

Using Variational Autoencoders to Reduce Noise in Speech Processing for Cochlear Implants

Marlous Nijman
s4551400
Artificial Intelligence
Radboud University
Nijmegen
27-6-2019

Bachelor's Thesis in Artificial Intelligence

Supervisor: Yagmur Güçlütürk
Artificial Intelligence, Donders Institute for Brain, Cognition and
Behaviour, Radboud University Nijmegen

ABSTRACT

In this thesis, I present a variational autoencoder model to reduce noise in speech processing for cochlear implants. The model was trained on noisy and clean speech samples that were converted to spectrograms, to be able to create clean speech from a noisy speech input. The variational autoencoder consists of a convolutional encoder and a deconvolutional decoder. Furthermore, a cochlear implant simulator was developed to allow for testing on normal hearing participants. Participants were asked to judge the quality of original noisy speech data and noisy speech data that was processed by the model, where both of which went through the cochlear implant simulator. Results indicate that the quality of speech was worse when it had went through the variational autoencoder compared to the original noisy speech.

1. INTRODUCTION

1.1. Cochlear Implants

Around 300.000 people worldwide have a registered cochlear implant [1]. A cochlear implant (CI) is a device that allows patients who suffer from sensorineural hearing loss, which is caused by damage to inner ear, either to the cochlear hair cells or to the VIIIth nerve, to regain some hearing functionality [2]. This is achieved through the stimulation of the nerve cells in the tonotopically organized basilar membrane in the cochlea, using an electrode array. The activation pattern of the electrode array depends on the sounds that are picked up from the environment using the CI's externally mounted microphone [2].

Although speech understanding performance is quite good in quiet environments, CI users struggle with understanding speech in noisy environments, compared to healthy individuals [3]. One way of improving performance in an environment with a low signal-to-noise ratio is to increase the number of channels [4]. Up until now, from around 4 up until 22 channels have been used in cochlear implants [5]. The focus of this research will be on another aspect of a cochlear implant to improve speech understanding performance in noisy environments. That is, the focus will be on the signal processor of the CI, which transforms the recorded sound into a signal for the electrode array. Some commonly used speech processing methods are continuous interleaved sampling (CIS), the spectral peak (SPEAK) strategy, and the advanced combination encoder (ACE) [5]. These methods convert the sounds coming from the environment to a stimulation pattern, but do not necessarily take into account possible noise coming from the environment. Different types of methods exists to

improve the quality of speech, but in this thesis the focus will be on the use of deep neural networks.

1.2. (Denoising) Autoencoders

Autoencoders are a type of neural network that learn a representation of its input data in an unsupervised manner. A denoising autoencoder is a type of autoencoder that is trained to reconstruct a clean output from a noisy, *corrupted* input [6]. Lu, Tsao, Matsuda, and Hori [7] developed a deep autoencoder (DEA) neural network to enhance speech using this method. The DEA estimates clean speech from noisy speech input. Their experiments show that the DEA performed better than a minimum mean square error based speech enhancement algorithm on all performance evaluations that were used, such as noise reduction and speech distortion. Furthermore, results showed that increasing the depth of the DEA improved performance as well [7].

1.3. Variational Autoencoders

Another variation on the autoencoder is the variational autoencoder (VAE). A variational autoencoder learns a complex distribution of a certain dataset in an unsupervised manner [8]. A property of VAEs that separates them from regular AEs is that the latent spaces that they learn are by design continuous [9]. This is what allows the random sampling from the learned distribution, and is achieved by the reparameterization trick: instead of a learning a latent vector between the encoder and the decoder, we learn two different vectors, a mean vector μ and a standard deviation vector σ [10]. From these vectors a sample is drawn, which then will be decoded, using the decoder part of the network. To achieve that the network learns to generate output that is as close as possible to the input, while also learning to follow a normal distribution in the mean

and standard deviation vectors, two different loss functions are combined. The mean squared error (MSE) loss is used to minimize the difference between the input and the output of the network. Furthermore, the Kullback-Leibler divergence (KLD) is used to minimize the difference between the probability distribution vectors μ and σ , and the desired distribution, usually a normal distribution [9].

VAEs are often used for generative modelling. For example, [11] used a variational autoencoder to analyze images. A deep Convolutional Neural Network (CNN) was used as an encoder, and a Deep Generative Deconvolutional Network (DGDN) was used as a decoder. The latent vector that was learned using this VAE model was later used to generate labels and captions for the images using other types of models. Their setup showed to be able to obtain results that are similar to results of state-of-the-art methods on a number of different tasks.

The application of these kinds of deep neural networks have mostly shown promising results in the field of generative modelling, but can also be used in other areas, such as speech enhancement. Bando, Mimura, Itoyama, Yoshii, and Kawahara [12] created a semi-supervised speech enhancement method called VAE-NMF, that uses a variational autoencoder to model speech, and non-negative matrix factorization (NMF) as a model for noise. Their results show that their method outperforms other supervised deep neural network methods, and is more robust to various types of noise [12].

Besides the previously mentioned study, there are not many other cases where a VAE has been used to denoise speech. Therefore, it might be interesting to further investigate the usefulness of variational autoencoders in noise reduction applications.

1.4. Convolutional Networks

Convolutional networks are often used in models for image processing and classification. Krizhevsky, Sutskever, and Hinton [13] have shown impressive results when using a deep convolutional neural network in image classification. Furthermore, in the (variational) autoencoder models created by [7] and [12], a speech signal was first converted to an image, e.g. a spectrogram, before feeding it into the network. Because of the fact that both convolutional networks and (variational) autoencoders work well with

images, and sound signals can easily be converted into spectrograms, it would be interesting to see how these two methods combined would perform on denoising speech signals. Therefore, this research will be combining a variational autoencoder with a deep convolutional network as its encoder, and a deep deconvolutional network as its decoder.

1.6. Research Question

All previously mentioned concepts and techniques will be used to answer the following research question: "Can the quality of speech perception in noisy environments be improved for cochlear implant users, using a variational autoencoder?".

2. METHODS

The data that was used for training and testing the model, comes from a noisy speech database for training speech enhancement algorithms [14]. The training set consists of matching clean and noisy speech samples recorded from 28 different speakers.

After some preprocessing, the speech samples were converted to spectrograms. These spectrograms were then fed into the network. When the VAE was properly trained, it was tested on unseen, noisy data. The spectrograms that were obtained during the testing phase were converted back to sound. Finally, to be able to test the success of the network on healthy, normal hearing subjects, a cochlear implant simulator was developed. The results from the testing phase of the network was fed through this simulator, as well as original noisy data coming from the dataset's test set. Finally, an online survey was conducted to rate the quality of speech in both of these settings.

2.1. Preprocessing

The first step done during preprocessing was the down sampling of the original speech data coming from the dataset from 48000 Hz to 16000 Hz. This was done to reduce the amount of data that needs to be processed, while maintaining the quality of the sound fragments.

Due to the fact that the dataset consists of speech samples of different durations, the second step in preprocessing was the splitting of the speech samples into chunks of equal size. The duration of the

chunks was defined to be 250 ms, where a silent audio segment was appended to chunks with a duration shorter than 250 ms. Following this, the spectrograms of the sound chunks were obtained using short-term Fourier transform (STFT), with a window size of 512, and a step size of 32, resulting in an image of size 112 x 256 (time domain x frequency domain). A python implementation developed by [15] was used to compute the STFT. This implementation is well designed to be used in neural networks, such as the VAE developed for this research. It contains a method to obtain the inversion (i.e. sound) of the spectrogram as well, that is needed to be able to evaluate the performance of the network in the experiment conducted on normal hearing subjects.

2.2. Model

2.2.1. Architecture

The model that was implemented is a variational autoencoder, which consists of an encoder, a reparameterization layer, and a decoder (Figure 1). The encoder that was implemented consists of several consecutive convolutional and max-pooling layers, where after each convolutional layer a rectified linear unit (ReLU) activation function is applied. The implementation details of the encoder can be found in Table 1. The decoder consists of consecutive deconvolutional and max-unpooling layers, and is a mirrored version of the encoder. The max-pooling layers in the encoder return indices that are used in the max-unpooling layers in the decoder. Again, ReLU activation functions were used after the deconvolutional layers, except for the last deconvolutional layer. After this last layer, a sigmoid activation function was used. This was done to ensure that the output values of the network are in the range 0 to 1. This is a requirement for the inversion function that is used to obtain the speech samples from the

outputted spectrograms. The implementation details of the decoder can be found in Table 2.

The layers between the encoder and the decoder are what separates a variational autoencoder from a regular autoencoder, and is what is called the *reparameterization layer*. Here, the network is forced to generate a latent vector that follows a normal distribution [8]. It does so by generating two different layers, one representing the mean, and the other representing the standard deviation of a normal distribution. From the mean and standard deviation a latent vector can be sampled [10]. From this latent vector, it is possible to retrieve back an image using the decoder of the VAE.

2.2.2. Loss Function

The objective of the network is to generate an image that matches the desired output, i.e. clean speech, as accurately as possible, as well as for the latent vector to follow a normal distribution as accurately as possible. To achieve this, two different loss functions were combined: the mean squared error (MSE) loss, and the Kullback-Leibler divergence (KLD) [8] [10]. The mean squared error was computed using the following formula:

$$MSE = \frac{1}{N} \sum_{i=1}^N (f_i - y_i)^2$$

Where N is the number of data points, f_i is the value returned by the model for data point i, and y_i is the desired value for data point i. The KLD was computed using the following formula [16]:

$$KLD = \frac{1}{2} \sum_{j=1}^J \left(1 + \log \left(\frac{\sigma_j^2}{\mu_j^2} \right) - \frac{\mu_j^2}{\sigma_j^2} - \frac{\sigma_j^2}{\mu_j^2} \right)$$

Where J is the dimension of the latent vector, μ the variational mean, and σ the variational standard deviation. The total loss function is then the sum of the KLD loss and the MSE loss.

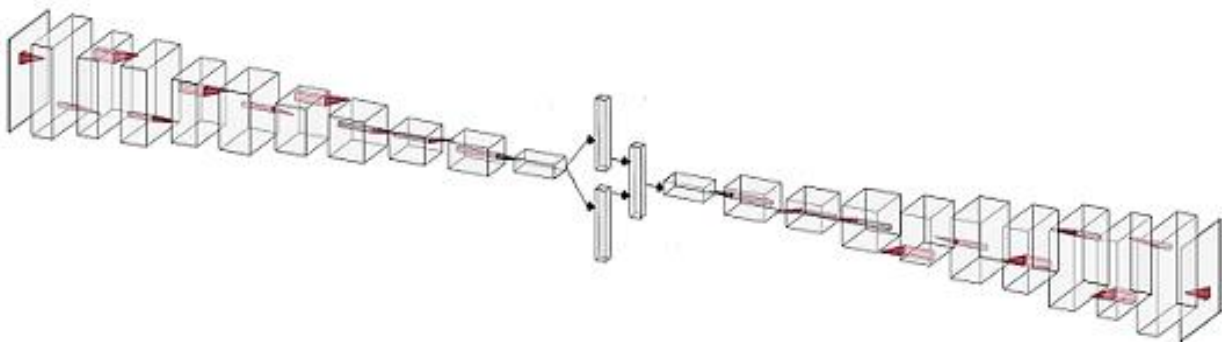


Figure 1. Variational Autoencoder Architecture.

Layer Type	Channels In	Channels Out	Kernel Size	Stride
Conv2d	1	16	(5, 5)	(1, 1)
MaxPool2d	-	-	(2, 2)	(2, 2)
Conv2d	16	32	(5, 5)	(1, 1)
MaxPool2d	-	-	(2, 2)	(2, 2)
Conv2d	32	64	(4, 4)	(1, 1)
MaxPool2d	-	-	(2, 2)	(2, 2)
Conv2d	64	128	(4, 4)	(1, 1)
MaxPool2d	-	-	(2, 2)	(2, 2)
Conv2d	128	256	(1, 4)	(1, 1)
MaxPool2d	-	-	(2, 2)	(2, 2)
Conv2d	256	512	(1, 3)	(1, 1)

Table 1. Encoder Implementation Details

Layer Type	Channels In	Channels Out	Kernel Size	Stride
ConvTranspose2d	512	256	(1, 3)	(1, 1)
MaxUnpool2d	-	-	(2, 2)	(2, 2)
ConvTranspose2d	256	128	(1, 4)	(1, 1)
MaxUnpool2d	-	-	(2, 2)	(2, 2)
ConvTranspose2d	128	64	(4, 4)	(1, 1)
MaxUnpool2d	-	-	(2, 2)	(2, 2)
ConvTranspose2d	64	32	(4, 4)	(1, 1)
MaxUnpool2d	-	-	(2, 2)	(2, 2)
ConvTranspose2d	32	16	(5, 5)	(1, 1)
MaxUnpool2d	-	-	(2, 2)	(2, 2)
ConvTranspose2d	16	1	(5, 5)	(1, 1)

Table 2. Decoder Implementation Details

2.2.3. Training

The VAE was then trained to minimize the loss using the Adam optimizer with a learning rate of 1e-4. Training was done for 15 epochs using a batch size of 1024.

2.2.4. Testing

After the VAE was trained, it could be tested on unseen data. In this case the noisy test set from [14] was used. The resulting spectrograms were stored after testing to be postprocessed later.

2.3. Postprocessing

The spectrograms generated during the testing phase were converted to sound using inverse short-term Fourier transform. This was done using the implementation developed by [15]. The resulting speech sample chunks were then appended in the correct order to reconstruct the original speech samples.

2.3. Simulator

In order for subjects to judge whether the developed model improves the intelligibility of speech in noisy environments, a simulator has to be implemented, that simulates how the processed sound coming from the environment would sound when hearing it through a cochlear implant. To achieve this, the obtained, denoised signal will have to be processed using a speech processing method, for example continuous interleaved sampling (CIS) [17] [18]. The output of this method will be a pulse train, that indicates how each channel in the CI’s electrode array has to be activated. Perceived sound can then be simulated by outputting sine waves for each channel at the frequency corresponding to that channel, as is done by [19].

For the implementation, first the number of channels was chosen and subsequently their stimulated frequency ranges were computed. As mentioned previously, the number of electrodes in cochlear implants typically range from 4 to 22 [5]. As it is more likely that the number of electrodes will increase instead of decrease in the future, the highest number of electrodes used today, i.e. 22, was chosen to be the number of electrodes in the CI simulator.

The frequency ranges corresponding to each of the channels were determined using the Greenwood formula, which is a cochlear frequency-position function [20]:

$$F = A(10^{ax} - k)$$

Here, F is the obtained frequency corresponding to position x in the cochlea (here expressed as a proportion of the basilar length). The constant A was set to 165.4, a to 2.1, and k to 1, as proposed in [20]. This formula was then used to compute the frequency ranges for each of the 22 channels, with the ranges over all channels being between 150 Hz and 8000 Hz, which is typically the case in cochlear implants [21]. In reality an upper bound frequency of 7999 Hz instead of 8000 Hz was used, due to the fact that all frequencies had to be smaller than the sampling rate (here, 16000 Hz) divided by two. The computed frequency ranges can be found in Appendix A.

The simulator was then implemented by outputting sine waves for each channel at its corresponding frequency [19], using continuous interleaved sampling (CIS). The current Python implementation was based on the MATLAB version

developed by [22], and the description of CIS by [17]. First, the signal was filtered by a FIR bandpass filter of length 1025 using a Hamming window, for each of the channels. The bandpass frequencies correspond to the frequency ranges of each of the channels. Then, the amplitude envelope of the filter output was computed, by first obtaining the analytical signal using the Hilbert transform, and then taking its absolute value. The obtained amplitude envelope was then multiplied with the sine wave corresponding to the same channel. This was done for all channels, and the sum of these multiplied sine waves was the final output of the simulator.

2.4. Experiment

Finally, an experiment was conducted to compare the results of the model with noisy speech samples from the test set. The experiment was an online survey consisting of 30 questions. Participants were asked to listen to a speech fragment once to judge the quality of the speech. The question that was asked was the following: "How would you rate the quality of the speech in this speech fragment?". To get a subjective rating from the participants, the most widely used opinion rating method was used: the Mean Opinion Score (MOS) [23]. Therefore, the questions were multiple choice, where participants could choose one of the following possible answers:

- Excellent
- Good
- Fair
- Poor
- Unsatisfactory

The question contained a clickable link, that played a speech fragment when clicked. Half of these speech fragments consisted of noisy speech samples that were processed by the cochlear implant simulator. The other half of the speech fragments consisted of noisy speech samples, that had been passed through the VAE, and subsequently passed through the simulator. The questions were asked in random order.

The first page of the survey consisted of a brief explanation of the experiment. It also contained an informed consent form, which contained a link to the general information brochure. Participants gave consent to participating in the experiment, by filling in the survey. The introductory page of the survey

and the general information brochure can be found in Appendix B.

3. RESULTS

3.1. Loss

Below, the losses during training of the model are shown. Figure 2 shows the average training and validation losses during training for all the epochs (epochs 1 to 15). Due to the average loss of the first epoch being relatively high, the trend of the losses after this epoch is unclear in this graph. Figure 3 shows the average training and validation losses for epochs 2 to 15, to give a better view on the learning curve of the network. What can be seen from these losses, is that the network learns a lot in the first epoch. However, in reality the learning curve was not as steep as it is portrayed in Figure 1. This is due to the fact that the average loss over the entire epoch is plotted, and not all the separate losses. In Figure 2., a more gradual learning curve can be observed, where the loss of the network gradually decreases. As can be seen, the training and validation loss are close to each other, with the validation loss sometimes being worse and sometimes being better than the training loss. This indicates that the network is not over- or underfitting on the training data.

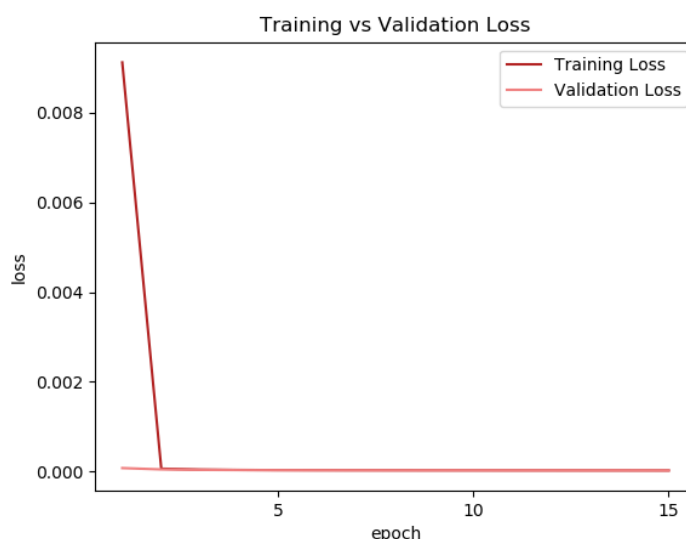


Figure 2. Average Training and Validation Losses During Training for Epochs 1 to 15.



Figure 3. Average Training and Validation Losses During Training for Epochs 2 to 15.

3.2. Survey Results

The online survey was completed by 10 participants, and the summarized results can be found in Table 3 (the raw data can be found in Appendix C). Table 3 shows the counts of the ratings for the different types of speech fragments that were included in the survey, as described in section 2.4. A visual representation of these results is shown in Figure 4.

	Original	Model
Excellent	14	0
Good	45	1
Fair	55	24
Poor	20	48
Unsatisfactory	16	77

Table 3. Rating Counts for Noisy Speech Samples Processed by the Cochlear Implant Simulator (Original) Compared to Noisy Speech Samples Processed by Both the VAE and the Cochlear Implant Simulator (Model).

A Chi-squared test was performed to analyze if the rating distribution for the different type of speech fragments are significantly different. The results indicate that there is a significant difference between the two distributions, as the p-value is significantly smaller than 0.05 ($X^2 = 119.79$, $df = 4$, $p\text{-value} < 2.2e-16$).

To improve statistical power, a second statistical analysis, i.e. a t-test, was performed. The qualitative ratings were converted to quantitative ratings, by giving the ratings a numerical score on a

scale from 1 to 5, with “Excellent” being given a score of 5, and “Unsatisfactory” a score of 1. The results of the t-test indicated a significant better average rating for the original noisy data ($M = 3.14$, $SD = 1.11$) compared to the noisy data that was processed by the model ($M = 1.66$, $SD = 0.77$), $t = 13.472$, $df = 265.48$, $p\text{-value} < 2.2e-16$.

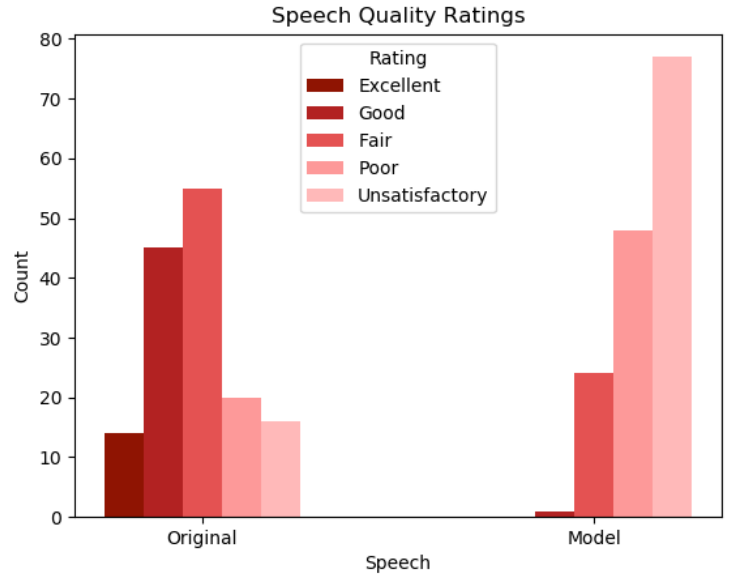


Figure 4. Rating Counts for Noisy Speech Samples Processed by the Cochlear Implant Simulator (Original) Compared to Noisy Speech Samples Processed by Both the VAE and the Cochlear Implant Simulator (Model).

4. DISCUSSION

As can be seen in the previous section, the variational autoencoder did not improve the speech quality in noisy environments in (simulated) cochlear implants. The developed model instead showed worse performance than the unprocessed speech samples. The following subsections investigate the factors that may have had an effect on the results.

First, combining spectrograms and convolutional layers may not have the same results as combining natural images with convolutional layers. As discussed in section 1., one could argue that spectrograms might work well with convolutional layers, since it represents some sort of image. However, there are some properties of spectrograms that do not hold for images, which could result in poorer results when combining them with convolutional layers. In the first place, the axes of a spectrogram do not have the same meaning. The x-axis represents time, whereas the y-axis represents

different frequencies (a possible solution to this will be discussed in the final paragraph of this discussion). Furthermore, as described by [24], the use of convolutional neural networks for images is partly based on the notion of translational invariance. Translational invariance means that an image feature is the same no matter where it is in the image [24]. This does not always hold for spectrograms. For example, when a feature is moved along the y-axis, the frequency of that feature will change, resulting in a completely different sound. Secondly, a pixel in a natural image usually corresponds to exactly one object. On the contrary, a pixel in a spectrogram can belong to an accumulation of different sources of sound coming from the environment [24]. This problem as well as the problem of translational invariance are inherent to the combination of spectrograms and convolutional layers. It does, however, not mean that it cannot be useful, as demonstrated by [25] and [26]. It would be interesting to find out whether spectrograms combined with fully connected linear layers would actually yield better results compared to convolutional layers, to see what the effects of these spectrogram properties really is.

Another cause of the poor ratings for the model output has likely to do with the methods used to convert spectrograms back to sound. When listening to the output of the model, one can clearly hear a burst of noise or loud sound at the point in time where the original noisy speech sample was cut into chunks. The sound is quite loud and very distracting of the actual speech in that sound fragment. A different method of obtaining and converting spectrograms back to sound may be a solution to this problem.

The use of a VAE in this particular application requires the use of the MSE loss. As mentioned by [27], when autoencoder (AE) models are applied to complex datasets of images, they tend to produce blurry images. Furthermore, they mention that this effect in AE and VAE models is associated with the use of the MSE loss [27]. This property of the MSE loss is not very likely to have caused the poor results of the model by itself. Similarly, the combination of convolutional layers and spectrograms is not likely to have caused a performance this poor by itself, especially considering the previous research where this combination did proof to be useful [25] [26]. However, a combination of all previously mentioned

factors is probably what explains the performance of the VAE in this research.

Finally, the number of subjects that participated in the online survey was quite small. However, it is highly unlikely that a larger number of participants would change the significant difference between the two speech rating distributions in a meaningful way. This is due to the fact that the p-value for both statistical tests was significantly smaller than 0.05. Both tests thus indicated that the null hypothesis, i.e. the hypothesis that the two distributions are equal, can be rejected.

Lastly, I will discuss some suggestions for future research. As mentioned previously, the combination of convolutional layers and spectrograms may be problematic. A possible solution to this problem is mentioned by [28]. Here, instead of treating the spectrogram as a 2D image consisting of a time and frequency axis as well as a 1-dimensional channel, it is treated as a 1 by the time domain image, with the channels being equal to the frequency dimensions. Another suggestion for future research would be to combine a variational autoencoder and a generative adversarial network, as was done by [29] and [30]. Instead of using an element-wise measure such as the MSE as reconstruction error, a GAN discriminator is used to measure similarity [29]. This approach also solves the problems of MSE and blurry images, that was mentioned before. This is confirmed by the results in [29], where the reconstructed faces using the VAE/GAN are much sharper compared to the results generated using just a VAE, and similar to the result of a GAN. In [30], a conditioned VAE/GAN was developed. Their work shows promising results on natural images, similar to those in [29], thus it might be worth investigating whether this kind of setup would work on spectrograms as well.

5. CONCLUSION

In this thesis, a variational autoencoder was proposed to increase speech perception quality in noisy environment speech samples for cochlear implant users. A cochlear implant simulator was developed as well to test the results on normal hearing participants. Experimental results have shown that the performance of the developed model is worse than actual noisy speech. The poor performance of the model can likely be attributed to

several factors, such as combining convolutional layers with spectrograms, the use of the MSE loss, and pre- and postprocessing methods. Suggestions for future research include a different type of representation of spectrograms in convolutional layers, as well as the proposal of a model that combines a variational autoencoder and a generative adversarial network.

6. REFERENCES

- [1] National Institute on Deafness and Other Communication Disorders (NIDCD), "Cochlear implants," 6 March 2017. [Online]. Available: <https://www.nidcd.nih.gov/health/cochlear-implants>. [Accessed 27 Februari 2019].
- [2] D. Purves, G. J. Augustine, D. Fitzpatrick, W. C. Hall, A.-S. LaMantia, J. O. McNamara and S. M. Williams, *Neuroscience*, 3rd ed., Sunderland, MA, U.S.A.: Sinauer Associates, 2004.
- [3] L. M. Friesen, R. V. Schannon, D. Baskent and X. Wang, "Speech recognition in noise as a function of the number as a function of the number of spectral channels: Comparisons of acoustic hearing and cochlear implants," *Acoustical Society of America*, pp. 1150-1163, 2001.
- [4] M. F. Dorman, P. C. Loizou, J. Fitzke and Z. Tu, "The recognition of sentences in noise by normal-hearing listeners using simulations of cochlear-implant signal processors with 6-20 channels," *The Journal of the Acoustical Society of America*, pp. 3583-3585, 1998.
- [5] B. S. Wilson and M. F. Dorman, "Cochlear implants: a remarkable past and a brilliant future," *Hearing research*, p. 3-21, 2008.
- [6] P. Vincent, H. Larochelle, Y. Bengio and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proceedings of the 25th International Conference on Machine Learning*, Helsinki, Finland, 2008.
- [7] X. Lu, Y. Tsao, S. Matsuda and C. Hori, "Speech enhancement based on deep denoising autoencoder," in *Interspeech*, Lyon, France, 2013.
- [8] C. Doersch, "Tutorial on variational autoencoders," no. arXiv preprint arXiv:1606.05908., pp. 1-23, 2016.
- [9] I. Shafkat, "Intuitively Understanding Variational Autoencoders," 4 Februari 2018. [Online]. Available: <https://towardsdatascience.com/intuitively-understanding-variational-autoencoders-1bfe67eb5daf>. [Accessed 26 June 2019].
- [10] K. Frans, "Variational Autoencoders Explained," 6 August 2016. [Online]. Available: <http://kvfrans.com/variational-autoencoders-explained/>. [Accessed 12 May 2019].
- [11] Y. Pu, Z. Gan, R. Henao, X. Yuan, C. Li, A. Stevens and L. Carin, "Variational autoencoder for deep learning of images, labels and captions," in *Advances in neural information processing systems*, Barcelona, Spain, 2016.
- [12] Y. Bando, M. Mimura, K. Itoyama, K. Yoshii and T. Kawahara, "Statistical speech enhancement based on probabilistic integration of variational autoencoder and non-negative matrix factorization," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Alberta, Canada, 2018.
- [13] A. Krizhevsky, I. Sutskever and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Advances in neural information processing systems*, pp. 1097-1105, 2012.
- [14] C. Valentini-Botinhao, *Noisy speech database for training speech enhancement algorithms and TTS models*, University of Edinburgh. School of Informatics. Centre for Speech Technology Research (CSTR), 2017.
- [15] T. Sainburg, "Spectrograms, MFCCs, and Inversion in Python," 7 July 2018. [Online]. Available: <https://timsainburg.com/python-mel-compression-inversion.html>. [Accessed 12 March 2019].

- [16] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," in *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*, Banff, Canada, 2014.
- [17] B. Somek, F. Sinša, A. Dembitz, M. Ivković and J. Ostojić, "Coding strategies for cochlear implants," *Automatika*, pp. 69-74, 2006.
- [18] B. S. Wilson, C. C. Finley, D. T. Lawson, R. D. Wolford and M. Zerbi, "Design and evaluation of a continuous interleaved sampling (CIS) processing strategy for multichannel cochlear implants," *Journal of Rehabilitation Research*, pp. 110-116, 1993.
- [19] M. F. Dorman, P. C. Loizou and D. Rainey, "Simulating the effect of cochlear-implant electrode insertion depth on speech understanding," *The Journal of the Acoustical Society of America*, pp. 2993-2996, 1997.
- [20] D. D. Greenwood, "A cochlear frequency-position function for several species—29 years later," *The Journal of the Acoustical Society of America*, pp. 2592-2605, 1990.
- [21] H. Jones, A. Kan and R. Y. Litovsky, "Comparing Sound Localization Deficits in Bilateral Cochlear-Implant Users and Vocoder Simulations With Normal-Hearing Listeners," *Trends in Hearing*, pp. 1-16, 2014.
- [22] Z. Fu, "YannyLaurel-with-CochlearImplant-Simulation," 13 April 2019. [Online]. Available: <https://github.com/fuzhenfz/YannyLaurel-with-CochlearImplant-Simulation>. [Accessed 5 May 2019].
- [23] K. Kondo, *Subjective quality measurement of speech: its evaluation, estimation and applications*, Berlin Heidelberg: Springer, 2012.
- [24] L. Wyse, "Audio spectrogram representations for processing with Convolutional Neural Networks," in *Proceedings of the First International Workshop on Deep Learning and Music joint with IJCNN*, Anchorage, US, 2017.
- [25] H. Zhang, I. McLoughlin and Y. Song, "Robust sound event recognition using convolutional neural networks," in *IEEE international conference on acoustics, speech and signal processing (ICASSP)*, South Brisbane, Queensland, Australia, 2015.
- [26] E. M. Grais and M. D. Plumbley, "Single channel audio source separation using convolutional denoising autoencoders," in *IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, Montreal, Canada, 2017.
- [27] S. Zhao, J. Song and S. Ermon, "Towards deeper understanding of variational autoencoding models," no. arXiv preprint arXiv:1702.08658., 2017.
- [28] D. Ulyanov and V. Lebedev, "Audio texture synthesis and style transfer," 13 December 2016. [Online]. Available: <https://dmitryulyanov.github.io/audio-texture-synthesis-and-style-transfer/>. [Accessed 16 June 2019].
- [29] A. B. Larsen, S. K. Sønderby, H. Larochelle and O. Winther, "Autoencoding beyond pixels using a learned similarity metric," in *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, New York, USA, 2016.
- [30] J. Bao, D. Chen, F. Wen, H. Li and G. Hua, "CVAE-GAN: fine-grained image generation through asymmetric training," in *Proceedings of the IEEE International Conference on Computer Vision*, Venice, Italy, 2017.

APPENDIX

A. Frequency Ranges

Channel	Frequency (Low)	Frequency (High)
1	150	200
2	200	258
3	258	326
4	326	404
5	404	495
6	495	600
7	600	722
8	722	864
9	864	1028
10	1028	1218
11	1218	1439
12	1439	1695
13	1695	1991
14	1991	2335
15	2335	2734
16	2734	3196
17	3196	3731
18	3731	4353
19	4353	5073
20	5073	5909
21	5909	6876
22	6876	7999

B. Survey

B.1. Introduction Text

Survey on the Quality of Speech in Cochlear Implants

Thank you for agreeing to take part in this survey! I am conducting this survey as part of my thesis research in Artificial Intelligence. The survey is being conducted to evaluate the quality of speech in cochlear implants. You will be asked to listen to short speech fragments that sound similar to how cochlear implant patients perceive sound. After listening to a speech fragment only once, you are asked to judge its quality (i.e. if the speech can be heard clearly, or you hear some noise) by choosing one of the following options:

- Excellent
- Good
- Fair
- Poor
- Unsatisfactory

To listen to the speech fragment you will have to click on the link that says "speech fragment". This link will open a new window. To answer the question you can simply return to the previous page in your browser.

This survey is anonymous, so no personal data will be saved. This survey should take approximately 10 minutes to complete.

By participating in this survey you confirm that:

- I was satisfactorily informed about the study both verbally and in writing, by means of the general information brochure and additional study specific information brochure(s) (version 2.1, December 2018).
- I have had the opportunity to put forward questions regarding the study and that these questions have been answered satisfactorily.
- I have carefully considered my participation in the experiment.
- I participate voluntarily.

agree that:

- My research data will be acquired and stored for scientific purposes as mentioned in the general information brochure until 10 years after the research has been finalized.
- Personal data is acquired for administrative and scientific purposes.
- The connection between my personal and research data is stored until maximally one month after finalization of this study.
- Demographic data or data concerning my health, background or preferences is collected for scientific purposes.
- My not directly identifiable experimental data will be made public for verification, re-use and/or replication.
- Regulatory authorities can access my data for verification purposes.

- I will be informed by a designated expert, my general practitioner or a general practitioner of the Academic General Practitioner Center Heyendaal about any information which is of clinical relevance to me.

understand that:

- I have the right to withdraw from the experiment at any time without having to give a reason.
- I have the right to request disposal of my experimental data up to 1 month after participation.
- My privacy is protected according to applicable European law (European General Data Protection Regulation (GDPR)).
- My consent will be sought every time I participate in a new experiment.

If you have any questions/comments regarding this survey you can send an email to M.Nijman@student.ru.nl

By clicking the arrow below you will start the survey.

B.2. General Information Brochure



GENERAL
INFORMATION
Version 2.1

This folder contains important information if you consider participating in one of the studies at the Donders Centre for Cognition of the Donders Institute. Please read the following information carefully.

The Donders Institute is a university research centre investigating the brain, cognition and behaviour. The Donders Centre for Cognition (DCC) is part of the Donders Institute and has at its disposal various techniques in order to measure behaviour and brain activity. For our research we need volunteers to participate in various experiments, e.g. language, perception, action or memory tasks. All our research and research methods have negligible to minimal risks.

Ethics check

Each study you may participate in has been reviewed and approved by an independent ethics committee (the 'Ethics Committee of Faculty of Social Sciences', Radboud University Nijmegen', <https://www.radboudnet.nl/socialsciences/research/ethics-committee-social-science/> or the medical ethics committee, www.cmoregio-a-n.nl).

Clinical data

Researchers at the DCC do not examine the data from a clinical perspective. Participation in any of the experiments can therefore not be considered as a clinical or screening test. In exceptional circumstances the data collected may give indications concerning your health conditions. Prior to participation in these kinds of experiments you are required to provide name and address of your general practitioner. Additionally, in these types of experiments, we will archive your name, your personal identification number (of our subject database), and/or your (email) address. In case of a possible finding which is of clinical relevance you will be informed by a designated specialist of the DCC, or your general practitioner. In case you do not have a general practitioner (in the Netherlands), you will be informed by the Academic General Practitioner Center Heyendaal, for which you will then have to register as a patient once. Your insurance policy will cover these costs; In case you are uninsured the Academic GP Center is required to charge you a minimal amount for consultation. If you do not wish to be informed about findings concerning your health, you cannot participate in experiments at the DCC.

Information about the experiment and giving consent

You will receive a study specific information brochure from the researcher sufficiently in advance (this means minimally 24 hours in advance) of participation in the experiment. This will allow you time to reflect on your participation. Prior to participation, you are asked to sign a study specific informed consent form in which you confirm that you have been informed satisfactorily and are willing and able to participate voluntarily. The researcher will also sign the form, confirming that you have been informed about the experiment satisfactorily. The researcher will also ensure your privacy and that the necessary privacy conditions will be met. You have the right to withdraw from the experiment at any time without giving a reason. You can request disposal of your experimental data up to 1 month after participation in the study. After that your data will be pseudonymized, this means not

directly identifiable, and stored in a repository. An example of the study specific “informed consent” is attached to the applicable study specific information brochure.

Insurance

On legal grounds a liability insurance and in some cases an additional subject insurance has been concluded for subjects participating in studies at the DCC as part of the Donders Institute. The subject insurance covers damage due to participation in the study, becoming apparent during participation in the study or within four years after termination of participation in the study.

Use and preservation of your data

For our research it is necessary to collect, use and preserve personal information. This concerns personal data like name, address, date of birth, email address or personal identification number of the participant database. Use and preservation of your personal data is necessary for administrative and scientific goals. These goals are: the documentation of consent for participation in research, payment for participation, granting request to destroy data, to approach in the case of incidental findings, and to approach for future research (in case consent has been given for this). In some cases it is necessary to collect demographical data or data concerning your health, background, or preferences for scientific purposes. If you do not agree with this, you cannot participate in this research.

Confidentiality of your data

The information you provide for the purpose of the study will be handled carefully and will only be accessible to employees who are authorized to do so. Your data will be treated confidentially. All your research data will be coded in order to protect your privacy. Your name and other information, which might lead to your identity, will be kept separate from the experimental data. Only with a so called key file your experimental data can be linked to your identity. To protect your privacy this key file will also be stored apart from the research data. Only members of the research team who are directly involved and for whom it is necessary can access your personal data and the key file. Other parties involved in the research receive only access to the coded research data and won't be able to identify you directly on the basis of this data. Reports or publications on the study will also only report your coded not directly identifiable research data.

In some studies additional audio, photo and/or video recordings will be obtained during the experiment. These are solely collected for scientific purposes. The experimenter will always inform you about this prior to participation, additionally asking for your approval. In all cases your privacy will be protected according to European law (European General Data Protection Regulation, GDPR)

Preservation time

Your data will be archived during for an established period of time, which is until 10 years after the research has been finalized. The connection between your personal data and your research data will be stored until maximally one month after finalization of the research.

Sharing your experimental data

Given the importance of verification, re-use and/or replication of research results, experimental data are shared or made public more often. Prior to this sharing the data will completely be pseudonymized (this means not traceable to your identity). In case of concerns regarding sharing your experimental data, you have the right to request disposal of your experimental data up to 1 month after participation.

Some experimental data cannot be pseudonymized completely due to its nature, e.g. video-, photo or audio recordings. You have the right to disapprove sharing these data

with other researchers beyond the scope of the study. This can be done via the study specific consent form.

Right to inspection

Few other people or agencies have the right to inspect your data, both personal and research data. This is necessary in order to check if the research was properly and reliably conducted. The persons or agencies who can obtain access to your data for the purpose of verification are: a controller who works for the responsible institute, and national or international regulatory bodies such as the Ministry of Health. They will protect and keep your personal information secret. You are requested to approve this right to inspection. In case you do not agree, you cannot participate in the study.

Future studies

After participation in an experiment at the DCC, it may be the case that we would like to approach you again for a future study. You can indicate on the consent form whether you agree with this. In case you consent to be approached we will store name, address, email address and your identification number from the participant database, if applicable. Also in these future experiments, participation is always voluntary and consent will be sought every time you participate in a new experiment.

Preparation of the experiment

Generally, no extra preparation is required before participation. It is important that you are fit, alert and that you did not drink alcohol or used drugs the night before.

Before the start of the actual experiment the researcher will explain the aims of the research and the applied measurement techniques to you. You will receive instructions about what you are asked to do during the experiment, such as watching a monitor, listen to sounds (possibly over a headset), perform a reaction task, make different movements or just lie still and relax. After everything has been fully explained, you will be asked to sign the consent form. Subsequent procedures depend on the research method that is being used. You can read more about this in the information brochures on EEG, moving chair, robot or EEG-FES.

Payment

Participation in experiments is reimbursed. The DCC gives this reimbursement by means of participant hours or VVV giftcards (<https://www.vvvcadeaubonnen.nl/>). In this latter case we need your name and address for administrative reasons.

Additional information, independent contact person and contact

If you are unable to make it to the appointment (on time), please inform the responsible researcher as soon as possible. You may also contact this researcher for additional information or if you would like to withdraw from participating.

After participation

We appreciate hearing about your experiences as a participant. You can give your feedback – with or without personal information - via this [webform](#). In case of questions or complaints about an experiment, contact the responsible experimenter first. You can also contact an independent contact person who is not involved in the study (independent contact person: Miriam Kos, of the Donders Centre for Cognition (dcclabcoordinator@socsci.ru.nl; tel 0243612650)) or fill out our webform. If applicable the independent contact person will contact you by phone.

More information concerning your rights for processing data

For more information with respect to compliance with your rights regarding the processing your personal data, you may contact the responsible entity for processing your data. The

Radboud University is responsible for compliance with the rights of processing personal data for this research. You may contact the office of the data Protection Officer of the Radboud University via privacy@ru.nl. More information about your rights regarding processing of personal data can be found at <https://www.ru.nl/privacy/english/> and on the website of the Dutch Data Protection Authority: <https://autoriteitpersoonsgegevens.nl/en>.

C. Survey Results

C.1. Original

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
Q1	Good	Excellent	Excellent	Excellent	Fair	Good	Excellent	Fair	Good	Good
Q2	Good	Good	Good	Excellent	Poor	Good	Good	Poor	Excellent	Fair
Q3	Fair	Good	Good	Fair	Poor	Fair	Poor	Fair	Poor	Fair
Q4	Fair	Good	Excellent	Excellent	Poor	Fair	Poor	Fair	Fair	Fair
Q5	Good	Good	Good	Fair	Fair	Fair	Good	Unsatisfactory	Fair	Good
Q6	Good	Good	Excellent	Excellent	Fair	Good	Fair	Unsatisfactory	Good	Good
Q7	Good	Fair	Fair	Fair	Fair	Poor	Fair	Unsatisfactory	Fair	Fair
Q8	Good	Good	Good	Good	Poor	Fair	Excellent	Unsatisfactory	Good	Good
Q9	Good	Good	Good	Fair	Fair	Poor	Good	Unsatisfactory	Fair	Fair
Q10	Fair	Fair	Good	Fair	Fair	Fair	Fair	Unsatisfactory	Good	Fair
Q11	Poor	Poor	Unsatisfactory	Fair	Unsatisfactory	Unsatisfactory	Unsatisfactory	Unsatisfactory	Unsatisfactory	Unsatisfactory
Q12	Poor	Fair	Fair	Good	Poor	Poor	Fair	Unsatisfactory	Poor	Poor
Q13	Fair	Fair	Excellent	Good	Fair	Good	Fair	Poor	Fair	Fair
Q14	Fair	Good	Excellent	Good	Poor	Poor	Fair	Unsatisfactory	Fair	Fair
Q15	Good	Fair	Excellent	Good	Fair	Fair	Good	Unsatisfactory	Good	Good

C.2. Model

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
Q16	Poor	Fair	Poor	Poor	Unsatisfactory	Unsatisfactory	Unsatisfactory	Unsatisfactory	Unsatisfactory	Unsatisfactory
Q17	Poor	Poor	Poor	Poor	Unsatisfactory	Unsatisfactory	Unsatisfactory	Unsatisfactory	Unsatisfactory	Unsatisfactory
Q18	Poor	Fair	Poor	Poor	Unsatisfactory	Unsatisfactory	Unsatisfactory	Unsatisfactory	Unsatisfactory	Poor
Q19	Fair	Fair	Poor	Poor	Unsatisfactory	Unsatisfactory	Poor	Unsatisfactory	Unsatisfactory	Unsatisfactory
Q20	Fair	Fair	Poor	Poor	Unsatisfactory	Unsatisfactory	Unsatisfactory	Unsatisfactory	Unsatisfactory	Unsatisfactory
Q21	Poor	Fair	Poor	Fair	Unsatisfactory	Unsatisfactory	Unsatisfactory	Unsatisfactory	Unsatisfactory	Unsatisfactory
Q22	Fair	Fair	Poor	Poor	Unsatisfactory	Unsatisfactory	Unsatisfactory	Unsatisfactory	Unsatisfactory	Unsatisfactory
Q23	Fair	Fair	Fair	Poor	Unsatisfactory	Unsatisfactory	Poor	Unsatisfactory	Poor	Poor
Q24	Fair	Good	Poor	Unsatisfactory	Unsatisfactory	Poor	Unsatisfactory	Unsatisfactory	Unsatisfactory	Poor
Q25	Poor	Fair	Unsatisfactory	Fair	Unsatisfactory	Unsatisfactory	Poor	Unsatisfactory	Unsatisfactory	Unsatisfactory
Q26	Fair	Poor	Poor	Fair	Unsatisfactory	Poor	Unsatisfactory	Poor	Unsatisfactory	Unsatisfactory
Q27	Fair	Fair	Poor	Unsatisfactory	Unsatisfactory	Unsatisfactory	Unsatisfactory	Unsatisfactory	Unsatisfactory	Poor
Q28	Fair	Fair	Poor	Poor	Unsatisfactory	Unsatisfactory	Unsatisfactory	Unsatisfactory	Poor	Poor
Q29	Poor	Poor	Poor	Poor	Unsatisfactory	Unsatisfactory	Unsatisfactory	Unsatisfactory	Unsatisfactory	Unsatisfactory
Q30	Poor	Fair	Fair	Poor	Poor	Poor	Unsatisfactory	Poor	Unsatisfactory	Unsatisfactory