CSC401 Assignment 3

Tutorial 1 of 4

2021-03-10

Based on the slides of previous years

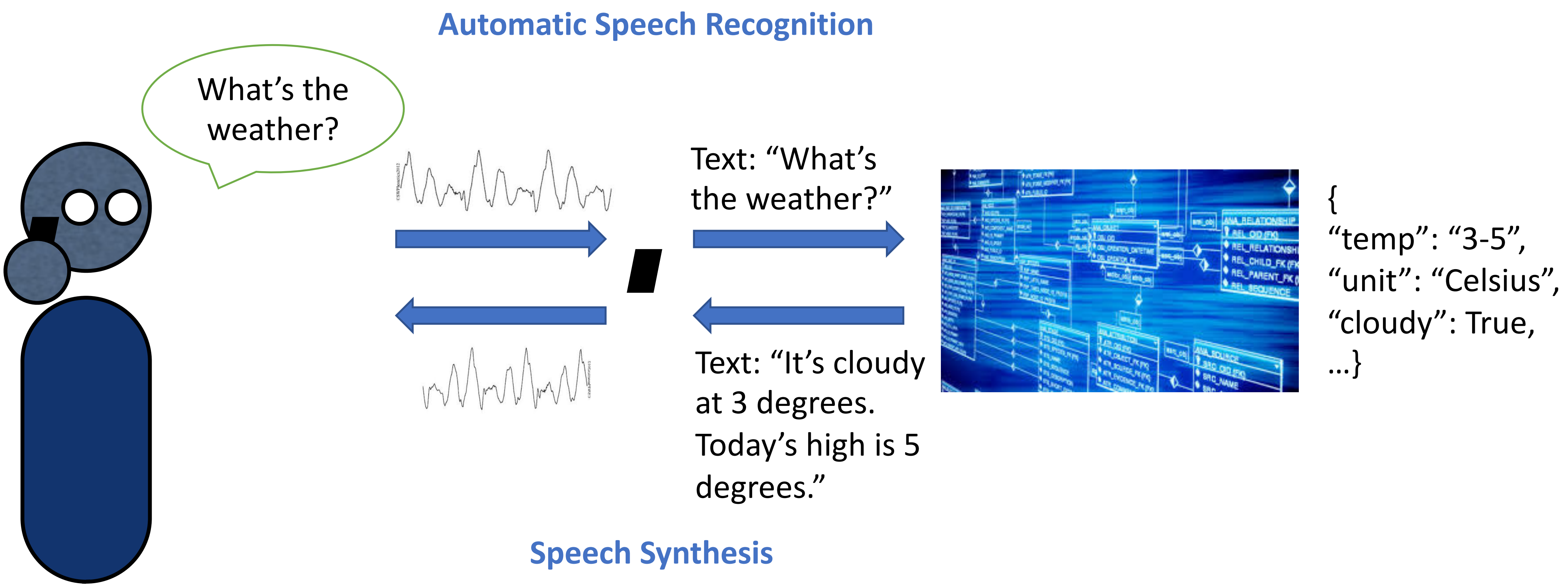
Computer Science
UNIVERSITY OF TORONTO

Agenda

- General introduction (← this tutorial)
  - Speech technology
  - Speech signal features, MFCC
  - Acoustic phonetics
- Speaker Recognition, Fitting to data, Gaussian Mixture Models
- Dynamic programming, WER, Levenshtein distance
- Misc. Q&A for A3

Applications of Speech Technology

Speech Technology: A Use Case



Challenges in speech data

- Co-articulation and dropped phonemes
- Intra- and Inter-speaker variability
- Lack of word boundaries
- Slurring, disfluency (e.g., ‘um’)
- Signal noise
- …

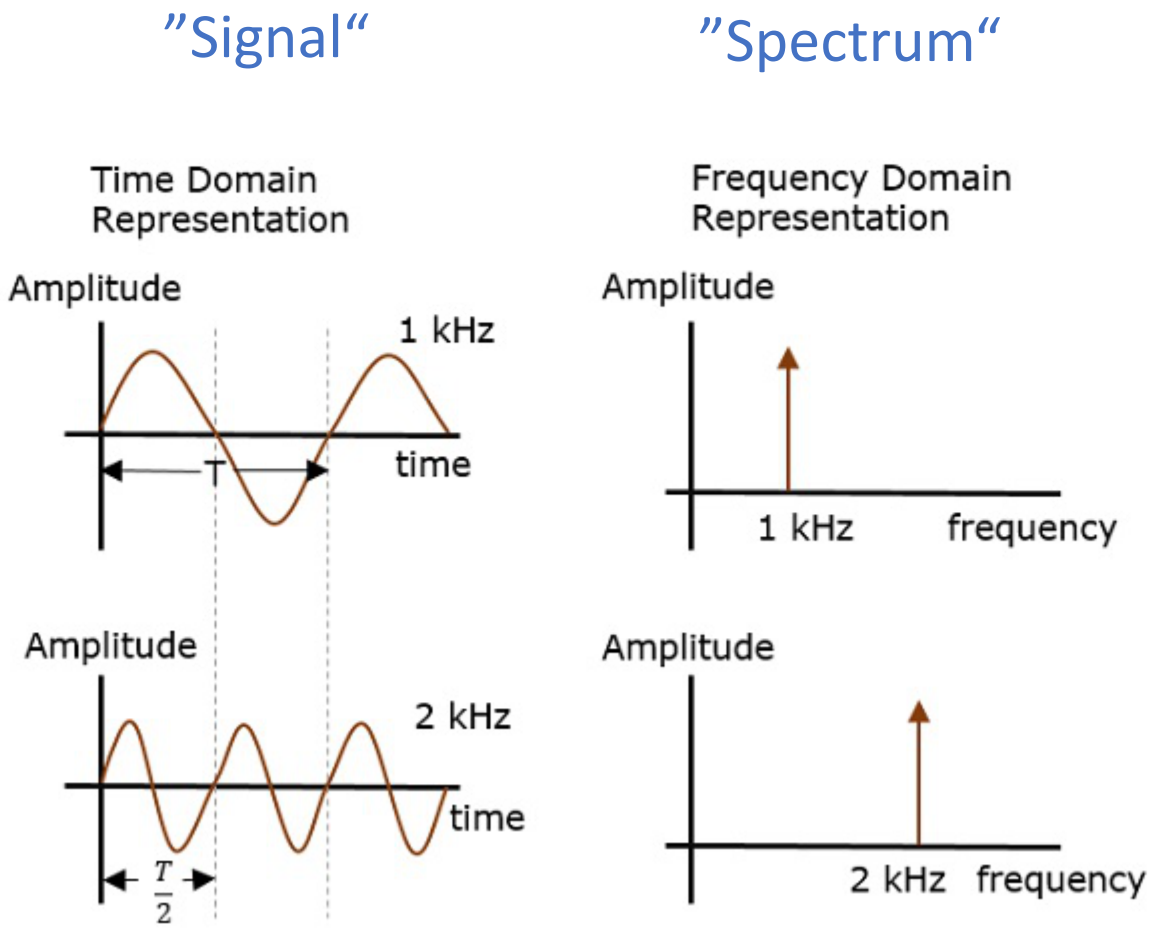
Automatic Speech Recognition

- Speech in, text out.
- This is done by machine learning:
  - Compute ("extract") features from acoustic signals.
  - Then use e.g., deep neural networks to predict the text.
- Some useful features:
  - Formants
  - MFCC
  - … …

This tutorial

Waveforms

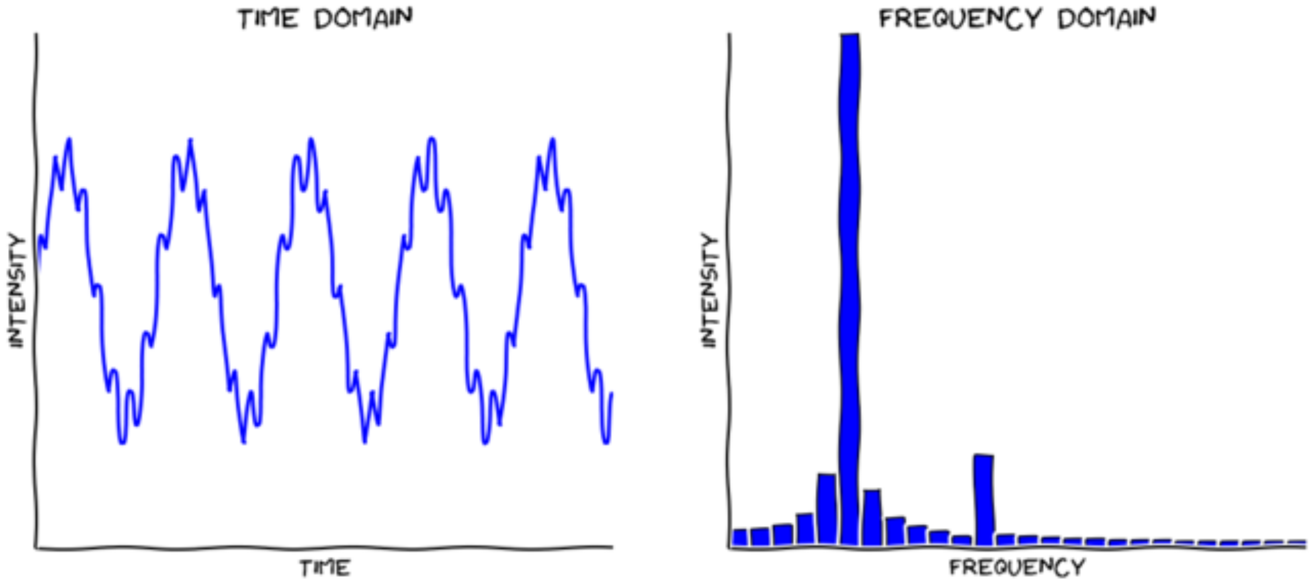
- Waveforms are recorded in the *time domain*.
- Signals can be conveniently analyzed in the *frequency domain*.
- Convert time-domain representation into frequency domain? Fourier transform.
  - FT computes the spectrum.



Computer Science
UNIVERSITY OF TORONTO

Fourier Transform Properties

- Linearity:
  - $F(ax_1(t) + bx_2(t)) = aF\left(x_1(t)\right) + bF(x_2(t))$
  - Complex functions can be analyzed in superpositions of simple ones.



Filtering a Signal

- Convolution theorem
  - Let $g(t) = x(t) * y(t)$
  - Let’s write Fourier Transform results as e.g., $G(\omega) = F(g(t))$ then:
  - Then $G(\omega) = X(\omega) \times Y(\omega)$

$$x(t) \Rightarrow \boxed{y(t)} \Rightarrow g(t)$$

- This process is “pass the signal x(t) through the filter $y(t)$ “

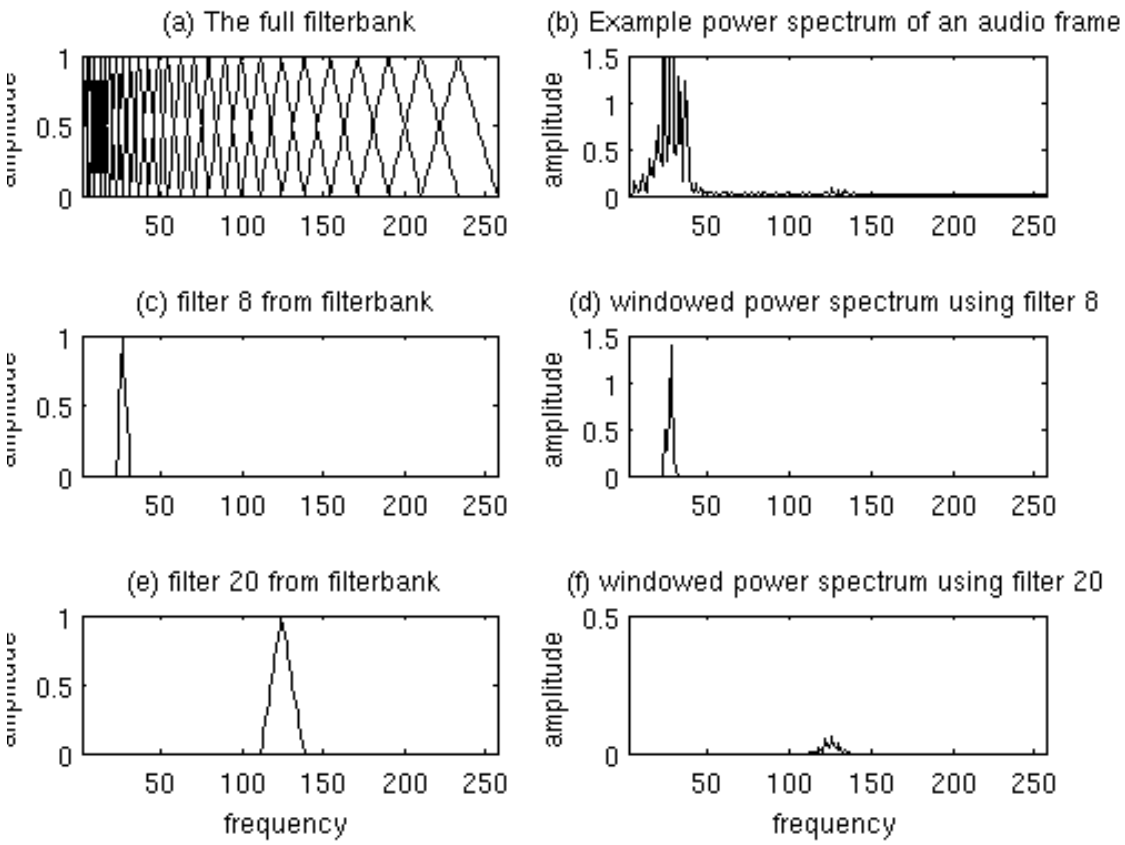
Computer Science
UNIVERSITY OF TORONTO

Frequency domain is about energy

- Parseval's theorem for Fourier Transform
  - Energy in time domain == Energy in frequency domain
  - $\int_{-\infty}^{\infty} x^2(t)dt = \int_{-\infty}^{\infty} X^2(\omega)d\omega$
  - $x(t)$ and $X(\omega)$ are two representations of the same signal.
  - *If the signal involves periodic waves, do a Fourier Transform. Computing energy is easier in the spectrum.*
- Now we can start deriving the MFCCs of a *short* speech sample.
  - For longer speech samples: segment into ~35ms long samples.
- **MFCC Step 1**: Fourier transform into the *frequency domain*.

Triangular Overlapping windows

- Real-world speech constitutes a mixture of various frequencies!
- **MFCC Step 2**: Filter with a mixture of overlapping *triangular windows*.
  - These windows are the filter banks.
  - Each filtered signal is approximately at one “pitch”.



Computer Science
UNIVERSITY OF TORONTO

The Mel Scale

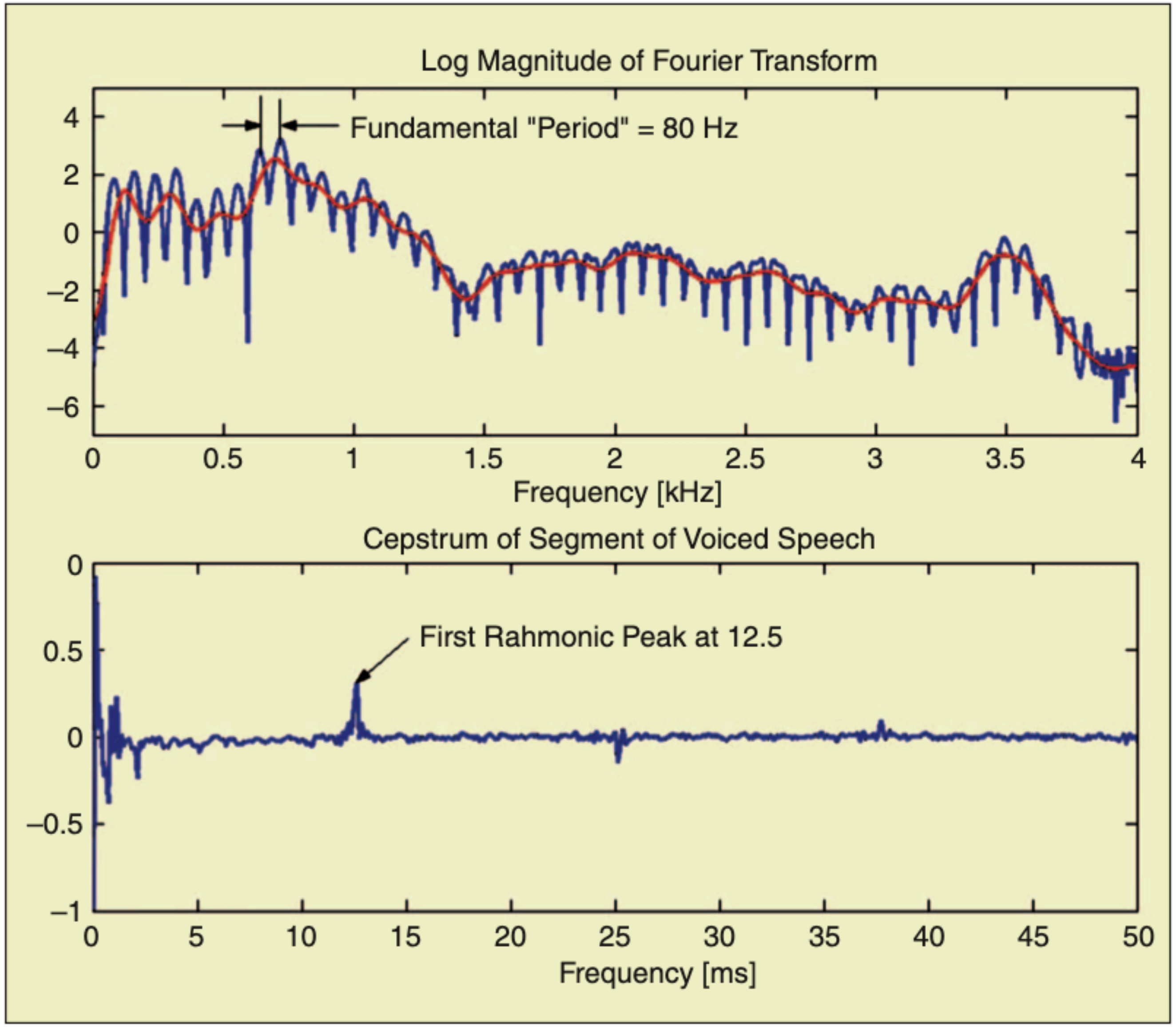
- Human hearing capacity has limited range.
  - Usually at 20 Hz to 20,000 Hz
- Human ears are more sensitive to changes in pitch at low frequencies.
  - … almost exponentially more sensitive.
  - This also explains why we take more windows at low frequency.
  - Intuition: take the log of the frequencies.

- **MFCC Step 3**: Convert the frequencies to Mel scale

$$M(f) = 1125 \ln(1 + \frac{f}{700})$$

Spectrum -> Cepstrum

- In voiced speech signals, there are *periodic signals* in spectrums!
- **MFCC Step 4**: Take the spectrum of the spectrum.
  - People called it the cepstrum
  - We are now in the *quefrency* domain (Bogert et al, 1963)
- *Voilà*, we got the Mel-Frequency Cepstral Coefficients.



Using MFCC Features

- What I actually do:

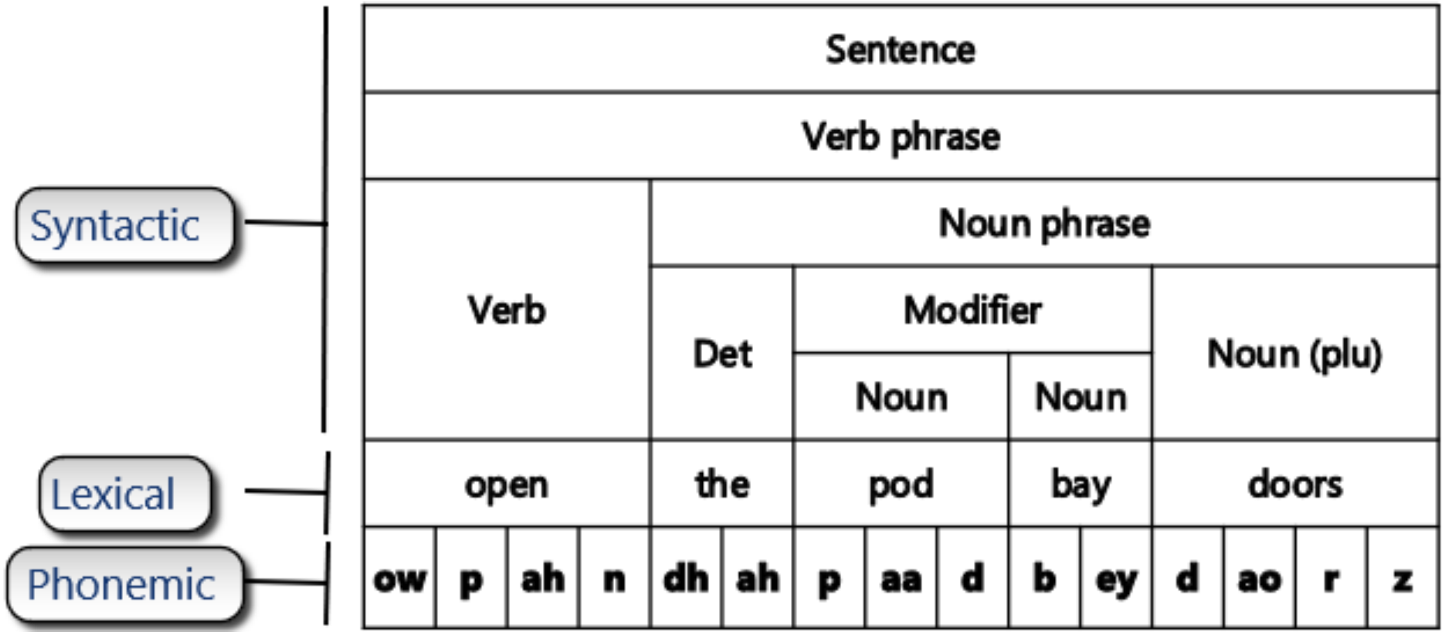
```
from python_speech_features import mfcc
from python_speech_features import logfbank
import scipy.io.wavfile as wav

(rate,sig) = wav.read("file.wav")
mfcc_feat = mfcc(sig,rate)
```

References

- The dummy’s guide to MFCC, by Pratheeksha
- MFCC tutorial by practical cryptography
- From frequency to Quefrency: a history of the Cepstrum

Acoustic Phonetics: Phonemes

- Words are formed by phonemes (aka ‘phones’),
  e.g., ‘pod’ = /p aa d/
- Words have different pronunciations. and in practice we can never be certain of which phones were uttered, nor their start/stop points.



Phonetic Alphabets

- International Phonetic Association (IPA)
  - Can represent sounds in all languages
  - Contains non-ASCII characters
- ARPAbet
  - One of the earliest attempts at encoding English for early speech recognition.
- TIMIT/CMU
  - Very popular among modern databases for speech recognition.

Example phonetic alphabets

IPA	CMU	TIMIT	Example	IPA symbol name
[ɑ]	AA	aa	father, hot	script a
[æ]	AE	ae	had	digraph
[ə]	AH0	ax	sofa	schwa (common in unstressed syllables)
[ʌ]	AH1	ah	but	turned v
[ɔ:]	AO	ao	caught	open o – Note, many speakers of Am. Eng. do not distinguish between [ɔ:] and [ɑ]. If your "caught" and "cot" sound the same, you do not.
[ɛ]	EH	eh	head	epsilon
[ɪ]	IH	ih	hid	small capital I
[i:]	IY	iy	heed	lowercase i
[ʊ]	UH	uh	hood, book	upsilon
[u:]	UW	uw	boot	lowercase u
[aɪ]	AY	ay	hide	
[aʊ]	AW	aw	how	
[eɪ]	EY	ey	today	
[oʊ]	OW	ow	hoed	
[ɔɪ]	OY	oy	joy, ahoy	
[ɚ]	ER0	axr	herself	schwar (schwa changed by following r)
[ɝ]	ER1	er	bird	reverse epsilon right hook

IPA	CMU	TIMIT	Example	IPA symbol name
[ŋ]	NG	ng	sing song	eng or angma
[ʃ]	SH	sh	sheet, wish	esh or long s
[tʃ]	CH	ch	cheese	
[j]	Y	y	yellow	lowercase j
[ʒ]	ZJ	zh	vision	long z or yogh
[dʒ]	JH	jh	judge	
[ð]	DH	dh	thee, this	eth

The other consonants are transcribed as you would expect

i.e., p, b, m, t, d, n, k, g, s, z, f, v, w, h

Summary

- Speech technology
- Speech signal features, MFCC
- Acoustic phonetics

Any questions?