CSC2518 – Spoken Language Processing – Fall 2014

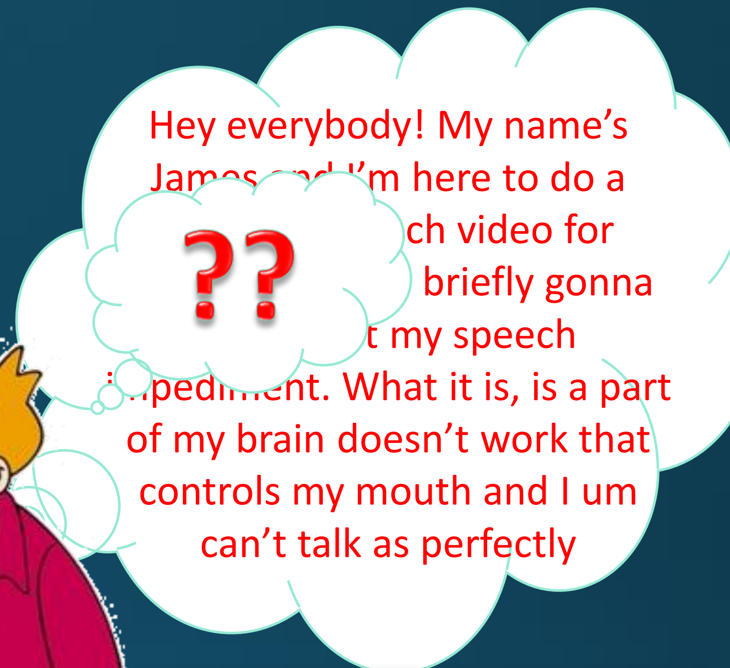
Lecture 1 Frank Rudzicz

University of Toronto

Speech in healthcare

Dysarthria

Neuro-motor articulatory disorders resulting in **unintelligible** speech.



7.5 million Americans have **dysarthria**

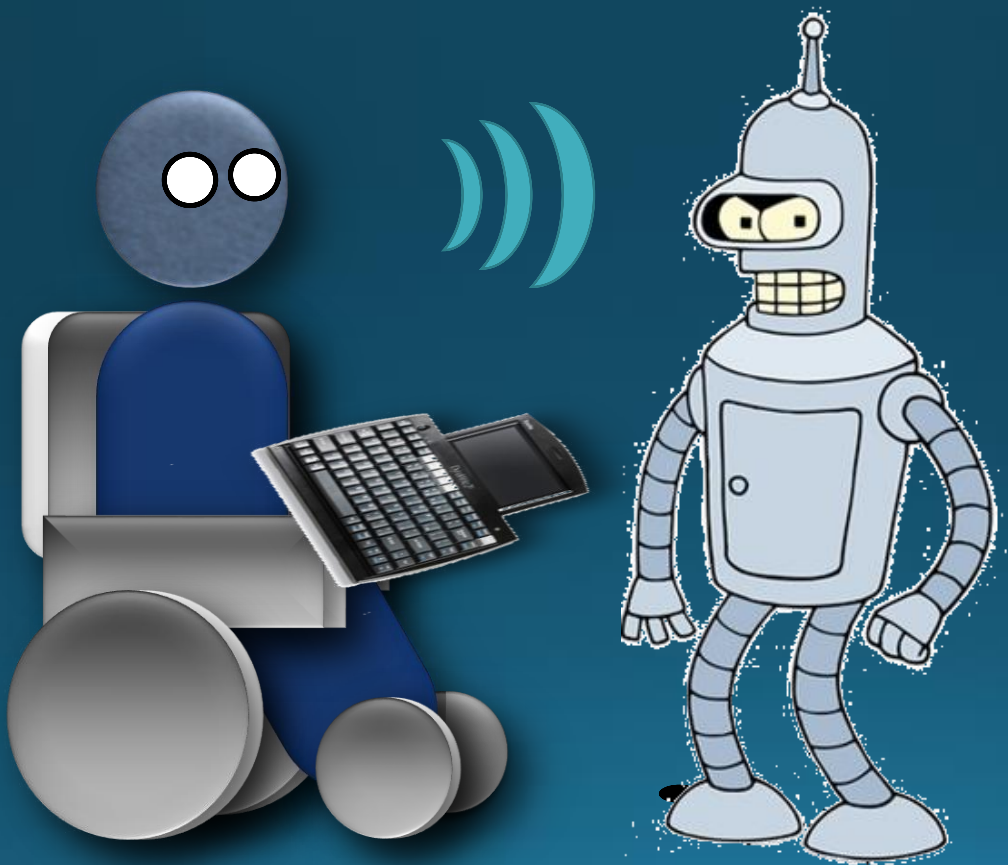
- Cerebral palsy,
- Parkinson's,
- Amyotrophic lateral sclerosis)

(National Institute of Health)

UNIVERSITY OF TORONTO

Dysarthria

The **broader** neuro-motor deficits associated with dysarthria can make **traditional** human-computer interaction difficult.



Can we use ASR for dysarthria?

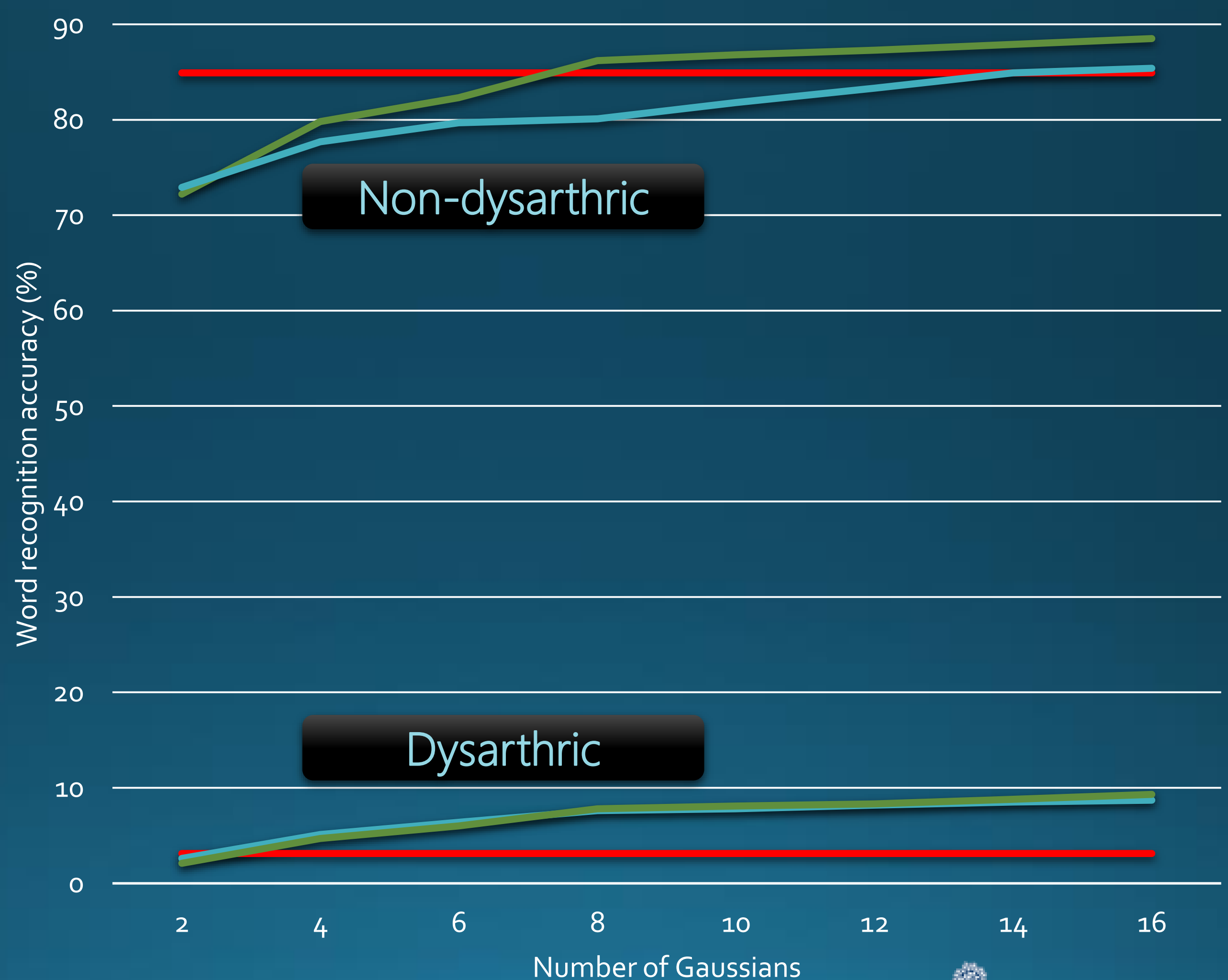
UNIVERSITY OF TORONTO

Adjusting to the individual

84.9%

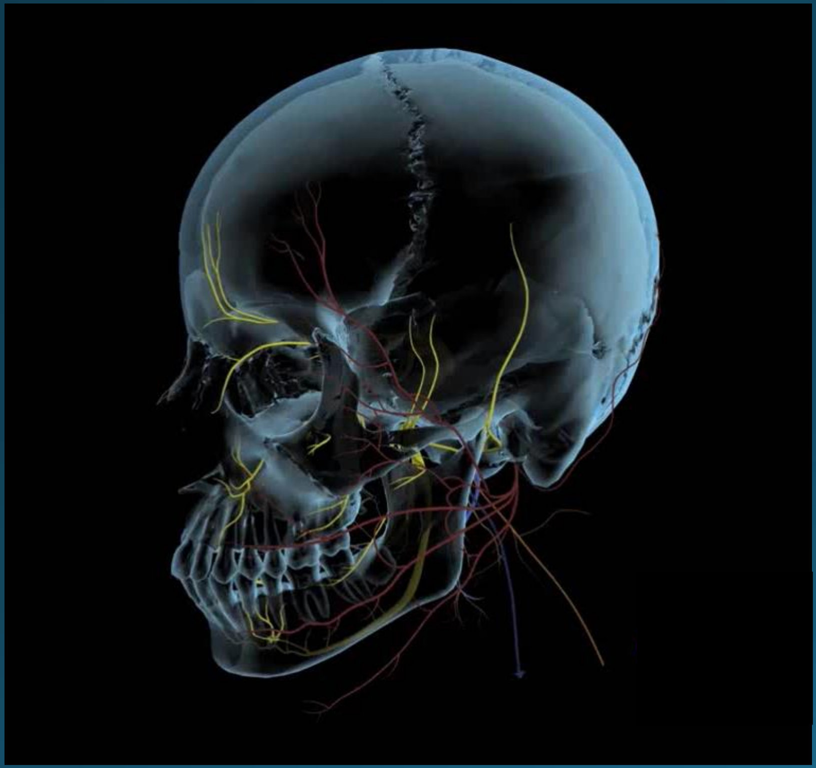
Traditional ASR
Speaker-dependent
Speaker-retrained

3.1%



Neural origins

- **Types** of dysarthria are related to **specific sites** in the subcortical nervous system.

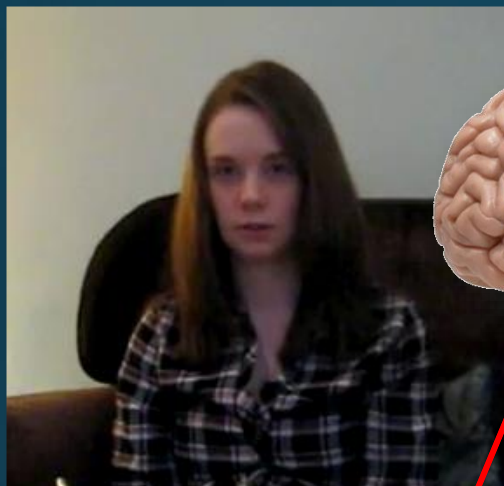


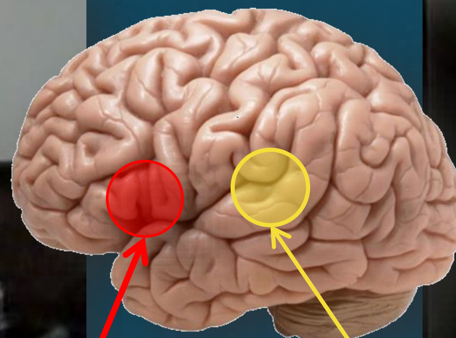
Type	Primary lesion site
Ataxic	Cerebellum or its outflow pathways
Flaccid	Lower motor neuron (≥1 cranial nerves)
Hypo-kinetic	Basal ganglia (esp. substantia nigra)
Hyper-kinetic	Basal ganglia (esp. putamen or caudate)
Spastic	Upper motor neuron
Spastic-flaccid	Both upper and lower motor neurons

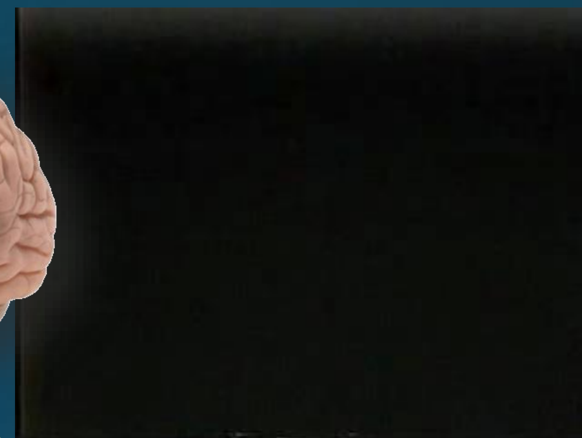
(After Darley *et al.*, 1969)

UNIVERSITY OF TORONTO

Deeper into the brain – Aphasia







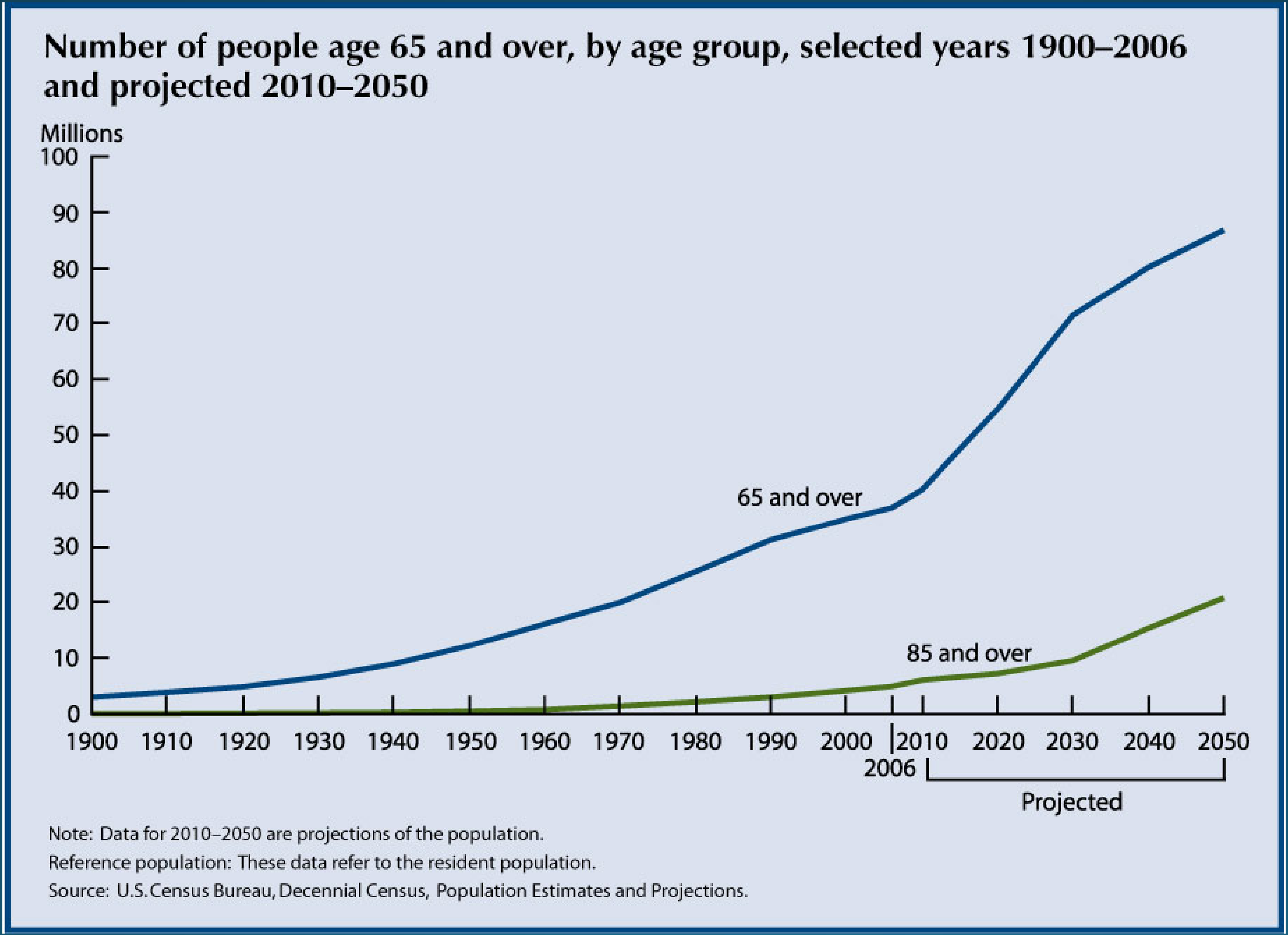
Broca's aphasia

Wernicke's aphasia

- **Reduced** hierarchical **syntax**.
- **Anomia**.
- **Reduced** "mirroring" between **observation** and **execution** of **gestures** (Rizzolatti & Arbib, 1998).

- **Normal** intonation/rhythm.
- **Meaningless** words.
- '**Jumbled**' syntax.
- **Reduced** comprehension.

Demographics



A future for speech diagnostics

- **Speech-language pathologists**: ~150,000 in USA.
  - This labour market is **growing faster** than the average and has recurrent software needs (Bureau of Labor Statistics, 2011).
- Between **8% and 10%** of the US population has some form of speech/language/hearing disorder (National Institute of Health).
  - This is increasing with the age of the population and the incidence of **stroke** and **dementia**.
- Caregivers often assist individuals with Alzheimer's disease (AD), either at home or in long-term care facilities.
  - >$100B are spent annually in the U.S. on caregiving for AD.
  - As the population **ages**, the incidence of **AD** may double or triple in the next decade (Bharucha *et al.*, 2009).
  - **Demographic crisis!**

Broad syllabus

- **Theme**: speech-based technology in healthcare.

- Automatic speech recognition in healthcare
  - E.g., dictation of medical records.
- Speech-based communication aids
  - E.g., synthetic speech, brain-computer interfaces.
- Speech-based diagnosis and monitoring
  - E.g., Parkinson's, post-stroke aphasia, cerebral palsy
- Clinically-relevant features, brains, et c.

Lecture 1

1. The nature of the course
   - (20%) **Participation**: 60 minutes of conference-style presentations.
   - (80%) A final course **project.**

2. Crash course in speech signal processing

3. Crash course in automatic speech recognition

Please subscribe to csc2518_2014 Google Group!

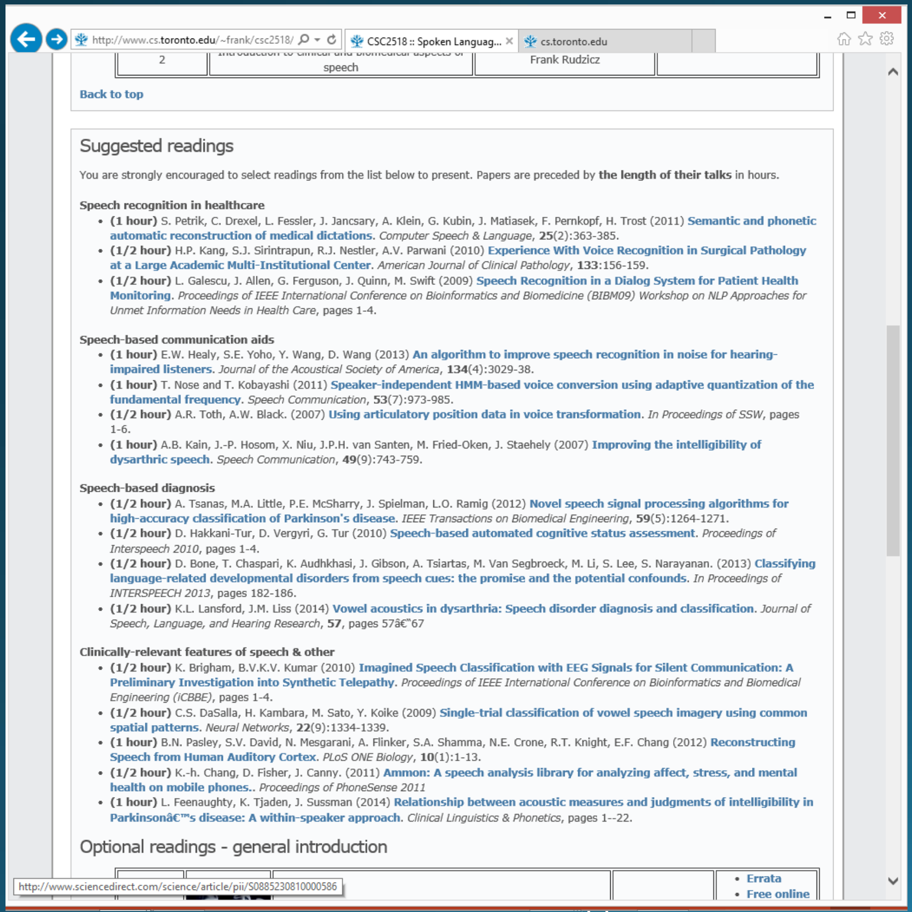
- Next week: Clinical/biomedical aspects of speech

1.Conference-style presentations

- Every student will deliver **conference-style presentations** for 60 minutes, either:
  - **Two papers, 30 minutes each** (25 min talk + 5 minute questions). Typically papers in conference proceedings, or
  - **One paper, 60 minutes** (50 min talk + 10 minute questions). Typically journal articles.
- Presentations can follow the structure of the paper, but should include a broad **overview**, **scientific context** (i.e., literature review), **methodology**, empirical **results**, and a **summary** of contributions.
- Though informal, students should be prepared to answer questions (15% of participation grade).

1.Conference-style presentations

- Students should **select papers** from the course website.



- **First come, first served**. **Email me** to volunteer for next available slot, then choose any of the remaining papers.
- Your lecture (+ any supplemental) materials will be **posted**.
- You should **meet with me** ≥1 week before your talk to go over your slides.

UNIVERSITY OF TORONTO

2. Project

- Get two birds stoned at once: get an A+ and a publication.

- Final report takes the form of a paper conforming to:

  - Transactions of the Association for Computational Linguistics
  - Interspeech
  - Neural Information Processing Systems

- Your report will be marked on 1) originality, 2) sufficient survey of existing work, 3) technical correctness, 4) empirical methods, 5) overall presentation.

2. Project

- Four components:
  - **Project proposal** (22 September). 5% of project grade. 1-2 pages.
    - Describe your goals.
    - Briefly describe 2-5 relevant papers.
    - Outline your plan to reach your goals (including schedule).
    - Outline a method to evaluate success.
  - **Midterm checkpoint** (mid-to-late October). Not marked.
    - You will meet with me to discuss progress.
  - **Project report** (15 December). 80% of final grade. ≥ 4 *tight, double-column* pages or equivalent.
    - You will be encouraged to submit this to a conference or journal.
  - ***N*-minute madness**. (15 December). 5% of project grade.
    - You will present your project for a brief $2 \leq N \leq 5$ minutes.

2. Project – data (e.g.)

- General speech:
  - **Switchboard**: telephone conversations, 8 kHz, 14 GB
  - **TIMIT**: phonemically-balanced, 16 kHz, 711 MB
  - **WSJ**: news broadcasts, 15 GB
- Pathological speech:
  - **DementiaBank**: dementia (and control), picture description, 13 GB
  - **TORGO**: cerebral palsy, articulation, phonemically balanced, 18 GB
  - **TBD**: Parkinson's disease, articulation, emotional speech, TBD
- EEG/MEG, robot
- …

UNIVERSITY OF TORONTO

Flexibility

- You can choose to **present a paper** other than those on the course page.

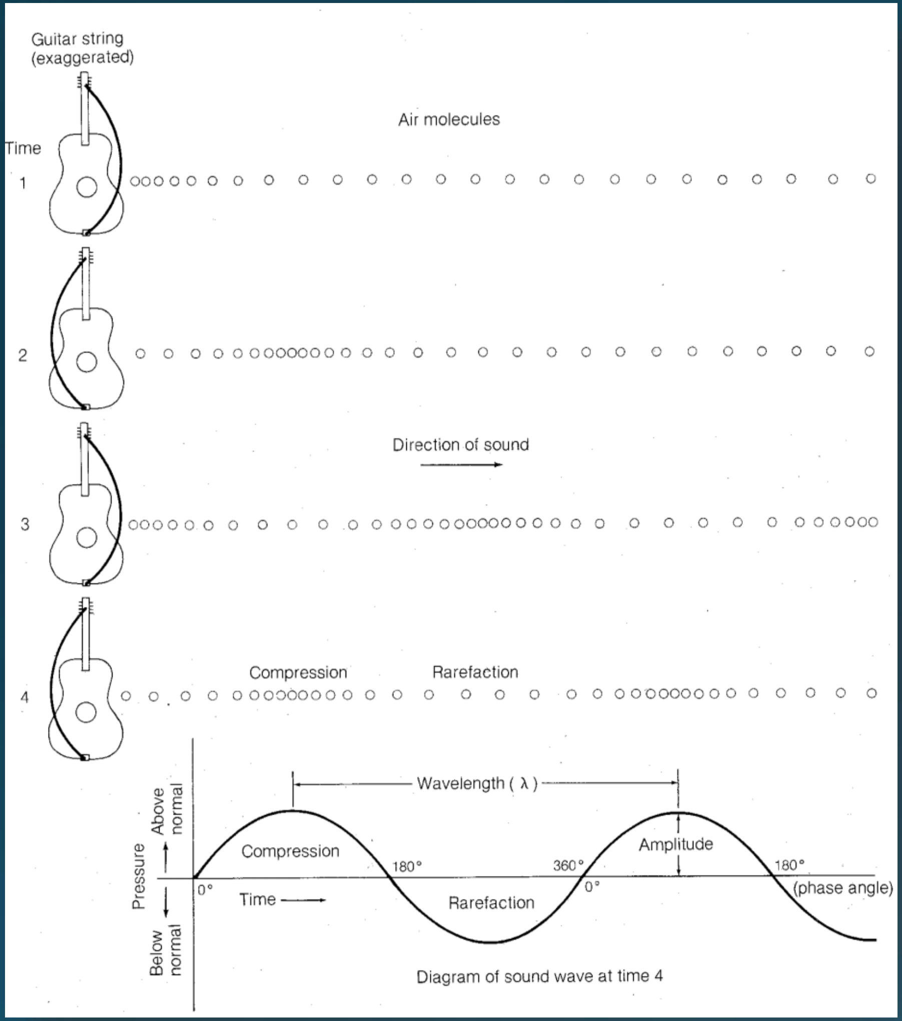
- You can choose to write in the **style** of another journal or conference.

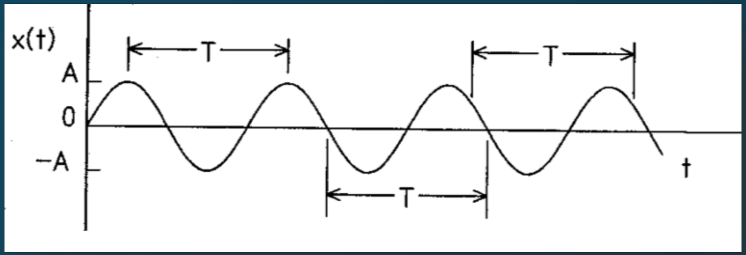
- You can choose to use **another set of data,** or **collect your own**.

*In all cases, **consult with me ASAFP***

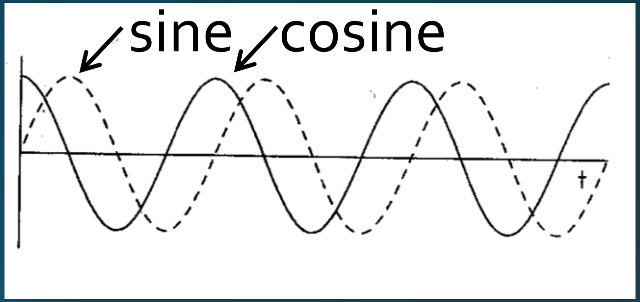
UNIVERSITY OF TORONTO

Sound signals





- **Frequency** $F = 1/T$

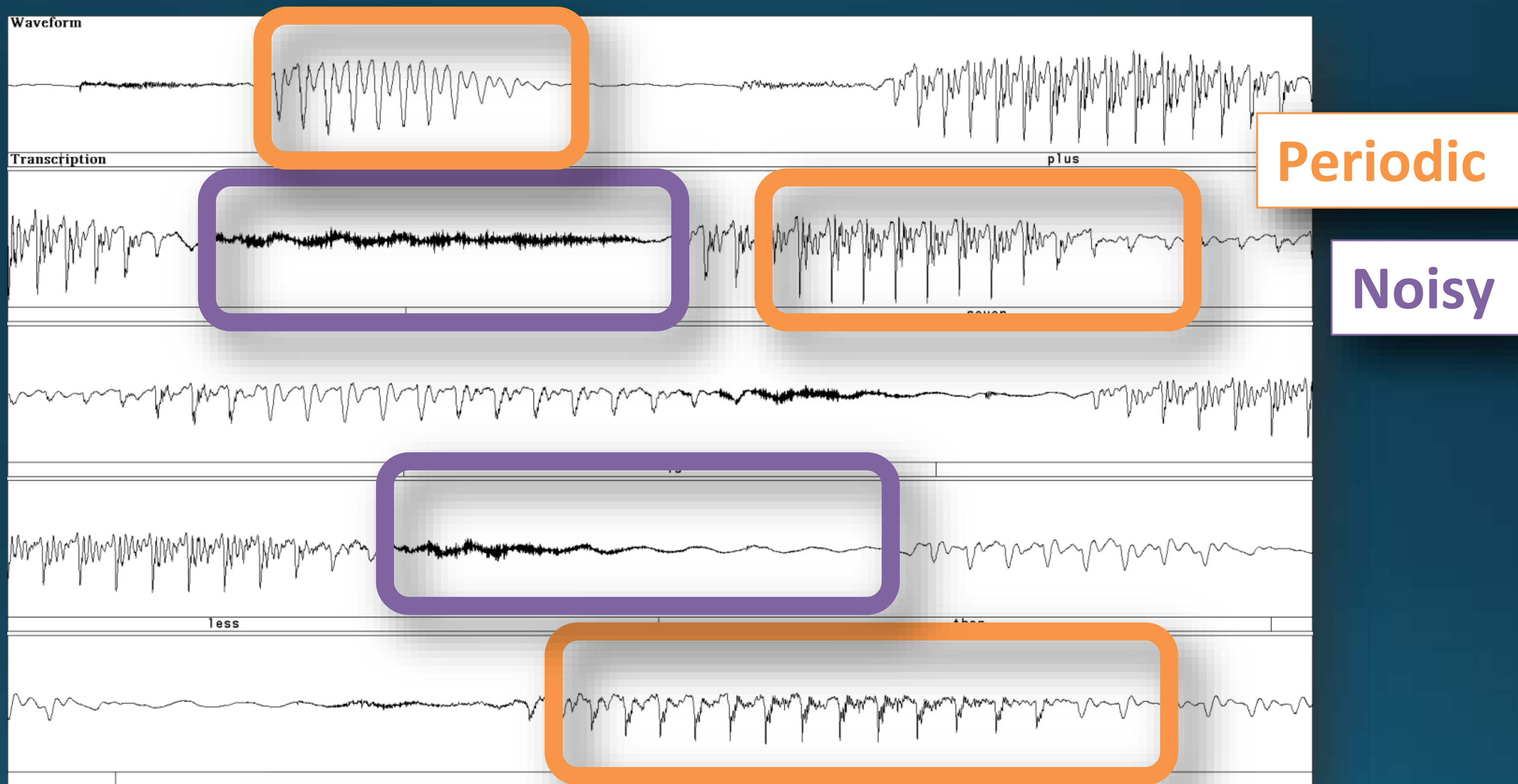


Given $\omega = 2\pi/T$, and **phase** $\phi = \pi/2$,

$$f(t) = A\sin(\omega t + \phi) = A\cos(\omega t)$$

UNIVERSITY OF TORONTO

Speech signals

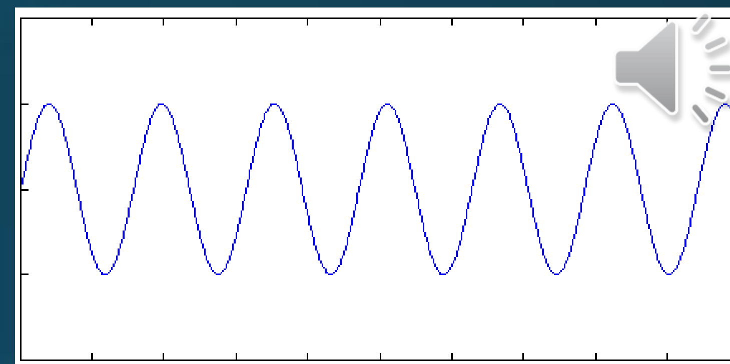


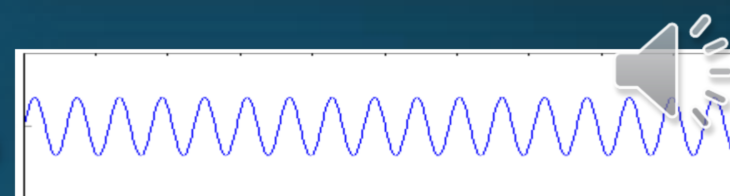
“Two plus seven is less than ten”

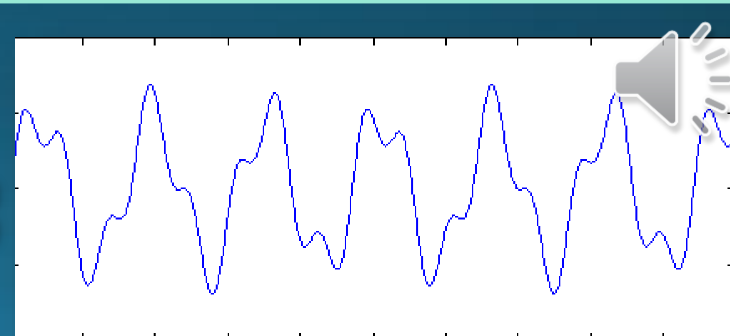
UNIVERSITY OF TORONTO

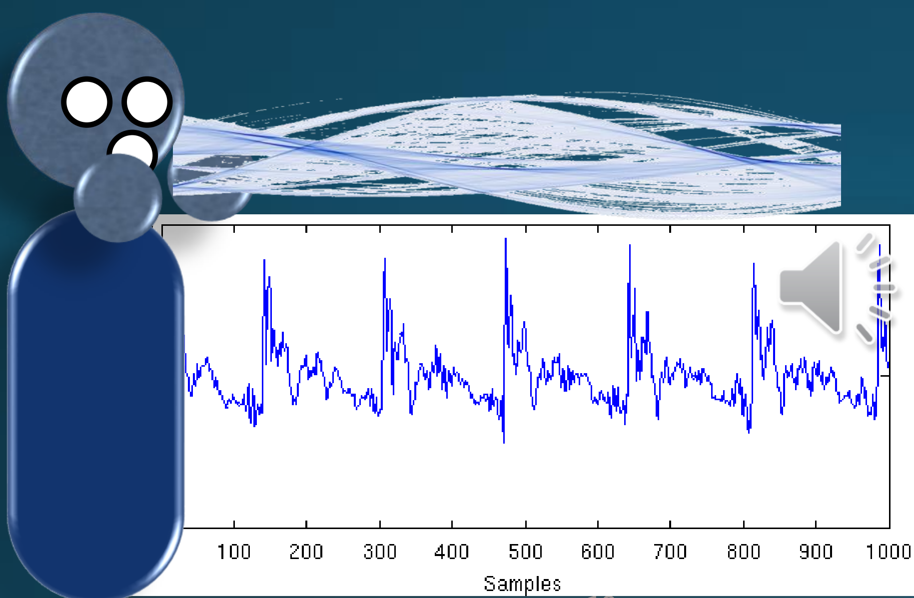
Signals as summed sinusoids

- Consider just the **periodic** segments.
- Fourier: $f(t) = \sum_{i=0}^{\infty} w_i f_i(t)$
  - Especially nice: $f_i(t) = \sin(\omega_i t + \phi_i)$

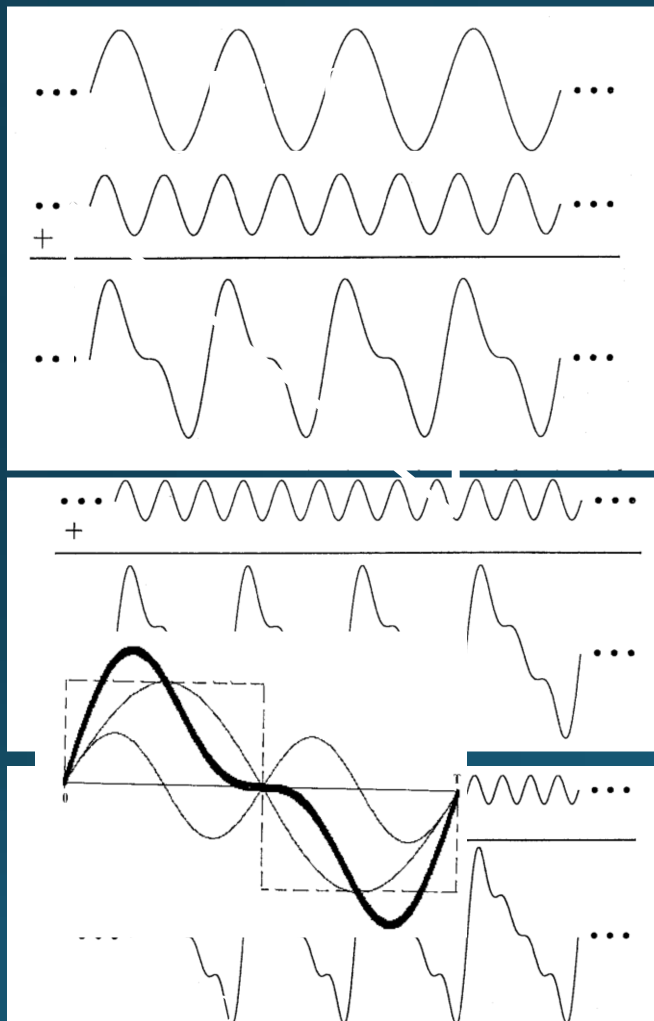


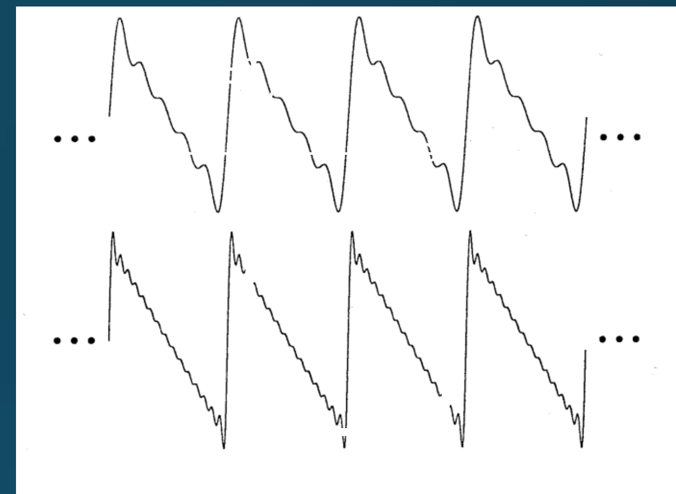




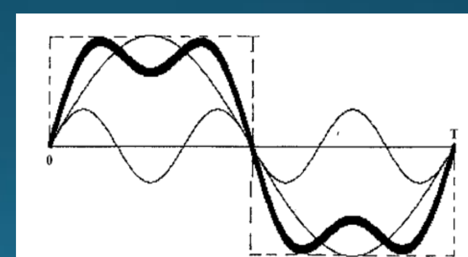


Signals as summed sinusoids





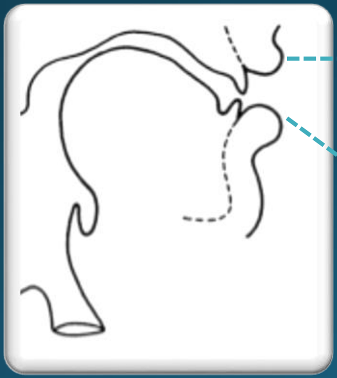
Et c. *ad infinitum*

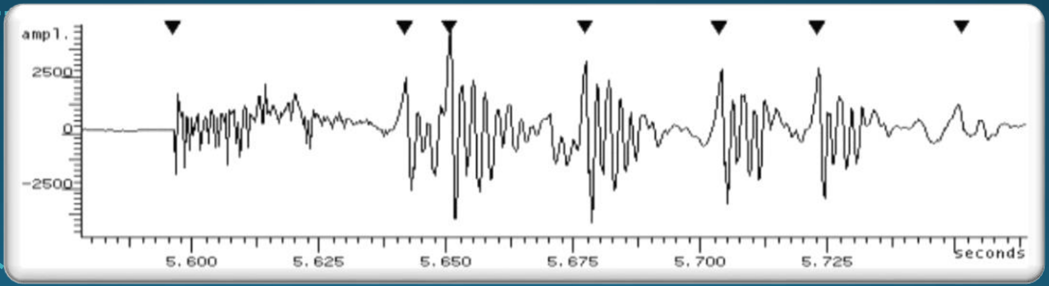


Et c. …

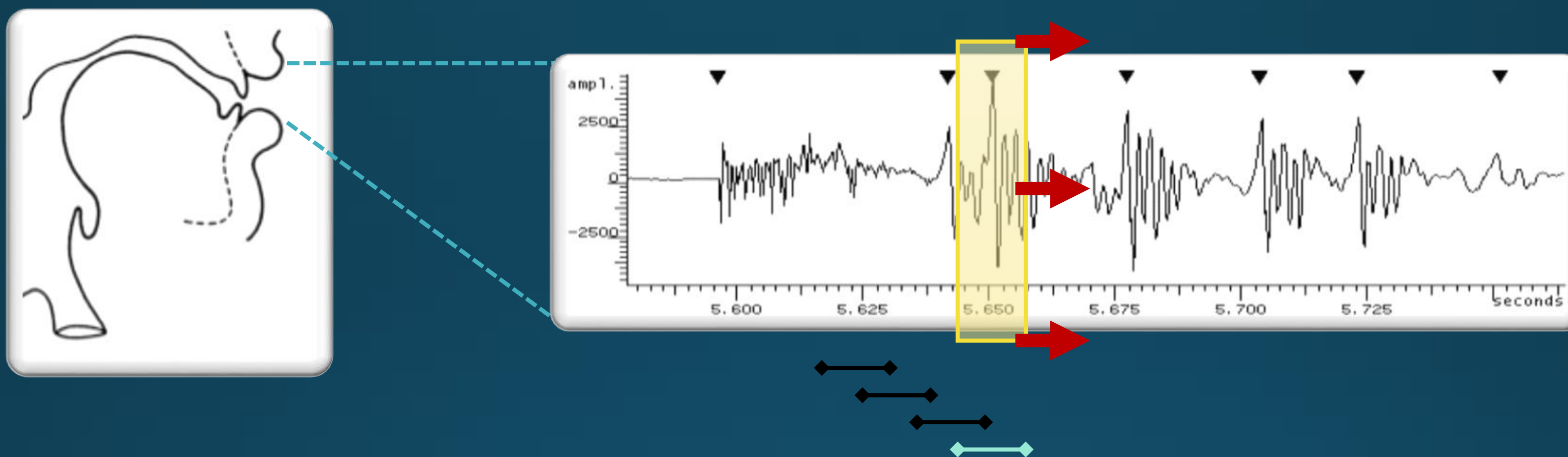
Extracting sinusoids from waves

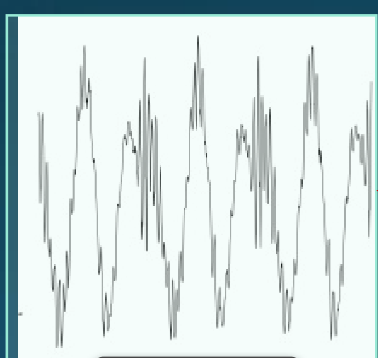
- As we will soon see, the relative **amplitudes** and **frequencies** of the sinusoids that combine in speech are often **extremely indicative** of the **phoneme** being uttered.
  - ∴ If we could **separate** the waveform into its component sinusoids, it would help us **classify** phonemes being uttered.





Short-time windowing

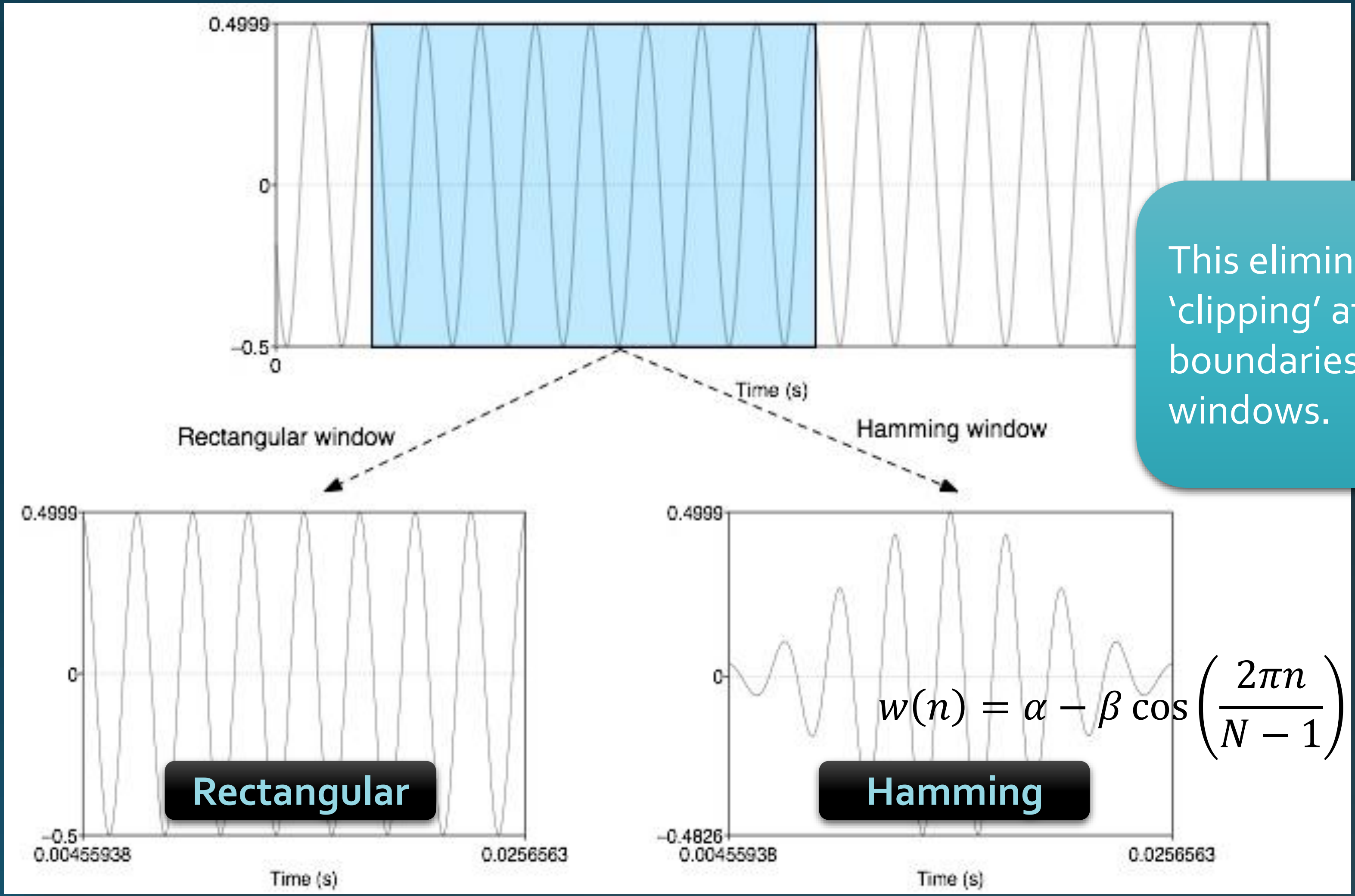




Frame

- Speech waveforms **change** drastically in time.
- We **move** a short analysis **window** (*assumed to be time-invariant*) across the waveform in time.
  - E.g. frame shift: 5—10 ms
  - E.g. frame length: 10—25 ms

Window types



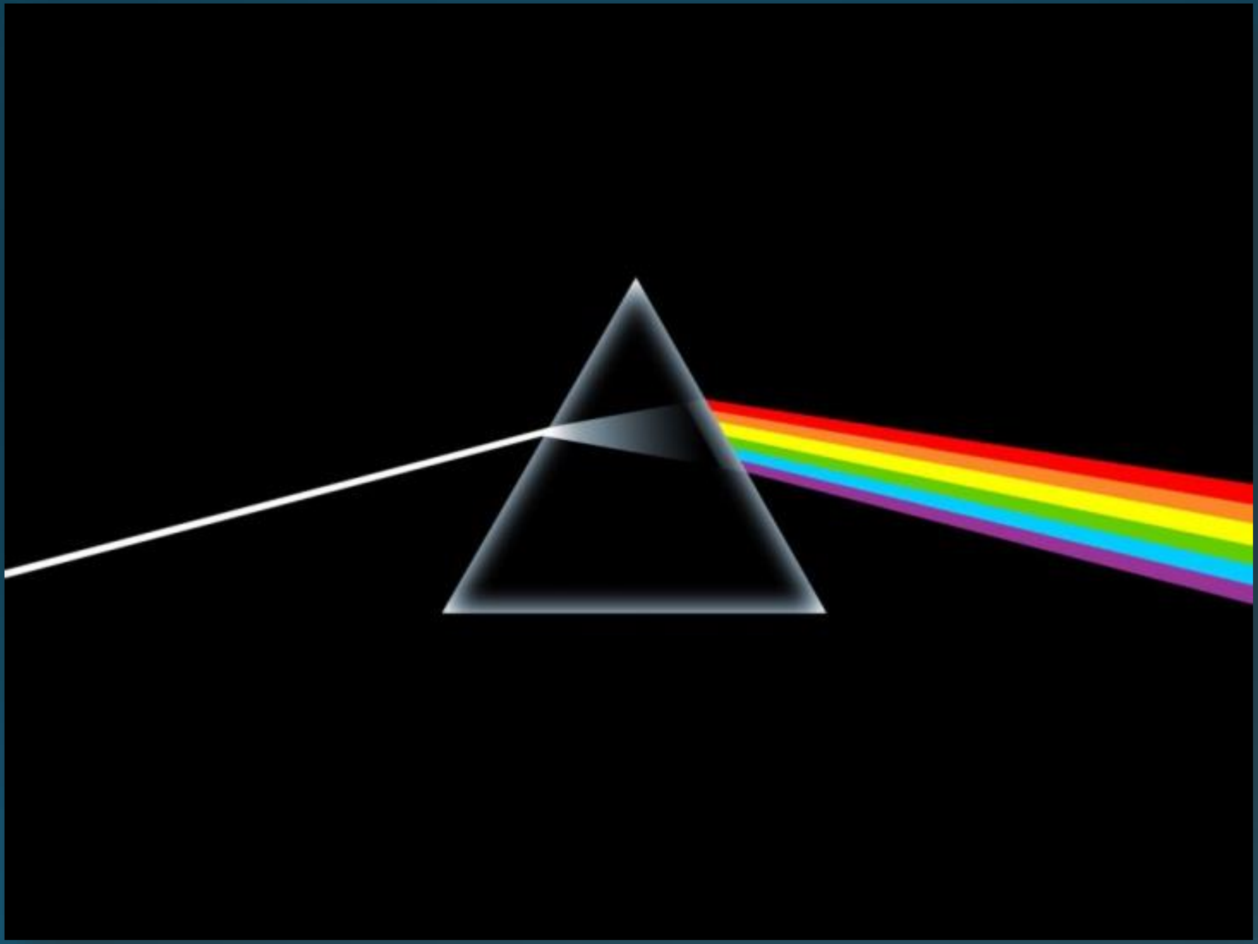
$$w(n) = \alpha - \beta \cos\left(\frac{2\pi n}{N-1}\right)$$

This eliminates 'clipping' at the boundaries of windows.

UNIVERSITY OF TORONTO

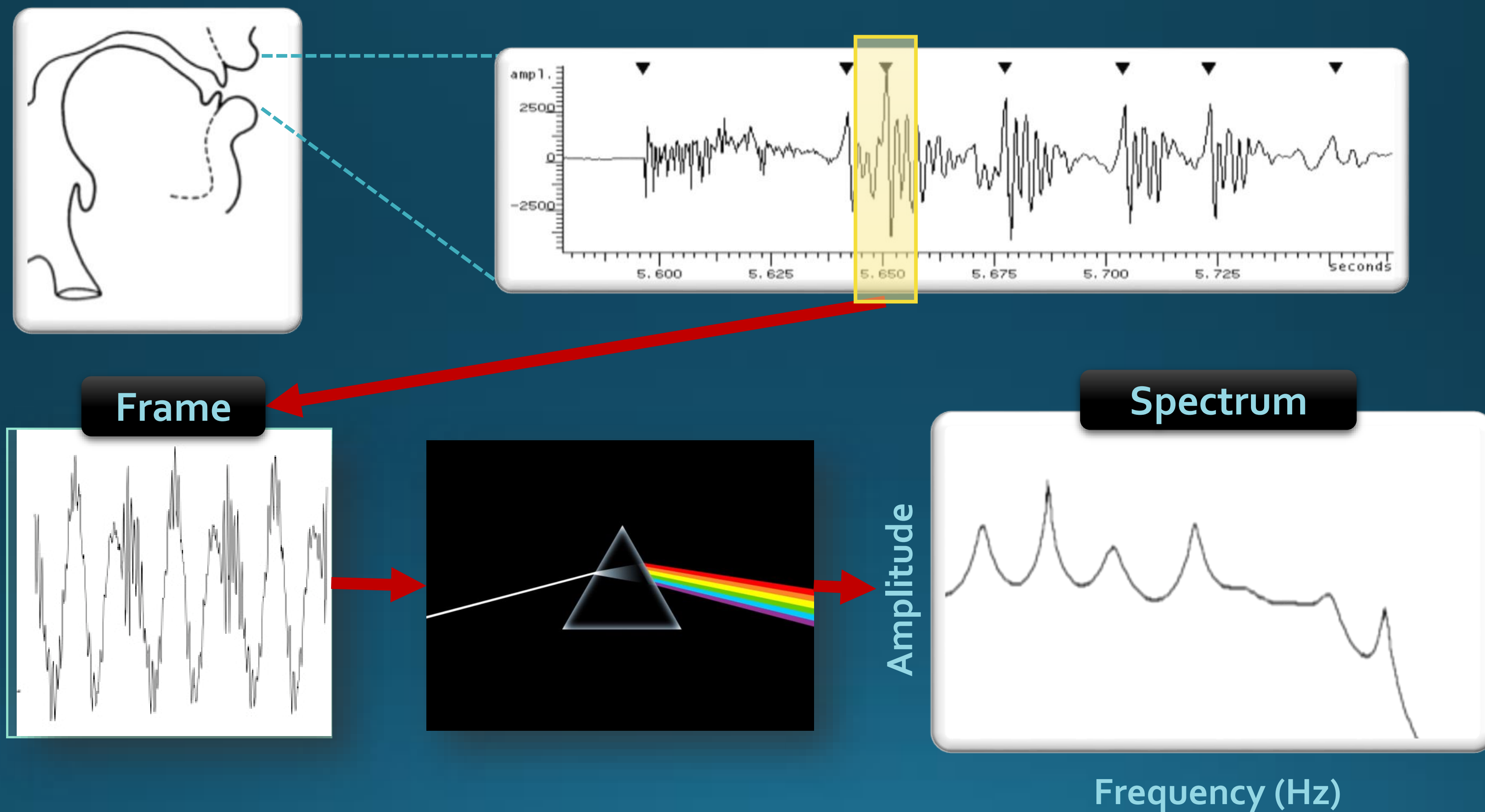
Spectrum

White light



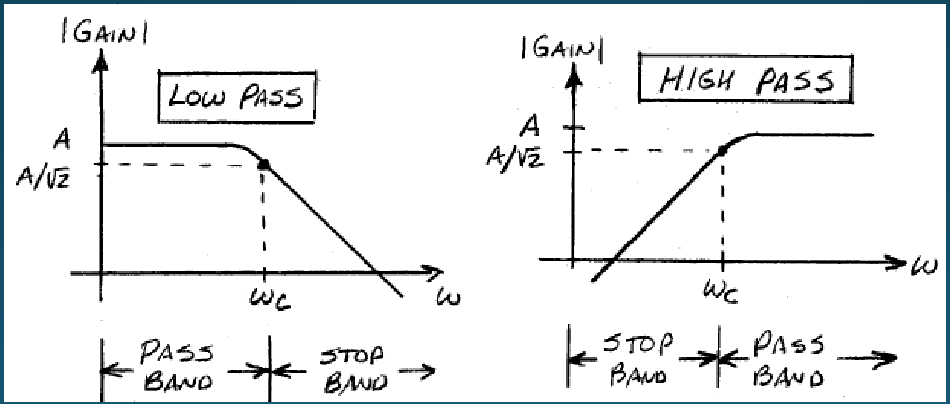
Any colour you like
(track 8)

Extracting a spectrum



Filtering

- Sometimes you only want part of a signal.
  - E.g., you have measurements of lip aperture over time – you know that they can't move > 5-10 Hz.
  - E.g., you know there's some low-frequency Gaussian noise in either the environment or transmission medium.



  - Low- and high-pass filters can be combined in series, yielding a **band-pass filter.**

UNIVERSITY OF TORONTO

Filtering

- **The Butterworth filter** is a transfer function designed to be maximally flat in the pass band.

n	Factors of Polynomial $B_n(s)$
1	$(s+1)$
2	$s^2 + 1.4142s + 1$
3	$(s+1)(s^2+s+1)$
4	$(s^2+0.7654s+1)(s^2+1.8478s+1)$

- The **transfer function** is

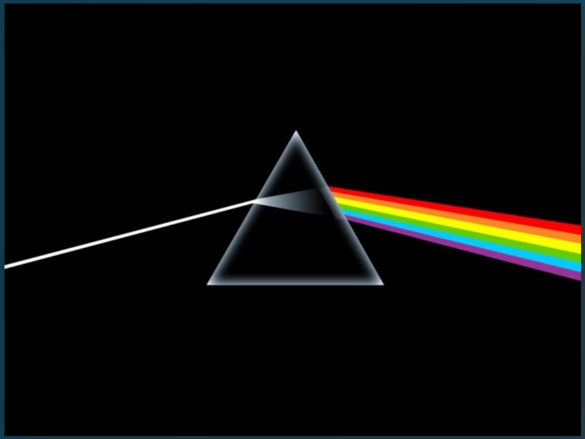
$$H(s) = G_0 / B_n(s/\omega_c)$$

where G_0 is the gain at zero frequency, and ω_c is the cutoff frequency.

- The **gain** of the n^{th}-order Butterworth filter is

$$G^2(\omega) = \frac{G_0^2}{1 + \left(\frac{\omega}{\omega_c}\right)^{2n}}$$

The continuous Fourier transform



- So we can **attenuate** frequencies above or below certain cut-offs.
- But, can we **measure** the actual **amount** of frequency F in a time signal $x(t)$?

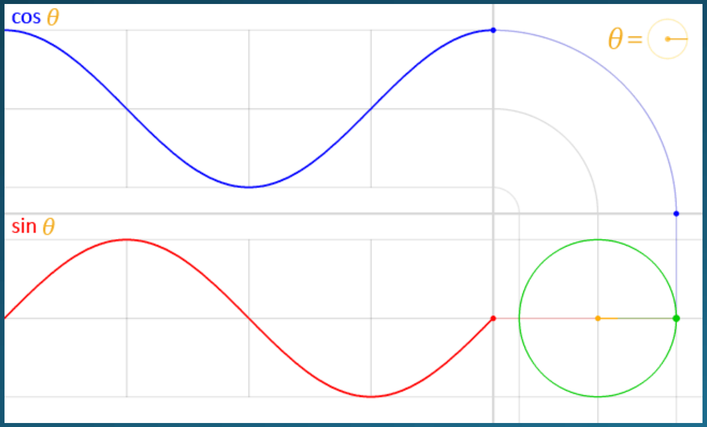
Euler's formula

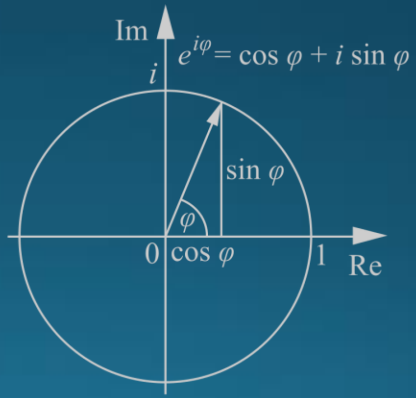
- Extracting spectra is made easier using **Euler's formula:**

$$e^{ix} = \cos(x) + i\sin(x) \qquad i^2 = -1$$

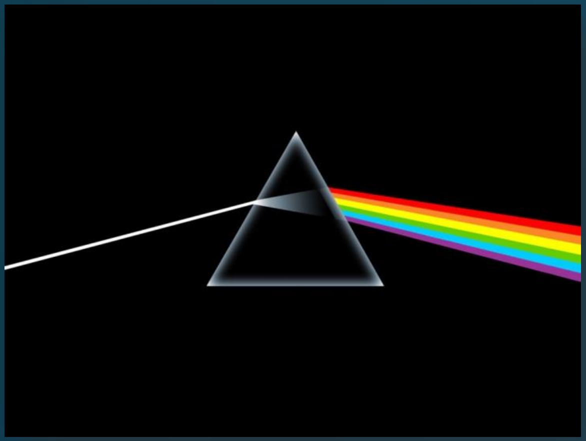
$$e^{i\pi} = -1$$

Euler's identity





The Fourier transform: intuition

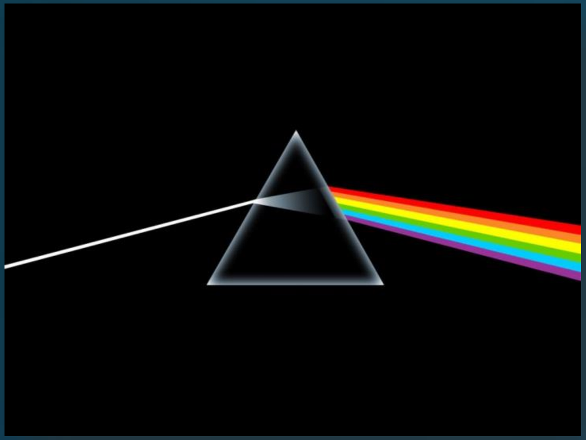


1. If we ignore phase, we only care about the real part, so
   $\cos(\omega \mathrm{t}) = e^{i\omega t}$ is **one** component.
2. How much '7' is in '**42**'?
   There is **42**/7= **6** 7s in **42**. Similarly,
   How much [18 Hz] is there in $\boldsymbol{x}(\boldsymbol{t})$?
   There is $\boldsymbol{x}(\boldsymbol{t})/[18Hz]$.
3. How much freq. ω is in $\boldsymbol{x}(\boldsymbol{t})$?
   $x(t)/\cos(\omega t) \quad = \quad x(t)/e^{i\omega t} \quad = \quad x(t)e^{-i\omega t}$
4. And over the entire signal?

$$\boldsymbol{X}(\boldsymbol{\omega}) = \int_{-\infty}^{\infty} \boldsymbol{x}(\boldsymbol{t}) e^{-i\omega t}\, dt$$

UNIVERSITY OF TORONTO

The continuous Fourier transform



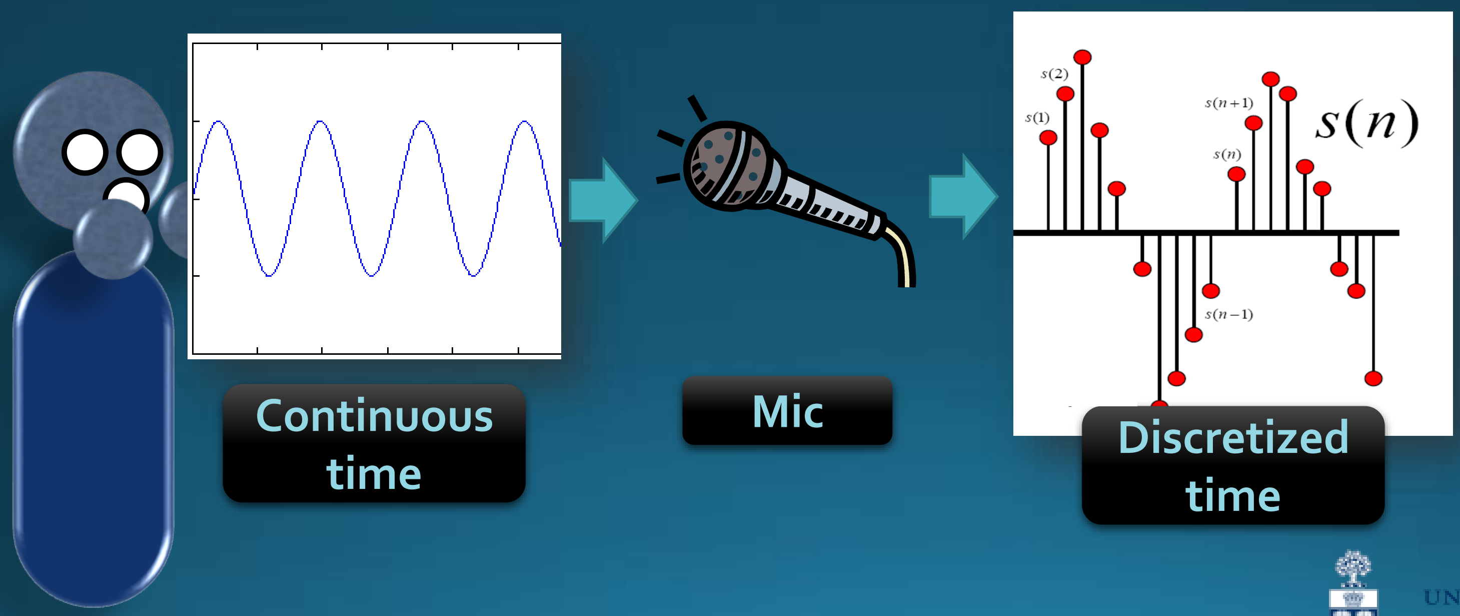
- **Input**: Continuous signal $x(t)$.
- **Output**: Spectrum $X(F)$ $(\omega = 2\pi F)$

$$X(F) = \int_{-\infty}^{\infty} x(t) e^{-i2\pi F t}\, dt$$

- It's **invertible**, i.e., $x(t) = \int_{-\infty}^{\infty} X(F) e^{i2\pi F t}\, dF$.
- It's **linear**, i.e., for $a, b \in \mathbb{C}$, **if** $h(t) = ax(t) + by(t)$, **then** $H(F) = aX(F) + bY(F)$
- ...
- It needs a **continuous** input $x(t)$....***uh oh?***

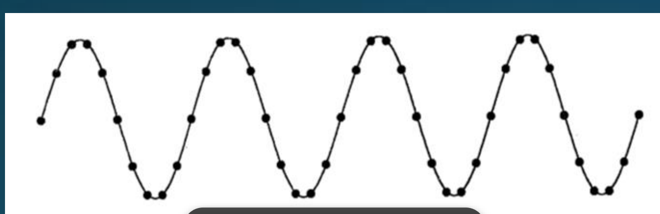
Discrete signals

- **Sampling**: *vbg.* measuring the amplitude of a signal at regular intervals.
  - e.g., 44.1 kHz (*CD*), 8 kHz (*telephone*).
  - These amplitudes are initially measured as **continuous** values at **discrete** time steps.

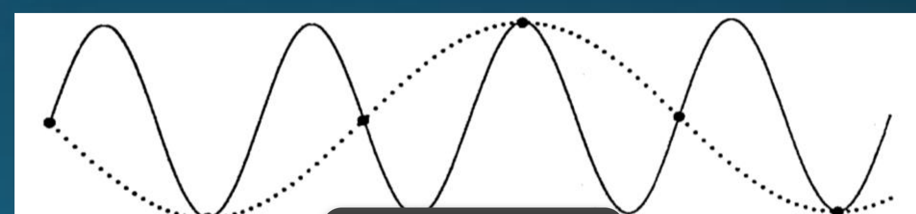


Discrete signals

- **Nyquist rate**: *n.* the **minimum** sampling rate necessary to preserve the **maximum** frequency.
  - i.e., **twice** the maximum frequency, since we need >2 samples/cycle.
  - Human speech is quite informative ≤ 4 kHz, $\therefore$ 8 kHz sampling.



Good sampling

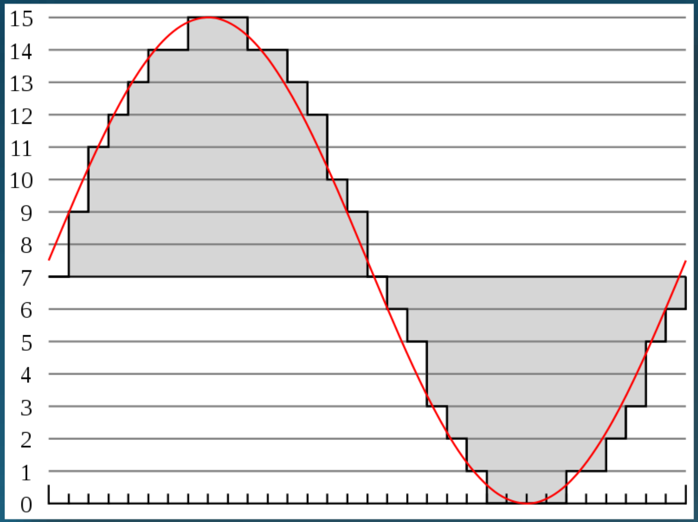


Under-sampling

Discrete signals

- **Quantization**: *n.* the conversion of **floating point** amplitude sample values to **integers**.
- **PCM**: *n.* (pulse code modulation) a method of quantization in which the analog amplitude is quantized at **uniform intervals** .

  (e.g., 8 bit (−128..127), 16 bit (−32768..32767).



UNIVERSITY OF TORONTO

Discrete Fourier transform (DFT)

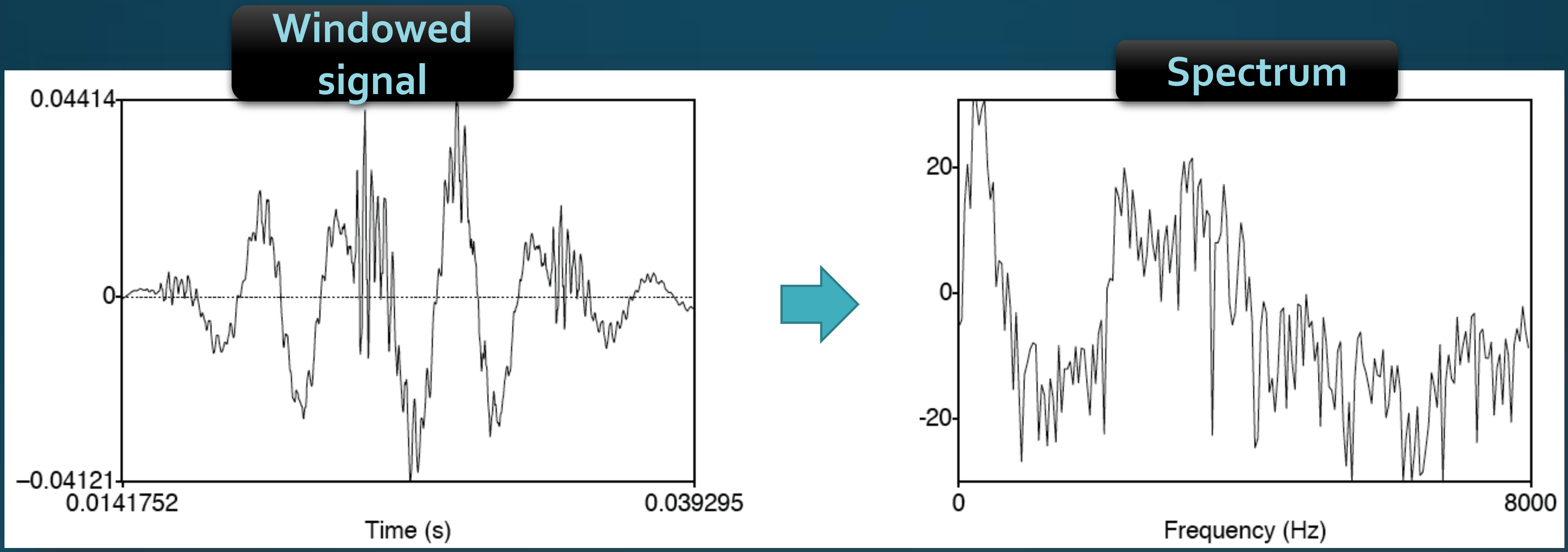
- **Input**: Windowed signal $x[0] \ldots x[N-1]$.
- **Output**: N complex numbers $X[k]$ $(k \in \mathbb{Z})$

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-i 2\pi k \frac{n}{N}}$$

- **Algorithm(s)**: the **Fast Fourier Transform** (FFT) with complexity $O(N \log N)$.
  - The **Cooley-Tukey algorithm** *divides-and-conquers* by breaking the DFT into smaller ones $N = N_1 N_2$.

Discrete Fourier transform (DFT)

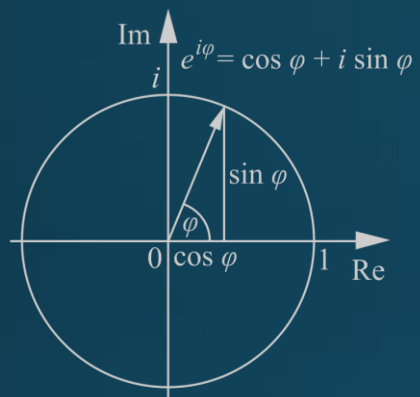
- Below is a 25 ms Hamming-windowed signal from */iy/*, and its spectrum as computed by the DFT.

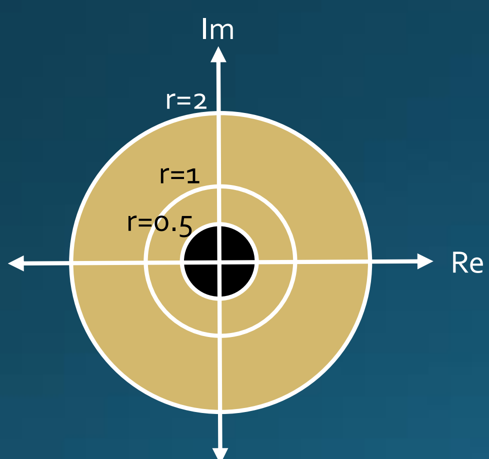


Recall: the Fourier transform is invertible

This really only covers a particular set of sinusoidal functions…

The z-transform





- What if we don't *need* the unit circle, $r = 1$?
- $X(z) = \sum_{n=-\infty}^{\infty} x[n]z^{-n}$,
    - where $z \in \mathbb{C}$ so $z = re^{i\omega}$
- Requires a **region of convergence** in the complex plane where the summation converges.
    - $RoC = \{z : |\sum_{n=-\infty}^{\infty} x[n]z^{-n}| < \infty\}$
- If **yellow region** on left is RoC, then discrete-time Fourier transform exists, since $r = 1$ is in the RoC.

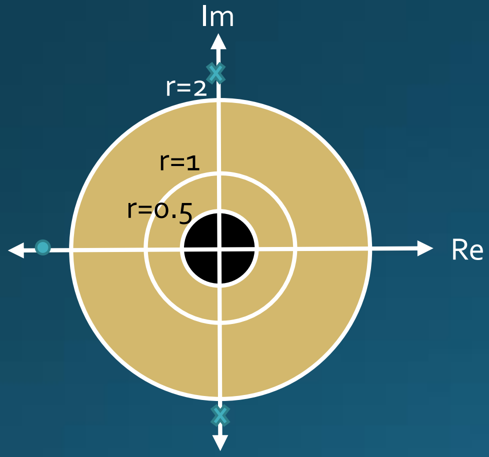
UNIVERSITY OF TORONTO

Poles and zeros

- **Transfer functions** of linear time-invariant (LTI) systems have this form:

$$H(s) = \frac{P(s)}{Q(s)} = \frac{G \cdot \sum_{m=0}^{M} b_m s^m}{s^N + \sum_{n=0}^{N-1} a_n s^m}$$

where G is the **gain**, M and N are **orders** of polynomials, and b_m & a_n are **coefficients** of those polynomials.

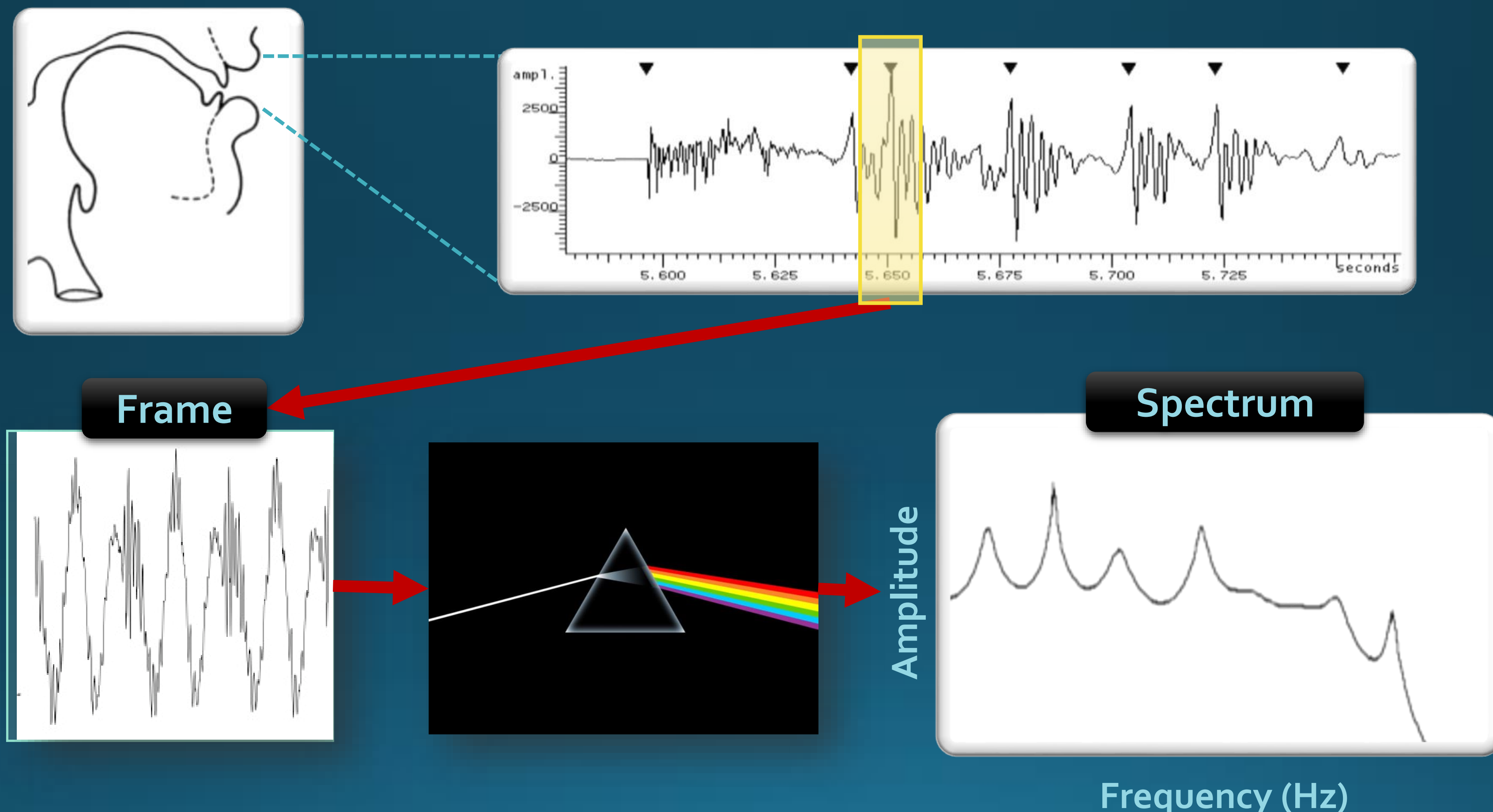


- Zeros occur when $P(s)|_{s=\beta_m} = 0$.
- Poles occur when $Q(s)|_{s=\alpha_n} = 0$.

- The RoC cannot contain any poles.

Q: Why do Polish airlines only fill half of their seats?
A: Because Poles on the right half of the plane are unstable.
(http://en.wikipedia.org/wiki/Nyquist_stability_criterion)

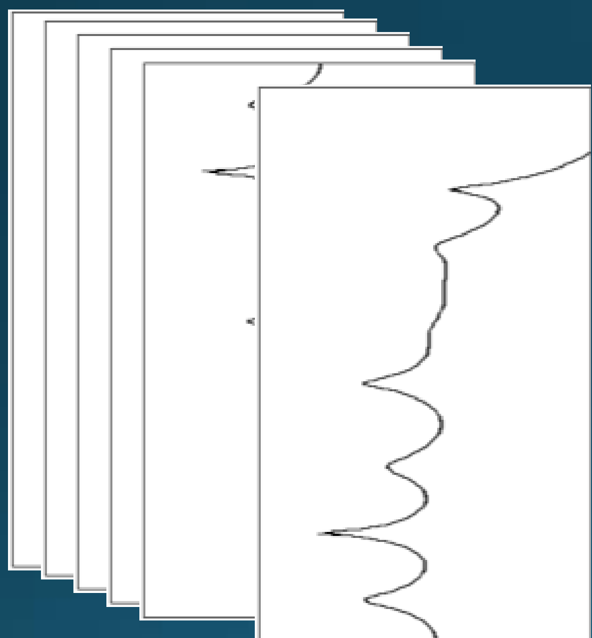
Extracting a spectrum



But in speech we need many successive windows…

Spectrograms

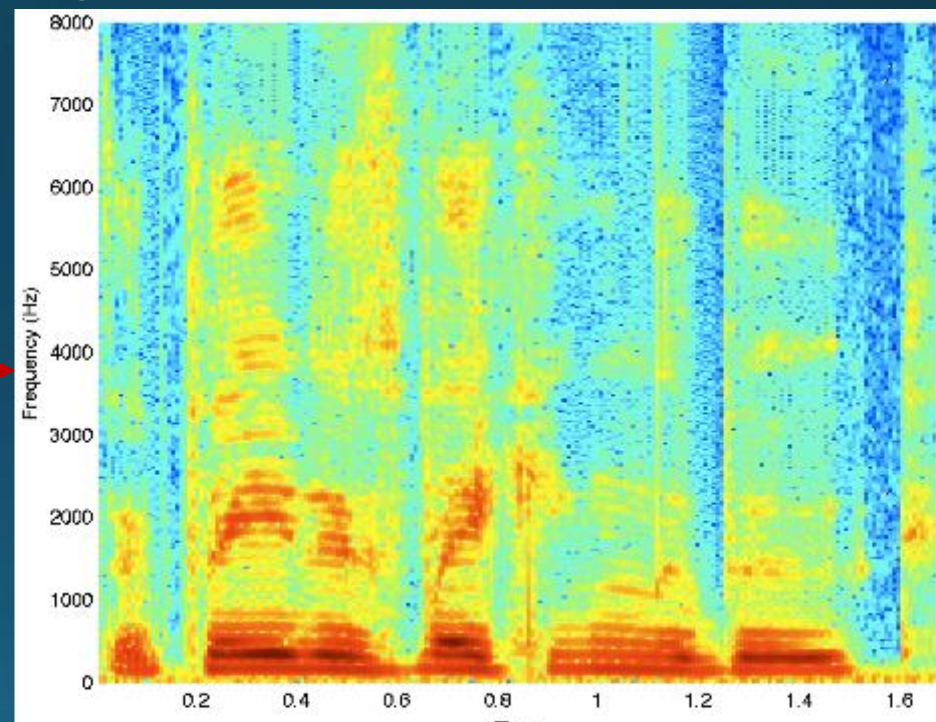
- **Spectrogram**: *n.* a 3D plot of **amplitude** and **frequency** over **time** (higher 'redness' → higher amplitude).



Frequency (Hz)

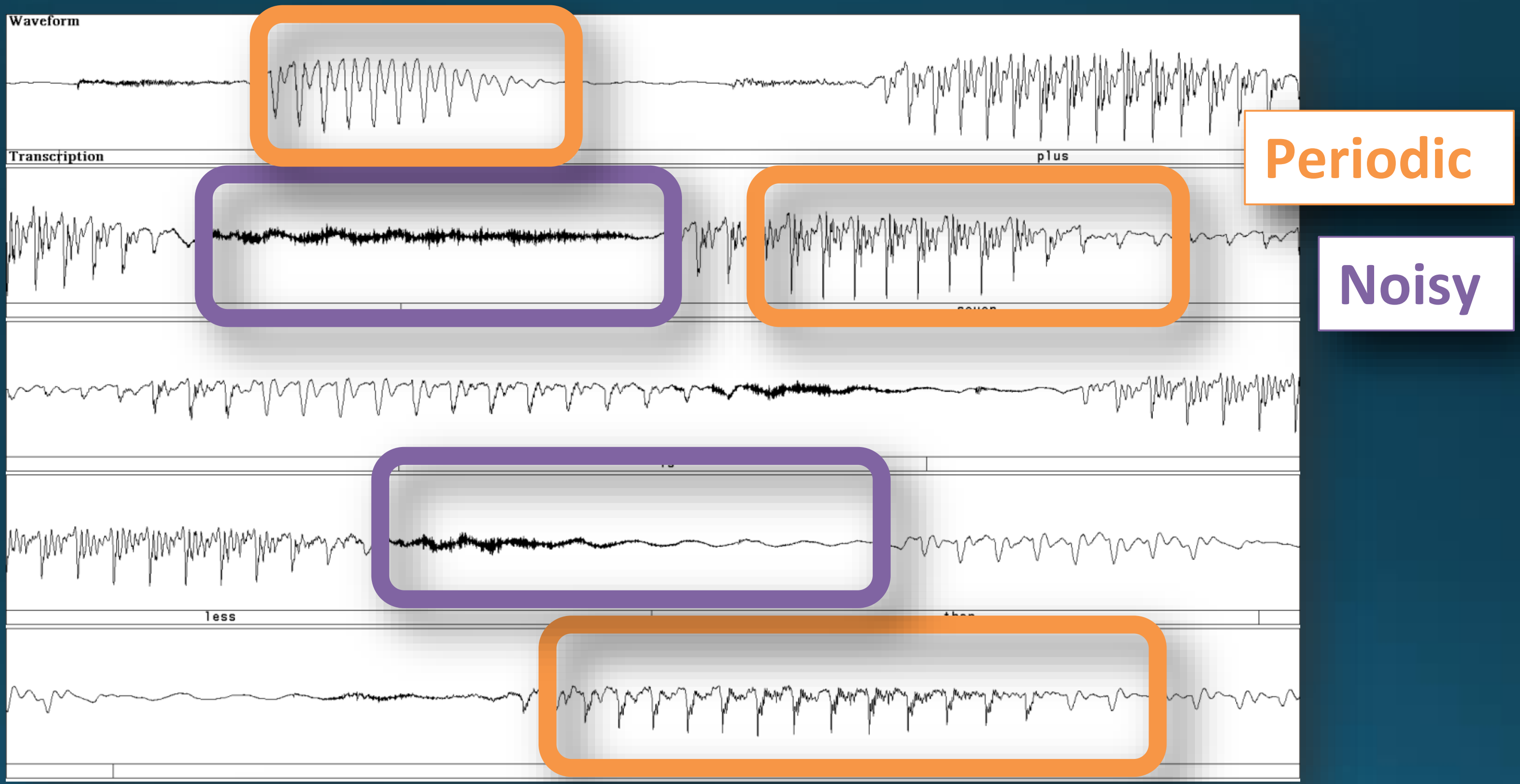
Amplitude

Frames



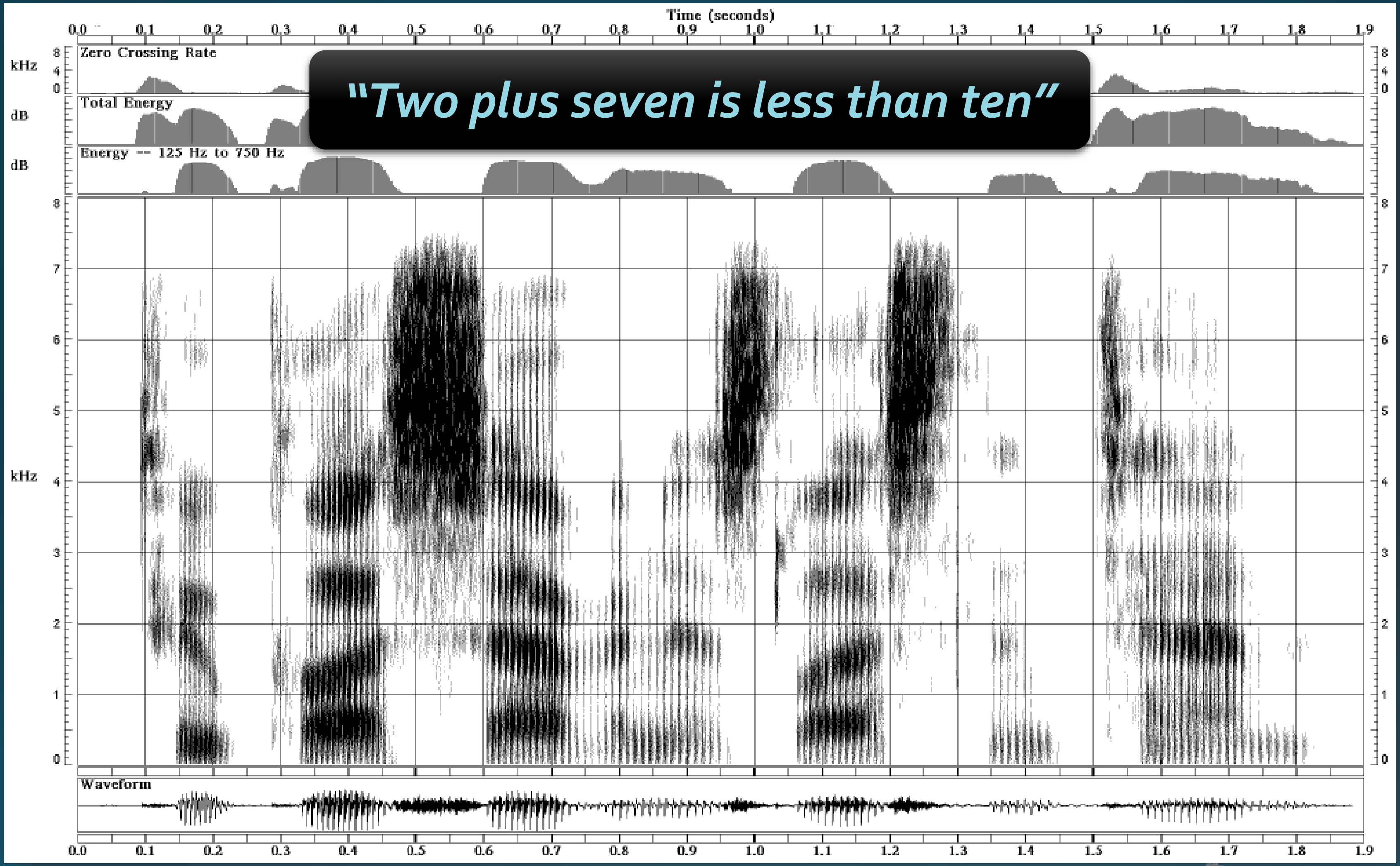
Spectrogram

Speech signals



"Two plus seven is less than ten"

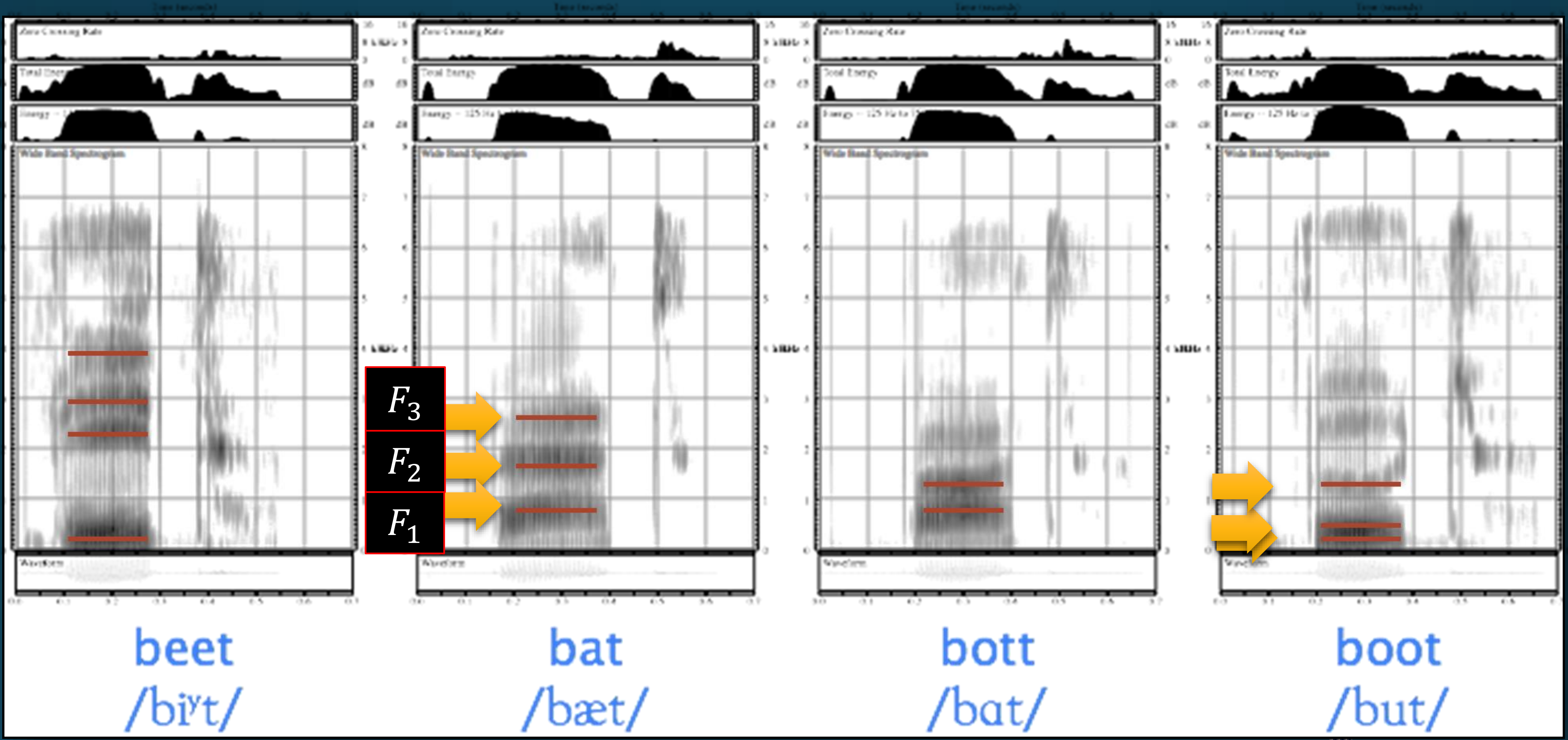
Spectrograms



UNIVERSITY OF TORONTO

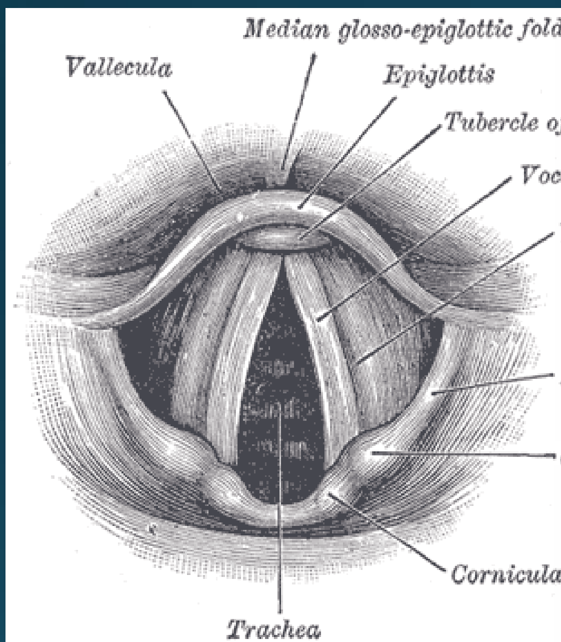
Formants and phonemes

- **Formant**: *n.* A concentration of energy within a frequency band. Ordered from low to high bands.

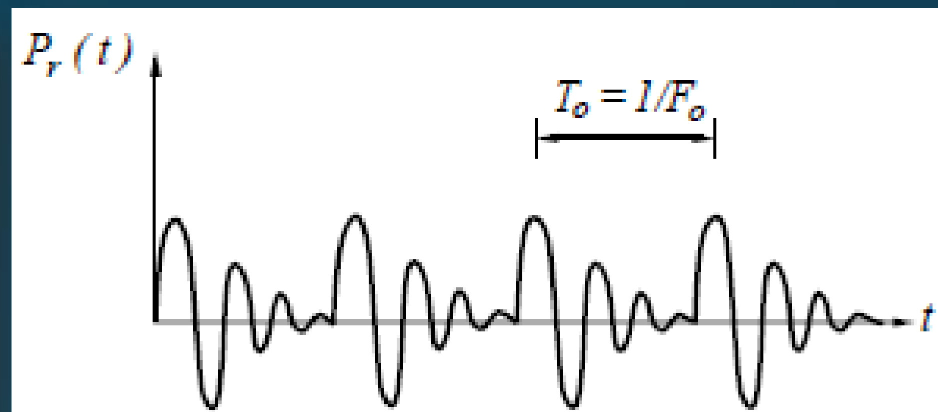


Fundamental frequency

- F_0: *n.* (**fundamental frequency**), the rate of vibration of the **glottis** – often very **indicative** of the speaker.



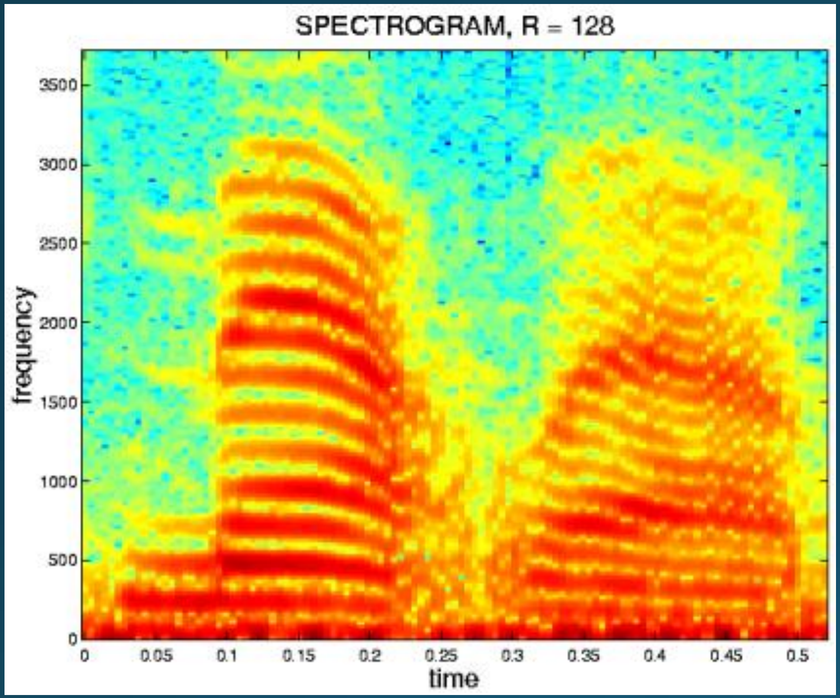
Glottis

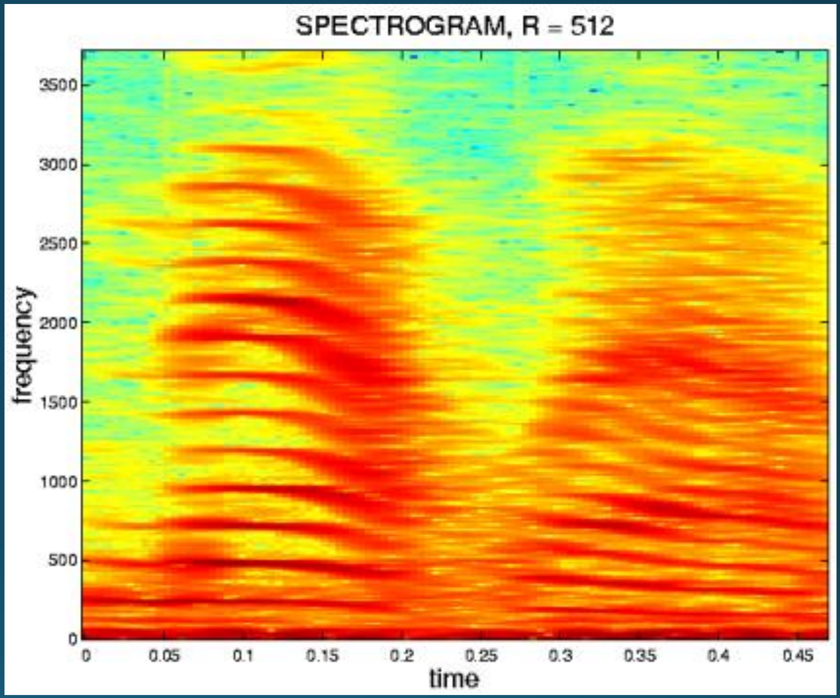


	Avg F_0 (Hz)	Min F_0 (Hz)	Max F_0 (Hz)
Men	125	80	200
Women	225	150	350
Children	300	200	500

Formants (should) occur at multiples of F_0

Effect of window length





Wide-band (better time resolution)

Narrow-band (better frequency resolution)

UNIVERSITY OF TORONTO

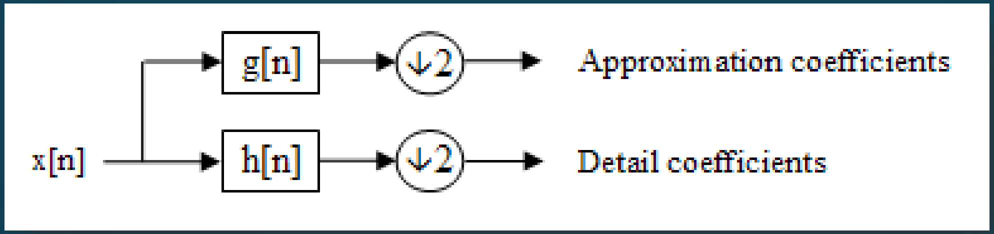
Wavelet transforms

- **Avoid** problem of resolution, and can **adapt** to changes in the signal over time (i.e., non-stationary signals).

- Wavelet transforms consist of **scaled** and **translated** versions ('daughter wavelets') of basis functions.

Morlet

Wavelet transforms



where, given low- and high-pass filters (g and h, respectively),

- **Approx**: $y_{low} = (x * g) \downarrow 2$, $y_{low}[n] = \sum_{k=-\infty}^{\infty} x[k]g[2n-k]$
- **Detail**: $y_{high} = (x * h) \downarrow 2$, $y_{high}[n] = \sum_{k=-\infty}^{\infty} x[k]h[2n-k]$

Convolution

Downsampled

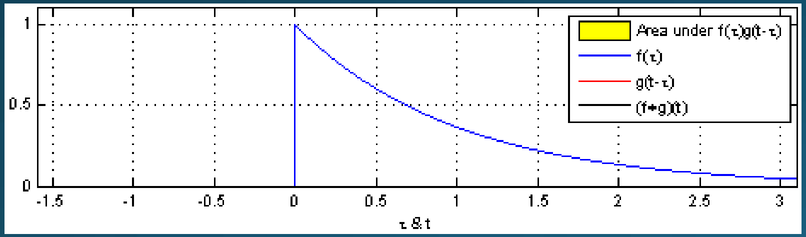
UNIVERSITY OF TORONTO

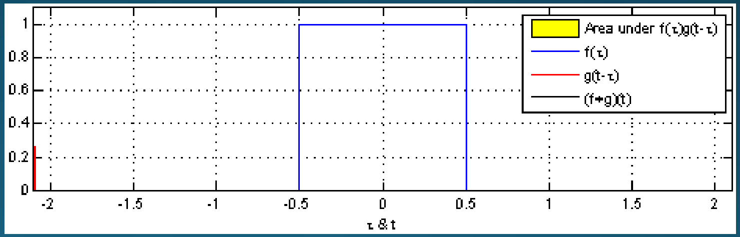
Convolution?

- The **convolution** of two functions, $f * g$, is the amount of **overlap** between two functions as one is **translated**.

- Discrete version:

$$(f * g)[n] = \sum_{m=-\infty}^{\infty} f[m]g[n-m]$$



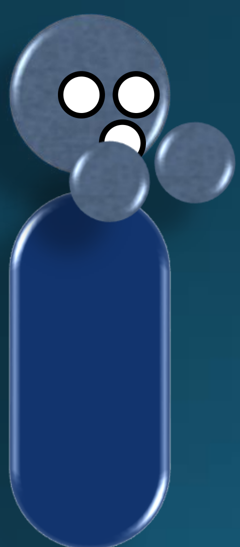


- It is related to cross-correlation, which is a measure of similarity.

UNIVERSITY OF TORONTO


Speech recognition

Speech as a sequence of phonemes



We want to convert
a series of acoustic observation vectors into
a sequence of phonemes or words.

The noisy channel model in ASR

Language model

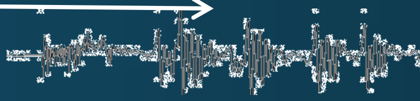
Source
$\boldsymbol{P(W)}$

W' →

Acoustic model

Channel
$\boldsymbol{P(X|W)}$

X' →



Decoder

$\boldsymbol{W^*}$ ←

← Observed $\boldsymbol{X}$

Word sequence W

Acoustic sequence X

$$\boxed{W^* = \underset{W}{\operatorname{argmax}}\, P(X|W)P(W)}$$

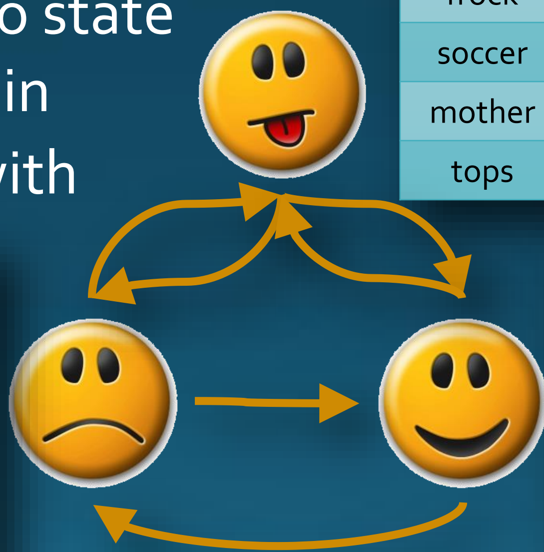
How to encode $P(X|W)$?

UNIVERSITY OF TORONTO

Reminder – discrete HMMs

- In **discrete Hidden Markov Models**, at each state we observe a discrete symbol.
- We **transition** from state s_i to state s_j with probability a_{ij}. While in state s we **observe** word w with probability $b_s(w)$.

word	P(word)
ship	0.1
pass	0.05
camp	0.05
frock	0.6
soccer	0.05
mother	0.1
tops	0.05



word	P(word)
ship	0.25
pass	0.25
camp	0.05
frock	0.3
soccer	0.05
mother	0.09
tops	0.01

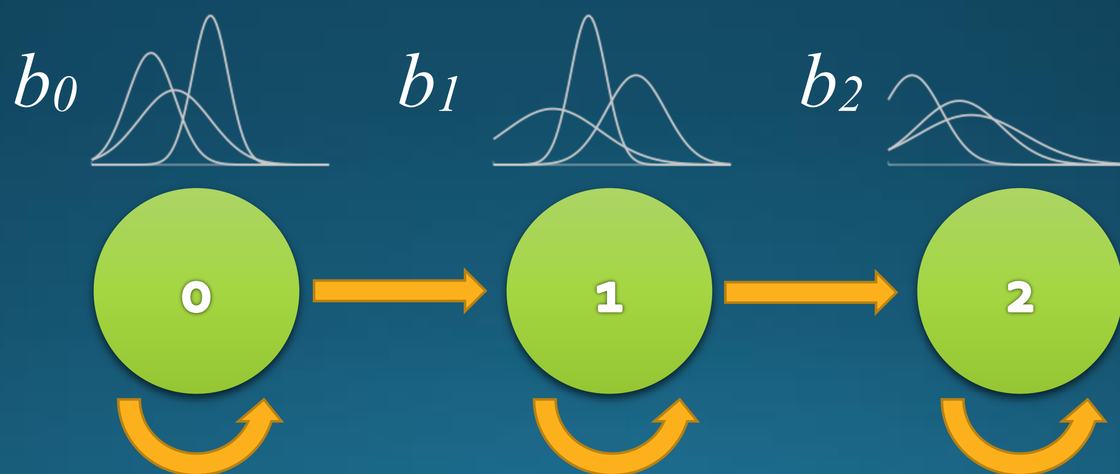
word	P(word)
ship	0.3
pass	0
camp	0
frock	0.2
soccer	0.05
mother	0.05
tops	0.4

But acoustics aren't discrete…

Continuous HMMs

- A **continuous HMM** has **continuous** output observations.
  - Observation probabilities, b_i, are also continuous.
  - E.g., here $b_0(\vec{x})$ tells us the probability of seeing the (multivariate) continuous observation $\vec{x}$ while in state 0.

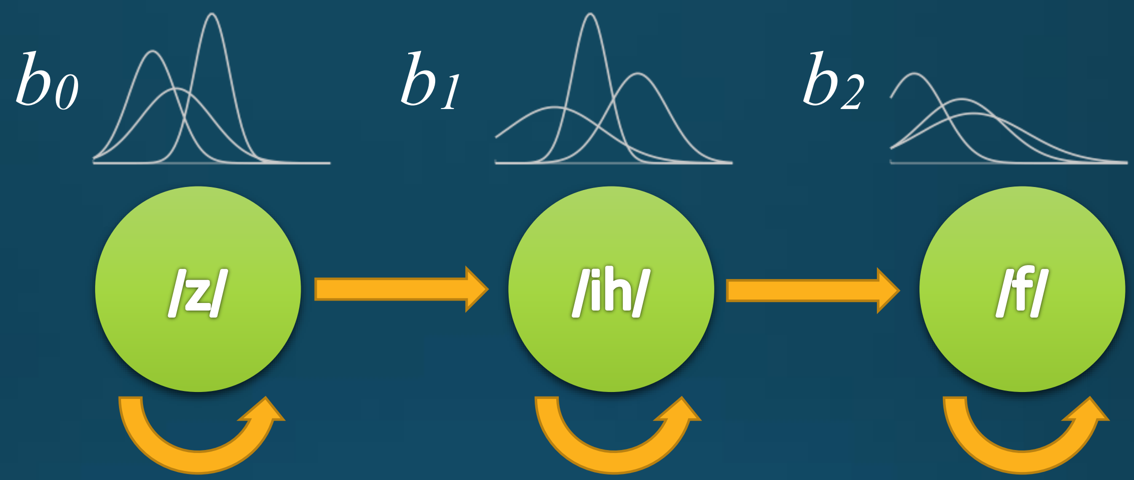
$$\vec{x} = \begin{pmatrix} 4.32957 \\ 2.48562 \\ 1.08139 \\ \dots \\ 0.45628 \end{pmatrix}$$



What do the states represent?

One HMM per word?

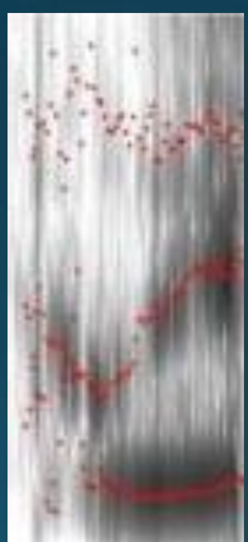
- In a word-level HMM, each state might be a phoneme.

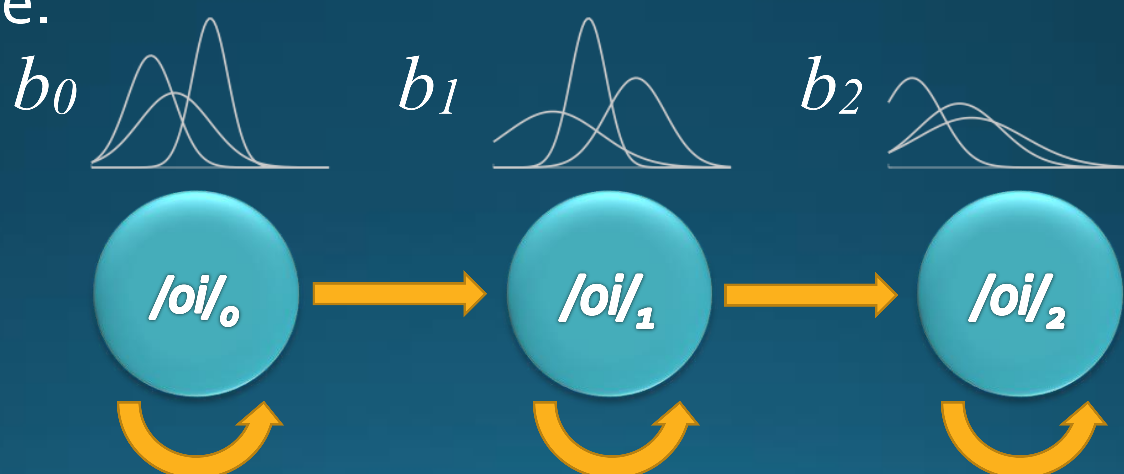


- Imagine that we want to learn an HMM for each word in our lexicon (e.g., **160K words** $\rightarrow$ **160K HMMs**).
- **No, thank you**! According to **Zipf's law**, we expect *many* words to occur *very* infrequently.
  - 1 (or a few) training examples of a word is ***not*** enough to train a model as highly parameterized as a CHMM.

One HMM per phoneme?

- Phonemes change over time – we model these dynamics by building one HMM for each phoneme.
  - **Tristate** phoneme models are popular.
    - The centre state is often the 'steady' part of the phoneme.





tristate phoneme model (e.g., */oi/*)

How do we learn these probabilities?

UNIVERSITY OF TORONTO

Training phoneme HMMs

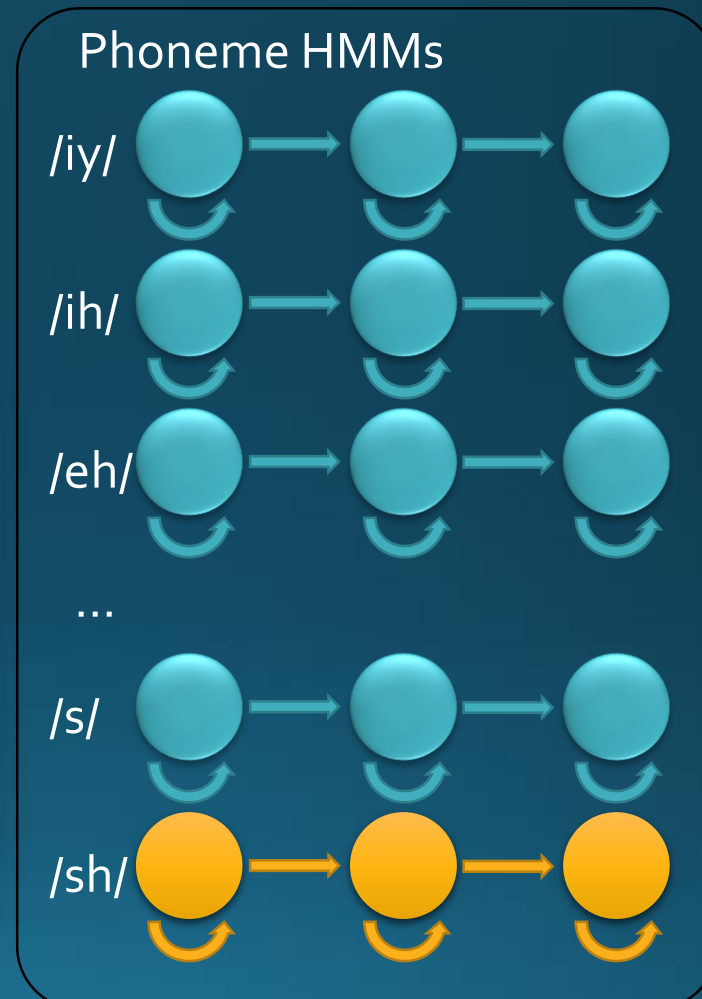
- Training data for a phoneme HMM come from ***all*** sequences of that phoneme.
  - *Even from different words.*

```
...
64 85   ae
85 96   sh
96 102  epi
102 106 m
...
```

annotation

		Time, t				
		…	85	…	96	…
Feature	1	…		…		…
	2	…		…		…
	3	…		…		…
	…	…	…	…	…	…
	42	…		…		…

observations



Combining HMMs

- We can learn an *N*-gram **language model** from word-level and phoneme-level annotations of speech data.
  - These models are discrete and are trained using MLE.
- Our phoneme HMMs together constitute an **acoustic model**.
  - Each phoneme HMM tells us how a phoneme 'sounds'.
- We can **combine** these models by **concatenating** together phoneme HMMs according to a known lexicon or phonemic dictionary.

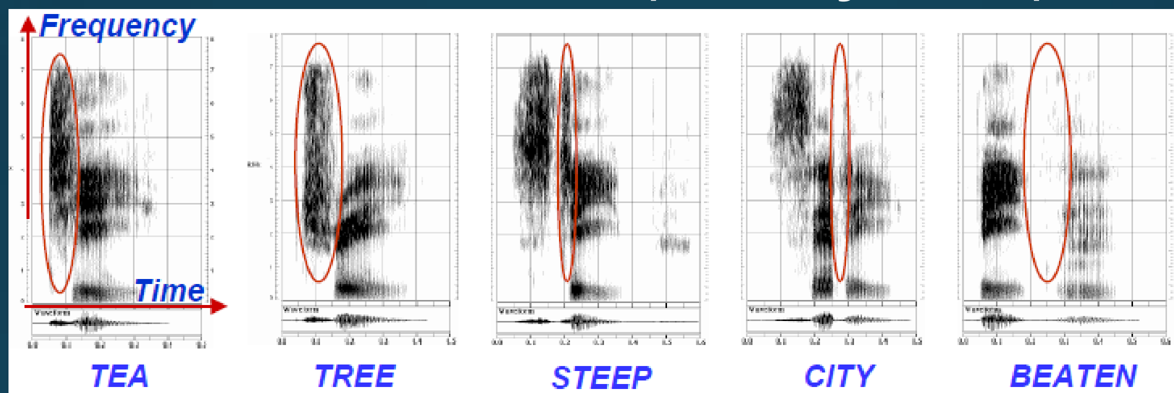
```
…
EVOLUTION                 EH2 V AH0 L UW1 SH AH0 N
EVOLUTION(2)              IY2 V AH0 L UW1 SH AH0 N
…
EVOLUTIONARY     EH2 V AH0 L UW1 SH AH0 N EH2 R IY0
```

Combining HMMs

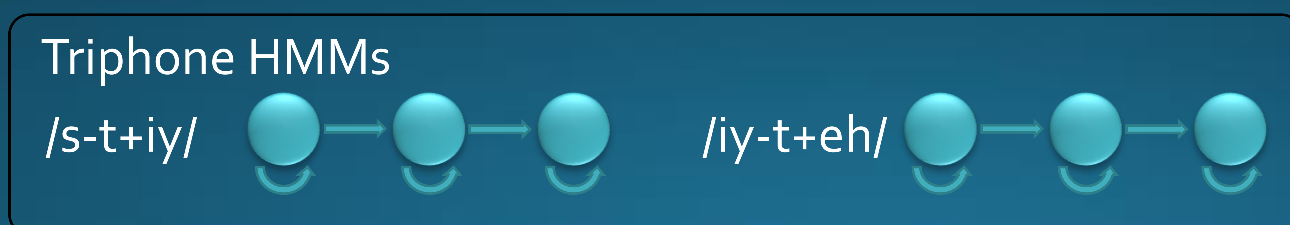
- If we know how phonemes combine to make words, we can simply **concatenate** together our phoneme models by inserting and **adjusting** transition weights.
  - e.g., *Zipf* is pronounced */z ih f/*, so…

Coarticulation and triphones

- **Co-articulation**: *n.* the situation when a phoneme is influenced by an adjacent phoneme.

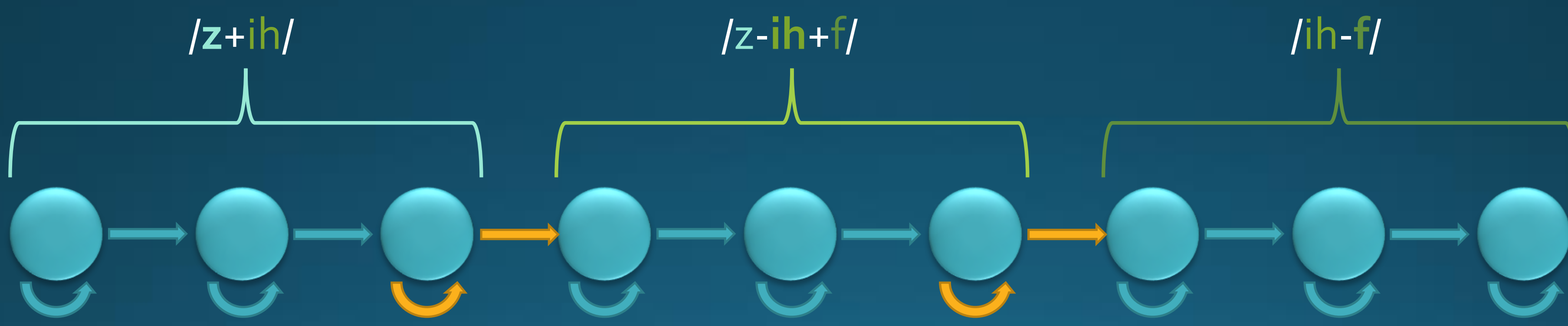


- A **triphone HMM** captures co-articulation but represents one phoneme.

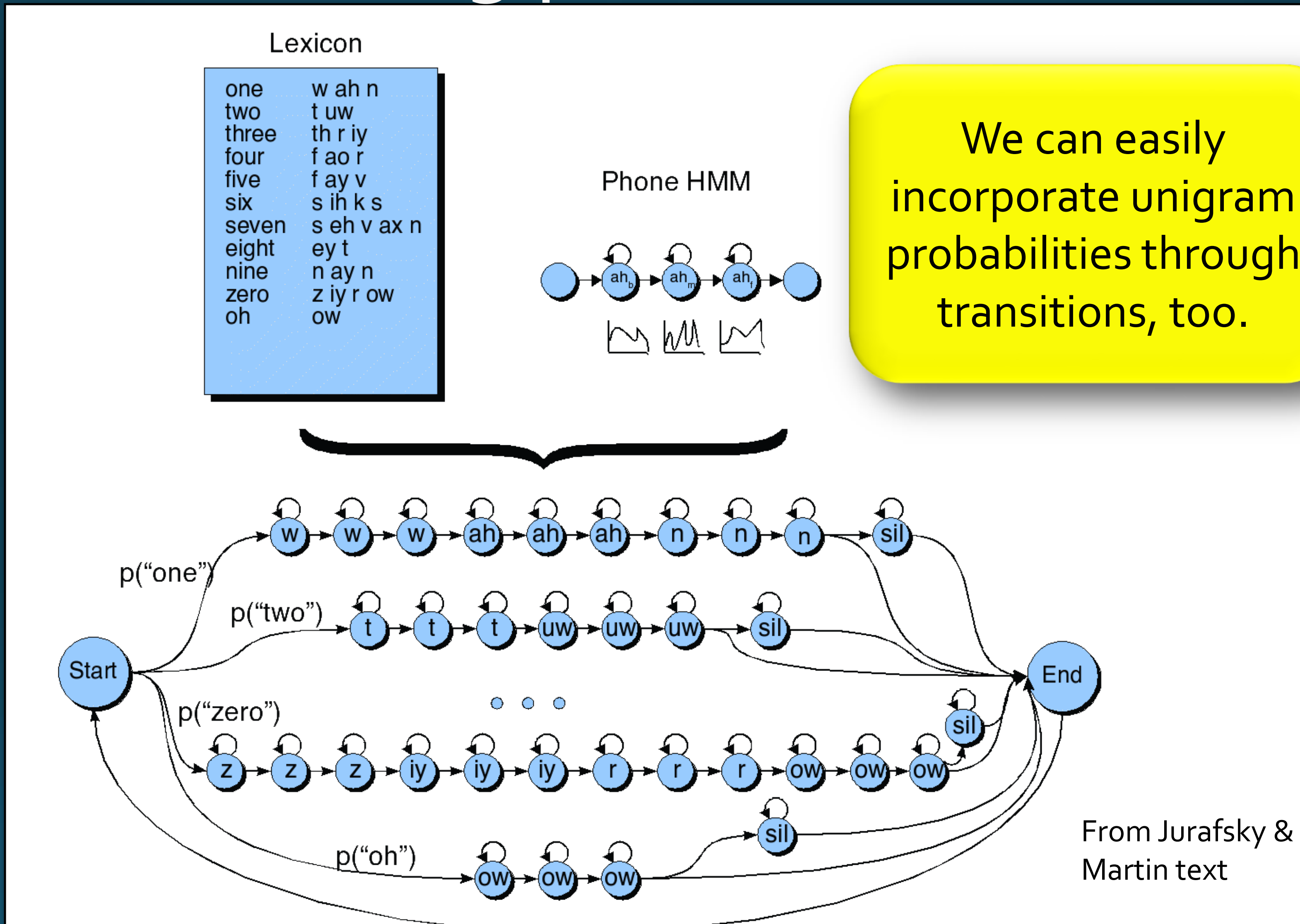


Combining triphone HMMs

- Triphone models can only connect to other triphone models that match the context.
  - Triphone model /a-b+c/ is the phoneme **b** that is preceded by **a** and followed by **c**.



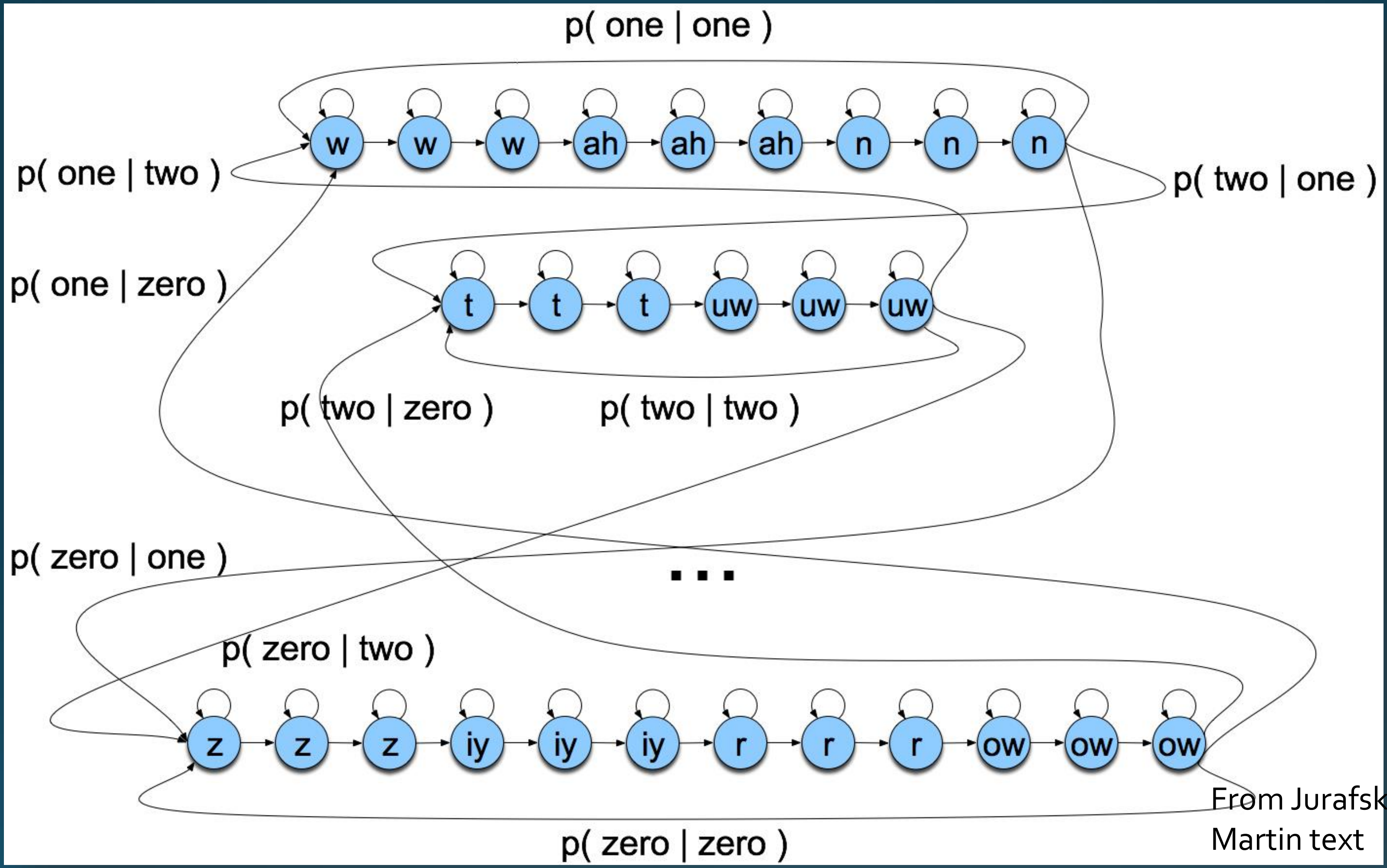
Concatenating phoneme models



We can easily incorporate unigram probabilities through transitions, too.

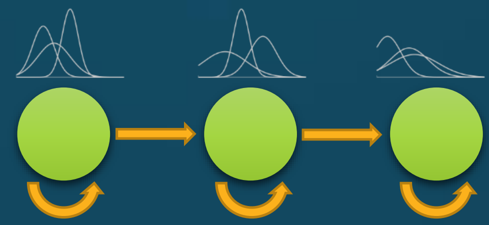
From Jurafsky & Martin text

Bigram models



From Jurafsky & Martin text

Using HMMs



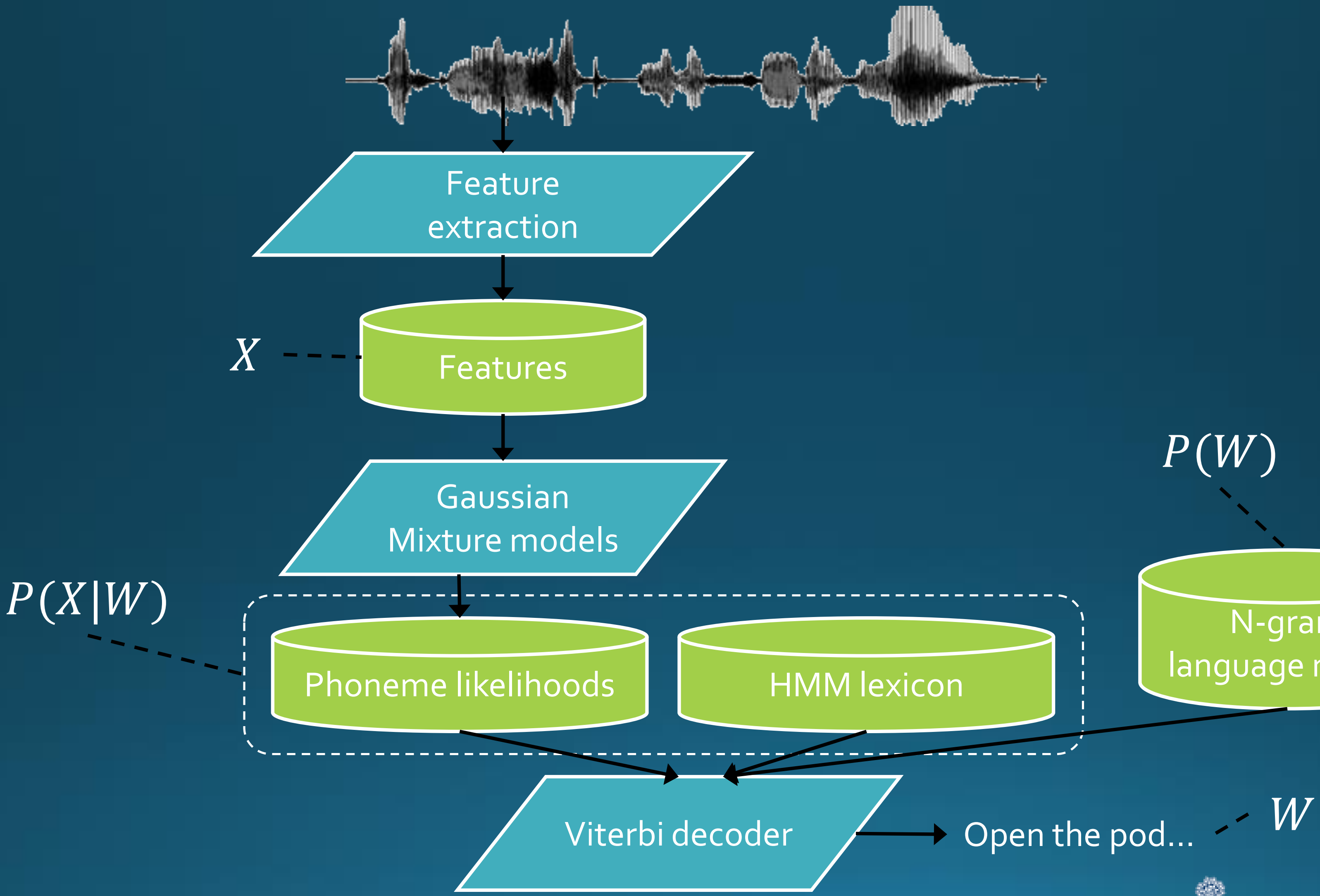
- HMMs are **generative** models that encode statistical knowledge of how output is **generated**.

- We **train** CHMMs with **Baum-Welch** (a type of Expectation-Maximization), as with discrete HMMs.
  - Here, the observation parameters, $b_i(\vec{x})$, are adjusted using another form of EM for GMMs.

- We find best state sequences using the **Viterbi** algorithm.
  - Here, the best state sequence returned gives us a **sequence of phonemes** and **words**.

UNIVERSITY OF TORONTO

ASR architecture



Aspects of ASR in the world

- **Speaking mode:** **Isolated** word (e.g., "*yes*") vs. **continuous** (e.g., "*Siri, sell my Apple stocks.*")
- **Speaking style**: **Read** speech vs. **spontaneous** speech; the latter contains many **dysfluencies** (e.g., stuttering, *uh*, *like*, …)
- **Enrolment**: **Speaker-dependent** (all training data from one speaker) vs. **speaker-independent** (training data from many speakers).
- **Vocabulary**: **Small** (<20 words) or **large** (>50,000 words).
- **Transducer**: Cell phone? Noise-cancelling microphone? Teleconference microphone?

Signal-to-noise ratio

- We are often concerned with the **signal-to-noise ratio** (SNR), which measures the **ratio** between the power of a **desired signal** within a recording (P_{signal}, e.g., the human speech) and **additive noise** (P_{noise}).
  - Noise typically includes:
    - **Background noise** (e.g., people talking, wind),
    - **Signal degradation**. This is *normally* 'white' noise produced by the medium of transmission.

$$SNR_{db} = 10 \, \log_{10}\left(\frac{P_{signal}}{P_{noise}}\right)$$

- High SNR_{db} is >30 dB. Low SNR_{db} is < 10 dB.

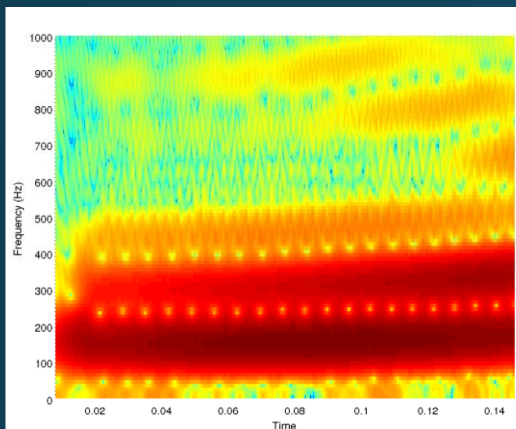
UNIVERSITY OF TORONTO

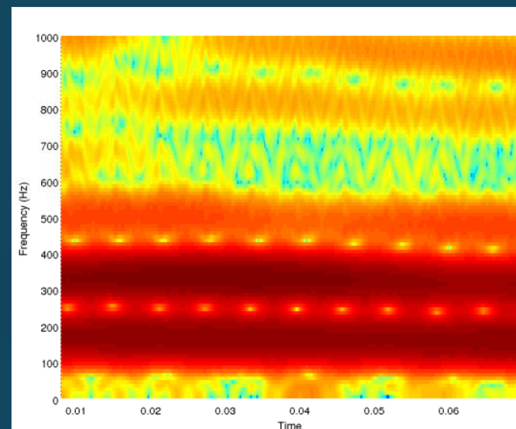
Audio-visual speech methods

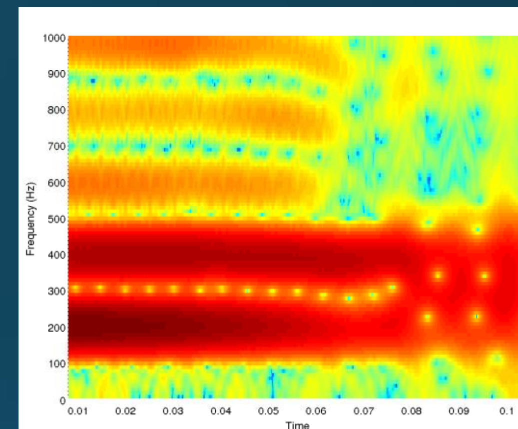
- Observing the **vocal tract** directly, rather than through inference, can be very helpful in ASR.
- The shape and aperture of the mouth gives some clues as to the phoneme being uttered.
  - Depending on the level of invasiveness, we can even measure the glottis and tongue directly.

Lip aperture and nasals

Acoustic spectrograms

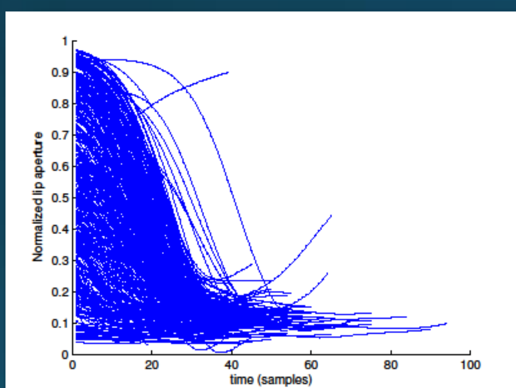


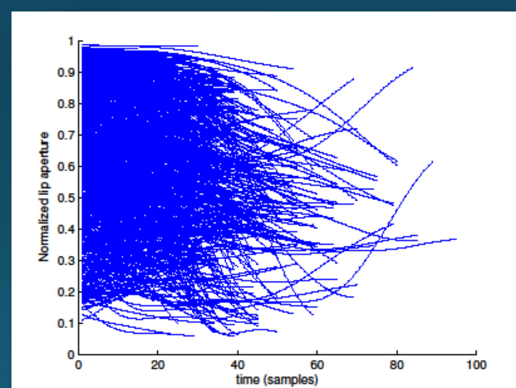


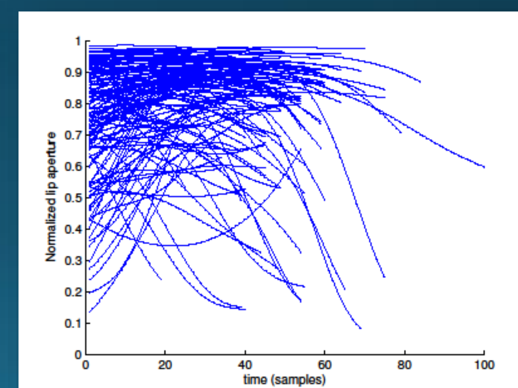


/m/ /n/ /ng/

Lip apertures over time

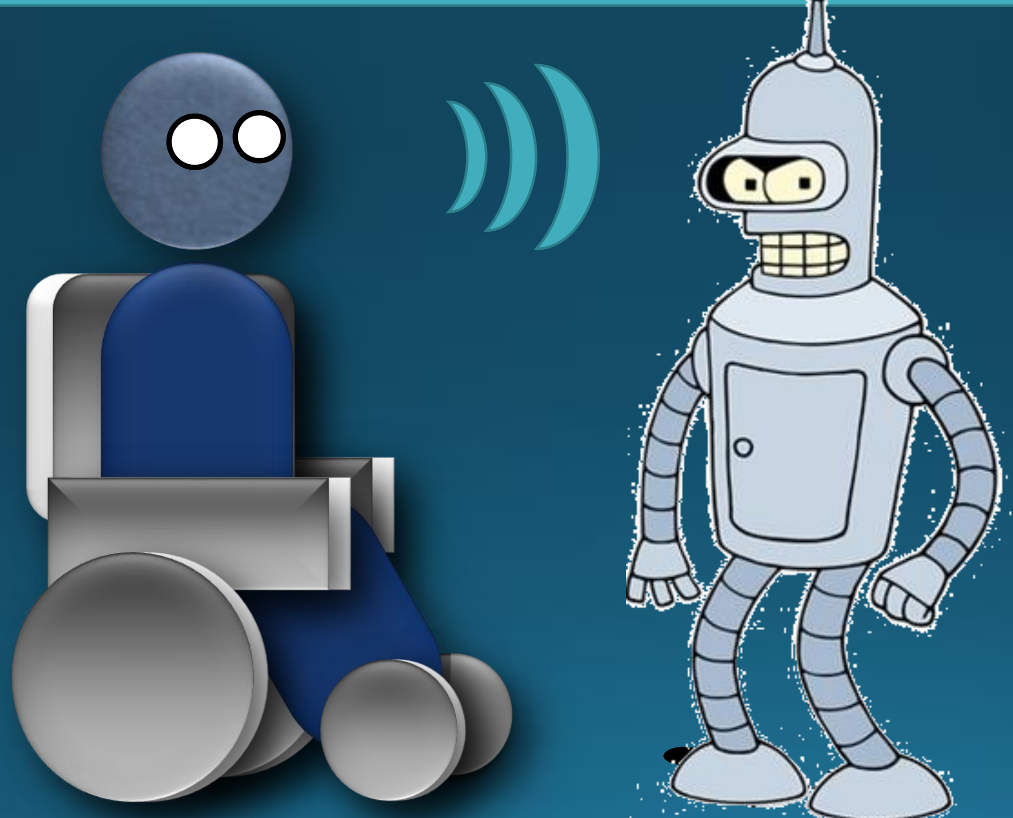






Dysarthria

Can we build models of atypical articulation? What are relevant features? How will technology be used? What about cognitive disorders?



Next week: clinical/medical aspects.

UNIVERSITY OF TORONTO