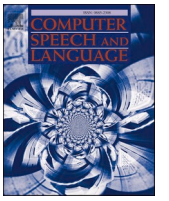
Computer Speech and Language 84 (2024) 101568

Contents lists available at ScienceDirect

Computer Speech & Language

journal homepage: www.elsevier.com/locate/csl



GTSO: Gradient tangent search optimization enabled voice transformer with speech intelligibility for aphasia

Ranjith R [a,*], Chandrasekar A [b]

[a] Research Scholar, Department of Computer Science and Engineering, St. Joseph's College of Engineering, OMR, Chennai 119, India.
[b] Professor, Department of Computer Science and Engineering, St. Joseph's College of Engineering, OMR, Chennai 119, India.

ARTICLE INFO

Keywords:
Median filter
Non-linear spectral subtraction
Gradient Descent (GD) Optimization
Voice transformer
Tangent Search Algorithm (TSA)

ABSTRACT

A frequent neurological condition known as aphasia is brought on by injury to language-related brain regions as well as possibly other regions of the brain involved in executive, memory, and attention functions. Due to a lack of speech-language pathologists and the vast expense of treatment, traditional therapy is difficult for aphasia-affected people to access. In this research work, speech intelligibility for aphasia is done by the proposed Gradient Tangent Search Optimization (GTSO) algorithm-enabled voice transformer. Here, the median filter is used for pre-processing the signal to reduce noise. The pre-processed voice signal is allowed for feature extraction and voice enhancement stages. Moreover, nonlinear spectral subtraction is used for voice enhancement and voice transformer is used for voice recognition. Also, the voice transformer is trained by GTSO, which is devised by hybridizing Gradient Descent (GD) Optimization and Tangent Search Algorithm (TSA). Then, the output obtained is fed to the language and pronunciation model for recognizing speech, and at last, the speech recognized is converted to text. Furthermore, the GTSO- enabled voice transformer is analyzed for its performance by three metrics, namely recognition accuracy, Positive Predictive Value (PPV), and Negative Predictive Value (NPV), with superior values of 0.919, 0.919, and 0.915.

1. Introduction

Aphasia (Damasio, 1992) is a common neurological disorder caused by damage to brain regions involved in language and possibly other cognitive abilities like memory, attention, and executive functioning. Aphasia can cause impairment in many areas of language that includes reading, writing, listening as well as speaking. Verbal production of Persons With Aphasia (PWA) is difficult to understand because of language problems like paraphasia, omitted words, or incorrect sentence construction. Aphasia has a profound impact on the lives of those affected. Aphasia is often accompanied by motor control disorders like apraxia and dysarthria, which can cause articulatory distortions and atypical prosody in addition to language disorders due to aphasia (Le et al., 2016). Difficulty can extend to various levels and components of the language system, including phonology, lexicon, syntax, and semantics (Adam, 2014). Speech analysis is an integral part of a comprehensive analysis, the primary purpose of which is to determine the type and/or severity of a PWA's impairment. This is done through acoustic and linguistic analysis of PWA speech obtained via narrative tasks (Qin et al., 2018). PWAs often face significant communication difficulties that can lead to frustration, loss of autonomy, and social isolation (Mahmoud et al., 2020). Assessment of narrative spontaneous speech, as a description of the picture, narration generated by PWA is an integral

* Corresponding author.
E-mail address: ranch890@gmail.com (R. R).

https://doi.org/10.1016/j.csl.2023.101568
Received 13 April 2023; Received in revised form 25 July 2023; Accepted 8 September 2023
Available online 9 September 2023

part of clinical assessment to assess the type and severity of aphasia. Content and fluency of unprepared narrative speech are considered informative indicators of the severity of the disorder (Yiu, 1992). Evaluation is performed by trained Speech Language Pathologists (SLP) along appropriate cultural and linguistic backgrounds. The reliability and accuracy of such subject assessment methods depend on the experience of the clinic (Qin et al., 2019).

Measurement of speech intelligibility is related to subjective listening, in which the percentage of the words understood correctly by people is estimated. However, such procedure is labor intensive and expensive, and is also influenced by the listener's perception of the patient's speech disorder and contextual or linguistic cues available in combined speech (Landa et al., 2014). So, time-effective and cost-effective automated comprehensible measurements provide a repeatable and reliable evaluation. During the past decade, a number of approaches is developed to assess pathological speech intelligibility, which can be broadly classified into triple categories, like Automatic Speech Recognition (ASR) enabled approaches, acoustic modeling-enabled approaches, and feature-enabled approaches. Feature-based approaches typically refer to a blind assessment of speech intelligibility by extracting several acoustic features such as pitch or percentage of audio frames (Janbakhshi et al., 2019). ASR is typically trained on people's voices without speech pathology and performance deteriorates while used on aphasic speech. Additionally, ASR is typically dependent on language and must be trained on hundreds of hours of transcribed speech. This feature prevents in many cases from extending their use to thousands of languages currently spoken in the world, and especially to aphasic speech recognition use cases, because there is not as much annotated data for more traditional educational recognition models guided learning methods (Torre et al., 2021). In recent years, the development of ASR technology and fully automated approaches are actively investigated (Le et al., 2018). Extraction of features is performed by time alignment and text output information produced by ASR. The design of the feature is largely based on expertise gained in clinical practice. In particular, textual statistics, such as counting word occurrences by parts of speech, and psycholinguistic information, like word-level familiarity and age of acquisition, is shown to be useful indicators of language decline (Qin et al., 2019).

ASR system was used to generate speech transcript to extract features of texture. The low recognition accuracy of this general-purpose ASR system in PWA speech limits its practical use in speech evaluation of PWA. A common problem in developing ASR systems for typical speech, including PWA speech, is the lack of adequate training data for speech content, speech style, etc (Qin et al., 2018). Moreover, "written" language *via* spoken language includes speech recognition training with various sets of exercises so that program recognizes certain features of user's speech as well as transcribes user's speech in an accurate manner (Estes and Bloom, 2011). The applicability of Deep Learning (DL) in speech impairment analysis was compared with widely utilized classical Machine Learning (ML). The DL-based method uses high-resolution Time Frequency (TF) images in Support Vector Machines (SVM) and a Gaussian Mixture Model (GMM) to automatically assess speech disorders (Mahmoud et al., 2020). Deeper networks are considered to be more efficient to learn than shallow networks (Gnanamanickam et al., 2021). Deep Auto Encoder (DAE) has been designed to train the architectures of deep networks. The challenge of DAE is the difficulty of generalizing the algorithm to every speech signal type. In addition to the traditional statistical techniques based on Minimum Mean Square Error (MMSE), inspection methods using deep neural networks have also been developed to augment huge and vast amounts of data on speech. These methods have been found to effectively and efficiently handle non-stationary sounds (Xu et al., 2014). To improve speech with known and unknown noise sources, it has been suggested to use several deep neural networks with one neural network (Gnanamanickam et al., 2021).

This research work is related to speech intelligibility for aphasia, which is a disorder affecting how to communicate. In this work, the input signal is taken from the Talkbank dataset, which is further allowed for the pre-processing stage. Also, pre-processing is done by a median filter by which unwanted noises in the input signal are removed. This pre-processed signal is fed for the feature extraction stage and voice enhancement phase. In the feature extraction stage, various features like spectral centroid, zero crossing rate, spectral roll-off, chromagram, Mel-Frequency Cepstral Coefficients (MFCC), as well as the probability of voicing are extracted. Moreover, voice enhancement is done by nonlinear spectral subtraction for pre-processed signals and extracted features. Further, voice recognition is done by a voice transformer that is trained using GTSO. This hybridized GTSO is formed by combining optimization algorithms like TSA and GD Optimization. Finally, the language and pronunciation model is applied to recognized voice taken from the voice recognition phase to convert speech into text.

Hybridized algorithm contributed to this paper is,

- Developed GTSO-enabled voice transformer: Aphasia patients who loss the ability to express or understand written or spoken language should be enhanced for speech intelligibility. Here, a voice from aphasia patients is recognized by a voice transformer that is trained by GTSO. This newly developed GTSO is formed by the combination of GD Optimization and TSA. Here, GTSO is highly efficient to gain optimal solutions and is capable to solve optimization problems in the real world.

The remaining paper is involved with various sections: Section 2 indicates motivation, challenges, and literature reviews of speech intelligibility for individuals with aphasia. Section 3 represents the GTSO-enabled voice transformer module for the final conversion of speech to text. Section 4 indicates the results and discussion of the model with various performance metrics, and Section 5 concludes the paper.

2. Motivation

Listeners frequently view PWA less favourably than their peers. These impressions contain false presumptions that could hinder fruitful social connections. Hence, it might be difficult to identify disordered speech due to a variety of factors, such as typical speech patterns, speaker unpredictability, and a lack of data. Hence, it is challenging to use conventional acoustic modelling and adaption techniques to disorganized speech. Changing the verbal output of PWA may also result in more favourable listener impressions of the

speech, speaker, and their own emotive response. Communication partner training has been proven to improve social outcomes relating to the listener. In this section, literature reviews along with challenges regarding speech intelligibility for aphasia are expressed.

2.1. Literature reviews

Le et al. (2016) used Deep Neural Network (DNN) acoustic model for automatic assessment of speech intelligibility for individuals with aphasia. Although this approach had high stability and reliability, this method failed to improve the accuracy of automatic transcript generation. Mahmoud et al. (2020) developed Convolutional Neural Network (CNN) for speech assessment of Mandarin-Speaking Aphasic Patients. This technique provided a novel basis for assessing aphasic patients' speech disability levels. However, this method failed to investigate other Chinese dialects and other international languages to enhance generalizability. Qin et al. (2019) designed DNN for the automatic assessment of speech impairment in Cantonese-speaking people with aphasia. Although this approach was highly robust and efficient in improving generalization capability, this technique failed to improvise the performance of ASR on aphasic speech to produce more robust features. Mahmoud et al. (2021) used ResNet-34 pre-trained CNN model for the assessment of aphasia. This framework had minimal computation resource requirement and complexity, but the performance of the model depends on the training dataset size, which is difficult because of the scarcity of domain data.

Qin et al. (2020) designed the CNN model for automatic speech assessment for Cantonese-speaking people with aphasia. CNN model used was capable of learning impaired acoustic patterns implicitly for speech assessment of PWA. However, it failed to learn semantic utterances sufficiently but concentrated on the acoustic impairment of PWA. Qin et al. (2018) used Machine Translation-Time Delay Neural Network – Bi-Directional Long Short-Term Memory (MT-TDNN-BLSTM) for automatic speech assessment for people with aphasia. This approach improved ASR performance successfully on impaired robustness and speech-of-text features. But, this method failed to consider the syntactic impairment of aphasic speakers in the assessment system to enhance the accuracy of recognition. Torre et al. (2021) presented Semi-Supervised Learning-based System for improving aphasia speech recognition on the aphasia bank for Spanish and English. This scheme was able to generalize learning of contextualized representations of speech at different types of speech, improvising the performance of ASR. However, this method failed to enhance results by fine-tuning specific methods for every aphasia severity level. Herath et al. (2022) introduced DNN Approach for classification into aphasia severity levels to recommend speech therapies. Here, the DNN model was successful in identifying aphasia severity levels very accurately and automatically, but this model suffered from low generalizability while the input chunk size was low. Zhao et al. (2022) implemented a Transformer-based ASR by integrating monotonic attention and sparse attention. Here, a learned sparsity approach was implemented by the sparse mechanism for fitting the corresponding head better. This model achieved offer better results in multi-head attention and self-attention. However, the complexity of the model was high. Liao et al. (2022) bidirectional context embedding (BCE) speech transformer model for ASR. This method avoided data leakage and the complexity of the model was low. However, this model was not applicable to large datasets. Gulati et al. (2020) implemented a transformer-based approach for ASR, called Conformer. Here, the transformers were combined with convolution neural networks, which exhibited a higher accuracy with less parameter when compared with previous methods. Anastasopoulos and Chiang (2018) implemented a multitask learning model. Here, the initial task offered higher level representations to the next task decoder, which offered better results than the single task approaches. This model was applicable for improving speech translation and transcription. Baevski et al. (2020) implemented wav2vec 2.0 for identifying powerful representations from the speech audio. This approach efficiently applied for speech recognition with minimum labeled data. Hsu et al. (2021) implemented a Hidden-Unit BERT (HuBERT) scheme to self-supervised speech representation learning. In this scheme, the prediction loss was applied to the masked regions that forced for learning a merged acoustic and language approach. Prabhavalkar et al. (2023) reviewed the ASR approaches in the last decade. It proved that end-to-end (E2E) schemes offered highly integrated, completely neural ASR approaches.

2.2. Challenges

Challenges by existing methods for automatic analysis of speech intelligibility for PWA are described as follows,

- DNN was proposed in Le et al. (2016) for speech intelligibility and was effective in bridging gap among humans as well as automatic intelligibility analysis, however, this technique needed further research for leveraging classification results to produce concrete feedback that PWAs can utilize for improvising their speech.
- The main challenge faced by the CNN model in Mahmoud et al. (2021) for speech assessment was that it failed to consider the collection of data for overcoming the scarcity of aphasia syndrome dataset type for improving the accuracy of CNN-enabled assessment and aphasia syndromes discrimination.
- CNN model proposed in Qin et al. (2020) for automatic speech assessment effectively saved a significant amount of manual work. However, it failed in establishing other neural network aiming at characterizing PWA language impairment for enhancing efficiency.
- The key issue faced by Semi-Supervised Learning Based System presented in Torre et al. (2021) for speech recognition was that it failed in improving the performance of the system by considering some other learning rate schedulers by tuning SpecAugment parameters, or by considering other hyperparameters configurations.
- Though many methods have been proposed for reliable speech assessment, general-purpose systems on ASR are not straightforwardly applied to impaired speech. Mis-matches in articulation, voice, as well as language usage lead to an error rate of the higher

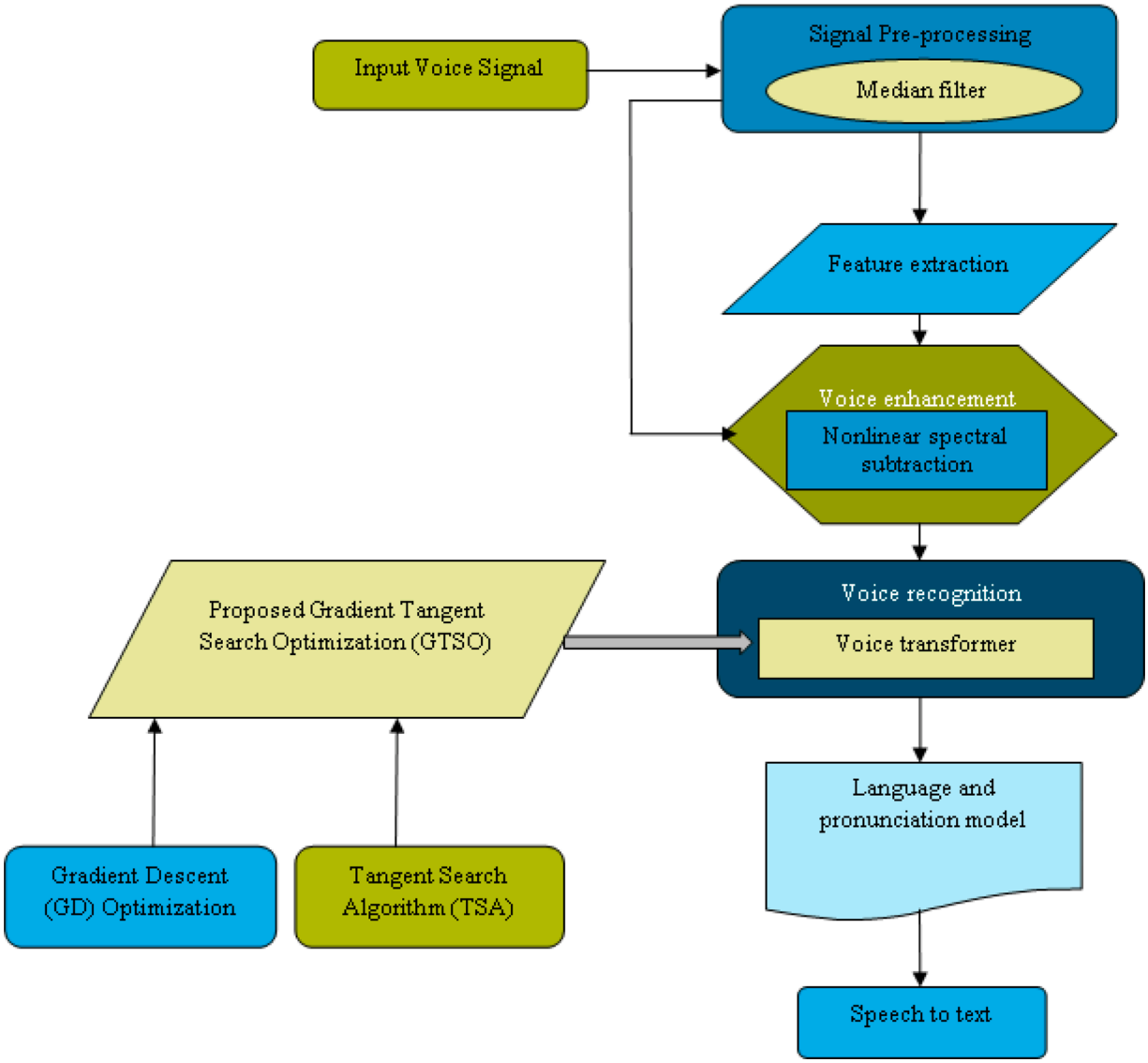


Fig. 1. Block diagram of GTSO-enabled voice transformer with speech intelligibility for aphasia.

words. Generating application-specific system of ASR with a higher rate of accuracy is a difficult task due to disease scarcity matched data training.

3. GTSO-enabled voice transformer

Aphasia is a brain disorder; where a person feels trouble in speaking and this proposed methodology provides an attractive solution for speech intelligibility. This technique for speech intelligibility for aphasia patients is applied in the following manner. At first, the input voice signal is acquired from the database (Talkbank dataset will be taken from, 2023) and is forwarded to signal pre-processing, where a median filter (Herzog, 2013) is used to reduce noise from the input voice signal. Then, the pre-processed voice signal is subjected to extraction of the feature phase as well as the voice enhancement phase. Here, features, such as zero crossing rate (Sandhya et al., 2020), spectral centroid (Sandhya et al., 2020), spectral roll-off (Sandhya et al., 2020), MFCC (Sandhya et al., 2020), chromagram (Sandhya et al., 2020), and probability of voicing (Nguyen et al., 2012) are extracted. Moreover, voice enhancement is performed on the pre-processed signal and extracted features using nonlinear spectral subtraction (Gnanamanickam et al., 2021), and the output obtained is forwarded to the voice recognition module. Here, voice recognition is carried out using a voice transformer (Dong et al., 2018), wherein the training process is carried out by using the proposed GTSO algorithm. This GTSO algorithm is devised by combining GD Optimization (Ruder, 2016; Gradient descent optimization is taken from, 2023) and TSA (Layeb, 2022). Hereafter, the output obtained is forwarded to the language and pronunciation model (Akita and Kawahara, 2009) for recognizing speech and

finally, the speech recognized is converted to text. Fig. 1 shows a block diagram of GTSO enabled voice transformer for the automatic assessment of speech intelligibility for PWA.

3.1. Data acquisition

Data acquisition is the first step considered in speech intelligibility for aphasia. The Talkbank dataset will be taken from (2023) is the source considered for data acquisition, which has many voice signals. Dataset Awith many voice signals is represented below,

$$\mathrm{A} = \left\{\mathrm{A}_1, \mathrm{A}_2, ..., \mathrm{A}_\alpha, ...\mathrm{A}_\beta\right\} \tag{1}$$

where, Ais a dataset with many voice signals, $\mathrm{A}_\alpha$is voice signal at the position α, andA_β is total voice signal at the last position β. Here, $\mathrm{A}_\alpha$is the input taken for further processing.

3.2. Signal pre-processing using median filter

The acquired signal from the dataset$\mathrm{A}_\alpha$is taken as input for pre-processing phase, which is carried out by a median filter (Herzog, 2013). Pre-processing is a method done to eradicate noises or artifacts from input signals. Median filtering is a non-linear operation and is highly robust. By median filter, Median Bdivides set into two equal-sized halves. If the sorted list of Cvalues $a(c)$is considered and Codd, then median $B$is middle element$a\left(\frac{C-1}{2}\right)$. Median filtering is defined by following formula,

$$B = \begin{cases} a\left(\dfrac{C-1}{2}\right) & C\ odd \\ \dfrac{1}{2}\left[a\left(\dfrac{C}{2}\right) + a\left(\dfrac{C}{2}+1\right)\right] & C\ even \end{cases} \tag{2}$$

Most commonly, median filters have odd length$C$. Non-linear nature of the median makes a closed-form description of its effects on audio signals. The median filter is robust to outliners and impulse like noise is removed by them. Also, the median filter completely suppresses impulses with huge magnitudes. The pre-processed signal is thus indicated by termD_α. This signal that is pre-processed is further allowed to the extraction of the feature phase and voice enhancement phase.

3.3. Feature extraction

The pre-processed signal $D_\alpha$is allowed for the feature extraction phase, in which features such as zero crossing rate (Sandhya et al., 2020), spectral centroid (Sandhya et al., 2020), spectral roll-off (Sandhya et al., 2020), MFCC (Sandhya et al., 2020), chromagram (Sandhya et al., 2020), and probability of voicing (Nguyen et al., 2012) are extracted. These are explained in detail below,

3.3.1. Zero crossing rate

Zero crossing rate (Sandhya et al., 2020; Nguyen et al., 2012) is count of times, where, signal indicating speech changes its polarity. This is useful measure in speech analysis, and this measure for Flength interval ending at $c = d$is given as,

$$E_1 = \frac{1}{2F}\sum_{c=e}^{d}|sign\{f(c)\} - sign\{f(c-1)\}|g(d-c) \tag{3}$$

Here, $e = d - F + 1$and $sign\{f(c)\} = \begin{cases} +1 & if\ \ f(c) \geq 0 \\ -1 & if\ \ f(c) < 0 \end{cases}$where, $E_1$is zero crossing rate.

3.3.2. Spectral centroid

This is the geometric centre or centre of mass of the spectrum, which is calculated as the average of frequencies in signal (Sandhya et al., 2020). Here, the scope of the frame $Gi|h|$is categorized to non-overlapping subbands, where every subband $j$is defined by an upper-frequency edge (u_j)as well as lower frequency edge (l_j)that is indicated as,

$$E_2 = \frac{\sum_{h=l_j}^{u_j} h|Gi[h]|^2}{\sum_{h=l_j}^{u_j}|Gi[h]|^2} \tag{4}$$

where, $E_2$is the spectral centroid.

3.3.3. Spectral roll-off

Spectral roll-off (Sandhya et al., 2020) provides a rough idea of higher frequency in signal and also provides frequency in which a certain quantity of energy is confined. The spectral roll-off frequency is utilized to distinguish among noisy sounds or above roll-off and harmonic or below roll-off, which is designated by termE_3.

3.3.4. MFCC

MFCC (Sandhya et al., 2020) is one of the widely utilized spectral characteristics in emotional speech recognition which amplifies the collection of coefficients providing information on the shape of the speech signal spectrum. This MFCC is represented by the below formula,

$$E_4 = 2595 * \log_{10}\left(1 + \frac{b}{700}\right) \tag{5}$$

where, $E_4$is MFCC, and$b$is frequency.

3.3.5. Chromagram

This feature relates to twelve various pitch classes (Sandhya et al., 2020). The main property of chroma features is that they help in capturing melodic and harmonic characteristics of sound, while being more robust to alterations in instrumentation and timbre. This feature is indicated as E_5.

3.3.6. Probability of voicing

The probability of voicing is provided for estimating voiced and unvoiced energy percentages for every harmonic within each of the plurality of bands of the speech signal spectrum. Pitch detection is highly accurate for the voiced pitch hypothesis and this performance degrades as deterioration of signal occurs. Hence, it is necessary to provide a probability of voicing and H_0value at the same time. The hypothesis is that firstly, voicing decision errors are manifested as absent pitch values; secondly, features like those indicating the shape of pitch contour are more robust to segmental misalignments; and thirdly voicing probability is more appropriate than the hard decision of 0 and 1, while used in statistical models (Nguyen et al., 2012). This is indicated by the termE_6.

Finally, extracted features are indicated in vector form as,

$$E_\alpha = \{E_1, E_2, ..., E_6\} \tag{6}$$

3.4. Voice enhancement using nonlinear spectral subtraction

Pre-processed signal D_α and extracted features$E_\alpha$are allowed for the voice enhancement process using nonlinear spectral subtraction (Gnanamanickam et al., 2021), where the enhanced signal is obtained. Nonlinear spectral subtraction is a primitive and famous speech-enhancing technique. This is suitable in situations, where boisterous environment contaminates real speech signals with similar bandwidth as of speech. This nonlinear spectral subtraction utilizes Signal-to-Noise Ratio (SNR) and over-subtraction factor in every frequency band. Initially, speech signal with noise as input is allowed to Fast Fourier Transform (FFT). Here, noisy speech is indicated by the below formula,

$$k(\alpha) = l(\alpha) + m(\alpha) \tag{7}$$

where, $k(\alpha)$is noisy speech, $l(\alpha)$is pure speech signal, and $m(\alpha)$is noise signal polluting pure signal. To get a relation in the spectral domain, power magnitude and Discrete Fourier Transform (DFT) with the assumption that speech and noise are uncorrelated are considered, which is expressed in relation as,

$$|K(n,o)|^2 = |L(n,o)|^2 + |M(n,o)|^2 \tag{8}$$

where, $n$is frame value, and$o$is frequency value. Assume that $|\widehat{M}(n,o)|$and $|M(n,o)|$can be estimated, and spectral subtraction is given by,

$$|\widehat{L}(n,o)|^2 = |K(n,o)|^2 - |\widehat{M}(n,o)|^2 \tag{9}$$

Spectral subtraction adapts damaged speech signals to short-term spectral magnitudes. The signal is changed as the synthesized signal feels as near to an unbroken voice signal. Here, noise power finds and subtraction rules are utilized for calculating spectral magnitudes' appropriate weighting. Moreover, rate of word error presents count of word error occurring at speech and is formulated by,

$$I = \mathrm{Z}s + \mathrm{Z}d + (\mathrm{Z}i/\mathrm{Z}n) \tag{10}$$

where, number of substitutions is Zs, number of deletions is Zd, number of insertions is Zi, and number of words in sentence is Zn. Thus, the enhanced voice signal is indicated by termJ_α.

3.5. Voice recognition using voice transformer model

Enhanced voice $J_\alpha$is further allowed for voice recognition using the voice transformer model (Dong et al., 2018). Voice recognition is the capacity of programme or machine to receive as well as interpret dictation or to comprehend and carry out spoken commands. Architecture representing the voice transformer is given below,

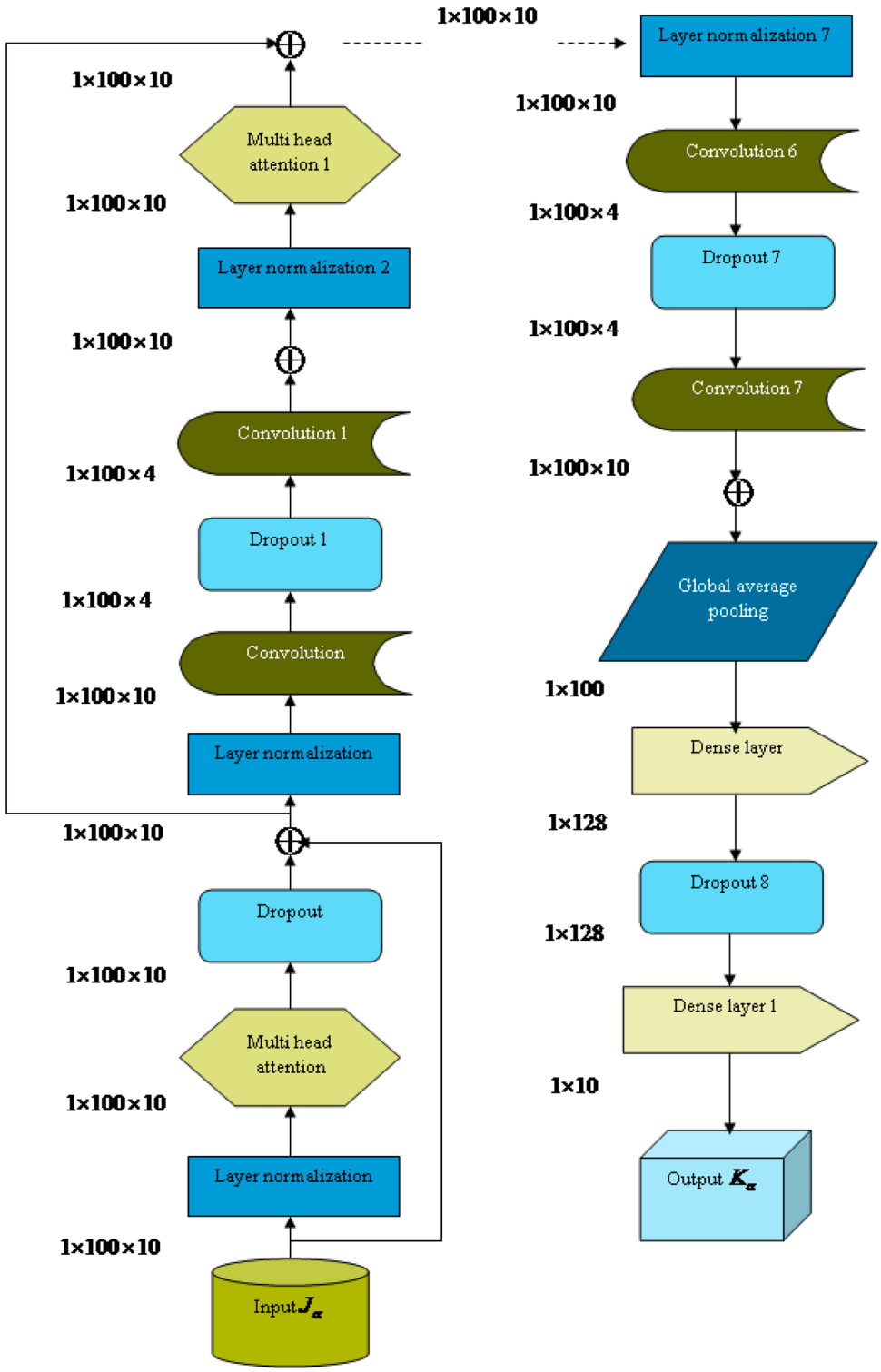


Fig. 2. Architecture model of voice transformer.

3.5.1. Architecture of voice transformer

Voice transformer (Dong et al., 2018) aims at converting sequence of speech feature to the corresponding sequence of characters. Feature sequence and character sequence are depicted as 2-dimensional spectrograms with frequency as well as time axes. Thus, convolutional networks are chosen for exploiting the structure locality of spectrograms and mitigating length mismatching by striding

with time. This architecture has an input layer, layer normalization, multi-head attention, dropout, convolution layer, global average pooling, and dense layers. This structure has 8 convolution layers for preventing GPU memory overflow and generates approximate hidden representation length along character length. Moreover, additional modules are stacked along with this for extracting more expressive representations for voice transformers. Here, input encoding is indicated as$p_{\mathrm{mod}\ el}$, and this is added to positional encoding for enabling the model to attend relative positions. Moreover, positional encoding is indicated by the below formula,

$$L_{(pos,q)} = \begin{cases} \sin\left(pos/10000^{2q/p_{\mathrm{mod}\ el}}\right) & 0 \le q < p_{\mathrm{mod}\ el}/2 \\ \cos\left(pos/10000^{2q/p_{\mathrm{mod}\ el}}\right) & p_{\mathrm{mod}\ el}/2 \le q < p_{\mathrm{mod}\ el} \end{cases} \tag{11}$$

where, $pos$is sequence position, $q$is qth dimension, and$L$is positional encoding. Here, the final encoded output is gained by inputting total of both input encoding and positional encoding to a stack of encoder blocks. Also, layer normalization as well as the residual connection is applied to every sub-block for efficient training, where, the corresponding output of sub-block inputs $r$is indicated by,

$$r + subblock(LayerNorm(r)) \tag{12}$$

Learned character level embedding is employed for converting character sequence to output encoding of the size$p_{\mathrm{mod}\ el}$ that is added along positional encoding. Then, the final output is obtained by the sum of stacked blocks. Finally, global average pooling is applied, which is further carried over to the dense layer, from which output $K_\alpha$is obtained from the voice transformer model. Fig. 2 indicates the architecture model of the voice transformer with various blocks.

3.5.2. Training process of voice transformer by GTSO

Voice transformer is trained by GTSO, which is formed by combining both TSA (Layeb, 2022) and GD (Ruder, 2016; Gradient descent optimization is taken from, 2021) Optimization. TSA is a population-enabled optimization algorithm to solve optimization issues. This algorithm uses a mathematical model related on the tangent function for converting a given solution to a better solution. Here, the tangent flight function is used which balances exploitation as well as exploration phases. TSA is very helpful in solving engineering problems and is capable of providing promising results on benchmarked functions. Similarly, GD Optimization is the most popular algorithm to optimize neural networks. GD is first-order optimization algorithm for finding local maxima and local minima of function. This GD reduces cost and loss of function and is highly employed because of its ease in implementation. When GD and TSA are united, they offered better outcomes in solving speech intelligibility in aphasia. The algorithmic procedure regarding GTSO is explained in detail below,

Step 1: Initialization

Initialization starts with generating a random population within the solution space, which is distributed in a uniform manner and is indicated as,

$$M_0 = N_{lb} + (N_{ub} - N_{lb}) * rand(O) \tag{13}$$

where, N_{lb} is the lower bound of the issue, N_{ub} is the upper bound of the issue, the random function is indicated as *rand*ranging $[0,1]$, and $O$is the dimension of the problem.

Step 2: Finding fitness

Fitness is calculated for finding maximum solutions for resolving optimization issues and utilizes outcomes of voice transformer along with targeted output. The fitness function is thus calculated by the below formula,

$$\mathrm{P} = \frac{1}{\beta}\sum_{\alpha=1}^{\beta}\left[Tr_\alpha^* - K_\alpha\right]^2 \tag{14}$$

where, Pis fitness function, total samples taken for processing is β, generated output from voice transformer is K_α, and targeted output is Tr_α^*.

Step 3: Intensification search

TSA begins by performing a local walk at random, directed as follows, and replacing the variables in the achieved solution with their corresponding values in the current optimal solution. Variables are replaced in the following ratios: 50 % for dimensions less than or equal to 4, and 20 % for dimensions more than 4. This is denoted in the below formula,

$$M_s^{t+1} = M_s^t + step * \tan(\phi) * \left(M_s^t - optQ_s^t\right) \tag{15}$$

$$M_s^{t+1} = M_s^t + step * \tan(\phi) * M_s^t - step * \tan(\phi) * optQ_s^t \tag{16}$$

Algorithm 1
Pseudo code of GTSO.

Sl. No.	Pseudo code of GTSO
1	***Input:*** M_s^t
2	***Output:*** Maximal solution M_s^{t+1}
3	**Start GTSO**
4	Initialize random population by Eq. (13)
5	Evaluate fitness function by Eq. (14)
6	**If** *rand* < *Uswitch*
7	Apply intensification search by Eq. (17)
8	**Hybridization of GD with TSA;**
9	Evaluate update equation of GTSO by Eq. (25)
10	**Else**
11	Apply exploration search by Eq. (26)
12	**End**
13	**If** *rand* < *Uesc*
14	Select agent search (Y)
15	Apply escape local minima by Eqs. (27) and (28)
16	**End**
17	Recalculate fitness function by Eq. (14).
18	**End GTSO**

$$M_s^{t+1} = M_s^t[1 + step * \tan(\phi)] - step * \tan(\phi) * optQ_s^t \tag{17}$$

where, $M_s^{t+1}$is the position of sth solution at iteration $(t+1)$, $step$is a function of step size that reduces as iteration $t$reduces, $\phi$is an angle, and $optQ$is the best current solution to guide the search process towards the best solution.

The basic formula of GD Optimization is,

$$M_s^{t+1} = M_s^t - \delta\nabla f\left(M_s^t\right) \tag{18}$$

$$M_s^t = \delta\nabla f\left(M_s^t\right) + M_s^{t+1} \tag{19}$$

where, $f$indicates the quasi function, $\delta$is a parameter for scaling gradient

Hybridization of GD Optimization with TSA is given by substituting Eq. (19) in Eq. (15),

$$M_s^{t+1} = \left(\delta\nabla f\left(M_s^t\right) + M_s^{t+1}\right)[1 + step * \tan(\phi)] - step * \tan(\phi) * optQ_s^t \tag{20}$$

$$M_s^{t+1} - M_s^{t+1}[1 + step * \tan(\phi)] = \delta\nabla f\left(M_s^t\right)[1 + step * \tan(\phi)] - step * \tan(\phi) * optQ_s^t \tag{21}$$

$$M_s^{t+1}[1 - 1 - step * \tan(\phi)] = \delta\nabla f\left(M_s^t\right)[1 + step * \tan(\phi)] - step * \tan(\phi) * optQ_s^t \tag{22}$$

$$M_s^{t+1} = \frac{\delta\nabla f\left(M_s^t\right)[1 + step * \tan(\phi)] - step * \tan(\phi) * optQ_s^t}{-step * \tan(\phi)} \tag{23}$$

$$M_s^{t+1} = \frac{-1}{-step * \tan(\phi)}\left[-\delta\nabla f\left(M_s^t\right)[1 + step * \tan(\phi)] + step * \tan(\phi) * optQ_s^t\right] \tag{24}$$

$$M_s^{t+1} = \frac{1}{step * \tan(\phi)}\left[step * \tan(\phi) * optQ_s^t - \delta\nabla f\left(M_s^t\right)[1 + step * \tan(\phi)]\right] \tag{25}$$

This is the basic equation of GTSO that train voice transformer for voice recognition. where, $M_s^t$is the position of sth solution at iteration t, and $\tan(\phi)$is an angle.

Step 4: Exploration search

TSA has a strong aptitude for exploration and creates global random walks using the tangent flight product and variable step size. The tangent function effectively facilitates search space research. Merging of global as well as local random walk is indicated in below formula representing the exploration search equation,

$$M_s^{t+1} = M_s^t + step * \tan(\phi) \tag{26}$$

When $\phi$is nearer to $uv/2$, then the tangent value is considered bigger, and gained solution remains far from the present solution, and when $\phi$is nearer to 0, then the tangent value is considered small, and gained solution remains nearer to the present solution. Also, $\phi = uv/2$ diverges TSA, and in exploration search ϕranges$[0, uv/3]$. Moreover, intensification and exploration search is based on parameter*Uswitch*.

Step 5: Escape local minima procedure

TSA utilizes specific means to escape from local minima stagnation issues. This procedure consists term *Uesc* and agent search *Y* is selected and followed in any one of the below equations,

$$M = M + Y * (optQ - rand * (optQ - M)) \tag{27}$$

$$M = M + \tan(\phi) * (N_{ub} - N_{lb}) \tag{28}$$

where, *M* is the current solution, *Y* is agent search, and*optQ*is the best current solution for guiding the search process toward an optimal solution.

Step 6: Termination

The fitness function is evaluated until an optimal solution is reached for voice recognition by a voice transformer, trained by GTSO. The fitness function is indicated as in Eq. (14) and thus the best solution is gained and the process is terminated. Algorithm 1 represents pseudo code of GTSO that trains the voice transformer.

Thus, the recognized speech from the voice transformer is indicated asK_α.

3.6. Language and pronunciation model

Recognized speech $K_\alpha$is allowed for language and pronunciation model in which speech is converted to text (Akita and Kawahara, 2009). Statistical properties of spontaneous speech are modeled independently from task-dependent factors in pronunciation and language models in order to represent spontaneous speech. Spoken-style features are being anticipated to fit Large Vocabulary Continuous Speech Recognition (LVCSR) target domain. In order to derive spoken style models, generic transitions between spoken and orthographic-document styles are modeled. Due to the framework's task independence, this transformation is used for a variety of tasks involving the recognition of spontaneous speech. By lining up phonetic and orthographic transcriptions, pronunciation differences are discovered, and variation patterns are recovered. The transformation model for pronunciation creates genuine pronunciation (surface form) entries from words present in the documents' orthodox pronunciation (baseform). This LVCSR model also forecasts and assigns pronunciation probability to each pronunciation entry. Here, correspondences between orthographic expressions and spontaneous speech events are primarily addressed, and they are retrieved in a broader way than word-level mappings. As a result, framework is anticipated to model transformation accurately and effectively with little quantity of training data. Basic formula for statistical transformation using Baye's rule is indicated below,

$$\mathrm{P(K\,/\,M)} = \frac{\mathrm{P(M/K)P(K)}}{\mathrm{P(M)}} \tag{29}$$

where, Kis the target language sentence, M is the source language sentence, and $\mathrm{P(M/K)}$is computed with the translation model.

4. Results and discussion

GTSO- enabled voice transformer is analyzed with many performance metrics and the results are explained in this section.

4.1. Experimental setup

GTSO- enabled voice transformer is implemented in a Python tool with three evaluation metrics utilizing Talkbank dataset.

4.2. Dataset description

This paper utilized the Talkbank dataset will be taken from (2023) for speech intelligibility in the case of aphasia. At Carnegie Mellon University, Brian MacWhinney created the TalkBank initiative with the help and support of hundreds of contributors and partners, including members of the TalkBank Governing Board. TalkBank's mission is to advance fundamental research in the field of human communication, with a focus on spoken language. TalkBank now offers repositories in 14 study fields, and its data have been supplied by hundreds of academics worldwide who are dedicated to the principles of open data sharing and who work in over 34 different languages. Several thousands of publications have been written as a result of the utilization of these data by thousands of scholars. The consistent XML-compatible CHAT representation of data in TalkBank enables automatic analysis and searching with free and open-source software. From 1999 until 2004, TalkBank was supported by a grant from the National Science Foundation to Carnegie Mellon University and the University of Pennsylvania.

4.3. Performance metrics

In this paper, the GTSO- enabled voice transformer was analyzed for its performance by three metrics, like recognition accuracy,

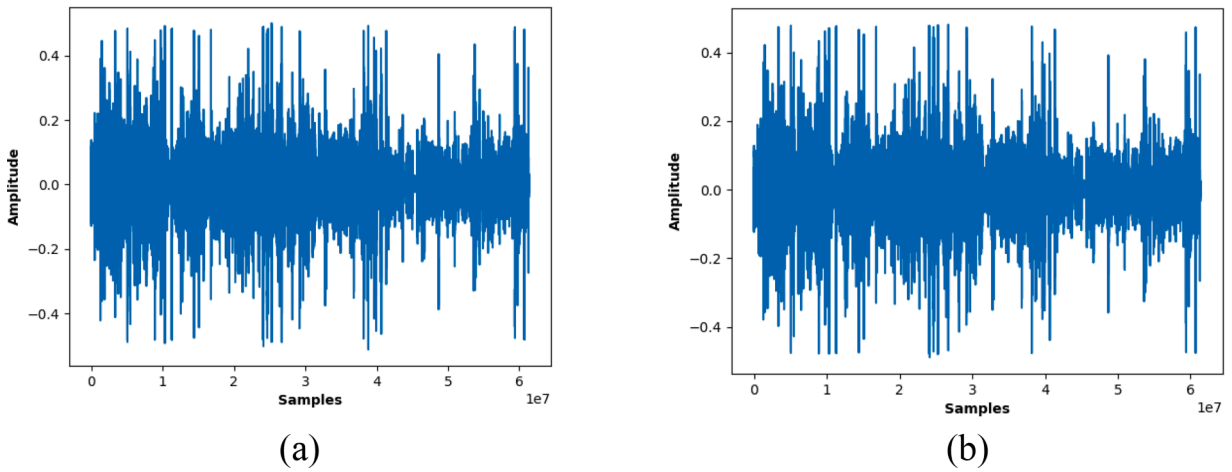


Fig. 3. Experimental outcomes of signal for speech intelligibility for aphasia (a) input signal, (b) pre-processed signal.

PPV, as well as NPV. These metrics are explained in detail below,

4.3.1. Recognition accuracy

An indicator of the model's performance across every class is accuracy. When all classes are important, accuracy is a highly intuitive metric. This is the ratio of the number of correct predictions to buy the total number of predictions regarding aphasia patients. This is formulated as,

$$R_{accu} = \frac{Pos_t + Neg_t}{Pos_t + Neg_t + Pos_f + Neg_f} \tag{30}$$

where, $R_{accu}$is recognition accuracy, $Pos_t$is true positive, $Neg_t$is true negative, $Pos_f$is false positive, and $Neg_f$is false negative.

4.3.2. PPV

It is the proportion of aphasia patients with actual positive diagnoses to all patients with positive test findings that include healthy subjects who were incorrectly diagnosed as patients. If the test is positive, this trait might indicate whether a person will actually be aphasia patient or not. This is represented in the below formula,

$$P_{pv} = \frac{Pos_t}{Pos_t + Pos_f} \tag{31}$$

where, $P_{pv}$is PPV.

4.3.3. NPV

NPV is the probability that subjects with negative screening tests truly don't have the disease. A person's chance of not having the disease, ailment, test-related biomarker, or gene mutation that led to a negative result. This NPV indicates how accurate a certain test may be determined and is represented as,

$$N_{pv} = \frac{Neg_t}{Neg_t + Neg_f} \tag{32}$$

where, N_{pv}indicates NPV.

4.4. Experimental outcomes

Fig. 3 indicates experimental outcomes of signal for speech intelligibility for aphasia. Fig. 3(a) represents the input signal, and Fig. 3 (b) indicates pre-processed signal by the median filter.

4.5. Comparative analysis

Various comparative methods are used for checking the performance of the GTSO- enabled voice transformer and the methods used are DNN acoustic model (Le et al., 2016), CNN (Mahmoud et al., 2020), DNN (Qin et al., 2019), and ResNet-34 pre-trained CNN model (Mahmoud et al., 2021). A comparison of the GTSO- enabled voice transformer is analyzed based on four videos in terms of three metrics. Here, an audio segment is taken on X-axis for analyzing the performance of the GTSO-enabled voice transformer.

Table 1
Comparative assessment based on video-1.

Methods/ Audio segments	DNN acoustic model Recognition accuracy	CNN	DNN	ResNet-34 pre-trained CNN model	GTSO- enabled voice transformer
1	0.693	0.719	0.733	0.747	0.774
2	0.712	0.725	0.770	0.805	0.826
3	0.726	0.769	0.795	0.841	0.860
4	0.773	0.786	0.826	0.845	0.883
5	0.800	0.828	0.888	0.893	0.919
PPV					
1	0.693	0.709	0.739	0.760	0.782
2	0.711	0.729	0.763	0.781	0.823
3	0.721	0.762	0.805	0.816	0.857
4	0.773	0.801	0.812	0.820	0.871
5	0.781	0.832	0.847	0.865	0.919
NPV					
1	0.671	0.703	0.737	0.758	0.788
2	0.718	0.727	0.765	0.802	0.834
3	0.729	0.766	0.799	0.832	0.850
4	0.772	0.784	0.831	0.857	0.887
5	0.799	0.833	0.862	0.877	0.915

Table 2
Comparative assessment based on video-2.

Methods/ Audio segments	DNN acoustic model Recognition accuracy	CNN	DNN	ResNet-34 pre-trained CNN model	GTSO- enabled voice transformer
1	0.687	0.709	0.724	0.756	0.800
2	0.714	0.723	0.742	0.789	0.821
3	0.731	0.767	0.797	0.847	0.870
4	0.776	0.794	0.814	0.851	0.887
5	0.800	0.842	0.867	0.877	0.911
PPV					
1	0.699	0.717	0.737	0.742	0.786
2	0.709	0.720	0.747	0.776	0.811
3	0.737	0.750	0.776	0.822	0.866
4	0.774	0.806	0.816	0.858	0.883
5	0.806	0.843	0.882	0.898	0.917
NPV					
1	0.699	0.708	0.727	0.758	0.806
2	0.712	0.722	0.761	0.798	0.818
3	0.737	0.754	0.785	0.810	0.869
4	0.774	0.809	0.830	0.851	0.874
5	0.787	0.845	0.880	0.890	0.916

Table 3
Comparative assessment based on video-3.

Methods/ Audio segments	DNN acoustic model Recognition accuracy	CNN	DNN	ResNet-34 pre-trained CNN model	GTSO- enabled voice transformer
1	0.694	0.710	0.735	0.742	0.788
2	0.705	0.732	0.753	0.790	0.832
3	0.723	0.764	0.806	0.821	0.861
4	0.779	0.797	0.835	0.846	0.886
5	0.806	0.831	0.890	0.890	0.917
PPV					
1	0.688	0.708	0.739	0.744	0.776
2	0.708	0.735	0.758	0.809	0.835
3	0.722	0.766	0.786	0.835	0.860
4	0.770	0.805	0.833	0.852	0.884
5	0.807	0.842	0.859	0.879	0.915
NPV					
1	0.688	0.715	0.728	0.754	0.803
2	0.701	0.721	0.761	0.792	0.840
3	0.721	0.747	0.779	0.823	0.859
4	0.764	0.784	0.831	0.844	0.873
5	0.806	0.843	0.883	0.890	0.912

Table 4
Comparative assessment based on video-4.

Methods/ Audio segments	DNN acoustic model Recognition accuracy	CNN	DNN	ResNet-34 pre-trained CNN model	GTSO- enabled voice transformer
1	0.686	0.718	0.737	0.746	0.789
2	0.719	0.729	0.757	0.803	0.827
3	0.723	0.757	0.787	0.811	0.866
4	0.760	0.786	0.836	0.866	0.887
5	0.798	0.837	0.875	0.888	0.915
PPV					
1	0.670	0.702	0.740	0.759	0.794
2	0.709	0.737	0.757	0.807	0.828
3	0.724	0.750	0.798	0.820	0.852
4	0.751	0.801	0.837	0.845	0.880
5	0.802	0.835	0.861	0.887	0.916
NPV					
1	0.699	0.700	0.733	0.768	0.799
2	0.716	0.735	0.760	0.774	0.832
3	0.726	0.746	0.782	0.839	0.858
4	0.777	0.802	0.824	0.850	0.882
5	0.796	0.825	0.865	0.872	0.919

4.5.1. Comparative assessment based on video-1

Table 1 indicates a comparative assessment of the GTSO- enabled voice transformer based on video-1. The highest recognition accuracy, PPV, and NPV obtained by the GTSO- enabled voice transformer is 0.919, 0.919, and 0.915, when considering the audio segments=5.

4.5.2. Comparative assessment based on video-2

Table 2 indicates a comparative assessment of the GTSO- enabled voice transformer based on video-2. The highest recognition accuracy, PPV, and NPV obtained by the GTSO- enabled voice transformer is 0.911, 0.917, and 0.916, when considering the audio segments=5.

4.5.3. Comparative assessment based on video-3

Table 3 indicates a comparative assessment of the GTSO- enabled voice transformer based on video-3. The highest recognition accuracy, PPV, and NPV obtained by the GTSO- enabled voice transformer is 0.917, 0.915, and 0.912, when considering the audio segments=5.

4.5.4. Comparative assessment based on video-4

Table 4 indicates a comparative assessment of the GTSO- enabled voice transformer based on video-4. The highest recognition accuracy, PPV, and NPV obtained by the GTSO- enabled voice transformer is 0.915, 0.916, and 0.919, when considering the audio segments=5.

5. Conclusion

The degree to which speaker's utterances is understood by listeners or speech intelligibility is a very important concept in speech language pathology. This speech intelligibility for aphasia is investigated in this work by utilizing a voice transformer. Here, the voice transformer is trained by GTSO, formed by combining both TSA and GD Optimization. Moreover, pre-processing phase is fulfilled by the application of the median filter for the input signal that is taken from the dataset. Pre-processing helps in removing unwanted noises from the input signal. Then, the voice enhancement is done by nonlinear spectral subtraction and then voice recognition is carried out by a voice transformer. This identified speech is forwarded toward the language and pronunciation model in which text is created from speech. Finally, the performance of the proposed GTSO- enabled voice transformer is analyzed with three evaluation metrics, such as recognition accuracy, PPV, and NPV with high values of 0.919, 0.919, and 0.915. In future, this work can be enhanced by applying some other better optimization algorithms for training voice transformers to recognize the perfect voice in case of aphasia patients.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgment

This work was supported by the AICTE, Government of India through Research Promotion Scheme File No. 8-100/FDC/RPS/POLICY-1/ 2021-22.

References

Adam, H., 2014. Dysprosody in aphasia: an acoustic analysis evidence from Palestinian Arabic. J. Lang. Linguist. Stud. 10 (11), 153–162.
Akita, Y., Kawahara, T., 2009. Statistical transformation of language and pronunciation models for spontaneous speech recognition. IEEE Trans. Audio Speech Lang. Process. 18 (6), 1539–1549.
Anastasopoulos, A., Chiang, D., 2018. Tied multitask learning for neural speech translation. In the. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics. New Orleans, Louisiana, pp. 82–91. Vol. 1.
Baevski, A., Zhou, H., Mohamed, A., Auli, M., 2020. wav2vec 2.0: a framework for self-supervised learning of speech representations. Comput. Lang. 12449–12460.
Damasio, A.R., 1992. Aphasia. N. Engl. J. Med. 326 (8), 531–539.
Dong, L., Xu, S., Xu, B., 2018. Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition. In: Proceeding of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5884–5888. April.
Estes, C., Bloom, R.L., 2011. Using voice recognition software to treat dysgraphia in a patient with conduction aphasia. Aphasiology 25 (3), 366–385.
Gnanamanickam, J., Natarajan, Y., K.R, Sri Preethaa, 2021. A hybrid speech enhancement algorithm for voice assistance application. Sensors 21 (21), 7025.
Gradient descent optimization is taken from, 2021 "https://towardsdatascience.com/gradient-descent-algorithm-a-deep-dive-cf04e8115f21".
Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., Pang, R., 2020. Conformer: convolution-augmented transformer for speech recognition. In the. In: Proceeding of Interspeech.
Herath, H.M.D.P.M., Weraniyagoda, W.A.S.A., Rajapaksha, R.T.M., Wijesekara, P.A.D.S.N., Sudheera, K.L.K., Chong, P.H.J., 2022. Automatic assessment of aphasic speech sensed by audio sensors for classification into aphasia severity levels to recommend speech therapies. Sensors 22 (18), 6966.
Herzog, S., 2013. Efficient DSP implementation of median filtering for real-time audio noise reduction. In: Proceedings of the international conference on Digital Audio Effects, pp. 1–6. September.
Hsu, W.-N., Bolte, B., Tsai, Y.-H.H., Lakhotia, K., Salakhutdinov, R., Mohamed, A., 2021. HuBERT: self-supervised speech representation learning by masked prediction of hidden units. Comput. Lang. 29, 3451–3460.
Janbakhshi, P., Kodrasi, I., Bourlard, H., 2019. Pathological speech intelligibility assessment based on the short-time objective intelligibility measure. In the. In: Proceeding of ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6405–6409. May.
Landa, S., Pennington, L., Miller, N., Robson, S., Thompson, V., Steen, N., 2014. Association between objective measurement of the speech intelligibility of young people with dysarthria and listener ratings of ease of understanding. Int. J. Speech Lang. Pathol. 16 (4), 408–416.
Layeb, A., 2022. Tangent search algorithm for solving optimization problems. Neural Comput. Appl. 34 (11), 8853–8884.
Le, D., Licata, K., Persad, C., Provost, E.M., 2016. Automatic assessment of speech intelligibility for individuals with aphasia. IEEE/ACM Trans. Audio Speech Lang. Process. 24 (11), 2187–2199.
Le, D., Licata, K., Provost, E.M., 2018. Automatic quantitative analysis of spontaneous aphasic speech. Speech Commun. 100, 1–12.
Liao, L., Kwofie, F.A., Chen, Z., Han, G., Wang, Y., Lin, Y., Hu, D., 2022. A bidirectional context embedding transformer for automatic speech recognition. Information 13 (2), 69.
Mahmoud, S.S., Kumar, A., Li, Y., Tang, Y., Fang, Q., 2021. Performance evaluation of machine learning frameworks for aphasia assessment. Sensors 21 (8), 2582.
Mahmoud, S.S., Kumar, A., Tang, Y., Li, Y., Gu, X., Fu, J., Fang, Q., 2020. An efficient deep learning based method for speech assessment of mandarin-speaking aphasic patients. IEEE J. Biomed. Health Inform. 24 (11), 3191–3202.
Nguyen, P., Tran, D., Huang, X., Sharma, D., 2012. A proposed feature extraction method for EEG-based person identification. In: Proceedings on the International Conference on Artificial Intelligence (ICAI), The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), p. 1.
Prabhavalkar, R., Hori, T., Sainath, T.N., Schlüter, R., Watanabe, S., 2023. End-to-end speech recognition: a survey. In the. In: Proceeding of the IEEE/ACM Transactions on Audio, Speech, and Language Processing.
Qin, Y., Lee, T., Feng, S., Kong, A.P.H., 2018. Automatic speech assessment for people with aphasia using TDNN-BLSTM with multi-task learning. In: Proceedings of the Interspeech, pp. 3418–3422. September.
Qin, Y., Lee, T., Kong, A.P.H., 2019. Automatic assessment of speech impairment in cantonese-speaking people with aphasia. IEEE J. Sel. Top. Signal Process. 14 (11), 331–345.
Qin, Y., Wu, Y., Lee, T., Kong, A.P.H., 2020. An end-to-end approach to automatic speech assessment for Cantonese-speaking people with aphasia. J. Signal Process. Syst. 92, 819–830.
Ruder, S., "An overview of gradient descent optimization algorithms", arXiv preprint arXiv:1609.04747, 2016.
Sandhya, P., Spoorthy, V., Koolagudi, S.G., Sobhana, N.V., 2020. Spectral features for emotional speaker recognition. In: Proceedings of the 2020 Third International Conference on Advances in Electronics, Computers and Communications (ICAECC), pp. 1–6. December.
Talkbank dataset will be taken from "https://talkbank.org/DB/", accessed on March 2023.
Torre, I.G., Romero, M., Álvarez, A., 2021. Improving aphasic speech recognition by using novel semi-supervised learning methods on aphasia bank for English and Spanish. Appl. Sci. 11 (19), 8872.
Xu, Y., Du, J., Dai, L.R., Lee, C.H., 2014. A regression approach to speech enhancement based on deep neural networks. IEEE/ACM Trans. Audio Speech Lang. Process. 23 (1), 7–19.
Yiu, E.M., 1992. Linguistic assessment of Chinese-speaking aphasics: development of a cantonese aphasia battery. J. Neurolinguistics 7 (4), 379–424.
Zhao, C., Wang, J., Wei, W., Qu, X., Wang, H., Xiao, J., 2022. Adaptive sparse and monotonic attention for transformer-based automatic speech recognition. In the. In: Proceeding of the IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA). Shenzhen, China.