IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 14 April 2024, accepted 6 May 2024, date of publication 13 May 2024, date of current version 20 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3400014

RESEARCH ARTICLE

A Computer-Assisted Interpreting System for Multilingual Conferences Based on Automatic Speech Recognition

JICHAO LIU[1], CHENGPAN LIU[2], BUZHENG SHAN[3], AND ÖMER S. GANIYUSUFOGLU[4]

[1]School of Foreign Studies, Tongji University, Shanghai 200092, China
[2]Sino-European Institute of Aviation Engineering, Civil Aviation University of China, Tianjin 300300, China
[3]Department of Mathematics, Texas A&M University, College Station, TX 77843, USA
[4]Qingdao International Academician Park, Qingdao 266199, China

Corresponding author: Chengpan Liu (leo_liuchengpan@163.com)

This work was supported by Tianjin Municipal Education Commission under Grant 2021SK035.

ABSTRACT Computer-aided interpreting(CAI) systems are software applied to one or more stages of interpreting tasks, which can directly promote the work of interpreters and improve the interpreting quality. Until now, studies on CAI in simultaneous interpretation(SI) have been limited, primarily focusing on the design and development of manual extraction models. Moreover, what's particularly noteworthy is the lack of thorough investigation into the development of fully automated CAI models for extracting terms (SI difficulties) and other related aspects. Based on the experimental research of existing ones, this study puts forward some new methods based on automatic speech recognition(ASR) and develops a CAI system–InterpretSIMPLE with user-friendly interface, which implements automatic retrieval and display of terms (pre-imported), numbers, etc. as well as other functions specialized for conference interpreters. Through the setting of three-line label control and underline panel control, the system realizes the attention allocation and positioning of the source text content at different levels. One-click import of commonly-used Excel glossary gives simple operation with no additional format conversion. Terminologies and numbers are displayed below the corresponding position while displaying the source text, so that interpreters could locate and solve these recognized SI difficulties. Through the ''exact matching'' or ''partial matching'' setting, it could meet the personalized requirements of terms matching. The experiment shows that after the system receives text information from Tencent Cloud, the real-time display rate of the pre-imported glossary reaches 98.92%. The research results could provide references for the research and development of in-process automated CAI tools.

INDEX TERMS Computer-aided interpreting, CAI, simultaneous interpreting, automatic speech recognition, artificial intelligence.

I. INTRODUCTION

Simultaneous Interpretation (SI) is a process wherein the interpreter translates the source language into the target language almost instantly as the speaker delivers their speech, without any interruptions. In SI, listening (source language input) and translation (target language output) are nearly synchronized, and interpreters need to process multi-level tasks such as understanding, memory, conversion and expression at the same time within a very limited time. Even those with professional training and rich experience make mistakes. Terminologies and numbers are regarded as ''problem triggers'' that affect the quality [1], [2]. In the same context, when interpreters fail to retrieve them from memory, it is often necessary to consult a pre-organized, printed paper or Excel electronic glossary, or to complete the process with the help of a partner.

In recent years, the existing research, combining the working process of SI and information retrieval technology, ascend the manual retrieval convenience of computer-assisted

The associate editor coordinating the review of this manuscript and approving it for publication was Yu-Da Lin.

interpreting (CAI) systems [3]. Although experimental studies [4], [5], [6] demonstrated that innovative manual retrieval methods improve the quality of interpretation such as terminology, however, in fast-paced, multi-task SI process, typing are considered ''unnatural'', ''time consuming and distracting'' behaviors [7], which can lead to short-term memory overload and affect the quality and continuity of interpretation.

Recent advancements in Artificial Intelligence (AI), particularly in deep learning and neural networks, have significantly improved Automatic Speech Recognition (ASR) quality [8]. Although these technological developments have paved the way for AI-powered SI, challenges persist, including ASR errors, dialect recognition issues, inaccurate punctuation, and sentence interpretation errors that directly affect the quality and comprehensibility of the translated text. Enhancing the quality and precision of ASR and Machine Translation (MT) remains an ongoing demand. However, the use of AI technology to achieve automatic retrieval of ''problem triggers'' in SI can further reduce cognitive stress [9]. Therefore, the academic community generally calls for the design and application of automatic retrieval CAI tools.

While technology, such as computer-assisted translation (CAT) tools, has transformed traditional translation methods, suggesting a future where translators will focus more on post-editing MT, research in CAI is still in its early stages [10]. With only a few researches on CAI tools in simultaneous interpreting (SI), all designed and developed by manual extraction model, there has been no study focusing on the detailed implementation of a real automated CAI model for extraction of terms, numbers(SI difficulties), etc [3], [11]. In response to the limited research, this paper proposes innovative methods of automatic retrieval and display of the difficulties, selecting the matching methods as needed, choosing whether to display punctuation as needed to avoid the impact of ASR errors in SI, presenting a newly-designed CAI system with a user-friendly interface. It is the first study to realize the complete research progress focusing on the implementation of a real automated CAI tool for in-process SI.

The academic paper consists of the following sections. Section II is the Literature Review, which includes a discussion on translation and AI-powered SI, with specific subtopics on ASR, MT, speech synthesis, and AI-powered SI. Section II also covers research on CAI, including its definition, categorization, and empirical studies. Section III is the Methodology section, which introduces a newly-designed CAI system, its design philosophy, and experiments conducted using this system. Section IV provides an overview of the system, including interface design and function implementation. Section V focuses on validation through experiments, discussing speech, experiment procedure, data analysis, and targeted system improvement.

II. LITERATURE REVIEW

To offer specific and targeted solutions for CAI systems based on Automatic Speech Recognition (ASR), this chapter carries out a comprehensive analysis of literature and empirical experiments. We have searched a body of relevant CAI literature, over the past 20 years, which also include Chinese literature achieved from resources such as the CNKI database given that the availability of English references is limited.

The literature review is segmented into three parts. The initial section delves into the historical context of interpretation and the cognitive analysis of simultaneous interpreting. The subsequent part encompasses the advancements in AI interpreting technologies, encompassing ASR and Machine Translation (MT). Lastly, the third part covers the progressive strides in CAI research.

A. TRANSLATION AND SIMULTANEOUS INTERPRETATION

Translation involves converting written or spoken content from one language into another. Among various forms of translation, interpreting, which involves orally conveying the words spoken in a different language, is unique and important, in view of its critical role in real-time communication, enabling immediate and direct interaction between individuals speaking different languages, fostering seamless dialogue, and facilitating effective understanding without the barriers of linguistic diversity.

Interpreting is integral to international communication between nations or regions. The enduring role of interpreting has been witnessed across diverse chapters of human history.

The historical development of SI, from its birth, underscores its evolution and significance in facilitating multilingual communication across critical international events. The contemporary mode of simultaneous interpretation can be traced back to the 1920s when Edward Filene, an American businessman and philanthropist, and Alan Gordon Finlay, an engineer and inventor, designed a device known as the ''Filene-Finlay simultaneous translator'', an innovation that included headphones and microphones, later refined and extended by them and IBM President Thomas Watson, for use at the League of Nations [12]. Subsequently, SI gained substantial recognition during post-war Nuremberg trials and has since been extensively utilized in multilateral organizations (such as UN, EU, etc.), international summits (such as APEC, G-20, etc.), legal proceedings, and press conferences. In 1947, the United Nations Resolution 152 designated SI as a permanent service that could either substitute for or complement consecutive interpreting [13]. The introduction of SI equipment fundamentally revolutionized the daily delivery of interpretation.

SI stands as one of the most intricate forms of language processing. From a cognitive standpoint, it involves exceptionally high mental demands, requiring interpreters to navigate multiple concurrent tasks [14], [15], [16], [17]. These encompass listening comprehension, managing short-term memory, segmenting sentences, producing translations,

and orchestrating overall coordination in real-time (Model I, Fig. 1). Confronted with the distinctiveness, complexity, and demanding nature of SI, even seasoned professional interpreters, despite their extensive training, may encounter translation errors. On that account, the precision and fluency of SI generally tend to be lower compared to consecutive interpretation and written translation.

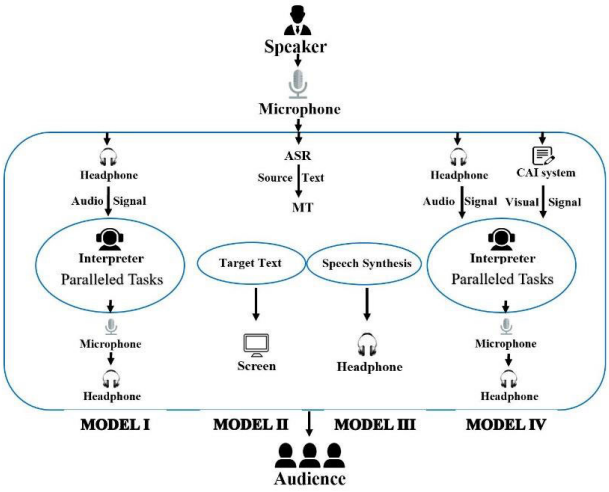

FIGURE 1. Different models of simultaneous interpretation.

Over the span of more than half a century of research, numerous scholars have explored the working process of SI from different perspectives. The consensus on SI as a cognitive task is largely characterized as ''complex'' [18], ''difficult'' [19], and ''demanding'' [20]. Be that as it may, the fundamental processing of SI remained largely unchanged until the advent of CAI and AI technology in the field of simultaneous interpretation.

B. AI-POWERED SIMULTANEOUS INTERPRETATION

AI-powered simultaneous interpreting involves a combination of various technologies to enable real-time translation of spoken language, among which ASR, MT are major players. The recent advancements have significantly propelled research in automatic speech translation, particularly applicable in live and streaming scenarios such as SI [21]. This technological synergy operates in a seamless process: the ASR model transcribes the spoken audio signal into the source language text, the MT model translates this text into the desired target language, and subsequently, the speech synthesis model converts the translated text into an audio signal (as depicted in Model II, Fig. 1). This comprehensive process ensures the fluid and real-time transformation of spoken language, facilitating efficient communication and interpretation across diverse linguistic contexts.

1) AUTOMATIC SPEECH RECOGNITION

Speech is the consistent auditory signal produced by the vocal organs of the human body by adhering to specific language principles. Speech recognition, also known as automatic speech recognition or speech-to-text, refers to the capacity of a computing system or machine to recognize and comprehend spoken language or verbal signals, transforming them into written text. The origins of speech recognition research date back to the 1950s, marked by the pioneering work of Davis et al. at Bell Labs, who delved into single-person speech digital recognition [22]. Over time, various methodologies emerged, including Dynamic Time Warping (DTW), Vector Quantization (VQ), and the Hidden Markov Model (HMM). Notably, in the 1980s, the statistical model-based algorithm represented by HMM gradually supplanted DTW, rising as the primary approach in speech recognition [23]. HMM is commonly amalgamated with the Gaussian Mixture Model (GMM), whereby the GMM-HMM model is created. Despite the dominance of GMM-HMM facilitated by model self-adaptation methods and various discriminative training criteria until the early 21st century, the practical applicability and overall efficacy of the speech recognition systems necessitated substantial enhancement.

Artificial Neural Networks (ANN), namely the predecessor of Deep Neural Networks (DNN), stepped into the realm of speech recognition during the 1980s. However, due to computational limitations and imperfect theoretical basis, its efficacy remained inferior to the established GMM-HMM approach. It wasn't until the advent of the deep belief network (DBN) proposed by Hinton et al. in 2006 that crucial obstacles such as local optimization and overfitting during DNN optimization were resolved [24]. Taking advantage of this achievement, the significant breakthrough arrived when DNN achieved success in small vocabulary continuous speech recognition in 2009, followed by their prowess in large vocabulary continuous speech recognition in 2011 [25]. This pivotal transition marked a momentous shift as DNN-based language recognition supplanted the traditional GMM-HMM framework, firmly establishing a new era of swift advancements in speech recognition.

Researchers made progressive advancements in speech recognition models by introducing the feedforward deep neural network (F)DNN) to substitute the GMM in the GMM-HMM model, thereby proposing the DNN-HMM approach. Subsequently, they introduced the recurrent neural network (RNN) framework to establish correlations between the information of the current moment and previous moment within the model, as well as introduced models such as Long Short-Term Memory (LSTM), Bi-directional Long Short-Term Memory (BLSTM) [26], and latency-control BLSTM [27]. Such innovations notably enhanced RNN-based speech recognition algorithm. On top of that, researchers adopted speech recognition framework based on convolutional neural network (CNN), which not only displays incredible robustness due to the local receptive field mechanism, but also processes both long-term historical information and future information. Both RNN and CNN have significantly propelled the advancement of speech recognition technology, as well as achieved success in practical scenarios.

Currently, the end-to-end speech recognition represents a focal point in speech recognition research, allowing for the direct transformation of speech into text sequences. This approach consolidates the input and output ends into a unified neural network model, offering a simpler realization of the model and higher accuracy. Despite such advantages, it still confronts issues related to stability, demanding further research and exploration.

2) MACHINE TRANSLATION

Machine translation (MT) is the automated method used for converting text or speech from one language to another through the utilization of computer software or algorithms. Ideally, its objective is to enable comprehension of content in diverse languages without requiring human intervention.

The evolution of machine translation spans different methodologies, from the early rule-based approaches rooted in Chomsky's transformational-generative grammar to more contemporary techniques such as example-based, statistics-based, and presently, neural network-based translation methods. Initially, the rule-based machine translation (RBMT), reliant on a set of meticulously formulated bilingual rules, used to be the cornerstone of machine translation, suited for sentences featuring standardized and lucid structures. However, their decline commenced in the late 1990s due to the difficulty in compiling comprehensive bases of rules, which could not accommodate non-standard language structures or emerging linguistic phenomena.

In the 1980s, Nagao [28] introduced the example-based machine translation method (EBMT), whereby sentences are matched with sentences in corpus based on their similarity, and then and translated into the target language. However, due to its reliance on the corpus scale and coverage, this approach was rather limited in its application, hindering its full potential for example optimization and full development.

IBM's Peter Brown et al. proposed a statistical machine translation (SMT) model in the early 1990s based on source channel model [29]. This model regards each sentence in the target language as a candidate, and selects the most probable translation based on the statistical model, thereby allowing full use of the parameters of the corpus learning model, and continual optimization and expansion with the growing corpus. This statistical model significantly enhanced translation accuracy, opening a new phase of rapid development in machine translation. Subsequent models like the logarithmic linear model [30], hierarchical phrase-based model [31], and others further enriched the SMT family. SMT algorithms founded on statistics have indisputably ushered in a new era marked by rapid development and flourishing progress in the realm of machine translation [32].

The introduction of DBN in 2006 accelerated the advancement of deep learning, triggering people's effort to utilize DNN in translation. The encoder-decoder structure for neural machine translation proposed in 2014 signaled the official entry of machine translation into the deep learning era. Various neural network models, including those based on RNN and CNN, have shown superior translation capabilities compared to traditional statistical models. Particularly, the attention-based transformer model introduced in 2017 by Vaswani et al. [33] demonstrated substantial performance improvements, soon rising as one of the leading models in neural network translation research. What's more, technologies such as pre-training and back-translation have further witnessed model efficiency and effectiveness.

3) AI-POWERED SIMULTANEOUS INTERPRETATION

Non-real-time speech translation, such as consecutive interpreting, involves the translation of the source text, which is converted by the MT part from ASR after the speaker has finished speaking. In contrast, in AI SI, the MT part needs to commence translation before the speaker concludes a sentence, as initiating translation after the speaker's completion would result in intolerable delays. On that account, this context demands the MT engine to strike a balance between delay time and translation quality. Initiating translation before receiving crucial content in the source language compromises translation quality, while waiting to translate after obtaining a substantial amount of source language content leads to unnecessary delays. In summary, the text processing and generation of MT is crucial, which determines the output quality and appropriateness of translation results in the SI process. Researchers have proposed various solutions to this situation, such as the initiation of translation until a certain number of words or units of text have been received (referred to as the "wait-k" strategy) [34] and the introduction of attention mechanisms [35].

In the context of speech-to-speech SI, the output translation must be stable and unmodifiable. However, some AI SI systems only consist of two models: ASR and MT (as depicted in Mode III, Fig. 1). The ASR model converts speech signals into text of the source language, and the MT translates the source text into target language text [36]. When the output form of the translated language is text, such as real-time subtitles, modifying the translated language becomes feasible. Compared to speech-to-speech, the re-translation strategy in speech-to-text is not limited to fixed translated language content, offering the advantage of low latency [37]. Of course, this mode also faces challenges in real-time translation display, including high load for computer processing, high rates of modification, and unstable display of subtitles.

In March 2018, the AI-powered SI debuted at the "Translating Automation User Society (TAUS) Asia Summit," marking the first global test of machine SI in authentic communication settings. Subsequently, during the Boao Forum for Asia in April and at the RISE in July of the same year (the largest technology summit in Asia), AI-powered SI was practically applied. However, several errors in ASR and MT during these events raised doubts among interpreters, the public, and the media. On that account, although the advancement in natural language processing technologies has

turned AI-powered SI into a reality, numerous unresolved issues persist.

Several crucial insights can be gleaned from the debut and subsequent application of AI-powered SI. Such tests signified a significant stride in the realm of machine interpretation. Nonetheless, the occurrence of multiple errors during these high-profile events led to skepticism among interpreters, the public, and the media. Despite the advancements in natural language processing technologies that validated AI-powered SI, it became evident that numerous issues and limitations persist, requiring further attention and resolution. These events underscored the importance of continuing research and development in refining the technology for more accurate and reliable real-world applications, and strategies for tackling unresolved issues within the machine interpreting system.

Besides, such real-life applications also evoke the thought about whether AI can replace interpreters. In this regard: we maintain the view that the current AI-powered SI may not be ready to replace human interpreters, a sentiment echoed by Ortiz and Cavallo [38]: ''Among the reasons that technology may not replace interpreters in the future are the complexities of nuances, linguistic variation, non-verbal communication, accents, emotional subtleties, between-the-lines comprehension, human adaptability, decision-making, reliability, cultural context, metaphors, intonation, irony, ambiguities, unpredictability, and judgment capabilities''.

Until natural language processing evolves into natural language understanding, further research is imperative to refine the application of AI-powered SI in practical scenarios, particularly focusing on the enhancement of ASR and MT models, and the formulation of reading and writing strategies.

C. RESEARCH ON COMPUTER-AIDED INTERPRETING

Based on the above literature review, it is established that the evolution of AI-powered systems, while significantly advancing machine interpretation, has not negated the indispensability of human interpreters in the field of simultaneous interpreting. Consequently, computer-aided interpreting (CAI) stands as the primary way to enhance simultaneous interpretation efficiency, ensuring improved accuracy and a more reliable performance within the field.

1) CLASSIFICATION OF COMPUTER-AIDED INTERPRETING

Approximately 15 years ago [39], CAI emerged with the objective of optimizing the interpreting process by integrating computer software to aid interpreters throughout various stages, starting from event preparation to in-process interpretation and subsequent tasks.

While research into CAI is still in its nascent stage, experts hold varied opinions regarding the classification of associated technologies. This paper adopts Fantinuoli's classification [40] as the basis for categorizing CAI, delineating technological tools into four distinct categories based on their direct relevance to conference interpreting. The first three categories encompass: 1) the enhancement of technical tools aimed at interpreting training, known as computer-assisted interpreting training (CAIT); 2) the provision of solutions and technologies catering to different forms of interpretation, inclusive of remote interpretation; and 3) the implementation of fully automatic interpretation technology, notably machine interpreting (MI). Notably, only technology specifically engineered to enhance the interpretation process and elevate interpreter performance is categorized under CAI.

Fantinuoli's differentiation serves to elucidate and define the precise objectives of CAI tools, setting them apart from skill-based training, remote interpretation, and AI interpreting tools. Consequently, CAI tools is meticulously designed and developed to aid interpreters, encompassing all forms of computer programs and mobile applications usable at various stages of the interpreting process [40]. This categorization serves to clarify the distinct role of CAI tools, distinguishing them from other related technologies and emphasizing their focused utility in supporting and refining the art of interpretation.

2) ROLES OF COMPUTER-AIDED INTERPRETING TOOLS IN DIFFERENT PHASES

The working phase of SI can be divided into two critical stages: advance preparation and the in-process phase. Interpreters routinely encounter challenges when dealing with specialized topics outside their expertise or qualifications. Accordingly, event preparation has been underscored in literature as a pivotal component in interpreting assignments [1]. The primary aim of advance preparation is to bridge knowledge gaps between conference participants and interpreters. This approach alleviates the cognitive load during the interpreting task, fostering more efficient management of the process and culminating in higher-quality interpretations [40].

To streamline the preparation process, CAI tools have integrated diverse functionalities, particularly in terminology acquisition. These tools automatically access terminology resources when constructing new glossaries, extract terms manually from parallel documents, automatically retrieve specialized terms from monolingual preparatory materials, and incorporate flash-card systems for the memorization of specialized terminologies ahead of an event [41]. This systematic approach rationalizes the interpreter's work with terminology, aiming ultimately to enhance the quality of their output, specifically in terms of terminological precision and adequacy.

In fast-paced, multi-task in-process SI, although experimental studies [4], [5], [6] suggested that innovative manual retrieval methods improve the quality of interpretation such as terminology, however, typing manually are considered ''unnatural'', ''time consuming and distracting'' behaviors [7], which can lead to short-term memory overload and affect the quality and continuity of interpretation.

ASR has been recognized for its substantial potential in reshaping interpreting practices [42]. Recent advancements have introduced new prospects for the in-process phase.

Integrating ASR into the interpreting workflow not only aids in event preparation but also serves as a tool during the interpreting process. Within the realm of SI, computer tools equipped with ASR have been proposed as invaluable companions, supplying real-time source text and suggesting solutions for challenging elements such as numbers, terminology, and named entities [9]. Up to now, there is still no study focusing on the implementation of a real in-process CAI tool in SI for automatic extraction of terms, numbers, etc [3], [11]. However, researchers have conducted experiments with different AI tools, experimental models, or simulation systems, which analyzed the feasibility and effectiveness of in-process CAI systems, thus providing references for the design of this study.

3) EMPIRICAL STUDIES OF COMPUTER-AIDED INTERPRETING

Various CAI tools and simulation systems (including PowerPoint and video simulations of ASR software), have been the subject of extensive experimentation within SI. These CAI tools offer different display modes, categorized into three types for interpreters' workflows, as outlined in Fig. 2.

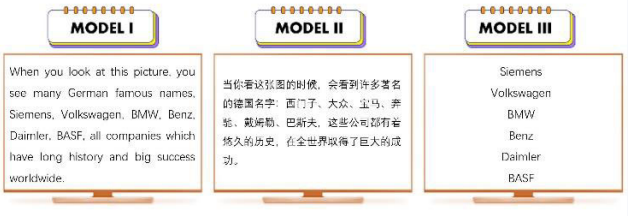

FIGURE 2. Different display models of computer-assisted interpreting tools.

The first type comprises ASR systems that present the entire recognized source text. Researchers, conducting SI experiments utilized ASR systems [43], [44], [45]. The findings from these experiments emphasized the significance of accuracy in numerical and terminological output for participants. Notably, the precision rate of ASR holds a direct influence on SI translation quality. Lin [46] highlighted that a precision rate below 85% negatively impacts interpreter performance. Conversely, a precision rate exceeding 95% reduces reaction times and latency compared to scenarios without CAI tools. Furthermore, researchers observed the detrimental impact of inaccurate punctuation on interpreters. This issue arises when impromptu speakers exhibit unusual speech patterns, such as extended pauses or explanatory remarks before completing a sentence. Such irregularities hinder speech recognition technology's ability to punctuate the transcript correctly, resulting in potential misunderstanding [47].

The second type is a system that integrates cascade ASR and MT, displaying complete translations of the target text or both the source text and the target text [48]. Regarding the effectiveness of presenting target texts on overall quality, there are different opinions. The experiment conducted by Sun et al. [48] suggests that participants using such tools generally perform slightly better than those who do not. However, Xiao and Wang [49] argues that simulating machine-assisted functions does not significantly improve the output quality of English to Chinese simultaneous interpretation. In Sun et al. [48] experiment, many participants believed that the negative impact of machine translation mainly lies in the dispersion of interpreters' focus.

The third type is systems that display specific parts of texts, such as terms, numbers, or names. In previous versions, interpreters manually looked up words during SI. Zhang [5] concluded that the software helps interpreters render terms more accurately. Zhou [4] also reached a similar conclusion and was convinced that it improves overall performance, despite some negative effects on the renditions of certain participants. Although empirical studies support the notion that interpreters in the booth may have the time and cognitive ability to manually look up specialized terms, an automated querying system would undoubtedly be a step forward in reducing the additional cognitive effort needed for this human-machine interaction [50].

In addition to the aforementioned types of CAI tools, some researchers have conducted experiments using mock-up systems. For example, Desmet et al. [2] conducted SI experiments using a number recognition mock-up system, which resulted in a 30% increase in the accuracy of number interpreting. Furthermore, Defrancq and Fantinuoli [51] used a mock-up system that displays the whole source text in ASR and enlarges numbers (including English word numbers) for SI. Although this system has drawbacks, such as the need to manually scroll the latest overflowed source text, it demonstrated a 22.5% improvement for SI from English to Dutch and a 41.5% improvement for SI. Wang [52] also showed improved accuracy in terms and numbers. Moreover, researchers and translators have argued that presenting the entire transcript, either in the original source text or the translated target text, may overwhelm the user with excessive visual information [51].

To summarize, multiple empirical experiments have demonstrated that CAI tools greatly benefit interpreters in two main ways. The comprehensive workflow chart of the CAI system tailored for SI interpreters, based on ASR, is delineated in Model IV, Fig. 1). Firstly, the accuracy of ASR technology improves ceaselessly, allowing tools to provide interpreters with source text that has a lower error rate. This transforms the interpreters' workflow from the traditional process of ''listening, processing, and translating'' to a more efficient mode ''listening, referring to the provided text, and then translating'', reducing the interpreters' workload and alleviates psychological pressure. Secondly, the complexity of rendering specialized terms, numbers, etc. poses a cognitive challenge for interpreters, but CAI tools help relieve this cognitive load, enabling interpreters to focus on other mentally demanding tasks.

III. METHODOLOGY

A. INTERPRETSIMPLE–A NEWLY-DESIGNED COMPUTER-AIDED INTERPRETING SYSTEM

Within diverse studies for SI, efforts have been directed towards evaluating these tools through various experiments. Owing to the nascent stage of in-process CAI systems, there is still no study focusing on the implementation of a real automated CAI tool [3], [11]. There persists a deficiency in an interpreter-friendly approach that directly addresses issues and materializes theory into a practical system. This study aims to amalgamate strengths and mitigate the shortcomings identified in earlier empirical research, culminating in the creation of a novel CAI system suitable for SI interpreters. The evaluation encompasses a scrutiny of the tool's functionalities.

The specific steps of the proposed method are detailed as follows:

Step 1: Collect and analyze the strengths and weaknesses identified in prior empirical research on simultaneous interpreting and CAI experiments, systematically summarizing the functionalities requisite for CAI tools.

Step 2: Design the tool's interface and settings, producing initial sketches.

Step 3: Execute specific functionalities through programming.

Step 4: Monitor and authenticate the real-time system status to evidence the successful execution of diverse functions and address any issues for enhancement.

The software is developed on the.NET Framework 4.7, utilizing Visual Studio Community 2022 for programming in C#. It necessitates a Windows system compatible with the installation of.NET Framework 4.5 or higher for operation. The Tencent Cloud Speech-to-Text API (https://www.tencentcloud.com/?lang=en&pg=) is chosen as the primary ASR technology.

We named this newly developed software 'InterpretSIMPLE', as 'SIMPLE' is short for 'Simultaneous Interpreting Magic Potion for Language Exchanges', hoping this study would not only provide reference for the research and development of CAI systems, but also make contributions to the evolution of SI working mode in multilingual conferences.

B. DESIGN PHILOSOPHY

The design philosophy of InterpretSIMPLE prioritizes efficiency and reduced cognitive load in simultaneous interpretation (SI). The system focuses on a simplified and clean interface, integrating features to minimize manual operations for interpreters working under high cognitive strain. Additionally, the integration of high-accuracy ASR technology aims to display only relevant, real-time segments of the source text to avoid visual overload. Furthermore, automated glossary and translation integration, along with the exclusion of punctuation in the display, facilitate real-time reference, alleviating cognitive strain and improving accuracy for the interpreters. This philosophy underscores the emphasis on simplicity, accuracy, and cognitive load reduction in enhancing the interpreting process.

1) EFFICIENCY IN SIMULTANEOUS INTERPRETATION THROUGH INTERFACE DESIGN

The clean interface with integrated buttons aims to reduce manual operations during work. Given the demanding nature of SI, interpreters grapple with high cognitive strain while engaging in multiple tasks simultaneously, including listening, comprehension, translation, text production, and monitoring [1]. To alleviate cognitive strain, CAI tools in SI should prioritize a simplified user interface, minimizing visual distractions, and reducing manual operations. Previous experimental methods requiring manual scrolling for page turning such as [51] should be replaced with automatic algorithms.

2) INCORPORATING HIGH-ACCURACY AUTOMATIC SPEECH RECOGNITION FOR DISPLAYING SOURCE TEXT IN REAL-TIME

ASR technology transforms the speaker's language into text, acknowledged for its positive impact on translation quality [2], [42], [43], [44], [45], [51]. However, the accuracy of ASR significantly influences its assistance in SI, demanding a minimum accuracy above 85%, preferably exceeding 95% [46].

3) OPTIMIZING SOURCE TEXT DISPLAY FOR ENHANCED REAL-TIME INTERPRETATION

Displaying extensive segments of source text generates visual overload, complicating an interpreter's information gathering [51]. Hence, the ASR display should feature only the latest sentence or a few sentences in a fixed position on the interface, ensuring interpreters can swiftly locate real-time information when necessary.

4) AUTOMATED INTEGRATION OF GLOSSARIES AND TRANSLATIONS FOR IMPROVED INTERPRETATION

To address interpreters dealing with varied specialized topics beyond their expertise [40], integrating glossaries and translations in ASR is critical. Preparation for terminology and phraseology is essential for interpretation, minimizing cognitive load and bridging linguistic gaps [1], [9]. Yet, interpreters cannot memorize all terms and might encounter unfamiliar ones, posing challenges. Numbers are also among the most dreaded source-text features in SI, with interpreters reporting them as an important stress factor [54]. Therefore, Terminology, numbers, acronyms, and proper names have always been seen as typical 'problem triggers' in SI [50]. Swift term queries and highlighting numbers in SI are known to enhance translation accuracy [2], [4], [5], [51]. Integrating glossaries and translations in ASR allows real-time reference to numerous terms, alleviating cognitive strain and enhancing interpretation accuracy.

5) ELIMINATING PUNCTUATION DISPLAY IN REAL-TIME INTERPRETATION

As simultaneous interpreters operate based on ''meaning group'' rather than full sentences, the role of punctuation is minimal. Errors in punctuation during speech-to-text conversion can disrupt interpreter understanding and translation [47], [53]. Thus, omitting punctuation marks in the display aims to avoid confusing interpreters and improve their work efficiency.

C. EXPERIMENTS WITH THIS SYSTEM

The comprehensive evaluation of InterpretSIMPLE, the newly developed CAI system was conducted via a meticulously organized real-time video speech recognition experiment. This experiment served as the pivotal method to gauge the system's proficiency and reliability in a dynamic environment. Through this carefully constructed assessment, several critical aspects of the system's performance were measured, providing a detailed understanding of its capabilities and areas for potential enhancement.

The primary focus of the evaluation was to assess the system's competence in four significant domains: Word Recognition Accuracy, Terminology Recognition and Matching, Number Recognition, and Punctuation Recognition, etc.

The meticulous assessment aimed to provide a comprehensive analysis of the system's performance in various crucial areas, offering insights into its strengths and areas for potential refinement in aiding the field of interpreting and real-time speech transcription.

To improve the system's performance, modifications were made to the display format of terminological controls and the inclusion of English numeral words in the terminology list.

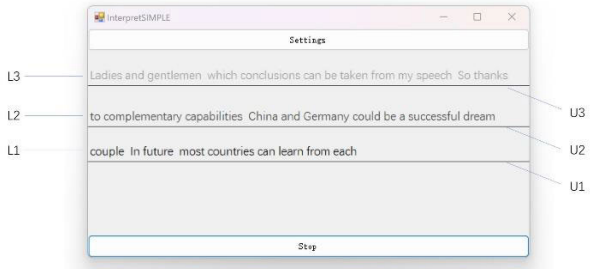

FIGURE 3. The MainForm interface.

IV. AN OVERVIEW OF THIS SYSTEM

A. INTERFACE DESIGN

1) THE MAINFORM

During SI, the system interface, functioning as the primary interface (the MainForm), is utilized for configuring options, initiating and terminating program execution, providing real-time displays of source text and terminology, executing terminology translation, and presenting numerical representations. In contrast to similar programs that often incorporate an array of setting buttons within the MainForm and lead to an overwhelming visual information load that hampers the efficiency of simultaneous interpreters, the MainForm of this program takes a streamlined approach. Aside from the standard options such as ''maximize,'' ''minimize,'' and ''close'' situated in the upper right corner, it features only two control buttons positioned at the top and bottom of the window, which are specifically labeled ''Settings'' and ''Start''. All configurable settings are incorporated within the ''Settings'' button, which, upon activation, opens the settings window. These buttons share the same width as the window, enhancing user identification and navigation. The ''Start'' button at the bottom is also of equal width to the window for users' convenience. A single click on this button swiftly transitions it to ''Stop,'' serving the dual purpose of initiating transcription and terminating recognition. Throughout the entire SI workflow, the ''Start/Stop'' button is the sole control requiring activation, thereby minimizing visual load and cognitive burden for interpreters.

In the context of real-time interpretation, where interpreters consistently concentrate on the speaker's most recent one or two sentences, the display of too much content poses challenges in swiftly locating and discerning crucial points. To address this, this program integrates three label controls (identified as L1, L2, L3) within the MainForm, specifically designed to retain the transcription content of the three lines from the original text. The most recent output of the ASR system is constantly displayed at the bottom line (L1), allowing interpreters to precisely acquire the latest information without the need for continuous vertical scrolling as the original text accumulates. As the content in the L1 line is fully displayed, it automatically shifts to the middle line (L2), and so on. To enhance reference without compromising interpreters' work and focus, the fonts of L1, L2, and L3 gradually decrease in thickness. Critical elements such as matched terminology, translated terminology, and numbers are highlighted in bold beneath L1 (as depicted in Fig. 3), allowing interpreters to selectively refer to them. Additionally, underlines (U1, U2, U3) are added below L1, L2, and L3. The thickness of these underlines can be adjusted through settings, facilitating interpreters in more effectively locating pertinent information.

2) THE OPTIONFORM

The OptionForm typically refers to an interface within the system, which is designed for configuring various options, parameters, or preference settings. This form provides a interface that allows users to customize the functions, appearance, or other behaviors of the software to meet individual needs or specific workflow requirements.

The OptionForm activates upon clicking the ''Settings'' button on the MainForm, enabling users to make necessary adjustments on parameters. Within the MainForm, the following controls are incorporated: (1) Four Label Controls, labeled as ''Audio Input devices'', ''Source Language Models'', ''Underline Thickness'' and ''Font Size''; (2) Five Button Controls, labeled as ''No Punctuation,'' ''Match Plurals,'' ''Finish,'' ''Import Term List'' and ''Clear Term List'';

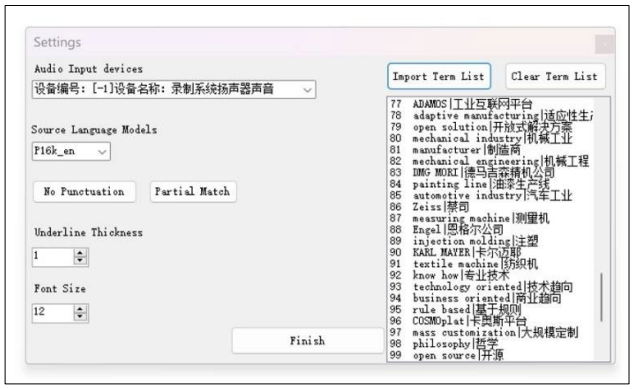

FIGURE 4. The OptionForm interface.

(3) A ListBox control, which is utilized for displaying the content of the user-imported terminology list.

Beneath the ''Audio Input devices'' and ''Source Language Models'', a ComboBox control is added for each, facilitating users to select from dropdown lists. For ''Audio Input Devices'', the dropdown list encompasses all audio devices detected by the program on the computer, allowing users to choose the input device for the source audio. For ''Source Language Models,'' the dropdown list includes 18 source languages provided by Tencent Cloud, comprising 23 Chinese dialects and other languages. Users can click to select the source language model. The ''With Punctuation'' button switches to ''No Punctuation'' upon clicking, allowing users to choose whether or not to display punctuations based on their preferences and different scenarios. Similarly, the ''Partial Match'' button switches to ''Exact Match'' to meet user preferences and specific matching requirements for terminology in different contexts. Beneath the ''Underline Thickness'' and ''Font Size'', a ''NumericUpDown'' control is added for each, facilitating users to select within the range of 0 to 100, where 0 indicates no display. A file selection window opens upon clicking the ''Import Term List'' button, allowing users to import a bilingual terminology list as needed. The selected content is displayed in the ListBox. By clicking the ''Clear Term List'', the imported terminology list and the content in the ListBox, which is used to show the user-imported terminology list, will be cleared. Once users have completed their settings, a click on the ''Finish'' button will apply the configurations and close the OptionForm.

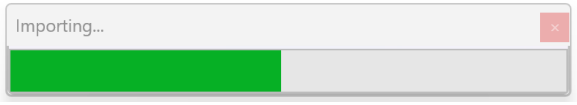

FIGURE 5. The ProcessBar interface.

3) THE PROCESSDIALOG

A ''ProcessBar'' within the ''ProcessDialog'' in this newly-designed system serves the purpose of visually indicating the progress of the ongoing operation. It provides interpreters with a clear visual indication of how much of the process is completed and how much is remaining. This visual feedback is crucial for managing interpreters' expectations and reducing uncertainty. Interpreters gain a sense of transparency and assurance regarding the progress of the task, which can enhance their experience, reduce frustration, and build trust in the system.

B. FUNCTIONS REALIZATION

1) MODULES

The system mainly includes three parts: Dynamic Link Library (DLL) module, MainForm module and OptionForm module.

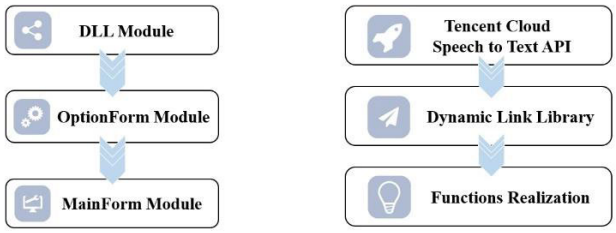

FIGURE 6. The modules and workflow of the system.

The implementation commences by establishing a Dynamic Link Library (DLL) that facilitates seamless integration with the Tencent Cloud Speech to Text API. The DLL, using the NAudio library (an open source audio processing class library for.NET platform) for audio processing, contains three classes, ''VoiceDevice'', ''VTTSentence'' and ''VoiceToText''. The VoiceDevice class represents audio devices including device index and device names. The VTTSentence class represents the parsing of the CVTT (Cloud Voice To Text) output sentence, defines different status and some related properties and methods related to the sentences such as status, null or not, status code, start time, end time, text content, etc. The VoiceToText class provides the functions of initializing and ending the real-time speech-to-text(STT), accessing the status of STT and the currently received message, as well as some related properties and methods such as maximum duration, audio input device number, APPID(Application Identification), SecretID, SecretKey, obtaining audio input devices list and generating random UUID (Universally Unique identifiers), etc.

The OptionForm module facilitates function selection and configuration through customizable buttons including ''Audio Input devices'', ''Source Language Models'', ''No/With Punctuation'', ''Partial/Exact Match'', ''Underline Thickness'', ''Font Size'', ''Import Term List'' and ''Clear Term List''.

The MainForm module enables the real-time display of text, terms and translations, and numbers. In the MainForm class, private variables and those interacting with other forms are defined, and three additional classes—TrWord, WordAndLocation, and TrWordLine—are created. The TrWord class stores numbers, terms and translations, access and manipulate words through properties; the WordAndLocation class stores the TrWord class objects and

their operation; the TrWordLine class stores TrWord objects and their corresponding location information for related operations and processing. Through the GetLocations method in the MainForm class, it searches the location of the terms or numbers in the given text, returning a list of location information. Subsequently, it updates the term list in the TrWordLine object and stores the location information of matching terms in the WordLocationList. By getting the term label control on the interface, it iterates through the pre-imported term list, updating the position of the term label control in the MainForm interface or creating a new term label control, thereby achieving the match and display of terms, term translations, numbers, and other functionalities like ''Partial/Exact Match,'' ''No/With Punctuation,'' etc.

2) WORKFLOW

First of all, by establishing the DLL, it realizes the seamless integration with Tencent Cloud Speech to Text (STT) API. The MainForm interface displays real-time STT, the match of terms, numbers and so on. While receiving the real-time speech-to-text(STT) information from Tencent Cloud, it judges whether it is a new line of text. Upon detection of a new line, it updates the three label controls on the MainForm interface (Fig. 3), assigning the latest text to L1, the original L1 text to L2, and the L2 text to L3. This arrangement ensures that interpreters can accurately locate the latest text in L1 without the need to shift their gaze up or down. Texts in L2 and L3 become lighter in turn, providing an auxiliary reference without disrupting the interpreter's attention during SI. Three panel controls, U1, U2 and U3, set under L1, L2 and L3 respectively, are used to display underlines(if the thickness is zero, no underlines displayed) to further assist interpreters in locating information.

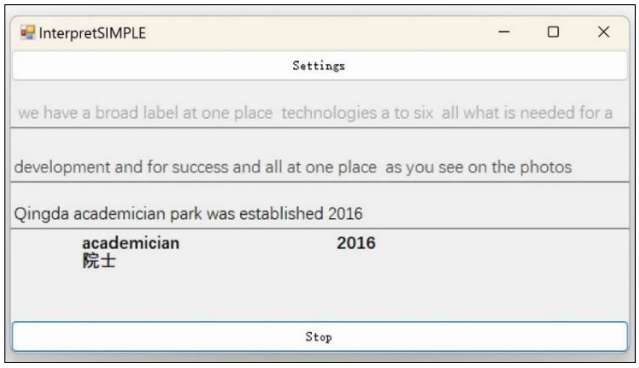

FIGURE 7. Terminology and number matching and punctuation display.

Glossary importing:

The LoadExcel method (for importing a glossary and displaying it in the OptionForm interface) is defined in the OptionsForm. To speed up the importing process, especially if the glossary contains a large number of terms and translations(such as thousands or tens of thousands), it creates a new thread and set the CPU multi-core startup thread, thereby expediting the loading process and mitigating memory consumption. It executes the LoadExcel method in the new thread:

Upon activating the ''Import Term List'' button, a file selection window pops up, allowing users to choose an excel file. It iterates each items in the file, adds each one to the WordList list of the MainForm, and displays it in the listBox control of the OptionForm interface.

Terminology and number matching:

The GetLocations method is defined in the MainForm class to access the location information for a specified term or number in the text message and subsequently return a list of location information:

Upon receiving the text string of Tencent Cloud Speech to Text API, it detects whether it is a number. When identifying numbers, it carries out the number matching by using regular expression to match the locations of numbers (including those presented as percentages or accompanied by percent signs) and iterates each of them. For each matching item, it calculates the width of the text preceding it, adds the numeric object and location to the WordLocationList list, returns the information in the WordLocationList list, updates the location of the term label control or create a new term label control, and display the bold number below the number.

In instances where the identified entity is not a number, the system proceeds to iterate through the term list, WordList for term matching. The location of the specified term is matched using regular expression, and each match is iterated, calculating the width of the text preceding it, adding the term object and location to the WordLocationList. Following the retrieval of information from the WordLocationList, the system updates the location of the term label control or creates a new term label control, showcasing the bolded term and its translation below it.

Partial/exact match

Given the existence of various forms of English words, including plurals, third-person singular, continuous tenses, and more, a precise match solely based on the exact letter-for-letter correspondence of each word is bound to result in failure. In response to this challenge, the program incorporates an optional feature within the GetLocations method:

Partial Match. In regular expression, it matched the word and the preceding space. For example, if there is ''person'' in the imported term list, it can match '' a lot of *persons* are…'', to match partial terms with suffixes such as singular, plural, continuous tense, etc.

It also provides an exact matching mode to facilitate the selection in different situations:

Exact Match. In regular expression, it matched the word and both the spaces preceding and followed.

For example, if there is ''person'' in the imported term list, it cannot match ''a lot of *persons* are…''.

V. VALIDATION VIA EXPERIMENT

In the realm of computer-aided interpreting systems, theoretical designs and simulated scenarios lay the groundwork, yet the ultimate test lies in real-world application. The necessity

of conducting experiments in real-life scenarios to evaluate the efficiency and functionality of a newly-designed computer-aided interpreting system cannot be overstated. Only through real-life trials can the system's adaptability, accuracy, and capacity to overcome real-world challenges be fully assessed, ultimately validating its applicability in multilingual, dynamic communication landscapes. The efficiency of the newly-designed computer-aided interpreting system underwent validation through a real-time video speech recognition experiment. This speech, conducted in English, focused on the topic of Industrial Internet, a notably popular research domain within multilingual conferences in recent times. The presentation was delivered by Prof. h.c. Dr.-Ing. Ömer Sahin Ganiyusufoglu, a member of German National Academy of Science and Engineering. The speech took place during the inaugural session of the 2021 World Industrial Internet Conference (WIIC) held in Qingdao, China, on October 27, 2021. Despite English not being the speaker's native language, the presence of accents is customary within multilingual gatherings and has no bearing on the objectivity of this particular experiment.

A. SPEECH

As the primary English-Chinese translator and interpreter for the conference, the first author of this paper acquired the video recordings of the speeches from the conference organizers. The original video duration spans 15 minutes and 38 seconds, with the experimental speech commencing at the 11-second mark, resulting in an experimental speech duration of 15 minutes and 27 seconds. This segment comprises a total of 1389 words, with an average speech rate of 90 words per minute. The speech incorporates a total of 119 terminologies(A term is counted as one instance irrespective of the number of words it encompasses; if a term appears multiple times, each occurrence is individually counted) and eight numbers.

B. EXPERIMENT PROCEDURE

The CAI program was installed on a laptop operating with a Windows 10 system integrated with .NET Framework 4.7. Within the ''Settings'' configuration, the ''Audio Input devices'' were specifically designated as ''System Sound,'' and the ''Source languages'' were set to ''16k_en: English General''. Additional preferences included choosing ''No Punctuation'', opting for ''Partial Matching'', and adjusting the ''Underline Thickness'' to ''1'' with the ''Font Size'' set at ''12.'' An Excel spreadsheet containing a compiled list of 75 English terminologies and their corresponding Chinese translations was imported and organized. Notably, adaptations were made to accommodate recognition variations, generating alternative forms for specific terms; for instance, the acronym BMW could be recognized by ASR as either BMW or B M W with spaces between letters. The categorization also included irregular plural forms and tense-related suffixes, like capability and capabilities, and terms displaying multiple forms, such as inter operability or interoperability (both considered as a singular term). Upon importing the terminology list, settings were saved by confirming the selection with ''OK.'' To initiate the video recording process, the ''Start'' button on the primary program interface was activated, and upon the completion of the video, the ''Stop'' button was pressed to conclude the recording. The comprehensive process captured both the computer screen and accompanying audio.

TABLE 1. Acquired data in the experiment.

	Time	Average Speed	Words	Terminologies	Numbers
Speech	15m 27s	90 words/s	1389	119	8
ASR	/	/	1335	93	8
Accuracy Rate of ASR	/	/	96.11%	78.15%	100%
Interpret-SIMPLE	/	/	/	92	8
Accuracy Rate of Interpret-SIMPLE	/	/	/	98.92%	100%

C. DATA ANALYSIS

We classify the accuracy of Tencent Cloud Speech to Text and the accuracy of InterpretSIMPLE. The ASR identification errors are as the following categories:

- Omission(the word is missing)
- Approximation(the alphabetical order are similar, eg. Ones become one)
- Phonological mistakes(phonological confusion in the source stimulus, e.g. ''ADAMOS'' becomes ''othermost'', a near-homophone in English)
- Other mistakes(miscellaneous errors that do not fit any of the other categories)

The total word count in the source speech amounted to 1389, and all the recognition words belong to the categories mentioned above(including approximation) are not counted among correctly recognized words. While accurately recognizing 1335 words, the Tencent Cloud Speech to Text achieved an accuracy rate of 96.11%(as in TABLE 1). This performance surpassed the minimum threshold of 85% and met the superior standard of 95% as outlined by Lin [46], affirming its supportive role for a CAI system in SI.

The total word count in the experimental video amounted to 1389, with Tencent Cloud Speech to Text accurately recognizing 1335 words, achieving an accuracy rate of 96.11%. This performance surpassed the minimum threshold of 85% and met the superior standard of 95% as outlined by Lin (2013), affirming its supportive role for SI. While the original video contained 119 terminologies, Tencent Cloud Speech to Text identified 93 terminologies. Recognition issues predominantly involved Chinese names and specific proper nouns. This CAI program successfully matched 92 terminologies, yielding an accuracy rate of 98.92% (rounded to two decimal

places). The sole discrepancy occurred with the proper noun ''Pingan,'' a Fortune Global 500 Chinese company, recognized as ''ping an'' by Tencent, leading to the CAI's failed match.

We divided the display types of numbers into the following two categories:

- Arabic numeral (eg. 100)
- English word (eg. hundred)

Out of the total 8 numerical instances in the original text, all were successfully identified by Tencent Cloud Speech to Text, with three identified as Arabic numerals and accurately matched by the program, achieving a 100% matching rate. The remaining five numbers were not recognized as Arabic numerals during the recognition process and were displayed in English word form, which were also fully matched. In terms of punctuations, the program displayed 0 instances of recognized punctuation, attaining a recognition rate of 100%.

The newly-developed CAI system demonstrated notable efficiency in supporting simultaneous interpreting. It showed a high level of accuracy in recognizing words and numbers. While encountering challenges with certain terminologies, particularly related to Chinese names and specific proper nouns, the system's overall performance in recognizing terms was commendable. It effectively matched the majority of terminologies, with only a few discrepancies noted, such as the recognition issue with the proper noun ''Pingan.'' In terms of numerical recognition, the system accurately identified Arabic numerals but faced challenges when numbers were displayed as English words. Remarkably, the system flawlessly recognized zero instances of punctuation. Overall, despite some minor recognition discrepancies, the system's performance marks a significant advancement in enhancing efficiency and accuracy in simultaneous interpreting.

VI. CONCLUSION AND FUTURE WORK

After a century of development, SI has evolved into an indispensable mode of translation for international conferences and collaborative exchanges. Recent strides in AI particularly in deep learning and neural networks, have significantly enhanced the quality of ASR, MT, and other natural language processing technologies, enabling AI-powered SI. However, issues such as accuracy and non-linguistic factors continue to impede AI's ability to replace human interpreters. Nevertheless, AI can support human interpreters by providing CAI, lessening cognitive load and enhancing overall translation quality, especially in terms of technical terminology, numerical data, and overall linguistic accuracy.

This academic paper introduces various methods designed to aid interpreters in SI through ASR technology and presents the development of a CAI system featuring user-friendly interaction and real-time functionalities such as instantaneous terminology and numerical matching. Experimental evaluations demonstrate that the system achieves over 98% accuracy in terminology matching and 100% accuracy in Arabic numeral matching. Additionally, improvements in the display of terminology and numerical information address issues related to dense terminologies and the accurate matching of English words and numbers. As the first detailed research of an automated CAI system available for in-process SI, the study proves the feasibility and high accuracy to automatically extract and display terms and numbers and to eliminate punctuations, and realizes a design of CAI tool with user-friendly interface. This study, by shedding fresh light on enhancing the efficiency and accuracy of SI, provide valuable insights into the complexities encountered and guiding potential improvements in CAI systems. The experimental findings offer a foundation for refining recognition processes within diverse linguistic and terminological contexts.

Be that as it may, this study has its own limitations. On that account, future research will concentrate on reducing the algorism time-complexity, introducing new statistical analysis to investigate the precision and distribution of the solutions [54], as well as devising matching solutions for different forms of terminology and refining more efficient approaches for numerical matching.

REFERENCES

[1] D. Gile, *Basic Concepts and Models for Interpreter and Translator Training*. Amsterdam, The Netherlands: John Benjamins Publishing Company, 2009.

[2] B. Desmet, M. Vandierendonck, and B. Defrancq, ''Simultaneous interpretation of numbers and the impact of technological support,'' in *Interpreting and Technology*, C. Fantinuoli, Ed. Berlin, Germany: Language Science Press, 2018, pp. 13–27.

[3] F. Frittella, *Usability Research for Interpreter-Centred Technology: The Case Study of SmarTerp*. Berlin, Germany: Language Science Press, 2023.

[4] L. Zhou, ''The impact of computer-aided interpreting tools on simultaneous interpreting performance: Taking InterpretBank as an example,'' MTI. thesis, Dept. Interpret. Eng., Xiamen Univ., Xiamen, China, 2019.

[5] J. Zhang, ''An experiment report on the impact of computer-aided interpreting tools on simultaneous interpreting,'' MTI. thesis, Dept. Interpret. Eng., China Foreign Affairs Univ., Beijing, China, 2021.

[6] T. Ge, ''Usability of terminology-assistance in Chinese to English simultaneous interpretation-taking InterpretBank as an example,'' MTI. thesis, Dept. Interpret. Eng., Beijing Foreign Studies Univ., Beijing, China, 2023.

[7] S. T. Winteringham, ''The usefulness of ICTs in interpreting practice,'' *Interpret. Newslett.*, vol. 23, no. 15, pp. 87–99, 2010.

[8] Y. Dong and D. Li, *Automatic Speech Recognition: A Deep Learning Approach*. Berlin, Germany: Springer, 2015.

[9] C. Fantinuoli, ''Computer-assisted preparation in conference interpreting,'' *Transl. Interpreting*, vol. 9, no. 2, pp. 24–37, Jul. 2017, doi: 10.12807/ti.109202.2017.a02.

[10] M. Guo, L. Han, and M. T. Anacleto, ''Computer-assisted interpreting tools: Status quo and future trends,'' *Theory Pract. Lang. Stud.*, vol. 13, no. 1, pp. 89–99, Dec. 2022, doi: 10.17507/tpls.1301.11.

[11] B. Prandi, *Computer-Assisted Simultaneous Interpreting: A Cognitive-Experimental Study on Terminology*. Berlin, Germany: Language Science Press, 2023.

[12] C. Fantinuoli, ''Interpreting and technology: The upcoming technological turn,'' in *Interpreting and Technology*, C. Fantinuoli, Ed. Berlin, Germany: Language Science Press, 2018, pp. 1–12.

[13] J. Delisle and J. Woodsworth, *Translators Through History*. Amsterdam, The Netherlands: John Benjamins Publishing Company, 2012.

[14] D. Gerver, ''Empirical studies of simultaneous interpretation: A review and a model,'' in *Translation: Applications and Research*, R. W. Bristlin, Ed. New York, NY, USA: Gardner Press, 1976, pp. 165–207.

[15] B. Moser-Mercer, ''Simultaneous interpretation: A hypothetical model and its practical application,'' in *Language interpretation and communication*, D. Gerver and H. W. Sinaiko, Eds. Boston, MA, USA: Springer, 1978, pp. 353–368.

[16] D. Gile, *Basic Concepts and Models in Interpreter and Translator Training*. Shanghai, China: Shanghai Foreign Language Education Press, 2011.

[17] R. Setton, *Simultaneous Interpreting: A Cognitive-Pragmatic Analysis*. Amsterdam, The Netherlands: John Benjamins Publishing Company, 1999.
[18] A. De Groot, "A complex-skill approach to translation and interpreting," in *Tapping and Mapping the Processes of Translation and Interpretin*, S. Tirkkonen-Condit and R. Jääskeläinen, Eds. Amsterdam, The Netherlands: John Benjamins, 2000, pp. 53–68.
[19] B. Moser-Mercer, "Beyond curiosity: Can interpreting research meet the challenge?" in *Cognitive Processes in Translation and Interpreting*, J. H. Danks, S. B. Fountain, M. K. McBeath, and G. M. Shreve, Eds. Thousand Oaks, CA, USA: Sage, 1997, pp. 176–195.
[20] J. Rinne, J. Tommola, M. Laine, B. Krause, D. Schmidt, V. Kaasinen, M. Teräs, H. Sipilä, and M. Sunnari, "The translating brain: Cerebral activation patterns during simultaneous interpreting," *Neurosci. Lett.*, vol. 294, no. 2, pp. 85–88, Nov. 2000, doi: 10.1016/S0304-3940(00)01540-8.
[21] H. Xiong, R. Zhang, C. Zhang, Z. He, H. Wu, and H. Wang, "Dutongchuan: Context-aware translation model for simultaneous interpreting," 2019, *arXiv:1907.12984*.
[22] K. H. Davis, R. Biddulph, and S. Balashek, "Automatic recognition of spoken digits," *J. Acoust. Soc. Amer.*, vol. 24, no. 6, pp. 637–642, Nov. 1952, doi: 10.1121/1.1906946.
[23] B. H. Juang and L. R. Rabiner, "Hidden Markov models for speech recognition," *Technometrics*, vol. 33, no. 3, pp. 251–272, Aug. 1991, doi: 10.1080/00401706.1991.10484833.
[24] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, Jul. 2006, doi: 10.1162/neco.2006.18.7.1527.
[25] D. Yu and L. Deng, "Deep learning and its applications to signal and information processing [exploratory DSP]," *IEEE Signal Process. Mag.*, vol. 28, no. 1, pp. 145–154, Jan. 2011, doi: 10.1109/MSP.2010.939038.
[26] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.
[27] Y. Zhang, G. Chen, D. Yu, K. Yao, S. Khudanpur, and J. Glass, "Highway long short-term memory RNNS for distant speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Sign. (ICASSP)*, Shanghai, China, Mar. 2016, pp. 5755–5759.
[28] M. Nagao, "A framework of a mechanical translation between Japanese and English by analogy principle," in *Proc. Int. NATO Symp. Artif. Human Intell.*, A. Elithorn and R. Banerji, Eds. Oct. 1984, pp. 173–180.
[29] P. F. Brown, J. Cocke, S. A. Pietra, V. Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, and P. S. Roossin, "A statistical approach to machine translation," *Comput. Linguist.*, vol. 16, no. 2, pp. 79–85, Jun. 1990.
[30] F. J. Och and H. Ney, "The alignment template approach to statistical machine translation," *Comput. Linguistics*, vol. 30, no. 4, pp. 417–449, Dec. 2004, doi: 10.1162/0891201042544884.
[31] D. Chiang, "Hierarchical phrase-based translation," *Comput. Linguistics*, vol. 33, no. 2, pp. 201–228, Jun. 2007, doi: 10.1162/coli.2007.33.2.201.
[32] Z. Feng, "The past and present of natural language processing," *Fore. Lang. Chin.*, vol. 5, no. 1, pp. 14–22, 2008, doi: 10.3969/j.issn.1672-9382.2008.01.003.
[33] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. 31st Interfaces Conf. Neural Inf. Proc. Sys.*, U. von Luxburg, I. Guyon, S. Bengio, H. Wallach R. Fergus, Eds. New York, NY, USA: Curran Associates, Dec. 2017, pp. 5998–6008.
[34] M. Ma, L. Huang, H. Xiong, R. Zheng, K. Liu, B. Zheng, C. Zhang, Z. He, H. Liu, X. Li, H. Wu, and H. Wang, "STACL: Simultaneous translation with implicit anticipation and controllable latency using prefix-to-prefix framework," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, Florence, Italy, 2019, pp. 3025–3036.
[35] N. Arivazhagan, C. Cherry, W. Macherey, C.-C. Chiu, S. Yavuz, R. Pang, W. Li, and C. Raffel, "Monotonic infinite lookback attention for simultaneous machine translation," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, Florence, Italy, 2019, pp. 1313–1323.
[36] R. Zhang and C. Zhang, "Dynamic sentence boundary detection for simultaneous translation," in *Proc. 1st Work Auto. Simul. Trans.*, Seattle, WA, USA, Jul. 2020, pp. 1–9.
[37] N. Arivazhagan, C. Cherry, W. Macherey, and G. Foster, "Re-translation versus streaming for simultaneous translation," 2020, *arXiv:2004.03643*.
[38] L. E. S. Ortiz and P. Cavallo, "Computer-assisted interpreting tools (CAI) and options for automation with automatic speech recognition," *Tradterm*, vol. 32, pp. 9–31, Dec. 2018, doi: 10.11606/issn.2317-9511.v32i0p9-31.
[39] B. Prandi, "An exploratory study on CAI tools in simultaneous interpreting: Theoretical framework and stimulus validation," in *Interpreting and Technology*, C. Fantinuoli, Ed. Berlin, Germany: Language Science Press, 2018, pp. 29–59.
[40] C. Fantinuoli, "Conference interpreting and new technologies," in *The Routledge Handbook of Conference Interpreting*, M. Albl-Mikasa and E. Tiselius, Eds. London, U.K.: Routledge, 2021, pp. 508–522.
[41] C. Fantinuoli, "InterpretBank: Redefining computer-assisted interpreting tools," in *Proc. 38th Conf. Trans. Comput.*, London, U.K., Nov. 2016, pp. 42–52.
[42] F. Pöchhacker, *Introducing Interpreting Studies*, 2nd ed., London, U.K.: Routledge, 2016.
[43] Y. Ma, "The impact of speech recognition software on C-E simultaneous interpreting based on Gile' s effort model: An exploratory study," MTI. thesis, Dept. Interpret. Eng., Beijing Foreign Studies Univ., Beijing, China, 2020.
[44] C. He, "An empirical study on how voice recognition technology influences Chinese-English simultaneous interpretation of numbers," M.A. thesis, Dept. Lingust. Eng., Shanghai Int. Studies Univ., Shanghai, China, 2018.
[45] J. Xiang, "Speech recognition software: Friend or foe for trainees of Chinese-English simultaneous interpretation?" M.A. thesis, Dept. Lingust. Eng., Beijing Foreign Studies Univ., Beijing, China, 2018.
[46] X. Lin, "An empirical study on computer aided interpretation from English to Chinese," M.A. thesis, Dept. Lingust. Eng., Shandong Normal Univ., Jinan, China, 2013.
[47] C. Xue, "Impact of real-time speech recognition on interpreting quality of E-C simultaneous interpretation of impromptu speech: An empirical study," MTI. thesis, Dept. Interpret. Eng., Beijing, China, 2022.
[48] H. Sun, K. Li, and J. Lu, "AI-assisted simultaneous interpreting—An experiment and its implication," *Tech. Enhan. Fore. Lang.*, vol. 43, no. 6, pp. 75–86, 2021.
[49] L. Xiao and Y. Wang, "An intervention study of simulated artificial intelligence auxiliary functions in E-C simultaneous interpretation," *J. Guangdong Univ. Educ.*, vol. 40, no. 6, pp. 52–57, 2020.
[50] C. Fantinuoli, "Speech recognition in the interpreter workstation," in *Proc. Transl. Comput.*, London, U.K., 2017, pp. 25–34.
[51] B. Defrancq and C. Fantinuoli, "Automatic speech recognition in the booth: Assessment of system performance, interpreters' performances and interactions in the context of numbers," *Target*, vol. 33, no. 1, pp. 73–102, 2021, doi: 10.1075/target.19166.def.
[52] Y. Wang, "Impacts des technologies d' interpretation assistee par ordinateur sur la qualite de l' interpretation simultanee francais-chinois," M.A. thesis, Beijing Forestry Univ., Beijing, China, 2023.
[53] J. Li, "Application of computer translation assistance technology in simultaneous transmission and its impact on the ecosystem of simultaneous transmission," *Chin. Transl. J.*, vol. 41, no. 4, pp. 127–132, 2020.
[54] M. Mollajafari and M. Shojaeefard, "TC3PoP: A time-cost compromised workflow scheduling heuristic customized for cloud environments," *Cluster Comput.*, vol. 24, pp. 2639–2656, Sep. 2021, doi: 10.1007/s10586-021-03285-5.

JICHAO LIU was born in Qingdao, Shandong, China, in 1990. He received the master's degree in interpreting from Nankai University, China, in 2014. He is currently pursuing the Ph.D. degree in translation studies with the School of Foreign Studies, Tongji University, China, with a focus on AI-assisted simultaneous interpreting.

From 2014 to 2018, he was a full time Engineer in international cooperation with the Institute of Oceanology, Chinese Academy of Sciences (CAS). He has studied scientific dissemination under the guidance of Prof. Shouyi Zheng, an academician of CAS, since 2018, and studied human–computer interaction in the Industrial Internet under the guidance of Prof. h.c. Dr.-Ing. Ö. S. Ganiyusufoglu, a member of the National Academy of Science and Engineering (acatech), since 2022. He was also employed as a part-time Teacher by China Translation Corporation with years of interpreting and translation experience. His research interests include computer-aided interpreting, translation, and dissemination of science and culture.

CHENGPAN LIU was born in Heze, Shandong, China, in 1989. He received the bachelor's degree in English studies from Harbin Institute of Technology, China, in 2012, and the master's degree in interpreting from Nankai University, China, in 2014.

He is currently a Lecturer with Sino-European Institute of Aviation Engineering, Civil Aviation University of China. He is hosting a research project funded by Tianjin Municipal Education Commission and completed a number of projects supported by Fundamental Research Funds for the Central Universities and Teaching Reform Projects. He has translated three books, authored one textbook, and published more than a dozen journal articles. His research interests include translation, interpretation, English for specific purposes, and cognitive linguistics.

BUZHENG SHAN was born in Qingdao, Shandong, China, in 2000. He received the B.Sc. degree in mathematics and applied mathematics from Tongji University, Shanghai, China, in 2022.

He is currently pursuing the Ph.D. degree in mathematics from Texas A&M University, College Station, TX, USA. His research interests include computational and applied mathematics and related areas.

ÖMER S. GANIYUSUFOGLU received the Bachelor of Science degree from the Technical University of Berlin, Germany, in 1979, and the Ph.D. degree in mechanical engineering from the Technical University of Berlin in 1984, with a focus on machine tools and manufacturing technology.

He worked as a Research Associate (Wissenschaftlicher Mitarbeiter) with Institut for Machine Tools and Manufacturing Technology and Fraunhofer Institut for Production Technology and Design (IPK–Berlin), directed by Prof. Dr. Spur. From 1985 to 1989, he worked with German CNC lathe manufacturer Traub, as the Head of Automation. In 1990, he joined Yamazaki Mazak Germany, and worked as the Managing Director, until 2005. He joined German CNC lathe manufacturer Index-Werke, in 2006, and moved to China, as the General Manager of Index Dalian Machine Tool Ltd., which is a Joint Venture between Dalian Machine Tool Group and Index-Werke Germany. In 2011, he joined Shenyang Machine Tool Group (SYMG) Company Ltd., Shenyang, China. As a Senior Consultant to the Chairman of the Group, he supported the company in terms of strategy, globalization and international cooperation. Since 2021, he is acting as an Industrial Development Consultant with Qingdao International Academician Park (QIAP). He is an Advisory Professor with the School Mechanical Engineering (SME), Tongji University, Shanghai, an Honorary Professor with Nanjing University for Aeronautics and Astronautics (NUAA), and a Visiting Professor with the Capital University for Economy and Business (CUEB), Beijing, China. Within the International Academy of Production Engineering (CIRP), he is representing as a Corporate Member with QIAP. From 2016 to 2019, he was the Chairman of the Corporate Members. He has been a member of the German Engineers' Association (VDI), since 1977. He is a recipient of the Best Labor Award of the City of Dalian, the Rose Prize of the City of Shenyang, and the Friendship Award of Liaoning Province in China, and the Friendship Award of Central Government in Beijing, and the highest award a foreigner can receive.