CHAPTER 1

INTRODUCTION TO SPEECH ENHANCEMENT

1.1 SPEECH COMMUNICATION AND PRODUCTION

Communication involves various methodologies such as sign language, facial expressions, gestures, and postures. In verbal communication, information is shared through writing or speaking. Speech is the most convenient and efficient way to express one's feelings and thoughts. Communication does not end when the person speaks; it ends when the speech is listened to by the listener without any disturbance. Communication stems from the act of speech that plays a considerable role in reality and helps to understand the emotions of others. In contrast, other modes of communication, like messages, a letter cannot do that. Speech is an advanced form of communication between two or more people. Speech aids in the peaceful resolution of conflicts in society, as well as the communication of messages and the presentation of critical ideas and structures of communication.

Speech is produced by an air stream that originates from the lungs, and the air is pushed upward through the trachea (windpipe) and oral and nasal cavities. It is during the passage of the air stream that various organs of speech modify it. These modifications produce different acoustic effects that are used to differentiate sounds. Four separate but interrelated processes are needed to produce a speech sound. First, airflow must be initiated in the lungs, then phonated in the larynx by moving the vocal folds. Then, the velum directs the air through the oral or nasal cavity to create sounds. Finally, its articulation occurs inside the oral cavity, as shown in Figure 1.1.

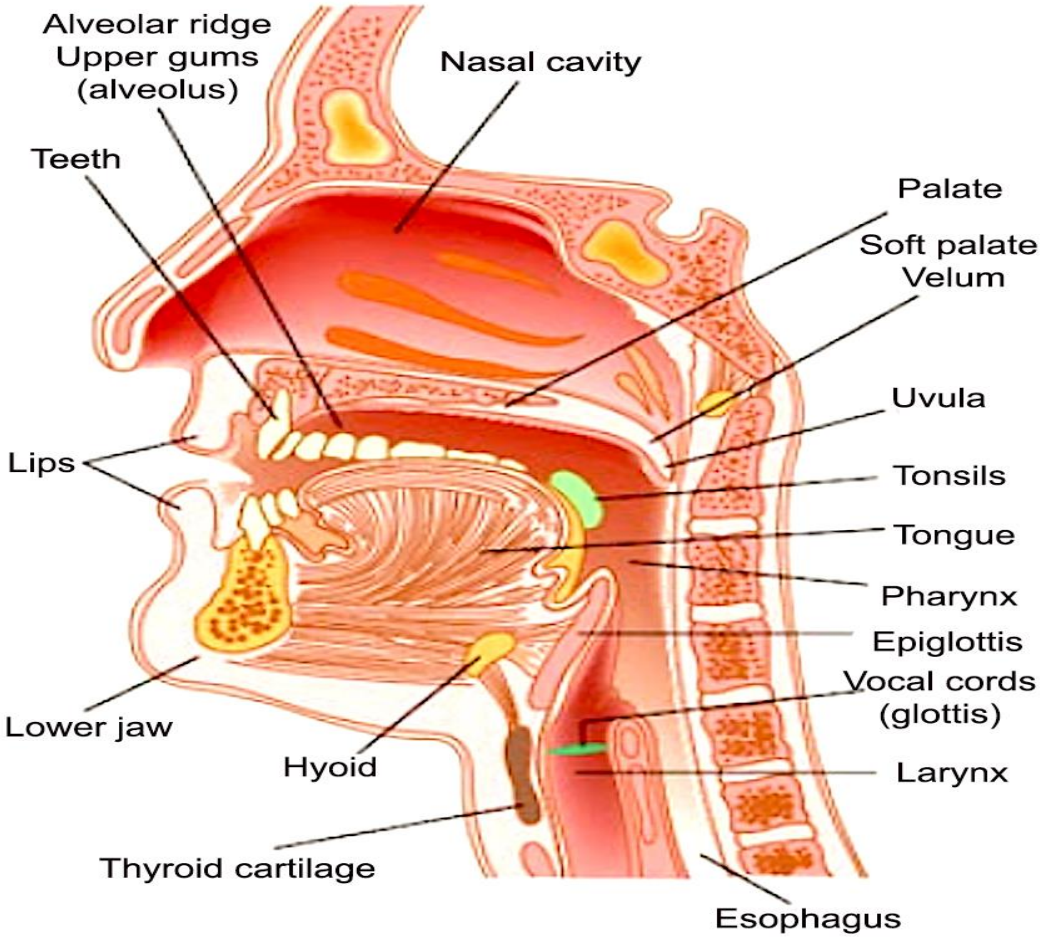

Figure 1.1 Organs involved in the Production of Speech

(Image Courtesy: MIT Open Courseware)

A speech signal is made up of a sequence of sounds. It is these sounds and the transition of sounds that represent the information. Speech applications can generally be divided into inter-human communication and human-machine interface. Acoustic communication between two people can be considered in two cases. Communication seems easy and accurate when people are close together in a noise-free environment. However, in another situation, they are far apart, and the background is noisy, which makes it difficult for the listener to understand. Speech in human-machine interfaces is generally recorded in electrical form, and the transmission media usually introduce distortion, making speech noisy. When the quality and intelligibility of speech are low, it poses more challenges.

In many cases, noise-induced distortion frequently lowers the quality or clarity of speech. Thus, it becomes necessary to use a speech enhancement system to remove noise from noisy speech. Since noise degrades the performance of speech recognition systems, these systems need to incorporate speech enhancement. In speech enhancement, the goal is to reduce noise levels, increase speech clarity, and reduce auditory fatigue by enhancing the intelligibility and quality of speech.

1.2 FEATURES OF NORMAL SPEECH

Everyday speech encompasses clear articulation, language phonology, and intelligibility. Appropriate pitch and tone convey meaning and emotion, while volume and rate adapt to the context. Pitch variations convey emotions and even gender distinctions, while tone modulation ensures speech aligns with the intended message. Intensity refers to the loudness or strength of speech and plays a significant role in conveying various aspects of a message. In normophonic speech, the accurate production and manipulation of formants contribute to clear and distinct vowel sounds, making words and speech more intelligible. Therefore, a clear and resonant voice quality enhances the effectiveness of communication.

1.3 SPEECH DISORDER

Speech is the process adopted to convey meaning to the listener by producing specific sounds. It is a speech disorder when a person cannot produce sounds that create words. The most effective way of communicating thoughts, ideas, and feelings is by speech. Speech disability is a problem that arises when articulating words. People need help in forming the speech sounds that are required for communication. It is possible to suffer from a speech disorder at any age. Some speech disorders include oesophageal speech, stuttering, Aphasia, and dysarthria. Recognizing a person's speech with any of these speech disorders is tiring as it creates lots of effort for the listener to recognize their speech. Therefore, speech enhancement plays a vital role before speech recognition is performed.

Different types of speech disorders make people experience different difficulties, such as adding sounds to words, pronouncing words, speaking with a raspy or hoarse voice, and soft speech. Some of the speech disorders are discussed below:

1.3.1 Oesophageal Speech

In many cases, laryngeal cancer can be treated conservatively through mouth surgery, chemotherapy, and radiation. A laryngectomy is performed when these treatments do not yield many results or cancer has advanced to the point where a normal function would be impaired. The larynx is removed during a laryngectomy, which divides the airways from the mouth, nose, and oesophagus. Oesophageal speech refers to the speech that is produced after the removal of the larynx. Laryngectomies are also carried out for head and neck malignancies. In addition to voice restoration and oral feeding, post-laryngectomy rehabilitation also focuses on taste and smell rehabilitation. After surgery, a person's quality of life may be negatively impacted. A total laryngectomy removes the larynx and vocal cords, and the way the person speaks changes after the laryngectomy. There will be a change in the person's voice as the voice will no longer be coming from the vocal cords. Oesophageal speech is a type of speech that is also called alaryngeal speech.

1.3.2 Stuttering

Stammering or stuttering is a condition that regularly and severely impairs the fluency and flow of speech. It is sometimes referred to as childhood-onset fluency disorder. People who stutter have difficulty expressing themselves despite knowing what they want to say. During communication, the person with stuttering may stop, repeat, or lengthen a word, syllable, vowel sound, or consonant. Young toddlers stutter as they begin to talk. Young children may stutter because they cannot keep up with the words they want to say. After all, their language and speech abilities need to be developed more. Children generally overcome this type of developmental stuttering. A chronic condition, stuttering lasts throughout adulthood. As a result, this type of stuttering may negatively affect self-esteem and interpersonal interactions.

1.3.3 Aphasia of speech

Aphasia of speech (AOS) is a neurological disorder that disturbs the brain pathways that coordinate movements to produce speech. The brain knows what it wants to say but cannot arrange and sequence the speech sound actions. The types of Aphasia are as follows:

In most cases, acquired AOS affects adults, although it can occur at any age. Acquired AOS is brought on by injury to the areas of the brain that are responsible for speech. AOS may also result from a head injury, stroke, or other problem affecting the brain. Acquired AOS can occur in conjunction with other disorders caused by nerve damage.

Childhood AOS affects children from their birth. Aphasia of speech, verbal Aphasia, and articulatory Aphasia are other names for Childhood AOS. According to the research findings, several genetic factors may contribute to the disorder.

1.3.4 Dysarthria

Dysarthria is a speech disability produced by muscular weakness. The damage to the nervous system results in motor speech disorders like dysarthria. The types of dysarthria are defined based on the part of the nervous system affected. Central dysarthria occurs due to damage to the brain, and peripheral dysarthria occurs due to damage to organs required for speech production. Developmental dysarthria occurs when the brain is damaged before birth or at birth. Acquired dysarthria may result from brain damage later in a person's life. It may occur due to a stroke, a brain tumor, or Parkinson's disease, and it is more common for adults to have acquired dysarthria.

1.4 MOTIVATION OF RESEARCH

In speech recognition, the signal needs good quality and intelligibility. As speech enhancement helps enhance speech by suppressing the background noise and omitting the distortions of speech sounds, this becomes more beneficial for people suffering from different types of speech disorders. With the speech enhancement process, it will be easier for people to understand the speech spoken by people with speech disorders. The enhancement systems give good results regarding the accuracy rate of words recognized by enhancing perceived speech quality and intelligibility and reducing hearing fatigue.

1.5 SPEECH RECOGNITION

In speech recognition or speech-to-text conversion, a machine identifies the words and sentences spoken loudly and clearly and converts them into readable text. Speech recognition software with a limited vocabulary can identify words clearly when they are spoken. Advanced software may be capable of handling natural speech, accents, and a variety of languages. In speech recognition, computer science, linguistics, and engineering have all played a role in research. Speech recognition refers to the identification of words in spoken language.

Speech recognition systems process and interpret spoken words into text using computer algorithms. The speech signal detected by the microphone is recognized by analyzing the audio, splitting it into parts, digitizing and converting it into a machine-readable format, and implementing a recognition algorithm to match the exact text for the audio detected. A speech recognition system must be robust to adapt itself to speech that is varying in nature. The recognition systems that organize audio into the text are trained on different speaking styles, patterns, accents, and phrasings as part of the training process. The software also incorporates speech enhancement that removes the background noise from the spoken audio accompanying the signal before speech recognition is performed.

A speech recognizer consists of a few components: a speech input, a feature extraction module, a feature vector, and a decoder. A decoder uses acoustic models, pronunciation dictionaries, and language models to determine the appropriate output. An evaluation of speech recognition technology is based on its accuracy rate in recognizing words or Word Error Rate (WER). Background noise, pronunciation, accent, pitch, and volume, can impact word error rates.

1.6 OBJECTIVE OF THE RESEARCH

Current speech enhancement techniques, particularly those employing traditional signal processing methods, often fail to maintain perceptual quality and intelligibility in noisy environments. Although traditional methods successfully remove noise, they are only achievable if they enhance speech quality and intelligibility, rendering the noise removal effort meaningful. Therefore, developing more sophisticated algorithms that effectively balance these two critical aspects and provide consistent results across diverse

acoustic scenarios is essential for advancing speech enhancement applications. The first part of this research work deals with enhancement of speech signals taken from a publicly available database, and the second part with patient data collected from laryngectomees.

The research investigates the efficiency of current algorithms that contribute to speech enhancement, focusing on improving the quality and intelligibility of speech signals. The primary aim is to implement deep learning algorithms for speech signals and compare them with conventional algorithms' performance. Furthermore, the quality and intelligibility of the enhanced speech are assessed through rigorous experimentation and analysis.

A speech enhancement system aims to enhance speech by improving its quality and intelligibility, enhancing the listening experience. Compared to the conventional techniques for speech enhancement, deep learning algorithms learn complex features directly from the speech signals, allowing them to identify and separate speech from noise effectively. It also adapts to diverse and varying noise conditions through extensive training, providing more robust and consistent results. Deep Neural Networks are applied for speech enhancement and denoising in single microphone captures. The algorithms considered for the research study are Deep Fully Connected Neural Networks, Deep Convolutional Neural Networks, Long Short Term Memory Networks (modified LSTM), and Fully Convolutional Recurrent Networks (modified FCRN). Deep Fully Connected Neural Network (DFNN) excels at modeling non-linear transformations and capturing intricate patterns in speech data, enhancing the clarity of the processed output. Deep Convolutional Neural Networks (Deep CNN) leverage spatial hierarchies in the input data, making them particularly effective for identifying local patterns in speech signals. The Long Short Term Memory (LSTM) network can remember patterns over long sequences, making them particularly well-suited for improving speech intelligibility by filtering out noise and preserving speech quality. Fully Convolutional Recurrent Network (FCRN) in speech enhancement combines the strengths of Convolutional Neural Networks (CNNs) for capturing spatial features and Recurrent Neural Networks (RNNs) for modeling temporal dependencies. This allows it to reduce noise while preserving speech details, making it well-suited for real-time speech enhancement applications.

In the first part of the research work, data taken from public databases are exposed to different types of noise at different levels. The noisy speech is enhanced by applying these deep-learning algorithms. The quantitative evaluation of both the quality and intelligibility of enhanced speech is performed through metrics such as Signal to Noise Ratio (SNR), Segmental SNR (segSNR), Scale Invariant Signal to Distortion Ratio (SI-SDR), Perceptual Evaluation of Speech Quality (PESQ), Short-Term Objective Intelligibility (STOI) and Deep Noise Suppression Mean Opinion Score (DNSMOS).

In the second part of the research work, speech recorded from laryngectomee patients implanted with speech prosthesis is considered. Alaryngeal speech, resulting from surgical procedures following prosthesis implantation like Bloom Singer, presents unique challenges in effective communication. The real-time data from patients implanted with a Blom Singer prosthesis was analyzed. Improving the quality and intelligibility of alaryngeal speech is crucial for enhancing the outcomes and quality of life for individuals with laryngeal dysfunction. Speech enhancement techniques are applied, and the paralinguistic features of alaryngeal speech are analyzed to achieve this. The effectiveness of the enhanced speech is validated by calculating the word error rate of speech processed by different algorithms.

The enhancement system could work as a pre-processor to increase the robustness of the speech recognition system. Speech enhancement attempts to reduce distortions and improve the clarity and intelligibility of speech by reducing listening effort and making speech sound more pleasant.

1.7 SPEECH DATABASE

The speech dataset for the speech enhancement system is obtained from the University of Edinburgh, Centre for Speech Technology Research (CSTR) (Valentini-Botinhao, 2017). This dataset comprises 400 speech sentences of both male and female speakers. From the 400 sentences, 320 speech sentences belonging to 84 speakers are taken for training in the first part of this research work. The remaining 80 utterances from the normal speaker dataset is taken for testing. The train and test split ratio is 80:20. In the process of creating noisy speech, the clean speech from the CSTR database is mixed with various types of noises taken from https://www.ee.columbia.edu/~dpwe/sounds/noise/ at different levels of noise, such as -10dB, -5dB, 0dB, 5dB, 10dB, and 15dB. For training and

testing, various noise types were used, such as Washing machine noise, Rainbow noise, Jet airplane noise, Train whistle noise from MATLAB\R2020b\toolbox\audio\samples Babble noise, Airport noise, Street noise, and Restaurant noise from the Columbia database. For evaluating the speech enhancement system, unseen noises such as car and subway noises were taken from https://www.ee.columbia.edu/~dpwe/sounds/noise/.

As per the survey taken from [Laryngeal Cancer - StatPearls - NCBI Bookshelf (nih.gov)], in India, Laryngeal cancer accounted for 13,150 new cases in 2017, constituting approximately one-third of all head and neck cancers, with 3,710 associated deaths. The mean age of patients diagnosed is 65 years, with a higher incidence observed in males than females. Notably, about 98% of laryngeal cancers originate in either the supraglottic or glottic regions, with early-stage cancers demonstrating high curability rates, ranging from 80% to 95%. Based on the latest estimates from the American Cancer Society [Throat Cancer Statistics | Cases of Throat Cancer Per Year | American Cancer Society] for laryngeal cancer, approximately 12,650 new cases are recorded in the U.S, with 10,030 occurring in men and 2,620 in women.

The Blom-Singer voice prosthesis aids laryngectomees in regaining their ability to speak by directing air from the lungs into the esophagus, which vibrates to produce sound. Bionic voice is generated using a prosthesis that restores speech to laryngectomees and can be controlled naturally by messages from the brain to the missing larynx. In the second part of the research work, the speech generated by the Blom Singer voice prosthesis is considered due to the availability of data because many patients have implanted Blom Singer prosthesis. Despite implanting the prosthesis, the alaryngeal speech has low intelligibility. The high prevalence of laryngeal cancer and the need to improve speech intelligibility emphasize the importance of research in enhancing alaryngeal speech.

The alaryngeal speech recordings were collected from 10 laryngectomee speakers, comprising 9 males and 1 female, aged 35 to 60 years. Recordings were conducted in an anechoic chamber using a Condenser Studio XLR microphone with echo cancellation, sampled at 16 kHz. To maintain optimal signal fidelity, the microphone was positioned 10 centimeters from each speaker. The data collected from laryngectomees serves as a valuable resource for examining the nuances of alaryngeal speech, supporting advanced analysis and modeling in speech processing research. Speech samples were selected from the Harvard Sentences corpus, known for its phonetically balanced content, to cover a

broad spectrum of phonetic features. Two sentences "Read verse out loud for pleasure" and "The stray cat gave birth to kittens" were chosen based on the relative ease of articulation for the speakers. Due to the inherent difficulty laryngectomees experience in producing consistent sounds, they took a sip of water to clear the mucus from their throat prior to speaking each sentence. Each speaker repeated these sentences two to three times based on their comfort level, resulting in a total of 54 recordings. 40 speech sentences were taken for training and 10 speech sentences for testing, adhering to 80:20 train-test split ratio. To facilitate robust performance evaluation, noisy speech was generated by combining the clean speech signals with various noise types and levels.

1.8 PERFORMANCE EVALUATION METRICS

The performance of the enhanced speech is evaluated based on the values of SNR, segSNR, PESQ, STOI and SI-SDR for quality and intelligibility. The denoising ability of deep noise suppression algorithms are evaluated by DNSMOS. The evaluation metrics are calculated as follows:

1.8.1 SNR

SNR (L.Wang et al. 2021) is calculated as the ratio of the root mean square of the speech signal to that of the root mean square of the noisy signal. It is expressed in decibels (dB). Equation 1.1 shows the formula for the calculation of SNR.

$$SNR = 20\ \log_{10} \frac{RMS\ of\ Speech\ signal}{RMS\ of\ Noise\ Level} \tag{1.1}$$

1.8.2 segSNR

segSNR (Saleem et al., 2021) is the signal power to the noise power for each segment in a speech signal. It is expressed in decibels (dB). Equation 1.2 shows the formula to calculate segSNR.

$$segSNR = 10 \log_{10} \frac{Sum(e^2)}{Sum(y^2)} \tag{1.2}$$

Where,

e is the Enhanced Speech Signal

y is the Noisy Speech Signal

1.8.3 PESQ

PESQ (Strake et al., 2020) is the subjective measure to determine the quality of the signal. The value is standardized by the International Telecommunications Union (ITU). It is measured on a scale of 0 to 5. It found that the value between 0-1 is called bad, 1-2 is termed poor, 2-3 is fair, 3-4 is Good, and 4-5 is considered excellent quality.

1.8.4 STOI

STOI (Tang et al. 2019) is a measure used to determine speech intelligibility. The value ranges between 0 and 1. A value closer to 1 indicates better speech intelligibility.

The following are the metrics used recently in speech enhancement to evaluate the performance of speech enhancement algorithms. These measures give a clear understanding of the effectiveness of speech enhancement algorithms.

1.8.5 SI-SDR

SI-SDR (Grzywalski & Drgas, 2022; Roy et al., 2021) is an evaluation metric to measure the improvement in the quality of the noise-separated signal compared to the original mixed signal. It is the ratio of the energy of the original source signal to the energy difference between the estimated source signal and the original source signal. It is expressed in decibels (dB). Equation. 1.3 shows the formula for the calculation of SI-SDR.

$$SI\ SDR = 10 \log_{10} \frac{\|x\|^2}{\|e-x\|^2} \tag{1.3}$$

Where,

$\|x\|^2$ – Energy of Clean Speech Signal

$\|e - x\|^2$ – Energy of Enhanced Speech Signal & Clean Speech Signal

e – Enhanced Speech Signal

1.8.6 DNSMOS

DNSMOS (A. Li et al., 2022) is a metric that evaluates the effectiveness of deep noise reduction algorithms in speech enhancement. It is measured on a scale of 1 to 5, where 5 represents the speech of best quality.

1.9 SCOPE OF THE WORK

The proposed work evaluates the performance of the speech enhancement algorithms in terms of quality and intelligibility to select a suitable algorithm for a speech recognition system. Various noises at different levels are added to the clean speech data taken from the Centre for Speech Technology Research (CSTR) database, University of Edinburgh, to analyze the efficiency of the Traditional and deep learning algorithms in noise removal. DNN-based speech enhancement approaches such as DFNN, Deep CNN, LSTM (modified LSTM) and FCRN (modified FCRN) are trained and tested with the data collected from the database.

The DFNN consists of multiple fully connected layers and has the ability to adapt to diverse noise conditions. The Deep CNN network comprising of multiple convolutional layers, each followed by batch normalization and an activation are considered as the technique to effectively enhance speech in different noise conditions. Similarly, LSTM based technique for speech enhancement is designed to optimize performance in enhancing noisy speech signals, with adjustments made to layer depth and network components. The performance evaluation based on adding more layers needs to be explored. Further the research work, explores the development of FCRN for speech enhancement, combining convolutional and recurrent layers to capture both local and long-term features. The network architecture is designed to balance speech quality and computational efficiency, with adjustments made to the depth of the convolutional and recurrent layers. Modifications in the layers is explored for better performance. The speech enhancement model among the four algorithms that demonstrate the best performance is ultimately chosen.

In today's world, people affected with laryngeal cancer are approximately 3 to 6% of all cancers. Laryngectomy is a surgical procedure involving the removal of the larynx, which contains the vocal cords. Patients undergoing this surgical procedure can no longer

speak the same way. The only option for laryngectomy patients to regain their speaking ability is through the implanted voice prosthesis. The speech is poor in quality and intelligibility and is embedded with background noise in real-world scenarios. SNR, segSNR, SI-SDR, PESQ, and STOI are computed to estimate the quality and intelligibility of the enhanced speech signal. DNSMOS evaluates the ability of deep noise suppression by deep learning algorithms.

The main contribution of this work is to identify an efficient DNN-based algorithm for speech enhancement, which is an essential step in a speech recognition system.

The novelty of this work is the creation of a unique resource consisting of alaryngeal speech with sentences spoken by subjects post-laryngectomy.

1.10 ORGANIZATION OF THE THESIS

This research work is elaborated in the remaining chapters of the thesis. This thesis is organized into seven different chapters:

Chapter 1 describes the speech communication and production process, features of normal speech, speech disorders, motivation of the research, speech recognition, objective of the research, speech database, performance evaluation metrics of speech, and scope of the work.

Chapter 2 interprets the literature review of the research work, focusing on the available concepts in the literature and related works.

Chapter 3 focuses on the traditional technique adopted for speech enhancement.

Chapter 4 explains the design and implementation of Deep Learning algorithms for speech enhancement.

Chapter 5 emphasizes the speech generation and speech changes caused by Laryngectomy.

Chapter 6 deals with alaryngeal speech enhancement, MATLAB Application for alaryngeal speech enhancement, and analysis of the paralinguistic features.

Chapter 7 summarizes the research and concludes with the best algorithm for enhancing alaryngeal speech.

This chapter is followed by an extensive literature survey of traditional algorithms, ML/DL algorithms, performance metrics, and the influence of different noises on speech enhancement.