Flynn et al, Gerontechnology (2026) 25:3
https://doi.org/10.30289/gt.202604_25(3).1320

Talking to robots: A research agenda on how to improve the understanding by social-assistive robots of older people

Ronan Flynn PhD[a], Katrin Paldán PhD[b,]*
[a]Technological University of the Shannon, Athlone, Ireland; [b]Vorarlberg University of Applied Sciences, Dornbirn, Austria

**Corresponding author: katrin.paldan@fhv.at*

Abstract
The world's population is aging; life expectancy today is much longer than it was fifty years ago. Advances in medicine and improved knowledge regarding healthy living have all contributed to people living longer. Increased human longevity is a cause for celebration but also requires supportive environments that enable older people to maintain a desired level of independence and to manage chronic health conditions, the prevalence of which increases with age. With increased age, there is often a need for care for older (over 65) and very old (over 85) people, be that at home or in health care facilities. This in turn has led to research and development of social-assistive robots, which are designed to interact with older people to provide company and to perform assistive care tasks. The modality most commonly used for people to communicate with assistive robots is speech. However, variations in spoken dialect and the changes in speech articulation due to aging can impact the robot's speech recognition accuracy and its ability to correctly respond to a spoken command or request. The result of this is frustration when interacting with the assistant robot and, therefore, a lack of acceptance. To improve the interaction with, and the acceptance of, assistant robots by older people, the speech recognition algorithms used must consider the speech characteristics of older people. Following a narrative review of the literature, this paper presents a conceptual framework and a research agenda for more integrated approaches that simultaneously consider both age-related and dialect-related variability in order not to disadvantage older people in speech recognition.

Keywords: social-assistive robot, speech recognition, dialect, data augmentation, older people

Introduction

Life expectancy for people across the world is increasing. The World Health Organisation predicts that by 2050 the number of persons aged 60 years and older will be double what it was in 2020, and that the number of persons aged 80 or over will triple (WHO, 2025). Over the last fifty years, a growth in the percentage of the population aged 65 years and above can be seen in many countries around the world. According to the World Bank, the percentage of the total population in Japan aged 65 years and above increased from 6% in 1960 to 30% in 2021 (The World Bank, 2025). India, Brazil and Canada, with 7%, 10% and 19% of their populations aged 65 years and older in 2021, respectively, are other examples of this increase (The World Bank, 2025).

The population of those aged 65 years and above (referred to as older people) in the 27 states of the European Union (EU-27) is expected to increase by 43% between 2019 and 2050; the population of very old people (those aged 85 years and above) is expected to more than double in the same period (European Commission. Statistical Office of the European Union., 2020). Within the EU-27, as a percentage of their total populations, 15% of the people in Ireland and 19.5% of those in Austria fell into the older people category in 2022 (OECD, 2024); these are projected to increase to approximately 24% and 27% for Ireland and Austria, respectively, by 2050 (European Commission. Statistical Office of the European Union., 2020). In the Vorarlberg region of Austria, older people made up 17.5% of the population in 2020; this compares with 13.7% for the Midlands region in Ireland (Eurostat, 2020).

Increased life expectancy, leading to an increase in overall population ageing, has resulted from a number of factors, which include improvements in healthcare, occupational changes and the adoption of healthy lifestyles (Mathers et al., 2015). The global increase in life expectancy has created both opportunities and challenges for societies worldwide. The growth

in the older demographic of the population is positive and brings benefits for older people, their families and societies as a whole (WHO, 2025). However, with people living longer there is the challenge of providing the necessary care for an ageing population.

As the number of older people relative to the working-age population increases, the size of the workforce available to provide the care required for older people will decline (Eurostat, 2020). The number of those providing long-term care is falling behind what is required to keep pace with the number of older people who require long-term care services; by 2040 there will need to be an increase of 60% in the number of care workers across the OECD to maintain the 2022 ratio of carers to older people (OECD, 2020). Long-term care providers, including nursing homes and home care services, are increasingly using technology to support their staff and to improve the engagement of residents. Where assistant robots have been introduced to hospitals and nursing homes, studies have shown that they help to deal with loneliness and to form social connections with older people (OECD, 2020), (Pu et al., 2019).

Social-assistive robots (SARs) have emerged as a promising technology to complement human caregiving by providing companionship, supporting daily tasks, and fostering independence among older adults (Abdi et al., 2019) (Naseer et al., 2025) (Pu et al., 2019). Unlike disembodied voice assistants such as Amazon Alexa or Google Home, robots are embodied agents: they combine speech interaction with physical presence, gestures, and expressions. This embodiment amplifies the social expectations placed on them and makes the quality of interaction particularly critical for acceptance (Deng et al., 2019). Furthermore, the embodiment shapes user expectations, creates richer social interactions, and may either mitigate or exacerbate frustrations with speech recognition.

Prior research in social robotics has highlighted how embodiment can increase trust, companionship, and engagement among older adults. However, the embodied presence of robots amplifies the consequences of recognition errors. Unlike a voice assistant, which may be perceived as "just a device," assistant robots carry a stronger sense of agency and social presence. Yet, the intersection of embodiment and speech-based accessibility challenges remains underexplored. When a robot fails to understand a request, users may experience not only frustration but also disappointment or even loss of trust. For older adults, especially those who may rely on such technologies for companionship, this can have a significant impact on acceptance and sustained use. Design and interaction quality must therefore be closely coordinated to avoid disappointment and rejection.

There are many different modalities that have been implemented for human-robot interaction, including body movement, head movement, eye movement, touch screen, tactile, speech, gesture, lights, sound, facial expressions, screen and projection (Mahdi et al., 2022). Of these, speech is the most natural modality for human–robot communication, especially for older adults who may be less comfortable with touchscreens or complex interfaces (Kim, 2021).

With the development of robotics, spoken language interaction is becoming increasingly important; however, there are many challenges to be addressed before humans can interact fully with robots by simply talking with them (Marge et al., 2022a). For users, for whom speech might be the preferred way to interact with a robot, ease of use in terms of a robot's speech recognition is essential; this is particularly important for older users, whose speech might be affected by age-related impairments (Prinzellner et al., 2022). While speech has been widely used for human-robot interaction, the diversity of languages, dialects, age, and gender, in addition to age-related impairments, must all be addressed as part of the speech recognition performance (Mahdi et al., 2022). In order for healthcare robots to interact with older people through a spoken dialogue, a high-performance recognition system for speech from older people is needed. For older people interacting with healthcare robotic assistants through speech, an accurate speech recognition system is very important because such interaction is expected to help maintain their cognitive abilities, while also providing them with companionship (Tanioka et al., 2017).

However, speech interaction presents challenges such as dialectal diversity, age-related changes in articulation and prosody, and the lack of representative training data for speech recognition systems. These challenges can lead to frustration with and abandonment of the technology, undermining its potential benefits.

The paper's contribution to tackle these challenges is a research agenda that focuses on the following:

- The unique difficulties older adults face when interacting with SARs through voice.
- The opportunities and risks of data augmentation methods, particularly the use of synthetic speech to create training datasets with greater linguistic diversity.

- Highlighting how assistant robots require both technical advances and human-centred design considerations to ensure inclusive use, together with ethical and social requirements.

By bringing together these research strands, human–robot interaction, speech technology and aging, this paper aims to place the challenges and opportunities of speech interaction with assistant robots in a broader interdisciplinary landscape.

Methodological approach

As this work does not constitute a systematic review, it does not claim to be exhaustive in terms of the literature covered. The selection was primarily based on relevance and theoretical compatibility to four conceptual domains: (1) age related speech variability, (2) speech recognition modelling strategies under conditions of data scarcity, (3) embodiment and speech processing in socially assistive robots, and (4) ethical, legal and interactional challenges in improving speech interaction of socially assistive robotics with data augmentation methods.

A ten-year window (2016-2025) was used for the review of the literature in relation to older-adult speech, SARs, and speech recognition. This wider search window was chosen to provide context and understanding of how addressing the challenges associated with the automatic recognition of older-adult speech has evolved. The selection of literature considered was targeted and theory-driven. It was based on relevant publications in the fields of SARs, speech recognition systems, and speech databases, which were identified through the Scopus, IEEE, ACM, PubMed, and Google Scholar databases. Representative search terms included the following: older or elderly persons, automatic speech recognition, linguistic or dialect variation, speech database or corpus, assistant or assistive robot, human robot interaction, and data augmentation. A brief description of the strategy for each database used follows; example database queries are provided in Appendix I:

- Scopus – Basic search over all fields based on two or more search fields, within specified date range; further refinement of search results using subject area and document type. When a basic search was not sufficient, an advanced search using an appropriate query string was used.
- IEEE – Advanced search within metadata using two or more keywords with Boolean operators, within a specified date range; further refinement of the search results based on document type and publication.
- ACM – Advanced search anywhere with two or more terms linked by Boolean operators, within a specified date range; further refinement of the search results using publication and date filters.
- PubMed – Advanced Search Builder to form a query over all fields using two or more search terms linked by Boolean operators; further refinement of the search results based on custom date range and article type.
- Google Scholar – Advanced search in which required words and phrases "anywhere in the article" are given, within a specified date range.

State of research

Research on speech interaction for older adults has predominantly focused on voice assistants such as Amazon Alexa, Google Home, or Echo Show. Studies in human–computer interaction (HCI) and accessibility have shown both the potential and the limitations of these devices in everyday life (Mitzner et al., 2024). For example, Pradhan et al. identified barriers to healthcare and quality-of-life improvements through voice assistants (Pradhan et al., 2020), while Corbett et al. explored older adults' reasons for adopting, limiting, or abandoning such systems (Corbett et al., 2021). More recently, Chen et al. examined how screen-based versus screenless voice assistants influence older adults' engagement and found that "while participants appreciated the visual outputs, they still preferred to respond via speech instead of touch" (Chen et al., 2023). These studies reveal a complex picture of enthusiasm, frustration, and abandonment, pointing to the importance of accessibility-oriented design in speech interfaces and future work on voice-interfaces.

Challenges for automatic speech recognition of older people

Variations in dialect for a particular language can compromise speech recognition performance. Speakers from regions with unique dialects, older speakers whose articulation may be less clear, and individuals with motor disorders that affect speech production can all generate what is often classified as non-typical speech. When such input is fed to an automatic speech recognition (ASR) system trained only on typical speech, a mismatch occurs, leading to poor recognition and, consequently, feelings of exclusion and a reduced quality of life for these users when interacting with voice-enabled technologies.

Improving ASR robustness to dialectal variation requires a speech database that contains examples of the expected language variations for the recogniser training process. Collecting and annotating this data demands time and resources, especially when few examples exist for a particular dialect group (Beckford Wassink et al.,

2022). For German, three multi-dialect corpora illustrate the current landscape. The SpeechDat-AT database for Austrian German (Baum et al., 2000) combines fixed-line and mobile telephone recordings and was designed for pronunciation studies and automatic dialect identification. The Regional Variants of German (RVG1) corpus (Burger & Schiel, 1998) offers read and spontaneous speech from nine German dialects across Austria, Germany and Switzerland. The Austrian German GRASS corpus (Schuppler et al., 2017) focuses on conversational, read, and command speech from six Austrian regions. Both RVG1 and GRASS are valuable for studying language variability and for robot-based conversational systems.

Research on dialect-aware ASR spans many languages, including Chinese (Liu et al., 2008), English (Das et al., 2021), Arabic (Najafian et al., 2017), and Japanese (Fukuda et al., 2023; Fukuda et al., 2019). Dialect identification systems enable ASR engines to adapt acoustic, pronunciation, and language models, thereby improving recognition accuracy (Biadsy et al., 2010). Typical identification methods use phonotactic (phoneme statistics), lexical (n gram word or character statistics), or acoustic (spectral or prosodic) features (Khurana et al., 2017). Nonetheless, distinguishing dialects is harder than identifying languages because dialects within the same language are often highly similar (Biadsy et al., 2010).

Aging introduces further acoustic variability. Age-related changes in the speech production apparatus can produce reduced speech rates, inarticulate pronunciation, and breathy sounds, altering the acoustic profile relative to younger speakers. Building an acoustic model that represents older adults is difficult because large-scale training data from this cohort are scarce (Yamamoto et al., 2021). Consequently, systems trained on younger speech perform poorly on older voices (Fukuda et al., 2019).

The three German corpora mentioned above illustrate a systematic bias toward younger speakers. SpeechDat-AT contains 1,959 speakers, but only 2 % are over 60 years of age, while 53 % are ≤30 years old. RVG1 shows a similar pattern: 2 % of its 533 speakers are over 60 and 74 % are ≤30. GRASS, though smaller (38 speakers), is also skewed towards younger speakers - 84 % are ≤37 years.

As life expectancy rises, ASR systems must adapt to older populations. Japan, now a "super aged society", has pioneered older-focused corpora. The S-JNAS corpus includes 301 participants aged 60–90 (average = 67.6) (Fukuda et al., 2019). A nursing robot dialogue corpus (Fukuda et al., 2018) initially gathered 144 speakers aged 66–99 (average = 81.2) and later expanded to 221 speakers (average = 79.2) for dialect-dependent Japanese ASR research (Fukuda et al., 2019). Similar initiatives for other languages will become increasingly important as populations age.

Compiling large-scale corpora of older speakers faces several obstacles. Researchers often must travel to participants, raising time and cost. Unfamiliarity with recording equipment can cause stress and lengthy recording sessions may lead to fatigue or emotional upset. These practical constraints typically yield limited speech data for older adults, hindering model training. To raise recognition accuracy from sparse older-speaker data, it is necessary to adapt models originally trained on younger speech (Yamamoto et al., 2021).

Gathering representative dialect data poses analogous challenges. Some dialects are spoken by small minorities in narrow geographic areas. In SpeechDat-AT, nine Austrian regions are represented, but speaker distribution is uneven: Vienna (25 %) and Lower Austria (19 %) dominate, while Vorarlberg (6 %) and Tyrol (5 %) are minimally covered. In RVG1, speakers span 36 subregions across nine primary dialect zones; the most represented are Südbairisch/Südösterreichisch (8.42 %), Niederrheinisch (7.95 %) and Oberöschisch (7.86 %). The least represented are Rheinfränkisch (1.14 %), Moselfränkisch (1.11 %) and Schleswigisch (0.98 %). GRASS is even more skewed, with Styria accounting for 61 % of its 38 speakers and Upper Austria 16 %. The factors outlined above—age, dialect, gender, age related impairments, articulation effects and background noise—must be considered when designing speech interfaces for older users of assistant robots. *Table 1* underscores the under representation of older speakers and dialectal variation in existing corpora. Few datasets capture both ageing-related speech characteristics and dialect diversity, limiting the development of ASR systems robust to these combined sources of variability. This scarcity threatens to reinforce technological barriers for ageing populations, highlighting the need for more inclusive data collection and modelling strategies to ensure equitable access to speech-based technologies.

When data for a target group are limited, data augmentation techniques can expand the training set and lower the word error rate (WER). Approaches include synthetic speech generation, dialect modelling and the creation

Table 1. Comparative overview of speech corpora and studies addressing aging and/or dialect variation in ASR

Study / Corpus	Language	Speaker Age Distribution	Dialect / Accent Coverage	Recording Conditions	ASR Approach	Evaluation Metrics	Main Findings / Implications for ASR
[Author Year / Corpus Name]	[e.g., German]	[e.g., 18–85; limited >70]	[e.g., Standard only / regional variants]	[lab / home / noisy]	[HMM / DNN / end-to-end]	[WER etc.]	[core insight + relevance for aging/dialect]
Baba et al. (2001)	Japanese	Elderly vs. adult speakers	Standard Japanese	Controlled recordings	Elderly acoustic models (HMM-based LVCSR)	WER	Age-specific acoustic models significantly improve recognition for elderly speech.
Baum et al. (2000) – SpeechDat-AT	Austrian German	Adults; no specific elderly focus	Austrian German regional variation	Telephone speech; real-world conditions	HMM-based ASR (database for training/testing)	Recognition performance; corpus characteristics/ WER	Large telephone corpus capturing regional variation; useful for robust ASR development but underrepresents older speakers → limited aging analysis.
Beckford Wassink et al. (2022)	English	Not age-focused	Ethnicity-related dialects (e.g., African American English)	Mixed datasets; natural speech	Commercial ASR systems evaluation	WER differences	Demonstrates systematic performance disparities across dialects → highlights algorithmic bias and need for dialect-aware training data.
Biadsy et al. (2010)	Arabic	Not specified	Arabic dialects	Speech corpora (varied sources)	Discriminative phonotactic modelling using context-dependent phone classifiers	Dialect classification accuracy	Phonotactic features improve dialect recognition; supports dialect-sensitive ASR pipelines but no aging focus.
Burger & Schiel (1998) – RVG-1	German	Adult speakers; not elderly-focused	Regional variants of contemporary German	Controlled recordings	Corpus development for ASR training	Corpus coverage statistics	Provides structured data on German regional variation enabling dialect modelling; lacks age diversity.
Das et al. (2021)	English	Not specified	Multiple English dialects	Benchmark speech datasets	Attention-based ensemble of experts (deep learning)	WER	Multi-dialect modelling improves performance across dialects → supports specialised models for dialect diversity.
Fukuda et al. (2023) – Super-elderly Japanese corpus	Japanese	Super-elderly (80+)	Limited dialect variation	Controlled recordings	Acoustic modelling for elderly speech	WER	Super-elderly speech shows distinct acoustic properties; dedicated models
Fukuda et al. (2019) – Elderly Japanese corpus	Japanese	Elderly speakers (65+)	Includes dialect-dependent analysis	Controlled recordings	Acoustic modelling and dialect-dependent ASR evaluation	WER	Both age and dialect influence ASR performance; targeted data required for robust recognition.
Fukuda et al. (2018) – Elderly Japanese corpus construction	Japanese	Elderly speakers (65+)	Limited dialect variation	Controlled recordings	Corpus construction for elderly ASR	Corpus statistics	Introduces elderly speech corpus; highlights scarcity of age-diverse datasets for ASR.
Jin et al. (2024)	Multiple	Elderly and dysarthric speech	None	Speech datasets	Adversarial data augmentation	WER	Personalised augmentation improves recognition of atypical speech.
Johnson et al. (2022)	English	Children's speech	Dialect variation	Speech datasets	LPC-based augmentation	WER	Speaker variation modelling transferable to dialect challenges.
Khurana et al. (2017) – QMDIS dialect identification system	Arabic	Not age-focused	Arabic dialects	Broadcast and speech datasets	Advanced dialect identification system	Dialect classification accuracy	Accurate dialect identification improves downstream ASR performance in multi-dialect settings.
Liu et al. (2008)	Chinese	Not specified	Chinese dialectal variation	Speech corpora	State-dependent phoneme-based model merging	Recognition accuracy	Model merging improves recognition of dialectal speech → supports adaptation to linguistic variability.
Lounnas et al. (2022)	Algerian Arabic	Adults	Algerian dialect	Speech recordings	Audio augmentation	Recognition accuracy	Augmentation improves dialect recognition.
Najafian et al. (2017)	Arabic	Not specified	Multi-genre broadcast speech including dialects	Broadcast media	DNN-based ASR	WER	Multi-genre training improves robustness but dialect variation remains challenging.
Schuppler et al. (2017) – Austrian German corpus	Austrian German	Adults; limited elderly representation	Austrian German regional variation	Read and conversational speech	Corpus for ASR training and evaluation	Corpus statistics; recognition performance	Provides high-quality conversational dialect data; useful for regional modelling but limited age diversity.
Ye et al. (2021)	English / Mandarin (DementiaBank)	Elderly with neurocognitive disorders	None	Clinical recordings	Elderly speech recognition system	WER / classification	Cognitive impairment and aging strongly affect ASR performance.

of inclusive corpora. While these methods can improve recognition across diverse user groups, they also raise concerns about bias amplification, loss of authenticity, and ethical implications of synthetic data.

In summary, dialectal variation and ageing-related speech changes both degrade ASR performance when not reflected in training data. Existing German corpora exhibit strong biases toward younger speakers and uneven dialect coverage; Japanese research demonstrates the benefits of dedicated older-speaker corpora. Overcoming collection challenges—through targeted recruitment, community-based recording and careful augmentation—will be essential for building ASR systems that serve all speakers fairly.

Data augmentation for speech diversity as an ASR modelling strategy

Collecting large speech corpora from older adults (≥ 65 yr) and the very old (≥ 85 yr) is

hampered by mobility limitations, fatigue and the logistical cost of travelling to participants. These constraints create a data-scarcity problem that directly harms the performance of ASR systems for ageing populations. Data-augmentation techniques therefore play a pivotal role (Jin et al., 2024). By artificially generating new utterances from existing recordings, augmentation can increase the size and diversity of training material, improving the generalisation, accuracy, and robustness of acoustic models (Yiwere et al., 2023; Gref et al., 2019.)

Typical augmentation operations alter the speed, timing, pitch, or spectral characteristics of a signal (Fucci et al., 2023; Lounnas et al., 2022; Ravi et al., 2024). Such perturbations introduce variability that mirrors real world differences in dialect, articulation and age-related vocal changes, thereby reducing the WER for under-represented speakers (Johnson et al., 2022). Many older speakers also use local dialects; when an ASR system relies on a standardised language model, dialectal mismatches can cause severe recognition failures (Cavallaro et al., 2024).

Vocal tract length perturbation (VTLP)
VTLP warps the Mel-frequency axis of each utterance by a randomly selected factor, producing a spectrally altered replica (Cui et al., 2015). Ko et al. (2015) combined VTLP with tempo perturbation (changing speech tempo without affecting pitch or spectral envelope). This speed perturbation pipeline yielded lower WERs than either technique alone. Subsequent studies have confirmed its utility: Gref et al. (2019) used speed perturbation to triple the amount of training data in a two-stage acoustic modeling system for robust German oral history ASR, coupling augmentation with transfer learning; Ye et al. (2021) expanded a small older-speaker corpus fourfold for neurocognitive disorder detection; Khurana et al. (2017) employed speed and volume perturbation to generate three times the original data for dialect identification.

Augmentation for low-resource languages and dialects
Alumäe et al. (2023) aligned conventional audio perturbations with text-to-speech synthesis to boost linguistic diversity in Indian language ASR, where both acoustic and textual resources are scarce. Lounnas et al. (2022) enhanced an Arabic digit recogniser for two Algerian speakers by stretching signals, adding white noise backgrounds, and shifting pitch, yielding noticeable accuracy gains. Feng et al., (2024) applied changes in speech speed, noise injection, and pitch shifting to triple the training set for dialectal Arabic ASR, improving deep learning model performance.

Bias arising from speaker and dialect homogeneity
Modern ASR systems are typically trained on native speakers using the most common language variant, inadvertently marginalising regional dialects and sociolinguistic groups (Feng et al., 2024). Age imbalance compounds this bias: training corpora dominated by speakers under 40 yr exclude the acoustic patterns of older adults (e.g., slower rates, longer vowel durations, altered fundamental frequency F_0). Data augmentation can partially remedy these gaps by synthetically expanding the under-represented cohort. An alternative route is deep transfer learning, where knowledge acquired from a source domain (e.g., ASR for younger speakers) is transferred to a target domain (e.g., ASR for speakers > 65 yr), enhancing robustness and generalisation (Kheddar et al., 2023).

Case study: older-speaker Japanese corpus
Fukuda et al. (2023) highlighted the need for age-specific data by creating a Japanese corpus with an average participant age of 83.1 yr. Two augmentation methods were employed: speech rate perturbation (factors 0.9 and 1.1) and SpecAug (Park et al., 2019), which masks portions of the spectrogram. The augmented corpus yielded measurable improvements over a non-augmented baseline. Future extensions could incorporate formant modification techniques (Kathania et al., 2022) to better emulate the vocal tract changes typical of the very old, especially for care home applications.

In summary, data augmentation offers a practical solution to the twin challenges of dialectal diversity and ageing-related speech variation. While speed perturbation and VTLP constitute well-established tools, emerging strategies—such as demographic-aware synthesis, combined acoustic-textual augmentation and targeted formant manipulation—promise to close the performance gap for older, dialect-rich users. Continued research should focus on evaluating the trade-offs between synthetic diversity and potential bias amplification, ensuring that augmentation pathways truly enhance inclusive, high-quality speech interaction for all speaker groups.

Embodiment and speech processing in socially assistive robots
Recent research in socially assistive robotics increasingly emphasises the role of embodiment—that is, a robot's physical and interactive presence—as a central factor shaping social effectiveness and user acceptance. Physically embodied robots elicit stronger social responses than virtual agents, improving likeability, engagement, trust, and communication quality (Céspedes et al., 2021; Gava et al., 2020).

Embodiment also directly influences how users perceive and respond to speech, emotional expression, and multimodal cues, with studies showing higher arousal and preference for embodied robots in speech-based interaction tasks (Ko et al., 2023; Klüber & Roesler, 2025). In care contexts, embodied socially assistive robots rely heavily on speech-based communication and require accurate speech recognition to maintain natural interaction, especially with older adults whose speech characteristics and dialectal diversity pose challenges for current ASR systems. Together, these findings highlight that embodiment and speech processing capabilities jointly shape the social presence, usability, and real-world acceptance of socially assistive robots.

Embodiment encompasses features such as physical expressiveness, gesture, spatial movement and the coordination of multimodal feedback signals. Recent studies show that these characteristics not only shape the subjective experience of interaction but also impose concrete requirements on technical modules such as speech recognition, dialogue management and attention or interaction management.

Ethical and legal considerations

The use of data-driven systems to provide social support to vulnerable groups is subject to ethical and legal obligations. The literature highlights, in particular, key requirements for the protection of personal data, the management of contextual data, the involvement of users in ongoing ethical assessment and the identification of risks arising from data augmentation processes. Together, these considerations form the basis for responsible, transparent and trust-building system design.

Personalisation and data protection

Speech recognition systems deployed in care environments may benefit from recognising and adapting to individual interaction partners. In settings such as care homes, where residents, care staff and visitors interact with the system in different ways, the ability to personalise speech recognition and interaction behaviour can improve usability and reduce misunderstandings. For example, personalised responses may enable the system to adapt to individual communication styles or preferences when performing actions such as playing music or providing reminders. However, such personalisation requires the storage and processing of user-related data (Moore & Pham, 2015). This raises important ethical and legal challenges regarding the protection of sensitive personal information. In care contexts, speech data may reveal health-related information, emotional states or daily routines, making robust data protection mechanisms essential.

Contextual data and privacy implications

Further improvements in speech recognition performance could be achieved through the integration of contextual information. Systems that support longer conversations or complex interactions may benefit from contextual awareness, such as information about the time of day, location, background noise or the number of people present in a room.

However, the continuous collection and storage of contextual data also raises privacy concerns. Such data can reveal detailed behavioural patterns and therefore require appropriate safeguards, including anonymisation, secure storage and, where possible, local on-device processing. Minimising the collection and retention of sensitive data is particularly important in environments involving vulnerable users.

User participation and ethical feedback mechanisms

Learning systems often rely on user feedback to improve performance over time. Involving older adults in such feedback processes can therefore be beneficial for improving speech recognition systems in real-world settings. At the same time, these processes must be designed carefully to respect ethical principles such as the avoidance of harm (World Medical Association, 2013).

In particular, feedback mechanisms should not require extensive effort or prolonged interaction that could become exhausting for older users. Furthermore, the design of interfaces should take into account potential audio, visual, motor, cognitive, or technological barriers that may affect older adults who are less familiar with digital technologies. Designing accessible interaction mechanisms can require substantial development effort and resources, which may present challenges for system developers.

Ethical risks of data augmentation

While data augmentation techniques can significantly improve the robustness and performance of speech recognition models, they also raise several ethical considerations. These include questions of fairness and representation, transparency of model training processes, and the potential misuse of synthetic or manipulated speech data.

For example, augmentation methods could be used to artificially manipulate speech characteristics or generate misleading audio content, which may pose risks in security-sensitive applications

such as speech biometrics (Firc et al., 2023); Bajwa et al., 2025). It is therefore essential to ensure that data used for augmentation has been collected and processed in an ethically and legally compliant manner (Romero Moreno, 2024), including the informed consent of data sources.

The ethical implications of optimising speech recognition systems are often difficult to classify as clearly beneficial or harmful. This is particularly true when considering the long-term societal effects of deploying such technologies in sensitive contexts such as elder care. Addressing these ethical concerns requires not only technical solutions but also appropriate regulatory guidance.

Ensuring transparency, fairness, and responsible data governance therefore remains essential when applying data augmentation techniques in speech recognition research and development. Current EU directives provide frameworks for the research and development of artificial intelligence (AI) systems and hold developers and researchers accountable (European Commission, 2019; European Commission, 2025).

Based on the narrative review and the state of research, we claim that there is a need to improve linguistic interaction between older people and social assistance robots.

Conceptual framework & research agenda

Figure 1 summarises the causal chain that drives our work:

1. Speaker variability – Older adults (≥ 65 yr) and speakers of regional dialects are systematically under-represented in most speech corpora. Age-related changes (slower rate, longer vowel durations, altered F0) and dialectal phonetic patterns create acoustic profiles that differ markedly from the "standard" data used to train most ASR systems.
2. ASR performance – The mismatch between training data and real-world input leads to higher WERs and erratic decoding, especially in noisy care home environments.
3. User experience – Degraded recognition reduces interaction fluidity, lowers trust, and jeopardises acceptance of SARs.
4. Interventions—Data augmentation and model adaptation can bridge the data gap.
5. Ethical and legal framework – Any interventions must respect all relevant ethical and legal constraints (privacy, consent, bias mitigation).

The research agenda in *Table 2* is organised around five interlocking strands that together aim to raise acoustic model accuracy for older speakers who use dialectal speech.

By following this tightly coupled framework—starting from a clear conceptual model, moving through systematic data enrichment and rigorous validation, and ending with a deployment strategy that balances accuracy, latency and privacy—we can build socially assistive robots that understand and respect the linguistic diversity of older adults.

Summary of actionable steps:

1. Audit existing corpora for age and dialect gaps; flag under-represented groups.
2. Collect a seed set of older-speaker recordings (target ≤ 30 min per dialect) in realistic care home environments.
3. Apply the augmentation toolbox (speed/VTLP, noise, SpecAug, formant shift) to expand the seed set by a factor of three, minimum.
4. Validate synthetic audio using both objective WER against a held-out real subset and MOS listening tests; discard artefact-laden samples.
5. Train acoustic models with the augmented corpus; employ transfer learning from a high-resource base model to retain generalisation.
6. Evaluate on three fronts: (a) standard WER; (b) robustness under simulated noise; (c) usability, based on the System Usability Scale (SUS) with older participants.
7. Report fairness metrics per age-dialect slice; if disparities exceed a predefined threshold, iterate augmentation or collect additional real data.
8. Deploy a hybrid cloud-edge ASR stack; monitor post-deployment performance in living-lab pilots and feed real usage data back into the augmentation loop.

Discussion

Across automatic speech recognition research, data augmentation is defined as any technique that expands the representational space when natural speech data are sparse (Bartelds et al., 2023). Whether the method perturbs tempo, reshapes the spectral axis (VTLP), injects noise, or synthesises new utterances with text-to-speech, the common purpose is to compensate for systematic under-representation of specific speaker groups. For older adults and regional dialect speakers, this scarcity is a primary cause of the pronounced drop in recognition accuracy reported in the literature (Feng et al., 2024; Liu et al., 2008). By deliberately injecting controlled variability, we obtain a synthetic expansion that mitigates the train-test mismatch described by Johnson et al. (2022).

Collecting a minimum seed corpus that captures the core acoustic traits of the target population (age, dialect, gender) is a prerequisite for any augmentation pipeline (Yamamoto et al., 2021). Recording seniors is costly: mobility limitations,



Figure 1. A framework for inclusive speech recognition.

fatigue, and the need to travel to geographically dispersed dialect zones restrict the number of usable recordings. A pragmatic approach is to record a modest, balanced set (≈ 30 minutes per dialect) with realistic care home acoustics and then increase it by a factor of 3–5 using augmentation methods such as speed/VTLP perturbation, noise injection, time-frequency masking, and formant or vocal tract length shifting (Cui et al., 2015; Fucci et al., 2023; Kathania et al., 2022). Even with a small seed, the augmented pool can cover most acoustic variations that cause mismatches—background noise, speech-rate slowdown, and age-related spectral shifts—while keeping computational and storage demands within the capacity of modern research clusters (Lang, 2023).

The choice of augmentation technique should be tailored to the deployment context. Speed and VTLP warping are especially effective for older speakers who tend to speak more slowly (Cui et al., 2015). Noise and spectral shifting techniques improve robustness to the HVAC and fan noises typical of care facilities (Fucci et al., 2023; Lima et al., 2025). SpecAug or other time-frequency masking forces the model to rely on broader context, which is beneficial when the seed set is limited (Park et al., 2019). Formant or VTLP directly mimics physiological changes of ageing (Kathania et al., 2022). When lexical diversity is also lacking, text-to-speech synthesis can generate entirely new utterances (Alumäe et al., 2023). The optimal mix of these methods is language- and dialect-dependent and should be identified through iterative ablation studies (Kheddar et al., 2023).

Before synthetic utterances are used for model training, they must be validated both objectively and subjectively. Objective validation includes computing WER on a held-out real subset and applying statistical similarity tests (Kolmogorov-Smirnov, Pearson correlation) to confirm that feature distributions of augmented and genuine speech overlap (Gupta et al., 2025). Subjective validation involves mean opinion score (MOS) listening tests with raters familiar with the target dialect; samples exhibiting clicks, filler artefacts or unnatural prosody are discarded (Soleymanpour et al., 2023) Any residual artefacts can be removed with speech enhancement pipelines (Xue et al., 2024). Only augmented utterances that pass both objective and subjective thresholds should be fed to the acoustic model.

While WER remains a useful baseline (Hu et al., 2024; Nasretdinov et al., 2025; Ni et al., 2025), it does not capture the user-centric qualities required for SARs. Additional evaluation dimensions are needed: robustness under simulated care home noises (Lima et al., 2025); usability measured with the SUS from older participants (Blavette et al., 2025); and acceptance or trust assessed through qualitative interviews that probe cultural expectations, robot appearance,

Table 2. Research Agenda with five inter locking stands

Strands	Goal	Core actions	Key references
1. Corpus enrichment	Provide representative training material.	• Gather targeted recordings from seniors in care settings. • Prioritise dialectal diversity (e.g., Austrian German, regional Japanese).	Beckford Wassink et al., 2022
2. Speech-data augmentation	Multiply and diversify the limited data.	• Speed, time, pitch, and frequency warping (VTLP, speed perturbation). • Additive white noise, spectral shifting, pitch-preserving stretch. • Time-frequency masking (SpecAug) and other perturbations.	Cui et al., 2015; Fucci et al., 2023; Johnson et al., 2022
3. Validation of synthetic speech	Ensure augmented utterances remain realistic and dialect-faithful.	• Objective: WER against a held-out real set; statistical similarity tests (Kolmogorov-Smirnov, correlation). • Subjective: MOS listening tests with trained raters. • Pre-processing: speech enhancement to remove artefacts.	Soleymanpour et al., 2023; Gupta et al., 2025; Xue et al., 2024).
4. Holistic system evaluation	Look beyond WER to the dimensions that matter for older users.	• Robustness – test under realistic background noises (HVAC, fans) • Usability & acceptance – system usability scale (SUS), qualitative interviews, iterative co-design. • Cultural fit – robot appearance, speech style, and interaction norms. • Report group-specific WER (age × dialect) and worst group performance.	Hu et al., 2024; Lee et al., 2025; Lima et al., 2025, Blavette et al., 2025;
5. Fairness & intersectionality	Detect and mitigate systematic disparities.	• Use small, targeted evaluation sets of older dialect speakers. • When real data are scarce, apply pitch modification, formant shifting, VTLP or voice conversion to simulate variability. • Continuous post-deployment monitoring in living-lab pilots to measure task success and adjust models.	Abboud & Oz, 2024; Nakatumba-Nabende et al., 2025; Kheddar et al., 2023; Kathania et al., 2022

and speech style (Lee et al., 2025; Zafrani et al., 2024). An iterative, user-centred design loop—where older adults test prototypes and provide feedback—ensures that technical gains translate into real-world adoption.

Because older dialect speakers are doubly under-represented, fairness must be assessed intersectionally (Nakatumba-Nabende et al., 2025). Studies should report group-specific WER (age × dialect) and highlight the worst group performance. Targeted evaluation sets composed solely of older dialect speakers can reveal systematic gaps even when overall data volumes are modest (Abboud & Oz, 2024). When real data are scarce, demographic-aware transformations—pitch modification, formant shifting, VTLP, or voice conversion—can simulate variability, though they never fully replace genuine recordings (Kathania et al., 2022). Post-deployment monitoring in living-lab pilots, with task success rates as a fairness indicator, provides continuous feedback for model refinement.

The deployment architecture introduces an additional trade-off. Cloud-based ASR offers state-of-the-art models, large vocabularies, and automatic updates, but incurs network latency that can exceed conversational thresholds (Marge et al., 2022) and raises privacy concerns because speech is transmitted externally (Farahani et al., 2023). On-device (edge) ASR eliminates latency and keeps data local, reducing privacy risks, yet it is constrained by processing power and memory, often yielding smaller vocabularies and lower accuracy. A hybrid solution—running a lightweight on-device decoder for routine commands and falling back to the cloud for complex or out-of-vocabulary utterances when connectivity permits—balances low latency, privacy, and performance, especially when the on-device model has been fine-tuned on the augmented older-dialect corpus.

In summary, improving speech recognition for older adults interacting with SARs requires (i) a modest yet representative seed corpus, (ii) a disciplined augmentation pipeline that respects the acoustic signatures of age and dialect, (iii) rigorous validation of synthetic data, (iv) evaluation that goes beyond WER to encompass robustness, usability and trust, (v) explicit intersectional fairness reporting, and (vi) a deployment strategy that reconciles accuracy, latency and privacy. Addressing these interconnected components moves the field from "recognition works for the majority" towards "recognition works for everyone, regardless of age or dialect."

Conclusion

Overall, the literature reveals a clear need for more inclusive speech resources and integrative modelling approaches that jointly consider aging, dialect diversity and real-world recording conditions.

The agenda outlined above translates the empirical shortcomings identified in the literature into a concrete, multi-disciplinary program of work. By jointly advancing acoustic modelling, validation protocols, fairness diagnostics and deployment strategies, we aim to deliver speech-enabled SARs that are accurate, trustworthy and inclusive for the older adults who stand to benefit most from assistive robotics.

We invite collaborations with speech technology labs, gerontology research groups and care facility partners to accelerate progress toward these goals.

References

Abboud, K., & Oz, G. (2024). Towards Equitable Natural Language Understanding Systems for Dialectal Cohorts: Debiasing Training Data. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024) (pp. 16487–16499). ELRA and ICCL. https://aclanthology.org/2024.lrec-main.1433/

Abdi, S., Spann, A., Borilovic, J., de Witte, L., & Hawley, M. (2019). Understanding the care and support needs of older people: A scoping review and categorisation using the WHO international classification of functioning, disability and health framework (ICF). BMC Geriatrics, 19(1), 195. https://doi.org/10.1186/s12877-019-1189-9

Alumäe, T., Kong, J., & Robnikov, D. (2023). Dialect Adaptation and Data Augmentation for Low-Resource ASR: TalTech Systems for the MADASR 2023 Challenge (arXiv:2310.17448). arXiv. https://doi.org/10.48550/arXiv.2310.17448

Baba, A., Yoshizawa, S., Yamada, M., Lee, A., & Shikano, K. (2001). Elderly acoustic model for large vocabulary continuous speech recognition. 7th European Conference on Speech Communication and Technology (Eurospeech 2001), 1657–1660. https://doi.org/10.21437/Eurospeech.2001-206

Bajwa, M. T. T., Tehreem, F., Farid, Z., Tahir, H. M. F., & Khalid, A. (2025). DEEPFAKE VOICE RECOGNITION: TECHNIQUES, ORGANIZATIONAL RISKS AND ETHICAL IMPLICATIONS. Spectrum of Engineering Sciences, 106–121.

Bartelds, M., San, N., McDonnell, B., Jurafsky, D., & Wieling, M. B. (2023). Making More of Little Data: Improving Low-Resource Automatic Speech Recognition Using Data Augmentation. 715–729. https://doi.org/10.48550/arXiv.2305.10951

Baum, M., Erbach, G., & Kubin, G. (2000). SpeechDat-

AT: A telephone speech database for Austrian German. Proc. LREC'2000 Satellite Workshop XLDB - Very Large Telephone Speech Databases.

Beckford Wassink, A., Gansen, C., & Bartholomew, I. (2022). Uneven success: Automatic speech recognition and ethnicity-related dialects. Speech Communication, 140, 50–70. https://www.sciencedirect.com/science/article/pii/S0167639322000486?via%3Dihub

Biadsy, F., Soltau, H., Mangu, L., Navratil, J., & Hirschberg, J. (2010). Discriminative Phonotactics for Dialect Recognition Using Context-Dependent Phone Classifiers. paper 44. https://www.isca-archive.org/odyssey_2010/biadsy10_odyssey.html#

Blavette, L., Dacunha, S., Alameda-Pineda, X., Hernández García, D., Gannot, S., Gras, F., Gunson, N., Lemaignan, S., Polic, M., Tandeitnik, P., Tonini, F., Rigaud, A.-S., & Pino, M. (2025). Acceptability and Usability of a Socially Assistive Robot Integrated With a Large Language Model for Enhanced Human-Robot Interaction in a Geriatric Care Institution: Mixed Methods Evaluation. JMIR Human Factors, 12, e76496. https://doi.org/10.2196/76496

Burger, S., & Schiel, F. (1998). RVG 1—A Database for Regional Variants of Contemporary German. https://www.researchgate.net/publication/2676232_RVG_1_-_A_Database_for_Regional_Variants_of_Contemporary_German

Cavallaro, A., Perillo, F., Romano, M., Sebillo, M., & Vitiello, G. (2024). Social robot in service of the cognitive therapy of elderly people: Exploring robot acceptance in a real-world scenario. Image and Vision Computing, 147, 105072. https://doi.org/10.1016/j.imavis.2024.105072

Céspedes, N., Irfan, B., Senft, E., Cifuentes, C. A., Gutierrez, L. F., Rincon-Roncancio, M., Belpaeme, T., & Múnera, M. (2021). A Socially Assistive Robot for Long-Term Cardiac Rehabilitation in the Real World. Frontiers in Neurorobotics, 15. https://doi.org/10.3389/fnbot.2021.633248

Chen, C., Lifset, E. T., Han, Y., Roy, A., Hogarth, M., Moore, A. A., Farcas, E., & Weibel, N. (2023). Screen or No Screen? Lessons Learnt from a Real-World Deployment Study of Using Voice Assistants With and Without Touchscreen for Older Adults. ASSETS. Annual ACM Conference on Assistive Technologies, 2023, 52. https://doi.org/10.1145/3597638.3608378

Corbett, C. F., Combs, E. M., Wright, P. J., Owens, O. L., Stringfellow, I., Nguyen, T., & Van Son, C. R. (2021). Virtual Home Assistant Use and Perceptions of Usefulness by Older Adults and Support Person Dyads. International Journal of Environmental Research and Public Health, 18(3), 1113. https://doi.org/10.3390/ijerph18031113

Cui, X., Goel, V., & Kingsbury, B. (2015). Data augmentation for deep neural network acoustic modeling. IEEE/ACM Trans. Audio, Speech and Lang. Proc., 23(9), 1469–1477. https://doi.org/10.1109/TASLP.2015.2438544

Das, A., Kumar, K., & Wu, J. (2021). Multi-Dialect Speech Recognition in English Using Attention on Ensemble of Experts. ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 6244–6248. https://doi.org/10.1109/ICASSP39728.2021.9413952

Deng, E., Mutlu, B., & Mataric, M. (2019). Embodiment in Socially Interactive Robots. Foundations and Trends® in Robotics, 7(4), 251–356. https://doi.org/10.1561/2300000056

European Commission. (2025). Living guidelines on the responsible use of generative AI in research. ERA Forum Stakeholders' document [2nd ed.]. European Commission. Directorate-General for Research and Innovation. https://research-and-innovation.ec.europa.eu/document/download/2b6cf7e5-36ac-41cb-aab5-0d32050143dc_en?filename=ec_rtd_ai-guidelines.pdf

European Commission, H.-L. E. G. on A. I. (2019). Ethics guidelines for trustworthy AI. European Commission. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

European Commission. Statistical Office of the European Union. (2020). Ageing Europe: Looking at the lives of older people in the EU : 2020 edition. Publications Office. https://data.europa.eu/doi/10.2785/628105

Eurostat. (2020). Eurostat / Regions and Cities Illustrated (RCI). Eurostat. https://ec.europa.eu/eurostat/cache/RCI/

Farahani, B., Tabibian, S., & Ebrahimi, H. (2023). Toward a Personalized Clustered Federated Learning: A Speech Recognition Case Study. IEEE Internet of Things Journal, 10(21), 18553–18562. https://doi.org/10.1109/JIOT.2023.3292797

Feng, S., Halpern, B. M., Kudina, O., & Scharenborg, O. (2024). Towards inclusive automatic speech recognition. Computer Speech & Language, 84, 101567. https://doi.org/10.1016/j.csl.2023.101567

Firc, A., Malinka, K., & Hanáček, P. (2023). Deepfakes as a threat to a speaker and facial recognition: An overview of tools and attack vectors. Heliyon, 9(4), e15090. https://doi.org/10.1016/j.heliyon.2023.e15090

Fucci, D., Gaido, M., Negri, M., Cettolo, M., & Bentivogli, L. (2023). No Pitch Left Behind: Addressing Gender Unbalance in Automatic Speech Recognition through Pitch Manipulation (arXiv:2310.06590). arXiv. https://doi.org/10.48550/arXiv.2310.06590

Fukuda, M., Nishimura, R., Nishizaki, H., Horii, K., Iribe, Y., Yamamoto, K., & Kitaoka, N. (2023). A new speech corpus of super-elderly Japanese for acoustic modeling. Computer Speech & Language, 77, 101424. https://doi.org/10.1016/j.csl.2022.101424

Fukuda, M., Nishimura, R., Nishizaki, H., Iribe, Y., & Kitaoka, N. (2019). A New Corpus of Elderly Japanese Speech for Acoustic Modeling, and a Preliminary Investigation of Dialect-Dependent Speech Recognition (p. 6). https://doi.org/10.1109/O-COCOSDA46868.2019.9041216

Fukuda, M., Nishizaki, H., Nishimura, R., Iribe, Y., & Kitaoka, N. (2018). Construction of a Corpus for

Elderly Japanese Speech Recognition. 2018 IEEE 7th Global Conference on Consumer Electronics (GCCE), 687–688. https://doi.org/10.1109/GCCE.2018.8574795

Gava, L., Grassi, L., Lagomarsino, M., Recchiuto, C., & Sgorbissa, A. (2020). Physical Embodiment of Conversational Social Robots. 2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), 456–463. https://doi.org/10.1109/RO-MAN47096.2020.9223591

Gref, M., Schmidt, C., Behnke, S., & Köhler, J. (2019). Two-Staged Acoustic Modeling Adaption for Robust Speech Recognition by the Example of German Oral History Interviews. 2019 IEEE International Conference on Multimedia and Expo (ICME), 796–801. https://doi.org/10.1109/ICME.2019.00142

Gupta, S., Khanna, A., Mukhopadhyay, B., & Dubey, S. (2025). WGAN Based Synthetic Data Augmentation of Biomedical Voice Measurements for Parkinson's Disease. International Conference on Education Technology Management. https://www.semanticscholar.org/paper/WGAN-Based-Synthetic-Data-Augmentation-of-Voice-for-Gupta-Khanna/c0a622c375a9664d199e7dcabeb4509409a65eaf

Hu, S., Xie, X., Geng, M., Jin, Z., Deng, J., Li, G., Wang, Y., Cui, M., Wang, T., Meng, H., & Liu, X. (2024). Self-Supervised ASR Models and Features for Dysarthric and Elderly Speech Recognition. IEEE/ACM Trans. Audio, Speech and Lang. Proc., 32, 3561–3575. https://doi.org/10.1109/TASLP.2024.3422839

Johnson, A., Fan, R., Morris, R., & Alwan, A. (2022). LPC Augment: An LPC-Based ASR Data Augmentation Algorithm for Low and Zero-Resource Children's Dialects (arXiv:2202.09529). arXiv. https://doi.org/10.48550/arXiv.2202.09529

Kathania, H. K., Kadiri, S. R., Alku, P., & Kurimo, M. (2022). A formant modification method for improved ASR of children's speech. Speech Communication, 136, 98–106. https://doi.org/10.1016/j.specom.2021.11.003

Kheddar, H., Himeur, Y., Al-Maadeed, S., Amira, A., & Bensaali, F. (2023). Deep transfer learning for automatic speech recognition: Towards better generalization. Knowledge-Based Systems, 277, 110851. https://doi.org/10.1016/j.knosys.2023.110851

Khurana, S., Najafian, M., Ali, A., Hanai, T. A., Belinkov, Y., & Glass, J. (2017). QMDIS: QCRI-MIT Advanced Dialect Identification System. 2591–2595. https://doi.org/10.21437/Interspeech.2017-1391

Kim, S. (2021). Exploring How Older Adults Use a Smart Speaker–Based Voice Assistant in Their First Interactions: Qualitative Study. JMIR mHealth and uHealth, 9(1), e20427. https://doi.org/10.2196/20427

Klüber, K., & Roesler, E. (2025). The Influence of Embodiment on the Emotional Contagion of Affective Speech in HRI. 2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 1408–1412. https://doi.org/10.1109/HRI61500.2025.10973848

Ko, S., Barnes, J., Dong, J., Park, C. H., Howard, A., & Jeon, M. (2023). The Effects of Robot Voices and Appearances on Users' Emotion Recognition and Subjective Perception. International Journal of Humanoid Robotics, 20(01), 2350001. https://doi.org/10.1142/S0219843623500019

Lang, N. (2023, May 27). What is Data Augmentation? | Data Basecamp. https://databasecamp.de/en/data/data-augmentations

Lee, S. H., Kim, J. S., & Yu, S. (2025). The impact of care robots on older adults: A systematic review. Geriatric Nursing, 65, 103507. https://doi.org/10.1016/j.gerinurse.2025.103507

Lima, M. R., Srinivasan, N., Daniels, S., Vaitheswaran, S., & Vaidyanathan, R. (2025). Cultural Feasibility of Conversational Robots for Dementia Care in India: Participatory Design Study. Journal of Participatory Medicine, 17, e80457. https://doi.org/10.2196/80457

Liu, L., Zheng, T. F., & Wu, W. (2008). State-dependent phoneme-based model merging for dialectal Chinese speech recognition. Speech Communication, 50(7), 605–615. https://doi.org/10.1016/j.specom.2008.04.006

Lounnas, K., Lichouri, M., & Abbas, M. (2022). Analysis of the Effect of Audio Data Augmentation Techniques on Phone Digit Recognition For Algerian Arabic Dialect. 2022 International Conference on Advanced Aspects of Software Engineering (ICAASE), 1–5. 2022 International Conference on Advanced Aspects of Software Engineering (ICAASE). https://doi.org/10.1109/ICAASE56196.2022.9931574

Mahdi, H., Akgun, S. A., Saleh, S., & Dautenhahn, K. (2022). A survey on the design and evolution of social robots—Past, present and future. Robot. Auton. Syst., 156(C). https://doi.org/10.1016/j.robot.2022.104193

Marge, M., Espy-Wilson, C., Ward, N. G., Alwan, A., Artzi, Y., Bansal, M., Blankenship, G., Chai, J., Daumé, H., Dey, D., Harper, M., Howard, T., Kennington, C., Kruijff-Korbayová, I., Manocha, D., Matuszek, C., Mead, R., Mooney, R., Moore, R. K., … Yu, Z. (2022). Spoken language interaction with robots: Recommendations for future research. Computer Speech & Language, 71, 101255. https://doi.org/10.1016/j.csl.2021.101255

Mathers, C. D., Stevens, G. A., Boerma, T., White, R. A., & Tobias, M. I. (2015). Causes of international increases in older age life expectancy. The Lancet, 385(9967), 540–548. https://doi.org/10.1016/S0140-6736(14)60569-9

Mitzner, T. L., Blocker, K. A., & Rogers, W. A. (2024). "Alexa, how do I...?": Older adults learning to use digital home assistants. Gerontechnology, 23(s), 1–13. https://doi.org/10.4017/gt.2025.24.1.1199.08

Moore, P., & Pham, H. (2015). Personalization and rule strategies in data-intensive intelligent context-aware systems. The Knowledge Engineering Review, 30, 140–156. https://doi.org/10.1017/S0269888914000265

Najafian, M., Hsu, W.-N., Ali, A., & Glass, J. (2017). Automatic speech recognition of Arabic multi-genre broadcast media. 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), 353–359. https://doi.org/10.1109/ASRU.2017.8268957

Nakatumba-Nabende, J., Kagumire, S., Kantono, C., & Nabende, P. (2025). A systematic literature review on bias evaluation and mitigation in automatic speech recognition models for low-resource African languages. https://hdl.handle.net/10625/65270

Naseer, F., Nasir Khan, M., Tahir, M., Addas, A., & Kashif, H. (2025). Enhancing Elderly Care With Socially Assistive Robots: A Holistic Framework for Mobility, Interaction, and Well-Being. IEEE Access, 13, 82698–82717. https://doi.org/10.1109/ACCESS.2025.3567331

Nasretdinov, R., Korostik, R., & Jukić, A. (2025). Robust Speech Recognition with Schrödinger Bridge-Based Speech Enhancement. https://doi.org/10.1109/ICASSP49660.2025.10890638

Ni, J., Wang, L., Zhang, Y., Qian, K., Gao, H., Hasegawa-Johnson, M., & Yoo, C. D. (2025). Towards Unsupervised Speech Recognition Without Pronunciation Models (arXiv:2406.08380). arXiv. https://doi.org/10.48550/arXiv.2406.08380

OECD. (2020). Who Cares? Attracting and Retaining Care Workers for the Elderly | OECD. https://www.oecd.org/en/publications/who-cares-attracting-and-retaining-elderly-care-workers_92c0ef68-en.html

OECD. (2024). Elderly population | OECD. https://www.oecd.org/en/data/indicators/elderly-population.html

Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D., & Le, Q. V. (2019). SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. Interspeech 2019, 2613–2617. https://doi.org/10.21437/Interspeech.2019-2680

Pradhan, A., Lazar, A., & Findlater, L. (2020). Use of Intelligent Voice Assistants by Older Adults with Low Technology Use. ACM Trans. Comput.-Hum. Interact., 27(4), 31:1-31:27. https://doi.org/10.1145/3373759

Prinzellner, Y., Sturm, N., Geyer, C., Salomon, G., Weiss, A., Zauchner, S., Plößnig, M., & Jung, O. (2022). Using an off-the-shelf platform to develop a gender-sensitive health care robot for older adults and chronically ill people. Aging and Health Research, 2(2), 100072. https://doi.org/10.1016/j.ahr.2022.100072

Pu, L., Moyle, W., Jones, C., & Todorovic, M. (2019). The Effectiveness of Social Robots for Older Adults: A Systematic Review and Meta-Analysis of Randomized Controlled Studies. The Gerontologist, 59(1), e37–e51. https://doi.org/10.1093/geront/gny046

Ravi, V., Wang, J., Flint, J., & Alwan, A. (2024). Enhancing accuracy and privacy in speech-based depression detection through speaker disentanglement. Computer Speech & Language, 86, 101605. https://doi.org/10.1016/j.csl.2023.101605

Romero Moreno, F. (2024). Generative AI and deepfakes: A human rights approach to tackling harmful content. International Review of Law, Computers & Technology, 38(3), 297–326. https://doi.org/10.1080/13600869.2024.2324540

Schuppler, B., Hagmüller, M., & Zahrer, A. (2017). A corpus of read and conversational Austrian German. Speech Communication, 94, 62–74. https://doi.org/10.1016/j.specom.2017.09.003

Soleymanpour, M., Johnson, M. T., Soleymanpour, R., & Berry, J. (2023). Accurate synthesis of Dysarthric Speech for ASR data augmentation (arXiv:2308.08438). arXiv. https://doi.org/10.48550/arXiv.2308.08438

Tanioka, T., Osaka, K., Locsin, R., Yasuhara, Y., & Ito, H. (2017). Recommended Design and Direction of Development for Humanoid Nursing Robots Perspective from Nursing Researchers. Intelligent Control and Automation, 8(2), 96–110. https://doi.org/10.4236/ica.2017.82008

The World Bank. (2025). World Bank Open Data. World Bank Open Data. https://data.worldbank.org

WHO. (2025). Ageing and health. https://www.who.int/news-room/fact-sheets/detail/ageing-and-health

World Medical Association. (2013). World Medical Association Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Subjects. JAMA, 310(20), 2191–2194. https://doi.org/10.1001/jama.2013.281053

Xue, H., Peng, X., & Lu, Y. (2024). Low-latency Speech Enhancement via Speech Token Generation (arXiv:2310.08981). arXiv. https://doi.org/10.48550/arXiv.2310.08981

Yamamoto, K., Ishiki, A., & Nakagawa, S. (2021, October 12). Improvement of Elderly Speech Recognition Using Gammatone Filterbank Adaptation. 2021 IEEE 10th Global Conference on Consumer Electronics (GCCE). Global Conference on Consumer Electronics (GCCE), Kyoto, Japan. https://ieeexplore.ieee.org/document/9622086

Ye, Z., Hu, S., Li, J., Xie, X., Geng, M., Yu, J., Xu, J., Xue, B., Liu, S., Liu, X., & Meng, H. (2021). Development of the Cuhk Elderly Speech Recognition System for Neurocognitive Disorder Detection Using the Dementiabank Corpus. ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 6433–6437. https://doi.org/10.1109/ICASSP39728.2021.9413634

Yiwere, M. Y., Barcovschi, A., Jain, R., Cucu, H., & Corcoran, P. (2023). Augmentation Techniques for Adult-Speech to Generate Child-Like Speech Data Samples at Scale. IEEE Access, 11, 109066–109081. https://doi.org/10.1109/ACCESS.2023.3317360

Zafrani, O., Nimrod, G., Krakovski, M., Kumar, S., Bar-Haim, S., & Edan, Y. (2024). Assimilation of socially assistive robots by older adults: An interplay of uses, constraints and outcomes. Frontiers in Robotics and AI, 11. https://doi.org/10.3389/frobt.2024.1337380

Appendix I: Example database queries

Queries for older persons or older people and speech recognition from 2016 to 2025, inclusive, for the five databases used in the writing of this paper.

SCOPUS:
(ALL ("older persons") OR ALL ("older people") AND ALL ("speech recognition")) AND PUBYEAR > 2015 AND PUBYEAR < 2026

IEEE Xplore:
("All Metadata":"older persons") OR ("All Metadata":"older people") AND ("All Metadata":"speech recognition") Filters Applied: 2016 - 2025

ACM:
[All: "older persons"] AND [All: or] AND [All: "older people"] AND [All: and] AND [All: "speech recognition"] AND [E-Publication Date: (01/01/2016 TO 31/10/2025)]

PubMed:
(("older people"[All Fields] OR "older persons"[All Fields]) AND "speech recognition"[All Fields]) AND ((2016/1/1:2025/10/30[pdat]) AND (english[Filter]))

Google Scholar:
https://scholar.google.com/scholar?as_q=older&as_epq=%22speech+recognition%22&as_oq=persons+people&as_eq=&as_occt=any&as_sauthors=&as_publication=&as_ylo=2016&as_yhi=2025&hl=en&as_sdt=0%2C5&as_rr=1