iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL
FOR RESEARCH
IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 Issue: XII Month of publication: December 2025

DOI: https://doi.org/10.22214/ijraset.2025.76006

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue XII Dec 2025- Available at www.ijraset.com

A Review on Transformer-Based Models for Disfluency Detection in Indian English and Code-Switched Speech

SMD Shafiulla
Assistant Professor, Department of Computer Science and Engineering, Scient Institute of Technology, Ibrahimpatnam, R.R Dist [India]

Abstract: As there is an advancements of multimedia technologies, human computer interfaces are in huge demand which is an predominant area for research. Vocal representations, facial expressions, lip movement are used to extract various types of information. In particular, the detection of disfluencies, i.e., interruptions in the normal flow of speech characterized by pauses, repetitions, and sound prolongations, is of interest not only for improving speech recognition systems but also for potentially identifying emotional aspects in audio. Several studies have aimed to define computational methods to identify and classify disfluencies, as well as appropriate evaluation methods in different languages. However, no studies have compiled the findings in the literature on this topic. This is important for both summarizing the motivations and applications of the research, as well as identifying opportunities that could guide new investigations. Our objective is to provide an analysis of the state of the art, the main limitations, and the challenges in this field. Most of the existing disfluency detection model are trained on American English datasets and peform poorly on Indian English and Code – switched speech. In this paper a comprehensive review is done on reviewing the Transformer based approach for identifying disfluencies in Indian English and Telugu – English Code switched speech.

Keywords: Disfluencies, Speech recognition, Spoken Dialogue, Natural Language Processing.

I. INTRODUCTION

Human computer conversational interfaces are becoming increasingly important and opening up a highly promising area of research. Vocal representations, along with facial, and body expressions, can be used to extract data with great potential for use in decision-making. In the context of vocal representations, human communication involves a complex and wide range of expressions, which vary according to grammatical rules, languages, accents, slang, disfluencies, and other events during speech. Disfluencies are spontaneous speech phenomena in which the flow is interrupted by pauses, repetitions, sound prolongations, and other occurrences. Topics related to disfluencies are studied across different disciplines, including clinical aspects [Williams et al., 2023; Schettino et al., 2023], second language learning [Belz et al., 2023; Li et al., 2022], and speech technology [Teleki et al., 2024; Kouzelis et al., 2023]. It is worth mentioning that the Disfluency in Spontaneous Speech (DiSS)1 conference, an interdisciplinary forum that has occurred since 1999, has provided opportunities to explore the phenomenon of disfluency by putting together researchers from different disciplines to discuss disfluencies and their implications [Lickley, 2015]. This continuous engagement keeps the topic under discussion and highlights its significance in understanding human communication. In this context, different techniques in statistics, mathematics, machine learning, and syntactic methods have been used to explore the identification and classification of disfluencies while considering different categories and objectives.

Previous reviews have considered disfluency analyses in specific areas, mainly in the context of stuttering. For example, in Barrett et al. [2022] a systematic literature review on machine learning approaches to detect stuttering showed that speech recognition combined with machine learning can be applied to the speech evaluation of people who stutter, where it provided reliable indicators on the presence and severity of stuttering. In Khara et al. [2018] a survey of techniques for extracting and classifying stuttering recognition features is presented, highlighting the growing importance of Automatic Speech Recognition (ASR) systems. Despite the contributions of these studies, they focus specifically on stuttering-related disfluencies. This is the key novelty of our article, which offers a comprehensive analysis of the state of-the-art, limitations, and challenges in disfluency detection across various applications. Disfluency detection and correction is becomes essential for Voice assistants, Meeting transcriptions, Conversational AI, Education technology, Medical and legal dictation systems.

The main contributions of this paper is

1) Classification of disfluencies
2) Challenges faced by Indian English & Code Switched Speech
3) Review on Transformer based approaches
4) Review on multimodal fusion architectures

The content of this article is organized into the following sections:Section 2 presents the classification of disfluencies.Section 3 presents the challenges faced by the Code switched speech.Section 4 presents the review of the proposed models for the disfluency detection and correction.

II. CLASSIFICATION OF DISFLUENCIES

Disfluency detection is a crucial component for robust Automatic Speech Recognition (ASR) system. The composition of disfluencies is independent of language. However, some factors interfere with this classification, such as word position, the presence of other disfluencies in the same sentence, sentence length, and even a combination of these factors. As such, the terminology that is frequently referenced in the publications was proposed by Shriberg [1994]. This structure is composed of four elements: “Reparandum (RM)”, “Interruption point (IP)”, “Interregnum (IM)”, and “Repair (RR)”, as shown in Figure 6. The “Reparandum” is the part of the utterance discarded or corrected by the next words. “Interruption point” is the instant at which the speaker interrupts the original utterance. “Interregnum” is the part used as a moment for the speaker to re-plan (without necessarily implying speech editing) and, finally, “Repair” is the part of the utterance that corresponds to the content of the “Reparandum”, whether it was able to correct it or not.

Overall, 10 disfluency categories were identified in the included articles. Table 1 summarizes the aggregated results according to category and its applicable terms, while the comprehensive distribution of disfluencies across the approved studies can be found in appendix (Table 5). Studies that investigated more than one type of disfluency were counted for all the disfluencies they explored. Some articles did not describe the studied disfluencies in detail, thus leaving interpretation open to the reader. In Table 1, applicable terms refer to terms that were used in a determined position. For example, in the “Interregnum” position, some articles used terms like “Fillers/Filled pauses” and “Interjections”. The most recurrent disfluency class was “Interregnum” (44 articles, 28%), characterized by the presence of “Fillers/Filled pauses”, “Interjections”, “Discourse markers” and “Editing terms”.

Disfluencies range from simple structures such as “Repetitions” that are exact or approximate copies of an utterance.These can be easily detected from a set of defined rules or more complex and arbitrary structures such as “Repairs” in speech that require further and more sophisticated processing. For example, in Miller and Schuler [2008] the authors highlight contributions that the syntactic structure of a sentence and some acoustic signals, such as pauses and prosodic contours, can help in the detection of “Repairs”. “Repairs” were studied in 29 (19%) of the included articles. As pointed out in Miller [2009], this type of disfluency

is a problem for speech recognizers and syntactic analysis applications, since in addition to detecting repairs, such systems need to know which words should be eliminated in order to form a correct grammatical structure. “Stuttering” was the less explored disfluency (2 articles, 1%). Its definition presented in Germesin et al. [2008] is syllables or consonants similar to the beginning of the next fully articulated word. This structure is similar to a “Word Fragment” disfluency that was studied in 5 papers (3%) in this survey. Despite the similarity between “Stuttering” and “Word Fragment”, we chose not to establish a synonym between them since most of the scientific community considers stuttering as a neurobiological disorder rather than a mere isolated occurrence that affects verbal fluency.

III. CHALLENGES FACED BY CODE SWITCHED SPEECH

Code-switched speech is using two or more languages or dialects within a single conversation, utterance, or even a single sentence.Some of the challenges are

1) Extremely accent diversity
2) Heavy code switching
3) Grammar mixing
4) Phonology clash
5) Massive disfluency rate
6) No publicly available dataset with disfluency tags for Indian English and Code switching

IV. REVIEW ON TRANSFORMER BASED APPROACHES

Transformers have become the dominant architecture for speech and language modeling due to their ability to model long-range dependencies through self-attention. Unlike RNNs and LSTMs, transformers process sequences in parallel and capture global context without sequential bottlenecks, making them ideal for disfluency detection where context surrounding a filler, repetition, or repair is crucial. Modern approaches use text-based transformers (BERT, RoBERTa), multilingual transformers (mBERT, MuRIL), speech transformers (Wav2Vec2, HuBERT, Whisper), and multimodal fusion models combining acoustic and linguistic features.

A. BERT for Disfluency Detection

BERT (Bidirectional Encoder Representations from Transformers) is widely used for tagging text with disfluency labels (BIO tagging). Bidirectional context captures both past and future words around disfluencies. Robust embedding space makes filler words separable. Excellent at spotting syntactic and semantic irregularities. BERT models are used in speech disfluency detection by treating it as a sequence-labeling problem to identify parts of a transcript that are disfluent, such as "um," "uh," or repeated words. They can be fine-tuned for this task, and some research focuses on creating smaller, faster BERT models to improve real-time performance on devices. These models use their attention mechanism to understand the relationships between words, allowing them to detect disfluencies even without being explicitly trained to do so, though fine-tuning significantly increases accuracy.

B. RoBERTa

RoBERTa is Robustly Optimized BERT Approach.This approach is an improved version of BERT with dynamic masking feature providing better contextual understanding, detecting hesitations and repetitions.RoBERTa is state-of-art natural language processing (NLP) model which is used in removal of Next Sentence Prediction(NSP),Dynamic Masking, Byte level Byte pair coding. These are widely used in Sentiment analysis , Question answering, Text classification tasks.

V. CONCLUSION

No single transformer handles both Indian English and disfluencies perfectly — multimodal models are needed. Whisper + MuRIL combination is currently the strongest baseline for Hinglish and Indian English disfluency detection. Fine-tuning on prosodic cues is essential, since Indian English disfluencies are heavily prosody-driven. Code-switch boundaries trigger more hesitations, which transformers handle well if trained on bilingual data.Future directions will be on multimodal transformer models which combine both acoustic and prosodic features of the speech.

REFERENCES

[1] ACDC (2024). Automated cardiac diagnosis challenge.

[2] Avanzi, M. (2024). A corpus-based approach to french regional prosodic variation. Cahiers de linguistique française, (31):309–323. DOI: 10.1093/oxfordhb/9780198865131.013.20.

[3] Bach, N. and Huang, F. (2019). Noisy BiLSTM-based models for disfluency detection. In Proc. Interspeech 2019, pages 4230-4234.DOI:10.21437/Interspeech.2019-1336.

[4] Barrett, L., Hu, J., and Howell, P. (2022). Systematic review of machine learning approaches for detecting developmental stuttering. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 30:1160–1172. DOI: 10.1109/TASLP.2022.3155295.

[5] Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., and Blackburn, J. (2020). The pushshift reddit dataset. Proceedings of the International AAAI Conference on Web and Social Media, 14(1):830–839. DOI: 10.1609/icwsm.v14i1.7347.

[6] Belz, M., Müller, M., and Mooshammer, C. (2023).

[7] Bertero, D., Wang, L., Chan, H. Y., and Fung, P. (2015). A comparison between a DNN and a CRF disfluency detection and reconstruction system. In Proc. Interspeech 2015, pages 844–848. DOI: 10.21437/Interspeech.2015-263.

[8] Bui, H. H., Phung, D. Q., and Venkatesh, S. (2004). Hierarchical hidden markov models with general state hierarchy. In Proceedings of the national conference on artificial intelligence, pages 324–329. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press;1999. Available at: https://cdn.aaai.org/AAAI/2004/AAAI04-052.pdf.

[9] Caines, A., Yannakoudakis, H., Allen, H., Pérez-Paredes, P., Byrne, B., and Buttery, P. (2022). The teacher-student chatroom corpus version 2: more lessons, new annotation, automatic detection of sequence shifts. In

[10] Alfter, D., Volodina, E., François, T., Desmet, P., Cornillie, F., Jönsson, A., and Rennes, E., editors, Proceedings of the 11th Workshop on NLP for Computer Assisted Language Learning, pages 23–35, Louvain-la-Neuve, Belgium. LiU Electronic Press. DOI: 10.3384/ecp190003.

[11] Caines, A., Yannakoudakis, H., Edmondson, H., Allen, H.,

[12] Pérez-Paredes, P., Byrne, B., and Buttery, P. (2020). The teacher-student chatroom corpus. In Alfter, D., Volodina, E., Pilan, I., Lange, H., and Borin, L., editors, Proceedings of the 9th Workshop on NLP for Computer Assisted Language Learning, pages 10–20, Gothenburg, Sweden. LiU Electronic Press. DOI: 10.3384/ecp2017510.

[13] Calhoun, S. et al. (2009). NXT switchboard annotations ldc2009t26. DOI: 10.35111/nn2p-v103.

[14] Canavan, A., Graff, D., and Zipperlen, G. (1997). Callhome american english speech ldc97s42. DOI: 10.35111/exq3- x930.

[15] Canavan, A. and Zipperlen, G. (1996). Callfriend american english-non-southern dialect ldc96s46. DOI: 10.35111/d37s-c536.

[16] Carletta, J., Kraaij, W., Ashby, S., Bourban, S., Flynn, M.,Guillemot, M., Hain,

[17] T., Kadlec, J., Karaiskos, V., Kronenthal, M., Lathoud, G., Lincoln, M., Lisowska, A., Post,

[18] W., Reidsma, D., Wellner, P., and McCowan, L. (2005). The AMI meeting corpus. In Proceedings of Symposium on Annotating and Measuring Meeting Behavior. DOI:10.1007/116774823.

[19] Chen, L. and Yoon, S.-Y. (2011). Detecting structural events for assessing non-native speech. In Proceedings of the 6th Workshop on Innovative Use of NLP for Building Educational Applications, IUNLPBEA ’11, page 38–45, USA. Association for Computational Linguistics. DOI: 10.21437/interspeech.2010-282.