Machine Learning Project : Emotion recognition(Schiller et al., 2020)

KADI Koussaila koussaila.kadi@etu.sorbonne-universite.fr
Master Ingénierie des Systèmes Intelligents
Sorbonne Université
Paris, France

GAN Ruyu ruyu.gan@etu.sorbonne-universite.fr
Master Ingénierie des Systèmes Intelligents
Sorbonne Université
Paris, France

ZAMBETTA Elisa elisa.zambetta@etu.sorbonne-universite.fr
Master Ingénierie des Systèmes Intelligents
Sorbonne Université
Paris, France

CUI Xinyi xinyi.cui@etu.sorbonne-universite.fr
Master Ingénierie des Systèmes Intelligents
Sorbonne Université
Paris, France

Editor: Avanced Machine Learning (2021-2022)

Abstract

The research and development of neural networks has had a wide impact not only in the field of image recognition, but has also led to innovations in other areas such as speech recognition and natural language understanding. Such as onvolutional neural networks are often used in computer vision problems to extract features and important information from images, and most of the time spectrogram representations of audio data are used in the majority of learning problems, especially for speech recognition, as these time-frequency representations carry a lot of information. However, in this work, we will try to use a convolutional SampleCNN based model on raw audio signals

Keywords: speech recognition, emotion recognition, features extraction

1. Introduction

Speech is an important means of conveying emotions and provides us with a wealth of emotional information. Speech recognition of emotions is expected to be used in a number of areas such as remote telephone services, driving systems and lie detection and security checks. In remote telephone services, monitoring a customer's emotions can instantly alert customer service staff to provide appropriate service and attention. While driving, the driver's mood can be understood at any time in order to adjust the temperature and sound level of the music in the car, ensuring safe driving and reducing the occurrence of dangerous driving. Emotion recognition technology in the field of lie detection and security checks can effectively improve the efficiency and quality of security checks by using the unique emotional characteristics of criminals.

In this paper, we will discuss a form of emotion recognition in raw speech. Based on an iterative reading of papers by authors such as Dominik Schiller, we have divided the main work into two main parts, the first focusing on the acoustic system and the second on the linguistic system. In the acoustic system, we will create models based on convolutional neural network structures. We chose

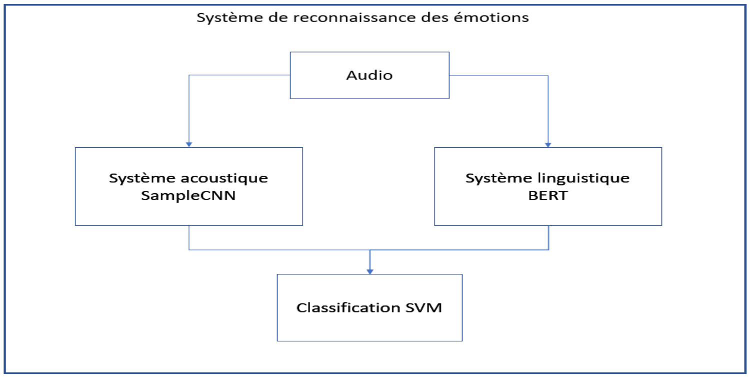

Système de reconnaissance des émotions
Audio
Système acoustique SampleCNN
Système linguistique BERT
Classification SVM

Figure 1: Emotion recognition system

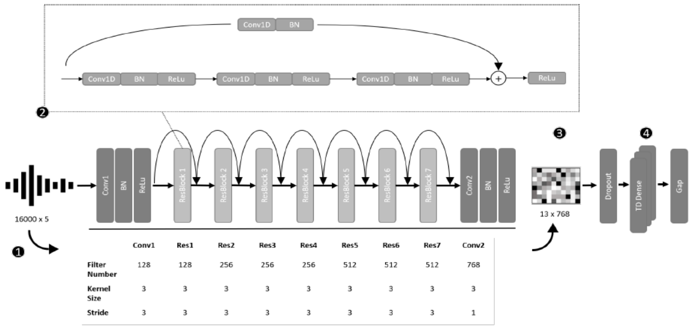
an open source audio file from Librispeech as our dataset and used the model we created to extract features and classify it. The natural language processing part of the linguistic system will use the bert model to extract and classify text features for emotion recognition.

2. Presentation of the Method

2.1 Acoustic System

The speech signal can be converted into a spectrogram after time-frequency analysis, and we can usually recognise speech on the basis of the speech spectrogram. However, the speech spectrogram is characterised by its structure. We need to overcome the diversity of the speech signal to improve the accuracy of speech recognition. The diversity of the speech signal includes the diversity of the speaker (the speaker himself, as well as the listener), and the diversity of the environment.

The advantage of the multilayer structure of CNNs is that they can extract information at several levels from the original speech signal. This means that the spectrogram obtained from the analysis of the speech signal is treated as if it were an image, and is recognised using the deep convolutional networks widely used in images.Our speech classification model is based on the SampleCNN model, which accepts the original audio waveform as input.



	Conv1	Res1	Res2	Res3	Res4	Res5	Res6	Res7	Conv2
Filter Number	128	128	256	256	256	512	512	512	768
Kernel Size	3	3	3	3	3	3	3	3	3
Stride	3	3	3	3	3	3	3	3	1

Figure 1: *Architecture of the acoustic classification model. The raw 16kHz waveform is fed to the fully convolutional model in 5 second chunks (1). To train a deeper network with smaller filter kernels effectively, residual connections are added to each convolutional block (2). The outputs of the convolutional network are the final extracted feature representations. During the training of the model those features are fed to a dense layer classification part (4).*

Figure 2: Convolutional network architecture(Schiller et al., 2020)

In psychology, there are two main sets of models for describing emotions in speech signals. One is the categorical approach and the other is the dimensional approach. The categorical approach is based on a set of primary emotions, the model divides the emotions into 8 basic components, which we can distinguish with different colours and shades (cf.Fig3).

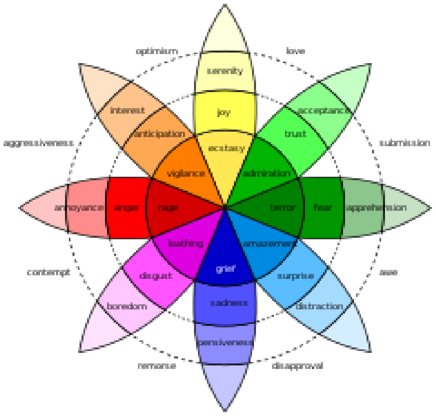


Figure 3: The categorical approach(Kerkeni, 2020)

The second approach, the dimensional approach, which is also the one we used for this project, presents emotions through two dimensions, valence and activation (or arousal). The role of valence is to classify emotions into two categories, positive or negative, and the role of activation is to subdivide these two categories into a hierarchy (cf.Fig4), according to which we can obtain a more detailed classification of emotions.(Kerkeni, 2020)

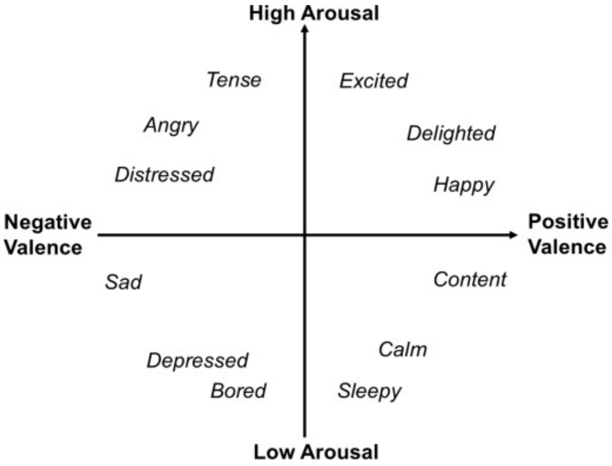


Figure 4: The dimensional approach(Du et al., 2020)

2.2 Linguistic System

In this linguistic system processing section we use the Bert model to implement pre-training on the text data. BERT is a new language representation model, which uses a bidirectional Transformer network to pre-train a language model on a large corpus, and fine-tunes the pre-trained model on other tasks. The task-specific BERT design is able to represent either a single sentence or a pair of sentences as a consecutive array of tokens. For a given token, its input representation is constructed by summing its corresponding token, segment, and position embeddings. For a classification task, the first word of the sequence is identified with a unique token [CLS], and a fully-connected layer is connected at the [CLS] position of the last encoder layer, finally a softmax layer completes the sentence or sentence-pair classification.

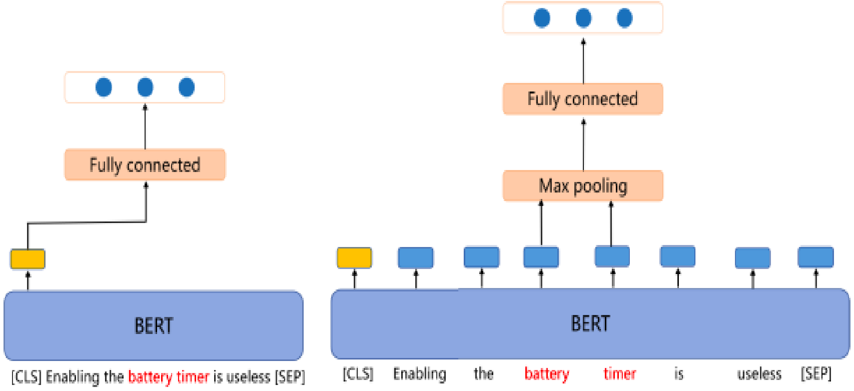


Figure 5: The architecture of BERT-FC (left) and TD-BERT (right).(Z. Gao and Wu, 2019)

First we train the AutoTokenizer model with the dataset and build a sentiment_analysis model to implement sentiment classification. After the text data processing, we obtain a dataframe containing the text and each of their sentences. We use a large amount of text as the entry point to train the sentiment_analysis model, which will calculate the prediction score corresponding to each sentence's classification (neutral, positive, or negative). Then, for each sentence, the highest prediction score is used to determine the type of the sentence. Finally, the number of all types is counted for each text, and the most occupied classification defines the type of the text.

3. Dataset

3.1 Acoustic System

In the dataset we distinguish 3 types of data: corpus of acted emotions, corpus of induced emotions and corpus of natural emotions. The database we used in our project is called Crema dataset, which is a corpus of acted emotions. According to the categorical approach there are 6 types of emotions (cf.Fig6). According to the dimensional approach we have 3 levels of emotion activation (cf.Fig7).

In this part, since we have a dataset with 7442 voice samples with different durations in average higher than 5 seconds, so to be able to load this dataset without saturating the RAM memory we were obliged to use another approach of loading data, this method is called data_generation which is based on multiprocessing and generates batches according to the need of the fit() method to train the model, so it loads at the same time a number of batch as the number of processes.

Our dataset at the beginning is labelled according to the categories of emotions, but our need is to have a database labelled according to the arousal values, so we converted the labels of each emotion with a value (low: for feelings that have low arousal values such as sad feeling 'sad', high: for feelings that have high arousal values such as angry feeling 'Anger', middle for neutral feelings) according to Plutchik's wheel of emotions (1980).

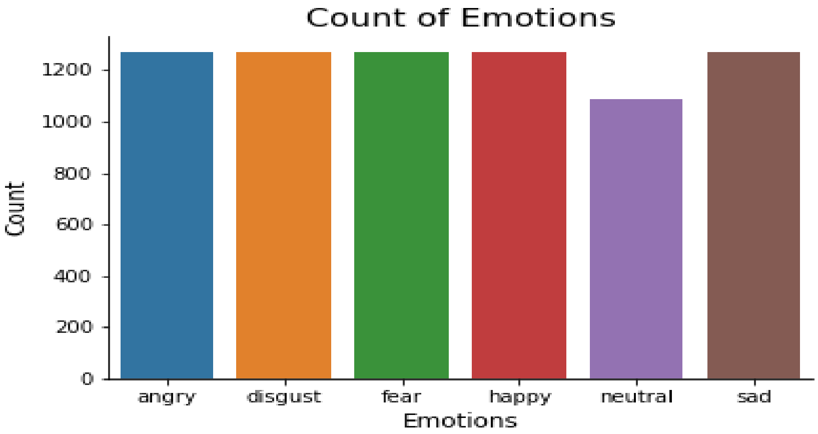


Figure 6: Type of emotion in dataset

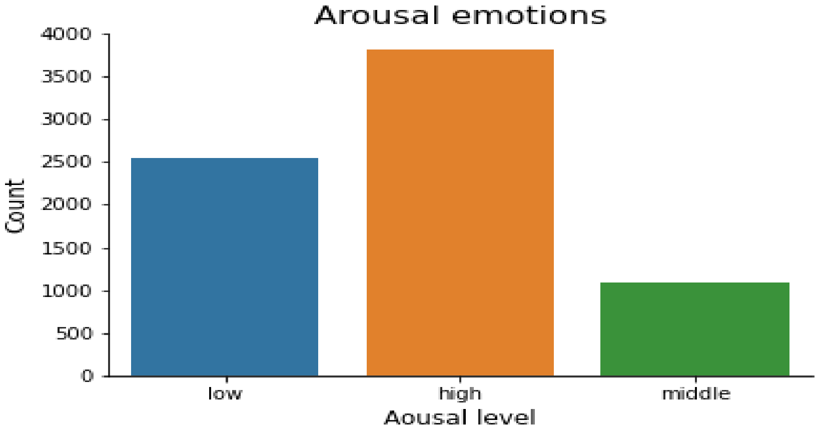


Figure 7: Arousal emotions

3.2 Linguistic System

For the part concerning the linguistic model we also used the LibriSpeech database, which also contains transcriptions of audio books. Each text file (cf. github *datasettext*) contains extracts of books from LibriSpeech's Gutenberg project.

4. Experiments

You can find the implementation of this article on the following link : `https://github.com/Koussailakadi/Speech_Input_for_Emotion_Recognition`

4.1 Results

4.1.1 Acoustic System

For the acoustic part, we trained the model on 10 epochs, and we notice that we have an over fitting from epoch 6. (cf.Fig9)

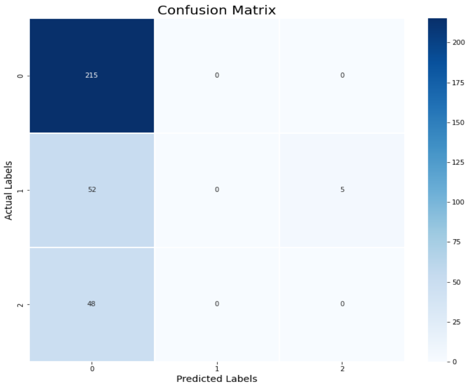


Figure 8: Confusion matrix

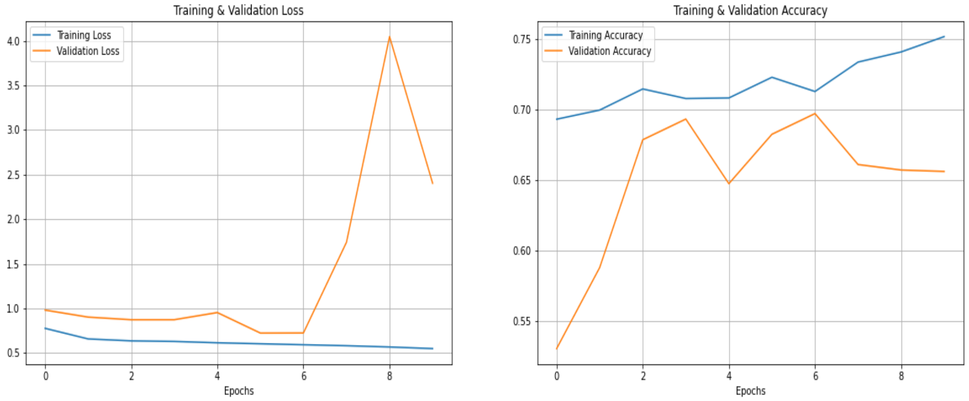
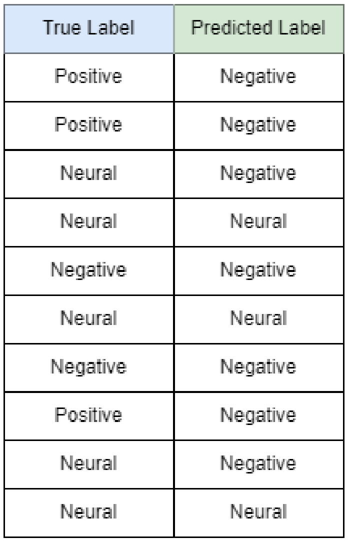


Figure 9: Comparison of accuracy and loss value

4.1.2 Linguistic System

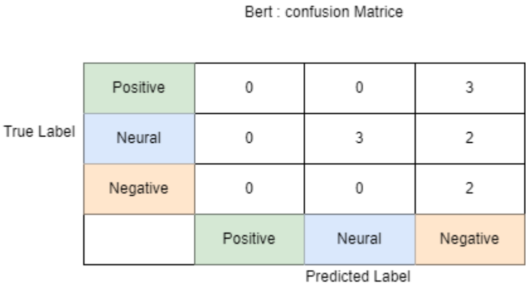
To evaluate our linguistic model we took 10 examples from our database and we manually evaluated the global class of each example according to three classes (positive, negative and neutral). We evaluate manually because we do not have the target of each example. In the following table you can see our results(cf.Fig.10):



True Label	Predicted Label
Positive	Negative
Positive	Negative
Neural	Negative
Neural	Neural
Negative	Negative
Neural	Neural
Negative	Negative
Positive	Negative
Neural	Negative
Neural	Neural

Figure 10: Predicted and true label

.



Bert : confusion Matrice

True Label				
	Positive	0	0	3
	Neural	0	3	2
	Negative	0	0	2
		Positive	Neural	Negative
			Predicted Label	

Figure 11: Confusion matrix for the linguistic model

.

We can see that our model have an acurracy around 50%(cf.Fig11).

4.2 Discussion

4.2.1 Acoustic System

The best accuracy obtained is 70 % on the validation basis, and 75 % on the training basis.

The confusion matrix (cf.Fig8) shows the number of well ranked examples out of the total number of test examples, we can see that our model is biased on the activation value 'High' which corresponds to a high arousal, as we have a lot of feelings with high arousal, as mentioned in the dataset part.

4.2.2 Linguistic System

For the reference article(cf.(Schiller et al., 2020)) they showed that for the linguistic system they found an accuracy of 52% (cf.Fig12). In the paper, they used a large dataset composed of German Wikipedia dumps, legal texts, and news articles to use the pre-trained BERT model.

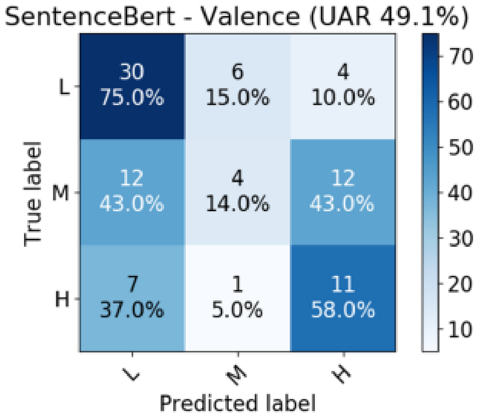


Figure 12: Confusion matrix for the linguistic model (Schiller et al., 2020)

.

Contrary to the article we have used the LibriSpeech database which is composed of audio books in English, and instead of using the BERT pre-training model for text classification we have used a derivative of this one which is called HeBERT ((Chriqui and Yahav, 2021) and (HeB)).We chose to use Hebert rather than the pretrained Bert model because we lacked the targets of the examples to use Bert. To find the class of a text example using the HeBERT model, it returns a class (negative, positive or neutral) for each sentence and the final classification of the text corresponds to the maximum classification of all sentences. We obtain the following results for our data with a performance of 50%(cf.Fig11). Despite the differences between the article and what we have implemented, we have a comparable performance.

5. Conclusion

Automatic feature extraction using the proposed CNN model performs better than manual feature extraction. However, the linguistic model provides a better performance on emotion classification. I noticed when classifying text that some positive words like ('joyful', 'positive' or 'happy') influence the classification. The final classification of the sentence tends towards the positive class whereas when reading the sentence it turns out to be neutral or negative. We notice the same thing for the negative class some words present in the sentence influence the classification like "unhappy" for example.

6. Bibliography

References

HeBERT: Pre-trained BERT for Polarity Analysis and Emotion Recognition, https://github.com/avichaychriqui/HeBERT.

Avihay Chriqui and Inbal Yahav. Hebert & hebemo: a hebrew bert model and a tool for polarity analysis and emotion recognition. *arXiv preprint arXiv:2102.01909*, 2021.

Na Du, Feng Zhou, Elizabeth M Pulver, Dawn M Tilbury, Lionel P Robert, Anuj K Pradhan, and X Jessie Yang. Examining the effects of emotional valence and arousal on takeover performance in conditionally automated driving. *Transportation research part C: emerging technologies*, 112: 78–87, 2020.

Leila Kerkeni. *Analyse acoustique de la voix pour la détection des émotions du locuteur*. PhD thesis, Le Mans, 2020.

Dominik Schiller, Silvan Mertes, and Elisabeth André. Embedded emotions–a data driven approach to learn transferable feature representations from raw speech input for emotion recognition. *arXiv preprint arXiv:2009.14523*, 2020.

X. Song Z. Gao, A. Feng and X. Wu. Target-dependent sentiment classification with bert. *IEEE Access, vol. 7, pp. 154290-154299, 2019, doi: 10.1109/ACCESS.2019.2946594*, 2019.