Int. J. Human-Computer Studies 109 (2018) 26–40

Contents lists available at ScienceDirect

International Journal of Human-Computer Studies

journal homepage: www.elsevier.com/locate/ijhcs

Evaluating a 3-D virtual talking head on pronunciation learning

Xiaolan Peng[a], Hui Chen[a,d,*], Lan Wang[b], Hongan Wang[a,c,d]

[a] Beijing Key Lab of Human-Computer Interaction, Institute of Software, Chinese Academy of Sciences, Building 5, No. 4, South Fourth Street, Zhong Guan Cun, Beijing 100190, PR China

[b] CAS Key Laboratory of Human-Machine Intelligence-Synergy Systems, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen University Town, 1068 Xueyuan Avenue, Shenzhen 518055, PR China

[c] State Key Laboratory of Computer Science, Institute of Software, Chinese Academy of Sciences, Building 5, No.4, South Fourth Street, Zhong Guan Cun, Beijing 100190, PR China

[d] University of Chinese Academy of Sciences, 100049, PR China

ARTICLE INFO

Keywords:
Evaluation
3-D talking head
Computer-aided pronunciation training

ABSTRACT

We evaluate a 3-D virtual talking head on non-native Mandarin speaker pronunciation learning under three language presentation conditions – audio only (AU), human face video (HF) and audio-visual animation of a three-dimensional talking head (3-D). An auto language tutor (ALT) configured with AU, HF and 3-D is developed as the computer-aided pronunciation training system. We apply both subjective and objective methods to study user acceptance of the 3-D talking head, user comparative impressions and pronunciation performance under different conditions. The subjective ratings show that the 3-D talking head achieved a high level of user acceptance, and both 3-D and HF were preferred to AU. The objective pronunciation learning improvements show that 3-D was more beneficial than AU with respect to blade-alveolar, blade-palatal, lingua-palatal, open-mouth, open-mouth(-i) and round-mouth. Learning with 3-D was better than learning with HF with respect to blade-alveolar, lingua-palatal and round-mouth, and the tones of falling-rising and falling. Learning with AU was better than learning with HF with respect to the falling-rising tone. Neither HF nor AU was superior to 3-D with respect to any of the initials, finals and tones.

1. Introduction

Learning how to pronounce Mandarin properly presents a number of challenges. Non-native Mandarin learners must practice sounds and tones that do not exist in their own language or are typically different from "European/Western" languages. Proper instruction is needed for Mandarin learning. Traditional second language learning methods depend largely upon printed text, audio and video materials (Kim and Gilman, 2008). Recent developments in teaching Mandarin visual speech through animated talking heads provide an appropriate means of facilitating language learning (Chen and Massaro, 2011; Liu et al., 2013).

Animated talking heads have been applied to computer aided language learning (CALL) as carriers of audio-visual speech (Hazan et al., 2005; Wang et al., 2012a), and they have been expected to enhance the computer aided language learning system by instructing the learners with visualized pronunciation animations. In particular, both internal and external articulator movements have been studied in a three-dimensional talking head to refine instruction for hearing-loss children or second-language learners (Badin et al., 2010; Gibert et al., 2015; Grauwinkel et al., 2007; Wang et al., 2012a).

Multimodal presentation conditions promote effective learning because humans process information through both visual and verbal channels (Mayer, 2009; Sweller et al., 1998). It is promising that a three-dimensional talking head-embedded computer-aided language learning system could promote learning through multimodal interaction. The three-dimensional talking head provides language learners audio-visual and face-to-face instruction. Although several studies have focused on the implementation of a three-dimensional talking head, few experiments have been performed to evaluate the effectiveness of the talking head on pronunciation learning (Theobald et al., 2008). There is a lack of comprehensive understanding of whether a three-dimensional talking head can efficiently enhance language learning and effectively instruct language learners.

[*] Corresponding author.
E-mail addresses: xiaolan@iscas.ac.cn (X. Peng), chenhui@iscas.ac.cn (H. Chen), lan.wang@siat.ac.cn (L. Wang), hongan@iscas.ac.cn (H. Wang).

http://dx.doi.org/10.1016/j.ijhcs.2017.08.001
Received 31 May 2017; Received in revised form 7 August 2017; Accepted 9 August 2017
Available online 10 August 2017
1071-5819/© 2017 Elsevier Ltd. All rights reserved.

1.1. Articulation in Mandarin

Mandarin is the standard Chinese spoken across most of northern and southwestern China. There are 900 million Chinese native speakers, which is greater than the number of native speakers of any other language in the world (Lewis et al., 2015). Each Mandarin character is spoken as one syllable, consisting of an initial and a final, encoded by a tone (Chen et al., 2013). Mandarin has a total of 21 initials and 39 finals that can be combined together to create more than 400 sounds. Most existing studies on visual speech have been performed on "European/Western" languages, particularly English (Chen and Massaro, 2011). Compared with English, Mandarin has some typical characteristics. For example, some Mandarin sounds do not typically appear in English, including the initials of blade-palatal and lingua-palatal, and the finals of close-mouth and round-mouth (Chen and Massaro, 2011). Moreover, the Mandarin supradental, blade-alveolar, and lingua-palatal are produced primarily with the tongue tip in constrast with those English sounds produced primarily with the tongue blade (Lee and Zee, 2003). Moreover, the perception and production of lexical tones are particularly difficult when learning Mandarin (Chen et al., 2013; Chiu et al., 2009). Chen and Massaro (2011) suggested that studying and applying Mandarin visual speech information presents contributions to segmental and tonal aspects of visual speech.

Many studies on Mandarin visual speech have been conducted from the perspectives of computer science and computer engineering (Chen et al., 2005; Pei and Zha, 2006; 2007; Wang et al., 2003; Wu et al., 2006; Zhou and Wang, 2007). However, very few evaluation works that specifically attempt to study the language-learning effectiveness of Mandarin visual speech have been published. In a study that evaluated synthetic and natural Mandarin visual speech, Chen and Massaro (2011) compared participants' visual speech perception responses and then improved the quality of the synthetic Mandarin consonants, vowels, and whole syllables conveyed by an animated talking head. To date, there remains a lack of evaluation of language learners' production of Mandarin initials, finals and tones after learning with a three-dimensional talking head.

1.2. Language learning with talking heads

Audio-visual articulatory instructions conveyed by talking heads are beneficial to language learning (Engwall, 2008; Fagel and Madany, 2008). Fagel and Madany (2008) used a 3-D virtual talking head with visualized articulators to train the German pronunciations of /s/ and /z/ for eight children with speech disorders, in which the children's pronunciations were recorded and scored manually. The results showed that six children could significantly enhance their speech production of the /s,z/ sound. Engwall (2008) used an animated virtual teacher to teach seven French subjects to pronounce nine Swedish words in a 5–10 min training program in which the subjects' acoustic and articulatory data were collected by an ultrasound scanner and an electromagnetic tracking system. The results showed that the subjects' pronunciation improvement was achieved through mimicking the articulations indicated by the virtual teacher (Engwall, 2008).

Massaro et al. (2008) used a between-subjects design in which a talking head was shown to the subjects with different presentation conditions, including audio, audio-visual, frontal and inside views of the vocal tract. It was demonstrated that visible speech contributed positively to the acquisition of new speech distinctions. In recent work of Wang et al. (2014), two groups of Mandarin speakers were trained to learn 9 single vowels using either an auditory or audiovisual talking head. The experiment showed that the audio-visual group outperformed the auditory group in the task of immediate repetition of vowels. Liu et al. (2007) conducted an online experiment to compare language learners' performance with different conditions of a talking head, human face and voice only, showing that learners in the talking head condition outperformed those in the voice only condition with respect to improvement on Mandarin finals, whereas no significant training condition effect was found on Mandarin initials. Hamdan et al. (2015) evaluated the effects of the realism level of talking-head characters on students' pronunciation training. Four groups of students learned 20 English words from different characters, showing that the group of students learning with the 3D non-realistic animation character obtained the best performance in the pronunciation tests, followed by learning with the actual human character, the 2-D animation character and the 3D realistic animation character.

Many factors can affect users' language learning performance when using talking heads, such as imprecise articulatory movements, over-realistic appearance, limited language training materials and a short language training period, along with a lack of commonly accepted evaluation criteria or evaluation methods until now. To evaluate talking heads on Mandarin pronunciation learning, there is a need for a well-designed talking head and a proper between-subjects design. In our study, we evaluate a three-dimensional talking head on Mandarin pronunciation learning. The talking head exhibits both the external and internal articulatory movements of speaking and instructs Mandarin learners' pronunciations. We developed an auto language tutor (ALT) configured with audio only (AU), human face video (HF) and audio-visual animation of a three-dimensional talking head (3-D). Sixty-nine non-native speakers were recruited to learn 60 Mandarin syllables under three conditions (AU, HF and 3-D). Comparative results under these conditions were collected and analyzed to provide a clear insight into 3-D talking head effects on Mandarin pronunciation learning.

1.3. Evaluation methods

Subjective evaluation is required to assess synthesized talking heads in terms of both visual speech synthesis intelligibility and naturalness in the LIPS2008 Visual Speech Synthesis Challenge (Theobald et al., 2008). Mattheyses et al. (2009) obtained participant ratings of visual speech naturalness and synchrony between audio and visual tracks using the LIPS2008 visual speech synthesis challenge database. The subjective ratings of preference and humor between the synthetic and natural talkers were collected in the study of Stevens et al. (2013) to evaluate modality effects on speech understanding and cognitive load. The subjective ratings of likeability with respect to different talking faces (a standard face, a texture mapped face and a sampled-based face) have also been used to evaluate synthetic talking faces for a simple interactive real-time system that provides information about theater shows (Pandzic et al., 1999).

Objective evaluation is usually applied when subjects' language learning performance is quantitatively measured. Word selecting accuracy (Fagel and Madany, 2008; Massaro and Light, 2004), pronunciation repeating accuracy (Calka, 2011), reaction time (Bailly, 2003; Stevens et al., 2013) and pronunciation naming accuracy (Ali et al., 2015) are the common quantitative measures employed to assess pronunciation learning performance. Ali et al. (2015) examined the effects of three different multimedia presentations on 3-D talking head Mobile-Assisted-Language-Learning (MALL). The objective pre-test and post-test pronunciation naming scores were utilized to determine participants' overall performance. The results showed that the participants in the 3-D talking head with spoken text and on-screen text MALL outperformed those in the 3-D talking head with spoken text alone MALL and those with spoken text with on-screen text MALL (Ali et al., 2015).

Converging methods based on both subjective and objective data have been conducted by Stevens et al. (2013), in which a dual-task paradigm was used to investigate the relative cognitive demand of perceiving Audio-only versus Audio-Visual speech produced by a talking head. They collected the objective measures of reaction time, shadowing accuracy and latency data, along with subjective ratings of quality, enjoyment and engagement. The results showed that the Audio-Visual modality had the advantage in speech understanding but created great cognitive load. D'Mello et al. (2010) demonstrated that spoken tutorial dialogues increased learning more than typed dialogues did in a