

Multilingual Speech-Based Human–Robot Interaction for Personalized Assistive Robotics in Smart Living Environments

Bowan Riley

Department of Computer Science, George Mason University, Fairfax, VA, USA.
rriley@gmu.edu

Lion Cehen

Department of Computer Science, University of Alabama at Birmingham, Birmingham, AL, USA.
leoncohen03@uab.edu

Abstract

Multilingual speech-based human-robot interaction is becoming a foundational capability for personalized assistive robotics in smart living environments. This paper presents a system-level examination of the architectural, infrastructural, and governance dimensions that shape such systems. Rather than focusing narrowly on algorithmic recognition performance, the analysis integrates advances in multilingual speech processing, natural language understanding, user modeling, and ambient intelligence. The discussion begins with the evolving role of speech as an interaction modality and reviews cross-lingual representation learning and transfer methods. It then develops a layered architecture that separates acoustic processing, language understanding, dialogue management, personalization, and smart environment orchestration. The paper emphasizes structural trade-offs between cloud and edge processing, synchronous and asynchronous dialogue, generic multilingual models and domain-specific adaptation, and individual personalization and collective fairness. Robustness concerns are examined through the lens of acoustic variability, code-switching, dialectal diversity, privacy exposure, and model bias. Governance and policy implications are considered in relation to data protection, medical assistive deployment, interoperability standardization, and long-term sustainability. Throughout, case illustrations from smart homes, elder care, rehabilitation, and ambient assisted living are used to highlight deployment realities. The paper concludes that sustainable multilingual assistive robotics depends not only on accuracy but on modular architectures, transparent governance, inclusive data practices, and continuous evaluation across heterogeneous living environments. Future work should integrate participatory design, federated personalization, and formal auditing mechanisms to balance utility, privacy, and fairness.

Keywords

human-robot interaction, multilingual speech processing, assistive robotics, smart living environments, natural language understanding, personalization, system architecture, governance.

1. Introduction

Human-robot interaction has matured from controlled laboratory demonstrations into practical deployments across homes, hospitals, and care facilities [1]. Speech is a central modality in

these environments because it allows users with limited mobility, reduced fine motor control, or low technical literacy to command a robot, request information, and express preferences without relying on physical interfaces. Spoken dialogue systems provide the interaction loop through which robotic actions can be negotiated, clarified, and revised [2]. When dialogue systems are integrated with assistive robots, they become more than command interfaces; they become mechanisms for collaborative planning, reminder management, social engagement, and environmental control. The quality of this collaboration depends on the system’s ability to interpret speaker intent across languages, dialects, accents, and interaction contexts [3].

Smart living environments add another layer of complexity. Sensors, actuators, appliances, and robots are increasingly connected through ambient intelligence platforms that observe activity patterns and adapt environmental conditions. In such settings, speech interaction cannot be treated as an isolated perception task. It must be coupled with situational awareness, user modeling, and safe actuation. Multilingual support is especially important in aging societies, multicultural households, and home care settings where users may speak different languages or alternate between languages. This paper therefore adopts a systems perspective rather than a component-level perspective. It examines how multilingual speech-based human-robot interaction can be structured for personalized assistive robotics while remaining robust, fair, governable, and sustainable. The discussion emphasizes architectural trade-offs, infrastructure choices, data governance, and policy implications rather than isolated model accuracy.

2. Background and Related Work

End-to-end deep speech architectures have substantially improved recognition across languages by directly learning acoustic and linguistic representations from large corpora [4]. Pretrained transformer language models then extended context-sensitive representation learning to many languages, enabling downstream systems to benefit from shared syntactic and semantic regularities [5]. Cross-lingual representation approaches allow multilingual transfer even without parallel corpora, reducing the need for expensive annotated data in every target language [6]. These developments are part of a broader shift toward layered neural representation learning [7], including convolutional architectures that transform raw signals into high-level features usable by subsequent components [8]. Together, these advances have made multilingual speech interfaces technically feasible, but their integration into assistive robotics raises distinct systems challenges.

Related work in assistive robotics has long recognized that social interaction is not merely a channel for task commands but a core part of user acceptance and long-term engagement. A robot that assists with medication reminders or mobility support must manage turn-taking, politeness, repair sequences, and emotional tone. When language is added to this social layer, errors in speech recognition can produce not only failed commands but also confusion, frustration, or unsafe actions. Multilingual interaction intensifies these concerns because the system must detect language changes, adapt pronunciation models, and resolve ambiguities across linguistic boundaries. Existing smart home research has examined activity recognition, sensor fusion, and automation rules, but relatively few deployments have combined multilingual spoken dialogue with personalized robotics in a coherent system design. The gap is not primarily algorithmic; it lies in the integration of language technologies with environmental control, user modeling, and governance structures.

3. System Architecture for Multilingual Speech-Based Human–Robot Interaction

A layered architecture is appropriate for multilingual speech-based assistive robotics because it decouples rapid perceptual processing from slower reasoning and personalization. Acoustic modeling using deep neural networks replaced Gaussian mixture models in many production systems, providing stronger robustness to speaker and environmental variability [9]. Recurrent architectures permitted temporal modeling of speech sequences, capturing coarticulation and prosodic cues that are important for understanding spontaneous commands [10]. Multilingual transformer encoders provide shared lexical and semantic spaces across languages, allowing a robot to process user input without necessarily maintaining separate monolingual stacks [11]. These capabilities can be organized into an interaction pipeline that includes audio capture, voice activity detection, automatic speech recognition, language identification, semantic parsing, dialogue state tracking, policy selection, robot action generation, and environment actuation.

The central architectural trade-off is between cloud-based and edge-based processing. Cloud processing offers access to large multilingual models, frequent updates, and centralized personalization, but it introduces latency, network dependency, and privacy risk. Edge processing improves responsiveness and keeps sensitive speech data within the home, but it constrains model size and adaptation capacity. Hybrid architectures are often more suitable for assistive robotics because they can run lightweight wake-word detection, language identification, and emergency command recognition locally while delegating complex semantic interpretation to cloud services under explicit user consent. A second trade-off concerns conversational control. Task-oriented dialogue systems can be designed with finite-state, frame-based, or statistical dialogue managers. Assistive scenarios benefit from a mixed approach in which safety-critical commands follow deterministic policies, while open-ended social conversation uses probabilistic dialogue management. This separation reduces the risk that a flexible language model produces unsafe robotic behavior.

The system architecture must also separate general multilingual competence from user-specific adaptation. Generic multilingual models can bootstrap understanding across languages, but they do not capture individual pronunciation patterns, vocabulary preferences, or assistive routines. A modular design therefore includes a personalization layer that stores user profiles, interaction histories, and preference weights separately from the shared speech and language components. This separation improves maintainability, facilitates privacy auditing, and allows personal models to be updated without retraining large foundation models. In this structure, the robot is not a monolithic agent but a coordination point among perception, interpretation, personalization, and environmental control services. Such modularity is essential for deployment in smart living environments where devices, sensors, and applications evolve independently.

4. Multilingual Speech Processing and Natural Language Understanding

Multilingual speech processing in assistive robotics requires more than translating a request from one language into another. It involves recognizing which language is being spoken, accommodating mixed-language utterances, and mapping surface variability into a stable representation of user intent. Massively multilingual sentence embeddings support cross-lingual transfer and zero-shot understanding across many languages, making it possible to build assistive systems that were not explicitly trained on every target language [12]. Transfer learning in natural language processing, including fine-tuning and adapter-based strategies, has shown that high-level linguistic knowledge learned in high-resource languages can improve performance in lower-resource settings [13]. These methods are especially relevant

for home care scenarios in which users speak regional dialects or heritage languages with limited digital resources.

Nevertheless, multilingual natural language understanding in smart environments faces challenges that are not fully addressed by transfer learning. Code-switching is common in multilingual households, where a user may begin a sentence in one language and complete it in another. A standard language identification front end may fail or introduce cascading errors if it forces a single language label on a mixed utterance. Dialectal variation, informal speech, and domain-specific vocabulary further complicate semantic parsing. Assistive domains often include medical terms, household object names, and personal abbreviations that are underrepresented in general corpora. Robust systems therefore require domain adaptation methods that combine small amounts of local data with large pretrained multilingual representations. This adaptation must occur continuously because user vocabulary and environmental context change over time. The design challenge is to balance adaptation speed with model stability, ensuring that personalization does not degrade the shared multilingual competence of the system.

5. Personalization and User Modeling

Personalization is a defining requirement of assistive robotics because user needs, preferences, and routines differ widely. Socially interactive robots have been studied as companions, tutors, and rehabilitation aids, and their effectiveness depends on the ability to interpret user state and adapt behavior accordingly [14]. Social intelligence in human-robot interaction involves recognizing not only what the user says but also how the interaction should proceed given social norms, emotional state, and long-term goals [15]. In multilingual speech-based systems, personalization extends to language choice, accent adaptation, preferred terms, and culturally specific politeness conventions. A user may prefer to receive medication reminders in one language but engage in social conversation in another. The system should infer these preferences from interaction history rather than requiring explicit configuration.

User modeling must be designed as a governance boundary as well as a technical component. Personal speech data provides rich signals for adaptation, but it also reveals health conditions, emotional states, daily routines, and social relationships. The system should separate stable user preferences from transient conversational content. Stable preferences can be stored in long-term profiles, while ephemeral speech content should be processed locally or deleted after use. Transparency is essential because users may not understand how their speech data influence robot behavior. Personalization mechanisms should include understandable summaries of what the system has learned and provide simple controls to correct or delete user models. Without such mechanisms, assistive robots risk becoming opaque and paternalistic rather than supportive and user-centered. The goal is not maximum adaptation but appropriate adaptation within negotiated boundaries.

6. Smart Living Environment Integration and Infrastructure

Smart living environments are characterized by distributed sensing, heterogeneous devices, and continuous activity. Ambient intelligence research has emphasized the need for context-aware services that blend into everyday life while supporting autonomy and well-being [16]. In assistive robotics, this means that speech interactions should be interpreted in relation to sensor events, user location, time of day, and ongoing activities. A command such as “turn on the light” is ambiguous without knowing which room the user occupies. A reminder about medication is more meaningful when the system knows whether the user is near the kitchen or

has already taken the dose. Multilingual natural language processing for smart home systems has begun to address the integration of language understanding with home automation patterns [17]. However, integration remains difficult because smart home platforms use different communication protocols, data formats, and security models.

Infrastructural choices strongly affect system performance and sustainability. A central smart home hub can simplify orchestration but creates a single point of failure and privacy concentration. Distributed middleware can improve fault tolerance but increases coordination complexity. In multilingual speech-based systems, the choice of infrastructure also affects where speech data are processed and how multilingual models are updated. Federated approaches can support personalization by training local models on edge devices while sharing only aggregated updates. This reduces privacy exposure and supports offline operation, but it complicates quality assurance and fairness monitoring. The deployment infrastructure must also account for power constraints, network variability, and long-term maintenance in ordinary homes. These practical factors are often more decisive than algorithmic sophistication in determining whether an assistive system remains usable over time.

Interoperability is another structural requirement. Assistive robots must communicate with lighting systems, door locks, thermostats, health devices, and emergency call platforms. Multilingual speech interfaces add a layer of semantic mapping between natural language and device actions. This mapping must be robust enough to handle paraphrases, incomplete commands, and multilingual expressions. Standards for semantic event representation and robot-to-device communication remain fragmented, making cross-vendor integration costly. A well-designed architecture therefore includes an abstraction layer that translates high-level user goals expressed in multiple languages into vendor-specific commands. This allows the language understanding components to remain independent from device protocols and supports incremental upgrades. Such architectural separation is essential for sustainable smart living deployments.

7. Robustness, Fairness, and Ethical Governance

Robustness in multilingual assistive robotics is not only a matter of speech recognition error rates. It includes the ability to recover from misunderstandings, reject unsafe actions, and maintain reasonable behavior under uncertain inputs. Fairness and machine learning scholarship has shown that automated systems can reproduce or amplify disparities when models are trained on unrepresentative data [18]. A multilingual speech system may perform well for standardized languages but poorly for dialects, accented speech, or older adults with atypical speech patterns. Bias audits in machine learning have documented systematic performance differences across demographic groups, and these differences can have serious consequences in care settings [19]. A speech interface that misunderstands a medication request or an emergency call is not merely inconvenient; it can be harmful.

Ethical governance must therefore be embedded in the system lifecycle rather than treated as a post-deployment concern. Language models can produce unsafe, toxic, or misleading content, and their deployment in assistive contexts requires careful risk assessment [20]. In robotic systems, language output is often coupled with physical action, which raises the stakes. A hallucinated instruction or a misrecognized command could lead to incorrect actuation. Governance mechanisms should include command verification, confirmation for high-risk actions, and fallback behaviors when confidence is low. In multilingual contexts, verification prompts must be available in the user's preferred language and should not assume that users

can respond fluently in a dominant language. Fairness considerations also extend to language policy: systems should not pressure users to shift to majority languages for reliable operation.

Transparency and contestability are essential governance principles. Users should be able to understand why a robot acted or failed to act, especially in safety-relevant situations. This requires logging interaction traces in a privacy-protecting manner and providing accessible explanations. Multilingual systems should offer explanations in multiple languages, but this introduces additional translation quality concerns. Independent auditing is necessary to evaluate whether fairness objectives are met across languages, age groups, accents, and care contexts. Auditing should examine not only aggregate accuracy but also failure modes that disproportionately affect vulnerable users. Governance frameworks should distinguish between low-risk social conversation and high-risk assistive actions, applying stricter oversight to the latter while preserving user autonomy and dignity.

8. Deployment, Sustainability, and Policy Implications

Deploying multilingual speech-based assistive robots in smart living environments involves a sequence of pilot studies, iterative refinement, and long-term support. Sustainability depends on maintenance procedures, local adaptation capacity, and integration with existing care networks. A system that requires continuous expert intervention is unlikely to remain viable in ordinary households. Policy frameworks must address data protection, medical device regulation, accessibility standards, and interoperability certification. In many jurisdictions, speech data collected in homes may qualify as personal health information when linked to care services. Policies should encourage privacy-preserving design while enabling enough data sharing for safety monitoring and quality improvement. Assistive robots are increasingly seen as part of social care infrastructure rather than consumer electronics, which changes the regulatory expectations around reliability and accountability.

Reimbursement and procurement policies influence deployment. If assistive robots are framed primarily as convenience devices, they may remain inaccessible to users with limited financial resources. If they are framed as medical devices, they face extensive certification processes that slow innovation. A balanced policy approach could support assistive robots through home care programs, insurance coverage, or public procurement while requiring evidence of safety, fairness, and multilingual competence. Policy should also promote open standards for vocabulary, intent representation, and device interoperability so that users are not locked into proprietary platforms. Open standards can lower deployment costs and support a competitive ecosystem of specialized components, including multilingual speech modules, safety monitors, and personalization services.

Long-term sustainability also requires attention to environmental and organizational factors. Multilingual models are computationally intensive, and continuous cloud inference can produce significant energy demands. Edge computing, model compression, and selective processing can reduce this burden, but they require sophisticated engineering trade-offs. Organizational sustainability involves training caregivers, family members, and users to interact with the system and understand its limitations. Assistive robots do not replace human care but redistribute tasks and introduce new forms of coordination. Policy should therefore support training and social infrastructure alongside technical deployment. By integrating technical architecture with governance, infrastructure planning, and policy incentives, multilingual speech-based human-robot interaction can become a practical and equitable component of smart living environments.

9. Conclusion

Multilingual speech-based human-robot interaction for personalized assistive robotics sits at the intersection of language technologies, social robotics, ambient intelligence, and care policy. The technical advances in multilingual speech recognition and natural language understanding now make such systems feasible, but feasibility alone does not ensure useful or fair deployment. A systems perspective reveals that architectural modularity, appropriate cloud-edge trade-offs, robust personalization boundaries, and context-aware environment integration are as important as model performance. Assistive robots must interpret user speech accurately while respecting privacy, handling multilingual variability, and coordinating safely with the broader smart living environment.

Sustainable deployment requires governance structures that address fairness, transparency, contestability, and accountability. Multilingual competence should not become a mechanism for excluding dialect speakers or pressuring users toward standardized languages. Instead, inclusive design should treat language diversity as a design resource and a governance requirement. Future work should integrate participatory design with users and caregivers, federated personalization that preserves privacy, and independent auditing frameworks that evaluate outcomes across heterogeneous populations. These efforts can support assistive robots that are not only capable of multilingual interaction but also trustworthy, adaptable, and genuinely supportive in the complex realities of everyday living.

References

1. Goodrich, M. A., & Schultz, A. C. (2007). Human-robot interaction: A survey. *Foundations and Trends in Human-Computer Interaction*, 1(3), 203-275.
2. Jurafsky, D., & Martin, J. H. (2009). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition* (2nd ed.). Prentice Hall.
3. Young, S., Gašić, M., Thomson, B., & Williams, J. D. (2013). POMDP-based statistical spoken dialogue systems: A review. *Proceedings of the IEEE*, 101(5), 1160-1179.
4. Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Chen, J., Chrzanowski, M., Coates, A., Damos, G., Elsen, E., Engel, J., Fan, L., Fougner, C., Hannun, A., Jun, B., LeGresley, P., ... Zhu, Z. (2016). Deep speech 2: End-to-end speech recognition in English and Mandarin. In *Proceedings of the 33rd International Conference on Machine Learning* (pp. 173-182). PMLR.
5. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019* (pp. 4171-4186). Association for Computational Linguistics.
6. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL 2020* (pp. 8440-8451). Association for Computational Linguistics.
7. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436-444.
8. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25* (pp. 1097-1105).

9. Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T., & Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Processing Magazine*, 29(6), 82-97.
10. Graves, A., Mohamed, A., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 6645-6649). IEEE.
11. Pires, T., Schlinger, E., & Garrette, D. (2019). How multilingual is multilingual BERT? In *Proceedings of ACL 2019* (pp. 4996-5001). Association for Computational Linguistics.
12. Artetxe, M., & Schwenk, H. (2019). Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7, 597-610.
13. Ruder, S., Peters, M. E., Swayamdipta, S., & Wolf, T. (2019). Transfer learning in natural language processing. In *Proceedings of NAACL-HLT 2019: Tutorials* (pp. 15-18). Association for Computational Linguistics.
14. Fong, T., Nourbakhsh, I., & Dautenhahn, K. (2003). A survey of socially interactive robots. *Robotics and Autonomous Systems*, 42(3-4), 143-166.
15. Dautenhahn, K. (2007). Socially intelligent robots: Dimensions of human-robot interaction. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362(1480), 679-704.
16. Cook, D. J., Augusto, J. C., & Jakkula, V. R. (2009). Ambient intelligence: Technologies, applications, and opportunities. *Pervasive and Mobile Computing*, 5(4), 277-298.
17. Chen, X. (2025). Research on the Application of Multilingual Natural Language Processing Technology in Smart Home Systems. *Journal of Computer, Signal, and System Research*, 2(5), 8-14.
18. Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and machine learning: Limitations and opportunities*. MIT Press.
19. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6), 1-35.
20. Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A., ... Gabriel, I. (2021). Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.