Feature Extraction

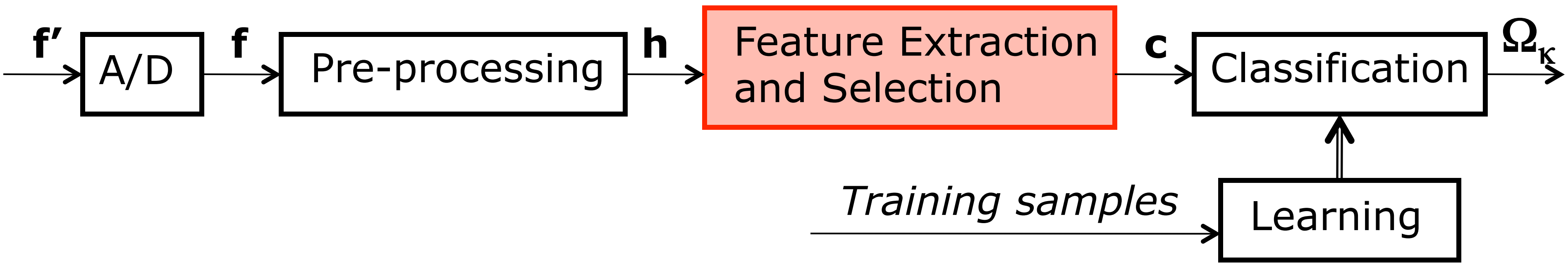
Linear Predictive Coding, Moments

Dr. Elli Angelopoulou

Lehrstuhl für Mustererkennung (Informatik 5)

Friedrich-Alexander-Universität Erlangen-Nürnberg

Pattern Recognition Pipeline



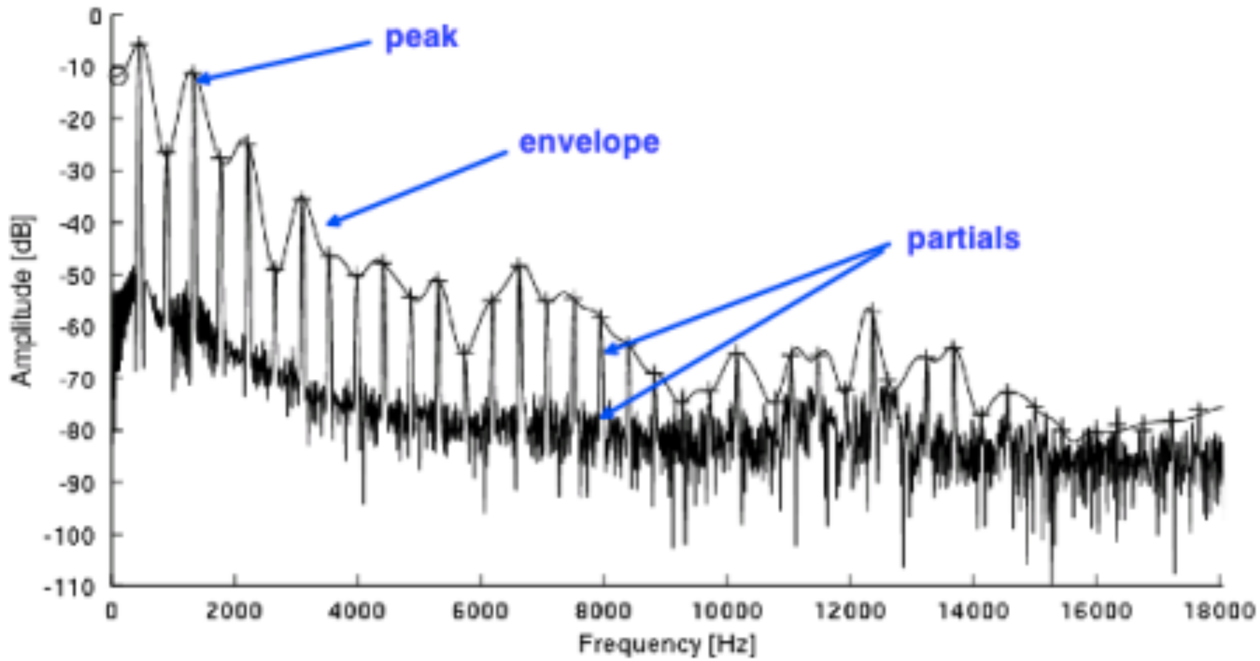
- One common method for heuristic feature extraction is the projection of a signal $\vec{h}$ or $\vec{f}$ on a set of orthogonal basis vectors (functions), $\Phi = \left[\vec{\varphi}_1, \vec{\varphi}_2, \ldots, \vec{\varphi}_M\right]$

$$\vec{c} = \Phi^T \vec{f}$$

Introduction to Linear Predictive Coding

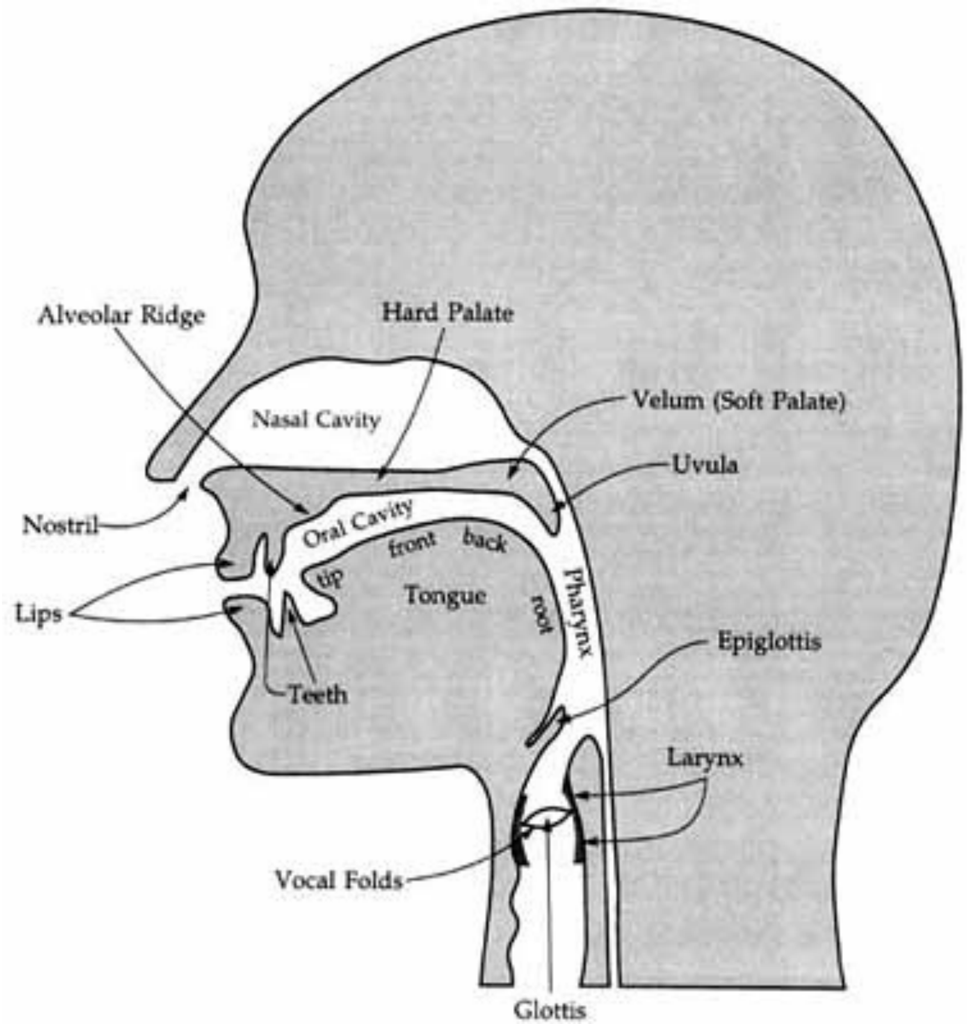
- Linear Predictive Coding (LPC) is a feature vector that is widely used in speech processing.
- It represents the spectral envelope of a digital signal of speech in a compressed form.
- LPC has been very successful in encoding good quality speech at a low bit rate.
- It also provides extremely accurate estimates of speech parameters.
- It is part of the GSM wireless communication standard.

Spectral Envelope



- The spectral envelope is a curve in the frequency domain.
- It is derived from the Fourier magnitude spectrum.
- It is a smooth function that passes through the prominent peaks of the spectrum.
- It changes over time.

Vocal Tract



- There are 3 key elements in the human vocal tract:
  - Vocal Cords
  - Pharynx
  - Oral/Nasal Cavity
- LPC assumes such an apparatus for voice/sound generation.

Abstract Model of Vocal Tract

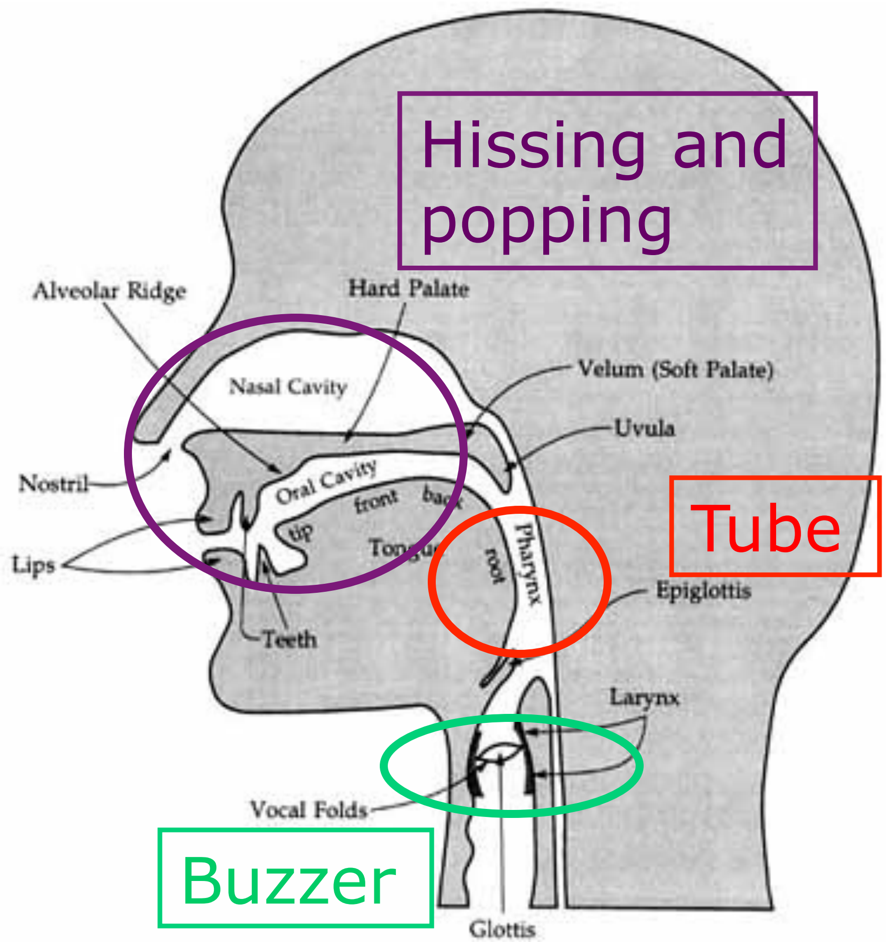
- An abstract model of the speech synthesis is often employed.
- Its key components are:
  - Buzzer
  - Tube
- The relationship between the vocal tract and the abstract model for speech production is:
  - Lungs
  - Trachia
  - Vocal cords -> Buzzer
  - Pharynx -> Tube
  - Oral cavity
  - Nasal cavity

Additional hissing and popping sounds

LPC and the Vocal Tract

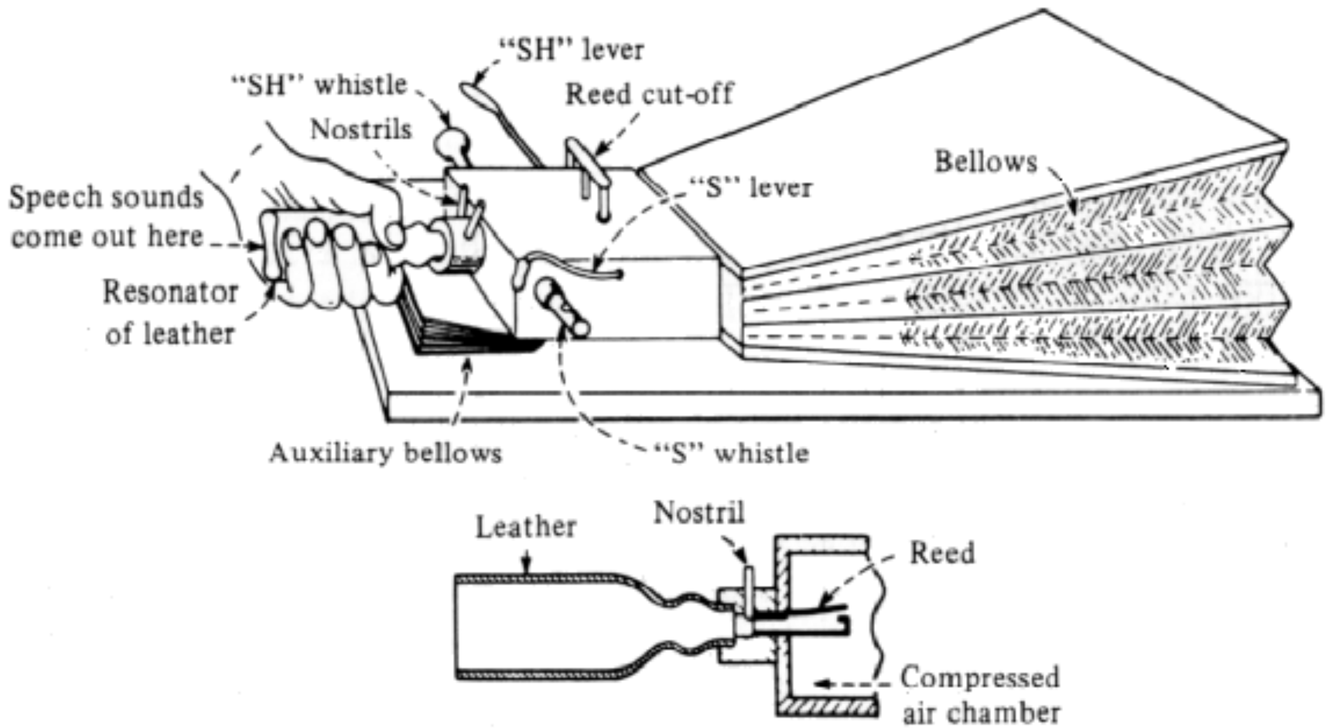
- LPC starts with the assumption that a speech signal is produced by a **buzzer** at the end of a **tube** (*voiced sounds*), with occasional added hissing and popping sounds (*sibilants and plosive sounds*).
- The glottis (the space between the vocal cords) produces the buzz, which is characterized by its *intensity* (loudness) and *frequency* (pitch).
- The pharynx forms the tube, which is characterized by its *resonances*, which are called *formants*.
- Hisses and pops are generated by the action of the tongue, lips and throat.

Vocal Tract



- There are 3 key elements in the human vocal tract:
  - Vocal Cords
  - Pharynx
  - Oral/Nasal Cavity
- They are abstracted to:
  - Buzzer
  - Tube
  - Hissing & Popping generator

An Early Speech Synthesizer



- Wheatstone's reconstruction of von Kempelen's speaking machine.
- Vowels were produced with vibrating reed and all passages were closed.
- Resonances were effected by deforming the leather resonator.
- Consonants, including nasals, were produced with turbulent flow trough a suitable passage with reed-off .

Formants

- In an acoustic signal formants are the peaks in the envelope of the sound signal. Such a peak may not be discernible in high-pitched sounds (kids, some women’s voices).
- Formants are the distinguishing frequency components in speech and singing.
- Vowels are identified by their distinct frequency content.
- Vowels have typically four or more distinguishable formants.

LPC and the Vocal Tract - continued

- LPC analyzes the speech signal by:
  - estimating the formants (the pharynx effects)
  - removing their effects from the speech signal
  - and estimating the intensity and frequency of the remaining buzz.
- LPC isolates the intensity and frequency of the buzz and the formants effects.
- Each (buzz effects and formant effets) can be stored (processed if needed) and transmitted separately.
- They are then recombined at the receiving end to create the speech signal.

Linear Predictive Model

- Assume that the present sample f_n of the speech is predicted by the past *m* speech samples so that

$$\hat{f}_n = a_1 f_{n-1} + a_2 f_{n-2} + \cdots + a_m f_{n-m} = \sum_{\mu=1}^{m} a_\mu f_{n-\mu}$$

  where $\hat{f}_n$ is the prediction of f_n, f_{n-i} is the sample of the i^th previous step, and the a_μ's are are the linear prediction coefficients (LPCs).

- The error between the actual sample and the predicted one is:

$$e_n = f_n - \hat{f}_n = f_n - \sum_{\mu=1}^{m} a_\mu f_{n-\mu}$$

- The best LPCs will result in $e_n = 0$.

Computation of the LPC-coefficients

- The prediction error is: $e_n = f_n - \hat{f}_n = f_n - \sum_{\mu=1}^{m} a_\mu f_{n-\mu}$
- Goal: Derive the LPCs a_μ that result in:

$$e_n = 0 \Rightarrow f_n - \sum_{\mu=1}^{m} a_\mu f_{n-\mu} = 0 \Rightarrow f_n = \sum_{\mu=1}^{m} a_\mu f_{n-\mu}$$

- How do we compute the values of the coefficients that satisfy

$$f_n = \sum_{\mu=1}^{m} a_\mu f_{n-\mu}$$

- Use additional *k* samples to obtain a system of linear equations from where one can compute a_μ.

System of Linear Equations

- From the last *k+1* samples we have:

$$f_n = \sum_{\mu=1}^{m} a_\mu f_{n-\mu}$$

$$f_{n+1} = \sum_{\mu=1}^{m} a_\mu f_{n+1-\mu}$$

$$\vdots$$

$$f_{n+k} = \sum_{\mu=1}^{m} a_\mu f_{n+k-\mu}$$

- We have k+1 equations which are all linear in a_μ.

Matrix Form

- Rewrite the system of equations in a matrix form:

$$\begin{bmatrix} f_n \\ f_{n+1} \\ \vdots \\ f_{n+k} \end{bmatrix} = \begin{bmatrix} f_{n-1} & f_{n-2} & \cdots & f_{n-m} \\ f_n & f_{n-1} & \cdots & f_{n+1-m} \\ \vdots & \vdots & \cdots & \vdots \\ f_{n+k-1} & f_{n+k-2} & \cdots & f_{n+k-m} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix}$$

$$\begin{bmatrix} f_n \\ f_{n+1} \\ \vdots \\ f_{n+k} \end{bmatrix} = A \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix} \Rightarrow \vec{f} = A\vec{a}$$

- *A* is a (k+1) x m matrix of observed signals.
- $\vec{f} \in R^{k+1}$.
- $\vec{a} \in R^m$.

Computing the Vector of LPC coefficients

- If $m = k + 1$, then *A* is a square matrix and thus it is invertible (assuming that $\det(A) \neq 0$).
- Hence the LPC coefficients are:

$$\vec{a} = A^{-1}\vec{f}$$

- If $m \neq k + 1$, then?
- We have to use the *pseudoinverse*: $A^{+} = (A^{T}A)^{-1}A^{T}$
- In this case the LPC coefficients are:

$$\vec{a} = A^{+}\vec{f}$$

- The best way to compute the pseudoinverse is to use singular value decomposition (SVD).

Alternative Estimation of LPC-coefficients

- Alternatively, we could define an objective function.

$$\varepsilon = \sum_{n=n_0}^{n_1} \left(f_n - \hat{f}_n \right)^2 =$$

$$\varepsilon = \sum_{n=n_0}^{n_1} \left(f_n - \sum_{\mu=1}^{m} a_\mu f_{n-\mu} \right)^2$$

- We then have to find the values of the LPC coefficients that minimize the error.

$$\frac{\partial \varepsilon}{\partial a_\nu} = 2 \sum_{n=n_0}^{n_1} \left(f_n - \sum_{\mu=1}^{m} a_\mu f_{n-\mu} \right) f_{n-\nu} = 0 \Rightarrow \sum_{n=n_0}^{n_1} f_n f_{n-\nu} = \sum_{\mu=1}^{m} a_\mu \sum_{n=n_0}^{n_1} f_{n-\mu} f_{n-\nu}$$

Three Remarks on LPC

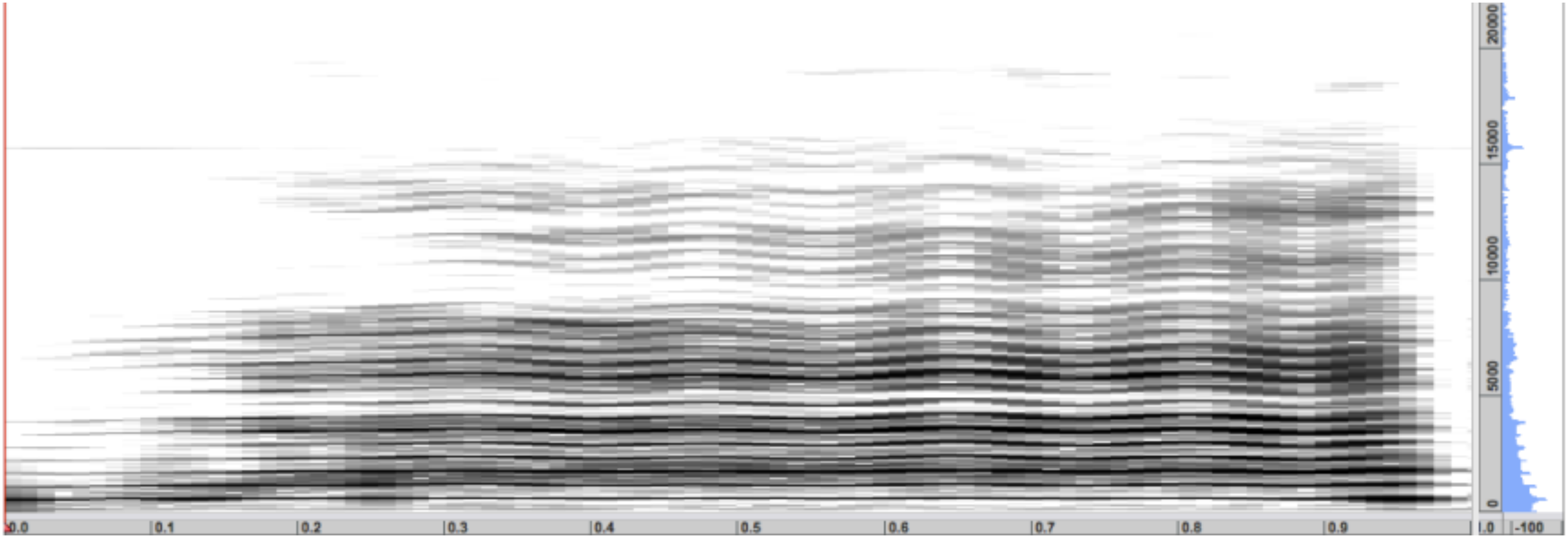
1. Rule of thumb for the number of coefficients:
   - m = 10 -15
   - The choice of m depends on the sampling frequency.
   - Let f_s be the sampling frequency in kHz, then
   - $m = 4 + f_s$ up to $m = 5 + f_s$
2. One can use the LPC coefficients to identify a person's voice.
   - LPC is particularly good at highlighting formant locations which have been shown to be significant in voice identification.
3. The vector of LPC coefficients can be used as a feature vector.

$$\vec{c} = \vec{a}$$

Discussion on *m*

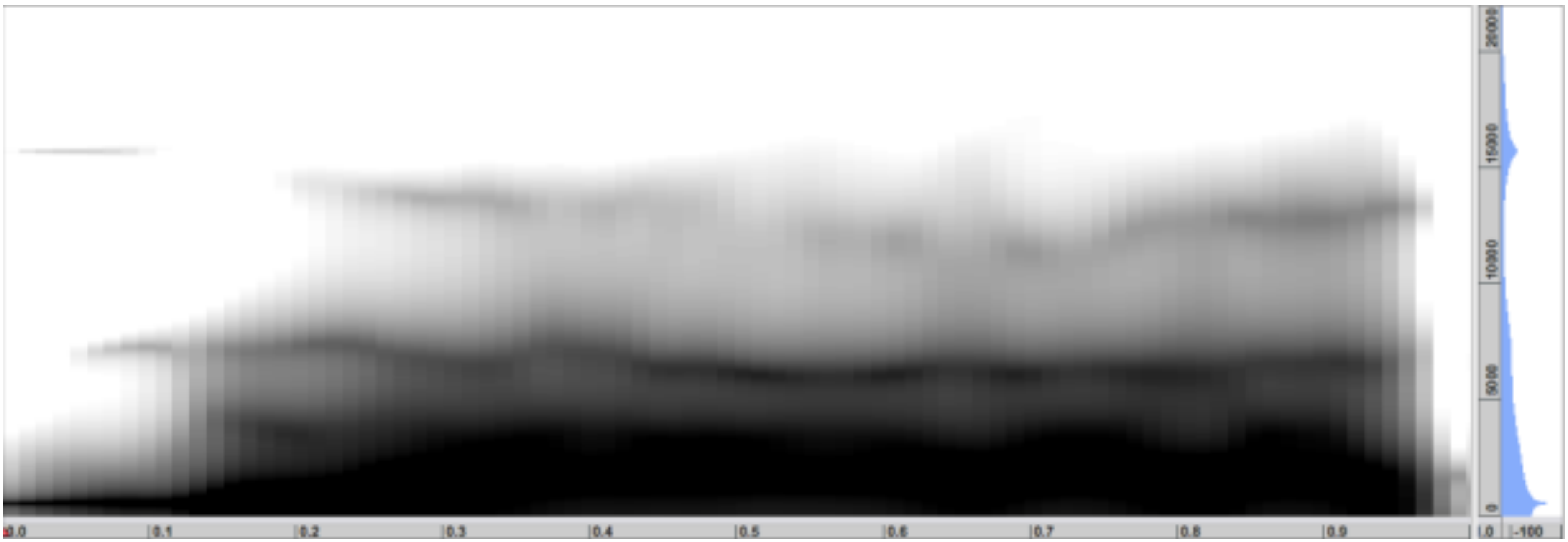
- *m* is also referred to as the order of the LPC model, or the number of poles of the LPC model.
- It is related to the number or formants.
- Typically we employ two poles per formant. Two to four additional poles are added to represent the overall source characteristics.
- For example:
  - 10kHz => *m* = 12-14 for males and 8-10 for females
  - 22kHz => *m* = 24-26 for males and 22-24 for females
- If no information is available *m* is set to 30.

LPC, *m* and FT



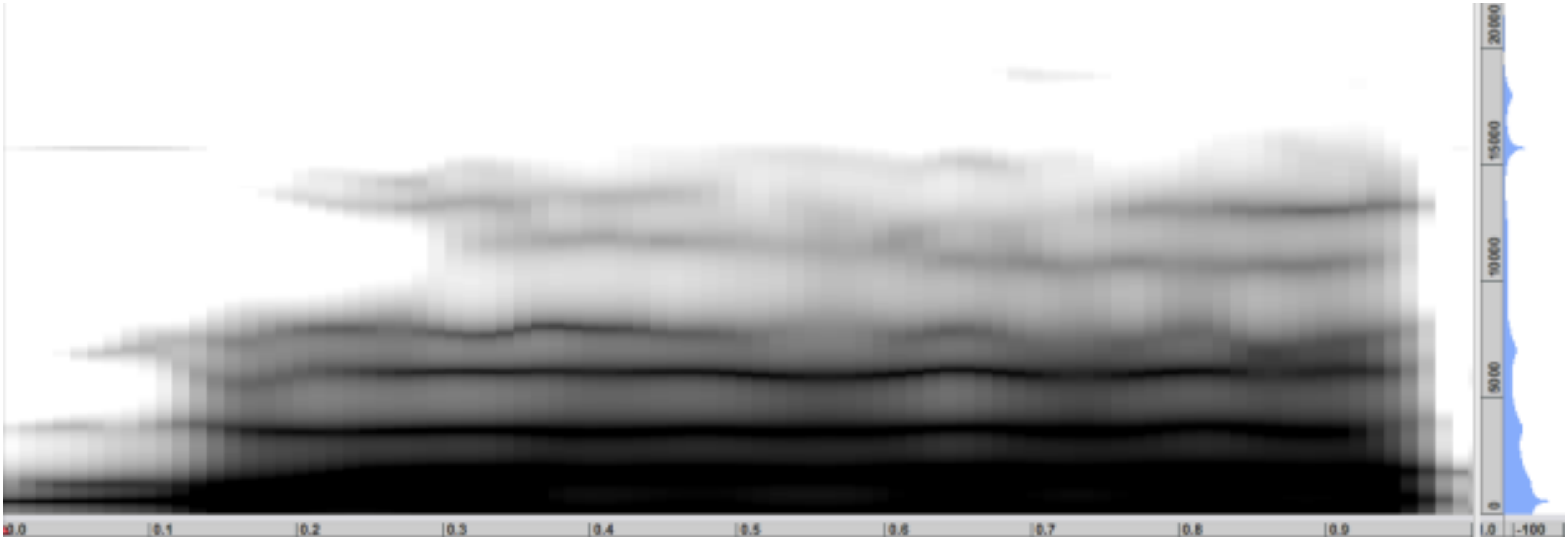
- The FFT of a vowel of a singing voice

LPC, *m* and FT



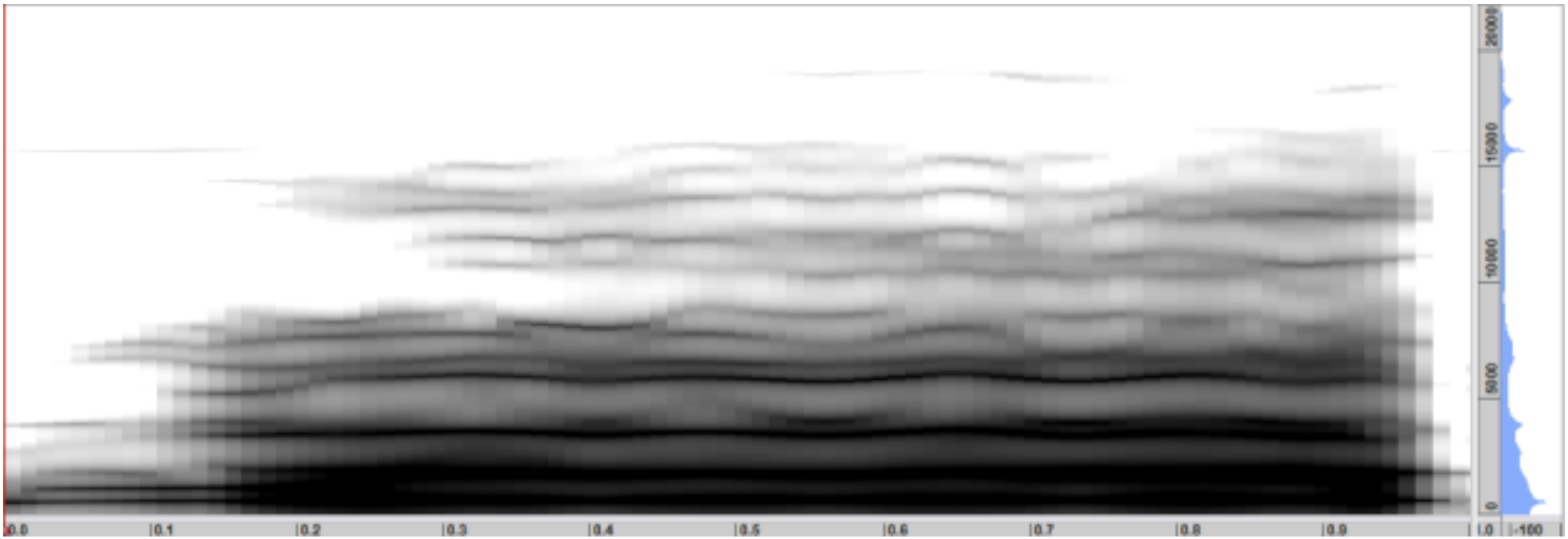
- The LPC of the same signal with m=10.

LPC, *m* and FT



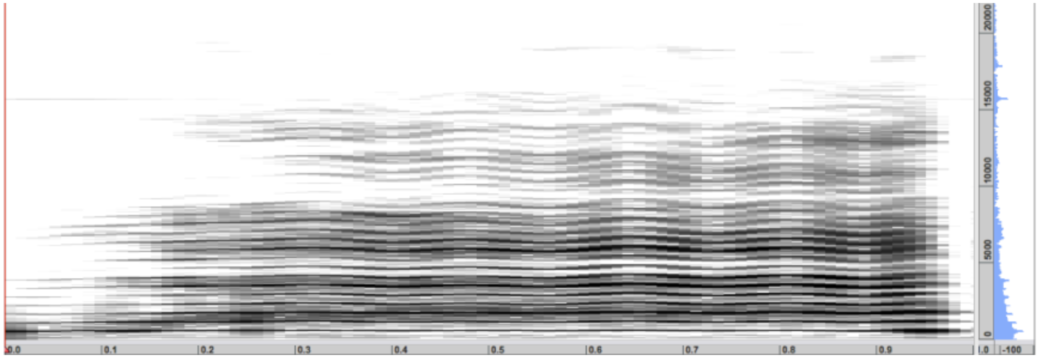
- The LPC of the same signal with m=20.

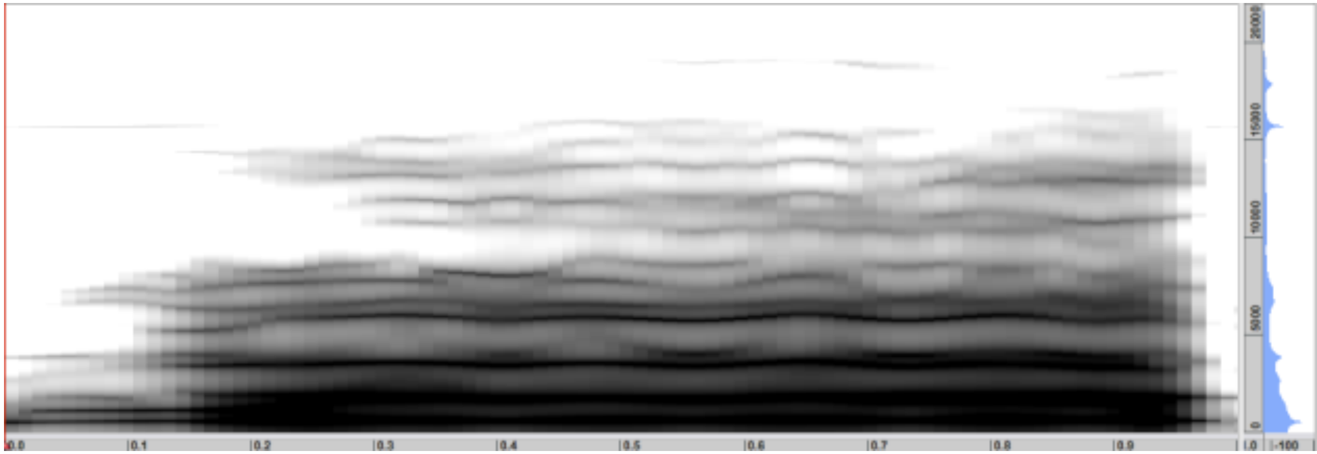
LPC, *m* and FT



- The LPC of the same signal with m=40.

LPC, *m* and FT





- The FFT of a signal and its LPC for m=40

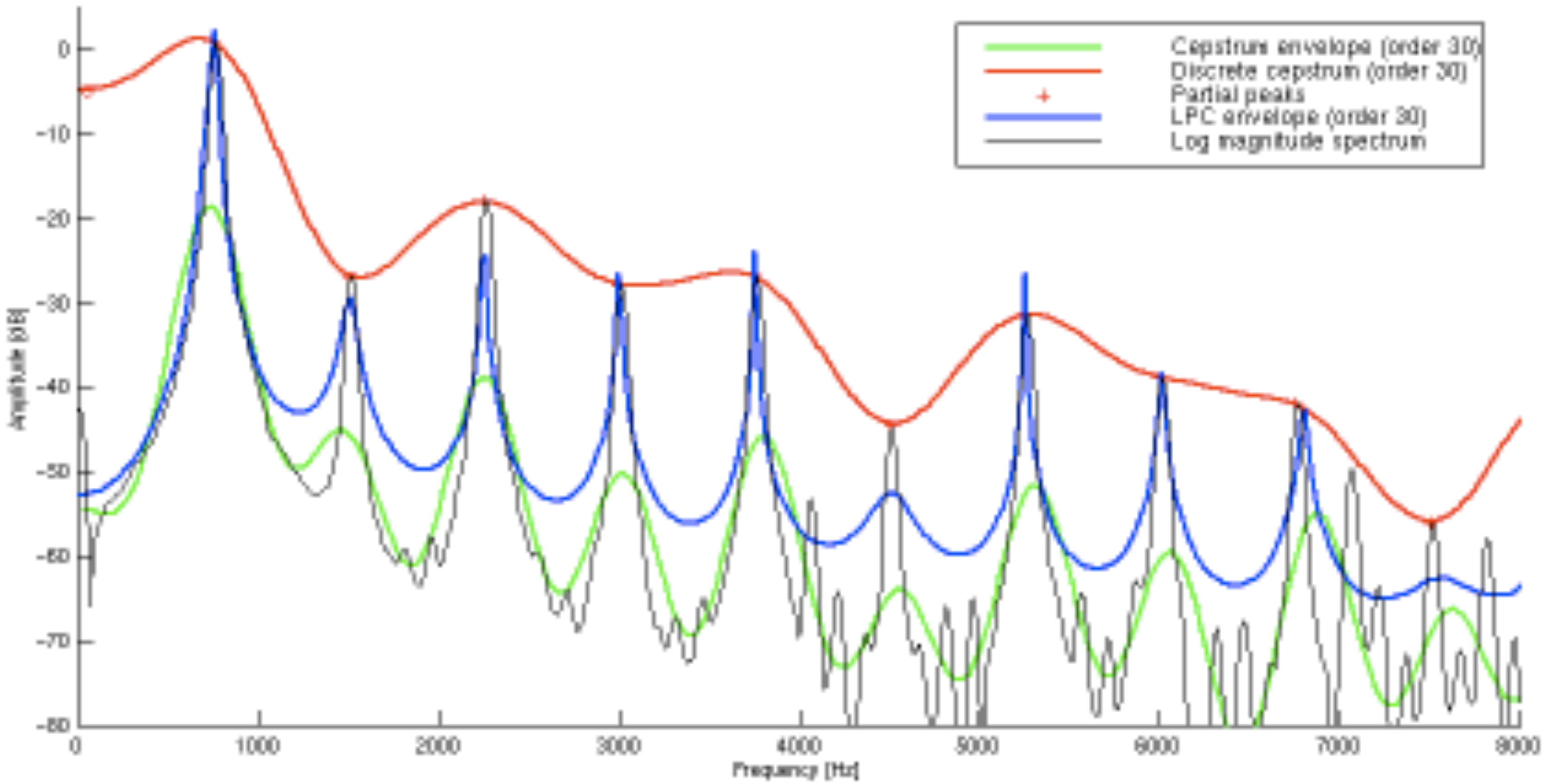
Final Remarks on LPC

4. One can use the LPC coefficients to compute the smoothed **Model Spectrum** of a signal.
   - The Model Spectrum is the Fourier Transform of the LPC coefficients.

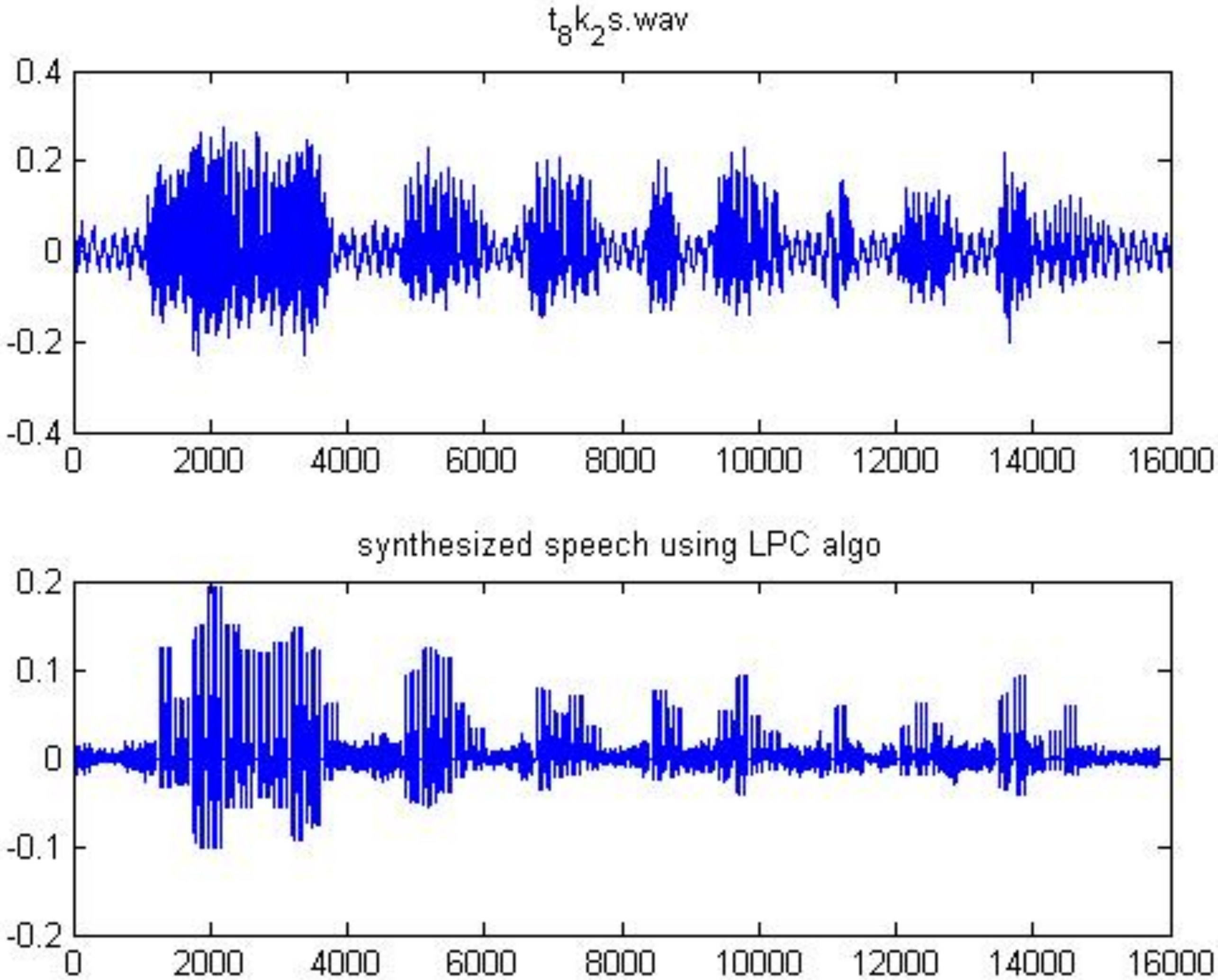
$$\text{ModelSpectrum}(\vec{a}) = \text{FT}(\vec{a})$$

   - It is a smooth spectrum of the speech signal.
   - Peaks in the Model Spectrum are formants.
   - Peaks in the frequency spectrum of a sound are caused by resonance (i.e. they are directly attributed to formants)
   - It has been shown that perceptually, formants is the information that humans use in distinguishing between different vowels.

LPC as Model Spectrum - Example



LPC for Speech Compression



Moments

- Given an image f(x,y), the geometric moments are defined as:

$$m_{pq} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^p y^q f(x,y)\,dxdy$$

- For the same image f(x,y) the central moments are defined as:

$$\mu_{pq} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x-\bar{x})^p (y-\bar{y})^q f(x,y)\,dxdy$$

where $\bar{x} = \frac{m_{10}}{m_{00}}$ and $\bar{y} = \frac{m_{01}}{m_{00}}$ are the center of mass.

Moments and Invariance

- An advantage of the central moments is that they are translation-invariant.
- We can compute another set of moments, the normalized central moments which are also scale-invariant.
- Given an image f(x,y), the normalized central moments are defined as:

$$\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}^{(1+0.5(p+q))}}$$

- Thus, the normalized central moments are translation- and scale-invariant.

Moment-Based Features

- One can also construct moments that are translation, scale and rotation invariant.
- A collection of such moments can be used as a feature vector $\vec{c}$.
- Each element c_i of the feature vector is a moment, i.e. $m_{pq}, \mu_{pq}, \eta_{pq}$ for any chosen value of *p* and *q*, or a combination of moments.
- A very popular set of moments used as a feature vector are the ones proposed by Hu. The are known as the Hu set of invariant moments.

Information Provided by Moments

- 1st order moments convey information about size, area, volume, or mass.
- 2nd order central moments are related to variance.
- 3rd order central moments provide information about the symmetry of an shape or distribution (skewness).
- 4th order central moments is a measure of whether the distribution is tall and skinny or short and squat, compared to the normal distribution of the same variance (kurtosis).
- In general in higher orders, central moments provide more intuitive information than moments about zero (raw geometric moments).

Hu Set of Invariant Moments (1 through 5)

$$I_1 = \eta_{20} + \eta_{02}$$

$$I_2 = \left(\eta_{20} - \eta_{02}\right)^2 + \left(2\eta_{11}\right)^2$$

$$I_3 = \left(\eta_{30} - 3\eta_{12}\right)^2 + \left(3\eta_{21} - \eta_{03}\right)^2$$

$$I_4 = \left(\eta_{30} + \eta_{12}\right)^2 + \left(\eta_{21} + \eta_{03}\right)^2$$

$$I_5 = \left(\eta_{30} - 3\eta_{12}\right)\left(\eta_{30} + \eta_{12}\right)\left[\left(\eta_{30} + \eta_{12}\right)^2 - 3\left(\eta_{21} + \eta_{03}\right)^2\right] +$$
$$\left(3\eta_{21} - \eta_{03}\right)\left(\eta_{21} + \eta_{03}\right)\left[3\left(\eta_{30} + \eta_{12}\right)^2 - \left(\eta_{21} + \eta_{03}\right)^2\right]$$

Hu Set of Invariant Moments (6 through 7)

$$I_6 = (\eta_{20} - \eta_{02})\left[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2\right] + 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03})$$

$$I_7 = (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})\left[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2\right] - (\eta_{30} - 3\eta_{12})(\eta_{21} + \eta_{03})\left[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2\right]$$

Some Remarks on the Hu Set

- J. Flusser and T. Suk showed that the Hu set of invariant moments is:

1. Not independent

   For example, I_2 and and I_3 are dependent so they provide no additional information.

2. Incomplete

   There is no independent 3rd order moment invariant. Low discriminating power.

- A 3rd order independent moment that can be used instead is:

$$I_8 = \eta_{11}\left[(\eta_{30}+\eta_{12})^2-(\eta_{03}+\eta_{21})^2\right]-(\eta_{20}-\eta_{02})(\eta_{30}+\eta_{12})(\eta_{03}+\eta_{21})$$

Sources

1. Spectral envelope plot courtesy of http://support.ircam.fr/docs/AudioSculpt/3.0/co/Spectral%20Intro.html
2. Vocal tract image by Jeff McNeill http://jcarreras.homestead.com/files/phoneticsvocaltract.jpg
3. The figure of Wheatstone’s speech synthesizer is from Sami Lemmetty http://www.acoustics.hut.fi/publications/files/theses/lemmetty_mst/chap2.html
4. Model spectrum image courtesy of http://support.ircam.fr/docs/AudioSculpt/3.0/co/LPC.html
5. Speech compression image courtesy of http://www.mathworks.com/matlabcentral/fx_files/13529/1/10.jpg