**SCT 211-0034/2018**

**PATRICK MAWEU**

**LITERATURE REVIEW ON LIVE TRANSCRIPTION IN VIDEO CONFERENCE APPLICATIONS**

**Introduction**

The production of automated captions for videos with audio tracks, entail the automatic recognition of speech in the audio data, providing automatic subtitles for the audio belonging to a video.

**APPROACHES TO SPEECH RECOGNITION:**

**1.The Acoustic-Phonetic Approach**

The first step in Acoustic Phonetic approach is pre-processing. In this, the features of each and every frame are calculated using Mel Frequency Cepstral Coefficients (MFCC) and Linear Predictive Coding (LPC) measures provide the spectral descriptions of the speech over time.

The next step is feature detection stage, where the spectral measurements are identified for each phonetic unit. There can be features other than spectral measurements, such as, voiced or unvoiced classification, formant location and so on. The third step is, in which the system finds the stable regions, i.e., where the spectral or other features remain approximately constant. These stable regions help in segmentation

The segmented regions are then labelled according to the phonetic units whose features most closely match the features of the region. In the final step, the word or word sequence that the identified phonemes match closely are outputted

**2. The Pattern Recognition Approach**

Use of an n-gram language model, an acoustic model and a dictionary, which specifies the phonetic units in the words present in the dictionary

The major obstacle in using pattern recognition approach is that a detailed analysis of the object or speech has to be carried out to theoretically find the invariant features that would help in an accurate recognition. The problem is that a lot of time and effort need to be invested into mastering these theoretical concepts. Even after that a good recognition rate cannot be guaranteed to any certain degree (Deep Learning solves that problem to a great extent. It employs Convolutional Neural Network (CNN) approach which can automatically learn the invariant features to distinguish and classify the objects)

**3. The Deep Learning Approach**

All current major commercial speech recognition systems are based on deep learning. The reason is that deep learning finally made speech recognition accurate enough to be useful outside of carefully-controlled environments. Raw audio data is converted to spectrograms which capture the nature of the audio as an image by decomposing it into the set of frequencies that are included in it. The spectrograms are then augmented and the data input into the various deep learning models:

* A regular convolutional network consisting of a few Residual CNN layers that process the input spectrogram images and output feature maps of those images.
* A regular recurrent network consisting of a few Bidirectional LSTM layers that process the feature maps as a series of distinct timesteps or ‘frames’ that correspond to our desired sequence of output characters. In other words, it takes the feature maps which are a continuous representation of the audio, and converts them into a discrete representation.
* A linear layer with SoftMax that uses the LSTM outputs to produce character probabilities for each timestep of the output.
* We also have linear layers that sit between the convolution and recurrent networks and help to reshape the outputs of one network to the inputs of the other.
* CTC (Connectionist temporal classification) algorithm takes character probabilities and derive the correct sequence of characters.

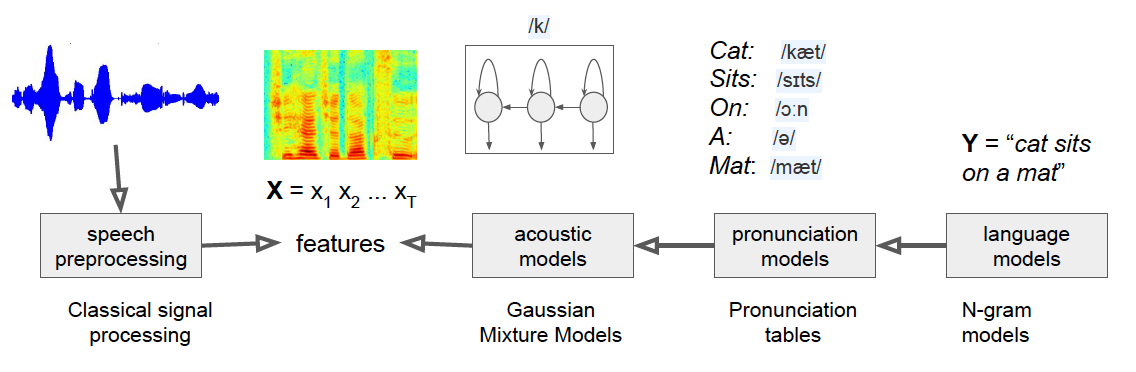
After training thenetwork, we must evaluate how well it performs. A commonly used metric for Speech-to-Text problems is the Word Error Rate (and Character Error Rate). It compares the predicted output and the target transcript, word by word (or character by character) to figure out the number of differences between them.

A Language model is created to capture how words are typically used in a language to construct sentences, paragraphs, and documents.

**A review of the technologies that have been used**

As observed above, the classic way of building a speech recognition system is to build a generative model of language. On the rightmost side, you produce a certain sequence of words from language models. And then for each word, you have a pronunciation model that says how this particular word is spoken. Typically it’s written out as the sequence of phonemes — which are basic units of sound, but for our vocabulary, we’ll just say a sequence of tokens — which represent a cluster of things that have been defined by linguistics experts.

Then, the pronunciation models are fed into an acoustic model, which basically defines how does a given token sounds. These acoustic models are now used to describe the data itself. Here the data would be x, which is the sequence of frames of audio features from x1 to xT. Typically, these features are something that signal processing experts have defined (such as the frequency components of the audio waveforms that are captured).

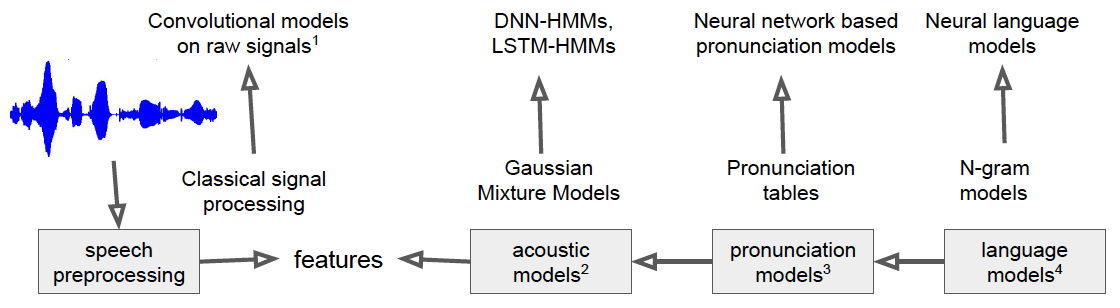


Each of these different components in this pipeline uses a different statistical model:

* In the past, language models were typically N-gram models, which worked very well for simple problems with limited speech input data. They are essentially tables describing the probabilities of token sequences.
* The pronunciation models were simple lookup tables with probabilities associated with pronunciations. These tables would be very large tables of different pronunciations.
* Acoustic models are built using Gaussian Mixture Models with very specific architectures associated with them.
* The speech processing was pre-defined.

Over time, researchers started noticing that each of these components could work more effectively if we used neural networks:

* Instead of the N-gram language models, we can build neural language models and feed them into a speech recognition system to restore things that were produced by a first path speech recognition system.
* Looking into the pronunciation models, we can figure out how to do pronunciation for a new sequence of characters that we’ve never seen before using a neural network.
* For acoustic models, we can build deep neural networks (such as LSTM-based models) to get much better classification accuracy scores of the features for the current frame.
* Interestingly enough, even the speech pre-processing steps were found to be replaceable with convolutional neural networks on raw speech signals.



**LSTM (Long ShortTerm Memory)**

LSTM is a special kind of RNN designed to solve the gradient disappearance and explosion problems when training of long sequences of data. Compared with conventional RNN, LSTM models do produce better generalization and prediction performances in longer sequences. Thus, this type of recurrent network holds great prospects for noise suppression.

**GRU (Gated Recurrent Unit)**

GRUs are improved version of standard recurrent neural networks. GRU has two gates (*reset* and *update* gates) where as an LSTM has three gates (namely *input*, *output* and *forget* gates). A GRU is slightly less complex but is approximately as good as an LSTM performance-wise.

GRUs are comparably easier to train and provides improved training efficiency. GRU, similar to LSTM, also controls the information flow by “gate”, but with one less gate than LSTM and without cell states.

**Related Works**

*Ibrahim, Odiketa & Ibiyemi, 2017*, examined a research on preprocessing of a speech recognition established on human computer interaction system. They discussed the procedures involved in preprocessing of speech recognition data which includes Noise removal, Voice Activity Detection, Pre-emphasis Framing and Windowing.

*Vyas & Suthar, 2017,* provided short reports concerning speech recognition application by incorporating several speech parameters, diverse techniques to speech identification, essentials of acoustic methods, language methods, complex algorithms and feature extraction approaches

*Gawali1, Gaikwad, Yannawar, Mehrotra, 2011*, postulated a Marathi Database and a confined characters recognition method. Mel-frequency Cepstral Coefficient (MFCC) and Distance Time Warping (DTW) were applied for the feature extraction method. Vocabulary comprises of Vowels of Marathi and restrained words that begins from a vowel and simple Marathi sentences. Marathi speech database used in this study was designed by using the computerized Speech lab for the feature mining. The y made used of thirty-five speakers and they were documented and all of the words happened to be recurred thrice. The study also offered the evaluation of recognition accuracy for DTW and MFCC.

**State of the art**

1. **Google Live Captions**

In 2019, Google introduced Live Captions an, Android feature that automatically captions media playing on your phone. The captioning happens in real time

When media is playing, Live Caption can be launched with a single tap from the volume control to display a caption box on the screen, building Live Caption for Accuracy and Efficiency

Live Caption works through a combination of three on-device deep learning models: a recurrent neural network (RNN) sequence transduction model for speech recognition (RNN-T), a text-based recurrent neural network model for unspoken punctuation, and a convolutional neural network (CNN) model for sound events classification. Live Caption integrates the signal from the three models to create a single caption track, where sound event tags, like [APPLAUSE] and [MUSIC], appear without interrupting the flow of speech recognition results. Punctuation symbols are predicted while text is updated in parallel.

**Successes**

* Worked in offline modes
* High accuracy in quiet rooms WER of 5.8% in low background noise areas
* Efficient Neural Network Design in (Used low memory and was time-efiicient)

**Challenges**

* Lower accuracy in noise filled areas WER of 9.8%
* Missing some audio snippets due to challenges in accents

1. **CMU Sphinx**

CMU Sphinx CMU Sphinx is the generic term to describe a group of speech recognition systems developed at Carnegie Mellon University. These include a set of speech recognizers and an acoustic model trainer. Currently, CMU Sphinx has an extensive vocabulary, speaker independent speech recognition code base and its code is available for download and use.

CMU Sphinx uses GMM-HMM model to predict the phonemes in the utterance to determine the word or set of continuous words that were spoken.

CMU sphinx uses an n-gram language model, an acoustic model and a dictionary, which specifies the phonetic units in the words present in the dictionary. The language model can also be a grammar.

1. **Facebook’s High-performance speech recognition**

**Wav2vec** Unsupervised is a way to build speech recognition systems that require no transcribed data at all. It rivals the performance of the best supervised models from only a few years ago, which were trained on nearly 1,000 hours of transcribed speech.

Wav2vec-U learns purely from recorded speech audio and unpaired text, eliminating the need for any transcriptions. This framework takes a novel approach compared with those of previous ASR systems: The method begins with learning the structure of speech from unlabeled audio. Using the facebook wav2vec 2.0 self-supervised model and a simple k-means clustering method, the voice recording is segmented into speech units that loosely correspond to individual sounds

To learn to recognize the words in an audio recording, a generative adversarial network (GAN) is trained consisting of a generator and a discriminator network (which is a neural network itself). The generator takes each audio segment embedded in self-supervised representations and predicts a phoneme corresponding to a sound in language. It is trained by trying to fool the discriminator, which assesses whether the predicted phonemes sequences look realistic. Initially, the transcriptions are very poor, but over time, and with the feedback of the discriminator, they become accurate.

**Successes**

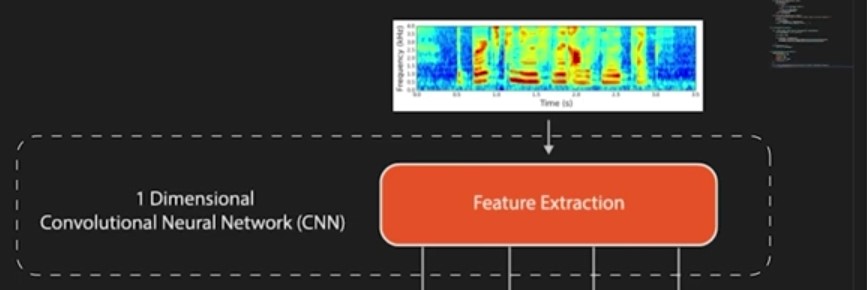
* Managed to achieve the highest word error rate of 11% in comparison to other unsupervised models in the state of the art
* Useful for non-labeled data

**Limitations**

* Unsupervised Learning is Computationally demanding

**METHODOLOGY**

Data Processing pipeline- Augmentation of the data



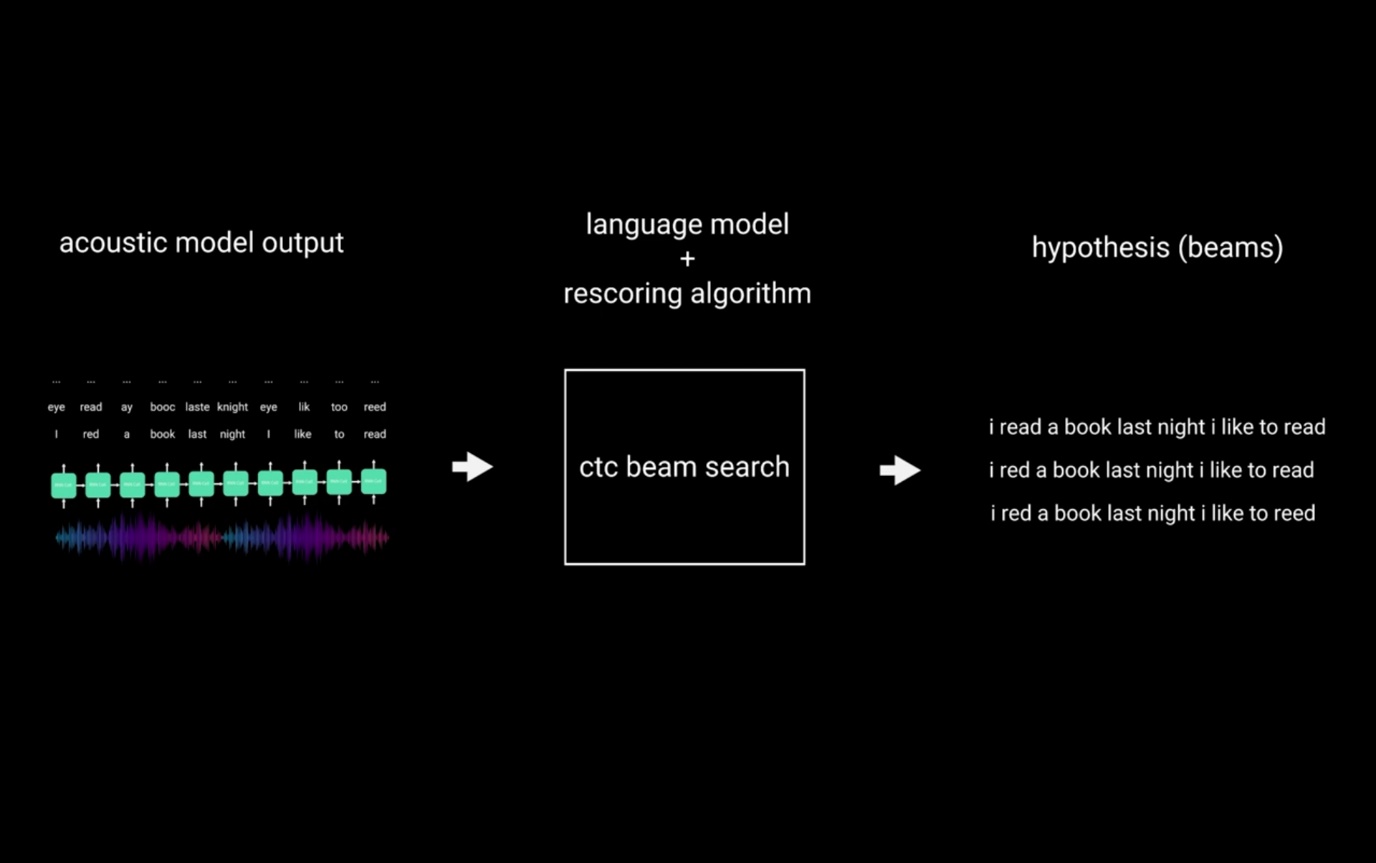
Acoustic Model

RNN CELL

RNN CELL

RNN CELL

RNN CELL

****

**Language Model + a Rescoring Algorithm (CTC Beam Search)**

**Conclusion**

All in all, there has been a lot of research recently that has gone into the field of natural language processing and specifically automatic speech recognition. Strides have been made in ensuring low word error rates and low compute software for NLP.

However, big challenges still remain to be solved. They include Accuracy of Transcriptions, Latency of network in real-time Automatic speech recognition systems and Computational Power.

In my work I will build a deep learning model that will allow us to transcribe English speech with low Word Error Rates and minimal losses due to Noise (This is accomplished by having dual layered recurrent networks built on LSTM and GRU for noise identification and elimination)

**References**

- *Akash, 2021* , “Ok, Google!”— Speech to Text in Python with Deep Learning in 2 minutes

-*Arkady Arkhangorodsky, Christopher Chu, Scot Fang*, 2021“MeetDot: Videoconferencing with Live Translation Captions”

- *Huang et al., 2021 ”*Automatic Simultaneous Translation Challenges, Recent Advances, and Future Directions “

- *Jacopo Giannetti* 2018 “Deep Learning Methods for Speech-to-Text Systems “

*-Bohouta, Gamal & Këpuska, Veton ,2017*. “Comparing Speech Recognition Systems (Microsoft API, Google API And CMU Sphinx)

-*Abhishek Dhankar* , 2019“Study of Deep Learning and CMU Sphinx in Automatic Speech Recognition”

*-Shipeng Han , Zhen Meng , Xingcheng Zhang and Yuepeng Yan,* 2021 “ Hybrid Deep Recurrent Neural Networks for Noise Reduction of MEMS-IMU with Static and Dynamic Conditions”

- *Kyunghyun Cho* et al. 2014 “Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation”

- *Yakubu A. Ibrahim , Juliet C. Odiketa* , 2017 “ pre-processing technique in automatic speech recognition for human computer interaction: an overview”