I will start by providing a brief background information about Speech emotion recognition.

#1 In our daily lives, we are surrounded by **contactless devices** that incorporate **speech technology**. For ex, **voice assistants like Siri or Google Assistant** on your phone are examples. The **demand** for contactless technology is **rapidly growing**, and **speech** serves as an important **means of facilitating human and computer interaction**.

**Speech emotion recognition** is a research area that aims to detect and classify emotions expressed in speech. Its **applications** can be found in various fields like **healthcare and customer service**. For example, through speech emotion analysis, a medical device can help identify conditions like dementia and depression.

**How speech emotion recognition system works** is that a speech signal goes through **feature extraction** and **selection**. Next, **Machine learning models are trained** on the speech data. This trained model is then used to classify new speech signals into different emotion categories.

#2 However, the development of SER faces **challenges,** particularly due to the **lack of speech-emotion databases** for model training. This issue has a more significant impact on less commonly used languages, such as **te reo Maori**. **Using the existing annotation tool**, creating large-scale databases is a very challenging task. To address this issue, **our project focuses** on developing a user-friendly web-based application for speech emotion annotation and visualisation.

--------------------------------------

#3 Next, I will present our **research findings** from the literature review.

During research, we explored different emotion theories, and the first one to consider is **categorical emotion theory.** The most common examples are **Ekman's model** and **Plutchik's Wheel model**. According to this theory, emotions are made up of a **fixed number of basic emotions**. However, this theory faces **limitations** as emotions are subjective and challenging to assign to a specific category.

#4 To address these limitations, **dimensional emotion models** have been proposed. The most widely used dimensional model is **Russell's circumplex model**, which represents emotions along **continuous dimensions**, such as valence and arousal. **Arousal** indicates how calm or excited an emotion is. while **Valence** shows how sad or happy an emotion is. On this plot, emotions like ‘**surprise’** are positioned on the top right with medium-high valence and high arousal.

There are also **three-dimensional models,** but our research findings suggest that considering three dimensions can introduce more complexity to emotion annotation. Thus, we have chosen to focus on the two-dimensional model for our tool.

#5 These are the **existing tool** that influenced our project. All the tools used a dimensional model for annotation, and the choice between **one or two-dimensional** emotion models depended on the researcher's **preference or research intentions**. Some tools were **only compatible with Windows** operating systems. we also noticed a growing trend **favouring web-based** applications over desktop applications. Also, these tools provided **annotation features**, but **Darma** was the only one offering visualisation of speech emotion data. Furthermore, none of these tools had live audio recording features.

**EmotionGUI** is the desktop application developed by another group last year at uoa. We have taken over this ongoing project and aim to transform it into a web application by incorporating the missing features identified in the existing tools.

Now, I will pass it onto Enuri.

#6 I will go through the **annotation webpage**. It’s got a **user-friendly layout** featuring video and audio players on the left side and annotation interface on the right side. They are laid out alongside each other so that user can annotate emotion easily while they watch video or listen to audio file simultaneously.

The **main feature** here is **providing three different emotion models** for users to choose from. Each model comes with its advantages and limitations, and we have incorporated all of them to **cater to the diverse needs** of our users.

#7 The **video player** we have here is a familiar tool for everyone. Users can choose a video file from their device and we’ve **intuitive playback controller** here, and displays the **status of the video player** at the bottom.

#8 The **audio player** has two panels that offer additional audio features. The top panel displays the **spectrogram**, representing the frequency over time, with different colours indicating the amplitude variations. And the bottom panel shows **waveform**, presenting amplitude change over time. We included these features **for researchers**, in particular, who are familiar with spectrograms and waveforms. They can also gain additional information on audio features while listening to the audio.

#9 This is the annotation user interface designed for users who wants to use a **categorical emotion model** for annotation. So here, After watching video or listening to audio, they can input their emotion into this text box or choose an emotion from the dropdown list. And then this data can be saved. While this model is the most **straightforward** approach, it may not capture all emotion characteristics. so many researchers prefer using dimensional model instead.

#10 This is designed for users who want the **two-dimensional emotion model**, which is a highly preferred choice for annotation. It has a **valence-arousal plotting canvas** where users can move their mouse within the valid region as they watch a video or listen to audio. The x and y data of the mouse position are **automatically** **saved every 20ms.** An **out-of-bound warning message** is displayed here whenever necessary. Users have options to clear or reannotate the data and save the data as CSV file by clicking these buttons. Some researchers have expressed that considering two dimensions simultaneously while listening to media files can be overwhelming. To reduce this **cognitive overload**, they prefer using a one-dimensional model instead.

#11 The last one is the interface for **one-dimensional model** designed with **three independent sliders with facial image aids**, each representing a different emotion dimension: **valence, arousal, and dominance**. Dominance is added for researchers who wish to work with all three dimensions. As users listen to the media file, they can **adjust the slider bars**, and the **positions of the sliders** are automatically saved every 20ms. As users can **only manipulate one slider bar at a time**, it simplifies the annotation process, but if they want to work on two dimensions, they would have to repeat the same process for the second dimension. So the choice between one or two-dimensional models will depend on the researchers' preferences and the specific goals of their studies.

#12 This is the **live audio webpage**. we have start and stop buttons here to control audio recording. **Just like the audio player**, we have pannels to display spectrogam and waveform. users can **visualise audio features in real-time** while recording. once recording is **complete**. they can **re-play** the recording using this audio player. And they can **download** both the audio and image files. This live audio recording feature offers a convenient way for users to create speech data directly that can be used for annotating emotions in the speech or fed into machine model for predicting emotions.