Perception Changes With and Without a Video Channel: A Study from a Speech-to-Speech, Machine Translation Mediated Map Task

Hayakawa Akira, Carl Vogel, Nick Campbell
School of Computer Science and Statistics
Trinity College Dublin, Dublin, Ireland
Email: {campbeak, vogel, nick}@tcd.ie

Saturnino Luz
Usher Institute of Population Health Sciences & Informatics
University of Edinburgh, Edinburgh, UK
Email: S.Luz@ed.ac.uk

***Abstract*—This study looks into the effect of the video channel, that provides realtime visual information of the subject's interlocutor in computer mediated multi-lingual map task dialogues. The addition of a video channel in long distance audio communication has been commercially available since 1964, pioneered by AT&T's Picturephone. However the complexity of adding an image channel to a task oriented dialogue has not penetrated the user audience enough to change the user expectation from a like-to-like alternative of Face-to-Face communication, to a new different communication style. This study reports the increase in visual cognitive state occurrences when communicating with a video channel and the different perception that this setting provided to the subjects of the ILMT-s2s corpus.**

I. Introduction

The promise that the modality of a visual channel would improve communication is a delicate matter. Since the introduction of the Picturephone by the American Telephone and Telegraph Company (AT&T) in 1964, that transmitted the user's image and voice simultaneously over telephone lines [1], industry has continued to add an image channel to the conversation modality, even though the Picturephone can be categorised as a complete commercial failure. Mobile phones that were capable of sending video images with the audio channel were first introduced in 1999, and today, it is more difficult to find a mobile phone that doesn't have this function. In a reference to the publication in Japanese of Yoshida and Kakuta [2, p. 149], a study of audio only telephone communication presented that visual information can be both a positive and negative attribute to communication.

> A study by Yoshida et al. showed that college students in and around Kyoto–Osaka (effective response 549) cited the following reasons for the phone's popularity: speed, available anytime, available anywhere, no visual information about the other end, and easier way to say what we want than in face-to-face conversation. They also cited the following disadvantages: no visual information about the other end and difficulty in conveying subtle emotion. How interesting that no visual information about the other end becomes both an advantage *and* disadvantage !

The authors [2] report the subjects considering telephone communication as a different alternative to Face-to-Face communication with its own communication style.

Video-conferencing devices are widely used for communicate over distance. From dedicated devices from Polycom Inc. to Skype of Microsoft Inc., image/video is now an easily accessible option for communication. Skype has recently added the Speech-to-Speech Machine-Translation (S2S-MT) function Skype Translator [3], but this will now create a conversation style that few users have adequate exposure and little research has been conducted to date [4]. It is inevitable that S2S-MT will now be rapidly added to this modality.

Conversation styles change with ease. High latency video conference systems produce fewer, but longer utterances. Lower latency systems, shorter but more utterances, a style closer to Face-to-Face communication [5]. Also the quality, low latency, of the audio is said to determine task effectiveness, implying that video has marginal importance [6]. A study into video enhanced communication in medical meetings reported that video enhanced the diverse needs of communication, but only after 8 months of use [7].

In this study, we look at how the user of a S2S-MT system, is affected by a video channel (w/ Video & w/o Video).

II. Material

We investigate the cognitive affect in S2S-MT communication of task oriented conversation. For the S2S-MT communication data, the fifteen dialogues from the ILMT-s2s corpus [8] and its 7 point Likert scale user survey results were used.

A. Data from the ILMT-s2s Corpus

The ILMT-s2s corpus contains fifteen dialogues between fifteen English and fifteen Portuguese subjects speaking to each other as pairs in their native language via a S2S-MT system (ILMT-s2s System). The dialogues are elicited using the Map Task technique, with maps 01 and 07 (Figure 1) of the HCRC Map Task corpus [9]. These maps were selected due to the simplicity to navigate, from their low mean deviation score, and longer mean duration within the HCRC Map Task corpus.

One subject is assigned the role of Information Giver (IG), and the other, the role of Information Follower (IF). The IG provides instructions to their interlocutor, the IF, so the IF can draw the same route as indicated on the IG's map, from the instruction/information provided by the IG.

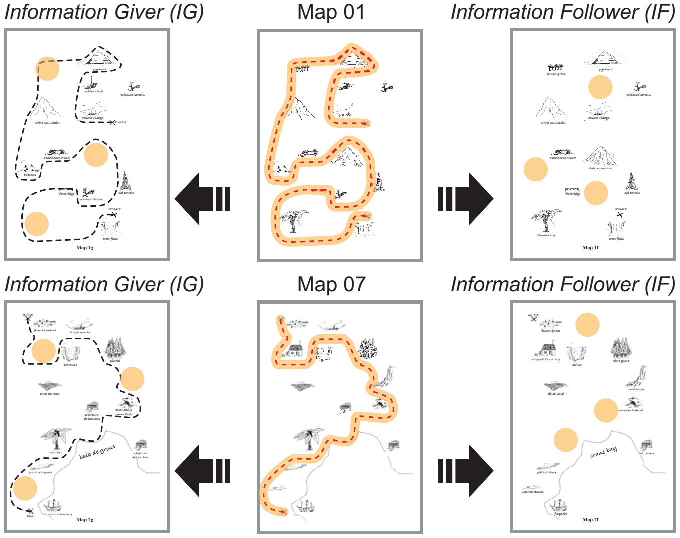


Fig. 1: Maps with differences highlighted — Left: Map used by IG, Centre: Map with all items, Right: Map used by IF

1) The ILMT-s2s System: Two subjects, seated in separate rooms, used the ILMT-s2s System (Figure 2) to communicate with each other. The ILMT-s2s System is a system that uses off-the-shelf components — Automatic Speech Recognition (ASR), Machine Translation (MT) and Text-to-Speech synthesis (TTS) — to perform S2S-MT. It is activated by a "Push-to-talk" button that the subject will click-and-hold for the duration of the utterance and release once the subject has finished. Neither subject can hear the other's voice, since the output of the ASR and MT is provided by a synthetic voice.

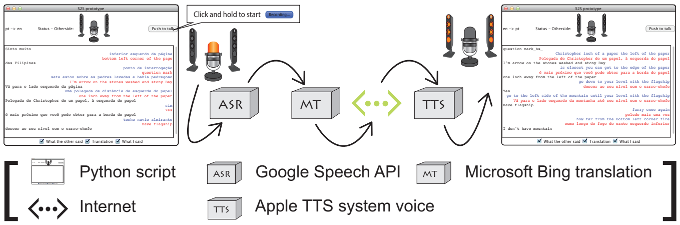


Fig. 2: ILMT-s2s System used to collect the data

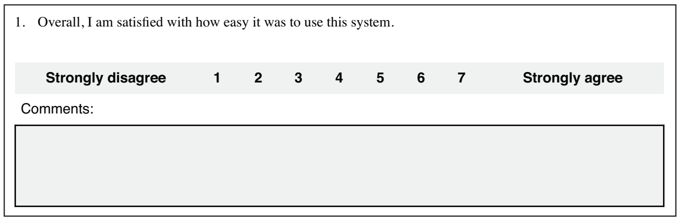
2) The Subjects and Recording Environment: The subjects (aged 18–45) were recruited from the Trinity College Dublin digital noticeboard and personal connections. Fifteen recordings of fifteen native English speakers (♀5, ♂10), and fifteen native Portuguese speakers (♀11, ♂4), were collected. One subject during each recording session was fitted with a biosignal recording device, while the other subject was not.[1]

3) The Fifteen 7 Point Likert Scale Statements: Thirteen of the fifteen statements (S.01 – S.12 & S.15 of Table I) were taken from "The Post-Study System Usability Questionnaire (PSSUQ)" [10, pp. 28–33] and two (S.13 & S.14 of Table I) were added specifically for characteristics of the ILMT-s2s System — the system's TTS voice and the system's output.

TABLE I: Survey statements of the ILMT-s2s corpus

7 point Likert scale statements
S.01: Overall, I am satisfied with how easy it was to use this system.
S.02: It was simple to use this system.
S.03: I could effectively complete the tasks using this system.
S.04: I was able to complete the task quickly using this system.
S.05: I was able to efficiently complete the task using this system.
S.06: I felt comfortable using this system.
S.07: It was easy to learn to use this system.
S.08: I believe I could become productive quickly using this system.
S.09: Whenever I made a mistake using the system, I could recover easily and quickly.
S.10: The interface of the system was pleasant.
S.11: I liked using the interface of this system.
S.12: This system has all the functions and capabilities I expected it to have.
S.13: I was satisfied with the voice of this system.
S.14: I was satisfied with the output of this system.
S.15: Overall, I am satisfied with this system.

The PSSUQ [10, p. 14], includes nineteen statement overall, with eight statements regarding the "System Usefulness", seven statements on "Information Quality", three statements on "Interface Quality", and finally one statement for the "Overall Satisfaction" of the system [10, p. 34]. For the fifteen statements used in the ILMT-s2s System user survey (Table I), statements S.01 to S.08 were taken from "System Usefulness", statement S.09 from "Information Quality",[2] statements S.10 to S.12 from "Interface Quality" and statement S.15 was from the "Overall Satisfaction". The additional two statements of S.13 and S.14 were added to supplement the S2S-MT aspect of the "Interface Quality" statements. Responses to the fifteen statements listed in Table I were given on a 7 point Likert scale (evenly spaced from 1 = Strongly disagree to 7 = Strongly agree), presented together with an open text field for possible further comments, as illustrated in Figure 3.



1. Overall, I am satisfied with how easy it was to use this system.

Strongly disagree 1 2 3 4 5 6 7 Strongly agree

Comments:

Fig. 3: 7 point Likert scale user survey actual layout example

4) Cognitive State [3] Annotations: The cognitive states "*Surprised*", "*Frustrated*", and "*Amused*" were annotated using the dedicated annotation tool ELAN [11] by two students that also transcribed the audio of the dialogues to text. The inter-coder agreement for the annotated cognitive state labels were calculated using the modified kappa feature of ELAN 4.9.0's "Inter-Annotator Reliability…" function on one of the dialogues and the results are well above 60%.[3]

[1] Data from the biosignal recordings were not used in this study.

[2] Of the seven "Information Quality" statements, only one was used since other statements were related to the system error feedback. Since the ILMT-s2s System did not have any help menu or detailed error messages, these statements were not added.

[3] Due to the limited size of the ILMT-s2s corpus, all cognitive state labels were verified after each dialogue was completed by the first author and ambiguous items were discussed with the annotators for verification.

TABLE II: Cognitive state label count comparison

	Utterances	All labels	After Utt.	After TTS	All else
All Cog.	–	1,706	827	601	278
Surprised	–	346	135	189	22
Frustrated	–	621	350	160	111
Amused	–	739	342	252	145
w/ Video	1,757	1,103	514	384	205
Diff.	(x1.21)	(x1.83)	(x1.64)	(x1.77)	(x2.81)
w/o Video	1,449	603	313	217	73

A cognitive state (*Surprised*, *Amused* or *Frustrated*) was assigned to locations where the annotator deemed the subject to be in one of these states, based on visual cues. While acknowledging that there are numerous cognitive states proposed by various annotation schemes [12], [13], [14], the initial annotation was limited to only the three states named. Since the data collection is task based, our understanding was that the subjects would be focused on providing and receiving clear instructions. However, given that each individual has their own understanding and preference for describing situations, clear communication is difficult even in Human-to-Human (H2H) situations [15]. By adding the *filters* (ASR, MT and TTS) of the ILMT-s2s System, further complication will arise. The assumption was that this would bring out the selected cognitive characteristic in interaction with the system.

TABLE III: Description of cognitive states annotation labels

Cognitive States [3]	Description
Surprised:	The subject is in a state of surprise.
Amused:	The subject is in a state of amusement.
Frustrated:	The subject is in a state of frustration.

B. Summary of the ILMT-s2s Corpus

As mentioned, the data from the ILMT-s2s corpus comprises fifteen dialogues with a total of thirty subjects. A previous study of the ILMT-s2s corpus has shown that the participants adapt their speech rate while speaking to the S2S-MT system at a relatively slower speed (Figure 4) [16], [17].

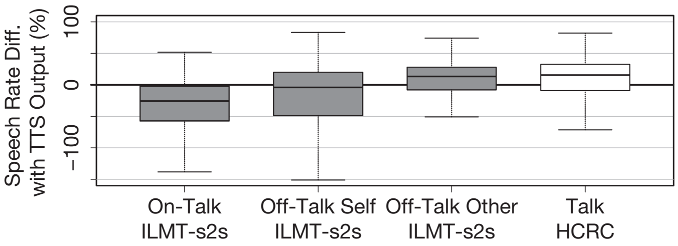


Fig. 4: Speech rates of HCRC Map Task, and ILMT-s2s corpus

Also as differentiated in Table IV, three types of utterances were distinguished within the corpus [18], [17]; *On-Talk* when the subject is using the S2S-MT system as a mediator to communicate with the interlocutor, *Off-Talk Self* when the subject is talking to him/herself, and *Off-Talk Other* when the subject is directly talking to a fellow human. In Figure 4, the speech rate[4] box plots of *Off-Talk Other* of the ILMT-s2s corpus and the dialogues of the subjects using maps 01 & 07 of the HCRC Map Task corpus show a similarity. Refer to Table IV for the median, means and *sd* values.

TABLE IV: Summary of ILMT-s2s, and HCRC Map Task corpus utterance count (all dialogues using maps 1 & 7) — for the HCRC corpus, with and without Eye-Contact

	Dialogue act count			Speech rate summary		
	All	w/ Vid.	w/o Vid.	Median	Mean	*SD*
HCRC	3,790	1,787	2,003	15.65	6.06	43.73
ILMT-s2s	3,809	2,230	1,579			
On-Talk	2,603	1,464	1,139	−25.54	−33.33	45.24
Off-Talk Self	859	450	409	−3.89	−27.94	85.09
Off-Talk Other	347	316	31	13.50	5.39	37.51

Since *Off-Talk Other* is defined as direct communication with a fellow human, the speech rate of *Off-Talk Other* was therefore compared with the speech rate values of the HCRC Map Task corpus. This comparison was made to clarify that there were no significant differences between the two H2H speech rates so as to indicate that the subjects of the ILMT-s2s corpus were not a fluke selection of slower speakers. As a result, no significant difference in the speech rate was observed (Mann-Whitney U test: $p = 0.2565$). Also the effect size was verified for good measure that resulted in a negligible estimate (Cliff's δ estimate: 0.051). This indicates that the non-mediated H2H communication in both corpora use similar speech rates and that the subjects of the ILMT-s2s corpus speak at a similar speech rate in direct H2H communication.

Furthermore, a comparison of dialogue acts used in the ILMT-s2s corpus and the sixteen dialogue that use the maps 01 & 07 in the HCRC Map Task corpus identified differences in the number of dialogue acts and the mean word count within the dialogue acts, fewer dialogue acts and more words per dialogue act being used in S2S-MT dialogues (Figure 5) [19]. The frequency that dialogue acts used also differ. The top three dialogue acts for the role of IF (*Acknowledge*, *Reply Y & Check*) that make a combined 73% in the HCRC Map Task corpus have only a share of 22% in the ILMT-s2s corpus.

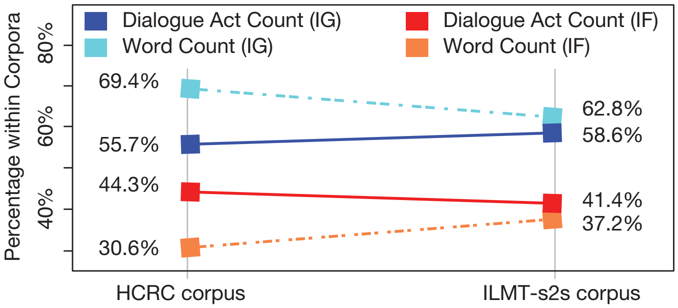


Fig. 5: Ratio comparison of dialogue act count and word count of roles in HCRC Map Task, and ILMT-s2s corpus

[4] The speech rate referred to here is the duration difference of the words uttered by the human subject compared to the duration of the same words uttered by the TTS system [16], [18], [17]. A negative rate indicates a duration that is longer (slower) than the TTS system voice speed setting of 180 wpm.

To obtain the same 73% in the ILMT-s2s corpus, five acts are required (*Explain*, *Query W*, *Reply Y*, *Acknowledge* & *Align*). A similar phenomenon also happens for the role of IG. Apart from the apparent top item of "*Instruct*" the next four items which make a combined 48% share are all moved lower than fifth position in the ILMT-s2s corpus indicating a change in the effort required to communicate [19]. These are style changes similar to that of high latency video conference systems [5].

III. Results

We present our analysis of the subject user survey scores ($n = 30$), and also the 1,706 annotated cognitive states.

A. Analysis of the 7 point Likert score

The overall median of all 7 point Likert scale results is 5.0 and looking at the individual statements, all fifteen statements have a median of 4.0 (neutral) or above. This indicates that overall, the ILMT-s2s System was perceived as a useful system with a good interface quality by the subjects since the five statements of S.02, S.07, S.10, S.11, and S.13 received a median of 6.0, and 5.0 for statements S.06 and S.12.

A different story starts to emerge once the results are differentiated by the subject groupings of Setting (w/ Video–w/o Video), Role (IG–IF), and Language spoken (Pt–En).[5] No overall median of the groupings drops below the neutral 4.0, but that cannot be said for individual items. A comparison of the median (and means) difference is indicated in Figure 6 with the X axis indicating the statement number and the Y axis indicating the difference within the groupings.

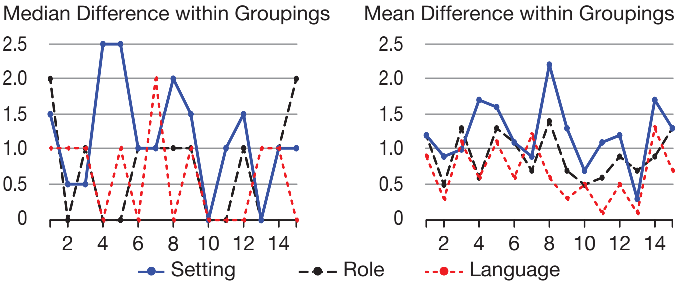


Fig. 6: 7 point Likert score result difference within groupings — Median and Mean ($n = 15$)

These differences evidently push some results below the median of 4.0, and show that the subjects who used the setting w/ Video, role of IF, and spoke English perceived the system more negatively. Furthermore, a very clear preference emerges with the combined grouping of "w/ Video–IF–Pt" providing the worst results and "w/o Video–IG–Pt" providing the best as indicated in Table V (for a visual perspective of these results, refer to the stacked bar chart in Figure 7 for the Portuguese subjects and Figure 8 for the English subjects).

B. Analysis of the annotated cognitive states

We begin with a previous study of the link between the ASR Word Error Rate (WER) and the subject's cognitive state, with results indicating that there were significant differences between the WER of each utterance *before* a cognitive state than for utterances *after* a cognitive state [20]. Different from the previous study [20], the cognitive states have been linked to the closest communication item — an utterance from the subject or additionally an output from the ILMT-s2s System. Even with this additional re-linkage, a significant difference emerged (ANOVA $F_{3,2348} = 8.577; p < 0.001$). As noticeable from Table VI, the significant difference was not identified in the w/o Video setting, but in the w/ Video setting. Within the w/ Video grouping, a significant difference was reported for both English (ANOVA $F_{3,670} = 10.61; p < 0.001$) and Portuguese (ANOVA $F_{3,640} = 11.52; p < 0.001$) subjects. Post-hoc comparisons (Tukey HSD test) revealed significant differences between *Amused* or *No Link* with *Frustrated*, in English subjects ($p < 0.01$) with an effect size of medium (Cohen's d estimate: 0.622) for *Amused–Frustrated* and large (Cohen's d estimate: 0.990) for *No Link–Frustrated*; for Portuguese subjects, *Amused* or *No Link* with *Frustrated* or *Surprised* ($p < 0.001$) with small (Cohen's d estimate: 0.355) to large (Cohen's d estimate: 1.126) effect sizes.

TABLE V: Ordered summary of 7 point Likert scale results

	Median	Mean	*SD*
w/ Video – IF – Portuguese	3.0	3.72	1.67
w/ Video – IF – English	4.0	3.65	1.93
w/ Video – IG – English	5.0	4.47	1.61
w/o Video – IF – English	5.0	5.07	1.60
w/o Video – IG – English	5.0	5.09	1.70
w/ Video – IG – Portuguese	5.5	5.17	1.36
w/o Video – IF – Portuguese	6.0	5.16	1.95
w/o Video – IG – Portuguese	7.0	6.25	1.07

Furthermore, what is clear from the a simple count is that there are more cognitive state labels from dialogues using the setting w/ Video. Taking into account the 1.2 times difference in the number of turns in the w/ Video setting, as indicated in Table II, there are still 1.3–2.3 times more cognitive states in the setting w/ Video. Most notably, the number of cognitive state labels that are not directly after a subject utterance (After TTS & All else) is more than double that of the setting w/o Video with a difference in the type of cognitive state linked to the languages spoken by the subject (Figure 7 & Figure 8).

IV. Discussion and Conclusion

From the observation of the user survey and the cognitive states reported in this paper, and adding the other results obtain from other papers published using the data of the ILMT-s2s corpus [16], [20], [21], [18], [19], [17], the conclusion is that simply adding video is not easing the task for the subject.

Previous analysis of the ILMT-s2s corpus reported that the subjects who use the S2S-MT system reduce their speech rate (Figure 4) when talking to the system [17]. This reduction is not a fluke example of slow speaking subjects, since the speech rate of the subjects talking directly to a fellow human is similar to that of the data of the HCRC Map Task subjects (Table IV). The main differentiator found in this study was that the speech rate difference between gender was affected by the system, but the setting of the system did not show a great difference (Mann-Whitney U test: $U = 240,930,\ p = 0.0019,$

[5] Grouping of gender and the map used were also verified, but the results from the two groupings were similar.

TABLE VI: Subject's cognitive states WER

Setting – Lang	Role	Cog. State	Mdn.	Mean	*SD*	Count
w/ Vid – Pt	IF	Surprised	0.625	0.645	0.39	19
w/ Vid – Pt	IG	Surprised	0.550	0.563	0.38	16
w/ Vid – Pt	IG	Frustrated	0.500	0.605	0.66	41
w/ Vid – Pt	IF	Frustrated	0.429	0.480	0.41	59
w/ Vid – Pt	IF	No Link	0.250	0.380	0.41	229
w/ Vid – Pt	IG	No Link	0.200	0.324	0.42	211
w/ Vid – Pt	IF	Amused	0.171	0.303	0.34	26
w/ Vid – Pt	IG	Amused	0.000	0.176	0.30	43
w/ Vid – En	IG	Frustrated	1.000	0.789	0.33	32
w/ Vid – En	IF	Frustrated	0.875	0.751	0.31	23
w/ Vid – En	IG	Surprised	0.679	0.671	0.32	9
w/ Vid – En	IG	Amused	0.571	0.538	0.42	72
w/ Vid – En	IF	Amused	0.523	0.523	0.38	50
w/ Vid – En	IF	Surprised	0.500	0.550	0.34	11
w/ Vid – En	IF	No Link	0.400	0.479	0.55	193
w/ Vid – En	IG	No Link	0.250	0.392	0.43	284
w/o Vid – Pt	IF	Surprised	0.625	0.662	0.48	8
w/o Vid – Pt	IG	Frustrated	0.343	0.360	0.24	24
w/o Vid – Pt	IF	Frustrated	0.297	0.439	0.38	36
w/o Vid – Pt	IF	No Link	0.286	0.387	0.43	138
w/o Vid – Pt	IG	Amused	0.268	0.380	0.38	20
w/o Vid – Pt	IG	Surprised	0.222	0.492	0.43	11
w/o Vid – Pt	IG	No Link	0.222	0.347	0.70	180
w/o Vid – Pt	IF	Amused	0.154	0.125	0.12	5
w/o Vid – En	IG	Amused	0.333	0.352	0.36	27
w/o Vid – En	IF	Surprised	0.333	0.333	0.47	2
w/o Vid – En	IF	Amused	0.200	0.423	0.80	19
w/o Vid – En	IG	No Link	0.200	0.324	0.40	305
w/o Vid – En	IG	Surprised	0.200	0.200	0.28	1
w/o Vid – En	IF	Frustrated	0.167	0.650	0.94	17
w/o Vid – En	IF	No Link	0.000	0.449	1.05	213
w/o Vid – En	IG	Frustrated	0.000	0.289	0.41	28
w/ Video Subtotal – Linked to Cog.			0.500	0.521	0.44	401
– Not Linked to Cog.			0.250	0.390	0.45	917
w/o Video Subtotal – Linked to Cog.			0.276	0.408	0.50	198
– Not Linked to Cog.			0.200	0.371	0.69	836

but a negligible effect size of $\delta = 0.099$). Whereas the role of the subject followed a similar pattern of a slower speech rate for the IG when compared to the IF. This indicates that the subject is required to adapt to the S2S-MT system, but it is not differentiated by the availability of the video channel.

A study of the dialogue acts used by the subjects of the ILMT-s2s corpus and the HCRC Map Task corpus has reported differences in the frequency of the dialogue acts used. As reported in a study of adaptation to video conferencing technology [5], backchannel "*Acknowledgement*" utterances reduced drastically in the dialogues of the ILMT-s2s corpus [19]. This reduction is similar to the reduction in backchannels due to the latency of slower video conferencing systems. However, even with improvement to latency matters, the characteristics of interpreting one language to another will continue to provide a latency that is uncommon in monolingual communication. Also, similarities with slower video conferencing system arise from the length of each utterance with fewer turns and longer utterances for systems with longer latencies [5, p.415] as was also present for in the utterance patterns of the IF in the ILMT-s2s corpus (Figure 5).

Where does the difference in cognitive label quantities come from? "When confronted with truly new technology that had not been an option before, consumers must find some way to match the unexpected with previous experience." [1, p. 56]. Looking at Figure 7 and Figure 8, it is possible to interpret that the subject of the ILMT-s2s corpus collection returned to the experience of Face-to-Face communication when they were provided with the video channel with the S2S-MT system. This interpretation is taken from the difference between the cognitive states that occur after the subject utterance to the system and the overall quantity of cognitive states. The difference can be interpreted as the subjects with the setting of w/ video are expressing a cognitive state to their interlocutor (from "All else" in Table II) as feedback. Though not directly

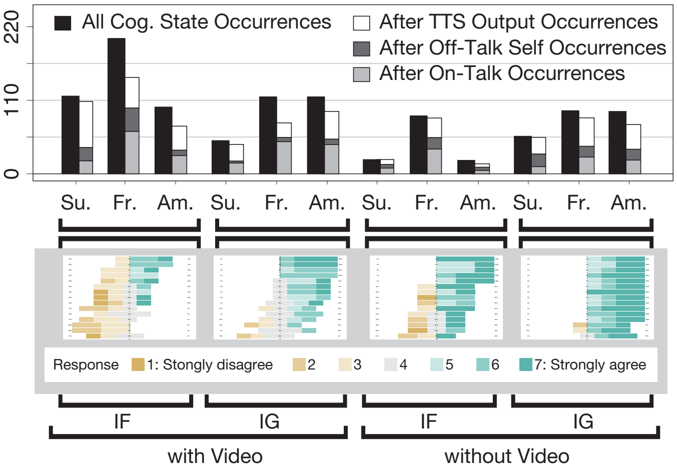


Fig. 7: Bar chart of Portuguese subject's cognitive state count (Su.: *Surprised*, Fr.: *Frustrated*, Am.: *Amused*), with a stacked bar chart of each grouping's user survey Likert score results

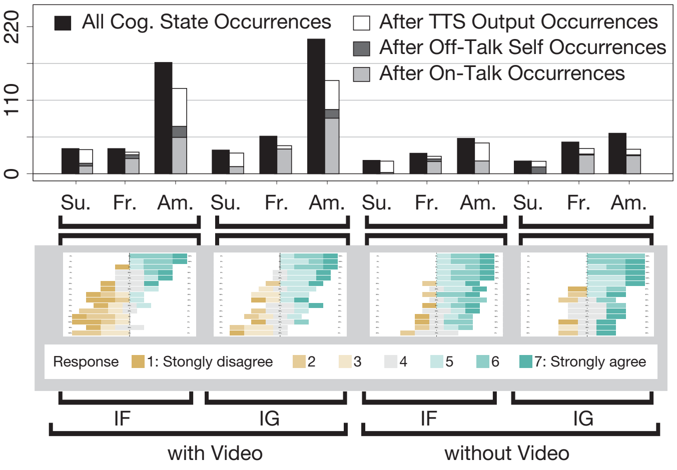


Fig. 8: Bar chart of English Subject's cognitive state count (Su.: *Surprised*, Fr.: *Frustrated*, Am.: *Amused*), with a stacked bar chart of each grouping's user survey Likert score results

asked, the subjects in the user survey comments suggested that the video channel was used as a feedback channel: "It's interesting because as a native speaker you can deduce what someone meant to say even if it comes out wrong, but the other person can't. There was no way for her to know that the word for temple ('templo') was very similar to the word for time ('tempo'), so whenever 'time' came up she looked immensely confused. It was funny but also frustrating" or "More or less, it helped that I was seeing the other person face, so I knew by his facial expressions when some words weren't matching". When this is compared with the number of cognitive state labels in the dialogues w/o Video and also the low number of cognitive states that are not linked to a turn (Figure 7 & Figure 8), it is intuitive to think that the subject is not using expressions of the cognitive state as a feedback channel, and is reacting to the task on hand. If the subject uses the w/ Video system setting interacts with the expectation that the ILMT-s2s System will imitate Face-to-Face communication but with the enhancement of an interpreter, this inevitably compares the strength of a modality and conversation style that one is familiar with to one that only seems familiar, but in reality pushes the user to adopt a conversation style that is very different, a style where the cost (effort of the subject) of communication is increased by the natural latency of interpretation (machine translation). The failure of the video channel to obtain higher user satisfaction may be better explained by the gap between the user's expectation and reality.

Machine mediated technological advancement will not be reversed, due to the possibility to enhance human life [22], [23]. However, talking to a computer interface is said to be similar for the subject as talking to a person who has hearing impairment, an effort that is more "exaggerated" than H2H communication due to the difficulty to understand it's limitation [24]. It is therefore critical that computer mediated communication systems have the capacity to adapt to variable human communication styles, by increasing the varying differences that we have [25] so that the transitional adaptation is as smooth as possible. An adaptation period of 8 months was required for highly motivated, error critical professionals to appreciate an image/video channel [26]. Now, how long will it take someone who is scared of change and technology?

Acknowledgment

This research is supported by Science Foundation Ireland through the CNGL Programme (Grant 12/CE/I2267) and the Research Centres Programme (Grant 13/RC/2106) in the ADAPT Centre (www.adaptcentre.ie) at Trinity College Dublin. The ADAPT Centre for Digital Content Technology is cofunded under the European Regional Development Fund.

References

[1] K. Lipartito, "Picturephone and the Information Age: The Social Meaning of Failure," *Technology and Culture*, vol. 44, no. 1, pp. 50–81, 2003.

[2] O. Morikawa and T. Maesako, "HyperMirror: Toward Pleasant-to-use Video Mediated Communication System," in *CSCW'98*, 1998, pp. 149–158.

[3] W. D. Lewis, "Skype translator: Breaking down language and hearing barriers," in *Translating and the Computer (TC37)*, 2015.

[4] K. Hara and S. T. Iqbal, "Effect of Machine Translation in Interlingual Conversation: Lessons from a Formative Study," in *CHI'15*, 2015, pp. 3473–3482.

[5] B. O'Conaill, S. Whittaker, and S. Wilbur, "Conversations Over Video Conferences: An Evaluation of the Spoken Aspects of Video-Mediated Communication," *Human–Computer Interaction*, vol. 8, no. 4, pp. 389–428, 1993.

[6] K. E. Finn, A. J. Sellen, and S. B. Wilbur, Eds., *Video-Mediated Communication*. Lawrence Erlbaum Associates, Inc., 1997.

[7] B. Kane and S. Luz, "Probing the use and value of video for multidisciplinary medical teams in teleconference," in *CBMS'06*, 2006, pp. 518–523.

[8] A. Hayakawa, S. Luz, L. Cerrato, and N. Campbell, "The ILMT-s2s Corpus — A Multimodal Interlingual Map Task Corpus," in *LREC 2016*, 2016, pp. 605–612.

[9] A. H. Anderson, M. Bader, E. G. Bard, E. Boyle, G. Doherty, S. Garrod, S. Isard, J. Kowtko, J. McAllister, J. Miller, C. Sotillo, H. S. Thompson, and R. Weinert, "The HCRC Map Task Corpus," *Language and Speech*, vol. 34, no. 4, pp. 351–366, 1991.

[10] J. R. Lewis, "IBM Computer Usability Satisfaction Questionnaires: Psychometric Evaluation and Instructions for Use," IBM Corporation, Boca Raton, FL, USA, Technical Report 54.786, 1993.

[11] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes, "ELAN: a Professional Framework for Multimodality Research," in *LREC 2006*, 2006, pp. 1556–1559.

[12] I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos *et al.*, "The AMI meeting corpus," in *Measuring Behavior 2005*, 2005, pp. 137–140.

[13] A. Popescu-Belis, "Dialogue acts: One or more dimensions," *ISSCO WorkingPaper 62*, 2005.

[14] P. Juel Henrichsen and J. Allwood, "Predicting the Attitude Flow in Dialogue Based on Multi-Modal Speech Cues," in *NEALT2012*, 2013, pp. 47–53.

[15] T. J. Taylor, *Mutual Misunderstanding: Scepticism and the Theorizing of Language and Interpretation*. Duke University Press, 1992.

[16] A. Hayakawa, L. Cerrato, N. Campbell, and S. Luz, "A Study of Prosodic Alignment in Interlingual Map-Task Dialogues," in *ICPhS XVIII*, 2015, paper 0760.1–5.

[17] A. Hayakawa, C. Vogel, S. Luz, and N. Campbell, "Speech Rate Comparison when Talking to a System and Talking to a Human: A study from a Speech-to-Speech, Machine Translation mediated Map Task," in *INTERSPEECH 2017*, 2017, In Press.

[18] A. Hayakawa, F. Haider, S. Luz, L. Cerrato, and N. Campbell, "Talking to a system and oneself: A study from a Speech-to-Speech, Machine Translation mediated Map Task," in *SP8*, 2016, pp. 776–780.

[19] A. Hayakawa, S. Luz, and N. Campbell, "Talking to a System and Talking to a Human: A Study from a Speech-to-Speech, Machine Translation Mediated Map Task," in *INTERSPEECH 2016*, 2016, pp. 1422–1426.

[20] A. Hayakawa, F. Haider, L. Cerrato, N. Campbell, and S. Luz, "Detection of Cognitive States and Their Correlation to Speech Recognition Performance in Speech-to-Speech Machine Translation Systems," in *INTERSPEECH 2015*, 2015, pp. 2539–2543.

[21] L. Cerrato, A. Hayakawa, N. Campbell, and S. Luz, *FETLT 2015, Seville, Spain, Nov. 19-20, 2015*. Cham: Springer International Publishing, 2016, ch. A Speech-to-Speech, Machine Translation Mediated Map Task: An Exploratory Study, pp. 53–64.

[22] P. Baranyi and Á. Csapó, "Definition and Synergies of Cognitive Infocommunications," *Acta Polytechnica Hungarica*, vol. 9, no. 1, pp. 67–83, 2012.

[23] P. Baranyi, Á. Csapó, and P. Várlaki, "An Overview of Research Trends in CogInfoCom," in *INES 2014*, 2014, pp. 181–186.

[24] H. P. Branigan, M. J. Pickering, J. Pearson, and J. F. McLean, "Linguistic alignment between people and computers," *Journal of Pragmatics*, vol. 42, no. 9, pp. 2355–2368, 2010.

[25] A. Esposito, A. M. Esposito, and C. Vogel, "Needs and challenges in human computer interaction for processing social emotional information," *Pattern Recognition Letters*, vol. 66, pp. 41–51, 2015.

[26] B. Kane and S. Luz, "Multidisciplinary Medical Team Meetings: An Analysis of Collaborative Working with Special Attention to Timing and Teleconferencing," *Computer Supported Cooperative Work*, vol. 15, no. 5–6, pp. 501–535, 2006.