I will start by providing a brief overview of the background information and the motivation behind our project.

The main objective of our project is to develop a web-based application tool that enables the annotation and visualization of emotions in human speech signals. In our daily lives, we are surrounded by contactless devices that incorporate speech technology. For ex, voice assistants like Siri, Google Assistant, and Alexa on your phone is an example of speech tech. The demand for contactless interaction between humans and computers is rapidly growing worldwide, and speech is a prevalent means of facilitating this interaction.

Researchers have been dedicated to developing Speech Emotion Recognition (SER) systems, which aim to recognize emotions in human speech. The goal is to enable computer systems to respond to users more effectively and empathetically.

SER has diverse applications. For ex In medical diagnosis, speech emotion analysis aids in identifying conditions like dementia and depression. Patient care can improve by understanding emotions and adapting responses to their needs.

However, the development of SER faces challenges, particularly in the availability of speech emotion databases. Creating large-scale databases using existing annotation tools is a challenging task. To address this issue, our project focuses on developing a user-friendly web-based application accessible to everyone.

Next, I will present our research findings from the literature review that are relevant to our project. I will discuss categorical and dimensional emotion theories and the evaluation of existing annotation tools. Here are ones that influenced our tool development the most. I will provide more details on this in following slides.

In categorical emotion model, Ekman's Six Basic Emotions and Plutchik's Wheel of Emotions are the two common frameworks used for categorizing human emotions. these categorical models have limitations in capturing all emotions within defined categories. Also Emotions are subjective and often challenging to assign to a specific category.

To address these limitations, dimensional emotion models are employed. One widely used dimensional model is Russell's circumplex model, which has two dimensions: Arousal and Valence. Arousal represents how calm or excited emotion is. while Valence represents how sad or happy an emotion is. Another three-dimensional model, Mehrabian's model, includes an additional dimension called Dominance. However, after reviewing relevant papers, we found that the third dimension did not contribute much2 to emotional features and added complexity during annotation. Hence, for our project, we decided to focus on the two-dimensional model.

After reviewing existing annotation tools for speech emotion, we discovered that most tools used a dimensional model and showed a preference for web-based applications over desktop installations. The choice between one or two-dimensional emotion models varied depending on the researcher's preference or research intentions. While all tools provided annotation features, Darma stood out by offering visualisation of speech emotion data on a 2D plot, also providing with additional graphical evaluation capabilities of multiple data.

EmotionGUI is a desktop application developed by another project group last year at uoa. We have taken over this ongoing project and aim to transform it into a web application. We will discuss the technical details and features of our tool later on in the presentation.

Now, I will pass it onto Enuri, who will discuss our project plan.

I will go through the annotation webpage and explain how the annotation features are laid out. On the left side, we have multimedia players, while the actual speech emotion annotation takes place on the right side.

Multimedia player : users can choose either a video or audio file, and we have a playback controller at the bottom. The audio player consists of two panels. The top panel displays the spectrogram, while the bottom panel shows the waveform.

On the right side, we offer three different models to cater to different user preferences. For the categorical model, users can input an emotion label into a text box after watching the video or listening to the audio. To ensure consistency in vocabulary, we plan to include examples of Ekman's or Plutchik's emotion categories.

The 1-dimensional model has three dimensions representing different emotion features: Arousal, Valence, and Dominance. Each dimension has a slider that users can adjust. We will align the minimum, maximum, and step values with the 2-dimensional model. The provided images serve as a guide, although we plan to replace them with our own facial images. Currently, the buttons to save annotated data are non-functional, but we will implement functionality to save data independently for each slider or collectively for all three sliders.

The 2-dimensional model features an arousal-valence plot. it is intended to record X and Y coordinates on the plot as soon as the video or audio plays. We are working on syncing media players with plotting.

Moving on to the live audio webpage, we have two panels to display the spectrogram and audio waveform. Start and stop buttons for voice recording. users can play back the recorded audio and we are working on displaying spec and audio waveform.

Enuri will continue with the rest of presentation.