International Journal of Translation, Interpretation, and Applied Linguistics
Volume 2 • Issue 2 • July-December 2020

Attention-Sharing Initiative of Multimodal Processing in Simultaneous Interpreting

Tianyun Li, Shandong Univerisity (Weihai), China
https://orcid.org/0000-0003-4950-7358

Bicheng Fan, Sogou Research Institute, Hangzhou, China
https://orcid.org/0000-0003-4676-9117

ABSTRACT

This study sets out to describe simultaneous interpreters' attention-sharing initiatives when exposed under input from both videotaped speech recording and real-time transcriptions. Separation of mental energy in acquiring visual input accords with the human brain's statistic optimization principle where the same property of an object is presented through diverse fashions. In examining professional interpreters' initiatives, the authors invited five professional English-Chinese conference interpreters to simultaneously interpret a videotaped speech with real-time captions generated by speech recognition engine while meanwhile monitoring their eye movements. The results indicate the professional interpreters' preferences in referring to visually presented captions along with the speaker's facial expressions, where low-frequency words, proper names, and numbers gained greater attention than words with higher frequency. This phenomenon might be explained by the working memory theory in which the central executive enables redundancy gains retrieved from dual-channel information.

KEYWORDS

Attention-Sharing, Multimodal Processing, Simultaneous Interpreting, Speech Recognition, Working Memory

INTRODUCTION

Simultaneous interpreting (SI) has been defined as one of the most difficult mental tasks since its emergence in the late 1940s. This cross-linguistic meaning transferring process involves numerous multi-channel cognitive sub-tasks, integrating active listening, memorization, streaming interpretation and coordination in a limited given chunk (Gile, 2009; Liu et al., 2004; Seeber, 2011; Stachowiak, 2014; Timarova, 2007). Drawn from existing theories, SI is concluded as an exhausting activity which pushed interpreters working close to saturation. The limited-capacity model propels interpreters to work close to cognitive saturation, like performing tightrope walking (Gile, 1999, 2009). It is plausible that workload survey results from AIIC members reported it burning out of their brain resources, leading to a maximum level of mental saturation, cognitive fatigue and stress when being on mike. Putting a single extra sub-task may result in cognitive overload, jeopardising the output quality in consequence. In the absence of an authentic benchmark in measuring quantified mental pressure, this

DOI: 10.4018/IJTIAL.20200701.oa4

local cognitive load may vary according to interpreter's working expertise, domain-specific knowledge accumulation, speaker's rate of speaking and task difficulty. While reading visual materials during an SI assignment will inevitably challenge the interpreter's multithreading mechanism, professional interpreters are still capable of handling this even-more intricate multi-tasking assignment, as simultaneous interpreting with text through accessing speech scripts or slides, or even sight translation pacing the speaker's rhythm, with background monitoring of the congruence between the speech flow and the textual version. In this case, the magnitude of cognitive load brought to a professional interpreter may not be measured in a static manner, like how many sub-tasks are generated varying from pure SI to SI with visual input.

Dual-channel input in simultaneous interpreting is a realistic solution for both the interpreting practitioner and the event organizer aiming for higher interpreting quality. This paper presents a probing study enquiring into professional interpreters' initiative in coping with multimodal processing in simultaneous interpreting with concurrent visual input, and to be more specific, their career intuition developed through long-term training in systematizing different information channels. Apart from the previous studies which put a major focus on "quality" or "performance" of simultaneous interpreting with text or other visual presentations, this present study turns to investigate the initial response under multiple-channel resources. It is also the first experimental study revealing some preliminary general characters of professional interpreters, who tackle with hearing thresholds that may arise due to problem triggers (for example, unfamiliar accent). The following sections will elaborate existing research on working memory and multimodal input, both in common language-learning scenarios and simultaneous interpreting. Then we will describe an analysis of experimental results on five conference interpreters who were asked to simultaneously interpret a videotaped speech with real-time captions generated by automatic speech recognition engine, in which their eye-movement data collected through a Tobii 4C eye tracker.

PROBLEM TRIGGERS AND PROCESSING LATENCY

Throughout a simultaneous interpreting assignment, an interpreter will inevitably encounter several stumbling blocks, for example, weak signal, an accented speaker and poor booth conditions. These elements are referred to as "problem triggers", which increases the interpreter's processing capacity in language comprehension or/and rendition production (Gile, 2009). A number of studies indicated the factors hindering the interpreter from an ideal rendition (Albl-mikasa, 2010; Cheung, 2013; Han & Riazi, 2017; I. H. I. Lin et al., 2013). Based on Gile's conceptual framework, Mankauskienė (2016, pp. 145–146) presented a structured list looking into the definition of "problem trigger" in simultaneous interpreting. According to this classified structure, problem triggers that may appear during an interpreting assignment are concluded into four general categories: 1) an increase in processing capacity requirement; 2) signal vulnerability; 3) language-specified related problems; and 4) the speaker factor (see Table 1).

Interestingly, if some of the above-mentioned obstructions are applied to the scenario of translation quality, most of them will not construct impediments towards a translator. In the study of Munro & Derwing (1995b, 1995a), native English speakers heard a set of English statements spoken by speakers with native English accent and speakers with Chinese Putonghua (Mandarin) accent. Though it appeared some hearing problems when the same content was spoken by Chinese-accented speakers, their utterances were also "perfectly intelligible" (Munro & Derwing, 1995b, p. 291). An increase in processing time is detected as side effects: for Chinese-accented speakers, native English listeners require more time in making semantic judgements. There is no in-depth investigation throughout this set of experiments, yet this phenomenon was interpreted as communicative "costs" in comprehending foreign-accented speeches. If we elaborate the latency data through the fuzzy logical model of speech perception (Jesse et al., 2000; Massaro, 1989), we may raise a hypothesis that problems occur at the "integration to decision" phase. At this stage, the incoming speech flow shall be segmented into

Table 1. Common problem triggers in simultaneous interpreting and mental alleviation through multimodal processing

Common Problem Triggers	Insufficient Knowledge or Result-In Lag	Alleviation Through Multimodal Processing With Speech Draft	Alleviation Through Multimodal Processing With Real-Time Captions
A high density of the information content	Lag	Yes (With prepared in advance in expanding working memory) No (Without prepared in advance)	No
A high rate of delivery	Lag	No (Additional effort required in fast navigation)	Yes (Provide piecemeal information in keeping track)
Noise	Lag	Barely (Additional effort required in fast navigation)	Yes (Multimodal integration)
Unfamiliar accents	Lag	Yes (With Multimodal integration)	Yes (Multimodal integration)
Syntactical difference between source language & target language	Insufficient knowledge &Lag	Yes (With prepared in advance for anticipation)	Slight (Expanding episodic buffer)
Culture-specific difficulties	Insufficient knowledge	Yes (With prepared in advance in expanding working/long-term memory)	No

coherent chunks, then paired with acoustic models stored in the native English listeners' long-term memory. However, syllables with Chinese accent diverged from the standard model, requesting the listeners' cognitive effort in disambiguation before decisions were made. In this case, it led to longer processing time before the English listeners "perfectly" comprehend the semantics.

However, given "simultaneous interpreting" is a time-critical cross-modality processing activity, where a simultaneous interpreter is compared as an air traffic controller coordinating fully-occupied aircraft flow (Zeier, 1997), a slight delay may lead to mental congestion. If an interpreter's central capacity resource is compared as the central processing unit of a computer (Kahneman, 1973; Seeber, 2011), then one may draw an analogy between the interpreter's episodic buffer as the random access memory, and a problem trigger as a bug leading to excessive computing power. When the simultaneous interpreting program encounters a problem trigger, it overruns the memory's capacity for a short time. In turn, such a process would delay the succeeding process for a short while. If the problem trigger has not been solved throughout the interpreting assignment, corresponding buffer overflow would occur, leading to a point when the interpreting programme ceases to respond to inputs.

SIMULTANEOUS INTERPRETING AND AUDIO-VISUAL INTEGRATION

At any single moment, one is being bombarded by multiple-sensory information. In perceiving a three-dimensional object, information from various cues was collected via different organs, like eyes, ears, nose, mouth and hand. These data, from two or more sensory modalities, were then merged in a systematized manner to a unified and robust percept. This process is referred to as multisensory integration (Ohshiro et al., 2011). A large number of behavioural studies demonstrated faster and more accurate judgements when exposed under multisensory integration. Among these modalities, perception from concurrent audio and visual perception unified through the human brain is defined

as audio-visual integration. Through a simple semantic consistency experiment, Giard and Peronnet (1999) found shortened multiple sensory coordination time under concurrent audio-visual stimuli through an event-related potentials study. In an fMRI study designed by Barutchu et al. (2018), participants with posterior cortical damage were exposed under both unimodal visual/auditory stimuli and congruent audio-visual stimuli, where it has demonstrated a stronger response in temporal lobe under bi-modal congruent input.

To explore the mechanism behind audio-visual integration in simultaneous interpreting, Massaro (1989) presented a three-stage illustration of the fuzzy logical model of perception. In this model, information from both audio and visual sources goes through the initial stage of perceptual recognition, it switched into psychological values like words and sentences, and support for decisions is the final stage for individuals in making response alternative. The above-mentioned processes utilize both long-term memories from storage and short-term memory for processing. Stretching back to the Working Memory Model proposed by Baddeley (1992, 2000), the episodic buffer serves as multi-dimensional storage for both visual-spatial sketchpad and phonological loop, providing coherent episodes for the central executive in information integration. Thus the interpreter is able to retrieve the resources from long-term memory, then execute information processing (do interpreting) through the central executive to generate rendition as complete information units (consecutive) or coherent information chunks (simultaneous). Therefore, some interpreting strategies can be well explained through the working memory model, including but not limited to attention switching, inhibition of distracters, resource allocation can be well-explained (Cowan, 2000; Liu et al., 2004; Timarova, 2007).

It seems redundant that the central executive collects homologous data from dual channels. For example, when the interpreter is listening to an official read-out speech, he/she may constantly refer to printed speech drafts. Such dual-channel information may be collected concurrently through the auditory modality together with visual modality and analysed through seemingly replicated manner through the central executive. In explaining this phenomenon, we can refer to biological indications from psychological research outcomes. Through years of elementary education, an adult is capable of pairing character/letter's visual image with its phonetic spelling. An fMRI study by Van Atteveldt et al. (2004) discovered higher temporal cortical response when participants exposed under congruent bimodal stimuli, this study also indicated activeness of Heschl's sulcus and planum temporale towards congruent bimodal stimuli, while these encephalic regions response much lower activeness towards incongruent bimodal stimuli. Raij et al. (2000) found audio-visual convergence at the time lag of 225ms in the superior temporal sulcus. It is worth noticing that redundancy signal effects were generated only through stimulus onset asynchronies (SOA) less than 250ms, facilitating multisensory integration, while this facilitating effect deteriorates under long SOA (Van der Stoep et al., 2015). In summary, the physiological structure of a human brain enables motivating effect on audio-visual integration, provided there is both temporal and spatial congruence between the two channels. If it fails to meet either criterion, then this facilitation effect deteriorates or even could cause negative effects towards the interpreting activity.

RESEARCH GAP IN EXISTING MULTIMODAL PROCESSING MODELS

At the current stage, there are two major theoretical accounts for simultaneous interpreting with complex visual modalities (text and the interpreter's vision): the Effort Model (Gile, 1999, 2009) and the Cognitive Load Model (Seeber, 2011, 2017). Both models discussed the interpreter's mental load in this scenario. In the Effort Model, simultaneous interpreting is divided into a sum of several subtasks: listening, memory, production and coordination, that is SI=L+M+P+C. Simultaneous interpreting with text imposes an additional reading task towards the interpreter, therefore, SI with text=R+L+M+P+C. Since the Effort Model was created for pedagogical purposes, its facilitating factor and negative factor has been elaborated in plain language. The visual modality "reduces memory problems and the effect of acoustic

difficulities" alongside with the "probability of failures due to insufficient processing capacity in the Listening and Analysis Effort", while additional cognitive load arises in following the vocal speech together with the printed text (Gile, 2009, p. 182). This explanation has make the Effort Model somewhat puzzling.

The Conflict Matrix is proposed as a key element in the Cognitive Load Model. Calculating through an aggregate of demand vectors and conflict coefficients for various overlapping modalities, the total interference score for simultaneous interpreting with visual/verbal input (text and vision) is 14.8, higher than 11.6 for simultaneous interpreting with visual input (vision). Thus, it presumes simultaneous interpreting with text more capacity-consuming with "pure" simultaneous interpreting, which goes against the Effort Model.

However, both of the models underestimated the complexity of simultaneous interpreting scenarios in reality. Throughout a simultaneous interpreting assignment, the speaker's non-native accent, speech rate, the difficulty of the material to-be-interpreted and the length for interpreter's preparation will pose volatility towards the interpreter's general mental load. For example, in the face of an entire read-out speech, a well-prepared conference interpreter may withstand much effortless cognitive interference than do pure simultaneous interpreting. In the event when texts were given to the interpreter on the last minute, while the speaker is jumping back and forth among the gist, then most interpreting instructors would suggest taking the eyes off the paper since it obviously costs more cognitive load. While the above-mentioned strains were neglected throughout the modelling process, the build of "demanding vectors" in the models was thus quite rigid, leading the Conflict matrix in simultaneous interpreting with visual modalities presented in an undifferentiated form (Seeber, 2017, p. 472). In this context, experiments implemented under the reference of either the Effort Model or the Cognitive Load Model may lead to deviation.

SPEECH RECOGNITION IN SIMULTANEOUS INTERPRETING: A NEW MODALITY

As we outlined in the previous section, a prolonged time lag may be caused due to problem triggers that occur throughout simultaneous interpreting. Given the audio-visual integration may be activated through both temporal and spatial congruence, while an entire document of speech draft does not satisfy its temporal requirement, a new type of solution shall be envisaged in providing accurate yet chunked visual support.

Real-time captions appearing simultaneously with the speaker's pace provides the perfect solution in the visual modality. Even in a decade ago when it is unable to reach technical maturity, researchers still regard speech recognition with "considerable potential for changing the way interpreting is practised" (Pöchhacker, 2004, p. 118). When applied in multimodal processing, monolingual or bilingual captions both supports the enhancement of listeners' comprehensibility, for example, noticeable rate of comprehension improved in sign language interpreting listeners (Debevc et al., 2015), and facilitating effects detected in English learning process (S. Liao et al., 2020).

The reason real-time transcribed caption has not been given due attention could be ascribed to its outdated engine, which may produce subtitles with low accuracy yet longer latency. Before the introduction of the Attention Model (Vaswani et al., 2017), connectionist temporal classification (CTC) model was weak in resistance to noise, accent and terminology-dense speech flow (Chan et al., 2016). Based on the artificial neural network, Listen Attend and Spell (LAS) model is more robust in comprehending terminology-dense and strong-accented speeches. Possible errors that would occur through LAS model recognition are mostly conversational terms rather than technical terms, making it easier for human viewers to distinguish (Chiu et al., 2018). Thus, it created a friendly environment being deployed for scientific research. To date, a limited number of theoretical foundations has been published in exploring real-time captions in simultaneous interpreting (Fantinuoli, 2017a, 2017b), and two computer-aided interpreting tools with automatic speech recognition have been developed (Fantinuoli, 2016; Li, 2018).

To summarize, according to the theoretical framework and empirical studies of audio-visual-integrated facilitation in multimodal processing (Baddeley, 1992, 2000; Chmiel et al., 2020; Giard & Peronnet, 1999; Jesse et al., 2000), access to visual input may facilitate listening comprehension, based on its enhancement, mental burdens in simultaneous interpreting may also be alleviated. Although the existing multimodal processing models in simultaneous interpreting paint a negative picture (Gile, 2009; Seeber, 2017), the current conditions are adequate in implementing empirical enquiries towards professional interpreters' reaction under multimodal input with real-time captions. This study will thus extend this line of research to explore how professional interpreters work with the visual prompts with initiatives of referring to the real-time captions or discard them as non-sensical disruption.

THE STUDY

The purpose of this study is to explore simultaneous interpreters' initiative in attention-sharing if exposed under multimodal input from both auditory modalities and twofold visual modalities. Five conference interpreters were asked to simultaneously interpret a videotaped speech from English to Chinese, where Chinese is their native language (A) and English is the working language (B). Captions simultaneously displayed with the speaker's pace is displayed on the lower part of the screen. Those captions were auto-transcribed through the YouTube video transcription system, which is built on the attention-based neural network (H. Liao et al., 2013; Soltau et al., 2017). At regards to data collection, we deployed a Tobii 4C eye-tracker in monitoring eye movements of the participants, through what their initiative in acquiring visual input could be detected.

In terms of setting a problem trigger in this experiment, as outlined in the previous sections, if a simultaneous interpreting assignment was implemented smooth enough without tight-roping the interpreter's mental load, then hardly the audio-visual integration may provide positive effect for the interpreter in multimodal processing. As shown in the study of Jesse et al. (2000), if the presentation of the auditory signal contains no "noise" in a broader sense, then visual text as an additional source will contribute bare improvement towards interpreting quality.

Participants

The participants were five conference interpreters (four males, one female), with an average age of 32.5 (SD=4.03) and average working experience of 9.4 years (SD=4.66). Three among them are university lecturers, while another two are freelance interpreters. They speak Chinese Putonghua (Mandarin) as their native language, English as their working language. All of the participants have acquired level two Certificate of China Accreditation Test for Translators and Interpreters (CATTI), three of them acquired level one Certificate. CATTI is the most authoritative interpreting proficiency qualification accreditation test, level two certificate holders are assumed with the capability of interpreting political speeches (Chen, 2009; Zhao & Gu, 2016). No one among them had been to East Africa or interpreted speeches from an East African speaker. All the participants have signed informed consent forms prior to the experiment.

Materials

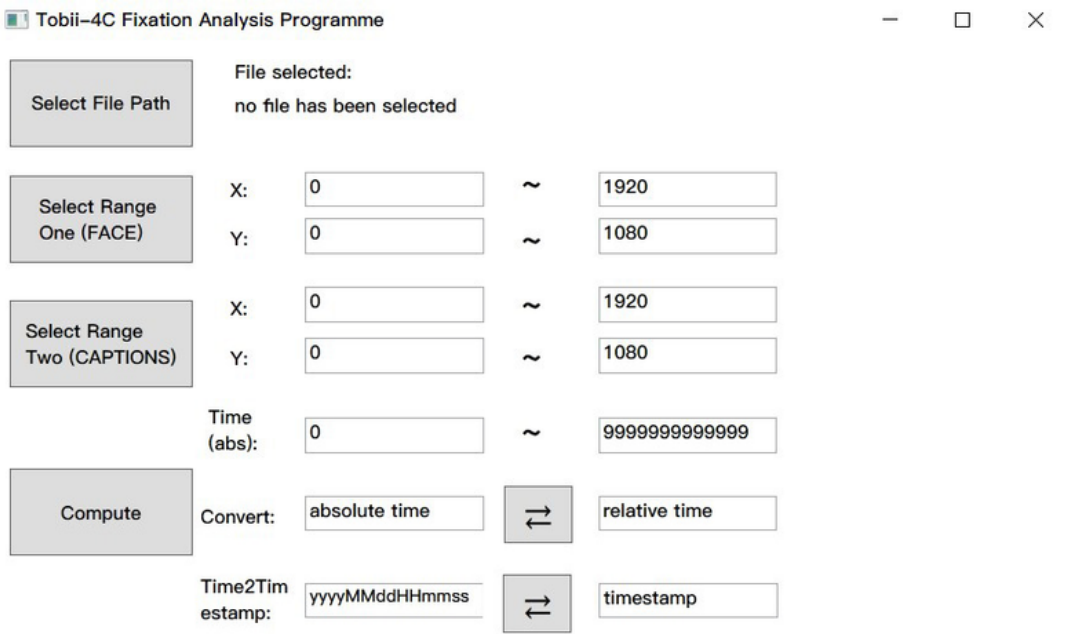
The speech material for this multimodal simultaneous interpreting experiment was based on a videotaped speech *Overcoming Kenya's Political Crisis and Advancing Democracy, Rule of Law and Stability* delivered at the Centre for Strategic and International Studies Forum, focusing on Kenya's democratic peace and stability, is assumed in coping with the participants' professional experience without further cultural or terminological preparation. The speaker Raila Odinga is the former Prime Minister of the Republic of Kenya, who received philological education in University of Leipzig, Germany. Raila has been keeping a close tie with European and Northern American politicians, which shaped his standard and rigorous English syntax. Raila speaks with a heavy East-African accent.

The experimental speech employed in this study starts from the beginning to around fifteen minutes, where Raila speaks at a pace of 99 words per minute. Manually transcribed speech remarks could be found on CSIS official website (CSIS, 2017). The speech has been computer-transcribed through YouTube speech recognition system. Fine adjustments have been made in both the written remarks and transcribed captions, for example, senseless fillers removed from both texts, and voiceover removed from the written remark. After these early-stage preparations, we used the TF-IDF algorithm, a numerical statistic frequently adopted in testing text similarities. The similarity between the two texts is 96.4%.

Eye-Tracker Preparation

Due to financial constraints, the eye-tracking data were recorded by the Tobii 4C eye-tracker. Tobii 4C has a satisfying frequency (90Hz) and acceptable accuracy. However, the lack of analysis software suite made it effortful in collecting and analysing data. We covered the default gaze trace, then managed to enable trace collection without drawing the participants' attention through Tobii Ghost. The eye movement data, the videotaped speech, participants' rendition has been recorded through three separate layers, then integrated through the screen recording software OBS. Through the Tobii-4C application programming interface, gaze data and fixation data have been collected for further analysis.

Figure 1. We wrote a data-collection programme in collecting eye-fixation data



Procedure

The participants were tested respectively, they were directed to sit in front of an Alienware laptop, the eye-tracker was connected through the USB port and fixed through the magnet. They were instructed to perform simultaneous interpreting. They were told that during the interpreting experiment, real-time captions will appear at the bottom of the screen, and they will have the initiative in referring to captions or not, even they can choose to close their eyes. All of the participants chose to use their own earphones.

Throughout each session, a single word appeared the moment after Raila spoke out each word. Therefore, one may only refer to the word from visual input immediately after he/she heard it, which differs simultaneous interpreting with real-time captions from simultaneous interpreting with text. A two-minute warm-up starts before the session in helping participants get acquainted with the experiment set, adjust the earphone volume and get familiar with the speaker's accent. The session started with a nine-point calibration procedure. Then the experimenter started OBS recording and data-collection programme simultaneous with the videotaped speech. Each experimental session lasted around twenty minutes.

After each participant accomplished the experimental session, a set of three raw data files were generated: gaze datasheet, the fixation datasheet and screen-recorded video with gaze trace and the participant's rendition. As the configuration set by Tobii 4C, one gaze point lasts 24ms to 40ms, which could be involuntarily eye movement. One fixation point consists of twenty continuous gaze points within the certain range, lasting 480ms to 800ms, therefore it is regarded as the meaningful eye gaze. Each participant generated around 65,000 gaze points and 3,300 fixation points, which indicated twelve-minute visual input within the fifteen-minute speech.

Data Analysis

For a qualitative analysis of the participants' initiative in referring to what specific content he/she glanced throughout the session, the experimenter invited a master's student from Master of Arts in Translation and Interpreting, Hong Kong Polytechnic University to transcribe each word one participant glanced at from the real-time captions. There are 1,515 words in this session. In the analysis of a word is "frequent" or not, the experimenter employed Corpus of Contemporary American English (COCA) as the yardstick, which provides a frequency list of the top 60,000 words. Words are divided into four categories: words with frequencies above 10,000 are referred to as "higher frequency", those beneath 10,000 as "lower frequencies", proper nouns which are not included in COCA as proper nouns, like "CSIS", and numbers (Table 2).

Table 2. Visually presented words referred to in this study, which were divided into four categories

Type	Proper Nouns	Numbers	Words With Frequencies Above 10,000	Words With Frequencies Beneath 10,000
Speech draft	74	9	1,234	198
Participant 1	56	8	518	142
Participant 2	48	7	340	121
Participant 3	34	6	351	123
Participant 4	37	6	312	124
Participant 5	39	6	278	93
Average	43	7	360	121

It is found that visually presented captions have drawn close attention from all five participants. We divided the words participants noticed into four categories: proper nouns, numbers, words with frequencies above 10,000 and words with frequencies beneath 10,000. Within these words, numbers have drawn the most attention from the participants (average: 7/9, 77.8%), followed by words with frequencies lower than 10,000 (average: 121/198, 61.1%). The participants also consciously put enough attention on proper nouns (average: 43/74, 58.1%), while words with higher frequencies have drawn comparatively least attention (average: 360/1234, 29.2%) (Table 3).

Table 3. Standard deviation analysis of visually presented words referred by the participants

Type	Proper Nouns	Numbers	Words With Frequencies Above 10,000	Words With Frequencies Beneath 10,000
Participant 1	75.7%	88.8%	42.0%	71.7%
Participant 2	64.8%	77.8%	27.6%	61.1%
Participant 3	45.9%	66.7%	28.5%	62.12%
Participant 4	50%	66.7%	25.28%	62.62%
Participant 5	52.7%	66.7%	22.52%	47.0%
Standard Deviation	0.12	0.10	0.08	0.09

The Standard Deviation suggested common features of all the participants' active behaviour in referring to visually-presented captions, we also found an interesting condition by the general sum of words participants referred to, showing that an approximation of one-third of the word appeared in captions were referred to, it could be explained as the maximation of mental resources, in which visual resources integrated without occupying processing capacities allocated to listening comprehension. However, this study is comparatively small in scale, yet similar experiments on speech rate, linguistic structure difference as variables should be implemented in further confirmation in the next batch.

CONCLUSION

Simultaneous interpreting is a cross-linguistic and cross-modality mental activity. When performing a simultaneous interpreting task with visual input, practitioners have to coordinate multimodal subtasks in maximizing information intake per time unit. This study firstly presents a theoretical framework of problem trigger and processing latency it brought to in the interpreting activity. Then we analysed the audio-visual integration effect from the perspective of neuroscience, with the conclusion of empirical findings matching this principle. With the technology advancement in speech recognition, it brings with a new processing modality if utilized in simultaneous interpreting as the visual assistance.

We further implemented an experiment in exploring professional interpreters' initiative when exposed under multichannel input from both the videotaped speech recording and the real-time captions generated by the speech recognition system. This study revealed professional interpreters' spontaneous motivation in accepting input from multiple resources. It also indicated interpreters' instinct in referring to a larger proportion of numbers, words with lower frequencies and proper names. In operating the experiment of simultaneous interpreting with real-time captions, we hope to broach this idea to some further research of multimodal processing in simultaneous interpreting.

ACKNOWLEDGMENT

This research was supported by IGI Global. We would like to thank the two anonymous reviewers for their insightful suggestions and comments.

FUNDING BODY INFORMATION

Open Access Funding Provided by Shandong University (Weihai), China.

REFERENCES

Albl-mikasa, M. (2010). Global English and English as a Lingua Franca (ELF): Implications for the interpreting profession. *Trans-Kom*, *3*(2), 126–148.

Baddeley, A. (1992). Working memory. *Science*, *255*(5044), 556–559. doi:10.1126/science.1736359 PMID:1736359

Baddeley, A. (2000). The episodic buffer: A new component of working memory? Trends in Cognitive Sciences, 4(11), 417–423. doi:10.1016/S1364-6613(00)01538-2

Barutchu, A., Spence, C., & Humphreys, G. W. (2018). Multisensory enhancement elicited by unconscious visual stimuli. *Experimental Brain Research*, *236*(2), 409–417. doi:10.1007/s00221-017-5140-z PMID:29197998

Chan, W., Jaitly, N., Le, Q., & Vinyals, O. (2016). Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2016*, 4960–4964. doi:10.1109/ICASSP.2016.7472621

Chen, J. (2009). Authenticity in accreditation tests for interpreters in China. *The Interpreter and Translator Trainer*, *3*(2), 257–273. doi:10.1080/1750399X.2009.10798791

Cheung, A. K. F. (2013). Non-native accents and simultaneous interpreting quality perceptions. *Interpreting*, *15*(1), 25–47. Advance online publication. doi:10.1075/intp.15.1.02che

Chiu, C. C., Tripathi, A., Chou, K., Co, C., Jaitly, N., Jaunzeikare, D., Kannan, A., Nguyen, P., Sak, H., Sankar, A., Tansuwan, J., Wan, N., Wu, Y., & Zhang, X. (2018). Speech recognition for medical conversations. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH,* 2972–2976. doi:10.21437/Interspeech.2018-40

Chmiel, A., Janikowski, P., & Lijewska, A. (2020). Multimodal processing in simultaneous interpreting with text. *Target. International Journal of Translation Studies*. Advance online publication. doi:10.1075/target.18157.chm

Cowan, N. (2000). Processing limits of selective attention and working memory. *Interpreting*, *5*(2), 117–146. doi:10.1075/intp.5.2.05cow

CSIS. (2017). *Overcoming Kenya's political crisis and advancing democracy, rule of law, and stability*. https://www.csis.org/events/raila-odinga-kenyan-elections

Davies, M. (2009). The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics*, *14*(2), 159–190. doi:10.1075/ijcl.14.2.02dav

Debevc, M., Miloševic, D., & Kožuh, I. (2015). A Comparison of comprehension processes in sign language interpreter videos with or without captions. *PLoS One*, *10*(5), 1–15. doi:10.1371/journal.pone.0127577 PMID:26010899

Fantinuoli, C. (2016). InterpretBank. Redefining computer-assisted interpreting tools. *Proceedings of the 38th Conference Translating and the Computer*.

Fantinuoli, C. (2017a). Computer-assisted interpreting: Challenges and future perspectives. In G. C. Pastor & I. Durán-Muñoz (Eds.), *Trends in E-Tools and Resources for Translators and Interpreters* (pp. 153–174). Brill/Rodopi., doi:10.1163/9789004351790_009

Fantinuoli, C. (2017b). Speech recognition in the interpreter workstation. *Proceedings of the Translating and the Computer*, *39*, 25–34. https://www.ibm.com/blogs/watson/2017/03/reaching-new-records-in-speech-recognition

García, A. M. (2019). The Neurocognition of Translation and Interpreting. The Benjamins Translation Library.

Giard, M. H., & Peronnet, F. (1999). Auditory-visual integration during multimodal object recognition in humans: A behavioral and electrophysiological study. *Journal of Cognitive Neuroscience*, *11*(5), 473–490. doi:10.1162/089892999563544 PMID:10511637

Gile, D. (1999). Testing the Effort Models ' tightrope hypothesis in simultaneous interpreting - A contribution. *Journal of Linguistics*, *23*(23), 153–172.

Gile, D. (2009). *Basic Concepts and Models for Interpreter and Translator Training: Revised edition*. https://www.amazon.com/Concepts-Models-Interpreter-Translator-Training/dp/9027224331%3FSubscriptionId%3D0JYN1NVW651KCA56C102%26tag%3Dtechkie-20%26linkCode%3Dxm2%26camp%3D2025%26creative%3D165953%26creativeASIN%3D9027224331

Han, C., & Riazi, M. (2017). Investigating the effects of speech rate and accent on simultaneous interpretation: A mixed-methods approach. *Across Languages and Cultures*, *18*(2), 237–259. Advance online publication. doi:10.1556/084.2017.18.2.4

Hsu, C. K., Hwang, G. J., & Chang, C. K. (2014). An automatic caption filtering and partial hiding approach to improving the english listening comprehension of EFL students. *Journal of Educational Technology & Society*, *17*(2), 270–283.

Jesse, A., Vrignaud, N., Cohen, M. M., & Massaro, D. W. (2000). The processing of information from multiple sources in simultaneous interpreting. *Interpreting*, *5*(2), 95–115. doi:10.1075/intp.5.2.04jes

Kahneman, D. (1973). *Attention and Effort*. Prentice Hall.

Lee, C. H., & Kalyuga, S. (2011). Effectiveness of on-screen pinyin in learning Chinese: An expertise reversal for multimedia redundancy effect. *Computers in Human Behavior*, *27*(1), 11–15. doi:10.1016/j.chb.2010.05.005

Li, T. (2018). Analysis of establishing a simultaneous interpreting system based on the interpreter's working model. *East Journal of Translation, 6*(5), 34-39+87.

Liao, H., McDermott, E., & Senior, A. (2013). Large scale deep neural network acoustic modeling with semi-supervised training data for YouTube video transcription. *2013 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2013 - Proceedings*, 368–373. doi:10.1109/ASRU.2013.6707758

Liao, S., Kruger, J., & Doherty, S. (2020). The impact of monolingual and bilingual subtitles on visual attention, cognitive load, and comprehension. *The Journal of Specialised Translation*, *33*, 70–98.

Lin, I. H. I., Chang, F. L. A., & Kuo, F. L. (2013). *The impact of non-native accented English on rendition accuracy in simultaneous interpreting*. Translation and Interpreting. doi:10.12807/ti.105202.2013.a03

Lin, X. (2013). *An empirical study on computer aided interpretation from English to Chinese* (Master's Thesis). Shandong Normal University.

Liu, M. (2019). Knowing what and knowing how: Teaching student interpreters research on interpreting. In *The Evolving Curriculum in Interpreter and Translator Education: Stakeholder perspectives and voices* (pp. 301–318). John Benjamins Publishing Company. doi:10.1075/ata.xix.14liu

Liu, M., Schallert, D. L., & Carroll, P. J. (2004). Working memory and expertise. *Interpreting*, *6*(1), 19–42. doi:10.1075/intp.6.1.04liu

Mankauskienė, D. (2016). Problem Trigger Classification and its Applications for Empirical Research. *Procedia: Social and Behavioral Sciences*, *231*(May), 143–148. doi:10.1016/j.sbspro.2016.09.083

Massaro, D. W. (1989). Testing between the TRACE model and the fuzzy logical model of speech perception. *Cognitive Psychology*, *21*(3), 398–421. doi:10.1016/0010-0285(89)90014-5 PMID:2758786

Munro, M. J., & Derwing, T. M. (1995a). Foreign Accent, Comprehensibility, and Intelligibility in the Speech of Second Language Learners. *Language Learning*, *45*(1), 73–97. doi:10.1111/j.1467-1770.1995.tb00963.x

Munro, M. J., & Derwing, T. M. (1995b). Processing Time, Accent, and Comprehensibility in the Perception of Native and Foreign-Accented Speech. *Language and Speech*, *38*(3), 289–306. doi:10.1177/002383099503800305 PMID:8816082

Ohshiro, T., Angelaki, D. E., & Deangelis, G. C. (2011). A normalization model of multisensory integration. *Nature Neuroscience*, *14*(6), 775–782. doi:10.1038/nn.2815 PMID:21552274

Pöchhacker, F. (2004). Introducing interpreting studies. In *Introducing Interpreting Studies*. Routledge. doi:10.4324/9780203504802

Raij, T., Uutela, K., & Hari, R. (2000). Audiovisual integration of letters in the human brain. *Neuron*, *28*(2), 617–625. doi:10.1016/S0896-6273(00)00138-0 PMID:11144369

Rennert, S. (2008). Visual input in simultaneous interpreting. *Meta*, *53*(1), 204–217. doi:10.7202/017983ar

Seeber, K. G. (2011). Cognitive load in simultaneous interpreting: Existing theories — new models. *Interpreting*, *13*(2), 176–204. doi:10.1075/intp.13.2.02see

Seeber, K. G. (2017). Multimodal Processing in Simultaneous Interpreting. In J. W. Schwieter & A. Ferreira (Eds.), *The Handbook of Translation and Cognition* (pp. 461–475). John Wiley & Sons, Inc., doi:10.1002/9781119241485.ch25

Shen, D. (2014). *A simulation study of speech recognition assisted simultaneous interpretation.* Xiamen University.

Soltau, H., Liao, H., & Sak, H. (2017). Neural speech recognizer: Acoustic-To-word LSTM model for largevocabulary speech recognition. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH,* 3707–3711. doi:10.21437/Interspeech.2017-1566

Stachowiak, K. (2014). Mind's Not lazy : On Multitasking in Interpreters and Translators. *Konińskie Studia Językowe*, *2*(3), 293–313.

Timarova, S. (2007). Working memory and simultaneous interpreting. *Translation and Its Others. Selected Papers of the CETRA Research Seminar in Translation Studies 2007, 2007*, 1–28.

Van Atteveldt, N., Formisano, E., Goebel, R., & Blomert, L. (2004). Integration of letters and speech sounds in the human brain. *Neuron*, *43*(2), 271–282. doi:10.1016/j.neuron.2004.06.025 PMID:15260962

Van der Stoep, N., Van der Stigchel, S., & Nijboer, T. C. W. (2015). Exogenous spatial attention decreases audiovisual integration. *Attention, Perception & Psychophysics*, *77*(2), 464–482. doi:10.3758/s13414-014-0785-1 PMID:25341648

Vaswani, A., Brain, G., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention Is All You Need. *Advances in Neural Information Processing Systems, Nips*, 5998–6008. http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf

Zeier, H. (1997). Psychophysiological stress research. *Interpreting*, *2*(1–2), 231–249. doi:10.1075/intp.2.1-2.09zei

Zhao, H., & Gu, X. (2016). China Accreditation Test for Translators and Interpreters (CATTI): Test review based on the language pairing of English and Chinese1. *Language Testing*, *33*(3), 439–446. doi:10.1177/0265532216643630

Tianyun Li is a doctoral student of School of Translation, Shandong University (Weihai). He received his MA (Distinction) in translation and bilingual communication from Hong Kong Baptist Unveristy. He has been an active conference interpreter alongside with his academic life. His main research interest concern the psycholinguistic aspects of simultaneous interpreting.

Bicheng Fan is a software engineer of AI chatbot project, Sogou (Hangzhou) Research Institute. His research interest lies in natural language processing, chatbot and automatic speech recognition. He has development experience in BERT, word2vec, FastText, ELMO, and TensorFlow framework.