Gaze and Speech in Multimodal Human-Computer Interaction: A Scoping Review

Anam Ahmad Khan
Industrial Design
KAIST
Daejeon, Republic of Korea
anam.khan@kaist.ac.kr

Florian Weidner
Glasgow University
Glasgow, United Kingdom
fweidner@sigchi.org

Jungwoo Rhee
Industrial Design
KAIST
Daejeon, Republic of Korea
jwoorhee@kaist.ac.kr

Yasmeen Abdrabou
Human-Centered Technologies for Learning
Technical University of Munich
München, Germany
yasmeen.e.mahmoud@gmail.com

Andrea Bianchi
Industrial Design
KAIST
Daejeon, Republic of Korea
andrea.whites@gmail.com

Eduardo Velloso
School of Computer Science
The University of Sydney
Sydney, New South Wales, Australia
eduardo.velloso@sydney.edu.au

Hans Gellersen
Lancaster University
Lancaster, United Kingdom
Aarhus University
Aarhus, Denmark
h.gellersen@lancaster.ac.uk

Joshua Newn
School of Computing Technologies
RMIT University
Melbourne, VIC, Australia
joshua.newn@rmit.edu.au

Abstract

Multimodal interaction has long promised to make interfaces more intuitive and effective by combining complementary inputs. Among these, gaze and speech form a compelling pairing: gaze provides rapid spatial grounding, while speech conveys rich semantic information. Together, they offer rich cues for understanding user behaviour and intent. Yet despite decades of exploration, the research remains fragmented, making this synthesis timely as these inputs mature and are integrated into consumer-ready devices. This scoping review examined 103 studies published between 1991 and 2025, organised into *explicit*, where users intentionally provide gaze and speech, and *implicit*, where systems leverage users' natural behaviours to support interaction. Across both, we identified recurring ways for combining gaze and speech to resolve ambiguity, ground references, and support adaptivity. We contribute a synthesis of research on their combined use while highlighting challenges of temporal alignment, fusion and privacy, offering guidance for future research toward richer multimodal human-computer interaction.

CCS Concepts

• **General and reference → Surveys and overviews**; • **Human-centered computing → Interaction techniques**; **Interaction devices**.

ACM ISBN 979-8-4007-2278-3/26/04
https://doi.org/10.1145/3772318.3791662

Keywords

Gaze, Speech, Multimodal Interaction, Human–Computer Interaction, Scoping Review

ACM Reference Format:
Anam Ahmad Khan, Florian Weidner, Jungwoo Rhee, Yasmeen Abdrabou, Andrea Bianchi, Eduardo Velloso, Hans Gellersen, and Joshua Newn. 2026. Gaze and Speech in Multimodal Human-Computer Interaction: A Scoping Review. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 24 pages. https://doi.org/10.1145/3772318.3791662

1 Introduction

In 1980, Richard Bolt's seminal work "Put-That-There" demonstrated one of the earliest and most well-known examples of multimodal interaction [19]. At a time when human-computer interaction was still primarily reliant on keyboards, the paper showed that combining speech with simultaneous pointing could resolve ambiguity in natural language commands. The idea that complementary input streams can make interaction more natural and efficient was ahead of its time and has remained a cornerstone of multimodal interaction in HCI. Since then, ongoing technological advances have steadily improved the availability and robustness of input modalities at scale, allowing researchers to continue exploring novel ways to use them in concert while addressing their inherent limitations.

We present a scoping review of 103 studies published between 1991 and 2025 that examine the combined use of gaze and speech in HCI—a compelling pairing that continues to show promise. Gaze input offers a continuous channel into attention and cognitive states [39, 54, 73, 142], while speech input provides an expressive channel for conveying meaning [31]. When combined, they help overcome each other's limitations, for example: gaze grounds and disambiguates speech, which can be noisy or spatially underspecified [128], while speech clarifies the intent behind gaze, which can be imprecise and prone to the 'Midas touch' problem [40, 53]. Together, these complementary channels support interactions with systems that more closely resemble natural human communication than either modality alone.

Over the past three decades, gaze and speech pairing has been investigated across diverse applications—from spatial command execution [49, 56, 146, 159] to user state inference and adaptive interfaces [64, 104, 167]. Yet research remains fragmented, partly because opportunities to study gaze and speech technologies have advanced independently, and partly because these modalities have been applied across domains, creating challenges in identifying overarching patterns and opportunities for new research directions. This review consolidates the literature to provide a unified perspective on their combined use, addressing the question: *How is the combination of gaze and speech used for interaction?*

The timeliness of this review is clear: consumer-ready devices that support both gaze and speech—such as XR headsets and smart glasses—have become increasingly accessible, and through advances in sensing, processing, and computing, can now easily support interaction in real-time in everyday settings. Gaze, once requiring controlled environments, can now be captured reliably in natural settings, while speech interfaces have been transformed by progress in natural language processing and large language models capable of contextual and semantic interpretation [2, 118].

Although gesture-based interaction also remains central in many XR systems (e.g., Apple Vision Pro), this review focuses on gaze and speech as a distinct and complementary pairing. Conceptually, these modalities occupy opposite ends of the interaction spectrum: gaze is fast, continuous, and low-effort but semantically limited, whereas speech is slower and more deliberate but highly expressive, with gestures typically lying between these extremes.

By focusing on gaze and speech as opposing yet complementary modalities, this review captures integration strategies, spanning explicit command-based interaction and implicit, context-driven inference. Practically, gaze and speech also form a hands-free input combination that is particularly well suited to head-worn devices, motivating a dedicated synthesis of how these modalities are combined for interaction. In venues across HCI, including CHI, we are beginning to see very promising implementations that combine gaze and speech for practical everyday use—for example, GazePointAR [75] for pronoun disambiguation in AR and G-VOILA [160] for everyday information querying. As these technologies become increasingly mainstream, it is therefore crucial to understand how gaze and speech can be combined to support effective interaction. This review systematically synthesises existing systems to reveal common multimodal interaction patterns, characterise their use across tasks and domains, and highlight remaining challenges and opportunities for future research.

We framed our synthesis according to the commonly observed forms of interaction: Explicit Gaze and Speech Interaction, where users intentionally provide gaze and speech as inputs, and Implicit Gaze and Speech Interaction, where systems leverage users' natural behaviour to support interaction. Figure 1 illustrates representative examples of these two forms of gaze and speech interaction. For

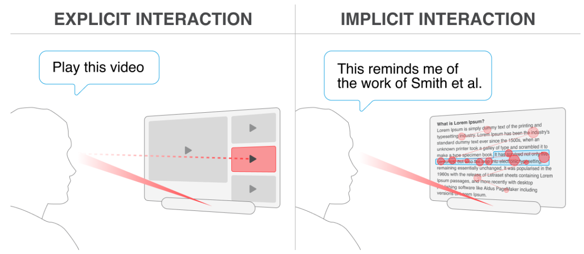

Figure 1: Examples of gaze and speech interaction. In explicit interaction, the user looks at a target and issues a direct speech command (e.g. "play this video"). In implicit interaction, gaze and speech are used as contextual cues, allowing the system to infer intent without an explicit command.

each, we apply two analytical lenses: the first examines how gaze and speech are fundamentally used and combined, and the second explores how this combination is applied across tasks and application domains in the literature. Our synthesis highlights three key insights. First, gaze and speech are leveraged through recurring patterns, behavioural feature representations, and integration strategies that reduce ambiguity and enable richer, hands-free interaction. Second, their complementarity—gaze for spatial precision and speech for semantic richness—is central in positioning them as an expressive and reliable multimodal input combination. Third, across the research landscape of gaze and speech, key challenges persist, including temporal alignment, generalisable pipelines, ethical and privacy concerns, evaluation beyond controlled settings, and the need to support diverse users and contexts.

This review advances the field through three key contributions by providing the first comprehensive synthesis focused specifically on the pairing of gaze and speech—two modalities whose combined use is becoming increasingly central in emerging technologies. While prior surveys have examined multimodal interaction [38], or reviewed gaze- or speech-only [e.g. 31, 134] systems, these treat the modalities independently and do not analyse gaze and speech as a unified input channel. Many benefits of this pairing—such as grounding, ambiguity reduction, and clearer intent signalling—and associated challenges emerge only when gaze and speech are considered together. To address these gaps, our review provides three contributions. First, we present a detailed analysis of recurring multimodal input combinations, integration strategies, and feature representations unique to gaze and speech, comparing patterns across explicit and implicit interaction paradigms. Second, we map the tasks and application domains that utilise gaze and speech, illustrating how the pairing facilitates control, grounding, and user-state inference. Third, we outline design challenges and research opportunities required to develop more robust and context-aware multimodal systems. Together, these contributions consolidate the literature and highlight the potential of combining gaze and speech to drive advances in multimodal interaction design.

2 Related Work

Researchers have long explored multimodal interaction in HCI, investigating the combination of multiple input channels for more effective interaction with computerised systems. Foundational works [e.g. 91, 99] established core principles and a design space for multimodal fusion, and focused on conceptual modality pairings rather than empirical systems, without reviewing specific input combinations such as gaze and speech. Later multimodal interaction surveys [e.g. 38, 55, 71, 92, 119] favour breadth—covering challenges, applications, and general fusion strategies across a wide range of potential pairings—and hence have not examined specific modality combinations in depth. For instance, Jaimes and Sebe [55] provide a vision-centred overview treating gaze and speech as independent channels, while XR-focused reviews highlight challenges for individual modalities (e.g. gesture, speech, gaze) without addressing how particular combinations function in interaction [71, 92, 119].

Several reviews that focus on gaze input acknowledge the combination of gaze with speech, including 2D displays [134], handheld devices [76], medical applications [18], and social collaboration [126], but do not examine the pairing in depth. XR-focused gaze reviews likewise mention gaze and speech among many multimodal combinations, but do not analyse how the modalities can be used together [108]. Speech-focused reviews [31] discuss long-standing challenges in speech technologies, but do not consider how gaze can support interaction by, for instance, grounding or clarifying spoken commands. Yet many of the core difficulties and opportunities emerge only when the two are used together—such as timing, fusion, and decisions about how each modality should contribute to the task. These issues are absent from both gaze- and speech-only surveys [e.g. 31, 108, 134], highlighting the need for a dedicated review that examines gaze and speech as a coordinated multimodal input. Recently, researchers have begun to examine other modality pairings in depth, but typically within specific environments, such as gesture and speech in augmented reality [e.g. 8]—further highlighting the value of modality pair-focused review.

Beyond formal reviews, some individual works that explore gaze and speech only mention prior studies briefly within their related-work sections [e.g. 64, 83], offering brief and often narrow summaries rather than in-depth analysis. Across the literature, we see repeated indications that gaze and speech complement one another. These works signal the value of examining the pairing more closely, motivating our systematic review to characterise how they are combined for interaction. Hence, this review addresses the gap by providing the first systematic synthesis focused specifically on how gaze and speech are combined for interaction across diverse environments and domains, consolidating fragmented findings and identifying the design patterns, computational approaches, task domains, and research opportunities unique to this pairing.

3 Method

3.1 Selection and Screening

Our methodology followed structured steps to ensure coverage and rigour: developing a comprehensive search query, applying inclusion and exclusion criteria, and screening articles through a multi-stage process. Figure 2 shows the adapted PRISMA [100] flowchart summarising the process.

We applied an initial keyword-based search on the ACM Digital Library (ACM-DL) and IEEE Xplore (IEEEX), Scopus, and Web of Science libraries, which capture the majority of computing publications and provide broad coverage across multiple disciplines. Then, we conducted a search query to match keywords against titles, abstracts, and author keywords. The Boolean query string was initially formulated using the terms “gaze” AND “speech” AND “interaction”. However, we observed that some key publications from venues (e.g. CHI and ICMI) were missing from the search results. We therefore expanded our query with a broader set of gaze-, speech-, and interaction-related keywords, which were refined by incorporating relevant keywords identified in the first 50 papers returned by the ACM-DL. The full keyword list is provided below, and the complete database queries appear in the Appendix.

> (“gaze” OR “eye”) AND (“speech” OR “voice” OR “audio” OR “vocal”) AND (“interact*” OR “communicat*” OR “input” OR “technique” OR “model” OR “system” OR “interface”)

The initial search query was conducted on September 29, 2024. The search results from all four databases were combined, resulting in 11504 records. We then removed duplicates by matching titles, authors, and publication years using a custom Python script, yielding 8,075 unique articles. Screening followed a two-stage process. In the first stage, records were assessed by title, abstract, and author keywords, and in the second, full-text reviews were performed alongside data extraction, guided by the following inclusion criteria (IC) and exclusion criteria (EC).

- **IC1** Studies use both gaze and speech as input modalities. *Rationale: Our focus is on multimodal interaction, where both inputs jointly influence system behaviour.*
- **IC2** Studies focus on single-user human–computer interaction, where a user interacts with a computer. *Rationale: Our focus is on how systems interpret gaze and speech for interaction, not on human–human interaction.*
- **EC1** Studies on computer-mediated human–human interaction rather than direct system input. *Rationale: These focus on social dynamics, not system design.*
- **EC2** Do not use gaze and speech as inputs from humans to facilitate interaction (e.g. papers focused on synthesising gaze and speech for interaction). *Rationale: Output or synthesis work does not inform multimodal system input.*
- **EC3** Are not journal articles or included in the main proceedings, such as adjuncts, posters, extended abstracts, companion proceedings, workshop proposals or editorials. *Rationale: To ensure quality of reviewed work.*
- **EC4** Are not written in the English language. *Rationale: Due to screening feasibility.*

During the first stage, the first author screened 8075 articles, resulting in 112 included, 20 marked unclear, and the remainder excluded. The unclear set was then discussed among authors, leading to 5 inclusions and 15 exclusions. The first round of screening resulted in 117 articles. To supplement the keyword search, we performed backward chaining, reviewing references in the 117 papers, which yielded 7 additional papers.

In the second stage, full-text screening and data extraction were performed. Four authors participated: 65% of articles were reviewed by the first author, 9% by the second, and the remaining 26% were split between the third and fourth authors. Each author assessed eligibility using the same inclusion and exclusion criteria as in the

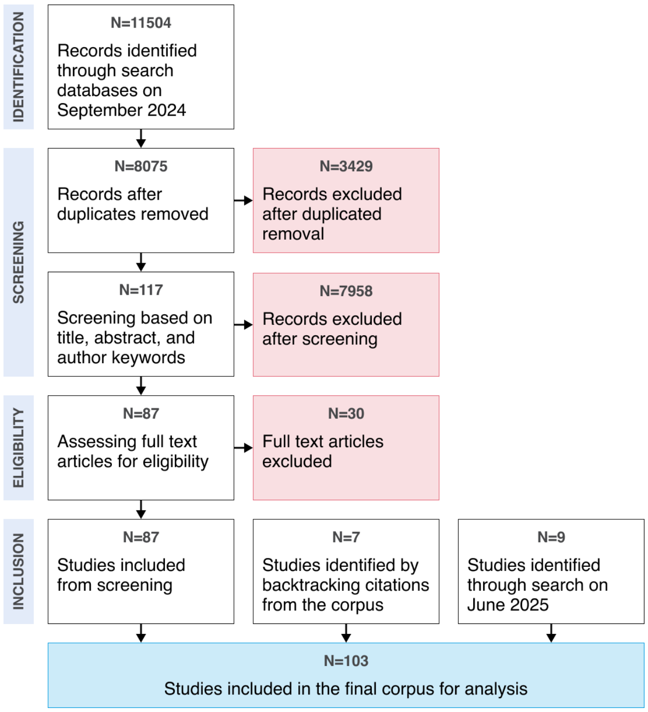

Figure 2: Adapted PRISMA [100] flowchart illustrating the process of selecting, screening, and reviewing papers for inclusion in the final corpus for analysis.

first stage, with any disagreements resolved through discussion. Data were extracted from all papers that met the criteria. This second screening phase led to the removal of 30 additional articles, leaving a set of 87 papers. Finally, to include the most recent work, we conducted an updated search on June 29, 2025, targeting papers published between September 2024 and June 2025. This yielded 1136 unique papers across the four databases. After applying the same inclusion/exclusion criteria, 9 additional papers were added. In total, the final corpus comprises 103 studies across 77 venues. The most frequent venues were the ACM International Conference on Multimodal Interaction (ICMI, 8), the ACM Conference on Human Factors in Computing Systems (CHI, 6) and the ACM Symposium on Eye Tracking Research & Applications (ETRA, 5).

3.2 Analysis and Review Structure

For analysis, we began by creating a structured spreadsheet and extracted key information from each paper, including whether gaze and speech cues were treated as *intentional inputs* or interpreted *implicitly* as behavioural traces, the specific gaze and speech *features* considered, the *interaction tasks* supported, the *type of interactive entity* involved (e.g. system or agent), how gaze and speech were *integrated* (sequentially or in parallel) and the *computational methods* used to interpret the multimodal signals. Building on the data extraction, we coded the studies by interaction type (explicit vs. implicit), the combination and integration strategy of gaze and speech, and their task and application domains. Coding categories were iteratively refined through pilot coding and discussion among the authors. The first author independently coded all papers, while the fourth author re-coded a random 25% subset to assess reliability. Inter-coder reliability was high (Cohen's $\kappa > 0.80$; [84]), and any discrepancies were discussed with all authors until full agreement was reached [23].

Through our inductive analysis, we identified two distinct research categories that structure this scoping review and capture fundamentally different ways in which gaze and speech are operationalised in interactive systems.

While the literature uses 'explicit' and 'implicit' to describe both interaction approaches and signal properties, we use these terms solely to distinguish whether the system responds to intentional actions or infers meaning from natural behaviour.

- Explicit Gaze and Speech Interaction (N=55) frames user actions as *intentional control signals*. Here, gaze and speech are combined in the interaction loop as forms of direct manipulation, enabling users to deliberately issue commands or manipulate systems.
- Implicit Gaze and Speech Interaction (N=48) frames gaze and speech as *behavioural cues* to be monitored and interpreted. In this case, systems model natural gaze or speech patterns to provide adaptive, context-aware support without requiring deliberate multimodal input from the user.

This explicit–implicit distinction follows prior gaze-interaction surveys [59, 108], reflecting established practice rather than a new taxonomy. Figure 1 illustrates how the two categories differ.

Although we organise the literature into these two categories, we emphasise that explicit and implicit interaction form a continuum rather than a strict dichotomy [133]. Users often interleave intentional and spontaneous behaviours, making the distinction between the two interactions inherently blurry [130, 133]. A paper is classified as explicit if the underlying system requires users to intentionally coordinate gaze and speech to perform an action (e.g. looking at a target and issuing a command), and implicit when it infers meaning, attention, or intent from spontaneous gaze or speech behaviour without requiring the user to align them intentionally. We further acknowledge that some task categories—such as Reference Resolution (§6.2.1) and Disambiguating Spoken Commands (§5.2.3)—can be realised through either explicit or implicit interaction. In these cases, we classify papers based on the interaction logic implemented by the system, reflecting how the modalities are operationalised rather than whether the user behaviour could be intentional.

After categorising papers into either explicit and implicit to provide top-level framing for our scoping review and to shape its overall structure, we then analyse them through two lenses. The first lens clarifies how gaze and speech are *fundamentally used* and combined—their functional roles, feature representations, and integration strategies—allowing us to compare and advance underlying approaches across studies. The second lens shows how the combination is *applied* in practice, highlighting the value of their combination in different contexts. Together, the two lenses separate the mechanisms that enable gaze and speech to work together from their practical use, offering a coherent way to synthesise otherwise fragmented research.

Before discussing the explicit and implicit interaction categories, we first outline the basic properties and signal features of gaze and speech in the following section. Establishing this foundation clarifies the individual contributions of each modality and sets the stage for understanding how they combine in both explicit and implicit interaction.

4 Gaze and Speech as Input Modalities

Understanding how gaze and speech function as input modalities requires examining both their high-level properties and their underlying signal features, which together form the foundation for multimodal interaction.

Gaze and speech offer complementary strengths that support human-computer interaction. Gaze provides fast, spatially precise input, making it effective for identifying targets [39, 53], but it is inherently low in expressiveness, as the eyes are primarily perceptual rather than communicative [53]. This makes it difficult to distinguish intentional input from exploratory visual scanning, leading to unintended activations—the so-called "Midas Touch" problem [53]. Speech, by contrast, provides flexible and semantically rich input, enabling natural language commands and data entry [89], but lacks spatial precision, especially when deictic terms like "this" or "there" require contextual grounding [70].

The complementarity of gaze and speech lies in combining their strengths: gaze can supply spatial grounding to resolve ambiguities in speech, while speech can express explicit intent and semantic detail that gaze alone cannot. For example, selecting an object with gaze while issuing a spoken command such as "delete" allows the system to infer user intent, reducing ambiguity and supporting more natural interaction than either modality alone [99].

This complementarity extends to their underlying signal features, which provide the building blocks for interpreting user state and behaviour. Speech features are commonly divided into *acoustic* and *linguistic* dimensions. Acoustic features describe the physical waveform and include *spectral* (energy distribution across frequencies), *prosodic* (rhythm, intonation, stress), and *voice quality* (clarity, stability, tension) components. Linguistic features are derived from content, capturing lexical choice, syntax, and semantics. Complementing this, gaze features such as *fixations* and *saccades* provide non-verbal insight into visual attention and cognitive processing [45, 120]. Fixations reflect stable focus for information processing, while saccades enable rapid exploration of the scene. Together, these features form reliable proxies for attention and intent.

By combining these high-level properties and signal characteristics, gaze and speech serve as rich, complementary channels for multimodal interaction, forming the foundation for both explicit and implicit uses of these modalities.

5 Explicit Gaze and Speech Interaction

Explicit gaze and speech interactions involve users deliberately directing their gaze and issuing spoken commands to explicitly perform *intentional* actions. This category depends on conscious attention and articulated intent, making it well-suited for tasks requiring spatial precision, semantic clarity, and hands-free control.

While this modality pairing has been widely explored for explicit interaction (N=55), the design choices and functional components of such systems remain inconsistently defined. To support a clearer understanding of research on explicit interaction, we organise this section around two analytical lenses: (1) *How gaze and speech are fundamentally used for explicit interaction* and (2) *How explicit gaze and speech are applied across task and application domains?.*

5.1 How gaze and speech are fundamentally used for explicit interaction?

Explicit interaction depends on the functional roles of gaze and speech and the ways in which these modalities are coordinated to form deliberate multimodal input. This subsection examines the functional components of each modality, how these components pair into recurring patterns, and the integration strategies that underpin these interaction patterns.

Gaze serves primarily two functions:

- **Pointing**: Gaze is commonly used for pointing—that is, referencing objects, locations, or interface elements by simply looking at them. It offers a fast and natural means of spatial referencing, particularly in contexts where manual input is not feasible (e.g. users with motor impairments or in hands-busy settings) [93, 154, 173], by leveraging the natural correlation between visual attention and user intent [53].
- **Confirmation:** In addition to pointing, gaze can also confirm user intent during interaction. Because the eye is primarily a perceptual organ, users often scan the interface without intending to select anything. However, if every fixation is interpreted as input, this can lead to unintended activations—known as the 'Midas touch' problem [53]. To address this, a separate confirmation mechanism is typically required to ensure that gaze-based selections are deliberate. Dwell-based selection—where a user must fixate on a target for a fixed duration—is the most widely used strategy to signal intentionality and reduce false positives [125, 149].

Similarly, speech affords two functional roles depending on its purpose in an interaction:

- **Command**: The most common use of speech is as an input mechanism to convey detailed user intentions, particularly action-oriented commands such as "delete" or "rotate". These commands express user intent and must be semantically interpreted to trigger appropriate system responses [89].
- **Data**: Speech can also serve as a source of content or information, allowing users to dictate content (e.g. filling text fields or entering numbers). Unlike commands, speech as data does not require semantic interpretation to infer intent, as the purpose of speech here is not to execute an action, but simply to provide data.

5.1.1 *Combinations of Gaze and Speech Input.*
Building on the functional roles, we identify three patterns in the literature for combining gaze and speech inputs. Figure 3 illustrates these explicit combinations, which leverage the strengths of each modality while addressing the limitations of the other. Each pattern is described in detail below.

Gaze to Point, Speech as Command. In this combination, gaze specifies the object or region of interest, while spoken input communicates the desired action. Speech commands can be simple triggers

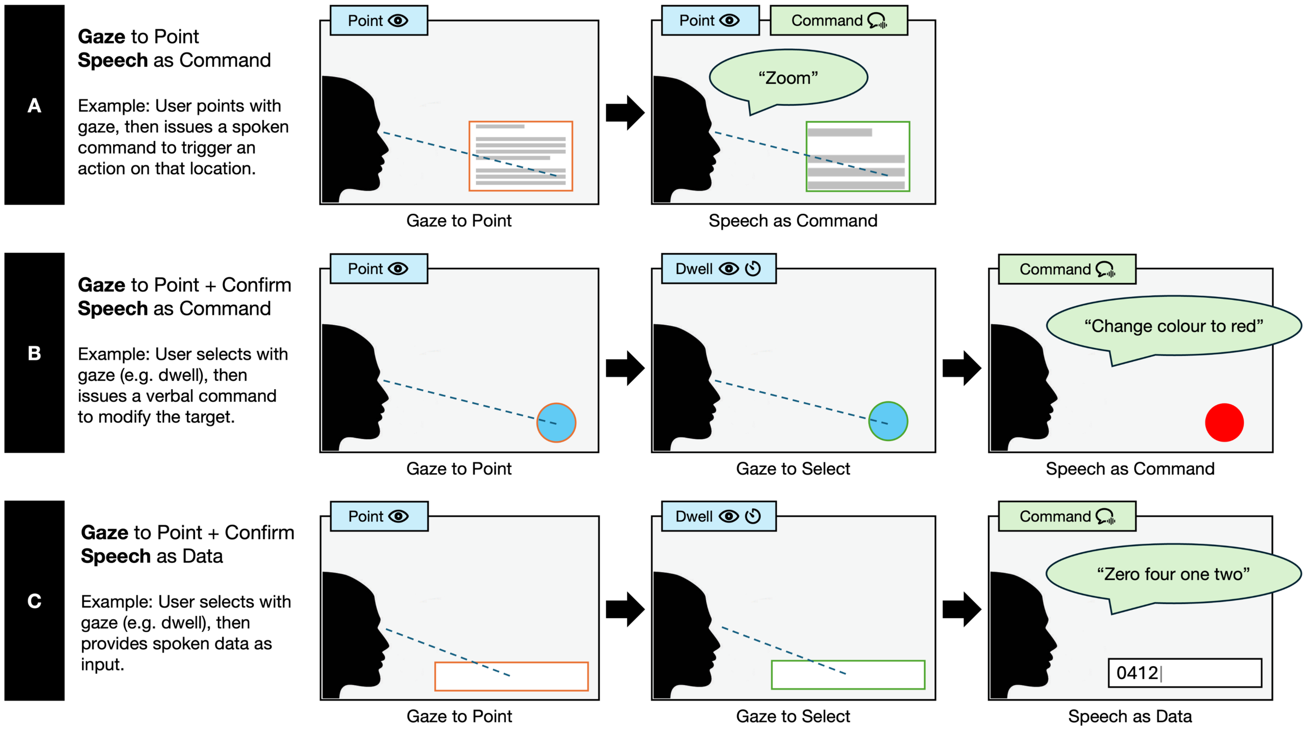

Figure 3: Three combinations of patterns of explicit gaze and speech: (A) gaze specifies a target while speech issues a command, (B) gaze identifies and confirms a target followed by a spoken command, and (C) gaze identifies and confirms a target followed by a spoken data input.

to simply select (e.g. "select") [11, 49, 132] or richer instructions (e.g. "rotate", "delete") [48, 60, 152, 161, 163, 175], but in both cases, they operate on the object the user is looking at when the command is issued. For example, a user points their gaze at a region of interest and says "Zoom In", with gaze indicating the target and speech conveying the action.

Gaze to Point + Confirm, Speech as Command. In this combination, gaze first indicates the object of interest and then confirms the selection, typically through a dwell, before any spoken input is processed. Unlike the previous combination, users do not need to keep fixating on the target during speech, as the object remains selected, allowing them to look elsewhere without losing the selection. The spoken input then communicates the action to perform. For example, a user may fixate on an object with a dwell to select it, and then issue the command "measure" to obtain its size [149].

Gaze to Point + Confirm, Speech as Data. Similar to the previous combination, gaze is used to point and confirm a target, but here the interaction leverages the richness of speech to provide data beyond commands. Speech is not interpreted as an action but instead populates the gaze-selected object with content. For example, as shown in Figure 3, a user can enter a sequence of numbers into a gaze-selected text field using speech [148].

5.1.2 Integration Strategies. To illustrate how combinations of gaze and speech map to integration strategies, we adopt the multimodal framework of Nigay and Coutaz [91] (see Appendix), which distinguishes integration by temporal use (sequential vs. parallel) and fusion method (independent vs. combined). Because our focus is on combining gaze and speech for interaction, independent use of modalities is not relevant; all pairings fall within the combined quadrants. We therefore analyse integration primarily in terms of temporal usage, contrasting *sequential* and *parallel* strategies.

In parallel integration, gaze and speech are processed within the same temporal window as a single fused action. This strategy aligns with combinations where gaze provides a real-time reference during speech (e.g. Gaze to Point, Speech as Command). Parallel designs enable a more fluid, conversational interaction style but assume that the referent is whatever the user is fixating on during the utterance, making them vulnerable to drift or timing mismatches. For example, Kaur et al. [60] found that the fixation most strongly associated with a spoken referent often precedes the utterance by about 630 ms, highlighting the need for temporal alignment models that account for gaze lead time.

In sequential integration, gaze and speech are processed in distinct stages: gaze first identifies and confirms the target, and only then is speech interpreted. This design underpins combinations that require explicit confirmation before command execution (e.g. Gaze to Point + Confirm, Speech as Command) or data input (e.g. Gaze to Point + Confirm, Speech as Data). By anchoring the system to a known referent before interpreting speech, sequential designs

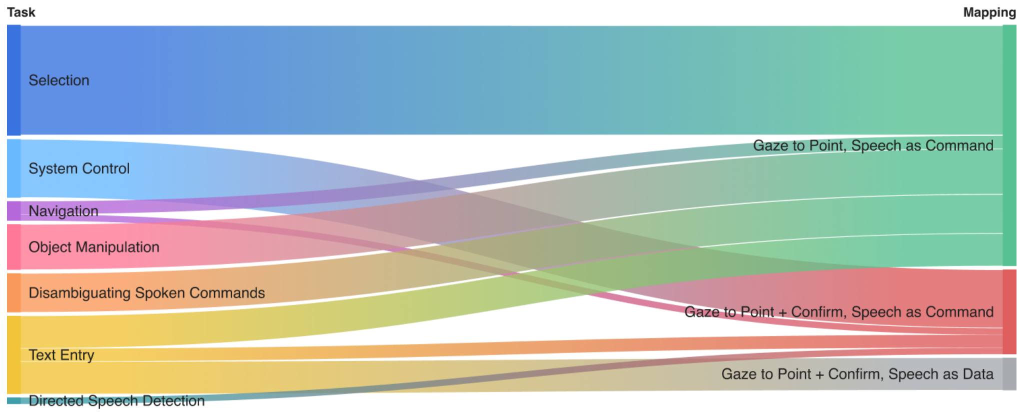

Figure 4: Sankey diagram illustrating the distribution of gaze and speech combination across different interaction tasks. The thickness of each flow corresponds to the number of studies employing that mapping for the given task.

reduce errors and allow users to move their gaze freely after selection, though the extra step may reduce efficiency in fast-paced interactions. Ultimately, choosing between sequential and parallel integration depends on task demands: sequential approaches suit error-sensitive contexts by minimising unintended actions, while parallel approaches favour speed and conversational flow.

5.2 How is explicit gaze and speech interaction applied across task and application domains?

Having identified how gaze and speech are fundamentally used for explicit interaction, we can now map these insights onto their application across various tasks and domains. We interpret these applications through different gaze and speech pairing combinations, highlighting how task demands shape the assignment of modality roles. Figure 4 presents a mapping between tasks with different explicit gaze and speech combinations, Table 1 below shows the distribution of studies following sequential or parallel integration of input, and Table 2 shows the distribution of studies across task categories and their sub-tasks.

Table 1: Number of studies using parallel and sequential integration across task categories for explicit gaze and speech interaction.

Task Category	Sequential	Parallel
Selection		17
System Control	10	
Disambiguating Spoken Commands		5
Navigation	1	2
Object Manipulation		7
Text Entry	7	5
Directed Speech Detection	1	

5.2.1 Selection and Control. The first category covers tasks where gaze establishes focus on an object or interface element, while speech provides the accompanying trigger or direct system actions.

Selection. Selection is a fundamental interaction task where users focus on an object of interest, and gaze has long been employed for this purpose [39, 142]. Traditional dwell-based techniques are effective but can cause fatigue with prolonged use, motivating the integration of gaze with speech to create more efficient multimodal selection methods. Seventeen studies explored this space using the *Gaze to Point, Speech as Command* combination, in which gaze anchors the target while speech specifies the action. Within this design, systems differ in how speech is treated. One group uses speech as a fixed verbal trigger (e.g. “select”, “click”) to confirm the target [1, 10, 12, 47, 56, 103, 107, 110, 124, 158, 159, 172]. In contrast, another group expands the role of speech, treating it as semantic input that carries information about the object itself, such as its colour or shape [85, 86, 170, 171], or the specific action to be performed, such as “left click” versus “right click” [78]. Together, both variants follow the same parallel integration strategy, but the choice between fixed triggers and semantic speech input shapes whether the design prioritises simplicity and speed or expressiveness and disambiguation.

System Control. Ten studies explored gaze and speech for system control across augmented reality (AR), physical environments, and 2D interfaces. System control interaction follows the *Gaze to Point + Confirm, Speech as Command* combination, where gaze is first used to establish the target spatially, while speech specifies the command to be executed on that target. In AR settings, this combination supports immersive manipulation of virtual objects and media. Users could fixate on a display or object to select it, then issue fixed verbal commands such as “rotate”, “enlarge” [93], or “Show CNN” [155] to control its state or content. The same principle extends into physical environments, where gaze operates as a natural pointer to devices or components located at a distance or in constrained spaces, while speech executes operations–from switching on an appliance [166], to disabling an LED [125], to measuring micro-objects with high precision [149]. Similarly, in 2D graphical interfaces, gaze-driven target selection paired with verbal commands supported applications ranging from document retrieval [151] and media inspection [42, 145] to hands-free digital drawing [154] and mode switching in multimodal speech recognisers [144].

Table 2: Task categories and sub-tasks of explicit gaze and speech interactions, showing the distribution of studies across selection and control, navigation and manipulation, and language input and editing.

Task Category	Sub-Task (# studies)	Studies
Selection & Control	Selection (17)	[1], [10], [12], [47], [56], [78], [85], [86], [103], [107], [110], [124], [158], [159], [170], [171], [172]
	System Control (10)	[42], [93], [125], [144], [145], [149], [151], [154], [155], [166]
Navigation & Manipulation	Navigation (3)	[5], [26], [164]
	Object Manipulation (7)	[28], [46], [48], [60], [152], [161], [175]
Language Input & Editing	Disambiguating Spoken Commands (5)	[70], [75], [83], [157], [163]
	Text Entry (12)	[11], [13], [49], [82], [97], [132], [136], [146], [147], [148], [156], [173]
	Directed Speech Detection (1)	[87]

A consistent advantage of system control applications is the reduction of speech recognition errors through contextual disambiguation: once gaze identifies the target, the system constrains expected commands, mitigating noise and ambiguity. This synergy improves robustness and user experience, reinforcing the value of gaze and speech integration for complex control tasks.

5.2.2 Navigation and Manipulation. This second category focuses on tasks where gaze provides spatial reference for movement or transformation, while speech directs how the action is carried out.

Navigation. Three studies have explored gaze and speech for hands-free navigation in both virtual and physical spaces. Two studies followed the *Gaze to Point, Speech as Command* combination, where gaze specifies the destination with speech issuing the navigation command [5, 26]. For instance, users could gaze at a target location to initiate teleportation in VR ("take me there") or control wheelchair direction with spoken cues like "start" and "stop". In contrast, Wu and Tanaka [164] employed the *Gaze to Point + Confirm, Speech as Command* combination, using sequential integration where gaze first confirms the navigation target (e.g. dwelling on a person) before speech triggers the action ("follow this person"). This design supports more deliberate target selection and assists decision-making in physical environments.

Object Manipulation. Object manipulation is a key interaction task that has been explored using the combination of gaze and speech, particularly for translation and scaling operations. This task consistently follows the *Gaze to Point + Speech as Command* combination, where gaze selects the object or destination and speech specifies the action. Translation refers to the task of changing the position of a virtual object. Five studies have explored the combination of gaze and speech for object translation, where gaze typically specifies the intended destination, while speech triggers the object movement [48, 60, 152, 161, 175]. Most studies relied on fixed verbal commands (e.g. "Move" or "Up") to initiate movement [60, 152, 161, 175], though some also experimented with non-verbal acoustic cues, such as a buzzing sound, offering quieter or alternative control modes [48]. Overall, while gaze- and speech-based translation tends to be slower and less accurate than speech-only [161] or touch-based interaction [152], it is often reported as more immersive and engaging for users. Scaling has been explored in two studies, where users perform zoom operations by gazing at a region and issuing commands such as "zoom in" or "zoom out" [28, 46], demonstrating accessibility benefits in hands-free contexts.

5.2.3 Language Input and Editing. This third category includes tasks where gaze supports entry, correction, and disambiguation of linguistic content, and speech contributes semantic commands or replacement data.

Disambiguating Spoken Commands. Gaze helps disambiguate spoken commands by providing spatial context that speech alone often lacks. For example, phrases like "move down" or deictic expressions such as "this" or "that" can be ambiguous without shared spatial reference [83]. To address this, systems adopt the *Gaze to Point, Speech as Command* pairing, where gaze anchors the referent while speech specifies the action. Some designs require users to fixate on an object while issuing the command, whereas others infer the referent from gaze behaviour such as direction or fixation points at the moment of speech [70, 75, 83, 157, 163]. In both cases, the combination of gaze and speech reduces ambiguity and enables more natural multimodal interaction.

Text Entry. Hands-free text entry is particularly valuable for accessibility and has been explored in eight studies that fall into two main approaches. The first follows the *Gaze to Point, Speech as Command* combination, where gaze selects characters on a virtual keyboard and speech simultaneously confirms input through fixed cues, either at a fine granularity (i.e., per character [11, 13]) or a coarser granularity (whole word, [49]). The second group aligns with the *Gaze to Point + Confirm, Speech as Data* combination, where gaze identifies and confirms the input region before speech provides the actual text [82, 97, 136, 147, 148]. For example, a user may fixate on a text field to select it and then dictate the intended input, which is directly transcribed into the chosen location [148].

Similarly, text correction has been investigated in four studies and shows the same division. Two studies use the *Gaze to Point, Speech as Command* combination, where users fixate on the erroneous word and issue a fixed verbal command (e.g. "fix") to confirm the selection, and then select replacements from a numbered list [132, 146]. Others adopt the *Gaze to Point + Confirm, Speech as Data* combination, where gaze identifies the error and speech provides the corrected content directly [156, 173]. For example, in [173], users select the word with their gaze and then speak the corrected form to substitute it. Across both text entry and correction, results consistently show that combining gaze for spatial precision with speech for confirmation or content reduces effort, increases efficiency, and improves robustness compared to unimodal input.

Directed Speech Detection. Gaze and Speech combination has also been used to address the challenge of distinguishing intentional voice commands from background speech. Instead of relying on wake words (e.g. "Hey Siri"), systems like *Look to Talk* [87] use gaze as an explicit intent signal: when users look at the interface, it enters an active listening state and speech is then processed as a command. This design follows the *Gaze to Point + Confirm, Speech as Command* combination—gaze confirms target first, followed by spoken command—and illustrates how gaze can operate not only as spatial input but also as a modality control mechanism within multimodal interfaces.

5.3 Takeaways

First, our synthesis shows that explicit gaze and speech combinations leverage the complementary strengths of the two modalities: gaze offers fast, precise spatial reference, while speech provides semantic content and intent. Together, they enable deliberate, hands-free actions across tasks from basic selection to complex control, typically through recurring input patterns (e.g. gaze to point, speech to command). Second, these combinations are implemented through either parallel integration, which supports fluid, conversational use but is sensitive to timing and gaze drift, or sequential integration, which improves reliability and allows gaze to shift during speech. Third, explicit systems tend to perform best in short, controlled tasks that require high accuracy. For designers, the synthesis highlights concrete input patterns linking modality roles and temporal fusion strategies, offering practical guidance for balancing speed, precision, and expressiveness in explicit multimodal systems.

6 Implicit Gaze and Speech Interaction

Implicit gaze and speech interactions involves systems passively interpreting users' natural gaze and speech behaviours to support tasks without requiring deliberate commands [133]. Here, users do not explicitly coordinate their gaze or speech [57]. Instead, interaction arises from spontaneous patterns of looking and speaking. Hence, unlike explicit interaction, which depends on conscious control, implicit interaction emphasises anticipation: systems infer cognitive states, attentional focus, or forthcoming actions from passive multimodal traces. This shifts the design paradigm from direct control toward adaptive systems that support interaction. To structure the research in this space, we apply two analytical lenses: (1) *How are gaze and speech fundamentally combined to enable implicit interactions?* and (2) *How is implicit gaze and speech applied across application domains?*

6.1 How are gaze and speech fundamentally combined to enable implicit interactions?

Designing implicit gaze and speech interaction requires shifting from direct control to inference: instead of responding to explicit commands, systems interpret behavioural traces to provide task-relevant support. At its core, this paradigm is typically modelled as a processing pipeline with two stages: a *representation* stage, where raw gaze and speech signals are transformed into structured features that capture behavioural or contextual patterns, and a *recognition* stage, where these features are interpreted by computational models to generate inferences that guide system behaviour. We structure this section around these two stages.

6.1.1 *Representation: Extracting Features from Gaze and Speech.*

Representation describes how raw gaze and speech data are transformed into structured features that capture behaviour and context. Our analysis reveals two broad approaches emerging across the literature: one analyses each modality independently, while the other analyses gaze and speech jointly, focusing on their synchrony across modalities.

Independent Feature Extraction. Independent feature extraction treats gaze and speech as separate channels of information, each offering a unique perspective on user behaviour. Rather than relying on cross-modal synchrony, this approach extracts what each modality can reveal independently. From gaze data, features are typically drawn at three levels. Fixations, measured by their count and duration, capture attentional focus and indicate which objects or regions the user finds important or cognitively demanding [4]. Saccades are characterised by their length, velocity, and duration and reflect attentional shifts and exploration strategies; long, fast saccades often signal scanning or searching behaviour, while shorter ones indicate local inspection [45, 58, 64]. Finally, scanpaths, which integrate sequences of fixations and saccades, provide higher-level structure, as illustrated in Figure 1.

These have been characterised using handcrafted metrics (e.g. path length) [96] or machine learning approaches for structured inference, such as detecting autism spectrum disorder [167]. While most of the 48 studies derive structured gaze features (e.g. fixations, saccades), 12 instead use raw low-level gaze signals (e.g. azimuth,

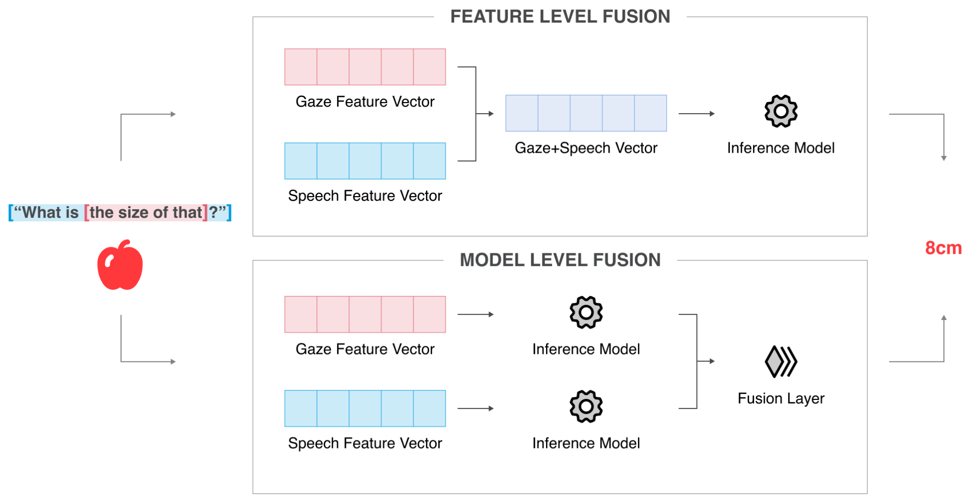

Figure 5: Feature-level versus model-level fusion of gaze and speech signals. In feature-level fusion, gaze and speech feature vectors are combined into a joint representation before inference. In model-level fusion, each modality is processed separately, and outputs are merged in a later fusion layer.

elevation) as continuous input to models, enabling fine-grained prediction but requiring heavier filtering [*e.g.* 44, 123].

Speech provides rich information from low-level acoustics to high-level semantics. At the signal level, acoustic features such as pitch (F0), rhythm, intensity, and jitter capture subtle cues about user state [4, 113] and intent [174]. Beyond these low-level cues, linguistic analysis offers a direct window into user goals. Syntactic features such as keyword analysis anchor utterances to tasks—for example, mapping "delete" to a deletion action or "start" to task initiation—though this approach is limited by predefined vocabularies [79]. Semantic analysis, in contrast, captures meaning beyond literal keywords, enabling flexible interpretation of varied expressions of intent [63]. For instance, "could you get rid of this?" can be recognised as equivalent to "delete". Together, linguistic features provide the 'what' of user intention, complementing the 'how' expressed through acoustic features.

Joint Feature Extraction. Unlike unimodal features, joint features are derived by analysing gaze and speech together, exploiting their synchrony to enable *joint interpretation.* Rather than treating modalities as parallel streams, joint features capture how gaze and speech mutually constrain each other in context—gaze narrowing possible referents and speech providing semantic specificity. Two main categories of joint features have emerged: *similarity-based* and *reference-informed*. Similarity-based features assess how closely spoken utterances align with gaze-fixated items. For instance, when multiple objects are being viewed at once, the semantic content of speech can be compared with the items under gaze, allowing the system to infer the intended referent [63, 111, 112, 114].

Reference-informed features instead use speech events as temporal anchors for gaze analysis. For instance, to map ambiguities in spoken commands, deictic expressions (e.g. "this") can define the window for examining fixation patterns, enabling systems to map ambiguous utterances to visual context [63]. Compared to analysing gaze and speech separately, joint feature extraction provides greater interpretive power. By modelling their synchrony, it reduces ambiguity, enriches user state inference, and enables more natural multimodal interaction.

6.1.2 *Recognition: Integrating and Interpreting Features.*

Recognition concerns how gaze and speech features are combined and interpreted to infer user states or intentions. This involves three key aspects: the *integration strategy*, which determines how multimodal features are brought together, the *computational model*, which interprets these features to generate inferences, and the *annotation and generalisability* process, which defines how data are labelled and how models are validated across participants and contexts to ensure reliability beyond the original training conditions.

Integration Strategies. Unlike explicit interaction, where gaze and speech are integrated at the input level through sequential or parallel timing, implicit interaction integrates them at the feature or model level. Here, synchrony is captured not through when the inputs occur, but through how their features align semantically and behaviourally. In inference tasks, gaze and speech features are typically integrated either at the feature- or model-level, as illustrated in Figure 5.

Feature-level integration is the most common approach, where features from both modalities are concatenated into a single joint vector and processed by one model [7, 63, 101, 112, 114]. It mirrors parallel integration in explicit interaction, treating modalities simultaneously and synchronously.

While efficient and capable of joint learning, this strategy assumes synchronous and semantically aligned features, which may not always reflect natural behaviour. In contrast, in the model-level integration, gaze and speech are processed independently by separate models, each producing task-specific inferences that are later combined through ensemble methods or decision rules[4, 167]. For example, Zhang [167] used an LSTM for gaze and a CNN for audio, then fused their output into an ensemble model for the detection of autism spectrum disorders. Although less common, model-level fusion offers robustness by allowing each modality to operate independently, at the cost of added complexity.

Although both strategies have proven effective for inference, most studies do not justify their choice. One exception is Alhargan et al. [4], who compared both strategies for affect recognition and showed that model-level fusion outperformed feature-level fusion. These findings highlight the need for more systematic evaluations of fusion strategies in implicit interaction.

Computational Models. The inference models used to interpret gaze and speech features vary widely depending on task requirements, data availability, and performance demands. When labelled data is limited, rule-based algorithms are often applied, using predefined thresholds or heuristics to interpret multimodal features [21, 104, 121, 160]. For instance, cheating might be inferred if a user's gaze remains off-screen while their speech intensity exceeds a set threshold [104]. These approaches are computationally lightweight and effective in constrained contexts but lack adaptability and generalisation to real-world variability.

To overcome these limitations, supervised machine learning models have been widely adopted for learning complex multimodal patterns. Support Vector Machines (SVMs), Logistic Regression, Naive Bayes, and XGBoost have all been applied to tasks such as affect recognition, cheating detection, and Automatic Speech Recognition (ASR) enhancement [4, 7, 33, 96, 112]. Among these, SVMs are the most frequently used due to their strong generalisation across tasks. Even though these models offer flexibility in handling linear and non-linear relationships within data, representing diverse input features—ranging from numerical to textual or structured data—and extending to different learning paradigms such as classification and regression, they often rely heavily on manual feature engineering [106, 131].

Recently, deep learning approaches have become prominent. LSTMs and CNNs are particularly effective in capturing the temporal dynamics of gaze and speech interaction and have been applied to tasks such as reference resolution [44], ASR enhancement [58] and detection of autism spectrum disorders [167]. By learning directly from raw data, these models reduce dependence on handcrafted features and exploit cross-modal synchrony more effectively. However, they require large amounts of labelled data, making them resource-intensive and less feasible in low-resource settings.

Despite these advances, model selection across the literature remains inconsistent and often poorly justified. Only a handful of studies explicitly compare approaches or explain their choice [63, 167]. This lack of transparency makes it difficult to identify best practices, underscoring the need for systematic evaluation of computational models and fusion strategies in implicit gaze and speech interaction.

Annotation and Generalisability of Computational Models. Annotation assigns labels to multimodal data to create structured training examples for inference models [16]. In implicit gaze and speech interaction, where behavioural cues map to latent states, annotation quality strongly constrains model generalisation. Across studies, annotation practices varied widely: most relied on manual annotation—for referent identification [63, 160], predicting affective states [67] or engagement events [7, 14]—while others used automated annotation [29] or task-defined labels embedded within experimental designs [69]. Manual annotation was generally performed by experimenters linking behavioural traces to user intentions [101, 160] or by deriving ground-truth labels from self-report questionnaires [139, 141]. Most studies involve only one or a small group of annotators [14, 121], with minimal reporting of coding reliability [17, 137, 143, 174].

In contrast, automated approaches relied on existing or algorithmically derived labels rather than human coding [44, 175]. Some studies reused public datasets—for ASD detection [81] or emotion recognition [96]—to train or validate multimodal models. Others generated labels through computational pipelines, such as aligning ASR timestamps with gaze fixations for referential grounding [44] or applying instance-segmentation models to identify gaze targets [113]. These methods reduce manual effort and scale well but often offer limited transparency about label quality. Across both manual and automated workflows, annotations were primarily used as supervised training targets and evaluation benchmarks, with only one study using them solely for inference validation [114].

Overall, annotation practices remain inconsistent and minimally reported, limiting comparability and model interpretability.

Model generalisability is central for ensuring robustness across tasks, contexts, and users, yet most studies provided limited evidence of such robustness. Evaluations were limited to narrow tasks or controlled settings, providing minimal insight into real-world performance [63, 90, 167]. Only a few studies evaluated model generalisability—for example, Kontogiorgos et al. [69] tested cross-domain transfer and Zhao et al. [174] explored within-task generalisability across simple and cluttered "bring object" settings.

To assess generalisability across participants, most studies relied on internal validation methods such as leave-one-participant-out cross-validation (LOOCV) [4, 63, 137], K-fold validation [15, 33, 115] and simple train–validation–test splits of the dataset [44, 167]. These approaches show how consistently a model performs across individuals within a dataset but offer limited insight into generalisation to entirely new participant groups. One exception is Madhavilatha and Krishna [81], where cross-demographic transfer was examined across age groups. Several other studies did not report participant-level evaluation at all [90, 123, 169], further limiting understanding of robustness across diverse users.

Overall, evidence for both cross-task and cross-participant generalisability remains limited, highlighting the need for more systematic and diverse evaluation frameworks.

6.2 How is implicit gaze and speech interaction applied across task and application domains?

Implicit gaze and speech interaction has been applied across diverse domains to improve the interaction itself or to infer user state. We group studies into two broad categories: Grounding and Communication, where gaze and speech resolve ambiguity, enhance recognition, and support dialogue and State and Intent Inference, where these features are modelled to infer cognitive, affective, or intentional states. Table 3 summarises the sub-tasks for each.

Table 3: Task categories and sub-tasks of implicit gaze–speech interactions, showing the distribution of studies across grounding & communication and state & intent inference.

Task Category	Sub-Task (# studies)	Studies
Grounding & Communication	Reference Resolution (9)	[21], [44], [62], [66], [90], [111], [112], [123], [160]
	Enhancing Speech Recognition (7)	[32], [33], [58], [115], [114], [137], [143]
	Improving Contextual Grounding & Retrieval (7)	[3], [63], [64], [65], [121], [168], [169]
	Managing Conversational Dynamics (4)	[17], [68], [69], [72]
State & Intent Inference	Affect Recognition (4)	[4], [67], [96], [101]
	Inferring Attention & Engagement (9)	[7], [14], [15], [41], [81], [104], [113], [153], [167]
	Inferring Personality (4)	[138], [139], [140], [141]
	Predicting User Intent (4)	[29], [77], [105], [174]

As acknowledged in Section 3.2, explicit and implicit gaze and speech interaction lie on a continuum rather than forming a strict divide. Within implicit interaction, some tasks are clearly implicit—such as modelling natural gaze and speech to track conversational dynamics or inferring cognitive state.

However, many tasks are ambiguous. For example, in tasks such as reference resolution or enhancing speech recognition, systems passively interpret users' visual attention as part of the interaction. Here, speech may be explicit, but gaze remains a spontaneous cue, and users are not required to fixate deliberately. Users can, nevertheless, naturally look at relevant objects, especially once they notice that their gaze influences system behaviour [133]. These tasks are still implicit because the system is designed to infer meaning from spontaneous behaviour rather than intentional multimodal coordination.

6.2.1 Grounding and Communication. This category groups studies that improve the robustness and fluidity of interaction by using gaze and speech to anchor references, contextualise spoken input, and maintain conversational flow.

Reference Resolution. Reference resolution is the task of identifying the specific item a user refers to in a spoken utterance [63]. For example, a command such as "colour this green" is uninterpretable unless the system can determine what "this" denotes. Nine studies have examined how gaze can support this process. The typical approach is to detect when an ambiguous reference occurs in speech by keyword spotting and then analyse gaze fixations to infer the referent, based on the assumption that users naturally look at the object they mention [21, 66, 90, 123, 160]. More advanced methods expand the analysis to a broader time window around the ambiguous expression, extracting gaze features [44] or combining gaze and speech features [62, 111, 112] to improve prediction accuracy. Together, this body of work shows that gaze-informed modelling enhances spatial disambiguation and increases the robustness of spoken command interpretation.

Enhancing Speech Recognition. Automatic speech recognition (ASR) remains difficult in real-world settings, where noise and speaker variability reduce accuracy [165]. To improve robustness, five studies use gaze to adapt language models through two approaches. One updates the model in real time based on the user's point of focus [33, 58, 143], offering immediate responsiveness. The other relies on gaze history to prioritise words from previously viewed objects, providing more stable interpretation [32, 137].

This same principle—gaze supplies contextual grounding for speech—extends beyond recognition accuracy to vocabulary growth. Out-of-vocabulary terms, particularly in domain-specific contexts, pose a persistent challenge. For instance, in medical dictation, ASR systems often fail to recognise newly introduced drug names not present in the training vocabulary [127]. To address this, Qu and Chai [114, 115] aligned gaze fixations with spoken utterances via semantic similarity, enabling novel words to be linked with visually attended objects. Here, gaze functions not only as an attentional marker but also as a lexical grounding signal, allowing passive behaviour to expand system vocabulary without explicit supervision.

Improving Contextual Grounding and Retrieval. As gaze offers spatial cues to attentional focus and speech contributes semantic intent, their combination allows systems to ground spoken content in context—linking utterances to referents, identifying genuine interest, and refining retrieval. This potential has been explored across two main applications: voice annotation and content-based information retrieval.

In voice annotation, utterances are linked to relevant objects or text. Rather than relying on single gaze points, systems such as *EyeDescribe* [121] and *GAVIN* [64] analyse fixation and saccade patterns within areas of interest (AoIs) to anchor speech to referents. While gaze alone can suggest likely targets, it is not always reliable. For example, as users may glance away while still speaking about a prior element or look ahead in anticipation, Khan et al. [63] addressed this by combining semantic features from speech with gaze-based measures, showing that multimodal integration yields more accurate annotation than either modality alone.

In content-based information retrieval (CBIR), the goal is to move beyond predefined queries by inferring genuine user interest to retrieve more relevant content. Gaze behaviour helps distinguish casual browsing from sustained attention, while speech further refines intent [3, 65] and validates retrieval outcomes [168, 169]. For example, fixation and saccadic patterns can implicitly evaluate whether retrieved images match user interest; when mismatches are detected, explicit speech corrections update content tags, improving current annotation and future retrieval effectiveness [168, 169].

Together, the work on voice annotation and CBIR demonstrates how gaze constrains possible referents or interest targets, while speech provides semantic specificity, enabling richer contextual grounding and more adaptive retrieval.

Managing Conversational Dynamics. Gaze and speech have been investigated as implicit cues for coordinating turn-taking and detecting conversational breakdowns, both essential for maintaining rhythm and balance in dialogue. Turn-taking has been a central challenge, as robots and conversational agents often fail to match the fluidity of human dialogue; interruptions or delays reduce perceived naturalness [34]. Models combining gaze direction with pitch and formant-based spectral features have been used to predict turn endings [17, 72], enabling smoother role transitions. Likewise, two studies integrated semantics of speech, prosodic signals, and gaze direction to predict breakdowns in task-oriented exchanges with robots [68, 69], allowing systems to trigger repair strategies rather than persist with unsuccessful interaction.

6.2.2 *State and Intent Inference.*

This category groups studies that leverage gaze and speech to reveal internal states and traits of the user—affect, attention, personality, and intent—enabling adaptive and personalised interaction.

Affective Recognition. Affect recognition uses behavioural and physiological cues to infer emotional states, with gaze and speech reflecting cognitive processing and expressive intent. Four studies have combined gaze features such as fixation duration and saccade dynamics with speech features including prosody and voice quality for continuous affect prediction, consistently finding that multimodal approaches outperform single modalities [4, 67, 96, 101].

However, most work still analyses gaze and speech separately—treating gaze mainly as an attention cue and speech as acoustic features—without capturing how they unfold together during affective episodes [25]. For example, heightened arousal may manifest through both erratic saccades and prosodic shifts, yet current systems rarely model these patterns together. Furthermore, studies are often limited to controlled lab settings with calibrated sensors and clean audio, raising questions about robustness in everyday contexts. Advancing affect recognition will require models that jointly capture gaze and speech as coupled affective signals and are validated in everyday contexts.

Inferring Attention and Engagement. Attention and engagement are central to interaction, shaping how individuals process information [95]. Nine studies have combined gaze and speech to infer attentional states in educational and diagnostic contexts [7, 14, 15, 41, 81, 104, 113, 153, 167]. For example, [153] proposed a rule-based method for e-learning that labelled users as attentive if they maintained visual focus on content and refrained from speaking.

Similarly, three studies measured the attention of students on online exams, where attention lapses from the exam content were treated as potential misconduct [7, 41, 104]. These approaches use gaze diversion to flag screen avoidance and acoustic or keyword analysis to detect collaboration, but their feature sets are limited. Gaze is reduced to a binary on/off-screen signal, overlooking that aversion may also reflect distraction or cognitive effort [24], while speech is treated only at the surface level, ignoring conversational context. These simplifications undermine robustness and increase false positives in real-world use.

More advanced work applies multimodal deep learning to detect attention-related conditions such as Autism Spectrum Disorder (ASD) [81, 167]. Zhang [167], for instance, combined scan path patterns with spectral speech features in a deep learning framework, showing a clear advantage over unimodal baselines.

Beyond education and diagnostics, research in human–robot interaction has used gaze and speech fusion to monitor engagement [14, 15], by tracking gaze toward the robot and integrating prosodic and spectral speech features into deep learning models to detect disengagement during spontaneous interaction. Together, this body of work highlights the potential of gaze and speech to assess attentional focus and sustained engagement across diverse HCI contexts.

Inferring Personality. Personality plays a key role in shaping how users interact with computers and robots, influencing engagement, acceptance, and overall interaction quality [162]. Four studies have investigated gaze and speech behaviour as indicators of the Big Five dimensions—Extroversion, Openness, Emotional Stability, Conscientiousness, and Agreeableness [43]. These studies have trained predictive models using multimodal features such as spectral speech descriptors and gaze fixations and direction toward the system [138–141]. While results suggest that combining gaze and speech can capture trait-level differences, feasibility for broader HCI remains limited. Most studies are conducted in tightly controlled settings (e.g. scripted Q&A sessions or constrained discussions), where user behaviours are shaped as much by the experimental design as by underlying personality traits. This task dependence limits generalisability, as cues identified in controlled settings may not transfer to open-ended interactions. Nevertheless, these studies provide a foundation for developing personality-aware systems.

Predicting User Intent. In multimodal interaction, gaze and speech function as complementary cues for anticipating user intent. Gaze typically precedes verbal action, with fixations closely tied to task demands and predictive of forthcoming behaviour [52, 122]. Speech, by contrast, offers an explicit channel for conveying intent, where even minimal verbal cues (e.g. keywords) can signal user goals [79]. Four studies have leveraged these properties to predict user intent when initiating actions [150] or selecting objects in robotic interaction [29, 105, 174]. These systems typically combine fixation patterns and keyword analysis to infer targets or actions without requiring explicit commands. In assistive robotics, gaze and speech fusion has also been used to assess input sufficiency, determining whether task specifications are complete or require clarification [77]. Collectively, this work shows that gaze signals early intent while speech conveys explicit goals, together supporting anticipation and completeness for adaptive interaction.

6.3 Takeaways

First, our synthesis shows that implicit interaction leverages natural gaze and speech as passive cues for inferring attention, cognitive state, and intent. Gaze offers temporal and spatial indicators of focus, while speech provides acoustic and semantic traces, enabling applications such as reference resolution and user-state inference that move beyond direct control toward adaptive support. Second, these systems typically rely on multimodal feature extraction—independent or joint—and integrate signals at either the feature or model level, each with trade-offs between synchronicity and robustness. Third, implicit interaction holds promise for anticipation and adaptation but remains sensitive to the quality of synchrony captured between modalities. Our synthesis highlights recurring strategies for representation and recognition in implicit systems, providing guidance for designing more robust, context-aware multimodal interactions.

7 Discussion

Human interaction with the world is inherently multimodal: we point while speaking, and conversations blend speech with gestures and facial expressions that convey intent and emotion. Systems that combine modalities similarly feel natural and intuitive [99]. Within this space, gaze and speech form a particularly effective pairing. Our review shows that the core message across both explicit and implicit interaction remains consistent: gaze provides a direct signal of attentional focus [39], while speech conveys expressive, goal-directed meaning [31]. Yet, each modality also carries limitations—gaze alone can be imprecise or ambiguous [12, 158], and speech alone can be noisy or spatially underspecified [83, 157].

By synthesising the literature, we demonstrate how their combination compensates for these weaknesses, with gaze grounding and disambiguating spoken input [70, 163], and speech clarifying semantic richness that gaze lacks [125, 149]. These complementary strengths, when combined, enable richer interaction supporting interaction tasks as varied as explicit commands (e.g. "look at an object and say delete") and implicit inference (e.g. gaze disambiguating an ambiguous "this" in reference resolution). Together, three decades of research show that this pairing reliably reduces ambiguity, increases expressiveness, and supports both precise control and adaptive system behaviour across diverse domains.

These insights are particularly relevant now as advances in eye tracking and large language models have made gaze and speech increasingly viable at scale. At the same time, people are becoming more reliant on technology in everyday life, often regarding computers not only as tools but as active partners. The rise of XR platforms, AI glasses, and mobile computing has also augmented the way we interact with devices, making traditional manual input less practical in many situations [108]. In these contexts—where users may be multitasking, have limited ability to use their hands, or require more flexible input—explicit combinations of gaze and speech can robustly support tasks such as selection, command, and navigation, while implicit combinations can enable adaptive and context-aware support, without adding interaction overhead.

Beyond efficiency, gaze and speech work particularly well for accessibility, offering an effective alternative for those who require hands-free interaction, such as people with disabilities [5, 77]. This community stands to gain further as these technologies become more robust, affordable, and widely available.

7.1 Explicit and Implicit Gaze and Speech Interaction: Similarities and Differences

Through our review of the literature on gaze and speech as combined modalities, we categorised the work into two types: *explicit* and *implicit* interaction. This section examines their similarities and differences in both fundamental input use and applications.

Explicit and implicit interaction share important similarities in how they fundamentally use gaze and speech. Both build on the same complementarity: gaze provides rapid, spatially precise grounding, while speech conveys semantic meaning. In explicit systems, these properties are formalised as functional roles (§5.1), while in implicit systems, they are extracted as behavioural features (§6.1.1), but in both cases, their value lies in working together to support interaction.

This shared foundation extends across tasks and application domains. Across both, tasks leverage the same complementary cues—gaze to establish or narrow the referent and speech to indicate intent—to facilitate interaction. For example, in explicit tasks such as selecting a file, a user might fixate on an icon and say "open this", deliberately coordinating the two modalities [47]. In implicit tasks such as note-taking or reading, gaze and speech offer subtler cues, revealing goals and engagement without overt commands [63, 64]. In essence, gaze and speech together form complementary channels for inferring user intent, whether explicit or implicit, providing a stronger and more reliable signal than either modality alone.

The two categories, however, diverge in how they leverage the modalities. Explicit interaction treats gaze and speech as intentional input channels, giving users direct control. To maximise accuracy, systems rely on predefined roles and structured input patterns. This reduces flexibility but ensures predictable behaviour, making interactions less error-prone in real time [124]. In contrast, implicit interaction treats gaze and speech as passive cues for modelling behavioural patterns and providing adaptive support [63]. Without explicit control, behaviour is more natural and less structured, requiring systems to capture features and interpret them through inference [9]. This shift from rigid to natural input makes such systems more difficult to build: they demand large training datasets, temporal modelling and robust sensing to manage noise and variability [61, 63]. However, this complexity also opens up greater opportunity, as inference-based approaches can reveal richer patterns of user state and enable more adaptive interaction.

These fundamental differences also shape the tasks and applications. Explicit systems are suited to short and high-control tasks where low error and quick response are essential [12, 152, 166]. For example, target selection can be achieved by combining gaze to indicate target with a spoken command to confirm selection, enabling precise outcomes in a single interaction [12, 56, 158]. Such systems also support straightforward recovery: if recognition fails on a command such as "select" the user can simply repeat the utterance, allowing interaction to proceed with minimal disruption.

In contrast, implicit systems operate over longer timescales and address tasks that rely on patterns unfolding gradually—how users look and speak over time—where sustained monitoring is needed to establish context [4, 96]. For instance, affect recognition depends on tracking changes in gaze dynamics and vocal prosody across extended interactions, rather than a single moment, to reliably infer emotions [4, 63, 64, 96]. Although such tasks require more complex pipelines, they add value by providing continuous background support without deliberate user effort.

Together, the findings across the two forms of interaction highlight a trade-off. Explicit gaze and speech interaction are reliable and straightforward for precise, short-burst input, but can feel rigid and less expressive. Implicit interactions align more closely with natural behaviour and enable continuous adaptation, but depend on sophisticated modelling to avoid errors. Given these trade-offs, our review synthesises how explicit and implicit gaze and speech interaction has been applied across domains and identifies the recurring patterns that emerge. These patterns offer a foundation for future research, helping designers choose how best to combine gaze and speech and balance precision, adaptability, and user experience in multimodal systems.

While we categorise the literature as explicit and implicit, some work does blur this distinction. In particular, intelligent user systems (IUIs)—although explicit in their interaction design—often leverage computational models to support and optimise the intentional input. For example, in EyeSayCorrect [173], users first utilise gaze to select a word on the screen and then provide a voice command to correct it. This design treats gaze and speech as intentional inputs but leverages computational models to infer user intent and context, enhancing the interaction's adaptability and efficiency. Similar hybrid approaches exist with other modalities [e.g. 50], but remain rare in gaze and speech research.

7.2 Open Challenges and Future Directions for Gaze and Speech Interaction

Across both explicit and implicit interaction, combining gaze and speech almost always outperforms unimodal baselines. Established pairings—such as *Gaze to Point + Confirm, Speech as Command*—recur across domains from text entry [49, 147] to AR system control [93], consistently reducing ambiguity and error rates.

Similarly, multimodal feature fusion improves inference accuracy in implicit systems [4, 63]. The synergistic value of gaze and speech integration is, therefore, well established. Yet this consistency has also led to a degree of repetition, with many studies revisiting the same logic through only minor variations in task or application domain. For instance, hands-free mouse emulation work often differs only in dwell thresholds or trigger words [10, 103, 172], while personality inference studies frequently rely on nearly identical data-processing pipelines [138–141]. Such work reinforces the case for multimodality but risks reinventing the wheel if it stops at proof-of-concept demonstrations.

The challenge, therefore, is not to show that gaze and speech can be combined to enhance human-computer interaction—*they clearly do*—but to tackle the deeper issues of robustness and scalability, while also exploring novel ways of using them beyond purely technical improvements. In the following, we outline the key challenges that must be addressed to move beyond repetition:

7.2.1 Challenge 1: Temporal Alignment. A recurring challenge across explicit and implicit paradigms is the temporal alignment of gaze and speech. Human communication is multimodal but not entirely synchronous: in deictic references, the fixation most strongly associated with the referent typically occurs about 630 ms before the spoken expression [60]. Gaze can therefore serve as a leading indicator of intent, but only if systems model this lead–lag rather than assuming synchrony. In explicit tasks such as disambiguating spoken commands (§5.2.3) and object manipulation (§5.2.2), many systems bind a command directly to the object under fixation at the time of speech [28, 46, 83, 157, 161, 175]. This simplification ignores natural gaze and speech timing offsets and risks applying actions to unintended objects. This challenge is also apparent in implicit tasks such as reference resolution, where studies analyse gaze and speech only at the moment a deictic is spoken [21, 66, 90, 123, 160], rather than across a temporal window to resolve deictic references. This increases vulnerability to natural gaze shifts before or during utterances, which can reduce inference accuracy.

We recommend that future work develop temporal alignment models that integrate gaze lead times, speech planning phases, and natural gaze shifts, treating timing as a core design parameter for robust integration.

7.2.2 Challenge 2: Limited Exploration of Fusion and Inference Strategies. The inference pipeline for implicit gaze and speech interaction faces two main methodological limitations. First, most studies process gaze and speech as independent streams, extracting unimodal features and combining them only at the inference stage. Although this approach improves performance over unimodal baselines, it overlooks the tightly coupled but non-synchronous dynamics of gaze and speech [60]. Evidence shows that modelling synchrony by analysing the data streams together—for example, in gaze-informed ASR adaptation [33, 143] or voice-note annotation [63]—yields a better understanding of interaction context and clear performance gains, but both such approaches and broader exploration remain rare as our review shows. Second, studies in our corpus rarely provide rationale for their choice of fusion strategy or inference model. Feature-level fusion is often adopted as the default despite evidence that model-level approaches may yield superior performance [4], and model selection is frequently left unexplained.

We recommend that future research should systematically benchmark fusion strategies and inference models as design variables rather than defaults. Without such rigour, implicit gaze and speech systems are likely to remain at the level of fragmented demonstrations rather than evolving into robust and generalisable pipelines.

7.2.3 Challenge 3: Lack of Reporting Annotation Method and Model Generalisability. The reliability of gaze and speech inference models depends on how data are annotated and model generalisation is evaluated. Yet across studies, annotation practices remain inconsistent and vaguely reported—manual labels rarely include inter-rater reliability [17, 72, 143], while automated pipelines do not describe label quality control [44, 113]. These gaps limit model interpretability and reproducibility, making it difficult to assess how annotation quality influences model behaviour. Likewise, most inference models are validated only within narrow tasks [140, 167, 168] or the same participant groups [90, 123, 153, 169], offering no evidence of robustness across users or contexts.

We recommend that future work establish transparent annotation and evaluation standards—including coder procedures, agreement metrics and explicit cross-task and cross-user generalisation evaluations—to enable more reliable and transferable multimodal inference systems.

7.2.4 Challenge 4: Ethical and Privacy Considerations. While gaze and speech integration enables flexible and adaptive interaction, it also raises ethical and privacy risks, as gaze reveals not only attention but also underlying cognitive states [24], and speech conveys cognitive state [37] as well as identity [129]. Combined, they can expose user intent [174] and broader traits such as affect [4] or personality [141], often extending beyond what users intend. For example, pipelines for reference resolution may inadvertently disclose what users attend to, how they express themselves, or when they hesitate [111, 112]. Moreover, most research has been conducted in laboratories under explicit consent, whereas real-world deployment entails continuous monitoring of spontaneous behaviour, often capturing both users and bystanders without their awareness. This raises critical questions: how to ensure transparency and user control over multimodal data, and what safeguards can protect the privacy of bystanders who do not wish to be monitored?

We recommend that future research embed privacy-by-design principles in gaze and speech systems, including on-device processing and user-facing controls to regulate or disable monitoring [51, 59]. Without such measures, the promise of natural multimodal interaction risks being undermined by mistrust.

7.2.5 Challenge 5: Evaluation Beyond Controlled Settings. We acknowledge that most studies in our corpus demonstrating the benefits of gaze and speech integration were conducted under controlled laboratory conditions. This reflects both the nature of research—where controlled, reproducible conditions are necessary for rigorous evaluation—and the focus of system design on socially acceptable private contexts, such as homes or offices, where speech and gaze are less intrusive and easier to study. However, many deployment scenarios will occur in public settings—for example, using smart glasses to compare products while shopping [160], exploring cultural exhibits in museums [36] and collaborating in shared AR workspaces [22]. In these environments, new challenges arise: gaze tracking becomes less stable due to movement and lighting variation [74], while speech is disrupted by noise, overlapping speakers, and conversational dynamics [30]. Such variability reduces the reliability of gaze as a referential cue and complicates the temporal alignment of gaze and speech. Current approaches are often tuned for controlled laboratory data, rarely account for these conditions, limiting their generalisability.

We recommend that future work prioritise in-the-wild evaluation, explicitly addressing how noisy gaze signals and dynamic speech patterns interact, and how integration strategies and inference pipelines can adapt to ensure robust multimodal interpretation.

7.2.6 Challenge 6: Supporting Diverse Users and Contexts. Although gaze and speech integration shows strong potential, it is not equally effective across users or contexts. Within our corpus, only two studies leveraged this combination for diagnostic inference in Autism Spectrum Disorder (ASD) [81, 167]. Beyond this, research remains limited for users with speech impairments, strong accents, low vocal intensity [20, 35], or visual and oculomotor disorders that compromise gaze tracking [80]. Social context further constrains feasibility: overt speech and visible tracking may feel intrusive in public settings [109], reducing acceptability even when technically functional. The open challenge is to design systems that do not assume uniform or "ideal" users.

Future work should broaden evaluation to neurodiverse, impaired, and socially constrained contexts, and develop adaptive models that accommodate variability in gaze and speech.

8 Limitations

This review has several limitations. First, although we applied broad search terms across multiple databases, some relevant works may have been missed due to alternative terminology. Second, we focused exclusively on studies where gaze and speech serve as input modalities in single-user human–computer interaction, deliberately excluding related areas such as multimodal output [27, 102] or computer-mediated communication [116, 117]. While this enabled a coherent synthesis, it narrows coverage and omits perspectives from fields like computer-mediated human–human interaction [94, 98, 135], which are growing in importance. Third, ten studies in our review examined embodied agent interactions, primarily for inferring personality or predicting user intent (§6.2). Many embodied-agent systems, however, operate in multi-user settings [6, 88], which fall outside our single-user HCI scope. Future work could extend beyond this boundary to examine gaze and speech interaction in multi-user embodied-agent contexts, where social dynamics and shared attention are central. Finally, our contribution centres on synthesising common feature representations, integration strategies, and task categories, rather than performing detailed comparative analyses of evaluation metrics. Future work could extend this foundation by systematically examining nuanced outcomes, such as performance measures and contextual effects.

9 Conclusion

In this scoping review, we systematically examined 103 studies that combine gaze and speech for human–computer interaction. We found that across the literature, this combination is used either explicitly, where gaze and speech function as deliberate input, or implicitly, where systems infer meaning from natural gaze and speech behaviour. Across both, we observed recurring patterns in how the modalities are combined, shaped by their functional roles, feature representations, and integration strategies. Moreover, their complementary strengths—with gaze providing spatial grounding and speech contributing meaning and explicit intent—enable systems to reduce ambiguity, enhance expressiveness, and support both precise control and adaptive interaction. Our synthesis consolidates a fragmented research landscape, offering a structured view of how gaze and speech work together across various application domains, and paving the way for future multimodal systems that are richer and more adaptive.

Acknowledgments

This work was supported by the IITP (Institute of Information & Communications Technology Planning & Evaluation)-ITRC (Information Technology Research Center) grant funded by the Korea government (Ministry of Science and ICT) (IITP-2026-RS-2024-00436398). Anam Ahmad Khan is thankful to the KAIST Jang Young Sil Postdoctoral Fellowship (2025-2026).

References

References marked with ✱ are included in the reviewed corpus.

[1] ✱ M. Aakay, I. Marsic, A. Medl, and Guangming Bu. 1998. A system for medical consultation and education using multimodal human/machine communication. *IEEE Transactions on Information Technology in Biomedicine* 2, 4 (1998), 282–291. https://doi.org/10.1109/4233.737584

[2] Harsh Ahlawat, Naveen Aggarwal, and Deepti Gupta. 2025. Automatic Speech Recognition: A survey of deep learning techniques and approaches. *International Journal of Cognitive Computing in Engineering* 6 (2025), 201–237. https://doi.org/10.1016/j.ijcce.2024.12.007

[3] ✱ Antti Ajanki, Mark Billinghurst, Hannes Gamper, Toni Järvenpää, Melih Kandemir, Samuel Kaski, Markus Koskela, Mikko Kurimo, Jorma Laaksonen, Kai Puolamäki, Teemu Ruokolainen, and Timo Tossavainen. 2011. An augmented reality interface to contextual information. *Virtual Real.* 15, 2–3 (June 2011), 161–173. https://doi.org/10.1007/s10055-010-0183-5

[4] ✱ Ashwaq Alhargan, Neil Cooke, and Tareq Binjammaz. 2017. Multimodal affect recognition in an interactive gaming environment using eye tracking and speech signals. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction* (Glasgow, UK) *(ICMI '17)*. Association for Computing Machinery, New York, NY, USA, 479–486. https://doi.org/10.1145/3136755.3137016

[5] ✱ JS Archana and K Gangadharan. 2015. A Real Time Personalized Self Assistive Technology for the Disabled People Based on Voice and EOG. *Asian Research Publishing Network (ARPN)* 10, 17 (2015).

[6] Mark Armstrong, Chi-Lan Yang, Kinga Skiers, Mengzhen Lim, Tamil Selvan Gunasekaran, Ziyue Wang, Takuji Narumi, Kouta Minamizawa, and Yun Suen Pai. 2024. SealMates: Improving Communication in Video Conferencing using a Collective Behavior-Driven Avatar. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 118 (April 2024), 23 pages. https://doi.org/10.1145/3637395

[7] ✱ Yousef Atoum, Liping Chen, Alex X. Liu, Stephen D. H. Hsu, and Xiaoming Liu. 2017. Automated Online Exam Proctoring. *IEEE Transactions on Multimedia* 19, 7 (2017), 1609–1624. https://doi.org/10.1109/TMM.2017.2656064

[8] Jin Bai, Mohd Shahrizal Sunar, and Norhaida Mohd Suaib. 2025. Augmented reality interaction: a comprehensive review of gesture and speech integration techniques. *Neural Comput. Appl.* 37, 17 (May 2025), 11347–11377. https://doi.org/10.1007/s00521-025-11190-w

[9] Michael Barz, Omair Shahzad Bhatti, and Daniel Sonntag. 2022. Implicit Estimation of Paragraph Relevance From Eye Movements. *Frontiers in Computer Science* Volume 3 - 2021 (2022). https://doi.org/10.3389/fcomp.2021.808507

[10] ✱ Glenn J. Beach, Charles J. Cohen, Jeff Braun, and Gary Moody. 1998. Eye tracker system for use with head mounted displays. In *SMC'98 Conference Proceedings. 1998 IEEE International Conference on Systems, Man, and Cybernetics (Cat. No.98CH36218)*, Vol. 5. 4348–4352 vol.5. https://doi.org/10.1109/ICSMC.1998.727531

[11] ✱ T. R. Beelders and P. J. Blignaut. 2012. Measuring the performance of gaze and speech for text input. In *Proceedings of the Symposium on Eye Tracking Research and Applications* (Santa Barbara, California) *(ETRA '12)*. Association for Computing Machinery, New York, NY, USA, 337–340. https://doi.org/10.1145/2168556.2168631

[12] ✱ T. R. Beelders and P. J. Blignaut. 2012. Using eye gaze and speech to simulate a pointing device. In *Proceedings of the Symposium on Eye Tracking Research and Applications* (Santa Barbara, California) *(ETRA '12)*. Association for Computing Machinery, New York, NY, USA, 349–352. https://doi.org/10.1145/2168556.2168634

[13] ✱ T. R. Beelders and P. J. Blignaut. 2014. *Gaze and Speech: Pointing Device and Text Entry Modality.* Springer International Publishing, Cham, 51–75. https://doi.org/10.1007/978-3-319-02868-2_4

[14] ✱ Atef Ben-Youssef, Chloé Clavel, and Slim Essid. 2021. Early Detection of User Engagement Breakdown in Spontaneous Human-Humanoid Interaction. *IEEE Transactions on Affective Computing* 12, 3 (2021), 776–787. https://doi.org/10.1109/TAFFC.2019.2898399

[15] ✱ Atef Ben-Youssef, Giovanna Varni, Slim Essid, and Chloé Clavel. 2019. On-the-Fly Detection of User Engagement Decrease in Spontaneous Human–Robot Interaction Using Recurrent and Deep Neural Networks. *International Journal of Social Robotics* 11, 5 (2019), 815–828. https://doi.org/10.1007/s12369-019-00591-2

[16] Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation Learning: A Review and New Perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* 35, 8 (Aug. 2013), 1798–1828. https://doi.org/10.1109/TPAMI.2013.50

[17] ✱ Miriam Bilac, Marine Chamoux, and Angelica Lim. 2017. Gaze and filled pause detection for smooth human-robot conversations. In *2017 IEEE-RAS 17th International Conference on Humanoid Robotics (Humanoids)*. 297–304. https://doi.org/10.1109/HUMANOIDS.2017.8246889

[18] Carmen Bisogni, Michele Nappi, Genoveffa Tortora, and Alberto Del Bimbo. 2024. Gaze analysis: A survey on its applications. *Image and Vision Computing* 144 (2024), 104961. https://doi.org/10.1016/j.imavis.2024.104961

[19] Richard A. Bolt. 1980. "Put-that-there": Voice and gesture at the graphics interface. In *Proceedings of the 7th Annual Conference on Computer Graphics and Interactive Techniques* (Seattle, Washington, USA) *(SIGGRAPH '80)*. Association for Computing Machinery, New York, NY, USA, 262–270. https://doi.org/10.1145/800250.807503

[20] Maria Borgestig, Jan Sandqvist, Richard Parsons, Torbjörn Falkmer, and Helena Hemmingsson. 2016. Eye gaze performance for children with severe physical impairments using gaze-based assistive technology—A longitudinal study. *Assistive Technology* 28, 2 (2016), 93–102. https://doi.org/10.1080/10400435.2015.1092182

[21] ✱ Riccardo Bovo, Steven Abreu, Karan Ahuja, Eric J Gonzalez, Li-Te Cheng, and Mar Gonzalez-Franco. 2025. EmBARDiment: an Embodied AI Agent for Productivity in XR. In *2025 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*. 708–717. https://doi.org/10.1109/VR59515.2025.00093

[22] Riccardo Bovo, Daniele Giunchi, Ludwig Sidenmark, Joshua Newn, Hans Gellersen, Enrico Costanza, and Thomas Heinis. 2023. Speech-Augmented Cone-of-Vision for Exploratory Data Analysis. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 162, 18 pages. https://doi.org/10.1145/3544548.3581283

[23] Elizabeth H. Bradley, Leslie A. Curry, and Kelly J. Devers. 2007. Qualitative Data Analysis for Health Services Research: Developing Taxonomy, Themes, and Theory. *Health Services Research* 42, 4 (2007), 1758–1772. https://doi.org/10.1111/j.1475-6773.2006.00684.x

[24] Andreas Bulling and Hans Gellersen. 2010. Toward Mobile Eye-Based Human-Computer Interaction. *IEEE Pervasive Computing* 9, 4 (2010), 8–12. https://doi.org/10.1109/MPRV.2010.86

[25] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Ebrahim (Abe) Kazemzadeh, Emily Mower Provost, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. IEMOCAP: Interactive Emotional Dyadic Motion Capture Database. *Language Resources and Evaluation* 42 (2008), 335–359. https://doi.org/10.1007/s10579-008-9076-6

[26] ✱ Davide Calandra, Filippo Gabriele Pratticò, and Fabrizio Lamberti. 2022. Comparison of Hands-Free Speech-Based Navigation Techniques for Virtual Reality Training. In *2022 IEEE 21st Mediterranean Electrotechnical Conference (MELECON)*. 85–90. https://doi.org/10.1109/MELECON53508.2022.9842994

[27] Ryan Canales, Eakta Jain, and Sophie Jörg. 2023. Real-Time Conversational Gaze Synthesis for Avatars. In *Proceedings of the 16th ACM SIGGRAPH Conference on Motion, Interaction and Games* (Rennes, France) *(MIG '23)*. Association for Computing Machinery, New York, NY, USA, Article 17, 7 pages. https://doi.org/10.1145/3623264.3624446

[28] ✱ Lizhou Cao, Huadong Zhang, Chao Peng, and Jeffrey T. Hansberger. 2023. Real-time multimodal interaction in virtual reality - a case study with a large virtual interface. *Multimedia Tools Appl.* 82, 16 (Feb. 2023), 25427–25448. https://doi.org/10.1007/s11042-023-14381-6

[29] ✱ Yu-Chen Chang, Nitish Gandi, Kazuki Shin, Ye-Ji Mun, Katherine Driggs-Campbell, and Joohyung Kim. 2023. Specifying Target Objects in Robot Teleoperation Using Speech and Natural Eye Gaze. In *2023 IEEE-RAS 22nd International Conference on Humanoid Robots (Humanoids)*. 1–7. https://doi.org/10.1109/Humanoids57100.2023.10375186

[30] G Charan, G Sivateja, M Karthik Sarma, and Sabyasachi Kumar. 2025. Unveiling the challenges of speech recognition in noisy environments: A comprehensive review of issues and solutions. *Challenges in Information, Communication and Computing Technology* (2025), 407–412. https://doi.org/10.13140/RG.2.2.24231.76966

[31] Leigh Clark, Philip Doyle, Diego Garaialde, Emer Gilmartin, Stephan Schlögl, Jens Edlund, Matthew Aylett, João Cabral, Cosmin Munteanu, Justin Edwards,

and Benjamin R Cowan. 2019. The State of Speech in HCI: Trends, Themes and Challenges. *Interacting with Computers* 31, 4 (09 2019), 349–371. https://doi.org/10.1093/iwc/iwz016

[32] ✱ Neil Cooke and Martin Russell. 2005. Using the focus of visual attention to improve spontaneous speech recognition. *9th European Conference on Speech Communication and Technology*, 1213–1216. https://doi.org/10.21437/Interspeech.2005-371

[33] ✱ Neil Cooke, Ao Shen, and Martin Russell. 2014. Exploiting a 'gaze-Lombard effect' to improve ASR performance in acoustically noisy settings. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 1754–1758. https://doi.org/10.1109/ICASSP.2014.6853899

[34] RH Cuijpers and VJP Van den Goor. 2017. Turn-taking cue delays in human-robot communication. In *2017 Workshop on Social Interaction and Multimodal Expression for Socially Intelligent Robots*. CEUR-WS. org, 19–29.

[35] Miguel Del Río, Corey Miller, Ján Profant, Jennifer Drexler-Fox, Quinn Mcnamara, Nishchal Bhandari, Natalie Delworth, Ilya Pirkin, Migüel Jetté, Shipra Chandra, et al. 2023. Accents in Speech Recognition through the Lens of a World Englishes Evaluation Set. *Research in Language* 21, 3 (2023), 225–244. https://doi.org/10.18778/1731-7533.21.3.02

[36] Piercarlo Dondi and Marco Porta. 2023. Gaze-Based Human–Computer Interaction for Museums and Exhibitions: Technologies, Applications and Future Perspectives. *Electronics* 12, 14 (2023). https://doi.org/10.3390/electronics12143064

[37] Georgios Drakopoulos, George Pikramenos, Evaggelos Spyrou, and Stavros Perantonis. 2019. Emotion Recognition from Speech: A Survey. In *Proceedings of the 15th International Conference on Web Information Systems and Technologies* (Vienna, Austria) *(WEBIST 2019)*. SCITEPRESS - Science and Technology Publications, Lda, Setubal, PRT, 432–439. https://doi.org/10.5220/0008495004320439

[38] Elias Dritsas, Maria Trigka, Christos Troussas, and Phivos Mylonas. 2025. Multimodal Interaction, Interfaces, and Communication: A Survey. *Multimodal Technologies and Interaction* 9, 1 (2025). https://doi.org/10.3390/mti9010006

[39] Andrew T. Duchowski. 2020. Eye-based interaction in graphical systems: 20 years later gaze applications, analytics, & interaction. In *ACM SIGGRAPH 2020 Courses* (Virtual Event, USA) *(SIGGRAPH '20)*. Association for Computing Machinery, New York, NY, USA, Article 18, 246 pages. https://doi.org/10.1145/3388769.3407492

[40] Anna Maria Feit, Shane Williams, Arturo Toledo, Ann Paradiso, Harish Kulkarni, Shaun Kane, and Meredith Ringel Morris. 2017. Toward Everyday Gaze Input: Accuracy and Precision of Eye Tracking and Implications for Design. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 1118–1130. https://doi.org/10.1145/3025453.3025599

[41] ✱ Aumkar Gadekar, Shreya Oak, Abhishek Revadekar, and Anant V. Nimkar. 2022. MMAP: A Multi-Modal Automated Online Proctor. In *Machine Learning and Big Data Analytics (Proceedings of International Conference on Machine Learning and Big Data Analytics (ICMLBDA) 2021)*, Rajiv Misra, Rudrapatna K. Shyamasundar, Amrita Chaturvedi, and Rana Omer (Eds.). Springer International Publishing, Cham, 314–325. https://doi.org/10.1007/978-3-030-82469-3_28

[42] ✱ Fotis Giariskanis, Yannis Kritikos, Eftychia Protopapadaki, Anthi Papanastasiou, Eleni Papadopoulou, and Katerina Mania. 2025. Dynamic Difficulty Adjustment in Audio Augmented Reality Games. *J. Comput. Cult. Herit.* 18, 2, Article 25 (April 2025), 19 pages. https://doi.org/10.1145/3718330

[43] Lewis R Goldberg. 1990. An alternative "description of personality": The Big-Five factor structure. In *Personality and personality disorders*. Routledge, 34–47. https://doi.org/10.1037/0022-3514.59.6.1216

[44] ✱ Amr Gomaa, Guillermo Reyes, Michael Feld, and Antonio Krüger. 2024. Looking for a better fit? An Incremental Learning Multimodal Object Referencing Framework adapting to Individual Drivers. In *Proceedings of the 29th International Conference on Intelligent User Interfaces* (Greenville, SC, USA) *(IUI '24)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3640543.3645152

[45] Ziad M. Hafed. 2011. Mechanisms for generating and compensating for the smallest possible saccades. *European Journal of Neuroscience* 33, 11 (2011), 2101–2113. https://doi.org/10.1111/j.1460-9568.2011.07694.x

[46] ✱ Jeffrey T. Hansberger, Chao Peng, Victoria Blakely, Sarah Meacham, Lizhou Cao, and Nicholas Diliberti. 2019. A Multimodal Interface for Virtual Information Environments. In *Virtual, Augmented and Mixed Reality. Multimodal Interaction*, Jessie Y.C. Chen and Gino Fragomeni (Eds.). Springer International Publishing, Cham, 59–70. https://doi.org/10.1007/978-3-030-21607-8_5

[47] ✱ Franz Hatfield and Eric A Jenkins. 1997. An interface integrating eye gaze and voice recognition for hands-free computer access. In *Proceedings of the CSUN 1997 Conference*. 1–7.

[48] ✱ Ramin Hedeshy, Chandan Kumar, Mike Lauer, and Steffen Staab. 2022. All Birds Must Fly: The Experience of Multimodal Hands-free Gaming with Gaze and Nonverbal Voice Synchronization. In *Proceedings of the 2022 International Conference on Multimodal Interaction* (Bengaluru, India) *(ICMI '22)*. Association for Computing Machinery, New York, NY, USA, 278–287. https://doi.org/10.1145/3536221.3556593

[49] ✱ Ramin Hedeshy, Chandan Kumar, Raphael Menges, and Steffen Staab. 2021. Hummer: Text Entry by Gaze and Hum. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 741, 11 pages. https://doi.org/10.1145/3411764.3445501

[50] Baosheng James Hou, Joshua Newn, Ludwig Sidenmark, Anam Ahmad Khan, and Hans Gellersen. 2024. GazeSwitch: Automatic Eye-Head Mode Switching for Optimised Hands-Free Pointing. *Proc. ACM Hum.-Comput. Interact.* 8, ETRA, Article 227 (May 2024), 20 pages. https://doi.org/10.1145/3655601

[51] Matthew B. Hoy. 2018. Alexa, Siri, Cortana, and More: An Introduction to Voice Assistants. *Medical Reference Services Quarterly* 37, 1 (2018), 81–88. https://doi.org/10.1080/02763869.2018.1404391

[52] Chien-Ming Huang, Sean Andrist, Allison Sauppé, and Bilge Mutlu. 2015. Using gaze patterns to predict task intent in collaboration. *Frontiers in Psychology* Volume 6 - 2015 (2015). https://doi.org/10.3389/fpsyg.2015.01049

[53] Robert J. K. Jacob. 1990. What you look at is what you get: eye movement-based interaction techniques. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Seattle, Washington, USA) *(CHI '90)*. Association for Computing Machinery, New York, NY, USA, 11–18. https://doi.org/10.1145/97243.97246

[54] Robert J. K. Jacob. 1991. The use of eye movements in human-computer interaction techniques: what you look at is what you get. *ACM Trans. Inf. Syst.* 9, 2 (April 1991), 152–169. https://doi.org/10.1145/123078.128728

[55] Alejandro Jaimes and Nicu Sebe. 2007. Multimodal human–computer interaction: A survey. *Computer Vision and Image Understanding* 108, 1 (2007), 116–134. https://doi.org/10.1016/j.cviu.2006.10.019

[56] ✱ Yvonne Kammerer, Katharina Scheiter, and Wolfgang Beinhauer. 2008. Looking my way through the menu: the impact of menu design and multimodal input on gaze-based menu selection. In *Proceedings of the 2008 Symposium on Eye Tracking Research & Applications* (Savannah, Georgia) *(ETRA '08)*. Association for Computing Machinery, New York, NY, USA, 213–220. https://doi.org/10.1145/1344471.1344522

[57] Melih Kandemir and Samuel Kaski. 2012. Learning relevance from natural eye movements in pervasive interfaces. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction* (Santa Monica, California, USA) *(ICMI '12)*. Association for Computing Machinery, New York, NY, USA, 85–92. https://doi.org/10.1145/2388676.2388700

[58] ✱ Ritwij Kashyap, Raghav Billore, Prateek Kumar, and Rahul Gupta. 2025. Cross-Modal Transfer Learning for Multimodal Enhancement in Human-Computer Interaction. In *2025 6th International Conference on Inventive Research in Computing Applications (ICIRCA)*. 1938–1944. https://doi.org/10.1109/ICIRCA65293.2025.11089658

[59] Christina Katsini, Yasmeen Abdrabou, George E. Raptis, Mohamed Khamis, and Florian Alt. 2020. The Role of Eye Gaze in Security and Privacy Applications: Survey and Future HCI Research Directions. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–21. https://doi.org/10.1145/3313831.3376840

[60] ✱ Manpreet Kaur, Marilyn Tremaine, Ning Huang, Joseph Wilder, Zoran Gacovski, Frans Flippo, and Chandra Sekhar Mantravadi. 2003. Where is "it"? Event Synchronization in Gaze-Speech Input Systems. In *Proceedings of the 5th International Conference on Multimodal Interfaces* (Vancouver, British Columbia, Canada) *(ICMI '03)*. Association for Computing Machinery, New York, NY, USA, 151–158. https://doi.org/10.1145/958432.958463

[61] Petr Kellnhofer, Adria Recasens, Simon Stent, Wojciech Matusik, and Antonio Torralba. 2019. Gaze360: Physically Unconstrained Gaze Estimation in the Wild. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. 6911–6920. https://doi.org/10.1109/ICCV.2019.00701

[62] ✱ Casey Kennington and David Schlangen. 2017. A simple generative model of incremental reference resolution for situated dialogue. *Computer Speech & Language* 41 (2017), 43–67. https://doi.org/10.1016/j.csl.2016.04.002

[63] ✱ Anam Ahmad Khan, Joshua Newn, James Bailey, and Eduardo Velloso. 2022. Integrating Gaze and Speech for Enabling Implicit Interactions. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 349, 14 pages. https://doi.org/10.1145/3491102.3502134

[64] ✱ Anam Ahmad Khan, Joshua Newn, Ryan M. Kelly, Namrata Srivastava, James Bailey, and Eduardo Velloso. 2021. GAVIN: Gaze-Assisted Voice-Based Implicit Note-taking. *ACM Trans. Comput.-Hum. Interact.* 28, 4, Article 26 (Aug. 2021), 32 pages. https://doi.org/10.1145/3453988

[65] ✱ Bumhwi Kim, Amitash Ojha, and Minho Lee. 2015. Active glass-type human augmented cognition system considering attention and intention. *Connection Science* 27, 4 (2015), 322–339. https://doi.org/10.1080/09540091.2015.1051513

[66] ✱ Norihide Kitaoka, Takuma Nakagawa, Ryota Nishimura, Yoshio Ishiguro, Shin'ichi Kojima, and Shin Ohsuga. 2020. *A Multimodal Control System for Autonomous Vehicles Using Speech, Gesture, and Gaze Recognition*. De Gruyter, Berlin, Boston, 101–112. https://doi.org/10.1515/9783110669787-007

[67] ✱ Shunta Konishi, Masaki Kuwata, Yoshio Matsumoto, Yuichiro Yoshikawa, Keiji Takata, Hideyuki Haraguchi, Azusa Kudo, Hiroshi Ishiguro, and Hirokazu Kumazaki. 2024. Self-administered questionnaires enhance emotion estimation of individuals with autism spectrum disorders in a robotic interview setting. *Frontiers in Psychiatry* Volume 15 - 2024 (2024). https://doi.org/10.3389/fpsyt.2024.1249000

[68] ✱ Dimosthenis Kontogiorgos, Andre Pereira, Boran Sahindal, Sanne van Waveren, and Joakim Gustafson. 2020. Behavioural Responses to Robot Conversational Failures. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction* (Cambridge, United Kingdom) *(HRI '20)*. Association for Computing Machinery, New York, NY, USA, 53–62. https://doi.org/10.1145/3319502.3374782

[69] ✱ Dimosthenis Kontogiorgos, Minh Tran, Joakim Gustafson, and Mohammad Soleymani. 2021. A Systematic Cross-Corpus Analysis of Human Reactions to Robot Conversational Failures. In *Proceedings of the 2021 International Conference on Multimodal Interaction* (Montréal, QC, Canada) *(ICMI '21)*. Association for Computing Machinery, New York, NY, USA, 112–120. https://doi.org/10.1145/3462244.3479887

[70] ✱ David B. Koons, Carlton J. Sparrell, and Kristinn R. Thórisson. 1991. Integrating simultaneous input from speech, gaze, and hand gestures. In *Proceedings of the 1991 International Conference on Intelligent Multimedia Interfaces* (Anaheim, CA, USA) *(IMI'91)*. AAAI Press, 257–276.

[71] Panagiotis Kourtesis. 2024. A Comprehensive Review of Multimodal XR Applications, Risks, and Ethical Challenges in the Metaverse. *Multimodal Technologies and Interaction* 8, 11 (2024), 98. https://doi.org/10.3390/mti8110098

[72] ✱ Divesh Lala, Koji Inoue, and Tatsuya Kawahara. 2019. Smooth Turn-taking by a Robot Using an Online Continuous Model to Generate Turn-taking Cues. In *2019 International Conference on Multimodal Interaction* (Suzhou, China) *(ICMI '19)*. Association for Computing Machinery, New York, NY, USA, 226–234. https://doi.org/10.1145/3340555.3353727

[73] Stephen RH Langton, Roger J Watt, and Vicki Bruce. 2000. Do the eyes have it? Cues to the direction of social attention. *Trends in Cognitive Sciences* 4, 2 (2000), 50–59. https://doi.org/10.1016/s1364-6613(99)01436-9

[74] Otto Lappi. 2015. Eye Tracking in the Wild: The Good, the Bad and the Ugly. *Journal of Eye Movement Research* 8, 5 (2015). https://doi.org/10.16910/jemr.8.5.1

[75] ✱ Jaewook Lee, Jun Wang, Elizabeth Brown, Liam Chu, Sebastian S. Rodriguez, and Jon E. Froehlich. 2024. GazePointAR: A Context-Aware Multimodal Voice Assistant for Pronoun Disambiguation in Wearable Augmented Reality. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 408, 20 pages. https://doi.org/10.1145/3613904.3642230

[76] Yaxiong Lei, Shijing He, Mohamed Khamis, and Juan Ye. 2023. An End-to-End Review of Gaze Estimation and its Interactive Applications on Handheld Mobile Devices. *ACM Comput. Surv.* 56, 2, Article 34 (Sept. 2023), 38 pages. https://doi.org/10.1145/3606947

[77] ✱ Zhi Li and Ray Jarvis. 2011. Multimodal interaction system for a household assistive robot. In *Australasian Conference on Robotics and Automation (ACRA)*.

[78] ✱ Soon-Bum Lim and Joo Hyun Park. 2022. Development of an eye-tracking and voice command interface to facilitate GUI operation for people with disabled upper limbs. *Univers. Access Inf. Soc.* 23, 1 (Nov. 2022), 329–343. https://doi.org/10.1007/s10209-022-00939-y

[79] Xiaomei Liu, Shuzhi Sam Ge, Rui Jiang, and Cher-Hiang Goh. 2016. Intelligent speech control system for human-robot interaction. In *2016 35th Chinese Control Conference (CCC)*. 6154–6159. https://doi.org/10.1109/ChiCC.2016.7554323

[80] Al Lotze, Kassia Love, Anca Velisar, and Natela M Shanidze. 2024. A low-cost robotic oculomotor simulator for assessing eye tracking accuracy in health and disease. *Behavior Research Methods* 56, 1 (2024), 80–92. https://doi.org/10.3758/s13428-022-01938-w

[81] ✱ Idimadakala Madhavilatha and Konda Hari Krishna. 2025. Federated Learning for Enhanced and Private Autism Spectrum Disorder Detection in Children Across Diverse Populations. In *2025 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE)*. 1–6. https://doi.org/10.1109/IITCEE64140.2025.10915458

[82] ✱ Diako Mardenbegi and Pernilla Qvarfordt. 2015. Creating gaze annotations in head mounted displays. In *Proceedings of the 2015 ACM International Symposium on Wearable Computers* (Osaka, Japan) *(ISWC '15)*. Association for Computing Machinery, New York, NY, USA, 161–162. https://doi.org/10.1145/2802083.2808404

[83] ✱ Sven Mayer, Gierad Laput, and Chris Harrison. 2020. Enhancing Mobile Voice Assistants with WorldGaze. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–10. https://doi.org/10.1145/3313831.3376479

[84] Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica* 22, 3 (2012), 276–282.

[85] ✱ Darius Miniotas, Ivan Tugoy, and I. MacKenzie. 2005. Extending the limits for gaze pointing through the use of speech. *Information Technology and Control* 34, 3 (2005).

[86] ✱ Darius Miniotas, Oleg Špakov, Ivan Tugoy, and I. Scott MacKenzie. 2006. Speech-augmented eye gaze interaction with small closely spaced targets. In *Proceedings of the 2006 Symposium on Eye Tracking Research & Applications* (San Diego, California) *(ETRA '06)*. Association for Computing Machinery, New York, NY, USA, 67–72. https://doi.org/10.1145/1117309.1117345

[87] ✱ B. Myers, R. Malkin, M. Bett, A. Waibel, B. Bostwick, R.C. Miller, Jie Yang, M. Denecke, E. Seemann, Jie Zhu, Choon Hong Peck, D. Kong, J. Nichols, and B. Scherlis. 2002. Flexi-modal and multi-machine user interfaces. In *Proceedings. Fourth IEEE International Conference on Multimodal Interfaces*. 343–348. https://doi.org/10.1109/ICMI.2002.1167019

[88] Yukiko Nakano and Yuki Fukuhara. 2012. Estimating conversational dominance in multiparty interaction. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction* (Santa Monica, California, USA) *(ICMI '12)*. Association for Computing Machinery, New York, NY, USA, 77–84. https://doi.org/10.1145/2388676.2388699

[89] Clifford Nass and Scott Brave. 2005. *Wired for Speech: How Voice Activates and Advances the Human-Computer Relationship*. The MIT Press.

[90] ✱ Robert Nebelrath, Mohammad Mehdi Moniri, and Michael Feld. 2016. Combining Speech, Gaze, and Micro-gestures for the Multimodal Control of In-Car Functions. In *2016 12th International Conference on Intelligent Environments (IE)*. 190–193. https://doi.org/10.1109/IE.2016.42

[91] Laurence Nigay and Joëlle Coutaz. 1993. A design space for multimodal systems: concurrent processing and data fusion. In *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems* (Amsterdam, The Netherlands) *(CHI '93)*. Association for Computing Machinery, New York, NY, USA, 172–178. https://doi.org/10.1145/169059.169143

[92] SS Muhammad Nizam, Rimaniza Zainal Abidin, Nurhazarifah Che Hashim, Meng Chun Lam, Haslina Arshad, and NAA Majid. 2018. A Review of Multimodal Interaction Technique in Augmented Reality Environment. *Int. J. Adv. Sci. Eng. Inf. Technol* 8, 4-2 (2018), 1460. https://doi.org/10.18517/ijaseit.8.4-2.6824

[93] ✱ Tom Novotny, Irma Lindt, and Wolfgang Broll. 2006. A multi modal table-top 3D modeling tool in augmented environments. In *Proceedings of the 12th Eurographics Conference on Virtual Environments* (Lisbon, Portugal) *(EGVE'06)*. Eurographics Association, Goslar, DEU, 45–52. https://doi.org/10.2312/EGVE/EGVE06/045-052

[94] Marc-Antoine Nüssli, Patrick Jermann, Mirweis Sangin, and Pierre Dillenbourg. 2009. Collaboration and abstract representations: towards predictive models based on raw speech and eye-tracking data. In *Proceedings of the 9th International Conference on Computer Supported Collaborative Learning - Volume 1* (Rhodes, Greece) *(CSCL'09)*. International Society of the Learning Sciences, 78–82.

[95] Heather L. O'Brien and Elaine G. Toms. 2008. What is user engagement? A conceptual framework for defining user engagement with technology. *Journal of the American Society for Information Science and Technology* 59, 6 (2008), 938–955. https://doi.org/10.1002/asi.20801

[96] ✱ Jonny O'Dwyer, Ronan Flynn, and Niall Murray. 2017. Continuous affect prediction using eye gaze and speech. In *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. 2001–2007. https://doi.org/10.1109/BIBM.2017.8217968

[97] ✱ Oliver Ohneiser, Malte Jauer, Jonathan R. Rein, and Matt Wallace. 2018. Faster Command Input Using the Multimodal Controller Working Position "TriControl". *Aerospace* 5, 2 (2018). https://doi.org/10.3390/aerospace5020054

[98] Jennifer K. Olsen, Kshitij Sharma, Nikol Rummel, and Vincent Aleven. 2020. Temporal analysis of multimodal data to predict collaborative learning outcomes. *British Journal of Educational Technology* 51, 5 (2020), 1527–1547. https://doi.org/10.1111/bjet.12982

[99] Sharon Oviatt. 1999. Ten myths of multimodal interaction. *Commun. ACM* 42, 11 (Nov. 1999), 74–81. https://doi.org/10.1145/319382.319398

[100] Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, Roger Chou, Julie Glanville, Jeremy M Grimshaw, Asbjørn Hróbjartsson, Manoj M Lalu, Tianjing Li, Elizabeth W Loder, Evan Mayo-Wilson, Steve McDonald, Luke A McGuinness, Lesley A Stewart, James Thomas, Andrea C Tricco, Vivian A Welch, Penny Whiting, and David Moher. 2021. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 372 (2021). https://doi.org/10.1136/bmj.n71

[101] ✱ Cristina Palmero, Mikel deVelasco, Mohamed Amine Hmani, Aymen Mtibaa, Leila Ben Letaifa, Pau Buch-Cardona, Raquel Justo, Terry Amorese, Eduardo González-Fraile, Begoña Fernández-Ruanova, Jofre Tenorio-Laranga, Anna Torp Johansen, Micaela Rodrigues da Silva, Liva Jenny Martinussen, Maria Stylianou Korsnes, Gennaro Cordasco, Anna Esposito, Mounim A. El-Yacoubi, Dijana Petrovska-Delacrétaz, M. Inés Torres, and Sergio Escalera. 2025. Exploring Emotion Expression Recognition in Older Adults Interacting With a Virtual Coach. *IEEE Transactions on Affective Computing* 16, 3 (2025), 2303–2320. https://doi.org/10.1109/TAFFC.2025.3558141

[102] Yifang Pan, Rishabh Agrawal, and Karan Singh. 2024. S3: Speech, Script and Scene driven Head and Eye Animation. *ACM Trans. Graph.* 43, 4, Article 47

(July 2024), 12 pages. https://doi.org/10.1145/3658172
[103] ✱ Mohsen Parisay, Charalambos Poullis, and Marta Kersten-Oertel. 2021. EyeTAP: Introducing a multimodal gaze-based technique using voice inputs with a comparative analysis of selection techniques. *Int. J. Hum.-Comput. Stud.* 154, C (Oct. 2021), 15 pages. https://doi.org/10.1016/j.ijhcs.2021.102676
[104] ✱ Priya N. Parkhi, Amna Patel, Dhruvraj Solanki, Himesh Ganwani, and Manav Anandani. 2024. Proficient Exam Monitoring System Using Deep Learning Techniques. In *ICT: Cyber Security and Applications*, Amit Joshi, Mufti Mahmud, Roshan G. Ragel, and S. Kartik (Eds.). Springer Nature Singapore, Singapore, 31–49. https://doi.org/10.1007/978-981-97-0744-7_3
[105] ✱ Xin Peng, Song Wang, Tong Wu, and Hao Long. 2025. Multimodal Interaction-Driven User Intent Perception Model and Application. In *2025 28th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*. 1208–1213. https://doi.org/10.1109/CSCWD64889.2025.11033301
[106] Derek A. Pisner and David M. Schnyer. 2020. Chapter 6 - Support vector machine. In *Machine Learning*, Andrea Mechelli and Sandra Vieira (Eds.). Academic Press, 101–121. https://doi.org/10.1016/B978-0-12-815739-8.00006-7
[107] ✱ Lucas Plabst, Florian Niebling, Sebastian Oberdörfer, and Francisco Ortega. 2025. Order Up! Multimodal Interaction Techniques for Notifications in Augmented Reality. *IEEE Transactions on Visualization and Computer Graphics* 31, 5 (2025), 2258–2267. https://doi.org/10.1109/TVCG.2025.3549186
[108] Alexander Plopski, Teresa Hirzle, Nahal Norouzi, Long Qian, Gerd Bruder, and Tobias Langlotz. 2022. The Eye in Extended Reality: A Survey on Gaze Interaction and Eye Tracking in Head-worn Extended Reality. *ACM Comput. Surv.* 55, 3, Article 53 (March 2022), 39 pages. https://doi.org/10.1145/3491207
[109] Salil Prabhakar, Sharath Pankanti, and Anil K. Jain. 2003. Biometric recognition: security and privacy concerns. *IEEE Security & Privacy* 1, 2 (2003), 33–42. https://doi.org/10.1109/MSECP.2003.1193209
[110] ✱ John David Prieto Prada, Myung Ho Lee, and Cheol Song. 2023. A Gaze-Speech System in Mixed Reality for Human-Robot Interaction. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*. 7547–7553. https://doi.org/10.1109/ICRA48891.2023.10161010
[111] ✱ Zahar Prasov and Joyce Y. Chai. 2008. What's in a gaze? the role of eye-gaze in reference resolution in multimodal conversational interfaces. In *Proceedings of the 13th International Conference on Intelligent User Interfaces* (Gran Canaria, Spain) *(IUI '08)*. Association for Computing Machinery, New York, NY, USA, 20–29. https://doi.org/10.1145/1378773.1378777
[112] ✱ Anna Prokofieva, Malcolm Slaney, and Dilek Hakkani-Tür. 2015. Probabilistic features for connecting eye gaze to spoken language understanding. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 5311–5315. https://doi.org/10.1109/ICASSP.2015.7178985
[113] ✱ Felix Putze, Dennis Küster, Timo Urban, Alexander Zastrow, and Marvin Kampen. 2020. Attention Sensing through Multimodal User Modeling in an Augmented Reality Guessing Game. In *Proceedings of the 2020 International Conference on Multimodal Interaction* (Virtual Event, Netherlands) *(ICMI '20)*. Association for Computing Machinery, New York, NY, USA, 33–40. https://doi.org/10.1145/3382507.3418865
[114] ✱ Shaolin Qu and Joyce Y. Chai. 2008. Incorporating temporal and semantic information with eye gaze for automatic word acquisition in multimodal conversational systems. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (Honolulu, Hawaii) *(EMNLP '08)*. Association for Computational Linguistics, USA, 244–253. https://aclanthology.org/D08-1026/
[115] ✱ Shaolin Qu and Joyce Y. Chai. 2010. Context-based word acquisition for situated dialogue in a virtual world. *J. Artif. Int. Res.* 37, 1 (Jan. 2010), 247–278. https://doi.org/10.1613/jair.2912
[116] Pernilla Qvarfordt, David Beymer, and Shumin Zhai. 2005. RealTourist – A Study of Augmenting Human-Human and Human-Computer Dialogue with Eye-Gaze Overlay. In *Human-Computer Interaction - INTERACT 2005*, Maria Francesca Costabile and Fabio Paternò (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 767–780. https://doi.org/10.1007/11555261_61
[117] Pernilla Qvarfordt and Matthew Lee. 2018. Gaze patterns during remote presentations while listening and speaking. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications* (Warsaw, Poland) *(ETRA '18)*. Association for Computing Machinery, New York, NY, USA, Article 33, 9 pages. https://doi.org/10.1145/3204493.3204540
[118] Mohaimenul Azam Khan Raiaan, Md. Saddam Hossain Mukta, Kaniz Fatema, Nur Mohammad Fahad, Sadman Sakib, Most Marufatul Jannat Mim, Jubaer Ahmad, Mohammed Eunus Ali, and Sami Azam. 2024. A Review on Large Language Models: Architectures, Applications, Taxonomies, Open Issues and Challenges. *IEEE Access* 12 (2024), 26839–26874. https://doi.org/10.1109/ACCESS.2024.3365742
[119] Ismo Rakkolainen, Ahmed Farooq, Jari Kangas, Jaakko Hakulinen, Jussi Rantala, Markku Turunen, and Roope Raisamo. 2021. Technologies for Multimodal Interaction in Extended Reality—A Scoping Review. *Multimodal Technologies and Interaction* 5, 12 (2021). https://doi.org/10.3390/mti5120081
[120] Keith Rayner. 1978. Eye movements in reading and information processing. *Psychological Bulletin* 85, 3 (1978), 618–660. https://doi.org/10.1037/0033-2909.85.3.618
[121] ✱ Kyle Reinholt, Darren Guinness, and Shaun K. Kane. 2019. EyeDescribe: Combining Eye Gaze and Speech to Automatically Create Accessible Touch Screen Artwork. In *Proceedings of the 2019 ACM International Conference on Interactive Surfaces and Spaces* (Daejeon, Republic of Korea) *(ISS '19)*. Association for Computing Machinery, New York, NY, USA, 101–112. https://doi.org/10.1145/3343055.3359722
[122] Constantin A. Rothkopf, Dana H. Ballard, and Mary M. Hayhoe. 2007. Task and context determine where you look. *Journal of Vision* 7, 14 (07 2007). https://doi.org/10.1167/7.14.16
[123] ✱ Leon J. M. Rothkrantz, Pascal Wiggers, Frans Flippo, Dimitri Woei-A-Jin, and Robert J. van Vark. 2004. Multimodal Dialogue Management. In *Text, Speech and Dialogue*, Petr Sojka, Ivan Kopeček, and Karel Pala (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 621–628. https://doi.org/10.1007/978-3-540-30120-2_78
[124] ✱ David Rozado, Alexander McNeill, and Daniel Mazur. 2016. VoxVisio–Combining Gaze and Speech for Accessible HCI. *Proceedings of the RESNA/NCART, Arlington, VA, USA* (2016), 10–14.
[125] ✱ David Rozado, Louis Stephen, and Navinda Kottege. 2014. Interacting with objects in the environment using gaze tracking glasses and speech. In *Proceedings of the 26th Australian Computer-Human Interaction Conference on Designing Futures: The Future of Design* (Sydney, New South Wales, Australia) *(OzCHI '14)*. Association for Computing Machinery, New York, NY, USA, 414–417. https://doi.org/10.1145/2686612.2686676
[126] Kerstin Ruhland, Christopher E Peters, Sean Andrist, Jeremy B Badler, Norman I Badler, Michael Gleicher, Bilge Mutlu, and Rachel McDonnell. 2015. A Review of Eye Gaze in Virtual Agents, Social Robotics and HCI: Behaviour Generation, User Interaction and Perception. *Computer Graphics Forum* 34, 6, 299–326. https://doi.org/10.1111/cgf.12603
[127] Wael Salloum, Erik Edwards, Shabnam Ghaffarzadegan, David Suendermann-Oeft, and Mark Miller. 2017. Crowdsourced Continuous Improvement of Medical Speech Recognition. In *AAAI-17 Workshop on Crowdsourcing, Deep Learning, and Artificial Intelligence Agents (WS-17-07)*.
[128] Stefan Schaffer, Robert Schleicher, and Sebastian Möller. 2015. Modeling input modality choice in mobile graphical and speech interfaces. *International Journal of Human-Computer Studies* 75 (2015), 21–34. https://doi.org/10.1016/j.ijhcs.2014.11.004
[129] Natalie Schilling and Alexandria Marsters. 2015. Unmasking Identity: Speaker Profiling for Forensic Linguistic Purposes. *Annual Review of Applied Linguistics* 35 (2015), 195–214. https://doi.org/10.1017/S0267190514000282
[130] Albrecht Schmidt. 2000. Implicit Human Computer Interaction through Context. *Personal Technologies* 4, 2 (2000), 191–199. https://doi.org/10.1007/BF01324126
[131] Bernhard Scholkopf and Alexander J. Smola. 2001. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.* MIT Press, Cambridge, MA, USA. https://doi.org/10.7551/mitpress/4175.001.0001
[132] ✱ Korok Sengupta, Sabin Bhattarai, Sayan Sarcar, I. Scott MacKenzie, and Steffen Staab. 2020. Leveraging Error Correction in Voice-based Text Entry by Talk-and-Gaze. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3313831.3376579
[133] Barış Serim and Giulio Jacucci. 2019. Explicating "Implicit Interaction": An Examination of the Concept and Challenges for Research. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3290605.3300647
[134] Asma Shakil, Christof Lutteroth, and Gerald Weber. 2025. A Taxonomy and Systematic Review of Gaze Interactions for 2D Displays: Promising Techniques and Opportunities. *ACM Comput. Surv.* 57, 12, Article 308 (July 2025), 37 pages. https://doi.org/10.1145/3736250
[135] Kshitij Sharma, Ioannis Leftheriotis, and Michail Giannakos. 2020. Utilizing Interactive Surfaces to Enhance Learning, Collaboration and Engagement: Insights from Learners' Gaze and Speech. *Sensors* 20, 7 (2020). https://doi.org/10.3390/s20071964
[136] ✱ Pavan Kumar Sharma and Pranamesh Chakraborty. 2025. Evaluation of data collection and annotation approaches of driver gaze dataset. *Behavior Research Methods* 57, 6 (2025), 172. https://doi.org/10.3758/s13428-025-02679-2
[137] ✱ Ao Shen, Neil Cooke, and Martin Russell. 2013. Selective use of gaze information to improve ASR performance in noisy environments by cache-based class language model adaptation. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 1844–1848. https://doi.org/10.21437/Interspeech.2013-454
[138] ✱ Zhihao Shen, Armagan Elibol, and Nak Young Chong. 2019. Inferring Human Personality Traits in Human-Robot Social Interaction. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 578–579. https://doi.org/10.1109/HRI.2019.8673124
[139] ✱ Zhihao Shen, Armagan Elibol, and Nak Young Chong. 2019. Nonverbal Behavior Cue for Recognizing Human Personality Traits in Human-Robot Social

Interaction. In *2019 IEEE 4th International Conference on Advanced Robotics and Mechatronics (ICARM)*. 402–407. https://doi.org/10.1109/ICARM.2019.8834279

[140] ✱ Zhihao Shen, Armagan Elibol, and Nak Young Chong. 2020. Understanding nonverbal communication cues of human personality traits in human-robot interaction. *IEEE/CAA Journal of Automatica Sinica* 7, 6 (2020), 1465–1477. https://doi.org/10.1109/JAS.2020.1003201

[141] ✱ Zhihao Shen, Armagan Elibol, and Nak Young Chong. 2021. Multi-modal feature fusion for better understanding of human personality traits in social human–robot interaction. *Robotics and Autonomous Systems* 146 (2021), 103874. https://doi.org/10.1016/j.robot.2021.103874

[142] Linda E. Sibert and Robert J. K. Jacob. 2000. Evaluation of eye gaze interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (The Hague, The Netherlands) *(CHI '00)*. Association for Computing Machinery, New York, NY, USA, 281–288. https://doi.org/10.1145/332040.332445

[143] ✱ Malcolm Slaney, Rahul Rajan, Andreas Stolcke, and Partha Parthasarathy. 2014. Gaze-enhanced speech recognition. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 3236–3240. https://doi.org/10.1109/ICASSP.2014.6854198

[144] ✱ Rainer Stiefelhagen and Jie Yang. 1997. Gaze Tracking for Multimodal Human-Computer Interaction. In *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 4. 2617–2620. https://doi.org/10.1109/ICASSP.1997.595325

[145] ✱ Yong Sun, Yu (David) Shi, Fang Chen, and Vera Chung. 2009. Building a Practical Multimodal System with a Multimodal Fusion Module. In *Human-Computer Interaction. Novel Interaction Methods and Techniques*, Julie A. Jacko (Ed.). Springer Berlin Heidelberg, Berlin, Heidelberg, 93–102. https://doi.org/10.1007/978-3-642-02577-8_11

[146] ✱ Meirav Taieb-Maimon and Luiza Romanovskii-Chernik. 2025. Improving Error Correction and Text Editing Using Voice and Mouse Multimodal Interface. *International Journal of Human–Computer Interaction* 41, 8 (2025), 4718–4741. https://doi.org/10.1080/10447318.2024.2352932

[147] ✱ Yeow Kee Tan, Nasser Sherkat, and Tony Allen. 2003. Error recovery in a blended style eye gaze and speech interface. In *Proceedings of the 5th International Conference on Multimodal Interfaces* (Vancouver, British Columbia, Canada) *(ICMI '03)*. Association for Computing Machinery, New York, NY, USA, 196–202. https://doi.org/10.1145/958432.958471

[148] ✱ Yeow Kee Tan, N. Sherkat, and T. Allen. 2003. Eye gaze and speech for data entry: a comparison of different data entry methods. In *2003 International Conference on Multimedia and Expo. ICME '03. Proceedings (Cat. No.03TH8698)*, Vol. 1. I–41. https://doi.org/10.1109/ICME.2003.1220849

[149] ✱ Zoltán Tomori, Peter Keša, Matej Nikorovič, Jan Kaňka, Petr Jákl, Mojmír Šerý, Silvie Bernatová, Eva Valušová, Marián Antalík, and Pavel Zemánek. 2015. Holographic Raman tweezers controlled by multi-modal natural user interface. *Journal of Optics* 18, 1 (nov 2015), 015602. https://doi.org/10.1088/2040-8978/18/1/015602

[150] Susanne Trick, Dorothea Koert, Jan Peters, and Constantin A. Rothkopf. 2019. *Multimodal Uncertainty Reduction for Intention Recognition in Human-Robot Interaction.* IEEE Press, 7009–7016. https://doi.org/10.1109/IROS40897.2019.8968171

[151] ✱ Luka Tummolini, Andrea Lorenzon, Giancarlo Bo, and Roberto Vaccaro. 2002. iTutor: A Wireless and Multimodal Support to Industrial Maintenance Activities. In *Human Computer Interaction with Mobile Devices*, Fabio Paternò (Ed.). Springer Berlin Heidelberg, Berlin, Heidelberg, 302–305. https://doi.org/10.1007/3-540-45756-9_27

[152] ✱ Cagkan Uludagli and Cengiz Acarturk. 2018. User interaction in hands-free gaming: A comparative study of gaze-voice and touchscreen interface control. *Turkish Journal of Electrical Engineering and Computer Sciences* 26 (07 2018). https://doi.org/10.3906/elk-1710-128

[153] ✱ Yücel Uğurlu. 2014. User attention analysis for e-learning systems using gaze and speech information. In *2014 International Conference on Information Science, Electronics and Electrical Engineering*, Vol. 1. 1–5. https://doi.org/10.1109/InfoSEEE.2014.6948154

[154] ✱ Jan van der Kamp and Veronica Sundstedt. 2011. Gaze and voice controlled drawing. In *Proceedings of the 1st Conference on Novel Gaze-Controlled Applications* (Karlskrona, Sweden) *(NGCA '11)*. Association for Computing Machinery, New York, NY, USA, Article 9, 8 pages. https://doi.org/10.1145/1983302.1983311

[155] ✱ Marius S. Vassiliou, Venkataraman Sundareswaran, S. Chen, Reinhold Behringer, Clement K. Tam, M. Chan, Phil T. Bangayan, and Joshua H. McGee. 2000. Integrated multimodal human-computer interface and augmented reality for interactive display applications. In *Cockpit Displays VII: Displays for Defense Applications*, Darrel G. Hopper (Ed.), Vol. 4022. International Society for Optics and Photonics, SPIE, 106 – 115. https://doi.org/10.1117/12.397779

[156] ✱ Keith Vertanen and David J.C. MacKay. 2010. Speech dasher: fast writing using speech and gaze. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, Georgia, USA) *(CHI '10)*. Association for Computing Machinery, New York, NY, USA, 595–598. https://doi.org/10.1145/1753326.1753415

[157] ✱ Chao Wang, Matti Krüger, and Christiane B. Wiebel-Herboth. 2020. "Watch out!": Prediction-Level Intervention for Automated Driving. In *12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (Virtual Event, DC, USA) *(AutomotiveUI '20)*. Association for Computing Machinery, New York, NY, USA, 169–180. https://doi.org/10.1145/3409120.3410652

[158] ✱ Jian Wang. 1995. Integration of eye-gaze, voice and manual response in multimodal user interface. In *1995 IEEE International Conference on Systems, Man and Cybernetics. Intelligent Systems for the 21st Century*, Vol. 5. 3938–3942 vol.5. https://doi.org/10.1109/ICSMC.1995.538404

[159] ✱ Jinge Wang, Minghua Zhu, Xiumin Fan, Xuyue Yin, and Zelin Zhou. 2020. Multi-Channel Augmented Reality Interactive Framework Design for Ship Outfitting Guidance. *IFAC-PapersOnLine* 53, 5 (2020), 189–196. https://doi.org/10.1016/j.ifacol.2021.04.098

[160] ✱ Zeyu Wang, Yuanchun Shi, Yuntao Wang, Yuchen Yao, Kun Yan, Yuhan Wang, Lei Ji, Xuhai Xu, and Chun Yu. 2024. G-VOILA: Gaze-Facilitated Information Querying in Daily Scenarios. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 2, Article 78 (May 2024), 33 pages. https://doi.org/10.1145/3659623

[161] ✱ Zhimin Wang, Haofei Wang, Huangyue Yu, and Feng Lu. 2021. Interaction With Gaze, Gesture, and Speech in a Flexibly Configurable Augmented Reality System. *IEEE Transactions on Human-Machine Systems* 51, 5 (2021), 524–534. https://doi.org/10.1109/THMS.2021.3097973

[162] Sarah Woods, Kerstin Dautenhahn, Christina Kaouri, René te Boekhorst, Kheng Lee Koay, and Michael L Walters. 2007. Are robots like people?: Relationships between participant and robot personality traits in human–robot interaction studies. *Interaction Studies* 8, 2 (2007), 281–305. https://doi.org/10.1075/is.8.2.06woo

[163] ✱ Marcelo Worsley, Kevin Mendoza Tudares, Timothy Mwiti, Mitchell Zhen, and Marc Jiang. 2021. Multicraft: A Multimodal Interface for Supporting and Studying Learning in Minecraft. In *HCI in Games: Serious and Immersive Games: Third International Conference, HCI-Games 2021, Held as Part of the 23rd HCI International Conference, HCII 2021, Virtual Event, July 24–29, 2021, Proceedings, Part II.* Springer-Verlag, Berlin, Heidelberg, 113–131. https://doi.org/10.1007/978-3-030-77414-1_10

[164] ✱ Like Wu and Jiro Tanaka. 2022. Enhancing Mall Security Based on Augmented Reality in the Post-pandemic World. In *Human Interface and the Management of Information: Applications in Complex Technological Environments: Thematic Area, HIMI 2022, Held as Part of the 24th HCI International Conference, HCII 2022, Virtual Event, June 26 – July 1, 2022, Proceedings, Part II.* Springer-Verlag, Berlin, Heidelberg, 296–314. https://doi.org/10.1007/978-3-031-06509-5_21

[165] Dong Yu and Li Deng. 2016. *Automatic Speech Recognition: A Deep Learning Approach* (1st ed.). Springer Publishing Company, Incorporated. https://doi.org/10.1007/978-1-4471-5779-3

[166] ✱ Naohiro Yuasa, Kohei Mitsui, Hiroki Sakakibara, Hiroshi Igaki, Masahide Nakamura, and Ken-ichi Matsumoto. 2008. Operating networked appliances using gaze information and voice recognition. In *Proceedings of the Third IASTED International Conference on Human Computer Interaction* (Innsbruck, Austria) *(HCI '08)*. ACTA Press, USA, 107–112. https://doi.org/10.5555/1722359.1722379

[167] ✱ Alec Zhang. 2023. A Novel Eye-tracking and Audio Hybrid System for Autism Spectrum Disorder Early Detection. In *2023 IEEE 3rd International Conference on Data Science and Computer Application (ICDSCA)*. 1495–1500. https://doi.org/10.1109/ICDSCA59871.2023.10393215

[168] ✱ He Zhang, Teemu Ruokolainen, Jorma Laaksonen, Christina Hochleitner, and Rudolf Traunmüller. 2011. Gaze- and Speech-Enhanced Content-Based Image Retrieval in Image Tagging. In *Artificial Neural Networks and Machine Learning – ICANN 2011*, Timo Honkela, Włodzisław Duch, Mark Girolami, and Samuel Kaski (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 373–380. https://doi.org/10.1007/978-3-642-21738-8_48

[169] ✱ He Zhang, Mats Sjöberg, Jorma Laaksonen, and Erkki Oja. 2011. A Multimodal Information Collector for Content-Based Image Retrieval System. In *Neural Information Processing*, Bao-Liang Lu, Liqing Zhang, and James Kwok (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 737–746. https://doi.org/10.1007/978-3-642-24965-5_83

[170] ✱ Qiaohui Zhang, Kentaro Go, Atsumi Imamiya, and Xiaoyang Mao. 2004. Robust object-identification from inaccurate recognition-based inputs. In *Proceedings of the Working Conference on Advanced Visual Interfaces* (Gallipoli, Italy) *(AVI '04)*. Association for Computing Machinery, New York, NY, USA, 248–251. https://doi.org/10.1145/989863.989905

[171] ✱ Qiaohui Zhang, Atsumi Imamiya, Kentaro Go, and Xiaoyang Mao. 2004. Resolving ambiguities of a gaze and speech interface. In *Proceedings of the 2004 Symposium on Eye Tracking Research & Applications* (San Antonio, Texas) *(ETRA '04)*. Association for Computing Machinery, New York, NY, USA, 85–92. https://doi.org/10.1145/968363.968383

[172] ✱ Darisy G. Zhao, Nikita D. Karikov, Eugeny V. Melnichuk, Boris M. Velichkovsky, and Sergei L. Shishkin. 2020. Voice as a Mouse Click: Usability and Effectiveness of Simplified Hands-Free Gaze-Voice Selection. *Applied Sciences* 10, 24 (2020). https://doi.org/10.3390/app10248791

[173] ✱ Maozheng Zhao, Henry Huang, Zhi Li, Rui Liu, Wenzhe Cui, Kajal Toshniwal, Ananya Goel, Andrew Wang, Xia Zhao, Sina Rashidian, Furqan Baig, Khiem Phi, Shumin Zhai, IV Ramakrishnan, Fusheng Wang, and Xiaojun Bi. 2022. EyeSayCorrect: Eye Gaze and Voice Based Hands-free Text Correction for Mobile Devices. In *Proceedings of the 27th International Conference on Intelligent User Interfaces* (Helsinki, Finland) *(IUI '22)*. Association for Computing Machinery, New York, NY, USA, 470–482. https://doi.org/10.1145/3490099.3511103

[174] ✱ Xiyuan Zhao, Huijun Li, Tianyuan Miao, Xianyi Zhu, Zhikai Wei, Lifen Tan, and Aiguo Song. 2024. Learning Multimodal Confidence for Intention Recognition in Human-Robot Interaction. *IEEE Robotics and Automation Letters* 9, 9 (2024), 7819–7826. https://doi.org/10.1109/LRA.2024.3432352

[175] ✱ Ao Zhou, Lei Han, and Yuzhen Meng. 2023. Multimodal Control of UAV Based on Gesture, Eye Movement and Voice Interaction. In *Advances in Guidance, Navigation and Control*, Liang Yan, Haibin Duan, and Yimin Deng (Eds.). Springer Nature Singapore, Singapore, 3765–3774. https://doi.org/10.1007/978-981-19-6613-2_366

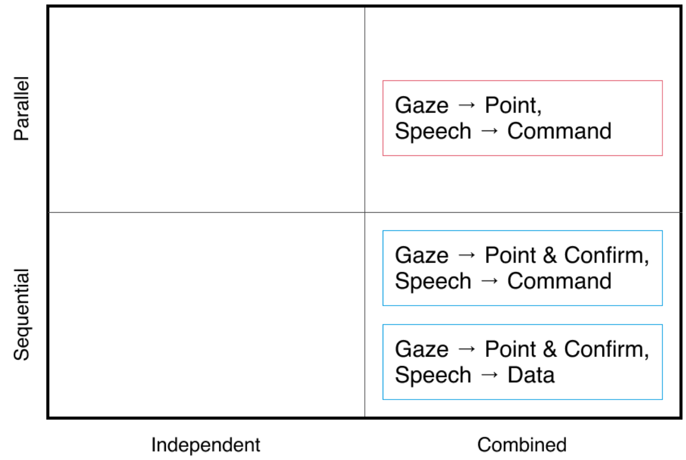

Figure 6: Design space of gaze and speech interaction, based on Nigay and Coutaz [91] multimodal framework, contrasting parallel vs. sequential fusion and independent vs. combined use of modalities.

Table 4: Search query string for ACM-DL, IEEE, Web of Science, Scopus

Database	Search Query
ACM Digital Library	(Abstract:(("gaze" OR "eye")) OR Title:(("gaze" OR "eye")) OR Keyword:(("gaze" OR "eye"))) AND (Title:(("speech" OR "voice" OR "audio" OR "vocal")) OR Abstract:(("speech" OR "voice" OR "audio" OR "vocal")) OR Keyword:(("speech" OR "voice" OR "audio" OR "vocal"))) AND (Abstract:(("interact*" OR "communicat*" OR "input" OR "technique" OR "model" OR "system" OR "interface")) OR Title:(("interact*" OR " communicat*" OR "input" OR "technique" OR "model" OR "system" OR "interface")) OR Keyword:(("interact*" OR " communicat*" OR "input" OR "technique" OR "model" OR "system" OR "interface")))
IEEEX	((("Document Title":"gaze" OR "Document Title":"eye" OR "Abstract":"gaze" OR "Abstract":"eye" OR "Author Keywords":"gaze" OR "Author Keywords":"eye") AND ("Document Title":"speech" OR "Document Title":"voice" OR "Document Title":"audio" OR "Document Title":"vocal" OR "Abstract":"speech" OR "Abstract":"voice" OR "Abstract":"audio" OR "Abstract":"vocal" OR "Author Keywords":"speech" OR "Author Keywords":"voice" OR "Author Keywords":"audio" OR "Author Keywords":"vocal") AND ("Document Title":"interact*" OR "Document Title":"communicat*" OR "Abstract":"interact*" OR "Abstract":"communicat*" OR "Author Keywords":"interact*" OR "Author Keywords":"communicat*" OR "Document Title":"input" OR "Document Title":"technique" OR "Document Title":"model" OR "Document Title":"system" OR "Document Title": "interface" OR "Abstract":"technique" OR "Abstract":"model" OR "Abstract":"system" OR "Abstract":"interface" OR "Abstract":"input" OR "Author Keywords":"technique" OR "Author Keywords":"model" OR "Author Keywords":"system" OR "Author Keywords":"interface" OR "Author Keywords":"input")))
Scopus	((TITLE (("gaze" OR "eye")) OR ABS (("gaze" OR "eye")) OR AUTHKEY (("gaze" OR "eye"))) AND (TITLE (("speech" OR "voice" OR "audio" OR "vocal")) OR ABS (("speech" OR "voice" OR "audio" OR "vocal")) OR AUTHKEY (("speech" OR "voice" OR "audio" OR "vocal"))) AND ((TITLE (("interact*" OR "communicat*" OR "input" OR "technique" OR "model" OR "system" OR "interface")) OR ABS (("interact*" OR "communicat*" OR "input" OR "technique" OR "model" OR "system" OR "interface")) OR AUTHKEY (("interact*" OR "communicat*" OR "input" OR "technique" OR "model" OR "system" OR "interface")))))
Web of Science	(TI=("gaze" OR "eye") OR AB=("gaze" OR "eye") OR AK=("gaze" OR "eye"))AND (TI=("speech" OR "voice" OR "audio" OR "vocal") OR AB=("speech" OR "voice" OR "audio" OR "vocal") OR AK=("speech" OR "voice" OR "audio" OR "vocal")) AND (TI=("interact*" OR "communicat*" OR "input" OR "technique" OR "model" OR "system" OR "interface") OR AB=("interact*" OR "communicat*" OR "input" OR "technique" OR "model" OR "system" OR "interface") OR AK=("interact*" OR "communicat*" OR "input" OR "technique" OR "model" OR "system" OR "interface"))

Table 5: Column descriptions of the table created for analyzing data extracted from selected papers

Column ID	Column Description
C1	What task is being facilitated through the combination of gaze and speech ?
C2	Who are the users interacting through the combination of gaze and speech (e.g., robot, user) ?
C3	Is the combination of gaze and speech employed for real-time interaction or for the offline analysis ?
C4	Are gaze and speech used as sequential or parallel modalities in the context of interaction ?
C5	What specific gaze and speech cues are used during the interaction ?
C6	Do the extracted gaze cues serve to facilitate interaction in an explicit or implicit manner ?
C7	Does the interaction require the fusion of gaze and speech modalities ?
C8	At what level (feature, or model) does the fusion of gaze and speech occurs ?
C9	What methodological techniques are used to combine gaze and speech cues for interaction ?
C10	What annotation practices are used for labelling ground truth data in implicit interaction ?
C11	How are models built and evaluated for inferences across task and users ?