This project explores the possibility of applying deep learning methods combining both image and voice recognition systems to identify a person when one of them is not working well. The results have shown that when both facial features and vocal features are used, accuracy of identifying the person improves significantly. Possible sources of error and inherent challenges of identifying a person are discussed to provide potential directions for further research.

The database is from VGGFace2 and AISHELL-1, which images and voice clips containing 200 people of different genders, races, and ages. With the help of deep learning methods, the significant features of faces and voice clips will be extracted and put into convolutional neural network for training as shown in Figure 3. By combining the confidence value of both voice and facial scores of the person to be identified, a more accurate result will be returned as shown in Figure 4.

As the formula shows under figure 3, the confidence level from the combined model of a person can be calculated as the sum of the confidence level of face over the maximum facial confidence level and the confidence level of the voice over the maximum voice confidence level.

For example, if a person is wearing a surgical mask. It usually covers both the nose and mouth of a person. A correct identification will be more dependent on the forehead and eyes of the person using the face model. However, the coverage on nose and mouth gives obstacles of identifying the person based on the face neural network alone. Hence the voice neural network is introduced to alleviate the potential errors the face model is likely to make while identifying the person, as the combined confidence value reduces the weightage of face model is the face is covered. According to the test data set, by combining both models, the accuracy of identifying a person when face is partially covered will be improved and the limitation that facial recognition is totally ineffective when the face is not detected can be further addressed. In this way, the accuracy of identifying a person can increasing to over 70% even though it is under undesired conditions. When the person is wearing a surgical mask, the accuracy of identifying the person is like to be 75% or above according to the tests done on the models as shown in the results recorded in the table.

Different models are characterized by very different statistical properties. Combined result using different modality is also capable to fill missing modality given the observed ones and show a good shared representations of multiple heterogeneous data modalities with a better prediction of missing input modality. By using combined accuracy, the information given by a more reliable modality can be given a larger weight and counter the results given by less reliable modalities. In the cases where facial recognition may fail, for example, the coverage on mouth gives obstacles of identifying the person based on the face neural network alone. Hence the voice neural network is introduced to alleviate the potential errors the face model is likely to make while identifying the person, as the combined confidence value reduces the weightage of face model is the face is covered, vice versa.

There are a few advantages of this project compared to the current measures. First of all, the accuracy of identification has a more than 10% increase using the new method under optimal conditions. Secondly, the new method of identification of people can have over 70% accuracy when the face is partially covered which traditional methods of identification always fail under such circumstances as shown in table in Figure 10. Last but not least, the new method makes the disguising more difficult as both images and voice clips are required to be successfully identified.

This project can be used widely under many situations such as attendance taking, security check, and transections & payments. With the help of this more advanced algorithm and method, personal privacy, properties, as well as safety can be better protected. The risk of crime and disguise can be highly reduced.

In the future, I would like to a to add noise cancelling features on the voice recognition part which would increase the effectiveness of the model under noisy conditions. Besides noise cancelling, I proposed some other ways for future work and improvement. 1. I would like to use greyscale instead of RBG while training the face model. Because the colour does not really play an important role in identifying a person. The use of greyscale instead of RGB reduces the dimension of the input from 3 layers to 1 layer. This should be able to reduce the complexity of computing significantly. 2. Besides, filter pruning can be used while designing the neural network to reduce the complexity and increase the speed. By removing every unimportant output filter set in every weight tensor. The total amount of hyperparameters and the size of the model needed to compute can be much less. This method is able to reduce the complexity of computing without much loss in the accuracy. 3. I would also like to apply channel pruning in the model. This reduces the number of channels for computation in layers. It allows the float point operations to be fewer while performing forward propagation. This method is then reducing the FLOPs to a larger extent without actually reducing the accuracy. Hence the trade-off between computational complexity and accuracy can be better balanced.

Thank you for spending your valuable time listening to my presentation, feel free to leave any questions.