UNSPOKEN SOUND: IDENTIFYING TRENDS IN NON-SPEECH AUDIO CAPTIONING ON YOUTUBE

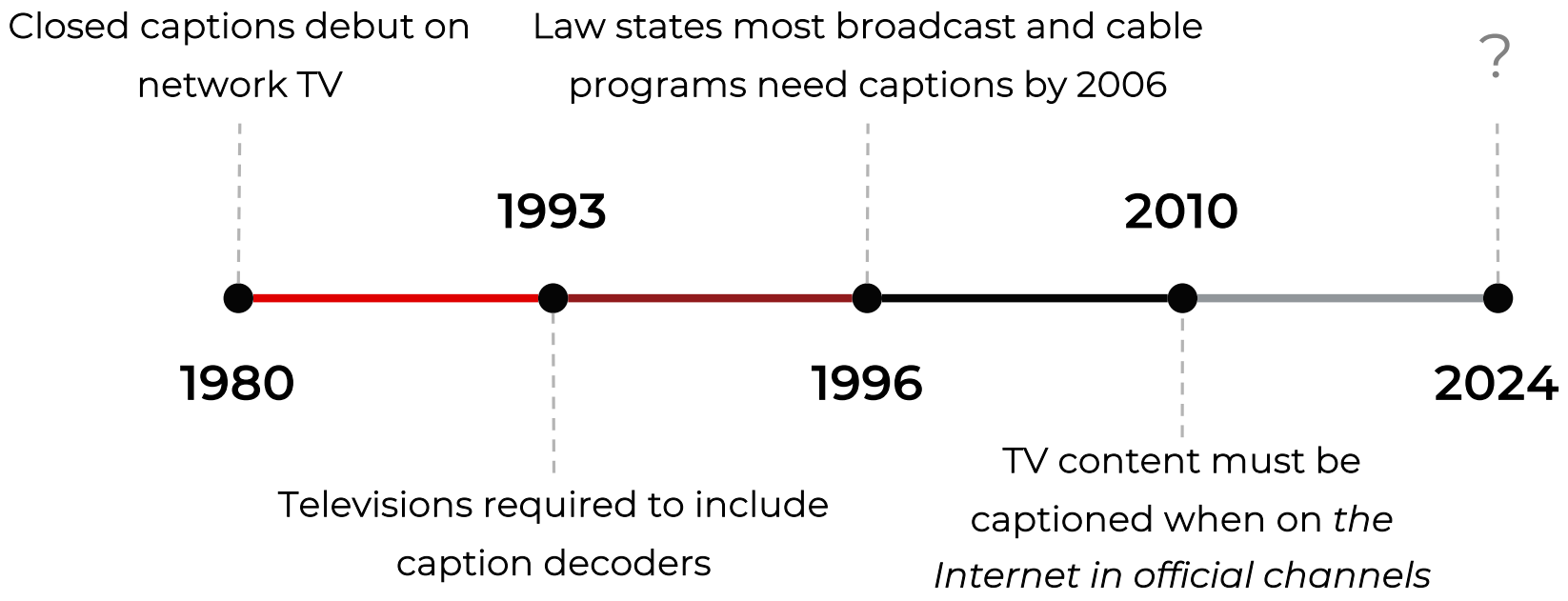
Lloyd May, Keita Ohshiro, Khang Dang, Sripathi Sridhar, Jhanvi Pai,
Magdalena Fuentes, Sooyeon Lee, *Mark Cartwright*

WHAT ARE CLOSED CAPTIONS?

Closed captions communicate a video's audio content with text, providing critical information for

- d/Deaf or hard-of-hearing (DHH) people
- hearing people who are in a situation (e.g., public place) where they cannot listen to the audio of video

HISTORY OF CLOSED CAPTIONING



...BUT WHAT ABOUT EVERYTHING ELSE?

For example,

?

- 500+ hours of video uploaded to YouTube every minute
- No caption requirements for this content
- YouTube mitigates this problem by using automatic speech recognition (ASR) to recognize *speech* in videos

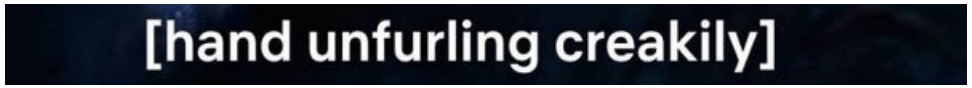
CAPTIONS *NOT* JUST TRANSCRIPTIONS OF SPEECH

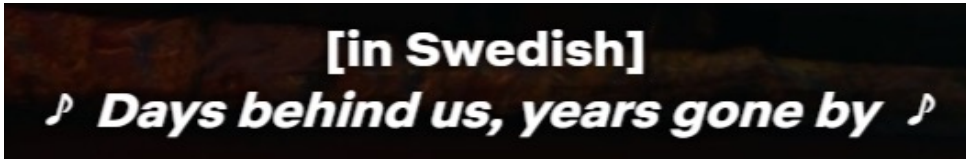
Non-speech information (NSI) captions convey information about sound in addition to simply what words were spoken.

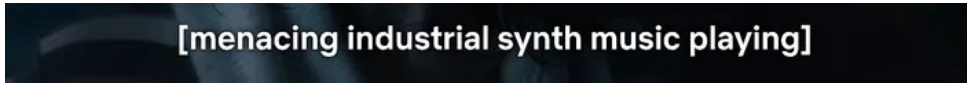
Types of NSI:

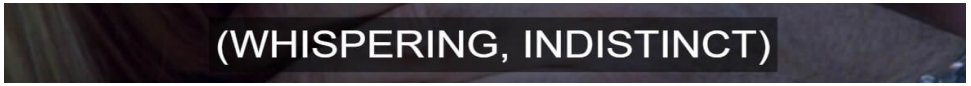
- Sound effects
- Environmental sounds
- Music
- Additional narrative information
- Extra-speech information
  - manner of speech e.g. “[Whispering] Oh no”)
  - speaker label (e.g. “[Juan] Oh no”)

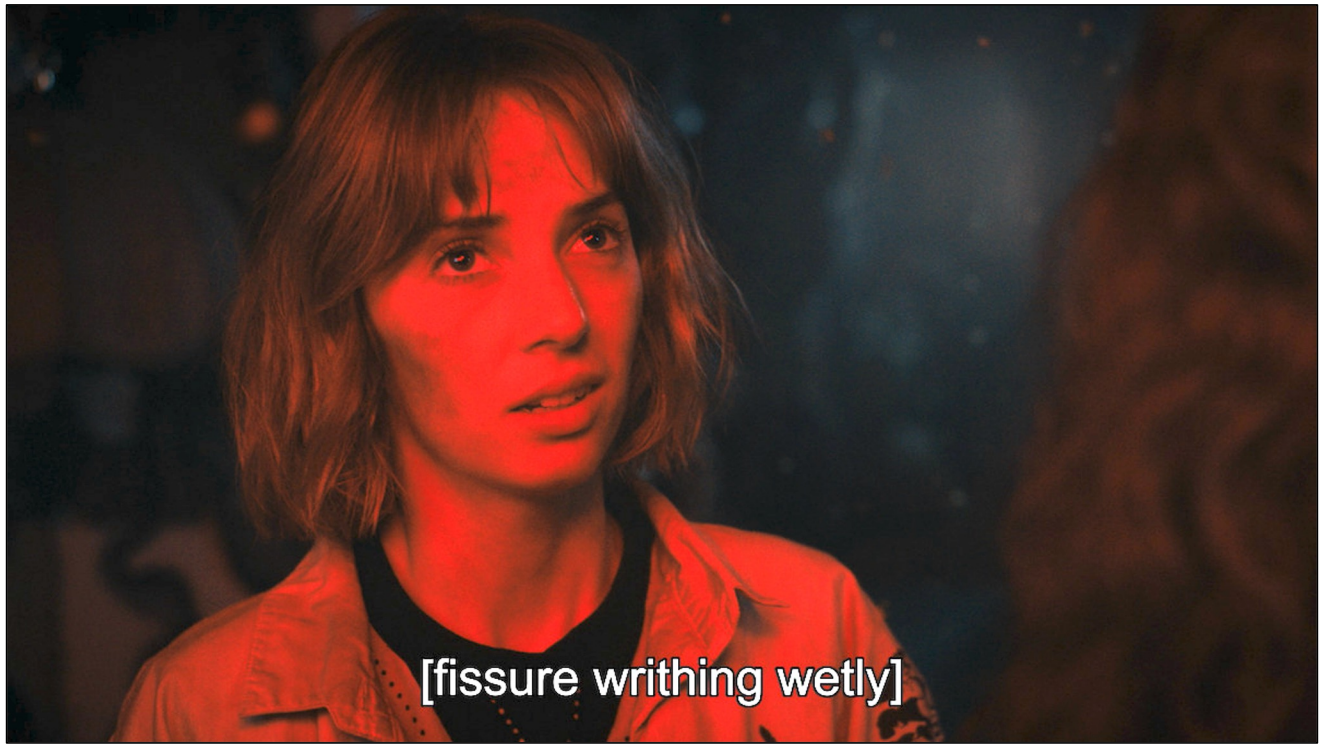
NSI CAPTIONING EXAMPLES











Example caption from *Stranger Things*


WIGAN
1-1
MAN UTD
80:42
PREMIERE LIVE
AIG


WIGAN
1-1
MAN UTD
80:42
PREMIERE LIVE
[Sneezes]

SUMMARY OF MOTIVATION

- NSI captions convey vital information about non-speech sound
- YouTube and other streaming platforms host an immense amount of video content whose captions are not regulated
- But automatic captioning solutions provided by YouTube focus on just the spoken words, not NSI

  (...with the exception of simple NSI [APPLAUSE], [MUSIC], [LAUGHTER} since 2017)

... so, how much of this important non-speech information is actually captioned?

RESEARCH QUESTIONS

1. What is the current and historical prevalence of non-speech information captioning on YouTube?
2. What factors may affect non-speech information captioning practices on YouTube?

METHODS

ANATOMY OF A YOUTUBE VIDEO

	One Category	News
28 Dec 2023	Upload Date	View Count 3,247 views
	Captions	Manual

CONSTRUCTING THE DATASET

Two samples from YouTube :

Popular Videos

- 2013–2022
- Most popular videos each month from each YouTube category

Studio Videos

- 2013–2022
- Most popular videos each month from channels run by large studios (value > $1B)

RESULTING DATASET

715k
Videos

273M
Lines of captions

6M
Estimated lines of NSI captions

ESTIMATING NSI

: < | ( ) [ ] “” > ♩ ♬ ♫ ♪ ♭ ♮ ♯ #

MANUAL ANNOTATION

Popular Videos	
300	2013
300	2018
300	2022

Studio Videos	
300	2013
300	2018
300	2022

1,800 Videos in Total

ANNOTATING NSI CAPTIONS

NSI Labels:

- Music
- Sound Effects
- Extra-Speech Information
  (Speaker Label, Manner of Speech)

Other Labels:

- Quoted Speech
- Not NSI
- Not English
- Additional Information

MEASURES OF NSI IN CAPTIONS

Annotated subset:

- NSI presence
- NSI count per minute if present (NSI CPMIP)

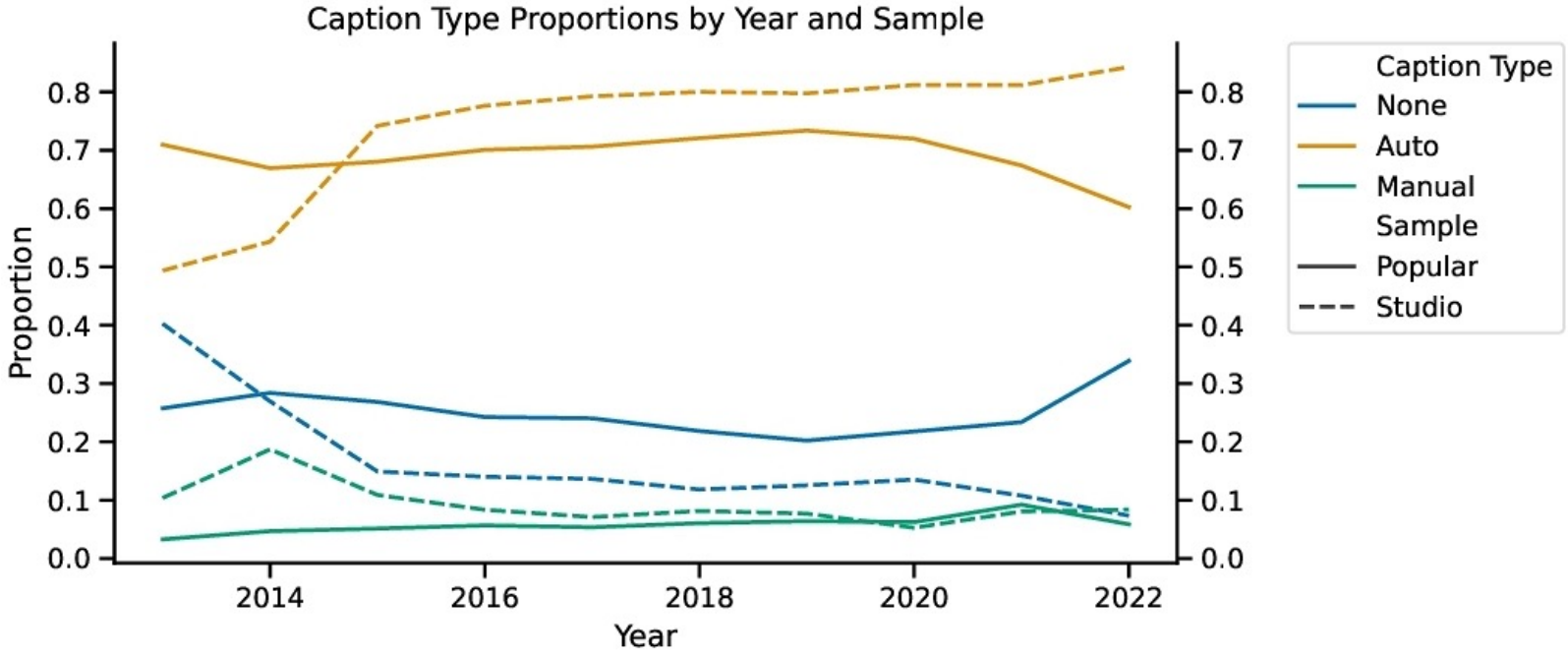
Full dataset:

- Estimated NSI presence
- Estimated NSI count per minute if present (Estimated NSI CPMIP)

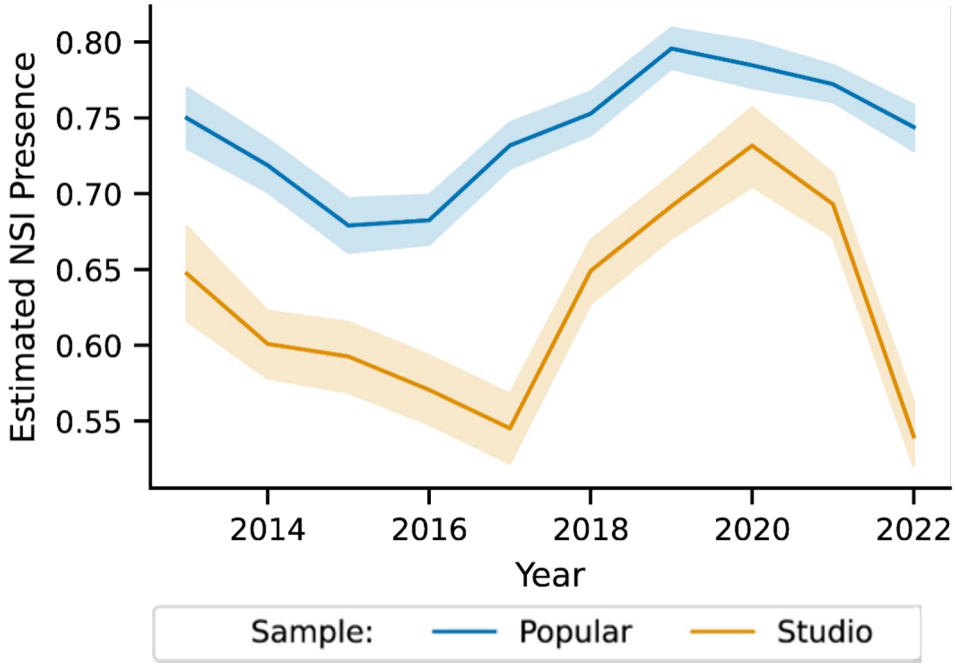
Precision of estimated NSI is ~0.9

RESULTS

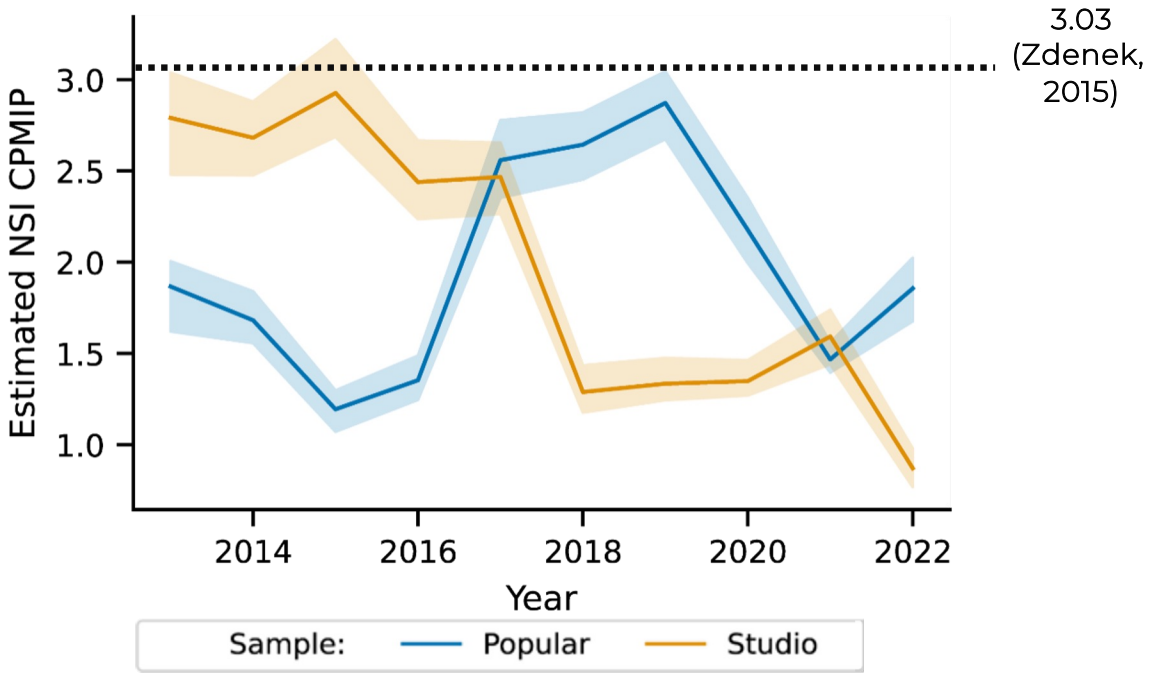
LESS THAN 10% OF VIDEOS HAVE MANUAL CAPTIONS



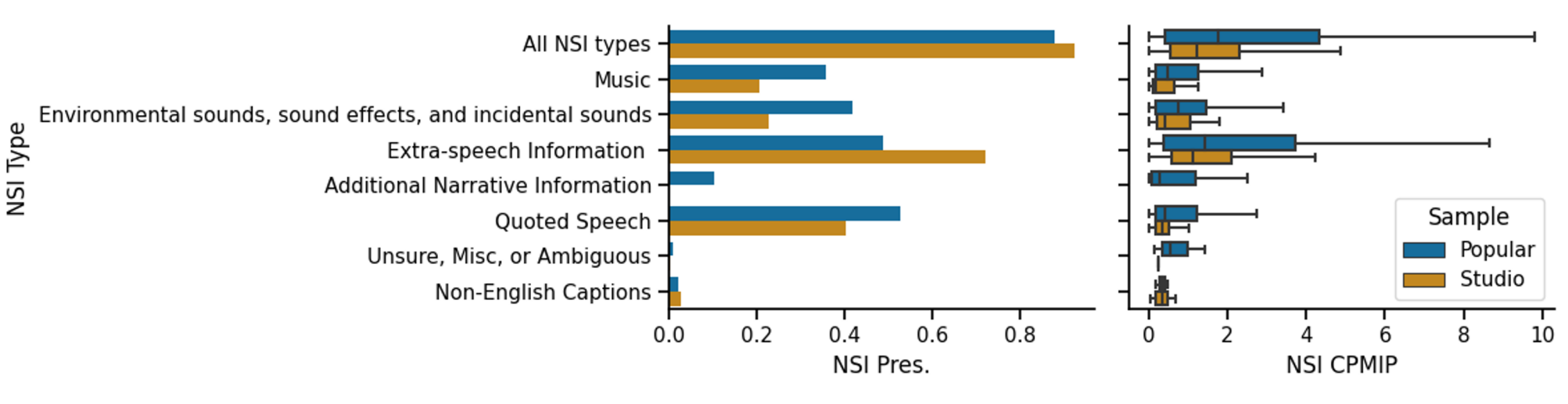
MOST MANUAL CAPTIONS CONTAIN NSI



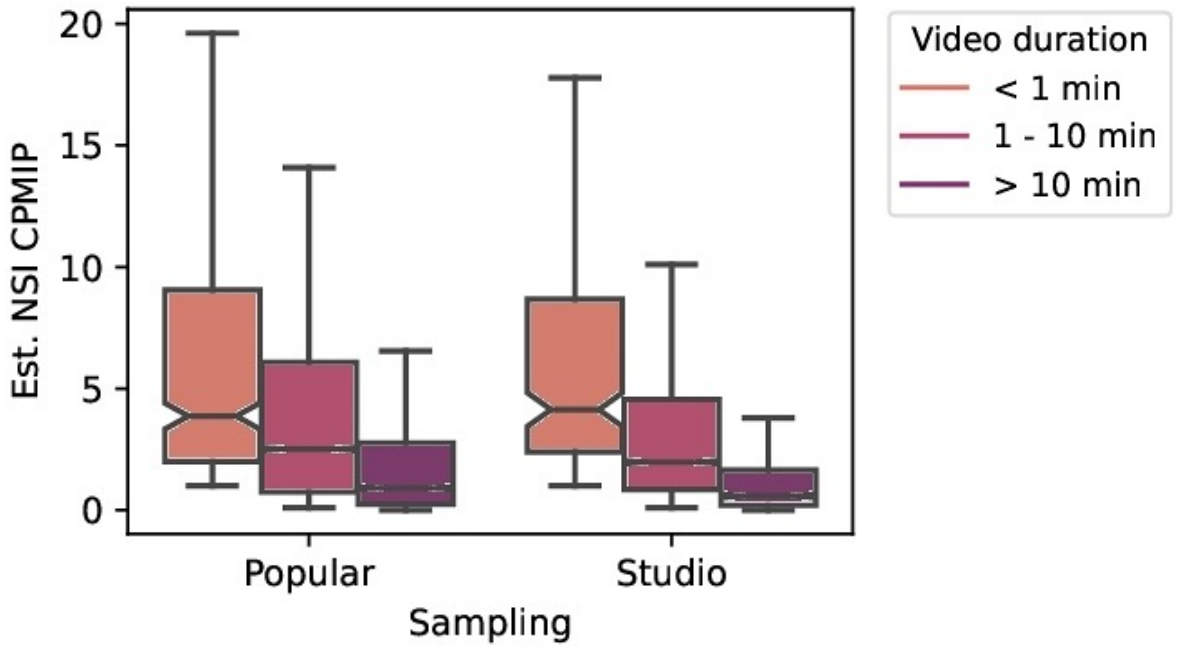
NSI CAPTION DENSITY IS LOW AND DECREASED IN STUDIO SAMPLE IN LAST DECADE



LESS MUSIC / SFX IN STUDIO NSI CAPTIONS



LONGER VIDEOS HAVE LESS NSI CAPTIONS PER MIN.



SUMMARY OF RESULTS

- Presence of manually-annotated NSI is quite low
- The density of manually-annotated NSI is lower than expected when it is present
- Video duration influences NSI caption density
- There are differences in large studio captioning practices, possibly due to increase reliance on ASR + editing in ‘manual’ captioning pipelines

RECOMMENDATIONS

- Develop more nuanced automatic NSI captioning systems
- Increase ease of manual editing of automatically generated captions
- Improve media processing pipelines and caption alignment for large studios
- Improve evaluation of NSI captioning

DATASET AVAILABLE ON ZENODO