Speech Recognition

DATA AND ARTIFICIAL INTELLIGENCE

speech-to-text

Learning Objectives

By the end of this lesson, you will be able to:

- Use speech to text mechanism
- Explain MFCC process
- Apply CNNs and RNNs for text recognition

DATA AND
ARTIFICIAL INTELLIGENCE

Introduction to speech-to-text

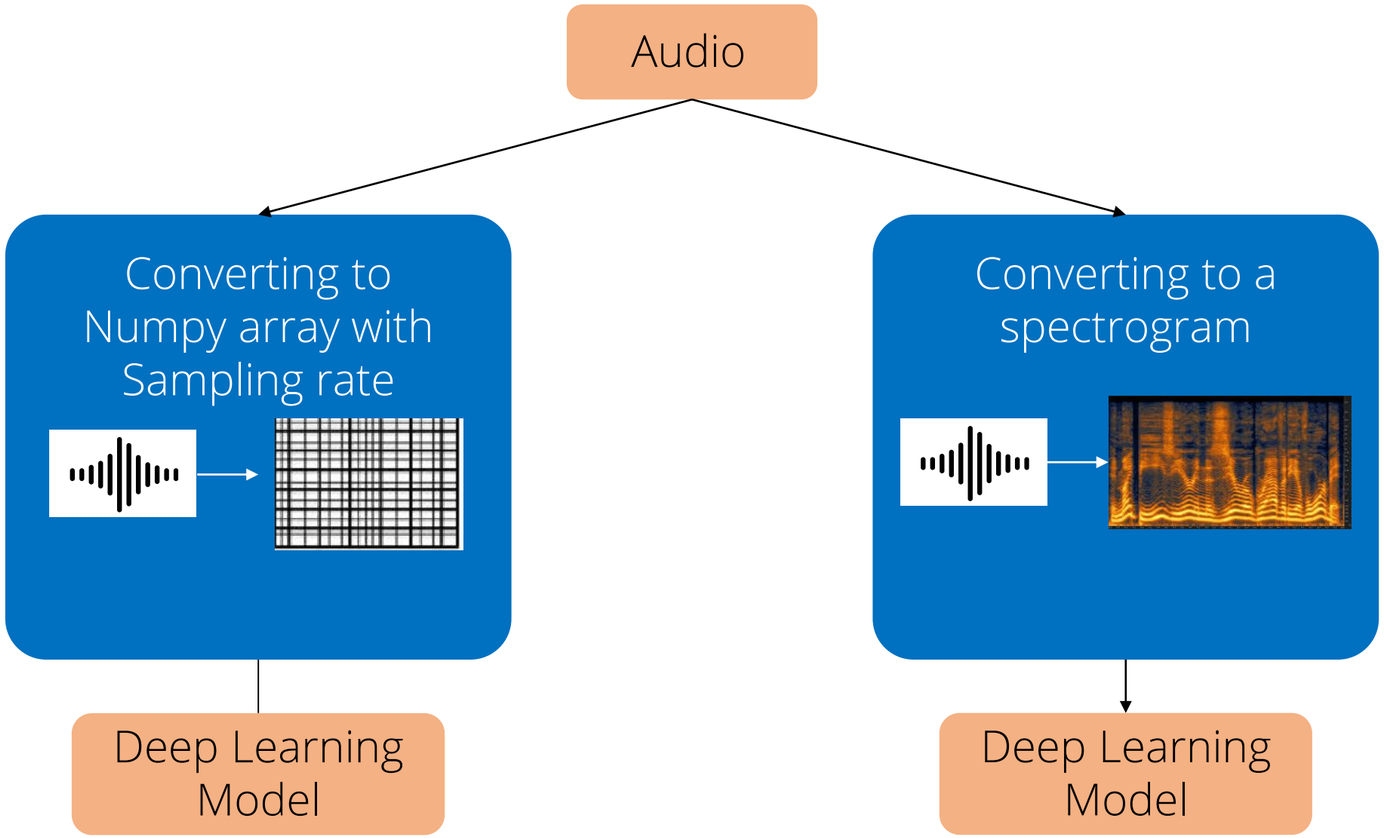
What Is Speech To Text Method?

In this method, you recognize the words spoken by a person and convert the voice to a written text.

Why Speech To Text?

1. Visually impaired people can control devices using only voice
2. Record meetings by converting the speech to text transcripts
3. Extract subtitles out of audio or video files
4. Translate words from one language into another

Speech as Data



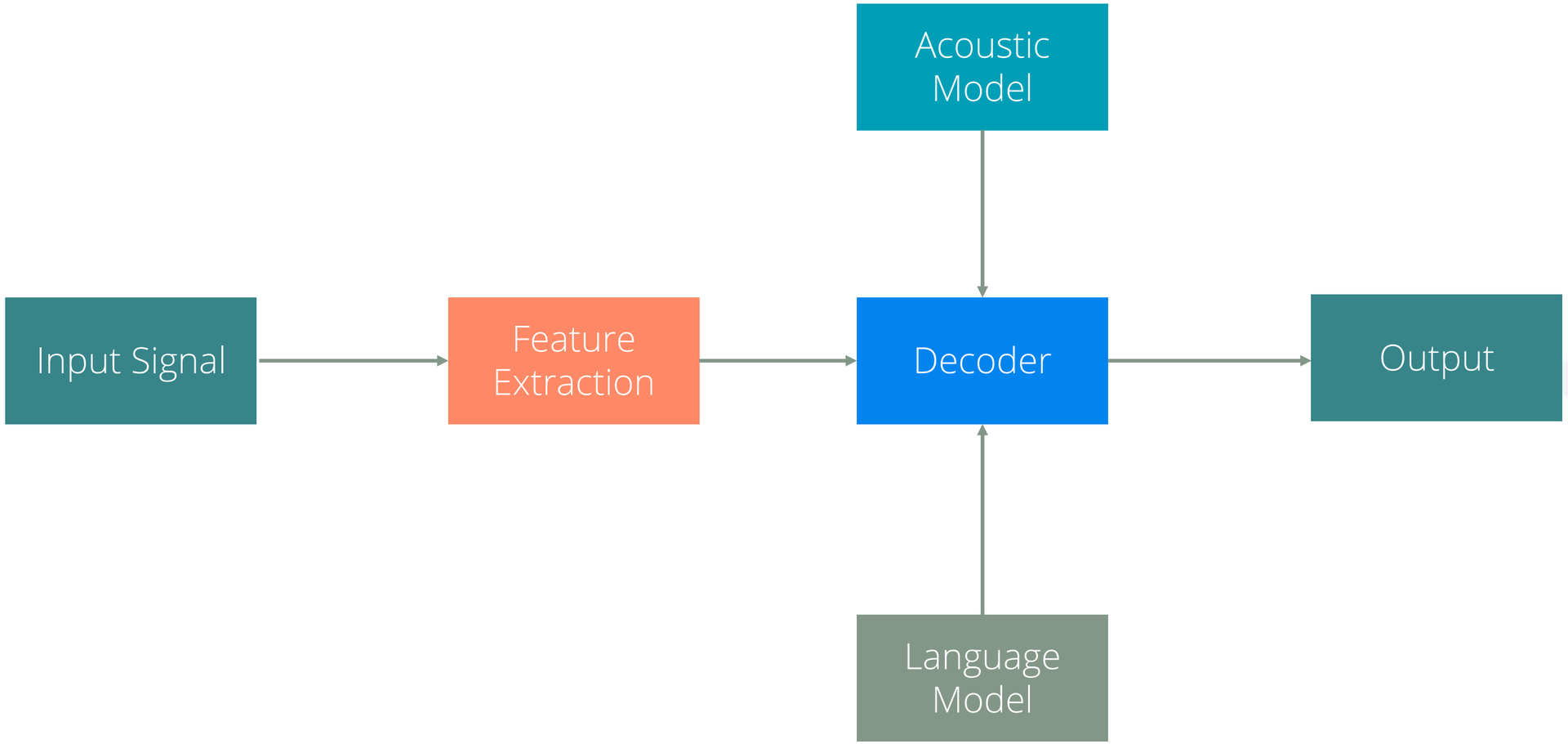
Note: Spectrogram also helps to convert speech to data.

DATA AND
ARTIFICIAL INTELLIGENCE

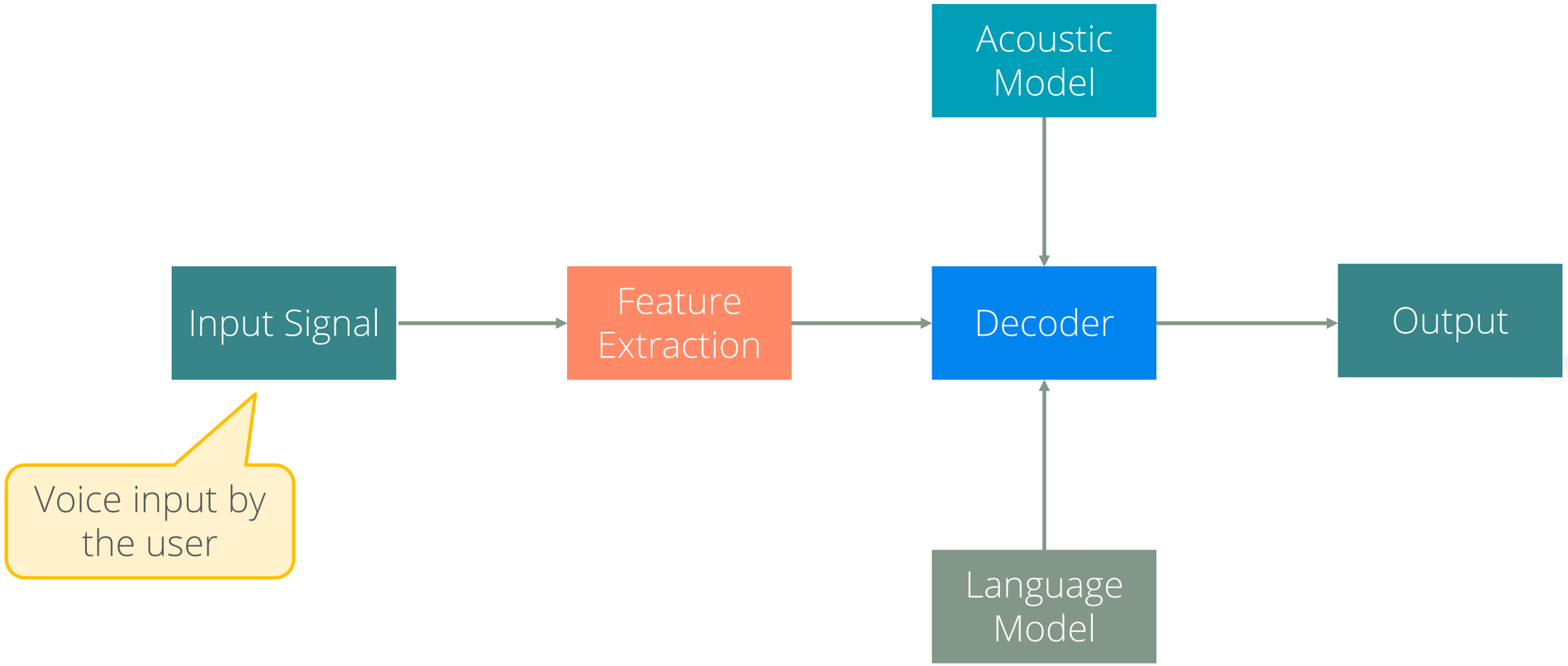
speech-to-text Model

speech-to-text Model

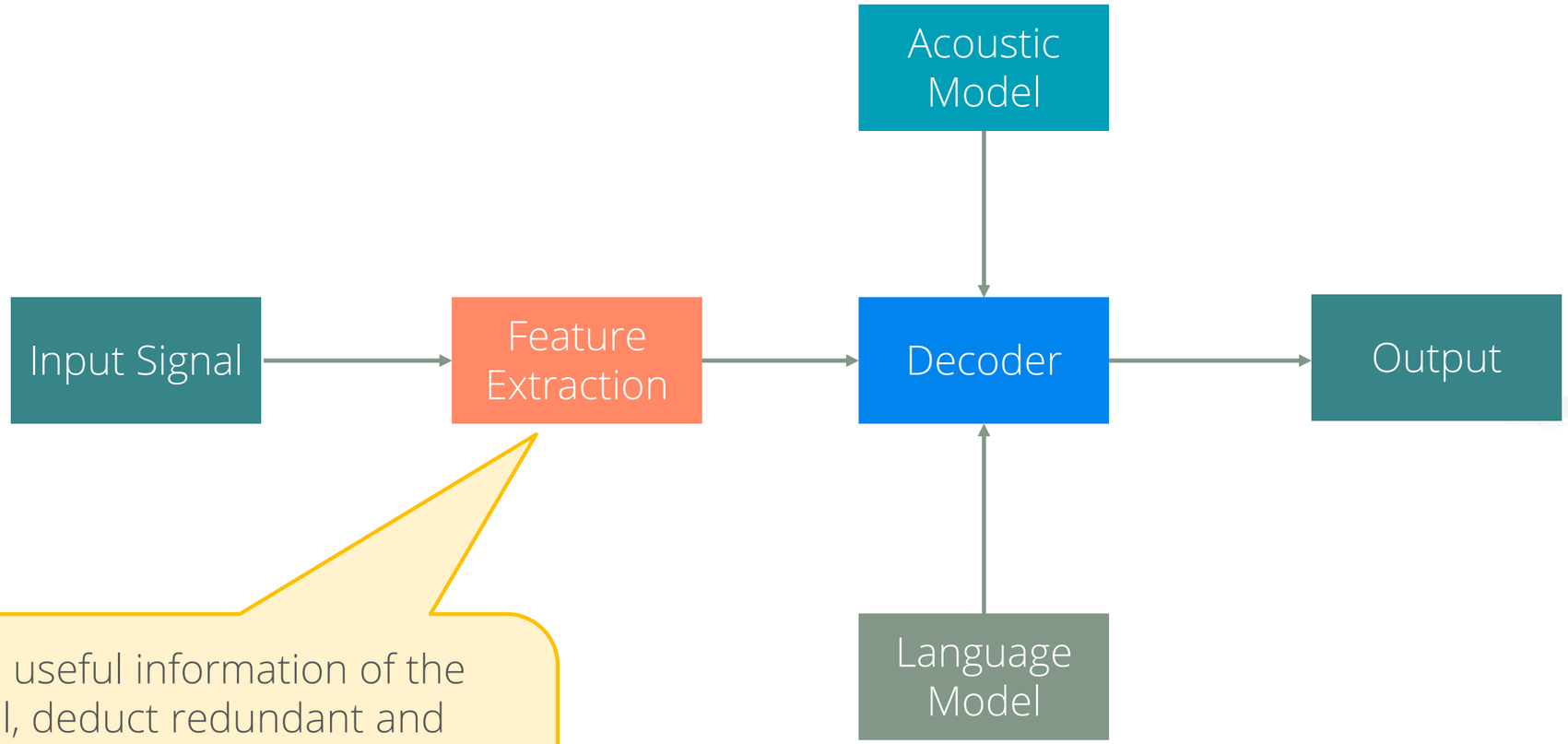
Below flowchart shows the block diagram of a typical speech to text model:



Input Signal

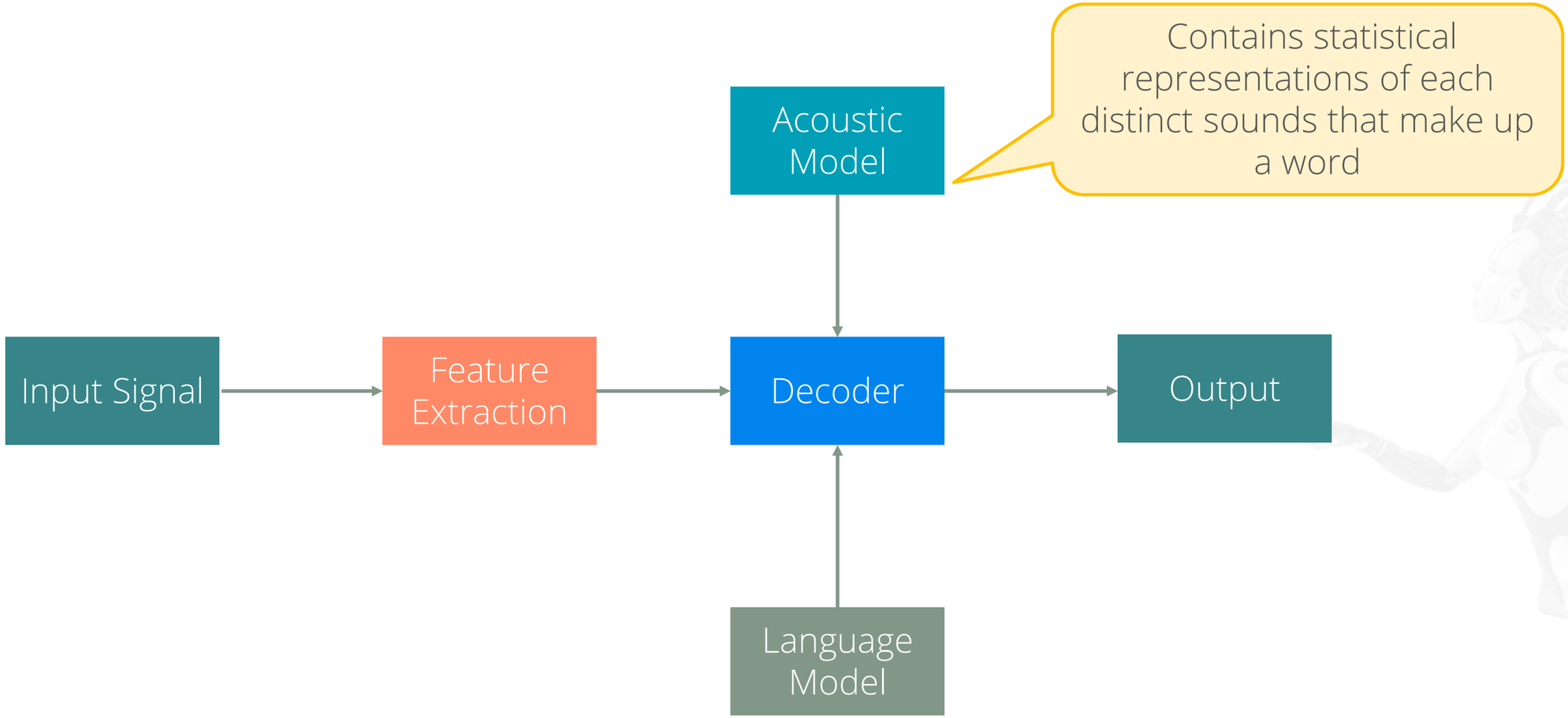


Feature Extraction

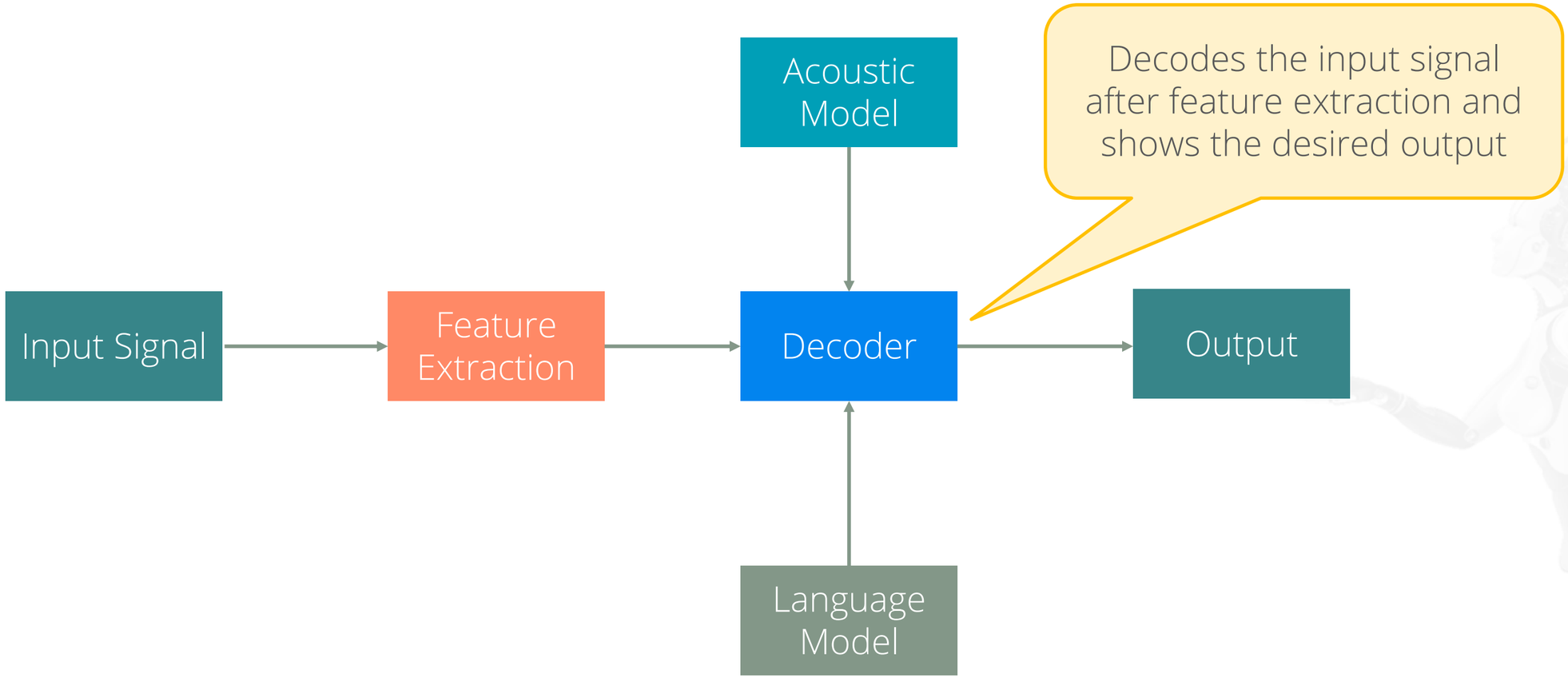


Retain useful information of the signal, deduct redundant and unwanted information, show less variation from one speaking environment to another

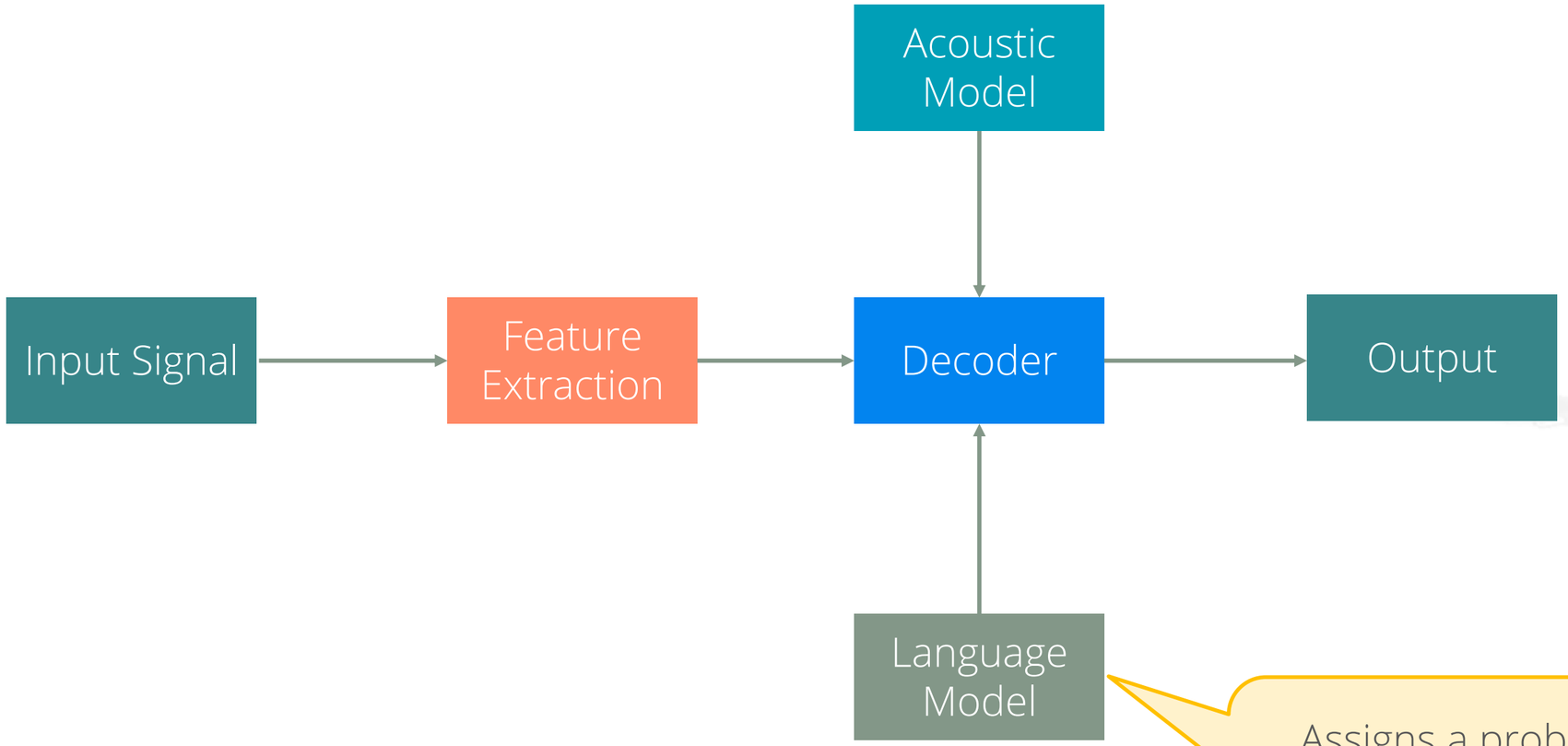
Acoustic Model



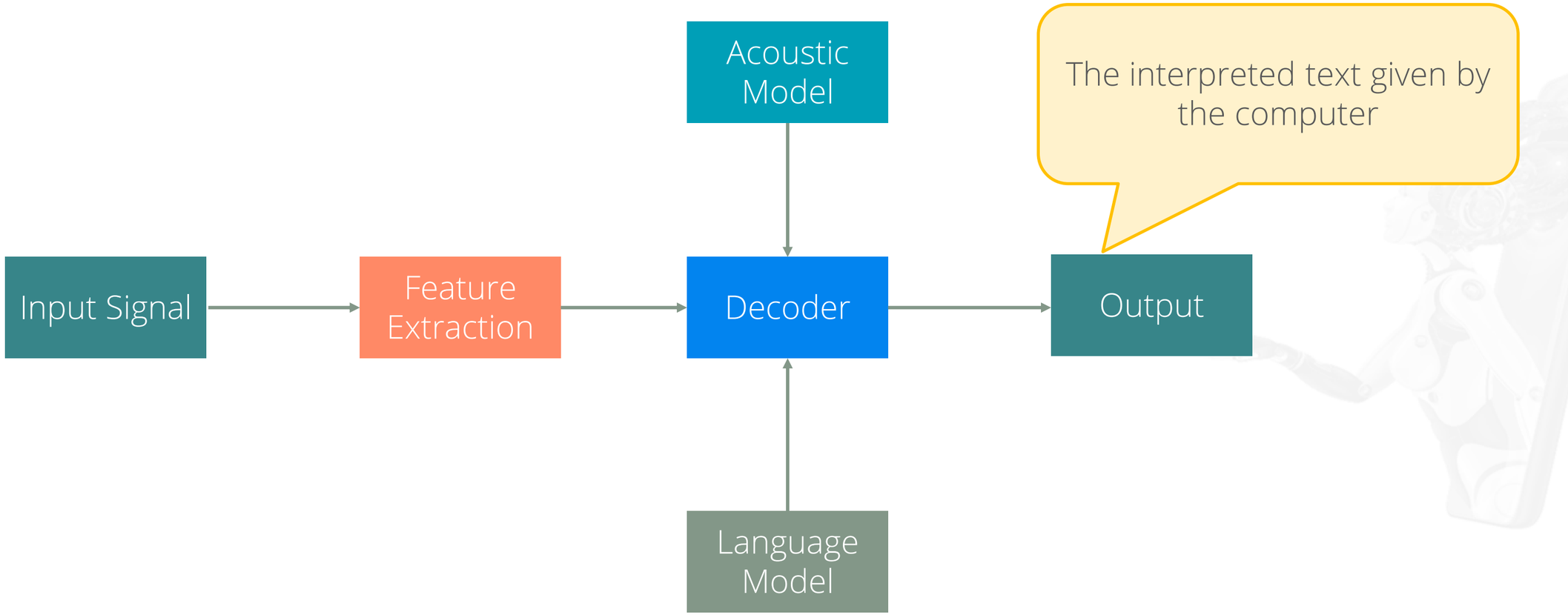
Decoder



Language Model



Output

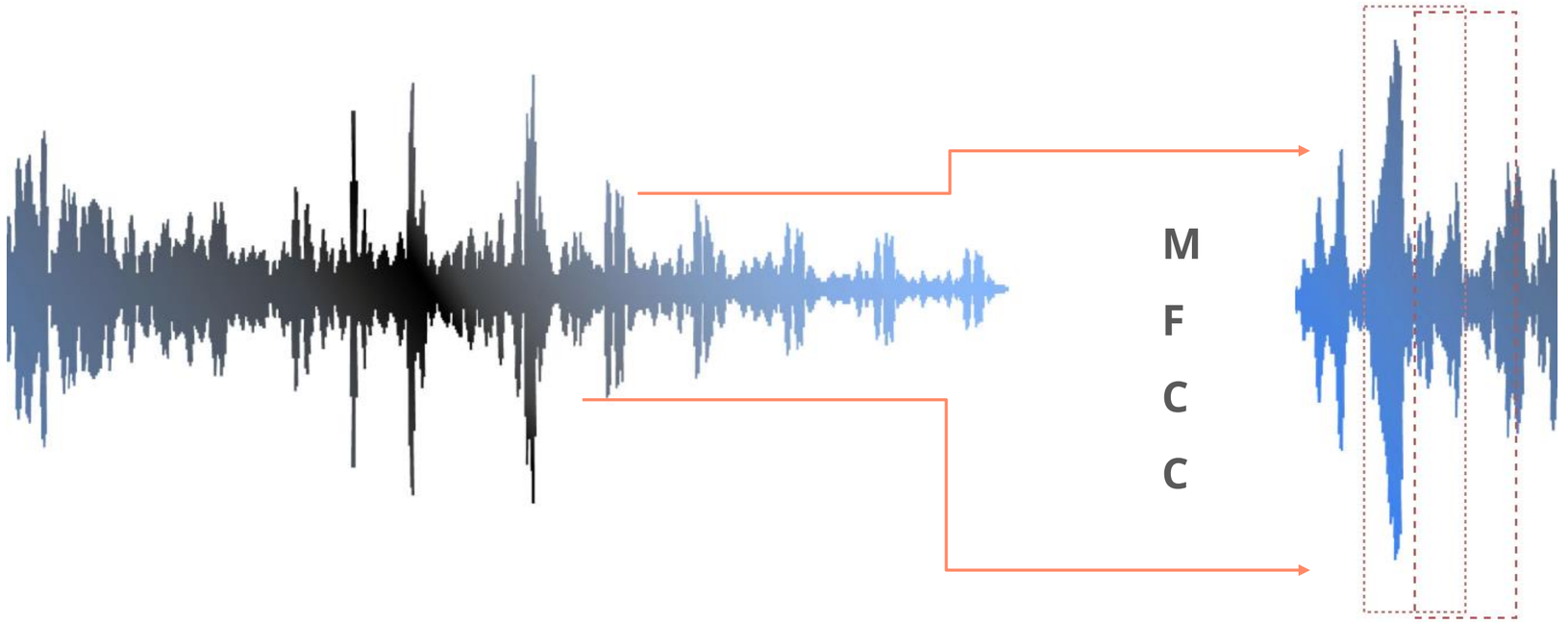


DATA AND
ARTIFICIAL INTELLIGENCE

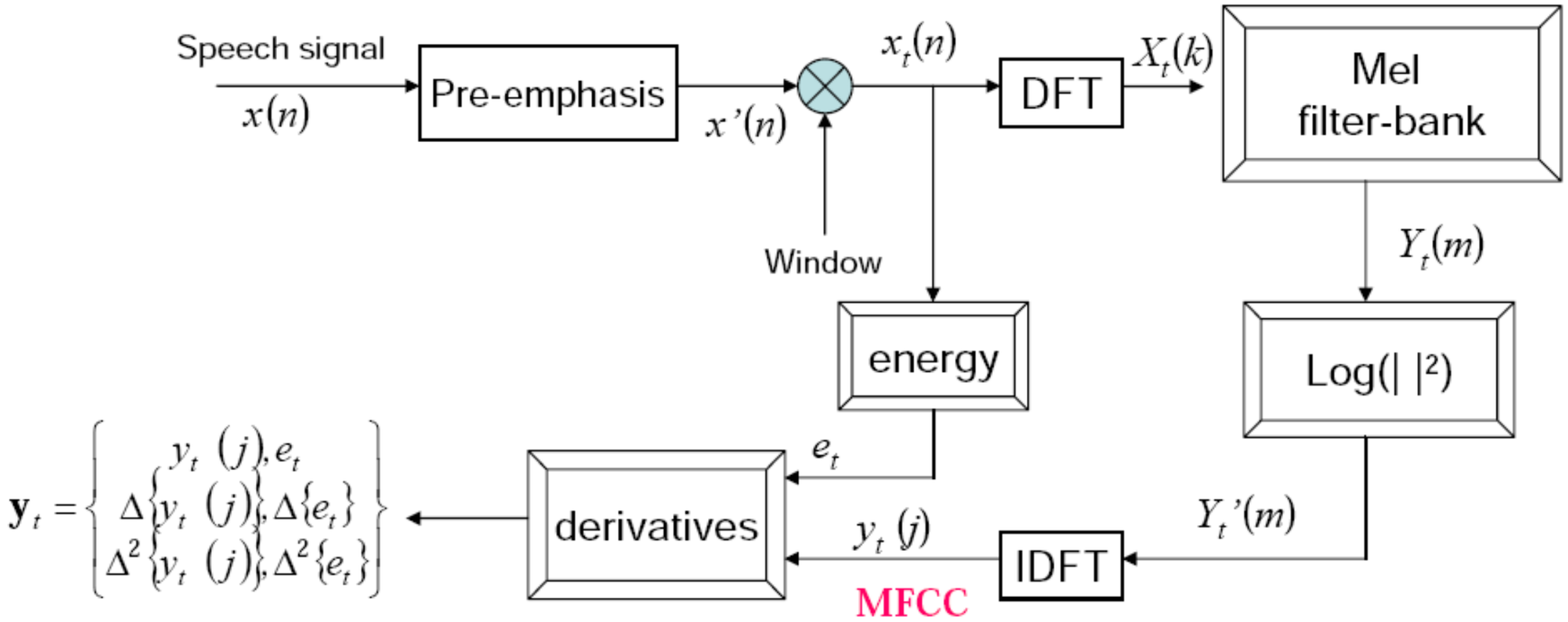
Feature Extraction

The Feature Extraction Process

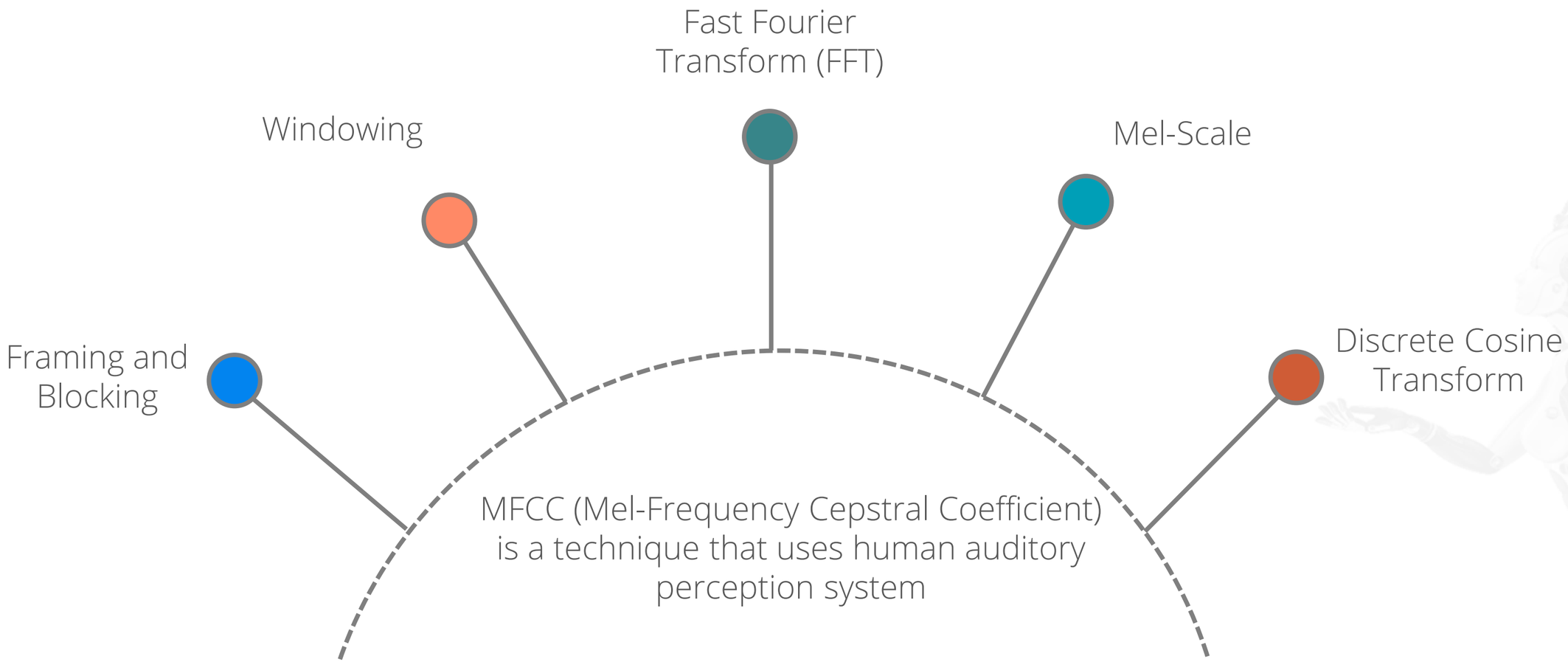
During this process, a set of features having correlation with speech is found. The features are computed by processing the acoustic waveform.



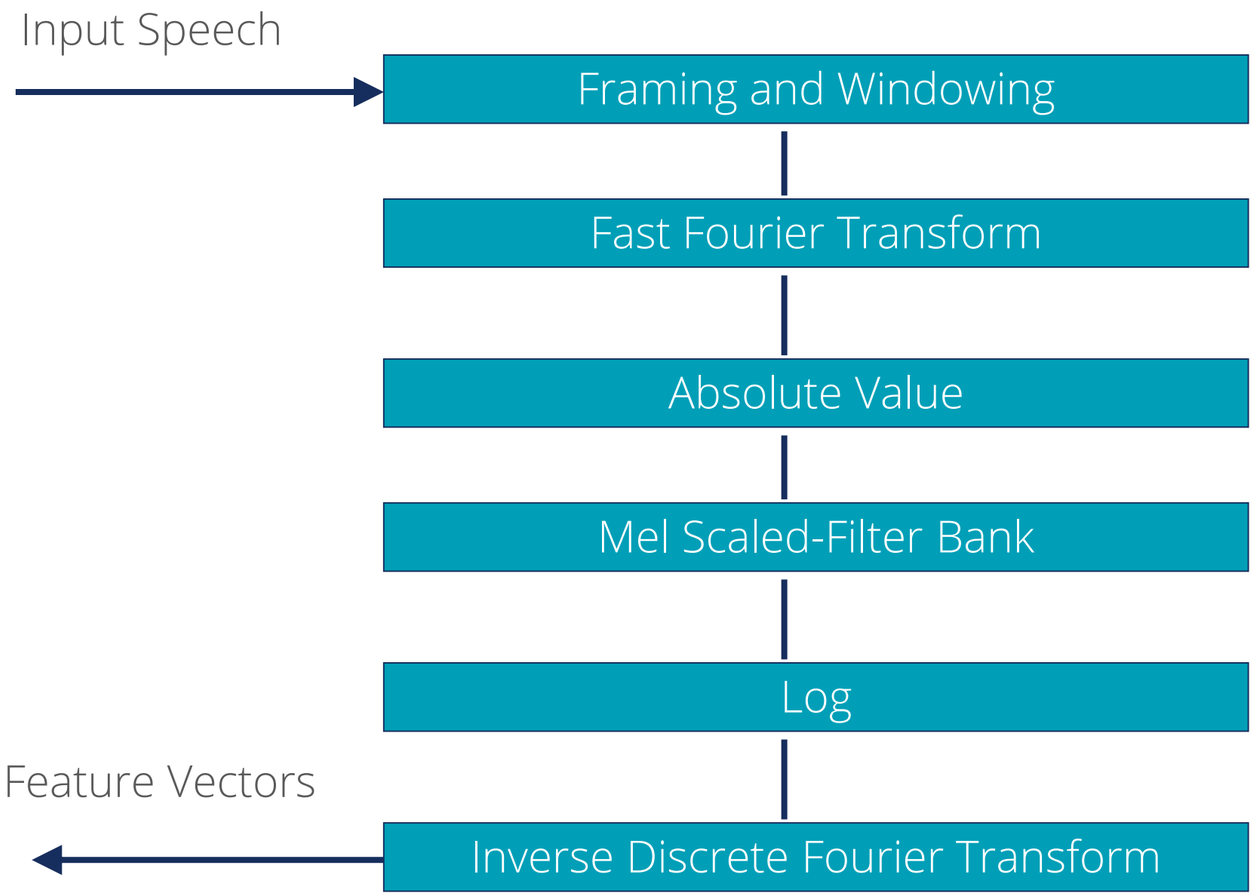
The MFCC Process



MFCC and Its Steps



The Feature Extraction Using MFCC

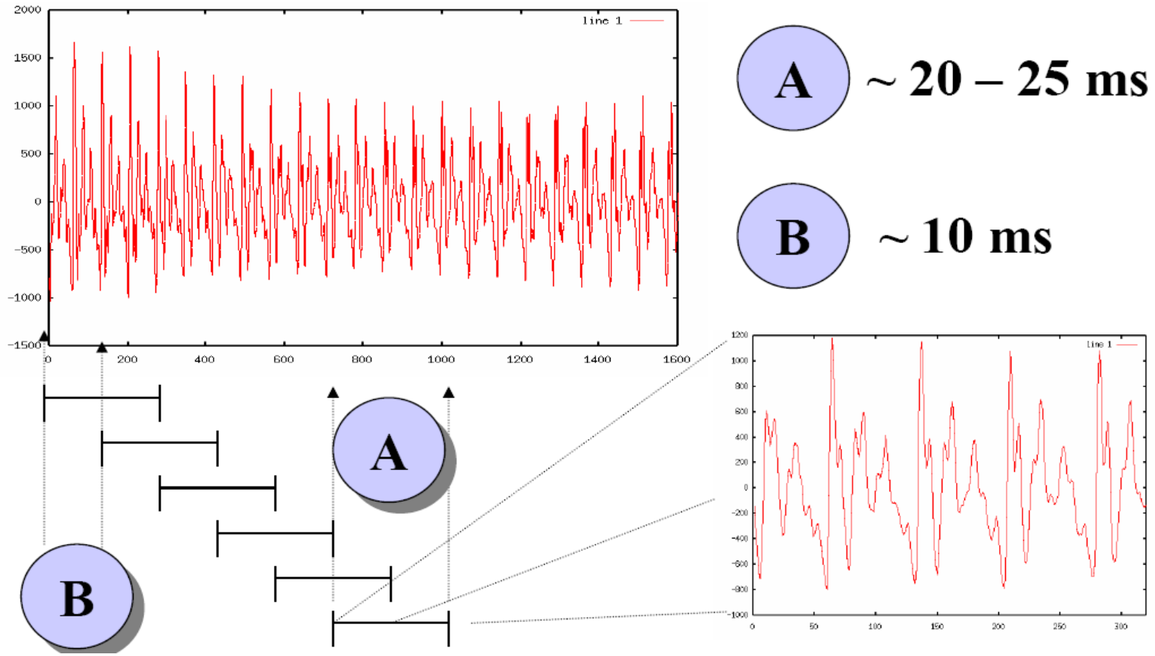


DATA AND
ARTIFICIAL INTELLIGENCE

Windowing

Windowing

It is used as a part of feature extraction to minimize the discontinuities in the signal.



Common Window Shapes

Rectangular Window

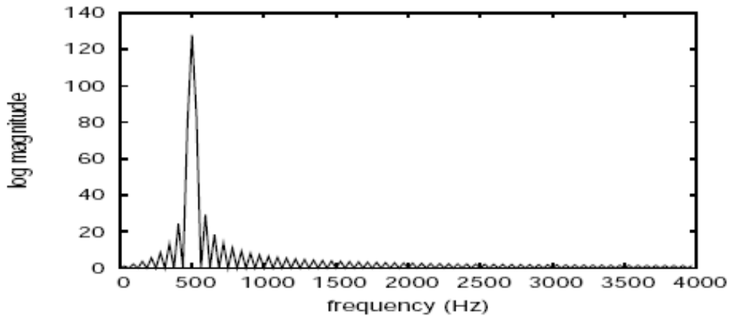
$$w[n] = \begin{cases} 1 & 0 \leq n \leq L-1 \\ 0 & \text{otherwise} \end{cases}$$

Hamming Window

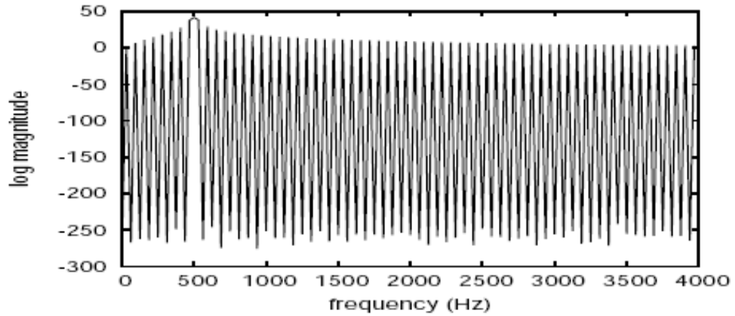
$$w[n] = \begin{cases} 0.54 - 0.46\cos\left(\frac{2\pi n}{L-1}\right) & 0 \leq n \leq L-1 \\ 0 & \text{otherwise} \end{cases}$$

Window in Time and Frequency Domain

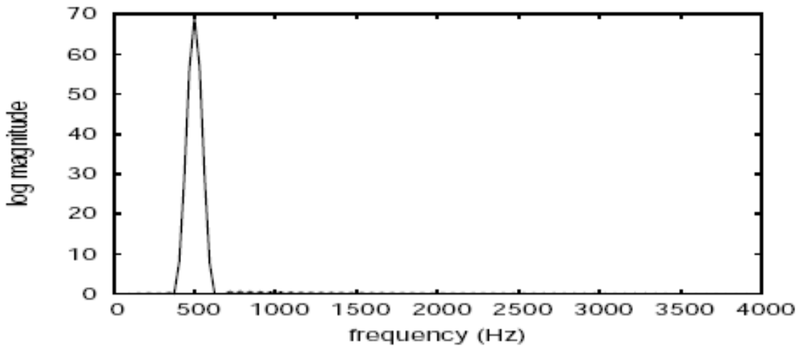
Rectangular Window



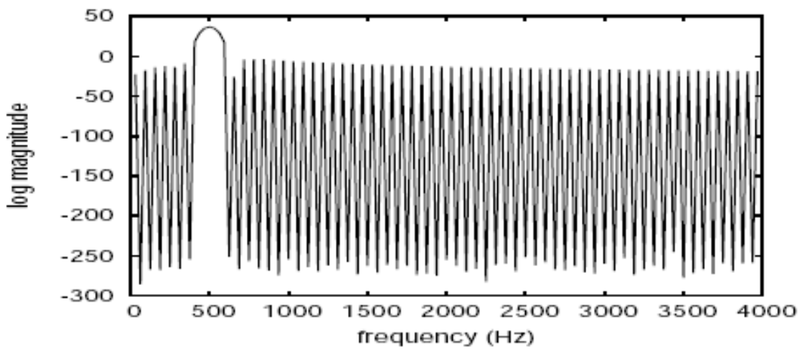
Rectangular Window



Hamming Window



Hamming Window



DATA AND
ARTIFICIAL INTELLIGENCE

Fourier Transform

Discrete Fourier Transform (DFT)

- **Input**
  - Windowed signal x[n]....x[m]
- **Output**
  - For each of N discrete frequency bands
  - A complex frequency X[k] representing magnitude and phase of that frequency component in the original signal
- **Discrete Fourier Transform**

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-j2\frac{\pi}{N}kn}$$

- **Standard Algorithm for Computing DFT**
  - Fast Fourier Transform (FFT) with complexity N*log(N)

Mel-Scale

What Is Mel-Scale?

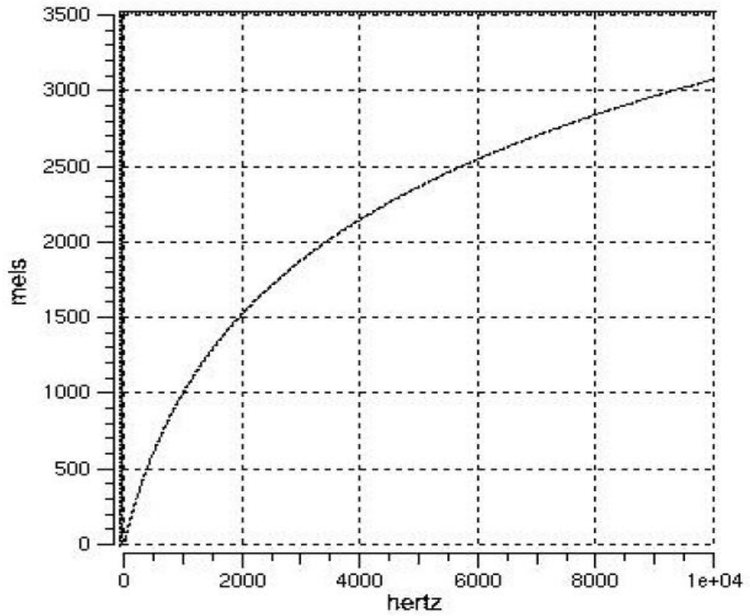
A unit of pitch that separates pairs of sounds with perceptually equidistant pitch. The Mel-Scale is approximately linear below 1kHz.

Mathematically:

$$\text{Mel (f)} = 2595 \log_{10}\left(1 + \frac{f}{100}\right)$$

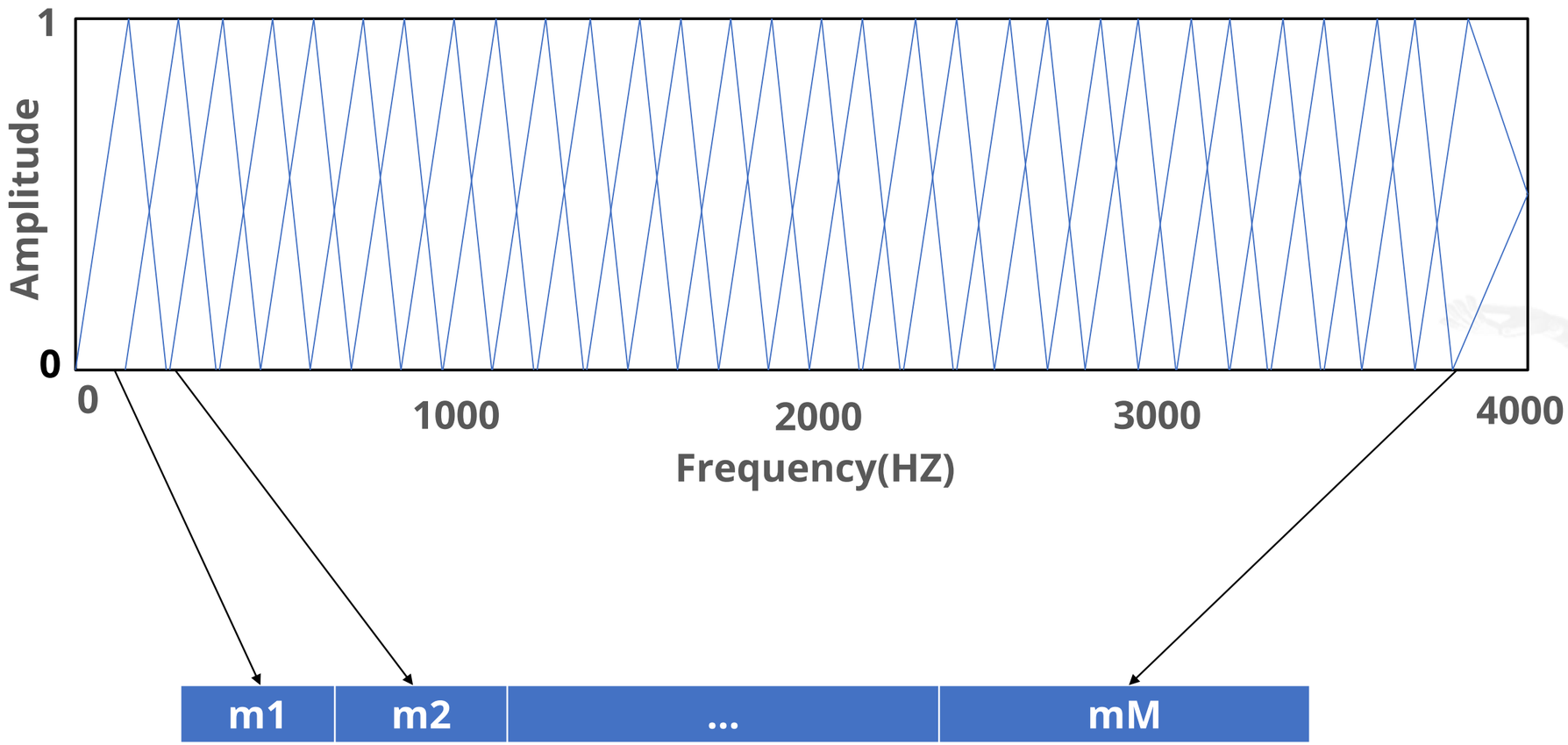
Using Mel-Scale

- Human hearing is not equally sensitive to all frequency bands
- It is less sensitive at higher frequencies that is roughly > 1000 Hz



What Is Mel Filter Bank?

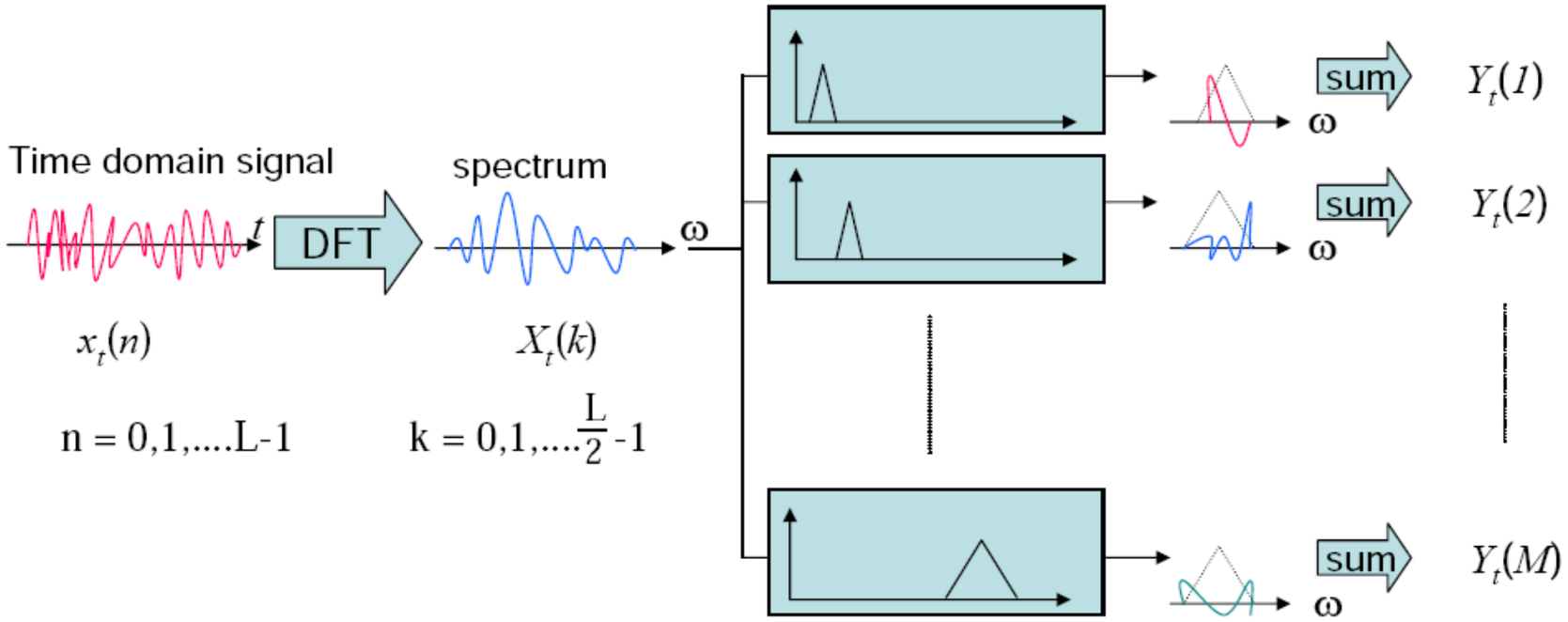
It is uniformly spaced before 1 kHz and logarithmically scaled after 1 kHz.



Mel Spectrum

Mel-Filter Bank Processing

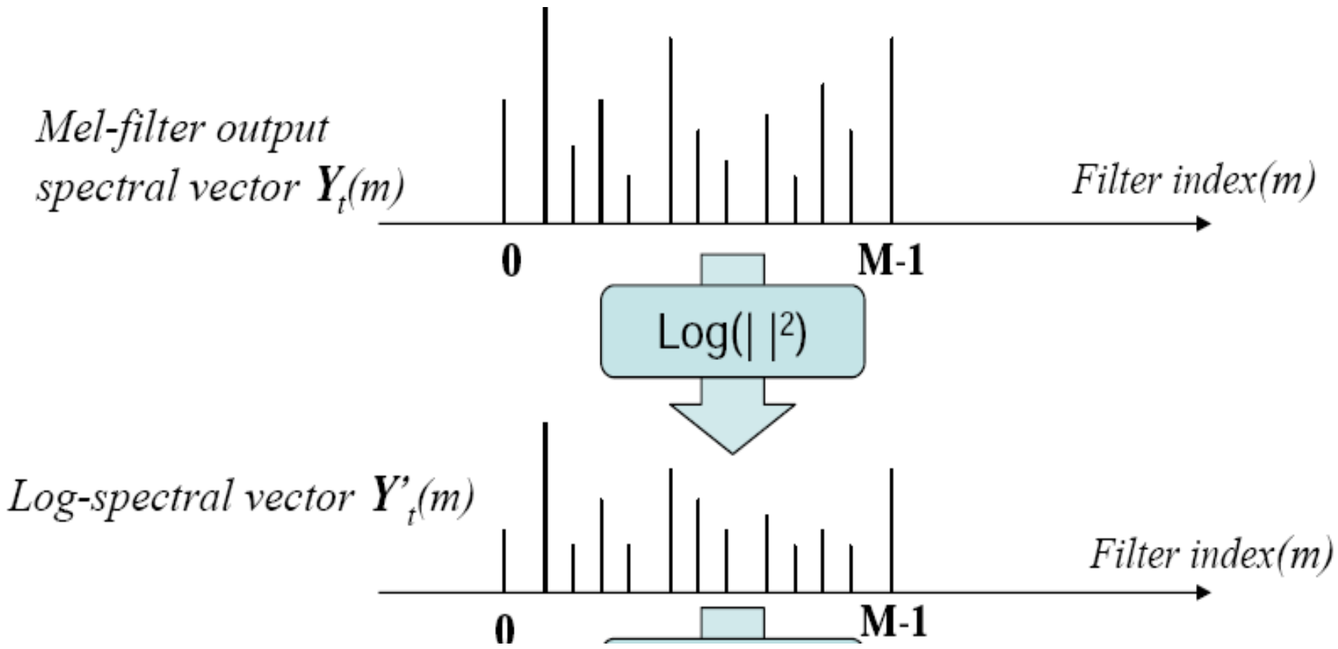
- Applies the bank of filters to the spectrum according to Mel-scale
- Each filter output is the sum of its filtered spectral components



Log Energy

The Log Energy Computation

It is computed by taking the logarithm of the square magnitude of the output of Mel-Filter bank.



Need for Log Energy Computation

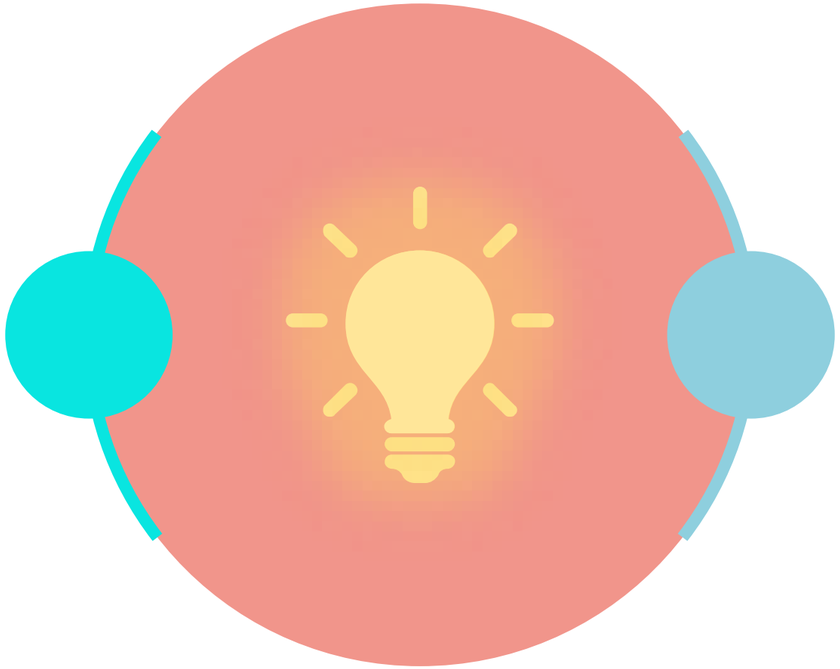
- Logarithm compresses dynamic range of values
- Human response to signal level is logarithmic
- It makes frequency estimates less sensitive to slight variations in input (power variation due to speaker's mouth moving closer to mic)
- Phase information is not helpful in speech

DATA AND
ARTIFICIAL INTELLIGENCE

Inverse Discrete Fourier Transform (IDFT)

Understanding Cepstrum

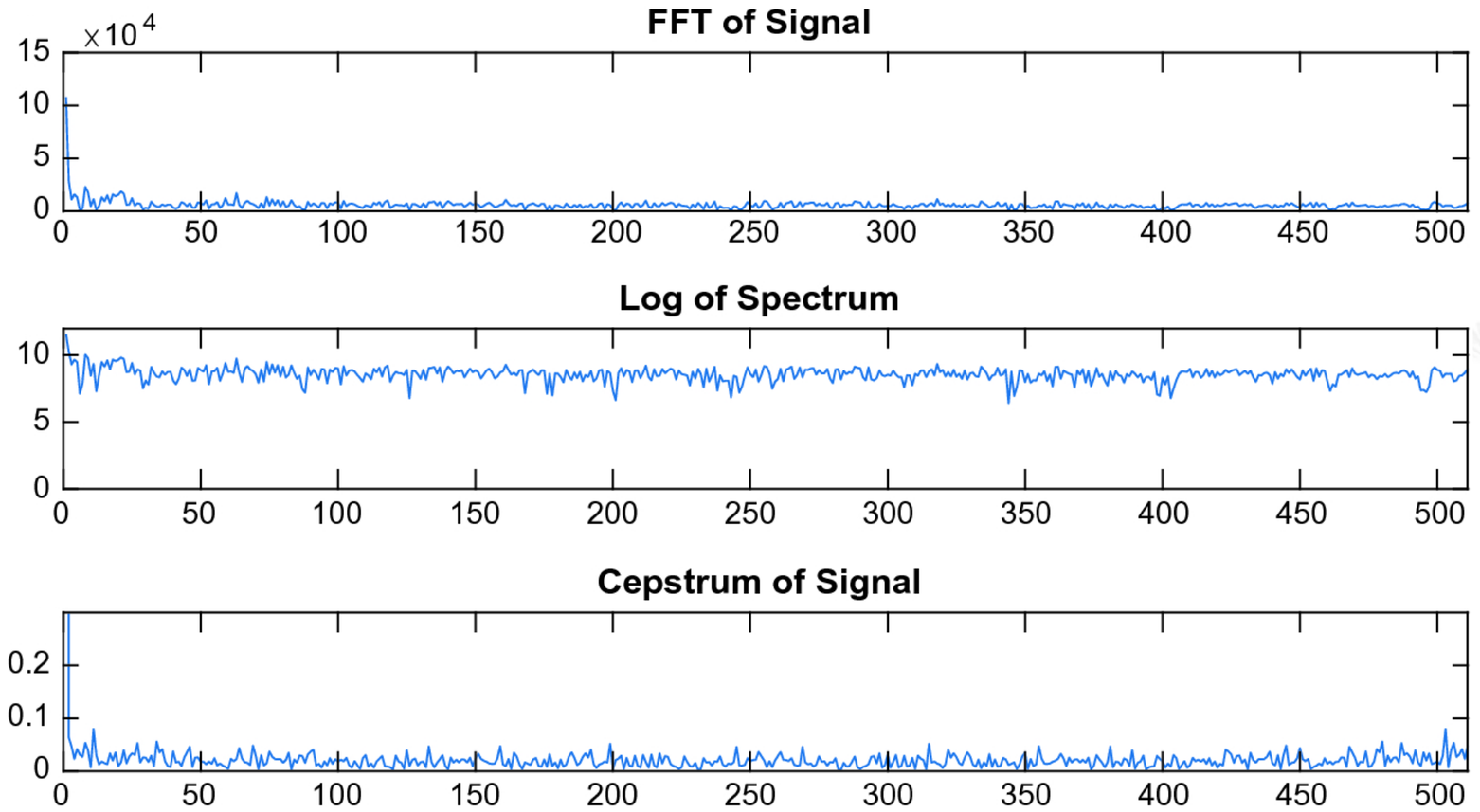
A speech waveform generated by a global source waveform



This waveform has a specific filtering characteristic when passed through a vocal tract

The Cepstrum

It is the spectrum of the log of the spectrum.



Mel-Frequency Cepstrum

- The cepstrum requires Fourier analysis but, you're going from frequency space back to time
- So, you apply inverse DFT

$$y_t[k] = \sum_{m=1}^{M} \log(|Y_t(m)|) \cos(k(m-0.5)\frac{\pi}{M}), \text{ k=0,...,J}$$

- Since the log power spectrum is real and symmetric, inverse DFT reduces to a **Discrete Cosine Transform (DCT)**

Dynamic Cepstral Coefficient

The cepstral coefficients do not capture energy. So, you add an energy feature.

$$Energy = \sum_{t=t_1}^{t_2} x^2[t]$$

Since the speech signal is not constant, you add changes (the slopes) in features. These are called **Delta** and **Double-Delta** features.

Why Is MFCC Popular?

- Computes efficiently
- Incorporates a perceptual Mel-Frequency scale
- Separates the source and the filter
- Decorrelates the features through IDFT (DCT)

DATA AND
ARTIFICIAL INTELLIGENCE

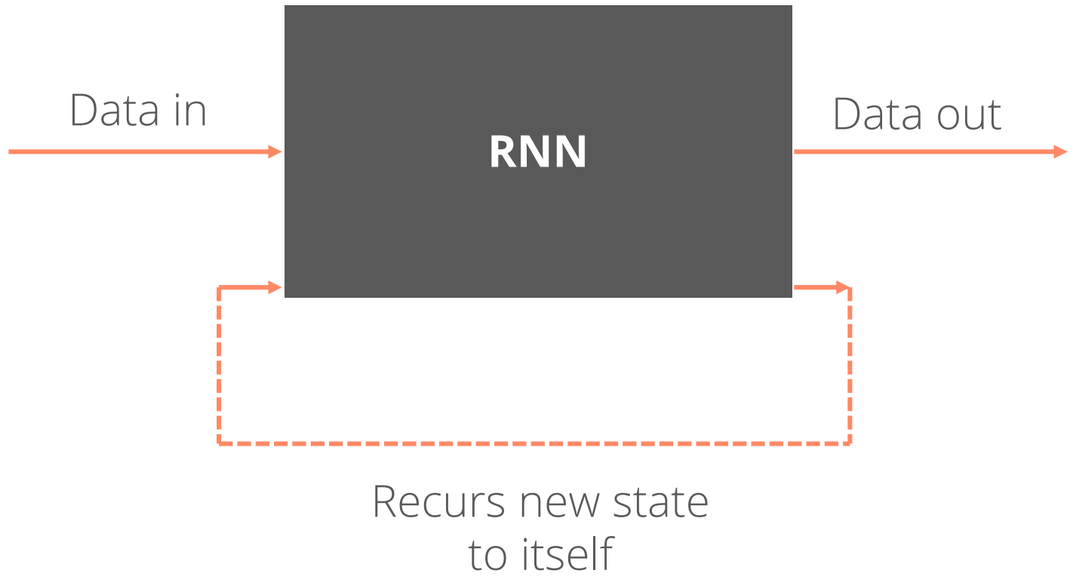
CNNs and RNNs for Speech Recognition

Visualizing Scene Text as an Image

Recognizing sequence-like objects such as scene text, handwriting can be done in the form of a sequence. Therefore, recognition of such objects can be cast as a sequence recognition problem.

RNNs for Sequential Data: Quick Recap

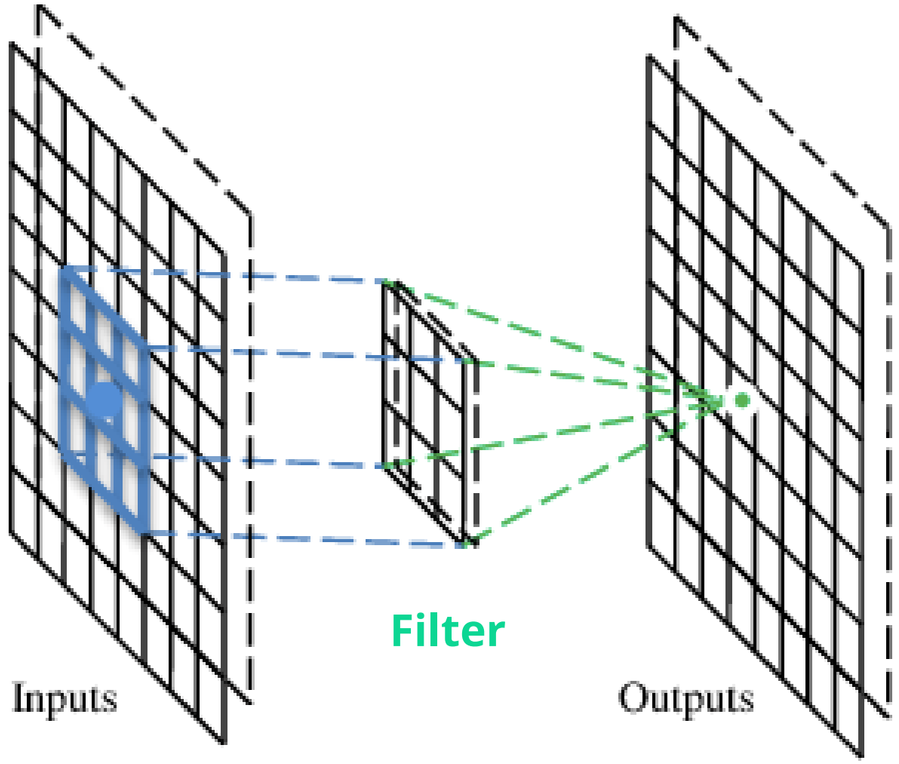
The RNN remembers the analysis done upto a given point by maintaining a **state**.



Note: You can think of the **state** as the **memory** of RNN which recurs into the net with each new input.

CNNs: Quick Recap

A CNN is a neural network with convolutional layers (among others). A convolutional layer has several filters that perform the convolution operation.

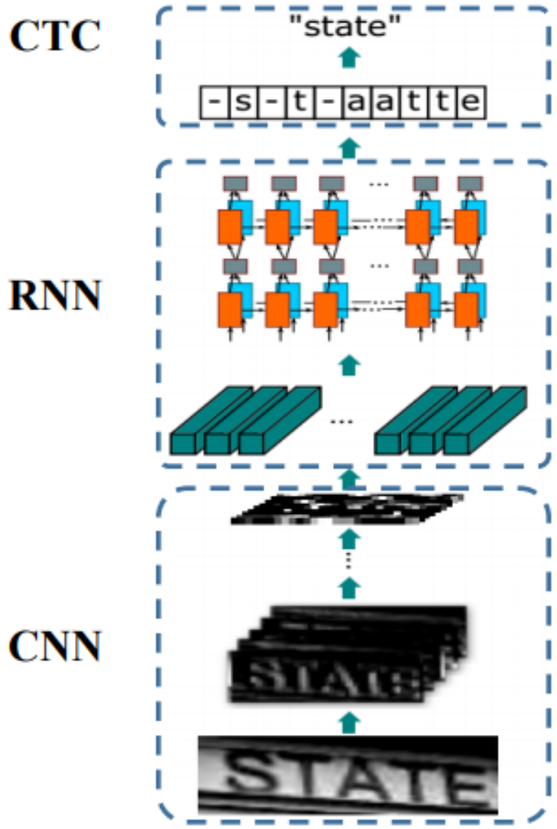


Convolutional Recurrent Neural Networks (CRNNs)

CNN encodes the information from the image and sends that data to the RNN-based decoder which decodes the data and outputs the corresponding text from the image.

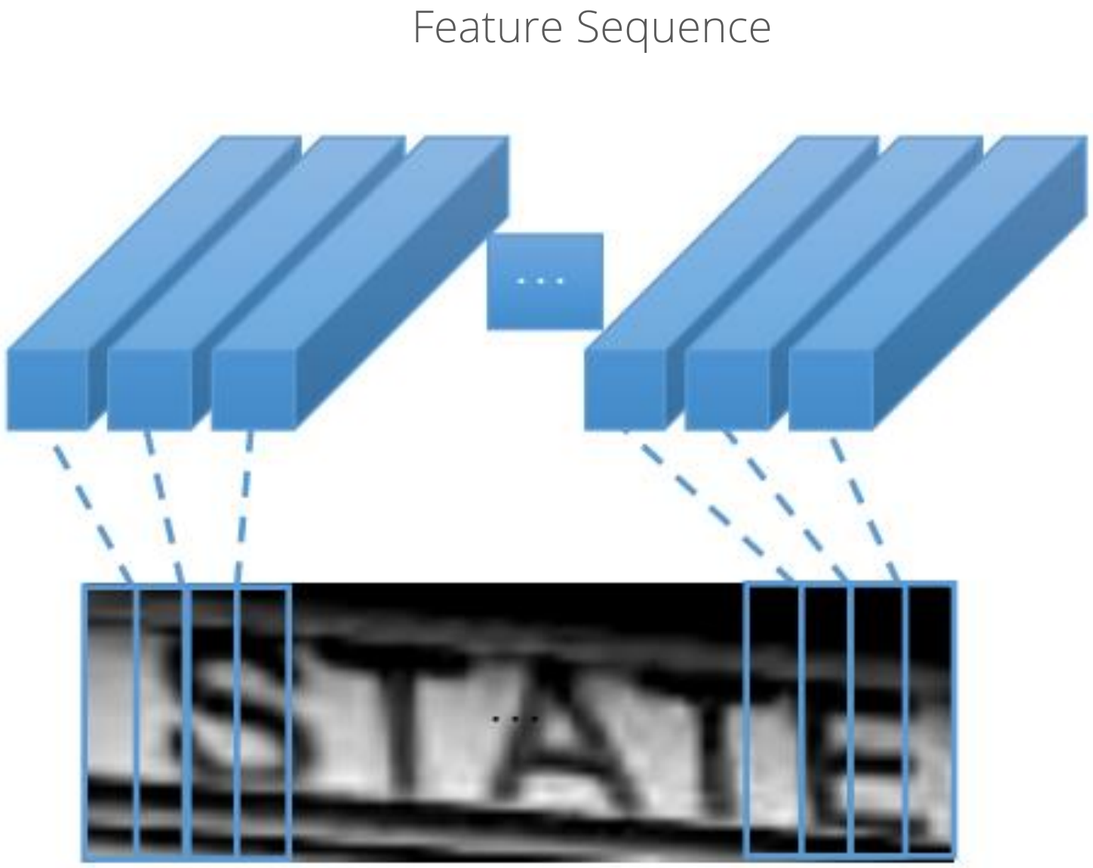
- Is the combination of DCNN and RNN
- Possesses properties of CNN and RNN
- Remains unconstrained to the lengths of sequence-like objects

CRNN Architecture



- Convolutional layers extract feature maps
- Convert feature maps into feature sequences
- Sequence labeling with LSTM
- Translate labels to text

Sequence Modeling with CRNN



Build a speech-to-text Model with TensorFlow Dataset

Problem Statement: Independent makers and entrepreneurs find it hard to build a simple speech detector using open data and code. Many voice recognition datasets require preprocessing before a neural network model can be built on them. Help entrepreneurs/makers by building a speech-to-text model with the assistance of the recently released TensorFlow speech commands dataset.

Objective: Use the Speech Commands Dataset to build an algorithm that understands simple spoken commands.

Access: Click on the Labs tab on the left side panel of the LMS. Copy or note the username and password that is generated. Click on the Launch Lab button. On the page that appears enter the username and password in the respective fields and click Login.

ASSISTED PRACTICE

Key Takeaways

- MFCC Process uses human auditory perception system
- CNN encodes the information from the image and sends that data to the RNN-based decoder
- An algorithm that understands simple spoken commands.

DATA AND
ARTIFICIAL INTELLIGENCE

Knowledge Check

Knowledge Check

1

Why is Fourier Transform used?

a. To convert analog signal to digital signal

b. To convert a signal from time domain to frequency domain

c. To convert a signal from frequency domain to time domain

d. To extract information out of signal

Knowledge Check

1

Why is Fourier Transform used?

a. To convert analog signal to digital signal

b. To convert a signal from time domain to frequency domain

c. To convert a signal from frequency domain to time domain

d. To extract information out of signal

The correct answer is **b**

Fourier Transform is used to convert a signal from time domain to frequency domain.

Knowledge Check

2

The CRNN model _____.

a. Reflects the functionality of CNN while rejecting the functionality of RNN

b. Reflects the functionality of RNN while rejecting the functionality of CNN

c. Reflects the functionalities of both CNN and RNN

d. All the above

Knowledge Check

2

The CRNN model _____.

a. Reflects the functionality of CNN while rejecting the functionality of RNN

b. Reflects the functionality of RNN while rejecting the functionality of CNN

c. Reflects the functionalities of both CNN and RNN

d. All the above

The correct answer is **c**

The CRNN model was modeled to give combined benefits of CNN and RNN models. Therefore, it reflects the functionalities of both.

Speech-to-Text Using Google Speech API

Problem Scenario: Nowadays, writing complex codes for getting text out of speech is a problem due to the complications and cost. To avoid these challenges, Google has created a great speech recognition API that converts spoken text (microphone) into written text (Python Strings). You can simply speak into a microphone and Google API will translate this into the written text.

Objective: To write a program to record audio from your microphone, send it to the Google's speech API, and return a Python String.

Access: Click the Practice Labs tab on the left panel. Now, click on the START LAB button and wait while the lab prepares itself. Then, click on the LAUNCH LAB button. A full-fledged jupyter lab opens, which you can use for your hands-on practice and projects.

LESSON-END PROJECT