ENHANCEMENT OF SLURRED SPEECH: A DEEP LEARNING APPROACH

A PROJECT REPORT

Submitted by

HARSHITAA YARRAMSETTI (195002048)

SRI KAVYA HARIHARAN (195002116)

in partial fulfillment for the award of the degree of

BACHELOR OF TECHNOLOGY

INFORMATION TECHNOLOGY

Department of Information Technology

Sri Sivasubramaniya Nadar College of Engineering

(An Autonomous Institution, Affiliated to Anna University)

Rajiv Gandhi Salai (OMR), Kalavakkam – 603 110

APRIL 2023

Sri Sivasubramaniya Nadar College of Engineering

(An Autonomous Institution, Affiliated to Anna University)

BONAFIDE CERTIFICATE

Certified that this Report titled "**ENHANCEMENT OF SLURRED SPEECH: A DEEP LEARNING APPROACH**" is the bonafide work of **Harshitaa Yarramsetti** (195002048) and **Sri Kavya Hariharan** (195002116) who carried out the work under my supervision.

Certified further that to the best of my knowledge the work reported hereindoes not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

SIGNATURE	**SIGNATURE**
Dr. Chandrabose Aravindan	Dr. Shahina A.
HEAD OF THE DEPARTMENT	**SUPERVISOR**
Head of the Department	Professor
Information Technology	Information Technology
SSN College of Engineering	SSN College of Engineering
Kalavakkam – 603 110	Kalavakkam – 603 110

Submitted for Project Viva-Voce Examination held on.....

EXTERNAL EXAMINER	**INTERNAL EXAMINER**

ABSTRACT

Existing methods of voice conversion have not proved to be efficient for speech samples that are not coherent. Hence, training dysarthric speech datasets in this criterion may not provide the desired results of enhancement. In this report, we propose a deep learning approach using neural style transfer to enhance slurred speech caused by dysarthria. We build a convolutional neural network based on mel spectrogram images given by the audio signals. The spectrograms are painted with the characteristic of the normal speech dataset thereby enhancing the way it sounds. This output image is then converted back to an audio signal using Griffin Lim algorithm as a moderate number of iterations allow us to recover the audio signal with efficiency. The final audio signal is cleaned using non stationery spectral gating. The model we have trained our dataset on is a one-shot learner that preserves the specific characteristics of individual speakers such as the pitch and rhythm of the speech even after the speech has been enhanced. Experimental results show that the evaluation metrics are highly suitable for this approach as the model is computationally inexpensive and is capable of

one-shot learning and thus works in the practical low data availability domain of dysarthric speech. The itakura saito (IS) distance helps understand the enhancement of dysarthic speech data by measuring the distance between the two mel specotgrams. As dysarthric speech is characterized by a range of distortions in the spectral envelope, IS distance has been found to be an effective measure for capturing these spectral distortions and comparing them to a target speech signal to estimate the degree of distortion. This has been tested against various convolutional models such as AlexNet, VGG16 and VGG19 to understand which gives the best results. The results confirm that VGG19 is the most suited for slurred speech enhancement.

ACNOWLEDGEMENTS

We would like to thank and express a deep sense of gratitude to our guide Dr. Shahina A., Professor, Department of Information Technology, for her valuable advice and suggestions as well as her continued guidance, patience and support that helped us to shape and refine our work.

Our sincere thanks to Dr. Chandrabose Aravindan, Professor and Head of the Department of Information Technology, for his words of advice and encouragement and we would like to thank our project Coordinator Dr. T. Sree Sharmila, Associate Professor, Department of Information Technology for her valuable suggestions throughout this first phase of project.

We express my deep respect to the founder Dr. SHIV NADAR, Chairman, SSN Institutions. We also express my appreciation to our Dr. V. E. Annamalai, Principal, for all the help that has rendered during this course of study.

We would like to extend my sincere thanks to all the teaching and non-teaching staffs of our department who have contributed directly and indirectly during the course of my project work. Finally, we would like to thank our parents and friends for their patience, cooperation and moral support throughout my life.

Harshitaa Yarramsetti Sri Kavya Hariharan

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

INTRODUCTION

1.1 BACKGROUND

Indistinct utterances of words due to their overlapping and incoherent nature is designated as slurring. Speech disorders can be characterized by the change in speed or rhythm of words being spoken and the level of clarity exhibited during a conversation. Dysarthria is one such voice pathology that is caused by nervous system disorders that may gradually cause facial paralysis and thus weaken the throat muscles and consequently affect the patient's speech. Accurate human perception of the slurred speech of dysarthric patients is extremely challenging for first time listeners. Patients afflicted with dysarthria gradually tend to lose control of their tongue and voice box thereby causing a slur during communication. It remains the most common neurodegenerative disease symptom.

Some other pathologies that have been known to cause dysarthria are Amyotrophic lateral sclerosis (ALS, or Lou Gehrig's disease), Cerebral palsy, Guillain-Barre syndrome, Huntington's disease, Lyme disease and so on. The terms slurring and dysarthria are often used interchangeably. Dysarthria can be an early symptom of amyotrophic lateral sclerosis (ALS) in about 30% of individuals, and nearly all patients eventually develop dysarthria in the later stages of the disease. Dysarthria is believed to impact around 70% to 100% of people with Parkinson's disease. However, diseases are not the only factors that bring about changes in the way a person talks. Fatigue, intoxication, and lack of teeth are few examples of what can cause slurred speech in a person.

The (State of the art) SOTA Automatic Speech Recognition Systems (ASR) such as Whisper do not perform well on moderate to severely dysarthric speech. Whisper ASR (Automatic Speech Recognition) is a technology aimed at recognizing and transcribing whispered speech. It is designed to work with whispered speech, which is a mode of speech characterized by reduced vocal fold vibration and decreased sound intensity.

Speech enhancement aims at improving the quality of a speech by suppressing the distortions and noises present in the signal. In recent times,

speech enhancement is being used in academia to reconstruct a speech signal's prosody thereby improving its intelligibility.

1.2 NEED FOR THE STUDY

Voice conversion has been a popularly employed speech enhancement approach to transfer between two independent domains. This approach has been adopted has been successfully implemented across several domains such as accent, gender, and emotion translation. However, when it comes to dysarthric to normal speech domains, the high levels of variability across speakers and genders presents itself as a major impediment to developing a reliable and consistent mapping. A second but equally important hurdle is the lack of a large corpus for dysarthric and normal parallel utterances. This has to do with the fact that patients are unable to utter more than a certain number of utterances without feeling the fatigue that is prevalent in cases of advanced stages of dysarthria. The smaller volume of data renders the distributions being modelled restrictive and non-generalizable to a wider range of utterances. On a separate note, the deployment of such a system elicits the need for a real time conversion and a low computational power system that can generate mappings. Also, dysarthric patients have varying slangs and accents, and each word is uttered at a different rhythm and pitch.

Furthermore, slurring is usually the result of some physical disabilities that cause communication difficulties and creates hurdles in using physical Augmentative and alternative communication (AAC). Enhancement is a good alternative for Augmentative and alternative communication (AAC) devices like keyboards, retina trackers used by patients. Usage of enhanced speech for automatic speech recognition (ASR) is the need of the hour as mild-to-severe Dysarthria gives lower accuracies in normal ASRs. In addition, various virtual assistant technologies remain inaccessible to individuals affected by dysarthria thereby requiring enhancement of their speech to improve accessibility. Accurate human perception of the slurred speech uttered by dysarthric patients is extremely challenging for listeners so there is an exigent need to bridge this gap for ease of communication.

Speech Generative adversarial network (SEGAN) is one of the most frequently used neural network that performs multistage enhancement

mapping to refine the speech signals. Improvisation of GAN has also resulted in various other models like ISEGAN, DESGAN, Cycle GAN which assist better in feature learning. A major drawback of deploying this on dysarthric speech is the variation in phonemes and pronunciations. As discussed above, there is only a limited amount of dysarthric speech available. Understanding the nuances of every word uttered is essential in speech enhancement and in addition, the Minimax architectures are computationally expensive and are data hungry models that might overfit to our sample of the population.

To solve the problems discussed above, we propose a neural style transfer approach to improve the intelligibility of dysarthric speech. Neural style transfer (NST) is an artistic image manipulation technique which blends a content image with the style image to adapt the style of one image onto the other. This has previously been used for voice conversions and cross lingual style transfer. Application of image transfer learning based neural style transfer on dysarthric speech to normalize it is a cost efficient, real-time approach as it is a one-shot learner that only requires parallel recordings of time continuous signals in the form of audio signals.

We utilize Mel spectrograms as the vocal tract representations of both the normal and dysarthric patient's utterances in the form of audio files which are inputted to the convolutional neural network. Using neural style transfer, we construct the Mel spectrogram of the enhanced utterance. We synthesize this utterance such that it preserves the specific characteristics of the dysarthric speaker prosody of the speech and simultaneously ensure that it also contains normal speaker characteristics such as correct syllable pronunciation and rhythm. This output graph is then converted back to an audio signal using Griffin Lim algorithm. The final audio signal is then denoised using non stationery spectral gating as the generated speech sample contains noise.

1.3 OBJECTIVE OF THE STUDY

We propose a novel approach to enhance dysarthric speech by employing Neural style transfer. We attempt to preserve speaker identity in such a way that the content representation of the control as well as normal patient is not compromised, thereby, enhancing speech intelligibility. The main objective

of the study is to make the source speaker's (dysarthric utterance) speech sound intelligible and similar to that of the target speaker (control utterance) while maintaining the linguistic content and the prosody of the original speech. Additionally, we try to capture the various minute variances in slurred speechand enhance them in a fine-grained manner by utilizing an appropriate vocal tract model that exhibits all necessary speaker characteristics.

Chapter 2

REVIEW OF LITERATURE

2.1 Dysarthric Speech Voice Conversion:

The findings in [8] cements voice conversion as a superior method over voice banking as the latter relies on building Text to Speech Synthesis Systems (TTS) for patients prior to speech deterioration. Voice Conversion ensures the preservation of linguistic content while converting across acoustic domains. Rule based VC does not provide stability across speaker frequency and temporal characteristics due to idiosyncratic speaker acoustic characteristics. However, frame-based mapping is not possible and neither is correction of severe dysarthric speech. The paper proposes an end-to-end VC for Dysarthric Speech Recognition. This consists of these components - a seq2seq TTS with encoder, attention and decoder that is trained on normal speech to generate character embeddings and a cross modal knowledge distillation system that consists of a speech encoder that converts dysarthric speech spectral features tospectral embeddings that are forced to resemble the character embeddings of the TTS speech encoder using adversarial loss. Metrics used were MOS and one human recognition score and WER of 9.33 onhigh dysarthria, 34.67 on medium, 43.32 on low and 44.67 on very low. Proposed model had the most naturalness.

The Gated CNN [9] can capture such long-range dependencies and the gated activation functions can be trained on the data so there is greater control on what information passes through the hierarchy. Parallel dysarthric and normal speech is used in the proposed Gated CNN. The dysarthric speech 80 dimensional mel spectral features are extracted and put into a speaker dependent ASR to get 74 dimensional PPGs which are then inputted to the gated CNN. The mel spectrograms output are converted to waveforms using WaveRNN. The Taiwan Mandarin hearing in noise test data is used with 1 dysarthric and one normal patient. Comparison models were CNN and BLSTM and the GatedCNN outperformed the Google ASR accuracy test with the enhanced speech ata rate of 85.9%.

While short term speech prosody modification does improve intelligibility of

speech, a better solution is an end-to-end VC. However, these cause loss of speaker identity. [10] proposes a style transfer approach by using cycle consistent loss in GANs to reduce the domain space of the possible mapping functions or outputs. During training, the normal and dysarthric sample spectrograms are given to the discriminator while the generator is made to distinguish between fake and real distributions. In the testing phase, the generator is made to convert the test samples to corrected speech spectrograms.The discriminator uses a Markovian PatchGan. QoLT (Quality of Life Technology) database was used with 100 dysarthric patients with cerebral palsy in the 30-40 age group. The waveform is converted to spectrogram using Short-Time Fourier Transform (STFT) with windows of 512 frames and 33% overlap. Then these are padded with white noise to form 128*128-pixel images. The Griffin Lim algorithm is used to convert the STFT to waveform again. The WER metric is used to evaluate on Google ASR. The dysarthric speechWER is 66.7 while the enhanced speech WER is 33.3.

[11] proposes an identity preserving Dysarthric to Normal seq2seq based VC model called Voice Transformer Network. It's a frame-by-frame model that uses local attribute preservation. Earlier works like Deep Convolution GANs, CycleGANS could not preserve speaker identity for normal to dysarthric VC. Indysarthric-to-dysarthric VC, an inflexible HL-VQ-VAE was described. The proposed model has two components, namely, the parallel seq2seq model and a nonparallel frame-wise model. In the seq2seq model, the training data for the VTN is from normal speakers, and the conversion is done to the target dysarthric speaker. Open-source VC software is used to implement VQVAE forthe second component. UA Speech dataset is used for the metrics such as P-ESTOI/P-STOI, PhonemeError Rate, Naturalness, severity and similarity.

2.2 General Speech Voice Conversion:

In the task of voice conversion (VC), the voice of the source speaker is converted into the voice of the target speaker, but the voice content does not change in this task. There has been some research in voice conversion and some progress and results have been achieved. A new voice conversion network based on GAN network is proposed, which is a voice conversion technique that relies on non-parallel data and is capable of converting samples of arbitrary duration. This network is called IVCGAN [12]. The network consists of a discriminator and a generator, in which the function of discriminator is to distinguish the real speech from the converted speech, and to classify the sourcespeakers corresponding to the speech, while the function of generator is to deceive the discriminator.

In [13], authors propose cross-lingual voice conversion (VC) to convert the sourcespeaker's voice to sound like that of the target speaker, when the source and target speakers speak different languages. In this approach, Generative Adversarial Networks (GANs) for cross-lingual voice-conversion are proposed. Further, Variational Autoencoding Wasserstein GAN (VAW-GAN) and cycle-consistent adversarial network (CycleGAN) are studied. These are known to be effective for mono-lingual voice conversion. As cross-lingual voice conversion needs to convert the voice across different phonetic system, itis more challenging than mono-lingual voice conversion. By using VAW-GANand CycleGAN, they successfully convert the speaker identity while carrying over the source speaker's linguistic content. The proposed idea is unique in the sense that it neither relies on bilingual data and their alignment, nor any externalprocess, such as ASR. Moreover, it works with limited amount of training data of any two languages.

2.3 Neural Style Transfer:

In [14], the authors propose to use image representations derived from Convolutional Neural Networks optimized for object recognition, which make high level image information explicit. We introduce A Neural Algorithm of Artistic Style that can separate and recombine the image content and style of natural images. The algorithm allows them to produce new images of high perceptual quality that combine the content of an arbitrary photograph with the appearance of numerous well- known artworks. Our results provide new insights into the deepimage representations learned by Convolutional Neural Networks and demonstrate their potential for high level image synthesis and manipulation.

Neural style transfer, which creates fantastically stylized images by separating and recombining the content and style, has received much attention in both academic and industrial applications. [15] proposes a novel content aware neural style transfer algorithm by taking account of variations of image contents (e.g., smooth, salient regions). To this end, their method first divides the content image into different regions based on salient information and then stylizes different regions discriminatively. Such a content-aware way of transfer retains more identifiability of the subject in the content image resulting in higher appreciable value in image quality than general transfer.

In [16], the authors propose a deep learning-based approach to perform voice conversion and speech style transfer between different speakers. They use a combination of Variational Auto-Encoder (VAE) and Generative Adversarial Network (GAN) as the main components of their model, followed by a WaveNet-based vocoder. The model is evaluated using three objective metrics, including ASVspoof 2019 for measuring the difficulty of differentiating between human and synthesized samples, content verification for transcription accuracy, and speaker encoding for identity verification. The results show the effectiveness of the proposed model in generating high-quality synthesized speech on the Flickr8k audio corpus.

In [17], the authors propose a novel approach for seen and unseen style transfer training on disjoint, multi-style datasets. The datasets consist of recordings of different styles, with each style recorded by one speaker in

multiple utterances. The authors introduce an inverse autoregressive flow (IAF) technique to improve variational inference for learning an expressive style representation. They also develop a speaker encoder network for learning a discriminative speaker embedding, which is jointly trained with the rest of the neural text-to-speech modules. The proposed approach uses six specifically-designed objectives, including reconstruction loss, adversarial loss, style distortion loss, cycle consistency loss, style classification loss, and speaker classification loss. Experiments show that the proposed approach is effective for seen and unseen style transfer tasks, both objectively and subjectively. The performance of the proposed approach is superior to and more robust than four other reference systems from prior art.

Dataset	Approach	Advantages	Disadvantages	Evaluation Metric	Speaker Depen-dency	One shot Learner
LJ Speech and UA Speech	TTS system trained with transcribed normal speech, cross modelled with transcribed dysarthric speech to train a speech encoder	First attempt to apply end-to-end Voice conversion based on knowledge distillation to the dysarthric speech reconstruction task	Difficult to preserve both speaker identity and content at the same time	35% reduction in word error rate	Yes	No
Corpus list adopted from Taiwan Mandarin hearing in noise test	Gated convolutional based voice conversion systems	Gated CNN improves the performance of a CNN model and number of parameters in this was only 35% to that in BLSTM.	Abundance of data is required	Accuracy 87.8%. Improves recognition rate from 17.1% to 80%	Yes	No
Quality of Life Technology	Cycle GAN - features automatically learnt in an unsupervised manner	Cycle consistency loss preserves translational cycle.	Cycle GAN can only learn one-to-one mapping	33.4% reduction in absolute word error rate	Yes	No
UA Speech dataset	Voice conversion model using voice trausformer network in a many-to-one seq2seq modelling	Mimic various characteristics of a speech signal	Loss of speaker identity	Lower phoneme error rate	Yes	No
AISHELL-3 Chinese dataset	IVCGAN , end to end network based on non-parallel data	To ensure learning of all samples in the domain and to eliminate the dependence of frequency spectrum characteristics	Huge gap between naturalness and source speech	Reduced MCD (Mel cepstral distortion) values	Yes	No
CMU database	Cross lingual voice conversion with variational auto encoders	Cycle GAN consistently outperforms VAW-GAN	Not feasible in evaluating output	High values of MCD for source speaker	Yes	No
NA	Neural algorithm of artistic style using CNN to perform style transfer	High level content of every object is captured	Weightage difference between images	NA	NA	NA
Flickr8k	VAE along with GAN	Improved style transfer of speech signals	Huge amounts of training data	WER of 10.36%	Yes	No
Chinese Corpus	Encoder-decoder neural network to perform seen and unseen style transfer on disjoint, multi style datasets	Improve performance on seen and unseen styles			Yes	No

Table 2.1: Gap Analysis

Chapter 3

PROBLEM DEFINITION AND ALGORITHM

3.1 TASK DEFINITION

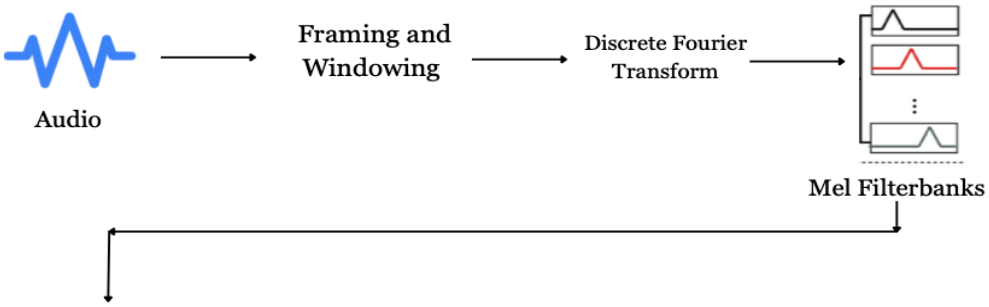

Fig. 3.1: Melspectrogram Extraction

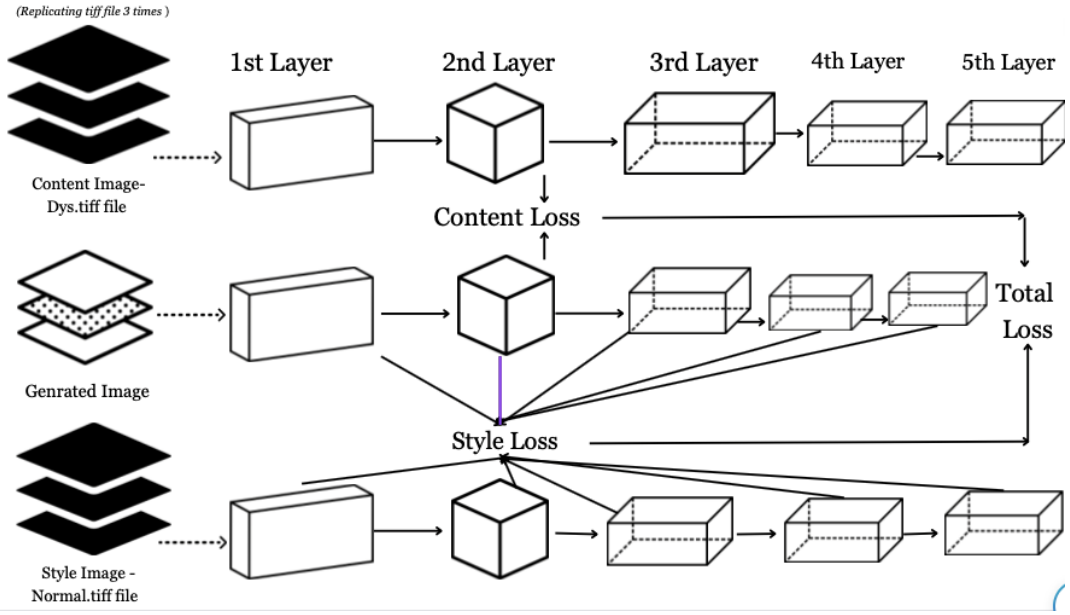

Fig. 3.2: Neural Style Transfer

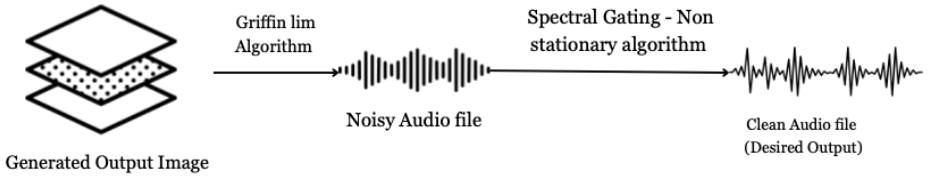

Fig. 3.3: Audio Extraction

We train a one-shot learner by using neural style transfer in order to enhance dysarthric speech. One-shot learning allows our model to generalize to new speakers with very few training examples. In traditional VC (voice conversion) approaches, a large amount of source and destination utterances is required to train a model that can perform well on unseen destination utterances. However, in voice conversion of normal to dysarthric utterances, it is not feasible to collect large amounts of data from each speaker, especially for severe dysarthria. One-shot learning approaches, are specifically designed to learn from a small number of training examples and generalize to new instances of the same task. This makes them well-suited for our task. In addition, this method is more effective than other methods because they can better capture the underlying structure of the speech signal, allowing for more accurate and natural-sounding voice conversion.

We input the audio samples of both the control and dysarthric utterance for either male/ female which are then represented as Mel spectrograms and inputted to the neural style transfer model. A mel spectrogram is a representation of the frequency content of an audio signal, where the frequency axis is divided into equally spaced mel frequency bins. The mel scale is a perceptually based frequency scale that is more aligned with the way humans perceive sounds than the linear frequency scale. The spectrogram is computed by taking the short-time Fourier transform (STFT) of the audio signal and applying a filterbank that is designed to mimic the human auditory system's frequency response.

Melspectrograms are utilized because they provide a compact and robust representation of the spectral content of an audio signal. They capture important features such as formants, which are responsible for the perceived vowel sounds, and the harmonic structure of the voice. The mel frequency scale also helps to reduce the dimensionality of the spectral representation, making it more suitable for the application at hand. In addition,

melspectrograms are invariant to certain transformations that are not meaningful for voice conversion, such as changes in the overall energy level of the signal. This makes them a more suitable input representations, which can learn to extract the relevant features for voice conversion without being distracted by irrelevant variations in the input signal.

The neural style transfer model we employ here utilizes transfer learning using VGG-19 Convolutional Neural Network. Transfer learning enables us to take advantage of Image Net weights without having to create a completely new model that requires millions of Mel spectrogram data graphs.

However, as Image Net is primarily an object detection dataset, we ensure that we utilize only the lower-level information from the feature maps generated. Transfer learning using pre-trained Convolutional Neural Networks (CNNs) such as ImageNet can be very effective for tasks such as sound classification using mel spectrogram data. This is because the lower-level features learned by CNNs on the ImageNet dataset, such as edges, corners, and blobs, are often relevant for other visual tasks as well, including those involving sound analysis. Mel spectrogram data represents sound as a visual image, where the y-axis represents frequency and the x-axis represents time. Therefore, the visual patterns in mel spectrogram data are very similar to those found in images. By using pre-trained CNNs that were originally trained on the ImageNet dataset, we can leverage the learned lower-level features to extract relevant features from the mel spectrogram data. In transfer learning, we can take advantage of the learned feature representations from the lower layers of the pre-trained CNN and fine-tune them for a new task using a smaller dataset. By doing so, we can reduce the amount of data needed for training and improve the generalization of the model on the new task.

The neural style transfer module outputs the enhanced Mel spectrogram which is converted back to audio by utilizing the Griffin Lim Algorithm. The Griffin-Lim algorithm is an iterative algorithm used for signal reconstruction in sound processing. It involves alternating between a time-frequency analysis step, where the short-time Fourier transform is applied to the input signal, and a synthesis step, where the signal is reconstructed from the magnitude and phase of the short-time Fourier transform using an iterative phase retrieval technique. This algorithm is commonly used in speech and music processing applications, where it can be used for tasks such as audio compression and synthesis.

Finally, the audio is denoised using the spectral gating algorithm. Spectral noise gating is a noise reduction technique used in audio signal processing that involves dividing the frequency spectrum of a signal into separate bands, and then applying a gating threshold to each band based on the level of noise in that band. If the noise level in a band exceeds the gating threshold, the signal in that band is attenuated. This technique is commonly used in audio production to reduce unwanted noise in recordings, and can be implemented using digital signal processing algorithms such as Fourier transforms and filters.

3.1 ALGORITHM DEFINITION

3.1.1 Neural style transfer

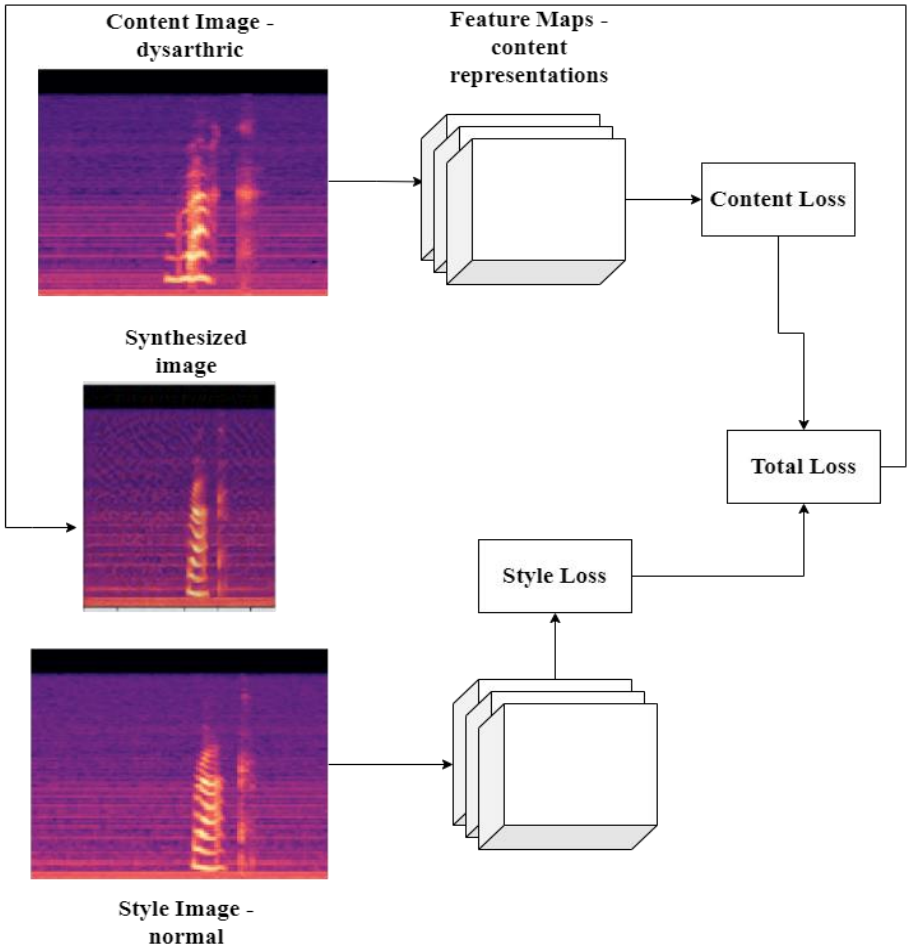

Fig. 3.4: Neural Style Transfer Methodology

The neural style transfer algorithm is a technique that combines the content of one image with the style of another image using a neural network. The main steps of the algorithm are as follows:

1. Define the content mel spectrogram and the style mel spectrogram as dysarthric and normal respectively.

2. Load a pre-trained neural network, with architecture similar to orVGG-19 or a similar network that has been trained on a large dataset of images.

3. Define a loss function that consists of two components: the content loss and the style loss. The content loss measures the difference between the feature maps of the content mel spectrogram and the generated mel spectrogram, while the style loss measures the difference between the Gram matrices of the style mel spectrogram and the generated mel spectrogram at different layers of the neural network.

4. Initialize a third mel spectrogram, called the generated mel spectrogram, as a copy of the content mel spectrogram.

5. Use an optimization algorithm (gradient descent), to iteratively update the generated mel spectrogram to minimize the loss function.

6. Repeat step 5 until the generated mel spectrogram converges to an audible melspectrogram.

During each iteration of the optimization algorithm, the neural network is used to compute the content and style features of the generated mel spectrogram. The loss function is then computed based on the difference between the features of the generated mel spectrogram, the content mel spectrogram, and the style mel spectrogram. The optimization algorithm then updates the generated mel spectrogram to reduce the loss function.

The final result is a mel spectrogram that combines the content of the content mel spectrogram with the style of the style mel spectrogram, as determined by the relative weighting of the content loss and style loss, and the selected layers of the neural network used to compute the loss. We use the Conv1_1, Conv2_1, Conv3_1, Conv4_1 and Conv5_1 layer as the style representation layers and the Conv2_1 layer as the content layer.

Let C be the content image, S be the style image, and G be the generated image. Let VGG-19 be the pre-trained neural network used for feature

extraction. The content loss L_C is defined as the Euclidean distance between the feature maps of the content image C and the generated image G at a selected layer l of the neural network:

$$L\_C = 1/2 * ||F\_l(C) - F\_l(G)||^2$$

where F_l(.) is the feature map of the l-th layer of VGG-19.

The style loss L_S is defined as the difference between the Gram matrices of the style image S and the generated image G at different layers l of the neural network:

$$L\_S = \sum\_l w\_l * ||G\_l(S) - G\_l(G)||^2$$

where G_l(.) is the Gram matrix of the feature maps at the l-th layer of VGG-19, and w_l is a weighting factor that controls the relative importance of each layer.

The total loss L is the weighted sum of the content loss and style loss:

$$L = \alpha * L\_C + \beta * L\_S$$

where α and β are hyperparameters that control the relative weighting of the content and style.

The optimization algorithm, such as gradient descent, is then used to update the generated image G to minimize the total loss L. At each iteration, the neural network is used to compute the content and style features of the generated image, and the loss function is then computed based on the difference between the features of the generated image, the content image, and the style image. The optimization algorithm then updates the generated image to reduce the loss function.

3.1.2 Mel spectrogram

Below are the steps to create mel spectrograms of audio signals

1. Divide the audio signal into frames of duration T with a hop length of H. The nth frame is denoted by x(n), with n=0,1,2,...,N-1, where N is the total number of frames.

2. Apply a window function w(n) to each frame, such as the Hamming

window. The windowed signal is denoted by **x_w(n) = x(n) * w(n),** where * denotes element-wise multiplication.

3. Compute the magnitude spectrum of each windowed frame using the Fast Fourier Transform (FFT):

   X_m(k, n) = |FFT(x_w(n))|^2

where k is the frequency bin index and n is the frame index.

4. Map the resulting frequency values onto the mel scale using a mel frequency filterbank. The filterbank consists of a set of M triangular filters that are applied to the magnitude spectrum to extract the mel spectrogram:

   H_m(k) =

   0, if k < f_{m-1}

   (k - f_{m-1}) / (f_m - f_{m-1}), if f_{m-1} <= k < f_m

   (f_{m+1} - k) / (f_{m+1} - f_m), if f_m <= k < f_{m+1}

   0, if k >= f_{m+1}

where f_m is the center frequency of the m-th mel filter, and H_m(k) is the filter response at frequency bin k.

5. Apply the mel filterbank to the magnitude spectrum of each frame to obtain the mel spectrogram:

   Y_m(n, m) = \sum_{k=0}^{K-1} H_m(k) X_m(k, n)

where m is the mel filter index, n is the frame index, and K is the total number of frequency bins in the magnitude spectrum.

6. Optionally, take the logarithm of the mel spectrogram to convert it into the decibel (dB) scale:

   Z_m(n, m) = 10 * log10(Y_m(n, m) + eps)

where eps is a small constant added to avoid taking the logarithm of zero.

7. Normalize the mel spectrogram by subtracting the mean and dividing by the standard deviation across all frames:

   S_m(n, m) = (Z_m(n, m) - mu_m) / sigma_m

where mu_m and sigma_m are the mean and standard deviation of the mth mel frequency band across all frames.

The resulting mel spectrogram S_m(n, m) is a two-dimensional array that represents the spectral content of the audio signal over time, with mel frequency bands arranged along the vertical axis and time intervals arranged along the horizontal axis.

3.1.3 Spectral Gating

The algorithm for noise reduction using non-stationary spectral gating involves the following steps:

1. Compute the short-time Fourier transform (STFT) of the noisy audio signal to obtain the complex spectrogram X.

2. Compute the magnitude spectrogram |X| by taking the absolute value of the complex spectrogram.

3. Estimate the noise power spectrum N by computing the median magnitude spectrogram over time.

4. Calculate the non-stationary threshold function T(n, k) at each time frame n and frequency bin k using:

   T(n, k) = α(n) * N(k) + β(n) * M(n, k)

   where M(n, k) is the magnitude spectrogram of the noisy signal at time frame n and frequency bin k, α(n) and β(n) are scalar values that control the trade-off between noise reduction and speech distortion, and N(k) is the noise power spectrum at frequency bin k.

5. Compute the binary mask Y(n, k) using:

   Y(n, k) = 1 if M(n, k) > T(n, k)

Y(n, k) = 0 otherwise

6. Apply the mask Y to the magnitude spectrogram of the noisy signal to obtain the magnitude spectrogram of the clean signal:

   Z(n, k) = Y(n, k) * M(n, k)

7. Estimate the phase of the clean signal using the phase of the noisy signal:

   P(n, k) = arg(X(n, k))

8. Reconstruct the clean signal by computing the inverse STFT of the product of the magnitude spectrogram and the phase:

   y = ISTFT(Z * exp(iP))

   where ISTFT is the inverse STFT operator and i is the imaginary unit.

The values of $\alpha(n)$ and $\beta(n)$ can be chosen based on the desired trade-off between noise reduction and speech distortion.

Chapter 4

EXPERIMENTAL EVALUATION

4.1 METHODOLOGY

4.1.1 Metrics

We utilize the Itakura Saito Distance as our objective metric between the enhanced, control and dysarthric melspectrograms to determine whether the speech has been enhanced appropriately. The Mean Opinion Score (MOS) is used to determine the subjective goodness of the enhanced audio. Further, the Linear Predictive (LP) Spectra are a graphical representation and comparison between the normal, enhanced and dysarthric audio.

Itakura-Saito distance is a measure of spectral distance between two speech signals. It is based on the Itakura-Saito distortion measure, which is a logarithmic representation of the squared error between two spectra. The distance measure considers the temporal and spectral properties of the signals, and it is particularly useful for speech signal analysis and synthesis. It is commonly used in speech processing tasks such as speech recognition, speaker recognition, and speech coding. The Itakura-Saito distance is calculated using dynamic programming and is widely used in various speech processing applications due to its good performance and low computational complexity. Itakura-Saito distance is a suitable choice for dysarthric-normal voice conversion (VC) because it can capture spectral distortions that often occur in dysarthric speech. Dysarthric speech is characterized by various types of distortions in the spectral envelope, such as changes in formant frequencies and bandwidhts, and these spectral distortions can affect the perception of speech. Itakura-Saito distance is designed to measure the spectral distance between two signals, and it is sensitive to these types of spectral distortions. Therefore, using Itakura-Saito distance as a cost function in the VC system can help to accurately map the spectral characteristics of the dysarthric speech onto the target normal speech, resulting in improved speech intelligibility and naturalness.

Mean Opinion Score (MOS) is a subjective quality assessment method used

to evaluate the quality of audio, video, or other multimedia content. MOS testing involves presenting test subjects with a set of audio or video stimuli and asking them to rate the stimuli on a scale from 1 to 5, where 1 indicates the lowest quality and 5 indicates the highest quality. The scores are then averaged to obtain a MOS score, which provides an overall assessment of the quality of the stimuli. MOS testing is widely used in industry and academia to evaluate the performance of multimedia processing algorithms and systems. MOS testing is a good choice for evaluating the performance of dysarthric-normal voice conversion (VC) systems because it provides a subjective quality assessment that reflects the perceptual quality of the converted speech. Dysarthric speech is often characterized by various types of spectral and prosodic distortions that can affect the intelligibility and naturalness of the speech. MOS testing can provide a measure of how well the VC system is able to mitigate these distortions and produce speech that is intelligible and natural sounding to listeners. MOS testing can also be used to evaluate the performance of different VC algorithms and configurations, and to optimize the system parameters to improve the overall quality of the converted speech. Therefore, MOS testing is a suitable choice for evaluating the performance of dysarthric-normal VC systems.

The Linear Prediction (LP) spectrum is a representation of the spectral envelope of a speech signal. It is obtained by applying linear prediction analysis to the speech signal, which models the speech signal as a linear combination of past samples. LP analysis estimates the filter coefficients that describe the spectral envelope of the speech signal, and the LP spectrum is obtained by evaluating the spectral magnitude response of this filter at different frequencies. The LP spectrum is commonly used in speech processing applications such as speech coding, speech enhancement, and speech synthesis, as it provides a compact representation of the spectral envelope that can be efficiently manipulated for these tasks. LP spectra is a good choice for dysarthric-normal voice conversion (VC) because it captures the spectral characteristics of speech signals in a compact form. Dysarthric speech is often characterized by changes in the spectral envelope due to articulation and motor control disorders. LP analysis can model these spectral characteristics effectively, and the resulting LP spectra can provide a compact representation of the spectral envelope that can be used to map the spectral characteristics of the dysarthric speech onto the target normal speech in the VC system. The LP spectra can also be used to estimate the vocal tract system transfer function, which is an important component for speech synthesis in

VC. Therefore, LP spectra is a suitable choice for dysarthric-normal VC as it can help to accurately model and map the spectral characteristics of dysarthric speech.

4.1.2 Hypothesis

Dysarthria is a motor speech disorder that affects the muscles used for speech production, resulting in reduced intelligibility and difficulty in communication. Voice conversion (VC) techniques have been developed to improve the intelligibility of dysarthric speech by converting it to normal speech. However, most existing VC techniques require a large amount of parallel training data, which can be difficult to obtain in the case of dysarthric speech.

This is where neural style transfer (NST) can be useful. NST is a deep learning technique that can transfer the style of one image onto another. In the context of speech, this means that NST can be used to transfer the "style" of normal speech onto dysarthric speech, effectively converting it to a more intelligible form.

The hypothesis is that CNN-based NST can be used as a one-shot learner for dysarthric-normal voice conversion. This means that the CNN model can be trained using a small amount of paired dysarthric and normal speech data, and then used to convert new dysarthric speech samples to normal speech without the need for additional training data.

To test this hypothesis, a dataset of paired dysarthric and normal speech samples would be required. The TORGO dataset is a good choice for this, as it contains both dysarthric and normal speech samples recorded from the same speakers. The dataset would need to be preprocessed to extract the relevant features, such as mel spectrograms, and then used to train the CNN model.

The architecture of the CNN model could be based on VGG19 or a similar architecture that has been shown to be effective for NST. The input to the model would be the mel spectrogram of the dysarthric speech, and the target output would be the mel spectrogram of the corresponding normal speech. The model would be trained using a loss function that encourages the output mel spectrogram to match the target mel spectrogram, while also preserving the content of the original dysarthric speech.

Once the CNN model has been trained, it can be used to convert new dysarthric speech samples to normal speech by feeding them through the model and obtaining the output mel spectrogram. The mel spectrogram can then be converted back to waveform using the Griffin Lim Algorithm, and the resulting speech can be evaluated using objective measures such as the Itakura-Saito distance and subjective measures such as MOS tests.

If the hypothesis is supported by the results, it would have significant implications for improving the intelligibility of dysarthric speech. The one-shot learning approach using CNN-based NST could enable VC without the need for large amounts of training data, making it more accessible and practical for use in clinical settings.

4.1.3 Dataset

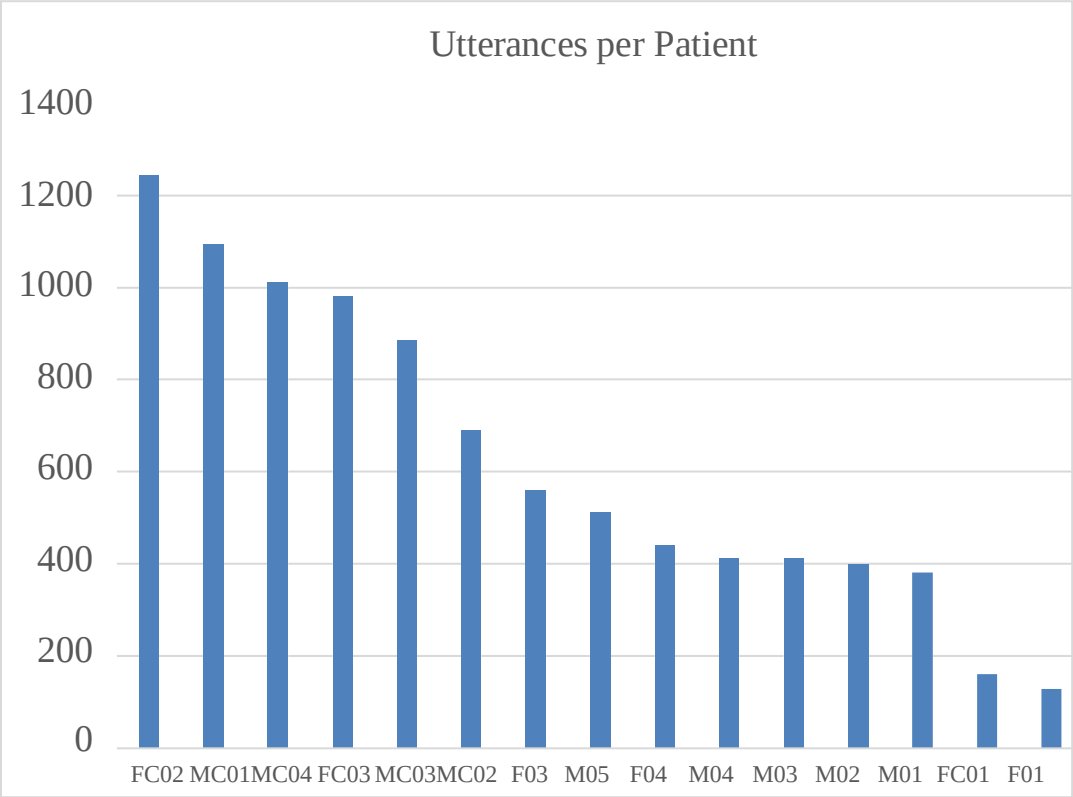

Fig: 4.1: TORGO utterances per patient

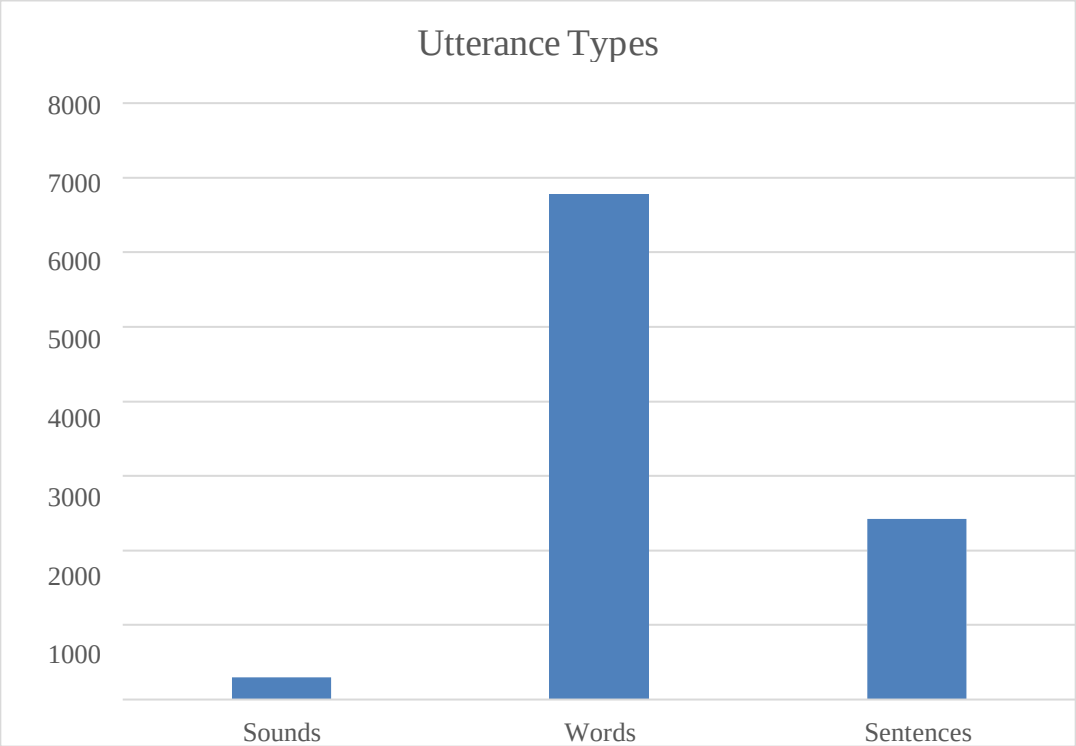

Fig. 4.2: TORGO Utterance Types

We use the TORGO dataset to perform the proposed methodology. The TORGO (Toronto and Groningen) dataset is a publicly available database of audiovisual recordings of speech for use in research on human-computer interaction, speech recognition, and speech processing. It was jointly developed by the University of Toronto and the University of Groningen, and consists of recordings of 24 speakers, including 12 males and 12 females, with varying degrees of dysarthria. The speakers were recorded in both audio-only and audiovisual modalities, under two different speaking conditions: (1) reading a set of standard sentences and (2) performing a set of spontaneous tasks such as describing pictures and telling jokes. The dataset includes both clean speech and speech with varying degrees of noise and reverberation. The audio recordings were manually transcribed and segmented at the word and phoneme level. The TORGO dataset has been widely used in research on speech processing, speech recognition, and voice conversion, and has contributed to the development of new algorithms and techniques for improving the quality of speech signals in noisy and reverberant environments, as well as for enhancing the intelligibility of speech produced

by dysarthric speakers.

The TORGO (Toronto and Groningen) dataset is a valuable resource for developing and evaluating dysarthric-normal voice conversion (VC) systems. Dysarthric speech is characterized by various types of distortions that affect its quality, intelligibility, and naturalness. The TORGO dataset includes recordings of speakers with varying degrees of dysarthria, speaking under different conditions and tasks, with both clean speech and speech with varying degrees of noise and reverberation. The dataset also includes manually segmented phoneme and word transcriptions, which provide a valuable resource for training and testing VC systems. By using the TORGO dataset, researchers can develop VC systems that can improve the intelligibility and naturalness of dysarthric speech, and convert it into speech that is more natural and easier to understand for listeners. The use of the TORGO dataset can help to optimize the performance of dysarthric-normal VC systems, and contribute to the development of new techniques for improving the quality of speech signals produced by dysarthric speakers.

4.1.4 Comparative Study

Neural Style Transfer (NST) is a technique in deep learning that involves generating an image that combines the content of one image with the style of another. NST can be performed using various pre-trained deep neural network architectures, including AlexNet, VGG19, and VGG16. Each of these architectures has its own unique characteristics, which can affect the quality of the stylized images produced.

AlexNet was one of the first deep neural networks that achieved significant success in the ImageNet Large Scale Visual Recognition Challenge in 2012. It has 8 layers, including 5 convolutional layers and 3 fully connected layers. The convolutional layers have filter sizes of 11x11, 5x5, and 3x3, and the fully connected layers have 4096 neurons each. The last layer of AlexNet is a softmax layer that outputs a probability distribution over the 1000 ImageNet classes.

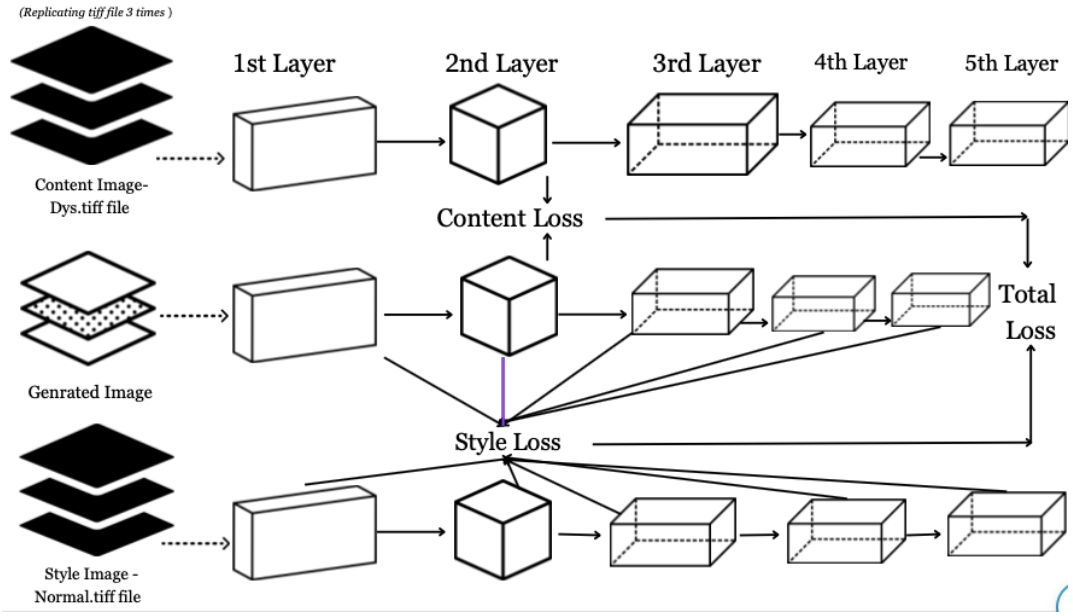

Fig 4.3: AlexNet NST Architecture

NST using AlexNet involves selecting one or more of the early convolutional layers to extract feature maps from the content and style images, and then applying the Gram matrix operation to these feature maps. The Gram matrix is a measure of the correlations between the activations of different filters in a layer. By minimizing the distance between the Gram matrices of the content and stylized images, the NST algorithm is able to generate a stylized image that combines the content of the content image with the style of the style image.

VGG19 and VGG16 are deeper architectures that were developed by the Visual Geometry Group at the University of Oxford. VGG19 has 19 layers, while VGG16 has 16 layers. Both architectures have a similar structure, with several blocks of convolutional layers followed by a few fully connected layers. VGG networks use only 3x3 filters for convolutional layers, which allows them to learn more complex features compared to AlexNet.

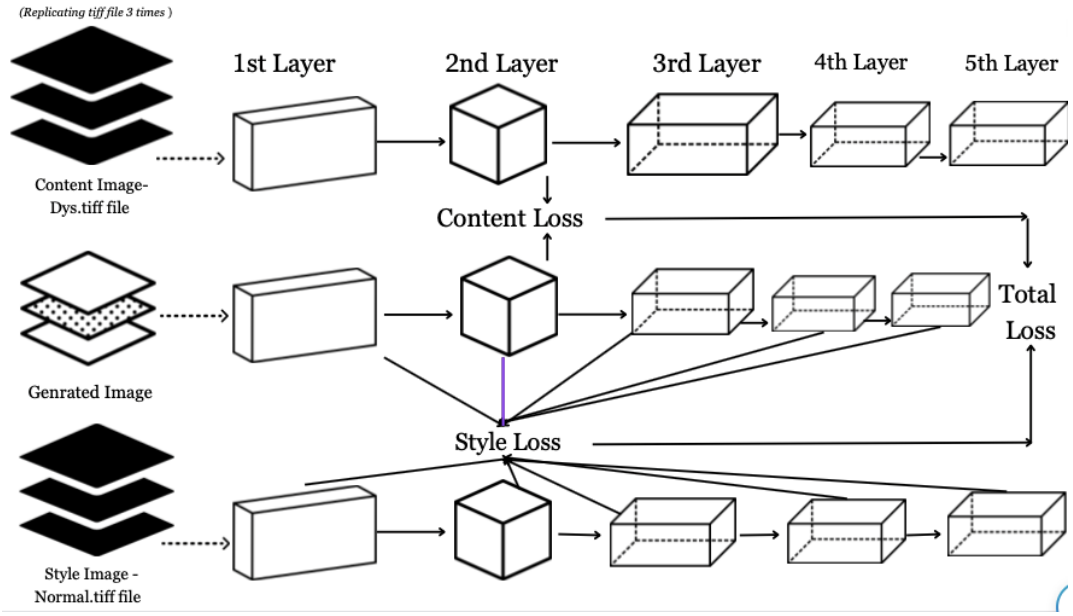

Fig 4.4: VGG16 NST Architecture

NST using VGG19 or VGG16 involves selecting one or more of the convolutional layers to extract feature maps from the content and style images, and then applying the Gram matrix operation to these feature maps. However, because VGG networks are deeper and have more layers than AlexNet, they can extract more detailed features from the input images. This allows them to produce more visually appealing stylized images, due to their ability to learn more complex features.

One of the key differences between VGG19 and VGG16 is their number of parameters. VGG19 has about 144 million parameters, while VGG16 has about 138 million parameters. This difference in the number of parameters can affect the computational resources required for NST using these architectures. Generally, VGG19 requires more computational resources compared to VGG16, due to its larger number of parameters.

Another difference between VGG19 and VGG16 is the number of convolutional layers in each architecture. VGG19 has 16 convolutional layers, while VGG16 has 13 convolutional layers. This difference in the number of convolutional layers can affect the level of detail that can be extracted from the input images. Generally, VGG19 is better suited for tasks that require more detailed features, while VGG16 is better suited for tasks that require a faster computation time.

AlexNet is a good choice for tasks that require less computational resources, but may not produce as visually appealing stylized images compared to VGG networks. VGG19 is better suited for tasks that require more detailed features, but may require more computational resources compared to VGG16. VGG16 is a good choice for tasks that require a faster computation time, but may not be able to extract as much detail from the input images compared to VGG19.

Input: The VGGNet takes in an image input size of 224×224. For the ImageNet competition, the creators of the model cropped out the center 224×224 patch in each image to keep the input size of the image consistent.

Convolutional Layers: VGG's convolutional layers leverage a minimal receptive field, i.e., 3×3, the smallest possible size that still captures up/down and left/right. Moreover, there are also 1×1 convolution filters acting as a linear transformation of the input. This is followed by a ReLU unit, which is a huge innovation from AlexNet that reduces training time. ReLU stands for rectified linear unit activation function; it is a piecewise linear function that will output the input if positive; otherwise, the output is zero. The convolution stride is fixed at 1 pixel to keep the spatial resolution preserved after convolution (stride is the number of pixel shifts over the input matrix).

Hidden Layers: All the hidden layers in the VGG network use ReLU. VGG does not usually leverage Local Response Normalization (LRN) as it increases memory consumption and training time. Moreover, it makes no improvements to overall accuracy.

Fully-Connected Layers: The VGGNet has three fully connected layers. Out of the three layers, the first two have 4096 channels each, and the third has 1000 channels, 1 for each class.

VGG19 is considered to be better than Alex Net in the following aspects

- Deeper network suggests learning of more complex features which is needed for a speech set like dysarthria which is already very difficult to decode.
- Smaller filter size as compared to Alex net. Smaller filters help capture more fine-grained features in an image, leading to better performance.
- VGG19 has a uniform architecture with 3x3 convolutional layers followed by max-pooling layers, making it easier to train and fine-tune

than AlexNet's architecture, which is more complex and varied.

Hence, VGG19 has been deployed to enhance slurred speech caused by dysarthria.

The layers involved in the architecture of VGG19 are as follows:

Let x be the input image or feature map with dimensions (W1, H1, D1).

Convolutional Layers (Block 1):

Conv1_1 = ReLU(Convolution(x, Wconv1_1, Bconv1_1))

Conv1_2 = ReLU(Convolution(Conv1_1, Wconv1_2, Bconv1_2))

Pool1 = Max_Pooling(Conv1_2, Kpool1, Spool1)

Convolutional Layers (Block 2):

Conv2_1 = ReLU(Convolution(Pool1, Wconv2_1, Bconv2_1))

Conv2_2 = ReLU(Convolution(Conv2_1, Wconv2_2, Bconv2_2))

Pool2 = Max_Pooling(Conv2_2, Kpool2, Spool2)

Convolutional Layers (Block 3):

Conv3_1 = ReLU(Convolution(Pool2, Wconv3_1, Bconv3_1))

Conv3_2 = ReLU(Convolution(Conv3_1, Wconv3_2, Bconv3_2))

Conv3_3 = ReLU(Convolution(Conv3_2, Wconv3_3, Bconv3_3))

Conv3_4 = ReLU(Convolution(Conv3_3, Wconv3_4, Bconv3_4))

Pool3 = Max_Pooling(Conv3_4, Kpool3, Spool3)

Convolutional Layers (Block 4): Conv4_1 = ReLU(Convolution(Pool3, Wconv4_1, Bconv4_1))

Conv4_2 = ReLU(Convolution(Conv4_1, Wconv4_2, Bconv4_2))

Conv4_3 = ReLU(Convolution(Conv4_2, Wconv4_3, Bconv4_3))

Conv4_4 = ReLU(Convolution(Conv4_3, Wconv4_4, Bconv4_4))

Pool4 = Max_Pooling(Conv4_4, Kpool4, Spool4)

Convolutional Layers (Block 5):

Conv5_1 = ReLU(Convolution(Pool4, Wconv5_1, Bconv5_1))

Conv5_2 = ReLU(Convolution(Conv5_1, Wconv5_2, Bconv5_2))

Conv5_3 = ReLU(Convolution(Conv5_2, Wconv5_3, Bconv5_3))

Conv5_4 = ReLU(Convolution(Conv5_3, Wconv5_4, Bconv5_4))

Pool5 = Max_Pooling(Conv5_4, Kpool5, Spool5)

Flatten:

FV = Flatten(Pool5)

Fully-connected Layers:

FC1 = ReLU(Fully_Connected(FV, Wfc1, Bfc1))

FC2 = ReLU(Fully_Connected(FC1, Wfc2, Bfc2))

FC3 = Softmax(Fully_Connected(FC2, Wfc3, Bfc3))

4.2 RESULTS

The Itakura Saito distance for various word utterances tested against various convolutional networks architectures are shown below:

Convolutional Network / Word Utterance	Itakura - Saito Distance
AIR (Male)	
Alex Net	1178
VGG16	1129
VGG19	931
DARK (Male)	
Alex Net	2180
VGG16	1457
VGG19	1315
FOXTROT (Female)	
Alex Net	894
VGG16	869
VGG19	757
GADGET(Female)	

Alex Net	1479
VGG16	1469
VGG19	1439

Table 4.1: Itakura-Saito distance

The MOS for various word utterances tested against various convolutional networks architectures are shown below:

Convolutional Network / Word Utterance	MOS	Quality
AIR (Male)		
Alex Net	3	Fair
VGG16	4	Good
VGG19	5	Excellent
DARK (Male)		
Alex Net	1	Bad
VGG16	2	Poor
VGG19	4	Good
FOXTROT (Female)		
Alex Net	2	Poor
VGG16	3	Fair
VGG19	5	Excellent
GADGET(Female)		

Alex Net	1	Bad
VGG16	3	Fair
VGG19	4	Good

Table 4.2: MOS

Linear Predictive Plots for various utterances are shown below:

Air Male:

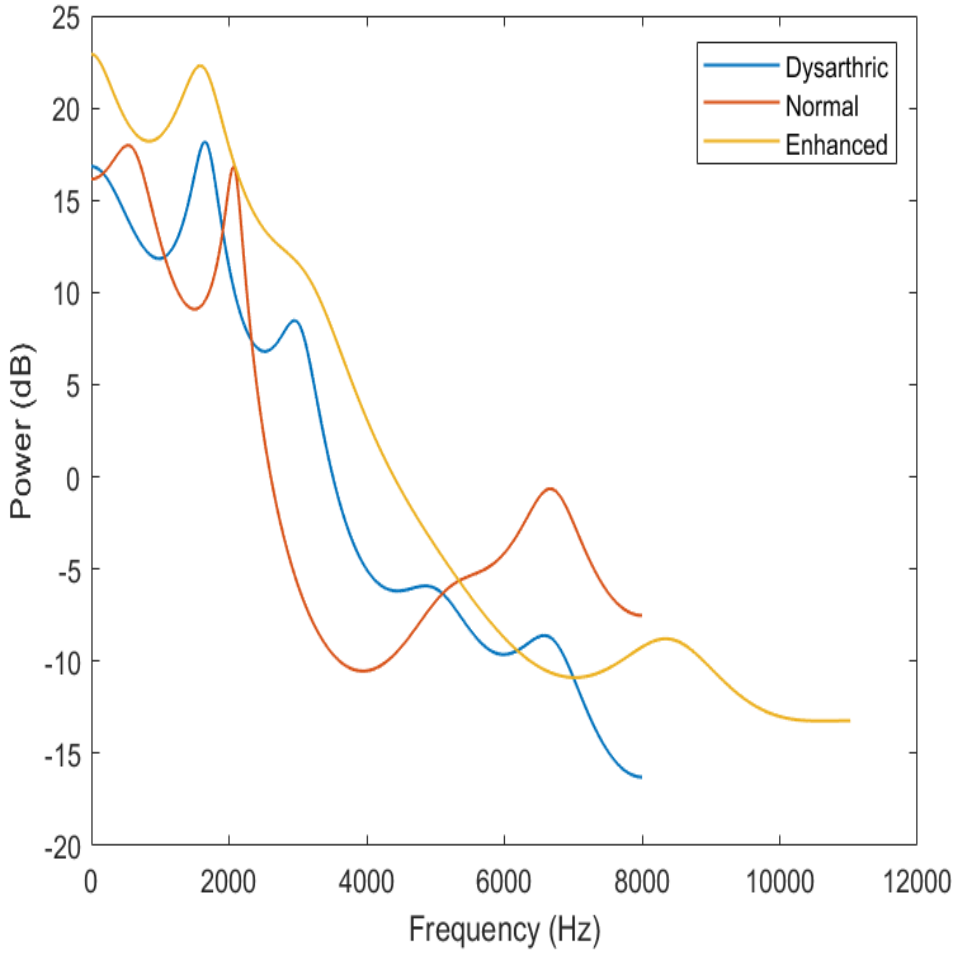

Fig: 4.5 LP Spectra Air male

Dark Male:

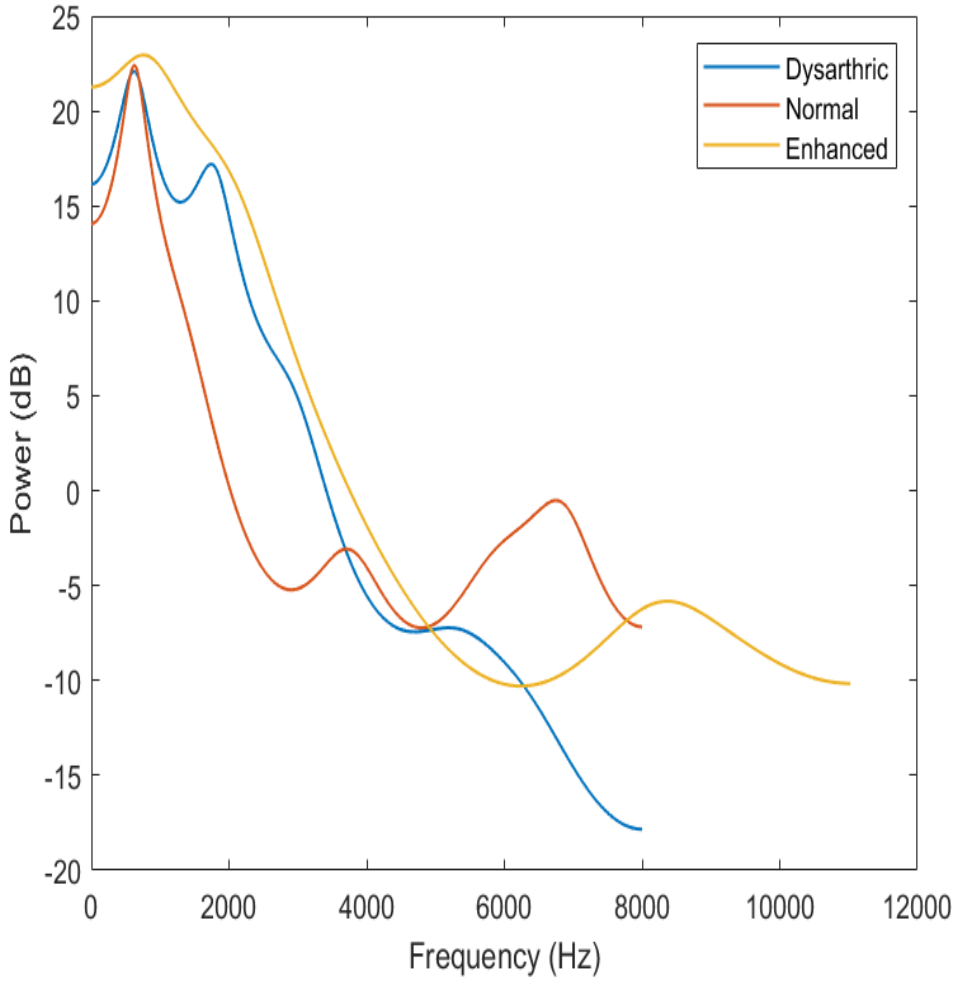

Fig: 4.6 LP Spectra Dark male

Foxtrot female:

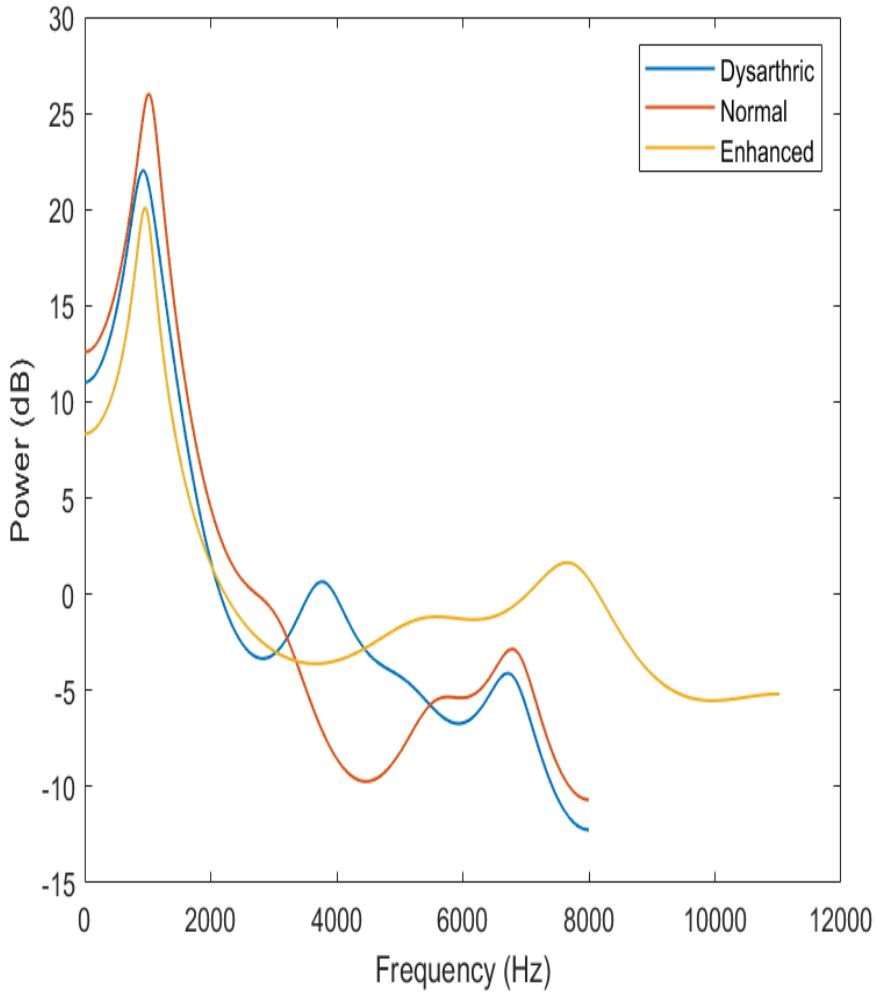

Fig: 4.7 LP Spectra Foxtrot female

Gadget female:

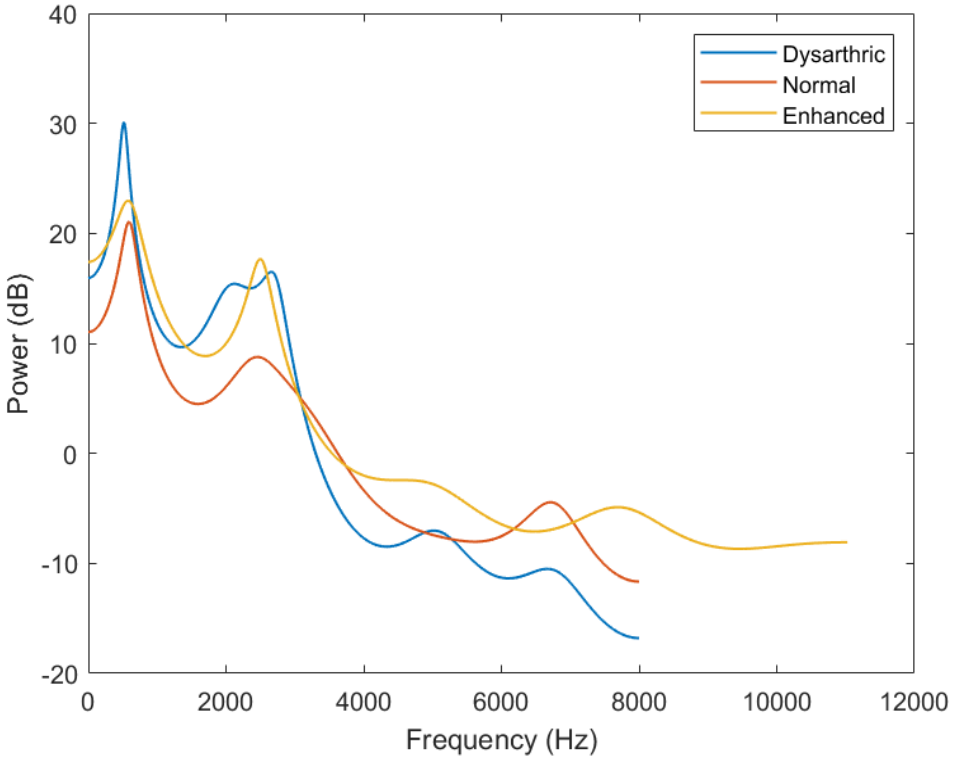

Fig: 4.8 LP Spectra Gadget female

4.3 DISCUSSION

The Itakura distance is a measure used to assess the difference between two probability distributions, especially in the context of frequency spectra. It is a nonsymmetric metric that is often employed to evaluate the effectiveness of models designed to generate results based on these distributions. In this study, we aimed to test the effectiveness of the Itakura distance against three different models for four different word pronunciations to understand how efficiently the results were being generated.

To evaluate the effectiveness of these models, we used the Itakura distance to compare the generated speech signal with the original input speech signal. We tested three different models, namely VGG19, AlexNet, and VGG16, on four different word pronunciations. The first word we examined was 'AIR', spoken by a male. The Itakura distance between this word spoken by a dysarthric person and the ehnaced images was shown to be 931 in the VGG19 model and the itakura distance between the normal speech and enhaced audio is 1129. This shows the minimal gap between these 2 speech sets thereby giving the most optimized results. The highest distance gap was observed in the AlexNet model, with a value of 1178 on the normal end and a value of 849 on the dysarthric end of the spectrum. The VGG16 model showed a distance gap that was centric to both the above mentioned approaches.

These results indicate that the VGG19 model is better suited for enhancing dysarthric speech than the other two models. The lower Itakura distance observed in the VGG19 model suggests that it is better able to capture the features of the original speech signal and generate a more intelligible speech signal. The higher Itakura distance observed in the AlexNet model suggests that it may not be as effective at capturing these features, resulting in a less intelligible generated speech signal.

Next, we examined the word 'GADGET', spoken by a female. The Itakura distance for this word was 1439 in the VGG19 model, with distances of 1479 and 1469 observed in the AlexNet and VGG16 models, respectively. Although the difference may not be significant, the results of running this model against various dysarthric words indicate that the VGG19 model consistently provides the lowest distance. As a result, it is the best fit for the task at hand.

The results of this study suggest that the Itakura distance is a useful metric for assessing the effectiveness of machine learning models designed to enhance

dysarthric speech. The VGG19 model consistently outperformed the other models when it came to reducing the distance between the generated speech signal and the original input speech signal. The depth of the VGG19 model is likely responsible for its ability to capture better features and generate more intelligible speech signals.

These findings have important implications for the development of machine learning models that can help enhance the communication abilities of individuals with dysarthria. By carefully selecting the appropriate model for a given task and using the Itakura distance as a measure of model performance, researchers can develop more effective models that can help dysarthric individuals communicate more effectively. Future research could explore the use of other metrics to assess the effectiveness of these models, as well as investigate other machine learning techniques that could be used to enhance dysarthric speech.

Chapter 6

FUTURE WORK

There is still much work that can be done in the field of dysarthria to normal one-shot learning. One potential avenue for future research is the exploration of transfer learning techniques for improving one-shot learning in this domain. Transfer learning involves training a model on a large dataset of related tasks, then fine-tuning it on a smaller dataset for a specific task. This approach has shown promise in improving one-shot learning in other domains, and could potentially be applied to dysarthria as well.

Another area of future research is the development of more sophisticated neural network architectures that are better suited to handling the unique challenges of dysarthric speech. For example, recurrent neural networks (RNNs) have been shown to be effective at modeling sequential data, and could potentially be used to improve the performance of one-shot learning models on dysarthric speech data. Other types of neural networks, such as convolutional neural networks (CNNs) and transformer-based models, could also be explored for their potential utility in this domain.

Another potential area of exploration is the development of new feature extraction techniques specifically tailored to dysarthric speech. Current approaches often rely on mel-frequency cepstral coefficients (MFCCs) and other common speech features, but these may not capture all of the relevant information in dysarthric speech. New techniques that are more sensitive to the unique characteristics of dysarthric speech, such as changes in articulation and prosody, could potentially lead to more accurate one-shot learning models.

Additionally, more work could be done to collect and curate larger datasets of dysarthric speech for training and evaluation purposes. While there are some datasets currently available, they are often small and limited in scope. A larger and more diverse dataset could help to improve the accuracy and generalizability of one-shot learning models for dysarthric speech.

Finally, there is potential for the integration of other types of data, such as video and physiological signals, into one-shot learning models for dysarthria. Video data could provide additional information about facial and gestural

cues that may be useful in identifying dysarthric speech patterns, while physiological signals such as electroencephalography (EEG) could provide insights into the underlying neural activity associated with dysarthria. By combining multiple modalities of data, it may be possible to improve the accuracy of one-shot learning models and better understand the mechanisms underlying dysarthria.

Recurrence plots are a powerful tool for analyzing time series data, which are frequently used in the study of dysarthria. A recurrence plot is a graphical representation of the recurrence of states in a time series. It provides a visual representation of the temporal patterns of the data, allowing researchers to identify patterns that may not be apparent in other analyses. In the context of dysarthria, recurrence plots can be used to analyze speech signals and identify patterns that are characteristic of the disorder. For example, studies have used recurrence plots to analyze the speech of dysarthric patients and identify patterns of dysfluency or other speech characteristics that are associated with the disorder.

Recurrence plots can also be used to compare the speech of dysarthric patients to that of normal speakers. By analyzing the recurrence plots of both groups, researchers can identify differences in the temporal patterns of the data that may be indicative of dysarthria. One advantage of recurrence plots is that they are able to capture nonlinear relationships in the data. Many traditional statistical analyses assume that the data are linear, which may not be the case in the study of dysarthria.

By using recurrence plots, researchers can identify nonlinear relationships that may be missed by other methods. Another advantage of recurrence plots is that they are relatively easy to generate and interpret. Unlike some other analyses, which may require specialized software or statistical expertise, recurrence plots can be generated using basic software tools and do not require extensive training to interpret.

Overall, recurrence plots are a valuable tool for analyzing time series data in the study of dysarthria. By providing a visual representation of the temporal patterns in the data, they can help researchers identify patterns that may be indicative of the disorder and compare the speech of dysarthric patients to that of normal speakers. With further research, recurrence plots may become an even more powerful tool for understanding dysarthria and developing effective interventions for the disorder.

In summary, there is still much work that can be done in the field of dysarthria to normal one-shot learning. Future research could explore transfer learning techniques, develop new neural network architectures and feature extraction methods, collect larger and more diverse datasets, and integrate multiple modalities of data. With these advancements, it may be possible to develop more accurate and effective one-shot learning models for dysarthric speech, ultimately improving the diagnosis and treatment of this disorder.

Chapter 7

CONCLUSION

To conclude, we have discussed, in this report, the detailed design and related algorithms for a project to transfer the "style" of normal speech onto dysarthric speech, effectively converting it to a more intelligible form. We have shown that a CNN-based NST can be used as a one-shot learner for dysarthric-normal voice conversion. The model can be trained using one paired dysarthric and normal speech data to generate the enhanced utterance. The enhanced audio thus generated retains speaker characteristics of dysarthric speaker while also preserving the content of the original dysarthric speech.

The architecture of the CNN model could be based on VGG19 or a similar architecture that has been shown to be effective for NST. The input to the model would be the mel spectrogram of the dysarthric speech, and the target output would be the mel spectrogram of the corresponding normal speech. The model would be trained using a loss function that encourages the output mel spectrogram to match the target mel spectrogram. The mel spectrogram can then be converted back to waveform using the Griffin Lim Algorithm, and the resulting speech can be evaluated using objective measures such as the Itakura-Saito distance and subjective measures such as MOS tests.

Thus, this experimental set up could make VC more accessible and practical for use in clinical settings, where resources are often limited. Secondly, the model could be trained to generalize to a wider range of speech styles and speakers. Thirdly, CNN-based NST could allow for greater flexibility in the conversion process. This can limit their ability to handle situations where a specific source or target speaker is not available. By using CNN-based NST, it is possible to transfer the style of any normal speech onto dysarthric speech, without the need for a specific source or target

APPENDIX

Neural Style Transfer

Neural Style Transfer (NST) is a machine learning technique used to blend two images, where one image serves as the content and the other as the style. It works by optimizing a cost function that aims to preserve the content of the original image while incorporating the style of the reference image. The result is a new image that retains the content of the original but is rendered in the style of the reference image.

VGG19

VGG19 is a convolutional neural network architecture used for image classification tasks. It consists of 19 layers, including 16 convolutional layers, 3 fully connected layers, and max-pooling layers. The network has a uniform architecture, with a fixed 3x3 filter size for all convolutional layers. It achieved state-of-the-art results in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in 2014. VGG19 is widely used as a base network for transfer learning and has been applied in various computer vision tasks, including object detection and segmentation.

VGG16

VGG16 is a convolutional neural network architecture used for image classification tasks. It consists of 16 layers, including 13 convolutional layers, 3 fully connected layers, and max-pooling layers. The network has a uniform architecture, with a fixed 3x3 filter size for all convolutional layers. It achieved state-of-the-art results in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in 2014. VGG16 is widely used as a base network for transfer learning and has been applied in various computer vision tasks, including object detection and segmentation.

AlexNet

AlexNet is a convolutional neural network architecture used for image classification tasks. It consists of 8 layers, including 5 convolutional layers, 3 fully connected layers, and max-pooling layers. It was the winning model in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in 2012. AlexNet introduced several key innovations, including the use of

Rectified Linear Units (ReLU) activation functions, dropout regularization, and data augmentation. It revolutionized the field of computer vision and paved the way for the development of deeper and more complex neural network architectures.

Convolutional Neural Network

A Convolutional Neural Network (CNN) is a type of neural network used primarily for image and video processing tasks. It consists of multiple layers of neurons that perform convolution operations on the input data, followed by pooling layers that reduce the size of the output. CNNs are designed to automatically learn features from the input data, enabling them to extract meaningful patterns and structures from images and videos. They have revolutionized the field of computer vision and are widely used in applications such as object detection, segmentation, and recognition.

Mel Spectrogram

Mel Spectrogram is a visualization of the frequency components of an audio signal. It is a 2D representation of the power spectral density of the audio signal, where the frequency domain is divided into multiple frequency bands using a Mel scale. Mel spectrograms are commonly used in speech and music analysis applications, such as speech recognition and genre classification. They allow researchers to analyze the frequency content of an audio signal, enabling them to extract meaningful features and patterns for further analysis.

Itakura-saito distance

Itakura-Saito Distance (ISD) is a distance metric used to compare two power spectra of signals. It is commonly used in speech processing and music information retrieval. The ISD measures the spectral distortion between two power spectra by computing the geometric mean of the ratio of the two spectra. The ISD is based on the auditory masking properties of the human ear, and it is particularly suited for measuring the perceptual differences between speech or music signals. It has applications in speech enhancement, speaker identification, and music genre classification.

Mean Opinion Score

Mean Opinion Score (MOS) is a method used to measure the quality of audio or video signals in subjective terms. MOS is obtained by conducting subjective listening tests where a group of human listeners rate the quality of the audio or video signals on a scale from 1 to 5 or 1 to 10. The mean value of the ratings is then computed, resulting in a MOS value that represents the overall quality of the signal. MOS is widely used in speech and audio processing applications to evaluate the performance of algorithms and systems.

Griffin Lim Algorithm

The Griffin-Lim algorithm is an iterative algorithm used for phase retrieval in audio signal processing. It is commonly used in the reconstruction of audio signals from their magnitude spectrograms. The algorithm works by initializing a random phase spectrogram and repeatedly transforming it back to the time domain to obtain a reconstructed audio signal. The reconstructed signal is then transformed back to the magnitude spectrogram, and the phase is updated based on the original magnitude spectrogram. This process is repeated until convergence, resulting in a reconstructed audio signal with a phase that is consistent with the original signal. The Griffin-Lim algorithm is widely used in speech and music processing applications, such as speech enhancement and music source separation.

Spectral Gating

Spectral gating is a signal processing technique used to selectively attenuate or remove certain frequency bands of an audio signal. It involves dividing the audio signal into frequency bands using a filter bank, and then applying a gating function to each frequency band to control the gain. The gating function is typically a threshold-based function that sets the gain of each frequency band to zero when the energy of the band falls below the threshold. Spectral gating is commonly used in audio processing applications such as noise reduction and speech enhancement, where it can be used to selectively remove unwanted background noise or reverberation from an audio signal.

Gradient Descent

Gradient descent is an optimization algorithm used to minimize the error or cost function of a machine learning model. It works by iteratively adjusting the model parameters in the direction of the steepest descent of the cost function, as determined by the gradient of the cost function. The gradient represents the direction of the fastest increase of the cost function, so moving in the opposite direction (negative gradient) should result in the fastest decrease of the cost function. Gradient descent is widely used in deep learning algorithms, such as artificial neural networks, to train models by minimizing the error between the predicted and actual output. There are different variations of gradient descent, including batch gradient descent, stochastic gradient descent, and mini-batch gradient descent, each with its own advantages and disadvantages.

Adam Optimization

Adam optimization is an algorithm used to update the parameters of machine learning models in an efficient manner. It is a variant of stochastic gradient descent (SGD) that adapts the learning rate for each parameter individually, based on the historical gradient and update steps. Adam combines the advantages of both the Adagrad and RMSprop optimization algorithms by using the first and second moments of the gradient to compute adaptive learning rates. This allows Adam to converge faster and with more accuracy than other optimization algorithms, especially in large-scale deep learning models. Adam is widely used in deep learning applications, such as image recognition, natural language processing, and speech recognition.

Fourier Transform

Fourier transform is a mathematical technique used to transform a signal from the time domain to the frequency domain. It works by representing the signal as a sum of sine and cosine waves of different frequencies and amplitudes. The Fourier transform allows us to analyze the frequency content of a signal and identify its underlying components. It has wide applications in signal processing, such as in audio and image processing, communication systems, and data compression. The inverse Fourier transform is used to transform a signal back to the time domain from the frequency domain. The discrete Fourier transform (DFT) and fast Fourier transform (FFT) are commonly used algorithms to compute the Fourier transform of digital signals.

Mel Filter Bank

A Mel filterbank is a set of triangular filters that are used to extract features from the frequency domain of an audio signal. The filters are spaced based on the Mel scale, which is a perceptual scale of pitches judged by listeners. The Mel filterbank is applied to the magnitude spectrum of the audio signal after it has been transformed from the time domain to the frequency domain using the Fourier transform. Each filter in the Mel filterbank covers a specific frequency range and emphasizes the energy within that range. The output of the Mel filterbank is typically used as input to a Mel frequency cepstral coefficients (MFCC) analysis, which is a widely used technique for speech and audio signal processing.

Dysarthria

Dysarthria is a motor speech disorder that affects the muscles used for speech production. It is caused by damage or injury to the parts of the brain or nervous system that control the muscles used for speech, such as the lips, tongue, vocal cords, and diaphragm. Dysarthria can cause slurred or slowed speech, difficulty with articulation, and changes in voice quality or pitch. It can be caused by a variety of conditions, such as stroke, traumatic brain injury, Parkinson's disease, cerebral palsy, and multiple sclerosis. Treatment for dysarthria may include speech therapy, physical therapy, medication, and surgery, depending on the underlying cause and severity of the condition.

REFERENCES

[1] Chen, A., & Garrett, C. G. (2005). Otolaryngologic presentations of amyotrophic lateral sclerosis. Otolaryngology-Head and Neck Surgery, 132, 500–504.

[2] Müller, J., Wenning, G. K., Verny, M., McKee, A., Chaudhuri, K. R., Jellinger, . . . Litvan, I. (2001). Progression of dysarthria and dysphagia in postmortem-confirmed Parkinsonian disorders. Archives of Neurology, 58, 259–264.

[3] Dutta, R., Tsiartas, A., Elsner, M., Frank, M., et al. (2020). Whispered speech recognition: A review. Computer Speech & Language, 65, 101086. https://doi.org/10.1016/j.csl.2020.101086

[4] Pascual, S., Bonafonte, A., Serra, J., et al. (2017). SEGAN: Speech enhancement generative adversarial network. In Proceedings of the 2017 Conference on Neural Information Processing Systems (NIPS) (pp. 569-579).

[5] Liu, R., Tang, Y., Weninger, F., et al. (2019). Iterative generative adversarial networks for robust speech enhancement. In Proceedings of the

2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 701-705)

[6] Zhu, J.-Y., Park, T., Isola, P., et al. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV) (pp. 2242-2251).

[7] Gatys, L. A., Ecker, A. S., & Bethge, M., et al. (2016). Image style transfer using convolutional neural networks. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 2414-2423).

[8] Wang, Disong, et al. "End-to-end voice conversion via cross-modal knowledge distillation for dysarthric speech reconstruction." ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020.

[9] Chen, Chen-Yu, et al. "Enhancing Intelligibility of Dysarthric Speech Using Gated Convolutional-Based Voice Conversion System." Interspeech. 2020.

[10] Yang, Seung Hee, and Minhwa Chung. "Improving dysarthric speech intelligibility using cycle-consistent adversarial training." arXiv preprint arXiv:2001.04260 (2020).

[11] Huang, Wen-Chin, et al. "Towards identity preserving normal to dysarthric voice conversion." ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022.

[12] Zhao, Wei, et al. "IVCGAN: An Improved GAN for Voice Conversion." 2021 IEEE 5th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC). Vol. 5. IEEE, 2021.

[13] Sisman, Berrak, et al. "On the study of generative adversarial networks for cross-lingual voice conversion." 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). IEEE, 2019.

[14] Gatys, Leon A., Alexander S. Ecker, and Matthias Bethge. "Image style transfer using convolutional neural networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[15] Yan, Xiyu, et al. "Neural style transfer with content discrimination." 2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW). IEEE, 2019

[16] AlBadawy, Ehab A., and Siwei Lyu. "Voice Conversion Using Speech-to-Speech Neuro-Style Transfer." Interspeech. 2020.

[17] An, Xiaochun, Frank K. Soong, and Lei Xie. "Disentangling style and

speaker attributes for tts style transfer." IEEE/ACM Transactions on Audio, Speech, and Language Processing 30 (2022): 646-65

[18] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105).

[19] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.