lissenote

An initiative to make lectures understandable, accessible and easy to remember.

Ministry of Electronics and Information Technology (MEITY)
PS Code: PK858
Problem Statement Title: Automated Notes Maker from Audio Recordings
Team Name: GNAR55
Team Leader Name: Akshay Warrier
Institute Code (AISHE): U-0802
Institute Name: Indian Institute of Information Technology, Kottayam
Theme Name: Smart Automation

Problem Statement

Conversion of voice based recordings of online lectures to PDF/word document.

Objectives

-

Proposed Solution

- Given an **audio file** (.mp3, .wav, .webm, .flac) or **video file**(.mp4), the program will output the notes as a **pdf file** or **docx file** which can be previewed directly on the website or can be downloaded.
- The notes will contain the text, and images which are extracted from the given video file.
- We have given an additional option of using a youtube video link instead of an audio file.
- In this case we directly get the closed captions from the video and write it to the pdf file.

Features

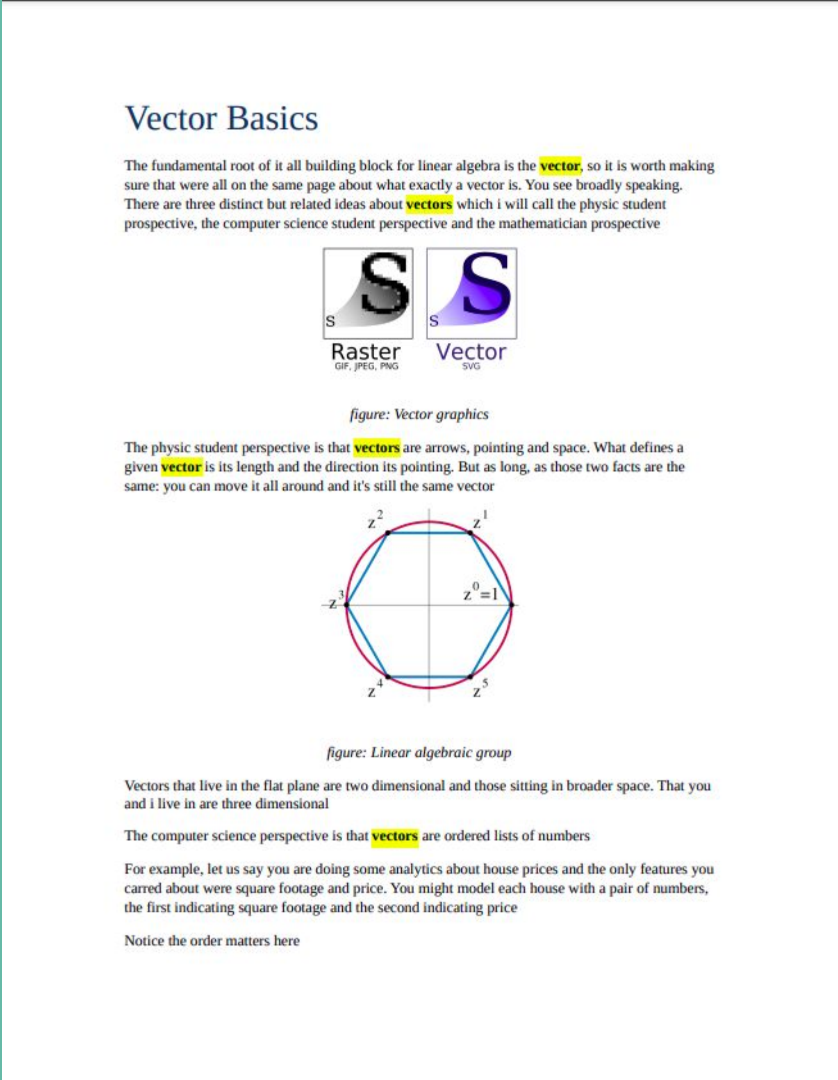
- Accepts inputs of the format audio, video and YouTube URL.
- The notes will contain the heading,images which are extracted from the video.
- We have given an additional option of using a youtube video link instead of an audio file.
- In this case we directly get the closed captions from the video and write it to the pdf file.
- The final notes can be downloaded in multiple languages(hindi, kannada, malayalam, tamil, telugu, bengali, gujarati, odia, punjabi, urdu) apart from english.
- The user can record audio live and get notes of the lecture/meeting.
- The notes contain screenshots of video for better access and understanding of lecture.

Technology Used

- Frontend : HTML/CSS, React.js
- Backend : Flask (Python)
- Speech-to-Text : wav2vec2 specially trained by GNAR55 (free and opensource)

Examples

Vectors | Chapter 1, Essence of linear algebra
3Blue1Brown



Vector Basics

The fundamental root of it all building block for linear algebra is the **vector**, so it is worth making sure that were all on the same page about what exactly a vector is. You see broadly speaking. There are three distinct but related ideas about **vectors** which i will call the physic student prospective, the computer science student perspective and the mathematician prospective

figure: Vector graphics

The physic student perspective is that **vectors** are arrows, pointing and space. What defines a given **vector** is its length and the direction its pointing. But as long, as those two facts are the same: you can move it all around and it's still the same vector

figure: Linear algebraic group

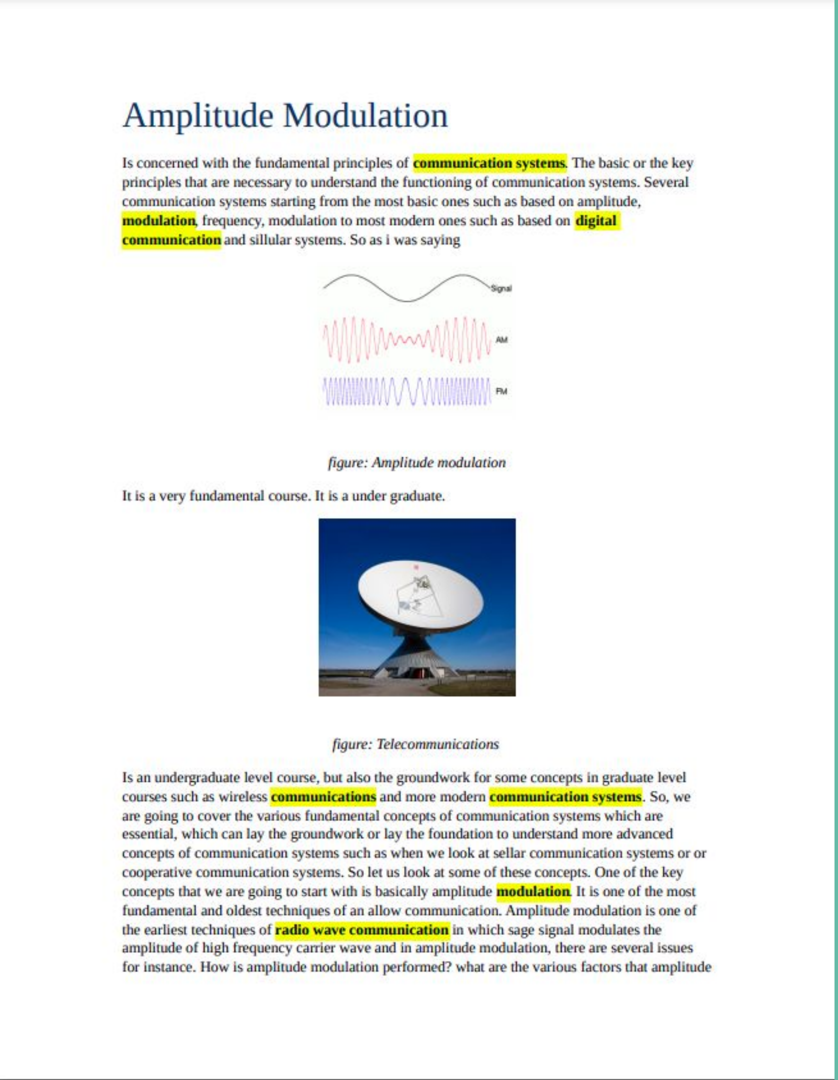
Vectors that live in the flat plane are two dimensional and those sitting in broader space. That you and i live in are three dimensional

The computer science perspective is that **vectors** are ordered lists of numbers

For example, let us say you are doing some analytics about house prices and the only features you carred about were square footage and price. You might model each house with a pair of numbers, the first indicating square footage and the second indicating price

Notice the order matters here

Examples



Amplitude Modulation

Is concerned with the fundamental principles of **communication systems**. The basic or the key principles that are necessary to understand the functioning of communication systems. Several communication systems starting from the most basic ones such as based on amplitude, **modulation**, frequency, modulation to most modern ones such as based on **digital communication** and sillular systems. So as i was saying

figure: Amplitude modulation

It is a very fundamental course. It is a under graduate.

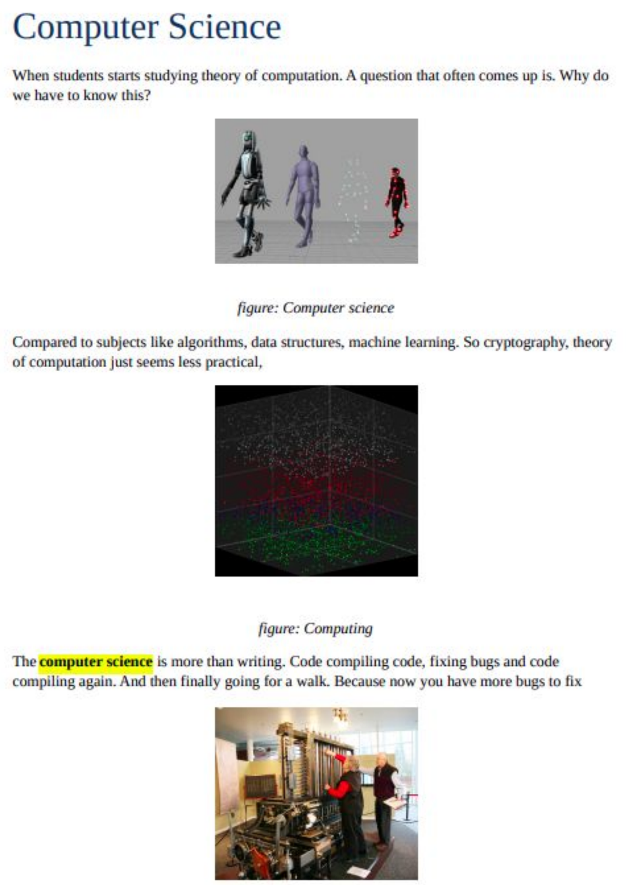
figure: Telecommunications

Is an undergraduate level course, but also the groundwork for some concepts in graduate level courses such as wireless **communications** and more modern **communication systems**. So, we are going to cover the various fundamental concepts of communication systems which are essential, which can lay the groundwork or lay the foundation to understand more advanced concepts of communication systems such as when we look at sellar communication systems or or cooperative communication systems. So let us look at some of these concepts. One of the key concepts that we are going to start with is basically amplitude **modulation**. It is one of the most fundamental and oldest techniques of an allow communication. Amplitude modulation is one of the earliest techniques of **radio wave communication** in which sage signal modulates the amplitude of high frequency carrier wave and in amplitude modulation, there are several issues for instance. How is amplitude modulation performed? what are the various factors that amplitude

Principles of Communication Systems -I - Introduction - Prof. Aditya K. Jagannatham Principles of Communication Systems-I

Examples

Why study theory of computation?
lydia



Computer Science

When students starts studying theory of computation. A question that often comes up is. Why do we have to know this?

figure: Computer science

Compared to subjects like algorithms, data structures, machine learning. So cryptography, theory of computation just seems less practical,

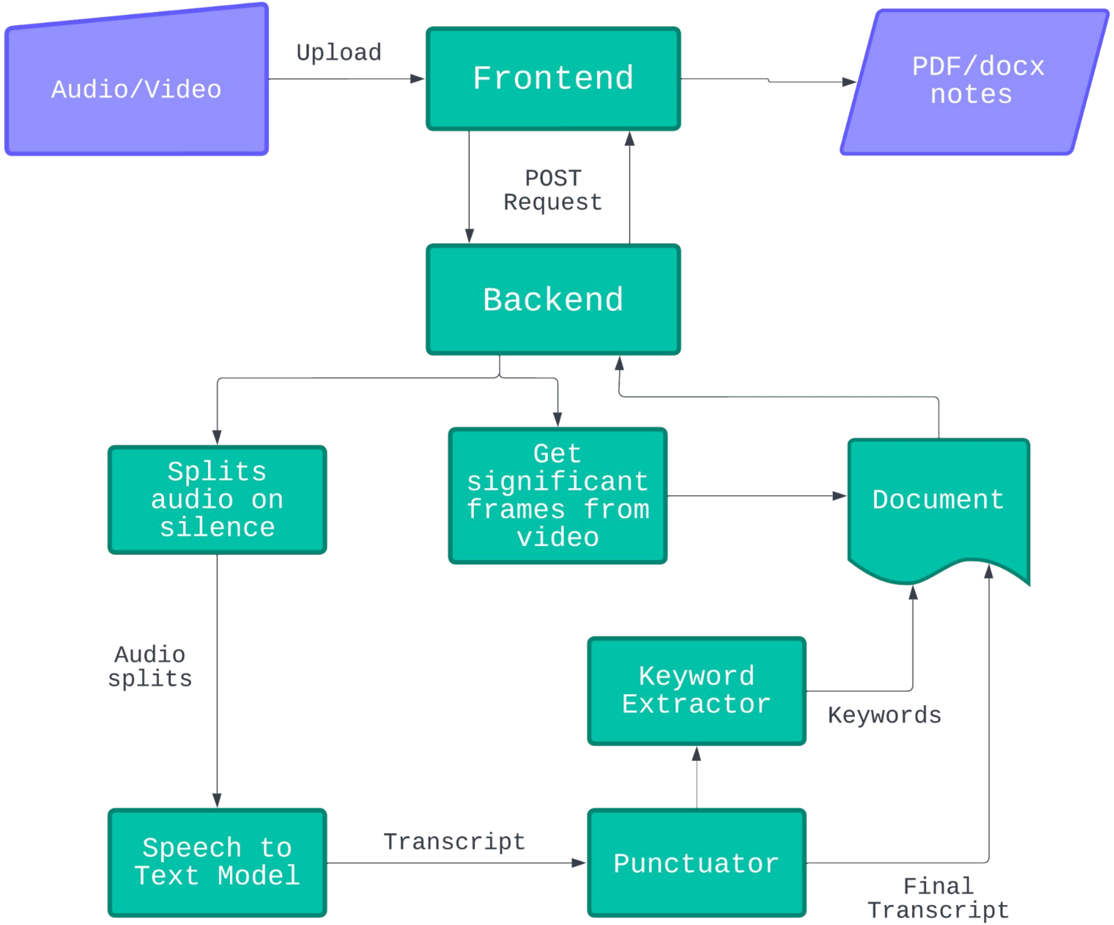
figure: Computing

The **computer science** is more than writing. Code compiling code, fixing bugs and code compiling again. And then finally going for a walk. Because now you have more bugs to fix

Use Cases

- Students can use lissenote to quickly revise their lectures and stay up-to date.
- Teachers and professors can use lissenote as a teaching tool.
- Students with poor internet connectivity who can’t join online classes can use lissenote.
- lissenote has made learning more accessible for deaf students.
- To get a quick rundown of professional meetings, lissenote can be used.
- Students have different paces of learning, and lissenote tries to bridge that gap.
- Teachers can take repeated remedial class using lissenote.

How It Works


Audio/Video
Upload
Frontend
PDF/docx notes
POST Request
Backend
Splits audio on silence
Get significant frames from video
Document
Audio splits
Keyword Extractor
Keywords
Speech to Text Model
Transcript
Punctuator
Final Transcript

How It Works

- When an audio file is uploaded, we send a POST request to one of the endpoints of the backend.
- The backend splits the audio file into smaller audio files. This is done using a python module which detects for silent moments in the audio file. This makes it easier for the speech to text model to transcribe the audio and also gives the different paragraphs in the notes.
- We are using **wav2vec2** as our STT model which was trained on the Common Voice dataset making the model very flexible for different kinds of accents. On top of this we finetuned the model on 8-10 hours of NPTEL lectures.
- After the different splits of the audio are transcribed to text, we use a machine learning model to insert punctuation in the text.

How It Works

- The keyword with the most weight is taken as the title of the notes. We use the Wikipedia-API to get links to the keywords and scrape the images from those wikipedia pages.
- Finally after we have all the content ready for the notes, they are written to a docx file and formatted neatly using the python docx module.
- The docx file is converted to a pdf, sent to the frontend as binary data and rendered in the frontend.

Suggestions and changes

1. Option to upload video files
   - Now drag and drop works with video formats as well
2. Extract diagrams drawn by lecturer
   - Frame by frame analysis
3. Option to work with live recordings
   - Live recording implemented
4. Option to translate notes to other languages
   - Lissenote can now translate to many languages

GNAR55's lissenote trained wav2vec2 speech-to-text model on 8-10 hours of NPTEL lectures

WER

28%

Our Results

Basis video : https://youtu.be/-k970mBzTQE

WER : 26.8%

Cosine Similarity : 68.5%

Jaccard Similarity : 58.5%

Our Competitors

AWS Transcribe

WER : 99.7%

Cosine Similarity : 50.9%

Jaccard Similarity : 38.8%

Limitations and Points of Improvement

- As of now lissenote takes about 4-5 minutes to process a 10 minute audio lecture. This is because the ML models like the STT model are being run on CPUs and they can be boosted majorly with GPUs.
- lissenote is constrained to English input but it can be extended to other languages like Hindi if training data for the same is available.
- The STT model gives a word error rate of 28% on the NPTEL lectures dataset with only a few hours of training.
- lissenote can be hosted and deployed to any cloud platform like AWS Lambda.
- lissenote is a web application now, but it can be converted to an app.