A Comprehensive Review of Stuttering Identification from Statistical Modelling to Deep Learning: Resources, Challenges and Future Directions

SHAKEEL A. SHEIKH*, Université de Lorraine, CNRS, Inria, LORIA, F-54000, Nancy, France
MD SAHIDULLAH, Université de Lorraine, CNRS, Inria, LORIA, F-54000, Nancy, France
FABRICE HIRSCH, Université Paul-Valéry Montpellier, CNRS, Praxiling, Montpellier, France
SLIM OUNI, Université de Lorraine, CNRS, Inria, LORIA, F-54000, Nancy, France

Stuttering which is also called stammering is a neuro-developmental speech disorder during which the flow of speech is interrupted by involuntary and silent pauses, by the disruption of unplanned prolongation and repetition of phrases, words, syllables or sounds. The conventional assessment of stuttering is to count manually the occurrences of stuttering types and indicate them as a proportion to the total number of words in a speech passage. The main drawback in this manual counting is that they are time consuming and subjective which makes it inconsistent and prone to error across different judges/ speech therapists. Approximately 70 million people suffer with stuttering problem worldwide which constitutes 1% of the world's population. Among them, the stuttering is significant in males which is approximately four-fifth. Stuttering is an interesting interdisciplinary domain research problem which involves pathology, psychology, acoustics, signal processing and deep learning that makes it hard and complicated to detect. Recent developments in machine and deep learning has dramatically revolutionized stuttering identification/classification problem. However on this exciting progress, there is a lack of comprehensive review. In this paper, we review comprehensibly the acoustics features, statistical and deep learning based stuttering/disfluency classification models with its challenges and possible solutions.

CCS Concepts: • **Speech Disorder Detection → Stuttering**; *Speech disfluency, speech disorder and deep learning*.

Additional Key Words and Phrases: stuttering detection, datasets, machine learning, deep learning, gaze detection, modality

ACM Reference Format:
Shakeel A. Sheikh, Md Sahidullah, Fabrice Hirsch, and Slim Ouni. 2021. A Comprehensive Review of Stuttering Identification from Statistical Modelling to Deep Learning: Resources, Challenges and Future Directions. *ACM Comput. Surv.* 54, 3, Article 20 (May 2021), 26 pages.

1 INTRODUCTION

Speech disorders or speech impairments are communication disorder in which a person has difficulties in creating and forming the normal speech sounds required to communicate with

Authors' addresses: Shakeel A. Sheikh, shakeeel-ahmad.sheikh@loria.fr, Université de Lorraine, CNRS, Inria, LORIA, F-54000, Nancy, France; Md Sahidullah, Université de Lorraine, CNRS, Inria, LORIA, F-54000, Nancy, France, md.sahidullah@inria.fr; Fabrice Hirsch, Université Paul-Valéry Montpellier, CNRS, Praxiling, Montpellier, France, fabrice.hirsch@univ-montp3.fr; Slim Ouni, Université de Lorraine, CNRS, Inria, LORIA, F-54000, Nancy, France, slim.ouni@loria.fr.

0360-0300/2021/5-ART20 $15.00
https://doi.org/

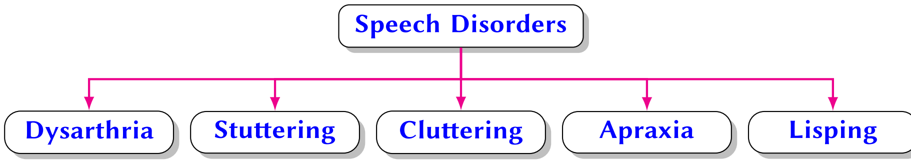

Fig. 1. Speech Disorders

others[1]. These disorders can take the form of dysarthria, apraxia, stuttering, cluttering, lisping, and so on [44] as shown in Figure 1. Only 5 to 10% of the world population are able to to produce proper speech sounds, the rest face some sort of speech disorder [44]. Dysarthria is defined as a speech disorder caused my muscle weakness (including face, lips, tongue, and throat) controlled by nervous system. The patients with dysarthria (PWD) produce slurred or mumbled sounds with aberrant speech patterns, such as flat intonation or very low or fast speech rate, which makes their speech very difficult to comprehend[2]. Cluttering is characterized by a patient's speech being too jerky, too rapid or both. Persons with cluttering (PWC) usually exclude/collapse most of the syllables, or aberrant rhythms or syllable stresses, and also contains excessive amounts of interjections such as *so*, hmm, *like*, *umm*, etc[3]. Apraxia is defined as a speech disorder when the neural path between the nervous system and the muscles responsible for speech production is obscured or lost. The persons with apraxia (PWA) knows what they want to speak, but can not speak due to the fact that the brain is unable to send exact message to the speech muscles which can articulate the intended sounds, despite of the fact speech muscle movements are working fine[3]. Usually, PWA doesn't speak words the same way everytime. They can speak shorter words more clearly than longer ones. Lisping speech disorder is defined as the incapability of producing sibilant consonants (*z or s*) correctly. The sibilant sounds are usually substituted by *th* sounds. For example, the persons with lisping speech disorder would pronounce the word *lisp* as *lithp*[4]. Stuttering is characterized by core behaviours which usually take the form of involuntary stoppages, repetition and prolongation of sounds, syllables, words or phrases. Of these speech impairments, stuttering - also called stammering/disfluency - is the most common one[5]. In this review, we will focus mainly on stuttering disorder detection in the context of machine and deep learning.

Fluency can be defined as the capacity to produce speech without any effort, at a fast rate [83]. A fluent speech requires linguistic knowledge in the spoken language and a mastery of the message content. Concerning physiological aspects, a precise respiratory, laryngeal and supraglottic movements control is necessary to maintain fluency [1]. When all these conditions are not met, speech disorder (stuttering) can emerge. They can take the form of silent or filled pauses, repetitions, interjections, revisions (content change or grammatical structure or pronunciation change), incomplete phrases,.. [75]. Generally, the normal speech is made up of mostly the fluent and some disfluent parts. Notice that normal disfluencies are useful in speech production, since they can be considered in time during which the speaker can correct or plan the upcoming discourse.

In some cases, like stuttering, disfluencies do not help the speaker to organize his/her discourse. Indeed, contrary to persons without any fluency disorder, persons who stutter (PWS) know what they want to pronounce but are momentarily unable to produce it [63].

[1]https://bit.ly/3ehBab7
[2]https://www.asha.org/public/speech/disorders/dysarthria/
[3]https://www.speechpathologygraduateprograms.org/2018/01/10-most-common-speech-language-disorders/
[4]https://bit.ly/3xjpzRI
[5]https://www.healthline.com/health/speech-disorders

Stuttering also called stammering[6] is a speech disorder which can be described as an abnormally and persistent duration of stoppages in the normal forward flow of speech [33]. These speech abnormalities are generally accompanied by unusual behaviours like head nodding, lip tremors, quick eye blinks and unusual lip shapes etc [73]. Stuttering can broadly be classified into two types:

- *Developmental Stuttering:* This stuttering is the most common one and it usually occurs in the learning phase of the language, *i.e. between 2 and 7.* Recent researches conclude that developmental stuttering is a multifactorial trouble including neurological and genetic aspects [25, 27]. Indeed, fMRI studies show anomalies in neural activity before speech, *i.e.* during the planning stages of speech production [91]. Furthermore, an atypical activation in the left inferior frontal gyrus and right auditory regions [7, 58] has been highlighted. Concerning the genetic aspects, [72] observe an unusual allele on chromosom 12 by PWS. Denis et al. [25] identify 87 genes which could be involved in stuttering, including one called GNPTAB, which was significantly present by a lot of PWS.
- *Neurogenic Stuttering:* This stuttering can occur after head trauma, brain stroke, or any kind of brain injury. This results in disfluent speech because of the incoordintaion of the different regions of the brain which are involved in speaking [60]. Even if neurogenic stuttering is rare, it can be observed by children and adults regardless of ages.

Globally, stuttering concerns 1% of the world's global population and its incidence rate is between 5% and 17% [98]. The difference between the prevalence and incidence rates can be explained by the fact that developmental stuttering disappears in 80% of the cases before adulthood either without any intervention or thanks to a speech therapy. Thus, about 70 million people suffer from this trouble which affects four times males than females [98].

The various factors that lead to stuttering which include stress, delayed childhood development, speech motor control abnormalities as there is a strong correlation between stress, anxiety and stuttering. As for normo-fluent speakers, fluency of PWS depends on several factors. Indeed, disfluencies are more frequent in stress or anxiety conditions, in dual tasks including speech and another cognitive charge and when they speak fast. Conversely, PWS produce less disfluencies when they sing in unison or speak with an altered auditory feedback [4]. As considered by the non-stuttering persons, the disfluency affects to the flow of speech only, however for PWS, it is greater than that. Several studies show that PWS are ignored, teased and/or bullied by normo-fluent [45]. The PWS are usually rated less popular than their non-stuttering peers and less likely to be considered leaders [45]. According to [6], 40% of the PWS have been repudiated school opportunity, promotion or job offers and it affects relationships. These data should be assessed in close conjunction with the fact that 85% of businessman consider stuttering as a negative element during a job interview and prefer offering a work which does not require a customer contact [48]. All these elements explain that PWS develop social anxieties and negative feelings (fear, shame,…) when they have to speak in public [10].

Concerning stuttering-like disfluencies, several types have been observed: repetitions, blocks, prolongations, interjections etc are detailed in Table 1. Some works try to link the locus of disfluencies and phonetic proprieties. M Blomgren et al [46], H. M. Chandrashekar et al. [9], and M Jayaram et al. [22] indicate that unvoiced consonants are more disfluent than their voiced counterparts. Furthermore, H. M. Chandrashekar et al. [9] notices that disfluencies are more frequent at the beginning of an utterance or just after a silent pause. Moreover, Ivana Didirkova. et al. [22] observes an important inter-individual variability concerning sounds and/or phonetic features which are the most disfluent. Studies based on motor data have been carried out about stuttering. E.G. Conture et al. [17, 18] observe inappropriate vocal folds abductions and adductions which lead to anarchic

[6]In this review, we will use the terms disfluency, stuttering and stammering interchangeably

Stutter Type	Definition	Example
Blocks	Involuntary pause before a word	I want **blockage/pause** to speak
Prolongations	Prolonged Sounds	**Sssssss**am is kind
Interjection	Insertion of sounds	uh, uhm
Sound Repetition	Phoneme repetition	He **w-w-w**-wants to write
Part-Word Repetition	Repetition of a Syllable	**Go-go-go** back
Word Repetition	Repetition of a Word	**Well, well**, I didn't get you
Phrase Repetition	Repetition of several successive words	**I have, I have** an iphone
Repetition–Prolongation	Repetition and Prolongation disfluencies occurring at the same time	**Ggg**o **b-b-b**back
Multiple	Multiple disfluencies in a word or phrase	**Tttttt-Tttttt-Tttttt**ariq **blockage/pause** is **kkkk**ind
False Start	Revision of a phrase or a word	**I had-** I lost my watch

Table 1. Various Stuttering Types

openings and closure of the glottis. Concerning the supraglottic level, ME Wingate. et al [96] hypothesizes that stuttering is not a trouble dealing with sounds production but a coarticulation trouble. He theorizes that disfluencies occur during a fault line, which corresponds to the interval when muscular activity due to a sound which have been produced is going off and muscular movements for the following sound is going on. More recently, Ivana et al.[23] show, thanks to EMA data, that stuttering is not only a coarticulation trouble. Furthermore, another study based on articulatory observations, note that stuttered disfluencies and non-pathological disfluencies do have common characteristics. However, stuttered disfluencies and non-pathological disfluencies produced by PWS present some particularities, mainly in terms of retention and anticipation, and the presence of spasmodic movements [24]. PWS tend to develop strategies allowing them to avoid sounds or words which can result in a disfluency; such strategies consist in using paraphrases or synonyms instead of the problematic segment [95].

Acoustic analyses have been carried out about stuttering, including speech rate, vowel(V)-consonant(C) transition duration, stop-gap duration, fricative duration, voice on-set time (VOT), CV transition duration, vowel duration, formants, glottis constriction, sharp increase in articulatory power and closure length elongation before the speech segmented is released [102]. Dehqan et al studied the second formant (F2) transitions of fluent segments of persian speaking PWS [21]. They concluded that the F2 frequency extent transitions are greater in stuttering speakers than fluent ones. They also reported that the PWS takes longer to reach vowel steady state, but the overall F2 formant slopes are similar for both stuttering speakers and normal ones [21]. The PWS generally exhibit slower speaking rates when compared to normal speakers.

Healey et al in [37] showed that for voiceless stops, chronic stuttering exhibits longer VOT when compared with normal persons. They showed that consonant and vowel duration were longer only in real-world phrases like *take the shape* in contrast with nonsense phrases like *ipi saw ipi* [37]. In [38], Hillman et al also found that the PWS reveals longer VOT for voiceless stops than fluent persons. In [2] Adams et al. found that not only voiceless stops exhibits longer VOT in PWS, but also, voiced stops displays longer VOT than non-stuttering persons.

Several other studies have investigated the CV formant transitions in stuttered speech. Yarus et al. examined the F2 transitions of children who stutter on syllable repetitions, and found no aberrant F2 transitions [100]. However Robb et al. analyzed the fluent speech segments of PWS, and showed that F2 fluctuations are longer for voiced and voiceless stops than normal speakers [74]. In a different study by Chang et al. [13], where 3-5 year aged children were analyzed in picture-naming task. The results showed that disfluent children produced smaller fluctuations of F2 transitions between alveolar and bilabial place of articulations than did fluent children, and the overall of

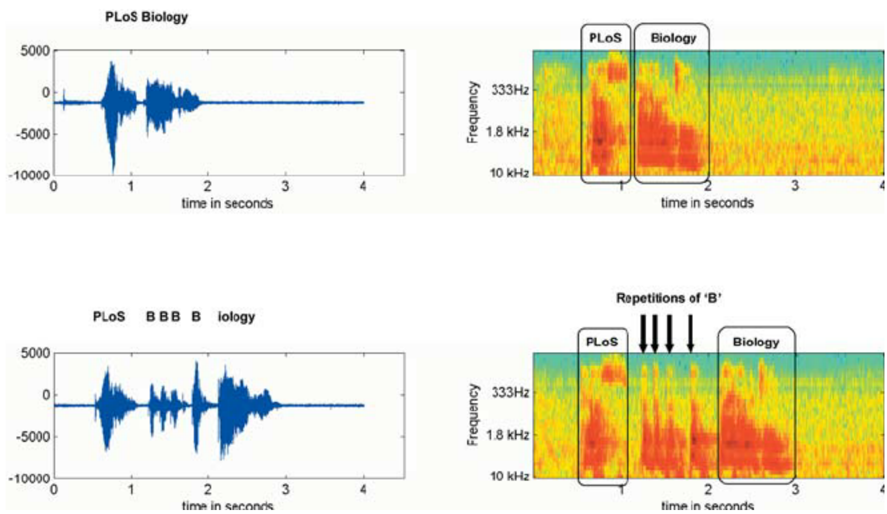

Fig. 2. Speech Waveforms and Spectrograms of a Speaker(Male) Saying "PLoS Biology" The left is waveforms (amplitude v/s time); the right is a time–frequency plot using a wavelet decomposition of these data. Top row is fluent speech; bottom row is stuttering (repetitions), occur at the "B" in "Biology." Four repetitions can be clearly identified by arrows in the spectrogram. (bottom right) [11]

.

degree of CV coarticulation is no different among stuttering and control groups. Subramanian et al. in [84] analyzed the F2 frequency fluctuations of voiceless stops, and revealed that near the onsets of CV, the stuttering children exhibited smaller F2 changes than the normal speakers.

If stuttering has been the subject of a lot of researches, additional studies on disfluencies should be carried out. Several dataset are available to produce such kind of works:

UCLASS. The most common concern in stuttering research is the lack of training data. University College Londons Archive of Stuttered Speech (UCLASS) public dataset (although very small) [39] is the most commonly used amongst the stuttering research community. The UCLASS comes in two releases from the UCL's department of Psychology and Language Sciences. This contains monologues, conversations, readings with a total audio recordings of 457. Some of these recordings contain transcriptions like orthographic, phonetic and standard ones. Of these, orthographic are the ones which are best suitable for stutter labelling. The UCLASS[7] release 1 contains 139 monologue samples from 81 PWS, aged from 5 to 47 years. The male samples are 120 and female samples are 18. The summary of UCLASS is given in the Table 2.

		Age				Gender	
Category	N	Range	Mean	STD	Median	Male	Female
UCL	139	5y4m-47y0m	13y2.86m	6y1m	12y1m	121	18

Table 2. The UCLASS Release 1 number all ages are given in NNyNNm format)

[7] url:http://www.uclass.psychol.ucl.ac.uk/uclass1.htm

LibriStutter. The availability of small amount of labelled stuttered speech led to synthetic LibriStutter's creation [50]. This LibriStutter is approximately of 20 hours and includes synthetic stutters for repetitions, prolongations and interjections. For each spoken word, T. Kourkounakis [50] used Google Cloud Speech-to-Text (GCSTT) API to generate timestamp correspondingly. Random stuttering was inserted within the 4 second window of each speech signal.

TORGO. This was developed by a collaboration between departments of Speech Language Pathology Computer Science at University of Toronto and the Holland-Bloorview Kids Rehab hospital [76]. This dataset comprises samples from seven persons, diagnosed with cerebralpalsy or amyotrophic lateral sclerosis including four males and three females aged between 16 to 50 years. In addition to this, it also contains samples from control speakers of the same age.

FluencyBank. This is a shared database for the study of fluency development which has been developed by Nan Bernstein Ratner (University of Maryland) and Brian MacWhinney (Carnegie Mellon University) [69]. The platform proposes audio and video files with transcriptions of adults and children who stutter.

The speech recognition systems (SRS) are working well for the fluent speech, but they fail to recognise the stuttered speech. So, it would not be feasible for a PWS to easily access virtual assistant tools like Alexa, Apple Siri etc [90].

Therefore, automatic stuttering identification systems (ASIS) is the need of an hour which provides an objective and consistent measurement of the stuttered speech. Consequently, with the recent developments in natural language processing (NLP), machine learning and deep learning, it became a reality to develop smart and interactive stuttering detection and therapy tools [50]. In-spite of the fact, that there are numerous applications of ASIS, very little attention has been given to this field.

We define an ASIS as a compilation of techniques and methodologies that takes audio speech signal as an input, pre-processes and categorizes them in order to identify the stuttering embedded in them. When we take a broad view of ASISs, we can express it into several domains as shown in Figure 3. It would be extremely useful to understand the stuttering better in order to enhance the stuttering classification process. The stuttering problem is still an open problem and it has been approached through several techniques, most of them fall in the supervised learning category. An ASIS system which consists of a classifier and a supervised learning loss function is trained on the data to recognize and identify stuttering types embedded in the audio speech signal. These supervised learning systems require the stuttering embedded labeled data. To feed the data to the model, it requires some preprocessing in order to extract useful features like Mel-frequency cepstral coefficients (MFCCs) which reduces the original data into its important characteristics that are essential for the classification purposes. In speech, these can be grouped into spectral, voice and prosodic features. The spectral ones are the mostly used in the literature. In addition to these, features from other modalities such as linguistic(textual) can also be incorporated to improve the classification performance. Deep learning based classifiers have become common these days for stuttering identification.

Some other relevant speech disorder problems, that have been tackled using deep learning incude - dysarthria, etc. Korzekwa et al. used encoder-decoder based approach for the detection and construction of dysarthric speech by successfully encoding the dysathric speech characteristics in the latent space [49]. The experiments were carried out using UA-Speech dataset from University of Illinois [49]. The model is trained jointly using the multimodal input including audio (mel spectrograms) and textual form of dysarthric speech. Gupta et al. [34] exploits residual network (ResNet) for the detection of dysarthric speech severity level using Universal Access (UA) corpus and reports an average accuracy of 98.90%. Chandrashekar et al. [12] evaluated the intelligibility of dysarthric speech with CNN and ANN on UA and Torgo datasets by investigating its spectro-temporal representation. In 2016, Chitralekha et al. investigated the use of voice parameters for dysarthric speech recognition. They showed that multi-taper spectral estimation based MFCC computation improves the recognition performance of unseen dysarthric speech. In another study, Chitralekha et al. used Time-Delay neural network based Denoising Autoencoder (TDNN-DAE) to enhance the dysarthric speech features to match that of normal speech and to make it recognizable for automatic speech processing (ASR) unit [8]. In a recent study by Juliette et al [56], instead of using hand-crafted acoustic features, they exploited raw speech directly for dysarthric detection with the attention based LSTM pipeline.

In this article, we give an up-to-date comprehensive literature survey of ASIS as shown in Figure 3. By providing this survey, we hope it would be a useful resource for the stuttering identification research community. The rest of the survey paper is organized as follows. The next two sections 2.1 and 2.2 provide a comprehensive present-day review of the earlier stuttering/disfluency detection works with the detailed analysis on experiments and results obtained. The challenges and future directions are listed under section 3 and concluding remarks in section 4.

2 AUTOMATIC STUTTERING IDENTIFICATION

2.1 Statistical Approaches

Stuttering identification, an interdisciplinary research problem in which a myriad number of research work (in-terms of acoustic feature extraction and classification methods) are currently going on with a focus on developing automatic tools for its detection and identification. This section provides in detail the comprehensive review of the various feature extraction and machine learning stuttering identification techniques, that have been used in the literature.

Acoustic Features: In case of developing any speech recognition or stuttering identification system, representative feature extraction and selection is extremely an important task that affects the system performance. The first common step in speech processing domain is the feature extraction. With the help of various signal processing techniques, we aim to extract the features that compactly represents the speech signal and approximates the human auditory system's response [43].

The various feature extraction methods that have been explored in the stuttering recognition systems are autocorrelation function and envelope parameters [40], duration, energy peaks, spectral of word based and part word based [41, 42, 47], age, sex, type of disfluency, frequency of disfluency, duration, physical concomitant, rate of speech, historical, attitudinal and behavioral scores, family history [29], duration and frequency of disfluent portions, speaking rate [61], frequency, 1^{st} to 3^{rd} formant's frequencies and its amplitudes [20, 47], spectral measure (Fast Fourier Transform (FFT) 512) [86, 88], Mel Frequency Cepstral Coefficients (MFCC) [16, 26, 35, 47], Linear Predictive Cepstral Coefficients (LPCCs) [35, 47], pitch, shimmer [53], zero crossing rate (ZCR) [47], short time average magnitude, spectral spread [47], Linear Predictive Coefficients (LPC), Weighted Linear Prediction Cepstral Coefficients(WLPCC) [35], Maximum Autocorrelation Value (MACV) [47], Linear Prediction-Hilbert transform based MFCC (LH-MFCC) [55], noise to harmonic ratio (NHR), shimmer harmonic to noise ratio (HNR), harmonicity, APQ(Amplitude Perturbation Quotient), formants and its variants (min, max, mean, median, mode, std), spectrum centroid [53], Kohonen's Self-organizing Maps [86], i-vectors [30], perceptual linear predictive (PLP) [26], respiratory biosignals [94], sample entropy feature [36]. With the recent developments in convolutional neural networks, the feature representation of stuttered speech is moving towards spectrogram representations from conventional MFCCs. One can easily discern the fluent and stuttered part of speech by analyzing the spectrograms as shown in Figure 2. T. Kourkounakis et al. in [50] exploited the

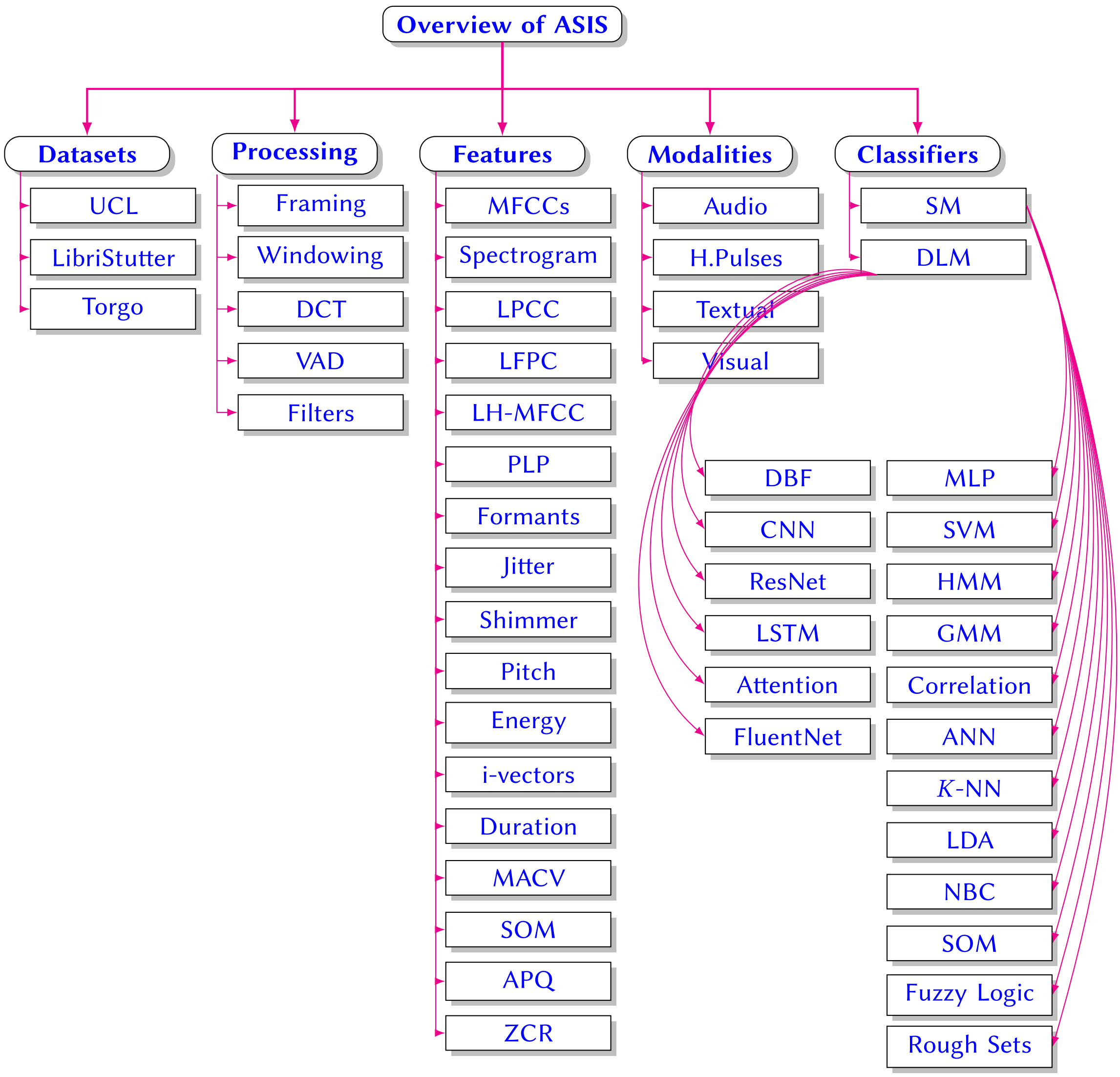

Fig. 3. Overview of Automatic Stuttering Identification Systems
DLM: Deep Learning Models SM: Statistical Models
MLP: Multi Layer Perceptron, SVM : Support Vector Machines
CNN: Convoluitonal Neural Network, HMM: Hidden Markow Models
RNN: Recurrent Neural Network, GMM: Gaussian Mixture Models
LSTM: Long Short Term Memory, LDA: Linear Discriminant Analysis
DBF: Deep Belief Neural Network, NBC: Naive Bayes Classifier
SOM: Self Organizing Maps
Features
MFCCs: Mel-Frequency Cepstral Coefficients
LPCC: Linear Prediction Cepstral Coefficients
LFPC: Log Frequency Power Coefficients
GFCC: Gammatone Frequency Cepstral Coefficients
DCT: Discrete Cosine Transforms
VAD: Voice Activity Detection

use of spectrograms (as a gray scale image) as sole feature extractors for stutter recognition and thus makes it suitable for the convolutional neural networks.

Different speech parameterization methods have their own benefits and drawbacks. P. Mahesha et al in [54] compared LPC, LPCC and MFCC for syllable repetition, word repition and prolongation and showed that LPCC based multi-class SVM (92% acc.) outperforms LPCC (75% acc) and MFCC(88% acc) based SVM stutter recognition models. In [35], M. Hariharan et al. discussed the effect of LPC, LPCC, and WLPCC features for stuttering (repetition and prolongation only) recognition events. They also discussed the effect of frame length and percentage of frame overlapping on stuttering recognition models (SRM). The authors conclude that the WLPCC feature based SRM outperforms LPC and LPCC. C.Y.Fook et al in [28] compared and analyzed the effect of LPC, LPCC, WLPCC and PLP features on the repetition and prolongation type of disfluencies and it has been shown that the MFCC feature based stuttering recognition models surpass the LPC, LPCC and WLPCC based ones. [5] used LPC and MFCCs as input features and concluded that MFCCs performs better than LPCs. [3] performs comparative study of LPCC and MFCC features in repetition and propagating stuttering and reports that LPCCs based ASIS outperforms MFFCs based ASIS slightly in varying frame length and frame overlapping. The optimal results of 94.51% and 92.55% accuracy on 21 LPCC & 25 MFCC coefficients respectively have been reported [3]. This can be due to the possibility of LPCCs are potential in capturing the salient cues from stuttering [3]. The use of spectrograms showed state-of-the-art performance in recognising the stuttering events [50]. The work by [50] didn't focus on the blocks and multiple stuttering types if present in a speech segment.

Machine Learning Classifiers: Stuttering detection systems process and classify underlying stuttering embedded speech segments. Including traditional classifiers, many statistical machine learning techniques have been explored in the automatic detection of stuttering. However, the studies are empirical, so there is no generally accepted technique that can be used. Table 3 lists chronologically the summary of stuttering classifiers including datasets, features, modality and stuttering type.

In ASIS, typically classification algorithms are used. A classification algorithm approximates the input X and maps it to output Y by learning procedure, which is then used to infer the class of new instance. The learning classifier requires annotated data for training which discerns the samples and their corresponding labels/classes. Once the training is finished, the performance of the classifier is evaluated on the remaining test data.

The traditional classifiers that have been explored ASIS include Support Vector Machines (SVM), Hidden Markov Models (HMM), Decision Trees (DT), Perceptron, Multi Layer Perceptrons (MLP), Gaussian Mixture Models (GMMs), k-Nearest Neighbor (k-NN), Naive Bayes Classifier (NBC), Rough Sets, Kohonen Maps (Self Organizing Maps (SOM)), Linear Discriminant Analysis (LDA), ANN (Artificial Neural Networks), Correlation.

Hidden Markov Models. HMMs lie at the heart of all contemporary speech recognition systems and has been successfully extended to disfluency classification systems. A simple and effective frame-work is provided by HMMs for modelling temporal sequences. Wisniewski et al [97] used euclidean distance as a codebook based on 20 MFCCs with HMMs. They reported an average recognition rate of 70% for 2 stuttering classes including blocks and prolongation with deleted silence and 60 frames of window length. T-S. Tan et al used 12 MFCC features with HMMs. The average recognition rate is 93% [89]. This tool recognizes only normal and stutter utterances and is

[8] Modality Considered: Audio Only
[9] Modality Considered: Audio and Textual
[10] Modality Considered: Audio, Visual and Textual
[11] Modality Considered: Bio-Respiratory Signals

Author and Year	Datasets	Features	Stutter Type	Model
Howell et al.[40](95)[8]	6 Speakers	EP, ACF-SC	(P),(R)	ANN
Howell et al.[41, 42](97)[9]	12 Speakers	Energy peaks, Duration	NA	ANN
Noth et al.[61](00)[8]	Northwind and Sun 37 Stutters, 16 Non-Stutters	Disfluent Frequency, Speaking rate, Duration	NA	HMMs
Geetha et al.[29](00)[9]	51 Stutters	Gender, Age Duration, Speech Rate	NA	ANNs
Czyzewski et al.[20](03)[8]	6-Normal, 6-SG Samples	Formants(1^{st} to 3^{rd}), Amplitude	(P),(R),(SG)	Rough Sets ANNs
Suszyński et al.[85] (03)[8]	NA	FFT	(P)	Fuzzy Logic
Śzczurowska et al.[88](06)[8]	8 PWS	FFT 512 Spectral Measure	(B)	MLP and SOM
M. Wisniewski et al[97](07)[8]	30 samples	MFCCs	NA	HMMs
TS Tan et al[89](07)[9]	UTM Skudai 10 Speakers (7M, 3F)	MFCCs	NA	HMMs
Ravi Kumar et al.[71](08)[8]	10 PWS	MFCCs, DTW for Score Matching	(SR)	Perceptron
Świetlicka et al.[86](09)[8]	8 PWS (Aged 10-23) 4 Fluent (2M, 2F)	FFT 512 Spectral Measure	NA	Kohonen based ML Kohonen based RBF
L.S. Chee [15](09)[8]	UCLASS	MFCCs	(R), (P)	*k*-NN, LDA
L.S. Chee [14](2009)[8]	UCLASS	LPCCs	(R), (P)	*k*-NN, LDA
Ravi Kumar et al.[70](09)[8]	15 PWS	MFCCs, DTW for score matching	(SR)	SVM
Yildirim et al.[101](09)[10]	10 CWS(Aged 4-6)	Duration, Pitch, Energy, Gestural, Linguistic	(R),(FS), (FP),(RP)	NBC
Pálfy et al.[65](11)[8]	UCLASS	MFCCs	(R)	SVM(Linear Kernel) SVM(RBF Kernel)
P. Mahesha et al.[54](13)[8]	UCLASS	LPCC, MFCC	(P),(WR),(SR)	SVM
Świetlicka et al.[87](13)[8]	19 PWS	FFT(512)	(B),(P), (SR)	Hierarchical ANN
Oue et al.[64](15)[8]	TORGO	MFCCs, LPCCs	(R)	DBN
P.Mahesha et al.[55](17)[8]	UCLASS	LH-MFCC	(P),(R),(I)	GMMs
Esmaili et al.[26](17)[8]	UCLASS	PLP	(P)	Correlation
Esmaili et al.[26](17)[8] Esmaili et al.[26](17)[8]	UCLASS Persian	WPT with entropy WPT with entropy	(P),(R) (P),(R)	SVM SVM
SA Ghonem et al.[30](17)[8]	UCLASS	I-Vectors	(R),(P), (RP)	*k*-NN, LDA
Santoso et al.[79](19)[8]	UUDB, PASD	Modulation Spectrum (Speech Rythm)	NA	BiLSTM
Santoso et al.[78](19)[8]	UUDB, PASD	Modulation Spectrum (Speech Rythm)	NA	BiLSTM + Attention
B. Villegas et al. [94](2019)[11]	69 Participants	Heart Rate Respiratory Air Volume Respiratory Air Flow	(B)	MLP
T. Kourkounakis et al.[50](20)[8]	UCLASS	Spectrograms	(WR),(I),(P), (SR),(RP), (FS)	ResNet + BiLSTM
T. Kourkounakis at al.[51](20)[8]	UCLASS, LibriStutter	Spectrograms	(WR),(I),(P), (SR),(R)	FluentNet
Shakeel et al.[81](21)[8]	UCLASS 128 PWS	MFCCs	(B),(P),(R),(F)	StutterNet

Table 3. Summary of several ASIS Systems in chronological order

not classifying different types of disfluencies. In 2000, Nöth et al. [61] used speech recognition system to evaluate the stuttering severity. This system can perform statistical counting and classification of three different types of disfluencies including repetition, pauses, and phoneme duration. Frequency of disfluent segments, speaking rate and disfluent durations are the measurable factors used to evaluate the stuttering severity during therapy sessions [14]

Support Vector Machines. SVMs gained substantial attention, have been widely used in the area of speech domain. SVM is a linear classifier that separates the data samples into its corresponding classes by creating a line or hyperplane. Mahesha et al.[54] used multiclass SVM to classify three stuttering disfluencies including prolongations, word repetitions and syllable repetitions. In this study, the different acoustic features including 12 LPC, LPCC and MFCCs are used. 75% average accuracy is obtained for LPC based SVM, whereas LPCC based SVM is 92% and for MFCCs based SVM is 88% [54]. K Ravikumar et al. used SVM to classify one disfluency type which is syllable repetitions [70]. The features used in [70] are MFCCs and DTW for score matching. An average accuracy of 94.35% is obtained on syllable repetitions. Pálfy et al. used SVM with two different kernel functions including linear and RBF [65]. In this case study, they used 16 audio samples from UCLASS [39] with eight males and eight females. 22 MFCC acoustic features with hamming window (25ms) with an overlap of 10ms are used in this case study [65]. 96.4% is the best recognition rate that has been reported with SVM when RBF is used as a kernel function [65]. With linear kernel based SVM, recognition rate is 98% [65]. I. Esmaili et al [26] used PLP features with a hamming window of 30ms and an overlap of 20ms to detect the prolongation type of stuttering based on correlation similarity measure between successive frames. 99% and 97.1% is the best accuracy that has been reported on UCLASS and persian datasets respectively [26]. In the same study they also evaluated the WPT+entorpy feature based SVM on UCLASS and persian stuttering datasets with 99% and 93.5% accuracies respectively [26].

Artificial Neural Networks (ANNs). They consist of several connected computing neurons that loosely model the biological neurons [31]. Like the synapses in biological neuron, each neuron can transmit a signal to other neurons via connections. A neuron receives a signal, processes it and can transmit signal to other connected neurons. The connections have weights associated with it which adjusts the learning procedure [31]. ANNs are trained by processing examples that maps input to its corresponding result by forming probability-weighted associations between the two. The training is conducted with the help of backpropagattion by optimizing the loss function by computing the error difference between the predicted output and its corresponding ground truth. Continuous weight adaptations will cause the ANNs to produce the similar output as the ground truth. After adequate number of weight adjustments, the training can be terminated once the optimization criteria is reached [31]. ANNs are essential tools both in the speech and speaker recognition. In recent times, ANNs play important roles in identifying and classifying the stuttering speech. P.Howell et al. used 2 ANNs for repetition and prolongation recognition [40]. The neural net is trained with 20 ACF, 19 vocoder coefficients of 10ms frame length and also with 20 frames of envelope coefficients. The networks are trained for with just 2 minutes of speech. The best accuracies of 82% and 77% are obtained for prolongations and repetitions when envelope parameters are used as an input features to ANNs [40]. ACF-SC based ANNs gives the best accuracy of 79% and 71% for prolongations and repetitions respectively [40]. P. Howell et al. [41, 42] designed a two stage recognizer for the detection of two types of disfluencies including prolongation and repetitions. The speech is segmented into linguistic units and then classified into its constituent category. The network is trained with the input features duration and energy peaks on a dataset of 12 speakers [41, 42]. The average accuracy on prolongations and repetitions obtained in this case study is 78.01% [41, 42]. Geetha et al. studied ANNs on 51 speakers to differentiate between stuttering children and normal disfluent children based on the features including disfluent type, rate of speech, disfluency duration, gender, age, family history and behavioral score [29]. They reported a classification accuracy of 92% [29]. I. Szczurowska et al. [88] used Kohonen based MLP to differentiate between non-fluent and fluent utterances. 76.67% accuracy has been reported on *blocks and stopped consonant repetition* disfluency types [88]. The Kohonen or Self Organizing Maps (SOM) are used first to reduce the feature dimensions of FFT 512 (with 21 digital 1/3-octave filters and a frame length of 23ms) input features, that later acts as an input to the MLP classifier. The model is trained on 8 PWS [88]. K Ravikumar et al. [71] proposed an automatic method by training a perceptron classifier for syllable repetition type of disfuency on 10 PWS with 12 MFCCs and DTW as the feature extraction methods. The best accuracy obtained for syllable repetition is 83% [71]. In 2003, Czyzewski et al. [20] addressed the stuttering problem by the help of stop-gaps detection, identification of syllable repetitions, detecting vowel prolongations. They applied ANNs and rough sets to recognize the stuttering utterances on the dataset of 6 fluent and 6 stop-gap based speech samples [20]. They reported that the average prediction accuracy of ANNs is 73.25% and rough-sets yielded an average accuracies of 96.67%, 90.00%, 91.67% on prolongations, repetitions and stop-gaps respectively [20]. W. Suszyński et al. [85] proposed a fuzzy logic based model for the detection and duration of prolongation type of disfluency. They used Sound Blaster card with a sampling frequency of 22KHz. 21 1/3 octave frequency bands with A filter and FFT features are used with the hamming window of 20ms. The features representing the prolongations are described by the use of fuzzy sets. Only the disfluent fricatives and nasals are considered in this study [85]. I. Świetlicka et al. [86] proposed an automatic recognition of prolongation type of stuttering by proposing Kohonen based MLP and RBF. From a dataset of 8 PWS and 4 fluent speakers, 118 (59 disfluent, 59 fluent), 118 total speech samples are recorded for the analysis. 21 1/3 octave filters with frequencies ranging from 100Hz to 10000Hz are used to parametrize the speech samples [86]. The parametrized speech samples are used as an input features to the Kohonen network that is expected to model the speech perception process. Thus, Kohonen is used to reduce the input dimensionality to extract salient features. These salient features are then fed to the MLP and RBF classifiers that are expected to model the cerebral processes, responsible for speech classification and recognition [86]. The method yielded a classification accuracy of 92% for Kohonen based MLP and 91% for Kohonen based RBF [86].

B. Villegas et al. in [94] introduced a respiratory bio-signals based stuttering classification method. They used respiratory patterns (air volume) and pulse rate as an input features to MLP. The dataset, developed at Pontifical Catholic University of Peru consists of 68 Latin American Spanish speaking participants with 27 PWS (aged 18-27 with mean of 24.4±5 years), 33 normal (aged 21-30 with mean of 24.3±2.3 years). The stuttering type studied in this research work is blocks with an accuracy of 82.6% [94].

In 2013, P. Mahesha et al. introduced a new Linear Prediction-Hilbert transform based MFCC (LH-MFCC) human perception feature extraction technique to capture the temporal, instantaneous amplitude and frequency characteristics of speech [55]. The study compares the MFCC and LH-MFFC features for three types of disfluencies including prolongation, repetition and interjection in combination with 64 Gaussian Mixture Models (GMM) components and reports a gain of 1.79% in average accuracy [55] with LH-MFCCs. The proposed LH-MFCC improves discriminatory ability in all classification experiments [55].

K-Nearest Neighbor and Linear Discriminant Analysis. **K**-NN, proposed by Thomas Cover is a non parametric model that can be used for both classification and regression. In **k**-NN classification, the output is described by the class membership and a sample is classified by the contribution of its neighbors. The sample is assigned to the class which is most common among its *k* $(k \geq 0)$ neighbors. This method relies on the distance metric for classification [57]

Linear discriminant analysis (LDA) also called normal discriminant analysis (NDA), or discriminant function analysis is a technique used in statistics and machine learning, to find a linear combination of features that separates two or more classes of samples. The resulting combination dimensionality reduction before classification or may be used as a linear classifier as well [57].

Chee et al. [15] presented an MFCC feature based *k*-NN and LDA classification models for repetition and prolongation types of disfluencies. The proposed models reports the best average accuracies of 90.91% for *k*-NN (with *k*=1) and 90.91% for LDA [15] on UCLASS [39] dataset. In 2009, Chee et al. [14] studied the effectiveness of LPCC features in prolongation and repetition detection with *k*-NN and LDA classfiers. The work achieved an average accuracy of 87.5% and the best average accuracy of 89.77% for LDA and *k*-NN respectively on the UCLASS [39] dataset. In 2017, SA Ghonem et al [30] introduced an I-vector (commonly used in speaker verification) feature based stuttering classification with *k*-NN and LDA methods. The technique reported an average accuracy of 52.9% among normal, repetition, prolongation and repetition-prolongation[12] stuttering events [30]. This is the first technique so far that has taken two disfluencies (occurring at same time) into consideration.

In 2009, S. Yildrim et al [101] proposed the first multi-modal disfluency boundaries detection model in spontaneous speech based on audio and visual modalities. The dataset used in this study was collected using Wizard of Oz (WoZ) tool. Audio recordings of high-quality were collected using a desktop microphone at 44.1 kHz. Two SonyTRV330 digital cameras, one focused from the front and the other capturing the child and the computer screen from the side were also used [101]. Three different classifiers including *k*-NN, Naive Bayes Classfier (NBC) and logistic model trees (LMT) have been utilised to evaluate the effectiveness of multi modal features on the collected dataset [101]. The stuttering types included in this case study are repetition, repair, false start and filled pauses [101]. In this work, the combination of three different modality based features including prosodic (duration, pitch and energy), lexical (hidden event posteriors) and gestural (optical flow) features were studied at feature level and decision level integration [101]. The work achieved the best accuracy for NBC among the three classifiers [101] and reports an accuracy of 80.5% and 82.1% at feature level integration and decision level feature integration respectively [101].

In 2005, Oue et al. [64] introduced deep belief network for the automatic detection of repetitions, non-speech disfluencies. 45 MFCC and 14 LPCC features from TORGO dataset [76] has been used in this case study for the detection of disfluencies [76]. The experimental results obtained showed that MFCCs and LPCCs produce similar detection accuracies of approximately 86% for repetitions and 84% for non-speech disfluencies [64].

The majority of statistical machine learning ASIS systems detailed above mostly focused only on either *prolongation* or *repetition* types of disfluencies with the most widely used features as MFCCs. Among the statistical techniques mentioned above, SVMs is the most widely used classifier in stuttering detection and identification.

2.2 Deep Learning Approaches

The majority of the state-of-the-art deep learning techniques combines several non-linear hidden layers as it can also reach to hundreds of layers as well, while a traditional ANNs consists of only one or two hidden layers. With the advancement in deep learning technology, the improvement in speech domain surpasses the traditional machine learning algorithms and hence the research in speech domain shifts towards the deep learning based framework and stuttering detection is

[12]repetition and prolongation disfluencies appearing at the same time

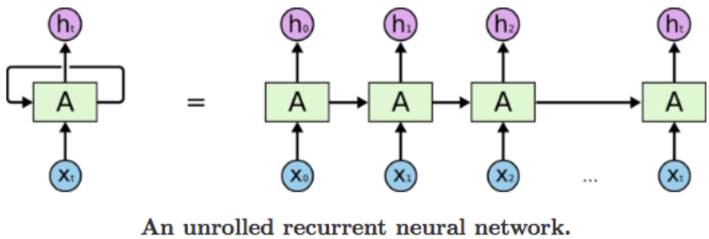

Fig. 4. Vanilla RNNs (Need to Add Link here for figure reference) [62]

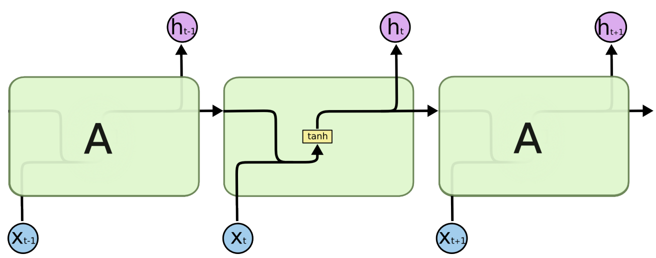

Fig. 5. RNN Processing Unit [62]

no exception. The salient advantage of these deep networks is automatic feature selection and extraction which avoids the cumbersome and tedious work of manual feature engineering step. The goal of these deep architecture classifiers is to approximate a mapping function f with $\mathbf{y} = f(\mathbf{X}; \theta)$ from input samples $\mathbf{X}$ to target labels $\mathbf{y}$ by adjusting its parameters θ. The most common deep learning architectures used in ASIS research domain are Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN).

Recurrent Neural Networks (RNNs). RNNs belong to a family of deep neural architectures where connections between neurons/nodes form a directed graph along a temporal sequence, thus allowing it to show temporal dynamic behaviour. RNNs consists of internal state (memory) that is used to process variable length input sequence. This structure makes RNNs good for modelling sequential tasks like time series, connected handwriting, video or speech recognition [31]. The other networks process inputs which are independent of each other, but in RNNs, inputs are related to each other as shown in Figures 4 and 5.

Initially, the RNN outputs h_0 by taking the first time step X_0 from the input sequence. This output h_0 together with X_1 is the input for the next time step producing h_1 as an output. Similarly, this h_1 together with X_2 will be the input for the next time step and so on which enables the RNNs to save the context while training [62]. The current hidden state can be computed by

$$h_t = f(h_{t-1}, X_t) \tag{1}$$

where f is any non-linear activation function like sigmoid, tanh, relu or softmax.

With the given input sequence $\mathbf{x} = (x_1, x_2, ...x_T)$, a vanilla RNN with the help of hidden vector sequence $\mathbf{h} = (h_1, h_2, ...h_T)$ computes the output vector sequence $\mathbf{y} = (y_1, y_2, ...y_T)$ by the following below mentioned equations in an iterative fashion from time $t = 1$ to T.

$$h_t = \psi(W_{xh}x_t + W_{hh}h_{t-1} + b_h) \tag{2}$$

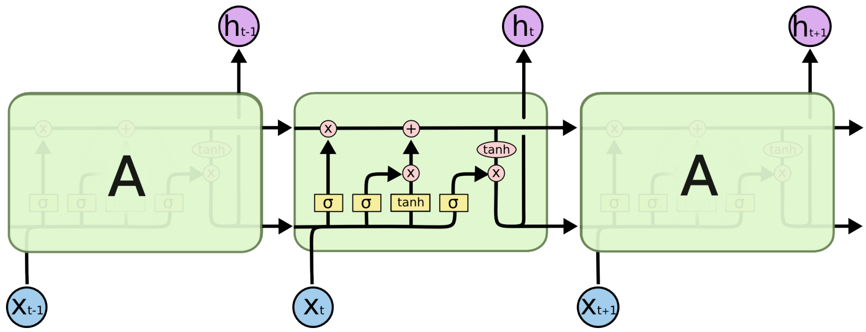

Fig. 6. LSTM Network [62]

$$y_t = (W_{hy}h_t + b_y) \tag{3}$$

where $W's$ are the corresponding shared weight matrices (W_{xh} corresponds to input-hidden weight matrix, W_{hh} corresponds to hidden-hidden weight matrix and W_{hy} is the corresponding hidden-output weight matrix) across the network. The ψ is the hidden non-linear activation function [32].

RNNs are believed with the concept that they might connect previous input information to the current one, such as using previous speech frames might assist in comprehending the current or future speech frames. At times, only a recent information is required to perform the current task. Consider an example of a language model that is trying to predict the next word in the sequence *"the sky is ..."*. It is pretty obvious from the context that the next word would be *sky*. RNNs can be extremely useful in learning the past information where the gap between the context information and the current time-step is small. In theory, RNNs are capable of capturing the "long-term dependencies", however, as the gap increases, it becomes difficult for the RNNs to learn to connect the previous context in practice which makes RNNs short time memory networks [62].

Long Short Term Memory Networks (LSTMs) introduced by Hochreiter & Schmidhuber (1997), a special type of RNN that is capable of capturing the long term dependencies in the temporal sequence [62].

Like standard RNNs, LSTM also consists of repeating chain like neural network modules as shown in Figure 5. However, vanilla RNNs have a very simple structure say of single relu layer, but instead of single layer, standard LSTMs consists of four interacting layers [62]. Figures 6 and 7 describes the behaviour of equations of all gates in the LSTM cell unit. The LSTM memory cell consists of 3 extra gates including input, forget and output gate, accompanied by three different weight matrices namely $W, U,$ & C. First the forget gate decides which information is to be discarded, input gate decides which value from the input should be used to change the memory, and output gate finally decides what should be forwarded to the next hidden state. W weight matrix connects the recurrent previous and current hidden layers. Weight matrix U connects the inputs to the hidden layer and C_t is the hidden cell state that is computed using the combination of previous memory state C_{t-1} multiplied with forget gate f_t and the input gate i_t multiplied with the newly computed candidate hidden state $\tilde{C}_t$ [62].

One of the major shortcomings of conventional LSTMs is that they only make use of the previous context. However, in disfluent speech, the current speech frame not only depends on the previous frame but also on future frame as well, and there seems no justification not to utilize future context as well. Bidirectional LSTMs (BiLSTMs) carry out this processing of the input data sequence in both directions (forward and backward) with two different hidden layers, which later on, are combined

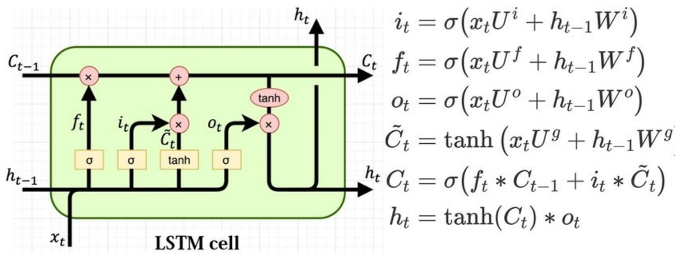

Fig. 7. LSTM Cell Structure with corresponding gate equations [92], *f, i, o and C* are respectively the forget gate, input gate, output gate and cell activation vectors, all of which are the same dimensions as hidden vector *h*.

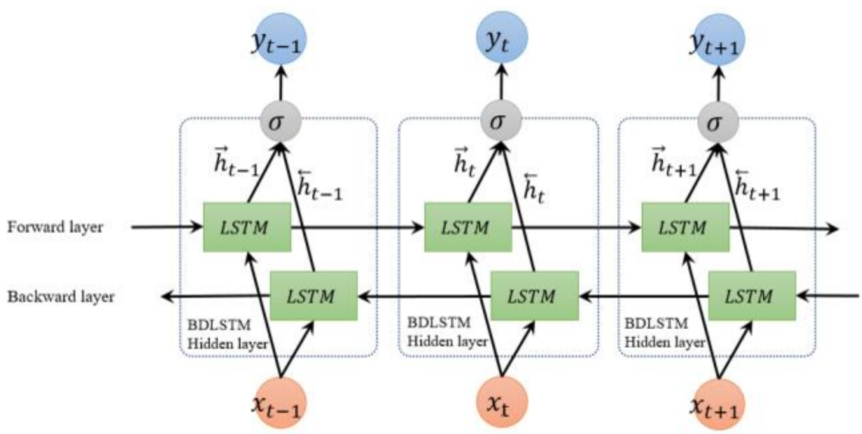

Fig. 8. BiLSTM Network [19]

to form a bidirectional context vector and then can be used for downstream tasks this [66]. As shown in Figure 8, BiLSTM produces the bidirectional context vector by computing the *forward* hidden sequence $\overrightarrow{h}$ from t = *1 to T* and the *backward* hidden sequence $\overleftarrow{h}$ from t = *T to 1*, thus updates the output layers according to:

$$\overrightarrow{h} = \psi(W_{x\overrightarrow{h}}x_t + W_{\overrightarrow{h}\overrightarrow{h}}\overrightarrow{h}_{t-1} + b_{\overrightarrow{h}}) \tag{4}$$

$$\overleftarrow{h} = \psi(W_{x\overleftarrow{h}}x_t + W_{\overleftarrow{h}\overleftarrow{h}}\overleftarrow{h}_{t+1} + b_{\overleftarrow{h}}) \tag{5}$$

$$y_t = W_{\overrightarrow{h_t}} + W_{\overleftarrow{h_t}} + b_y \tag{6}$$

The computed bidirectional context can either be used for output sequence problems or for simple classification by feeding it to other feedforward network. This can be done by taking the average of the context vector sequence or by simply using the last vector from the bidirectional context sequence [32].

In 2019, J. Santoso et al. [78] proposed modulation spectrum feature based BiLSTM to detect the causes of errors in speech recognition systems. The method is tested on the Japanese dataset of 20 speakers with 10 males and 10 females [78]. The experiment used 640-dimensional modulation spectrum feature vector with a block length of 320ms [78]. The method achieved an F-score of

0.381 for successfully detecting the stuttering events in the speech [78]. The proposed model used the overall utterance for the stuttering error detection, however recognition errors arise only from a small part of the full utterance. In order to address this issue, J. Santosa et al. [79] introduced attention based BiLSTM classifier for stuttering event detection. The best F-score of 0.691 is attained by taking the block length of 32ms [79].

Convolutional Neural Networks (CNN). CNNs are special type of neural nets that work with grid-structured data like images, audio spectrograms, video frames etc. A CNN consists of several layers in pipeline: convolution, pooling and fully-connected layers. With the help of several feature maps, CNNs are successful in capturing the spatial and temporal dependencies from the input data.

Convolution layer, a core component of the CNNs, is comprised of a set of learnable parametric kernels (filters) that transforms an input image into several number of small receptive fields [31]. In forward pass, a dot product is performed between the entries of an input image and filter resulting in an activate map of that filter [31]. This dot product is also known as convolution operation, defined by the following equation:

$$feature\ map = y[i, j] = input \circledast kernel = \sum \sum X[i - m, j - n].h[m, n] \tag{7}$$

where i, j indices related to image and m, n are concerned with the kernel , X represents the audio spectrogram or image matrix which is to be convolved with the filter h.

Due to parameter sharing of the convolutional operation, divergent feature or activation maps can be extracted, thus makes the CNNs translation invariance architectures [31]. Pooling, a down-sampling dimensionality reduction layer partitions the input matrix into a set of translational invariant non-overlapping combination of features. There are many methods to implement pooling operation, the most common among which is *average* pooling, computes the average value from each sub-region of the feature maps [31]. Fully connected (FC) layers, a global operation unlike convolution and pooling, usually used at the end of the network, connects every neuron in one layer to every neuron in another layer [31]. The FC layer takes the non-linear combination of selected features, which is later used for downstream tasks like classification [31].

Most of the existing work identify stuttering either by language models or by automatic speech recognition systems, which first converts the audio signals into its corresponding textual form, and then by the application of language models, detects or identifies stuttering. This procedure of stuttering identification seems a subsidiary computational step and could also be a potential source of error [50]. In order to address this, T. Kourkounakis et al. proposed a CNN based model to learn [50] stutter-related features. They approached this problem of stuttering identification by formulating it a binary classification problem, where they used the same architecture for identifying different types of stuttering. They used residual network to capture the disfluency-specific features from the spectrograms [50], that are the sole input features used in this study. Each audio speech sample is first segmented into several 4-second audio clips, and then annotated according to a specific type of stuttering present in these audio clips. The spectrogram features are extracted every $10ms$ on a window of $25ms$. The experimental analysis has been carried out on UCLASS release 1 dataset with a model architecture of 6 convolutional blocks, where each block is comprised of 3 convolutional layers, a total of 18 layered deep residual network [50]. The model is trained with batch norm and ReLu activation function [50]. In order to capture, the temporal aspect of the stuttered speech, the stuttering-specific learned representations by residual network are fed as an input to two recurrent BiLSTM layers [50]. Each BiLSTM layer consists of 512 BiLSTM units with a dropout rate of 0.2 and 0.4 is applied to recurrent layers respectively [50]. The proposed model reported an average accuracy of 91.15% and average miss rate of 10.03% (surpasses the state-of-the-art by almost 27%) on 6 different types of stuttering: revision, prolongation, interjection, phrase repetition, word repetition, sound repetition [50]. T. Kourkounakis et al. in [51] proposed a FluentNet as shown in Figure 10. that combines Squeeze-and-Excitation Residual Network (SE-ResNet) with BiLSTM networks, where SE-ResNet (8 blocks) is used to learn the stutter-specific spectral frame-level representations. Each audio speech is first segmented into 4 second audio clips, then acoustic features (spectrograms) are extracted, which are fed to SE-ResNet in order to capture stutter-specific spectral features, followed by a global attention based two layered BiLSTM (512 units) network, that helps in capturing effective temporal relationships [51]. The proposed model is trained uisng a root mean square propagation (RMSProp) optimizer on a binary cross entropy loss function with a dropout of 0.2 and a learning rate of 10^{-4}. In order to tackle the issue of stuttered speech data scarcity, they developed a synthetic stuttered speech dataset (LibriStutter) from a fluent LibriSpeech datatset [51]. The proposed FluentNet model reports an average accuracy of 91.75% and 86.7% on UCLASS and LibriStutter datasets respectively. Six different disfluency types are considered in this experimental study including phoneme repetition, word repetition, phrase repetition, interjection, prolongation, and revisions [51].

The stuttering identification methods discussed above consider only a small subset of disfluent speakers in their experimental studies, so it can not be said with certainty that the discussed models which performed very well on small speakers can also generalize to large set of stuttered speakers. In order to evaluate this, Shakeel et al. proposed a *StutterNet* [81], a time delay neural network based stuttering detection method shown in Figure 9. They addressed this problem by formulating it a multi-class classification problem. Only the core behaviours (blocks, repetition and prolongation) and fluent segments of the speech were considered in this case study. 128 speakers from the UCLASS dataset were used in this case study, thus makes it the first experimental study to be evaluated on the large set of disfluent speakers. Each audio sample is initially divided into 4-second audio segments, then acoustic features (MFCCs) are extracted, which are then fed to the *StutterNet*. The features are generated after every 12ms on a 25ms window for each 4 sec audio sample. On this larger set of disfluent speakers, they compared this study with the ResNet+BiLSTN [50] based ASIS system and reported an overall average accuracy of 50.79% and MCC of 0.23, in comparison to ResNer+BiLSTM based system comprising of 46.10% overall average accuracy and 0.21 MCC.

Among the DL based ASIS systems described above in detail, for a small set of disfluent speakers, the FluentNet classifier proposed by [51] and the spectrogram feature representations of stuttered speech are the most effective, that gives promising classification results on disfluency identification. However for a large set of stutterted speakers, *StutterNet* is the most effective one.

3 CHALLENGES & FUTURE DIRECTIONS

This section describes various challenges faced by ASIS systems and their possible solutions, which can be explored in the field of stuttering research.

Dataset Issue: Although there have been several developments in the automatic identification of disfluency, there are still several impediments that need to be addressed for robust and effective identification of stuttering. One of the most common barriers that needs to be addressed is the issue of scarcity of data on stuttering. There are only few natural stuttered datasets including UCLASS [39], TORGO [76] and a synthetic one LibriStutter, that has been made public recently [51].

A first difficulty related to data collection is the control of textual and linguistic content. Indeed, in order to make a fine analysis across several speakers, it is appropriate to have the same content (same list of sentences, for example). Unfortunately, in practice, when a PWS is asked to read a list of sentences, the disfluency effects are greatly reduced. For this reason, more spontaneous speech is used to hope to induce disfluencies. Moreover, depending on the speaker, the presence of disfluency in a recording is more or less important for several reasons: emotional state, speaking in public

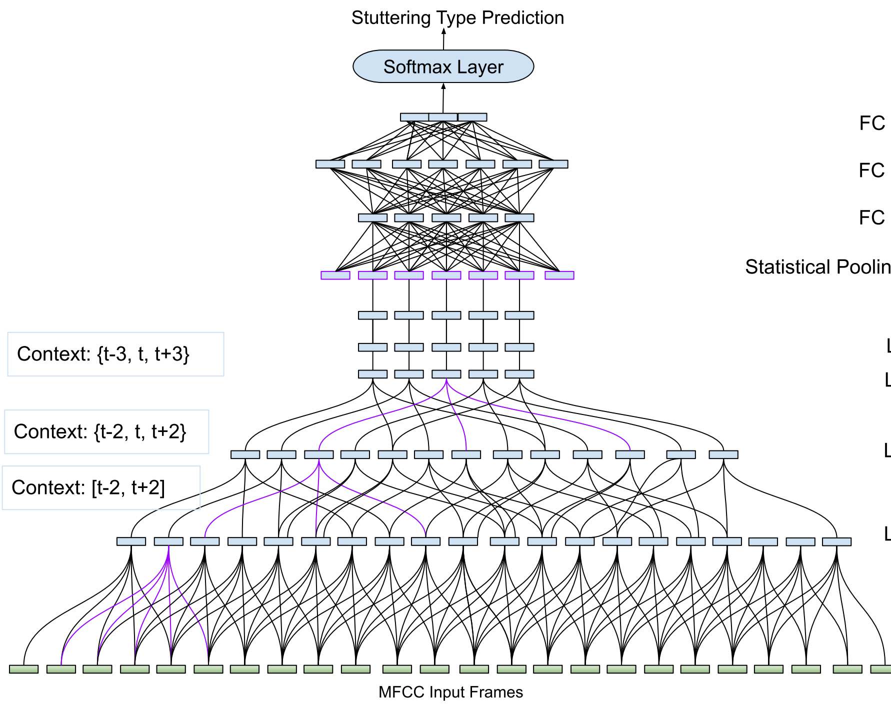

Fig. 9. StutterNet [81]

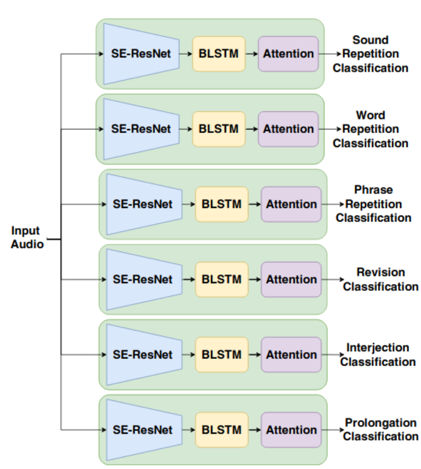

Fig. 10. FluentNet Model for Stuttering Classification [51]

or alone, spontaneous or read speech,etc. This makes the collection of a corpus difficult and its size from one speaker to another can be variable if one aims at having a comparable number of examples of disfluencies. Moreover, it is extremely difficult, if not impossible, to collect a corpus that contains the same number of examples of each type of disfluency. It is even more challenging to achieve high variability in gender, race, speaker, language and dialect. It should be noted that the recording of spontaneous speech must be well controlled to comply with the legislation. Indeed, it must be ensured that there is no personal information that can identify the speaker or that could be harmful to him in any way. Of course, we are not dealing with anonymization, as the voice could identify the speaker, but a minimum effort in this direction is required.

In order to identify stuttering using deep learning models, the data must be properly labelled and unbiased. Different background noises can corrupt the stuttered speech data. Likewise, the noise of recording equipments can also degrade the speech signal. Noise injection techniques can be exploited to learn reliable stutter-specific features from the noisy corrupted data. Deep Learning models like denoising auto encoders (DAEs), imputation AEs [52] can also be utilized to learn robust stutter-specific features from corrupted data. Training and testing data distribution mismatch is a significant challenge for noise robust ASIS systems. As mentioned above, the it is extremely difficult to get a stuttering dataset, because of its scarcity, so it is even harder to get the ample variability of gender, race, speaker, language and dialect in the annotated training speech.

Since stuttering datasets are scarce, we can attempt to solve this problem by enlarging the training data size and its diversity by generative models. Deep generative models like Generative Adversarial Networks (GANs) and Variational Autoencoders(VAE) etc., can be utilized for data augmentation [68] to generate more stuttered speech samples with the aim of improving the stuttering identification systems.

Data Annotation Issue: It is no doubt that DL has led to the enormous advancement in ASIS performance, nonetheless it demands a large amount of labelled data, and also, the dataset bias has plagued current ASIS methods. Annotating the stuttered speech requires expert speech pathologists/therapists, thus is expensive and laborious. Unsupervised learning (UL) enables to capture the underlying innate structure/pattern(s) in the data distribution [52]. In the context of stuttered speech, it can capitalize the unlabelled data to create understandings and learn good stutter specific feature representations, which later on, can be used to enhance the performance of ASIS systems in a supervised fashion. Semi-supervised learning (SSL) can also be exploited to solve this problem by employing unlabelled data, in conjunction with the annotated data to develop better classification models. Due to the unavailability of annotated and limited size of stuttering data, it becomes extremely difficult for the deep models to generalize. Self supervision, where the main idea is to find a proxy or pretext task for the deep models to learn without any explicit annotations, but rather, the data's innnate patterns provides the labels [82], is a compelling approach to address this paucity of stuttered data by capturing the innate compositions of the disfluency data.

Hand-Engineered Features: The another issue in the stuttering related speech domain is the need of hand-engineered features, which approximates the human auditory system. MFCCs are the principal set of hand-engineered acoustic features that have been used mainly for stuttering identification tasks. The main drawback of this approach is that by being manual it is cumbersome and requires human knowledge. Over the past few years in speech domain, the use of hand-engineered acoustic features is gradually changing and representation learning is acquiring recognition as an effective alternative to learn and capture task specific features directly from raw speech signals, thus circumvents the hand-engineered feature extraction module from the pre-processing pipeline [52]. In addition, Restricted Boltzmann Machines (RBMs) have shown to be successful and effective in learning hidden features from speech, and can learn more discriminate features when compared to MFCCs [52]. In [77] Sailor et al. have showed that Unsupervised Deep Auditory Model (UDAM) (stacked 2 convolutional RBMs) can learn human auditory processing relevant features like filterbanks from raw speech. This idea could be exploited to learn and capture the stuttered-specific features directly from the raw speech signal, which later on can be used for down stream tasks like classification, prediction etc.

Domain Adaptation: Additionally, most of the existing ASIS techniques proposed so far are neither language nor speaker invariant. This could be due to the fact that existing ASIS techniques depend on a probabilistic model to capture language and speaker specific factors, so that any alteration in the input speech distribution could have a significant impact (in terms of language or speakers) at the time of inference.

It is yet to be explored that, how well an ASIS technique performs across cross-language environment. There could be two possible scenarios of cross-language issue: the first is when the model is trained with a specific-language data, but tested in other languages; the other scenario could be, during training, a disfluent person registered in one language, but evaluated in a different language at the test time. Learning stutter-specific features that are invariant to variabilities in language, speakers, recording conditions, etc., could improve the performance of ASIS systems. These invaraint representations can be learned via various domain adaptation techniques [52].

Multi-Task Learning: One more issue with the ASIS systems is the generalization of trained models. Several techniques such as early stopping, regularization, dropout have been used to improve generalization [67]. The main drawback of these techniques is that they are limited by the identification/recognition task. This problem can be solved by the Multi-Task learning (MLT) startegy, i.e., if the model is forced to learn some auxiliary tasks in parallel in addition to its main task. Language classification and gender classification are two auxiliary tasks, that can be learned together with the stutter identification task on the same input feature space to improve generalization.

Multi-Modal Learning: In stuttering, identification, DL have been successfully applied to single modalities like text and audio. Inspired from the human brain, where the perceptions are carried out through the integration of information from several sensory organs including vision, hearing, smell etc., Ngiam et al. in [59] proposed a multi-modal (audio visual (AV)) learning and showed how to train deep models that learn effective shared representations across the modalities. The stuttering itself exhibits as an AV problem. Cues are present both in the visual (e.g. head nodding, lip tremors, quick eye blinks and unusual lip shapes) as well as in the audio modality [60]. This multi-modal learning paradigm from [59] could be helpful in learning robust stutter-specific hidden representations across the cross-modality platform, and could also help in building robust ASIS systems. Self supervised learning can also be exploited to capture acoustic stutter-specific representations based on guided video frames. As proposed in [82], this framework could be helpful in learning stutter-specific features from audio signal guided by visual frames or vice-versa.

Data Imbalance: Stuttering datasets also suffer from the data imbalance problems, i.e., the distribution of different disfluency categories is not uniform. The model trained on imbalanced dataset is biased toward the major classes. In order to address this problem, several techniques can be exploited, including resampling, reweighting and metric learning. Self supervision as proposed recently by Yang et al. [99] can also be used to address the problem of labeling bias effect in learning on imbalanced disfluent data.

Parallel Computation: One major issue with RNN based ASIS systems is that these sequential models can not be trained in parallel because the rnns are recursive in nature and the current hidden state depends on the previously computed hidden state [93]. In order to address this issue, transformers, originally, proposed in the context of neural machine translation (NMT), aims to eschew recursion as a means to allow parallel training [93], can be exploited instead of rnns, that would reduce training time of ASIS systems.

Multi-Stuttering Identification: One more issue with the proposed ASIS systems is that, they are not suitable for identifying a stuttering cluster (a stuttering cluster is defined when there is more than one stuttering type present in an utterance like *d-d-d—–dog dog is big*, consists of syllable repetition, prolongation and word repetition types of disfluencies [80]). To the best of our knowledge, Ghonem et al's work [30] is the only study, that has been carried out to identify a *repetition-prolongation* stuttering cluster. The ASIS systems require more robust techniques in order to detect and identify stuttering clusters.

4 CONCLUSION

Stuttering is a very complex disorder during which the flow of speech is interrupted by involuntary blocks, prolongations and repetitions. In the past two decades, a lot of research work has been performed in the automatic identification of disfluencies. In this paper, we give an up-to-date comprehensive review of the various datasets, acoustic features and ASIS classification models, that have been used by various researchers for the identification and recognition of stuttering disfluency. This paper also discussed several challenges with possible solutions that need to be addressed for future work. These ASIS systems demand the training data among which the most common dataset, that have been used in the stuttering research is UCLASS [39]. The audio speech is prepossessed to extract the acoustic features that replicate the human auditory system. The most common acoustic features that yield better results in the ASIS systems include MFCCs and spectrograms. The results can further be enhanced by appending features from other modalities, such as visual features.

Once the relevant acoustic features are extracted, ASIS systems have an extensive collection of classification techniques to select from. The majority of the classification models that are used for ASIS systems belong to statistical machine learning domain, however, there is an increasing surge towards the adoption of deep learning paradigm for ASIS system such as CNNs, LSTMS, attention networks.

Due to the challenges discussed in the section 3, ASIS systems are not yet available for real-time stutter identification, unlike SRS, that are easily accessible on portable mobile devices. To achieve this goal, ASIS systems demand more powerful models so that stuttering identification rate increases in cross language and cross speaker platforms with no labelled or very few annotated data.

REFERENCES

[1] Martin R Adams. 1974. A physiologic and aerodynamic interpretation of fluent and stuttered speech. *Journal of Fluency Disorders* 1, 1 (1974), 35–47.

[2] Martin R Adams. 1987. Voice onsets and segment durations of normal speakers and beginning stutterers. *Journal of Fluency Disorders* 12, 2 (1987), 133–139.

[3] Ooi Chia Ai, M Hariharan, Sazali Yaacob, and Lim Sin Chee. 2012. Classification of speech dysfluencies with MFCC and LPCC features. *Expert Systems with Applications* 39, 2 (2012), 2157–2165.

[4] Elena A Antipova, Suzanne C Purdy, Marjorie Blakeley, and Shelley Williams. 2008. Effects of altered auditory feedback (AAF) on stuttering frequency during monologue speech production. *Journal of fluency disorders* 33, 4 (2008), 274–290.

[5] KN Arjun, S Karthik, D Kamalnath, Pranavi Chanda, and Shikha Tripathi. 2020. Automatic Correction of Stutter in Disfluent Speech. *Procedia Computer Science* 171 (2020), 1363–1370.

[6] National Stuttering Association et al. 2009. The experience of people who stutter: A survey by the National Stuttering Association. *New York, NY: Author* (2009).

[7] Michel Belyk, Shelly Jo Kraft, and Steven Brown. 2015. Stuttering as a trait or state–an ALE meta-analysis of neuroimaging studies. *European Journal of Neuroscience* 41, 2 (2015), 275–284.

[8] Chitralekha Bhat, Biswajit Das, Bhavik Vachhani, and Sunil Kumar Kopparapu. 2018. Dysarthric Speech Recognition Using Time-delay Neural Network Based Denoising Autoencoder. In *Proc. Interspeech 2018*. 451–455. https://doi.org/10.21437/Interspeech.2018-1754

[9] M Blomgren, M Alqhazo, and E Metzger. 2012. Do speech sound characteristics really influence stuttering frequency. In *Proceedings of the 7th World Congress of Fluency Disorders, CD-ROM.*

[10] Gordon W Blood and Ingrid M Blood. 2016. Long-term consequences of childhood bullying in adults who stutter: Social anxiety, fear of negative evaluation, self-esteem, and satisfaction with life. *Journal of fluency disorders* 50 (2016),

72–84.

[11] Christian Büchel and Martin Sommer. 2004. What Causes Stuttering? *PLoS biology* 2 (03 2004), E46. https://doi.org/10.1371/journal.pbio.0020046

[12] H. M. Chandrashekar, Veena Karjigi, and N. Sreedevi. 2020. Spectro-Temporal Representation of Speech for Intelligibility Assessment of Dysarthria. *IEEE Journal of Selected Topics in Signal Processing* 14, 2 (2020), 390–399. https://doi.org/10.1109/JSTSP.2019.2949912

[13] Soo-Eun Chang, Ralph N Ohde, and Edward G Conture. 2002. Coarticulation and formant transition rate in young children who stutter. *Journal of Speech, Language, and Hearing Research* (2002).

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

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

[16] Lim Sin Chee, Ooi Chia Ai, and Sazali Yaacob. 2009. Overview of automatic stuttering recognition system. In *Proc. International Conference on Man-Machine Systems, no. October, Batu Ferringhi, Penang Malaysia*. 1–6.

[17] Edward G Conture, Gerald N McCall, and David W Brewer. 1977. Laryngeal behavior during stuttering. *Journal of Speech and Hearing Research* 20, 4 (1977), 661–668.

[18] Edward G Conture, Howard D Schwartz, and David W Brewer. 1985. Laryngeal behavior during stuttering: A further study. *Journal of Speech, Language, and Hearing Research* 28, 2 (1985), 233–240.

[19] Zhiyong Cui, Ruimin Ke, Ziyuan Pu, and Yinhai Wang. 2018. Deep bidirectional and unidirectional LSTM recurrent neural network for network-wide traffic speed prediction. *arXiv preprint arXiv:1801.02143* (2018).

[20] Andrzej Czyzewski, Andrzej Kaczmarek, and Bozena Kostek. 2003. Intelligent processing of stuttered speech. *Journal of Intelligent Information Systems* 21, 2 (2003), 143–171.

[21] Ali Dehqan, Fariba Yadegari, Michael Blomgren, and Ronald C Scherer. 2016. Formant transitions in the fluent speech of Farsi-speaking people who stutter. *Journal of Fluency Fisorders* 48 (2016), 1–15.

[22] Ivana Didirkova. 2016. *Parole, langues et disfluences: une étude linguistique et phonétique du bégaiement*. Ph.D. Dissertation. Université Paul Valéry-Montpellier III.

[23] Ivana Didirková and Fabrice Hirsch. 2020. A two-case study of coarticulation in stuttered speech. An articulatory approach. *Clinical linguistics & phonetics* 34, 6 (2020), 517–535.

[24] Ivana Didirkova, Sébastien Le Maguer, and Fabrice Hirsch. 2020. An articulatory study of differences and similarities between stuttered disfluencies and non-pathological disfluencies. *Clinical Linguistics & Phonetics* (2020), 1–21.

[25] Dennis Drayna and Changsoo Kang. 2011. Genetic approaches to understanding the causes of stuttering. *Journal of Neurodevelopmental Disorders* 3, 4 (2011), 374–380.

[26] Iman Esmaili, Nader Jafarnia Dabanloo, and Mansour Vali. 2017. An automatic prolongation detection approach in continuous speech with robustness against speaking rate variations. *Journal of Medical Signals and Sensors* 7, 1 (2017), 1.

[27] Andrew C Etchell, Oren Civier, Kirrie J Ballard, and Paul F Sowman. 2018. A systematic literature review of neuroimaging research on developmental stuttering between 1995 and 2016. *Journal of Fluency Disorders* 55 (2018), 6–45.

[28] Chong Yen Fook, Hariharan Muthusamy, Lim Sin Chee, Sazali Bin Yaacob, and Abdul Hamid Bin Adom. 2013. Comparison of speech parameterization techniques for the classification of speech disfluencies. *Turkish Journal of Electrical Engineering & Computer Sciences* 21, Sup. 1 (2013), 1983–1994.

[29] YV Geetha, Karanth Pratibha, Rao Ashok, and Shetty K Ravindra. 2000. Classification of childhood disfluencies using neural networks. *Journal of Fluency Disorders* 25, 2 (2000), 99–117.

[30] Samah A Ghonem, Sherif Abdou, Mahmoud A Esmael, and Nivin Ghamry. 2017. Classification of stuttering events using i-vector. *The Egyptian Journal of Language Engineering* 4, 1 (2017), 11–19.

[31] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. 2016. *Deep learning*. Vol. 1. MIT press Cambridge.

[32] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. 2013. Speech recognition with deep recurrent neural networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 6645–6649.

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

[34] Siddhant Gupta, Ankur T. Patil, Mirali Purohit, Mihir Parmar, Maitreya Patel, Hemant A. Patil, and Rodrigo Capobianco Guido. 2021. Residual Neural Network precisely quantifies dysarthria severity-level based on short-duration speech segments. *Neural Networks* 139 (2021), 105–117. https://doi.org/10.1016/j.neunet.2021.02.008

[35] Muthusamy Hariharan, Lim Sin Chee, Ooi Chia Ai, and Sazali Yaacob. 2012. Classification of speech dysfluencies using LPC based parameterization techniques. *Journal of Medical Systems* 36, 3 (2012), 1821–1830.

[36] M Hariharan, Vikneswaran Vijean, CY Fook, and Sazali Yaacob. 2012. Speech stuttering assessment using sample entropy and Least Square Support Vector Machine. In *Proc. 2012 IEEE 8th International Colloquium on Signal Processing and its Applications*. IEEE, 240–245.

[37] E Charles Healey and Peter R Ramig. 1986. Acoustic measures of stutterers' and nonstutterers' fluency in two speech contexts. *Journal of Speech, Language, and Hearing Research* 29, 3 (1986), 325–331.

[38] Robert E Hillman and Harvey R Gilbert. 1977. Voice onset time for voiceless stop consonants in the fluent reading of stutterers and nonstutterers. *The Journal of the Acoustical Society of America* 61, 2 (1977), 610–611.

[39] Peter Howell, Stephen Davis, and Jon Bartrip. 2009. The university college london archive of stuttered speech (uclass). *Journal of Speech, Language, and Hearing Research* (2009).

[40] Peter Howell and Stevie Sackin. 1995. Automatic recognition of repetitions and prolongations in stuttered speech. In *Proc. of the first World Congress on Fluency Disorders*, Vol. 2. University Press Nijmegen Nijmegen, The Netherlands, 372–374.

[41] Peter Howell, Stevie Sackin, and Kazan Glenn. 1997. Development of a two-stage procedure for the automatic recognition of dysfluencies in the speech of children who stutter: I. Psychometric procedures appropriate for selection of training material for lexical dysfluency classifiers. *Journal of Speech, Language, and Hearing Research* 40, 5 (1997), 1073–1084.

[42] Peter Howell, Stevie Sackin, and Kazan Glenn. 1997. Development of a two-stage procedure for the automatic recognition of dysfluencies in the speech of children who stutter: II. ANN recognition of repetitions and prolongations with supplied word segment markers. *Journal of Speech, Language, and Hearing Research* 40, 5 (1997), 1085–1096.

[43] Xuedong Huang, Alex Acero, Hsiao-Wuen Hon, and Raj Reddy. 2001. *Spoken language processing: A guide to theory, algorithm, and system development.* Prentice hall PTR.

[44] I Hurjui, S Pete, and I Bostan. 2016. Spatial distribution and the prevalence of speech disorders in the provinces of Iran. *Journal of Medicine and Life* 9, 1 (2016), 56.

[45] Lisa Iverach, Mark Jones, Lauren F McLellan, Heidi J Lyneham, Ross G Menzies, Mark Onslow, and Ronald M Rapee. 2016. Prevalence of anxiety disorders among children who stutter. *Journal of Fluency Disorders* 49 (2016), 13–28.

[46] M Jayaram. 1983. Phonetic influences on stuttering in monolingual and bilingual stutterers. *Journal of communication disorders* 16, 4 (1983), 287–297.

[47] Shweta Khara, Shailendra Singh, and Dharam Vir. 2018. A comparative study of the techniques for feature extraction and classification in stuttering. In *Proc. 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*. IEEE, 887–893.

[48] Joseph F Klein and Stephen B Hood. 2004. The impact of stuttering on employment opportunities and job performance. *Journal of fluency disorders* 29, 4 (2004), 255–273.

[49] Daniel Korzekwa, Roberto Barra-Chicote, Bozena Kostek, Thomas Drugman, and Mateusz Lajszczak. 2019. Interpretable deep learning model for the detection and reconstruction of dysarthric speech. *arXiv preprint arXiv:1907.04743* (2019).

[50] Tedd Kourkounakis, Amirhossein Hajavi, and Ali Etemad. 2020. Detecting Multiple Speech Disfluencies Using a Deep Residual Network with Bidirectional Long Short-Term Memory. In *Proc. ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6089–6093.

[51] Tedd Kourkounakis, Amirhossein Hajavi, and Ali Etemad. 2020. FluentNet: End-to-End Detection of Speech Disfluency with Deep Learning. *arXiv preprint arXiv:2009.11394* (2020).

[52] Siddique Latif, Rajib Rana, Sara Khalifa, Raja Jurdak, Junaid Qadir, and Björn W Schuller. 2020. Deep representation learning in speech processing: Challenges, recent advances, and future trends. *arXiv preprint arXiv:2001.00378* (2020).

[53] K López-de Ipiña, U Martinez-de Lizarduy, PM Calvo, B Beitia, J García-Melero, E Fernández, M Ecay-Torres, M Faundez-Zanuy, and P Sanz. 2018. On the analysis of speech and disfluencies for automatic detection of Mild Cognitive Impairment. *Neural Computing and Applications* (2018), 1–9.

[54] P Mahesha and DS Vinod. 2013. Classification of speech dysfluencies using speech parameterization techniques and multiclass SVM. In *Proc. International Conference on Heterogeneous Networking for Quality, Reliability, Security and Robustness*. Springer, 298–308.

[55] P Mahesha and DS Vinod. 2017. Lp-hillbert transform based mfcc for effective discrimination of stuttering dysfluencies. In *Proc. 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*. IEEE, 2561–2565.

[56] Juliette Millet and Neil Zeghidour. 2019. Learning to Detect Dysarthria from Raw Speech. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2019), 5831–5835.

[57] Kevin P Murphy. 2012. *Machine learning: a probabilistic perspective.* MIT press.

[58] Nicole E Neef, TN Linh Hoang, Andreas Neef, Walter Paulus, and Martin Sommer. 2015. Speech dynamics are coded in the left motor cortex in fluent speakers but not in adults who stutter. *Brain* 138, 3 (2015), 712–725.

[59] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. 2011. Multimodal deep learning. In *ICML*.

[60] NIDCD. 2015. Stuttering. (2015). :https://www.nidcd.nih.gov/health/stuttering/

[61] Elmar Nöth, Heinrich Niemann, Tino Haderlein, Michael Decher, Uwe Eysholdt, Frank Rosanowski, and Thomas Wittenberg. 2000. Automatic stuttering recognition using hidden Markov models. In *Proc. Sixth International Conference on Spoken Language Processing*.

[62] Christopher Olah. [n.d.]. Understanding LSTM Networks. http://colah.github.io/posts/2015-08-Understanding-LSTMs/. Accessed: 2020-12-24.

[63] World Health Organization et al. 1977. *Manual of the international statistical classification of diseases, injuries, and causes of death: based on the recommendations of the ninth revision conference, 1975, and adopted by the Twenty-ninth World Health Assembly.* World Health Organization.

[64] Stacey Oue, Ricard Marxer, and Frank Rudzicz. 2015. Automatic dysfluency detection in dysarthric speech using deep belief networks. In *Proc. of SLPAT 2015: 6th Workshop on Speech and Language Processing for Assistive Technologies.* 60–64.

[65] Juraj Pálfy and Jiří Pospíchal. 2011. Recognition of repetitions using support vector machines. In *Signal Processing Algorithms, Architectures, Arrangements, and Applications SPA 2011.* IEEE, 1–6.

[66] Paperswithcode. [n.d.]. Bidirectional LSTM. https://paperswithcode.com/method/bilstm. Accessed: 2020-12-24.

[67] Gueorgui Pironkov, Stephane Dupont, and Thierry Dutoit. 2016. Multi-task learning for speech recognition: an overview.. In *ESANN*.

[68] Yanmin Qian, Hu Hu, and Tian Tan. 2019. Data augmentation using generative adversarial networks for robust speech recognition. *Speech Communication* 114 (2019), 1–9.

[69] Nan Bernstein Ratner and Brian MacWhinney. 2018. Fluency Bank: A new resource for fluency research and practice. *Journal of fluency disorders* 56 (2018), 69.

[70] KM Ravikumar, R Rajagopal, and HC Nagaraj. 2009. An approach for objective assessment of stuttered speech using MFCC. In *Proc. The International Congress for Global Science and Technology.* 19.

[71] KM Ravikumar, Balakrishna Reddy, R Rajagopal, and H Nagaraj. 2008. Automatic detection of syllable repetition in read speech for objective assessment of stuttered disfluencies. *Proc. of World Academy of Science, Engineering and Technology* 36 (2008), 270–273.

[72] Naveeda Riaz, Stacy Steinberg, Jamil Ahmad, Anna Pluzhnikov, Sheikh Riazuddin, Nancy J Cox, and Dennis Drayna. 2005. Genomewide significant linkage to stuttering on chromosome 12. *The American Journal of Human Genetics* 76, 4 (2005), 647–651.

[73] Patricio Riva-Posse, Laura Busto-Marolt, Ángeles Schteinschnaider, Lucia Martinez-Echenique, Ángel Cammarota, and Marcelo Merello. 2008. Phenomenology of abnormal movements in stuttering. *Parkinsonism & related disorders* 14, 5 (2008), 415–419.

[74] Michael Robb, Michael Blomgren, and Yang Chen. 1998. Formant frequency fluctuation in stuttering and nonstuttering adults. *Journal of Fluency Disorders* 23, 1 (1998), 73–84.

[75] Patricia M Roberts, Ann Meltzer, and Joanne Wilding. 2009. Disfluencies in non-stuttering adults across sample lengths and topics. *Journal of communication disorders* 42, 6 (2009), 414–427.

[76] Frank Rudzicz, Aravind Kumar Namasivayam, and Talya Wolff. 2012. The TORGO database of acoustic and articulatory speech from speakers with dysarthria. *Journal of Language Resources and Evaluation* 46, 4 (2012), 523–541.

[77] Hardik B Sailor and Hemant A Patil. 2016. Unsupervised Deep Auditory Model Using Stack of Convolutional RBMs for Speech Recognition.. In *INTERSPEECH*. 3379–3383.

[78] Jennifer Santoso, Takeshi Yamada, and Shoji Makino. 2019. Categorizing error causes related to utterance characteristics in speech recognition. *Proc. NCSP* 19 (2019), 514–517.

[79] Jennifer Santoso, Takeshi Yamada, and Shoji Makino. 2019. Classification of causes of speech recognition errors using attention-based bidirectional long short-term memory and modulation spectrum. In *Proc. 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC).* IEEE, 302–306.

[80] Jean Sawyer. 2010. By the Numbers: Disfluency Analysis for Preschool Children who Stutter. In *Proc. International Stuttering Awareness Day Online Conference.*

[81] Shakeel Ahmad Sheikh, Md Sahidullah, Fabrice Hirsch, and Slim Ouni. 2021. StutterNet: Stuttering Detection Using Time Delay Neural Network. In *EUSIPCO 2021 – 29th European Signal Processing Conference.* Dublin, Ireland. https://hal.inria.fr/hal-03227223

[82] Abhinav Shukla, Konstantinos Vougioukas, Pingchuan Ma, Stavros Petridis, and Maja Pantic. 2020. Visually guided self supervised learning of speech representations. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 6299–6303.

[83] C Woodruff Starkweather. 1987. *Fluency and stuttering.* Prentice-Hall, Inc.

[84] Anu Subramanian, Ehud Yairi, and Ofer Amir. 2003. Second formant transitions in fluent speech of persistent and recovered preschool children who stutter. *Journal of Communication Disorders* 36, 1 (2003), 59–75.

[85] Waldemar Suszyński, Wiesława Kuniszyk-Jóźkowiak, Elżbieta Smołka, and Mariusz Dzieńkowski. 2015. Prolongation detection with application of fuzzy logic. *Annales Universitatis Mariae Curie-Sklodowska, sectio AI–Informatica* 1, 1 (2015), 1–8.

[86] Izabela Świetlicka, Wiesława Kuniszyk-Jóźkowiak, and Elżbieta Smołka. 2009. Artificial neural networks in the disabled speech analysis. In *Computer Recognition Systems 3*. Springer, 347–354.

[87] Izabela Świetlicka, Wiesława Kuniszyk-Jóźkowiak, and Elżbieta Smołka. 2013. Hierarchical ANN system for stuttering identification. *Computer Speech & Language* 27, 1 (2013), 228–242.

[88] Izabela Szczurowska, Wiesława Kuniszyk-Jóźkowiak, and Elżbieta Smołka. 2014. The application of Kohonen and Multilayer Perceptron Networks in the speech nonfluency analysis. *Archives of Acoustics* 31, 4 (S) (2014), 205–210.

[89] Tian-Swee Tan, AK Ariff, Chee-Ming Ting, Sh-Hussain Salleh, et al. 2007. Application of Malay speech technology in Malay speech therapy assistance tools. In *Proc. 2007 International Conference on Intelligent and Advanced Systems*. IEEE, 330–334.

[90] USA TODAY TECH. [n.d.]. For people who stutter, the convenience of voice assistant technology remains out of reach. https://eu.usatoday.com/story/tech/2020/01/06/voice-assistants-remain-out-reach-people-who-stutter/2749115001/. Accessed: 2020-12-24.

[91] Sarah Vanhoutte, Marjan Cosyns, Pieter van Mierlo, Katja Batens, Paul Corthals, Miet De Letter, John Van Borsel, and Patrick Santens. 2016. When will a stuttering moment occur? The determining role of speech motor preparation. *Neuropsychologia* 86 (2016), 93–102.

[92] Savvas Varsamopoulos, Koen Bertels, and Carmen G Almudever. 2018. Designing neural network based decoders for surface codes. *arXiv preprint arXiv:1811.12456* (2018).

[93] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*. 5998–6008.

[94] Bruno Villegas, Kevin M Flores, Kevin José Acuña, Kevin Pacheco-Barrios, and Dante Elias. 2019. A Novel Stuttering Disfluency Classification System Based on Respiratory Biosignals. In *Proc. 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 4660–4663.

[95] David Ward. 2008. *Stuttering and cluttering: frameworks for understanding and treatment*. Psychology Press.

[96] ME Wingate. 1969. Stuttering as phonetic transition defect. *Journal of Speech and Hearing Disorders* 34, 1 (1969), 107–108.

[97] Marek Wiśniewski, Wieslawa Kuniszyk-Jóźkowiak, Elzbieta Smołka, and Waldemar Suszyński. 2007. Automatic detection of disorders in a continuous speech with the hidden Markov models approach. In *Computer Recognition Systems 2*. Springer, 445–453.

[98] Ehud Yairi and Nicoline Ambrose. 2013. Epidemiology of stuttering: 21st century advances. *Journal of Fluency Fisorders* 38, 2 (2013), 66–87.

[99] Yuzhe Yang and Zhi Xu. 2020. Rethinking the value of labels for improving class-imbalanced learning. *arXiv preprint arXiv:2006.07529* (2020).

[100] J Scott Yaruss and Edward G Conture. 1993. F2 transitions during sound/syllable repetitions of children who stutter and predictions of stuttering chronicity. *Journal of Speech, Language, and Hearing Research* 36, 5 (1993), 883–896.

[101] Serdar Yildirim and Shrikanth Narayanan. 2009. Automatic detection of disfluency boundaries in spontaneous speech of children using audio–visual information. *IEEE Transactions on Audio, Speech, and Language Processing* 17, 1 (2009), 2–12.

[102] Patricia M Zebrowski, Edward G Conture, and Edward A Cudahy. 1985. Acoustic analysis of young stutterers' fluency: Preliminary observations. *Journal of Fluency Disorders* 10, 3 (1985), 173–192.