



# Bilingual Speakers Exhibit Cognitive Fatigue: A Speech Disfluencies Case Study on Research Talks

Ashwin Ram<sup>1</sup>, Marisol Muñoz<sup>2</sup>, Zoi Gkalitsiou<sup>2</sup>, Alexandros G. Dimakis<sup>3,4</sup>

<sup>1</sup>University of Texas at Austin, USA; <sup>2</sup>California State University, East Bay, USA; <sup>3</sup>University of California, Berkeley, USA; <sup>4</sup>BespokeLabs.AI, USA

ashwin.ram@utexas.edu, mmunoz51@horizon.csueastbay.edu, zoi.gkalitsiou@csueastbay.edu, alexdimakis@berkeley.edu

## Abstract

Speech disfluencies are vital for understanding cognitive processes and improving speech recognition systems. We curate a dataset with annotated text and labeled speech disfluencies from more than 20 hours of speech from monolingual and bilingual speakers. Furthermore, we illustrate a large-scale validation of a bilingual cognitive fatigue phenomenon that seems to be independent of the two spoken languages of the speakers. That is, after navigating a lexically complex word, bilingual speakers tend to use a disfluency, such as a filled pause or repair, followed by a phonetically simpler word in order to possibly regain momentum for subsequent utterance segments. We conclude by exploring how our research can help speech pathologists by revealing distinct bilingual cognitive strategies and how they manifest in speaker disfluencies.

**Index Terms:** speech disfluencies, bilinguals, machine learning, statistical analysis

## 1. Introduction and Related Work

Speech disfluencies, such as repetitions, revisions, pauses, and fillers, are prevalent in both typical and disordered speech and serve as key signals of underlying cognitive processes [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]. Such disfluencies occur across the lifespan [6, 11], can signal processing difficulties [12], and often vary in form, depending on the speaker’s linguistic background [13, 14], age [2, 6], and clinical profiles [15, 16, 17]. Disfluencies are also related to cognitive load, revealing how speakers plan and monitor their utterances [7, 18, 19, 20] and offering insights for speech and language production [5, 8].

With regard to bilingual speech production, disfluencies can be indicative of increased cognitive demands or challenges during language production due to the navigation of two languages [21]. These challenges may be due to insufficient knowledge and practice of the speaker’s second language. Namely, when using their second language, speakers are recruiting more cognitive resources that can trigger more disfluencies and repairs as opposed to using their first language, which requires routine cognitive processes [14].

Recent advances in *automatic* speech recognition have motivated robust approaches to labeling disfluencies and nonlexical sounds [22, 23], with end-to-end frameworks [24, 25] and hierarchical methods [26] emerging as state-of-the-art. Techniques such as self-supervised and multimodal architectures [27, 28] have further improved sensitivity to disfluent segments, while corpus-centric comparisons [29] underscore the importance of model selection in handling spontaneous speech variability. Additional research has explored the synergy between accent modeling [30] and lexical difficulty.

Disfluency metrics are part of diagnostic and/or treatment

protocols for various communication disorders in children and adults [1, 2, 3, 6]. Unfortunately, bilingual speakers are largely evaluated using disfluency measures designed for monolingual English speakers, which places them at a higher risk for misdiagnosis of speech or language disorders [31, 32, 33]. Given the reported overlapping as well as distinctive speech patterns between bilingual and monolingual speakers, recent investigations suggest that understanding disfluency variability across languages contexts [5, 17] can improve clinical decision making. Therefore, integration of data-driven approaches to improve the distinction of typical versus atypical speech patterns in bilingual speakers is warranted. Several studies have extensively analyzed disfluencies in monolingual speech [1, 2, 11] and in controlled bilingual experiments [13, 18] but not at a wide range of languages or long-duration speech samples.

Our work contains the following contributions:

1. We develop an automatic machine learning methodology for generating labeled speech disfluency datasets with between-word disfluencies.
2. We curate and open-source a large-scale, naturalistic speech dataset on highly technical academic talks drawn from an online lecture repository in a replicable manner.
3. We capture rich utterance-level insights that reveal a novel bilingual compensatory strategy: after encountering a demanding lexical item, bilingual speakers deploy a disfluency and then revert to simpler words, suggesting a mechanism of cognitive fatigue and recovery. This phenomenon, absent in most automatic disfluency detection studies [24, 25, 26], underscores the need to consider deeper the linguistic background and cognitive load in future speech modeling frameworks.

## 2. Data

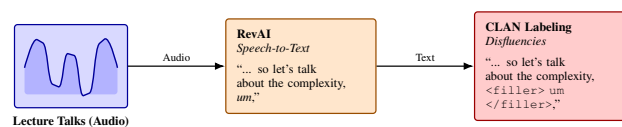


Figure 1: **Pipeline Overview:** Lecture Talks (Audio) are passed to RevAI for transcription, resulting in text that preserves disfluencies. The text is then processed using CMU CLAN [34] to insert explicit disfluency labels.

### 2.1. Dataset Construction

We introduce a novel pipeline for annotating speech data: Figure 1 depicts the three main stages. First, we obtain audio and

convert the speech to text via RevAI<sup>1</sup>, and then label the disfluencies with the Computerized Language ANalysis (CLAN) guidelines introduced by Dr. MacWhinney’s group at Carnegie Mellon University [34]. This approach preserves the disfluencies, making the resulting annotations more reflective of natural speech. In addition, our pipeline allows for simple substitution of the transcription engine or annotation method, allowing for broad applicability to multiple domains and language varieties.

To ensure that all speakers are in a controlled setting and that the method of collecting and annotating data is replicable, we use online scientific lectures (approximately 45 to 75 minutes long, in the same setting) of various researchers. These are lectures where speakers are frequently interrupted with questions, as opposed to scripted speeches, and hence contain naturally occurring speech disfluencies.

## 2.2. Dataset Statistics

Using the pipeline illustrated in Figure 1, we curate an open-source dataset of roughly 21 hours of lectures, each extracted from publicly available online sources and carefully transcribed to include explicit between-word disfluencies. The dataset spans native English speakers, who speak only English (which is the monolingual group) and non-native English speakers, in each of the following bilingual groups: Greek-English, Spanish-English, and Hebrew-English.

We chose not to annotate *silent pauses*, since they heavily depend on recording conditions and individual speaking speeds, or within-word *lengthenings*, because their acoustic boundaries are similarly difficult to consistently automatically identify. Instead, we focus on between-word fillers, repetitions, and revisions, as these disfluencies can be reliably and clearly identified across all talks.

As shown in Table 1, each transcript has annotated *fillers*, *repetitions*, and *revisions* as defined below:

- **Fillers:** Vocalizations that do not add semantic content but fill pauses (e.g., “*um*,” “*uh*”).
- **Repetitions:** Unintentional repeats of words or phrases (e.g., “*I think—I think we should proceed*”).
- **Revisions:** Corrections of lexical or grammatical changes mid-sentence (e.g., “*We need to— we should try a different approach*”).

We apply this uniform speech-to-text and disfluency labeling process across all lecture recordings. As such, the resulting corpus preserves a range of disfluencies and utterances that are of particular interest to speech pathologists, allowing them to investigate and compare disfluency patterns among speakers with different linguistic backgrounds. Because this methodology is reproducible, it extends to the study of cross-linguistic variance in speech production, providing researchers and clinicians with a rich, real-world data set to explore both language-specific and broader universal traits of disfluency.

## 2.3. Human Annotator Scoring

A Speech Language Pathology (SLP) research assistant trained in disfluency annotation validated the performance of our method. The instructions were to watch and annotate the disfluencies in 10-15 minutes of one speaker in each of the four linguistic backgrounds. The research assistant watched the corresponding four lecture videos and manually marked false positives (FP) and false negatives (FN) of the three types of marked

between-word disfluency in the transcripts: fillers, revisions, and repetitions. Table 1 shows the scores received by our data curation method compared to human annotations.

Despite the inherent variability in speaking styles and the natural occurrence of between-word disfluencies, the overall precision of 0.966 demonstrates the high reliability of our system in accurately detecting fillers, revisions, and repetitions. The Recall scores (ranging from 0.733 for Hebrew to 0.873 for Spanish) suggest that while most disfluencies are correctly identified by our model, there remains room for improvement in capturing all instances, particularly for languages with smaller datasets or more nuanced speech patterns. However, strong F1 scores across languages highlight the effectiveness of our approach, especially since the data construction method does not require any manual effort.

# 3. Methodology

## 3.1. Explanation of Features

### 3.1.1. Lexical Features

We include three main lexical properties for each word. First, we estimate *word frequency* by consulting a comprehensive language usage database [35]. Each word’s frequency is represented as a numerical value indicating how commonly it appears across multiple text sources. For instance, high-frequency words (e.g., “*the*”) will be assigned a large value, whereas technical or domain-specific words (e.g., “*hypergraph*”) appear less often and therefore receive a lower frequency value.

Second, we identify whether the word belonged to a class of closed-category *stop words* as opposed to open-category (content words) using the `spacy` library. Stop words are typically short, functional words (such as articles, prepositions, or pronouns) that carry relatively little lexical meaning [36]. A binary indicator reflected if a given token was a stop word or not.

Next, we mark each word as *content-bearing* or otherwise by examining its part-of-speech using the `nltk` library. In particular, words falling into open-class categories (e.g., nouns, verbs, adjectives, adverbs) were treated as *content words*, whereas other grammatical categories were not [36]. Combining these three lexical attributes offered a broad characterization of each token’s role in the language (frequent vs. rare, functional vs. contentful).

Finally, we incorporate *neighborhood density* as an additional lexical feature. Neighborhood density refers to the number of words that differ from a target word by only one phoneme [37] (e.g., for the target word “*cat*”, “*bat*” is an example of a neighbor because it differs by just one phoneme). To obtain these values, we used a web-based interface from Kansas University that calculates phonotactic probability for English words and nonwords [38]. If the interface did not provide data for a particular token, we assigned a neighborhood density of 0. Including this measure helps capture how many similar-sounding words may compete with or facilitate recognition and production.

### 3.1.2. Index of Phonetic Complexity (IPC)

To assess each word’s phonetic complexity, we use the Index of Phonetic Complexity, a standard measure that assigns a numerical score based on various phonological sub-elements [39]. In this framework, a word’s complexity increases whenever it exhibits features, such as multiple consonantal places of articulation, a greater number of syllables, or the presence of hetero-

<sup>1</sup>Can be accessed here: <https://www.rev.ai/>

Table 1: Combined View of the Main Dataset and Human Validation Metrics

Lectures Dataset						
Language Group	Total Words	Filler	Repetition	Revision	Disfluency Ratio	
English	40019	2745	74	44	0.0715	
Greek	34303	2298	46	60	0.0701	
Hebrew	22735	1324	45	19	0.0611	
Spanish	24881	1705	31	120	0.0746	

Human Validation						
	TP	FN	FP	Precision	Recall	F1-Score
<i>Overall</i>	<b>607</b>	<b>140</b>	<b>21</b>	<b>0.966</b>	<b>0.813</b>	<b>0.884</b>
<i>Spanish</i>	186	27	4	0.979	0.873	0.924
<i>Hebrew</i>	135	49	1	0.993	0.733	0.844
<i>Greek</i>	119	35	5	0.961	0.773	0.855
<i>English</i>	167	29	11	0.938	0.852	0.893

ganic consonant clusters (i.e., adjacent consonants produced at different places in the vocal tract). That is, the IPC is computed by summing eight binary components, each reflecting whether the word meets a particular phonological criterion. For example, if the word ends in a consonant, it contributes one point toward the total. Similarly, if the vowel component contains a diphthong or rhotic segment, it adds another point, and so on. Higher IPC scores thus indicate that a word combines more intricate phonological attributes. More concretely, if a word has a higher IPC score, it is a more complex word in terms of *phonetics*.

When computing the IPC score, we first convert each word into a phonemic representation using the CMU Pronouncing Dictionary<sup>2</sup> and then apply the IPC criteria to the resulting phoneme sequence. This allows us to identify details such as dorsal fricatives, rhotic vowels, or three-syllable-plus shapes. The final IPC value thus captures how “demanding” the word is from a phonological standpoint. In turn, this enriched phonological metric complements our lexical analysis, enabling a deeper examination of how speakers navigate complex word forms when disfluencies occur.

### 3.2. Statistical Analysis

In order to assess which linguistic features predict the likelihood of a word preceding and following a disfluency, we perform a logistic regression of the form

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \sum_{k=1}^K \beta_k X_{ki}, \quad (1)$$

where  $p_i$  is the probability that word  $i$  is either pre or post-disfluent,  $X_{ki}$  represents the  $k$ th feature (e.g., *IPC*, *Word Frequency*, *Stop Word*, *Content Word*, etc.) for word  $i$ , and  $\beta_k$  is the coefficient for that feature. Note that all features are normalized to have zero mean and unit variance. Table 2 summarizes the regression coefficients, standard errors, and  $p$ -values for each feature in both the *Monolingual* and *Bilingual* groups.

We include both *preceding* and *following* word features, motivated by the hypothesis that the phonetic complexity, frequency, or grammatical class of neighboring words might influence where disfluencies occur. Logistic regression was chosen

for its interpretability: each coefficient  $\beta_k$  can be viewed as the change in the *log-odds* of a word being pre-disfluent or post-disfluent, per unit change in feature  $X_k$ .

Statistical significance was tested at  $p < 0.05$  (or  $p < 0.001$  where indicated). All analyses were conducted in Python with the `statsmodels` library, reporting standard errors, Wald  $z$ -statistics, and  $p$ -values.

## 4. Results and Discussion

### 4.1. Overview of Findings

Table 2 presents a side-by-side comparison of how lexical and phonetic features predict both pre-disfluency and post-disfluency boundaries for *monolingual* and *bilingual* speakers. This analysis indicates that while both groups are sensitive to whether the adjacent word is a *stop word* or a word with more neighbors, bilingual speakers exhibit additional significant effects tied to word frequency and phonetic complexity. In particular, *Word Frequency Preceding*, *Neighbor Density Preceding*, and *IPC Following* significantly shape bilingual disfluency behavior, highlighting the interplay of cognitive load, lexical retrieval, and phonological planning.

### 4.2. Monolingual Patterns

Focusing on the monolingual model, the only features reaching significance (*Stop Word Following* and *Neighborhood Density Following*) suggest that monolingual speakers place disfluencies differently depending on the grammatical category of neighboring words. When a stop word directly follows a disfluency, the model assigns a higher probability of disfluency, possibly reflecting a quick pivot from one unit of meaning to another. Moreover, if the following word is less rare and has more neighbors, monolinguals also appear more likely to place a disfluency boundary. However, phonetic complexity (IPC) and word frequency show no significant effect in the monolingual model, implying that monolingual speakers are comparatively less burdened by lexical retrieval difficulties when transitioning between words.

### 4.3. Bilingual Patterns

By contrast, bilingual speakers display a more intricate set of interactions between lexical and phonetic attributes. Namely, for

<sup>2</sup>[www.nltk.org/\\_modules/nltk/corpus/reader/cmudict.html](http://www.nltk.org/_modules/nltk/corpus/reader/cmudict.html)

Table 2: *Logistic Regression Results for Monolingual vs. Bilingual on the Online Lectures Dataset. Significant features for the Monolingual model are Neighbor Density Following ( $p < 0.01$ ) and Stop Word Following ( $p < 0.05$ ). Significant features for the Bilingual model are IPC Following ( $p < 0.05$ ), Word Frequency Preceding ( $p < 0.001$ ), Neighbor Density Preceding ( $p < 0.001$ ), Stop Word Following ( $p < 0.001$ ), Content Word Preceding ( $p < 0.001$ ), and Content Word Following ( $p < 0.001$ ).*

Feature	Monolingual			Bilingual		
	Coeff.	Std Err	p-val	Coeff.	Std Err	p-val
IPC Preceding	0.0590	0.060	0.329	-0.0678	0.039	0.085
IPC Following	0.0259	0.072	0.718	<b>-0.0855</b>	<b>0.042</b>	<b>0.040</b>
Word Frequency Preceding	-0.1165	0.081	0.150	<b>-0.2167</b>	<b>0.057</b>	<b>&lt;0.001</b>
Word Frequency Following	-0.0754	0.067	0.262	-0.0318	0.042	0.453
Neighbor Density Preceding	-0.0889	0.063	0.160	<b>-0.1806</b>	<b>0.042</b>	<b>&lt;0.001</b>
Neighbor Density Following	<b>0.1627</b>	<b>0.058</b>	<b>0.005</b>	-0.0093	0.069	0.893
Stop Word Preceding	-0.0702	0.078	0.369	-0.0610	0.048	0.208
Stop Word Following	<b>0.1637</b>	<b>0.076</b>	<b>0.032</b>	<b>0.3266</b>	<b>0.045</b>	<b>&lt;0.001</b>
Content Word Preceding	0.1641	0.084	0.051	<b>0.2469</b>	<b>0.053</b>	<b>&lt;0.001</b>
Content Word Following	0.1100	0.074	0.136	<b>0.3915</b>	<b>0.046</b>	<b>&lt;0.001</b>

**Word Frequency Preceding ( $p < 0.001$ ) and Neighborhood Density Preceding ( $p < 0.001$ ),** we see a strong negative coefficient for both features, which indicates that having just produced a **rarer** word (i.e., a word that has a *low-frequency* or that is from a *smaller neighborhood*) correlates with an elevated probability of a disfluency on the *subsequent* utterance. We interpret this as a signal of heightened cognitive load or “fatigue” after retrieving an infrequent lexical form. This finding aligns with prior research [20] that found that bilingual speakers may experience greater lexical access costs and thus insert disfluencies more frequently to regain planning capacity. Similarly, for **IPC Following ( $p < 0.05$ )**, the bilingual model shows a significant negative relationship, meaning that if the *next* word is *less* phonologically complex (i.e., it has a lower IPC score), the likelihood of a preceding disfluency boundary is higher. Interestingly, we find that rather using a disfluency *before* a difficult word (as one might expect), bilinguals appear to experience a “cognitive release” when an upcoming word is simpler. In essence, having successfully navigated a demanding or difficult prior word, they may allow a disfluency to surface in the subsequent transition to a more straightforward or simpler word.

Overall, the bilingual findings reveal a resource-depletion mechanism that significantly influences where disfluencies emerge: cognitively taxing preceding words (i.e., rare or lexically challenging) precipitate higher disfluency probabilities, while simpler upcoming words function as natural points of disfluency release. Such patterns are absent among monolinguals, underscoring the distinct speech-planning burden that bilinguals face. Clinically, these insights may inform diagnostic protocols wherein disfluencies are used as diagnostic markers of communication disorder, particularly in bilingual populations. Speech-language pathologists could leverage the knowledge that bilingual speakers exhibit disfluencies in systematic ways, tied to specific lexical and phonological contexts, to refine assessments and interventions.

## 5. Conclusion and Future Work

In this paper, we develop an automated pipeline for collecting and labeling fine-grained disfluencies from online lecture talks, spanning both monolingual and bilingual speakers. Our logistic regression analyses reveal distinct patterns that connect cognitive load, particularly for bilingual speakers, to predictable lo-

cations of disfluencies. Lower frequency preceding words and less phonetic complexity in following words emerged as key triggers for bilingual disfluency boundaries, suggesting a *cognitive fatigue–release* cycle.

Moving forward, **future work** can expand the corpus to include *additional languages* (e.g., Chinese-English, Arabic-English) and broader *speaker demographics*. This would enable further testing of our hypothesized load–release mechanism across more diverse linguistic contexts. Moreover, *multimodal methods* incorporating audio, articulatory signals, and real-time EEG or fMRI could offer deeper insights into the neural basis of bilingual disfluencies. Finally, we aim to integrate our findings into advanced *disfluency detection* and *automatic speech recognition* systems. By accounting for bilingual lexical-phonological factors, further research may achieve more accurate alignment of disfluent speech and yield novel clinical applications for diagnosing fluency disorders in diverse populations.

## 6. References

- [1] R. R. Martin, S. K. Haroldson, and P. Kuhl, “Disfluencies of young children in two speaking situations,” *Journal of Speech and Hearing Research*, vol. 15, no. 4, pp. 831–836, 1972. [Online]. Available: <https://pubs.asha.org/doi/abs/10.1044/jshr.1504.831>
- [2] E. Yairi and N. F. Clifton, “Disfluent speech behavior of preschool children, high school seniors, and geriatric persons,” *Journal of Speech and Hearing Research*, vol. 15, no. 4, pp. 714–719, 1972. [Online]. Available: <https://pubs.asha.org/doi/abs/10.1044/jshr.1504.714>
- [3] E. Yairi and B. Lewis, “Disfluencies at the onset of stuttering,” *Journal of Speech, Language, and Hearing Research*, vol. 27, no. 1, pp. 154–159, 1984. [Online]. Available: <https://pubs.asha.org/doi/abs/10.1044/jshr.2701.154>
- [4] R. Soma, “Speaking disfluency of an english major students at one public university in indonesia,” *PPSDP International Journal of Education*, vol. 2, no. 2, pp. 427–435, 2023.
- [5] K. N. Johnson and M. T. Mills, “Exploratory examination of speech disfluencies in spoken narrative samples of school-age bidialectal children,” *American Journal of Speech-Language Pathology*, vol. 32, no. 3, pp. 1182–1194, 2023. [Online]. Available: [https://pubs.asha.org/doi/abs/10.1044/2023\\_AJSLP-21-00158](https://pubs.asha.org/doi/abs/10.1044/2023_AJSLP-21-00158)
- [6] E. J. Beier, S. Chantavarin, and F. Ferreira, “Do disfluencies increase with age? evidence from a sequential corpus study of dis-

- fluencies,” *Psychology and Aging*, vol. 38, no. 3, pp. 203–218, 2023.
- [7] K. Rapoeye, R. J. Hartsuiker, and A. Pistono, “Semantic interference affects speech production by increasing disfluencies, not errors,” *Royal Society Open Science*, vol. 10, p. 230006, 2023. [Online]. Available: <http://doi.org/10.1098/rsos.230006>
- [8] M. M. Laske and F. D. D. Reed, “Um, so, like, do speech disfluencies matter? a parametric evaluation of filler sounds and words,” *Journal of Applied Behavior Analysis*, vol. 57, no. 3, pp. 574–583, 2024.
- [9] E. E. Shriberg, “Preliminaries to a theory of speech disfluencies,” Ph.D. dissertation, University of California, Berkeley, Berkeley, CA, USA, 1994. [Online]. Available: <https://citeseerx.ist.psu.edu/documentrepid=rep1&type=pdf&doi=c0ca94051f549f08e0bb4be7694540460fd47f1b>
- [10] R. Eklund, “Disfluency in swedish human–human and human–machine travel booking dialogues,” Ph.D. dissertation, Linköping University, Linköping, Sweden, 2004.
- [11] S. J. Owens, J. M. Thacker, and S. A. Graham, “Disfluencies signal reference to novel objects for adults but not children,” *Journal of Child Language*, vol. 45, no. 3, pp. 581–609, 2018.
- [12] A. Sugiura, Z. Alqatan, Y. Nakai, T. Kambara, B. H. Silverstein, and E. Asano, “Neural dynamics during the vocalization of ‘uh’ or ‘um’,” *Scientific Reports*, vol. 10, p. 11987, 2020. [Online]. Available: <https://doi.org/10.1038/s41598-020-68606-x>
- [13] Z. Gkalitsiou and D. Werle, “Speech disfluencies in bilingual greek–english young adults,” *Journal of Fluency Disorders*, vol. 78, p. 106001, 2023. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0094730X2300044X>
- [14] R. Chakraborty, N. Morales, K. Fritsch, and M. D. Gonzales, “Second language proficiency and maze: Marathi–english bilinguals,” *Clinical Archives of Communication Disorders*, vol. 2, no. 2, pp. 103–115, 2017. [Online]. Available: <https://doi.org/10.21849/cacd.2017.00101>
- [15] J. S. Yaruss and E. G. Cature, “Stuttering and phonological disorders in children: Examination of the covert repair hypothesis,” *Journal of Speech, Language, and Hearing Research*, vol. 39, no. 2, pp. 349–364, 1996. [Online]. Available: <https://pubs.asha.org/doi/abs/10.1044/jshr.3902.349>
- [16] J. V. Borsel, E. Geirnaert, and R. V. Coster, “Another case of word-final disfluencies,” *Folia Phoniatrica et Logopaedica*, vol. 57, no. 3, pp. 148–162, 2005.
- [17] I. Balčiūnienė and A. N. Kornev, “Linguistic disfluencies in russian-speaking typically and atypically developing children: Individual variability in different contexts,” *Clinical Linguistics and Phonetics*, vol. 38, no. 4, pp. 287–306, 2023.
- [18] R. Karniol, “Stuttering, language, and cognition: A review and a model of stuttering as suprasegmental sentence plan alignment (spa),” *Psychological Bulletin*, vol. 117, no. 1, pp. 104–124, 1995.
- [19] G. Daras, N. Raouf, Z. Gkalitsiou, and A. G. Dimakis, “Multi-tasking models are robust to structural failure: A neural model for bilingual cognitive reserve,” in *Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS)*. Red Hook, NY, USA: Curran Associates Inc., 2022, p. 2546.
- [20] N. Raouf, Y. Wu, C. Bonilla, J. J. Li, S. M. Grasso, A. G. Dimakis, and Z. Gkalitsiou, “Modeling bilingual disfluencies with large language models,” in *Proceedings of the ICML Workshop on Large Language Models and Cognition*, 2024. [Online]. Available: <https://openreview.net/pdf?id=rrNAqNYRLA>
- [21] C. Bergmann, S. A. Sprenger, and M. S. Schmid, “The impact of language co-activation on I1 and I2 speech fluency,” *Acta Psychologica*, vol. 161, pp. 25–35, October 2015.
- [22] P. Mihajlik, Y. Meng, M. S. Kadar, J. Linke, B. Schuppler, and K. Mády, “On disfluency and non-lexical sound labeling for end-to-end automatic speech recognition,” in *Proceedings of Interspeech 2024*, 2024, pp. 1270–1274.
- [23] H. Nakashima and K. Shimada, “Disfluency detection with context information from real utterances and generative utterances,” in *2023 14th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)*, 2023, pp. 462–467.
- [24] X. Zhou, A. Kashyap, S. Li, A. Sharma, B. Morin, D. Baquirin, J. Vonk, Z. Ezzes, Z. Miller, M. Tempini, J. Lian, and G. Anumanchipalli, “Yolo-stutter: End-to-end region-wise speech disfluency detection,” in *Interspeech 2024*, 2024, pp. 937–941.
- [25] A. Romana, K. Koishida, and E. M. Provost, “Automatic disfluency detection from untranscribed speech,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 4727–4740, 2024.
- [26] J. Lian and G. Anumanchipalli, “Towards hierarchical spoken language disfluency modeling,” in *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, Y. Graham and M. Purver, Eds. Association for Computational Linguistics, March 2024, pp. 539–551.
- [27] Y.-J. Shih, Z. Gkalitsiou, A. G. Dimakis, and D. Harwath, “Self-supervised speech models for word-level stuttered speech detection,” in *Proceedings of the IEEE Spoken Language Technology Workshop (SLT)*, 2024, pp. 937–944.
- [28] Y. Li, G. K. Anumanchipalli, A. Mohamed, P. Chen, L. H. Carney, J. Lu, J. Wu, and E. F. Chang, “Dissecting neural computations in the human auditory pathway using deep neural networks for speech,” *Nature Neuroscience*, vol. 26, pp. 2213–2225, 2023. [Online]. Available: <https://doi.org/10.1038/s41593-023-01418-y>
- [29] M. Teleki, X. Dong, S. Kim, and J. Caverlee, “Comparing asr systems in the context of speech disfluencies,” in *Proceedings of Interspeech 2024*, 2024, pp. 4548–4552.
- [30] G. K. Anumanchipalli, L. C. Oliveira, and A. W. Black, “Accent group modeling for improved prosody in statistical parametric speech synthesis,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 6890–6894.
- [31] C. T. Byrd, “Assessing Bilingual Children: Are Their Disfluencies Indicative of Stuttering or the By-product of Navigating Two Languages?” *Semin. Speech Lang.*, vol. 39, no. 4, pp. 324–332, Sep. 2018.
- [32] C. De Lamo White and L. Jin, “Evaluation of speech and language assessment approaches with bilingual children,” *Int. J. Lang. Commun. Disord.*, vol. 46, no. 6, pp. 613–627, Nov–Dec 2011.
- [33] L. Fabiano-Smith and K. Hoffman, “Diagnostic accuracy of traditional measures of phonological ability for bilingual preschoolers and kindergarteners,” *Lang. Speech Hear. Serv. Sch.*, vol. 49, no. 1, pp. 121–134, Jan. 2018.
- [34] B. MacWhinney, *The CHILDES Project: Tools for Analyzing Talk*, 3rd ed. Mahwah, NJ: Lawrence Erlbaum Associates, 2000.
- [35] J. D. Anderson, “Phonological neighborhood and word frequency effects in the stuttered disfluencies of children who stutter,” *Journal of Speech, Language, and Hearing Research*, vol. 50, no. 1, pp. 229–247, February 2007.
- [36] A. Bell, D. Jurafsky, E. Fosler-Lussier, C. Girand, M. Gregory, and D. Gildea, “Effects of disfluencies, predictability, and utterance position on word form variation in english conversation,” *Journal of the Acoustical Society of America*, vol. 113, no. 2, pp. 1001–1024, February 2003.
- [37] M. S. Vitevitch and M. S. Sommers, “The facilitative influence of phonological similarity and neighborhood frequency in speech production in younger and older adults,” *Memory & Cognition*, vol. 31, no. 4, pp. 491–504, Jun. 2003.
- [38] M. S. Vitevitch and P. A. Luce, “A web-based interface to calculate phonotactic probability for words and nonwords in english,” *Behavior Research Methods, Instruments, & Computers*, vol. 36, no. 3, pp. 481–487, Aug. 2004.
- [39] G. A. Coalson, C. T. Byrd, and B. L. Davis, “The influence of phonetic complexity on stuttered speech,” *Clinical Linguistics and Phonetics*, vol. 26, no. 7, pp. 646–659, July 2012.