Journal of the Brazilian Computer Society, 2025, 31:1, doi: 10.5753/jbcs.2025.4443

Identification and classification of speech disfluencies: A systematic review on methods, databases, tools, evaluation and challenges

Alana S. Luna [**University of São Paulo** | ***alana.luna@usp.br***]
Ariane Machado-Lima [**University of São Paulo** | ***ariane.machado@usp.br***]
Fátima L. S. Nunes [**University of São Paulo** | ***fatima.nunes@usp.br***]

Escola de Artes, Ciências e Humanidades da Universidade de São Paulo, Rua Arlindo Béttio, 1000 - Ermelino Matarazzo, São Paulo, SP, 03828-000, Brazil.

Received: 24 April 2024 • **Accepted:** 18 October 2024 • **Published:** 24 February 2025

Abstract With the advancement of multimedia technologies, human-computer conversational interfaces are becoming increasingly important and are emerging as a highly promising area of research. Vocal representations, facial expressions, and body language can be used to extract various types of information. In the context of vocal representations, the complexity of human communication involves a wide range of expressions that vary according to grammatical rules, languages, accents, slang, disfluencies, and other speech events. 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. Eighty articles were extracted from four databases and analyzed through a systematic review. Our results show that research into the detection of disfluencies has been conducted for various purposes. Some aimed to improve the performance of translation tools, while others focused on the summarization of spoken dialogues, speaker diarization, and Natural Language Processing. Most of the research was oriented toward the English language. F-score, precision, and recall were the most commonly used evaluation measures for the reported methods. Statistical and machine learning techniques were widely applied, with CRFs (Conditional Random Fields), MaxEnt (Maximum Entropy), Decision Trees, and BLSTM (Bidirectional Long Short-Term Memory) being especially prominent. In general, newer approaches, such as BERT and BLSTM, have demonstrated higher performance. However, several challenges remain, opening up new research opportunities.

Keywords: Disfluencies, Speech Recognition, Spoken Dialogue, Natural Language Processing, Rich Transcription

1 Introduction

With the advancement of multimedia technologies, 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 tech-

[1] `https://filledpause.org/diss/`

niques 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. This objective is achieved through a systematic review, in which we analyze 80 articles published between 1996 and 2023. Thus, our main contributions are:

- Presentation of a disfluency categorization while grouping main denominations and their synonyms;
- Compilation of the main databases, transcription tools, and explored languages;
- Analysis of computational techniques used to detect and classify disfluencies, as well as their evaluation methods;
- Indication of the main challenges and research opportunities considering the computer science area.

The content of this article is organized into the following sections: Section 2 presents the systematic review protocol. Section 3 provides a global analysis of selected publications, while Section 4 categorizes the most commonly explored disfluencies. Section 5 presents an analysis of the techniques used, and Section 6 lists the databases and transcription tools employed. Main evaluation measures are discussed in Section 7. The limitations we encountered and the current challenges are presented in Section 8. Finally, the conclusion of this article is presented in Section 9.

2 Research methodology

The systematic review presented in this article consists of four phases (Planning, Conducting, Data Extraction, and Analysis) to answer the following research questions:

- What are the main existing methods and techniques for processing speech disfluencies from audio or text?
- How are the methods evaluated?
- What tools/technologies and databases are used?
- What types of disfluencies are processed?

The summary of each step is shown in Figure 1.

An exploratory analysis was conducted using Google Scholar to identify the main terms used in this study area. We considered the following keywords: Acoustic model; Automatic speech analysis; Automatic speech recognition; Dialog systems; Disfluencies; Disfluency classification; Disfluency detection; Disfluency event; Prosodic information; Prosody labeling; Prosody modeling; and Speech recognition. According to the articles obtained in the exploratory analysis, the scientific databases *IEEE*[2], *ACM*[3], *ACL*[4] and *ISCA's*[5] were used as sources. In the latter database the search was restricted to the main conference (InterSpeech).

The keywords found in this preliminary phase contributed to the following search string: ("disfluenc*" OR "dysfluenc*") AND ("classification" OR "detection" OR "annotation" OR "perception" OR "corpus processing" OR "sequence-tagging"). Complete studies with results that included the search terms in their titles, abstracts, or keywords were selected. We initially found 268 articles. The adopted inclusion (I) and exclusion (E) criteria for selection of studies are presented below:

- (I1) Articles that propose or evaluate techniques for processing disfluencies from audios or texts;
- (E1) Articles that propose or evaluate techniques for processing disfluencies, but that do not use audio or texts as input data;
- (E2) Articles that propose or evaluate techniques for processing audio or texts but that do not consider disfluencies;
- (E3) Articles that evaluate disfluencies outside the computational context;
- (E4) Papers that score below the average of the assigned quality grades (detailed following).

The 117 studies that advanced to the extraction phase were fully analyzed in order to obtain the following data: experiments performed; assessment measures applied; language; dataset; transcription tool, when used; disfluency categorization; technology used to process disfluencies.

We defined the following criteria to evaluate the quality of each study: presentation of transcription tool; presentation or evaluation of a disfluency classification technique; mention of disfluency classification categories; presentation of assessment measures; detailed presentation of the database; and execution of experiments. Each criterion was assigned a score of one when it was fully met, 0.5 when it was partially met, and zero when it was not met. The average of the scores was calculated, and the minimum quality threshold was set as the average score across all studies. Only studies with an average score above this threshold were included. After this step, 80 studies were accepted for analysis.

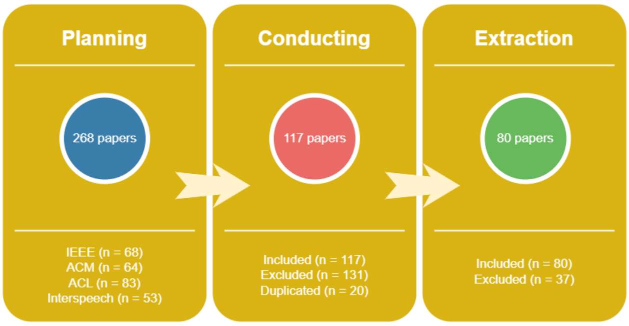

Figure 1. Flow diagram summarizing the steps in developing this systematic review.

3 Overall analysis

The main aspects extracted from the accepted articles in this systematic review are presented in the appendix (Table 5).

[2] `https://ieeexplore.ieee.org/`
[3] `https://dl.acm.org/`
[4] `https://aclanthology.org/`
[5] `https://www.isca-speech.org/`

These publications focused mainly on the detection of disfluencies (58 articles, approximately 73%), followed by removing disfluencies (7 articles, approximately 9%), and prediction (4 articles, approximately 5%), as shown in Figure 2.

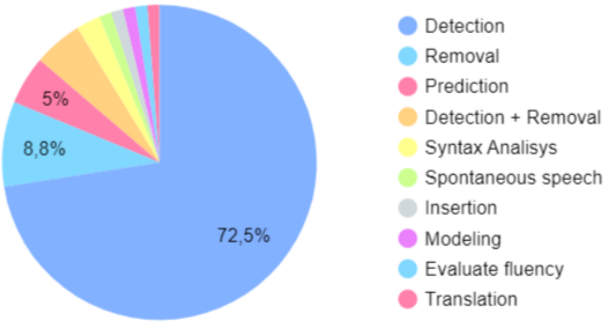

Figure 2. Overall proportion of research objectives.

Disfluencies have been studied since 1996, with some variation in the number of publications over time (Figure 3).

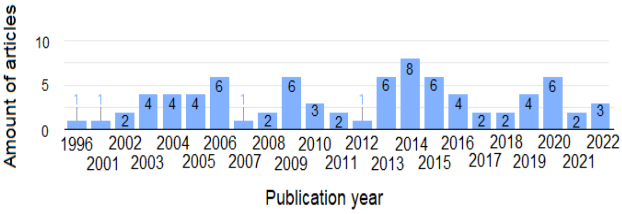

Figure 3. Distribution of publications over the years.

When analyzing the evolution of techniques used over time, there is a noticeable trend. In the early years, traditional statistical techniques such as N-grams, Hidden Markov models (HMMs), and Maximum Entropy (MaxEnt) were commonly used. Over time, machine learning techniques, including Decision Trees, Support Vector Machines (SVMs), and Max-Margin Markov Networks (M3Ns), gained prominence. In the context of machine learning, the last decade has been marked by the rise of Long Short-Term Memory (LSTM), Bidirectional Long Short-Term Memory (BLSTM), and Bidirectional Encoder Representations from Transformers (BERT). This trend is also realized in other computer science areas and reflects not only advances in computational power but also the increasing availability of more robust and accessible datasets (Figure 4), as well as the availability of libraries and other artifacts that contribute to faster development of programs. Additional details can be found in the appendix (Table 5). Conferences stood out as the most frequent publication vehicle (90% of occurrences), possibly due to a faster publication cycle.

We noted that there was no single trend in the motivation of studies seeking to detect disfluencies. Those who focused on removing disfluencies were mainly motivated to improve the performance of translation tools, summarizing of spoken dialogues, speaker diarization, and Natural Language Processing (NLP). In the context of syntactic analysis, the main interest was obtaining input that enabled a grammatical structure from a set of words, as well as exploring syntactic and semantic information in order to identify repairs (Section 4).

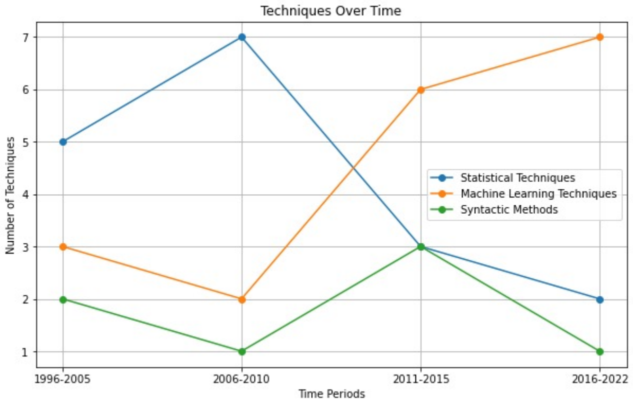

Figure 4. Categories of techniques employed for detecting disfluencies over time.

In this context, the study Stolcke and Shriberg [1996], for example, calculates the probability of a specific sequence of words appearing by taking into account the presence or absence of disfluencies. The work of Lin and Wang [2020] address the joint prediction of spelling scores and disfluencies in transcribed speech. They also highlight the mutual influence they exert on each other. For example, the repetition of "to" in "She decided to, to go home" might be a sign of disfluency or could emphasize the speaker's choice, introduce a new idea, or signal a moment of hesitation. The combination of these tasks is important for error reduction potential and improving the performance of speech recognizers and NLP tasks [Wang *et al.*, 2014b].

Since English is considered an international language, most of the research considers this idiom (60 articles, approximately 70%). Nevertheless, there is a significant portion of articles focused on Mandarin (8 articles, approximately 9%) and Chinese (4 articles, approximately 5%) as shown in Figure 5. It is important to remember that papers exploring multiple languages were counted multiple times, so these percentages may not reflect the number of distinct articles.

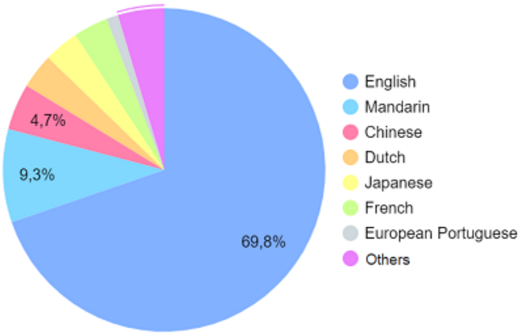

Figure 5. Overall proportion of languages explored within this systematic review.

4 Terminology and categorization of disfluencies

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". The number of studies focused on this topic may be related to the fact that "Interregnum" does not contribute to the semantic content of a speech; thus, their identification and cleaning improves the readability of transcription tools and speech recognizers.

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.

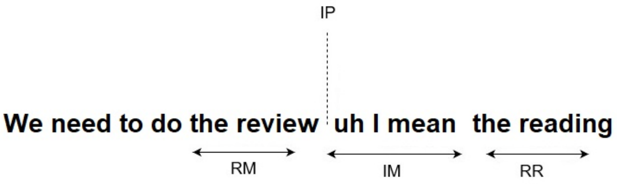

Figure 6. Example of Shriberg's terminology.

5 Techniques

Figure 7 shows the techniques used in the included articles according to their objectives. The most applied Statistical and Mathematical techniques were CRFs (Conditional Random Fields) and Maxent (Maximum Entropy). Among machine-learning techniques, Decision Tree and BLSTM (Bidirectional Long Short-Term Memory) are the most prominent ones. Some authors have proposed the application of techniques little explored in the literature, such as the Boosting algorithm.

In addition to the analysis and the purpose of this study, disfluency processing techniques were divided into single approaches when only one technique was used and hybrid approaches when two or more techniques were employed, as detailed in the following subsections.

5.1 Single approach techniques

Figure 8 presents an overview of the study areas in which single techniques were employed. The Machine Learning field is the one that stands out the most than others. The main techniques are detailed below.

5.1.1 Statistics and Mathematics

This category encompasses techniques that use approaches from statistics and/or mathematics in collecting and analyzing data for decision-making.

CRFs (Conditional Random Fields) CRFs are a powerful statistical modeling framework designed specifically for sequence labeling tasks [Fitzgerald *et al.*, 2009]. Unlike traditional classifiers that treat each label independently, CRFs consider the global context of the sequence, allowing them to capture interdependencies between labels. This makes them particularly well-suited for tasks where the prediction of one label can influence the prediction of subsequent labels, such as named entity recognition. A feature function maps the input sequence and its features to a vector representation, capturing relevant information for label prediction. A conditional probability distribution then estimates the probability of each label sequence given the feature vector. This allows the model to learn the relationships between labels and make predictions that are consistent with the overall context of the sequence. As an example, the approach developed in Shahih and Purwarianti [2016] uses CRFs during the pre-processing phase of a machine translation system dealing with disfluencies, and in Ostendorf and Hahn [2013] this technique is employed to improve speech transcripts readability.

Table 1. Summary of disfluency categories (*** means that no additional applicable terms were identified for this category).

Disfluency	Applicable terms	Description	Citations
Interregnum	Filler, filled pause, interjection, discourse marker, edit term	Optional part of speech that indicates reasoning development or confusion. May involve abnormal lengthening of words. Includes expressions such as “um”, “eh”, “ah”, “oh”, “uh”, “well”, “you know”, “I mean”, “anyways”, “basically”, “let’s see”, “like”, “actually”, “so”, “okay”, “now”, “moreover”. In the structure of Shriberg [1994], it marks the end of the reparandum. Example: “Show me flights from Boston on **uh, I mean** from Denver on Monday”	44 (28%)
Repair	Correction, revision, restart	In the structure of Shriberg [1994], it is the part of the utterance that corrects and modifies the original utterance. Example: “Show me flights from Boston on uh, I mean **from Denver** on Monday”	29 (19%)
Edition	***	It is the *reparandum – interregnum – repair* structure defined in Shriberg [1994]. Example: “Show me flights [(from Boston on) * uh, I mean from Denver on] Monday” where the expression “(from Boston on)” is the reparandum, “*” marks the interruption point, “uh, I mean” is the interregnum and “from Denver on” is the repair	22 (14%)
Repetition	***	Exact or approximate copy of a statement. Ignoring the interregnum (filling in), occurs when the reparandum equals the repair. Example: “**This is this is** an example”	19 (12%)
Interruption point	***	It marks the end of the reparandum, it is the moment when the speaker interrupts an utterance, which may involve a revision, repetition, or abandonment of content in the sequence. In the structure of Shriberg [1994], it marks the end of the reparandum. Example: “We need to do the review * uh, I mean the reading”	12 (8%)
Reparandum	***	In the structure of Shriberg [1994], it is the part of the utterance that is discarded or corrected by the next words. Example: “Show me flights **from Boston on** uh, I mean from Denver on Monday”	12 (8%)
False start	Exclusion, rupture	Moment when the speaker abruptly starts a new sentence without completing the previous sentence, generating abandoned and incomplete clauses. Example: “**We’ll never** what about next month?”	8 (5%)
Word fragment	***	Occurs when a word is interrupted in mid-speech. Example: “**th-**that was”	5 (3%)
Insertion	***	Occurs when a word is inserted without context into the sentence. Example: “**Of a** that’s really good”	3 (2%)
Stuttering	***	Occurs when a spoken word is part of the following words. Example: “This is an **exa** example”	2 (1%)

HMM (Hidden Markov Model) This technique is a powerful statistical tool for understanding and predicting sequential data. They are particularly useful when the data contains hidden states that are not directly observable but influence the observed outcomes. HMMs leverage the Markov chain property, assuming that the probability of transitioning to the next state depends only on the current state, not on the entire history of the sequence. This allows them to efficiently capture temporal dependencies and make predictions about future observations. In Liu *et al*. [2006] this technique was explored with the purpose of enriching speech recognition output and in Liu *et al*. [2005] the performance of the HMM approach was compared to a Maxent and a CRF model for disfluency detection.

HHMM (Hierarchical Hidden Markov Model) HHMM is a sophisticated extension of the traditional Hidden Markov Model (HMM), designed to address the complexities of stochastic and dynamic processes in various real-world applications. The HHMM introduces a structure where hidden states are organized into a hierarchy, allowing the model to represent more complex dependencies and transitions across different levels [Bui *et al*., 2004]. In Miller and Schuler [2008] and Miller [2009], HHMM is used to capture the hierarchical organization inherent in spoken language while also accommodating disfluencies. In Miller and Schuler [2008], HHMM is used to better capture this structure, while Miller [2009] focuses on how HHMM enables efficient, incremental parsing of spontaneous speech with disfluencies by adding random variables for speech repair, enhancing robustness.

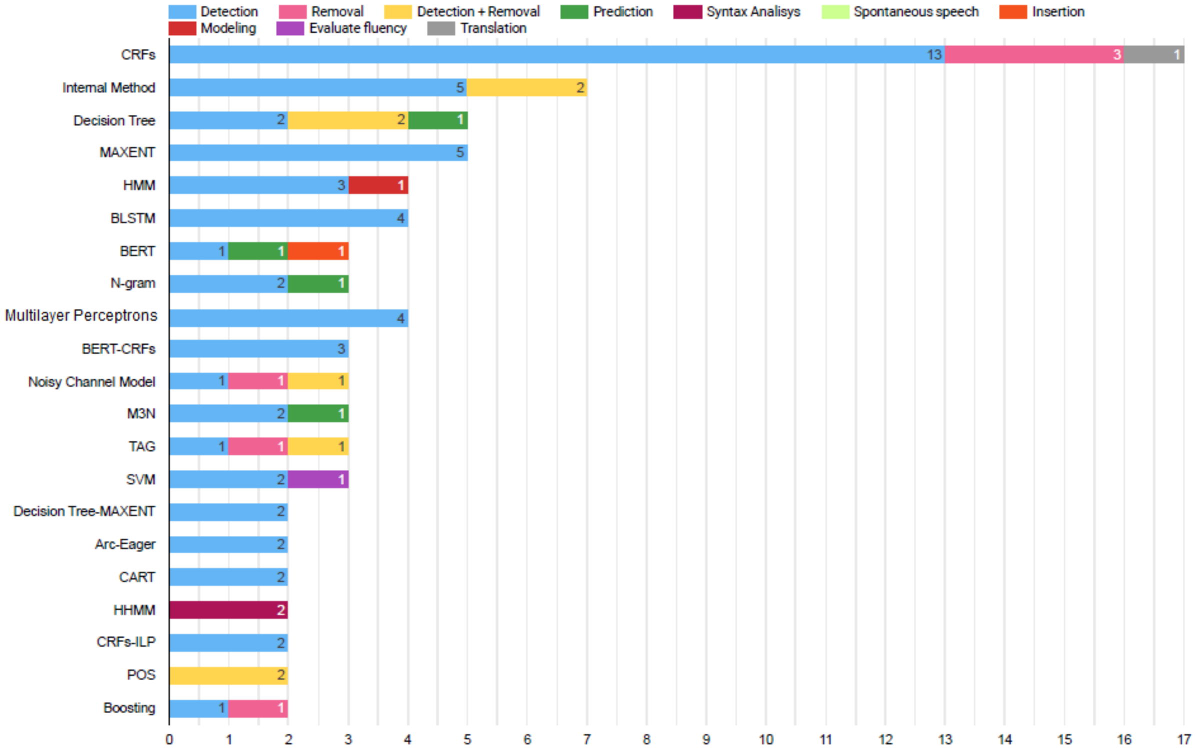

Figure 7. Main objectives and techniques from the publications in this review.

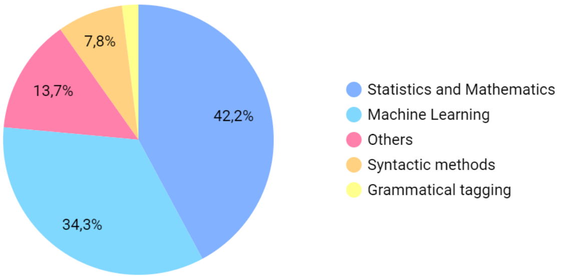

Figure 8. Distribution of single techniques by field of study.

Maxent (Maximum Entropy) MaxEnt models are widely used in NLP tasks like text segmentation, part-of-speech tagging, and named entity recognition. The principle of MaxEnt suggests that in situations of restricted information, the most impartial approach involves embracing uncertainty and refraining from making assumptions about unobserved data. The intuition lies in the concept of information gain. Each observation provides some information about the underlying process that generated it. MaxEnt seeks to maximize this information gain by choosing a model that is consistent with the data but avoids making unnecessary assumptions about the unobserved aspects of the process. This approach leads to robust models that are less likely to overfit the training data and generalize better to unseen examples. One challenge lies in formulating appropriate constraints that capture the essential features of the problem. Additionally, finding the maximum entropy distribution can be computationally expensive for complex systems. The research in Chen and Yoon [2011] employed this method to detect disfluencies in non-native English speakers and in Lin and Lee [2005] it is used to identify interruption points in Mandarin idiom.

NCM (Noisy Channel Model) Widely used in statistical speech recognition and automatic translation [Zwarts and Johnson, 2011], NCM assumes that when a message is created and transmitted over a communication channel, it is subject to degradation that corrupts the original message. The NCM is characterized by two primary probabilities. The first is *P(x|s)*, representing the conditional probability of receiving message x given that the source message was s. This probability encapsulates the impact of noise on the channel. The second probability is *P(s)*, denoting the prior probability of the source message being s. This probability incorporates any pre-existing knowledge about the types of messages the source is inclined to generate. Leveraging these probabilities, the NCM employs diverse decoding algorithms to reconstruct the original message [Honal and Schultz, 2005]. NCM was applied in Zwarts *et al*. [2010] to detect repairs and help speech recognizers interrupt or correct a speaker´s mistake. In Zwarts and Johnson [2011] the detection of repairs was investigated through an NCM that proposes a list of the 25 best hypotheses for disfluency.

N-grams N-grams are a fundamental concept in NLP and other fields that deal with sequential data. They represent sequences of *n* consecutive elements, such as words, characters, or syllables. N-grams are used to capture the statistical dependencies between these elements, which are crucial for understanding the structure and meaning of the data. There are various types of n-grams, depending on the value of *n*. Unigrams are individual elements, such as single words or characters. Bigrams are composed of consecutive pairs of elements, while Trigrams are sequences of three consecutive elements. Higher-order n-grams extend this pattern where "n" holds any higher value. They capture increasingly complex dependencies but may be computationally more expensive to process. In Stolcke and Shriberg [1996] a model is developed to predict disfluencies. In Germesin *et al.* [2008] this approach was used to develop a system capable of dealing with various types of disfluency, focusing on improving the structure of documented meetings.

5.1.2 Machine learning

In this category, algorithms for classification or prediction based on models learned from data were included as follows.

DT (Decision Tree) This approach relies on sequential assessments of information organized within a tree-like structure. The process initiates from a root node representing the entire dataset, and at each level of the tree, a division is made based on a chosen attribute to homogenize the samples in the leaves. Each leaf represents a decision or estimation, depending on whether the tree is used for classification or regression. Additionally, Decision Trees can handle both categorical and numerical data, providing versatility in addressing diverse datasets. Overfitting occurs when the tree captures noise in the data, resulting in a model that performs well on training data but poorly on new, unseen data. Techniques like limiting the tree's depth or implementing minimum sample requirements per leaf node aid in preventing overfitting. Moreover, ensemble methods such as Random Forests or Gradient Boosting combine multiple trees to enhance predictive accuracy and generalization. These techniques mitigate the limitations of individual Decision Trees and bolster the robustness of the model in handling diverse datasets. By using this technique, existing problems in summarizing spoken dialogues were addressed in Zechner [2001], mainly in corporate meeting conversations, sales, and customer support. Another example is presented in Womack *et al.* [2012], where medical narratives were analyzed, and the use of fillers in speech was compared between experienced doctors and doctors in training.

CART (Classification and Regression Trees) CART is a particular method for building decision trees, focusing on both classification and regression tasks. Classification trees are employed when the target variable consists of a finite number of unordered values. On the other hand, regression trees are used to handle continuous or ordered discrete target variables [Loh, 2011]. The studies in both Liu [2003] and Medeiros *et al.* [2013] employed CART to analyze speech features. In Liu [2003], CART helped identify the acoustic-prosodic features most influential in predicting word fragments. Similarly, in Medeiros *et al.* [2013], CART not only achieved the best overall classification of disfluent sequences but also helped researchers understand which acoustic and prosodic features were most crucial for identifying different parts of a disfluency.

SVM (Support Vector Machine) SVM represents a powerful and versatile class of supervised learning algorithms employed in both classification and regression tasks. As a linear classifier, SVM seeks to delineate optimal hyperplanes within the feature space to segregate distinct categories. It accomplishes this by identifying support vectors—data points lying closest to the decision boundary—and determining the hyperplane that maximally separates these vectors, thus establishing a clear margin between classes. SVM's efficacy extends beyond linear scenarios, as it can be adapted through kernel functions to address non-linear relationships in the data. This technique was highlighted in Deng *et al.* [2020] in the context of second language fluency assessment methods. Another example is presented in Tsvetkov *et al.* [2013], where an SVM classifier was built to identify word fragments after the degradation caused by disfluencies in the performance of speech recognizers.

BERT (Bidirectional Encoder Representations from Transformers) Represents a significant advancement in NLP due to its remarkable ability to understand the context and meaning of words within a sentence [Lin and Wang, 2020]. Unlike previous models, which relied solely on left-to-right word order, BERT is "bidirectional", meaning it analyzes the entire sentence simultaneously, allowing it to capture deeper semantic relationships within text, leading to superior performance on tasks requiring a nuanced understanding of language. For example, in Rocholl *et al.* [2021] BERT was used to ensure high performance in the detection of reparandums and speech markers, and in Yang *et al.* [2020] generated disfluent sentences from fluent sentences while seeking to understand how the generated sentences could help in the task of detecting disfluencies.

BLSTM (Bidirectional Long Short-Term Memory) Is a powerful recurrent neural network architecture designed to address the challenges of processing sequential data with long-range dependencies. While LSTMs process information sequentially, from the beginning to the end of a sequence, BLSTMs utilize two LSTMs operating in opposite directions. One LSTM analyses the sequence from left to right, while the other reads it from right to left. This allows BLSTMs to capture contextual information from both the past and the future, significantly enhancing their ability to handle long-range dependencies. BLSTM was applied in Zayats *et al.* [2016] in order to understand the connections between disfluencies and cognitive load, as well as a speaker's social context. Another example of BLSTM application is presented in Wang *et al.* [2016], which reinforces the importance of detecting disfluencies in natural language comprehension while treating the problem through a sequence-to-sequence approach.

Multilayer Perceptrons Neural networks are inspired by the human brain and designed to learn by processing information. They are built from simple processing units called neurons, connected in layers. Each connection has a weight that determines its influence. As the network processes data, the weights between neurons are adjusted to improve future performance. This ongoing process allows the network to learn and establish relationships between the data it receives and the results it produces [Abdi *et al.*, 1999]. When there are more than two layers, the network is referred to as "deep". Previous research has explored this technique for self-training a neural disfluency detection model [Jamshid Lou and Johnson, 2020] and jointly modeling punctuation prediction and disfluency detection [Cho *et al.*, 2015].

M3Ns (Max-Margin Markov Networks) M3Ns are a powerful framework for structured prediction tasks, where the output space is not simply a set of independent labels but rather a complex structure with dependencies between its components. This framework combines the strengths of both discriminative and generative models, leveraging the advantages of both approaches while minimizing their limitations [Taskar *et al.*, 2004]. It was described in Wang *et al.* [2014a] as a kind of SVM analogous to the CRF seeking to maximize the probabilistic difference between a true sequence and an incorrectly predicted sequence. Researches carried out in Wang *et al.* [2014a] and Qian and Liu [2013] use this technique, highlighting the importance of improving transcript readability and improving the performance of tasks inserted into NLP.

5.1.3 Syntactic methods

This category considers techniques that applied sequential encodings to represent structural patterns.

TAG (Tree Adjoining Grammar) Represents linguistic constructions as trees, allowing for the recursive combination of elementary trees through tree adjoining operations. These operations facilitate the modeling of syntactic relationships and dependencies, offering a nuanced approach to language analysis. Initial trees represent elements such as noun phrases and verb phrases. Adjoining trees, on the other hand, represent substructures that can be "adjoined" to initial trees at specific locations, enriching the overall syntactic structure. This allows TAGs to capture complex linguistic phenomena. The research presented in Lease *et al.* [2006] shows the importance of enriched transcripts that consider disfluencies by using the TAG technique. In Lease and Johnson [2006] the authors investigated the benefit of deleting speech completions in advance to improve the performance of text-based processing.

Arc-Eager-based parser Initially, a single root node is established to represent the entire sentence. Subsequently, in the shift operation, tokens from the input buffer are successively added to the stack, becoming the current node. If possible, a dependency arc is formed between the top two nodes on the stack, establishing the current node as the child and the preceding node as the parent—this step is termed Arc-Left. Upon the current node having all its children connected, it is removed from the stack and incorporated into the dependency tree—a process termed Reduce. This sequence—Shift, Arc-Left, Reduce—is reiterated until the input buffer is exhausted, and solely the root node remains within the stack, encapsulating the complete dependency structure of the sentence [Rasooli and Tetreault, 2014]. In Yoshikawa *et al.* [2016] a syntactic analysis method based on the Arc-Eager algorithm was proposed to deal with disfluencies directly from speech recognizer outputs. In Honnibal and Johnson [2014] the technique was used to detect repairs and restarts.

5.2 Hybrid techniques

As shown in figure 9, the results highlight the dominance of Machine Learning combined with Statistics and Mathematics, showcasing their powerful synergy in this field. This trend is evident in various combinations like Decision Tree-Maxent [Lin and Lee, 2009], BLSTM-CRFs [Tanaka *et al.*, 2019], and LSTM-Noisy Channel Model [Jamshid Lou and Johnson, 2017]. These approaches leverage the strengths of different methods to achieve superior performance. As pointed out in Hristea [2011], the application of statistical methods and machine learning algorithms to solve NLP problems has been occurring since the 90s. This is due to their prediction robustness based on some input and data inference on different variable relationships.

While less common, statistical and mathematical techniques combined with linear programming also emerged as a significant approach. This is seen in the integration of methods like Maxent and CRFs with linear programming [Georgila *et al.*, 2010; Georgila, 2009].

Finally, the combination of machine learning and linear programming, exemplified by HELM-ILP [Georgila, 2009], appears to be a less developed area, suggesting that there is potential for further exploration.

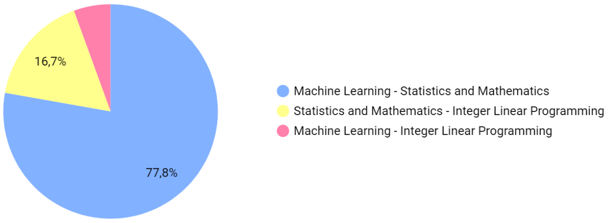

Figure 9. Distribution of single techniques by field of study.

6 Databases and transcription tools

A challenge in the context of disfluency analysis is the composition of databases that are adequate for evaluating methods and techniques. In this context, the EARS (Effective, Affordable, Reusable Speech-to-Text) program was created by the North American research agency DARPA, with the aim of advancing the state of the art in speech recognition. Under the administration of the NIST (National Institute of Standards and Technology), the program's efforts focused on research in conversion from speech to text (Speech-to-text / STT) and the readability of this information, also known as metadata extraction (Metadata Extraction / MDE) [Strassel, 2004].

Among the supporters of the EARS program, it is worth highlighting the Linguistic Data Consortium [LDC, 2024]. Founded in 1992, the LDC brings partnering universities, corporations, and research laboratories together in order to collect and distribute linguistic resources while offering a variety of transcribed and annotated databases annotated to a series of phenomena. Most of the studies included in this systematic review used some dataset distributed by the LDC, with Switchboard being the most reported one.

Table 2 provides details of the databases used in the included articles, such as citation number, distribution, language, composition, transcription availability, and disfluency annotation. The databases were categorized according to availability, being classified as “Public” (when the data used was publicly available), “Private” (when the data used was not publicly available) and “Commercial private” (when the data used was available publicly but under conditions of some license payment). The acronym “NI” was used when the data was not informed while “NA” when not applied. As shown, 43% of the databases have public distribution, 27% did not inform, 23% are of the private-commercial type and 7% are private. Studies using more than one database were counted for all databases used.

Considering that the main databases already made the audio recordings transcripts available, most studies (67 articles) did not use a transcription tool to test proposed models. Among the studies that used transcription technologies and mentioned the name of the tool, 20% opted for the Hidden Markov Model Toolkit. Other studies used Kaldi (10%), Iflyrec (10%), Audimus.Media (10%), BBN Byblos (10%), LIUM (10%) and Microsoft Research S2S (10%). An overview of the transcription tools used is presented in Table 3. The acronym “NI” was used when the data was not informed.

7 Evaluation Measures

All included studies carried out some type of evaluation of the proposed method, using measures like F-score, precision, and recall (as highlighted in Figure 10). Studies were counted for each evaluation method used.

Disregarding traditional measures and taking into account specific measures on the present study scenario, it is pertinent to highlight the Edit Error Rate and IP Error Rate. This is established as the average number of missed detections and false alarms for edits and interruption points, respectively. Such measures were defined by the NIST (National Institute of Standards and Technology) in RT-04F (Rich Transcription Fall 2004 Evaluation). For this reason, they are also referred to in the literature as the NIST Error Rate [Wang *et al.*, 2010; Liu *et al.*, 2006].

Another very important measure in speech recognition is the Word Error Rate (WER), consisting of the sum of insertions, deletions, and substitutions divided by the original number of words. For this rate, the smaller the value, the better the tool performs.

8 Limitations, challenges and opportunities

8.1 Limitations and challenges

As most studies relied on transcripts developed by humans, this implies the availability of a transcribed and annotated database for disfluencies. Nevertheless, the main cited databases that have transcription and annotation for disfluencies are commercial (Section 6). This fact is probably due to the cost involved in building such bases, which requires the action of many people in order to obtain reliable results. Making public databases available with annotations so that various research groups can benefit is still a research challenge.

We noted that transcription tools have the default behavior of removing disfluencies or trying to correct for sentences that are grammatically correct, or close to it.

To enhance the reader’s understanding of the comparative effectiveness of various disfluency detection methods, we conducted a performance analysis across eight studies (Table 4), focusing on the reported F-score measure of these methods. In order to accomplish this objective, we selected the most similar studies, i.e., those that utilized the same language (English) and the same dataset (Switchboard). Even so, the fact that some of these studies involved multiple languages and databases introduces variability that may affect the comparability of results. Moreover, different definitions of disfluencies, along with variations in dataset size, quality, and preprocessing methods, can introduce biases. Despite these challenges, analyzing the data reveals that newer techniques, such as BERT and BLSTM, tend to outperform older methods, suggesting that advancements in technology and research have led to improved performance. BLSTM, in particular, achieves the highest F-score of 91.80, as reported in Bach and Huang [2019].

8.2 Opportunities

Some questions to be explored in the future were raised. The first is to understand the applicability of techniques developed in languages other than the one for which the technique was designed. The idea is to evaluate the performance of the techniques, verifying if there is any loss of performance when changing a language or not. As presented in Table 5 of the appendix, approximately 70% of the included articles were developed for the English language. While this fact is a limitation of the current state of the art, it is also a valuable opportunity for research. It is also a research opportunity for configurable tools that enable the choice of whether to detect disfluencies or not, as well as the type of disfluencies to be detected.

It is worth further exploring automatic speech recognizer outputs, assessing the impacts of their use and generating input for the development of transcription tools. Additionally, the potential for improving disfluency detection was noticeable in the detection of structural events such as spelling punctuation (Section 3). As such, it is interesting to study the contribution of other conversation elements such as visual

Table 2. Summary of databases

Database	Citations	Distribution	Language	Composition	Transcript	Annotates disfluency?
Switchboard	46 (49%)	Godfrey and Holliman [1993b,a]; Graff *et al.* [1998, 1999, 2001a,b,c, 2002, 2004]; Calhoun *et al.* [2009]	English	Phone Conversations and Wall Street Journal Collections	Yes	Yes
Mandarin Conversational Dialogue Corpus (MCDC)	5 (5%)	S.-C. [2004]	Mandarin	27 hours of dialogue	Yes	Yes
Fisher	4 (4%)	Cieri *et al.* [2004]	English	984 hours of recorded phone conversations	Yes	No
CallHome	3 (3%)	Canavan *et al.* [1997]; Kingsbury *et al.* [1997]	English	120 30-minute phone conversations	Yes	No
Spoken Dutch Corpus	3 (3%)	Oostdijk [2002]	Dutch	Speech samples from different regions of Holland and Flanders (Belgium)	Yes	Yes
SCOTUS	3 (3%)	Oyez [2022]	English	US Supreme Court Cases, recordings with about 1h	Yes	No
RT-04 MDE	3 (3%)	Walker *et al.* [2005]; Lee and Strassel [2005]	English	Approximately 60 hours of phone conversations	Yes	Yes
Corpus of Spontaneous Japanese (CSJ)	3 (3%)	Maekawa [2003]	Japanese	Dialogues containing about 7 million words	Yes	Yes
Financial Crisis Inquiry Commission (FCIC)	2 (2%)	Zayats *et al.* [2014]	English	Recorded Audiences	Yes	Yes
Spontaneous Speech Reconstruction (SSR)	2 (2%)	Fitzgerald and Jelinek [2008]	English	Fisher extension, 6400-phrase set	Yes	Yes
French Broadcast News	1 (1%)	Dufour *et al.* [2009]	French	11:37 minutes collected from five radio programs	Yes	No
DARPA-TransTac	1 (1%)	Stallard *et al.* [2011]	English	Dialogues in different contexts	Yes	No
Lectra	1 (1%)	Trancoso *et al.* [2008]	European Portuguese	1 semester of classes between 60-90 minutes	Yes	Yes
Mandarin Chinese CallHome (MCC)	1 (1%)	Honal and Schultz [2004]	Mandarin	100 recorded dialogs	Yes	Yes
English Verbmobil	1 (1%)	Honal and Schultz [2004]	English	127 recorded dialogs	Yes	Yes
TOEFL Practice Online Test	1 (1%)	Chen and Yoon [2011]	English	1,066 responses collected from a proficiency test	Yes	Yes
Rapport	1 (1%)	Gratch *et al.* [2007]	English	Interaction with a virtual agent	No	Yes
AMI Meeting	1 (1%)	Carletta *et al.* [2005]	English	100 hours of recorded meetings	Yes	No
GroupMeeting	1 (1%)	Zechner [2002]	English	Carnegie Project Group Meetings-Mellon University	NI	NI
Voxfactory	1 (1%)	Clavel *et al.* [2013]	French	1000 hour recordings set from a call center	Yes	Yes
CrossFire	1 (1%)	Zechner [2002]	English	Excerpts from the CrossFire television show from CNN (1998)	NI	NI
GALE Ontonotes Chinese	1 (1%)	Wang *et al.* [2010]	Chinese	3,230 sentences	Yes	Yes
Phonologie du Français Contemporain	1 (1%)	Avanzi [2014]	French	Text readings and interviews that add up more than 11 hours long	Yes	Yes
NewsHour	1 (1%)	Zechner [2002]	English	Excerpts from the TV show PBS NewsHour (1998)	NI	NI
IWSLT	1 (1%)	Lin and Wang [2020]	NI	NI	NI	NI
Pushshift Reddit	1 (1%)	Baumgartner *et al.* [2020]	English	Comments from a community on the Reddit platform	NA	No
CallFriend	1 (1%)	Canavan and Zipperlen [1996]	English	Phone conversations	Yes	No
MAT Speech Database	1 (1%)	Yeh and Wu [2006]	Mandarin	NI	NI	NI
IARPA Babel project	1 (1%)	Tsvetkov *et al.* [2013]	Cantonese and Turkish	71 hours of speaking in Cantonese and 40 hours in Turkish	Yes	Yes
PhoATIS	1 (1%)	Dao *et al.* [2022]	Vietnamese	5871 fluent utterances	Yes	Yes

and gestural movements. In the same way that some studies used acoustic and prosodic resources for detecting disfluencies (Section 5), it is pertinent to understand that other resources can be extracted for improved detection, and may even provide inputs for the diagnosis of emotions such as anxiety, fear, doubt and others.

As mentioned, annotated databases are still challenging in the area. A promising approach to address this issue is using shared tasks, as defined in SIGEDU [2024]. Shared tasks can involve collaborative research and competition, where

Table 3. Summary of main transcription tools used

#	Tool	Description	Availability	Cited in
1	Hidden Markov Model Toolkit (HTK)	Software designed to cover general speech processing activities	Public	Yeh and Wu [2006];Yeh *et al.* [2007]
2	Kaldi	A toolkit developed in C++	Public	Yoshikawa *et al.* [2016]
3	Iflyrec	Supported transcription platform for 9 languages that promises up to 97.5% accuracy	Private	Wang *et al.* [2017]
4	Audimus.Media	Speech recognizer for European Portuguese. Developed in Meinedo *et al.* [2003]	NI	Medeiros *et al.* [2013]
5	Byblos	Speech recognizer designed to handle large vocabularies Chow *et al.* [1987]	NI	Gupta *et al.* [2014]
6	LIUM	Speech recognizer based on the Sphinx system	NI	Dufour *et al.* [2009]
7	Microsoft Research S2S	Speech Recognition and Machine Translation system	NI	Hassan *et al.* [2014]

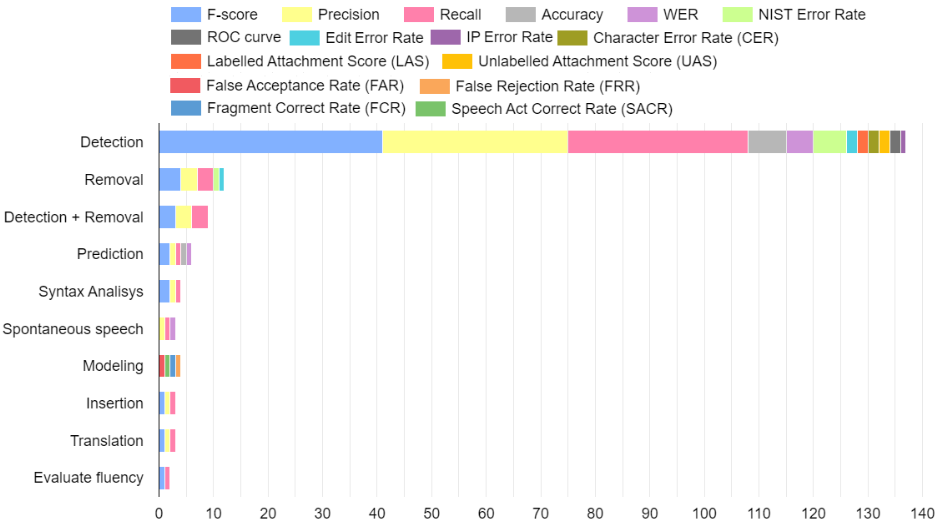

Figure 10. Main measures used and objectives.

Table 4. Performance comparison across studies considering the F-score measure and studies that consider English as idiom and Switchboard dataset.

Reference	Technique	F-score
Bach and Huang [2019]	BLSTM	91.80
Rocholl *et al.* [2021]	BERT	90.90
Qian and Liu [2013]	M3N	84.10
Zwarts and Johnson [2011]	NCM	83.80
Zayats *et al.* [2014]	CRFs	82.80
Rasooli and Tetreault [2014]	Arc-Eager	82.60
Johnson and Charniak [2004]	TAG	79.70
Miller and Schuler [2008]	HHMM	77.15

researchers and practitioners work together to solve shared problems using common data and evaluation measures.

Some recent examples of this practice share databases, code, and experiences. In Tack *et al.* [2023] is presented a contest aimed to evaluate the ability of generative language models to act as AI teachers, replying to a student in a teacher-student dialogue. Eight teams participated, employing various state-of-the-art models. All participants worked with training and test samples derived from the Teacher-Student Chatroom Corpus [Caines *et al.*, 2020, 2022], which provided dialogue contexts and reference teacher responses for training. While the results were promising, the study emphasized the need for more suitable evaluation metrics for educational contexts.

Another example can be found in Volodina *et al.* [2023], where researchers were tasked with developing systems to detect grammatical errors in five languages: Czech, English, German, Italian, and Swedish. The participants were provided with training, development, and test datasets for each language, and were allowed to incorporate additional publicly available resources, such as monolingual data, artificial data, pre-trained models, and syntactic parsers. This effort aimed to address the lack of representation of certain languages in grammatical error detection research. By focusing on multilingual datasets, the task encouraged the creation of models that are both more robust and adaptable across different languages.

Competitions have played a significant role in promoting dataset sharing. As highlighted by Ribeiro and Nunes [2022], events like SATA [2024], Kaggle [2024], and ACDC [2024] have significantly influenced the development of left ventricle segmentation methods, allowing comparisons among different approaches trained and tested with common datasets. Similarly, shared tasks could contribute to the analysis and classification of disfluencies by providing common datasets.

Although shared tasks have proven beneficial in both academia and industry, some challenges remain, including issues with transparency, secretive practices, conflicts of interest, and unequal access to resources. To address these challenges, it is necessary to encourage participants to disclose their systems, resources, successes, and failures, as stated by Nissim *et al.* [2017].

9 Conclusion

Building human-computer conversational interfaces is a challenging task due to the wide diversity of expressions and events in spontaneous speech. With the understanding that disfluencies directly impact readability for speech recognizers, this article presented a systematic review identifying the main challenges along with research opportunities in the area.

Researchers have made significant progress in developing methods to detect disfluencies using various techniques, with a focus on English. Statistical and machine learning approaches have been dominant, and evaluation measures like F-score, precision, and recall are widely used. A comparison we conducted among the most similar studies (with the same idiom and the same dataset) has shown that the more recent approaches, such as BERT and BLSTM, have provided higher performance. However, limitations and opportunities remain.

A major challenge is the lack of publicly available, annotated databases for disfluencies. Existing resources are often commercial, hindering collaboration and comparison of research findings. Additionally, automatic transcription tools often remove disfluencies, further complicating analysis.

Future research should explore the applicability of disfluency detection techniques across languages. Developing configurable tools that allow users to choose the type and level of disfluency detection would also be valuable. Another promising direction lies in integrating visual and gestural cues with audio data to enhance detection capabilities. Much like the current utilization of acoustic and prosodic features, these supplementary elements could greatly enhance disfluency recognition and potentially provide valuable insights into emotional states.

Declarations

Authors' Contributions

ASL contributed to conceptualization, data curation, formal analysis, investigation, methodology, software and writing – original draft. AML contributed to conceptualization, validation, writing – review & editing. FLSN contributed to conceptualization, validation, writing – review & editing. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

The authors thank the volunteers who participated in the experiment. This study was partially funded by the National Council for Scientific and Technological Development (CNPq) - process 307710/2022-0) and São Paulo Research Foundation (FAPESP) – National Institute of Science and Technology – Computer Assisted Medicine Scientific (INCT-MACC) – grant process 2014/50889-7.

Availability of data and materials

The datasets (and/or softwares) generated and/or analysed during the current study will be made upon request.

References

Abdi, H., Valentin, D., and Edelman, B. (1999). *Neural networks*. Number 124. Sage. DOI: 10.4135/9781412985277.

ACDC (2024). Automated cardiac diagnosis challenge. Last accessed 05 December 2024.

Avanzi, M. (2014). A corpus-based approach to french regional prosodic variation. *Cahiers de linguistique française*, (31):309–323. DOI: 10.1093/oxfordhb/9780198865131.013.20.

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.

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.

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.

Belz, M., Müller, M., and Mooshammer, C. (2023). How consistent are non-native speakers in their usage of filler particles when talking to native speakers? In *Disfluency in Spontaneous Speech (DiSS) Workshop 2023*, pages 53–57. DOI: 10.21437/DiSS.2023-11.

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.

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`.

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 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.

Caines, A., Yannakoudakis, H., Edmondson, H., Allen, H., 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.

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

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

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

Carletta, J., Kraaij, W., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kronenthal, M., Lathoud, G., Lincoln, M., Lisowska, A., Post, W., Reidsma, D., Wellner, P., and McCowan, L. (2005). The AMI meeting corpus. In *Proceedings of Symposium on Annotating and Measuring Meeting Behavior*. DOI: $10.1007/11677482_3$.

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.

Chen, Q., Chen, M., Li, B., and Wang, W. (2020). Controllable time-delay transformer for real-time punctuation prediction and disfluency detection. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8069–8073. DOI: 10.1109/ICASSP40776.2020.9053159.

Cho, E., Kilgour, K., Niehues, J., and Waibel, A. (2015). Combination of NN and CRF models for joint detection of punctuation and disfluencies. In *Proc. Interspeech 2015*, pages 3650–3654. DOI: 10.21437/Interspeech.2015-724.

Chow, Y., Dunham, M., Kimball, O., Krasner, M., Kubala, G., Makhoul, J., Price, P., Roucos, S., and Schwartz, R. (1987). Byblos: The BBN continuous speech recognition system. In *ICASSP '87. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 12, pages 89–92. DOI: 10.1109/ICASSP.1987.1169748.

Christodoulides, G. and Avanzi, M. (2015). Automatic detection and annotation of disfluencies in spoken french corpora. In *Proc. Interspeech 2015*, pages 1849–1853. DOI: 10.21437/Interspeech.2015-69.

Cieri, C. *et al.* (2004). Fisher english training speech part 1 transcripts ldc2004t19. DOI: 10.35111/w4bk-9b14.

Clavel, C., Adda, G., Cailliau, F., Garnier-Rizet, M., Cavet, A., Chapuis, G., Courcinous, S., Danesi, C., Daquo, A.-L., and Suignard, P. (2013). Spontaneous speech and opinion detection: Mining call-centre transcripts. *Language Resources and Evaluation*, 47:1–37. DOI: 10.1007/s10579-013-9224-5.

Dao, M. H., Truong, T., and Nguyen, D. Q. (2022). From disfluency detection to intent detection and slot filling. In *Proc. Interspeech 2022*, pages 1106–1110. DOI: 10.21437/Interspeech.2022-10161.

Demuynck, K., Duchateau, J., Van Compernolle, D., and Wambacq, P. (2000). An efficient search space representation for large vocabulary continuous speech recognition. *Speech communication*, 30(1):37–53. DOI: 10.1016/s0167-6393(99)00030-8.

Deng, H., Lin, Y., Utsuro, T., Kobayashi, A., Nishizaki, H., and Hoshino, J. (2020). Automatic fluency evaluation of spontaneous speech using disfluency-based features. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 9239–9243. DOI: 10.1109/ICASSP40776.2020.9053452.

Duchateau, J., Demuynck, K., and Van Compernolle, D. (1998). Fast and accurate acoustic modelling with semi-continuous HMMs. *Speech Communication*, 24(1):5–17. DOI: 10.1016/s0167-6393(98)00002-8.

Dufour, R., Estève, Y., Deléglise, P., and Béchet, F. (2009). Local and global models for spontaneous speech segment detection and characterization. In *2009 IEEE Workshop on Automatic Speech Recognition Understanding*, pages 558–561. DOI: 10.1109/ASRU.2009.5372928.

Dutrey, C., Clavel, C., Rosset, S., Vasilescu, I., and Adda-Decker, M. (2014). A CRF-based approach to automatic disfluency detection in a french call-centre corpus. In *Proc. Interspeech 2014*, pages 2897–2901. DOI: 10.21437/Interspeech.2014-601.

Ferguson, J., Durrett, G., and Klein, D. (2015). 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*, pages 257–262, Denver, Colorado. Association for Computational Linguistics. DOI: 10.3115/v1/N15-1029.

Fitzgerald, E., Hall, K., and Jelinek, F. (2009). Reconstructing false start errors in spontaneous speech text. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, page 255–263, USA. Association for Computational Linguistics. DOI: 10.3115/1609067.1609095.

Fitzgerald, E. and Jelinek, F. (2008). Linguistic resources for reconstructing spontaneous speech text. Available at: `http://www.lrec-conf.org/proceedings/lrec2008/pdf/874_paper.pdf`.

Georgila, K. (2009). Using integer linear programming for detecting speech disfluencies. In *Proceedings of Human Language Technologies: The 2009 Annual Con-*

ference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers, NAACL-Short '09, page 109–112, USA. Association for Computational Linguistics. DOI: 10.3115/1620853.1620885.

Georgila, K., Wang, N., and Gratch, J. (2010). Cross-domain speech disfluency detection. In *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, SIGDIAL '10, page 237–240, USA. Association for Computational Linguistics. Available at: `https://aclanthology.org/W10-4343.pdf`.

Germesin, S., Becker, T., and Poller, P. (2008). Domain-specific classification methods for disfluency detection. In *Proc. Interspeech 2008*, pages 2518–2521. DOI: 10.21437/Interspeech.2008-624.

Ghosh, S., Kumar, S., Kumar, Y., Ratn Shah, R., and Umesh, S. (2022). Span classification with structured information for disfluency detection in spoken utterances. In *Proc. Interspeech 2022*, pages 3998–4002. DOI: 10.21437/Interspeech.2022-11242.

Godfrey, J. J. and Holliman, E. (1993a). Switchboard-1 release 2 ldc97s62. DOI: 10.35111/sw3h-rw02.

Godfrey, J. J. and Holliman, E. (1993b). Switchboard credit card ldc93s8. DOI: 10.35111/cmtf-v363.

Graff, D., Canavan, A., and Zipperlen, G. (1998). Switchboard-2 phase i ldc98s75. DOI: 10.35111/c7th-nf28.

Graff, D., Miller, D., and Walker, K. (2002). Switchboard-2 phase iii audio ldc2002s06. DOI: 10.35111/ydsv-hw57.

Graff, D., Walker, K., and Canavan, A. (1999). Switchboard-2 phase ii ldc99s79. DOI: 10.35111/5qpg-1r82.

Graff, D., Walker, K., and Miller, D. (2001a). Switchboard cellular part 1 audio ldc2001s13. DOI: 10.35111/a74g-hy08.

Graff, D., Walker, K., and Miller, D. (2001b). Switchboard cellular part 1 transcribed audio ldc2001s15. DOI: 10.35111/3wcn-6c29.

Graff, D., Walker, K., and Miller, D. (2001c). Switchboard cellular part 1 transcription ldc2001t14. DOI: 10.35111/8j7x-fx86.

Graff, D., Walker, K., and Miller, D. (2004). Switchboard cellular part 2 audio ldc2004s07. DOI: 10.35111/mgp6-4j96.

Gratch, J., Wang, N., Gerten, J., Fast, E., and Duffy, R. (2007). Creating rapport with virtual agents. volume 4722, pages 125–138. DOI: 10.1007/978-3-540-74997-$4_1$2.

Gupta, N. and Bangalore, S. (2003). Segmenting spoken language utterances into clauses for semantic classification. In *2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No.03EX721)*, pages 525–530. DOI: 10.1109/ASRU.2003.1318495.

Gupta, N. K. and Bangalore, S. (2002). Extracting clauses for spoken language understanding in conversational systems. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, EMNLP '02, page 273–280, USA. Association for Computational Linguistics. DOI: 10.3115/1118693.1118728.

Gupta, R., Ananthakrishnan, S., Yang, Z., and Narayanan, S. S. (2014). Variable span disfluency detection in ASR transcripts. In *Proc. Interspeech 2014*, pages 2892–2896. DOI: 10.21437/Interspeech.2014-600.

Hassan, H., Schwartz, L., Hakkani-Tür, D., and Tur, G. (2014). Segmentation and disfluency removal for conversational speech translation. In *Proc. Interspeech 2014*, pages 318–322. DOI: 10.21437/Interspeech.2014-76.

Honal, M. and Schultz, T. (2004). Correction of disfluencies in spontaneous speech using a noisy-channel approach. DOI: 10.21437/eurospeech.2003-741.

Honal, M. and Schultz, T. (2005). Automatic disfluency removal on recognized spontaneous speech - rapid adaptation to speaker-dependent disfluencies. In *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, volume 1, pages I/969–I/972 Vol. 1. DOI: 10.1109/ICASSP.2005.1415277.

Honnibal, M. and Johnson, M. (2014). Joint incremental disfluency detection and dependency parsing. *Transactions of the Association for Computational Linguistics*, 2:131–142. DOI: 10.1162/tacl$_{a0}$0171.

Horii, K., Fukuda, M., Ohta, K., Nishimura, R., Ogawa, A., and Kitaoka, N. (2022). End-to-end spontaneous speech recognition using disfluency labeling. In *Proc. Interspeech 2022*, pages 4108–4112. DOI: 10.21437/Interspeech.2022-281.

Hough, J. and Schlangen, D. (2015). Recurrent neural networks for incremental disfluency detection. In *Proc. Interspeech 2015*, pages 849–853. DOI: 10.21437/Interspeech.2015-264.

Hristea, F. T. (2011). *Statistical Natural Language Processing*, pages 1452–1453. Springer Berlin Heidelberg, Berlin, Heidelberg. DOI: 10.1007/978-3-642-04898-$2_8$2.

Jamshid Lou, P., Anderson, P., and Johnson, M. (2018). Disfluency detection using auto-correlational neural networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4610–4619, Brussels, Belgium. Association for Computational Linguistics. DOI: 10.18653/v1/D18-1490.

Jamshid Lou, P. and Johnson, M. (2017). Disfluency detection using a noisy channel model and a deep neural language model. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 547–553, Vancouver, Canada. Association for Computational Linguistics. DOI: 10.18653/v1/P17-2087.

Jamshid Lou, P. and Johnson, M. (2020). Improving disfluency detection by self-training a self-attentive model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3754–3763, Online. Association for Computational Linguistics. DOI: 10.18653/v1/2020.acl-main.346.

Jamshid Lou, P., Wang, Y., and Johnson, M. (2019). Neural constituency parsing of speech transcripts. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2756–2765, Minneapolis, Minnesota. Association for Computational Linguistics. DOI: 10.18653/v1/N19-1282.

Johnson, M. and Charniak, E. (2004). A tag-based noisy channel model of speech repairs. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, ACL '04, page 33–es, USA. Association for Computational Linguistics. DOI: 10.3115/1218955.1218960.

Kaggle (2024). Data science bowl cardiac challenge data. Available at: `https://www.kaggle.com/c/second-annual-data-science-bowl`. Last accessed 05 December 2024.

Khara, S., Singh, S., and Vir, D. (2018). 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)*, pages 887–893. DOI: 10.1109/ICICCT.2018.8473099.

Kingsbury, P. *et al.* (1997). Callhome american english transcripts ldc97t14. DOI: 10.35111/z1z4-ep76.

Kouzelis, T., Paraskevopoulos, G., Katsamanis, A., and Katsouros, V. (2023). Weakly-supervised forced alignment of disfluent speech using phoneme-level modeling. In *INTERSPEECH 2023*, pages 1563–1567. DOI: 10.21437/Interspeech.2023-1887.

LDC (1992-2024). LDC - linguistic data consortium. Available at: `https://www.ldc.upenn.edu/about`. Last accessed 05 December 2024.

Lease, M. and Johnson, M. (2006). Early deletion of fillers in processing conversational speech. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, NAACL-Short '06, page 73–76, USA. Association for Computational Linguistics. DOI: 10.3115/1614049.1614068.

Lease, M., Johnson, M., and Charniak, E. (2006). Recognizing disfluencies in conversational speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(5):1566–1573. DOI: 10.1109/TASL.2006.878269.

Lee, D., Ko, B., Shin, M. C., Whang, T., Lee, D., Kim, E., Kim, E., and Jo, J. (2021). Auxiliary sequence labeling tasks for disfluency detection. In *Proc. Interspeech 2021*, pages 4229–4233. DOI: 10.21437/Interspeech.2021-400.

Lee, H. and Strassel, S. (2005). RT-04 MDE training data speech ldc2005s16. DOI: 10.35111/27r9-h809.

Lendvai, P. (2003). Learning to identify fragmented words in spoken discourse. In *Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics - Volume 2*, EACL '03, page 25–32, USA. Association for Computational Linguistics. DOI: 10.3115/1067737.1067742.

Li, X., Ishi, C. T., Fu, C., and Hayashi, R. (2022). Prosodic and voice quality analyses of filled pauses in japanese spontaneous conversation by chinese learners and japanese native speakers. In *Speech Prosody 2022*, pages 550–554. DOI: 10.21437/SpeechProsody.2022-112.

Lickley, R. J. (2015). Fluency and disfluency. *The handbook of speech production*, pages 445–469. DOI: 10.1002/9781118584156.ch20.

Lin, B. and Wang, L. (2020). Joint prediction of punctuation and disfluency in speech transcripts. In *Proc. Interspeech 2020*, pages 716–720. DOI: 10.21437/Interspeech.2020-1277.

Lin, C.-K. and Lee, L.-S. (2005). Improved spontaneous mandarin speech recognition by disfluency interruption point (IP) detection using prosodic features. In *Proc. Interspeech 2005*, pages 1621–1624. DOI: 10.21437/Interspeech.2005-533.

Lin, C.-K. and Lee, L.-S. (2009). Improved features and models for detecting edit disfluencies in transcribing spontaneous mandarin speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(7):1263–1278. DOI: 10.1109/TASL.2009.2014792.

Lin, C.-K. and shan Lee, L. (2006). Latent prosodic modeling (LPM) for speech with applications in recognizing spontaneous mandarin speech with disfluencies. In *Proc. Interspeech 2006*, pages paper 1901–Thu1FoP.9. DOI: 10.21437/Interspeech.2006-599.

Liu, Y. (2003). Word fragment identification using acoustic-prosodic features in conversational speech. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Proceedings of the HLT-NAACL 2003 Student Research Workshop - Volume 3*, NAACLstudent '03, page 37–42, USA. Association for Computational Linguistics. DOI: 10.3115/1073416.1073423.

Liu, Y., Shriberg, E., Stolcke, A., and Harper, M. (2004). Using machine learning to cope with imbalanced classes in natural speech: evidence from sentence boundary and disfluency detection. In *Proc. Interspeech 2004*, pages 1525–1528. DOI: 10.21437/Interspeech.2004-573.

Liu, Y., Shriberg, E., Stolcke, A., and Harper, M. (2005). Comparing HMM, maximum entropy, and conditional random fields for disfluency detection. In *Proc. Interspeech 2005*, pages 3313–3316. DOI: 10.21437/Interspeech.2005-851.

Liu, Y., Shriberg, E., Stolcke, A., Hillard, D., Ostendorf, M., and Harper, M. (2006). Enriching speech recognition with automatic detection of sentence boundaries and disfluencies. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(5):1526–1540. DOI: 10.1109/TASL.2006.878255.

Loh, W.-Y. (2011). Classification and regression trees. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, 1(1):14–23. DOI: 10.1002/widm.8.

Maekawa, K. (2003). Corpus of spontaneous japanese: Its design and evaluation. *Proceedings of SSPR*. Available at: `https://www.isca-archive.org/sspr_2003/maekawa03_sspr.html#`.

Maskey, S., Zhou, B., and Gao, Y. (2006). A phrase-level machine translation approach for disfluency detection using weighted finite state transducers. In *Proc. Interspeech 2006*, pages paper 1886–Tue1A1O.2. DOI: 10.21437/Interspeech.2006-262.

Medeiros, H., Moniz, H., Batista, F., Trancoso, I., and Nunes, L. (2013). Disfluency detection based on prosodic features for university lectures. In *Proc. Interspeech 2013*, pages 2629–2633. DOI: 10.21437/Interspeech.2013-605.

Meinedo, H., Caseiro, D., Neto, J., and Trancoso, I. (2003).

Audimus.(media): A broadcast news speech recognition system for the european portuguese language. pages 9–17. DOI: 10.1007/3-540-45011-4$_2$.

Miller, T. (2009). Word buffering models for improved speech repair parsing. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2 - Volume 2*, EMNLP '09, page 737–745, USA. Association for Computational Linguistics. DOI: 10.3115/1699571.1699609.

Miller, T. and Schuler, W. (2008). A syntactic time-series model for parsing fluent and disfluent speech. In *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*, COLING '08, page 569–576, USA. Association for Computational Linguistics. DOI: 10.3115/1599081.1599153.

Nissim, M., Abzianidze, L., Evang, K., van der Goot, R., Haagsma, H., Plank, B., and Wieling, M. (2017). Sharing is caring: The future of shared tasks. *Comput. Linguist.*, 43(4):897–904. DOI: 10.1162/COLI$_{a0}$0304.

Oostdijk, N. (2002). The design of the spoken dutch corpus. In *New frontiers of corpus research*, pages 105–112. Brill. DOI: 10.1163/9789004334113$_0$08.

Ostendorf, M. and Hahn, S. (2013). A sequential repetition model for improved disfluency detection. In *Proc. Interspeech 2013*, pages 2624–2628. DOI: 10.21437/Interspeech.2013-604.

Oyez (2022). Oyez. Available at: https://www.oyez.org/.

Qian, X. and Liu, Y. (2013). 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*, pages 820–825, Atlanta, Georgia. Association for Computational Linguistics. Available at: https://aclanthology.org/N13-1102/.

Rasooli, M. S. and Tetreault, J. (2013). Joint parsing and disfluency detection in linear time. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 124–129, Seattle, Washington, USA. Association for Computational Linguistics. Available at: https://www.aclweb.org/anthology/D13-1013.

Rasooli, M. S. and Tetreault, J. (2014). Non-monotonic parsing of fluent umm i mean disfluent sentences. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pages 48–53, Gothenburg, Sweden. Association for Computational Linguistics. Available at: https://www.aclweb.org/anthology/E14-4010.

Ribeiro, M. A. O. and Nunes, F. L. S. (2022). Left ventricle segmentation in cardiac MR: A systematic mapping of the last decade. *ACM Comput. Surv.* Just Accepted. DOI: 10.1145/3517190.

Rocholl, J. C., Zayats, V., Walker, D. D., Murad, N. B., Schneider, A., and Liebling, D. J. (2021). Disfluency detection with unlabeled data and small BERT models. In *Proc. Interspeech 2021*, pages 766–770. DOI: 10.48550/arXiv.2104.10769.

Rohanian, M. and Hough, J. (2020). Re-framing incremental deep language models for dialogue processing with multi-task learning. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 497–507, Barcelona, Spain (Online). International Committee on Computational Linguistics. DOI: 10.18653/v1/2020.coling-main.43.

S.-C., T. (2004). Processing spoken mandarin corpora. *Traitement Automatique des Langues*, 45(2):89–108. available at: https://www.isca-archive.org/iscslp_2004/tseng04_iscslp.html.

SATA (2024). SATA segmentation challenge. Available at: https://www.cardiacatlas.org/lv-segmentation-challenge/. Last accessed 05 December 2024.

Schettino, L., Maffia, M., De Micco, R., and Tessitore, A. (2023). Disfluency and speech management in italian patients with early-stage parkinson's disease. In *Disfluency in Spontaneous Speech (DiSS) Workshop 2023*, pages 23–27. DOI: 10.21437/DiSS.2023-5.

Shahih, K. M. and Purwarianti, A. (2016). Utterance disfluency handling in indonesian-english machine translation. In *2016 International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA)*, pages 1–5. DOI: 10.1109/ICAICTA.2016.7803104.

Shriberg, E. E. (1994). *Preliminaries to a theory of speech disfluencies*. PhD thesis, University of California, Berkeley. Thesis.

SIGEDU (2024). SIGEDU - special interest group on building educational applications. Available at: https://sig-edu.org/sharedtasks. Last accessed 05 December 2024.

Snover, M., Dorr, B., and Schwartz, R. (2004). A lexically-driven algorithm for disfluency detection. In *Proceedings of HLT-NAACL 2004: Short Papers*, HLT-NAACL-Short '04, page 157–160, USA. Association for Computational Linguistics. DOI: 10.3115/1613984.1614024.

Stallard, D., Prasad, R., Natarajan, P., Choi, F., Saleem, S., Meermeier, R., Krstovski, K., Ananthakrishnan, S., and Devlin, J. (2011). *The BBN TransTalk Speech-to-Speech Translation System*. DOI: 10.5772/19405.

Stolcke, A. and Shriberg, E. (1996). Statistical language modeling for speech disfluencies. In *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, volume 1, pages 405–408 vol. 1. DOI: 10.1109/ICASSP.1996.541118.

Stouten, F. and Martens, J. (2003). A feature-based filled pause detection system for dutch. In *2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No.03EX721)*, pages 309–314. DOI: 10.1109/ASRU.2003.1318459.

Stouten, F. and Martens, J.-P. (2004). Coping with disfluencies in spontaneous speech recognition. In *Proc. Interspeech 2004*, pages 1513–1516. DOI: 10.21437/Interspeech.2004-570.

Strassel, S. (2004). Linguistic resources for effective, affordable, reusable speech-to-text. In *LREC*. Available at: http://www.lrec-conf.org/proceedings/lrec2004/pdf/762.pdf.

Tack, A., Kochmar, E., Yuan, Z., Bibauw, S., and Piech,

C. (2023). The BEA 2023 shared task on generating AI teacher responses in educational dialogues. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 785–795, Toronto, Canada. Association for Computational Linguistics. DOI: 10.18653/v1/2023.bea-1.64.

Tanaka, T., Masumura, R., Moriya, T., Oba, T., and Aono, Y. (2019). 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)*, pages 1009–1013. DOI: 10.1109/APSIPAASC47483.2019.9023119.

Taskar, B., Guestrin, C., and Koller, D. (2004). Max-margin markov networks. In Thrun, S., Saul, L., and Schölkopf, B., editors, *Advances in Neural Information Processing Systems*, volume 16. MIT Press. Available at: `https://proceedings.neurips.cc/paper_files/paper/2003/file/878d5691c824ee2aaf770f7d36c151d6-Paper.pdf`.

Teleki, M., Dong, X., Kim, S., and Caverlee, J. (2024). Comparing ASR systems in the context of speech disfluencies. In *Interspeech 2024*, pages 4548–4552. DOI: 10.21437/Interspeech.2024-1270.

Trancoso, I., Martins, R., Moniz, H., Mata, A., and Viana, C. (2008). The LECTRA corpus - classroom lecture transcriptions in european portuguese. Available at: `http://www.lrec-conf.org/proceedings/lrec2008/pdf/359_paper.pdf`.

Tsvetkov, Y., Sheikh, Z., and Metze, F. (2013). Identification and modeling of word fragments in spontaneous speech. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7624–7628. DOI: 10.1109/ICASSP.2013.6639146.

Volodina, E., Bryant, C., Caines, A., De Clercq, O., Frey, J.-C., Ershova, E., Rosen, A., and Vinogradova, O. (2023). MultiGED-2023 shared task at NLP4CALL: Multilingual grammatical error detection. In Alfter, D., Volodina, E., François, T., Jönsson, A., and Rennes, E., editors, *Proceedings of the 12th Workshop on NLP for Computer Assisted Language Learning*, pages 1–16, Tórshavn, Faroe Islands. LiU Electronic Press. Available at: `https://aclanthology.org/2023.nlp4call-1.1`.

Walker, C. *et al.* (2005). RT-04 MDE training data text/annotations ldc2005t24. DOI: 10.35111/qwyc-cw15.

Wang, F., Chen, W., Yang, Z., Dong, Q., Xu, S., and Xu, B. (2018). Semi-supervised disfluency detection. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3529–3538, Santa Fe, New Mexico, USA. Association for Computational Linguistics. Available at: `https://aclanthology.org/C18-1299/`.

Wang, S., Che, W., and Liu, T. (2016). A neural attention model for disfluency detection. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 278–287, Osaka, Japan. The COLING 2016 Organizing Committee. Available at: `https://aclanthology.org/C16-1027/`.

Wang, S., Che, W., Zhang, Y., Zhang, M., and Liu, T. (2017). Transition-based disfluency detection using lstms. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2785–2794, Copenhagen, Denmark. Association for Computational Linguistics. DOI: 10.18653/v1/D17-1296.

Wang, W., Stolcke, A., Yuan, J., and Liberman, M. (2013). A cross-language study on automatic speech disfluency detection. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 703–708, Atlanta, Georgia. Association for Computational Linguistics. Available at: `https://aclanthology.org/N13-1083/`.

Wang, W., Tur, G., Zheng, J., and Ayan, N. F. (2010). Automatic disfluency removal for improving spoken language translation. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5214–5217. DOI: 10.1109/ICASSP.2010.5494999.

Wang, X., Ng, H. T., and Sim, K. C. (2014a). A beam-search decoder for disfluency detection. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1457–1467, Dublin, Ireland. Dublin City University and Association for Computational Linguistics. Available at: `https://aclanthology.org/C14-1138`.

Wang, X., Sim, K. C., and Ng, H. T. (2014b). Combining punctuation and disfluency prediction: An empirical study. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 121–130, Doha, Qatar. Association for Computational Linguistics. DOI: 10.3115/v1/D14-1013.

Williams, S., Lancaster, C., and Tanner, C. (2023). Inhibitory control and the production of disfluencies in speakers with alzheimer's disease. In *Disfluency in Spontaneous Speech (DiSS) Workshop 2023*, pages 18–22. DOI: 10.21437/DiSS.2023-4.

Womack, K., McCoy, W., Alm, C. O., Calvelli, C., Pelz, J. B., Shi, P., and Haake, A. (2012). Disfluencies as extra-propositional indicators of cognitive processing. In *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics*, ExProM '12, page 1–9, USA. Association for Computational Linguistics. Available at: https://aclanthology.org/W12-3801/.

Wu, C.-H. and Yan, G.-L. (2005). Speech act modeling and verification of spontaneous speech with disfluency in a spoken dialogue system. *IEEE Transactions on Speech and Audio Processing*, 13(3):330–344. DOI: 10.1109/TSA.2005.845820.

Wu, S., Zhang, D., Zhou, M., and Zhao, T. (2015). Efficient disfluency detection with transition-based parsing. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 495–503, Beijing, China. Association for Computational Linguistics. DOI: 10.3115/v1/P15-1048.

Yang, J., Yang, D., and Ma, Z. (2020). Planning and generating natural and diverse disfluent texts as augmentation for disfluency detection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1450–1460, Online. Association for Computational Linguistics. DOI: 10.18653/v1/2020.emnlp-main.113.

Yeh, J.-F. and Wu, C.-H. (2006). Edit disfluency detection and correction using a cleanup language model and an alignment model. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(5):1574–1583. DOI: 10.1109/TASL.2006.878267.

Yeh, J.-F., Wu, C.-H., and Wu, W.-Y. (2007). Disfluency correction of spontaneous speech using conditional random fields with variable-length features. In *Proc. Interspeech 2007*, pages 2157–2160. DOI: 10.21437/Interspeech.2007-582.

Yildirim, S. and Narayanan, S. (2009). Automatic detection of disfluency boundaries in spontaneous speech of children using audio–visual information. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(1):2–12. DOI: 10.1109/TASL.2008.2006728.

Yoshikawa, M., Shindo, H., and Matsumoto, Y. (2016). Joint transition-based dependency parsing and disfluency detection for automatic speech recognition texts. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1036–1041, Austin, Texas. Association for Computational Linguistics. DOI: 10.18653/v1/D16-1109.

Zayats, V. and Ostendorf, M. (2019). Giving attention to the unexpected: Using prosody innovations in disfluency detection. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 86–95, Minneapolis, Minnesota. Association for Computational Linguistics. DOI: 10.18653/v1/N19-1008.

Zayats, V., Ostendorf, M., and Hajishirzi, H. (2014). Multi-domain disfluency and repair detection. In *Proc. Interspeech 2014*, pages 2907–2911. DOI: 10.21437/Interspeech.2014-603.

Zayats, V., Ostendorf, M., and Hajishirzi, H. (2016). Disfluency detection using a bidirectional LSTM. In *Proc. Interspeech 2016*, pages 2523–2527. DOI: 10.21437/Interspeech.2016-1247.

Zechner, K. (2001). Automatic generation of concise summaries of spoken dialogues in unrestricted domains. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, page 199–207, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/383952.383989.

Zechner, K. (2002). Automatic summarization of open-domain multiparty dialogues in diverse genres. *Comput. Linguist.*, 28(4):447–485. DOI: 10.1162/089120102762671945.

Zwarts, S. and Johnson, M. (2011). The impact of language models and loss functions on repair disfluency detection. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 703–711, Portland, Oregon, USA. Association for Computational Linguistics. DOI: 10.18653/v1/p17-2087.

Zwarts, S., Johnson, M., and Dale, R. (2010). Detecting speech repairs incrementally using a noisy channel approach. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, page 1371–1378, USA. Association for Computational Linguistics. DOI: 10.3115/1218955.1218960.

Table 5. Summary of approved articles

#	Reference	Goal	Dataset	Language	Transcribes	Disfluencies	Technique	Evaluation
1	Shahih and Purwarianti [2016]	Translation	Sani, A. U. F. (2013)	Indonesian	No	Filler; Discourse Marker; Repetition; Repair; Stuttering	CRFs	PREC, RECALL, F-SCORE
2	Wang *et al*. [2010]	Removal	GALE Ontonotes	Mandarin	No	Filler; Edition; Interruption Point	HMM-Heuristic Rules; CRFs; Heuristic Rules	TER score; NIST Error Rate
3	Jamshid Lou *et al*. [2019]	Detection	Switchboard	English	No	Edition	Multilayer Perceptrons	PREC, RECALL, F-SCORE
4	Honnibal and Johnson [2014]	Detection	Switchboard	English	No	Repair	Transition-based Dependency Parsing	PREC, RECALL, F-SCORE, Unlabelled Attachment Score (UAS), Labelled Attachment Score (LAS)
5	Yildirim and Narayanan [2009]	Detection	Internal Data	English	No	Repetition; Repair; False Start; Filler	Internal Method	ACC, PREC, RECALL, F-SCORE
6	Lin and Lee [2009]	Detection	MCDC	Mandarin	No	Edition	Decision Tree-MAXENT	IP error rate, Edit error rate
7	Deng *et al*. [2020]	Evaluate fluency	CSJ	Japanese	No	Filler; Fragments	SVM	RECALL, F-SCORE
8	Tsvetkov *et al*. [2013]	Detection	IARPA Babel project	Cantonese	No	Fragments	SVM	PREC, WER, Character Error Rate (CER)
9	Honal and Schultz [2005]	Removal	Verbmobil, MCC	English, Mandarin	NI	Discourse Marker; Repetition; False Start	Noisy Channel Model	RECALL, PREC, F-SCORE
10	Yeh and Wu [2006]	Removal	MCDC, MAT	Mandarin	HTK	Edition	Logarithmic linear model	PREC, RECALL, Edit error rate
11	Tanaka *et al*. [2019]	Detection	CSJ	Japanese	No	Filler; Fragments	BLSTM-CRFs	PREC, RECALL, F-SCORE
12	Wu and Yan [2005]	Modeling	Internal Data	English	NI	Repetition; Filler	HMM	Fragment Correct Rate (FCR); Speech Act Correct Rate (SACR); False Acceptance Rate (FAR); False Rejection Rate (FRR)
13	Chen *et al*. [2020]	Detection	Internal Data	Chinese	No	Reparandum; Filler	CT-Transformer	PREC, RECALL, F-SCORE
14	Stolcke and Shriberg [1996]	Prediction	Switchboard	English	No	Filler; Repetition; False Start	N-gram	WER
15	Liu *et al*. [2006]	Detection	RT-04 MDE	English	No	Edition; Interruption Point; Filler; Discourse Marker	HMM; MAXENT; CRFs	WER; NIST Error Rate
16	Lease *et al*. [2006]	Detection	Switchboard, Fisher	English	NI	Repair; Filler; Interruption Point	TAG	Edit error rate
17	Gupta and Bangalore [2003]	Removal	Internal Data	English	No	Filler; Discourse Marker; Edition	Boosting	PREC, RECALL, F-SCORE
18	Stouten and Martens [2003]	Detection	Spoken Dutch Corpus	Dutch	No	Filler	Internal Method	PREC, RECALL, ROC curve
19	Dufour *et al*. [2009]	Spontaneous speech	French Broadcast News	French	LIUM ASR	Filled pause; Repetition	Max. likelihood	PREC, RECALL, WER, NCE
20	Zechner [2002]	Detection + Removal	Switchboard, CallHome, CallFriend, NewsHour, CrossFire, GroupMeeting	English	No	Filled Pause; Repair; False Start	POS; Decision Tree; Internal Method	PREC, RECALL, F-SCORE
21	Johnson and Charniak [2004]	Detection + Removal	Switchboard	English	No	Repair	TAG	PREC, RECALL, F-SCORE
22	Zechner [2001]	Detection + Removal	Switchboard	English	No	Filler; Repair; False Start	POS; Decision Tree; Internal Method	F-SCORE, PREC, RECALL
23	Georgila *et al*. [2010]	Detection	Rapport corpus, Switchboard	English	No	Repetition; Repair	CRFs; CRFs-ILP	PREC, RECALL, F-SCORE, NIST Error Rate
24	Lendvai [2003]	Detection	Spoken Dutch Corpus	Dutch	No	Fragments	TiMBL 4.3, RIPPER (Cohen, 1995)	ACC, PREC, RECALL, F-SCORE
25	Chen and Yoon [2011]	Detection	TOEFL Practice Online Test	English	No	Interruption Point	MAXENT; CRFs	ACC, PREC, RECALL, F-SCORE
26	Gupta and Bangalore [2002]	Detection	Switchboard	English	No	Repair	N-gram; Boosting	RECALL, PREC
27	Georgila [2009]	Detection	Switchboard	English	No	Repetition; Repair	HELM; MAXENT; CRFs; HELM-ILP; MAXENT-ILP; CRFs-ILP	F-SCORE, NIST Error Rate
28	Zwarts *et al*. [2010]	Detection + Removal	Switchboard	English	No	Repair	Noisy Channel Model	Responsiveness Measure
29	Liu [2003]	Detection	Switchboard	English	No	Fragments	CART	PREC, RECALL, ACC
30	Miller and Schuler [2008]	Syntax Analisys	Switchboard	English	No	Repair	HHMM	F-SCORE; edit-finding
31	Snover *et al*. [2004]	Detection	Switchboard	English	No	Edition; Filler	Transformation-Based Learning	LER
32	Miller [2009]	Syntax Analisys	Switchboard	English	No	Repair	HHMM	PREC, RECALL, F-SCORE
33	Fitzgerald *et al*. [2009]	Detection	SSR	English	No	Filler; Edition	CRFs	PREC, RECALL, F-SCORE
34	Lease and Johnson [2006]	Removal	Switchboard	English	No	Filler	TAG	F-SCORE
35	Womack *et al*. [2012]	Prediction	Internal Data	English	No	Filler	Decision Tree	ACC
36	Wang *et al*. [2013]	Detection	RT-04 MDE	Mandarin, English	No	Filler; Edition; Interruption Point	CRFs	NIST Error Rate
37	Yang *et al*. [2020]	Insertion	Switchboard	English	No	Insertion; Repetition	BERT	PREC, RECALL, F-SCORE
38	Wu *et al*. [2015]	Detection	Switchboard, Internal Data	English, Chinese	No	Repair	R2L parsing	PREC, RECALL, F-SCORE, Unlabelled Attachment Score (UAS), Labelled Attachment Score (LAS)
39	Rasooli and Tetreault [2013]	Detection	Switchboard	English	No	Reparandum; Discourse Marker; Filler	Arc-Eager	PREC, RECALL, F-SCORE
40	Jamshid Lou *et al*. [2018]	Detection	Switchboard	English	No	Repair	ACNN; CNN	PREC, RECALL, F-SCORE
41	Wang *et al*. [2017]	Detection	Switchboard, Internal Data	English, Chinese	iflyrec	Reparandum	LSTM; BLSTM	PREC, RECALL, F-SCORE
42	Jamshid Lou and Johnson [2017]	Detection	Switchboard	English	No	Repair	LSTM-Noisy Channel Model	F-SCORE; error rate
43	Wang *et al*. [2016]	Detection	Switchboard, Internal Data	English, Chinese	No	Repair	BLSTM	PREC, RECALL, F-SCORE
44	Yoshikawa *et al*. [2016]	Detection	Switchboard	English	Kaldi	Repair; Discourse Marker; Filler	Joint Dependency Parsing	PREC, RECALL, F-SCORE
45	Rohanian and Hough [2020]	Detection	Switchboard	English	No	Repair; Edition	LSTM-CRFs	F-SCORE
46	Zwarts and Johnson [2011]	Detection	Switchboard	English	No	Repair	Noisy Channel Model	F-SCORE
47	Qian and Liu [2013]	Detection	Switchboard	English	No	Edition	M3N	F-SCORE
48	Zayats and Ostendorf [2019]	Detection	Switchboard	English	No	Reparandum; Repair	LSTM-CRFs	F-SCORE
49	Wang *et al*. [2014a]	Detection	Switchboard	English	No	Edition	M3N	F-SCORE

50	Ferguson *et al.* [2015]	Detection	Switchboard	English	No	Reparandum; Repair	Prosody-Semi-CRFs	PREC, RECALL, F-SCORE
51	Wang *et al.* [2018]	Detection	Switchboard	English	No	Reparandum; Filler	Semi-supervised model	PREC, RECALL, F-SCORE
52	Wang *et al.* [2014b]	Prediction	Switchboard	English	No	Edition; Filler	M3N	F-SCORE
53	Jamshid Lou and Johnson [2020]	Detection	Switchboard, Fisher	English	No	Edition	Multilayer Perceptrons	PREC, RECALL, F-SCORE
54	Rasooli and Tetreault [2014]	Detection	Switchboard	English	No	Reparandum	Arc-Eager	PREC, RECALL, F-SCORE
55	Liu *et al.* [2004]	Detection	Switchboard	English	No	Interruption Point	Prosody model; HMM; Prosody-HMM	Classification Error Rate
56	Stouten and Martens [2004]	Detection	Spoken Dutch Corpus	Dutch	Demuynck *et al.* [2000]; Duchateau *et al.* [1998]	Filler; Repetition	Adaptive search technique	WER
57	Lin and Lee [2005]	Detection	MCDC	Mandarin	No	Interruption Point	Decision Tree; MAXENT; Decision Tree-MAXENT	ACC
58	Liu *et al.* [2005]	Detection	RT-04 MDE	English	NI	Edition; Interruption Point	HMM; MAXENT; CRFs	NIST Error Rate, WER
59	Maskey *et al.* [2006]	Detection	Switchboard	English	No	Repetition; Repair; Filler	Fast phrase-based statistical translation	PREC, RECALL, F-SCORE
60	Lin and shan Lee [2006]	Detection	MCDC	Mandarin	No	Interruption Point	Latent Prosodic Modeling	ACC
61	Yeh *et al.* [2007]	Removal	MCDC	Mandarin	HTK	Edition	CRFs	Loss rate, false alarm rate and error rate
62	Germesin *et al.* [2008]	Detection	AMI Meeting Corpus	English	No	Filler; Stuttering; False Start; Tongue slip; Discourse Marker; Edition; Insertion; Repetition; Repair; Erro; Order; Omission	Internal Method; Decision Tree; N-gram	F-SCORE; ACC
63	Ostendorf and Hahn [2013]	Detection	Switchboard, SCOTUS	English	No	Reparandum; Repetition	CRFs	F-SCORE
64	Medeiros *et al.* [2013]	Detection	LECTRA	European Portuguese	Audimus	Interruption Point; Filler; Repair	CART	PREC, RECALL, F-SCORE, NIST Error Rate
65	Hassan *et al.* [2014]	Removal	Switchboard, Fisher	English; Spanish	Microsoft Research S2S	Filler; Discourse Marker; Repair; Repetition	CRFs	F-SCORE
66	Gupta *et al.* [2014]	Detection	DARPA TransTac	English	BBN Byblos	Repetition; Repair	Internal Method	WER, ROC curve
67	Dutrey *et al.* [2014]	Detection	Voxfactory	French	No	Edition	CRFs	PREC, RECALL, F-SCORE, Slot Error Rate (SER)
68	Zayats *et al.* [2014]	Detection	Switchboard, CallHome, SCOTUS, FCIC	English	No	Repair	CRFs	F-SCORE
69	Christodoulides and Avanzi [2015]	Detection	Phonologie du Français Contemporain	French	No	Filler; Repetition; Interruption Point; Edition	SVM; CRFs; SVM-CRFs	PREC, RECALL, F-SCORE
70	Bertero *et al.* [2015]	Detection	SSR	English	No	Filler; Repetition; False Start	CRFs; Multilayer Perceptrons	PREC, RECALL, F-SCORE
71	Hough and Schlangen [2015]	Detection	Switchboard	English	No	Edition; Repair	RNN	Precision metrics; Time metrics; Diachronic metrics
72	Cho *et al.* [2015]	Detection	Internal Data	English	No	Filler; Repetition; Interruption Point	CRFs; Multilayer Perceptrons; CRFs-Multilayer Perceptrons	PREC, RECALL, F-SCORE
73	Zayats *et al.* [2016]	Detection	Switchboard	English	No	Edition	BLSTM	PREC, RECALL, F-SCORE
74	Bach and Huang [2019]	Detection	Switchboard, CallHome, SCOTUS, FCIC, interviews	English	No	Insertion; False Start; Repetition	BLSTM	PREC, RECALL, F-SCORE
75	Lin and Wang [2020]	Prediction	Switchboard, English IWSLT dataset	English	No	Filler; Edition	BERT	PREC, RECALL, F-SCORE
76	Rocholl *et al.* [2021]	Detection	Switchboard, Fisher, Pushshift Reddit	English	No	Reparandum; Filler	BERT	PREC, RECALL, F-SCORE
77	Lee *et al.* [2021]	Detection	Switchboard	English	No	Reparandum	Transformer-CRFs; BERT-CRFs; ELECTRA-CRFs	PREC, RECALL, F-SCORE
78	Dao *et al.* [2022]	Detection	PhoATIS	Vietnamese	No	Reparandum; Filler	BERT-CRFs	F-SCORE
79	Ghosh *et al.* [2022]	Detection	Switchboard	English	No	Reparandum	BERT-CRFs; ELECTRA-CRFs	PREC, RECALL; F-SCORE
80	Horii *et al.* [2022]	Detection	CSJ	Japanese	No	Filler	Internal Method	Character Error Rate (CER); Sentence Error Rate (SER)