A Review of A Regression Approach to Speech Enhancement Based on Deep Neural Networks

Introduction

Hearing impairment is an incredibly pervasive issue, affecting more than 1.5 billion people worldwide, and is estimated to grow to 2.5 billion people by 2050[1]. Relative to the US, about 40 million Americans are afflicted by hearing impairment[2]. Individuals may experience hearing impairment due to many different causes, like trauma to the ear, genetic predisposition, proximity effects of another disease/virus, etc. With this, hearing impairment can intensify over time, due to environmental exposure to sounds and one's lived experience. For reference, while only 2-3 Americans out of 1000 are born with hearing impairment in one ear[1], 1 in 8 Americans over the age of 12 experience hearing impairment in both ears[3]. In many cases of hearing impairment, it is challenging for individuals to perceive sounds. One reason for this is that those inflicted by hearing impairment have a narrower dynamic range compared to someone with normal hearing. Additionally, individuals experiencing hearing impairment may struggle to differentiate the relative intensity of sounds. In effect, individuals with hearing impairment can experience trouble understanding expressions within language and lateralizing sounds in crowded areas. This can make it challenging for individuals to connect within the world of hearing, diminishing their quality of life. While hearing impairment can be caused by many different ailments, it commonly affects the inner ear or the hearing nerve, as sensorineural loss is the most common type of hearing impairment[4]. Current solutions, like hearing aids and cochlear implants, function to amplify sounds either mechanically or via electrical stimulation of nerve endings. However, current solutions are not perfect solutions as they struggle to detect dynamic speech patterns and perform in noisy environments.

Our project focuses on developing a machine learning-based solution to enhance the speech quality and intelligibility of audio for hearing-impaired individuals. The model will utilize deep neural networks to enhance the auditory experience by improving background noise filtering, reducing distortion, and adapting to the user's hearing needs. This will be achieved by utilizing open-source, training corpus data of speech recordings encompassing different conditions and environments to adapt an automatic speech recognition system. Furthermore, this model could be integrated into various platforms like hearing aids, cochlear implants, and other hearing assistive devices to improve the user's outcome.

The current state-of-the-art in speech enhancement includes various techniques like spectral subtraction, Wiener filtering, and MMSE estimators. However, these methods often produce artifacts like "musical noise" and struggle with non-stationary noise types. The state of the art method we use to compare objective metrics, such as Short-Time Objective Intelligibility (STOI) scores and Perceptual Evaluation of Speech Quality (PESQ), with the DNN performance is logMMSE. LogMMSE is notable for several key reasons: most importantly, its robustness in noise reduction is highly regarded. It employs a statistical approach to estimate the clean speech signal from noisy observations, which is crucial for enhancing speech intelligibility and quality

in noisy environments. Additionally, logMMSE effectively balances noise suppression with speech signal distortion. This balance is essential in applications such as mobile communications and hearing aids, where clarity and quality are both critical. Another significant advantage of logMMSE is its ability to minimize the generation of musical noise, a common and undesirable artifact produced by many noise reduction techniques. Furthermore, the logMMSE method's widespread use and extensive study over the years have established it as a reliable benchmark for evaluating new speech enhancement techniques[5]. In the context of the paper by Xu et al, the comparison of the Deep Neural Network (DNN)-based approach with the traditional logMMSE method serves to benchmark the improvements offered by DNNs[6]. This comparison is especially pertinent in handling a wide range of noise types, including the challenging non-stationary noises, where DNNs potentially offer significant advancements over conventional methods like logMMSE. Such comparative analyses help in understanding the strengths and limitations of both the traditional and the emerging techniques in real-world applications.

This project underlines the significance and prevalence of hearing impairment in the US and worldwide, and its effects on individuals suffering from them. As such, this project aims to bridge the communication gap for individuals with hearing impairment, significantly improving their quality of life and enabling them to engage more fully in the world around them. By utilizing neural deep learning, this project will train and adapt an automatic speech recognition system to enhance auditory experience. This project represents a promising step towards a more inclusive and equitable society for those affected by hearing impairment.

Methods

Preprocessing: Creating training, validation, and test sets with noisy and clean speech data

In this study, our preprocessing methodology was designed to address real-world challenges by incorporating noise data from the OpenSLR Room Impulse Noise Database[7] and clean speech data from the ARU Speech Corpus at the University of Liverpool[8]. The ARU speech corpus includes recordings of IEEE sentences spoken by British English speakers. To construct our training dataset, we leveraged the clean speech data available in the ARU Speech Corpus. A sample of 4620 utterances from this data was systematically corrupted with the noise types from the OpenSLR database at two distinct SNR levels: 0 dB, and -5 dB. This process resulted in the creation of a multi-condition training set, comprising pairs of clean and noisy speech utterances. The inclusion of noise data from the OpenSLR Room Impulse Noise Database ensured the representation of acoustic environments. Our training set, constituting 60% of the overall dataset, encompassed approximately three hours of data. The remaining 40% of the dataset was divided into a validation set (20%) and a test set (20%) for the model.

Introduction to Methodology

The project employs a Deep Neural Network (DNN)-based framework for speech enhancement, primarily aimed at improving the quality and intelligibility of speech for individuals with hearing impairments. The methodology revolves around the principle of regression, where a mapping function between noisy and clean speech signals is established using DNNs. This approach is a deviation from the conventional Minimum Mean Square Error (MMSE)-based noise reduction techniques. The DNN-based approach is proposed to overcome these limitations by leveraging a large dataset encompassing a wide range of noise types and deep learning models for robust noise suppression.

Data Preprocessing during the Training and Enhancement

The data preprocessing primarily occurs between the input of the speech and noise data and the DNN in the form of a feature detector. In the time domain alone, it is impossible to fully separate mixed clean speech and noisy sound data due to the periodic nature of consonants and noise. In order to give the DNN the ability to distinguish between speech and sound, the data is preprocessed in a similar way to how the human body decodes sound). In speech processing, when we train deep neural networks (DNNs) to enhance speech, the DNN also aims to align the data representation with the limits of human auditory perception. Humans have a limit to how finely they can resolve temporal differences in sound, typically not perceiving gaps or changes in sound that occur in less than 3 milliseconds. To mimic this in a DNN, we frame the audio data into small time segments or frames. This is counterintuitive because digitized sound is discrete, but by sampling the sound in 3ms frames instead of extremely tiny chunks (an artifact of sound data through an extremely high sampling rate such as 16,000 Hz), incomprehensible to a human, we train the neural network more in line to how humans would perceive the sound. By restricting the model to sample data using frames, we effectively overwhelm the network with less information, ensuring the network only makes use of the same information that human brains are working with, when they are decoding sound. However, we don't want to make each frame too large. Doing so will cause the spectral resolution of the sound frequency data to be muddled. Ultimately, the DNN described in the paper decided on a frame length of 32 milliseconds.

Windowing is used in tandem with framing to address spectral leakage. The spectral leakage is an artifact of using frames and a discrete fourier transform, where the edges of the frames fail to translate well under a normal fourier transform. Windowing solves this problem by weighting the function in the time domain with a Gaussian distribution to emphasize the weights at the center of each frame and emphasize the weights at the edges of each frame. The use of windowing and framing is a critical preprocessing step before feeding data into a DNN to preserved a signal from the artifacts of a normal discrete fourier transform on discrete, subsequently occurring frames.

In addition to processing the speech data in frames, a Discrete Fourier Transform is performed on each frame in the speech and noise file pair during the training stage, where each frame is decomposed into their amplitude and phase components. In the enhancement stage, only the mixed-noise frames are fed into the DNN through the feature extractor. This time, unlike in

the fine-tuning training section of the model, where the learning is supervised and the phases of speech and noisy frames are extracted, the phase of the mixed-noisy frames is preserved throughout the speech enhancement. “Although phase information is important in human speech recognition… phase was extracted directly from the noisy signal considering that our ears are insensitive to small phase distortions or global spectral shifts” (Xu et al.).

While the phase is preserved from the noisy signal, further preprocessing is necessary to prepare the sound data to be visualized by the DNN. The aptitude-frequency spectrum is squared to create the power spectrum, which describes the power of the signal for each set of frequency bands. However, because humans process sound logarithmically, allowing us to begin to compare and distinguish between extremely loud and soft noises, the log of the power spectrum is taken to linearize the power spectrum, allowing the DNN to have the information necessary to decode sound data similar to how we do. This is ultimately how we preprocess our sound data.”

Method Architecture Description

During the DNN training procedure, different learning rates were employed for the pre-training and fine-tuning phases. For the unsupervised pre-training part, where the model was prepared by stacking multiple restricted Boltzmann machines (RBMs), the learning rate was set to 0.0005 (Xu et al., 2015). This phase involved training each layer of the DNN to maximize the likelihood over the training samples. In the subsequent fine-tuning stage, which was supervised, the initial learning rate was set higher at 0.1 for the first 10 epochs. After these initial epochs, the learning rate was then decreased by 10% after each subsequent epoch (Xu et al., 2015). This strategy was employed to refine the DNN's parameters to reduce the mean squared error between the estimated and reference clean speech features.

In the DNN architecture, multiple layers with non-linear features are used, aiding to create a complex function that can convert noisy speech into clear speech. During the fine-tuning section, the model uses a Mean Squared Error (MSE) criterion in the log-power spectral domain. MSE is a popular method in speech processing and machine learning because it provides a way to compare different models. It's especially useful because it gives more weight to larger errors than smaller ones, reducing noticeable mistakes or distortions in the clearer speech. Using MSE in the log-power spectral domain also matches well with how humans hear. By employing MSE, the DNN model mimics this logarithmic perception, evaluating errors in a way that mirrors how humans judge the loudness and pitch of sounds. .

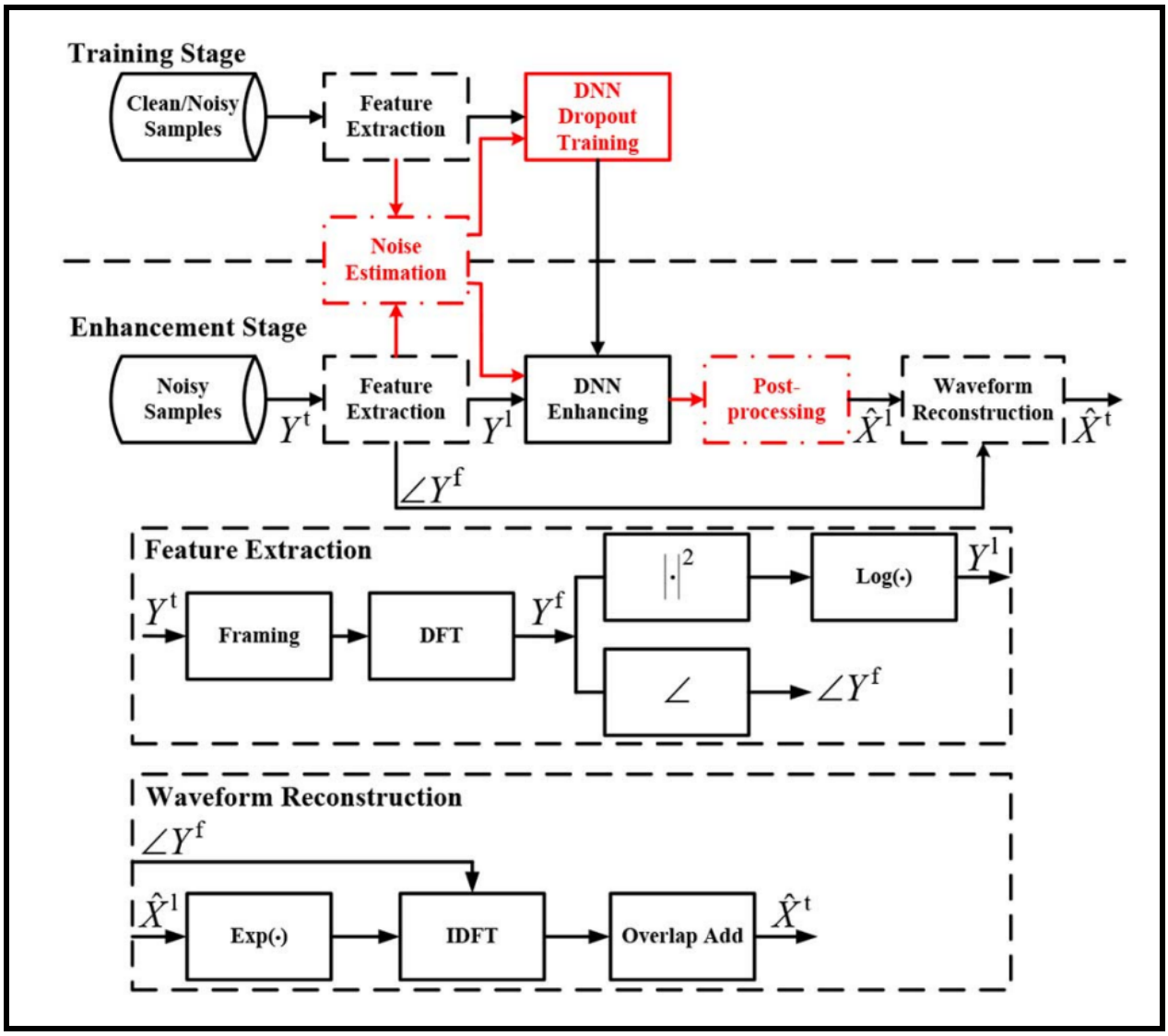

Figure 1. A block diagram of the proposed DNN-based speech enhancement system (Xu et al.).

Dropout training was also used to improve the model's generalization capability, and NAT was adopted to provide the DNN with additional noise information for better prediction accuracy (Xu et al.). Dropout training allowed the DNN to be robust to several different types of noise, and this allows the DNN to effectively distinguish between speech and noise during the enhancement stage. To further enhance the DNN model's performance, an equalization of the global variance (GV) was employed as part of the post-processing. Global variance equalization reduces the over-smoothing problem which is often observed in DNN outputs, and over-smoothing can make the enhanced speech sound muffled or unnatural. An equalization of GV works by adjusting the variance, a measure of how spread out the speech signal is, of the DNN's output to match the variance of the original, clean speech signal, scaling the DNN’s output higher or lower to closely resemble that of natural speech.

The final section of the enhancement stage involves transforming the frequency denoised sound data into the time domain, via an exponential function and an inverse fourier transform. Overlap-add is used to ensure continuity between successive frames and to reconstruct a continuous time signal, and the end result is an enhanced wav file. An enhanced speech file is now extracted from the original mixed noise signal.

Evaluation Metrics

The model's performance was evaluated using objective metrics like Short-Time Objective Intelligibility (STOI) and Perceptual Evaluation of Speech Quality (PESQ). The STOI is obtained by comparing the correlation coefficient of the enhanced, denoised files with the clean speech files, to determine the model's performance. The clean speech files are kind of being used as ground truth values to evaluate the model. For STOI, the values fall in a range of 0-1. PESQ returns perceptual evaluation of the speech quality obtained by comparing the enhanced audio files with that of the clean speech files. For PESQ, the values fall in a range of -0.5 to 4.5. These metrics return efficient assessment of the intelligibility and quality of the model when compared to the clean speech files.

Results

Table 1. PESQ and STOI Results for Raw Noisy Data, Enhanced LogMMSE Data, and Enhanced DNN .wav Files

	Noisy	logMMSE (State of the Art)	DNN
-5dB	STOI: 0.372 PESQ : 1.102	STOI: 0.286 PESQ: 1.169	STOI: 0.136 PESQ: 1.143
0dB	STOI: 0.350 PESQ : 1.081	STOI: 0.280 PESQ: 1.072	STOI: 0.130 PESQ: 1.154

After training the model with the dataset we are utilizing, the evaluation metrics for each model was evaluated at the two distinct SNR levels (-5dB and 0dB). Furthermore, to assess the results of each model, their performance was compared to that of unprocessed noisy files. From Table 1, it can be observed that for both -5dB and 0dB, the STOI values of the regular noisy files returned the higher values when compared to the other models, with the DNN model having the lowest STOI values. Additionally, for the PESQ values, the LogMMSE had the highest value for SNR level of -5dB and the DNN had the highest value for SNR level of 0dB.

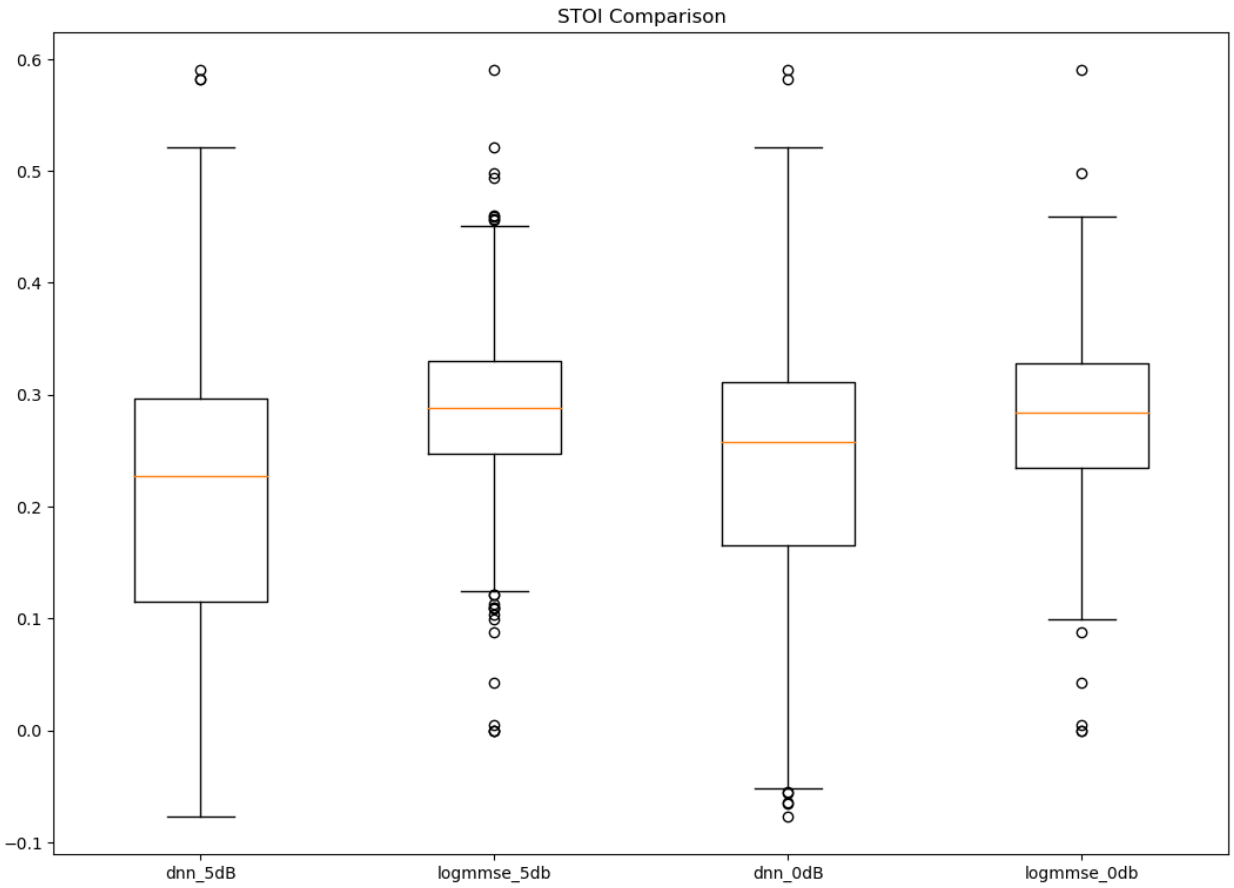

Figure 2. Five number summary of the STOI results for the -5dB and 0dB for state of the art LogMMSE model and the proposed DNN model

The STOI results for both the state of the art LogMMSE model and proposed DNN model are summarized in the box-and-whisker plot in Figure 2. It can be noted that the LogMMSE model had an overall higher median and more compressed interquartile range (IQR) range for both SNR levels. Additionally, the LogMMSE model also had much more outliers in the data. The whiskers of the DNN model had much more variability, as indicated by their wide range. The DNN for -5dB has less outliers than the DNN for 0dB. In addition to this, the median STOI is lower for the DNN with the 5dB SNR compared to the DNN with the 0dB SNR. Similarly, the median STOI is lower for the LogMMSE with the 5dB SNR compared to the DNN with the 0dB SNR. Additionally, the LogMMSE for 0dB has less outliers than the LogMMSE for -5dB. The outliers of the LogMMSE for -5dB SNR are grouped quite closely to the data compared to the LogMMSE for -5dB.

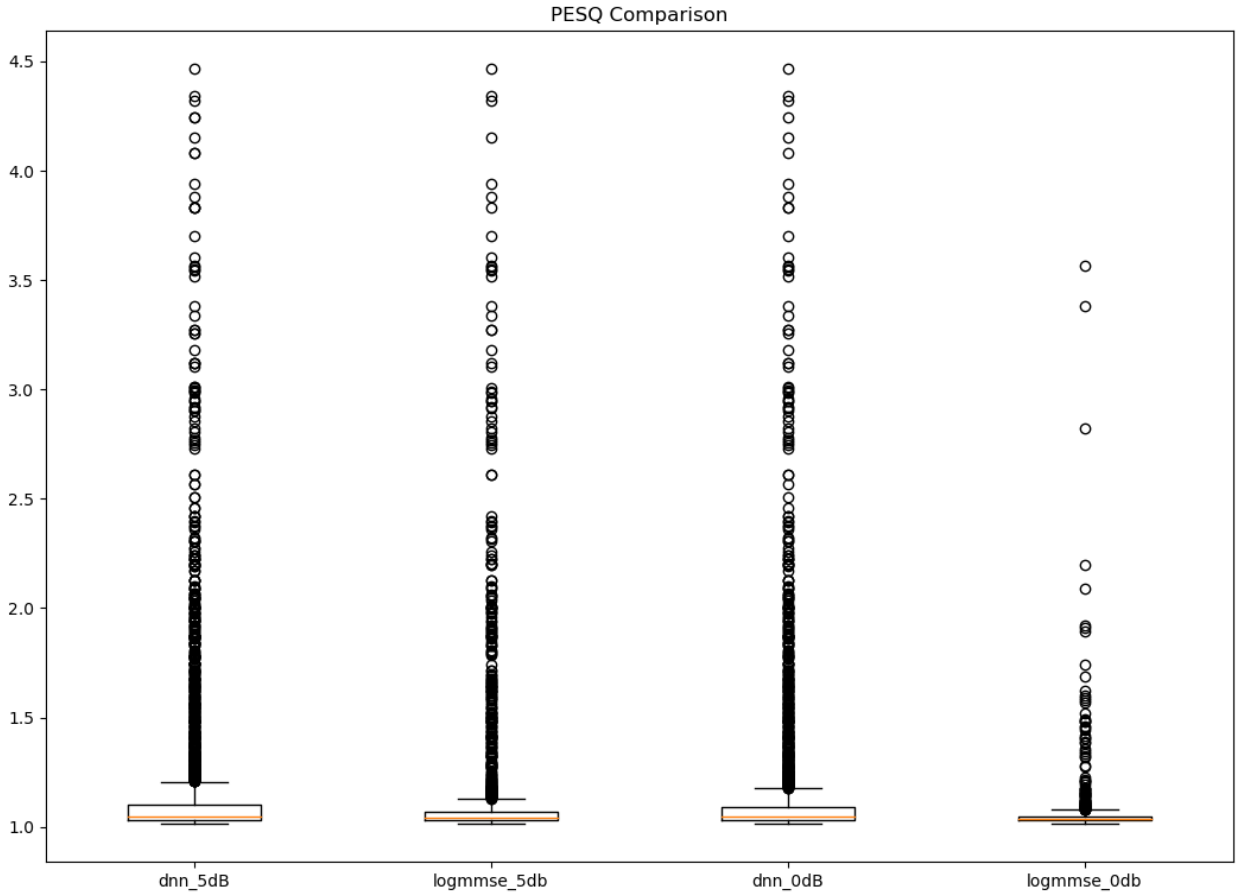

Figure 3. Five number summary of the PESQ results for the -5dB and 0dB for state of the art LogMMSE model and the proposed DNN model with outliers

The LogMMSE model and the DNN model PESQ values are summarized using a box-and-whisker plot, as displayed in Figure 3. There are an extensive number of outliers associated with each model, all of which are greater than the median. With this, there are no outliers lower than the bound of the box and whisker plot for any of the 4 instances. Besides that, for the SNR level at -5, it can be noted that while the DNN model had a higher median value, it also had a larger IQR range indicating more variability. Similar behavior can be observed for the 0dB SNR level. Additionally, it can be viewed that there are less outliers for the LogMMSE model than the DNN model, for both SNR levels. Lastly, all of the diagrams are positively skewed.

Discussion

Based on the results mentioned from computing the model with the dataset we are utilizing, we conclude that the DNN model performed worse than the state of the art method LogMMSE model and the regular noisy files. This is an interesting finding compared to what the original paper was able to generate through their own dataset, which was that the DNN model performed

significantly better than anything else. After further investigation, a major difference between our training of the model and the paper's is in the number of time spent training the data. The paper spent 650 hours worth of data training to obtain their results[6]. When putting this amount of data trained by the paper in comparison to our ~3 hours worth of data training, the results start making more sense. Furthermore, one major consequence of under training the data is over smoothing of the enhanced files. This simply means that the model did not have enough training to be able to distinguish between speech and noise. As such, when it tried to denoise the data, it ended up removing some parts of the speech as well, resulting in an overall increased noise in the files. This would explain why the evaluation metrics for the model with our data returned worse values when compared with both the noise and LogMMSE models.

With that said, The main shortcomings of this model that we discovered is how heavily it is reliant on very large datasets in order for it to even generate an enhanced speech file. There should be a minimum size of dataset for the model to not generate even more noise by over smoothing, which is the other major disadvantage of using this model. Building on that thought, the model also consumes a very long time for it to train and run, which might make it a bit less applicable. Besides that, for both datasets, the one used in paper and our dataset, the model was trained solely in the English language, making it harder to evaluate its efficiency when used for different languages.

On the other hand, the model does have some advantages that makes it more desirable and applicable, given proper and intensive training. One of its main advantages is that it can handle very large datasets. This means that the larger the training set that is being fed into the model, the better the model will perform up to a certain degree. In addition to that, in the paper, the proposed DNN model was found to perform well for non-stationary noises in the real-world[6], which is crucial for the model if it would be implemented into real-life applications. Also, the way the preprocessing is handled in the model resembles that of the way humans decode sound through utilizing frequency-based decoding, making it applicable with cochlear implants. Besides that, the model battles overfitting by utilizing dropout training to ensure no neurons in the model end up over relying on other neurons. Lastly, in the paper, they also found that the model was fairly robust to noise given the limited noise types they used for mixing with speech files. This is interesting to note as it shows the versatility of the model for a large array of noise types only given a limited sample of it.

Overall, the model is very promising, but has a lot of areas for improvement. One thing that can be improved is the data being fed into the model for training purposes. Increasing the size of the dataset would reduce the overall margins of error and emphasize on the patterns between speech and noise files. This is crucial as that would also help eliminate over smoothing of the mixed audio files. Also, expansion of the model for speech data from varied dialects could allow for more diverse applications. Another thing that can be improved is the model itself. We could do

this by adopting a Gammatone filterbank that would allow for better simulation models of human cochlea. We could also implement a Multi-Resolution CochleaGram (MRCG) which captures local and contextual information. Lastly, we could use a Dynamic noise adaptation scheme, which would improve tracking of non-stationary noises. Future application of this research could be seen in many industries. One industry is film entertainment where this model could be used to filter out unwanted noises captured in dynamic scenes. Another industry is medical, where this model could integrate into hearing aids and cochlear implants to enhance noise reduction.

Conclusion

Our implementation of the paper's model showed that the DNN model did not perform better than the LogMMSE when implemented on our dataset. This can be seen from our results where we recorded lower STOI and PESQ metrics for both denoising models compared to the noisy data. We believe that this is due to overprocessing of the noise and our small dataset. To explain, our training set was roughly 3 hours long, compared to the original studies 650 hours training set. Regardless, future optimization could lead to implementation of this model in many industries, if proper improvements were added.

References

1. World Health Organization. Deafness and hearing loss. www.who.int. Published 2023. https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss

2. State of Hearing Loss in America | NCOA.org. National Council on Aging. Accessed September 23, 2023. https://www.ncoa.org/adviser/hearing-aids/hearing-loss-america/#:~:text=About%2015.5%25%20of%20American%20adults

3. NIH. Quick Statistics About Hearing. NIDCD. Published August 18, 2015. https://www.nidcd.nih.gov/health/statistics/quick-statistics-hearing#:~:text=About%202%20to%203%20out

4. Types of Hearing Loss. www.hopkinsmedicine.org. Published November 1, 2022. https://www.hopkinsmedicine.org/health/conditions-and-diseases/hearing-loss/types-of-hearing-loss#:~:text=about%20each%20type.-

5. Asbai N, Zitouni S, Bounazou H, Yahi A. Noisy speech enhancement based on correlation canceling/log-MMSE hybrid method. Multimed Tools Appl 2023;82:5803–21. https://doi.org/10.1007/s11042-022-13591-8.

6. Xu Y, Du J, Dai L-R, Lee C-H. A Regression Approach to Speech Enhancement Based on Deep Neural Networks. IEEE/ACM Transactions on Audio, Speech, and Language Processing 2015;23:7–19. https://doi.org/10.1109/TASLP.2014.2364452.

7. OpenSLR. (n.d.). Room Impulse Response and Noise Database [Audio database]. Retrieved from https://www.openslr.org/28/#:~:text=Summary%3A%20A%20database%20of%20simulated,rate%20and%2016%2Dbit%20precision.

8. Hopkins, C., Graetzer, S., & Seiffert, G. (2019). ARU speech corpus (University of Liverpool) [Data Collection]. University of Liverpool. DOI: 10.17638/datacat.liverpool.ac.uk/681

Musiek, F., & Niemczak, C. (2022, August 15). Part 1 – Gap Detection: The Past, Present, and Future. Hearing Health Matters. Retrieved from https://hearinghealthmatters.org/pathways-society/2022/part-1-gap-detection-the-past-present-and-future