Towards a High-Quality and Well-Controlled Finnish Audio-Visual Speech Synthesizer

Mikko Sams, Janne Kulju, Riikka Möttönen, Vili Jussila, Jean-Luc Olivés, Yongjun Zhang, Kimmo Kaski
Helsinki University of Technology,
Laboratory of Computational Engineering,
P.O. Box 9400, FIN-02015 HUT, Finland

Päivi Majaranta, Kari-Jouko Räihä
University of Tampere,
Department of Computer Science,
P.O. Box 607, FIN-33101 Tampere, Finland

ABSTRACT

We have constructed an audio-visual text-to-speech synthesizer for Finnish by combining a facial model with an acoustic speech synthesizer. The quality of the visual speech synthesizer has been evaluated twice. In addition, we have started to refine the facial model taking a more physiologically and anatomically based approach. The synthesizer will be used to produce stimuli for studies of neurocognitive mechanisms of audio-visual speech perception. This sets requirements for flexibility and full controllability for the synthesis. We are also developing applications for the synthesizer.

Keywords: audio-visual speech, facial animation, multimodality, physically-based model, speech synthesis.

1. INTRODUCTION

The perception of speech is normally audiovisual. By using both auditory and visual modalities we can understand the message better than by relying on audition only. The visual component improves the intelligibility of speech especially when speech is exposed to noise [6], bandwidth limitation [5], hearing limitations or other disturbances. The two modalities convey complementary information; while some utterances (for example /ba/ and /da/) can be difficult to distinguish based on auditory information only, they are visually clearly distinguishable. On the other hand, /pa/ and /ma/ are visually very similar but they are easy to discriminate on the basis of auditory signal. Visual speech perception has its natural limits. We can't perceive the whole vocal tract visually but have to rely on information primarily from lips, tongue and teeth. Visual information is also crucial in determining the identity and emotional state of the talker, the reaction of listener and in conducting a fluent dialogue between two or more people.

We have constructed our first version of a Finnish text-to-audiovisual-speech synthesizer [1], which can produce real-time speech from unlimited written text. The visual part is a descendant of Parke's facial model [3], and it is synchronized with an acoustic text-to-speech synthesizer.

The intelligibility of our synthesizer has been evaluated twice [2,8]. The model without a tongue was used in the first evaluation. The second one was performed when the tongue model was added to the visual speech synthesizer and some phoneme articulations were improved on the basis of the first evaluation. Our objective is to further improve the quality of synthesis and to use the synthesizer as a stimulus generator for speech perception experiments. We are also developing applications for the synthesizer. It will be used, e.g., in teaching lip-reading. To achieve these goals an appropriate user interface for controlling the synthesizer is required.

2. THE CURRENT FACIAL MODEL

Our facial model is presented in Fig. 1. The geometry of the model is defined with slightly less than 1000 vertices that are used to construct about 1500 polygons. These figures do not contain the tongue vertices and polygons, because the amount of them can not be unambiguously stated. The facial geometry is controlled with 49 parameters 12 of which are used for visual speech. The parameters used in speech production are based on used coordinate system rather than physiological properties of the face.
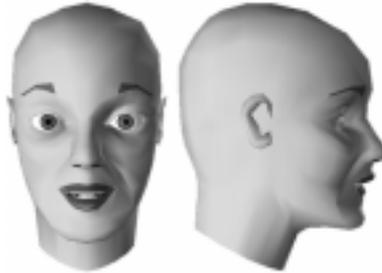
Figure 1. A front and side view of our facial model.

We have constructed a tongue model with flexible low-level parameterization (Fig. 2). The present tongue model does not comply with physiological realism. There are two fixation points connecting the tongue to the rest of the facial model: tongue root position (R) and the initial position of tongue tip (T). They are expressed in terms of facial model coordinate system.
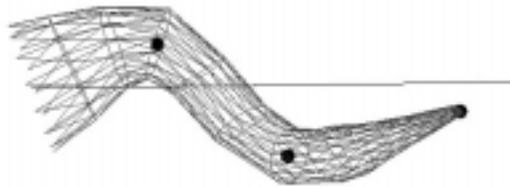
Figure 2. A wire-frame image of the tongue model. The control joints are marked with black dots.

The tongue geometry is implemented with elliptic layers. The amount of layers and amount of vertices in one layer can be freely decided. The width and height of each layer can be adjusted separately. In addition, the relative distance in R-T axis from R for each layer can be determined.

The low level control mechanism for tongue model is established by using joints. A relative distance from R and a three-dimensional offset is associated to each joint. The relative distance is similar to layer's relative distance, and the offset is expressed in terms of deviation from R-T axis in facial model's coordinate units. The amount of joints and their relative distances are modifiable, although there always exists one joint at the tip of the tongue. The actual position of tongue tip depends on T and tip joint offset. The tongue median line, *i.e.* the line passing through the center of each elliptic layer, is calculated by using cardinal spline fitting.

This low-level control mechanism allows the definition of higher-level control parameterizations. A viseme (a visual analogue of a phoneme) level control mechanism involves the decision of joint amount and offset definitions for each joint within each viseme. The tongue positions for the present viseme set were adjusted by using John Webster's x-ray microbeam database [9] and heuristic evaluation.

3. FROM TEXT TO SPEECH

Until now there was no appropriate audiovisual database for determining the Finnish phoneme articulation (viseme) set. Therefore, the original set used in our model was based on inspection of the modeler's own articulation and subjectively evaluating the result. The letter-to-phoneme correspondence is very strong in Finnish language, and therefore a letter-to-viseme mapping is used to determine the viseme sequence needed to produce a specific utterance instead of phoneme-to-viseme mapping. One letter of a written text corresponds to one viseme with one exception: written "nk" and "ng" are mapped to one viseme (Finnish */η/*).

The timing information for each viseme is obtained from acoustic synthesis. One viseme consists of "ideal" viseme data and coarticulation data for each 12 parameters. The ideal data is used as a basis that is modified by the coarticulation data of surrounding visemes. However, proper parameter values for coarticulation are still lacking and facial animation is created by linear interpolation between the ideal visemes. There were four exceptions in the model studied in the first intelligibility study [2]: velar /k/, /g/ and /η/ and glottal /h/ were estimated to have so context-dependent realizations that no ideal visemes were defined for them. When the tongue model was added, /g/, /η/ and /h/ were defined so that the tongue returns to the default position when these phonemes are uttered. A new viseme with proper tongue position was constructed for /k/ before the second intelligibility study.

Each ideal viseme is defined by 12 parameter values, each of which is an addition to the corresponding parameter value. The tongue model is controlled by additional parameter values. By using added instead of absolute values, it is possible to

separate the effect of speech from other facial gestures. During animation, the base parameter values are stored, and during each viseme values specific to it are added to the stored base values. Non-speech actions can affect the base values and thus change the expression during speech.

Coarticulation

The coarticulation model assumes that three properties are associated with each parameter of a viseme. Two of the properties are the forward and backward dominance, which describe the strength of the effect of prevailing viseme to next (forward) and previous (backward) viseme. The third property is a sensitivity parameter, which describes the degree to which a parameter of the prevailing viseme is modified by the coarticulatory effects of surrounding visemes. The sensitivity can be anything between 0 and 1. Zero means that the parameter has no context sensitivity. The final parameter target value p is calculated as

$$p_{final} = p_{base} + p_{viseme}$$

in which

$$p_{viseme} = p_{ideal} + s \cdot (fd_{prev} \cdot \Delta p_{prev} + bd_{next} \cdot \Delta p_{next})$$

$\Delta p_{prev/next}$ is the difference between ideal parameter value of prevailing and previous/next viseme for the parameter in question. This method does not take time into account, and we aim at adopting a more sophisticated coarticulation method in the future.

Acoustic synthesis and hardware

Currently our visual synthesizer can be used with two acoustic synthesizers: Synte V, which is being developed in the Laboratory of Acoustics and Audio Signal Processing of the Helsinki University of Technology, and modified version of MikroPuhe 4.1 by TimeHouse Ltd. The visual synthesizer can be run in any hardware environment with OpenGL support, and the existing combined syntheses can be used in 32-bit Windows (PC) or IRIX (SGI) operating systems. In IRIX environment it is possible to view the stereoscopic version of the visual synthesis by using appropriate eyewear.
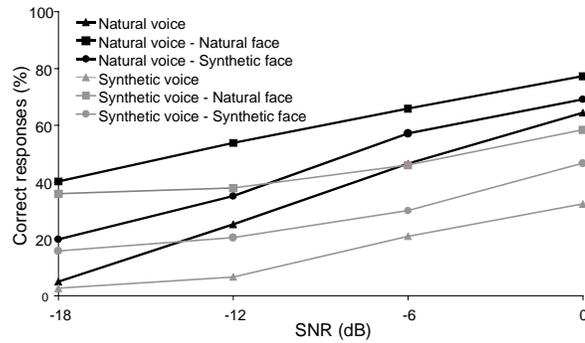
4. INTELLIGIBILITY

The First Intelligibility Study

The first intelligibility study for our synthesizer [2] was carried out before the tongue model was added to the visual speech synthesis. The test corpus consisted of 39 VCV words (13 consonants in , ε, and ψ contexts) which were presented in six conditions: 1) natural voice + natural face, 2) synthetic voice + synthetic face, 3) natural voice only, 4) synthetic voice only, 5) natural voice + synthetic face, and 6) synthetic voice + natural face. Signal-to-noise ratios (SNRs) of the stimuli were 0, -6, -12 and –18 dB. The subjects were 11 male and 9 female native Finnish speakers. The global intelligibility is depicted in Fig. 3. The facial animation improved the intelligibility of both the synthetic and natural acoustic speech. The mean improvement was about 15% being somewhat larger with smaller SNRs. However, the normal face improved the intelligibility another 15%.

The Second Intelligibility Study

We repeated the intelligibility study when the viseme set was improved by using the results of the first intelligibility study and a small database for Finnish visual speech [8]. The tongue model was added to the visual speech synthesizer after the first intelligibility study. The corpus included VCV words with 12 Finnish consonants in symmetric context and 8 VV words with all Finnish vowels. The utterances spoken by a natural talker (a female student of logopedics) and by the visual speech synthesizer were presented on a computer screen intermixed, in random order. Only visual speech was presented to 10 subjects. Their task was either to recognize consonants or vowels depending on the experimental condition. In addition the study included conditions, where subjects wore the CrystalEyes from StereoGraphics in order to see the visual speech synthesizer as three-dimensional.

Figure 3. The mean intelligibility scores of VCV words for the six different stimulus types (from the first intelligibility study).

The synthetic / /C/ / utterances were better recognized by lipreading in the second intelligibility study than in the first one. When in the first study 25% of the synthetic consonants were correctly identified in symmetric / /-context (at SNR of -18 dB), in the second one 33% of the synthetic consonant articulations were correctly identified. The proportion of correct natural consonant identifications was 54% in the second study. Moreover, in the second study 74% of the natural and 51% of the synthetic vowel articulations were correctly identified. The 3D presentation did not have a significant effect on the intelligibility of neither synthetic vowels nor consonants.

A hierarchical cluster analysis was performed to both natural and synthetic vowel and consonant confusions in order to find out how many viseme categories our subjects were able to recognize. One viseme category consists of phoneme articulations, which are frequently confused with each other in lipreading. The subjects were able to distinguish five consonant viseme categories from natural visual speech: bilabials (p, m), labiodental (v), incisives (s, t), alveolar dentals & velars (d, j, l, n, r, k) and glottal (h). Our subjects were able to identify these categories also from synthetic speech with the accuracy of 60% at least, except the glottal, which was poorly recognized because of the misleading tongue movement. The confusions among natural vowel articulations formed six viseme categories: ( ), (—, ɛ), (ι), (o), (↓), (υ, ψ). Four vowel viseme categories were distinguished from the synthetic visual speech: (,—, ɛ), (ι), (o,↓), (υ, ψ).

The intelligibility of our visual speech synthesizer is now almost at the stage where normal subjects without special training in lipreading can distinguish the same viseme categories from synthetic and natural speech. However, the individual phoneme articulations are not yet as informative as natural ones. In addition to fine-tuning the individual visemes, we expect that implementation of the coarticulation rules will significantly improve the intelligibility in general. To have more relevant baseline data for the further improvements, we shall test our model with hearing impaired persons who are expert lipreaders. The construction of the original visemes and their improvements has until now been done by comparing the viseme models to the articulation of normal talkers and making the necessary adjustments manually. We have now collected a small Finnish audiovisual speech database as the basis for improvements. We also aim at developing automatic extraction of the facial movements from the database.

5. PHYSICALLY-BASED FACIAL MODEL

In order to produce as natural looking facial animation as possible we have started developing a head that is motivated by physiology and anatomy of a real head [4]. The head geometry is obtained from magnetic resonance images (MRIs) of one subject. Its functionality is based on three main components: the skin model, the muscle model, and the skull model.

Skin model

The skin model is composed of a two-layered lattice [4] of point masses, connected by springs. The model is a simplification of the real human skin, which also has a layered structure consisting of the epidermis, a superficial layer consisting of dead cells, and the fatty dermis. Newtonian mechanics can be used to integrate the motion of point masses trough time. This involves calculating the positions, velocities and accelerations for point masses at time steps t, $t+\Delta t$, $t+2\Delta t$, $t+3\Delta t$, ..., where t is the starting time of the simulation and Δt is the difference between the time steps.

Muscle model

We have implemented a linear muscle model [7] in which a muscle is a straight line which extends from some top layer point mass to the sub layer. When the muscle is contracted, it applies an attractive force in direction of its tail, to the point mass it is attached to. This force is further reflected through the spring lattice to all nearby point masses. The result is that the skin
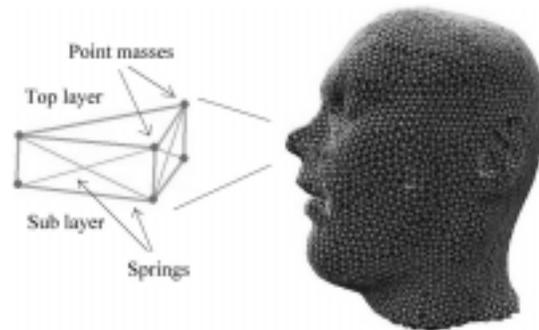
Figure 4. Wire-frame version of the head model shows the form of the spring lattice. The head consists of about 7000 vertices. On the left is an enlarged side view of a single lattice element.

around the attachment point of the muscle is displaced toward the other end of the muscle. Currently the head model contains only two muscles at the corners of the mouth. These are used for testing purposes but they can also be used to make the face to smile (Fig. 5).

Figure 5. On the left the face is in its initial form. On the right the muscles at the corners of the mouth have been contracted and the face smiles.

Skull model

Forces acting on the skin during muscle contraction can make the skin to collapse or sink inside the head. To improve such situations, we implemented a skull model, which restricts the movement of the point masses so that they cannot go through the skull. The skull is a down-scaled version of the head and at every time step the movement path of each point mass is checked against the surface of the skull. If the movement path intersects with the surface it is modified so that the point mass will slide on top of the skull instead of going through it. In the future we probably use real skull data from CT measurements to further increase the quality of the model.

6. TOOLS FOR EASY CONTROL OF THE HEAD

It is important that the head is under a strict and easy control of an experimentalist or a speech therapist. For this reason, we have developed software tools to modify the head, its expressions and speech. With the graphical user interface it is easy to create well-controlled stimulus sequences.

The user interface provides both high and low level controls for the head. Low level controls can be used to control the individual parameters, e.g., the size of the pupil. High-level controls change several parameters simultaneously, e.g., those that make the face look happy. There are about 100 controls for the head.

Figure 6. User interface for the head.

With the tools it is possible to change the whole facial appearance and the appearance of the dynamic facial features. This means, that any single facial parameter can be controlled separately. The tools give also means to control the orientation and the movements of the head, eyes and tongue.

With the tools both the auditory and visual speech can be controlled separately. The user can, e.g., define the correspondences between visemes and phonemes. The user can also change the individual parameters of the auditory speech, such as rate and pitch, and fine tune the visemes or create user-defined viseme libraries.

The head is able to produce facial displays, which can be controlled by the user interface. The face can smile, look angry, sad, etc. Also the intensities of the expressions can easily be modified. The user can create and save new expressions, and load them during the speech.

The user can create timed speech sequences. All the audiovisual parameters, orientation and expressions can be controlled during the speech. The actions during the speech are controlled using tag commands like <happy 100> (look happy with intensity of 100%). The user does not have to remember the exact syntax of the tags, because the interface includes an easy to use tag wizard.

Figure 7. The face can express basic facial emotions.

7. CONCLUSIONS

We have developed a Finnish audio-visual speech synthesizer that can be controlled flexibly and accurately by a graphical user interface. Our evaluation has shown that the visual speech synthesis significantly improves the intelligibility of acoustic speech presented in noise. We are currently improving the set of visemes, implementing the coarticulation rules and improving the synchrony between the visual and acoustic speech. These changes will make the visual speech more natural and further increase the intelligibility of the synthesizer. In parallel, we have started to develop a physiologically and anatomically based model consisting of the skull, the skin, and the facial muscles. This new model is computationally very demanding, but the visual speech and facial expressions are expected to be more natural. The audio-visual synthesizers are used to produce well-controlled stimuli to study the neurocognitive mechanisms of audiovisual speech. We are also developing applications. As an example, the synthesizer is included in a lip-reading tutorial, which will be used by speech therapists.

8. REFERENCES

[1] J. Kulju, M. Sams, K. Kaski. A Finnish-Talking Head. *Proceedings of the Finnic Phonetics Symposium, Linguistica Uralica* XXXIV, 3: 329-333, 1998.

[2] J.-L. Olives, R. Möttönen, J. Kulju, M. Sams. Audio-visual Speech Synthesis for Finnish, *Proc. AVSP '99*: 157-162, 1999.

[3] F. Parke. Parameterized Models for Facial Animation, *IEEE Comp. Graph.*, 2: 61-68, 1982.

[4] F. Parke, K. Waters. *Computer Facial Animation*, A K Peters, Ltd. 1996.

[5] A. Risberg, J. Lubker. Prosody and Speechreading, *Quarterly Progress & Status Report 4*, KTH, Speech Transmission Lab, Stockholm: 1-16, 1978.

[6] W. Sumby, I. Pollack. Visual Contribution to Speech Intelligibility in Noise. *J. Acoust. Soc. Am.*, 26, 2: 212-215, 1954.

[7] D. Terzopoulos, K. Waters. Physically-based Facial Modelling, Analysis, and Animation. *J. Vis. Comp. Anim.*, 1: 73-80, 1990.

[8] R. Möttönen, J-L. Olivés, J. Kulju and M. Sams. Parameterized Visual Speech Synthesis and Its Evaluation. *Proc. of EUSIPCO 2000*, Tampere, Finland, in press.

[9] J. R. Westbury, "X-ray Microbeam Speech Production Database User's Handbook", Waisman Center on Mental Retardation Human Development, University of Wisconsin, USA, 1994.