CLASSIFICATION OF SPEECH FLUENCY THROUGH TRANSFER LEARNING ON SPECTROGRAM

S. Mohamed Mansoor Roomi[1]*, K. Priya[2], V. Natchu[3], Faazelah Mohamed Farook[4]

Abstract
Fluency recognition from speech signals plays a vital role in computer-assisted voice analysis. The proposed work presents a computational framework using an audio processing system capable of classifying the fluency of speech such as fluency, non-fluency pause, and non-fluency stammer. The proposed model comprises pre-processing, spectrogram generation, and classification of speech fluency by the VGG16 pre-trained model. This model consists of convolutional layers and these layers extract discriminative features from spectrogram images of the speech signal. In this work, speech datasets such as Libri Speech, Crosslinguistic Corpus of Hesitation Phenomena (CCHP) English, and University College London's Archive of Stuttered Speech (UCLASS) were used to find speech fluency. The performance of the proposed model was compared with the existing pre-trained network and state of art methods.

Keywords: Convolutional layers, Spectrogram, Speech fluency, Transfer Learning, VGG16

[1*,2,3,4]Electronics and Communication Engineering Thiagarajar College of Engineering Madurai, TamilNadu, India. [1]Email: smmroomi@tce,edu, [2]Email:priya5586@gmail.com, [3]Email: natchu@student.tce.edu , [4]Email: faazelah@student.tce.edu

***Corresponding Author:** S. Mohamed Mansoor Roomi
***Electronics and Communication Engineering Thiagarajar College of Engineering Madurai, Tamil Nadu, India. Email: smmroomi@tce.edu

DOI: 10.48047/ecb/2023.12.si5a.0272

I. Introduction

The most natural form of human communication is speech signals. Breathing, phonation, and articulation are carefully coordinated muscular actions that result in the sounds of normal speech [1]. Speech Fluency Recognition (SFR) has been an active research area where the goal is to recognize the fluency of the speech signal. The definition of fluency is overall speaking proficiency with smoothness and ease of speech delivery. Another definition of fluency is non-stuttered and forward-moving speech in regards to being both content and productive [2]. Fluency may be affected due to persons having a problem delivering a speech.

Many speech disorders, including apraxia, dysarthria, and stuttering or stammering, have an impact on speech fluency, which can cause issues with speech fluency processing. Stuttering is one of these speech problems, affecting 1% (70 million) of the world's population [3]. It is a neuro develop mental problem that develops when neurological connections that support language, voice, and emotional function alter quickly. Repetition, extension, silent extended pauses, and filled gaps all break up the flow of speech. Repetition of a sound, word, or sentence is the one of these that comes closest to stammering. Filled pauses, which are a mixture of phonetics such as a, am, ahh, e, em, u, umm, ohh, etc., also affect the fluidity of speech. The accuracy of the ASR is decreased by filled pauses and speech stutters in gadgets like Google Home, Amazon Echo, Microsoft Cortana, etc. The sample speech wave of types of speech fluency waveform is shown in figure 1. To distinguish between speech that is fluent and that that is not, both conventional and deep learning techniques can be used.

Mel Frequency Cepstral Coefficients (MFCC) and Dynamic Time Warp (DWT) features classified by K Nearest Neighbor (KNN) classifier were used by Monica Mundada et al. [4] to provide a fluency classification model.

Jing Jiang et al provided stuttering types based on brain activity. In this work, the pattern classifier was performed to find the more typical or less typical stuttering of the patient using MRI image [5]. Michael Yi-Chao Jiang et al [6] suggested Automatic Speech Recognition (ASR) method for language classrooms. The phonological correctness, speed fluency, and repair fluency of English language learners are improved by this method.

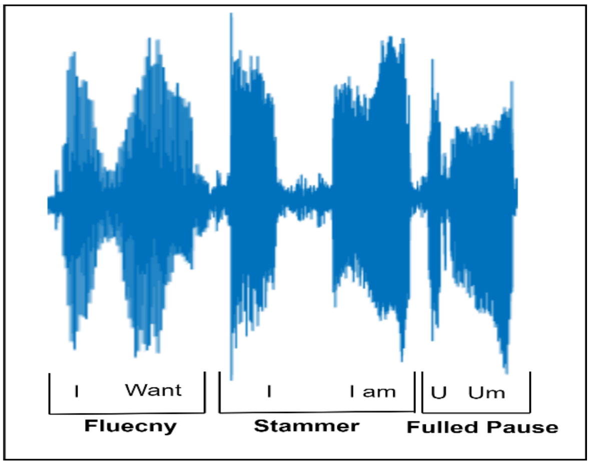

Fig.1. Types of speech waveform

Karthik et al [7] presented work for filled pause detection using vocal tract variations and concluded the formant is stable during a filled pause signal. Megdalena et al [8] worked for speaker recognition systems to detect filled pauses, audible breath, and correlation between the temporal structure of breath. Kiettiphong et al [9] proposed work for pause classification using MFCC and Long Short-Term Memory (LSTM). Using an artificial neural network, Dash et al. [10] published a work to recognize the speech signal for the text-speech system with acceptable stuttering speech (ANN). An ANN for stuttering categorization using auditory features including MFCC, formants, and Zero Crossing Rate was proposed by Savin et al. [11]. (ZCR). For a real-time online application, Mishra et al. [12] presented a Neural Network (NN) to identify stutter or not. A hierarchical NN system was suggested by Izabella et al. [13] to identify speech signals that were stuttering. The input vector's dimension was reduced by this network's design, and the neurons at the output level produced better results more quickly.

The ASR devices cannot be effectively used for stammering or filled pause speech. From the literature survey, the existing work focus on only pause or stammer detection. Detecting such stammering or filled pause speech would aid them to use those devices effectively and communicate with others through this device without any hesitation. So, the proposed method detects the speech signal as fluency, non-fluency-filled pause, and non-fluency stammer to improve the ASR accuracy.

The main contributions of the proposed works are

- Collection and Compilation of speech databases into classes like Fluency, Filled Pause, and stammer.

- Proposal of fluency classification model using spectrogram and pre-trained deep learning model.
- Performance comparison of the proposed method against various ANN classifiers, and state of art.

Sections 2 and 3 of the paper explained dataset collections and the proposed methodology. Section 4 described the results and discussions of the proposed method. Section 5 concludes the paper.

II. Speech Database Collection

There are no available databases for speech fluency recognition such as fluency, Non-Fluency Pause, and Non-Fluency Stammer. So that collects various speech corpus such as Libri Speech Corpus [14], Crosslinguistic Corpus of Hesitation Phenomena (CCHP) [15], and University College London Archive of Stuttered Speech (UCLASS) [16] to satisfy the need for fluency classification. Libri Speech corpus contains 1000 hours of speech which is a reading book with 2 to 3 min length for the Libri Vox project. This Speech is delivered without any hesitation and pauses so that it is used for fluency speech class. CCHP is a corpus that contains three reading practices reading aloud, picture description, and topic narrative. Among these, picture descriptions and topic narrative recordings have a pause voice which is taken for Non-fluency pause class. UCLASS speech is collected from the university college London and specially designed for stuttering speech with 2 to 3 minutes in length. It is taken for Non-fluency stammer class. The details of the collected databases are shown in Table 1. All three datasets contain 2 to 4 minutes of speech. This speech signal split into non overlapped 5 seconds in length. After segmenting the speech files the number of the sample speech signal in each class is 2038.

Table 1: Collected Database Descriptions

Speech Fluency Class	Corpus Name	Sampling Rate (kHz)	Length(Minutes)
Fluency	Libri speech	16	2 - 3
Non-Fluency - Pause	CCHP English	48	3 - 4
Non-Fluency Stammer	UCLASS	44.1	2 - 3

III. Proposed Methodology

In proposed work consists of signal pre-processing, spectrogram generation, and classification through training the pre-trained model as shown in figure 2. In pre-processing the silent removal [17] and pre-emphasis [18] filtering was applied to the collected speech signals. Then generate a spectrogram [19] of the pre-processed signal. The spectrogram of the speech signal was the input of the pre-trained model [20].

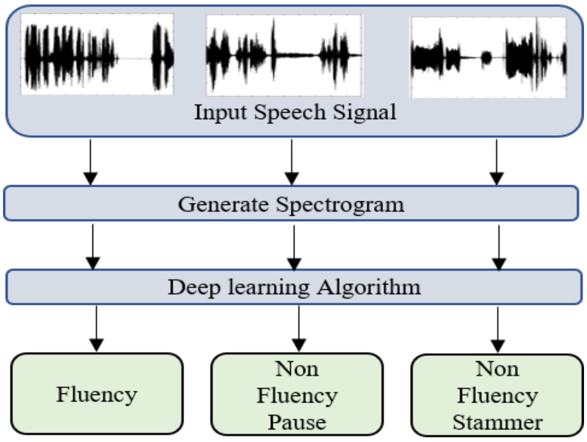

Fig. 2. Flow of proposed method

The quiet or noise in the segmented voice signal may be present, and this portion is information-free. In comparison to the vocal signal, the energy of noise or stillness is quite low. Therefore, silence reduction is a crucial step in increasing classification accuracy. By using a Voice Activity Detection (VAD) block with a threshold, the suggested method eliminates the silent portion. Then the silence-removed signal is applied to the pre emphasize filter to boost the high-frequency signals.

The two-dimensional graph of the spectrogram has time on the horizontal axis and frequency on the vertical axis and it was the Short Time Fourier Transform (STFT) of the pre-processed speech signal. The amplitude of the speech signal was represented by the strength of the color of the spectrogram. First, the speech signal is divided into several sections using equation 1

$$Nsc=N/4.5 \qquad (1)$$

Where N is the number of samples in the speech signal. These sections multiplied with a hamming window with a 50% overlapping as represented in equation 2.

$$w_n = 0.5 - 0.46\cos\left[\left(\frac{2\pi}{N}\right)n\right] \qquad (2)$$

Where w(n) is the hamming window of the signal. Then Compute the FFT using equation 3.

$$X_{k=}\sum_{n=0}^{N-1} w_n x_n e^{\frac{-i2\pi kn}{N}} \qquad k = 0,1,2 \ldots . N-1 \qquad (3)$$

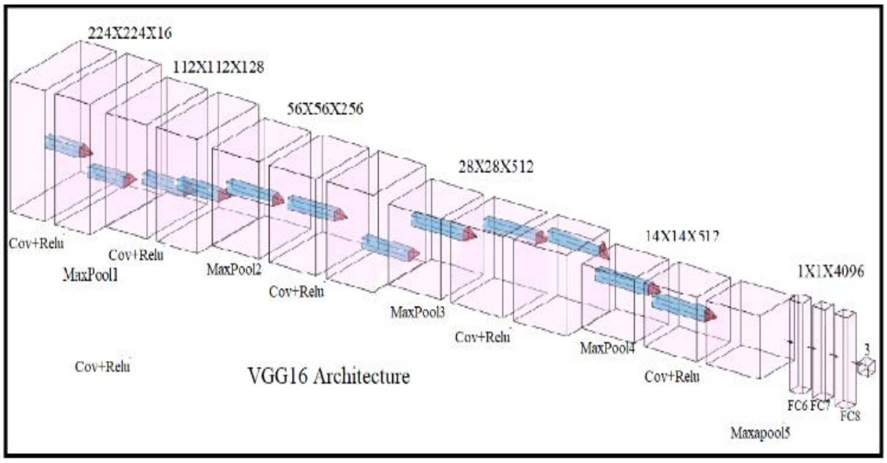

Fig. 3. VGG Architecture

Then the spectrogram of the images was trained with VGG 16 pre-trained model. VGG16 architecture contains 16 layers as shown in figure 3. The input image size of the model was 224*224*3. The layer description of the VGG16 model is listed in table 3.

Table 2: Layer Description of VGG16

Layer		Feature Map	Size	Kernel Size	Stride	Activation
Input	Image	1	224×224×3	-	-	-
1	2×Convolution	64	224×224×64	3×3	1	Relu
	Max Pooling	64	112×112×64	3×3	2	Relu
3	2×Convolution	128	112×112×128	3×3	1	Relu
	Max Pooling	128	56×56×128	3×3	2	Relu
5	2×Convolution	256	56×56×256	3×3	1	Relu
	Max Pooling	256	28×28×256	3×3	2	Relu
7	3×Convolution	512	28×28×512	3×3	1	Relu
	Max Pooling	512	14×14×512	3×3	2	Relu
10	3×Convolution	512	14×14×512	3×3	1	Relu
	Max Pooling	512	7×7×512	3×3	2	Relu
13	Fully Connected	-	25088	-		Relu
14	Fully Connected	-	4096	-		Relu
15	Fully Connected	-	4096	-		Relu
Output	Fully Connected	-	3	-		Softmax

IV. Results and Discussions

In this section, the results of the proposed method have been explained. The 80% samples of the speech signals were used to train the VGG16 model and the remaining samples were used for testing. The speech signal is silence removed by the thresholding algorithm and then applied to the pre-emphasis filter. Figure 4 shows the sample speech signal with a stammering voice. Figure 5 and 6 represents the silence-removed waveform and pre-emphasized waveform. The spectrogram of the pre-processed signal is obtained for training as shown in figure 7.

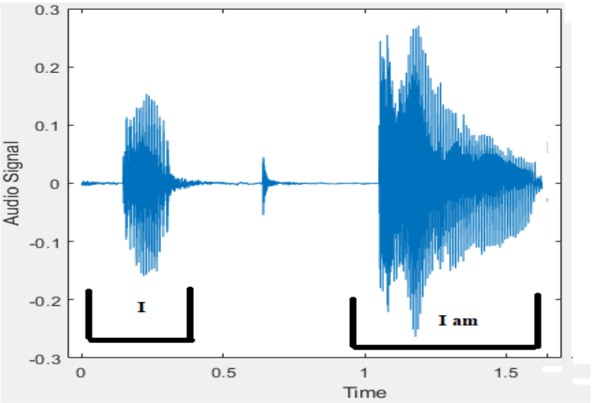

Fig. 4. Sample Input Stammer Speech Waveform

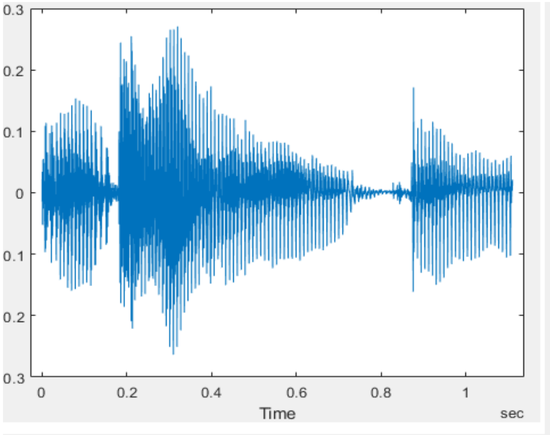

Fig. 5. Silence Removed Waveform

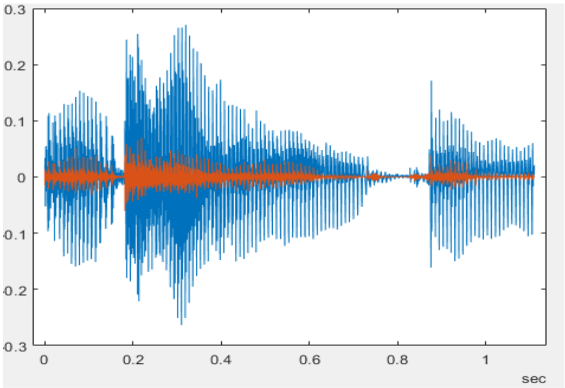

Fig. 6. Pre-Emphasized Waveform

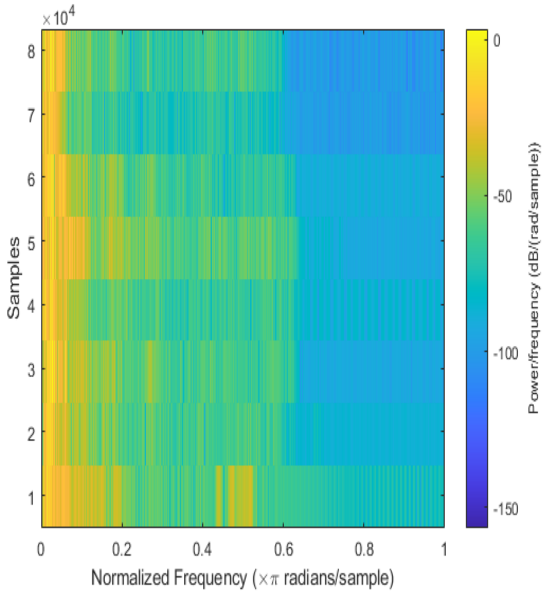

Fig. 7. Spectrogram of Waveform

The spectrogram of the images is trained using the VGG16 model. Table 3 listed the hyperparameter tuning of the model. Among these parameters, the minibatch size, learning rate, and epochs are tuned to increase the speech fluency recognition as listed in table 4. The proposed method achieves 89.59% of speech fluency recognition.

Table 3: Hyper Tuning Parameters of the VGG16 model

Parameters	**Name/Value**
Optimizer	Stochastic Gradient Descent with Momentum (SGDM)
Mini Batch Size	8
Device Type	GPU
Learning Rate	0.01
l2Regularization	0.001
Gradient clipping	Gradient Threshold
Epoch	10
Verbose Frequency	50

Table 4: Tuning Parameters of the VGG16 model

Epoch	Learning rate	Testing Accuracy (%)
5	0.1	74.23
5	0.01	77.84
5	0.001	78.51
10	0.1	81.67
10	0.01	82.45
10	0.001	84.74
15	0.1	85.12
15	0.01	87.84
15	0.001	89.59
20	0.1	88.45
20	0.01	88.32

The performance of the proposed work can be evaluated using the confusion matrix as shown in figure 8. Classifications for speech fluency include fluency, filled pauses, and stutters. Fluency speaking is given a class 1 classification. 1747 speech signals in the fluency class are accurately identified as speech signals, 1828 samples as filled pauses, and 1908 samples as stammers. Precision, recall, and the F1 score are used to gauge the performance of the classifier. The proposed model performs better because 89.59% of the speech samples were overall accurately sorted into the appropriate class.

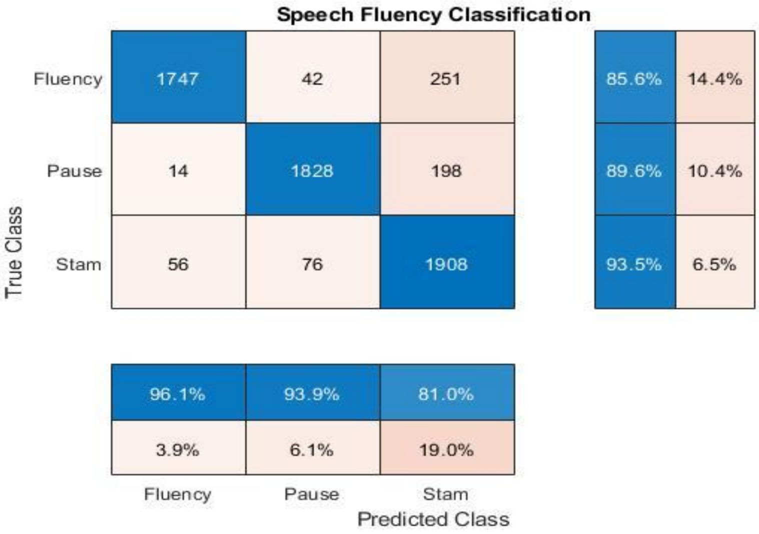

Fig. 8. Confusion matrix of the speech fluency recognition

Table 5: VGG16 Classification Performance

Fluency class	Precision (%)	Recall (%)	F1 Score (%)	Accuracy (%)
Fluency	85.5	96.1	90.5	89.59
Non-Fluency -Pause	89.6	93.9	91.7	
Non-Fluency -Stam	93.5	81.0	86.8	

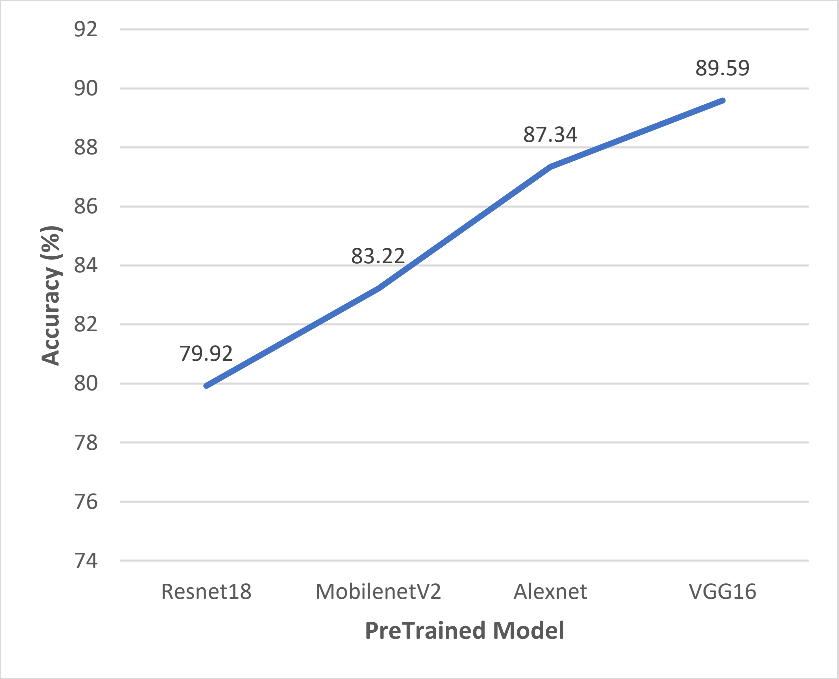

Fig. 9. Performance Evaluation of the Proposed Method with various Pre-trained Models

The performance of the proposed model compared with other pre-trained models such as Resnet18, MobilenetV2, Alexnet, and VGG16 is shown in figure 9. Among these, the proposed VGG16 achieves higher classification accuracy. Table 7 listed the comparison of the proposed method against current fluency recognition approaches. In the current approaches, fluency recognition varies in the range of 75% to 87.39%. The prosed method achieved better recognition accuracy than the existing methods.

Table 7: Comparison of State of Art Methods

Reference	Database	Features	Classifier	Detected Speech	Accuracy (%)
Mishra et al (2021)	UCLASS	MFCC, RMSE	NN	Stammer	87
Dash et al (2018)	Local Dataset	MFCC, Formants, Zero Crossing Rate (ZCR)	ANN	Stammer	87.39
Kiettiphong (2018)	LOTUS BN	MFCC	LSTM NN	Pause	74.7
Magdalena (2016)	Spontaneous Polish Speech	MFCC	XG Boost	Pause	75
Kartik et al (2009)	IBM Call Center Data set	Formant	Cepstral Based Method	Pause	78
Proposed Method	Libri Speech, CCHP, UCLASS	Deep features from the spectrogram	VGG16 pre-trained model	Fluent, Stammer and Pause	95.59

V. CONCLUSION

In this work, various databases such as Libri speech, CCH English and, UCLASS are collected for fluency classes such as fluency, Non-fluency pause and, Non-fluency stammer because of a lack of fluency databases. The proposed model consists of signal pre-processing, spectrogram generation, and fluency classification by the VGG16 model. The proposed model achieved better fluency recognition of 89.59%. The performance of the proposed method is compared with various pre-trained models and state of art methods. As a result, the proposed method is highly recommended for fluency recognition based on the speech signal.

ACKNOWLEDGEMENT

The authors express their gratitude to the Thiagarajar College of Engineering (TCE) for supporting us to carry out this research work. Also, the financial support from TCE under the Thiagarajar Research Fellowship scheme (File.no: TRF/Jul-2021/04) is gratefully acknowledged.

REFERENCES

1. Paul, D.R. (2013). Fluency and Fluency Disorders. In: Volkmar, F.R. (eds) Encyclopedia of Autism Spectrum Disorders. Springer, New York, NY. https://doi.org/10.1007/978-1-4419-1698-3_1931
2. Edwards, S. (1989). Book reviews : Fluency and stuttering C. Woodruff Starkweather Englewood Cliffs, NJ: Prentice-Hall, 1987. xiv+272 pp. Child Language Teaching and Therapy, 5(1), 99–101. https://doi.org/10.1177/026565908900500114
3. Stuttering Facts and Information. url: http://www.stutteringhelp.org/faq.
4. Monica Mundada, Sangramsing Kayte, Sumegh Tharewal, Dr. Bharti Gawali, Classification of Fluent and Dysfluent Speech Using KNN Classifier, International Journal of Advanced Research in Computer Science and Software Engineering, IJARCSSE, Volume 4, Issue 9, September 2014 ISSN: 2277 128X
5. Jiang J, Lu C, Peng D, Zhu C, Howell P. Classification of types of stuttering symptoms based on brain activity. PLoS One. 2012;7(6):e39747. doi: 10.1371/journal.pone.0039747. Epub 2012 Jun 25. PMID: 22761887; PMCID: PMC3382568.
6. Michael Yi-Chao Jiang, Morris Siu-Yung Jong, Wilfred Wing-Fat Lau, Ching-Sing Chai, Na Wu, Exploring the effects of automatic speech recognition technology on oral accuracy and fluency in a flipped classroom, Journal of Computer Assisted Learning, 22 August 2022, https://doi.org/10.1111/jcal.12732.
7. Kartik Audhkhasi, Kundan Kandhway, Om D. Deshmukh, Ashish Verma, Formant-Based Technique for Automatic Filled-Pause Detection In Spontaneous Spoken English, ICASSP 2009, 978-1-4244-2354-5/09/$25.00 ©2009 IEEE.
8. Magdalena Igras-Cybulska, Bartosz Ziółko, Piotr Żelasko and Marcin Witkowski, Structure of pauses in speech in the context of speaker verification and classification of speech type, Journal on Audio, Speech, and Music Processing (2016) 2016:18 DOI 10.1186/s13636-016-0096-7.
9. Kiettiphong Manovisut, Pokpong Songmuang, and Nattanun Thatphithakkul, An Improved Short Pause Based Voice Activity Detection Using Long Short-Term Memory Recurrent Neural Network, Springer Nature Singapore Pte Ltd. 2018 J. Chen et al. (Eds.): KSS 2018, CCIS 949, pp. 267–274, 2018. https://doi.org/10.1007/978-981-13-3149-7_20268.
10. A. Dash, N. Subramani, T. Manjunath, V. Yaragarala, and S. Tripathi, “Speech

recognition and correction of a stuttered speech," in 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2018, pp. 1757–1760.

11. P. S. Savin , Pravin B. Ramteke, Shashidhar G, "Recognition of Repetition and Prolongation in Stuttered Speech Using ANN", " Proceedings of 3rd International Conference on Advanced Computing, Networking and Informatics", pp 65- 71, Volume 1, 08 October 2015.
12. Mishra, N., Gupta, A. & Vathana, D. Optimization of stammering in speech recognition applications. Int J Speech Technol 24, 679–685 (2021). https://doi.org/10.1007/s10772-021-09828-w.
13. Izabela Swietlicka and Wieslawa Kuniszyk, zkowiak and Elzbieta Smolka, Hierarchical ANN system for stuttering identification Comput. Speech Lang 2013, volume 27, pages 228-242.
14. Vassil Panayotov, Guoguo Chen, Daniel Povey, Sanjeev Khudanpur, Librispeech: An Asr Corpus Based On Public Domain Audio Books. https://librivox.org/
15. See https://filledpause.org/chp/cchp/
16. See https://www.uclass.psychol.ucl.ac.uk/
17. T. Giannakopoulos. (2009) A method for silence removal and segmentation of speech signals, implemented in matlab. University of Athens, Athens.
18. Vergin, R., & O'Shaughnessy, D. (1995). Pre-emphasis and speech recognition. Proceedings 1995 Canadian Conference on Electrical and Computer Engineering, pp. 1062-1065 vol.2, doi: 10.1109/CCECE.1995.526613.