ESCA. EUROSPEECH '95. 4th European Conference on Speech Communication and Technology. Madrid, September 1995.

RULE-BASED VISUAL SPEECH SYNTHESIS

Jonas Beskow
beskow@speech.kth.se
Department of Speech Communication and Music Acoustics
KTH, Stockholm, Sweden

ABSTRACT

A system for rule based audiovisual text-to-speech synthesis has been created. The system is based on the KTH text-to-speech system which has been complemented with a three-dimensional parameterized model of a human face. The face can be animated in real time, synchronized with the auditory speech. The facial model is controlled by the same synthesis software as the auditory speech synthesizer. A set of rules that takes coarticulation into account has been developed. The audiovisual text-to-speech system has also been incorporated into a spoken man-machine dialogue system that is being developed at the department.

1. INTRODUCTION

The visual channel in speech communication is of great importance, as has been demonstrated by for example McGurk [6]. A view of the face can improve intelligibility of both natural and synthetic speech, especially under degraded acoustic conditions [5]. Moreover, visual signals can express emotion, add emphasis to the speech and support the interaction in a dialogue situation through e.g. turn-taking signals and back-channeling. This makes the use of a computer-synthesized face to create visual speech synthesis an important complement to traditional speech synthesis, especially in applications for hearing impaired people, in noisy environments and in speech based multimodal user interfaces.

By extending the KTH rule-based text-to-speech synthesis system [2,3] with a real-time animated 3D model of a human face, a system for audiovisual speech synthesis has been developed. The face is controlled from the same text-to-speech rule compiler that controls the auditory speech synthesizer. This provides a unified and flexible framework for development of audiovisual text-to-speech synthesis that allows the rules controlling the two modalities to be written using the same notation.

2. THE FACIAL MODEL

The face synthesis is based on a model developed by Parke [7]. The model consists of a three dimensional mesh with 800 vertices, connected together by about 800 polygons, which approximate the surface of a human face, see Figure 1. The shape of the facial surface is controlled by a set of parameters, which operates directly on the coordinates of the vertices. The topology of the polygon network remains constant.

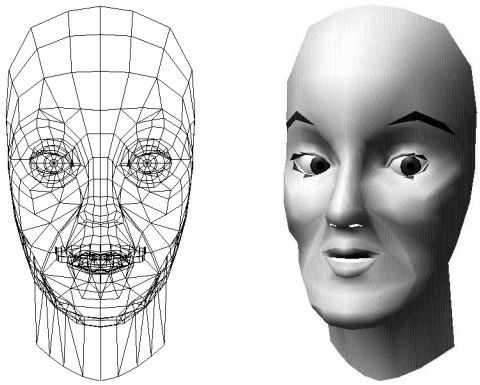


Figure 1: Wireframe and shaded representations of the face

There are about 50 parameters, which can be divided into two groups:

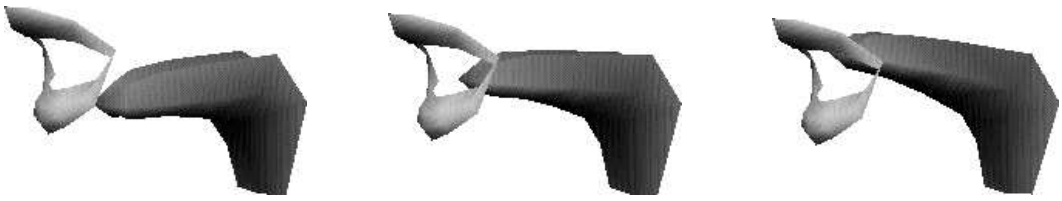
- expression parameters, that control the articulation and mimic of the face. These include jaw rotation, eyebrow shape and position, various mouth shape parameters etc.
- conformance parameters, that control static features in the face like for example nose length and jaw width.

A number of modifications have been made to the original model, in order to make it more suitable for speech synthesis. These modifications include introduction of a tongue and creation of a new set of parameters to control lip movements.

2.1. A model of the tongue

Human speech production relies heavily on tongue movements. Tongue actions are important not only auditorily but also visually. Visibility of the tongue during speech is speaker dependent, due to for example individual articulation style. Clearly visible tongue movements can however raise the intelligibility of speech. In order to allow synthesis of visual speech without omitting any potentially important visual information carrier, a tongue was modeled and added to the face.

In natural speech, the apically articulated phonemes are responsible for most visually perceivable tongue actions.



(apex = 0 %) *(apex = 50 %)* *(apex = 100 %)*

Figure 2: View of lips and tongue with different apex values

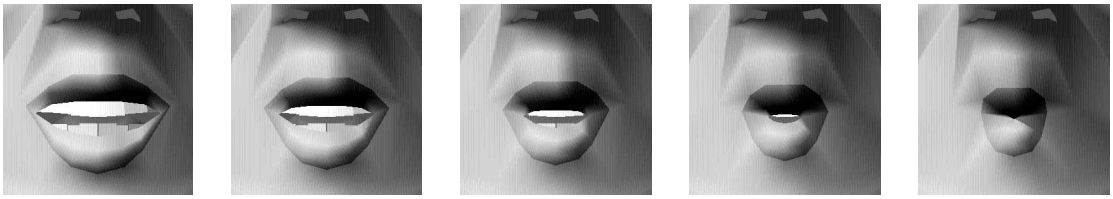


Figure 3: Lip shape for different values of the lip rounding parameter: 0, 25, 50, 75 and 100 % rounding. Jaw rotation is 5 degrees.

Thus, for the purposes of visual speech, the tongue can be modeled quite simply, by only allowing tip movement and disregarding the shape of the tongue body. The tongue was created using 64 polygons, under control of four parameters: length, width, thickness and apex. The value of the apex parameter is given in percent, and determines the vertical position of the tongue tip relative to the hard palate. A value of 100 % always corresponds to maximum elevation of the tip, regardless of the current jaw rotation. See Figure 2 for an illustration of the tongue and the apex parameter. By adjustment of the tongue length, there is a limited way of simulating retroflex articulation. A more accurate simulation of such tongue actions could be achieved by introducing a parameter dedicated to determine the place of articulation, from dental to retroflex.

2.2. Improved control of lip movements

In the original Parke-model [7], synthesis of speech-like lip movements is quite difficult. Therefore, a new set of lip control parameters was developed. The most important additions are the parameters for lip rounding, bilabial occlusion and labiodental occlusion. Like the apex, the parameters are dimensionless.

2.2.1. Lip rounding

This parameter facilitates synthesis of rounded vowels. By calculating the center of the lip opening, and moving all lip vertices towards that point, a funneled lip shape is achieved, see Figure 3.

2.2.2. Bilabial occlusion

Although bilabial occlusion can be achieved by adjusting the parameters that control the vertical positions of the lips, this is impractical, since it requires three parameters (upper lip raise, lower lip depression and jaw rotation) to be given the correct values, in order for the lips to close tight.

The new parameter works similar to the lip rounding, and allows bilabial occlusion to be controlled with only one parameter independently of any others, by pulling the lips towards each other.

2.2.3. Labiodental occlusion

One of the most distinctive visual speech actions is the labiodental occlusion, as in /f/ and /v/. This movement has also been assigned its own parameter, which pulls the lower lip towards the edge of the upper front teeth.

2.3. Implementation

The polygon surface can be rendered using illumination and interpolated (gouraud-) shading, which gives the surface a smooth appearance. The model is implemented using a standard 3D graphics library called PEX, which makes it portable to most UNIX environments.

The model can be animated at a rate of approximately 20 frames per second on a HP 715/100 workstation. This allows for real time visual speech synthesis.

3. AUDIOVISUAL SPEECH SYNTHESIS

By connecting the facial model to an existing text-to-speech system [2,3], an audiovisual speech synthesis system was created.

3.1. System description

The audiovisual text-to-speech system consists of three parts, as shown in Figure 4.

- The auditory speech synthesizer, GLOVE [3]. This is a formant based speech synthesizer, controlled by 40 parameters.
- The visual speech synthesizer, which is the facial model described in section 2 of this paper. Ten of the models fifty parameters are used in the production of visible articulatory actions (see Table 1).

Jaw rotation
Lip rounding
Lip protrusion
Mouth width
Bilabial occlusion
Labiodental occlusion
Upper lip raise
Lower lip depression
Apex
Tongue length

Table 1: Parameters used for articulatory control of the face.

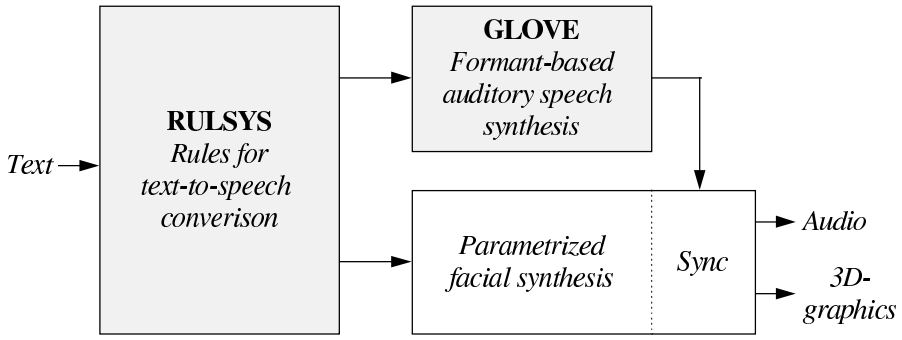


Figure 4: Schematic overview of the audiovisual text-to-speech system. The shaded parts correspond to the original text-to-speech system.

- The rule synthesis framework, RULSYS [2]. This is a generic rule synthesis system, that transform input text (orthographic or phonetic) to synthesizer control parameters. The system itself is independent of both language and synthesizer used to produce output. The language- and synthesizer-specific functionality lies entirely in the rules.

The process works as follows: RULSYS operates on the incoming text according to the morphologic, syntactic and phonetic rules, and outputs a multi channel data file. This file contains parameter values for both the auditory and the visual synthesizer, sampled every 10th millisecond.

The GLOVE synthesizer module creates an audio sample file based on the parameters in the data file. The face synthesis module then plays back the audio file, and animates the face model in real time, synchronized with the audio playback.

RULSYS controls both modalities on the parameter level. This ensures correct timing and coherence between auditory and visual representations of the synthetic speech signal.

3.2. Controlling articulation

During visual speech synthesis, the articulatory movements are controlled by the rule synthesis system (RULSYS). First, orthographic text is transformed to a phonetic string. To this point, the process is the same for both the auditory and the visual modality.

The transformation from phonetic representation into control parameters for the face model is based on visemes. From 45 Swedish phonemes, 21 visemes were created by grouping visually equivalent or similar phonemes. These visemes were then translated to parameter settings in the face model, by interactive adjustment of the parameters in Table 1.

For each viseme, each of the parameters were either assigned a value, or were left undefined. If a parameter is left undefined, it simply means that the viseme is independent of that particular parameter. (For example, /r/ can be either rounded or unrounded, depending on the context, thus the lip rounding parameter is left undefined for this viseme.) This is a simple way of accounting for coarticulation effects.

When each segment in the phonetic string has been assigned parameter values, the undefined parameters are given values by linear interpolation between the nearest segments where the parameter is defined. In this way, both forward and backward coarticulation are modeled.

In RULSYS the output parameters can be filtered using different time variable filters, in order to get the right dynamic properties. The face synthesis parameters are assigned different time constants depending on e.g. the mass of the corresponding speech organ.

3.3. Visual prosody

The visible articulatory movements are mainly those of the lips, jaw and tongue. However, these are not the only visual information carriers in the face during speech. Much information related to phrasing, stress, intonation and emotion are expressed by for example nodding of the head, raising and shaping of the eyebrows, eye movements and blinks.

This kind of facial actions should also be taken into account in a visual speech synthesis system, not only because they may transmit important nonverbal information, but also because they make the face look alive.

These movements are more difficult to model in a general way than the articulatory movements, since they are optional, and highly dependent on the speakers personality, mood, purpose of the utterance etc.

Nevertheless, a few general rules apply to most speakers. For example, it's quite common to raise the eyebrows at the end of a question and to raise the eyebrows at a stressed syllable. There have been attempts to apply such rules to facial animation systems [8]. A few such visual prosody rules have been implemented in our visual speech synthesis system.

However, to make the face live, one does not necessarily have to synthesize meaningful nonverbal facial actions. By introducing random eyeblinks and very faint eye and head movements, the face looks much more alive, and becomes more pleasant to watch. This is especially im-

portant when the face is not talking, e.g. during silent periods in a dialogue application (see section 4.1).

4. APPLICATIONS

There are several possible applications of an audiovisual text-to-speech system. As mentioned in section 1, the greatest benefit from the visual modality in speech perception can be expected during acoustically degraded conditions, regardless of whether the degradation is due to external noise or to hearing impairment.

Thus, audiovisual speech synthesis has many applications to hearing impaired people. It can for example be used as a tool for training of lipreading. A face with semi-transparent skin [4] can be used to visualize tongue positions in speech training for deaf children.

Synthesized audiovisual speech also has a great potential for information systems in public areas, such as airports and train stations and in multimodal dialogue systems in general.

4.1. Visual speech synthesis in the "Waxholm" dialogue system

A system for spoken man-machine dialogue is currently being developed at the department [1]. The system deals with information on boat traffic, restaurants and accommodation in the Stockholm archipelago. Input to the system is speech and output is in the form of synthetic speech and graphics, such as charts, tables and maps.

Recently, the visual speech synthesis module has been incorporated into the dialogue system. This is expected to raise intelligibility of the systems responses and questions. But the addition of the face into the dialogue system has many other exciting implications. Facial nonverbal signals can be used to support turn taking in the dialogue, and to direct the users attention in certain ways. (For example by letting the head turn towards time tables, charts etc. that appear on the screen during dialogue. This feature has recently been implemented.)

The dialogue system also provides an ideal framework for experiments with nonverbal communication and facial actions at prosodic level, as discussed in part 3.3, since the system has a much better knowledge of the situation than is the case in plain text-to-speech synthesis.

5. CONCLUSIONS

This paper describes the first effort on visual speech synthesis at KTH. The parameterized face can be viewed as a fully compatible extension to the existing KTH text-to-speech system. Like the formant synthesis, it is a terminal analogue technique, and it's controlled by the same software and uses the same rule notation as the acoustic speech synthesis.

The main effort so far has been on the system itself, i.e. the introduction of a new modality into the text-to-speech system. The result is a platform for experiments with and development of audiovisual synthetic speech by rules.

The set of rules, on the other hand, does have to be developed further. The grouping and parameter dependency of the visemes, and values and dynamic properties of the parameters are all empirically determined and based on relatively limited observations. In order to improve the rules, more visual articulatory data is needed. Such data might be gathered using automatic measurements of a real speakers speech movements.

6. ACKNOWLEDGEMENT

This work has been supported by grants from the Swedish National Language Technology Program.

7. REFERENCES

[1] Bertenstam, J., Beskow, J., Blomberg, M., Carlson, R., Elenius, K., Granström, B., Gustafson, J., Hunnicutt, S., Högberg, J., Lindell, R., Neovius, L., Nord, L., de Serpa-Leitao, A. and Ström, N. (1995): "The Waxholm system - a progress report", *Proc. ESCA Workshop on Spoken Dialogue Systems*, Vigsø, Denmark.

[2] Carlson, R., Granström, B., and Hunnicutt, S. (1982): "A multi-language text-to-speech module," *Proc. ICASSP-Paris*, Paris, Vol. 3, pp 1604-1607.

[3] Carlson, R., Granström, B., Karlsson, I. (1991), "Experiments with voice modeling in speech synthesis", *Speech Communication* 10, pp 481-489.

[4] Cohen, M. M. and Massaro, D. W. (1993): "Modeling coarticulation in synthetic visual speech", In N. M. Thalmann & D. Thalmann (Eds.) *Models and Techniques in Computer Animation*. Tokyo: Springer-Verlag.

[5] Le Goff, B., Guiard-Marigny, T., Cohen, M.M., Benoît, C. (1994): "Real-time analysis-synthesis and intelligibility of talking faces", Proceedings of the second ESCA/IEEE Workshop on Speech Synthesis, New Paltz, New York, USA.

[6] McGurk, H., MacDonald, J. (1976): "Hearing lips and seeing voices", *Nature*, 264, pp 746-748.

[7] Parke, F. I. (1982): "Parametrized models for facial animation", *IEEE Computer Graphics*, 2(9), pp 61-68.

[8] Pelachaud, C. (1991): "Communication and Coarticulation in Facial Animation", *Ph.D. dissertion*, University of Pennsylvania.