Automatic Lip Synchronization by Speech Signal Analysis

Goranka Zoric, Aleksandra Cerekovic, Igor S. Pandzic

Department of Telecommunications, Faculty of Electrical Engineering and Computing

{Goranka.Zoric, Aleksandra.Cerekovic, Igor.Pandzic}@fer.hr

Abstract

In this paper a system for the automatic lip synchronization of virtual 3D human based only on the speech input is described. The speech signal is classified into viseme classes using neural networks. Visual representation of phonemes, visemes, defined in MPEG-4 FA, is used for face synthesis.

Index Terms: lip syncing, facial animation, MPEG-4 FBA, virtual characters, speech processing, neural networks

1. Introduction

A human speech consists of both auditory and visual component. By incorporating both components into the human-computer interaction interfaces, we achieve a lot on the naturalness. Our goal is create a system for automatic lip synchronization in the real-time capable of generating an animation of 3D human face model only by analyzing a speech signal in the real time. Such system may find applications from film production and advertising to games, teleconferencing, messaging, news delivery and in advanced user interfaces (i.e. for education and commerce). Although such systems already exist, most often they do not work in the real time, what might be essential for some applications.

2. System overview

The whole process is completely automatic and starts from the speech signal. The system must analyse an audio signal containing speech and classify it into lip shape categories (visemes) in order to synchronize the lips of a computer generated face with the speech. Our system for automatic lip synchronization is suitable for real-time and offline applications. It is speaker independent and multilingual. Visemes, defined in MPEG-4 FA, is used for face synthesis. Database used for viseme classification is not audio-visual but auditory only. The initial version of this lip sync system has been implemented in [1] and in the collaboration with [2].

3. Conclusions

Our lip synchronization system works in real-time; it is language independent and easy to use. In order to achieve more precise results phoneme database might be bigger. Adding facial gestures remains important issue in creating believable virtual humans, which is in fact our current research challenge.

4. Acknowledgements

The work was partly carried out within the research project "Embodied Conversational Agents as interface for networked and mobile services" supported by the Ministry of Science, Education and Sports of the Republic of Croatia.

5. References