Detecting Agitated Speech: A Neural Network Approach

Kevin Hetterscheid
University of Twente
P.O. Box 217, 7500AE Enschede
The Netherlands
k.j.t.hetterscheid@student.utwente.nl

ABSTRACT

Agitation is common across neuropsychiatric disorders such as dementia. It is considered as a symptom of distress which contributes to disability, institutionalization, and diminished quality of life for patients and their caregivers. Literature suggests that agitation can be monitored or detected by abnormal vocal and physical activities. For the scope of this research, voice-based activities were used. The aim is to construct a suitable neural network that can classify these voice activities. Several datasets (RAVDESS[1], TESS[2] and ElderReact[3]) are used to train and test a Recurrent Neural Network (RNN). This network is build using Long Short Term Memory (LSTM) layers and Bidirectional LSTM layers. Several combinations of these are used and compared to find the most suitable combination. The proposed model reaches an accuracy of 86%, which is in line with other state-of-the-art approaches.

Keywords

Neural network, Speech recognition, Agitated speech, RNN

1. INTRODUCTION

Agitation is very common in elderly people suffering from dementia. Agitation in the case of dementia is defined by Cohen-Mansfield [4, p. 309] as ”inappropriate verbal, vocal or motor activity that is not judged by an outside observer to result directly from the needs or confusion of the agitated individual”. It contributes to a diminished quality of life for both patient and their caregivers [5]. Agitation is operationalised by using traditional agitation measuring scales like Cohen-Mansfield et al.[6], and a scale for the observation of agitation in persons with dementia of the Alzheimer type (SOAPD)[7]. These scales mandates the presence of caregivers to observe the patients continuously. To avoid the undependability of agitation monitoring or placing the burden of detection on caregivers, technical interventions were explored by Valembois et al.[8]. The aim of these technical interventions was to detect agitation in earlier stages. But these state-of-the-art technical interventions are either wearable or invade the privacy of the user[9]. For example, a fitness band worn on the wrist monitors the physiological activities of a person. A camera based system for monitoring physical activities of patients is also an option. These systems not only creates discomfort for the user and but also demands attention of user. Therefore, an unobtrusive approach to detect agitation in early stage is required.

SOAPD[7] and the Cohen-Mansfield scale[6] has shown that agitation can be demarcated by physical and voice based activities. Of interest to this work is voice based agitation activities or agitated speech. In SOAPD[7], three type of vocal activities are outlined: high-pitched or loud noise, repetitive vocalization and negative words. Among these three categories we will only be taking activities that can be monitored by prosodic features of voice. Prosodic features can be categorised in auditory terms (pitch, loudness, timber) and in acoustic terms (fundamental frequency, duration, intensity, spectral characteristics). Further, there are two main factors that needs to be considered while detecting agitated speech. These are: the characteristics of agitated speech and a suitable method which can grasp such subtle details of agitated speech. In this work we want to go a step ahead of traditional neural network approaches by using recurrent neural networks (RNNs). RNNs are a popular architecture of Neural Network, used for sequential or contextual data analysis. Unlike traditional neural network where all the inputs and outputs are independent of each other, an RNN's output from the previous step are fed as input to the current step. Considering the complexity of human speech, RNN's are useful for voice recognition as they remembers each and every information with respect to time i.e. long short term memory (LSTM). The 2D input features representing Mel Frequency Cepstral Coefficients (MFCCs) and Mel-spectrograms will be fed to the RNN. MFCCs takes into account human perception for sensitivity at appropriate frequencies by converting the conventional frequency to the mel scale[10]. It is a representation of the power spectrum of a sound, converted to the mel scale[10]. It was first introduced by Mermelstein[11]. A mel spectrogram is a spectrogram where the frequencies are converted to the mel scale[10]. These features are converted to the mel scale[10], which simulates human sound perception. As a consequence of this, they are often used in the area of speech processing, and thus will be beneficial to this task.

Apart from using spectral voice features, various configurations of Long Short Term Memory (LSTM) layers and Bidirectional LSTM layers will be tested in this work. This will be done to find the best suited configuration for this specific task and will be compared with other state-of-the-art results. This leads to the following research questions:

- Which neural network is best suited for recognising agitated speech?

33rd Twente Student Conference on IT July. 3rd, 2020, Enschede, The Netherlands.

- What configurations of the neural network will yield the best results for recognising agitated speech?

2. RELATED WORK

The optimal features that can be used for voice analysis are always debatable. In the survey conducted by, El Ayadi, Kamel and Karray[12] it was shown that there is no consensus about which features to use for emotion recognition considering each feature having its own strengths and weaknesses. Having said that, optimal features depends on the application of voice processing. Shah et al[13] stated the benefits of using chromagrams in speech detection where pitch of the sound is major factor as chromograms are more processed version compared to other voice representations. Another interesting approach is proposed by Bachu et al[14]. They pose a technique to identify which parts of audio clips are voiced and which parts are unvoiced. This is done by using two specific features, namely the zero-crossing rate (ZCR) and the energy. When the ZCR is low and the energy is high, the clip is most likely voiced. This inverse is also true. Further, one of the major application of voice detection is emotions detection. Issa et al.[15] used a deep convolutional neural network to recognise emotions in speech. Their results shows an impressive success rate (64% and higher) over 3 different datasets (RAVDESS[1], EMO-DB[16] and IEMOCAP[17]). The RAVDESS[1] model and the IEMOCAP[17] model outperformed the state-of-the-art approaches. The EMO-DB[16] model outperformed all but one. To predict emotions they used five audio features namely: mel-frequency cepstral coefficients (MFCC), mel-scaled spectrogram, chromagram, spectral contrast feature, and Tonnetz representation. Han et al.[18] use a deep neural network to show that it is very promising to use it for emotion recognition, boasting an 20% improvement over other approaches. In the work by Lalitha et al.[19] MFCCs were used as prominent features that resulted in 80% accuracy rate. In the work by Badshah et al.[20] spectograms were fed into deep convolutional neural networks (CNN) for emotion recognition . Their results are satisfactory for most emotions except fear when using a newly-trained CNN model. Inspiring from this, and considering the role of emotions like anger, fear and sadness in agitated speech detection, we will be feeding three prevalent features Mel Frequency Cepstral Coefficients (MFCCs), Mel-spectrogram, and Chromagram to our neural network.

To decide which type of neural network to use, three similar studies were considered. Yao et al.[21] show that a fusion between three learning-based classifiers yields a higher accuracy rate then the same three individually. They tested a deep neural network (DNN), a convolutional neural network (CNN) and a recurrent neural network (RNN). Each of the classifiers had different input features, in order to get the most out of the data. Mel-spectrograms were used by the CNN, the RNN used low-level descriptors and the DNN was given high-level statistical functions. Combined, their accuracy was significantly higher than the accuracy of each individual classifier (a three to eight percent increase). Similarly, Huang et al.[22] also propose a combined model between a RNN model and a CNN model. Their model, however, also takes non-verbal sounds into consideration. As with Yao et al[21], they find that a RNN model outperforms a CNN model, but the combination of these two models outperforms its components with an accuracy increase of 8%. Lastly, Bhowmik et al[23] compare a CNN model with multiple RNN models, adding their own variation of a RNN model. They find that the CNN model underperforms with respect to the RNN models. Their proposed model (Ultra Long Short Term Memory RNN) has the highest accuracy with 90%. Therefore we will be using RNNs with different configuration of layers.

3. METHOD

3.1 Gathering and preparing the dataset

Three open source datasets were acquired. They are, Ryerson Audio-Visual Database of Emotional Speech and Song dataset (RAVDESS)[1], Toronto Emotional Speech Set (TESS)[2] and ElderReact dataset[3].

The RAVDESS dataset consists of 7000+ files containing the voices of 24 different male and female actors (aged 21 to 40 years). Of these 7000 files, around 2500 are audio files further divided into speech and songs. In total, there are 1400 speech-only files which can be used for this project. Each of these files have a different emotion and intensity. These emotions include anger, fear, disgust, happiness, surprise, sadness, calmness and a neutral emotion. For this paper, only anger, fear, disgust, sadness and neutral are used. Happiness and sadness are dropped, and calmness is added to the neutral category. This was done to make the data set more representative of agitation.

TESS is a female-only dataset containing 2800 audio clips from two actresses (aged 28 and 64 years). The dataset consist of seven different emotions (anger, disgust, fear, happiness, pleasant surprise, sadness and neutral) each emotion having 200 words. Similarly to RAVDESS dataset, happiness and surprise were dropped and they were combined in order to increase the size of the training data.

The ElderReact dataset was obtained by using clips from the REACT YouTube channel. It consists of 1323 video clips of 46 elderly people and annotated these with six emotions (anger, disgust, fear, happiness, sadness and surprise). In this dataset one clip to can have multiple emotions. To make this set compatible with the other sets, they amount needed to be brought down to one. This was done by using taking the most TODO: explain why certain emotions were picked

Finally, three test-sets were created from these three datasets. One used the RAVDESS dataset[1] combined with the TESS dataset[2], as these were very similar (Test-set 1). This set consists of 2864 samples. Secondly, the RAVDESS dataset[1] was used in combinations with half of the TESS dataset[2]. The dataset was split by extracting all files related to the 64 year old actress, using their clips as validation to get as close a possible to an accurate representation of an elderly person (Test-set 2). Lastly, the ElderReact dataset[3] was used (Test-set 3), which consists of 1323 clips.

3.2 Neural Network

3.2.1 Features Extraction

For training and testing the neural network, Mel Frequency Cepstral Coefficients (MFCCs) and Mel-spectrograms were extracted from the sound clips. This was done by using librosa[24], a Python-based audio and music analysis library.

3.2.2 Parameters variation in RNN

The creation of the Recurrent Neural Network is done in Python[25] using Keras[26]. Six different type of RNNs were trained and tested having different combinations of Long Short Term Memory (LSTM) and Bidirectional LSTM (BDLSTM) layers. To understand LSTM layers, one first needs to understand recurrent neural networks (RNN). A RNN receives an input and generates an output. However, it also keep track of this decision and uses it when making the next decision. This way, RNNs combine two sources of input, the present and the recent past. However, this leads to a flaw; the *vanishing* or *exploding gradient problem*[27]. The RNN uses it previous decision in order to update the weight of its neurons. However, this gradient can become (vanishingly) small and in turn, the weight will get updated accordingly. This means that the updated weight has barely changed, which leads to the network not learning anything.

In order to combat this, LSTM layers were proposed. RNNs change their weight by using multiplication, which can lead to exponential results. LSTM layers, on the other hand, exchanged this multiplication for addition, dampening the vanishing problem. Bidirectional LSTM layers (BDLSTM) were also added. Normal LSTM layers preserve information from the past as those are the only input known to it. BDLSTM layers, however, preserve information from both the past and the future and thus can make more informed decisions. They do this by using two different layers, processing data in both directions. The output of this is then combined and fed to the output layer. Because of this, BDLSTM layers should reach a higher accuracy than their single direction counterparts. With these two layers, six different combinations were tested. These combinations can be found in Figure 1.

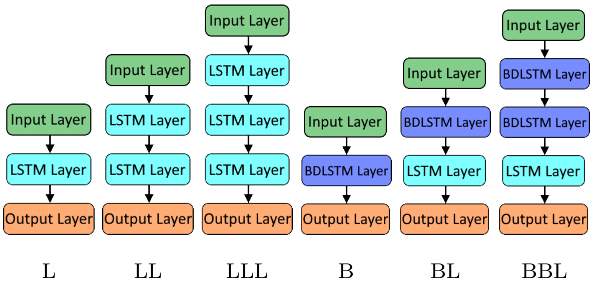


Figure 1: Layer configurations

These configurations were tested by all three training and validation sets. This means that there are also $6 * 3 = 18$ confusion matrices to compare and analyse.

A big problem of neural networks is overfitting. This is when the network corresponds too closely to the training set, making it less general. This was also a problem with this neural network. In order to prevent this, tweaks to certain parameters were made. Firstly, the learning rate of the network was changed. This resulted in comperatively better performance. Secondly, the number of *units* in the LSTM layers was modified. This altered the amount of neurons in the LSTM. As with the changes to the learning rate, it helped the performance, but did not solve the overfitting. As another prevention method, dropout layers were added. These layers help prevent overfitting by randomly setting input units to zero (dropping). The rate of this could also be changed, and has been to find the optimum.

Finally, after applying various combinations, the following values were chosen for these parameters; a learning rate of 0.001, the number of units was 16 and the dropout rate was 20%. Using these parameters, the network was then trained over 100 epochs. However, checks were in place with the purpose of detecting stagnation in learning. When it detected this, the network would stop at that epoch, resulting in the best possible result. These changes helped in reducing the overfitting problem of the algorithm on the data.

3.2.3 Training and Validating

Three training and validation sets were used. Firstly, using the RAVDESS dataset[1] in conjunction with the TESS dataset[2] (Test-set 1). 67% of the dataset was used for training, 33% was used for validating. Secondly, the RAVDESS dataset[1] in conjunction with the TESS dataset[2] was used as the training set (Test-set 2), while one of the actors from the TESS dataset was used as validation. This was done as this particular actor is close in age to the target demographic. At first, this was supposed to be the ElderReact dataset, however, the results of this were not adequate enough. Consequently, this decision was made. Lastly, using only the ElderReact dataset[3]. Again, 67% was used for training, leaving the remaining 33% to be used for validating (Test-set 3).

4. RESULTS

The results of each training and validation session were compared to each others for drawing conclusions on the best configurations as compared to state-of-the-art neural networks.

4.1 Confusion Matrices

The Confusion Matrices can be found in Appendix A.1. They show the guesses of the neural network against the actual classes. The columns represent the guesses, while the rows denote the actual classes. The 'Total' column shows the number of instances of that specific class. The 'Percent' column shows the percentage of correct guesses in that class. Similarly, the 'Percent' row denotes the percentage of correct guesses out of all the guesses of that class. The colour of a cell represents the amount of guesses where a deeper colour means more guesses. This is done so one can easily spot anomalies.

As said before, three test-sets were used. One, using only the RAVDESS dataset combined with the TESS dataset (found in Appendix A.1.1, called Test-set 1). Two, using the RAVDESS dataset for training and the TESS dataset for validation (found in Appendix A.1.2, called Test-set 2). And three, using only the ElderReact dataset (found in Appendix A.1.3, called Test-set 3).

Each of these three categories have six different confusion matrices, one for each configuration. Each configuration has an abbreviation (which can be found in Figure 1). This abbreviation can also be found in the matrices themselves.

The results obtained from three different test-sets are as follows:

Test-set 1 and configuration L: By using one LSTM layer, an accuracy of 84% is achieved.

Test-set 1 and configuration LL: By using two LSTM layers, an accuracy of 83% is achieved.

Test-set 1 and configuration LLL: By using three LSTM layers, an accuracy of 83% is achieved.

Test-set 1 and configuration B: By using one BDLSTM layer, an accuracy of 86% is achieved.

Test-set 1 and configuration BL: By using one BDLSTM layer and one LSTM layer, an accuracy of 86% is achieved.
Test-set 1 and configuration BBL: By using two BDLSTM layers and one LSTM layer, an accuracy of 86% is achieved.

Test-set 2 and configuration L: By using one LSTM layer, an accuracy of 56% is achieved.
Test-set 2 and configuration LL: By using two LSTM layers, an accuracy of 63% is achieved.
Test-set 2 and configuration LLL: By using three LSTM layers, an accuracy of 71% is achieved.
Test-set 2 and configuration B: By using one BDLSTM layer, an accuracy of 60% is achieved.
Test-set 2 and configuration BL: By using one BDLSTM layer and one LSTM layer, an accuracy of 64% is achieved.
Test-set 2 and configuration BBL: By using two BDLSTM layers and one LSTM layer, an accuracy of 76% is achieved.

Test-set 3 and configuration L: By using one LSTM layer, an accuracy of 28% is achieved.
Test-set 3 and configuration LL: By using two LSTM layers, an accuracy of 32% is achieved.
Test-set 3 and configuration LLL: By using three LSTM layers, an accuracy of 37% is achieved.
Test-set 3 and configuration B: By using one BDLSTM layer, an accuracy of 26% is achieved.
Test-set 3 and configuration BL: By using one BDLSTM layer and one LSTM layer, an accuracy of 28% is achieved.
Test-set 3 and configuration BBL: By using two BDLSTM layers and one LSTM layer, an accuracy of 31% is achieved.

Each of these results have an associated confusion matrix, which can be found in Appendix A.1. The results are summarised in Table 1. A bar-graph has also been made in order to visualise the results easier. It can be found in Appendix A.2.

It can be observed from the table 1 that test-set 1 shows average accuracy of 84.6% , test-set 2 shows average accuracy of 65.0% , and test-set 3 shows average accuracy of 30.3% .

	Test-set 1	Test-set 2	Test-set 3	Average
L	84%	56%	28%	56.0%
LL	83%	63%	32%	59.3%
LLL	83%	71%	37%	63.6%
B	86%	60%	26%	57.3%
BL	86%	64%	28%	59.3%
BBL	86%	76%	31%	64.3%
Average	84.6%	65.0%	30.3 %	

Table 1: Configuration accuracies

4.2 Comparison with State-of-the-art papers

Figure 2 shows the accuracy of this model compared to similar models from similar studies.

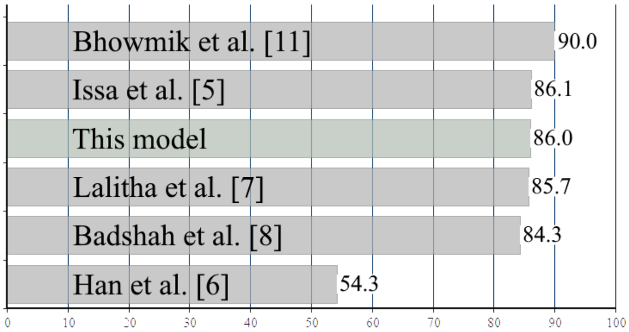


Figure 2: Comparison with previous work

5. CONCLUSION

When detecting agitation or emotion, a large amount of features can be used. Here, Mel-frequency Cepstral Coefficients, Mel Spectrogram were used together. The combination of these features resulted in an acceptable RNN. We tested six configurations (see Figure 1), and their results can be seen in Section 4. From these, it can be gathered that, when using a combination of the RAVDESS dataset[1] and the TESS dataset[2], a promising accuracy of 86% can be achieved. When looking at the other two test-sets, one can conclude that adding more LSTM layers improves the accuracy. Similarly, using multiple Bidirectional LSTM layers improves the accuracy beyond that. Consequently, it can be concluded that a Bidirectional LSTM outperforms its single-direction counterpart. The model proposed is in lines with the current state-of-the-art, outperforming some, and thus, could be used as agitation detection.

A drastic change in accuracy was seen when comparing the test datasets. It shows the need of better agitation related datasets for predicting agitation related emotions/behaviour clearly. Furthermore, from these results we can comment on the importance of these features but to improve the accuracy more, more features can be added. Also, only a RNN was used, but literature suggest that a combination of a RNN model with a CNN model yielded the highest results. Due to time constraints, only a LSTM-RNN was used in this paper. More research needs to be done in order to implement a CNN-RNN model.

6. LIMITATIONS AND FUTURE WORK

The datasets used for this project were suitable, but could be better. The ElderReact dataset[3] had a large amount of background noise due to the nature of the REACT channel. The RAVDESS dataset[1], on the other hand, was clear. However, it uses actors aged between 20 and 40 years which is not the target demographic for this paper. Similarly, the amount of trainable data was on the low-end. Future work could use or create better suited dataset, such as using recording from nursing homes.

This paper used MFCCs and mel-frequency spectrograms as trainable features, but there are more which can also be used. These include features such as the energy, the zero-crossing rate[14] and the cepstrum[19]. These are also used in speech and emotion recognition and could prove useful in future work.

The configuration of the network was explored in this paper. However, there is still more work to be done in this regard. One can make a deeper configuration, or use different layers. For instance, as the literature study showed, a CCN paired with a RNN could prove fruitful.

The neural network model was chosen using a literature study. However, better models are already known. Due to time constraints, these could not be implemented and as such, they pose a good starting-off point for further research.

In this paper, the only input used was audio, this was chosen because it is non-intrusive. This works fine for some kinds of agitation, but agitation can also present itself in the form of physical acts like aggression or constant walking[6]. As these activities make almost no sound, this solution cannot pick up on them.

Similarly, he ElderReact dataset[3] consists of video files. Using this dataset for audio-only purposes lead to a loss of information. Taking visual data into account should prove useful with this dataset. As such, further research into agitation detection using visual data could be an interesting direction.

7. ACKNOWLEDGEMENTS

I would like to thank my supervisor Nikita Sharma for her invaluable feedback and help regarding this thesis. I would also like to thank K. Ma, X. Wang, X. Yang, M. Zhang, J.M. Girard and L.-P. Morency for granting access to the ElderReact dataset[3].

8. REFERENCES

[1] S. R. Livingstone and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)," Apr. 2018. Funding Information Natural Sciences and Engineering Research Council of Canada: 2012-341583 Hear the world research chair in music and emotional speech from Phonak.

[2] M. K. Pichora-Fuller and K. Dupuis, "Toronto emotional speech set (TESS)," 2020.

[3] K. Ma, X. Wang, X. Yang, M. Zhang, J. M. Girard, and L.-P. Morency, "Elderreact: A multimodal dataset for recognizing emotional response in aging adults," in *2019 International Conference on Multimodal Interaction*, ICMI '19, (New York, NY, USA), pp. 349–357, Association for Computing Machinery, 2019.

[4] J. Cohen-Mansfield, "Conceptualization of agitation: Results based on the cohen-mansfield agitation inventory and the agitation behavior mapping instrument," *International Psychogeriatrics*, vol. 8, no. S3, p. 309âĂŞ315, 1997.

[5] A. Bankole, M. Anderson, T. Smith-Jackson, A. Knight, K. Oh, J. Brantley, A. Barth, and J. Lach, "Validation of noninvasive body sensor network technology in the detection of agitation in dementia," *American Journal of Alzheimer's Disease & Other Dementias*, vol. 27, no. 5, pp. 346–354, 2012. PMID: 22815084.

[6] J. Cohen-Mansfield, J. Marx, M. S., and A. S. Rosenthal, "A description of agitation in a nursing home.," *Journal of Gerontology: Medical Sciences*, vol. 44, no. 3, pp. M77 – M84, 1989.

[7] A. Hurley, L. Volicer, L. Camberg, J. Ashley, P. Woods, G. Odenheimer, W. Ooi, K. McIntyre, and E. Mahoney, "Measurement of observed agitation in patients with dementia of the alzheimer type," vol. 5, pp. 117–132, 01 1999.

[8] L. Valembois, C. Oasi, S. Pariel, W. Jarzebowski, C. Lafuente-Lafuente, and J. Belmin, "Wrist actigraphy: A simple way to record motor activity in elderly patients with dementia and apathy or aberrant motor behavior," *The journal of nutrition, health and aging*, vol. 19, 05 2015.

[9] S. S. Khan, B. Ye, B. Taati, and A. Mihailidis, "Detecting agitation and aggression in people with dementia using sensorsâĂŤa systematic review," *Alzheimer's and Dementia*, vol. 14, no. 6, pp. 824 – 832, 2018.

[10] S. S. Stevens, J. Volkmann, and E. B. Newman, "A scale for the measurement of the psychological magnitude pitch," *The Journal of the Acoustical Society of America*, vol. 8, no. 3, pp. 185–190, 1937.

[11] P. Mermelstein, "Distance measures for speech recognition, psychological and instrumental," *Pattern Recognition and Artificial Intelligence*, pp. 374–388, 1976.

[12] M. E. Ayadi], M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572 – 587, 2011.

[13] A. Shah, M. Kattel, A. Nepal, and D. Shrestha, "Chroma feature extraction," 01 2019.

[14] R. Bachu, S. Kopparthi, B. Adapa, and B. Barkana, "Voiced/unvoiced decision for speech signals based on zero-crossing rate and energy," in *Advanced Techniques in Computing Sciences and Software Engineering* (K. Elleithy, ed.), (Dordrecht), pp. 279–282, Springer Netherlands, 2010.

[15] D. Issa, M. F. Demirci], and A. Yazici, "Speech emotion recognition with deep convolutional neural networks," *Biomedical Signal Processing and Control*, vol. 59, p. 101894, 2020.

[16] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of german emotional speech," vol. 5, pp. 1517–1520, 01 2005.

[17] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower Provost, S. Kim, J. Chang, S. Lee, and S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, pp. 335–359, 12 2008.

[18] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," 09 2014.

[19] S. Lalitha, D. Geyasruti, R. Narayanan, and S. M, "Emotion detection using mfcc and cepstrum features," *Procedia Computer Science*, vol. 70, pp. 29 – 35, 2015. Proceedings of the 4th International Conference on Eco-friendly Computing and Communication Systems.

[20] A. M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik, "Speech emotion recognition from spectrograms with deep convolutional neural network," in *2017 International Conference on Platform Technology and Service (PlatCon)*, pp. 1–5, 2017.

[21] Z. Yao, Z. Wang, W. Liu, Y. Liu, and J. Pan, "Speech emotion recognition using fusion of three multi-task learning-based classifiers: Hsf-dnn, ms-cnn and lld-rnn," *Speech Communication*, vol. 120, pp. 11 – 19, 2020.

[22] K. Huang, C. Wu, Q. Hong, M. Su, and Y. Chen, "Speech emotion recognition using deep neural network considering verbal and nonverbal speech sounds," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5866–5870, 2019.

[23] S. Bhowmik, A. Chatterjee, S. Biswas, R. Farhin, and G. Yasmin, "Speech-based emotion classification for human by introducing upgraded long short-term

memory (ulstm)," *Advances in Intelligent Systems and Computing*, vol. 1120, pp. 101–112, 2020. cited By 0.

[24] B. McFee, C. Raffel, D. Liang, D. P. W. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," 2015.

[25] G. Van Rossum and F. L. Drake, *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace, 2009.

[26] F. Chollet *et al.*, "Keras," 2015.

[27] S. Hochreiter, "The vanishing gradient problem during learning recurrent neural nets and problem solutions," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 6, pp. 107–116, 04 1998.

APPENDIX

A. RESULTS

A.1 Confusion Matrices

Please note that these are not the final results. The network is still being trained and tweaked.

A.1.1 Using only RAVDESS+TESS

Table 2: RAVDESS+TESS dataset using one LSTM layer

L	Neutral	Sad	Angry	Fearful	Disgust	Total	Percent
Neutral	78	8	0	1	4	91	86%
Sad	2	112	3	11	7	135	83%
Angry	1	3	110	2	12	128	86%
Fearful	1	16	2	105	7	131	80%
Disgust	2	7	5	5	112	131	85%
Percent	93%	77%	92%	85%	79%	-	-

Table 3: RAVDESS+TESS dataset using two LSTM layers

LL	Neutral	Sad	Angry	Fearful	Disgust	Total	Percent
Neutral	79	7	0	3	2	91	87%
Sad	3	109	5	12	6	135	81%
Angry	0	2	107	2	17	128	84%
Fearful	0	12	2	106	11	131	81%
Disgust	1	6	3	9	112	131	85%
Percent	95%	80%	91%	80%	76%	-	-

Table 4: RAVDESS+TESS dataset using three LSTM layers

LLL	Neutral	Sad	Angry	Fearful	Disgust	Total	Percent
Neutral	83	4	0	0	4	91	91%
Sad	5	111	4	8	7	135	82%
Angry	1	1	111	2	13	128	87%
Fearful	3	13	7	96	12	131	73%
Disgust	2	3	8	5	113	131	86%
Percent	88%	84%	85%	86%	76%	-	-

Table 5: RAVDESS+TESS dataset using one BDLSTM layer

B	Neutral	Sad	Angry	Fearful	Disgust	Total	Percent
Neutral	80	6	0	0	5	91	88%
Sad	5	118	1	2	9	135	87%
Angry	1	1	109	2	15	128	85%
Fearful	0	11	4	100	16	131	76%
Disgust	2	3	2	1	123	131	94%
Percent	91%	85%	94%	95%	73%	-	-

Table 6: RAVDESS+TESS dataset using one BDLSTM layer and one LSTM layer

BL	Neutral	Sad	Angry	Fearful	Disgust	Total	Percent
Neutral	82	3	0	3	3	91	90%
Sad	3	116	2	12	2	135	86%
Angry	0	1	105	5	17	128	82%
Fearful	1	8	3	114	5	131	87%
Disgust	4	5	3	4	115	131	88%
Percent	91%	87%	93%	83%	81%	-	-

Table 7: RAVDESS+TESS dataset using two BDLSTM layers and one LSTM layer

BBL	Neutral	Sad	Angry	Fearful	Disgust	Total	Percent
Neutral	83	1	0	3	4	91	91%
Sad	6	105	2	11	11	135	78%
Angry	1	2	111	3	11	128	87%
Fearful	2	6	4	108	11	131	82%
Disgust	0	1	4	3	123	131	94%
Percent	90%	91%	92%	84%	77%	-	-

A.1.2 Using RAVDESS+TESS and validation with a TESS actress

Table 8: RAVDESS+TESS and validation with TESS actress using one LSTM layer

L	Neutral	Sad	Angry	Fearful	Disgust	Total	Percent
Neutral	171	28	0	0	0	199	86%
Sad	2	198	0	0	0	200	99%
Angry	0	62	21	51	66	200	10%
Fearful	0	21	0	96	83	200	48%
Disgust	0	95	0	28	77	200	39%
Percent	99%	49%	100%	55%	34%	-	-

Table 9: RAVDESS+TESS and validation with TESS actress using two LSTM layers

LL	Neutral	Sad	Angry	Fearful	Disgust	Total	Percent
Neutral	129	69	0	1	0	199	65%
Sad	5	188	0	0	7	200	94%
Angry	0	20	31	133	16	200	15%
Fearful	0	0	0	199	1	200	99%
Disgust	0	76	0	42	82	200	41%
Percent	96%	53%	100%	53%	77%	-	-

Table 10: RAVDESS+TESS and validation with TESS actress using three LSTM layers

LLL	Neutral	Sad	Angry	Fearful	Disgust	Total	Percent
Neutral	196	0	0	3	0	199	98%
Sad	21	101	1	0	77	200	51%
Angry	0	3	154	30	13	200	77%
Fearful	0	56	1	106	37	200	53%
Disgust	0	31	0	19	150	200	75%
Percent	90%	53%	99%	67%	54%	-	-

Table 11: RAVDESS+TESS and validation with TESS actress using one BDLSTM layer

B	Neutral	Sad	Angry	Fearful	Disgust	Total	Percent
Neutral	197	1	0	0	0	199	99%
Sad	3	187	0	0	0	200	94%
Angry	0	129	16	43	43	200	8%
Fearful	0	109	0	87	87	200	43%
Disgust	1	46	0	38	38	200	57%
Percent	98%	40%	100%	52%	81%	-	-

Table 12: RAVDESS+TESS and validation with TESS actress using one BDLSTM layer and one LSTM layer

BL	Neutral	Sad	Angry	Fearful	Disgust	Total	Percent
Neutral	194	0	0	5	0	199	87%
Sad	51	116	0	3	30	200	53%
Angry	0	11	89	99	1	200	45%
Fearful	0	53	1	145	1	200	72%
Disgust	0	60	0	40	100	200	50%
Percent	79%	48%	99%	50%	76%	-	-

Table 13: RAVDESS+TESS and validation with TESS actress using two BDLSTM layers and one LSTM layer

BBL	Neutral	Sad	Angry	Fearful	Disgust	Total	Percent
Neutral	199	0	0	0	0	199	100%
Sad	7	179	0	1	13	200	90%
Angry	0	0	63	137	0	200	32%
Fearful	0	0	22	178	0	200	89%
Disgust	1	51	0	7	141	200	70%
Percent	96%	78%	74%	55%	92%	-	-

A.1.3 Using only ElderReact

Table 14: ElderReact dataset using one LSTM layer

L	Neutral	Sad	Angry	Fearful	Disgust	Total	Percent
Neutral	1	3	27	3	6	40	3%
Sad	2	2	11	1	2	18	11%
Angry	7	8	44	6	10	75	59%
Fearful	1	0	10	1	2	14	7%
Disgust	4	3	19	3	3	32	9%
Percent	7%	12%	40%	7%	13%	-	-

Table 15: ElderReact dataset using two LSTM layers

LL	Neutral	Sad	Angry	Fearful	Disgust	Total	Percent
Neutral	3	2	27	2	6	40	7%
Sad	3	2	8	0	5	18	11%
Angry	3	10	43	5	14	75	57%
Fearful	1	0	7	3	3	14	21%
Disgust	3	2	16	4	7	32	22%
Percent	23%	12%	43%	21%	20%	-	-

Table 16: ElderReact dataset using three LSTM layers

LLL	Neutral	Sad	Angry	Fearful	Disgust	Total	Percent
Neutral	4	0	26	4	6	40	10%
Sad	3	0	10	0	5	18	0%
Angry	9	4	51	1	10	75	68%
Fearful	0	0	11	1	2	14	7%
Disgust	4	1	16	1	10	32	31%
Percent	20%	0%	45%	14%	30%	-	-

Table 17: ElderReact dataset using one BDLSTM layer

B	Neutral	Sad	Angry	Fearful	Disgust	Total	Percent
Neutral	2	3	26	3	6	40	5%
Sad	1	0	11	0	6	18	0%
Angry	7	10	36	4	18	75	48%
Fearful	1	0	10	1	2	14	7%
Disgust	5	3	15	2	7	32	22%
Percent	12%	0%	37%	10%	18%	-	-

Table 18: ElderReact dataset using one BDLSTM layer and one LSTM layer

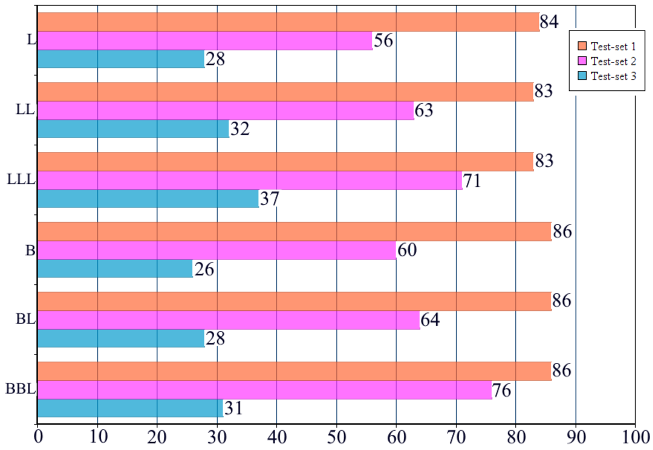
BL	Neutral	Sad	Angry	Fearful	Disgust	Total	Percent
Neutral	3	3	27	1	6	40	7%
Sad	2	1	11	0	4	18	6%
Angry	6	8	42	3	16	75	56%
Fearful	1	0	10	0	3	14	0%
Disgust	2	4	19	3	4	32	12%
Percent	21%	6%	39%	0%	12%	-	-

Table 19: ElderReact dataset using two BDLSTM layers and one LSTM layer

BBL	Neutral	Sad	Angry	Fearful	Disgust	Total	Percent
Neutral	2	2	28	3	5	40	5%
Sad	5	2	8	1	2	18	11%
Angry	4	9	48	4	10	75	64%
Fearful	1	0	12	0	1	14	0%
Disgust	4	3	20	2	3	32	9%
Percent	12%	12%	41%	0%	14%	-	-

A.2 Additional graphs

Figure 3: Configuration Accuracies


L
LL
LLL
B
BL
BBL
Test-set 1
Test-set 2
Test-set 3