**SPEECH ENHANCEMENT AND MOOD DETECTION**

Project Synopsis

**NIMISH AGRAWAL, 140905596**

**KRITIK MATHUR, 140905586**

**PROBLEM STATEMENT**

With the advent of virtual assistants in computers, speech recognition has become more and more important. Speech recognition has been around for a lot of years but it still is a field that observes continuous improvements like accent recognition, language identification, real time translation, etc.

One such area where the field is still wanting of quality is speech recognition in noisy environments. Every virtual assistant suffers the problem of being inaccurate in a crowded, or a loud environment which renders them useless.

The goal of this project is to enhance human speech from noisy audio clips and detect the mood of the speaker.

**INTRODUCTION**

With this project, we are trying to alleviate the problem virtual assistants face in a noisy environment by segmenting human speech from other noise. Testing will be done against audio clips from crowded environments, like a party hall.

The algorithm will classify audio as *human speech* or *not human-speech*.

This will be followed by analysing the human speech and preparing frequency and amplitude distributions from the audio clip. Such distributions will be used to predict the mood of the speaker.

*“Emotional prosody refers to the melodic and rhythmic components of speech that listeners use to gain insight into a speaker's emotive disposition.*”

Like, for example, studies have shown that sad emotions are produced with a higher pitch, less intensity but more vocal energy (2000 Hz), longer duration with more pauses, and a lower first formant. Such emotional prosody based studies will be the basis for our mood prediction algorithms.

Thus, our project will consist of the following two stages:

1. Speech enhancement stage
2. Mood prediction stage

**DATASET DESCRIPTION**

All experiments will be carried out on audio clippings recorded by us and taken from video games and movies. The clippings contain environments like a mess, a conference hall, hostel corridor, etc. Audio will be classified as *speech* or *not-speech* only, in the sense the application will not differentiate between different kinds of noises or silence. The format of these clippings will either be *MPEG-2 Audio Layer III* (.mp3) or *Wave* (.wav).

This data will be the input for the first stage – the speech enhancement stage – of our project. The output generated by this stage will serve as the input for the next stage of our project, i.e. the mood prediction stage.

As suggested by the earlier example, for mood prediction algorithms, we have collected some data based on emotional prosody centred studies which predicts a person’s mood based on the attributes of the speech like frequency, amplitude, variation, peak duration, etc.

Thus, for a given audio sample, two outputs will be produced, one will be the enhanced speech clip that is free of noise, and the other will be the mood of the speaker as represented in the speech.