A Novel Multi-task Learning based Automatic Speech Impairment Assessment Algorithm

Yu Ge, Tianlei Wang, Jiuwen Cao†, *Senior Member, IEEE, Siyu Xu*
Machine Learning and I-health International Cooperation Base of Zhejiang Province,
Artificial Intelligence Institute, Hangzhou Dianzi University, Zhejiang, China
Email: geyu9805@163.com, tianlei.wang.cn@gmail.com, jwcao@hdu.edu.cn, 2524686112@qq.com

2022 China Automation Congress (CAC) | 978-1-6654-6533-5/22/$31.00 ©2022 IEEE | DOI: 10.1109/CAC57257.2022.10054791

***Abstract*—Speech impairment assessment is crucial to the treatment evaluation of speech therapy. The current evaluation methods mainly depend on speech-language pathologists (SLPs). Automatic speech impairment assessment (ASIA), especially the regression of severity scores, has not received enough attention. So we present a novel ASIA algorithm which based on the multi-task leaning with joint severity level classification and score regression. Owing to the auxiliary classification task, the precision of the severity score prediction can be improved effectively. In addition, the residual network (ResNet) and the long short-term memory (LSTM) are cascaded as the backbone. The performance of the model is demonstrated on the Mandarin AphasiaBank dataset and the experiments show that the algorithm achieves promising performance.**

***Index Terms*—Speech impairment assessment, Multi-task learning, Severity score regression, Severity level classification.**

I. Introduction

Speech impairment is a dysfunction in specific brain regions usually caused by stroke or other neurological diseases. The common types of speech impairment include aphasia, dysarthria, stuttering, etc. The symptom of speech impairment is that the vocal tract, tongue, mouth and jaw can't cooperate with each other to produce comprehensible words, which makes it hard for people to communicate properly [1]. Typical speech impairment is always treated by speech-language pathologists (SLPs). Helm-Estabrooks *et al.* [2] showed that the treatment must reach a certain intensity and frequency to ensure effectiveness and thus a precise speech impairment assessment (SIA) is crucial to the treatment evaluation of speech therapy [3]. But the current evaluation methods mainly depend on the experience and abilities of SLPs [4], which lacks a quantitative descriptions and indicators.

Therefore, some researchers turn their attention to the automatic speech impairment assessment (ASIA). The ASIA can be seen as a pattern recognition problem replying on feature engineering. Tsanas *et al.* [5] explored the relation between speech impairment and Parkinson's disease, where 132 dysphonia feasures were extracted along with 4 different feature selection algorithms for dimension reduction. Marina *et al.* [6] extracted varieties of prosodic, acoustic and conversational features and fed features into deep neural network for autism severity score assessment. The autocorrelation and entropy features from different frequency bands were extracted for automatic pathology speech detection and classification in [7]. Huang *et al.* [8] used cross-correlation matrix, Gaussian distribution and linear subspace for speech representation learning and an ensemble classifier was constructed for severity assessment.

Recently, the deep neural networks based ASIAs attracted much attention due to the exceptional successes of deep learning. Specially, the combinations of convolutional neural network (CNN) and speech spectrogram were researched widely [9]–[11]. The gate recurrent unit (GRU) and CNN model were established to assess pathological speech, respectively [12]. Gupta *et al.* [13] used the improved ResNet to precisely quantify dysarthria severity level. Duc Le *et al.* [14] designed a feature set through an acoustic model to capture pronunciation, rhythm and intonation features. Qin *et al.* [15] decoded the patients' audios through an automatic speech recognition (ASR) model to extract text and acoustic features, which can comprehensively reflect impairments in vocabulary, grammar, pronunciation, fluency, etc.

However, the aforementioned ASIA methods focused on classifying the severity level and little attention has been paid to severity scores regression. The severity scores can quantify the severity better and provide a rationale for subsequent targeted rehabilitation. Thus, a regression task for evaluation of severity scores is first built in this paper. At the meantime, the classification of severity levels is still reserved resulting multi-task learning framework for speech impairment. The classification task provides additional constraints on the model and thus a better capability of representation learning can be achieved. In addition, a novel backbone consisting of the cascaded ResNet and LSTM are constructed. The ResNet is first trained by the multitasking loss function and the outputs of fully connected (FC) layer of ResNet are then fed to LSTM for training and prediction. The effectiveness is verified on the Mandarin AphasiaBank dataset and the comparisons with other methods are provided.

II. The proposed ASIA Algorithm

Fig. 1 presents the flowchart of the ASIA system where the network with the cascaded ResNet and LSTM is constructed as the backbone. The ResNet can extract frequency and energy variation features from Mel-frequency spectrograms (MFSs) and the LSTM is used to learn the incoherent and ambiguity

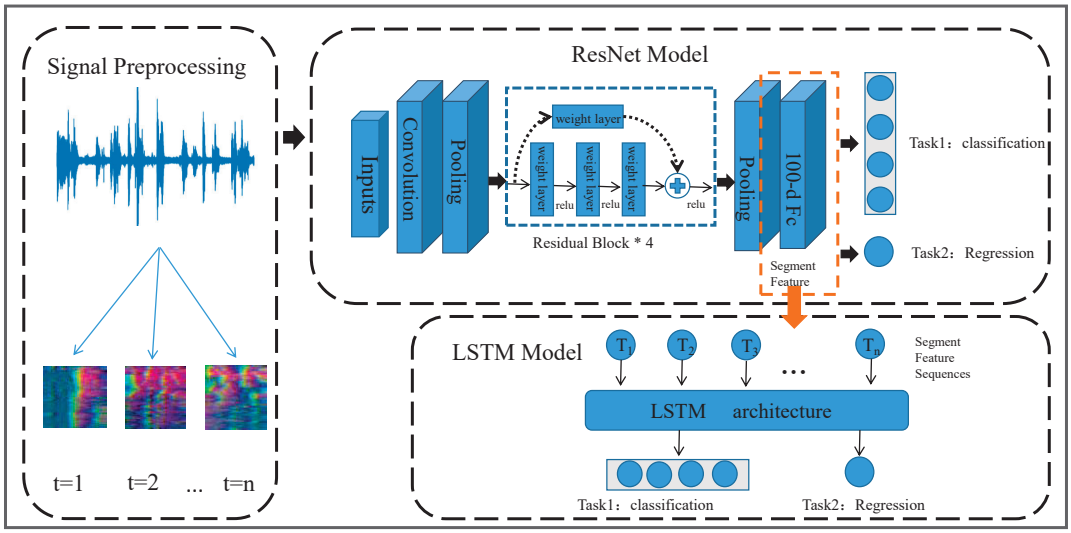


Fig. 1. The flowchart of speech impairment assessment algorithm

in context [16]. The flowchart shown in Fig. 1 can be divided into: 1) Signal preprocessing, which transforms the time-domain signals to three-channel MFSs and then splits them into several sequence blocks along with timeline; 2) Training the ResNet based on multi-task loss function for feature extraction from the sequence blocks; 3) Constructing LSTM to capture the temporal structure and obtain the severity score and level. More details are given in the following sections.

A. Dataset

The Mandarin AphasiaBank[1] collected by the Rehabilitation Medicine Center of the First Affiliated Hospital of Nanjing Medical University [17] is adopted as the data in this paper. The dataset is an original acquisition of videos of stroke patients, where the background is in a quiet and bright room. The stroke patients are guided by professional speech therapists to complete the speech assessment tasks that includes five parts: greetings, picture speaking, story setting, process telling and free speaking. All patients are evaluated comprehensively using the Western Aphasia Battery (WAB) [18] including many sub-tests that measure expressive fluency and naming ability [19]. The Aphasia Quotient (AQ) [20] represents the severity score of speech impairment.

In this paper, the audio data is extracted from the videos and the four-second audio is chosen as a sample. Specially, the audios are divided into four severity levels based on the AQ scores to construct the auxiliary classification task. Table I gives the specification of the constructed dataset.

B. Signal preprocessing

The MFS is employed to transform the acoustic signals to images. The four-second audio is first enframed with 25ms-length Hamming window and frame shift of 10ms. The short-time amplitude spectrum $X^m(k)$ of the m-th frame signal is then derived using the discrete Fourier transform (DFT). Next the l-th Mel-frequency per frame is obtained by

$$s^m(l) = \ln(\sum_{k=0}^{N-1} |X^m(k)|^2 H_l(k)). \quad (1)$$

Here, $l = 1, \cdots, L$, which means the quantity of Mel-filters and $m = 1, \cdots, M$, which represents the amount of frames, N means the length of each frame, and $H_l(k)$ means the output of the l-th Mel-filter.

Fig. 2 shows the MFSs of four different severity levels, where the MFSs is 400×64 representing 400 consecutive frames and 64 Mel-frequencies. Voice signals with different severity levels are distributed in different color regions. The increase of severity levels leads to the lower energies when the speaker is difficult to pronounce.

TABLE I
MANDARIN APHASIABANK DATASET

Severity level	Patient	Gender	Age	AQ	Number of data
Mild	JL01b	M	48	91	145
	JL01c	M	48	87	148
	JL05a	F	27	91	139
	JL07a	M	40	92	241
	LN04b	M	51	94	66
	LN09	F	67	88	41
	LN10	M	51	87	136
	LN12	F	34	93	93
Moderate	JL01a	M	48	78	182
	JL06a	F	23	80	55
	JL10a	M	54	81	97
Serious	JL03a	M	54	68	175
	JL08a	F	41	74	55
	JL09a	M	40	71	75
	LN04a	M	51	75	71
Critical	JL11a	M	65	61	33
	Jn03a	M	52	54	124
	LN11	F	63	63	137
	LN17	M	61	55	68

[1]https://aphasia.talkbank.org/access/Mandarin/Aphasia/JiangLin.html

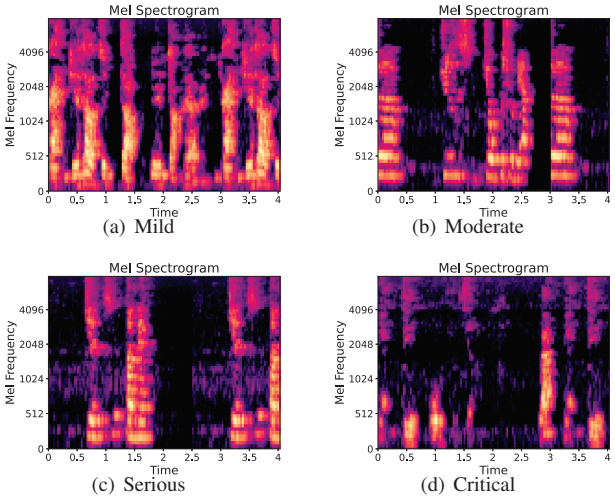


Fig. 2. MFSs of different severity levels.

Considering the input restriction of ResNet and LSTM, the MFSs are then segmented into several sections using 64×64 rectangular window with horizontal shift of 30 frames, as shown in Fig. 3. Subsequently, the deltas and delta-deltas features [21] are further derived on the basis of (1). The three-channel MFS is generated by overlapping consecutive frames of Mel-frequency bands, deltas and delta-deltas.

C. ResNet and LSTM

The ResNet50 [22] combining with the multi-task loss function is trained for feature extraction. First, the last fully connected (FC) layer with 1000 neurons is reduced to 100 neurons. Then a new FC named "Fc2" is added as the output layer that includes one neuron for severity score regression and four neurons for severity level classification. The final network structure is shown in Table II.

TABLE II
RESNET50 NETWORK STRUCTURE

Layers	ResNet50	Output Size
conv1	7×7, 64, stride 2	112×112
conv2	3×3 max pooling, stride 2 $\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} \times 3$	56×56
conv3	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{bmatrix} \times 4$	28×28
conv4	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{bmatrix} \times 6$	14×14
conv5	$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{bmatrix} \times 3$	7×7
Fc1	average pool, 100-d fc	100×1
Fc2	4-d fc, softmax and 1-d fc	1×1

The ResNet50 is initialized with pre-trained weights on ImageNet and then the multi-task loss function is constructed to tune the pre-trained ResNet50. At last, the outputs of Fc1 are taken as speech segment-level features and fed into the LSTM for training and prediction. Fig. 4 shows the extracted features of different layers of the trained ResNet50. It can be seen that the energetic parts of the MFSs can be highlighted as the layer deepen.

The features extracted by ResNet is then fed to the LSTM for prediction. A three-layer LSTM network with 64 neurons in each layer is built. Similar to ResNet, the output layer of LSTM consists of one neuron for severity score regression and four neurons for severity level classification. The Multitasking loss function is also employed to train the LSTM.

D. Multitasking loss function

The multitasking loss function is the key component of the proposed ASIA system. It consists of the mean square error (MSE) based regression term and the cross entropy (CE) based classification term, i.e.,

$$\text{Loss} = \alpha \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}_i)^2 + \beta[-\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m} y_{ij} \log(\bar{y}_{ij})] \quad (2)$$

where α and β are the trade-off parameters between the regression and classification, n means the quantity of samples, m means the quantity of severity levels. x and $\bar{x}$ represent the label value and predicted value corresponding to the regression task, y and $\bar{y}$ represent the true value and predicted value corresponding to the classification task.

III. EXPERIMENTS AND DISCUSSIONS

A. Experimental setup

This part, the effectiveness of the ASIA system is verified by experiments. We use 90% of the Mandarin AphasiaBank for training and 10% for testing. The ResNet and LSTM are trained by the adaptive moment estimation (ADAM) with the learning rate 0.0001. All batch sizes in the network are 64, while the epoch of ResNet is 80 and the epoch of LSTM is set to 300. The R-Squared, Root Mean Squard Error (RMSE) and Mean Absolute Error (MAE) are the performance metric for score regression, where the definitions are expressed as follows.

$$R - Squared = 1 - \frac{\frac{1}{n} \sum_{i=1}^{n} (\bar{y}_i - y_i)^2}{\frac{1}{n} \sum_{i=1}^{n} (y_i - y_{mean})^2}, \quad (3)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\bar{y}_i - y_i)^2}, \quad (4)$$

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |(\bar{y}_i - y_i)|, \quad (5)$$

where n means the amount of samples, y_i means the labeled values of samples, $\bar{y}_i$ is the model predicted values of samples, y_{mean} represents the mean of the labeled values. For severity

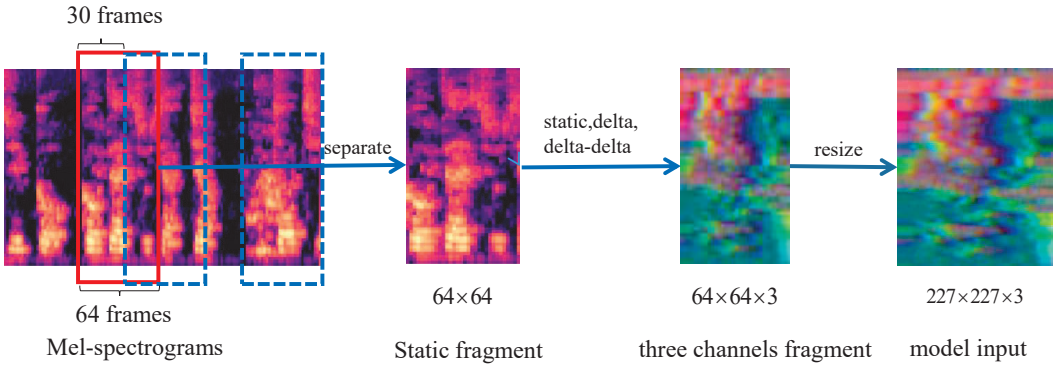


Fig. 3. Segmentation processing of MFS.

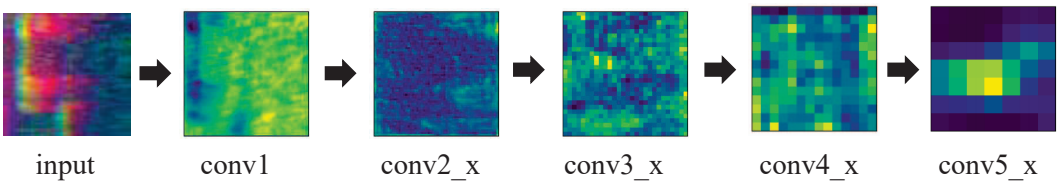


Fig. 4. Feature visualizations of different layers of ResNet50

level classification, the F1 Score(F1), Recall(Rec) and Precision(Prec) are adopted and the definitions are given as follows.

$$Prec = \frac{TP}{FP + TP} \tag{6}$$

$$Rec = \frac{TP}{TP + FN} \tag{7}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{8}$$

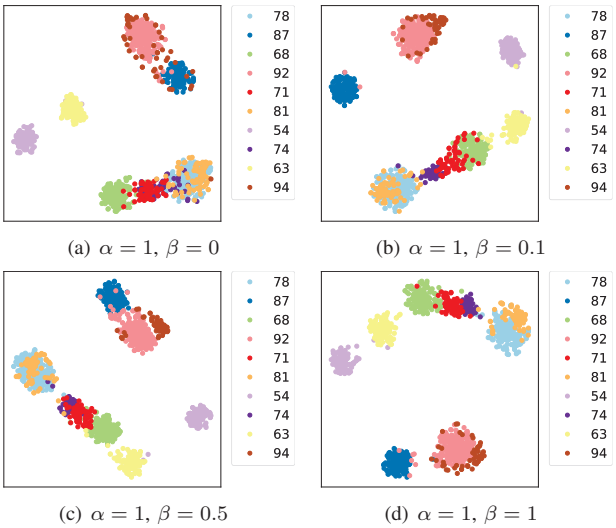
B. Selection of Hyper Parameters

As shown in (2), the settings of α and β are two crucial parameters in model training. Fig. 5 plots the TSNE visualization of feature representation by ResNet, where the samples belonging to 10 different AQ groups are selected for visualization. α is fixed to be 1 and β is changed in $\{0, 0.1, 0.5, 1\}$. It can be seen that the features extracted by ResNet have the better discrimination when $\alpha = 1$, $\beta = 0.5$.

Fig. 6 shows the regression performance of LSTM with different β values ($\alpha = 1$). Similarly, the best performance is obtained when $\alpha = 1$, $\beta = 0.5$.

C. Experiment comparisons on different MFSs

The three-channel MFS using the first- and second-order difference coefficients can capture more dynamic information than single-channel MFS. Table III and Table IV give the comparisons between single-channel MFS and three-channel MFS using the proposed ASIA method. Specially, Table III only uses score regression to train ResNet50 and LSTM but Table IV uses the multi-task loss function. By comparison, the three-channel MFS achieves better performance in 3 different



(a) $\alpha = 1$, $\beta = 0$ (b) $\alpha = 1$, $\beta = 0.1$

(c) $\alpha = 1$, $\beta = 0.5$ (d) $\alpha = 1$, $\beta = 1$

Fig. 5. TSNE visualization of feature representations by ResNet.

metrics. In addition, the adopted multi-task loss function in Table IV also performs better than only using regression loss function in Table III.

TABLE III
COMPARISONS OF DIFFERENT MFSS WITH SINGLE TASK.

Spectrograms	R-square	RMSE	MAE
Single-channel MFS	0.77	4.18	3.22
Three-channel MFS	0.83	3.74	2.86

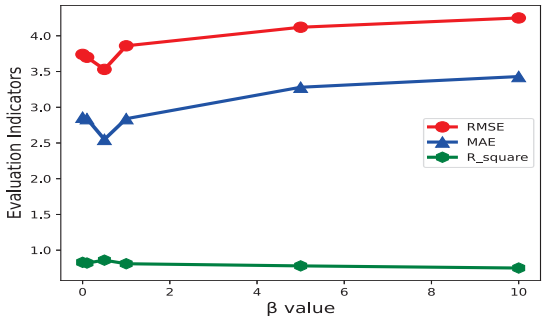


Fig. 6. Performance evaluation of LSTM with different β values

TABLE IV
COMPARISONS OF DIFFERENT MFSS WITH MULTI-TASK LEARNING.

Spectrograms	R-square	RMSE	MAE
Single-channel MFS	0.83	3.87	2.91
Three-channel MFS	0.86	3.53	2.55

D. Comparison study

Comparing Table III and Table IV, the auxiliary classification task can improve the performance of the severity score prediction effectively. Fig. 7 further shows the effectiveness of the multi-task loss function and our proposed method restrains most outliers but the single regression task fails. The reasons behind the results are that the categories of the classification task are obtained according to the range of score intervals, which enable the regression task to learn the features of specific score intervals in a targeted manner. Table V gives the results of severity level classification and our proposed method also obtains the high classification accuracy.

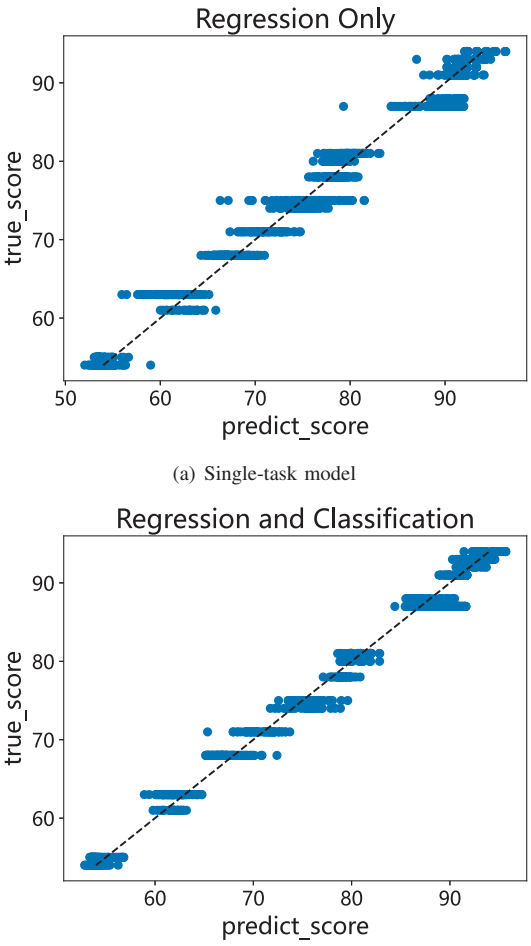
TABLE V
CLASSIFICATION ACCURACY OF THE PROPOSED ASIA METHOD (%).

	Precision	Recall	F1 Score
Mild	0.998	1.0	0.999
Moderate	0.87	0.993	0.927
Serious	0.99	0.661	0.793
Critical	0.821	1.0	0.902

Table VI provides the comparisons of our method with the state-of-the-art (SOTA) methods. It shows that our method obtains the best performance. Specially, our method improves 0.11 on R-square, reduces 1.03 on RMSE and 1.17 on MAE over ResNet [13].

IV. CONCLUSIONS

A novel multi-task learning based automatic speech impairment assessment system (ASIA) has been developed. The three-channel Mel-frequency spectrogram (MFS) has been adopted and then split to several sequence blocks as inputs. The ResNet50 and LSTM were cascaded as the backbone, and the multitasking loss function are are carried out to train them.



(a) Single-task model

(b) Multi-task model

Fig. 7. Scatter plots of the true and the predicted scores.

TABLE VI
COMPARISONS WITH THE SOTA METHODS

Method	R-square	RMSE	MAE
ResNet [13]	0.75	4.56	3.72
ResNet+Attention [23]	0.76	4.22	3.38
Feature fusion [24]	0.84	3.96	2.87
Ours	0.86	3.53	2.55

The outputs of penultimate FC layer of the ResNet50 were fed to LSTM for training and prediction. The validity of the ASIA algorithm was proved in Madarin AphasiaBank dataset and its performance was compared with those of several representative methods. Experiments show the superiority of the proposed ASIA system.

V. ACKNOWLEDGMENT

Our work is supported by the General Research Projects of Education Department of Zhejiang Province (Y202146999), the Fundamental Research Funds for the Provincial Universities of Zhejiang (GK219909299001411,

GK219909299001017), the Zhejiang Provincial Natural Science Foundation (LQ22F030006, LZ22F030002), the Open Research Projects of Zhejiang Lab (2021MC0AB04).

REFERENCES

[1] David Frank Benson and Alfredo Ardila. *Aphasia: A clinical perspective*. Oxford University Press on Demand, 1996.

[2] Nancy Helm-Estabrooks, Martin L Albert, and Marjorie Nicholas. *Manual of aphasia and aphasia therapy*. Pro-ed, 2014.

[3] Katherine Kelly, Steven Cumming, Anna Corry, Kerry Gilsenan, Claire Tamone, Kylie Vella, and Hans Bogaardt. The role of speech-language pathologists in palliative care: Where are we now? a review of the literature. *Progress in Palliative Care*, 24(6):315–323, 2016.

[4] Ronald Prins and Roelien Bastiaanse. Analyzing the spontaneous speech of aphasic speakers. *Aphasiology*, 2004.

[5] Athanasios Tsanas, Max A Little, Patrick E McSharry, Jennifer Spielman, and Lorraine O Ramig. Novel speech signal processing algorithms for high-accuracy classification of parkinson's disease. *IEEE Transactions on Biomedical Engineering*, 59(5):1264–1271, 2012.

[6] Marina Eni, Ilan Dinstein, Michal Ilan, Idan Menashe, Gal Meiri, and Yaniv Zigel. Estimating autism severity in young children from speech signals using a deep neural network. *IEEE Access*, 8:139489–139500, 2020.

[7] Ahmed Al-Nasheri, Ghulam Muhammad, Mansour Alsulaiman, Zulfiqar Ali, Khalid H Malki, Tamer A Mesallam, and Mohamed Farahat Ibrahim. Voice pathology detection and classification using auto-correlation and entropy features in different frequency regions. *IEEE Access*, 6:6961–6974, 2017.

[8] Dong-Yan Huang, Minghui Dong, and Haizhou Li. Combining multiple kernel models for automatic intelligibility detection of pathological speech. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6485–6489. IEEE, 2016.

[9] HM Chandrashekar, Veena Karjigi, and N Sreedevi. Spectro-temporal representation of speech for intelligibility assessment of dysarthria. *IEEE Journal of Selected Topics in Signal Processing*, 14(2):390–399, 2019.

[10] HM Chandrashekar and Veena Karjigi. Investigation of different time-frequency representations for intelligibility assessment of dysarthric speech. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 28(12):2880–2889, 2020.

[11] Laiba Zahid, Muazzam Maqsood, Mehr Yahya Durrani, Maheen Bakhtyar, Junaid Baber, Habibullah Jamal, Irfan Mehmood, and Oh-Young Song. A spectrogram-based deep feature assisted computer-aided diagnostic system for parkinson's disease. *IEEE Access*, 8:35482–35495, 2020.

[12] Ying Qin, Tan Lee, Yuzhong Wu, and Anthony Pak Hin Kong. An end-to-end approach to automatic speech assessment for people with aphasia. In *2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 66–70. IEEE, 2018.

[13] Siddhant Gupta, Ankur T Patil, Mirali Purohit, Mihir Parmar, Maitreya Patel, Hemant A Patil, and Rodrigo Capobianco Guido. Residual neural network precisely quantifies dysarthria severity-level based on short-duration speech segments. *Neural Networks*, 139:105–117, 2021.

[14] Duc Le, Keli Licata, Carol Persad, and Emily Mower Provost. Automatic assessment of speech intelligibility for individuals with aphasia. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(11):2187–2199, 2016.

[15] Ying Qin, Tan Lee, and Anthony Pak Hin Kong. Automatic assessment of speech impairment in cantonese-speaking people with aphasia. *IEEE Journal of Selected Topics in Signal Processing*, 14(2):331–345, 2019.

[16] Zhaoci Liu, Zhiqiang Guo, Zhenhua Ling, and Yunxia Li. Detecting alzheimer's disease from speech using neural networks with bottleneck features and data augmentation. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7323–7327. IEEE, 2021.

[17] Brian MacWhinney, Davida Fromm, Margaret Forbes, and Audrey Holland. Aphasiabank: Methods for studying discourse. *Aphasiology*, 25(11):1286–1307, 2011.

[18] Cynthia M Shewan and Andrew Kertesz. Reliability and validity characteristics of the western aphasia battery (wab). *Journal of Speech and Hearing Disorders*, 45(3):308–324, 1980.

[19] Jessica D Richardson, Sarah Grace Hudspeth Dalton, Davida Fromm, Margaret Forbes, Audrey Holland, and Brian MacWhinney. The relationship between confrontation naming and story gist production in aphasia. *American Journal of Speech-Language Pathology*, 27(1S):406–422, 2018.

[20] Andrew Kertesz and Elizabeth Poole. The aphasia quotient: the taxonomic approach to measurement of aphasic disability. *Canadian Journal of Neurological Sciences*, 31(2):175–184, 2004.

[21] Ziping Zhao, Qifei Li, Zixing Zhang, Nicholas Cummins, Haishuai Wang, Jianhua Tao, and Björn W Schuller. Combining a parallel 2d cnn with a self-attention dilated residual network for ctc-based discrete speech emotion recognition. *Neural Networks*, 141:52–60, 2021.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[23] Huijun Ding, Zixiong Gu, Peng Dai, Zhou Zhou, Lu Wang, and Xiaoxiao Wu. Deep connected attention (dca) resnet for robust voice pathology detection and classification. *Biomedical Signal Processing and Control*, 70:102973, 2021.

[24] Lang He and Cui Cao. Automated depression analysis using convolutional neural networks from speech. *Journal of Biomedical Informatics*, 83:103–111, 2018.