A Talking Head for Speech Tutoring

Priya Dey, Steve Maddock, Rod Nicolson *
University of Sheffield

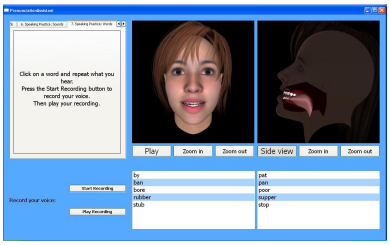

Figure 1: *Pronunciation Assistant*

1 Introduction

This work applies a viseme-driven talking head in a pronunciation training system. The aim is to create a pronunciation assistant to complement traditional methods and to assist the work of a human language tutor. Visual speech can be valuable in speech tutoring applications because vision benefits human speech perception, for three reasons as suggested by Summerfield (Summerfield, 1987): It helps speaker localization, it contains speech segmental information that supplements the audio, and it provides complementary information about the place of articulation. This study aims to elucidate the benefits of visual speech in language learning.

2 Approach

A talking head, Tara (Talking Articulation Assistant), was implemented using viseme-driven speech animation, with the creation of key mouth shapes for each speech sound. Face models for 16 visemes were created using Facegen. Loquendo TTS (Loquendo, 2008) generated acoustic speech and outputted phonetic labels and durations. Each phonetic label was mapped to a mesh for the corresponding viseme.

A coarticulation model was implemented using a dominance function to represent the influence over time of each viseme on a speech utterance, blending the dominance functions of each segment to generate a speech trajectory (Cohen and Massaro 1993). The animation frames were compared against video frames of a real person saying the same words, and the coefficients were tuned to give the closest match that could be found by observation. Principal Component Analysis was used on the viseme polygon meshes to reduce data dimensionality (Lazalde, Maddock et al 2008). Dominance functions were applied to 15 Principal Components (PCs), which were reconstructed into meshes during the generation of frames for animation. This reduced the computation time because the dominance functions were applied to only a small number of PCs instead of to every vertex of a mesh. Synchronisation between audio and video was achieved by using the audio playback loop to determine which frame to display at each time step.

The visual speech was evaluated by 32 participants in a word identification test, and was found to be more intelligible for isolated words than acoustic speech alone, and almost as intelligible as video of a real speaker under similar noise conditions. Overall the visemes were identifiable, so the talking head was determined to be sufficiently realistic to be used to demonstrate pronunciation in a tutoring system. The head was integrated into a speech tutoring application which demonstrates how to pronounce sounds at phoneme, word and sentence level, displaying the appropriate mouth movements, and displays a transverse cross-section though the head, showing the movement of internal parts such as the tongue during speech (Figure 1).

*e-mail: {p.dey, s.maddock, r.nicolson} @sheffield.ac.uk

3 Evaluation

A pilot trial was run with 5 native Arabic speakers, learning English as a second language, who worked through a session lasting one hour with a repeat session one week later. 3 were presented with the complete software, and 2 were presented with the software with no talking head. A pre-test and post-test of pronunciation were carried out, in which the subjects read aloud isolated words and sentences in English and their speech was recorded. A listening test was also carried out, in which the participants listened to acoustic speech of isolated words and identified which words they heard. A human judge evaluated the pronunciation tests, and the number of correct pronunciations were counted to give an overall score.

Generally, there was an improvement in speaking and listening, from the first test to the final test, for both groups. Overall the talking head gave a more consistent improvement than audio alone. Further studies will compare the effects of various aspects of the animation of the talking head, such as the impact of more natural facial expressions. Future tests will be carried out on larger groups of participants, to determine whether the use of talking heads can be of benefit in learning pronunciation.

References

COHEN, M. M., AND MASSARO., D. W. 1993. Modelling coarticulation in synthetic visual speech. In *Models and Techniques in Computer Animation*, N. M. Thalmann and D. Thalmann, Eds. Springer-Verlag, 139–156.

LAZALDE, O. M., MADDOCK, S., AND MEREDITH, M. 2008. A constraint-based approach to visual speech for a Mexican-Spanish talking head. *International Journal of Computer Games Technology*, 3.

LOQUENDO, 2008. Loquendo. //www.loquendo.com.

SUMMERFIELD, A. 1987. Some preliminaries to a comprehensive account of audio-visual speech perception. In *Hearing by Eye: The Psychology of Lip-Reading*, Dodd, B., and R. Campbell, Eds. Lawrence Erlbaum Associates, 3–51.