Objective and Subjective Speech Quality Assessment of Amplification Devices for Patients With Parkinson's Disease

Amr Gaballah, *Student Member, IEEE*, Vijay Parsa, *Member, IEEE*, Monika Andreetta, and Scott Adams

***Abstract*— This paper investigated subjective and objective assessment of Parkinsonian speech quality. Speech stimuli were recorded from 11 Parkinsonian and 10 age-matched normal control participants under different amplification and environmental conditions. Quality ratings of the recorded stimuli were obtained from naïve listeners. For objective assessment, feature vectors were derived from the speech recordings based on temporal, spectral, and/or cepstral parametrization. These feature vectors were subsequently mapped to the predicted quality scores through several regression methods, including support vector regression, Gaussian process regression, and deep learning. Analyses of subjective speech quality ratings showed that Parkinsonian speech quality was significantly poorer than control subjects' speech quality, and that the amplification devices differentially affected perceived quality of Parkinsonian speech. Objective analyses revealed disparity in performance among feature vectors and mappers, with some feature vector and mapper combinations exhibiting statistically similar correlations with subjective ratings. A set consisting of cepstral, spectral, and modulation domain speech features when combined with Gaussian process regression or deep learning resulted in the highest correlation of 0.85 with the subjective data.**

***Index Terms*— Parkinson disease, speech quality, speech amplifiers, speech analysis, machine learning.**

I. Introduction

PARKINSON'S disease (PD) is the second most common neurodegenerative disease (the first being Alzheimer's disease), with an estimated prevalence of between 1 and 3 per 100 for people ages 65 years and older [1]. Worldwide, 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 increase in the $65-85$ year old age group to around $150-300$ persons [2]. There is therefore a significant research and clinical interest in effective diagnosis, treatment, and rehabilitation options for PD [1].

Statistics show that nearly 90% of people impaired with PD develop voice and speech disorders during the course of their disease [2]. 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 [2]. Hypophonia is considered the most frequent speech symptom in PD [3]. It is hypothesized that hypophonia may be attributed to a sensorimotor deficit in the self-perceived loudness of the individual's own speech [4]. 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-to-Noise Ratios (SNR) and reduced intelligibility when individuals with PD are conversing in moderate levels (65-70 dB SPL) of multi-talker noise [5], [6]. Furthermore, imprecise articulation and abnormal speech rate lead to a blurring of speech components, which also impact the understanding of PD speech [7]. 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, etc. PD speech, even when intelligible, may be perceived as abnormal given its harsh and breathy qualities [7].

There are many ways to treat speech impairments in PD; of these techniques three methods receive the most attention: the use of perceptually-based behavioral speech therapy, instrumentally-based biofeedback therapy, and prosthetic or assistive speech devices [8]. These treatment procedures aim to increase the speech intensity, improve the speech prosody, reduce rapid speech, and increase articulatory mobility and precision [8]. Although the first two methods have proved to be effective in the treatment of PD speech impairments, they lack the ability to transfer the treatment outside the clinical environment [9]. In other words, people impaired by PD show negligible improvements when they

Manuscript received February 15, 2019; revised April 10, 2019; accepted May 2, 2019. Date of publication May 7, 2019; date of current version June 6, 2019. The work of V. Parsa was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant. This paper was presented in part at the IEEE Engineering in Medicine and Biology Society (EMBS) Conference, HI, USA, in July 2018, and in part at the Canadian Electrical and Computer Engineering Conference, Quebec, Canada, in May 2018. *(Corresponding author: Amr Gaballah.)*

A. Gaballah is with the Department of Electrical and Computer Engineering, University of Western Ontario, London, ON N6A 5B9, Canada (e-mail: agaballa@uwo.ca).

V. Parsa is with the Department of Electrical and Computer Engineering, University of Western Ontario, London, ON N6A 5B9, Canada, and also with the School of Communication Sciences and Disorders, University of Western Ontario, London, ON N6A 5B9, Canada (e-mail: vparsa@uwo.ca).

M. Andreetta and S. Adams are with the School of Communication Sciences and Disorders, University of Western Ontario, London, ON N6A 5B9, Canada (e-mail: mschel@uwo.ca; sadams@uwo.ca).

Digital Object Identifier 10.1109/TNSRE.2019.2915172

leave the clinical treatment [9]. 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.

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 [8]. Amplification devices are used primarily to increase voice intensity and loudness, which leads to an improvement in the perceived speech intelligibility [3]. 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 [3]. Several amplification devices are commercially available for this purpose. The electroacoustic performance of these amplifiers is typically characterized using measures of frequency response, sensitivity, distortion etc., which are found in the device specification sheets. While these parameters are useful in basic 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 [3].

Andreetta *et al.* [3] 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 [3] 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.

While subjective assessment of amplifier 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 favor of objective, instrumental assessment of speech quality, where computational algorithms are used to quantify the speech quality without requiring the involvement of human subjects [10]. While such objective metrics are routinely used for evaluating telecommunication and assistive hearing devices [10], few studies have applied them to Parkinsonian speech quality assessment. 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.* [11] extracted features such as shimmer, jitter, and harmonic to noise ratio (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.* [12] 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 and Lowit [13] collected recordings of sustained vowel /a/ and continuous speech from 43 speakers with PD and 10 participants with ataxia. These recordings were perceptually rated on grade, roughness, breathiness, and asthenia dimensions by a trained listener. Acoustic measures extracted from recordings included Cepstral Peak Prominence (CPP), HNR, and jitter and shimmer – related measures. The CPP correlated best with the subjective ratings; however, the absolute correlation with ratings of more ecologically valid continuous speech was significantly lower than correlation with sustained vowel ratings (0.54 vs. 0.86 for the grade or overall quality rating respectively).

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. Existing objective methods of Parkinsonian speech quality assessment correlate poorly with subjective judgements. This paper investigates several alternative objective quality metrics and reports new results that show enhanced prediction of perceived Parkinsonian speech quality. The rest of the paper is organized as follows: Section. II provides the methodological details of speech recordings and their subjective quality evaluation, feature extraction, and mapping of the computed features to the predicted quality score. Results from subjective and objective analyses are presented in Section. III. Finally, Section. IV draws conclusions from this research and discusses future work.

II. Methods

A. Speech Recordings and Subjective Evaluation

Subjective data collection procedures outlined in this paper 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 [14] 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 [3], [8]. 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 [15]. For speech recordings in noisy environment, 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) [8]. Thus, a database of $21(11\ \text{PD} + 10\ \text{control speakers}) \times 2$ (conversation and sentence speech tasks) $\times 2$ (quiet and noisy environments) $\times 8$ (amplification options) = 672 speech recordings was 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.

B. Features & Their Computation

As the focus is on 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 [13]. In particular, the smoothed CPP value was computed from each speech recording following the algorithm given in [16]. In addition to the CPP, speech signal parametrization through filterbank analyses, modulation domain analyses, and Linear Prediction (LP) analyses was also explored, computational details of which are given in the following subsections.

1) Filterbank-Based Features: The MFCCs are popularly used as features in speech recognition systems [17], and have been shown to perform well in objective speech quality prediction [18]. The computation of MFCCs followed the procedure used in automatic speech recognition (ASR) research [17]. 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 paper, 40 filters constituted the mel filterbank, where the first 13 filters were linearly spaced and the last 27 filters were logarithmically spaced [17]. The log filterbank energies were decorrelated using the discrete cosine transform (DCT) and the lower 13 coefficients were retained [12], [17]. 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 comparsion 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 [19]. 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 [20]. Following Shao et al.'s [19] ASR research, 30 of the frame – averaged lower GFCCs and their first-order time differences were included in the GFCC feature set.

2) Modulation-Based Features: The speech-to-reverberation masking ration (SRMR) is an objective technique that was developed by Falk *et al.* [21] 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 [21].

In this method [21], 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 [21].

Another measure based on modulation-domain analysis is Modulation Area (ModA) parameter. There are some similarities between ModA and SRMR [10]. 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 down sampled 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 [22].

3) Linear Prediction – Based Features: We followed the LP-based feature extraction methodology in Low Complexity Quality Assessment (LCQA) proposed by Grancharov *et al.* [23]. 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.* [23] for computational formulae). These five quantities together with their first order differences constituted the 10-dimensional parameter vector per frame [23], [24]. 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 [24].

C. 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. To express this mathematically, we have [25]:

$$\mathbf{y} = f(\theta, \mathbf{X}) + \mathbf{b}, \tag{1}$$

where θ represents the parameters and functions associated with the feature mapper, $\mathbf{X}$ is the feature matrix that has size $m \times n$, m is the number of training samples, n is the size of the feature vector, $\mathbf{y}$ are the subjective scores corresponding to the training samples, and $\mathbf{b}$ is the prediction error. Commonly used feature mappers include linear regression (LR), the support vector regression (SVR) [26], and the Gaussian Process Regression (GPR) [27].

Recent developments in machine learning for classification and regression have focused on deep learning. In deep learning [28], the learning process is divided into multiple layers where features are extracted from each layer. Deep learning then uses the back propagation 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 [29], [30]. The DNN structure used in this paper 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 were kept small to avoid overfitting of the model.

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 finetuning the parameters and hyperparameters of the feature mapper.

D. 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. 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.

III. Results

A. Subjective Results

Fig. 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. 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. 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),

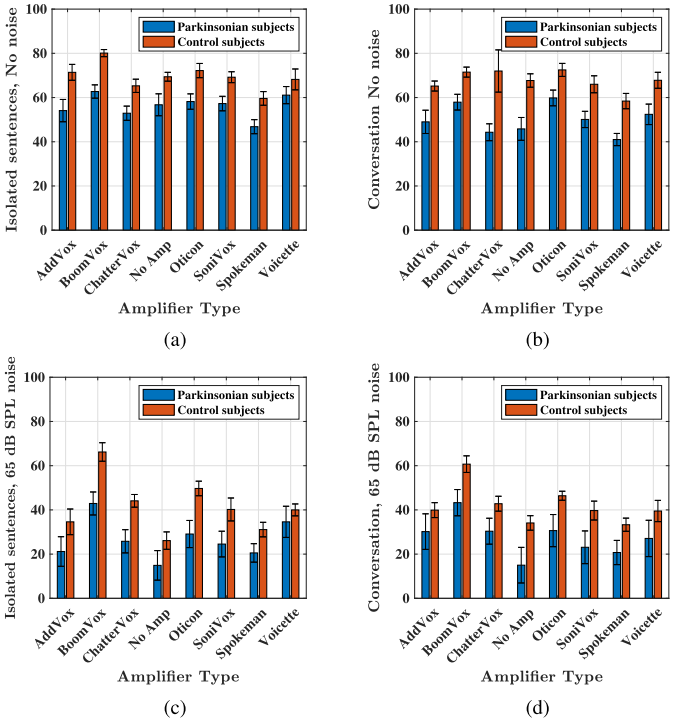
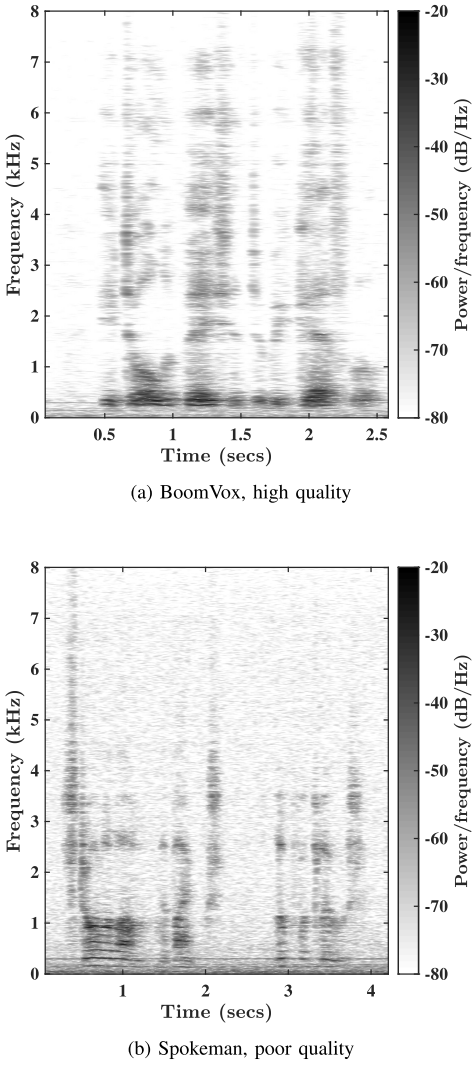


Fig. 1. Averaged subjective speech quality ratings for control and Parkinsonian speech samples, with the error bars denoting one standard deviation.

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 [8]. 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 ($F(1, 18) = 25.26$, $p < 0.001$, $\eta_p^2 = 0.584$), background noise ($F(1, 18) = 227.75$, $p < 0.001$, $\eta_p^2 = 0.927$), and device type ($F(7, 126) = 37.16$, $p < 0.001$, $\eta_p^2 = 0.674$). There was no significant main effect of speech task ($F(1, 18) = 1.55$, $p = 0.229$, $\eta_p^2 = 0.079$), 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 ($F(1, 18) = 0.002$, $p = 0.964$, $\eta_p^2 = 0.00$), speaker group by speech task ($F(1, 18) = 0.878$, $p = 0.361$, $\eta_p^2 = 0.046$), and speaker group by device ($F(7, 126) = 1.30$, $p = 0.254$, $\eta_p^2 = 0.068$), 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 ($F(3.90, 70.21) = 7.80$, $p < 0.001$, $\eta_p^2 = 0.302$). This interaction stemmed from the differential quality ratings associated with the ChatterVox device. As can be seen from Fig. 1, the ChatterVox device received lower quality ratings than no amplification in quiet conditions (Fig. 1a & Fig. 1b), but higher ratings in conditions involving background noise (Fig. 1c & Fig. 1d). Finally, no significant three-way or four-way interactions were found.



(a) BoomVox, high quality

(b) Spokeman, poor quality

Fig. 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".

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 [8].

B. Objective Results

Fig. 2 displays sample spectrograms associated with speech samples collected from a subject with Parkinson's disease in quiet condition. Fig. 2a displays the spectrogram of an isolated sentence produced by the subject while utilizing the BoomVox amplifier. Fig. 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 of the subjective speech quality scores, and ρ is the correlation coefficient between the true and predicted quality scores [31].

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. 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.

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 respectively, which were mapped separately using four learning algorithms *viz.* LR, SVR, GPR, and DNN.

Table I 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. 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 [32] 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.

3) Reduced Multiple Features Objective Metrics: By applying the feature selection and reduction method mentioned in subsection. II-D, the number of features for LCQA, MFCC, and GFCC were reduced to 7, 16, and 11 respectively. Fig. 4

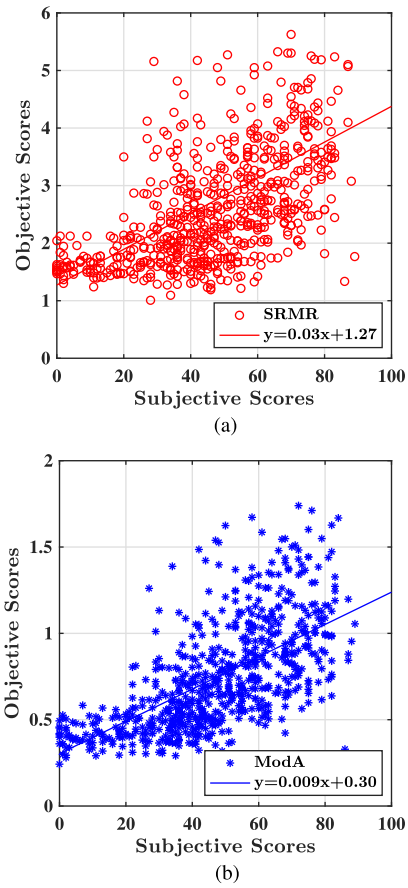


Fig. 3. Scatter plot between the objective and subjective scores for all the speech recordings in the database. Data was plotted for the four conditions, *viz* isolated sentences in quiet and in 65 dB SPL background noise (labeled SQT (no noise) and SQT (noise) respectively), and sentences extracted from conversation in quiet and in 65 dB SPL background noise. (a) SRMR. (b) ModA.

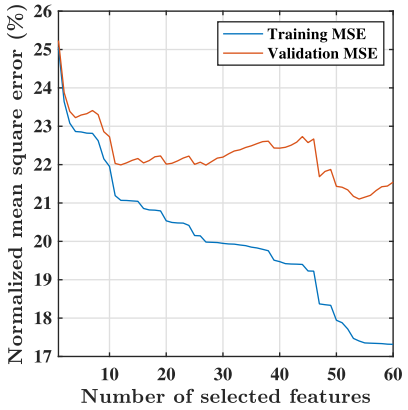


Fig. 4. The normalized mean square error (MSE) between actual and predicted speech quality scores as a function of the number of GFCC features. The MSE data for training and validation datasets are shown separately, and greater separation between these two lines is a potential indicator of overfitting.

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

TABLE I

CORRELATION COEFFICIENTS AND SDPE VALUES BETWEEN OBJECTIVE AND SUBJECTIVE DATA. BOLD CORRELATION COEFFICIENTS REPRESENT FEATURE VECTOR AND MAPPER COMBINATIONS THAT PERFORMED STATISTICALLY SIMILAR WITH THE TEST DATASET

Metric	Non-reduced feature set				Reduced feature set			
	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

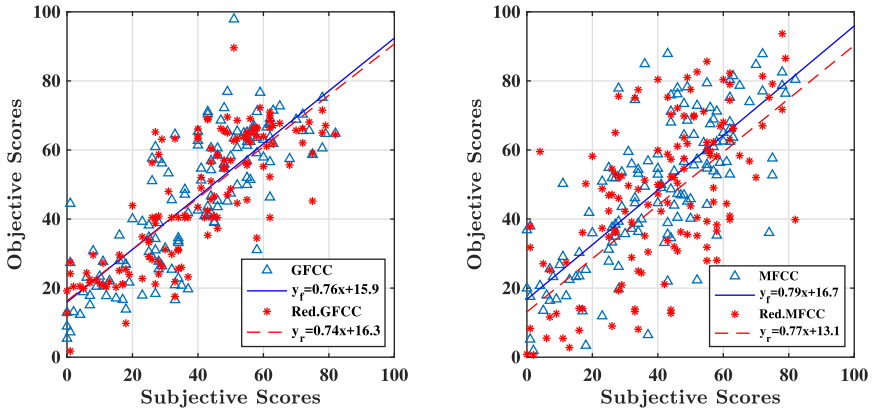


Fig. 5. Scatter plot between the objective and subjective data using deep learning for the test dataset. The data is plotted for both unreduced and reduced GFCC and MFCC feature vectors. y_f is the vector of obtained scores from the linear regression for the non reduced model, while y_r is the vector of obtained scores from the linear regression of the reduced model. (a) GFCC. (b) MFCC.

TABLE II

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

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 chances of overfitting. This implied that the selected 11 features will avert overfitting and potentially lead to better results.

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

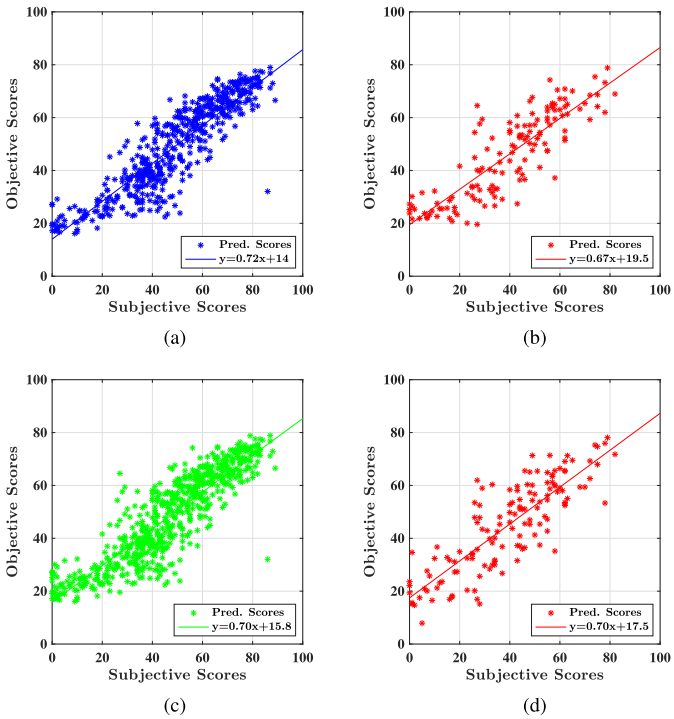


Fig. 6. Subjective scores against objective scores for the combined metric using GPR. (a) Training dataset. (b) Test dataset. (c) Full dataset. (d) Test dataset after noise training.

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 [10], [17], in that the GFCCs appear to capture perceptually salient features perhaps due to their better approximation of the auditory filterbank characteristics. Fig. 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) 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 II 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. 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, the test, and the full 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 Fig. 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. 6d shows 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.

IV. 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 describe 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 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 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 I. These results highlight the robustness of the learning model in predicting the quality of Parkinsonian speech. Showed that the original and the new objective scores are statistically similar.

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 paper 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.

REFERENCES

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

[2] R. Pahwa and K. E. Lyons, Eds., *Handbook of Parkinson's Disease*, 5th ed. Boca Raton, FL, USA: CRC Press, 2013.

[3] M. D. Andreetta, S. G. Adams, A. D. Dykstra, and M. Jog, "Evaluation of speech amplification devices in Parkinson's disease," *Amer. J. Speech-Lang. Pathol.*, vol. 25, no. 1, pp. 29–45, Feb. 2016. [Online]. Available: http://ajslp.pubs.asha.org/article.aspx?doi=10.1044/2015_AJSLP-15-0008

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

[5] 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," *J. Med. Speech-Lang. Pathol.*, vol. 14, no. 4, pp. 221–229, Dec. 2006.

[6] S. G. Adams, A. Dykstra, M. Jenkins, and M. Jog, "Speech-to-noise levels and conversational intelligibility in hypophonia and Parkinson's disease," *J. Med. Speech-Lang. Pathol.*, vol. 16, no. 4, pp. 165–173, May 2019.

[7] K. Tjaden, "Speech and swallowing in Parkinson's disease," *Topics Geriatric Rehabil.*, vol. 24, no. 2, p. 115, 2008.

[8] M. Andreetta, "A comparison of speech amplification devices for individuals with Parkinson's disease and hypophonia," M.S. thesis, School Commun. Sci. Disorders, Western Univ., London, ON, Canada, 2013.

[9] S. Wight and N. Miller, "Lee silverman voice treatment for people with Parkinson's: Audit of outcomes in a routine clinic," *Int. J. Lang. Commun. Disorders*, vol. 50, no. 2, pp. 215–225, Mar. 2015.

[10] T. H. Falk *et al.*, "Objective quality and intelligibility prediction for users of assistive listening devices: Advantages and limitations of existing tools," *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 114–124, Mar. 2015.

[11] 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 Gener. Comput. Syst.*, vol. 66, pp. 36–47, Jan. 2017.

[12] A. Benba, A. Jilbab, and A. Hammouch, "Discriminating between patients with Parkinson's and neurological diseases using cepstral analysis," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 24, no. 10, pp. 1100–1108, Oct. 2016.

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

[14] 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," *J. Psychiatric Res.*, vol. 12, no. 3, pp. 189–198, Nov. 1975.

[15] K. Yorkston, D. Beukelman, and R. Tice, *Sentence Intelligibility Test*. Lincoln, NE, USA: Tice Technologies, 1996.

[16] J. Hillenbrand and R. A. Houde, "Acoustic correlates of breathy vocal quality: Dysphonic voices and continuous speech," *J. Speech, Lang., Hearing Res.*, vol. 39, no. 2, pp. 311–321, 1996.

[17] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*. Upper Saddle River, NJ, USA: Prentice-Hall, 2001.

[18] M. Narwaria, W. Lin, I. V. McLoughlin, S. Emmanuel, and C. L. Tien, "Non-intrusive speech quality assessment with support vector regression," in *Proc. Int. Conf. Multimedia Modeling*, Jan. 2010, pp. 325–335.

[19] Y. Shao, Z. Jin, D. Wang, and S. Srinivasan, "An auditory-based feature for robust speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Apr. 2009, pp. 4625–4628. [Online]. Available: http://ieeexplore.ieee.org/abstract/document/4960661/

[20] M. Slaney, "Auditory toolbox," Interval Res. Corp., Interval Res. Corp., Palo Alto, CA, US, Tech. Rep. 1998-010, 1998, vol. 10, p. 1998. [Online]. Available: http://www.tka4.org/materials/lib/Articles-Books/Speech%20Recognition/AuditoryToolboxTechReport.pdf

[21] T. H. Falk, C. Zheng, and W.-Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 7, pp. 1766–1774, Sep. 2010. [Online]. Available: http://ieeexplore.ieee.org/document/5547575/

[22] F. Chen, O. Hazrati, and P. C. Loizou, "Predicting the intelligibility of reverberant speech for cochlear implant listeners with a non-intrusive intelligibility measure," *Biomed. Signal Process. Control*, vol. 8, no. 3, pp. 311–314, May 2013. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S1746809412001218

[23] V. Grancharov, D. Y. Zhao, J. Lindblom, and W. B. Kleijn, "Low-complexity, nonintrusive speech quality assessment," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 6, pp. 1948–1956, Nov. 2006.

[24] H. Salehi and V. Parsa, "On nonintrusive speech quality estimation for hearing aids," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, Oct. 2015, pp. 1–5. [Online]. Available: http://ieeexplore.ieee.org/abstract/document/7336897/

[25] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer Series in Statistics), 2nd ed. New York, NY, USA: Springer, 2009.

[26] V. Vapnik, *Statistical Learning Theory*. Hoboken, NJ, USA: Wiley, 1998.

[27] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. Cambridge, MA, USA: MIT Press, 2012.

[28] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.

[29] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep Learning*, vol. 1. Cambridge, MA, USA: MIT Press, 2016.

[30] D. P. Kingma and J. Ba. (2014). "Adam: A method for stochastic optimization." [Online]. Available: https://arxiv.org/abs/1412.6980

[31] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 1, pp. 229–238, Jan. 2008.

[32] J. H. Steiger, "Tests for comparing elements of a correlation matrix," *Psychol. Bull.*, vol. 87, no. 2, p. 245, Mar. 1980.