

FluentNet: End-to-End Detection of Speech Disfluency with Deep Learning

Tedd Kourkounakis, Amirhossein Hajavi, and Ali Etemad

Index Terms-Speech, stutter, disfluency, deep learning, squeeze-and-excitation, BLSTM, attention.

I. INTRODUCTION

CLEAR and comprehensive speech is the vital backbone to strong communication and presentation skills [1]. Where some occupations consist mainly of presenting, most careers require and thrive from the ability to communicate effectively. Research has shown that oral communication remains one of the more employable skills in both the perception of employers and new graduates [2]. Simple changes to ones speaking patterns such as volume or appearance of disfluencies can have a huge impact on the ability to convey information effectively. By providing simplified, quantifiable data concerning ones speech patterns, as well as feedback on how to change ones speaking habits, drastic improvements could be made to anyone's communication skills [3].

In regard to presentation skills, disfluent speech remains one of the more common factors [4]. Any abnormality or generally uncommon component of one's speech patterns is

referred to as a speech disfluency [5]. There are hundreds of different speech disfluencies often grouped together alongside language and swallowing disorders. Of these afflictions, stuttering proves to be one of the most common and most recognized of the lot [5].

Stuttering, also known as stammering, as a disorder can be generally defined as issues pertaining to the consistency of the flow and fluency of speech. This often involves involuntary additions of sounds and words, and the delay or inability to consistently progress through a phrase. Although labelled as a disorder, stuttering can occur in any persons speech, often induced by stress or nervousness [6]. These cases however do not correlate with stammering as a disorder, but are caused by performance anxiety [7]. The use of stutter detection does not only apply to those with long term stutter impairments, but can appeal to the majority of the world as it can help with the improvement of communication skills.

As the breadth of applications using machine learning techniques have flourished in recent decades, they have only recently began to be utilized in the field of speech disfluency and disorder detection. While deep learning has dominated many areas of speech processing, for instance speech recognition [8] [9], speaker recognition [10] [11], and speech synthesis [12] [13], very little work has been

done toward the problem of speech disfluency detection. Disfluencies, including stutters, are not easily definable; they come in many shapes and variations. This means that factors such as gender, age, accent, and the language themselves will affect the contents of each stutter, greatly complicating the problem space. As well, there are many classes of stutter, each with their own sub-classes and with wildly different structures, making the identification of all stutter types with a single model a difficult task. Even a specific type of stutter applied to a single word can be conducted in a wide variety of ways. Where people are great at identifying stutters through their experience with them, machine learning models have historically struggled with this (as we show in Section III).

Another common issue is the sheer lack of sufficient training data available. Many previous works often rely on their own manually recorded, transcribed, and labelled datasets, which are often quite small due to the work involved in their creation [14] [15] [16] [17]. There is only one commonly used public dataset, UCLASS [18], that is widely used amongst works in this area, though it still is also quite small.

Many disfluency detection solutions provide some form of filler word identification, flagging and counting any spoken interjections (e.g. 'okay', 'right', etc.). However, upon further investigation, these applications simply request a list of interjections from the user and use Speech-to-Text (STT) tools in order to match the spoken word with any interjections in the list. Though this may work fine for interjections such as 'um' and 'uh' (assuming the used STT tool has the necessary embeddings), this can lead to serious overall errors in classification for most other utterances that are actual words, such as 'like', which is commonly used as a filler word in the English language.

Early works in stutter detection, realizing the challenges mentioned above, first sought out to test the viability of identifying stutters from clean speech. These models primarily focused on machine learning models with very small datasets, consisting of a single stutter type, or even a single word [14], [19]. In more recent years, and due to the rise of automatic speech recognition (ASR), language models have been used to tackle stutter recognition. These works have proven to be strong at identifying certain stutter types, and have been showing ever improving results [17], [16]. However, due to the uncertainty surrounding relations between cleanly spoken and stuttered word embeddings, it remains difficult for these models to generalize across multiple stutter types. It is hypothesized that by bypassing the use of language

All the authors are with the Department of Electrical and Computer Engineering, Queen's University, Kingston, ON, K7L3N6 Canada, e-mail: tedd.kourkounakis@queensu.ca.

models, and by focusing solely on phonetics through the use of convolution networks, a model can be created that both maintains a strong average accuracy while also being effective across all stutter types.

In this paper, we propose a model capable of detecting speech disfluencies. To this end, we design **FluentNet**, a deep neural network (DNN) for automated speech disfluency detection. The proposed network does not apply any language model aspects, but instead focuses on the direct classification of speech signals. This allows for the avoidance of complex and time consuming ASR as a pre-processing steps in our model, and would provide the ability to view the scenario as an end-to-end solution using a single deep neural network. We validate our model on a commonly used benchmark dataset UCLASS [18]. To tackle the issue of scarce stutter-related speech datasets, we also develop a synthetic dataset based on a non-stuttered speech dataset (LibriSpeech [20]), which we entitle LibriStutter. This dataset is created to mimic stuttered speech and vastly expand the amount of data available for use. Our end-to-end neural network takes spectrogram feature images as inputs, and uses Squeeze-and-Excitation residual (SE-ResNet) blocks for learning the speech embedding. Next, a bidirectional long short-term memory (BLSTM) network is used to learn the temporal relationships, followed by an attention mechanism to focus on the more salient parts of the speech. Experiments show the effectiveness of our approach in generalizing across multiple classes of stutters while maintaining a high accuracy and strong consistency between classes on both datasets.

The key contributions of our work can be summarized as follows: (1) We propose **FluentNet**, an end-to-end deep neural network capable of detection of several types of speech disfluencies; (2) We develop a synthesized disfluency dataset called LibriStutter based on the publicly available LibriSpeech dataset by artificially generating several types of disfluencies, namely sound, word, and phrase repetitions, as well as prolongations and interjections. The dataset contains the output labels that can be used in training deep learning methods; (3) We evaluate our model (**FluentNet**) on two datasets, UCLASS and LibriStutter. The experiments show that our model achieves state-of-the-art results on both datasets outperforming a number of other baselines as well as previously published work; (4) We make our annotations on the existing UCLASS dataset, along with the entire LibriStutter dataset and its labels, publicly available ¹ to contribute to the field and facilitate further research.

This is an extension of our earlier work titled "Detecting Multiple Speech Disfluencies using a Deep Residual Network with Bidirectional Long Short-Term Memory", published in the 2020 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). This paper focused on tackling the problem of detection and classification of different forms of stutters. The model used a deep residual network and bidirectional long short-term memory layers to classify different types of stutters. In this extended work, we replace the previously used residual blocks of the spectral

encoder with residual squeeze-and-excitation blocks. Additionally, we add an attention mechanism after the recurrent network to better focus the network on salient parts of input speech. Furthermore, we develop a new dataset, which we present in this paper and make publicly available. Lastly, we perform thorough experiments, for instance through additional benchmark comparisons and ablation studies. Our experiments show the improvements made by **FluentNet** over our preliminary work, as validated on both the UCLASS dataset (previously used) as well as the newly developed dataset. This new model provides greater advancement towards end-to-end disfluency detection and classification.

The rest of this paper is organized as follows; a discussion of previous contributions towards stutter recognition in Section III followed by our methodology including a breakdown of the model in Section III, the datasets and benchmark models applied in Section IV, a discussion of our results in Section V, and our conclusion in the final section.

II. RELATED WORK

There has recently been increasing interest in the fields of deep learning, speech, and audio processing. However, as discussed earlier in section 1, there has been minimal research targeting automated detection of speech disfluencies including stuttering, most likely as a result of insufficient data and smaller number of potential applications in comparison to other speech-related problems such as speech recognition [21] [9] and speaker recognition [10] [11]. In the following sections we first provide a summary of the type of disfluencies commonly targeted in the area, followed by a review of the existing work that fall under the umbrella of speech disfluency detection and classification.

A. BACKGROUND: TYPES OF SPEECH DISFLUENCY

There are a number of different stuttering types, often categorized into four main groups: repetitions, prolongations,

interjections, and blocks. A summary of all these disfluency types and examples of each have been presented in Table I. The descriptions for each of these categories is as follows.

Repetitions are classified as any part of an utterance repeated at quick pace. As this definition still remains general, repetitions are often further sub-categorized [5]. These subcategories have been used in previous works classifying stutter disfluencies [22] [23] [17], which include sound, word, and phrase repetitions, as well as revisions. Sound repetitions (S) are repetitions of a single phoneme, or short sound, often consisting of a single letter. Part-word, or syllable repetitions (PW), as its name suggests, are classified as the repetition of syllables, which can consist of multiple phonemes. Similarly, word repetitions (W) are defined as any repetition of a single word, and phrase repetitions (PH) are the repetition of phrases, consisting of multiple consecutive words. The final repetition type disfluency is revision (R). Similar to phrase repetitions, they consist of repeated phrases, where the

¹1 <http://aiimlab.com/resources.html>

repeated segment is rephrased, containing new or different information from the first iteration. A rise in pitch may accompany this disfluency type [24].

Interjections (I), often referred to as filler words, consist of the addition of any utterance that does not logically belong in the spoken phrase. Common interjections in the English language include exclamations, such as 'um' and 'uh', as well as discourse markers such as 'like', 'okay', and 'right'.

Prolongation (PR) stutters are presented as a lengthened or sustained phoneme. The duration of these prolonged utterances vary alongside the severity of the stutter. Similar to repetitions, this disfluency is often accompanied by a rise in pitch.

The final category of stuttering are silent blocks (B), which are sudden cutoffs of vocal utterances. These are often involuntary and are presented as pauses within a given phrase.

B. STUTTER RECOGNITION WITH CLASSICAL MACHINE LEARNING

Before the focus of stutter recognition targeted maximizing accuracy in classification of stammers, a number of works were performed toward testing the viability of stutter detection. In 1995, Howell et al. [14], who later helped to create the UCLASS dataset [18] used in this paper, employed a set of pre-defined words to identify repetition and prolongation stutters. From these, they extracted the autocorrelation features, spectral information, and envelope parameters from the audio. Each was used as an input to a fully connected artificial neural network (ANN). Findings showed that the model achieved its strongest classification results against severe disfluencies,

and was weakest for mild ones. These models were able to achieve a maximum detection rate of 0.82 on severe prolongation stutters. Howell et al. [15] later furthered their work using a larger set of data, as well as a wider variety of audio parameters. This work also introduced an ANN model for both repetition and prolongation types, and more judges were used to identify stutters with strict restrictions towards agreement of disfluency labeling. Results showed that the best parameters for disfluency classification were fragmentation spectral measures for whole words, as well as duration and supralexical disfluencies of energy in part-words.

Tan et al. [19] worked on testing the viability of stutter detection through a simplified approach in order to maximize the possible results. By collecting audio samples of clean, stuttered, and artificially generated copies of single pre-chosen words, they were able to reach an average accuracy of 96% on the human samples using a hidden Markov model (HMM). This served as a temporary benchmark towards the possible best average results for stutter detection.

Ravikumar et al. have attempted a variety of classifiers on syllable repetitions, including an HMM [25] and support vector machine (SVM) [26] using Mel-frequency cepstral coefficients (MFCCs) features. Their best results were obtained when classifying this stutter type using the SVM on

15 participants, achieved an accuracy of 94.35%. No other disfluency types were considered.

A detailed summary of previously attempted stutter classification methods, including some of the aforementioned classical models, is available in the form of a review paper in [27]. This paper provides background on the use of three different models (ANNs, HMMs and SVM) towards the application of stutter recognition. Of the works considered in that review paper in 2009, it was concluded that HMMs achieve the best results in stutter recognition.

C. STUTTER RECOGNITION WITH DEEP LEARNING

With the recent advancements in deep learning, disfluency detection and classification has seen an increase in popularity within the field with a higher tendency towards end-to-end approaches. ASR has become an increasingly popular method of tackling the problem of disfluency classification. As some stuttered speech results in repeated words, as well as prolonged utterances, these can be represented by word embeddings and sound amplitude features, respectively. To exploit this concept, Alharbi et al. [17] detected sound and word repetitions, as well

as revision disfluencies using task-oriented finite state transducer (FST) lattices. They also utilized amplitude thresholding techniques to detect prolongations in speech. These methods resulted in an average 37% miss rate across the 4 different types of disfluencies.

Dash et al. [16] have used an STT model in order to identify word and phrase repetitions within stuttered speech. To detect prolongation stutters, they integrated a neural network capable of finding optimal cutoff amplitudes for a given speaker to expand upon simple thresholding methods. As these ASR works required full word embeddings to classify repetitions, they either fared poorly against, or did not attempt sound or part word repetitions.

Deep recurrent neural networks (RNN), namely BLSTM, have been used to tackle stutter classification. Zayats et al. [28] trained a BLSTM with Integer Linear Programming (ILP) [32] on a set of MFCC features to detect repetitions with an F-score of 85.9. Similarly, a work done by Santoso et al. applied a BLSTM followed by an attention mechanism to perform stutter detection based on input MFCC features, obtaining an maximum F-score of 69.1 [30]. More recently in a study by Chen et al., a Controllable Time-delay Transformer (CT-Transformer) has been used to detect speech disfluencies and correct punctuation in real time [31]. In our initial work on stutter classification, we utilized spectrogram features of stuttered audio and used a BLSTM [33] to learn temporal relationships following spectral frame-level representation learning by a ResNet. This model achieved a 91.15% average accuracy across six different stutter categories.

In an interesting recent work, Villegas et al. utilized respiratory biosignals in order to better detect stutters [29]. By correlating respiratory volume and flow, as well as heart rate measurements correlating to the time when a stutter occurs, they were able to classify block stutters with an accuracy of 95.4% using an MLP.

A 2018 summary and comparison of different features and classification methods for stuttering has been conducted by Khara et al. [34]. This work discusses and compares different popular feature extraction methods, classifiers and their uses, as well as their advantages and shortcomings. The paper discusses that MFCC feature extraction has historically provided the strongest results. Similarly, ANNs provide the most flexibility and adaptability compared to other models, especially linear ones.

Table II provides a summary of the related works on disfluency detection and classification. It can be observed and concluded that disfluency classification has been progressing in one of two fronts i) end-to-end speech-based methods, or ii) language-based models relying on an ASR pre-processing step. Our work in this paper is positioned in the first category in order to avoid the reliance on an ASR step. Moreover, from Table II. we observe that although progresses is being made in the area of speech disfluency recognition, the lack of available data remains a hindrance to potential further achievements in the field.

III. PROPOSED METHOD

A. PROBLEM AND SOLUTION OVERVIEW

Our goal in this section is to design and develop a system that can be used for detecting various types of disfluencies. While one approach to tackle this concept is to design a multiclass problem, another approach is to design an ensemble of single-disfluency detectors. In this paper, given the relatively small size of available stutter datasets, we use the latter approach which can help reduce the complexity of each binary task. Accordingly, the goal is to design a single network architecture that can be trained separately to detect different disfluency types with each trained instance, where together they could detect a number of different disfluencies. Figure 1 shows the overview of our system. The designed network should possess the capability of learning spectral frame-level representations as well as temporal relationships. Moreover, the model should be able to focus on salient parts of the inputs in order to effectively learn the disfluencies and perform accurately.

B. PROPOSED NETWORK: FLUENTNET

We propose an end-to-end network, **FluentNet**, which uses the short-time Fourier transform (STFT) spectrograms of audio clips as inputs. These inputs are passed through a Squeeze-and-Excitation Residual Network (SE-ResNet) to learn framelevel spectral representations. As most stutter types have distinct spectral and temporal properties, a bidirectional LSTM network is introduced to learn the temporal relationships

present among different spectrograms. An attention mechanism is then added to the final recurrent layer to better focus on the necessary features needed for stutter classification. **FluentNet**'s final output reveals a binary classification to detect a specific disfluency type that it has been trained for. The architecture of the network is presented in Figure 2(a).

In the following, we describe each of the components of our model in detail.

- 1) Data Representation: Input audio clips recorded with a sampling rate of 16 khz are converted to spectrograms using STFT with 256 filters (frequency bins) to be fed to our end-to-end model. A sample spectrogram can be seen in Figure 2 where the colours represent the amplitude of each frequency bin at a given frame, with blue representing lower amplitudes, and green and yellow representing higher amplitudes. Following the common practice in audio-signal processing, a 25 ms frame has been used with an overlap of 10 ms .
- 2) Learning Frame-level Spectral Representations: **FluentNet** first focuses on learning effective representations from each input spectrogram. To do so, CNN architectures are often used. Though both residual networks [35] and squeeze-andexcitation (SE) networks [36] are relatively new in the field of deep learning, both have proven to improve on previous state-of-the-art models in a variety of different application areas [37], [38]. The ResNet architecture has proven a reliable solution to the vanishing or exploding gradient problems, both common issues when back-propagating through a deep neural network. In many cases, as the model depth increases, the gradients of weights in the model become increasing smaller, or inversely, explosively larger with each layer. This may eventually prevent the gradients from actually changing the weights, or from the weights becoming too large, thus preventing improvements in the model. A ResNet, overcomes this by utilizing shortcuts all through its CNN blocks resulting in norm-preserving blocks capable of carrying gradients through very deep models.

Squeeze-and-excitation modules have been recently proposed and have shown to outperform various DNN models using previous CNN architectures, namely VGG and ResNet, as their backbone architectures [36]. SE networks were first proposed for image classification, reducing the relative error compared to previous works on the ImageNet dataset by approximately 25% [36].

Every kernel within a convolution layer of a CNN results in an added channel (depth) for the output feature map. Whereas recent works have focused on expanding on the spectral relationships within these models [39] [40], SE-blocks build stronger focus on channel-wise relationships within a CNN. These blocks consist of two major operations. The squeeze operation aggregates a feature map across both its height and width resulting in a one-dimensional channel descriptor. The excitation operation consists of fully connected layers providing channel-wise weights, which are then applied back to the original feature map.

To exploit the capabilities of both ResNet and SE architectures and learn effective spectral frame-level representations from the input, we use an SE-ResNet model in our end-toend network. The network consists of 8 SE-ResNet blocks, as shown in Figure 2(a). Each SE-ResNet block

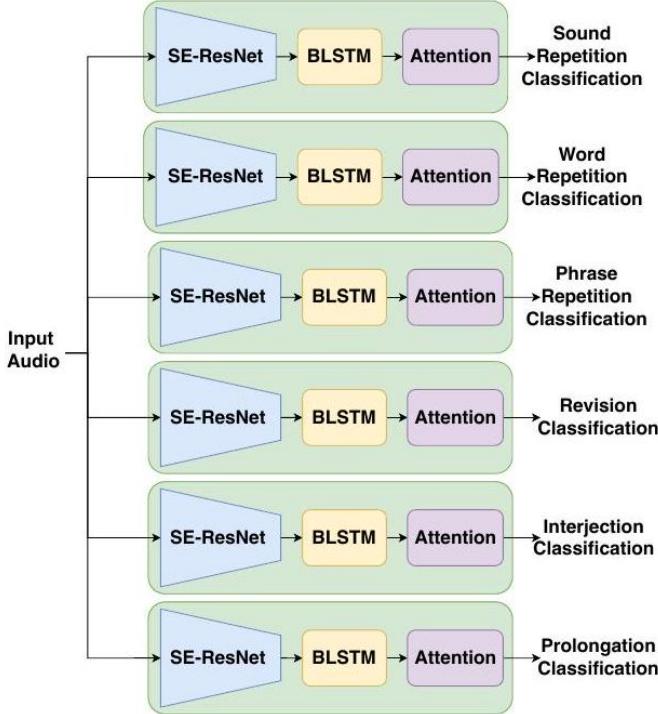


Fig. 1: Full model overview using **FluentNet** for disfluency classification.

in **FluentNet**, illustrated in Figure 2(b), consists of three sets of twodimensional convolution layers, each succeeded by a batch normalization and Rectified Linear Unit (ReLU) activation function. A separate residual connection shares the same input as the block’s non-identity branch, and is added back to the non-identity branch before the final ReLU function, but after the SE unit (described below). Each residual connection contains a convolution layer followed by batch normalization. The Squeeze-and-Excitation unit within each SE-ResNet block begins with a global pooling layer. The output is then passed through two fully connected layers: the first followed by a ReLU activation function, and the second succeeded with a sigmoid gating function. The main convolution branch is scaled with the output of the SE unit using channel-wise multiplication.

3) Learning Temporal Relationships: In order to learn the temporal relationships between the representations learned from the input spectrogram, we use an RNN. In particular, LSTM [41] networks have shown to be effective for this purpose in the past and are widely used for learning sequences of spectral representations obtained from consecutive segments of time-series data [42] [43] [44].

Each LSTM unit contains a cell state, which holds information contained in previous units allowing the network to learn temporal relationships. This cell state is part of the LSTM’s memory unit, where there lie several gates that together control which information from inputs, as well as from the previous cell and hidden states, will be used to generate the current cell and hidden states. Namely, the forget gate, f_t , and input gate, i_t , are utilized to learn what information from each of these respective states will be

saved within the new current state, C_t . This is shown by the following equations:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$

$$C_t = f_t * C_{t-1} + i_t * \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$$

where σ represents the sigmoid function, and the $*$ operator denotes point-wise multiplication. This new cell state, along with an output gate, o_t , are used to generate the hidden state of the unit, h_t , as represented by:

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$$

$$h_t = o_t * \tanh(C_t)$$

The cell state and hidden state are then passed to successive LSTM units, allowing the network to learn long-term dependencies.

We used a BLSTM network [45] in **FluentNet**. BLSTMs consist of two LSTMs advancing in opposite directions, maximizing the available context from relationships of both the past and future. The outputs of these two networks are multiplied together into a single output layer. **FluentNet** consists of two consecutive BLSTMs, each utilizing LSTM cells with 512 hidden units. A dropout [46] of 20% was also applied to each of these recurrent layers. To avoid overfitting given the size of the dataset, the randomly masked neurons caused by dropout forces the model to be trained using a sparse representation of the given data.

4) Attention: The recent introduction of attention mechanisms

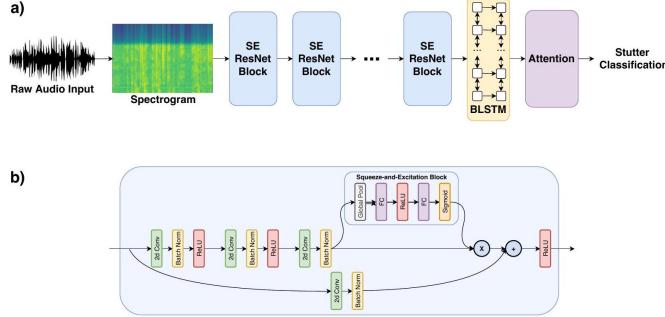


Fig. 2: a) A full workflow of **FluentNet** is presented. This network consists of 8 SE residual blocks, two BLSTM layers, and a global attention mechanism. b) The breakdown of a single SE-ResNet block in **FluentNet** is presented.

[47] and its subsequent variations [48] have allowed for added focus on more salient sections of the learned embedding. These mechanisms have recently been applied to speech recognition models to better focus on strong emotional characteristics within utterances [49] [50], and have similarly been used in **FluentNet** to improve focus on specific parts of utterances with disfluent attributes. **FluentNet** uses global attention [51], which incorporates all hidden state values of the encoder (in this case the BLSTM). A diagram showing the attention model is presented in Figure 3.

The final output value of the second layer of the BLSTM, h_t , as well as a context vector, C_t , derived through the use of the attention mechanism are used to generate **FluentNet**'s final classification, \tilde{h}_t . This is done by applying a tanh activation function, as shown by:

$$\tilde{h}_t = \tanh(W_c [C_t; h_t])$$

The context vector of the global attention is the weighted sum of all hidden state outputs of the encoder. An alignment vector, generated as a relation between h_t and each hidden state value is passed through a softmax layer, which is then used to represent the weights to the context vector. Dot product was used as the alignment score function for this attention mechanism. The calculation for the context vector can be represented by:

$$C_t = \sum_{i=1}^t \bar{h}_i \left(\frac{e^{h_t^\top \cdot \bar{h}_i}}{\sum_{i=1}^t e^{h_t^\top \cdot \bar{h}_i}} \right)$$

where \bar{h}_i represents the i th BLSTM hidden state's output.

C. IMPLEMENTATION

FluentNet was implemented using Keras [52] with a Tensorflow [53] backend. The model was trained with a learning rate of 10^{-4} yielded the strongest results. A root mean square propagation (RMSProp) optimizer, and a binary cross-entropy loss function were used. All experiments were trained using an Nvidia GeForce GTX 1080 Ti GPU. Python's Librosa library [54] was used for audio importing and manipulation towards creating our synthetic dataset as described later. Each STFT spectrogram was generated using four-second audio clips. This length of time can encapsulate

any stutter apparent in the dataset, with no stutters lasting longer than four seconds.

IV. EXPERIMENTS

A. DATASETS

Despite an abundance of datasets for speech-related tasks such as ASR and speaker recognition [20] [55] [56], there is a clear lack of corpora that are focused on speech disfluencies. An ideal speech disfluency dataset would require the labelling and categorization of each existing disfluent utterance. In this paper, to tackle this problem, in addition to using the UCLASS dataset which is a commonly used stuttered speech corpus [57] [58] [17], a second dataset was created through adding speech disfluencies into clean speech. This synthetic corpus contributes a drastic expansion to the available training and testing data for disfluency classification. Through the following subsections, we describe the UCLASS dataset used in our study, as well as the approach for creating the synthetic dataset, LibriStutter, which we created using the original nonstuttered LibriSpeech dataset.

- 1) UCLASS: The University College London Archive of Stuttered Speech (UCLASS) [18] is the most commonly used dataset for disfluency-related studies with machine learning. This corpus came in two releases, in 2004 and 2008, from the university's Division of Psychology and Language Sciences. The dataset consists of 457 audio recordings including monologues, readings, and conversations of children with known stutter disfluency issues. Of those recordings, a select few contain written transcriptions of their respective audio files; these were either standard, phonetic or orthographic transcriptions. Orthographic format is the best option for manual labelling of the dataset for disfluency as they try to transcribe the exact sounds uttered by the speaker in the form of standard alphabet. This helps to identify the presence of disfluency in an utterance more easily. The resulting applicable data consisted of 25 unique conversations between an examiner and a child between the ages of 8 and 18, totalling to just over one hour of audio.

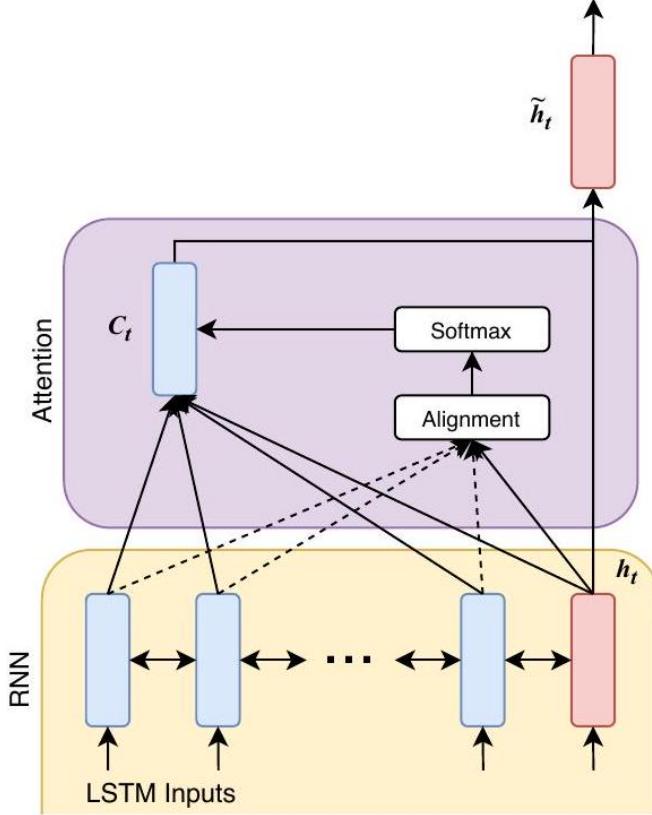


Fig. 3: Global attention addition to binary classifier of recurrent network.

In order to pair the utterances with their transcriptions, each audio file and its corresponding orthographic transcription were passed through a forced time alignment tool. The resulting table related each alphabetical token in the transcription to its matching timestamp within the audio. This process was then manually checked for outlaying utterances not matching their transcriptions.

The provided orthographic transcriptions only flagged the existence of disfluencies (through the use of capitalization), but gave no information towards a disfluency type. To build a more detailed dataset and be able to classify the type of disfluency, all utterances were manually labelled as one of the seven represented classes for our model. These included clean (no stutter), sound repetitions, word repetitions, phrase repetitions, revisions, interjections, and prolongations. The annotation methods applied in [22] and [23] were used as guidelines when manually categorizing each utterance. Out of the 8 disfluencies, 6 were used: sound, word, and phrase repetitions, as well as revisions, interjections, and prolongations.. Of the usable audio in the dataset, only three instances of 'part-word repetitions' appeared, lacking sufficient positive training samples to feasibly classify these types of stutters. As 'block disfluencies' exist as the absence of sound, they could not feasibly be represented in the orthographic transcriptions, which represent how utterances are performed.

2) LibriStutter: The 2015 LibriSpeech ASR corpus by Panayotov et al. [20] includes 1000 hours of prompted English

speech extracted from audio books derived from the LibriVox project. We used this dataset as the basis for our synthetic stutter dataset, which we name LibriStutter. LibriStutter's creation compensates for two shortcomings of the UCLASS corpus: the small amount of labelled stuttered speech data available and the imbalance of the dataset (several disfluency types in UCLASS consisted of a small number of samples).

To allow for a manageable size for LibriStutter and feasible training times, we used a subset of LibriSpeech and set the size of LibriStutter to 20 hours. LibriStutter includes synthetic stutters for sound repetitions, word repetitions, phrase repetitions, prolongations, and interjections. We generated these stutter types by sampling the audio within the same utterance, the details of which are described below. Revisions were excluded from LibriStutter, as this disfluency type requires the speaker to change and revise what was initially said. This would require added speech through the use of complex language models and voice augmentation tools to mimic the revised phrase, both of which fall out of scope for this project.

For each audio file selected from the LibriSpeech dataset, we used the Google Cloud Speech-to-Text API [59] to generate a timestamp corresponding to each spoken word. For every four-second window of speech within a given audio file, either a random disfluency type was inserted and labelled accordingly, or alternatively left clean. Each disfluency type underwent a number of processes to best simulate natural

stutters.

All repetition stutters relied upon copying existing audio segments already present within each audio file. Sound repetitions were generated by copying the first fraction of a random spoken word within the sample and repeating this short utterance a several times before said word. Although repetitions of sounds can occur at the end of words, known as word-final disfluencies, this is rarely the case [60]. One to three repeated sound utterances were added in each stuttered word. After each instance of the repeated sound, a random empty pause duration of 100 to 350ms was appended as this range sounded most natural. Inserted audio may leave sharp cutoffs, especially part-way through an utterances. To avoid this, interpolation was used to smooth the added audio's transition into the existing clip.

Both word and phrase repetitions underwent similar processes to that of sound repetitions. For word repetitions we repeated one to two copies of a randomly selected word before the original utterance. For phrase repetitions, a similar approach was taken, where instead of repeating a particular word, a phrase consisting of two to three words were repeated. The same pause duration and interpolation techniques used for sound repetitions were applied to both word and phrase repetition disfluencies.

Prolongations consist of sustained sounds, primarily at the end of a word. To mimic this behaviour, the last portion of a word was stretched to simulate prolonged speech. For a randomly chosen word, the latter 20% of the signal was stretched by a factor of 5. This prolonged speech segment replaced the original word ending. As applying time stretching to audio results in a drop in pitch, pitch shifting was used to realign the pitch with the original audio. The average pitch of the given speech segment was used to normalize the disfluent utterance.

Unlike the aforementioned classes, interjection disfluencies cannot be created from existing speech within a sample as it requires the addition of filler words absent from the original audio (for example 'umm'). Multiple samples of common filler words from the UCLASS were isolated and saved separately to create a pool of interjections. To create interjection disfluencies, a random filler word from this pool was inserted between two random words, followed by a short empty pause. The same pitch scaling and normalization method as used for prolongations was applies to match the pitches between the interjection and audio clip. Interpolation was used as in repetition disfluencies to smooth sharp cutoffs caused by the added utterance.

To ensure that sufficient realism was incorporated into the dataset, a registered speech language pathologist was consulted for this project. Nonetheless, it should be mentioned that despite our attention to creating a perceptually valid and realistic dataset, the notion of "realism" itself is not a focus of this dataset. Instead, much like synthetic datasets in other areas such as image processing, the aim is for the dataset to be valid enough such that machine learning and deep learning methods can be trained and evaluated with, and later on transferred to real large-scale datasets [in the future] with

little to no adjustments to the model architectures.

Figure 4 displays side by side comparisons of spectrograms of real stuttered data from the UCLASS dataset, and artificial stutters from LibriStutter. Each pairing represents a single stutter type, with the same word or sound being spoken in each. It can be observed that the UCLASS stutters samples and their corresponding LibriStutter examples show clear similarities. Moreover, to numerically compare the samples, cosine similarity [61] was calculated between the UCLASS and LibriStutter spectrogram samples shown earlier. To add relevance to these values, a second comparison was performed for each UCLASS spectrogram with respect to 100 random samples from the LibriStutter dataset, and the average score was used as the represented comparison value. These scores are summarized in Table III. We observe that the UCLASS cosine similarity scores corresponding to the matching LibriStutter samples are noticeably (approximately between 10× to 30×) greater than those compared to random audio samples, confirming that the disfluent utterances contained in LibriStutter share phonetic similarities with real stuttered samples, empirically showing the similarity between a few sample real and synthesized stutters.

The LibriStutter dataset consists of approximately 20 hours of speech data from the LibriSpeech train-clean-100 (training set of 100 hours "clean" speech). In turn, LibriStutter shares a similar make up to that of its predecessor. It consists of disfluent prompted English speech from audiobooks. It also contains 23 male and 27 female speakers, with an approximate 53% of the audio coming from males, and 47% from females. There are 15000 disfluencies in this dataset, with equal counts for each of the five disfluency types: 3000 sound, word, and phrase repetitions, as well as prolongations and interjections. Each audio file has a corresponding CSV file containing each word or utterance spoken, the start and end time of the utterance, and its disfluency type, if any.

B. BENCHMARKS

For a thorough analysis of our results, we compare the results obtained by the proposed **FluentNet** to a number of other models. In particular, we employ two type of solutions for comparison purposes. First, we compare our results to related works and the state-of-the-art as follows:

Alharbi et al. [17]: This work conducted classification of sound repetitions, word repetitions, revisions, and prolongations on the UCLASS dataset through the application of two different methods. First, an original speech prompt was aligned, and then passed to a task-oriented FST to generate word lattices. These lattices were used to detect repeated part-words, words, and phrases within the sample. This method scored perfect results on word repetition classification, though the results on sound repetitions and revisions proved much weaker. To classify prolongation stutters, an autocorrelation algorithm consisting of two thresholds was used: the first to detect speech with similar amplitudes (sustained speech), and another dynamic threshold to decide whether the duration of similar speech would be considered a prolongation.

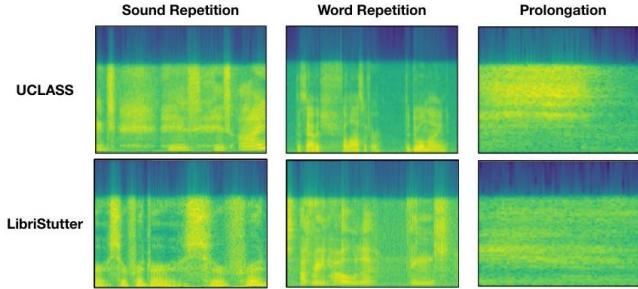


Fig. 4: Spectrograms of the same stutters found in the UCLASS dataset and generated in the LibriStutter dataset.

Using this algorithm, perfect prolongation classification was achieved.

Chen et al. [31]: A CT-Transformer was designed to conduct repetition and interjection disfluency detection on an in-house Chinese speech dataset. Both word and position embeddings of a provided audio sample were passed through a series of CT self attention layers and fully connected layers. This work was able to achieve an overall disfluency classification miss rate of 38.5% (F1 score of 70.5). Notably, this is one of the few works to have attempted interjection disfluency classification, yielding a miss rate of 25.1%. Note that the performance on repetition disfluencies encompasses all repetition-type stutters, including sound, word, and phrase repetitions, as well as revisions.

Kourkounakis et al. [33]: As opposed to other current models focusing on ASR and language models, our previous work proposed a model relying solely on acoustic and phonetic features, allowing for the classification of several multiple disfluencies types without the need for speech recognition methods. This model applied a deep residual network, consisting of 6 residual blocks (18 convolution layers) and two bidirectional long short-term memory layers to classify six different types of stutters. This work achieved an average miss rate of 10.03% on the UCLASS dataset, and sustained strong accuracy and miss rates across all stutter types, prominently word repetitions and revisions.

Zayats et al. [28]: A recurrent network was used to classify repetition disfluencies within the Switchboard corpus. It consists of a BLSTM followed by an ILP post processing method. The input embedding to this network consisted of a vector containing each word's index, part of speech, as well as 18 other disfluency-based features. The method achieved a miss rate of 19.4% across all repetitions disfluencies.

Villegas et al. [29]: This model was used a reference to compare the effectiveness of repository signals towards stutter classification. These features included the means, standard deviations, and distances of respiratory volume, respiratory flow, and heart rate. Sixty-eight participants were used to generate the data for their experiments. The best performing model in this work was an MLP with 40 hidden layers, resulting in a 82.6% average classification accuracy between block and non-block type stutters.

Dash et al. [16]: This method passed the maximum amplitude of the provided audio sample through a neural network

to generate a custom threshold for each sample, trained on a set of 60 speech samples. This amplitude threshold was used to remove any perceived prolongations and interjections. The audio was then passed the audio through a SST tool, which allowed for the removal of any repeated words, phrases, or characters, achieving an overall stutter classification of 86% on a test set of 50 speech segments.

Note that the latter three works only provide results on a group of disfluency types [28], a single disfluency type [29], or overall stutter classification [16]. As such, only their average disfluency classification results could be compared. Moreover, these works ([31], [28], [29], and [16]) have not used the UCLASS dataset, therefore the comparisons should be taken cautiously.

Next, we also compare the performance of our solution to a number of other models for benchmarking purposes. These models were selected due to their popularity for timeseries learning and their hyperparameters of these models are all tuned to obtain the best possible results given their architectures. These benchmarks are as follows: (i) VGG-16 (Benchmark 1): VGG-16 [62] consists of 16 convolutional or fully connected layers, comprised of groups of two or three convolution layers with ReLU activation, with each grouping being followed by a max pooling layer. The model concludes with three fully connected layers and a final softmax function. (ii) VGG-19 (Benchmark 2): This network is very similar to its VGG-16 counterpart, with the only difference being an addition of three more convolution layers spread throughout the model. (iii) ResNet-18 (Benchmark 3): ResNet18 was chosen as a benchmark, which contains 18 layers: eight consecutive residual blocks each containing two convolutional layers with ReLU activation, followed by an average pooling layer and a final fully connected layer.

V. RESULTS AND ANALYSIS

A. VALIDATION

In order to rigorously test **FluentNet** on the UCLASS dataset, a leave-one-subject-out (LOSO) cross validation method was used. The results of models tested on this dataset are represented as the average between 25 experiments, each consisting of audio samples from 24 of the participants as training data, and a unique single participant's audio as a test set. A 10 -fold cross validation method was used on the LibriStutter dataset with a random 90% subset of the samples

from each stutter being used for training along with 90% of the clean samples chosen randomly. The remaining 10% of both clean and stuttered samples were used for testing. All experiments were trained over 30 epochs, with minimal change in loss seen in further epochs.

The two metrics used to measure the performance of the aforementioned experiments were miss rate and accuracy. Miss rate ($1 - \text{recall}$) is used to determine the proportion of disfluencies which were incorrectly classified by the model. To balance out any bias this metric may hold, accuracy was used as a second performance metric.

B. PERFORMANCE AND COMPARISON

The results of our model for recognition of each stutter type are presented for the UCLASS and LibriStutter datasets in Table IV. **FluentNet** achieves strong results against all the disfluency types within both datasets, outperforming nearly all of the related work as well as the benchmark models.

As some previous works have been designed to tackle specific disfluency types as opposed to a general solution for detecting different types of disfluencies, a few of **FluentNet**'s individual class accuracies do not surpass previous works', namely word repetitions and prolongation. In particular, the work by Alharbi et al. [17] offers perfect word repetition classification, as word lattices can easily identify two words repeated one after the other. Amplitude thresholding also proves to be a successful prolongation classification method. It should be noted that **FluentNet** does achieve strong results for these disfluency types as well. Notably, our work is one of the only ones that has attempted classification of interjection disfluencies. These disfluent utterances lack the unique phonetic and temporal patterns that, for instance, repetition or prolongation disfluencies contain. Moreover, they may be present as a combination of other disfluency types, for example an interjection can be both prolonged or repeated. For these reasons, interjections remain the hardest category, with a 24.05% and 29.78% miss rate on the UCLASS and LibriStutter datasets, respectively. Nonetheless, **FluentNet** provides good results, especially given that interjections have been historically avoided.

The task-oriented lattices generated in [17] show strong performance on word repetitions and prolongations, but struggle to detect sound repetitions and revision. Likewise, as is presented in [31], the CT-Transformer yields a comparable interjection classification miss rate to that of **FluentNet**. However, when the same model is applied to repetition stutters, the performance of the model drops severely, hindering its overall disfluency detection capabilities. The use of an attention-based transformer proves a viable method of classifying interjection disfluencies, however as the results suggest, the convolutional and recurrent architecture in **FluentNet** allows for effective representations to be learned for interjection disfluencies alongside repetitions and prolongations.

FluentNet's achievements surpass our previous work's across all disfluency types on the Libristutter dataset, and all but word repetition accuracy on the UCLASS dataset. The results show a greater margin of improvement against the

LibriStutter dataset as compared to UCLASS between the two models. Notably, word repetitions and prolongation relay a decrease in miss rate of approximately 20% between **FluentNet** and [33]. This implies the SE and attention mechanisms assist in better representing the disfluent utterances within stuttered speech found in the synthetic dataset.

An interesting observation is that LibriStutter proves a more difficult dataset compared to UCLASS as evident by the lower performance of all the solutions including **FluentNet**. This is likely due to the fact that given the large number of controllable parameters for each stutter type, LibriStutter is likely to contain a larger variance of stutter styles and variations, resulting in a more difficult problem space.

Table V presents the overall performance of our model with respect to all disfluency types on UCLASS and LibriStutter datasets. The results are compared with other works on respective datasets, and the benchmarks which we implemented for comparison purposes. We observe that **FluentNet** achieves average miss rates and accuracy of 9.35% and 91.75% on the UCLASS dataset, surpassing the other models and setting a new state-of-the-art. A similar trend can be seen for the LibriStutter dataset where **FluentNet** outperforms the previous model along with all the benchmark models.

The BLSTM used in [28] yields successful results towards repetition stutter classification by learning temporal relationships between words, however it remains impaired by its reliance solely on lexical model inputs. On the other hand, as shown by the results, **FluentNet** is better able to learn these phonetic details through the spectral and temporal representations of speech.

The work from [16] uses similar classification techniques to [17], however improves upon the thresholding technique with the addition of a neural neural network. Though achieving an average accuracy of 86% across the same disfluency types used in this work, **FluentNet** remains a stronger model given its effective spectral frame-level and temporal embeddings. Nonetheless, the results of this work contains only a single overall accuracy value across all of repetition, interjection, and prolongation disfluency detection. Little is discussed on the origin and makeup of the dataset used.

Of the benchmark models without an RNN component, ResNet performs better than both VGG networks for both datasets, indicating that ResNet-style architectures are able to learn effective spectral representations of speech. This further justifies the use of a ResNet as the backbone of our

model. Moreover, the addition of the LSTM component to the benchmarks shows that learning the temporal relationships through an RNN contributes to the performance.

To further demonstrate the performance of **FluentNet**, the Receiver Operator Characteristic (ROC) curves were generated for each disfluency class on the UCLASS and LibriStutter datasets, as shown in Figures 5(a) and 5(b), respectively. It can be seen that word repetitions, phrase repetitions, revisions, and prolongations reveal very strong classification on both datasets. Both sound repetitions and interjections classification fair weakest, with the LibriStutter dataset, proving to be a more difficult dataset for **FluentNet**,

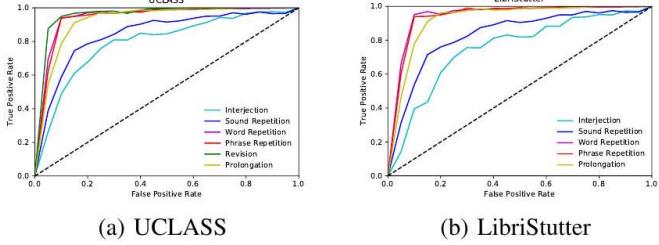


Fig. 5: ROC curves for each stutter type tested on the UCLASS and LibriStutter datasets.

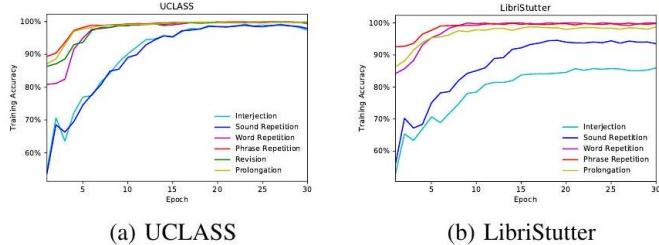


Fig. 6: Average training accuracy for **FluentNet** on the considered stuttered types for the UCLASS and LibriStutter datasets.

as previously observed and discussed.

C. PARAMETERS

Multiple parameters have been tuned in order to maximize the accuracy of **FluentNet** and the baseline experiments on both datasets. These include convolution window sizes, epochs, and learning rates, among others. Each has been individually tested in order to find the optimal values for the given model. Note that all of **FluentNet**'s hyper-parameters remain the same across all disfluency types.

Thorough experiments were performed to obtain the optimum architecture of **FluentNet**. For the SE-ResNet component, we tested a different count of convolution blocks, ranging between 3 to 12, with each block consisting of 3 convolutional layers. Eight blocks were found to be the approximate optimal depth for training the model on the

UCLASS dataset. Similarly, we experimented with the use of different number of BLSTM layers, ranging between 0 to 3 layers. The use of 2 layers yielded the best results. Moreover, the use of bi-directional layers proved slightly more effective than uni-directional layers. Lastly, we experimented with a number of different values and strategies for the learning rate where 10^{-4} showed the best results.

Figures 6(a) and 6(b) show **FluentNet**'s performance for each stutter type against different epochs on the UCLASS and LibriStutter datasets, respectively. It can be seen that the training accuracy stabilizes after around 20 epochs. Whereas all disfluencies types in the UCLASS dataset approach perfect training accuracy, training accuracy plateaus at much lower accuracies for interjections and sound repetitions within the LibriStutter dataset.

D. ABLATION EXPERIMENTS

To further analyze **FluentNet**, an ablation study was done in order to systematically evaluate how each component con-

tributes towards the overall performance. Both the SE portion and attention mechanisms were removed, individually and together, in order to analyse the relationship between their absences, and how these affect both accuracy and miss rates for each disfluency class. The ablation results for both the UCLASS and LibriStutter datasets can be seen summarized in Table VI. Overall, **FluentNet** shows stronger accuracy and lower miss rates across both datasets and all stutter types, compared to the three variants. Although the drops in performance varies across different stutter types with the removal of each element, the experiment shows the general advantages of the different components of **FluentNet**.

The results show that across both datasets, the SE component and the attention mechanism both individually benefit the model for most stutter types. Removal of the SE component yields the greatest drop in the accuracy and increase in miss rates across nearly all stutter types. The removal of the SE components from **FluentNet** has the most negative impact. The removal of the global attention mechanism as the final stage of the model, also reduces the classification accuracy of **FluentNet**. Similarly, with both the SE component and attention removed, the model showed a decline in accuracy and miss rates across all classes tested. Note that the results of these ablation experiments hold similar conclusions for both the UCLASS and our synthesized dataset (with a slightly higher impact observed on UCLASS vs. LibriStutter), thereby

reinforcing the validity of LibriStutter's similarity to real stutters.

VI. CONCLUSION

Of the measurable metrics of speech, stuttering continues to be the most difficult to identify as their diversity and uniqueness make them challenging for simple algorithms

to model. To this end, we proposed a deep neural network, **FluentNet**, to accurately classify these disfluencies.

FluentNet is an end-to-end deep neural network designed to accurately classify stuttered speech across six different stutter types: sound, word, and phrase repetitions, as well as revisions, interjections, and prolongations. This model uses a Squeeze-and-Excitation residual network to learn effective spectral frame-level speech representations, followed by recurrent bidirectional long shortterm memory layers to learn temporal relationships from stuttered speech. A global attention mechanism was then added to focus on salient parts of speech in order to accurately detect the required influences. Through comprehensive experiments, we demonstrate that **FluentNet** achieves state-of-the-art results on disfluency classification with respect to other works in the area as well as a number of benchmark models on the public UCLASS dataset. Given the lack of sufficient data to facilitate more in-depth research on disfluency detection, we developed a synthetic dataset, LibriStutter, based on the public LibriSpeech dataset.

Future works may include improving on LibriStutter's realism, which could constitute conducting further research into the physical sound generation of stutters and how they translate to audio signals. Whereas this work focuses on the educational and business applications of speech metric analysis, further work may focus towards medical and therapeutic use-cases.

ACKNOWLEDGMENT

The authors would like to thank Prof. Jim Hamilton for his support and valuable discussion throughout this work. We also wish to acknowledge Adrienne Nobbe for her consultation towards this project.

REFERENCES

- [1] S. H. Ferguson and S. D. Morgan, "Talker differences in clear and conversational speech: Perceived sentence clarity for young adults with normal hearing and older adults with hearing loss," *Journal of Speech, Language, and Hearing Research*, vol. 61, no. 1, pp. 159-173, 2018.
- [2] J. S. Robinson, B. L. Garton, and P. R. Vaughn, "Becoming employable: A look at graduates' and supervisors' perceptions of the skills needed for employability," in *NACTA Journal*, vol. 51, 2007, pp. 19-26.
- [3] Mayo Foundation for Medical Education and Research. (2017) Stuttering. [Online]. Available: <https://www.mayoclinic.org/diseases-conditions/stuttering/diagnosis-treatment/drc-20353577>
- [4] H. Trinh, R. Asadi, D. Edge, and T. Bickmore, "Robocop: A robotic coach for oral presentations," *ACM Conference on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 1, no. 2, p. 27, 2017.
- [5] ASHA. (2020) Childhood fluency disorders. [Online]. Available: <https://www.asha.org/Practice-Portal/Clinical-Topics/Childhood-Fluency-Disorders>
- [6] United Kingdom National Health Service. (2019) Stammering. [Online]. Available: <https://www.nhs.uk/conditions/stammering/>
- [7] Anxiety and Depression Association of America. (2019). [Online]. Available: <https://adaa.org/understanding-anxiety/social-anxiety-disorder/treatment/conquering-stage-fright>
- [8] Y. Wang, A. Mohamed, D. Le, C. Liu, A. Xiao, J. Mahadeokar, H. Huang, A. Tjandra, X. Zhang, F. Zhang et al., "Transformer-based acoustic modeling for hybrid speech recognition," in *ICASSP 20202020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6874-6878.
- [9] Y. He, T. N. Sainath, R. Prabhavalkar, I. McGraw, R. Alvarez, D. Zhao, D. Rybach, A. Kannan, Y. Wu, R. Pang et al., "Streaming end-to-end speech recognition for mobile devices," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6381-6385.
- [10] A. Hajavi and A. Etemad, "A deep neural network for short-segment speaker recognition," *INTERSPEECH*, 2019.
- [11] D. Snyder, D. Garcia-Romero, G. Sell, A. McCree, D. Povey, and S. Khudanpur, "Speaker recognition for multi-speaker conversations using x-vectors," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5796-5800.
- [12] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3617-3621.
- [13] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, "Neural speech synthesis with transformer network," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 6706-6713.
- [14] P. Howell and S. Sackin, "Automatic recognition of repetitions and prolongations in stuttered speech," *Proceedings of the First World Congress on Fluency Disorders*, 011995.
- [15] P. Howell, S. Sackin, and K. 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*, 101997.
- [16] A. Dash, N. Subramani, T. Manjunath, V. Yaragarala, and S. Tripathi, "Speech recognition and correction of a stuttered speech," in *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 2018, pp. 1757-1760.
- [17] S. Alharbi, M. Hasan, A. Simons, S. Brumfitt, and P. Green, "A lightly supervised approach to detect stuttering in childrens speech," *INTERSPEECH*, pp. 3433-3437, 2018.
- [18] P. Howell, S. Davis, and J. Bartrip, "The university college london archive of stuttered speech (uclass)," *Journal of Speech, Language, and Hearing Research*, vol. 52, pp. 556-569, 2009.

- [19] T. Tan, Helbin-Liboh, A. K. Ariff, C. Ting, and S. Salleh, "Application of malay speech technology in malay speech therapy assistance tools," International Conference on Intelligent and Advanced Systems, pp. 330334, 2007.
- [20] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015, pp. 5206-5210.
- [21] N. Zeghidour, Q. Xu, V. Liptchinsky, N. Usunier, G. Synnaeve, and R. Collobert, "Fully convolutional speech recognition," arXiv preprint arXiv:1812.06864, 2018.
- [22] E. Yairi and N. G. Ambrose, "Early childhood stuttering i: persistency and recovery rates," Journal of Speech, Language, and Hearing Research, vol. 42, 1999.
- [23] F. S. Juste and C. R. F. de Andrade, "Speech disfluency types of fluent and stuttering individuals: Age effects," International Journal of Phoniatrics, Speech Therapy and Communication Pathology, vol. 63, 2011.
- [24] Stuttering Foundation. (2020) Differential diagnosis. [Online]. Available: <https://www.stutteringhelp.org/differential-diagnosis>
- [25] K. Ravikumar, S. Kudva, R. Rajagopal, and H. Nagaraj, "Development of a procedure for the automatic recognition of disfluencies in the speech of people who stutter," in International Conference on Advanced Computing Technologies, Hyderbad, India, 2008, pp. 514-519.
- [26] H. K.M Ravikumar, R.Rajagopal, "An approach for objective assessment of stuttered speech using mfcc features," Digital Signal Processing Journal, vol. 9, pp. 19-24, 2019.
- [27] L. S. Chee, O. C. Ai, and S. Yaacob, "Overview of automatic stuttering recognition system," in Proc. International Conference on Man-Machine Systems, no. October, Batu Ferringhi, Penang Malaysia, 2009, pp. 1-6.
- [28] V. Zayats, M. Ostendorf, and H. Hajishirzi, "Disfluency detection using a bidirectional 1stm," INTERSPEECH, pp. 2523-2527, 2016.
- [29] B. Villegas, K. M. Flores, K. Jos Acua, K. Pacheco-Barrios, and D. Elias, "A novel stuttering disfluency classification system based on respiratory biosignals," in 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2019, pp. 46604663.
- [30] J. Santoso, T. Yamada, and S. Makino, "Classification of causes of speech recognition errors using attention-based bidirectional long shortterm memory and modulation spectrum," in 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2019, pp. 302-306.
- [31] Q. Chen, M. Chen, B. Li, and W. 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), 2020, pp. 8069-8073.
- [32] K. Georgila, "Using integer linear programming for detecting speech disfluencies," in Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers, 2009, pp. 109-112.
- [33] T. Kourkounakis, A. Hajavi, and A. Etemad, "Detecting multiple speech disfluencies using a deep residual network with bidirectional long shortterm memory," in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020, pp. 6089-6093.
- [34] S. Khara, S. Singh, and D. Vir, "A comparative study of the techniques for feature extraction and classification in stuttering," in 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), 2018, pp. 887-893.
- [35] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," IEEE Conference on Computer Vision and Pattern Recognition, pp. 770-778, 2016.
- [36] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," CoRR, vol. abs/1709.01507, 2017. [Online]. Available: <http://arxiv.org/abs/1709.01507>
- [37] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," arXiv preprint arXiv:1602.07261, 2016.
- [38] A. G. Roy, N. Navab, and C. Wachinger, "Concurrent spatial and channel squeeze & excitationin fully convolutional networks," in International conference on medical image computing and computer-assisted intervention. Springer, 2018, pp. 421-429.
- [39] S. Bell, C. Lawrence Zitnick, K. Bala, and R. Girshick, "Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2874-2883.
- [40] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in European conference on computer vision. Springer, 2016, pp. 483-499.
- [41] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735-1780, 1997.
- [42] Y. Ma, H. Peng, and E. Cambria, "Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive lstm." in Aaai, 2018, pp. 5876-5883.
- [43] H. Y. Kim and C. H. Won, "Forecasting the volatility of stock price index: A hybrid model integrating lstm with multiple garch-type models," Expert Systems with Applications, vol. 103, pp. 25-37, 2018.
- [44] P. Li, M. Abdel-Aty, and J. Yuan, "Real-time crash risk prediction on arterials based on 1stm-cnn," Accident Analysis & Prevention, vol. 135, p. 105371, 2020.
- [45] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," IEEE Transactions on Signal Processing, vol. 45, no. 11, pp. 2673-2681, 1997.
- [46] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent

- neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929-1958, 2014.
- [47] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," arXiv preprint arXiv:1409.0473, 2014.
- [48] A. Hajavi and A. Etemad, "Knowing what to listen to: Early attention for deep speech representation learning," arXiv preprint arXiv:2009.01822, 2020.
- [49] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017, pp. 2227-2231.
- [50] T. Sun and A. A. Wu, "Sparse autoencoder with attention mechanism for speech emotion recognition," in 2019 IEEE International Conference on Artificial Intelligence Circuits and Systems (AICAS), 2019, pp. 146-149.
- [51] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," arXiv preprint arXiv:1508.04025, 2015.
- [52] F. Chollet et al. (2015) Keras. [Online]. Available: <https://keras.io>
- [53] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard et al., "Tensorflow: A system for largescale machine learning," in 12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16), 2016, pp. 265-283.
- [54] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in Proceedings of the 14th python in science conference, vol. 8, 2015.
- [55] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "Darpa timit acoustic-phonetic continuous speech corpus cd-rom. nist speech disc 1-1.1," STIN, vol. 93, p. 27403, 1993.
- [56] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," arXiv preprint arXiv:1706.08612, 2017.
- [57] L. S. Chee, O. C. Ai, M. Hariharan, and S. Yaacob, "Mfcc based recognition of repetitions and prolongations in stuttered speech using k-nn and lda," in 2009 IEEE Student Conference on Research and Development (SCOReD), 2009, pp. 146-149.
- [58] O. C. Ai, M. Hariharan, S. Yaacob, and L. S. Chee, "Classification of speech dysfluencies with mfcc and lpc features," *Expert Systems with Applications*, vol. 39, no. 2, pp. 2157-2165, 2012.
- [59] (2020) Google cloud speech-to-text. [Online]. Available: <https://cloud.google.com/speech-to-text/>
- [60] J. Van Borsel, E. Geirnaert, and R. Van Coster, "Another case of wordfinal disfluencies," *Folia phoniatrica et logopaedica*, vol. 57, no. 3, pp. 148-162, 2005.
- [61] J. Han, M. Kamber, and J. Pei, *Data Mining*, 3rd ed. Elsevier Inc., 2012.
- [62] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.