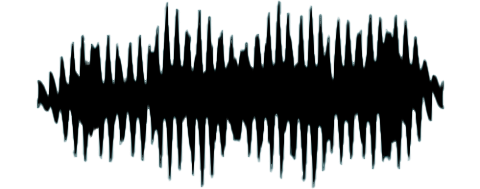
**Acoustic Modeling (Automatic Speech Recognition ASR)**

How does the computer know that the wave shown below translate to “Never Gonna Give You Up” by Rick Ashley? This is done using method called Acoustic Modeling. Acoustic modeling is special method of extracting texts from speech. This works by predicting what the parts of wave translate to and what sound it makes.



Time complexity (Training): O(N) where N is size of training data

Time Complexity (Decoding): O(M) where M is length of speech input

Space Complexity: O(P) where P is size of Model Parameters

**Steps involved in ASR/AM**

1. Data Collection, Labeling

The first step is to collect the data. The data collected are the waveforms of the audio that are to be labeled. The process of labeling is done to tell the computer what a section of a waveform sounds like.

1. Acoustic Model Architecture

Selection of architecture is done to say the computer what method to use when decoding the waveform. The famous architectures are Convolutional Neural Network, Recurrent Neural Network, or hybrid models. In our case, we are making use of LSTM (Long Short Term Memory) under RNN. More about LSTM later.

1. Training

Training process is the most important part of ASR module. In this step, as the name suggests, this process involves making the neural network understand what different waveform translates to.

1. Validation

Now, validation is the manual process where we, programmer, validate if the text recognized by the module corresponds with what was in the speech.

1. Decoding

Once training is complete, it can be now used to decode unseen speeches and audio. It will also work for different linguistic speeches as well. It is done using LSTM or Viterbi Algorithm.

**Viterbi Algorithm**

Viterbi algorithm is dynamic algorithm and is used, in our case, to give possible texts that can be heard in the speech. Working mechanism of Viterbi Algorithm:

1. Problem Formulation

Viterbi helps in problem formulation. Problems in our case using Hidden Markov Model. HMM is the technique that helps us in understanding what the hidden models in speeches are and what observable models are. Hidden models can be phrases, words in the speech and observable models can be tone of the speech and audacity.

1. Initialization

This algorithm has a set of possibilities for each state of the speech. The set of possibilities are stored in table which is also known as “Viterbi Trellis”.

1. Recursion

Viterbi is a recursive algorithm, which means that, with each iteration, it gives us a set of possible phrases that have been said in the part of the audio. An audio is in linear sequence, meaning it has various state. Viterbi goes through each state and checks what phrase can come after the previous state and based on the confidence i.e. number of votes for each phrase, it gives us the result. Recursion uses technique like **LSTM**, to get all possible results.

1. Termination

After recursion and after detecting all the possible phrases, the algorithm then identifies the final state with the highest probability value.

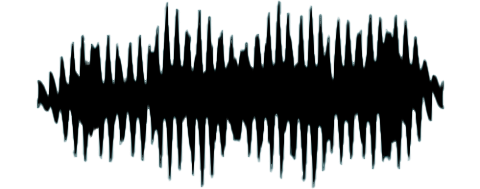
1. Backtracking

At the end, starting from final state, all previous states are backtracked, terminated and text with highest probability for each state is finalized to be the result.

**LSTM (Long Short Term Memory)**

LSTM is the process used in neural network to make the module remember all important stuffs and forget about unnecessary details. In our scase, LSTM helps in getting the appropriate result from the speech.

In our case, lets say we have following wave:



And our Viterbi module was able to get following result as “Nevet Gonna Giv You Upp”, which we can clearly see has a lot of mistake. Now, LSTM is what takes this phrase and checks if there are any words that can be formed using the above phrase. It has a collection of useful data that can easily tell that the phrase should be “Never Gonna Give You Up” and not “Nevet Gonna Giv You Upp”.