Combined Approach of Speech Intelligibility for Calming Children with Speech Disorders

VISHALI MURALIDHARAN[1], R. JOSPHINELEELA[2]

[1] *M.E. Computer Science and Engineering, Panimalar Engineering College, Chennai, Tamil Nadu*

[2] *professor, Dept of Computer Science and engineering, Panimalar Engineering College, Chennai, Tamil Nadu*

Abstract— Evaluating the degree of dysarthria's severity can help pathologists plan therapy, help automated dysarthric speech recognition systems, and give insight into how well the patient is improving. This article presents comparative research on the use of several deep learning algorithms and acoustic characteristics for the categorization of dysarthria severity levels. First, we assess the fundamental architecture options, including the convolutional neural network, DNN, GRU, and LSTM, utilizing fundamental characteristics Subsequently, DNN models are used to assess aspects related to speech disorders.

Index Terms- Speech Disorder, Deep Neural Network, Severity

I. INTRODUCTION

Speech disorder is a motor speech disease brought on by either a breakdown in the speech production subsystems or inadequate coordination. It either develops together with any neuro-degenerative illness or results from a neurological damage such as cerebral palsy [1]. It causes uneven speech pace, aberrant prosody, poor audibility, and inaccurate articulation, all of which worsen speech quality. The patients would sound harsher when speaking, have poor facial responses, hypernasality, and increased weariness. Thus, dysarthric individuals are unable to phonetically construct or speak syntactically proper sentences, even when they can

be conceived. Their social life is impacted and they start speaking incoherently as a result. Due to their poor muscular coordination, dysarthric patients experience physical limitations like shaky hands, which reduces interactive apps. normal ASRs created have substantial mistake rates when utilized by disordered speakers.

II. RELATED WORK

Because of their familiarity with the patient, professionals' perceptions of their evaluations would differ depending on their background and listening abilities. Nonetheless, monitoring customers during recuperation. establishes necessity of an automated technique for classifying the severity of dysarthria. Additionally, ASR systems designed for people with dysarthria may function better as a result of this categorization.

Speech processing uses a wide range of perceptual features. Through the application of machine learning classifiers, MFCCs have demonstrated their value in the literature and for identifying the severity of dysarthria. MFCCs demonstrate their effectiveness that is similar for the identification. A multilayer perceptron (MLP) is used to classify the severity of dysarthria. The advantages of DBN features are negligible. The basic fared better than the glottal features utilized in [8] when it came to recognizing solitary perform better. These studies inspired us to examine the effectiveness of many deep learning models for dysarthric severity estimate using the fundamental MFCC features in order to see whether any appreciable gains over machine learning classifiers could be made. CQCCs have been a great option having first been suggested. The authors of [11] have demonstrated the effectiveness of CQT in dysarthria severity detection by demonstrating how the intensity of formants and harmonics in CQT spectrograms diminishes as the intelligibility level drops. Additionally, CQCCs have shown promising outcomes when included as baseline characteristics in [12]. These results have encouraged us to examine CQCCs' suitability for the suggested job. Numerous other novel and established features have been investigated in the literature to improve the accuracy of dysarthria severity identification. These include the use of audio descriptors in [14], the introduction of the PE-SFCC in [12], and the use of breath-iness indices in [13].

III. INCLUSION

- Using MFCCs and CQCCs, we performed performance analyses of the fundamental deep learning architectures, including CNN, DNN, GRU, and LSTM. Our first stage of research using MFCCs is documented.
- Prosodic, glottal, phonetic, and articulatory aspects are evaluated using DNN classifiers. The concatenated feature set is further subjected to dimensionality reduction, with the outcomes being analyzed.
- Putting into practice a "two-level learning classifier" that classifies data using DNNs at the secondlevel. Trials use leave-one-speaker-out (LOSO) round-robin cross validation

IV. DATASETS

UA-Speech is used. for assessing the suggested task. The audio three-dimensional characteristics eight dysarthric patients and seven healthy speakers make up the TORGO database. Only the words utilized in this study are included in the corpus, which also includes non-words, words, and limited and unrestricted phrases. Thirteen speakers in good health and nineteen patients with dysarthria are represented in UA-Speech. Only 15 patients' worth of data are accessible, though. 155 popular terms are used three times.

matching the 100 frequent words in the Brown corpus, computer instructions, international radio alphabets, and English numbers. The training data consists of 465 frequent words per speaker, or 6975 utterances in total. Additionally, each speaker in the corpus contains 300 unique unusual words that were chosen to optimize variety. These words were taken from children's novels that Project Gutenberg had digitized [27]. These are tested (a total of 4500 unseen words) in order to assess how resilient the models are. The sixth channel's data, at fs = 16 kHz in the microphone array, was utilized. Based on the intelligibility reports from five unsuspecting listeners for UA-Speech, the severity levels are rated as very low, low, medium, and high. In terms of TORGO, these are according to the FDA, given by SLP

V. EXPERIMENTAL DESIGN

- MFCC AND CQCC

Speech intelligibility is influenced by coordination, and MFCCs can record abnormal motions or the absence or alterations [4]. CQCCs are produced by coupling between the conventional cepstral analysis and CQT. With these insights, we carry out the first experiment (E1), in which MFCCs [25] and CQCCs are used as features and the fundamental deep learning techniques—DNN, CNN, GRU, and LSTM—are used for classification.

VI. FEATURE EXTRACTION

Analysis is done on how well they emphasize the paralinguistic elements of speech. When diagnosing dysarthria, one of the most obvious signs that doctors see that goes along with it. Both the imprecise and delay in the lip, tongue, jaw motions are explained by articulatory aspects. Variations in phonation can account for the deterioration of voice quality in dysarthric individuals with respect to stability and periodicity [28]. Thus, in this sense, phonetic characteristics pertaining to perturbation are retrieved. The aberrant variations in pitch, loudness, and length that characterize dysarthric speech the identification and diagnosis of dysarthria. These Disturbances hinder the expression of appropriate emotion and cadence in speech, and they can be measured using prosodic characteristics. This study uses DNN characteristics. A more comprehensible representation is produced by concatenating them and using dimensionality reduction. This is done using FA approach. As a result, factors are constructed from the concatenated feature set to indicate their shared variance or correlation. With this method, a succinct and satisfying explanation of the multi-variate data may be produced. It is possible to think of it as a more complex and advanced version. Consequently, the feature lowered in the work suggested.
FA is typically used in conjunction with machine learning classifiers to pick the optimal features by

eliminating duplicate representations. As a result, the classifiers would be forced to handle less complexity, which would enhance their overall performance. On the other hand, feature capture all of the therein are inherent to deep learning models. As a result, dimensionality reduction methods are not frequently applied in tandem. However, we investigate if this inclusion can lead to any improvements.

An estimated average is used to determine the number of frames over the total number of utterances in the sample (400 for UA-Speech and 180 for TORGO). The bigger utterances are clipped, and the smaller ones are zero padded. The obtained by computing the average horizontal axis. Derivatives are not employed and MFCCs are provided frame-wise to the RNN, GRU, and Long Short Term Memory. Since these networks are able to learn temporal information on their own, introducing deltas would introduce redundancy and may highlight speech qualities that aren't important. The frequency range is restricted to 100Hz - 8kHz (fs/2) and there are 48 bins in an octave. A sampling period is used for the resampling. the acoustic parameterization is carried out utilizing, together with their first two deltas, in order to extract i-vectors. Next, the auxiliary database including sound audio samples of UA-Speech is used to train the UBM ten times over using the expectation-maximization (EM) technique. After computing the i-vectors consideration. By applying the Eigen-Voice Adaption approach, the Target (dysarthric) GMM is adapted from the UBM.

$$M = m + T w$$

VII. BASELINE CLASSIFIERS

The foundational classifiers in machine learning are random forest (RF) and support vector machines (SVM). Radial basis function (RBF) and linear basis function (SVM) kernels were also supported, with the ideal regularization parameter, c, which is adjusted between 1 and 10. The best results were obtained with c = 1 for E3, c = 6 for E1 and E2, and 30% of training the data for E3. The RF were designed with (ntree) being tweaked between 10 and 150 in the validation data. The best results were achieved with ntree = 50,125, and 100 for E1, E2, and E3. Together with these classifiers, the PLDA scoring system is applied for E3.

VIII. DL CLASSIFIERS

A model is learns and modeling the high-level abstractions seen in the feature sets. In Keras, Deep Neural Network models are constructed by building dense layers. of the activation of ReLU. neuron count is intended to increase in powers of two in tandem with the model depth. The number of nodes in the first layer is the same as the product of the two closest. Since 39 MFCCs are utilized for E1, there are 32 nodes in the first layer, 64 in the second, and so on. There is a layer with a dropout value of 0.4 after the thick layers. There is softmax activation on the output layer. Every DNN undergoes training using a learning batch size of 32.

CNN consists of alternating convolution and pooling layers. Every speech frame in the front end is thirteen provide the two-dimensional use. As a result, the frame-wise feature representation's contained variabilities may be effectively used to retrieve local information. n layered are used to create CNN models. Each layer is followed by a batch-normalization layer. Similar to DNN models, grows in two. resultant flattened result is transmitted to the dense layers, where n is the number of units that decrease in powers of 2. Here, just MFCC functionalities are utilized. It has been demonstrated that recurrent neural networks (RNNs) are effective in capturing the temporal relationships for sequential tasks. One of its variants, LSTM, can adaptably capture long-range relationships by solving the standard RNNs' vanishing gradient issue. Three gates regulate the information in the network are used to accomplish this. They also contain information.

IX. RESULTS

The CNN and DNN are adjusted. Figures 1 plot the findings for the MFCC along with CQCC features, correspondingly. The highest layers of the model identify effective feature representations that perform well across datasets as it becomes more and more complex. As a result, when employing MFCCs with DNNs, an improvement in accuracy was shown up to n = 5 for both databases [25]. UAS is the term used to refer to UA-Speech when labeling the graphs. Upon beyond four layers, the overall accuracy of categorization declines. This is due to the fact that as model complexity increased, generalization ability dropped. The network overfits the training set and is unable to provide wise choice based on the test data that wasn't viewed. As can be observed in Fig. 1(a), the best results. Similar to the DNN models, accuracy decreased with additional increase. A similar tendency was noted for models that used CQCCs, but with lower accuracy ratings than those obtained with MFCCs. The graphs clearly show that there is a 20% variation in the accuracy. One reason might be because the speaker's speech pattern, in addition to their physical attributes, both demonstrate the severity of dysarthria. The variations in the monotonicity pattern displayed by patients with dysarthria

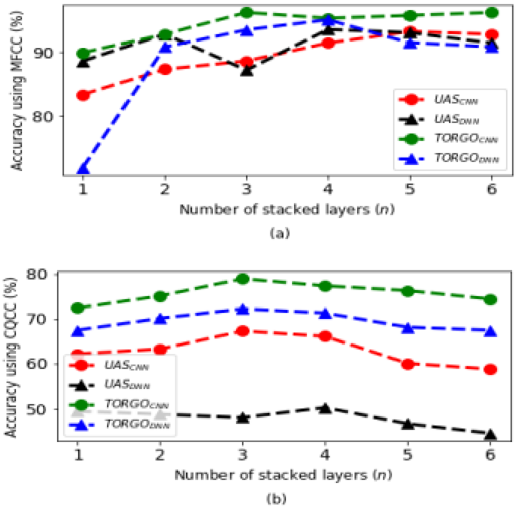

Fig 1 Variation of classification accuracy

X. DISCUSSION

SLPs' use of auditory perceptual measurements to identify the evolution of dysarthria has been automated in E1 utilizing speech perceptual characteristics such as MFCCs and CQCCs. The outcomes demonstrate that they may be utilized in conjunction with effective classifiers to offer an objective assessment of the degree of dysarthria. When utilizing MFCCs, CNN and DNN. However, GRU handled CQCCs more skillfully, suggesting the significance of their temporal dependencies. In the SD test scenario, MFCCs fared better on features than CQCCs, but CQCCs exhibit less speaker-overfitting, which suggests that they will make better SID models. When i-vectors were used, iMFCCs outperformed all other features in the SD scenario and improved classification accuracy over raw MFCCs in SID systems by about 20%. vector-PLDA paradigm provides a notable improvement over the traditional one i-vectors. When employing gain margin can be enhanced and is worth the extra processing work. The models' subpar performance in the SID scenario is indicative of their training limitation—having just a small number of subjects each class. It would be advantageous to develop UBM for the i-vector, just like with deep learning models. All pioneer efforts, however, have been conducted usingUA-Speech.

Although the speech disorder specific traits have been widely employed in the literature to distinguish between healthy and disordered speech, they have not shown to be as effective in modeling the severity levels of dysarthria. The 28-dimensional phonation characteristics came in second. It demonstrated that accuracy need not be impacted by the feature set's size, supporting the conclusions of [42]. In the automated assessment of Parkinson's disease patients, articulation traits have been shown to be more effective than the others. in the references [17]. To choose the best feature descriptors. PFE can be used to quantify the differences and similarities between the various feature sets within and between classes. This would score the characteristics according to how well they classified the severity levels and assess each feature's capacity for discrimination. This would also explain why, when all the features taken into account for the study, the DNN classifiers are unable to produce good results.

CONCLUSION

classifying the severity of speech disorder using distinct acoustic cues. Also, we have implemented DNNs for a advanced level feature analysing. Out of all the characteristics that have been studied, MFCCs have the min. computational complexity. Nonetheless, the DNN-iMFCC structure must be applied if accuracy is the main consideration. Investigation of the application of ETEO in future research to distinguish between the various degrees of dysarthria severity is needed.

REFERENCES

[1] F. Rudzicz, "Articulatory knowledge in the recognition of dysarthric speech," IEEE Trans. Audio, Speech, Language Process., vol. 19, no. 4, pp. 947–960, May 2021.

[2] H. M. Chandrashekar, V. Karjigi, and N. Sreedevi, "Spectro-temporal representation of speech for intelligibility assessment of dysarthria," IEEE J. Sel. Topics Signal Process., vol. 14, no. 2, pp. 390–399, Feb. 2020.

[3] M. J. Kim, J. Yoo, and H. Kim, "Dysarthric speech recognition using dysarthria severity-dependent and speaker-adaptive models," in Proc. Interspeech, 2020, pp. 3622–3626.

[4] J. I. Godino-Llorente, P. Gomez-Vilda, and M. Blanco-Velasco, "Dimensionality reduction of a pathological voice quality assessment system based on Gaussian mixture models and short-term cepstral parameters," IEEE Trans. Biomed. Eng., vol. 53, no. 10, pp. 1943–1953, Oct. 2019.

[5] C. Bhat and H. Strik, "Automatic assessment of sentence-level dysarthria intelligibility using BLSTM," IEEE J. Sel. Topics Signal Process., vol. 14, no. 2, pp. 322–330, Feb. 2020.

[6] G. Vyas, M. K. Dutta, J. Prinosil, and P. Harár, "An automatic diagnosis and assessment of dysarthric speech using speech disorder specific prosodic features," in Proc. IEEE 39th Int. Conf. Telecommun. Signal Process., vol. 2019, pp. 515–518.

[7] A. Farhadipour, H. Veisi, M. Asgari, and M. A. Keyvanrad, "Dysarthric speaker identification with different degrees of dysarthria severity using deep belief networks," ETRI J., vol. 40, no. 5, pp. 643–652, Oct. 2019.

[8] N. P. Narendra and P. Alku, "Dysarthric speech classification using glottal features computed from non-words, words and sentences," in Proc. Interspeech, Sep. 2018, pp. 3403–3407.

[9] M. S. Paja and T. H. Falk, "Automated dysarthria severity classification for improved objective intelligibility assessment of spastic dysarthric speech," in Proc. 13th Annu. Conf. Int. Speech Commun. Assoc., 2020, pp. 1–4.

[10] M. Todisco, H. Delgado, and N. W. Evans, "A new feature for automatic speaker verification anti-spoofing: Constant Q cepstral coefficients," in Odyssey. Bilbao, Spain, 2016, pp. 283–290.

[11] C. H M, V. Karjigi, and N. Sreedevi, "Investigation of different timefrequency representations for intelligibility assessment of dysarthric speech," IEEE Trans. Neural Syst. Rehabil. Eng., vol. 28, no. 12, pp. 2880–2889, Dec. 2020.

[12] K. Gurugubelli and A. K. Vuppala, "Perceptually enhanced single frequency filtering for dysarthric speech detection and intelligibility assessment," in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), May 2019, pp. 3403–3407.

[13] H. M. Chandrashekar, V. Karjigi, and N. Sreedevi, "Breathiness indices for classification of dysarthria based on type and speech intelligibility," in Proc. Int. Conf. Wireless Commun. Signal Process. Netw. (WiSPNET), Mar. 2019, pp. 266–270.

[14] C. Bhat, B. Vachhani, and S. K. Kopparapu, "Automatic assessment of dysarthria severity level using audio descriptors," in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Mar. 2017, pp. 5070–5074.

[15] K. Kadi, S. Selouani, B. Boudraa, and M. Boudraa, "Discriminative prosodic features to assess the dysarthria severity levels," in Proc. World Congr. Engg., vol. 3, 2019, pp. 1–5.

[16] E. A. Belalcázar-Bolanos, J. R. Orozco-Arroyave, J. F. Vargas-Bonilla, T. Haderlein, and E. Nöth, "Glottal flow patterns analyses for parkinson's disease detection: Acoustic and nonlinear approaches," in Proc. Int. Conf. Text, Speech, Dialogue. Cham, Switzerland: Springer, 2020, pp. 400–407.

[17] J. C. Vásquez-Correa, J. R. Orozco-Arroyave, T. Bocklet, and E. Nöth, "Towards an automatic evaluation of the dysarthria level of patients with Parkinson's disease," J. Commun. Disorders, vol. 76, pp. 21–36, Nov. 2020.

[18] J. R. Orozco-Arroyave et al., "NeuroSpeech: An open-source software for Parkinson's speech

analysis," Digit. Signal Process., vol. 77, pp. 207–221, Jun. 2021.

[19] P. Verma and P. K. Das, "I-vectors in speech processing applications: A survey," Int. J. Speech Technol., vol. 18, no. 4, pp. 529–546.

[20] D. Martínez, E. Lleida, P. Green, H. Christensen, A. Ortega, and A. Miguel, "Intelligibility assessment and speech recognizer word accuracy rate prediction for dysarthric speakers in a factor analysis subspace," ACM Trans. Accessible Comput., vol. 6, no. 3, pp. 1–21.

[21] C. Espana-Bonet and J. A. Fonollosa, "Automatic speech recognition with deep neural networks for impaired speech," in Proc. 3rd Int. Conf. Adv. Speech Lang. Technol. Iberian Lang., 2016, pp. 97–107.

[22] S. Gupta et al., "Residual neural network precisely quantifies dysarthria severity-level based on short-duration speech segments," Neural Netw., vol. 139, pp. 105–117, Jul. 2021.

[23] M. Perez et al., "Classification of Huntington disease using acoustic and lexical features," in Proc. Interspeech, 2018, p. 1898.

[24] A. Tripathi, S. Bhosale, and S. K. Kopparapu, "Improved speaker independent dysarthria intelligibility classification using deepspeech posteriors," in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), May 2020, pp. 6114–6118.

[25] A. A. Joshy and R. Rajan, "Automated dysarthria severity classification using deep learning frameworks," in Proc. 28th Eur. Signal Process. Conf. (EUSIPCO), Jan. 2021, pp. 116–120.

[26] F. Rudzicz, A. K. Namasivayam, and T. Wolff, "The TORGO database of acoustic and articulatory speech from speakers with dysarthria," Lang. Resour. Eval., vol. 46, no. 4, pp. 523–541, Dec. 2022.

[27] H. Kim et al., "Dysarthric speech database for universal access research," in Proc. Interspeech, 2020, pp. 1741–1744.

[28] T. Arias-Vergara, J. C. Vásquez-Correa, and J. R. Orozco-Arroyave, "Parkinson's disease and aging: Analysis of their effect in phonation and articulation of speech," Cognit. Comput., vol. 9, no. 6, pp. 731–748, Dec. 2019. JOSHY AND RAJAN: AUTOMATED DYSARTHRIA SEVERITY CLASSIFICATION 1157

[29] R. A. Reyment and K. Jvreskog, Applied Factor Analysis in the Natural Sciences. Cambridge, U.K.: Cambridge Univ. Press, 2016.

[30] N. Dehak, P. J. Kenny, R. Dehak, D. Pierre, and O. Pierre, "Front-end factor analysis for speaker verification," IEEE Trans. Audio, Speech, Language Process., vol. 19, no. 4, pp. 788–798, May 2021.

[31] A. Kanagasundaram, R. Vogt, D. Dean, S. Sridharan, and M. Mason, "I-vector based speaker recognition on short utterances," in Proc. Interspeech, 2021, pp. 2341–2344.

[32] A. Lozano-Diez et al., "Analysis and optimization of bottleneck features for speaker recognition," in Odyssey. Bilbao, Spain, 2019, pp. 352–357.