Transfer Learning based Disfluency Detection using Stuttered Speech

Sparsh Garg[1], *Utkarsh Mehrotra*[1], *Gurugubelli Krishna*[1], *Anil Kumar Vuppala*[1]

[1]Speech Processing Laboratory, International Institute of Information Technology, Hyderabad, India

sparsh.garg@research.iiit.ac.in, utkarsh.mehrotra@research.iiit.ac.in, krishna.gurugubelli@research.iiit.ac.in, anil.vuppala@iiit.ac.in

Abstract

Spontaneous speech is characterized by the presence of hesitations, which result in the breaking of normal speech flow. These hesitations are referred to as speech disfluencies. Disfluencies can provide information regarding the speaking style, speaker identity and language fluency, which can be useful for several speech-based applications. For automatic speech recognition (ASR) systems, the presence of these disfluencies leads to a higher word error rate, since most ASR systems are developed on non-spontaneous read speech data. Thus, the detection of disfluencies in spontaneous speech becomes an essential task for many applications. This paper presents a transfer learning approach to detect disfluencies in spontaneous lecture mode speech using two frame-level automatic disfluency detection systems trained on stuttered speech. The model is tested for four types of disfluencies - filled pause, prolongation, part-word repetition and word repetition. We obtain an accuracy of 92.73% for filled pause and 83.43% for part-word repetition detection using MFCC features as input to the Deep Neural Network (DNN) based disfluency detection model. Overall, the transfer learning method gives an average improvement of 2.69% and 2.00% in detection accuracy using DNN based and Bidirectional LSTM (BiLSTM) based model, respectively, across all disfluencies over the baseline results on the IIITH-IED dataset.

Index Terms: disfluency detection, transfer learning, stuttered speech

1. Introduction

Spontaneous speech is a particular type of speech setting where a speaker speaks without preparing in advance. This makes the speaker think about what to say on the spot, formulate the utterances and then produce the speech. Such a setting often leads to abrupt breaks or discontinuities in the normal conversation flow because of the following reasons - language complexity, time taken by the speaker to decide what to say, nervousness while speaking, etc. [1,2]. These discontinuities or breaks are known as speech disfluencies. Knowledge about the presence of disfluencies and their duration can be useful for many speech-based applications. For the task of language learning, the number of disfluencies produced by the speaker and the related duration of each disfluency can help in evaluating language proficiency [3]. The types and frequency of disfluencies produced by a speaker can be used as features for speaker recognition and language identification systems. For ASR systems, the presence of disfluencies in the speech signal can adversely affect systems' performance since most ASRs are built on read speech. Hence, the detection of speech disfluencies becomes important to enhance the performance of a lot of speech-based systems.

The type of speech disfluency produced varies depending on the type of discontinuity, interrupting the normal speech flow. Some of the most common types of speech disfluencies are described below with examples :

1. **Filled Pause** - Pauses in speech which have a filler word in them like 'um', 'uh', etc. The filler word does not provide additional meaning to the utterance. example - I am going to uh Delhi.
2. **Prolongation** - The lengthening of a particular word or part of a word, which produces a discontinuity in normal speech flow. example - Whoooose pen is it ?
3. **Repetition** - This type of disfluency occurs when the speaker repeats a part of the utterance, disrupting the flow speech flow. On the basis on the repeated unit, repetition is further classified into three types. Examples of each repetition type are given below.
   (a) Part-word repetition - Wh-what is your name ?
   (b) Word repetition - Please pass me the the book.
   (c) Phrase repetition - I like I like ice cream.

The detection of speech disfluencies has been explored extensively in the literature [4–6]. In general, most of the detection methods belong to one of the following categories - as a post-processing step after the ASR output by using text-based features along with speech [7, 8] or as a pre-processing step before giving input to a speech system by using signal level processing [9, 10]. Though the ASR-based approaches are effective and give encouraging results, they depend on how accurate the ASR is. In this work, we use only speech-based features for the disfluency detection problem. Early works on disfluency detection focused on identifying cues in speech signals such as formants, vowel-lengthening, and nasality for individual disfluencies [11–13]. In [5], Conditional Random Field (CRF) and Hidden Markov Model (HMM) classifiers were used. Recurrent Neural Networks (RNN), with their ability to capture temporal dependencies efficiently, have been used in recent works with considerable success for disfluency detection task [7, 14–16].

Stuttered speech is another primary source of disfluencies. The disfluencies in stuttered speech are similar to those present in conversational speech, but their frequency of occurrence is higher. This serves as the motivation to use transfer learning in the current work - using stuttered speech data, we try to detect disfluencies in conversational, lecture-mode speech. The contributions of this work are listed below :

- Exploration of the use of stuttered speech data for detection of the disfluencies using transfer learning. To the best of our knowledge, this has not been explored in the literature yet.
- Evaluating the Transfer learning based disfluency detection method for 4 types of disfluencies.

The rest of the paper is organised as follows - in Section 2, we give a brief overview of stuttered speech and its relation to disfluencies in conversational speech. The experimental setup used in this study is described in Section 3. Experimental results and discussion are presented in Section 4. Finally, the conclusion and future scope of this work are provided in Section 5.

2. Disfluencies: Stuttering – Spontaneous Speech

Stuttering is a speech disorder where hesitations that break the flow of speech occur involuntarily [17]. Audible or inaudible prolongation of words, excessive use of fillers (like 'um', 'uh', etc.) and uncontrolled repetitions are some of the main characteristics of stuttered speech. In humans, both physiological and neurogenic causes can lead to stuttering, with some of the reasons being - increased dopamine levels in muscles causing inhibitory effects, problems with auditory processing and motor disorders pertaining to the basal ganglia [18]. From the perspective of speech, stuttering is sub-divided into various forms, depending on the type of hesitation, which led to the stutter. Some of the different forms of stutter are - interjections (fillers), prolongations, repetitions and revisions.

On exploring the literature pertaining to stuttered speech and disfluencies in spontaneous speech, we found that the acoustic and linguistic basis by which different forms of stutters are categorised is very similar to the categorisation of disfluencies. Even the set of audio features used for automatic stutter classification and disfluency detection are also overlapping. This observation has led to the hypothesis used in our work that a classifier trained on stuttered data can help in detecting disfluencies in conversational speech. Hence, a transfer learning approach to detect disfluencies was explored.

Various works have been done on automatic stutter classification. Early works focused on extracting acoustic and signal level features from input audio combined with classical machine learning methods such as GMM, LDA, k-NN, etc. [19–21]. In [19], Mel Frequency Cepstral Coefficients (MFCC) features were used with LDA and k-NN as classifiers to detect repetition and prolongations in stuttered speech. In [20], the same authors used LPCC based features with LDA and k-NN to detect repetitions and prolongations. But LPCC takes an average of 3.73 sec more than MFCC for giving a decision. To better understand the human perception of speech, both spectral and temporal characteristics of the signal are important. With MFCC, short-term spectral features of speech are captured effectively but the information about temporal behaviour is not perceived considerably. So, to capture the temporal, instantaneous amplitude and frequency characteristics of signals, in [22], LP-Hilbert Envelope Based MFCC features were used for the detection of prolongations, repetitions and interjections.

In recent studies, deep learning architectures are being used with both text and signal level features for the detection of stuttering events [23–25]. In [23], a lightly supervised approach was used with task-oriented lattices to recognise stuttering events in children's speech and provide a complete verbatim output of stuttered speech to help diagnose the disorder. In [25], the Sequence labelling approach was employed with Conditional Random Fields (CRF) and BiLSTM for the detection of stuttering events in both manual and automatically generated transcripts (by ASR). In [26], spectrogram features were used with a Deep residual network and BiLSTM to classify a 4-second stutter file into one of the six types of major stuttered disfluencies.

Some recent works have also utilised non-speech related features to detect stuttering events [27, 28]. In [27], based on respiratory biosignal activity, stuttering events were classified into blocks and non-block states of speech using Multi-Layer Perceptron (MLP). In [28], with Artificial Intelligence (AI) aided Convolutional Neural Network (CNN) and facial movement patterns, expected speech is classified as fluent or stuttered.

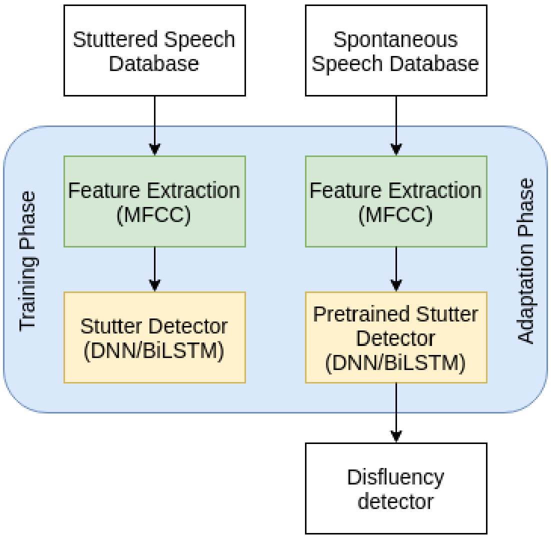
Stuttered Speech Database
Spontaneous Speech Database
Training Phase
Feature Extraction (MFCC)
Feature Extraction (MFCC)
Adaptation Phase
Stutter Detector (DNN/BiLSTM)
Pretrained Stutter Detector (DNN/BiLSTM)
Disfluency detector

Figure 1: *Transfer Learning based disfluency detection pipeline*

3. Experimental Setup

In order to test our hypothesis, stutter data from the UCLASS corpus is used in experiments here. A disfluency detection model is trained on data from the UCLASS corpus, and the trained model is then used to detect related disfluencies in the IIITH-IED dataset. Figure 1 shows the disfluency detection system using transfer learning used here. Details about the UCLASS and IIITH-IED datasets are presented in the next subsections, followed by a description of the BiLSTM and DNN based frame-level disfluency detection models.

3.1. UCLASS Dataset

The University College London's Archive of Stuttered Speech (UCLASS) Dataset is one of the most popular resources for studies on stuttered speech. It was introduced in [29]. The dataset consists of stuttered speech recordings and corresponding annotations in British English. The UCLASS dataset has two main releases. Here we have used speech recordings from Release One of the UCLASS dataset to prepare the pre-trained models for transfer learning experiments. This dataset consists of monologue speech recordings from children of age 8 to 18 years, who were diagnosed with stuttering disorder of varying severity. Out of the 139 recordings available, 25 were used because of the availability of corresponding transcriptions. The transcriptions were forced aligned with the audio to generate timestamps for each word. Each recording was then annotated for 7 types of stutter disfluencies - filled pause, prolongation, sound repetition, part-word repetition, word repetition, phrase repetition and revision, as done in [26]. The annotation was carried out similar to [30].

3.2. IIITH-IED Dataset

The IIITH-IED dataset was introduced in here[1]. Ten hours of lecture-mode speech in Indian English were transcribed to prepare this dataset. Speech recordings from the freely available lectures under the NPTEL initiative of the Government of India were used to make this dataset. Since lecture-mode speech is prepared, there are instances where the lecturer has to explain a topic spontaneously, this type of speech is categorized as semi-spontaneous. The IIITH-IED dataset consists of speech from 60 speakers - 30 male and 30 female. A 10-minute recording of a lecture from each speaker is annotated for both words as well as disfluencies manually. Each speech recording from a speaker is further segmented into speech files of length 8 to 12 seconds, with a sampling rate of 16000 Hz. After segmentation, annotation is also performed at the signal level to identify the starting and ending time of disfluencies present in each segmented file. The number of occurrences of each disfluency type used from this dataset in our experiments is shown in Table 1. More details about the dataset can be found here [2].

Table 1: *Number of occurrences of each disfluency type in IIITH-IED Dataset*

Disfluency Type	**# of Occurences**
Filled Pause	1428
Prolongation	71
Part-word Repetition	164
Word Repetition	211

3.3. Detection Model

In order to test the transfer learning hypothesis, frame-level automatic disfluency detection systems are used. These systems are used to detect whether or not a particular disfluency type is present in a speech frame of 10 ms. The disfluencies considered for the experiments here are - filled pause, prolongation, part-word repetition and word repetition. For every disfluency type, detection was set up as a binary classification problem - a speech frame either belongs to the disfluency type or it does not.

The features used for the task of disfluency detection are MFCC features. The MFCC features consisted of the following - the first 13 cepstral coefficients, the 0th cepstral coefficient and the energy of the frame. Windowed speech frames having a length of 25 ms, with a 10 ms frame shift are used for MFCC feature extraction here. The delta and delta-delta MFCC coefficients are also computed and used. The size of the feature vector obtained then was 45-dimensional per frame.

To produce better disfluency detection results, stacking up features from neighbouring frames using a context window is proven beneficial in [31,32]. So, window lengths of ±1, ±2 and ±3 frames were used to experiment with the MFCC features extracted above. The dimensions of the final feature vectors for each frame obtained using ±1, ±2 and ±3 frames window lengths were 135, 225 and 315, respectively. As best classification results were obtained by stacking features from 3 frames before and after each individual frame, final disfluency detection results are reported using this configuration only.

Two disfluency detection systems are then trained using the MFCC features extracted. The first system is a DNN-based detector. The network used here has 2 hidden layers. The number of hidden units in the layers are 50 and 100 respectively. The next system is a BiLSTM-based detector. Two bidirectional recurrent layers, each having 7 units are then used to learn temporal dependency between features and then classify whether the disfluency type is present or not. Dropout rates for both the recurrent layers are set at 0.2 and 0.4 to avoid overfitting in the BiLSTM.

4. Experiments and Results

Disfluency detection for each of the four disfluencies was setup as a binary classification task. The baseline disfluency detection systems were developed here on the IIITH-IED dataset using the DNN and BiLSTM architectures defined in the previous section. Four binary classifiers were trained for each of the architectures to detect the four disfluencies. The baseline detection accuracy and F1-score obtained on the UCLASS Dataset are shown in Table 3 and the results for the IIITH-IED dataset for the four disfluencies are shown in Table 4. For both DNN and BiLSTM classifiers, a learning rate of 10^{-3} was used for training, with the binary cross-entropy loss function and RMSprop optimizer. The number of training epochs used for DNN were 50, while for the BiLSTM-classifiers, the number of training epochs used were 10. For every 10 ms speech frame, the MFCC features extracted for that frame were used as input to the models, which then predicted whether that speech frame belongs to the disfluency type or not. Since IIITH-IED is a biased dataset in terms of the number of disfluencies present for each class, so for evaluating the model more effectively, *stratified K-fold cross-validation* was used so that the ratio of samples from each class is the same in train and test sets. Here, the value of K was set to 10, and 9-folds were used for training, and 1 fold was used for testing the model.

Table 2: *Cosine Similarity between a stutter type and the closest related disfluency*

Disfluency Type	**Stutter type**	**Cosine Similarity**
Filled Pause	Filler	4.23e-2
Prolongations	Prolongation	3.19e-2
Part-word Repetitions	Sound Repetition	5.76e-3
Word Repetitions	Word Repetition	8.55e-4

Further, the proposed transfer learning based disfluency detection systems were trained using the UCLASS corpus and the IIITH-IED dataset. Transfer Learning refers to the learning in a target domain by transferring knowledge from another related task [33]. In this approach, we trained the disfluency detection models to detect a particular type of disfluency using the UCLASS corpus and then validated this model for the related disfluency in the IIITH-IED dataset. In order to find how closely related the occurrences of each disfluency type are in the UCLASS and IIITH-IED datasets, the cosine similarity metric was used. Table 2 shows the cosine similarity values obtained for each of the pairs. The cosine similarity is calculated by first taking the dot product (similarity measure) between the frame-level features for each pair of frames. The final value is obtained by averaging across all frames. As can be seen from the table, a close correspondence is found between disfluencies in the two

[1] https://www.dropbox.com/s/hwflfjnrtnnya1h/NCC_Paper.pdf

[2] https://bit.ly/3fAc3mb

Table 3: *Baseline disfluency detection results for the four types of disfluencies in the UCLASS Dataset. Here F1 refers to the F1-score.*

	DNN		BiLSTM	
Disfluency Type	**Accuracy**	**F1**	**Accuracy**	**F1**
Filler	91.37	0.912	92.54	0.924
Prolongations	89.91	0.900	91.82	0.919
Sound Repetitions	84.70	0.846	80.93	0.809
Word Repetitions	85.21	0.851	83.41	0.832

Table 4: *Baseline disfluency detection results for the four types of disfluencies in the IIITH-IED Dataset. Here F1 refers to the F1-score.*

	DNN		BiLSTM	
Disfluency Type	**Accuracy**	**F1**	**Accuracy**	**F1**
Filled Pause	89.92	0.892	90.17	0.891
Prolongations	88.26	0.887	91.07	0.911
Part-word Repetitions	82.47	0.817	78.72	0.778
Word Repetitions	80.73	0.805	77.97	0.774

Table 5: *Detection results obtained using the proposed transfer learning approach. Here, Acc. refers to Accuracy and F1 refers to the F1-score.*

	DNN		BiLSTM	
Disfluency Type	**Acc.**	**F1**	**Acc.**	**F1**
Filled Pause	92.73	0.927	91.60	0.910
Prolongations	94.90	0.949	94.01	0.941
Part-word Repetitions	83.43	0.826	80.45	0.799
Word Repetitions	81.16	0.807	79.87	0.795

datasets, indicating that using stuttered speech data might help in the task of disfluency detection.

While testing, the pre-trained models developed on the UCLASS dataset are evaluated on the IIITH-IED dataset using stratified 10-fold cross-validation so that the network has some amount of learning experience on our data as well, and uniformity is maintained. The learning rate, loss function and the optimizer used in the experiments are the same as in the baseline experiments. A batch size of 32 was taken while training. The performance of the proposed transfer learning based detection systems for all the disfluency types is shown in Table 5. The average accuracy and average F1-score obtained across all folds for each disfluency were used as the metrics to compare the performance.

As can be seen from Tables 4 and 5, using the proposed transfer learning approach, an increase in detection accuracy and F1-score was obtained for all four types of disfluencies, with both the classification methods i.e. DNN based and BiLSTM based. From Table 3 and Table 5, we an see clearly that the knowledge is being transferred from the stutter domain to spontaneous speech disfluency domain, as should be the case in transfer learning. Especially for filled pause and prolongation disfluencies, the increase in performance is significant. In the filled pause case, the detection accuracy obtained using transfer learning was 92.73% using the DNN classifier, with the F1-score being 0.927. An absolute increase of 2.79% is obtained for the DNN classifier using the transfer learning approach compared to the baseline result. This increase can be attributed to the fact that two forms of filled pause are most common in the UCLASS and IIITH-IED datasets - ‘um’ and ‘uh’. The increase in the number of samples of these two types of filled pause in the training set leads to an increase in performance. The highest detection F1-score is obtained for prolongations using the transfer learning method, with 0.949 being the F1-score using the DNN classifier. The absolute increase in detection accuracy is also the highest for prolongation when compared to the baseline results. This is because the majority of occurrences of prolongations in both datasets correspond to the lengthening of vowels (especially vowels /o/ and /a/). Also, the number of occurrences of prolongation in the UCLASS dataset are a lot more than the IIITH-IED dataset, which leads to significant improvements using the transfer learning approach.

In the case of part-word and word repetitions, the improvements obtained using the transfer learning setup are much less than filled pause and prolongation. This is because the occurrences of these two types of disfluencies can take up many forms, leading to high intra-class variance. This makes it difficult to model the samples belonging to part-word repetition and word repetition using the transfer learning setup. Hence, marginal improvements of 1.1% and 0.3% are obtained in the detection performance for part-word repetition and word repetition, respectively, using the DNN-based disfluency detection system.

5. Conclusion

In this work, we proposed a transfer learning approach to detect disfluencies in conversational speech using a model trained on stutter data. Disfluency detection is done for four types of disfluencies as a binary classification task. Two types of model i.e. DNN based and BiLSTM based were trained for classification of a particular type of stutter using MFCC features as input. These trained models were then tested for detecting the disfluency, which is most similar to that stutter-type. Using this approach, we obtained an average relative improvement of 2.69% and 2.00% in detection accuracy across all four disfluencies over the baseline experiment (when only the samples from the IIITH-IED dataset were taken) using DNN and BiLSTM based models, respectively.

Our future works will be aimed at developing a single classification model for multiple disfluencies. We also plan to incorporate text-based features into this pipeline so as to study the effect of combination of text and speech features on this disfluency detection architecture.

6. References

[1] Brown, Aaron, and David A. Patterson. ”To err is human.” Proceedings of the First Workshop on evaluating and architecting system dependability (EASY'01). 2001.

[2] Corley, Martin, and Oliver W. Stewart. ”Hesitation disfluencies in spontaneous speech: The meaning of um.” Language and Linguistics Compass 2.4 (2008): 589-602.

[3] Lu, Yiting, Mark JF Gales, Katherine Knill, Potsawee Manakul, and Yu Wang. ”Disfluency Detection for Spoken Learner English.” In SLaTE, pp. 74-78. 2019.

[4] Lickley, Robin J. ”Detecting disfluency in spontaneous speech.” PhD diss., University of Edinburgh, 1994.

[5] Liu, Yang, Elizabeth Shriberg, Andreas Stolcke, Dustin Hillard, Mari Ostendorf, and Mary Harper. ”Enriching speech recognition with automatic detection of sentence boundaries and disfluencies.” IEEE Transactions on audio, speech, and language processing 14, no. 5 (2006): 1526-1540.

[6] Lou, Paria Jamshid, and Mark Johnson. ”End-to-End Speech Recognition and Disfluency Removal.” arXiv preprint arXiv:2009.10298 (2020).

[7] Zayats, Vicky, Mari Ostendorf, and Hannaneh Hajishirzi. ”Disfluency detection using a bidirectional lstm.” arXiv preprint arXiv:1604.03209 (2016).

[8] Lu, Yiting, Mark JF Gales, Katherine Knill, Potsawee Manakul, and Yu Wang. ”Disfluency Detection for Spoken Learner English.” In SLaTE, pp. 74-78. 2019.

[9] Salesky, Elizabeth, Matthias Sperber, and Alex Waibel. ”Fluent translations from disfluent speech in end-to-end speech translation.” arXiv preprint arXiv:1906.00556 (2019).

[10] Hamzah, Raseeda, and Nursuriati Jamil. ”Investigation of Speech Disfluencies Classification on Different Threshold Selection Techniques Using Energy Feature Extraction.” Malaysian Journal of Computing 4.1 (2019): 178-192.

[11] Wu, Chung-Hsien, and Gwo-Lang Yan. ”Acoustic feature analysis and discriminative modeling of filled pauses for spontaneous speech recognition.” Real World Speech Processing. Springer, Boston, MA, 2004. 17-30.

[12] Audhkhasi, Kartik, Kundan Kandhway, Om D. Deshmukh, and Ashish Verma. ”Formant-based technique for automatic filled-pause detection in spontaneous spoken English.” In 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 4857-4860. IEEE, 2009.

[13] Kaushik, Mayank, Matthew Trinkle, and Ahmad Hashemi-Sakhtsari. ”Automatic detection and removal of disfluencies from spontaneous speech.” Proceedings of the Australasian International Conference on Speech Science and Technology (SST). Vol. 70. 2010.

[14] Lou, Paria Jamshid, and Mark Johnson. ”Disfluency detection using a noisy channel model and a deep neural language model.” arXiv preprint arXiv:1808.09091 (2018).

[15] Dong, Qianqian, Feng Wang, Zhen Yang, Wei Chen, Shuang Xu, and Bo Xu. ”Adapting translation models for transcript disfluency detection.” In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, no. 01, pp. 6351-6358. 2019.

[16] Wang, Shaolei, Zhongyuan Wang, Wanxiang Che, and Ting Liu. ”Combining Self-Training and Self-Supervised Learning for Unsupervised Disfluency Detection.” arXiv preprint arXiv:2010.15360 (2020).

[17] Guitar, Barry. Stuttering: An integrated approach to its nature and treatment. Lippincott Williams & Wilkins, 2013.

[18] Büchel, Christian, and Martin Sommer. ”What causes stuttering?.” PLoS Biol 2.2 (2004): e46.

[19] Chee, Lim Sin, Ooi Chia Ai, M. Hariharan, and Sazali Yaacob. ”MFCC based recognition of repetitions and prolongations in stuttered speech using k-NN and LDA.” In 2009 IEEE Student Conference on Research and Development (SCOReD), pp. 146-149. IEEE, 2009.

[20] Chee, Lim Sin, Ooi Chia Ai, M. Hariharan, and Sazali Yaacob. ”Automatic detection of prolongations and repetitions using LPCC.” In 2009 international conference for technical postgraduates (TECHPOS), pp. 1-4. IEEE, 2009.

[21] Wiśniewski, Marek, Wiesława Kuniszyk-Jóźkowiak, Elżbieta Smołka, and Waldemar Suszyński. ”Automatic detection of prolonged fricative phonemes with the hidden Markov models approach.” Journal of Medical Informatics & Technologies 11 (2007).

[22] Mahesha, P., and D. S. Vinod. ”Lp-hillbert transform based mfcc for effective discrimination of stuttering dysfluencies.” In 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), pp. 2561-2565. IEEE, 2017.

[23] Alharbi, Sadeen, Madina Hasan, Anthony JH Simons, Shelagh Brumfitt, and Phil Green. ”A lightly supervised approach to detect stuttering in children's speech.” In Proceedings of Interspeech 2018, pp. 3433-3437. ISCA, 2018.

[24] Santoso, Jennifer, Takeshi Yamada, and Shoji Makino. ”Classification of causes of speech recognition errors using attention-based bidirectional long short-term memory and modulation spectrum.” 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). IEEE, 2019.

[25] Zayats, Vicky, Mari Ostendorf, and Hannaneh Hajishirzi. ”Disfluency detection using a bidirectional lstm.” arXiv preprint arXiv:1604.03209 (2016).

[26] Kourkounakis, Tedd, Amirhossein Hajavi, and Ali Etemad. ”Detecting Multiple Speech Disfluencies Using a Deep Residual Network with Bidirectional Long Short-Term Memory.” ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020.

[27] Villegas, Bruno, Kevin M. Flores, Kevin José Acuña, Kevin Pacheco-Barrios, and Dante Elias. ”A novel stuttering disfluency classification system based on respiratory biosignals.” In 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 4660-4663. IEEE, 2019.

[28] Das, Arun, Jeffrey Mock, Henry Chacon, Farzan Irani, Edward Golob, and Peyman Najafirad. ”Stuttering Speech Disfluency Prediction using Explainable Attribution Vectors of Facial Muscle Movements.” arXiv preprint arXiv:2010.01231 (2020).

[29] Howell, Peter, Stephen Davis, and Jon Bartrip. ”The university college london archive of stuttered speech (uclass).” (2009).

[30] Juste, Fabiola Staróbole, and Claudia Regina Furquim De Andrade. ”Speech disfluency types of fluent and stuttering individuals: age effects.” Folia Phoniatrica et Logopaedica 63, no. 2 (2011): 57-64.

[31] Oue, Stacey, Ricard Marxer, and Frank Rudzicz. ”Automatic dysfluency detection in dysarthric speech using deep belief networks.” Proceedings of SLPAT 2015: 6th Workshop on Speech and Language Processing for Assistive Technologies. 2015.

[32] Riad, Rachid, Anne-Catherine Bachoud-Lévi, Frank Rudzicz, and Emmanuel Dupoux. ”Identification of primary and collateral tracks in stuttered speech.” arXiv preprint arXiv:2003.01018 (2020).

[33] Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. ”Deep Learning (Adaptive Computation and Machine Learning Series).” DOI 10 (2016): 1762-1766.