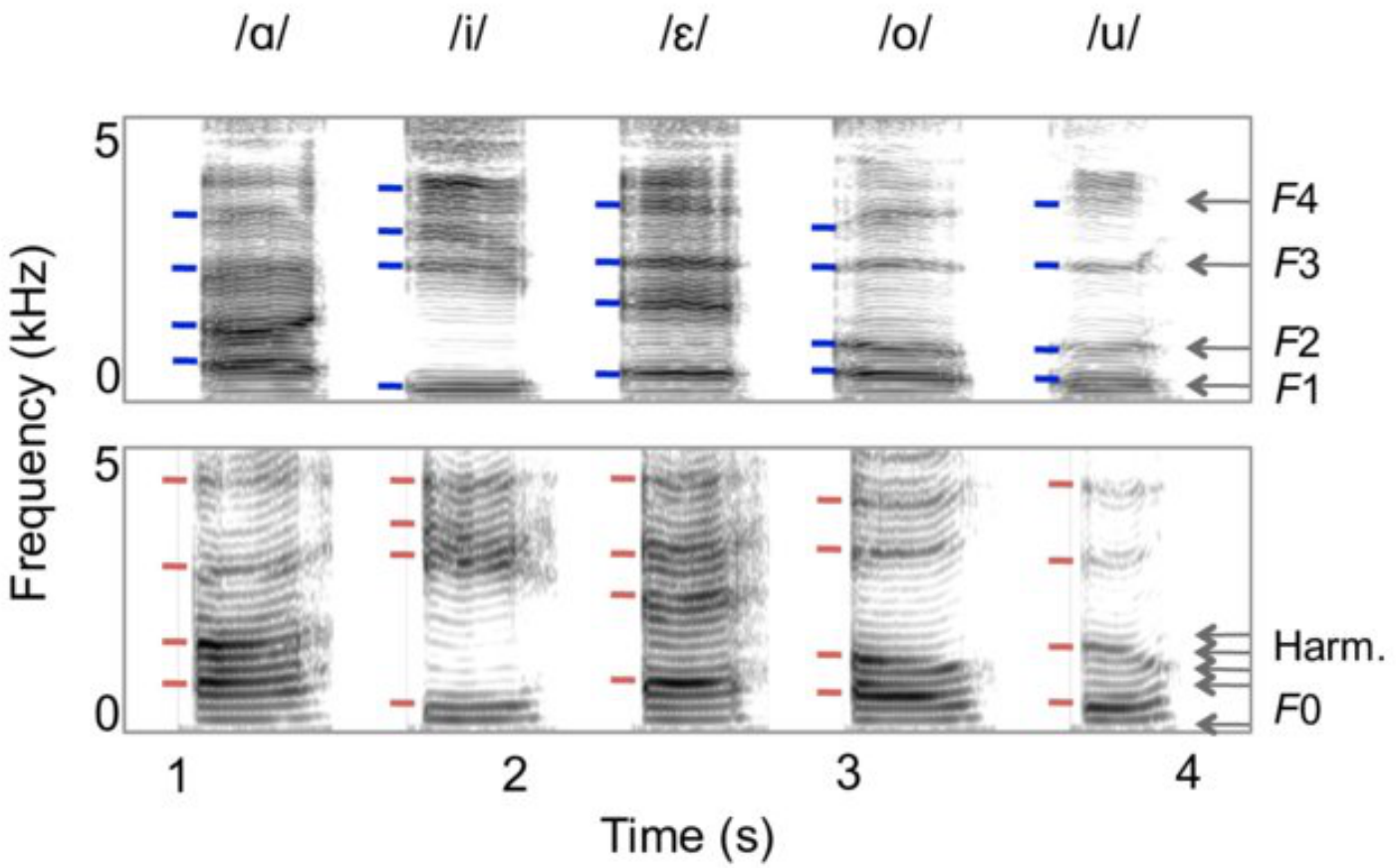
- **“Visual Features” - Spectrograms**: plots which show, by means of color intensities or tonalities, the energy of a signal as a function of time and frequency. Usually, horizontal and vertical axes represent time and frequency, respectively. “Hot” or intense colors indicate considerable amounts of energy. The plots can be obtained based on ordinary Fourier analysis. Examples follow.



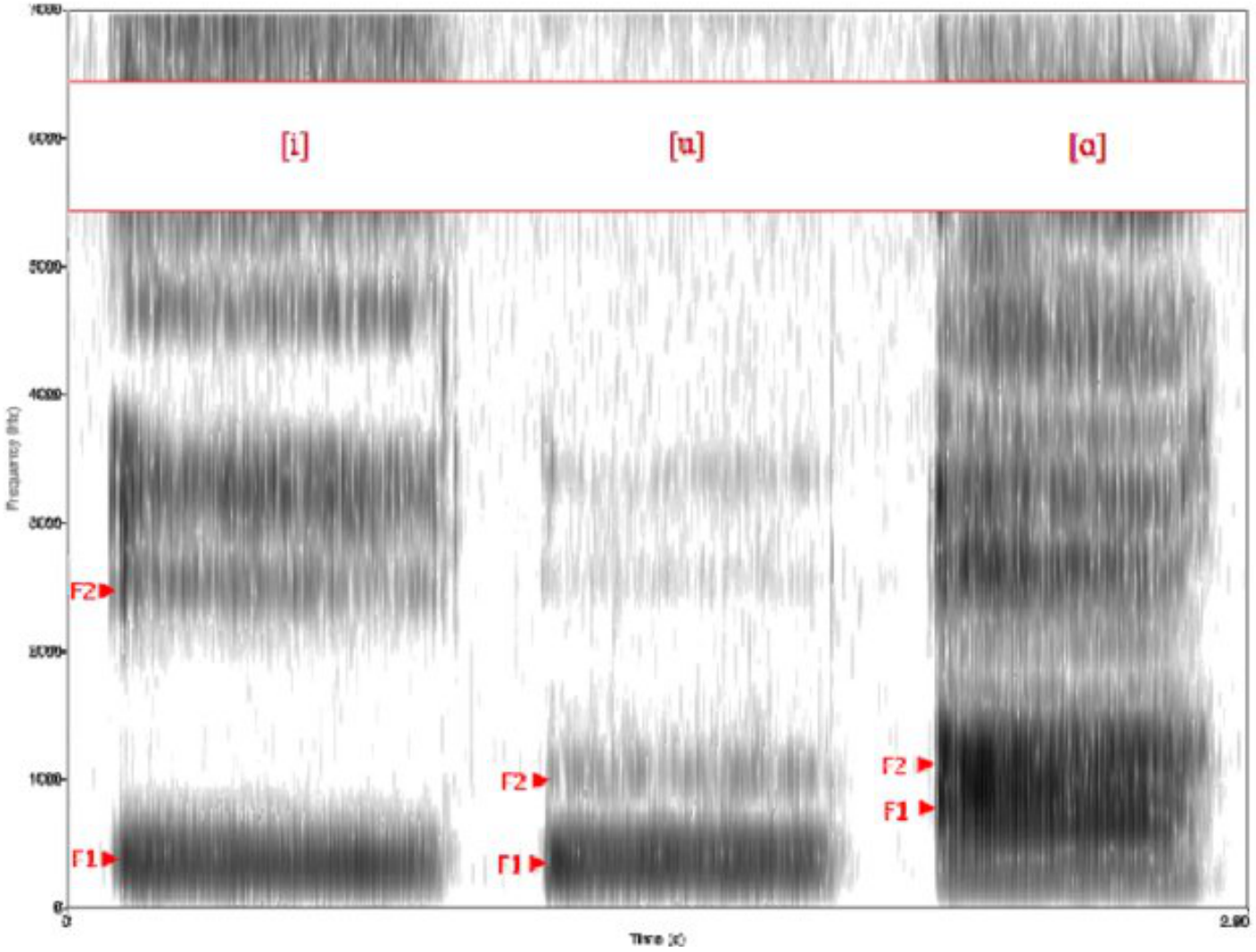
- - - Speech Processing - - -



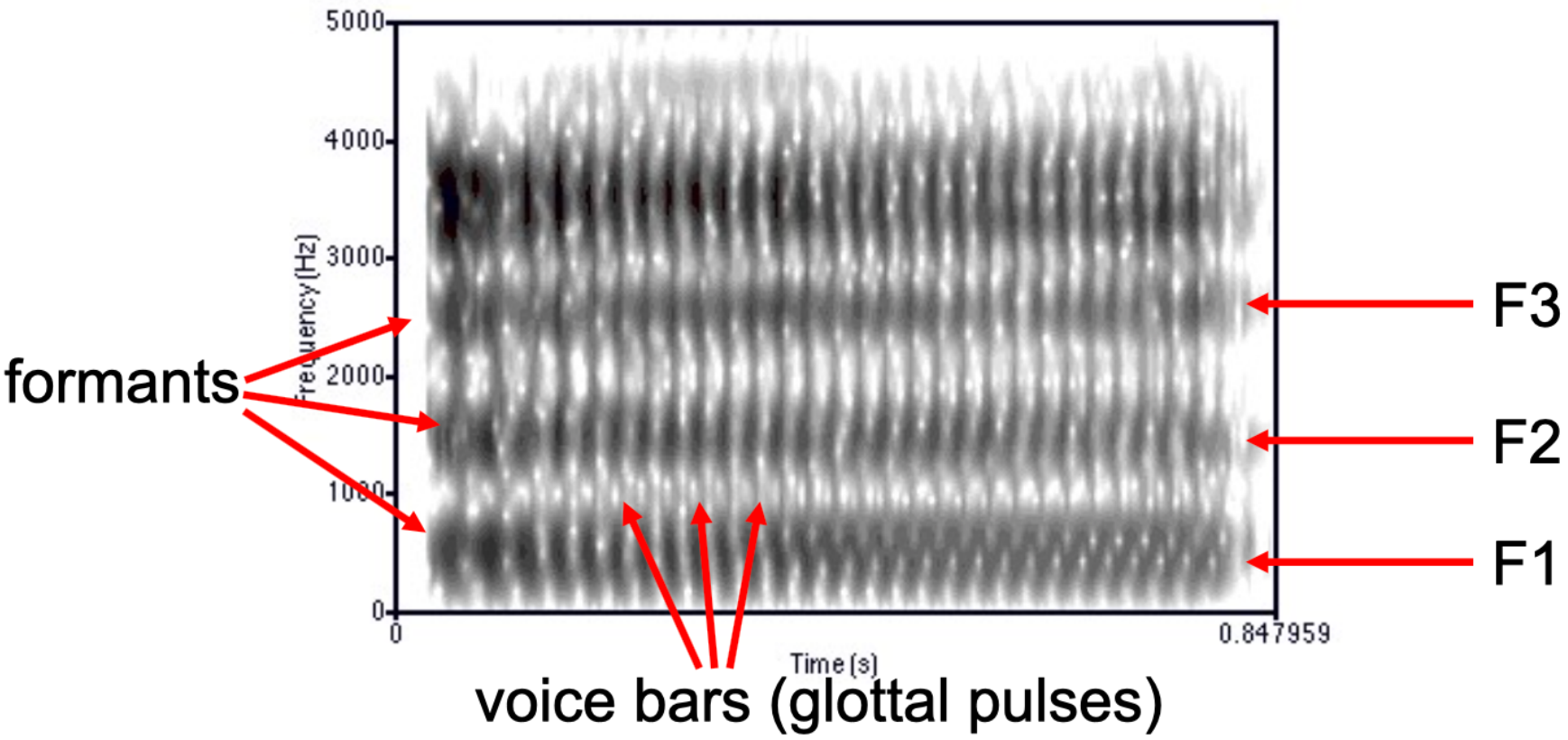
- - - Speech Processing - - -



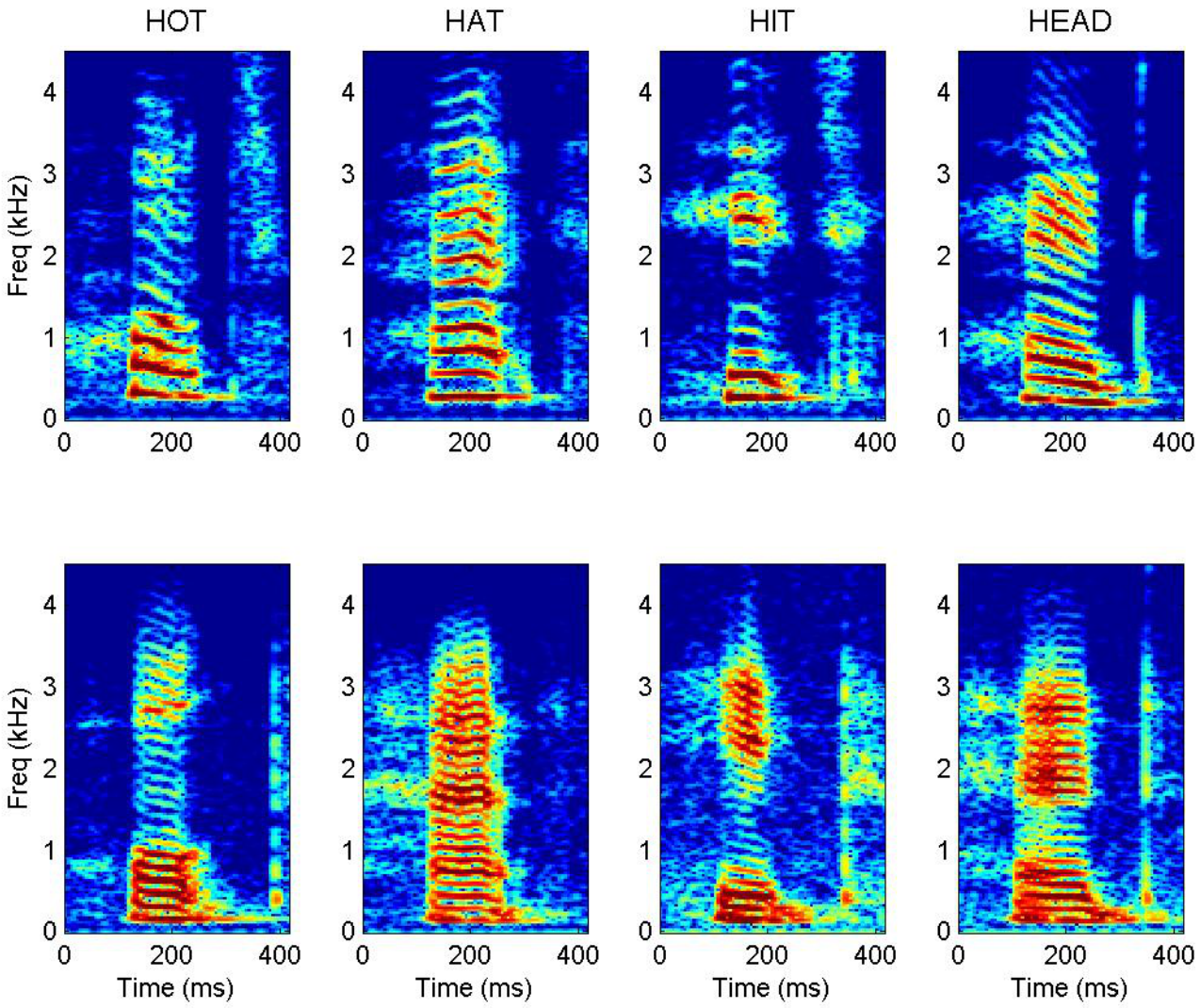
- - - Speech Processing - - -



- - - Speech Processing - - -

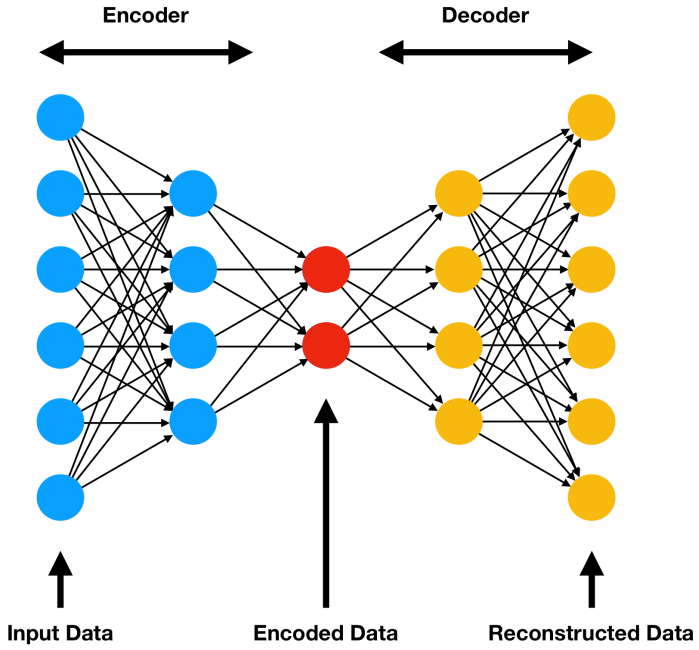


- - - Speech Processing - - -



- - - Speech Processing - - -

- **Feature Learning**: in this case, the features are not extracted based on specific concepts. Instead, a neural network model is used to learn the features directly from speech data. One of the possibilities to do so is known as autoencoder, as shown below. The number of hidden layers and the encoded data dimension vary, being usually determined experimentally. This topic is detailed in Artificial Intelligence and Machine Learning courses.



- - - Speech Processing - - -

- **Hidden Markov Model (HMM) Classifier**: Decades ago, HMMs played a very important role in speech processing applications, primarily for speech recognition purposes. Nowadays, since the advent of deep learning-based strategies, it is no longer the strongest model. Nevertheless, it is still used and, thus, it is worth to comment on its general structure.

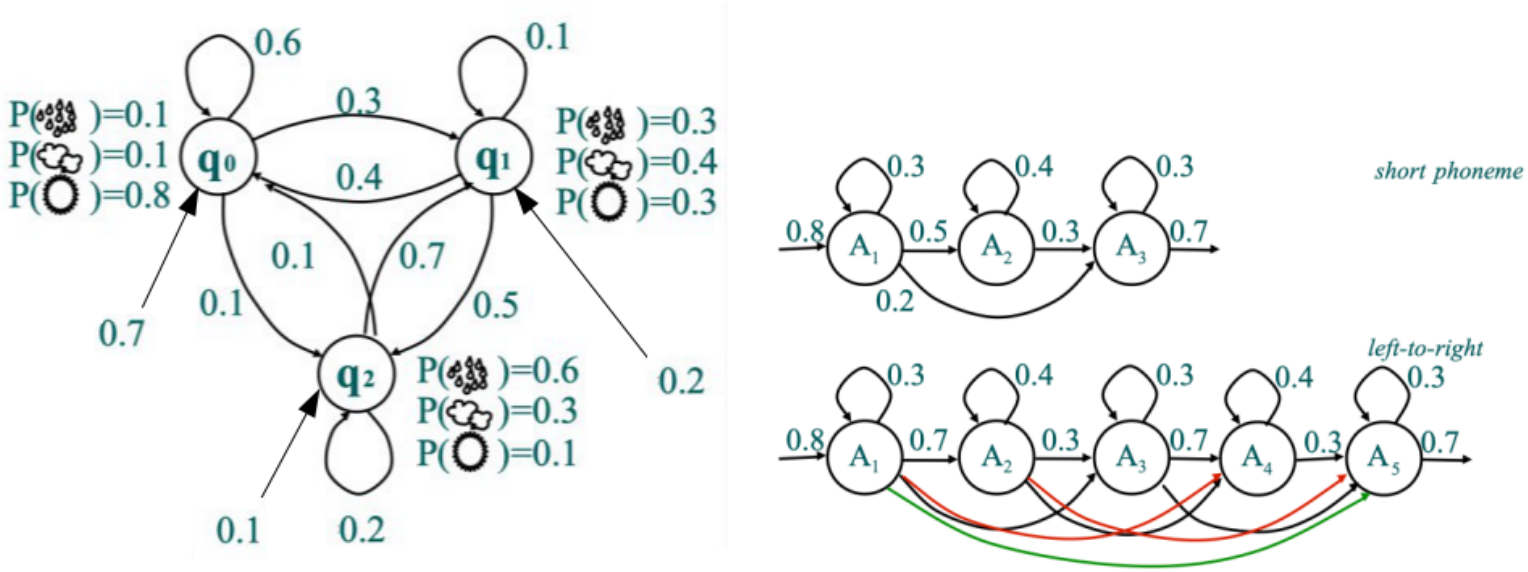


Figure: [left]: a general HMM model; [right]: HMM typical topologies used in speech recognition applications.