Age Prediction of Speaker's Voice

Team01

林育頡 *109062232*　林子崵 *109062225*　呂廷洋 *109062229*　劉元愷 *109062303*

I. Abstract

Speech signals are used in many real life applications, like speech emotion detection and HCI. In order for intelligent machine to capture more information about the speaker, it will require some signal processing techniques to analyze the sound wave, which is a tough task when applying traditional frequency analysis. In this study, we proposed a convolutional neural network, which takes a short-time Fourier transform (STFT) of captured sound wave as its inputs, and predict the age of the speaker from it. Since the sound wave of different gender varies, we also proposed **another DNN model** to differentiate them, so that we could further improve the model precision by using two models specialized for male and female age prediction. Our proposed model achieved 92% accuracy score on gender prediction, and 52%, 50% on female and male age prediction respectively.

II. Introduction

This is a project of machine learning from NTHU CS students. Voice is often consider as a method of identifying age. We've saw many machine learning project conducted on identifying people's age via picture of there faces, body parts. However, few were done via voice. Voice in data usually illustrated as waveform - signal in time domain. However, we are more interested in "frequency" domain. Thus, we used Fourier's Transform to obtain frequency distribution at each time point, and after combining frequency distribution at each point we have *Spectrogram*, a frequency distribution - time graph. This method is also called - short-time Fourier transform (STFT). First, we use only average frequency to verify the gender of voices. We found out that the fundamental frequency, which represents the lowest frequency at which a periodic sound appears, is a great identifier of gender. Second, we use the *Spectrogram* as data, building two convolutional neural networks (CNN) for each gender.

III. Methods

We used the *supervised models* such as CNN & NN to try to predict the age and gender of the input voice. That is, we need the dataset that has the voice files with the labels of gender and age. Hence, we used the common voice dataset on the *Kaggle* to train our prediction model.

A. Data collection

As mentioned, we used the **common voice** dataset to train our model, however, it is a huge dataset which contains over 300K data in it. Furthermore, there are three subsets in the dataset, which are *valid*, *invalid* and *other*, which means if the data has been verified by at least 2 people. Hence, we decided to use *valid* subset as our training and validation dataset, which contains about 196K data.

B. Data Preprocessing

For preprocessing, we found out that the **age** and **gender** column in the *.csv* file for labeling are not fully labeled, thus, we first filtered out the completely labeled data and collected them as the final training and validation dataset. Then for the preprocessing of voice files, we first transform the *.mp3* file into wave plot, such as Fig.1

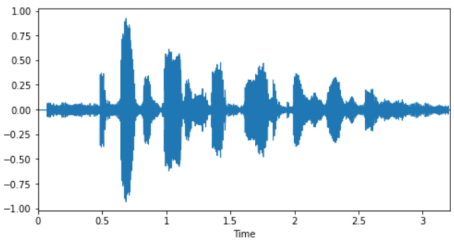

Fig. 1. Original voice wave plot.

1) Cleaning the Wave Plot: After we drew the wave plot of the voice, We observed that there exists some frequency that exists all the time in the plot, which we thought it is the **background noise**, thus, we want to reduce it to get the more pure voice from human. We use a python package called **noisereduce** [1] to reduce these noise that below specific dB, and the wave plot after noise reducing would be like Fig.2. Furthermore, we also found out that in the beginning and at last of the wave plot, there often exists some time at which barely no sound. Therefore, we also trimmed those time stamp to get more explicit human sound. The result is shown as Fig.3.

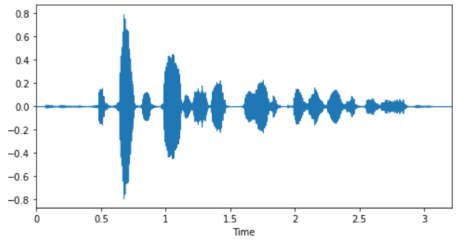

Fig. 2. Wave plot after noise reducing.

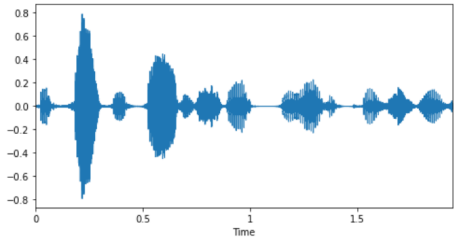

Fig. 3. Wave plot after trimming.

2) Transforming the Wave Plot: Then since we want to have the better insight to the sound, we tried to transform the wave plot into **Spectrogram**, which is graph that can both has *time* and *freqency* domain so that we can better analyze the voice.

First spectrogram we tried was **Short-time Fourier transform** [2], which is actually using **Fast-Fourier transform** for small time window.

After transforming by the **STFT**, the spectrogram would be like Fig.4 below.

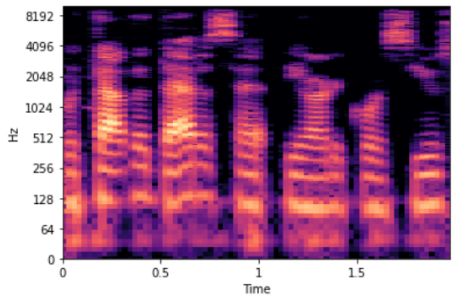

Fig. 4. STFT spectrogram.

Another spectrogram we tried to use was **Mel spectrogram** [5], which is based on the *mel scale* [4]. Mel spectrogram is similiar to the STFT spectrogram but different on the scale of **y-axis**. And it looks like Fig.5.

After testing above two spectrogram results, we decide to choose *STFT spectrogram* as our input to train our model since the freqency of human sound is about 85Hz to 255Hz, and we observed that STFT spectrogram has higher resolution on this span of frequency, hence, we thought it could better fit our goal and have better results.

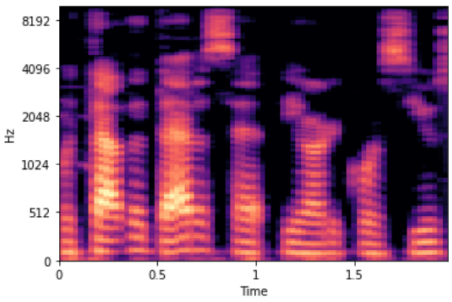

Fig. 5. STFT spectrogram.

C. Gender Model

We thougt the biggest difference of the different gender's voice is the **pitch**, so called the **fundamental frequency** of the sound, which is the lowest frequency in the sound that periodically appear. Therefore, we choose to extract the fundamental frequency using probabilistic [6] and compute the following value of it as features:

1) mean
2) median
3) Standard deviation
4) 5-% percentile value
5) 95-% percentile value

After calculating these values, we use them as input to the **NN** model, whose structure is shown as Fig.6. And the output, we used *one-hot encoded* with two groups which is male and female.

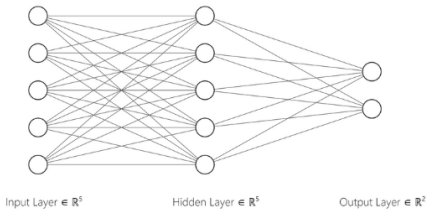

Fig. 6. Gender prediction model.

D. Age Model (Male and Female)

We constructed two *CNN* models and trained them into male-age prediction and female-age prediction model respectively. We use the label file and the pre-processed STFT spectrogram as input to the models, which size is $(216, 334, 3)$. As for the output, we used *one-hot encoded* label and predict the age into four groups, which is

1) 0 to 30 years old.
2) 30 to 50 years old.
3) 50 to 70 years old.
4) older than 70 years old.

The reason we divided it like that was not only for the prediction but also for the data balancing, since we have

counted the number of data in each age, and we wanted to balance before training the model to prevent overfitting to specific age group. If we didn't balance it, the final prediction would always falls in the group that has the most training data.
And Fig.7 is our CNN model's structure.

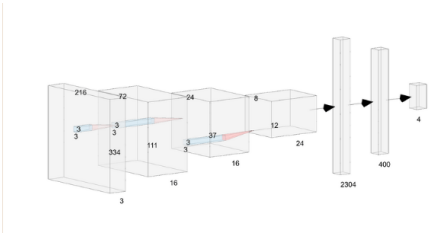

Fig. 7. Age prediction model.

IV. Results

A. Result of Gender Model

We use *binary cross entropy* as our loss function and Fig.8 is the loss curve and accuracy curve repectively.

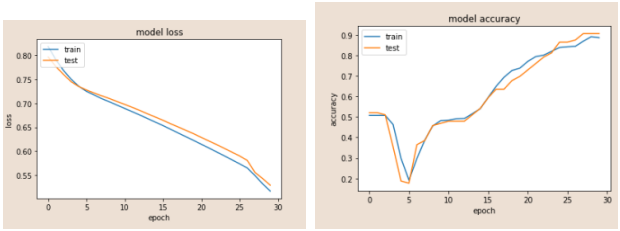

Fig. 8. loss and accuracy curve of gender model.

As we can see, the accuracy achieved is about 90%, which is quite accurate.

B. Result of Age Model

We used *categorical cross entropy* as our loss function, and Fig.9 is our loss curve and accuracy curve.

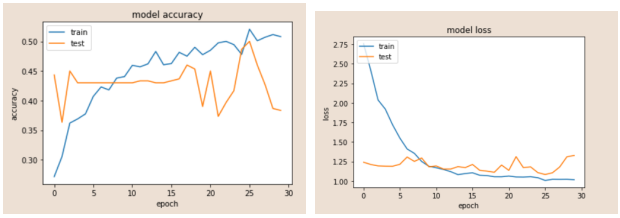

Fig. 9. loss and accuracy curve of gender model.

Here, we use the result of **male** prediction model, whose accuracy is slightly lower than the female model.
We use **categorical cross entropy** as our loss function to train the models, and as for the result, the accuracy can achieve about 50%. Although the result might not be so satisfying , it still twice higher than random guess.

V. Discussion/Conclusion

A. Gender Prediction Accuracy

According to our gender prediction accuracy, we can conclude that the difference on *fundamental frequency* of voice between male and female might be obvious since after we using the fundamental frequency related data into model, the accuracy of age prediction is more satisfying than using image to predict both age and gender.

B. Lower Accuracy on Age prediction

As for age prediction, the results of accuracy of either male or female is not really optimistic. So we think that age and voice might not have much relation to each other. Perhaps, if we tried to use different math method such as original **Fourier transform**, we might get some different results
Nevertheless, by our method, which is **Short-time Fourier transform** ,there seems to still exist some common factors for the prediction accuracy is twice higher than just guess randomly (25%).

C. Different Accuracy between Different Gender

In the process of training age prediction model, we observed that the accuracy of female model is easily to be higher than the accuracy of male model, that is, we didn't need to put as much effort such as *dropout*, *data augmentation* and *batch normalization* to prevent overfitting and make higher accuracy as we do for male model.
For this phenomenon, we could only guess that the voice of female varies more than male between different age.

VI. Author Contribution Statements

林育頡 : Data collection, Data preprocessing, combine the models to do final prediction and construct male age prediction model.
林子嵎 : Environment set up, combine the models to do final prediction and construct male age prediction model.
呂廷洋 : Data collection, Data preprocessing, combine the models to do final prediction and construct male age prediction model.
劉元愷 : Model drop-out mechanism, combine the models to do final prediction and construct male age prediction model.

References

[1] https://pypi.org/project/noisereduce/
[2] https://kevinsprojects.wordpress.com/2014/12/13/short-time-fourier-transform-using-python-and-numpy/
[3] https://en.wikipedia.org/wiki/Short-time_Fourier_transform
[4] https://en.wikipedia.org/wiki/Mel_scale
[5] https://librosa.org/doc/main/generated/librosa.feature.melspectrogram.html
[6] https://librosa.org/doc/main/generated/librosa.pyin.html