Interspeech 2024
1-5 September 2024, Kos, Greece

Comparing ASR Systems in the Context of Speech Disfluencies

Maria Teleki[1], *Xiangjue Dong*[1], *Soohwan Kim*[1], *James Caverlee*[1]

[1]Texas A&M University, USA

{mariateleki,xj.dong,cocomox26,caverlee}@tamu.edu

Abstract

In this work, we evaluate the disfluency capabilities of two automatic speech recognition systems – Google ASR and WhisperX – through a study of 10 human-annotated podcast episodes and a larger set of 82,601 podcast episodes. We employ a state-of-the-art disfluency annotation model to perform a fine-grained analysis of the disfluencies in both the scripted and non-scripted podcasts. We find, on the set of 10 podcasts, that while WhisperX overall tends to perform better, Google ASR outperforms in WIL and BLEU scores for non-scripted podcasts. We also find that Google ASR's transcripts tend to contain closer to the ground truth number of edited-type disfluent nodes, while WhisperX's transcripts are closer for interjection-type disfluent nodes. This same pattern is present in the larger set. Our findings have implications for the choice of an ASR model when building a larger system, as the choice should be made depending on the distribution of disfluent nodes present in the data.

Index Terms: automatic speech recognition, disfluency, annotations

1. Introduction

Disfluencies are "interruptions in the fluent speech stream" [1], occurring at a rate of approximately 4-6% in regular speech [2, 3, 4, 5]. They typically take the form of specific tokens such as *um* or *uh*, phrases that start and then restart a sentence (e.g., *let's go to the store, wait no, the movies today*), and others.

Naturally, many automatic speech recognition (ASR) systems treat disfluencies as *noise* and aim to remove them, since the presence of disfluencies may hinder interaction with existing systems such as voice assistants [6, 7] or hurt the quality of tasks such as summarization [8]. Yet, carefully preserving the presence of disfluencies can be a critical signal for important applications like conversational tutoring systems for education [9, 10, 11]. And new research shows how disfluencies can be helpful for memory by bringing extra attention on upcoming material [12]. Further, in 2022, 24% of people got a smart speaker to "[h]elp with a disability" [7], and yet ASR systems generally do not work well for people with stuttering disorders [13] or dysarthic speech [14].

Hence, the choice of automatic speech recognition (ASR) system is one that a system designer must make carefully. In practice, however, there is little fine-grained analysis of existing ASR systems with a special focus on disfluencies. We aim to fill this gap through a comprehensive study of two representative ASR systems – Google ASR [15] and WhisperX, a new state-of-the-art ASR system [16] – through a study of 10 human-annotated podcasts and a larger set of 82,601 podcast episodes (see Figure 1). Our study is organized around three research questions: **(RQ1)** How does the choice of ASR system impact performance – specifically across *scripted* and *non-scripted* podcasts? **(RQ2)** How does the choice of ASR system impact the specific disfluency types which are transcribed? **(RQ3)** And are these findings consistent at a large-scale?

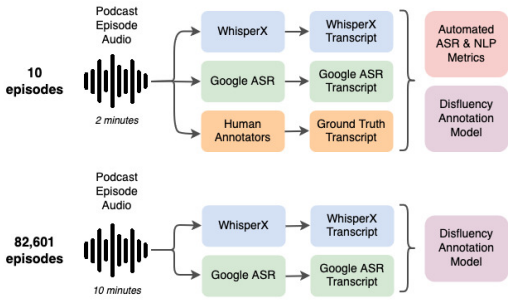

Figure 1: *We conduct a fine-grained comparison of two key ASR systems – WhisperX and Google ASR – using human annotations over 10 podcast episodes. Additionally, we compare the two ASR systems over a larger set of 82,601 episodes.*

We find that WhisperX achieves a superior word error rate of 7.19% on the podcasts, versus 10.47% for Google ASR. However, we find that specifically in the case of disfluency, Google ASR outperforms WhisperX in a few key areas. We obtain detailed annotations for 10 podcast segments, and we find that, overall, WhisperX performs better in terms of automatic metrics such as WER, BLEU, and others. And, using a parsing-based disfluency annotation model [17], we find that WhisperX also performs better in transcribing interjection nodes (such as "*uh*" and "*um*"). However, we find that for non-scripted podcasts, Google ASR achieves a better WER and BLEU score, and an edited node count which is closer to the ground truth. We additionally use the disfluency annotation model on a set of 82,601 podcasts, and find a similar pattern: WhisperX transcribes more interjection nodes, while Google ASR transcribes more edited nodes. We hypothesize that this may be due to the vocabulary diversity of WhisperX. We make the code available at `https://github.com/mariateleki/Comparing-ASR-Systems` and the full annotation guidelines and token distributions for the disfluent nodes available at `www.comparing-asr-systems.com`.

2. Experimental Settings

We first introduce the Spotify Podcast dataset [18], the ASR systems [15, 16], and our ground truth annotation process.

10.21437/Interspeech.2024-1270

2.1. The Spotify Podcast Dataset

To ground our study in a natural representation of human speech and communication, we adopt the Spotify Podcast dataset. These podcasts are typically non-scripted, colloquial, less-structured, and conversational in style. The Spotify Podcast Dataset[1] was released in 2020 as part of the TREC Podcasts Track [18], consisting of over 100,000 podcasts and 60,000 hours of audio. The podcasts are heterogeneous, spanning many different topics, conversation styles, and types of shows (e.g., interviews, stories, news, gossip, meditations, and more).

We create a **small-scale dataset of 10 podcasts with fine-grained disfluency annotations**, and a **separate large-scale dataset of 82,601 podcasts** for our analysis. We arrive at 82,601 podcasts by applying the following filters to the dataset: (i) following Reddy et al. [19], we truncate episodes to 10 minutes for normalization, (ii) we remove episodes $<$10 minutes in duration for normalization, (iii) we remove podcasts identified by WhisperX as non-English,[2] and (iv) we remove podcasts which contain $<$10 words in their transcript (we observe that these are mostly ASMR podcasts).

2.2. ASR Systems

Google ASR is a proprietary, widely-used automatic speech recognition system [15]. The Spotify Podcast Dataset was originally transcribed using Google ASR in 2020, and we study these transcriptions in this work [18]. Google ASR is representative of commercial-type ASR systems, as it tends to filter out certain disfluent tokens, such as "*uh*" and "*um*."

WhisperX is a state-of-the-art open-source ASR which iteratively improves on Whisper [20] by OpenAI, achieving superior performance over Whisper in many instances [16]. Additionally, WhisperX achieves a speedup of 12x over Whisper using Voice Activity Detection (VAD) Cut & Merge – a strategy which allows for simultaneous transcription of batched audio. We re-transcribe the podcast audio with WhisperX.

2.3. Ground Truth Annotations

Annotation Process. We ask three annotators to transcribe the first 2 minutes of the 10 randomly-selected podcasts by hand. One of the authors served as the annotator coordinator for these three annotators, to iteratively update the annotation guidelines in response to annotator feedback. We used a **three-round approach** to annotate the transcripts.

In *Round 1*, the annotators focused on two main tasks:

1. **Scripted/Non-Scripted Classification**: The annotators classified each podcast as sounding *scripted* (i.e., the speaker is reading out loud completely or mostly) or *non-scripted* (i.e., the audio does not sound scripted – this may mean that the audio is more conversational in style).
2. **Ground Truth Transcription**: Each annotator transcribed each of the first 2 minutes of audio for the 10 podcasts. The annotators especially focused on transcribing disfluencies – we provide the **abbreviated annotation guidelines** in Figure 2.[3] They transcribed each word and sound (with the exception of music) – as an ideal ASR model would. The annotators were instructed to notify the coordinator of any cases not in the guidelines as they arose, so that the coordinator could set a standard and update the guidelines.

Abbreviated Annotation Guidelines

- No punctuation, no capitalization
- Numbers as numeric (ex: *21* not *twenty-one*)
- Ignore pauses and music
- Choose the word that sounds most like what they said
  - Ex: *them* vs. *em* → *em*
  - Ex: *going to* vs. *gonna* → *gonna*
- Spell it out for hashtags and URLs (ex: *hashtag x y* not *#xy*)
- [IDK] Special Token
  - Can't understand word or words, but know that they are words and not sounds
  - Names that are hard to spell
- [INAUDIBLE] Special Token
  - Can't really hear what they're saying
  - Making sounds that aren't mapped to words (counterexample: uh and um are easily mapped to words)
  - Ex: *s- s- something* → *[INAUDIBLE] something*
  - Grunts that aren't distinct
- Classify each podcast as scripted or not scripted:
  - *Scripted*: They sound like they're reading from a script/document.
  - *Not Scripted*: The conversation may (or may not be) guided by topics/questions, but it isn't scripted; it's more conversational.

Figure 2: *Abbreviated annotation guidelines; for full guidelines, see* `www.comparing-asr-systems.com`.

In *Round 2*, the [IDK][4] tokens were resolved. The annotators viewed each others' annotations for the [IDK] tokens and the surrounding context (provided by the coordinator), and then decided whether to keep their original annotation, or change it based on what the other annotators wrote for that segment.

In *Round 3*, the coordinator combined the transcripts from three annotators to form the **ground truth transcripts**, using the key rules: (i) if $\geq \frac{2}{3}$ of the annotators wrote down word w_i, then w_i was included in the ground truth transcript – this involved human judgement to align the word order (see following paragraphs for interrater disagreement), (ii) resolved spelling, (iii) resolved apostrophe style, and (iv) re-capitalized "I."

Interrater Disagreement. For the *scripted/non-scripted classification* task, we measure inter-rater *disagreement* as a simple percentage of the annotators who labeled podcast i differently (e.g., 2 annotators label podcast i as *scripted*, while 1 annotator labels podcast i as *non-scripted*), and we find a disagreement of $\kappa = 0\%$ for all of the labels for the 10 podcasts. We note that the 4 *scripted* podcasts have a mean duration of $18.15_{\pm 14.38}$ minutes, whereas the 6 *non-scripted* podcasts have a much larger mean duration of $37.03_{\pm 19.47}$ minutes – indicating that non-scripted podcasts tend to run longer, perhaps due to their conversational, and sometimes informal, nature.

For the *ground truth transcription* task, we measure the inter-rater *disagreement* between these 3 annotators using an aggregated difference measure. We calculate the difference using the built-in python library *difflib*, which uses a modified version of gestalt pattern matching [23]. We calculate the *percent difference* by averaging the pairwise differences amongst the 3 annotators, and normalizing based on the string length.[5]

[1] Creative Commons Attribution 4.0 International License

[2] Clifton et al. [18] originally used *langid.py* for language identification – we used WhisperX [16].

[3] This is not an easy task, as Shriberg describes: "[a] remarkable aspect of [disfluencies] is that they go largely unnoticed in everyday comprehension. Human listeners are so apt at filtering out [disfluencies] that the task of recording what was actually said in utterances containing [disfluencies] is a difficult and unnatural one, often requiring many passes at transcription" [2]. In Section 3.4.3 of her work, Shriberg notes that there were "[a] number of errors involved [in the] transcription of the filled pauses "uh" and "um". These were either missed entirely, or misplaced" in the original Switchboard transcriptions (which Shriberg then corrected in that work) [2, 21, 22]. Hence, our 2 minute transcriptions of 10 podcasts was indeed a significant effort.

[4] "I.D.K." stands for "I don't know." Hence, the [IDK] special token.

[5] Other interrater agreement metrics such as Cohen's κ and Fliess'

We find that the *scripted* podcasts have an average percent difference of $5.22_{\pm 3.12}$, whereas the non-scripted podcasts have a much larger percent difference $14.24_{\pm 5.94}$. This indicates that the annotators tended to agree *more* as to the transcription of the scripted podcasts, and *less* on the transcriptions of the non-scripted podcasts. We attribute this to the more conversational nature of the non-scripted podcasts.

3. Experiments

We first experiment on the 10 randomly-selected podcasts to evaluate the transcription quality of Google ASR and WhisperX, as compared to our human-annotated ground truth. Then, we further compare the two ASRs on the 82,601 podcasts.

3.1. Quality Measures

To comprehensively assess the quality of the transcriptions from Google ASR and WhisperX, we use automatic metrics at the *character-level* – Character Error Rate (CER), *word-level* – Word Error Rate (WER) and Word Information Lost (WIL), and *sentence-level* – BLEU, BERTScore, and ROUGE-L.[6]

Character Error Rate (CER) calculates the error rate of the ASR system: $CER = \frac{S+I+D}{N}$, where S is the number of character substitutions, D is the number of character deletions, I is the number of character insertions, and N is the total number of characters. **Word Error Rate (WER)** is calculated the same way as CER, however, at the word-level. **Word Information Lost (WIL)** is calculated as follows: $\text{WIL} = 1 - \frac{B}{G} * \frac{B}{N}$, where B is the number of words which correctly occur in both the ground truth and the ASR output, G is the number of words in the ground truth, and N is the number of words in the ASR output. **BLEU** is typically used for evaluating machine translation. We apply BLEU to order the "candidate translations" of Google ASR and WhisperX versus the Ground Truth [24]. **BERTScore** calculates the cosine similarity between two texts using BERT embeddings [25]. **ROUGE-L** calculates the longest common substring overlap between two texts [26].

3.2. Disfluency

Types of Disfluencies. We use a state-of-the-art disfluency annotation model [17] to obtain parse trees for the podcast transcripts. There are three types of nodes which are considered *disfluent nodes* in the parse trees: interjection nodes (e.g., *um*, *uh*), parenthetical nodes (e.g., *you know*, *I mean*), and edited nodes (e.g., *the store* in the phrase *let's go to the store, wait no, the movies today*). We additionally focus on two specific interjections, "*uh*" and "*um*", which are notable as they signal to listeners that a delay is about to occur, which aids in human speech processing [27, 28]. Additionally, insertion-type disfluencies (vs. deletion or substitution-type) are the most common type of disfluency amongst people with disfluency disorders [6].

Punctuation Preprocessing. In order to automatically obtain punctuation for the ground truth transcripts – as the annotators are instructed to focus on transcribing the words[3] – we employ ChatGPT [29][7] with 3 different prompts: *Add [maximal/minimal/] punctuation to the following text, do not remove any tokens, do not add "..." and keep the text in a single paragraph: {text}*.[8] We run the disfluency annotation model on all 3 versions of the transcript, and report the average of the results.

Disfluency Annotation Model. We employ a specialized parsing-based disfluency annotation model, *english-fisher-annotator*,[9] to obtain the parse trees for the Google ASR and WhisperX transcripts [17]. We use a top-down recursive approach to count the number of nodes of each disfluency type (INTJ, PRN, EDITED), taking the topmost disfluent label. The model performs multiclass classification (label l) on all string spans (from position i to position j), then scores the parse tree, $s(T)$, by summing $s(i, j, l)$ as follows:

$$s(T) = \sum_{(i,j,l) \in T} s(i, j, l). \quad (1)$$

From the parse trees, the highest-scoring parse tree ($\hat{T}$) is selected as the parse tree for that sentence using *argmax*.

We choose to use a parsing-based disfluency detection system [17, 30, 31] – rather than a translation-based system [32, 33] or other type of system – to compare the outputs of the two ASR systems using the fine-grained disfluency types (i.e., the relative quantities of edited, interjection, parenthetical nodes).

4. Results

4.1. RQ1: How does the choice of ASR system impact performance – across *scripted* and *non-scripted* podcasts?

In Table 1, we use automated metrics to analyze the similarity between (i) the Google ASR transcripts and the ground truth transcripts, and (ii) the WhisperX transcripts and the ground truth transcripts. Starting with the scripted podcasts, we observe that WhisperX obtains better performance across all of the metrics: at the character-level, word-level, and sentence-level. However, moving to the non-scripted podcasts, we notice that Google ASR outperforms WhisperX in two metrics: the word-level WIL, and the sentence-level BLEU score. Across all the podcasts – scripted and non-scripted – WhisperX shows superior performance for all the automated metrics.

Word Error Rate Comparison. For Google ASR, we find a WER of 10.47%, whereas Clifton et al. [18] reports a sample WER of 18.1% for Google ASR on the Spotify podcasts. For WhisperX, we find a WER of 7.19% on the Spotify podcasts, whereas, comparably, Whisper shows a WER of 13.8% on the Switchboard dataset [20, 34]. In both cases, the different error rates are due to (i) our sample size of 10 with the 2 minute limitation, and (2) our active focus on disfluency transcription.[3]

4.2. RQ2: How does the choice of ASR system impact the specific disfluency types which are transcribed?

4.2.1. Transcription of "uh" and "um"

In Table 2, we take two common interjections – "*uh*" and "*um*" – and we compare how much Google ASR and WhisperX transcribe these tokens as compared with the ground truth human-annotated transcriptions. Starting with the *scripted* podcasts, we see that there are no "*uh*" or "*um*" tokens contained in the

κ, are not suitable for measuring text agreement, as they require categorical ratings data.

[6] We use the *evaluate* library from HuggingFace and the *torcheval* library from PyTorch.

[7] The model version is *gpt-4-0125-preview*.

[8] We average the punctuation counts for these 3 prompts: overall the podcasts obtain $61.78_{\pm 5.37}$. The scripted podcasts $43.92_{\pm 3.79}$, and the non-scripted podcasts $54.63_{\pm 4.63}$. Hence, we find that the non-scripted podcasts (1) tend to receive more punctuation from ChatGPT, and (2) have a slightly larger variation in the punctuation amount.

[9] The model checkpoint is *swbd_fisher_bert_Edev.0.9078* from `github.com/pariajm/english-fisher-annotations`.

Table 1: *Character-level, word-level, and sentence-level metrics across the Google ASR and WhisperX transcripts, as compared to the ground truth human-annotated transcripts for the 10 podcasts.*

		Character-level	Word-level		Sentence-level		
		CER (↓)	WER (↓)	WIL (↓)	ROUGE-L (↑)	BERTScore (↑)	BLEU (↑)
Scripted	**Google ASR**	$3.46_{\pm 2.07}$	$7.39_{\pm 2.99}$	$15.02_{\pm 1.67}$	$93.83_{\pm 2.46}$	$97.66_{\pm 1.31}$	$85.09_{\pm 0.32}$
	WhisperX	$\mathbf{1.87}_{\pm 1.49}$	$\mathbf{3.36}_{\pm 1.37}$	$\mathbf{14.01}_{\pm 2.45}$	$\mathbf{97.41}_{\pm 0.93}$	$\mathbf{99.03}_{\pm 0.53}$	$\mathbf{86.24}_{\pm 1.12}$
Non-Scripted	**Google ASR**	$8.87_{\pm 5.95}$	$12.98_{\pm 6.96}$	$\mathbf{15.03}_{\pm 0.67}$	$90.48_{\pm 5.06}$	$96.29_{\pm 2.07}$	$\mathbf{84.85}_{\pm 1.56}$
	WhisperX	$\mathbf{6.05}_{\pm 3.77}$	$\mathbf{9.74}_{\pm 5.32}$	$15.32_{\pm 0.97}$	$\mathbf{93.34}_{\pm 3.29}$	$\mathbf{97.40}_{\pm 1.25}$	$84.71_{\pm 1.96}$
All	**Google ASR**	$6.71_{\pm 5.37}$	$10.47_{\pm 6.18}$	$15.02_{\pm 1.09}$	$91.82_{\pm 4.39}$	$96.84_{\pm 1.86}$	$84.95_{\pm 1.18}$
	WhisperX	$\mathbf{4.38}_{\pm 3.64}$	$\mathbf{7.19}_{\pm 5.21}$	$\mathbf{14.79}_{\pm 1.73}$	$\mathbf{94.97}_{\pm 3.27}$	$\mathbf{98.05}_{\pm 1.30}$	$\mathbf{85.32}_{\pm 1.78}$

ground truth transcriptions, nor either of the ASR transcriptions. This makes sense, as the podcasts are scripted.

Next, we look to the *non-scripted* podcasts. Here, we observe that WhisperX is closest to the ground truth, with mean values of 0.33 and 0.67, compared to the ground truth mean values of 1.67 and 1.33 for "*uh*" and "*um*" – which we note, are still much greater than the WhisperX mean values. Google ASR, however, is unable to transcribe these tokens, as the mean values are both 0. We observe the same trend with the aggregation of the scripted and the non-scripted podcasts: Google ASR does not transcribe the tokens, while WhisperX does, but less than the ground truth amount. We hypothesize that this may be due to the data WhisperX was trained on, as Whisper (which WhisperX is built on) was trained on human-generated transcript data [16, 20], and people often skip over disfluencies when transcribing them [2]. These findings indicate that there is a difference between scripted and non-scripted content, and that a defining characteristic of non-scripted content is the prevalence of disfluencies – specifically, "*uh*" and "*um*".

4.2.2. Transcription of INTJ, PRN, and EDITED Nodes

In Table 2, we compare the amount of interjection (INTJ), parenthetical (PRN), and edited (EDITED) nodes which were identified by the disfluency annotation model [17] for the transcripts produced by Google ASR and WhisperX, and the ground truth transcripts.[10] We focus on which of the ASR systems is *closer to the ground truth*, rather than the absolute number of each node which was transcribed – as we recognize (i) the disfluency annotation model is not perfect, and (ii) the disfluency annotation model was trained on Switchboard [22].

First, for the *scripted* podcasts, we notice that Google ASR's average number of interjection nodes, 1.00, is the same as in the ground truth, whereas WhisperX averages 0.75 interjection nodes per scripted podcast. The two ASRs tie in transcribing parenthetical nodes. For the edited nodes, WhisperX is closer to the ground truth. For the *non-scripted* podcasts, we notice that the case is the opposite: WhisperX transcribes closer to the number of ground truth interjection nodes, while Google ASR transcribes closer to the ground truth number of edited nodes. Thus, these results suggest that it may be beneficial to select an ASR based on the distribution of disfluent node types present in the data to maximize downstream performance.

4.3. RQ3: Are these findings consistent at a large-scale?

In Table 3, we compare the two ASRs' performance at a larger scale. In the absence of ground truth transcripts, we compare their relative performance to each other, and their performance in the small-scale experiment. First, we notice that WhisperX again transcribes, on average, more "*um*" and "*uh*" tokens than Google ASR. Next, we notice that, again, WhisperX also transcribes on average more interjections, while Google ASR's transcriptions, on average, tend to result in more edited nodes. However, the two ASR systems differ in their mean number of parentheticals, with Google ASR transcribing more parentheticals than WhisperX. Again, at the larger scale, these results suggest that it may be beneficial to select an ASR based on the distribution of disfluent node types present in the data.

Table 2: *The mean count of "um" and "uh" tokens, and INTJ, PRN, and EDITED nodes for the 10 podcasts.*

		$C_{\text{"uh"}}$	$C_{\text{"um"}}$	C_{INTJ}	C_{PRN}	C_{EDITED}
Scripted	**Ground Truth**	0	0	$1.00_{\pm 0.82}$	$0.25_{\pm 0.50}$	$0.58_{\pm 0.81}$
	Google ASR	0	0	$\mathbf{1.00}_{\pm 0.82}$	**0**	$1.50_{\pm 1.91}$
	WhisperX	0	0	$0.75_{\pm 0.96}$	**0**	$\mathbf{0.75}_{\pm 1.50}$
Non-Scripted	**Ground Truth**	$1.67_{\pm 1.97}$	$1.33_{\pm 1.21}$	$9.06_{\pm 6.81}$	$2.00_{\pm 2.38}$	$5.33_{\pm 4.25}$
	Google ASR	0	0	$6.33_{\pm 5.32}$	$\mathbf{2.17}_{\pm 2.93}$	$\mathbf{5.33}_{\pm 2.50}$
	WhisperX	$\mathbf{0.33}_{\pm 0.82}$	$\mathbf{0.67}_{\pm 0.82}$	$\mathbf{7.83}_{\pm 6.40}$	$\mathbf{2.17}_{\pm 2.40}$	$3.67_{\pm 2.73}$
All	**Ground Truth**	$1.00_{\pm 1.70}$	$0.80_{\pm 1.14}$	$5.83_{\pm 6.59}$	$1.30_{\pm 2.02}$	$3.43_{\pm 4.04}$
	Google ASR	0	0	$4.20_{\pm 4.85}$	$\mathbf{1.30}_{\pm 2.45}$	$\mathbf{3.80}_{\pm 2.94}$
	WhisperX	$\mathbf{0.20}_{\pm 0.63}$	$\mathbf{0.40}_{\pm 0.70}$	$\mathbf{5.00}_{\pm 6.04}$	$\mathbf{1.30}_{\pm 2.11}$	$2.50_{\pm 2.68}$

Table 3: *The mean count of "um" and "uh" tokens, and INTJ, PRN, and EDITED nodes for the 82,601 podcasts.*

	$C_{\text{"uh"}}$	$C_{\text{"um"}}$	C_{INTJ}	C_{PRN}	C_{EDITED}
Google ASR	$0.09_{\pm 0.35}$	$0.25_{\pm 0.70}$	$48.02_{\pm 37.12}$	$\mathbf{12.71}_{\pm 11.29}$	$\mathbf{30.26}_{\pm 13.71}$
WhisperX	$\mathbf{1.38}_{\pm 3.03}$	$\mathbf{1.69}_{\pm 3.14}$	$\mathbf{50.90}_{\pm 39.88}$	$10.84_{\pm 9.79}$	$16.71_{\pm 9.58}$

5. Impact and Conclusion

In conclusion, we find that in terms of automated metrics, WhisperX tends to perform best – however, there are a few metrics in specifically the non-scripted case where Google ASR tends to obtain better performance. In terms of disfluency, WhisperX is overall better at transcribing interjection nodes, and in the large-scale case also parenthetical nodes, whilst Google ASR is better at transcribing edited nodes. These results suggest that it may be beneficial to select an ASR based on the distribution of disfluent node types present in the data.

6. Ethics and Limitations

Annotators. Two of the annotators are student volunteers, and one is an author who also received course credit. The annotators risked exposure to the content in the 10 randomly-selected podcasts. The three annotators were managed by the annotation coordinator, also an author of this work.

Limitations. In this work, we study 2 ASR systems, limit our human annotations to 2 minutes, and sample 10 of the podcasts for our analysis. Future work could expand these aspects.

[10] We release the disfluent token distributions on the project website: `https://www.comparing-asr-systems.com/disfluent-token-distributions`.

7. Acknowledgments

We thank Anwesha Basu and Eric Nunes for serving as volunteer annotators for the project. We also thank Haoran Liu and Zhouer Wang for their feedback and comments on the work.

8. References

[1] E. Diachek and S. Brown-Schmidt, "The Effect of Disfluency on Memory for What Was Said," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 49, no. 8, pp. 1306–1324, 2022.

[2] E. Shriberg, "Preliminaries to a Theory of Speech Disfluencies," Ph.D. dissertation, 1994.

[3] H. Branigan, R. Lickley, and D. McKelvie, "Non-linguistic influences on rates of disfluency in spontaneous speech," in *International Conference of Phonetic Sciences*, 1999, p. 387–390.

[4] E. Shriberg, "Disfluencies in switchboard," in *International Conference on Spoken Language Processing*, vol. 96, 1996, pp. 11–14.

[5] J. F. Tree, "The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech," *Journal of Memory and Language*, vol. 34, no. 6, p. 709–738, 1995.

[6] V. Mitra, Z. Huang, C. Lea, L. Tooley, S. Wu, D. Botten, A. Palekar, S. Thelapurath, P. Georgiou, S. Kajarekar, and J. Bigham, "Analysis and tuning of a voice assistant system for dysfluent speech," in *Interspeech*, 2021.

[7] National Public Radio and Edison Research, "The smart audio report," 2022. [Online]. Available: https://www.nationalpublicmedia.com/insights/reports/smart-audio-report/

[8] M. Teleki, X. Dong, and J. Caverlee, "Quantifying the Impact of Disfluency on Spoken Content Summarization," in *Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, 2024.

[9] A. F. Wise, R. Martinez-Maldonado, I. Hilliger, M. Xia, Y. Zhao, M. H. Erol, J. Hong, and J. Kim, "Understanding distributed tutorship in online language tutoring," in *International Learning Analytics and Knowledge Conference*, 2022, p. 164–174.

[10] R. Schmucker, M. Xia, A. Azaria, and T. Mitchell, "Ruffle&Riley: Towards the Automated Induction of Conversational Tutoring Systems," in *Neural Information Processing Systems*, 2023.

[11] J. Paladines and J. Ramirez, "A systematic literature review of intelligent tutoring systems with dialogue in natural language," *IEEE Access*, vol. 8, p. 164246–164267, 2020.

[12] E. Diachek and S. Brown-Schmidt, "The effect of disfluency on memory for what was said." *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 49, no. 8, p. 1306, 2023.

[13] C. Lea, Z. Huang, J. Narain, L. Tooley, D. Yee, D. T. Tran, P. Georgiou, J. P. Bigham, and L. Findlater, "From user perceptions to technical improvement: Enabling people who stutter to better use speech recognition," in *Conference on Human Factors in Computing Systems*, 2023, p. 1–16.

[14] Z. Qian and K. Xiao, "A Survey of Automatic Speech Recognition for Dysarthric Speech," *Electronics*, vol. 12, no. 20, p. 4278, 2023.

[15] Google Cloud, "Speech-To-Text: Automatic Speech Recognition," 2024. [Online]. Available: https://cloud.google.com/speech-to-text

[16] M. Bain, J. Huh, T. Han, and A. Zisserman, "WhisperX: Time-Accurate Speech Transcription of Long-Form Audio," in *Interspeech*, 2023.

[17] P. J. Lou and M. Johnson, "Improving Disfluency Detection by Self-Training a Self-Attentive Model," in *Association for Computational Linguistics*, 2020, p. 3754–3763.

[18] A. Clifton, S. Reddy, Y. Yu, A. Pappu, R. Rezapour, H. Bonab, M. Eskevich, G. Jones, J. Karlgren, B. Carterette, and R. Jones, "100,000 Podcasts: A Spoken English Document Corpus," in *Conference on Computational Linguistics*, 2020, pp. 5903–5917.

[19] S. Reddy, M. Lazarova, Y. Yu, and R. Jones, "Modeling language usage and listener engagement in podcasts," in *Association for Computational Linguistics Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and International Joint Conference on Natural Language Processing*, 2021.

[20] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*, 2023.

[21] M. Mitchell, B. Santorini, M. A. Marcinkiewicz, and A. Taylor, "Treebank-3 LDC99T42 Web Download," 1999. [Online]. Available: https://catalog.ldc.upenn.edu/LDC99T42

[22] A. Taylor, M. Marcus, and B. Santorini, "The Penn Treebank: An Overview," *Treebanks: Building and Using Parsed Corpora*, pp. 5—22, 2003.

[23] J. W. Ratcliff and D. E. Metzener, "Pattern Matching: the Gestalt Approach." 1988, [Online]. Available: https://www.drdobbs.com/database/pattern-matching-the-gestalt-approach/184407970?pgno=5

[24] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A Method for Automatic Evaluation of Machine Translation," in *Association for Computational Linguistics*, 2002, p. 311–318.

[25] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," in *International Conference on Learning Representations*, 2020.

[26] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," in *Association for Computational Linguistics*, 2004.

[27] H. H. Clark and J. E. F. Tree, "Using uh and um in spontaneous speaking," *Cognition*, vol. 84, no. 1, pp. 73–111, 2002.

[28] J. E. F. Tree, "Listeners' Uses of Um and Uh in Speech Comprehension," *Memory & Cognition*, vol. 29, no. 2, pp. 320–326, 2001.

[29] OpenAI, "ChatGPT," 2023. [Online]. Available: https://cdn.openai.com/papers/gpt-4.pdf

[30] N. Kitaev and D. Klein, "Constituency parsing with a self-attentive encoder," in *Association for Computational Linguistics*, 2018.

[31] P. J. Lou, Y. Wang, and M. Johnson, "Neural constituency parsing of speech transcripts," in *North American Chapter of the Association for Computational Linguistics*, 2019.

[32] S. Wang, Z. Wang, W. Che, S. Zhao, and T. Liu, "Combining self-supervised learning and active learning for disfluency detection," *Transactions on Asian and Low-Resource Language Information Processing*, vol. 21, no. 3, p. 1–25, 2021.

[33] Q. Dong, F. Wang, Z. Yang, W. Chen, S. Xu, and B. Xu, "Adapting translation models for transcript disfluency detection," in *AAAI Conference on Artificial Intelligence and Innovative Applications of Artificial Intelligence Conference and AAAI Symposium on Educational Advances in Artificial Intelligence*, 2019.

[34] Linguistic Data Consortium. (2002, 1) 2000 HUB5 English Evaluation Transcripts. [Online]. Available: https://catalog.ldc.upenn.edu/LDC2002T43