Survey Paper on
Noise Cancellation and Speech Enhancement using Digital Audio Processing Techniques

Ayush Pathak
Student, BE Computer,
Marathwada Mitra Mandal's
College of Engineering
ayushpathak2017.comp@mmcoe.edu.in

Bhakti Kulkarni
Student, BE Computer,
Marathwada Mitra Mandal's
College of Engineering
bhaktikulkarni2017.comp@mmcoe.edu.in

Juilee Niphadkar
Student, BE Computer,
Marathwada Mitra Mandal's
College of Engineering
juileeniphadkar2017.comp@mmcoe.edu.in

Riya Thorbole
Student, BE Computer,
Marathwada Mitra Mandal's
College of Engineering
riyathorbole2017.comp@mmcoe.edu.in

Under the Guidance of:
Mrs Pooja Dhule,
Asst. Prof, Department of Computer Engineering,
Marathwada Mitra Mandal's
College of Engineering

Abstract- Hearing aids are one of the most commonly used machines which work actively to support human functioning and elevate quality of life for millions of people over the world. Even though hearing aid technology has progressed leaps and bounds over the past few decades, the experience of hearing aids still isn't seamless. Problems like "The Cocktail Party Problem" are still to be solved to a satisfying extent. This paper proposes to use the latest advancements in audio processing techniques to create a pipeline that will take in raw audio from any kind of noisy environment, clean it, enhance the speech in the audio and relay it to the user in real time.

Keywords- Audio Processing, Noise Cancellation, Speech Enhancement, Hearing Aids, Deep Neural Networks, Spectral Domain

I. INTRODUCTION

The most commonly used hearing aids have three main components, the microphone, the amplifier, and the speaker. In the most basic of hearing aids, all sounds picked up by the microphone are simply made louder, and replayed to the user via the speaker. Such amplification of all sounds regardless of source or pitch causes major discomfort to the user, forcing them to adjust the speaker volume incessantly.

In addition to that, a problem that is faced by even the high-end Hearing Aid technology is known as the Busy Restaurant Dilemma. When the user is in an environment with multiple speakers coupled with background noise, primitive noise cancellation turns out to be insufficient. In these cases, it becomes necessary for the hearing aid to pinpoint whether the speaker intends to speak to the user of hearing aid.

The goal of this project is to create a device that solves these issues using Machine Learning Techniques.

The proposed system will have 3 major components:

I. Hardware

The enhanced hearing aid will need to have 3 basic components, a microphone (to capture all sounds around the user), a processor (to filter and amplify sounds), and a speaker (to deliver the processed sounds to the user). To realize selective speech amplification, it is necessary to detect whether the speaker is speaking directly to the user. To achieve that, two microphones on either ear can determine the spatial position of the speaker with respect to the listener. The captured sound can then be transformed by the processor and finally played to the user via the speakers.

II. Audio Pre-processing

The audio collected by the device then has to be processed by segregating it according to the sources of the sound. Background noises can then be eliminated and speech can be separated by its speaker.

Traditionally, speech separation is studied as a signal processing problem. A more recent approach formulates speech separation as a supervised learning problem, where the discriminative patterns of speech, speakers, and background noise are learned from training data. The recent introduction of deep learning to supervised speech separation has dramatically accelerated progress and boosted separation performance.

III. Selective Amplification

Finally, using segregated speech and spatial information, selective amplification can be applied, in an attempt to solve the “Cocktail Party Problem”.

II. LITERATURE SURVEY

[1] A Soft Thresholding-based Approach discusses denoising of audio data using image processing and compares three methods namely - The Soft Thresholding Method, The Hard Thresholding Method and The Fft Thresholding Method. The study concludes that The Soft Thresholding Method achieves the best performance with the largest SNR improvement, The Hard Thresholding Method has a smaller improvement, and The Fft Thresholding Method hardly suppresses noise.

[2] In an overview of a challenging problem in auditory perception, the cocktail party phenomenon, they discuss the three major computational methods aimed at solving The Cocktail Party Problem namely - (1) blind source separation (BSS) and independent component analysis (ICA), (2) temporal binding and oscillatory correlation, and (3) cortronic network.

[3] A supervised learning approach to monaural segregation of reverberant voiced speech is proposed. A major source of signal degradation in real environments is room reverberation. This paper tries to improve upon spectral subtraction, Wiener filtering, minimum mean square error (MMSE) estimation, and subspace analysis by dealing with a general acoustic background. A reverberant mixture is processed in a three-stage system. The first stage extracts pitch-based features within each time–frequency unit. In the second stage, multilayer perceptron (MLP) is trained in every channel to associate those features with the grouping cues. T–F units are then labeled according to a criterion based on the MLP output. The last stage performs segmentation and grouping.

[4] Applies convolutional deep belief networks to audio data and empirically evaluates them on various audio classification tasks. In the case of speech data, it shows that the learned features correspond to phones/phonemes. The proposed process includes, conversion of time-domain signals into spectrograms, training on unlabeled speech dataset, visualization of the first layer bases, showing how the bases relate to the phonemes by comparing visualization. Applications of this method relevant to our project are speaker identification, gender classification, phone classification.

[5] This paper presents a historical review about some speech estimation techniques and states the difference between their theoretical background. This paper compares 6 speech enhancement techniques after providing them common conditions of noise variance and speech to noise ratio. The conclusion of this paper is that Wiener filter and Lotter’s method gave comparatively good results in comparison to the spectral subtraction method.

[6] In this study, a novel SNR-based progressive learning framework to improve the performance of regression DNN based speech enhancement in low SNR environments is proposed. The direct mapping from noisy to clean speech is decomposed into multiple stages with SNR increasing progressively by guiding hidden layers in the DNN architecture to learn targets explicitly. The effectiveness of the proposed framework in single-SNR and multi-SNR training conditions is tested under three unseen noise environments. Experimental results demonstrate that this approach can effectively improve the enhancement performance and reduce parameters by 50% when compared with the conventional DNN approach.

[7] A two-stage algorithm can be used to deal with the confounding effects of noise and reverberation separately, where denoising and dereverberation are conducted sequentially using deep neural networks. They employ a DNN with 3 hidden layers to predict the IRM to remove the noise from noisy and reverberant speech. Secondly, we use the IRM-processed magnitude spectrogram of noisy and reverberant speech for feature extraction to train the dereverberation DNN. Finally, after two DNN sub-systems are utilized to perform denoising and dereverberation separately, a coherent system is formed by joint optimization.

[8] A speech enhancement algorithm based on single- and multi-microphone processing techniques estimates a time-frequency mask which represents the target speech and uses masking-based beamforming to enhance corrupted speech. Experimental results show that, as a frontend, the proposed algorithm greatly improves ASR performance. The ASR results significantly outperform the current best system on the CHiME-3 dataset . However, this ASR system is relatively simple, and includes no speaker adaptation. By including advanced techniques in ASR, we believe that its performance can be further improved.

[9] The paper titled - "WAVE-U-NET: A MULTI-SCALE NEURAL NETWORK FOR END-TO-END AUDIO SOURCE SEPARATION" proposes the Wave-U-Net, an adaptation of the U-Net to the one-dimensional time domain, which repeatedly resamples feature maps to compute and combine features at different time scales. As indicated by their experiments, it outperforms the state-of-the-art based U-Net when trained under comparable settings. They highlight the lack of a proper temporal input context in recent separation and models, which can hurt performance and create artifacts, and propose a simple change to the padding of convolutions as a solution. Cons for this approach have been blending of signals over each other, hence effectively cancelling each other out.

[10] An overview of the research on deep learning based supervised speech separation in the last several years, discusses three main components of supervised separation: learning machines, training targets, and acoustic features. Much of the overview is on separation algorithms where they review monaural methods, including speech enhancement (speech-nonspeech separation), speaker separation (multi-talker separation), and speech dereverberation, as well as multi-microphone techniques. This paper also discusses The Cocktail Party Problem and suggests a different, concrete measure: a solution to the cocktail party is a separation system that elevates speech intelligibility of hearing-impaired listeners to the level of normal-hearing listeners in all listening situations. However, the DNN based speech enhancement described in the paper has met the criterion in limited conditions only.

[11] In this paper, three system designs that apply the idea of smart hearing aids are proposed. The meaning of smart hearing aids is that the hearing aid would have the ability to

detect important noise: fire alarm, car horn, etc. and make it audible, so as to avoid catastrophic events that might happen to the hearing impaired if these noises are not heard. The first proposed solution for the integrated speech enhancement and alerting system is a DNN based speech enhancement network, re-trained with different training set input in order to add the noise classification smart feature. In the second proposed system, a CNN classifier is added to work in parallel with the speech enhancement network. In the third proposed solution there are two main processes that the system performs in parallel: speech enhancement and desired noise enhancement. In this system, instead of using a classifier to classify the noise as desired, or undesired, a desired noise enhancement algorithm is added. The second and third system achieved a better performance due to the fact that they carried out the speech enhancement and noise classification processes in two separate independent networks, so the speech quality is not affected. Moreover, they are more flexible than the first system, as we can use any speech enhancement approach to reduce network complexity, or enhance the system performance.

[12] A causal system to address deep learning based speech separation incorporates a convolutional recurrent network (CRN) and a recurrent network with long short-term memory (LSTM). The CRN takes the real and imaginary spectrograms of input signals , while LSTM2 takes the magnitude spectrograms of them as input features. Evaluation results of this paper show that the proposed method effectively removes acoustic echo and background noise in the presence of nonlinear distortions for both simulated and measured room impulse responses.

[13] An all-inclusive survey on unsupervised single-channel speech enhancement algorithm divides 10 speech enhancement algorithms into 5 classes for not only 0db SNR like the survey paper titled “Single Channel Speech Enhancement Techniques in Spectral Domain” but also 10db and 15db variants in 4 types of noise environments making the comparison much more extensive. The conclusion of the survey results indicates no single algorithm is categorized as the best, and several speech enhancement algorithms performed equally well across SNRs situations and noise types. In terms of the speech quality, MMSESPU, LMMSE, WF, EMD-H and MBSS performed equally well across all SNRs situations. The MMSE-SPU, LMMSE, MBSS and WF performed especially well for speech intelligibility.

III.DISCUSSION

Title	**Overview**	**Limitations**
The cocktail party problem.	In an overview of a challenging problem in auditory perception, the cocktail party phenomenon, they discuss the three major computational methods aimed at solving The Cocktail Party Problem namely - (1) blind source separation (BSS) and independent component analysis (ICA), (2) temporal binding and oscillatory correlation, and (3) cortronic network.	It provides a primitive yet arguably con-vincing basis for how the human auditory system solves the cocktail party problem. However, in its present form, the learning rule is not equipped to deal with questions pertaining to computational auditory scene analysis.
Spectrogram enhancement algorithm: A soft thresholding-based approach.	A Soft-Thresholding based approach discusses denoising of audio data using image processing and compares three methods namely - The Soft Thresholding Method, The Hard Thresholding Method and The Fft Thresholding Method.	It is not difficult to reverse the process and generate a copy of the original signal from a spectrogram
Dual-Signal Transformation LSTM Network for Real-Time Noise Suppression	This paper introduces a dual-signal transformation LSTM network (DTLN) for real-time speech enhancement as part of the Deep Noise Suppression Challenge (DNS-Challenge). This approach combines a short-time Fourier transform (STFT) and a learned analysis and synthesis basis in a stacked-network approach with less than one million parameters.	They assume that STFT features provide a higher robustness for noisy input since phase information - which is not useful in high-noise conditions
Divide and Conquer: A Deep CASA Approach to Talker-independent Monaural	This paper addresses talker-independent monaural speaker separation from the perspectives of deep learning and	Despite considerable effort, monaural (single-microphone) algorithms

Speaker Separation	computational auditory scene analysis (CASA). The proposed deep CASA approach optimizes frame-level separation and speaker tracking in turn, and produces excellent results for both objectives.	capable of increasing the intelligibility of speech in noise have remained extremely hard to achieve.

The major problem in digital audio processing for hearing aids is described in the paper titled, " The Cocktail Party Problem" which goes into detail about the said problem in theoretical terms and defines the units of measurement and standards to be maintained in the experimentation environment. The first approach we considered for noise cancellation was from the paper titled, "Spectrogram enhancement algorithm: A soft thresholding-based approach." which proposes transforming audio data into spectrograms and applying image processing algorithms on them. After implementing this approach we found that reconversion of the spectrograms into audio causes major losses making the gains from noise cancellation insignificant.
The second and more natural approach turned out to be in the paper titled, "Dual-Signal Transformation LSTM Network for Real-Time Noise Suppression" which deals with the noise in the input audio by transforming and retransforming it using Short Term Fourier transforms and using thresholds of noise allowance. This approach is much faster and more reliable than the spectrogram approach due to its low overheads and large availability of STFT algorithms.

In order to achieve speaker specific enhancement, first it is necessary to split the human audio by the speaker who is generating it, in order to achieve this we refer to the paper titled, "Divide and Conquer: A Deep CASA Approach to Talker-independent Monaural Speaker Separation". This paper proposes a deep learning solution to the speaker separation problem.

IV.CONCLUSION

In conclusion, this paper focuses on finding the most appropriate noise cancellation and speech enhancement algorithms that can work together efficiently. We have studied various research and survey papers which focus on solving these problems using mainly deep learning solutions. The aim of this paper is to find algorithms that will become part of a pipeline which can take in any kind of noisy audio, clean it, separate the speech and enhance it to achieve a smooth hearing aid experience.

IV. REFERENCES

[1] Liu, Bin, Yuanyuan Wang, and Weiqi Wang. "Spectrogram enhancement algorithm: A soft thresholding-based approach." Ultrasound in medicine & biology 25.5 (1999): 839-846.

[2] Haykin, Simon, and Zhe Chen. "The cocktail party problem." Neural computation 17.9 (2005): 1875-1902.

[3] Jin, Zhaozhang, and DeLiang Wang. "A supervised learning approach to monaural segregation of reverberant speech." IEEE Transactions on Audio, Speech, and Language Processing 17.4 (2009): 625-638.

[4] Lee, Honglak, et al. "Unsupervised feature learning for audio classification using convolutional deep belief networks." Advances in neural information processing systems 22 (2009): 1096-1104.

[5] Kawamura, Arata, Weerawut Thanhikam, and Youji Iiguni. "Single channel speech enhancement techniques in spectral domain." ISRN Mechanical Engineering 2012 (2012).

[6] Gao, Tian, et al. "SNR-Based Progressive Learning of Deep Neural Network for Speech Enhancement." INTERSPEECH. 2016.

[7] Zhao, Yan, Zhong-Qiu Wang, and DeLiang Wang. "A two-stage algorithm for noisy and reverberant speech enhancement." 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2017.

[8] Zhang, Xueliang, Zhong-Qiu Wang, and DeLiang Wang. "A speech enhancement algorithm by iterating single-and multi-microphone processing and its application to robust ASR." 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2017.

[9] Stoller, Daniel, Sebastian Ewert, and Simon Dixon. "Wave-u-net: A multi-scale neural network for end-to-end audio source separation." arXiv preprint arXiv:1806.03185 (2018).

[10] Wang, DeLiang, and Jitong Chen. "Supervised speech separation based on deep learning: An overview." IEEE/ACM Transactions on Audio, Speech, and Language Processing 26.10 (2018): 1702-1726.

[11] Nossier, Soha A., et al. "Enhanced smart hearing aid using deep neural networks." Alexandria Engineering Journal 58.2 (2019): 539-550.

[12] Zhang, Hao, Ke Tan, and DeLiang Wang. "Deep Learning for Joint Acoustic Echo and Noise Cancellation with Nonlinear Distortions." INTERSPEECH. 2019.

[13] Saleem, Nasir, Muhammad Irfan Khattak, and Elena Verdú. "On Improvement of Speech Intelligibility and Quality: A Survey of Unsupervised Single Channel Speech Enhancement Algorithms." International Journal of Interactive Multimedia & Artificial Intelligence 6.2 (2020)