Does Speech Translation Meet Users' Needs? An English to Portuguese Study Across Demographics

Giuseppe Attanasio[*Λ], **Beatrice Savoldi**[Φ], **Daniel Chechelnitsky**[Π], **Matteo Negri**[Φ], **Marine Carpuat**[Υ], **André F.T. Martins**[ΣΛT]

[Λ] Instituto de Telecomunicações, Lisbon, Portugal
[Φ] Fondazione Bruno Kessler, Trento, Italy
[Σ] Instituto Superior Técnico, Lisbon, Portugal
[Π] Carnegie Mellon University, Pittsburgh, USA
[Υ] University of Maryland, College Park, USA
[T] TransPerfect

Abstract

This paper introduces Ouvia, a research project to assess user-perceived usability and reliability of modern speech translation tools in En→Pt scenarios. The project centers on a user study in which we simulate real-life daily interactions by recruiting crowdworkers online from different sociodemographic groups. We collect their spoken requests and self-assessments about quality, satisfaction, and reliability. Here, we describe the project's motivation and objectives, the study design, and the expected outcomes we will provide to speech translation practitioners.

1 Introduction

While machine translation is consolidating its popularity among laypeople, scholarly accounts have advocated rethinking a new way to evaluate this technology—one centered on people, their communication needs, and real-world use contexts (Savoldi et al., 2025; Carpuat et al., 2025).

This project investigates **whether modern AI-based speech translation systems can meet people's communication needs** across diverse real-life scenarios. It centers on an online user study and data collection that mimics a one-to-one interaction: a person starts a conversation with a request in English, and an AI-based system translates it for a Portuguese listener. The project's goal is to measure success along two axes. First, we center the evaluation on user-perceived aspects—"I trust the AI translation to convey my message" or "I would use this AI translation in a real-world situation" are two of the statements we ask participants to rate their level of agreement with. Second, we replicate the study over different demographic groups, stratifying our speakers on first language (native English or not), gender, and ethnic group. This dimension aligns with recent calls for AI-enabled speech technologies to be equitable across different demographics (Attanasio et al., 2024).

Objectives. This project has three main objectives. We will (**O1**) establish whether current speech translation AIs provide outputs that end users would be willing to rely on in real-world En→Pt scenarios, (**O2**) measure the extent to which language variants and demographic factors affect user-perceived reliability, and (**O3**) assess whether current translation quality metrics capture such user-centered perceptions.

We will compile all collected data, including speaker recordings, annotations, and self-reported assessments, in a new speech recognition and translation benchmark. We plan to recruit participants (speakers) from three language groups (native and non-native English), three broad ethnic groups, and three gender identities. We will release all artifacts (data, annotations, code) in a GDPR-compliant bundle, licensed under CC BY 4.0.

Funding Agency and Partners. The project is a joint effort from Instituto de Telecomunicações (IT; leading), Instituto Superior Técnico (IST), Fondazione Bruno Kessler, Carnegie Mellon University, and the University of Maryland. The duration is 12 months, starting May 2025. Funds for the entire data collection are provided by the European Association for Machine Translation, awarded through the *2025 EAMT Sponsorship of Activities* call. IT provides computational and logistic support. The user study obtained ethical ap-

[*]Correspondence to giuseppe.attanasio@lx.it.pt.

proval from IST's Ethics Committee.

2 Study Design

We conduct a four-stage online study to mimic a one-to-one exchange in which an English *speaker* interacts with a Portuguese *receiver*, with the speaker's voice translated by an AI.

Round 1. The speaker reads aloud and records a passage we provide them with. This *conversation starter* is 40-60 words long and contains key information, such as named entities or quantities, that a system should translate correctly. An example is "Hi, I've had a persistent rash on my arms and torso for five days. [...]" **Round 2.** We translate the speaker's recording automatically and provide the receiver with the outcome. Then, we ask the receiver to reply to a set of questions grounded in the original passage, e.g., "For how many days has the person had symptoms?" This approach borrows from QA-based machine translation evaluation protocols. **Round 3.** We recruit a third participant, who speaks both English and Portuguese fluently, to validate the translation in two ways: A scalar quality score and a binary judgment of which question was answered correctly by the receiver. We use several heuristics to check these assessments. **Round 4.** The initial speaker receives the validation outcome and answers a qualitative survey designed to capture self-assessed satisfaction, effectiveness, and reliability.

3 Preliminary Results and Conclusions

At the time of writing, our main outcomes are (**i**) a set of conversation starters and associated questions used in Rounds 1 and 2, (**ii**) a web-based platform to handle data collection and user participation, and (**iii**) data from 1200+ interactions.

We compiled (**i**) with healthcare-related and mundane (e.g., traveling) interactions to cover both high- and low-stakes scenarios, sourcing most starters from existing dialogue benchmarks. To increase linguistic coverage, we developed an AI-assisted pipeline in which we first prompt a flagship language model to generate synthetic starters, then manually validate and apply heuristics to quality-check them. All participants interacted through a custom-developed web app (**ii**), which streamlines users and recordings management and admin tooling. We recruited all crowdworkers through `prolific.com`. We collected scripted recordings from more than 100 US residents who self-identified as female or male and as white or black/African American (self-declared attributes), totaling (**iii**) 1,240 unique interactions (speaker recordings, receiver answers, validator quality assessments, and speaker final self-assessment).

Our data, platform, and cross-demographic findings will help identify blind spots in current speech translation systems and inform future development priorities. They will foster research across speech and demographic groups, laying the groundwork for focused inquiries into fairness and equity, as well as studies on how automatic metrics relate to user-defined reception and usability.

4 Acknowledgments

GA is supported by the Portuguese Recovery and Resilience Plan through project C645008882-00000055 (Center for Responsible AI) and by cofunded EU funds under UID/50008: Instituto de Telecomunicações; AM by the project DECOLLAGE (ERC-2022-CoG 101088763). BS is funded by the Horizon Europe research and innovation programme, under grant agreement No 101135798, project Meetween.

References

[Attanasio et al.2024] Attanasio, Giuseppe, Beatrice Savoldi, Dennis Fucci, and Dirk Hovy. 2024. Twists, humps, and pebbles: Multilingual speech recognition models exhibit gender performance gaps. In Al-Onaizan, Yaser, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of EMNLP 2024*, Miami, Florida, USA, November. Association for Computational Linguistics.

[Carpuat et al.2025] Carpuat, Marine, Omri Asscher, Kalika Bali, Luisa Bentivogli, Frédéric Blain, Lynne Bowker, Monojit Choudhury, Hal Daumé III, Kevin Duh, Ge Gao, Alvin Grissom II, Marzena Karpinska, Elaine C. Khoong, William D. Lewis, André F. T. Martins, Mary Nurminen, Douglas W. Oard, Maja Popovic, Michel Simard, and François Yvon. 2025. An interdisciplinary approach to human-centered machine translation. In Christodoulopoulos, Christos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of EMNLP 2025*, Suzhou, China, November. Association for Computational Linguistics.

[Savoldi et al.2025] Savoldi, Beatrice, Alan Ramponi, Matteo Negri, and Luisa Bentivogli. 2025. Translation in the hands of many: Centering lay users in machine translation interactions. In Christodoulopoulos, Christos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of EMNLP 2025*, Suzhou, China, November. Association for Computational Linguistics.