SPEECH PROCESSING
Time-frequency Analysis

Patrick A. Naylor
Spring Term 2018-19

Learning Objectives

- This lecture is concerned with the **spectrogram** as a representation of how the frequency components within a signal vary with time.
  - Define what we mean by normalized time and frequency
  - Define the short-term discrete Fourier transform
  - Look at the effect of different **window lengths** on time and frequency resolution
  - Derive a quantitative description of the **frequency resolution** of the short term DFT and compare the performance of common windows
  - Derive a quantitative description of the **time resolution** of the short term DFT
  - Explain the **uncertainty principle** and illustrate its effects with some examples

Normalized Time

- With sampled data systems it is customary to change the time scale and work in units of the sample period.
  - In normalized units the sampling frequency (f_s) and period (T) both equal 1. They can therefore be omitted from equations: any such omissions can be deduced using dimensional consistency arguments.
  - The Nyquist frequency is 1/2 Hz = π radians/second
  - To convert back to real units:
    - any time quantity must be multiplied by the real sample period (or divided by the real sample frequency)
    - any frequency or angular frequency quantity must be multiplied by the real sample frequency (or divided by the real sample period)

DFT Properties

$$X_k = \sum_{m=0}^{N-1} x_m \exp\left(-\frac{2\pi j}{N} km\right) = X(z) \text{ evaluated at } z = \exp\left(2\pi j \frac{k}{N}\right)$$

- Exact line-spectrum of a periodic signal $\{x_m\}$
- Sampled continuous spectrum of zero-extended $\{x_m\}$
- Sampled continuous spectrum of infinite $\{x_m\}$ convolved with spectrum of rectangular window
- FFT is an *algorithm* for calculating DFT in time $\propto N \log N$

Energy Conservation (Parseval's theorem)

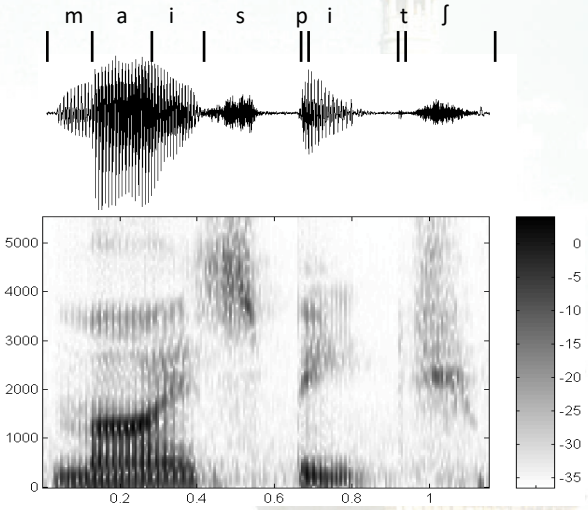
$$E_x = \sum_{m=0}^{N-1} x_m^2 = \frac{1}{N} \sum_{k=0}^{N-1} X_k^2$$

Symmetries $\{x_m\} \Rightarrow \{X_k\}$

Discrete	$\Rightarrow$	Periodic: $X_{k+N} = X_k$
Real	$\Rightarrow$	Hermitian: $X_{-k} = X^*_k$
Periodic: $x_{m+N/r} = x_m$	$\Rightarrow$	Discrete: $X_k = 0$ for $k \neq ir$
Skew Periodic: $x_{m+N/2r} = -x_m$	$\Rightarrow$	Odd Harmonics: $X_k = 0$ for $k \neq (2i+1)r$
Even: $x_m = x_{N-m}$	$\Rightarrow$	Real
Odd: $x_m = -x_{N-m}$	$\Rightarrow$	Purely Imaginary
Real & Even	$\Rightarrow$	Real & Even
Real & Odd	$\Rightarrow$	Purely Imaginary and Odd

Spectrogram

- The spectrogram shows the **energy** in a signal at each **frequency** and at each **time**.
- We calculate this by evaluating the short-term discrete Fourier transform.
- Dark areas of spectrogram show high intensity.



Short-Term Discrete Fourier Transform

- We often want to estimate the "power spectrum" of a non-stationary signal at a particular instant of time.
  - Multiply by a finite length window and take the DFT.

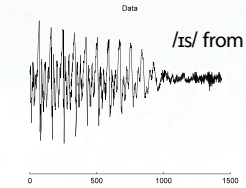
- For a window of length N ending on sample m we have:

$$X(k,m) = \sum_{i=0}^{N-1} w(i)x(m-i)\exp\left(-\frac{2\pi j}{N}k(m-i)\right)$$

  - $|X(k,m)|^2$ gives the power at a frequency of k/N Hz (normalized) for a window centred at m-$(N$-$1)/2$.
  - the $(m$-$i)$ term in the exponent means that the phase origin remains consistent by cancelling out the linear phase shift introduced by a delay of m samples.
  - the window samples are numbered backwards in time (for convenience later) hence the summation is performed backwards in time.
  - the values $X(k,m)$ are based on the N signal values from m–N+1 to m.
  - the frequency resolution is $1/N$ Hz (normalized units)
  - the spectrum is periodic and (since w and x are real) conjugate symmetric:

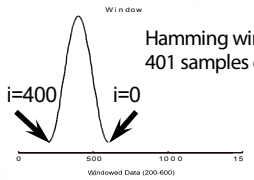
$$X(k) = X(k+N) = X^*(-k)$$

  - there are only $N/2$+1 independent frequency values



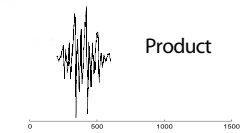
/ɪs/ from "my speech"

$$X(k,m) = \sum_{i=0}^{N-1} w(i)x(m-i)\exp\left(-\frac{2\pi j}{N}k(m-i)\right)$$

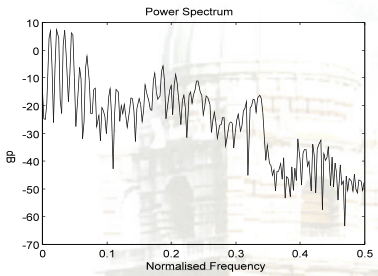


Hamming window of length 401 samples centred on 400.

i=400 i=0



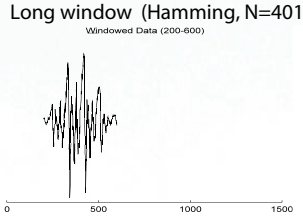
Product

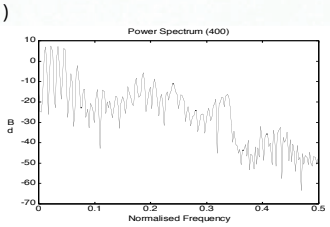


Spectrum from larynx oscillation (at about 0.01 normalized Hz) is superimposed on vocal tract resonances.

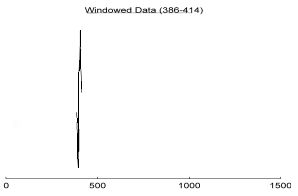
Effect of Window Length

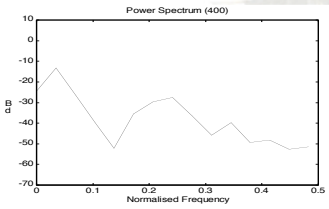
Long window (Hamming, N=401)





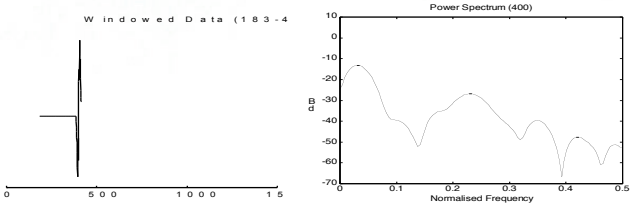
Short window eliminates the fine detail (Hamming, N=29):





Effect of Zero-padding

Zero padded window gives more spectrum points and an illusion of more detail (N=232). This is the normal case for a spectrogram.



Analysis

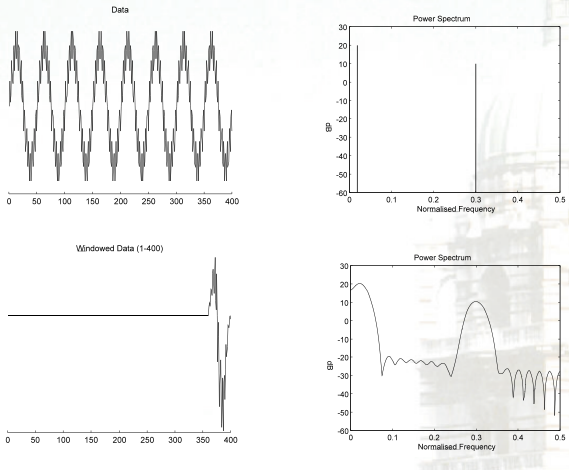
$$X(k,m) = \sum_{i=0}^{N-1} w(i)x(m-i)\exp\left(-\frac{2\pi j}{N}k(m-i)\right)$$

- By setting $r = N-1-i$ we can rewrite this as:

$$X(k,m) = \exp\left(-\frac{2\pi j(m-N+1)}{N}k\right) \times \sum_{r=0}^{N-1} y_m(r)\exp\left(-\frac{2\pi j}{N}kr\right)$$
$$\text{where} \quad y_m(r) = w(N-1-r)x(m-N+1+r)$$

- This is a standard DFT multiplied by a phase-shift term that is proportional to k: this compensates for the starting time of the window: $m–N+1$
- $y(r)$ is a product of two signals so its DFT is the convolution of the DFT's of $w(N–1–r)$ and $x(m–N+1+r)$

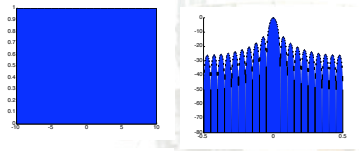
Illustrative Example Plots



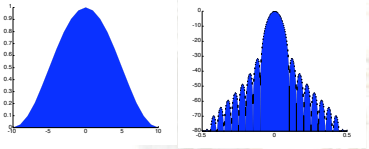
Defining Characteristics of Windows

- For an N-point window
  - the –6dB normalised bandwidth = a/N
  - and $-\infty$ normalised bandwidth = b/N
- Common windows and values for a and b are shown below.

Rectangular:
a=1.21, b=2
Sidelobe = -13dB



Hanning:
a=1.65, b=4
Sidelobe = –23dB

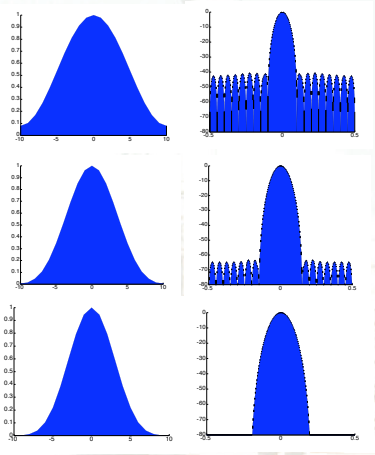


(… contd)

Hamming: $0.54 + 0.46c_1$
a=1.81, b=4
Sidelobe = –43dB

Blackman-Harris: 3 term
$0.423 + 0.498c_1 + 0.079c_2$
a=1.81, b=6
Sidelobe = –67dB

Blackman-Harris: 4 term
$0.359 + 0.488c_1 + 0.141c_2 + 0.012c_3$
a=2.72, b=8
Sidelobe = –92dB

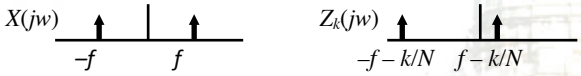


In the formulae
$c_k = \cos(2k\pi(n - \frac{1}{2}N)/N$

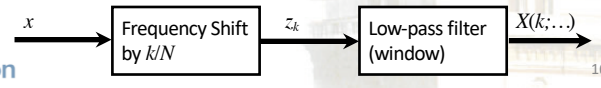
Time Resolution: Filter bank viewpoint

- Consider one particular value of k $\quad z_k(r) = x(r)\exp\left(-\frac{2\pi j}{N}kr\right)$
- $z_k(r)$ is just the same as $x(r)$ but shifted down in frequency by k/N.
  - E.g. if x is a complex exponential at frequency f

$$x(r) = \exp(2\pi jfr) \quad \Rightarrow \quad z_k(r) = \exp(2\pi j(f - k/N)r)$$

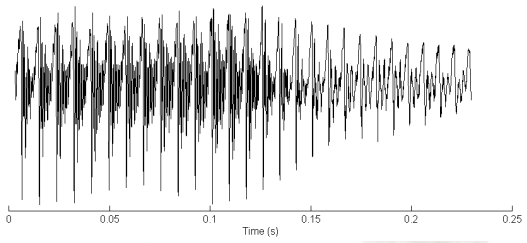


- Hence $\quad X(k,m) = \sum_{i=0}^{N-1} w(i)x(m-i)\exp\left(-\frac{2\pi j}{N}k(m-i)\right) = \sum_{i=0}^{N-1} w(i)z_k(m-i)$
- Thus the k^{th} frequency bin is a filtered version of z_k in which the filter has an impulse response of $w(i)$.
  - From the previous slide, this is a low-pass filter with a –6 dB bandwidth of $1/2(a/N)$
  - Time resolution = $2N/a$ (normalized)

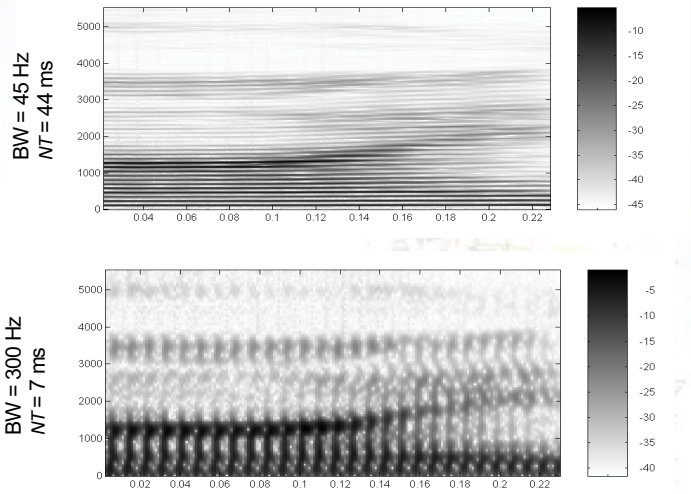


Spectrogram Example

- Consider the following example of /aɪ/ from "my" with 45Hz and 300Hz bandwidth spectrograms
- Time domain waveform

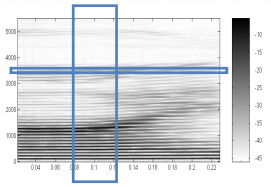


Narrow-band and Wide-band Spectrograms

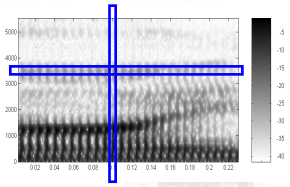


Time-frequency Resolution Trade-off

/aɪ/ from "my" with 45Hz and 300Hz bandwidth spectrograms

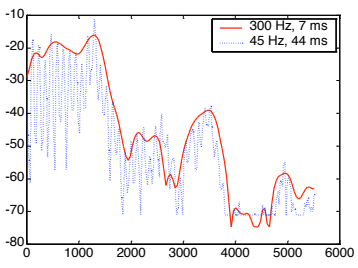


45 Hz, 44 ms

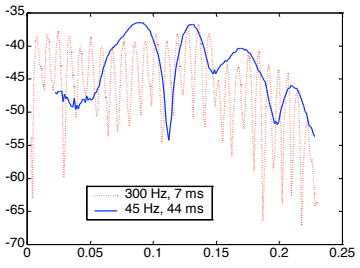


300 Hz, 7 ms

Vertical slice through spectrogram:
$mT = 0.1$ s
45 Hz gives finer frequency resolution



Horizontal slice through spectrogram:
$k/NT = 3.5$ kHz
300 Hz gives finer time resolution

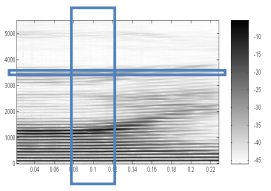


Uncertainty Principle

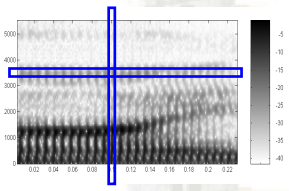
- You cannot get good time resolution and good frequency resolution from the same spectrogram.
  - Duration of window = N/f_s
  - Frequency resolution = $f_s \times a/N$
    - Equal amplitude frequency components with this separation will give distinct peaks
    - Most windows have $a \approx 2$ (see earlier)
  - Time resolution = $2N/af_s$
    - Amplitude variations with this period will be attenuated by 6 dB.
- For all window functions, the product of the time and frequency resolutions is equal to 2.
- Linguistic analysis typically uses a window length of 10–20 ms. The transfer function of the vocal tract does not change significantly in this time.

Overlapping Windows

- To keep all the information about time variation of spectral components:
  - sample the spectrum twice the rate the spectral magnitudes are varying.
- Using normalized frequencies:
  - $-\infty$ dB bandwidth for an N-point window $= b/N$
    - $b = 4$ for a Hamming window
  - Significant variation of spectral components occurs at frequencies below $\frac{1}{2}b/N$
  - we must sample the spectrum at a frequency of b/N
    - the separation between spectral samples is the window width divided by b



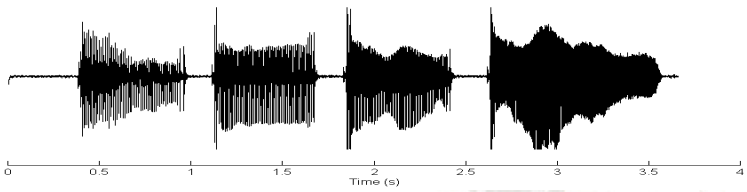
45 Hz, 44 ms



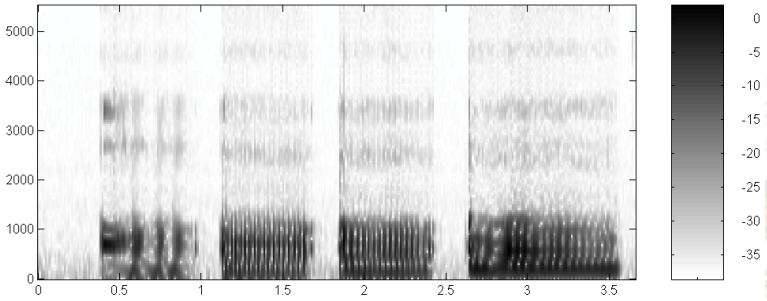
300 Hz, 7 ms

Examples of Singing

/A/ (“ah”) sung as an arpeggio



300 Hz Bandwidth: Note constant formant frequencies



50 Hz Bandwidth: Note harmonic spacing increases + f_x warbles