Towards Multi-Modal Conversational Information Seeking

Yashar Deldjoo
Polytechnic University of Bari
Italy
yashar.deldjoo@poliba.it

Johanne R. Trippas
University of Melbourne
Australia
johanne.trippas@unimelb.edu.au

Hamed Zamani
University of Massachusetts Amherst
United States
zamani@cs.umass.edu

ABSTRACT

Recent research on conversational information seeking (CIS) mostly focuses on uni-modal interactions and information items. This perspective paper highlights the importance of moving towards developing and evaluating multi-modal conversational information seeking (MMCIS) systems as they enable us to leverage richer context, overcome errors, and increase accessibility. We bridge the gap between the multi-modal and CIS research and provide a formal definition for MMCIS. We discuss potential opportunities and research challenges in designing, implementing, and evaluating MMCIS systems. Based on this research, we propose and implement a practical open-source framework for facilitating MMCIS research.

CCS CONCEPTS

• **Information systems** → *Specialized information retrieval*;

ACM Reference Format:
Yashar Deldjoo, Johanne R. Trippas, and Hamed Zamani. 2021. Towards Multi-Modal Conversational Information Seeking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21), July 11–15, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3404835.3462806

1 INTRODUCTION

Interactivity is core to information-seeking tasks, and human conversation is the most natural communication tool. This has motivated researchers and practitioners to imagine conversational interactions with information-seeking systems for many decades [11, 44]. Recent advances in automatic speech recognition (ASR) and deep learning models for language understanding and generation, including the popularity of devices such as smartphones, have created an increasing interest in the area of conversational information seeking (CIS). Despite the main focus of previous work on uni-modal interactions and information seeking in conversational environments, it is widely known that human conversations are multi-modal. We communicate with each other not only by speech but also using a multitude of modes. Nevertheless, searching for information is still mainly conducted over a visual channel (i.e., typed queries and lists of search results). These properties call for developing CIS systems that provide multi-modal items and interact with users through channels with multiple modalities.

SIGIR '21, July 11–15, 2021, Virtual Event, Canada

ACM ISBN 978-1-4503-8037-9/21/07…$15.00
https://doi.org/10.1145/3404835.3462806

In this perspective paper, we study multi-modal conversational information seeking (MMCIS) tasks from multiple perspectives. We define MMCIS which extends and combines the advantages of multi-modal and conversational search interactions. We use the notion of "conversation" as an information exchange with more than two turns instead of commands such as setting a timer or turning on the light. Furthermore, we focus on the topical transactional role of search versus the social chit-chat [9, 60]. Despite the growing interest in making everything "conversational", little work has focused on establishing how conversational search interactions can be used in a multi-modal system. In addition, little is know about the variety of devices, contexts, and tasks for MMCIS. We aim to address this gap by investigating: (1) *Why* using MMCIS, (2) *Which* tasks to support in MMCIS, (3) *When* to integrate multiple modalities and conversations to reinforce MMCIS, and (4) *How* to research multiple modalities and conversations to enable MMCIS.

The goal of this paper is to act as a starting point for discussions between several research areas, including information retrieval (IR), recommender systems (RS), multi-media (MM), and human-computer interaction (HCI), and link to psychological and cognitive sciences [29]. The intersection of the research areas as an interdisciplinary field to enable people to search for information through multi-modal conversations has not received focused attention [10]. Imagine the following example: a person is cycling along the road on their way to work. She is planning her day, including tasks from presenting a budget, hosting a new client, picking up their children after school, and making dinner. The cyclist passes a flower on the sideroad, which caught her eye and wanted to know what this plant is. Since she is cycling on a busy road, she quickly stops, takes a photo, and keeps riding. Meanwhile, she asks her earbuds to tell her which plant that was by a spoken query such as "*what was that plant and is it edible?*". The MMCIS system combines the GPS location, photo, and query (input modalities). The system senses via the heart-rate monitor that she is still cycling and therefore does not provide visual information but instead speaks into the earbuds (output modality selection). Next, the cyclist may ask for information on their work task on budget presentation. We hypothesize a range of modality combinations to inform and specify the information need through conversational interactions appropriately.

Our approach is the following. We study past work on multi-modal interactions and conversations, and propose a set of definitions related to MMCIS systems. In this paper, we provide a background on modality and multi-modality levels for the IR community (cf. Section 3), and make the following contributions:

- We suggest a formal definition for MMCIS (Section 4).
- We outline IR challenges and bridge the gap between IR and other fields such as HCI and MM (Section 5).

- We propose practical ways for developing and evaluating MMCIS systems. More importantly, we extend the Macaw architecture [64] by supporting different MMCIS tasks and publicly release the resulted platform, named Macaw-MMCIS (Section 6).

Assumptions: For simplicity of discussion, the term information-seeking (IS) is used to refer to a set of relevant systems, including (and not limited to) question answering (QA), recommendation, and IR systems. We assume all these systems take as input a representation of the user's information need, e.g., a keyword query, the user profile, or a natural language question, to provide information access to any documents, answers, or information units that satisfy the user's information need.

2 RELATED WORK

In this section, we introduce the main concepts of conversational information seeking and multi-modal information seeking. We make a distinction between the two main systems, *(i)* uni-modal information seeking and *(ii)* multi-modal information seeking (MMIS).

2.1 Uni-Modal Information Seeking

Users often search for information by interacting with a traditional browser-based search engine such as Google, Bing, or Yandex. Commonly, users write queries into a search box representing their information need. These written queries are usually a short statement of the user's knowledge gap [5]. The system uses these queries to match, rank, retrieve, and present documents to help users solve their information needs. This process is often completed with solely textual information and can be classified as *uni-modal* information seeking in which both input and output are presented in text.

More recently, with the developments around machine learning and natural language processing (NLP), the search interactions in which both input and output are presented via speech has received considerable attention [33, 58]. This more natural way of expressing an information need via speech is considered an advantage of the spoken conversational search interaction paradigm. Other benefits include that users can access information even when there is no screen or keyboard available, on mobile devices, or on-the-go. Furthermore, people with limited literacy skills may benefit from non-visual interactions. Nevertheless, presenting search results and documents without overwhelming the user with information over a speech-only channel is challenging [57, 59, 62]. Even though all information should be able to be presented over audio for accessibility reasons, much different information formats benefit from a non-audio form (i.e., images or tables) [67, 69]. Spoken conversational search can still be seen as a uni-modal way of interaction. That is because, although the input and output are presented via audio in a spoken conversational search setting, ASR and speech synthesis techniques still transform speech into a text format.

Increasingly, major search companies are using the advantages of multi-modal interactions. For example, instead of a search query submitted in a query box, users can now search by submitting an image to retrieve similar images, names, and locations. Furthermore, with personalizations through accounts, GPS locations, or multiple devices, everyday search interactions move away from the traditional desktop search to an even more ubiquitous activity.

2.2 Multi-Modal Information Seeking

MMIS systems are often divided into two sub-systems: *(i)* classical multi-modal search and RS, and *(ii)* multi-modal information seeking through spoken dialogs. We discuss these two system types.

Classical multi-modal information-seeking: It has been shown that searching through keywords or recommendations often benefit from multi-modal signals, from contextual item recommendation [32, 63], to visual-based and multimedia recommendation [16, 18, 41], cold-start dilemmas [14, 17, 45], or explaining and visualizing recommendation outcomes [55]. Furthermore, many IS tasks can be enhanced by using the available multi-modal domain knowledge. For instance, what does Persian or Ancient Egyptian architecture look like, what makes Schindler's List's story development strong, or what kind of outfit is suited for a holiday resort?

A recent survey by Deldjoo et al. [20] provides a frame of reference for multi-modal RS highlighting how multimedia content (here referred to as audio, visual, and textual content) can be useful in real-world recommendation problems. These systems are built in two recommendation scenarios, both accepting multimedia input but providing different types of output:

- to suggest a specific *media item* to a user–for example, a music track or a video; or
- to suggest a *non-media item* by exploiting the media (e.g., product images) associated with the information items–for example, to recommend fashion items or food based on the visual appearance of the images associated with the items.

From an algorithmic point-of-view, a major issue in MMIS relates to the fusion of several modalities (e.g., text and image) to obtain a meaningful representation. Recent state-of-the-art techniques employ joint representation techniques to find a latent space in which multiple modality information can be projected and compared [26]. This can be a challenging task, e.g., while content data such as text and images might be well-aligned, obtaining a latent space to perform joint content and user-preference representation (e.g., ratings, clicks, social-media information) might not be trivial. Textual (uni-modal) embedding techniques such as word2vec [42], glove [49], and BERT [21], have been a driving force to design multi-modal embedding techniques e.g., based on graphs [22], deep neural networks [8], and general-purpose techniques [27]. These approaches have now found a long-standing position across various tasks related to MMCIS, such as QA, recommendation, cross-media retrieval [8], and multi-modal dialogue systems [43].

Multi-modal dialogues: Due to the advances in deep learning, research at the intersection of vision and language is unified. This has led to increasing demand for multi-modal dialogue agents, which unlike MMCIS that mostly focuses on open-domain information seeking tasks, they mostly focus on chit-chat conversations or task-oriented dialogues in a specific domain. The primary bottleneck for advancing research in (deep) multi-modal dialogues is the availability of large-scale datasets in modalities beyond text. As a leading study, Saha et al. [53] presented a multi-modal benchmark dataset for the fashion domain. The authors developed two multi-modal neural models in the encode-attend-decode paradigm and demonstrated the proposed systems' efficacy on two relevant sub-tasks, namely text response generation and best image response selection. The MMD dataset paved the path for interesting new

challenges on task-oriented dialog systems and visually grounded dialogues. Later, Liao et al. [38] built a knowledge-aware multimodal dialog system (KMD), based on deep reinforcement learning, and a hierarchical neural system to produce more substantive responses. Cui et al. [13] present a KMD system in which the authors pay more attention to user explicit requirement in the attributes by dynamically encoding the dialog history based on user attention. Facebook AI released SIMMC [12], the situated interactive multimodal conversation dataset, which improves classical task-oriented dialogues (based on basic tasks such as question-and-answer or call-and-response) by situating the user in an online store based on AR/VR and run complex multi-modal actions, such as changing the view of an object in the scene, searching or adding to cart.

3 MULTI-MODAL INTERACTIONS

Multi-modality is a highly inter-disciplinary concept; it concerns integrating different information sources, enabling expanding knowledge from only one source. In the following, we present the definition of different modality types (Section 3.1), and the modality principles, linking to psychological and cognitive science (Section 3.2).

3.1 Modality Types

Several terms pertinent to multi-modality include multi-modal interactions, multi-modal interfaces, modes and modality, channels, platforms, multi-sensory, and multimedia. These vocabularies have adopted different implications in various communities [39, 61]. There are two views on multi-modal interaction: *human-centered* and *system-oriented* view [52]. The human-centered view relates the term modality to human sensory modalities (the five senses of vision, hearing, touch, taste, and smell), and multi-modal interaction refers to the capacity of the human to *(i)* receive, *(ii)* process, and *(iii)* deliver information from/to the outside using more than one sensory modality. In contrast, the system-oriented view focuses on the system and regards multi-modal interactions as systems that accept many different inputs combined in a meaningful manner. We combine these two views in Figure 1, showing a multi-modal process starting from the human, hardware, software to the human. The process displays the different components/features involved:

- The user *inputs human actions* through the activation of muscles (e.g., vocal cords, hand), corresponding to several human biological/sensorial modalities.
- The user communicates with the computer using several physical *input devices* (keyboard, mouse), or more advanced ones such as motion or eye-gaze tracking sensors. These input devices correspond to different *interaction channels* (see Section 3.1.1).
- The information sensed by the machine's input devices produce different representations of data in audio, text, images, video, or the *presentation mediums*. These data provide distinctive levels of understanding of user intention at diverse semantic levels (i.e., low-level, semantic). Audio, image, and text correspond to different processing modalities (see Section 3.1.2).
- The computer processes the information coming from constituting modalities through applying various computer vision, NLP, audio analysis, and data fusion, corresponding again to the *processing modalities* (see Section 3.1.2).
- The computer outputs the message through appropriate devices (e.g., screen, loudspeaker). The computer may send statistically raw data (e.g., static images, audio files, or video clips) or data generated dynamically from abstract representations (such as generation of text, graphics, or speech synthesis) (see Section 3.1.3).
- Eventually, several user senses are stimulated by the system output (e.g., vision, hearing).

Multi-modality can be analyzed with respect to the **communication/interaction channel**, **processing modalities**, **presentation mode**, or a combination.

3.1.1 Interaction Channel. A channel is a pathway through which information is conveyed. We use this term to refer to its HCI usage, i.e., an interaction technique that exploits a specific combination of the user's ability and device capability to enable human-computer communication. This modality refers to *technologies* or *tasks* related to communication between a person and a system, as input (to the system) or output (to a human).

- *Input channels:* they involve various input types based on pointing, touch, speech, body gesture; for instance, mice, keyboard (traditional input devices), or interaction devices based on touch or vision such as gesture recognition and motion tracking (modern input devices), where the latter allows user to interact with a computer more naturally and intuitively.
- *Output channels:* such as traditional 2D screens, audio output, to holograms.

Example: The cyclist example in Section 1 uses multiple sensors/devices (GPS, heart-rate monitoring), together with the photo and the query, which constitute different interaction channels. This system is a multi-modal system from the channel perspective.

3.1.2 Processing Modality. Processing modalities describe the *processes* performed by the system and the *data representation* of information items. They are often based on human biological senses (or sensory modalities), namely visual, auditory, tactile, smell, and taste. From a processing point-of-view, we can recognize the following primary processing modalities, visual (V), text (T), audio (A), touch (To), others (O).[1] Note that emerging technologies such as non-evasive sensing of neural activities via a brain-computer interface may become an indispensable part of multi-modal systems in the future [61]. We consider signals obtained from such technologies as part of the category "others".

Example: A QA system that combines text and images input (question) or output (answer) level is a multi-modal QA system at the processing level, where the constituting modalities include T+V.

3.1.3 Presentation Mode. This level is helpful to reason about modality's psychological/cultural impact or explains the modality effects based on the human brain. The presentation mode of communication refers to whether the presented stimulus to human is *verbal* or *non-verbal*.

- *Verbal communication (VC):* It is the process of sending and receiving messages through the use of *words*.
- *Non-Verbal communication (NVC):* It is the *wordless* process of conveying messages, e.g., using facial expressions, body language, gesture, and posture.

[1] Note that although "text" does not correspond to an individual human sensory modality due to its predominant use we consider it as a processing modality.

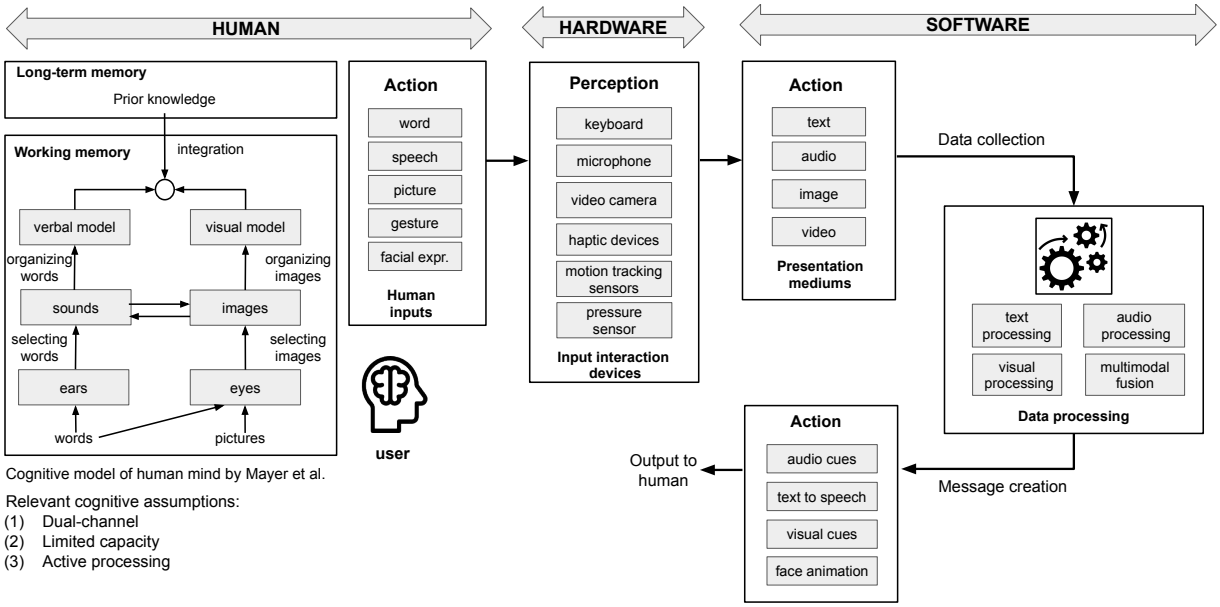

Figure 1: A conceptual design of a multi-modal system and its different elements (Figure is adapted from [35, 40]).

VC enables humans to convey messages clearly and fast, while NVC often adds to or complements the verbal message. When people interact, the nature of their relationships is affected by non-verbal factors, e.g., their distance from each other, eye contact, posture, or facial expression. VC and NVC together provide a rich source of information to develop rapport and trust in communications. For example, NVC can be vital for communicating with people with limited communication abilities, such as aphasia [23]. These people heavily rely on NVC, such as gesturing and drawing.

Example: We can further categorize VC and NVC according to vocal (spoken) and non-vocal (written). For instance, the spoken words "put that there" are the VC's vocal elements, whereas its writing is the VC's non-vocal element. On the other hand, the tone of the voice, sighs, and vocal qualities such as the rate of speech are the vocal element of the NVC, and lastly, the sign language of "put that there" is the non-vocal part of the NVC. Human-human communication has a higher chance of success when VC and NVC work in harmony together.

Table 1 provides a list of examples of different interaction channels, involved modalities from a processing and presentation mode.

3.2 Modality Principles

Mayer and Moreno [40] formalized the cognitive theory of multimedia learning, defined as learning from text and images, based upon three principles, *(i) dual-channel*: human mind uses two processing channels, for individual processing of visual/pictorial information; *(ii) limited capacity*: each channel has limited capacity for processing information; and *(iii) active processing*: that involves the coordination of the cognitive processes (selecting and organizing relevant words and pictures into coherent representations and integrating them with prior knowledge).

According to Mayer and Moreno [40] (see left side Figure 1), when words and pictures are presented to the learner, they enter via sensory organs (ears and eyes). Based on the learner's attention level, she captures some of the auditory and visual sensations, labeled as "selecting words/images". The learner constructs a coherent verbal and pictorial representation from the incoming words and images represented with "organizing words/images". Finally, the verbal and visual models are mentally connected and with relevant prior knowledge activated from long-term memory. Meaningful learning from words and images happens when the learner engages in these central cognitive processes during learning.

Example: In an information-seeking scenario, where animation is used together with the concurrent subtitles (dual-channel), this creates the so-called *split-attention effect*, which means, due to limited processing capacity, the learner's visual attention is split between viewing the animation and reading the on-screen text. The situation can be remedied if the text is *narrated* by the animation since it would off-load the visual channel into the auditory channel, increasing the effective working memory capacity. Furthermore, according to active processing theory, the best multimedia learning occurs when cognitive processing occurs both in the verbal and non-verbal channels. We can observe that multi-modality in presentation mode, i.e., verbal and non-verbal, improves multimedia learning quality.

4 MMCIS: DEFINITIONS AND APPLICATIONS

Recent conversational approaches to finding information through dialogue have mainly been viewed from natural language interactions (i.e., search in which input and output are mediated via "conversations" [33, 51, 60]). Even though several mentions towards conversational multi-modal systems suggested an expansion in the in- and output options [28, 46], limited research has been devoted to interacting with IS systems in a range of modalities which complement each other. We aim to address this.

4.1 Multi-Modality in CIS

Section 3 introduced modality types in multi-modal systems. However, multi-modality becomes even more complex when it comes to MMCIS systems, mainly due to the *multi-turn* and *information access* nature of MMCIS systems. Therefore, given the multi-modality basics presented earlier, we define MMCIS as follows.

Definition 4.1 (Multi-modality in CIS). Multi-modality in CIS can be defined based on three dimensions: (I) processing modality in

Table 1: Comparison of different modality types (refer to Section 3.1). Input/Output represent the input and output channels to the system. The information was originally obtained from Glinert and Blattner [25], Turk [61], and extended thereby.

Input/ Output	Interaction channel	Processing modality					Presentation mode		
		Visual	Textual	Audio	Touch	Others	Verbal		Non-verbal
							Vocal	Non-vocal	Non-vocal
I	Structured layouts (forms, lists)		✓					✓	
I	Speech		✓	✓			✓		
I	Facial expression, gestures, lip reading	✓							✓
I	Emotion recognition from EEG					✓			✓
I	Eye/gaze tracking	✓							✓
I	Pressure				✓				✓
I	Interactive map	✓			✓				✓
I	Motion capture (non-visual)					✓			✓
O	Animation + on-screen text (subtitle)	✓	✓					✓	✓
O	Narrating animation	✓	✓	✓			✓		✓

* Note that although speech is by definition a multi-modal signal (spoken words+atmosphere noise), we refer to it as a uni-modal signal.
** The vocal element of NVC may include tone of the voice, sighs and other vocal qualities, which are not used in this table.

conversation (C), (II) multi-modality in user-system interaction ($\mathcal{I}$), and (III) multi-modality in processing and accessing information items ($\mathcal{D}$). Therefore, multi-modality in each MMCIS system can be formally represented as:

$$\text{Multi-Modality in MMCIS} = C + \mathcal{I} + \mathcal{D} \tag{1}$$

We define the multi-modality dimensions C, $\mathcal{I}$, and $\mathcal{D}$ as follows:

Dimension I: Processing Modality in conversation (C). Let $C = [c_1, c_2, \cdots, c_m]$ denote a conversation with m conversation interactions, between the user and the system, where c_i contains all the information about the i^{th} interaction, including the actor (user(s) or system(s)), the content, and the context (e.g., time, location, or device). We define multi-modality in processing the conversation C based on two concepts of processing modality alternation and combination as follows:

- **Processing modality alternation in conversation**: If each conversation interaction uses a single processing modality, but the processing modality between two adjacent interactions alters, then the conversation is multi-modal by processing modality alternation. Formally, the conversation C is multi-modal by alternation if both of the following conditions are satisfied.

$$\begin{cases} |\text{modality}_{\text{p}}(c_i)| = 1 & \forall i : 1 \leq i \leq m \\ \text{modality}_{\text{p}}(c_i) \neq \text{modality}_{\text{p}}(c_{i+1}) & \exists i : 1 \leq i < m \end{cases} \tag{2}$$

  where $\text{modality}_{\text{p}}(\cdot)$ denotes the processing modality of a given conversation interaction. Note that the definition of processing is provided in Section 3. Example: a user asks a natural language question in a conversation and the system responds with an image.
- **Processing modality combination in conversation**: If one conversational interaction in a conversation consists of multiple processing modalities, then the conversation is multi-modal by combination. Formally, the conversation C is multi-modal by combination if:

$$|\text{modality}_{\text{p}}(c_i)| > 1 \qquad \exists i : 1 \leq i \leq m \tag{3}$$

  Example: a user selects an image of a plant and asks a question about the image, therefore, the interaction consists of two processing modalities.

In the above equations, uni- or multi-modality of a conversation interaction c_i are defined on the processing modalities reviewed in Section 3.1. If there is no (at least one) modality alternation or combination in the conversation C, then C is uni-modal (multi-modal) with respect to the first multi-modality dimension (C).

Dimension II: Multi-modality in user-system interaction ($\mathcal{I}$). Independent from the processing modality of conversational interactions in C, if the interaction channel or the result presentation mode (see Section 3.1) involves multiple interaction modalities, then the conversational information access system is multi-modal with respect to Dimension II or user-system interactions. Example: a user interacts with the system using a speech interface and/or a visual screen.

Dimension III: Multi-modality in processing and accessing information items ($\mathcal{D}$). If the information items used in the information access systems (for example, the documents that are retrieved or recommended) require different processing modalities (see Section 3.1) or the modality of the information items and the conversation interactions in C are different, then the system is multi-modal with respect to Dimension III ($\mathcal{D}$). Example: the system retrieves music in response to a keyword search query of the user in a multi-turn conversation.

In summary, C represents what information the system is receiving from the user during the conversation. Thus, C focuses on the *system-side* of the interaction in which multi-modality refers to processing modality. $\mathcal{I}$ represent all the interaction channels used by the user to interact with the system and by the system to interact with the user. $\mathcal{D}$ focuses on the information items and processing data. Thus, $\mathcal{D}$ has a system oriented view of the data collection and is centered on processing modality.

4.2 Why use MMCIS

In essence, conversations are multi-modal. Humans interact with their environment and fellow humans through a myriad of modalities. We talk, look, and touch our physical world concurrently. We observe a friend's tone of voice, facial expressions, and hand movements to understand which message they convey beyond the spoken language. All these different inputs provide us with a holistic view of the conversation's topic, our friend's sentiment towards the topic, or their mental state. However, all this information is lost in a simple interface which we typically navigate in a uni-modal way with a mouse, keyboard, or touchscreen. The simplicity of uni-modality leaves us as humans wanting more satisfactory, rich, and human-like interactions, which leads us to consider how multi-modal interactions can be incorporated into search.

In particular, for search formulations, users can expand their information need input modality from a typically written query to multi-modal input. The different ways of expressing an information need may overcome the difficulty to express our thoughts or information gap [5, 56].

4.2.1 Advantages of MMCIS. We highlight some advantages to searching over a multi-modal channel, incorporating **context**, overcoming **errors**, improving **learning**, and enhancing **accessibility**.

Context. Much work has been done to include context into search; however, advanced multi-modal interactions and data should increasingly be included in future models. In addition, intentionally incorporation contextual features enables the move from sequential (uni-modal) to parallel design (multi-modal). Furthermore, information needs can be expressed in more than keywords, facilitating the use of suitable modalities for particular needs and thus enabling to interact beyond spoken language. Instead of relying on the context which can be sensed by a system, users can actively disclose their context by sharing photos or videos as query (i.e., the context) including voice as a proxy of the user's mental or emotional state. The increased usage of context through multi-modal interactions also can be more human-like (including "socialness"), that is, because all the extra context features can be included in the model.

Errors and Accuracy. The multi-modal system can help overcome errors and increase accuracy from both input and output from systems. For input signals, multi-modal input can help with overcoming errors (i.e., multi-modal can deal with speech disfluencies better by combining speech recognition and lipreading [24, 69]). For output, a combination of speech and subtitles can overcome issues of presenting results in a noisy environment.

Learning. Since human learning is a complex multidimensional activity, it makes sense to consuming information through multi-modal interactions (see Section 3.2). This novel interaction mode can enhance users' different kinds of thinking and reasoning, adapting to the information seeker's needs.

Accessibility. People differ in capabilities, needs, or preferences. Even though a particular task can be achieved with a specific modality, providing users with several modalities and the opportunity to switch between modalities enhances equal information access. For example, a person with dyslexia may be very capable of typing a keyword; however, being able to verbalise it to a system may overcome spelling difficulties. Different modalities have different benefits, and it is often easier to point to an object instead of describing it. Finally, multi-modal output can adapt to the most informative mediums overcoming the limitations of a single-medium output [24] and thus making it more "natural" to interact with MMCIS systems.

4.3 When to Integrate Multiple Modalities and Conversations to support MMCIS

Natural language statements or short queries are not always suitable to search. Hence, MMCIS is suitable in the following conditions:

- the person who is searching has **device(s)** available which allow for more than one interaction mode (multi-device and multi-modal),
- when the task's **context** is important and can be captured with a device in a suitable modality enhancing personalization,
- when **task complexity** can be supported by the mode of device interaction,
- when the results can be returned in an appropriate **output modality** given the device, context, and complexity.

Next, we illustrate these differences and expand on when MMCIS may not be or less suitable.

Example task 1: Classic information retrieval. Let us consider a classic IR task of the Robust TREC 2004 topic. This task has been identified by Bailey et al. [4] as a medium complexity task as per Taxonomy of Learning objectives [3]. Let us imagine a user is in their home environment at their desk while listening to a podcast.

- `Topic ID`: 314
- `Title`: Marine Vegetation
- `Description`: Commercial harvesting of marine vegetation such as algae, seaweed and kelp for food and drug purposes.

Since our user has access to a desktop (device) at their desk (context), and the task involves more than one keyword search (complexity), it makes sense to conduct the search over a keyword browser-based retrieval. Nevertheless, the engine may pro-actively converse to narrow down or specify the search [58, 65]. Furthermore, depending on *how* the user anticipates the results, the browser could use visual or auditory information to satisfy the information need (e.g., video fragments from news broadcasts). Thus, making use of the multi-modal response presentation techniques and conversational feedback loops to refine results.

Example task 2: On-the-go and longitudinal information seeking. Let us revisit the example stated in Section 1, in which a person was cycling (non-stationary context), had multiple mobile electronic devices available, and several information needs with ranging complexity. The user instigated a conversation with the system by submitting numerous queries (photo and voice query) to build their information need. However, many of these information needs were not addressed on the spot while the person was cycling but instead queued until the user was able to receive the information through the appropriate modality. It was thus enabling the user to gather and access information over a long period.

Example task 3: Multi-party or collaborative search. The previous examples illustrate single-party information needs with one

or multiple devices and information needs. The following example demonstrates the advantages of multi-modal interactions for collaborative (or multi-party) CIS.

Imagine you are driving with some friends for a day out, suddenly your car has a warning sign indicating you need to stop and pull over. You and your friends start to search online by submitting the information need to diagnose the problem. You start submitting keywords via voice and photographs to a search engine, watch videos to establish the problem, and upload the car diagnostics to your dealer. Simultaneously, one of the friend's films and photographs the car and uploads it to a forum on cars to ask experts for advice while searching for alternative transportation to get back home. The other friend just hung up the phone with the car mechanic, who is now adding their search terms and found the information to the search session. The use of different devices to submit the multiple information needs to support the accomplishment of several tasks at the same time while triangulating, fusing, or "meshing" sources to improve overall recognition robustness [36, 50]. This complex information task allows you, your friends, and your car mechanic to corroborate information from multiple devices and modalities while supporting sense-making of the issue. Furthermore, this example illustrates the utility of both synchronous and asynchronous searching while the curated conversation with information through the system can alleviate the cognitive burden of the information-seeking task. Again, different modalities to consume the information may be appropriate at given times. Furthermore, we hypothesize that the systems will play an active role in selecting, organisation, and presenting the information as the knowledge expert.

Summary. As seen in the examples, different electronic devices (e.g., desktop, mobile device, smart speaker, or car) supporting multiple modalities, user contexts (stable or changing context), and task complexities (simple to complex) can be accommodated in a multi-modal conversation with information. The examples show that searching through the web with ever-growing corpora is not an easy task. MMCIS is inherently an interactive process and we believe that supporting cooperative user–system conversations will improve existing IR systems.

Scenarios Which May Not Invite For MMCIS. MMCIS may not always be a good idea and can add burden for simple information needs, especially when existing channels work well [7]. Furthermore, some disadvantage of the multi-modal conversations include strong computational and continuous network capabilities. That is, users may not be able to search when they are not connected over the internet. On a user level, there is the possibility of cognitive overload users by too much information, device-switching, or modality changes. Challenges such as privacy concerns about all the gathered data need to be investigated. Furthermore, on a more technical level, multi-modal research is expensive to collect, build, and test. It has been suggested that interdisciplinary know-how is needed to optimise the complex and challenging needs of a multi-modal system [31].

5 MMCIS CHALLENGES

In this section, we discuss challenges in designing and building MMCIS systems. These challenges are divided into five categories.

5.1 Multi-Modal Conversational Interactions

Multi-modal conversational input interactions result in several research challenges that do not exist or has been overlooked in uni-modal conversational systems which are addressed next.

Designing devices that support different interaction channels. Each interaction channel requires unique sensors, processing units, and user interfaces. Some of them are common in existing devices, such as smartphones and speech-only intelligent assistants. However, there exist several multi-modal interactions that are not supported by current devices. For instance, even though eye-tracking is extensively used for understanding user interactions research [6], it has not been widely deployed in popular devices. We believe that advancing technology, reducing deployment cost, and increasing applications are the key to introducing new interaction channels to the devices for MMCIS. One can imagine that interacting with virtual and augmented reality devices for information seeking purposes becomes widely accessible soon.

Recognizing interactions. MMCIS systems should recognize multi-modal interactions. For instance, for speech interactions, it is often difficult to find the answer to the user request in the form of speech signals, and this is why ASR is used to transcribe speech interactions. Different interaction modalities require unique models for recognizing the interactions, and developing these models is necessary to advance MMCIS research. An important research question is what representation should be used for different interaction modalities. Converting all modalities to text is not optimal and producing a unified representation across modalities seems an unavoidable research challenge. Moreover, some modalities are context-rich and useful information can be inferred from the context. For instance, the background noise in speech interactions can provide useful information, e.g. user's location. Similarly, speech can provide rich cues about user's emotion and sentiment, a useful piece of information omitted in text-based CIS. Extracting and inferring such information from context introduce new research problems for MMCIS systems. Note that using such information may have privacy implications and user consent may be required.

Correcting the recognized interactions and error mitigation. Automatic recognition of multi-modal interactions is not errorless. Different techniques, such as language modeling or computing the probability of observing each recognized interaction, are required for correcting these mistakes. If the system cannot accurately correct the recognized interaction, it may bring the user in the loop by asking a clarifying question, e.g., see [54].

Discoverability of interaction channels. Like most new technologies, users can be educated on how and when to use different interaction channels in the conversation. This can be simply ignored with the hope that users will discover the capabilities of the system themselves. However, different models can be developed to make this process more efficient, thus resolving the tension between exploration and instructions [47].

5.2 Multi-Modal Conversational Understanding

Conversational understanding in information seeking conversations refers to a process of accurate representation of user information need in a multi-turn user–system conversation. Topic tracking, co-reference, and ellipsis resolutions are major challenges in conversation understanding [2]. Different levels of multi-modality in conversation, as explained in Section 3.1, makes conversation understanding tasks challenging. We review a number of these challenges that should be addressed for developing successful MMCIS systems:

Resolving co-references and ellipsis across modalities. Co-reference resolution, i.e., finding all expressions that refer to the same entity, and ellipsis resolution, i.e., identifying all omissions from a clause of one or more words, are at the core of conversation understanding. Existing work mainly focuses on co-reference and ellipsis resolutions from a conversation in the form of text [37]. These problems also exist in multi-modal conversations and yet to be appropriately explored.

Multi-modal query rewriting. Rewriting the last user request within the context of the existing conversation in order to produce a history-independent request (query) is one of the common tasks in conversation understanding. Multi-modal query rewriting models should be able to draw the connection across modalities that appeared in the conversation. For instance, modelling such connections, the connection between the text and images used in a conversation, is a challenging and essential task to be explored.

Learning conversation representation across modalities. Conversation understanding models are mainly trained based on user–system interactions. Learning from different conversations, each in different modalities, is a challenging task. A straightforward solution is to train different models for each modality. However, this is not an optimal solution. Transferring knowledge across modalities is an exciting and essential challenge in MMCIS systems.

Conversation understanding for cold-start modalities. Advancing in technology leads to the development of new sensors, devices, and interfaces, and thus new interaction modalities. Adding a new modality to an existing MMCIS system is yet another challenge in conversation understanding. We call this problem studying cold-start modalities and this can be another case of transferring knowledge across modalities.

5.3 Multi-Modal Conversational Ranking and Generation

Multi-modality further results in various research challenges in conversational result ranking and generation. They include computing the similarity between conversation representation and the retrieved items. If the modality of items in the collection is different from the conversation modality, the MMCIS system should bridge this gap by learning shared representations or transforming one modality to another. Due to the nature of retrieval tasks, these solutions should be efficient and scalable. Besides, generating multi-modal results requires developing new generative models [19] to keep the connection between different modalities in a generation.

5.4 Multi-Modal Response Presentation

As frequently mentioned throughout the paper, conversational systems create challenging research problems related to result presentation. Here we review these challenges:

Selecting output modality. In the case of multiple output modalities, deciding which modality to use for results presentation is important. Selecting the output modality can depend on the type of request and response, user preferences, system properties (e.g., screen size), and situational context (e.g., using speech outputs while driving a car).

Changing the modality of the retrieved or generated response. If the selected output modality is different from the retrieved or generated response, a model should be employed to convert its modality to the selected one. Some examples include automatic speech generation (converting text to speech), generating text from images and diagrams and vice versa.

Response presentation in multiple modalities. A response can be presented in multiple different modalities. For example, the response to a user request may be an image (or diagram) in addition to a text or speech description of that image (or diagram). Presenting the results using multiple modalities may need further research in terms of user interface and response ranking and generation.

5.5 MMCIS Evaluation Challenges

Evaluating IIR models is challenging. The reusable test collections for CIS tasks are built based on several simplifying assumptions about the system abilities and user behavior. For instance, the TREC Conversational Assistance Track [15] assumes that users always ask related natural language questions in each session and the system can only retrieve several passages. As another example, the Qulac dataset [1] considers clarifying questions in response to search queries with the assumption that users always submit a single keyword query in each session. Such assumptions do not often hold in real-life settings. This is why online evaluation of CIS systems is critical. However, large-scale online evaluation is expensive and time-consuming and is only accessible to a small fraction of researchers. Therefore, building reusable test collections is still one of the most important parts of CIS research. All the mentioned facts are relevant to all types of CIS systems, including MMCIS. That being said, evaluating MMCIS systems requires some unique properties. They include evaluating the system's ability to *(i)* represent and utilize different input modalities and item modalities in the collection, *(ii)* present the responses in different modalities, and *(iii)* select the most appropriate modality for response presentation. Little is known about the guidelines for evaluating MMCIS systems. Methodologies used for creating reusable test collections presented in previous work for CIS research [1, 15, 48, 65, 67, 68], user simulation [66], and online evaluation [30, 34] are generally applicable to MMCIS evaluation.

6 A PLATFORM FOR MMCIS RESEARCH

We introduce a platform for developing and evaluating MMCIS models by extending the Macaw's platform [64] to handle multiple interaction and processing modalities. Macaw is an open-source

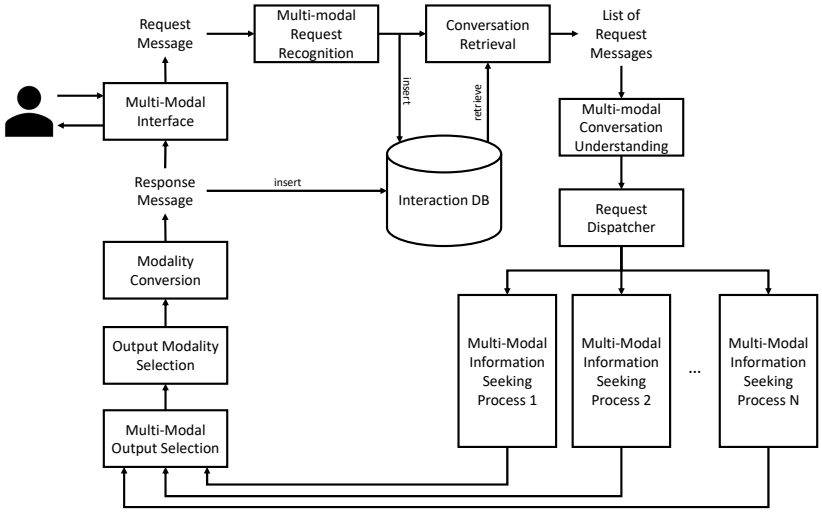

Figure 2: A high-level architecture for the Macaw-MMCIS platform.

platform implemented in Python for CIS research. It uses the Telegram interface that supports multi-modal interactions, however, its original architecture fails at handling different types of multi-modal interactions. This has motivated us to extend its architecture by multi-modal components. We open-source our implementation, called Macaw-MMCIS, for research purposes.[2]

The high-level architecture of Macaw-MMCIS is depicted in Figure 2. The new components, with respect to the Macaw's original architecture include:

- Multi-modal Request Recognition: The goal of this component is to convert the user's Request Message to a unified format, e.g., converting everything to text or producing a high-dimensional latent shared representation for all modalities. The output of this component is a Message that extends the Request Message with the recognized unified representation. Note that each Message in Macaw is a JSON object.
- Multi-modal Conversation Understanding: Once the whole conversation history for the current request is retrieved, this component is responsible for helping the downstream information-seeking tasks utilize the conversation. This may include multi-modal co-reference and ellipsis resolution and/or learning a joint representation from the given conversation.
- Multi-Modal Information Seeking Processes: These components are responsible for response retrieval or generation. Each may use different algorithms or perform additional information seeking tasks. Note that these components are equivalent to Actions in the original Macaw's architecture with the difference that they can retrieve or generate multi-modal items.
- Output Modality Selection: This component is a classifier that selects the modality that should be used for presenting the produced response to the user.
- Modality Conversation: If the selected output is different from the produced response modality, this component converts the response's modality and returns a Response Message object. This component can be as simple of speech generation when the produced result is `text` and the selected modality is `speech`.

For the rest of components, we refer the reader to Zamani and Craswell [64], as they have not been significantly modified from their original implementations in Macaw. The developed platform supports different modalities, including text, speech, image, and video, in addition to other interaction channels such as click.

7 CONCLUSION AND FUTURE WORK

This perspectives paper explored information-seeking through multi-modal conversations. We consolidated existing research on multi-modal interactions and CIS to define MMCIS. We implemented a clear definition of modality types (i.e., interaction channel, processing modality, and presentation mode), including examples enabling a proper understanding of multi-modality and its significance. We consider the contribution of the MMCIS definition and dimensions (i.e., multi-modality in conversations, user–system interactions, and processing and accessing information items) to support the research and development for dedicated MMCIS systems. Furthermore, by illustrating possible MMCIS tasks and research challenges, we identified conditions in which this emerging interaction paradigm is suitable. Additionally, we released a research platform for MMCIS. To the best of our knowledge, no formal MMCIS definitions have been proposed. Finally, overcoming the identified research challenges with our proposed MMCIS research platform, MACAW-MMCIS, is a critical extension of this work. We argue that this highly multi-disciplinary research can bridge research communities and positively impact information accessibility.

8 ACKNOWLEDGEMENTS

This work was supported in part by the Center for Intelligent Information Retrieval. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

[2]Macaw-MMCIS is integrated into the Macaw project and is available at https://github.com/hamed-zamani/macaw.

REFERENCES

[1] Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, and W. Bruce Croft. 2019. Asking Clarifying Questions in Open-Domain Information-Seeking Conversations. In *SIGIR*. ACM, 475–484.

[2] Avishek Anand, Lawrence Cavedon, Hideo Joho, Mark Sanderson, and Benno Stein. 2020. Conversational search (dagstuhl seminar 19461). In *Dagstuhl Reports*, Vol. 9. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

[3] Lorin W. Anderson, David R. Krathwohl, and B. S. Bloom. 2001. *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. Longman.

[4] Peter Bailey, Alistair Moffat, Falk Scholer, and Paul Thomas. 2015. User variability and IR system evaluation. In *Proceedings of Conference on Research and Development in Information Retrieval (SIGIR)*. 625–634.

[5] Nicholas J. Belkin. 1980. Anomalous states of knowledge as a basis for information retrieval. *Canadian Journal of Information Science* 5, 1 (1980), 133–143.

[6] Valeria Bolotova, Vladislav Blinov, Yukun Zheng, W. Bruce Croft, Falk Scholer, and Mark Sanderson. 2020. Do People and Neural Nets Pay Attention to the Same Words: Studying Eye-Tracking Data for Non-Factoid QA Evaluation. In *Proceedings of International Conference on Information and Knowledge Management (CIKM)*. 85–94.

[7] Lawrence Cavedon, Bernd Fröhlich, Hideo Joho, Ruihua Song, Jaime Teevan, Johanne Trippas, and Emine Yilmaz. 2020. Scenarios that Invite Conversational Search. In *Conversational Search (Dagstuhl Seminar 19461)*, Avishek Anand, Lawrence Cavedon, Hideo Joho, Mark Sanderson, and Benno Stein (Eds.). Dagstuhl, 66–69.

[8] Jingjing Chen, Chong-Wah Ngo, Fuli Feng, and Tat-Seng Chua. 2018. Deep Understanding of Cooking Procedure for Cross-modal Recipe Retrieval. In *2018 ACM Multimedia Conference on Multimedia Conference, MM 2018, Seoul, Republic of Korea, October 22-26, 2018*. ACM, 1020–1028.

[9] Leigh Clark, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, Cosmin Munteanu, Vincent Wade, and Benjamin R. Cowan. 2019. What Makes a Good Conversation? Challenges in Designing Truly Conversational Agents. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. 1–12.

[10] W. Bruce Croft. 2019. The Importance of Interaction for Information Retrieval. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'19)*. 1–2.

[11] W. B. Croft and R. H. Thompson. 1987. I3R: A new approach to the design of document retrieval systems. *Journal of the American Society for Information Science* 38, 6 (1987), 389–404.

[12] Paul A. Crook, Shivani Poddar, Ankita De, Semir Shafi, David Whitney, Alborz Geramifard, and Rajen Subba. 2019. SIMMC: Situated Interactive Multi-Modal Conversational Data Collection And Evaluation Platform. *CoRR* abs/1911.02690 (2019).

[13] Chen Cui, Wenjie Wang, Xuemeng Song, Minlie Huang, Xin-Shun Xu, and Liqiang Nie. 2019. User Attention-guided Multimodal Dialog Systems. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*. ACM, 445–454.

[14] Peng Cui, Zhiyu Wang, and Zhou Su. 2014. What Videos Are Similar with You?: Learning a Common Attributed Representation for Video Recommendation. In *Proceedings of the ACM International Conference on Multimedia, MM '14,*, Kien A. Hua, Yong Rui, Ralf Steinmetz, Alan Hanjalic, Apostol Natsev, and Wenwu Zhu (Eds.). ACM, 597–606.

[15] Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. 2019. CAsT 2019: The Conversational Assistance Track Overview. In *Proceedings of the Twenty-Eighth Text REtrieval Conference, TREC 2019, Gaithersburg, Maryland, USA*.

[16] Yashar Deldjoo, Mihai Gabriel Constantin, Hamid Eghbal-Zadeh, Bogdan Ionescu, Markus Schedl, and Paolo Cremonesi. 2018. Audio-visual encoding of multimedia content for enhancing movie recommendations. In *Proceedings of the 12th ACM Conference on Recommender Systems*. 455–459.

[17] Yashar Deldjoo, Maurizio Ferrari Dacrema, Mihai Gabriel Constantin, Hamid Eghbal-zadeh, Stefano Cereda, Markus Schedl, Bogdan Ionescu, and Paolo Cremonesi. 2019. Movie genome: alleviating new item cold start in movie recommendation. *User Model. User Adapt. Interact.* 29, 2 (2019), 291–343.

[18] Yashar Deldjoo, Tommaso Di Noia, Daniele Malitesta, and Felice Antonio Merra. 2021. A Study on the Relative Importance of Convolutional Neural Networks in Visually-Aware Recommender Systems. In *CVPRW-CVFAD 2021 :The 4th CVPR Workshop on Computer Vision for Fashion, Art, and Design*. CVPR Proceedings.

[19] Yashar Deldjoo, Tommaso Di Noia, and Felice Antonio Merra. 2021. A survey on adversarial recommender systems: from attack/defense strategies to generative adversarial networks. *ACM Computing Surveys (CSUR)* 54, 2 (2021), 1–38.

[20] Yashar Deldjoo, Markus Schedl, Paolo Cremonesi, and Gabriella Pasi. 2020. Recommender Systems Leveraging Multimedia Content. *ACM Comput. Surv.* 53, 5 (2020), 106:1–106:38.

[21] J. Devlin, M. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proc. of NAACL*.

[22] Aleksandr Farseev, Ivan Samborskii, Andrey Filchenkov, and Tat-Seng Chua. 2017. Cross-Domain Recommendation via Clustering on Multi-Layer Graphs. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 195–204.

[23] Marlen Fröhlich, Christine Sievers, Simon W Townsend, Thibaud Gruber, and Carel P van Schaik. 2019. Multimodal communication and language origins: integrating gestures and vocalizations. *Biological Reviews* 94, 5 (2019), 1809–1829.

[24] Dafydd Gibbon, Inge Mertins, and Roger K Moore. 2012. *Handbook of multimodal and spoken dialogue systems: Resources, terminology and product evaluation*. Vol. 565. Springer Science & Business Media.

[25] Ephraim P. Glinert and Meera Blattner. 1996. Multimodal Interaction. *IEEE Multim.* 3, 4 (1996), 13.

[26] Iva Gornishka, Stevan Rudinac, and Marcel Worring. 2019. Interactive Search and Exploration in Online Discussion Forums Using Multimodal Embeddings. *CoRR* abs/1905.02430 (2019).

[27] Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable Feature Learning for Networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 855–864.

[28] Alex Hauptmann, Joao Magalhaes, Ricardo G. Sousa, and Joao Paulo Costeira. 2020. MuCAI'20: 1st International Workshop on Multimodal Conversational AI. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20)*. 4767–4768.

[29] Alejandro Jaimes and Nicu Sebe. 2007. Multimodal human–computer interaction: A survey. *Computer vision and image understanding* 108, 1-2 (2007), 116–134.

[30] Jiepu Jiang, Ahmed Hassan Awadallah, Rosie Jones, Umut Ozertem, Imed Zitouni, Ranjitha Gurunath Kulkarni, and Omar Zia Khan. 2015. *Automatic Online Evaluation of Intelligent Assistants*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 506–516.

[31] Kristiina Jokinen and Antti Raike. 2003. Multimodality–technology, visions and demands for the future. In *Proceedings of the 1st Nordic Symposium on Multimodal Interfaces*. 239–251.

[32] Marius Kaminskas, Francesco Ricci, and Markus Schedl. 2013. Location-aware music recommendation using auto-tagging and hybrid matching. In *Seventh ACM Conference on Recommender Systems, RecSys '13, Hong Kong, China, October 12-16, 2013*. ACM, 17–24.

[33] Johannes Kiesel, Arefeh Bahrami, Benno Stein, Avishek Anand, and Matthias Hagen. 2018. Toward Voice Query Clarification. In *Proceedings of Conference on Research and Development in Information Retrieval (SIGIR)*. 1257–1260.

[34] Julia Kiseleva, Kyle Williams, Jiepu Jiang, Ahmed Hassan Awadallah, Aidan C. Crook, Imed Zitouni, and Tasos Anastasakos. 2016. Understanding User Satisfaction with Intelligent Assistants. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval (CHIIR '16)*. 121–130.

[35] Multimodal Interaction Lab. 2015. Multimodal Interaction - Human Car Interaction. http://humancarinteraction.com/multimodal-interaction.html. (2015). (Last accessed February 4, 2021).

[36] Jennifer Lai. 2000. Conversational Interfaces. *Commun. ACM* 43, 9 (Sept. 2000), 24–27.

[37] Xiaolong Li and Kristy Boyer. 2016. Reference Resolution in Situated Dialogue with Learned Semantics. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, Los Angeles, 329–338. DOI: http://dx.doi.org/10.18653/v1/W16-3642

[38] Lizi Liao, Yunshan Ma, Xiangnan He, Richang Hong, and Tat-Seng Chua. 2018. Knowledge-aware Multimodal Dialogue Systems. In *2018 ACM Multimedia Conference on Multimedia Conference, MM 2018, Seoul, Republic of Korea, October 22-26, 2018*. ACM, 801–809.

[39] Richard E. Mayer. 2005. *The Cambridge handbook of multimedia learning, 1st Edition*. Cambridge Univ. Press. http://www.worldcat.org/oclc/57526976

[40] Richard E Mayer and Roxana Moreno. 2003. Nine ways to reduce cognitive load in multimedia learning. *Educational psychologist* 38, 1 (2003), 43–52.

[41] Julian J. McAuley, Christopher Targett, Qinfeng Shi, and Anton van den Hengel. 2015. Image-Based Recommendations on Styles and Substitutes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, Santiago, Chile, August 9-13, 2015*. ACM, 43–52.

[42] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[43] Liqiang Nie, Wenjie Wang, Richang Hong, Meng Wang, and Qi Tian. 2019. Multimodal Dialog System: Generating Responses via Adaptive Decoders. In *Proceedings of the 27th ACM International Conference on Multimedia, MM 2019, Nice, France, October 21-25, 2019*. ACM, 1098–1106.

[44] R. N. Oddy. 1977. Information retrieval through man-machine dialogue. *Journal of Documentation* 33, 1 (1977), 1–14.

[45] Sergio Oramas, Oriol Nieto, Mohamed Sordo, and Xavier Serra. 2017. A Deep Multimodal Approach for Cold-start Music Recommendation. In *Proceedings of the 2nd Workshop on Deep Learning for Recommender Systems, DLRS@RecSys 2017, Como, Italy, August 27, 2017*. ACM, 32–37.

[46] Sharon Oviatt and Laure Soulier. 2020. Conversational Search for Learning Technologies. In *Conversational Search (Dagstuhl Seminar 19461)*, Avishek Anand, Lawrence Cavedon, Hideo Joho, Mark Sanderson, and Benno Stein (Eds.).

Dagstuhl, 69–74.

[47] Stephen J Payne. 2007. Mental models in human-computer interaction. *The Human-Computer Interaction Handbook* 20071544 (2007), 63–76.

[48] Gustavo Penha and Claudia Hauff. 2020. Challenges in the evaluation of conversational search systems. In *KDD 2020 Workshop on Conversational Systems Towards Mainstream Adoption (KDD-Converse 2020)*, Vol. 2666. CEUR-WS.

[49] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[50] Martin Porcheron, Joel E. Fischer, Stuart Reeves, and Sarah Sharples. 2018. Voice Interfaces in Everyday Life. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, 1–12.

[51] Filip Radlinski and Nick Craswell. 2017. A Theoretical Framework for Conversational Search. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval (CHIIR '17)*. 117–126.

[52] Roope Raisamo. 1999. *Multimodal Human-Computer Interaction: a constructive and empirical study*. Tampere University Press.

[53] Amrita Saha, Mitesh M. Khapra, and Karthik Sankaranarayanan. 2018. Towards Building Large Scale Multimodal Domain-Aware Conversation Systems. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18)*. AAAI Press, 696–704.

[54] Svetlana Stoyanchev, Alex Liu, and Julia Hirschberg. 2014. Towards Natural Clarification Questions in Dialogue Systems. In *AISB '14*, Vol. 20.

[55] Pongsate Tangseng and Takayuki Okatani. 2020. Toward Explainable Fashion Recommendation. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2020, Snowmass Village, CO, USA, March 1-5, 2020*. IEEE, 2142–2151.

[56] Robert S. Taylor. 1962. The process of asking questions. *American Documentation* 13, 4 (1962), 391–396.

[57] Paul Thomas, Daniel McDuff, Mary Czerwinski, and Nick Craswell. 2017. MISC: A data set of information-seeking conversations. In *SIGIR 1st International Workshop on Conversational Approaches to Information Retrieval (CAIR'17)*. 6 pages.

[58] Johanne R. Trippas, Damiano Spina, Lawrence Cavedon, Hideo Joho, and Mark Sanderson. 2018. Informing the Design of Spoken Conversational Search. In *Proceedings of Conference on Information Interaction and Retrieval (CHIIR)*. 32–41.

[59] Johanne R. Trippas, Damiano Spina, Mark Sanderson, and Lawrence Cavedon. 2015. Towards Understanding the Impact of Length in Web Search Result Summaries over a Speech-only Communication Channel. In *Proceedings of Conference on Research and Development in Information Retrieval (SIGIR)*. 991–994.

[60] Johanne R. Trippas, Damiano Spina, Paul Thomas, Mark Sanderson, Hideo Joho, and Lawrence Cavedon. 2020. Towards a Model for Spoken Conversational Search. *Information Processing & Management* 57, 2 (2020), 102162.

[61] Matthew A. Turk. 2014. Multimodal interaction: A review. *Pattern Recognit. Lett.* 36 (2014), 189–195.

[62] Alexandra Vtyurina, Charles LA Clarke, Edith Law, Johanne R. Trippas, and Horatiu Bota. A Mixed-Method Analysis of Text and Audio Search Interfaces with Varying Task Complexity. In *Proceedings of International Conference on the Theory of Information Retrieval (ICTIR)*. 61–68.

[63] Suhang Wang, Yilin Wang, Jiliang Tang, Kai Shu, Suhas Ranganath, and Huan Liu. 2017. What Your Images Reveal: Exploiting Visual Contents for Point-of-Interest Recommendation. In *Proceedings of the 26th International Conference on World Wide Web, WWW 2017*. ACM, 391–400.

[64] Hamed Zamani and Nick Craswell. 2020. Macaw: An Extensible Conversational Information Seeking Platform. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20)*. Association for Computing Machinery, 2193–2196.

[65] Hamed Zamani, Susan T. Dumais, Nick Craswell, Paul N. Bennett, and Gord Lueck. 2020. Generating Clarifying Questions for Information Retrieval. In *Proceedings of the 29th International Conference on World Wide Web (WWW '20)*.

[66] Shuo Zhang and Krisztian Balog. 2020. Evaluating Conversational Recommender Systems via User Simulation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '20)*. Association for Computing Machinery, 1512–1520.

[67] Shuo Zhang, Zhuyun Dai, Krisztian Balog, and Jamie Callan. 2020. *Summarizing and Exploring Tabular Data in Conversational Search*. 1537–1540.

[68] Yongfeng Zhang, Xu Chen, Qingyao Ai, Liu Yang, and W. Bruce Croft. 2018. Towards Conversational Search and Recommendation: System Ask, User Respond. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM '18)*. Association for Computing Machinery, 177–186.

[69] Victor W. Zue and James R. Glass. 2000. Conversational interfaces: advances and challenges. *Proc. IEEE* 88, 8 (2000), 1166–1180.