Capstone Project

Speech emotion recognition using convolutional neural networks

Author:

Thomas Stoll
April 15, 2023

Table of Contents

1. Project context

What is speech emotion recognition?

Speech emotion recognition is the process of identifying and classifying the emotions conveyed in human speech. Emotions are expressed in speech through a combination of acoustic features such as pitch, loudness, and speech rate, as well as linguistic features such as word choice and syntax.

Speech emotion recognition systems use machine learning algorithms to analyse these features and classify the emotional state of the speaker. The most common emotions that are recognized are happiness, sadness, anger, fear, surprise, and neutral.

There are different approaches to speech emotion recognition, including the use of acoustic features, linguistic features, and a combination of both. The accuracy of speech emotion recognition systems can vary depending on the quality of the data, the complexity of the emotions being recognized, and the robustness of the algorithms used.

Why is speech emotion recognition important?

Speech emotion recognition has numerous applications, including improving communication between individuals, diagnosing mental health conditions, personalizing marketing and customer service, and improving human-computer interaction.

2. Stakeholders

The key stakeholders for speech emotion recognition systems may vary depending on the specific context and application of the technology, e.g. businesses and organizations may use speech emotion recognition systems to analyse customer feedback, evaluate employee performance, or monitor the emotional state of individuals in various contexts (e.g. speech therapists, psychologists).

3. Business question

How can speech emotion recognition systems be improved to classify the emotions conveyed in human speech more accurately?

4. Data question

Are the datasets of sufficient quantity, quality and complexity to accurately classify emotions conveyed in human speech?

5. Data summary

Two emotional speech datasets were downloaded from Kaggle. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) and the Toronto Emotional Speech Set (TESS).

For this project only a portion of the **RAVDESS** was used, namely the speech audio-only modality which has **1440 audio files**. It contains 24 professional actors (12 female, 12 male), vocalizing two lexically matched statements in a neutral North American accent. Speech emotions includes calm, happy, sad, angry, fearful, surprise, and disgust expressions. Each expression is produced at two levels of emotional intensity (normal, strong), with an additional neutral expression.

In the **TESS** data a set of 200 target words were spoken in the carrier phrase *"Say the word ___ "* by two actresses (aged 26 and 64 years) and recordings were made of the set portraying each of seven emotions (anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral). The TESS data contains **2800 audio files**.

Both datasets combined resulted in a **total 4240 .wav audio files**.

6. Data science process

Exploring audio data

The number of audio files for each emotion is plotted in Figure 1. Emotion 'calm' is under-represented (5% of total files), while all other emotions show equal representation (12-14% of total files).

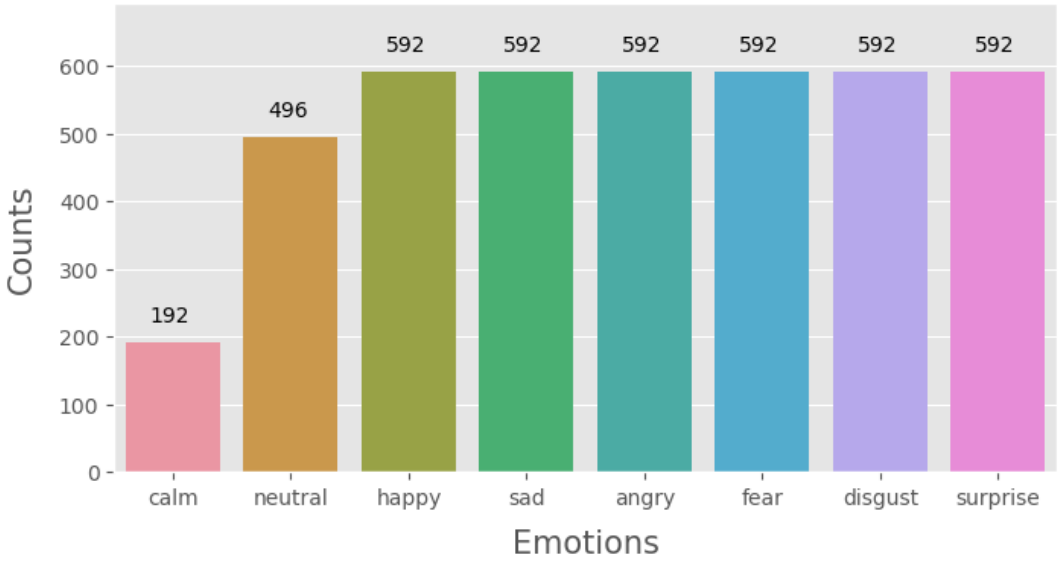

Fig. 1: Number of audio files for each emotion in the combined dataset

Visualizing audio data

Audio data is obtained by sampling the sound wave at regular time intervals and measuring the intensity or amplitude of the wave at each sample. The sampling rate which is the number of samples per second determines the sound frequency range. Audio signals can be plotted as waveplots and spectrograms. Figure 2 shows various visualizations. **Waveplots** (2D) show time and amplitude (intensity) of an audio signal, while **spectrograms** (3D) show the spectrum of frequencies changing with time. **MFCCs** (3D) represent the spectral characteristics of an audio signal. The latter two representations are also used in audio feature extraction for modelling.

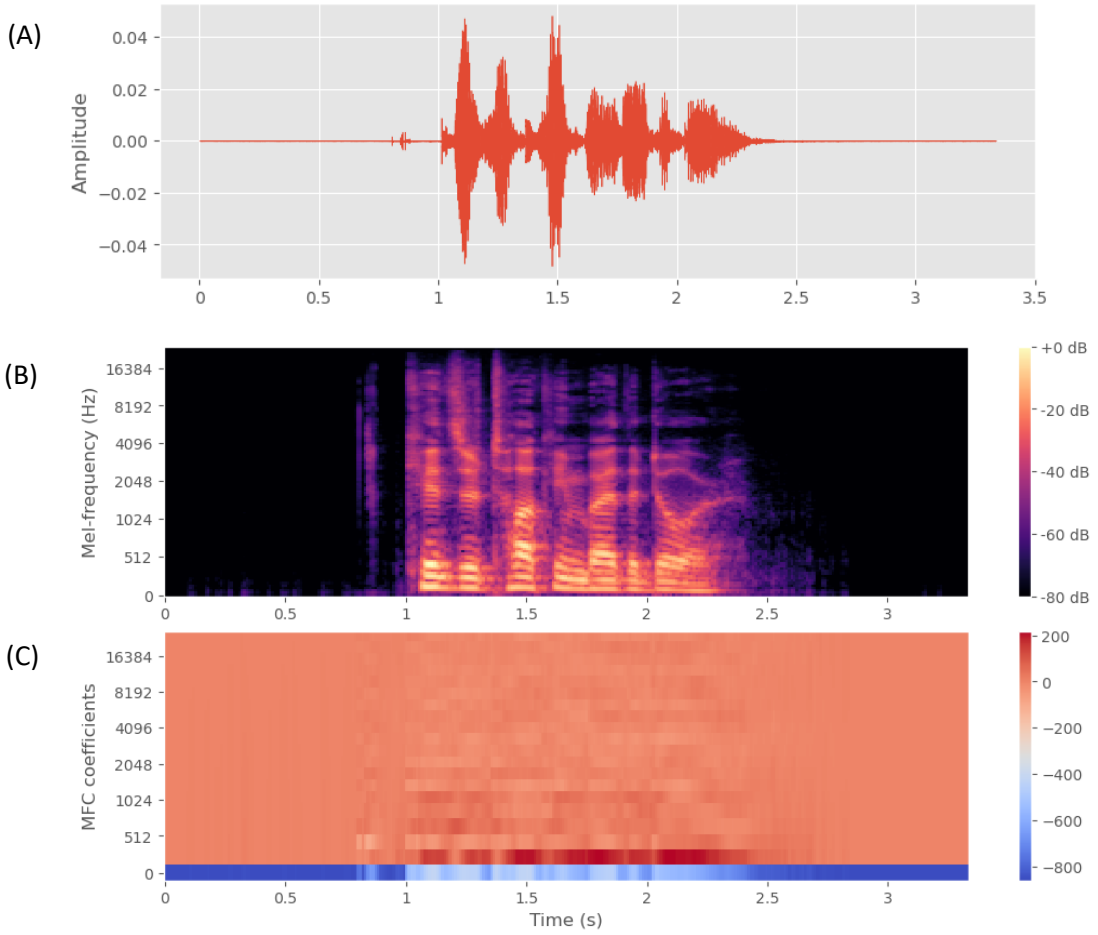

Fig. 2: Visualizing audio data for emotion ‘neutral’ as (A) waveplot (B) spectrogram and (C) MFCCs

Audio data augmentation

Data augmentation is the process by which new synthetic data samples are created by adding small perturbations to the initial data. To generate syntactic data for audio, **noise** injection, shifting **time**, changing **pitch** and **speed** can be applied. The objective is to make the model invariant to those perturbations and enhance its ability to generalize. However, adding perturbations must conserve the label of the original data. Figure 3 shows waveplots of the original and the augmented audio using noise injection

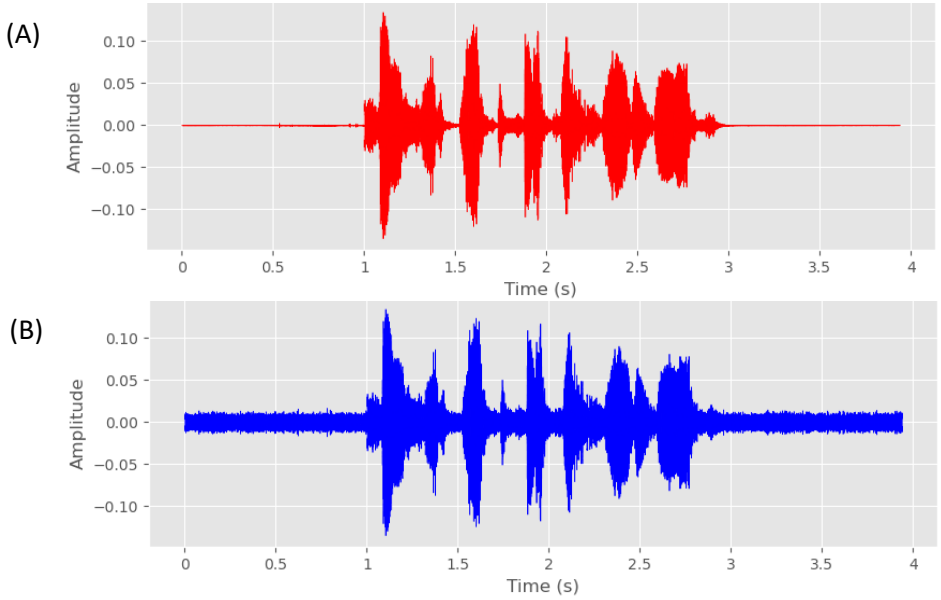

Fig. 3: Waveplots for emotion 'angry' of (A) the original and (B) the augmented audio using noise injection

Audio feature extraction

For feature extraction, sampling rate and audio file duration are required to be of equal size for each sample. Figure 4 shows the distributions of the native sampling rate and the audio file duration. The RAVDESS dataset has a sampling rate of 48 kHz and a mean duration of about 3.5 s, while the TESS dataset has a mixture of sampling rates and a mean duration of about 2 s. Prior to audio feature extraction, all audio files were **re-sampled** to 22050 Hz and **re-sized** to a length 5 s.

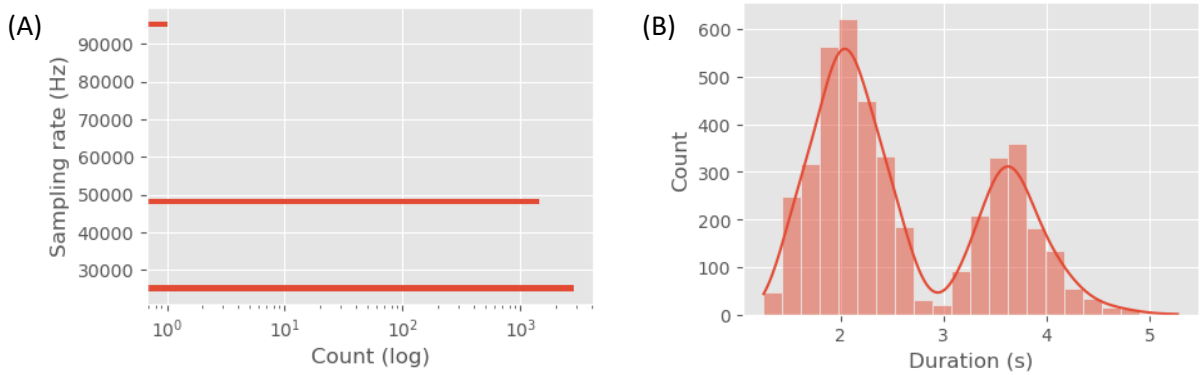

Fig. 4: Distribution of (A) native sampling rate and (B) audio file duration

Four different methods were employed to extract audio features (Table 1). Method #1 extracts mean MFCCs and stores the output for each sample as a 1D array. Method #2 extracts and concatenates multiple features (MFCCs, ZCR, RMSE) into a 1D array, then performs audio data augmentation before repeating the extraction process. The augmentation process resulted in a multiplication of the sample size output. Methods #3 and 4 extract MFCCs and (mel) spectrograms, respectively, and store the output for each of the 4240 samples as a 2D feature map.

#	Feature extraction method	Output of samples	Output dimension of feature map
1	Mean MFCCs	4240	1D (40,)
2	Combination of features (MFCCs, ZCR, RMSE) plus augmentation (noise, pitch)	16960	1D (9072,)
3	MFCCs	4240	2D (40, 216)
4	(Mel) Spectrogram	4240	2D (64, 216)

Table 1: Summary of methods employed for audio feature extraction. Definitions: MFCCs (Mel-frequency cepstral coefficients) represent the spectral characteristics of an audio signal. ZCR (Zero Crossing Rate) shows the rate at which an audio signal changes. RMSE (Root-Mean Square Energy) shows the average loudness of an audio signal over time.

Modelling

The following **data pre-processing** steps were performed prior to modelling.
The feature data was

- split into 80% train and 20% test with stratification enabled
- scaled into a range between 0 and 1
- re-shaped to fit the CNN model input shape

In addition, target values were encoded to a binary class matrix (label encoding).

For each feature extraction method, a variety of CNN models were investigated. The model architecture was optimized by trial and error using learning curves as assessment tools to minimise under- and overfitting. The best model architecture for each feature extraction method was finally evaluated and predictions were performed on the test data. Accuracy, precision, recall and F1-score were used to determine the model performance. In this report, only the results for the **best method/model-combination** are presented. Table 2 and Figure 5 show the performance metrics and learning curves for the best CNN model using feature extraction method #2, respectively. A neural network with 4 convolutional and 2 fully connected layers showed the best performance (for model details refer to Table 3), resulting in an accuracy of 95% and F1-scores of ≥94% for all emotions (except 'calm'). Note, emotion 'calm' was under-represented in this dataset, which explains the lower performance scores. Learning curves show only minimal overfitting.

Emotion	Precision	Recall	F1-score	# Samples test data
Neutral	0.90	0.97	0.94	397
Calm	0.79	0.82	0.81	153
Happy	0.96	0.93	0.95	473
Sad	0.96	0.93	0.95	473
Angry	0.96	0.96	0.96	474
Fear	0.98	0.94	0.96	474
Disgust	0.97	0.95	0.96	474
Surprise	0.97	0.98	0.97	474

Table 2: Performance metrics for the best CNN model using feature extraction method #2

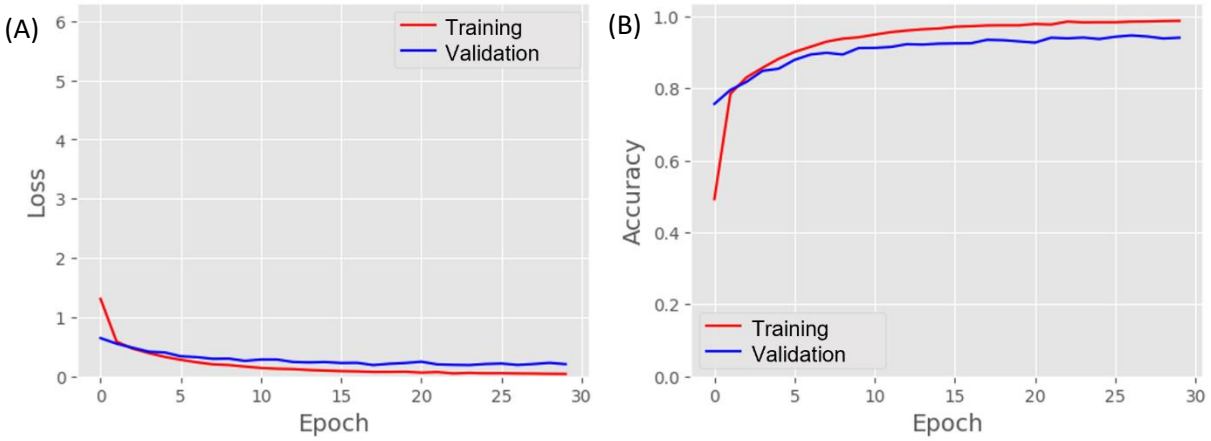

Fig. 5: Learning curves showing (A) loss and (B) accuracy of the best CNN model using feature extraction method #2

Model: Sequential	
Input	(9072, 1)
4 blocks of layers: Conv1D MaxPool1D Dropout	 nodes per block: 128, 128, 64, 32; kernel: 5; activation: relu pool: 2 0.2
Flatten layer	
Dropout layer	0.5
Dense layer	nodes: 256; activation: relu
Dense output layer	nodes: 8; activation: softmax
Optimizer	adam (learning rate=0.001)
Loss function	categorical_crossentropy
Metrics	accuracy
Epochs	30

Table 3: Details of the best performing CNN model

7. Outcomes

The accuracy of speech emotion recognition can be improved by data quantity and diversity, choice of feature extraction method and possibly augmentation. Models tend to overfit with small datasets – e.g. when using the RAVDESS only – even when augmented data was added. The best feature extraction method was the extraction of multiple features plus augmentation, which resulted in a 95% accuracy of classifying emotions from human speech when combined with a convolutional neural network

8. Implementation

Depending on the specific applications of the speech emotion recognition systems, an user interface will be developed in consultation with stakeholders and consumers.

9. Data answer

Combining two datasets resulted in an overall large and diverse enough dataset to accurately classify emotions conveyed in human speech. F1-scores for all emotions was equal or larger than 94% except for ‘calm’, which can be attributed to its underrepresentation in the dataset.

10. Business answer

The business question was answered adequately as outlined above. The proposed solution provides 95% accuracy to classify emotions conveyed in human speech.

11. Response to stakeholders

The project is very promising and expected to achieve its objective of improving the accuracy of speech emotion recognition. It is recommended to proceed with the implementation of the project.

12. End-to-end solution

Implementing the end-to-end solution requires the necessary technical infrastructure and computational power for data acquisition, modelling and storage. Deploying and integrating the end-to-end solution into the target application requires adapting the model to the specific use case and integrating it into the existing system architecture. Ongoing monitoring, optimization and maintenance is important to ensure continued effectiveness and reliability over time.

13. References

Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)
https://www.kaggle.com/datasets/uwrfkaggler/ravdess-emotional-speech-audio

Toronto Emotional Speech Set (TESS)
https://www.kaggle.com/datasets/ejlok1/toronto-emotional-speech-set-tess