

## **Procedures for collecting and processing the data:**

Here I will highlight some of the procedures and research Methodology that I have followed to collect the spontaneous speech samples, in which I explained and answered two questions, how the data was collected and how the data was processed.

Speech dataset was collected from the native speaker, which addresses the research aims, objectives, meets our needs, and will provide valid result "Analysis of social speaker characteristics in synthetic voices", such data could be found in human-to-human conversations, which could be TV dialog, talk show, and interview for jobs, this kind of speech known as spontaneous conversation, in which speakers are allowed to freely choose their own words, in other words spoken are produced from the mind.

## **Data sources:**

Collecting data is valuable, gives us the opportunity to study and analysis it, moreover, developers can create more personalized and robust technology that helps larger groups of people.

YouTube is most popular video sharing website, which can be used in many fields of search, so I captured the data from YouTube which available online for download.

## **Data Collection Tool:**

To collect the relevant data, there are various methods of collecting primary data, by using online data collection tools, you can easy gather raw data,

Mp3 converter tool one of the most crucial tools for collecting data. we can convert you tube video into mp3 audio. to accomplish this task, we have to do the following steps:

- 1- Determine your video which contains your data.
- 2- Select the video.
- 3- Copy the video address and paste it in mp3 converter
- 4- Select a file type, mp3, mp4, wav.
- 5- Click download Button to complete the process.

### **Primary data processing:**

After collecting the raw data, we need to process and perform a certain operation in order to obtain the speech samples, moreover, to satisfy the requirements like the statistical information and social characteristics of the speaker.

There are many tools that can be used to process the speech dataset. I have used a smart tool called Audacity, which is open source and a cross platform to extract my target data, in this case is guest's speech, I have done some treatments by using a variety of tools for instance, remove the laughs, overlapping speech, and anchor's speech.

To accomplishing the pervious treatments with Audacity, I have implemented some commands such as,

- Selecting the audio segmentation
- Apply some effects
- Rearrange audio
- Export the results

As result, I have collected spontaneous speech from one native English speaker, it is nearly duration of one hour of audio with .mp3 format, vary in length from 5 to 15 second, and the transcription is provided for each speech data.