# Speech Emotion Recognition using Convolutional Neural Network for Health Care

Breno van Tricht
The Hague University of Applied Sciences
The Hague, Netherlands
b.a.vantricht@student.hhs.nl

Koen de Bruijn
The Hague University of Applied Sciences
The Hague, Netherlands
k.debruijn@student.hhs.nl

Jaap van Gestel
The Hague University of Applied Sciences
The Hague, Netherlands
j.h.a.vangestel@student.hhs.nl

Julian Groen
The Hague University of Applied Sciences
The Hague, Netherlands
j.g.groen-1@student.hhs.nl

Yuri Lamijo
The Hague University of Applied Sciences
The Hague, Netherlands
y.a.lamijo@student.hhs.nl

Zahir Bholai
The Hague University of Applied Sciences
The Hague, Netherlands
z.r.bholai@student.hhs.nl

***Abstract*— An increasingly ageing population taxes the healthcare institutions in the Netherlands [1]. In order to allow people with dementia to live longer independently and alleviate these institutions, the Smart Teddy project by Dr Hani Al-Ers aims to remotely detect the Quality of Life (QoL) of senior citizens with dementia. One of the indicators the Smart Teddy uses in determining QoL is the utterances of emotions. Specifically, the emotions anger, sad, neutral, and happy, which are the emotions that are most often still expressed by people with dementia [2], are looked at. This paper proposes a system to detect these four classes of emotions using a Convolutional Neural Network which it can with a precision of 80% using publicly available and independently reviewed datasets.**

***Keywords—Emotion recognition, Speech, Convolutional Neural Network (CNN), Household ambient sounds, Dementia, Health Care, Smart Teddy***

## I. Introduction

Speech emotion recognition has gained much attention in the last decade. Researchers [3] [4] have proven that machine learning models and deep learning techniques can recognize emotions from speech. However, these models and techniques have not yet been applied to help people with dementia.

One approach to helping people with dementia is a patient monitoring system that utilizes a Speech Emotion Detector (SED). The SED sends the detected emotion to the patient monitoring system which visualizes the information for clinicians and caregivers. This information gives indications about the quality of life or health-related issues of the patient. Additionally, it can also be beneficial for the patients themselves and improve their quality of life. Clinicians and caregivers using this system can maintain an overview of their patients' health to intervene more effectively if necessary.

In [3], the authors proposed a solution for emotion detection using speech signals. They utilized the RAVDESS [5] dataset to train their model. Their proposed model is a 2D convolutional neural network (CNN). This model reported, with unaltered audio, an accuracy of 78% for female audio and 71% for male audio in detecting the emotions angry, calm, disgust, fearful, happy, sad and surprised. They improved these results by normalizing the speech signals with the European Broadcasting Union Standard R128 Normalization (EBU), Root Mean Square (RMS) and Peak normalization techniques. Which resulted in an 85% accuracy score for female signals and 81% for male signals. Furthermore, by performing data augmentation the results improved to accuracies of 95% for female signals and 93% for male signals. They developed this model for an Internet of Things (IoT) system that can be used in the health care sector, specifically aimed at elderly people in nursing homes.

The authors of [4], proposed another solution for emotion detection using speech audio. They proposed a Support Vector Machine (SVM). They also utilized the RAVDESS dataset to train their model. In this paper, they have used Mel-frequency Cepstral Coefficients (MFCCs) and the zero-crossing (ZCR) features. The final SVM model resulted in a test accuracy of 77%. Adding ambient sounds of people talking, walking by (footsteps), and birds chirping to the audio resulted in a test accuracy of 64%. They developed this model to classify emotions for children that have autism.

From related research [3], convolutional neural networks are considered a promising deep learning technique for classifying emotions from speech. A possible challenge with detecting emotions from audio data is the environment in which the audio was produced. In real-world scenarios, audio is often produced in conjunction with ambient sounds. A vacuum cleaner, a dishwasher or the phone ringing can influence the results of a prediction. This paper proposes a solution for speech emotion detection using Convolutional Neural Networks (CNN) to assist healthcare employees with improving and monitoring the quality of life of dementia patients.

Due to the unavailability of speech audio data of dementia patients, this paper utilizes a dataset that consists of 4 different datasets where professional actors and actresses recorded sentences expressing the emotions anger, disgust, fear, happy, neutral, sad, calm and surprise.

The sections of the paper are structured as follows. Section II explains the context of dementia and the research project. Section III describes the systems architecture of the convolutional neural network. It also describes the datasets and emotions used in this paper. Section IV reports the results of the convolutional neural network. Section V concludes the paper and offers discussion points. Section VI contains future work.

## II. BACKGROUND

*Dementia is the loss of cognitive functioning — thinking, remembering, and reasoning — to such an extent that it interferes with a person's daily life and activities. Some people with dementia cannot control their emotions, and their personalities may change. Dementia ranges in severity from the mildest stage, when it is just beginning to affect a person's functioning, to the most severe stage, when the person must depend completely on others for basic activities of living* [6].

Although senior citizens in the late stages of dementia certainly require constant supervision, people in the early stages of the disease are perfectly capable of living independently at home [7]. Currently, the people with the mildest stages are often placed unnecessarily in nursing homes or other supervised facilities [8]. A possible reason for this is because it is unclear whether the senior citizen can live independently at home without supervision.

Dr Hani Al-Ers has commissioned research for his project "The Smart Teddy" to help improve the detection of quality of life of senior citizens with dementia. Smart Teddy is a therapeutic companion located in the homes of senior citizens within the early stages of dementia [9]. It will be used to monitor the quality of life without having a full-time caregiver checking up on the senior. Like a real pet, the Smart Teddy observes the senior with various sensors, which combined with the software in the base station produces reports. These reports will be, in part, based on audio data which is detected in the senior citizens' homes. The detected audio will be put through trained supervised learning models that recognize eating/drinking sounds, recognize which person is speaking in a conversation and classify certain emotions. A summary of these recognitions will be shown in a dashboard for caregivers and/or family members [10]. The goal of the Smart Teddy is extending the time that the seniors can live independently in their house by remotely monitoring their quality of life.

## III. METHODOLOGY

### A. Datasets

The following four datasets have been used in this research: Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [5], Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D) [11], Toronto emotional speech set (TESS) [12] and Surrey Audio-Visual Expressed Emotion (SAVEE) [13].

These datasets contain numerous audio and video samples recorded under varying circumstances. This paper focuses on detecting emotion from speech. Resultingly, only audio samples were used for the experiments. Table 1 summarizes the information on the datasets. These specific datasets are commonly used in emotion recognition models and have been reviewed for quality by parties that are not involved in the creation of the datasets nor have any financial interest in them. [5] [11]

| | **RAVDESS** | **CREMA-D** | **TESS** | **SAVEE** |
|---|---|---|---|---|
| Number of samples | 7356 | 7442 | 2800 | 480 |
| Number of audio samples | 1440 | 7442 | 2800 | 480 |
| Number of actors | 24<br>(12 male, 12 female) | 91<br>(48 male, 43 female) | 2<br>(0 male, 2 female) | 4<br>(4 male, 0 female) |
| Emotions | Anger<br>Disgust<br>Fear<br>Happy<br>Neutral<br>Sad<br>Calm<br>Surprise | Anger<br>Disgust<br>Fear<br>Happy<br>Neutral<br>Sad | Anger<br>Disgust<br>Fear<br>Happy<br>Neutral<br>Sad<br>Surprise | Anger<br>Disgust<br>Fear<br>Happy<br>Neutral<br>Sad<br>Surprise |

Table 1: Datasets summarization

The RAVDESS and CREMA-D datasets contain samples spoken by both male and female actors. TESS and SAVEE contain samples spoken by respectively female and male actors. All samples are spoken in English by a variety of races and ethnicities (British, Canadian, African American, Asian, Caucasian and Hispanic). The CREMA-D dataset has 12 sentences covering 6 emotions spoken in four different emotion levels low, medium, high, and unspecified. The RAVDESS dataset has 2 sentences covering 8 emotions. TESS has 200 sentences covering 7 emotions. Lastly, the SAVEE dataset has 15 sentences per emotion excluding neutral.

Combining all four datasets resulted in one dataset with 12.162 audio samples. Excluding the emotion 'surprise', the number of samples per emotion is generally balanced well. The emotions angry, disgust, fear, happy and sad each have 1923 samples, neutral has 1895 samples and surprise has 652 samples. The duration of the audio samples from the combined dataset varies between 1 and 5 seconds.

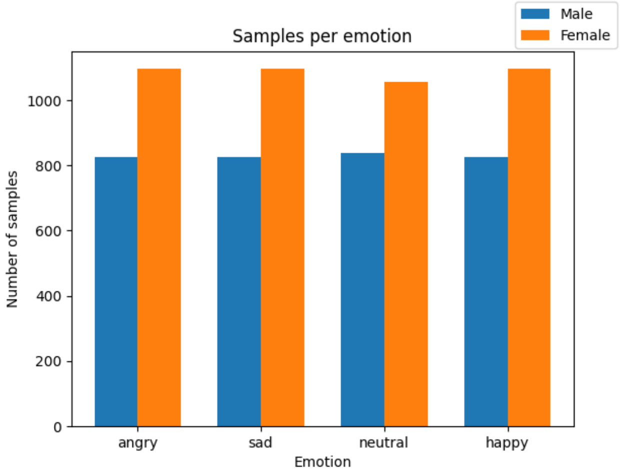

Figure 1: samples balance per emotion and gender

Not every dataset contains all emotions. Additionally, some emotions are more difficult to detect for supervised learning models than others. One of the goals of the Smart Teddy project is to detect the increase or decrease in certain emotions uttered by senior citizens over a period of time. In consultation with Dr Hani Al-Ers, this research focuses on the emotions: angry, sad, neutral, and happy. Figure 1 demonstrates the balance between these emotions. For training purposes, data has been split into 60% train data, 20% validation data and 20% test data.

## B. *Preprocessing the data*

The Python libraries librosa and Pydub are utilized for audio preprocessing. Librosa is built for music and audio analysis. It provides the ability to create visual representations of audio, can remove silence gaps and manipulate audio for data augmentation [14]. Pydub offers the ability to manipulate audio with a simple high-level interface. It is also possible to overlap audio with different audio using this library [15].

Several audio samples in the combined dataset contain silence at the beginning and the end. Using the trim function from librosa these silences are removed. The silence threshold for this is set to 20 decibels. This threshold has been chosen by following the recommendations in the documentation of librosa [14] and Audacity [16].

To reproduce the ambience of a senior citizen's home, household ambient sounds from online videos [17] [18] which will be referred to as ambient sound, were collected. This ambient sound contains a wide range of noises i.e., vacuuming, cooking, showering, washing dishes, and the sound of a washing machine. Combining both videos into a single audio file resulted in a file with a duration of 1 hour and 25 minutes. To match the maximum duration of samples from the dataset containing emotions through speech and to increase performance while augmenting data, the ambient sounds have been split into samples of 5 seconds.

To prevent the proposed model from learning one specific ambient sound, a random ambient sound has been selected out of all available ambient sound files with a duration of 5 seconds. to be overlayed with the emotion sample. Out of this randomly selected sound, a random interval of time has been sliced with the same duration as the trimmed emotion sample. The sliced ambient sound has been lowered in volume by 20 dB to match the emotion samples. Finally, these two sounds have been overlayed to create a sample of emotion speech combined with an ambient sound. Figure 2 demonstrates a waveform of an emotion sample overlayed with an ambient noise sample.

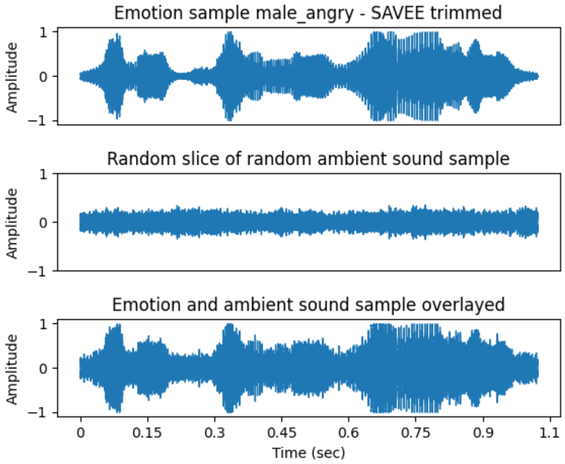

Figure 2: emotion sample – ambient noise sample - overlayed

Deep learning models are better in detecting patterns from images than raw audio files [19]. Therefore, the raw audio files were transformed with librosa into log spectrograms. Research has shown that log scaling the spectrograms can affect the results of pattern detection positively [20].

## C. *System architecture*

This paper proposes a solution to recognize emotion from speech utilizing a convolutional neural network which will be referred to as a CNN. The CNN proposed utilizes transfer learning to improve its performance. Transfer learning provides the ability to utilize a pre-trained CNN model. The pre-trained model that has been applied is the DenseNet-161 model. The authors of [21], concluded that the DenseNet-161 performs better on analyzing audio spectrograms in comparison to the Inception and ResNet pre-trained models. Figure 3 shows a general overview of the DensNet-161 model. Additionally, the proposed CNN contains a fully connected linear layer of (1000, 4), to downscale the transfer learning output of the DenseNet model to 4 chosen emotions.

The proposed CNN utilizes the hyperparameters: batch size, number of epochs, learning rate and log spectrograms image dimensions. The default values of 32 and 0.0001 were applied for the batch size and learning rate. The log spectrograms have image dimensions (256, 128). These hyperparameters were used in 50 epochs.

## D. *Evaluation method*

To evaluate the performance of the model, 'precision' was used as an evaluation metric. Precision gives insight about the number of false positives. In context of detecting emotions from senior citizens with dementia, this signifies that when a certain emotion, for example, anger is detected there is a high certainty that the citizen was indeed angry and not happy.

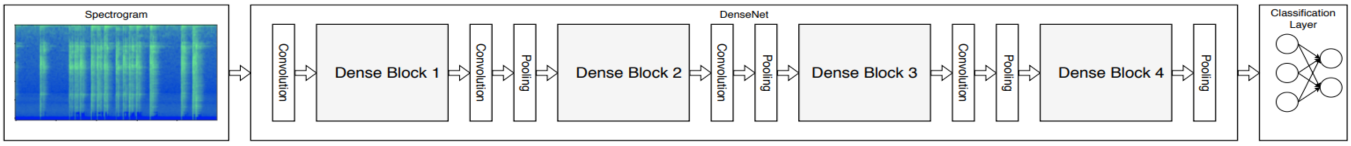

Figure 3: DenseNet Architecture

The disadvantage of using precision as an evaluation metric compared to recall is that the model will detect less emotion but is more certain about the quality of its prediction. Using recall as an evaluation metric increases the number of predictions but decreases the quality of the predictions. When classifying (mental) health issues it is especially important to be certain about the quality of predictions.

The Smart Teddy is designed to measure an increase or decrease in preselected emotions uttered by senior citizens with dementia. Training the model for recall will alter this measurement more heavily than when training for precision. After all, recall increases the quantity of predictions. Whenever a significant increase or decrease in emotions uttered is detected, medical professionals are needed to manually check-up on the citizen to confirm his or her capability to live independently at home. This check-up taxes the healthcare sector. Resultingly, a high level of certainty in a significant increase or decrease of emotions uttered is desired.

## IV. RESULTS

Two experiments have been executed on the proposed system architecture to measure the performance on the emotions: angry, sad, neutral and happy. The first experiment utilized the combined dataset of emotions without the added ambient sounds. This experiment concluded with an precision of 84% on the unbiased test set. The second experiment utilized the combined dataset of emotions with added ambient sounds. The experiment resulted in a precision of 80%. The results show that when ambient sounds are applied there is a small negative impact on the precision.

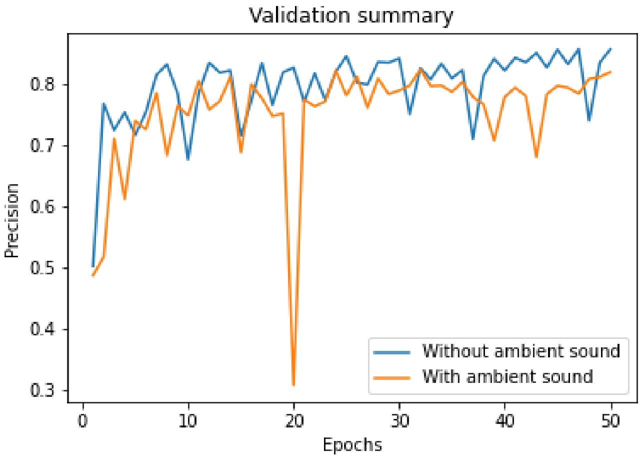

Figure 4: Validation summary experiment one and two

Figure 4 shows the learning curve of the two different experiments.

## V. CONCLUSION & DISCUSSION

The proposed system can detect the emotions angry, sad, neutral and happy with ambient noise with a precision of 80%. The system can reliably be used by medical professionals to remotely detect changes in the emotion uttered by senior citizens with dementia. This data can also be used to determine the quality of life for the senior citizen which in turn can help determine whether or not the citizen can live independently at home. The ability to remotely detect a change in the number of emotions uttered can decrease the strain on an already taxed healthcare system.

Although precision was used as an evaluation metric to prevent unnecessary check-ups, medical professionals should still be consulted regularly to determine the quality of life of a senior citizen with dementia.

The proposed system demonstrates that classifying emotion is possible with high precision using a 2D convolutional neural network (CNN). It also shows simulating a household environment can have a minimal negative impact on the precision of its predictions.

Due to the unavailability of audio data from seniors suffering from dementia, it remains uncertain how the system will perform in a real-world scenario within the home of a citizen with dementia.

The proposed system was trained by adding ambient sounds artificially. This action may also not accurately represent the situation inside a senior citizen’s home.

Cultural differences and the language spoken may impact the way of expressing emotions, which in turn could affect the detection of emotions from audio. Additionally, people express emotions in various ways. Some people are very expressive whilst others are not. Resultingly, the performance of the system might be affected.

One of the research papers mentioned in the introduction [3] showed an accuracy of 95% for females and 93% for males in detecting emotions spoken. Their methodology varies from the one used in this research in several ways. By combining four separate datasets into one, a more diverse dataset in terms of quality and quantity of samples is created in comparison with RAVDESS dataset used in [3]. Additionally, the evaluation method of [3] is based on accuracy and not precision.

## VI. FUTURE WORK

The research that has been conducted can be further extended in the future. The most promising methods are discussed in this chapter.

The number of neutral samples in the dataset could be increased to improve the precision of neutral emotions. Consequently, the dataset becomes unbalanced but there is a probability that this will improve the precision of detecting the emotions happy, angry, and sad indirectly.

Additionally, one of the ways the research could be extended, and the results can be improved is by adding more valid datasets. Initially, this paper used only the RAVDESS and CREMA-D datasets. Once the TESS and SAVEE datasets were added it increased the precision.

Currently, there are a limited number of open emotional speech datasets. In addition to the four sets that have been used in this paper, the eNTERFACE [22] dataset is a possible candidate for addition to the experiments. Other possible candidates are EMO-DB [23], DES [24] and SUSAS [25]. Some of these datasets contain samples that are not spoken in English. This could potentially adversely affect the model depending on the differences in culture, intonation etc.

The results can be further improved once the Smart Teddy project is in use in the homes of senior citizens. Once operational the model can be trained with data obtained from real-world situations.

Hyperparameter tuning can also be explored as an avenue to improve the results. Tuning the learning rate or batch size could improve the speed and results for finding the right optimum.

Finally, the research could be expanded by adding or removing emotions that are classified by the model. Medical professionals should be consulted when changing the emotions used in the classification model.

## VII. REFERENCES

[1] B. D. Y. G. E. M. M. Rechel, „How can health systems respond to population ageing?”.

[2] C. C. C. G. D. M. C. &. C. C. Magai, „Emotional Expression During Mid- to Late-Stage Dementia,” International Psychogeriatrics, 1996.

[3] S. K. S. a. Y. L. Z. Tariq, „Speech Emotion Detection using IoT based Deep Learning for Health Care,” 2019 IEEE International Conference on Big Data (Big Data), 2019.

[4] R. M. a. D. Valles, „A Speech Emotion Recognition Solution-based on Support Vector Machine for Children with Autism Spectrum Disorder to Help Identify Human Emotions,” 2020 Intermountain Engineering, Technology and Computing (IETC), 2020.

[5] S. R. Livingstone and F. A. Russo, „The Ryerson Audio-Visual Database of emotional speech and Song (RAVDESS): A Dynamic, multimodal set of facial and vocal expressions in North American English,” PLOS ONE, 2018.

[6] N. I. o. Aging, „What Is Dementia? Symptoms, Types, and Diagnosis,” [Online]. Available: https://www.nia.nih.gov/health/what-is-dementia#:~:text=What%20Is%20Dementia%3F-,Symptoms%2C%20Types%2C%20and%20Diagnosis,and%20their%20personalities%20may%20change. [Geopend 04 01 2022].

[7] dementiacarecentral.com, „Stages of Alzheimer’s & Dementia: Durations & Scales Used to Measure Progression (GDS, FAST & CDR),” 24 04 2020. [Online]. Available: https://www.dementiacarecentral.com/. [Geopend 05 01 2022].

[8] K. Yaffe, „Patient and Caregiver Characteristics and Nursing Home Placement in Patients With Dementia,” 2002.

[9] „Smart Teddy taking care of senior citizens,” The Hague University, [Online]. Available: https://www.thehagueuniversity.com/research/centre-of-expertise/projectdetails/smart-teddy-taking-care-of-senior-citizens. [Geopend 05 01 2022].

[10] T. H. U. o. A. Science, „Smart Teddy,” [Online]. Available: https://bigdata-thuas.eu/projects/smart-teddy/. [Geopend 05 01 2022].

[11] D. G. C. M. K. K. R. C. G. A. N. a. R. V. H. Cao, „Crema-d: Crowd-sourced emotional Multimodal Actors Dataset,” IEEE Transactions on Affective Computing, 2014.

[12] M. K. Pichora-Fuller en K. Dupuis, „Toronto emotional speech set (TESS),” University of Toronto Dataverse, 2020.

[13] kahlan.eps.surrey.ac.uk, „Surrey Audio-Visual Expressed Emotion (SAVEE) Database,” [Online]. Available: http://kahlan.eps.surrey.ac.uk/savee/. [Geopend 04 01 2022].

[14] librosa, „librosa — librosa 0.8.1 documentation,” [Online]. Available: https://librosa.org/doc/. [Geopend 05 01 2022].

[15] Pydub, „jiaaro/pydub @ GitHub,” [Online]. Available: https://pydub.com/. [Geopend 05 01 2022].

[16] Audacity, „Truncate Silence - Audacity Manual,” [Online]. Available: https://manual.audacityteam.org/man/truncate_silence.html. [Geopend 05 01 2022].

[17] MobiusASMR, „ASMR Washing Dishes and Household Sounds,” 24 01 2015. [Online]. Available: https://www.youtube.com/watch?v=J1l8RMR-qXM. [Geopend 05 01 2022].

[18] H. H. TV, „1 hour of HOUSEWORK SOUND ~ Vacuuming, Washing Up, Cooking, Cleaning, Washing Machine WHITE SOUND,” 17 03 2018. [Online]. Available: https://www.youtube.com/watch?v=pnETk_wlwb8. [Geopend 05 01 2022].

[19] T. E. Writer, „Audio classification using CNN-coding example,” Medium, 22 03 2019. [Online]. Available: https://medium.com/x8-the-ai-community/audio-classification-using-cnn-coding-example-f9cbd272269e. [Geopend 09 01 2022].

[20] K. F. G. C. K. S. M. Choi, „A Comparison of Audio Signal Preprocessing Methods for Deep Neural Networks on Music Tagging,” Cornell University, 2017.

[21] K. S. D. Y. A. Palanisamy, „Rethinking CNN Models for Audio Classification,” 2020.

[22] M. L. a. L. C. H. M. Fayek, „Towards real-time Speech Emotion Recognition using deep neural networks,” 2015 9th International Conference on Signal Processing and Communication Systems (ICSPCS), 2015.

[23] F. P. A. R. M. S. W. F. &. W. B. Burkhardt, „A database of German emotional speech,” Interspeech 2005, 2015.

[24] C. Liu, M. Osama en A. De Andrade, „DENS: A dataset for multi-class emotion analysis,” arXiv, 2019.

[25] J. H. L. Hansen, „SUSAS,” Linguistic Data Consortium, Philadelphia, 1999.