© The Author(s) 2019. Published by Oxford University Press on behalf of The British Computer Society.
All rights reserved. For Permissions, please email: journals.permissions@oup.com
Advance Access publication on 11 September 2019 doi: 10.1093/iwc/iwz016

The State of Speech in HCI: Trends, Themes and Challenges

Leigh Clark[1,*], Philip Doyle[1], Diego Garaialde[1], Emer Gilmartin[2], Stephan Schlögl[3], Jens Edlund[4], Matthew Aylett[5], João Cabral[6], Cosmin Munteanu[7], Justin Edwards[1] and Benjamin R. Cowan[1]

[1]*School of Information and Communication Studies, University College Dublin, Belfield, Dublin 4, Ireland*

[2]*Speech Communication Laboratory, Trinity College Dublin, 7-9 South Leinster Street, Dublin 2, Ireland*

[3]*MCI Management Center, Universitätsstraße 15, 6020 Innsbruck, Innsbruck, Austria*

[4]*Department of Speech, Music and Hearing, KTH Royal Institute of Technology, Lindstedtsvägen 24, 114 28 Stockholm, Sweden*

[5]*CereProc Ltd, Codebase, Argyle House, 3 Lady Lawson Street, Edinburgh, EH3 9DR, United Kingdom*

[6]*School of Computer Science and Statistics, Trinity College Dublin, College Green, Dublin 2, Ireland*

[7]*Institute of Communication, Culture, Information and Technology, University of Toronto Mississauga, 3359 Mississauga Road, Mississauga, ON L5L 1C6, Canada*

Corresponding author: leigh.clark@ucd.ie

Speech interfaces are growing in popularity. Through a review of 99 research papers this work maps the trends, themes, findings and methods of empirical research on speech interfaces in the field of human–computer interaction (HCI). We find that studies are usability/theory-focused or explore wider system experiences, evaluating Wizard of Oz, prototypes or developed systems. Measuring task and interaction was common, as was using self-report questionnaires to measure concepts like usability and user attitudes. A thematic analysis of the research found that speech HCI work focuses on nine key topics: system speech production, design insight, modality comparison, experiences with interactive voice response systems, assistive technology and accessibility, user speech production, using speech technology for development, peoples' experiences with intelligent personal assistants and how user memory affects speech interface interaction. From these insights we identify gaps and challenges in speech research, notably taking into account technological advancements, the need to develop theories of speech interface interaction, grow critical mass in this domain, increase design work and expand research from single to multiple user interaction contexts so as to reflect current use contexts. We also highlight the need to improve measure reliability, validity and consistency, in the wild deployment and reduce barriers to building fully functional speech interfaces for research.

RESEARCH HIGHLIGHTS

- Most papers focused on usability/theory-based or wider system experience research with a focus on Wizard of Oz and developed systems
- Questionnaires on usability and user attitudes often used but few were reliable or validated
- Thematic analysis showed nine primary research topics
- Challenges identified in theoretical approaches and design guidelines, engaging with technological advances, multiple user and in the wild contexts, critical research mass and barriers to building speech interfaces

Keywords: speech interfaces; speech HCI; review; speech technology; voice user interfaces; intelligent personal assistants

Handling Editor: Javier Bargas-Avila

Received 19 April 2018; Revised 22 March 2019; Editorial Decision 30 May 2019; Accepted 30 May 2019

Downloaded from https://academic.oup.com/iwc/article/31/4/349/5560013 by Georgia Institute of Technology user on 23 September 2025

1. INTRODUCTION

Speech has become a more prominent way of interacting with automatic systems. In addition to long established telephony-based or interactive voice response systems (IVRSs), voice-enabled intelligent personal assistants (IPAs) like Amazon Alexa, Apple Siri, Google Assistant and Microsoft Cortana are widely available on a number of devices. Home-based devices such as Amazon Echo, Apple HomePod and Google Home are increasingly using speech as the primary form of interaction. The market for IPAs alone is projected to reach between $4.61 billion (Kamitis, 2016) and $9 billion (BusinessWire, 2018) by the early 2020s. The technical infrastructures underpinning speech interfaces have advanced rapidly in the recent years and is the subject of extensive research in the speech technology community (Chan *et al.*, 2016). Research on the user side of speech interfaces is said to be limited by comparison (Aylett *et al.*, 2014), and we know little about the current state of speech interface work in the human–computer interaction (HCI) field. Our work aims to map out this research, giving a comprehensive review of empirical research published in core HCI sources on speech interface interaction. We particularly concentrate on uncovering the methods and approaches used for studying speech interaction in HCI and the topics covered in this research.

Our review finds that research in this domain is usability/theory-focused or explores wider system experiences. Research uses a variety of Wizard of Oz (WoZ) systems, interactive prototypes or established commercial systems. Evaluation of these systems often focus on task performance and self-report questionnaires, measuring concepts like usability and user attitudes towards interaction, though the use of standard validated questionnaires is low. Most studies use bespoke measures that are not adequately tested for reliability or validity. The reviewed papers also highlighted a lack of generalizable design guidelines for speech interface interaction studies. The research conducted also tends to converge on nine topics, with the study of system speech production, design insight and modality comparisons accounting for the majority of papers. Based on the review, we propose a number of challenges for the HCI field to address in future speech-based research. Research challenges include taking into account recent technological advancements, improving theories of speech-based interaction and striving towards a research critical mass, increasing design guideline work and exploring multi-user contexts. Similarly, evaluation challenges are proposed to improve questionnaire validity and reliability, increasing in the wild evaluations and reducing barriers to building speech interface prototypes.

2. WHAT ARE SPEECH INTERFACES?

Speech interfaces are systems that use speech, either pre-recorded ('canned') or synthesized speech to communicate or interact with users (Weinschenk and Barker, 2000). Speech can be used as a primary input or output or in a two-way dialogue. Systems using speech (canned or synthesized) as their primary medium for communicating information to the user (Aylett *et al.*, 2015) are analogous to audio user interfaces (Weinschenk and Barker, 2000). In this type of interaction speech is monologic, produced only by the system as output. Speech can also be used solely as an input modality, often seen in 'command and control' of systems or devices (Cohen *et al.*, 2013). In this, user input is recognized by *automatic speech recognition* (ASR) and *natural language understanding* (NLU) components to identify user intents or commands. Perhaps the most common example of speech interfaces is spoken dialogue systems (SDSs), where the system and user can interact through spoken natural language. The dialogue involved can range from highly formulaic question–answer pairs where the system keeps the initiative and users respond, through HMIHY ('How may I help you?') systems where the user can formulate wider queries and the system computes the optimal next move, to systems that appear to allow user initiative by permitting users to interrupt ('*barge-in*') and change task. SDSs generally follow a common pipeline design to engage in dialogue with users. They first detect that a user is addressing the system. This can include identifying the user addressing it from a range of possible users (speaker diarization). The system's ASR then recognizes what is said and passes recognition results to the NLU component. The function of the NLU is to identify the intent behind the user's utterance and express it in machine-understandable form, often as a *dialogue act*. The system *dialogue manager* (DM) then selects an appropriate action to take based on the identified intent, considering factors such as the current state of the dialogue and recent dialogue history. A *natural language generation* component then generates a natural response, which is outputted by the system as artificial speech using *text-to-speech synthesis* (Jokinen and MacTear, 2010; Lison and Kennington, 2016). Some speech interfaces are used in conjunction with other input/output modalities to facilitate interaction (Weinschenk and Barker, 2000). For example, speech can be used as either input or output along with graphical user interfaces (GUIs), commonly seen in speech dictation for word-processing tasks or in screen-reading technology to support the navigation of websites. Indeed, SDS's such as IPAs (e.g. Siri) often display content and request user input through the screen as well as through speech (Cowan *et al.*, 2017).

2.1. Research aims

While interest in speech interfaces has been growing steadily (Cohen *et al.*, 2016; Munteanu *et al.*, 2017; Munteanu and Penn, 2014) there is no clear idea of what forms the core of speech-based work in the field of HCI. This makes it difficult to identify novel areas of research and the challenges faced in the HCI field, particularly for those new to the topic. As speech

Downloaded from https://academic.oup.com/iwc/article/31/4/349/5560013 by Georgia Institute of Technology user on 23 September 2025

TABLE 1. Review search terms.

Speech interface; voice user interface; voice system; speech-based; voice-based; speech-mediated; voice-mediated; human computer dialog *; human machine dialog*; natural language dialog* system; natural language interface; conversational interface; conversational agent; conversational system; conversational dialog* system; automated dialog* system; interactive voice response system; spoken dialog* system; spoken human machine interaction; human system dialog*; intelligent personal assistant

Asterisks (*) denote truncation to account for alternative spellings e.g. *dialog* or *dialogue*.

gains in popularity as an interface modality, it is important that the state of speech research published in the HCI community is clearly mapped, so that those who come to the research have a clear idea of the major trends, topics and methods. The current paper aims to achieve this by reviewing empirical work published across a range of leading HCI venues. Our aim is not to cover all work that covers both HCI and speech interface research, rather it is to cover research focused on speech interfaces within the field of HCI—specifically core HCI venues defined in Section 3.2. We hope this work may guide and inform future endeavours in the field, by identifying opportunities for further research. Below we report the method used to conduct the review and our findings and discuss challenges for future research efforts based on our results.

3. METHOD

3.1. Scope

We reviewed 99 publications on user interactions with speech as either a system output (e.g. Alm *et al.*, 1993), user input (e.g. Harada *et al.*, 2009) or in a dialogue context (e.g. Cowan *et al.*, 2015; Porcheron *et al.*, 2017). Papers were selected using adapted PRISMA guidelines, similar to the adapted QUORUM procedures in previous reviews (Bargas-Avila and Hornbæk, 2011; Mekler *et al.*, 2014).

3.2. Search procedure

Three databases were searched for relevant publications in January 2018: ACM Digital Library (ACM DL), ProQuest (PQ) and Scopus (SP). Each database was searched using terms generated from keywords in existing speech literature and from a survey of 11 leading researchers in speech HCI and speech technology (see Table 1). The terms were searched as exact phrases and, where possible, combined using Boolean operators (e.g. 'OR'). Otherwise, terms were searched for individually. Searches were limited to terms appearing in the title, abstract or publication keywords and were also limited to journal articles and conference papers, excluding other sources such as technical reports (see Section 3.3 for further details). No date range limitation was imposed on the searches. Similar to previous HCI-based reviews (e.g. Hornbæk, 2006), our search results were limited to core HCI publication venues. This list of publication venues was established primarily by combining top HCI publication sources listed on Google Scholar (Human Computer Interaction)[i] and Thomson Reuters (Computer Science, Cybernetics)[ii] rankings, with duplicate venues being removed. The SCImago Journal and Country Rank (SJCR) (Human-Computer Interaction)[iii] was consulted to ensure wide coverage, with additional from SJCR added if they were deemed relevant and appropriate by the authors. Conference venues were included from other sources but excluded from the SJCR search as the database ranks these by individual years rather than the conference as a whole. Overall, this provided a list of 48 unique HCI publication venues. This list is available in the supplementary material. Searches were then imported into the reference management software Mendeley.

3.3. Inclusion and exclusion criteria

The search resulted in a total of 1616 unique entries, following the removal of duplicates (ACM DL = 918, PQ = 173, SP = 525). Several inclusion and exclusion criteria were then applied to ensure relevance to the aims of the research. Inclusions were based on the following criteria: (1) *Only papers primarily investigating speech input (user utterances to system), speech output (from system to user) and/or dialogue (two-way interaction) were included*. (2) *Only full papers that were published in conferences and journals and written in English were included*. Because of the similarity in status with full papers, CHI notes were also included.

Exclusions were based on the following criteria: (1) *Papers investigating embodied interfaces were excluded*. Papers that discussed embodied interfaces like embodied conversational agents (e.g. Bickmore *et al.*, 2009) and robots (e.g. Strait *et al.*, 2015) were removed. This was to avoid studies where speech is confounded by issues of embodiment, such as gesture and emotive facial expressions (Breazeal, 2003), that can affect users' interactions and experiences (Bruce *et al.*, 2002; Kuno *et al.*, 2007). (2) *Papers without empirical measurement or evaluation of interaction with users were excluded*. We excluded technical discussions of systems, models or methods related to speech

1 goo.gl/Vmp5DM.
2 goo.gl/ngBucp.
3 goo.gl/PrnJta

Downloaded from https://academic.oup.com/iwc/article/31/4/349/5560013 by Georgia Institute of Technology user on 23 September 2025

interfaces that had little to no user evaluation (Han *et al.*, 2013). (3) *Non-full or non-peer-reviewed papers were excluded.* We excluded works in progress and extended abstracts (e.g. Aylett *et al.*, 2015; Cowan *et al.*, 2016), debates or panel discussions (e.g. Cohen *et al.*, 2016), workshop papers (e.g. Munteanu *et al.*, 2017) and magazine articles, for instance those from Communications of the ACM (e.g. Shneiderman, 2000) or ACM Interactions (e.g. Shneiderman and Maes, 1997).

The lead author and a co-author filtered the results independently based on these selection criteria. There was a strong level of agreement between the authors in the final set of papers chosen [Kappa = 0.856 ($P < 0.001$). After applying these criteria, 99 papers were selected for final analysis. Following Mekler *et al.* (2014) we extracted the research aims, interaction type being researched, methodologies and results from each paper. We also extracted publication sources, research topics, types of interfaces used, aspects of speech interfaces being assessed, type of research conducted, experiment procedures and measures and participant demographics. This information is available in the supplementary material.

4. RESULTS: PUBLICATION TRENDS AND RESEARCH METHODOLOGIES

4.1. Publication trends

Speech-based HCI work was published almost evenly between conferences and journals, with 50 papers published in conferences and 49 papers published in journals. While there were 17 unique publication venues overall (see Table 2), CHI was by far the most common venue for work reviewed here, accounting for 32.3% (N = 32) of all papers. The *International Journal of Human-Computer Studies* was the most common journal publication to feature, accounting for 18.2% of papers reviewed (N = 18).

TABLE 2. Number of papers reviewed by publication source.

Conference on Human Factors in Computing Systems (CHI)[1]	32
International Journal of Human-Computer Studies (IJHCS)[2]	18
International Conference on Multimodal Interaction (ICMI)	9
Behaviour & Information Technology	8
Computers in Human Behavior	5
Universal Access in the Information Society	5
ACM Transactions on Computer-Human Interaction (TOCHI)	4
Conference on Human-Computer Interaction with Mobile Devices and Services (Mobile HCI)	3
International Journal of Human-Computer Interaction	3
Interacting with Computers	3
Computer-Supported Cooperative Work and Social Computing (CSCW)	2
Intelligent User Interfaces (IUI)	2
ACM Transactions on Interactive Intelligent Systems (TIIS)	1
Designing Interactive Systems (DIS)	1
Human-Computer Interaction	1
IEEE International Conference on Systems, Man and Cybernetics	1
User Modeling and User-Adapted Interaction	1

[1]This includes one publication at the INTERCHI '93 conference - a joint conference between INTERACT and CHI.
[2]Includes papers published under the journal's previous title "International Journal of Man-Machine Studies".

4.2. Research methods

This section provides a summary of the methodological approaches used in the 99 papers reviewed, highlighting the systems used in interactions, communicative contexts and measures to research speech in HCI (see Table 3). Some papers contained more than one type of system in the categories presented (e.g. Medhi *et al.*, 2009) and as such the multiple systems are included in the totals.

4.2.1. Systems tested

One third of systems in the research (N = 33) were *mock systems*, in which users interacted with a WoZ system. In WoZ studies participants are led to believe they are interacting with a system, however, a (usually unseen) confederate controls the system's output (Dahlbäck *et al.*, 1993). These studies are generally designed to assess user responses to different types of output from hypothetically possible systems (e.g. Knutsen *et al.*, 2017; Oviatt *et al.*, 2008). Seventy-two of the papers reviewed either focused on *existing systems* or *working prototypes.* These included existing IPAs like Amazon Alexa and Siri (e.g. Cowan *et al.*, 2017; Porcheron *et al.*, 2017), prototype speech systems (e.g. Yankelovich *et al.*, 1995) and custom-built interfaces for delivering user-directed speech (e.g. Clark *et al.*, 2016). Two systems reported in the papers, but excluded from Table 3, were *hypothetical interactions and past experiences*. One of these papers had a more general focus on computers that use speech as a whole (Buchheit and Moher, 1990) rather than exploring a specific type of system, while the second used surveys to assess public attitudes towards IVRS and similar systems (Katz *et al.*, 1997).

4.2.2. Device contexts

We discovered four main types of device contexts that researchers used to deliver speech interfaces. A total of 49.5% (N = 49) researched *computer-based interactions*: existing

Downloaded from https://academic.oup.com/iwc/article/31/4/349/5560013 by Georgia Institute of Technology user on 23 September 2025

TABLE 3. Frequency of systems tested.

Device contexts	Mock systems			Existing systems & working prototypes		
Type of speech speech interface	*User speech input*	*System speech output*	*User-system dialogue*	*User speech input*	*System speech output*	*User-system dialogue*
Computer-based	5	1	8	22	9	4
Telephone-based	-	-	7	-	2	15
Mobile applications/IPAs	2	3	1	3	2	5
Vehicle-based	-	2	4	-	2	2
Other	-	-	-	3	1	2

or customized software running on desktops or laptops to accomplish a range of specific tasks. In these interactions, users sat with a computer and interacted with the speech interface for the duration of the experiment. Tasks included completing picture-naming tasks with multiple partner conditions (Cowan *et al.*, 2015) or listening to computer-media news (Kallinen and Ravaja, 2005). A total of 24.2% (N = 24) of systems explored *telephone-based dialogues* as an interaction, for example in healthcare appointment booking (Wolters *et al.*, 2009) and telephone banking systems (e.g. Wilkie *et al.*, 2005). Sixteen speech systems were *mobile applications* or *IPAs* on smartphones, including existing mobile IPAs (e.g. Porcheron *et al.*, 2017) and a mobile interface using speech input (Corbett and Weber, 2016). Ten of the systems were *vehicle-based* systems, where studies examined the user responses to system speech within driving simulation tasks (e.g. Truschin *et al.*, 2014). Finally, six were classified as *other types* of device contexts that did not readily fit into the preceding categories. Porayska-Pomsta and Mellish (2013) evaluated a computer output model based on human tutor feedback, while the second examined the use of a hands-free audio device with a speech interface in a healthcare setting (Sammon *et al.*, 2006). Piper and Hollan (2008) explored the use of tabletop displays to support medical conversations between deaf and hearing individuals, allowing a medical doctor to provide spoken input that was converted into text. Johnson and Coventry s(2001) included the evaluation of a prototype Automated Teller Machine (ATM) in exploring self-service user interfaces, while Berglund and Johansson (2004) investigated the use of speech and dialogue in controlling an interactive television. Finally, Vashistha *et al.* (2017) explored a crowd-powered, voice-based transcription system.

4.2.3. *Direction of communication*

We also categorized the direction of communication being explored in each paper, showing the breakdown of the papers across our three types of speech interface outlined in the introduction: *user speech input* (monologic from user to system), *system speech output* (monologic from system to user) and *user-system dialogue* (two-way speech interaction). Of the monologic interactions, 22 systems explored only system speech output. For example, asking users to evaluate voices (Dahlbäck *et al.*, 2007) or assess an auditory web-browser (Sato *et al.*, 2011). Thirty-five systems used only user speech input, including the use of an alternative interaction model for dictation tasks (Kumar *et al.*, 2012) or in assessing elderly users' interactions with an IVRS (Murata and Takahashi, 2002). The most commonly explored direction of communication (N = 48) was some form of dialogue between system and user, in which varying degrees of two-way spoken communication was explored. Examples included assessing language production when interacting with mocked up speech interfaces (e.g. Amalberti *et al.*, 1993; Cowan *et al.*, 2015), interactions with existing commercially available IPAs (e.g. Luger and Sellen, 2016; Porcheron *et al.*, 2017) and telephone-based IVRS (e.g. Wilke *et al.*, 2007; Wilkie *et al.*, 2005).

TABLE 4. Frequency of objective and subjective measurement of concepts.

Concepts measured	Objective	Subjective	Total
Task performance	54	11	65
User attitudes	-	53	53
Perceived usability	-	39	39
Lexis & syntax	27	2	29
System usage	19	4	23
User recall	6	1	7
Cognitive load	-	7	7
Physiological data	5	-	5
Other	8	9	17

4.2.4. *Measures*

In measuring user interactions with systems, the majority of papers (N = 63) used a combination of objective (e.g. user speech choices, task completion time) and subjective (e.g. self-report questionnaire) measures. Fewer papers used only subjective (N = 21), while 15 papers relied solely on objective measures.

Downloaded from https://academic.oup.com/iwc/article/31/4/349/5560013 by Georgia Institute of Technology user on 23 September 2025

A number of concepts were measured across the research reviewed (see Table 4). *Task performance* was most commonly measured. These included measures like the total number of conversational turns (Le Bigot *et al.*, 2007), percentages of tasks completed correctly (Oviatt *et al.*, 2008) and task completion time (Patel *et al.*, 2009). *User attitudes* were also common and included attitudes towards an interface's voice and/or speech content with research measuring concepts such as likeability and human likeness (e.g. Clark *et al.*, 2016). *Perceived usability* measures concentrated on quantifying concepts like perceived ease of use and learnability (e.g. Evans and Kortum, 2010), generally through Likert scale questionnaires. *Lexis and syntax* choices were also measured in the research reviewed (Cowan *et al.*, 2015; Piper and Hollan, 2008). *System usage* measures aimed to identify and quantify what people used speech interfaces for (e.g. Cowan *et al.*, 2017) and how they used them (e.g. Schaffer *et al.*, 2015). Memory-based research included measures of *user recall* in specific aspects of interaction (e.g. recall of system outputs; Knutsen *et al.*, 2017). *Cognitive load* on the user sought to measure aspects including physical, mental and temporal demands (e.g. Truschin *et al.*, 2014). *Physiological* data was assessed with measures including eye tracking (Hofmann *et al.*, 2014) or loudness of speech (Lunsford *et al.*, 2005). *Other* measures included system recognition of user input (e.g. Oviatt *et al.*, 2008), user agency (Limerick *et al.*, 2015), assessing user creativity and self-disclosure (Wang and Nass, 2005) and user compliance towards system suggestions (Takayama and Nass, 2008).

4.2.5. *Data collection methods*

Many papers included a combination of measures in conducting their research. *Questionnaires* were used in the majority of papers (N = 62), assessing aspects such as user attitudes (e.g. Suhm *et al.*, 2002) and usability (e.g. Perugini *et al.*, 2007). These tended to be administered post-interaction. The majority of these scales were *custom-built* (N = 58) and varied from single- to multiple-item questionnaires. Some of these custom-built scales were based on pre-existing measures. For instance, Le Bigot *et al.* (2006) developed a set of custom items based on the NASA-TLX scale for measuring cognitive load and Wolters *et al.* (2009) formed a scale based on the ITU-T Rec. P.815 evaluation for telephone-based dialogue systems (Möller *et al.*, 2009). The NASA-TLX scale for measuring cognitive load appeared eight times (e.g. Kumar *et al.*, 2012). Usability measures were occasionally assessed using existing scales, with three papers using the System Usability Scale (e.g. Evans and Kortum, 2010). Three other existing scales were used in the papers reviewed. The Subjective Assessment of Speech System Interface (SASSI) and the Driving Activity Load Index were both used once within the same paper (Hofmann *et al.*, 2014). One further paper used the Working Memory Span Test to assess participants' working memory capacity in operating speech-capable mobile phone services (Howell *et al.*, 2006). *Observations* of participants' interaction behaviour were equally common as the use of questionnaires (N = 62). Examples include lexical choices (e.g. Porcheron *et al.*, 2017) and modality use (e.g. Melichar and Cenek, 2006). *Interviews* were less common (N = 30) and mainly focused on participants reporting and reflecting on their interaction experiences (e.g. Luger and Sellen, 2016). *System measures* like system logs appeared in the same amount of papers as *interviews* (N = 30) and were useful for measuring what people used systems for (Sammon *et al.*, 2006). *Other* data collection methods, which did not fit easily into the other categories, appeared 10 times in the papers. These included corpus data (Derriks and Willems, 1998), focus groups (Cowan *et al.*, 2017), user modelling (Schaffer *et al.*, 2015), perceived sense of agency (Limerick *et al.*, 2015), physiological data (Kallinen and Ravaja, 2005) and cognitive walkthroughs (Johnson and Coventry, 2001). One paper was not explicitly clear in the methods for evaluating the iterative design of a tutor system, though it is assumed some form of interview was used (Hakulinen *et al.*, 2004). The frequency of each data collection method is included in Table 5.

TABLE 5. Frequency of data collection methods.

Data collection methods	Number of papers
Questionnaire	62
Observations	62
Interview	30
System measures	30
Other	10
Not explicit	1

5. RESULTS: RESEARCH TOPICS

As part of our review we also categorized the types of work reviewed (Section 5.1) as well as the primary research themes covered across the papers (Section 5.2).

5.1. Type of work

The papers reviewed generally divided into two types of work. The majority of papers (N = 59) were categorized as *usability/theory-based research*. These papers comprised of (a) those exploring how a particular design choice or behaviour impacted usability, systems or user performance and/or UX measures and (b) those exploring concepts and theories from research in human communication (e.g. linguistics, psychology) in the context of speech interface interaction. These two approaches often overlapped, and a number of papers in this category used theory or theory-based concepts to inspire the design of specific aspects of a speech interface, rather than researching complete systems. For example, Wilkie *et al.* (2005) used politeness theory (Brown and Levinson,

Downloaded from https://academic.oup.com/iwc/article/31/4/349/5560013 by Georgia Institute of Technology user on 23 September 2025

1987) to design greetings with specific politeness strategies in voice enabled phone banking systems and assessed usability and user attitudes towards the system based on these strategies. Similarly, other papers in this category have modified specific components, such as synthesized speech, while also exploring theories borrowed from human communication (e.g. alignment (Cowan *et al.*, 2015) and vague language (Clark *et al.*, 2016)).

The remaining papers (N = 40) were classified as *system experience research*. These papers explored user interactions with either working prototypes or existing systems, rather than specific system components and did not focus on theory-based concepts. This incorporated the vast majority of exploratory work that often used semi-structured interviews, ethnography and other qualitative analysis techniques to identify user views and issues. For example, recent work by Luger and Sellen (2016) and Cowan *et al.* (2017) explored users' experiences and past interactions with IPAs through interviews and focus groups respectively. A number of system focused research papers explored the deployment of bespoke and prototypical systems. Corbett and Weber (2016), for example, conducted usability studies of a system designed specifically for motor-impaired users. Similar approaches were adopted in exploring speech interface systems for users with other specific accessibility requirements (e.g. Harada *et al.*, 2009; Piper and Hollan, 2008), healthcare workers (Lai and Vergo, 1997; Sammon *et al.*, 2006) and identifying more general challenges towards designing an experimental system (Yankelovich *et al.*, 1995).

5.2. Research themes

Inductive Thematic Analysis (Braun and Clarke, 2006) was conducted on the final search results to categorize the research themes discussed in each publication. Themes were initially coded independently by two of the authors. After initial coding, perceived inconsistencies and variation in themes were resolved by discussion between the two authors and an independent observer, who had familiarity with the papers but had not contributed to the initial coding. Table 6 shows the breakdown of papers for each of the research topics and, if appropriate, their respective sub-topics. In some cases, the focus of papers overlapped topics and sub-topics, and so papers are categorized by their primary research topic, which was judged collaboratively to be the main focus of the paper. The topics and findings for the 99 papers are summarized in the sections below.

5.2.1. *System speech production*

Papers reviewed focused on the topic of *system speech production*—that is, system speech directed towards its users—and its effects on interaction behaviour and/or UX accounted for 22.2% (N = 22) of the review. This was the largest category observed, and was organized into a further four sub-topics, each focusing on a more specific area of system speech production.

TABLE 6. Number of papers in each research topic and sub-topic.

Research topic	Sub-topic		Totals
System speech production	Synthesis	8	22
	Content	8	
	Spatiotemporal aspects	3	
	General system speech production	3	
Design insight	Iterative design and bespoke systems	9	22
	Interface navigation	8	
	Dialogue modelling and design	3	
	ASR design	2	
Modality comparison	Keyboard and/or mouse	13	20
	Digital pen	3	
	Graphical	3	
	Gesture	1	
Experiences with IVRS	Dialogue and menu styles	6	8
	Usability and aging	1	
	Public attitudes	1	
Assistive Technology and Accessibility	Prototypes for physical impairments	2	7
	Conversation participation	2	
	Auditory and sonic modifications	2	
	Age and spatial ability	1	
User speech production	General user speech production	3	6
	Addressee identification	2	
	Alignment	1	
Speech Technologies for Development			6
IPA Experience			4
User Memory			3
Miscellaneous			1

The first of these sub-topics contains papers investigating the effects of specific manipulations to elements of *speech synthesis* on user perceptions and behaviour. Two papers examined the effects of synthesized speech on perceived personality. In the context of a book-buying website, Nass and Lee (2000) discovered that personality traits of extraversion and introversion were observed in synthesized speech, much as they are in human speech. Furthermore, they observed evidence of *similarity-attraction effects*, in which participants displayed positive preferences and attitudes to synthesized speech personalities that matched their own. In a similar paper,

Downloaded from https://academic.oup.com/iwc/article/31/4/349/5560013 by Georgia Institute of Technology user on 23 September 2025

Lee and Nass (2003) also observed that matching a user's personality with that of synthesized speech increased feelings of social presence, as did matching the personality of synthesized speech with textual content in contrast to when these were mismatched. However, the authors also observed stronger effects of social presence in extroverted personality conditions in contrast to introverted conditions. The similarity-attraction effect was also researched by Dahlbäck *et al.* (2007), in which choosing a voice that matches the accent of the user, rather than an accent related to the information being described, led people to view a vocal source as more informative and likeable, overriding any perceived expertise effects. Other research on medical IVRS interactions found no effects of either user gender or voice personality conditions on user behaviours such as self-disclosure (Evans and Kortum, 2010), which suggests that manipulating certain voice characteristics may be less impactful in some interaction scenarios than others.

Effects of voice-matching were also investigated by Truschin *et al.* (2014). In research using speech to read emails during driving, an interface that used synthesized voices that matched email sender characteristics led to better comprehension scores than when a single voice was used for all emails. However, this negatively impacted driving performance. Another paper assessed the impact of voice familiarity on user reactions and levels of interruption and comprehension (Bhatia and McCrickard, 2006). The authors found that while unfamiliar voices are the least interruptive, perhaps due to being filtered out more than familiar voices, quicker reactions were observed when participants heard their own voice. A further two papers examined different aspects of speech synthesis. Gong and Lai (2001) compared users' task performance and perceptions when interacting with only synthesized speech or a combination of synthesized and human speech in a virtual assistant interface. The results showed that while participants preferred the combined voice, believing they performed better on tasks in this condition, their actual performance was better when interacting in the synthesized speech only condition. Kallinen and Ravaja (2005) compared perceptions and physiological reactions towards hearing news read on a computer at a slow or fast speech rate, observing the fast rate as more arousing yet less understandable than the slow rate. Furthermore, the fast rate was judged more positively by younger participants, while older participants were more positive towards slow paced news delivery.

The second most common sub-topic in system speech production included papers that focused on the *content* of speech, as opposed to the vocal qualities in the speech synthesis research described above. These papers were interested in the message and language used by a system, including lexical choices and temporal features and how these impacted on users. Four of these papers comparatively explored different styles of dialogue presented to users and how it is controlled. Hu *et al.* (2007) compared two approaches to presenting information in voice-based browsing during simulated driving tasks. The first approach was 'summarize and refine (SR)' where the system groups and summarizes a number of options into small clusters based on shared attributes. The second was 'user-modelled summarize and refine (UMSR)', which combines SR with user-modelling (UM), in that the system presents options that best match a user's preferences while also clustering them according to common attributes. Both algorithms were based on participants' completion of a flight booking task, in which they were asked to assume the persona of a business traveller and book flights according to specific persona-based criteria. In the driving simulation task, the authors found USMR led to more efficient information retrieval than the SR approach. Hofmann *et al.* (2014) also explored information presentation in driving simulations. Their research compared two versions of an in-car speech dialogue system—one command-based and one conversational—on usability and driver distraction. In the command-based dialogue strategy, simple singular commands are issued by the user and the system subsequently guides the interaction step-by-step to elicit further information, with the user providing further simple responses. For example, the user may initially start the interaction with 'book a hotel', which the system will then use as a starting point to request related information from the user (e.g. location and duration of stay). In the conversational dialogue strategy, the initiative switches between the user and system during the interaction, and users are able to speak in whole sentences and request or provide multiple parameters during one interaction turn (e.g. location and duration of stay in the same phrase rather than separate as in the command-based strategy). The authors argue this allows for dialogue to be 'more natural, flexible and efficient' (Hofmann *et al.*, 2014, p.216). In driving simulation studies comparing these strategies, results suggested the command-based dialogue was accepted slightly more positively than conversational dialog, possibly a consequence of the system correctly interpreting users' utterances more frequently. The authors did not find any evidence for differences in driver distraction. In other driving simulation tasks, Vetek and Lemmelä (2011) compared the effects of guided and open-dialogue strategies on drivers' cognitive load when carrying out secondary tasks like making hotel, restaurant and flight bookings. These were similar to the different strategies employed by Hofmann *et al.* (2014). Guided dialogue strategies were similar to the command-based strategies—these were system-initiated and guided users step-by-step through dialogue, focusing on one aspect (e.g. location, date) of reservations, for example, at a time. Open-dialogue strategies allowed users to respond more freely in discussing multiple aspects at a time within one utterance (e.g. both location and date of reservations). In comparing the effects of these two dialogue strategies on users' cognitive load, the authors found that while open dialogues resulted in lower overall task load than guided dialogues for speech-only tasks, this difference was not observed in combined speech

Downloaded from https://academic.oup.com/iwc/article/31/4/349/5560013 by Georgia Institute of Technology user on 23 September 2025

and driving tasks. Both dialogue strategies were also observed to inhibit driving task performance. Walker *et al.* (1998) had a somewhat similar focus with a spoken language interface for email. In comparing dialogues that were either system controlled (system initiative dialogue) or flexibly controlled by the user or system (mixed initiative dialogue), the authors found that system-initiative dialogue was preferred, although mixed-initiative dialogue was more efficient in terms of time and number of turns to complete email tasks.

Three content-focused papers also explored expectations and perceived appropriateness of system language use. Porayska-Pomsta and Mellish (2013) evaluated a model of tutorial feedback in comparison to responses by a human tutor, finding there was no significant difference between a system's best preferred output and that of a human tutor. Buchheit and Moher (1990) assessed people's expectations of a hypothetical computer partner compared to a human partner in terms of their levels of assertiveness. In comparing these hypothetical scenarios, the authors found that people expect computer partners to be more assertive than human partners and are less likely to expect computers to use indirect language during interaction. For example, when offering an alternative choice of drink, indirect responses of 'you could try root beer' were expected more by human partners, whereas direct responses of 'you'll have to take root beer' were expected more by computer partners (Buchheit and Moher, 1990, p.113). Another paper exploring studies on system politeness found that the use of different politeness strategies led people to have less positive attitudes towards an automated voice banking service (Wilkie *et al.*, 2005). Finally, one content-focused paper assessed evaluations of a SDS designed to tailor its use of acknowledgements in accordance with the perceived 'ephemeral emotions' of its user, such as pleasure, confidence, confusion and dependency (Tsukahara and Ward, 2001). Results indicated that participants preferred the system that had this feature to the system than without it.

Three papers focused on *spatiotemporal aspects* of a system's speech: *when*, *how* and *from where* information should be delivered. Iqbal *et al.* (2011) looked at proactive road condition alerts for drivers engaged in phone conversations. They showed that interventions lowered collisions and turning errors compared to trials without interventions. Moreover, more descriptive messages that were explicit about upcoming conditions were perceived more positively than shorter, more general messages. Kousidis *et al.* (2014) also investigated dialogue systems in driving simulation tasks, when testing a situationally aware system that refrained from interruptions when higher driver attention was required. Their results showed improvements in driving performance and attention to the content of speech, compared to when situational awareness was not available, but noted similar levels of performance and attention when drivers could control the system. Takayama and Nass (2008) also explored driving simulation tasks in comparing the proximity of two sources of information—in-car and wireless computer systems. Participants were intermittently reminded of which type of system they were interacting with. The authors note that system proximity can have effects on both user behaviour and attitudes. Results showed that participants felt more engaged with the in-car system, leading to safer driving behaviour, yet they also drove faster than in interactions with the wireless system. In-car systems also resulted in greater feelings of contentment and self-disclosure towards systems.

Finally, three papers in the review sample were categorized as investigating *general system speech production*, as the papers demonstrated a focus on both synthesis and content produced by a system. Clark *et al.* (2016) examined people's responses to an interface using vague language (e.g. phrases including *kind of*, *more or less* and *a little bit*), while also exploring the effects across different synthesized and human voices. The participants perceived vague language as more appropriate when used by a human recorded voice compared to any of two synthesized voices, with the authors suggesting synthesized speech may be less capable in executing the social functions of vague language. Jeon *et al.* (2015) explored the effectiveness of enhanced menu cues in dual-task experiments, one of which included participants conducting a driving simulation task. In doing so, the authors looked at two different types of cues. The first—*spearcons*—are compressed and sped up spoken phrases that are no longer comprehensible as individual words or phrases but retain a unique audio fingerprint. The second—*spindex*—create auditory associations between the first letters or phonemes of menu items and specific phrases. The authors provide an example of the spindex cue for 'all of the above' being based on the spoken sound of 'A' (p.3). Findings showed that the use of enhanced auditory cues may improve drivers' abilities to operate in-vehicle menus more efficiently while also driving more safely, though the authors acknowledge that other aspects such as task difficulty and interaction style may need to be take into account. Finally, Wolff and Brechmann (2015) investigated different feedback styles produced by systems in a computer-assisted learning environment, focusing on modes of generating feedback (pre-recorded and synthesized speech) and prosody (neutral and motivational speech). Participants were tasked with listening to a series of experimental tones varying in properties of duration and frequency modulation, with each participant given a target group of one of four possible combinations of tone properties. They were then tasked with categorizing a series of experimental tones as originating from the 'target' or 'non-target' groups, with different feedback about the correctness presented in permutations of feedback generation mode and prosody (neutral and pre-recorded, motivational and pre-recorded, neutral and synthesized). Findings showed a learning advantage for participants receiving pre-recorded neutral over synthesized neutral feedback, as well as benefits for participants receiving motivational versus neutral pre-recorded feedback. The authors highlight these properties of speech as being valuable for computer-assisted learning environments.

Downloaded from https://academic.oup.com/iwc/article/31/4/349/5560013 by Georgia Institute of Technology user on 23 September 2025

5.2.2. *Design insight*

Papers aimed at generating *design insight* accounted for 22.2% (N = 22) of the review. These included works aimed at developing generally applicable guidelines for speech interface interactions and iterations of developments for specific systems.

Nine of these papers explored *iterative design and bespoke systems*. These papers analysed multiple versions of a system in developing it based on various types of testing and feedback, or looked at one version of a specific, bespoke system tailored to accomplish a specific task. Hakulinen *et al.* (2004) explored the design and iterative development of an integrated tutor for teaching new users to operate a speech-based email reading application, while monitoring their actions to they learn how to use it. The authors presented three iterations of the tutor and presented several findings for possible future design for similar systems, including being explicit in telling users what to do, teaching one thing at a time and giving users the opportunities to control the pacing of the system's tutoring. Yankelovich *et al.* (1995) focused on reporting the design of an experimental speech interface, *SpeechActs*, which aimed to access existing GUI-based applications (e.g. email and calendar tasks). Interface testing led to a number of design guidelines including the need to develop effective conversational structures, lexicons, information flow and organization. One paper developed a system designed for healthcare workers to interact with a hands-free mobile voice agent for functions including phone calls, paging, voice messaging and queries on staff locations (Sammon *et al.*, 2006). Wang *et al.* (2008) presented a pen and speech-based storytelling system for Chinese children using a scenario-based and user-centered approach. Findings indicated generally good usability, with the authors offering considerations of adaptive user interfaces for future iterations of similar systems. Harada *et al.* (2008) presented the *VoiceLabel* system, whereby users can collect and label activity data (e.g. walking, going upstairs) using speech. The system then generates labels that can then be reviewed and edited using offline desktop software. The authors found the system was moderately effective at automatically labelling pre-tagged activities and noted participants found the system usable though not always reliable. The paper highlights possible avenues for future work, including different modalities such as graphical interfaces.

Moran *et al.* (2013) investigated team responses to an agent's instructions in a large-scale outdoor pervasive game, whereby a rule-based agent provided players with game instructions using their mobile phones as the interface. The authors provided potential future design recommendations on concepts including trust, compliance and reliability. The authors presented the results of its deployment in a hospital-based user study, noting that while there was some value in the voice agent, participants noted usability problems on not understanding how to use some of its features and noting that a speech-driven interface may not be sufficient for their needs. Johnson and Coventry (2001) explored the concept of speech interactions with a proposed ATM system. Cognitive walkthroughs, heuristic evaluations, focus groups and user trials were conducted, with the user trial of the 'ATM' conducted through a personal digital assistant device. The system struggled with recognizing user input and, though the authors noted participants were generally positive towards future speech technology, they identified issues around privacy and the need for multiple evaluations of novel systems. One paper (Vashistha *et al.*, 2017) introduced *Respeak*—a voice-based, crowd-powered system using crowdsourcing and ASR to transcribe audio files. Findings showed that speech input tasks were quicker and contained fewer errors than typing, with the authors affirming its ability to improve transcription quality and enhance low-income populations' earning potential. Finally, Sivaraman *et al.* (2016) introduced *SimpleSpeech*—a prototype system for recording and editing asynchronous voice messages, providing automatically-generated transcriptions of voice messages which can be edited through word-processing. Mixed-methods evaluations showed users were favourable to text-style editing of speech input, with a decrease in perceived cognitive load for student users. The formality of this text-edited speech also emerged as somewhere between typical spoken and text media.

A further eight papers categorized under design insight focused on *interface navigation*. These papers looked at speech as an alternative modality for navigating through system interfaces, or how speech may be used as a navigation modality in novel and prototype systems. Sears *et al.* (2002) presented two experiments on this topic. The first compares speech for using standard and predictive cursor controls. Few differences were observed between the two controls, though predictive cursors were associated with higher errors when used for diagonal movements. Larger target sizes were shown to reduce errors and reduce target selection times along with shorter distances. The second experiment examined delays that were associated with speech input, highlighting drawbacks in the time taken to produce spoken commands in controlling the cursor. Work by Dai *et al.* (2005) also explored cursor control through two grid-based approaches. One approach provided users with a single cursor and the second provided users with nine cursors. Results were mixed in target selection tasks, with the nine-cursor approach resulting in faster selection time, while the single cursor led to reduced error rates. Sears *et al.* (2003) focused on examining speech navigation for dictation, comparing target-oriented and direction-oriented navigation approaches. Results of dictation and email composition tasks showed similar failure rates between these two approaches. Users experience increased difficulty when issuing direction-based commands and were more likely to switch to target-based navigation than the reverse.

Feng and Sears (2004) took a different approach in understanding the use of confidence scores in the error correction process for navigation systems in document proofreading. These scores are numeric values given by ASR engines that represent the level of certainty or confidence that the recognition engine has of a particular word being correct. However, the

Downloaded from https://academic.oup.com/iwc/article/31/4/349/5560013 by Georgia Institute of Technology user on 23 September 2025

authors note that some ASR errors will have high confidence scores, and some correctly recognized words will have low confidence scores. In attempting to facilitate the detection of errors, the authors determined that both raw confidence scores and differences in confidence scores for words are unlikely to be effective. Instead, the paper shifts focus to users' navigation of interfaces, with the assumption that users are responsible for detecting words that are to be corrected and how their efforts in navigating to these words can be facilitated. This work develops a *navigation anchor* approach. Instead of existing target- and direction-based navigation approaches, users can navigate from words within documents using distinct commands (next and previous) as well as some directional commands (up, down, left and right). Navigation anchors—the specific phrases being selected as starting points—were selected based on the authors' evaluation of differences in confidence scores to present appropriate recognition errors to users. In evaluating this navigation anchor approach for proofreading documents, participants were able to specify erroneous words efficiently and encountered low failure rates. These results outperformed comparisons to existing techniques (target- and direction-based), though the authors note that using differences in confidence scores may not necessarily be the best possible solution in designing navigation anchors. Feng *et al.* (2011) explored further comparisons in email composition tasks, finding that the target- or direction-based and navigation anchor approaches were similar in terms of efficiency, though anchor-based approaches allowed for greater usability and quality of text production. In another paper, Feng *et al.* (2006) conducted longitudinal evaluations of users conducting speech-based navigation activities. Findings showed that efficiency in navigating improved with experience and that strategy selections changed over time, with a shift towards more cognitively demanding but efficient strategies. The authors highlight insights for new technology users, noting that users need time to develop efficient strategies, as well as being provided multiple options for single solutions, thus allowing users to discover which strategies are appropriate for specific tasks.

Löhr and Brügge (2008) discuss *conversation-and-control*—an alternative approach to navigating graphical interfaces via manipulation of widget functions through speech commands. This approach is based on modelling peoples' interactions with a graphical interface and used specific single commands for each widget function, reducing the number of commands needed per interaction. This approach was compared with a traditional command-and-control approach, which offers users the ability to activate or navigate away from widgets that a system specifically focuses on (e.g. invoking commands such as *push* or *cancel*). In a study, the conversational approach resulted in shorter task completion times. The authors suggest there may be higher cognitive load for the conversation-and-control approach, suggesting it may be directed more towards expert users, with command-based approaches better suited to novice users.

Finally, speech navigation has also been examined in a prototype interactive television system (Berglund and Johansson, 2004). Participants were given seven task scenarios similar to real life situations (e.g. finding specific films or television programmes) using speech. Results showed generally positive attitudes towards the system and its efficiency, though the authors noted the need to keep the remote control device even if speech remains a modality option. The authors noted that participants were liable to overestimate the system's capabilities due to speech being available, highlighting the need to inform users with clear definitions of the system's actual capabilities.

Three papers in this category explored *dialogue modelling and design.* Derriks and Willems (1998) adopted a corpus analysis approach to identify and classify negative feedback phenomena in human-machine information dialogues, with a view to constructing more acceptable future dialogue models in detecting and resolving difficulties at any time during the dialogue. Hone and Baber (2001) also explored system dialogue design in terms of *habitability*—the congruence between the language people use towards a system and the language a system can understand and accept. Results in interactions based on ATMs and telephone banking systems indicate that habitability cannot necessarily be achieved by visually displaying an input language. Instead, goals of habitability may be improved by providing menus of words indicating a system's capabilities or providing spoken system, prompts to users about specific tasks. The authors suggest that spoken menus may prove more habitable than visual menus for speech input. Work by Jokinen (2006) explores user modelling in a speech-based email application *AthosMail* to facilitate more natural interaction through the dialogue control and amount of information presented to users. The paper describes an adaptive model based on the perceived level of expertise of its users, differentiating between three levels of expertise based on user behaviour. The author notes that the evaluation setup was not long enough for reliable assessment of the system's ability to adapt to its user and notes the difficulties adapting systems appropriately in this context. However, the paper also argues that user modelling may prove valuable in allowing users to complete specific tasks efficiently.

Finally, two papers in this category discussed aspects of *ASR design* and improving system speech recognition. Murray *et al.* (1996) examined data entry tasks with auditory-only systems and auditory systems with visual prompts in understanding requirements for ASR design. In their first experiment, the authors analysed the use of closed word-sets or syntax in enhancing recognition accuracy for auditory only interfaces. They found using partial syntax accommodated users' error-correction attempts better than full syntax or no syntactic constraints. In their second experiment using an auditory interface with visual prompts, the authors investigated two styles of prompts that provided feedback for the users' data entry. Results showed that displaying the full set of available options (option prompts) led to more efficient performances

Downloaded from https://academic.oup.com/iwc/article/31/4/349/5560013 by Georgia Institute of Technology user on 23 September 2025

than displaying current fieldnames with expansions to include further relevant options (fieldname prompts). The authors conclude with several guidelines for the design of both styles of interface investigated, including that overly-constrained syntaxes can lead to more user errors despite higher recognition rates and that user deletions should be accompanied by auditory confirmations to provide necessary feedback to users. The second paper in this category explored the effects of motion on ASR for mobile devices (Price *et al.*, 2006). Participants performed dictation tasks while seated and while walking on a treadmill. Recognition error rates were shown to increase when walking but may be reduced by having users complete enrolment—having users read a set of predefined text for a system to recognize before engaging in the task at hand. When this enrolment is conducted while seated, recognition error rates increased when participants performed dictation tasks while walking. However, when enrolment was conducted while participants were walking, recognition error rates decreased when seated. The authors note that ASR error rates may be reduced by having users complete enrolment periods under more demanding and stressful conditions than those that will actually be experienced during the system's usage.

5.2.3. *Modality comparison*

Twenty papers in the review compared the use of different modalities, exploring concepts including user performance and UX. Example comparisons include comparisons of single modalities (e.g. Begany *et al.*, 2015; Murata and Takahashi, 2002; Oviatt *et al.*, 2008), as well as unimodal and multimodal combinations (e.g. speech and gesture; Hauptmann and McAvinney, 1993).

The results of these modality comparisons are mixed. Thirteen papers explored speech, keyboard/text and mouse input. Eight papers explore just speech and keyboard/text input. Three of these report a negative impact of using speech, with speech reducing ease of system use (Begany *et al.*, 2015) and a user's sense of agency (Limerick *et al.*, 2015) compared to more traditional input modalities such as keyboard and mouse. Molnar and Kletke (1996) compared the use of a spreadsheet software package with either keyboard and mouse input alone, or with keyboard and mouse input augmented by speech input. Results showed that the inclusion of speech led to slower task completion and less favourable user attitudes compared to when using only a keyboard and mouse. Conversely, two papers report improvements when using speech as an input modality. Greater benefits were observed among elderly compared to younger users, particularly for those not accustomed to keyboard input (Murata and Takahashi, 2002). Furthermore, with a voice-driven video learning interface, voice was seen as more useful and was used more than typing (Culbertson *et al.*, 2017). Two further papers compared speech and text within a dialogue interaction. Le Bigot *et al.* (2006) found performance with an information retrieval system improved over time, regardless of using speech or text. However, similar experiments comparing speech and written modalities observed lower efficiency when speech was used (Le Bigot *et al.*, 2007). Both papers also reported effects on language use, with higher uses of pronouns in user speech input. Another paper compared speech and text as system outputs when users answered open-ended questions to determine their levels of creativity and self-disclosure, with higher levels of creativity observed in users' answers when exposed to text output (Wang and Nass, 2005).

Five papers explored the use of a mouse alongside speech, either with or without a keyboard, and observed mixed preferences in modality choice. In comparing unimodal and multimodal systems, participants were shown to prefer using a mouse for navigating within a graphical interface but using natural language for entering data (Melichar and Cenek, 2006). A different multimodal system for radiologists showed users preferring speech unless their hands were already on the keyboard or mouse. However, the results identified instances of users' low tolerance for speech recognition errors, as well as the difficulty of remembering some spoken commands (Lai and Vergo, 1997). Bekker *et al.* (1995) examined the use of only speech and mouse input with a document-annotation system, observing more errors with speech and a user preference for mouse input. Price and Sears (2005) observed participants completing data entry for handheld devices using server-side speech recognition. Tasks including setting meetings and dictating details were conducted using either multi-tap (representing a telephone keypad) or soft-keyboard (representing touch screen) interfaces alongside speech dictation. While no significant differences in data entry speed were observed, the authors noted the soft-keyboard interface appeared to be more intuitive. A follow-up study investigating the impact of practice on the soft-keyboard system did not show improved performance, with recognition errors continuing to impact data entry. Another paper compared a standard email interface with a speech-based email system (Liapis, 2011). The speech system was observed to be 220% quicker than the traditional system and showed significantly less cognitive load on users.

Three papers in this category also explored the use of *digital pen* input alongside other modalities. Two of these papers observe the unimodal and multimodal use of pen and speech input. Suhm *et al.* (2001) examined error correction for speech interfaces, finding a preference for multimodal correction. Specifically, skilled typists preferred keyboard and mouse input, but the authors predict higher performance could be generated for all users with speech and pen input instead. When speech and pen inputs were available during a map-based task for coordinate emergency resources during a major flood, a shift from unimodal to multimodal communication was observed when cognitive load increased (Oviatt *et al.*, 2004). A further paper showed cognitive load was best managed with speech input compared to pen input in the context of using a simulated tutoring system (Oviatt *et al.*, 2008).

Downloaded from https://academic.oup.com/iwc/article/31/4/349/5560013 by Georgia Institute of Technology user on 23 September 2025

Another four papers explored speech in combination with *graphics* or *gesture*. Three assessed the use of *graphical* modalities. When using a prototype application for booking restaurant reservations in a mobile multimodal interaction, decreasing graphical input efficiency resulted in higher speech usage (Schaffer *et al.*, 2015). In comparison with only graphical output, the higher the level of spoken feedback present with a timetable information system the more that participants rated the system as human-like (Qvarfordt *et al.*, 2003). A third paper evaluated route-following using a wrist-worn mobile device specifically designed for people with cognitive impairments (Fickas *et al.*, 2008). Graphical, text-based and audio modalities were compared, with audio instructions scoring highest in both navigation scores of confidence and accuracy as well as user preferences. Finally, in comparing speech with *gesture*, as well as with a combination of the two, no significant differences were found when using them to manipulate graphical objects (Hauptmann and McAvinney, 1993), although user preferences were observed for speech and gesture combined.

5.2.4. *Experiences with IVRS*

Eight papers assessed *experiences with IVRS*, exploring design choices including dialogue strategies, vocal characteristics and methods of directing calls. In this category, six papers were classified as explore *dialogue and menu styles*. Perugini *et al.* (2007) examined the use of *out-of-turn* interaction which allows users to make unsolicited utterances when unsure how to best respond to system prompts. This means that a user is able to specify they want to engage in specific sub-menus (e.g. regarding password settings) even if that option is not presented to them by the system in higher level menu settings. Findings showed when this was compared to a system without the feature, out-of-turn interactions allowed for quicker task completion and improvements in usability and user attitudes. Wilke *et al.* (2007) investigated somewhat of a similar approach in looking at the effect of providing a hidden menu option rather than explicit menu option in telephone-based banking tasks. Specifically, an overdraft service could be accessed by saying the keyword 'overdraft', which was introduced to participants verbally by the system during the call-flow but was not explicitly listed as one of the menu options. The authors found this strategy resulted in 37% of users being unable to complete the overdraft request, noting that non-explicit menu options may prevent some users from successfully completing elements of telephone-based banking tasks. Litman and Pan (2002) evaluated an online train schedule retrieval system accessed via telephone, revealing performance improvements when the system predicts ASR problems and adapts to more conservative dialogue strategies. In another paper, a field study compared call routing in a cell centre environment (Suhm *et al.*, 2002). The results showed improvements in usability for natural language call routing compared to traditional touch-tone menus, as well as the potential for reducing call centre costs. Howell *et al.* (2005) examined the effects of underlying spatial metaphors (metaphors that structure people and information in relation to one another and within space) and private or public contexts on using a mobile city guide service. The underlying metaphor was based on an office filing system. Results showed that participants who used the office filing system metaphor performed tasks faster and interrupted prompts more frequently than non-metaphor participants, though no significant differences were observed in user attitudes or public and private settings. A follow up study introduced the metaphor of a computer desktop to the two existing conditions (Howell *et al.*, 2006). Results showed no overall differences for performance or attitudes in the three conditions. However, participants who were able to visualize services performed significantly better than participants who did or could not.

Of the remaining two papers, one focused on *usability and ageing* in experiences with IVRS (Dulude, 2002). Usability ratings were found to be lower for older users than younger users, with confusing choices, speech rate and length of options accounting for the majority of problems for older users. Finally, Katz *et al.* (1997) conducted a national survey on *public attitudes* in the United States towards IVRS. The authors found that the quality of recent IVRS experiences was the most significant predictor of indicating a liking for the technology, with older respondents having more negative attitudes, particularly in relation to perceptions of convenience and frustration.

5.2.5. *Assistive technology and accessibility*

Speech interface usability and user experience was explored in different specific communities in the reviewed papers—i.e. users with specific needs or requirements, or users interacting with a specific type of system. Seven of the papers reviewed were categorized as researching *assistive technology and accessibility*. Assistive technology is generally defined as 'Adapted or specially designed equipment, products and technologies that assist people in daily living' (World Health Organization, 2001, p.174). Accessibility often includes the use of assistive technology in designing environments for improving usability with specific people (LaPlante, 1992).

Two papers explored designs of *prototypes for physical impairments*. One of the papers assessed a speech interface with a smartphone (Corbett and Weber, 2016), finding that existing approaches in voice interaction design do not necessarily translate into mobile interaction spaces. The authors highlight the need to explore mobile-specific theories, rather than always borrowing those from desktop-based interaction. Another paper compared users' interactions with a voice-based mouse emulator *Vocal Joystick*, with motor impaired and non-impaired individuals (Harada *et al.*, 2009). Users were observed to learn the mapping of vowel sounds and directions, and marked improvements were seen by the end of a longitudinal study.

A further two papers examined using technology to allow for *conversation participation*. Alm *et al.* (1993) assessed the

Downloaded from https://academic.oup.com/iwc/article/31/4/349/5560013 by Georgia Institute of Technology user on 23 September 2025

usability of an assistive dialogue prototype for users who have severe physical disabilities and do not speak. In using a system that helps these users take part in conversation, one subject was observed to increase their number of words in conversation, compared to using a more traditional word board and a speech device that stored specific phrases and names. Piper and Hollan (2008) adapt a participatory design process in creating an interactive tabletop display with audio input to assist conversations between physicians, deaf patients and interpreters. They discussed the potential for using tabletop displays for enhancing privacy and independence for deaf patients, for example in interactions with doctors or lawyers regarding personal matters.

Auditory and sonic modifications were investigated in two papers. Sato *et al.* (2011) conducted studies with blind participants, analysing the benefits of a voice-based web browser plugin, which uses a secondary voice to provide contextual information alongside the primary voice. When comparing this to traditional screen reader software, results showed the plugin system brought the users increased confidence, though did not improve task performance. Mascetti *et al.* (2016) explored two auditory approaches of *sonification* techniques in guiding people with visual impairments or blindness across roads—mono sonification (single channel suitable for playback through a device's speaker) and stereo sonification (dual channel through headphones indicating left and right placement of audio). These sonification techniques were compared with a speech-based guiding mode. Results showed that 75% of participants preferred one of the sonification techniques to the speech-based guiding mode, though authors discuss that higher user effort is required to decode the sonified information.

Finally, Pak *et al.* (2008) also examined auditory computer interfaces for *age and spatial ability* in auditory processing tasks performance, using IVRS to complete 24 information obtaining interactions in banking or electric utility contexts. The results showed evidence of age-related differences in spatial ability contributing to age-related differences in the IVRS task performance. Consequently, the authors suggest future system designs may need to reduce spatial ability demands on older users.

5.2.6. *User speech production*

Six papers focused on *user speech production*. Three of these papers were categorized as investigating *general user speech production*. Amalberti *et al.* (1993) showed that people adapt their language choices according to their partner models but noted that differences between human and computer speech choices decreased as people got more familiar with the interaction. People also tended to use fewer fillers (e.g. 'um', 'err'), request confirmation and repetition more and use fewer topic shifts in computer compared to human interaction. Kumar *et al.* (2012) compared existing dictation with 'Voice Typing'—a speech interaction model that transcribes users' utterances as they are produced, allowing for error identification in real-time. In using this, their study showed a reduction in error rate and certain cognitive demands compared to dictation. Another paper explored the impact of spoken translation software on cross-lingual dialogues (Hara and Iqbal, 2015). During experiments, participants were observed adapting their speech and comprehension due to imperfections in system-produced translations, and the authors accordingly formulated a set of design guidelines for such systems.

Two papers examined *addressee identification*. In aiming to establish better models for differentiating between computer and human-directed utterances by users, Lunsford *et al.* (2005) observed that participants reduced the loudness of their voice to indicate self-talk, compared to higher amplitudes used for system-directed speech. Lunsford and Oviatt (2006) investigated the accuracy of people's judgements of whether a speaker's intended interlocutor is a human or computer. Participants were more accurate in identifying computer-intended interlocutors than human, particularly when presented with visual information alone, and both gaze and tone of voice were important factors in making judgements.

Finally, one paper focused on the relationship between *alignment*—a convergence of dialogue partners' speech choices in conversation—and *partner models*—users' perceived working models of a system's communicative ability (Cowan *et al.*, 2015). The authors observed that, although people's partner models were affected by the humanness of the speech synthesis used, this did not significantly impact levels of syntactic alignment (i.e. structural arrangement of phrases) between human or computer partners.

5.2.7. *Speech technologies for development*

Six papers explored the domain of *speech technologies for development*. There were similarities in these papers with those focusing on modality comparisons, however, these four papers had a distinct focus on speech interface use to support rural communities, novice users, or people with low levels of literacy. Medhi *et al.* (2009) compared mobile speech-based, text-based and rich multimedia money-transfer interfaces with non-literate and semi-literature participants. The authors highlighted that while rich multimedia interfaces containing audio-visual output provided better task completion, speech-based interfaces afforded greater speed and required less assistance when using them. Similarly, Medhi *et al.* (2011) compare a text-based interface with a live operator for patients discussing symptoms with health workers over the phone, observing that the live operator was up to 10 times more accurate than the text-based interface. Patel *et al.* (2009) compared dialled and speech input interfaces for farmers accessing agricultural information over the phone, discovering that dialled input outperformed speech in both task completion rates and perceived user difficulty. Raza *et al.* (2013) had a somewhat different approach and explored the challenges of virally spreading awareness of, and training people in, speech-based services for users in communities with low levels of literacy. The authors noted that, while the system would spread to new users quickly, a rapid

Downloaded from https://academic.oup.com/iwc/article/31/4/349/5560013 by Georgia Institute of Technology user on 23 September 2025

declining interest saw the majority of people using the system for only a few days.

DeRenzi *et al.* (2017) conducted a 12-month longitudinal study of a prototype mobile voice- and web-based application to provide community health workers, mostly with low levels of literacy, with feedback regarding their work. Findings showed that participants used the voice-based system more frequently in contrast to the web-based system and that this was due to its easier perceived usability. Cuendet *et al.* (2013) presented *VideoKheti*—a mobile system for rural Indian farmers with low levels of literacy. The system allows these farmers to watch agricultural videos in their own language and dialect. While participants were enthusiastic about using the system, the addition of speech to a tactile and graphical interface did not show significant improvements. Participants still encountered difficulties related to their level of literacy—those with lower levels encountered more speech-based errors and were shown to use the tactile interface more.

5.2.8. *IPA experience*

Four papers explored people's experiences with current IPAs. Leahu *et al.* (2013) discuss the flexible, rather than static, nature of 'human' and 'machine' identity categories in this context and possible experimental design implications. Luger and Sellen (2016) describe power users' experiences with IPAs, highlighting a dissonance between people's mental models and the reality of what an IPA can do. They also found that people lack trust in IPAs' ability to conduct tasks effectively and that users saw speech interaction as something that had to be learned. Porcheron *et al.* (2017) report how IPAs are used by multiple users at once, mapping the collaborative mechanism and structure of the interaction, as well as highlighting the challenges in mutual social silences. Finally, Cowan *et al.* (2017) examined barriers to the more frequent use of IPAs, including usability difficulties, social embarrassment with public use and the negative impact of human likeness on UX.

5.2.9. *User memory*

Three papers assessed the effects of interface design on *user memory.* In a study investigating user recall of menu options in IVRS, recall was significantly impaired when five or more options were presented, or if system messages were followed by suffixes (i.e. 'pseudowords', natural-language prompts or system beeps) impairing memorization regardless of message (Le Bigot *et al.*, 2013). A study on cognitive load for speech interface users scheduling health appointments showed that working memory span did not affect appointment recall, nor did strategies used to reduce working memory such as reducing the number of menu options or providing confirmation of the system interactions (Wolters *et al.*, 2009). Another study reported that more content was recalled when information was provided by a human speaker rather than a machine (Knutsen *et al.*, 2017).

5.2.10. *Miscellaneous*

One paper could not be categorized. Ramanarayanan *et al.* (2017) tested the efficacy of crowd source evaluations of engagement levels between non-native English speakers and a computer-assisted language learning system. While findings showed consistent ratings of engagement when the human, computer or both were communicating to one another in the videos, there were low correlation levels between self-evaluations and third-party evaluations of engagement.

6. DISCUSSION

Our review reveals the state of empirical speech interface research within the field of HCI at present. We show that there is a very similar number of publications in both conferences and journals. Methodologically, papers tend to explore users' performance with, and behaviours and attitudes towards, a range of interactive prototypes, commercial or WoZ systems. Analysing aspects of task and interaction performance with systems was common, as were self-report questionnaires, particularly those measuring concepts like usability and user attitudes. However, many of the questionnaires used lack reliability or validity testing and are not consistently used across the research reviewed, creating issues with the validity of measurement. More studies in the review tended to compare aspects of design choices or investigate theory-inspired research questions (*usability/theory-based* research), than exploring wider working systems or prototypes (*system experience* research). Our categorization of research topics found that speech interface work published in HCI venues coalesced around nine main topics. Based on these findings we point to a number of key challenges in the research, methodological approaches and evaluation of speech interfaces to be addressed by future research efforts.

6.1. Research challenges

6.1.1. *Engaging with technological advances in speech-based HCI research*

While notable advances in speech technology have developed in recent years, these have not necessarily been a primary focus in the more current papers discussed in this review. The implementation of deep neural networks and cognitive computing that facilitate artificial intelligence, for example, are accredited as having a significant influence on the growth of everyday use of speech interfaces (McTear *et al.*, 2016). These approaches, combined with advanced speech synthesis techniques like WaveNet[iv], allow for the addition of layers of humanlike qualities in speech (e.g. more advanced emotion, personality and improved conversational abilities). In the more recent papers discussed in this review, there are few direct references as to how these technological advancements

[4] https://deepmind.com/blog/wavenet-generative-model-raw-audio/

Downloaded from https://academic.oup.com/iwc/article/31/4/349/5560013 by Georgia Institute of Technology user on 23 September 2025

impact on users in speech-based HCI. Technical discussions of systems were excluded from this review, partially accounting for this absence. It might also be explained by the distinct separation between a system's technological mechanics and the behaviours they produce.

While features such as WaveNet are being implemented into existing commercial IPAs (specifically Google Assistant; Oord *et al.*, 2017), the papers discussed in this review focus more on general usage and usability aspects of these systems (Luger and Sellen, 2016; Cowan *et al.*, 2017) as well as the anatomy of multi-user interaction (Porcheron *et al.*, 2017). Another important point to consider is while the systems described in more recent papers (e.g. those researching IPAs) have advanced significantly beyond what was conceivable in research from decades prior, some of the underlying research concepts and theories have persisted. For example, research on partner models was discussed in papers decades apart (Amalberti *et al.*, 1993; Cowan *et al.*, 2015). Similarly, both Buchheit and Moher (1990) and Clark *et al.* (2016) discussed elements of assertive language in HCI partners, noting that expectations of assertiveness and indirectness are influenced by the machinelike or humanlike nature of their interaction partners. This highlights the importance of accounting for the historical nature of speech-based HCI research, although this may be truer for theoretically-grounded research. It is important to consider the extent to which earlier research can be applied to current interactions with speech interfaces—past research needs to be critically accounted for. Technological advancements may be significant moderators in how past research should be applied to future efforts. Indeed, future speech-based HCI research may need to include more recent developments in speech technology in order to ground work in the context of state-of-the-art systems and some of their specific features.

6.1.2. *Developing theories of speech interface interaction*

Our review highlights that speech interface work has been successful in producing theories and effects that have been widely adopted by the field (e.g. the *similarity-attraction effect* (Dahlbäck *et al.*, 2007; Nass and Lee, 2000); *consistency-attraction* (Lee and Nass, 2003)). These concepts focus on explaining user attitudes, allowing researchers to predict, hypothesize and interpret the impact of particular design decisions. Yet our review showed a similar set of concepts or theories for understanding user language choices in speech interface interactions is lacking in the literature. Hone and Baber (2001) did discuss the concept of *habitability*—the congruence between users' language and the language a system can actually process, suggesting that systems may need to be designed with habitability in mind. Design strategies such as spoken prompts can provide users with a better understanding of what types of utterances a system can understand and consequently may help understand what language users might produce given certain design specifications. This type of insight is critical as understanding what influences users' language input is as fundamental to speech HCI as understanding touch or selection-based behaviours in other interaction modalities. Although few other formal effects or theories have been explored in the literature, our review does illuminate a consensus and emerging debate that could be used as a foundation for additional theory building in this regard. A number of papers reviewed propose that language in speech interface interaction is driven by the assumptions users have of partner abilities (i.e. our partner models) and that they adapt speech choices accordingly (e.g. Amalberti *et al.*, 1993), similar to mechanisms proposed in human-human communication (Branigan *et al.*, 2011; Brennan and Clark, 1996). The evidence for this in our review is mostly based on the differences in speech production between human-human and human-computer dialogue, although a study that looks at the impact of system design debates the influence of partner models on all language choices (Cowan *et al.*, 2015). As well as influencing user language choices, these expectations may also affect evaluation of system output. One paper demonstrated the relationship between an interface's voice and the perceptions of politeness in receiving computer instructions, based on expectations of linguistic capabilities (Clark *et al.*, 2016). A key debate around the role of partner models seems to be persisting in HCI based publications (Amalberti *et al.*, 1993; Cowan *et al.*, 2015), which reflects wider debates on more general language production in psycholinguistics (Brennan and Clark, 1996; Horton and Keysar, 1996; Keysar *et al.*, 1998). An opportunity for the HCI community lies in building on these types of research. From this a more robust theory of language production and interlocutor perceptions may grow, which will be able to inform design and interaction science research (Howes *et al.*, 2014) in the speech domain.

6.1.3. *Achieving critical mass*

The review also shows the somewhat fragmented nature of research within the area, highlighting a need for developing critical mass in a number of these topics. Three primary topics dominated the papers, with the remaining areas seeing fewer than eight papers in each category. Even within the more common topics the work tends to be fractured into specific sub-topics with few papers representing these, making it hard to identify a consensus or a summary of results. For example, system speech production research (Section 5.2) explores four distinct sub-topics; speech synthesis (e.g. Nass and Lee, 2000), content (e.g. Hofmann *et al.*, 2014), spatiotemporal aspects (e.g. Iqbal *et al.*, 2011) and general speech production (e.g. Clark *et al.*, 2016), with each study varying in the concepts explored within those sub-topics. Similar variation was observed for papers concerning design insight and comparing speech with other modalities. While diversity across speech HCI research is encouraging, themes discussed in this paper point to a noticeable fragmentation of topics explored. However, this was not the case for all

Downloaded from https://academic.oup.com/iwc/article/31/4/349/5560013 by Georgia Institute of Technology user on 23 September 2025

areas. Papers discussing speech technologies for development (Section 5.2.7), for example, tended to hold a common thread of exploring interface use with novice users and people with lower levels of literacy. They also provide example benefits and drawbacks of speech interface design and use with these communities (e.g. Medhi *et al.*, 2009; Raza *et al.*, 2013). The growing interest in speech interfaces should be used to contribute greater cohesion to the current body of knowledge in speech-based HCI. Indeed, increased focus could help develop and embed new and existing theories, concepts and paradigms (e.g. those in Section 6.1.1.) across the numerous research topics within the area.

6.1.4. Further speech interaction design work is required

Our review showed that design work was one of the more common research topics being addressed, with some design recommendations also emerging in papers categorized as investigating other primary topics. The scope of design work and the generalizability of guidelines that were developed varied in these papers. Recommendations on improving ASR (e.g. Murray *et al.*, 1996), for example, are more widely applicable than considerations for improving speech-based ATM interactions (Johnson and Coventry, 2001). A number of papers took more of a usability engineering-based approach, comparing interface designs e.g. menu structures (e.g. Wilkie *et al.*, 2005; Wilke *et al.*, 2007) and observing their effect on user attitudes and performance. While some of these findings are useful for specific interaction contexts, more generalizable design guidelines were lacking. While such guidelines in HCI exist for graphical interfaces (e.g. Nielsen, 1994; Norman, 2013; Shneiderman *et al.*, 2016), some of which have been revised over the years, the same guidelines for speech interfaces were scarce in this review. Though texts exist to support the design of voice user interfaces (Pearl, 2016), there is still a need to establish design considerations and robust heuristics for developing user centred speech interactions. We also need to identify what design methods and theories work best in a speech context, rather than solely importing those that have worked on other modalities and contexts without critical reflection (Corbett and Weber, 2016). This can be achieved if the design community in HCI embraces speech interfaces and its ongoing technological developments more widely in the future.

6.1.5. Investigating multiple user contexts

The majority of papers in the review researched single user interactions. Yet with IPAs like Alexa or Google Home being used in social spaces, the interaction opportunities and challenges of multiple user scenarios need to be further understood. Research in the HCI community has begun to shed light on group-based IPA interactions, in particular the dynamics of user behaviour in this experience (e.g. Porcheron *et al.*, 2017). The technical challenges in terms of recognizing who is speaking (speaker diarization) and whether they are addressing the system are well researched (Batliner *et al.*, 2008). The HCI community is well placed to inform these technical advancements, particularly by understanding users' expectations, behaviours, social dynamics and the interaction barriers that may limit the effectiveness of speech interfaces in group situations. The study of situations that include multiple users and multiple speech interfaces would be highly valuable as this is likely to become more common in the near future.

6.2. Methodological and evaluation challenges

6.2.1. Improving measure reliability, validity and consistency

Across the papers in our review, there was a range of objective and subjective concepts being measured. A high proportion of research in our review used self-report questionnaires to measure other concepts like user satisfaction, usability, user attitudes towards speech interfaces and general user experience. A number of these self-report measures lacked any reliability or validity testing. Self-report questionnaires specific to speech interfaces that have been assessed for internal reliability do exist in the literature (e.g. SASSI (Hone and Graham, 2000)), although these need further validity testing (Hone, 2014; Hone and Graham, 2000). We were surprised to see that, although SASSI was deployed once in the reviewed papers (Hofmann *et al.*, 2014) and subscales used in others (e.g. Hone and Baber, 2001), the scale is not widely used in HCI publications. The field therefore needs to make a concerted effort in developing well-validated, reliable subjective measures for more UX related dimensions in speech interaction. This is made more challenging as our review shows a lack of consistency in the concepts being measured across papers, which presents difficulties for the domain in building robust measures and a body of knowledge around specific concepts or paradigms. Discussion within the HCI field is needed to map out the measures, concepts and paradigms that are pertinent to this research area, so as to identify the priority areas for efforts of metrics development to focus on. This type of mapping could also bring more cohesion to speech HCI work as a whole.

6.2.2. Evaluating speech interfaces in real world contexts

Throughout the papers in the review, there was a strong tendency to use laboratory-based approaches to investigate speech interface interactions. These laboratory experiments often use controlled prototypes and/or WoZ simulated systems. This work is essential in developing the necessary science in the field. Yet further effort is also needed in understanding how these lab-based findings transfer to real-world contexts, as well as comparing interactions across multiple contexts. As Corbett and Weber (2016) highlight, theories may not automatically translate from one context of interaction to another. More studies using *in the wild* experimental approaches would be useful in this regard. Indeed, we feel that more investigative work exploring wider issues in speech interface interaction—but not necessarily focused on comparing experimental

Downloaded from https://academic.oup.com/iwc/article/31/4/349/5560013 by Georgia Institute of Technology user on 23 September 2025

conditions—should be encouraged. This type of work creates valuable insight into the impact of context on speech interface use, while also shedding light on what types of situations these interfaces are used in and what types of UX issues emerge. Our review indicates that there has been some recent progress towards this (e.g. Luger and Sellen, 2016; Porcheron *et al.*, 2017; Cowan *et al.*, 2017). Future work should look to build on these efforts.

6.2.3. Reducing barriers to building speech interfaces

As mentioned above we found a lack of design related work on speech interfaces. The scarcity of design research may speak to a perceived high barrier in developing operational speech interface prototypes to explore in these types of studies. A number of packages and toolkits do exist that can help in developing prototypical speech interfaces such as OpenDial (Lison and Kennington, 2016), IrisTK (www.iristk.net) or Aspect Prophecy (www.aspect.com/). Amazon's Skills Kit (https://developer.amazon.com/alexa-skills-kit) also allows people to develop skills for Alexa that could be deployed as prototypes for interaction studies. Highly flexible and easy to use tools to develop speech interfaces for research and prototyping purposes should be further encouraged to facilitate speech interface prototyping and design.

6.3. Limitations

The aim of this research was to identify the main trends, themes and methods used in speech research published specifically in core HCI venues. To do this we focused on a comprehensive list of the top HCI publication venues, based on Google Scholar, Thomson Reuters and SJCR rankings. This allows us to focus specifically on papers within HCI as a field, similar to approaches used in other HCI reviews. However, there are a number of studies published in other fields that are relevant to speech HCI work, including cognitive psychology, ergonomics, linguistics and speech technology (e.g. Branigan *et al.*, 2011; Large et al., 2017; Cowan and Branigan, 2015; Mendelson and Aylett, 2017). Although these areas are beyond the scope of the review, similar topics are present in papers published in these areas. Future reviews should look to concentrate on these fields and compare directly with the findings from this research. Our research also focused on empirical studies, and thus omitted research without user evaluation or user-based data collection that may have been influential to the field or that focus more on recent technological advancements in speech technology. Similarly, as this review spans decades of work, it is important to consider the age of research when applying their findings to current speech interface interactions. We also decided to exclude research on embodied conversational agents and robotics in speech interfaces, because of the danger of embodiment-based factors confounding speech-related effects. Future reviews should focus on these domains.

7. CONCLUSION

This paper maps out the trends and findings of speech research published in the field of HCI, in the hope of stimulating further speech-based research by describing the current state of the field. Based on the review of 99 papers from core HCI publication venues, we highlight nine primary research topics and identify the key methodological approaches taken in the research reviewed. From this we identify key research, methodological and evaluation challenges in developing further theory and design work in this area, expanding contexts of human-system evaluation, reducing technical barriers to research and improving consistency and rigor in evaluation measures.

FUNDING

New Horizons grant from the Irish Research Council entitled 'The COG-SIS Project: Cognitive effects of Speech Interface Synthesis' (R17339).

REFERENCES

Alm, N., Todman, J., Elder, L. and Newell, A. F. (1993) *Computer Aided Conversation for Severely Physically Impaired Non-speaking People,* In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, pp. 236–241. ACM Press.

Amalberti, R., Carbonell, N. and Falzon, P. (1993) User representations of computer systems in human-computer speech interaction. *Int. J. Man-Mach. Stud.*, 38, 547–566.

Aylett, M. P., Kristensson, P. O., Whittaker, S. and Vazquez-Alvarez, Y. (2014) *None of a CHInd: Relationship Counselling for HCI and Speech Technology.* In *CHI'14 Extended Abstracts on Human Factors in Computing Systems*, pp. 749–760. ACM Press.

Aylett, M. P., Vazquez-Alvarez, Y. and Baillie, L. (2015) *Interactive Radio: A New Platform for Calm Computing,* In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, pp. 2085–2090. ACM Press.

Bargas-avila, J. A. and Hornbæk, K. (2011) Old wine in new bottles or novel challenges? A critical analysis of empirical studies of user experience, In *Proceedings of the SIGCHI conference on human factors in computing systems*.

Batliner, A., Hacker, C. and Nöth, E. (2008) To talk or not to talk with a computer. *J. Multimodal User In.*, 2, 171.

Begany, G. M., Sa, N. and Yuan, X. (2015) Factors affecting user perception of a spoken language vs. textual search interface: A content analysis. *Interact. Comput.*, 28, 170–180.

Bekker, M. M., van Nes, F. L. and Juola, J. F. (1995) A comparison of mouse and speech input control of a text-annotation system. *Behav. Inf. Technol.*, 14, 14–22.

Berglund, A. and Johansson, P. (2004) Using speech and dialogue for interactive TV navigation. *Universal Access Inf.*, 3, 224–238.

Bhatia, S. and McCrickard, S. (2006) *Listening to your inner voices: Investigating means for voice notifications,* In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1173–1176. ACM Press.

Downloaded from https://academic.oup.com/iwc/article/31/4/349/5560013 by Georgia Institute of Technology user on 23 September 2025

Bickmore, T. W., Pfeifer, L. M. and Jack, B. W. (2009) *Taking the time to care: empowering low health literacy hospital patients with virtual nurse agents,* In *Proceedings of the SIGCHI conference on human factors in computing systems* , pp. 1265–1274. ACM Press.

Branigan, H. P., Pickering, M. J., Pearson, J., McLean, J. F. and Brown, A. (2011) The role of beliefs in lexical alignment: Evidence from dialogs with humans and computers. *Cognition*, 121, 41–57.

Braun, V. and Clarke, V. (2006) Using thematic analysis in psychology. *Qual. Res. Psychol.*, 3, 77–101.

Breazeal, C. (2003) Emotion and sociable humanoid robots. *Int. J. Hum. Comput. Stud.*, 59, 119–155.

Brennan, S. E. and Clark, H. H. (1996) Conceptual pacts and lexical choice in conversation. *J. Exp. Psychol. Learn. Mem. Cogn.*, 22, 1482–1493.

Brown, P. and Levinson, S. C. (1987) *Politeness: Some Universals in Language Usage*. Cambridge University Press.

Bruce, A., Nourbakhsh, I. and Simmons, R. (2002) *In Proceedings 2002 IEEE International Conference on Robotics and Automation (Cat. No. 02CH37292)*, (Vol. 4, pp. 4138–4142).

Buchheit, P. and Moher, T. (1990) Response assertiveness in human-computer dialogues - Science Direct. https://www.sciencedirect.com/science/article/pii/S0020737305801786 (accessed April 12, 2018.

BusinessWire. (2018) Global intelligent virtual assistant market, 2018–2023. https://www.businesswire.com/news/home/20180723005506/en/Global-Intelligent-Virtual-Assistant-Market-2018-2023-Market (accessed September 21, 2018).

Chan, W., Jaitly, N., Le, Q. and Vinyals, O. (2016) Listen, attend and spell: a neural network for large vocabulary conversational speech recognition. *In 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4960–4964.

Clark, L., Ofemile, A., Adolphs, S. and Rodden, T. (2016) A multimodal approach to assessing user experiences with agent helpers. *ACM Trans. Interact. Intell. Syst.*, 6, 1–29.

Cohen, P., Cheyer, A., Horvitz, E., El Kaliouby, R. and Whittaker, S. (2016) On the future of personal assistants. In *Proceedings of the 2016 CHI conference extended abstracts on human factors in computing systems*, pp. 1032–1037. ACM, New York, NY, USA.

Cohen, P. R., Buchanan, M. C., Kaiser, E. J., Corrigan, M., Lind, S. and Wesson, M. (2013) *Demonstration of sketch-thru-plan: a multimodal interface for command and control*, pp. 69–70. ACM Press.

Corbett, E. and Weber, A. (2016) What can I say? Addressing user experience challenges of a mobile voice user interface for accessibility. In *Proceedings of the 18th international conference on human-computer interaction with mobile devices and services*, pp. 72–82. ACM, New York, NY, USA.

Cowan, B. R. and Branigan, H. P. (2015) *Does voice anthropomorphism affect lexical alignment in speech-based human-computer dialogue?, Proc. 3rd Annual ACM SIGGRAPH Symposium*, pp. 155–159.

Cowan, B. R., Branigan, H. P., Obregón, M., Bugis, E. and Beale, R. (2015) Voice anthropomorphism, interlocutor modelling and alignment effects on syntactic choices in human−computer dialogue. *Int, J. Hum. Comp. Stud.*, 83, 27–42.

Cowan, B. R., Gannon, D., Walsh, J., Kinneen, J., O'Keefe, E. and Xie, L. (2016) *Towards Understanding How Speech Output Affects Navigation System Credibility,* In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pp. 2805–2812. ACM Press.

Cowan, B. R., Pantidi, N., Coyle, D., Morrissey, K., Clarke, P., Al-Shehri, S. and Bandeira, N. (2017) '*What Can I Help You With?': Infrequent Users' Experiences of Intelligent Personal Assistants, Proc. 3rd Annual ACM SIGGRAPH Symposium,* In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services*, pp. 1–12. ACM Press.

Cuendet, S., Medhi, I., Bali, K. and Cutrell, E. (2013) VideoKheti: making video content accessible to low-literate and novice users. In *Conference on human factors in computing systems-proceedings*, pp. 2833–2842.

Culbertson, G., Shen, S., Jung, M. and Andersen, E. (2017) *Facilitating Development of Pragmatic Competence through a Voice-driven Video Learning Interface,* In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pp. 1431–1440. ACM Press.

Dahlbäck, N., Jönsson, A. and Ahrenberg, L. (1993) *Wizard Of Oz Studies—Why And How. Intelligent User Interfaces,* In *Proceedings of the International Conference on Intelligent User Interfaces (IUI 1993)*. 8. pp. 193–200. ACM, New York, NY, USA.

Dahlbäck, N., Wang, Q., Nass, C. and Alwin, J. (2007) Similarity is more important than expertise: accent effects in speech interfaces. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 1553–1556. ACM, New York, NY, USA.

Dai, L., Goldman, R., Sears, A. and Lozier, J. (2005) Speech-based cursor control using grids: modelling performance and comparisons with other solutions. *Behaviour and Information Technology*, 24, 219–230.

DeRenzi, B., Dell, N., Wacksman, J., Lee, S. and Lesh, N. (2017) Supporting community health workers in India through voice- and web-based feedback. In *Proceedings of the 2017 CHI conference on human factors in computing systems*, pp. 2770–2781. ACM, New York, NY, USA.

Derriks, B. and Willems, D. (1998) Negative feedback in information dialogues: identification, classification and problem-solving procedures. *Int. J. Hum. Comput. Stud.*, 48, 577–604.

Dulude, L. (2002) Automated telephone answering systems and aging. *Behaviour & Information Technology*, 21, 171–184.

Evans, R. E. and Kortum, P. (2010) The impact of voice characteristics on user response in an interactive voice response system. *Interact. Comput.*, 22, 606–614.

Feng, J. and Sears, A. (2004) Using confidence scores to improve hands-free speech based navigation in continuous dictation systems. *ACM Trans. Comput. Hum. Interact.*, 11, 329–356.

Feng, J., Sears, A. and Karat, C.-M. (2006) A longitudinal evaluation of hands-free speech-based navigation during dictation. *Int. J. Hum. Comput. Stud.*, 64, 553–569.

Feng, J., Zhu, S., Hu, R. and Sears, A. (2011) Speech-based navigation and error correction: a comprehensive comparison of two solutions. *Univers. Access Inf. Soc.*, 10, 17–31.

Fickas, S., Sohlberg, M. and Hung, P.-F. (2008) Route-following assistance for travelers with cognitive impairments: a comparison

Downloaded from https://academic.oup.com/iwc/article/31/4/349/5560013 by Georgia Institute of Technology user on 23 September 2025

of four prompt modes. *Int. J. Hum. Comput Stud.*, 66, 876–888.

Gong, L. and Lai, J. (2001) *Shall we mix synthetic speech and human speech? Impact on users' performance, perception, and attitude,* In *proceedings of the ACM Conference on Human Factors in Computing Systems (CHI 2001)*, pp. 158–165.

Hakulinen, J., Turunen, M., Salonen, E.-P. and Räihä, K.-J. (2004) *Tutor Design for Speech-Based Interfaces,* In *Proceedings of the 5th conference on Designing interactive systems: processes, practices, methods, and techniques*, pp. 155–164. ACM Press.

Han, S., Philipose, M. and Ju, Y.-C. (2013) *NLify: Lightweight Spoken Natural Language Interfaces via Exhaustive Paraphrasing,* In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, pp. 429–438. ACM Press.

Hara, K. and Iqbal, S. T. (2015) *Effect of Machine Translation in Interlingual Conversation: Lessons from a Formative Study,* In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pp. 3473–3482. ACM Press.

Harada, S., Lester, J., Patel, K., Saponas, T. S., Fogarty, J., Landay, J. A. and Wobbrock, J. O. (2008) VoiceLabel: using speech to label mobile sensor data. In *Proceedings of the 10th international conference on multimodal interfaces*, pp. 69–76. ACM, New York, NY, USA.

Harada, S., Wobbrock, J. O., Malkin, J., Bilmes, J. A. and Landay, J. A. (2009) *Longitudinal Study of People Learning to Use Continuous Voice-Based Cursor ControlProc.* In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 347–356. ACM Press.

Hauptmann, A. G. and McAvinney, P. (1993) Gestures with speech for graphic manipulation. *Int. J. Man Mach. Stud.*, 38, 231–249.

Hofmann, H., Tobisch, V., Ehrlich, U., Berton, A. and Mahr, A. (2014) Comparison of speech-based in-car HMI concepts in a driving simulation study. In *Proceedings of the 19th international conference on intelligent user interfaces*, pp. 215–224. ACM, New York, NY, USA.

Hone, K. (2014) *Usability measurement for speech systems: SASSI revisited, Designing Speech and Language Interactions Workshop, CHI 2014*, 4.

Hone, K. S. and Baber, C. (2001) Designing habitable dialogues for speech-based interaction with computers. *Int. J. Hum. Comput. Stud.*, 54, 637–662.

Hone, K. S. and Graham, R. (2000) Towards a tool for the subjective assessment of speech system interfaces (SASSI). *Nat. Lang. Eng.*, 6, 287–303.

Hornbæk, K. (2006) Current practice in measuring usability: challenges to usability studies and research. *Int. J. Hum. Comput. Stud.*, 64, 79–102.

Horton, W. S. and Keysar, B. (1996) When do speakers take into account common ground? *Cognition*, 59, 91–117.

Howell, M., Love, S. and Turner, M. (2005) The impact of Interface metaphor and context of use on the usability of a speech-based mobile city guide service. *Behaviour & Information Technology*, 24, 67–78.

Howell, M., Love, S. and Turner, M. (2006) Visualisation improves the usability of voice-operated mobile phone services. *Int. J. Hum. Comput. Stud.*, 64, 754–769.

Howes, A., Cowan, B. R., Janssen, C. P., Cox, A. L., Cairns, P., Hornof, A. J. and Pirolli, P. (2014) *Interaction Science SIG: Overcoming Challenges*, pp. 1127–1130. ACM Press.

Hu, J., Winterboer, A., Nass, C. I., Moore, J. D. and Illowsky, R. (2007) *Context & usability testing: user-modeled information presentation in easy and difficult driving conditions* In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 1343–1346. ACM Press.

Iqbal, S. T., Horvitz, E., Ju, Y.-C. and Mathews, E. (2011) *Hang on a sec!: effects of proactive mediation of phone conversations while driving,* In *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 463–472. ACM Press.

Jeon, M., Gable, T. M., Davison, B. K., Nees, M. A., Wilson, J. and Walker, B. N. (2015) Menu navigation with in-vehicle technologies: auditory menu cues improve dual task performance, preference, and workload. *Int. J. Hum. Comput. Int.*, 31, 1–16.

Johnson, G. I. and Coventry, L. (2001) 'You talking to me?' Exploring voice in self-service user interfaces. *Int. J. Hum. Comput. Int.*, 13, 161–186.

Jokinen, K. (2006) Adaptation and user expertise modelling in AthosMail. *Univers. Access Inf. Soc.*, 4, 374–392.

Jokinen, K. and MacTear, M. (2010) *Spoken Dialogue Systems*. Morgan & Claypool, San Rafael, Calif.

Kallinen, K. and Ravaja, N. (2005) Effects of the rate of computer-mediated speech on emotion-related subjective and physiological responses. *Behaviour & Information Technology*, 24, 365–373.

Kamitis (2016) *Intelligent Personal Assistant-Products, Technologies and Market: 2017–2022*.

Katz, J., Aspden, P. and Reich, W. A. (1997) Public attitudes toward voice-based electronic messaging technologies in the United States: a national survey of opinions about voice response units and telephone answering machines. *Behaviour and Information Technology*, 16, 125–144.

Keysar, B., Barr, D. J. and Horton, W. S. (1998) The egocentric basis of language use: insights from a processing approach. *Curr. Dir. Psychol. Sci.*, 7, 46–50.

Knutsen, D., Le Bigot, L. and Ros, C. (2017) Explicit feedback from users attenuates memory biases in human-system dialogue. *Int. J. Hum. Comput. Stud.*, 97, 77–87.

Kousidis, S., Kennington, C., Baumann, T., Buschmeier, H., Kopp, S. and Schlangen, D. (2014) A multimodal in-car dialogue system that tracks the driver's attention. In *Proceedings of the 16th international conference on multimodal interaction*, pp. 26–33. ACM, New York, NY, USA.

Kumar, A., Paek, T. and Lee, B. (2012) *Voice typing: a new speech interaction model for dictation on touchscreen devices,* In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 2277–2286. ACM Press.

Kuno, Y., Sadazuka, K., Kawashima, M., Yamazaki, K., Yamazaki, A. and Kuzuoka, H. (2007) *Museum guide robot based on sociological interaction analysis,* In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 1191–1194. ACM Press.

Lai, J. and Vergo, J. (1997) MedSpeak: report creation with continuous speech recognition. In *Proceedings of the ACM SIGCHI conference on human factors in computing systems*, pp. 431–438. ACM, New York, NY, USA.

Downloaded from https://academic.oup.com/iwc/article/31/4/349/5560013 by Georgia Institute of Technology user on 23 September 2025

LaPlante, M. P. (1992) Assistive technology devices and home accessibility features: prevalence, payment, need, and trends. *Adv Data Vital Health Stat*. 217, 1–11.

Large, D. R., Clark, L., Quandt, A., Burnett, G. and Skrypchuk, L. (2017) Steering the conversation: a linguistic exploration of natural language interactions with a digital assistant during simulated driving. *Applied ergonomics*, 63, 53–61.

Le Bigot, L., Jamet, E., Rouet, J.-F. and Amiel, V. (2006) Mode and modal transfer effects on performance and discourse organization with an information retrieval dialogue system in natural language. *Comput. Human Behav.*, 22, 467–500.

Le Bigot, L. L., Caroux, L., Ros, C., Lacroix, A. and Botherel, V. (2013) Investigating memory constraints on recall of options in interactive voice response system messages. *Behaviour & Information Technology*, 32, 106–116.

Le Bigot, L., Terrier, P., Amiel, V., Poulain, G., Jamet, E. and Rouet, J.-F. (2007) Effect of modality on collaboration with a dialogue system. *Int. J. Hum. Comput. Stud.*, 65, 983–991.

Leahu, L., Cohn, M. and March, W. (2013) How categories come to matter. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 3331–3334. ACM, New York, NY, USA.

Lee, K. M. and Nass, C. (2003) Designing social presence of social actors in human Computer Human Interaction. (2003) *New Horizons*, 5, 289–94.

Liapis, C. (2011) A primer to human threading. *Comput. Hum. Behav.*, 27, 138–143.

Limerick, H., Moore, J. W. and Coyle, D. (2015) *Empirical evidence for a diminished sense of agency in speech interfaces,* In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pp. 3967–3970. ACM Press.

Lison, P. and Kennington, C. (2016) OpenDial: a toolkit for developing spoken dialogue systems with probabilistic rules. In *Proceedings of ACL-2016 system demonstrations*, pp. 67–72. Association for Computational Linguistics, Berlin, Germany http://anthology.aclweb.org/P16-4012 (accessed February 22, 2019).

Litman, D. J. and Pan, S. (2002) Designing and evaluating an adaptive spoken dialogue system. *User Model. User-Adapt. Interact.*, 12, 111–137.

Löhr, A. and Brügge, B. (2008) Mixed-initiative dialog management for speech-based interaction with graphical user interfaces. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 979–988. ACM, New York, NY, USA.

Luger, E. and Sellen, A. (2016) Like having a really bad PA': the gulf between user expectation and experience of conversational agents. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pp. 5286–5297. ACM, New York, NY, USA.

Lunsford, R. and Oviatt, S. (2006) *Human perception of intended addressee during computer-assisted meetings,* In *Proceedings of the 8th international conference on Multimodal interfaces*, pp. 20–27. ACM Press.

Lunsford, R., Oviatt, S. and Coulston, R. (2005) *Audio-visual cues distinguishing self- from system-directed speech in younger and older adults,* In *Proceedings of the 7th international conference on Multimodal interfaces*, p. 167–174. ACM Press.

Mascetti, S., Picinali, L., Gerino, A., Ahmetovic, D. and Bernareggi, C. (2016) Sonification of guidance data during road crossing for people with visual impairments or blindness. *Int. J. Hum. Comput. Stud.*, 85, 16–26.

McTear, M., Callejas, Z. and Griol, D. (2016) *The conversational interface: talking to smart devices*. Springer.

Medhi, I., Gautama, S. N. N. and Toyama, K. (2009) *A comparison of mobile money-transfer UIs for non-literate and semi-literate users,* In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1741–1750. ACM Press.

Medhi, I., Patnaik, S., Brunskill, E., Gautama, S. N. N., Thies, W. and Toyama, K. (2011) Designing mobile interfaces for novice and low-literacy users. *ACM Trans. Comput. Hum. Interact.*, 18, 1–28.

Mekler, E. D., Bopp, J. A., Tuch, A. N. and Opwis, K. (2014) *A systematic review of quantitative studies on the enjoyment of digital entertainment games,* In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*, pp. 927–936. ACM Press.

Melichar, M. and Cenek, P. (2006) *From vocal to multimodal dialogue management,* In *Proceedings of the 8th international Conference on Multimodal interfaces*, pp. 59–67. ACM Press.

Mendelson, J. and Aylett, M. P. (2017) Beyond the listening test: an interactive approach to TTS evaluation, In *INTERSPEECH*, pp. 249–253. ISCA.

Moller, S., Engelbrecht, K.-P., Kuhnel, C., Wechsung, I. and Weiss, B. (2009) A taxonomy of quality of service and Quality of Experience of multimodal human-machine interaction. In *2009 International Workshop on Quality of Multimedia Experience. IEEE*, 7–12.

Molnar, K. K. and Kletke, M. G. (1996) The impacts on user performance and satisfaction of a voice-based front-end Interface for a standard software tool. *Int. J. Hum. Comput. Stud.*, 45, 287–303.

Moran, S., Pantidi, N., Bachour, K., Fischer, J. E., Flintham, M., Rodden, T. and Johnson, S. (2013) *Team reactions to voiced agent instructions in a pervasive game,* In *Proceedings of the 2013 international conference on Intelligent user interfaces*, pp. 371–382. ACM Press.

Munteanu, C., Irani, P., Oviatt, S., Aylett, M., Penn, G., Pan, S. and Nakamura, K. (2017) Designing speech, acoustic and multimodal interactions. In *Proceedings of the 2017 CHI conference extended abstracts on human factors in computing systems*, pp. 601–608. ACM, New York, NY, USA.

Munteanu, C. and Penn, G. (2014) Speech-based interaction: myths, challenges, and opportunities. In *CHI'14 extended abstracts on human factors in computing systems*, pp. 1035–1036. ACM, New York, NY, USA.

Murata, A. and Takahashi, Y. (2002) Does speech input system lead to improved performance for elderly? Discussion of problems when using speech interfaces for elderly. *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, https://okayama.pure.elsevier.com/en/publications/does-speech-input-system-lead-to-improved-performance-for-elderly (accessed February 17, 2019).

Murray, A. C., Jones, D. M. and Frankish, C. R. (1996) Dialogue design in speech-mediated data-entry: the role of syntactic constraints and feedback. *Int. J. Hum. Comput. Stud.*, 45, 263–286.

Nass, C. and Lee, K. M. (2000) *Does computer-generated speech manifest personality? an experimental test of similarity-attraction,* In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pp. 329–336. ACM Press.

Downloaded from https://academic.oup.com/iwc/article/31/4/349/5560013 by Georgia Institute of Technology user on 23 September 2025

Nielsen, J. (1994) Enhancing the explanatory power of usability heuristics. In *Proceedings of the SIGCHI conference on human factors in computing systems, Proc. 3rd Annual ACM SIGGRAPH Symposium*, pp. 152–158. ACM.

Norman, D. A. (2013) *The Design of Everyday Things (revised and expanded edition).* Basic Books, New York.

van den Oord, A., Li, Y., Babuschkin, I., Simonyan, K., Vinyals, O., Kavukcuoglu, K. and Hassabis, D. (2017) Parallel WaveNet: fast high-fidelity speech synthesis. https://ai.google/research/pubs/pub46540 (accessed September 21 2018).

Oviatt, S., Coulston, R. and Lunsford, R. (2004) When do we interact multimodally? Cognitive load and multimodal communication patterns. In *Proceedings of the 6th international conference on multimodal interfaces*, pp. 129–136. ACM, New York, NY, USA.

Oviatt, S., Swindells, C. and Arthur, A. (2008) *Implicit user-adaptive system engagement in speech and pen interfaces,* In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 969–978. ACM Press.

Pak, R., Czaja, S. J., Sharit, J., Rogers, W. A. and Fisk, A. D. (2008) The role of spatial abilities and age in performance in an auditory computer navigation task. *Comput. Hum. Behav.*, 24, 3045–3051.

Patel, N., Agarwal, S., Rajput, N., Nanavati, A., Dave, P. and Parikh, T. S. (2009) *A comparative study of speech and dialed input voice interfaces in rural India,* In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 51–54. ACM Press.

Pearl, C. (2016) *Designing Voice User Interfaces: Principles of Conversational Experiences* (1st edn). O'Reilly Media, Inc.

Perugini, S., Anderson, T. J. and Moroney, W. F. (2007) *A study of out-of-turn interaction in menu-based, IVR, voicemail systems,* In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 961–970. ACM Press.

Piper, A. M. and Hollan, J. D. (2008) Supporting medical conversations between deaf and hearing individuals with tabletop displays. *CSCW 08- conference proceedings, 2008 ACM conference on computer supported cooperative work*.

Porayska-Pomsta, K. and Mellish, C. (2013) Modelling human tutors' feedback to inform natural language interfaces for learning. *Int. J. Hum. Comput. Stud.*, 71, 703–724.

Porcheron, M., Fischer, J. E. and Sharples, S. (2017) Do Animals Have A ccents? Do animals have accents?: talking with agents in multiparty conversation. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*, pp. 207–219. ACM Press.

Price, K. J. and Sears, A. (2005) Speech-based text entry for mobile handheld devices: an analysis of efficacy and error correction techniques for server-based solutions. *Int. J. Hum. Comput. Interact.*, 19, 279–304.

Price, K. J., Lin, M., Feng, J., Goldman, R., Sears, A. and Jacko, J. A. (2006) Motion does matter: an examination of speech-based text entry on the move. *Universal Access Inf. Soc.*, 4, 246–257.

Qvarfordt, P., Jönsson, A. and Dahlbäck, N. (2003) The role of spoken feedback in experiencing multimodal interfaces as human-like, In *Proceedings of the 5th international conference on Multimodal interfaces* 8, 250–257.

Ramanarayanan, V., Leong, C. W., Suendermann-Oeft, D. and Evanini, K. (2017) *Crowdsourcing ratings of caller engagement in thin-slice videos of human-machine dialog: benefits and pitfalls, Proc. 3rd Annual ACM SIGGRAPH Symposium,* In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pp. 281–287. ACM Press.

Raza, A. A., Ul Haq, F., Tariq, Z., Pervaiz, M., Razaq, S., Saif, U. and Rosenfeld, R. (2013) Job opportunities through entertainment: virally spread speech-based services for low-literate users. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 2803–2812. ACM, New York, NY, USA.

Sammon, M. J., Brotman, L. S., Peebles, E. and Seligmann, D. D. (2006) MACCS: enabling communications for mobile workers within healthcare environments. In *Proceedings of the 8th conference on human-computer interaction with mobile devices and services*, pp. 41–44. ACM, New York, NY, USA.

Sato, D., Zhu, S., Kobayashi, M., Takagi, H. and Asakawa, C. (2011) Sasayaki: augmented voice web browsing experience. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 2769–2778. ACM, New York, NY, USA.

Schaffer, S., Schleicher, R. and Möller, S. (2015) Modeling input modality choice in mobile graphical and speech interfaces. *Int. J. Hum. Comput. Stud.*, 75, 21–34.

Sears, A., Lin, M. and Karimullah, A. S. (2002) Speech-based cursor control: understanding the effects of target size, cursor speed, and command selection. *Universal Access Inf. Soc.*, 2, 30–43.

Sears, A., Feng, J., Oseitutu, K. and Karat, C.-M. (2003) Hands-free, speech-based navigation during dictation: difficulties, consequences, and solutions. *Hum. Comput. Interact.*, 18, 229–257.

Shneiderman, B. (2000) The limits of speech recognition. *Communications of the ACM,* 43, 63–65.

Shneiderman, B. and Maes, P. (1997) Direct manipulation vs. interface agents. *Interactions*, 4, 42–61.

Shneiderman, B., Plaisant, C., Cohen, M., Jacobs, S., Elmqvist, N. and Diakopoulos, N. (2016) *Designing the User Interface: Strategies for Effective Human-Computer Interaction* (6th). Pearson.

Sivaraman, V., Yoon, D. and Mitros, P. (2016) Simplified audio production in asynchronous voice-based discussions. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pp. 1045–1054. ACM, New York, NY, USA.

Strait, M., Vujovic, L., Floerke, V., Scheutz, M. and Urry, H. (2015) *Too much humanness for human-robot interaction: exposure to highly humanlike robots elicits aversive responding in observers*, pp. 3593–3602. ACM Press.

Suhm, B., Bers, J., McCarthy, D., Freeman, B., Getty, D., Godfrey, K. and Peterson, P. (2002) *A comparative study of speech in the call center: natural language call routing vs. touch-tone menus*, p. 283. ACM Press.

Suhm, B., Myers, B. and Waibel, A. (2001) Multimodal error correction for speech user interfaces. *ACM Transactions on Computer-Human Interaction*, 8, 60–98.

Takayama, L. and Nass, C. (2008) Driver safety and information from afar: an experimental driving simulator study of wireless vs. in-car information services. *Int. J. Hum. Comput. Stud.*, 66, 173–184.

Truschin, S., Schermann, M., Goswami, S. and Krcmar, H. (2014) Designing interfaces for multiple-goal environments: experimental insights from in-vehicle speech interfaces. *ACM Trans. Comput.-Hum. Interact.*, 21, 1–7.

Tsukahara, W. and Ward, N. (2001) *Responding to subtle, fleeting changes in the user's internal state, Proc. 3rd Annual ACM SIG-*

Downloaded from https://academic.oup.com/iwc/article/31/4/349/5560013 by Georgia Institute of Technology user on 23 September 2025

GRAPH Symposium, In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 77–84. ACM Press.

Vashistha, A., Sethi, P. and Anderson, R. (2017) Respeak: a voice-based, crowd-powered speech transcription system. In *Proceedings of the 2017 CHI conference on human factors in computing systems*, pp. 1855–1866. ACM, New York, NY, USA.

Vetek, A. and Lemmelä, S. (2011) Could a dialog save your life? Analyzing the effects of speech interaction strategies while driving. In *Proceedings of the 13th international conference on multimodal interfaces*, pp. 145–152. ACM, New York, NY, USA.

Walker, M. A., Fromer, J., Di Fabbrizio, G., Mestel, C. and Hindle, D. (1998) *What can I say? Evaluating a spoken language interface to email,* In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* , pp. 582–589. ACM Press/Addison-Wesley Publishing.

Wang, D., Li, J., Zhang, J. and Dai, G. (2008) A pen and speech-based storytelling system for Chinese children. *Comput. Human Behav.*, 24, 2507–2519.

Wang, Q. and Nass, C. (2005) Less visible and wireless: two experiments on the effects of microphone type on users' performance and perception. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 809–818. ACM, New York, NY, USA.

Weinschenk, S. and Barker, D. T. (2000) *Designing Effective Speech Interfaces* (Vol. 1). Wiley, New York.

Wilke, J., McInnes, F., Jack, M. A. and Littlewood, P. (2007) Hidden menu options in automated human – computer telephone dialogues: dissonance in the user's mental model. *Behaviour & Information Technology*, 26, 517–534.

Wilkie, J., Jack, M. A. and Littlewood, P. J. (2005) System-initiated digressive proposals in automated human–computer telephone dialogues: the use of contrasting politeness strategies. *Int. J. Hum. Comput. Stud.*, 62, 41–71.

Wolff, S. and Brechmann, A. (2015) Carrot and stick 2.0: the benefits of natural and motivational prosody in computer-assisted learning. *Comput. Human Behav.*, 43, 76–84.

Wolters, M., Georgila, K., Moore, J. D., Logie, R. H., MacPherson, S. E. and Watson, M. (2009) Reducing working memory load in spoken dialogue systems. *Interact. Comput.*, 21, 276–287.

World Health Organization. (2001) *International Classification of Functioning, Disability and Health: ICF*.

Yankelovich, N., Levow, G.-A. and Marx, M. (1995) Designing speech acts: issues in speech user interfaces. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 369–376. ACM Press/Addison-Wesley Publishing Co, New York, NY, USA.

Downloaded from https://academic.oup.com/iwc/article/31/4/349/5560013 by Georgia Institute of Technology user on 23 September 2025