Speech Analysis
Elementary Speech Analysis

- **Speech analysis, or speech signal processing** transforms a speech waveform into a form which is more suitable for analysis:
  - **Human visual inspection** - for example by a phonetician, speech scientist, speech therapist or forensic phonetician
  - **Computer analysis** - for example for automatic speech recognition, speaker recognition or paralinguistic processing
Speech Analysis Continued

- **Suitable** might mean:
  - Amenable to human visual interpretation
  - Requiring a small number of bits per second for transmission across a comms channel, or storage
  - Compatible with the assumptions in a particular speech model for speech recognition
  - Analogous with the analysis performed in the human peripheral auditory system...
The Short-Term Spectrum

- Let $t$ be a particular time

- A vertical ‘slice’ through the spectrogram represents distribution of power with respect to frequency over a time interval centred at $t$.

- Called the short-term spectrum at time $t$
  - From the perspective of the source-filter model, it tells us about the shape of the vocal tract at time $t$
  - From the perspective of human speech perception, we know that a similar analysis is performed in the cochlea in the initial stages of human speech perception
Calculating the short-term spectrum

- The short-term spectrum can be computed in various ways. However, they all involve:
  - Low-pass filtering
  - Analogue-to-Digital (A/D) conversion - convert analogue signal into a digital signal
  - Windowing - select a short section of speech centred at time $t$, and smooth its edges
  - Frequency Analysis - estimate distribution of power w.r.t frequency at time $t$. 
A-D Conversion

- PCM (Pulse Code Modulation) measures and encodes the speech signal at regular sampling points.
Sample Rate

- **Nyquist’s theorem** – sample rate of $2N$ samples/s needed to encode a signal band limited at $N$ Hz
- Human ear sensitive to frequencies up to around 20,000 Hz (hence 44 samples per second CD rate)
- But for **speech**:
  - High-quality $\Rightarrow$ 10,000 Hz bandwidth $\Rightarrow$ 20,000 samples/s
  - Bandwidth can be reduced to 4,000Hz (telephone bandwidth $\sim$ 3,750Hz) $\Rightarrow$ 8,000 samples per second
  - E.G: Some civil telephony uses 8-bit PCM at 8K samples per second - 64K bits per second
Calculation of short-term spectrum

![Graph showing short-term spectrum calculation with window size and time (t) axes.](image-url)
Windowing

Original signal \( s(n) \)

Hamming window

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

Windowed signal

\[
s'(n) = w(n)s(n)
\]
Frequency Analysis

- **Discrete Fourier Transform** (DFT) applied to windowed digital waveform \( \{s(n) : n = 1, \ldots, N\} \).

- Assuming \( N \) sample window, this results in an \( N/2 \) point **complex** spectrum \( \{S(f) : f = 1, \ldots, N/2\} \).

- Take **modulus** - \( N/2 \) point **power spectrum** \( \{P(f) = |S(f)| : f = 1, \ldots, N/2\} \). (phase ignored)

- Take **logarithm** to compress dynamic range, - **log-power spectrum** \( \{LP(f) = \log|S(f)| : f = 1, \ldots, N/2\} \).

- The log-power spectrum, computed over a short window centred at time \( t \), is referred to as the **short-term (Fourier) spectrum at time \( t \).**
Frequency Analysis

- This is **not** the only way to compute a short-term power spectrum
- Other approaches include:
  - **Filter-bank** analysis (based on a set of **band-pass** filters)
  and
  - **Linear Predictive Coding** (LPC-Spectrum) (c.f. tomorrow’s lecture)
Band-pass filter

- Filters encountered in speech processing typically result from, or simulate, physiological processes

Equivalent Rectangular Bandwidth (ERB)

Frequency (Hz)
Filter-bank

- Spectrum can be estimated as a vector of outputs from a bank of band-pass filters
- $y = y_1, \ldots, y_N$ where $y_n$ is the output of the $n$th filter
Time - Frequency Resolution

- Back to the DFT…

- If the window is long then
  - number of points $N$ in frequency analysis is large
    $\Rightarrow$ the number of points in the spectrum is large,
    $\Rightarrow$ fine frequency resolution, poor temporal resolution
  - long window $\Rightarrow$ narrow-band frequency analysis - narrow-band spectra.
Time - Frequency Resolution

- If the window is **short** then
  - poor frequency resolution, but **fine temporal resolution**

  **short** window $\Rightarrow$ **broad-band** frequency analysis - **broad-band spectra**.

- In summary:
  - **short** window $\Rightarrow$ **broad-band** frequency analysis
  - **long** window $\Rightarrow$ **narrow-band** frequency analysis
Bandwidth for implicit DFT filters

- The value of the spectrum at a particular frequency $f$ can be thought of as the output of a band-pass filter, with bandwidth dependent on window size (in seconds)

- If the sample rate is $N$ samples per second and the window length is $L$ samples, then for a Hamming window, the implicit filter bandwidth is $2N/L$
Wide- and Narrow-band Spectrograms
Wide- and Narrow-band Spectrograms
Bandwidth of speech signals

- CD quality speech sampled at 44kHz, giving 22kHz bandwidth
- In the case of speech, almost all of the relevant information lies below 20kHz, so a sample rate of 20kHz gives good quality
- Restricting the bandwidth to 3.75kHz results in intelligible speech, but quality is degraded
- Intelligibility compromised at bandwidths below 3.75kHz
Speech (22kHz bandwidth)
Speech (11kHz bandwidth)
Speech (5.6kHz bandwidth)
Speech (2.8kHz bandwidth)
Continuity of Speech Patterns
Window Size and $F_0$ - adult male

- Consider a low-pitch adult male speaker
- $F_0$ typically between 50Hz and 200Hz
- Typical analysis window is around 20ms
- Hence, for a low-pitch adult male speaker with 50Hz $F_0$, window size corresponds to approximately one excitation pulse
- Spectrum shape depend on precise position of window w.r.t. pulse
- …Pitch synchronous analysis window?
Window size and $F_0$

- **CASE 1**: Adult male, 20ms analysis window

![Graph showing window 1 and window 2 with F0 values.](image)
Window Size and $F_0$ - female

- Now consider a high-pitch female speaker, or child
- $F_0$ typically between 100Hz and 400Hz
- Typical analysis window ~20ms
- Hence, for a female speaker with high, 400Hz $F_0$, window size corresponds to approximately 8 larynx pulses
- Spectrum shape will reflect vocal tract filter shape plus excitation spectrum
Window Size and $F_0$ - female

- **CASE 2**: female speaker, 20ms analysis window

![Graph showing 20ms analysis window](image-url)
Effect of Glottal Waveform

- For a voiced speech sound, the ‘measured’ short-term spectrum will be the result of two phenomena:
  - The vocal tract shape
  - The excitation signal from the larynx
- More precisely, the measured spectrum will be the **product** of:
  - The vocal tract filter
  - The spectrum of the excitation signal
Speech spectral analysis

VOCAL CORD PULSES

RADIATED WAVE

SOURCE SPECTRUM

VOCAL TRANSMISSION

SPECTRUM OF RADIATED VOWEL ɑ

RELATIVE LEVEL IN DB

FREQUENCY IN KILOCYCLES

Speech Technology Lab

EEM4R: Introduction to speech analysis
Narrow band spectrum
Wide band spectrum
Summary

- Brief introduction to speech signal processing
- Main goal was to review spectral analysis
- More details throughout the course