INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

# Whisper Features for Dysarthric Severity-Level Classification

*Siddharth Rathod*[1], *Monil Charola*[1], *Akshat Vora*[1], *Yash Jogi*[2], *Hemant A. Patil*[1]

[1]Speech Research Lab, Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), Gandhinagar-382007, Gujarat, India
[2]Sprinklr, India

siddharth_rathod@daiict.ac.in, monil_charola@daiict.ac.in, akshat_vora@daiict.ac.in, 201601202@daiict.ac.in, hemant_patil@daiict.ac.in

## Abstract

Dysarthria is a speech disorder caused by improper coordination between the brain and the muscles that produce intelligible speech. Accurately diagnosing the severity of dysarthria is critical for determining the appropriate treatment and directing speech to suitable Automatic Speech Recognition systems. Recently, various methods have been employed to investigate the classification of dysarthria severity-levels using advanced features, including STFT and MFCC. This study proposes utilizing Web-scale Supervised Pretraining for Speech Recognition (WSPSR), also known as Whisper, encoder module for dysarthric severity-level classification using transfer learning approach. Whisper model is an advanced machine learning model used for speech recognition, which is trained on a large scale of 680,000 hours of labeled audio data. The proposed approach demonstrated a high accuracy rate of $98.02\%$, surpassing the accuracies achieved by MFCC ($95.2\%$) and LFCC ($96.05\%$).

**Index Terms**: Dysarthria, Encoder-Decoder Transformer, WSPSR (Whisper).

## 1. Introduction

Dysarthria is a common speech disorder that can affect the dynamic movements of the articulators, responsible for the production of intelligible speech, and upper respiratory system, resulting in difficulty producing natural speech. This disorder can occur as a result of various neurological conditions, such as cerebral palsy, muscular dystrophy, stroke, brain infection, brain injury, facial paralysis, tongue or throat muscular weakness, and nervous system disorders. These conditions cause an imbalance in coordination between the brain and the muscles involved in speech production, essential for speech production mechanism, leading to a range of speech disorders, including dysarthria, stuttering, apraxia, and dysprosody [1].

Accurate classification of dysarthric severity-level has important clinical applications. Personalised treatment plans can be developed by healthcare professionals based on the severity-level, ultimately leading to improved outcomes for individuals suffering from dysarthria. Moreover, such systems may be able to detect dysarthria on an early stage providing an alternative to diagnosis in localities, where healthcare services are unavailable. In addition, the severity-level classification has applications in Automatic Speech Recognition (ASR) systems, diverting a speech signal based on dysarthric severity-level to an appropriate ASR system [2].

In the recent past, researchers have extensively utilized the Short-Time Fourier Transform (STFT) [3] and several other acoustical parameters to classify the severity-levels of dysarthria [4]. To capture the global spectral envelope information of speech signals, state-of-the-art feature sets, such as Mel Frequency Cepstral Coefficients (MFCC) have been commonly used [5]. Additionally, glottal excitation source parameters from quasi-periodic sampling of the vocal tract system have also been used [6]. These feature sets have been selected because they are known to capture perceptual information and effective in characterizing dysarthric speech. Recently, transfer learning approaches have been explored for the problem. In particular, Bidirectional Long-Short Term Memory (BLSTM) have been used to classify dysarthric speech into intelligible (I) and non-intelligible (NI) [7]. In [8], ResNet-50 model pretrained on the ImageNet dataset is used for a transfer learning approach on a CNN classifier to classify speech into two classes as done in [7]. However, pretrained ResNet-50 doesn't encompass the sequential information as it treats the audio signal as an image, moreover, the training is done using ImageNet, which is a visual dataset. This study proposes a transfer learning approach making use of the pretrained **W**eb-scale **S**upervised **P**retraining for **S**peech **R**ecognition [9], also referred to as **Whisper**, for dysarthric severity-level classification. Whisper model is historically trained for the purpose of speech recognition.

The whisper model is trained by scaling weakly supervised audio paired with its transcripts scraped from the Internet. As a result, the dataset produced is extremely diversified, encompassing a wide range of sounds from several different environments, recording setups, speakers, and languages, making it a suitable model for speech applications using transfer learning. This study proposes using the transformer encoder of the Whisper model, which is pretrained on a large speech dataset, to classify dysarthric speech from UA Speech Corpus, into four classes based on severity. The contributions of this paper are as follows:

- Proposes end-to-end pretrained Whisper transformer encoder, using a transfer learning approach to classify dysarthria into four classes of severity.
- Experimental performance evaluation on three whisper architectures, namely, tiny, base, and small.
- Accurate diagnosis of the severity-level is crucial in determining the course of treatment for dysarthria, and it is necessary to be able to achieve this even for shorter durations of speech. Therefore, our study includes an analysis of latency periods, and a comparison with state-of-the-art feature sets.
- Since research on dysarthria requires high performance, it is crucial to evaluate the precision of model retraining. This study reports the experiments for this purpose.


10.21437/Interspeech.2023-1891

# 2. Proposed Work

## 2.1. Introduction to Whisper Model

Whisper is pre-trained on a massive amount of labelled audio-transcription data, in contrast to many of its predecessors, such as *wav2vec 2.0* [10], which is pre-trained on unlabelled audio data. Whisper is an open source pre-trained automated speech recognition (ASR) model released in September 2022, at `https://github.com/openai/whisper`. Whisper is derived from the acronym **WSPSR**, which stands for **W**eb-scale **S**upervised **P**retraining for **S**peech **R**ecognition [9]. Whisper essentially highlights the fact that training on a substantial and varied supervised dataset and focusing zero-shot transfer significantly improves the endurance and performance of the system.

The Whisper model is an encoder-decoder Transformer architecture similar to that described in [11]. It has been trained to perform various tasks, such as transcription, voice activity detection, alignment, translation, and language identification on audio samples. The input audio is broken into $30seconds$ segments, and if necessary, padded before being resampled at a frequency of 16 kHz. A Log-Mel Spectrogram with 80 channels is then computed using a window length of $25ms$ and a stride of $10ms$, as outlined in [9].

## 2.2. Whisper models

Whisper has five models, each having increasing model size, namely, tiny, base, small, medium, and large. Whisper models of different number of trainable parameters and number of transformer encoder-decoder layers are shown in Table 1. Whisper encoder features are fixed dimensional vectors obtained at the end of the encoder module of the size $1 \times 1500 \times 384$, $1 \times 1500 \times 512$, $1 \times 1500 \times 768$, $1 \times 1500 \times 1024$, and $1 \times 1500 \times 1280$ for tiny, base, small, medium, and large model, respectively. The size of vectors obtained increases as the size of the whisper model increases.

The second dimension of the fixed vector remains the same for all models as it encompasses the temporal values for the input audio signal. In Section 4, we have analysed the effect of Whisper model size on the performance of our system using tiny, base, and small models.

Table 1: *Whisper Models. After [9]*

| Whisper Model | Layers | Width | Heads | Parameters |
|---|---|---|---|---|
| Tiny | 4 | 384 | 6 | 39M |
| Base | 6 | 512 | 8 | 74M |
| Small | 12 | 768 | 12 | 244M |
| Medium | 24 | 1024 | 16 | 769M |
| Large 1 | 32 | 1280 | 20 | 1550M |

## 2.3. Dataset used to train Whisper

The dataset created to train the whisper model used consists of 680,000 hours of audio from which 117,000 hours covers other languages, and 125,000 hours of the dataset is translation from other languages to English [9]. This results in a diversified dataset, encompassing a wide range of sounds from several different environments, recording setups, speakers, and languages. The huge volume and enormous variety in audio quality certainly helps in the training the model with high performance and robustness.

This study proposes a transfer learning approach for the classification of dysarthric severity-levels. Specifically, we hypothesize that the Transformer Encoder module of the pre-trained Whisper model captures all relevant information for this task. To test this hypothesis, we utilized the pre-trained Whisper encoder combined with a CNN acting as a classifier, utilizing the learned representations from the Whisper encoder's last layer, Whisper encoder features.

Transfer learning has proven effective in various natural language processing and speech recognition tasks. Leveraging the pre-trained Whisper encoder enables us to benefit from its ability to extract high-level features from audio data. We chose this specific approach as the variability of the dataset used for training makes the model more robust and suitable for our problem. The proposed transfer learning approach allows us to leverage the model's ability to generalize unseen data, which is essential for our task of dysarthric severity classification.

## 2.4. Transfer Learning

A machine learning approach called transfer learning uses information from a related activity to speed up learning for a new task. Transfer learning happens when a model that has already been trained is retrained using a different dataset while preserving the information learned from the original dataset by freezing some trainable hyperparameters and neurons [12].

The training pipeline of our work is shown in Fig. 1. The speech signal is preprocessed and made ready to be given as input to the Whisper encoder block. Upon this, the input is processed by two convolution layers of kernel width 3 [9]. In order to help the Whisper encoder learn the relative positions within the input speech signal, sinusoidal embeddings are applied to it [11]. The processed signal is then directed to the Whisper encoder block, which depending upon the size of Whisper model, gives a vector output of fixed dimensions in its last hidden state. This output is then taken as an input by a CNN, which classifies the speech signal into four classes of dysarthric-severity.

For training process, the weights of the Whisper encoder are kept frozen, and only the weights of CNN classifier are updated during back propagation. To further study the effect of deep neural network architecture appended at the end of the pipeline, we have performed similar experiments using a ResNet instead of a CNN in Section 4 of this work.

# 3. Experimental Setup

## 3.1. Datasets Used

The Universal Access Dysarthric Speech (UA-Speech) corpus is used in this study [13]. We have adopted our baseline from [3]. The classifier models were trained on Whisper's Encoder's output features by freezing the Encoder part of the pipeline, from each speaker's microphone arrays: $M3$, $M5$, and $M6$. Apart from this, from a total of 765 utterances, 465 were used. For training the model, 837, 837, 833, and 676 utterances belonging to each class, which constitutes $90\%$ of the total data were used. The remaining $10\%$ of the data was used for the evaluation of the trained model, which had 354 utterances. Additionally, TORGO dataset was used [14], from which a total 1982 utterances belonging to three classes of severity-level were taken. The same data distribution, i.e, $90\%$ training data and $10\%$ testing data, was employed for TORGO dataset as well.

Table 2: *Class-wise patient details.*

| | UA Corpus [13] | TORGO [14] |
|---|---|---|
| Very Low | F05, M08, M09, M10, M14 | F04, M03 |
| Low | F04, M05, M11 | F01, M05 |
| Medium | F02, M07, M16 | M01, M04 |
| High | F03, M01, M04, M12 | - |

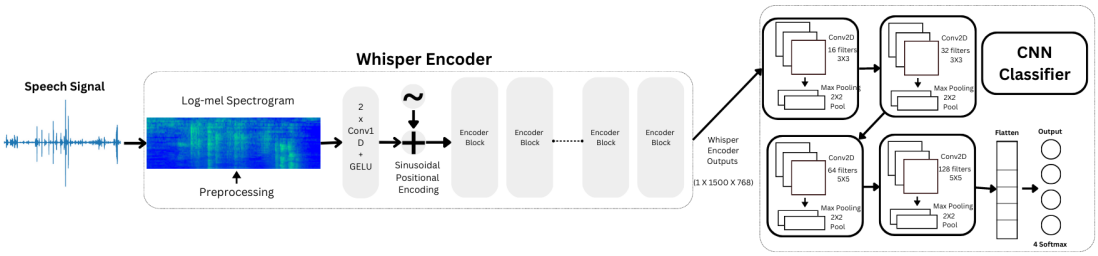


Figure 1: *Functional Block Diagram of Proposed Whisper Encoder Transfer Learning Pipeline in tandem with CNN classifier*

### 3.2. Details of the Feature Sets Used

In this study, the performance of whisper encoder-based method is compared with the state-of-the-art feature sets, such as MFCC [15], and Linear Frequency Cepstral Coefficients (LFCC) [15]. The parametric details of these features are given in Table 3. The widespread use of MFCC & LFCC in speech pathology detection in literature makes them suitable for comparison with the proposed method, hence they are taken as baseline feature sets.

Table 3: *Details of Parameters of the Various Feature Sets Used*

| Parameters | Whisper | MFCC | LFCC |
|---|---|---|---|
| **Frequency Scale** | - | Mel | Linear |
| **Subband Filter** | - | 40 | 40 |
| **Feature Dimension** | 1×1500×512 | 42 | 120 |

### 3.3. Convoluitonal Neural Network (CNN)

A Convolutional Neural Network (CNN) was employed as a classifier, given its ability to replicate the way human brain perceives images or visual features. The model consisted of four convolutional layers and one fully-connected (FC) layer, with respective convolution kernel sizes of $3 \times 3$, $3 \times 3$, $5 \times 5$, and $5 \times 5$. Rectified Linear activation units (ReLU) [16] was applied to the model, along with max-pool layers after each convolutional layer. Additionally, 2D spatial dropout layers with a probability of $0.225$ were included after each convolution layer to avoid overfitting. Stochastic gradient descent optimizer was used during the training of the CNN model [17]. The proposed model was trained for $100$ epochs, in which the loss was estimated using a categorical cross-entropy function with a learning rate of $0.01$ for the first $20$ epochs, which was later reduced to $0.003$.

### 3.4. Residual Neural Network (ResNet)

ResNet is used as another DNN classifiers to make sure that the encoder output is not biased against a given classifier. In the present study, the ResNet50, a state-of-the-art convolutional neural network (CNN) comprising 50 layers, was initially employed. However, due to the limited number of utterances in the dataset, the issue of overfitting was encountered. To mitigate this problem, we opted to downscale the model by reducing the number of layers to four. Specifically, each layer was composed of two standard ResNet blocks as opposed to the original architecture, which consisted of three, four, six, and three blocks, respectively. The use of Stochastic Gradient Descent (SGD) was made as an optimizer with the default learning rate of $0.003$.

### 3.5. Performance Evaluation

For evaluating our model's performance, we used some of the widely accepted metrics like the F1-Score [18], Jaccard's Index, which measures the similarity and dissimilarity of two classes [19], Mathew's Correlation Coefficient (MCC), which shows degree of association between the expected and the actual class [20], and Hamming Loss, which is calculated on the basis of number of samples that are inaccurately predicted [21], are used.

## 4. Experimental Results

This study involved conducting rigorous experiments to assess the resilience of the proposed approach, by using two different transfer learning architectures, one using CNN and the other using ResNet, on both the datasets. Moreover, this study entailed conducting experiments to evaluate the feasibility of implementing our approach, which involved analyzing latency periods and performing precision retraining. Additionally, we conducted a comparative analysis of the performance of our approach with that of MFCC and LFCC.

### 4.1. Effect of Size of Whisper Model

We have tested three whisper models, namely, tiny, base, and small, the specifications of which are given in Table1, on both the datasets using both the pipelines, one with CNN and the other with ResNet. Fig. 2 indicates that the testing accuracy of the data increases with an increase in the number of trainable parameters as the size of the whisper model grows, with consistent performance observed across both the classifiers. Notably, the Whisper-Small Model yielded the highest accuracy, and was thus selected for the remaining experiments presented in this paper.

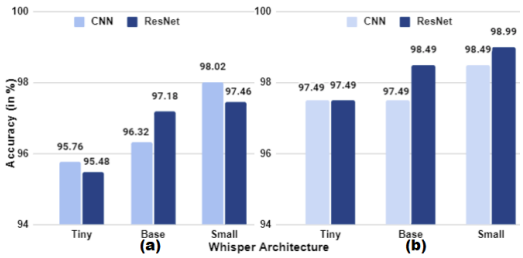


Figure 2: *Performace Comparison between different Whisper Models, a) UA Speech Corpus, b) Torgo Corpus*

### 4.2. Effect of classifiers

In this study, we employed two distinct DNN classifiers, namely, CNN [5] and ResNet [22], in order to address any potential classifier model bias during performance evaluation. The performance achieved by both the models was found to be almost identical, although ResNet exhibited a slightly better performance due to its deeper architecture.

### 4.3. Comparison with existing feature sets

The classification accuracy on the baseline MFCC, LFCC, and whisper-encoder method on CNN and ResNet are shown in the Fig. 3. Clearly, it can be discerned that our proposed method of

transfer learning outperforms the baseline MFCC and LFCC, for both the datasets and for both the DNN classifiers. The peak Testing accuracy achieved by the proposed methodology is found to be 98.02% and 98.49% on CNN classifier, and, 97.46% and 98.99% on ResNet classifier, on UA Corpus and TORGO dataset, respectively.

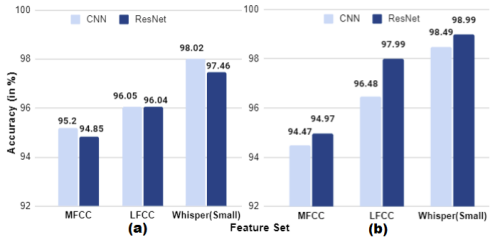


Figure 3: *Comparison of Performance with Other Feature Sets, a) UA Speech Corpus, and b) Torgo Corpus*

Table 4 presents the confusion matrix computed for MFCC, LFCC, and the proposed method utilizing the CNN pipeline. The results indicate that the proposed method outperforms the others, with the lowest per class error observed across all the classes. Additionally, the proposed method performs better on other widely accepted statistical parameters, such as, F1-Score, Jaccard's Index, MCC score and Hamming Loss when compared to MFCC and LFCC, using UA Speech dataset, as can be observed from Table 5.

Table 4: *Confusion Matrix of Baseline MFCC, LFCC, and Whisper(Small)*

| **MFCC** | High | Medium | Low | Very Low |
|---|---|---|---|---|
| High | 70 | 2 | 2 | 1 |
| Medium | 1 | 88 | 3 | 1 |
| Low | 1 | 1 | 88 | 3 |
| Very Low | 1 | 1 | 0 | 91 |

| **LFCC** | High | Medium | Low | Very Low |
|---|---|---|---|---|
| High | 68 | 4 | 3 | 0 |
| Medium | 2 | 88 | 2 | 1 |
| Low | 0 | 2 | 91 | 0 |
| Very Low | 0 | 0 | 0 | 93 |

| **Whisper (Small)** | High | Medium | Low | Very Low |
|---|---|---|---|---|
| High | **72** | 2 | 1 | 0 |
| Medium | 1 | **91** | 1 | 0 |
| Low | 2 | 0 | **91** | 0 |
| Very Low | 0 | 0 | 0 | **93** |

Table 5: *Performance Evaluation for Various Feature Sets*

| **Feature Set** | **Accuracy** | **F1-Score** | **MCC** | **Jaccard Index** | **Hamming Loss** |
|---|---|---|---|---|---|
| **MFCC** | 95.20 | 0.91 | 0.88 | 0.84 | 0.087 |
| **LFCC** | 96.05 | 0.96 | 0.96 | 0.93 | 0.034 |
| **Whisper** | **98.02** | **0.98** | **0.96** | **0.97** | **0.019** |

### 4.4. Analysis of Precision for Retraining

We further conducted experiments using our pipeline of the CNN model on the UA Speech corpus, to analyse the retrainability of our proposed work. The experiments were repeated five times for 100 epochs and in each of the runs, a maximum accuracy of 98.02% was obtained. Moreover, it can be observed from Fig. 4 that the performance eventually converges to the same value with minor variations due to randomness in the model and seed values.

### 4.5. Analysis of Latency Period

Fig. 5 shows that the proposed method performs significantly better than LFCC and MFCC. Whisper encoder-based method performs well even for speech signals of $100ms$ duration, making it more suitable for practical implementation.

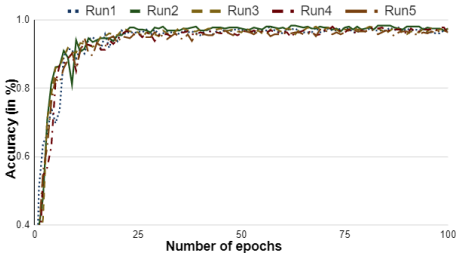


Figure 4: *Precision for Retraining*

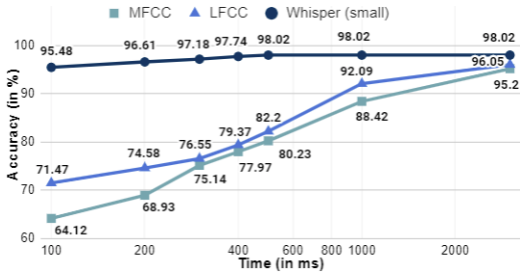


Figure 5: *Analysis of Latency Period*

## 5. Summary and Conclusion

This study investigated significance of whisper encoder-based features for classification of dysarthric severity-level. In conclusion, the utilization of a pretrained encoder-decoder transformer sequential model, which has been trained on a diverse range of audio data characterized by varying environmental settings, microphones, languages, and configurations, exhibits a significant improvement in the task of dysarthric severity-level classification. The authors believe that such a model is robust to variability, allowing it to extract meaningful features that are invariant to the wide variation in dysarthric speech based on the severity of the condition, hence capturing difference between normal and dysarthric speech with increasing levels of severity. Additionally, the model can learn to extract high-level features that capture important aspects of the speech signal, such as pitch, intonation, and spectral characteristics, which are found to be significant for dysarthric severity-level classification. Due to the small number of dysarthric patients, the proposed method's extra computation cost is not substantial. Furthermore, the method can be readily implemented as web API for medical use and doesn't require powerful computing hardware.

The Proposed transfer learning methodology is found to perform relatively better for several evaluations factors, such as comparison with state-of-the-art MFCC and LFCC features, comparison of two pipelines (one with CNN and the other with ResNet), latency period analysis, and analysis of precision should the model be retrained. Due to limited resources we were not able to evaluate performance for whisper medium and large models. Our future work would be directed towards utilising the proposed pipeline with appropriate modifications for dysarthric speech recognition and designing a numerical measure which would give more accurate diagnosis of the severity-level, which is socially relevant assistive speech technology.

## 6. Acknowledgements

The authors sincerely thank MeitY, Govt. of India, for the project 'Speech Technologies in Indian Languages' BHASHINI', (Grant ID: 11(1)2022-HCC (TDIL)).

# 7. References


[1] P. Lieberman, "Primate vocalizations and human linguistic ability," *The Journal of the Acoustical Society of America (JASA)*, vol. 44, no. 6, pp. 1574–1584, 1968.

[2] M. J. Kim, J. Yoo, and H. Kim, "Dysarthric speech recognition using dysarthria-severity-dependent and speaker-adaptive models." in *INTERSPEECH, Lyon, France*, 2013, pp. 3622–3626.

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

[4] B. A. Al-Qatab and M. B. Mustafa, "Classification of dysarthric speech according to the severity of impairment: An analysis of acoustic features," *IEEE Access*, vol. 9, pp. 18 183–18 194, 2021.

[5] A. A. Joshy and R. Rajan, "Automated dysarthria severity classification using deep learning frameworks," in $28^{th}$ *European Signal Processing Conference (EUSIPCO), Amsterdam, Netherlands*, 2021, pp. 116–120.

[6] S. Gillespie, Y.-Y. Logan, E. Moore, J. Laures-Gore, S. Russell, and R. Patel, "Cross-database models for the classification of dysarthria presence," in *INTERSPEECH, Stockholm, Sweden*, 2017, pp. 3127–31.

[7] C. Bhat and H. Strik, "Automatic assessment of sentence-level dysarthria intelligibility using blstm," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 2, pp. 322–330, 2020.

[8] S. R. Mani Sekhar, G. Kashyap, A. Bhansali, A. A. A., and K. Singh, "Dysarthric-speech detection using transfer learning with convolutional neural networks," *ICT Express*, vol. 8, no. 1, pp. 61–64, 2022.

[9] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," *arXiv preprint arXiv:2212.04356*, 2022, {Last Accessed: March 6, 2023}.

[10] Y. Iwamoto and T. Shinozaki, "Unsupervised spoken term discovery using wav2vec 2.0," in *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Tokyo, Japan*, 2021, pp. 1082–1086.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems (NIPS), Long Beach, USA*, vol. 30, 2017.

[12] L. Torrey and J. Shavlik, "Transfer learning," in *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*. IGI Global, 2010, pp. 242–264.

[13] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. Gunderson, T. Huang, K. Watkin, and S. Frame, "Dysarthric speech database for universal access research," *INTERSPEECH, Brisbane, Australia*, pp. 1741–1744, 2008.

[14] F. Rudzicz, A. K. Namasivayam, and T. Wolff, "The torgo database of acoustic and articulatory speech from speakers with dysarthria," *Language Resources and Evaluation*, vol. 46, pp. 523–541, 2012.

[15] O. M. Strand and A. Egeberg, "Cepstral mean and variance normalization in the model domain," in *Robustness Issues in Conversational Interaction, Norwich, United Kingdom*, pp. 30-31 August, 2004.

[16] A. F. Agarap, "Deep learning using rectified linear units (relu)," *CoRR*, vol. abs/1803.08375, 2018, {Last Accessed: Feb 6, 2023}. [Online]. Available: http://arxiv.org/abs/1803.08375

[17] Z. Zhang, "Improved adam optimizer for deep neural networks," in *2018 IEEE/ACM 26th IWQoS*. IEEE, 2018, pp. 1–2.

[18] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.

[19] M. Bouchard, A.-L. Jousselme, and P.-E. Doré, "A proof for the positive definiteness of the Jaccard index matrix," *International Journal of Approximate Reasoning*, vol. 54, no. 5, pp. 615–626, 2013.

[20] B. W. Matthews, "Comparison of the predicted and observed secondary structure of t4 phage lysozyme," *Biochimica et Biophysica Acta (BBA)-Protein Structure*, vol. 405, no. 2, pp. 442–451, 1975.

[21] K. Dembczyński, W. Waegeman, W. Cheng, and E. Hüllermeier, "Regret analysis for performance metrics in multi-label classification: the case of Hamming and subset zero-one loss," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2010, pp. 280–295.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vega, USA*, 2016, pp. 770–778.