Machine Learning for Stuttering Identification: Review, Challenges & Future Directions

Shakeel Sheikh, Md Sahidullah, Fabrice Hirsch, Slim Ouni

Machine Learning for Stuttering Identification: Review, Challenges & 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 is a speech disorder during which the flow of speech is interrupted by involuntary pauses and repetition of sounds. Stuttering identification is an interesting interdisciplinary domain research problem which involves pathology, psychology, acoustics, and signal processing that makes it hard and complicated to detect. Recent developments in machine and deep learning have dramatically revolutionized speech domain, however minimal attention has been given to stuttering identification. This work fills the gap by trying to bring researchers together from interdisciplinary fields. In this paper, we review comprehensively acoustic features, statistical and deep learning based stuttering/disfluency classification methods. We also present several challenges and possible future directions.

CCS Concepts: • **Computing methodologies** → **Artificial intelligence**; **Machine learning**; **Natural language processing**.

Additional Key Words and Phrases: speech disfluency, stuttering detection, datasets, features, modality, machine learning, deep learning, challenges, future directions.

ACM Reference Format:
Shakeel A. Sheikh, Md Sahidullah, Fabrice Hirsch, and Slim Ouni. 2021. Machine Learning for Stuttering Identification: Review, Challenges & Future Directions. *ACM Comput. Surv.* 54, 4 (2021), 27 pages.

1 INTRODUCTION

Speech disorders or speech impairments are communication disorders in which a person has difficulties in creating and forming the normal speech sounds required to communicate with others [1, 2]. These disorders can take the form of dysarthria, apraxia, stuttering, cluttering, lisping, and so on [1–5].

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* 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 [6]. *Cluttering* is characterized by a patient's speech being too jerky, too rapid, or both. *Persons with cluttering* usually exclude/collapse most of the syllables, or aberrant rhythms or syllable stresses, and also contain excessive amounts of interjections such as *so*, hmm, *like*, *umm,* etc [6]. *Apraxia* is defined as a speech disorder when

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/-ART $15.00
https://doi.org/

the neural path between the nervous system and the muscles responsible for speech production is obscured or lost. The *persons with apraxia* 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 [6]. *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* [6]. Stuttering speech impairment is different than other speech disorders because it can be cured, if early intervention is being made to help the persons who stutter (PWS) develop normal fluency [1]. Of all these speech impairments, stuttering is the most common one[1].

Stuttering - also called stammering/disfluency[2] - is a neuro-developmental disorder that commences when neural connections supporting language, speech, and emotional functions are quickly changing [7], and is characterized by core behaviours which usually take the form of involuntary stoppages, repetition and prolongation of sounds, syllables, words or phrases. Stuttering can be described as an abnormally and persistent duration of stoppages in the normal forward flow of speech [1]. These speech abnormalities are generally accompanied by unusual behaviours like head nodding, lip tremors, quick eye blinks and unusual lip shapes, etc [8]. Fluency can be defined as the capacity to produce speech without any effort, at a normal rate [9]. 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 control movement is necessary to maintain fluency [10]. 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 [11]. 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, PWS know what they want to pronounce but are momentarily unable to produce it [12].

Stuttering can broadly be classified into two types [1]:

- *Developmental Stuttering:* This stuttering is the most common one and it usually occurs in the learning phase of the language, *i.e. between two and seven.* Recent researches conclude that developmental stuttering is a multifactorial trouble including neurological and genetic aspects [13, 14]. Indeed, fMRI studies show anomalies in neural activity before speech, *i.e.* during the planning stages of speech production [15]. Furthermore, an atypical activation in the left inferior frontal gyrus and right auditory regions [16, 17] has been highlighted. Concerning the genetic aspects, Riaz et al. [18] observe an unusual allele on chromosom 12 by PWS. Drayna and Kang [14] identify 87 genes which could be involved in stuttering, including one called GNPTAB, which was significantly present in 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 [19]. Even if neurogenic stuttering is rare, it can be observed in children and adults regardless of their age.

Globally, stuttering concerns 1% of the world's global population and its incidence rate is between 5% and 17% [20]. 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

[1]https://www.healthline.com/health/speech-disorders

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

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 [20]. As considered by the non-stuttering persons, the disfluency affects 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 people [21]. The PWS are usually rated less popular than their non-stuttering peers and less likely to be considered leaders [21]. According to national stuttering association [22], 40% of the PWS have been repudiated school opportunities, promotion or job offers and it affects relationships. The 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 [23]. All these elements explain that PWS develop social anxieties and negative feelings (fear, shame, etc.) when they have to speak in public [24].

Stuttering appears to be complex and mysterious. Several factors that lead to stuttering includes: stress, delayed childhood development, speech motor control abnormalities, as there is a strong correlation between stress, anxiety, and stuttering [1]. 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 [25]. In a recent study, Smith and Weber [7], postulated the mutifactorial dynamic pathways theory, where they asserted that the stuttering actually occurs because of the failure of central nervous system in generating the necessary patterns of motor commands for fluent speech. Thus, stuttering shows impairment in sensorimotor processes that are responsible for speech production, and its orientation throughout the life of PWS is strongly affected by linguistic and emotional aspects.

In conventional stuttering assessment, the speech language pathologists (SLP) or speech therapists (ST) manually analyze either the PWS' speech or their recordings [26]. The stuttering severity is usually measured by taking the ratio of disfluent words/duration to the total words/duration [1]. The most conventional speech therapy sessions involve helping the PWS observe and monitor their speech patterns in order to rectify them [1]. The speech therapeutic success rate recoveries have been reported to be 60-80% when dealt in early stage [27]. This convention of detecting stuttering severity and its improvement due to therapeutic sessions is very demanding and time consuming, and is also biased and prejudiced towards the subjective belief of SLPs. Due to the nature of stuttering, its therapeutic sessions are very intense course, that usually, extends to several months (several years in some cases), which necessitates PWS to see the SLP regularly [28]. Usually, the speech therapy sessions are private and are very expensive, thus makes it unaffordable to some PWS. Thus, it is important to develop interactive automatic stuttering detection systems.

The automatic speech recognition systems (ASR) 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. [29]. The stuttering detection may help in adapting the virtual assistant tools for the disfluent persons.

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, machine learning and deep learning, it became a reality to develop smart and interactive stuttering detection and therapy tools [30]. 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 2. 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 [31–33]. 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.

Different researchers have used different acoustic features and classifiers for stuttering detection. As such, there is no systematic study which mentions the detailed analysis and challenges of various stuttering detection methods. In this paper, we give an up-to-date comprehensive survey by unifying all the stuttering methods. We also give a basic summary of various acoustic characteristics of stuttering, that can be exploited in stuttering detection. This paper will be a valuable resource for the researchers, engineers and practitioners. This paper discusses some stuttering datasets that can be referenced by various researchers. Despite the recent advancements by deep learning (DL) in speech recognition [34], emotion detection [35], speech synthesis[36], etc, it has not been much used in the stuttering domain. Due to the data scarcity in stuttering domain, the DL has not been explored in a great extent in stuttering detection. We recently proposed *StutterNet* trained on small dataset, that shows promising results in stuttering detection. In this paper, we also present several challenges faced by various stuttering detection methods and their possible solutions to show that how we can efficiently use DL to boost stuttering detection from voice.

There are several modalities that are considered in stuttering detection, which include: speech [31], visual [37], text [38], bio-respiratory signals [39], and functional near-infrared spectroscopy [40, 41]. This paper focuses mainly on the speech modality. Most of the stuttering modalities are very expensive to collect besides speech modality, because that is cheap and can be collected remotely by a simple mobile application or a web interface.

The remainder of the paper is organised as follows. Section 2 discusses the various acoustics properties of stuttered speech. This section describes how the stuttered speech impacts the various characteristics like formant transitions, pitch, VOT, etc. Section 3 presents stuttering datasets that have been used in the ASIS. Section 4.1 describes various acoustic features like MFCC, LPCC, etc, that have been exploited in stuttering detection. This section also discusses all statistical machine learning methods, that have been used in various stuttering detection events. Section 4.2 describes, how deep learning can be used to model and detect various types of stuttering disfluencies. This section also provides some preliminary studies of DL in stuttering detection, and discusses the use of spectrograms in stuttering detection. Section 5 discusses various challenges that the current ASIS systems are facing. Among them, few are data scarcity, hand engineered features, cross domain, etc. This section also describes their possible solutions, that can be exploited to address the mentioned challenges, and, finally, the concluding remarks are provided in Section 6.

2 CHARACTERISTICS OF STUTTERED SPEECH

Most brain scan studies show that, during fluent speech and silent rest, there is no difference in cerebral activity between PWS and normal fluent speakers [42]. However, during stuttering, there is a dramatic change in cerebral activity. The right-hemisphere areas which are normally not active during normal speech becomes active, and the left-hemisphere areas, which are active during normal speech becomes less active [5, 42]. It has also been found that there is under-activity in

Stutter Type	Definition	Example
Blocks	Involuntary pause before a word	I w **blockage/pause** ant 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	**Ttttt-Ttttt-Ttttt**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

the central auditory processing area. In 1997, a study by Kehoe and Contributors [5], suggests that adult PWS have an inability to integrate somatic and auditory processing. A brain study by Foundas et al. [43] found that PWS have rightward asymmetry in planum temporale (PT), i.e, their right PT is larger than their left PT, on the contrary, normal people's PT is larger in the left side of their brains. Studies based on motor data have been carried out about stuttering. The PWS also show over-activity in the speech motor control area [5], in particular in the *left caudate nucleus* area. Conture et al. [44, 45] observe inappropriate vocal folds abductions and adductions which lead to anarchic openings and closure of the glottis. Concerning the supraglottic level, Wingate [46] hypothesizes that stuttering is not a trouble dealing with sound 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. However in a recent study, Didirková and Hirsch [47] show, thanks to EMA data, that stuttering is not a coarticulation trouble. They found correct coarticulatory patterns in the fluent and stuttered utterances. Furthermore, another study based on articulatory observation, notes 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 [48]. 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 [4].

Concerning stuttering-like disfluencies, several types have been observed: repetitions, blocks, prolongations, interjections, etc, which are detailed in Table 1. Some works try to link the locus of disfluencies and phonetic proprieties. Jayaram [49], Blomgren et al. [50], and Didirkova [51] indicate that unvoiced consonants are more disfluent than their voiced counterparts. Furthermore, Blomgren et al. [50] notices that disfluencies are more frequent at the beginning of an utterance or just after a silent pause. Moreover, Didirkova [51] observes an important inter-individual variability concerning sounds and/or phonetic features which are the most disfluent.

Acoustic analysis has been carried out about stuttering, including speech rate, stop-gap duration, vowel(V)-consonant(C) transition 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 [52].

Dehqan et al. [53] studied the second formant (F2) transitions of fluent segments of persian speaking PWS. They concluded that the F2 frequency extent transitions are greater in stuttering speakers than normal fluent ones. The PWS takes longer to reach vowel steady state, but the

overall F2 formant slopes are similar for both stuttering speakers and normal ones [53]. The PWS generally exhibit slower speaking rates when compared to normal speakers. Several other studies have investigated the CV formant transitions in stuttered speech. Yaruss and Conture [54] examined the F2 transitions of children who stutter on syllable repetitions, and found no aberrant F2 transitions. However, Robb et al. [55] analyzed the fluent speech segments of PWS, and showed that F2 fluctuations are longer for voiced and voiceless stops than normal speakers. In a different study by Chang et al. [56], 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 degree of CV coarticulation is no different among stuttering and control groups. Subramanian et al. [57] 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. Blomgren et al. [58] found that PWS and normal speaker in /CVt/ token exhibit no differences in the F1 (average). The stutters show significantly lower F2 in /Cit/ tokens than the control groups. The formant spacing for /i/ is significantly lower in PWS than fluent persons [58]. Hirsch et al. [59] conducted a study by analysing the first two formants (vowel space) in CV sequences between the stuttered and normal group. At a normal speaking rate, stuttering group shows reduction in the vowel space, in contrary to the fast speaking rate, where, the latter shows no noticeable deviation.

VOT is the duration of time between the release of a stop consonant and the beginning of vocal fold vibrations [1]. Healey and Ramig [60] showed that for voiceless stops, chronic stuttering exhibits longer VOT when compared with normal fluent 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* [60]. Hillman and Gilbert [61] also found that the PWS reveals longer VOT for voiceless stops than fluent persons. Adams [62] found that not only voiceless stops exhibits longer VOT in PWS, but also, voiced stops displays longer VOT than non-stuttering persons. No significant VOT differences have been found in control and PWS groups [63]. In another study by Jäncke [64], the PWS show strong variability in repeated production of VOT for voiceless stops, however, there is no significant difference between the two groups. In a study carried by De Nil and Brutten [65], it shows that the stuttering children exhibit more variability in VOT than their counterparts. Celeste and de Oliveira Martins-Reis [66] also found that the stuttering group shows higher VOT for voiceless stops. Brosch et al. [67] examined the VOT in stuttering children. They found that the children with severe stuttering have higher values of VOT. Borden et al. [68] examined the VOT of fluent utterances from PWS. Their study showed that the fluent utterances of PWS exhibit no statistical differences in VOT and are within the limits of normal groups.

Fosnot and Jun [69] examined the prosodic characteristics of PWS and fluent children. They found that the variability in pitch is greater in stuttered group, but slightly differ from the normal group. In another study, it has been shown that the normal group and PWS show same patterns in f0 deviation [70]. The stuttering occurs less significantly in low-pitched condition as compared to high-pitched condition [70]

3 DATASETS FOR STUTTERING DETECTION RESEARCH

Data is an indispensable component of any DL model. DL saves feature engineering costs by automatically generating relevant features, however require substantial amounts of annotated data. Most stuttering identification studies so far are based on in-house datsets [3, 30, 33, 71, 72] with limited speakers. In stuttering domain, there is a lack of datasets and several stuttering datasets that have been collected so far are discussed below:

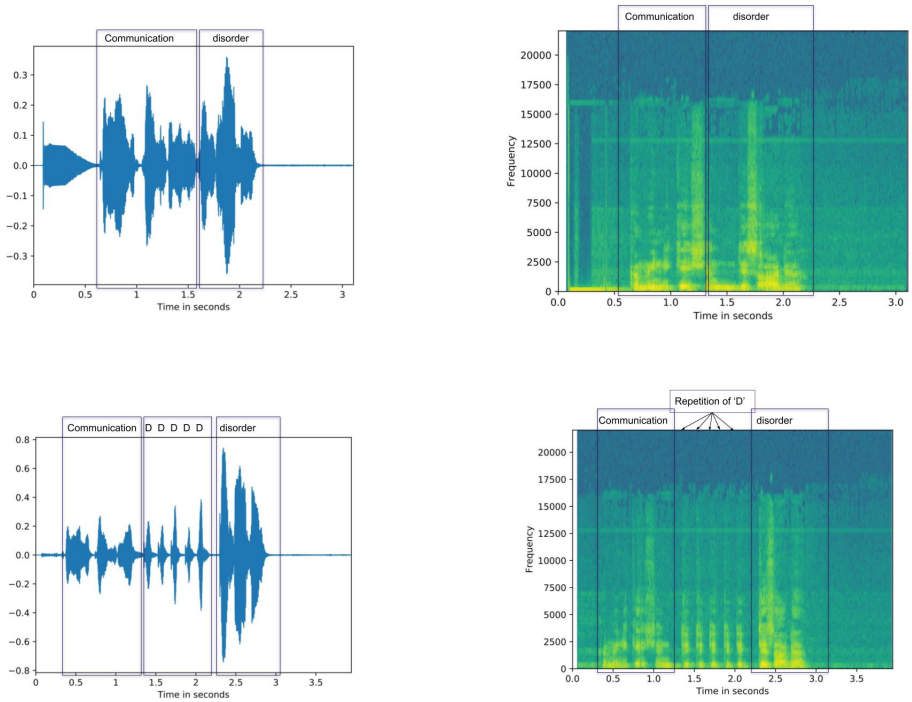

Fig. 1. Speech waveforms and spectrograms of a speaker (male) saying *Communication disorder*. The left is waveforms (amplitude v/s time); the right is a time–frequency plot using a wavelet decomposition of these data. Top row is fluent speech; bottom row is stuttering (repetitions), occur at the "D" in "disorder." Five repetitions can be clearly identified by arrows in the spectrogram. (bottom right)

.

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) [71] 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[3] release 1 contains 139 monologue samples from 81 PWS, aged from five to 47 years. The male samples are 121 and female samples are 18.

LibriStutter. The availability of small amount of labelled stuttered speech led to synthetic LibriStutter's creation [30]. The LibriStutter consists of 50 speakers (23 males and 27 females), and is approximately of 20 hours and includes synthetic stutters for repetitions, prolongations and interjections. For each spoken word, Kourkounakis et al. [30] used Google Cloud Speech-to-Text (GCSTT) API to generate timestamp correspondingly. Random stuttering was inserted within the four second window of each speech signal.

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

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 [72]. 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) [3]. The platform proposes audio and video files with transcriptions of adults and children who stutter. The FluencyBank is an interview data of 32 PWS.

SEP-28k. The public stuttering datasets are too small to build well generaizable ASISs. So in order to address this, Colin et al [33] recently curated a public version of *Stuttering Events in Podcasts (SEP-28k)* dataset. This dataset contains a total samples of 28,177. The SEP-28k dataset is the first publicly available annotated dataset with stuttering labels including (prolongations, repetitions, blocks, interjections, fluent(no disfluencies)) and non-disfluent labels including (natural pauesm unitelligible, unsure, no speech, poor audio quality and music)

4 AUTOMATIC STUTTERING IDENTIFICATION

4.1 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. Most of the existing work detect and identify stuttering either by language models [73, 74] or by ASR systems [75, 76], which first converts the audio signals into its corresponding textual form, and then by the application of language models, detects or identifies stuttering. This section provides in detail the comprehensive review of the various acoustic based 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 [77].

The various feature extraction methods that have been explored in the stuttering recognition systems are autocorrelation function and envelope parameters [78], duration, energy peaks, spectral of word based and part word based [79–81], age, sex, type of disfluency, frequency of disfluency, duration, physical concomitant, rate of speech, historical, attitudinal and behavioral scores, family history [38], duration and frequency of disfluent portions, speaking rate [26], frequency, 1^{st} to 3^{rd} formant's frequencies and its amplitudes [81, 82], spectral measure (fast Fourier transform (FFT) 512) [83, 84], mel frequency cepstral coefficients (MFCC) [81, 85–87], Linear Predictive Cepstral Coefficients (LPCCs) [81, 86], pitch, shimmer [88], zero crossing rate (ZCR) [81], short time average magnitude, spectral spread [81], linear predictive coefficients (LPC), weighted linear prediction cepstral coefficients (WLPCC) [86], maximum autocorrelation value (MACV) [81], linear prediction-Hilbert transform based MFCC (LH-MFCC) [89], noise to harmonic ratio, shimmer harmonic to noise ratio , harmonicity, amplitude perturbation quotient, formants and its variants (min, max, mean, median, mode, std), spectrum centroid [88], Kohonen's self-organizing Maps [84], i-vectors [90], perceptual linear predictive (PLP) [87], respiratory biosignals [39], and sample entropy feature [91]. With the recent developments in convolutional neural networks, the feature representation of

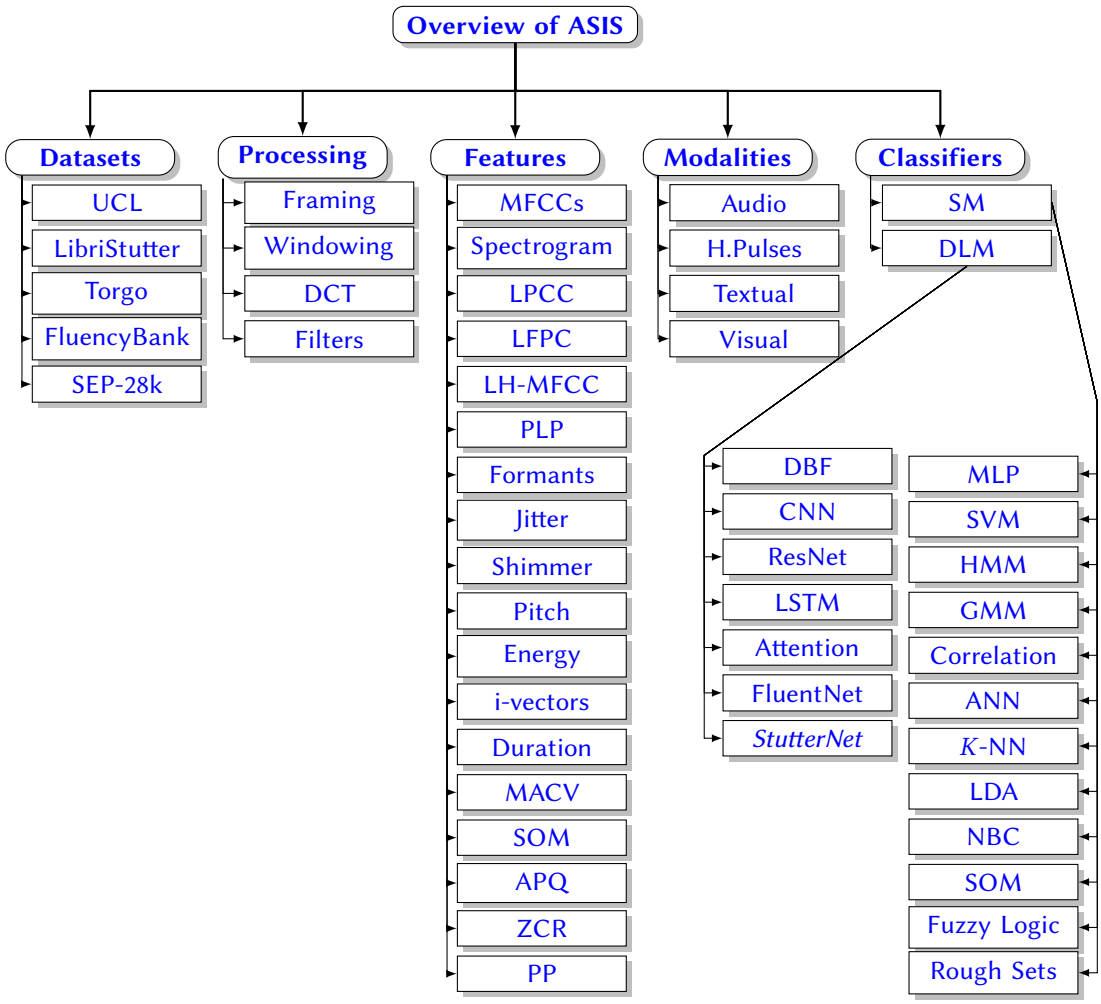

Fig. 2. Overview of automatic stuttering identification systems

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 blassifier

SOM: Self organizing maps

Features

MFCCs: Mel-frequency cepstral coefficients

LPCC: Linear prediction cepstral coefficients

LFPC: Log frequency power coefficients

DCT: Discrete cosine transforms

PP: Phoneme probabilities

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 1. Kourkounakis et al. [30] exploited the 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. Mahesha and Vinod [92] 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. Hariharan et al. [86] 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. Fook et al. [93] 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. Arjun et al. [94] used LPC and MFCCs as input features and concluded that MFCCs performs better than LPCs. Ai et al. [95] 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 [95]. This can be due to the possibility of LPCCs are potential in capturing the salient cues from stuttering [95]. The use of spectrograms showed state-of-the-art performance in recognising the stuttering events [30]. The work by Kourkounakis et al. [30] 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 2 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 explore stuttering identification include support vector machines (SVM), hidden Markov models (HMM), perceptron, multi layer perceptrons (MLP), Gaussian mixture models (GMM), k-nearest neighbor (k-NN), naive Bayes classifier (NBC), rough sets, Kohonen maps (self organizing maps (SOM)), linear discriminant analysis (LDA), artificial neural networks (ANN)

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 framework is provided by HMMs for modelling temporal sequences. Wiśniewski et al. [97] used euclidean distance as a codebook based on 20 MFCCs with HMMs. They reported an average recognition

[4]Modality Considered: Audio Only
[5]Modality Considered: Audio and Textual
[6]Modality Considered: Audio, Visual and Textual
[7]Modality Considered: Bio-Respiratory Signals

Author and Year	Datasets	Features	Stutter Type	Model
Howell and Sackin [78](95)[4]	6 Speakers	EP, ACF-SC	(P),(R)	ANN
Howell et al. [79, 80](97)[5]	12 Speakers	Energy peaks, Duration	NA	ANN
Nöth et al. [26](00)[4]	Northwind and Sun 37 Stutters, 16 Non-Stutters	Disfluent Frequency, Speaking rate, Duration	NA	HMMs
Geetha et al. [38](00)[5]	51 Stutters	Gender, Age Duration, Speech Rate	NA	ANNs
Czyzewski et al. [82](03)[4]	6-Normal, 6-SG Samples	Formants(1^{st} to 3^{rd}), Amplitude	(P),(R),(SG)	Rough Sets ANNs
Suszyński et al. [96] (03)[4]	NA	FFT	(P)	Fuzzy Logic
Szczurowska et al. [83](06)[4]	8 PWS	FFT 512 Spectral Measure	(B)	MLP and SOM
Wiśniewski et al. [97](07)[4]	30 samples	MFCCs	NA	HMMs
Tan et al. [98](07)[5]	UTM Skudai 10 Speakers (7M, 3F)	MFCCs	NA	HMMs
Ravikumar et al. [99](08)[4]	10 PWS	MFCCs, DTW for Score Matching	(SR)	Perceptron
Świetlicka et al. [84](09)[4]	8 PWS (Aged 10-23) 4 Fluent (2M, 2F)	FFT 512 Spectral Measure	NA	Kohonen based ML Kohonen based RBF
Chee et al. [100](09)[4]	UCLASS	MFCCs	(R), (P)	*k*-NN, LDA
Chee et al. [101](2009)[4]	UCLASS	LPCCs	(R), (P)	*k*-NN, LDA
Ravikumar et al. [102](09)[4]	15 PWS	MFCCs, DTW for score matching	(SR)	SVM
Yildirim and Narayanan [37](09)[6]	10 CWS(Aged 4-6)	Duration, Pitch, Energy, Gestural, Linguistic	(R),(FS), (FP),(RP)	NBC
Pálfy and Pospíchal [103](11)[4]	UCLASS	MFCCs	(R)	SVM(Linear Kernel) SVM(RBF Kernel)
Mahesha and Vinod [92](13)[4]	UCLASS	LPCC, MFCC	(P),(WR),(SR)	SVM
Świetlicka et al. [104](13)[4]	19 PWS	FFT(512)	(B),(P), (SR)	Hierarchical ANN
Oue et al. [105](15)[4]	TORGO	MFCCs, LPCCs	(R)	DBN
Mahesha and Vinod [89](17)[4]	UCLASS	LH-MFCC	(P),(R),(I)	GMMs
Esmaili et al. [87](17)[4]	UCLASS	PLP	(P)	Correlation
Esmaili et al. [87](17)[4]	UCLASS	WPT with entropy	(P),(R)	SVM
Esmaili et al. [87](17)[4]	Persian	WPT with entropy	(P),(R)	SVM
Ghonem et al. [90](17)[4]	UCLASS	I-Vectors	(R),(P), (RP)	*k*-NN, LDA
Santoso et al. [106](19)[4]	UUDB, PASD	Modulation Spectrum (Speech Rythm)	NA	BiLSTM
Santoso et al. [107](19)[4]	UUDB, PASD	Modulation Spectrum (Speech Rythm)	NA	BiLSTM + Attention
Villegas et al. [39](2019)[7]	69 Participants	Heart Rate Respiratory Air Volume Respiratory Air Flow	(B)	MLP
Kourkounakis et al. [30](20)[4]	UCLASS	Spectrograms	(WR),(I),(P), (SR),(RP), (FS)	ResNet + BiLSTM
Kourkounakis et al. [32](20)[4]	UCLASS, LibriStutter	Spectrograms	(WR),(I),(P), (SR),(R)	FluentNet
Sheikh et al. [31](21)[4]	UCLASS 128 PWS	MFCCs	(B),(P),(R),(F)	StutterNet

Table 2. Summary of several ASIS Systems in chronological order

rate of 70% for two stuttering classes including blocks and prolongation with deleted silence and 60 frames of window length. Tan et al. [98] used 12 MFCC features with HMMs. The average recognition rate is 93% [98]. This tool recognizes only normal and stutter utterances and is not classifying different types of disfluencies. In 2000, Nöth et al. [26] 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 [101]

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 and Vinod [92] 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% [92]. Ravikumar et al. [102] used SVM to classify one disfluency type which is syllable repetitions. The features used are MFCCs and DTW for score matching. An average accuracy of 94.35% is obtained on syllable repetitions. Pálfy and Pospíchal [103] used SVM with two different kernel functions including linear and radial basis function (RBF). In this case study, they used 16 audio samples from UCLASS [71] with eight males and eight females. 22 MFCC acoustic features with hamming window (25ms) with an overlap of 10 ms are used in this case study [103]. 96.4% is the best recognition rate that has been reported with SVM when RBF is used as a kernel function [103]. With linear kernel based SVM, recognition rate is 98% [103]. Esmaili et al. [87] used PLP features with a hamming window of 30 ms and an overlap of 20 ms 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 [87]. 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 [87].

Artificial Neural Networks (ANNs). They consist of several connected computing neurons that loosely model the biological neurons [108]. 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 [108]. 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 [108]. 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. Howell and Sackin [78] used two ANNs for repetition and prolongation recognition. The neural net is trained with 20 ACF, 19 vocoder coefficients of 10 ms frame length and also with 20 frames of envelope coefficients. The networks are trained for with just two 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 [78]. ACF-SC based ANNs gives the best accuracy of 79% and 71% for prolongations and repetitions respectively [78]. Howell et al. [79, 80] 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 [79, 80]. The average accuracy on prolongations and repetitions obtained in this case study is 78.01% [79, 80]. Geetha et al. [38] 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. They reported a classification accuracy of 92% [38]. Szczurowska et al. [83] 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 [83]. 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 eight PWS [83]. Ravikumar et al. [99] 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% [99]. In 2003, Czyzewski et al. [82] 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 six fluent and six stop-gap based speech samples [82]. 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 [82]. Suszyński et al. [96] 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 22 kHz. 21 1/3 octave frequency bands with *A* filter and FFT features are used with the hamming window of 20 ms. The features representing the prolongations are described by the use of fuzzy sets. Only the disfluent fricatives and nasals are considered in this study [96]. Świetlicka et al. [84] proposed an automatic recognition of prolongation type of stuttering by proposing Kohonen based MLP and RBF. From a dataset of eight PWS and fourfluent speakers, 118 (59 disfluent, 59 fluent), 118 total speech samples are recorded for the analysis. 21 1/3 octave filters with frequencies ranging from 100 Hz to 10000 Hz are used to parametrize the speech samples [84]. 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 [84]. The method yielded a classification accuracy of 92% for Kohonen based MLP and 91% for Kohonen based RBF [84].

Villegas et al. [39] 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% [39].

In 2013, Mahesha and Vinod [89] 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. 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 model (GMM) components and reports a gain of 1.79% in average accuracy [89] with LH-MFCCs. The proposed LH-MFCC improves discriminatory ability in all classification experiments [89].

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≥0)* neighbors. This method relies on the distance metric for classification [109]

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 [109].

Chee et al. [100] 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 [100] on UCLASS [71] dataset. In 2009, Chee et al. [101] 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 [71] dataset. In 2017, Ghonem et al. [90] 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[8] stuttering events [90]. This is the first technique so far that has taken two disfluencies (occurring at same time) into consideration.

In 2009, Yildirim and Narayanan [37] 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 [37]. Three different classifiers including *k*-NN, NBC and logistic model trees have been utilised to evaluate the effectiveness of multi modal features on the collected dataset [37]. The stuttering types included in this case study are repetition, repair, false start and filled pauses [37]. 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 [37]. The work achieved the best accuracy for NBC among the three classifiers [37] and reports an accuracy of 80.5% and 82.1% at feature level integration and decision level feature integration respectively [37].

In 2005, Oue et al. [105] introduced deep belief network for the automatic detection of repetitions, non-speech disfluencies. 45 MFCC and 14 LPCC features from TORGO dataset [72] has been used in this case study for the detection of disfluencies [72]. The experimental results obtained showed that MFCCs and LPCCs produce similar detection accuracies of approximately 86% for repetitions and 84% for non-speech disfluencies [105].

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.

4.2 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

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

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 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 and recurrent neural networks.

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 [108]. The other networks process inputs which are independent of each other, but in RNNs, inputs are related to each other.

Long short term memory networks (LSTMs) introduced by Hochreiter and Schmidhuber [110], is a special type of RNN, capable of capturing the long term dependencies in the temporal sequence.

In 2019, Santoso et al. [107] proposed modulation spectrum feature based BiLSTM (Bidirectional LSTM) 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 [107]. The experiment used 640-dimensional modulation spectrum feature vector with a block length of 320 ms [107]. The method achieved an F-score of 0.381 for successfully detecting the stuttering events in the speech [107]. 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, Santoso et al. [106] introduced attention based BiLSTM classifier for stuttering event detection. The best F-score of 0.691 is attained by taking the block length of 32 ms [106].

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 [108]. 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 [108]. This dot product is also known as convolution operation, defined by the following equation:

$$feature\ map = y[i, j] = input \circledast kernel = \sum \sum X[i - m, j - n].h[m, n] \tag{1}$$

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 [108]. 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 [108]. 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 [108]. The FC layer takes the non-linear combination of selected features, which is later used for downstream tasks like classification [108].

As discussed in Section 4.1 that most of the existing stuttering identification work either depend on language models or on ASR systems. This procedure of stuttering identification seems a subsidiary computational step and could also be a potential source of error. In order to address this, Kourkounakis et al. [30] proposed a CNN based model to learn stutter-related features. They formulated it a binary classification problem, where they used the same architecture for identifying different types of stuttering. They used residual based CNN and BiLSTM (for temporal aspect) to capture the disfluency-specific features from the spectrograms [30], that are the sole input features used in this study. The model is trained with batch norm and ReLU activation function [30]. Each BiLSTM layer is followed by a dropout rate of 0.2 and 0.4 respectively [30]. 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 six different types of stuttering: revision, prolongation, interjection, phrase repetition, word repetition, and sound repetition [30]. Kourkounakis et al. [32] proposed a FluentNet as shown in Figure 4. that combines squeeze-and-excitation residual network (SE-ResNet) with BiLSTM networks, where SE-ResNet (eight blocks) is used to learn the stutter-specific spectral frame-level representations. Each audio speech is first segmented into four 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 [32]. The 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 [32]. 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 [32].

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, we recently proposed a *StutterNet* [31], a time delay neural network based stuttering detection method shown in Figure 3. We 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 four second audio segments, then acoustic features (MFCCs) are extracted, which are then fed to the *StutterNet*. The features are generated after every 10 ms on a 20 ms window for each 4-sec audio sample. On this larger set of disfluent speakers, we compared this study with the ResNet+BiLSTN [30] based ASIS system and reported an overall average accuracy of 50.79% and Mathews correlation coefficient (MCC) of 0.23, in comparison to ResNer+BiLSTM based system comprising of 46.10% overall average accuracy and 0.21 MCC. The comparative results are shown in Table 3 and Table 4.

Among the DL based ASIS systems described above in detail, for a small set of disfluent speakers, the FluentNet classifier proposed by Kourkounakis et al. [32] 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.

	Precision				Recall				F1-Score			
Method	Rept	Pr	B	F	Rept	Pr	B	F	Rept	Pr	B	F
ResNet+BiLSTM [30]	0.33	0.42	0.43	**0.63**	0.20	**0.23**	**0.53**	0.55	0.22	**0.28**	0.44	0.52
StutterNet (Baseline)	**0.36**	**0.43**	0.42	0.59	**0.28**	0.17	0.42	0.67	**0.30**	0.23	0.42	0.62
StutterNet **(Optimized)**	0.35	0.31	**0.47**	0.59	0.24	0.13	0.47	**0.70**	0.27	0.16	**0.46**	**0.63**

Table 3. Results in precision, recall and F1-score (B: Block, F: Fluent, Rept: Repetition, Pr: Prolongation)

Method	Accuracy				Tot. Acc.	MCC.
	Rept	Pr	B	F		
Resnet+BiLSTM [30]	20.39	**23.17**	**53.33**	55.00	46.10	0.20
StutterNet *(Baseline)*	**27.88**	17.13	42.43	66.63	49.26	0.21
StutterNet **(Optimized)**	23.98	12.96	47.14	**69.69**	**50.79**	**0.23**

Table 4. Results in accuracies and MCC(B: Block, F: Fluent, Rept: Repetition, Pr: Prolongation)

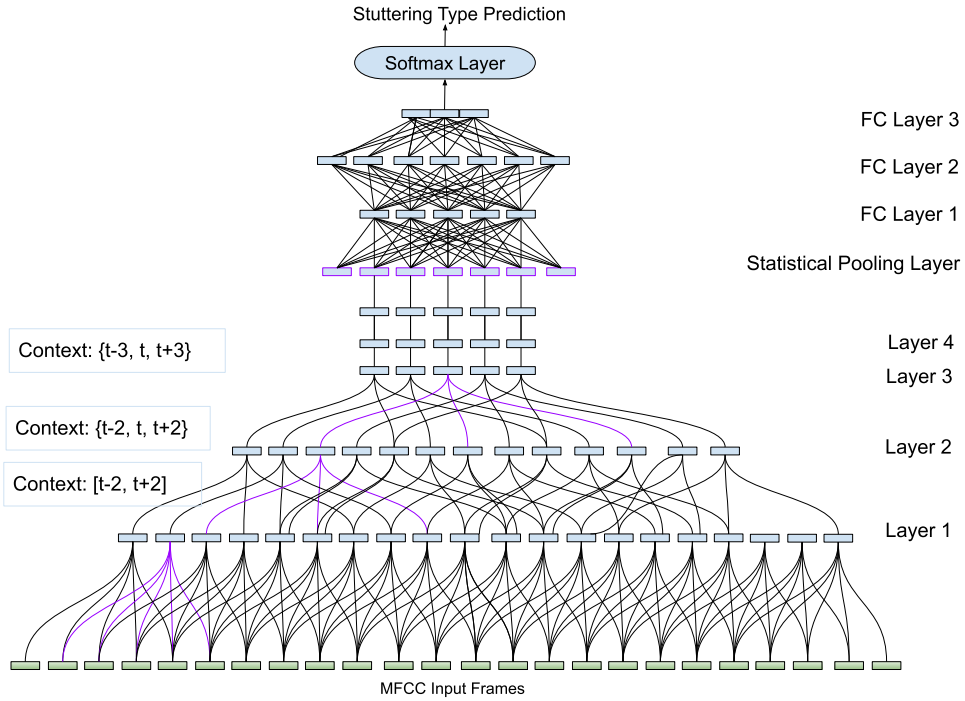

Fig. 3. StutterNet (reproduced with permission taken from authors) [31].

In a recent study by Lea et al. [33], they curated a large stuttering dataset named *SEP–28k* and employed ConvLSTM model to detect various types of stuttering. In addition to 40 MFCC input features, the model also takes pitch and articulatory features as an input, and reports a weighted accuracy of 83.6, F1 of 83.6 on the *FluencyBank*. On *SEP-28k*, they reported F1-scores of 55.9, 68.5, 63.2, 60.4, and 71.3 in block, prolongation, sound repetition, and word repetition respectively. They also evaluated their proposed model on 41 dim phoneme probabilities extracted from pre-trained

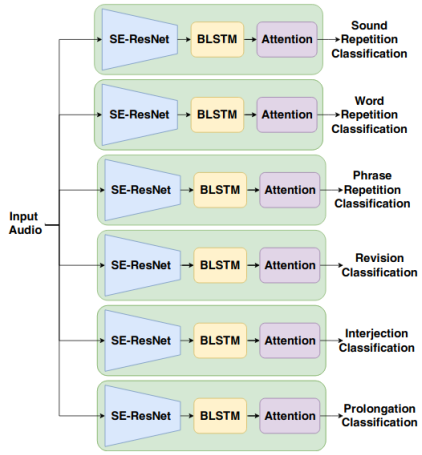

Fig. 4. FluentNet model for stuttering classification (reproduced with permission taken from authors) [32].

time-depth separable CNN on LibriSpeech, and reports F1 score of 74.8 on *FluenyBank*. The model was trained with a mini-batch size of 256 and cross entropy loss function.

5 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. Although there have been several developments in the automatic identification of stuttering, there are still several impediments that need to be addressed for robust and effective identification of stuttering.

5.1 Dataset Collection

One of the most common barriers that needs to be addressed is the issue of scarcity of data on stuttering identification research. There are only few natural stuttered datasets as discussed in Section 3. Usually, the medical data collection is expensive and very demanding, and stuttering is no exception. Thus, having ample speaker and sentence variation adds more complexity in stuttering domain. In order to make a fine analysis across several speakers, it is appropriate to have the same content (same list of sentences). 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 or alone, spontaneous or read speech, etc. This makes the collection of a corpus even more 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, language and dialect. It should be noted that the recording of spontaneous speech must be well controlled to comply with the legislation. Due to the sensitivity of medical data and privacy concerns, it can not applied at large scale. Currently, 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. Different background noises can corrupt the stuttered speech data. Likewise, the noise of recording equipments can also degrade the speech signal. Noise injection techniques [111] can be exploited to learn reliable stutter-specific features from the noisy corrupted data. DL models like denoising auto encoders, imputation auto encoders [112] can also be utilized to learn robust stutter-specific features from corrupted data. Training and testing data distribution mismatch is a significant challenge for ASIS systems to be robust to noise.

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 such as generative adversarial networks can be utilized for data augmentation [113] to generate more stuttered speech samples with the aim of improving the stuttering identification systems.

5.2 Data Annotation Issue

It is no doubt that DL has led to the enormous advancement in speech domain, 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 [108] can be used to exploit unlabelled data from different distributions, and these learned representation can later on be used in various (limited data, e.g., stuttering) downstream classification tasks. Unsupervised learning such as contrastive predictive coding [114] enables to capture the underlying innate structure/pattern(s) in the data distribution [112]. 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 [115] 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 [116], is a compelling approach to address this paucity of stuttered data by capturing the innate compositions of the disfluency data.

5.3 Data Imbalance

Stuttering datasets also suffer from the data imbalance problems, i.e., the number of samples available for different disfluenct categories is not uniform. It is mentioned that in stuttering, the repetitions are the most frequent ones followed by prolongations, and blocks [1]. However, in UCLASS datset, the block type is present in majority followed by repetitions and prolongations. The model trained on this type of imbalanced dataset is biased towards the majority class. In order to address this, several techniques can be exploited, including resampling [117], reweighting [118] and metric learning [119]. Self supervision as proposed recently by Yang et al. [120] can also be used to address the problem of labeling bias effect in learning on imbalanced disfluent data.

5.4 Lack of Appropriate Acoustic Representation

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 [112]. Liu et al. [121] recently proposed a learnable MFCCs for speaker verificaton. Sailor and Patil [122] showed that unsupervised deep auditory model can learn human auditory processing relevant features like filterbanks from raw speech. In addition, Millet and Zeghidour [123] used raw signal to learn the filterbanks for dysathria detection. This 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.

5.5 Domain Adaptation

Most of the existing ASIS techniques proposed so far are evaluated on dataset of specific language consisting of limited speakers. The existing ASIS techniques depend on a probabilistic model to capture domain specific factors, so that any alteration in the input speech domain 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 on cross-domain or cross-language environment. There could be two possible scenarios for 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. Domain adaptation techniques have been successfully applied in various speech tasks such as, emotion recognition [124], speaker verification [125], and ASR [126]. However in stuttering domain research, it has not been exploited yet. Several domain adaptation techniques such as [127, 128] culd be exploited to learn domain-invaraint stutter representations. These domain-invariant stutter specific representations can later on be used to improve the performance of various ASIS systems.

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 [129]. 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 strategy, (i.e., if the model is forced to learn some auxiliary tasks in parallel, in addition to it's main task). Language and gender classifications are two auxiliary tasks, that can be learned together with the stutter identification task on the same input feature space to improve generalization.

5.6 Multimodal 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 inputs including vision, hearing, smell etc., Ngiam et al. [130] proposed a multi-modal (audio visual) learning and showed how to train deep models that learn effective shared representations across the modalities. The stuttering itself exhibits as an audio visual 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 [19]. This multimodal learning paradigm from 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 by Shukla et al. [116], this framework could be helpful in learning stutter-specific features from audio signal guided by visual frames or vice-versa. Altinkaya and Smeulders [131] recently presented the first audio-visual stuttered dataset which consists of 25 speakers (14 male,

11 female). They trained ResNet based RNN (gated recurrent unit) on audio-visual modality for the detection of block stuttering type.

5.7 Attention Approach

The nature of stuttering is that it either usually happens on specific words or part of words, or part of sounds, syllables or phrases and thus, is contained only in certain frames. The attention networks have been successfully applied in speech emotion detection [132], speaker verification [133], speech recognition [134]. Attention networks [134, 135] which imitates the cognitive attention can be exploited to force the neural networks to focus on the particular stutter-embedded frames that may improve the detection performance of ASIS systems.

5.8 Multi-Stuttering Identification

Most of the ASIS studies focus on utterances, which consists of only one type of stuttering. However, the speech utterance can contain a mix of stuttering such as *d-d-d—–dog dog is big*, which consists of syllable repetition, prolongation and word repetition types of disfluencies [136]. There is a lack of studies in detecting multiple stuttering types if present in an utterance, and to the best of our knowledge, Ghonem et al. [90] is the only study, that has been carried out to detect multiple stuttering types (*repetition-prolongation*) in an utterance.

6 CONCLUSION

Stuttering is a speech disorder during which the flow of speech is interrupted by involuntary blocks, prolongations and repetitions. 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/STs. 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 identification is a complex interdisciplinary problem which involves speech processing, signal processing, neuroscience, psychology, pathology, and machine learning. The recent advancements in machine and DL has significantly transformed the speech domain. However in stuttering detection, it has not been explored eminently. This work tries to fill the gap by trying to bring researchers together from interdisciplinary fields. In the past two decades, a lot of research work has been performed in the automatic identification of stuttering. This paper gives an up-to-date comprehensive review of the various datasets, acoustic features and ASIS classification methods, that have been used by various researchers for the identification and recognition of stuttering disfluencies. In this paper, We 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 [71].

Due to the challenges discussed in the Section 5, ASIS systems are not yet available for real-time stutter identification, unlike ASR, 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.

ACKNOWLEDGMENTS

This work was made with the support of the French National Research Agency, in the framework of the project ANR BENEPHIDIRE (18-CE36-0008-03). Experiments presented in this paper were carried out using the Grid'5000 testbed, supported by a scientific interest group hosted by

Inria and including CNRS, RENATER and several universities as well as other organizations(see https://www.grid5000.fr) and using the EXPLOR centre, hosted by the University of Lorraine.

REFERENCES

[1] Barry Guitar. *Stuttering: An Integrated Approach to its Nature and Treatment.* Lippincott Williams & Wilkins, 2013.

[2] Joseph R Duffy. *Motor Speech Disorders-E-Book: Substrates, Differential Diagnosis, and Management.* Elsevier Health Sciences, 2013.

[3] Nan Bernstein Ratner and Brian MacWhinney. Fluency bank: A new resource for fluency research and practice. *Journal of Fluency Disorders*, 56:69, 2018.

[4] David Ward. *Stuttering and Cluttering: Frameworks for Understanding and Treatment.* Psychology Press, 2008.

[5] Thomas D Kehoe and Wikibooks Contributors. *Speech Language Pathology-Stuttering.* Kiambo Ridge, 2006.

[6] Joseph Kalinowski, Sandra Noble, Joy Armson, and Andrew Stuart. Pretreatment and posttreatment speech naturalness ratings of adults with mild and severe stuttering. *American Journal of Speech-Language Pathology*, 3(2):61–66, 1994.

[7] Anne Smith and Christine Weber. How stuttering develops: The multifactorial dynamic pathways theory. *Journal of Speech, Language, and Hearing Research*, 60(9):2483–2505, 2017.

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

[9] C Woodruff Starkweather. *Fluency and Stuttering.* Prentice-Hall, Inc, 1987.

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

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

[12] World Health Organization et al. *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, 1977.

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

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

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

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

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

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

[19] NIDCD. Stuttering. 2015. URL :https://www.nidcd.nih.gov/health/stuttering/.

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

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

[22] National Stuttering Association NSA. The experience of people who stutter: A survey by the national stuttering association. *New York, NY: Author*, 2009.

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

[24] Gordon W Blood and Ingrid M Blood. 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*, 50:72–84, 2016.

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

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

Spoken Language Processing, 2000.

[27] Tim Saltuklaroglu and Joseph Kalinowski. How effective is therapy for childhood stuttering? dissecting and reinterpreting the evidence in light of spontaneous recovery rates. *International Journal of Language & Communication Disorders*, 40(3):359–374, 2005.

[28] Megan York Roberts. *Using emprical benchmarks to assess the effects of a parentimplemented language intervention for children with language impairments*. Vanderbilt University, 2011.

[29] USA TODAY TECH. 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.

[30] Tedd Kourkounakis, Amirhossein Hajavi, and Ali Etemad. 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)*, pages 6089–6093. IEEE, 2020.

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

[32] Tedd Kourkounakis, Amirhossein Hajavi, and Ali Etemad. Fluentnet: End-to-end detection of speech disfluency with deep learning. *arXiv preprint arXiv:2009.11394*, 2020.

[33] Colin Lea, Vikramjit Mitra, Aparna Joshi, Sachin Kajarekar, and Jeffrey P. Bigham. Sep-28k: A dataset for stuttering event detection from podcasts with people who stutter. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6798–6802, 2021. doi: 10.1109/ICASSP39728.2021.9413520.

[34] Ali Bou Nassif, Ismail Shahin, Imtinan Attili, Mohammad Azzeh, and Khaled Shaalan. Speech recognition using deep neural networks: A systematic review. *IEEE access*, 7:19143–19165, 2019.

[35] Mehmet Berkehan Akçay and Kaya Oğuz. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Communication*, 116:56–76, 2020.

[36] Yishuang Ning, Sheng He, Zhiyong Wu, Chunxiao Xing, and Liang-Jie Zhang. A review of deep learning based speech synthesis. *Applied Sciences*, 9(19):4050, 2019.

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

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

[39] Bruno Villegas, Kevin M Flores, Kevin José Acuña, Kevin Pacheco-Barrios, and Dante Elias. 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)*, pages 4660–4663. IEEE, 2019.

[40] Rahilsadat Hosseini, Bridget Walsh, Fenghua Tian, and Shouyi Wang. An fnirs-based feature learning and classification framework to distinguish hemodynamic patterns in children who stutter. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 26(6):1254–1263, 2018. doi: 10.1109/TNSRE.2018.2829083.

[41] Soo-Eun Chang. Research updates in neuroimaging studies of children who stutter. In *Seminars in Speech and Language*, volume 35, page 67. NIH Public Access, 2014.

[42] Roger J Ingham, Peter T Fox, Janis C Ingham, Frank Zamarripa, Charles Martin, Paul Jerabek, and John Cotton. Functional-lesion investigation of developmental stuttering with positron emission tomography. *Journal of Speech, Language, and Hearing Research*, 39(6):1208–1227, 1996.

[43] AL Foundas, AB Lane, DM Corey, M Hurley, and KM Heilman. Anomalous anatomy in adults with persistent developmental stuttering: A volumetric mri study of cortical speech and language areas. In *Neurology*, volume 56, pages A157–A158. LIPPINCOTT WILLIAMS & WILKINS 530 WALNUT ST, PHILADELPHIA, PA 19106-3621 USA, 2001.

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

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

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

[47] Ivana Didirková and Fabrice Hirsch. A two-case study of coarticulation in stuttered speech. an articulatory approach. *Clinical Linguistics & Phonetics*, 34(6):517–535, 2020.

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

[49] M Jayaram. Phonetic influences on stuttering in monolingual and bilingual stutterers. *Journal of Communication Disorders*, 16(4):287–297, 1983.
[50] M Blomgren, M Alqhazo, and E Metzger. Do speech sound characteristics really influence stuttering frequency. In *Proceedings of the 7th World Congress of Fluency Disorders, CD-ROM*, 2012.
[51] Ivana Didirkova. *Parole, langues et disfluences: une étude linguistique et phonétique du bégaiement.* PhD thesis, Université Paul Valéry-Montpellier III, 2016.
[52] Patricia M Zebrowski, Edward G Conture, and Edward A Cudahy. Acoustic analysis of young stutterers' fluency: Preliminary observations. *Journal of Fluency Disorders*, 10(3):173–192, 1985.
[53] Ali Dehqan, Fariba Yadegari, Michael Blomgren, and Ronald C Scherer. Formant transitions in the fluent speech of farsi-speaking people who stutter. *Journal of Fluency Disorders*, 48:1–15, 2016.
[54] J Scott Yaruss and Edward G Conture. F2 transitions during sound/syllable repetitions of children who stutter and predictions of stuttering chronicity. *Journal of Speech, Language, and Hearing Research*, 36(5):883–896, 1993.
[55] Michael Robb, Michael Blomgren, and Yang Chen. Formant frequency fluctuation in stuttering and nonstuttering adults. *Journal of Fluency Disorders*, 23(1):73–84, 1998.
[56] Soo-Eun Chang, Ralph N Ohde, and Edward G Conture. Coarticulation and formant transition rate in young children who stutter. *Journal of Speech, Language, and Hearing Research*, 2002.
[57] Anu Subramanian, Ehud Yairi, and Ofer Amir. Second formant transitions in fluent speech of persistent and recovered preschool children who stutter. *Journal of Communication Disorders*, 36(1):59–75, 2003.
[58] Michael Blomgren, Michael Robb, and Yang Chen. A note on vowel centralization in stuttering and nonstuttering individuals. *Journal of Speech, Language, and Hearing Research*, 41(5):1042–1051, 1998.
[59] Fabrice Hirsch, Fayssal Bouarourou, Béatrice Vaxelaire, Marie-Claude Monfrais-Pfauwadel, Marion Bechet, Jean Sturm, and Rudolph Sock. Formant Structures of Vowels Produced by Stutterers in Normal and Fast Speech Rates. In *8th International Seminar On Speech Production*, page NC, France, December 2008. URL https://halshs.archives-ouvertes.fr/halshs-00716583.
[60] E Charles Healey and Peter R Ramig. Acoustic measures of stutterers' and nonstutterers' fluency in two speech contexts. *Journal of Speech, Language, and Hearing Research*, 29(3):325–331, 1986.
[61] Robert E Hillman and Harvey R Gilbert. 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):610–611, 1977.
[62] Martin R Adams. Voice onsets and segment durations of normal speakers and beginning stutterers. *Journal of Fluency Disorders*, 12(2):133–139, 1987.
[63] Ben C Watson and Peter J Alfonso. A comparison of lrt and vot values between stutterers and nonstutterers. *Journal of Fluency Disorders*, 7(2):219–241, 1982.
[64] Lutz Jäncke. Variability and duration of voice onset time and phonation in stuttering and nonstuttering adults. *Journal of Fluency Disorders*, 19(1):21–37, 1994.
[65] Luc F De Nil and GJ Brutten. Voice onset times of stuttering and nonstuttering children: The influence of externally and linguistically imposed time pressure. *Journal of Fluency Disorders*, 16(2-3):143–158, 1991.
[66] Leticia Correa Celeste and Vanessa de Oliveira Martins-Reis. The impact of a dysfluency environment on the temporal organization of consonants in stuttering. *Audiology-Communication Research*, 20(1):10–17, 2015.
[67] Sibylle Brosch, Andrea Häge, and Helge S Johannsen. Prognostic indicators for stuttering: The value of computer-based speech analysis. *Brain and Language*, 82(1):75–86, 2002.
[68] Gloria J Borden, Thomas Baer, and Mary Kay Kenney. Onset of voicing in stuttered and fluent utterances. *Journal of Speech, Language, and Hearing Research*, 28(3):363–372, 1985.
[69] Susan Meyers Fosnot and S Jun. Prosodic characteristics in children with stuttering or autism during reading and imitation. In *Proceedings of the 14th International Congress of Phonetic Sciences*, pages 1925–1928, 1999.
[70] Peter R Ramig and Martin R Adams. Vocal changes in stutterers and nonstutterers during high-and low-pitched speech. *Journal of Fluency Disorders*, 6(1):15–33, 1981.
[71] Peter Howell, Stephen Davis, and Jon Bartrip. The university college london archive of stuttered speech (uclass). *Journal of Speech, Language, and Hearing Research*, 2009.
[72] Frank Rudzicz, Aravind Kumar Namasivayam, and Talya Wolff. The torgo database of acoustic and articulatory speech from speakers with dysarthria. *Journal of Language Resources and Evaluation*, 46(4):523–541, 2012.
[73] Vicky Zayats, Mari Ostendorf, and Hannaneh Hajishirzi. Disfluency detection using a bidirectional lstm. In *Interspeech 2016*, pages 2523–2527, 2016. doi: 10.21437/Interspeech.2016-1247. URL http://dx.doi.org/10.21437/Interspeech.2016-1247.
[74] Qian Chen, Mengzhe Chen, Bo Li, and Wen Wang. Controllable time-delay transformer for real-time punctuation prediction and disfluency detection. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8069–8073, 2020. doi: 10.1109/ICASSP40776.2020.9053159.

[75] Sadeen Alharbi, Madina Hasan, Anthony J H Simons, Shelagh Brumfitt, and Phil Green. A lightly supervised approach to detect stuttering in children's speech. In *Proc. Interspeech 2018*, pages 3433–3437, 2018. doi: 10.21437/Interspeech.2018-2155. URL http://dx.doi.org/10.21437/Interspeech.2018-2155.

[76] Sadeen Alharbi, Madina Hasan, Anthony J H Simons, Shelagh Brumfitt, and Phil Green. Sequence labeling to detect stuttering events in read speech. *Computer Speech Language*, 62:101052, 2020. ISSN 0885-2308. doi: https://doi.org/10.1016/j.csl.2019.101052. URL https://www.sciencedirect.com/science/article/pii/S0885230819302967.

[77] Xuedong Huang, Alex Acero, Hsiao-Wuen Hon, and Raj Reddy. *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. Prentice hall PTR, 2001.

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

[79] Peter Howell, Stevie Sackin, and Kazan Glenn. 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):1073–1084, 1997.

[80] Peter Howell, Stevie Sackin, and Kazan Glenn. 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):1085–1096, 1997.

[81] Shweta Khara, Shailendra Singh, and Dharam Vir. 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)*, pages 887–893. IEEE, 2018.

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

[83] Izabela Szczurowska, Wiesława Kuniszyk-Jóźkowiak, and Elżbieta Smołka. The application of kohonen and multilayer perceptron networks in the speech nonfluency analysis. *Archives of Acoustics*, 31(4 (S)):205–210, 2014.

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

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

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

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

[88] 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. On the analysis of speech and disfluencies for automatic detection of mild cognitive impairment. *Neural Computing and Applications*, pages 1–9, 2018.

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

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

[91] M Hariharan, Vikneswaran Vijean, CY Fook, and Sazali Yaacob. 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*, pages 240–245. IEEE, 2012.

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

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

[94] KN Arjun, S Karthik, D Kamalnath, Pranavi Chanda, and Shikha Tripathi. Automatic correction of stutter in disfluent speech. *Procedia Computer Science*, 171:1363–1370, 2020.

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

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

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

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

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

[100] Lim Sin Chee, Ooi Chia Ai, M Hariharan, and Sazali Yaacob. 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)*, pages 146–149. IEEE, 2009.

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

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

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

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

[105] Stacey Oue, Ricard Marxer, and Frank Rudzicz. 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*, pages 60–64, 2015.

[106] Jennifer Santoso, Takeshi Yamada, and Shoji Makino. 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)*, pages 302–306. IEEE, 2019.

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

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

[109] Kevin P Murphy. *Machine Learning: a Probabilistic Perspective*. MIT press, 2012.

[110] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, November 1997. ISSN 0899-7667. doi: 10.1162/neco.1997.9.8.1735.

[111] Shi Yin, Chao Liu, Zhiyong Zhang, Yiye Lin, Dong Wang, Javier Tejedor, Thomas Fang Zheng, and Yinguo Li. Noisy training for deep neural networks in speech recognition. *EURASIP Journal on Audio, Speech, and Music Processing*, 2015(1):1–14, 2015.

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

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

[114] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[115] Xiaojin Zhu and Andrew B Goldberg. Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning*, 3(1):1–130, 2009.

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

[117] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

[118] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9260–9269, 2019. doi: 10.1109/CVPR.2019.00949.

[119] Nan Wang, Xibin Zhao, Yu Jiang, Yue Gao, and KLISS BNRist. Iterative metric learning for imbalance data classification. In *IJCAI*, pages 2805–2811, 2018.

[120] Yuzhe Yang and Zhi Xu. Rethinking the value of labels for improving class-imbalanced learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 19290–19301. Curran Associates, Inc., 2020.

[121] Xuechen Liu, Md Sahidullah, and Tomi Kinnunen. Learnable mfccs for speaker verification. In *2021 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–5, 2021. doi: 10.1109/ISCAS51556.2021.9401593.

[122] Hardik B Sailor and Hemant A Patil. Unsupervised deep auditory model using stack of convolutional rbms for speech recognition. In *INTERSPEECH*, pages 3379–3383, 2016.

[123] Juliette Millet and Neil Zeghidour. Learning to detect dysarthria from raw speech. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5831–5835. IEEE, 2019.

[124] Mohammed Abdelwahab and Carlos Busso. Supervised domain adaptation for emotion recognition from speech. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5058–5062, 2015. doi: 10.1109/ICASSP.2015.7178934.

[125] Md Jahangir Alam, Gautam Bhattacharya, and Patrick Kenny. Speaker verification in mismatched conditions with frustratingly easy domain adaptation. In *Proc. Odyssey 2018 The Speaker and Language Recognition Workshop*, pages 176–180, 2018. doi: 10.21437/Odyssey.2018-25. URL http://dx.doi.org/10.21437/Odyssey.2018-25.

[126] Lahiru Samarakoon, Brian Mak, and Albert Y.S. Lam. Domain adaptation of end-to-end speech recognition in low-resource settings. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 382–388, 2018. doi: 10.1109/SLT.2018.8639506.

[127] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning*, pages 1180–1189. PMLR, 2015.

[128] Qing Wang, Wei Rao, Sining Sun, Leib Xie, Eng Siong Chng, and Haizhou Li. Unsupervised domain adaptation via domain adversarial training for speaker recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4889–4893, 2018. doi: 10.1109/ICASSP.2018.8461423.

[129] Gueorgui Pironkov, Stephane Dupont, and Thierry Dutoit. Multi-task learning for speech recognition: an overview. In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, 2016.

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

[131] Mehmet Altinkaya and Arnold W.M. Smeulders. A dynamic, self supervised, large scale audiovisual dataset for stuttered speech. In *Proceedings of the 1st International Workshop on Multimodal Conversational AI*, MuCAI?20, page 9–13, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450381567. URL https://doi.org/10.1145/3423325.3423733.

[132] Yuanyuan Zhang, Jun Du, Zirui Wang, Jianshu Zhang, and Yanhui Tu. Attention based fully convolutional network for speech emotion recognition. In *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1771–1775. IEEE, 2018.

[133] Shi-Xiong Zhang, Zhuo Chen, Yong Zhao, Jinyu Li, and Yifan Gong. End-to-end attention based text-dependent speaker verification. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, pages 171–178. IEEE, 2016.

[134] Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. Attention-based models for speech recognition. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper/2015/file/1068c6e4c8051cfd4e9ea8072e3189e2-Paper.pdf.

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

[136] Jean Sawyer. By the numbers: Disfluency analysis for preschool children who stutter. In *Proc. International Stuttering Awareness Day Online Conference*, 2010.