Beyond Technical Efficiency: A Sociotechnical Approach to Measuring Crisis Speech Tech Impact

BELU TICONA, George Mason University, USA
AMNA LIAQAT, George Mason University, USA
ANTONIOS ANASTASOPOULOS, George Mason University, USA.

Crisis is a high-stake domain that requires accurate, timely, and trustworthy speech communication. Across the world, crisis affects populations from diverse languages and cultures, particularly groups that are linguistically and socio-economically more vulnerable. In this context, deployment of *effective* speech artificial intelligence (AI) technologies for crisis management is urgent for communities who primarily use their language in spoken forms (e.g. non-institutional, indigenous, and vernacular languages, among others). Interdisciplinary and cross-institutional collaboration is crucial to identify how to deliver *impactful* speech technology, considering the needs, uses and challenges from crisis ecosystem. In this position paper, we argue that current evaluation practices for speech technologies fail to capture real-world impact, which is primordial in high-stakes and vulnerable settings. This work is informed by our experiences building speech infrastructure for minoritized language communities as part of two collaborations with non-profit organizations: Mozilla Foundation and CLEAR Global. Ultimately, we call on the academic community to move beyond narrow technical efficiencies towards multidimensional situated measurement that support decision-making in the process of technology development.

CCS Concepts: • **Human-centered computing** → **Human computer interaction (HCI)**; **Accessibility**; • **General and reference** → **Evaluation**; • **Computing methodologies** → *Speech recognition.*

Additional Key Words and Phrases: Crisis Management, Speech Data Collection, Decision Impact Study

1 Introduction

Measuring real-world impact of language technologies (LTs) initiatives is little explored by the academic community, despite being of high relevance to stakeholders, practitioners, end-users, and the general public. For instance, in the ACL Anthology[1] only 0.1% of published work report impact evaluations of any form, which is crucial to move beyond metric evaluation towards understanding the effects of technology to society [21]. This is particularly important in high-stake domains, such as crisis management, where technology can have life-saving consequences for most vulnerable communities.

The rapid adoption of Large Language Models (LLMs) as general purpose machines has increased NLP applications to serve crisis responders, covering domain-specific tasks for all stages of disaster management [15, 23]. However, most work relies on text-only data, leaving spoken languages behind. For instance, D. Lewis et al. [9] analyzes crisis social media datasets, finding that half of corpora are English-only. In this sense, speech technology is utmost need to serve communities whose preferred communication form is oral, such as displaced populations, refugees, and indigenous communities [11, 12].

In this position paper, we propose to measure impact as part of decision-making in the development of technology, especially when serving vulnerable communities. Our position is informed by two collaborations with NGOs that address linguistic inequality and social good impact by building speech infrastructure for crisis scenarios. We hope

[1]https://aclanthology.org

Authors' Contact Information: Belu Ticona, mticonao@gmu.edu, George Mason University, Fairfax, Virginia, USA; Amna Liaqat, aliaqat@gmu.edu, George Mason University, Fairfax, Virginia, USA; Antonios Anastasopoulos, antonis@gmu.edu, George Mason University, Fairfax, Virginia, USA..

to incentive the academic community to reflect on *who* and *how* is impacted by LTs, sharing their experiences to understand how better serve society.

2 Related Work

This work lies at the intersection of speech technology and human-computer interaction (HCI), more specifically focusing on the social and real-world impact of data collection projects for automatic speech recognition (ASR) systems when working with spoken languages. ASR systems are known to have different performance according the speakers' socio-linguistic context [10, 16, 18]. Recent work analyzes the limitation of common ASR metrics and calls for HCI and ethnographics perspectives on how to better evaluate systems (e.g. using intersectional benchmarks, qualitative error analysis, among others) [17]. From a HCI perspective, limited work study ASR for communities of low-resource language languages, exploring design methods for community engagement [2] and user studies on how to adapt ASR to Global South populations [5]. Related work by Reitmaier et al. [22] explore participatory methods with a farming community in India to create speech-based information retrieval systems, reflecting on how to co-create tools for unwritten languages. In parallel, NLP efforts to address linguistic inequality have gained increasing relevance in recent years [7, 13, 19], examining the influence of socio-economic and demographic factors on technology performance [6, 8]. This trend has raised concerns about the social relevance, utility, and impact of NLP initiatives [1, 14, 20], motivating studies on positive impact [4] and reflections on impact evaluation [21]. Furthermore, Reiter [21] argues that the NLP community should include such evaluations to assess the utility of academic efforts and accelerate technology adoption through methods such as key performance indicators, A/B studies, and before-and-after studies.

3 Studies on Speech Technology for Vulnerable Communities

3.1 Crowd-sourcing Indigenous Speech Data: the Qom Case

Indigenous communities from low socio-economic backgrounds and with limited access to technology are among the most vulnerable in crisis and disaster scenarios. Our first informative study is based on our collaboration with a Qom community, a South American Indigenous group who traditionally inhabit the Gran Chaco region [3][2]. We created the first speech corpus for the Qom/Toba language using the Mozilla Common Voice (MCV) platform[3], as part of the MCV Spontaneous Speech Initiative for voice-based AI inclusion. More precisely, we worked with two families from Argentina to collect 10 hours of transcribed speech through prompt answering, conducting linguistic fieldwork and training speakers on how to use the crowd-sourcing platform. We adapted our methods of involvement and interaction with speakers based on the diverse socioeconomic and technological literacy backgrounds of our Qom collaborators. By doing so, we provided various engaging experiences, such as in-person meetings, virtual training, phone calls, and daily/weekly instant messages, which provided distinct social impacts at individual and community levels (e.g., project ownership, resonance between family and friends, identity pride, motivation to pursue higher education). These first-hand experiences allowed us to reflect on the meaning of *impactfulness* in speech data collection projects, as a subjective dimension that can refer to data resourcefulness, capacity building, application utility, or a type of socio-affective experience for speakers.

[2]This region encompasses Northern Argentina, Southern Paraguay, and Bolivia. In particular, our collaborators are originally from Chaco province, the region with the highest level of poverty in Argentina
[3]https://commonvoice.mozilla.org/

3.2 Crisis Speech Technology: A Decision Impact Study

With the goal of supporting vulnerable speech communities in crisis scenarios, and grounding our decision-making process in the first-hand lived experiences of those interacting with affected communities, we established a collaboration with CLEAR Global, a non-profit organization focused on helping communities be heard in critical domains such as health, humanitarian response, and human rights. This NGO has broad experience supporting communities across the globe, who could potentially benefit from crisis speech infrastructure. However, due to time and funding constraints, our collaboration can only serve a limited number of language communities. How to identify which communities would benefit the most from crisis speech technology? Or in other words, how to estimate the real-life *impact* of speech AI for each potential served community? To answer this question, we designed a composite language-community risk index to guide decision-making on how speech infrastructure would be most beneficial. Specifically, we considered technological, linguistic, socio-economic, and crisis risk indicators, using language, country, and type of crisis as the combinatorial unit of analysis. Based on this index, we identified ten of the most in-need communities to support through the collection of in-domain, crowdsourced speech data.

Our current work-in-progress consists of validating this index using a mixed-methods approach, combining: (1) a quantitative breadth survey intended for crisis translators, and (2) an in-depth qualitative interview study directed at various crisis actors (such as community leaders, government representatives, radio and broadcasting organizations, crisis tech companies, and NGOs). In this way, we aim to map the landscape of crisis communication, focusing on how voice tools can be designed for maximum benefit and successful adoption within the crisis management ecosystem.

4 The What, Who and How of Measurement in Crisis Speech Technology

Our experience working with vulnerable communities and NGOs that support crisis responders shows that the creation of speech technology in the crisis domain demands a reflection on the impact of our academic practices, moving measurement beyond system performance towards. We argue that in high-stakes domains, where interdisciplinary and cross-institutional collaboration is crucial, metric-supported decision-making is determinative for generating a *positive* impact on communities we aim to serve. The definition of measurement artifacts can provide transparency on the desired effects to society our technology is designed to produce. This requires understanding how actors within the ecosystem adopt technology, the characteristics of the speech communities we aim to support, and the specific type of impact desired. While metrics in speech research are usually restricted to task performance (e.g., Word Error Rate for automatic speech recognition), the development of speech applications requires an understanding of the domain's complexity to define which dimensions need to be measured, why they are relevant, and how to measure them. In the following, we describe our insights to answer these questions, focusing on the crisis management domain and informed by our previously described studies.

Impact Measurement. Initiatives for technology inclusion target communities traditionally underserved in the tech space, whether in academia or industry. Even when researchers aim to address socially relevant projects, their perspectives might neglect the communities they aim to support due to a lack of insight into how these communities perceive the problem and how they wish to be assisted. Our experiences show that a reflection on the desired impact needs to be addressed for effective inclusion, grounding our approaches in lived experiences with target communities and collaborating with domain experts and organizations experienced in supporting affected populations. Our engagement with the Qom community shows that initiatives for speech inclusion might overlook the socio-affective effects of how we interact with Indigenous communities, focusing on data collection and relying the speakers' engagement into the

projects' leads decision (generally academics non-community members). This experience enabled us to reflect on the impact we aim to create and the actual experiences of the speakers. Our second study, with an organization deeply involved in the crisis and humanitarian response domain, allowed us to define criteria to quantify the potential *impact* of our efforts to build a crisis speech infrastructure, forming our decisions on how to best allocate resources and capacity. It is important to note that practitioners, funding institutions, and governments might have different criteria to identify how impactful technology could be, reflecting particular interests and positions.

The Crisis Ecosystem. Communication within crisis management involves multiple actors, each with distinct perspectives on the utility, requirements, and challenges of speech applications. For instance, government organizations focused on crisis response prioritize the broadcasting of official announcements and national alerts (one-to-many), while radio and emergency stations may focus on interactive speech tools (one-to-one or few-to-few). In contrast, NGOs often target marginalized populations, necessitating support for local language varieties. The design of the quantitative survey and qualitative interviews in our second study aims to define the specific speech tools and data types required to support these various crisis actors. Furthermore, we expect to identify early indicators of the potential reception and adoption of speech AI technology, which may carry a variety of connotations for end-users depending on their specific language and location. Moreover, while the composite crisis index was constructed considering socioeconomic and technically relevant dimensions informed by the NGO's expertise, a similar methodology could potentially be used to serve other stakeholders.

Language-Community Specificity. Traditional metrics for speech tasks were primarily designed and tested on high-resource languages, often neglecting the particularities of language varieties and dialects. For example, Word Error Rate (WER) may be useful for non-agglutinative languages where sentences are formed of multiple distinct words. However, it fails for languages where a single, long word can function as a full sentence by combining multiple morphemes, each carrying a unique meaning. Furthermore, automatic metrics usually involve a comparison with a gold standard, which is only viable for languages with a unified orthographic system. Our experience with the Qom community, specifically during the validation and transcription of the speech corpus, illustrates how speakers, even from the same extended family, may disagree on transcription decisions. The selection of one language variety as representative of a language, or the creation of a unified or standardized orthography, can carry political and power implications that may affect how speakers perceive the usefulness of end applications and how willing they are to adopt them. In addition, community characteristics such as culture and literacy also shape speech data collection practices. For example, in our first study we had to adapt and revise our initial set of prompts (which had originally been designed in Spanish), so they would better fit Qom speakers' culture.

Application Performance. The deployment of an application requires decision-making with regard to usability and reliability. Even when applications rely on the same underlying technology, the performance threshold for determining suitability may differ based on the context of use. Moreover, the measurement of application performance is dependent on the level of Human-AI collaboration. For instance, speech translation for situational awareness may be ready for deployment if it accelerates the identification of "where" and "what" in crisis scenarios, potentially completing end-to-end tasks independently. However, in a health crisis the accurate translation of concepts, such as body parts, instructions, and symptoms, is crucial. In such high-stakes environments, a tool must support rather than undermine aid provision. In these cases, the performance of a speech application should be measured by quantifying its utility in assisting human experts, rather than focusing on independent task completion.

5 Conclusion

In this work, we have argued that measurement in speech technology cannot focus solely on system performance, especially when addressing vulnerable communities traditionally underrepresented in technology-making spheres. Our formative studies with these communities illustrate how the effectiveness of speech infrastructure is a multidimensional construct. We propose establishing interdisciplinary and cross-institutional collaborations to align the intended impact with the needs, uses, and challenges of stakeholders in the domain ecosystem. Essentially, we understand measurement is non-neutral act, but a reflection of the values we prioritize within the socio-technical ecosystem. As we continue our crisis tech impact index validation and the analysis of crisis actors' perspectives on speech AI tools, we call on collaborations between the HCI and LTs community to reflect on how measurement can be redesigned as a tool for accountability of technology.

Acknowledgments

This work is supported by the National Science Foundation under award CIRC 2346334. We are also thankful to our collaborators: Francis Tyers (Mozilla Foundation); Paola Cúneo (Consejo Nacional de Investigaciones Científicas y Técnicas - Universidad de Buenos Aires, Instituto de Lingüística); Toba (Qom) Daviaxaiqui Community, located at Presidente Derqui (provincia de Buenos Aires, Argentina); and Alp Öktem, Polly Harlow, and Lisa Reiners (CLEAR Global).

References

[1] [n. d.]. *Beyond Good Intentions: Reporting the Research Landscape of NLP for Social Good - ACL Anthology.* https://aclanthology.org/2023.findings-emnlp.31/

[2] [n. d.]. Opportunities and Challenges of Automatic Speech Recognition Systems for Low-Resource Language Speakers. https://dl.acm.org/doi/epdf/10.1145/3491102.3517639. doi:10.1145/3491102.3517639

[3] [n. d.]. *The Situation of Indigenous Peoples in Argentina | United Nations Special Rapporteur on the Rights of Indigenous People.* https://un.arizona.edu/search-database/situation-indigenous-peoples-argentina

[4] Katherine Atwell, Laura Biester, Angana Borah, Daryna Dementieva, Oana Ignat, Neema Kotonya, Ziyi Liu, Ruyuan Wan, Steven Wilson, and Jieyu Zhao (Eds.). 2025. *Proceedings of the Fourth Workshop on NLP for Positive Impact (NLP4PI).* Association for Computational Linguistics. doi:10.18653/v1/2025.nlp4pi-1.0

[5] Gifty Ayoka, Giulia Barbareschi, Richard Cave, and Catherine Holloway. 2024. Enhancing Communication Equity: Evaluation of an Automated Speech Recognition Application in Ghana. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24).* Association for Computing Machinery, New York, NY, USA, 1–16. doi:10.1145/3613904.3641903

[6] Elisa Bassignana, Amanda Cercas Curry, and Dirk Hovy. 2025. The AI Gap: How Socioeconomic Status Affects Language Technology Interactions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 18647–18664. doi:10.18653/v1/2025.acl-long.914

[7] Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic Inequalities in Language Technology Performance across the World's Languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Dublin, Ireland, 2022-05), Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, 5486–5505. doi:10.18653/v1/2022.acl-long.376

[8] Amanda Cercas Curry, Giuseppe Attanasio, Zeerak Talat, and Dirk Hovy. 2024. Classist Tools: Social Class Correlates with Performance in NLP. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 12643–12655. doi:10.18653/v1/2024.acl-long.682

[9] William D. Lewis, Haotian Zhu, Keaton Strawn, and Fei Xia. 2025. Tapping into Social Media in Crisis: A Survey. In *Proceedings of the Fourth Workshop on NLP for Positive Impact (NLP4PI)*, Katherine Atwell, Laura Biester, Angana Borah, Daryna Dementieva, Oana Ignat, Neema Kotonya, Ziyi Liu, Ruyuan Wan, Steven Wilson, and Jieyu Zhao (Eds.). Association for Computational Linguistics, Vienna, Austria, 306–331. doi:10.18653/v1/2025.nlp4pi-1.27

[10] Anuj Diwan, Rakesh Vaideeswaran, Sanket Shah, Ankita Singh, Srinivasa Raghavan, Shreya Khare, Vinit Unni, Saurabh Vyas, Akash Rajpuria, Chiranjeevi Yarra, Ashish Mittal, Prasanta Kumar Ghosh, Preethi Jyothi, Kalika Bali, Vivek Seshadri, Sunayana Sitaram, Samarth Bharadwaj, Jai

Nanavati, Raoul Nanavati, Karthik Sankaranarayanan, Tejaswi Seeram, and Basil Abraham. 2021. Multilingual and Code-Switching ASR Challenges for Low Resource Indian Languages. In *Interspeech 2021*. 2446–2450. arXiv:2104.00235 [cs] doi:10.21437/Interspeech.2021-1339

[11] Sunita Joann Rebecca Healey, Nafiseh Ghafournia, Peter D Massey, Karinne Andrich, Joy Harrison, Kathryn Taylor, and Katarzyna Bolsewicz. 2022. Ezidi Voices: The Communication of COVID-19 Information amongst a Refugee Community in Rural Australia- a Qualitative Study. *International Journal for Equity in Health* 21 (Jan. 2022), 10. doi:10.1186/s12939-022-01618-3

[12] Amanda Howard, Kylie Agllias, Miriam Bevis, and Tamara Blakemore. 2017. "They'll Tell Us When to Evacuate": The Experiences and Expectations of Disaster-Related Communication in Vulnerable Groups. *International Journal of Disaster Risk Reduction* 22 (June 2017), 139–146. doi:10.1016/j.ijdrr.2017.03.002

[13] Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 6282–6293. doi:10.18653/v1/2020.acl-main.560

[14] Grace LeFevre, Qingcheng Zeng, Adam Leif, Jason Jewell, Denis Peskoff, and Rob Voigt. 2025. Good Intentions Beyond ACL: Who Does NLP for Social Good, and Where?. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing* (Suzhou, China, 2025-11), Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (Eds.). Association for Computational Linguistics, 5138–5150. doi:10.18653/v1/2025.emnlp-main.259

[15] Zhenyu Lei, Yushun Dong, Weiyu Li, Rong Ding, Qi R. Wang, and Jundong Li. 2025. Harnessing Large Language Models for Disaster Management: A Survey. In *Findings of the Association for Computational Linguistics: ACL 2025*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 14528–14551. doi:10.18653/v1/2025.findings-acl.750

[16] Andreas Liesenfeld, Alianda Lopez, and Mark Dingemanse. 2023. The timing bottleneck: Why timing and overlap are mission-critical for conversational user interfaces, speech recognition and dialogue systems. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, Svetlana Stoyanchev, Shafiq Joty, David Schlangen, Ondrej Dusek, Casey Kennington, and Malihe Alikhani (Eds.). Association for Computational Linguistics, Prague, Czechia, 482–495. doi:10.18653/v1/2023.sigdial-1.45

[17] Nina Markl and Catherine Lai. 2021. Context-Sensitive Evaluation of Automatic Speech Recognition: Considering User Experience & Language Variation. In *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, Su Lin Blodgett, Michael Madaio, Brendan O'Connor, Hanna Wallach, and Qian Yang (Eds.). Association for Computational Linguistics, Online, 34–40.

[18] Mumtaz Begum Mustafa, Mansoor Ali Yusoof, Hasan Kahtan Khalaf, Ahmad Abdel Rahman Mahmoud Abushariah, Miss Laiha Mat Kiah, Hua Nong Ting, and Saravanan Muthaiyah. 2022. Code-Switching in Automatic Speech Recognition: The Issues and Future Directions. *Applied Sciences* 12, 19 (Jan. 2022), 9541. doi:10.3390/app12199541

[19] Surangika Ranathunga and Nisansa de Silva. 2022. Some Languages are More Equal than Others: Probing Deeper into the Linguistic Disparity in the NLP World. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Yulan He, Heng Ji, Sujian Li, Yang Liu, and Chua-Hui Chang (Eds.). Association for Computational Linguistics, Online only, 823–848. doi:10.18653/v1/2022.aacl-main.62

[20] Surangika Ranathunga, Nisansa De Silva, Dilith Jayakody, and Aloka Fernando. 2024. Shoulders of Giants: A Look at the Degree and Utility of Openness in NLP Research. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 519–529. doi:10.18653/v1/2024.acl-short.48

[21] Ehud Reiter. 2025. We Should Evaluate Real-World Impact. *Computational Linguistics* 51, 4 (Dec. 2025), 1419–1431. doi:10.1162/coli.a.18

[22] Thomas Reitmaier, Dani Kalarikalayil Raju, Ondrej Klejch, Electra Wallington, Nina Markl, Jennifer Pearson, Matt Jones, Peter Bell, and Simon Robinson. 2024. Cultivating Spoken Language Technologies for Unwritten Languages. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–17. doi:10.1145/3613904.3642026

[23] Fengyi Xu, Jun Ma, Nan Li, and Jack C.P. Cheng. 2025. Large Language Model Applications in Disaster Management: An Interdisciplinary Review. *International Journal of Disaster Risk Reduction* 127 (Sept. 2025), 105642. doi:10.1016/j.ijdrr.2025.105642