Assessment of Amplified Parkinsonian Speech Quality Using Deep Learning

Amr Gaballah, Vijay Parsa, Monika Andreetta, and Scott Adams
Western University, London, Ontario, Canada
Emails: {agaballa, vparsa, mschel, sadams}@uwo.ca

***Abstract*—In this paper, deep neural networks (DNNs) are applied to features extracted from Parkinsonian speech recordings to predict their perceived quality. This procedure was also used to benchmark the electroacoustic characteristics of speech amplifiers used by people afflicted with Parkinson Disease (PD). Speech recordings were obtained from 11 PD subjects and 10 normal controls, with and without the assistance of 7 different speech amplifiers, and their quality was assessed subjectively by normal hearing listeners. Mel-frequency and Gammatone frequency cepstral coefficients (MFCCs and GFCCs respectively) and their first order derivatives were extracted as features, and given as input to the DNN. Two optimizers were used to train the neural network, namely stochastic gradient descent (SGD) and Adam optimizers. The paper also shows the effect of feature reduction in enhancing the performance of the objective predictors. Experimental results showed that a trained DNN with reduced set of GFCC features outperforms other objective metrics in terms of correlation with the subjective measures.**

I. Introduction

The worldwide rise in the mean population age has led to an increase in the prevalence of neurological disorders such as the Parkinson's Disease (PD) [1]. PD causes the production rate of dopmaine, which is a chemical substance responsible for the control of the human motor system, to drop down [1]. As a result of that, symptoms such as rigidity, slowness of movement, difficulty with walking, and speech impairments begin to appear [1]. PD speech is commonly characterized by monotonous pitch and loudness, short rushes of speech, imprecise articulation, overall increased speech rate, and lowered overall loudness [2]. Reduced speech intensity, termed hypophonia, is a result of lessened respiratory support caused by increased rigidity of the chest wall and abdomen [3]. Hypophonia reduces the clarity and the intelligibility of PD speech, especially in noisy environments with inferior Signal-to-Noise Ratios (SNRs) [3], [4].

Amplification devices are typically used to increase the voice intensity and loudness with concomitant improvement in PD speech intelligibility. Moreover, amplification devices decrease vocal effort by PD speakers and enhance self-perception and correction of their speech [4]. SNR, frequency response, Total Harmonic Distortion (THD), etc. are all parameters that quantify the performance of the amplifiers, but they do not provide information on the perceived intelligibility and quality of the amplified PD speech. Thus, systematic benchmarking of these devices in a perceptually relevant manner is important from both device development and clinical viewpoints [4]. The focus of this paper is on the assessment of the impact of amplification devices on perceived PD speech quality.

In general, speech quality is assessed subjectively, wherein a group of people listen to the speech recordings and evaluate each recording on a rating scale. The so-called Mean Opinion Score (MOS) is the average of these ratings that designates the perceived quality of the recording [5]. While the subjective ratings have high face validity and can be considered the gold standard, they are also time- and resource-intensive [5]. Thus objective, instrumental metrics that estimate the perceived quality without the need for a human intervention are attractive [5]. Such objective indices are routinely used in telecommunications and audio engineering fields, but are sparsely used in disordered speech quality estimation. In particular, the application of objective speech quality metrics in predicting the perceived quality of amplified PD speech has not been investigated before. As objective metrics commonly embody many features, a need for a regression technique that transforms these many features into a single numerical predicted quality score is needed. In this paper deep learning is used for such purpose.

In the recent years, deep neural networks (DNNs) have received a significant attention as an authoritative supervised machine learning techniques for regression and classification. As a machine learning algorithm, it is widely used in several applications such as image classification and processing, natural language processing (NLP), and speech recognition [6]. This paper exploits the recent advances in deep learning regression techniques wherein a set of features extracted from PD speech recordings are mapped to their perceived subjective quality scores using two different optimizers. The paper is organized as follows, in Section II, a brief description of the objective evaluation procedure is given, then Section III gives a summarized introduction of deep learning and some of its optimization techniques that are used in this paper. Afterwards, the methodology followed throughout this work is presented in Section IV. The obtained results are discussed in Section V. Finally, Section VI concludes this paper, with a brief description of future work.

II. Objective Evaluation

Objective evaluation methods can be broadly classified into two categories, intrusive and non-intrusive methods. Intrusive techniques need a clean reference signal to compare with the measured signal in order to calculate the metric. In the

978-1-5386-2410-4/18/$31.00 ©2018 IEEE

context of PD speech evaluation however, a clean reference signal is unavailable. As such, objective estimates of PD speech intelligibility, quality, and loudness must rely on "non-intrusive" techniques [5]. A typical non-intrusive metric extracts a number of features from the speech recording, and the subsequent features vector is mapped to a predicted quality score using a features' mapping technique. The following subsections present a brief description of two of commonly used feature sets, while the next section describes the deep learning-based feature mapping in greater detail.

A. Mel Frequency Cepstrum Coefficients (MFCC)

MFCCs are derived by transforming the short-term speech spectra into the nonlinear mel scale. The input speech signal is framed where each frame is 256 samples each, then each frame is windowed by a Hamming window and transformed to the frequency domain using the Fast Fourier Transform (FFT). The narrowband spectra are processed by the triangular mel-scale filterbank which can be expressed as [7]:

$$H_m(k) = \begin{cases} 0 & f_k < f(m-1) \\ \frac{f_k - f(m-1)}{f(m) - f(m-1)} & f(m-1) \le f_k \le f(m) \\ \frac{f(m+1) - f_k}{f(m+1) - f(m)} & f(m) < f_k \le f(m+1) \end{cases}, \quad (1)$$

where $f(\cdot)$ is the list of mel linearly spaced frequencies, m is the filter number, f_k is the frequency at FFT bin k. The mel filterbank is constructed such that there are 13 linearly spaced and 27 logarithmically spaced filters spanning the $13-6854$ Hz frequency range [8]. MFCCs are then derived by computing the log of the mel-filtered spectrum and applying the discrete cosine transform (DCT).

B. Gammatone Frequency Cepstrum Coefficients (GFCC)

The Gammatone filterbank has a better performance in modeling the auditory filterbank than the mel filterbank; hence, cepstral coefficients extracted using the Gammatone filter bank have a better speech recognition performance than MFCCs [9]. In GFCC, the equivalent rectangular bandwidth (ERB) scale is used. The impulse response of the Gammatone filter is given by,

$$g(t) = \frac{a t^{n-1} \cos\left(2\pi f_c t + \varphi\right)}{e^{2\pi b t}}, \quad (2)$$

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, and t is the time in seconds. The value of the filter order and carrier phase are set to be $n = 4$, $b = 1.019$ ERB, $\varphi = 0$. Laplace transform and the impulse invariance method are applied to transform the impulse response of the Gammatone filter into a discrete time equivalent filter. Extracted digital filters are applied to the framed speech to extract the energies of the filter banks [8]. GFCCs are then computed by applying DCT to the log of Gammatone filtered spectra.

III. Deep learning

In deep learning [10], the learning process is divided into multiple layers where features are extracted from each layer. Deep learning then uses 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 of a human intervention to design these layers of neurons, since they are learned from the input data solely.

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 paper, the $Relu$ 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 $n \times m$ features 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 follows [6]

$$Z^{[i]} = W^{[i]} A^{[i-1]} + b^{[i]} \quad (3)$$

$$A^{[i]} = g\left(Z^{[i]}\right), \quad (4)$$

where i is the layer number that ranges from 1 to l, and $g(\cdot)$ is the activation function. The value of $A^{[0]}$ equals to the input features matrix X, while the value of $A^{[l]}$ equals to the objective scores vector $\hat{y}$. In this paper, the least mean square error (LMSE) is calculated as [6]:

$$J = \frac{1}{2m} \sum (y - \hat{y})^2, \quad (5)$$

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 Eq. (5). The stochastic gradient descent backward propagation can be expressed as:

$$W^{[i]} = W^{[i]} - \alpha \frac{\delta J}{\delta W^{[i]}}, \quad (6)$$

where i is the layer number. After modifying the weights, the forward and backward propagations are iterated to minimize the cost function. The learning rate α must be chosen carefully so that the output of the cost function does not overshoot its global minimum.

The Adaptive Momentum Estimation (Adam) modifies the weights using the first and second moment of gradients and can be expressed as,

$$W^{[i]} = W^{[i]} - \alpha \frac{\hat{V}^{[i]}_{\delta W}}{\sqrt{\hat{S}^{[i]}_{\delta W} + \epsilon}}, \quad (7)$$

where $V_{\delta W}^{[i]}$ is the first gradient moment, $\hat{S}_{\delta W}^{[i]}$ is the second gradient moment, and ϵ is a correction factor [11].

IV. Methodology

A. Subjective Database

Subjective data were collected in the Speech Movement Disorders Laboratory at the University of Western Ontario after obtaining ethics approval from the University's health sciences research ethics board [12]. Eleven PD subjects with an age range of 58 to 80 years participated in the study. Moreover, recordings from 10 normal subjects were collected as control data. Each subject gave recordings with the presence of 4 acoustic scenarios, two of these scenarios incudes repeating a sentence with and without the presence of 65 dB SPL noise, while the other two are to perform a natural conversation with the same noise codtions [12]. Since the goal of the study was to assess the performance of speech amplifiers, the above four speech tasks were carried out with no amplification, and with seven 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) [4]. The speech recordings were played back to 10 normal hearing listener, and a quality rating between 0 and 1 was obtained. The quality ratings served as labels for the dataset which was separated randomly into two groups. The first group contained 80% of the data and served as a training dataset, while the remaining 20% of the database were used to test the performance of the obtained neural network.

B. Feature selection and reduction

A higher dimensionality of the feature vector may cause overfitting . In such situations, extracted number of features for each metric must be reduced before applying to the neural network to avoid overfitting. To accomplish this goal, the correlation between each feature group and the subjective scores is obtained, and then, the features are rearranged according to their correlation values. Subsequently, a Monte Carlo algorithm is applied to extract the minimum number of features that minimizes the cost function for each of the training and the test datasets. This algorithm takes the the rearranged features' matrix and the subjective scores vector as two inputs. Afterwards, the data is split into two groups, training and test data set, where they are selected randomly. The training dataset contains 80% of the data, and the test dataset contains 20%. 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.

C. Deep neural network (DNN)

A unified structure of DNN is used for each of GFCC and MFCC regression. The first layer was the input layer which consisted of 11 features. Next to that, 2 hidden layers are deployed, where the first hidden layer is formed of 25 neurons while the second layer contains 12 neurons. Finally, the output layer has 1 neuron which represents the objective quality of the speech signal. It is noticed that in hidden layers, a small number of neurons were used to avoid overfitting of the model.

V. Results

By applying the feature selection and reduction method mentioned in Subsection IV-B, it was found that MFCC features can be reduced from 26 features to 11 features only, and GFCC features are cut from 60 features to 11 features. Fig. 1 shows that a subset of GFCC 11 features of achieves the minimum MSE for the test dataset.

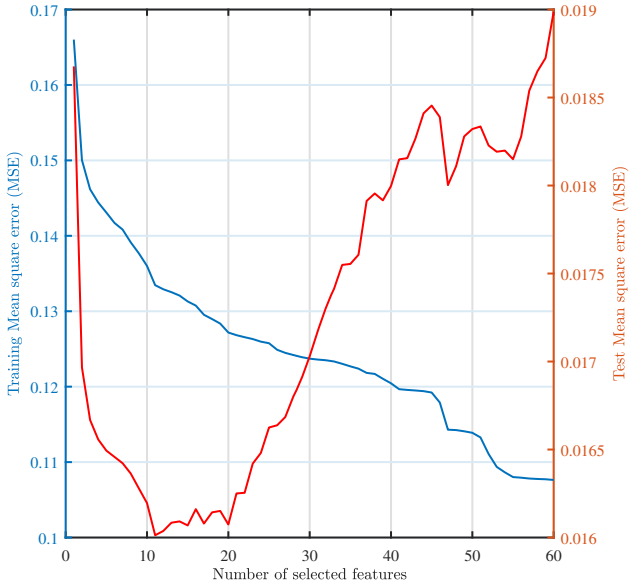

Fig. 1. Mean square error for GFCC training and test datasets

Table. I presents the correlation values between the subjective scores and objective scores for each DNN regression algorithm. It can be noticed that the Adam optimizer has a much higher performance than SGD with both MFCC and GFCC feature sets. The second observation comes from the fact that GFCC neural networks have higher correlation results with the subjective scores more than MFCC networks. Moreover, GFCC networks are prone to less overfitting than MFCCs. The better performance of GFCC feature set is in line with ASR research. GFCCs appear to capture perceptually salient features perhaps due to their better approximation of the auditory filterbank characteristics. To reduce overfitting and to enhance correlation with subjective measures, MFCCs may require a greater number of training samples. Finally, it is noticed that reducing the number of features enhanced the correlation results of MFCC when the optimizer used was SGD, however, it did the opposite when the feature reduction was applied on networks with Adam optimizer. On the other hand, GFCC correlation values did not change significantly

TABLE I
CORRELATION VALUES OF OBJECTIVE METRICS.

Metric	Correlation (Training Dataset)	Correlation (Test Dataset)
MFCC-SGD	0.52	0.52
GFCC-SGD	0.70	0.70
MFCC(Red)-SGD	0.67	0.67
GFCC(Red)-SGD	0.80	0.78
MFCC-Adam	0.95	0.75
GFCC-Adam	0.83	0.80
MFCC(Red)-Adam	0.98	0.63
GFCC(Red)-Adam	0.81	0.81

TABLE II
AVERAGED SUBJECTIVE AND OBJECTIVE SCORES PER DEVICE

Metric	Averaged subjective score	Averaged objective score
ADDvox	0.45	0.41
BoomVox	0.60	0.60
ChatterVox	0.47	0.40
No device	0.40	0.34
Oticon	0.52	0.53
SoniVox	0.45	0.44
Spokeman	0.39	0.42
Voicette	0.49	0.50

when feature reduction was applied to its neural networks, indicating more robustness.

Fig. 2 shows the plot of the subjective against the objective scores that were obtained when applying the Adam optimizer. The DNN with reduced GFCC feature set resulted in the highest correlation and less dispersion with the subjective scores.

Table. II shows the averaged values for all the test dataset quality scores for subjective and reduced GFCC neural network metric per device. There is a high similarity between objective metric and subjective metric scores. As such, it is this objective metric is a promising candidate for benchmarking speech amplifiers intended for use with PD patients.

VI. DISCUSSION & CONCLUSIONS

MFCC and GFCC objective metrics were applied to determine the quality scores of recordings from 11 PD subjects and 10 normal people. It was shown that using Adam optimizer for regression is more efficient than using SGD optimizer. The results showed that GFCC has a better performance in evaluating the quality of Parkinsonian speech than MFCC metric. There is scope for improvement, however, as the highest correlation obtained for the test dataset was 0.81. Perhaps this can be achieved by increasing the size of the training dataset. As collecting more recordings from PD subjects may not be feasible, a potential alternative is transfer learning wherein a pre-trained neural network that has been derived for speech recognition applications is incrementally adapted to predict

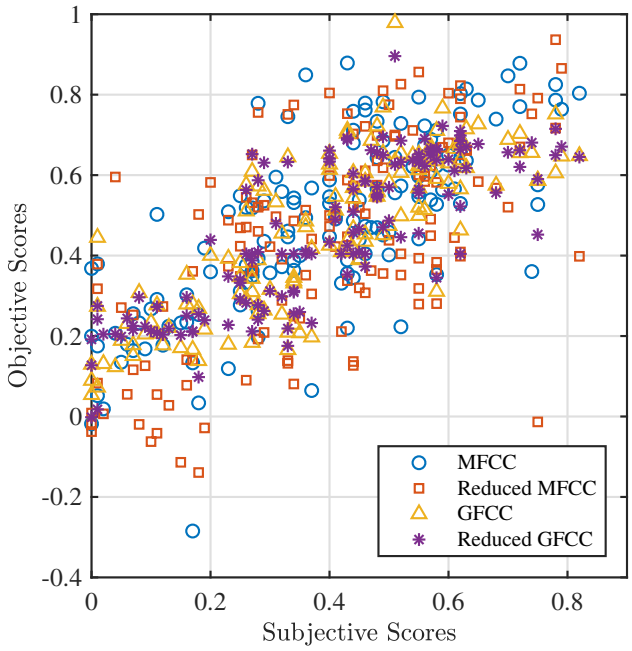

Fig. 2. Objective scores obtained by Adam optimizer.

amplified speech quality. This idea will be investigated as an extension to this work in future research.

REFERENCES

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

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

[3] N. Sadagopan and J. E. Huber, "Effects of loudness cues on respiration in individuals with Parkinson's disease," *Movement Disorders*, vol. 22, no. 5, pp. 651–659, Apr. 2007.

[4] 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, no. 1, p. 29, Feb. 2016.

[5] 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, no. 2, pp. 114–124, Mar. 2015.

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

[7] 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.

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

[9] 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*. IEEE, 2009, pp. 4625–4628.

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

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

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