Available online at www.sciencedirect.com

SciVerse ScienceDirect

Int. J. Human-Computer Studies 71 (2013) 440–454

International Journal of Human-Computer Studies

www.elsevier.com/locate/ijhcs

Evaluating a synthetic talking head using a dual task: Modality effects on speech understanding and cognitive load

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

MARCS Institute, University of Western Sydney, Australia

Received 1 September 2011; received in revised form 7 December 2012; accepted 19 December 2012
Communicated by P. Mulholland
Available online 3 January 2013

Abstract

The dual task is a data-rich paradigm for evaluating speech modes of a synthetic talking head. Three experiments manipulated auditory–visual (AV) and auditory-only (A-only) speech produced by text-to-speech synthesis from a talking head (Experiment 1—single task; Experiment 2—dual task), and natural speech produced by a human male similar in appearance to the talking head (Experiment 3—dual task). In a dual task, participants perform two tasks concurrently with a secondary reaction time (RT) task 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 under AV (dynamic face with lips moving) or A-only (static face) speech modes. First, it was hypothesized that category naming is more difficult than shadowing. The hypothesis was supported in each experiment with significantly longer latencies on the primary task and slower RT on the secondary task. Second, an AV advantage was hypothesized and supported by significantly shorter latencies for the AV modality on the primary task of Experiment 3 and with partial support in Experiment 1. Third, it was hypothesized that while the AV modality helps it also creates great cognitive load. Significantly longer RT for AV presentation in the secondary tasks supported this hypothesis. The results indicate that task difficulty influences speech perception. Performance on a secondary task can reveal cognitive demand that is not evident in a single task or self-report ratings. A dual task will be an effective evaluation tool in operational environments where multiple tasks are conducted (e.g., responding to spoken directions and monitoring displays) and an implicit, sensitive measure of cognitive load is imperative.

Keywords: Evaluation; Avatar; Dual task; Divided attention; Reaction time; Shadowing

1. Introduction

Evaluation is a crucial phase in the development of any new or modified complex system with an increasing demand for evaluation of synthetic talking heads as more avatars and speech, face, and emotion models are developed. It is appealing to apply rigorous experimental methods to evaluate usability, perceptual quality or intelligibility of local and/or global aspects of a synthetic talking head. Ideally, the method veils from users the hypothesis under investigation and returns quantitative data that can be tested for statistical significance. It would be efficacious if the same evaluation shell could be used in a range of settings for systematic comparison of different modules or systems; for example, combined with the LIPS2008 visual speech synthesis challenge (Theobald et al., 2008). Finally, evaluation needs to take place under conditions of varying demand where, for example, user attention is divided across multiple tasks. These are the goals of the present proof of concept. In a dual task, participants perform two unrelated tasks concurrently with performance on one task being an indicator of cognitive demand of responding to various instantiations of the talking head in the other task. Experimental hypotheses

*Correspondence to: School of Social Sciences & Psychology and MARCS Institute, University of Western Sydney, Locked Bag 1797, Penrith, NSW 2751, Australia. Tel.: +61 2 9772 6324; fax: +61 2 9772 6040.
E-mail address: kj.stevens@uws.edu.au (C.J. Stevens).
URL: http://marcs.uws.edu.au/ (C.J. Stevens).

1071-5819/$ - see front matter
http://dx.doi.org/10.1016/j.ijhcs.2012.12.003

are tacit and the objective behavioural accuracy and reaction time measures recorded in response to the cognitive tasks can be correlated with more explicit, subjective, ratings of avatar engagement, ease of understanding and likeability.

Methods of evaluation will be reviewed followed by a rationale for the application of a dual task paradigm as an implicit evaluative technique in the context of auditory–visual speech perception. We then report the results of three dual task experiments in which auditory only (A-only; speech plus static face) and auditory–visual (AV; speech plus dynamic face) modes of a synthetic talking head or human were compared. Speech understanding was gauged from performance accuracy and latency on shadowing and word categorisation tasks, and ease of processing inferred from reaction time on a concurrent task under levels of increasing cognitive load.

2. Methods for evaluating synthetic talking heads: Implicit and explicit perceptual tasks

A detailed scheme for perceptual evaluation of video-realistic speech has been developed by Geiger et al. (2003). They distinguish between two types of experiments. Those that involve explicit perceptual discrimination such as Turing tests where experiment participants *distinguish* (visually) between real and synthetic image sequences of the same utterances, and implicit perceptual discrimination where researchers infer visual speech *recognition* by comparing lip reading performance of real and synthetic sequences of the same utterances. In their study, Geiger et al. found that neither real nor synthetic stimuli were better distinguished. However, using the lip reading task, they observed better recognition for real than for synthetic utterances. Geiger et al. concluded that the latter implicit perceptual discrimination task is more sensitive as an evaluative method.

Similarly, in their proposal of the LIPS2008 Visual Speech Synthesis Challenge, Theobald et al. (2008) argue that "synthesized talking faces require subjective evaluation" emphasizing the need for perception tests that shed light on what is perceptible. The LIPS challenge involves evaluation of visual speech synthesis intelligibility and naturalness. Sentence level utterances – phonetically-balanced semantically unpredictable sentences (Benoit et al., 1996) – are used as stimuli which participants then transcribe. The task yields accuracy but no response time (i.e., cognitive processing time) data and is an explicit task with the goal of speech intelligibility obvious to participants. As an example of the approach, Mattheyses et al. (2009) used the LIPS2008 visual speech synthesis challenge database and obtained participant ratings of visual speech naturalness and synchrony between audio and visual tracks.

While rating scales provide insight into subjective assessment of aspects of a synthetic talking head they are explicit with the intent of the task in full view to participants. One risk associated with hypotheses being overt through ratings is that participants attempt to provide responses that they think the experimenter is seeking (Dell et al., 2012; Orne, 1962). Moreover, assigning a rating is a form of introspection and insensitive to more covert cognitive processes that, through learning, may have become automatic or are difficult to verbalise (e.g., creative thinking, problem solving, inductive or deductive reasoning). Thus, there is a need for more implicit evaluation methods that minimize demand characteristics (Orne, 1962) and where cognitive processes can be inferred and quantified from behaviour. For example, Ito and Speer (2006) gauged listeners' perceptual and cognitive processing of intonational prominence from eye movement latencies and concluded that eye movements are an effective online task with respect to prosody processing.

Shadowing is another indirect method that is sensitive to task manipulations and cognitive processing. The close shadowing technique used by Bailly (2003) provides an online quantitative measure of speech intelligibility. Shadowing requires an experiment participant to repeat immediately what has been spoken. Normative data obtained from a comparison of natural stimuli and text to speech synthesis (TTS) indicated an average delay of 70 ms in response to natural stimuli and more than 100 ms for TTS (Bailly, 2003). The basis for the greater delay to TTS is inappropriate or impoverished prosody (Bailly, 2003, p. 11). A small number of shadowers (four) were used in the study; they shadowed continuous speech and knew the sentences. These factors would contribute further to the relatively short latencies obtained.

In the present experiments, we will use shadowing as a tool for evaluating synthetic speech and anticipate relatively long latencies when discrete words are shadowed in the absence of a sentence context. Shadowing latencies will be investigated under A-only and AV single task conditions (Experiment 1) and dual task A-only (lips static) and AV (lips moving) conditions (Experiments 2 and 3). The present study also accords with the need for consistency in the use of test utterances and evaluation metrics (Theobald et al., 2008). We implement a perceptual task that can add to the current suite of evaluation tools and eventually be adapted to work with the test utterances of LIPS2008 and be used to accumulate population norms; it also includes the addition of a less explicit perceptual task to evaluate user performance when attention is divided and tasks vary in difficulty.

An evaluation technique that builds on the collection of both objective and subjective data is the application of the experimental method wherein particular variables of theoretical interest or design relevance are manipulated systematically (e.g., Bailly et al., 2010; Weiss et al., 2010, 2011). Buisine et al. (2004), for example, adopted an experimental evaluative approach obtaining both ratings and recall data. Three different multimodal strategies were attributed to different looking 2D embodied conversational agents (ECAs). This design enabled evaluation of the effects of the multimodal strategy independent of ECA