Detecting Multiple Disfluencies from Speech using Pre-linguistic Automatic Syllabification with Acoustic and Prosody Features

Utkarsh Mehrotra, Sparsh Garg, Gurugubelli Krishna and Anil Kumar Vuppala
International Institute of Information Technology, Hyderabad, India
E-mail: utkarsh.mehrotra, sparsh.garg, krishna.gurugubelli@research.iiit.ac.in, anil.vuppala@iiit.ac.in Tel/Fax: +91-9424320192

***Abstract*—In this paper, a new method to detect disfluencies directly from speech is explored. The method makes use of pre-linguistic automatic syllabification - the process of segmenting input speech signals into perceptually distinct syllable-like regions, to develop syllable-level disfluency detection systems. Statistical prosody features related to fundamental frequency, energy and duration are extracted from each syllable-like region and used to train a DNN classifier for automatic detection of speech disfluencies. Further, complementary information useful for the task of disfluency detection is added to the pipeline with the help of acoustic features. A BiLSTM feature extractor is used to get complex acoustic representation from the baseline MFCC features for each syllable-like region. This acoustic representation is concatenated with the prosody features and used in the proposed system for detecting multiple speech disfluencies. Experiments are conducted for four types of disfluencies in the UCLASS and the IIITH-IED datasets to test the proposed disfluency detection system. Overall, it is found that the proposed system gives a detection accuracy of 88.75% for the disfluencies in the UCLASS dataset, whereas for the IIITH-IED dataset, the accuracy obtained is 91.24%, showing the effectiveness of considering perceptually distinct syllable-like regions as representational units for detecting disfluencies.**

I. Introduction

Speech is one of the most important modes of communication for human beings. The efficiency with which humans are able to transfer information through speech depends on its fluency. Fluency is defined as the ease with which a person can enunciate sounds and words to form a message [1]. But, there are instances when abrupt breaks or disruptions affect the normal flow of speech. These disruptions are referred to as speech disfluencies. Disfluencies arise due to many different reasons. For spontaneous or unprepared speech, the speaker has to formulate the message on the fly, thinking and speaking at the same time. In such a scenario, the speaker might need time to think about what has to be spoken next or need to correct an utterance that was spoken wrongly [2]. Such instances lead to the presence of speech disfluencies. Another case is that of stuttered speech. Stuttering is a speech impediment in which the forward flow of a speaker's speech is disrupted by the unintentional presence of disfluencies [3]. Some of the common types of disfluencies occurring in both spontaneous and stuttered speech are as follows -

1) Prolongation - The lengthening of a particular word or part of a word. For example: I want theeee pen (lengthening of /e/ in the).
2) Filled Pause - Filler words like 'um' and 'uh' which do not add any semantic meaning to the utterance. For example: I want um an ice cream.
3) Part-word repetition - Repeating a particular part of a word in order to maintain continuity with what is being spoken. For example: I want th-the pen.
4) Word repetition - Repeating an entire word to maintain the speech flow. For example: Give give me the book, please.
5) Phrase repetition - Repetition of a phrase of the utterance. For example: I am I am going to get it.

Detection of disfluencies in speech is an important task for many applications. In the case of stuttered speech, automatic detection of disfluencies can help Speech Language Pathologists (SLP) to gauge the severity of a person's stutter and recommend proper treatment [4]. For ASR systems, the presence of disfluencies leads to higher word error rates (WER), deteriorating the performance of the systems [5]. Hence, the efficient detection of disfluencies in speech is important for the proper functioning of such applications.

The existing methods for disfluency detection in the literature can be broadly classified into the following three categories - using lexical features extracted from the orthographic transcription for disfluency detection [6], [7]; using features extracted from speech signal to detect disfluencies directly in speech [8], [9]; using a combination of lexical and speech-based features for disfluency detection [10], [11]. Methods relying on lexical features depend heavily on the availability of the correct orthographic transcription. However, these might not always be available due to multiple reasons like unavailability of a proper ASR system, very high WER in the obtained transcriptions etc. In such cases, we have to rely only on the speech signal for performing the detection task. This work focuses on the use of speech-based features to detect disfluencies. A number of works have focused on the detection of disfluencies directly from speech [8], [12]. Prosody features like signal-to-noise ratio (SNR), fundamental frequency (F0) and duration were used to develop a rule-based system for the detection of four types of speech disfluencies

in [13]. The stability of the first four formants and the nasality effect were explored in the voiced regions of speech in [8], [14] to detect filled pause and vowel lengthening. Frame-level automatic disfluency detection was performed in [15] using log Mel-filterbank and F0 contour features for multiple speech disfluencies using SVM and DNN classifiers. In [16], [17] MFCC features and multiple modifications of the normal MFCCs were used for developing and improving the performance of disfluency detection systems. In [18] utterance level detection experiments were performed directly from the speech signal for six types of disfluencies by feeding the spectrogram representation of 4-second audio clips to a Deep Residual Network with BiLSTM classifier.

In this work, we present a new approach for automatic detection of speech disfluencies at the syllable level. Syllables as representational units have been used for disfluency detection in [19], [20] since the duration of a syllable region (approx. 100-800ms) is similar to the duration of disfluent regions in speech. Thus, syllables can be used to model the characteristics of many speech disfluencies (like filled pause, part-word repetition etc.). This is not the case with the conventional 10-30 ms speech frames (too small to capture the characteristics of disfluencies) or the entire speech utterance (too big, disfluencies present only in a small region of the utterance). In this work, instead of using the orthographic transcription to get syllable units, we perform pre-linguistic automatic syllabification to get perceptually distinct syllable-like units directly from the speech signal. Each input audio recording is segmented into chunks of perceptually different syllable-like units. Then for each of these syllable-like units, statistical prosody features relating to intonation pattern, the energy of the unit and duration are extracted to develop baseline automatic disfluency detection systems for four types of speech disfluencies. The disfluencies considered for this work are - filled pause (interjection), prolongation, part-word repetition and word repetition. Further, the prosody features used in the baseline models are combined with acoustic representations extracted for each syllable-like unit to give the proposed disfluency detection models. The detection experiments are performed for the UCLASS and IIITH-IED datasets to evaluate the system performance for stutter as well as spontaneous speech disfluencies. The main contributions of this work are -

- Using pre-linguistic automatic syllabification to develop syllable-level disfluency detection systems. This has not been explored yet in the literature to the best of our knowledge.
- Developing syllable-level automatic disfluency detection systems using prosody features for four types of speech disfluencies. We further combine the prosody features with acoustic representations learned at the syllable level to obtain the proposed disfluency detection models.

The rest of the paper is organised as follows - details about the pre-linguistic automatic syllabification process are presented in Section 2. Section 3 describes the experimental setup used here. This includes the datasets used, the prosody features extracted, and the disfluency detection systems developed. Experiments performed in this study and the results obtained are discussed in Section 4. We conclude by providing a summary of this work and discussing plans for future works in Section 5.

II. Pre-Linguistic Automatic Syllabification

Studies on language acquisition and speech perception consider syllable units as one of the fundamental representations to model the underlying pre-linguistic structure of speech [21], [22]. Pre-linguistic automatic syllabification refers to identifying and segmenting the speech signal into syllable-like chunks, which are perceptually distinct from one another [23]. The identification of phonological syllables (actual syllables in orthographic transcription) from speech would require the incorporation of formal linguistic representation in terms of language-specific rules and language-based constraints, but since pre-linguistic information available from speech only is being used, the syllable-like chunks obtained do not precisely align with the phonological syllable. The perceptual differences in the identified chunks, however, provide important cues for many speech and language-based tasks [24].

Pre-linguistic automatic syllabification has been explored in [23], [25], [26]. Speech features like intensity and voicing property were used in [27] to highlight the perceptual difference in various speech sounds. Existing automatic syllabification algorithms like [28], [29] use low-pass filtered energy and amplitude envelopes to determine the syllable boundaries in the signal. In this work, we use the method proposed in [23] for the task of automatic speech syllabification. The idea behind the work in [23] is the use of sonority as the perceptual correlate to identify syllable boundaries. Sonority is defined as the relative audibility of speech sounds. Different definitions of a 'syllable' have been used in various works [30], [31], but there is a consensus that syllable units are related to the rhythmic fluctuations in the sonority level of speech. A syllable usually consists of a sonority maxima (corresponding to the syllable nucleus, i.e. a vowel) and the sonority decreases as we move from the nucleus to the edges of a syllable. This idea has been used to detect syllable boundaries in the current work.

Since sonority is a perceptual quantity, physical correlates of sonority are needed to identify the syllable structure of speech from sonority. Here, the amplitude modulations in the speech signal have been for this purpose, as done in [23]. The idea is to use harmonic oscillators to entrain the amplitude modulations in speech, modelling the speech parsing mechanism in the cortex of the human ear. Figure 1 shows a block diagram of the automatic syllabification process used here.

First, the raw audio input (sampling frequency 16000 Hz) is filtered using Gammatone filterbank, which consists of 20 logarithmic filters spaced in the region of 100 to 7500 Hz [32] such that the filterbank can capture the auditory filtering

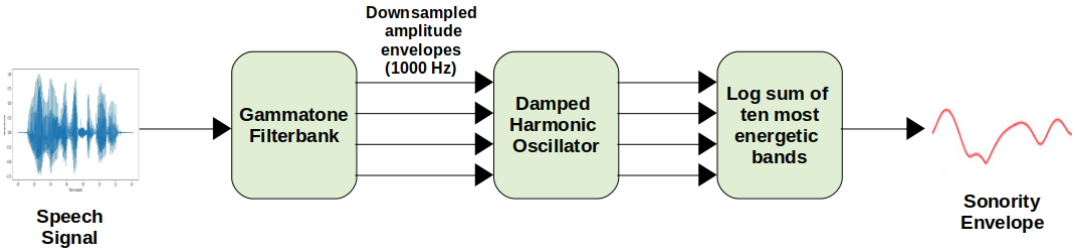

Fig. 1. Block diagram showing the different stages of the syllabification process.

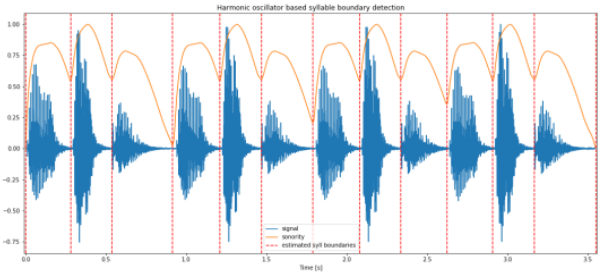

Fig. 2. Figure showing syllabification for a speech file. The orange curve shows the sonority envelope, and the dashed red lines show the syllable boundaries.

characteristics of the human ear. Let $s(t)$ be the speech input. Then -

$$e_c(t) = s(t) * g_c(t), \quad c = 1, 2....., 20 \tag{1}$$

where $g_c(t)$ and $e_c(t)$ are the Gammatone filter and amplitude envelope obtained for the c^{th} frequency band of the filterbank respectively. The amplitude envelopes are then downsampled to f_s = 1000 Hz and passed through a damped harmonic oscillator ($f_0 = 5Hz$ and $\Delta f = 6Hz$), where f_0 is the centre frequency and Δf is the bandwidth of the oscillator. This is done to capture the rhythmic variations in the envelopes—the same oscillator parameters used for each amplitude envelope. The following equations drive the harmonic oscillator:

$$a_c(t) = a_c(t-1) + \frac{u_c(t)}{f_s}, \tag{2}$$

$$u_c(t) = u_c(t-1) + \frac{f_c(t)}{f_s m}, \tag{3}$$

$$f_c(t) = e_c(t) - k a_c(t-1) - d u_c(t-1), \tag{4}$$

where $a_c(t)$, $u_c(t)$ and $f_c(t)$ are the amplitude, velocity and force of the oscillator for frequency band c at the time t. The parameters m, k and d correspond to the mass, spring constant and the damping coefficient of the oscillator, respectively. These parameters are kept fixed for all bands. From the 20 oscillator amplitudes $a_c(t)$ obtained (one for each frequency band) at each time step, the product of 10 most energetic bands is taken to get the sonority envelope $S(t)$. Since the relative sonority values have to be compared to obtain the sonority maxima and minima for syllable boundary detection, the logarithmic sum of amplitudes is taken instead of the product of the amplitudes. Before applying the log function, an offset value ϵ is added to $a_c(t)$, such that the value inside the log function is positive.

$$\sum_{n=1}^{10} \log_{10}(a_n(t) + \epsilon) = S(t) \tag{5}$$

The sonority envelope is then normalized to a range of 0-1,

$$\hat{S}(t) = \frac{S(t) - min(S(t))}{max(S(t)) - min(S(t))}, \tag{6}$$

where $\hat{S}(t)$ is the normalized sonority envelope. From the normalized envelope, each local minima, which is preceded by a local maxima, is marked as the boundary for a syllable-like chunk. Figure 2 shows the sonority envelope and the syllable boundaries detected for the speech signal.

III. Experimental Setup

This section presents presents the details of the UCLASS and IIITH-IED datasets used for the disfluency detection experiments, the prosody features extracted for each syllable-like region, and the disfluency detection system used.

A. UCLASS Dataset

To perform the disfluency detection experiments for stuttered speech, we use the UCLASS Dataset (University College London's Archive of Stuttered Speech) [33]. Speech samples are taken from Release One of the UCLASS dataset, which consists of monologue recordings from participants aged 8 to 18 years in British English. All participants suffer from the stuttering disorder of varying severity. The speech recordings and transcriptions are force aligned to get timestamps for every word and stutter spoken. The recordings are then manually annotated for multiple speech disfluencies using a method similar to [34]. Samples of four types of disfluencies: filled pause, prolongation, part-word repetition (sound repetition samples also included) and word repetition are used for the experiments here.

B. IIITH-IED Dataset

The IIITH-Indian English Disfluency (IIITH-IED) Dataset is a spontaneous speech dataset for the study of speech disfluencies in Indian English. It was introduced in [35]. The dataset consists of 10 hours of speech in Indian English. Lecture recordings available under the NPTEL initiative of the Government of India have been used for the preparation of the dataset. The lectures used in this dataset belong to the following domains: Artificial Intelligence, Computer Science, Database Management etc. Ten-minute audio recordings from 60 lecturers (30 male and 30 female) are taken and manually annotated for multiple types of speech disfluencies. The annotation is performed in the dataset at two levels: the word level (identifying where the disfluencies occur in the transcription) and the signal level (start time and end time for each disfluency occurrence noted). Table I shows the number of occurrences in the IIITH-IED dataset of the four disfluencies used in the experiments in this work. Further details about this dataset can be found here [1].

TABLE I
NUMBER OF OCCURRENCES OF THE FOUR DISFLUENCY TYPES IN IIITH-IED DATASET.

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

C. Prosody Features

The input audio is first passed through the automatic syllabification system, and syllable-like chunks are obtained. For each chunk, statistical prosody features are extracted to develop the baseline disfluency detection system. Local prosody features have been used in [13], [36] for the task of detecting disfluencies. Here, we use features related to the fundamental frequency (F0) contour, energy and duration of each syllable-like unit. Table II shows the list features extracted per syllable region.

The fundamental frequency is estimated here using Praat's auto-correlation based F0 extraction algorithm. A frame size of 10 ms and shift of 5 ms is used for the computation of F0. Energy computation is performed using the Short-time Fourier transform (STFT), with a frame size of 20 ms and a frame shift of 10 ms. The energy of each syllable-like region is obtained by summing up the energies of individual speech samples lying within the region. As we can see from Table II, 15 features corresponding to F0, 9 features corresponding to the energy in the syllable unit and 8 features corresponding to duration and ratios of silence, voiced and unvoiced regions are extracted. This 32-dimensional feature vector is then used as input to our disfluency detection system to decide whether the syllable-like unit has a disfluency in it or not.

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

TABLE II
PROSODY FEATURES EXTRACTED FOR EVERY SYLLABLE REGION. HERE STD. DEV. STANDS FOR STANDARD DEVIATION, AND MSE STANDS FOR MEAN SQUARED ERROR.

Feature	Statistics Computed	Dimension
F0	F0 contour - Average, Std. dev., Maximum, Minimum, Skewness	5
	Tilt of F0 contour - Average, Std. dev., Maximum, Minimum, Skewness	5
	MSE of F0 contour - Average, Std. dev., Maximum, Minimum, Skewness	5
Energy	Energy Contour - Average, Std. dev., Skewness	3
	Tilt of energy contour - Average, Std. dev., Skewness	3
	MSE of energy contour - Average, Std. dev., Skewness	3
Duration	Pause Duration - Average, Std. dev., Minimum, Maximum, Skewness	5
	Ratio of Voiced to unvoiced duration	1
	Ratio of Voiced to pause duration	1
	Ratio of Unvoiced to pause duration	1
Total		**32**

D. Disfluency Detection System

The disfluency detection systems used here are syllable-level detection systems. These systems detect whether a syllable-like unit corresponds to a particular disfluency or not. The detection of each type of disfluency is set up as a binary classification problem to study how efficient the developed systems are for detecting each disfluency type. The classifier used is a DNN with 2 hidden layers, as used in [35]. The first hidden layer has 100 units and the second hidden layer has 50 units in it, with the same architecture being used for all four disfluencies. A 90:10 train-test split is used in the experiments. During training, 10-fold stratified cross-validation is performed, with 9-folds used for training and 1-fold used for validation. Hyperparameter tuning is done on the validation set using grid search, and an optimal learning rate of 10^{-2} is obtained, with the Adam optimizer and binary cross-entropy loss function. Early stopping is applied to the validation loss to avoid overfitting.

The UCLASS and IIITH-IED datasets are biased since the number of syllables corresponding to normal speech are greater than disfluent syllables. So, to ensure that the systems developed are unbiased, Synthetic Minority Oversampling Technique (SMOTE) [37] is used to make the number of samples of each class equal.

E. Evaluation Metrics

To properly assess the performance of the systems developed, two metrics have been used - Accuracy and F1-score. The F1-score is calculated as follows:

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall}, \tag{7}$$

where,

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives}, \quad (8)$$

and,

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives}. \quad (9)$$

IV. Experiments and Results

A. Baseline

The first set of experiments performed are for the baseline results of disfluency detection using the aforementioned statistical prosody features with the DNN disfluency detection system, as described above. For each disfluency type, the experiments are set up as a binary classification task to determine whether a particular syllable-like audio chunk belongs to the disfluency type or not. Table III shows the results obtained for the detection of the four disfluency types in the UCLASS and IIITH-IED datasets. An accuracy of 81.26% is obtained for the detection of filled pause in case of stuttered speech (UCLASS dataset), whereas in the case of spontaneous speech (IIITH-IED dataset), the accuracy obtained is 84.28%.

TABLE III
Baseline results obtained using prosody features for the UCLASS and IIITH-IED Dataset. Here F1 shows the F1-Score.

Disfluency Type	UCLASS		IIITH-IED	
	Accuracy	F1	Accuracy	F1
Filled Pause	81.26	0.812	84.28	0.840
Prolongation	83.71	0.831	87.33	0.869
Part-word repetition	79.45	0.794	80.18	0.794
Word repetition	72.93	0.722	71.19	0.712

From the results, we can observe that the baseline system gives good detection performance in the case of mono-syllabic disfluencies like a filled pause. Most of the prolongations in the UCLASS and IIITH-IED datasets are mono-syllabic as well, so a high detection accuracy is obtained. However, the performance of the system deteriorates in the case of multi-syllabic disfluencies (like word repetition). This is because the features belonging to only one particular syllable region have been used for classification without taking into account the context of neighbouring syllables. So, to consider the context of neighbouring syllables, we stack up features from one syllable region before and after the current syllable (window of ±1 syllables). Hence, a 96-dimensional feature vector ($32 * 3 = 96$) is now used to represent each syllable region. The results obtained by stacking up features with the same detection system are shown in Table IV.

We can see from Table III and Table IV that taking neighbouring syllables into account improves the detection performance for all the four disfluencies. However, the increase in performance is much greater for multi-syllabic disfluencies. For word repetitions in the case of stuttered speech, an improvement of 6.65% in the detection accuracy is obtained when using stacked features compared to the case when no stacking is done, whereas, for the IIITH-IED dataset, the improvement in the case of word repetitions is 5.96%. Improvements are observed for filled pause and prolongation as well, but they are marginal (in the case of stuttered speech, the improvement for filled pause and prolongation are 1.66% and 1.4%, respectively).

TABLE IV
Results obtained by stacking up prosody features from neighbouring syllable regions for the UCLASS and IIITH-IED Dataset. Here F1 shows the F1-Score

Disfluency Type	UCLASS		IIITH-IED	
	Accuracy	F1	Accuracy	F1
Filled Pause	82.92	0.828	85.63	0.855
Prolongation	85.11	0.848	88.97	0.890
Part-word repetition	82.73	0.826	83.41	0.831
Word repetition	79.58	0.796	77.15	0.769

B. Combination with Acoustic Features

Although prosody features provide important correlates for detecting disfluencies, certain other features such as stability of formants, nasality effect of speech etc., which have been shown to be helpful in disfluency detection [8], [9], [14], cannot be captured using statistical prosody features. Frame-level acoustic information has to be added to the system for this purpose. So, to incorporate this complementary information, we use an acoustic representation learned for every syllable-like unit. The following procedure is used to get the acoustic representation:

1) MFCC features are extracted for the input audio using a Hamming window of size 25 ms with 10 ms shift. The first 13 coefficients, their Δ and $\Delta\Delta$ make up the 39-dimensional MFCC features for each speech frame of size 10 ms.
2) MFCC features of the speech frames belonging to the current syllable region, its predecessor and successor (window of ±1 syllables) are concatenated to form a 39xT dimensional feature vector, T being the number of frames.
3) Now, to get a complex acoustic representation of fixed dimensions for each region, a BiLSTM feature extractor is used. The BiLSTM feature extractor helps to take into account the forward and backward context while learning the acoustic representation. The BiLSTM network consists of 2 recurrent layers, each having 90 units. A dropout rate of 0.2 is set for the first recurrent layer to avoid overfitting.
4) Using this feature extraction network, a 90-dimensional acoustic representation is obtained for each syllable-like unit.

The acoustic features are then combined with the statistical prosody features to give a 186-dimensional feature vector. This final feature vector is then used with the DNN classifier to detect disfluencies. Figure 3 shows the entire disfluency

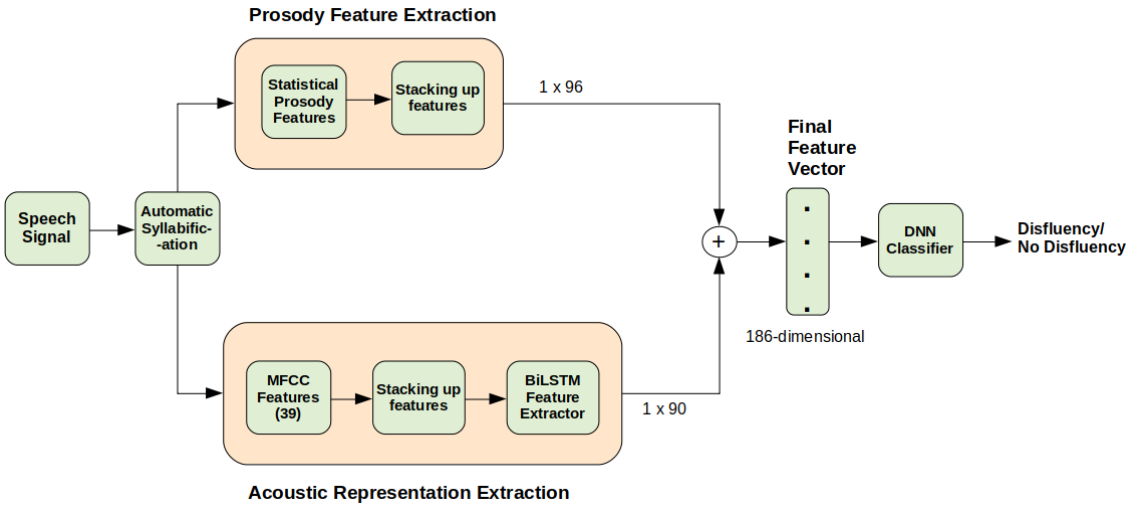

Fig. 3. Pipeline of the proposed disfluency detection system. Prosody and acoustic features are combined to produce the final feature vector.

detection pipeline used when combining acoustic and prosody features. The results obtained by using this setup are reported in Table V.

TABLE V
RESULTS OBTAINED BY COMBINING PROSODY AND ACOUSTIC FEATURES FOR THE UCLASS AND IIITH-IED DATASET. HERE F1 SHOWS THE F1-SCORE

Disfluency Type	UCLASS		IIITH-IED	
	Accuracy	F1	Accuracy	F1
Filled Pause	90.14	0.897	92.46	0.921
Prolongation	92.21	0.919	93.39	0.924
Part-word repetition	87.04	0.866	86.28	0.855
Word repetition	86.21	0.853	81.53	0.802

As can be seen from Table V, the detection performance is enhanced by adding the acoustic features for all the four disfluencies in both datasets. This shows that complementary information is being added to the disfluency detection pipeline through the acoustic features. For filled pause, an absolute improvement of 7.22% and 6.83% is obtained in the detection accuracy for stuttered and spontaneous speech, respectively, compared to the baseline systems. In the case of prolongation, the obtained improvement is 7.1% for stutter disfluencies and 4.42% for spontaneous speech disfluencies. Improvements are obtained in the detection accuracy of repetition type disfluencies as well, but due to high intra-class variability in the samples of repetition disfluencies, the detection performance is lesser as compared to filled pause and prolongation.

Comparing the performance of the proposed method to [18], which uses acoustic features as well to detect disfluencies in the UCLASS dataset, it is found the current system outperforms [18] by a margin of 8.74% in the detection of filled pause. This is because in [18], audio files of fixed duration (4 seconds) are used to extract features using a ResNet. This 4-second duration of each file is large as compared to the duration of filled pause (usually between 100ms - 500ms), so robust modelling of filled pause is difficult. In our proposed method, since acoustic and prosody features are extracted for each syllable-like region (duration of each region is usually between 100 ms - 800 ms), the modelling of filled pause using the extracted features is better. In the case of prolongation and part-word repetition, the performance of our system is comparable to that of [18]. Our proposed system performs better by a small margin of 2.94% for part-word repetition and for prolongation [18] performs better marginally (1.87%). However, [18] outperforms our system in the case of word repetition because occurrences of word repetition can exceed the 3-syllable duration used for extracting features in our experiments. This causes more samples to be misclassified. High detection accuracies are obtained for filled pause and prolongation in [38], [39] as well, but a very small subset of the UCLASS dataset was used. This prevents us from performing a fair comparison with our proposed system.

C. Disfluency vs Non-Disfluency Classification

The proposed disfluency detection system is also evaluated on its performance in discriminating disfluent syllables from non-disfluent or normal syllables. For this, syllable-like units belonging to any of the four disfluency types are considered as one class and the normal syllable units as the other class. Binary classification is then performed to check the system performance in discriminating disfluent syllables from normal syllables. The results of the experiments are reported in Table VI.

The results show that the model performs well in classifying disfluent and normal syllable-like units. A combination of acoustic and prosodic features gives a detection accuracy of 88.75% for the UCLASS dataset. For the IIITH-IED dataset, the obtained accuracy is 91.24%.

TABLE VI
RESULTS OBTAINED FOR THE DISFLUENCY VS NON-DISFLUENCY CLASSIFICATION IN THE UCLASS AND IIITH-IED DATASETS. HERE F1 SHOWS THE F1-SCORE

Input Features	UCLASS		IIITH-IED	
	Accuracy	F1	Accuracy	F1
Prosodic	81.80	0.821	83.89	0.836
Prosodic + Acoustic	**88.75**	**0.886**	**91.24**	**0.914**

V. CONCLUSION AND FUTURE WORKS

In this paper, a new method to detect disfluencies in speech was discussed. Syllable-level disfluency detection was performed using a pre-linguistic automatic syllabification system to segment an input speech signal into perceptually distinct syllable-like units. Statistical prosody features related to F0, energy and duration were extracted from each syllable-like unit and used to train baseline automatic disfluency detection systems for four types of disfluencies. The experiments were conducted for stuttered speech disfluencies (UCLASS dataset) as well as spontaneous speech disfluencies (IIITH-IED dataset). To further enhance the performance of the baseline detection systems, acoustic representations learned using MFCC features with a BiLSTM network were used along with the prosody features to provide complementary information to the system. Detection accuracy of 88.75% was obtained for all disfluencies using the proposed system in the case of the UCLASS dataset, whereas for the IIITH-IED dataset, the detection accuracy obtained was 91.24%.

Future works will be aimed at exploring different methods for automatic syllabification from speech signals to improve the disfluency detection pipeline further. We also plan to explore better prosody representations for each disfluency type and perform model-level adaptations to enhance the system performance.

VI. ACKNOWLEDGEMENT

The authors would like to thank Ministry of Electronics and Information Technology (MeitY), Government of India for their support by funding this research.

REFERENCES

[1] Judit Kormos and Mariann Dénes, "Exploring measures and perceptions of fluency in the speech of second language learners," *System*, vol. 32, no. 2, pp. 145–164, 2004.

[2] Robin J Lickley, "Detecting Disfluency in spontaneous speech", *Ph.D. thesis*, University of Edinburgh, 1994.

[3] Ooi Chia Ai, M Hariharan, Sazali Yaacob, and Lim SinChee, "Classification of speech dysfluencies with mfcc and lpcc features," *Expert Systems with Applications*, vol. 39, no. 2, pp. 2157–2165, 2012.

[4] J Scott Yaruss, "Clinical measurement of stuttering behaviors," *Contemporary Issues in Communication Science and Disorders*, vol.24, no. Spring, pp. 27–38,1997.

[5] Sharon Goldwater, Dan Jurafsky and Christopher D Manning, "Which words are hard to recognize? Prosodic, lexical and disfluency factors that increase speech recognition error rates," *Speech Communication*, vol. 52, no. 3, pp. 181–200, 2010.

[6] Qianqian Dong, Feng Wang, Zhen Yang, Wei Chen, Shuang Xu and Bo Xu, "Adapting translation models for transcript disfluency detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019,vol. 33, pp. 6351–6358.

[7] Shaolei Wang, Wangxiang Che, Qi Liu, Pengda Qin, Ting Liu and William Yang Wang, "Multi-task self-supervised learning for disfluency detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, vol. 34, pp. 9193–9200

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

[9] Vasilisa Verkhodanova, Vladimir Shapranov and Irina Kipyatkova, "Hesitations in spontaneous speech: acoustic analysis and detection," in *International Conference on Speech and Computer*. Springer, 2017, pp. 398–406.

[10] Tomohiro Tanaka, Ryo Masumura, Takafumi Moriya, Takanobu Oba and Yushi Aono, "Disfluency detection based on speech-aware token-by-token sequence labeling with blstm-crfs and attention mechanisms," in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2019, pp. 1009–1013.

[11] Sheena Christabel Pravin and M Palanivelan, "A hybrid deep ensemble for speech disfluency classification," *Circuits, Systems, and Signal Processing*, pp. 1–28, 2021.

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

[13] Elizabeth Shriberg, Rebecca Bates and Andreas Stolcke, "A prosody only decision-tree model for disfluency detection," in *Fifth European Conference on Speech Communication and Technology*, 1997.

[14] Chung-Hsien Wu and Gwo-Lang Yan, "Acoustic feature analysis and discriminative modeling of filled pauses for spontaneous speech recognition," in *Real World Speech Processing*, pp. 17–30. Springer, 2004.

[15] Rachid Riad, Anne-Catherine Bachoud-Lévi, Frank Rudzicz and Emmanuel Dupoux, "Identification of primary and collateral tracks in stuttered speech," 2020, unpublished.

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

[17] S Gupta, RS Shukla and RK Shukla, "Weighted mel frequency cepstral coefficient based feature extraction for automatic assessment of stuttered speech using bidirectional lstm," *Indian Journal of Science and Technology*, vol. 14, no. 5, pp. 457–472, 2021.

[18] Tedd Kourkounakis, Amirhossein Hajavi and Ali Etemad, "Detecting multiple speech disfluencies using a deep residual network with bidirectional long short-term memory," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6089–6093.

[19] G Diwakar and Veena Karjigi, "Improving speech to text alignment based on repetition detection for dysarthric speech," *Circuits, Systems, and Signal Processing*, vol. 39, no. 11, pp. 5543–5567, 2020.

[20] Nirmal Sugathan and Santosh Maruthy, "Nonword repetition and identification skills in kannada speaking school-aged children who do and do not stutter," *Journal of fluency disorders*, vol. 63, pp. 105745, 2020.

[21] Pierre Perruchet and Barbara Tillmann, "Exploiting multiple sources of information in learning an artificial language: Human data and modeling," *Cognitive Science*, vol. 34, no. 2, pp. 255–285, 2010.

[22] Stephan Meylan, Chigusa Kurumuda, Mike Frank, Ben-jamin Borschinger and Mark Johnson, "Modeling online word segmentation performance in structured artificial languages," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, 2012, vol. 34.

[23] Okko Räsänen, Gabriel Doyle and Michael C Frank, "Pre-linguistic segmentation of speech into syllable-like units," *Cognition*, vol. 171, pp. 130–150, 2018.

[24] Christoph Daube, Robin AA Ince and Joachim Gross, "Simple acoustic features can explain phoneme-based predictions of cortical responses to speech," *Current Biology*, vol. 29, no. 12, pp. 1924–1937, 2019.

[25] Yannick Marchand, Connie R Adsett and Robert I Damper, "Automatic syllabification in english: A comparison of different algorithms," *Language and speech*,vol. 52, no. 1, pp. 1–27, 2009.

[26] Leena Mary, Anil P Antony, Ben P Babu and SR Mahadeva Prasanna, "Automatic syllabification of speech signal using short time energy and vowel onset points," *International Journal of Speech Technology*, vol. 21, no.3, pp. 571–579, 2018.

[27] Carmen Jany, Mathew Gordon, Carlos M Nash and Nobutaka Takara,

"How universal is the sonority hierarchy?: a cross-linguistic acoustic study," in *Proceedings of the ICPhS*. Citeseer, 2007, pp. 1401–1404.

[28] Nicolas Obin, François Lamare and Axel Roebel, "Syll-o-matic: An adaptive time-frequency representation for the automatic segmentation of speech into syllables," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp.6699–6703.

[29] Victoria Leong and Usha Goswami, "Acoustic-emergent phonology in the amplitude envelope of child-directed speech," *PloS one*, vol. 10, no. 12, pp. e0144411, 2015.

[30] Stephen George Parker, "Quantifying the sonority hierarchy", *Ph.D. thesis*, University of Massachusetts at Amherst, 2002

[31] John Harris, "The phonology of being understood: Further arguments against sonority," *Lingua*, vol. 116, no.10, pp. 1483–1494, 2006

[32] Roy D Patterson, KEN Robinson, John Holdsworth, Denis McKeown, C Zhang and Michael Allerhand, "Complex sounds and auditory images," in *Auditory physiology and perception*, pp. 429–446. Elsevier, 1992.

[33] 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.

[34] F. S. Juste and C. R. Furquim de Andrade, "Speech disfluency types of fluent and stuttering individuals: Age effects," *International Journal of Phoniatrics, Speech Therapy and Communication Pathology*, vol. 63, 2011.

[35] Sparsh Garg, Utkarsh Mehrotra, Gurugubelli Krishna and Anil Kumar Vuppala, "Towards a Database For Detection of Multiple Speech Disfluencies in Indian English", *NCC 2021*, in press.

[36] James Ferguson, Greg Durrett and Dan Klein, "Disfluency detection with a semi-markov model and prosodic features," in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2015, pp. 257–262.

[37] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall,and W Philip Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002

[38] P Mahesha and DS Vinod, "Automatic segmentation and classification of dysfluencies in stuttering speech," in *Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies*, 2016, pp. 1–6.

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