Improving Naturalness of Visual Speech Synthesis

László Czap, János Mátyás

Department of Automation and Communication-Technology,
University of Miskolc, Miskolc, Hungary
czap@mazsola.iit.uni-miskolc.hu

Abstract
Facial animation has progressed significantly over the past few years and a variety of algorithms and techniques now make it possible to create highly realistic characters. Based on the author’s visual feature database for speechreading and the development of 3D modelling, a Hungarian talking head has been created. Our general approach is to use both static and dynamic observations of natural speech to guide the facial animation. A three level dominance model has been introduced that takes co-articulation into account. Each articulation feature has been grouped to dominant, flexible or uncertain classes. Analysis of the standard deviation and the trajectory of features served the evaluation process. The acoustic speech and the articulation are linked to each other by a synchronising process. Natural head movements, eyebrow rising, blinking and expressing emotions are demonstrated.

Index Terms: AV speech synthesis, improving naturalness, expressing emotions

1. Introduction

Visible features are meaningful gestures, especially hard of hearing people can read it. In dialogs gestures can support the turn taking, the lift of eyebrows can indicate paying attention, nodding can show acknowledgement. Studying the head movements of professional speakers, moderate nodding, tilting and blinking are introduced. Series of subjective tests served the fine tuning of parameters.

2. Improving naturalness

There are a couple of features added to improve the naturalness of articulation:

1. Pre-articulation. Prior to utterance a silence period is inserted – imitating breathing by opening the mouth – then the first dominant viseme is progressed from the neutral starting position.

2. A filtering and smoothing algorithm has been developed for adaptation to the tempo of either the synthesized or natural speech and for realizing the temporal asynchrony effect.

3. Head movement, gaze, eyebrow rising, and eye blink can be controlled semi-randomly and manually.

4. Basic emotions defined by Ekman can be expressed in a scalable and blended manner.

2.1. Pre-Articulation

Prior to utterance there is an about 300 ms silence period inserted – imitating breathing by opening the mouth – then the first dominant viseme is progressed from the neutral starting position. By this pre-articulation the mouth is formed beforehand the sound is emitted in like manner as natural speech. If the last sound of the sentence is bilabial, the mouth is opened slightly after the sound fade (post-articulation).

2.2. Adapting to the Tempo of Speech and Filtering

During the synchronization to human or synthesized speech we have faced different tempo of speech. When the speech is slow viseme features approach their nominal value, while fast speech is articulated roughly in natural speech. For flexible features the round off is stronger in fast speech. A median filter is applied for interpolation of flexible features: the values of neighbouring frames are sorted and the median is chosen. A feature is formed by the following steps:

- linear interpolation among values of dominant and flexible features neglecting the uncertain ones
- in the neighbouring of flexible features median filtering is performed
- these values are then filtered by the weighted sum of the two previous frames, the actual and the next one

The weights of the filter are fixed, not depending on the speech tempo. The smoothing filter refines the movements and reduces the peaks for fast speech.

2.3. Facial Gestures

Head movement, gaze, eyebrow rising, and eye blink can be controlled semi-randomly or manually. Automatic generation of facial gestures is organized in a semi-random manner. Tilting-nodding head movement is related to the short time (200 ms) average energy of the acoustic speech. At sentence accent a downward head movement is observed. Imitating this, the bigger the average sound energy is, the higher the probability of a downward tilting. Moderate and slow head turning (pan) and side inclination are controlled randomly. Amplitude of these head movements is not more than 2–3°. Gaze is controlled to compensate the head movements to keep looking into the camera (the observer’s eyes).

According to our observations, blinking is occurred in about 1.5 to 3 second period of time. Blinking is controlled semi-randomly. Higher average energy makes the blinking more frequent. Probability of blinking is increased at long vowels and the first vowel of the word, as in Hungarian word-stress is always put on the first syllable of the word.

2.4. Expressing Emotions

After Ekman, in our system the basic emotions can be selected in a scalable manner. These emotions include anger, disgust, happiness, sadness, fear, and surprise.

During the utterance of a sentence, expression of the face is progressing from a neutral look to the target display of emotion. Emotion can be controlled in a scalable and blended manner. E.g. 20&+30$ means 20% fear and 30% surprise, while 20*+30$ evolves 20% happiness and 30% surprise.