Speech Processing

Undergraduate course code: LASC10061
Postgraduate course code: LASC11065

All course materials and handouts are the same for both versions.
Differences: credits (20 for UG, 10 for PG); exam/coursework weightings; marking criteria
All course materials are available via Learn

Slide pack 1 of 3: Introduction

What is speech processing?

- Communicating with machines via speech
  - Speech input ("speech-to-text") → *automatic speech recognition*
  - Speech output ("text-to-speech") → *speech synthesis*
- But also processing human-human communication
- Applications of speech recognition
  - dictation; audio archive searching; voice dialling; command-and-control
- Applications of speech synthesis
  - telephone services; reading machines; eyes-free applications; computer games; voice communication aids; announcement systems
- End-to-end applications
  - spoken dialogue systems; conversational agents; speech-to-speech translation

Course structure

- Three blocks
  - Introduction
  - Speech synthesis
  - Speech recognition
- Each week you should attend
  - One lecture
  - One of the lab-based tutorial sessions
- See Learn for schedule
  - You will have tasks to complete both *before* and *after* each lecture!

Timetable

- **Lecture** - split into two parts
  - Thursdays 9.00-9.50 + 10:00-10:50
- **Labs** - multiple groups (number of groups varies with class size)
  - See Learn for times
- **Things you need to do immediately**
  - Sign up on Learn for one lab group
  - Return the lab access form
  - Get a linguistics computer account - go to the lab after the first lecture
  - Optional: attend an “Introduction to Unix” session - sign up on Learn

Syllabus

- Basics
  - Waveform, spectrum, spectrogram
  - Speech production, speech perception
  - Acoustic phonetics
- Speech synthesis
  - Components of a Text-to-speech synthesiser. Text analysis; lexicons, phrasing accents, pitch; waveform generation and prosodic manipulation
- Speech Recognition
  - Components of a recogniser. Dynamic time warping, Probability distributions. Hidden Markov models. Bayes' Theorem. Viterbi algorithm for recognition. Training HMMs. Simple language models.

Practicalities

- Computer accounts
  - All practicals are done on the iMac computers, running OS X
  - We are not able to support use of your own computer
- Unix/Linux/command line OS X
  - Basics: Terminal, mv, cp, cd, mkdir, starting programs
  - Never switch off machines, just log out
- Lab access
  - Via matriculation card at any time (PIN required out of hours)

Assessment

- Two practical assignments and an exam
  - 20% (PG) / 25% (UG) - speech synthesis practical write-up
  - 20% (PG) / 25% (UG) - speech recognition practical write-up
  - 60% (PG) / 50% (UG) - closed book exam
- Coursework due dates are given on Learn
- Exam: December

Reading

- ***Speech and Language Processing (SECOND EDITION!)*, Daniel Jurafsky and James H. Martin.** ***Many copies on short loan, main library***
- *Speech Synthesis*, Paul Taylor. *Main library, or available in electronic form*
- *Spoken language processing*, Xuedong Huang, Alex Acero and Hsiao-Wuen Hon. *Optional reading only*
- *Speech Synthesis and Recognition,* John N. Holmes and Wendy J. Holmes (2nd edition). *Main library, or available in electronic form*
- *Fundamentals of Speech Recognition*, Lawrence R. Rabiner and Biing-Hwang Juang. *Optional reading only*
- *Elements of Acoustic Phonetics*, Peter Ladefoged. 2nd edition (1996). *Many copies on short loan, main library*

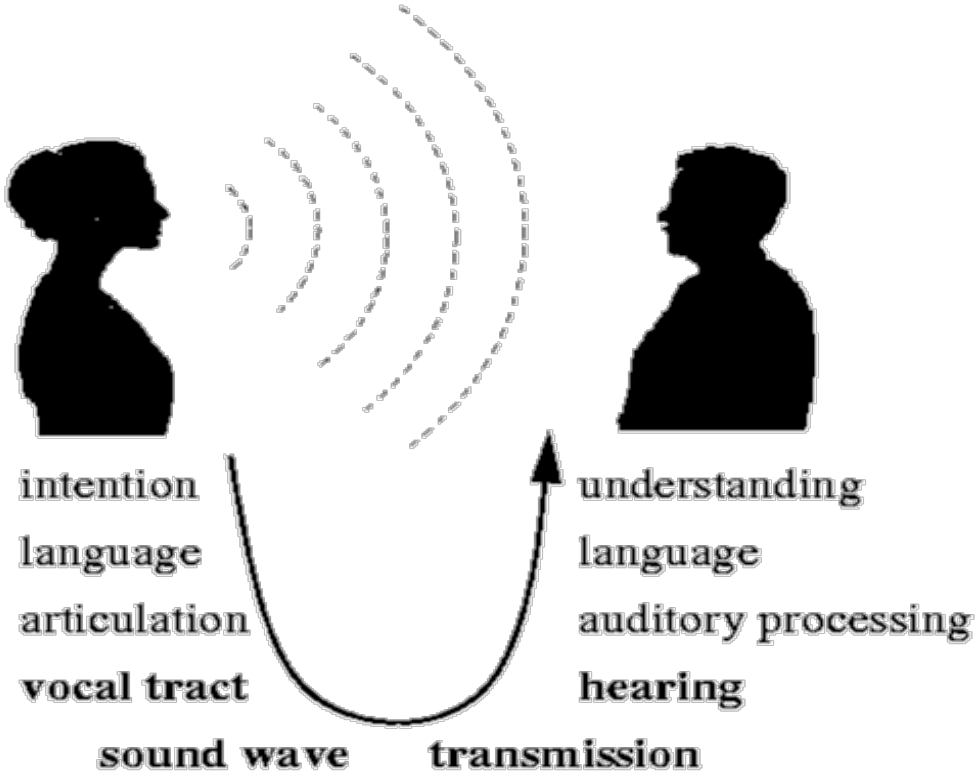
Please co-operate and share library copies!

Disciplines

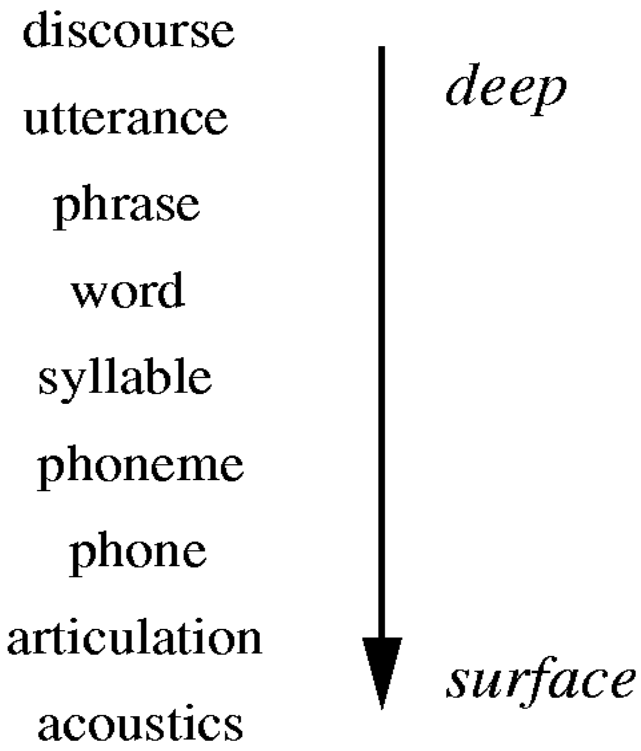
- This course involves:
  - Linguistics: Phonetics, phonology, intonation, (perhaps syntax)
  - Mathematics: statistics and probability, parameter estimation
  - Engineering: practical implementations, empirical findings
  - Computer science: algorithms, efficient implementation

The speech chain

- Some jargon:
  - ASR – automatic speech recognition
  - TTS – text-to-speech



Levels of representation

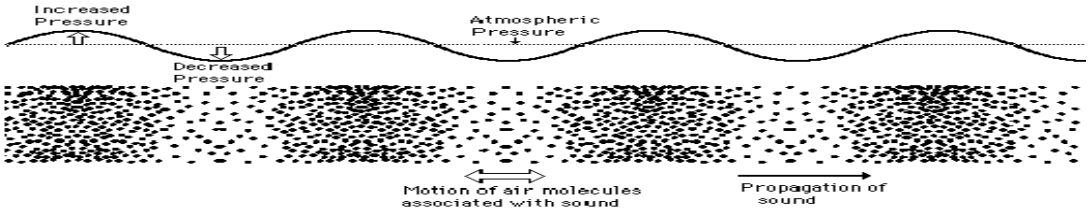


Some basic concepts

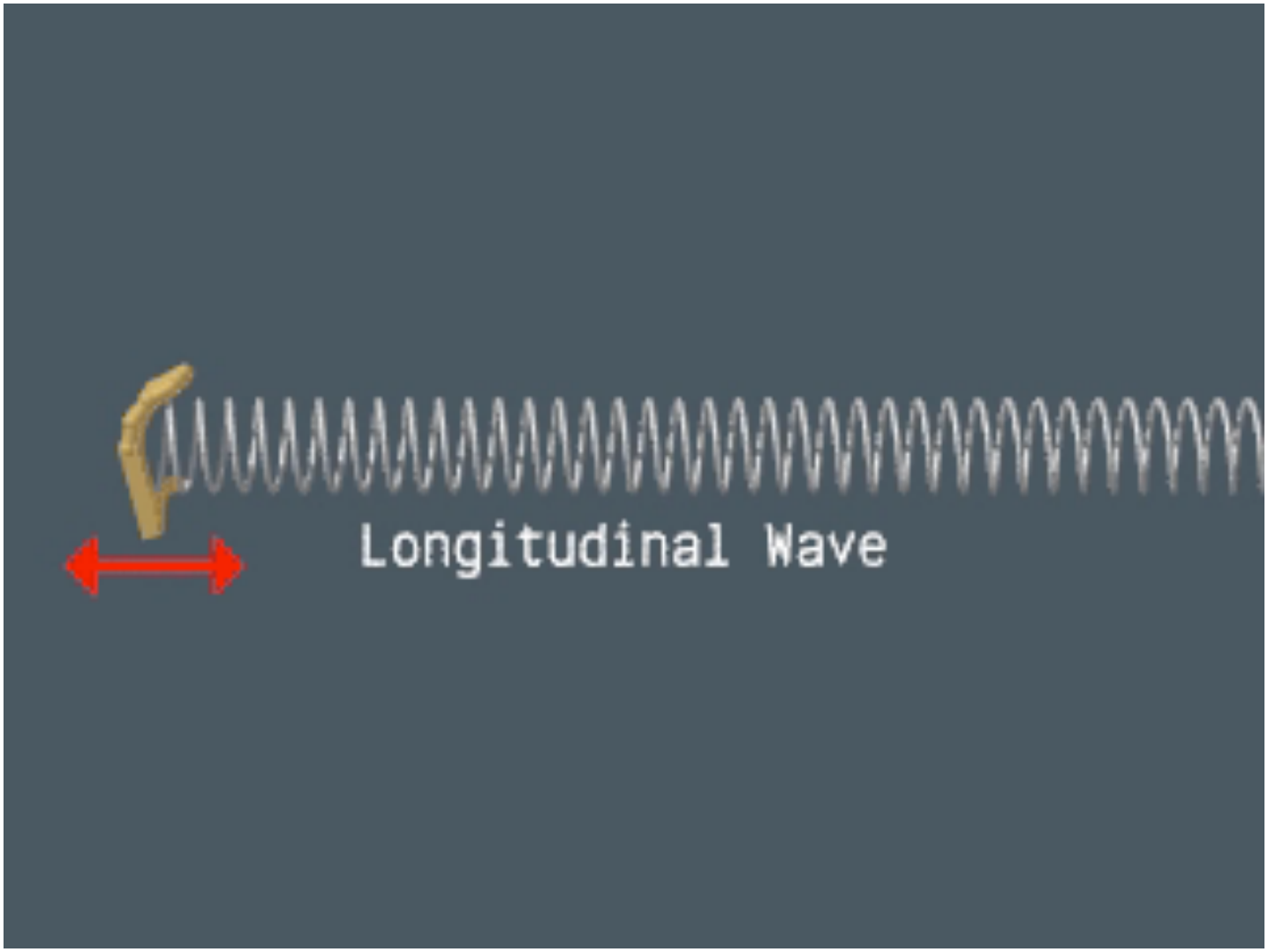
- Before going on, we need to understand some concepts
- Basics:
  - What is sound?
  - The speech waveform
- Not so basic
  - The frequency domain
  - Spectrum
  - Spectrogram

What is sound?

- Pressure waves transmitted through a medium -- e.g. air
- Analogy: a spring – regions of compression and expansion



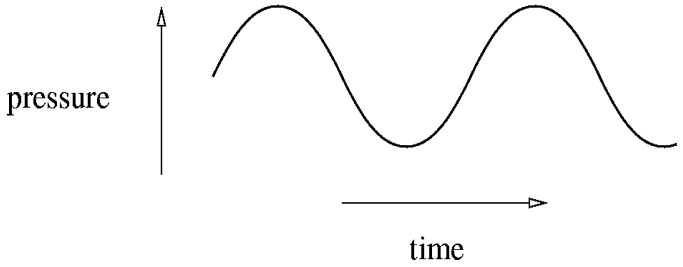
Measure pressure with a microphone
Can plot pressure against time



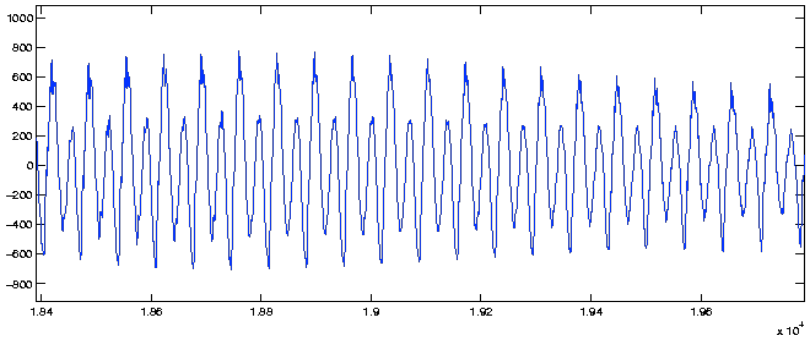
14

Waveforms

- Simple waveform



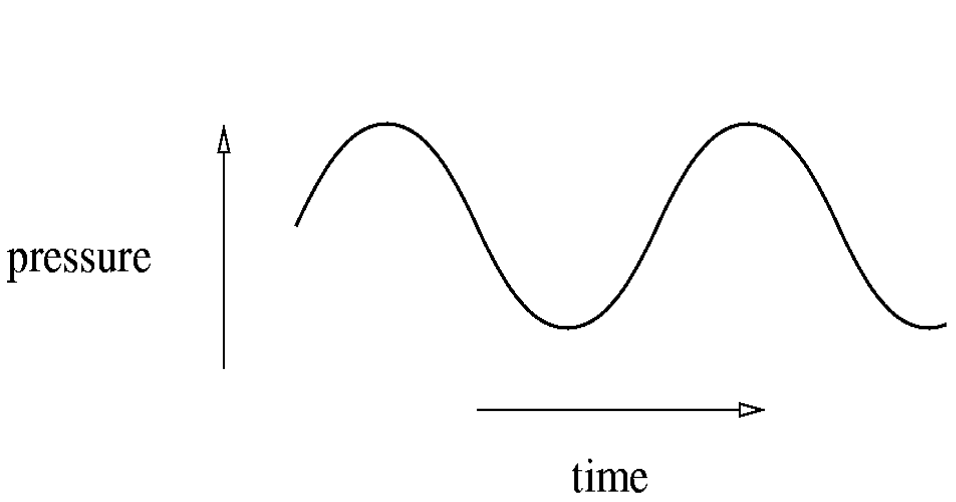
- Speech waveform

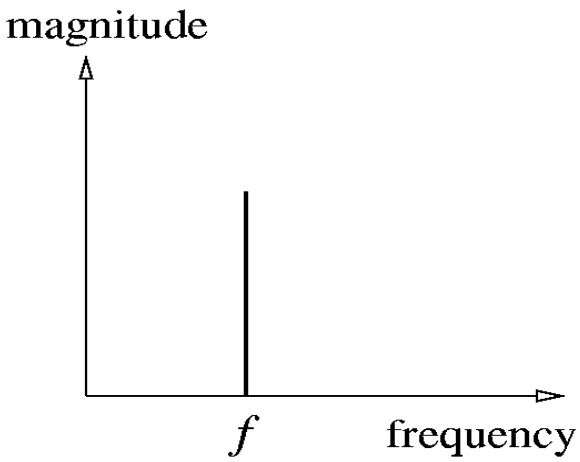


Why is the speech waveform more complicated?

Concept: Spectrum

- This is a pure tone – it contains a single frequency. We can plot the signal in the frequency domain - we call that the spectrum





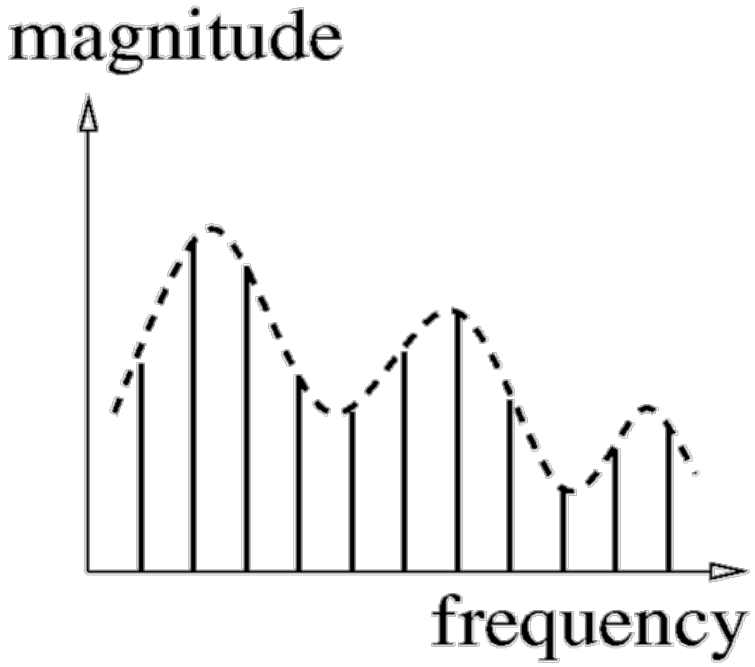
Analogy: a prism splits light into its component colours.

What does the spectrum of the speech signal look like?

Spectra and the Fourier principle

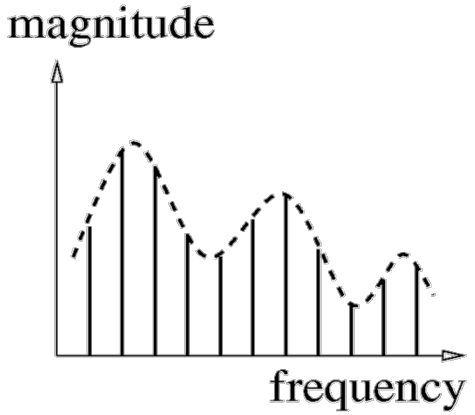
- The Fourier principle tells us that any periodic signal can be decomposed into a sum of simple signals (sine waves)
    - Fourier analysis tells us which sine waves we need to add together, to make the original signal
    - The amplitudes of those sine waves reveal the frequency content of the original signal
- Real world signals tend not to be perfectly periodic
    - but we can often assume that they are over some short period of time
    - so we perform Fourier analysis on short regions of the signal

Spectrum of a voiced sound



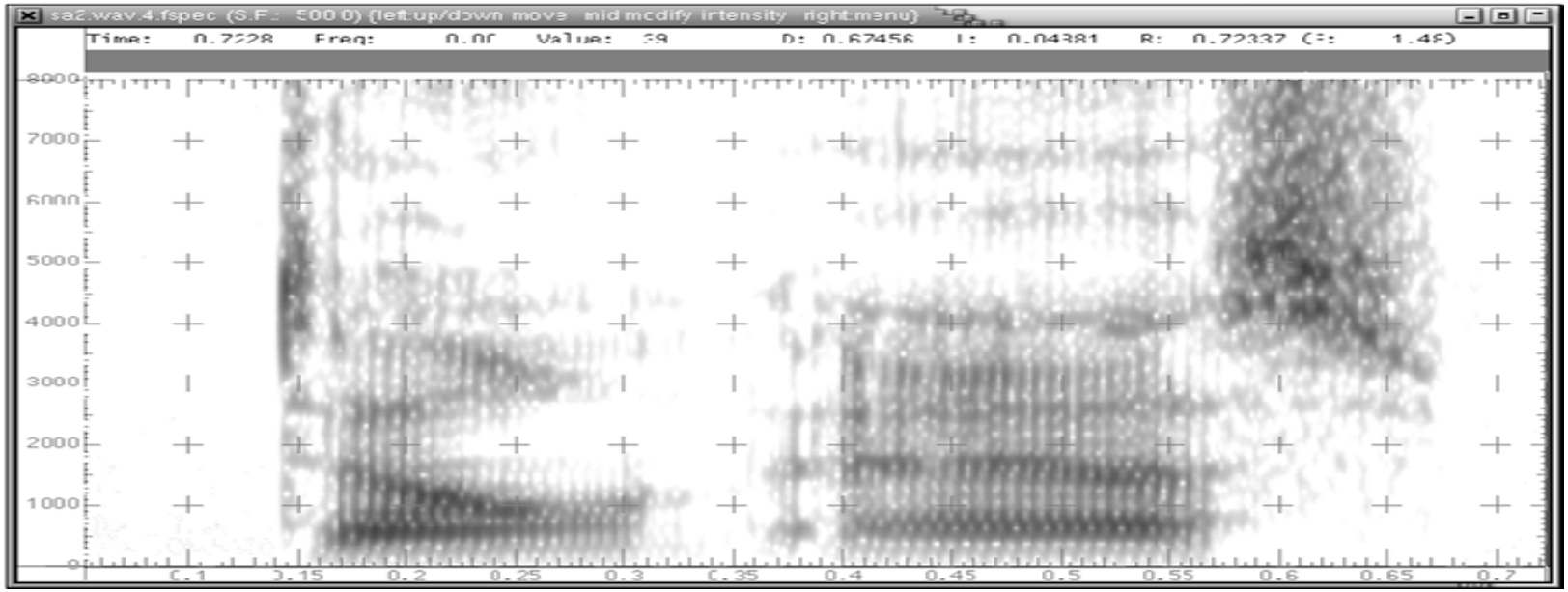
Analysis of the speech spectrum

- Two distinct components in voiced speech
  - Overall shape (spectral envelope)
  - Spectral detail
- Can we explain these in terms of the speech production mechanism?
- What about other classes of sound?



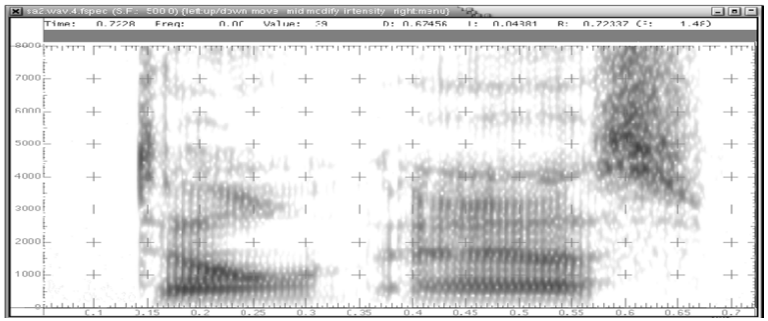
Concept: Spectrogram

- A spectrum is a snapshot of the frequency content of a waveform at one “instant” in time (or over some short region of time)
- A spectrogram shows how the spectrum changes over time



Analysing the speech spectrogram

- There is clearly more than one class of sound
- Can you
  - segment the spectrogram into regions ?
  - group similar regions into classes of sound ?



Exercises

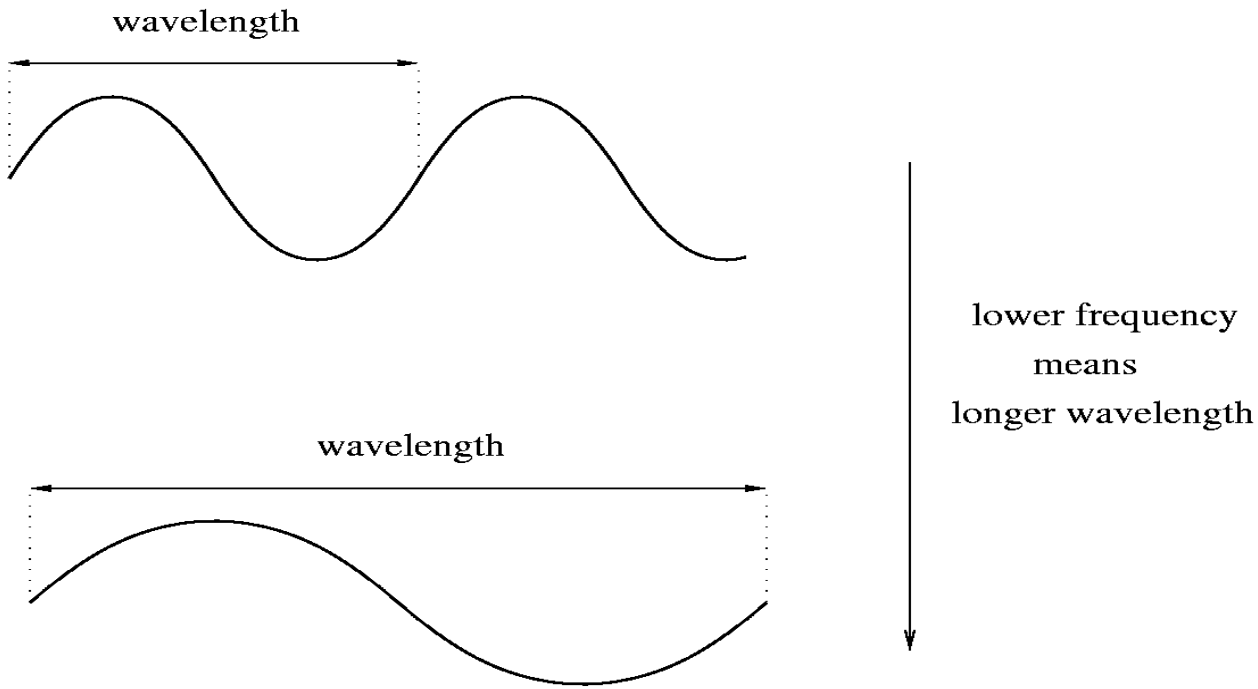
- Examine waveform, spectrum and spectrogram of
  - Pure tones
  - Pulse trains
  - Speech
- Examine speech spectrogram and
  - Try to segment into regions
  - Group regions into classes
  - Work out how each class of sounds was produced

Frequency, period and wavelength

- The speed of sound in a given medium is constant (about 350 ms^{-1} in air at sea level)
- FREQUENCY is the number of cycles per second.
  - peaks per second observed at some fixed position in space.
  - measured in Hz (Hertz), which is the same as 1/seconds, or s^{-1}
- Time between peaks is the PERIOD (unit: seconds, s)
- Distance between peaks in the WAVELENGTH (unit: meters, m)

Frequency and wavelength

- Higher frequency means pressure peaks are closer together
- i.e. at higher frequencies wavelength is shorter



Resonance

- Resonant systems will oscillate when energy is input at the right frequency
- Examples:
  - Clock pendulum, child's swing
  - Mass + spring
  - Air in a tube, e.g., an organ pipe, a bottle, the vocal tract

Analysing resonance: air in a tube

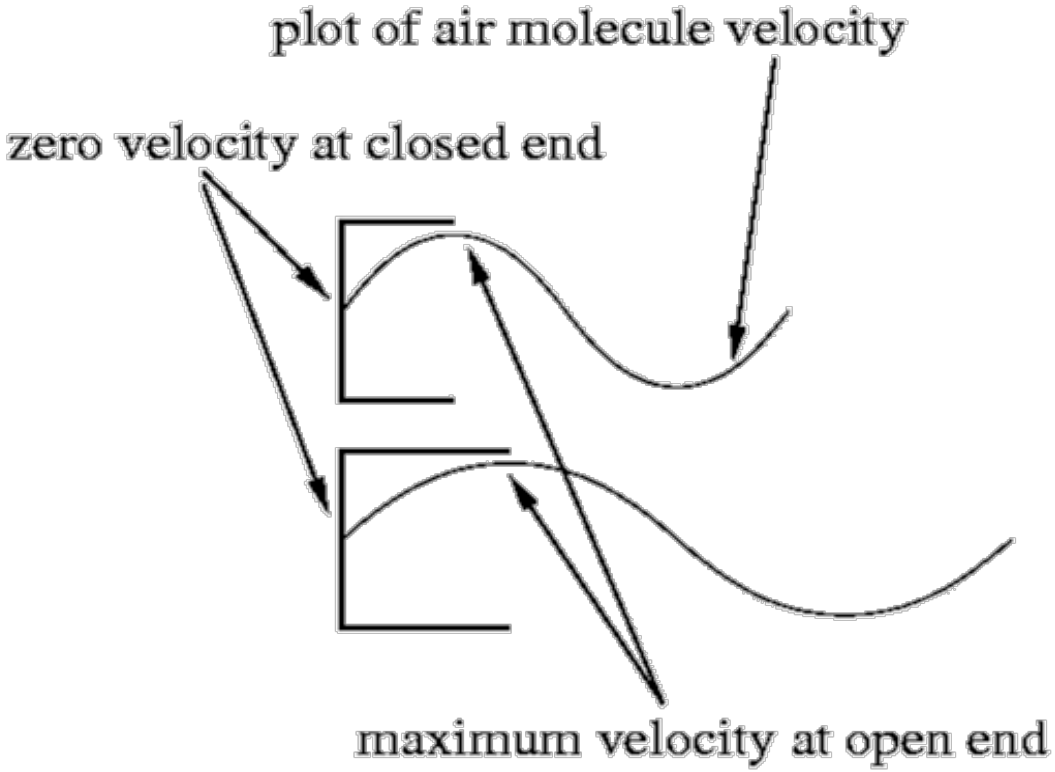
- Some periodic sound sources generate pressure waves
- Pressure waves propagate (travel) along the tube, and are reflected when they reach the end
- Resonance will occur if reflected pressure waves are “in step” with new waves produced by the sound source
  - “in step” waves add up and reinforce one another
  - amplitude builds up

Standing waves

- When reflected waves coincide and resonance occurs:
  - a fixed pattern of pressure waves is set up within the tube
  - this pattern of pressure peaks and troughs is called a standing wave
  - the individual waves do not stand still, but they create a stationary pattern

  - http://www.walter-fendt.de/ph14e/stlwaves.htm
  - http://paws.kettering.edu/~drussell/Demos/waves/wavemotion.html

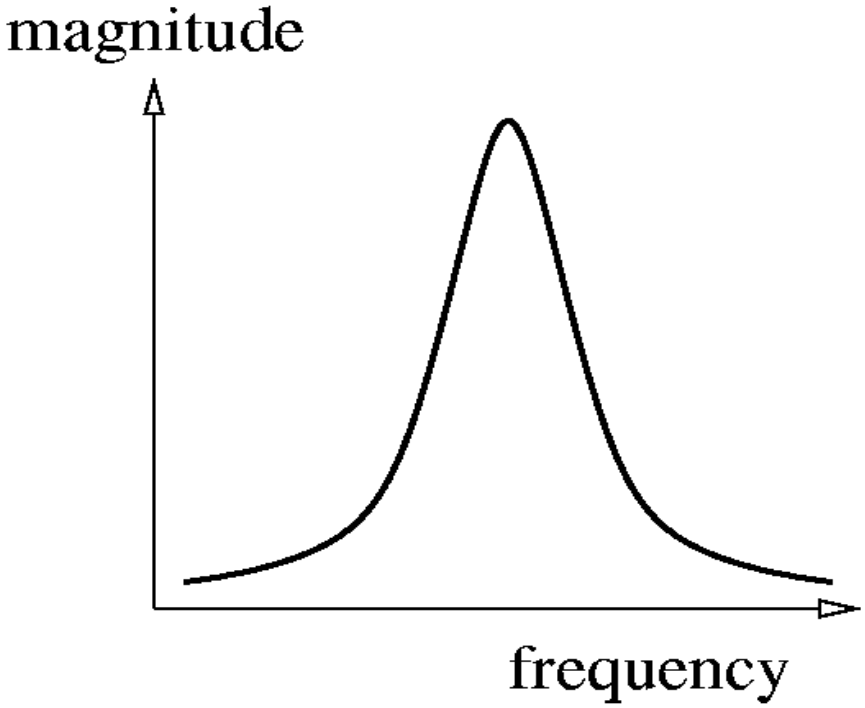
Resonance: tubes of different lengths



fundamental wavelength is
4 times the tube length

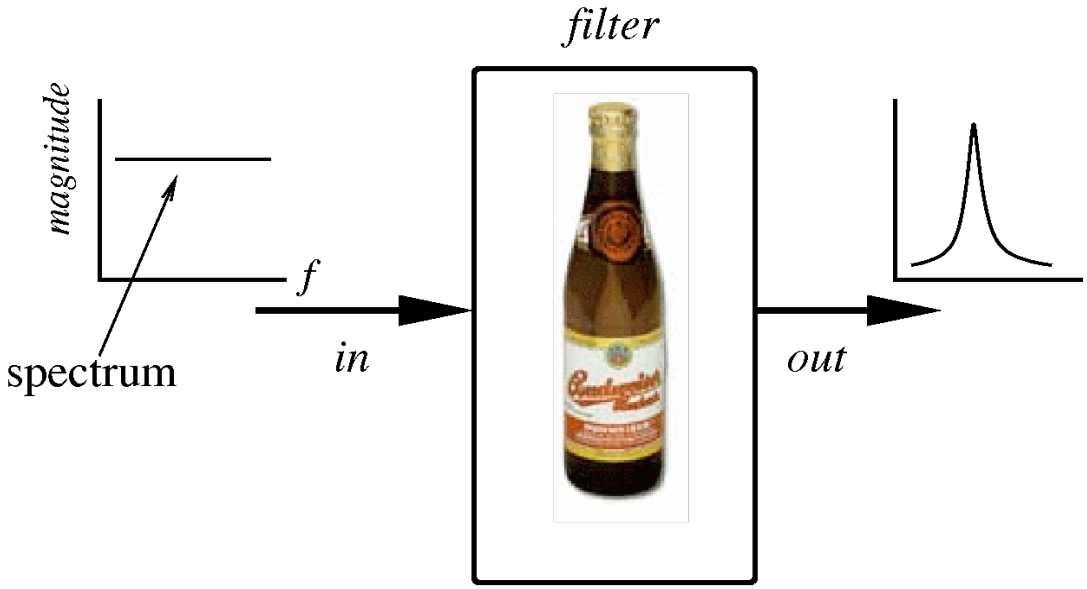
Resonance and filtering

- A simple resonator responds to a certain input frequency
- Sounds waves at (or close to) the resonant frequency will be amplified



For example a bottle

- We put energy into the bottle by blowing:



- What comes out is energy only at the resonant frequencies of the bottle.

Relationship between frequency and wavelength

- Standing waves occur inside the tube. The relationship between frequency and wavelength is:

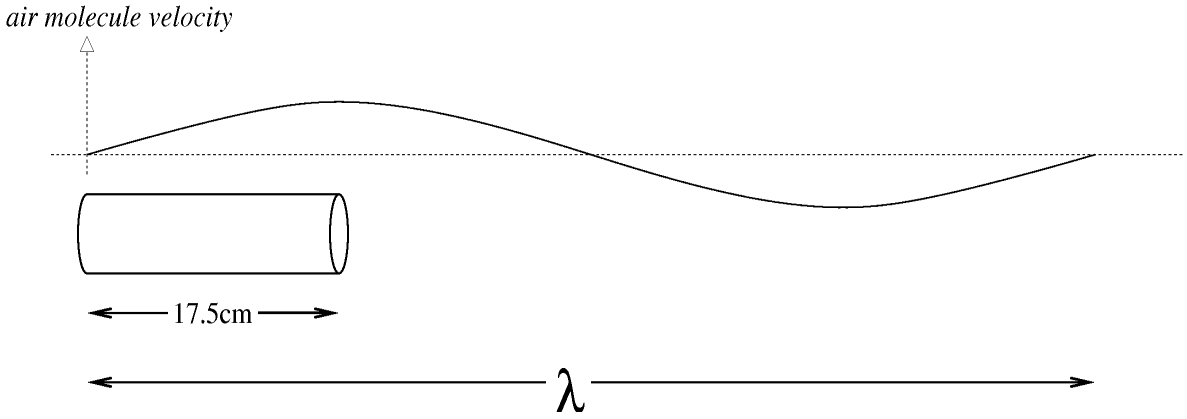
$$f\lambda = c$$

f is frequency, λ is wavelength

c is about $350 ms^{-1}$ (metres per second)

Fitting waves into the tube

- The wavelength of the lowest resonance (i.e. longest wavelength) for this tube has a wavelength of 4 times the length of the tube

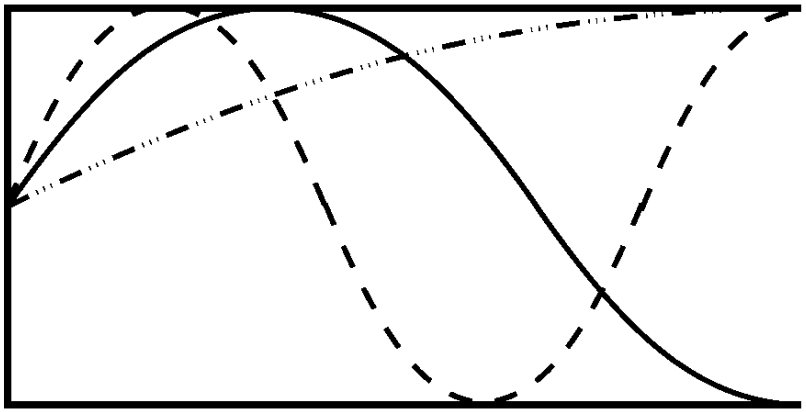


$$\lambda = 4 \times 17.5\ \text{cm} = 70\ \text{cm} = 0.7\text{m}$$

$$f = \frac{c}{\lambda} = \frac{350\ \text{ms}^{-1}}{0.7\ \text{m}} = 500\ \text{Hz}$$

Multiple resonant frequencies in a single tube

- There are other resonances of the uniform tube



- The other wavelengths that fit into this tube have wavelengths 1/3, 1/5, … of the longest wavelength
- So they have frequencies 3, 5, … times the lowest frequency.
- i.e. 1500Hz, 2500Hz, …

Speech production – sound source

- Simple breathing does not produce speech
- Need a source of sound energy
- Vocal folds (vocal chords)
- Make some vowel sounds
  - feel your vocal folds vibrating
  - and feel the airflow coming out of your mouth.
- Air flowing through the glottis (the space between the vocal folds) makes them vibrate: we call this VOICING
  - homework: find some online videos of the vocal folds
- Can you make sounds without using your vocal folds?

Speech production – other sounds

- Make some unvoiced sounds
- Vocal folds are not vibrating
- Still airflow out of mouth though

- Where is the sound source now?

- Make some nasals (/n/ and /m/ for example)
- What are your vocal folds doing?
- Is there airflow out of your mouth?

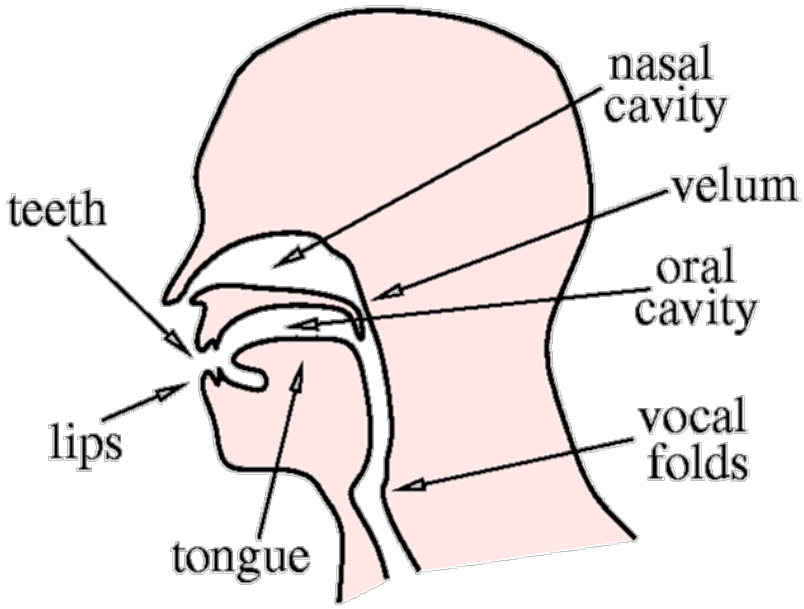
Speech production apparatus

- Make the vowels in the following English words:

- Bard
- Bead
- Boot

- What controls the difference between the vowels?

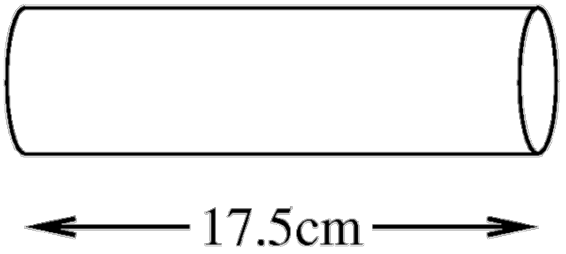
Articulators



What makes vowel sounds different from one another?

The neutral vowel – schwa

- With the articulators in the relaxed, neutral position, we get the vowel schwa
- We can model this as a simple tube, length 17.5 cm. We already saw that the fundamental wavelength for this tube is: λ=0.7m, f=500Hz

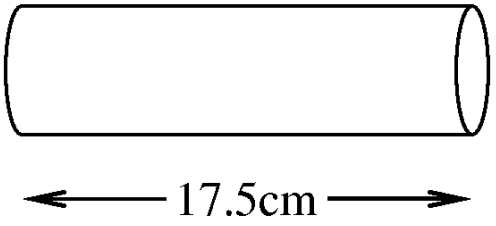


The other wavelengths that fit into this tube have frequencies 3, 5, … times the fundamental, i.e., 1500Hz, 2500Hz, …

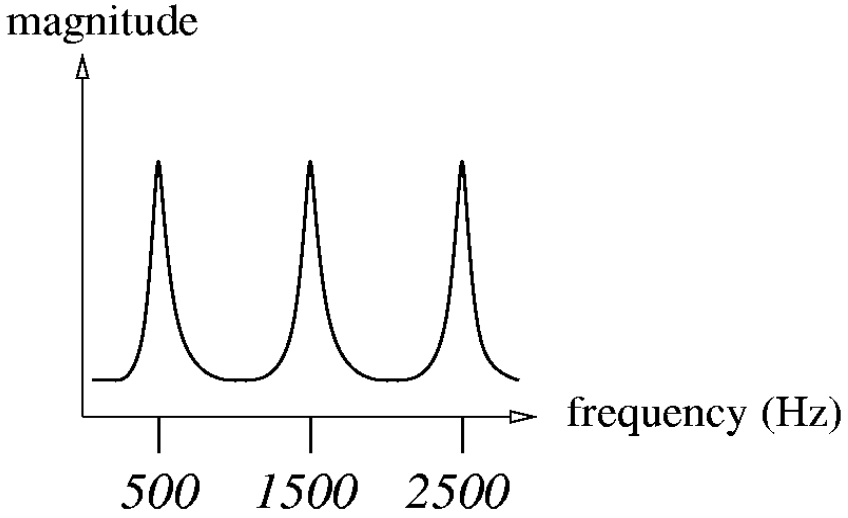
Our model predicts that the first three formants of schwa are 500Hz, 1500Hz and 2500Hz

From tube length to frequency response

- The frequency response we calculated for this simple tube

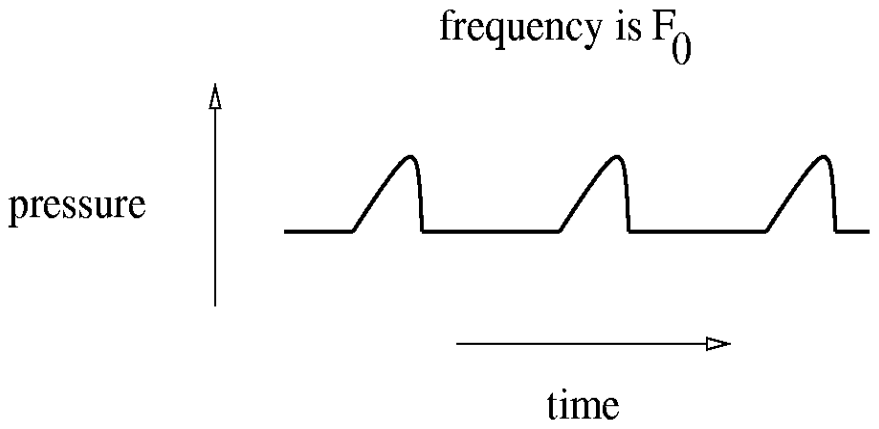


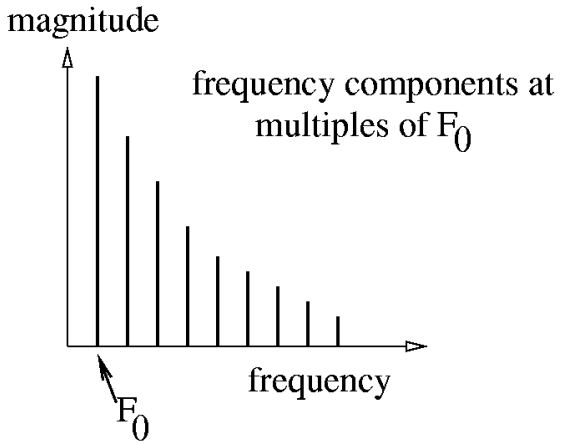
- looks like this



The glottal pressure wave

- Contains energy at frequency F_0
  - and at every multiple of F_0: 2 x F_0, 3 x F_0, 4 x F_0, and so on
- These are the harmonics of F_0



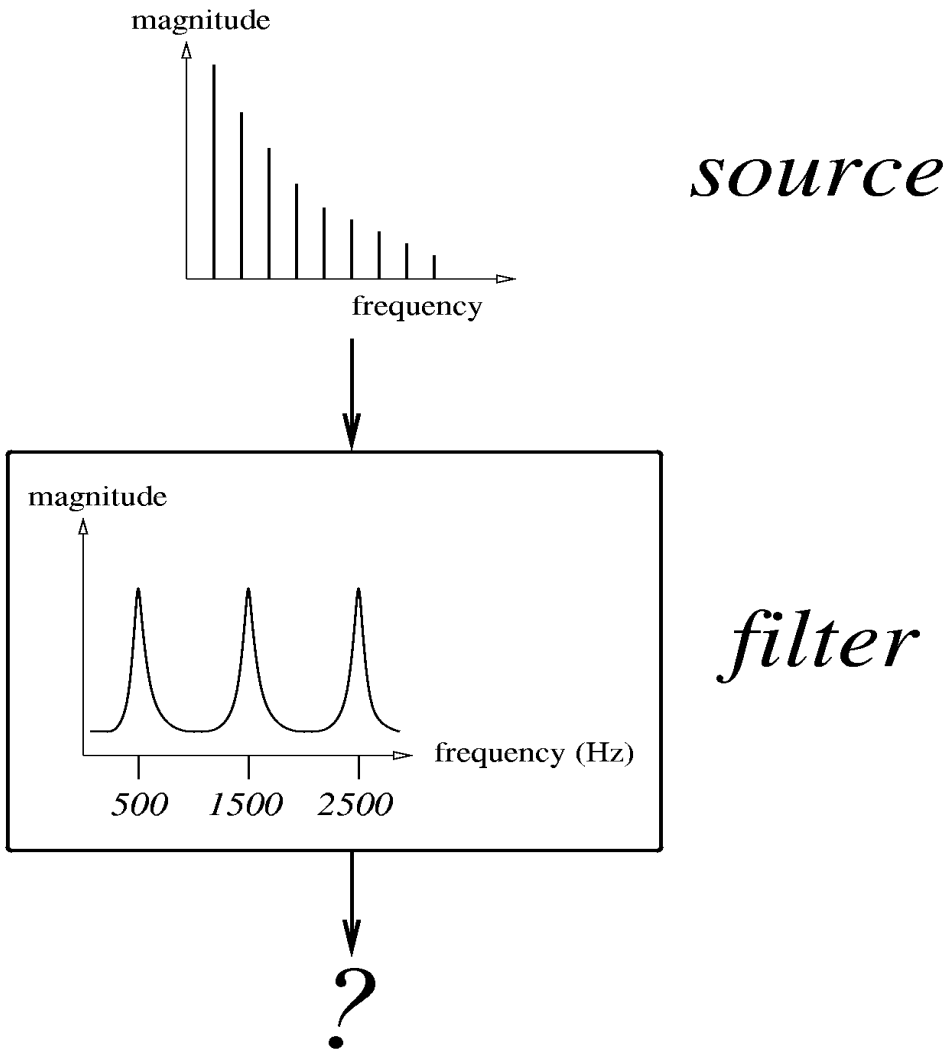


Putting the source and filter together

- We know the frequency response of the filter
  - It has peak corresponding to the resonances
  - Positions of the peaks depend on tube configuration
  - Which depends on the articulator positions
- We know the spectrum of the sound wave generated by the vocal folds
  - It has energy at F_0 and every multiple of F_0

- So .... what is the spectrum of a speech signal?

Spectrum of schwa

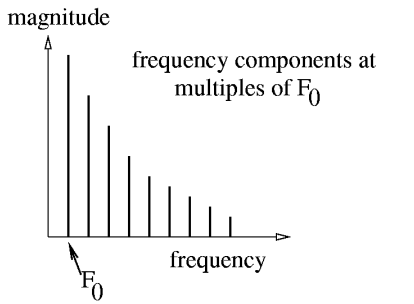


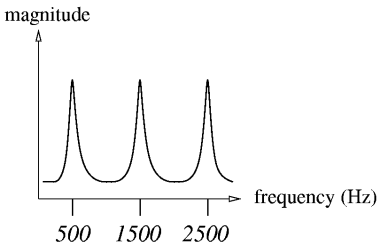
Multiply the spectra

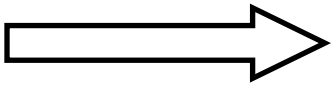
- The effect the filter has on the input signal
  - is linear, which means we can consider each frequency in the spectrum independently
- For speech, this means that the vocal tract
  - affects each harmonic of F0 independently of the other harmonics
  - can only reduce or increase the amplitude of each harmonic
  - it cannot move the frequency of a harmonic
  - it cannot add energy at new frequencies

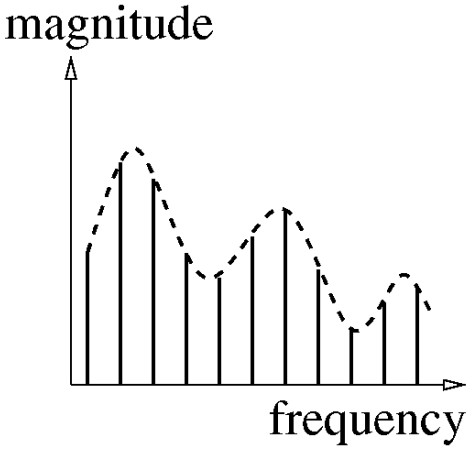
Spectrum of a vowel

- Overall shape (“envelope”) has peaks
  - Due to the vocal tract frequency response
- Fine structure is the harmonics of F_0
  - Due to the source (vocal folds)



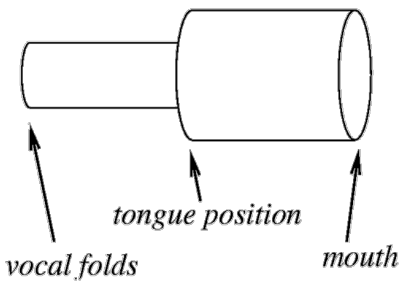


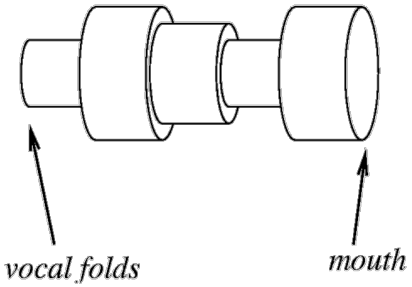




Vocal tract – more complex models

- Vocal tract is not always a simple tube
- The articulators vary its shape
- We can use more complex models:

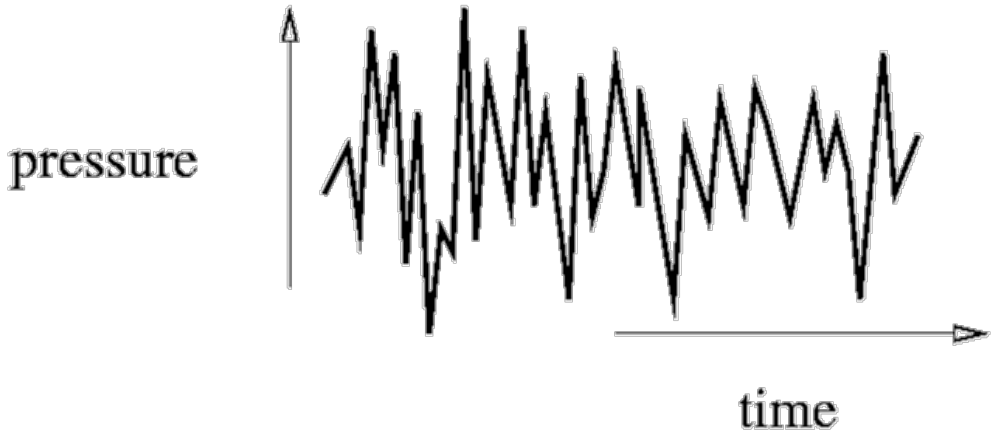




The resonance patterns depend on lengths of the different tubes, and to an extent, the interaction between tubes

Turbulence and fricatives

- For unvoiced speech, the source is turbulent airflow:
- E.g., /s/ /f/ /S/
- What controls which fricative is produced?

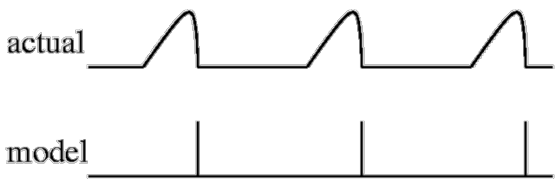


Putting this all together in terms of speech production

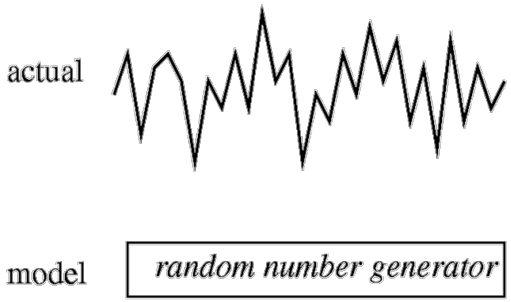
- Vocal folds open and close abruptly
- Produce a sound wave containing many different frequencies (all multiples of some fundamental frequency)
- This signal passes through the vocal tract
- Certain frequencies are amplified by the vocal tract resonances
- Vocal tract resonances are called formant frequencies or simply formants

Modelling speech waveforms

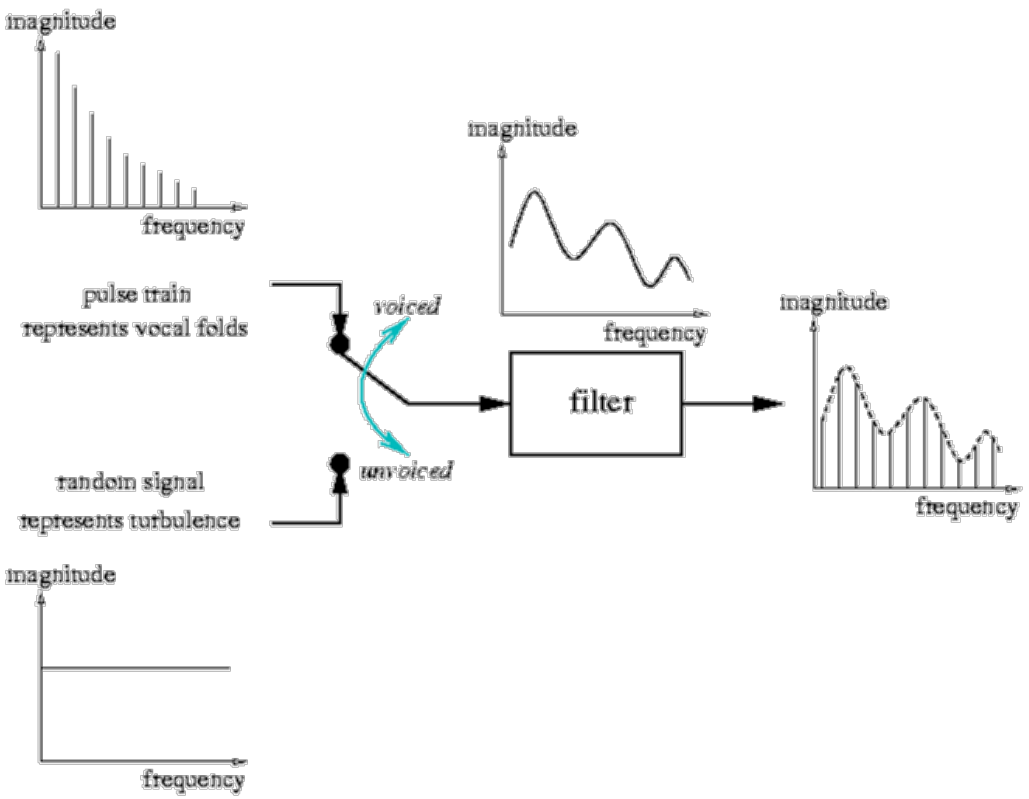
- Vocal folds



- Frication



The source-filter model



What can we do with the source filter model

- There are algorithms for determining filter parameters from the speech signal
- Calculate vocal tract shape and use this information in
  - phonetics research
  - speech therapy
- Obtain a smooth spectrum free from the effects of F_0
- Separate the source from the filter
  - then modify each independently (speech synthesis, speech modification)
  - automatic speech recognition uses only the filter shape (for non-tone languages)

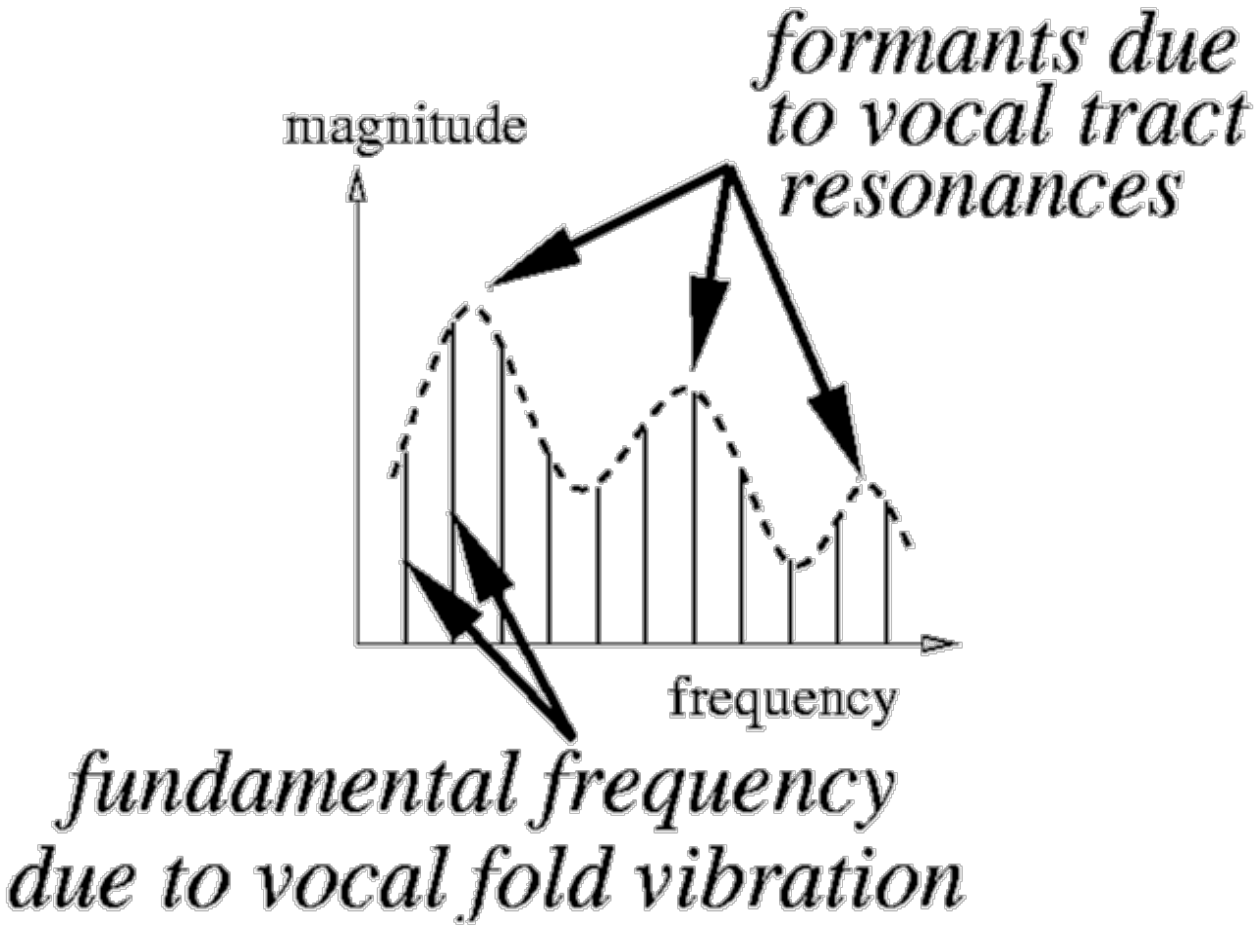
Synthesis with a source-filter model

- Need to control pitch (source) independently of segment identity
  - Varying the source frequency with a fixed filter allows us to control pitch
- Need to control duration
  - We can stretch segments without changing the pitch
- The source-filter model can give us independent control over:
  - pitch
  - duration
  - segment identity

F_0, pitch, formants: some clarification

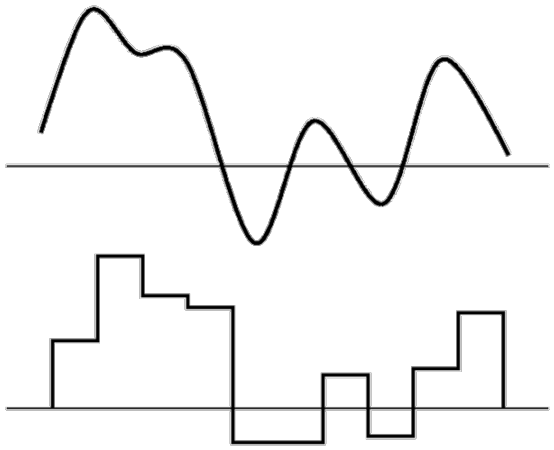
- Fundamental frequency (written as F_0, F0, f0, ...)
  - frequency at which the vocal folds vibrate
  - perceived as pitch
  - (think of different musical notes)
- Formants (called F1, F2, F3, …)
  - resonances of the vocal tract
  - the main cues to which sound we perceive (for vowels, at least)
  - (think of different shaped musical instruments)
- F0 is a different type of thing to a formant
  - don’t be confused by the notation (F0, F1, F2,…)

F_0 vs. formants



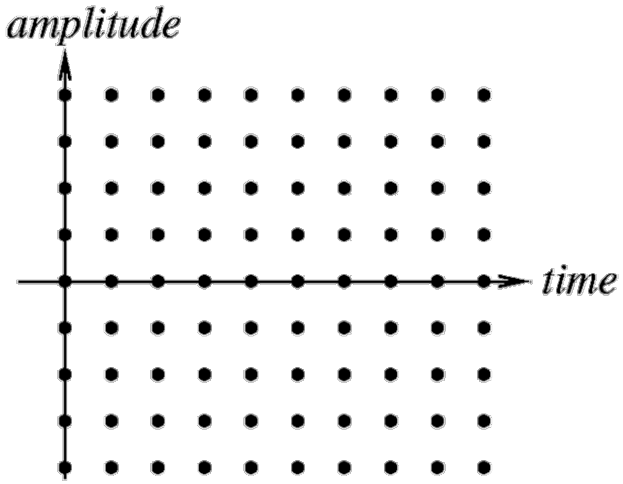
Concepts: sampling and quantisation

- To represent sound pressure waves in the computer, we need to convert continuous values to discrete (digital) representations
  - Time axis: the sound pressure is sampled at fixed intervals (thousands of times per second)
  - Vertical axis: continuous value (representing sound pressure) is encoded as one of a fixed number of discrete levels



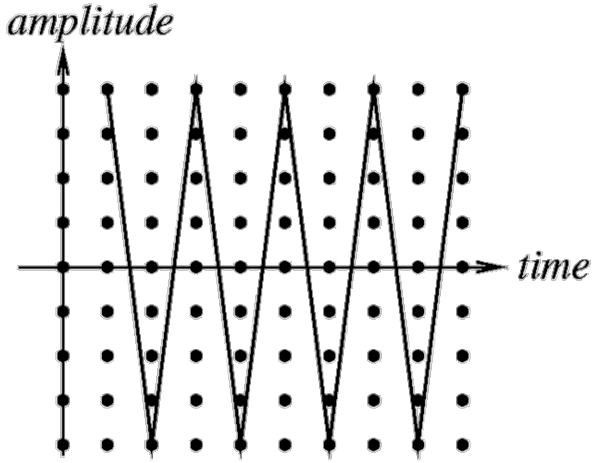
The effect of sampling

- The grid below represents the resolution at which we can sample.
- What is the highest frequency waveform that we can draw using only the available points?



The Nyquist frequency

- Definition: The sampling frequency is the number of times per second we record the value of the waveform



We can only represent frequencies up to half the sampling frequency. This is called the Nyquist frequency.

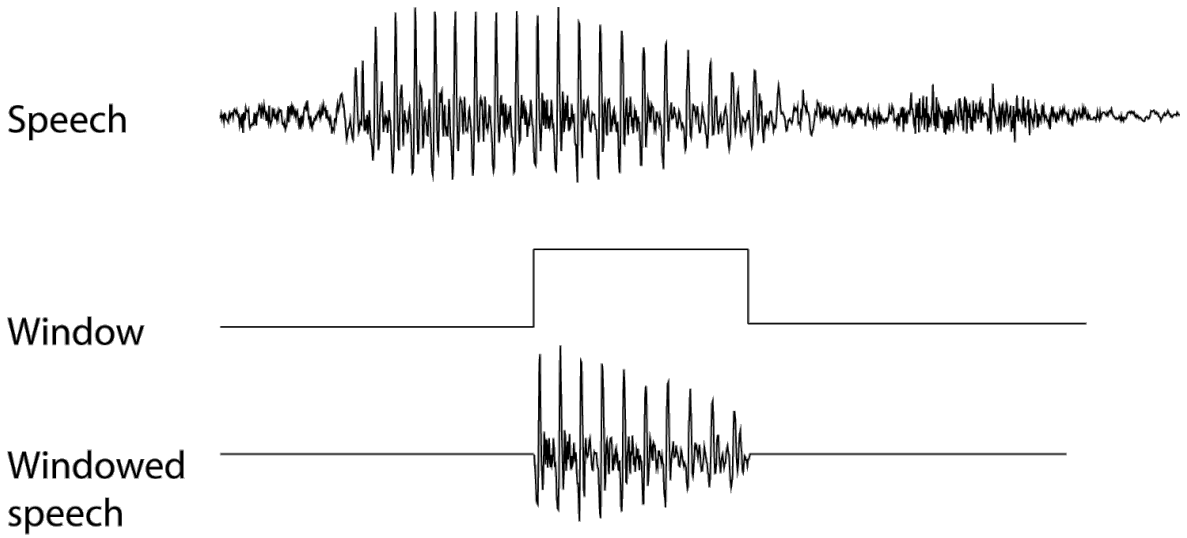
Sampling rates and bit depth

- To capture frequencies up to 8kHz we must sample at (a minimum of) 16kHz.
- CDs use a 44.1kHz sampling rate.
- Current studio equipment records at 48, 96 or 192 kHz

- Each sample is represented as a binary number
- Number of bits in this number determines number of different amplitude levels we can represent
- Most common bit depth is 16 bits
  - $2^{16} = 65536$

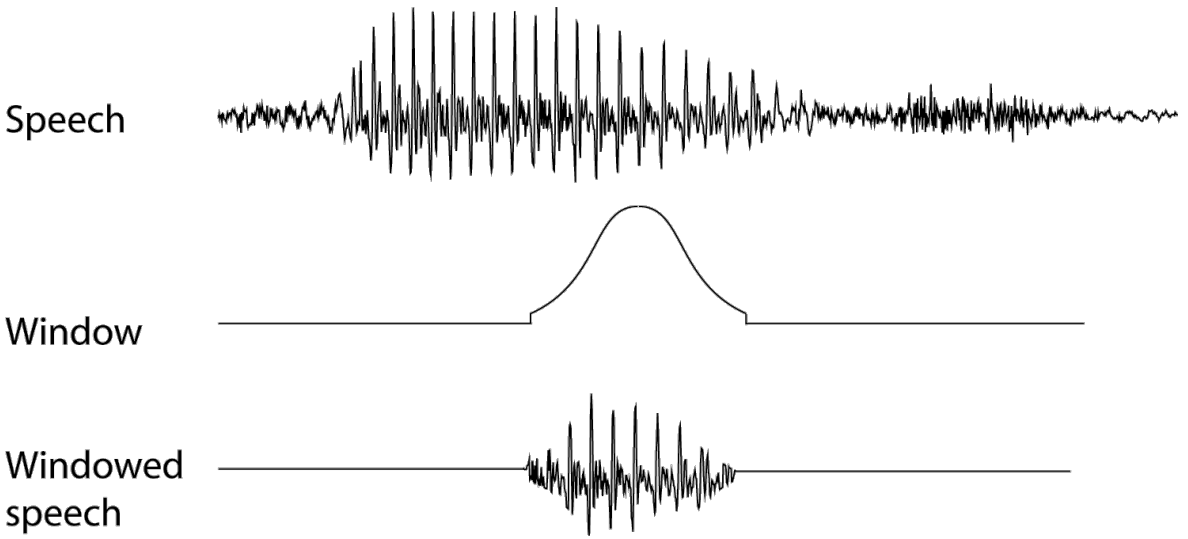
Short-term analysis, frames and windowing

- Most analysis techniques operate on short regions of speech, because they must assume properties (F_0, formants, etc) are constant over this duration
- The simplest approach is to simply cut out the bit of speech we want to analyse, like this



Hamming Window

- The rectangular window can create artefacts because of the abrupt starting and stopping of the signal
- There are various better types of windows, which smooth the edges, like this



Time domain and frequency domain

- A sound signal can be represented in either the time domain or the frequency domain
  - Waveform and spectrum, respectively

- A filter is probably easier to think about in the frequency domain, but it can also be represented in the time domain
  - Frequency response and impulse response, respectively