Towards a Database For Detection of Multiple Speech Disfluencies in Indian English

by

Sparsh Garg, Utkarsh Mehrotra, Gurugubelli Krishna, Anil Kumar Vuppala

Report No: IIIT/TR/2021/-1

Centre for Language Technologies Research Centre
International Institute of Information Technology
Hyderabad - 500 032, INDIA
July 2021

Towards a Database For Detection of Multiple Speech Disfluencies in Indian English

Sparsh Garg*
Speech Processing Lab
LTRC, KCIS, IIIT Hyderabad
Hyderabad, India
sparsh.garg@research.iiit.ac.in

Utkarsh Mehrotra*
Speech Processing Lab
LTRC, KCIS, IIIT Hyderabad
Hyderabad, India
utkarsh.mehrotra@research.iiit.ac.in

Gurugubelli Krishna
Speech Processing Lab
LTRC, KCIS, IIIT Hyderabad
Hyderabad, India
gurugubelli.krishna@research.iiit.ac.in

Anil Kumar Vuppala
Speech Processing Lab
LTRC, KCIS, IIIT Hyderabad
Hyderabad, India
anil.vuppala@.iiit.ac.in

Abstract*—The detection and removal of disfluencies from speech is an important task since the presence of disfluencies can adversely affect the performance of speech-based applications such as Automatic Speech Recognition (ASR) systems and speech-to-speech translation systems. From the perspective of Indian languages, there is a lack of studies pertaining to speech disfluencies, their types and frequency of occurrence. Also, the resources available to perform such studies in an Indian context are limited. Through this paper, we attempt to address this issue by introducing the IIITH-Indian English Disfluency (IIITH-IED) Dataset. This dataset consists of 10-hours of lecture mode speech in Indian English. Five types of disfluencies - filled pause, prolongation, word repetition, part-word repetition and phrase repetition were identified in the speech signal and annotated in the corresponding transcription to prepare this dataset. The IIITH-IED dataset was then used to develop frame-level automatic disfluency detection systems. Two sets of features were extracted from the speech signal and then used to train classifiers for the task of disfluency detection. Amongst all the systems employed, Random Forest with MFCC features resulted in the highest average accuracy of 89.61% and F1-score of 0.89.

***Index Terms*—speech disfluencies, acoustic features, binary classification, recall**

I. Introduction

Spontaneous speech refers to the scenario when a speaker is expected to speak without undergoing any preparation in advance. The speaker thinks and formulates utterances on the go. Such a setting often leads to abrupt breaks and discontinuity in normal conversation flow because of hesitations in the speaker's speech. These discontinuities in speech occur due to a variety of reasons, for instance - language complexity, nervousness while speaking, taking time to formulate thoughts while speaking, etc. [1], [2]. These abrupt breaks or hesitations in the normal flow of speech are referred to as speech disfluencies. Disfluencies can take up various forms depending on the type of discontinuity and how the speaker overcomes or corrects that discontinuity. Speech disfluencies include - filled pause, prolongation, repetitions and revisions. TABLE I discusses the most common types of disfluencies observed in spontaneous speech, along with their examples.

In recent years, the development of robust speech-based applications, which can be used in multiple settings, has been the focus of a number of studies [3], [4]. The presence of disfluencies in speech can adversely affect the performance of many such applications. For ASR systems, disfluencies lead to higher word error rates (WER) since most ASRs are developed for read and non-spontaneous speech. In the case of a system like speech-to-speech translation, this error is further propagated to downstream applications (machine translation and text to speech systems), increasing the total error of the pipeline by many folds [5]. Hence, removing disfluencies from speech signal before giving it as input to these applications becomes crucial for getting appreciable performance.

On exploring the existing literature on speech disfluencies and their detection, it was found that most works were performed for British and American English [6]–[8]. India has one of the largest English speaking populations globally, with around 83 million people using it as their second language [9]. Hence, the study of disfluencies from the perspective of Indian English becomes imperative. However, it was observed that existing studies had not paid as much attention to disfluencies in Indian English as the language deserves. Also, there is a lack of resources available to perform such a study.

To address this issue of lack of freely available data for disfluency-based studies in Indian English, we introduce the IIITH-IED dataset. Five major types of speech disfluencies were identified and annotated in 10-hours of lecture mode speech in Indian English for the preparation of this dataset. Since the lecturer prepares some key points of his/her lecture in advance, but there are instances where the lecturer has to explain a topic spontaneously, we characterize this type of speech as semi-spontaneous. The prepared dataset was then used here to develop automatic disfluency detection systems.

* Equal Contribution

978-1-6654-4177-3/21/$31.00 ©2021 IEEE

TABLE I
DESCRIPTION OF DIFFERENT TYPES OF DISFLUENCIES.

Disfluency Type	**Description**
Filled Pause	Pauses in speech with filler words. eg.- uhh, umm
Prolongations	Lengthening of a particular sound or syllable. eg.- whoooose book it is
Word Repetition	Repetition of complete word. eg.- small small
Part-word Repetition	Repetition of particular phoneme. eg.- th-this
Phrase Repetition	Repetition of more than one words or phrase. eg.- I am I am
Revisions	Amending the original utterance by using the similarly structured phrase. eg.- I went to London uhh I went to Sydney

The systems developed here were frame-level disfluency detection systems, which predict whether a particular speech frame belongs to a specific disfluency type or not. The main contributions of this work are as follows -

- Introduction of the IIITH-IED dataset in order to facilitate studies on speech disfluencies in Indian English.
- Developing automatic disfluency detection systems using the IIITH-IED dataset to present results of the disfluency detection systems on our dataset.

The rest of this paper is organised as follows - in Section II we discuss prior works done in the field of disfluency detection and the resources available for speech disfluency studies. Details about the annotation and analysis of the IIITH-IED dataset are presented in Section III. Section IV describes the disfluency detection systems developed on this dataset: the features used for disfluency detection, the classifiers and hyperparameters used for training and the results obtained for every model. We conclude by summarizing our work and discussing it's future scope in Section V.

II. Related Works

Various datasets have been created for studying disfluencies in different speech settings. The Switchboard speech corpora [10] has been used for studying disfluencies in spontaneous conversational speech. It contains 2400 two-sided telephonic conversations in American English, annotated with disfluency events in the transcripts. In [11], the UCLASS dataset was introduced. This dataset is in British English and deals with disfluencies present in stuttered speech. Speech disfluencies have been the subject of research in languages other than English as well. In [12], read speech data from children was collected in Portuguese, and 9 disfluency events were annotated. The HESITA speech corpus was introduced in [13]. Disfluency events like filled pause, repetitions, substitutions, vocalic extensions and truncated words were identified and annotated in speech recordings from 30 daily news programs in European Portuguese. In [14], segment prolongations were studied in the Hebrew language from spontaneous speech recordings collected from 36 speakers and amounting to 97 minutes. Phonetic and structural properties of segment prolongations in German were studied in [15] using data from 18 speakers, amounting to 4 hours. Automatic disfluency detection has been the focus of many works [16]–[19]. In general, most of the speech disfluency detection methods belong to one of the following categories - as a post-processing step after the ASR output, which use text-based features along with speech for disfluency detection [20], [21] or as a pre-processing step before the ASR. Such methods use only signal-level features to detect disfluencies [19], [22]. In [23], log-energy Mel scale filters and pitch based features were shown to perform well in detecting disfluencies at frame level using SVM and DNN classifiers. Formant information and nasality effect were used as cues in [24], [25] for automatic detection of individual disfluencies. [26] addressed the issue of detecting disfluencies at the utterance level using speech features. Disfluencies were detected in a 4-second speech file using spectrogram as an input feature to Deep Residual network with Bidirectional LSTM (BiLSTM). Sequence tagging using lexical features is an effective approach to detect disfluencies and has been used in [20], [27]. In [28], disfluency detection based on lexical features was done using a neural machine translation model. In [29], disfluency detection was performed using a noisy training approach, BiLSTM and self-attention model and outperformed the state of the art BERT model [30].

In the context of Indian English, little work has been done for analyzing disfluencies. In [24], a formant-based thresholding system was discussed, where the first 2 formants and duration information was used to detect filled pause in 96 minutes speech in Indian English. A similar method for filled pause and repetition detection was described in [25]. Here, the stability of the first four formants was used to identify disfluencies in the speech signal. A small dataset of 60 English sentences was considered for this work.

III. Database

We discuss the details about the IIITH-IED Dataset here. This dataset deals with speech disfluencies in Indian English. India has a large English speaking population, with English being used as the primary teaching medium in most higher education institutions. The lectures delivered in these institutions are an excellent source for studying the characteristics and frequency of occurence of disfluencies since the lecturer is expected to explain a particular topic, sometimes on the go, during these lectures, which leads to the occurrence of disfluencies along with normal speech. So, to prepare this dataset, freely available lectures under the government of India's NPTEL initiative were used. The lectures used for the preparation of this dataset belonged to the following domains - Computer Science, Artificial Intelligence, Electronics and Communication, and Electrical engineering. Speech recordings from 60 speakers were used in the preparation of the IIITH-IED dataset. Out of the 60 speakers, 30 were male and 30 were female, to minimize any gender-based imbalance that might be present. For every speaker, a 10-minute audio recording was taken from a lecture to capture the variability in the speaking style and disfluencies that might be present in the speech. This 10-minute audio recording was further split into smaller audio files of length 8 to 12 seconds, with a sampling rate of 16000 Hz. The smaller audio files were segmented by ensuring that the segmentation does not lead to the chopping of words and phrases. Two annotators then listened to these audio files and marked the positions of disfluencies in the corresponding transcripts. After marking at the transcript level, the annotation was performed on the speech signal, to identify the starting time and ending time of each occurrence of disfluency. Audacity open-source toolkit was used to perform this signal level annotation and generate the corresponding labels for each audio file. This timing information in the generated labels was used later to develop frame-level automatic disfluency detection systems on the IIITH-IED dataset. Fig. 1 show an example of the annotation performed here.

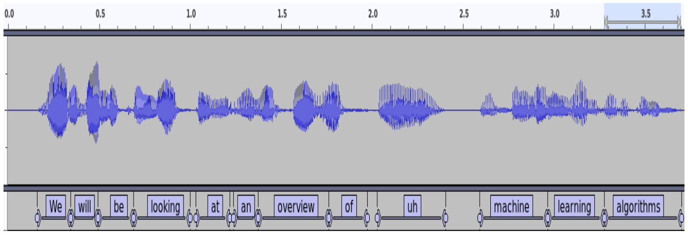

Fig. 1. Signal-level Transcription of the following sentence - We will be looking at an overview of uh machine learning algorithms.

The annotation format used here is similar to [31]. In order to maintain consistency, all the audio files of one particular speaker were annotated at both the word level and the signal level by the same annotator. The other annotator then verified this annotation and the final timestamps for each disfluency occurrence were obtained. For the IIITH-IED dataset, five different types of speech disfluencies were considered. They were - filled pause, prolongations, part-word repetitions, word repetitions and phrase repetitions. The number of occurrences and the average duration of each type of disfluency present in the IIITH-IED dataset is mentioned in TABLE II.

TABLE II
NUMBER OF OCCURRENCES OF EACH DISFLUENCY TYPE IN IIITH-IED DATASET.

Disfluency Type	**# of Occurences**	**Avg. Duration**
Filled Pause	1428	0.395 sec
Prolongation	71	0.553 sec
Part-word Repetition	164	0.954 sec
Word Repetition	211	0.890 sec
Phrase Repetition	76	1.365 sec

As can be seen, filled pause was the most common type of disfluency occurring in the dataset. It was observed that, like British English, two forms of filled pause, 'um' and 'uh', were the most common [32]. Out of these two, the number of occurrences of 'uh' (1265) were greater than that of 'um' (163). There was a gender-based difference here, with the number of 'um' filled pause produced by female speakers (106) being much greater than what their male counterparts produced (57). In the case of part-word repetitions, most instances had word-initial repetition in them; that is, the repetition took place at the initial syllable position of the words (for example: w-what). Word repetitions were the second most common type of disfluency in the IIITH-IED dataset. Most instances of word repetitions were for commonly used words like - and, of, the, for, etc. Phrase repetitions and prolongations were the rarest disfluency types, having 76 and 71 occurrences, respectively. The instances of prolongations observed in the dataset were caused due to vowel lengthening. This lengthening occurred either at the word-initial position (for example: lengthening of /o/ in of) or at the word-final position (for example: lengthening of /o/ in to). Middle-word lengthenings leading to prolongation disfluencies were very rare in the dataset.

In total, the IIITH-IED dataset consists of 3386 wavfiles of duration 8-12 seconds. Each wavfile has zero, one, two or more disfluencies in it. TABLE III presents the statistics about the number of audio files having a certain number of disfluencies in them.

TABLE III
NUMBER OF WAVFILES CORRESPONDING TO THE NUMBER OF DISFLUENCIES IN IIITH-IED DATASET

# of Disfluencies	**Number of wavfiles**
Zero	2013
One	934
Two	326
Three	91
More than three	22

IV. EXPERIMENTS AND RESULTS

In order to develop disfluency detection systems in Indian English, the IIITH-IED dataset was used. The systems developed here are frame-level automatic disfluency detection systems. The models used rely only on acoustic features, i.e., the information extracted from the speech signal for disfluency detection. The aim of developing such systems was to detect whether a particular type of disfluency was present in a 10 ms frame of speech or not, as done in [23] and [33]. Experiments were performed here for the 5 types of disfluencies making up the IIITH-IED dataset - filled pause, part-word repetitions, word-repetitions, phrase repetitions and prolongation. The detection of every type of disfluency was set up as a binary classification task - the speech frame either belong to that disfluency type or does not. Another set of experiments were performed to decide whether a particular speech frame is disfluent or not. In this experiment, speech frames belonging to all disfluency types were considered as one class while normal speech frames having no disfluency were considered as the other class. Binary classification was then performed to decide whether a speech frame is fluent or disfluent.

A. Feature Extraction

Two different sets of acoustic features were used for the task of disfluency detection directly from the speech signal. In [23], log Mel-filterbank features were used for disfluency detection and produced high levels of accuracy. Hence, the first set of features used here was a combination of log Mel-filterbank features and the fundamental frequency calculated for each frame. 40 log Mel-filterbank features were extracted by taking a 25 ms frame of the speech signal with a 10 ms shift. The filterbank features were then mean-variance normalized for every audio file, using the VAD information. Besides this, the signal's fundamental frequency in that frame was calculated and used along with the filterbank features, giving a feature vector having 41-dimensions per frame. This set of features will be referred to as the Filterbank features from here on.

The next set of features used here were the MFCC input features. The MFCC features were made up of 13 cepstral coefficients per frame, the 0th cepstral coefficient and the energy of the frame. The delta and delta-delta coefficients were also computed for the MFCC features, giving a 45-dimensional feature vector per frame. Windowed frames of length 25 ms and frame shift of 10 ms were used for computing the MFCC features.

In [33], stacking up features from frames using a context window was shown to produce better detection results. So, for both sets of features here, window lengths of ±1, ±2 and ±3 frames were experimented with.

It was observed that stacking up features from 3 frames before and after every particular frame gave the best classification results. Hence, this configuration was used for reporting the final disfluency detection results.

B. Detection Models

Using the features extracted above, 3 systems were trained for the task of disfluency detection. First, SVM with a linear kernel and penalty term of 0.25 was used for the binary classification task. Next, the Random Forest based ensemble model was taken, with the maximum number of trees being 100 and the maximum depth of every tree being 20 in the ensemble. Finally, a DNN classifier with 2 hidden layers was considered for disfluency detection. The number of hidden units in each layer were 100 and 50, respectively. The optimizer used was Adam optimizer. Hyperparameter tuning was performed as well for training the DNN. An optimal learning rate of 0.001 and an optimal batch size of 32 was used here. A train-test split of 80:20 was used for the experiments here.

An important point considered was that disfluency datasets tend to be very imbalanced since the number of frames corresponding to disfluencies are much less than fluent speech frames. So, in order to ensure that the disfluency detection models are not biased, random undersampling [34] and SMOTE (Synthetic Minority Oversampling Technique) based oversampling [35] techniques were used while training each model. For random undersampling, a large number of frames belonging to the majority class were removed randomly so that the number of frames of both the classes is equal. On the other hand, using SMOTE, the number of samples of the minority class were made equal to that of the majority class by generating new samples of the minority class using k-nearest neighbour approach.

C. Results

The first set of results presented here are for the detection of a single disfluency class. TABLE IV shows the results obtained for filled pause, prolongations, part-word repetition, word repetition and phrase repetition detection using Filterbank features. The metrics chosen for the evaluation of classification models here are - accuracy, recall and F1-score. For filled pause, an F1-score of 0.938 was obtained with the Random Forest classifier in the oversampling condition. For prolongation detection, a high recall (0.955) and high F1-score (0.951) was obtained. There was an improvement in the detection accuracy using all three classifiers in the oversampling approach compared to undersampling, which shows that further improvements can be obtained by using the same experimental setup with more extensive datasets so that the variance in samples of each disfluency can be captured effectively. As compared to filled pause and prolongation, the accuracy and F1-score obtained for repetition type disfluencies was lesser. This can be attributed to the longer average duration of these types of disfluencies, which provides scope for greater variance in the samples of these classes. Thus, the best results for part-word and word repetitions are obtained using oversampling, with the best F1-scores obtained being 0.835 for part-word repetition and 0.852 for word repetition. In case of phrase repetitions, a lower recall is obtained in all the experiments. This can be because the number of samples of phrase repetition in the dataset are not enough to efficiently model this class. The highest F1-score obtained for phrase repetition detection was 0.767.

Detection results obtained using MFCC features for individual disfluencies are presented in TABLE V. Compared to Filterbank features, MFCC features give better outcomes for all the five disfluencies considered here. With DNN as a classifier, a definite improvement of 3.76% was observed

TABLE IV
RESULTS OBTAINED FOR INDIVIDUAL DISFLUENCY DETECTION USING FILTERBANK FEATURES. HERE ACC. SHOWS ACCURACY, REC. SHOWS RECALL AND F1 SHOWS F1-SCORE

Models	**Filled Pause**			**Prolongations**			**Part-Word Repetitions**			**Word Repetitions**			**Phrase Repetitions**		
	Acc.	**Rec.**	**F1**	**Acc.**	**Rec.**	**F1**	**Acc.**	**Rec.**	**F1**	**Acc.**	**Rec.**	**F1**	**Acc.**	**Rec.**	**F1**
SVM (undersampled)	87.12	0.842	0.866	82.77	0.828	0.821	68.23	0.621	0.638	60.94	0.547	0.585	62.95	0.563	0.615
RF (undersampled)	89.28	0.887	0.890	95.21	0.942	0.950	81.22	0.803	0.812	82.01	0.824	0.815	78.90	0.761	0.758
DNN (undersampled)	85.83	0.866	0.859	74.65	0.693	0.712	66.43	0.610	0.632	57.88	0.556	0.535	60.85	0.531	0.594
SVM (SMOTE)	88.91	0.902	0.892	86.24	0.848	0.865	74.14	0.773	0.797	70.12	0.686	0.690	64.58	0.580	0.626
RF (SMOTE)	91.32	0.944	0.938	95.87	0.955	0.951	84.64	0.812	0.835	86.33	0.881	0.852	79.78	0.773	0.767
DNN (SMOTE)	86.18	0.901	0.878	81.30	0.799	0.824	79.71	0.815	0.833	65.82	0.704	0.728	64.79	0.664	0.672

TABLE V
RESULTS OBTAINED FOR INDIVIDUAL DISFLUENCY DETECTION USING MFCC FEATURES. HERE ACC. SHOWS ACCURACY, REC. SHOWS THE RECALL AND F1 SHOWS THE F1-SCORE

Models	**Filled Pause**			**Prolongations**			**Part-Word Repetitions**			**Word Repetitions**			**Phrase Repetitions**		
	Acc.	**Rec.**	**F1**	**Acc.**	**Rec.**	**F1**	**Acc.**	**Rec.**	**F1**	**Acc.**	**Rec.**	**F1**	**Acc.**	**Rec.**	**F1**
SVM (undersampled)	89.91	0.882	0.895	84.22	0.856	0.845	71.23	0.709	0.698	72.44	0.702	0.713	66.12	0.654	0.662
RF (undersampled)	89.14	0.892	0.889	94.64	0.932	0.945	82.86	0.838	0.826	82.53	0.824	0.821	79.03	0.773	0.785
DNN (undersampled)	88.97	0.831	0.828	80.95	0.812	0.809	77.41	0.757	0.775	75.13	0.763	0.748	71.33	0.662	0.700
SVM (SMOTE)	90.12	0.887	0.892	86.24	0.864	0.851	73.41	0.713	0.728	73.35	0.732	0.724	69.63	0.681	0.702
RF (SMOTE)	93.84	0.936	0.937	98.21	0.978	0.979	92.95	0.900	0.927	93.32	0.913	0.931	81.02	0.791	0.814
DNN (SMOTE)	89.94	0.894	0.898	88.35	0.891	0.883	82.33	0.809	0.819	80.86	0.799	0.803	73.52	0.722	0.736

for filled pause detection in the oversampling condition using MFCC compared to Filterbank features. The best F1-score obtained for filled pause and prolongation using MFCC features was 0.937 and 0.978, respectively. The accuracy obtained for repetition type disfluencies was higher as well in this case. The best detection results are obtained using the Random Forest classifier. For word repetitions, a high recall value of 0.913 and high accuracy of 93.3% was obtained in the oversampling condition. In the case of part-word repetitions, high variability in the samples of this disfluency leads to lower classification accuracy in the undersampled condition. Further, the oversampling scenario improved the results in terms of accuracy and F1-score of 92.9% and 0.927, respectively. For phrase repetitions, using MFCC features with DNN gave an absolute improvement of 8.73% in accuracy. The highest F1-score for phrase repetition was obtained using the Random Forest classifier and was 0.814.

TABLE VI
RESULTS OBTAINED CONSIDERING ALL DISFLUENCIES AS SINGLE CLASS

Models	**Fiterbank + F0**			**MFCC**		
	Acc.	**Rec.**	**F1**	**Acc.**	**Rec.**	**F1**
SVM (undersampled)	73.28	0.755	0.743	80.34	0.805	0.812
RF (undersampled)	85.26	0.825	0.847	86.69	0.848	0.863
DNN (undersampled)	72.86	0.768	0.738	82.25	0.814	0.820
SVM (SMOTE)	73.57	0.753	0.757	80.94	0.810	0.816
RF (SMOTE)	88.69	0.904	0.883	89.61	0.862	0.891
DNN (SMOTE)	73.11	0.787	0.770	83.26	0.811	0.828

TABLE VI shows the results obtained when all the disfluencies were considered a single class, and binary classification was performed to determine if a speech frame was disfluent or not. In this case, as well, MFCC features outperformed Filterbank features in terms of F1-Score, accuracy and recall. In the oversampling condition, an absolute difference of 10.15% was observed in the accuracy of the DNN classifier using MFCC features compared to Filterbank features. The best results are obtained with Random Forest classifier using Filterbank and MFCC features in terms of F1-scores of 0.883 and 0.891, respectively in oversampling condition. In addition, high recall values are also obtained using these features with DNN and Random Forest classifier, showing the ability of these models to minimize the number of false alarms obtained in classification.

V. CONCLUSION

In this work, the issue of lack of resources for the study of speech disfluencies in Indian English was addressed. The IIITH-IED dataset was introduced in this paper. 10-hours of lecture mode speech in Indian English was collected and annotated for five main types of disfluencies occurring in spontaneous speech. The disfluencies were identified, and their start and end times in the speech signal were noted for the preparation of this dataset. Analysis related to the frequency of occurrence of each disfluency, their type and duration in the dataset was performed. The collected data was then used to develop frame-level automatic disfluency detection systems. Two sets of features - Filterbank and MFCC features were used here for developing the disfluency detection systems with 3 different types of classifiers. High F1-scores of 0.883 and 0.891 were obtained for disfluency detection using Filterbank and MFCC features respectively, with the Random Forest based ensemble classifier.

Future works in this domain will be aimed at developing utterance level disfluency detection systems in Indian English. Also, model-level improvements will be incorporated in the disfluency detection systems to enhance their performance. Finally, these experiments would also be extended to study disfluencies in other Indian languages, analyse their forms and frequencies and carrying out feature analysis to develop efficient automatic disfluency detection systems in the Indian scenario.

References

[1] E. Shriberg, “To’errrr’is human: ecology and acoustics of speech disfluencies,” *Journal of the International Phonetic Association*, pp. 153–169, 2001.

[2] M. Corley and O. W. Stewart, “Hesitation disfluencies in spontaneous speech: The meaning of um,” *Language and Linguistics Compass*, vol. 2, no. 4, pp. 589–602, 2008.

[3] C. Donahue, B. Li, and R. Prabhavalkar, “Exploring speech enhancement with generative adversarial networks for robust speech recognition,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5024–5028.

[4] D. Cai, W. Cai, and M. Li, “Within-sample variability-invariant loss for robust speaker recognition under noisy environments,” in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6469–6473.

[5] S. R. Maskey, Y. Gao, and B. Zhou, “Disfluency detection for a speech-to-speech translation system using phrase-level machine translation with weighted finite state transducers,” Dec. 28 2010, uS Patent 7,860,719.

[6] X. Qian and Y. Liu, “Disfluency detection using multi-step stacked learning,” in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2013, pp. 820–825.

[7] V. Zayats and M. Ostendorf, “Robust cross-domain disfluency detection with pattern match networks,” *arXiv preprint arXiv:1811.07236*, 2018.

[8] S. Alharbi, M. Hasan, A. J. Simons, S. Brumfitt, and P. Green, “A lightly supervised approach to detect stuttering in children’s speech,” in *Proceedings of Interspeech 2018*. ISCA, 2018, pp. 3433–3437.

[9] S. M. Mathews, “Language skills and secondary education in india,” *Economic and Political Weekly*, vol. 53, no. 15, pp. 20–22, 2018.

[10] J. J. Godfrey, E. C. Holliman, and J. McDaniel, “Switchboard: Telephone speech corpus for research and development,” in *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, vol. 1. IEEE Computer Society, 1992, pp. 517–520.

[11] P. Howell, S. Davis, and J. Bartrip, “The university college london archive of stuttered speech (uclass),” 2009.

[12] J. Proença, D. Celorico, S. Candeias, C. Lopes, and F. Perdigão, “Children’s reading aloud performance: A database and automatic detection of disfluencies,” in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[13] S. Candeias, D. Celorico, J. Proença, A. Veiga, and F. Perdigão, “Hesita (tions) in portuguese: a database,” in *Sixth Workshop on Disfluency in Spontaneous Speech*, 2013.

[14] V. Silber-Varod, M. Gósy, and R. Eklund, “Segment prolongation in hebrew,” in *The 9th Workshop on Disfluency in Spontaneous Speech*, 2019, p. 47.

[15] S. Betz, R. Eklund, and P. Wagner, “Prolongation in german,” in *DiSS 2017 The 8th Workshop on Disfluency in Spontaneous Speech, KTH, Royal Institute of Technology, Stockholm, Sweden, 18–19 August 2017*. KTH Royal Institute of Technology, 2017, pp. 13–16.

[16] R. J. Lickley, “Detecting disfluency in spontaneous speech,” Ph.D. dissertation, University of Edinburgh, 1994.

[17] Y. Liu, E. Shriberg, A. Stolcke, D. Hillard, M. Ostendorf, and M. Harper, “Enriching speech recognition with automatic detection of sentence boundaries and disfluencies,” *IEEE Transactions on audio, speech, and language processing*, vol. 14, no. 5, pp. 1526–1540, 2006.

[18] C.-H. Wu and G.-L. Yan, “Acoustic feature analysis and discriminative modeling of filled pauses for spontaneous speech recognition,” in *Real World Speech Processing*. Springer, 2004, pp. 17–30.

[19] E. Salesky, M. Sperber, and A. Waibel, “Fluent translations from disfluent speech in end-to-end speech translation,” *arXiv preprint arXiv:1906.00556*, 2019.

[20] V. Zayats, M. Ostendorf, and H. Hajishirzi, “Disfluency detection using a bidirectional lstm,” *arXiv preprint arXiv:1604.03209*, 2016.

[21] Y. Lu, M. J. Gales, K. Knill, P. Manakul, and Y. Wang, “Disfluency detection for spoken learner english.” in *SLaTE*, 2019, pp. 74–78.

[22] R. Hamzah and N. Jamil, “Investigation of speech disfluencies classification on different threshold selection techniques using energy feature extraction,” *Malaysian Journal of Computing*, vol. 4, no. 1, pp. 178–192, 2019.

[23] R. Riad, A.-C. Bachoud-Lévi, F. Rudzicz, and E. Dupoux, “Identification of primary and collateral tracks in stuttered speech,” *arXiv preprint arXiv:2003.01018*, 2020.

[24] K. Audhkhasi, K. Kandhway, O. D. Deshmukh, and A. Verma, “Formant-based technique for automatic filled-pause detection in spontaneous spoken english,” in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2009, pp. 4857–4860.

[25] M. Kaushik, M. Trinkle, and A. Hashemi-Sakhtsari, “Automatic detection and removal of disfluencies from spontaneous speech,” in *Proceedings of the Australasian International Conference on Speech Science and Technology (SST)*, vol. 70, 2010.

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

[27] J. Ferguson, G. Durrett, and D. 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.

[28] Q. Dong, F. Wang, Z. Yang, W. Chen, S. Xu, and B. Xu, “Adapting translation models for transcript disfluency detection,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 6351–6358.

[29] N. Bach and F. Huang, “Noisy bilstm-based models for disfluency detection.” in *INTERSPEECH*, 2019, pp. 4230–4234.

[30] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.

[31] F. S. Juste and C. R. F. De Andrade, “Speech disfluency types of fluent and stuttering individuals: age effects,” *Folia Phoniatrica et Logopaedica*, vol. 63, no. 2, pp. 57–64, 2011.

[32] G. Tottie, “On the use of uh and um in american english,” *Functions of Language*, vol. 21, no. 1, pp. 6–29, 2014.

[33] S. Oue, R. Marxer, and F. 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.

[34] G. Lemaître, F. Nogueira, and C. K. Aridas, “Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning,” *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 559–563, 2017.

[35] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “Smote: synthetic minority over-sampling technique,” *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.