RESEARCH ARTICLE OPEN ACCESS

An Analysis on Enhancing Speech Clarity for Cerebral Palsy Individuals

Pradeep Kumar KG[*], Rachitha M N[**], Sameeksha S Poojary[**], Sandra S Pillai[**], Yashaswini A G[**]

[*]Faculty, Department of CSE, Vivekananda College of Engineering & Technology, Puttur, Karnataka

[**]Students, Department of CSE, Vivekananda College of Engineering & Technology, Puttur, Karnataka

ABSTRACT

Cerebral palsy is a neurological disorder that impairs movement and muscle coordination, often leading to speech difficulties that hinder effective communication and social interactions. These speech impairments, caused by the lack of motor control, vary in severity, making traditional therapeutic approaches less efficient and inconsistent in their outcomes. Verbal therapy for those whose suffer from mental illness generally requires long-term intervention, but the results can be unpredictable and slow. With recent advances in intelligent systems and algorithmic learning, automation technology now has the power to address these issues more effectively. Due to this procedural redundancy, data-based models can sift through new samples of your speech to pinpoint where you struggle and make adjustments in real time to help even out your delivery much faster and more consistently than traditional persons. Machine-learning routines using algorithms such as the k-Nearest Neighbors (KNN) have the ability to classify both new and previous speech samples and apply algorithmic adjustments to correct for troubled areas of speech as they happen. This makes the intervention much more automatic and faster to elevate the overall well-being for individuals surviving with cerebral palsy.

Keywords — Cerebral Palsy, Machine Learning, KNN, Android Application, Speech Enhancement.

I. INTRODUCTION

Cerebral palsy is a bunch of neurological disorders characterized by abnormal muscle tone, poor balance and coordination, and impaired movement. The main factor contributing to the significantly poor standard of life for people with cerebral palsy (CP) is the consequence of motor impairments on communication. When you struggle with saying what you want to say, it's harder to fulfil that most fundamental requirement of being human – perhaps even the most important – quality to communicate with others and engage in social interactions. Speech impairments in CPP are motor disorders; in other words, the origin of difficulty speaking clearly resides in muscles that are weak, abnormally flexible or otherwise compromised. The consequence of such physical constraints may vary from the relatively mild – such as requiring slight tweaks to pronunciation – to the more profound (eg, almost unintelligible speech). Of course, anything to improve people's speech clarity should be pursued. This is because, among other reasons, speaking clearly makes us more empathetic towards others: slow and clear speech conveys our care and concern. In the past, the effort to enhance the quality of speech in CPP involved a great deal of typical, therapy-driven work; namely, doctors spending their time with patients, directly working on muscles to improve speech clarity. Exercises include overcoming resistance (repeatedly opening a door), saying words that improve coordination (saying 'up' and 'down' with specific timing), and craniofacial adjustments to improve speech. All of these efforts have their drawbacks. First, they are time-consuming. Additionally, the outcomes of traditional therapy can change according to the individual's specific condition, and progress is not always consistent.

For those affected by cerebral palsy, for example, it can take a considerable amount of time before they experience any significant improvements in speech clarity resulting from therapy. In recent times, advances in technology – specifically, artificial intelligence and machine learning – have begun to offer new ways of addressing these issues. Machine learning procedures may be utilized to automate speech analysis, when and how speech clarity is impaired. These techniques might generate faster, more robust results than manual analysis alone. Automating parts of the process could cut down drastically the amount of time and effort that needs to be spent on improving speech, and still yield similar results.

A speech-clarity-enhancing system based on the k-Nearest Neighbours (KNN) algorithm can apply several of the criteria above to improve speech intelligibility. KNN is commonly utilized for classification purposes. It operates by comparing new data points to a training set comprising known examples. Subsequently, new points are classified based on their similarity to the k-nearest neighbours in the training set. This approach can be used to compare new speech inputs with previously labelled examples: a KNN model can classify which examples the new input is most similar to. For the task of speech-clarity enhancement, KNN can be exploited to categorize speech samples on different fuzzy classes of intelligibility. If we have a dataset of speech sound from cerebral palsy patients, we can create annotation along a fuzzy continuum of clarity. So, a person rating the audio data can annotate each speech sample with a level of intelligibility. Next, we can train a KNN model on this dataset using the annotated samples as training examples. The KNN model can

then infer when speech is unclear and can apply corrections to these samples based on clearer versions. We must begin by extracting the most significant features of the speech signals. Speech is a multichannel time-variant signal that contains data regarding its source, content, context, situation etc. However, not every feature of speech are needed, or useful, for intelligibility. So, we create feature vectors consisting of certain features that is applicable to enhance clarity. One such feature is pitch: pitch is the perception of the speech signal's frequency of change, and it has a significant effect on intelligibility. Another feature is duration, which is the measure of the segment of the speech signal. Duration can be problematic for intelligibility as overly long or short segments are difficult to process. Similarly, amplitude (the volume of the speech signal) can affect intelligibility by altering sensory measures of speech. Formants are the resonant frequencies produced by the vocal tract; it specializes in distinguishing speech sounds and is vital for clarity. Let us say we have all these features and encoded them for a machine. These features, whether segmental or suprasegmental, can be considered as feature vectors. We can then feed them to the KNN model. The model can classify new speech input by comparing them to the k-nearest neighbours in the dataset. If the model can identify that new speech input is similar to other unclear examples, then simple corrections can be applied based on clearer examples. For instance, if the speech is slow, we could speed up various segments in the new input by training on clearer examples to bring it closer to typical, clearer speech. If the pitch is excessively high or too low, then corresponding adjustments attainable to bring the speech to typical ranges of pitch if it is too monotonous.

Instead of classifying speech samples, the KNN model is used to correct unclear segments of speech by performing in real time some sort of speech-feature editing. For example, segments of an utterance that were deemed unclear would have their speech features (such as pitch values, timing, amplitude) edited to make them sound clearer (while still retaining the speaker's style of speaking), and this process could be completely automated. (In this regard, it is similar to the BLE Teleprompter.) Unlike BLE, which still requires ultimate human approval, this type of feedback and correction could happen in real time and thus could speed up the learning and improvement process. As another benefit, the computational nature of this approach potentially allows for use with machine-learning techniques, which could take the automation to new levels. The ability to make inferences from samples of unlabelled speech data and then provide feedback or correction could drastically lower the costs of providing speech clarity enhancement for citizens with cerebral palsy, especially those in areas where traditional face-to-face and hands-on therapy is not easily available. Critics could argue that the existence of machine-learning algorithms that enhance speech clarity helps us see the unique conditions of our time in a way that might also be influential. They might urge us to approach more recent technologies, such as machine learning with some humility and to avoid treating them as solutions to everything that ails us.

More importantly, these models, such as the KNN machines mentioned above, get better all the time by feeding them more data: the speech samples get bigger and better, and the labeling improves until the model recognizes not just the examples it was initially trained with, but similar samples having a wider range of patterns and impairments. So, in addition to benefiting the cerebral palsy population, the system could get more accurate and light-touch with more use, providing better results in years to come. Automating cue enhancement will have positive impacts on communicative competence, individual and family fulfilment, and QoL. A major reason for communicative behavior is to socialize, and coherent speech is a key method for doing so. Current difficulties understanding the speech of individuals with CP limits their education, employment and, likely, their QoL. Improved speech clarity as a result of machine learning could bring greater independence to those with CP, widen opportunities for social interaction, and enhance QoL. At the same time, machine learning automates the task of cue enhancement to be more efficient and effective than traditional methods operating alone.

II. REVIEW OF LITERATURE

By eliminating the requirement for time alignment in discrete utterance recognition, Shore et al.[1] groundbreaking method, they use information-theoretic spectral distortion measurements, minimal cross-entropy pattern classification, and vector quantization for rate-distortion voice coding. For a vocabulary of 20 words, they generate individual vector quantization codebooks. They achieve a remarkable 99% accuracy rate for speaker-dependent recognition, which includes the 10 digits, and an 88% accuracy rate for speaker-independent recognition. The authors show that significant performance can be achieved without time sequence information, refuting the widely held notion that it is essential for voice recognition. Their results highlight the importance of cross-entropy minimization in classification efficacy and make important connections to information theory through the use of the Itakura-Saito distortion measure. Notably, they imply that accurate codebooks can be created with fewer codewords, which encourages more research into methods to boost recognition performance. Their preliminary findings point to the possibility of improvements through the selective incorporation of time-sequence data, suggesting more study into identifying typical codeword trajectories to maximize classification precision and computational effectiveness. In addition to making a contribution to the field of voice recognition, this study lays the groundwork for further research at the nexus of speech technology and information theory.

In their exploration of the complexities of invariable approximation in the context of speech recognition, Lalit R. Bahi et al. [2] offer a technique based on maximizing mutual information (MMI) between related word sequences and

acoustic data. They start by explaining the standard design of speech recognition systems, in which a linguistic decoder converts voice waveforms into a word sequence after an acoustic processor produces a series of noteworthy features. The authors draw attention to the shortcomings of conventional maximum likelihood estimation (MLE) techniques, which, since they rely on oversimplified assumptions about probability distributions, may find it difficult to appropriately describe the link between spoken words and acoustic data. They respond by presenting their MMI estimation method, which deliberately optimizes the mutual information and improves the model's capacity to precisely represent the relationship between the characteristics and the desired word sequence. A momentous rise in the log probability of the correct script in training data and a noticeable drop in recognition errors during testing are two examples of experimental results that demonstrate the efficacy of this approach. In particular, their MMI method improves the log probability from -190.11 (using MLE) to -1.16, which translates to 64 recognition mistakes as opposed to 78 when using regular MLE. In addition to demonstrating how MMI can improve recognition performance, this study expressed, critical improved estimating methods are to the continuous advancement of reliable speech recognition systems.

The complexities of building high-accuracy continuous voice identification methodologies with Hidden-Markov-Models(HMMs) are examined by Young et al. [3], who concentrate on the twin problems of model complexity and sparse training data. The enormous number of possible triphones in big vocabulary systems, particularly those requiring cross-word context-dependent modeling, can result in a serious data insufficiency problem, the authors note. In order to address this, they present a tree-based state tying technique that makes use of phonetic decision trees. This technique efficiently groups related states according to the phonetic context, improving the model's capacity to generalize across unseen triphones. In addition to preserving the acoustic deviations that are indispensable for precise speech recognition, clustering technique makes sure that the training data is utilized as efficiently as possible. By contrasting their tied-state methodology with more conventional model-based techniques maintain computing efficiency while achieving better recognition results. The technique enables flexibility in modeling complex speech patterns and adapting the model's complexity according to available data by using continuous density mixing Gaussian distributions. By presenting experimental data from the Resource Management and Wall Street Journal assignments, their method simplifies the model creation process while producing state-of-the-art outcomes. They stress the fundamental significance that HMM techniques have played in the development of speech recognition technologies and end with a request for more study to improve these methods.

ASR paradigm supported on auditory-scene-analysis(ASA), which focuses on isolated speech signs from background noise, is proposed by Cooke et al. [4]. They address the difficulties of recognizing occluded speech to investigate approaches to machine-driven speech recognition. Two main approaches are examined by the authors: one uses unsupervised learning to handle missing-component spectral vectors, while the other modifies the Viterbi algorithm to handle occluded speech. The result demonstrated that incomplete information can be efficiently used for voice recognition, as the system can continue to conduct recognition even when more than 50% of the observation vector is occluded. Similar to how babies create auditory-phonetic representations in noisy surroundings, this study shows how voice recognition algorithms can be trained with partial data.

When insufficient adaptation data is available, Kuan-ting Chen et al. [5] improve upon the conventional maximum likelihood linear-regression (MLLR) procedures with their novel eigenspace-based approach to quick speaker adaptation. In order to develop eigen-matrices, the main concept is to use the prior knowledge gained by training speakers using PCA to build an eigenspace for MLLR's full regression matrices. The suggested approach successfully lowers the number of free parameters while preserving an accurate depiction of inter-dimensional correlations among model parameters by limiting the regression matrices of external speakers within the span of the first k eigen-matrices. Specifically, the eigenspace-based MLLR significantly outperformed conventional techniques with only 10 seconds of adaptation data, indicating its potential to improve speaker adaptation in speech recognition systems.

Tan Lee et. al[6] presented CU Corpora, to create large speech corpora for Cantonese, a dialect spoken by more than 60 million people, mostly in southern-china and Hong Kong. The dearth of significant spoken-language-resources for cantonese is addressed in this work, which is important considering the growing need for voice synthesis and recognition technologies in a variety of applications. Being the first extensive datasets created to facilitate Cantonese speech processing, the CU Corpora are special because they cover a wide variety of linguistic units, including continuous sentences, polysyllabic words, and isolated syllables. By developing corpora that comprise both general-purpose datasets with rich phonetic content (CUSYL, CUWORD, CUSENT) and application-specific data, including command recognition (CUCMD) and digits (CUDIGIT), the authors highlighted the significance of phonetic diversity. Every dataset was recorded in controlled settings and rigorously annotated by hand, guaranteeing the thorough orthographic and phonetic details required to create text-to-speech (TTS) and automatic speech recognition (ASR) systems with high performance. The authors proved that the CU Corpora offer a balanced representation of Cantonese speech by a comprehensive statistical study of the phonetic content. This is essential for the efficient modeling of a variety of speech occurrences. The creation of a connected-syllable recognizer, a broad vocabulary continuous speech recognizer, and a TD-PSOLA TTS synthesis system demonstrated the usefulness of these corpora. To improve the capabilities of multilingual

spoken language technologies and foster more efficient human-computer interaction in a mixture of cantonese-speaking contexts, CU Corpora are necessary, they highlight the unique phonological characteristics of cantonese and offer a wealth of resources for research and technology development. These resources accessibility for both academic and commercial purposes greatly advances the field of speech technology and encourages more advancements in cantonese voice processing. Through their dedicated website, CU Corpora's comprehensive documentation and license policies are accessible, promoting wider use and participation in the industry and research community.

Seltzer et al[7] made a substantial contribution to the field of speech recognition. Their work tackles the crucial problem of speech recognition systems noise degradation, which frequently results in subpar performance when trained only on clean speech data. The methodology may adjust to changing conditions by putting forth a technique that categorizes the dependability of spectrographic components without imposing limiting assumptions regarding the type of corrupting noise. By differentiating between trustworthy and faulty areas of the spectrogram, the method enables more precise identification even in difficult situations. The study underlines the necessity for adaptational algorithms for considering dynamic character of real-world audio and draws attention to the shortcomings of conventional noise estimating techniques, which are predicated on stationary noise assumptions. The bayesian classifier's durability and efficacy were demonstrated by experimental findings that demonstrated its performance over traditional mask estimation techniques across a range of noise types and signal-to-noise ratios. The authors also address the intricacy of their methodology because of the large number of parameters involved and recommend that future studies concentrate on optimizing these parameters, especially in situations like music where speech and harmonic noise overlap.This supports an integrated approach that combines mask estimation with recognition performance optimization.

In order to provide a standardized, five-level classification of gross motor function in kids with CP, Jan Willem Gorter et al[8] extended the Gross Motor Function Classification System(GMFCS). This system is intended to direct therapy, research, and professional and caregiver communication. From Level-I, which indicates children who can walk freely, to Level-V, which includes children who require substantial assistive technology or physical support for mobility, each level in the GMFCS represents unique functional skills. The framework is extremely relevant across various age groups and motor development stages because it places a higher priority on functional activities than age-based milestones. In order to guarantee consistent classification across practitioners, they used extensive inter-rater reliability and validity assessments, which included a sizable and diverse cohort. This assured the GMFCS appropriately captures age-related and functional differences in motor abilities. Because of its reliability, the GMFCS may be a significant tool in clinical and research settings for tracking development, establishing personalized treatment plans, and improving results for kids with cerebral palsy. As a worldwide standard, the GMFCS now improves communication and treatment quality throughout the CP community by facilitating a common understanding of motor abilities and aligning rehabilitation programs.

CU2C is a dual-condition cantonese voice database created by Nengheng Zheng et al[9] especially for speaker recognition studies. With the use of this task-oriented database, which includes recordings of cantonese digit strings, Hong Kong ID numbers, and entire phrases, speaker identification systems for a variety of applications can be developed. CU2C is notable for its parallel data collection under two different acoustic conditions: a wideband desktop microphone and a public fixed-line telephone channel. Because it enables researchers to evaluate how various recording conditions affect recognition accuracy, this dual-condition methodology is essential for researching channel effects in speaker recognition. A large dataset for analysis is provided by the database, which includes recordings from 84 target speakers and 23 imposters. Each speaker contributed 18 sessions over a period of 4 to 9 months. The significance of appropriately gathered and annotated voice databases for the advancement of spoken language technologies in cantonese is further supported by preliminary evaluations, which showed that the baseline performance attained using CU2C is comparable to similar databases in other languages. The study also highlighted the need for varied datasets that replicate real-world situations in order to improve speaker recognition systems resilience.

Hong-Kwang Jeff Kuo et al[10] take a novel approach to speech recognition by moving away from the conventional bayesian framework that depends on Hidden Markov Models(HMMs). They suggest the maximum entropy Markov model (MEMM), which makes it easier to include overlapping and asynchronous characteristics by directly calculating the probability of a state or word sequence from observation sequences. Even when utilizing traditional acoustic features, this innovative modeling technique allows for the efficient use of a wide range of linguistic information, leading to a notable decrease in word mistake rates when compared to HMMs. Additionally, when paired with HMM and language model scores, the MEMM shows encouraging results, indicating that the direct modeling strategy not only improves performance but also provides flexibility in feature integration. The discriminative character of the model, which enables enhanced robustness in voice recognition, is credited by the authors with its greater performance. They support more research to add more contextual and suprasegmental characteristics to the MEMM, which could improve speech recognition systems' accuracy and versatility.The work, funded by the DARPA Babylon voice to Speech Translation Program, pushes the limits of traditional voice recognition techniques and creates opportunities for more efficient acoustic modeling techniques.

The effects of extensive dysarthria therapy on older children with cerebral palsy were examined by Pennington et al[11], who concentrated on the speech, language, and communication difficulties that severely restrict these children's social and educational chances. Six participants in the trial, ages 10 to 18, received individualized therapy designed to maximize overall effort in speech production, maintain consistent speech volume, and improve breath support. Assessments of intelligibility in single words and connected speech were used in this focused approach. They were performed prior to therapy, right after, and seven weeks later. The findings showed significant increases in single-word intelligibility but insignificant increases in connected speech, underscoring the complex nature of dysarthria and the varied effects of CP on speech ability. To promote meaningful speech development, the researchers underlined the necessity of specialized therapeutic approaches that go beyond conventional classroom settings and incorporate intensive practice. The study also brought up important issues about the best frequency and intensity of therapy, participant characteristics that may affect results, and reliable methods for evaluating speech intelligibility. To investigate these factors and enhance therapeutic strategies, the authors contend that thorough follow-up studies are essential.

William Byrne[12] explores sophisticated approaches to improve large-vocabulary continuous-speech-recognition(LVCSR) systems. He highlights the importance of minimal risk approximation and decoding strategies, especially with lattice sectionalization techniques, which improve the approximation of parameters in HMMs while enabling the designation of smaller, more manageable recognition tasks. The study challenges the traditional dependence on HMMs and proposed that better performance in LVCSR systems may result from a move toward more creative modeling and decoding techniques. In his explanation of the statistical modeling framework utilized in automatic speech processing, Byrne emphasizes the necessity of striking a balance between linguistic and acoustic models in order to accomplish successful recognition. He highlights two crucial strategies that allow the system to reduce recognition risk rather than only maximize likelihood: lattice segmentation and Pinched-Lattice-Minimum-Bayes Risk-Discriminative-Training(PLMBRDT). The research points out that robust performance requires an appropriate training set size and composition. Byrne comes to the conclusion that his suggested approaches offer a viable path toward creating fresh approaches that might greatly improve the effectiveness and precision.

A thorough examination of HMMs and its fundamental function in contemporary large vocabulary continuous-speech-recognition (LVCSR) systems is given by Mark Gales et al[13]. They start by going over the architecture of HMM-based recognizers, highlighting the important parts that process and interpret spoken language together, like language models, acoustic models, and decoding algorithms. The authors point out the intrinsic drawbacks of simple HMM implementations, which frequently fail to adjust to a range of speaker traits and contextual circumstances. To defeat these obstacles, they support advanced improvements such as gaussian-mixture-models, which improve the model's capacity to faithfully represent the elaboration of speech signals in various settings. The grandness of feature extraction methods is covered in detail, with a focus on Mel Frequency Cepstral Coefficients (MFCCs), especially helpful for capturing the spectrum characteristics for speech analysis. With a direction on their significance in maximizing model performance using exacting data-driven strategies, training methodologies like maximum likelihood estimation and discriminative training approaches are also examined. To ensure that the models can adapt to individual variations and noise, the paper also discusses adaption mechanisms, both speaker-specific and environmental, for preserving system reliability in practical applications. The utility of noise resilience approaches, including as spectrum subtraction and model combination procedures, in improving recognition accuracy in difficult-to-hear conditions is investigated. In order to improve overall performance, multi-pass recognition architectures are introduced as a way to repeatedly refine recognition results by utilizing various models.

Gorter JW, et al[14], who emphasized the need for reclassification at age 2 or older. 77 infants with CP of different forms, including unilateral spastic, bilateral spastic, and dyskinetic types (mean age 19.4 months; 41 boys, 36 girls) participated in the study. A moderate general agreement (linear weighted kappa = 0.70) was found when researchers assessed the stability of GMFCS classifications across time. They also observed that 42% of the children's GMFCS levels changed, with the majority shifting to less functional categories. For newborns who were first placed in GMFCS Levels I, II, and III, the positive predictive value was a noteworthy 96%, suggesting a high probability of staying in the same classification throughout early childhood. However, the study found that newborns initial classifications were less accurate than those of older children, indicating that reclassification becomes increasingly important as more clinical data becomes available. This recognizes the essential evaluation at or after the age of two while highlighting the significance of using the GMFCS early on. In the end, the findings support a dynamic and responsive classification system that reflects the changing nature of each child's condition and better supports their developmental trajectory. They emphasize the necessity for clinicians to modify their classification approach as individual circumstances change, particularly in light of emerging associated conditions like epilepsy.

A thorough registry-based study by Guro Andersen et al[15] looked at the use of Augmentative and Alternative Communication (AAC) and the prevalence of speech issues in Norwegian children with cerebral palsy (CP). According to the study, which used data from the Norwegian CP Registry, which comprised 564 children born between 1996 and 2003, a noteworthy 35% of these children had indistinct, very

indistinct, or no speech, with the highest prevalence observed in children with severe gross motor impairments and dyskinetic CP. The results showed that only 54% of the 197 children with speech problems used AAC in some capacity, which raises serious questions regarding the suitability and accessibility of communication aids. Notably, hand signs were the most common method, but they were frequently of low quality. Additionally, compared to their preterm colleagues, infants delivered at term had more severe speech issues, which suggests that the significant brain abnormalities associated with term deliveries may make communication difficulties worse. By suggesting the adoption of a standardized registry form to gather crucial data, such as CP subtype and eating challenges at the time of diagnosis, the authors underlined the critical necessity for early identification of AAC needs. This proactive approach seeks to support prompt treatments that can improve social interaction, communication outcomes, and ultimately the quality of life for kids with cerebral palsy. The conclusions highlight the need for medical practitioners to give AAC resources and assistance top priority in order to guarantee that all children with cerebral palsy can acquire efficient communication skills, which are essential for their social and cognitive growth.

A speaker recognition system was created by Ning Wang et al[16] to address the difficulties presented by noisy surroundings, which frequently impair the effectiveness of conventional recognition methods. By using a feature estimation method that efficiently removes noise-specific components prior to feature extraction using spectral subtraction, their creative methodology captures both vocal source and vocal tract features from spoken utterances. Significant gains in recognition performance, especially at low signal-to-noise ratios (SNRs), were shown by the study's analytical derivations and simulation results. For applications in real-world settings with high noise levels, the system consistently reduced both identification error rates and equal error rates throughout a range of SNRs from 0 to 15 dB. By combining traditional vocal tract features with voice source-related features, the study highlighted how important it is to maintain speaker-discriminative information even under unfavorable acoustic circumstances. Additionally, the results showed that reliable parameter estimation not only improves speaker recognition systems performance but also creates exciting opportunities for further study, especially in the area of sophisticated noise reduction methods that could improve recognition accuracy and dependability even more.

By combining Vector Quantization (VQ) for effective classification and Mel Frequency Cepstral Coefficients (MFCC) for feature extraction, Kashyap Patel et al[17] created a comprehensive method for speaker and gender recognition. By using the "melcepst" function to compute the melcepstrum—a representation of crucial speech signal features—their method used MFCC to record speakers' distinctive vocal qualities. By using VQ in conjunction with the K-means algorithm to cluster these features into a codebook that had representative vectors for every speaker, it was possible to effectively identify unfamiliar speakers through similarity matching that minimized Euclidean distance. They used MATLAB to compute autocorrelation and the Harmonic Product Spectrum(HPS) technique to assess pitch in order to reliably discern between male and female voices based on pitch characteristics. High accuracy in speaker and gender recognition was shown in extensive testing on voice commands, confirming the system's usefulness in practical applications and demonstrating the strong potential of MFCC and VQ techniques in improving the accuracy and dependability of speech recognition technologies.

In order to address the severe speech difficulties that people with cerebral palsy (CP) encounter because of their impaired control and disrupted neural connectivity—which frequently require repetition in communication—Mohd Hafidz Mohamad Jamil et al[18] created an adaptive speech-to-text recognition system especially for CP patients. This cutting-edge system uses Dynamic Time Warping (DTW) to compare these features with pre-stored templates customized to the distinct speech patterns of CP users, Zero Crossing Rate (ZCR) to differentiate between sound and silence, and Mel-Frequency Cepstral Coefficients (MFCC) to extract essential and detailed speech features. With results ranging from 78% to 97%, the system achieves great accuracy in speech recognition by allowing it to "learn" from each user's speech features. This demonstrated its potential to help CP people engage in more fluid and productive everyday interactions. Furthermore, by reducing communication barriers and enhancing accessibility in regular discussions, individualized voice recognition systems can help people with speech impairments and eventually enable them to participate more completely in social settings.

Prasanth P. S[19] reports a study on speaker recognition that focuses on automatically identifying and validating people based on their voice traits. In order to improve speaker classification accuracy, the study highlights the importance of robust feature estimate employing vocal tract data, particularly pitch and Mel Frequency Cepstral Coefficients (MFCC). It explains text-independent speaker classification techniques that make use of cepstral coefficients, which accurately capture the spectral properties of speech, and makes a careful distinction between identification and verification modes. The study combines pitch and MFCC to greatly enhance recognition performance, despite the difficulties caused by inaccurate pitch estimate. By training on MFCC features obtained from spoken utterances, the Support Vector Machine (SVM) is used as a classifier and produces hyperplanes for binary speaker categorization. They examined how noise affects classification accuracy, using MATLAB to examine the connection between Signal-to-Noise Ratio (SNR) and recognition performance.

The relationship between preterm infant’s cognitive and language results and adult language exposure in the NICU was investigated by Caskey M et al[20]. The study used 16-hour digital recordings at 32 and 36 weeks postmenstrual age

(PMA) using a Language Environment Analysis (LENA) device to track 36 preterm infants with birth weights ≤1250g. The findings showed that, regardless of birth weight, more adult word counts were associated with better Bayley-III cognitive and language scores at 7 and 18 months' corrected age. In particular, adult word exposure at 32 weeks PMA explained 26% of the cognitive variance at 7 months, 12% of the language variance, and 20% of the expressive communication variance at 18 months. These results imply that language exchanges between parents and newborns in the NICU may be a useful early intervention for promoting preterm infants' cognitive and linguistic development.

In a study connected to several Dutch institutions, Rimke C. Vos et al[21] investigated the expressive and receptive communication developmental trajectories of children and young adults with cerebral palsy (CP). The Vineland Adaptive Behavior Scales were used to measure communication in 418 participants (ages 1–24) over a period of 2–4 years. Participants were categorized by type of CP (unilateral spastic (USCP), bilateral spastic (BSCP), and non-spastic (NSCP)) and intellectual disability. According to the results, people without intellectual disabilities performed the best in receptive language, whereas those with USCP had the best expressive outcomes. The results support customized interventions by showing that expressive communication is more in line with the type of motor disorder, but receptive communication is associated with intellectual ability.

In order to overcome the difficulties caused by background noise in voice signals, Abd El-Fattah et al[22] suggest an adaptive wiener filtering technique for improving speech quality. In order to account for the time-varying character of speech, the approach uses local mean and variance statistics of the speech signal to adjust the filter transfer function sample by sample. The suggested method functions in the time domain as opposed to the conventional frequency-domain wiener filtering. Using a range of voice quality indicators, the authors contrasted their approach with well-known methods like wavelet denoising and spectral subtraction. The findings reveal that the adaptive wiener filter works better than conventional techniques in situations with colored noise and Additive White Gaussian Noise (AWGN).

The study found adaptive impulse response greatly improves voice quality and intelligibility without requiring input other than the noisy signal.

The efficacy of PROMPT (Prompts for Restructuring Oral Muscular Phonetic Targets) therapy in enhancing the accuracy of speech output was assessed by Ward R et al[23] in six children with moderate-to-severe speech deficits linked to cerebral palsy, ages 3 to 11. A baseline phase (A1), an intervention phase aimed at the first priority in the PROMPT hierarchy (B), and a higher-level goal phase (C) comprised the study's single-subject research design. To evaluate both taught and untrained words, weekly speech probes were given with an emphasis on perceptual accuracy and motor-speech movement parameters. All participants made considerable progress between phases A1 and B, and four out of six showed improvement between phases B and C, according to the data, which showed statistically significant gains in speech output accuracy. These results highlight the advantages of sensory augmentation and motor learning in improving speech intelligibility, providing early evidence in favor of the application of dynamic systems theory in speech therapies for children with cerebral palsy.

Early detection and intervention techniques for newborns with cerebral palsy (CP) or at risk for it were examined by Herskind et al[24]. The study emphasizes the potential advantages of early intervention, especially when started within the first six months following term age. The authors stress the importance of accurately identifying newborns exhibiting early indicators of cerebral palsy (CP), such as aberrant movement patterns and motor milestone delays, as identified by neuro-imaging and the movements assessment. The authors support the potential advantages of early intervention despite difficulties in proving its substantial long-term effects, emphasizing that future research should give preference to infants chosen based on reliable neuro-developmental assessments.

A comprehensive evaluation was carried out by Morgan et al[25] to assess the efficacy of motor therapies for infants with cerebral palsy (CP) or at high risk of developing it from birth to two years of age. The review focused on 34 studies, including 10 randomized controlled trials, and incorporated literature found in several journal databases. The results showed that the most often studied intervention, either as the experimental or control assignment, was neuro-developmental therapy. Task-specific instruction, contextual alteration, and child-initiated movement were common themes among the interventions that had moderate to substantial benefits on motor outcomes. The authors pointed out that the evidence for early motor intervention is limited because there aren't enough high-quality trials. In order to better understand the efficacy of motor therapies for this vulnerable population, the study emphasized the significance of carrying out additional research, especially randomized controlled trials with explicit descriptions of interventions.

A population-based investigation on the prevalence and features of language impairment in children with cerebral palsy (CP) between the ages of 5 and 6 was carried out by Cristina Mei et al[26]. The team, examined information on 84 kids from the Victorian Cerebral Palsy Register (VCPR). 61% of participants in the study had language impairments, with 24% not being able to speak. The most common type of language impairments were combined receptive and expressive (44%), with isolated receptive (7%) or expressive (5%) impairments being less common. Cognitive deficits were the strongest predictor in multivariable models, and severe cognitive impairment along with Gross Motor Function Classification System levels IV and V were linked to higher impairment rates. Language deficits were found to span multiple domains, indicating generalized impairment. Given the significant frequency of language impairment, the authors recommend clinical language examinations for early

intervention. However, they point out potential drawbacks, including the PLS-4 test's liberal cut-off (>1SD below mean) and motor demands, which may have understated some talents. In order to develop evidence-based interventions specifically designed for children with cerebral palsy, they advise more study.

A thorough analysis of several voice recognition algorithms was conducted by Vadwala et al[27], who emphasized the speech as a basic communication method and its consequences for human-computer interaction. They divided recognition techniques into two categories: connected word recognition, which allows for more natural speech patterns with fewer pauses, and isolated word recognition, which necessitates pauses between uttered words. The study described the inherent difficulties in voice recognition, such as contextual dependencies, background noise, accent differences, and speech rates, which greatly impede accurate speech-to-text conversion. The authors emphasized how artificial intelligence (AI) is revolutionizing speech recognition, especially with hybrid systems that combine acoustic phonetics and pattern recognition to get more accurate results. They went into further detail on the usefulness of Artificial Neural Networks (ANNs), which may use a mixture of learning approaches, including supervised, unsupervised, and reinforcement learning, to enhance recognition accuracy by dynamically learning and adapting to various speech patterns. The knowledge-based method, which makes use of linguistic and phonetic norms generated from expert knowledge, was also emphasized in the survey. However, because it is difficult to acquire and efficiently encode such knowledge, it confronts practical implementation challenges. To improve the accuracy and usefulness of speech recognition technologies in a assortment of real-world contexts, the incorporation of speech interfaces into consumer electronics and accessibility tools for citizens with disabilities, they underlined the importance of handling the constraints presented by ambient noise and speech variability.

The difficulties and guiding ideas of rehabilitating persons with cerebral palsy (CP), a nonprogressive neurological disorder, are covered by Mintaze Kerem Günel et al[28]. Although cerebral palsy is stable in and of itself, people with CP may have functional and health deficits as they age, primarily as a result of changes in motor and postural control and problems with muscle tone. These modifications frequently result in compensatory postures or movements, which might have an additional negative influence on general function and health. The authors stress that many people with CP are able to participate in social activities, learn new skills, seek work, and live independently with the help of comprehensive medical management, rehabilitation, adaptive approaches, recreational activities, and assistive technology. However, a considerable portion of people with CP need continuous help because to the complexities of aging. In order to enhance evidence-based clinical practice and create strong care systems that cater to the particular requirements of this population, they promote interdisciplinary collaboration, highlighting the significance of specialized approaches to promote their well-being.

According to Xiong F et al[29], a more reliable transfer learning framework designed to create more reliable tailored speech models for speakers who have dysarthria. Using minimum speaker data for the dysarthria, a CNN-TDNN-F-based state-of-the-art acoustic model is first pretrained on source domain data before being transferred to the target domain based on weight adaption for the neural network. The findings demonstrate that linear weights are necessary for the neural layers to represent speech from dysarthric people more well, resulting in relative recognition increases of 11.6% and 7.6% over conventional speaker-dependent training and data-combination methods, respectively. The authors present the idea of an utterance-based data selection method in which the posterior probability's entropy—which has a gaussian distribution—is statistically examined. The further absolute recognition performance gain from close to 2% above the baseline of transfer learning for moderate to severely dysarthric speakers was the result of this method's ability to obtain true positives in source domain data that are beneficial.

Speech Vision (SV) is an ASR system that uses a sophisticated deep learning framework and was developed by Shahamiri et al[30] to help individuals with dysarthria, a motor speech disorder that results in unclear articulation due to paralysis of the voice-producing muscles. The work, which focuses on three main problems with dysarthric ASR—phoneme distortion, a lack of dysarthric speech data, and mislabeled phonemes. By employing visual features rather than conventional phoneme recognition, Speech Vision introduces a novel visual approach to ASR. This indicates that the model "sees" the word shapes that people with dysarthria pronounce. To improve training, SV also makes use of synthetic dysarthric speech and data augmentation. To increase the recognition of dysarthric word forms, it uses a number of strategies, such as transfer learning from healthy speech data. It was evaluated using the UA-Speech dataset. 67% of the UA-Speech speakers, particularly those with severe dysarthria, showed improvement, indicating that SV performed better than the state-of-the-art systems at the time. Although it has shown promise, its primary limitations are the extremely limited creation of synthetic samples and the less than ideal outcomes for mild cases of dysarthria, which can be addressed with additional architectural advancements.

Zhengjun Yue et. al[31] presented the use of multi-stream architectures made up of convolutional, recurrent, and fully linked layers, aimed to improve ASR by fusing articulatory information with acoustic data. At different levels of abstraction, these architectures enable the best possible feature fusion. The value of genuine articulatory data in enhancing recognition accuracy was demonstrated by evaluations on the TORGO dysarthric speech database, which showed that the suggested strategy dramatically decreased word error rates(WER) for dysarthric speakers by up to 4.6% absolute (9.6% relative). The authors recognized that more research into different modalities and more complex fusion methods

was necessary to improve ASR performance for dysarthric speech.

In order to evaluate accessibility issues, Jaddoh et al[32] reviewed how individuals with dysarthria interact with automated speech recognition (ASR) systems. The results show that using dysarthric speech as training material enhances ASR performance, but they also indicate ongoing difficulties in ASR system interactions for users with dysarthria. However, the deficiency of variety in utterances and the usage of recurrent samples from the same users make it difficult to compile large datasets of dysarthric speech. The authors stress that in order to increase accessibility, it is necessary to keep improving ASR systems, include dysarthric people in the design and testing stages, and create interaction strategies specifically for this demographic.

A systematic evaluation of the use of machine learning (ML) in cerebral palsy (CP) studies concerning diagnosis, subtype classification, and treatment planning was conducted by Nahar A et al[33]. The results indicate that RF and DT are very good at classifying subjects' movements, that RF and SVM are highly accurate at evaluating exercise, and that 94% of the time, RF was used to analyze gait patterns. Because BCF predicts orthopedic and neurological outcomes with 74% accuracy and neural networks diagnose CP from eye pictures with 94.17% accuracy, machine learning also makes it possible to create a treatment plan tailored to a particular patient. They believe that, more extensive and carefully selected datasets are still needed to increase the model's dependability across a range of demographics. Other ethical issues include patient permission and data security, emphasizing that clinical scenarios should be responsible for ML integration.

III. PROPOSED METHODOLOGY

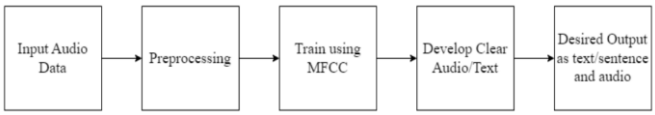

Fig. 1. Process Diagram

The process(Fig.1) can begin with the meticulous collection of audio samples from those with cerebral palsy, guaranteeing that the recordings reflect a diversity of speech impairments and levels of clarity. This rich data can then form the basis of a model that a machine-learning researcher can train so that it generalises from one set of speech features to another. The second stage could be to preprocess the audio so that it sounds better and is easier for the model to learn from. You could normalise the audio levels so that everything in the dataset is the same volume. This adds consistency, and helps prevent the model from learning specific differences in volume.

With normalisation, background noise reduction techniques can be utilized to remove any unwanted noises that could partially obscure the speech and interfere with how accurately it can be analysed. Removing the non-speech elements from the signal will remove a mass of performance noise, ensuring that the features extracted from the audio are truly representative of that speech alone. Splitting up the audio into smaller clips can allow for more granular analysis, as the model can focus on smaller segments or regions with different patterns and features.

After the preprocessing, Mel-frequency cepstral coefficients (MFCC) extraction takes place, which extracts primary features of the signal, transforming it from a signal into a mathematical representation that is processable by the machine learning model. Coefficients are clustered, and three outliers are removed, leaving a total of 52 coefficients. These features, as a vector, can now be fed into the KNN algorithm to train the samples. Owing to the ease of the classification task, the KNN algorithm has been shown to possess high efficiency and accuracy.

During training, a KNN model can learn how to map the features of different exemplars of a kind of speech (eg, saying ‘a thorough examination’ in various speech prostheses) to the way those speech samples were actually rated for intelligibility (eg, highly clear, somewhat clear, very impaired). For any new speech sample, it can then use the answers to 100 similarity questions (‘Was ‘a thorough examination’ in speech prosthesis x similar to ‘a thorough examination’ in speech prosthesis y?’) – answered under the constraint that k neighbours are shared between x and y – to classify this new speech sample as more or less clear, and then manipulate the input to improve its intelligibility. The value of k can vary, rely on the data: small values were used to differentiate between two different vocoder types, but the researchers had to use 25 nearest neighbours to distinguish between a large set of vocoder types. The curse of dimensionality also means that the choice of distance metric can have a major impact on the classification, even when using the same hyperparameter k.

Armed with this training, the KNN model can be brought to bear on the analysis of new audio samples in real time – the model can take in the features of an incoming speech signal, and use the associations it has learned to categorise it. The output can then be generated in two forms: a text-based transcript of the enhanced speech, and an audio file that speaks the same audio, augmented with the improvements. This dual output ensures the end-users that, they are able to not just understand the content of their speech expressed in words written out before them, yet they also received a reference audio that exemplifies the improvements made on their speech.

Through this workflow, ultimately, we provide cognitive time to implement PALS technology in a way that slows down speech, increasing the window of opportunity for the words to be correctly pronounced and giving persons with cerebral-palsy increased speech intelligibility, leveling the language and articulation playing field in general social participation, in addition to instrumental everyday activity. From this approach, we provide further grounds to study the possibilities of using machine learning to enhance the intelligibility of stutterers and persons with chronic dysarthria. We could reach a point where

persons with stutter and dysarthria no longer have to contend with their struggles, a reality that will ultimately promote inclusion and free the world for those persons to live the life they desire.

IV. CONCLUSIONS

Enhancing speech clarity for individuals with cerebral palsy represents a significant advancement in communication methods through the use of machine learning algorithms. A system can be created based on a set of attributes derived from the speech samples and the K-Nearest Neighbours (k-NN) model, which can help us to identify the various groups of speech sounds and allows to improvise the clarity by avoiding the unspeakable combinations of the sounds. All-together, this entire approach to the problem can be done including the data acquisition, pre-processing, the development of the KNN model as well as making an evaluation of that system. Either an android or standalone application will be useful in real-time.

REFERENCES

- Shore, J., Burton, D.: Discrete Utterance Speech Recognition without Time Alignment. IEEE Transactions on Information Theory 29(4), 473–491 (1983).
- L. R. Bahl, P. F. Brown, P. V. de Souza, and R. L. Mercer, "Maximum mutual information estimation of hidden Markov model parameters for speech recognition," in Proceedings of ICASSP, pp. 49–52, Tokyo, 1986.
- S. J. Young, J. J. Odell, and P. C. Woodland, "Tree-based state tying for high accuracy acoustic modelling," in Proceedings of Human Language Technology Workshop, pp. 307–312, Plainsboro NJ, Morgan Kaufman Publishers Inc, 1994.
- M. Cooke, P. D. Green, and M. D. Crawford, "Handling missing data in speech recognition," in Proceedings of ICSLP, pp. 1555–1558, Yokohama, Japan,1994
- K. T. Chen, W. W. Liao, H. M. Wang, and L. S. Lee, "Fast speaker adaptation using eigenspace-based maximum likelihood linear regression," in Proceedings of ICSLP, Beijing, China, 2000.
- Tan Lee, W.K. Lo, P.C. Ching and Helen Meng, "Spoken language resources for Cantonese speech processing," Speech Communication, Vol. 36, pp.327-342, 2002.
- M. Seltzer, B. Raj, and R. Stern, "A Bayesian framework for spectrographic mask estimation for missing feature speech recognition," Speech Communication, vol. 43, no. 4, pp. 379–393, 2004.
- Gorter JW, Rosenbaum PL, Hanna SE, et al, "Limb distribution, motor impairment, and functional classification of cerebral palsy", Dev Med Child Neurol 2004; 46: 461–67.
- N. Zheng, C. Qin, T. Lee, and P. C. Ching, "CU2C: A dual-condition Cantonese speech database for speaker recognition applications", Proc. Oriental-COCOSDA, 2005, pp. 67–72.
- H.-K. Kuo and Y. Gao, "Maximum entropy direct models for speech recognition", IEEE Transactions on Audio Speech and Language Processing, vol. 14, no. 3, pp. 873–881, 2006.
- Pennington L, Smallman CE, Farrier F. "Intensive dysarthria therapy for older children with cerebral palsy: findings from six cases", Child Language Teaching & Therapy 2006; 22: 255-273.
- W. Byrne, "Minimum Bayes risk estimation and decoding in large vocabulary continuous speech recognition," IEICE Transactions on Information and Systems: Special Issue on Statistical Modelling for Speech Recognition, vol. E89-D(3), pp. 900–907, 2006.
- Gales, M., & Young, S., "The application of hidden Markov models in speech recognition", Foundations and trends in signal processing, 1(3), 195-304.
- Gorter JW, Ketelaar M, Rosenbaum P, Helders PJ, Palisano R., "Use of the GMFCS in infants with CP: the need for reclassification at age 2 years or older", Dev Med Child Neurol 2009; 51: 46–52
- Andersen G, Mjøen TR, Vik T, "Prevalence of speech problems and the use of augmentative and alternative communication in children with cerebral palsy: a registry-based study in Norway", Perspect Augment Altern Commun 2010; 19: 12–20.
- C. Ching, Nengheng Zheng, Tan Lee, "Robust Speaker Recognition Using Denoised Vocal Source and Vocal Tract Features" , IEEE Transactions on Audio, Speech, and Language Processing, Vol. 19, No. 1, January 2011.
- Mr. Kashyap Patel Dr. R. K. Prasad "Speech Recognition and Verification Using MFCC & VQ" International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 5, May 2013 ISSN: 2277 128X.
- Abd Manaf, A., Zeki, A., Zamani, M., Chuprat, S., El-Qawasmeh, E. (eds) Informatics Engineering and Information Science. ICIEIS 2011. Communications in Computer and Information Science, vol 251. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-25327-0_5
- Prasanth P.S. "Speaker Recognition Using Vocal Tract Features", International Journal of Engineering Inventions e-ISSN:2278 7461, P-ISSN: 2319-6491 Volume 3, Issue 1 (August 2013) PP. 26-30.
- Caskey M, Stephens B, Tucker R, Vohr B. "Adult talk in the NICU with preterm infants and developmental out-comes", Pediatrics 2014; 133: e578-84
- Vos RC, Dallmeijer AJ, Verhoef M, et al, "Developmental trajectories of receptive and expressive communication in children and young adults with cerebral palsy", Dev Med Child Neurol 2014; 56: 951–59.

- Abd El-Fattah, M. A., Dessouky, M. I., Abbas, A. M., Diab, S. M.,El-Rabaie, E. S. M., Al-Nuaimy, W., Alshebeili, S. A., & Abd El-Samie, F. E. "Speech enhancement with an adaptive Wiener filter", International Journal of Speech Technology, 17(1), 53–64.
- Ward R, Leitão S, Strauss G, "An evaluation of the eIectiveness of PROMPT therapy in improving speech production accuracy in six children with cerebral palsy", International Journal of Speech Language Pathology 2014;16(4):355-71.
- Herskind A, Greisen G, Nielsen , "Early identification and intervention in cerebral palsy", Dev Med Child Neurol 2015; 57: 29–36
- Morgan C, Darrah J, Gordon AM and others, "Effectiveness of motor interventions in infants with cerebral palsy: a systematic review", Dev Med Child Neurol 2016; 58:900–09
- Mei C, Reilly S, Reddihough D, Mensah F, Pennington L, Morgan A, "Language outcomes of children with cerebral palsy aged 5 years and 6 years: a population based study", Dev Med Child Neurol 2016; 58:605–11
- Ayushi Y. Vadwala, Krina A. Suthar, Yesha A. Karmakar, Nirali Pandya, "Survey Paper on Different Speech Recognition Algorithm: Challenges and Techniques", International Journal of Computer Applications, October 2017, Volume 175.
- Günel, M. K., Karadag, Y. S., & Anlar, "Rehabilitation Principles of Adults with Cerebral Palsy", Cerebral Palsy-Springer, Cham. October 2018. Volume 343-348.
- F. Xiong, J. Barker, Z. Yue, and H. Christensen, "Source domain data selection for improved transfer learning targeting dysarthric speech recognition", IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 7424–7428.
- Seyed Reza Shahamiri, "Speech vision: An end-to-end deep learning-based dysarthric automatic speech recognition system," IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 29, pp. 852-861, 2021.
- Z. Yue, E. Loweimi, Z. Cvetkovic, H. Christensen, and J. Barker, "Multi-modal acoustic-articulatory feature fusion for dysarthric speech recognition," in ICASSP. IEEE, 2022.
- Jaddoh, Aisha & Loizides, Fernando & Rana, Omer, "Interaction between people with dysarthria and speech recognition systems: A review", Assistive Technology. 35. 1-9. 10.1080/10400435.2022.2061085.
- Anjuman Nahar, Sudip Paul, Manob Jyoti Saikia, "A systematic review on machine learning approaches in cerebral palsy research", PeerJ, 12:e18270, 2024.
- Pradeep Kumar KG, Dr. Karunakara K, and Dr. Thyagaraju G. S., "Automated Identification of Diabetic Retinopathy: A Survey", International Journal on Recent and Innovation Trends in Computing and Communication (IJRITCC), June 17, Volume 5, Issue 6, ISSN: 2321-8169, PP: 514 – 520