3PLAY MEDIA

THE STATE OF AUTOMATIC SPEECH RECOGNITION (ASR)

✋ Type questions in the Q&A window during the presentation

 This webinar is being recorded & will be available for replay

 To view live captions, please click the CC icon

 www.3playmedia.com | @3playmedia | #a11y

HELLO!

Elisa Lewis

Content Marketing Manager

- Passionate about web accessibility
- Love dogs and crafting

Tessa Kettelberger

Research and Development Engineer

AGENDA

- Introduction
- Focus on Innovation

- The Research and Testing Process
- Research Findings
- Examples of ASR
- Key Takeaways for Your Business
- Questions

INTRO

- What is the Annual State of ASR report?
- Why do we conduct this research?

Research investigates the current state of automatic speech recognition technology with regard to captioning accuracy.

Published annually as the State of ASR report to share improvements in ASR.

With 3Play's focus on innovation, the data and findings allow us to improve our process.

FOCUS ON INNOVATION

What does this mean? What is 3Play Media's relationship with innovation?

We have 11 patents on our processes, and use machine learning and artificial intelligence (AI) heavily in everything we do.

THE UNIQUE CHALLENGE OF CAPTIONING

Variety of environments

Music, background noise, number of speakers

Variety of subjects

Cannot constrain vocabulary, focus on a topic

Length

Long video & audio content, not short commands with immediate feedback

Readability

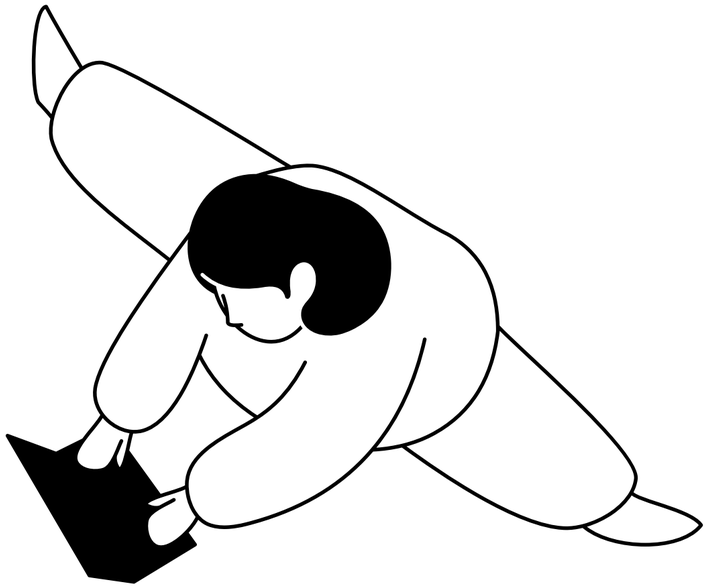
Captions and transcripts are consumed by humans and need to be understandable

Timing

Captions are time-aligned

RESEARCH

The study tested 6 of the most relevant ASR technologies as well as our own ASR service, which uses Speechmatics with custom post-processing.



3Play Media
ASR available through SMX API V1 with 3Play post-processing

Speechmatics +
ASR available through SMX API V2 with no post-processing

Speechmatics
ASR available through SMX API V1 with no post-processing

IBM

Microsoft

Google

Rev

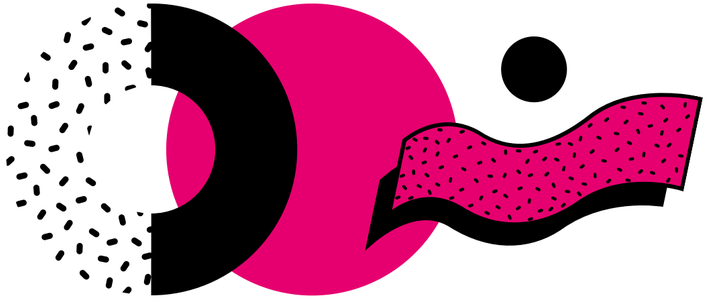
VoiceGain

TESTING

All testing used real content, representative of the content that we receive at 3Play.

We tested:

- 490 files
- 65 hours
- 670,000 words



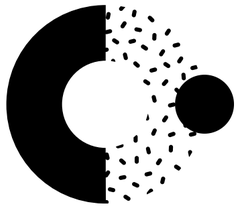
The breakdown of data by primary industry is:

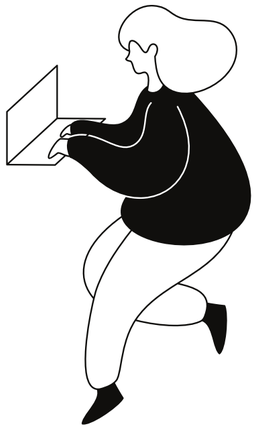
- 28% Education
- 16% Online video
- 15% Entertainment
- 13% Other
- 12% eLearning
- 12% Corporate
- 2% Government
- 1.5% Fitness
- > 1% Societies & Associations
- > 1% Faith

Additional diversity of content:

The duration, number of speakers, audio quality, and speaking style (e.g. scripted vs. spontaneous) varies greatly across this data.

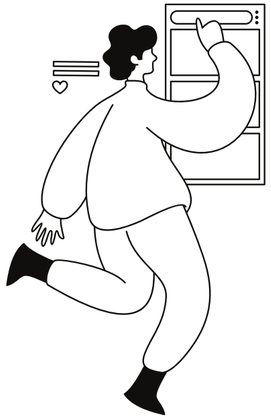
POLL TIME!





WER V. FER

When it comes to captioning accuracy, it's important to consider both Formatted Error Rate (FER) and Word Error Rate (WER).

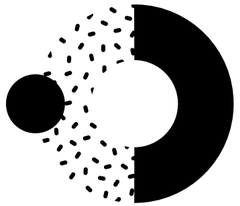


Word Error Rate

- Takes into account when the word is incorrect.

Formatting Error Rate

- Takes into account errors that relate to formatting elements such as punctuation, grammar, speaker identification, non-speech elements, capitalization, and other notations



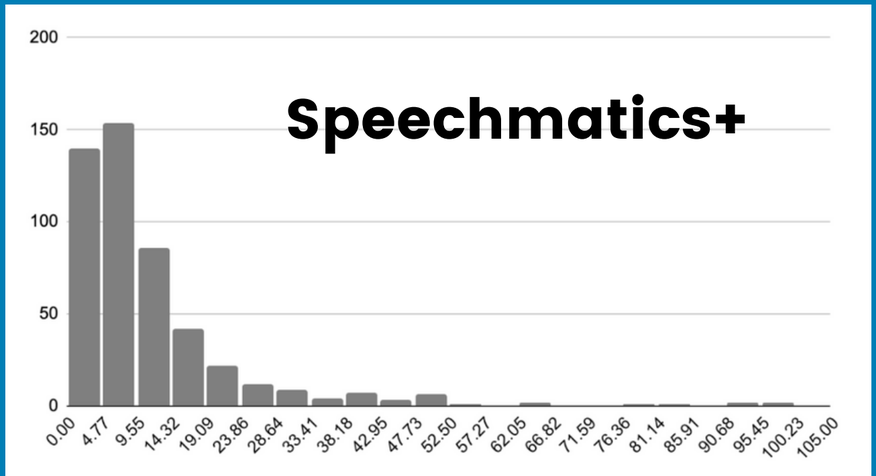
WORD ERROR RATES

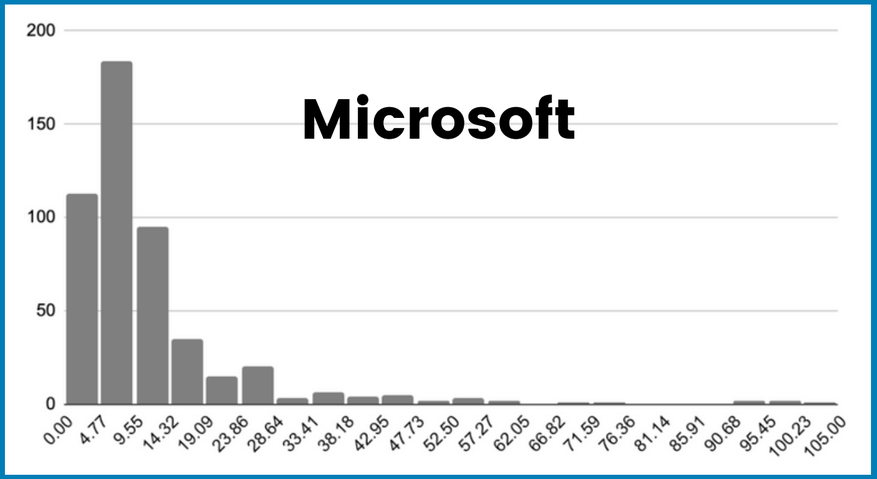
	ERR	CORR	SUB	INS	DEL
3PM	13.1	90.2	5.93	3.32	3.88
SMX	14.1	89.8	6.33	3.90	3.88
SMX+	13.0	90.3	5.93	3.22	3.82
IBM	26.3	80.1	12.8	6.37	7.11
MIC	13.4	90.1	5.83	3.48	4.08
GOO *	20.9	85.1	7.13	6.01	7.75
REV	15.3	90.2	5.65	5.50	4.15
VG	15.8	86.6	8.28	2.36	5.12

FORMATTING ERROR RATES

	ERR	CORR	SUB	INS	DEL
3PM	24.9	78.0	16.1	2.91	5.94
SMX	27.2	76.7	19.5	3.82	3.80
SMX+	24.7	78.5	17.8	3.15	3.74
IBM**	41.8	64.4	28.7	6.16	6.91
MIC	25.7	77.8	18.3	3.41	4.01
GOO*	36.1	69.8	22.6	5.89	7.63
REV	26.4	79.0	16.9	5.42	4.07
VG	28.2	74.1	20.9	2.79	5.05

HOW GOOD CAN ASR BE?





These graphs show the distribution of Word Error Rate on each individual file for the two best-scoring technologies that we tested.
The numbers we've been discussing are averages. There is a lot of variance in performance based on audio quality, duration, and content.

KEY FINDINGS

Our research data allows us to draw several conclusions about the current state of ASR.

We're seeing exciting improvements in many of the technologies we tested this year

Most improvement is driven by improved data and training, rather than breakthroughs in ASR technology itself.

The success of ASR is heavily dependent on audio quality and content difficulty

We are confident that we are using the best technology available for our problem

In terms of FER, no one is providing an output close to sufficient for compliance.

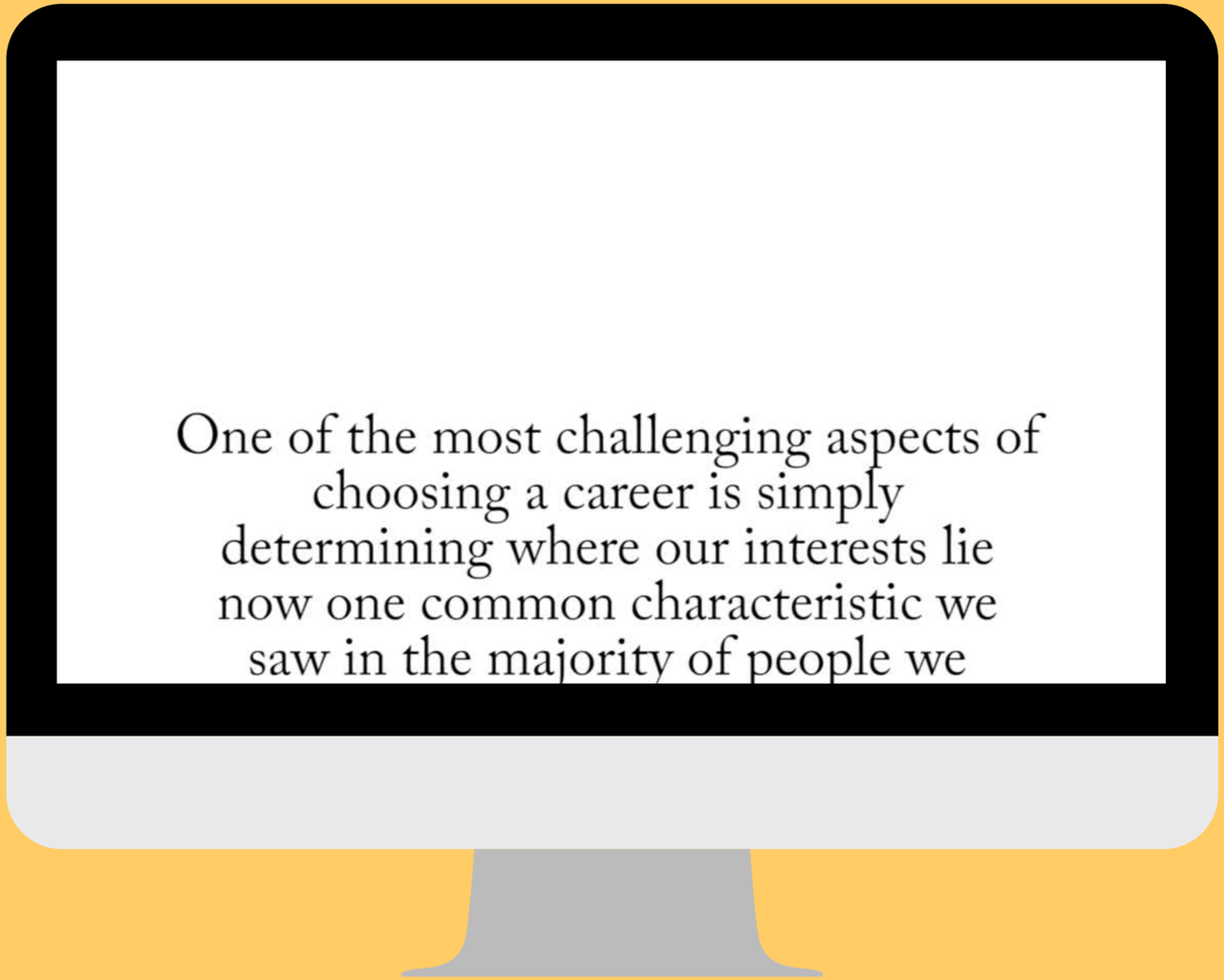
WHAT THIS MEANS FOR YOU

While technology continues to improve, there is still a significant leap to real accuracy from even the best speech recognition engines, making humans a crucial part of creating accurate captions.

Let's take a look at some examples.

ASR EXAMPLE

Type in the chat window: What errors do you notice?



ASR ERRORS

Common ASR errors include:

- Speaker labels
- Punctuation
- Grammar
- Relevant non-speech elements
- No [INAUDIBLE] tags
- Acoustic errors
- “Function” words

FER

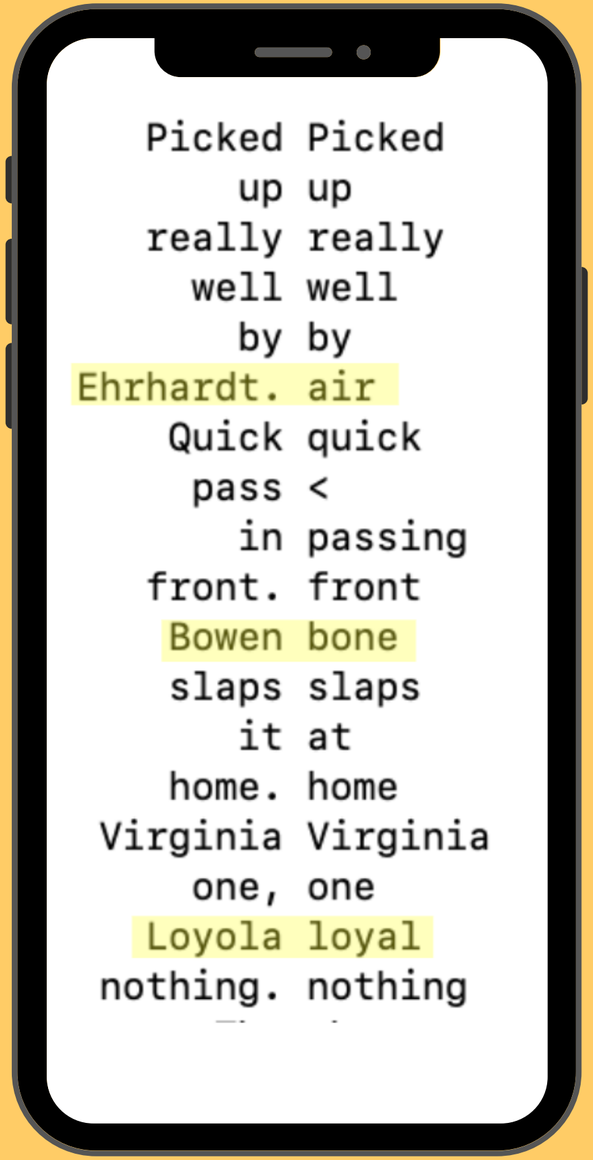
Punctuation and capitalization are crucial to relaying the correct message.

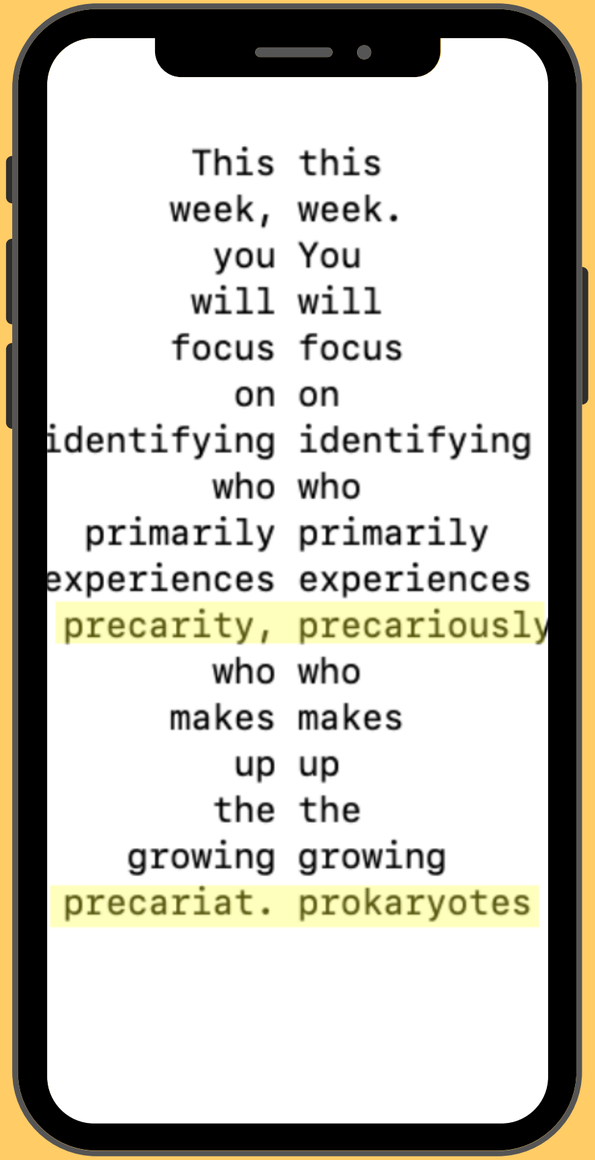
And to Grandma's safety!

ASR EXAMPLE

Common causes of ASR errors include:

- Multiple speakers or overlapping speech
- Background noise
- Poor audio quality
- False starts
- Complex vocabulary





LAST YEAR V. THIS YEAR

We have identified several noteworthy findings around the state of ASR in 2020 as compared to 2019.

SMX+ performed the best compared with the other speech engines, at an accuracy rate of 90.3%.

Speechmatics V1 (SMX) paired with our 3Play Media post-processing (3PM) followed closely.

Speechmatics showed a 7% reduction in WER from V1 (SMX) to V2 (SMX+) thus leading us to move to SMX+.

TAKEAWAYS

The application of captioning is unique in regard to AI and automatic speech recognition technology.

Several improvements in technology and training capabilities in the last year.

The best ASR systems can achieve accuracy rates in the high ‘80s and into the low ’90s.

When it comes to FER, none of the solutions we tested are sufficient alone.

WHAT'S NEXT?

Explain the pricing method for each variation of your product or service.

State of Captioning Report Coming Soon

WBN: Global Outlook for A11y Compliance - March 18

Allied Podcast Launch

Coming Soon

THANK YOU

Questions?