

# Herausforderungen des Crowworkings zu Forschungszwecken

## Spam-Erkennung und Validierung in Amazon-Mechanical-Turk-Studien

*Stefan Taubert, Maximilian Eibl*

Technische Universität Chemnitz, Deutschland  
{[stefan.taubert](mailto:stefan.taubert@informatik.tu-chemnitz.de), [eibl](mailto:eibl@informatik.tu-chemnitz.de)}@informatik.tu-chemnitz.de

### Abstract

Crowdworking-Plattformen wie Amazon Mechanical Turk (AMT) ermöglichen eine schnelle Datenerhebung, sind jedoch mit erheblichen Herausforderungen verbunden. Im Rahmen eines Forschungsprojekts zur Sprachsynthese wurden auf AMT eine Alphastudie und eine Hauptstudie durchgeführt, um die Verständlichkeit und Natürlichkeit synthetisierter Sätze zu bewerten. Die Alphastudie zeigte, dass zahlreiche Bots an der Umfrage teilnahmen, wodurch die erhobenen Daten unbrauchbar wurden. Daraufhin wurde ein überarbeitetes Studiendesign mit einer vorgeschalteten Qualifizierung entwickelt, um Bots auszuschließen und die Qualität der Bewertungen zu erhöhen. Die Hauptstudie konnte die Teilnahme von Bots unterbinden und eine Korrelation zwischen subjektiven und objektiven Ergebnissen herstellen.

**Keywords:** AMT; Sprachsynthese; TTS; Crowdsourcing

## 1 Einleitung

Crowdworking ist eine Form der Arbeitsorganisation, bei der Aufgaben über digitale Plattformen an eine große, oft globale Gruppe von Arbeitskräften vermittelt werden. Dadurch können sehr rasch einfache Daten erhoben werden, die auch im Rahmen von Forschungsprojekten erhoben werden, wie etwa Annotationen oder Bewertungen. In diesem Posterbeitrag wird der Ein-

satz einer sehr breit eingesetzten Plattform, Amazon Mechanical Turk (AMT), beschrieben, wobei wir vor allem auf die Schwierigkeiten der Verwendung eingehen wollen.

Im Rahmen eines Forschungsprojekts zu Sprachsynthesystemen wurden auf AMT eine Alphastudie und eine Hauptstudie durchgeführt, um die Verständlichkeit und Natürlichkeit synthetisierter Sätze mithilfe von Meinungsskalen zu bewerten. Die Alphastudie diente dabei als erste Annäherung an die Befragungsmethode und es wurden kaum Einschränkungen hinsichtlich der Abgabe von Bewertungen implementiert. Dabei stellte sich aufgrund der Art und Geschwindigkeit der Antworten schnell heraus, dass zahlreiche Bots an der Umfrage teilgenommen hatten, wodurch die Ergebnisse unbrauchbar waren. Um ungültige Bewertungen zu identifizieren, wurden umfassende Analysen durchgeführt.

In der Hauptstudie wurden Maßnahmen ergriffen, um ungültige Antworten bereits im Vorfeld zu verhindern. Dafür wurde ein komplexes Studiendesign entwickelt, das aus zwei Umfragen bestand: einer Qualifizierungsumfrage und einer Hauptumfrage zur Bewertung der Audiobeispiele. Durch das verbesserte Studiendesign konnten die Teilnahme von Bots sowie die Abgabe ungültiger Antworten wirksam unterbunden werden. Während die subjektiven Bewertungen der ersten Umfrage keine Übereinstimmung mit objektiven Ergebnissen zeigten, war in der zweiten Umfrage eine Korrelation zwischen subjektiven und objektiven Bewertungen feststellbar.

## 2 AMT

AMT ist ein seit 2005 existierender Crowdworking-Marktplatz, der es Einzelpersonen und Unternehmen erleichtert, ihre Prozesse und Aufgaben an verteilte Arbeitskräfte auszulagern, die diese Aufgaben virtuell ausführen können. Umfragendurchführer profitieren von AMT, da sie Studien kostengünstig, schnell und flexibel durchführen können.

Mit über 100.000 registrierten und ca. 2450 aktiven Arbeitern gibt es einen Zugang zu einer breiten Arbeiterbasis (Difallah et al. 2018: 142 f.). Die in Auftrag gegebene Arbeit wird dabei in einzelne Aufgaben (Human Intelligence Tasks, HITs) unterteilt, um jederzeit eine Pause zu ermöglichen. In der Literatur ist AMT die am häufigsten zur Sprachsynthese-Evaluation eingesetzte

Plattform (Eskenazi et al. 2013: 257), siehe z.B. Ribeiro et al. (2011), Cambre et al. (2020: 2) oder Kim et al. (2020: 6).

### 3 Alphastudie

In der Alphastudie wurden lediglich die *Approval Rate* und das Herkunftsland der Teilnehmer als Voraussetzungen für die Teilnahme festgelegt. Die *Approval Rate* beschreibt das Verhältnis zwischen akzeptierten und zurückgewiesenen HITs und weist bei niedrigen Werten auf häufig unbrauchbare Ergebnisse hin, da fehlerhafte oder unplausible Arbeiten von Durchführern zurückgewiesen werden können. Bereits innerhalb der ersten 20 Minuten nach Veröffentlichung der Studie hatten 299 der insgesamt 318 teilgenommenen Arbeiter 505 von 540 Aufgaben bearbeitet. Dieses Verhalten war äußerst unwahrscheinlich, da es sehr unplausibel war, dass eine so große Anzahl von Arbeitern zeitgleich auf die neuen HITs stößt und sie unverzüglich korrekt bearbeitet. Auffällig war zudem, dass 203 Arbeiter (64%) lediglich ein einzelnes HIT bearbeiteten.

Nachträglich wurden verschiedene Kriterien angewendet, um die Echtheit der Antworten zu überprüfen und verdächtige Verhaltensmuster zu identifizieren:

- auffällig niedrige Bewertungen von Audiobeispielen aus der *Ground Truth*,
- übermäßig schnelle Bearbeitungen,
- auffälliges Verhalten, wie das gleichzeitige Bearbeiten mehrerer HITs.

Wenn ein Arbeiter eines dieser Kriterien in einem seiner HITs erfüllte, wurden in der Postbearbeitung der Alphastudie alle HITs dieses Arbeiters herausgefiltert. Verbleibende Ergebnisse, die deutlich von objektiven Bewertungen abwichen, wurden ebenfalls als unplausibel gewertet und Bots zugeschrieben. Zum Vergleich wurde die Metrik Mel-Cepstral-Distanz berechnet (Kubichek 1993; Wagner et al. 2019: 107 f.). Diese Metrik wird häufig zur objektiven Evaluation von Sprachsynthesystemen eingesetzt, wie unter anderem in Skerry-Ryan et al. (2018) oder Lee und Kim (2019).

## 4 Hauptstudie

Um die Probleme der Alphastudie zu beheben, wurde ein völlig neues Studiendesign entwickelt, das sich an der Richtlinie ITU-T P.808 (ITU-T 2021) orientierte. Diese Richtlinie definiert ein standardisiertes Verfahren zur Bewertung der Sprachqualität mithilfe von Crowdworking. Die ITU-T ist ein Hauptsektor der International Telecommunication Union (ITU), der für die Standardisierung im Bereich der Telekommunikation zuständig ist. Die ITU selbst ist eine Sonderorganisation der Vereinten Nationen.

In dieser zweiten Studie wurden neben der Festlegung auf die USA als Teilnahmeort auch strengere Kriterien für die Zulassung definiert. Es wurde eine höhere *Approval Rate* vorausgesetzt und nur Arbeiter mit mindestens 10.000 akzeptierten HITs wurden zugelassen. Vor der Hauptumfrage absolvierten die Arbeiter einen Qualifizierungstest zur Eignungsprüfung, wobei beispielsweise Arbeiter aus dem Bereich Sprachverarbeitung ausgeschlossen wurden. Anschließend mussten sie in einem Hör- und Englischtest eine Mindestpunktzahl erreichen.

In der Hauptaufgabe wurden Fangfragen eingebaut, um die Aufmerksamkeit der Arbeiter zu überprüfen. In diesen Fangfragen wurde nur der Anfang eines Audios abgespielt, bevor eine andere Stimme sie dazu aufforderte, eine spezifische Auswahl zu treffen. HITs, in denen für diese Fangfrage die falsche Auswahl getroffen wurde, wurden automatisch zurückgewiesen. Zusätzlich wurde ein Mechanismus integriert, bei dem jeweils ein Audio pro HIT doppelt innerhalb der Umfrage auftauchte, um die Konsistenz der Bewertungen zu überprüfen. Große Abweichungen in der Bewertung desselben Audios deuteten auf unzuverlässige Ergebnisse hin.

Zusätzlich wurden verschiedene Faktoren gemessen und analysiert, um die Antwortqualität der Teilnehmer zu bewerten:

- Anzahl der mehrfach angehörten Audios,
- gleichzeitige Bearbeitung mehrerer HITs,
- Unterbrechungen während des Anhörens,
- Inaktivitätsdauer bei der Bearbeitung,
- Bewertungen vor vollständigem Abspielen der Audiodatei,
- Korrelation mit den Ergebnissen anderer Teilnehmer.

Das verbesserte Studiendesign führte zu Ergebnissen, die mit den objektiven Evaluierungen und Erwartungen übereinstimmten und stellte sicher, dass

ausschließlich Menschen teilnahmen. Dennoch wiesen acht der 33 Teilnehmer negative Korrelationen oder Werte nahe Null auf, was die Notwendigkeit weiterer Optimierungen in der Methodik, wie gezieltere Auswahlkriterien oder verbesserte Anreize, aufzeigt.

## Literatur

- Cambre, J.; Colnago, J.; Maddock, J.; Tsai, J.; Kaye, J. (2020): Choice of Voices: A Large-Scale Evaluation of Text-to-Speech Voice Quality for Long-Form Content. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. <https://doi.org/10.1145/3313831.3376789>
- Difallah, D.; Filatova, E.; Ipeirotis, P. (2018): Demographics and Dynamics of Mechanical Turk Workers. In: *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. <https://doi.org/10.1145/3159652.3159661>
- Eskenazi, M.; Levow, G.-A.; Meng, H.; Parent, G.; Suendermann, D. (2013): *Crowdsourcing for speech processing. Applications to data collection, transcription and assessment*. Wiley.
- ITU-T (2021): International Telecommunication Union Recommendation P.808 – Subjective evaluation of speech quality with a crowdsourcing approach. <https://www.itu.int/rec/T-REC-P.808-202106-I/en>
- Kim, J.; Kim, S.; Kong, J.; Yoon, S. (2020): Glow-TTS: A Generative Flow for Text-to-Speech via Monotonic Alignment Search. In: *Advances in Neural Information Processing Systems*.
- Kubichek, R. (1993): Mel-cepstral distance measure for objective speech quality assessment. In: *Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing*. <https://doi.org/10.1109/PACRIM.1993.407206>
- Lee, Y.; Kim, T. (2019): Robust and Fine-grained Prosody Control of End-to-end Speech Synthesis. In: *ICASSP 2019 – 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. <https://doi.org/10.1109/ICASSP.2019.8683501>
- Ribeiro, F.; Florêncio, D.; Zhang, C.; Seltzer, M. (2011): CrowdMOS: An approach for crowdsourcing mean opinion score studies. In: *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. <https://doi.org/10.1109/ICASSP.2011.5946971>
- Skerry-Ryan, R. J.; Battenberg, E.; Xiao, Y.; Wang, Y.; Stanton, D.; Shor, J.; Weiss, R.; Clark, R.; Saurous, R. A. (2018): Towards End-to-End Prosody Transfer for

Expressive Speech Synthesis with Tacotron. In: *Proceedings of the 35<sup>th</sup> International Conference on Machine Learning*.

Wagner, P.; Beskow, J.; Betz, S.; Edlund, J.; Gustafson, J.; Eje Henter, G.; Le Maguer, S.; Malisz, Z.; Székely, É.; Tännander, C.; Voße, J. (2019): Speech Synthesis Evaluation – State-of-the-Art Assessment and Suggestion for a Novel Research Program. In: *10<sup>th</sup> ISCA Workshop on Speech Synthesis (SSW 10)*. <https://doi.org/10.21437/SSW.2019-19>

In: M. Eibl (Hrsg.): Datenströme und Kulturoasen – Die Informationswissenschaft als Bindeglied zwischen den Informationswelten. Proceedings des 18. Internationalen Symposiums für Informationswissenschaft (ISI 2025), Chemnitz, Deutschland, 18.–20. März 2025. Glückstadt: Verlag Werner Hülsbusch, S. 446–451.  
DOI: <https://doi.org/10.5281/zenodo.14925656>