Available online at www.sciencedirect.com
SCIENCE DIRECT®
Speech Communication 45 (2005) 115–137
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Training Baldi to be multilingual: A case study for an Arabic Badr

Slim Ouni *, Michael M. Cohen, Dominic W. Massaro

Perceptual Science Laboratory, University of California at Santa Cruz, CA, USA

Received 31 January 2004; received in revised form 14 October 2004; accepted 8 November 2004

Abstract

In this paper, we describe research to extend the capability of an existing talking head, Baldi, to be multilingual. We use parsimonious client/server architecture to impose autonomy in the functioning of an auditory speech module and a visual speech synthesis module. This scheme enables the implementation and the joint application of text-to-speech synthesis and facial animation in many languages simultaneously. Additional languages can be added to the system by defining a unique phoneme set and unique phoneme definitions for the visible speech for each language. The accuracy of these definitions is tested in perceptual experiments in which human observers identify auditory speech in noise presented alone or paired with the synthetic versus a comparable natural face. We illustrate the development of an Arabic talking head, Badr, and demonstrate how the empirical evaluation enabled the improvement of the visible speech synthesis from one version to another.

Keywords: Talking head; Avatar; Visible and visual speech synthesis; Text-to-speech; Auditory; Arabic; Multilingual

1. Introduction

Research during the past several decades proves that the face presents visual information during speech that supports effective communication. Movements of the lips, tongue and jaw enhance intelligibility of the acoustic stimulus, particularly when the auditory signal is noisy (Jesse et al., 2000; Sumby and Pollack, 1954). Given this important dimension of speech, our persistent goal has been to develop and evaluate an animated agent Baldi® (Fig. 1) to produce accurate visible speech (Massaro, 1998). Baldi has a promising potential to benefit virtually all individuals, but especially those with hearing problems (28,000,000 in the USA alone), including the millions of people who acquire age-related hearing loss every year

* Corresponding author. Address: LORIA, Speech Group, 615 rue du jardin botanique, 54600 Villers-lès-Nancy, France. Tel.: +33 3 83 59 20 22; fax.: +33 3 83 27 83 19.
E-mail address: slim@fuzzy.ucsc.edu (S. Ouni).

0167-6393/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2004.11.008

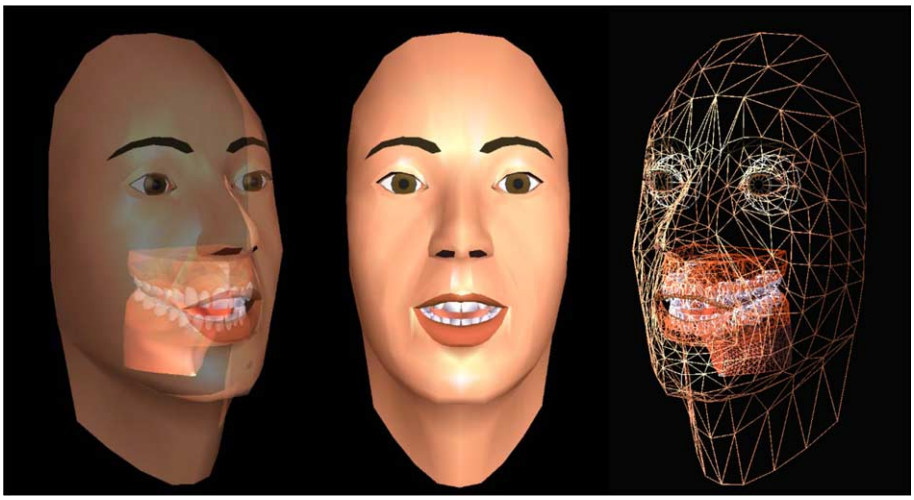

Fig. 1. Talking head Baldi. Three different views. In the middle, the standard Baldi; to the left, semi-transparent Baldi (which allows to see the inner articulation: tongue, palate and teeth); to the right, the wire frame.

(http://www.nidcd.nih.gov/health/statistics/hearing.asp), and for whom visible speech takes on increasing importance. One of many applications allows the training of individuals with hearing loss to "read" visible speech, and thus facilitate face-to-face oral communication in many situations (educational, social, work-related, etc). Baldi can also function effectively as a spoken language tutor, a reading tutor, or personal agent in human machine interaction.

For the past 10 years, the Perceptual Science Laboratory (PSL) has been improving the accuracy of visible speech produced by Baldi (e.g., Cohen et al., 2002). Research has also shown that language learning exercises featuring Baldi can improve both speech perception and production of hard of hearing children (Massaro and Light, 2004b). Baldi has also been used effectively to teach vocabulary to children with hearing loss (Barker, 2003; Massaro et al., 2003; Massaro and Light, 2004a). The same pedagogy and technology has been employed for language learning with autistic children (Bosseler and Massaro, 2003).

While Baldi's facial and tongue animation probably represent the state of the art in real-time visible speech synthesis, experiments have shown that Baldi's visible speech is not quite as effective as its human counterpart is (Massaro, 1998, Chapter 13). Preliminary observations strongly suggest that the specific segmental and prosodic characteristics are not defined optimally. One of our continual goals, therefore, is to significantly improve Baldi's communicative effectiveness. In this paper, we present our work to extend the capability of Baldi to be multilingual. We begin with an overview of facial animation and visible speech synthesis. Then, we present the general scheme to make Baldi multilingual using a client/server architecture. Finally, we present our perceptual evaluation of the Arabic version of the multilingual talking head and how this evaluation was used to improve its articulation.

2. Facial animation and visible speech synthesis

Visible speech synthesis is a sub-field of the general areas of speech synthesis and computer facial animation (Massaro, 1998, Chapter 12, organizes the representative work that has been done in this area). The goal of the visible speech synthesis at PSL has been to develop a polygon (wireframe) model with realistic motions (but not to duplicate the musculature of the face to control this mask). We call this technique, terminal analogue synthesis because its goal is to simply duplicate the observable articulation of speech production (rather than illustrate the