A Flexible Dual Task Paradigm for Evaluating an Embodied Conversational Agent: Modality Effects and Reaction Time as an Index of Cognitive Load

Catherine J. Stevens, Guillaume Gibert, Yvonne Leung, and Zhengzhi Zhang

MARCS Auditory Laboratories, University of Western Sydney,
Locked Bag 1797, Penrith, NSW 2751, Australia
{kj.stevens,g.gibert,y.leung,z.zhang}@uws.edu.au

Abstract. A new experimental method based on the dual task paradigm is used to evaluate speech intelligibility of an embodied conversational agent (ECA). The experiment consists of the manipulation of auditory-visual (AV) versus auditory-only (A) presentation of speech. In the dual task, participants perform two tasks concurrently. The secondary task is sensitive to cognitive processing demands of the primary task. In the primary task participants either shadowed words or named the superordinate categories to which words belonged, as the word items were spoken by the ECA under A or AV conditions. Reaction time (RT) on the secondary task–swatting a fly on the ECA face–was affected by the difficulty of the concurrent task. The secondary RT was affected by modality of presentation of the primary task. Using a relatively primitive ECA, RT on the secondary task was significantly slower when shadowing occurred in AV versus A conditions. The benefits of this evaluation system, that returns quantitative behavioural data and self-report ratings, are discussed.

Keywords: Evaluation, Embodied Conversational Agent, Dual Task, Divided Attention, Reaction Time, Shadowing.

1 Introduction

There has been increasing interest and demand for ECA evaluation as more agents and speech, face, and emotion models have been developed. Taxonomies [e.g., 1] and frameworks [e.g., 2] have been proposed often emphasizing the need to distinguish features of user, agent, and task. It is more common now for evaluation of ECAs, or component models such as natural language generation or text to speech (TTS) synthesis systems, to consist of both objective and subjective measures [2-7]. There are still instances, however, where data are collected in the absence of the manipulation of specific variables (comparison conditions) or without a control condition [e.g., 8]. Interpretation of such data without a baseline reference or comparison group is necessarily limited. A promising technique that builds on the collection of both objective and subjective data is the application of an experimental method wherein particular variables of theoretical interest or design relevance are manipulated systematically [e.g., 9, 10].

H. Högni Vilhjálmsson et al. (Eds.): IVA 2011, LNAI 6895, pp. 331–337, 2011.

The present study develops a dual task paradigm to gauge indirectly and sensitively the cognitive demand or mental workload imposed by the presence of a very basic ECA model. While improvements to the AV speech, facial expression, and attention models are in progress, the basic model is used to illustrate the logic and flexibility of the dual task paradigm to elicit a range of quantifiable and interpretable behavioural responses and its potential for systematic comparison of different models within or across different ECAs.

1.1 The Architecture and Logic of the Dual Task Paradigm

The dual task paradigm is a useful method to investigate dividing attention across two tasks. The paradigm involves performing two tasks concurrently resulting in impaired behavioural performance on one or both tasks [11]. The general assumption is that attention is finite–either limiting the extent to which two tasks can be carried out at the same time [12] or more flexible with attentional allocation occurring moment to moment depending on task instructions and priorities [11,13,14].

In the present study, participants perform a cognitive word-based primary task and secondary reaction time (RT) task at the same time. The primary task has two levels of difficulty. The easy version involves shadowing or saying aloud the word that was uttered by the ECA–the spoken word being a sensory cue. The more difficult version of the primary task requires the participant to name the superordinate category to which the word belongs–here the spoken word is a semantic cue. With a flexible view of attention, relatively early selection (shadow the word) is possible with a sensory cue but a later mode of selection (categorise the word) is necessary when the word serves as a semantic cue. The secondary task requires a button press response to a visual target on the ECA's face; the target is a small fly. The secondary task is used to measure potential capacity expended on the cognitive task. The rationale is that the greater the capacity allocated to the cognitive task the less capacity available for monitoring the fly and the longer the RTs on the secondary task should be [13]. This is regardless of whether the two tasks involve the same or multiple modalities [14]. Attentional capacity expended is akin to mental workload [15].

We compare the facilitation or impediment on processing achieved by the presence of an ECA producing the primary task sensory or semantic cues. In the auditory-visual (AV) condition, the ECA utters individual words and a participant sees the ECA utter the words. In the auditory only (A) condition, the ECA is present but there are no lip movements, only the voice uttering the individual word items. If the ECA AV model is effective and intelligible then this should facilitate shadowing and we should see equal or reduced RTs on the secondary task in the AV versus A condition. Conversely, if the AV model is ineffective then there will be no difference or poorer secondary task RTs on the AV versus A conditions. The relatively demanding category naming task is included to investigate any interaction between primary task demand and multi- versus uni-modal stimuli on secondary task RTs. A baseline of RTs on the fly swatting task is obtained by presenting the secondary task on its own, serving as a reference from which to measure the capacity (RT) required for the cognitive task. The secondary task RT ordering should be: baseline < shadowing < category naming.

2 Method

2.1 Participants, Stimuli, Equipment, and Procedure

Forty-seven female first year psychology students (M = 20.60 years, SD = 6.42) from the University of Western Sydney (UWS) participated in the study for course credit.

Thirty words from each superordinate category (Cooking, Animal, Seascape) were used as sensory or semantic cues in the shadowing and category-naming version of the primary task, respectively. A one-way analysis of variance (ANOVA) showed that there was no significant difference in word frequency between categories, $F(2,87)=.16$, $p=.90$, η^2_p =.004. Thirty-seven words had one syllable, 51 had two syllables, and two had three syllables. The nine rating scales consisted of five steps labelled from "totally disagree" (1) through to "totally agree" (5).

The ECA was displayed on a Cueword Teleprompter with a colour CCTV video camera and a shotgun microphone for videorecording. Two laptops were connected with a network switch for sending commands from the event manager program on one laptop to another that displayed the ECA and sent the image to the teleprompter. The audio from the ECA was transferred from the laptop to the USB Audio Capture and sent to the headphones and an Ultra Low-noise 8-input 2-Bus Mixer. The mixer also received audio input from the participants and sent the voice of both the ECA (IBM Viavoice) and participants to a DV capture device that transferred all audio input to the recording program. The video camera also sent images directly to the program.

Participants started with the baseline (simple RT only) task while the order of performing the shadowing and category naming tasks was counterbalanced. In the baseline RT task, participants looked at the ECA (static face) and pressed the spacebar as soon as they saw a static fly appearing. RT to the fly was measured from fly onset time. In the shadowing task, participants were instructed to repeat the word that the ECA said (primary task-sensory cue) while concurrently performing the RT (secondary) task. The ECA pronounced 90 words one by one with a 2 s inter-stimulus interval (ISI) between word items. Participants repeated the word as the word was uttered by the ECA. At the same time, they had to press the spacebar whenever they saw a fly appearing on the screen. In the category-naming task (primary task-semantic cue), the ECA pronounced the same 90 words as in the shadowing task and at the same rate of presentation but in a new order. This time, participants were asked to name one of three superordinate categories to which the spoken word belonged while performing the RT task concurrently. In the auditory-only condition, participants looked at a static face version of the ECA with auditory output throughout the experiment. In the auditory-visual condition, a dynamic face of the ECA (with lip movements somewhat correlated with spoken items) was presented in the shadowing and category naming tasks. At the end of the experiment, participants assigned ratings to different qualities of the ECA and the interaction.

3 Results

3.1 Secondary Task

3.1.1 Reaction Time

RTs refer to correct responses on the secondary (fly swatting) task and reported as milliseconds (ms). There was a significant main effect of task, $F(2, 2254)=845.28$, $p<.001$, $\eta^2_p=.43$. Pairwise comparisons showed the ordering of tasks to be as expected: Baseline (M=429.13, SE=3.45) significantly faster than Shadowing (M=581.80, SE=4.58), which was significantly faster than Category naming (M=672.36, SE=5.58). There was a main effect of modality, $F(1, 1127)=22.49$, $p=.001$, $\eta^2_p=.02$ with significantly faster RTs recorded on the secondary task in the A-only condition (M=546.52, SE=0.97) compared with the AV condition (M=575.24, SE=0.95); see Figure 1. There was a significant task by condition interaction, $F(2, 2254)=7.11$, $p=.001$, $\eta^2_p=.006$. All levels of task (baseline versus shadowing, baseline versus category, and shadowing versus category) differed significantly from each other in both the Auditory and Auditory-Visual modality conditions, $p<.001$. Modality had the greatest impact on RT during the shadowing task relative to baseline and category naming.

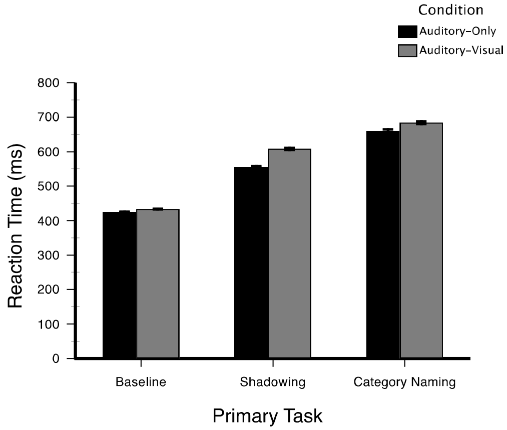

Fig. 1. Mean RT (ms) on the secondary (fly swatting) task shown as a function of Auditory-only and Auditory-Visual conditions and three levels of the primary task

3.1.2 Accuracy

In the secondary (fly swatting) task, it should be the case that accuracy is highest during the baseline condition, followed by shadowing and then the category naming condition. There was a significant main effect of task, $F(2,37)=5.80$, $p=.006$, $\eta^2_p=.24$. Accuracy on the secondary task was significantly higher during the baseline (M=.99, SE=.003) than the category naming task (M=.96, SE=.009), $p=.004$ and higher during the shadowing (M=.99, SE=.005) than the category naming task, $p=.02$. The mean accuracy scores on the secondary task, all > 95%, indicate that participants attended to the primary fly swatting task diligently and accurately. Accuracy on the secondary task did not differ across AV and A conditions.

3.2 Primary Task

3.2.1 Shadowing and Category Naming Latencies

Shadowing latencies were measured from the onset of the word spoken by the ECA to the onset of the shadowing response. The mean shadowing latency in the A condition was 372.52 ms (*SD*=158.82) and the AV condition was 366.14 ms (*SD*=151.48); there was no significant difference.

3.2.2 Accuracy

Mean accuracy in response to a sensory cue to shadow the word was .92 (*SD*=.04) in A and .91 (*SE*=.03) in the AV condition with no significant difference between these conditions. The overall accuracy exceeding 90% indicates that the individual word items uttered by the ECA were generally intelligible. Category naming (*M*=.86, *SD*=.08) was more difficult than shadowing (*M*=.91, *SD*=.03), *F*(1,38)=13.68, *p*=.001, η^2_p=.27. There was no word task x modality interaction.

3.3 Self-report Ratings

Table 1 shows the mode self-report ratings assigned to the nine rating scale items for A and AV conditions. The ratings differ significantly from the midpoint of the scale for both A-only *t*(8)=30.61, *p*<.001 and for AV conditions *t*(8)=33.16, *p*<.001; ratings did not differ significantly from each other. There was no effect of modality on mean ratings, A-only (*M*=3.66, *SD*=0.36) and AV (*M*=3.52, *SD*=0.32).

Table 1. Mode Ratings of ECA and Interaction Quality, Enjoyment and Engagement for auditory-only (A-only) and auditory-visual (AV) conditions; minimum possible rating is 1 ("totally disagree") and maximum possible rating is 5 ("totally agree")

Item	A-Only	AV
I find the Head likeable	4	4
I find the Head engaging	4	4
I find the Head easy to understand	2	2
I find the Head life-like	5	4
I find the Head humorous	3	4
The Head kept my attention	4	4
I would like to interact with the Head again	3	4
I enjoyed interacting with the Head	3	4
I felt as if the Head was speaking just to me	5	5

4 Discussion

The dual task paradigm has been developed as the means to evaluate components of an ECA using the experimental method. Results reveal that the paradigm works. In the present case, where a relatively primitive version of an ECA has been evaluated, the results indicate that the AV speech model does not enhance user perception. Indeed, under some circumstances when task demand is high and the concurrent task relies on speech perception, e.g., shadowing, performance in response to the current AV model impedes RT relative to the auditory only condition. Importantly, performance on the primary task, reflected in shadowing accuracy and latency, are not affected by modality with comparable results in AV and auditory only conditions. Thus, the existing AV model is intelligible, yielding 91-92% shadowing accuracy, but the relatively poor AV integration of spoken word and visemes has a significant processing cost reflected in slower RTs recorded on the secondary task when concurrently shadowing items presented AV compared with presentation of just audio versions of the items. In other words, the poorly integrated visual cues to the spoken items are distracting. The secondary task RT modality x task interaction indicates greater cognitive load or processing cost performing the concurrent task when shadowing (but not category naming) is in response to the AV model. The baseline of RT on the secondary (fly-swatting) task serves as a reference from which we can estimate relative capacity (RT) required for the different levels of the cognitive task. There was no significant effect of modality on baseline RT on the secondary task suggesting that the modality effect observed in shadowing and category naming is not simply overload from the presentation of concurrent visual stimuli.

The evaluation paradigm is a shell into which different modules or systems can be incorporated and systematically and quantitatively compared. The secondary task is sensitive to demands of the primary task and facilitation or impediment from different ECA models or component parts. A comparison with human video has recently been conducted with results currently being analysed.

Acknowledgments. This research is supported by the Thinking Head project and the Australian Research Council and the National Health and Medical Research Council (TS0669874). We thank Stelarc, Damith Herath, and Staci Parlato-Harris.

References

1. Ibister, K., Doyle, P.: Design and evaluation of embodied conversational agents: a proposed taxonomy. In: Proc 1st Intl Joint Conf Autonomous Agents and Multi-Agent System, AAMAS 2002, Bologna, Italy (2002)
2. Catrambone, R., Stasko, J., Xiao, J.: ECA as user interface paradigm: experimental findings within a framework for research. In: Ruttkay, Z., Pelachaud, C. (eds.) From Brows to Trust: Evaluating Embodied Conversational Agents, pp. 239–267. Springer, The Netherlands (2005)
3. Buisine, S., Abrillian, S., Martin, J.-C.: Evaluation of multimodal behaviour of embodied agents. In: Ruttkay, Z., Pelachaud, C. (eds.) From Brows to Trust: Evaluating Embodied Conversational Agents, pp. 217–238. Springer, The Netherlands (2005)

4. Koller, A., Striegnitz, K., Byron, D., Cassell, J., Dale, R., Dalzel-Job, S., Oberlander, J., Moore, J.: Validating the web-based evaluation of NLG systems. In: Proc ACL-IJCNLP, Singapore (2009)
5. Ruttkay, Z., Pelachaud, C. (eds.): From brows to trust: evaluating embodied conversational agents. Kluwer Academic Publishers, Dordrecht (2005)
6. Sharp, H., Rogers, Y., Preece, J.: Interaction design: Beyond human-computer interaction, 2nd edn. John Wiley & Sons, Ltd., Chichester (2007)
7. Stevens, C., Lees, N., Vonwiller, J., Burnham, D.: On-line experimental methods to evaluate text-to-speech (TTS) synthesis: Effects of voice gender and signal quality on intelligibility, naturalness and preference. Computer Speech & Language 19, 129–146 (2005)
8. Kopp, S., Gesellensetter, L., Krämer, N.C., Wachsmuth, I.: A conversational agent as museum guide – design and evaluation of a real-world application. In: Panayiotopoulos, T., Gratch, J., Aylett, R.S., Ballin, D., Olivier, P., Rist, T. (eds.) IVA 2005. LNCS (LNAI), vol. 3661, pp. 329–343. Springer, Heidelberg (2005)
9. Bailly, G., Raidt, S., Elisei, F.: Gaze, conversational agents and face-to-face communication. Speech Comm. 52, 598–612 (2010)
10. Badin, P., Tarabalka, Y., Elisei, F., Bailly, G.: Can you 'read' tongue movements? Evaluation of the contribution of tongue display to speech understanding. Speech Comm. 52, 493–503 (2010)
11. Karatekin, C., Couperus, J.W., Marcus, D.J.: Attention allocation in the dual-task paradigm as measured through behavioral and psychophysiological responses. Psychophysiol. 41, 175–185 (2004)
12. Pashler, H., Johnston, J.C.: Attentional limitations in dual-task performance. In: Pashler, H. (ed.) Attention, pp. 155–189. Psychology Press, East Sussex (1998)
13. Johnston, W.A., Heinz, S.P.: Flexibility and capacity demands of attention. J. Experimental Psychol.: General 107, 420–435 (1978)
14. Wickens, C.D.: Multiple resources and performance prediction. Theoretical Issues in Ergonomic Science 3, 159–177 (2002)
15. Fisk, A.D., Derrick, W.L., Schneider, W.: A methodological assessment and evaluation of dual-task paradigms. Current Psychological Research & Reviews 5, 315–327 (1986)