Available online at www.sciencedirect.com
ScienceDirect
Signal Processing 86 (2006) 3644–3656
SIGNAL PROCESSING
www.elsevier.com/locate/sigpro

Real-time language independent lip synchronization method using a genetic algorithm

Goranka Zorić*, Igor S. Pandžić

Department of Telecommunications, Faculty of Electrical Engineering and Computing, University of Zagreb, Unska 3, HR-10000 Zagreb, Croatia

Received 1 July 2005; received in revised form 5 December 2005; accepted 1 February 2006
Available online 24 May 2006

Abstract

Lip synchronization is a method for the determination of the mouth and tongue motion during a speech. It is widely used in multimedia productions, and real time implementation is opening application possibilities in multimodal interfaces. We present an implementation of real time, language independent lip synchronization based on the classification of the speech signal, represented by MFCC vectors, into visemes using neural networks (NNs). Our implementation improves real time lip synchronization by using a genetic algorithm for obtaining a near optimal NN topology. The automatic NN configuration with genetic algorithms eliminates the need for tedious manual NN design by trial and error and considerably improves the viseme classification results. Moreover, by the direct usage of visemes as the basic unit of the classification, computation overhead is reduced, since only visemes are used for the animation of the face. The results are obtained in comprehensive validation of the system using three different evaluation methods, two objective and one subjective. The obtained results indicate very good lip synchronization quality in real time conditions and for different languages, making the method suitable for a wide range of applications.

Keywords: Lip synchronization; Lip sync; Facial animation; MPEG-4 FBA; Human–computer interaction; Virtual characters; Speech processing; Neural networks; Genetic algorithms

1. Introduction

A human speech is bimodal in its nature [1]. A speech that is perceived by a person depends not only on the acoustic information, but also on the visual information such as lip movements or facial expressions. In noisy environments, a visual component of a speech can compensate for a possible loss in speech signal. This combination of the auditory and visual speech recognition is more accurate than only auditory or only visual. Use of multiple sources generally enhances a speech perception and understanding. Consequently, there has been a large amount of research on incorporating bimodality of a speech into the human–computer interaction interfaces. Lip synchronization is one of the research topics in this area.

The goal is to animate the face of a speaking avatar (i.e. a synthetic 3D human face) in such a way that it realistically pronounces the given text, which is based only on the speech input. Especially

*Corresponding author. Tel.: +385 1 6129 801; fax: +385 1 6129 832.
E-mail addresses: Goranka.Zoric@fer.hr (G. Zorić), Igor.Pandzic@fer.hr (I.S. Pandžić).

0165-1684/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.sigpro.2006.02.038

important component of facial animation is the movement of lips and the tongue during speech. For a realistic result, lip movements must be perfectly synchronized with the audio. However, in the real time use, some time delay must be accepted, since a speech has to be spoken before it can be classified.

In this section the problem of the lip synchronization is introduced. Next section gives the background information and related work. Section 3 explains the proposed lip synchronization algorithm. Implementation of our system is briefly described in Section 4, while achieved results and description of the system behaviour in different conditions is presented in Section 5. The paper closes with the conclusion and the discussion of the future work.

2. Background

Lip synchronization is the determination of the motion of the mouth and tongue during a speech [2]. Intonation characteristics, a pitch, an amplitude and voiced/whispered quality, are dependent on the sound source, while the vocal tract determines the phoneme. A phoneme is the basic unit of the acoustic speech. A visual representation of the phoneme is called viseme. There are many sounds that are visually ambiguous when pronounced. Therefore, there is a many-to-one mapping between phonemes and visemes. To make lip sync possible, position of the mouth and tongue must be related to characteristics of the speech signal. Positions of the mouth and tongue are functions of the phoneme and are independent of intonation characteristics of a speech.

The basic idea of lip synchronization is shown in Fig. 1. The process of the automatic lip sync consists of two main parts. The first one, audio to visual mapping, or more specific speech to lip shape mapping, is the key issue in the bimodal speech processing. In this first phase a speech is analysed and classified into viseme categories. In the second part, calculated visemes are then used for the animation of virtual character's face. The animation is not the topic of the interest in this work as it is already implemented in the Visage Technologies [3] software on which our application is based, so it is only briefly described in this paper.

The problem of converting a speech signal to the lip shape information can be solved on several different levels, depending on the speech analysis that is being used [4]. These levels are:

- Front end (signal level)
- Acoustic model (phoneme level)
- Language model (word level)

Each of the three levels can be applied within the speech-driven face animation system. However, the choice will depend on a specific application, considering characteristics of the individual solution. In addition, a balance between time needed for the signal processing and the quality to be achieved must be found.

A signal level concentrates on a physical relationship between the shape of the vocal tract and the sound that is produced. The speech signal is segmented into frames. A mapping is then performed from acoustic to visual feature, frame by frame. This method uses a large set of audio-visual parameters to train the mapping. There are many algorithms that can be modified to perform such mapping—Vector Quantization (VQ), the Neural Networks (NN), the Gaussian Mixture Model (GMM), etc.

At the second level, speech is observed as a linguistic entity. The speech is first segmented into a sequence of phonemes. Mapping is then found for each phoneme in the speech signal using a lookup table, which contains one visual feature set for each phoneme. The standard set of visemes is specified in

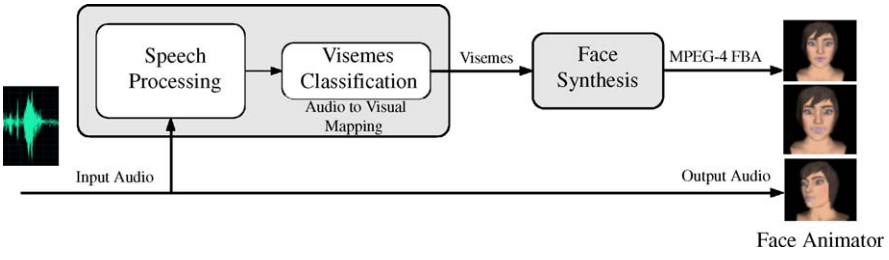

Fig. 1. The basic idea of lip sync.