Myanmar Speech Recognition

Than Lwin Aung

Speech Recognition

Speech Recognition is actually one of the earliest attempts of machine automation even before the advent of Computer Systems.

Although the machine recognition of human speech in any situation is far from being solved, dictation to machines (computers) has been achieved with high accuracy nowadays.

There are two types of Speech Recognition Systems: **Modular Systems** and **End-to-End Systems**.

Traditionally, Speech Recognition Systems have been implemented with Modular Systems.

However, with the widespread of Machine Learning, End-to-End Systems have become more and more popular.

Speech Signals

Human Speech is generally in the form of audio (electric) signals, which are analog (continuous) in nature.

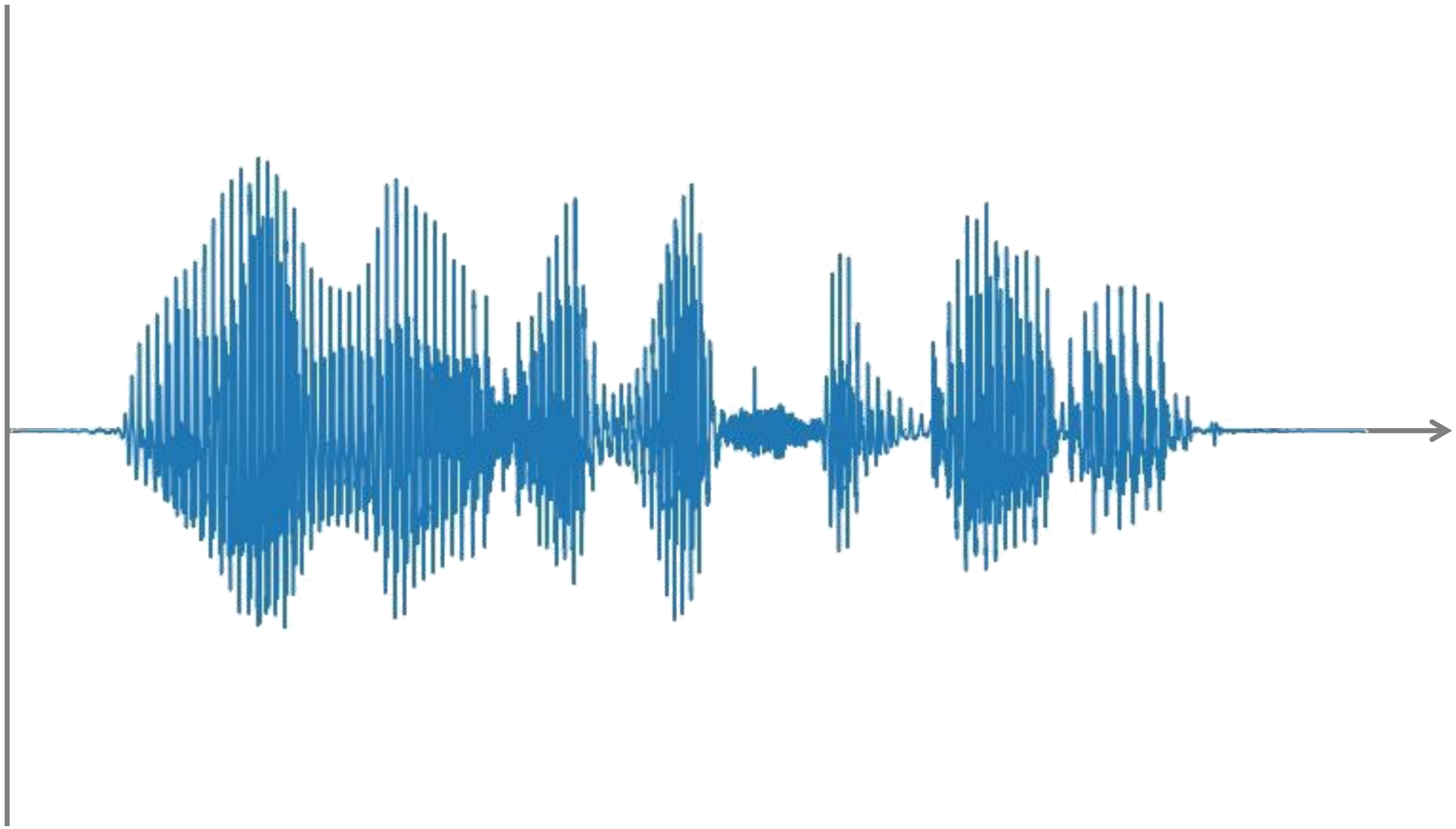
Therefore, it is important to transform Analog Audio Signals to Digital Audio Signals before performing any recognition task.

Analog to Digital Conversion is generally carried out with **PCM** (Pulse Code Modulation), which includes **Sampling** and **Quantization**.

There are different **Sampling Rate** (Nyquist Frequency): 8 KHz, 16 KHz, 44.1 KHz.

Samples are quantized with either Integers or Floats (8 bit, 16 bit) .

16 Bit PCM Speech Signals



16 Bit PCM Samples

Sampling Rate: 44100 Hz

-23	**0**	**12**	**34**	**15**	**0**	**-13**	**-23**	**-15**	**0**

Spectrogram

Generally, Speech Signals are not directly used for Speech Recognition.

Spectrograms (Audio Features) are normally extracted from Speech Signals before performing any Machine Learning Tasks.

Extracting Spectrograms from Speech Signals involves:

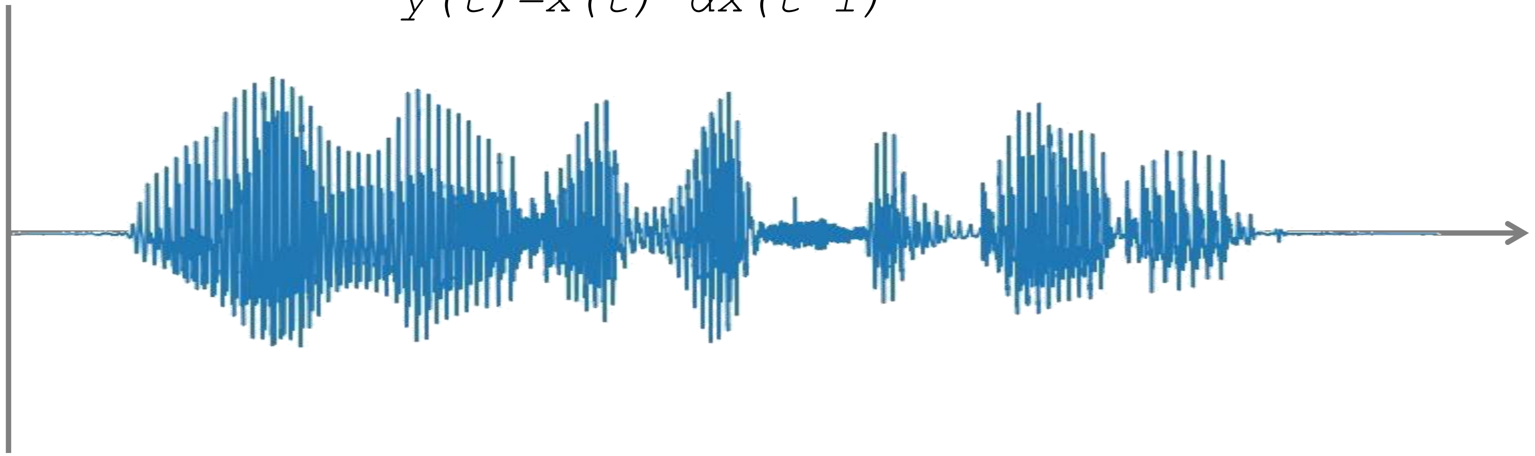
- Pre-Emphasis Filtering
- Framing
- Windowing
- Fourier Transformation
- Generating Filter Banks with Mel Power Spectrum

Pre-Emphasis Filtering

In Speech Signals, higher frequency components usually have lower magnitude than lower frequency components, Pre-Emphasis Filter is to balance the overall frequency components in Speech Signals.

Pre-Emphasis Filter is a Time Domain Filter.

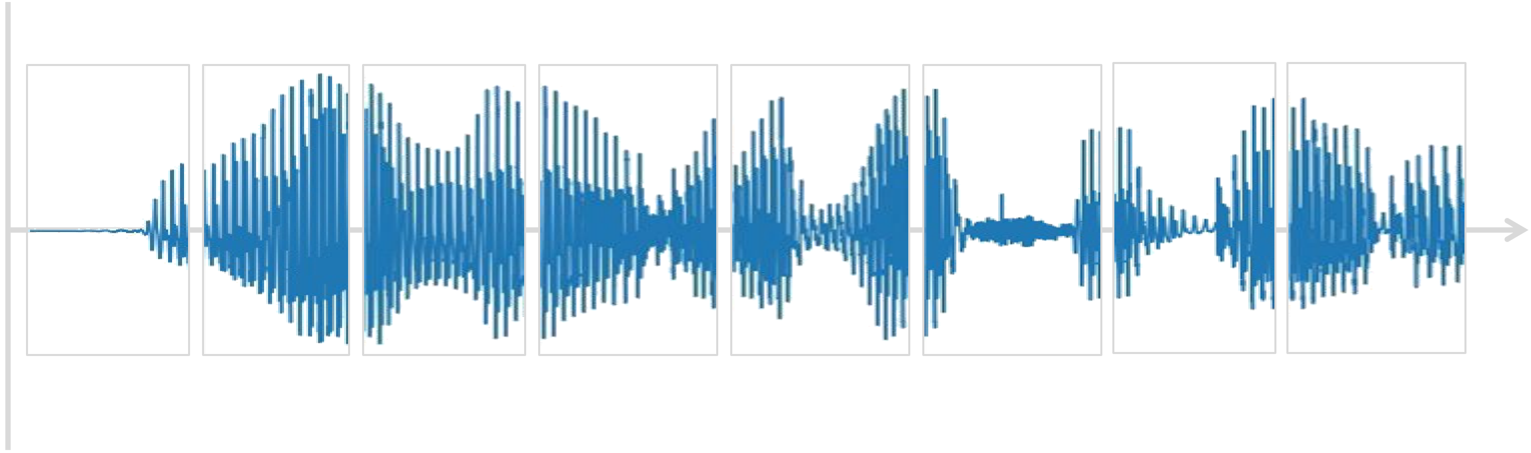
$$y(t) = x(t) - \alpha x(t-1)$$



Framing

Speech Signals is split into Short-Time Frames, 20 ms (Milliseconds). It is assumed that Frequency Components are relatively stationary in Short-Time Frame, which can be easily computed for Frequency Contour with Fourier Transform.

Frames are usually split with 20ms to 40ms Time Frame with 50% overlap between consecutive frames.

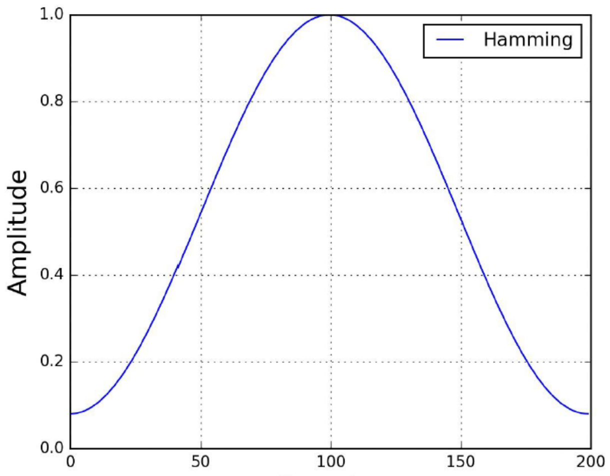


Windowing

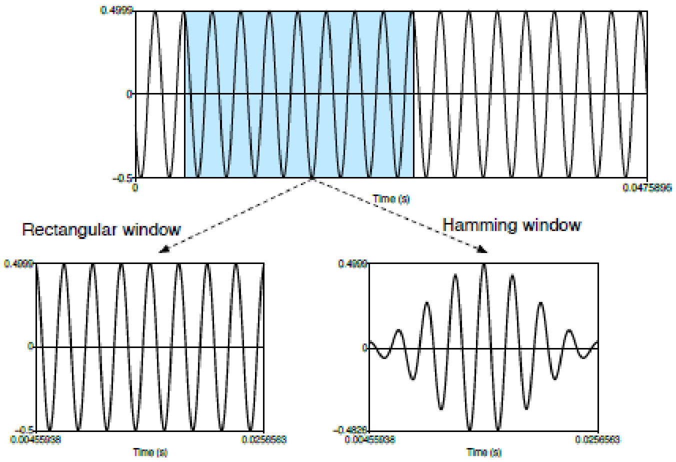
Frames are normally discontinuous at the boundary, which is not desirable.

Therefore, Filtering Window, such as Hamming Window, is applied to each frame to create a continuous consecutive frame at the boundary.

$$w[n]=0.54-0.46cos(2\pi nN-1)$$



Hamming Window



Windowing

Fourier Transform

Fourier Transform is performed on each window Frame. Normally, STFT (Short Time Fourier Transform) is used to transform from Time Domain Signal to Frequency Domain.

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

$$P=|FFT(xi)|2N$$

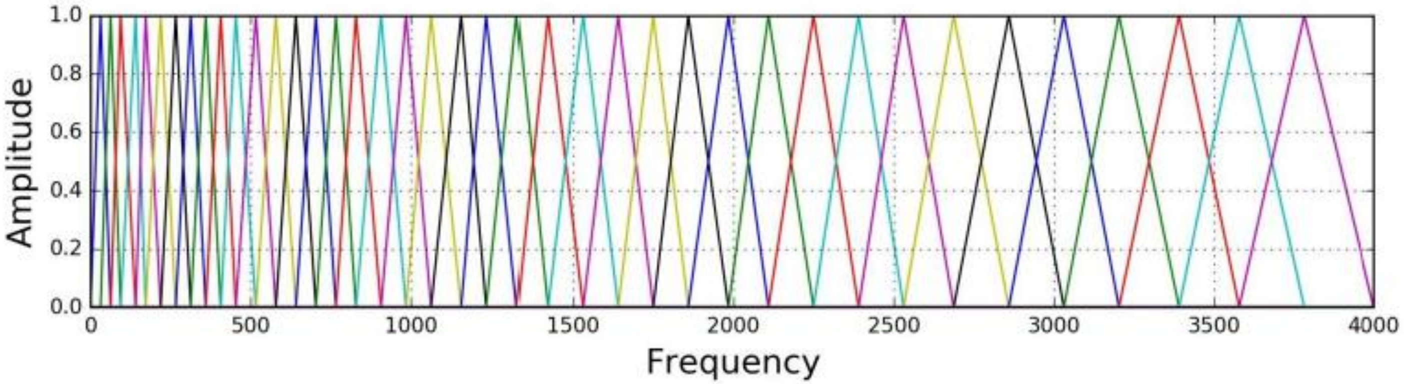
Filter Bank with Mel Spectrum

Frequency (Hz) can be converted into Mel (m). Mel Scale is aimed to mimic the human’s perception of sound.

Each Frequency Components is filtered with a Filter Bank, which contain 40 Filters. Each filter in Filter Bank is a triangular filter with a **center frequency** in the middle and **cut-off frequency** at both ends.

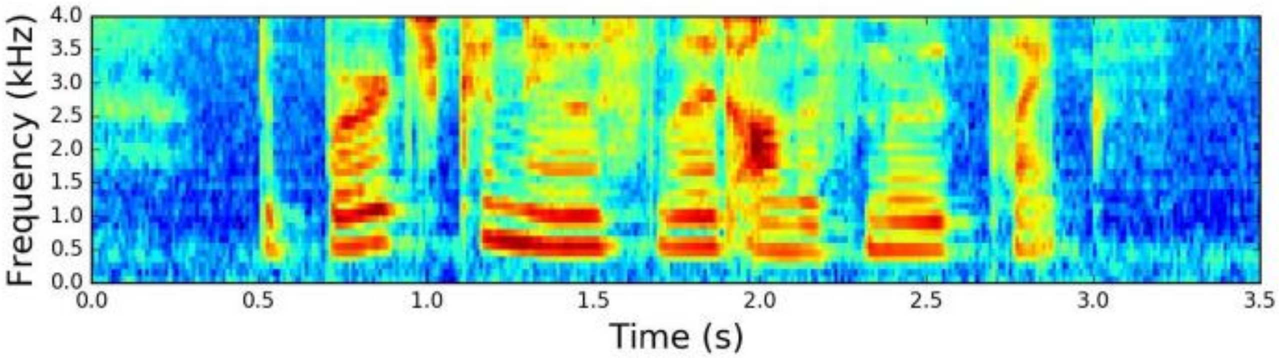
$$m=2595log10(1+f700)$$

$$f=700(10m/2595-1)$$



Speech Spectrogram

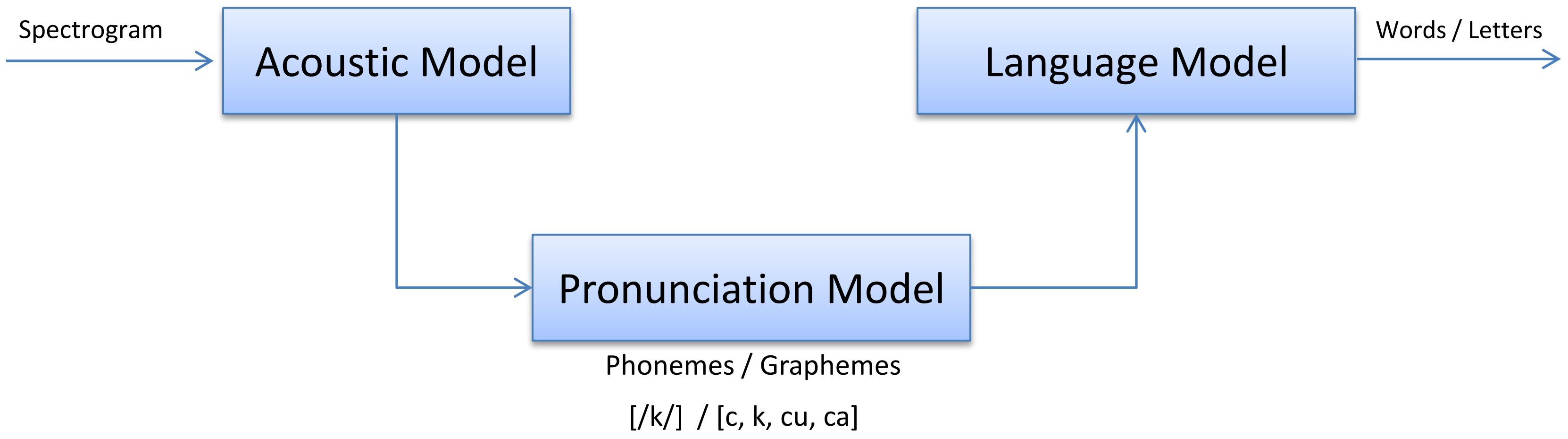
After being filtered by Filter Bank, the Speech Signal is converted into Spectrogram with Mel Scale, which can be used for Speech Recognition or other Machine Learning Tasks.



Modular Systems

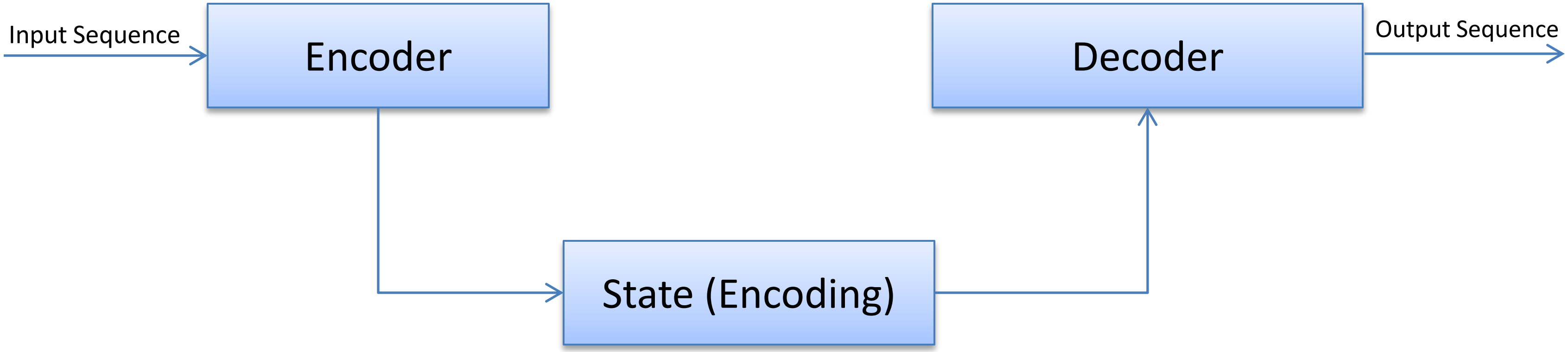
Modular System are mainly composed of 3 Modules:

- **Acoustic Model**
- **Pronunciation (Prosodic) Model**
- **Language Model**



Encoder Decoder Model

Encoder-Decoder Model is a Machine Learning Model, which accepts and encodes a **Sequence of Inputs** into some form of **Encoding** (Encoded States) and decodes those encodings into a **Sequence of Outputs**.



End-to-End Systems

End-to-End Speech Recognition Systems are more or less based on Encoder-Decoder Model. To date, there are 3 primary variants of End-to-End Systems:

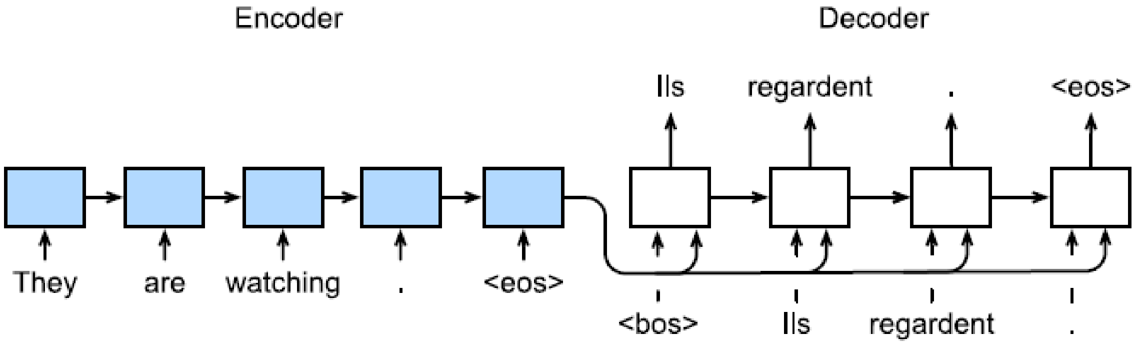
- Attention Encoder-Decoder Variants: LAS (Listen, Attend, Spell), Transformer
- Independent Encoder-Decoder Variants: CTC (Connectionist Temporal Classification)
- Dependent Encoder-Transducer Variants: RNN-T

Sequence-to-Sequence Learning

Sequence-to-Sequence Learning is primarily used to learn the mapping between input sequences and output sequences, such as Language Translation.

Speech Recognition can be thought of as a Sequence-to-Sequence Learning, in which **Audio Input Sequences** are mapped to **Text Output Sequences**.

In essence, Speech Recognition is another form of Machine Translation Task, in which **Speech Utterances** are translated into **Words**.

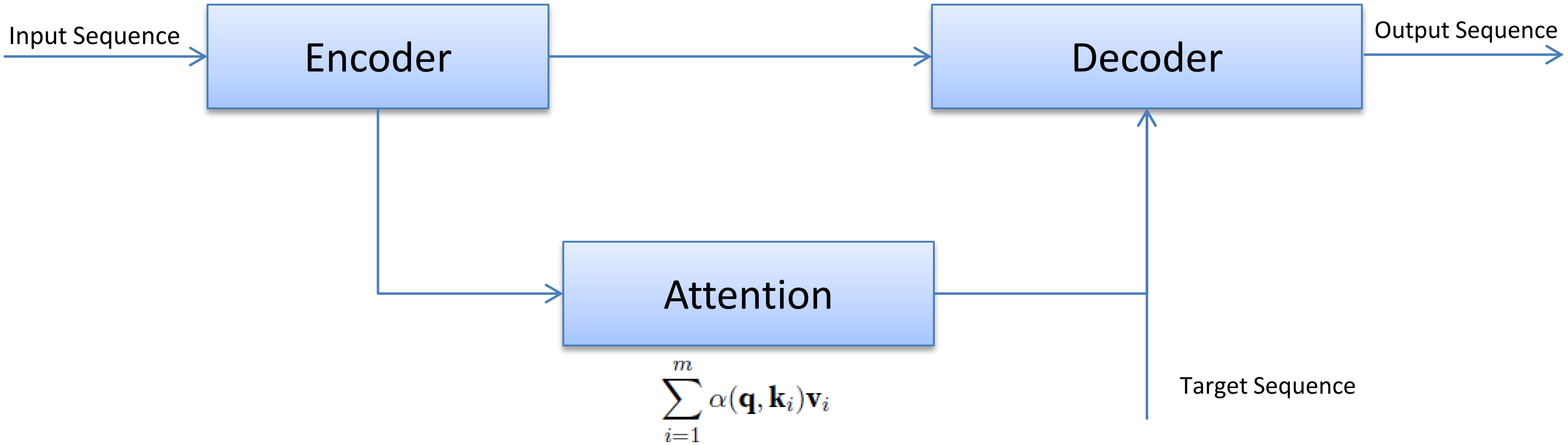


Attention Encoder Decoder

Attention Encoder Decoder is a Sequence-to-Sequence Encoder Decoder Model with **Attention Pooling**.

Attention Pooling is a **Parametric Weighted Average Pooling** of Values, based on Query, Key and Value, in which weighted average is calculated by a Parametric Function of Query and Key.

$$p(y_1,\ldots,y_n) = \prod_{i=1}^{n} p(y_i|y_1,\ldots,y_{i-1},X)$$



$$\sum_{i=1}^{m} \alpha(\mathbf{q},\mathbf{k}_i)\mathbf{v}_i$$

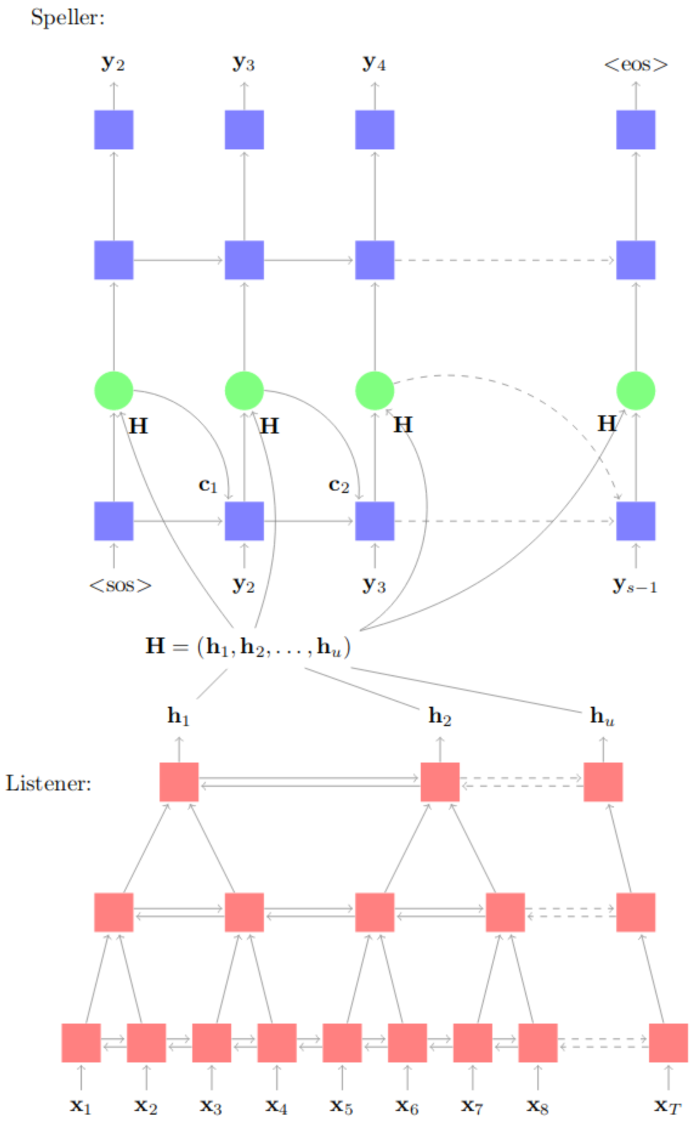
Listen, Attend, Spell

LAS is a Sequence to Sequence Encoder-Decoder with Attention Mechanism.

LAS contains Listener and Speller.

Listener is a Pyramid LSTM Encoder and Speller is an Attention based LSTM Transducer.

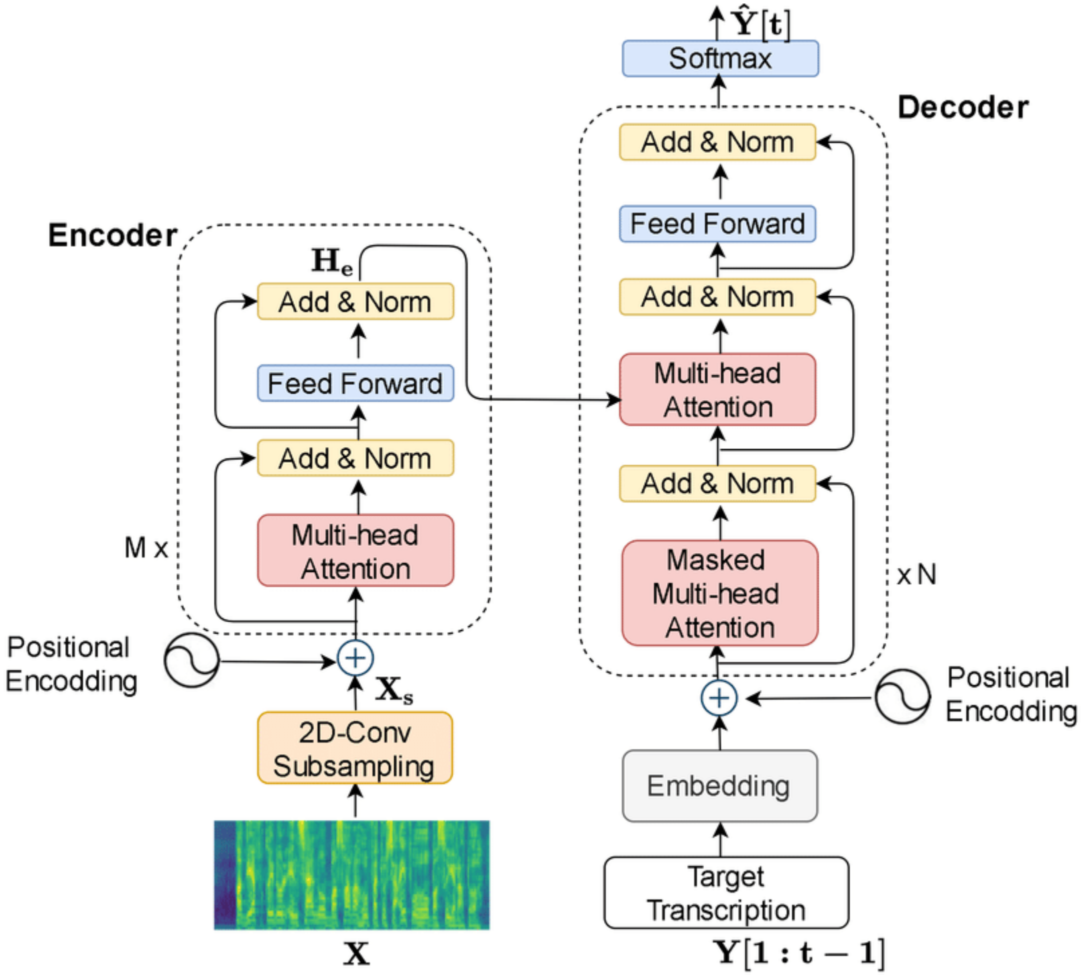
Beam Search is generally used for the inference of output sequences.



Transformer

Transformer has been highly popular with Machine Translation Task. However, it is proved to be useful in Speech Recognition as well.

There are many variants of Speech Transformer.



Independent Encoder Decoder

Unlike Sequence-to-Sequence Encoder Decoder Model, Independent Encoder Decoder is based on the assumption that **Sequences of Input and Output** are **independent**.

Although Sequence-to-Sequence Model has many advantages, there is a major drawback: **Streaming**.

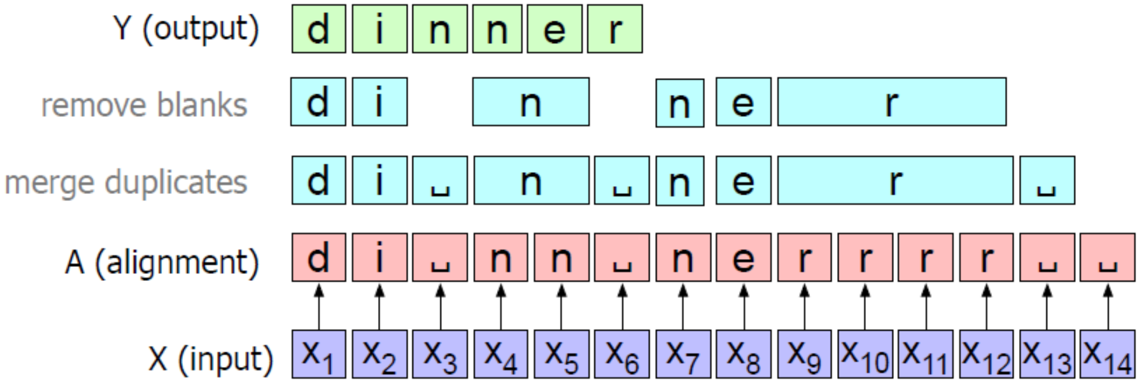
In fact, Independent Encoder is more suitable for Streaming of Sequences.

CTC (Connectionist Temporal Classification)

CTC is one of the Independent Encoding Models, which is based on the assumption that **Sequences of Input and Output** are **independent**.

Input Sequences (Spectrograms) are assumed to be independent, and Output Sequences (Letters / Graphemes) are generated for each Input Frame. However, there is a special **Blank** symbol for Empty Input Frame, which allows encoding to separate Input Sequences.

Repeated Output Sequences are then collapsed into a Single Output Sequence.



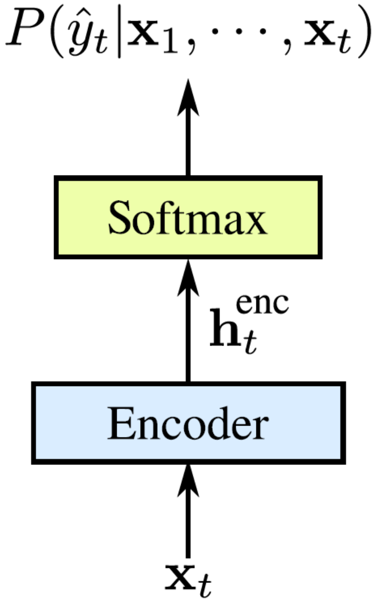
Y (output)
d i n n e r
remove blanks
d i n n e r
merge duplicates
d i ␣ n ␣ n e r ␣
A (alignment)
d i ␣ n n ␣ n e r r r r ␣ ␣
X (input)
x_1 x_2 x_3 x_4 x_5 x_6 x_7 x_8 x_9 x_{10} x_{11} x_{12} x_{13} x_{14}

CTC Model

CTC Model is Machine Learning Model which uses CTC as an alternative to Sequence-to-Sequence Model.

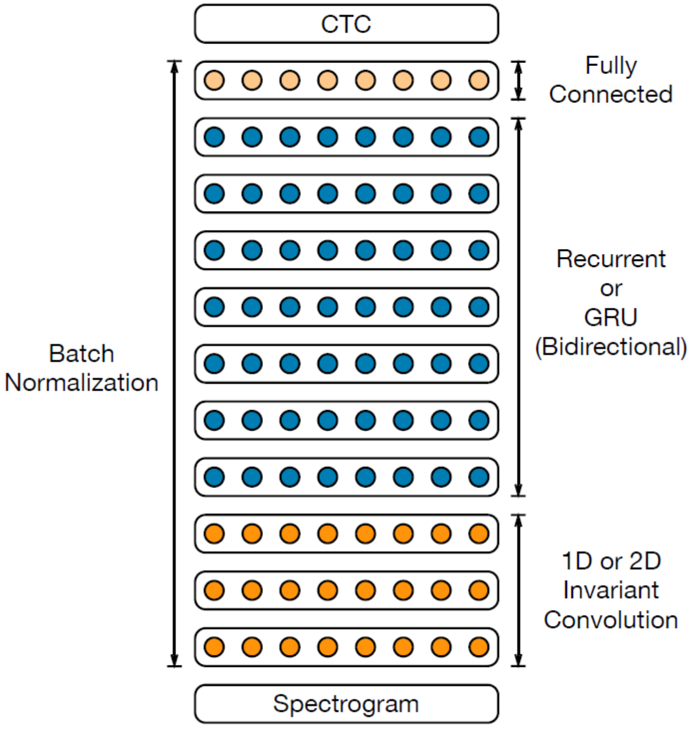
Output Sequences are generally generated with CTC Inference, based on the assumption of Independent Sequence.

Although CTC Model is very convenient for Streaming, it is not as accurate as Sequence-to-Sequence Model, such as LAS or Transformer.

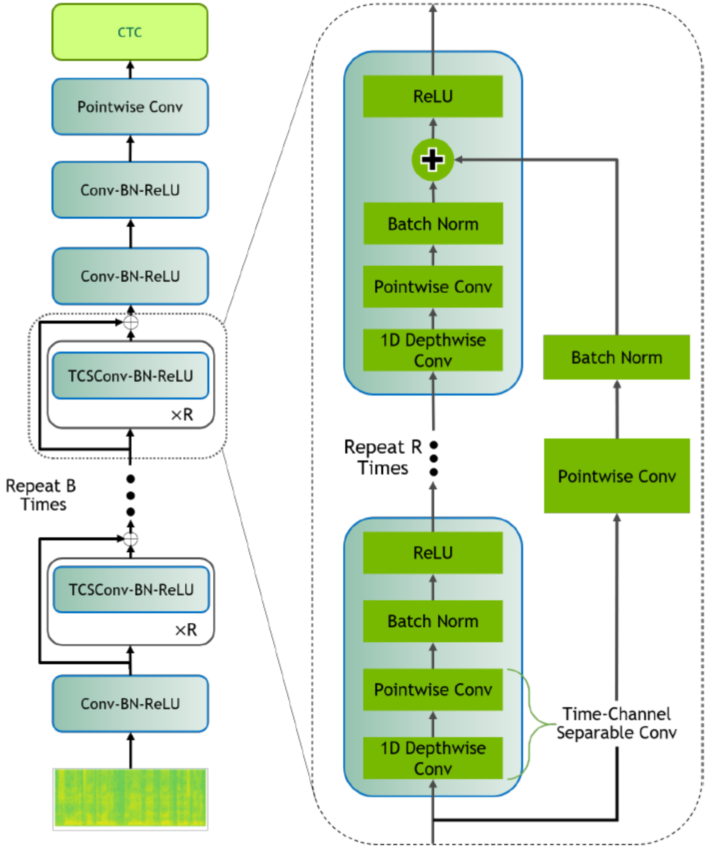

$P(\hat{y}_t|\mathbf{x}_1, \cdots, \mathbf{x}_t)$
Softmax
$\mathbf{h}_t^{\text{enc}}$
Encoder
$\mathbf{x}_t$

CTC Model Variants

Encoder in CTC Model is generally implemented in many different ways with CNN, RNN, or Attention or their combinations.


CTC
Fully Connected
Recurrent or GRU (Bidirectional)
Batch Normalization
1D or 2D Invariant Convolution
Spectrogram

DeepSpeech2 Model



QuartzNet Model

Encoder Transducer

There are, in fact, 2 Architectures: Seq2Seq which has higher accuracy but not convenient for Streaming, and CTC which has lower accuracy but good for Streaming.

However, luckily, it is possible to combine these 2 Architectures, which is Better Streaming Model with Higher Accuracy.

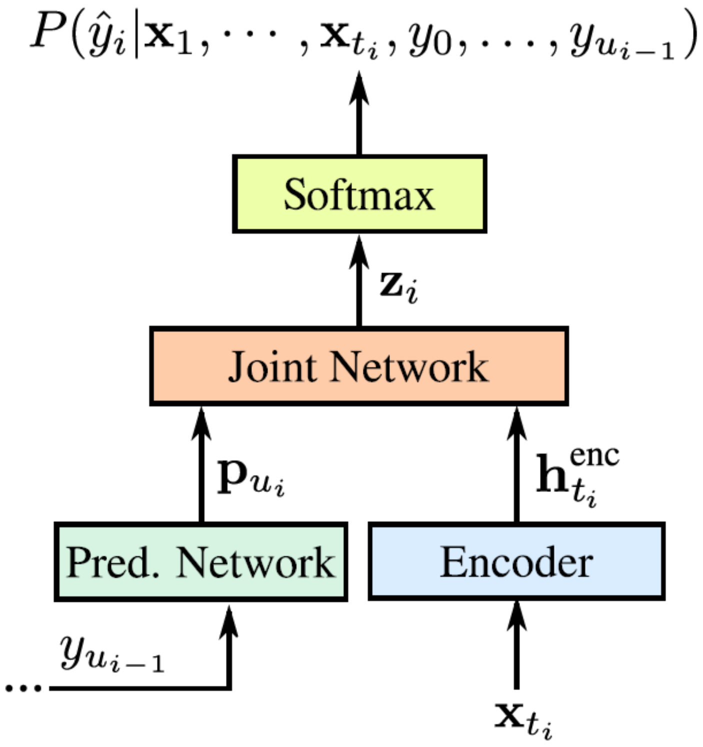
One of the integrated implementations is RNN-T Architecture.

RNN-T Model

RNN (Encoder) - T (Transducer) (RNN-T) is a integration of Sequence-to-Sequence Model and Independent Encoding Model.

Encoder is generally is CTC Encoding Model which is independent in nature and Prediction Network is generally is Sequence-to-Sequence Model.

The output sequence is generated based on the Joint Network of CTC and Seq2Seq Model.


$P(\hat{y}_i|\mathbf{x}_1, \cdots, \mathbf{x}_{t_i}, y_0, \ldots, y_{u_{i-1}})$
Softmax
$\mathbf{z}_i$
Joint Network
$\mathbf{p}_{u_i}$
$\mathbf{h}_{t_i}^{\text{enc}}$
Pred. Network
Encoder
$y_{u_{i-1}}$
...
$\mathbf{x}_{t_i}$

RNN-T Variants

There are many RNN-T variants generally implemented in many different ways with CNN, RNN, Attention, Transformer or their combinations.

Transformer-Transducer is a RNN-T variant, which uses Transformer (instead of RNN Encoder).

Conformer is another RNN-T variant, which uses Conformer (instead of RNN Encoder).

ContextNet is also another RNN-T variant, which uses CNN-LSTM (instead of RNN Encoder).

Design Architectures

Generally speaking, there are many different Machine Learning Architecture which are more advantageous in certain situations.

CNN is generally used for capturing Locality. To capture Global Features, Max Pooling or Average Pooling or Global Pooling is used along with CNN.

RNN is generally used for capturing Global Contextual Dependency.

Attention is generally used for capturing Special Locality in Global Context. Therefore, in most cases, Attention can replace both CNN and RNN. (That’s why it is known “Attention is all you need.”)

MLP is generally used for capturing Generality. Therefore, the more MLPs, the more general the model is.

To date, **CNN**, **RNN**, **Attention** and **MLP** are the most basic building blocks in most Supervised Machine Learning Models.

Research in Progress

It is still an ongoing research on Myanmar Speech Recognition. So far, it is the reviews of available State of the Arts Speech Recognition Systems proposed by Google, IBM, Biadu, Facebook, Amazon and Microsoft.