8-22-2019 1:45 PM

Parkinsonian Speech and Voice Quality: Assessment and Improvement

Amr Gaballah, *The University of Western Ontario*

Supervisor: Parsa, Vijay, *The University of Western Ontario*

A thesis submitted in partial fulfillment of the requirements for the Doctor of Philosophy degree in Electrical and Computer Engineering

Abstract

Parkinson's disease (PD) is the second most common neurodegenerative disease. Statistics show that nearly 90% of people impaired with PD develop voice and speech disorders. Speech production impairments in PD subjects typically result in hypophonia and consequently, poor speech signal-to-noise ratio (SNR) in noisy environments and inferior speech intelligibility and quality. Assessment, monitoring, and improvement of the perceived quality and intelligibility of Parkinsonian voice and speech are, therefore, paramount.

In the first study of this thesis, the perceived quality of sustained vowels produced by PD patients was assessed through objective predictors. Subjective quality ratings of sustained vowels were collected from 51 PD patients, on and off the Levodopa medication, and 7 control subjects. Features extracted from the sustained vowel recordings were combined using linear regression (LR) and support vector regression (SVR). An objective metric that combined linear prediction and harmonicity features resulted in a high correlation of 0.81 with subjective ratings, higher than the performance reported in the literature.

The second study focused on the prediction of amplified Parkinsonian speech quality. Speech amplifiers are used by PD patients to counteract hyperphonia. To benchmark the amplifier performance, subjective ratings of the quality of speech samples from 11 PD patients and 10 control subjects using 7 different speech amplifiers in different background noise conditions were collected. Objective quality predictors were then developed in combination with machine learning algorithms such as deep learning (DL). It was shown that the speech amplifiers differentially affect Parkinsonian speech quality and that the composite objective metric resulted in a correlation of 0.85 with subjective speech quality ratings.

In the third study, a new signal-to-noise feedback (SNF) device was designed and developed to help PD patients control their speech SNR, intelligibility, and quality. The proposed SNF device contained dual ear-level microphones for estimating the speech SNR, a throat accelerometer for reliable voice activity detection, and visual/auditory alarms when the produced speech was below a certain threshold. Performance evaluation of this device in noisy environments demonstrated significant improvements in speech SNR, perceived intelligibility, and predicted quality, especially in high background noise levels.

Keywords: Parkinson disease, speech quality, machine learning, speech analysis, speech amplifiers, speech-to-noise feedback device

Lay Summary

Nearly 90% of people impaired with Parkinson's disease (PD) develop voice and speech disorders during the course of their disease. Parkinsonian speech is typically accompanied with deterioration in loudness, intelligibility (i.e. how many words of the sentence can the listener understand?), and quality (a multi-dimensional perceptual phenomenon that encompasses attributes such as clarity, pleasantness, and naturalness). Therefore, there is a clinical need to assess the speech quality for people impaired with PD. Traditionally, voice and speech quality are evaluated by a panel of listeners (subjective assessment). While this may be the gold standard, objective estimation of speech quality through computational models is more robust and time and cost-efficient approach. This thesis focuses on the development and evaluation of such objective Parkinsonian speech and voice quality estimators.

In this study's first contribution, multiple objective metrics are developed to assess the quality of the Parkinsonian sustained vowels. First, acoustic features of the vowels' records are extracted to estimate the quality. There is a need to machine learning algorithms to map the acoustic characteristics to the behavioral assessment of speech quality. Investigations in this study led to an automatic quality estimator that predicts the Parkinsonian voice quality with a correlation value (a mathematical measure of similarity) of 0.81 with the subjective scores.

In the second contribution of this study, another set of objective metrics are developed to assess the quality of Parkinsonian running speech. This research study also benchmarked the effect of speech amplifiers on Parkinsonian speech quality. Different acoustic features and various and more sophisticated machine learning algorithms like deep learning are used to extract these objective metrics. The correlation value between the subjective and objective estimations of the Parkinsonian speech quality reached 0.85.

In the third contribution of this study, a new device is developed and presented to help people impaired with PD to enhance and improve their speech quality and intelligibility. This device is designed to be portable, cost-effective, and easy to build. The performance of the device has been tested in noisy environments. The results showed enhancements of the intelligibility and quality of subjects' speech with the help of the device.

Epigraph

"I often say now
I don't have any choice
whether or not I have
Parkinson's, but surrounding
that non-choice is a million
other choices
that I can make."
- Michael J. Fox, a celebrity and
a Parkinson's disease survivor

Dedication

To the most magnificent human beings, my father, Mahmoud Gaballah, and my mother, Awatef Elmaleh.

Without you, I would have never been here, and without you, I will never go anywhere.

Acknowledgements

All praise is due to our God (Allah) who has been bestowing me with his great bounties and enabled me to complete my dissertation.

I would like to express my sincere gratitude to my advisor Dr Vijay Parsa for his continuous support of my PhD study and related research, for his guidance, and for his patience. Through this journey, I learned a lot from Dr Parsa about how to be an outstanding researcher, a role model instructor, and a great person. I will carry the lessons I learned from Dr Parsa for the rest of my life.

I appreciate all the help I got from Dr Scott Adams and his student Daryn Cushnie-Sparrow to complete the research work in this thesis and collect the subjective data.

I would like to thank my advisory committee, Dr Hanif Ladak and Dr Ilia Polushin, for their insightful comments and encouragement. I am grateful to all the members of the examination committee for accepting to review my thesis. Their comments further strengthen the thesis and enhance its quality.

I would like to thank my brothers in arms, Anas Saci and Moftah Ali. Anas is a person who has a lot of knowledge, and he does not hesitate to share it with anyone who needs it. I benefited from his expertise throughout these years, and I am in a high debt to him. Moftah is a big brother to me who always gives me valuable advice, and I enjoy his friendship.

My gratitude extends to Steve Beaulac and David Grainger for investing their knowledge and their skills to help and support the work of the new device system. I would like to extend my gratitude to all members of the National Centre of Audiology and the Electrical and Computer Engineering Department at Western University for their cooperation and their support during my studies at Western University.

My sincere gratitude and my love are due to my parents for their continuous guidance, encouragement, support, and prayers during my life. I would also like to extend my gratitude to my siblings, Nada, Mohamed, and Omar for their help and cooperation. Thanks are due to my wife, Heba Askar, for all the support she gave me and all the sacrifices she made to let me achieve success in this journey. She is my partner in this success and my partner in life.

Finally, I want to express my endless love, my joy, and my gratitude for the presence of my little angel Dana in my life. She is the fresh air that blew over my soul and turned my life into a piece of heaven.

Table of Contents

List of Tables

List of Figures

Nomenclature

AAC	Augmentative and Alternative Communincation
AAVS	Articulatory-Acoustic Vowel Space
Adam	Adaptive Moment Estimation Optimizer
APM	Ambulatory Phonation Monitor
AVQI	Acoustic Voice Quality Index
BBB	Blood-Brain Barrier
CAPE-V	Consensus Auditory Perceptual Evaluation of Voice
CASA	Computation Auditory Sense Analysis
CPP	Cepstral Peak Prominence
CPPs	Smoothed Cepstral Peak Prominence
DCT	Discrete Cosine Transform
DL	Deep Learning
DNN	Deep Neural Network
ERB	Equivalent Rectangular Bandwidth
GFCC	Gaussian Frequency Cepstrum Coefficients
GMM	Gaussian-Mixture Model
GPR	Gaussian Process Regression
GRBAS	Grade, Roughness, Breathiness, Asthenia, and Strain.

HNR	Harmonic-to-Noise Ratio
ICC	Intraclass Correlation Coefficient
LB	Lewy Bodies
LCQA	Low Complexity Quality Assessment
LPC	Linear Prediction Coefficients
LR	Linear Regression
MARS	Multivariate Adaptive Regression Splines
MFCC	Mel Frequency Cepstrum Coefficients
ModA	Modulation Area
PCA	Princpal Component Analysis
PD	Parkinson disease
PSD	Power Spectral Density
PSE	Power Spectrum Envelope
SNF	Speech-to-Noise Feedback
SNR	Signal-to-Noise Ratio
SPL	Sound Pressure Level
SRMR	Speech-to-Reverberation Modulation Energy Ratio
SVR	Support Vector Regression
TL	Spectral Tilt
UPDRS	Unified Parkinson's Disease Rating Scale
VAD	Voice Activity Detection

Chapter 1

Introduction

1.1 Parkinson's disease (PD)

Parkinson disease (PD) and its symptoms were first described in a medical context in 1817 by James Parkinson, a general practitioner in London, UK [1]. In his paper named "An Essay on the Shaking Palsey", he described the disease as an involuntary tremulous motion, with lessened muscular power, in parts not in action and even when supported; with a propensity to bend the trunk forwards, and to pass from a walking to a running pace with the senses and intellects being uninjured [2]. James Parkinson called the disease that was causing these symptoms as paralysis agitans [2]. In 1877, Jean-Martin Charcot, who is also known as the father of neurology, had a significant impact on the study of PD. He identified the main features of PD and was the first one to give the disease its name after its first discoverer James Parkinson [1, 3].

Today, PD is the second most common neurodegenerative disease (the first being Alzheimer's disease). Much research has been conducted to find out the causes of PD and develop methods and medications for the treatment. However, PD's root causes are still unknown, and a final cure has not been reached yet. Today, the increase in the average age of the worldwide population led to the growth of the incidence, prevalence, and mortality rates of PD [4]. The annual incidence rate or the number of new cases of PD occurring is estimated to be 100 / $100,000$ for the age group $70 - 80$ years old [4]. The prevalence or the total number of PD cases reaches 565 / $100,000$ in the same age group [4]. Although PD is not a direct cause of death, it has been found that there is an increased risk of death to the people impaired with PD due to the severe motor dysfunction [4]. In the Canadian context, an estimated 0.2% of the Canadian population as a whole, and 4.9% of Canadians living in long-term care facilities are afflicted by PD [5]. Studies have shown that among a group of $100,000$ persons from all ages, 10 to 20 cases are diagnosed with PD every year, and this number increases in

the $65-85$ year old age group to around $150-300$ persons. There is, therefore, a significant research and clinical interest in effective diagnosis, treatment, and rehabilitation options for PD [6].

Pathological symptoms of PD are severe loss of dopaminergic neurons in the nigrostriatal region of the brain and the appearance of cytoplasmic inclusions known as Lewy bodies (LBs) [7]. A reduction in dopamine production leads to the appearance of rest tremors, akinesia, cogwheel rigidity, and postural instability. Besides, PD leads to voice and speech impairments, which are discussed next.

1.2 Speech and voice impairments related to PD

Statistics show that nearly 90% of people impaired with PD develop voice and speech disorders during their time with the disease [1]. PD speech is often characterized with reduced voice volume (hypophonia); harsh voice quality (dysphonia); imprecise consonant and vowel articulation due to the reduced range of articulatory movements (hypokinetic articulation); and a tendency of these movements to decay and/or accelerate towards the end of a sentence [1]. Hypophonia is considered the most frequent speech symptom in PD [8]. It is hypothesized that hypophonia may be attributed to a sensorimotor deficit in the self-perceived loudness of the individual's speech [9]. Hypophonia can make it very challenging for listeners to understand individuals with PD, particularly in conditions with background noise. Previous studies have demonstrated abnormally reduced speech Signal-to-Noise Ratios (SNRs) and reduced intelligibility when individuals with PD are conversing in moderate levels (65-70 dB sound pressure level (SPL)) of multi-talker noise [10, 11].

Furthermore, imprecise articulation and abnormal speech rate lead to a blurring of speech components, which also impact the understanding of PD speech [12]. While speech intelligibility is an indicator of how well the message carried by the speech stimulus was understood (e.g., how many words in a sentence were correctly interpreted), speech quality is a multi-dimensional perceptual phenomenon that encompasses attributes such as clarity, pleasantness, naturalness. PD speech, even when intelligible, may be perceived as abnormal, given its harsh and breathy qualities [12]. Evidence exists that the intelligibility and quality aspects of PD speech deteriorate with the advance of the disease [1, 12, 13] and as such evaluation of intelligibility and quality of PD speech is of paramount research and clinical interest.

1.3 Treatment of speech and voice impairments

1.3.1 Medications

Levodopa is a medical treatment for PD that improves the motor symptoms of the disease [14]. Levodopa is considered to be the most effective drug to alleviate the motor symptoms of PD and the only one documented to improve the life span in this disease [14]. Levodopa crosses the blood-brain barrier (BBB) and increases the production of dopamine [14]. This process reduces the effect of the dopamine production drop caused by PD and enhances the efficiency of the motor features of the PD subject [14]. There is a profound enhancement of the motor systems when using levodopa at the early stages of PD, and despite the fact that levodopa is considered to be an old treatment (the first report of improvement in the motor symptoms due to the use of levodopa was published in 1961), no other medication surpasses its benefit in the improvement of the motor features of PD [14].

Despite its therapeutic effects in the treatment of motor deficits of PD, levodopa does not have the same healing effect on PD speech. The magnitude, consistency, and long-term effects of levodopa are far from satisfactory [15, 16, 17]. This raises the need for other therapeutic and assistive methods to mitigate the deterioration of quality and intelligibility for Parkinsonian speech. One of these methods is the use of assistive devices to increase the Parkinsonian speech intensity and enhance its quality and intelligibility.

1.3.2 Assistive devices

Three methods receive the most attention in the treatment of voice and speech impairments in PD. These methods are: perceptually-based behavioral speech therapy, instrumentally-based biofeedback therapy, and prosthetic or assistive speech devices [18]. These treatment procedures aim to increase the speech intensity, improve speech prosody, reduce rapid speech, and increase articulatory mobility and precision [18]. Although the first two methods have proved to be effective in the treatment of PD speech impairments, they cannot transfer the treatment outside the clinical environment [19]. In other words, people impaired by PD show negligible improvements when they leave clinical treatment [19]. This raises the need for a solution that can be easily transferred outside the clinic, such that people with PD continue to benefit from the treatment in their daily life. The third approach comprising of assistive amplification devices provides such a solution to people impaired by PD.

1.3.2.1 Voice amplifiers

Amplification devices for PD subjects are categorized among the augmentative and alternative communication (AAC) devices, which are used to compensate for impairment and disability patterns [18]. Amplification devices are used primarily to increase voice intensity and loudness, which leads to an improvement in the perceived speech intelligibility [8]. Moreover, when individuals with PD use an amplification device, they may expend less vocal effort and experience more successful communication with fewer requests to repeat their messages [8]. Several amplification devices are commercially available for this purpose. The electroacoustic performance of these amplifiers is typically characterized using measures of some attributes such as frequency response, sensitivity, and distortion, which are found in the device specification sheets. While these parameters are useful in essential performance characterization and quality control, they do not capture the effects of amplification on perceived intelligibility and quality. Consequently, there is a need to benchmark the amplification devices in a perceptually relevant manner when they are used by people impaired by PD [8].

1.3.2.2 Speech-to-Noise Feedback (SNF) type devices

There have been some papers that investigated the use of assistive devices to measure particular speech characteristics by utilizing sensors that are attached to the human body. For example, Boudreaux *et al.* [20], presented a system called the Ambulatory Phonation Monitor (APM). APM was used to find if there was any difference between the values of the phonation time, amplitude, and the mean fundamental frequency between non impaired people and people impaired with PD. The system comprised of a sensor that did not use a microphone to record speech; instead it measured speech characteristics by detecting the neck skin vibration. This approach enhanced the measurement process by making it objective and independent of the ambient noise. This paper showed that people with PD have a lower mean fundamental frequency, amplitude, and phonation time than people who are not impaired with the disease. However, not using a microphone made APM incapable of measuring characteristics like SNR because of the non-measurement of the noise signal.

1.4 Assessing speech and voice impairments

Traditionally, subjective methods are used to assess the quality and intelligibility of speech [21]. Subjective measures need human judges to assess the quality and the intelligibility of the speech stimuli. A group of participants listen to the speech recordings and give their

subjective assessment in terms of intelligibility and quality, then the mean score of their ratings is the value that represents the perceived quality or the intelligibility [21].

While subjective assessment of sound quality has high face validity and can be considered as the gold standard, it is not efficient in terms of time and resources. This weighs in favour of objective, instrumental assessment of speech quality, where computational algorithms are used to quantify the speech quality without requiring the involvement of human subjects [21]. While such objective metrics are routinely used for evaluating telecommunication and assistive hearing devices [21], few studies have applied them to Parkinsonian speech quality assessment.

1.5 Problem statement and thesis scope

As mentioned before, PD leads to the deterioration of speech quality and intelligibility. Speech amplifiers are used to compensate these disability patterns, but their performance in terms of quality and intelligibility needs to be benchmarked when they are used by people impaired with PD. Although subjective measurements of speech quality and intelligibility are considered the gold standard, they are not efficient in terms of time and cost. This weighs in favor of the idea of automating the process of measurements, to establish an objective assessment of speech quality and intelligibility [21].

Previous research work utilized traditional acoustic measures such as shimmer, jitter, harmonics-to-noise ratio (HNR), and cepstral peak prominence (CPP) to classify the Parkinsonian speech, but to the best of our knowledge, there has not been a research conducted to apply other acoustic features in assessing the quality and intelligibility of PD speech like mel frequency cepstrum coefficients (MFCC), low complexity quality assessment (LCQA), and Gammatone frequency cepstral coefficients (GFCCs). The research aims to utilize these objective metrics to enhance the performance of the objective PD speech quality predictors and obtain a higher correlation with the subjective scores. This research aims to develop objective metrics to assess the quality of both sustained vowels and running speech. The research aims to benchmark the performance of speech amplifiers used with PD in terms of speech quality and intelligibility. Machine learning algorithms including support vector regression (SVR), deep learning (DL), and Gaussian process regression (GPR) are used to map the extracted acoustic features to the subjective scores, thereby leading to the development of new objective metrics.

Finally, a new ambulatory assistive device to help people impaired with PD is presented in this research. The new device helps people impaired with PD to control their SNR, especially in noisy environments which will enhance the intelligibility and quality of their

speech. This device is easy to use and more cost-effective than other similar devices. The impact of this device in enhancing quality and intelligibility is explored.

1.6 Thesis organization and contributions

This thesis has been divided into six chapters, as follows:

In Chapter 2, a review of the previous research work in quality assessment and the treatment of Parkinsonian speech is presented. The chapter contains an explanation of the structure of non-intrusive quality measures. The acoustic features explored in this thesis are reviewed. This chapter includes a more in-depth look on machine learning algorithms, including SVR, DL, and GPR.

Chapter 3 outlines a study on the objective evaluation of the quality of sustained vowels of PD patients. The database contains records collected from 51 patients with and without taking the levodopa medication, in addition to 10 control subjects. Data were collected from the patients when they were on medication and when they were off medication. The whole database contained 113 voice samples. The study used HNR, smoothed CPP, GFCC, and LCQA as acoustic features for the objective models. Linear regression (LR) and SVR were used to map these acoustic features to the perceived subjective scores. The new presented combined metric has a high correlation of 0.81 with the subjective scores.

In Chapter 4, a study to assess the quality of amplified PD running speech using subjective and objective metrics is presented. Assessing the quality of the amplified Parkinsonian speech is used to benchmark the performance of the speech amplifiers used by people impaired with PD. The study compares between the performance of several objective metrics by utilizing different acoustic features that include MFCC, GFCC, LCQA, speech-to-reverberation modulation energy ratio (SRMR), modulation area (ModA), and CPP. Machine learning algorithms are applied on these acoustic features to develop the objective metrics to estimate the quality of Parkinsonian speech. The combined objective metric developed in this chapter resulted in a correlation of 0.85 with the subjective measurements of the quality of running speech.

Chapter 5 presents a new ambulatory and cost-effective device that is used by people impaired with PD in their daily life to help them enhance the intelligibility of their speech, especially in noisy environments. It is shown that the use of the new SNF device led to the enhancement of quality, SNR, and intelligibility measurements of all the speech tasks at high background noise environments.

Finally, Chapter 6 concludes the research in this thesis and presents suggestions for future work.

Chapter 2

Literature Review

2.1 Introduction

Typical measures of phonatory aspects of PD speech (computed with sustained vowel or voiced components extracted from running speech recordings) include: (a) jitter, which quantifies the short-time perturbations in the fundamental frequency or pitch; (b) shimmer, which measures the short-time variations in the intensity; and (c) harmonics-to-noise ratio (HNR), which calculates the relative level of the harmonic components to aperiodic background noise [13]. Typical measurements of speech rate and fluency include the number of syllables per second during a reading task or the number of inappropriate silence intervals [13]. Articulatory aspects of PD speech are typically measured using the vowel space area or its variants, which measure the area spanned by the first and second formants (F_1, F_2) for the four corner vowels [13]. Efforts have been made to extend this approach to formant tracks extracted from running speech. For example, Whitfield *et al.* [22] extracted the formant tracks from the first sentence of rainbow passage recorded from 12 subjects with PD and 10 neurologically healthy controls. The so-called articulatory-acoustic vowel space (AAVS) extracted from these formant tracks in the F_1, F_2 space was found to be reduced for the PD subjects and correlated well with perceptual ratings of speech clarity. However, this technique (and others that depend on extracting voiced components from running speech for further analysis) required accurate identification of voiced portions and manual pruning of formant tracks to remove spurious components.

More recently, measures based on cepstrum analysis have been applied for characterizing the PD speech. Since this is an essential part of the proposed research, a more detailed description of cepstrum analysis is given next.

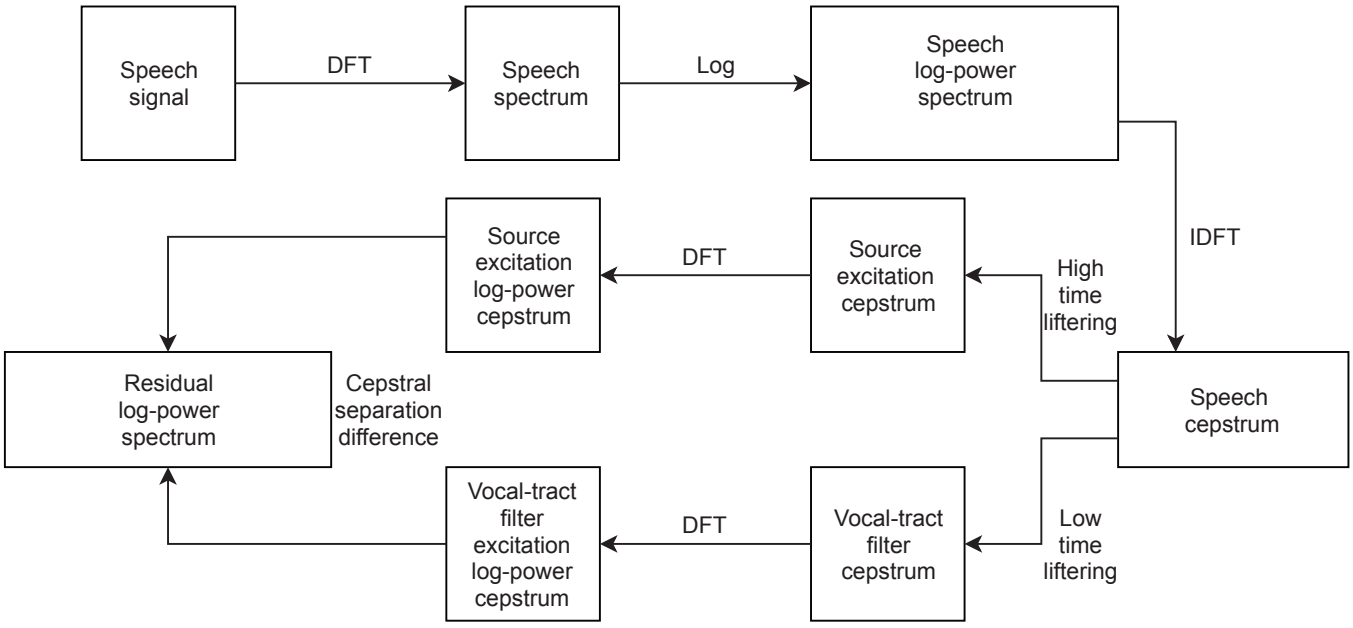


Figure 2.1: Cepstrum analysis and the calculation of Cepstral Separation Distance (CSD) [23]

2.2 Cepstrum Analysis and Related Features

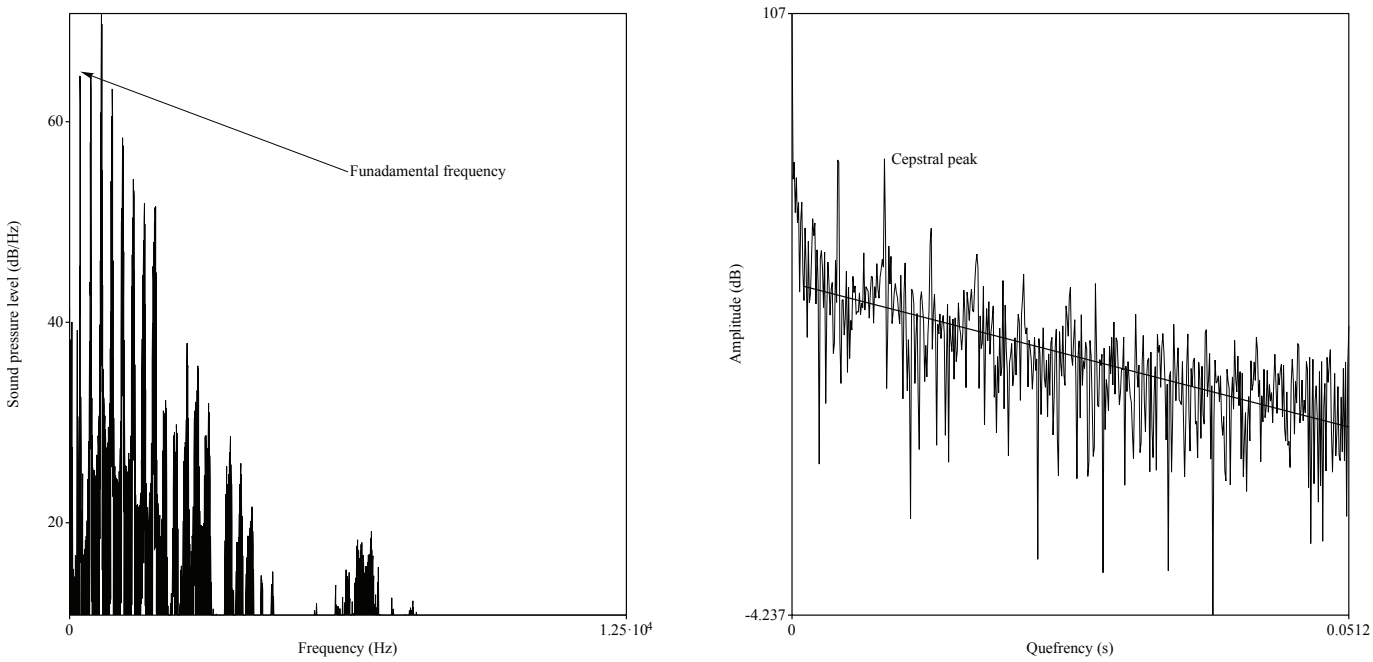
The first four blocks of Fig. 2.1 represent the steps in computing the cepstrum, which is defined as [23, 24]:

$$c(n) = \left| F^{-1}\left\{ \log\left(\left| F\left\{ s(n) \right\} \right| \right) \right\} \right| \tag{2.1}$$

where $s(n)$ is the discrete time speech signal, $F\{\cdot\}$, $F^{-1}\{\cdot\}$ represent the Fourier transform and its inverse respectively. As seen in Fig. 2.1, cepstrum analysis allows decoupling of excitation and vocal-tract filtering components in the source-filter speech production model.

Several indices derived from cepstrum analysis have been applied for characterizing PD speech. For example, Khan *et al.* [23] developed the Cepstral Separation Difference (CSD) index to predict the PD speech severity. As shown in Fig. 2.1, the CSD captures the difference between the excitation log power spectrum and the filter log power spectrum, with larger values of CSD observed for speech signals recorded from subjects with more advanced PD. Using a database of 240 clinically rated speech samples collected from 60 PD subjects and 20 healthy controls, Khan *et al.* [23] reported a correlation of 0.78 between CSD index and Unified Parkinson's Disease Rating Scale (UPDRS) speech symptom severity scores.

Another commonly used metric based cepstrum analysis is the cepstral peak prominence (CPP), and an example of how to calculate it is shown in Fig.2.2. CPP represents the differ-



(a) Spectrum (b) Cepstrum

Figure 2.2: Spectral and cepstral representations of sustained vowel /a/ [25]

ence between the peaks of the cepstrum and the linear regression function of the cepstrum [25]. CPP was shown to correlate well with subjective ratings of dysphonic voice quality [25].

An alternative approach to cepstrum analysis and subsequent parameterization is the utilization of a filterbank. The Mel-Frequency Cepstral Coefficients (MFCCs), the central features in many commercial speech recognition systems [26], are computed using this approach as described in the next subsection.

2.3 Mel Frequency Cepstral Coefficients

In MFCC, we use a scale to mimic the performance of the human ear as humans are sensitive to small changes at low frequencies. The formula of changing the frequency axis to Mel scale axis is as follows [26]:

$$M(f) = 1125 \ln\left(1 + \frac{f}{700}\right) \tag{2.2}$$

$$M^{-1}(m) = 700\left(\exp\left(\frac{m}{1125}\right) - 1\right) \tag{2.3}$$

Applying Eq. (2.2) to the frequency values change them into mel bins, while Eq. (2.3) is used to transform the the mel bins into frequencies. The first step in forming the mel filter bank is to divide the whole mel scale between the 2 mels that correspond to 0 Hz and $f_s/2 = 8000$ Hz into 40 equal linear spacings. The corresponding mel range is between 0 and 2834.99 mels. Afterwards, these mel bins are converted back into their corresponding frequencies. These frequencies serve as the band limits of the mel filters that are applied on each frame of the wave record. Each filter has three points, the first and the third points have 0 output, while the middle point is the point that has the maximum output value of the filter. The second frequency point is the first point of the next filter. The filter transfer function can be expressed as follows [26]:

$$H_m(k) = \begin{cases} 0 & f\left(k\right) < f(m-1) \\ \frac{f(k)-f(m-1)}{f(m)-f(m-1)} & f(m-1) \leq f\left(k\right) \leq f(m) \\ \frac{f(m+1)-f(k)}{f(m+1)-f(m)} & f(m) < f\left(k\right) \leq f(m+1) \end{cases} \tag{2.4}$$

where $f\left(k\right)$ is the measured frequency, $f\left(m\right)$ and $f\left(m-1\right)$ are the filter boundaries. After applying MFCC filter bank, we sum the energy at each filter band and take the logarithm of them. Afterwards, we apply discrete cosine transform to reduce the correlation between the filter bank energies as they are overlapped. Finally, inverse Fourier transform is done to calculate MFCC coefficients [26].

Fig. 2.3 shows the graph of the resulting MFCC filters. After processing, we take into consideration only the first 13 filters as they contain most of the energy [26].

2.4 Gammatone Frequency Cepstrum Coefficients (GFCC)

The Gammatone filterbank has a greater ability to mimicking the auditory filterbank than the mel filterbank. As such, cepstral coefficients extracted using gammatone filterbank have been shown to produce better speech recognition performance than MFCCs [27]. Calculating GFCC coefficients has the same steps as calculating MFCC coefficients, but we use different scale and different filter transfer functions. In GFCC we use the equivalent rectangular bandwidth (ERB) scale. There are two equations for ERB scale as follows [28]:

$$\mathrm{ERB}\left(f\right) = \begin{cases} 24.7 + \frac{f}{9.265} & f > 2\,\mathrm{KHz} \\ 9.265 \ln\left(1 + \frac{f}{228.8455}\right) & 27\,\mathrm{Hz} < f < 2.2\,\mathrm{KHz} \end{cases} \tag{2.5}$$

The impulse response of the Gammatone filter is given by,

$$g(t) = \frac{at^{n-1}\cos\left(2\pi f_c t + \varphi\right)}{e^{2\pi bt}} \tag{2.6}$$

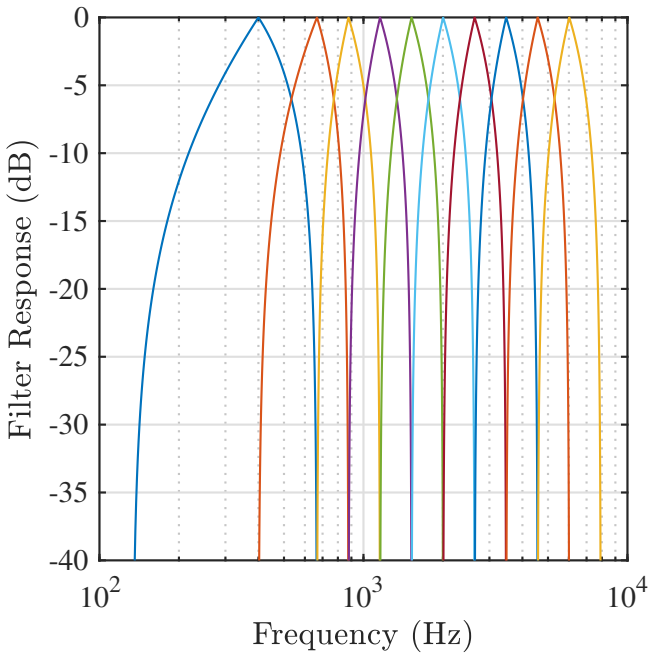


Figure 2.3: MFCC filter bank [26]

where f_c is the filter center frequency, φ is the phase of the carrier, a is the amplitude, n is the filter order, b is the filter bandwidth in Hz, and t is time. We usually set the value of $n = 4$, $b = 1.019$ ERB, $\varphi = 0$. As we can see, Eq. (2.6) is in the time domain. To get the frequency response of this filter, we follow the procedure in Slaney *et al.* [29]. First, we get the Laplace transform of the function $g(t)$ as,

$$G(s) = \mathcal{L}\{g(t)\} \tag{2.7}$$

where the $\mathcal{L}\{\cdot\}$ is the Laplace transform process, that is given by definition as,

$$\mathcal{L}\{t^n f(t)\} = (-1)^n \frac{d}{ds^n} F(s) \tag{2.8}$$

Then, by computing (2.7) using the definition in (2.8), we get

$$G(s) = \frac{6\left(-B^4 - 4B^3 s - 6B^2 s^2 - 4Bs^3 - s^4 + 6B^2\omega^2 + 12Bs\omega^2 + 6s^2\omega^2 - \omega^4\right)}{\left(B^2 + 2Bs + s^2 + \omega^2\right)^4} \tag{2.9}$$

This is an eighth order filter which can be transformed into four second order cascaded filters. The digital equivalent of this analog filter is obtained through bilinear transform. The frequency response of the Gammatone filter bank is shown in Fig. 2.4. The spectrum is divided into 128 bands and take into consideration the first 40 filters.

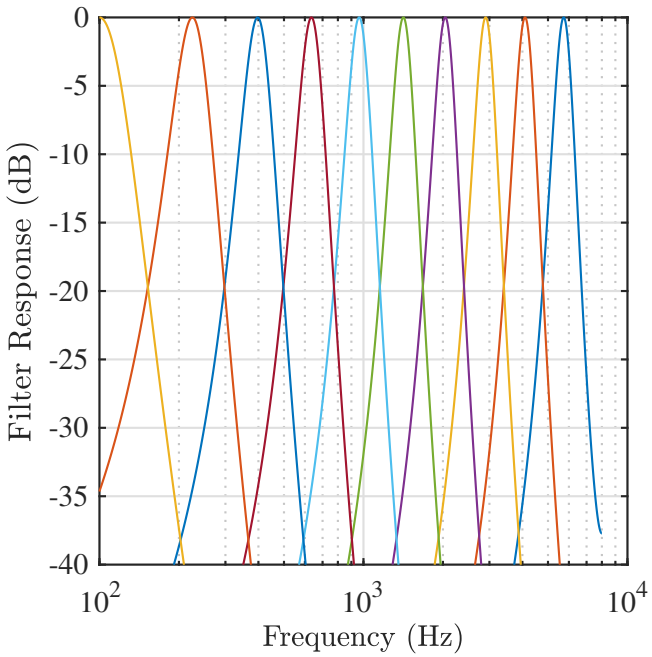


Figure 2.4: GFCC filter bank [29].

It is pertinent to note here that GFCCs have not been applied for estimating the intelligibility or quality of PD speech. In this research, GFCC is utilized to predict the quality of sustained vowels and running speech of people impaired with PD. The results show that the GFCCs outperform MFCCs in predicting the quality of amplified PD speech.

2.5 Features based on Speech Envelope

2.5.1 Speech-to-reverberation modulation energy ratio (SRMR)

Fig. 2.5 shows a block diagram of the SRMR method. In this method, the speech signal is applied to Gammatone filter bank [30] – a series of overlapped filters that have center frequencies that range from 125 Hz to half the sampling rate. Afterwards, Hilbert transform is applied to each of the filter-bank output to evaluate temporal envelope of each filter output; these envelopes have frequencies that range from 0 to 128 Hz. At this point, each envelope is filtered into eight overlapping modulation bands with center frequencies from $4 - 128$ Hz. Finally, SRMR is computed as a ratio between the energy stored in the first four filters, which they contain most of the speech energy and the last four filters which contain the background noise.

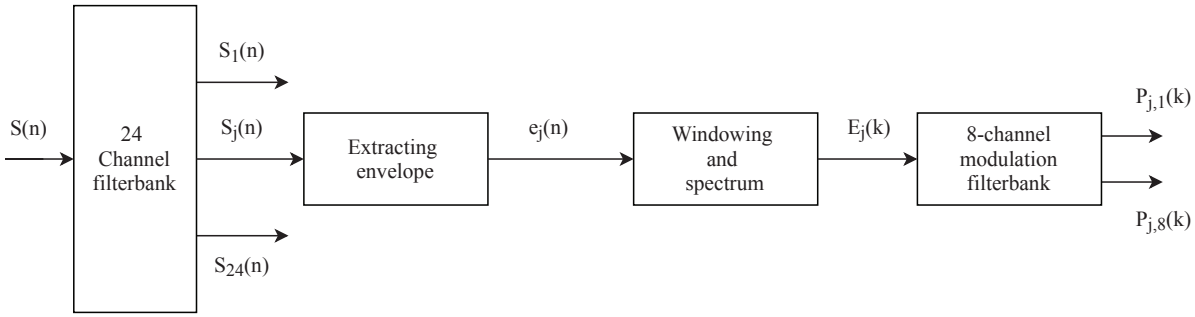


Figure 2.5: SRMR block diagram [30].

2.5.2 Modulation Spectrum Area (ModA)

There are some similarities between ModA and SRMR [31]. While SRMR depends on calculating the ratio between the energy in the lowest temporal bands and the highest temporal bands, ModA depends on the idea that reverberation smears the speech signal envelope that will lead to a decrease of the modulation area. Fig. 2.6 shows the change of modulation areas due to reverberation. Unlike SRMR, the speech signal is decomposed to only 4 filters, and then Hilbert transform is applied to each acoustic band to get the temporal envelope for each acoustic band. Each envelope is down sampled to 20 Hz then applied to 1/3 octave filters with center frequency ranges of .5–8 Hz. The indices from those filters are summed up to get the area under each acoustic band, and then those areas are averaged by dividing them by 4 to get the metric index.

While both SRMR and ModA have been evaluated with noisy and reverberant speech [30, 31], they have not been specifically investigated with PD speech. As discussed earlier, PD speech is characterized by short rushes and abnormal rate, which impact the profile of the speech envelope. Recent work by Fletcher [32] has shown that certain envelope features, such as envelope energy in the region of $3-6$ Hz for the 250 Hz frequency channel and envelope energy in the $0-10$ Hz band for the 500 Hz frequency channel, contributed the most in predicting the intelligibility of PD speech.

2.6 Low Complexity Quality Assessment (LCQA)

Fig. 2.7 shows the structure of LCQA algorithm used in [33]. The key idea of LCQA is to extract linear prediction coefficient (LPC) features per frame and derive their statistical characterization. The speech signal is framed, and a vector of 11 features is extracted from each frame, then the speech quality is assessed from the statistical properties of this vector such as mean, variance, skewness, and kurtosis. These properties per frame are then applied

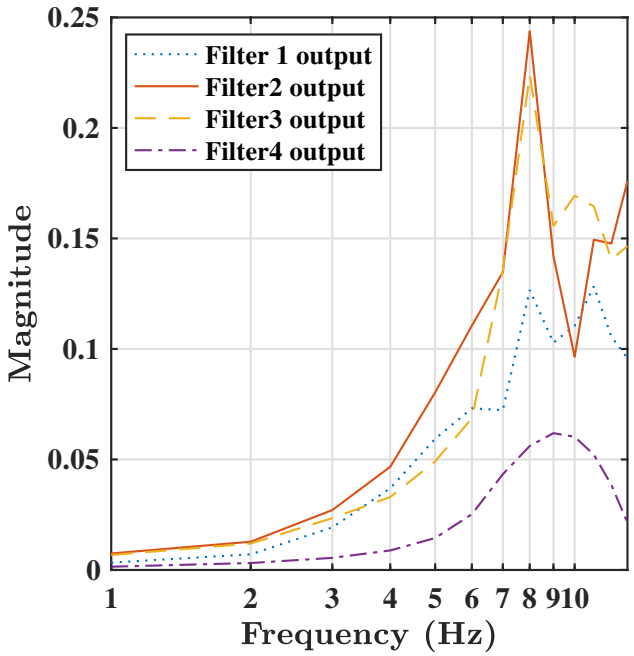


Figure 2.6: Areas under modulation curves in ModA [31].

to Gaussian-mixture probability model (GMM) to assess the quality of the speech signal [34]. The eleven features extracted from each frame include six features and the first order derivatives of five of them. These features are spectral flatness, spectral dynamics, spectral centroid, excitation variance, speech variance, and pitch period [33]. The spectral flatness measures the shape of the power spectral density (PSD) and is related to the strength of the resonant structure in the power spectrum [33]. Spectral dynamics feature quantifies the dynamics of the power spectrum envelope (PSE), which play an essential role in the perceived distortion [33]. The spectral centroid indicates the perceptual brightness, while speech and excitation variances measure the energies of input speech and the error residual after linear prediction coding [33]. Pitch period is another indicator of quality assessment [33]. In addition to these six features, all their first order derivatives are selected as part of the feature vector except for the spectral dynamics. In [33], some dimensionality reduction was used before feature mapping to reduce the number of features used by the following two methods,

- Principal Component Analysis (PCA) followed by linear regression.
- Multivariate Adaptive Regression Splines (MARS).

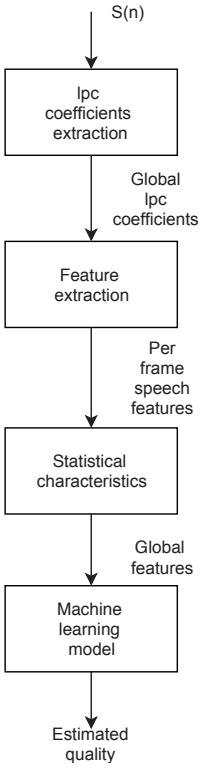


Figure 2.7: LCQA block diagram [34].

2.7 Feature Mapping

Mapping algorithms aim to generate a function that assimilates the collected data (such as GFCC) to match the subjective scores. To express this mathematically, if speech data extracted from the patients' recordings are x^i, where i is the patient, while y is the clinical rating, then an equation that links x, y will be as follows[35]:

$$\hat{y} = f(\theta, X) + b, \tag{2.10}$$

where X is the feature matrix that has size $m \times n$, m is the number of training samples, θ is the matrix containing the weights of the machine learning algorithm, and n is the number of extracted features. The function $f(\theta, x)$ obtains the required objective scores, and the coefficients of this model are updated such that they minimize the mean square error function:

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (\hat{y} - y)^2, \tag{2.11}$$

where y is the subjective scores' vector that the regression model needs to approach, y has size $m \times 1$. Applying the gradient descent algorithm updates the model coefficients such

that:

$$\theta^{'} = \theta - \alpha \frac{dJ(\theta)}{d\theta} \tag{2.12}$$

where α is the gradient descent learning rate that controls the speed performance of the regression model, and $\theta^{'}$ is the updated coefficients matrix. There are other models that incorporate some modifications to enhance the regression performance. In this research, three methods are used, support vector regression (SVR), Gaussian process regression (GPR), and deep learning. The following subsections present brief explanations of each technique.

2.7.1 Support vector regression (SVR)

SVR was developed in 1995 [36], and this technique relies on intuition that achieving flatness in Eq.2.11 is desirable [37]. Minimizing the norm $\|\theta\|^2$ achieves such purpose, which means:

$$\text{minimizing } J(\theta) = \|\theta\|^2 + C \sum_{i=1}^{m} (\xi_i + \xi_i^{*)} \tag{2.13}$$

$$\begin{gathered} y_i - \theta^T x - b \leq \epsilon + \xi \\ \theta^T x + b - y_i \leq \epsilon + \xi^* \\ \xi, \xi^* > 0 \end{gathered} \tag{2.14}$$

$$|\xi| = \begin{cases} 0 & \text{if } |\xi| \leq \epsilon \\ |\xi| - \epsilon & \text{otherwise} \end{cases} \tag{2.15}$$

Constraints in Eqs. (2.14),(2.15) present a duality problem [37], and to solve it, new parameters called Lagrange coefficients ($\alpha, \alpha^* > 0$) are added for each sample to the objective function:

$$J(\alpha) = \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) x_i^T x_j + \epsilon \sum_{i=1}^{m} (\alpha_i + \alpha_i^*) + \sum_{i=1}^{m} y_i (\alpha_i - \alpha_i^*) \tag{2.16}$$

Smola *et al.* [37], explain the proof of Eq.(2.16) in detail. The hypothesis function is modified to be

$$f(\theta, x) = (\alpha_i - \alpha_i^*) x_i^T x + b \tag{2.17}$$

$(\alpha_i - \alpha_i^*)$ is called a support vector, and using $x_i^T x_j$ is called a linear kernel. Another

kernel to be used is called Gaussian kernel, where we use

$$G(x) = exp\left(-\frac{(x_i - x_j)^2}{2\sigma^2}\right) \tag{2.18}$$

where σ^2 is the variance of the SVR model.

2.7.2 Gaussian Process Regression (GPR)

A Gaussian process is defined as a set of random variables that has a joint Gaussian distribution [38]. The Gaussian process is characterized by its mean function $m(x)$ and covariance function $k\left(x, x^{'}\right)$. The covariance function measures the correlation between the input features of the database and they can be called kernels [39]. The kernel function in Eq. (2.18) is an example that can be used in GPR models. When the regression model trains data that are noiseless, the training dataset is $D = \{(x_i, y_i), i = 1 \dots m\}$. Given the test data set of features X_* of size $m_* \times n$ where m_* is the number of test samples, and n is the number of features per sample, it is required to evaluate the corresponding values of y_*. The joint Gaussian probability distribution function (PDF) is as follows [38, 39]:

$$\begin{pmatrix} y \\ y_* \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu \\ \mu_* \end{pmatrix}, \begin{pmatrix} K & K_* \\ K_*^T & K_{**} \end{pmatrix}\right) \tag{2.19}$$

where μ, K are the mean and the covariance of the dataset, μ_*, K_{**} are the mean and covariance of the test dataset, and K_* is the joint covariance of the training and the test dataset. The conditional PDF of the Gaussian process f_* can be expressed as [38]:

$$f_*|f \sim \left(\mu_* + K_*^T K^{-1}(f - \mu), K_{**} - K_*^T K^{-1} K_*\right) \tag{2.20}$$

In the case of noisy regression, i.e. $y = f(x) + \epsilon$, where ϵ is an additive noise with 0 mean and σ_n^2 variance, the training dataset covariance K changes to be $K_y = K + \sigma_n^2 I_m$, where m_* is the number of training dataset samples.

2.7.3 Deep learning

In deep learning [40], the learning process is divided into multiple layers where features are extracted from each layer. Deep learning then uses the backpropagation method to train these multilayer architectures and adapt them to extract new features to minimize the error function. One of the key characteristics of deep learning is that there is no need for a human intervention to design these layers of neurons since they are learned from the input data alone. Deep neural networks (DNNs) have proved to be a state-of-the-art tool in speech

recognition, and hence, they are adopted in this research.

DNN extracts hidden layers based on input samples applied to the mapper. The output of each neuron in every hidden layer is applied to an activation function to reduce the summation effect of the neurons' values. In this research, the rectified linear unit (*Relu*) function is applied to all the neurons' outputs in the hidden layers, while the *tanh* function is applied to the last output layer. Assuming that the input features are x_1, x_2, $\cdots$, x_n , X is $n \times m$ feature matrix, where n is the number of features and m is the number of training samples, W^1, $W^{[2]}$, $\cdots$, $W^{[l]}$ are the weights of the hidden layers, and b^1, $b^{[2]}$, $\cdots$, $b^{[l]}$ are the bias correcting vectors. The corresponding values are calculated as [41]:

$$Z^{[i]} = W^{[i]} A^{[i-1]} + b^{[i]} \tag{2.21}$$

$$A^{[i]} = g\left(Z^{[i]}\right), \tag{2.22}$$

where i is the layer number that ranges from 1 to l, and $g(\cdot)$ is the activation function. The calculation of the output of each neuron moving from the input layer to the output layer is called forward propagation, where the output of the DNN is applied to the cost function. The value of $A^{[0]}$ equals the input features matrix X, while the value of $A^{[l]}$ equals to the objective scores vector $\hat{y}$. In this research, the least mean square error (LMSE) is calculated as [41]:

$$J = \frac{1}{2m} \sum (y - \hat{y})^2, \tag{2.23}$$

where J is the mean square error through m number of samples, y is the subjective score or label vector, and $\hat{y}$ is the neural network output or the objective score vector. After calculating the error function, backward propagation is applied through the neural network from the output layer to the input layer to modify the network weights, which is called optimization, and this process is iterated to reduce the cost function. In this research, adaptive moment estimation (Adam) optimizer is used for training the model; more details about Adam optimizer can be found in [41, 42].

2.7.4 Feature selection and reduction

A higher dimensionality of the feature vector may cause overfitting. In such situations, extracted numbers of features for each metric must be reduced before applying the machine learning algorithm to avoid overfitting. One option for addressing this issue is through the principal component analysis (PCA). PCA is used to reduce the dimensionality of the input features of the machine learning algorithm and enhance the interpretation of the features

[43]. This dimensionality reduction or feature reduction has to be done in a way that keeps the information contained in the input features [43]. PCA utilizes the eigenvalues and the eigenvectors to come up with new features that have smaller dimensionality but maximizes the variance of the dataset [43].

In another approach to accomplish the goal of dimensionality reduction, the correlation between every single feature and the subjective scores is obtained, and then the features are rearranged according to their correlation values from the highest to the lowest. Subsequently, a Monte Carlo algorithm is applied to extract the maximum number of features that minimizes the cost function for each of the training and the test datasets. This algorithm takes the rearranged features' matrix and the subjective scores vector as two inputs [44]. At this point, the data is split into a training dataset and a test dataset. The algorithm applies linear regression to a subset of the datasets to find which subset achieves the minimum square error (MSE) with the subjective scores.

2.8 Evidence on the effectiveness of the acoustic measures

In the PD research context, a substantial number of acoustic analysis studies have focused on the classification of Parkinsonian speech from normal speech. For example, Al Mamun et al., [45] extracted features such as shimmer, jitter, and HNR and trained a deep neural network (DNN) to classify Parkinsonian speech based on these features with an accuracy of 97%. Similarly, Benba *et al.* [46] computed mel-frequency cepstral coefficients (MFCCs) from the speech waveforms and employed the support vector machines (SVMs) for discriminating between Parkinsonian and normal speech, with an accuracy of 90%. The few studies investigating the acoustic correlates of Parkinsonian speech quality have reported low correspondence with subjective scores. For example, Jannetts *et al.* [47] collected recordings of sustained vowel /a/ and continuous speech from 43 speakers with PD and 10 participants with ataxia. These recordings were perceptually rated using two subjective scales: GRBAS (G: the grade or overall dysphonia severity, R:roughness, B:breathiness, A:asthenia or weakness, and S:strain), and the other one is consensus auditory perceptual evaluation of voice (CAPE-V). Acoustic measures extracted from recordings included CPP, HNR, jitter, and shimmer – related measures. The CPP correlated best with the subjective ratings; it was noted that CPP and smoothed CPP (CPPs) correlation with subjective ratings increased 20% more than jitter correlation, double more than shimmer correlation, and 40 % more than HNR. However, the absolute correlation with ratings of more ecologically valid contin-

uous speech was significantly lower than the correlation with sustained vowel ratings (0.54 vs 0.86 for the grade or overall quality rating respectively).

Cushnie-Sparrow *et al.,* [48] conducted their research to investigate the effect of the levodopa medication on the acoustic measurement of perceived quality. Data were collected from 51 subjects impaired with Parkinson disease in addition to 11 healthy control individuals. This yielded a sustained vowels recording database of 113 samples. Measured acoustic metrics included jitter, shimmer, HNR, CPP, and acoustic voice quality index (AVQI). The authors in this paper measured the correlation of each acoustic feature with the perceived quality scores. They found that the highest obtained correlation value was between the subjective scores and the HNR with a value that reached 55%. Measurement of the perceived quality was obtained from a panel of 3 listeners, and the mean opinion score of the whole group of listeners served as a metric for the sound quality for each recording.

2.9 Assistive devices

Andreetta et al., [8] evaluated the performance of seven amplification devices with PD subjects. Isolated sentences and unscripted conversation from PD subjects and age-matched normal controls were recorded in the presence and absence of background noise, and with and without the use of amplification devices. These recordings were later played back to normal hearing listeners, and their perceived intelligibility ratings for the recorded stimuli on a scale of 0 - 100 were collected. Results showed that while all amplifiers enhanced the perceived intelligibility of PD speech, there was a differential effect in that the intelligibility rating for the best performing amplification device was approximately 30 basis points higher than the score associated with the lowest ranked amplifier. Although Andreetta et al.,'s study [8] does not report the perceived quality of amplified PD speech, it does highlight the need for assessing the amplifier performance in a manner that relates to speech perception.

Mehta *et al*. [49], investigated the correlation between characteristics derived from recordings obtained by the microphone and those obtained from accelerometer signal analysis. Subjects repeated sustained vowels: /*a*/, /*i*/, /*u*/, and they consisted of normal people and people impaired with voice disorders. A BU-27135 Knowles accelerometer was needed to detect the signal from the larynx of the subject, a head-mounted microphone recorded the speech signal, and a smart phone system was used as a processing device to analyze the data. They investigated the correlation of the characteristics: jitter, shimmer, HNR, spectral tilt (TL), and CPP. It was found that there was a very high correlation in jitter and CPP between signals of the accelerometer and the mic that reached to 90%. Shimmer, HNR, and TL had a correlation factor that ranged from 50% to 80%.

Borsky *et al.* [50], introduced a technique to classify the voice modes using BU-27135 Knowles accelerometer. The paper classified the voice recordings by analyzing the signals obtained from the accelerometer by extracting MFCC features and find out if the voice quality was modal, breathy, pressed, or rough. The classifier had an accuracy that ranged from 80% to 90%.

2.10 Summary

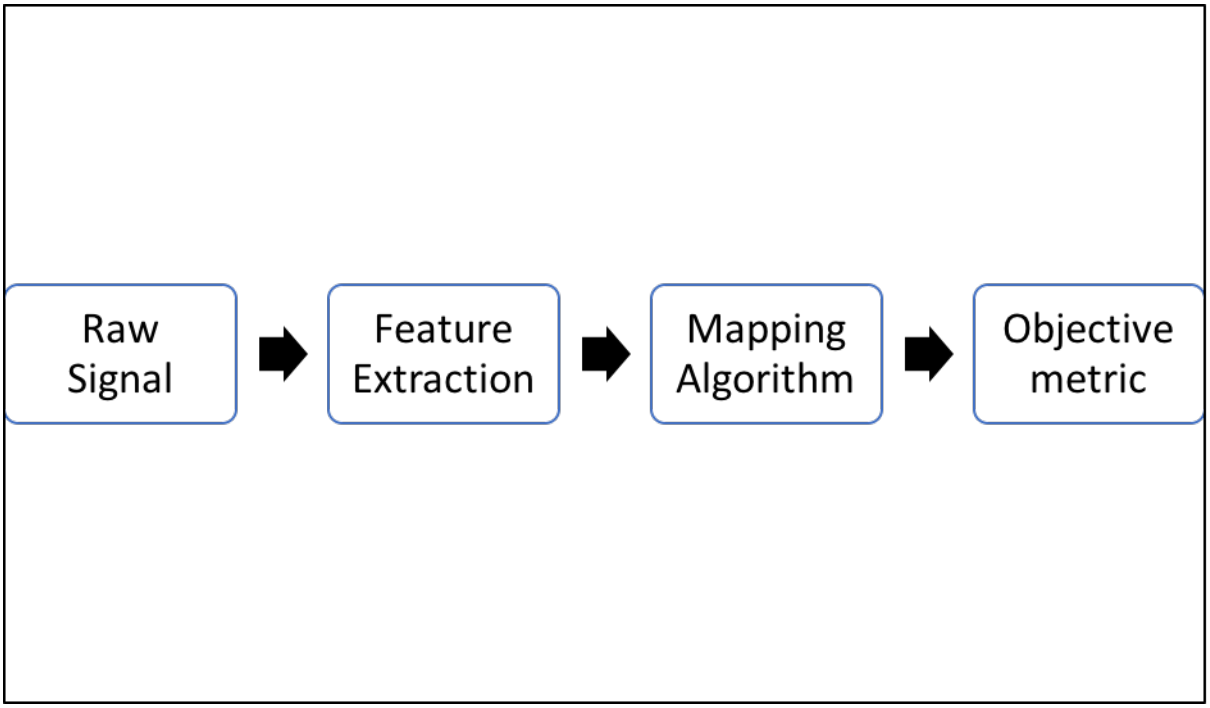


Figure 2.8: Unified framework for objective metrics

Given that one of the manifestations of PD is abnormal speech/voice characteristics, there has been research impetus towards the objective evaluation of PD speech. However, a majority of the published studies based their objective analysis on sustained vowel samples, rather than more ecologically valid running speech samples. Besides, several studies incorporated machine learning techniques for PD speech classification. Beyond classifying the PD speech, it is desirable to estimate the quality attributes, which perhaps have more clinical relevance. While some articles have investigated the relationship between objective speech measures and the severity of PD speech (mainly measured through the UPDRS scale), there is scope for improvement.

The goal of this thesis is to build a system for accurately estimating the quality of PD speech. To accomplish this, a simple unified framework, as shown in Fig.2.8 will be followed.

From the raw PD speech recording, a set of features will be extracted. Loudness, intelligibility, and quality are inter-related perceptual phenomena, and as such, commonality in the feature set is expected. The relative weighting of these features, though, will be different for estimating each of the three perceptual indices. This relative weighting is determined in the feature mapping block, wherein machine learning techniques are utilized to assimilate the feature set and map to the respective perceptual index.

Another goal for this thesis is to present a new assistive device to help the people impaired with PD to control their SNR. The new device is designed to be easy to build, versatile, and cost-effective.

Chapter 3

Evaluation of The Quality of Sustained Vowels of Parkinson's Disease

3.1 Introduction

Statistics show that nearly 90% of people impaired with PD develop voice and speech disorders during the course of their disease [1, 15]. The classic characteristics of Parkinsonian speech and voice include reduced vocal loudness (hypophonia), with a tendency of the voice to fade out; reduced prosodic pitch inflection (hypoprosodia); breathy or hoarse voice; imprecise articulation of consonants and vowels; and mumbled speech [1, 15]. Glottal incompetence and reduced respiratory support are considered reasons for hypophonia; these symptoms are caused by the rigidity of the chest wall [51]. Usually, the speech impairments of Parkinsonian speech are named hypokinetic dysarthria [15]. While speech articulation and fluency problems appear at later stages of PD, voice abnormalities may appear at earlier stages of the course of the disease [15], and thus require early attention. Voice quality is a multi-dimensional perceptual phenomenon that encompasses attributes such as clarity, pleasantness, naturalness, smoothness, and richness. It is of significant clinical and research interest to know whether PD medication such as Levodopa enhances PD voice quality.

In this chapter, non-intrusive objective metrics are developed to estimate the perceived quality of sustained vowels produced by PD subjects. Machine learning algorithms are deployed to combine multiple features to enhance the performance of the presented objective metrics. The correlation between the estimated quality scores and the subjective quality measurements served as an indicator of the performance of these objective metrics.

3.2 Methods

3.2.1 Voice recordings and subjective evaluation

In this research, subjective data collected by Cushnie-Sparrow *et al.* [48] were used to develop and benchmark the performance of non-intrusive objective metrics. A brief description of subjective data collection procedure is given here for the sake of completion. Sustained vowel samples were collected from 51 PD subjects with ages ranging from 45 to 85 years of age. Patients were evaluated off and on the medication for PD, levodopa. In addition, recordings were collected from 11 subjects who were non-impaired with PD; these recordings served as a control of the measurement process. All recordings were collected using a high quality headset microphone at 44100 Hz sampling rate and 16 bits/sample quantization. A total of 113 vowel recordings were collected through this procedure. Two-second samples from the middle of each vowel recording were extracted for analysis, and perceptual judgments of each segment were provided by 3 listeners (graduate students in the Speech Language Pathology program at Western University). The average of the three listener ratings served as the overall quality rating of the vowel recordings. More details can be found in [48].

3.2.2 Features and their computation

Subjective ratings obtained through the procedure outlined in the previous section were used to benchmark the performance of the objective measures. Prior to feature extraction, the sustained vowel recordings were decimated to a 16 kHz sample rate. Features for the objective measures were taken from GFCC, LCQA, CPPs, and HNR. One of the objective metrics included the 60 features of GFCC. Another metric contained the 40 features of LCQA. A third metric was formed by adding HNR and CPPs to the LCQA group, and it is named the combined metric. The 113 sample database was divided into 2 datasets. The first dataset contained 80% of the whole dataset or 91 samples, while the test dataset contained 20% of the data or 22 voice samples. Machine learning algorithms were applied to the three aforementioned metrics to estimate the quality of the Parkinsonian vowel records. Moreover, PCA and the feature reduction methods were employed to reduce the number of input features and reduce the overfitting.

3.2.2.1 Filterbank based features

GFCC coefficients are mainly used in computation auditory sense analysis (CASA) studies to transform signals into time-frequency (T-F) domain to perform robust speech recognition [52]. The recorded signal was segmented into frames of 256 samples, with a frame overlap

of 100 samples. Afterwards, the power spectrum of each frame was obtained after multiplying with a Hamming window. The equivalent rectangular bandwidth (ERB) filterbank was applied to the frame power spectra. In this research, 128 filters constituted the ERB filterbank, and the log filterbank energies are decorrelated using the discrete cosine transform (DCT) [52]. The frame averaged GFCCs and their first-order time differences ("delta" values) resulted in the final GFCC feature set that contained 60 features.

3.2.2.2 Modulation based features

As mentioned earlier in Chapter 2, speech-to-reverberation modulation energy ratio (SRMR) and Modulation area (ModA) are envelope based features [21]. The envelope of the waveform is extracted and passed through filterbanks specific for SRMR and ModA [30, 31]. The ratio between the low band filters, which are assumed to contain the speech energies and the high order filters which are assumed to contain the noise energies represent the quality of the sustained vowel signal.

In SRMR [30], the speech signal is processed through a 23-channel Gammatone filterbank with center frequencies ranging from 125 Hz to half the sampling rate. Hilbert transform was then applied to the filterbank outputs, to extract the temporal envelope in each channel. These envelopes have frequencies that range between 0 to 128 Hz. At this point, each envelope was filtered into eight overlapping modulation bands, with center frequencies ranging from 4-128 Hz. Finally, SRMR was computed as a ratio between the energy stored in the first four filters, which contain most of the speech energy, and the last four filters, which contain the background noise [30].

In ModA, the speech signal was decomposed using only 4 bandpass filters, and filtered signals had Hilbert transform applied to derive the band-specific temporal envelopes. Each envelope was subsequently downsampled to 20 Hz, then processed through a 1/3 octave filterbank with center frequencies ranging between 0.5 – 8 Hz. The filterbank output energies were then used to derive the area under each acoustic band, and then those areas are averaged to produce the ModA metric [31].

3.2.2.3 Traditional acoustic measures

Traditional acoustic measures include jitter, shimmer, HNR, and CPP. Jitter is defined as the cycle-to-cycle variation of the fundamental frequency, while the relative jitter is the ratio between the absolute jitter and the average fundamental frequency [53]. Shimmer is defined as the variability of the peak to peak amplitude in decibels, while relative shimmer is the ratio between the absolute shimmer and the average amplitude [53]. HNR measures the

ratio between the periodic and non-periodic components of the signal [54]. It quantifies the relationship between the periodic components (Harmonics) and the aperiodic components (noise) of the signal [48]. Finally, the cepstral peak prominence (CPP) is defined as the difference between the peaks of the cepstrum and its linear regression function [25].the cepstrum is defined as a homomorphic transformation, transforming the convolution of a source and a filter into a sum, which can efficiently separate them [26]. A commonly used feature extraction method based on cepstrum analysis is the cepstral peak prominence (CPP). CPP is defined as the difference between the peaks of the cepstrum and its linear regression function [25].

These traditional acoustic measures were computed from the sustained vowel records using the Praat software package [55]. The records were analyzed using a Praat script, and a report of voice characteristics was generated. Traditional acoustic features were extracted from the voice report.

3.2.2.4 Linear prediction – based features

The LP-based feature extraction methodology is presented in Low Complexity Quality Assessment (LCQA) proposed by Grancharov et al. [34]. The central idea of LCQA is to extract statistical features of the speech signal [34]. Each speech recording was segmented into 20 ms non-overlapping frames, and an 18th order LP model was computed for each frame, and a vector of features is extracted from each frame. This feature' vector incorporates 10 features which are the spectral flatness, the excitation variance, the signal variance, the spectral centroid, and the spectral dynamics for each frame in addition to the first derivative of each of the aforementioned features [33]. At this point, the statistical properties of each one of the 10 features are calculated across all the frames; these statistical features include mean, variance, skew, and kurtosis [34]. This yields to the formation of a vector of size 40×1 for each speech signal record[33, 52].

3.2.3 Feature mapping

While HNR, CPP, SRMR, and ModA are single numbers that represent the predicted speech quality, the GFCC, and LCQA are multi-dimensional feature vectors. Mapping algorithms aim to generate a function that assimilates the multi-dimensional feature vectors to match the subjective scores. Linear regression (LR) and the support vector regression (SVR) [56] were applied to map the acoustic features to the subjective scores of the quality of Parkinsonian sustained vowels. Principal component analysis (PCA) and the feature selection and reduction method were used to reduce the dimensionality of the input features and reduce

the overfitting of the obtained objective scores. [43, 44]

3.3 Results

3.3.1 Subjective results

Intra-rater reliability of the perceptual judgement of voice quality was assessed using intraclass correlation coefficient (ICC) [48]. Each rater was assessed using average agreement in two-way mixed model. The average ICC across all raters = 0.754 (95% CI:0.378 − 0.903) [48], which is considered to be a moderate intra-rater reliability. Inter-rater reliability across the 3 subjective estimators was assessed using average consistency in a two-way random model, average ICC = 0.826 (95% CI:0.770 − 0.870) [48], which can be interpreted as good inter-rater reliability.

Paired sample t-tests showed that there were no statistically significant differences between PD vowel quality ratings on and off Levodopa. In other words, when the PD patient cohort was considered as a whole, the PD medication did not have any influence on their vowel quality. An interesting finding does emerge, however, when PD group is divided into two groups: those with poor perceived voice quality and those with good perceived voice quality in the off-medication condition. There was a significant improvement in perceived vowel quality for the poor quality group with the administration of medication. This interaction between off-medication voice quality and the improvement post-medication is shown in Fig. 3.1. It can be seen that patients who have low sustained vowel quality ratings before taking levodopa have a high improvement in voice quality after taking the medication. On the other hand, people who have high voice quality ratings before taking the medication have a statistically insignificant change in voice quality. These results highlight the need for either subjective or objective assessment of PD voice quality, in order to predict the effectiveness of levodopa medication on voice quality.

3.3.2 Objective results

Fig. 3.2 displays sample spectrograms associated with sustained vowel samples collected from 2 subjects in the database. Fig. 3.2c displays the spectrogram of normal control subject with high subjective quality score. This record has a relative jitter of 0.29, a relative shimmer of 2.99, a CPPs value of 11 dB, and HNR value of 22.8 dB. Fig. 3.2d displays the spectrogram of a subject impaired with PD. This subject has been off Levodopa medication and has a low sustained vowel quality. This record of the Parkinsonian subject has a relative jitter of 1.019, a relative shimmer of 12.128, a CPPs value of 15.5 dB, and HNR value of 14.23 dB.

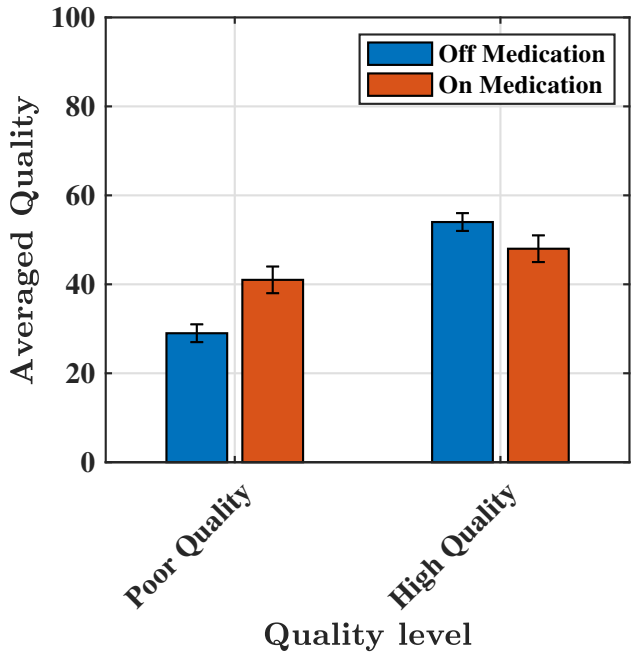
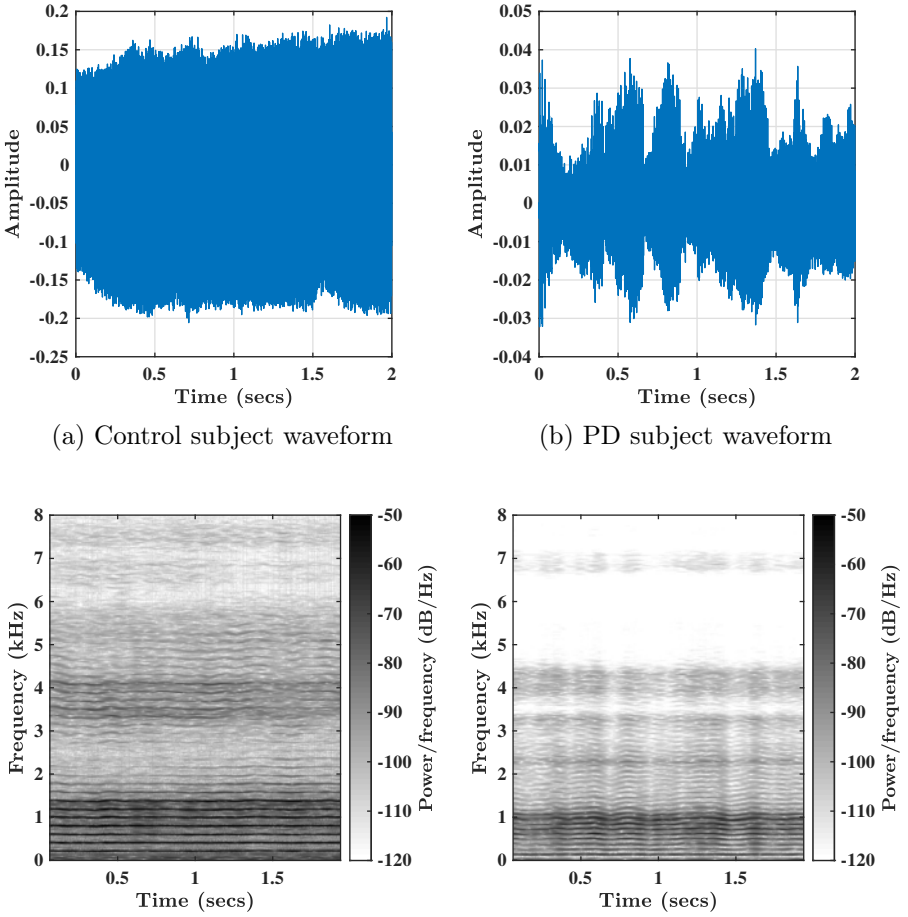


Figure 3.1: Averaged quality subjective scores off and on medication

Jitter and shimmer had poor correlation values with the subjective scores of the quality of sustained vowel records, and therefore, they were not considered to be reliable objective metrics of the quality of Parkinsonian sustained vowels.

Table. 3.1 shows (a) the correlation values between the objective scores and the subjective perceived quality ratings using different metrics, and (b) standard deviation of prediction error (SDPE) given by $SDPE = \hat{\sigma}_s\sqrt{1-\rho^2}$, where $\hat{\sigma}_s$ is the standard deviation of the subjective speech quality scores, and ρ is the correlation coefficient between the true and predicted quality scores [57]. It must be noted here that while high correlation coefficients between objective and subjective measures is desirable, a big difference between the correlation coefficients for training and test datasets is an indication of overfitting.



(a) Control subject waveform

(b) PD subject waveform

(c) Control subject, high quality

(d) Parkinsonian subject, poor quality

Figure 3.2: Waveforms and spectrograms of selected sustained vowels recordings from subjects in the database. Panel (a) and (c) represent the waveform and the spectrogram for the sustained vowel "\$a$" collected from a normal control subject, while (b) and (d) represent the waveform and the spectrogram for "\$a$" sustained vowel collected from a PD subject who has been off his medication

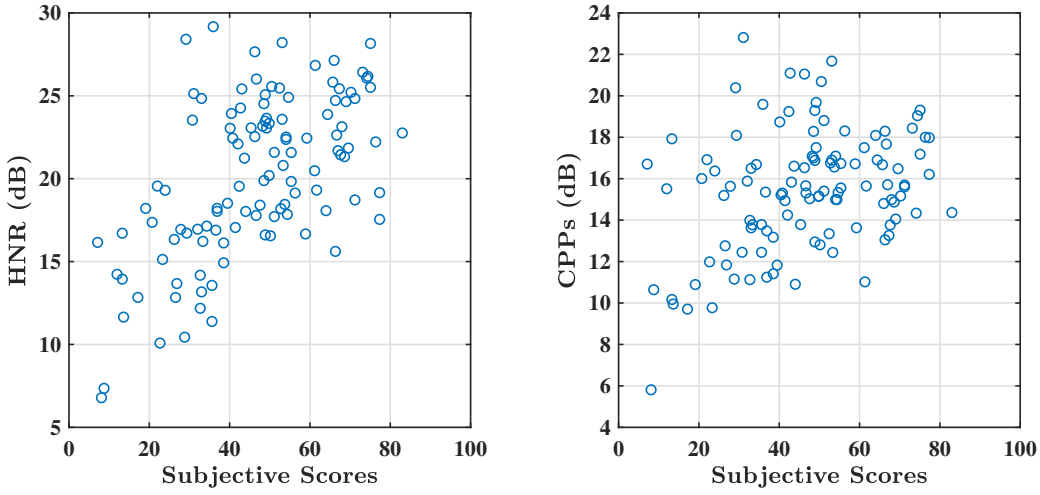
3.3.2.1 Unmapped objective metrics

For the first four objective metrics, SRMR, ModA, CPPs, and HNR each of them contains one feature only, which means there is no need to apply machine learning algorithms on them. Both of SRMR and ModA have low correlation values with the quality of sustained vowels. An explanation of that is the envelope of the high quality sustained vowel does not have a lot of variations, while the envelope of the low quality sustained vowel is exposed to high variation. Fig. 3.2a and Fig. 3.2b are examples of the high quality and low quality waveforms. This contradicts with the way SRMR and ModA measure the quality of running

Table 3.1: Correlation values of objective metrics

Metric	Full Set				PCA				Reduced			
	Correlation (Training)	SDPE (Training)	Correlation (Test)	SDPE (Test)	Correlation (Training)	SDPE (Training)	Correlation (Test)	SDPE (Test)	Correlation (Training)	SPDE (Training)	Correlation (Test)	SDPE (Test)
SRMR	_	_	0.24	17.455	_	_	_	_	_	_	_	_
ModA	_	_	−0.33	16.97	_	_	_	_	_	_	_	_
CPPs	_	_	0.34	16.91	_	_	_	_	_	_	_	_
HNR	_	_	0.59	14.52	_	_	_	_	_	_	_	_
GFCC-LR	0.86	9.15	−0.25	17.44	0.65	13.63	0.42	16.35	0.56	14.86	0.55	15.05
GFCC-SVR	0.60	14.35	0.06	17.98	0.60	14.35	0.30	17.19	0.54	15.10	0.52	15.39
LCQA-LR	0.82	10.27	0.46	16.00	0.76	11.66	0.55	15.05	0.75	11.87	0.75	11.92
LCQA-SVR	0.66	13.48	0.66	13.54	0.73	12.26	0.53	15.28	0.66	13.48	0.66	15.28
Combined-LR	0.87	8.85	0.51	15.50	0.73	12.26	0.70	12.87	0.81	10.52	0.80	10.81
Combined-SVR	0.75	11.87	0.77	11.50	0.71	12.63	0.66	13.54	0.71	12.63	0.82	10.31

speech in which the variations of the envelope and the ratio between energies in the low band and the energies in the high bands of the envelope indicates the quality of the waveform. Fig. 3.3 shows the scatter plot for HNR and CPPs against the subjective scores. The correlation between HNR and the subjective scores was 60%, while the correlation between the subjective measurements and the CPPs scores was 34% only.



(a) HNR (b) CPPs

Figure 3.3: Subjective scores vs single feature objective metrics

3.3.2.2 Objective metrics with multiple features

The full set metrics are the metrics that contained multiple features and the whole set of features are trained without any feature reduction. Using machine learning algorithms on the GFCC features did not yield a reliable metric to estimate the Parkinsonian voice quality while applying SVR on the LCQA features resulted in LCQA-SVR metric that has 66% correlation value with the subjective scores, which is higher than the correlation values obtained in previous work.

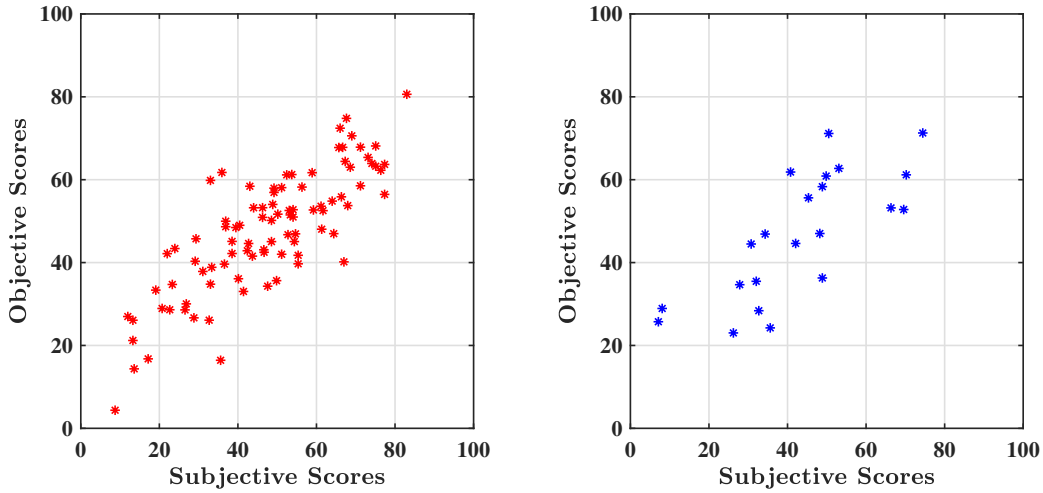
3.3.2.3 Reduced multiple feature objective metrics

Applying PCA and feature selection and reduction method to the GFCC and LCQA methods led to a reduction of the number of dimensions of the GFCC metric from 60 features to 3 features only. It also led to reducing the number of LCQA features from 40 features to 16 features only. It is noted that the metrics resulted from the feature reduction method led to higher performance than the PCA method. This enhanced the performance of most of the metrics and reduced the overfitting effect. Applying LR to the LCQA metric led to obtaining an objective metric that has a 75% of correlation with the subjective scores.

3.3.2.4 A composite objective voice quality estimator

A new metric was derived by augmenting the HNR and CPPs features with LCQA features and applying SVR and LR to extract the quality scores. The combined metric, which included 42 features, resulted in objective scores that have 77% correlation values with the subjective measures. This is noteworthy in that it is higher than all the other multiple feature metrics.

Afterwards, PCA method was applied to the features before training the model to estimate the vowel quality. The number of dimensions was reduced to 23 features, which explained 95% of the data variance. Finally applying the feature reduction method had greater improvement of the performance more than using PCA. Applying LR to the reduced combined feature set resulted in a model that estimated the quality of the vowels with 80% correlation with the subjective scores. Fig. 3.4 shows the scatter plot of the subjective voice quality scores on the X-axis against the objective scores on the Y-axis for the training and the test datasets using linear regression on the reduced combined features.



(a) Training data (b) Test data

Figure 3.4: Subjective scores vs objective scores using the reduced combined LR metric

3.4 Discussion & Conclusion

In this chapter, the perceived quality of the sustained vowels of patients with PD. Vowel samples were collected by the researchers in the School of Communications and Speech Disorders in [48] from 51 patients before and after taking the medication Levodopa. Vowel samples were also collected from 11 healthy control subjects, which resulted in the formation of a database of 113 vowels. A panel of 3 listeners rated the perceived quality of these vowel recordings [48].

Although subjective assessment of the perceived quality of sustained vowels is considered the gold standard, it is considered to be time and cost consuming process. Therefore, this research investigated the application of objective methods to evaluate the vowels quality automatically. Predictive features of the quality included GFCC, LCQA, HNR, and smoothed CPP. Machine learning algorithms SVR and LR were applied on these multidimensional features to estimate the quality objectively. Some of the features mentioned above were blended to form an objective metric to measure the quality with higher performance than the other metrics. Moreover, PCA and feature reduction were applied to reduce the number of input features to machine learning algorithms to reduce the overfitting and enhance the performance of the objective metrics. While GFCC was used in other studies to measure the quality of Parkinsonian speech and had a good performance [44, 52], this was not the case for estimating the perceived quality for Parkinsonian sustained vowels. The best performance for GFCC objective metric after feature reduction resulted in 55% of correlation between the

subjective and the objective scores. For the non-reduced full set category, applying SVR to the combination of the 40 LCQA feature, HNR, and smoothed CPP was the best objective metric of this category with a correlation value of 77% for the test dataset. The difference between the training and the test dataset was the minimum, which meant that the effect of overfitting is minimum. Applying PCA to the features led to the enhancement of most of the objective metrics. It is noted that using LR and SVR on the PCA reduced combined metric yields statistically similar results. Applying the feature reduction method on the objective features yielded a great enhancement in the performance of the objective metrics. Applying LR and SVR to the reduced combined metric yielded to statistically similar results. However, the metric with linear regression had a smaller difference between the training and the test datasets which means the smaller effect of overfitting. As a result of that, the reduced combined metric with linear regression is considered to be the best objective metric for estimating the quality of the Parkinsonian sustained vowels. In conclusion, this study presented new objective metrics to measure the perceived quality automatically. While the objective metrics presented in this chapter are shown to be more reliable in estimating the vowels perceived quality, future research involving a larger dataset of Parkinsonian sustained vowels will enhance the obtained results and ensure more generalizability. A larger dataset will also make the utilization of other machine learning algorithms such as deep learning more feasible to develop a more precise and generic metric for evaluating the quality of Parkinsonian vowels.

3.5 Summary

This chapter investigated the performance of several objective metrics to evaluate the quality of sustained vowels of people impaired with Parkinson disease. The database contained recordings collected from 51 patients before and after being on the Levodopa medication, in addition to records collected from 11 healthy subjects regarded as a control for the experiment. A panel of listeners was hired to evaluate the obtained recordings, and the mean opinion scores of the panel were the subjective evaluations of the quality of the vowels. In order to objectively estimate the quality of the vowels, acoustic and statistical features were evaluated to give a numerical estimate of the perceived quality. These features included Gammatone Frequency Cepstral Coefficients (GFCC), Low complexity quality assessment (LCQA), Harmonic to noise ratio (HNR), and Smoothed Cepstral peak prominence (CPPs). To assimilate these features to a number that represents the quality and correlates to the subjective scores, machine learning algorithms were applied on these features to estimate the vowels quality numerically. These algorithms included linear regression (LR) and support

vector regression (SVR). Moreover, feature reduction methods were applied to the features to avoid objective model overfitting and enhance the prediction capabilities for the test dataset. As a result of the procedure described in this chapter, an objective metric for estimating the quality of sustained vowels was obtained with objective scores that had a correlation with the subjective scores with a value that reached 80%.

Chapter 4

Evaluation of The Quality of Running Speech of Parkinson's Disease

4.1 Introduction

The first study reported in the previous chapter investigated objective and subjective assessment of Parkinsonian voice quality through computational and behavioural evaluation of sustained vowels respectively. Sustained vowels are often used in Speech Language Pathology due to several reasons: (a) they are controlled and free from prosodic features such as intonation, rhythm, and stress; (b) they are easy to produce and analyze; and (c) less affected by dialect. However, humans do not communicate through sustained vowels alone. Natural running or continuous speech can therefore be considered as a more ecologically valid option for both objective and subjective assessment. Similar to sustained vowels, continuous speech from PD patients has a deficit in quality compared to normal speech, as it is usually perceived as harsh and breathy [12].

Amplification devices for PD subjects are used to increase the intensity and loudness of Parkinsonian speech [18], with a potential concomitant increase in its perceived clarity and overall quality. Different amplification devices are benchmarked according to their electroacoustic performance by measuring attributes such as frequency response, sensitivity, and distortion, but these features do not quantify the effect of amplification on the perceived speech quality [8]. This necessitates more perceptually-relevant performance evaluation of the amplification devices when they are used in PD speech context [8]. Assessing the quality of amplified Parkinsonian speech quality helps with the clinical research of PD speech impairment treatment, gives a method to provide a better assessment of the patient's needs from the amplification devices, and helps to enhance the lifestyle of people impaired by PD

[18].

While objective speech quality metrics are routinely used to assess the quality of telecommunication and assistive hearing aid devices [21], few studies have applied objective quality metrics to characterize Parkinsonian speech [45, 47]. In addition, objective assessment of the quality of Parkinsonian speech in conjunction with the amplification devices has not been investigated before.

In summary, quality assessment of Parkinsonian speech is important in evaluating the effectiveness of the clinical treatment of PD speech impairments, and in characterizing the impact of assistive amplification devices on Parkinsonian speech [58]. Existing objective methods of Parkinsonian speech quality assessment correlate poorly with subjective judgements. This chapter introduces the methodology and the system model for the subjective and objective assessment of amplified Parkinsonian speech. Results from this study are discussed in Section. 4.3. Finally, Section. 4.4 concludes the research in this chapter and discusses future work.

4.2 Methods

4.2.1 Speech recordings and subjective evaluation

Subjective data collection procedures outlined in this chapter received ethics approval from Western University's health sciences research ethics board.

This study included 11 individuals with mild to moderate hypophonia and mild to moderate idiopathic PD (aged 58-80 years; M = 70.9 years; 10 men, 1 woman). The average number of years since diagnosis of PD was 6.7 years (range = 1-16 years). Participants with PD were tested approximately 1 hr after their regularly scheduled anti-Parkinson medication. Two of the participants with PD were not on anti-Parkinson medications, whereas all other participants were on levodopa-carbidopa medication. None of the participants with PD had been previously prescribed a speech amplification device. The participants had no prior history of speech, language, or hearing problems. The Mini Mental State Examination [59] was used to exclude participants with dementia (cutoff score = 26/30). All participants with PD passed a bilateral 30 dB HL hearing screening at 500, 1000, and 2000 Hz. None of the participants with PD had received surgical treatment for their PD (i.e. deep brain stimulation).

Speech recordings were collected from eleven PD subjects and ten age-matched normal controls in different environmental and amplification conditions [8, 18]. The control group had an age range of $59 - 86$ years (mean = 71.4 years). Both PD and control speakers

were seated in a sound-treated booth and completed two speech tasks in two environmental conditions: unscripted conversation in quiet and in the presence of background noise, and reciting a given sentence in quiet and noisy environments. For the sentence recordings, the subjects repeated a sentence consisting of 5 to 15 words, which was selected randomly from a database that contains 1100 sentences [60]. For speech recordings in noisy environments, multi-talker babble was generated from two loudspeakers that were placed at a constant distance from the subject. The background noise level was calibrated to 65 dB SPL at the recording microphone, which was placed 4 m from the subject. The examiner was at a fixed interlocuter distance of 1.5 meters throughout the experiment. The participants received no feedback about their speech during the experiment. All speech recordings were sampled at 16 kHz and quantized at 16 bits/sample.

The aforementioned speech recordings were obtained with no amplification, and with the aid of seven different amplification devices: Addvox (Addvox, Waltham, MA), Boomvox (Griffin Laboratories, Temecula, CA), Chatterbox (Connections Unlimited, Nashville, TN), Oticon Amigo (Oticon, SmØrum, Denmark), Sonivox (Griffin Laboratories, Temecula, CA), Spokeman (KEC Innovations, Singapore), and Voicette (Luminaud Inc., Mentor, OH)[18]. Thus, a database of 21(11 PD + 10 control speakers) $\times 2$ (conversation and sentence speech tasks) $\times 2$ (quiet and noisy environments) $\times 8$ (amplification options) = 672 speech recordings were created for this study.

Ten normal hearing naive listeners with an age range of $21 - 25$ (mean = 22.7 years) evaluated the quality of each of the 672 recordings. For the conversation samples, a single 5 to 15-word sentence was extracted for the speech quality rating. Listeners were asked to rate the perceived quality of the recording on a visual analogue scale, with 0 and 100 representing the lowest and highest sound quality scores, respectively. For reliability purposes, 20% of the sentences were re-rated by the listeners. Intra-rater and inter-rater reliability, based on correlations (ICC), was found to be 0.90 and 0.97, respectively.

4.2.2 Features & their computation

As the focus is on the objective estimation of continuous speech quality, traditional measures such as jitter, shimmer, and HNR were not considered. The CPP measure was included in this investigation as it showed promise in earlier studies with Parkinsonian speech [47]. In particular, the smoothed CPP value was computed from each speech recording following the algorithm given in [61]. In addition to the CPP, speech signal parametrization through filterbank analyses, modulation domain analyses, and Linear Prediction (LP) analyses were also explored, computational details of which are given in the following subsections.

4.2.2.1 Filterbank-based features

The MFCCs are popularly used as features in speech recognition systems [26], and have been shown to perform well in objective speech quality prediction [62]. The computation of MFCCs followed the procedure used in automatic speech recognition (ASR) research [26]. The speech signal was segmented into frames of 256 samples, with a frame overlap of 100 samples. The power spectrum of each frame was then obtained after multiplying with a Hamming window. The triangular mel filterbank was applied to the frame power spectra. In this chapter, 40 filters constituted the mel filterbank, where the first 13 filters were linearly spaced, and the last 27 filters were logarithmically spaced [26]. The log filterbank energies were decorrelated using the discrete cosine transform (DCT) and the lower 13 coefficients were retained [46, 26]. The frame-averaged MFCCs and their first-order time differences ("delta" values) comprised the final MFCC feature set.

In addition to the MFCCs, cepstral coefficients extracted using the Gammatone filterbank were also utilized as a separate feature set. The Gammatone filterbank better approximates the auditory filterbank in comparison to the mel filterbank. As such, the cepstral coefficients extracted using the Gammatone filterbank have been shown to produce better speech recognition performance than MFCCs [27]. The computation of Gammatone frequency cepstral coefficients (GFCCs) followed the same steps as that of MFCC, except the mel filterbank was replaced by the Gammatone filterbank, which was generated using Malcolm Slaney's auditory toolbox [29]. Following Shao et al.'s [27] ASR research, 30 of the frame – averaged lower GFCCs and their first-order time differences were included in the GFCC feature set.

4.2.2.2 Modulation-based features

The speech-to-reverberation masking ration (SRMR) is an objective technique that was developed by Falk *et al.* [30] to measure the intelligibility of reverberant speech. The authors assume that the change of slow temporal envelope modulations provides a useful objective estimation of speech quality and intelligibility. It is known that clean speech has temporal envelopes with frequencies ranging from $2 - 20$ Hz, with peaks at around 4 Hz, which represent the syllabic rate of natural speech [30].

In this method [30], the speech signal is applied to a 23-channel Gammatone filterbank with center frequencies ranging from 125 Hz to half the sampling rate. Hilbert transform is then applied to the filterbank outputs, to extract the temporal envelope in each channel. These envelopes have frequencies that range between 0 to 128 Hz. At this point, each envelope is filtered into eight overlapping modulation bands, with center frequencies ranging from 4-128 Hz. Finally, SRMR is computed as a ratio between the energy stored in the first

four filters, which contain most of the speech energy, and the last four filters, which contain the background noise [30].

Another measure based on modulation-domain analysis is Modulation Area (ModA) parameter. There are some similarities between ModA and SRMR [21]. While SRMR depends on calculating the ratio between the energy in the lowest temporal bands and the highest temporal bands, ModA accommodates the reality that reverberation smears the speech signal envelope, which will lead to a decrease in the modulation area. Unlike SRMR, the speech signal is decomposed into only 4 filters, and then Hilbert transform is applied to derive the band-specific temporal envelope. Each envelope is subsequently downsampled to 20 Hz, then processed through a 1/3 octave filterbank with center frequencies ranging between 0.5 – 8 Hz. The filterbank output energies were then used to derive the area under each acoustic band, and then those areas are averaged to produce the ModA metric [31].

4.2.2.3 Linear prediction – based features

We followed the LP-based feature extraction methodology in Low Complexity Quality Assessment (LCQA) proposed by Grancharov et al. [34]. Each speech recording was segmented into 20 ms non-overlapping frames, and an 18th order LP model was computed for each frame. The LP model parameters were then used to calculate the frame-wise spectral flatness, the excitation variance, the signal variance, the spectral centroid, and the spectral dynamics (see Grancharov et al. [34] for computational formulae). These five quantities, together with their first order differences, constituted the 10-dimensional parameter vector per frame [34, 33]. The statistical properties of these parameters across the entire sentence (*viz.* mean, variance, skewness, and kurtosis) resulted in the final 40×1 LCQA feature vector for each speech recording [33].

4.2.3 Feature Mapping

While SRMR, ModA, and CPP are single numbers that represent the predicted speech quality, the MFCC, GFCC, and LCQA are multi-dimensional feature vectors. Mapping algorithms aim to generate a function that assimilates the multi-dimensional feature vectors to match the subjective scores. Commonly used feature mappers include linear regression (LR), the support vector regression (SVR) [36], and the Gaussian Process Regression (GPR) [39].

Recent developments in machine learning for classification and regression have focused on deep learning. In deep learning [40], the learning process is divided into multiple layers where features are extracted from each layer. Deep learning then uses the backpropagation method

to train these multilayer architectures and adapt them to extract new features to minimize the error function. One of the key characteristics of deep learning is that there is no need for a human intervention to design these layers of neurons since they are learned from the input data alone. Deep neural networks (DNNs) have proved to be a state-of-the-art tool in speech recognition, and hence, they have been investigated in this research. In this research, adaptive moment estimation (Adam) optimizer was used in the learning stage; more details about Adam optimizer can be found in [41, 42]. The DNN structure used in this chapter had four layers: (a) the input layer which intook the feature vectors; (b) two hidden layers where the first hidden layer was formed of 25 neurons while the second layer contained 12 neurons; and (c) the output layer had 1 neuron which resulted in the predicted quality of the speech signal under test. It is pertinent to note that the number of hidden layers and neurons per layer was kept small to avoid overfitting of the model. Adam optimizer computes individual adaptive learning rates for different parameters from estimates of first and second moments of the gradients, which leads to a faster training of the neural network and anoidance of reaching a local minimum of the cost function [42].

The features extracted from the 672 speech recordings and their corresponding subjective quality scores were divided into two sets, with 80% of the speech stimuli comprising the training dataset and the remaining 20% comprising the test dataset. This partitioning was done randomly, and the test dataset was isolated from the training dataset to ensure the generalization of the machine learning algorithm. Five-fold cross-validation was performed within the training dataset for fine-tuning the parameters and hyper-parameters of the feature mapper.

4.2.4 Feature selection and reduction

A higher dimensionality of the feature vector may cause overfitting. In such situations, the feature dimensionality of the training set must be reduced before applying the machine learning algorithm to avoid overfitting. To accomplish this goal, the correlation coefficient between each feature in the training set and the subjective scores was obtained, and then the features were rearranged according to their correlation values from the highest to the lowest [44, 52]. A ten-fold cross-validation procedure was then followed, wherein the training dataset was randomly split into a training subset and a validation subset for each fold. The minimum mean square error (MSE), post-fitting for both the training and validation subsets across the ten folds, was logged as the number of features were varied. The feature subset that resulted in the lowest difference between the training and the validation MSE values was chosen as the reduced model feature set.

4.3 Results

4.3.1 Subjective results

Fig. 4.1 displays the averaged speech quality scores for speech samples collected from control and Parkinson's subjects in the four experimental scenarios. The following key observations can be deduced from Fig. 4.1: (i) speech from subjects with Parkinson's disease received lower quality ratings in comparison to control subjects' speech, (ii) speech quality ratings were lower in the presence of background noise, and (iii) speech quality ratings were impacted by the amplification device.

Intra-rater reliability of the perceptual judgements of speech quality was assessed using intraclass correlation coefficient (ICC) [18]. Each rater was assessed using average agreement in two-way mixed model. The average ICC across all raters $= 0.755$ (95% CI:$0.638 - 0.894$) [18], which is considered to be a good intra-rater reliability. Inter-rater reliability across the 3 subjective estimators was assessed using average consistency in a two-way random model, average ICC $= 0.801$ (95% CI:$0.708 - 0.868$) [18], which can also be interpreted as a good inter-rater reliability.

Repeated measures ANOVA was performed on the subjective speech quality data to assess the statistical significance of the results, with the speech task (sentences vs conversation), background noise (no noise vs 65 dB SPL multi-talker babble), and speech amplification device as the within-group variables, and the speaker type (control vs. Parkinson's) as the between-groups variable [18]. Greenhouse-Geisser corrections were applied when the sphericity condition, as assessed by Mauchly's test, was violated.

ANOVA results showed that there were significant main effects of the speaker group $\big(F(1,18) = 25.26, p < 0.001, \eta_p^2 = 0.584\big)$, background noise $\big(F(1,18) = 227.75, p < 0.001, \eta_p^2 = 0.927\big)$, and device type $\big(F(7,126) = 37.16, p < 0.001, \eta_p^2 = 0.674\big)$. There was no significant main effect of speech task $\big(F(1,18) = 1.55, p = 0.229, \eta_p^2 = 0.079\big)$, indicating that the raters were consistent in judging the talker speech quality whether it was an isolated sentence or a sentence extracted from the conversation. There were no significant two-way interactions between speaker group by noise $\big(F(1,18) = 0.002, p = 0.964, \eta_p^2 = 0.00\big)$, speaker group by speech task $\big(F(1,18) = 0.878, p = 0.361, \eta_p^2 = 0.046\big)$, and speaker group by device $\big(F(7,126) = 1.30, p = 0.254, \eta_p^2 = 0.068\big)$, indicating that none of these variables differentially affected the perceived quality of speech from control and Parkinsonian subjects. There was a significant two-way interaction between the device type and noise variables $\big(F(3.90, 70.21) = 7.80, p < 0.001, \eta_p^2 = 0.302\big)$. This interaction stemmed from the differential quality ratings associated with the ChatterVox device. As can be seen from Fig. 4.1, the

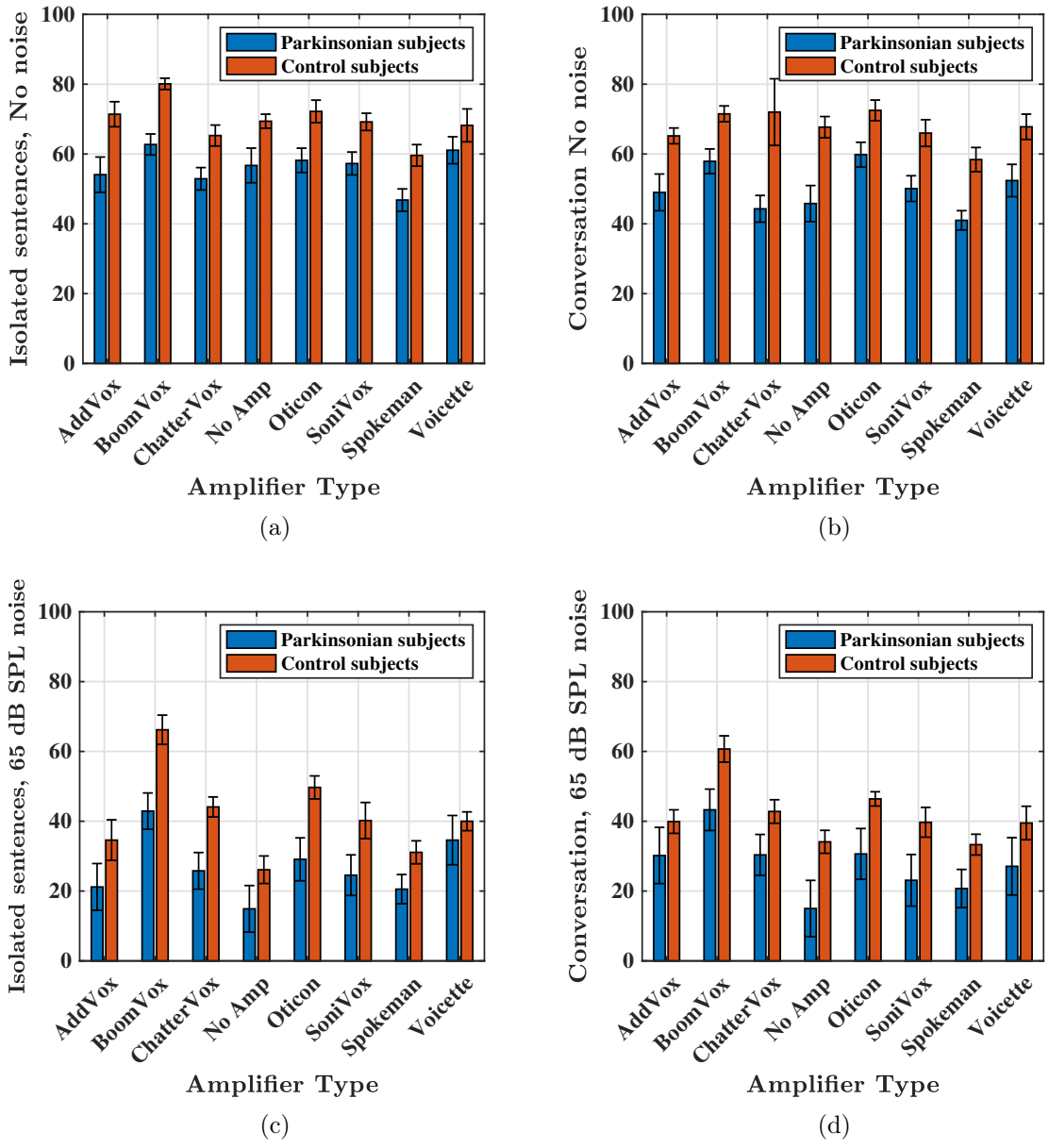
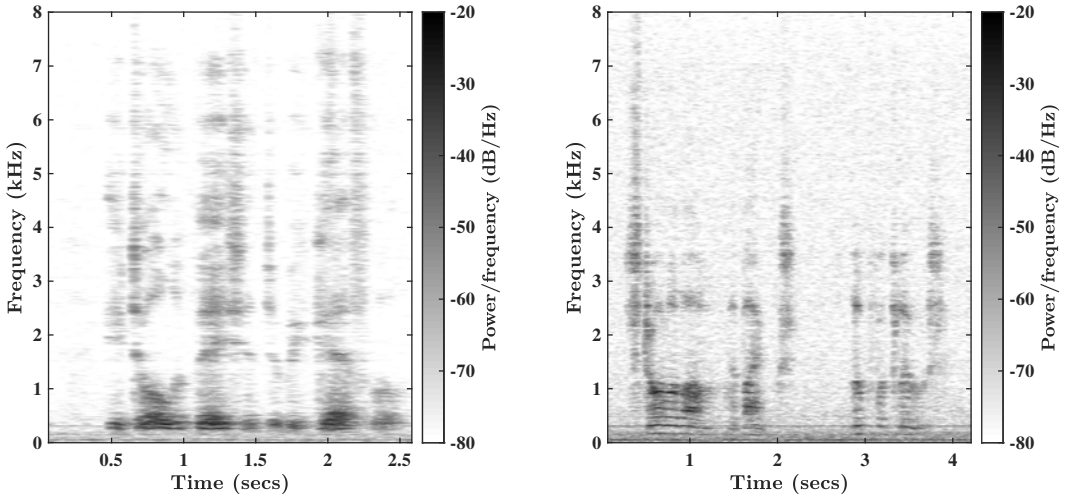


Figure 4.1: Averaged subjective speech quality ratings for control and Parkinsonian speech samples, with the error bars denoting one standard deviation



(a) BoomVox, high quality

(b) Spokeman, poor quality

Figure 4.2: Spectrograms of selected speech recordings from a Parkinson's subject in quiet condition. Panel (a) represents the spectrogram for "He told the patient to be careful", and panel (b) represents the spectrogram for "Stroll along the banks, look for clues"

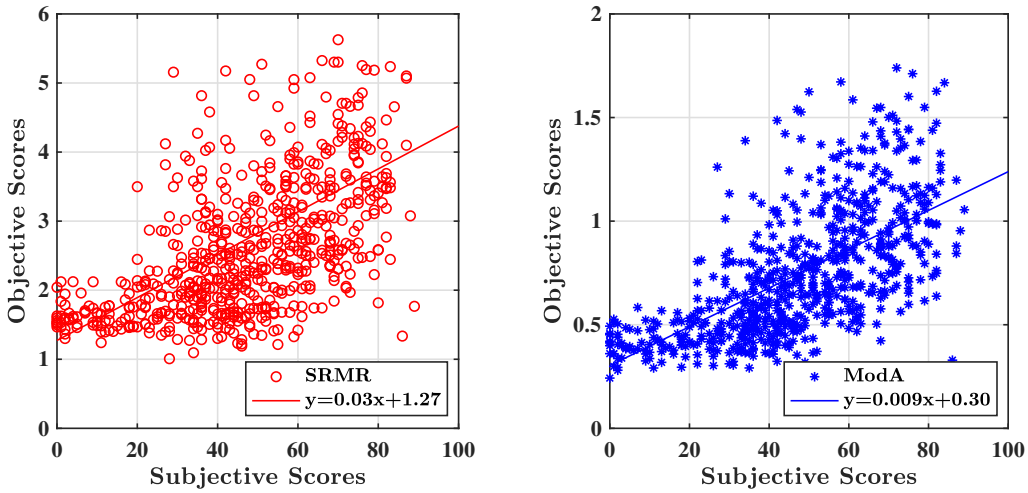
ChatterVox device received lower quality ratings than no amplification in quiet conditions (Fig. 4.1a & Fig. 4.1b), but higher ratings in conditions involving background noise (Fig. 4.1c & Fig. 4.1d). Finally, no significant three-way or four-way interactions were found.

Post-hoc analyses with Bonferroni corrections revealed that the BoomVox received significantly higher speech quality rating than other devices and that there were no statistically significant differences among the devices with the three lowest quality scores [18].

4.3.2 Objective results

Fig. 4.2 displays sample spectrograms associated with speech samples collected from a subject with Parkinson's disease in the quiet condition. Fig. 4.2a displays the spectrogram of an isolated sentence produced by the subject while utilizing the BoomVox amplifier. Fig. 4.2b displays the spectrogram of a different isolated sentence produced by the same talker, but with the Spokeman amplifier, which received a poorer speech quality rating. Visual inspection of these spectrograms reveals broadband background noise with the Spokeman amplifier.

The subjective scores of speech recordings served as a reference for benchmarking the objective metrics in this research. Two figures of merit were used: (a) the Pearson correlation coefficient between the true and predicted subjective scores, and (b) standard deviation of prediction error (SDPE) given by $SDPE = \hat{\sigma}_s\sqrt{1-\rho^2}$, where $\hat{\sigma}_s$ is the standard deviation



(a) SRMR (b) ModA

Figure 4.3: Scatter plot between the objective and subjective scores for all the speech recordings in the database. Data were plotted for the four conditions, *viz* isolated sentences in quiet and in 65 dB SPL background noise (labeled SQT (no noise) and SQT (noise) respectively), and sentences extracted from conversation in quiet and in 65 dB SPL background noise

of the subjective speech quality scores, and ρ is the correlation coefficient between the true and predicted quality scores [57].

4.3.2.1 Unmapped objective metrics

As indicated earlier, SRMR, ModA, and CPP report a single number predictive of the subjective speech quality. As such, no mapping algorithm was applied to these metrics. Analyses showed that the correlation coefficient between the SRMR scores and the subjective scores was only 0.5. However, when averaging the scores per device and the background noise conditions, the correlation increased to 0.89. As for the ModA technique, the overall correlation with the subjective metrics was 0.64, but it reached 0.88 when the scores where averaged per device and background noise conditions. In the case of CPP, the correlation between the objective scores and subjective scores was 0.35, and it reached 0.59 when the scores were conditionally averaged. Fig. 4.3 shows the scatter plots between the SRMR and ModA scores against the subjective scores for the entire database. The greater dispersion in the scatter plot, and the estimator bias for the poorer quality subjective scores are evident in this figure.

4.3.2.2 Objective metrics with multiple features

Multiple features objective metrics are those metrics in which each has a group of features to represent the quality of Parkinsonian speech. As such, a machine learning algorithm has to be applied to map the feature vector extracted from each speech recording to the corresponding subjective scores. The feature vector dimensions for LCQA, MFCC, and GFCC were 40, 26, and 60,f respectively, which were mapped separately using four learning algorithms *viz.* LR, SVR, GPR, and DNN.

Table 4.1 shows the correlation coefficients between the true subjective scores and predicted subjective scores through feature mapping for all feature vector - feature mapping combinations and for both training and test datasets. The corresponding SDPE values were also included in this table. It can be seen that the LR method has high overfitting for all the non-reduced objective metrics because of the gap between the correlation values associated with the training dataset and the test dataset. A similar phenomenon can be noted with the MFCC-GPR and MFCC-DNN conditions. As expected, the SDPE values are substantially higher for the test dataset in overfitting cases (e.g., MFCC-DNN).

A number of feature vector-feature mapping combinations have resulted in similar correlation values for the test dataset. The Steiger's Z test [63] was therefore employed to assess the statistical significance of differences between different correlation coefficients. Results from this analyses showed that MFCC-GPR, GFCC-GPR, ad GFCC-DNN performed statistically similar in predicting subjective scores. Of these, GFCC-DNN had the lowest difference in the correlation coefficient between training and test datasets.

4.3.2.3 Reduced multiple features objective metrics

By applying the feature selection and reduction method mentioned in subsection. 4.2.4, the number of features for LCQA, MFCC, and GFCC were reduced to 7, 16, and 11, respectively. Fig. 4.4 displays the mean square error (MSE) between the true and predicted subjective speech quality scores for both the training and the test databases when plotted against the number of selected features from the GFCC feature vector. As expected, the MSE for the training dataset continues to decrease, while the test dataset error decreases until the number of selected features become 11. After this point, the test dataset MSE increases, which means that increasing the number of features beyond this point will yield to an increased chance of overfitting. This implied that the selected 11 features would avert overfitting and potentially lead to better results.

The last four columns in Table 4.1 show the correlation values resulting from feature set reduction for different combinations of feature vectors and mappers and for both training

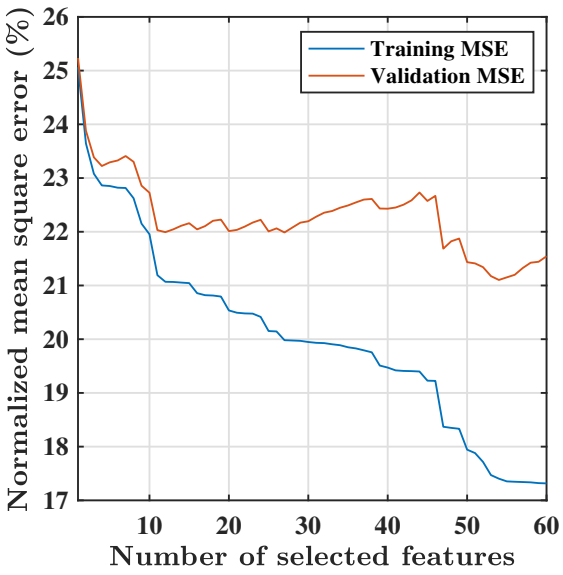


Figure 4.4: The normalized mean square error (MSE) between actual and predicted speech quality scores as a function of the number of GFCC features. The MSE data for training and validation datasets are shown separately, and greater separation between these two lines is a potential indicator of overfitting

and test datasets. It can be observed that using the feature selection and reduction enhanced the performance LR-based metrics significantly. For example, GFCC test correlation was increased from 0.70 to 0.78, while the overfitting between the training and the test datasets was reduced from 0.16 to 0.02. This was also the case for MFCC and LCQA where their test correlation values increased from 0.66 and 0.73 to 0.75 and 0.75, respectively. It is noted that the performance of metrics utilizing SVR and DNN mappers was not affected significantly when applying feature reduction. The performance of GFCC remained at 0.80 correlation for both the reduced and non-reduced versions. Feature selection and reduction improved the performance of the metrics using GPR in terms of overfitting reduction. The overfitting between the training and the test dataset was reduced from around 0.1 to 0.03 only in the case of GFCC, overfitting was reduced from 0.12 to 0.07 in the case of MFCC, and it was reduced from 0.08 to 0.01 in the case of LCQA.

Statistical analyses using Steiger's Z test showed that more feature vector and feature mapper combinations resulted in statistically similar performances with reduced feature sets. Once again, GFCC-DNN had the lowest difference in correlation coefficients between training and test datasets, as well as a lower SDPE value. This finding is consistent with speech quality and automatic speech recognition research [21, 26], in that the GFCCs appear to capture perceptually salient features perhaps due to their better approximation of the audi-

Table 4.1: Correlation coefficients and SDPE values between objective and subjective data. Bold correlation coefficients represent feature vector and mapper combinations that performed statistically similar to the test dataset

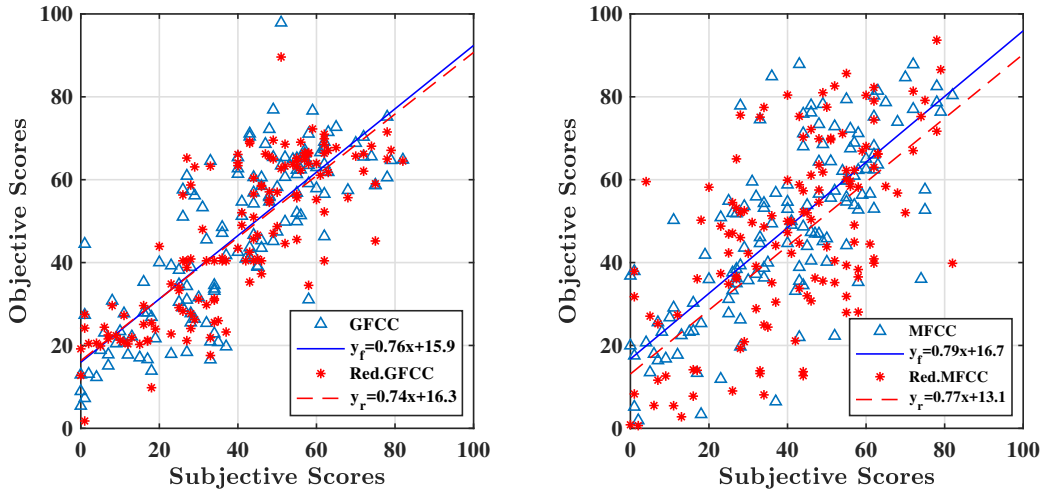
	Non-reduced feature set				Reduced feature set			
Metric	Correlation (Training)	Correlation (Test)	SDPE (Training)	SDPE (Test)	Correlation (Training)	Correlation (Test)	SDPE (Training)	SDPE (Test)
LCQA-LR	0.81	0.66	0.12	0.16	0.76	0.75	0.14	0.14
MFCC-LR	0.80	0.73	0.13	0.14	0.78	0.75	0.13	0.14
GFCC-LR	0.86	0.70	0.11	0.15	0.80	0.78	0.13	0.13
LCQA-SVR	0.77	0.72	0.13	0.14	0.79	0.77	0.13	0.13
MFCC-SVR	0.79	0.74	0.13	0.14	0.88	**0.80**	0.10	0.13
GFCC-SVR	0.89	0.77	0.10	0.13	0.82	0.78	0.12	0.13
LCQA-GPR	0.86	0.77	0.11	0.13	0.81	**0.80**	0.12	0.13
MFCC-GPR	0.93	**0.81**	0.10	0.12	0.89	**0.82**	0.10	0.11
GFCC-GPR	0.90	**0.79**	0.10	0.13	0.83	**0.80**	0.12	0.13
LCQA-DNN	0.81	0.78	0.12	0.13	0.81	**0.79**	0.13	0.12
MFCC-DNN	0.95	0.75	0.07	0.14	0.81	**0.79**	0.12	0.13
GFCC-DNN	0.83	**0.80**	0.12	0.13	0.81	**0.81**	0.12	0.12

tory filterbank characteristics. Fig. 4.5 shows GFCC-DNN and MFCC-DNN scores against the subjective quality scores with and without the feature reduction, for the test dataset. It is evident that the feature reduction led to a reduction in the data spread and variability for both MFCC-DNN and GFCC-DNN.

4.3.2.4 A composite objective speech quality estimator

In this section, a metric was derived by augmenting the GFCC feature vector with CPP, LCQA, SRMR, and ModA parameters and applying the feature mapping procedure. The combined feature set, which included 103 features, was first subject to feature reduction in a similar manner as described in the previous section. The number of features was reduced from 103 to 22 features through feature reduction, which included 7 features from GFCC, 12 features from LCQA, and the CPP and ModA values. Table 4.2 shows the correlation coefficient and SDPE values between the scores obtained by this composite objective metric and subjective scores for both the training and the test datasets. It is noted that this model has a higher test correlation value of this dataset more than any other metric mentioned in the previous sections, which was statistically significant. Fig. 4.6 shows the plot of the subjective scores on the x-axis against the composite objective scores for the y-axis for each of the training and the test datasets when GPR was utilized as the feature mapper.

It can be observed from the scatter plots between the objective and subjective data that there sometimes is a bias in estimating the poorer quality Parkinsonian speech, especially for those which have subjective quality value less than 0.2. This effect can be observed clearly in



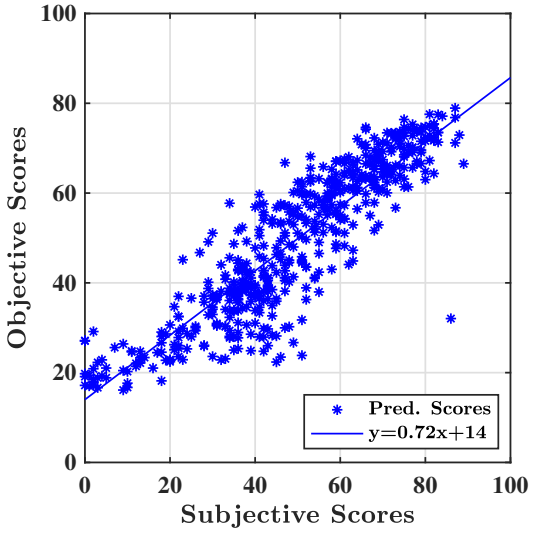
(a) GFCC (b) MFCC

Figure 4.5: Scatter plot between the objective and subjective data using deep learning for the test dataset. The data is plotted for both unreduced and reduced GFCC and MFCC feature vectors. y_f is the vector of obtained scores from the linear regression for the non-reduced model, while y_r is the vector of obtained scores from the linear regression of the reduced model

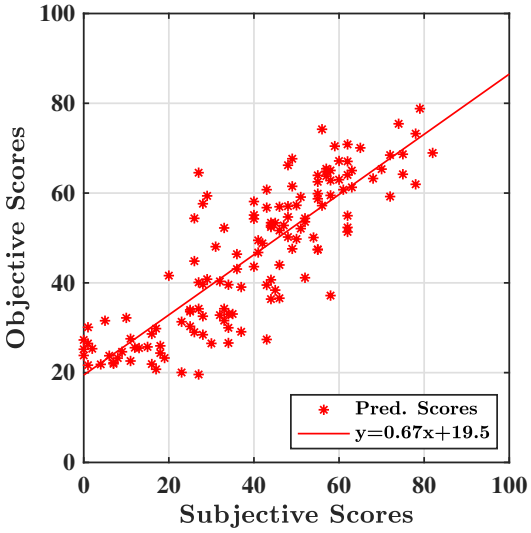
Table 4.2: Correlation values of the combined features metric

Regression algorithm	Correlation (Training dataset)	Correlation (Test dataset)	SDPE (Training dataest)	SDPE (Test dataset)
GPR	0.89	0.85	0.10	0.11
SVR	0.80	0.85	0.13	0.11
DNN	0.89	0.84	0.10	0.11

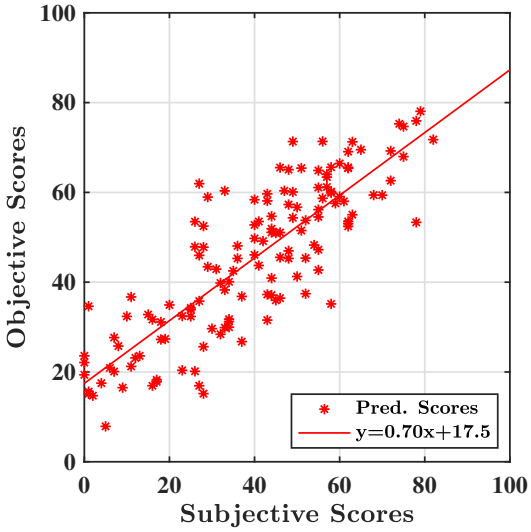
Fig. 4.6, where there are no speech recordings that have an objective (i.e. predicted) score less than 0.2. After investigation, it was discovered that this was related to the characteristics of the background noise in which the speech recordings were obtained. The noise used while collecting the speech recordings was non-stationary multi-talker babble with overlapping temporal modulation and spectral properties with natural speech. As such, the model was unable to predict low subjective speech quality scores associated with environmental conditions where SNR was 0 dB or less. In other words, the recording dominated by the multi-talker babble had similar modulation and spectral features as natural speech. In order to overcome this effect, a synthetic collection of 430 records that contained only multi-talker babble was added to the training dataset with given subjective scores of 0. It is noted that training the new database led to an enhancement of the prediction capabilities of the model towards speech recordings that have less than 0.20 subjective quality scores. Fig. 4.6d shows



(a) Training dataset



(b) Test dataset



(c) Test dataset after noise training

Figure 4.6: Subjective scores against objective scores for the combined metric using GPR

the scatter plot of the test dataset against the subjective scores after including the multi-talker babble in training. It is noted that the bias at the low quality records is reduced with the new training dataset, and the correlation value improved to 0.86. This point highlights the need for proper training database in order to effectively predict the perceived speech quality across the entire rating scale.

4.4 Discussion & Conclusion

Speech amplifiers are typically employed by people with Parkinson's disease to overcome hypophonia. In this study, the perceived quality of Parkinson's speech before and after amplification was assessed in a number of different test conditions. Speech samples from 11 Parkinson's patients and 10 age-matched healthy controls were recorded in quiet and noisy environments, with and without the aid of seven commercially available voice amplifiers. Naive listeners rated the perceived quality of these recordings. Statistical analyses of the quality rating data revealed that the quality ratings for Parkinsonian speech were significantly lower than speech quality ratings for age-matched controls. In general, the voice amplifiers enhanced the quality of Parkinsonian speech, but there were significant differences in the ratings associated with different devices. This study, therefore, highlights the need for benchmarking voice amplifiers in a perceptually relevant manner.

While subjective assessment of voice amplifier performance has high face validity, it is also time- and resource-intensive. As such, this study investigated the applicability of objective, instrumental predictors of perceived quality. Among these the CPP, SRMR, and ModA metrics are single feature objective metrics that did not require a feature mapping algorithm, which displayed modest performance in estimating the perceived quality of the amplification devices. On the other hand, LCQA, MFCC, and GFCC procedures resulted in multi-dimensional feature vectors that needed a feature mapping algorithm. In addition to these objective metrics, a composite objective metric was developed by gathering and combining a subset of the feature sets described above.

The LR, GPR, SVR, and DNN algorithms were utilized as the feature mappers. For the non-reduced multiple features objective metrics category, it was noted that applying the deep learning algorithm on the GFCC features yielded to the best performance of this category with a correlation value of 0.83 for the training set and 0.80 for the test set. The difference between the training and the test correlation values was the minimum, which implied that it was the most generalized model and the least prone to overfitting effect. As such, this metric would be more preferred than a metric applying GPR to MFCC features which resulted in 0.81 correlation value for the test dataset but had a higher difference between the training

and the test correlation values, which again is an indication of overfitting. For the reduced feature objective metrics, the metric obtained from applying the deep learning algorithm to the GFCC features was selected to be the best metric to estimate the Pakinsonian speech quality because it was least prone to overfitting.

It is noted that the reduction of features contributed to the enhancement of the correlation values obtained from applying SVR to all LCQA, MFCC, and GFCC with a significant statistical difference. In the case of applying feature reduction to the metrics using GPR, the enhancement in the test dataset correlation values was statistically similar. However, there was an enhancement in the overfitting effect by reducing the difference between the training and the test datasets. The composite metric had a statistically superior performance when compared to all the other measures explored in this study.

In order to further probe the robustness of the presented models, an additional experiment was conducted by separating the training and the test datasets such that the test set included all the data points from 4 randomly chosen subjects. This was performed to address the concern that the learning model may be influenced by the data/scores from a few subjects. This analysis was performed with the GFCC and MFCC feature sets using the GPR machine learning algorithm, in a similar manner as before. The new GFCC-GPR metric resulted in correlation values of 0.85 and 0.75 for the new training and test datasets, while the corresponding correlation values for the MFCC-GPR combination were 0.94 and 0.80. The Steiger's Z analysis revealed that the correlation coefficients obtained with this new data partitioning were statistically similar to the corresponding values in Table 4.1. These results highlight the robustness of the learning model in predicting the quality of Parkinsonian speech.

In conclusion, this study showed the differential impact of speech amplifiers on perceived Parkinsonian speech quality. It also demonstrated the applicability of instrumental metrics for benchmarking the speech amplifiers in a perceptually relevant manner. While the results presented in this chapter are promising, future research involving a larger quality rating dataset of amplified Parkinsonian speech is warranted for assessing the robustness and generalizability of objective measures investigated in this research. A larger dataset also facilitates better training and optimization of the deep learning models, leading to better speech quality prediction performance.

4.5 Summary

In this chapter, the quality of the amplified Parkinsonian speech was assessed using subjective and objective methodology. Subjective quality scores were obtained for speech samples

collected from Parkinsonian subjects under different environmental conditions, which were subsequently used to benchmark the performance of the objective metrics. The feature extraction methods included CPP, SRMR, ModA, MFCC, GFCC, and LCQA. Some of the aforementioned methods require a mapping procedure to assimilate a group of features into a predicted quality score. Linear regression, support vector regression, Gaussian process regression, and deep learning are all regression techniques that are deployed to map the extracted features to the subjective measures. The integration of some selected features from GFCC, CPP, LCQA, SRMR, and ModA methods led to the development of a metric that had the highest correlation of 0.85 with the subjective data.

Chapter 5

Design and Evaluation of A New Speech-to-Noise Feedback Device

5.1 Introduction

As discussed earlier, individuals with Parkinson's Disease (PD) suffer from hypophonia, leading to a poor understanding of their speech in noisy environments. To provide better context for this chapter, relevant research studies are reviewed here which systematically investigated the relationship between the production level and intelligibility of Parkinsonian speech. For example, Adams *et al.* [10] studied the effect of background noise level variation on speech production intensity in participants with hypophonia due to Parkinson's disease. The study included 10 Parkinsonian subjects and 10 normal control subjects, and involved three conditions: a conversation task in a multi-talker background noise at 50, 55, 60, 65, and 70 dB SPL; a level matching task where the participants were asked to imitate three speech intensity targets of 60, 70, and 80 dB SPL; and a speech production task where the participants were asked to produce their maximum intensity. While both PD and control subjects increased their speech intensity when the noise level increased indicating a positive Lombard effect, the speech intensity increase for the PD participants was significantly lower than the increase in the normal subjects. Similarly, the PD participants were lower by 3 – 4 dB in matching the speech intensity targets, and 6 – 7 dB lower than normal controls in generating the maximum intensity.

In another study conducted by Adams *et al.* [11], the effect of background noise on both the Signal-to-Noise Ratio (SNR) and the speech intelligibility was investigated. This study included 25 PD subjects and 15 normal control participants. Three levels of noise were included in the experiment (60, 65, and 70 dB SPL), and the noise was generated

through a single loudspeaker at a distance of 115 cm from the PD subject and an angle of 45 degrees. Speech recordings were collected from 2 microphones, one attached to the head of the participant and the other was at located at a distance of 115 cm from the PD subject. The latter microphone served as the "listener", with the subject, the noise source (the loudspeaker), and the "listener" microphone forming an equilateral triangle of a side length of 115 cm. The study showed that in addition to the SNR reduction, there was a reduction of approximately 20 - 30 % in the intelligibility of PD participants at all speech tasks. The intelligibility dropped to below 50 % when the SNR level fell below 1.8 dB.

In a follow-up paper, Adams *et al.* [64], employed two microphones to measure speech SNR from PD and normal control participants in background noise levels ranging between 50 – 70 dB SPL. The first microphone was attached to the subject's head while the other microphone was attached to the subject's throat. Similar to studies reviewed above, results from this study showed that the PD participants were consistently and significantly lower than normal controls in their speech intensity levels across all background noise levels, leading to lower speech SNRs. In addition, this study showed that the head microphone was more sensitive to the effects of the hypophonia on the SNR variations than the throat microphone. However, the throat microphone was found to be useful as it offered excellent noise isolation and good speech detection. This study, therefore, suggested the use of both a throat microphone and a head microphone measures speech SNR at high noise levels.

Finally, Dykstra *et al.* [65], showed that there is a big drop in the intelligibility of PD subjects when exposed to high noise levels. For example, average conversational intelligibility of PD speech was about 57% in the presence of 70 dB SPL noise level, while it was about 89% in the presence of no noise. It is worthwhile to highlight here that the conversational intelligibility of speech produced by normal controls in the 70 dB SPL background noise condition was reported as 85%. All these studies showed that PD speech suffers from reduced SNR in high background noise levels, which negatively affects its intelligibility. It is therefore imperative to enhance PD speech, either during production or soon after, to improve its intelligibility by listeners.

In the previous chapter, the utility of voice amplifiers in this context has been explored. Results from the previous chapter demonstrated that the perceived quality of Parkinsonian speech can be improved with a voice amplifier, in both quiet and noisy environments. However, as shown in the previous chapter, this improvement is dependent on the voice amplifier characteristics, necessitating electroacoustic verification of the amplifier performance by the patient or the clinician. An alternative approach is to enable PD patients to consciously increase their voice level in challenging environments through appropriate feedback. The design and evaluation of such a feedback system form the focus of this chapter. In particular, a

new wearable device is developed to measure the SNR level between speech and background noise. The wearable device aims to give feedback to the subject to increase their voice level when the SNR falls below a certain level. By this way, the SNR of the participants is kept above the level that maintains the speech to be intelligible to the listener. Since SNR estimation is central to the wearable device, the issues associated with realtime speech SNR estimation are discussed next.

5.2 Speech SNR estimation

The discrete-time noisy mixture, $x(n)$, recorded through a measurement microphone is given by;

$$x\left(n\right) = s\left(n\right) + v\left(n\right) \tag{5.1}$$

where $s\left(n\right)$ is the speech signal, and $v\left(n\right)$ is the background noise uncorrelated with the speech signal. For the rest of this chapter, the noise is assumed to be additive and environmental reverberation effects are not included. In realtime SNR estimation, the noisy recording is analyzed on a frame-by-frame basis, and in order to estimate the frame-wise speech SNR, there must be a method to detect voice activity from talkers through a voice activity detector (VAD). The speech SNR for the j^{th} frame can then be estimated as:

if VAD_j

$$SNR_{dB,j} = 10 * \log_{10}\left(\frac{\sigma_{x,j}^2 - \hat{\sigma}_{n,j-1}^2}{\hat{\sigma}_{n,j-1}^2}\right) \tag{5.2}$$

$$\hat{\sigma}_{n,j}^2 = \alpha\hat{\sigma}_{n,j-1}^2 \tag{5.3}$$

else

$$\hat{\sigma}_{n,j}^2 = \beta\hat{\sigma}_{n,j-1}^2 + \left(1-\beta\right)\sigma_{x,j}^2 \tag{5.4}$$

where $\sigma_{x,j}^2$ is the variance of the noisy mixture in the j^{th} frame, $\hat{\sigma}_{n,j}^2$ is the estimate of the noise variance in the j^{th} frame, and α and β are weighting constants respectively.

It is evident from the above that the VAD plays a crucial part in speech SNR estimation. Several approaches have been undertaken to accurately detect speech activity based solely on the noisy speech recordings (e.g., through statistical analysis of recorded data [66, 67] or a decision-directed parameter estimation method with a likelihood ratio test [68]). In general, the performance of these VAD algorithms is dependent on the noise type and true SNR. Figure 5.1 highlights this issue for a sample VAD [68]. Figure 5.1a shows the clean speech

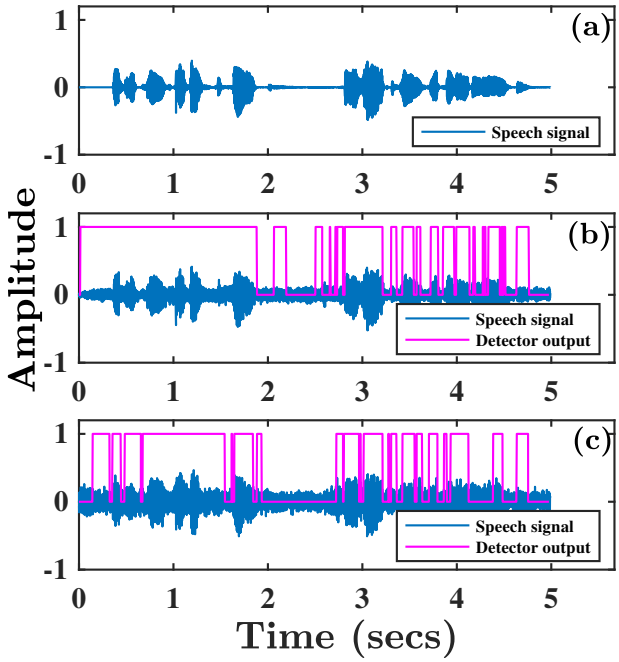


Figure 5.1: Statistical VAD. (a) clear speech signal, (b) the speech signal at SNR= 5 dB (c) the speech signal at SNR = 0 dB.

sentence, while Figures 5.1b and 5.1c depict the same sentence corrupted by an additive multi-talker babble noise at 5 dB and 0 dB SNRs respectively. The magenta lines in Figures 5.1b and c indicate the binary decisions rendered by the VAD algorithm [68], with a value of "1" representing speech presence and "0" indicating speech absence. The high variability in the VAD performance is evident in this figure, across both SNRs. While some of this variability can be mitigated through post-processing and careful tuning of the detection algorithm parameters, its performance cannot be generalized across different noise types (in other words, a different set of fine-tuned parameters may be required for a different noise source). Recent efforts on incorporating DNNs to classify the type of background noise and apply appropriately tuned parameters in the VAD are interesting [69], but the computational complexity of these approaches may preclude their implementation in wearable devices. This thesis takes an alternate approach, wherein a dedicated accelerometer placed on the subject's throat area reliably detects speech activity, irrespective of the type of background noise or its level. The block diagram of the proposed device and its functionality are described next.

5.3 Proposed Speech-to-Noise Feedback (SNF) device

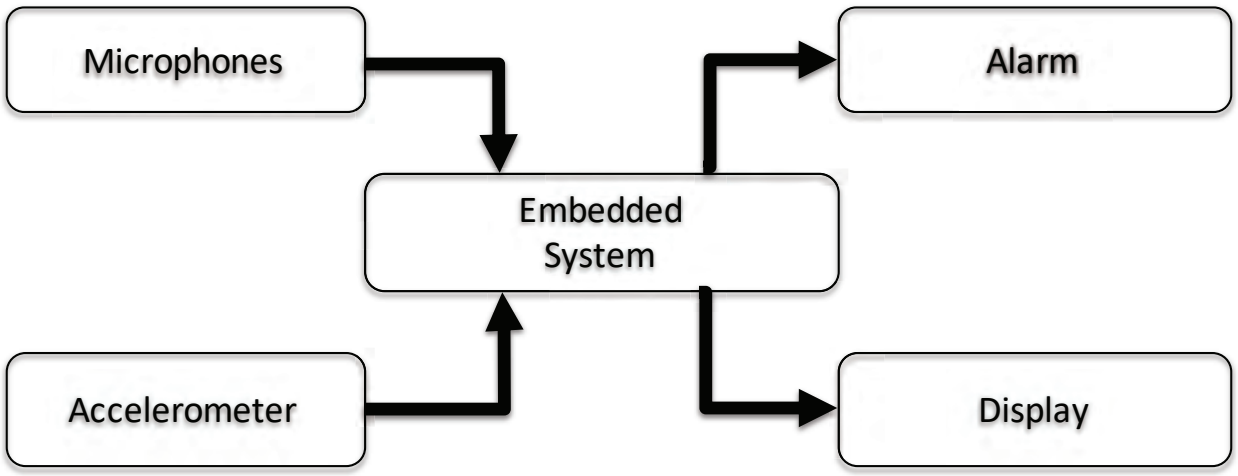


Figure 5.2: System block diagram

Figure 5.2 displays the block diagram of the proposed SNF device, which is built using off-the-shelf components. Key features of this device include:

- Dual ear-level microphones – Two microphones, one above the left ear and the other above the right ear, serve as the measurement microphones for the proposed system. The dual-channel or binaural recording design is an innovative feature of this system, as it allows for better SNR estimation in environments with asymmetric noise sources. For example, a background noise source spatially located on the right side of the wearer will generate better (higher) SNR on the left side and vice versa, due to the head shadow effect. Integrating SNRs estimated at left and right ears will offer a more robust measure of the environmental SNR, as shown in the experimental results later in the Chapter.

- VAD accelerometer – a dedicated accelerometer forms the third data acquisition channel, conveying information on wearer's speech activity. As discussed earlier, previous studies have used the Knowles BU-27135 for logging the voice information (e.g., ambulatory phonation monitor [20]). However, the Knowles BU-27135 outputs an analog signal, which necessitates an analog-to-digital converter (ADC) interface, prior to accelerometer data analyses. An alternate model, MPU-6050, is used in this design as it outputs digital data and is more cost-effective than the Knowles BU-27135.

- Low cost embedded system – The Raspberry Pi 3 b+ is used as the computing unit in this device. It has a 64-bit quad core processor running at 1.4 GHz. It has an SDRAM

of 1 GB, and 40 pin-GPIO header, which are adequate for the current device.

- Display and alarming system – The system features a graphical user interface (GUI) that allows for initial setup of the SNF software. The GUI also facilitates the visualization of the realtime SNR. The alarming system alerts the user when the estimated SNR falls below a threshold for a pre-determined period of time, both visually (by color-coding the estimated SNRs red) and auditorily (by playing a beep).

5.3.1 Initial hardware prototype

During the initial development of the SNF device, a pair of Polsen OLM-10 omnidirectional microphones were employed as the recording microphones. Aquisition of the analog signals generated by these electret microphones required a sound card to be installed on the top of the Raspberry Pi micro-controller. The Audio Injector sound card was deployed for this purpose. Fig. 5.3 displays the initial version of the prototype, with the Audio Injector card installed on top of Raspberry PI. The stereo microphone inputs are connected at the top RCA ports, while the stereo audible alarm connection was at the right. The MPU 6050 accelerometer was interfaced to the Raspberry Pi through the I2C port.

Pilot testing with the initial prototype revealed low sensitivity by the OLM-10 microphones, especially when worn at ear-level. The low sensitivity precluded accurate speech SNR estimate at high background noise levels. The OLM-10 microphones were therefore replaced by the Electret Microphone Amplifier MAX 4466s, which were more compact and had higher sensitivity. While the MAX 4466s solved the problem of poor measurements of SNR, additional problems arose when converting from a monaural system to a binaural system. The binaural system added a processing load to the sound card, causing problems in the latency and the synchronization of the processing system. In particular, the data collected from the two microphones and the accelerometer were not synchronized (as the system incurred latency through the acquisition of dual microphone data). This lead to the erroneous computation of the SNR and subsequent malfunctioning of the audible and visual alarming system. As such, there was a need to remove the sound card, which required the use of digital microphones that interfaced directly to the Raspberry Pi. The Adafruit I2S MEMS microphones have an onboard digitizer whose output can be connected directly to the Raspberry Pi system without the need of a sound card. This solved the problem of latency errors.

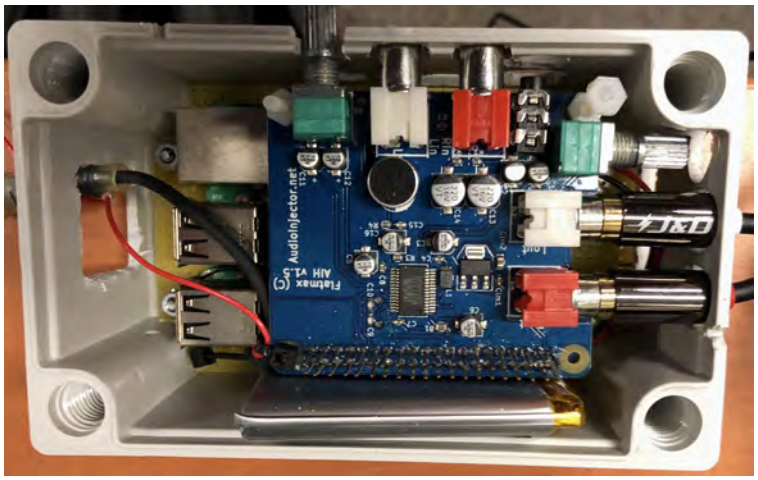


Figure 5.3: Initial prototype with sound card

5.3.2 Refined prototype

Fig. 5.4 shows the components and the finished version of the refined prototype. The Adafruit SPH0645 is the compact, low cost MEMS microphone with a bandwidth of 50 Hz – 15 kHz that interfaces to the Raspberry Pi through the I2S interface. These microphones were encased in a plastic case and placed above the left and right ear of the participants. The MPU 6050 accelerometer is connected to the Raspberry Pi through the I2C interface. It was held securely against the throat of the wearer using Velcro. Removal of the sound card ensured the processing portion of the device was compact, even with the addition of a 7" LCD touchscreen display. The processing system is placed in a belt pack and can be worn around the waist, as shown in Fig. 5.4.

Fig. 5.4 shows images taken for the SNF device before assembly, after assembly, and after attaching the device to the participant.

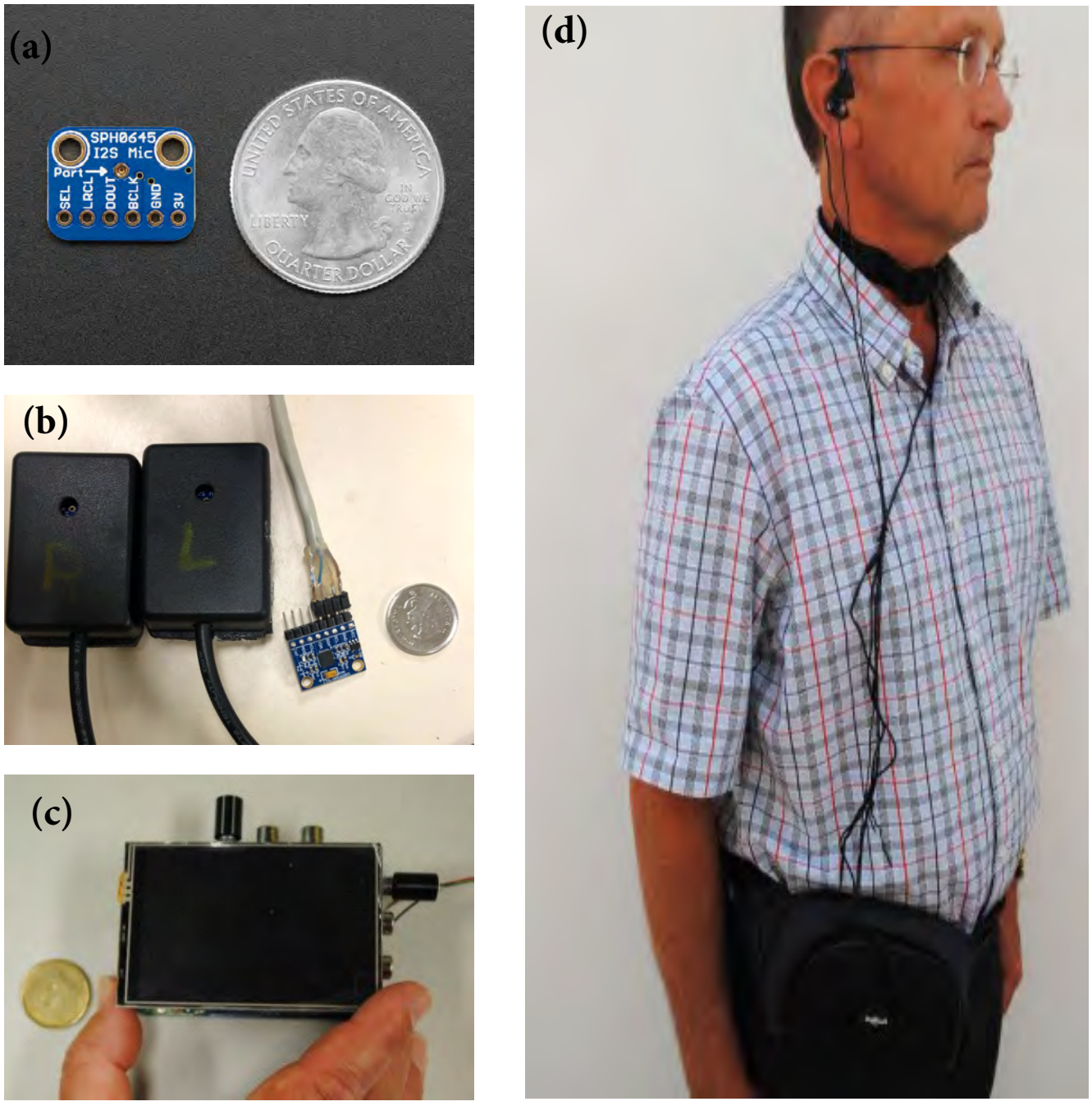


Figure 5.4: SNF device. Rotating counter-clockwise from the upper left corner (a) The MEMS microphone, (b) The 2 microphones in plastic caskets and the MPU 6050 accelerometer, (c) the device including the Raspberry Pi , and (d) The device assembled and attached to the user

5.3.3 Software design

Custom software routines were written in Python 3.6 for data collection, analysis, display, and alert. The software ran two threads: the main thread that acquired stereo data from

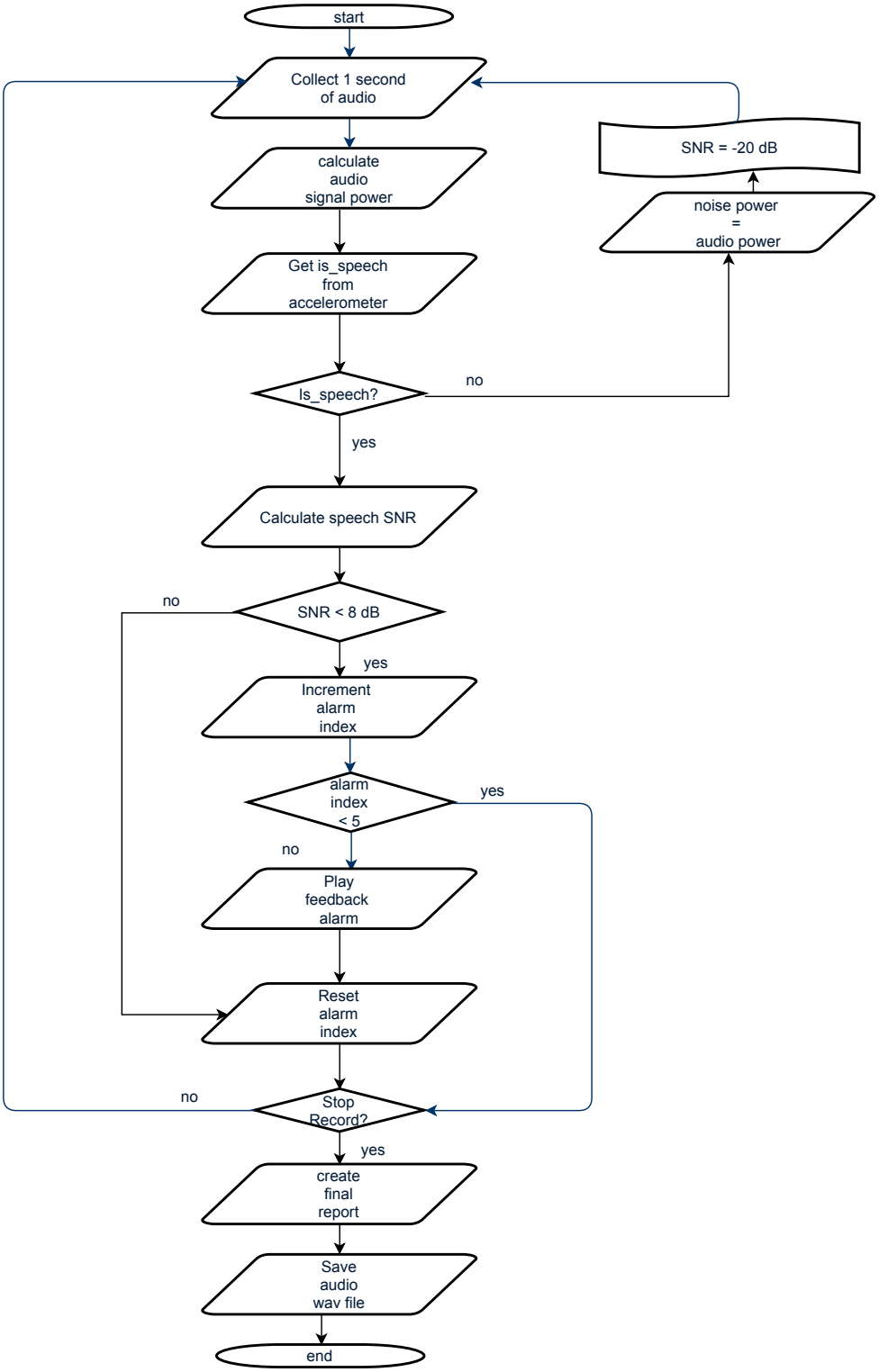


Figure 5.5: Flow chart of the device software

the MEMS microphones, conducted all the SNR estimation calculations, and performed the realtime display and alerting, while the second thread continuously acquired the data from the accelerometer. Thread synchronization was accomplished through proper locking and unlocking mechanisms. Fig. 5.5 shows the flow chart of the script running the device. The device alerts the wearer when the measured SNR is below 8 dB for 5 seconds. The steps of the algorithm are as follows:

- An audio frame of 1 second is collected at 44100 Hz sampling rate.
- The power (i.e. variance) of the signals recorded from the left and right microphones is calculated.
- The data coming from the accelerometer is analyzed to indicate larynx movement.
- If the accelerometer data shows no speech from the subject, the calculated power represents the noise power for the next audio frame. The SNR was set to -20 dB, and the next audio frame is collected.
- If the accelerometer data was above an empirically determined threshold to indicate the presence of speech, the SNR was calculated as follows:

$$SNR_{dB,j} = 10 * \log_{10}\left(\frac{\sigma^2_{x,j} - \hat{\sigma}^2_{n,j-1}}{\hat{\sigma}^2_{n,j-1}}\right) \tag{5.5}$$

where $\sigma^2_{x,j}$ is the variance of the noisy mixture in the j^{th} frame, $\hat{\sigma}^2_{n,j-1}$ is the estimate of the noise variance in the j^{th} frame

- If both the left or right channel SNR(i) > 8 dB, the current frame speech SNR is higher than the desired SNR level. Reset the alarm index and check if this is the last audio frame.
- If either the left or right channel SNR(i) < 8 dB, the current frame speech SNR is lower than the desired SNR level. Increment the alarm index by 1. This ensures that noise coming from one side of the wearer is captured properly by the ipsilateral microphone.
- If the alarm index is less than 5 seconds, move to record the next audio frame.
- If the alarm index is higher than 5 seconds, play the feedback alarm, reset the alarm index, and move to record the next audio frame.
- If this is the last audio frame, stop recording, generate a report containing the recorded SNR values, and save the audio wav file.

Fig. 5.6 displays a sample screenshot of the device output. When the measured SNR is below the desired threshold, the measured SNR bar appears in red, but when the measured SNR is higher than the desired threshold, the measured SNR bar appears in green. An alarm is played after 5 consecutive red bars. When there is no speech detected, the measured SNR bar is yellow and placed at a value of -20 dB. It must be noted here that the program parameters including the SNR threshold, voicing threshold for the accelerometer signal, the frame size for SNR calculations, and the number of consecutive below-threshold SNRs to activate the alarm, are all configurable through the software.

The designed prototype was tested as a proof-of-concept with three control subjects. The experimental methodology for device performance validation is described in the next section.

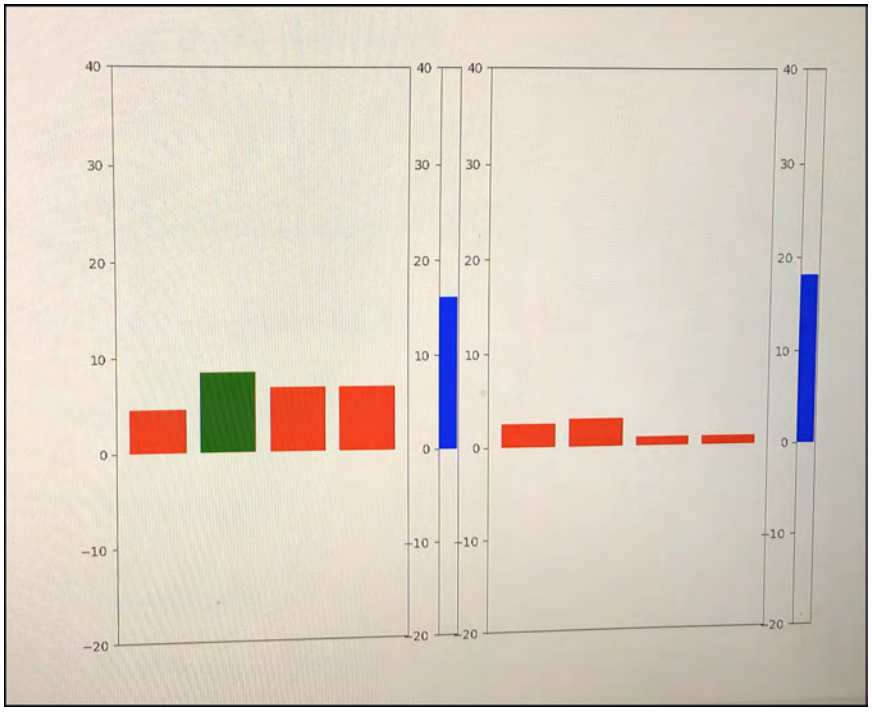


Figure 5.6: Screenshot of the device output

5.4 Methodology

Subjective data collection procedures outlined in this chapter received ethics approval from Western University's Health Sciences Research Ethics Board.

This study included collecting speech samples from 3 normal individuals in different noise conditions, with and without the assistance of the new SNF device. Each participant was seated in a lab environment and completed 4 different speech tasks. The first task was to produce three repetitions of the sustained vowel /a/, the second task was to repeat six

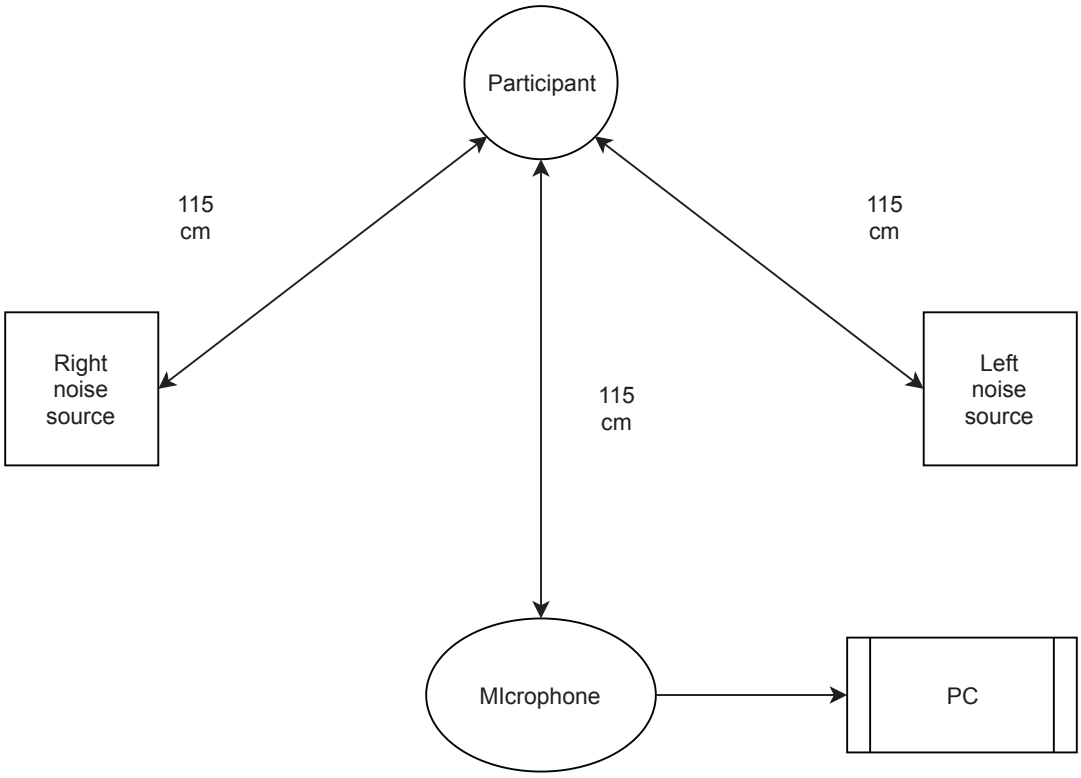


Figure 5.7: Methodology block diagram

selected sentences, the third task was to read the phonetically balanced rainbow passage [70], and the fourth task was to engage in a normal conversation with the experimenter about the weather, hobbies, and daily life. These tasks were repeated at different background noise levels, no noise or quiet environment (although there is ambient room noise that was measured to be 55 dB SPL) and multi-talker babble noise at 65 dB SPL and 75 dB SPL. The multi-talker babble noise was generated in from two loudspeakers that were positioned 115 cm on either side of the participant, as shown in Fig. 5.7. A measurement microphone was positioned at a distance of 115 cm in front of the participant, and served as the "listener" (see Fig. 5.7). The loudspeaker and the microphone levels were calibrated before each session. The multi-talker babble noise at both 65 and 75 dB SPL was presented in three separate conditions: noise from the right loudspeaker only, noise from the left loudspeaker only, and noise from both loudspeakers. These experimental conditions were realized to compare the device operation in monaural (i.e. use of only one SNF device microphone) and binaural (use of both SNF device microphones) modes. The whole experiment was conducted twice, once with the SNF device active, and the other with the SNF device inactive. These two conditions were counterbalanced across the participants. For each participant, this procedure led to the collection of 2 device conditions (active/inactive) $\times$ 2 noise levels (65 and 75 dB

SPL) × 3 noise modes (left, right, or both) × 11 stimuli (3 sustained vowels, 6 repeated sentences, rainbow passage, conversation) +2 device conditions × 11 stimuli in the "quiet" mode, for a total of 154 recordings per participant.

The recorded data was analyzed to assess the effect of using the device on the performance of the participants' speech in terms of SNR, intelligibility, and quality. In particular, the following procedure was followed for analyzing the recorded data at the "listener" location:

- The combined objective measure developed in Chapter 3 was applied to the sustained vowel recordings. This allowed for quantifying the impact of SNF device on predicted PD vowel quality in different noisy environments.

- Two sentences were selected out of the six repeated sentence recordings. For rainbow passage and the conversation recordings, 4 seconds around the midpoint were selected. The speech SNR was estimated for these recordings and compared between device active and inactive situations. The assessment of the SNR of the records are done by averaging the instantaneous SNR of each record.

- A panel of three listeners was recruited to subjectively evaluate the intelligibility of selected speech material. This subjective evaluation allowed for investigating the SNF device impact on perceived intelligibility, and also to gauge the relation between the estimated speech SNR and perceived intelligibility.

- The selected speech material was separately assessed using the composite objective measure developed in Chapter 4. This enabled the measurement of the SNF device influence on predicted speech quality in different background noise conditions.

5.5 Results

Fig. 5.8 displays the sample output data from the SNF device for a participant performing the speech tasks. The output waveform recorded at one of the headset microphones is shown in Fig. 5.8.a, while Fig. 5.8.b shows the output waveform recorded from the accelerometer that is attached to the subject's larynx. It is noted that the background noise has no effect on the accelerometer wave, which means that this accelerometer wave can be used to detect the occurrence of speech or not whatever is the background noise. Fig. 5.8.c shows the measured instantaneous SNR that is displayed on the device's screen. It must be noted here that the speech SNR is computed in realtime, and on the SNF device screen, the estimated speech SNR is displayed in a scrolling format.

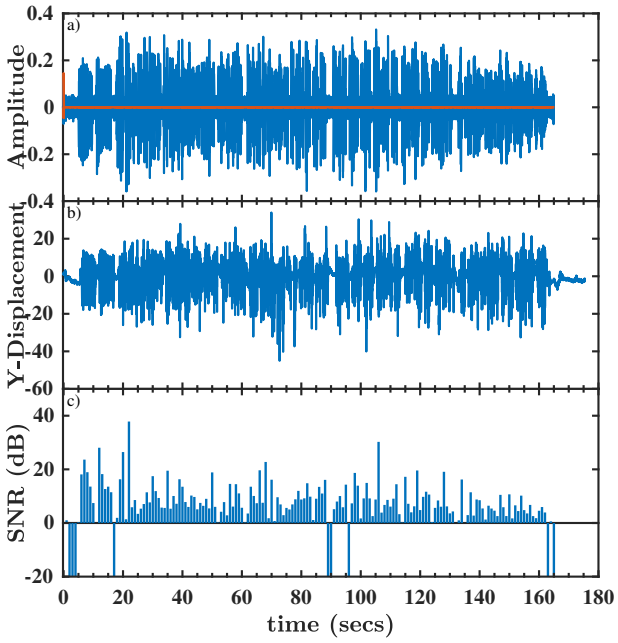
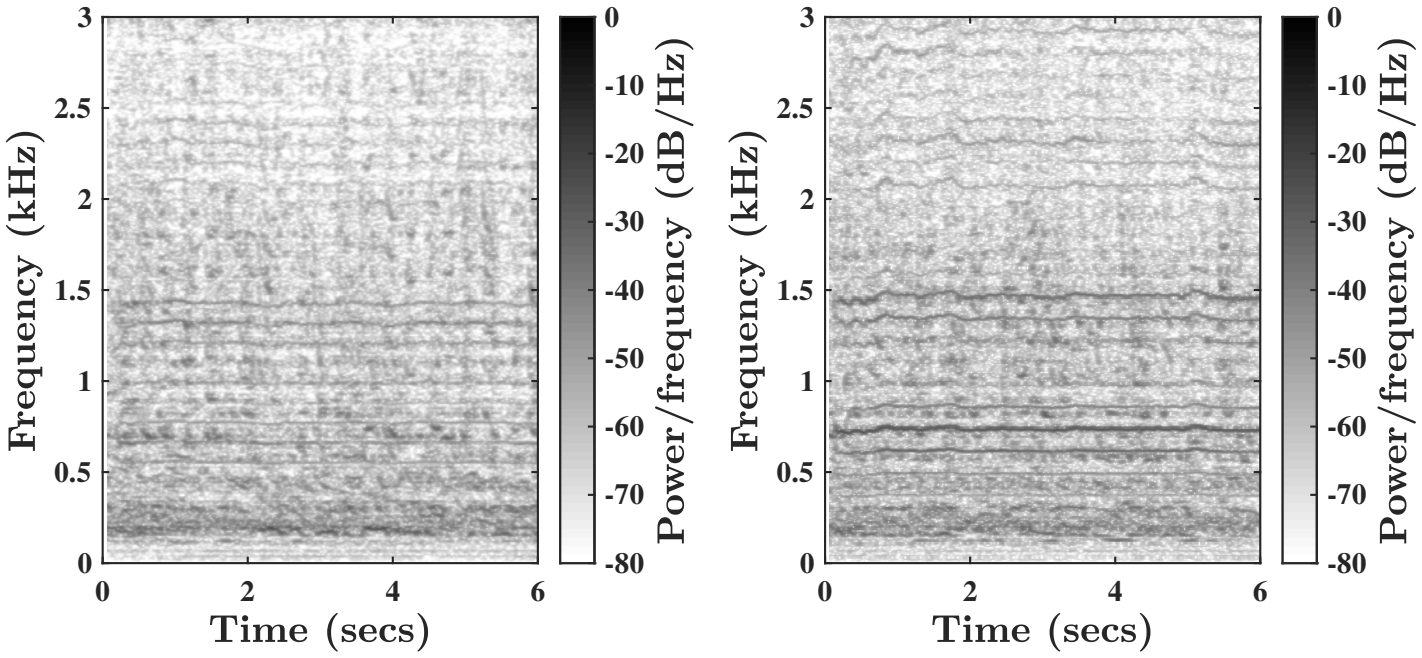


Figure 5.8: The output data of the device (a) The output waveform from the device microphones (b) The output signal from the accelerometer (c) The instantaneous measure SNR

5.5.1 Analysis of sustained vowel recordings

In the experiments conducted in this study, each participant produced the sustained vowel /a/ three times at the different background noise conditions mentioned in the methodology section. Fig. 5.9 shows the spectrograms of two sustained vowel /*a*/ recordings collected from the same participant in the presence of 75 dB SPL multi-talker babble background noise, with and without the SNF device active. Fig. 5.9a shows the spectrogram of the sustained vowel recording when the SNF device was inactive. It is evident in this figure that the vowel fundamental frequency and its harmonics are blurred by the background noise, and which influences its perceived quality (later analyses showed that this recording had one of the lowest predicted quality scores). In contrast, Fig. 5.9b shows the spectrogram of the sustained vowel recording from the same participant in the same background noise environment, but with the audible feedback from the SNF device. The relative enhancement of the harmonic content due to an increase in the vocal intensity can be seen in this spectrogram. Subsequent analyses showed that this record that had a higher predicted vowel quality, suggesting that the feedback alert from the SNF device resulted in tangible improvements in perceived quality of the sustained vowel in noisy environments.



(a) Before using SNF device

(b) After using SNF device

Figure 5.9: Spectrogram of 2 sustained vowels for a participant before and after using the SNF device

As the next step, the combined objective metric developed and described in Chapter 3, was used to predict the quality of the all sustained vowel recordings before and after the use of the SNF device. The relevant features were extracted from each of the sustained vowel recording, and the trained machine learning model from Chapter 3 was applied to assimilate the features to a predicted quality score. Fig. 5.10 shows the averaged predicted quality scores across the same background noise conditions and participants, before and after the use of the SNF device, with the error bars in Figure. 5.10 denoting the standard error of the average. It can be seen that the feedback through the SNF device has enhanced the predicted quality of the sustained vowels across all noise conditions. The biggest quality enhancement was observed at 75 dB SPL noise condition, highlighting the merit of the SNF device in more challenging noisy environments.

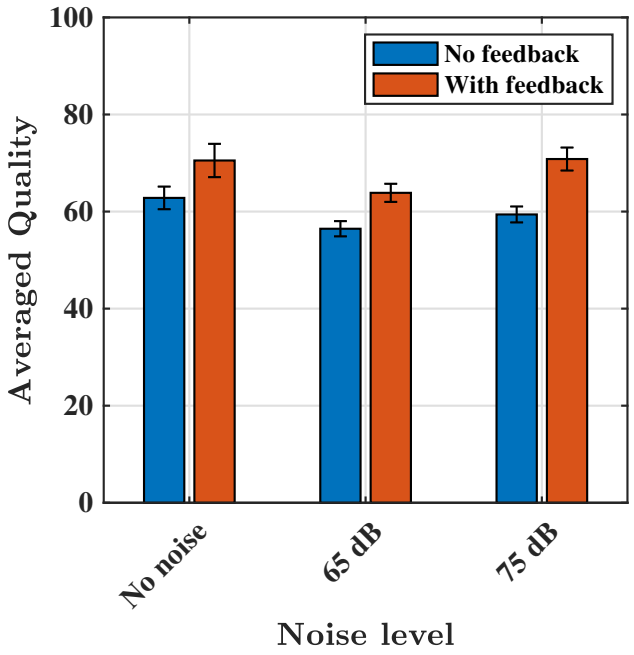
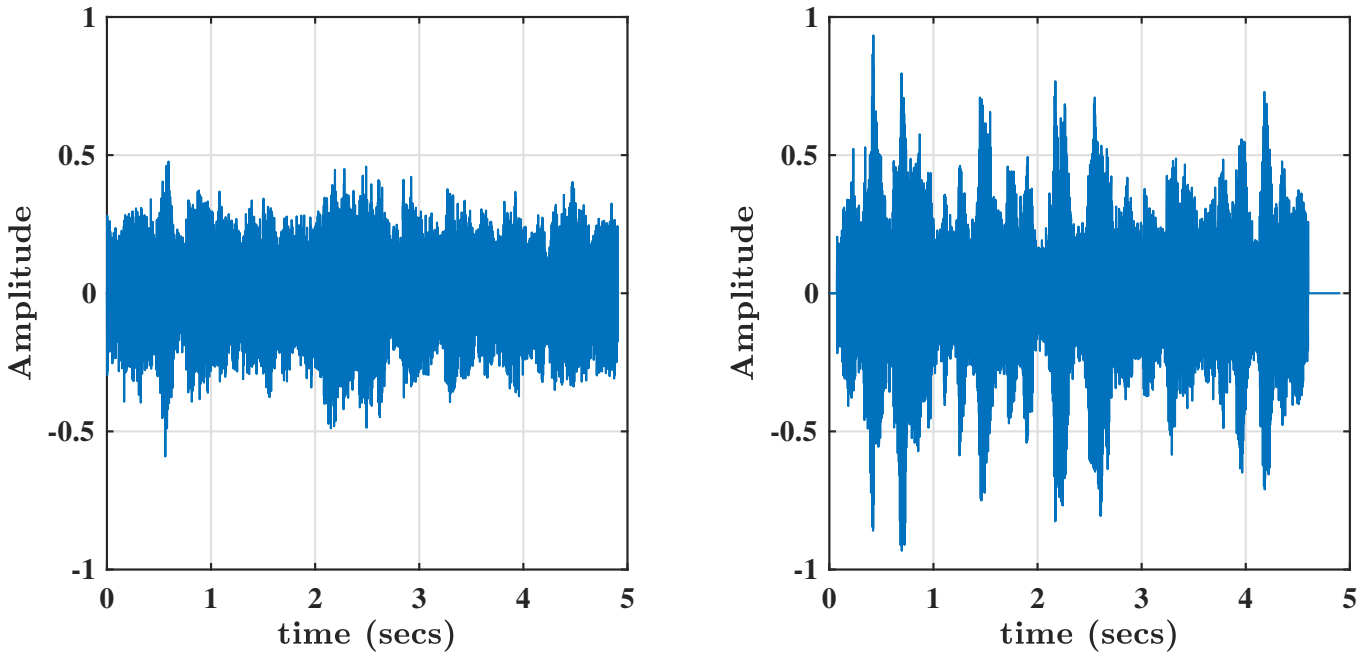


Figure 5.10: Quality of sustained vowels

5.5.2 Assessment of speech SNR

For the continuous speech stimuli, the changes in estimated speech SNR with device activation was investigated first. Fig. 5.11 shows the waveforms of the same speech utterance from a participant in 75 dB SPL background noise recorded at the listener location, before and after using the SNF device. The waveforms are associated with a selected sentence in the rainbow passage viz. "The rainbow is the division of white light into many beautiful colors". Fig. 5.11a shows the produced speech waveform when no feedback was provided. It is evident in this figure that a substantial portion of speech information is masked by the background noise, which will impact its intelligibility. The maximum speech SNR value estimated for this waveform was about -4.6 dB. On the other hand, activating the SNF led to an increase of the speech production level, which consequently raised the SNR level. This is evident in Figure 5.11b, where speech components and envelope are distinguishable above the background noise. The estimated SNR value of the recorded waveform after using the SNF device was about 1.9 dB.

Fig. 5.12 displays the estimated speech SNR values before and after using the device for repeated sentences, rainbow passage, and normal conversation speech stimuli, in the absence of loudspeaker noise (room noise level is 55 dB SPL), and at 65 dB SPL, and 75 dB SPL multi-talker babble loudspeaker noise. For each record, the maximum instantaneous SNR

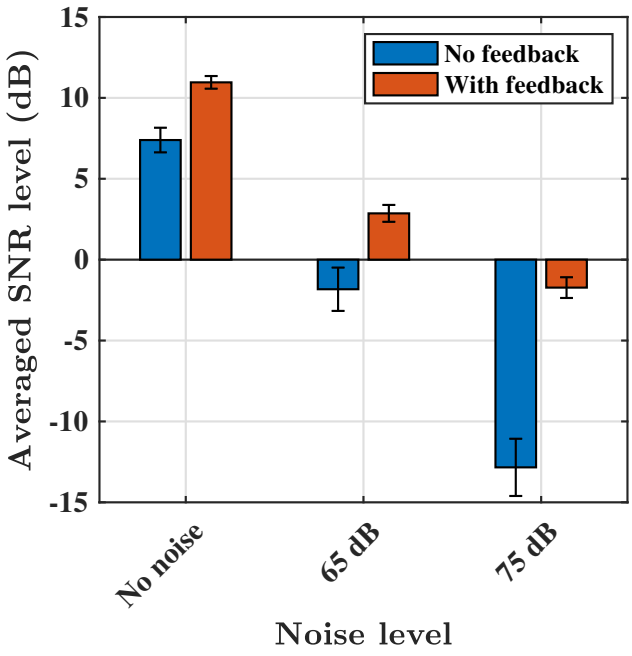


(a) Before using SNF device

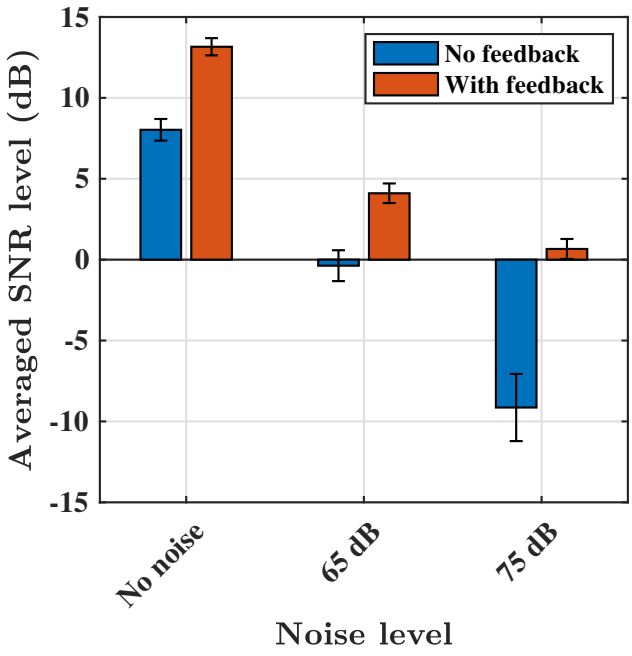
(b) After using the SNF device

Figure 5.11: The waveforms of a selected sentence from the rainbow passage at 75 dB SPL

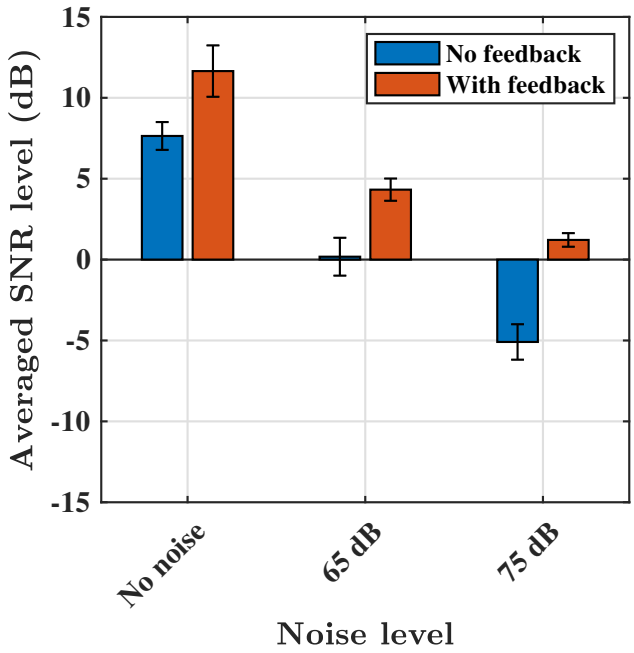
was selected to represent the speech SNR, since the background noise is the non-stationary multi-talker babble. It is noted that the use of the SNF device led to a big increase in the SNR for all the speech tasks at different noise levels. It is also noted that the SNR at 75 dB SPL was very low without using the device. The device clearly enhanced the SNR level when used by participants during conversation and reading rainbow passage tasks. Although the SNR increased in the case of the repeated sentences task, it was still below zero. The reason for that is that the device alerts the participant after 5 consecutive seconds of speech SNR below the preset threshold. The repeated sentence are approximately 2 seconds in duration, which might not give the device enough time to alert the subject.



(a) Sentences



(b) Rainbow passage

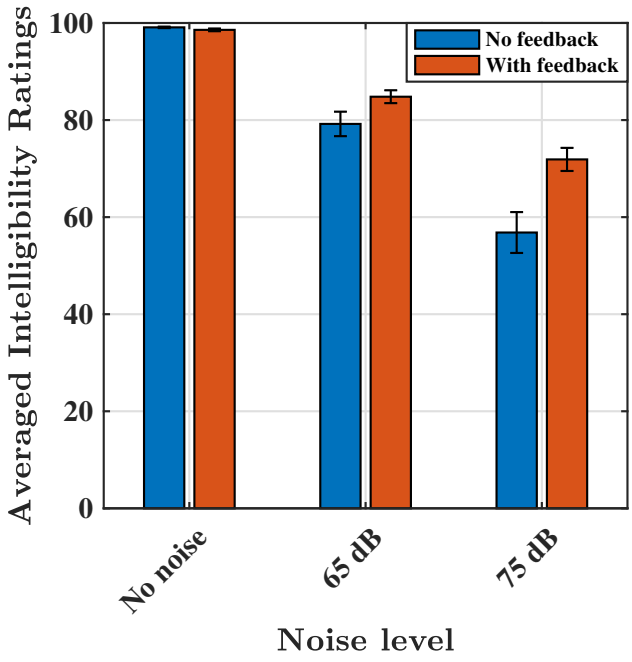


(c) Conversation

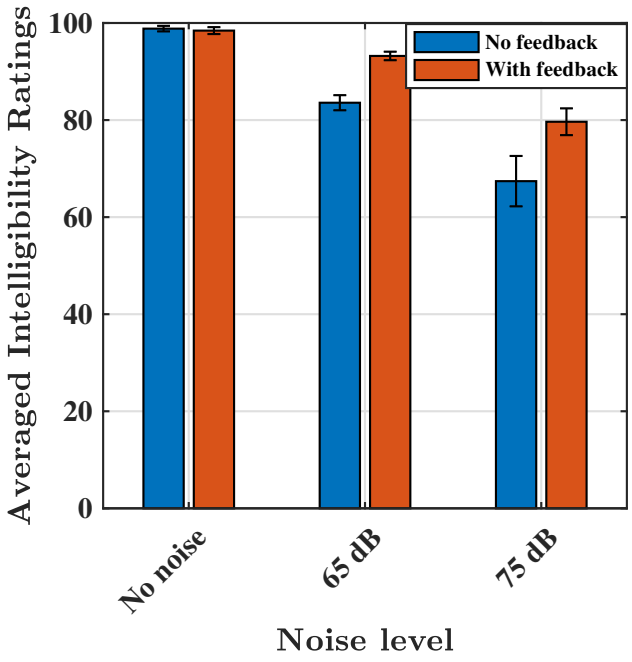
Figure 5.12: Assessment of SNR before and after using the SNF device

5.5.3 Assessment of speech intelligibility

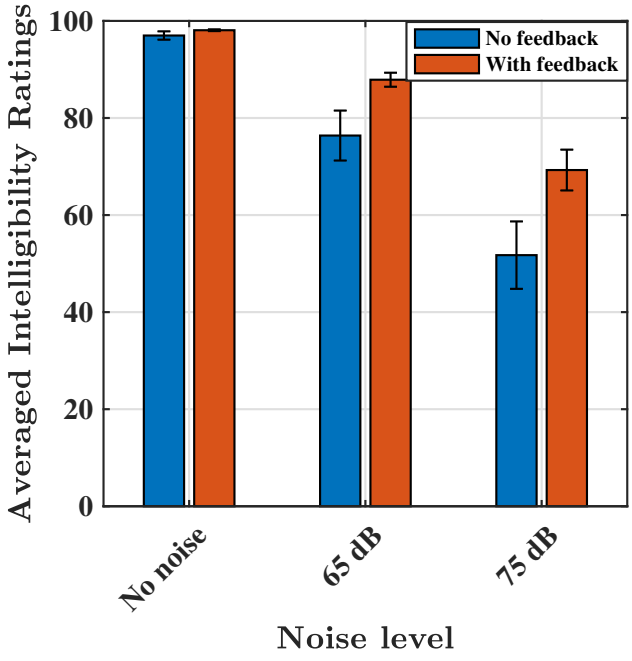
As described in the methodology section, the selected speech recordings were evaluated by three listeners for their intelligibility. The listeners rated the intelligibility of each sample on a visual analog scale ranging between 0 – 100, with 0 representing no intelligibility and 100 representing perfect intelligibility. The intelligibility ratings were averaged across the three listeners for each condition. Fig. 5.13 shows the assessment of the averaged intelligibility before and after the use of the SNF device, with the error bars once again denoting the standard error of the average. Regardless of the speech material, it is clear that there is a significant drop of intelligibility with an increase of the background noise level. When there is no noise added, the intelligibility is near 100%, and there is no significant impact of the SNF device on the intelligibility. At higher levels of background noise, there is a drop of intelligibility at all speech tasks; this is where the SNF device has the greatest impact in enhancing the intelligibility of the speech of the participants. Similar to the SNR data, the greatest intelligibility improvements were observed for the 75 dB SPL background noise condition. The averaged intelligibility improvements were 15%, 12%, and 18% for the repeated sentences, rainbow passage, and the conversational speech respectively.



(a) Sentences



(b) Rainbow passage

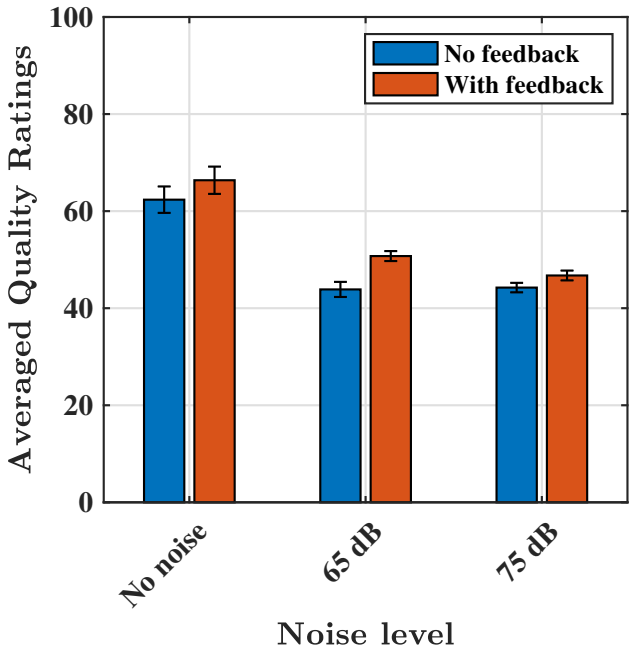


(c) Conversation

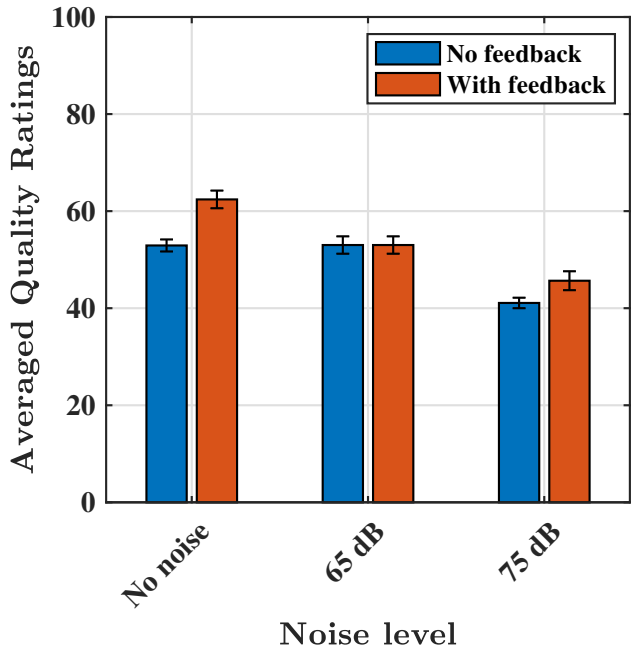
Figure 5.13: Assessment of intelligibility before and after using the SNF device

5.5.4 Assessment of speech quality

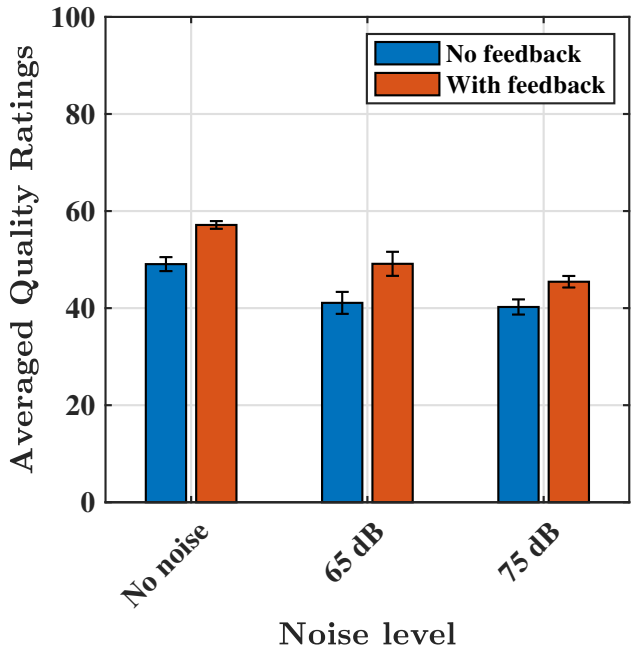
The selected speech recordings were subsequently assessed in terms of their predicted quality. The same speech stimuli from the subjective speech intelligibility experiment were employed in this evaluation as well. Relevant features were extracted from each of the speech recording and the trained Gaussian Process Regression (GPR) machine learning model from Chapter 4 was applied to output the predicted quality score. Fig. 5.14 displays the predicted speech quality scores in different conditions, averaged across the three participants, with the error bars denoting the standard error of the average. In general, the quality results are in line with speech intelligibility and speech SNR results: (a) there was a decrease in predicted quality in the presence of background noise, and (b) there was an increase in the predicted quality with the feedback from the SNF device. The improvements in predicted quality with the activation of the SNF device are not as strong as those seen with intelligibility or speech SNR data. This potentially stems from the perception of speech intelligibility and quality, and can be explained from the example waveform in Figure 5.11 . The underlying speech components are more visible and perceptible in Figure 5.11 b, which will enhance the intelligibility of this recording. However, the continued presence of background noise, would still negatively impact its perceived quality. Thus, a significant improvement in intelligibility can be expected between the two waveforms in Fig. 5.11, but with only a modest increase in the perceived speech quality.



(a) Sentences



(b) Rainbow passage



(c) Conversation

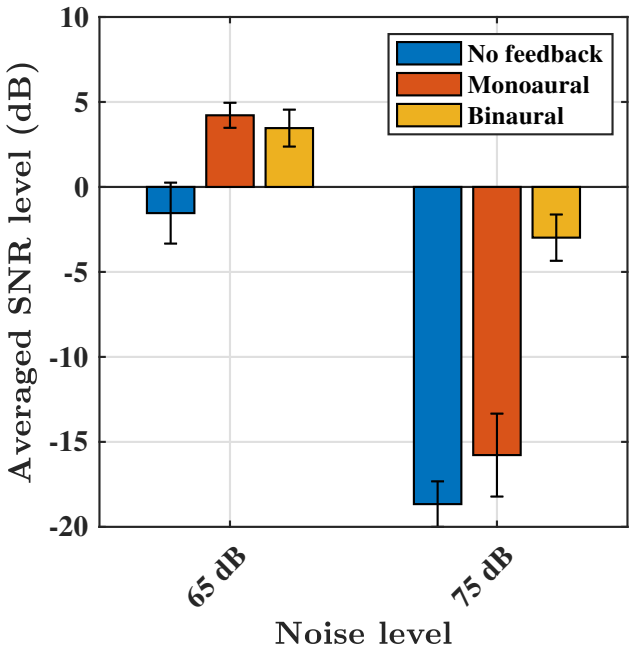
Figure 5.14: Assessment of quality before and after using the SNF device

5.5.5 Assessment of the binaural microphones on SNR and intelligibility

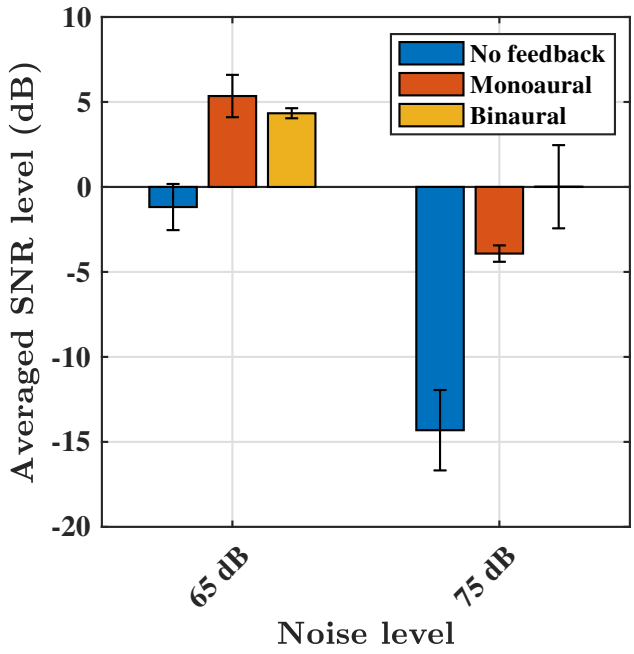
An interesting feature of the new SNF device is its binaural measurement capability, which means that it assesses the SNR level from the right and the left sides of the participant. Comparing the performance of the binaural device to the monaural device showed that at high background noise levels, the binaural device exhibits a better performance than the monaural device in terms of SNR levels and intelligibility. To evaluate the SNR and intelligibility for the monaural case, only one noise source played the multi- talker babble noise at 65 or 75 dB SPL from either the right side or the left side of the participant. Only one microphone was detecting the SNR level and that was on the contralateral side of the noise source.

Fig. 5.15 shows the averaged SNR levels for the speech recordings when there was no feedback device, when the device was operating in monaural mode, and when the device was functioning in the binaural mode. It is noted that the difference between the monaural and the binaural cases is negligible at 65 dB SPL background noise. On the other hand, there is a big improvement in the SNR level in the case of 75 dB SPL for repeated sentences, rainbow passage, and the conversational speech.

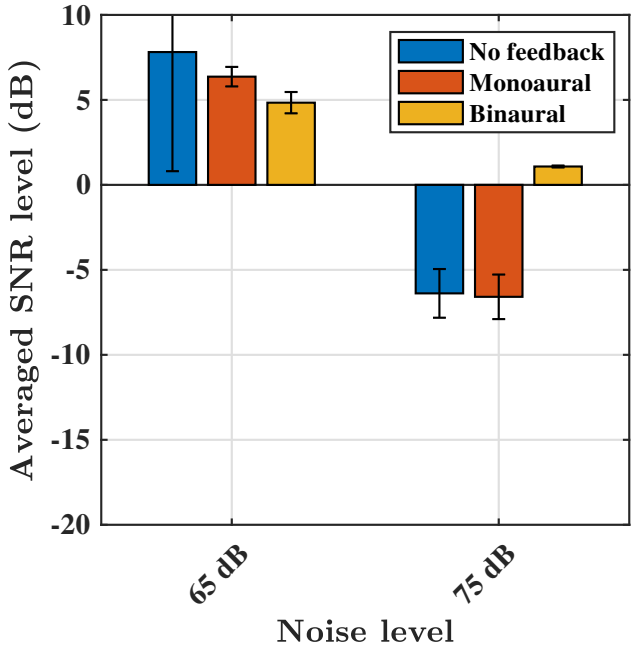
Fig. 5.16 shows the averaged intelligibility of the records in the cases of no feedback, monaural feedback, and binaural feedback. As the case in SNR measurements, the enhancement in intelligibility with the device operating in the binaural mode is clear when the background noise was at 75 dB SPL. Thus, the binaural operation appears to be more beneficial in more challenging background noise conditions.



(a) Sentences

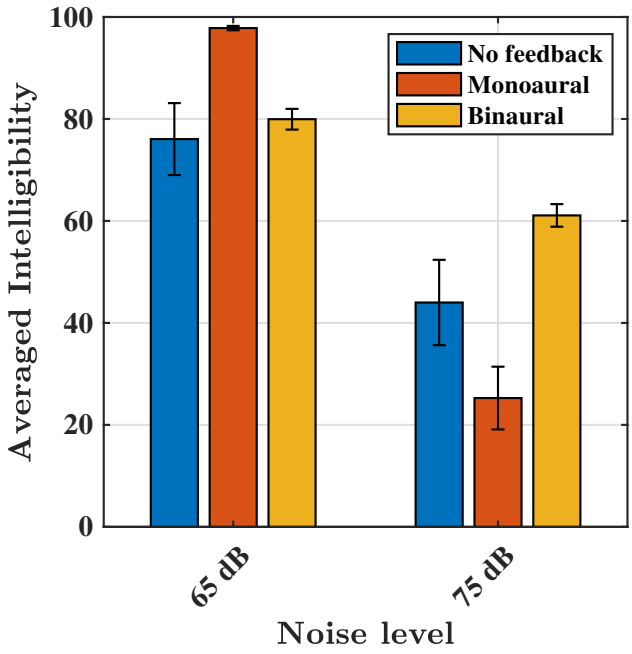


(b) Rainbow passage

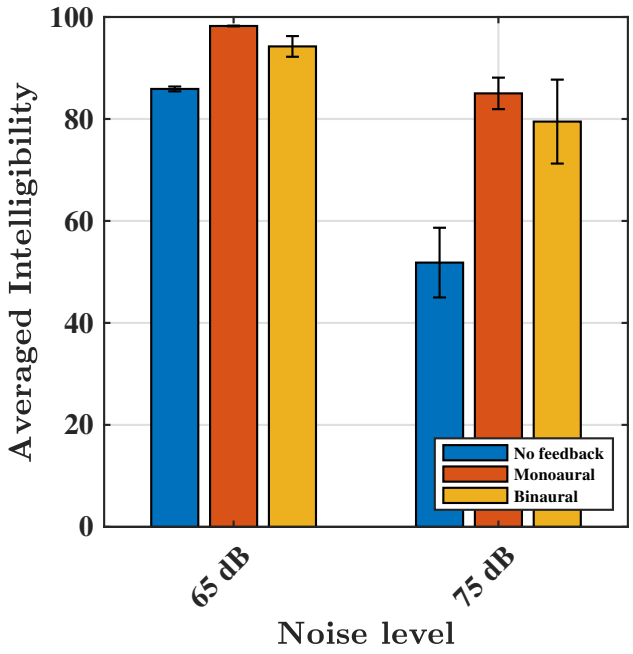


(c) Conversation

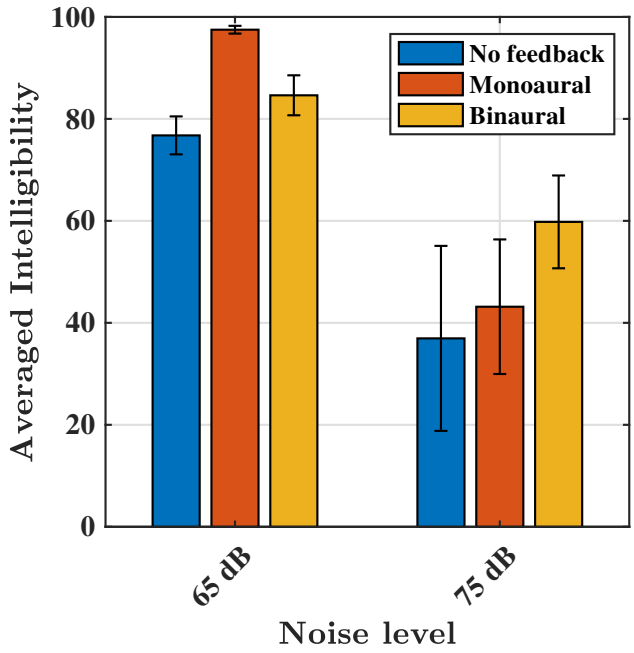
Figure 5.15: Assessment of SNR before using the SNF device, monaural, and binaural cases



(a) Sentences



(b) Rainbow passage



(c) Conversation

Figure 5.16: Assessment of intelligibility before using the SNF device, monaural, and binaural cases

5.6 Discussion and conclusion

SNF device is using the accelerometer to detect the speech even in highly noisy environments because the accelerometer signal is not effected by the noise signal. The use of 2 microphones makes the SNF device binaural, which enhances the SNR feedback, especially in environments with multiple noise sources at high levels. The SNF device was found to enhance the user produced speech in terms of SNR, intelligibility, and quality.

The use of binaural microphones in the SNF device system had proved to be beneficial, especially in the case of high background noise levels. The use of a binaural system leads to higher scores of SNR levels and intelligibility more than the use of monaural microphones.

There is a drawback in the design of the device that it cannot detect a big variation in the value of SNR that is caused by a sudden big change in the value of the noise level. The design of the SNF device depends on recoding the value of the previous instantaneous noise level to predict the value of the current SNR. When the participant keeps talking, the value of the noise level does not change. The problem arises when the noise level changes suddenly while the subject talks. In this case, the device will not be able to record the new noise level until the subject stop talking. Although the case of big variation of noise is not common in daily life, more investigation is needed to overcome this problem.

This device has been tested on 3 normal participants. As the device is proved to increase the SNR of speech for its users, it is expected to be useful to people impaired with PD and other speech disorders and to enhance their speech's SNR, intelligibility, and quality. More experiments are to be conducted on impaired people to see the effect of the device on their speech performance during their daily life.

5.7 Summary

In this chapter, a new device is presented to help people impaired with PD to enhance the intelligibility and quality of their speech. This device uses an accelerometer that is attached to the larynx of the subject to detect the occurrence of speech even in noisy environments. The device is tested on three normal participants who performed the tasks of repeating sentences, reading the rainbow passage, and engaging in normal conversations at 3 different noise levels, 55, 65, and 75 dB SPL. The SNF device was found to enhance the SNR, the intelligibility, and the quality of the speech of its users. Future research must include more participants, and further enhancements to the software and hardware design.

Chapter 6

Conclusions and Future Directions

In this chapter, the contributions of this work are given. Recommendations for future work are also included.

6.1 Contributions

- Five methods were presented for feature extraction from the Parkinsonian sustained vowels. While the first two methods, CPP and HNR, were single feature methods, the other three methods, GFCC, LCQA, and the combined method needed linear regression and SVR machine learning techniques to predict the quality of the sustained vowels objectively. The combined method represents an improved objective Parkinsonian vowel quality estimator that incorporates LCQA, HNR, and CPP together as a set of features. The combined method had a higher correlation value more than the correlation obtained from using the LCQA, HNR, and CPP separately. The correlation value of the combined method reached 77% for the test dataset.

- To reduce the overfitting effect in the machine learning technique, PCA and the feature selection and reduction method were used to reduce the dimensionality of the input features and reduce the overfitting of the obtained objective scores. This led to a reduction of the number of input features and to enhancement of the correlation scores for the test datasets. When applying linear regression to the combined method, the correlation value reached 80% between the subjective and objective scores.

- To assess the quality of running Parkinsonian speech, 7 feature extraction methods were utilized, SRMR, ModA, CPP, GFCC, MFCC, LCQA, and the combined method. The methods SRMR, ModA, and CPP are single feature methods, and they do not need a machine learning technique to map its objective scores to the corresponding

subjective scores. On the other hand, GFCC, MFCC, and LCQA are multi-feature methods, and a machine learning algorithm must be applied to extract the objective scores. This work incorporated the use of deep learning and GPR to assess the quality of Parkinsonian speech in addition to SVR. The combined method feature set included GFCC, CPP, LCQA, SRMR, and ModA features. The use of the combined method with deep learning and GPR achieved a correlation value of 85%.

- The use of multi-talker babble noise led to the appearance of a bias effect at low quality ratings. This effect was mitigated by adding noise samples that have 0 quality scores to the training database. This led to the reduction of the bias effect and the enhancement of the prediction capabilities. The correlation value obtained from the combined method raised from 85% to 86%. The application of the feature reduction method to the objective methods led to an increase of the correlation values between the subjective scores and the objective scores of speech quality.

- A new signal-to-noise feedback (SNF) system has been introduced to enhance the intelligibility and the quality of speech in noisy environments. The new system utilizes an accelerometer, MPU6050, that is a cost-effective and has a digital output, which means it does not need an interface. This leads to simplicity in the design and a reduction of complexity. The device uses Raspberry Pi 3, which make it portable and easy to use in the patient's daily life. Unlike previous systems, the device uses 2 microphones instead of one microphone to make the system binaural and more effective, especially in multiple, high level noisy environments.

- The use of the new system led to enhancements in the SNR, intelligibility, and quality of the speech for participants. The use of the SNF device led to an increase in the SNR for all the speech tasks at different noise levels. It is also noted that the SNR at the no noise level was low because the participants tended to speak at low levels when there is no background noise, while this changed when a background noise was added and the SNR increased due to the natural Lombard effect that normal speakers tend to raise their voices when there is a background noise that is higher than 50 dB SPL. At higher levels of background noise, when there is a drop of intelligibility at all speech tasks, this is where the SNF device has the most significant impact in enhancing the intelligibility of the speech of the participants. When using the presented combined objective metric to assess the quality of speech before and after using the device, it is noted that the quality is acting proportionally with the intelligibility rating with a clear enhancement of the quality ratings when using the SNF device.

6.2 Study limitations and future work

Based on the work presented in this thesis, a number of recommendations exist for future work:

- This study focused on the objective assessment of the quality of Parkinsonian running speech. Although intelligibility and quality are two correlated attributes, a future study should focus on the objective assessment of the intelligibility of Parkinsonian speech. A study must investigate which acoustic features may have a higher representation of speech intelligibility, and which machine learning techniques may be more suited to match the objective intelligibility scores to their corresponding subjective results.

- This study did not investigate the objective assessment of loudness in Parkinsonian speech because there is an existing objective model for estimating loudness that is proved to be efficient in [71]. However, a future study should investigate the efficiency of this model in estimating the loudness of Parkinsonian speech.

- Future studies should collect a larger dataset of Parkinsonian speech. Increasing the size of the dataset will lead to higher accuracy in the estimation of Parkinsonian speech quality.

- • In applying deep learning machine learning, the size of the deep neural network was limited due to the limitation of the database's size. Increasing the size of the database will lead to implementing a deeper neural network that does not need extracted acoustic features as input, but it will estimate the quality directly from the running speech samples.

- The device has been tested on 3 control subjects. The device has been proved to be efficient in enhancing the SNR, intelligibility, and quality of the participants' speech. However, a future study must study the effect of using the device in outdoor environments by Parkinsonian subjects.

- There is a limitation on the work of the SNF device in severely varying environments. The device cannot detect a big variation in the value of SNR that is caused by a sudden big change in the value of the noise level. The problem arises when the noise level changes suddenly while the subject talks. In this case, the device will not be able to record the new noise level until the subject stop talking. Although the case of big variation of noise is not common in daily life, more investigation is needed to overcome this problem.

Bibliography

[1] R. Pahwa and K. E. Lyons, eds., *Handbook of Parkinson's disease*. Boca Raton, FL: CRC Press, Taylor & Francis Group, fifth edition ed., 2013. OCLC: ocn847481353.

[2] J. Parkinson, "An essay on the shaking palsy," *The Journal of neuropsychiatry and clinical neurosciences*, vol. 14, no. 2, pp. 223–236, 2002.

[3] D. R. Kumar, F. Aslinia, S. H. Yale, and J. J. Mazza, "Jean-martin charcot: the father of neurology," *Clinical medicine & research*, vol. 9, no. 1, pp. 46–49, 2011.

[4] M. Korell and C. M. Tanner, "Epidemiology of Parkinson's disease: An overview," in *Parkinson's Disease*, pp. 31–47, CRC Press, 2012.

[5] S. L. Wong, H. Gilmour, and P. L. Ramage-Morin, "Parkinson's disease: prevalence, diagnosis and impact," *Health reports*, vol. 25, no. 11, p. 10, 2014.

[6] K. Wirdefeldt, H.-O. Adami, P. Cole, D. Trichopoulos, and J. Mandel, "Epidemiology and etiology of Parkinson's disease: a review of the evidence," *European journal of epidemiology*, vol. 26, no. 1, p. 1, 2011.

[7] M. Toft and Z. K. Wszolek, "Overview of the genetics of Parkinsonism," in *Parkinson's Disease*, pp. 117–129, CRC Press, 2012.

[8] M. D. Andreetta, S. G. Adams, A. D. Dykstra, and M. Jog, "Evaluation of Speech Amplification Devices in Parkinson's Disease," *American Journal of Speech-Language Pathology*, vol. 25, p. 29, Feb. 2016.

[9] J. P. Clark, S. G. Adams, A. D. Dykstra, S. Moodie, and M. Jog, "Loudness perception and speech intensity control in Parkinson's disease," *Journal of Communication Disorders*, vol. 51, pp. 1–12, 2014.

[10] S. Adams, B.-H. Moon, A. Dykstra, K. Abrams, M. Jenkins, and M. Jog, "Effects of multitalker noise on conversational speech intensity in Parkinson's disease," *Journal of Medical Speech-Language Pathology*, vol. 14, no. 4, pp. 221–229, 2006.

[11] S. G. Adams, A. Dykstra, M. Jenkins, and M. Jog, "Speech-to-noise levels and conversational intelligibility in hypophonia and Parkinson's disease," *Journal of Medical Speech-Language Pathology*, vol. 16, no. 4, pp. 165–173, 2008.

[12] K. Tjaden, "Speech and swallowing in Parkinsons disease," *Topics in geriatric rehabilitation*, vol. 24, no. 2, p. 115, 2008.

[13] S. Skodda, W. Gronheit, N. Mancinelli, and U. Schlegel, "Progression of voice and speech impairment in the course of Parkinsons disease: a longitudinal study," *Parkinsons Disease*, vol. 2013, 2013.

[14] J. L. Goudreau and E. Ahlskog, "Symptomatic treatment of Parkinson's disease: Levodopa," in *Parkinson's Disease*, pp. 847–863, CRC Press, 2012.

[15] S. Sapir, L. Ramig, and C. Fox, "Speech therapy in the treatment of Parkinson's disease," in *Parkinson's Disease*, pp. 974–987, CRC Press, 2012.

[16] L. O. Ramig, C. Fox, and S. Sapir, "Speech treatment for Parkinson's disease," *Expert Review of Neurotherapeutics*, vol. 8, no. 2, pp. 297–309, 2008.

[17] A. M. Goberman and L. W. Elmer, "Acoustic analysis of clear versus conversational speech in individuals with Parkinson disease," *Journal of Communication Disorders*, vol. 38, no. 3, pp. 215–230, 2005.

[18] M. Andreetta, "A comparison of speech amplification devices for individuals with Parkinson's disease and hypophonia," Master's thesis, University of Western Ontario, 2013.

[19] S. Wight and N. Miller, "Lee silverman voice treatment for people with Parkinson's: audit of outcomes in a routine clinic," *International journal of language & communication disorders*, vol. 50, no. 2, pp. 215–225, 2015.

[20] D. M. Boudreaux, "Using the ambulatory phonation monitor to measure the vocal parameters of older people with and without Parkinson's disease," 2011.

[21] T. H. Falk, V. Parsa, J. F. Santos, K. Arehart, O. Hazrati, R. Huber, J. M. Kates, and S. Scollie, "Objective Quality and Intelligibility Prediction for Users of Assistive Listening Devices: Advantages and limitations of existing tools," *IEEE Signal Processing Magazine*, vol. 32, pp. 114–124, Mar. 2015.

[22] J. A. Whitfield and A. M. Goberman, "Articulatory–acoustic vowel space: Application to clear speech in individuals with Parkinson's disease," *Journal of communication disorders*, vol. 51, pp. 19–28, 2014.

[23] T. Khan, J. Westin, and M. Dougherty, "Cepstral separation difference: A novel approach for speech impairment quantification in Parkinson's disease," *Biocybernetics and Biomedical Engineering*, vol. 34, no. 1, pp. 25–34, 2014.

[24] R. B. Randall, "A history of cepstrum analysis and its application to mechanical problems," *Mechanical Systems and Signal Processing*, 2016.

[25] Y. D. Heman-Ackah, D. D. Michael, M. M. Baroody, R. Ostrowski, J. Hillenbrand, R. J. Heuer, M. Horman, and R. T. Sataloff, "Cepstral peak prominence: a more reliable measure of dysphonia," *Annals of Otology, Rhinology & Laryngology*, vol. 112, no. 4, pp. 324–333, 2003.

[26] X. Huang, A. Acero, and H.-W. Hon, *Spoken language processing: a guide to theory, algorithm, and system development*. Upper Saddle River, NJ: Prentice Hall PTR, 2001.

[27] Y. Shao, Z. Jin, D. Wang, and S. Srinivasan, "An auditory-based feature for robust speech recognition," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pp. 4625–4628, IEEE, 2009.

[28] T. Necciari, P. Balazs, N. Holighaus, and P. L. Søndergaard, "The erblet transform: An auditory-based time-frequency representation with perfect reconstruction," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 498–502, May 2013.

[29] M. Slaney, "Auditory toolbox," *Interval Research Corporation, Tech. Rep*, vol. 10, p. 1998, 1998.

[30] T. H. Falk, C. Zheng, and W.-Y. Chan, "A Non-Intrusive Quality and Intelligibility Measure of Reverberant and Dereverberated Speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, pp. 1766–1774, Sept. 2010.

[31] F. Chen, O. Hazrati, and P. C. Loizou, "Predicting the intelligibility of reverberant speech for cochlear implant listeners with a non-intrusive intelligibility measure," *Biomedical Signal Processing and Control*, vol. 8, pp. 311–314, May 2013.

[32] A. R. Fletcher, *Predicting treatment outcomes in dysarthria through speech feature analysis.* PhD thesis, University of Canterbury, 2016.

[33] H. Salehi and V. Parsa, "On nonintrusive speech quality estimation for hearing aids," in *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2015 IEEE Workshop on*, pp. 1–5, IEEE, 2015.

[34] V. Grancharov, D. Zhao, J. Lindblom, and W. Kleijn, "Low-Complexity, Nonintrusive Speech Quality Assessment," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, pp. 1948–1956, Nov. 2006.

[35] T. Hastie, R. Tibshirani, and J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. Springer series in statistics, New York, NY: Springer, 2nd ed ed., 2009.

[36] V. Vapnik, *Statistical learning theory. 1998*. Wiley, New York, 1998.

[37] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and computing*, vol. 14, no. 3, pp. 199–222, 2004.

[38] C. E. Rasmussen, "Gaussian processes in machine learning," in *Advanced lectures on machine learning*, pp. 63–71, Springer, 2004.

[39] C. Robert, "Machine learning, a probabilistic perspective," 2014.

[40] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, p. 436, 2015.

[41] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*, vol. 1. MIT press Cambridge, 2016.

[42] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[43] I. T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2065, p. 20150202, 2016.

[44] A. Gaballah, V. Parsa, M. Andreetta, and S. Adams, "Assessment of amplified Parkinsonian speech quality using deep learning," in *2018 IEEE Canadian Conference on Electrical & Computer Engineering (CCECE)*, pp. 1–4, IEEE, 2018.

[45] K. A. Al Mamun, M. Alhussein, K. Sailunaz, and M. S. Islam, "Cloud based framework for Parkinson's disease diagnosis and monitoring system for remote healthcare applications," *Future Generation Computer Systems*, vol. 66, pp. 36–47, 2017.

[46] A. Benba, A. Jilbab, and A. Hammouch, "Discriminating between patients with Parkinson's and neurological diseases using cepstral analysis," *IEEE transactions on neural systems and rehabilitation engineering*, vol. 24, no. 10, pp. 1100–1108, 2016.

[47] S. Jannetts and A. Lowit, "Cepstral analysis of hypokinetic and ataxic voices: correlations with perceptual and other acoustic measures," *Journal of Voice*, vol. 28, no. 6, pp. 673–680, 2014.

[48] D. Cushnie-Sparrow, S. Adams, A. Abeyesekera, M. Pieterman, G. Gilmore, and M. Jog, "Voice quality severity and responsiveness to levodopa in Parkinson's disease," *Journal of communication disorders*, vol. 76, pp. 1–10, 2018.

[49] D. D. Mehta, J. H. Van Stan, and R. E. Hillman, "Relationships between vocal function measures derived from an acoustic microphone and a subglottal neck-surface accelerometer," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 24, no. 4, pp. 659–668, 2016.

[50] M. Borsky, M. Cocude, D. D. Mehta, M. Zañartu, and J. Gudnason, "Classification of voice modes using neck-surface accelerometer data," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5060–5064, IEEE, 2017.

[51] N. Sadagopan and J. E. Huber, "Effects of loudness cues on respiration in individuals with Parkinson's disease," *Movement Disorders*, vol. 22, pp. 651–659, Apr. 2007.

[52] A. Gaballah, V. Parsa, M. Andreetta, and S. Adams, "Objective and subjective assessment of amplified Parkinsonian speech quality," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 2084–2087, IEEE, 2018.

[53] M. Farrús, J. Hernando, and P. Ejarque, "Jitter and shimmer measurements for speaker recognition," in *Eighth annual conference of the international speech communication association*, 2007.

[54] J. Fernandes, F. Teixeira, V. Guedes, A. Junior, and J. P. Teixeira, "Harmonic to noise ratio measurement-selection of window and length," *Procedia computer science*, vol. 138, pp. 280–285, 2018.

[55] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (version 5.1. 05)[computer program]. retrieved may 1, 2009," 2009.

[56] V. N. Vapnik, *The nature of statistical learning theory*. Statistics for engineering and information science, New York: Springer, 2nd ed ed., 2000.

[57] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on audio, speech, and language processing*, vol. 16, no. 1, pp. 229–238, 2008.

[58] A. Gaballah, V. Parsa, M. Andreetta, and S. Adams, "Objective and subjective speech quality assessment of amplification devices for patients with Parkinson's disease," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2019.

[59] M. F. Folstein, S. E. Folstein, and P. R. McHugh, ""mini-mental state": a practical method for grading the cognitive state of patients for the clinician," *Journal of psychiatric research*, vol. 12, no. 3, pp. 189–198, 1975.

[60] K. Yorkston, D. Beukelman, and R. Tice, "Sentence intelligibility test," *Lincoln, NE: Tice Technology Services*, 1996.

[61] J. Hillenbrand and R. A. Houde, "Acoustic correlates of breathy vocal quality: dysphonic voices and continuous speech," *Journal of Speech, Language, and Hearing Research*, vol. 39, no. 2, pp. 311–321, 1996.

[62] M. Narwaria, W. Lin, I. V. McLoughlin, S. Emmanuel, and C. L. Tien, "Non-intrusive speech quality assessment with support vector regression," in *International Conference on Multimedia Modeling*, pp. 325–335, Springer, 2010.

[63] J. H. Steiger, "Tests for comparing elements of a correlation matrix.," *Psychological bulletin*, vol. 87, no. 2, p. 245, 1980.

[64] S. G. Adams, A. D. Dykstra, and M. Jog, "A comparison of throat and head microphones in a pda-based evaluation of hypophonia in Parkinson's disease," *Journal of Medical Speech-Language Pathology*, vol. 20, no. 4, pp. 1–7, 2012.

[65] A. D. Dykstra, S. G. Adams, and M. Jog, "Examining the conversational speech intelligibility of individuals with hypophonia associated with Parkinson's disease," *Journal of Medical Speech-Language Pathology*, vol. 20, no. 4, pp. 53–59, 2012.

[66] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[67] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Transactions on speech and audio processing*, vol. 9, no. 5, pp. 504–512, 2001.

[68] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE signal processing letters*, vol. 6, no. 1, pp. 1–3, 1999.

[69] P. Papadopoulos, A. Tsiartas, S. Narayanan, P. Papadopoulos, A. Tsiartas, and S. Narayanan, "Long-term snr estimation of speech signals in known and unknown channel conditions," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 24, no. 12, pp. 2495–2506, 2016.

[70] D. D. Mehta, P. C. Chwalek, T. F. Quatieri, and L. J. Brattain, "Wireless neck-surface accelerometer and microphone on flex circuit with application to noise-robust monitoring of lombard speech.," in *INTERSPEECH*, pp. 684–688, 2017.

[71] B. C. J. Moore, B. R. Glasberg, A. Varathanathan, and J. Schlittenlacher, "A Loudness Model for Time-Varying Sounds Incorporating Binaural Inhibition," *Trends in Hearing*, vol. 20, pp. 1–16, Sept. 2016.

Curriculum Vitae

Name:	Amr Gaballah
Post-Secondary Education and Degrees:	Western University London, ON, Canada 2019 Doctor of Philosophy in Electrical and Computer Engineering
	Western University London, ON, Canada 2015 Master of Engineering
	Zagazig University Zagazig, Egypt 2008 Bachelor of Engineering
Honours and Awards:	Western Graduate Research Scholarship (WGRS) Western University 2015 - 2019
	ECE Graduate Travel Award Western University 2018
Certifications	Western Certificate in University Teaching and Learning Western University 2018
	Deep Learning Specialization Coursera 2018
	Machine Learning Coursera 2017

Certifications **(continued)**:	Alcatel-Lucent Network Routing Specialist Alcatel-Lucent 2013 Alcatel-Lucent NGN Platform Alcatel-Lucent, Lanyon, France 2009
Related Work Experience:	Graduate Teaching and Research Assistant University of Western Ontario 2014 - 2019 Internet Protocol (IP) and Data Access Engineer Alcatel-Lucent Egypt 2010 - 2014 Network Switching Subsystem (NSS) Field Engineer Alcatel-Lucent Egypt 2009 - 2010
Publications:	
Journal Publications	A. Gaballah, V. Parsa, M. Andreetta and S. Adams Objective and Subjective Speech Quality Assessment of Amplification Devices for Patients with Parkinson's Disease. in IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 27, no. 6, pp. 1226-1235, June 2019.
Conference Proceedings	Gaballah, A., Parsa, V., Andreetta, M., and Adams, S. (2018a). Assessment of Amplified Parkinsonian Speech Quality Using Deep Learning. 2018 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE), Quebec City, QC, Canada Gaballah, A., Parsa, V., Andreetta, M., and Adams, S. (2018b). Objective and Subjective Assessment of Amplified Parkinsonian Speech Quality. 2018 40^{th} Annual International Conference of The IEEE Engineering in Medicine and Biology Society (EMBC), Honolulu, HI, USA