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.

Crown Copyright © 2012 Published by Elsevier Ltd. All rights reserved.

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

appearance and the factorial experimental design elicited informative results. On the one hand, multimodal strategies (redundant, complementary, or specialized) influenced subjective ratings of the quality of explanation, especially for male users. On the other hand, the ECA appearance affected likeability and recall. The sample of participants was small (two groups of nine) with corresponding low statistical power. Pandzic et al. (1999) too used converging operations in evaluating synthetic talking faces. They compared effects of different face animation techniques on speech understanding and in optimal and noisy conditions and gleaned subjective benefits such as general appeal to the user, improving satisfaction, and so on.

Converging methods will be used here with collection of quantitative reaction time and shadowing accuracy and latency data together with subjective ratings of likeability and humour of the synthetic and natural talkers. These ratings can be correlated with task accuracy to see whether accuracy of performance is positively or negatively associated with user engagement. For example, if captivated by the talking head, accuracy and reaction time performance on a concurrent task may either suffer (i.e., engaged and therefore distracting) or be enhanced (engaged and therefore task-motivated and vigilant). If it is the case that participants are unaware of, or unable to, articulate the cognitive demand associated with responding to a talking head, then we may see changes in behavioural indicators of cognitive processing (e.g., lower accuracy, slower reaction time) in the absence of changes in ratings.

Ouni et al. (2007) report an experiment-based evaluation of visual speech in animated talking heads. A significant contribution is the "relative visual contribution metric" that can be used to compare performance across different experiments. The experimental task, however – speech perception in noise – again makes the goal of the experiment explicit. The one-way experimental design elicited results that showed speech perception to be poorest under unimodal conditions. We will use a less explicit dual task paradigm and, by crossing variables, will potentially see interactions between factors such as unimodal versus bimodal speech mode and task difficulty. The need for inclusion of modality and task difficulty variables is corroborated by Weiss et al. (2011) who argue that the impact of modality is influenced by factors such as scenario and degree of interactivity. Like Weiss et al. (2010) we will include a number of dependent variables or measures.

In the spirit of applying the experimental method and manipulating key variables of interest, the present study develops a dual task paradigm to gauge sensitively the cognitive demand imposed by the presence of a synthetic talking head. The paradigm records both objective (perception) and subjective (rating scale) measures.

3. The architecture and logic of the dual task paradigm

The dual task paradigm is a useful method to investigate dividing attention across two tasks (e.g., Campana et al., 2010; Kahneman, 1973; Stevens et al., 2007). The paradigm involves performing two tasks concurrently resulting in impaired behavioural performance on one or both tasks (Karatekin et al., 2004; Salvucci and Taatgen, 2008). The general assumption underlying the paradigm is that attention is finite – either limiting the extent to which two tasks can be carried out at the same time (Pashler and Johnston, 1998) or more flexible with attentional allocation occurring moment to moment depending on task instructions and priorities (Johnston and Heinz, 1978; Karatekin et al., 2004; Meyer and Kieras, 1997), and activation of multiple versus shared modalities and codes (Wickens, 2002).

Adopted previously by Campana et al. (2010) to evaluate interface decisions in spoken dialogue systems, the dual task paradigm offers an unbiased and fine-grained measure of usability (Campana et al., 2010). In the study reported by Campana, systems were compared by manipulating the primary task (using the spoken interface), while holding the secondary task constant across conditions (flicker detection). The results revealed that generating natural referring expressions that are dependent on discourse context reduces cognitive load, i.e., faster responding to the secondary flicker detection task. As Campana et al. (2010) note, limited capacity cognitive resources are closely related to usability and a system will be most usable if understanding the speech they generate consumes fewest cognitive resources (p. 317). Rather than asking participants about cognitive load and making the hypothesis explicit, cognitive load is inferred from performance on behavioural tasks. The experimental design also enables manipulation of task demand to see whether speech clarity is influenced not only by modality of presentation but also difficulty of the task.

In the present experiments, participants perform a cognitive word-based primary task and, in Experiments 2 and 3, a secondary reaction time (RT) task at the same time. The primary task has two levels of difficulty. The easy version involves shadowing, i.e., saying aloud the word that was uttered by the talking head (Experiments 1 and 2) or human (Experiment 3)—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—in this case the spoken word is a semantic cue. Shadowing is a relatively simple perceptual task whereas category naming is more cognitively demanding requiring access to knowledge of the word and associated categories in long-term memory.

The secondary task, introduced in Experiments 2 and 3, requires a button press response to a visual target on the talking head or human's face; the target is an image of a small fly. The secondary task is used to measure potential capacity expended on the primary task. The rationale is that the greater the capacity allocated to the primary task the less capacity available for monitoring the fly target and the longer the RTs on the secondary task should be (Johnston and Heinz, 1978). This is regardless of whether the two tasks involve the same or multiple modalities

(cf., Kahneman, 1973; Wickens, 2002). Attentional capacity expended can be thought of as mental workload (Fisk et al., 1986/7). A visual target was used in the RT task as an auditory target may mask the words spoken by the talking head. Detecting a visual target, as has been used in the past (e.g., Stevens et al., 2007), also enables independent, varied, and unpredictable timing of the presentation of targets in the secondary task compared with regular and predictable timing of the spoken items in the primary task. It is anticipated that, as category naming is more demanding than shadowing, RT on the secondary fly detection task will be slower while concurrently category naming than concurrently shadowing. In addition to the primary tasks increasing in complexity, we will also manipulate mode of presentation in the primary task as A-only or AV.

4. Modality effects in speech perception

It is well established that when perceiving natural speech, performance is superior in response to auditory–visual speech than to audio-only or video-only speech (Bailly et al., 2002; Sumby and Pollack, 1954; van Wassenhove et al., 2005). Initially it was thought that the benefits of watching the talker occurred when the acoustic information was suboptimal, i.e., in noisy settings. However, the advantage gained from seeing the talker is apparent at all noise levels (Campbell, 2008). Identification of target syllables has been found to be more rapid when presented in AV than in A-only mode (Besle et al., 2004). The McGurk effect – i.e., when presented with incongruent auditory and visual consonant–vowel syllables, participants report a percept different from either the auditory or visual signal – demonstrated that perception of certain speech segments can be influenced by vision (Campbell, 2008; Gibert et al., 2010). In a recent review, Campbell (2008) notes two modes that underpin the seen speech advantage: a complementary mode where vision provides information about some aspects of speech that are hard to hear and which may depend on visibility of the lower face (e.g., see de Paula et al., 2006) and a correlated or redundancy mode where there are regions of similar dynamic patterning across auditory and visual channels.

In the context of synthesized visual speech, Bailly et al. (2002) compared natural articulatory trajectories with synthetic trajectories that had been computed by different movement generation systems from phonetic input. Point light displays were paired with natural acoustic stimuli and participants rated on a five-point scale the degree of coherence between acoustic and facial motion. In line with the benefits of AV speech, the original AV stimuli received the highest coherence ratings. There was little variation in response latency. We will introduce a more challenging task – a primary task with two levels of difficulty (Stevens et al., 2007) – to induce variation in response latencies under AV and A-only conditions. The secondary task will quantify cognitive load in shadowing versus categorising synthetic speech spoken by a synthetic and a natural face under AV and A-only conditions.

5. Experiment 1—Talking head single task

In Experiment 1, participants completed the single primary task of shadowing and category naming in AV and A-only speech modes. Independent variables were speech mode (AV, A-only; between subjects) and word task (shadowing, category naming; within subjects) and the dependent variables were shadowing and category naming accuracy, latencies, and talking head quality ratings. In the auditory–visual (AV) speech mode, the talking head utters individual word items and a participant sees the dynamic talking head utter the words. In the auditory only (A-only) speech mode, the talking head is present but there are no lip movements, only the voice uttering the individual word items. Auditory–visual speech facilitates speech perception generally (Davis and Kim, 2004; Sumby and Pollack, 1954) and particularly in degraded or noisy environments (Kim and Davis, 2004). Accordingly, it is hypothesized that there is an advantage of AV over A-only speech mode on clarity (greater accuracy and shorter latencies) in the primary task. Task difficulty will be reflected in greater accuracy and shorter latencies in the shadowing (sensory cue) task than the category (semantic cue) naming task.

5.1. Method

5.1.1. Participants

A sample of 32 female first year psychology students ($M=21.97$ years, Range = 17–44 years) from the University of Western Sydney completed Experiment 1 and received course credit for participating. All reported having normal or corrected vision and normal hearing. There were 16 participants assigned randomly to the AV speech mode and 16 assigned randomly to the A-only speech mode. Gender of participant was controlled as statistical power would be reduced if there was an interaction between the gender of the talking head and of the participant. All participants had self-reported normal or corrected vision and normal hearing. The research complied with the requirements of the University of Western Sydney Human Research Ethics Committee (H7776).

5.1.2. Stimuli

The synthetic talking head was based on the prosthetic head created in the likeness of performance artist Stelarc (http://stelarc.org). The avatar, shown in Fig. 1, is an animated head displayed on an LCD screen which subtended 24.35° visual angle. The visual front-end is a three-dimensional computer-graphic representation of a male face which is capable of visual speech movements and of displaying basic emotional expressions. The animation component works as a text-to-AV synthesis system: it receives text data intended as speech for the animated face, and generates the speech and corresponding face

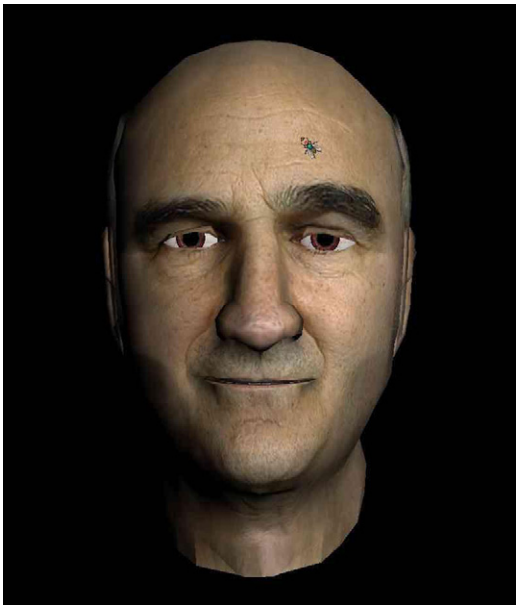

Fig. 1. Screen capture of the talking head (Experiments 1–3) and the visual target on the secondary task, a fly (Experiments 2 and 3).

motion as output. The facial animation is performed by interpolation between a set of 16 visemes; no prephonatory gestures are implemented in this animation model. The system consists of a TTS module; a phoneme-to-face motion database; a phoneme-to-face animation generator; and a face animation module (Burnham et al., 2008). The voice of the talking head is IBM Viavoice text to speech (TTS) synthesis.

The primary task was conducted consisting of A-only and AV presentation of spoken words which participants shadowed and categorized in counterbalanced blocks. Either a static or a dynamic image of the talking head was displayed on the screen and participants listened to the auditory output of the words (synthetic TTS voice) through the headphones.

Thirty words from each superordinate category (Cooking, Animal, and Seascape) were used as sensory or semantic cues in the shadowing and category naming version of the primary task, respectively (see Appendix A). Mean SublexUS word frequency (Brysbaert and New, 2009) of words in Cooking, Animal and Seascape were 308.77 (SD=176.58), 327.57 (SD=172.36) and 331.93 (SD=149.57), respectively, ranging from 108 to 701 across categories. 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_p^2=.004. Thirty-seven words had one syllable, 51 had two syllables, and two had three syllables. Mean word duration was 560.33 ms, SD=130.92. Rating scales consisted of five steps labelled from "totally disagree" (1) through to "totally agree" (5). There were nine rating scale items designed to gauge quality of the synthetic talking head (Table 1).

5.1.3. Equipment

The talking head was displayed on a Cueword Teleprompter (Xpose VGA input monitor) with a colour CCTV video camera (Panasonic WVCL934) installed at the back and a shotgun microphone (Beyer Dynamics MCE86 II) at the side for video recording. Two laptops (Lenovo T500, Microsoft Window XP Professional v.2002) were connected with a network switch (D-Link 10/100 Fast Ethernet switch) for sending commands from the Event Manager programme on one laptop to another which displayed the talking head and sent the image to the teleprompter. The audio sound of the talking head was transferred from the laptop to the USB Audio Capture (EDIROL by Roland UA-25EX) and then sent to the headphones (Sennheiser HD650) and a Ultra Low-noise design 8-input 2-Bus Mixer (Eurorack UB802). The mixer also received audio input from the participants during the recording. It then sent the voice of both the talking head (IBM Viavoice text to speech (TTS) synthesis) and the participants to a DV capture device (Canopus ADVC-55) which transferred all the audio input to the recording programme (Adobe Premiere Pro 2.0) in a computer. The video camera also sent the recorded images directly to the programme.

5.1.4. Procedure

Participants read an information sheet and signed a consent form. All participants were tested in a sound-attenuated booth and were video recorded. Participants were instructed to look at the talking head presented on the computer screen while performing different kinds of tasks. The order of performing the shadowing and category naming tasks was counterbalanced across participants.

In the shadowing task, participants were instructed to repeat the word that the talking head said (primary task-sensory cue). The talking head pronounced 90 words one by one with a 1500 ms inter-stimulus interval (ISI) between word items. Participants were asked to repeat the word immediately, loudly and clearly, as the word was uttered by the talking head. Participants started the task with practice trials of 21 word presentations (different words from the main task).

In the category naming task (primary task-semantic cue), the talking head 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. Practice trials were given at the beginning of the task with 21 different words.

In the A-only condition, participants looked at a static face version of the talking head with auditory output throughout the experiment. In the AV condition, a dynamic face of the talking head was presented in the shadowing and category naming tasks. At the end of the experiment, participants assigned ratings to different qualities of the talking head. The experiment took 30 min.

5.2. Results

5.2.1. Shadowing and category naming latencies

Only correct responses were analyzed from the shadowing (81.01%) and category naming (80.14%) tasks. Latencies

were calculated from the word unicity point to the onset of the response. The unicity point was calculated for each word using the English Lexicon Project database from the Washington University in St. Louis, URL http://elexicon.wustl.edu/query14/query14.asp. The phonological neighbours with the greatest number of shared phonemes were identified and, from those, the neighbour with the highest frequency of occurrence was chosen. The phoneme was then identified at the point where the word became different from the neighbour phonologically.

A mixed 2×2 repeated measures ANOVA was conducted on the response latencies in each word task across A-only and AV speech modes. There was a significant main effect of task, $F(1, 2737) = 3957.50$, $p < .01$, partial $\eta^2 = .59$ with significantly faster responding when shadowing ($M = 457.40$ ms, $SD = 157.57$) than when category naming ($M = 734.82$ ms, $SD = 240.82$). The effect of speech mode was not significant. There was a significant task × speech mode interaction, $F(1,2737) = 13.05$, $p = .001$, partial $\eta^2 = .005$, see Fig. 2a. Participants in the AV speech mode responded more quickly than A only in the shadowing task, $F(1,30) = 184.47$, $p < .001$, with the reverse pattern of responding in the categorization task, $F(1,30) = 155.85$, $p < .001$.

5.2.2. Accuracy

A mixed 2×2 repeated measures ANOVA was conducted on the accuracy rate in each word task across A-only and AV speech modes. There was a significant main effect of task, $F(1, 30) = 11.56$, $p = .002$, partial $\eta^2 = .28$ with accuracy significantly greater in the shadowing task ($M = .92$, $SD = .04$) than in the category naming task ($M = .88$, $SD = .07$). There was no main effect of speech mode and no task × modality interaction.

5.2.3. Self-report ratings

Mode (highest frequency) self-report ratings are shown in Table 1. The results of t-tests conducted on the ratings indicate that the ratings all differ significantly from the midpoint of the scale (3: neither agree nor disagree) for both A-only $t(15) = 18.75$, $p < .001$ and AV modes $t(15) = 20.35$, $p < .001$. A one-way between-subjects ANOVA showed no effect of speech mode on mean ratings, A-only (Mean = 3.34, $SD = .71$) and AV (Mean = 3.45, $SD = .68$). Correlations were calculated separately between ratings of engagement and understanding with the response accuracy in shadowing or naming categories; none were significant.

5.3. Discussion

Experiment 1 consisted of the primary task of shadowing and category naming performed in AV or A-only mode. There was evidence of an AV advantage in latencies on shadowing but not in the more difficult categorization task and no advantage in accuracy data. Because the advantage is evident in only the simpler of the two word tasks, task difficulty and context, as discussed by Weiss et al. (2010, 2011), appear to have an impact. The AV advantage in shadowing is likely to be from the additional cues provided in the visual speech. Such cues to acoustics of the word should facilitate a perceptual task such as shadowing. However, the results suggest that the semantic task of categorizing words spoken by a talking head does not benefit from visual cues. The significant interaction shows that the AV mode helps in shadowing as the visual cues are related to the word reading response. By contrast, the AV mode is a hindrance in category naming as the word reading response needs to be inhibited and the visual cues to the word slow correct responding to the category. The primary task will take the same form in Experiment 2 and be performed concurrently with a secondary RT task.

6. Experiment 2—Talking head dual task

Using the dual task paradigm in Experiment 2, we compare the facilitation or impediment on processing achieved by the presence of a talking head producing the primary task sensory or semantic cues. It is hypothesized that there is a visual speech advantage with shorter latencies on the primary task in both shadowing and category naming when speech is dynamic AV than static A-only. A visual speech advantage may also reduce demand and be reflected in RT on the secondary fly

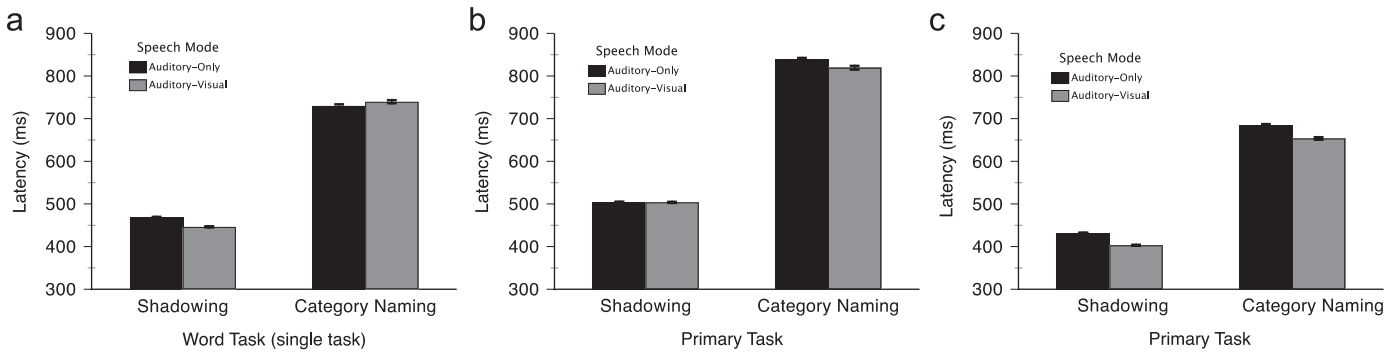

Fig. 2. (a) Experiment 1 (control) single word task: mean shadowing and category naming latencies interacting with speech mode. (b) Experiment 2 (talking head) primary word task: mean shadowing and category naming latencies; there is no interaction with speech mode. (c) Experiment 3 (human video) primary word task: mean shadowing and category naming task latencies. Error bars refer to standard error of the mean.

detection task with faster RT in AV than A-only speech modes. Alternatively, if the mouth is fixated in the AV speech mode then RT will be longer and accuracy poorer on the concurrent visual secondary task in the AV compared with the A-only speech mode.

The relatively demanding category-naming task is included to investigate any interaction between primary task difficulty and multi- versus uni-modal stimuli on the secondary task RT. A baseline of RT on the fly swatting task will be obtained by presenting the secondary task on its own. This serves as a reference from which to measure the capacity (RT) required for the cognitive task. The secondary task RT ordering is hypothesized as: baseline < shadowing < category naming.

Clarity of the talking head speech model can be gauged from shadowing accuracy on the primary task. Accuracy on the secondary RT task will reflect vigilance on that task. Self-report ratings of talking head likeability, engagement, etc., will also be obtained. As there may be a systematic relationship between secondary task RT and ratings (e.g., high ratings of engagement associated with slow secondary task RT), correlations between ratings and secondary task RT will also be calculated. If cognitive demand is not evident or able to be articulated by participants then there will be changes in accuracy and RT performance without any corresponding change in the ease-of-understanding rating scale item.

The aim of Experiment 2 was to investigate the relative cognitive demand of perceiving AV versus A-only speech produced by a talking head. The independent variables were speech mode (AV, A-only) and word task difficulty (shadow word-sensory cue, categorise word-semantic cue) with the former variable between subjects and the latter variable within subjects. A secondary task consisted of baseline (single task) and dual task conditions involving simple RT to a visual target (fly). Dependent variables consisted of secondary task accuracy and RT, primary task shadowing or categorization accuracy, shadowing or categorisation latency, and nine ratings of talking head quality.

6.1. Method

6.1.1. Participants

A new sample of 47 female first year psychology students ($M = 20.60$ years, $SD = 6.42$) from the University of Western Sydney participated in the study for course credit. Data from an additional seven students were excluded from the analysis due to technical issues with stimulus presentation. Participants had English as their first language. There were 20 participants assigned randomly to each of the AV and A-only speech modes. All participants had self-reported having normal hearing; 39 self-reported having normal vision and one self-reported having corrected vision.

6.1.2. Stimuli

Experiment 2 used the same word stimuli and talking head as Experiment 1.

6.1.3. Equipment

The same equipment and setup from Experiment 1 was used.

6.1.4. Procedure

Participants read an information sheet and signed a consent form. All participants were tested in a sound-attenuated booth and were video recorded. They were asked to look at the talking head presented on the computer screen while performing different kinds of tasks. All participants started with the baseline (simple RT only) task while the order of performing the shadowing and category naming tasks was counterbalanced across participants. In the baseline RT task, participants were asked to look at the talking head (static face) and press the spacebar as quickly as possible when they saw an image of a static fly appearing at random intervals on the screen within the face region. The dimensions of the fly were 32×32 pixels. Reaction time of responding to the fly was measured from fly onset time. The fly disappeared once the participant had responded or after 3000 ms if no response was given. Practice trials were provided at the beginning of the task with 10 fly appearances presented at random time intervals ranging from 1 s to 3 s. There were 30 fly appearances presented in random 1–3 s time intervals in the baseline condition and during the shadowing and category (dual task) conditions; for one third of the time there was a fly present.

In the shadowing task, participants were instructed to repeat the word that the talking head said (primary task-sensory cue) while concurrently performing the RT (secondary) task. The talking head pronounced 90 words one by one with a 1500 ms inter-stimulus interval (ISI) between word items. Participants were asked to repeat the word immediately, loudly and clearly, as the word was uttered by the talking head. At the same time, they had to press the spacebar whenever they saw a fly appearing on the screen. Participants started the task with practice trials of 21 word presentations (different words from the main task) and seven fly presentations.

In the category naming task (primary task-semantic cue), the talking head 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. Practice trials were given at the beginning of the task with 21 different words and seven fly presentations. In the dual task conditions, participants were asked to perform the two tasks at once.

As in Experiment 1, in the A-only condition, participants looked at a static face version of the talking head with auditory output throughout the experiment. In the AV condition, a dynamic face of the talking head was presented in the shadowing and category naming tasks. At the conclusion of the task, participants assigned ratings to different qualities of the talking head. The experiment took 30 min.

6.2. Results

The dual-task paradigm yielded a number of behavioural measures: accuracy and reaction time on the secondary (fly detection) task, accuracy and latencies on the primary (word) task shadowing and categorising conditions, and ratings of engagement, humour, likeability, etc. Each measure will be reported separately with reference to specific hypotheses.

6.2.1. Secondary task

6.2.1.1. Reaction time. Reaction times refer to RTs on correct responses and reported as milliseconds (ms). Correct RTs less than 100 ms and greater than 1000 ms were regarded as outliers and removed (281/1200 responses; 23%); the majority of these (16%) occurred in the relatively difficult category naming condition. As a check of the manipulation of task complexity, we expect RT on the secondary task to be ordered baseline < shadowing < category. As hypothesized, there was a significant main effect of task, $F(2, 1860) = 723.65$, $p < .01$, $\eta_p^2 = .44$. Pairwise comparisons, with Sidak adjustment for multiple comparisons, showed the ordering of tasks to be as expected: baseline ($M = 419.71$, $SD = 111.10$) significantly faster than shadowing ($M = 556.95$, $SD = 93.10$), which was significantly faster than category naming ($M = 617.08$, $SD = 92.77$).

There was a main effect of speech mode, $F(1, 930) = 30.26$, $p = .01$, $\eta_p^2 = .03$ with significantly faster RTs recorded on the secondary task in the A-only speech mode ($M = 517.05$, $SD = 135.14$) compared with the AV speech mode ($M = 545.45$, $SD = 126.96$); see Fig. 3a. There was a significant task by speech mode interaction, $F(2, 1860) = 8.20$, $p = .01$, $\eta_p^2 = .01$. A simple effects analysis showed that in both A only and AV speech modes, the RT during the concurrent shadowing task was significantly longer than during the baseline task but shorter than during the concurrent category naming task. Speech mode had the greatest impact on RT during the shadowing task relative to category naming.

6.2.1.2. Accuracy. In the secondary (fly detection) task, and as a check on the manipulation of task complexity, it should be the case that accuracy is highest during the baseline condition, followed by shadowing and then the category naming condition. As hypothesized, there was a significant main effect of task, $F(2,76) = 41.76$, $p < .01$, $\eta_p^2 = .52$. Pairwise comparisons revealed that accuracy on the secondary task was significantly higher during the baseline ($M = .99$, $SD = .02$) than the category naming task ($M = .83$, $SD = .28$), $p = .004$ and higher during the shadowing ($M = .94$, $SD = .08$) than the category naming task, $p = .02$. The mean accuracy scores on the secondary task, all > 80%, indicate that participants attended to the secondary fly detection task with a good level of accuracy. Accuracy on the secondary task did not differ significantly across AV and A-only speech modes.

6.2.2. Primary task

6.2.2.1. Shadowing and category naming latencies. Shadowing and category-naming latencies were measured from the unicity point of the word spoken by the talking head to the onset of the shadowing or category response made by participants. Only items named correctly were analysed (91%). As shown in Fig. 2b the mean shadowing latency in the A-only speech mode was 504.08 ms ($SD = 164.13$) and the AV mode was 503.76 ms ($SD = 157.34$). The mean category-naming latency in the A-only speech mode was 838.09 ms ($SD = 285.92$) and the AV speech mode was 819.42 ms ($SD = 272.38$). A mixed ANOVA on latencies for word tasks × speech mode revealed a significant main effect of task, $F(1, 2978) = 4324.31$, $p < .01$, $\eta_p^2 = .59$ with latencies longer when participants categorised ($M = 828.75$, $SD = 279.33$) than

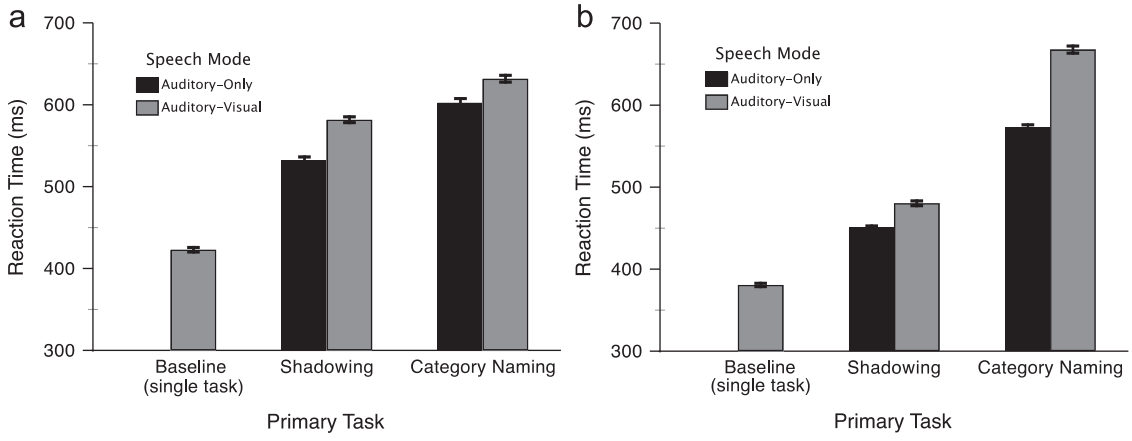

Fig. 3. (a) Experiment 2 (talking head) and (b) Experiment 3 (human video): mean RT (ms) on the secondary (fly detection) task shown as a function of A-only and AV speech modes, the baseline (single task) RT, and the two levels of the primary task. Error bars refer to standard error of the mean.

shadowed ($M=503.92$, $SD=160.74$) words. There was no main effect of speech mode and no task × mode interaction.

6.2.2.2. Accuracy. Mean accuracy in response to a sensory cue to shadow the word was .92 ($SD=.04$) in the A-only speech mode and .91 ($SD=.04$) in the AV speech mode with no significant difference between these unimodal and multimodal conditions. The overall accuracy exceeding 90% indicates that the individual word items uttered by the talking head were generally intelligible. A main effect revealed that category naming ($M=.86$, $SD=.06$) was significantly more difficult than shadowing ($M=.91$, $SD=.06$), $F(1,38)=13.68$, $p=.001$, $\eta_p^2=.27$. There was no word task × speech mode interaction.

6.2.3. Self-report ratings

Table 1 shows the mode (highest frequency) self-report ratings assigned to the nine rating scale items for A-only and AV speech modes. The results of *t*-tests conducted on the ratings indicate that the ratings all differ significantly from the midpoint of the scale (3: neither agree nor disagree) for both A-only $t(8)=30.61$, $p<.001$ and for AV modes $t(8)=33.16$, $p<.001$; ratings did not differ significantly from each other. A one-way between-subjects ANOVA showed no effect of speech mode on mean ratings, A-only (Mean=3.66, $SD=.36$) and AV (Mean=3.52, $SD=.32$). Correlations were calculated separately between ratings of engagement and understanding with secondary task RT whilst also shadowing or naming categories; none of the four correlations was significant.

6.3. Discussion

The dual task paradigm has been applied here as the means to evaluate a talking head using direct and indirect measures of perception. On the primary task – shadowing discrete words spoken by the talking head or naming the category to which the word belonged – accuracy was high (91–92% shadowing) with, as expected, significantly longer latencies in category naming than shadowing. Differences in shadowing and categorizing speed of responding occur because the former is a perceptual task that requires participants repeat the word spoken by the talking head whereas categorizing requires a semantic analysis of the word – recognizing the word, its meaning, and then recognizing a semantic association between the word and one of the three categories. The two levels of this task have been shown to increase task difficulty (e.g., Stevens et al., 2007 adapted from Johnston and Heinz, 1978) and, in the talking head context, the two levels permit a reliable manipulation of cognitive demand. There was no significant difference between A-only and AV speech modes in either primary task latency or accuracy. Thus, under quiet laboratory (zero noise) conditions there is no advantage for visual speech on the primary task. If the task was undertaken in noise, the AV advantage may emerge. As the speech model included 16 visemes but no prephonatory gestures, a different system based on hidden Markov models or concatenation may contribute to a seen speech advantage.

A benefit of the dual task paradigm is the range of dependent variables that is obtained and the way performance is inferred from behaviour. The diagnostic value of the dual task is evident in results on the secondary RT task. As hypothesized, RT on the target detection task was fastest for baseline followed by shadowing and category naming concurrent task conditions. Similarly, accuracy increased from baseline to shadowing to category naming. Thus, we have clear results where the secondary task RT has been slowed by increasingly difficult concurrent tasks. RT on the secondary task was also affected by speech mode of the concurrent task. Specifically, RT on the secondary task was slightly but significantly faster when the concurrent task speech mode was A-only compared with AV and this effect was most pronounced during category naming. Notably, we have detected an effect of speech mode that was not evident in the primary task or in the ratings. Vigilance on the primary task was unaffected

Table 1
Experiments 1, 2 and 3: Mode ratings of 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	Experiment 1 (primary task) talking head		Experiment 2 (dual task) talking head		Experiment 3 (dual task) human video	
	A-Only	AV	A-Only	AV	A-Only	AV
I find the talking head likeable	3	4	4	4	3	4
I find the talking head engaging	4	4	4	4	2	4
I find the talking head easy to understand	4	3	2	2	4	5
I find the talking head life-like	4	4	5	4	4	5
I find the talking head humorous	2	2	3	4	2	1
The talking head kept my attention	4	4	4	4	4	4
I would like to interact with the talking head again	3	4	3	4	3	3
I enjoyed interacting with the talking head	5	4	3	4	3	3
I felt as if the talking head was speaking just to me	4	3	5	5	4	4

by speech mode but the secondary task RT indicates greater cognitive load (slower RT) in the AV mode. Subjective ratings too showed no effect of speech mode. Dividing attention across tasks thus brings into relief the cognitive demand of the talking head modes.

It remains for comparison data to be collected when a human presents the word stimuli. Experiment 3 was designed to investigate patterns of responding when the agent is a video of a human rather than a software-rendered talking head. The human should have a natural and correlated phoneme–viseme AV speech production system and, unlike our talking head, prephonatory gestures. To match as closely as possible characteristics of age, gender, face of the talking head, and a human speaker, the installation artist Stelarc, on whom the talking head was modelled, video recorded the word stimuli for Experiment 3.

7. Experiment 3—Human reference dual task

The design of Experiment 3 was identical to that of Experiment 2 except that, instead of a software-rendered talking head, word stimuli were presented from video footage of a human male of similar age and appearance to the talking head.

7.1. Method

7.1.1. Participants

A new sample comprised 40 female first year psychology students ($M=25.90$ years, $SD=13.33$, Range=17–73 years) from the University of Western Sydney with the majority receiving course credit for participating, and five participants receiving AU$20 for their travel expenses to the laboratory. As in Experiments 1 and 2, gender of participant was controlled to avoid statistical interactions between the gender of the reference (male) and the gender of the participant. All participants had self-reported normal or corrected vision and normal hearing.

7.1.2. Stimuli

The video for each AV trial showed the reference saying a word once; in the A-only condition a static image of the human reference was displayed. An image of a static fly appeared on the video in 30 out of 90 trials randomly. The same list of words was used with 30 words from each superordinate category (Cooking, Animal, and Seascape) in the shadowing and category naming primary tasks. Mean word duration was 794.51 ms, $SD=169.84$ with an ISI of 1700 ms. Words were presented in random order across participants. The same nine-item rating scale from Experiments 1 and 2 was used to evaluate the quality of the human communication.

7.1.3. Equipment

The same equipment from Experiments 1 and 2 was used. Visual angle of the reference subtended 23.53°.

7.1.4. Procedure.

The procedure was identical to that of Experiment 2.

7.2. Results

Accuracy and reaction time data were analysed in separate mixed analyses of variance. As epsilon values were high, Huynh–Feldt corrections were applied. Reaction times refer to correct responses only; RTs < 100 ms or > 1000 ms were treated as errors and removed with 146 of 1200 (12.17%) data points removed.

7.2.1. Secondary task

7.2.1.1. Reaction time. There was a significant main effect of task, $F(1.83, 1927.96)=1497.15$, $p<.01$, partial $\eta^2=.59$. Pairwise comparisons showed significantly faster responding to the fly target in the baseline task ($M=388.84$, $SD=85.37$), slower responding in the shadowing task ($M=464.74$, $SD=100.83$), and slowest responding in the category-naming task ($M=617.59$, $SD=143.47$). There was a significant main effect of speech mode, $F(1, 1052)=65.54$, $p>.01$, partial $\eta^2=.06$ with significantly faster responding in the A-only mode ($M=473.09$, $SD=101.50$) compared with the AV mode ($M=509.56$, $SD=112.23$). Evident in Fig. 3b, there was a significant task × speech mode interaction, $F(1.83, 1927.96)=83.51$, $p<.01$ partial $\eta^2=.07$, with simple effects for both A-only, $F(1.83, 1012.10)=517.91$, $p<.01$, partial $\eta^2=.48$ and AV modes, $F(1.84, 916.02)=970.16$, $p<.01$, partial $\eta^2=.66$. In both A-only and AV speech modes, the RT in the shadowing task was significantly longer than in the baseline task but shorter than in the category-naming task.

7.2.1.2. Accuracy. A significant main effect of task was found, $F(2, 76)=23.82$, $p=<.001$, partial $\eta^2=.39$. Pairwise comparisons showed that the difference in accuracy between baseline ($M=.997$, $SD=.007$) and shadowing conditions ($M=.996$, $SD=.006$) was not significant, but there was significantly higher accuracy in the shadowing than in the category-naming ($M=.97$, $SD=.042$) task. There was also a significant main effect of speech mode, $F(1, 38)=4.05$, $p=.05$, partial $\eta^2=.10$, indicating that overall, there was significantly higher accuracy in the A-only speech mode ($M=.99$, $SD=.016$) compared to the AV ($M=.98$, $SD=.02$) mode.

There was a significant task × speech mode interaction, $F(2, 76)=5.60$, $p=.02$, partial $\eta^2=.13$, see Fig. 4. Simple effect analyses were conducted separately for A-only and AV modes. The simple effects for A-only were significant, $F(2, 38)=4.44$, $p=.04$, partial $\eta^2=.19$, showing that only accuracy in category-naming was significantly lower than the baseline, $F(1, 19)=5.12$, $p=.04$, partial $\eta^2=.21$. There was no significant difference in accuracy between baseline and shadowing conditions on the secondary task. The simple effects for AV were also significant, $F(2, 38)=20.40$,

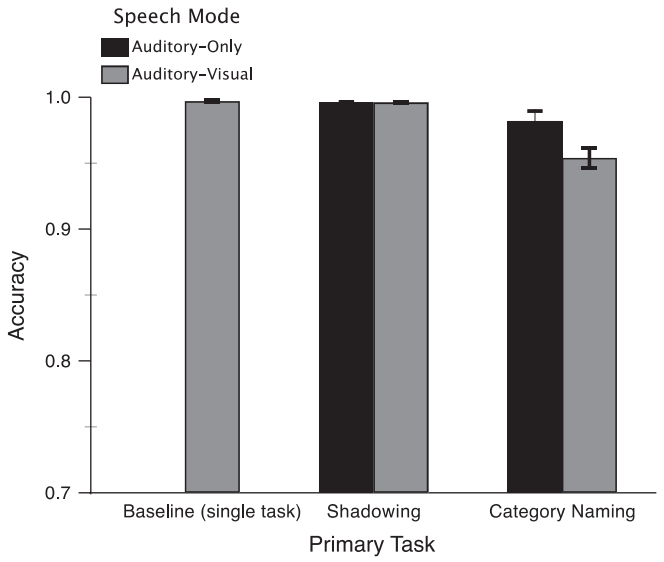

Fig. 4. Experiment 3: Accuracy (proportion correct) on the secondary (fly detection) task comparing the A-only and AV speech modes, the baseline (single task), and the two levels of the primary task. Error bars refer to standard error of the mean.

$p < .001$, partial $\eta^2 = .52$, showing that only accuracy in category naming was significantly lower than the baseline task, $F(1, 19) = 19.84$, $p < .001$, partial $\eta^2 = .51$. There was no significant difference in accuracy between baseline and shadowing tasks in the AV speech mode.

7.2.2. Primary task

7.2.2.1. Shadowing and category naming latencies. Only correct responses were included in the analyses of response latency data obtained from the shadowing (99.4%) and category-naming (97.6%) tasks. Latencies refer to the difference between the unicity point of each word and response onset. As shown in Fig. 2c, 2×2 mixed repeated measures ANOVA revealed a significant main effect of modality, $F(1, 3513) = 24.69$, $p < .01$, partial $\eta^2 = .007$ with faster responding in AV ($M = 528.31$ ms, $SD = 206.88$) than A only mode ($M = 558.13$, $SD = 201.73$). There was also a main effect of task, $F(1, 3513) = 4841.60$, $p < .01$, partial $\eta^2 = .58$ with significantly faster responding on shadowing ($M = 417.46$, $SD = 167.96$) than category naming ($M = 669.02$, $SD = 241.82$). There was no modality $\times$ task interaction.

7.2.2.2. Accuracy. Mean accuracy in shadowing words in the A-only speech mode was .995 ($SD = .01$) and in the AV mode .98 ($SD = .02$). Mean accuracy in naming the category in the A-only speech mode was .99 ($SD = .01$) compared with .98 ($SD = .02$) in the AV mode. One-way ANOVAs revealed that there were no main effects of speech mode on either shadowing or category-naming tasks.

7.2.3. Self-report ratings

Mode self-report ratings are shown in Table 1. The results of *t*-tests conducted on the ratings indicate that the ratings all differ significantly from the midpoint of the scale (3: neither agree nor disagree) for both A-only, $t(19) = 20.94$, $p < .001$, and AV speech modes, $t(19) = 33.21$, $p < .001$. A one-way ANOVA revealed that a difference between A-only ($M = 3.25$, $SD = .69$) and AV ($M = 3.59$, $SD = .48$) speech modes fell short of significance, $F(1, 38) = 3.31$, $p = .08$, partial $\eta^2 = .08$, with higher ratings recorded in the AV mode.

Four correlations were computed between ratings of engagement and shadowing RT, understanding and shadowing RT, engagement and category naming RT, and understanding and category naming RT. There were significant positive correlations between understanding and shadowing, $r = .31$, $p < .05$, and understanding and category naming, $r = .32$, $p < .05$. That is, higher ratings of understanding were associated with longer RTs while performing concurrently the shadowing and category naming primary tasks. These correlations suggest that understanding is judged to be better when more time is taken to respond.

7.3. Discussion

In Experiment 3 we replaced the talking head with a video of a human male with similar appearance to provide baseline data for the dual task evaluation paradigm. As in Experiment 2, the secondary RT task elicited the fastest RTs in the baseline condition followed by the shadowing and then category-naming tasks. The primary task conditions – shadowing and category-naming – therefore increased in cognitive demand as anticipated; performing the primary and secondary tasks concurrently slowed RT reliably in both experiments. In Experiment 3, accuracy was also significantly greater in the shadowing than the category-naming condition although unlike Experiment 2 there was no significant difference between secondary task accuracy under baseline versus concurrent shadowing conditions. RTs were significantly slower and accuracy lower in the AV speech mode compared with the A-only mode on the secondary task. There are at least three possible explanations. First, resources were directed to the primary task over the secondary task. As accuracy in both primary and secondary tasks is $>97\%$ this is unlikely to be the case although a speed–accuracy trade-off cannot be ruled out. Second and relatedly, the location of the visual target on the face, particularly on the upper part of the face, may have elicited slow RT or poor accuracy when participants fixated the mouth of the dynamic version of the human video to respond to the spoken words. de Paula et al. (2006) for example, have demonstrated that the lower half of the face and mouth region are fixated during speech perception. There is evidence for this interpretation in the visual speech advantage in both shadowing and categorisation latencies. If participants fixated the mouth

region for additional cues they may miss or be slow responding to the fly target. Third, participants in the A-only condition may have learned quickly that the face was always static and therefore diverted their visual attention to the secondary task and auditory attention to the primary task. Resources would then be available to perform the concurrent tasks accurately and efficiently (Wickens, 2002).

Accuracy rates on the shadowing condition of the primary task of 99% indicate that speech clarity was very high as in Experiment 2 and was not affected by speech mode. On the primary task, both shadowing and category-naming were significantly faster in the AV mode compared with the A-only mode. Such a result conforms with AV speech perception research has demonstrated facilitation for speech intelligibility when visual as well as acoustic cues are available (Davis and Kim, 2004; Sumby and Pollack, 1954). The seen speech advantage is explained by vision providing information about aspects of speech that are hard to hear (de Paula et al., 2006) and redundancy where acoustic and visual channels match (Campbell, 2008). It is noteworthy that an advantage for visual speech occurs in both the perception and semantic conditions of the primary task when the agent is a human. By contrast, the results of Experiment 1 reveal a visual speech advantage only for the talking head in the shadowing task and no advantage when performing concurrent tasks (Experiment 2). Task demand is a significant influence on talking head clarity. The results of Experiment 3 provide a human speaker baseline from which improved AV speech models can be compared.

Higher self-report ratings were recorded in the AV than the A-only speech mode and this accords with accuracy and speed of performance on both primary and secondary tasks. In some instances, mode ratings were slightly higher for the talking head than for the human (see Table 1) on items such as "I felt as if the person/talking head was speaking just to me". This pattern of responding likely reflects presence or relative intimacy achievable in the talking head setting and contrasting with the video situation where participants would have been aware that the video was not a live feed.

8. General discussion

A dual task paradigm has been used as a flexible laboratory method for evaluating talking head systems or component modules. An unchanging secondary RT task provided a gauge of cognitive demand imposed by concurrent tasks that draw on features of the talking head, for example, speech clarity under unimodal and multimodal conditions. In Experiment 1 a seen speech advantage was evident in latencies in the shadowing task but not in category naming or in accuracy. An explanation is that with more resources available, participants focused on the mouth region with some benefits in shadowing from the additional visual cues to the word. By contrast, in the semantic task there was no benefit from visual cues. In Experiment 2 we observed that while subjective ratings across conditions may not differ, quantitative indicators of cognitive demand do differ.

The secondary task in both Experiments 2 and 3 begins to provide benchmark data with which modifications to the present modules or new systems can be compared. Experiment 2 provides data for speech clarity of a synthetic talking head while Experiment 3 provides a human-to-human (video) baseline. The advantage in extant research for visual speech (e.g., Besle et al. 2004; Davis and Kim, 2004) has been observed in latencies in the human-to-human setting and in the synthetic talking head setting when shadowing as a single task. Accuracy in shadowing words in Experiment 2 indicates that speech clarity of the talking head was reasonable (91%); clarity of human speech in Experiment 3 was 99%. The absence of prephonatory gestures likely contributed to long shadowing latencies in experiments with the talking head. A visual speech advantage may occur in shadowing the talking head while performing a secondary task where AV speech production is based on a hidden Markov model or concatenation model.

Experiment 1 without the secondary task suggests that with resources available participants use visual cues in shadowing. In Experiment 2, dividing attention across tasks eliminated the seen speech advantage on shadowing latency. The A-only advantage in the secondary task RT is therefore likely the result of ignoring the static visual stimulus and directing visual attention to fly detection and auditory attention to shadowing (Wickens, 2002). By contrast, in Experiment 3, where the human produces speech with strong viseme–phoneme correlation, participants fixate the mouth and respond more quickly to visual speech but relatively slowly to the secondary visual target. By comparing performance under single and dual tasks, and talking head and human conditions, similar patterns of responding (i.e., the A-only RT advantage in Experiments 2 and 3), can be seen to potentially have differing causes.

Performance on the primary word task in Experiments 1 and 3 (Fig. 2) reveals that there is an interaction between task and speech mode on latencies in the single task (Fig. 2a) but separate main effects of task and speech mode in the dual task (Fig. 2c). Overall, latencies are shorter in Experiment 3 with a general visual speech advantage emerging in the faster latencies (coupled with higher accuracy) whereas such an advantage only occurs in the simpler sensory cue/shadowing task in the talking head context of Experiment 1. One could conclude from results of the single task in Experiment 1 that the talking head has good speech capabilities. In fact, the AV latency advantage over A-only may be only due to either the avatar providing a cue at the beginning of each stimulus or an increase in engagement due to multimodality. The use of a dual task allows separating real speech capabilities of the avatar from task engagement. Indeed, in Experiment 2, the AV

latency advantage disappeared whereas it is present for the human condition (Experiment 3). In the same way as a speech in noise experiment, adding cognitive load via the dual task paradigm enables a thorough investigation of effects. Unlike speech in noise experiments, the dual task paradigm can be adapted to various components, not just the speech module, of an avatar. For example, the paradigm may be used to compare human responding to two (or more) different techniques for generating visual emotion and expression, whole body gestures, faces, or eye models.

Task complexity is a useful manipulation as it influences additive or multiplicative effects of speech mode depending on a talking head or video set-up. Similarly, differences in complexity of the primary task reveal differences in accuracy when performing the secondary fly detection task, with accuracy reducing significantly while concurrently categorising in the AV mode of the video experiment, whereas there was no such effect in the talking head experiment. Recording accuracy and RT to both primary and secondary tasks can yield insights into speed–accuracy patterns and trade-offs.

Comparing the pattern of results across experiments, Fig. 3 shows that responding is generally slower during the concurrent category naming task in the human video (Experiment 3) context but this is offset by less of a difference in latencies between shadowing and categorising in human than the talking head experiment (Fig. 2b versus c). Caution is warranted when comparing across different samples and experiments but, as participants acted as their own controls in the primary tasks, we can conclude that task difficulty (shadowing versus categorising), speech mode, and the nature of the interface interact.

The measures used here are sensitive with potential for detecting performance differences not evident in single tasks and/or where the goal of the study is explicit. This is crucial if avatars are to be used in operational environments such as transport or security where user attention is divided across multiple tasks and maximum cognitive resources are required when an unexpected event or error occurs. For example, processing a spoken direction while monitoring a rail or traffic network or scanning CCTV feeds. In settings such as language learning and math tutoring the load of processing AV versus A-only speech could also be evaluated using a secondary task. Where emotional expression or posture models of virtual agents have been created (e.g., Courgeon et al., 2011; Ruttkay and Pelachaud, 2005) a secondary task could quantify the load imposed by different versions of the model or the avatars in degraded or visually complex (e.g., multi-agent) settings.

Context makes a difference (Weiss et al., 2010). For example, lower ratings for understanding and higher ratings of intimacy were assigned to the talking head in the more demanding dual task than in the single task. The absence of a significant effect of speech mode on ratings in Experiments 1 and 2 counters the possibility that participants were disappointed by the presence of a static face. The speech mode variable was also a between subjects factor minimising demand characteristics that may occur where different versions of a system are presented to the same participants, they may deduce the hypothesis and rate the stimuli accordingly.

While minimising demand characteristics (Dell et al., 2012; Orne, 1962) by not disclosing hypotheses to participants, the dual task also yields a range of human performance data. Behavioural responses, accuracy and RT, recorded from the secondary task provide objective indicators of the ease of talking head processing that is independent of explicit and introspective self-reports of talking head intelligibility. Future experiments could record other relevant measures of cognitive processing, cognitive load or stress such as eye fixations, saccades, skin conductance, or heart rate variability.

Limitations of the study include the use of a single talking head and speech model; the reliance on a visual target in the secondary task; and, to maintain statistical power, limiting the sample to female participants. In future, improved or degraded systems could be compared; another stimulus modality (e.g., vibrotactile) used as the secondary task target; and interaction between participant and talking head gender investigated. Finally, a more interactive task could be developed to extend the dual task paradigm to ECAs.

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 talking head models. The video-based experiment enables analysis of the relative contribution and strengths of the auditory, visual and integrated AV systems of a talking head and future manifestations could cross, for example, face, voice, or emotion systems to systematically evaluate modules either in isolation or interacting with other modules or modalities. Furthermore, the word task can be developed for more interactive and/or realistic scenarios where a talking head may be trialled, for example, recording the medical history of a user (i.e., person responds ‘‘yes’’ or ‘‘no’’ to having had particular medical conditions) or having users respond to items from a museum exhibit, math or language learning situation, railway monitoring, or transport booking scenario.

Acknowledgements

This research was supported by the Thinking Head project, a Special Initiative scheme of the Australian Research Council and the National Health and Medical Research Council (TS0669874). We thank Stelarc, Damith Herath for coordinating the programming of the experiment, Staci Parlato-Harris for assistance with data analysis, and three reviewers for helpful comments.

Appendix A. Word stimuli organised according to superordinate category

Animal	Cooking	Seascape
Ape	Bacon	Anchor
Beaver	Bean	Breeze
Camel	Bun	Coral
Cattle	Cabbage	Cruiser
Cheetah	Carrot	Dive
Cobra	Cracker	Dock
Deer	Curry	Drift
Donkey	Garlic	Ferry
Eagle	Gravy	Float
Goat	Ham	Flood
Hamster	Jam	Harbour
Hawk	Lemon	Horizon
Hen	Loaf	Hull
Kitten	Meatball	Lighthouse
Leopard	Muffin	Marina
Lizard	Mustard	Naval
Mule	Noodle	Pier
Owl	Olive	Pirate
Ox	Onion	Raft
Panda	Pancake	Reef
Panther	Pasta	Ridge
Parrot	Peanut	Rudder
Peacock	Pepper	Sail
Pigeon	Pies	Shell
Pony	Plum	Starboard
Pup	Popcorn	Stream
Sheep	Pumpkin	Surf
Sparrow	Sausage	Tide
Squirrel	Stew	Voyage
Zebra	Veal	Yacht

References

Bailly, G., 2003. Close shadowing natural versus synthetic speech. International Journal of Speech Technology 6, 11–19.

Bailly, G., Gibert, G., Odisio, M., 2002. Evaluation of movement generation systems using the point-light technique. In: Proceedings of the IEEE Workshop on Speech Synthesis. IEEE. Santa Monica, CA, pp. 27–30.

Bailly, G., Raidt, S., Elisei, F., 2010. Gaze, conversational agents and face-to-face communication. Speech Communication 52, 598–612.

Benoit, C., Grice, M., Hazan, V., 1996. The SUS test: a method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences. Speech Communication 18, 381–392.

Besle, J., Fort, A., Delpuech, C., Giard, M.-H., 2004. Bimodal speech: early suppressive effects in human auditory cortex. European Journal of Neuroscience 20, 2225–2234.

Brysbaert, M., New, B., 2009. Moving beyond Kučera and Francis: a critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. Behavior Research Methods 41, 977–990.

Buisine, S., Abrillian, S., Martin, J.-C., 2004. Evaluation of multimodal behaviour of embodied agents. In: Ruttkay, Z., Pelachaud, C. (Eds.), From Brows to Trust: Evaluating Embodied Conversational Agents. Springer, The Netherlands, pp. 217–238.

Burnham, D., Abrahamyan, A., Cavedon, L., Davis, C., Hodgins, A., Kim, J., Kroos, C., Kuratate, T., Lewis, T., Luerssen, M., Paine, G., Powers, D., Riley, M., Stelarc, Stevens, K., 2008. From talking to thinking heads: report 2008. In: Proceedings of the International Conference on Auditory–Visual Speech Processing (AVSP). AVSP. Queensland, Australia.

Campana, E., Tanenhaus, M.K., Allen, J.F., Remington, R., 2010. Natural discourse reference generation reduces cognitive load in spoken systems. Natural Language Engineering 17, 311–329.

Campbell, R., 2008. The processing of audio-visual speech: empirical and neural bases. Philosophical Transactions of the Royal Society B 363, 1001–1010.

Courgeon, M., Clavel, C., Tan, N., Martin, J.-C., 2011. Front view vs. side view of facial and postural expressions of emotions in a virtual character. Journal Transactions on Edutainment, VI, LNCS 6758, 132–143 Springer.

Davis, C., Kim, J., 2004. Audio-visual interactions with intact clearly audible speech. The Quarterly Journal of Experimental Psychology 57A, 1103–1121.

Dell, N., Vaidyanathan, V., Medhi, I., Cutrell, E., Thies, W., 2012. "Yours is better!" Participant response bias in HCI. In: Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI '12). ACM Press, New York, pp. 1321–1330.

de Paula, H., Yehia, H.C., Shiller, D., Jozan, G., Munhall, K.G., Vatikiotis-Bateson, E., 2006. Analysis of audiovisual speech intelligibility based on spatial and temporal filtering of visible speech information. In: Harrington, J., Tabain, M. (Eds.), Speech Production: Models, Phonetic Processes, and Techniques. Psychology Press, New York, pp. 135–147.

Fisk, A.D., Derrick, W.L., Schneider, W., 1986/7. A methodological assessment and evaluation of dual-task paradigms. Current Psychological Research & Reviews 5, 315–327.

Geiger, G., Ezzat, T., Poggio, T., 2003. Perceptual Evaluation of Video-Realistic Speech. Massachusetts Institute of Technology, Cambridge, MA (CBCL Memo 224).

Gibert, G., Fordyce, A., Stevens, C.J., 2010. Role of form and motion information in auditory-visual speech perception of McGurk combination and fusion stimuli. In: Proceedings of the 9th International Conference on Auditory-Visual Speech Processing—AVSP2010. Hakone, Japan.

Ito, K., Speer, S.R., 2006. Immediate Effects of Intonational Prominence in a Visual Search Task. Speech Prosody. ISCA, Dresden, Germany.

Johnston, W.A., Heinz, S.P., 1978. Flexibility and capacity demands of attention. Journal of Experimental Psychology: General 107, 420–435.

Kahneman, D., 1973. Attention and Effort. Prentice-Hall, Englewood Cliffs, NJ.

Karatekin, C., Couperus, J.W., Marcus, D.J., 2004. Attention allocation in the dual-task paradigm as measured through behavioral and psychophysiological responses. Psychophysiology 41, 175–185.

Kim, J., Davis, C., 2004. Investigating the audio-visual speech detection advantage. Speech Communication 44, 19–30.

Mattheyses, W., Latacz, L., Verhelst, W., 2009. On the importance of audiovisual coherence for the perceived quality of synthesized speech signals. EURASIP Journal on Audio, Speech, and Music Processing 2009, 169819. http://dx.doi.org/10.1155/2009/169819.

Meyer, D.E., Kieras, D.E., 1997. A computational theory of executive cognitive processes and multiple task performance: Part 1. Basic mechanisms. Psychological Review 104, 3–65.

Orne, M.T., 1962. One the social psychology of the psychological experiment: with particular reference to demand characteristics and their implications. American Psychologist 17, 776–783.

Ouni, S., Cohen, M.M., Ishak, H., Massaro, D.W., 2007. Visual contribution to speech perception: measuring the intelligibility of animated talking heads. EURASIP Journal on Audio, Speech, and Music Processing 2007, 047891. http://dx.doi.org/10.1155/2007/47891.

Pandzic, I., Ostermann, J., Millen, D., 1999. Users evaluation: synthetic talking faces for interactive services. The Visual Computer 15, 330–340.

Pashler, H., Johnston, J.C., 1998. Attentional limitations in dual-task performance. In: Pashler, H. (Ed.), Attention. Psychology Press, East Sussex, UK, pp. 155–189.

Ruttkay, Z., Pelachaud, C. (Eds.), 2005. Kluwer Academic Publishers, Dordrecht.

Salvucci, D.D., Taatgen, N.A., 2008. Threaded cognition: an integrated theory of concurrent multitasking. Psychological Review 115, 101–130.

Stevens, C., Walker, G., Boyer, M., Gallagher, M., 2007. Severe tinnitus and its effect on selective and divided attention. International Journal of Audiology 46, 208–216.

Sumby, W.H., Pollack, I., 1954. Visual contribution to speech intelligibility in noise. Journal of the Acoustical Society of America 26, 212–215.

Theobald, B.-J., Fagel, S., Bailly, G., Elisei, F., 2008. LIPS2008: visual speech synthesis challenge. In: Proceedings of Interspeech 2008. Interspeech, Brisbane, Australia, pp. 2310–2313.

van Wassenhove, V., Grant, K.W., Poeppel, D., 2005. Visual speech speeds up the neural processing of auditory speech. Proceedings of the National Academy of Sciences 102, 1181–1186.

Weiss, B., Kühnel, C., Wechsung, I., Fagel, S., Möller, S., 2010. Quality of talking heads in different interaction and media contexts. Speech Communication 52, 481–492.

Weiss, B., Möller, S., Wechsung, I., Kühnel, C., 2011. Quality of experiencing multi-modal interaction. In: Minker, W., Lee, G.G., Nakamura, S., Mariani, J. (Eds.), Spoken Dialogue Systems Technology and Design. Springer, New York, pp. 213–230.

Wickens, C.D., 2002. Multiple resources and performance prediction. Theoretical Issues in Ergonomic Science 3, 159–177.