Speech Emotion Recognition of Intelligent Virtual Companion for Solitudinarian

Mutaz Alnahhas(✉), Tan Wooi Haw, and Ooi Chee Pun

Digital Home and Lifestyle Centre, Faculty of Engineering, Multimedia University, Cyberjaya, Malaysia
1211400101@student.mmu.edu.my

Abstract. Human emotions are essential to recognize the behaviour and the state of mind of a person. Emotion detection through speech signals has started to receive more attention lately. Living alone could be hard for some people due to the lack of social interaction, as they might develop a series of negative emotions daily. Furthermore, there are some unavoidable circumstances when family members need to live away from their families, leaving their old parents to live alone. These circumstances may cause parents to experience anxiety or a decline in mental health, which is a major cause for concern for their children. This is where assisted living technology can come in to support. This research proposes the design and development of a speech emotion recognition system for solitary people to detect and monitor their state of mind as well as their daily emotional behaviour. The research has three main contributions. First, to implement a real-time system based on audio where we can predict emotions from recorded human voices via deep learning. Secondly, a model has been designed to use data normalization and data augmentation techniques for advanced classification. Finally, a speech emotion detection system has been created using a Long Short Term Memory (LSTM) recurrent neural network. This research aims to study solitary person activities at any time at home. The resulting system will be used for mental health monitoring.

Keywords: Speech Emotion Recognition · MLP classifier · MFCC · Emotion Recognition · Long Term Short Memory (LTSM)

1 Introduction

Speech is the most natural way to express ourselves as humans. We rely on it so much that when we have to utilize other modes of communication, such as emails or text messages, we realize how important it is. It is no wonder that emoticons have grown popular in text messages, as they can be misinterpreted, and we would like to convey emotion along with the text, just like how we do in conversation. As emotions assist us to better understand one another, it is only logical that we extend this knowledge to computers. Thanks to smart mobile devices that can accept and respond to voice commands with synthetic speech, speech recognition is already part of our daily lives. Speech emotion recognition (SER) could be used to enable them to detect our emotions, as well.

M. Y. bin Alias (Ed.): MECON 2022, AER 214, pp. 132–138, 2023.
https://doi.org/10.2991/978-94-6463-082-4_14

For more than two decades the SER has been discussed and its applications in human-computer interaction, robots, mobile services, call centres, and psychological assessment. Nevertheless, emotions are subjective which makes their recognition a challenging task.

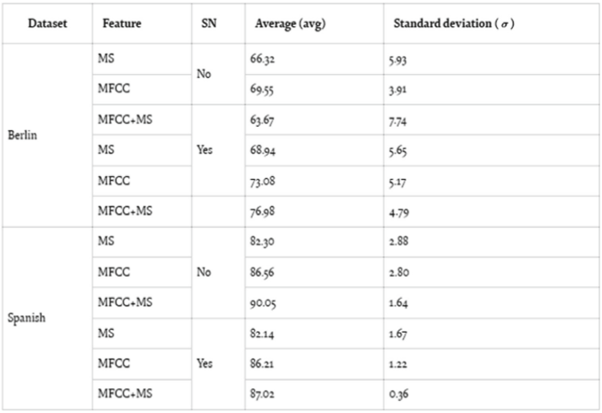
2 Literature Review

There have been many previous works focused on emotion recognition from acoustic features and mainly using speech as an input. Previous works have used various datasets such as RAVEDESS [1], IEMOCAP [2] and CREMA-D [3] dataset. Multiple algorithms have been used as well, such as SVM [4], CNN [3], RNN [5] and Multilayer Perceptron (MLP) Classifier [6].

Joy et al. [6], used an MLP classifier on RAVDESS dataset, their work showed an accuracy of 70.28% when tested on test samples and newly added data. The results also showcased the inference time in which it took around 77 s to predict the emotion using the Logistic activation function while having a slower result when using Relu activation function at 80 s.

Kerkeni et al. [5] used a different approach where they used a Recurrent Neural Network classifier along with using 2 different audio datasets separately (Berlin database, Spanish database). The features used in the paper were MFCC and modulation spectral (MS) while studying the results of applying speaker normalization. The achieved results can be showcased in Fig. 1.

Another approach is taken by Nguyen et al. [7] where they used transfer learning to mitigate the challenge of insufficient annotated emotion datasets which makes the trained models limited in their generalization capability. The use of transfer learning raises a flag where the fine-tuned knowledge might overwrite the important knowledge learned from the pre-trained models, in which they addressed this issue by proposing a PathNet-based transfer learning method to improve the quality. The paper not only used audio features, but they tested their method as well in visual emotion detection. Their proposed system



Dataset	Feature	SN	Average (avg)	Standard deviation (σ)
Berlin	MS	No	66.32	5.93
	MFCC		69.55	3.91
	MFCC+MS	Yes	63.67	7.74
	MS		68.94	5.65
	MFCC		73.08	5.17
	MFCC+MS		76.98	4.79
Spanish	MS	No	82.30	2.88
	MFCC		86.56	2.80
	MFCC+MS		90.05	1.64
	MS	Yes	82.14	1.67
	MFCC		86.21	1.22
	MFCC+MS		87.02	0.36

Fig. 1. Achieved results

achieved an accuracy of 85% when predicting emotion from speech, and its improved to 91% by further training it using transfer learning on different dataset. Another method is used by Ning et al. [8] where they were using speech signals and its implementation in real-time using Internet of things (IoT) based deep learning. The human voice has been recorded and emotion detection is created by integrating deep learning model using a 2D convolutional neural networks (CNN). The reported accuracy by this method was approximately 95% which outperformed all state-of-the-art approaches. Sharma et al. [9] used a different approach by creating two models for speech emotion recognition. Mel Frequency Ceptral Coefficient (MFCC) were used for feature extraction from audio files. The first model has been created using Multi-Layer Perceptron (MLP) classifier which gave an accuracy of 57.29%. The second model was created in Long Short-Term Memory (LSTM) and gave a good accuracy of 92.88%. RAVDESS dataset were used for classification.

3 Proposed Methodology

3.1 Dataset Selection

In this paper, we used RAVDESS dataset to train a Multilayer Perceptron Classifier. The dataset contains 1440 wav files of 24 actors (12 males, 12 females) to record eight (8) different emotions like happy, sad, calm, angry, surprise, disgust, neutral and fear. The total number of utterances is 1440 wav files with a 48,000 Hz sampling rate. All actors' age range is between 21–33 years old. The actors were in studio environment and were given a different set of utterances and recorded 10 times for each actor. The RAVDESS originally contains Video and Audio Samples, but we discarded the Video samples since our focus is on using speech features to determine a person's emotions. Some papers as in [2] used IEMOCAP dataset since it contains a transcription of the data as well, and since our focus is to only use the Acoustic feature, we didn't see a need in using the IEMOCAP dataset.

3.2 Feature Selection

In this paper, we will be using 3 acoustic features to predict the person's emotions from their voice signals. Which is MFCC, Chroma and Mel Frequency Cepstrum as speech features instead of using raw audio waveform which may contain unnecessary information that will affect the classification process.

MFCC. Mel Frequency Cepstral Coefficient is a representation of the short-term power of sound. Five common processes are used by MFCC to turn a raw audio sample into a more accurate representation for mathematical operations. [6].

- The Discrete Fourier Transform (DFT) of the signal inside the chosen window size (20 ms) is obtained.
- The spectrum is then perfectly matched to the mel scale.
- The natural logarithm is applied to all acquired mel frequencies.
- On the resulting log values, discrete cosine transform is done assuming that the log data represent a signal. The resulting values are the amplitudes of the final spectrum.

Chroma. In music, the Chroma feature is used to capture the harmonic melodic characteristic of the music. Thus gives the possibility in using it in emotion detection as each emotion follows a different melody [8].

Mel. Mel spectrograms are visual representation of audio signal where they are obtained by converting the frequency into the Mel scale. By applying Short-Time Fourier transform on the audio signal.

The formula of Converting Frequency into Mel scale is:

$$M(f) = 1125\ln\left(1 + \frac{f}{700}\right)$$

3.3 Emotion Classification

In our study, multiplayer perception (MLP) classifier is used. The (MLP) model is utilized to compute a suitable output from a different sets of input data. Three layers make up an MLP model. Input layer, hidden layer and output layer. Where hidden layer may be more than one. The structure of the MLP model is similar to a connected graph, where each layer's nodes are completely connected to one another via weighted edges. Multiple nodes make up each layer. There are two functions in each node: an input function and an output function. A supervised learning method called back propagation is used by MLP to train the network. [11] The input layer receives various speech features as input, and each node of the subsequent layer receives an input as a weighted sum of each node of the preceding layer. Each node in MLP generates output using a nonlinear function known as an activation function. There are various approaches to build this output function (Fig. 2).

The model parameter that determined by the grid search were used. It is a fully connected neural network with single layer that contains 300 units, the size of the batch used is 256 with iterations of 500 and an adaptive learning rate (Fig. 3).

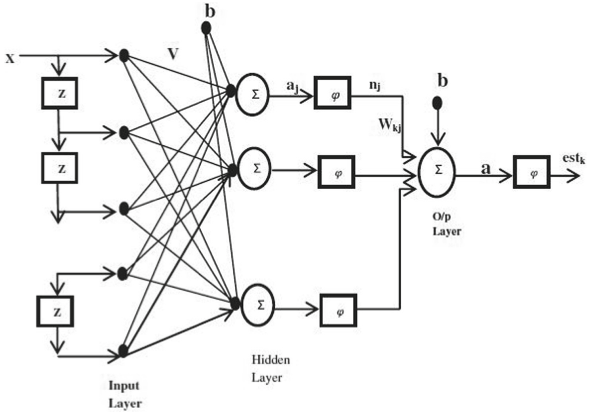
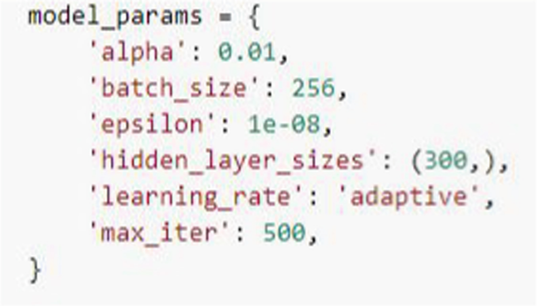


Fig. 2. The structure of MLP



```
model_params = {
    'alpha': 0.01,
    'batch_size': 256,
    'epsilon': 1e-08,
    'hidden_layer_sizes': (300,),
    'learning_rate': 'adaptive',
    'max_iter': 500,
}
```

Fig. 3. Model Parameter

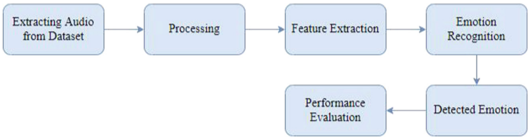
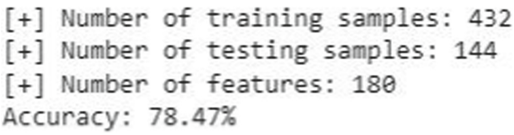


Fig. 4. System Architecture flow chart



```
[+] Number of training samples: 432
[+] Number of testing samples: 144
[+] Number of features: 180
Accuracy: 78.47%
```

Fig. 5. Training and testing samples

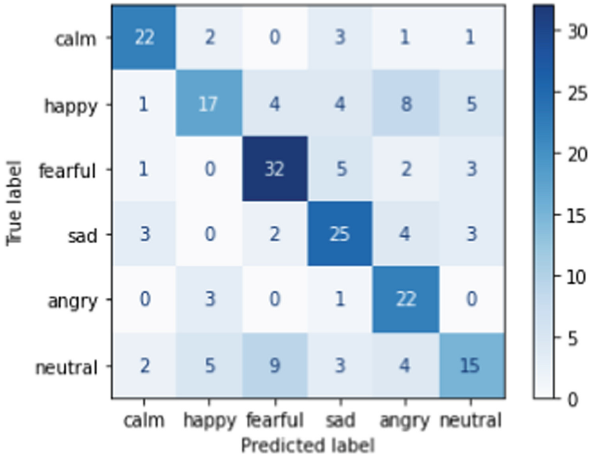
3.4 System Architecture

The overall system Architecture starts from the database where the raw audio data is saved in. The audio is extracted from the database to be pre-processed to get rid of any unnecessary silence. The 5 features are then extracted and fed into the model in which the model will predict the most suitable emotion and the resulted output will be stored back into the database as notation of the person feelings. Figure 4 shows the proposed system architecture.

4 Results

We only trained the model on the following emotions “Anger, Neutral, Fearful, Happy, Calm, Sad” other emotions have been discarded for them been irrelevant in this project.

The model achieved an accuracy of 78.47% in which the model was able to give similar results when tested on new samples. The model’s ability in predicting each emotion varied where it shows strong capabilities in detecting “fearful, sad, calm and angry emotions” and a poorer result when it comes to “happy, calm” emotions (Fig. 5).



True label \ Predicted label	calm	happy	fearful	sad	angry	neutral
calm	22	2	0	3	1	1
happy	1	17	4	4	8	5
fearful	1	0	32	5	2	3
sad	3	0	2	25	4	3
angry	0	3	0	1	22	0
neutral	2	5	9	3	4	15

Fig. 6. Confusion matrix

The model achieved better accuracy when combining four emotions and shows significant results when combining two emotions. A confusion matrix has been constructed in Fig. 6 to give a clearer image of the model capabilities.

5 Conclusion

The performance of the model can be further improved by further training the model on new data, and in this case, we suggest a healthcare audio database AVEC 2013 (The Continuous audio/visual Emotion and Depression Recognition challenge) [12] which will help the model specialize in it.

Transfer learning can also be used to further train a pre-trained model which proven to be an effective way of increasing the overall model performance as in [7]. Another way of enhancing the of the robustness of the emotion recognition system is by fusion of classifiers. Or multimodal systems in which multiple models are trained secretly and fused together for a better performance.

Feature selection is also another direction in which we can improve to further increase the accuracy and the robustness of the model. Acoustic signals contain a huge amount of parameter that reflect the emotional characteristic of the speaker. One example is the usage of Modulation spectral features in [5] which showed an increase in the performance of the model when this feature is added.

The goal of our work was to develop a system that's capable of determining person emotions using their acoustic features. We aim to achieve much higher results by further improving and testing.

Acknowledgments. The authors would like to thank Multimedia University for supporting the research as well as the anonymous reviewers for their insightful suggestions and comments.

References

1. C. Huang, W. Gong, W. Fu, D. Feng, and S. Balint, “A Research of Speech Emotion Recognition Based on Deep Belief Network and SVM,” 2014, doi: https://doi.org/10.1155/2014/749604.
2. D. Issa, M. Fatih Demirci, and A. Yazici, “Speech emotion recognition with deep convolutional neural networks,” Biomed. Signal Process. Control, vol. 59, p. 101894, May 2020, doi: https://doi.org/10.1016/j.bspc.2020.101894.
3. L. Cristian Dut and A. Radoi, “Emotion Recognition System from Speech and Visual Information based on Convolutional Neural Networks Nicolae-C˘ at˘ alin Ristea.”
4. Y. Chavhan, “SPEECH EMOTION RECOGNITION USING SUPPORT VECTOR MACHINE.”
5. L. Kerkeni, Y. Serrestou, M. Mbarki, K. Raoof, M. Ali Mahjoub, and C. Cleder, “Automatic Speech Emotion Recognition Using Machine Learning,” in Social Media and Machine Learning, IntechOpen, 2020.
6. J. Joy, A. Kannan, S. Ram, and S. Rama, “Speech Emotion Recognition using Neural Network and MLP Classifier,” IJESC, 2020, doi: https://doi.org/10.5772/intechopen.80419.
7. D. Nguyen, S. Sridharan, T. Nguyen, S. Denman, D. Dean, and C. Fookes, “Meta Transfer Learning for Emotion Recognition,” 2020.
8. E. Glenn Schellenberg, A. M. Krysciak, and R. Jane Campbell, “Perceiving emotion in melody: interactive effects of pitch and rhythm,” Music Percept., vol. 18, no. 2, pp. 155–171, 2000, doi: https://doi.org/10.2307/40285907.
9. Sharma, S. (2021, January). Emotion Recognition from Speech using Artifi-cial Neural Networks and Recurrent Neural Networks. In 2021 11th Interna-tional Conference on Cloud Computing, Data Science & Engineering (Con-fluence) (pp. 153–158). IEEE.
10. Ning, J. I. A., & Zheng, C. (2021, April). Emotion Recognition of Depressive Patients Based on General Speech Information. In 2021 6th International Conference on Intelligent Computing and Signal Processing (ICSP) (pp. 618–621). IEEE.
11. S. Basu, J. Chakraborty, A. Bag and M. Aftabuddin, "A review on emotion recognition using speech," 2017 International Conference on Inventive Communication and Computational Technologies (ICICCT), 2017, pp. 109-114, doi: https://doi.org/10.1109/ICICCT.2017.7975169.
12. Valstar, M., Schuller, B., Smith, K., Eyben, F., Jiang, B., Bilakhia, S., … Pantic, M. (2013). AVEC 2013. Proceedings of the 3rd ACM International Workshop on Audio/visual Emotion Challenge - AVEC ’13. doi:https://doi.org/10.1145/2512530.2512533