Talking heads and pronunciation training: a review

Valerie Hazan
Speech, Hearing and Phonetic Sciences, UCL, UK
v.hazan@ucl.ac.uk

Abstract

This special session will include talks describing the use of talking heads in pronunciation training programs for second-language learners and clinical populations. This introductory talk will provide a brief review of the field.

Index Terms: audiovisual synthesis, speech training.

1. Introduction

Some 200 years ago, the French journalist Rivarol set out his vision of a mechanical talking head in every household that would teach native and foreign speakers to speak with a ‘standardised’ French accent. Rivarol could not have envisaged that these talking heads would be virtual 3D displays rather than made out of wood and leather but the general concept of an easily-accessible aid to pronunciation training is now becoming a reality. This special session will give a state-of-the-art account of current systems that use talking heads for pronunciation training. The aim of this introductory talk is to provide a review of how the field of audiovisual synthesis has developed over the past 20 years.

2. Embodied conversational agents vs talking heads with accurate articulation

Embodied conversational agents (ECA) are becoming increasingly widespread in computer applications and gaming and aim to emulate face-to-face communication (see [1] for a review). Although great effort is put into the development of features that increase the naturalness of the interaction (e.g. facial expressions, eye movements, head/hand gestures), the accuracy of the visible articulations of these ECAs is relatively poor. Although approximate articulations may be adequate for many applications, they are not acceptable in applications such as pronunciation training for second language learners or children with language, hearing or speech disorders which need to provide highly accurate articulatory information.

3. Talking heads with accurate articulation

Early prototypes of fully-synthetic talking heads capable of accurate visible articulations were developed in the early 1990s. Two notable models from that time, Massaro and Cohen’s Baldi [2] and ‘August’ and successors developed by Beskow [3], have continued to undergo developments to this day. Both are based on Parke’s early model for facial animation [4]. Ongoing modifications to these talking head models over the last 10 to 15 years have focused on the development of tongue movement, further refinement of control parameters and the addition of controls for paralinguistic information. Substantial developments in audiovisual synthesis have also emanated from work done by Bailly, Badin and colleagues at the ICP in France [e.g. 5]. The development of accurate models for tongue movement [6, 7] has been especially important for the application of talking heads as articulation trainers.

4. Talking heads for language learning

In the last ten years or so, there has been increasing focus on integrating these talking heads within applications for pronunciation training. Baldi, used for many years as a ‘virtual teacher’ to support vocabulary learning for deaf children, is integrated within software programs to train speech production and perception in deaf children [8] and second language learners. The articulation training system being developed at KTH (ARTUR) builds on the previous work described above and uses audiovisual-to-articulatory inversion to compare the user’s articulatory gestures to a ‘correct’ version [9]. Speech trainers involving 3D talking heads aimed at children with pronunciation difficulties are being developed in Germany. These and other systems will be presented at this special session.

Developers of such training programs face many challenges, as accurate and informative feedback is of paramount importance in any computer-based speech training application. Another key issue is the evaluation of the efficacy of such systems, and of the relative contribution of various components of the training (talking head, methodology, used, speech materials, etc).

5. References