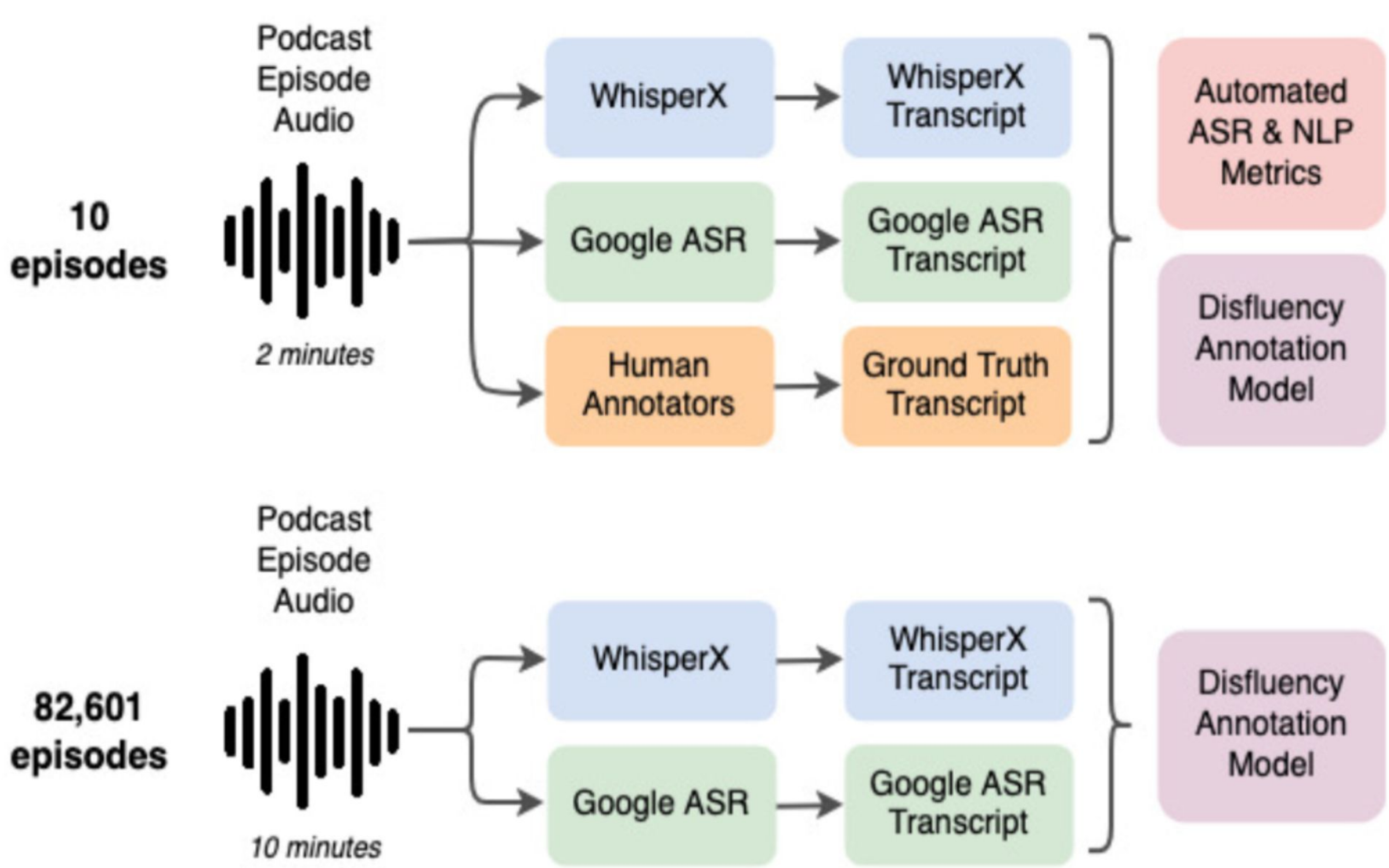
Comparing ASR Systems in the Context of Speech Disfluencies

Maria Teleki, Xiangjue Dong, Soohwan Kim, James Caverlee
Texas A&M University

Introduction & Experimental Settings



- We evaluate the disfluency capabilities of two ASR systems – WhisperX [1] and Google ASR [2] – in terms of their interactions with a parsing-based disfluency annotation model [3].
  - Why? It's natural for application developers to plug an ASR-created transcript into a disfluency removal model.
  - 4-6% of non-scripted speech is disfluent [4].
  - We use [3] to annotate the **3 types of disfluencies**:
    - (i) interjections (**INTJ**) – *ex: let's go to the uh store today*
    - (ii) parentheticals (**PRN**) – *ex: let's to go the store, wait no, the movies today*
    - (iii) edited nodes (**EDITED**) – *ex: let's to go the store, wait no, the movies today*
- We use the **Spotify Podcasts Dataset** [5] for our analysis.
  - We obtain ground truth transcripts via human annotations (N=3).

RQ1 *(10 episodes)*: How does the choice of ASR system (WhisperX, Google ASR) impact performance (as compared to the human-annotated ground truth)?

- While WhisperX performs better overall in terms of automated metrics, Google ASR outperforms WhisperX in WIL and BLEU for specifically non-scripted 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	$\mathbf{4.38}_{\pm 3.64}$	$\mathbf{7.19}_{\pm 5.21}$	$\mathbf{14.79}_{\pm 1.73}$	$\mathbf{94.97}_{\pm 3.27}$	$\mathbf{98.05}_{\pm 1.30}$	$\mathbf{85.32}_{\pm 1.78}$

RQ2 *(10 episodes)*: How does the choice of ASR system (WhisperX, Google ASR) impact the specific disfluency types which are transcribed (as compared to the human-annotated ground truth)?

- WhisperX transcribes closer to the ground truth number of *uhs*, *ums*, and INTJ nodes than Google ASR.
  - The ground truth number of *uh*s and *ums* is higher.
- Google ASR transcribes closer to the ground truth number of EDITED nodes than WhisperX.
- WhisperX and Google ASR transcribe the same number of PRN nodes.

		$C_{\text{"uh"}}$	$C_{\text{"um"}}$	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	$\mathbf{0.20}_{\pm 0.63}$	$\mathbf{0.40}_{\pm 0.70}$	$\mathbf{5.00}_{\pm 6.04}$	$\mathbf{1.30}_{\pm 2.11}$	$2.50_{\pm 2.68}$

RQ3 *(82,601 episodes)*: Are these findings consistent at a large scale?

- Google ASR transcribes hardly any *uhs* or *ums*.
- Same trend as small-scale in RQ2: WhisperX transcribes more INTJ nodes, while Google ASR transcribes more EDITED nodes – however Google ASR transcribes more PRN nodes.
  - We hypothesize this is due to the *vocabulary diversity of WhisperX versus that of Google ASR.*

	$C_{\text{"uh"}}$	$C_{\text{"um"}}$	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	$\mathbf{1.38}_{\pm 3.03}$	$\mathbf{1.69}_{\pm 3.14}$	$\mathbf{50.90}_{\pm 39.88}$	$10.84_{\pm 9.79}$	$16.71_{\pm 9.58}$

References

[1] M. Bain, et al., "WhisperX: Time Accurate Speech Transcription of Long-Form Audio," in Interspeech, 2023.
[2] Google Cloud, "Speech-To-Text: Automatic Speech Recognition," 2024. [Online]. Available: https://cloud.google.com/ speech-to-text
[3] P. J. Lou and M. Johnson, "Improving Disfluency Detection by Self-Training a Self-Attentive Model," in ACL, 2020.
[4] E. Shriberg, "Preliminaries to a Theory of Speech Disfluencies," Ph.D. dissertation, 1994.
[5] A. Clifton, et al., "100,000 Podcasts: A Spoken English Document Corpus," in COLING, 2020.

Conclusion

These results suggest that it may be **beneficial to select an ASR system based on the distribution of disfluent node types** present in the data.