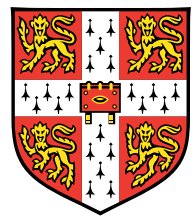
MPhil in Machine Learning and Machine Intelligence

Module MLMI2: Speech Recognition

L2: Speech Signal Processing—Spectral Analysis

Phil Woodland
pcw@eng.cam.ac.uk

Michaelmas 2021



Cambridge University Engineering Department

Outline

Frequency Domain Analysis

- Intro to Fourier analysis and the frequency domain
- Spectrum via correlation with sine/cosine
- Discrete Fourier Transform (DFT)
- Windowing
- Fast Fourier Transform
- Spectral Properties of Speech Sounds
- The Spectrogram

Other Spectral Estimation Methods

- Source filter model and linear prediction
- Power spectrum from linear prediction model
- Mel-frequency filterbanks and cepstral representations

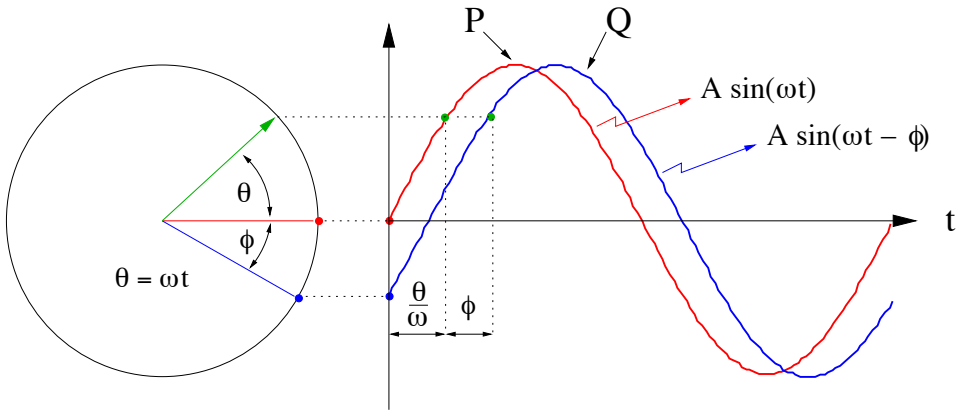
Many books cover the material in a more mathematical fashion (or assume more background) than the current lecture. Relevant books include:

- Paul Taylor, *Text-to-Speech Synthesis*
- Gold & Morgan, *Speech and Audio Signal Processing*
- Huang, Acero & Hon, *Spoken Language Processing*

Sinusoids and Phase

General sinusoid: $A\sin(\omega t + \phi)$.
Described by

- Amplitude (A) (height of peak, length of rotating vector)
- Phase (ϕ)
- Frequency f number of cycles per second (Hertz) $f = 1/\tau$
  (or angular frequency $\omega = 2\pi f$)



Arguments described in radians.

P leads Q by ϕ radians of phase difference.

- When $\phi = 0$ we use the term **sine wave** and denote it by

$$A\sin(\omega t)$$

- When $\phi = \pi/2$, we have a **cosine wave**

$$A\sin(\omega t + \pi/2) = A\cos(\omega t)$$

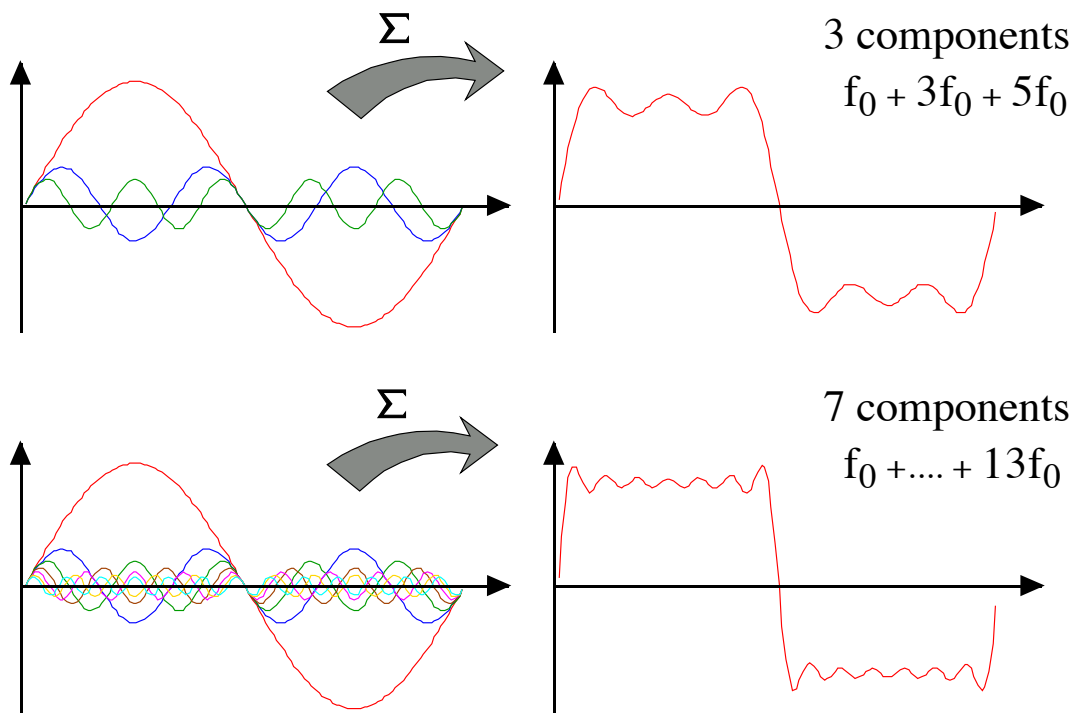
Fourier Analysis

Theorem: Any periodic signal of frequency f_0 can be constructed exactly by adding together sinusoids with frequencies

$$f_0, 2f_0, 3f_0, 4f_0, 5f_0, \ldots$$

each with the appropriate amplitude and phase. f_0 is called the **fundamental frequency** and $2f_0, 3f_0$ etc are the **harmonics**.

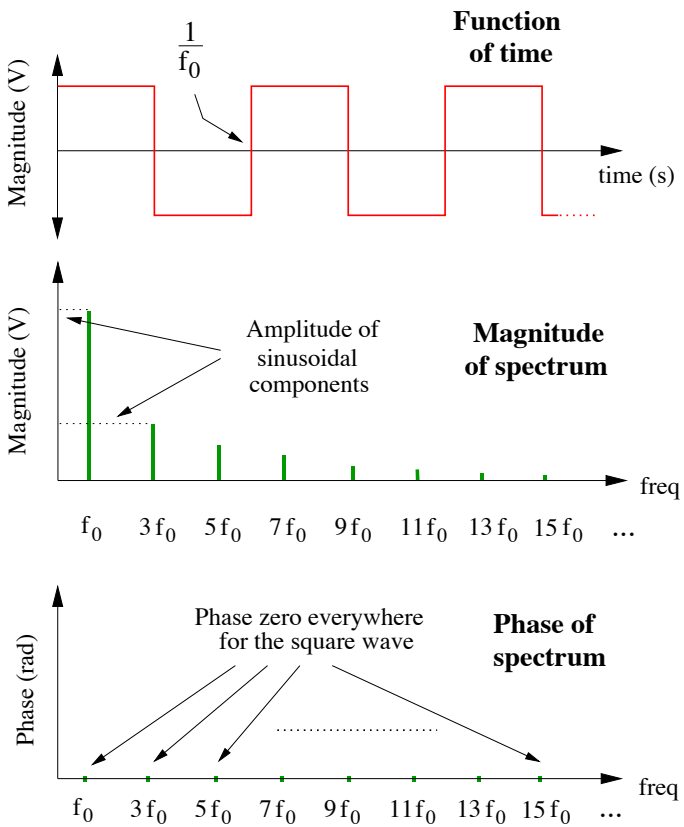
Example: a square wave (odd harmonics only)



Any periodic function can be characterised by the amplitude and phase of its sinusoidal components. This is the **spectrum**.

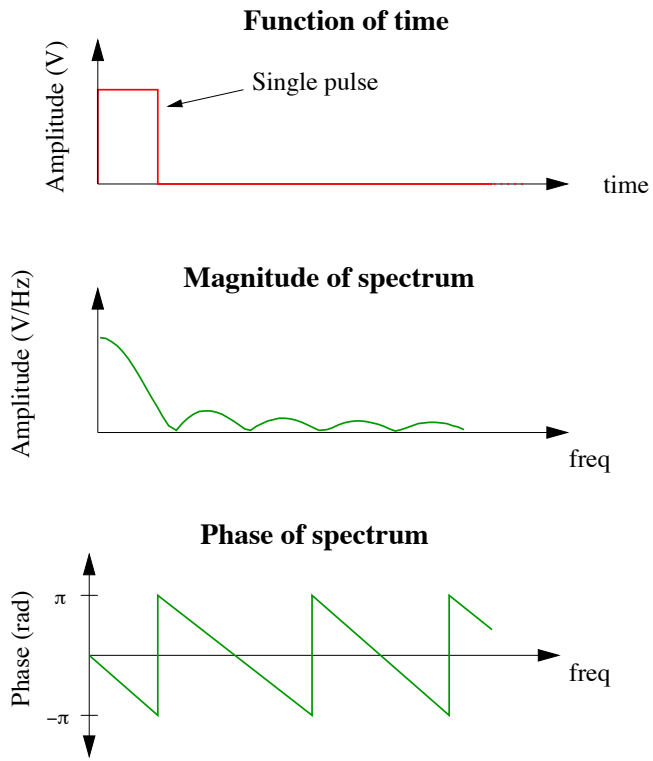
Periodic signals have spectral components at integer multiples of the fundamental frequency.

Example: a square wave and its spectrum



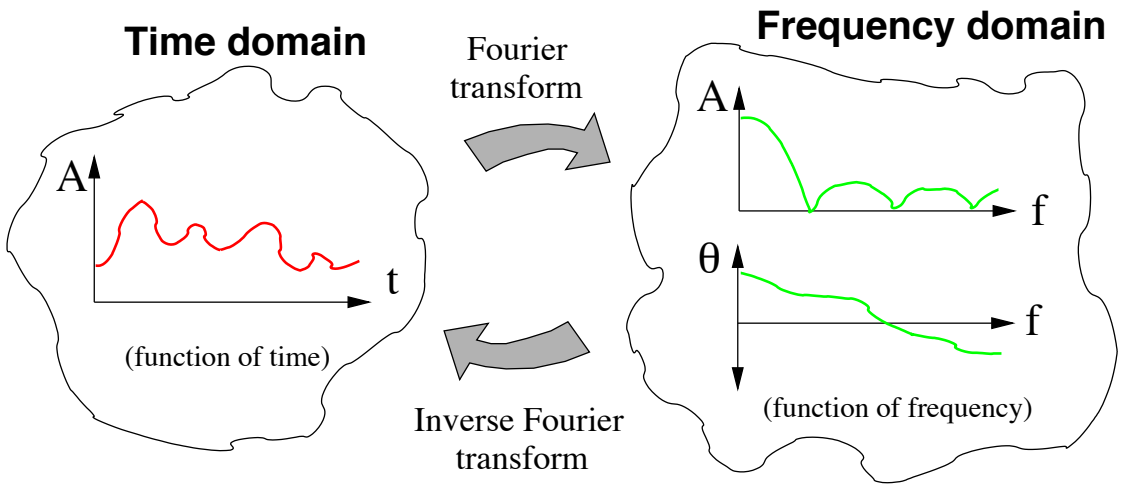
Aperiodic and stochastic signals have spectra that are **continuous** functions of frequency i.e. at all possible frequencies.

Example: An aperiodic signal and its spectrum



The Fourier Transform

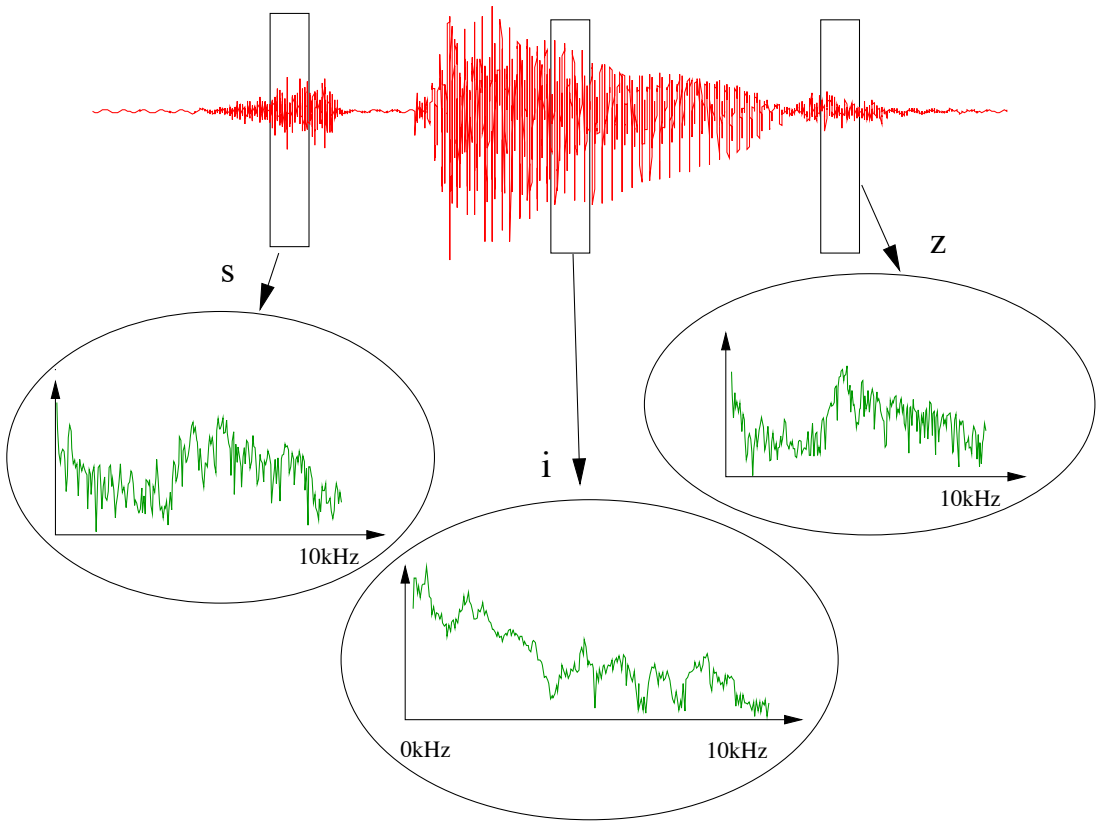
- The spectrum is obtained from the time signal by means of the **Fourier Transform**
- The magnitude and phase of every sinusoidal component is given by the spectrum
- The time signal is recovered from (complex) spectrum by the **Inverse Fourier Transform**
- The FT allows us to move between the **time domain** and the **frequency domain**



- The spectrum and the time-domain signal contain the same information expressed differently
- The frequency domain representation is particularly relevant for characterising speech sounds (cochlea performs a frequency-based analysis)

Application of FT to a Speech Signal

- Divide speech into 10 ms segments
- Find the magnitude spectrum of each segment
- For the word “skills”



- Spectral “envelope” is most important for describing speech sounds
- Spectrograms can be used for visualising a sequence of spectra

Correlation with Cosines

Correlation between two discrete-time signals x and y with elements x_i,y_i over an N point interval $[0, \ldots, N-1]$

$$q = \sum_{i=0}^{N-1} x_i . y_i$$

Compute correlation between two cosine waves: signal cosine with angular frequenciy ω_s and a test cosine with angular frequency ω_t. It can be shown that

$$\boxed{q = \alpha A \text{ if } \omega_s = \omega_t \text{ and zero otherwise}}$$

Now consider the signal to be a sum of many component cosine waves:

- The correlation with a cosine yields a method to extract the cosine components of an arbitrary signal.
- The amplitude of each test cosine frequency may be measured by correlation. However, it must be ensured that the correlation is computed over an integer number of cycles

Set test frequencies to ensure p cycles of test cosine in the summation

$$\omega_t = \frac{2\pi p}{NT} \qquad p = 0, 1, \ldots, N-1 \qquad \text{or} \qquad f_t = \frac{p}{NT} \qquad p = 0, 1, \ldots, N-1$$

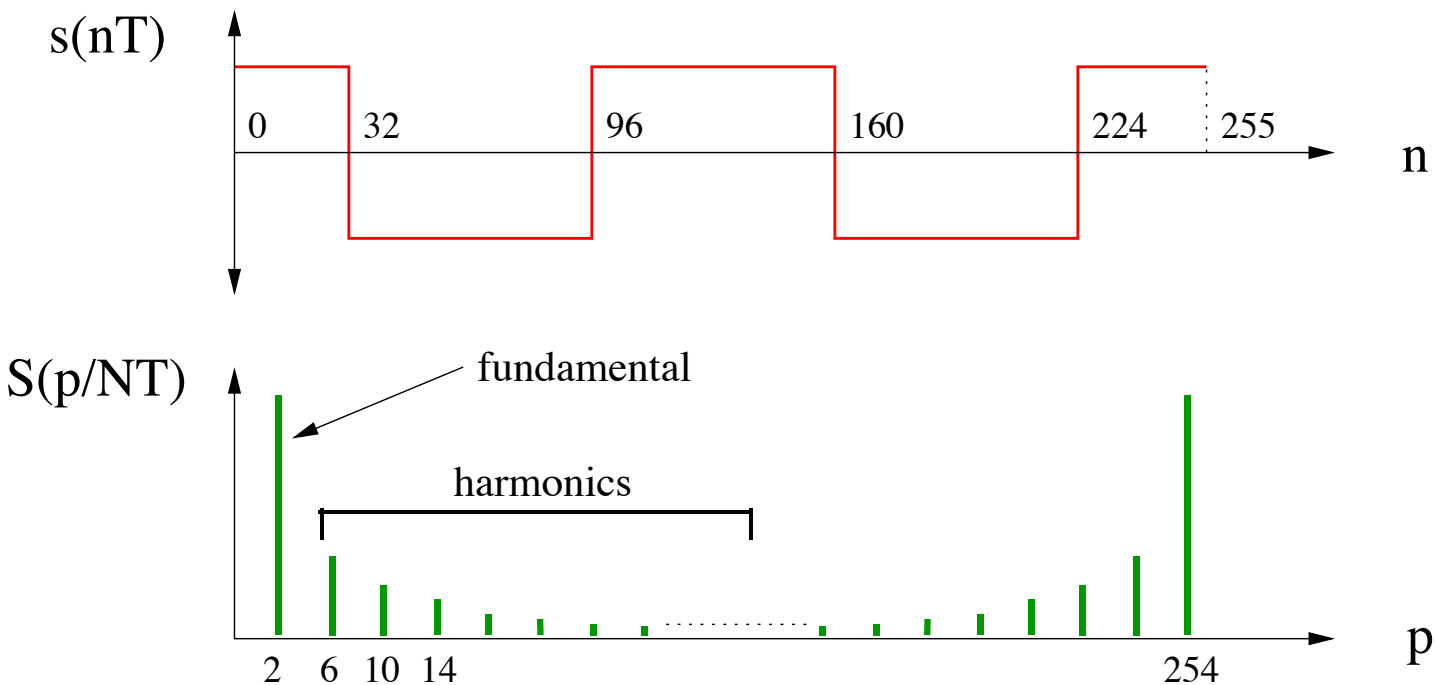
where T is the sampling period.

Apply cosine correlation to arbitrary N point signal: find amplitude of each cosine component

$$S_p = \sum_{n=0}^{N-1} s_n \cos\left(\frac{2\pi np}{N}\right) \qquad p = 0 \ldots N-1$$

Note that $\frac{p}{NT}$ is the actual frequency (in Hz) of each test frequency.

Example: square wave

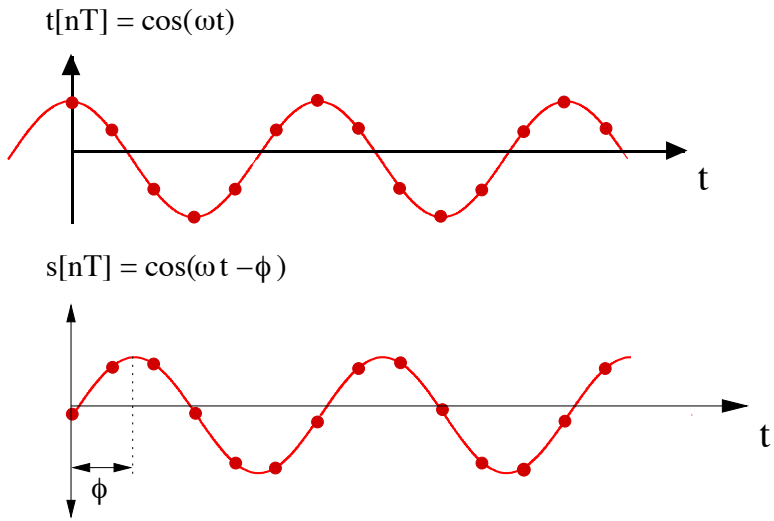


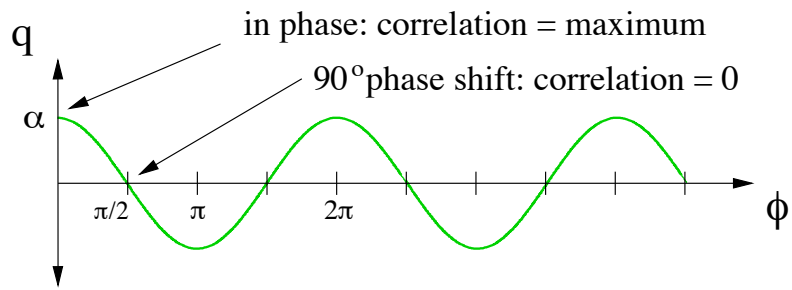
Cosine correlation will now be extended to full spectral analysis, including phase.

Phase in Spectral Analysis

Finding the frequency components of a square wave works using only cosine correlation because all components have **zero phase**

Correlation of a test cosine $t[nT]$ with a delayed signal cosine $s[nT]$ of the same frequency:





The correlation is also a cosine: correlation with a test cosine is unable to detect a $\pi/2(90^o)$ phase shift.

Need to also compute the **sine correlation** (out of phase by by $\pi/2$ with cosine correlation of same frequency). The combination allows the computation of both the magnitude and phase of the each frequency component.

Amplitude and Phase

Correlations with sine and cosine

$$\begin{aligned} \text{cosinecorrelation} &= \alpha \cos \phi \\ \text{sinecorrelation} &= \alpha \sin \phi \end{aligned}$$

and

$$[\alpha \cos \phi]^2 + [\alpha \sin \phi]^2 = \alpha^2$$

- By adding the squares of the sine and cosine correlation and taking the square root, obtain desired amplitude of the sinusoid component at frequency ω independent of the phase ϕ
- The phase of this component is given by

$$\tan \phi = \frac{\alpha \sin \phi}{\alpha \cos \phi} = \frac{\text{sinecorrelation}}{\text{cosinecorrelation}}$$

We have now found a general method for calculating the magnitude and phase of the spectrum.

General Spectral Analysis Algorithm

Any periodic signal can be expressed as the sum of a fundamental sinusoid and its harmonics.

$$s(nT) = a_0 + a_1 \cos(\omega nT + \phi_1) + a_2 \cos(2\omega nT + \phi_2) + \ldots$$

Individual components at a frequency $\Omega = p\omega$ can be found by correlating $s(nT)$ with $\cos(\Omega nT)$ and $\sin(\Omega nT)$. Let $c(\Omega)$ be the cosine correlation and $s(\Omega)$ the sine correlation with signal

$$\begin{aligned} c(\Omega) &= \sum_{n=0}^{N-1} s(nT) \cos\left(\frac{2\pi np}{N}\right) \qquad p = 0, 1, \ldots, N-1 \\ s(\Omega) &= \sum_{n=0}^{N-1} s(nT) \sin\left(\frac{2\pi np}{N}\right) \qquad p = 0, 1, \ldots, N-1 \end{aligned}$$

then

$$\begin{aligned} a_p &= \sqrt{\left[c^2(\Omega) + s^2(\Omega)\right]} \\ \phi_p &= \tan^{-1}\left[\frac{s(\Omega)}{c(\Omega)}\right] \end{aligned}$$

This is the **Discrete Fourier Transform (DFT)**.

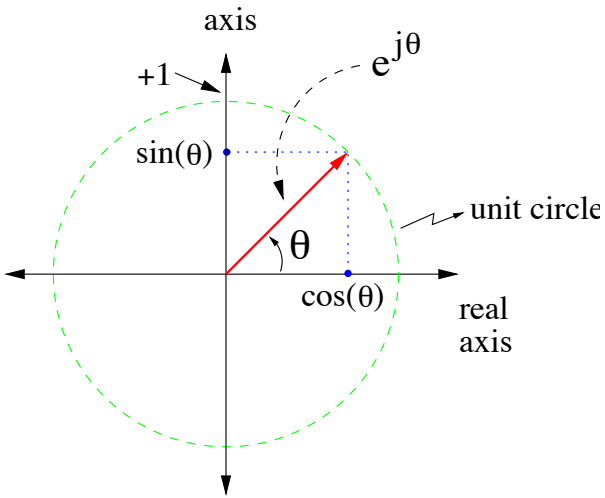
Complex Formulation of the DFT

Often you will see DFT expressed using complex number notation.

Complex number $z = x + jy$ where $j = \sqrt{-1}$. Magnitude $|z| = \sqrt{x^2 + y^2}$, Phase $\theta = \tan^{-1}\left[\frac{y}{x}\right]$

$$e^{j\theta} = \cos\theta + j\sin\theta$$

$$\begin{aligned} S_p &= \sum_{n=0}^{N-1} s(nT)\left[\cos\left(\frac{2\pi np}{N}\right) - j\sin\left(\frac{2\pi np}{N}\right)\right] \\ &= \sum_{n=0}^{N-1} s(nT)e^{-j\left(\frac{2\pi np}{N}\right)} \end{aligned}$$



Hence can express the DFT as

$$S_p = \sum_{n=0}^{N-1} s(nT)e^{-j\left(\frac{2\pi np}{N}\right)} \qquad p = 0, 1, 2, \ldots, N-1$$

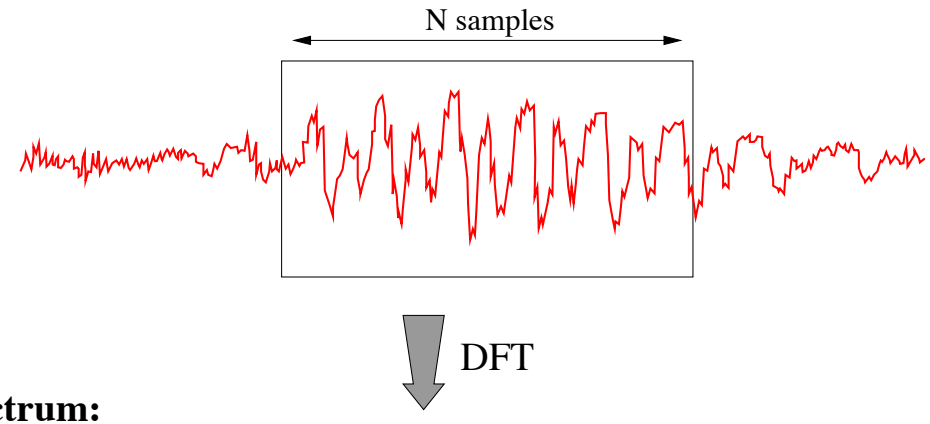
where S_p gives magnitude and phase of spectrum at frequency $\frac{p}{NT}$ Hz.

Note also **Inverse DFT**

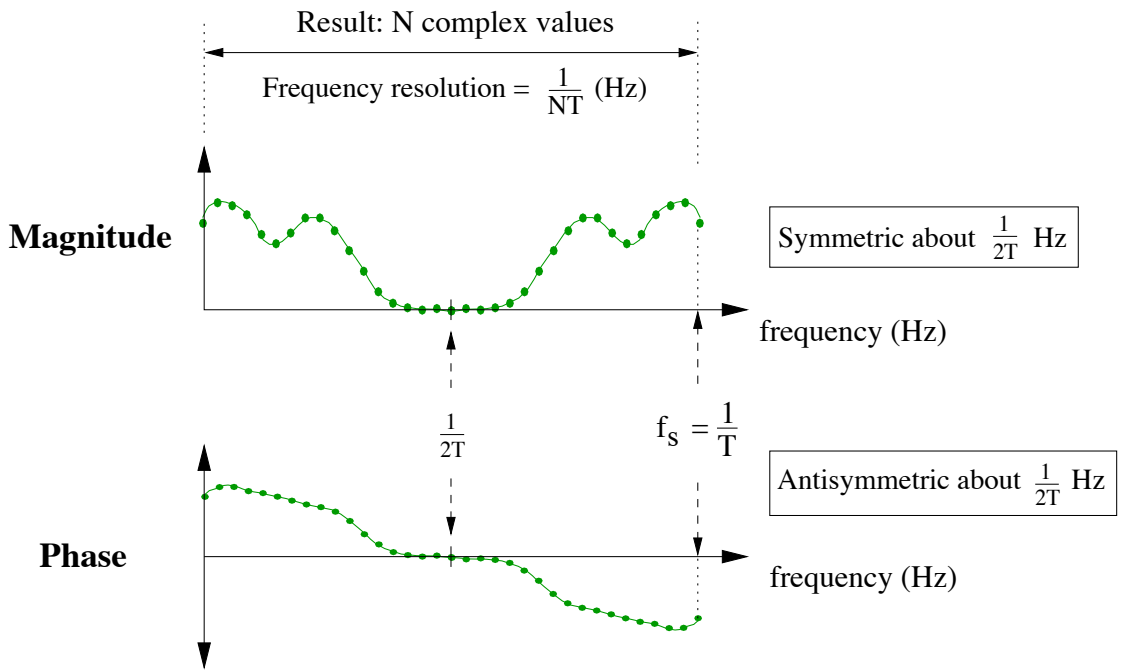
$$s(nT) = \frac{1}{N}\sum_{p=0}^{N-1} S_p e^{j\left(\frac{2\pi np}{N}\right)} \qquad n = 0, 1, 2, \ldots, N-1$$

Using the DFT

Signal:



Spectrum:



Time vs Frequency Resolution

As the length of analysis frame (N) increases we assume that signal characteristics remain the same over progressively longer intervals

- This leads to a reduced ability to respond to sudden changes in the signal: poor **time resolution**

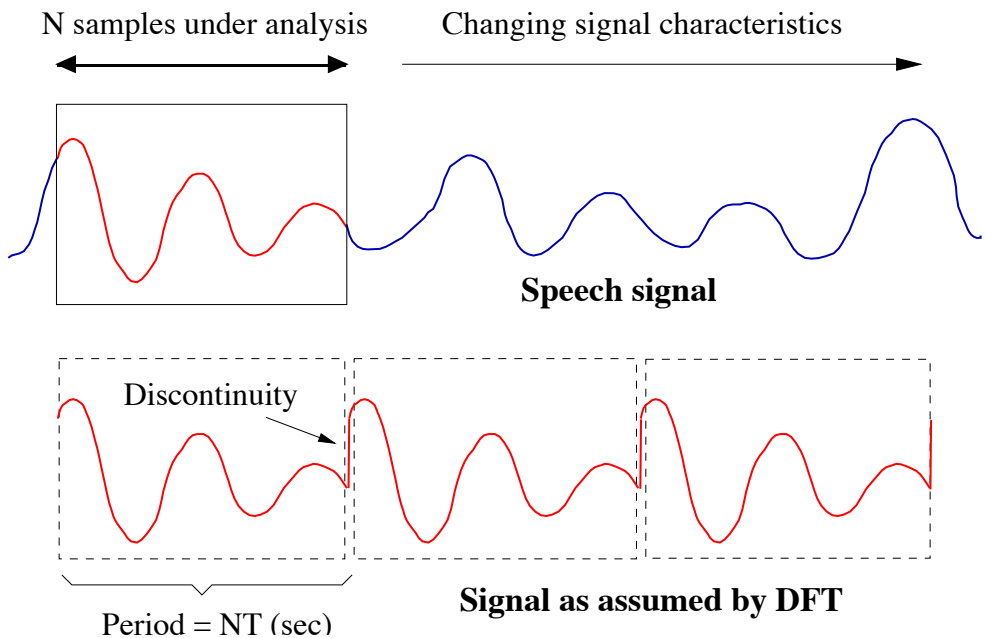
As the length of the analysis frame (N) increases, spacing between the spectral components from the DFT $\frac{1}{NT}$ decreases.

- Hence can better determine the signal frequencies present more accurately: improved **frequency resolution**

	Spectral Resolution	Time Resolution
Large Window N	Good	Poor
Small Window N	Poor	Good

Implicit Periodicity with the DFT

- DFT evaluates spectrum at N evenly spaced discrete frequencies
- Only periodic functions have discrete spectra, aperiodic/stochastic signals have a continuous spectrum
- DFT assumes periodicity outside the analysis frame, with period equal to window length



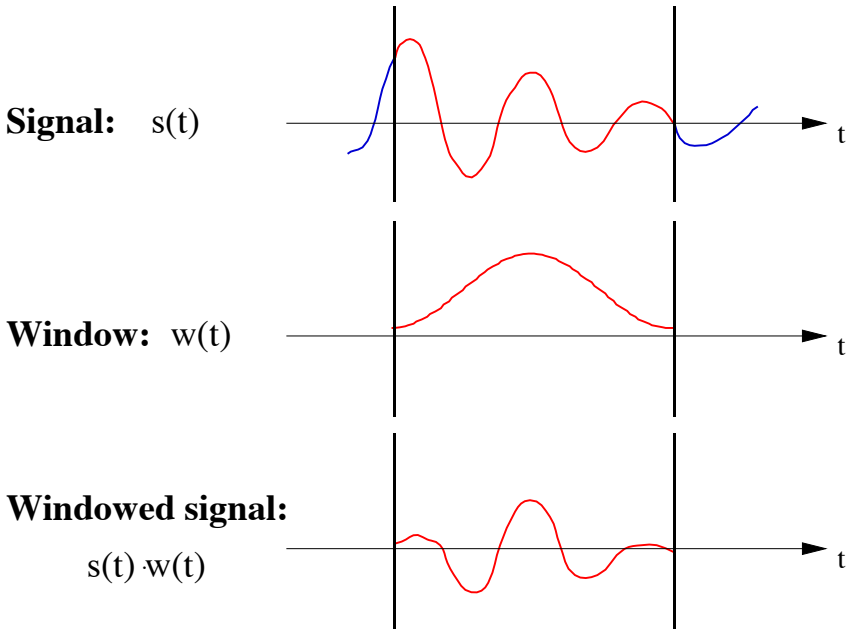
- Discontinuity gives rise to undesirable other components (mainly high-frequency)

Windowing

The discontinuity introduced by segmenting the signal into frames distorts the spectrum. The distortion can be reduced by multiplying the speech frame by a window function

The most common window function is the **Hamming Window**

$$w(nT) = 0.54 - 0.46 \cos\left[\frac{2\pi n}{N-1}\right]$$



- Shape of window (e.g. Hamming vs rectangular) causes different effects (spectrum of window convolved with signal spectrum)
- Hamming window attenuates the components caused by the discontinuity but also smears the spectral peaks peaks.

Fast Fourier Transform (FFT)

- Direct implementation of the DFT requires the order of N^2 multiply-add operations
- By exploiting symmetry can devise an algorithm requiring only order $N \log_2 N$ multiply-add operations: the FFT

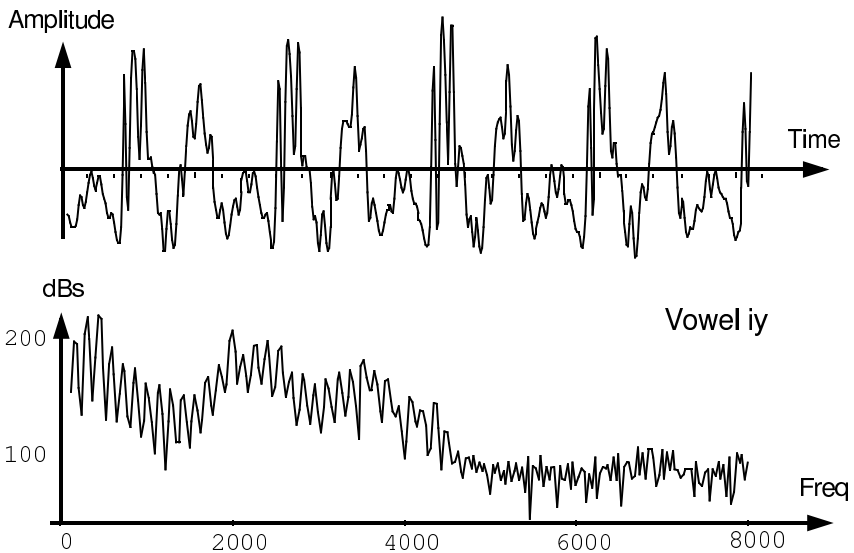
For comparison

N	N^2	$N \log_2 N$	Ratio
32	1024	160	6
64	4096	384	11
128	16384	896	18
256	65536	2048	32
512	262144	4608	57
1024	1048576	10240	102

- Standard FFT requires that the window be a power of 2 samples in size.
- This can be achieved by appropriate choice of analysis size and/or zero-padding (after windowing) the frame to a power of two.

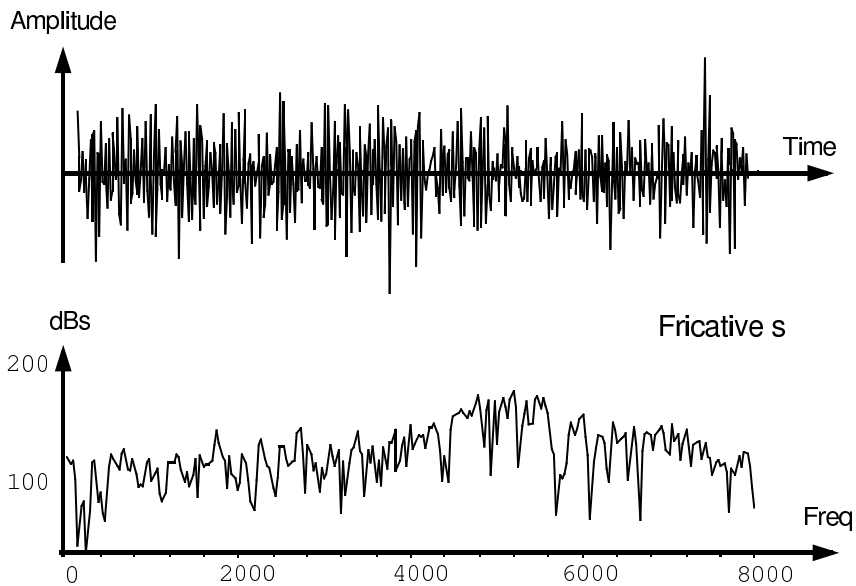
Spectral Properties of Speech

Vowel



- Waveform/Fourier magnitude spectrum of a fragment of the vowel iy. Waveform contains 512 points and it is 32ms in duration (16kHz sampling frequency).
- Vowel **time domain** waveform is approximately **periodic** with $f_0 = 130$Hz.
- Periodic excitation is clearly visible in the spectrum as a high frequency ripple: about 7.5 cycles of ripple per 1000Hz. Confirms time domain pitch estimate.

Fricative



- A time-domain segment and its spectrum for the fricative s
- time domain shows **no periodicity** and the spectrum has only random variations at much higher frequency with peak about 5kHz.

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.

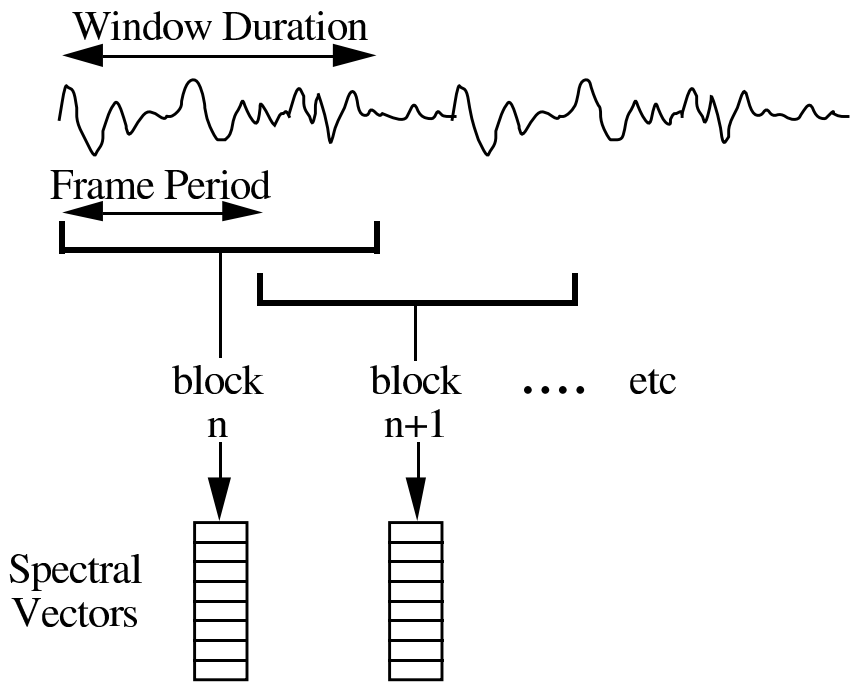
Nasals have a strong low 1st formant around 250Hz and weak higher formants. There is often energy around 2.5kHz.

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.

Block Processing

- For a complete waveform, a spectral estimate must be computed about every 10 ms.
- Since this is rather a short duration to calculate a spectrum (especially with windowing), analysis windows are allowed to overlap
- Typically 25 ms windows are used.

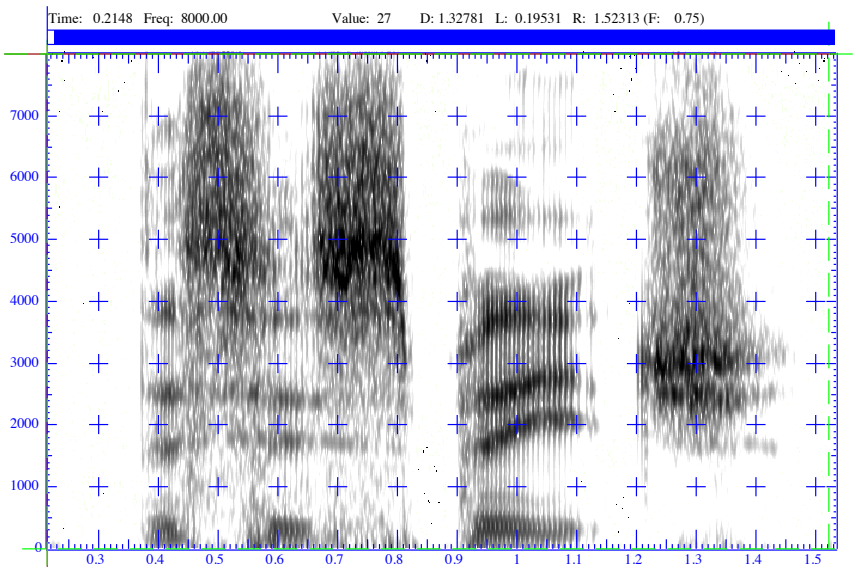


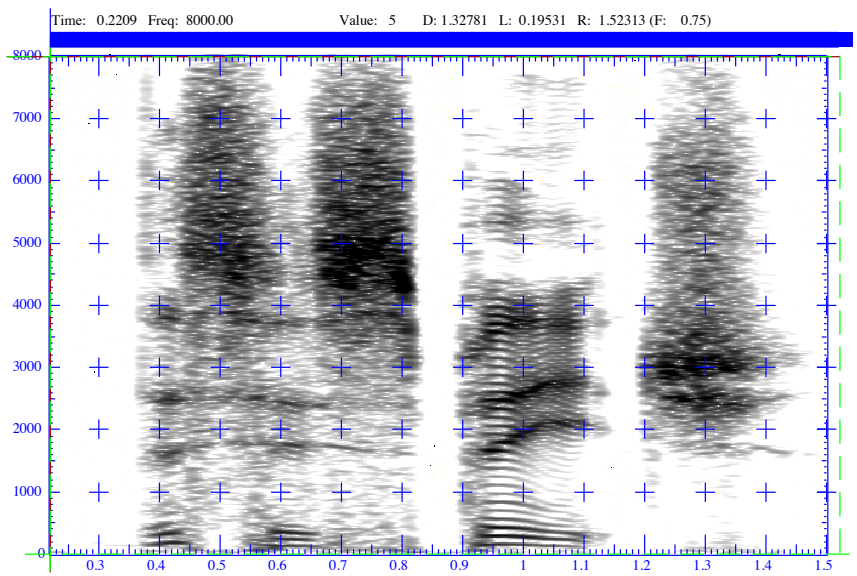
The above figure illustrates this **block processing** technique.
Note that each segment of speech is often referred to as a **frame**

Spectrograms

Sequence of spectra from block processing speech displayed as a grey-scale image with dimensions of time and frequency and with spectral energy represented by the intensity of the image: the **spectrogram**.

Using different length FFTs, different trade-offs between *time-resolution* and **frequency resolution**. Short-window (wide-band) and long-window (narrow-band) spectrograms are shown below.





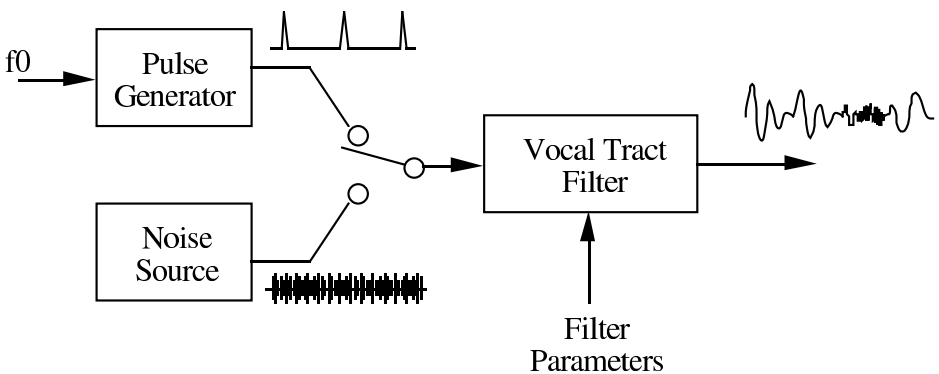
A short analysis window gives good time but poor frequency resolution. In the wide-band spectrogram (left), the pitch periods are visible whereas in narrow-band spectrogram (right) the harmonics of the fundamental frequency can be observed.

Source-Filter Model & Linear Prediction

Vocal tract is a complex time-varying non-linear filter which is excited by a number of different energy sources. A number of simplifying assumptions can be made:

1. Lossless linear time variant filter with single input.
2. Excitation is either a periodic pulse train or noise, depending on basic sound classes.
3. Filter and excitation characteristics are stationary over periods of the order of 10 ms

This leads to the **source-filter model**.



- Explicit estimation of vocal tract filter parameters via **Linear prediction analysis**
- In LP analysis, speech divided into segments of 10-25 ms frames
- Used in low-bit-rate coding (more complex excitation model)

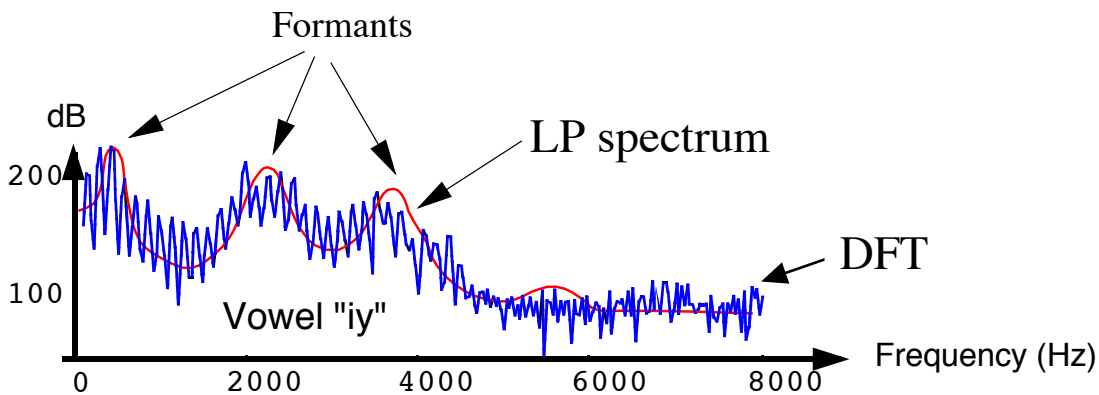
LP analysis uses:

- **All-pole** filter representation
- Typical number of filter parameters p = 10-14 (p is the number of **poles**)
- Computed by autocorrelation analysis
- Minimises sum-squared prediction error from previous p samples in time domain

LP Spectrum

- The frequency response of the LP filter for each frame can be used to estimate the speech spectrum.
- Each pair of poles can model a peak in the spectrum (i.e. a formant)
- LP spectrum is a smoothed approximation of the speech spectrum that is aimed at just modelling the vocal tract
- It can be computed using the DFT (or directly finding filter frequency response)

The resulting LP spectrum is smooth and shows formant structure



- Linear prediction can be viewed as **power spectrum matching**
- Autocorrelation based equations used in LP analysis can be derived in frequency domain

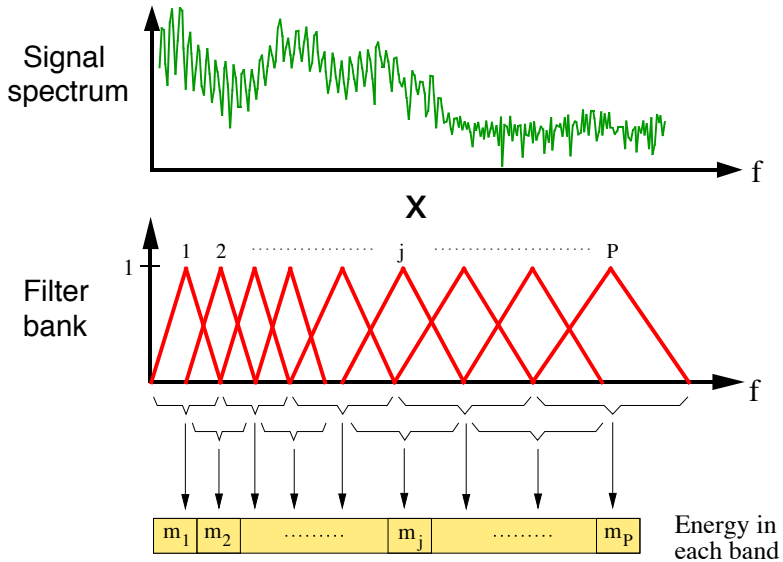
Mel-Scale Filterbanks

Reduce frequency resolution and analysis to model ears spectral resolution.

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

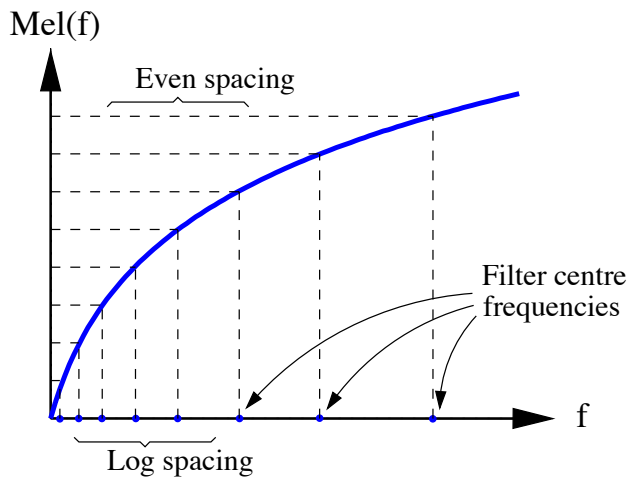
The spacing of the centre frequencies is based on the **Mel-scale**.
The Mel-scale is defined as

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



The Mel frequency scale is shown right.

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

Note that for Deep Neural Network acoustic models, the mel scale filterbank energies can be used directly as features to represent each frame (or can use MFCCs).

Cepstral Representations

- The **cepstrum** of a signal is defined as the IDFT of the log spectrum.
- Aims to separate the vocal tract response and excitation of the signal by filtering the log spectrum
- Cepstrum computed in **quefrency** domain and filtering in this domain is called **liftering**.
- The MFCC cosine transform is similar (hence termed cepstral)

There are a number of alternatives to computing cepstral representations including computing them from a linear prediction analysis (an efficient recursion is possible: don't have to evaluate LP spectrum) which are used in speech recognition systems.

Alternative cepstral representations that use a non-linear frequency scale include using a type of linear prediction analysis termed **perceptual linear prediction** or (**PLP**).

- Compute power spectrum on non-linear frequency scale (Bark scale, similar to Mel scale).
- Compress power spectrum (e.g. with a power-law compression) and other compensation for frequency sensitivity of human hearing
- Compute linear prediction model from autocorrelation coefficients found from power spectrum (normally done in time domain)
- Cepstral coefficients then obtained by applying recursion to linear prediction model (rather than finding log LP spectrum and IDFT).

Summary

- Fourier analysis gives frequency content (amplitude and phase) for periodic signals
- Developed Discrete Fourier Transform (DFT) from correlation with sine and cosine waves
- Given N samples of the signal, DFT calculates the phase, magnitude at N regularly spaced intervals
- Trade off between time and frequency resolution
- Window functions may be used to reduce the spectral distortion introduced by segmentation into frames

- Speech sounds are represented by particular spectral features
- First two formants largely responsible for vowel quality
- Block processing into overlapping frames used for spectral analysis
- Spectrograms give a convenient way to visualise speech

- Linear prediction is an all-pole filter model of speech from source filter model
- Alternative method of computing the power spectrum
- Mel-frequency filterbanks and cepstral coefficients (MFCC) used as ASR representations
- Alternative cepstral representations include PLP cepstra