Emotional Activity Detection using Behavioural Sounds

Romme Knol
University of Twente
P.O. Box 217, 7500AE Enschede
The Netherlands
romme.knol@student.utwente.nl

ABSTRACT

Emotion detection is a desirable feature for monitoring systems in healthcare. Many data sources have been examined, like facial expressions, physiological signals, and speech. A yet unused data source are sounds produced by human behaviours. When human behaviour is related to emotion, these behavioural sounds can be used for emotion detection. In this study, a convolutional neural network was trained to recognise seven different emotional behaviours, based on a newly collected data set. The classification performance was evaluated in terms of class-wise F1 scores, which ranged between 0.54 and 0.90. The study demonstrates the feasibility of using behavioural sounds for emotion detection and gives some first guidelines for implementation. In particular behaviours with regular patterns and impacts on hard surfaces lend themselves well to detection, and the model performs better when the distance between the recording device and the person is short.

Keywords

Emotion Detection, Activity Recognition, Behavioural Sounds, Deep Learning, Convolutional Neural Networks

1. INTRODUCTION

E-health is the area where healthcare and electronic systems meet [2]. It includes monitoring systems capable of estimating the emotional state of patients. Emotional wellbeing is vital in maintaining a healthy mental state and dealing with negative events [11]. A monitoring system can be used to keep track of a patient's response to their environment, condition, and treatment. This can save human resources as well as allow monitoring while leaving the patient undisturbed.

For a system to detect and classify emotions in a person, it must examine data generated by the person as a result of these emotions. Many different modalities have been examined, most notably video in the form of facial expressions, audio in the form of speech, and physiological signals [4].

One source of data which has gone unused so far is non-speech audio, which this study addresses. Certain behaviours which relate to a person's emotional state, such as pacing, finger tapping, or yawning [4] produce sounds. Such sounds will henceforth be described as behavioural sounds, and are the focus of this study, in which deep learning is used to detect these emotional behaviours through their corresponding sounds. Adding this modality to emotion detection systems is beneficial, because multimodal systems, which detect emotions based on multiple types of data, have been shown to outperform unimodal ones [12]. It is, therefore, worthwhile to examine all possible sources of information and create systems which combine them.

Behavioural sounds have advantages compared to other modalities when it comes to privacy and obtrusiveness - two closely linked issues that need to be considered in the development of emotion detection systems. Monitoring patients should not infringe on their privacy any more than necessary, and the system that performs the detection should blend into the environment as much as possible, so patients do not feel 'watched'.

These issues influence the suitability of potential modalities. For example, examining a person's facial expression necessitates processing video data of a person's face, which is a significant privacy issue, as the information it contains can be used for much more than just emotion detection. A system that analyses speech could also extract the meaning of words, another privacy issue. Physiological signals require sensor devices to be attached to the body, making such a system very obtrusive.

Compared to other modalities, the issue is of privacy is easier to deal with when using behavioural sounds, as they provide little information aside from the behaviour they stem from. The microphones required to capture the data can be small and placed out of sight, making it an unobtrusive system.

2. RESEARCH QUESTIONS

The question this study seeks to answer is the following:

How can behavioural sounds be analysed using deep learning to detect behaviours that convey information regarding a person's emotional state?

Two sub-questions have been identified, laying out concrete goals and expected results.

- *How do different behaviours and their acoustic properties affect the classification performance of the deep learning models?*
- *How does the distance between a person and the recording device affect the classification performance of the deep learning models?*

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

3. RELATED WORKS

In their survey from 2018, Jadhav and Sugandhi reviewed the current state of the art in the field of emotion detection [4]. The studies they examined cover a wide range of modalities and machine learning techniques. One of the survey's findings was that convolutional neural networks and recurrent neural networks outperform deep neural networks and deep belief networks on average.

As a contribution to elderly care, Tariq et al. implemented an IoT system that detects emotions from speech, using a 2D convolutional neural network [10]. They reported a maximum accuracy of *95%*.

Jain et al. proposed a hybrid convolution-recurrent neural network method for facial expression recognition [5]. The model was tested on the JAFFE [7] and MMI datasets, where it performed competitively with other state-of-the-art methods, with *94.91%* and *92.07%* accuracy respectively.

To the author's best knowledge, emotion detection using behavioural sounds has not been attempted. However, non-speech audio has been used to detect other things with success.

Mendoza et al. developed a wireless sensor network to detect audio events [8]. They used convolutional neural networks and achieved an accuracy of *83.79%* on the Urban8k dataset [9] by extracting constant-Q transform features as system inputs.

Jung and Chi developed a recognition model based on sound for daily indoor activities [6]. These activities, such as sleeping or showering, have a longer overall duration than the behaviours examined in this study, which is reflected in the respective length of the used audio fragments in each study. Jung and Chi used fragments of 10 seconds, while this study uses fragments of 2 seconds. Jung et al. used the Log Mel-filter bank energies methods for feature extraction and trained a residual neural network with 34 convolutional layers, achieving an *87.6%* accuracy.

In conclusion, there is a solid foundation in the literature for research into sound-based detection of behaviours and activities. This study aims to provide the first step in applying this type of detection to the field of emotion recognition by investigating how and which emotionally informative behaviours can best be detected using deep learning.

4. DATA COLLECTION

For the study, audio data of a set of emotional behaviours were collected, as there was no suitable pre-existing data set. With the goal of an application in healthcare in mind, the selection of behaviours was based on literature regarding agitation in dementia patients [1][3]. Seven behaviours were selected for study. Listed below are the behaviours' labels and descriptions.

1. Shifting - Repeatedly shifting in a chair
2. Chair - Repeatedly getting up from and sitting down in a chair
3. Fingers - Tapping fingers on a surface
4. Feet - Tapping feet on the ground
5. Pacing - Pacing back and forth
6. Palm - Slamming one's palm on a surface
7. Door - Slamming a door shut

Participants	Age Range	Female/Male
11	16-76	2/9

Table 1: Dataset Metadata

Of the eleven participants, four participated in supervised data collection. The other seven participants received detailed instructions to collect the data in their own home environments. The instructions specified which behaviours to perform and how to record them.

Participants were asked to perform each behaviour for a duration of two and a half minutes. Participants received guidelines explaining how each behaviour should be performed. However, details such as the force with which the participant slammed a door, or the exact speed at which they paced were left up to the participants, to approximate a behaviour that felt natural to them. In addition to the seven behaviours, participants were asked not to move or speak for two and a half minutes, to obtain a baseline of the noise in the environment.

For each of the behaviours and the baseline, two simultaneous recordings were made. One with a recording device next to the participant, and the other with another recording device at a two-metre distance. The near and far recordings each serve as a separate dataset. Participants used different smartphones, tablets and pc's for the sound and video recordings.

All data collection was video recorded, so that proper adherence to the instructions could be verified. Metadata about the data set is shown in table 1.

4.1 Data Pre-Processing

After reviewing the video recordings for adherence to the instructions, the audio files were subsequently processed into uniform shape. Since the data was collected on varying mobile phones, the collected data was spread across multiple file types (mp3, m4a, wav) and number of channels (stereo, mono). All files were converted to mono-channel wav files. The files were then split into fragments of two seconds long. This duration was chosen so that most fragments of the non-continuous behaviours - door slamming and palm slamming - contain only one instance of that behaviour. As a result, the near and far data sets each consist of 6600 audio fragments, spread evenly across the seven behaviours and the baseline.

5. METHODOLOGY

Using the PyTorch library, two convolutional neural networks were trained on the spectrogram images of the audio fragments. Section 5.1 explains the spectrogram images, section 5.2 describes the CNN architecture, and section 5.3 covers the training process and its parameters.

5.1 Data Representation

Each two-second audio fragment was converted to a spectrogram of featuring time on the x-axis and frequency on the y-axis, and intensity being represented through colour. The images, initially having a pixel resolution of 515x389, were resized to squares. Four different image resolutions were considered: 64x64, 128x128, 256x256, and 512x512. A comparison in terms of F1 scores (averaged across the classes) and training duration was performed to determine the most suitable size. Table 2 shows that models trained on images of 256x256 pixels achieved the highest F1 scores in the shortest amount of time. This resolution was used in the experiments refining the architecture and training

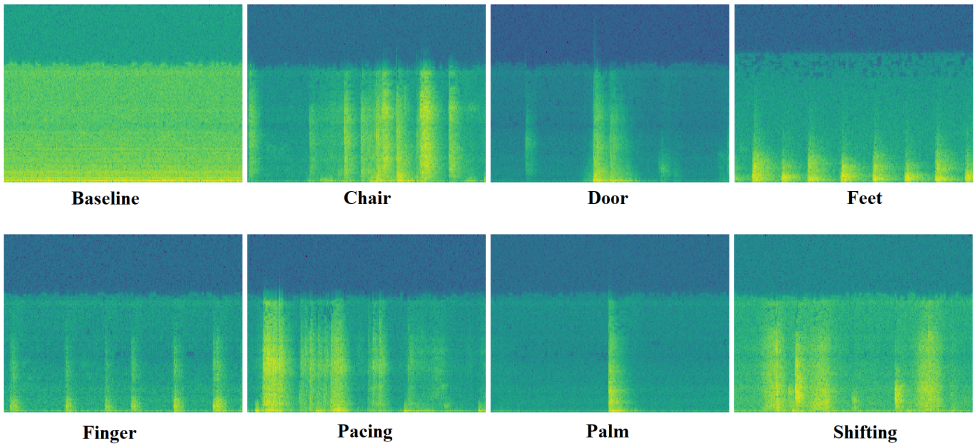


Figure 1: Spectrogram Images

Spectrogram Size	Average F1	Training Time (s)
64x64	0.65	2555
128x128	0.67	3448
256x256	0.69	3213
512x512	0.61	5420

Table 2: Comparison of Models Trained on Differently Sized Spectrograms

Class	Far	Near
Baseline	0.77	0.86
Chair	0.40	0.56
Door	0.74	0.86
Feet	0.47	0.82
Fingers	0.77	0.89
Pacing	0.61	0.69
Palm	0.84	0.87
Shifting	0.55	0.58

Table 3: Class-Wise F1 Scores on the Far and Near Datasets

parameters, described in sections 5.2 and 5.3.

A representative spectrogram from the far dataset for each of the classes is shown in Figure 1. The spectrograms from the near dataset are similar in regard to discernible patterns, differing mainly on intensity. The spectrogram images were then used to train a convolutional neural network to classify the behaviours.

5.2 Model Architecture

The initial architecture was based on the sequential CNN in the work of Mendoza et al. [8], which was also used to classify sounds based on spectrogram images. By experimenting with different variations of the architecture, in terms of the number of convolutional layers (1-4) and kernel sizes (3x3, 5x5, 7x7), the architecture was refined to achieve the maximum classification performance.

The final architecture is visualised in Figure 2. It consists of three convolutional layers with a 7x7 kernel and a stride of 1x1, each followed by a rectified linear unit and max pooling with a 5x5 kernel and a stride of 1x1. The results of the convolutional layers are subsequently fed through two linear layers, reducing the number of features to 96 in the first layer, and classifying those into the eight classes in the second layer.

5.3 Training Parameters

The full dataset was divided into a training set and a validation in an 8:2 ratio. This ratio led to higher classification performance than alternatives with a larger or smaller validation set. The contents of each set were sampled randomly from the full set (without replacement).

The optimal training parameters were obtained experimentally. The number of epochs was chosen by evaluating the training and validation loss , as well as the average of the class-wise F1 scores after each epoch. Batch sizes of 4, 8 and 16 were tested, and evaluated by the average F1 score. Learning rates of 1e-2, 1e-4, 1e-5, and 5e-5 were tested in the same manner. Because training times never became prohibitive, the impact of the varying parameters on it was not evaluated explicitly.

The best results were obtained by training the network for 40 epochs using a batch size of 8, the cross-entropy loss function, and a stochastic gradient descent optimiser with a 5e-5 learning rate and 0.9 momentum. The training was performed on an NVIDIA Geforce RTX 2080, and took approximately one hour to finish.

6. RESULTS

The classification performance of the models was evaluated in terms of class wise F1 scores and normalised confusion matrices. Table 3 shows the class-wise F1 scores for both the far and near datasets. Figures 3a and 3b show the confusion matrices for the far and near datasets.

Table 3 shows that the F1 scores diverge substantially between classes. In the far dataset, the behaviours 'chair', 'feet', and 'shifting' have low F1 scores, while 'baseline', 'door', 'fingers', and 'palm' have high scores, with 'pacing' being in between those groups.

Table 3 also shows that the F1 scores are higher for the near dataset than the far dataset for every class. For 'palm' and 'shifting' the score improvement is minor, for 'baseline', 'chair', 'door', 'fingers', and 'pacing' it is substantial, and for 'feet' it is enormous, placing 'feet' in the group of well detected classes.

The confusion matrix of the far data set (Figure 3a) shows that, the model has great difficulty in telling apart 'shifting' and 'chair'. Additionally, 'pacing' is often misclassified as 'chair'. The predictions for samples belonging to the 'feet' class vary widely, with the largest share of mis-

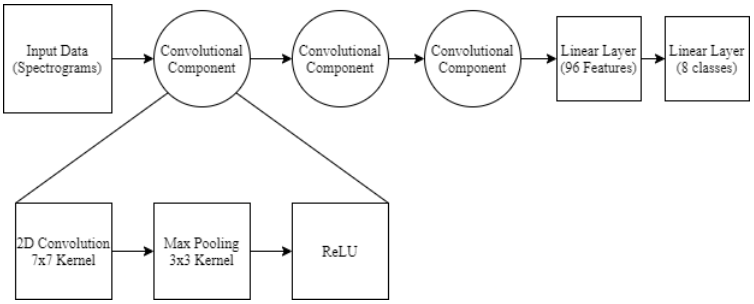
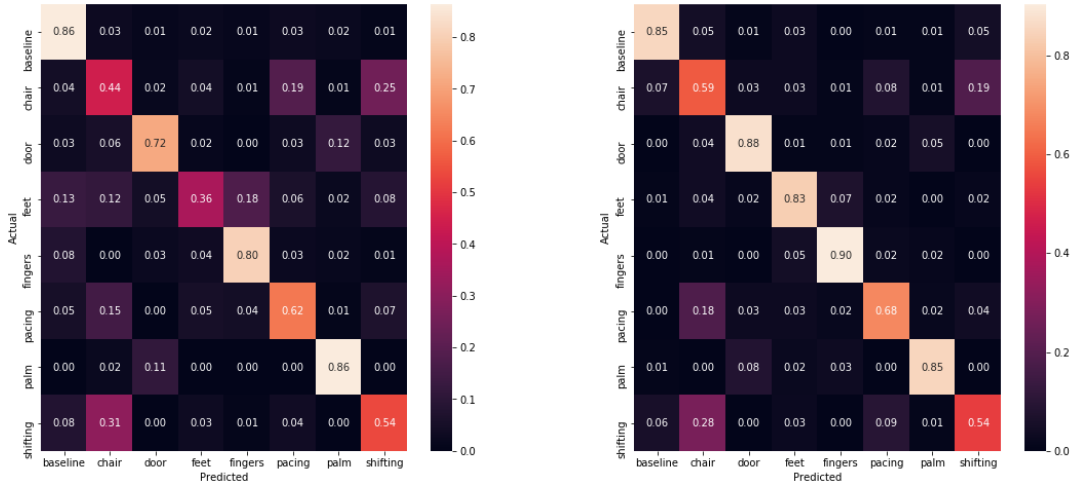


Figure 2: Model Architecture



(a) **(b)**

Figure 3: Confusion Matrices on the Far (a) and Near (b) datasets

interpretations being predicted as 'fingers'.

The confusion matrix of the near data set (Figure 3b) shows that the model still has trouble telling apart 'shifting' and 'chair', even though the recording has been done from a short distance. It still often misclassifies 'pacing' as 'chair'. However, confusion is lower across all categories, and especially 'feet' is much more accurately classified than when recording is done from farther away.

7. DISCUSSION

From the results, it can be deduced that the distance between the recording device and the person performing the behaviours has a significant effect on the classification performance of the model. This is probably because at a higher distance, the volume of the recorded audio is lower, resulting in spectrograms with less distinctive sound patterns of a behaviour due to the lack of differentiation in intensity between frequencies.

The behaviours that are most accurately detected by the model share two apparent properties. Firstly , they involve impacts on hard surfaces, often with significant force. This applies to finger tapping, pacing, door slamming and palm slamming, all of which score high in the near dataset. Secondly, they involve a regular pattern, observable in each spectrogram. This applies to finger tapping, pacing, door slamming, palm slamming, and the baseline, all of which have high F1 scores in the near dataset. The behaviours with the lowest F1 scores - 'chair' and 'shifting' - have irregular patterns and do not involve impacts on hard surfaces.

7.1 Comparison to Literature

Since this study is focused on detection of emotional behaviours from sound and did not extend to the detection of concrete emotions from these behaviours, a direct comparison of classification performance with emotion detection models is not useful. However, it can be noted that translating detected behaviours to emotions is likely to result in a loss of classification performance. Given that the accuracies of state of the art emotion detection already exceed *90%* [10][5], behaviour detection will need to be almost perfect to compensate the loss of classification performance incurred by mapping behaviours to emotions and still be competitive with other modalities.

The comparison with sound-based activity recognition, on the other hand, is more straightforward. Jung and Chi [6] report F1 scores between 0.786 and 0.937 for each of their activity classes. In this study, such high scores are achieved for only one of the behaviours on the far dataset, and four of the behaviours on the near dataset.

Mendoza et al. [8] do not explicitly report F1 scores, though they can be derived from their confusion matrix. Doing so reveals that the F1 scores range between 0.83 and 0.92 for most of the classes. Two classes have lower scores, 0.66 and 0.71. A similar disparity between the the F1 scores of classes is also found in the results of this study.

8. CONCLUSIONS

This study set out to investigate how behavioural sounds can be analysed using deep learning, to detect emotional behaviours, to explore the potential of non-speech audio as a modality for emotion detection.

The results show that the behaviours which are best detected involve impacts on hard surfaces and regular patterns, exemplified by behaviours such as finger tapping and slamming doors. Behaviours consisting of sliding motions and/or involving soft surfaces, such as shifting in a chair and repeatedly getting up and sitting down in one, do not lend themselves well to detection by their sounds.

Furthermore, the results show that the overall performance of the model is substantially influenced by the distance of the recording device to the source of the sound. For all researched behaviours, classification performance is higher when the distance is smaller. In some cases, the pattern of a behaviour may be completely undetectable beyond a certain distance.

In conclusion, this study shows the feasibility of detecting emotional behaviours using deep learning techniques, as the classification performance is in line with those of earlier sound-based activity recognition studies.

9. LIMITATIONS AND FUTURE WORK

This study is limited to detecting certain emotional behaviours from their corresponding sounds. To develop behavioural sounds as a full-fledged modality for emotion detection, the connection between behaviours and emotions (either in discrete or dimensional terms) needs to be established, so that a system may translate a detected behaviour into a prediction regarding a person's emotional state.

Future work should also seek to improve the classification performance for the behaviours, to ensure that behavioural sounds are either competitive with or a worthwhile contribution to existing modalities. Investigating more complex CNN architectures, other neural network architectures, and data representations would be a logical step in this direction.

This study did not test detection in a real-world setting. In such a setting, multiple people may be producing sounds simultaneously, strong background noise may be present, and behaviours may not be performed exactly as expected. The extent to which such factors hinder proper detection will have to be investigated so that they may be mitigated.

10. REFERENCES

[1] J. Cohen-Mansfield, M. S. Marx, and A. S. Rosenthal. A description of agitation in a nursing home. *Journals of Gerontology*, 44(3), 1989.

[2] V. Della Mea. What is e-health (2): The death of telemedicine?, 6 2001.

[3] A. C. Hurley, L. Volicer, L. Camberg, J. Ashley, P. Woods, G. Odenheimer, W. L. Ooi, K. McIntyre, and E. Mahoney. Measurement of observed agitation in patients with dementia of the Alzheimer type. *Journal of Mental Health and Aging*, 5(2):117–132, 1999.

[4] N. Jadhav and R. Sugandhi. Survey on Human Behavior Recognition Using Affective Computing. In *Proceedings - 2018 IEEE Global Conference on Wireless Computing and Networking, GCWCN 2018*, pages 98–103. Institute of Electrical and Electronics Engineers Inc., 3 2018.

[5] N. Jain, S. Kumar, A. Kumar, P. Shamsolmoali, and M. Zareapoor. Hybrid deep neural networks for face emotion recognition. *Pattern Recognition Letters*, 115:101–106, 11 2018.

[6] M. Jung and S. Chi. Human activity classification based on sound recognition and residual convolutional neural network. *Automation in Construction*, 114:103177, 6 2020.

[7] M. Lyons, J. Budynek, and S. Akamatsu. Automatic classification of single facial images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(12):1357–1362, 1999.

[8] J. M. Mendoza, V. Tan, V. Fuentes, G. Perez, and N. M. Tiglao. Audio event detection using wireless sensor networks based on deep learning. In *Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, LNICST*, volume 264, pages 105–115. Springer Verlag, 10 2019.

[9] J. Salamon, C. Jacoby, and J. P. Bello. A dataset and taxonomy for urban sound research. In *MM 2014 - Proceedings of the 2014 ACM Conference on Multimedia*, pages 1041–1044, New York, New York, USA, 11 2014. Association for Computing Machinery, Inc.

[10] Z. Tariq, S. K. Shah, and Y. Lee. Speech Emotion Detection using IoT based Deep Learning for Health Care. In *Proceedings - 2019 IEEE International Conference on Big Data, Big Data 2019*, pages 4191–4196. Institute of Electrical and Electronics Engineers Inc., 12 2019.

[11] S. Tivatansakul, M. Ohkura, S. Puangpontip, and T. Achalakul. Emotional healthcare system: Emotion detection by facial expressions using Japanese database. In *2014 6th Computer Science and Electronic Engineering Conference, CEEC 2014 - Conference Proceedings*, pages 41–46. Institute of Electrical and Electronics Engineers Inc., 11 2014.

[12] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou. End-to-End Multimodal Emotion Recognition Using Deep Neural Networks. *IEEE Journal on Selected Topics in Signal Processing*, 11(8):1301–1309, 12 2017.