Gist and Verbatim: Understanding Speech to Inform New Interfaces for Verbal Text Composition

Brinda Mehra
City University of Hong Kong
Hong Kong, China
University of Michigan
Ann Arbor, Michigan, United States
brinda@umich.edu

Kejia Shen
ChengDu Planning Information Technology Center
ChengDu, China
City University of Hong Kong
Hong Kong, China
kejiashen920@gmail.com

Hen-Chen Yen
City University of Hong Kong
Hong Kong, China
University of Waterloo
Waterloo, Ontario, Canada
ryanyen2-c@my.cityu.edu.hk

Can Liu*
City University of Hong Kong
Hong Kong, China
canliu@cityu.edu.hk

ABSTRACT

Recent interest in speech-to-text applications has found speech to be an efficient modality for text input. However, the spontaneity of speech makes direct transcriptions of spoken compositions effortful to edit. While previous works in Human-Computer Interaction (HCI) domain focus on improving error correction, there is a lack of theoretical ground around the understanding of speech as an input modality. This work explores literature from Cognitive Science to synthesize relevant theories and findings for the HCI audience to reference. Motivated by the literature indicating a fast memory decay of speech production and a preference towards gist abstraction in memory traces, an experiment was conducted to observe users' immediate recall of their verbal composition. Based on the theories and findings, we introduce new interaction concepts and workflows that adapt to the characteristics of speech input.

CCS CONCEPTS

• **Human-centered computing → Interaction paradigms**; **HCI theory, concepts and models**; **Interaction techniques**.

KEYWORDS

speech, speech-to-text, STT, text composition, text editing, dictation, text entry

ACM Reference Format:
Brinda Mehra, Kejia Shen, Hen-Chen Yen, and Can Liu. 2023. Gist and Verbatim: Understanding Speech to Inform New Interfaces for Verbal Text Composition. In *ACM conference on Conversational User Interfaces (CUI '23), July 19–21, 2023, Eindhoven, Netherlands.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3571884.3597134

1 INTRODUCTION

Recent advances in automatic speech recognition (ASR) technology have reignited interest in speech as an input modality for text production - particularly in mobile scenarios and have appeared as both stand-alone products (Otter.ai, Dragon NaturallySpeaking) or integrated into existing offerings (Microsoft Word, Notes) [69]. While the spontaneous, rapid, and incorporeal nature of speech allows for an increased level of mobility both in the context of the environment it is produced in and the content itself, the very same characteristics may also be acting as barriers to the mass adoption of speech-to-text (STT) as an input modality.

A major barrier is the fact that STT systems transcribe spoken output literally, often causing the transcribed text to resemble a stream-of-consciousness piece rather than a coherent piece of text. The structure of spoken output inherently differs from the more rigid skeleton followed in traditional writing and often, the produced output must undergo heavy editing before it can be shared or used as a store of information [35]. This is further compounded by the arduous process of editing on current STT systems, which involves users locating, reviewing, and editing any incorrectly transcribed text and removing any discourse markers [9].

Given this, it is natural that the bulk of research focused on STT systems and speech as an input modality has focused on developing the necessary technical infrastructure or using speech as a supporting or replacement input modality. The majority of current literature focused on STT systems explores its use for specific scenarios related to language or disability support or is focused on advancing editing mechanisms through natural language processing or machine learning [34]. As a consequence, however, there has been fairly limited research focusing on the design considerations surrounding the features and functions of STT systems that take into account fundamental characteristics of speech[2].

Building on previous works that warn against the direct transportation of traditional graphical user interface (GUI) design heuristics into the design and development of voice/speech-based ones,

[*] Corresponding Author

CUI '23, July 19–21, 2023, Eindhoven, Netherlands

ACM ISBN 979-8-4007-0014-9/23/07…$15.00
https://doi.org/10.1145/3571884.3597134

this paper focuses on addressing two key demands made in Clark et al's State of Speech in HCI – (a) developing qualitative or design-centered studies to design more intuitive speech-based interfaces and (b) summarizing theoretical understandings of speech as a way to inform interface design [12]. In line with Shneiderman's comments on the importance of understanding speech and its memory issues to design STT applications, we identify and attempt to provide a starting overview of relevant characteristics of speech for the HCI community [63].

Following an examination of speech as a text input modality and the state of the art in developing interface support for editing spoken text in HCI, this paper examines literature in cognitive science and linguistics to lay theoretical grounding for the properties of speech production and its recall and storage to circumvent current limitations of speech UIs. Through exploring how users store and recall their own compositions, this paper indicates support for more gist-level processing for the storage and recall of speech rather than exact, verbatim storage and recall. The implications of the finding are subsequently used to inform the research direction of new workflows for using speech to support fast and iterative drafting processes, instead of directly editing spoken text like writing. We challenge the direct transplant of text editor interfaces used for dictation and suggest a paradigm shift towards interacting with text based on its gist rather than verbatim representations.

We believe this paper provides three key contributions to the HCI community:

- Establishing a theoretical ground differentiating speaking from writing based on literature from cognitive science;
- Observing how the relevant theories manifest in an HCI context of verbal text composition through an experiment;
- Proposing new interaction concepts, workflows, and research directions for verbal composition and interaction with spoken content.

2 RELATED WORKS

In this section, we summarize previous research on dictation and efforts for improving text composition and editing with speech.

2.1 Dictation and Text Composition

Speech has long been considered a more "natural and primary" channel of interaction - being familiar, convenient, and producible in a variety of environmental conditions that make it interesting as an input modality for human-computer interaction [21, 44]. The "untethered and device-independent" nature of speech as a communication channel allows for speech to be used concurrently and separately as an input modality alongside tasks requiring mobility or in eyes- and hands-free scenarios [13, 24, 70].

STT systems have been used to support text composition through the transcription of dictated information in various settings such as healthcare [49], enhancing learning [62] and supporting individuals with disabilities [18, 67]. Speech is increasingly being considered as an attractive input modality for situations where screen size is limited or they need to move around in their environment [63, 74]. Speaking is also considered to be natural, and can be eyes-free and hands-free, allowing people to perform multiple tasks at the same time without significant cognitive effort [30]. Previous research showed that experienced writers found "using one's voice is more congenial than using one's fingers", and that the hands-and-eyes free capabilities allow them to think more fluidly and freely[31]. Using speech for text input is also fast, with humans being able to produce around 200 words per minute when speaking, compared to about 30-80 words per minute for handwriting or typing [40]. Additional research indicates speech has also been found to be three times as fast as typing for text entry on mobile devices across languages, allowing for a greater rate of text production than typing in the same amount of time [33, 55]. However, despite these advantages and the wide availability of dictation software on smartphones, 75% of users still prefer typing for text input [19]. The bottleneck in adopting dictation as a mode of text input seems to be caused by the loss of productivity gain caused by error correction, editing, and revision activities – with previous research indicating users spend 66% of their time correcting the speech recognition output on desktop dictation systems [35]. This suggests a need to closely examine existing editing mechanisms for these interfaces and identify possible issues and their solutions to better support text composition and production through dictation.

2.2 Re-Dictation as an Editing Technique

Although speech input has the advantage of providing eyes-and-hands-free experience, the need for editing transcribed text often brings users back to the keyboard. The steps and spatial referencing involved in text editing (awareness of the error, locating the error, correcting the error) often require heavy visual engagement, making speech suitable for text input but not necessarily for feedback and outcome evaluation [47, 60]. The spontaneous and ephemeral nature of speech does not naturally lend itself to identifying and delineating "where and how much of the text needs to be changed", thus requiring the support of other modalities like vision and touch [8, 24]. To address this bottleneck, there is an increasing interest in HCI on how speech can be leveraged not only as an input modality but also as a key modality for editing.

Recent works have developed various techniques to improve speech-based text editing using a multimodal approach. For example, EyeSayCorrect used eye gaze to select words and then allowed users to speak the new phrase or provided a series of options that could be selected on a touch-screen to replace the incorrectly transcribed word [75]. Gaze'N'Touch found that users often preferred to use gaze when selecting the text they want to edit, finding it less physically effortful than traditional touch-based selection but increasing gaze demand as users must remain focused on the screen [53]. Voice Typing used a marking menu that presents a touch-based list of alternative candidates for the selected word as well as the option to re-speak the text [39]. ReType proposed a new technique for common editing operations using gaze that allowed users to simultaneously use the keyboard [66]. While ReType was found to have both a better user experience and to beat the speed of mouse-based interaction for small text edits, it reduced the ability to use speech in mobile scenarios given the heavy gaze demand. Similarly, while EYEditor used voice to modify text for on-the-go text editing, a wearable ring-mouse was used for text navigation and selection on smart glasses [26]. Talk-and-Gaze proposed a method for using eye gaze to provide spatial information and select erroneous words or the selection of errors through specific voice commands [61]. Other examples of multimodal strategies for text editing with voice include combining handwriting or gesture typing with speech or using touch to indicate word boundaries [64, 65, 68].

However, there has been a concerted effort to develop editing mechanisms that can reduce the gaze demand when using STT systems and do not require users to switch between modalities, allowing them to compose and edit through speech. Two key strategies that were explored as potential mechanisms for editing via speech were (a) *Commanding* - where verbal commands like "Add" or "Delete" were used to carry out simple editing tasks; and (b) *Re-Dictation* - restating the original utterance to directly replace the erroneously transcribed text [73]. However, remembering commands placed an additional burden on cognitive resources while *Re-Dictation* was found to be both easier to carry out and more efficient for long, more complex edits [25, 27].

The original idea of *re-dictation* as an editing mechanism involves users simply selecting and restating their original utterance until it is transcribed correctly [42]. Recent research on *Re-dictation* has focused largely on the development of intelligent infrastructure to support the process. For example, Just Speak It [20] introduces neural networks and pre-trained models to understand user intentions based on semantics and allowed users to remove colloquial inserts automatically and edit by simply speaking out target words.

While much technical innovation has taken place to improve speech recognition accuracy, address speech repair and ease the effort of editing, the underlining workflow being used for speech-based text composition remains unchanged from current processes used in writing or typing. This inherently demands heavy visual engagement and precise content manipulation, which cripple most of the eyes-and-hands-free benefits offered by speech. In this work, we attempt to rethink the fundamental differences between speaking and writing/typing through examining literature from cognitive science and propose an alternative workflow for speech-based text composition by encouraging iterative re-speaking and reducing precise text editing based on our findings.

3 UNDERSTANDING SPEECH: PERSPECTIVES FROM COGNITIVE SCIENCE AND LINGUISTICS

Based on the issues defined in the related works, this section outlines relevant literature from cognitive science and linguistics to understand the theoretical differences in the properties and cognitive processes of speaking and writing as the first step to developing more intuitive interactions with speech-based interfaces.

3.1 Production and Feedback Mechanisms of Speaking versus Writing

Speaking and writing are both considered key channels for communication, and they share the same initial stages of cognitive processes, namely conceptual preparation. In this stage, thoughts are placed in a framework for verbalization and grammatical or phonological encoding, which involves the extraction of syntactic or semantic information from the mental lexicon [58].

However, once relevant information is extracted, the processes of speech and writing diverge as a function of the neurophysiological processes involved in the execution of the respective tasks. Speaking, for example, directly moves to a stage of vocal articulation, where phonetic information is converted into motor commands for the manipulation of the larynx, tongue, and jaw to produce audible sounds [48, 50]. In contrast, writing requires fine motor control of arms, wrists, and hands to produce graphemic representations of the content [54]. Similarly, typing requires the execution of motor commands through parsing words into characters and then typing them out through the keyboard in sequential order to create similar graphemic representations [57]. However, unlike speech, writing and typing involve a simultaneous and consecutive stage of "sub-vocalization", wherein motor commands convert graphemic representations into phonetic ones through motor commands to the larynx, tongue, and jaw that does not result in audible productions of speech but instead facilitates "internal speech" as a "sound code to assist word identification" and aid reading comprehension [17, 41, 43].

The presence of both motor execution and sub-vocal articulation makes it so that writing, for people with normal vision, inherently provides both visual and haptic feedback down to each letter during the process of composition as well as visual and motor memory traces. In contrast, the inherent feedback for speaking is only auditory and in phonetic units. Research comparing the different types of memory has found that auditory memory tends to be weaker than its visual or motor counterparts [29], with auditory representations often being less precise and detailed than visual memories in particular and deteriorating with greater speed over time [3, 28, 36].

As a result, even though speech is often considered the more "natural" medium of communication, previous research efforts have often focused on writing as the main modality for the communication and verbalization of experiences, largely because of the visual nature of the written modality [16, 46]. Unlike the ephemerality of speech, written output is attributed to have a certain amount of permanence because of the graphemic representation of the content, which can act as a "functional bridge" mapping phonology to orthography [7]. In contrast, the limited spatial radius and faster decay rate of speech do not provide any forms of visible retrieval support that persist across space and time and could act as a cue for recall [22, 32].

3.2 Organization and Recursivity

The spontaneous nature of speech and the lack of time for elaborate pre-planning that forces people to "activate ideas off the top of their head" [14] reduces the ability for any alteration or correction of the produced utterance. In contrast, the slower process of writing, with a production rate of 40 words per minute (WPM) against the 200 WPM for speech, allows for the revision and reshaping of the composed text in a process known as "working over", which involves the review and revision of content in a cyclic process and possibly contributes to the perception of writing as the main modality for communication because of the perception that "working over" produces a better quality of text than speech [10, 72].

Speaking and traditional writing also differ in their level of recursivity. Recursivity is a concept existing in writing, as "the lack of a sequential or orderly approach" [38], meaning how writers go back to their previous text to continue producing new content or revise.

Writing has a higher level of recursivity, meaning the content is composed and revised in small loops already before being put down on paper or screen whereas in speaking the produced content is rather the "raw" verbalization produced on the fly.

3.3 Memory and Recall

With the differences in production and feedback mechanisms between speaking and writing, we expect the memory traces of produced content in them also differ. By drawing on a key theory of memory from cognitive science, this section attempts to explore how the strength of different types of memory traces come to play.

The Fuzzy Trace Theory is a key mental model theory for the storage and recall of information and posits that any event or information elicits two types of memory traces [45]. The first kind is *verbatim memory trace*, defined as the exact representation of the event's surface form. The second kind is *gist memory trace*, which is defined as the fuzzy representation of meaning or substance [51]. Although both memory traces are derived in parallel, a closer examination of covariance in the recall of verbatim and gist memories found a dramatic level of independence between the two traces, supporting the presence of independent processing, storage, and retrieval [5, 52, 52]. The independence of the two memory traces is also supported by a divergence in their properties, particularly concerning the accessibility and malleability of the two memory traces. Gist memory traces were found to be easier to access, slower to decay, and less likely to distort for the essence of the information [51]. In contrast, verbatim memory traces begin decaying almost immediately after encoding unless prompted otherwise, resulting in a higher amount of information loss in comparison to gist memory traces [4, 23].

Existing research indicates a tendency of gist-level recall of verbal content, with studies indicating that verbatim memory of spoken content is lost as soon as it has been understood except for cases where the stimuli are affectively charged, short, and tested immediately or the subject is explicitly aware that their recall needs to be verbatim [1, 23]. Another study examining the likelihood of verbatim recall found that when tested for the retention of information from provided passages, subjects only stored the original, surface components of the sentences in the passage for the time required to comprehend their meaning, following which only the essence or the information stored within the sentence was captured [56]. Similarly, testing recognition of closely related spoken sentences to see if any false positives occurred found that subjects tended to reorganize semantically related sentences into one "holistic idea1" and also falsely recognize sentences that were not originally presented but contained the "combined meaning of multiple individual sentences" that were experienced previously [6]. Schweppe further argued that the processing of a sentence for verbatim recall is a very cognitively costly process since an exact recall would not only require verbal competence but also usurp a "substantial amount of general attention resources" on top of the cognitive processing power that is already devoted to processing the sentence – causing a performance breakdown and difficulties in recall [11, 59].

The disparity in the speed of text produced using speaking and writing and the lack of graphemic representation in spoken output can affect how well information is stored and retrieved. The presence of multimodal feedback mechanisms in traditional writing may actively enhance writers' memory of their produced text in comparison to speech, with research suggesting that spoken content is more prone to major distortions in recall than writing, with the properties and feedback mechanisms of the latter reducing retrieval efforts in the reconstruction of a text [37].

3.4 Implications for STT Systems

Based on the above-mentioned theory and empirical research, we hypothesize that the lack of inherent graphemic and haptic feedback on its alphabetic production, the faster rate of production, its spontaneity, and low recursivity could all contribute to a faster decay of verbatim recall of spoken compositions and result in a stronger tendency to recall the gist of the content, rather than the exact utterances.

Although STT systems or dictation interfaces provide visual feedback with real-time transcription, the feedback likely has a *disconnection* with the speech itself, due to its graphemic and alphabetic format different from its phonetic format of production and gist-level memory. Moreover, the presentation of transcribed text often has delay and inconsistency caused by speech recognition time and error, resulting in increased confusion or mental demand for the users as they try to reconcile errors in the transcribed text with their intended utterance [10]. In addition, commercially available STT systems such as Google Voice Typing or Otter.ai, display a dynamic presentation of interim text transcription and autocorrect it based on sentence-level context. While these features have good reasons to stay, they could nevertheless further enlarge the disconnection between speech and its visual feedback and distract users' attention.

Establishing clear boundaries between the properties and processes of feedback, storage, and recall between speech and writing has major implications for the design of speech-based interfaces. Existing research indicates that current speech-based interfaces draw heavily upon GUI design principles without altering them to suit the input modality, which could have ramifications for the extent to which the systems can support users in composing text through speech that is capable of being disseminated [15]. The implications of differences in the production time and form of writing and speech, as well as the respective attentional demand and ability to edit compositions, have been only peripherally touched on in the development of speech-based interfaces. Even in cognitive science and linguistics, the focus on writing as the key communication modality has resulted in "very few studies in the literature" directly collecting data on spoken output, leading to the assumption that cognitive processes associated with writing can be generalized to speech [37].

After synthesizing relevant theories from the cognitive science literature above, we conducted a study observing users' verbal composition and self-recall to seek evidence and more insights on how the tendency of gist-level memory comes to play in the context of using speech for text composition.

4 STUDY: RECALL OF VERBAL COMPOSITION

We have learned from the literature that speech production is prone to memory decay and there is a tendency towards gist abstraction when memories fade. This leads to a hypothesis that we will see differences in the verbatim comparison between the composed text and their immediate recall. To test this hypothesis and gain a better understanding of how gist abstraction in memory plays out in a verbal text composition context, our experiment tests users’ recall of self-produced compositions to identify both the level of accuracy of their recall and key content manipulation patterns across different types and lengths of produced content.

4.1 Experimental Design

The study paradigm involved a 2 × 3 within-subject experiment consisting of paired tasks where participants were asked to verbally produce spoken compositions and immediately recall their compositions afterward. The two independent variables were: Content Type [Opinion, Experience] and Length [Short (≈15-20 words), Medium (≈50-60 words), Long (≈250-300 words)]. For the purposes of this study, Opinion-based compositions required participants to produce a composition based on beliefs, values or judgments about a phenomenon, such as what they think about soft drinks, abortion right, etc. Experience-based compositions required participants to produce descriptive compositions based on events they had personally experienced, such as talking about a memorable event that happened over the weekend.

Based on the theories and our analysis of speech input, we hypothesized:

- H1. The accuracy of recall decreases with the increased length of the composition.
- H2. The genre of composed content affects how accurately content can be recalled.

Besides testing these hypotheses, we analyzed the differences between the original compositions and the recalled text, to observe patterns of memory decay and recomposition.

4.1.1 Participants. 12 bilingual English speakers aged 18 – 21 (5 males) were recruited from a local university using convenience sampling. All participants received the majority of their education in English and were pursuing their undergraduate degrees in a broad range of fields ranging from Computer Science to Psychology. Additionally, participants had a preliminary understanding of and experience with speech-based interfaces, although neither was required. All participants were compensated for their participation with a gift card to a nearby supermarket.

4.1.2 Materials and Procedure. Following a brief introduction to the study, participants were asked to carry out a small training task where they produced and recalled a short composition of ≈30-40 words in response to a self-selected prompt from a list of prompts or composed something on-the-spot, which was also used for the main study (see experimental protocol and prompts in Appendix A). Participants were instructed to produce a composition of a certain number of words in estimation and then asked to recall it as accurately as possible. They were quickly shown a printout of pseudo text for each Length condition to get a sense of the number of words to produce for each condition. Participants were not allowed to use any supporting devices such as notes throughout the experiment and were requested to prevent re-using prompts across trials. The experimenter used Otter.AI and roughly estimated word counts to monitor the length of production in real-time and signaled to the participants with a hand gesture after they reach the length of text for the condition. The experimental trials were blocked by Length. The orders of both Length and Content Type were fully counterbalanced across participants. A follow-up interview was conducted after the participants completed all the trials. The entire experiment took around one hour.

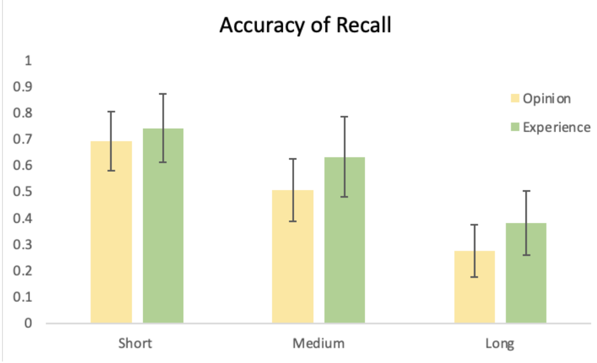


Figure 1: Accuracy of recall across three Lengths and two Content Types.

Effect	Predictor	n, d	F	p	η^2
Recall Accuracy	Content Type	1, 11	15.64	<.005	.587
	Length	2, 22	66.12	<.001	.857
	ContentType×Length	2, 22	.897	.422	.075

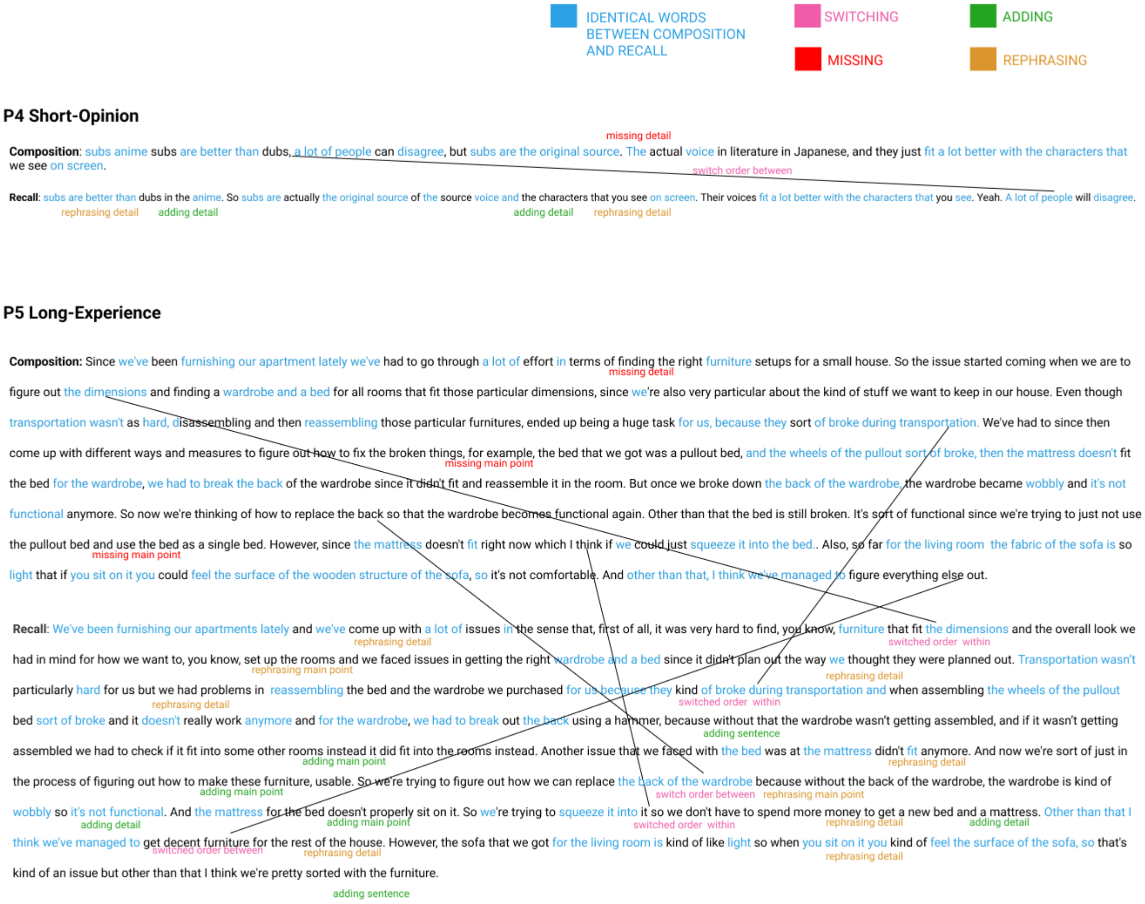
Table 1: Effects of Length and Content Type on Recall Accuracy

4.1.3 Data Collection. Overall, we collected 2 Lengths × 3 Content-Type × 12 Participants = 72 original composition-recall sets throughout this experiment. All composition and recalled text were recorded on Otter.AI and manually corrected for later analysis. A short follow-up interview was carried out with each participant after the culmination of the experiment and explored strategies and the impact of text length and type on composition, recall, and perceived quality.

4.2 Quantitative Results

This section examines the effect of content type and length on the accuracy of recall as well as on the patterns of content distortion through a comparison of the composition and the recalled text. Differences between the two texts were coded and then analyzed using repeated-measures ANOVA and Friedman’s test with post hoc Wilcoxon Signed Ranks Test on SPSS ver.29.

4.2.1 Accuracy of Recall. To calculate the accuracy of recall in each of the paired tasks, transcriptions of the composition and recall were compared to count the number of identical words that appeared in both the composition and recall. To avoid over-complicating this measure, accuracy was calculated, using the bag-of-words approach in NLP, as a percentage of the number of identical words within the total number of words in the recalled text. Although this method tolerates to some extent the reordering of semantic units - meaning it does not count reordering as differences, we noticed


IDENTICAL WORDS BETWEEN COMPOSITION AND RECALL
SWITCHING
ADDING
MISSING
REPHRASING
missing detail
switch order between
rephrasing detail
adding detail
adding detail
rephrasing detail
missing detail
missing main point
missing main point
rephrasing detail
switched order within
rephrasing main point
rephrasing detail
rephrasing detail
switched order within
adding sentence
adding main point
rephrasing detail
adding main point
switch order between
rephrasing main point
adding detail
adding main point
switched order within
rephrasing detail
adding detail
switched order between
rephrasing detail
rephrasing detail
adding sentence

P4 Short-Opinion

Composition: subs anime subs are better than dubs, a lot of people can disagree, but subs are the original source. The actual voice in literature in Japanese, and they just fit a lot better with the characters that we see on screen.

Recall: subs are better than dubs in the anime. So subs are actually the original source of the source voice and the characters that you see on screen. Their voices fit a lot better with the characters that you see. Yeah. A lot of people will disagree.

P5 Long-Experience

Composition: Since we've been furnishing our apartment lately we've had to go through a lot of effort in terms of finding the right furniture setups for a small house. So the issue started coming when we are to figure out the dimensions and finding a wardrobe and a bed for all rooms that fit those particular dimensions, since we're also very particular about the kind of stuff we want to keep in our house. Even though transportation wasn't as hard, disassembling and then reassembling those particular furnitures, ended up being a huge task for us, because they sort of broke during transportation. We've had to since then come up with different ways and measures to figure out how to fix the broken things, for example, the bed that we got was a pullout bed, and the wheels of the pullout sort of broke, then the mattress doesn't fit the bed for the wardrobe, we had to break the back of the wardrobe since it didn't fit and reassemble it in the room. But once we broke down the back of the wardrobe, the wardrobe became wobbly and it's not functional anymore. So now we're thinking of how to replace the back so that the wardrobe becomes functional again. Other than that the bed is still broken. It's sort of functional since we're trying to just not use the pullout bed and use the bed as a single bed. However, since the mattress doesn't fit right now which I think if we could just squeeze it into the bed.. Also, so far for the living room the fabric of the sofa is so light that if you sit on it you could feel the surface of the wooden structure of the sofa, so it's not comfortable. And other than that, I think we've managed to figure everything else out.

Recall: We've been furnishing our apartments lately and we've come up with a lot of issues in the sense that, first of all, it was very hard to find, you know, furniture that fit the dimensions and the overall look we had in mind for how we want to, you know, set up the rooms and we faced issues in getting the right wardrobe and a bed since it didn't plan out the way we thought they were planned out. Transportation wasn't particularly hard for us but we had problems in reassembling the bed and the wardrobe we purchased for us because they kind of broke during transportation and when assembling the wheels of the pullout bed sort of broke and it doesn't really work anymore and for the wardrobe, we had to break out the back using a hammer, because without that the wardrobe wasn't getting assembled, and if it wasn't getting assembled we had to check if it fit into some other rooms instead it did fit into the rooms instead. Another issue that we faced with the bed was at the mattress didn't fit anymore. And now we're sort of just in the process of figuring out how to make these furniture, usable. So we're trying to figure out how we can replace the back of the wardrobe because without the back of the wardrobe, the wardrobe is kind of wobbly so it's not functional. And the mattress for the bed doesn't properly sit on it. So we're trying to squeeze it into it so we don't have to spend more money to get a new bed and a mattress. Other than that I think we've managed to get decent furniture for the rest of the house. However, the sofa that we got for the living room is kind of like light so when you sit on it you kind of feel the surface of the sofa, so that's kind of an issue but other than that I think we're pretty sorted with the furniture.

Figure 2: Examples of how content distortion was analyzed for composition-recall data for Short and Long length conditions

that some changes in wording tended to occur when semantic units get reshuffled.

A Two-Way Repeated Measures Analysis of Variance (RMANOVA) was conducted to examine significant differences in accuracy based on Length and Content Type. Before conducting the RMANOVA, normality, and sphericity were assessed using histograms and Mauchly's W test, and indicated non-significant values for accuracy which were in line with the assumptions of RMANOVA.

The results support both H1 and H2. We found both Length and Content-Type had a significant effect on the accuracy of the recall, with their interaction being none-significant (Fig. 1). Posthoc pairwise comparison with Bonferroni corrections showed all pairwise comparisons between the levels of Length being statistically significant ($p < 0.001$), except for the comparison between *short* and *medium* ($p = 0.287$) for Experience-based compositions. Similarly, a significant main effect of Content Type on recall accuracy was observed, with the mean difference between the *opinion* and *experience* content types being statistically significant ($p < 0.05$). Overall, a decreasing level of accuracy was observed for longer texts than shower, and Opinion-based compositions than Experience-based.

4.2.2 Patterns of Content Distortion. We analyzed the content distortion in memory by comparing the composition and recalled text and categorizing the differences. Using a deductive Thematic Analysis approach, two researchers reviewed 12 trials of the gathered participant data across all conditions to observe common patterns. They iteratively refined the definition of the four categories of content distortion, including *Switching*, *Missing*, *Adding* and *Rephrasing Information*, until reaching consensus. Afterwards one researcher coded the rest of the data following the definition. One example text with markups of the content distortion categories is shown in Figure 2.

We identified both micro and macro levels of all the four categories of content distortion. For *Switching information*, we distinguished the reordering of information unit *within-sentence* and *between-sentence*. For *Missing*, *Adding* and *Rephrasing information*, they are named as the *Detail* level (words and short phrases) and *Main point* (long clauses or sentences). The occurrences of marked deviations from normality in the form of positively skewed data and high kurtosis violated the baseline assumptions of ANOVA, so a non-parametric Friedman's test was used for each measure instead.

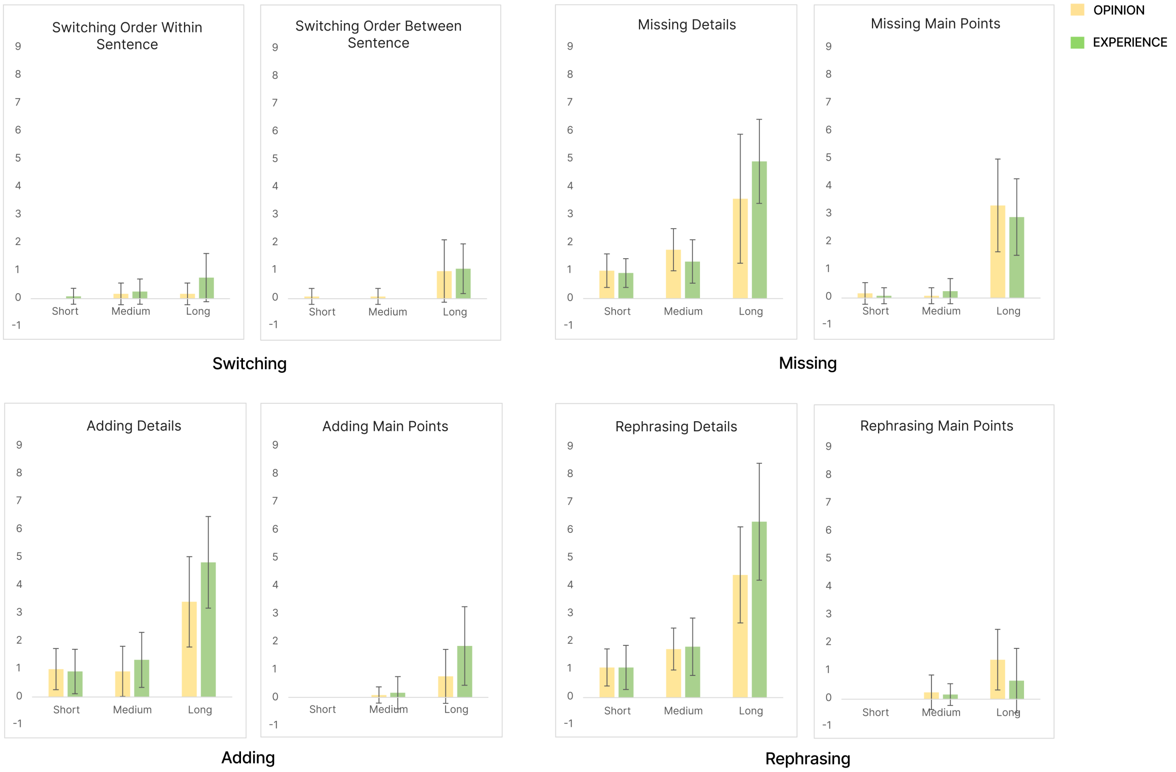


Figure 3: Occurrences of four types of content distortion in recall at micro and macro levels

Wilcoxon signed-rank tests were conducted for pair-wise comparison with Bonferroni correction. We report the main findings in the following, with illustrations in Figure 3.

Switching Information. For this experiment, *Switching Information* was defined as discrepancies in the placement order of a piece of information, such as a subject, an object, an action, or an expression, between the composition and recall.

For *within sentences*, statistically significant differences in the occurrence of switching information were found between the Long and Short conditions (Z = -2.438, p = 0.015) and between Opinion and Experience (Z = -2.070, p = 0.038). For *between sentences*, statistically significant differences in the occurrence of switching information was found between the Long and Short conditions (Z = -3.499, p < 0.001) and the Long and Medium conditions (Z = -3.499, p < 0.001) but not between Opinion and Experience (Z = -.159, p = 0.873).

In summary, a longer composition was found to significantly increase the order switching of information in recall both within and between sentences. The recall of Experience-based composition was more likely to have order switching than Opinion-based, but only within sentences.

Missing Information. As mentioned above, we identified two types of omission of information: *Missing Details* referring to words or small phrase level, and *Missing Main Points* referring to long clause or sentence level.

For *Missing Details*, a statistically significant difference was found between the Long and Short conditions (Z = -4.010, p < 0.001), Long and Medium (Z = -3.942, p < 0.001) and Short and Medium (Z = -2.636, p = 0.008) but not between Opinion and Experience (Z = -.601, p = 0.548). For *Missing Main Points*, a statistically significant difference was found between the Long and Short conditions (Z = -4.305, p < 0.001) and the Long and Medium conditions (Z = -4.126, p < 0.001) but not between Opinion and Experience (Z = -.246, p = 0.806).

In summary, longer compositions had more information loss in recall, at both the micro and macro levels. The content type did not make a significant difference for occurrences of information loss in recall.

Adding Information. Similar to missing information, participants also showed a tendency towards adding information during recall. We also identified *Adding Details*, referring to words or small phrase level, and *Adding Main Points*, referring to long clause or sentence level.

For the occurrence of *Adding Details*, a statistically significant difference was found between the Long and Short conditions (Z = -4.131, p < 0.001) and Long and Medium (Z = -4.328, p < 0.001) but not between Opinion and Experience (Z = -.808, p = 0.419). For the occurrence of *Adding Main Points*, a statistically significant difference was found between the Long and Short conditions (Z = -3.572, p < 0.001) and the Long and Medium conditions (Z = -3.328, p < 0.001) as well as between Opinion and Experience (Z = -2.338, p = 0.017), with Experience having a higher rate.

In summary, longer compositions were found to have more occurrences of new information being added in the recall, although short and medium lengths did not significantly differ. While different content types did not make a difference in adding details, Experience-based composition had more tendency to have major information addition in recall.

Rephrasing Information. The last category of memory distortion was found to be information that remained from composition to recall but got rephrased with new wordings. Again we distinguished between *Rephrasing Details* referring to words or small phrase level, and *Rephrasing Main Points*, referring to long clause or sentence level.

For the occurrence of *Rephrasing Details*, a statistically significant difference was found between the Long and Short conditions (Z = -4.119, p < 0.001), Long and Medium (Z = -4.213, p < 0.001) and Short and Medium (Z = -2.584, p = 0.010) but not between Opinion and Experience (Z = -1.734, p = 0.083). For the occurrence of *Rephrasing Main Points*, a statistically significant difference was found between the Long and Short conditions (Z = -3.384, p < 0.001) and the Long and Medium conditions (Z = -2.830, p < 0.001), with longer compositions having higher rates. There is also a significant difference between Opinion and Experience (Z = -2.496, p = 0.013), with Opinion having a higher rate of alternation.

In summary, longer compositions tended to have more information being rephrased, in both micro and macro levels. Opinion-based compositions had more major information being rephrased than Experience-based ones.

Overall, the patterns of content distortion provide additional insights on how verbatim memory decay in immediate recall of spoken composition. The fact that content got added, missed, or rephrased both on a word/phrase level and clause/sentence level provide evidence of multiple levels of gist extraction in memory and its reproduction in recall. All categories of content distortion showed significant and large differences between Long and Short compositions, but the differences between Short and Medium were not significant except for rephrasing details. The content type only affected order switching and rephrasing, not adding and missing information. We also found opinion-based composition less likely to have details reshuffled within a sentence than experience-based, but more likely to have major points rephrased across sentences.

4.3 Qualitative Results

Following the experiment, all participants were asked to take part in a short interview that encouraged participants to discuss their efforts and strategies for performing the task. Deductive thematic analysis was used to identify key themes and categories emerging from the interviews around participants' perception of their performance, effort, and their strategies The following section details these findings in the context of both compositions and recalls to provide an overview of their perceptions.

Strategies for Spoken Composition and Recall. Participants often relied on key points to help organize the content mentally for both composition and recall, using the points as a “scaffold to assist in recall” (P9). 5 participants drew overt parallels between the key points strategy used in this experiment and their strategies for memorization in other contexts. P4 compared their strategy for the composition-recall task to how they study for exams, with both strategies relying on "remembering the keywords and points". However, a key difference in the strategies for recall was attributed to the different types of content, with P12 stating that the "compositions dealing with experience were often structured and recalled more chronologically while opinion-based compositions often relied on perceptions or associated emotions".

Effort of Recall Across Content-Length and Type. Consistent with the quantitative findings, the ability to accurately recall spoken compositions decreased as a function of the length of the text. All participants found recall to be the hardest for the Long composition condition both because of the amount of content that needed to be remembered and the reduced interval between composition and recall that prevented the rehearsal and storage of content. Common perceptions of recalling longer content included "switching up words a lot" and "difficulties in recalling what was said exactly" (P5). Similarly, 10 participants found their recall for Experience-based compositions to be easier than for Opinion-based compositions, in alignment with the quantitative results. Explanations for the ease in recalling Experience-based compositions suggested that the personal nature of the content allowed participants to use preexisting memories of the event to guide their recall because it is “something [they] have lived through” (P4) whereas composing and forming opinions required participants to “identify, combine and formulate disparate pieces of information” (P7). In line with the quantitative findings, the chronological nature of Experience-based composition resulted in a tendency to “add or revise information as [they] recalled the event in greater detail” during retrieval (P9) while the more objective nature of Opinion-based compositions reduced the possibilities of missing, adding or rephrasing details.

Perceptions of Recall Quality. 8 participants preferred their recalled text over their original compositions, perceiving them to be superior in quality despite the occurrence of content manipulation in the form of removing, adding or rephrasing information. Participants found their recalled text to be “clear and effective, with improvements in grammar and vocabulary so that they could better convey their points”, suggesting that the content manipulation patterns acted as a form of sub-conscious revision and allowed participants to fill in any gaps in their original composition or remove unnecessary information to convey their points more succinctly (P11). However, participants were cognizant of the fact that their recall tended to be “less structured and organized than their initial composition”, with the addition or removal of some content often drastically affecting the substance of the text itself (P10).

4.4 Summary

The findings of this study suggest that length is a key factor influencing the recall of spoken compositions, with accuracy decreasing the content distortion occurring in both detailed information to major points.

The type of content also affected how memory decayed and the participants' efforts and strategies for recall. These findings build on previous research examining the differential decay rates of semantic and verbatim representation of sentences during storage and recall [71]. The experimental results are consistent with research exploring the decay of verbal stimuli in memory [6, 56, 59], as well as dual-trace theories of memory that suggest the elicitation of both verbatim and gist memory traces to encode the form and the substance of spoken information [4, 5, 51].

The findings of the study supplement the arguments made based on the background literature by showing *how* gist-based memory distortion looks like in the context of using speech for text composition. It provides concrete observations for designers and researchers of new systems to refer to, including the granularity of information units for gist extraction and how they change over the length of the composition.

Results from the study provided important keynotes for designing speech-to-text interfaces. First, the rapid decay of verbatim memory indicates the need for support from the interface during the process of editing. Second, individuals tend to recall the gist of their propositions rather than their exact utterances, providing support for the premise that speech-based interfaces should operate on higher elements of information rather than words or phrases. Lastly, the recall of spoken content often transitions into a revision and reproduction of the content, resulting in improvements in the quality and conciseness of the composition. The next section introduces the conceptualization of new directions of research, informed by the theories and findings of this study.

5 TOWARDS NEW INTERACTION PARADIGMS FOR STT INPUT

Based on the theories about speech production and memory as well as the findings from our study, we propose new directions for developing speech-to-text interfaces based on the characteristics of speech as a main input modality. This section summarizes them in the following directions.

5.1 Speaking as an Iterative Drafting Method

The transient nature of speech and the rapid decay of verbatim memory of one's spoken utterances indicate the need for a mechanism that supports the revision of spoken composition without requiring precise verbatim recall from the user. With a slower speed and constant visual and haptic feedback, traditional writing supports a more fine-grained recursivity, meaning that writers can perform micro-revisions, perhaps multiple times in a short sentence. In contrast, speaking produces content at a higher speed but is impromptu, with a relatively coarse recursivity which manifests as repetition and speech repair.

Other aspects of text composition and editing via speech – such as its temporal demand, spontaneous nature, and faster rate of production may prevent deliberate planning and subsequently increase the disorganization of the produced composition and further compound the difficulty of "locating and delimiting" what needs to be edited [24]. Despite the success of re-dictation as a more natural editing mechanism compared to voice commands, further exploration of this technique found that users tend to simply try to select and restate only the erroneous words [20], which follows the strategy of editing on a text editor that optimizes speed by typing the minimum amount. The process of *re-dictation* introduces a certain amount of rigidity into the process of producing spoken compositions. The ensuing effort and mental demand involved in recalling and re-dictating the target text could affect the flow of the user's composition and create a sense of having "irretrievably lost ideas" as users pause to edit their text [14].

To address this, we believe there is a need to rethink text editing with speech holistically without being constrained by "small error correction" and our inherited habits of using a text editor by typing. One direction we propose is to support the entire text authoring task as an *iterative drafting* process by adopting new workflows. Instead of treating the first "blurt" coming out of speech as a piece of writing to be precisely edited, we encourage iterative production and reproduction of longer text to improve its quality before precise editing takes place. Novel interfaces can be designed to support this workflow. Potentially this could reduce the need for precise editing caused by repetition, disorganization, and lack of clarity in spoken content.

5.2 Gist-Based Interaction With Text

Based on the known effects of faster memory decay with speech, the theory indicating a preference for gist-level recall over verbatim, and our observation of how this manifested as rephrasing information points in recalling one's verbal composition, we suggest a potential paradigm shift for interacting with spoken text, from focusing on precise interaction with characters to coarse interaction with chunks of text based on their gist. Such interfaces should help users easily segment text in various levels of abstraction and support the visualization and manipulation of the gist of text segments. The representation of the gist should also serve as a memory aid of users' spoken composition to mitigate memory decay.

This direction potentially allows us to address the verbose issue of speech production, bypass some challenges in correcting speech recognition errors, and support a fast and interactive text authoring process that drives users' attention to meaning production instead of being distracted by frequent precise editing. Recent advancement in natural language processing and understanding enables opportunities to support such interfaces with text segmentation and summarization capabilities, as well as multimodal representation of textual content. The design space for how to represent the gist of text segmentations needs to be explored in future research.

5.3 Supporting Non-Linear Composition

The cognitive process of speech production and its temporal demand determine a lack of deliberate planning. Concurrent activation of new ideas and topics often occur in verbal composition and this leads to disorganization of content in a spoken draft, which takes much time to manually organize and edit. Apart from this, our study also demonstrates the discrepancies in recall around information in one's own spoken text, with users often mixing up the original order of events or details, indicating need for support. With the rapid advancement of Natural Language Understanding, future systems could help users classify their spoken content into meaningful segments. We suggest that future dictation interfaces support non-linear verbal composition by allowing users to speak freely while the system assists in the categorization and organization of the produced content.

5.4 Conclusion

Echoing Shneiderman's early statements about the importance of understanding memory for STT applications [63], we highlight the lack of knowledge and adoption of relevant theories and findings from other fields that study the cognitive aspects of speech. One important message this paper attempts to convey is the implication of *gist extraction* and *verbatim memory decay*. The interfaces today used for dictation or voice typing are inherited from GUI editors that rely heavily on a verbatim representation of text and assume the text being produced at the first go is the "writing" to be revised with precision. Both theoretical and empirical findings in this work showed this may conflict with users' mental models of their composition and incur cognitive load when trying to reconcile the text and find ways to edit. We are not aware of any research in HCI that attempts to understand this in depth.

In this work, we introduced a body of research from Cognitive Science and Linguistics to HCI to establish a theoretical and empirical understanding of the characteristics of speech as a text input modality. By teasing out its differences from writing/typing in its production and feedback mechanism as well as observing how memory decay manifests in a dictation context, we identified why our old habits in using a text editor do not translate well to speech input. Based on these, we propose new interface concepts and directions for future research for supporting verbal composition and interaction with spoken text.

ACKNOWLEDGMENTS

This research was supported by the Hong Kong Research Grants Council - ECS scheme under project number CityU 21209419. We thank Scott Jenson and Shumin Zhai from Google LLC for their constructive feedback and insightful discussions.

REFERENCES

[1] John R Anderson. 1974. Verbatim and propositional representation of sentences in immediate and long-term memory. *Journal of Verbal Learning and Verbal Behavior* 13, 2 (1974), 149–162.

[2] Matthew P Aylett, Per Ola Kristensson, Steve Whittaker, and Yolanda Vazquez-Alvarez. 2014. None of a CHInd: relationship counselling for HCI and speech technology. In *CHI'14 Extended Abstracts on Human Factors in Computing Systems*. 749–760.

[3] James Bigelow and Amy Poremba. 2014. Achilles' ear? Inferior human short-term and recognition memory in the auditory modality. *PloS one* 9, 2 (2014), e89914.

[4] Charles J Brainerd and Johannes Kingma. 1985. On the independence of short-term memory and working memory in cognitive development. *Cognitive Psychology* 17, 2 (1985), 210–247.

[5] Charles J Brainerd and Valerie F Reyna. 2005. *The science of false memory.* Oxford University Press.

[6] John D Bransford and Jeffery J Franks. 1971. The abstraction of linguistic ideas. *Cognitive psychology* 2, 4 (1971), 331–350.

[7] Alfonso Caramazza. 1991. Some aspects of language processing revealed through the analysis of acquired aphasia: The lexical system. *Issues in reading, writing and speaking* (1991), 15–44.

[8] Stuart K Card, Thomas P Moran, and Allen Newell. 1980. Computer text-editing: An information-processing analysis of a routine cognitive skill. *Cognitive psychology* 12, 1 (1980), 32–74.

[9] Wallace Chafe. 1994. *Discourse, consciousness, and time: The flow and displacement of conscious experience in speaking and writing.* University of Chicago Press.

[10] Wallace Chafe and Deborah Tannen. 1987. The relation between written and spoken language. *Annual review of anthropology* 16 (1987), 383–407.

[11] Harald Clahsen and Claudia Felser. 2006. How native-like is non-native language processing? *Trends in cognitive sciences* 10, 12 (2006), 564–570.

[12] Leigh Clark, Philip Doyle, Diego Garaialde, Emer Gilmartin, Stephan Schlögl, Jens Edlund, Matthew Aylett, João Cabral, Cosmin Munteanu, Justin Edwards, et al. 2019. The state of speech in HCI: Trends, themes and challenges. *Interacting with Computers* 31, 4 (2019), 349–371.

[13] Jan Cuřín, Martin Labskỳ, Tomáš Macek, Jan Kleindienst, Hoi Young, Ann Thyme-Gobbel, Holger Quast, and Lars König. 2011. Dictating and editing short texts while driving: Distraction and task completion. In *Proceedings of the 3rd International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. 13–20.

[14] Susan De La Paz. 1999. Composing via dictation and speech recognition systems: Compensatory technology for students with learning disabilities. *Learning Disability Quarterly* 22, 3 (1999), 173–182.

[15] Wanyu Du, Zae Myung Kim, Vipul Raheja, Dhruv Kumar, and Dongyeop Kang. 2022. Read, Revise, Repeat: A System Demonstration for Human-in-the-loop Iterative Text Revision. *arXiv preprint arXiv:2204.03685* (2022).

[16] Konrad Ehlich. 1989. Deictic expressions and the connexity of text. *Text and discourse connectedness* (1989), 33–52.

[17] John F Ehrich. 2006. Vygotskyan inner speech and the reading process. (2006).

[18] Lisa B Elliot, Michael S Stinson, Barbara G McKee, Victoria S Everhart, and Pamela J Francis. 2001. College students' perceptions of the C-Print speech-to-text transcription system. *Journal of deaf studies and deaf education* 6, 4 (2001), 285–298.

[19] Eric Enge. 2020. Mobile voice usage trends in 2020. https://www.perficient.com/insights/research-hub/voice-usage-trends

[20] Jiayue Fan, Chenning Xu, Chun Yu, and Yuanchun Shi. 2021. Just Speak It: Minimize Cognitive Load for Eyes-Free Text Editing with a Smart Voice Assistant. In *The 34th Annual ACM Symposium on User Interface Software and Technology*. 910–921.

[21] Kengo Fujita and Tsuneo Kato. 2011. Design and development of eyes-and hands-free voice interface for mobile phone. In *International Conference on Human Centered Design*. Springer, 207–216.

[22] Paul L Garvin. 1989. Professor Vachek (revisited)-some contemporary issues in the study of speech and writing. (1989).

[23] Morton Ann Gernsbacher. 1985. Surface information loss in comprehension. *Cognitive psychology* 17, 3 (1985), 324–363.

[24] Debjyoti Ghosh. 2021. Voice-based Interactions for Editing Text On The Go. In *2021 Joint Workshop of the German Research Training Groups in Computer Science*. 143.

[25] Debjyoti Ghosh, Pin Sym Foong, Shengdong Zhao, Di Chen, and Morten Fjeld. 2018. EDITalk: towards designing eyes-free interactions for mobile word processing. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–10.

[26] Debjyoti Ghosh, Pin Sym Foong, Shengdong Zhao, Can Liu, Nuwan Janaka, and Vinitha Erusu. 2020. Eyeditor: Towards on-the-go heads-up text editing using voice and manual input. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[27] Debjyoti Ghosh, Can Liu, Shengdong Zhao, and Kotaro Hara. 2020. Commanding and Re-Dictation: Developing Eyes-Free Voice-Based Interaction for Editing Dictated Text. *ACM Transactions on Computer-Human Interaction (TOCHI)* 27, 4 (2020), 1–31.

[28] Michele E Gloede and Melissa K Gregg. 2019. The fidelity of visual and auditory memory. *Psychonomic Bulletin & Review* 26 (2019), 1325–1332.

[29] Michele E Gloede, Emily E Paulauskas, and Melissa K Gregg. 2017. Experience and information loss in auditory and visual memory. *Quarterly Journal of Experimental Psychology* 70, 7 (2017), 1344–1352.

[30] Florian Habler, Marco Peisker, and Niels Henze. 2019. Differences between smart speakers and graphical user interfaces for music search considering gender effects. In *Proceedings of the 18th International Conference on Mobile and Ubiquitous Multimedia*. 1–7.

[31] James Hartley, Eric Sotto, and James Pennebaker. 2003. Speaking versus typing: a case-study of the effects of using voice-recognition software on academic correspondence. *British Journal of Educational Technology* 34, 1 (2003), 5–16.

[32] Charles F Hockett and Charles D Hockett. 1960. The origin of speech. *Scientific American* 203, 3 (1960), 88–97.

[33] Lee Honeycutt. 2003. Researching the use of voice recognition writing software. *Computers and Composition* 20, 1 (2003), 77–95.

[34] Biing-Hwang Juang and Lawrence R Rabiner. 2005. Automatic speech recognition–a brief history of the technology development. *Georgia Institute of Technology. Atlanta Rutgers University and the University of California. Santa Barbara* 1 (2005), 67.

[35] Clare-Marie Karat, Christine Halverson, Daniel Horn, and John Karat. 1999. Patterns of entry and correction in large vocabulary continuous speech recognition systems. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. 568–575.

[36] Azlina Amir Kassim, Rehan Rehman, and Jessica M Price. 2018. Effects of modality and repetition in a continuous recognition memory task: Repetition has no effect on auditory recognition memory. *Acta Psychologica* 185 (2018), 72–80.

[37] Ronald T Kellogg. 2007. Are written and spoken recall of text equivalent? *The American Journal of Psychology* 120, 3 (2007), 415–428.

[38] Stephen D Krashen. 1984. *Writing, research, theory, and applications.* Pergamon.

[39] Anuj Kumar, Tim Paek, and Bongshin Lee. 2012. Voice typing: a new speech interaction model for dictation on touchscreen devices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* 2277–2286.

[40] Eric Lambert and Pauline Quémart. 2019. Introduction to the special issue on the dynamics of written word production: methods, models and processing units. *Reading and Writing* 32, 1 (2019), 1–12.

[41] Yuan-Hsuan Lee. 2015. The effectiveness of using inner speech and communicative speech in reading literacy development: A synthesis of research. *International Journal of Social Science and Humanity* 5, 8 (2015), 720.

[42] Arthur E McNair and Alex Waibel. 1994. Improving recognizer acceptance through robust, natural speech repair. In *Third International Conference on Spoken Language Processing.*

[43] James Moffett. 1982. Writing, inner speech, and meditation. *College English* 44, 3 (1982), 231–246.

[44] Cosmin Munteanu and Gerald Penn. 2017. Speech-based interaction: Myths, challenges, and opportunities. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems.* 1196–1199.

[45] Timothy N Odegard and James M Lampinen. 2005. Recollection rejection: Gist cuing of verbatim memory. *Memory & Cognition* 33, 8 (2005), 1422–1430.

[46] Walter J Ong. 2013. *Orality and literacy.* Routledge.

[47] Sharon Oviatt, Phil Cohen, Lizhong Wu, Lisbeth Duncan, Bernhard Suhm, Josh Bers, Thomas Holzman, Terry Winograd, James Landay, Jim Larson, et al. 2000. Designing the user interface for multimodal speech and pen-based gesture applications: State-of-the-art systems and future research directions. *Human-computer interaction* 15, 4 (2000), 263–322.

[48] Joseph S Perkell. 2012. Movement goals and feedback and feedforward control mechanisms in speech production. *Journal of neurolinguistics* 25, 5 (2012), 382–407.

[49] Thomas G Poder, Jean-François Fisette, and Véronique Déry. 2018. Speech recognition for medical dictation: overview in Quebec and systematic review. *Journal of medical systems* 42, 5 (2018), 1–8.

[50] Jeremy J Purcell, Peter E Turkeltaub, Guinevere F Eden, and Brenda Rapp. 2011. Examining the central and peripheral processes of written word production through meta-analysis. *Frontiers in psychology* 2 (2011), 239.

[51] Valerie F Reyna. 2012. A new intuitionism: Meaning, memory, and development in Fuzzy-Trace Theory. *Judgment and Decision making* (2012).

[52] Valerie F Reyna and Charles J Brainerd. 1995. Fuzzy-trace theory: An interim synthesis. *Learning and individual Differences* 7, 1 (1995), 1–75.

[53] Radiah Rivu, Yasmeen Abdrabou, Ken Pfeuffer, Mariam Hassib, and Florian Alt. 2020. Gaze'N'Touch: Enhancing Text Selection on Mobile Devices Using Gaze. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems.* 1–8.

[54] David A Rosenbaum. 2010. Human motor control. (2010).

[55] Sherry Ruan, Jacob O Wobbrock, Kenny Liou, Andrew Ng, and James Landay. 2016. Speech is 3x faster than typing for english and mandarin text entry on mobile devices. *arXiv preprint arXiv:1608.07323* (2016).

[56] Jacqueline Strunk Sachs. 1967. Recognition memory for syntactic and semantic aspects of connected discourse. *Perception & Psychophysics* 2, 9 (1967), 437–442.

[57] Timothy A Salthouse. 1986. Perceptual, cognitive, and motoric aspects of transcription typing. *Psychological bulletin* 99, 3 (1986), 303.

[58] Herbert Schriefers and Gabriella Vigliocco. 2015. Speech production, psychology of [repr.]. In *International Encyclopedia of the Social & Behavioral Sciences (2nd ed) Vol. 23.* Elsevier, 255–258.

[59] Judith Schweppe, Sandra Barth, Almut Ketzer-Nöltge, and Ralf Rummer. 2015. Does verbatim sentence recall underestimate the language competence of near-native speakers? *Frontiers in psychology* 6 (2015), 63.

[60] Andrew Sears, Jinhuan Feng, Kwesi Oseitutu, and Claire-Marie Karat. 2003. Hands-free, speech-based navigation during dictation: difficulties, consequences, and solutions. *Human-computer interaction* 18, 3 (2003), 229–257.

[61] 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.* 1–11.

[62] Rustam Shadiev, Wu-Yuin Hwang, Nian-Shing Chen, and Yueh-Min Huang. 2014. Review of speech-to-text recognition technology for enhancing learning. *Journal of Educational Technology & Society* 17, 4 (2014), 65–84.

[63] Ben Shneiderman. 2000. The limits of speech recognition. *Commun. ACM* 43, 9 (2000), 63–65.

[64] Khe Chai Sim. 2010. Haptic voice recognition: Augmenting speech modality with touch events for efficient speech recognition. In *2010 IEEE spoken language technology workshop.* IEEE, 73–78.

[65] Khe Chai Sim. 2012. Speak-as-you-swipe (SAYS) a multimodal interface combining speech and gesture keyboard synchronously for continuous mobile text entry. In *Proceedings of the 14th ACM international conference on Multimodal interaction.* 555–560.

[66] Shyamli Sindhwani, Christof Lutteroth, and Gerald Weber. 2019. ReType: Quick text editing with keyboard and gaze. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.* 1–13.

[67] Michael S Stinson, Lisa B Elliot, Ronald R Kelly, and Yufang Liu. 2009. Deaf and hard-of-hearing students' memory of lectures with speech-to-text and interpreting/note taking services. *The Journal of Special Education* 43, 1 (2009), 52–64.

[68] Bernhard Suhm, Brad Myers, and Alex Waibel. 2001. Multimodal error correction for speech user interfaces. *ACM transactions on computer-human interaction (TOCHI)* 8, 1 (2001), 60–98.

[69] Oanh Thi Tran and Viet The Bui. 2021. Neural text normalization in speech-to-text systems with rich features. *Applied Artificial Intelligence* 35, 3 (2021), 193–205.

[70] Rein Turn. 1974. Speech as a man-computer communication channel. In *Proceedings of the May 6-10, 1974, national computer conference and exposition.* 139–143.

[71] Ovid J Tzeng. 1975. Sentence memory: Recognition and inferences. *Journal of Experimental Psychology: Human Learning and Memory* 1, 6 (1975), 720.

[72] Josef Vachek. 1973. Written language: General problems and problems of English Mouton. *The Hague* (1973).

[73] Keith Vertanen and Per Ola Kristensson. 2010. Getting it right the second time: Recognition of spoken corrections. In *2010 IEEE Spoken Language Technology Workshop.* IEEE, 289–294.

[74] Karen Ward and David G Novick. 2003. Hands-free documentation. In *Proceedings of the 21st annual international conference on Documentation.* 147–154.

[75] Maozheng Zhao, Henry Huang, Zhi Li, Rui Liu, Wenzhe Cui, Kajal Toshniwal, Ananya Goel, Andrew Wang, Xia Zhao, Sina Rashidian, et al. 2022. EyeSayCorrect: Eye Gaze and Voice Based Hands-free Text Correction for Mobile Devices. In *27th International Conference on Intelligent User Interfaces.* 470–482.

A APPENDIX

Example Topics for Inspiration If you require some assistance coming up with a topic for your oral composition, please feel free to either use the prompts provided below or use the prompts as an inspiration to come up with your own topic. However, please ensure that you read the phrasing of the prompts carefully and that you do not mix and match the prompts across types of compositions. Please also refrain from re-using topics for your following compositions.

Once you have decided on your topic, please return this sheet to the experimenter.

For Oral Compositions about Experiences

- Describe a recent event that happened to you at work
- Describe a recent event that happened to you at school
- Describe a recent event that happened to you when you were with your family
- Describe a recent event that happened to you when you were with your friends
- Describe a recent event that happened to you at a party
- Describe a recent event that happened to you over the weekend

For Oral Compositions about Opinions

- Give your opinion on sodas and soft-drinks
- Give your opinion on social media usage by young people
- Give your opinion on the usefulness of studying abroad for young people
- Give your opinion on the political situation of the United States of America
- Give your opinion on whether standardized tests are a good idea
- Give your opinion on whether universities should allow fraternities/sororities