Research Transcriptions

How Safe is AI For Qualitative Research?

2024

Research Transcriptions

How Safe is AI For Qualitative Research?

See real data that shows where AI can be used effectively, where it can threaten results, and where confidentiality is risked.

by Rob Foley

I. Introduction

In recent years, Automated Speech Recognition (ASR) technology has made significant strides, promising to revolutionize how we transcribe and analyze spoken language. For professionals engaged in qualitative research, the allure of ASR is strong: it offers the potential to dramatically reduce the time and cost associated with transcribing interviews and focus groups. However, despite technological advancements, ASR remains an unreliable tool for qualitative research, potentially compromising the integrity of results across all fields.

This report aims to explore the current state of ASR technology, its limitations, and the specific dangers it poses to qualitative research. By examining recent data, comparing different ASR services, and considering the unique requirements of qualitative analysis, we will demonstrate why researchers should approach ASR with caution and skepticism.

II. Background on Automated Speech Recognition (ASR)

ASR technology has come a long way since its inception in the 1950s and 60s when innovators were using speech recognition systems to identify people speaking digits out loud (Carter, 2023). Modern ASR systems use complex machine learning algorithms, particularly deep neural networks, to convert spoken language into text. Recent developments in end-to-end (E2E) models have further improved ASR capabilities, simplifying the transcription pipeline and enabling larger training datasets through self or semi-supervised learning (Chung et al. 2021; Xu et al. 2021; Zhang et al. 2022).

Despite these advancements, ASR still faces significant challenges in accurately transcribing natural speech, especially in real-world conditions. Background noise, accents, speaking speed, and specialized vocabulary can all impact ASR performance.

III. Measuring Transcription Accuracy

The most common metric for evaluating transcription accuracy is the Word Error Rate (WER). WER represents the minimum edit distance between a transcription and the reference solution, quantifying the relative amount of errors to the total number of words in a text (Woodard and Nelson 1982).

But WER has limitations. **It doesn't reflect text understanding and only weakly correlates with human judgment of a transcript's quality** (Favre et al. 2013; Mishra et al. 2011; Wang et al. 2003). Alternative measures like Character Error Rate (CER) and Automated-Caption Evaluation (ACE) have been proposed, but they, too, have drawbacks (Kafle and Huenerfauth 2017; Wells et al. 2022).

IV. Current State of ASR Accuracy

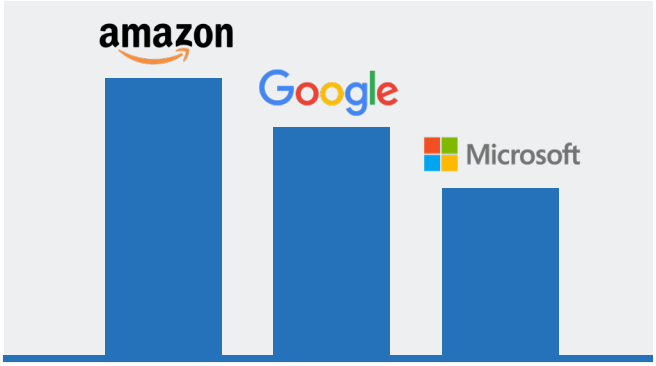
Recent studies paint a complex picture of ASR accuracy. While some researchers report WERs as low as 2.5% on isolated datasets (Meta AI 2023), real-world performance can vary significantly.

A comprehensive study by Dubois et al. (2024) found significant variations in accuracy across different ASR platforms and speaker characteristics.

Key findings from recent research include:

1. Accuracy varies widely between vendors and individual audio samples (Dubois et al. 2024).
2. There's a significant drop in quality for streaming ASR, which is crucial for live events (Dubois et al. 202 a4).
3. **Gender bias:** ASR performs differently for male and female speakers (Koenecke et al. 2020).
4. **Racial disparities:** some systems show almost twice as high WER for Black American speakers compared to White American speakers (Koenecke et al. 2020).
5. **Non-native speakers and those with regional accents experience lower accuracy rates** (Tadimeti et al. 2022; Cumbal et al. 2021).

Benchmarks published in 2021 found that the leading ASR systems continue to struggle with accuracy:



- Amazon's speech-to-text technology had an error rate of 18.42%
- Microsoft's error rate was 16.51%
- Google Video was ranked at 15.82% (Carter, 2023)

These error rates are significant, especially for qualitative research where nuance and exact wording are crucial.

V. ASR in Challenging Audio Environments

Loakes (2024) conducted a study comparing various ASR systems, including OpenAI's Whisper, on both good-quality and poor-quality (forensic-like) audio. The results are telling:

1. For good-quality audio, most systems performed well, with error rates between 0 and 18%.
2. **For poor-quality audio, even the (Whisper) had a WER of 50%. best-performing system**
3. Other systems performed significantly worse, with some recognizing only a fraction of the spoken words.

These findings highlight the significant challenges ASR faces in real-world research settings, where audio quality may be less than ideal.

IV. ASR in Qualitative Research: Challenges and Risks

For qualitative researchers, verbatim transcription is crucial. Every word, pause, and nuance can be significant in analysis. ASR's limitations pose several risks:

1. **Misinterpretation of Data:**
   Transcription errors can lead to misinterpretation of partici pants' responses, potentially skewing research findings
2. **Loss of Nuance:**
   ASR may miss subtle cues in speech such as tone, emphasis, or hesitation, which are often crucial for proper punctuation to obtain accurate qualitative analysis.
3. **Bias Amplification:**
   The biases present in ASR systems (e.g., lower accuracy for certain accents or demographics) can lead to systematic errors in data collection, potentially amplifying existing biases in research.

VII. Case Study: Zoom's Transcription Feature

Zoom, a popular platform for conducting remote interviews and focus groups, offers an ASR feature. However, it has **several limitations** that make it unsuitable for rigorous qualitative research:

1. **Accuracy:**
   Zoom's automatic closed captioning is generally estimated to be around 80% accurate for typical use cases, but actual performance can vary significantly depending on the context and audio quality. This general estimate suggests that one in every five words could be incorrect in a typical scenario.
2. **Study-Based Accuracy:**
   A more specific study by Waddell (2020) found that Zoom's auto-captions had an average error rate of about 8 per 100 words. This translates to approximately **92% accuracy** which is better than the general estimate **but still problematic** for research purposes.
3. **Bias:**
   The study found a **3.6% average accuracy gap between native and non-natives peakers when using Zoom's auto-captions (Waddell, 2020)**
4. **Inconsistency:**
   Waddell's study revealed significant **vairiability in Zoom's transcription accuracy. At it's best, Zoom had just two errors per 100 words, while at it's worst, it mistranscribed nearly every third word (Waddell, 2020**
5. **Speaker Identification:**
   The saved transcript doesn't automatically identify individual speakers, though it may identify them if they introduce themselves or are referenced by name (Waddell, 2020).
6. **Limited Features:**
   Transcripts lack features for summarizing or creating action plans, and can only be downloaded in VTT format (Waddell, 2020).
7. **Accessibility:**
   The feature isn't available with free Zoom accounts, and enabling it can be challenging (Waddell, 2020).
8. **Language Support:**
   Zoom offers limited language support (Waddell, 2020).
   These limitations highlight the risks of relying on general-purpose ASR tools like Zoom for research-grade transcript tion.

Concerns about Zoom Transcription:

* Bias toward native English speakers
* Cannot distinguish speakers
* Unreliable accuracy rates

VIII. Accuracy of Other ASR Tools

Major ASR players aiming to promote their technology have come out with studies showing accuracy rates over 90%. However, these figures are proven to be inflated when tested in independent studies.

1. **Google Cloud Speech-to-Text:**
   Google claims a word accuracy rate of up to 95% for certain use cases. However, independent studies have shown lower accuracy rates, especially for non-standard accents and noisy environments (Carter, 2023).
2. **Amazon Transcribe:**
   Amazon reports accuracy rates of up to 90% for certain scenarios. However, in the study by Loakes (2024), Amazon Transcribe performed poorly with forensic-like audio, transcribing only a small fraction of the total words.
3. **Microsoft Azure Speech to Text:**
   Microsoft claims up to 95% accuracy rates for certain languages and scenarios. However, the accuracy can drop significantly in real-world tests, especially with non-native speakers or noisy environments (Carter, 2023).
4. **IBM Watson Speech to Text:**
   IBM reports accuracy rates of up to 90% for certain use cases. However, like other systems, its performance can vary widely depending on audio quality and speaker characteristics.
5. **AssemblyAI:**
   AssemblyAI claims to make 43% fewer errors on noisy data than other systems. However, specific accuracy rates for different scenarios are not publicly available (Carter, 2023).

6. **OpenAI's Whisper:**
   While WER performed best among the systems tested by Loakes (2024), it still had a 50% Word Error Rate with poor-quality audio. This means half of the words were either incorrect or missing.

It's important to note that these accuracy rates are often based on ideal conditions and may not reflect real- world performance, especially in challenging research environments.

IX. Implications for Different Research Fields

Business and Marketing Research:
Misinterpreting consumer feedback due to transcription errors could lead to misguided business decisions or ineffective marketing strategies.

Healthcare and Medical Research:
Errors in transcribing patient interviews or focus groups could result in misunderstanding symptoms, experiences, or treatment effects, potentially impacting patient care or drug development.

Social Sciences and Education:
Inaccurate transcription could misrepresent participants' views or experiences, compromising the validity of studies on social phenomena or educational outcomes.

Healthcare and Medical Research:
Errors in transcribing patient interviews or focus groups could result in misunderstanding symptoms, experiences, or treatment effects, potentially impacting patient care or drug development.

Social Sciences and Education:

Inaccurate transcription could misrepresent participants' views or experiences, compromising the validity of studies on social phenomena or educational outcomes.

X. Ethical Considerations

Using ASR in qualitative research raises several ethical concerns:

1. **Consent and Privacy:**
   Participants may not be aware that their words are being processed by AI, raising questions about informed consent.
2. **Responsibility for Errors:**
   When transcription errors lead to misinterpretation, who bears responsibility – the researcher or the ASR provider?
3. **Accessibility:**
   While ASR can make research more accessible to deaf and hard-of-hearing researchers and participants, its inaccuracies may create new barriers.

XI. Alternative Approaches and Best Practices

Given the risks associated with ASR, researchers should consider alternative approaches:

1. **Human Transcription:**
   While more time-consuming and expensive, professional human transcription remains the gold standard for accuracy in qualitative research.
2. **Hybrid Approaches:**
   Using ASR for initial transcription followed by human editing can balance efficiency and accuracy. However, this approach still carries risks if errors are missed during editing. Furthermore, the time and cost may make human transcription the more sensible option from the start.
3. **Guidelines for ASR Use:**
   If ASR must be used, researchers should implement strict quality control measures, including:

   - Manual review of all transcripts
   - Clear documentation of the ASR system used and it'sknown limitations
   - Transparency about the use of ASR in research methods and potential impacts on findings

XII. Future of ASR in Qualitative Research

While ASR technology continues to improve, it's unlikely to fully replace human transcription in qualitative research any time soon. Ongoing developments in natural language processing and machine learning may address some current limitations, but the nuanced nature of qualitative data will likely continue to require human oversight.

Potential improvements in ASR that could benefit qualitative research include:

- Better handling of accents and non-native speech
- Improved speaker diarization (identifying who is speaking)
- More accurate capture of paralinguistic features (tone, emphasis, etc.)

Even as these improvements materialize, researchers must remain critical and vigilant about using ASR in qualitative studies.

XIII. Conclusion

While Automated Speech Recognition offers tantalizing benefits in terms of speed and cost-efficiency, its current limitations make it a dangerous tool for qualitative research. The risks of misinterpretation, data loss, and bias amplification are too high to ignore.

Recommendations for researchers:

Prioritize accuracy over speed and cost when it comes to transcription. If using ASR, implement rigorous quality control measures and be transparent about its use.

Stay informed about developments in ASR technology and its limitations. Consider the ethical implications of using ASR in research, particularly regarding consent and accessibility. As Loakes (2024) concludes, "While the results of this study, for Whisper in particular, are a marked improvement in performance compared to the systems trialled on the same audio in Loakes (2022), this study advocates for the use of human transcription done in a measured and systematic manner." For researchers prioritizing accuracy and confidentiality, services like Research Transcriptions offer an alternative to ASR. They specialize in 100% human transcription based in the US, ensuring that AI never touches the data. Their approach combines field-specific transcription specialists with a proprietary accuracy process, aiming to capture the nuances that AI may miss.

As ASR technology evolves, it's crucial that the research community continues to critically evaluate its appropriateness for qualitative studies. The integrity of our research and the voices of our participants must always take precedence over technological convenience. Whether opting for advanced ASR systems or human transcription services, researchers must carefully weigh the trade-offs between efficiency, accuracy, and the specific needs of their qualitative studies.

Contact Research Transcriptions to learn about accurate human transcription.