- - - Speech Processing - - -

- **Cepstrum**: used in practice to find F_0, as one of the most traditional methods, it was developed while studying echoes in seismic signals, in the 60's. Cepstrum is defined just as the Inverse Fourier Transform of the log magnitude of the Fourier Transform of a certain input speech signal $s[n]$, i.e., $C(s[n]) = F^{-1}(log(|F(s[n])|))$. The reason for that transformations becomes evident when we recall that the speech signal ($Y(\omega)$) is the result of filtering the input signal, i.e., the glotal pulse ($X[z]$), by using the vocal tract response ($H[z]$), i.e.,

$$Y(j\omega) = X(j\omega) \cdot H(j\omega)$$

$$\log(Y(j\omega)) = \log\Big(X(j\omega) \cdot H(j\omega)\Big)$$

$$\log(Y(j\omega)) = \log(X(j\omega)) + \log(H(j\omega)) \quad .$$

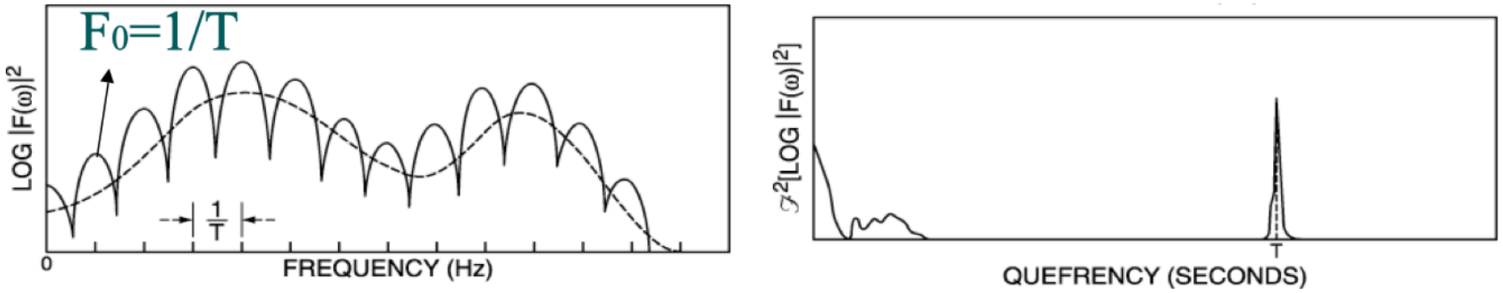
Thus, the speech signal in the log frequency domain can be easily separated into two log frequency signals.

- - - Speech Processing - - -

- Once the Inverse Fourier Transform is taken, we get:

$$F^{-1}\Big(\log(Y(j\omega))\Big) = F^{-1}\Big(\log(X(j\omega)) + \log(H(j\omega))\Big)$$
$$= F^{-1}\Big(\log(X(j\omega))\Big) + F^{-1}\Big(\log(H(j\omega))\Big) \quad .$$

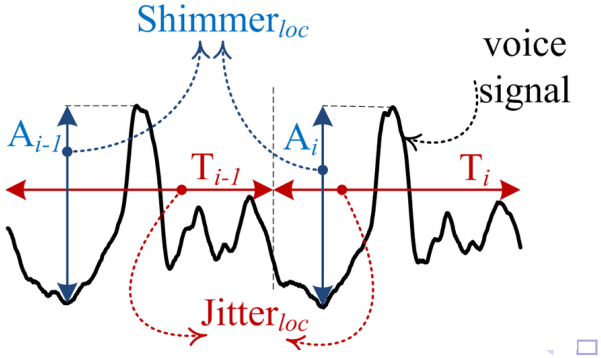
Thus, each one of the components in a more "dilated" scale is distant from the other, being "lifting" the name of such a separation. That scale is called "quefrency" and is measured in seconds.



Consequently, based on cepstrum, we can easily find the pitch frequency, i.e., $F_0 = \frac{1}{T}$ for voiced speech signals or frames.

- - - Speech Processing - - -

- **F_0-related deviations over time**: dozens of possibilities exist to measure F_0-related deviations over time. Among them, the following are the most common, with applications focused on voice pathology detection and speaker identification:
  - **jitter**: $\frac{\sum_{i=1}^{N-1} |P_i - P_{i+1}|}{N-1}$, where each P_i corresponds to the value of F_0 in the i-th frame under analysis, considering that N frames exist.
  - **shimmer**: $\frac{\sum_{i=1}^{N-1} |A_i - A_{i+1}|}{N-1}$, where each A_i corresponds to the amplitude value of F_0 in the i-th frame under analysis, considering that N frames exist.



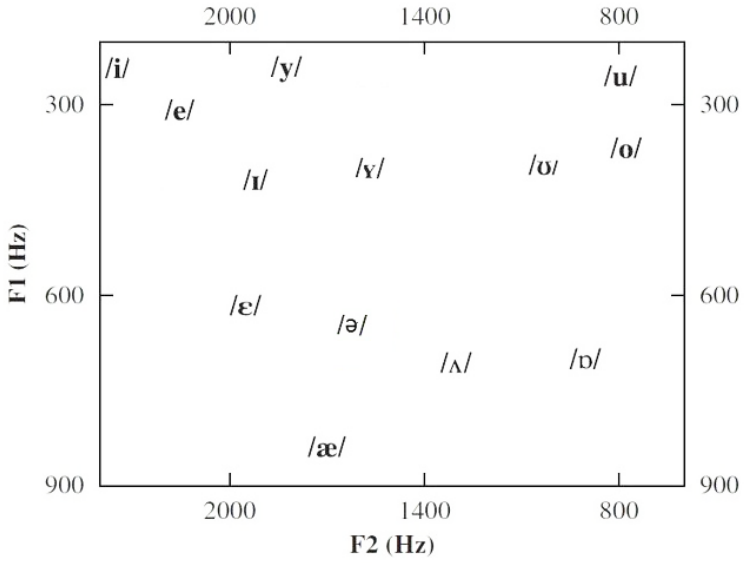
- - - Speech Processing - - -

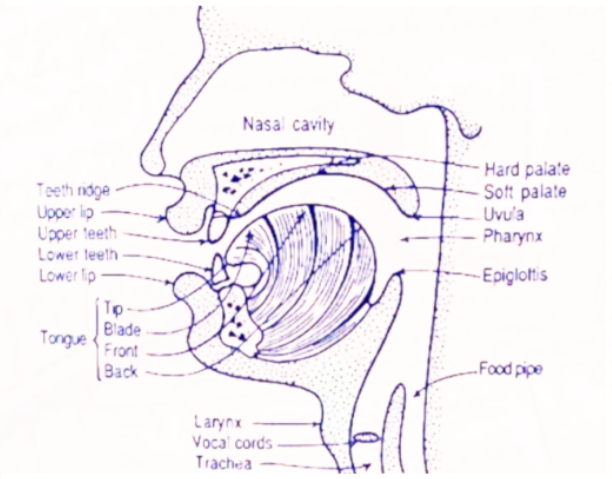
- **entropy**: also known as Shannon's entropy (H), it may be understood as a measure of unpredictability of information content, whereas it equals zero upon a concrete and fully predictable outcome. It is defined as $H = -\sum_{i=0}^{K-1} p_i \log_\beta(p_i)$, where β is the basis adopted, such as 2 or 10, and p_i is the probability of the i-th distinct datum, i.e., symbol, in a set, or array, of size M with K distinct symbols. A detailed description of entropy, with examples and applications might be found in paper "GUIDO, R.C. A Tutorial-review on Entropy-based Handcrafted Feature Extraction for Information Fusion. *Information Fusion*, n.41, pp.161-175, (2018)". Entropy can be extracted from speech frames and can be applied mainly to voice pathology detection, speech recogition, speaker identification, emotion recognition, and so on.

  Example: calculate the entropy of the set $\{6, 4, 4, 3, 8, 4, 3, 4\}$, using $\beta = 2$ as an example.

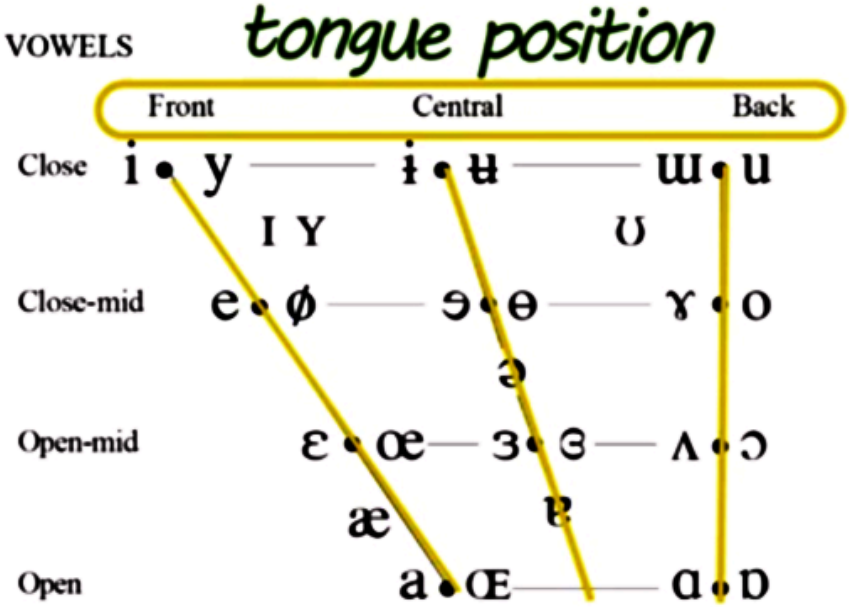
- - - Speech Processing - - -

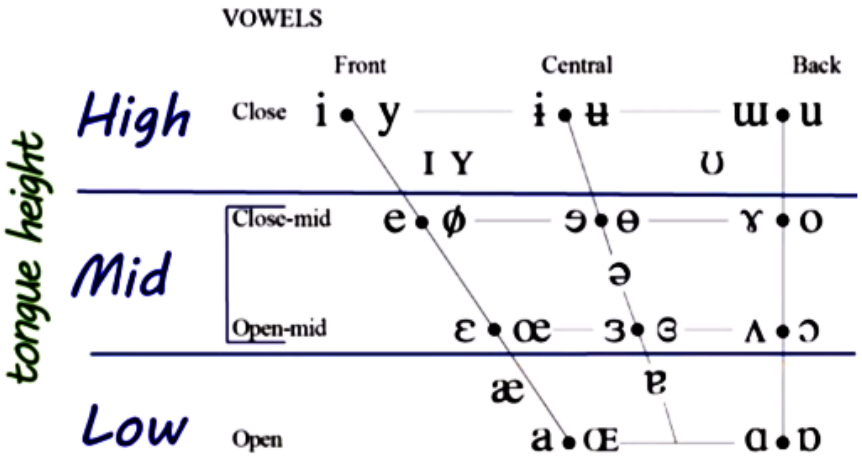
- **Formant Frequencies ($F_1, F_2, F_3, ...$)**: formants are very important parameters for speech recognition and speaker identification. We may find formants based on two main approaches, i.e., either by extracting the envelope of a speech frame or by using linear predictive coding (LPC). For the former, different possibilities exist, such as the Least Squares Method (LSM) or the Hilbert Transform. Disregarding the specific technique used, we can observe that the joint distribution of F_1 and F_2 reveals voiced sounds, as in the example below.





- - - Speech Processing - - -





- **Today's Short Test (ST11)**: Based on AMDF, find F_0 for the signal $s[n] = \{0, 1, 3, 0, 1, 3, 0, 1, 3, 0\}$ that was sampled at r samples per second.

- - - Speech Processing - - -

- **Least Mean Squares (LMS)**: among many other applications in science and engineering, LMS method might be used for formant tracking. To do so, we might derive an LMS-based approximation for the speech spectrum by using a polynomial of degree $8 \sim 10$, more or less, as mentioned during the class. Nevertheless, this is not the most used possibility in practice.
- **Linear Predictive Coding (LPC)**: LPC has been intensively used in digital speech processing, since the 70's, for formant tracking and compression, being, for instance, the basis for voice over IP (VoIP) development. LPC allows for the spectral envelope to be obtained from the speech spectrum, as detailed during the class and as follows. Considering $y[n]$ is the speech signal under analysis and $e[n]$ is the prediction error, we have:

$$\begin{aligned} y[n] &= (-a_1 y_{n-1} - a_2 y_{n-2} - a_3 y_{n-3} - ....) + e_n \\ Y[z] &= (-a_1 Y[z]z^{-1} - a_2 Y[z]z^{-2} - a_3 Y[z]z^{-3} - ...) + E[z] \end{aligned}$$

- Thus, we have:

- - - Speech Processing - - -

$$Y[z] + a_1 Y[z]z^{-1} + a_2 Y[z]z^{-2} + a_3 Y[z]z^{-3} = E[z]$$

$$Y[z]\left(1 + a_1 z^{-1} + a_2 z^{-2} + a_3 z^{-3} + ...\right) = E[z]$$

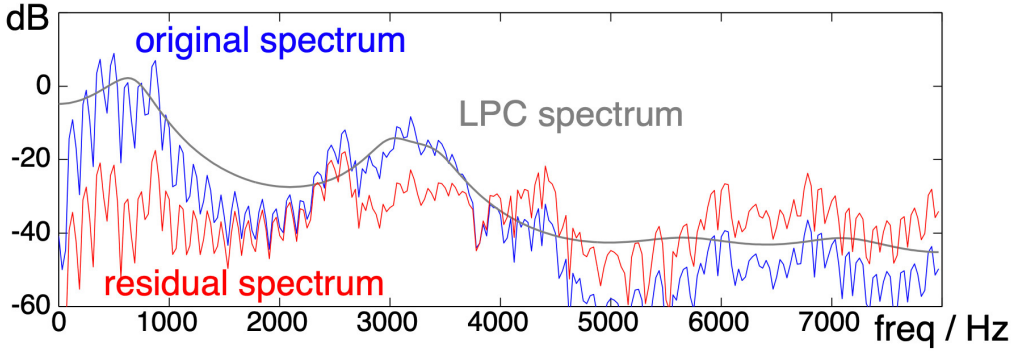
$$\frac{Y[z]}{E[z]}\left(1 + a_1 z^{-1} + a_2 z^{-2} + a_3 z^{-3} + ...\right) = 1$$

$$\frac{Y[z]}{E[z]} = \frac{1}{\left(1 + a_1 z^{-1} + a_2 z^{-2} + a_3 z^{-3} + ...\right)}$$

- Comparing the previous expression with the model we derived for voiced speech, on page 58, we note that:
  - $E[z]$ is just the excitation signal $X[z]$
  - a_0 is normalized to the unit
  - $a_1, a_2, a_3, ...$ are the main coefficients
- Consequently, once we somehow find $a_1, a_2, a_3, ...$, we might easily obtain $\frac{Y[z]}{X[z]} = H[z]$, i.e., the system transfer function which corresponds to the vocal tract frequency response.

- - - Speech Processing - - -

- From that frequency response curve, we just find the peaks, which correspond to the formant frequencies, i.e., $F_1, F_2, F_3, ...$ In addition, we observe that LPC allows us to predict the n-th sample of the speech signal as a linear combination of its past samples, considering an error signal $e[n]$. In practice, we use a predictor with order $12 \sim 18$ together with the LSM to find the LPC coefficients $\{a_1, a_2, a_3, ...\}$. Let us see an example.



- **Example**: Assume that the hypothetical speech signal $y[n] = \{2, 4, 6, \overline{7, 8, 5, 9}\}$. Let us use, just in this example expected to be performed by hand, a 3^{rd} order predictor to estimate the coefficients $\{a_1, a_2, a_3\}$ of the LPC method.