Korean J Neurotrauma. 2024 Sep;20(3):168-179
https://doi.org/10.13004/kjnt.2024.20.e30
pISSN 2234-8999·eISSN 2288-2243

Laboratory Investigation

Feasibility Study of Parkinson's Speech Disorder Evaluation With Pre-Trained Deep Learning Model for Speech-to-Text Analysis

Kwang Hyeon Kim [1], **Byung-Jou Lee** [2], **and Hae-Won Koo** [2]

[1]Clinical Research Support Center, Inje University Ilsan Paik Hospital, Inje University College of Medicine, Goyang, Korea
[2]Department of Neurosurgery, Inje University Ilsan Paik Hospital, Inje University College of Medicine, Goyang, Korea

Received: Apr 26, 2024
Revised: Jun 18, 2024
Accepted: Jun 23, 2024
Published online: Sep 23, 2024

Address for correspondence:
Hae-Won Koo
Department of Neurosurgery, Inje University Ilsan Paik Hospital, Inje University College of Medicine, 170 Juhwa-ro, Ilsanseo-gu, Goyang 10380, Korea.
Email: hwkoo@paik.ac.kr

ORCID iDs
Kwang Hyeon Kim
https://orcid.org/0000-0003-0434-1905
Byung-Jou Lee
https://orcid.org/0000-0002-4030-3618
Hae-Won Koo
https://orcid.org/0000-0001-7014-3005

Funding
No funding was obtained for this study.

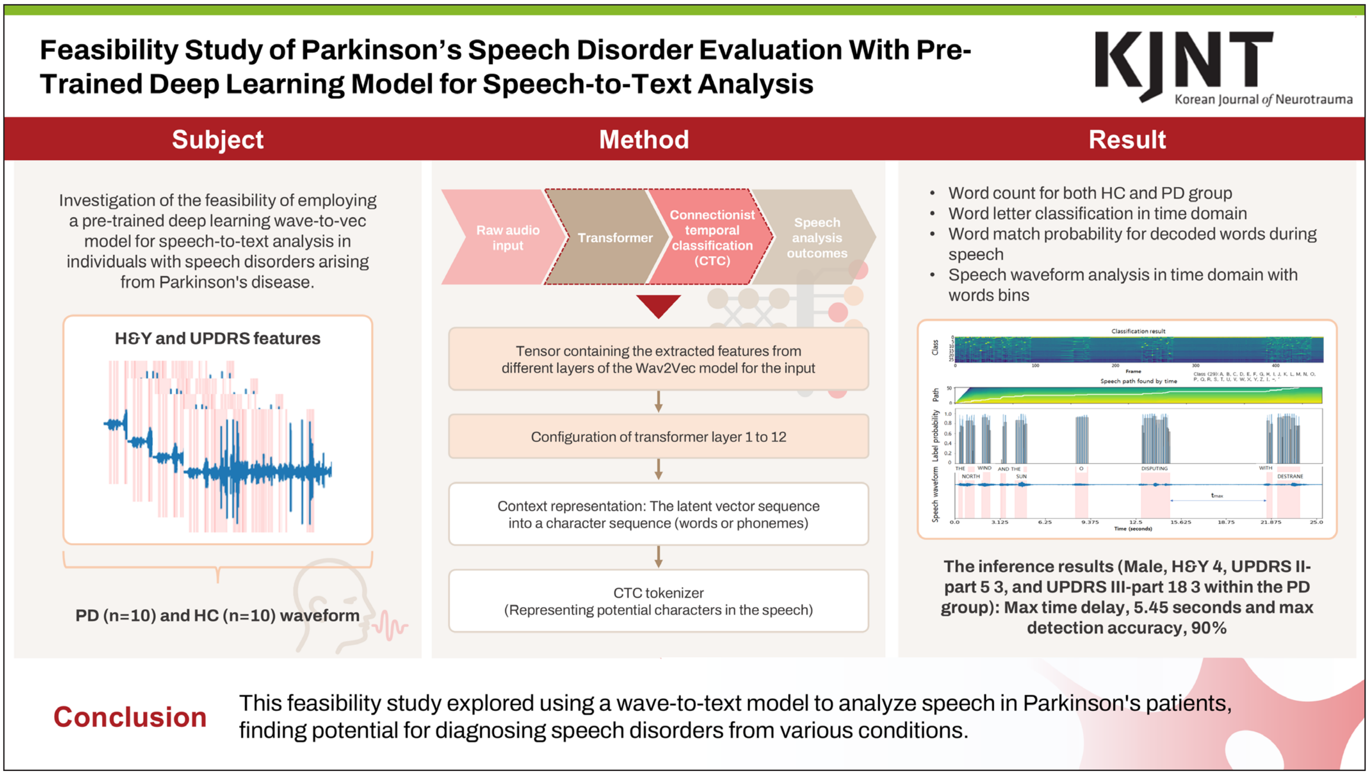
Conflict of Interest
The authors have no financial conflicts of interest.

ABSTRACT

Objective: This study investigates the feasibility of employing a pre-trained deep learning wave-to-vec model for speech-to-text analysis in individuals with speech disorders arising from Parkinson's disease (PD).
Methods: A publicly available dataset containing speech recordings including the Hoehn and Yahr (H&Y) staging, Movement Disorder Society Unified Parkinson's Disease Rating Scale (UPDRS) Part I, UPDRS Part II scores, and gender information from both healthy controls (HC) and those diagnosed with PD was utilized. Employing the Wav2Vec model, a speech-to-text analysis method was implemented on PD patient data. Tasks conducted included word letter classification, word match probability assessment, and analysis of speech waveform characteristics as provided by the model's output.
Results: For the dataset comprising 20 cases, among individuals with PD, the H&Y score averaged 2.50±0.67, the UPDRS II-part 5 score averaged 0.70±1.00, and the UPDRS III-part 18 score averaged 0.80±0.98. Additionally, the number of words derived from decoded text subsequent to speech recognition was evaluated, resulting in mean values of 299.10±16.79 and 259.80±93.39 for the HC and PD groups, respectively. Furthermore, the calculated degree of agreement for all syllables was based on the speech process. The accuracy for the reading sentences was observed to be 0.31 and 0.10, respectively.
Conclusion: This study aimed to demonstrate the effectiveness of wave-to-vec in enhancing speech-to-text analysis for patients with speech disorders. The findings could pave the way for the development of clinical tools for improved diagnosis, evaluation, and communication support for this population.

Keywords: Speech disorders; Parkinson disease; Speech-to-text; Traumatic brain injury; Wav-to-Vec; Deep learning

Graphical Abstract



Feasibility Study of Parkinson's Speech Disorder Evaluation With Pre-Trained Deep Learning Model for Speech-to-Text Analysis

KJNT Korean Journal *of* Neurotrauma

Subject

Investigation of the feasibility of employing a pre-trained deep learning wave-to-vec model for speech-to-text analysis in individuals with speech disorders arising from Parkinson's disease.

H&Y and UPDRS features

PD (n=10) and HC (n=10) waveform

Method

Raw audio input → Transformer → Connectionist temporal classification (CTC) → Speech analysis outcomes

Tensor containing the extracted features from different layers of the Wav2Vec model for the input

Configuration of transformer layer 1 to 12

Context representation: The latent vector sequence into a character sequence (words or phonemes)

CTC tokenizer (Representing potential characters in the speech)

Result

- Word count for both HC and PD group
- Word letter classification in time domain
- Word match probability for decoded words during speech
- Speech waveform analysis in time domain with words bins

The inference results (Male, H&Y 4, UPDRS II-part 5 3, and UPDRS III-part 18 3 within the PD group): Max time delay, 5.45 seconds and max detection accuracy, 90%

Conclusion This feasibility study explored using a wave-to-text model to analyze speech in Parkinson's patients, finding potential for diagnosing speech disorders from various conditions.

Informed Consent

This type of study does not require informed consent.

Ethics Approval

This research did not require ethical approval as it does not involve human subjects, their data, or biological samples.

INTRODUCTION

Speech disorders can arise from a variety of medical conditions affecting the neurological, structural, or functional components of speech production.[10,27] Neurological disorders such as stroke, traumatic brain injury (TBI), and Parkinson's disease (PD) disrupt the central or peripheral nervous system's control over the muscles involved in speech articulation, resulting in dysarthria characterized by slurred speech, imprecise articulation, and reduced intelligibility. Structural abnormalities like cleft lip and palate, vocal fold paralysis, and structural damage from trauma or surgery interfere with the anatomical integrity or mobility of the speech organs, leading to articulatory difficulties, phonatory disturbances, or resonance issues.[4,13,21,30] Functional disorders such as developmental speech and language disorders, stuttering, and psychogenic speech disorders stem from cognitive, psychological, or behavioral factors affecting speech production, manifesting as disruptions in fluency, rhythm, or prosody.[4,13] These diverse etiologies disrupt the complex processes of speech production, including phonation, articulation, resonance, and prosody, resulting in a wide range of speech impairments that require comprehensive assessment and intervention tailored to the underlying pathology and individual needs of the affected individuals.[27]

In diagnosing PD, clinicians often employ standardized assessment tools such as the Hoehn and Yahr (H&Y) scale and the Unified Parkinson's Disease Rating Scale (UPDRS) to evaluate various aspects of motor function, including speech disorders.[10,12,24,28] The H&Y scale categorizes PD progression into stages based on motor symptoms' severity, ranging from stage 1 (mild symptoms affecting one side of the body) to stage 5 (severe symptoms affecting both sides of the body and possibly requiring assistance for mobility).[28] The UPDRS further assesses specific motor and non-motor symptoms, including speech disorders, through subjective ratings and objective evaluations.[28] Speech disorders in PD commonly include hypophonia (reduced speech volume), dysarthria (impaired articulation and clarity), and monotone speech.[10] Clinicians utilize standardized tasks such as reading a standard passage or counting to assess speech intelligibility, prosody, and phonation quality. Additionally, acoustic analysis tools may quantify speech characteristics such as pitch variability, speech rate, and pauses. Integration of H&Y and UPDRS scores, along with detailed speech assessment findings, contributes to a comprehensive diagnosis and monitoring of PD progression, facilitating tailored treatment strategies and interventions to address speech impairments and enhance overall quality of life for individuals living with PD.

Recent advancements in artificial intelligence (AI) have shown promise in aiding the diagnosis and grading of PD.[5] Khan et al.[17] employed machine learning algorithms to examine video recordings (13 patients), aiming to quantify tapping symptoms by analyzing the motion of index fingers (finger tapping; FT). Their approach stands out for its utilization of facial characteristics to adjust tapping intensity, ensuring normalization of distance discrepancies between the camera and the subject. In their study, the cross-correlation analysis of the normalized peaks revealed a robust Guttman correlation (μ_2=−0.80) with the clinical ratings. Employing a support vector machine (SVM) classifier and implementing 10-fold cross-validation, the classification of tapping characteristics accurately categorized patient samples across different UPDRS-FT levels with an 88% accuracy rate. Moreover, research has been conducted to explore the potential application of phonetic approaches for diagnostic purposes, involving the analysis of distinctive features among individuals afflicted with speech disorders.[19,23]

Additionally, AI-based approaches offer the advantage of scalability and accessibility, as they can be deployed through smartphone applications or web-based platforms for remote monitoring and screening of PD symptoms.[5]

The aim of this study is to investigate the feasibility of integrating speech recognition deep learning alongside acoustic analysis for converting speech to text among individuals experiencing speech impairments attributable to PD, stroke, multiple sclerosis, and TBI.

MATERIALS AND METHODS

Dataset

We used the open-source dataset of Mobile Device Voice Recordings at King’s College London from both early and advanced PD patients and healthy controls (HC).[16] The dataset was collected at King's College London Hospital, situated in Denmark Hill, Brixton, London, during the timeframe spanning from September 26th to September 29th, 2017. Utilizing a conventional examination facility encompassing approximately 10 m^2 and characterized by a typical reverberation duration of approximately 500 ms, voice recordings were conducted.[16] The dataset has been partitioned into segments consisting of reading text and spontaneous dialogue. In this investigation, a collective sum of 20 datasets, encompassing 10 subjects from a healthy cohort and 10 from a cohort afflicted with PD, were employed from the reading text segment of the dataset (n=20).

Dataset has the evaluation metrics for H&Y staging, Movement Disorder Society UPDRS Part I (Mentation, Behavior, and Mood), and UPDRS Part II (Activities of Daily Living) scores.[16] First, H&Y stage is a 5-point scale that describes the overall severity of PD based on functional limitations.[12,24] For example, an H&Y rating of 1 indicates early-stage PD. People at this stage may experience mild tremors or stiffness, but they are generally independent in their daily activities. Second, UPDRS is a more detailed scale that assesses various aspects of PD, including: UPDRS Part I for scores here evaluate non-motor symptoms like depression or cognitive decline.[12] For example, a score of 2 is relatively low and suggests minimal impairment. And UPDRS Part II is a part assesses how PD affects daily tasks like dressing, bathing, and eating.[24] A score of 2 suggests some mild difficulties but likely still manageable independence. Furthermore, UPDRS Part III is the most focused on movement and evaluates things like tremor, rigidity, and balance.[28]

Speech-to-text analysis via Wav2Vec deep learning model

We implemented speech-to-text analysis method using Wav2Vec model in patient cases with PD dataset (**FIGURE 1**).[3] And the process shows obtaining speech-to-text analysis outcomes using the Wav2Vec inference model to analyze the characteristics between PD and HC. The Wav2Vec model, fine-tuned for automatic speech recognition (ASR) tasks, is a model library that can perform feature extraction and classification in one step.[3] To initiate the process of speech recognition for the dataset using Wav2Vec, the first step involves inputting the healthy group and PD data into the system in the second pipeline of **FIGURE 1**. This typically entails obtaining audio recordings of speech utterances, which can be sourced from various recording devices such as microphones or digital recorders. The quality of the recordings is crucial, as clear

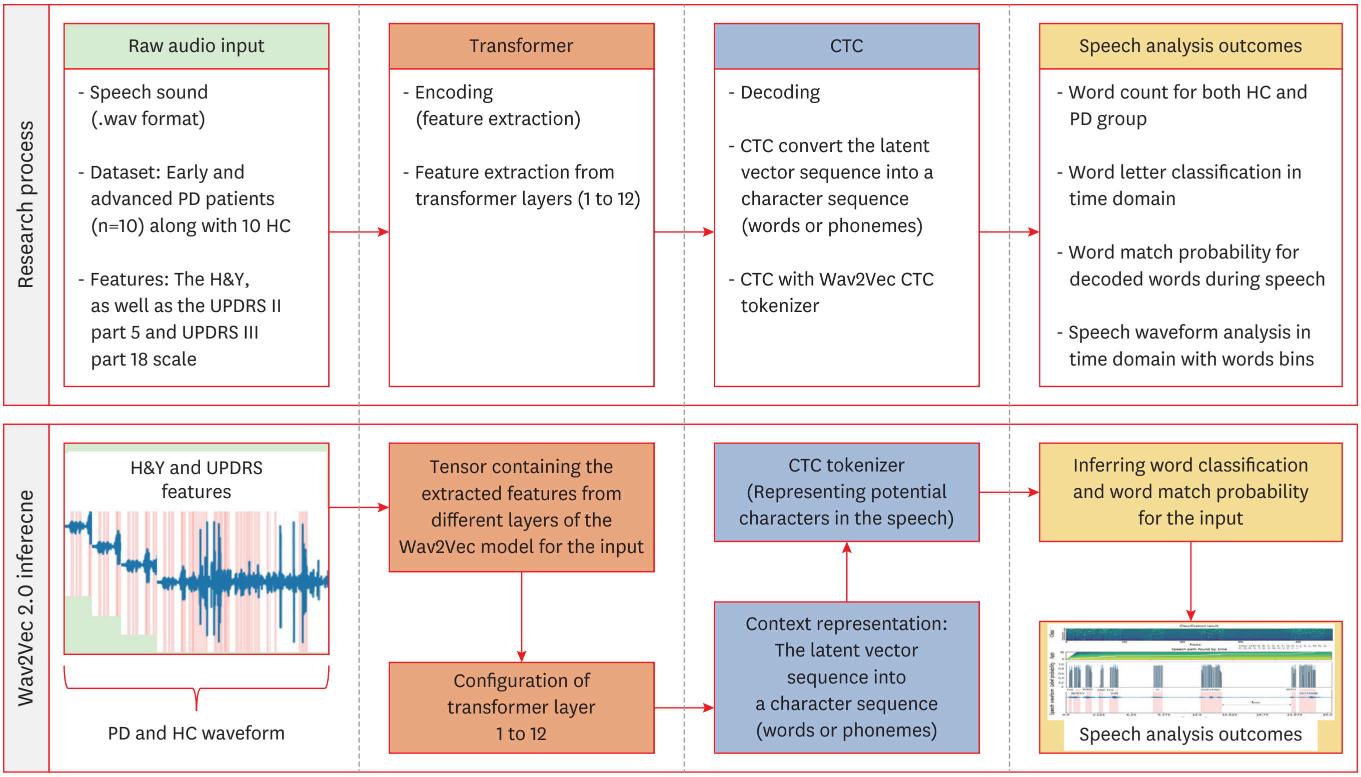

Research process
Raw audio input
- Speech sound (.wav format)
- Dataset: Early and advanced PD patients (n=10) along with 10 HC
- Features: The H&Y, as well as the UPDRS II part 5 and UPDRS III part 18 scale
Transformer
- Encoding (feature extraction)
- Feature extraction from transformer layers (1 to 12)
CTC
- Decoding
- CTC convert the latent vector sequence into a character sequence (words or phonemes)
- CTC with Wav2Vec CTC tokenizer
Speech analysis outcomes
- Word count for both HC and PD group
- Word letter classification in time domain
- Word match probability for decoded words during speech
- Speech waveform analysis in time domain with words bins
Wav2Vec 2.0 inferecne
H&Y and UPDRS features
PD and HC waveform
Tensor containing the extracted features from different layers of the Wav2Vec model for the input
Configuration of transformer layer 1 to 12
CTC tokenizer (Representing potential characters in the speech)
Context representation: The latent vector sequence into a character sequence (words or phonemes)
Inferring word classification and word match probability for the input
Speech analysis outcomes

FIGURE 1. Research flow chart and the process of obtaining speech-to-text analysis outcomes using the Wav2Vec inference model to analyze the features between PD and HC for this study. HC and PD group were used and encoding and decoding process were employed in speech analysis procedure.
HC: healthy control, PD: Parkinson's disease, H&Y: Hoehn and Yahr, UPDRS: Unified Parkinson's Disease Rating Scale, CTC: Connectionist Temporal Classification.

and intelligible speech facilitates more accurate recognition outcomes. Once the audio data is acquired, it needs to be preprocessed to ensure compatibility with the Wav2Vec2 model.[2,3] Preprocessing of the voice data involves several steps aimed at preparing the audio signals for input into the Wav2Vec model. This includes standardizing the sampling rate of the audio files to match the requirements of the model, typically 16 kHz or 48 kHz. Additionally, any background noise or artifacts present in the recordings may need to be removed or attenuated through noise reduction techniques to enhance the clarity of the speech signal. Furthermore, the audio data may be segmented into smaller chunks or frames to facilitate processing by the Wav2Vec model. Once the voice data is appropriately preprocessed, it is ready for processing with the Wav2Vec model (also known as Wav2Vec2, short for Waveform-to-Vector 2). It operates by transforming raw audio waveforms into high-dimensional feature vectors, which are then fed into a neural network for further processing (**FIGURE 1**). The model consists of multiple layers of convolutional and transformer-based architectures, which learn to extract meaningful representations of speech features directly from the raw waveform for the PD and HC.

During inference, the preprocessed voice data is input into the Wav2Vec model, which generates a sequence of feature vectors representing the acoustic characteristics of the speech signal (**FIGURE 1**). These feature vectors are then passed through the model's encoder layers, where they are transformed into a higher-level representation that captures the temporal and contextual information of the speech. Finally, the output of the Wav2Vec model can be further processed by downstream components, such as a language model or a classifier, to perform tasks as word letter classification, word match probability, and speech waveform analysis (**FIGURE 1**).

Programming environment

This study utilized Python version 3.8.3, PyTorch version 1.2.0, and TorchAudio version 0.6.0, coupled with a GeForce RTX 4060 GPU for computational acceleration. The statistical analysis was performed using SPSS 21.0 for windows (SPSS Inc., Chicago, IL, USA).

RESULTS

Baseline characteristics

For the dataset comprising 20 cases, speech waveforms were processed using the Wav2Vec model followed by word count analysis. Subsequently, the baseline statistical findings are summarized, as presented in **TABLE 1**. Statistical analysis revealed significant effects (*p*-value

TABLE 1. Dataset characteristics (n=20)

Category	HC	*p*-value*	PD	*p*-value*
Number	10	N/A	10	N/A
Gender		0.686		0.642
Man	1 (10.00)		7 (70.00)	
Woman	9 (90.00)		3 (30.00)	
H&Y scores (mean, SD)	0.00±0.00	N/A	2.50±0.67	0.011
UPDRS II-part 5	0.00±0.00	N/A	0.70±1.00	0.000
UPDRS III-part 18	0.00±0.00	N/A	0.80±0.98	0.014
Speech time (minutes:seconds)	02:21±00:16	N/A	02:14±00:29	0.084
Number of words decoded from speech	299.10±16.79	0.119	259.80±93.39	N/A

Values are presented as number (%) or mean ± standard deviation.
HC: Healthy control, PD: Parkinson's disease, H&Y: Hoehn and Yahr, UPDRS: Unified Parkinson's Disease Rating Scale, N/A: not applicable.
*The *p*-value was obtained through the analysis of variance.

<0.05) for H&Y score, UPDRS part II score, and UPDRS part III score within the PD cohort. The analysis encompassed 10 participants each from the HC and PD cohorts. Among individuals with PD, the H&Y score averaged 2.50±0.67, the UPDRS II-part 5 score averaged 0.70±1.00, and the UPDRS III-part 18 score averaged 0.80±0.98. Additionally, the number of words derived from decoded text subsequent to speech recognition was evaluated, resulting in mean values of 299.10±16.79 and 259.80±93.39 for the HC and PD groups, respectively.

Speech-to-text analysis in time domain for HC and PD

The inference outcomes for the speech sound model output are depicted in **FIGURES 2** & **3**. **FIGURE 2** illustrates the inference findings for instances identified as female, H&Y 0, UPDRS II-part 5 score of 0, and UPDRS III-part 18 score of 0 within the HC group. We visually represented the corresponding class for each of 500 frames along the time axis of the waveform (**FIGURE 2A**). In this context, the classification is informed by the phonetic and linguistic attributes of the English language. Each symbol correlates with specific phonemes, which serve as the fundamental units of sound differentiation within a language, distinguishing one word from another. The alphabetical characters A to Z denote the

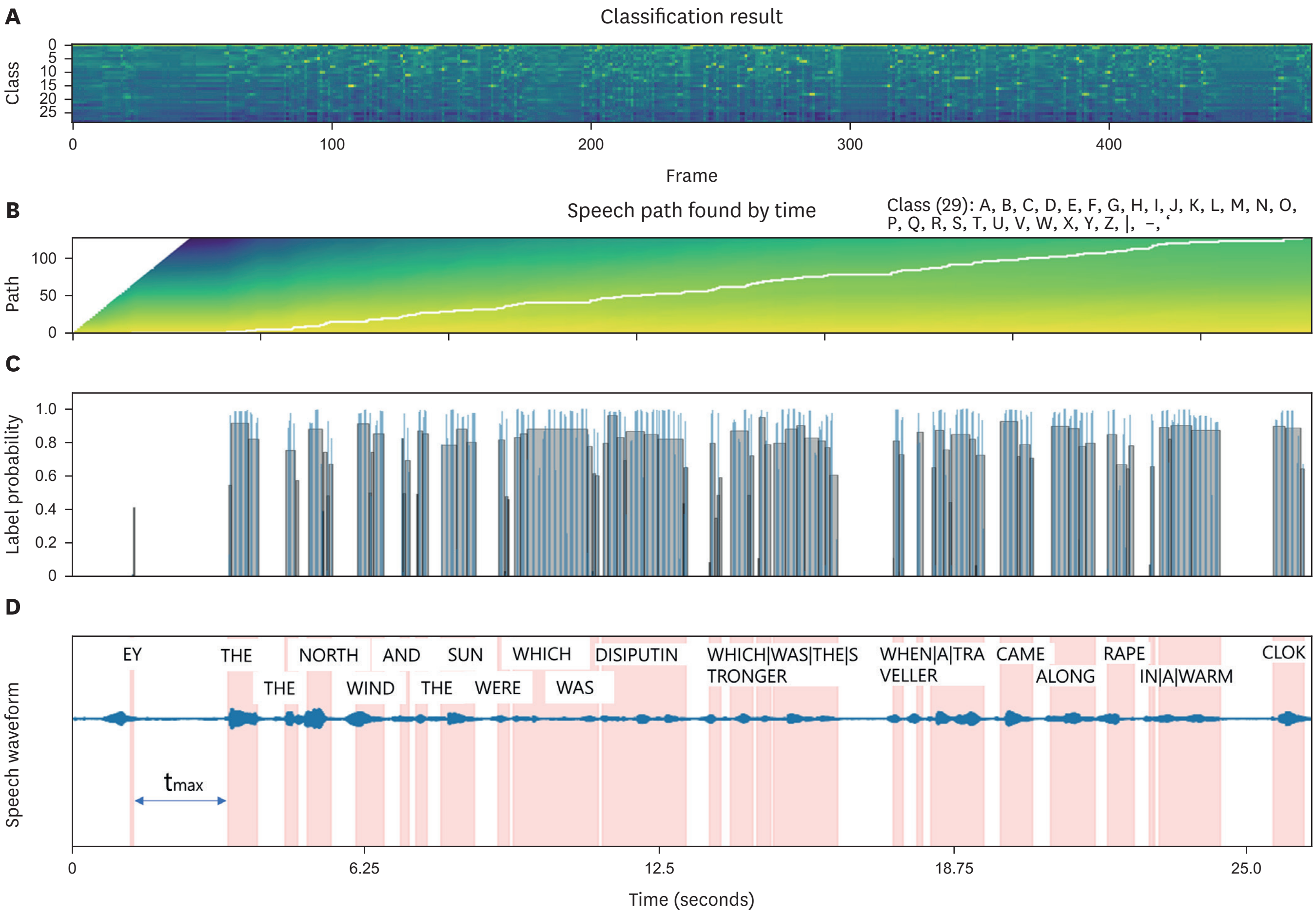


FIGURE 2. Visualization and analysis of inference results for speech sound model output in HC. (A) The inference findings for instances identified as female, H&Y 0, UPDRS II-part 5 score of 0, and UPDRS III-part 18 score of 0 within the HC group. The corresponding class for each of 500 frames along the time axis of the waveform. (B) The trajectory for detecting words decoded from vocalized sounds is visually depicted. (C) The probability at the detection juncture in the time domain. (D) The decoded words are superimposed onto the waveform within the time domain. And a maximum delay (t_{max}), 3.15 seconds was identified.
HC: healthy control, H&Y: Hoehn and Yahr, UPDRS: Unified Parkinson's Disease Rating Scale, t_{max}: maximum delay time.

recognition of individual speech sounds, encompassing consonants and vowels in English vocabulary. Furthermore, symbols such as | (pipe), - (hyphen), and ' (apostrophe) may signify diverse linguistic elements including word delimiters, pauses, or distinctive speech phenomena such as glottal stops or aspiration. In **FIGURE 2B**, the trajectory for detecting words decoded from vocalized sounds is visually depicted. The probability at the detection juncture in the time domain is presented in **FIGURE 2C**, showcasing a detection accuracy ranging between 60% to 80%, barring the initial 'EY' word (**FIGURE 2C**). Subsequently, the decoded words are superimposed onto the waveform within the time domain (**FIGURE 2D**). And a maximum delay (t$_{\text{max,}}$ maximum delay time), 3.15 seconds was identified.

The inference outcomes of a case among PD for the speech sound model output are depicted in **FIGURE 3**. **FIGURE 3** illustrates the inference findings for instances identified as male, H&Y 4, UPDRS II-part 5 score of 3, and UPDRS III-part 18 score of 3 within the PD group. We visualized the corresponding class for each of 500 frames along the time axis of the waveform (**FIGURE 3A**). In **FIGURE 3B**, the trajectory for detecting words decoded from vocalized sounds is visually depicted. The probability at the detection juncture in the time domain

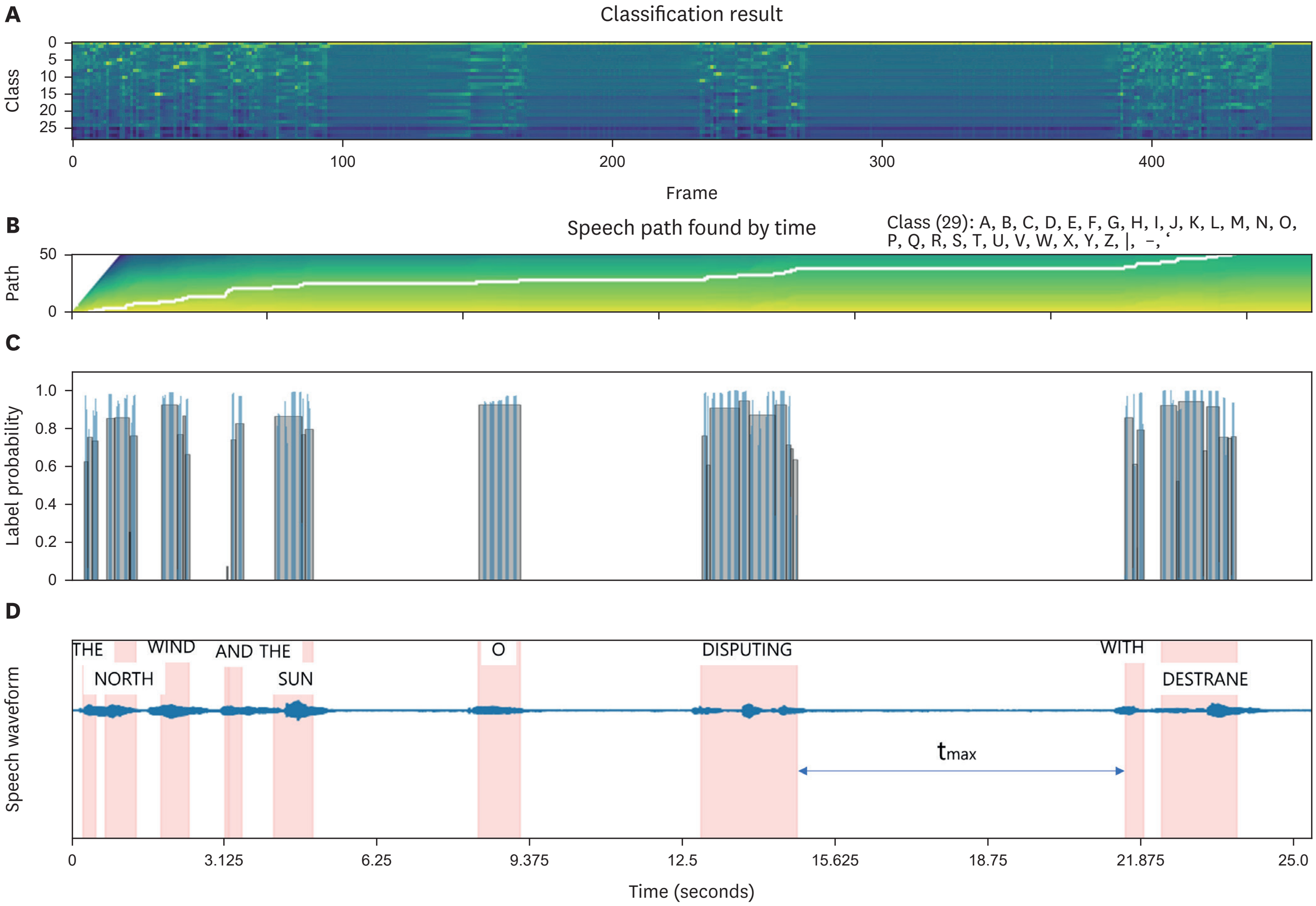


FIGURE 3. Visualization and analysis of inference results for speech sound model output in PD. (A) The inference findings for instances identified as male, H&Y 4, UPDRS II-part 5 score of 3, and UPDRS III-part 18 score of 3 within the PD group. The corresponding class for each of 500 frames along the time axis of the waveform. (B) The trajectory for detecting words decoded from vocalized sounds is visually depicted. (C) The probability at the detection juncture in the time domain. (D) The decoded words are superimposed onto the waveform within the time domain. In contrast to the scenario observed in the HC group, a maximum time delay, 5.45 (t$_{max}$) seconds was identified.
PD: Parkinson's disease, H&Y: Hoehn and Yahr, UPDRS: Unified Parkinson's Disease Rating Scale, HC: healthy control, t$_{max}$: maximum delay time.

TABLE 2. The speech-decoded text from a case of HC (H&Y stage 0) and PD (H&Y stage 4) and its corresponding accuracy

<table>
<tr><th>Sentences</th><th>Orthographic version</th><th>Cohort</th><th>Speech-decoded text of Case 8</th><th>Accuracy</th></tr>
<tr><td rowspan="2">Sentence 1</td><td rowspan="2">“The North Wind and the Sun were disputing which was the stronger, when a traveler came along wrapped in a warm cloak.”</td><td>HC</td><td>THE|NORTH|WIND|AND|THE|SUN|WERE|DISPUTING|WHICH|WAS|D|STRONGER|EH|WHEN|A|TRAVELER|CAME|ALONG|WRAPPED|IN|A|WARM|CLOAK|</td><td>0.90</td></tr>
<tr><td>PD</td><td>WE|THE|NORTH|WIND|AND|THE|SUN|O|DISPUTING|WITH|DISTRANGER|LOR|EA|TICKING|WHEN|TRABLA|CAME|ALONG|WITH|THO|RA|IN|TO|WON|CLO|</td><td>0.31</td></tr>
<tr><td rowspan="2">Sentence 2</td><td rowspan="2">“They agreed that the one who first succeeded in making the traveler take his cloak off should be considered stronger than the other.”</td><td>HC</td><td>THEY|AGREED|THAT|THE|ONE|WHO|
FIRST|SUCCEEDED|IN|MAKING|D|
TRAVELER|TAKE|HIS|CLOAK|OFF|SHOULD|BE|CONSIDERED|STRONGER|THAN|THE|OTHER|</td><td>0.95</td></tr>
<tr><td>PD</td><td>THE|AGREE|THA|THE|FIRST|E|MAKING|TRAMBLING|WE|CLOCK|WE|TESIDE|NI|C|TOYO|</td><td>0.10</td></tr>
</table>

Orthographic version: sentence from the “North Wind and the Sun.”[13)]
Accuracy = (Number of Words Matching the Orthographic Version Among Words Detected by Speech Recognition)/(Total Number of Words in the Orthographic Version Sentence).
HC: healthy control, PD: Parkinson’s disease, H&Y: Hoehn and Yahr.

TABLE 3. Total mean accuracy for the first and second sentence decoded both HC (n=10) and PD (n=10)

Sentences	HC	PD
Decoded sentence 1	0.94±0.03	0.66±0.18
Decoded sentence 2	0.94±0.02	0.64±0.23

Values are presented as mean ± standard deviation. Sentences: sentences from the “North Wind and the Sun” and “Computer Applications in Geography Snippet.”[13)]
HC: healthy control, PD: Parkinson’s disease.

is presented in **FIGURE 3C**, showcasing a detection accuracy ranging between 60% to 90% (**FIGURE 3C**). Subsequently, the decoded words are superimposed onto the waveform within the time domain (**FIGURE 3D**). Notably, in contrast to the scenario observed in the HC group, a maximum delay (t_{max}), 5.45 seconds was identified (**FIGURE 3C & D**).

Speech decoding to text and accuracy calculation

Speech decoding was performed on one female and one male participant from the HC and PD cohorts. The female participant had a H&Y score of 0, a UPDRS II-part 5 score of 0, and a UPDRS III-part 18 score of 0. The male participant had H&Y score of 4, UPDRS II-part 5 score of 3, and UPDRS III-part 18 score of 3. Accuracy analysis was subsequently conducted (**TABLE 2**). The first column represents the original text read aloud by the participant, while the second column displays a compilation of words extracted through waveform decoding of the recorded speech. Subsequently, the third column presents the calculated degree of agreement for the entire syllable based on this process. The accuracy for the 2 sentences was noted as 0.90 and 0.95 for a female in the HC, compared to 0.31 and 0.10 for a male in the PD group, respectively (**TABLE 2**).

Participants in the HC and PD cohort were a mix of those who read “North Wind and the Sun” and those who read “Computer Applications in Geography Snippet.”[16)] The accuracy results were calculated for the first and second decoded sentences, respectively (**TABLE 3**). The HC group showed the mean accuracy of 0.94±0.03 and 0.94±0.02 when reading both sentences, whereas the PD group exhibited accuracy levels of 0.66±0.18 and 0.64±0.23 for the respective sentences.

DISCUSSION

ASR has the potential to revolutionize how clinicians interact with patients suffering from speech disorders.[25,26,31,33)] Wave-to-vec, used in this study, employs deep learning models to directly learn informative representations from raw speech waveforms.[3)] This bypasses the need for domain-specific feature engineering, which can be ineffective for speech disorders due to the unpredictable nature of the variations.[3)] The core concept of wave-to-vec involves training a neural network on large amounts of speech data. The network architecture

https://doi.org/10.13004/kjnt.2024.20.e30

typically consists of convolutional layers that progressively extract local features from the waveform. These features are then passed through recurrent layers, allowing the network to capture the temporal dependencies within the speech signal. Finally, the network outputs a vector representation that encodes the essential characteristics of the speech segment. This learned wave-to-vec representation can then be used in conjunction with various ASR techniques. The Connectionist Temporal Classification decoder utilizes the wave-to-vec embedding to predict the most likely sequence of phonemes or words that corresponds to the input speech. The wave-to-vec method holds particular promise for speech recognition in patients with speech disorders. By learning directly from the speech waveform, the model can potentially capture the underlying speech information even when traditional acoustic features are unreliable. This paves the way for more robust and accurate ASR systems that can effectively recognize the speech of individuals with communication difficulties.

Meanwhile, Wav2Vec 2.0 model itself is language-agnostic. It processes raw audio waveforms to generate numerical representations. The pre-tokeninzer in the study is converting audio into English words or recognizing word classes in English. The applicability and precision of this approach may vary across languages. For instance, Korean and English exhibit substantial disparities in their linguistic structures. A primary distinction lies in the formation of phonemes: Korean operates at the syllable level, whereas English operates at the letter level. In Korean, syllables are segmented into initial, medial, and final consonants, with each syllable comprising a blend of consonants and vowels.[8)] Conversely, English relies on individual alphabet letters to construct words. In multi-language applications, a tailored pre-tokenization solution could be necessary. This involves training on multilingual datasets covering the intended languages and developing a pre-tokenizer capable of accommodating the unique phonetic elements and structural intricacies of each language.

Given the presence of speech impairments in PD, non-invasive speech analysis offers a valuable tool for facilitating the diagnosis and tracking disease progression.[29,32)] This methodology complements conventional assessments like the UPDRS, notably UPDRS II (Assessment of Mind and Activities of Daily Living) and UPDRS III (Motor Examination), by accentuating speech-related parameters. For instance, UPDRS includes items pertaining to voice characteristics, such as tone (UPDRS II 1.2).[29)] Various features, not only perceptual rating including intelligibility, perceptual score, fluency, and prosody but acoustic analysis including pitch, loudness, and jitter, can be extracted from speech samples utilizing techniques such as Mel-frequency cepstral coefficient or gammatone filter banks.[1,22)] Statistical scrutiny of these features may elucidate alterations in prosodic elements, such as diminished pitch modulation (monotonous speech) or reduced loudness (hypophonia), commonly observed in PD patients.[10)] Additionally, classification of speech samples according to PD likelihood can be accomplished through machine learning algorithms like SVM or Random Forest, trained on speech data correlated with predetermined UPDRS scores.[14)]

The functionality for continuous text reading (UPDRS III 18) permits the transcription of spoken text employing ASR systems such as the dataset used in this study. This facilitates the automated derivation of metrics like reading speed (words per minute), fluency (number of pauses and hesitations), and pronunciation errors. Comparative analysis of these metrics against normative values or established UPDRS scores in healthy cohorts may delineate potential language impairments associated with PD. Moreover, within the domain of text reading speed and latency (UPDRS III 18), pauses can be delineated and subjected to detailed analysis regarding duration and frequency. Furthermore, the algorithmic detection of hesitations and repetitions may signify bradykinesia (slowness of movement), which impinges upon speech production in individuals afflicted with PD. Examining the analytical outcomes presented in **FIGURES 2** & **3**, it was observed that the female participant classified as HC in **FIGURE 2** exhibited delays between each word ranging from 0.5 seconds to less than 1 second while reading a sentence. Conversely, the analysis of the male participant categorized as PD in **FIGURE 3** revealed delays ranging from 3.5 to 5 seconds. This acoustic analysis approach serves as an illustration of a metric capable of discerning the advancement of PD. Furthermore, it was ascertained that the accuracy, as analyzed in **TABLE 2**, exhibited a lower value in the PD group compared to a female in the HC group (**TABLE 2**, accuracy 0.31 and 0.10, respectively). PD can impact speech communication, though it might not manifest in classic pronunciation errors or reading/speaking mistakes.[10,12,29,32] Clinical observations and academic research point towards a decline in speech intelligibility for PD patients.[6] This doesn't necessarily mean they're mispronouncing words or making grammatical errors. Instead, the core issue lies in how PD affects the physical aspects of speech production. People with PD often exhibit softer voices, imprecise articulation of consonants, and reduced variation in pitch and volume.[7] These subtle changes can make it harder for listeners to understand the speech content, even if the words chosen and sentence structure are correct.[20] This can lead to communication breakdowns and frustrations for both the PD patient and the listener struggling to decipher the message. It's important to note that the severity of these speech changes can vary considerably depending on the individual and disease progression.

Although label probabilities were computed in **FIGURES 2C** & **3C**, further research is needed to improve the accuracy of word recognition in recorded speech.

While more comprehensive perceptual evaluation and acoustic analysis are warranted, our findings suggest the potential of this approach as a clinical decision support tool for diagnosing and evaluating speech disorders resulting from conditions such as PD, stroke, multiple sclerosis, and TBI.

There are few challenges in this field of study. Speech characteristics can vary due to factors like age, gender, and dialect. And normalization techniques or speaker-specific models might be needed for robust analysis. Finally, training machine learning models often requires large datasets of speech recordings with corresponding UPDRS scores. And the data collection efforts focused on PD patients are crucial. Speech analysis has to be integrated with other modalities like facial expression recognition for a more comprehensive assessment.

Pre-trained models can struggle with diverse languages due to limited training data.[9,18] However, this hurdle can be addressed through several techniques. Kim et al.[18] studied to overcome the problem of using Wav2Vec 2.0, an English-centric speech recognition model, for Korean speech analysis. The study includes a multi-task architecture for syllables and phonemes, and a joint decoder to handle out-of-vocabulary words. The approach utilizes both fine-tuning and cross-lingual pre-training on the target language data, achieving state-of-the-art performance on speech recognition tasks. Multilingual fine-tuning broadens the model's competency by retraining it on a dataset encompassing multiple languages.[9] Adapter modules offer a more efficient approach by adding lightweight extensions to the model for specific languages.[15]

Speech analysis tools that can be easily integrated into clinical settings could be developed.[11] Investigation of the potential of deep learning architectures for more complex feature extraction and disease classification can be conducted. Furthermore, exploration of the use of smartphone-based speech analysis applications for remote monitoring of PD patients can be performed.

CONCLUSION

This study investigates the feasibility of employing a pre-trained deep learning wave-to-vec model for speech-to-text analysis in individuals with language impairments. We evaluated the specific cases' outcomes by inferring text from the audio waveforms of HC and patients diagnosed with PD. Additionally, we visualized the model's detection probability within the time domain. Notably, male participants in the PD group with H&Y stage 4, UPDRS II-part 5 score of 3, and UPDRS III-part 18 score of 3 exhibited voice delays ranging from 3.5 seconds to exceeding 5 seconds. Furthermore, the reading accuracy of the particular male in PD was 0.31 and 0.10 for 2 sentences, respectively, which fell below the mean of 0.94 observed among the HC participants. While more comprehensive perceptual evaluation and acoustic analysis are warranted, our findings suggest the potential of this approach as a clinical decision support tool for diagnosing and evaluating speech disorders resulting from conditions such as PD, stroke, multiple sclerosis, and TBI.

REFERENCES

1. Adiga A, Magimai M, Seelamantula CS. Gammatone wavelet cepstral coefficients for robust speech recognition in Proceedings of the 2013 IEEE International Conference of IEEE Region 10 (TENCON 2013). New York, NY: IEEE, 2013
2. Arora SJ, Singh RP. Automatic speech recognition: a review. **Int J Comput Appl 60:**34-44, 2012 CROSSREF
3. Baevski A, Zhou Y, Mohamed A, Auli M. Wav2vec 2.**0:** a framework for self-supervised learning of speech representations in Proceedings of the 34th International Conference on Neural Information Processing System. Red Hook, NY: Curran Associates Inc., pp12449-12460, 2020
4. Barnett C, Armes J, Smith C. Speech, language and swallowing impairments in functional neurological disorder: a scoping review. **Int J Lang Commun Disord 54:**309-320, 2019 PUBMED | CROSSREF
5. Belić M, Bobić V, Badža M, Šolaja N, Đurić-Jovičić M, Kostić VS. Artificial intelligence for assisting diagnostics and assessment of Parkinson's disease-a review. **Clin Neurol Neurosurg 184:**105442, 2019 PUBMED | CROSSREF
6. Carvalho J, Cardoso R, Guimarães I, Ferreira JJ. Speech intelligibility of Parkinson's disease patients evaluated by different groups of healthcare professionals and naïve listeners. **Logoped Phoniatr Vocol 46:**141-147, 2021 PUBMED | CROSSREF
7. Chiu YF, Neel A. Predicting intelligibility deficits in Parkinson's disease with perceptual speech ratings. **J Speech Lang Hear Res 63:**433-443, 2020 PUBMED | CROSSREF
8. Choi JM, Kim JD, Park CY, Kim YS. Automatic word spacing of Korean using syllable and morpheme. **Applied Sciences 11:**626, 2021 CROSSREF
9. Conneau A, Khandelwal K, Goyal N, Chaudhary V, Wenzek G, Guzmán F, et al. Unsupervised cross-lingual representation learning at scale. **arXiv** April 8, 2020. https://doi.org/10.48550/arXiv.1911.02116 CROSSREF
10. Critchley EM. Speech disorders of Parkinsonism: a review. **J Neurol Neurosurg Psychiatry 44:**751-758, 1981 PUBMED | CROSSREF
11. Dimauro G, Caivano D, Bevilacqua V, Girardi F, Napoletano V. VoxTester, software for digital evaluation of speech changes in Parkinson disease in Proceedings of the 2016 IEEE International Symposium on Medical Measurements and Applications (MeMeA). New York, NY: IEEE, 2016
12. Movement Disorder Society Task Force on Rating Scales for Parkinson's Disease. The Unified Parkinson's Disease Rating Scale (UPDRS): status and recommendations. **Mov Disord 18:**738-750, 2003 PUBMED | CROSSREF
13. Duffy JR. Functional speech disorders: clinical manifestations, diagnosis, and management. **Handb Clin Neurol 139:**379-388, 2016 PUBMED | CROSSREF

14. Govindu A, Palwe S. Early detection of Parkinson's disease using machine learning. **Procedia Comput Sci 218:**249-261, 2023 CROSSREF

15. Guo J, Zhang Z, Xu L, Chen B, Chen E. Adaptive adapters: an efficient way to incorporate BERT into neural machine translation. **IEEE-ACM T Audio Spe 29:**1740-1751, 2021 CROSSREF

16. Jaeger H, Trivedi D, Stadtschnitzer M. Mobile Device Voice Recordings at King's College London (MDVR-KCL) from both early and advanced Parkinson's disease patients and healthy controls. **Zenodo** May 17, 2019. https://doi.org/10.5281/zenodo.2867216 CROSSREF

17. Khan T, Nyholm D, Westin J, Dougherty M. A computer vision framework for finger-tapping evaluation in parkinson's disease. **Artif Intell Med 60:**27-40, 2014 PUBMED | CROSSREF

18. Kim J, Kang P. K-wav2vec 2.**0:** automatic speech recognition based on joint decoding of graphemes and syllables. **arXiv** October 11, 2021. https://doi.org/10.48550/arXiv.2110.05172 CROSSREF

19. Lamba R, Gulati T, Jain A, Rani P. A speech-based hybrid decision support system for early detection of Parkinson's disease. **Arab J Sci Eng 48:**2247-2260, 2023 CROSSREF

20. Levy ES, Moya-Galé G, Chang YHM, Freeman K, Forrest K, Brin MF, et al. The effects of intensive speech treatment on intelligibility in Parkinson's disease: a randomised controlled trial. **EClinicalMedicine 24:**100429, 2020 PUBMED | CROSSREF

21. Lillian A, Zuo W, Laham L, Hilfiker S, Ye JH. Pathophysiology and neuroimmune interactions underlying Parkinson's disease and traumatic brain injury. **Int J Mol Sci 24:**7186, 2023 PUBMED | CROSSREF

22. Naing HM, Miyanaga Y, Hidayat R, Winduratna B. Filterbank analysis of MFCC feature extraction in robust children speech recognition in Proceedings of the 2019 International Symposium on Multimedia and Communication Technology (ISMAC). New York, NY: IEEE, 2019

23. Quan C, Ren K, Luo Z. A deep learning based method for Parkinson's disease detection using dynamic features of speech. **IEEE Access 9:**10239-10252, 2021 CROSSREF

24. Ramaker C, Marinus J, Stiggelbout AM, Van Hilten BJ. Systematic evaluation of rating scales for impairment and disability in Parkinson's disease. **Mov Disord 17:**867-876, 2002 PUBMED | CROSSREF

25. Sadeghian R, Schaffer JD, Zahorian SA. Towards an automatic speech-based diagnostic test for Alzheimer's disease. **Front Comput Sci 3:**624594, 2021 CROSSREF

26. Schultz BG, Tarigoppula VS, Noffs G, Rojas S, van der Walt A, Grayden DB, et al. Automatic speech recognition in neurodegenerative disease. **Int J Speech Technol 24:**771-779, 2021 CROSSREF

27. Shriberg LD, Austin D, Lewis BA, McSweeny JL, Wilson DL. The speech disorders classification system (SDCS): extensions and lifespan reference data. **J Speech Lang Hear Res 40:**723-740, 1997 PUBMED | CROSSREF

28. Shulman LM, Gruber-Baldini AL, Anderson KE, Fishman PS, Reich SG, Weiner WJ. The clinically important difference on the unified Parkinson's disease rating scale. **Arch Neurol 67:**64-70, 2010 PUBMED | CROSSREF

29. Skodda S, Grönheit W, Mancinelli N, Schlegel U. Progression of voice and speech impairment in the course of Parkinson's disease: a longitudinal study. **Parkinsons Dis 2013:**389195, 2013 PUBMED | CROSSREF

30. Smith M, Ramig L. Neurological disorders and the voice. **NCVS Status Prog Rep 7:**207-227, 1994

31. Tóth L, Hoffmann I, Gosztolya G, Vincze V, Szatlóczki G, Bánréti Z, et al. A speech recognition-based solution for the automatic detection of mild cognitive impairment from spontaneous speech. **Curr Alzheimer Res 15:**130-138, 2018 PUBMED | CROSSREF

32. Tsanas A, Little M, McSharry P, Ramig L. Accurate telemonitoring of Parkinson's disease progression by noninvasive speech tests. **IEEE Trans Biomed Eng 57:**884-893, 2010 PUBMED | CROSSREF

33. Zhou L, Fraser KC, Rudzicz F. Speech Recognition in Alzheimer's Disease and in Its Assessment in Proceedings of the Interspeech 2016. [place unknown]: ISCA; 2016