# Virtual speaker

LÁSZLÓ CZAP
*Department of Automation, University of Miskolc*
*czap@mazsola.uni-miskolc.hu*

JÁNOS MÁTYÁS
*North Hungarian Regional Training Centre, Miskolc*
*matyasj@mail.erak.hu*

***Reviewed***

***Key words: facial animation, talking head, dynamic speech features, speechreading***

*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 speechreading study 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 facial modelling. The evaluation of Hungarian consonants and vowels is presented for classifying visemes - the smallest perceptible visual units of the articulation process. A three level dominance model has been introduced that takes coarticulation into account. Each articulatory feature has been grouped to dominant, flexible or uncertain classes. The analysis of the standard deviation and the trajectory of the features served the evaluation process. Acoustic speech and articulation are linked with each other by a synchronising process. A filtering and smoothing algorithm has been developed for the adaptation either to the tempo of the synthesized or natural speech.*

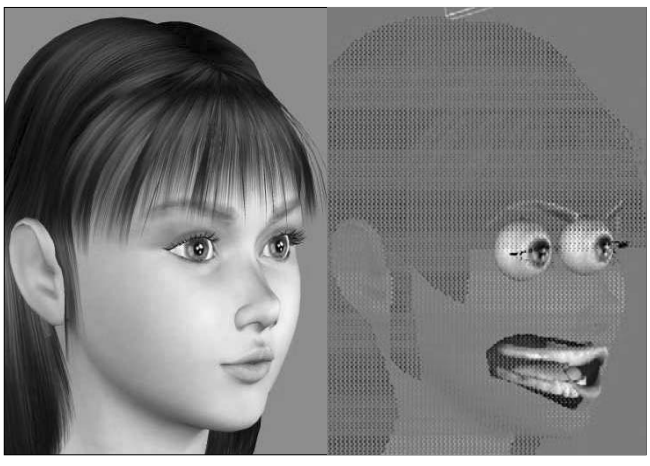
## 1. Introduction

The intelligibility of speech can be improved by showing the articulation of the speaker. This visual support is essential in noisy environment and for hearing impaired people. An artificial talking head can be a natural supplement to the sophisticated acoustic speech synthesis. The pioneer work of face animation for modelling the articulation started about two decades ago. The development of 3D body modelling, the evolution of computers and the advances at the analysis of human utterance enabled the development of realistic models.

Since the last decade the area has been developing dynamically and more and more applications have appeared. The audio-visual speech recognition and synthesis can open up a new prospect in the human-machine interface.

Virtual speakers and actors can improve the freedom of artists in multimedia applications. Teaching hearing impaired people to speak can be aided by an accurately articulating virtual speaker, which can make its face transparent and show the details of show the utterance better than a human speaker.

*Figure 1.*
*Photorealistic and transparent visualization*

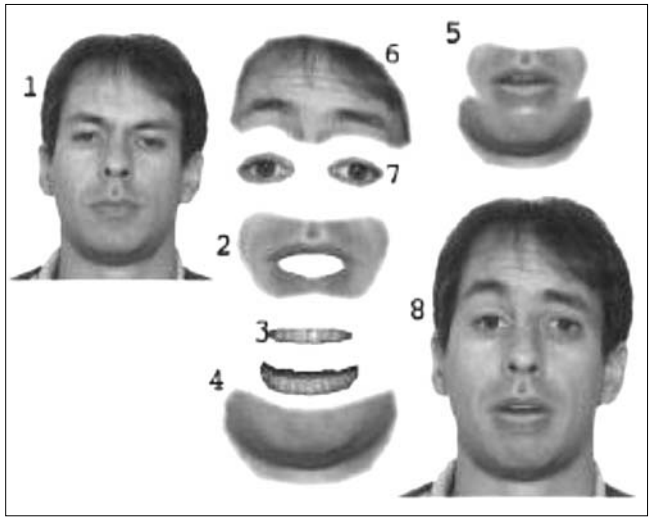


## 2. Speech animation

The first visual speech synthesizers were based on a 2D head model, recalling beforehand stored images of a speaker. Phases between keyframes sometimes were produced by image morphing. A 2D model can hardly provide head movements, gestures and emotions.

The progress at solid modelling directed the researchers' interest to the three-dimensional modelling. Either type of 3D models simulates facial expressions by tensing muscles. They produce realistic results, but the analysis of real muscular tensions is difficult. Surface

*Figure 2.*
*Elements of a 2D head model [1]*



models seem to be promising by acting textured polygons. Their features can be analysed on human speakers.

### 2.1. The visual unit of speech

The visual parallel of shortest acoustic unit, a phoneme is called *viseme*. The set of visemes has fewer elements than that of phonemes as utterances of several phonemes are visually the same. E.g. the voiced quality is invisible and the voices of the same place of articulation that are different only in duration or intensity belong to the same viseme class. The static positions of the speech organ for Hungarian phonemes can be found in essential publications. *Figure 3* shows the similarity of the same viseme on the speaker's photograph [5] and the 3D model [6].

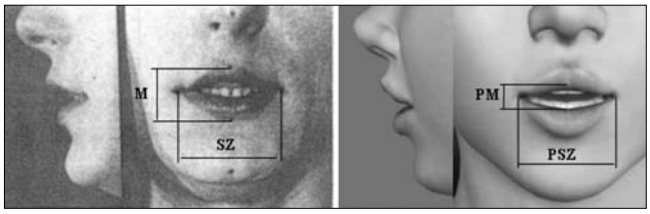



*Figure 3.*
*A photograph of a speaker and the 3D model of the same viseme.*

The features of Hungarian visemes have been created according to the word specimens of [4]. *Table 1.* shows the resultant groups of visemes represented by their Sampa codes.

| Vowels | Consonants |
|---|---|
| E | b, p, m |
| e: | f, v |
| ii | t, d, n |
| 2, o | r |
| y, u | s, z, ts, dz |
| A: | l |
| O | S, Z, tS, dZ |
| | t', d', j, J |
| | k, g |
| | h |

*Table 1. The Hungarian viseme classes*

Remarks to the grouping:

- Viseme classes are based on the lip shapes, the invisible tongue position can be different e.g. o – 2, u – y.
- The lip opening of unlisted vowels are narrower than that of their short counterparts.
- For synchronization an enlarged selection is used.

The main features of visemes can be adapted from the published sound maps [4] and albums [5,6]. These features are the foundation of keyframes that the articulation is based on [7].

Features controlling the lips and tongue are crucial. Basic lip properties are the opening and width, their rate is related to lip round. The lip opening and the visibility of teeth are referred to the jaw movement. The tongue is described by its horizontal and vertical position, its bend and the shape of the tongue tip *(Figure 4.)*.

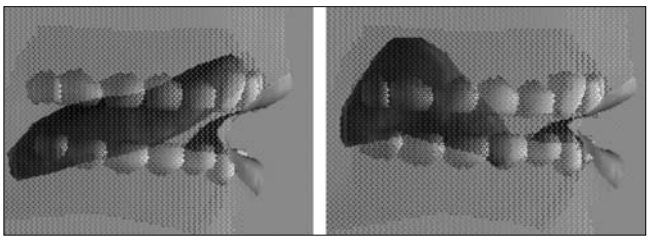


*Figure 4.*
*Illustrative tongue positions for sounds* n *(left) and* k-g *(right).*

Based upon the static features, the articulation parameters characteristic to the stationary section can be set.

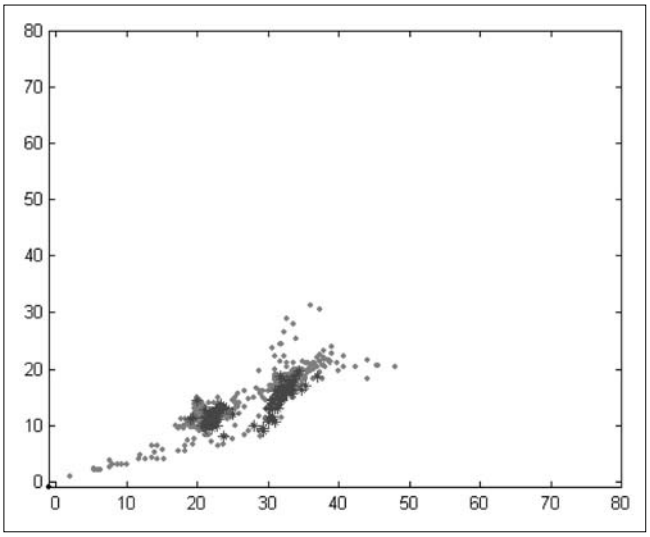
### 2.2. Dynamic operation

The dynamic features of continuous Hungarian utterances have not been described yet. The usage of motion phases represented in voice albums are limited, and can be related only to the particular word of specimen. The other source of dynamic analysis are my own studies in speechreading [8]. Trajectories of width and height of lips and the visibility of teeth and tongue are derived from there. These data drive the interpolation between the motion phases.

Some features take their characteristic value, while others do not reach their nominal value during utterance. All features of the visemes (eg. lip shape, tongue position) were classified according to their dominance. The categorization is based on the standard deviation of the speechreading data. Viseme features can be divided into three grades:

- *dominant* – coarticulation has no effect on them
- *flexible* – the neighbouring visemes affect them
- *uncertain* – the neighbourhood determines the feature.

*Figure 5.*
*Lip open and lip width of transitional (.) and stationary (*) phases of S.*



Besides the standard deviation, the distribution of transitional and stationary periods of visible features help to determine the grade of dominance. In *Figure 5.* the lip sizes of transitional and quasi stationary phases of sound *S* can be seen. Among the transitional states, determined by the neighbouring sounds, the features of the middle frames cover a restricted area.

The trajectory of viseme features can be also essential for determining the dominance classes. *Figure 6/a.* shows the trajectory of lip sizes of viseme *E*. Nevertheless, the curves cannot be traced one by one, but it is observable that they go through a dense area regardless of the starting and final states. The dominant nature of vowels' lip shape is obvious.

In contrast, the uncertain features do not tend to a certain value. The trajectory of *h* can be seen in *Figure 6/b.* (To be able to track them, only a couple of curves are represented.)

*Figure 6.*
*Trajectory of lip sizes of viseme* E *(a) and* h *(b)*

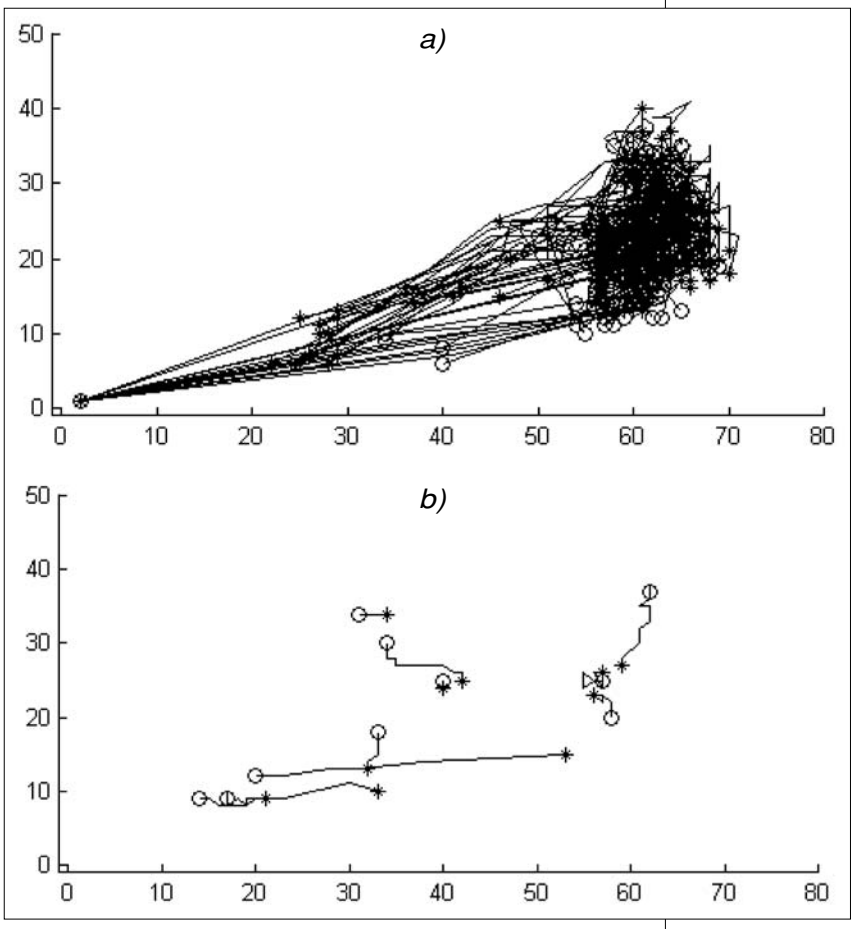



*Table 2.* describes the dominance classes of lip shapes while *Table 3.* shows that of horizontal position of tongue.

*Table 2. Dominance grades of lip shape*

| | |
|---|---|
| **Dominant** | vowels, S, Z, tS, dZ |
| **Uncertain** | k, g, r, h |
| **Mixed** | p, b, m, l, j, n, J, f, v, s, z, ts, dz,, d, t, t', d' (lip opening is dominant, lip width is uncertain) |

| | |
|---|---|
| **Dominant** | t, d, n, r, l, t', d', j, J, S, Z, tS, dZ, s, z, ts, dz |
| **Flexible** | vowels |
| **Uncertain** | p, b, m, f, v, k, g, h |

*Table 3.*
*Dominance grades of horizontal position of tongue*

The dominance grades of visemes control the interpolation of the features. Other improvements – as inserting a permanent phase into long vowels – refine the articulation.

## 3. Improving the naturalness

Studying the head movements of professional speakers, moderate nodding, tilting and blinking were introduced in a semi-random manner. Algorithm for head movement and mimicry can hardly be created according to prosody – these features are manually set by tags (e.g. lifting eyebrows at sentence accent, or control the glance). In dialog systems gestures can support the turn taking, the lift of eyebrows can indicate paying attention, nodding can mean acknowledgement.

### 3.1. Pre-articulation and filtering

Prior to utterance there is an apt. 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 before the sound is emitted in like manner as natural speech.

During the synchronization to natural or synthesized speech we were faced with different tempo of speech. When the speech is slow viseme features approach their nominal value, while fast speech is articulated roughly. For flexible features the round off is stronger in fast speech. A median filter is applied to 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.

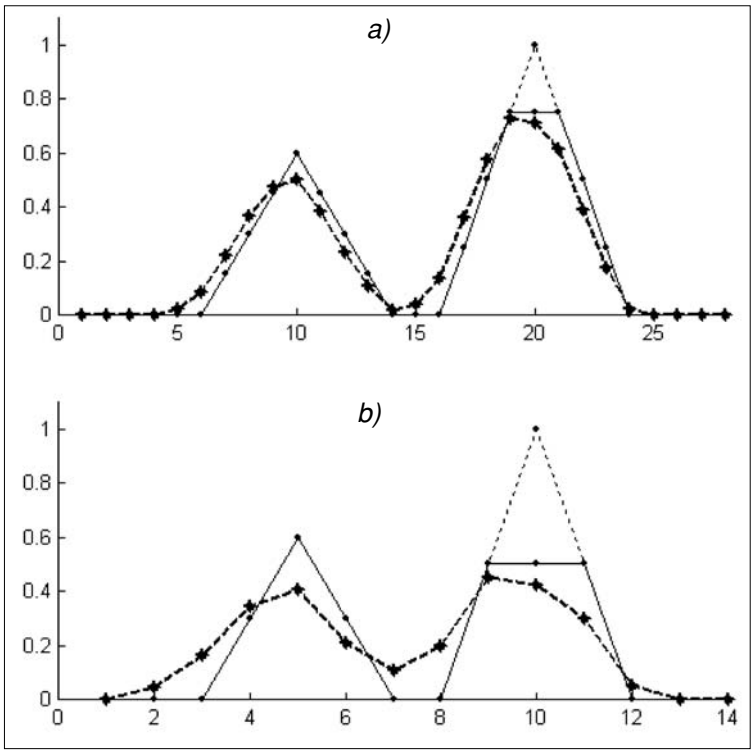



*Figure 7. The interpolation of dominant (first peak) and flexible (second peak) features for slow (a) and fast (b) speech after linear interpolation (...), median filtering (—) and smoothing (---).*

*Figure 7.* depicts the effect of median filtering and smoothing. In the example the slow speech contains double as many frames as the fast one.

### 3.2. Expressing of emotions

In multimodal speech we can confirm or disprove the verbal message by gestures and body language. After Ekman, the basic emotions can be selected in a scalable manner: anger, disgust, fear, enjoyment, sadness, surprise. *Figure 8.* depicts a couple of examples.

*Figure 8. Expression of disgust and enjoyment*

## 4. Conclusions

This paper describes the results of several years of research and development work that aim at working out a Hungarian audio-visual text-to-speech system. In this phase further refinement of co-articulation is performed. Due to the time consuming rendering process the virtual speaker can be utilized for reading pre-recorded messages. In the near future the results are going to be transferred to a real-time rendering platform.

Sample videos can be found:
http://mazsola.iit.uni-miskolc.hu/~czap/mintak

### References

[1] Cosatto E., Grafat H. P. (1998)
2D Photo-realistic Talking Head.
Computer Animation,
Philadelphia, Pennsylvania, pp.103–110.
[2] Massaro, D.W.: (1998)
Perceiving Talking Faces.
The MIT Press Cambridge,
Massachusetts, London,
England, pp.359–390.
[3] Bernstein, L.E., Auer, E.T.: (1996)
Word Recognition in Speechreading.
Speechreading by Humans and Machines.
Springer-Verlag, Berlin Heidelberg,
Germany, pp.17–26.
[4] Molnár J.: (1986)
The Map of Hungarian Sounds.
Tankönyvkiadó, Budapest.
[5] Bolla K.: (1995)
A Phonetic Conspectus of Hungarian.
Tankönyvkiadó., Budapest.
[6] Bolla K.: (1980)
Hungarian Sound Album.
MTA Nyelvtudományi Intézet, Budapest.
[7] Mátyás J.: (2003)
Visual speech synthesis. M.Sc. Thesis,
University of Miskolc.
[8] Czap, L.: (2000)
Lip Representation by Image Ellipse. ICSLP 2000
Bejging, China, Proceedings Vol. IV., pp.93–96.
[9] Ekman, P., Friesen, W.: (1978)
Facial Action Coding System.
Consulting Psychologists Press. Inc.