Comparing speech samples with a Convolutional Neural Network to detect audio with the same speaker

Björn Appehl
Amber Bezouwen
Olaf Bolleurs
Maria Hoendermis
David Hollander
Leander Loomans

I. Abstract

With a growing number of dementia patients, it becomes important to monitor indicators of a declining condition such as changes in social interaction. Better monitoring will allow caregivers to act quickly. This would let dementia patients live longer independently at home. The aim of this study was to research how data science techniques can be used to detect if there is a conversation between at least two people by analyzing audio files. In order to do this, WAV audio files were transformed into Mel Frequency Cepstral Coefficients, which were used as input for two Convolutional Neural Networks. The first neural network detects instances of speech, which it then passes on to to the second neural network. This second neural network would then determine if two instances of speech are the same speaker or not. The speech detection model has an accuracy of 89% and the speaker differentiation model has an accuracy of 94%. These models are used for the final product which gives a percentage for the time when speech is present in the audio and the amount of times the speaker changes. The results obtained by the final product could be used by experts and/or caregivers of the dementia patients to make a conclusion if a conversation has taken place.

II. Introduction

As the average age of the global population is steadily increasing, the number of dementia patients is expected to rise to 78 million in 2030 and 139 million in 2050 [1]. Allowing dementia patients to safely live in their own home for as long as possible can be beneficial for multiple reasons. For a dementia patient, moving to a new home can be extremely stressful [2] and living at home for longer can free up caregiver resources to other patients in need. A patient's dementia progression can be monitored to assess whether he or she is capable of living at home. One factor that measures the severity of dementia is the person's levels of social interaction. A change in the amount of social interactions a person with dementia has, can be an important indicator of their dementia progressing [3].

In order to take measurements of a patient's daily life, data needs to be collected. A device is being developed by researchers at The Hague University of Applied Sciences specifically for this purpose called the Smart Teddy Bear. This device gathers data for measuring a number of quality of life factors by using sensors such as microphones or cameras. This way the dementia patients will be able to retain a safe and familiar environment at home for a longer time [4]. The data from these sensors can be analyzed by applying data science techniques to assess a dementia patient's quality of life.

Data science techniques can be used on audio data in order to determine whether conversation is taking place. The scope of this study is to use data science techniques on audio data and detect when there is conversation to reliably measure levels of social interaction for dementia patients. Detecting conversations can be split up into two parts. First, a trained model is used to detect speech in audio. Second, a separate model is used to compare the voices in the audio containing speech, to determine whether the two audio parts are spoken by the same person.

This paper consists of an introduction, where the research problem and domain is introduced to the reader. The method section covers the datasets used, along with the machine learning techniques that were applied to the datasets. In the results section, the results of the machine learning models are explained, along with metrics for evaluation and the final product results. These results are interpreted in the discussion section, and the recommendation section serves to guide further research.

A. Background

What makes this study unique, is the measuring of conversation with neural networks to provide a quality of life indicator for dementia patients. This is for the purpose of prolonging independent living. Many other studies are leaning towards either diagnosing dementia by using speech recognition [5] [6] [7] [8] or are specializing in a complete smart home suite for assisted living [9]. Unlike these previously mentioned studies, the goal of this study is to provide a useful indicator for monitoring dementia, rather than diagnosing dementia. This while also using only audio data gathered from a single smart device in a household environment.

A literature study has been conducted to determine the relevance of this project in regards to previous research. One of the studies found was a study by Udin *et al.* [10] which summarizes projects that utilize different sorts of sensors for patient monitoring in elderly homes. Out of the eight studies that use sound sensors, only Vacher *et al.* [9] uses the audio data to detect daily activity. However, the study performed by Vacher *et al.* distinguishes itself by only analyzing vocal orders or utterances of distress for assistance purposes. The study by Udin *et al.* indicates that using audio data as an approach for detecting daily activities such as conversations is not very common.

B. Research problem

The primary goal for this study is the following: "How can data science techniques detect a conversation between at least two people by analyzing audio files?".

A conversation in the context of this paper is defined as speech between at least two speakers. However, using conversation as the basis of a quality of life indicator can be challenging. If the dementia patient is for instance having a conversation by using their phone, a microphone may not pick up the audio of the other person talking. Another scenario would be if the dementia patient has visitors who are all having conversation, and the dementia patient is not present. This situation constitutes a false positive in the context of this paper. There is also the aspect of background noise, which can negatively impact the accuracy of determining conversation. In order to detect speech reliably, the algorithms have to be somewhat resilient to noise, since speech might overlap with environmental sounds.

III. Methods

To simplify the problem, the decision was made to split the task into two separate models. The first model, which will recognise voice activity, is used as a filter. This model will filter out all audio samples that have no speech in it and will pass on the samples with speech to the second model. The second model will then compare two samples to determine if the samples are from the same speaker or from different speakers, a similar approach as [11]. Ultimately the two models will be combined to determine if there is a conversation present in an audio sample.

A. Dataset

During the course of this study, multiple datasets were used. See below for a quick overview of the main attributes of these datasets:

1) AVA-Speech [12]
   - 45 hours of audio extracted from movies on YouTube
   - divided into 4 labels, for distribution see Table I
   - Stereo, 44.1kHz sampling rate
2) LibriSpeech (train-clean-100) [13]
   - 100 hours of audio excerpts from 251 different speakers
   - Mono, 16kHz sampling rate
   - English speech

Label	Amount (% of time)
NO_SPEECH	47.68%
CLEAN_SPEECH	14.55%
SPEECH_WITH_MUSIC	13.46%
SPEECH_WITH_NOISE	24.32%

TABLE I: Distribution of labels in AVA-Speech

For the first model which has to recognize speech activity, the *AVA-Speech* dataset was used. A selection was made, consisting of 15,000 seconds of *NO_SPEECH*, and 5,000 seconds of each of the other labels, resulting in a new dataset of 30,000 seconds with varying data and a 1:1 ratio between 'voice' and 'no voice', or *true* and *false* respectively.

Besides this positive and negative speech audio, another dataset is required to be able to distinguish between different speakers in the second model, since the *AVA-Speech* dataset has no labels for speaker identity. For this purpose, the audiobook excerpts from the *LibriSpeech* dataset were used, which do have speaker identity labels. This way, two samples by the same speaker can be labeled as *true* and different speakers as *false*. This dataset does not contain audio without speech.

B. Testing Dataset

To test the final product, another dataset was necessary to avoid testing on data that was used to train either of the models. For the *true* samples, three unused voices from *LibriSpeech* [13] were taken to provide audio with a speaker ID as label. This audio needed to be mixed with *false* samples, such as samples not containing speech. For this purpose the *CHIME-Home* [14] dataset was used. These together resulted in 10 minutes of audio, labeled with either silence or a corresponding speaker ID.

CHIME-Home

- 1946 4-second audio excerpts
- Mix of voice and non-voice audio
- Mono, 16kHz sampling rate
- human-labeled
- audio labeled c, m or f (the voice audio) was left out

C. Data Preparation

To detect conversation, the audio needs to be transformed to more suitable data. Following similar research [11], both models use an inputstream of small intervals of audio data. Smaller intervals mean lower chances of overlapping speakers, which would complicate speaker differentiation. Both 1 and 0.5 second methods were used for comparison. However, this showed no significant difference in accuracy score, which is why 0.5 second intervals were used.

Each fragment of 0.5 second was resampled to 44.1kHz to ensure homogeneity of audio between different sources. For every 0.5 second fragment, Mel Frequency Cepstral Coefficients (MFCCs) were created to represent the audio data. Using MFCCs as input for models is conventional in the field of speech/voice recognition[15]. This process resulted in a 40x44 shaped MFCC for each half second of audio. Each sample in the dataset consists of a pairing of two MFCCs for the model to compare. Like the first model, the labels are *true* or *false* for every MFCC-pair. Here however, *true* means that the two MFCCs are by the same speaker, whilst *false* means two different speakers.

As mentioned in the III-A dataset subsection, a custom dataset had to be created for performing the speaker differentiation. For this, a custom script [16] was written to dynamically generate combinations of speakers based on the *LibriSpeech* dataset. The output of this script is a dataset where each sample has two half-second audio fragments and a label which is *true* when the samples belong to the same speaker, and *false* when they belong to two different speakers. The generated dataset consists of 50% *true* and 50% *false* samples. The dataset contains the first 100 speaker IDs from *LibriSpeech*, which means each speaker occupies 1% of the dataset. For each speaker there is an equal amount of positive and negative samples.

Both datasets are split into a 60% train, 20% validation and 20% test set [16].

D. Neural Networks

In existing work for similar tasks, Convolutional Neural Networks (CNN) are often used to classify MFCC data [17]. That is why in this study CNNs are used.

The input for training the first model is the dataset based on *AVA-speech*. Each MFCC in the set has a *true* or *false* label, indicating whether it contains voice data. The input for training the second model is the dataset based on *LibriSpeech*.

The output from the two models together predict the presence of speech, and whether it is spoken by the same or a different person. Running audio data through these two models, using the output from the first CNN as input for the second CNN, would then output the information needed to determine if the audio meets the requirements for the definition of a conversation. If the probability of speech is high, and if two or more speakers are consistently present, a conversation is likely taking place.

E. Model Architecture

For the architecture of the models inspiration was found in [11] [18] [19]. These papers make use of shrinking kernel sizes. This can not negatively impact the output of the model, because if the model cannot make use of the bigger kernel size, the weight will be set to zero. This is the reason why both models apply this shrinking kernel size concept [20].

Both models make use of a similar architecture. The difference between the two models is that the model for speaker differentiation needs two audio samples in order to make the differentiation. That is why the fc layers of the speaker differentiation model are twice as big.

Because each row of the input data represents a frequency band, the model uses 1D convolutional layers to process them. Therefore the input data is interpreted by each frequency band. The convolutional layers are being used in combination with a MaxPool. The architecture of the model can be found in Figure 1. All models are trained with the Adam optimiser [21] and a binary cross-entropy loss function.

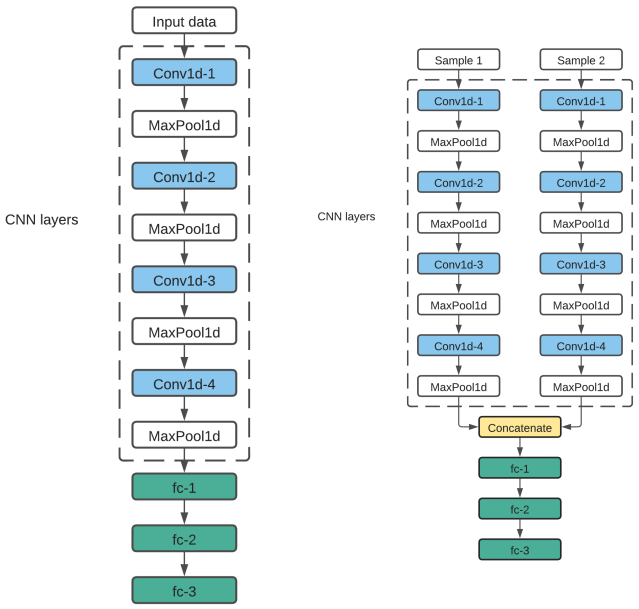

Fig. 1: Architecture of speech detection model (on the left) and speaker differentiation model (on the right)

F. Final product

The final product works by aggregating the results from the two models. First, audio data is converted to MFCCs. Second, the MFCC data is handed to the speech detection model. Third, the results of the speech detection model are then prepared to fit the speaker differentiation model. Fourth, data is passed to the speaker differentiation model and finally the results are displayed. The preparation of the data and the models has already been discussed in chapter III-C and have been put together to form the final product. This, along with the preparation of the results of the speech detection model to fit the speaker differentiation model and finally displaying the results.

The final product is able to detect voices and will return how often there is speech within the audio samples provided. The samples which contain speech are then put into the speaker differentiation model. To compare speakers, two lists are made: one with the original samples, and another list with the same samples but with their position shifted by 1. Comparing these two lists makes it possible to detect if the samples are spoken by the same person or not. When this comparison results in a false, it

will be known that the person speaking has changed. This occurrence is saved and the amount of times during the speech audio that the speaker has changed, is displayed.

IV. Results

The results show that using MFCCs in combination with a CNN will make it possible to detect speech and differentiate between different speakers. For the purpose of the study, the models focus on high precision. With higher precision, false positives will occur less often. The chance of a dementia-patient getting lonely because the algorithm overestimates their social activity, will then be reduced. Getting more false negatives is less problematic than false positives, as it will only make sure the caretakers will give too much care for the patients. This in contrast to providing not enough care.

At first, spectrograms were used to represent audio fragments. Switching to MFCCs improved accuracy for the speech detection model by 10% compared to using spectrograms, so MFCCs were used from this point on.

A. Results for Speech Detection

The results of the speech detection model are displayed in Table II. Notice that the model performs better on detecting no voice, than voice. This can be perceived when looking at the precision score. This is a desired outcome, since this way the speaker differentiation model will not receive a lot of false positive samples.

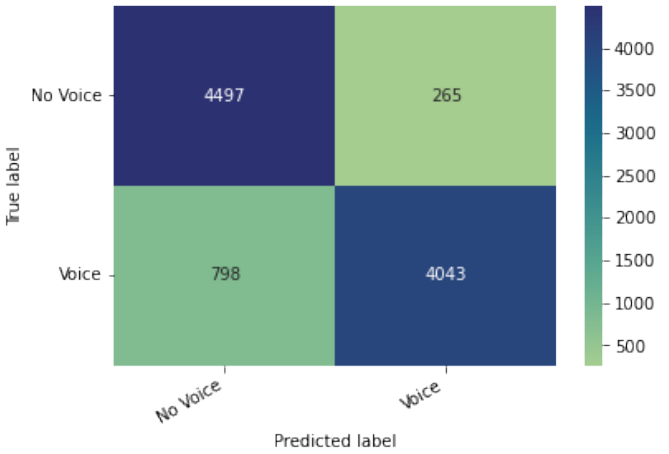

Fig. 2: Confusion matrix Speech detection on test set

	Accuracy	Precision	Recall	f1	Samples
No Voice	-	0.85	0.94	0.89	4762
Voice	-	0.94	0.84	0.88	4841
Total	0.89	-	-	-	9603

TABLE II: Results speech detection model on test set

B. Results for Speaker Differentiation

The results of the second model are displayed in Table III. The results show that it is possible to accurately predict if there is the same or different person speaking.

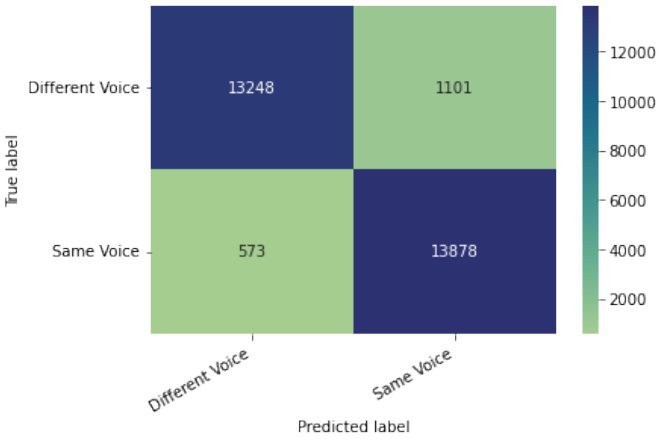

Fig. 3: Confusion matrix Speaker Differentiation on test set

	Accuracy	Precision	Recall	f1	Samples
Different Voice	-	0.96	0.92	0.94	14349
Same Voice	-	0.93	0.96	0.94	14451
Total	0.94	-	-	-	28800

TABLE III: Results Speaker Differentiation model on test set

C. Results for the Final product

As seen in figure 4, there is voice present in about 50% of the audio.

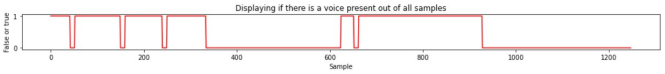

Fig. 4: Displaying when a voice is present in the audio data

The results of the speech detection model can be seen in figure 5.

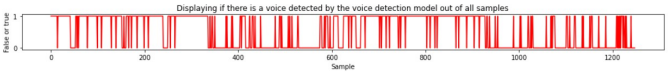

Fig. 5: Displaying when speech is detected by the speech detection model in the audio data

This was done with 87.90% accuracy and the figure 5 can be compared to figure 4. When all the speech samples are found, the speaker differentiation model will execute. The accuracy of the differentiation was 85.14% and can be seen in figure 6.

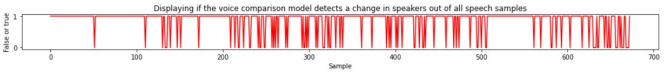

Fig. 6: Displaying when a change in voice is detected by the speaker differentiation model

Finally, the product will display the percentage of the time there was speech and the amount of times the speaker changed within the audio. This can be seen in figure 7

```
53.93% of the audio contains speech.
During this speech audio, 114 times the speaker changed.
```

Fig. 7: Displaying the results of the final product

V. Discussion

In order to detect social interaction in audio data, experiments have been conducted with two CNN models. These two models were combined in a final product. To validate the results, the accuracy score has been used. To avoid overfitting, a test set with data with different speakers than in the training set is necessary. The dataset used in this study does not contain unique speakers in the training, validation and test set. A limitation of this research can be that the model is overfitting to specific speakers.

The results indicate that detecting different voices in audio can lead to determining whether there is a conversation between people or not. When executing the final product, it will display the percentage of the time there was speech and the amount of times the speaker changed within the audio. This will be of use to the end user, which will be an expert and/or a caregiver of the dementia patient. The user of the final product will be able to conclude if there has been a conversation. The parameters for when there is a conversation are outside the scope of the project. The final product provides input for the experts to make a conclusion.

This current study can be an addition to the study by Vacher *et al.*[9] which focuses on detecting daily activity of elderly in order to record vocal orders or utterances of distress. Since this study centres around elderly dementia patients living at home alone, the results can be of great use for the Vacher study. It must be taken into account that this study only focuses on data that does not contain music or television sounds. As soon as these variations are added, the results could lead to different outcomes.

VI. Recommendation

One thing that could help improve the results of this study would be to further verify the models accuracy in more specific scenarios. One such scenario could be to tell the difference between speech spoken by people present in the room, and speech coming from a device's speakers, such as audio coming from a person speaking through a television or phone. More accurate results would be achieved by further testing on how good the current model already is at determining the difference in this type of scenario and if its accuracy is currently insufficient, improving it. Perhaps this could be done by adding another filtering model specifically trained to filter out the speech coming from device speakers.

Another possible point of improvement would be if the voice of the dementia patient could be identified. Right now if a conversation happens in the room between two people then it gets identified as a conversation even if it might not include the patient. Ensuring that the model can identify if the dementia patient participates in the conversation, would increase the accuracy of the quality of life indicator.

Both models could benefit from repeating the training process with a new dataset, consisting of audio gathered by the Smart Teddybear itself, once this data exists. Combining audio from different datasets was necessary but not ideal.

Another aspect that is worth testing further, would be to look at the outcome using different time intervals. Currently 0.5 seconds intervals are used for the inputstream but trying more different time intervals would be useful to see if it affects the results.

Finally, the implementation of the second model can be extended by comparing more samples than just consecutive ones. This can be used as a verification, or possibly even to determine an exact amount of speakers.

VII. Conclusion

The goal of this study was to answer the question: "How can data science techniques detect if there is a conversation between at least two people by analyzing audio files?" In order to answer this question, two CNNs were made that take MFCCs as input. The returning audio samples that contain speech from the first CNN model were used as input for the second CNN model. This model compares the audio samples and tells whether it is the same or a different voice. The results were promising, with both models having a high accuracy score of 89% and 94%, meaning the CNNs can reliably classify the audio data.

The final product combines the outcome of the two models and gives a percentage of how many audio samples contain speech and how often the speaker changes. So, this study shows that social interaction between people can be detected by analyzing audio data. The results of this study can be used to provide a quality of life indicator and would be suitable as a base to improve upon.

VIII. Acknowledgement

Now that the project has come to an end, we'd like to acknowledge a few people. First of all, we would like to say thank you to Dr. H. Al-Ers for guiding us throughout the project and giving feedback to keep us on the right track. We would also like to thank the teachers (Mr. T. Andrioli, Mr. R. Vermeij and Mr. J.B.P. Vuurens) for the lectures and feedback that helped us make the models and the coaching for the last 20 weeks.

References

[1] WHO. Rates of dementia. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/dementia

[2] Alzheimer's Association. Moving a person with dementia into a caregiver's home. [Online]. Available: https://www.alz.org/media/greatermissouri/moving_a_person_with_dementia_into_a_caregiver_s_home.pdf

[3] R. A. Hackett, A. Steptoe, D. Cadar, and D. Fancourt, "Social engagement before and after dementia diagnosis," *English Longitudinal Study of Ageing*, 2019.

[4] K. H. Innovation. Slimme knuffel 'smart teddy' past op ouderen. [Online]. Available: https://www.dehaagsehogeschool.nl/onderzoek/kenniscentra/projectdetails/slimme-knuffel-smart-teddy-past-op-ouderen
[5] B. Mirheidari, D. Blackburn, T. Walker, A. Venneri, M. Reuber, and H. Christensen, "Detecting signs of dementia using word vector representations," September 2018.
[6] R. Haulcy and J. Glass, "Classifying alzheimer's disease using audio and text-based representations of speech," January 2021.
[7] E. L. Campbell, L. Docio-Fernandez, J. J. Raboso, and C. García-Mateo, "Alzheimer's dementia detection from audio and text modalities," August 2020.
[8] T. Searle1, Z. Ibrahim1, and R. Dobson, "Comparing natural language processing techniques for alzheimer's dementia prediction in spontaneous speech," September 2020.
[9] M. Vacher, D. Istrate, F. Portet, T. Joubert, and T. Chevalier, "The sweet-home project: Audio technology in smart homes to improve well-being and reliance," August 2011.
[10] Z. Uddin, W. Khaksar, and J. Torresen, "Ambient sensors for elderly care and independent living: A survey," June 2018.
[11] H. Salehghaffari, "Speaker verification using convolutional neural networks," Aug 2018.
[12] S. Chaudhuri, J. Roth, D. P. W. Ellis, A. Gallagher, L. Kaver, R. Marvin, C. Pantofaru, N. Reale, L. G. Reid, K. Wilson, and Z. Xi, "Ava-speech: A densely labeled dataset of speech activity in movies," 2018.
[13] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur. (2015) Librispeech: An asr corpus based on public domain audio books. [Online]. Available: http://www.danielpovey.com/files/2015_icassp_librispeech.pdf
[14] P. Foster, S. Sigtia, S. Krstulovic, and J. P. Barker. (2015, Oct) Chime-home: A dataset for sound source recognition in a domestic environment. [Online]. Available: https://www.researchgate.net/publication/308732345_CHiME-Home_A_dataset_for_sound_source_recognition_in_a_domestic_environment
[15] S. Gupta, J. Jaafar, Ahmad, W. Wan, and A. Bansal, "Feature extraction using mfcc," *Signal & Image Processing: An International Journal*, vol. 4, no. 4, pp. 101–108, 2013.
[16] O. Bolleurs, L. Loomans, B. Appehl, D. Hollander, M. Hoendermis, and A. Bezouwen. (2021) Dialogue project github repository. [Online]. Available: github.com/ambervb/dialogue-proj
[17] A. Ashar, M. S. Bhatti, and U. Mushtaq, "Speaker identification using a hybrid cnn-mfcc approach," in *2020 International Conference on Emerging Trends in Smart Technologies (ICETST)*, 2020, pp. 1–4.
[18] N. Wilkinson and T. Niesler, "A hybrid cnn-bilstm voice activity detector," *arXiv preprint arXiv:2103.03529*, 2021.
[19] Y. Eom and J. Bang, "Speech emotion recognition using 2d-cnn with mel-frequency cepstrum coefficients," september 2021.
[20] Y. Gong, B. Liu, W. Ge, and L. Shi, "Ara: Cross-layer approximate computing framework based reconfigurable architecture for cnns," *Microelectronics Journal*, vol. 87, pp. 33–44, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0026269218307055
[21] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017.