Engineering Part IIB: Module 4F11
Speech and Language Processing
Lectures 2/3 : Speech Analysis

Phil Woodland: pcw@eng.cam.ac.uk

Lent 2016

Cambridge University Engineering Department

Speech Analysis

These two lectures cover the following topics

Spectral Analysis of Speech

- DFT/Windowing
- Spectral Properties of Speech Sounds
- Spectrogram

Linear Prediction Models

- All-pole model of speech
- Parameter estimation
- Spectral properties

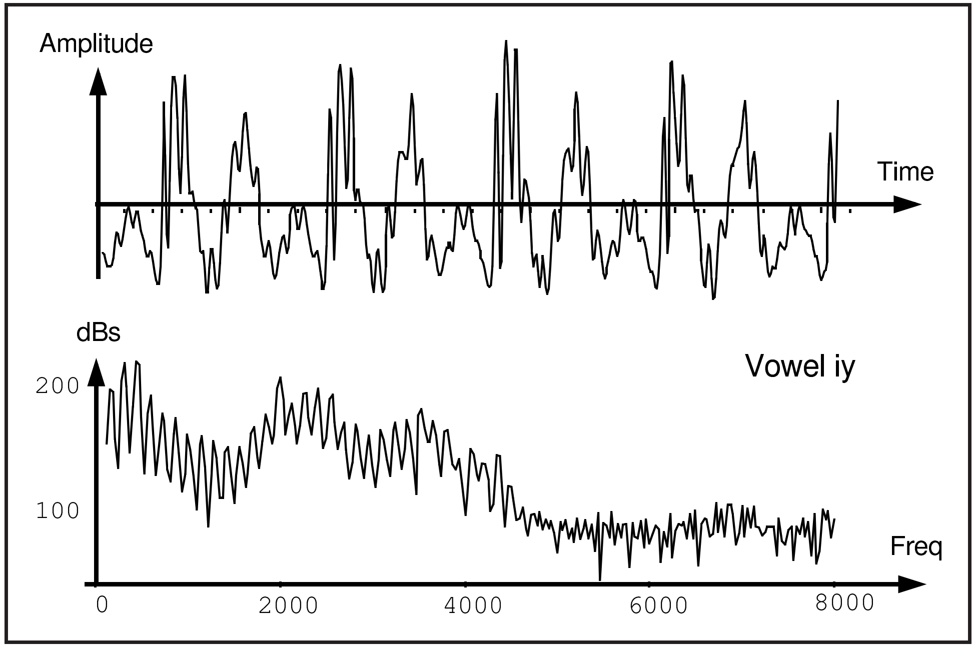
Cepstral Analysis

- Homomorphic filtering
- Filterbank-based cepstral representations

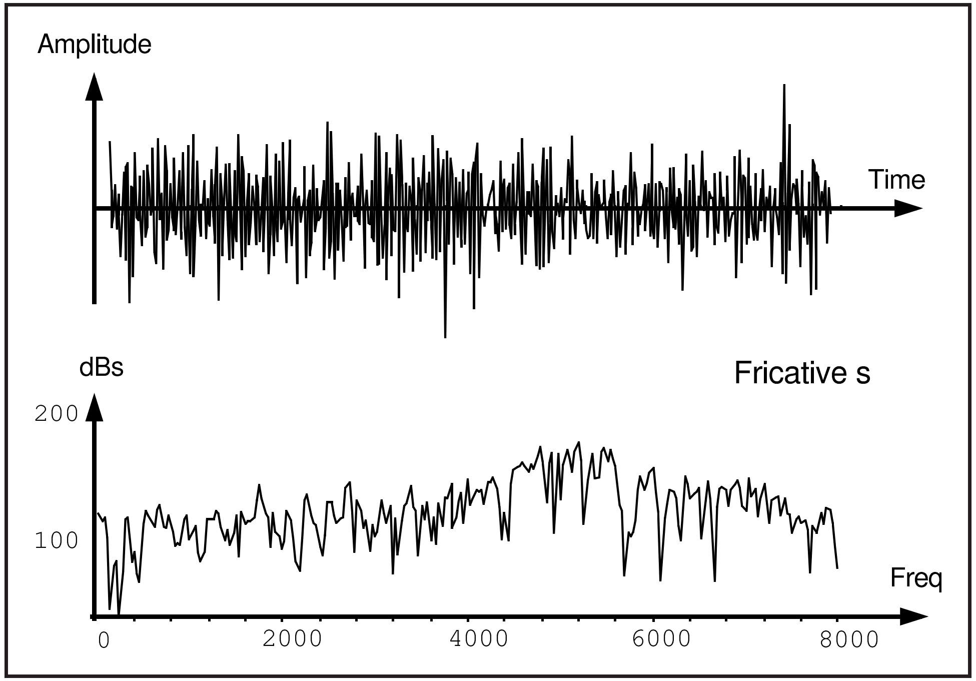
Spectral Properties of Speech

- Most information in speech is encoded by movement of the articulators (lips, tongue, jaw, etc) resulting in variations in the short term spectrum.

- Figure shows the waveform and Fourier magnitude spectrum of a fragment of the vowel iy as computed by the DFT/FFT. The waveform contains 512 points and it is 32msec in duration (16kHz sampling frequency).



- A time-domain segment and its spectrum for the fricative s is shown below.



The sampling theorem dictates that the highest present in the speech waveform must not exceed 8kHz to avoid aliasing. The spectra shown cover the full range from 0 to 8kHz. Both waveforms were sampled at 16kHz with 16 bits of precision. This is the norm for clean wide bandwidth speech processing. For telephone speech, 8-bit μ-law or A-law sampling at 8kHz is usually used.

- The vowel **time domain** waveform is approximately *periodic* with fundamental frequency around 130Hz. The periodic excitation is clearly visible in the spectrum as a high frequency ripple. There are about 7.5 cycles of this ripple per 1000Hz confirming the pitch frequency estimate from the time domain. In contrast, the time domain waveform for the fricative shows *no periodicity* and the spectrum has only random variations at much higher frequency.

- There is little information above 5kHz in the vowel **spectrum**. By contrast, the fricative spectrum has a broad single peak centred at about 5kHz. Thus, for high quality speech a bandwidth of 8kHz is just adequate. The bandwidth of telephony channels are limited to the range 300Hz to 3400Hz. This is just sufficient to avoid seriously impairing intelligibility but it does nevertheless reduce human ability to make distinctions between certain fricatives and stops.

Spectral Features of Sounds

Vowel sounds are characterised by the first 3 spectral peaks (**formants**). In the above spectrum of the vowel iy, the formant locations are at 250Hz, 2100Hz and 3300Hz. A low F1 and high F2 is typical of a high front vowel. There is a simple relationship between the tongue and jaw positions, and the values of F1/F2.

	Tongue Front	Tongue Back
High Jaw	F1 Low - F2 High	F1 Low - F2 Low
Low Jaw	F1 High - F2 High	F1 High - F2 Low

Liquids are characterised by formant position also but in this case the dynamics are important and the overall energy is lower than for vowels.

Nasals have a strong low 1st formant around 250Hz and weak higher formants. There is often energy around 2.5kHz.

Fricatives have most energy in higher frequencies. Voiced fricatives also show weak formant structure.

Stops are characterised by silence optionally followed by a burst of high energy.

DFT and Windowing

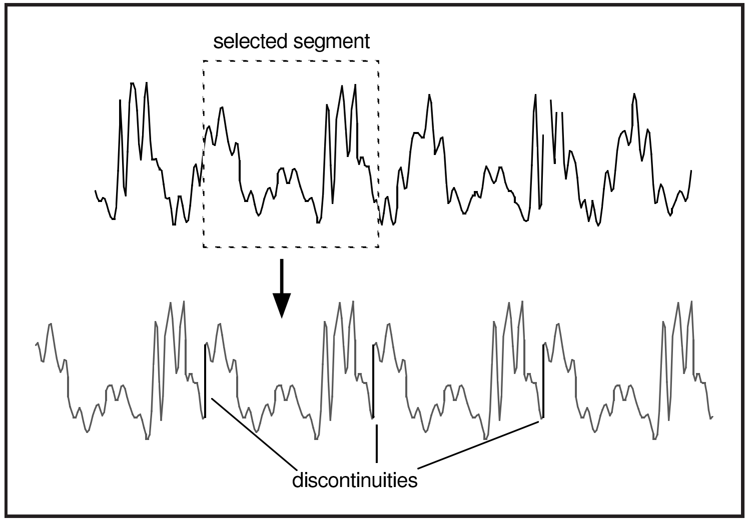
A frequency-discrete representation of the spectrum of a finite length signal $s_0 \rightarrow s_{N-1}$ is given by the Discrete Fourier Transform (DFT):

$$S(e^{j\frac{2\pi p}{N}}) = \sum_{n=0}^{N-1} s_n \, e^{-j\frac{2\pi np}{N}}$$

where the spectrum $\omega = 0 \rightarrow \pi/T$ has been divided into $N/2+1$ equally spaced discrete frequency points (including those at 0 and π/T) and the

$$\text{angular frequency of point } p = \frac{2\pi p}{NT}$$

Since the spectrum computed by the DFT of a finite segment of speech is that of a periodic wave formed by repeating the segment, discontinuities at the segment boundaries lead to unwanted artifacts.



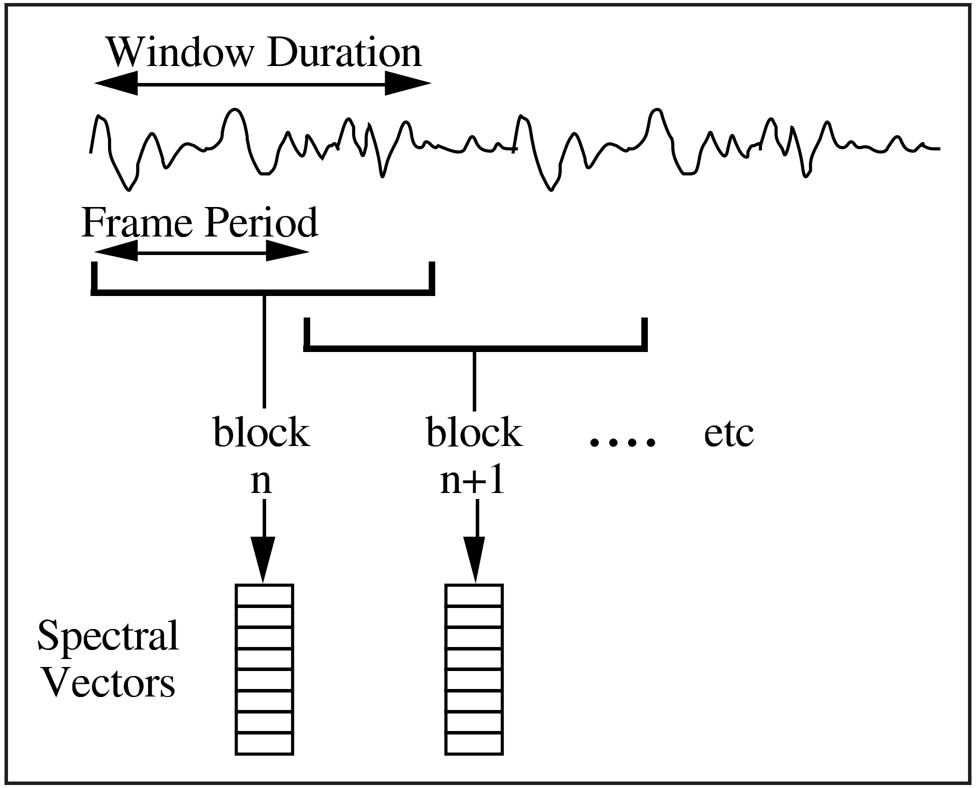
The commonly used solution is to multiply the speech segment by a tapered window. This reduces the discontinuities at the segment boundaries but, of course, it distorts the signal itself. A common choice is the Hamming window

$$w_n = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right)$$

where the DFT is now applied to the sequence $w_n s_n$ for $n = 0 \rightarrow N-1$.

Block Processing

For a complete waveform, a spectral estimate must be computed about every 10 msecs. Since this is rather a short duration to calculate a spectrum, analysis windows are allowed to overlap so that typically 25 msec windows are used.

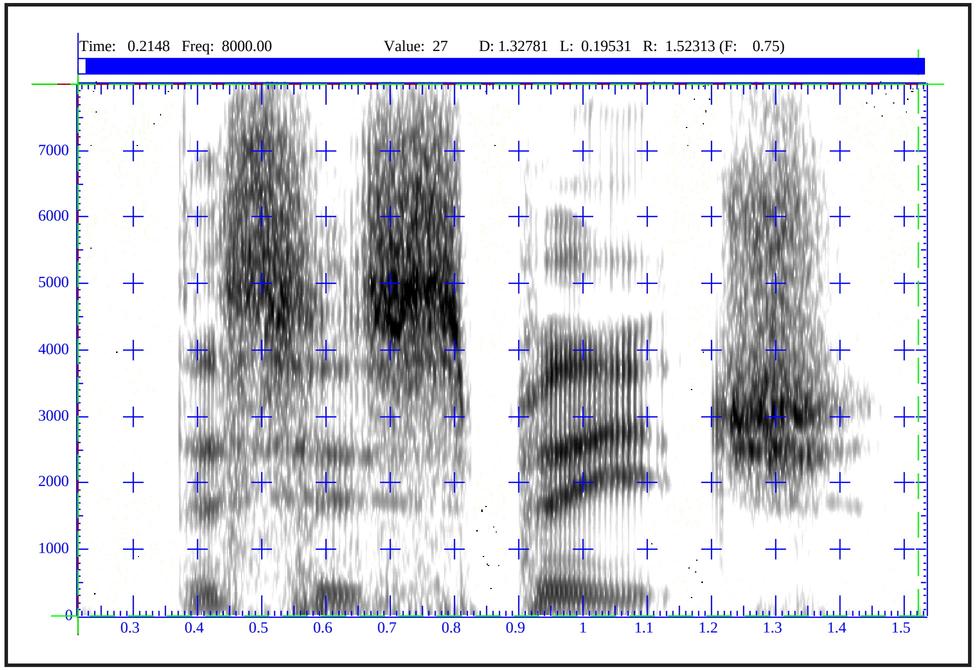


The above figure illustrates this *block processing* technique. Note that each segment of speech is often referred to as a *frame*.

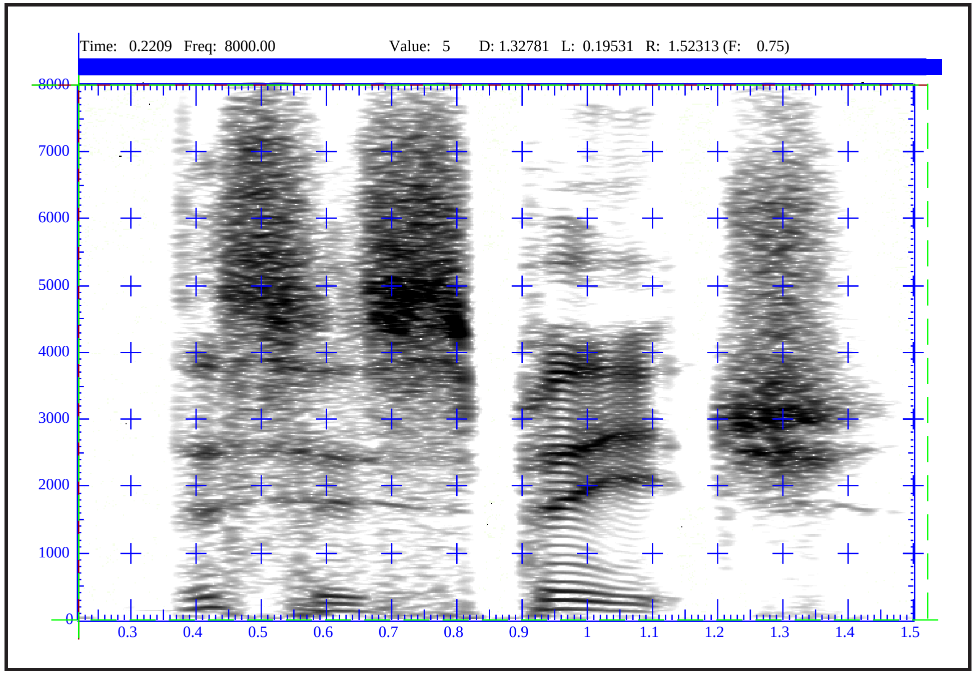
Spectrograms

The sequence of spectra derived by block processing continuous speech can be displayed as a grey-scale image with dimensions of time and frequency and with the spectral energy represented by the intensity of the image. This representation is called a **spectrogram**.

By using different length FFTs, the trade-off between *time-resolution* and *frequency resolution* can be investigated. Short-window (wide-band) and long-window (narrow-band) spectrograms are shown below.



A short analysis window gives good time but poor frequency resolution. In the wide-band spectrogram above, the pitch periods are visible whereas in narrow-band spectrogram below the harmonics of the fundamental frequency can be observed.



Linear Prediction Analysis

Linear prediction analysis of speech is historically one of the most important speech analysis techniques. The basis is the source-filter model. It assumes that a particular speech sample in a frame can be *predicted* as a *weighted-sum* of the previous p samples (typically in range $p = 10$ to $p = 15$) i.e.

$$\begin{aligned} \hat{s}_n &= a_1 s_{n-1} + a_2 s_{n-2} + \ldots + a_p s_{n-p} \\ \hat{s}_n &= \sum_{i=1}^{p} a_i s_{n-i} \end{aligned}$$

The prediction error for a particular sample e_n, is:

$$\begin{aligned} e_n &= s_n - \hat{s}_n \\ &= s_n - \sum_{i=1}^{p} a_i s_{n-i} \end{aligned}$$

Taking z-transforms of both sizes we obtain

$$E(z) = \left[1 - \sum_{i=1}^{p} a_i z^{-i}\right] S(z)$$

Then the filter transfer function from the prediction error sequence to the speech is

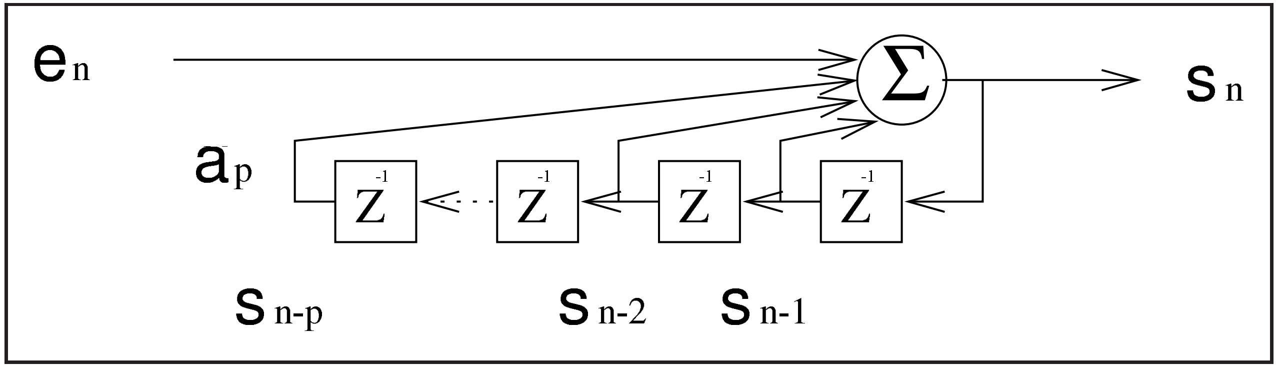
$$\begin{aligned} &= \frac{S(z)}{E(z)} \\ \frac{1}{A(z)} &= \frac{1}{1 - \sum_{i=1}^{p} a_i z^{-i}} \end{aligned}$$

Now in the source filter model we assume that the input to the filter is spectrally flat (e.g. from a **unit** impulse train). In that case we can write the transfer function of the filter as

$$
\begin{aligned}
H(z) &= \frac{G}{1 - \sum_{i=1}^{p} a_i z^{-i}} \\
&= \frac{G}{A(z)}
\end{aligned}
$$

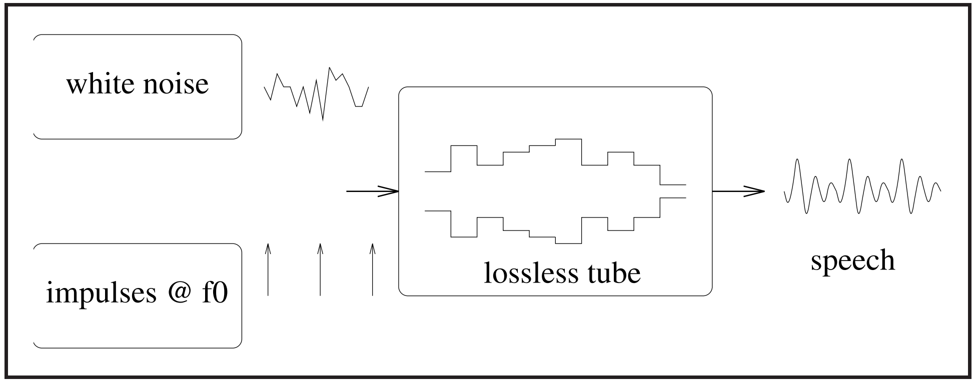
where G can be estimated to match the overall energy. The spectral shape of the the speech is given by $1/A(z)$ which, in fact, models the combined effect of the excitation source and the vocal tract transfer function.

Note that the linear prediction filter defined above is **all-pole**, i.e. it uses past samples of the output to compute the current sample and only the current sample of the input as shown below



Motivation from lossless tubes

Note that the transfer function of a lossless tube made up of sections of constant cross-sectional area can be described by an all-pole model. ... But



- Vocal tract is not built of cylinders of constant cross-sectional area
- Vocal tract is not lossless
- Vocal tract has a side passage (the nasal cavity)
- fricatives (e.g. /s/) are generated near the lips

Nevertheless, with sufficient parameters the LP model can make a reasonable approximation to the spectral envelope for all speech sounds (note that can approximate any spectral shape with enough poles!)

Parameter estimation

Given N samples of speech, we would like to compute estimates to a_i that result in the best fit. One reasonable way to define "best fit" is in terms of mean squared error [1] .

The summed squared prediction error E_T

$$\begin{aligned} E_T &= \sum_n e_n^2 \\ &= \sum_n (s_n - \sum_{k=1}^{p} a_k s_{n-k})^2 \end{aligned}$$

The minimum of E_T occurs when the derivative is zero with respect to each of the parameters, a_k. Note that we have not yet defined the range of the summation over n. To minimise E_T find the solution to $\partial E_T / \partial a_k = 0$.

Hence differentiating E_T with respect to a_j and setting equal to zero gives the

[1] These can also be regarded as "most probable" parameters if it is assumed the distribution of errors is Gaussian and a priori there were no restrictions on the values of a_i.

set of p equations:

$$\frac{\partial E}{\partial a_j} = 0 \quad = \quad -\sum_n \left(2(s_n - \sum_{k=1}^{p} a_k s_{n-k}) s_{n-j} \right)$$

$$= \quad -2 \sum_n s_n s_{n-j} + 2 \sum_n \sum_{k=1}^{p} a_k s_{n-k} s_{n-j}$$

Rearranging gives the set of p simultaneous linear equations $($normal equations)for values of j from 1 to p:

$$\sum_n s_n s_{n-j} \quad = \quad \sum_{k=1}^{p} a_k \sum_n s_{n-k} s_{n-j} \qquad (1)$$

So far the range of the summations over n was from $-\infty$ to $+\infty$, which is undesirable for speech processing. Different possibilities give rise to variants of linear prediction analysis. The two most widely used are:

1. **covariance method**
   the limits are $n = 0$ to $n = N - 1$

2. **autocorrelation method**
   the limits are $\pm\infty$ and hence the speech requires windowing to select the portion of data to analyse. We will discuss this option further below since it is widely used.

Autocorrelation method

In the autocorrelation method the summation is taken over all samples. Thus

$$\sum_{n=-\infty}^{\infty} s_{n-i}s_{n-j} = \sum_{n=-\infty}^{\infty} s_n s_{n+i-j}$$

This is the autocorrelation sequence r_{i-j}.

Therefore if we process a windowed version of the data in which

$$s_n = 0 \ \text{ if } \ n < 0 \ \text{ or } \ n >= N$$

we also have

$$r_k = \sum_{n=0}^{N-1-k} s_n s_{n+k}$$

Now the normal equations are as shown below (a Toeplitz matrix):

$$\begin{pmatrix} r_1 \\ r_2 \\ \ldots \\ r_p \end{pmatrix} = \begin{pmatrix} r_0 & r_1 & r_2 & \cdots & r_{p-1} \\ r_1 & r_0 & r_1 & \cdots & r_{p-2} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ r_{p-1} & r_{p-2} & \cdots & \cdots & r_0 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \ldots \\ a_p \end{pmatrix}$$

Due to the special structure of the above matrix, efficient solution methods exist, one of which is **Durbin's algorithm**.

Autocorrelation Method: Durbin's Algorithm

Denoting the values of the LP parameters at iteration i by $a_k^{(i)}$ and the sum-squared predictor error (or residual energy) by $E_T^{(i)}$ ($E_T^{(0)} = r_0$) for i = 1, 2, ...

$$\begin{aligned}
k_i &= \left(r_i - \sum_{j=1}^{i-1} a_j^{(i-1)} r_{i-j} \right) / E_T^{(i-1)} \\
a_i^{(i)} &= k_i \\
a_j^{(i)} &= a_j^{(i-1)} - k_i a_{i-j}^{(i-1)} \qquad 1 \le j < i \\
E_T^{(i)} &= (1 - k_i^2) E_T^{(i-1)}
\end{aligned}$$

Example of Durbin's Algorithm (1 of 2)

For an example waveform the first auto-correlation coefs are:

$$\begin{aligned} r_0 &= 2.4470\ 10^8 \\ r_1 &= 2.2466\ 10^8 \\ r_2 &= 1.7823\ 10^8 \end{aligned}$$

Therefore on the first iteration:

$$\begin{aligned} k_1 &= r_1/E_T^{(0)} = 0.9181 \\ a_1^{(1)} &= k_1 \\ &= 0.9181 \\ E_T^{(1)} &= (1 - k_1^2)E_T^{(0)} \\ &= (1 - 0.9181 * 0.9181)2.4470\ 10^8 \\ &= 0.38442\ 10^8 \end{aligned}$$

Example of Durbin's Algorithm (2 of 2)

And on the second iteration:

$$
\begin{aligned}
k_2 &= \left(r_2 - a_1^{(1)} r_1\right) / E_T^{(1)} \\
&= \left(1.7823\ 10^8 - 0.9181 * 2.2466\ 10^8\right) / 0.38442\ 10^8 \\
&= -0.72915 \\
a_2^{(2)} &= k_2 = -0.72915 \\
a_1^{(2)} &= a_1^{(1)} - k_2 a_1^{(1)} \\
&= 0.9181 - -0.729150.9181 \\
&= 1.58753 \\
E_T^{(2)} &= (1 - k_2^2) E_T^{(1)} \\
&= (1 - -0.72915 * -0.72915) 0.38442\ 10^8 \\
&= 0.18004\ 10^8
\end{aligned}
$$

Properties of Durbin's Algorithm

The parameters k_i are known as the reflection coefficients and are always in the range ± 1 (and are used in an acoustic tube model of the vocal tract).

Note that:

- As Durbin's algorithm proceeds, all intermediate order predictors are calculated
- This method also provides the reflection coefficients and the error energies of all intermediate order predictors
- The resulting filter is guaranteed to be stable (which is useful for synthesis/coding!)
- The value of the squared prediction residual, $E_T^{(i)}$ is also computed and is guaranteed to decrease (or remain constant) on each iteration
- Compared to the covariance method substantially larger windows are required
- Most commonly used

Spectral Interpretation

The poles of the LP filter are either real (which give a spectral tilt) or occur in complex-conjugate pairs and model resonances. If all poles are complex, filter can be viewed as a cascade of $p/2$ 2-pole resonators.

The autocorrelation method also has a spectral interpretation. By Parseval's Theorem

$$\begin{aligned}
E_T &= \frac{1}{2\pi}\int_{-\pi}^{\pi} \left|E(e^{j\omega T})\right|^2 \, d\omega T \\
&= \frac{1}{2\pi}\int_{-\pi}^{\pi} \left|S(e^{j\omega T})\right|^2 \left|A(e^{j\omega T})\right|^2 \, d\omega T \\
&= \frac{1}{2\pi}\int_{-\pi}^{\pi} P(e^{j\omega T}) \left|A(e^{j\omega T})\right|^2 \, d\omega T
\end{aligned}$$

where $P(e^{j\omega T})$ is the speech power spectrum.

Now the LP model approximation of the speech power spectrum can be written as

$$\hat{P}(e^{j\omega T}) = \left|\frac{G}{A(e^{j\omega T})}\right|^2$$

and hence

$$E_T = \frac{G^2}{2\pi}\int_{-\pi}^{\pi} \frac{P(e^{j\omega T})}{\hat{P}(e^{j\omega T})}\, d\omega T$$

Thus minimising E_T minimises this integrated ratio with the constraint that $\int_{-\pi}^{\pi} P(e^{j\omega T})\, d\omega T = \int_{-\pi}^{\pi} \hat{P}(e^{j\omega T})\, d\omega T$.

Hence linear prediction can be viewed as power spectrum matching. In fact, starting from the frequency domain interpretation of the error, the autocorrelation equations can be derived in the frequency domain.

Pre-emphasis

- The LP filter so far presented attempts to fit an all-pole model using a least-squares measure.
- The lower formants contain more energy and therefore are modelled more accurately than the higher frequency formants
- A simple pre-emphasis filter,

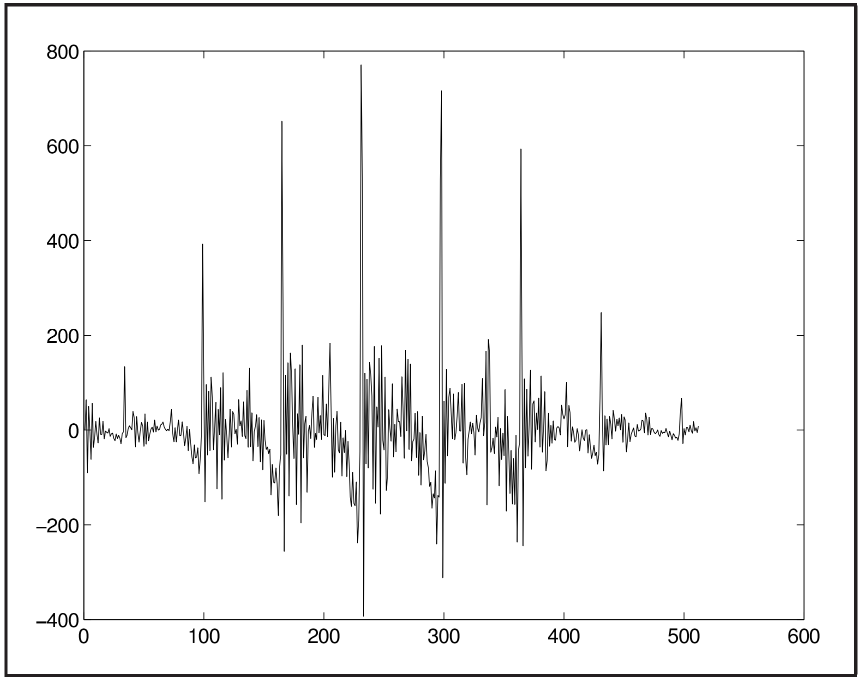
$$s'_n \quad = \quad s_n - a_1 s_{n-1} \tag{2}$$

  is often used to boost the higher frequencies. Typically $0.96 \leq a_1 \leq 0.99$, or the optimal pre-emphasis $a_1 = r_1/r_0$ is used.
- If reconstructing the speech the inverse of the pre-emphasis filter should be used:

$$s_n \quad = \quad s'_n + a_1 s_{n-1} \tag{3}$$

The residual signal

Plotting the error signal e_n for the example waveform



Finding the error signal or prediction residual is known as *inverse filtering*.

- Most short term correlations seem to be lost in the error signal
- The residual contains long term correlations due to pitch pulses
- There is a spike at the pitch periods when prediction is poor

LP Spectrum

The frequency response of the LP filter is one way to estimate the speech spectrum. Since the detailed excitation is mainly modelled by the error signal, the LP filter frequency response is a smoothed all-pole approximation to the speech spectrum. As such it is of use in calculating e.g. formant frequencies.

Recall that the filter can be represented as

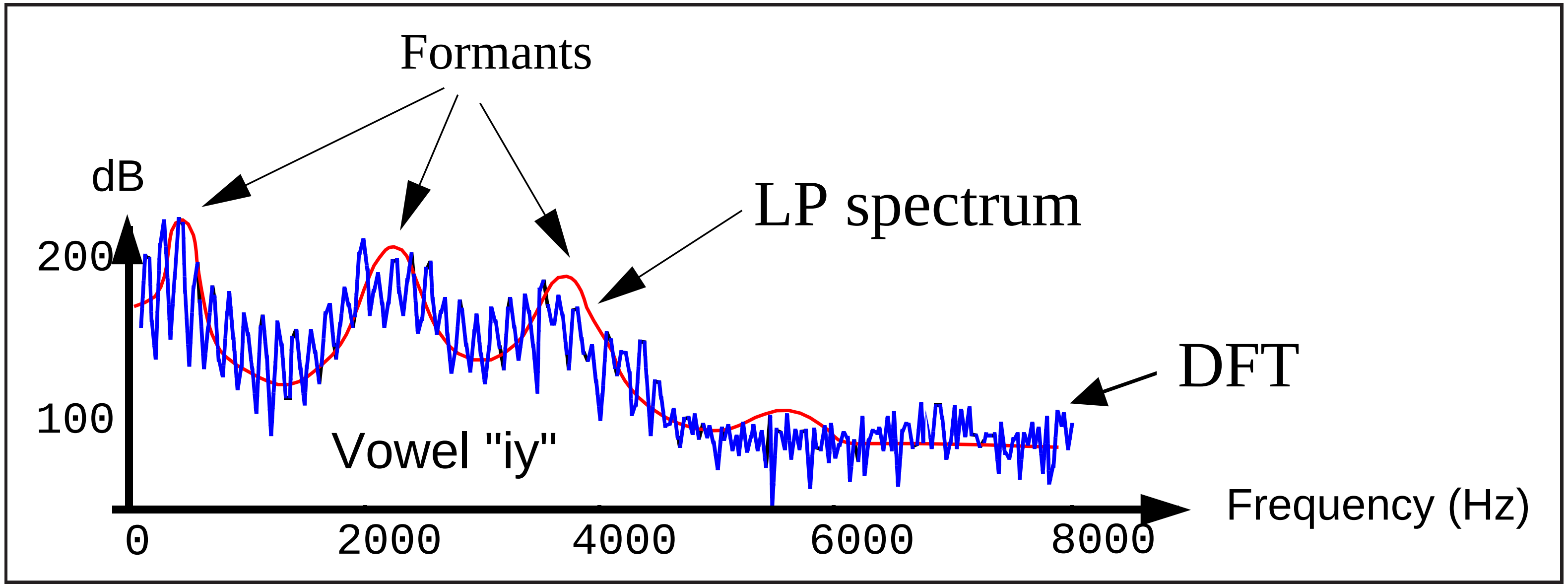
$$H(z) = \frac{G}{A(z)}$$

Then the LP filter frequency response is found by setting $z = e^{j\omega T}$. The resulting LP spectrum is smooth and shows formant structure.

- Poles near the unit circle produce peaks
- There are no zeros
- Typically formants are very sharp

LP Spectrum Example

An example of the LP spectrum is shown below. It can be seen that relative to the DFT spectrum the effect of the excitation has been removed and a smooth representation of the vocal tract frequency response remains.



Summary of LP Analysis

- All-pole filter model of speech
- Filter associated with vocal tract
- Error signal associated with excitation
- Filter parameters estimated to minimise sum-squared prediction error
- Autocorrelation method uses Durbin's algorithm for efficient solution
- Typical order of analysis is 10-15 (about 2 poles per formant, plus others for spectral shape)
- Often LP analysis is performed after **pre-emphasis** to flatten spectrum
- Autocorrelation analysis yields reflection coefficients with links to lossless acoustic tube model
- LP filter frequency response is the"LP spectrum"

Cepstral Analysis

Cepstral analysis which is another way (apart from LP analysis) to separate the vocal tract frequency response from the excitation and can also obtain a smooth representation of the vocal tract frequency response.

We will discuss

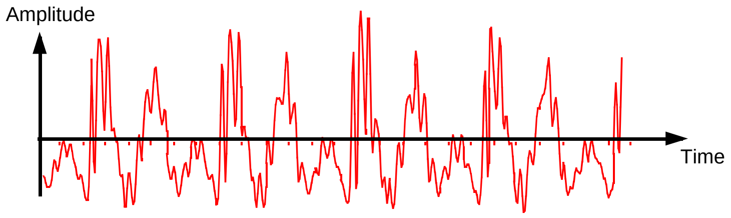
- Homomorphic filtering
- The cepstrum
- Applications of cepstral analysis
- Mel-scale filterbanks
- Discrete cosine transform

The final part leads to Mel-scale cepstral coefficients which are an important representation of the speech signal used in speech recognition.

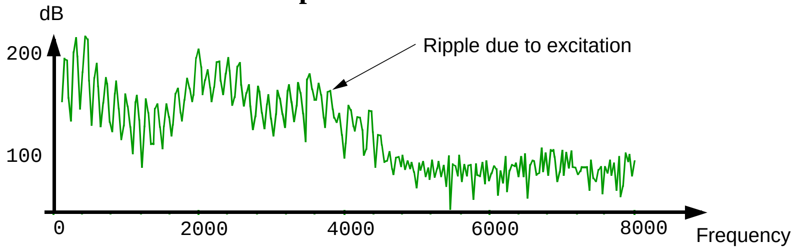
Aim of Cepstral Analysis

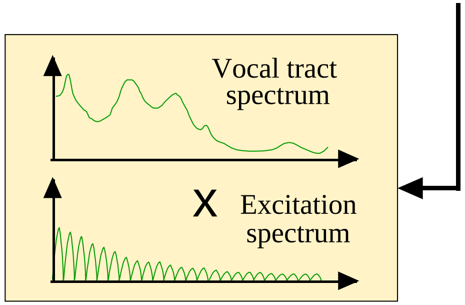
The source-filter model regards the spectrum as the product of the excitation spectrum and the vocal tract frequency response. We aim to separate these

Time−domain speech signal: the vowel "iy"



Spectrum





Separation of

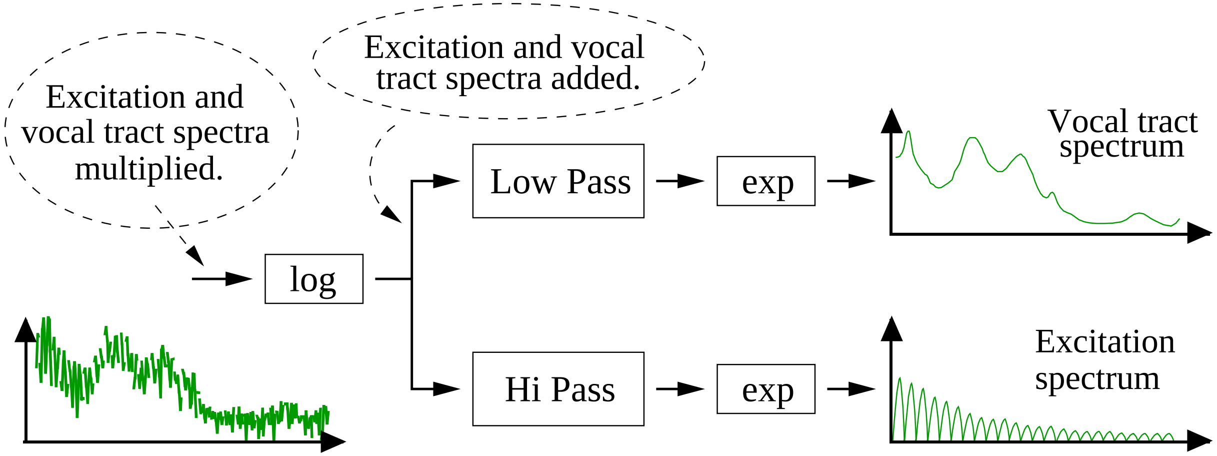
- excitation
- vocal tract frequency response
- other filtering effects

Homomorphic Filtering

The excitation gives rise to a quickly varying ripple in the spectrum. If the vocal tract frequency response and excitation were **added** then the signals could be separated. However the signals have been **multiplied**!

The solution is to take logs to convert multiplication to addition:

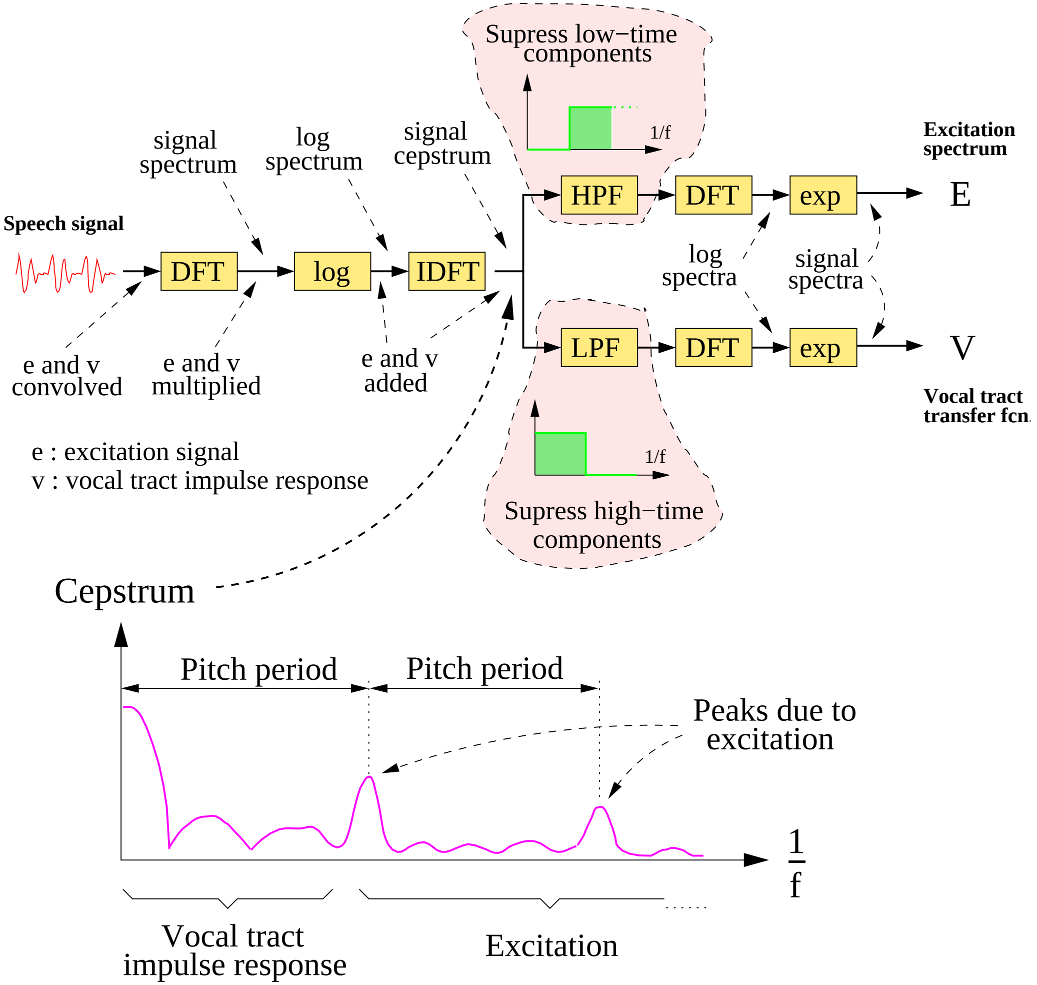
$$\begin{aligned} \log(a \cdot b) &= \log a + \log b \\ e^{\log y} &= y \end{aligned}$$


Excitation and vocal tract spectra multiplied.
Excitation and vocal tract spectra added.
log
Low Pass
Hi Pass
exp
exp
Vocal tract spectrum
Excitation spectrum

This approach is called **homomorphic filtering**. We are filtering the log spectrum as we would normally filter in the time domain.

The Cepstrum

Homomorphic filtering usually employs the DFT. Note that for the real cepstrum the log is applied to the magnitude spectrum.

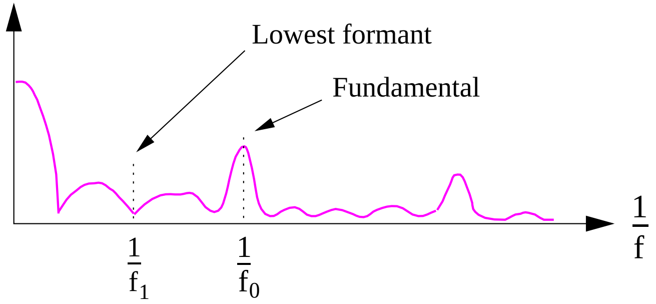


In the cepstrum vocal tract impulse response decays rapidly and can be separated (by windowing) from the excitation.

The cepstrum is computed in **quefrency** domain and filtering in this domain is called **liftering**. Note that taking the IDFT of the cepstrum doesn't return to the time domain because of the non-linear log operation.

If have lowest formant at frequency f_1 and the fundamental (pitch) at f_0 then cepstral coefs:

- $c_0 \rightarrow c_h$ encodes vocal tract response if $h \geq \frac{1}{f_1 T}$
- $c_p \rightarrow c_{N-1}$ include the major pitch peak if $p \geq \frac{1}{f_0 T}$



- If $f_0 < 250Hz$ and $f_1 > 500Hz$, and $T = 62.5\mu s$ then $h = 32$, $p = 64$

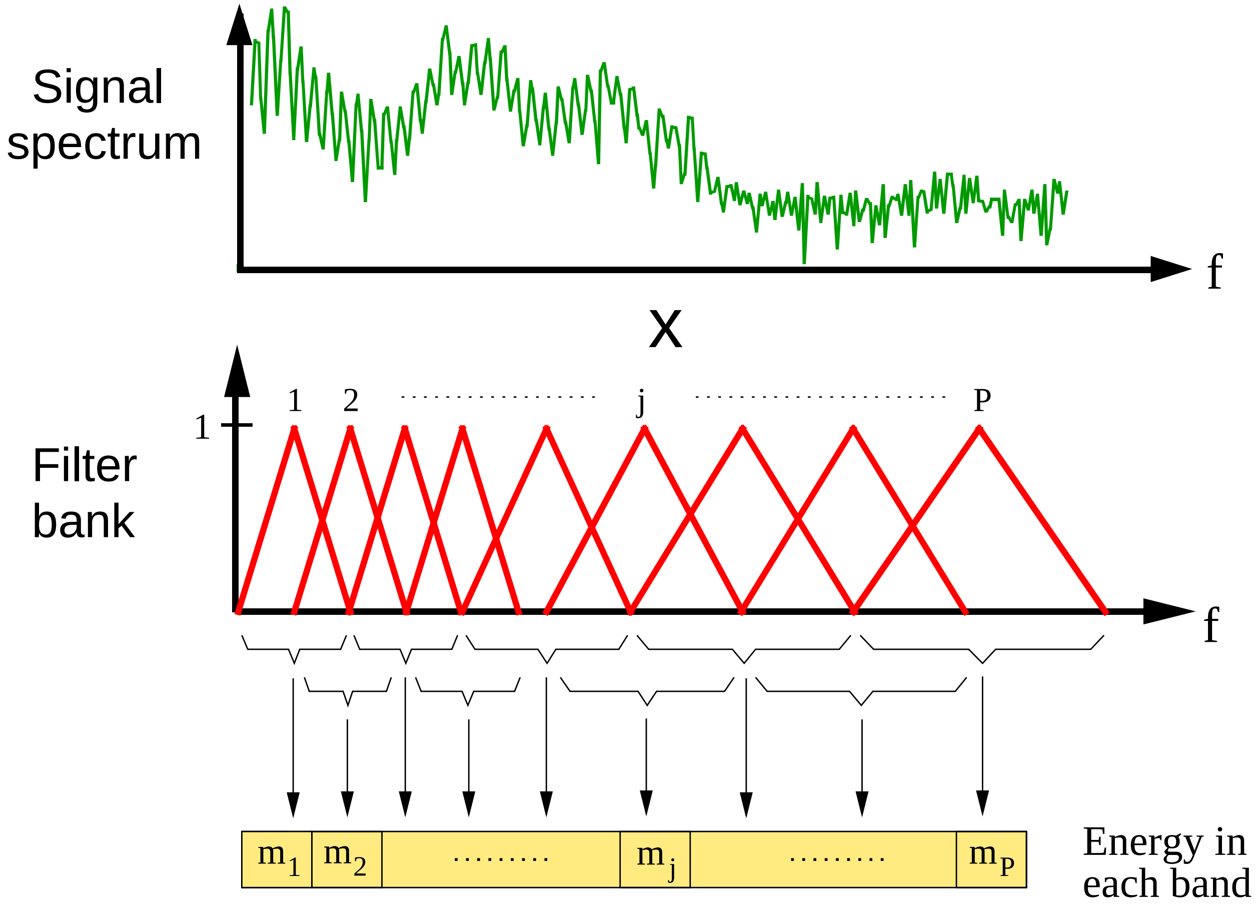
Making h smaller increases the smoothing over the whole spectrum.

Applications of Cepstral Analysis

- **Pitch estimation**
  Find the peak cepstral value in the range $c_p \rightarrow c_{N-1}$.
  If the peak is at c_n then fundamental frequency is $\frac{1}{nT}$
- **Smoothed spectrum**
  Take a DFT of (zero-padded) $c_1 \rightarrow c_h$. This is an alternative to the LP spectrum.
- **Vocoding**
  Find fundamental frequency and transmit speech as a sequence of frames $\{n, c_0, c_1, \cdots, c_h\}$.
- **Recognition**
  Use $c_1 \rightarrow c_h$ as a representation of the (smoothed) spectrum in recognition. In fact (forms of) cepstral parameters are the standard representation used in current speech recognition systems. These often used a non-linear frequency scale that roughly corresponds to the frequency resolution of the ear.

Mel-Scale Filterbanks

Reduce frequency resolution and analysis to model ears spectral resolution.



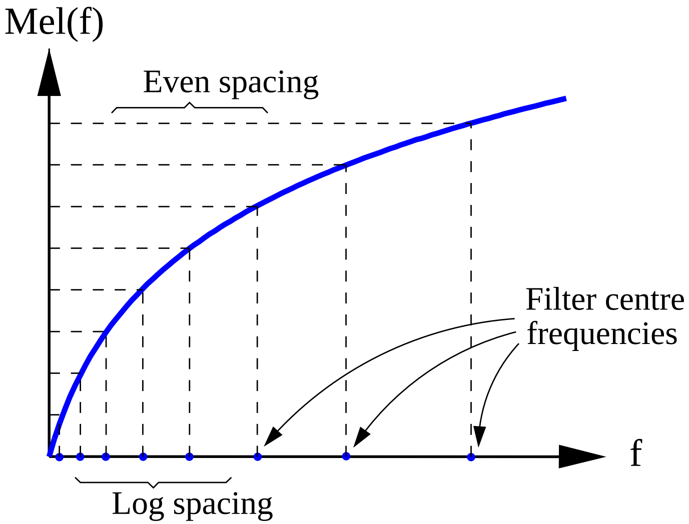
The energy in each frequency band is computed from the DFT.

The spacing of the center frequencies is based on the **Mel-scale**.

The Mel-scale is defined as

$$\mathrm{Mel}(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right)$$

This frequency scale is shown below:



The scale is often regarded as being approximately linear up to 1kHz and logarithmic thereafter.

Discrete Cosine Transform

Cepstral coefficients can be derived from the Mel filterbank energies using a simplified version of the DFT known as the discrete cosine transform (DCT). This uses the fact that the log magnitude spectrum is real-valued, symmetric with respect to 0 and periodic in frequency.

$$c_n = \sqrt{\frac{2}{P}} \sum_{i=1}^{P} m_i \cos\left[\frac{n(i-\frac{1}{2})\pi}{P}\right]$$

where P is the number of filterbank channels.

The representation found in this way is known as **Mel-frequency cepstral coefficients** (or **MFCCs**).

- The DCT decorrelates the spectral coefficients and allows them to be modelled with diagonal Gaussian distributions
- The number of parameters needed to represent a frame of speech is reduced. This in turn reduces memory and computation requirements.
- Note that c_0 is a measure of the signal energy

Other Cepstral Representations

There are a number of alternatives to computing cepstral representations which are used in speech recognition systems. These include using a cepstral representation of linear prediction coefficients.

This can be computed using the LP spectrum but there is also a direct and efficient method to obtain these parameters from the predictor coefficients.

Recall the LP transfer function:

$$H(z) = \frac{G}{1 - \sum_{i=1}^{p} a_i z^{-i}}$$

The p cepstral coefficients can be computed using the recursion:

$$\begin{aligned} c_0 &= \log G^2 \\ c_n &= a_n + \frac{1}{n}\sum_{j=1}^{n-1} j c_j a_{n-j} \qquad 0 < n \leq p \end{aligned}$$

There are also several alternatives of cepstral represntations that use a non-linear frequency scale. These include using a type of linear prediction analysis termed **perceptual linear prediction** or (**PLP**).

First the power spectrum is computed on a non-linear frequency scale (Bark scale, similar to Mel scale). This is then generally compressed (e.g. with a power-law compression) and other compensation for the frequency sensitivity of human hearing is applied.

The autocorrelation coefficients can be obtained as the inverse DFT of the power spectrum. Autocorrelation analysis LPC can be computed using Durbin's algorithm, and from there cepstral coefficients obtained.

Cepstral Analysis Summary

- Cepstral analysis is a method to separate
  - the excitation
  - the vocal tract frequency response
  - other filtering effects
- The IDFT of the log spectrum is termed the cepstrum
- Can use the cepstral analysis for
  - pitch estimation
  - vocal-tract frequency response estimation
  - vocoding
  - speech recognition
- Most current speech recognition systems use a form of cepstral analysis to represent speech. Generally compute these use a non-linear filterbank modelled to roughly match human auditory perception