Research Article

Open Access CrossMark

Improving voice detection in real life scenarios: differentiating television and human speech at older adults' houses

Abstract

The use of voice-operated robots in real-life settings introduces multiple issues as opposed to the use of them in controlled, laboratory conditions. In our study, we introduced conversation robots in the homes of 18 older adults' homes to increase the conversation activities of the participants. A manual examination of the audio data the robot considered a human voice showed that a considerable amount was from television sounds present in the participants' homes. We used this data to train a neural network that can differentiate between human speech and speech-like sounds from television, achieving high metrics. We extended our analysis into how the voices of the participants contain inherent patterns that can be general or uncommon and how this affects performance of our algorithm in our attempts to identify human speech with or without these patterns.

Keywords: voice identification, human robot interaction, conversation robot, machine learning, real-life scenario

Volume 9 Issue 1 - 2023

David Figueroa,[1] Shuichi Nishio,[2] Ryuji Yamazaki,[2] Hiroshi Ishiguro[1]

[1]Graduate School of Engineering Science, Osaka University, Japan

[2]Institute for Open and Transdisciplinary Research Initiatives, Osaka University, Japan

Correspondence: David Figueroa, Graduate School of Engineering Science, Osaka University, Japan, Tel +81 6-6850-8330, Email figueroa.david@irl.sys.es.osaka-u.ac.jp

Co-Correspondence: Shuichi Nishio, Osaka University, Institute for Open and Transdisciplinary Research Initiatives, Japan, Email nishio@botransfer.org

Received: December 21, 2022 | **Published:** March 01, 2023

Abbreviations: VAD, voice activity detector; CNN, convolutional neural network; FFT, fast Fourier transform

Introduction

The use of voice activated devices is rising due to the simplicity of their use, avoiding the need of learning how to use specific interfaces in order to use these devices while also allowing a natural way of interaction with machines. This facilitates the use of them by a wider audience, including people with disabilities and older adults.[1] Companion robots with voice interfaces can be used to provide simple conversation with users, especially when the users do not engage in social activities on a regular basis. In our study, we placed small conversation robots in the houses of multiple older adults living alone in order to encourage conversation and in an effort to increase the social activities of the participants.[2] Each robot had the capability of recording what the participant said in order to collect audio data to be used to enhance the conversation system and improve the overall conversation experience. When we examined the audio data collected, we noticed that a big amount of it was not composed of participant speech but sounds from other sources. Most of the misrecognized sounds came from televisions, which the older adults seem to use for most part of the day. Many of the voices that did not come from the actual participant were also identified as valid voices, which the robot misrecognized as coming from a person present in the room and produced an unexpected response. Real world environments, opposed to controlled conditions in the laboratory, have many different sources of noise, and voices that are not part of the interaction with the human but which can be identified as such, possibly causing the interaction to degrade.

To prevent the robots to answer to sounds and voices that are not coming from a person physically near them, we need a way to identify the sources of the voices so we can make sure that the robots interact with participants only. Most of the literature involving voice processing systems targets the security issue: how specific signals produced by different electronic appliances capable of playing audio can be used to issue unwanted commands to the devices. For example, Abdullah et al.[3] developed multiple classes of perturbations that are unintelligible to humans but can successfully inject undesired commands in both proprietary and public voice processing systems. Similarly, research by Vaidya et al.[4] found that it is possible to craft sounds intelligible to humans that speech recognition algorithms identify as a command.

This continues to be an active research area, with the caveat that does not address the security issue highlighted beforehand. Recent improvements in the voice activity detector (VAD) algorithms, a preprocessing step to identify voices in audio samples, differ from classical approaches[5-7] and achieve high accuracy and robustness even in noisy environments,[8,9] but does not concern with the source of the voice, either produced by a human or by electronic equipment.

On the other hand, research focused on distinguishing human-generated voices from voices generated by electronic speakers has also been addressed, but with important limitations. Wang et al.[10] focused on studying the differences in the relative phase extracted from Fourier spectrums to classify human speech and spoofed speech, achieving high accuracy, with the limitation that the dataset used contained no noise, which is far from our real-world conditions. Blue and his research team[1] discovered that an interesting phenomenon occurs in all electronic speakers, which is that there is a sub-bass over-excitation frequency due to the enclosing of the devices which can be used to differentiate human speech from other sources. To be able to use this detection technique, the microphones of the voice-activated device must be able to accurately detect frequencies between 20 and 80 Hertz, which could not be achieved by commercially available robot companions; and the microphones must have a known frequency response curve, information that might not be available when using off-the-shelf devices. More complex efforts, like detecting audio *deepfakes*,[11] simulated speech designed to sound like a specifically targeted individual, require previous knowledge of which audio belongs to a real user and which is simulated, i.e., such conditions to be met that hardly can be controlled in a real-world scenario.

While security is a very important topic when users interact with voice-activated devices, our focus falls toward voice source identification to improve the interaction with a conversation robot. Gong et al.,[12] proposed a defense strategy based on developing a machine learning model that can be used to distinguish the human speaker from a playback device, specifically, between a loudspeaker, an iPod, headphones, and an actual human. Based on this idea, we aim to develop a system capable of differentiating between a human voice and other playback devices. We gathered data from Japanese older adults living alone having short conversations with a social robot. For this research, we focused specifically on sounds produced by televisions and no other devices, such as radios, because these are not usually used by older adults. Therefore, our focus is not to create a system that can distinguish between any playback device and human voices, but it is limited to implementing a system that can differentiate between playback produced exclusively by television and no other devices. After creating such a system, we also want to investigate if there are common patterns in the human speech samples to facilitate the proper recognition of human voice activity and improve the robots' conversation system. Discovering hints about patterns present in different houses could facilitate the conversation in the case where new participants use our system.

Method

Our study placed Sharp's RoBoHoNs,[13] a humanoid companion robot, with customized software developed by our research team, directly inside the homes of older adults living alone. The trial was done in Osaka, Japan, and was based on approval from the Ethics Committee at the Graduate School of Engineering Science, Osaka University (approval code 31-3-4). Our software allowed us to keep track of the robots' functionality and monitor their proper behavior, while also collecting data in order to improve the conversation experience. The overall behavior of the robot was composed of three main blocks: detecting voice, processing the speech and generating a response. To detect a voice, the robot was continuously listening while running WebRTC's Voice Activity Detection (VAD)[14] in order to detect the presence of a voice. When the VAD detects a voice, the robot starts recording audio until the VAD signals the voice activity is finished. The recorded audio is sent to an external server where speech recognition is performed, and the result is sent back to the robot. Using this result, the robot generates a response via a mixture of different types of motion alongside generated dialogue related to the user's detected speech. The recorded audio obtained while the VAD system detects there is the presence of a voice is what we assumed to be speech coming from the robot's user. A manual examination of the data showed that multiple television voices and sounds were interpreted as valid voices coming from the participants, so the robot reacted to them, which was an undesired effect.

Data acquisition

We acquired sound data from a whole day of robot usage from 17 different participants, and three days of usage from another participant, due to low samples in one day only, giving us a total of 18 sets of conversations for use in our analysis. We used *Pyannote* VAD[15,16] as a first step for filtering our data and removing some of the audio samples without voices in them. This process reduced our dataset from 11593 audio samples to 8755. Then, we manually annotated the data, labeling each audio sample as containing participant speech, television sound, other environmental sounds, and human speech overlapped with television sound. Data that had both labels were not used in this experiment, nor data that contained only environmental noise. The total amount of data used for this study was comprised of 7469 audio samples. Table 1 shows the detail of the data we obtained.

Table 1 Detail of data obtained from robots placed at participants homes. Total samples show the data we collected from each participant, in number of recorded samples. Total samples show the amount of data before our first filtering process. The remaining filtered samples were labeled as data containing only speech or only TV sound. The difference between filtered samples and available data are due to audio coming from environment noise and/or to mixed samples with human speech and TV sounds

Participant ID	Total Samples	Filtered Samples	Human Speech	TV sound	Available data
1	656	597	302	114	416
2	167	121	116	0	116
3	364	342	336	0	336
4	285	226	95	71	166
5	660	583	142	359	501
6	540	304	283	0	283
7	521	447	238	93	331
8	105	98	26	47	73
9	434	359	313	29	342
10	1104	922	645	101	746
11	691	530	155	364	519
12	440	401	356	35	391
13	266	224	128	27	155
14	769	597	257	184	441
15	448	399	336	42	378
16	165	144	143	0	143
17	2332	2167	104	1793	1843
18	422	294	159	130	289
Total	**11539**	**8755**	**4134**	**3335**	**7469**

Data processing

For classifying our data into the classes of human speech or TV sound, we chose to use an artificial neural network, more precisely, a convolutional neural network (CNN). CNNs are used for pattern recognition and are useful to get accurate results when processing images. To use audio data with CNNs, an image representation is needed. This is why we transformed the audio samples into spectrograms, and later into mel-spectrograms. This approach has been used before to classify emotional states from voice samples,[17] achieving high accuracy. We processed our dataset using the Fast Fourier Transformation (FTT) to obtain information in the frequency domain in the form of spectrograms. We opted for a narrow-band spectrogram, which allows a fine, detailed resolution regarding frequencies.[18] We then mapped the spectrograms into mel-spectrograms to obtain better information about the voices present in the samples. We used a length of 2048 for the FFT window, 512 as hop length, and 2084 as the size of the Hanning window. We then used a filter bank of 128 bins to obtain the mel-spectrogram representation of the samples. Figure 1 shows a sample of the mel-spectrograms obtained from TV sound and human speech from the same household. The figure shows that even though both images have clear frequency resolution, represented by the horizontal white lines, there are clear differences in their resolution and differences in power in the harmonics. There is more power in defined frequencies and their harmonics in human speech than in TV sound. In our process, we aim to detect these frequency differences in order to classify the audio as human or spoofed speech. The mel-spectrograms we used were reduced to one channel, creating a black-and-white image, to reduce the computational cost in terms of storage and future processing.

Model architecture & training

The model created is composed of three blocks of 2D convolutional layers with 32 filters and reLU activation,[19] followed by maximum 2D pooling layers.[20] After the three blocks, the output goes through a flatten layer, a fully connected layer with 128 units with reLU activation, and finally, the result is fed into a sigmoid activation layer[21] to obtain the probability of an audio sample to be either TV sound or human speech. To train the model, we resized the input mel-spectrogram images to 128 by 128 pixels, used a batch size of 32, Adam optimizer,[22] and binary cross-entropy as loss function, so the output is the probability of the sample to be human speech and one minus the output is the probability of the sample to be TV sound. Figure 2 shows the network structure, as well as the input and output of the whole process.

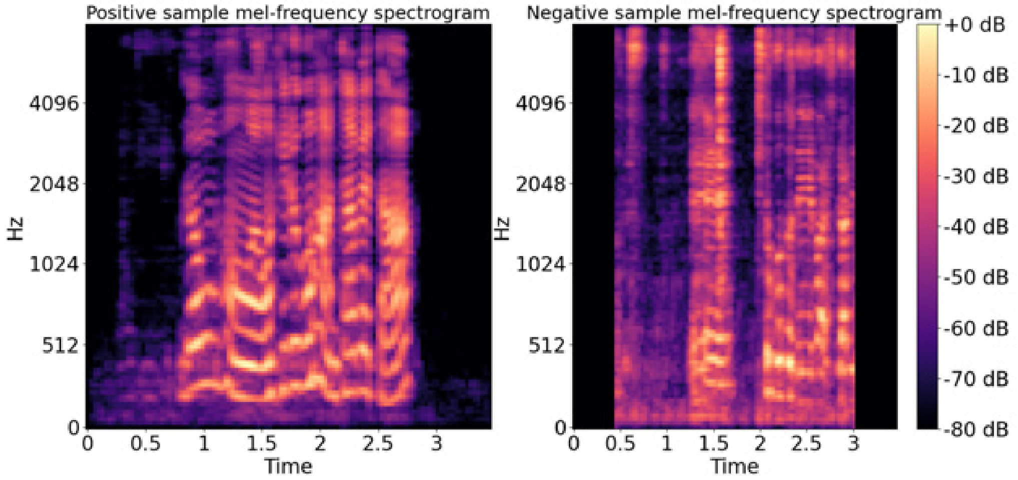


Figure 1 Samples of mel-spectrograms obtained from a sample with human speech (left) and TV sound (right).

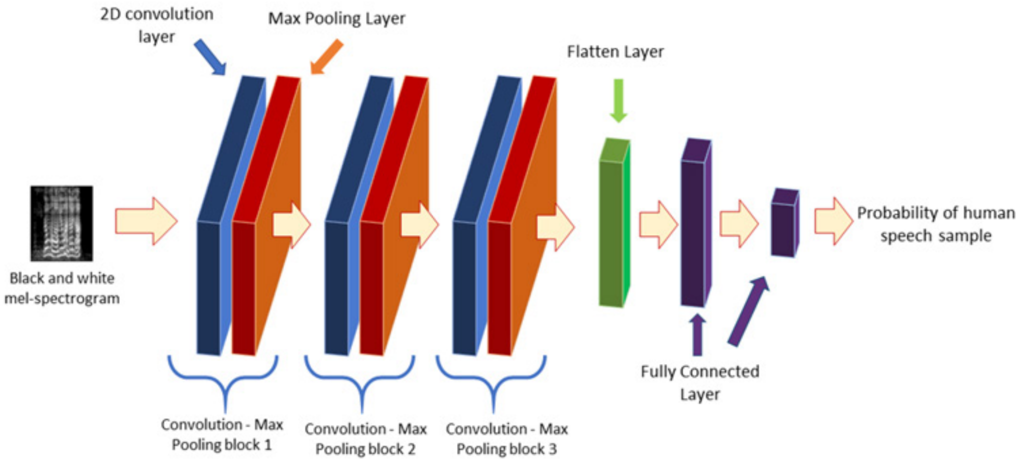

2D convolution layer
Max Pooling Layer
Flatten Layer
Black and white mel-spectrogram
Probability of human speech sample
Fully Connected Layer
Convolution - Max Pooling block 1
Convolution - Max Pooling block 2
Convolution - Max Pooling block 3

Figure 2 Network architecture used for differentiation of human speech and television speech-like sounds. As input, we use the black and white mel-spectrogram of an audio sample from our dataset. The sample goes through 3 blocks of convolutional and max pooling layers before going through a flatten layer and a fully connected layer with reLU activation, and finally to a dense layer which outputs the probability of the sample to be human speech or not.

The same architecture and hyperparameters were used to train models in two different cases. In the first case, as our dataset is relatively small, we used k-fold validation with 5 folds, and we calculated the average of the metrics. For each fold, 80% of the data was used for training and 20% was left out for testing, resulting in five different models.

As the amount of data from each house had different sizes, we also wanted to know how much each house influenced the result, which could also lead us to discover common patterns in our target group's speech. In an ideal case, where most of the samples should be equally distributed among all the households, the output metrics should be close to the k-fold validation case. In order to examine this effect, we held the data from one house for testing and used the remaining 17 houses' data for training. This resulted in 18 trained models, from which we calculated the test metrics and computed the average among all of them.

As a final case, we wanted to use the capabilities of CNNs to identify patterns and examine if there is any particular trend to differentiate human speech from TV sounds. To do so, we held data from two participants and used the remaining 16 participants data for training. Repeating this process of holding all the different possible pairs yielded 153 trained models. For the sake of completeness, we used the data held out in each case for testing and computed the average metrics.

To evaluate if there were similarities among the held-out pairs, we used data from one of the pairs as testing set for each case. As the trained model has no information regarding the excluded pair, using one of them for testing can give us some insights into how similar both participants' data are. If the accuracy is high, the corresponding pair might have little or no influence on the data used for testing. On the other hand, if the accuracy is low, the corresponding pair might have a high similarity with the testing data. We used each of the excluded paired data for testing, so we ended up testing each model twice, which gave us 306 results.

Results

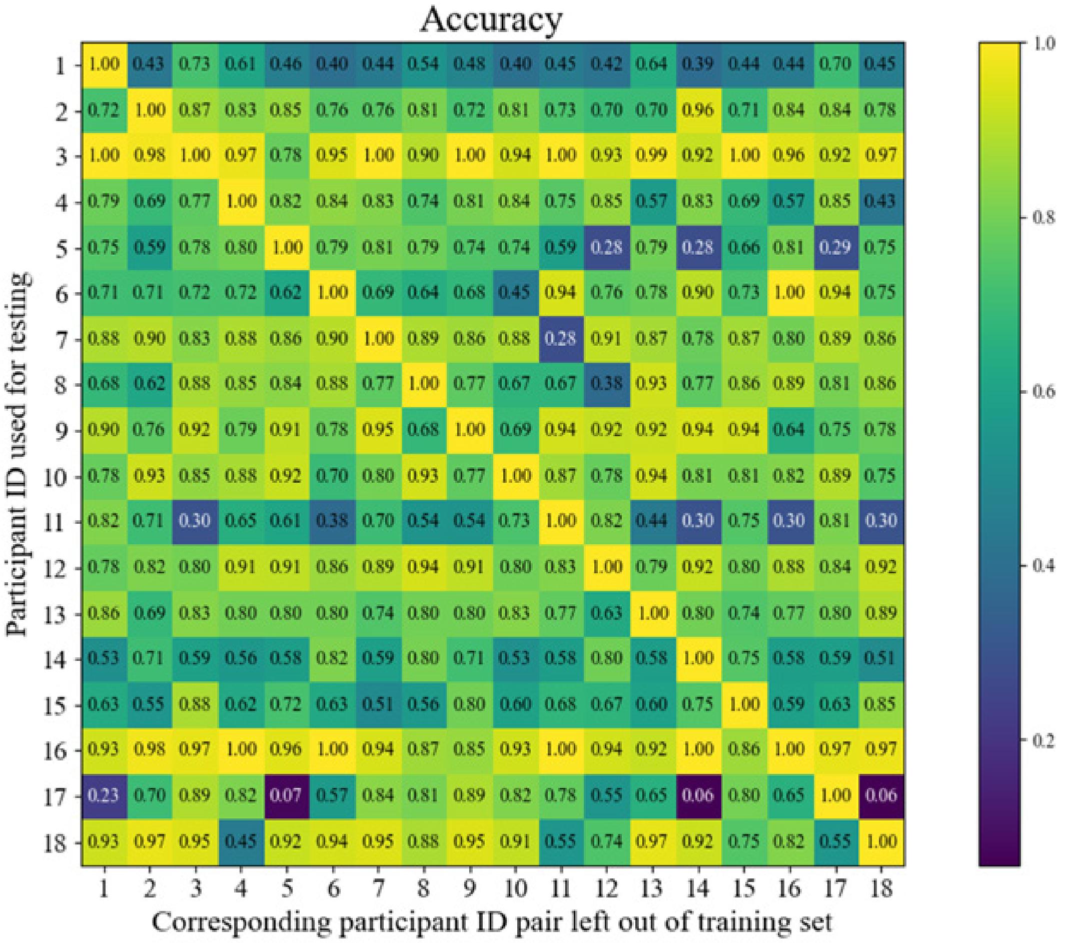
Table 2 shows the average metrics of the 5-fold cross-validation case and the metrics of the 153 models tested over pairs of houses left for testing.

Table 2 Average metrics of the two cases studied: 5-fold cross validation using the whole dataset and combinations of pairs of houses left for testing

	Metrics		
Case	**Accuracy**	**Precision**	**Recall**
5-fold cross validation	89.96%	89.14%	93.33%
One house for testing	77.91%	83.42%	76.72%
Two houses for testing	73.85%	78.64%	79.92%

The metrics when using the whole dataset and using 5 folds cross-validation are all close to 90%, while the metrics for the case when we use a pair of houses for testing decrease, accounting for a difference in accuracy, precision, and recall of 16.11, 10.50 and 13.41 percentual points respectively.

Using the accuracy metric, we built a matrix that shows the score of each trained model using one of the pairs for testing, which can be seen in figure 3. Each square in the matrix corresponds to the accuracy obtained when data from the ID indicated in the vertical axis was used for testing, while the corresponding ID in the horizontal axis is the pair left out from both training and testing. We trained 153 models, so each model is tested twice, one with each left out pair. Therefore, for example, the square with ID 1 in the horizontal axis and ID 2 in the vertical axis uses the same model as the square with ID 1 in the vertical axis and ID 2 in the horizontal axis, changing only which of the two was used for testing. In our case, the matrix is not symmetric, which suggests the left-out pairs does not have a symmetric influence with one another. The main diagonal has no value as it was added for keeping the matrix in a squared shape.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	1.00	0.43	0.73	0.61	0.46	0.40	0.44	0.54	0.48	0.40	0.45	0.42	0.64	0.39	0.44	0.44	0.70	0.45
2	0.72	1.00	0.87	0.83	0.85	0.76	0.76	0.81	0.72	0.81	0.73	0.70	0.70	0.96	0.71	0.84	0.84	0.78
3	1.00	0.98	1.00	0.97	0.78	0.95	1.00	0.90	1.00	0.94	1.00	0.93	0.99	0.92	1.00	0.96	0.92	0.97
4	0.79	0.69	0.77	1.00	0.82	0.84	0.83	0.74	0.81	0.84	0.75	0.85	0.57	0.83	0.69	0.57	0.85	0.43
5	0.75	0.59	0.78	0.80	1.00	0.79	0.81	0.79	0.74	0.74	0.59	0.28	0.79	0.28	0.66	0.81	0.29	0.75
6	0.71	0.71	0.72	0.72	0.62	1.00	0.69	0.64	0.68	0.45	0.94	0.76	0.78	0.90	0.73	1.00	0.94	0.75
7	0.88	0.90	0.83	0.88	0.86	0.90	1.00	0.89	0.86	0.88	0.28	0.91	0.87	0.78	0.87	0.80	0.89	0.86
8	0.68	0.62	0.88	0.85	0.84	0.88	0.77	1.00	0.77	0.67	0.67	0.38	0.93	0.77	0.86	0.89	0.81	0.86
9	0.90	0.76	0.92	0.79	0.91	0.78	0.95	0.68	1.00	0.69	0.94	0.92	0.92	0.94	0.94	0.64	0.75	0.78
10	0.78	0.93	0.85	0.88	0.92	0.70	0.80	0.93	0.77	1.00	0.87	0.78	0.94	0.81	0.81	0.82	0.89	0.75
11	0.82	0.71	0.30	0.65	0.61	0.38	0.70	0.54	0.54	0.73	1.00	0.82	0.44	0.30	0.75	0.30	0.81	0.30
12	0.78	0.82	0.80	0.91	0.91	0.86	0.89	0.94	0.91	0.80	0.83	1.00	0.79	0.92	0.80	0.88	0.84	0.92
13	0.86	0.69	0.83	0.80	0.80	0.80	0.74	0.80	0.80	0.83	0.77	0.63	1.00	0.80	0.74	0.77	0.80	0.89
14	0.53	0.71	0.59	0.56	0.58	0.82	0.59	0.80	0.71	0.53	0.58	0.80	0.58	1.00	0.75	0.58	0.59	0.51
15	0.63	0.55	0.88	0.62	0.72	0.63	0.51	0.56	0.80	0.60	0.68	0.67	0.60	0.75	1.00	0.59	0.63	0.85
16	0.93	0.98	0.97	1.00	0.96	1.00	0.94	0.87	0.85	0.93	1.00	0.94	0.92	1.00	0.86	1.00	0.97	0.97
17	0.23	0.70	0.89	0.82	0.07	0.57	0.84	0.81	0.89	0.82	0.78	0.55	0.65	0.06	0.80	0.65	1.00	0.06
18	0.93	0.97	0.95	0.45	0.92	0.94	0.95	0.88	0.95	0.91	0.55	0.74	0.97	0.92	0.75	0.82	0.55	1.00

Figure 3 Accuracy matrix using models trained without the data from the house ID in the horizontal and vertical axes and tested with the data from the house ID in the vertical axis. A score closer to one indicates a high similarity with the corresponding pair left out of the training and testing data.

Discussion

The classifier trained with all the available data achieved nearly 90% in terms of accuracy, precision and recall. Reviewing the total amount of data described in Table 1, this means that from the initial 3335 TV sound samples the robot reacted to, using our classifier, the robot would react to approximately only 333. If we account that this inaccuracy came from data from 18 different participants, this would mean that the robot would react to the TV sound on an average of around 24 voice-like sounds, which is a big improvement. Reducing the amount of human speech misrecognitions, the robot behavior could be improved to focus on the actual interaction we are interested in, the one performed with the older adults taking part in our research. This could also improve the overall interaction experience from the participants perspective, as the robot could be perceived as focused and interested in the human, also leading to higher robot usage.

As the data we gathered is not balanced, i.e., each participant' house contained different amount of audio samples in both the human speech and TV sound categories, we assume that each household data contributes in different ways to the classifier network. The results of our models tested with data from one house only and trained with the remaining data show that there is an actual decrease in the classifying performance. We can think about two different explanations of this phenomenon. The first cause could be that the imbalanced amount of data causes the classifier to overfit and the network has not enough information to correctly classify unseen data. A clear case of this could be the case where participant 17's data is used for testing. The total amount of audio samples from participant 17 is 1843, divided in 1739 samples of TV sound and 104 samples of human speech. Participant might have the robot placed too close to the television, or had the television set to high volume, so the robot received many more samples of TV noise. Removing this amount of data from the training set would clearly cause the classifier to underperform and overfit the training set.

The second explanation could be that some houses contain general patterns common to the other households, causing these data to have a bigger contribution to the generalization power of the classifier. This would mean that what is affecting the network's performance is not exactly the amount of data, but the underlying patterns in each house, which the classifier can learn and generalize to unseen data. To further examine this possible explanation, we trained models leaving out a pair of houses data and using only one in testing, and observed how accurate the network is in generalizing unseen data without another house data. We can have a rough idea of the influence of the left-out house over the test house. This is summarized in figure 3. This figure shows that some testing data seem to achieve high accuracy regardless which other participant data was excluded. The clearest example of this is participant 3. When participant 3 is used as testing data (row-wise direction) we can see that the accuracy is high, suggesting that this set of data might contain very general patterns which the classifier can learn from the rest of the data, resulting in a good generalization. When participant 3 is the left-out pair (column-wise direction), the majority of test sets achieve an accuracy higher than 75%, suggesting that the general pattern of this participant does not have a big influence in the classifier as the general pattern can be learned from the train data. There are, of course, a few exceptions, the most noticeable being when participant 11 is used for testing. Data from participant 3 seems to have many similarities with participant 11, as the general patterns of participant 3 seem to be very close to the ones of participant 11 and the network fails to generalize unseen data and reports a low accuracy score of only 30%.

In broad terms, we can see similar cases of this explanation also when participants 9 and 10 are used for testing: an approximate overall accuracy of over 75%, with specific households actually affecting the resulting accuracy. These specific households suggest that both participants might have similar patterns, making the pair similar if the accuracy is low, or independent, sharing general patterns with other participants data, when the accuracy is high.

An interesting result is that some of the left-out pairs seem to have a one-way influence only. For example, the pair composed by participant 17 and 14. The classifier have not seen data from these participants during training, yet using participant 17 as testing set yields an accuracy of only 6%, very different from using participant 14 as the testing set for the same model, reaching an accuracy of 59%. The low accuracy obtained from using participant 17 for testing would suggest that participants 17 and 14 are similar, yet using participant 14 for testing achieves a much higher accuracy. Counter-intuitive results like this suggest that each dataset might contain more than one underlying pattern. In such a case, for example, we could think that participant 14 has two patterns, one general and one similar to participant 17, while participant 17 might have one pattern only, and directly related to 14, which could be why we see these kinds of one-way similarity.

Using the whole data available achieved our desired objective of classifying sound samples as human speech or TV sound. However, one limitation of the current system is that, while we are using real-world data containing noise and gathered outside laboratory-controlled conditions, we have discarded data with *both* human speech and TV sound. Based on the successful results obtained in the present study, further research can be conducted in order to manage this kind of data too. Also, focusing efforts on analyzing in depth the different patterns that each participant's household contains and their relation with other houses could allow us to create different classifiers focused specifically on certain patterns, in order to reduce the robots' misrecognitions when dealing with new participants. We could train several models specialized on a type or pattern, and use them with the first phrases from the human speaking to the robot in order to choose which voice pattern group the new participant belongs to and improve the overall conversation experience.

The CNN used in our research seems to learn the underlying patterns in the audio samples to differentiate between the television sounds and the actual human speech. We can suppose of the characteristics the network learns can relate to previous research, such as high power in frequencies at the low end of the spectrum. Another pattern that might help in the classification is the presence of different effects that voices produced in television shows often have, which are not present in human speech. The CNN operates mostly as a black box, which makes the inherent weights and structures difficult to interpret. This topic can be included in future research, which can be used to optimize the network to have better detection and allow the system to operate embedded in the robot, allowing real-time recognition.

Conclusion

The neural network based on CNNs proposed in our research achieved high performance in differentiating between television voice-like sounds and voices from the human participants in proximity to the robot. When using the complete dataset we gathered from our field experiment, the classifier achieves metrics of 89.96% in accuracy, 89.14% in precision and 93.33% in recall. As gathering data from real-life situations present many challenges, we got a different amount of data from the participants. Because of this, we extended our analysis in an effort to understand how different voice patterns affect the results of the neural network. From this analysis, we observed that data collected from one participant's household can have a non-intuitive one-way influence with data from other participants' homes. This result suggests that a particular voice might have multiple different patterns that the neural network can identify, i.e., some patterns being common, which do not affect the rest of the voices of different users, while some others seem to be unusual and have a high effect in specific speech patterns.

Acknowledgments

This work was partially supported by Innovation Platform for Society 5.0 at MEXT.

Conflicts of interest

The authors declare there is no conflict of interest.

References

1. Blue L, Vargas L, Traynor P. *Hello, is it me you're looking for? Differentiating between human and electronic speakers for voice interface security*. Proceedings of the 11th ACM Conference on Security & Privacy in Wireless and Mobile Networks: New York, USA; 2018. 123–133 p.
2. Yamazaki R, Nishio S, Nagata Y, et al. *A Preliminary Study of robotic media effects on older adults with mild cognitive impairment in solitude*. Proceedings of International Conference on Social Robotics: Singapore; 2021. 10–13 p.
3. Abdullah H, Garcia W, Peeters C, et al. *Practical hidden voice attacks against speech and speaker recognition systems*. Proceedings of the 26th Network and Distributed System Security Symposium: San Diego, USA; 2019. 24–27 p.
4. Vaidya T, Zhang Y, Sherr M, et al. *Cocaine noodles: Exploiting the gap between human and machine speech recognition*. Proceedings of the 9th USENIX Conference on Offensive Technologies: Denver, USA; 2015. 16 p.
5. Hughes T, Mierle K. *Recurrent neural networks for voice activity detection*. Proceedings of the 38th International Conference on Acoustics, Speech and Signal Processing: Vancouver, Canada; 2013. 26–31 p.
6. Ming J, Hazen T, Glass R, et al. Robust speaker recognition in noisy conditions. *IEEE Transactions on Audio, Speech and Language Processing*. 2017;15(5):1711–1723.
7. Germain F, Sun D, Mysore G. *Speaker and noise independent voice activity detection*. Proceedings of the 14th Annual Conference of the International Speech Communication Association: Lyon, France; 2013. 25–29 p.
8. Braun S, Tashev I. *On training targets for noise-robust voice activity detection*. Proceedings of the 29th European Signal Processing Conference: Dublin, Ireland; 2021. 23–27 p.
9. Sarkar E, Prasar R, Magimai-Doss M. *Unsupervised voice activity detection by modeling source and system information using zero frequency filtering*. Proceedings of the 22nd Annual Conference of the International Speech Communication Association: Incheon, Korea; 2022. 18–22 p.
10. Wang L, Yoshida Y, Kawakami Y, et al. *Relative phase information for detecting human speech and spoofed speech*. Proceedings of the 16th Annual Conference of the International Speech Communication Association: Dresden, Germany; 2015. 6–10 p.
11. Blue L, Warren K, Abdullah H, et al. *Who are you (I really wanna know)? Detecting audio deepFakes through vocal tract reconstruction*. Proceedings of the 31st USENIX Security Symposium: Carlsbad USA; 2022. 11–13 p.
12. Gong Y, Poellabauer C. *Protecting voice controlled systems using sound source identification based on acoustic cues*. Proceedings of the 27th International Conference on Computer Communication and Networks: Hangzhou, China; 2019.
13. Sharp corporation. Product information; 2020.
14. Google Inc. WebRTC.
15. Bredin H, Yin R, Coria J, et al. *Pyannote.audio: Neural building blocks for speaker diarization*. Proceedings of the 47th International Conference on Acoustics, Speech, and Signal Processing: Barcelona, Spain; 2020. 4–8 p.
16. Bredin H, Laurent A. *End-to-end speaker segmentation for overlap-aware resegmentation*. Proceedings of the 21st Annual Conference of the International Speech Communication Association: Brno, Czech Republic; 2021.
17. Dossou B, Gbenou Y. FSER: *Deep convolutional neural networks for speech emotion recognition*. Proceedings of the 18th International Conference on Computer Vision Workshops: Montreal, Canada; 2021. 11–17 p.
18. Cheung S, Lim J. *Combined multi-resolution (Wide-band/narrowband) spectrogram*. Proceedings on the 1991 International Conference on Acoustics, Speech, and Signal Processing: Toronto, Canada; 1991. 14–17 p.
19. Nair V, Hinton, G. *Rectified linear units improve restricted boltzmann machines*. Proceedings of the 27th International Conference on Machine Learning: Haifa, Israel; 2010. 21–24 p.
20. Murray N, Perronnin F. *Generalized max pooling*. Proceedings of the 27th Conference on Computer Vision and Pattern Recognition: Columbus, USA; 2014. 23–28 p.
21. Pratiwi H, Windarto A, Susliansyah S, et al. Sigmoid activation function in selecting the best model of artificial neural networks. *Journal of Physics*: *Conference Series*. 2020;1471(1):1–7.
22. Kingma D, Ba J Adam. *A method for stochastic optimization*. In Proceedings of the 3rd International Conference on Learning Representations: San Diego, USA; 2015. 7–9 p.

Citation: Figueroa D, Nishio S, Yamazaki R, et al. Improving voice detection in real life scenarios: differentiating television and human speech at older adults' houses. *Int Rob Auto J.* 2023;9(1):8–13. DOI: 10.15406/iratj.2023.09.00255