FSER: Deep Convolutional Neural Networks for Speech Emotion Recognition

Bonaventure F. P. Dossou
Jacobs University, Germany
f.dossou@jacobs-university.de

Yeno K. S. Gbenou
Drexel University, USA
yg428@drexel.edu

Abstract

Using mel-spectrograms over conventional MFCCs features, we assess the abilities of convolutional neural networks to accurately recognize and classify emotions from speech data. We introduce FSER, a speech emotion recognition model trained on four valid speech databases, achieving a high-classification accuracy of 95.05%, over 8 different emotion classes: anger, anxiety, calm, disgust, happiness, neutral, sadness, surprise. On each benchmark dataset, FSER outperforms the best models introduced so far, achieving a state-of-the-art performance. We show that FSER stays reliable, independently of the language, sex identity, and any other external factor. Additionally, we describe how FSER could potentially be used to improve mental and emotional health care and how our analysis and findings serve as guidelines and benchmarks for further works in the same direction.

1. Introduction

Emotions are integral parts of daily communication between humans. Whether oral or written, properly understanding the interlocutor is key for a good communication. Even though emotions can be easily concealed in written communication, several psychological studies have shown that it is more difficult for humans to hide their feelings physically or vocally. More generally, facial expressions and voice tones are very good indicators of one’s emotional state. Many psychological and physiological studies [2, 6, 8, 10, 12, 16] (not limited to the ones cited here) have proved, that emotions make us feel and act; stimulating and influencing both our facial expressions and voice tone. For instance, adrenaline is released in fearful situations to help us run away from danger, as excitement or joy can be expressed while we are talking with friends, family, cuddling our pets or practicing risky sports or activities such as mountaineering or skydiving.

The recognition of emotional states in speech, so called Speech Emotion Recognition (SER) is not a new concept in the fields of Artificial Intelligence and Machine Learn-
| Emotion class | Amount of data (%) |
|--------------|--------------------|
| Anger        | 463 (15.22%)       |
| Anxiety      | 405 (13.31%)       |
| Calm         | 273 (8.97%)        |
| Disgust      | 382 (12.55%)       |
| Happiness    | 407 (13.37%)       |
| Neutral      | 379 (12.45%)       |
| Sadness      | 398 (13.09%)       |
| Surprise     | 336 (11.04%)       |
| Total        | 3043 (100%)        |

Table 1: Distribution of emotion classes with the numbers of speech samples of each class.

Savee is a British English language database of public speaking, made of speech recordings with seven emotions: happiness, sadness, anger, fear, disgust, surprise and neutral. The recordings were made by four English male actors, generating 15 samples for each emotion. The entire dataset contains 480 speeches.

Ravdess has been created by 24 authors: 12 men and 12 women with eight emotions: neutral, calm, happy, sad, angry, scared, disgusted, surprised. The entire dataset contains 1440 speeches.

Despite the fact that speeches across the datasets are sampled at the same frequency, they are structured in different and specific ways. Hence, we had to classify them appropriately and combine the same emotions from all datasets. The average length of speech samples is 3-4 seconds.

After the collection and classification of all emotions from the four different datasets, we got 10 classes of emotions: anger, anxiety, fear, boredom, calm, disgust, happiness, neutral, sadness, surprise. Due to the small number of fear and boredom emotions samples, we decided to add them respectively to anxiety and calm emotion classes. This was also motivated by the fact that, those speech samples have similar pitches and amplitudes. This results in 8 classes of emotion.

We chose to disregard gender information and to focus only on emotions. The final distribution of emotion classes with the numbers of speech samples of each class, are stated in Table 1.

3. Data Processing and Visualisation

Figure 1 shows the amplitudes of a random speech sample along the time measured in seconds (s). Speech signals are made up of amplitudes and frequencies. To get more information from our speech samples, we decided to map them into the frequency domain using the Fast Fourier Transformation (FFT) \[3, 4\].

![Figure 1: Speech sample visualisation using the Librosa library](image)

![Figure 2: Speech sample characteristics using Librosa, and scipy.fft libraries.](image)

![Figure 3: Speech sample converted into mel-spectrogram, with time, frequency and amplitudes information.](image)

Applying the FFT to a sample using the \texttt{scipy.fft} and \texttt{librosa} packages, gives Figure 2 which shows the initial speech plot, the double-side (negative and positive) FFT spectrum, and the positive FFT spectrum.

However, this is still lacking of time information. To remedy this, and make sure we preserve frequencies, time and amplitudes information about the speech samples, in reasonable and adequate range, we decided then to use mel-
Figure 4: Mel-Spectrograms of speakers speaking a sentence with anger (a) and anxiety (b).

spectrograms.

In a mel-spectrogram \[^4\] the abscissa represents time, the ordinate axis represents frequency, and amplitudes are showed by the darkness of a precise frequency at a particular time: low amplitudes are represented with a light-blue color, and very high amplitudes are represented by dark red (see Figure 3).

There are two types of spectrograms: broad-band spectrograms and narrow-band spectrograms. In our study, we used narrow-band spectrograms because they have higher frequency resolution, and larger time interval for every spectrum than broad-band spectrograms: this allows the detection of very small differences in frequencies. Moreover, they show individual harmonic structures, which are vibration frequency folds of the speech, as horizontal striations.

In Figure 4 are provided examples of mel-spectrograms of speakers speaking a sentence with anger \[^4\](a) and anxiety \[^4\](b). We converted all original speech signals to mel-spectrograms using the \[^5\]librosa melspectrogram module of the \[^6\]librosa library, and transformed the SER task into a pattern recognition and image classification problem.

We used 512 as length of the FFT window, 512 as the hop-length (number of samples between successive frames) and a hanning windows size is set to the length of FFT window.

Dataset Splitting: 2434 images (80%) have been allocated for the training phase. The remaining 609 (20%) have been used as testing set, to evaluate the model performance. As a standard in the machine learning field, we then split the initial training into two subsets: training set (80% - 1947 images) and validation set (20% - 487 images). To increase the size of the final training set, we applied a very common technique that has shown much success in recent years with CNNs: data augmentation \[^13\] [18].

Data Augmentation: Using the algorithm \[^1\] we implemented the data augmentation (DA) method, generating 20 additional pictures from every original input picture of our training set. To each input image (mel-spectrogram) object, we applied width and height shifting, zooming and horizontal flipping, using the \[^6\]keras.preprocessing.image module from the \[^6\]Keras deep learning framework. After the DA, our training set size increased to 40887 images (mel-spectrograms).

|               | Precision | Recall | F1-Score | Support | AUC-Score |
|---------------|-----------|--------|----------|---------|-----------|
| Anger         | 96        | 96     | 96       | 91      | 99.84     |
| Anxiety       | 94        | 97     | 95       | 80      | 99.85     |
| Calm          | 93        | 97     | 95       | 54      | 99.92     |
| Disgust       | 97        | 92     | 94       | 77      | 99.72     |
| Happiness     | 95        | 93     | 94       | 83      | 99.75     |
| Neutral       | 94        | 94     | 94       | 77      | 99.74     |
| Sadness       | 95        | 94     | 94       | 78      | 99.76     |
| Surprise      | 94        | 95     | 94       | 69      | 99.82     |

|               |            |      |          |         |           |
| accuracy      | -          | 95   | 609      | -       |           |
| macro avg     | 95         | 95   | 609      | 99.79   |           |
| weighted avg  | 95         | 95   | 609      | 99.79   |           |

Table 2: Classification Report of the Predictions of FSER on the Testing set

[^4]: http://www.glottopedia.org/index.php/Spectrogram
[^5]: https://librosa.org/doc/main/generated/librosa.feature.melspectrogram.html
[^6]: https://keras.io/api/preprocessing/image/
4. Model Architecture, Training and Results

4.1. Model Architecture and Training

Figure 5 presents the architecture of FSER. The input is an RGB image of shape (64, 64, 3). Our model is made of four blocks of convolutional layers (CLs) with ReLU activation [11], maximum 2D pooling layers (MPLs) [19] and Dropouts [5] interposed between them. The output of the four CLs is flattened and fed into a series of three fully-connected layers, which is finally fed into a softmax activation unit [7] to output the probability of the input to belong to each class of emotion.

Hyper-parameters and Training: We used a batch size of 64, a learning rate of 0.001 and Stochastic Gradient Descent [14, 15] as optimizer. The loss function is the categor-
Table 4: Comparison of our work to other relevant related works on Emodb, Savee and Ravdess datasets.

| Relevant related works ↓ | Speech Dataset | Features | Accuracies (%) |
|--------------------------|----------------|----------|----------------|
|                          | Emodb | Emovo | Savee | Ravdess | Mel-Spectrogram (*) | MFCC (**) |
| Qayyum et al. [1]        | -     | -     | +     | -      | -                  | +         | -       | 83.61  |
| Kannan, V. and Haresh, R. [17] | -     | -     | -     | +      | +                  | +         | 86      | 53     |
| Zhao et al. [20]         | +     | -     | -     | -      | -                  | +         | 95.33   | -      |
| Our work (Emodb)         | ✓     | -     | -     | -      | ✓                  | -         | ✓       | 97.74  | X      |
| Our work (Savee)         | -     | -     | ✓     | -      | ✓                  | -         | ✓       | 99.10  | X      |
| Our work (Ravdess)       | -     | -     | -     | ✓      | ✓                  | -         | ✓       | 98.67  | X      |

Table 4: Comparison of our work to other relevant related works on Emodb, Savee and Ravdess datasets.

4.2. Results and potential use case

Tables 2 and 3 present respectively the classification report of FSER on the SER task and the normalized confusion matrix. As shown in Tables 2 and 3 and in Figure 6, FSER has a nearly-perfect score for every emotion class.

Additionally, Table 4 shows the comparison of our work to other works done on the SER task, with the datasets of interest. We can interpret Table 4 as follow:

- On Emodb, our FSER (97.44) outperformed the mel-spectrograms-based approach proposed by Zhao et al. [20] (95.53).
- On Savee, our FSER (99.10) outperformed the traditional MFCCs-based approach proposed by Qayyum et al. [1] (83.61).
- On Ravdess, our FSER (98.67) outperformed both mel-spectrograms and traditional MFCCs approaches proposed by Kannan, V. and Haresh, R. [17] (86 and 53 respectively).

To test the robustness of FSER, we collected real-life audio samples from French, Fon, and Igbo speakers. Each of them was asked to imitate as much as they can those emotions, without providing any guidelines. In total, we had 24 audios (3 samples per emotion class), out of which FSER predicted correctly 20 audios. The 4 misclassified audios were found to belong to the classes fear and boredom that we had to respectively add and consider as anxiety and calm classes. Recalling that FSER has been trained in English, German, and Italian, its performance on these real-time audios, shows that FSER stays reliable, independently of the language, sex identity, and any other external factor.

Potential use case: For humans, emotions constitute an important factor to personal and global health. Being emotionally unstable can affect not only mental wellness but also physical health. During this pandemic, many health reports demonstrated a significant increase in feelings of loneliness and emotional instability leading to and increased rate of suicide for example. Being emotionally healthy is hence crucial for our well-being. As AI-based systems are nowadays helping doctors to quickly diagnose diseases such as cancers, FSER-like systems could also be useful to doctors, particularly psychologists, in identifying the emotional state of their patient. It is no news that some people have trouble expressing their emotions with words, so FSER brings a potential solution to help physicians to better understand their patients, and provide them more efficient treatments. They could also be useful to organizations providing mental care and support to families, clients, or communities at large.

5. Conclusion

In the current study, we evaluated the abilities of CNNs to effectively recognize and classify emotions from speech data. We introduced FSER, which outperformed all models that have been introduced for SER task. We showed that FSER is reliable, with no regards to the language spoken, the sex identity of the individual, or any other external factor. FSER-like systems could be beneficial to the healthcare system, as they can help to better understand patients, to provide them better and quicker treatments. It is worth mentioning the limitations of this work, notably the data augmentation we had to perform to deal with the limited number of samples.

References

[1] A. B. Abdul Qayyum, A. Arefeen, and C. Shahnaz. Convolutional neural network (cnn) based speech-emotion recognition. In 2019 IEEE International Conference on Signal Processing, Information, Communication Systems (SPICSCON), pages 122–125, 2019.
[2] Daniel C. Albas, Ken W. McCluskey, and Cheryl A. Albas. Perception of the emotional content of speech: A comparison
from keras.preprocessing.image import *
import numpy as np

for image in images:
    # images is the list of all images
    image = load_img(image)
    image = img_to_array(image)
    image = np.expand_dims(image, 0)

    # construct the image data
    # generator for data augmentation
    aug = ImageDataGenerator(*args)
    imageGen = aug.flow(image, *args)
    total = 0

    # loop over image data generator
    for image_generated in imageGen:
        total += 1
        if total == 20:
            break

Algorithm 1: Algorithm for the generation of new samples for Data Augmentation