Department of Biomedical Engineering

19ECT303-Artificial Intelligence and Machine Learning

III Year : V Semester

TITLE: Applications – Speech Recognition.

Speech recognition, or speech-to-text, is **the ability of a machine or program to identify words spoken aloud and convert them into readable text.**

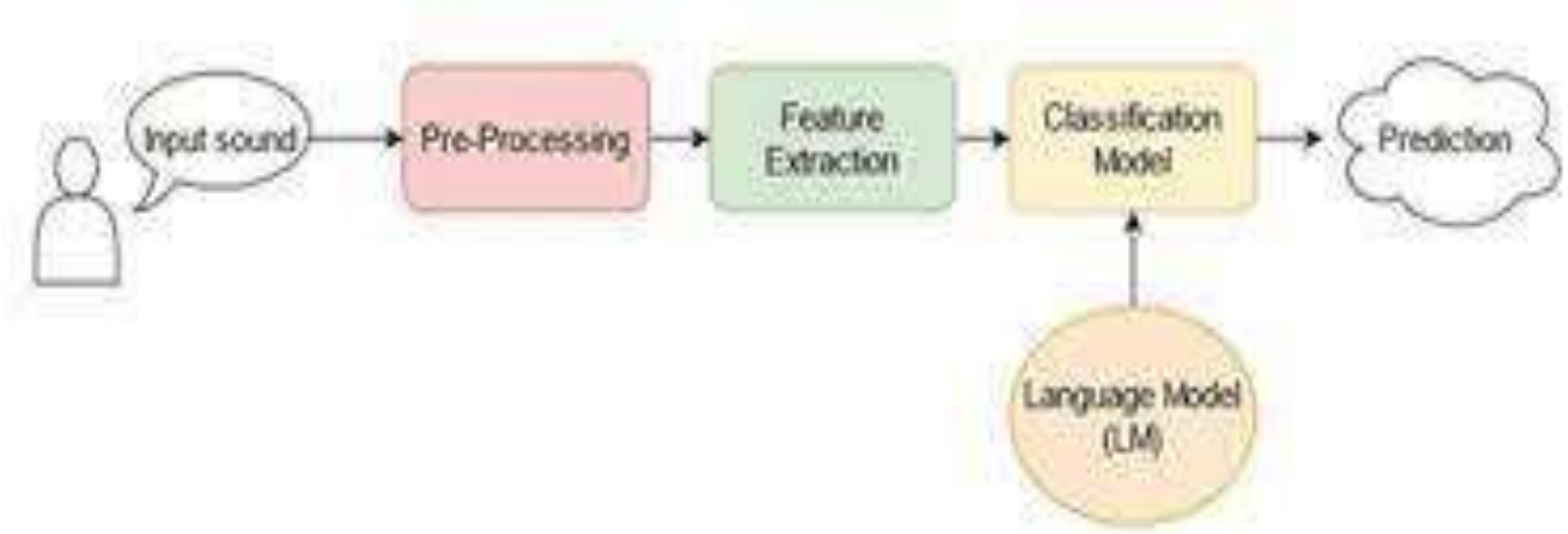
Speech Recognition with Deep Learning

Automatic speech recognition (ASR) refers to the task of recognizing human speech and translating it into text.

- This research field has gained a lot of focus over the last decades. It is an important research area for human-to-machine communication.
- Early methods focused on manual feature extraction and conventional techniques such as Gaussian Mixture Models (GMM), the Dynamic Time Warping (DTW) algorithm and Hidden Markov Models (HMM).
- More recently, neural networks such as recurrent neural networks (RNNs), convolutional neural networks (CNNs) and in the last years Transformers, have been applied on ASR and have achieved great performance.

Speech Recognition with Deep Learning

The overall flow of ASR can be represented as shown below:



Overview of an ASR system

Deep Learning Architecture

There are many variations of deep learning architecture for ASR. Two commonly used approaches are:

- A CNN (Convolutional Neural Network) plus RNN-based (Recurrent Neural Network) architecture that uses the CTC Loss algorithm to demarcate each character of the words in the speech. eg. Baidu's Deep Speech model.

- An RNN-based sequence-to-sequence network that treats each 'slice' of the spectrogram as one element in a sequence eg. Google's Listen Attend Spell (LAS) model.

Deep Learning Architecture

CNN performs better than RNN in Speech Recognition?

At Image Processing and Speech Emotional Recognition(generally Speech Recognition)due to application of filters and MaxPooling which leads to elimination of lighter pixels in Image Processing(compared to darker); and noise and ...(compared to the voice phonemes)in Speech Recognition; accordingly, CNNs work by reducing an image/speech to its key features and using the combined probabilities of the identified features appearing together to determine a classification. While RNNs deals with, only, sequential and there is no elimination and filtering which has made RNN fall back compared to CNN in aforementioned tasks. The link below is a brief description of CNN.

Here are some key aspects of speech recognition applications in deep learning:

- Long Short-Term Memory (LSTM) Networks
- Gated Recurrent Unit (GRU) Networks
- Connectionist Temporal Classification (CTC)
- Transformer-based Models
- End-to-End Models
- Large Datasets and Pre-training
- Online and Continuous Speech Recognition
- Multilingual and Accented Speech Recognition

Vision Title 3

1.Long Short-Term Memory (LSTM) Networks:

LSTMs are a type of RNN designed to address the vanishing gradient problem. They are capable of learning and remembering information over longer sequences, making them more effective for speech recognition tasks where context over time is crucial.

2.Gated Recurrent Unit (GRU) Networks:

GRUs are another type of RNN that, like LSTMs, addresses the vanishing gradient problem. GRUs are simpler than LSTMs and are sometimes preferred in applications where computational resources are limited.

1.Connectionist Temporal Classification (CTC): CTC is a technique often used in deep learning for speech recognition. It allows the model to learn sequences directly without the need for alignment between input and output sequences during training.

2.Transformer-based Models: Transformers, originally designed for natural language processing tasks, have also been successful in speech recognition. They allow for parallelization of training and have proven effective in capturing long-range dependencies in sequential data.

3.End-to-End Models: Traditional speech recognition systems involve multiple stages, such as feature extraction, acoustic modeling, pronunciation modeling, and language modeling. End-to-end models, on the other hand, attempt to directly map input audio signals to transcriptions without explicit intermediate stages. This simplifies the training process and often leads to more accurate models.

1. **Large Datasets and Pre-training:** Deep learning models, especially those with a large number of parameters, benefit from large amounts of labeled data. Pre-training on large datasets for related tasks (transfer learning) has proven effective in improving the performance of speech recognition models.
2. **Online and Continuous Speech Recognition:** Some applications require the ability to recognize speech in real-time or continuously. Deep learning models can be adapted and optimized for such scenarios, providing low-latency and high-throughput solutions.
3. **Multilingual and Accented Speech Recognition:** Deep learning models can be trained to handle various languages and accents, making them versatile for applications in diverse linguistic environments.

Speech recognition has found several applications in the medical field, streamlining various processes and improving overall efficiency. Here are some notable examples:

1. Medical Transcription
2. Clinical Documentation
3. Radiology Reporting
4. Virtual Assistants for Doctors
5. Surgical Documentation
6. Dictation for Pathologists
7. Remote Patient Monitoring
8. Language Processing for Clinical Research

THANKYOU