This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2022.3158840, IEEE Journal of Biomedical and Health Informatics



# Continuous TBI monitoring from Spontaneous Speech using Parametrized Sinc Filters and a Cascading GRU

Apiwat Ditthapron, Adam C. Lammert, and Emmanuel O. Agu

***Abstract*— Traumatic Brain Injury (TBI) is caused by a head injury that affects the brain, impairing cognitive and communication function and resulting in speech and language disorders. Over 80,000 individuals in the US suffer from long-term TBI disabilities and continuous monitoring after TBI is essential to facilitate rehabilitation and prevent regression. Prior work has demonstrated the feasibility of TBI monitoring from speech by leveraging advancements in Artificial Intelligence (AI) and speech processing technology. However, much of prior work explored TBI detection using audio captured using a mobile device while subjects performed scripted speech tasks such as diadochokinesis tests or read a passage. Such scripted approaches require active user involvement that significantly burdens participants. Moreover, they are episodic and do not provide a longitudinal picture of the user's TBI condition, which is useful in monitoring recovery trajectory. This study proposes a continuous TBI monitoring from changes in acoustic features of spontaneous speech collected passively using the smartphone. Low-level acoustic features are extracted using parametrized Sinc filters (pSinc) that are then classified TBI (yes/no) using a cascading Gated Recurrent Unit (cGRU). The cGRU model utilizes a cell gate unit in the GRU to store and incorporate each individual's prediction history as prior knowledge into the model. In rigorous evaluation, our proposed method outperformed prior TBI detection methods on a dataset containing conversational speech recorded during patient-therapist discourses following TBI, achieving 83.87% balanced TBI classification accuracy. Furthermore, unique words that are important in TBI prediction were identified using SHapley Additive exPlanations (SHAP). A correlation was also found between features acquired by the proposed method and coordination deficits following TBI.**

***Index Terms*— Traumatic brain injury, Continuous monitoring, Acoustic features, Deep learning, Smartphone**

## I. Introduction

This material is based on research sponsored by DARPA under agreement number FA8750-18-2-0077. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA or the U.S. Government.

Apiwat Ditthapron and Emmanuel O. Agu are with the Computer Science Department, Worcester Polytechnic Institute, Worcester, MA 01609 USA (e-mail: aditthapron@wpi.edu, emmanuel@wpi.edu).

Adam C. Lammert is with the Biomedical Engineering Department, Worcester Polytechnic Institute, Worcester, MA 01609 USA (e-mail: alammert@wpi.edu).

Corresponding author: Emmanuel O. Agu.

TRAUMATIC Brain Injury (TBI) is a complex neurological condition stemming from a physical insult or injury to the head, which causes a wide range of physiological, neurological, psychological, and behavioral issues. In the United States, more than 1.4 million people are diagnosed with TBI each year, with 80,000 of them developing permanent disabilities [1]. While most TBI patients receive immediate treatment and are discharged from the hospital, many patients experience lingering issues that require clinician follow-up or monitoring of symptoms in order to steer treatment and reduce the risks of rehospitalization and premature death. These adverse events are often caused by infectious, neurological, and neurosurgical disorders, which are particularly significant in moderate-to-severe cases [2]. Individuals with *concussion*, or mild TBI (*mTBI*, can have long-term TBI-associated sequelae, such as cognitive and communication impairment that disrupts normal life activities [2]–[4]. The large and increasing number of TBI cases is a source of concern in public health. There is urgent need for automated monitoring of TBI symptoms and ailment trajectories to prevent adverse events and fatalities, and assist patients in the recovery from long-term TBI sequelae.

Speech and language disorders are common communication impairments following TBI and are often used as TBI biomarkers [4]. Speech contains rich paralinguistic (e.g., affective and health-related) information, in addition to its linguistic content [5], making it an effective biomarker for detecting neurological disorders including depression, bipolar disorder, and TBI [6]–[13]. Gathering speech samples for analyses is non-invasive and can be performed passively outside the clinic using ubiquitous sensing devices such as smartphones. Moreover, passive recording of speech on smartphones significantly reduces participant burden, increasing participation rates and provides access to the 85 percent of Americans who own a smartphone [14], [15]. Previous work proposed TBI detection systems based on speech captured using a mobile tablet with TBI classification performed on a server [11]–[13]. However, these previous work required active user involvement and collected speech while users were performing assigned speech assessment tasks, such as diadochokinetic rate and passage reading, or elicitation, such as describing a picture. Approaches that require active user involvement present more burden than the completely passive approach we explore and are episodic and do not monitor participants' TBI condition continuously, which is important in recovery monitoring. In

2168-2194 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Worcester Polytechnic Institute. Downloaded on April 08,2022 at 15:12:57 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2022.3158840, IEEE Journal of Biomedical and Health Informatics



this study, we propose a TBI detection model (binary classification) that continuously evaluates TBI status passively from spontaneous speech collected on the smartphone. Our TBI speech analyses pipeline is illustrated in Figure 1.

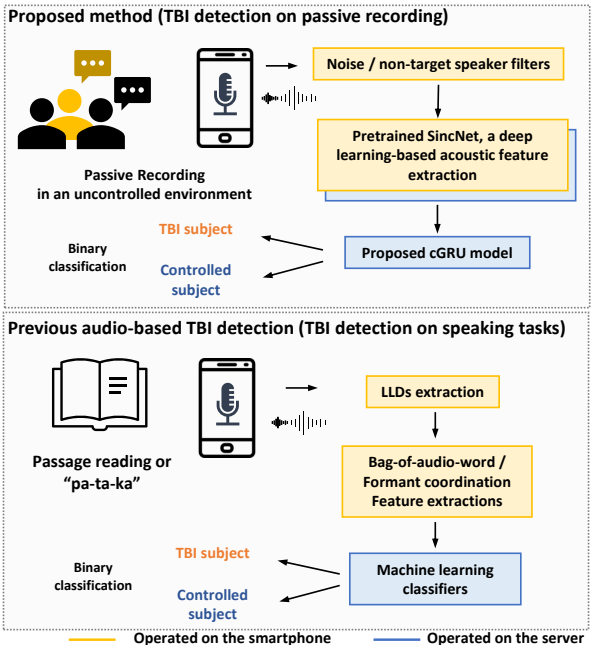



Fig. 1: Pipeline in the proposed continuous TBI monitoring and conventional TBI detection

To passively and continuously monitor TBI using smartphones, two primary challenges are: mitigating interfering noise and assessing speech from continuous recordings. Noise reduction is a critical issue when processing speech captured in an uncontrolled environment that has background noise or cross-talk. Noise mitigation for speech assessment were previously addressed in [12], [16]. This study addresses the issue of assessing subjects from continuous speech, which is an important component to monitor TBI recovery passively. In contrast with previous works [8]–[13] that evaluate speech within a discrete interval of interest (e.g. before and after high TBI risk activity), our proposed method can assess TBI risk continuously at various monitoring intervals (e.g., hourly, daily or weekly) throughout the day without user engagement. Our system facilitates taking into account and fully integrating temporal relationships between assessments. Inspired by prior work [8], [13] that utilize the time-delayed feature to capture changes in the strengths of formant coupling to detect TBI detection, we propose a cascading Gated Recurrent Unit (cGRU) that continuously classifies the smartphone user's current TBI state from continuous speech using speaker-dependent prior knowledge stored within the *cell unit* of GRU. cGRU analyzes a cascaded representation of the speech signal at different scales, which is done using a sliding window and skipped layer connections, allowing data at different scales to pass through the GRU's cell unit. Another advantage of using the cGRU to process time-series data is that it eliminates the need for additional data storage, as only the cell unit is retained and overridden in each time frame.

The cGRU performs on acoustic speech features, in which a parametrized Sinc filter (pSinc) feature, Convolutional Neural Network (CNN) feature, Low-Level-Descriptors (LLD) of speech, and formant frequencies are considered. In rigorous evaluation, we found that the pSinc feature yielded the best TBI detection performance and enabled model interpretation in the frequency domain. pSinc and CNN features have previously been proposed for the speaker recognition [17] and Automatic Speech Recognition (ASR) tasks [18], respectively. Conceptually, acoustic features extracted from pSinc are similar to those extracted from a band-pass filter. They have an advantage over CNN feature in that they learn to capture complex spectral features that benefit TBI detection. Our architectures of pSinc and CNN differ from the previous works in that we incorporate a skip layer connection, which yielded significantly improved TBI detection performance. Additionally, we improved the efficiency of TBI detection by employing Vocal-Tract-Length-Normalization (VTLN) [19] to suppress inter-speaker acoustic variability between subjects in Coelho's corpus [20]. VTLN reduces speaker to speaker variability, which is an important factor in speaker recognition but reduces the performance of speaker independent tasks such as automatic speech recognition and speech assessment.

We compared cGRU to traditional GRU, Long short-term memory (LSTM), Support Vector Machine (SVM), Random Forest (RF), and Multi-Layer Perceptrons (MLP) on spontaneous speech collected during speech and language discourse following TBI (i.e. Coelho's corpus [20]). Additionally, pSinc features are compared to three types of features previously proposed for passive and continuous assessment features, for the TBI detection task, namely Low-Level Descriptors (LLD), Bag-Of-Audio-Word (BOAW) and formant coordinations. To further interpret and analyze learned features, we used SHapley Additive exPlanations (SHAP) method [21] to estimate the effect of each input time-slice (or so-called attribution), on the final TBI prediction. cGRU with pSinc feature was found to have a correlation with formant coordination features. Certain spoken words were also found to be more prominent among TBI subjects vs controls based on the attribution scores in the temporal and spectrotemporal domains.

The contribution of this work can be summarized as follows.

1) We propose and evaluate a cGRU for continuous TBI assessment, which successfully detects and classifies TBI based on acoustic features.
2) We demonstrate the benefits of TBI classification using pSinc features with skipped layer connections and VTLN over traditional CNN and LLD.
3) We analyze the features learned within the proposed cGRU to determine its association with conventional acoustic features.
4) We extend the SHAP method to analyze input speech in the temporal and spectral domains to find characteristics of speech that are prevalent following TBI.

The rest of this paper is organized as follows. Previous works in TBI detection and acoustic features are presented in Section II. Then, the cascading GRU and experiment setup are described in IV. The evaluation results of the proposed

2168-2194 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Worcester Polytechnic Institute. Downloaded on April 08,2022 at 15:12:57 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2022.3158840, IEEE Journal of Biomedical and Health Informatics



method are reported in Section V, followed by a discussion and application of this study in Section VI. Finally, we conclude our study in Section VII.

## II. RELATED WORK

### A. Speech and language disorders following TBI

Speech and language disorders are commonly identified as communication impairments that impair an individual's ability to function normally after an injury. In severe TBI cases with cranial nerves damaged, flaccid paralysis of the muscles supplied by cranial nerves V (trigeminal), VII (facial), X (vagus), or XII (hypoglossal) may present permanently, resulting in various types of motor speech disorders [22]. Dysarthria, weakness or incoordination of speech musculature, was indicated as a common alignment of TBI, with a prevalence rate ranging from 8% to 100% depending on incident severity and on-set time after the injury [22]. In a case that has a cerebral lesion, apraxia of speech, another common motor speech disorder associated with TBI that affects motor planning, sequencing, control, and timing, may be observed [23]. Common symptoms of apraxia are increased speech initiation difficulty and prosodic disturbance. Dysarthria and apraxia of speech are often diagnosed in cohort studies of TBI [3].

Apart from the deficits in speech production, individuals with TBI possibly manifests a degree of difficulty in formulating and comprehending speech, which are results of linguistic and cognitive impairments [24]. Previously, the severity of language disorders was determined by the nomenclature of aphasia, which occurs when the Broca or Wernicke areas of the brain are damaged. According to Normal *et al.*, aphasia is the most prevalent communication disorder among veterans with a history of TBI, with the majority occurring in moderate and mild TBI [3]. Individuals with TBI exhibit a wide range of speech and language deficits, many of which are assumed to be sub-clinical but nonetheless are present (e.g., [8]) and can be used to detect TBI non-invasively and passively.

### B. Previous acoustic-based TBI detection

Speech acoustics have recently gained more attention in research into TBI assessment technologies because they are non-invasive and can be collected passively on the smartphone. In traditional TBI assessment, individuals suspected of having TBI are often subjected to a diadochokinesis speech rate test to determine their syllable alternating motion rate, which was found to associate with TBI [25], through repetition of the trisyllabic sequences *pa-ta-ka* [11]–[13]. Poellabauer *et al.* collected speech from short sentences and trisyllabic sequences in order to determine motor speech disorder following TBI [11]. The recording includes seven measures, the first three of which assess prosody, stress, and standard syllabic rate, while the last four assess motion rate and sustained vowels. The authors used existing speech and phonetic recognition tools to locate word and vowel boundaries before extracting spoken word speed, stressed word duration, fundamental frequency, intensity, and diadochokinetic speech rate as features for a linear classification model with a classification result greater than 70%. Similarly, Daudet *et al.* [12] examined spectral-domain and temporal-domain features using only the first six measures for TBI detection. Using Logistic Regression, [12] was able to achieve Area Under receiver operating characteristic Curve (AUC) of 0.86 by combining spectral and temporal features.

Among the spectral features, formant frequency have received much attention in automatic TBI detection due to their well-understood relationship with movement of the speech articulators [8], [10], [13]. Formant frequency indicates a particular frequency where acoustic energy is concentrated and corresponds to the resonances of the vocal tract. Helfer *et al.* considered auto- and cross-correlations of first three formants' (i.e., F1, F2, and F3) velocities and accelerations over short periods to estimate ImPACT scores, an FDA-cleared concussion assessment, of 32 high school athletes using a Grandfather passage [8]. The classification was performed using a Support Vector Machine (SVM) with Principal Component Analysis (PCA) as a pre-processing step and archived an AUC of 0.95. Similarly, Talkar *et al.* considered the same formant correlation features set with gate features, to detect mTBI using Gaussian Mixture Model (GMM) and a Convolutional Neural Network (CNN) [13]. Additionally, they considered free speech, read speech, and diadochokinetic and indicated that read speech with CNN outperforms other setups with an AUC of 0.90, and combining all three tasks improved the AUC to 0.96.

In another study, Falcone *et al.* collected the pronunciation of ten digits from 105 male athletes before and following a boxing tournament using a directional microphone [10]. The speech recorded prior to the tournament was used to train a one-class SVM classifier which was then evaluated using speech recorded after the tournament. Acoustic features used to train the classifier included formant frequency, jitter, shimmer, harmonic to noise, and pitch frequency. Formant frequency was one of the top three most significant features across digit words, with a maximum F1 score of 87.5.

Previous work has established that acoustic features of speech are useful for detecting TBI, as evidenced by the high detection accuracies in Table I. However, TBI assessment that fully integrates temporal information available in acoustic features into a system for passive and continuous monitoring has not been demonstrated in the literature.

### C. Acoustic features

Speech is typically represented as a waveform that represents changes in sound pressure over time. Waveforms are recorded at a high sampling rate (typically $>$ 16kHz), which immediately introduces the *curse of dimensionality* problem that causes training issues in machine learning. Consequently, the previous TBI detections discussed in Section II-B do not detect TBI directly on the waveform but its derived acoustic features, such as Low-Level Descriptor (LLD) and Formant coordination. We also consider Bag-Of-Audio-Word (BOAW) feature as it demonstrated ground-breaking success in language disorder assessment [26].

*1) Low-Level Descriptor:* Low-Level Descriptor (LLD) is a set of acoustic features that characterizes human speech over a short duration. There are several descriptors, but they are typically classified into prosodic, spectral, cepstral, and voice

2168-2194 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Worcester Polytechnic Institute. Downloaded on April 08,2022 at 15:12:57 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2022.3158840, IEEE Journal of Biomedical and Health Informatics



TABLE I: Previous speech-based TBI detection studies

| Method | Recording task | Feature | Classifier | TBI subject ratio | Performance |
|---|---|---|---|---|---|
| Poellabauer *et al.* and Daudet *et al.* [11], [12] | Short sentence, pa-ta-ka | Prosody, stress, motion rate, sound sustaining | LR | 16% (n=581) | AUC: 0.86 |
| Falcone et al. [10] | Digit words | Formants, jitter, shimmer, HNR, pitch | SVM | 7% (n=105) | F1: 87.51% |
| Helfer *et al.* [8] | Passage reading | Formant auto- and cross-correlations | SVM | 29% (n=32) | AUC: 0.95 |
| Talkar *et al.* [13] | Passage reading<br>Free Speech<br>Pa-ta-ka<br>Previous three combined | Formant auto- and cross-correlations | CNN | 55% (n=21) | AUC:0.90<br>AUC:0.89<br>AUC:0.89<br>AUC:0.96 |

It should be noted that the evaluation metrics in the performance column are not directly comparable since the studies examined different populations.

quality features. Prosodic features, such as F0, energy, zero-crossing, and loudness, are used to characterize an individual speaker's style, gender, dialect, and phonological factors. As an alternative to the time-domain waveform, audio may be represented in the frequency domain, which can be represented as descriptors that contain information about the energy, centroid, harmonicity, skewness, and kurtosis of each spectral band. Voice quality features express the variety of speech created by the larynx, such as formant, jitter, and shimmer. Determining a collection of LLDs features is application-dependent and typically requires clinical expert knowledge. We include LLD as a baseline for acoustic feature as it has been used to represent speech in sensor modalities to monitor bipolar disorder [27] passively on the smartphone.

*2) Formant coordination:* To capture changes in articulatory coordination following TBI, a time-delayed auto- and cross-correlation of formant frequency was developed [8], [13]. Cross-correlation is computed between the first three formants, and auto-correlation is applied on multi-scales time-delay formants to construct a covariance matrix, which is used as a feature to represent a short duration of speech. Although formant coordination features have not yet applied to passive recordings, [13] demonstrated that formant coordination works effectively on conversational speech.

*3) Bag-Of-Audio-Word (BOAW):* BOAW is a dimension reduction method for LLD using the Bag-Of-Word (BOW) algorithm. In order to reduce the number of parameters in LLDs, some LLDs are selected as a reference point, and the distance between each sample to the reference points is used as an LLD representation. BOAW has been used in aphasia detection [26], which is one of the co-morbidities associated with TBI, but it has not yet been validated for TBI detection.

## III. BACKGROUND

### A. Parametrized Sinc filter: a DNN-based acoustic feature extraction

Prior work has demonstrated that DNNs can be used to extract acoustic features from raw audio (waveform) leading to output superior to that of conventional acoustic features in speaker recognition and Automatic Speech Recognition (ASR) tasks [17], [18], [28]. This study examines how well the parametrized Sinc filter (pSinc) learns acoustic features for TBI detection. Each pSinc filters out frequency components outside a trainable frequency range ($f_{c_1}$,$f_{c_2}$), which is determined by a subtraction of two sinc filters, which are described mathematically in Equation 1. Multiple pSinc filters are used in the first layer of the TBI detection model in order to learn a set of frequency ranges that are relevant to TBI speech, (e.g., formant frequency.)

$$pSinc[n, f_{c_1}, f_{c_2}] = [2f_{c_2} sinc(2f_{c_2} n) - 2f_{c_1} sinc(2f_{c_1} n)] \quad (1)$$

$$sinc(x) = \frac{\sin(x)}{x} \quad (2)$$

$$s_i[n] = x[n] * pSinc[n, f_{c_1}, f_{c_2}] \quad (3)$$

We used pSinc filter to extract spectral features of speech ($s_i$) from audio ($x$) in the $n^{th}$ sliding window, followed by CNNs that extract a higher level of representation of speech. Using a learning algorithm such as stochastic gradient descent, $f_{c_1}$ and $f_{c_2}$ in a band-pass filter (Equation 1) were trained as a Finite Impulse Response (FIR) convolution filter on $x[n]$ to allow only signals within frequency $f_{c_1}$ and $f_{c_2}$ to pass through to subsequent CNN layers.

The primary benefit of using pSinc over CNN in the first layer is that it forces the network to learn spectral features of speech using only two parameters ($f_{c_1}$ and $f_{c_2}$) regardless of the kernel size, while the number of parameters in CNN depends on the kernel size. Additionally, pSinc enables interpretability in the frequency domain, allowing in-depth analysis of acoustic features learned by the model. In terms of performance, [17] demonstrated that pSinc outperforms CNN-based architecture, such as VGGish [28], in speaker recognition and phone recognition tasks with a higher degree of network interpretability in the frequency domain. In speech assessment, pSinc outperformed the LLDs in assessments of neurodegenerative disorders and cognitive impairment based on speech [29].

### B. Gated Recurrent Unit (GRU)

A recurrent Neural Network (RNN) is a type of DNN that specializes in capturing and learning patterns from sequential data in the temporal domain. As such, the RNN has frequently been employed to learn long-term temporal dependencies of speech features [30], [31]) However, when processing long sequential data, the vanilla RNN often encounters vanishing gradients problem, caused by updating small gradients in internal loop over a long sequential input [32]. We utilize the Gated Recurrent Unit (GRU), a variant of the RNN that solves the vanishing gradient problem and and has fewer parameters than the Long Short-Term Memory (LSTM) [32], to process sequential acoustic features. GRU enhances memory utility

2168-2194 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Worcester Polytechnic Institute. Downloaded on April 08,2022 at 15:12:57 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2022.3158840, IEEE Journal of Biomedical and Health Informatics



in the Long Short-Term Memory (LSTM) by combining the forget gate and state unit into a single gating unit $h_i^t$ (for time step $t$ and cell $i$) as illustrated in Figure 2 with $y_i^t = h_i^t = u_i^{t-1}h_i^{t-1} + (1-u_i^{t-1})tanh\big(b_i + \sum_j U_{i,j}x_j^t + W_{i,j}r_j^{t-1}h_j^{t-1}\big)$ The update gate $u_i^t$ and reset gate $r_i^t$ are computed from $u_i^t = \sigma\big(b_i^u + \sum_j U_{i,j}^u x_j^t + W_{i,j}^u h_j^t\big)$ and $r_i^t = \sigma\big(b_i^r + \sum_j U_{i,j}^r x_j^t + W_{i,j}^r h_j^t\big)$ respectively. The term $W, U$, and $b$ denote recurrent weight, input weight, and bias, respectively. Each gate contains a sigmoid function ($\sigma$), which is a dedicated mechanism to learn when to *update*, *reset*, hidden state $h^t$, or when to skip the irrelevant input $x^t$. Gating mechanisms can improve learning performance in TBI assessment, as not all acoustic features are prominent in TBI speech.

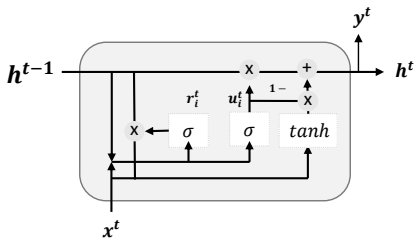



Fig. 2: Connections within a GRU

### C. Vocal Tract Length Normalization

Vocal Tract Length Normalization (VTLN) has been widely used to reduce interspeaker acoustic variability in modern ASR systems by warping the frequency axis associated with cepstral features. Rather than determining the exact vocal tract length of each speaker, [19] proposed an expectation–maximization algorithm for estimating a vocal tract length-related warping factor ($\alpha$) that produces the lowest error in ASR. Specifically, $\alpha$ was optimized by $\arg\max_{\alpha} Pr(X_i^{\alpha} \mid \lambda, W_i)$ to increase phoneme classification performance ($W$) using a Gaussian Mixture Model (GMM) and Hidden Markov Model (HMM), denoted as $\lambda$, on cepstrum features ($X$) of each speaker $i$ in the coelho corpus [19].

## IV. Methodology

### A. Coelho corpus

The proposed cGRU was evaluated on the Coelho corpus [20], which contains conversational speech collected during discourses following TBI. The discourse included story retelling, story generation, and conversation collected from 55 native English speakers with non-penetrating head injuries and 52 native English speakers with no brain injury (control subjects). Here, we only used speech gathered during conversational discourse and analyzed topic initiation, topic maintenance, and response appropriateness of the participant. Proficiency in these three skills are fundamental to fluent communication, which individuals with TBI experience some levels of difficulty.

*Study protocol:* Each subject participated in a 10-15 minutes conversation discourse examined by a speech-language pathologist. The first conversation was initiated by the examiner with the question "Why are you here at the hospital/rehabilitation centre today?", and was then shifted by either a subject or the examiner. Speaking turns and utterances were counted to analyze topic maintenance. Response appropriateness was measured in terms of responses and comments to the previous conversation. The study statistically indicated that TBI subjects produced more utterances in their response to the examiner, produced less comments, and required more story shifting from the examiner [20].

*Deep learning dataset creation:* All 55 subjects with TBI met the criteria of having recovered a high level of functional language – they had achieved the ability to converse fluently and did not demonstrate appreciable deficits on traditional clinical language tests. The dataset includes a diverse demographic range in gender (Male:39, Female:16), age (28.53±12.31), education level (13.01±2.38 years), working class (Unskilled:19, Skilled:18, Professional:18), TBI severity (Coma duration: 16.95±22.10 days) and recording on-set time after the accident (10.35±17.78 months). The causes of brain injury for subjects in this corpus are motor vehicle accident (45 subjects), fall (6 subjects), struck by car (3 subjects), and other (1 subjects).

### B. Proposed cascading GRU detection model

As demonstrated in [8], [13], changes in the complexity of correlation between acoustic features is associated with speech production after TBI. These previous studies employed autocorrelation to capture time-delayed linear correlation in time series of formant frequencies [8], [13]. Motivated by prior work, we incorporated the GRU, which is capable of capturing linear and non-linear correlations of acoustic features, and pSinc to monitor TBI. The specific architecture we utilized is called cascading GRU (cGRU) as it passes on individual hidden states throughout the continuous TBI monitoring. The cGRU is illustrated in Figure 3.

**Pre-trained pSinc:** Our proposed cascading GRU model is composed of three pre-trained layers from the SincNet model, which includes 64 pSinc with a kernel size of 251 and Leaky ReLu activation in the first layer. Following the pSinc, two layers of CNNs with 32 filters, kernel size of 5 and max pooling with a kernel size of 3 are included in the second and third layers. These configurations were fine-tuned from the list shown in Table II using grid search. The model was pre-trained using the Librispeech corpus [33] with tuned hyperparameters from [17]. pSinc features were extracted on a short duration of audio using a sliding window of 200 ms with 25 ms intervals, which was fine-tuned as explained in Section IV-E. Then, features were accumulated over 20 sliding windows as an input to the cGRU for TBI detection. This step is crucial, as audio within a single sliding window is too short to extract high-level features of TBI.

Prior to cascading features through the temporal domain, two skip layer connections are applied to aggregate acoustic features learned by pSinc and CNN at various stages. Through experimentation, we discovered that skip layer connections enhance TBI detection efficiency and avoid the vanishing gradient problem.

2168-2194 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Worcester Polytechnic Institute. Downloaded on April 08,2022 at 15:12:57 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2022.3158840, IEEE Journal of Biomedical and Health Informatics



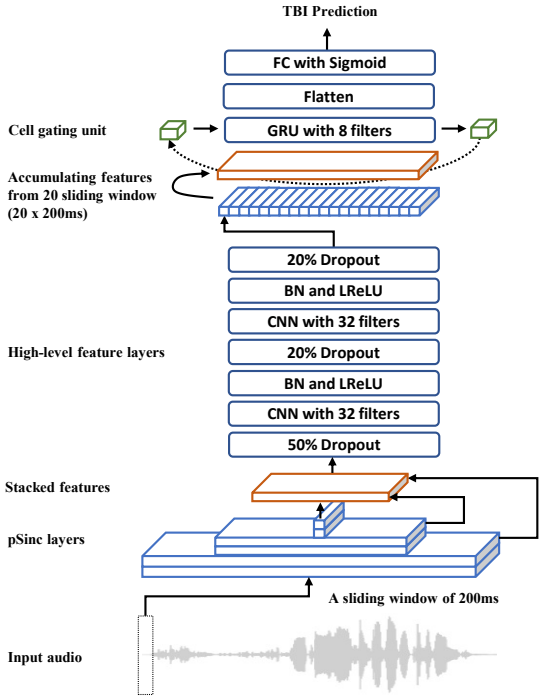



Fig. 3: Architecture of the cascading GRU model for TBI detection

**Domain adaptation layer:** The aggregating layer is preceded by a dropout layer to mitigate overfitting, as we discovered that regularization is critical for pSinc layers even though transfer learning was used. Then, two CNN layers consisting of 32 filters with BN, LReLU, and 20% dropout are included to extract high-level audio features for TBI detection. This latent feature at this step is used as an input ($x_t$) to the GRU.

**TBI classification using GRU:** The parameters in the cell gating unit ($h^{t-1}$) were randomly assigned before training for each subject using Glorot initialization. Any subsequent hidden state from the same subject will adopt $h^t$ from the previous hidden state and replace it with the current $h^{t+1}$. To perform TBI classification, the continuous outputs $y^t$ from 20 sliding windows (4 seconds each) are combined or flattened into a single dimension using a fully connected layer followed by a Sigmoid activation function with a threshold set to 0.5 for positive TBI speech. We tuned the proposed model's hyperparameters as described in Table II.

The proposed cGRU detects TBI over each 4-second window of speech using a hidden state stored within GRU's cell gating unit, leveraging temporal information across multiple recordings. The cell gating unit, which attempts to solve the data storage problem while retaining high TBI prediction accuracy, is the only additional parameter to store when tracking TBI over time.

### C. Data preprocessing and implementation of cascading GRU model

The speech recording includes a dialogue between an interviewer and a participant, from which we extracted only participant speech using the Coelho corpus's transcript with time boundaries. Due to poor sound quality, three TBI and one control subject were omitted from the experiments. The audio signals were downsampled to 16000 Hz and then normalized by the maximum value of each subject's absolute value. Additionally, we applied VTLN to suppress the inter-speaker variation from vocal tract length.

**VTLN:** We performed VTLN on spectrogram using a bilinear transform with warping factor $\alpha$ calculated, between 0.6 and 1.4 with 0.05 incremental, for each speaker over all utterances. The frequency-warped spectrogram was then converted back to its original waveform using the Griffin-Lim algorithm [34] that iteratively estimates audio phase based on the redundancy of Fourier transform.

**Implementation and training:** The cascading GRU model was implemented in Python [1] using the Pytorch library [35] and trained on two NVIDIA Tesla V100 GPUs. Mini-batch optimization was performed using the RMSprop optimizer and Binary Cross-Entropy loss $\big(-\frac{1}{N}\sum_{i=0}^{N} y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)\big)$.

As the optimization mechanism that performs gradient descent over a subsample, we discovered that it works best to train the proposed model using a mini-batch of 16 samples ($N$), each with 4000 ms audio length to predict the TBI class likelihood ($\hat{y}_i$) , from different subjects. The model was trained for a total of 120 epochs, but pre-trained pSinc layers were not optimized in the first 20 epochs as a domain adaptation phase. At epoch 21, fine-tuning starts, which optimized all parameters as an end-to-end model. The learning rate was fine-tuned to be 0.003 with a decay rate of $1e-6$.

Although the Coelho corpus contains almost equal numbers of subjects with TBI subject and healthy controls, subjects spoke for different lenghts of time, resulting in an unbalanced class problem in each mini-batch. We alleviated this problem by constructing each training batch to have equal numbers of TBI and healthy control classes, randomly at the beginning of each training epoch, and discarding any remaining samples. We found that discarding the remaining audio as an alternative to sampling with replacement resulted in a higher detection performance.

### D. Baselines

We used the acoustic features discussed previously in Section II-C, namely LLD, BOAW and formant coordination, as baseline features for pSinc and five classifiers, namely Support Vector Machine (SVM) [36], Random Forest (RF) [37] and Multi-Layer Perceptron (MLP) [38] and LSTM as Baseline models for cGRU.

#### 1) Baseline features:

**LLD:** LLDs were previously proposed to detect TBI by extracting LLD around vowels and feeding them into an SVM as acoustic features [10], [11]. This study used vowel boundary detection, a Praat script [2] by Hugo Quené with some modifications to locate vowel boundary in speech. The script was implemented following [39], which determines the vowel by identifying the peak in the pressure contour. LLDs

[1] http://github.com/aditthapron/speech-based-tbi-detection-using-psinc-cgru
[2] http://phonetics.linguistics.ucla.edu/facilities/acoustic/

2168-2194 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Worcester Polytechnic Institute. Downloaded on April 08,2022 at 15:12:57 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2022.3158840, IEEE Journal of Biomedical and Health Informatics



TABLE II: A list of tuned hyperparameters in all models.

| Model | Parameter(/Kernel) | Values |
|---|---|---|
| Proposed model | Sliding window (s) | 1,2,4,8 |
| | Number of pSinc filter | 16,32,64,128 |
| | pSinc filter size | 101,151,201,251,301,351 |
| | Number of CNN layer | 1,2,3 |
| | Number of CNN filter | 16,32,64,128 |
| | Dropout rate (%) | 0,10,20,30,40,50 |
| | Number of GRU layer | 1,2 |
| | GRU filter size | 4,8,16 |
| SVM | $C$ / All | $0.001, 0.01, 0.1, 1, 10, 100$ |
| | $\gamma$ / All | $0.001, 0.005, 0.01, 0.05, 0.1$ |
| | $d$ / Poly | $2, 3, 4, 5$ |
| RF | Max features | 1-10 |
| | Number of estimator | 100 - 500 with a step of 50 |
| | Max depth | 5 - 21 with a step of 3 |
| MLP | [(layer),(unit)] | [(0,1,2,3),(4,8,16,32,64)] |
| | Dropout | 0, 0.1, 0.2, 0.3, 0.4, 0.5 |

were extracted using the OpenSMILE library [40] with the COMPARE 2016 configuration [41]. A total of 130 LLD features were extracted over 20 ms audio length with a 10 ms time step. The LLDs within the vowel boundary were combined as an instance followed by a Principal Component Analysis (PCA) of 12 components, experimentally tuned, to reduce the dimension of LLD's.

**BOAW:** BOAW was used in place of PCA to minimize the dimensionality of LLD. After extracting the COMPARE 2016 features, two BOAW codebooks were created using OpenXBOW [42]. Each codebook contained 1000 audio words and was used to vectorize COMPARE 2016 features.

**Formant coordination:** Praat was used to extract the first three formant tracks from speech. We followed the steps in [43] to build a matrix of correlation and covariance coefficients for the first three formants over four time-delayed scales of 10 ms, 30 ms, 70 ms, and 150 ms. At the final step, total power and entropy constant values were estimated and included to the feature set. PCA was applied to reduce data dimensionality with 8 components.

#### 2) Baseline models:

**Machine learning classifier:** As a machine learning baseline, we considered Support Vector Machine (SVM) [36], Random Forest (RF) [37] and Multi-Layer Perceptron (MLP) [38]. Moreover, the GRU in cGRU was replaced with LSTM to quantify the benefit of GRU. All machine learning classifiers were implemented using the Scikit-learn library [44] with hyperparameters configured as reported in Table II.

**Proposed cGRU without sharing cell gating unit and cLSTM:** To examine the cell gating unit utility, we trained and evaluated the cGRU with different sequence sizes, i.e., the hidden state was re-initilized when the model accessed $n$ samples from the same subjects where $n$ is an independent variable. For cLSTM, we replace GRU with LSTM, which has an additional gate, to compare with the GRU.

### E. Evaluation method and Metrics

The Coelho corpus is balanced in terms of class between TBI and controlled, but is imbalanced in terms of gender, age, and education. These demographics have a significant impact on results. Consequently, we used multi-label classification method [45] that preserves the distribution of data across 10 folds. The subjects in Coelho's corpus were stratified into ten groups using the process introduced in [45]. Each model configuration and hyperparameters were determined using nested cross-validation, i.e. the testing fold was left out during this step. Specifically, the inner cross-validation preserves 8 folds for training and 1 fold was used to select the hyperparameter and to execute early stopping. The trained model with the lowest validation loss was then evaluated on the testing set that was left out in the outer fold. We report Balanced Accuracy (BAC=(Sensitivity + Specificity)/2), F1 score (F1 = $(2 \times \text{Precision} \times \text{Recall})/(\text{Precision} + \text{Recall})$), AUC score, sensitivity (recall) and specificity of testing set with standard error ($\sum_{i=1}^{\text{m}}(E_i - \overline{E})^2/\sqrt{m}$), where $E_i$ denotes estimated performance of fold $i$ from $m$ folds. To compare the model's effectiveness, the Wilcoxon signed-rank test was used to compare the rank between two distributions of measurements in all experiments.

### F. DNN interpretation

Although DNN has exhibited state-of-the-art performance, outperforming conventional algorithms, in various domains, it is regarded as a blackbox model that is difficult to interpret. In this study, we applied SHapley Additive exPlanations (SHAP) method [21] to the proposed model in order to analyze the acoustic features learned by the cGRU. The *impact score* or *contribution score* of speech components to the TBI prediction was estimated at each time-slice of the audio. Based on the contribution score, determined correlations of the input at specific timepoints to the TBI target labels, allowing us to perform analysis at word-level. We chose the attribution score over the TBI likelihood predicted by the Sigmoid activation in the final layer of cGRU because the Sigmoid activation predicts values over 4 seconds of audio, which is too coarse for the following experiment.

**SHAP:** We considered individual time-slices as features to the cGRU, which were compared in terms of impact to the TBI prediction using the SHAP method. The SHAP approach is based on the Shapley value from coalitional game theory, which distributes a fair gain or loss to a set of players in coalition. SHAP defines explanation model $g(z')$ over a coalition vector $z'$ in a form of $\phi_0 + \sum_{i=1}^{M}\phi_i z'_i$ where and $\phi_i$ is a feature attribution of feature $i$ from $m$ feature, which can be estimated from $\sum_{z' \subseteq x'} \frac{|z'|!(M-|z'|-1)}{M!}(f_x(z') - f_x(z'|z'=0))$. $\phi_i$ is computed on the the original model $f$ (proposed TBI detection model) and input sample $x$ over all $z'$ that are subsets of the non-zero $x'$, which is a simplified binary mapping of $x$, when feature $z'$ is not provided $(z'|z'=0)$.

We used the GradientSHAP [21] explainability method implemented in Captum library [46] to estimate the attribution score from the trained cGRU model. Gradient SHAP is a gradient method to compute SHAP values. The final SHAP values represent the expected value of gradients * (inputs - baselines). By averaging the scores within word boundaries, we conducted statistic analysis to evaluate the discrepancies between TBI and control subjects based on the spoken word.

2168-2194 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Worcester Polytechnic Institute. Downloaded on April 08,2022 at 15:12:57 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2022.3158840, IEEE Journal of Biomedical and Health Informatics



**Spectral-temporal attribution scores:** Given the presence of pSinc in the first layer of our model, additional attribution scores could be extracted in the spectral domain, allowing us to examine important components of a spectrogram. We considered the latent vector after the pSinc layer as an input to the model. This latent vector is similar to the filterbank feature, which can be used to estimate a frequency's attribution scores as follows. Let $l_j$ and $l_j$ be low frequency band and high frequency band learned filter $j$ of $N$ filters. Attribution $\phi_{freq}$ can be estimated from $\sum_{j=1}^{N} \phi'_j(\frac{1}{s[n]})$ using pre-computed $\phi'_j$ that corresponds to filter $j$ with $s[n] = 2f_c sinc(2f_c n)$. To analyze spectral feature, we summarized the attribution score for each work in the spectral domain and compared them between TBI and controlled subjects.

**Relationship between learned cGRU's features and other acoustic features:** All baseline acoustic features were derived from a short segment of audio, which can be matched with attribution scores from the cGRU using timestamps. For each acoustic feature, we computed correlations of TBI likelihood predicted by the acoustic feature with SVM to the corresponding SHAP score of the cGRU model. This analysis provides an import link between features developed by experts in speech processing to the features auto-learned by the cGRU.

## V. Results

### A. An evaluation of Cascading GRU for TBI classification

**Hyperparameter tuning results:** The proposed cGRU's hyperparameters are the number of CNN layers, which are located after the pretrained pSinc layer, and the number of GRU layers. The CNN and GRU configurations listed in Section IV were determined from a variety of configurations, as shown in Table II. The CNN and GRU hyperparameters were tuned together. The reported performances in CNN configuration were obtained from a GRU with 8 filters without dropout, while the reported performance for the GRU configuration were obtained from two layers of CNN with 32 filters with a 20% dropout rate. When tuning CNN hyperparameters, a configuration with two layers of CNN, 32 filters, and 20% dropout rate produces the highest balanced accuracy, as shown in Table III. We discovered that increasing the number of layers and the size of the filters could result in the overfitting problem, even when a dropout is used as a regularizer. The other two configurations in two layers of CNN also performed well in terms of F1 score and AUC, but we consider BAC as the primary metric. We also observed that a single layer of GRU with a filter size of 8 without dropout offered the highest balanced accuracy. This configuration was specifically tuned for the Coelho dataset, which had 50 subjects from each class. In a scenario in which more training data is available, increasing the number of GRU layers or substituting LSTM for GRU should be considered. The sliding window length is adjusted to 4 seconds as it maximizes TBI detection BAC. Compared to 2 seconds and 8 seconds, statistical significance is met at the 0.01 level using Wilcoxon's signed-rank test.

**cGRU evaluation:** To measure the benefit of cell gating unit in GRU, we plotted BAC across temporal steps, on which the cell gating unit is passed to the next sample in Figure 4. The temporal step in Figure 4 is a step between speech instance not the internal temporal step within the GRU cell. Without the prior cell gating unit, the performance of the cGRU is reduced to 63% BAC. The performance of cGRU is stable when the model has access to at least 50 samples, i.e., 200 seconds of spontaneous speech is required by the proposed method.

**cGRU enhancement:** The cGRU is enhanced further by including VTLN and skip layer connections, as shown in Table IV. VTLN normalizes the differences in vocal tract that may results from gender, age and other demographics whereas skipping layer connections provide different scale and complexity of temporal and spectral acoustic features to the latter layers in cascading GRU. We found that skip layer connections and VTLN substantially improve the cGRU at the 0.01 and 0.05 statistically significance levels respectively using Wilcoxon's signed-rank test. When cGRU is used in conjunction with skip layer connections and VTLN, the maximum output of 83.87 percent BAC is obtained (Wilcoxon signed-rank test, p-value $< 0.01$). VTLN normalizes vocal tract differences caused by gender, age, and other demographic variables, whereas skip layer connections provide different scale and complexity of temporal and spectral acoustic features to the subsequent layers in cGRU. All baselines are outperformed by cGRU with VTLN and skip layer connections. It can be observed that using the cell gating unit as an individual history improves the TBI detection BAC by 21.83 %.

**Compared with baseline models:** Furthermore, we compared the cGRU with cLSTM, SVM and MLP on three different features which were pSinc, LLD and formant frequency as shown in Figure 5. For the pSinc feature, cGRU outperforms cLSTM (Wilcoxon signed-rank test, p-value 0.05) and other models (Wilcoxon signed-rank test, p-value $< 0.01$). For formant frequency features, the performance of cLSTM and cGRU are similar and outperform other models (Wilcoxon signed-rank test, p-value $< 0.01$).

**Compared with baseline feature:** Using cGRU and cLSTM, pSinc significantly outperforms LLD and formant frequency (Wilcoxon signed-rank test, p-value $< 0.01$). pSinc with cGRU model are indicated as the most effective TBI detection in this study. Among the baselines, formant coordination stood out from the others and are closed to cGRU. We speculated that formant coordination features may share some similarity with cGRU, which was explored as an additional experiment below.

### B. Interpretation of the cGRU's network

**Temporal and Spectral features:** We investigated interpreting the acoustic features learned within cGRU using the SHAP attribution score for two distinct inputs to the model: 1) audio input (temporal attribution), and 2) pSinc features (spectral and temporal attributions).

Based on the audio input, We analyzed which words increased or decreased the likelihood of TBI prediction using the attribution score. The twenty most frequently spoken words in Coelho dataset were analyzed, namely, *the*, *and*, *i*, *a*, *to*, *it*, *yeah*, *uh*, *in*, *that*, *was*, *so*, *like*, *it's*, *of*, *they*, *you*, *um*, *he*, and *but*. The attribution scores between TBI subjects and healthy

2168-2194 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Worcester Polytechnic Institute. Downloaded on April 08,2022 at 15:12:57 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2022.3158840, IEEE Journal of Biomedical and Health Informatics



TABLE III: Cascading GRU for TBI detection: hyperparameter tuning

| cGRU configuration | | BAC | F1 | AUC | TPR | TNR | FNR | FPR |
|---|---|---|---|---|---|---|---|---|
| 1 CNN layer of | 16 filters with 0% dropout | 62.74 (1.22) | 69.26 (1.25) | 65.62 (0.87) | 74.83 (2.16) | 52.53 (2.25) | 25.17 (2.16) | 47.47 (2.25) |
| | 32 filters with 0% dropout | 64.33 (1.23) | 70.84 (1.28) | 63.32 (0.74) | 72.27 (2.00) | 58.59 (1.92) | 27.73 (2.00) | 41.41 (1.92) |
| | 64 filters with 20% dropout | 64.84 (1.23) | 70.14 (1.35) | 64.22 (0.80) | 75.72 (2.29) | 57.25 (2.18) | 24.28 (2.29) | 42.75 (2.18) |
| 2 CNN layers of | 16 filters with 0% dropout | 65.05 (1.47) | **71.42 (1.22)** | 79.30 (1.02) | 75.02 (2.16) | 56.03 (1.94) | 24.98 (2.16) | 43.97 (1.94) |
| | 32 filters with 20% dropout | **68.23 (1.41)** | 71.38 (1.42) | 79.08 (0.89) | 72.62 (2.34) | **63.92 (2.09)** | 27.38 (2.34) | **36.08 (2.09)** |
| | 64 filters with 20% dropout | 66.78 (1.45) | 70.29 (1.53) | **79.52 (1.00)** | 70.06 (2.08) | 63.18 (2.19) | 29.94 (2.08) | 36.82 (2.19) |
| 3 CNN layers of | 16 filters with 20% dropout | 65.73 (1.50) | 70.12 (1.46) | 70.45 (1.03) | **77.07 (2.26)** | 55.14 (2.07) | **22.93 (2.26)** | 44.86 (2.07) |
| | 32 filters with 20% dropout | 66.18 (1.51) | 70.43 (1.35) | 69.73 (0.90) | 72.73 (2.22) | 61.54 (2.30) | 27.27 (2.22) | 38.46 (2.30) |
| | 64 filters with 40% dropout | 63.52 (1.41) | 68.94 (1.13) | 69.12 (1.03) | 70.43 (2.14) | 55.26 (2.26) | 29.57 (2.14) | 44.74 (2.26) |
| 1 GRU layer of | 4 filters with 0% dropout | 56.35 (1.24) | 63.41 (1.11) | 70.13 (1.15) | 68.45 (1.86) | 50.31 (1.06) | 31.55 (1.86) | 49.69 (1.06) |
| | 8 filters with 0% dropout | **68.23 (1.41)** | 71.38 (1.42) | **79.08 (0.89)** | 72.62 (2.34) | 63.92 (2.09) | 27.38 (2.34) | 36.08 (2.09) |
| | 16 filters with 0% dropout | 65.85 (1.46) | 71.75 (1.50) | 76.41 (1.77) | 75.42 (1.41) | **69.84 (2.66)** | 24.58 (1.41) | **30.16 (2.66)** |
| 2 GRU layers of | 4 filters with 0% dropout | 58.72 (1.48) | 68.52 (1.08) | 71.51 (0.95) | 66.62 (1.98) | 52.62 (1.18) | 33.38 (1.98) | 47.38 (1.18) |
| | 8 filters with 0.3% dropout | 63.26 (1.68) | **72.41 (1.64)** | 78.26 (1.06) | **75.61 (2.62)** | 57.33 (2.06) | **24.39 (2.62)** | 42.67 (2.06) |
| | 16 filters with 0.3% dropout | 58.15 (1.73) | 70.63 (1.58) | 78.60 (1.10) | 75.57 (2.73) | 52.82 (1.98) | 24.43 (2.73) | 47.18 (1.98) |
| Sliding window length of | 1 second | 52.73 (2.08) | 68.53 (1.52) | 65.11 (0.85) | 56.33 (2.04) | 49.90 (2.43) | 43.67 (2.04) | 50.10 (2.43) |
| | 2 seconds | 56.64 (2.63) | 70.12 (1.55) | 72.41 (1.11) | 66.42 (1.13) | 52.78 (1.29) | 33.58 (1.13) | 47.22 (1.29) |
| | 4 seconds | **68.23 (1.41)** | **71.38 (1.42)** | **79.08 (0.89)** | **72.62 (2.34)** | **63.92 (2.09)** | **27.38 (2.34)** | **36.08 (2.09)** |
| | 8 seconds | 62.98 (2.25) | 71.21 (1.63) | 74.12 (1.48) | 65.11 (2.31) | 60.22 (1.88) | 34.89 (2.31) | 39.78 (1.88) |

Note: All results reported in this table does not apply VTLN on the input and do not have skipping layer connections.

TABLE IV: TBI detection results

| Model Configuration | | BAC | F1 | AUC | TPR | TNR | FNR | FPR |
|---|---|---|---|---|---|---|---|---|
| **cGRU:** | | 68.23 (1.01) | 71.38 (1.42) | 79.08 (0.89) | 72.62 (2.34) | 63.92 (2.09) | 27.38 (2.34) | 36.08 (2.09) |
| with VTLN | | 72.02 (1.51) | 74.91 (2.11) | 75.37 (1.20) | 78.15 (1.57) | 66.14 (1.62) | 21.85 (1.57) | 33.86 (1.62) |
| with skipping layer connections | | 77.35 (1.46) | 83.91 (1.58) | **89.52 (1.10)** | 82.15 (1.31) | 75.24 (1.10) | 17.85 (1.31) | 24.76 (1.10) |
| **with VTLN and skipping layer connections** | | **83.87 (1.65)** | **85.09 (2.58)** | 87.44 (1.21) | **90.55 (2.98)** | **77.02(1.97)** | **9.45 (2.98)** | **22.88 (1.97)** |
| **LLDs:** | **SVM** (RBF,0.01,0.05) | **56.37 (1.46)** | 60.14 (1.14) | 59.62 (1.20) | **62.62 (1.81)** | 51.11 (1.98) | **37.38 (1.81)** | 48.89(1.98) |
| | RF (4, 250, 8, 8, 2) | 52.18 (0.96) | **77.66 (1.30)** | **76.27 (1.23)** | 51.14 (0.83) | **52.75 (1.13)** | 48.86 (0.83) | **47.25 (1.13)** |
| | MLP (2,8,0.1) | 52.85 (1.21) | 74.32 (1.43) | 74.12 (1.15) | 53.41 (0.97) | 51.96 (1.09) | 46.59 (0.97) | 48.04 (1.09) |
| **BOAW:** | **SVM** (RBF,1,0.01) | **66.05 (1.36)** | **74.44 (1.21)** | 71.41 (2.00) | 71.03 (1.22) | **60.25 (2.49)** | 28.97 (1.22) | **39.75 (2.49)** |
| | RF (3, 100, 6, 11, 2) | 50.14 (1.46) | 73.72 (1.72) | 72.17 (1.15) | 47.02 (0.95) | 52.15 (1.32) | 52.98 (0.95) | 47.85 (1.32) |
| | MLP (2,4,0) | 62.97 (1.23) | 66.74 (1.20) | **75.28 (1.49)** | **78.08 (1.03)** | 57.04 (1.62) | **31.92 (1.03)** | 42.96 (1.62) |
| **Formant coordination:** | **SVM** (RBF,10,0.01) | **76.95 (1.56)** | **82.93 (1.87)** | **86.92 (1.90)** | 79.26 (1.94) | **69.42 (1.88)** | 20.74 (1.94) | **30.58 (1.88)** |
| | RF (4, 150, 8, 11, 5) | 66.53 (1.78) | 80.24 (1.55) | 72.55 (2.11) | 78.22 (1.18) | 62.41 (1.06) | 21.78 (1.18) | 37.59 (1.06) |
| | MLP (2, 8, 0.1) | 72.73 (1.62) | 80.37 (1.00) | 75.84 (1.30) | **79.33 (1.02)** | 64.03 (2.50) | **20.67(1.01)** | 35.97 (2.50) |
| Proposed method without sharing cell gating unit | | 68.84 (1.42) | 75.22 (2.08) | 76.90 (2.24) | 76.39 (1.88) | 67.90 (1.64) | 23.61 (1.88) | 32.10 (1.64) |

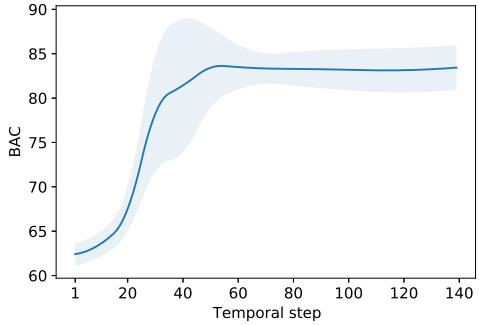



Fig. 4: Temporal step (data sample) of cGRU. The number of temporal steps is plotted against TBI detection accuracy.

controls are statistically different on the words *the*, *uh*, and *um* (Mann Whitney U Test, p-value $< 0.001$).

Additionally, we analyzed the attribution of these words in the spectral-temporal domains, but found no statistically important results. The example of spectral-temporal is shown in Figure 6. Due to the vast differences in vocal tract and word sustaining, we were unable to stack words from different speakers as in the previous experiment on the spectrogram. However, the attribution score can be used to examine segments that contribute to the TBI prediction, which are mostly located in high frequency bands.

**Correlation between cGRU and other acoustic features:** Scatter plots of pSinc attribution score and likelihood of TBI prediction from other acoustic features are shown in Figure 7. Pearson's correlation coefficients of 0.001, 0.192 and 0.910 are obtained for LLD features, BOAW and formant coordination, features respectively. According to the correlations, LLD and BOAW have no relationships with the cGRU. However, the BOAW plot demonstrates some patterns, which have two clusters of correct prediction and a narrow band of incorrect prediction when the attribution score of the cGRU is close to zero. The correlation indicates a strong relationship between cGRU and the formant coordination feature, which is consistent with the high TBI detection results. This analysis underlines the importance of continuous or delayed acoustic features.

## VI. DISCUSSION

### A. Principal findings

The proposed cGRU with pSinc features effectively detects TBI from spontaneous speech, outperforming previously proposed TBI detection methods and features by 6.92-27.5%

2168-2194 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Worcester Polytechnic Institute. Downloaded on April 08,2022 at 15:12:57 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2022.3158840, IEEE Journal of Biomedical and Health Informatics



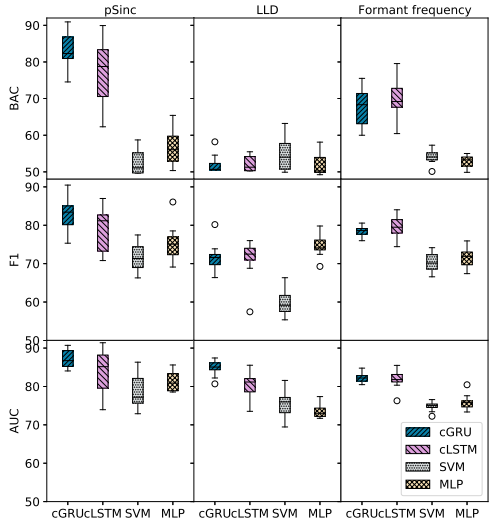



Fig. 5: TBI detection performance of cGRU, cLSTM, SVM and MVP using pSinc, LLD and formant frequency features

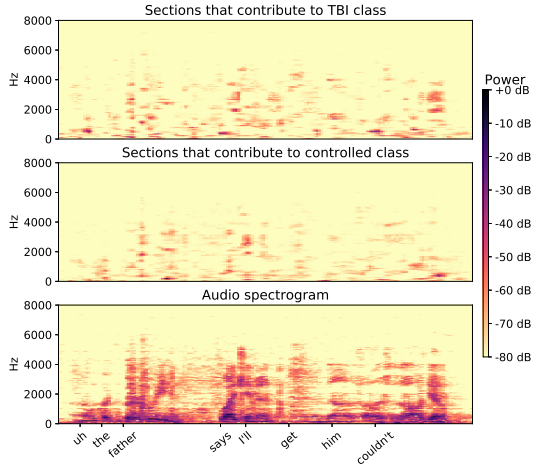



Fig. 6: Top: components of a spectrogram that the proposed model considered to be normal, center: components of a spectrogram that the proposed model considered to be TBI, bottom: input spectrogram

BAC. We demonstrated that the TBI detection benefits from continuous speech as carrying GRU's hidden state of the current sample to the next sample from the same subject increases BAC. Figure 4 demonstrates that once 60 samples (240 seconds of audio) are provided to the cGRU, the improvement ceases and the signal becomes steady. The model still benefits from data with fewer than 60 samples (240 seconds of speech) , but there is considerable volatility. A larger sliding window should be considered when more training data are available, allowing the model to run less infrequently and improve time efficiency. The main advantages of GRU come from the use of skipping layer connections followed by a hard dropout, which enables subsequent CNN layers to learn on audio representations extracted at different scales, and the use of VTLN, which reduces inter-speaker variation, allowing the model to learn TBI features that are less related to individual speech components. VTLN requires an additional ASR run to determine the warping factor for a new speaker, but the cost is negligible because it only needs to run once per speaker.

Among the baselines, formant coordination manifests the highest TBI detection accuracy where we found it has a substantial link to our proposed method based on TBI prediction likelihood. We speculate that the pSinc is capable of learning to extract formant frequency features or other spectral features, while the cGRU is capable of extracting linear and non-linear correlations between these features and across temporal domain, which are corresponding to the feature extraction steps in formant coordination.

## B. Limitations and passive speech recording concerns

*1) Speech quality in smartphone recording:* Although the cGRU was not tested on speech captured from a smartphone, this section discusses any degradations in recording efficiency as audio recording systems have shown some effects on speech assessment performance [6], [7].

Stasak *et al.* conducted an assessment, comparing three smartphone models in terms of PHQ-9 score and mental state predictions using LLD features [6]. The findings indicate that the accuracy difference between voice recorded using an acoustic cardioid and a smartphone in closed talks is less than 5%, that there is no statistical difference between feature extraction approaches, but that there is a difference between machine learning models.

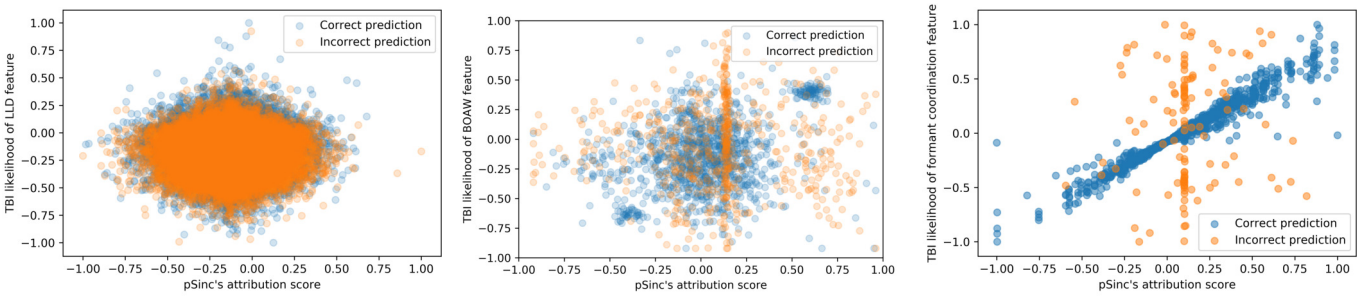
Smartphones have shown promising results in the mental and neurological disorders detection [6], [7]. However, the use of smartphones to assess TBI and monitor its recovery stage is still in its infancy due to multiple challenges which this work addresses. Specifically, we mitigate problems stemming from passive recording by using pSinc features and proposed using the cGRU to learn temporal relationships in acoustic features. With an ultimate goal of passively monitoring signs of diseases following TBI using speech collected on the smartphone, other major challenges are speaker privacy, noise, cross-talk, limited available data, language, and accents, which we aim to study in future.

*2) Privacy in speech feature:* Given the importance of speaker privacy, the use of passive recording in previous speech processing is now discussed. The audio recording of speech contains rich information on speech characteristics and linguistic content. Speech characteristics may implicitly reveal a speaker's biometric identity, personality, physical traits, age, gender, and health condition. Various encryption methods for speaker authentication were categorized by [47] according to three criteria. These three criteria posit that speech data, in order to be considered privacy-preserving, need to be: *unlinkable*, *irreversible*, and *renewable*. Unlinkability prevents speech collected at different scenarios to be related. Irreversibility prevents features from being reverted back to raw audio. Renewability is specific for forensic aspects of speaker authentication, which does not apply in the study at this step. Consider that all acoustic features in this work, including the baseline methods, are extracted on the smartphone and classified on the server. The main question is "To what extent does the server have access to the user's private speech?"

2168-2194 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Worcester Polytechnic Institute. Downloaded on April 08,2022 at 15:12:57 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2022.3158840, IEEE Journal of Biomedical and Health Informatics






(a) LLD feature (b) BOAW feature (c) Formant coordination feature

Fig. 7: Correlation between (a) LLD, (b) BOAW, (c) formant coordination features and TBI target labels classified using cGRU

TABLE V: Privacy in speech feature

| Acoustic feature | Unlinkable | Irreversible | Average |
|---|---|---|---|
| Raw audio | 1 | 1 | 1 |
| LLDs | 2 | 2 | 2 |
| BOAW | 1 | 3 | 2 |
| Formant coordination | 2 | 3 | 2.5 |
| pSinc feature | 1 | 2 | 1.5 |

In Table V, we conceptually categorize the privacy level of each acoustic feature. We rate the two criteria based on their fundamental and previous works on a scale of one to three. If the fundamental of acoustic feature does not meet the criteria, a score of one is assigned. A score of two is assigned if prior work has been conducted to demonstrate that the acoustic function does not meet the requirements. A score of three is assigned if no prior work has been conducted to determine if the acoustic function violates the criteria. This privacy-focused analyses of features is particularly important because due to IRB restrictions, in order to maintain speaker privacy, speech features are used in place of raw as a privacy preservation mechanism. Specifically, in many speech assessment systems, raw audio is featurized right on the smartphone. Thereafter only the speech features and not raw audio, are stored on servers and analyzed. Thus, it is important to fully understand how well each feature type preserves the speaker's privacy.

As a data source, raw audio is scored as one in all criteria as. LLDs features include Mel-Frequency Cepstral Coefficients (MFCCs), which are used for speaker recognition [48] and can be inverted to raw audio [49]. BOAW algorithm incorporate unsupervised clustering based on LLD feature, which violates the unlinkable property by its fundamental although the cluster's centroid is unknown, making it irreversible. For formant coordination, formant track has been used for speaker recognition [50], which shows that speech is linkable through formant, but there is no work showing that speech can be reconstructed from the formant to the best of our knowledge. pSinc feature was developed and trained for speaker recognition task, which shows that the feature is linkable [17]. Even though audio reconstruction of pSinc has not been examined, previous work has shown that audio can be reconstructed from spectral information using an autoencoder [51].

### C. Future work

This study proposed a method to detect TBI in continuous, spontaneous speech. However, before the framework can be evaluated as an fully passive TBI assessment on smartphones, it would need speech pre-processing step to minimize noise in addition to our previously proposed work that isolates speech during cross-talk [16]. We previously claimed that cGRU benefited from hidden state information taken from previous data recordings, but the evaluation in this study was not performed on longitudinal data. It is possible that the hidden state may fail to adapt on new speech recorded in a different context or with a long interval to previous input.

## VII. Conclusion

This study proposed a continuous TBI assessment for spontaneous speech as an alternative to conversational speech assessment that performs episodically during specific periods of interest. The model performs on acoustic features using learnable parametrized Sinc filter with GRU to cascade speech overtime. The primary advantages of cGRU are due to the cell gating unit, which provides the model with prior information about each subject. We innovatively adopted the pSinc function, which has previously learned to extract acoustic features from raw audio for the TBI detection task. cGRU outperforms all baseline methods in TBI detection with a balanced detection accuracy of 83.87%. The proposed network's interpretation shows that features learned in pSinc and GRU are correlated with formant coordination, and some filler words are prominent among TBI subjects. In future work, the framework proposed in this study can be extended to passively monitor subject recovery following TBI using speech recorded on the smartphone.

## References

[1] M. Faul, M. M. Wald, L. Xu, and V. G. Coronado, "Traumatic brain injury in the united states; emergency department visits, hospitalizations, and deaths, 2002-2006," 2010.

[2] F. M. Hammond, J. D. Corrigan, J. M. Ketchum, J. F. Malec, K. Dams-O'Connor, T. Hart, T. A. Novack, J. Bogner, M. N. Dahdah, and G. G. Whiteneck, "Prevalence of medical and psychiatric comorbidities following traumatic brain injury," The Journal of head trauma rehabilitation, vol. 34, no. 4, pp. E1–E10, 2019.

[3] R. S. Norman, C. A. Jaramillo, M. Amuan, M. A. Wells, B. C. Eapen, and M. J. Pugh, "Traumatic brain injury in veterans of the wars in iraq and afghanistan: Communication disorders stratified by severity of brain injury," Brain injury, vol. 27, no. 13-14, pp. 1623–1630, 2013.

2168-2194 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Worcester Polytechnic Institute. Downloaded on April 08,2022 at 15:12:57 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2022.3158840, IEEE Journal of Biomedical and Health Informatics




[4] M. McHenry, "Acoustic characteristics of voice after severe traumatic brain injury," The Laryngoscope, vol. 110, no. 7, pp. 1157–1161, 2000.

[5] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, "A review of depression and suicide risk assessment using speech analysis," Speech Communication, vol. 71, pp. 10–49, 2015.

[6] B. Stasak and J. Epps, "Differential performance of automatic speech-based depression classification across smartphones," in IEEE ACIIW. IEEE, 2017, pp. 171–175.

[7] J. Almeida, P. Rebouas, T. Carneiro, W. Wei, R. Damaeviius, R. Maskelinas, and V. H. de Albuquerque, "Detecting parkinson's disease with sustained phonation and speech signals using machine learning techniques," Pattern Recognition Letters, vol. 125, pp. 55–62, 2019.

[8] B. S. Helfer, T. F. Quatieri, J. R. Williamson, L. Keyes, B. Evans, W. N. Greene, T. Vian, J. Lacirignola, T. Shenk, T. Talavage et al., "Articulatory dynamics and coordination in classifying cognitive change with preclinical mtbi," in ISCA, 2014.

[9] A. C. Lammert, J. R. Williamson, A. Hess, T. Patel, T. F. Quatieri, H. J. Liao, A. Lin, and K. J. Heaton, "Noninvasive estimation of cognitive status in mild traumatic brain injury using speech production and facial expression," in IEEE ACII, 2017, pp. 105–110.

[10] M. Falcone, N. Yadav, C. Poellabauer, and P. Flynn, "Using isolated vowel sounds for classification of mild traumatic brain injury," in IEEE ICASSP. IEEE, 2013, pp. 7577–7581.

[11] C. Poellabauer, N. Yadav, L. Daudet, S. L. Schneider, C. Busso, and P. J. Flynn, "Challenges in concussion detection using vocal acoustic biomarkers," IEEE Access, vol. 3, pp. 1143–1160, 2015.

[12] L. Daudet, N. Yadav, M. Perez, C. Poellabauer, S. Schneider, and A. Huebner, "Portable mtbi assessment using temporal and frequency analysis of speech," IEEE JBHI, vol. 21, no. 2, pp. 496–506, 2016.

[13] T. Talkar, S. Yuditskaya, J. R. Williamson, A. Lammert, H. Rao, D. Hannon, A. O'Brien, G. Vergara-Diaz, R. DeLaura, D. Sturim et al., "Detection of subclinical mild traumatic brain injury (mtbi) through speech and gait," INTERSPEECH, pp. 135–139, 2020.

[14] B. N. Renn, A. Pratap, D. C. Atkins, S. D. Mooney, and P. A. Areán, "Smartphone-based passive assessment of mobility in depression: Challenges and opportunities," Mental health and physical activity, vol. 14, pp. 136–139, 2018.

[15] P. R. Center, "Demographics of mobile device ownership in the united states 2021," Pew Research Center: Internet, Science & Tech, 2021.

[16] A. Ditthapron, E. O. Agu, and A. C. Lammert, "Privacy-preserving deep speaker separation for smartphone-based passive speech assessment," IEEE OJEMB, 2021.

[17] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with sincnet," in IEEE SLT. IEEE, 2018, pp. 1021–1028.

[18] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," arXiv preprint arXiv:1904.05862, 2019.

[19] L. Lee and R. C. Rose, "Speaker normalization using efficient frequency warping procedures," in IEEE ICASSP, vol. 1, 1996, pp. 353–356.

[20] C. A. Coelho, K. M. Youse, and K. N. Le, "Conversational discourse in closed-head-injured and non-brain-injured adults," Aphasiology, vol. 16, no. 4-6, pp. 659–672, 2002.

[21] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," Advances in Neural Information Processing Systems, vol. 30, pp. 4765–4774, 2017.

[22] B. E. Murdoch and D. G. Theodoros, "Traumatic brain injury: Associated speech, language, and swallowing disorders," 2001.

[23] Y. Lebrun, "Apraxia of speech: a critical review," Journal of Neurolinguistics, vol. 5, no. 4, pp. 379–406, 1990.

[24] A. S.-L.-H. Association et al., "Guidelines for speech-language pathologists serving persons with language, socio-communicative, and/or cognitive-communicative impairments," ASHA, vol. 33, no. Suppl. 5, pp. 21–18, 1991.

[25] Y.-t. Wang, R. D. Kent, J. R. Duffy, J. E. Thomas, and G. Weismer, "Alternating motion rate as an index of speech motor disorder in traumatic brain injury," Clinical Linguistics & Phonetics, vol. 18, no. 1, pp. 57–84, 2004.

[26] C. Kohlschein, M. Schmitt, B. Schüller, S. Jeschke, and C. J. Werner, "A machine learning based system for the automatic evaluation of aphasia speech," in IEEE Healthcom, 2017, pp. 1–6.

[27] A. Grünerbl, A. Muaremi, V. Osmani, G. Bahle, S. Oehler, G. Tröster, O. Mayora, C. Haring, and P. Lukowicz, "Smartphone-based recognition of states and state changes in bipolar disorder patients," IEEE JBHI, vol. 19, no. 1, pp. 140–148, 2014.

[28] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold et al., "Cnn architectures for large-scale audio classification," in IEEE ICASSP, 2017, pp. 131–135.

[29] Y. Pan, B. Mirheidari, Z. Tu, R. O'Malley, T. Walker, A. Venneri, M. Reuber, D. Blackburn, and H. Christensen, "Acoustic feature extraction with interpretable deep neural network for neurodegenerative related disorder classification," INTERSPEECH, pp. 4806–4810, 2020.

[30] M. Kim, B. Cao, K. An, and J. Wang, "Dysarthric speech recognition using convolutional lstm neural network," INTERSPEECH, pp. 2948–2952, 2018.

[31] E. Rejaibi, A. Komaty, F. Mériaudeau, S. Agrebi, and A. Othmani, "Mfcc-based recurrent neural network for automatic clinical depression recognition and assessment from speech," arXiv preprint arXiv:1909.07208, 2019.

[32] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," arXiv preprint arXiv:1412.3555, 2014.

[33] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in IEEE ICASSP. IEEE, 2015, pp. 5206–5210.

[34] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform," IEEE Transactions on acoustics, speech, and signal processing, vol. 32, no. 2, pp. 236–243, 1984.

[35] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in Advances in Neural Information Processing Systems 32. Curran Associates, Inc., 2019, pp. 8024–8035.

[36] C. Cortes and V. Vapnik, "Support-vector networks," Machine learning, vol. 20, no. 3, pp. 273–297, 1995.

[37] A. Liaw, M. Wiener et al., "Classification and regression by randomforest," R news, vol. 2, no. 3, pp. 18–22, 2002.

[38] P. Werbos, "Beyond regression:" new tools for prediction and analysis in the behavioral sciences," Ph. D. dissertation, Harvard University, 1974.

[39] F. Cummins and R. Port, "Rhythmic constraints on stress timing in english," Journal of Phonetics, vol. 26, no. 2, pp. 145–171, 1998.

[40] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in ACM international conference on Multimedia, 2010, pp. 1459–1462.

[41] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi et al., "The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in INTERSPEECH, 2013.

[42] M. Schmitt and B. Schuller, "Openxbow: introducing the passau open-source crossmodal bag-of-words toolkit," The Journal of Machine Learning Research, vol. 18, no. 1, pp. 3370–3374, 2017.

[43] J. R. Williamson, D. Young, A. A. Nierenberg, J. Niemi, B. S. Helfer, and T. F. Quatieri, "Tracking depression severity from audio and video based on speech articulatory coordination," Computer Speech & Language, vol. 55, pp. 40–56, 2019.

[44] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[45] K. Sechidis, G. Tsoumakas, and I. Vlahavas, "On the stratification of multi-label data," in ECML KDD. Springer, 2011, pp. 145–158.

[46] N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, J. Reynolds, A. Melnikov, N. Lunova, and O. Reblitz-Richardson, "Pytorch captum," 2019.

[47] A. Nautsch, A. Jiménez, A. Treiber, J. Kolberg, C. Jasserand, E. Kindt, H. Delgado, M. Todisco, M. A. Hmani, A. Mtibaa et al., "Preserving privacy in speaker and speech characterisation," Computer Speech & Language, vol. 58, pp. 441–480, 2019.

[48] V. Tiwari, "Mfcc and its applications in speaker recognition," International journal on emerging technologies, vol. 1, no. 1, pp. 19–22, 2010.

[49] B. Milner and X. Shao, "Prediction of fundamental frequency and voicing from mel-frequency cepstral coefficients for unconstrained speech reconstruction," IEEE transactions on audio, speech, and language processing, vol. 15, no. 1, pp. 24–33, 2006.

[50] M. Chougala and S. Kuntoji, "Novel text independent speaker recognition using lpc based formants," in IEEE ICEEOT. IEEE, 2016, pp. 510–513.

[51] J. Engel, C. Resnick, A. Roberts, S. Dieleman, M. Norouzi, D. Eck, and K. Simonyan, "Neural audio synthesis of musical notes with wavenet autoencoders," in ICML. PMLR, 2017, pp. 1068–1077.


2168-2194 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Worcester Polytechnic Institute. Downloaded on April 08,2022 at 15:12:57 UTC from IEEE Xplore. Restrictions apply.