Speech Disfluency, Repetition and Reduplication: A Survey

Arif Ahmad, Pushpak Bhattacharyya

CFILT, Indian Institute of Technology Bombay
190110010@iitb.ac.in, pb@cse.iitb.ac.in

Abstract

This survey presents a comprehensive analysis of speech disfluency, repetition, and reduplication within the context of Automatic Speech Recognition (ASR) and natural language processing (NLP). The paper reviews various types of disfluencies, their implications for speech recognition systems, and methods to detect and correct them. We explore both the technological advances in handling disfluencies and the ongoing challenges posed by the diverse nature of human speech. Additionally, the survey discusses the impact of linguistic diversity on ASR performance, highlighting the need for more inclusive and adaptable systems to ensure equitable technology access across different languages and dialects.

1 Introduction

The rapid evolution of Automatic Speech Recognition (ASR) technologies, propelled by advancements in deep learning and robust computational platforms, has significantly enhanced the quality of speech-to-text translation. This progress has broadened the scope of ASR applications, encompassing voice-controlled systems, personal assistants, and automated transcription services, yielding impressive results particularly in well-resourced languages such as English and Chinese (Amodei et al., 2016). Despite these advancements, ensuring equitable performance across diverse demographic groups remains a substantial challenge. Studies have consistently shown that ASR systems perform variably based on the user's dialect, gender, and ethnicity, often disadvantaging non-white speakers and those speaking English as a second language (Wheatley and Picone, 1991; Meyer et al., 2020; Koenecke et al., 2020). This disparity raises significant ethical concerns about linguistic justice and representation within ASR technologies (Blodgett et al., 2020; Hovy and Prabhumoye, 2021; Markl and McNulty, 2022).

Moreover, ASR systems' performance in low-resource languages presents additional complexities. These languages often lack substantial transcribed speech and parallel text data, crucial for training effective models. In such scenarios, end-to-end speech-to-text systems, which directly convert spoken language into another language's text, have shown potential despite the substantial resources still required for optimal performance (Weiss et al., 2017; Bansal et al., 2018). The gap between high-resource and low-resource languages underscores the necessity for innovative approaches, such as multi-task learning, to leverage limited data more effectively (Anastasopoulos and Chiang, 2018).

Additionally, the presence of speech disfluencies like filled pauses, repetitions, and corrections complicates the processing of spontaneous speech, posing significant challenges for downstream NLP applications that typically depend on clean, fluent speech data (Liu et al., 2006; Johnson and Charniak, 2004). These disfluencies, while often overlooked, are prevalent in everyday conversation and can severely impact the performance of dialogue systems and question-answering models trained on disfluency-free data (Rajpurkar et al., 2018; Raffel et al., 2020). Addressing these issues requires dedicated disfluency detection models that can refine ASR outputs for more accurate downstream processing (Zayats et al., 2016; Dong et al., 2019).

Despite the clear challenges posed by disfluencies and the demographic variability in speech, ASR systems continue to be integral in various applications, pushing the boundaries

of what's possible with speech recognition technologies. However, the journey towards truly inclusive and robust ASR systems is ongoing, demanding continued research into both the technological advancements and the socio-ethical implications of speech recognition technology.

This ongoing journey highlights the need for a deeper understanding of the underlying linguistic and acoustic features that differentiate speakers and dialects. For instance, linguistic studies have shown that phonological variations such as consonant voicing and vowel height have significant acoustic correlates, which can influence ASR accuracy (Koenecke et al., 2020; Wassink et al., 2022). Such insights are crucial for refining ASR models to handle sociolinguistic variation and L2 transfer effectively, thereby improving system robustness across different speaker profiles (Corder, 1983; Dechert and Raupach, 1989; Best et al., 1994).

In parallel, there is a growing recognition of the need to develop ASR and NLP technologies that can adapt to the natural, unscripted nature of human speech, which frequently includes disfluencies. These spontaneous elements, although challenging for speech recognition systems, are integral to human communication and cannot be ignored. Addressing them requires innovations in both system architecture and training methodologies. For instance, end-to-end models that integrate disfluency detection directly into the speech recognition process are showing promise. These models can leverage the paralinguistic features inherent in the speech signal, offering a more holistic approach to understanding and processing spoken language (Zayats and Ostendorf, 2019).

Furthermore, the quality of ASR is traditionally assessed by metrics such as Word Error Rate (WER), which, while useful, do not fully capture the nuances of speech recognition errors (Specia et al., 2013; Ogawa and Hori, 2017). Recent research has begun exploring more sophisticated error analysis techniques, such as detecting and classifying different types of ASR errors using sequence models. These approaches aim to provide a more detailed understanding of where and why ASR systems fail, leading to more targeted improvements (Seigel and Woodland, 2014; Ghannay et al., 2015; Tam et al., 2014).

The necessity for enhanced datasets cannot be overstated. Most existing speech datasets tend to focus on specific domains like telephone conversations or broadcast news, offering a limited view of the full range of human speech (Godfrey and Holliman, 1993; Zayats et al., 2014). A richer, more diverse collection of speech data, encompassing a wider array of disfluencies and speaker demographics, is essential for training and testing the next generation of ASR systems. Only then can we hope to achieve the level of linguistic inclusivity and accuracy required for ASR technologies to be genuinely useful across all sections of society. The development of such datasets would not only help in training more robust ASR systems but also enable a more comprehensive evaluation of these systems across various real-world scenarios.

As we move forward, the integration of NLP methodologies to handle disfluencies becomes increasingly crucial. Innovative approaches such as using neural language models to predict the likelihood of disfluencies based on contextual predictability are promising. These models can help elucidate the complex interactions between speech production processes and linguistic structures, potentially leading to breakthroughs in how speech systems handle spontaneous language.

Moreover, the incorporation of fairness and bias mitigation strategies in ASR development is essential. By addressing the disparities in ASR performance among different demographic groups, researchers and developers can ensure that speech technologies do not perpetuate or exacerbate social inequalities. This involves not only technical improvements but also a concerted effort to understand and integrate the ethical implications of deploying these technologies in diverse environments.

In conclusion, the field of ASR is at a pivotal juncture. The potential to create inclusive, accurate, and efficient speech recognition technologies is immense, but realizing this potential requires tackling the complex challenges of variability in speech, disfluencies, and demographic biases. By embracing a multidisciplinary approach that includes advancements in machine learning, linguistics, and

Language	Word (Meaning)	Reduplicated Word (Meaning)
Indonesian/Malay	orang (person)	orang-orang (people)
Tagalog	bili (buy)	bili-bili (to buy here and there)
Tamil	kaal (leg)	kaal-kaal (legs)
Punjabi	xushii (happy)	xushii-xushii (happily)
Mandarin Chinese	妈 (mā, mother)	妈妈 (māma, mommy)
Hawaiian	wiki (quick)	wiki-wiki (very quick)
Samoan	pili (cling)	pili-pili (to cling repeatedly)
Turkish	ev (house)	ev-ev (every house)
Basque	txiki (small)	txiki-txiki (very small)

Table 1: Examples of Morphological Reduplication in Various Languages Demonstrating Pluralization, Intensification, and Other Grammatical or Semantic Changes

ethics, the next generation of ASR technologies can truly meet the needs of all users, irrespective of their language, dialect, or sociocultural background. This will not only enhance the usability of ASR systems but also their acceptance and reliability in everyday applications, making them indispensable tools in our increasingly interconnected world.

2 Reduplication as a Multiword Expression

In the domain of natural language processing (NLP), multiword expressions (MWEs) are identified for their complex and often idiomatic nature, which presents unique challenges in linguistic analysis. Effective handling of MWEs within NLP systems is pivotal, demanding sophisticated computational strategies to unravel their syntactic and semantic complexities intricacies (Baldwin and Kim, 2010; Sag et al., 2002).

Reduplication serves as a quintessential example of MWEs and has been studied across various languages including Bengali, Cantonese, and Mandarin Chinese, showcasing its ability to express grammatical or semantic nuances (Chakraborty and Bandyopadhyay, 2010; Lam, 2013; Chen et al., 1992). These studies highlight the global relevance of reduplication, with substantial progress in modeling these phenomena through the use of two-way finite-state transducers (2-way FSTs), which adeptly capture the processes involved in reduplication across languages (Dolatian and Heinz, 2018). Furthermore, the advent of finite-state buffered machines (FSBMs) has refined our computational understanding of reduplication (Wang, 2021).

The practical ramifications of these advancements are evident in their incorporation into machine translation systems, where they significantly enhance translation accuracy (Doren Singh and Bandyopadhyay, 2011). Moreover, the creation of the RedTyp database marks a significant stride in providing a structured resource for reduplication studies, facilitating a deeper understanding of this linguistic feature across different languages (Dolatian and Heinz, 2019). Despite these technological advances, there is a persistent need for more comprehensive datasets specifically designed for reduplication research, alongside the application of modern NLP techniques to better process these complex structures.

In exploring the concept of iconicity within reduplication, various scholars propose that the form-meaning pairings in languages exhibit different types of iconicity, which may interact with each other (Peirce, 1955; Regjer, 2022; Rozhanskiy, 2015). Lǐ and Ponsford (2018) further this discussion by identifying five features that characterize the form of fully reduplicated words, which correlate iconically with certain semantic aspects marked by total reduplication. They categorize these into five dimensions of iconicity: identity, magnitude, discreteness, proximity, and sequentiality, each reflecting a different aspect of how form and meaning align in reduplication. For instance, in Balinese, the use of pluractional markers illustrates how different types of pluractionality are marked either by reduplication or non-reduplicative affixation, reveal-

ing an underlying iconic relationship between the form of expression and its semantic intent (Arka and Dalrymple, 2017).

For many languages pluractionality is marked distinctly through either reduplication or affixation, not based on repetition in meaning but rather on the complexity of the expression (Conathan and Wood, 2003). This analysis serves as a proof-of-concept for applying model-theoretic approaches to linguistic structure, demonstrating how the complexities in linguistic forms and meanings can be effectively modeled and understood. The theoretical discussions provide a solid foundation for a computational representation of pluractionality and reduplication, enriching our understanding of these linguistic phenomena and their applications in modern NLP systems.

3 Disfluency

Speech utterances can be broadly categorized into read speech and conversational speech. "Read speech" refers to situations where content is delivered verbatim from a written text. In contrast, conversational speech is characterized by its spontaneity, with speakers formulating and articulating thoughts in real-time. This type of speech often includes various irregularities, known as disfluencies, which do not typically contribute meaningful content to the discourse. Common types of disfluencies include fillers, phrase repetitions, abrupt topic shifts, and self-corrections.

Consider the following example where disfluencies are emphasized in italicized and bold text:

So, it was ***like***, ***um*** I was trying to explain my point, and then ***uh***, ***you know***, I totally lost my train of thought.

While often perceived as disruptions or irregularities, disfluencies frequently occur in everyday speech and are generally overlooked in casual interactions. However, they have garnered considerable attention in computational linguistics due to their prevalence. A seminal study from 1994 using the Switchboard corpus highlighted the regular occurrence of disfluencies in conversational speech. The study found that sentences containing 10-13 words had a 50% chance of including a disfluency, with this likelihood increasing with sentence length. This correlation emphasizes how disfluencies can complicate sentence structure, diminish semantic clarity, and degrade the overall fluidity of speech.

These effects pose significant challenges for downstream applications such as machine translation, which rely on accurate and clear interpretation of spoken content. The presence of disfluencies can lead to reduced effectiveness in these technologies, highlighting the importance of understanding and managing disfluencies in computational systems.

4 Types of Disfluencies

In our discussion of disfluencies, we focus on those that occur within single sentences rather than spanning multiple sentences, with variations in annotation across different corpora. Following the classifications described by Honal and Schultz (2003), disfluencies range from simple to complex types. This section outlines these types, providing a structured overview and examples for each.

The simplest forms of disfluencies are filled pauses and discourse markers. Filled pauses, such as "uh", "um", and "ah", serve no semantic purpose but act as placeholders in speech. Similarly, discourse markers like "well", "yeah", "alright", "you know", and "okay" structure discourse but add no semantic content. They are essential in managing turns during conversation or acknowledging previous statements. Often, words like "yeah" and "okay" are also categorized as filled pauses due to their function in maintaining the flow of speech rather than conveying concrete information.

More intricate disfluencies include interjections, repetitions or corrections, false starts, and edits. Interjections like "oops", "ugh", "uh-huh" convey spontaneous reactions or emotions and are typically non-lexical but impactful. Repetitions or corrections occur when a speaker repeats or slightly alters a phrase without changing its syntactical structure significantly. These adjustments generally maintain the original train of thought.

False starts are characterized by the abandonment of a phrase, followed by the initiation of a new phrase with a different syntactic structure and semantic content, reflecting

Figure 1: Examples showing the four regions of any disfluency: Redarandum, Interruption Point, Interregnum, and Repair. Not all parts are necessary to be present in every example of a disfluency; as can be seen in Example (b) in the Figure, with no interregnum.

a shift in the speaker's thought process. Edits are corrective phrases that explicitly refer to previous parts of the discourse to indicate a deviation from the intended message, often leading to the correction or abandonment of the initial utterance.

4.1 Classification and Examples

Each type of disfluency serves a distinct function in speech, as summarized in the table below:

Understanding these types is crucial for enhancing the performance of speech recognition and natural language processing systems, as it allows for more accurate modeling of natural language patterns and improves the handling of spontaneous speech in computational applications.

5 Structure of Disfluency

Shriberg (1994) introduces the disfluency structure, which comprises three key components: the Reparandum, Interregnum, and Repair. Although the Interruption Point, marking the moment of interruption, is integral to this structure, it is not typically included in transcripts as it is part of the acoustic signal rather than the textual representation. This omission is considered in our modeling strategy. Figure 1 shows an example with the structural components of a disfluency.

The Reparandum-Interregnum-Repair (RiR) structure forms the foundation of our classification approach, essential for identifying and differentiating patterns of reduplication and repetition:

- **Reparandum**: The original segment of speech that is subject to modification through either repetition or reduplication.
- **Interregnum**: An optional component that may include disfluent markers, playing a crucial role in distinguishing between repetition and reduplication.
- **Repair**: The segment where the initial speech (Reparandum) is repeated or reduplicated, often with variations.

5.1 Significance of the RiR Structure

The integration of the RiR structure into our classification methodology is instrumental in decoding complex linguistic phenomena, particularly useful when reduplication and repetition occur simultaneously within a single utterance. This is illustrated by the structured formula:

$$[\ reparandum\ +\ \{interregnum\}\ +\ repair\ ]$$

5.1.1 Examples

- **Example 2:**

  वह [नीला + { नहीं } नीला नीला] फूल है।
  It [blue + *no* + blue blue] flower is.

  Translation: "It is a blue, no blue-blue flower."

  In this instance, the interregnum "नहीं" (no) indicates a repetition between the first and second instances of नीला (blue), separated by a negation marker. This contrasts with the reduplication seen between the second and third नीला (blue).

- **Example 3:**

  वह [नीला + { } नीला] फूल है।
  It [blue + { } + blue] flower is.

  Translation: "It is a blue, blue flower."

  The lack of an interregnum in this example suggests a straightforward reduplication where the Reparandum is repeated without any intervening elements.

6 Multilingual Disfluency Studies

In this section we discuss several studies across different languages to explore the characteristics of disfluency in bilingual and multilingual settings, providing insights into how disfluency

Type	Description	Examples
Filled Pause	Non-lexicalized sounds with no semantic content.	”uh”, ”um”, ”ah”
Interjection	Non-lexicalized sounds indicating affirmation or negation.	”uh-huh”, ”ugh”, ”oops”
Discourse Marker	Words that assist in managing turns or serve as acknowledgments without adding semantic content.	”well”, ”you know”, ”okay”
Repetition or Correction	Repetition or slight modification of previously uttered words, maintaining the original idea.	”If I can’t don’t know the answer myself, I will find it.”
False Start	Abandonment of a phrase followed by a new phrase with different syntax and semantics.	”We’ll never find a day what about next month?”
Edit	Words following a disfluency that indicate the preceding words were unintended.	”We need two tickets, I’m sorry, three tickets for the flight to Boston.”

Table 2: Types of Disfluencies with descriptions and examples (Honal and Schultz, 2003)

manifests differently depending on linguistic and cultural contexts.

Al’Amri and Robb (2021) investigated the disfluency characteristics in Omani Arabic-English bilingual speakers, particularly focusing on bilinguals who stutter (BWS). The study compared the frequency of overall disfluencies and stuttering-like disfluencies (SLDs) across the two languages in both oral reading and conversational speech. It found equivalent levels of overall disfluencies in both languages during conversation, but a higher incidence of SLDs in Arabic during reading due to the linguistic complexity of formal Arabic. Carias and Ingram (2006) analyzed disfluency patterns in Spanish-English bilingual children, observing that the children exhibited more disfluencies in one language compared to the other. This variation correlated with the mean length of utterance, indicating that increased linguistic complexity could lead to higher disfluency rates. The type of disfluency also varied between the two languages, underscoring language-specific effects on speech patterns. Brundage and Rowe (2018) examined typical disfluency rates in Spanish–English simultaneous bilinguals, noting significant differences in disfluency rates between the two languages. Interestingly, typical disfluency rates were lower in bilingual children compared to monolinguals, suggesting unique linguistic processing in bilingual environments. The study also highlighted the influence of mean length of utterance and vocabulary diversity on disfluency rates, particularly in English.

Moving towards disfluency detection, Kundu et al. (2022) introduced a zero-shot disfluency detection model for Indian languages, utilizing a pretrained multilingual model fine-tuned on English disfluencies. The approach was validated by generating synthetic disfluent text in four Indian languages, demonstrating the potential of transfer learning in low-resource language settings. Bhat et al. (2023) presented DISCO, a large-scale human-annotated corpus for disfluency correction across four Indo-European languages. The study showcased the improvement in downstream language processing tasks when disfluencies were systematically removed, highlighting the critical role of disfluency correction in enhancing the quality of automated speech recognition outputs. Kochar et al. (2024) focused on creating annotated corpora for Indian languages, emphasizing

the need for a nuanced understanding of linguistic properties unique to these languages. The study proposed synthetic generation of disfluent data to improve model training for disfluency detection. Lastly, Dao et al. (2022) conducted the first empirical study on Vietnamese disfluency detection, creating a manually annotated dataset and testing various baseline models. The study found that language-specific word segmentation significantly enhances disfluency detection, with the best results achieved by fine-tuning a monolingual pre-trained model.

Cho et al. (2016) explored a multilingual approach to disfluency removal using neural machine translation (NMT), suggesting that a joint representation of disfluencies across languages could effectively address the data scarcity issue in disfluency annotation. Their multilingual NMT system demonstrated improved performance over single-language systems and enhanced outcomes in downstream applications.

Together, these studies illustrate the diverse manifestations of disfluency across languages and the effectiveness of multilingual approaches in understanding and mitigating disfluencies in speech processing.

7 Disfluency Detection

This section covers prior research in the area of disfluency detection. We explore various methodologies including sequence tagging, parsing-based models, and noisy channel approaches, which have significantly advanced our understanding and capabilities in this field.

7.1 Approaches to Disfluency Detection

We discuss four prominent ways of modelling the task of disfluency detection in the current literature.

7.1.1 Sequence Tagging Models

Sequence Tagging Models are prevalent in disfluency detection, leveraging a variety of machine learning techniques to classify words as fluent or disfluent. These models employ technologies such as Hidden Markov Models (HMM) (Liu et al., 2006), Conditional Random Fields (CRF) (Liu et al., 2006; Georgila et al., 2010; Ostendorf and Hahn, 2013; Zayats et al., 2014), Max-Margin Markov Networks (M3N) (Qian and Liu, 2013), Recurrent Neural Networks (RNN) (Hough and Schlangen, 2015), Bidirectional Long Short-Term Memory networks (Bi-LSTM) (Zayats et al., 2016, 2014), and Transformers (Wang et al., 2020). These models tag each word with labels indicating whether it is fluent or disfluent, or use the Begin-Inside-Outside (BIO) tagging format to denote the structure of disfluencies (Ferguson et al., 2015; Wang et al., 2018; Qian and Liu, 2013; Hough and Schlangen, 2015).

7.1.2 Parsing-based Models

Parsing-based Models integrate disfluency detection with syntactic parsing, focusing on the structural elements of disfluencies such as reparandum and filled pauses to analyze and interpret the syntactic structure of sentences. These models, such as those by Rasooli and Tetreault (Rasooli and Tetreault, 2013), Honnibal and Johnson (Honnibal and Johnson, 2014), Wu et al. (Wu et al., 2015), Yoshikawa et al. (Yoshikawa et al., 2016), and Jamshid Lou and Johnson (Jamshid Lou and Johnson, 2020) focus on detecting disfluencies along with identifying the syntactic structure of the sentence. In particular, Jamshid Lou and Johnson (2020b) (Jamshid Lou and Johnson, 2020) focus on joint disfluency detection and constituency parsing of transcriptions, providing insights into how these elements can be integrated for a more in-depth understanding of both the syntax and disfluencies present in speech.

7.1.3 Noisy Channel Models

Noisy Channel Models conceptualize disfluency as noise added to an otherwise fluent sentence. The objective of these models is to reconstruct the original fluent sentence from its noisy counterpart, making assumptions about the nature of the noise and the methods of its introduction (Johnson and Charniak, 2004; Zwarts and Johnson, 2011; Jamshid Lou and Johnson, 2017).

7.2 Motivation for Sequence Tagging

Drawing inspiration from the aforementioned studies, Ahmad et al. (2024)'s approach primarily focuses on sequence tagging based mod-

eling. This method offers direct and explicit tagging of disfluencies at the word level, which is crucial for fine-grained detection and classification. Such capabilities are indispensable for the development of robust speech recognition systems, allowing for precise control and correction of disfluent speech segments. The sequence tagging framework's ability to handle complex disfluency patterns effectively supports the decision to adopt this approach as the cornerstone in Ahmad et al. (2024)'s disfluency detection strategy.

8 Metrics for Disfluency Measurement

Evaluating disfluency detection models has predominantly relied on precision, recall, and F1 score at the token level. These metrics measure the accuracy of models in identifying and correcting disfluencies within speech or text data. However, some studies also consider the BLEU score to evaluate the fluency of generated sentences against fluent reference sentences.

8.1 Standard Evaluation Metrics

The majority of research in disfluency detection utilizes traditional metrics such as precision, recall, and F1 score to assess performance. These metrics are calculated based on the correct identification of disfluent versus fluent tokens and provide a balanced measure of model accuracy through the harmonic mean of precision and recall (Liu et al., 2006). Additionally, some studies report the NIST Error Rate, which includes errors from insertions, deletions, and substitutions, normalized by the length of the reference utterance (Georgila et al., 2010).

8.2 Metrics for Specific Contexts

Some recent studies have developed more nuanced metrics to better capture the effects of disfluency on downstream tasks like summarization and machine translation. For instance, ROUGE scores, which are traditionally used in summarization, have been adapted to measure how disfluencies affect the quality of generated summaries from spoken content (Teleki et al., 2024). This approach allows for a more detailed understanding of how disfluencies impact the performance of NLP models beyond simple error rates.

Mohapatra et al. (2022) address the challenge of disfluency detection in stuttering within speech, emphasizing the balance between model accuracy and the computational efficiency of using limited data. Georgila et al. (2010) compare the performance of Integer Linear Programming (ILP) models to Conditional Random Fields (CRFs) for disfluency detection, using metrics like F-score and NIST Error Rate to demonstrate the significant superiority of ILP, especially in contexts with limited domain-specific training data. This highlights the effectiveness of ILP in environments where data scarcity could otherwise compromise model performance.

8.3 Metrics for Integrated Systems

Lou and Johnson (2020) tackles end-to-end models for simultaneous speech recognition and disfluency removal, proposing new metrics to evaluate these integrated systems. They suggest that traditional metrics like BLEU and METEOR may not adequately reflect the nuances of disfluency removal as they are sensitive to sequence length and exact n-gram matches. Therefore, they introduce metrics specifically designed to assess the fluency of generated transcripts, aiming to provide a more targeted evaluation of disfluency detection within speech recognition systems. Chen et al. (2022) introduces a novel BERT-based sequence tagging model for real-time disfluency detection, focusing on metrics that evaluate both accuracy and operational efficiency in streaming contexts. This includes metrics like Time-to-Detection (TTD) and Edit Overhead (EO), which measure the latency and stability of predictions in dynamic, real-time environments. These metrics are crucial for applications requiring immediate feedback, such as interactive voice-responsive systems. Salesky et al. (2019) tackle the challenge of translating disfluent speech into fluent text, utilizing BLEU and METEOR scores to compare model outputs against clean, edited references. This study underscores the importance of both metrics for evaluating how well models maintain semantic integrity while effectively removing disfluencies, providing insights into the dual objectives of translation accuracy and disflu-

ency removal in speech translation applications.

Overall, the selection of metrics for evaluating disfluency detection models is guided by the specific challenges and goals of each study, ranging from improving summarization in the presence of spoken disfluencies to enhancing the responsiveness of real-time disfluency detection systems. These metrics not only assess the accuracy of disfluency identification but also the practical effectiveness of models in applications where real-time processing is critical.

9 Summary and Conclusions

This survey has extensively discussed the phenomenon of disfluency in speech, emphasizing its significance in both human communication and ASR technologies. Key points from the survey include:

- The identification and classification of different types of disfluencies and their common occurrences in natural speech.
- The challenges disfluencies pose to ASR systems, particularly in terms of accuracy and efficiency, and the various methodologies developed to detect and mitigate these disruptions.
- Innovations in modeling techniques, such as sequence tagging and parsing-based models, that have improved our ability to handle disfluencies.
- The influence of linguistic diversity on the performance of ASR systems, underscoring the importance of developing adaptive models capable of handling a variety of speech patterns and dialects.
- The necessity for integrating disfluency detection and correction into NLP tasks to enhance the performance of downstream applications such as machine translation and dialogue systems.

The review underscores that while substantial progress has been made in understanding and processing disfluencies, significant work remains to fully integrate these insights into ASR and NLP systems. Future research should focus on:

- Enhancing the robustness of ASR systems against disfluencies by employing more sophisticated machine learning models and larger, more diverse training datasets.
- Developing real-time disfluency detection models that can operate efficiently in live conversations, thus broadening the applicability of ASR technologies.
- Addressing the ethical dimensions of ASR technology, particularly ensuring that these systems do not perpetuate or exacerbate linguistic biases.

In conclusion, advancing our handling of speech disfluencies will not only improve the technical capabilities of ASR systems but also their usability and accessibility, making them more effective and equitable tools for global communication.

References

Arif Ahmad, Mothika Gayathri Khyathi, and Pushpak Bhattacharyya. 2024. Looks can be deceptive: Distinguishing repetition disfluency from reduplication. *arXiv preprint arXiv:2407.08147*.

Fathiya Al'Amri and Michael P Robb. 2021. Disfluency characteristics of omani arabic-english bilingual speakers. *Clinical Linguistics & Phonetics*, 35(7):593–609.

Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, Jie Chen, Jingdong Chen, Zhijie Chen, Mike Chrzanowski, Adam Coates, Greg Diamos, Ke Ding, Niandong Du, Erich Elsen, Jesse Engel, Weiwei Fang, Linxi Fan, Christopher Fougner, Liang Gao, Caixia Gong, Awni Hannun, Tony Han, Lappi Johannes, Bing Jiang, Cai Ju, Billy Jun, Patrick LeGresley, Libby Lin, Junjie Liu, Yang Liu, Weigao Li, Xiangang Li, Dongpeng Ma, Sharan Narang, Andrew Ng, Sherjil Ozair, Yiping Peng, Ryan Prenger, Sheng Qian, Zongfeng Quan, Jonathan Raiman, Vinay Rao, Sanjeev Satheesh, David Seetapun, Shubho Sengupta, Kavya Srinet, Anuroop Sriram, Haiyuan Tang, Liliang Tang, Chong Wang, Jidong Wang, Kaifu Wang, Yi Wang, Zhijian Wang, Zhiqian Wang, Shuang Wu, Likai Wei, Bo Xiao, Wen Xie, Yan Xie, Dani Yogatama, Bin Yuan, Jun Zhan, and Zhenyao Zhu. 2016. Deep speech 2 : End-to-end speech recognition in english and mandarin. In

Proceedings of The 33rd International Conference on Machine Learning, volume 48 of *Proceedings of Machine Learning Research*, pages 173–182, New York, New York, USA. PMLR.

Antonios Anastasopoulos and David Chiang. 2018. Tied multitask learning for neural speech translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 82–91, New Orleans, Louisiana. Association for Computational Linguistics.

I Wayan Arka and Mary Dalrymple. 2017. Nominal, pronominal, and verbal number in balinese. *Linguistic Typology*, 21(2):261–331.

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. *Handbook of natural language processing*, 2:267–292.

Sameer Bansal, Herman Kamper, Karen Livescu, Adam Lopez, and Sharon Goldwater. 2018. Low-resource speech-to-text translation. *arXiv preprint arXiv:1803.09164*.

Catherine T Best et al. 1994. The emergence of native-language phonological influences in infants: A perceptual assimilation model. *The development of speech perception: The transition from speech sounds to spoken words*, 167(224):233–277.

Vineet Bhat, Preethi Jyothi, and Pushpak Bhattacharyya. 2023. Disco: A large scale human annotated corpus for disfluency correction in indo-european languages. *Preprint*, arXiv:2310.16749.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Shelley B Brundage and Hannah Rowe. 2018. Rates of typical disfluency in the conversational speech of 30-month-old spanish–english simultaneous bilinguals. *American Journal of Speech-Language Pathology*, 27(3S):1287–1298.

Sofia Carias and David Ingram. 2006. Language and disfluency: Four case studies on spanish-english bilingual children. *Journal of Multilingual Communication Disorders*, 4(2):149–157.

Tanmoy Chakraborty and Sivaji Bandyopadhyay. 2010. Identification of reduplication in bengali corpus and their semantic analysis: A rule based approach. In *Proceedings of the 2010 Workshop on Multiword Expressions: from Theory to Applications*, pages 73–76.

Angelica Chen, Vicky Zayats, Daniel D Walker, and Dirk Padfield. 2022. Teaching bert to wait: Balancing accuracy and latency for streaming disfluency detection. *arXiv preprint arXiv:2205.00620*.

Feng-yi Chen, Ruo-ping Mo, Chu-Ren Huang, and Keh-Jiann Chen. 1992. Reduplication in mandarin chinese: Their formation rules, syntactic behavior and icg representation. In *Proceedings of rocling v computational linguistics conference v*, pages 217–233.

Eunah Cho, Jan Niehues, Thanh-Le Ha, and Alex Waibel. 2016. Multilingual disfluency removal using NMT. In *Proceedings of the 13th International Conference on Spoken Language Translation*, Seattle, Washington D.C. International Workshop on Spoken Language Translation.

Lisa Conathan and Esther Wood. 2003. Repetitive reduplication in yurok and karuk: Semantic effects of contact. *Algonquian Papers-Archive*, 34.

Stephen Pit Corder. 1983. A role for the mother tongue. *Language transfer in language learning*, 1:85–97.

Mai Hoang Dao, Thinh Hung Truong, and Dat Quoc Nguyen. 2022. Disfluency detection for Vietnamese. In *Proceedings of the Eighth Workshop on Noisy User-generated Text (W-NUT 2022)*, pages 194–200, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Hans W Dechert and Manfred Raupach. 1989. Transfer in language production. *(No Title)*.

Hossep Dolatian and Jeffrey Heinz. 2018. Modeling reduplication with 2-way finite-state transducers. In *Proceedings of the Fifteenth Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 66–77, Brussels, Belgium. Association for Computational Linguistics.

Hossep Dolatian and Jeffrey Heinz. 2019. Redtyp: A database of reduplication with computational models. *Society for Computation in Linguistics*, 2(1).

Qianqian Dong, Feng Wang, Zhen Yang, Wei Chen, Shuang Xu, and Bo Xu. 2019. Adapting translation models for transcript disfluency detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6351–6358.

Thoudam Doren Singh and Sivaji Bandyopadhyay. 2011. Integration of reduplicated multiword expressions and named entities in a phrase based statistical machine translation system. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 1304–1312, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

James Ferguson, Greg Durrett, and Dan Klein. 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.

Kallirroi Georgila, Ning Wang, and Jonathan Gratch. 2010. Cross-domain speech disfluency detection. In *Proceedings of the SIGDIAL 2010 Conference*, pages 237–240, Tokyo, Japan. Association for Computational Linguistics.

Sahar Ghannay, Yannick Esteve, and Nathalie Camelin. 2015. Word embeddings combination and neural networks for robustness in asr error detection. In *2015 23rd European Signal Processing Conference (EUSIPCO)*, pages 1671–1675. IEEE.

John Godfrey and Edward Holliman. 1993. Switchboard-1 release 2 ldc97s62. *Linguistic Data Consortium*, 34.

Matthias Honal and Tanja Schultz. 2003. Correction of disfluencies in spontaneous speech using a noisy-channel approach. In *Interspeech*.

Matthew Honnibal and Mark Johnson. 2014. Joint Incremental Disfluency Detection and Dependency Parsing. *Transactions of the Association for Computational Linguistics*, 2:131–142.

Julian Hough and David Schlangen. 2015. Recurrent neural networks for incremental disfluency detection. In *Proc. Interspeech 2015*, pages 849–853.

Dirk Hovy and Shrimai Prabhumoye. 2021. Five sources of bias in natural language processing. *Language and linguistics compass*, 15(8):e12432.

Paria Jamshid Lou and Mark Johnson. 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.

Paria Jamshid Lou and Mark Johnson. 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.

Mark Johnson and Eugene Charniak. 2004. A tag-based noisy-channel model of speech repairs. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 33–39.

Chayan Kochar, Vandan Vasantlal Mujadia, Pruthwik Mishra, and Dipti Misra Sharma. 2024. Towards disfluency annotated corpora for Indian languages. In *Proceedings of the 7th Workshop on Indian Language Data: Resources and Evaluation*, pages 1–10, Torino, Italia. ELRA and ICCL.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the national academy of sciences*, 117(14):7684–7689.

Rohit Kundu, Preethi Jyothi, and Pushpak Bhattacharyya. 2022. Zero-shot disfluency detection for Indian languages. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4442–4454, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Charles Lam. 2013. Reduplication across categories in cantonese. In *Proceedings of the 27th Pacific Asia Conference on Language, Information, and Computation (PACLIC 27)*, pages 277–286.

Yang Liu, Elizabeth Shriberg, Andreas Stolcke, Dustin Hillard, Mari Ostendorf, and Mary Harper. 2006. Enriching speech recognition with automatic detection of sentence boundaries and disfluencies. *IEEE Transactions on audio, speech, and language processing*, 14(5):1526–1540.

Paria Jamshid Lou and Mark Johnson. 2020. End-to-end speech recognition and disfluency removal. *arXiv preprint arXiv:2009.10298*.

Yuèyuán Lǐ and Dan Ponsford. 2018. Predicative reduplication: Functions, their relationships and iconicities. *Linguistic Typology*, 22(1):51–117.

Nina Markl and Stephen Joseph McNulty. 2022. Language technology practitioners as language managers: arbitrating data bias and predictive bias in asr. *arXiv preprint arXiv:2202.12603*.

Josh Meyer, Lindy Rauchenstein, Joshua D Eisenberg, and Nicholas Howell. 2020. Artie bias corpus: An open dataset for detecting demographic bias in speech applications. In *Proceedings of the twelfth language resources and evaluation conference*, pages 6462–6468.

Payal Mohapatra, Akash Pandey, Bashima Islam, and Qi Zhu. 2022. Speech disfluency detection with contextual representation and data distillation. In *Proceedings of the 1st ACM international workshop on intelligent acoustic systems and applications*, pages 19–24.

Atsunori Ogawa and Takaaki Hori. 2017. Error detection and accuracy estimation in automatic

speech recognition using deep bidirectional recurrent neural networks. *Speech Communication*, 89:70–83.

Mari Ostendorf and Sangyun Hahn. 2013. A sequential repetition model for improved disfluency detection. In *Proc. Interspeech 2013*, pages 2624–2628.

Charles S Peirce. 1955. *Philosophical writings of Peirce*, volume 217. Courier Corporation.

Xian Qian and Yang Liu. 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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Mohammad Sadegh Rasooli and Joel Tetreault. 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.

Terry Regjer. 2022. Reduplication and the arbitrariness of the sign. In *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society*, pages 887–892. Routledge.

Fedor Ivanovich Rozhanskiy. 2015. Two semantic patterns of reduplication: Iconicity revisited. *Studies in Language. International Journal sponsored by the Foundation "Foundations of Language"*, 39(4):992–1018.

Ivan A Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. In *Computational Linguistics and Intelligent Text Processing: Third International Conference, CICLing 2002 Mexico City, Mexico, February 17–23, 2002 Proceedings 3*, pages 1–15. Springer.

Elizabeth Salesky, Matthias Sperber, and Alex Waibel. 2019. Fluent translations from disfluent speech in end-to-end speech translation. *arXiv preprint arXiv:1906.00556*.

Matthew Stephen Seigel and Philip C Woodland. 2014. Detecting deletions in asr output. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2302–2306. IEEE.

Elizabeth Ellen Shriberg. 1994. *Preliminaries to a theory of speech disfluencies*. Ph.D. thesis, Citeseer.

Lucia Specia, Kashif Shah, Jose G.C. de Souza, and Trevor Cohn. 2013. QuEst - a translation quality estimation framework. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 79–84, Sofia, Bulgaria. Association for Computational Linguistics.

Yik-Cheung Tam, Yun Lei, Jing Zheng, and Wen Wang. 2014. Asr error detection using recurrent neural network language model and complementary asr. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2312–2316. IEEE.

Maria Teleki, Xiangjue Dong, and James Caverlee. 2024. Quantifying the impact of disfluency on spoken content summarization. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13419–13428, Torino, Italia. ELRA and ICCL.

Feng Wang, Wei Chen, Zhen Yang, Qianqian Dong, Shuang Xu, and Bo Xu. 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.

Shaolei Wang, Wangxiang Che, Qi Liu, Pengda Qin, Ting Liu, and William Yang Wang. 2020. Multi-task self-supervised learning for disfluency detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9193–9200.

Yang Wang. 2021. Recognizing reduplicated forms: Finite-state buffered machines. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 177–187, Online. Association for Computational Linguistics.

Alicia Beckford Wassink, Cady Gansen, and Isabel Bartholomew. 2022. Uneven success: automatic speech recognition and ethnicity-related dialects. *Speech Communication*, 140:50–70.

Ron J. Weiss, Jan Chorowski, Navdeep Jaitly, Yonghui Wu, and Zhifeng Chen. 2017. Sequence-to-sequence models can directly transcribe foreign speech. *CoRR*, abs/1703.08581.

Barbara Wheatley and Joseph Picone. 1991. Voice across america: Toward robust speaker-independent speech recognition for telecommunications applications. *Digital Signal Processing*, 1(2):45–63.

Shuangzhi Wu, Dongdong Zhang, Ming Zhou, and Tiejun Zhao. 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.

Masashi Yoshikawa, Hiroyuki Shindo, and Yuji Matsumoto. 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.

Vicky Zayats and Mari Ostendorf. 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.

Vicky Zayats, Mari Ostendorf, and Hannaneh Hajishirzi. 2016. Disfluency Detection Using a Bidirectional LSTM. In *Proc. Interspeech 2016*, pages 2523–2527.

Victoria Zayats, Mari Ostendorf, and Hannaneh Hajishirzi. 2014. Multi-domain disfluency and repair detection. In *INTERSPEECH*, pages 2907–2911.

Simon Zwarts and Mark Johnson. 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.