Interpretable Self-Supervised Facial Micro-Expression Learning to Predict Cognitive State and Neurological Disorders

Arun Das, [1] **Jeffrey Mock,** [2] **Yufei Huang,** [1] **Edward Golob,** [2] **Peyman Najafirad** [1]*

[1] Secure AI and Autonomy Laboratory, University of Texas at San Antonio
[2] Cognitive Neuroscience Laboratory, University of Texas at San Antonio
{arun.das, jeffrey.mock, yufei.huang, edward.golob, paul.rad}@utsa.edu

Abstract

Human behavior is the confluence of output from voluntary and involuntary motor systems. The neural activities that mediate behavior, from individual cells to distributed networks, are in a state of constant flux. Artificial intelligence (AI) research over the past decade shows that behavior, in the form of facial muscle activity, can reveal information about fleeting voluntary and involuntary motor system activity related to emotion, pain, and deception. However, the AI algorithms often lack an explanation for their decisions, and learning meaningful representations requires large datasets labeled by a subject-matter expert. Motivated by the success of using facial muscle movements to classify brain states and the importance of learning from small amounts of data, we propose an explainable self-supervised representation-learning paradigm that learns meaningful temporal facial muscle movement patterns from limited samples. We validate our methodology by carrying out comprehensive empirical study to predict future speech behavior in a real-world dataset of adults who stutter (AWS). Our explainability study found facial muscle movements around the eyes (p <0.x001) and lips (p <0.001) differ significantly before producing fluent vs. disfluent speech. Evaluations using the AWS dataset demonstrates that the proposed self-supervised approach achieves a minimum of 2.51% accuracy improvement over fully-supervised approaches.

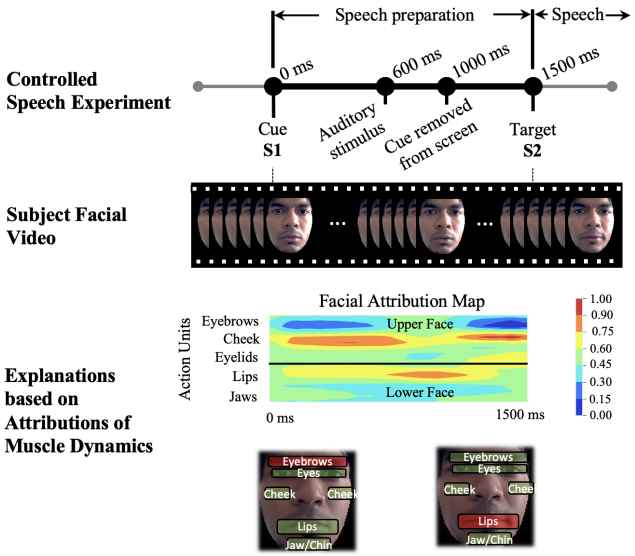

Figure 1: Facial muscle movement dynamics of individuals with speech disorders indicates large variations in upper and lower muscle groups during speech disfluency. We learn to distinguish these subtle micro-expressions using our proposed self-supervised explainable algorithm.

Introduction

Every action is preceded by information flow between specific sensory and motor networks responsible for producing the desired movement. This information flow resides in distributed neuronal networks that are in a constant state of flux. Research has shown micro involuntary movements also accompany many desired voluntary movements. To date, most research on human behavior focuses on the easy to quantify factors of voluntary movements like reaction times and accuracy while ignoring the information present in the involuntary movements. Involuntary movements have been historically ignored due to the difficulty in identifying, quantifying and interrupting the information present in them.

The face is a known body region to be rich with both voluntary and involuntary movements. It is evident that specific combinations of voluntary facial muscle activations could even induce real emotions (Ekman, Levenson, and Friesen 1983) varying slightly between cultures (Ekman 1992). Facial activity is a treasure-chest of information which aids psychologists and neuropsychologists to examine and diagnose various diseases (Wu et al. 2014; Bandini et al. 2016). These facial activities are usually micro-expressions (Oh et al. 2018) which are subtle involuntary facial expressions happening briefly (Ekman 2009; Verma et al. 2019) in the upper and lower facial regions.

The successes in encoding facial muscle patterns as facial Action Units (AUs) (Ekman and Rosenberg 2005; Friesen and Ekman 1978) and using AUs to study attention, affect (Sayette et al. 2002; Hamm et al. 2011; Lints-Martindale et al. 2007), and pain (Kunz, Meixner, and Lautenbacher

*Corresponding Author

2019) demonstrates the use of data-driven study of facial activities in estimating cognitive states. The recent surge in using Deep Learning (DL) based Artificial Intelligence (AI) algorithms in modeling facial activities to detect micro-expressions and diseases (Bandini et al. 2017; Jiang et al. 2020) affirm the use of data-driven approaches for automated analysis of the face in the field of medicine. Numerous studies also establish relationships between upper and lower facial muscle movements in affect (Wang, Wang, and Ji 2013; Wehrle et al. 2000; Mehu et al. 2012) and speech motor preparation (Meng, Han, and Tong 2017). Together, these provide a new venue to explore facial muscle movement dynamics for various neuro-cognitive disorders, especially in studying speech disfluencies.

Supervised DL algorithms, however, require a large amount of data to optimize the parameters of deep neural networks such as Convolutional Neural Networks (CNNs). Several authors have used more than 100,000 labeled images to train CNNs for disease prediction (Gulshan et al. 2016; Das et al. 2019). However, manual labeling of medical data is resource intense, takes up many human work hours, and could induce data bias. To mitigate this, techniques such as self-supervised learning have been proposed to impute existing data (Cao et al. 2020) and learn better representations from a limited amount of labeled data (Li et al. 2020). Also, there is a need to explain the decisions of DL classifiers for mission-critical tasks such as in healthcare (Das and Rad 2020).

In this paper, motivated by the use of facial AUs in learning human behavior and the importance of learning from few samples, we propose a self-supervised pre-training algorithm to learn dense representations of facial muscle movements without the need for a large amount of labeled data. Here, we define self-supervised pre-training (pretext) tasks to learn meaningful residual dynamics and micro-expressions from facial muscle movement information. Once the pretext model is trained, we adapt the pre-trained model for a downstream task of predicting the cognitive states of Adults-Who-Stutter (AWS) based on facial muscle movement information. We validate our methodology on a video dataset to detect stuttering speech disfluency from facial muscle movements of AWS. More specifically, we try to predict the future onset of stuttering based on temporal facial AU data available before the speech vocalization by training classifiers on the embeddings extracted using the pre-trained pretext models. In summary, our main contributions include:

- We propose a self-supervised learning scheme to extract meaningful micro-expression facial muscle movement representations for predicting cognitive states and neurological disorders.
- We collected the first real-world multimodal behavioral dataset of AWS subjects and performed the first explainable self-supervised stuttering disfluency pilot study. We show that our self-supervised approach can train with a very limited amount of data while providing information about muscle movement dynamics between fluent and disfluent trials.
- We present our findings which affirm prior knowledge of secondary motor behaviors during the onset of stuttering disfluency, especially in the muscles around eye brows and lips, which could generalize speech-motor behavior of AWS.

Related Work

Self-Supervised Learning. To learn meaningful representations from available unlabeled data and apply the knowledge to improve the performance of another domain, self-supervised learning methods usually consist of a pretext and downstream task. Pretext tasks are designed to learn and convert the features and feature correlations to dense vector representations which can later be used by downstream tasks (Sheng et al. 2020). Here, the downstream tasks may be trained for similar or slightly different data problems. State-of-the-art self-supervised methods are now able to learn temporal correspondence in videos (Li et al. 2019; Tschannen et al. 2020), speech representations (Shukla et al. 2020), predicting retinal diseases (Rivail et al. 2019), disfluency detection from the text (Wang et al. 2019), and many more.

Stuttering Disfluencies and Facial Muscle Activity. Speech disfluencies, such as stuttering, are disruptions to the normal flow of speech, and include word or syllable prolongations, silent blocks, and part-word repetitions. Among children under the age of five, 2.5% stutter (Proctor, Duff, and Yairi 2002; Yairi and Ambrose 1999). Of those who stutter as children, about 20% continues to stutter as adults (Craig and Tran 2005). Stuttering reflects multiple stable factors of the individual, attempted speech message, and speaking context (Bloodstein and Ratner 2008).

Recent advances in AI, especially in DL, have enabled research in various data-driven learning approaches to detect stuttering disfluency. Some of the data modalities include respiration rate (Villegas et al. 2019), and audio (Zhang, Dong, and Yan 2013) during speech vocalization. A few studies have looked at supervised DL algorithms to detect stuttering disfluency from pre-speech EEG data during speech preparation in AWS (Myers et al. 2019).

Initial sound or syllable of speech contributes to over 90% of speech disfluency (Sheehan 1974). This highlights the importance of understanding pre-speech secondary behaviors to ultimately understand the neuro-psychological aspects and cerebral activities of an individual as they prepare to speak. People who stutter often have "secondary behaviors" while speaking, such as eye blinking, involuntary movements of head or limbs, jerks in the jaw, etc. (Prasse and Kikano 2008). Secondary behaviors may be due to imprecise motor control in the brain, termed "motor overflow" (Hoy et al. 2004). Facial and vocal articulators are represented near each other in human motor cortex (Penfield and Boldrey 1937), suggesting that speech motor overflow may influence facial muscle activity. Considering various secondary behaviors and facial changes on or before speech vocalization, we see an opportunity to study the facial muscle movement patterns before speech utterance to classify stuttering disfluency. The present study will test this possibility in AWS.

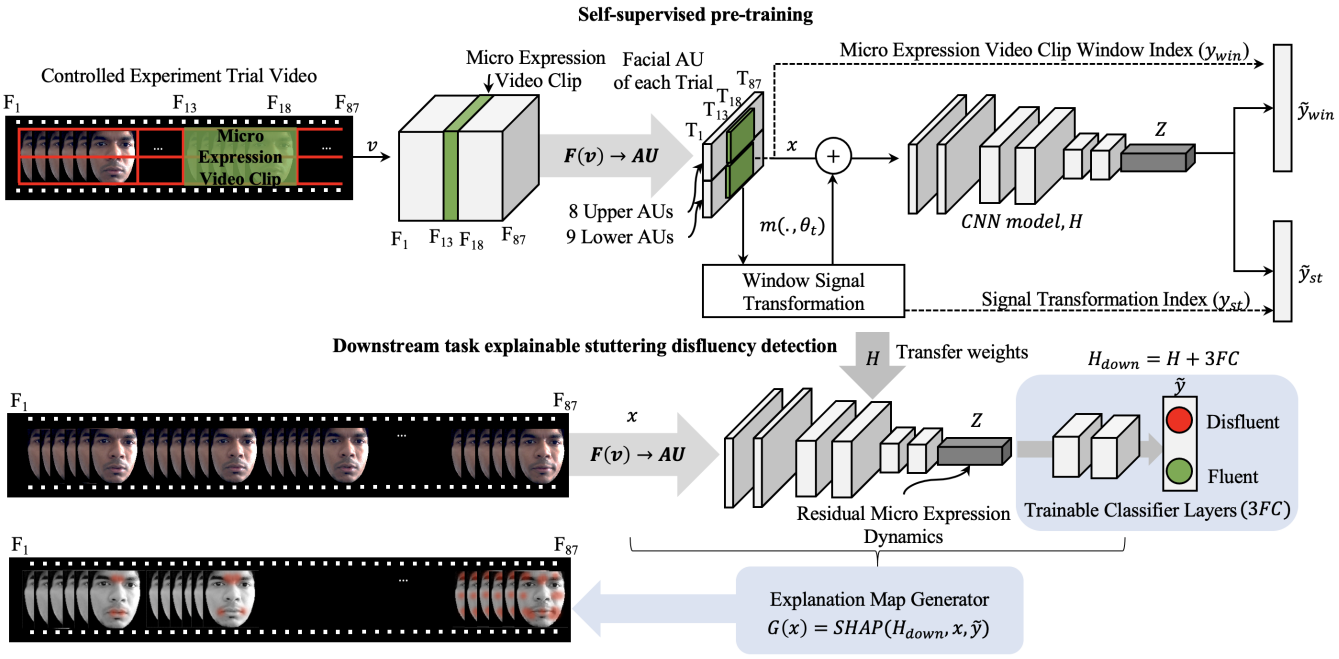

Figure 2: A high-level diagram of our method. We propose two pretext tasks to learn muscle movement dynamics and micro-expressions without labeled data. Once pre-trained, we use the learned embeddings to carry out stuttering disfluency classification. DeepSHAP method is applied to generate explanations of the muscles involved in fluent and disfluent speech.

Self-Supervised Learning for Stuttering Disfluency. At the time of writing, no published studies have looked at the facial muscle activities during or before speech vocalization to study developmental stuttering. We hypothesize that the facial muscle movement activities during a few seconds of speech preparation encode enough secondary behavior information to classify a future speech vocalization as either fluent or disfluent. However, we still face the challenge of collecting real-world patient data for stuttering-based studies. Hence, there is a requirement to learn good facial muscle movement representations which can model stuttering disfluency from a small number of data samples.

Methods

In order to evaluate the effectiveness of our self-supervised learning paradigm on a task that is not explored in published research, we construct a data-driven approach to study learning performance of the proposed solution. Here, we present the self-supervised learning methodology and discuss the architecture and design considerations for the pretext and downstream task for stutter disfluency classification. A high-level diagram is illustrated in Figure 2. Hyper parameters and training details are summarized in individual subsections.

Self-Supervised Representation Learning

Our self-supervised representation learning approach is driven by the primary psychological observation that muscle movements in the face are driven by universal changes (Ekman and Rosenberg 2005; Friesen and Ekman 1978) with upper and lower facial muscle movements showing different aspects of anticipation and motor behavior (Wang, Wang, and Ji 2013; Meng, Han, and Tong 2017). Stuttering, as a speech-motor disease, shows numerous secondary behaviors on or before speech utterance (Prasse and Kikano 2008) as we discussed above.

Self-supervised Pre-training. We use a classical machine learning algorithm to extract facial AUs to model facial muscle movements from HOG features and face geometry (Baltrusaitis, Mahmoud, and Robinson 2015). Let v be a video collected from an individual for a given time-window T, we extract facial action units AU_i for each frame F_n where $i \in I$ is the action units extracted and $n \in N$ is the number of frames. In this study, I's cardinal number is 17 and includes upper AUs 1, 2, 4, 5, 6, 7, 9, and 45, and lower AUs 10, 12, 14, 15, 17, 20, 23, 25, and 26. Facial AU data becomes $\{x|x \in \mathrm{R}^{I \times N}\}$ for each input video segment, with upper facial AUs and lower facial AUs grouped together. The goal of the self-supervised pre-training is to learn temporal correlations between the AUs in x using different pretext tasks. Towards this goal, we present the following tasks for pre-training:

Micro-expression Window Finder: In this task, we select a small 100 ms time-window from the AU input map x, enough to include micro-expressions, and ask the DL network to predict which window was selected. At any given time, we select AUs from either the upper or lower face and never together. This helps to disentangle the representations and lessen the reliance on specific areas of the face. After selecting a window, we carry out certain signal transforma-

tions. This way, the DL algorithm could spatio-temporally find the correct window chosen. If no window was chosen (no signal transformations) we label it differently to avoid mistakes. Altogether if we have W windows, window finder task will have $W + 1$ labels named y_{win}.

Micro-expression Window Loss: The loss function of window finder task should reinforce the models' ability to disentangle upper and lower facial regions as well as understanding areas-of-interest. Considering the DL pre-training model as H and the selected window as w, the loss term associated with window finder task is defined as:

$$\begin{aligned} \mathcal{L}_{\text{win}}(w) &= -\log P_{win}(w; H) \\ P_{win}(w; H) &= P(\tilde{y}_{win} = y_{win} | w; H) \end{aligned} \tag{1}$$

where y_{win} is the label for window w, $\tilde{y}_{win}$ is the prediction by DL model, and P_{win} is the probability of picking the correct window.

Signal Transformations: The AU map may have inherent errors due to sudden jerks, movements, occlusions, etc. of the subjects' faces during data collection. A robust and meaningful face representation from the temporal AU data should be invariant to sudden changes in a large group of AUs. Due to the limited available data samples, the pre-training model might not be able to learn to mitigate such anomalies. Hence, for a selected window w of the temporal AU map, we augment the data with transformations such as scaling, Gaussian noise, and zero-filling with predefined parameters of augmentation.

Each transformation has a neurological or data robustness relation. For example, scaling the AUs in a specific time window relates to sudden changes in the facial muscle regions over a period of 100 ms. Gaussian noise could simulate random jitters, partial occlusions, etc. while zero-filling could simulate complete occlusion of either upper or lower facial region such as during a hand moving across the face. The actual AU map is also kept as an input without applying transformations. In this case, as discussed above, the label of the window is chosen to indicate that there is no window to look for. The labels for the signal transformation task is based on various transformations and their parameters as summarized in Table 1.

Signal Transformation Loss: For each input, given a window w and a signal transformation st with signal manipulation parameter θ_{st} for cases in $m(\cdot, \theta_{st})$, we follow the degradation identification loss presented in (Sheng et al. 2020), and define the loss term for the signal transformation task st to reinforce the model to learn various signal degradations, sharp increases in facial movement, etc. applied to the window:

$$\begin{aligned} \mathcal{L}_{\text{st}}(w, \theta_{st}) &= -\log P_{st}(w; H) \\ P_{st}(w; H) &= P(\tilde{st} = st | m(w, \theta_{st}); H) \end{aligned} \tag{2}$$

Total loss function: To improve the learning of micro-expressions and variations of facial activities, we formulate the final self-supervised pre-training as a linear combination of the window loss and signal transformation loss as described below:

Signal Transformation	Parameters θ_{st}
Scaling	{0.25, 0.5, 1.25, 1.75}
Gaussian Noise	{0.1, 0.25, 0.5, 0.75, 0.9}
Zero Fill	-
None	-

Table 1: Signal transformation tasks and parameters used to improve pretext tasks.

$$\mathcal{L} = \alpha \cdot \mathcal{L}_{\text{win}}(w) + \beta \cdot \mathcal{L}_{\text{st}}(w, \theta_{st}) \tag{3}$$

where α and β are used to balance the loss term.

Stuttering Disfluency Downstream Task. Understanding the nuances of stuttering disfluency from a pre-speech facial muscle movement perspective is previously unexplored. Learning from facial muscle movements is a hard problem since pre-speech data have more micro-expressions than large upper or lower facial muscle movements. Since real-world subject data is limited, we plan to use the self-supervised pre-trained model as a way of improving the performance of stuttering disfluency classification. We describe here some of the information regarding the disfluency study.

Subjects: We collected video data from a cohort of AWS subjects with a mean age of 23 (18-31 years). Each subject self-reported to have developmental stuttering from childhood and was diagnosed by a speech-language pathologist for verification. Each subject was invited to attend data collection sessions where they were presented with an experimental task. On average, the subjects attended 3.7 sessions. All studies were done under strict protocols of the Institutional Review Board and the Declaration of Helsinki.

AWS Speech Study and Hardware-Software Design: The modulations in brain state of AWS have been studied by introducing a delay that separates speech preparation from speech execution, with the objective of identifying brain activity during speech preparation that predicts later execution of fluent vs. disfluent speech. We have designed a controlled experiment to measure AWS speech preparation tasks using auditory and visual cues. A small delay is imposed using a specially designed 'S1-S2' task to separate speech preparation and utterance. The subject is seated in a sound booth and is asked to read pseudo word-pairs from a computer monitor in front of them. The word-pairs are chosen to mimic English language phonetically while removing emotional responses or meaning of words to normalize the ratio of fluent and disfluent trials across subjects.

Individual trials of S1-S2 tasks span only 1500 ms, where the subject is shown with a visual stimuli at S1 and is expected to speak at S2. According to set paradigms, the subject either sees the pseudo-word-pairs at S1 or at S2. This is done to study the effect of having the word in memory towards speech disfluency. The paradigm where the subject sees a word first is termed "Word-Go" (WG). Here, the subject sees a word-pair at S1 and "!!!" symbol at S2. In "Cue-Word" (CW), the subject sees the symbol "+" at S1 and the word-pair at S2. Two Logitech C920 high-definition cameras are used to collect data from both the face of the subject

and the monitor at 58 frames-per-second (FPS). EEG and eye-tracking data were also collected during the experiments but are not used for the current study.

Data Processing and Statistics: Face videos collected were split based on the S1-S2 task for each 1500 ms trials, resulting in N=87 frames, per each video v. Each trail was diagnosed for stuttering disfluency by a certified speech pathologist and marked as either disfluent or fluent. A label-balanced dataset was created from 3700 trial studies, resulting in 1850 disfluent and 1850 fluent 1500 ms video trials. AUs were extracted using the OpenFace method described in (Baltrusaitis, Mahmoud, and Robinson 2015).

Explaining Classifier Decisions: We use DeepSHAP method (Lundberg and Lee 2017) to study the impact of each facial AU at specific time-windows towards the classification of stuttering disfluency. Here, we consider the input to the neural network as an image with each pixel corresponding to 17.24 ms of facial muscle movements. DeepSHAP generates an explanation map $e(x) \in \mathrm{R}^{17\times87}$ with positive and negative correlations of each AU towards the classifier output. We render these explanation maps on the corresponding input video frame according to FACS rules and present them in visual format, Figure 2. By averaging the explanation maps across different trials and studying the mean and standard deviations, we could start to generalize the muscle movement behaviors and the neural network learning performance across subjects with the possibility of personalized downstream models for each subject.

Experiments

Our training pipeline consists of two parts as illustrated in Figure 2: a self-supervised pre-training stage to learn meaningful AWS facial muscle movement representations and a downstream task adaptation stage to study usefulness of the learned representations for stutter disfluency classification, both from a limited amount of data. Since one of the input dimensions $x \in \mathrm{R}^{17\times87}$ is smaller than 32×32, traditional models such as ResNet do not work due to unmatched sizes. Hence, we provide custom architectures as described below.

Pretext Task Network Design: To compare learning performance and generalization of pre-training stage, we designed a customized CNN architecture (CNN-A) to account for the temporal nature of the data and compare them with two CNN architectures (CNN-B and CNN-C) designed similar to the popular VGG network corrected for number of layers to accommodate for the shape variation. Our proposed CNN-A architecture (30k parameters) consists of 4 Convolutional (Conv) layers with {16, 32, 64, 64} kernels respectively, all shaped 1×17. Average pooling is applied after the first two Conv layers. A depth-wise Conv with 128 kernels is then applied to compress the data across the AUs. Afterward, a separable Conv with 64 kernels is used to reduce the dimensions and output is flattened to an embedding of 1×256 dimensions. Two fully connected dense layers are added for the model output, one each for $\mathcal{L}_{\mathrm{win}}$ and $\mathcal{L}_{\mathrm{st}}$.

CNN-B architecture (280k parameters) consists of 3 Conv layers with {16, 32, 64} kernels respectively, all shaped 3×3. To study the impact of additional layers, CNN-C (317k parameters) has an additional Conv layer with 64 kernels of size 3×3. Batch normalization is applied after each Conv layer. Average pooling and dropout is applied on all but the first convolutional layer. Final convolutional layer output is flattened to an embedding of 1×256 dimensions similar to CNN-A with similar output layers, one each for $\mathcal{L}_{\mathrm{win}}$ and $\mathcal{L}_{\mathrm{st}}$. All CNNs were developed in the Tensorflow framework (Abadi et al. 2016).

Downstream Task Network Design: We compare the facial muscle movement representations generated by each pre-training stage network for two downstream networks: 1) three fully connected (3FC) Multi-Layer Perceptron (MLP) layers and 2) a random forest (RF) classifier with 500 trees. 3FC was trained on Tensorflow and the RF models were trained on ScikitLearn (Pedregosa et al. 2011). Training details of both tasks are detailed below.

Hyper Parameters and Training Details: For the pretext task, we trained all CNN classifiers on NVIDIA GPU systems on Jetstream Cloud (Stewart et al. 2015) with Adam optimizer at an initial learning rate of 0.01, scheduled to reduce to half the value at every 5 epochs according to multi-head validation loss (equation 3). Batch normalization, a 50% dropout of nodes, and early stopping were applied to curb overfitting. The minimum expected decrease in validation loss for early stopping criteria was in the order of 10^{-3} for every recurring 3 epochs. α was set to 0.25 of β. Since we clip 100 ms video information from the 1500 ms AU map to learn micro-expressions, we can pick 30 such windows and an extra *None* category for cases with no signal transformations. Hence, the output labels y_{win} is $\mathrm{R}^{1\times31}$. Considering different signal transformations and parameters in Table 1 as separate labels, y_{st} is $\mathrm{R}^{1\times11}$.

For the downstream task involving 3FC, learning rate schedule is done per 20 epochs and early stopping checks for a change within 50 epochs. We use binary cross-entropy as the loss term for downstream task as the final output classification is either one of fluent or disfluent trial. To train 3FC network, we froze the pretext networks and trained the fully connected layers on the embeddings of pretext network. To train the RF model, we pre-extracted the embeddings of each pretext network and trained the RF classifier on the embeddings directly. 10-fold cross validation was carried out to randomly split the data to training and validation data. A separate hold-out test set was used to evaluate the performance of each method. Each downstream method was trained for different percentages of training data to study the impact of learned pretext latent space on downstream task from a small amount of data. We provide results for 10%, 25%, 50%, 75%, and 100% of available data, all with an equal split between stutter and fluent trials. To compare learning performance of fully-supervised counterparts, we also train CNN-A, -B, and -C with 3FC from scratch.

Statistical Significance: To study the statistical significance of the experiments and the impact of upper and lower facial muscles in stuttering disfluency classification, we carry out ANOVA tests on the explanation maps of the best performing classifier. We also study the impact of individual AUs and the impact of specific time-zones (0-500 ms, 500-1000 ms, 1000-1500 ms) to understand the significance of muscle movements across time for our S1-S2 task.

Data %	CNN-A			CNN-B			CNN-C		
	AUC	F1	Acc	AUC	F1	Acc	AUC	F1	Acc
Self-Supervised CNN Pre-training + Downstream Fully Connected Network (3FC)									
100	***0.82±0.01***	0.73±0.01	***75.27±1.11***	***0.82±0.01***	***0.74±0.02***	74.83±1.15	***0.82±0.01***	*0.73±0.03*	74.53±1.61
75	**0.80±0.01**	**0.69±0.03**	**73.59±1.17**	0.79±0.01	0.68±0.03	72.17±1.31	0.79±0.02	0.68±0.04	71.51±1.60
50	0.75±0.01	**0.70±0.02**	**71.93±1.70**	0.76±0.02	0.68±0.02	69.75±1.72	**0.77±0.01**	0.68±0.02	70.64±1.21
25	0.75±0.03	0.68±0.05	**71.50±3.21**	0.78±0.02	**0.70±0.03**	71.20±2.62	**0.80±0.03**	0.69±0.02	70.80±3.12
10	0.69±0.03	0.68±0.05	**71.75±3.92**	0.74±0.03	0.67±0.04	67.00±2.84	**0.77±0.03**	**0.70±0.03**	68.50±4.89
Self-Supervised CNN Pre-training + Downstream Random Forest (RF)									
100	0.74±0.02	0.71±0.03	73.75±2.04	**0.75±0.03**	**0.73±0.03**	***75.17±2.59***	0.74±0.02	0.71±0.02	73.51±2.40
75	0.74±0.03	***0.72±0.04***	74.16±3.46	**0.75±0.03**	**0.72±0.04**	**74.56±3.42**	0.74±0.03	0.72±0.03	74.20±3.75
50	0.71±0.02	0.68±0.03	71.11±2.39	**0.75±0.03**	**0.72±0.05**	**74.72±2.96**	0.74±0.04	0.71±0.06	73.81±3.80
25	0.71±0.07	0.67±0.10	70.99±6.38	**0.75±0.04**	**0.72±0.05**	**74.86±4.79**	0.74±0.03	0.71±0.05	73.89±3.49
10	0.72±0.08	**0.69±0.08**	70.58±7.97	**0.73±0.07**	0.68±0.10	**72.92±7.92**	0.67±0.10	0.66±0.12	67.24±9.05
Fully Supervised CNN Training									
100	0.74±0.05	0.66±0.04	72.76±1.38	0.81±0.02	**0.72±0.03**	74.53±1.11	**0.81±0.04**	0.70±0.04	***74.63±1.70***
75	0.71±0.02	0.62±0.01	70.20±0.68	0.78±0.04	0.68±0.03	72.83±1.20	**0.79±0.02**	**0.70±0.03**	**72.96±1.26**
50	0.71±0.01	0.63±0.01	**70.40±0.54**	0.74±0.03	**0.67±0.03**	66.93±3.74	**0.76±0.01**	0.63±0.05	68.71±1.28
25	0.70±0.03	0.60±0.03	68.60±1.52	0.73±0.02	0.61±0.03	67.40±1.82	**0.74±0.02**	**0.63±0.03**	**69.60±2.07**
10	0.69±0.03	0.60±0.06	65.50±2.09	**0.70±0.04**	0.61±0.03	66.00±5.48	**0.70±0.04**	**0.65±0.02**	**68.00±3.26**

Table 2: Stuttering disfluency classification generalization performance for self-supervised and fully-supervised methods. Bold font represents best performers across CNN-A, -B, and -C and *Italics* across different methods.

Results and Discussions

In this study, we designed AWS speech preparation experiments and collected a real-world dataset to evaluate whether facial micro-expression can predict cognitive states and near future speech behavior. Our pretext tasks were trained on a 10% randomly picked subset of available data. Our results of different self-supervised and fully-supervised experiments on all three CNN's are reported in Table 2 evaluated for Area-Under-Curve (AUC), F1 score, and accuracy.

Self-supervised pre-training of carefully designed network with less parameters performed better than full-supervision on same amount of data. It is evident that self-supervised CNN-A (CNN-A$_{self}$) maintains higher performance than fully-supervised counterpart (CNN-A$_{full}$). CNN-A$_{self}$ also generated the best overall AUC and accuracy out of all our experiments. With only 30k parameters, this shows that depending on the task and data modality, a carefully designed small neural network could learn better latent representations with self-supervised pre-training than trained fully-supervised from scratch.

RF for downstream task gave good average performance but is highly unstable for small number of data. Our most stable results, in terms of average accuracy, is generated by CNN-B RF method. However, as we see from the standard deviations, RF method is highly unstable during training. Our 10-fold cross-validation training proved 3FC networks to be substantially stable than RF for downstream tasks.

Large number of parameters could help during self-supervised pre-training. Even though for 3FC networks CNN-C$_{self}$ under-performed than CNN-A$_{self}$ in terms of accuracy, AUC was more stable for CNN-C$_{self}$. RF methods, though unstable, generated better metrics for CNN-B. Since our task involves learning micro-expressions from a 1500ms time-window, we designed CNN-A to respect the temporal nature of expressions. A future study focused on larger CNN-A architectures could help generalize this idea.

Model Explanation Analysis (MEA): MEA was carried out using DeepSHAP method on corresponding test datasets of CNN-A$_{self}$ FC trained on 100% data and compared it with the results achieved for CNN-A$_{self}$ FC on 10% data. Results showed little variation between the two models. We present here an analysis on the CNN-A$_{self}$ FC trained on 10%. For AWS subjects with high stutter rate (SR), we found high statistical correlation of upper (F=22.09, p <0.001) and lower (F=14.84, p <0.001) facial AUs towards classifying the disfluency from pre-speech muscle movements. Both upper (F=18.58, p <0.001) and lower (F=42.83, p <0.001) facial AUs showed large significance for the first 0-500 ms of S1-S2 task. The considerable impact of AUs around the eye region was found in 0-500 ms window with large correlations of eye blinker (AU 45; F=26.34, p <0.001). Eyelid raiser (AU 7) was prominent in AWS subjects with a low stuttering rate (F=42.48, p <0.001). These findings are of high importance since AWS subjects have huge jumps in cognitive states and show numerous secondary behaviors including eye blinking before, during, or after word utterance. Here, our results suggest that AWS subjects show signs of anticipation at S1 as they are subjected to the S1-S2 task.

In the lower facial region, a considerable impact is found for muscles around the lip area including upper lip raiser (AU 10; F=45.77, p <0.001 [0-1500 ms]), lip stretcher (AU 20; F=29.81, p <0.001 [0-1500 ms]), and lip corner puller (AU 12; F=51.97, p <0.001 [0-500 ms]). Large correlations of muscle activity towards the beginning and end of the S1-S2 task suggests that secondary behaviors could be showing up as anticipation in the upper face and jitters and micro expressions in the lower facial region, especially around the lips. Furthermore, we present Figure 3 as a way of visually

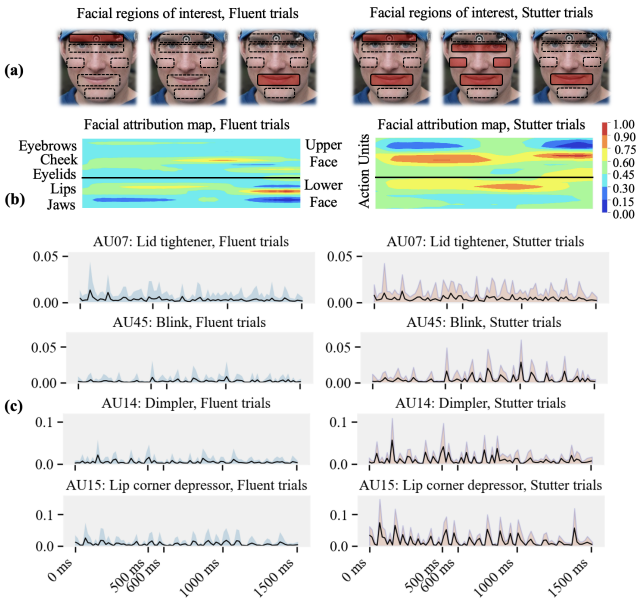

Figure 3: Facial regions of interest (a), attribution maps (b), and average muscle movement attributions (c) across AWS with high stuttering rate (SR) is visualized. Large reliance on muscles around eye and lip regions can be seen as a distinguishing factor between disfluent and fluent speech.

interpreting the large variations in the attributions of various facial muscle movement regions. We can see that individuals with AWS who stutter more tend to blink more during stutter trials after 500 ms than fluent trials. Also, we can see highly intense muscle attributions in the lower facial region. This is consistent with the behaviors visually observable in some AWS individuals as they are about to speak.

Ablation Study: To assess the impact of various parameters used for $\mathcal{L}_{\text{st}}$ during pre-training, we carry out an ablation study by reducing the range of factors used for the signal transformation. The degradation of accuracy for different data percentages is summarized in Table 3. The large decrease in accuracy, especially for a lower amount of data affirms our choice of using larger ranges of signal transformations for the pretext task. A future research opportunity might be to explore other signal transformations and ranges that would be beneficial to learn from temporal datasets.

Impact of embedding dimensions: Facial representations learned by the pretext models could be influenced by the size of the final embedding z. To study the effect of embeddings with larger dimensions, we compare the learning performance of our downstream tasks for our models with 1×256 dimensions against embeddings of size 1×970. Since output dimensions of CNN-A before embedding layer was only 1×64, we omitted CNN-A from this study. From Table 4, we can see that the 3FC models did not benefit much from the larger embedding with the exception of CNN-B performing better with larger embeddings on smaller data. In contrast, RF methods performed well with larger embeddings for smaller amount of data, especially for CNN-C RF with one extra CNN layer than CNN-B RF. The choice of downstream classifier thus depends on the dimensions of the pretext embedding, especially for RF classifiers.

Method	Data @ x (%)		
	x=100%	x=50%	x=10%
CNN-A 3FC	$\sim$ 1.9%	$\sim$ 3.0%	$\sim$ 6.9%
CNN-B 3FC	$\sim$ 3.0%	$\sim$ 6.0%	$\sim$ 6.1%
CNN-C 3FC	$\sim$ 0.8%	$\sim$ 5.0%	$\sim$ 3.5%
CNN-A RF	$\sim$ 4.0%	$\sim$ -0.8%	$\sim$ 6.8%
CNN-B RF	$\sim$ 3.2%	$\sim$ 2.8%	$\sim$ 8.2%
CNN-C RF	$\sim$ 0.6%	$\sim$ 2.1%	$\sim$ 3.2%

Table 3: The degradation in performance caused by reducing the range of factors used for the signal transformations.

Method	Data @ x (%)		
	x=100%	x=50%	x=10%
CNN-A 3FC	-	-	-
CNN-B 3FC	$\sim$ -3.7%	$\sim$ -0.9%	$\sim$ 6.9%
CNN-C 3FC	$\sim$ -4.8%	$\sim$ -4.5%	$\sim$ 0.6%
CNN-A RF	-	-	-
CNN-B RF	$\sim$ 2.1%	$\sim$ 0.6%	$\sim$ -0.5%
CNN-C RF	$\sim$ 2.9%	$\sim$ 2.7%	$\sim$ 11.0%

Table 4: The performance changes caused by increasing the embedding dimensions from 256 to 970 in the pretext task.

Conclusion

In this paper, we presented an interpretable self-supervision methodology to learn meaningful facial muscle movement representations to investigate the possibility of using these learned representations to classify neurological disorders. We successfully showcased that the face representations could be used to detect future onset of stuttering disfluency. We found that self-supervised pre-training of our carefully designed neural network performed better than full-supervision on same amount of data. Fully connected networks were more stable than random forest classifiers for the downstream stuttering classification task. Our model explainability analysis found a considerable influence of muscles around the eye (eye blinker AU 45, eyelid raiser AU 7) and lips regions (lip corner puller AU 12, lip stretcher AU 20) towards identifying stuttering disfluency. Large correlations of muscle activity towards the beginning and end of each trial suggest anticipation, possible secondary behaviors, and micro expressions throughout the trial. Large variance and amplitude of disfluent trials compared to fluent trials suggests the large facial activity of AWS subjects which algorithms can learn to classify.

Although it is hard to generalize with smaller populations of data, our study reveals an automated way to streamline the process of marking pre-speech facial muscle movement information of a future vocalization as fluent or disfluent. Study should be expanded on larger populations of AWS to generalize the claim. Also, future studies could explore task personalization to learn facial muscle movement patterns of individual AWS subjects.

Acknowledgements

This work was partly supported by the Open Cloud Institute (OCI) at University of Texas at San Antonio (UTSA) and partly by the National Institutes of Health (NIH) under Grant DC016353. The authors gratefully acknowledge the use of the services of Jetstream cloud.

References

Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M.; et al. 2016. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)*, 265–283.

Baltrusaitis, T.; Mahmoud, M.; and Robinson, P. 2015. Cross-dataset learning and person-specific normalisation for automatic Action Unit detection. In *2015 11th IEEE Int. Conf. Work. Autom. Face Gesture Recognit.*, 1–6. IEEE. ISBN 978-1-4799-6026-2.

Bandini, A.; Orlandi, S.; Escalante, H. J.; Giovannelli, F.; Cincotta, M.; Reyes-Garcia, C. A.; Vanni, P.; Zaccara, G.; and Manfredi, C. 2017. Analysis of facial expressions in parkinson's disease through video-based automatic methods. *Journal of neuroscience methods* 281: 7–20.

Bandini, A.; Orlandi, S.; Giovannelli, F.; Felici, A.; Cincotta, M.; Clemente, D.; Vanni, P.; Zaccara, G.; and Manfredi, C. 2016. Markerless analysis of articulatory movements in patients with Parkinson's disease. *Journal of Voice* 30(6): 766–e1.

Bloodstein, O.; and Ratner, N. B. 2008. *A handbook on stuttering.* Clifton Park (N.Y.): Thomson/Delmar Learning.

Cao, B.; Zhang, H.; Wang, N.; Gao, X.; and Shen, D. 2020. Auto-GAN: Self-Supervised Collaborative Learning for Medical Image Synthesis. In *AAAI*, 10486–10493.

Craig, A.; and Tran, Y. 2005. The epidemiology of stuttering: The need for reliable estimates of prevalence and anxiety levels over the lifespan. *Advances in Speech Language Pathology* 7(1): 41–46.

Das, A.; and Rad, P. 2020. Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey. *arXiv preprint arXiv:2006.11371* .

Das, A.; Rad, P.; Choo, K.-K. R.; Nouhi, B.; Lish, J.; and Martel, J. 2019. Distributed machine learning cloud teleophthalmology IoT for predicting AMD disease progression. *Future Generation Computer Systems* 93: 486–498.

Ekman, P. 1992. Facial expressions of emotion: an old controversy and new findings. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 335(1273): 63–69.

Ekman, P. 2009. *Telling lies: Clues to deceit in the marketplace, politics, and marriage (revised edition)*. WW Norton & Company.

Ekman, P.; Levenson, R. W.; and Friesen, W. V. 1983. Autonomic nervous system activity distinguishes among emotions. *science* 221(4616): 1208–1210.

Ekman, P.; and Rosenberg, E. L. 2005. *What the Face Reveals Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*. Oxford University Press.

Friesen, E.; and Ekman, P. 1978. Facial action coding system: a technique for the measurement of facial movement. *Palo Alto* 3.

Gulshan, V.; Peng, L.; Coram, M.; Stumpe, M. C.; Wu, D.; Narayanaswamy, A.; Venugopalan, S.; Widner, K.; Madams, T.; Cuadros, J.; et al. 2016. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama* 316(22): 2402–2410.

Hamm, J.; Kohler, C. G.; Gur, R. C.; and Verma, R. 2011. Automated Facial Action Coding System for dynamic analysis of facial expressions in neuropsychiatric disorders. *Journal of Neuroscience Methods* 200(2): 237–256.

Hoy, K. E.; Fitzgerald, P. B.; Bradshaw, J. L.; Armatas, C. A.; and Georgiou-Karistianis, N. 2004. Investigating the cortical origins of motor overflow. *Brain Res. Rev.* 46(3): 315–327. ISSN 01650173. doi:10.1016/j.brainresrev.2004.07.013.

Jiang, C.; Wu, J.; Zhong, W.; Wei, M.; Tong, J.; Yu, H.; and Wang, L. 2020. Automatic Facial Paralysis Assessment via Computational Image Analysis. *Journal of Healthcare Engineering* 2020.

Kunz, M.; Meixner, D.; and Lautenbacher, S. 2019. Facial muscle movements encoding pain—a systematic review. *Pain* 160(3): 535–549.

Li, X.; Jia, M.; Islam, M. T.; Yu, L.; and Xing, L. 2020. Self-supervised Feature Learning via Exploiting Multi-modal Data for Retinal Disease Diagnosis. *IEEE Transactions on Medical Imaging* .

Li, X.; Liu, S.; De Mello, S.; Wang, X.; Kautz, J.; and Yang, M.-H. 2019. Joint-task self-supervised learning for temporal correspondence. In *Advances in Neural Information Processing Systems*, 318–328.

Lints-Martindale, A. C.; Hadjistavropoulos, T.; Barber, B.; and Gibson, S. J. 2007. A Psychophysical Investigation of the Facial Action Coding System as an Index of Pain Variability among Older Adults with and without Alzheimer's Disease. *Pain Med.* 8(8): 678–689.

Lundberg, S. M.; and Lee, S.-I. 2017. A unified approach to interpreting model predictions. In *Advances in neural information processing systems*, 4765–4774.

Mehu, M.; Mortillaro, M.; Bänziger, T.; and Scherer, K. R. 2012. Reliable facial muscle activation enhances recognizability and credibility of emotional expression. *Emotion* 12(4): 701–715.

Meng, Z.; Han, S.; and Tong, Y. 2017. Listen to your face: Inferring facial action units from audio channel. *IEEE Transactions on Affective Computing* .

Myers, J.; Irani, F.; Golob, E.; Mock, J.; and Robbins, K. 2019. Single-Trial Classification of Disfluent Brain States in Adults Who Stutter. *Proc. - 2018 IEEE Int. Conf. Syst. Man, Cybern. SMC 2018* 57–62.

Oh, Y.-H.; See, J.; Le Ngo, A. C.; Phan, R. C.-W.; and Baskaran, V. M. 2018. A survey of automatic facial micro-expression analysis: databases, methods, and challenges. *Frontiers in psychology* 9: 1128.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. 2011. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* 12: 2825–2830.

Penfield, W.; and Boldrey, E. 1937. Somatic motor and sensory representation in the cerebral cortex of man as studied by electrical stimulation. *Brain* 60(4): 389–443.

Prasse, J. E.; and Kikano, G. E. 2008. Stuttering: an overview. *American family physician* 77(9): 1271–1276.

Proctor, A.; Duff, M.; and Yairi, E. 2002. Early childhood stuttering: African Americans and European Americans. *Journal of Speech, Language, and Hearing Research* 4(15): 102.

Rivail, A.; Schmidt-Erfurth, U.; Vogel, W.-D.; Waldstein, S. M.; Riedl, S.; Grechenig, C.; Wu, Z.; and Bogunovic, H. 2019. Modeling disease progression in retinal OCTs with longitudinal self-supervised learning. In *International on Predictive Intelligence In Medicine*, 44–52. Springer.

Sayette, M. A.; Cohn, J. F.; Wertz, J. M.; Perrott, M. A.; and Parrot, D. J. 2002. A Psychometric Evaluation of the Facial Action Coding System for Assessing Spontaneous Expression Michael A. Sayette, Jeffrey F. Cohn, Joan M. Wertz, Michael A. Perrott, and Dominic J. Parrott University of Pittsburgh. *Journal of Nonverbal Behavior* .

Sheehan, J. G. 1974. Stuttering behavior: A phonetic analysis. *Journal of Communication Disorders* 7(3): 193–212.

Sheng, K.; Dong, W.; Chai, M.; Wang, G.; Zhou, P.; Huang, F.; Hu, B.-G.; Ji, R.; and Ma, C. 2020. Revisiting Image Aesthetic Assessment via Self-Supervised Feature Learning. In *AAAI*, 5709–5716.

Shukla, A.; Vougioukas, K.; Ma, P.; Petridis, S.; and Pantic, M. 2020. Visually guided self supervised learning of speech representations. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6299–6303. IEEE.

Stewart, C. A.; Cockerill, T. M.; Foster, I.; Hancock, D.; Merchant, N.; Skidmore, E.; Stanzione, D.; Taylor, J.; Tuecke, S.; Turner, G.; et al. 2015. Jetstream: a self-provisioned, scalable science and engineering cloud environment. In *Proceedings of the 2015 XSEDE Conference: Scientific Advancements Enabled by Enhanced Cyberinfrastructure*, 1–8.

Tschannen, M.; Djolonga, J.; Ritter, M.; Mahendran, A.; Houlsby, N.; Gelly, S.; and Lucic, M. 2020. Self-Supervised Learning of Video-Induced Visual Invariances. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13806–13815.

Verma, M.; Vipparthi, S. K.; Singh, G.; and Murala, S. 2019. LEARNet: Dynamic imaging network for micro expression recognition. *IEEE Transactions on Image Processing* 29: 1618–1627.

Villegas, B.; Flores, K. M.; Jose Acuna, K.; Pacheco-Barrios, K.; and Elias, D. 2019. A Novel Stuttering Disfluency Classification System Based on Respiratory Biosignals. In *2019 41st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 4660–4663. IEEE.

Wang, S.; Che, W.; Liu, Q.; Qin, P.; Liu, T.; and Wang, W. Y. 2019. Multi-task self-supervised learning for disfluency detection. *arXiv preprint arXiv:1908.05378* .

Wang, Z.; Wang, S.; and Ji, Q. 2013. Capturing Complex Spatio-temporal Relations among Facial Muscles for Facial Expression Recognition. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, 3422–3429. IEEE.

Wehrle, T.; Kaiser, S.; Schmidt, S.; and Scherer, K. R. 2000. Studying the dynamics of emotional expression using synthesized facial muscle movements. *Journal of Personality and Social Psychology* 78(1): 105–119.

Wu, P.; Gonzalez, I.; Patsis, G.; Jiang, D.; Sahli, H.; Kerckhofs, E.; and Vandekerckhove, M. 2014. Objectifying facial expressivity assessment of Parkinson's patients: preliminary study. *Computational and mathematical methods in medicine* 2014.

Yairi, E.; and Ambrose, N. G. 1999. Early childhood stuttering I: Persistency and recovery rates. *Journal of Speech, Language, and Hearing Research* 42(5): 1097–1112.

Zhang, J.; Dong, B.; and Yan, Y. 2013. A computer-assist algorithm to detect repetitive stuttering automatically. *Proc. - 2013 Int. Conf. Asian Lang. Process. IALP 2013* 249–252.