API vs. human: Comparing the performance of speech-to-text transcription using voice answers from a smartphone survey

Höhne[1,2] & Lenzner[3]

[1] Leibniz University Hannover

[2] German Center for Higher Education Research and Science Studies (DZHW)

[3] GESIS – Leibniz Institute for the Social Sciences

General Online Research (GOR) Conference

Cologne (Germany) – February 21 to 23, 2024

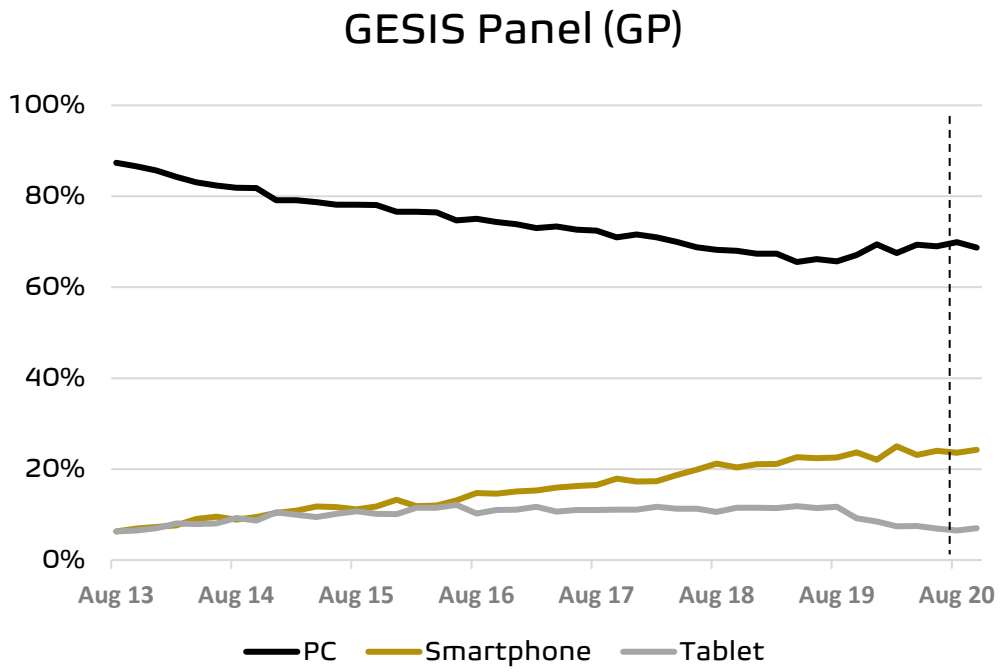
Digitalization and Research Potential

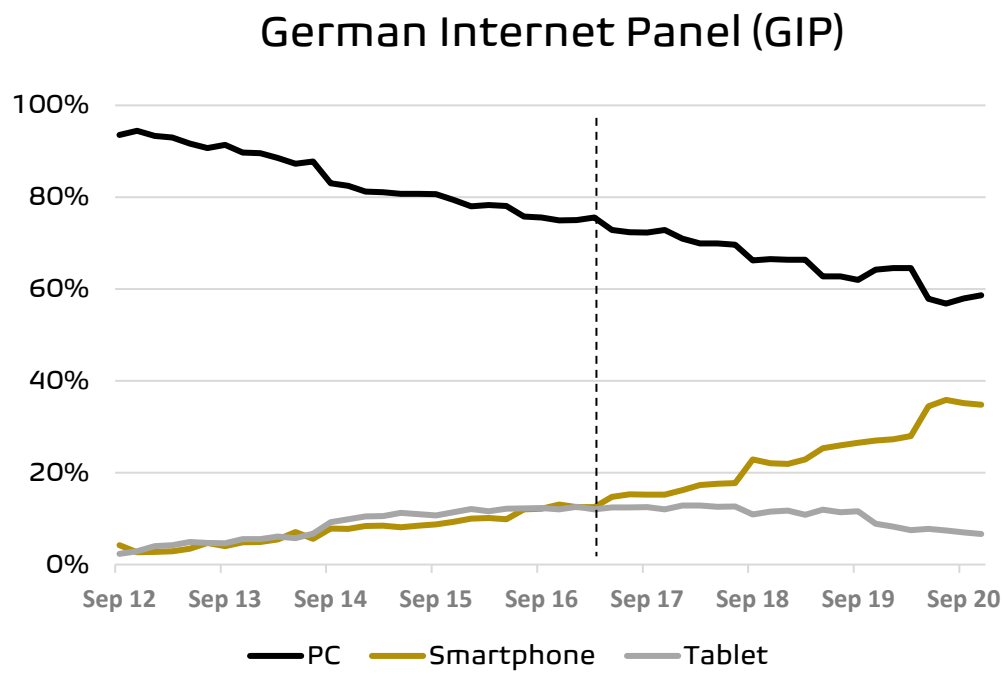
- Global digitalization tendency
  - *Increase in internet use* (Pew Research Center 2016, 2019a)
  - *Increase in smartphone ownership* (Pew Research Center 2019b)
- New opportunities for researching (social) reality
  - *People leave traces and produce data in digital spheres* (Struminskaya et al. 2020)
- Transformation of social and behavioral sciences
  - *New conferences: BigSurv and Mobile Apps and Sensors in Surveys*
  - *New journals: Frontiers in Big Data and Journal of Computational Social Science*

Web Surveys and Digital Innovations

- Increase of web-based surveys
  - *Academia: ANES, CRONOS, EVS, GESIS Panel, GIP, HRS, LISS Panel etc.*
  - *Public/private sector: Facebook, Google, UNESCO, World Bank etc.*
- Increase of mobile device use in web-based surveys
  - *Mobile optimized layouts as default* (Revilla et al. 2016)
- Emergence of digital intersections
  - *Ex ante data linkage (e.g., sensors)* (Elevelt et al. 2021; Höhne & Schlosser 2019)
  - *Ex post data linkage (e.g., trace data)* (Pasek et al. 2020; Stier et al. 2020)

Devices in Web Surveys





Country: Germany. Prob-based online panels. Vertical lines indicate the introduction of mobile-optimized layouts. Calculations: Gummer et al. (2023).

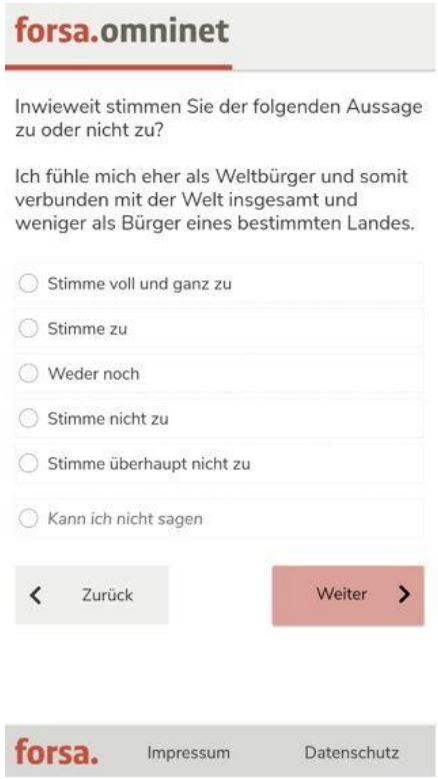
Smartphones and Voice Answers

- New communication channels because of smartphones
  - *Linking established methods with technological innovations*
- Voice answer to (open) questions
  - *Closeness to daily conversation* (Tourangeau et al. 2000)
  - *Rich information due to narrations* (Gavras & Höhne 2022; Gavras et al. 2022)
- Technological requirements of voice answers are met
  - *Even in web surveys with large N*
- Nonetheless, answer transcription is still required
  - *Human transcription is burdensome and time consuming*
  - *APIs may not be entirely ready*

Research Questions (RQs)

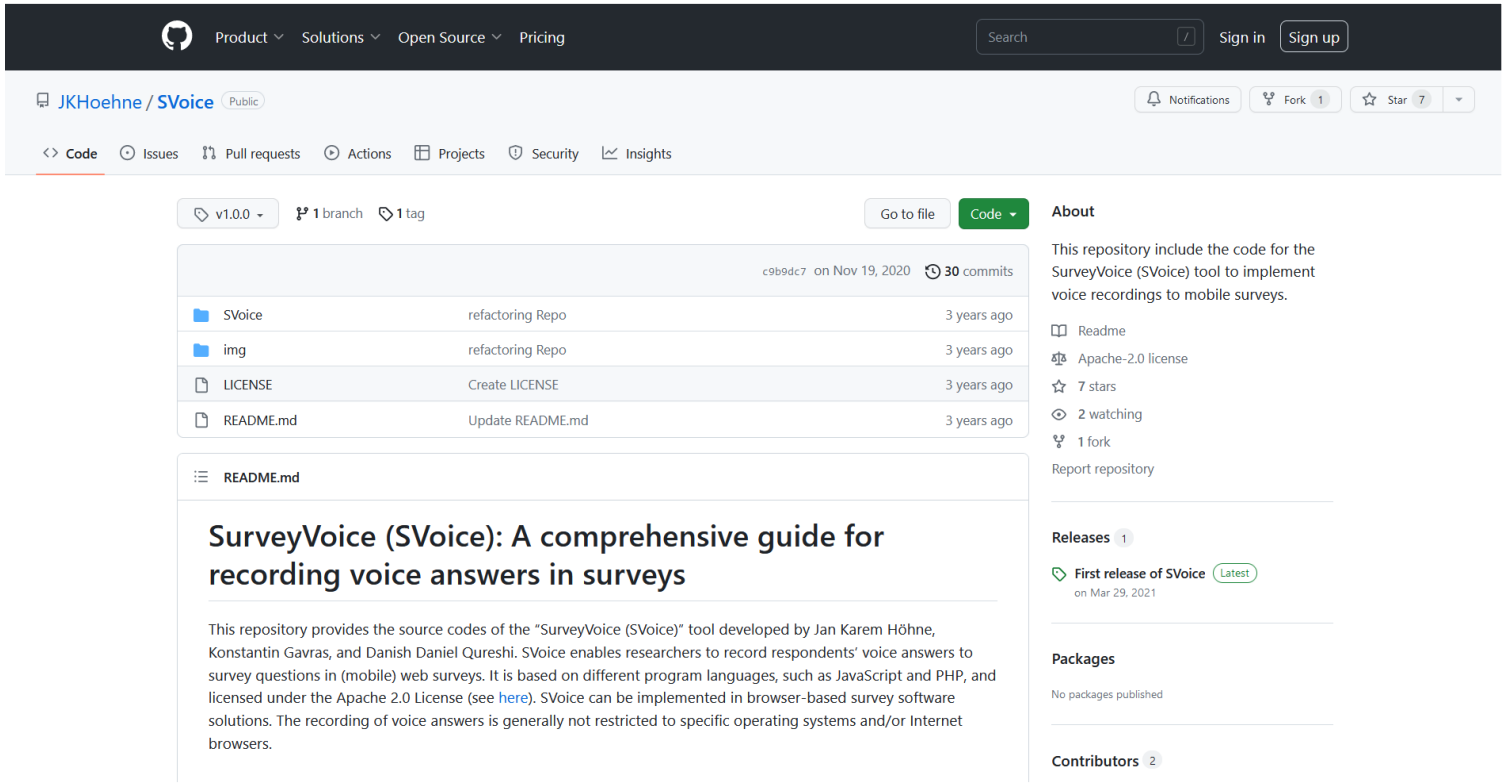
- RQ1: What is the transcription quality of APIs?
- RQ2: What types of errors occur in API transcription?
- RQ3: How long does transcription by APIs and humans take?

Method: Study Design



- Cross-quota sample
  - *Age and gender (3x2) plus education (3)*
- 2 questions plus probes
  - *Relationships between citizens and state (ISSP 2013, 2014)*
  - *Comprehension probes*
- No recording time restrictions
  - *Overall, we have 609 voice answers for analysis*
  - *These answers vary between 1 and 295 seconds*
- Extended replication study (Lenzner & Neuert 2017)

Method: Collecting Voice Data



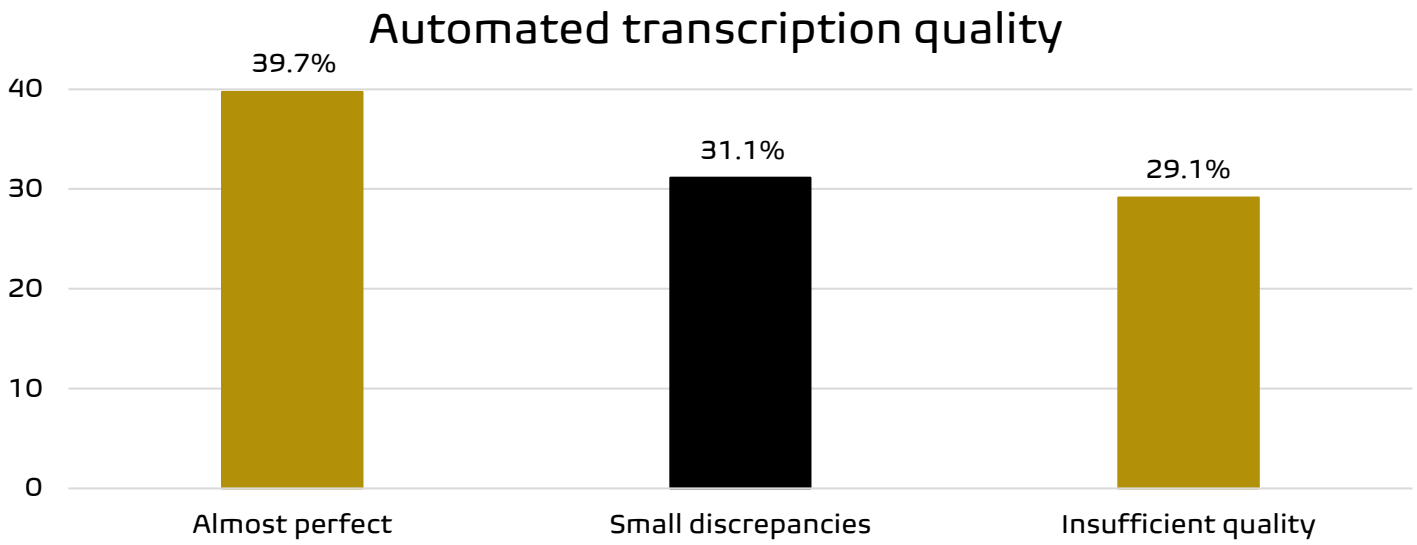
See https://github.com/JKHoehne/SVoice/tree/v1.0.0

- SurveyVoice (Svoice) tool (Höhne et al. 2021)
- Open-Source
  - *Apache 2.0 license*
- JavaScript, CSS, HTML, and PHP
- Implementable in browser-based smartphone surveys

Method: Analytical Strategy

- We used the Google "Cloud Speech-to-Text" API
  - *Costs: $0.024 per minute (without data logging – standard)*
- Transcription quality
  - *1 Almost perfect, 2 small discrepancies (minor errors), and 3 insufficient quality (major errors)*
  - *Intercoder reliability: Agreement > 90% (Kappa > 0.84)*
- Transcription error types
  - *1 No mistakes, 2 misspellings, 3 word separation error, 4 word transcription error, 5 missing words, 6 incorrect grammatic, 7 words added by mistake, and 8 words replaced by numbers*
  - *Intercoder reliability: Agreement > 82% (Kappa > 0.78)*

Results: Research Question 1



We have less than 1% of poor-quality recordings that cannot be transcribed.

Results: Research Question 1

Almost perfect

ja Weltbürger wäre dass ich mich überall zu Hause fühlen würde das oder so ähnlich

ich finde den Begriff zivile Ungehorsam doof und ich habe es auch dreimal gelesen um es einigermaßen zu verstehen aber ich will keine Beispiele nennen aber ich denke einfach dass wir anderer Meinung sind also die Menschen da anderer Meinung sind und schon deshalb einfach Ungehorsam aber der Begriff ziviler Ungehorsam der gehört hier nicht her widerstrebt mir finde ich auch nicht gut

Small discrepancies

ein Weltbürger ist quasi überall auf der Welt zu Hause vielleicht auch nirgendwo zu 100% zu Hause und ja kann man sich nicht an eine bestimmte Kultur oder Herkunft ist offen für er alle möglichen Kulturen und auch ich freue mich flexibel also möglichst viel gereist und hat auch ja einen großen Teil seines Lebens vielleicht im Ausland verbracht und dort gelebt

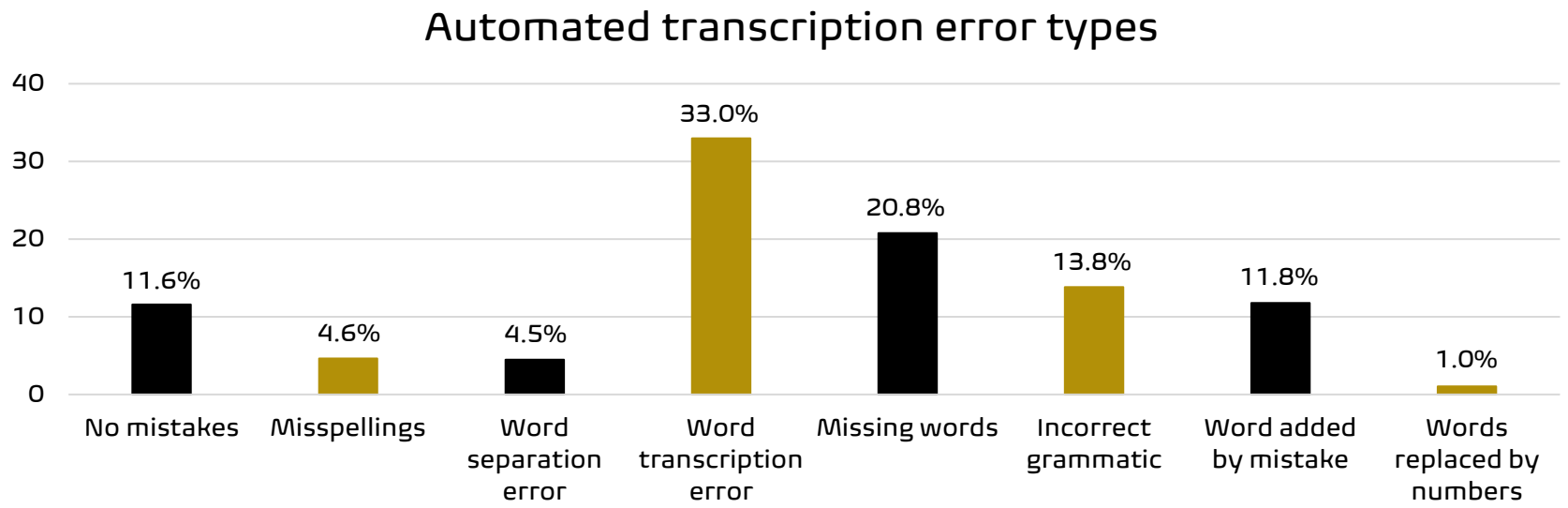
Beispiel ich bin zivil und gehorche nicht in Medien und Politik

Insufficient quality

es ist eine schöne und mein schöner Garten Schüler glaube oder sonst du für Weltbürger ist letztendlich wird die Mehrheit der Menschheit immer erst zum 80. somit auch auf ihr eigenes Land und werde snap macht der geh zu Grunde dafür gibt genügend Beispiele in der Geschichte

ja heute kleine Aufkleber wie ging das System also sich nicht an alles halten was die einem vorschreiben möchten weil es kann ja nicht sein dass das da das ist zwar die Meinungsfreiheit los heute beginnt jetzt immer Kasperl und so entscheiden 19 Zoll für dich richtig ist aber naja und das Badezimmer zwei Leute hatten da ich glaube ich würde hier die auch mal besprechen

Results: Research Question 2



Results: Research Question 3

Questions	File number	Coder 1 time (20%)	Coder 2 time (20%)	Mean time (20%)	Estimated time (100%)
1	63	126 min	112 min	119 min	595 min
2	60	93 min	95 min	94 min	468 min

Google "Cloud Speech-to-Text" API needs about 40 seconds for transcribing 1 min voice input. In total, the API needed about 153 minutes for all voice answers.

Discussion and Conclusion

- The API makes a great job in 40% of the cases
  - *In the other cases, there are (some) quality concerns*
  - *Combination of API and humans may be most promising*
- There are various error types
  - *Word transcription error and missing words are most prominent*
  - *Misspellings and word separation errors are only minor threats*
- API transcription is (at least) 3.1 times faster than humans
  - *Its console can be easily operated through R*
  - *Open-source solutions are developing (e.g., Whisper)*
- Take home message
  - *APIs are not entirely ready to take transcriptions over*
  - *Transcription effort needs to be considered in studies with voice answers*

Many thanks for your attention!

www.jkhoehne.eu
@jkhoehne

Literature

- Elevelt, A., Bernasco, W., Lugtig. P., Ruiter, S., & Toepoel, V. (2021). Where you at? Using GPS locations in an electronic time use diary study to derive functional locations. Social Science Computer Review, 39, 509–526.
- Gavras, K., & Höhne, J.K. (2022). Evaluating political parties: Criterion validity of open questions with requests for text and voice answers. International Journal of Social Research Methodology, 25, 135-141.
- Gavras, K., Höhne, J.K., Blom, A., & Schoen, H. (2022). Innovating the collection of open-ended answers: The linguistic and content characteristics of written and oral answers to political attitude questions. Journal of the Royal Statistical Society (Series A), 185, 872-890.
- Gummer, T., Höhne, J.K., Rettig, T., Roßmann, J., & Kummerow, M. (2023). Is there a growing use of mobile devices in web surveys? Evidence from 128 web surveys in Germany. Quality and Quantity. DOI: 10.1007/s11135-022-01601-8
- Höhne, J.K., Gavras, K., & Qureshi, D.D. (2021). SurveyVoice (SVoice): A comprehensive guide for collecting voice answers in surveys. Zenodo. DOI: 10.5281/zenodo.4644590
- Höhne, J.K., & Schlosser, S. (2019). SurveyMotion: What can we learn from sensor data about respondents' completion and response behavior in mobile web surveys? International Journal of Social Research Methodology, 22, 379–391.
- Pasek, J., McClain, C.A., Newport, F., & Marken, S. (2020). Who's tweeting about the president? What big survey data can tell us about digital traces? Social Science Computer Review, 38, 6
- Revilla, M., Toninelli, D., Ochoa, C., & Loewe, G. (2016). Do online access panels need to adapt surveys for mobile devices? Internet Research, 26(5), 1209–1227.
- Stier, S., Breuer, J., Siegers, P., & Thorson, K. (2020). Integrating survey data and digital trace data: Key issues in developing an emerging field. Social Science Computer Review, 38, 503–516.
- Tourangeau, R., Rips, L.J., & Rasinski, K. (2000). The psychology of survey response. Cambridge, UK: Cambridge University Press. 33–650.