https://doi.org/10.3176/lu.1998.3.28

JANNE KULJU, MIKKO SAMS, KIMMO KASKI (Helsinki)

A FINNISH-TALKING HEAD

In face-to-face communication speech perception is both visual and auditory. Under very noisy conditions the visual information from the talkers articulation can help to understand hardly audible speech. We have started developing a Finnish audio-visual speech synthesizer. In our current model, we have combined a three-dimensional facial model based on work by F. Parke (1982) with a commercial audio text-to-speech synthesizer. The visual speech is based on straightforward letter-to-viseme mapping, in which each letter of a written text corresponds to a viseme. Visual speech is animated by linear interpolation between visemes. Coarticulation has not been modeled. Audio-visual speech synthesizer can be used to prepare well-controlled stimuli for speech research and cognitive neuroscience. In addition, various application areas will benefit of high-quality audio-visual speech synthesis including telecommunication, human-computer interfaces, and speech therapy.

Introduction

The perception of the speech is normally audio-visual. By both hearing and seeing the talker, we can infer the contents of the message, identity and emotional state of the talker, and the talker's reaction to our speech better than with only auditory information. The contribution of visual information is twofold. It improves the intelligibility of auditory speech especially when speech is exposed to noise (Sumby, Pollack 1954), bandwidth limitations (Breeuwer, Plomp 1985), hearing limitations or other disturbances. On the other hand, the acoustic and visual components of speech are complementary: while some utterances (for example */ba/* and */da/*) can be difficult to distinguish based on auditory information only, they are visually clearly distinguishable. Visual speech perception has its natural limits, though. We can't perceive the whole vocal tract visually but have to rely on information primarily from lips, tongue and teeth. Therefore some utterances (for example */ma/* and */pa/*) are hard, if not impossible to be distinguished visually.

In speaker recognition, the visual information is naturally crucial. Also such information which is traditionally treated as paralinguistic, like emotions, can be recognized more precisely by integrating auditory and visual modality. The neurocognitive mechanism of integration is known very crudely, but evidently there exists a strong and intricate synergy between these two modalities. The gain obtained by integration can be huge. According to A. Risberg and J. Lubker (1978), the percentage of correctly recognized words in low-pass filtered speech combined with speaker's video image was

much greater (45%) than sum of percentages that were gained by speechreading without sound (1%) and mere sound (6%). The influence of visual modality is not limited to clarifying distorted auditory speech signal. Contradicting auditory and visual stimuli may be perceived something that is neither of initial stimuli. For example, visual /*ga*/ mixed with auditory /*ba*/ is often perceived as /*da*/ (McGurk, MacDonald 1976).

Although not as effective as natural faces, synthetic faces increase the intelligibility of both natural and synthetic auditory speech (e.g., Beskow, Dahlquist, Granström, Lundeberg, Spens, Öhman 1997). Therefore many application areas, such as telecommunications, human-computer interfaces and speech therapy would benefit from audio-visual speech synthesizer of good quality. Currently, there are at least English (Parke 1982; Waters 1987; Pearce, Wyvill, Wyvill, Hill 1986), French (Le Goff, Benoît 1997) and Swedish (Beskow, Dahlquist, Granström, Lundeberg, Spens, Öhman 1997) audio-visual speech synthesizers.

We have constructed the first version of Finnish text-to-audiovisual-speech synthesizer by combining a commercial text-to-speech synthesizer (modified version of MikroPuhe 4.1 by Timehouse Inc.) with a three-dimensional facial model based on work by F. Parke (1982). We aim at using the talking head for studying the neurocognitive mechanisms of the speech perception, in teaching lipreading and in speech therapy.

Computer-based synthetic visual speech

Both auditory and visual speech are the result a single physiological process. Then perhaps the best possible audio-visual synthesis would probably be obtained by modeling that process. Unfortunately, constructing an accurate model of vocal tract and face is currently beyond the capabilities of researchers, because there is not enough information of the accurate dynamics of articulatory tract during speech. Therefore, simplified and computationally effective methods have to sought for.

A fundamental part of visual speech synthesis is the facial animation. In computer-based facial animation roughly two approaches have been used: concatenative and modeling animation. A similar division describes also acoustic synthesis methods.

In concatenative approach, the animation is created by interpolating between stored images. One simple way of doing this is called key-frame or key-pose animation. In this approach, certain facial expressions are stored, and visual speech is created by interpolating between these expressions. A novel method resembling this approach is morphing video images based on prerecorded databases to create natural-looking visual speech (Ezzat, Poggio 1997; Bregler, Covell, Slaney 1997).

The modeling approach is based on face model, in which the controlling is carried out with appropriate parameters, whose values combined with initial face geometry define the facial expression. Facial animation is created by changing the values of control parameters and redrawing the face by using new values. The parametrization depends on the nature of the model. The model can be a topological model with little physiological basis (for example Parke 1974; 1982; Pearce, Wyvill, Wyvill, Hill 1986), or it can simulate the anatomical, physiological and dynamic properties of human face to a certain degree of accuracy (for example Platt, Badler 1981; Waters 1987). F. Parke and K. Waters (1996) make a distinction between parametric models and muscle-based models. Since muscle-based models try to simulate the actual human speech process, they would be the best way to model facial movements during speech. However, the improvement of the accuracy and reality of the model means increment in difficulty of constructing such model. Other drawback is that such models need much more computing power than topological

models, which restricts their usage in applications. Facial Action Coding System, FACS, (Ekman, Friesen 1978) is a widely used method in describing facial expressions. Although it is originally developed for observing purposes, it has been also used as control parametrization in several facial models (for example Platt, Badler 1981). The basic unit of FACS is an action unit, AU, which is defined as a minimal noticeable action that can't be expressed as a combination of smaller actions.

The positions of visible articulators are dependent of the uttered phoneme. Therefore the viseme to be produced at certain time and the synchronization information can be received from timed phoneme sequences, obtained from an acoustic speech synthesizer or from a speech recognition tool. According to Q. Summerfield (1992), a 160 ms delay in audio signal strongly reduces the benefit of visual modality, but up to 80 ms the effect of delay is usually negligible. Therefore, we assume that the allowed maximum error in synchronizing audio and visual speech is few tens of milliseconds. Though the basic idea of using acoustic synthesizer for timing and synchronization of visual synthesis is straightforward, the practical implementation of animation and synthesizer and the used hardware determine the difficulty of the task.

The rate of change of articulators' positions is limited by physical properties of vocal tract. Rapid changes are anticipated beforehand and the initial state can be traced back after a rapid change. This context-sensitivity which is called coarticulation needs to be taken into account while modeling audio-visual speech. M. Cohen and D. Massaro (1993) have studied modeling of coarticulation and suggest the use of dominance functions, in which an exponentially damping function is attached to each visual speech segment to represent its forward and backward dominance on other speech segments.

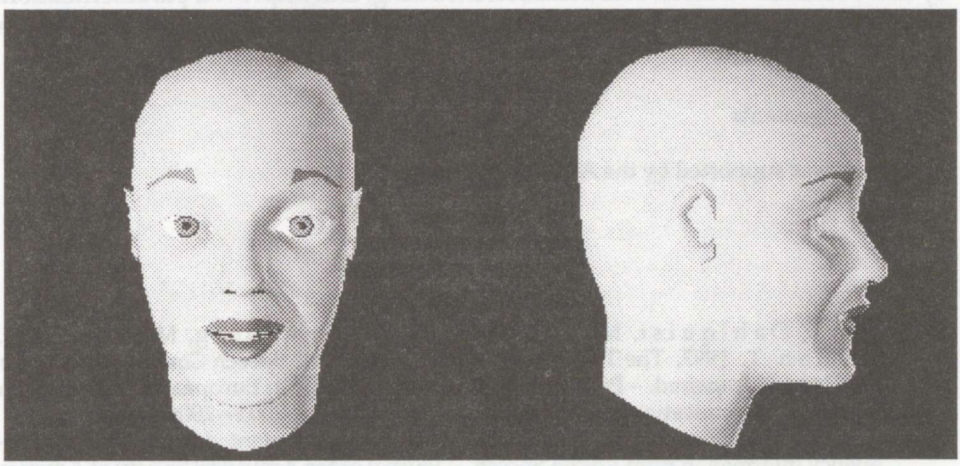

Figure 1. **Left: a front view of our facial model uttering a /r/. Right: a side view of the same model uttering a /v/.**

The implementation of our model

Our facial model (see Figure 1), a descendant of F. Parke's (1982) model, is a parameter-controlled topological model. It is currently controlled using 49 parameters 12 of which are used for visual speech affecting the jaw rotation and lip formation. The parameters used in speech production are based on used coordinate system rather than physiological properties of face. The audio-visual speech synthesizer consists of the facial model and MikroPuhe 4.1 audio speech synthesizer (Time-

house Inc.), which has been modified slightly in order to get signals needed in synchronization. Synchronization information, currently active viseme as well as previous and subsequent visemes are gained from audio synthesizer.

The 12 viseme parameter values are based on heuristic trial and error. One letter in text corresponds to one viseme with the exception of *nk* and *ng*, which are represented as one viseme. Coarticulation has not been taken into account except for *h*, *k*, *g*, *nk* and *ng*, which depend of preceeding and following viseme. This primitive approach in viseme determination is mainly due to the lack of appropriate audio-visual speech database. Due to the used parametrization and lack of tongue in the model, the available viseme collection is limited.

The input to synthesizer is given as text. The visual speech is animated by linear interpolation with 1—5 intermediate steps between parameter values corresponding the initial and final visemes. The synthesizer can reproduce unlimited Finnish text. A 3D impression can be simulated with the aid of stereographic glasses. The graphics is implemented using libraries that are available in multiple hardware platforms, and currently our model can be run in SGI IRIX and PC (Windows) environments.

Conclusions and future work

We have constructed our first version of Finnish audio-visual speech synthesizer. Intelligibility tests have not yet been performed, so objective evaluation of its quality can not be presented. This is one of the first things to do in the future. Other future enhancements include improving the quality of visual synthesis by more realistic visemes, coarticulation modeling, and improved parametrization. The visual outlook of the model will be improved via tongue addition and texture mapping.

Acknowledgements

This study was supported by the Academy of Finland.

REFERENCES

Beskow, J., Dahlquist, M., Granström, B., Lundeberg, M., Spens, K., Öhman, T. 1997, The Teleface Project — Multimodal Speech Communication for the Hearing Impaired. – Proceedings of Eurospeech'97, 5th European Conference on Speech Communication and Technology, vol 4, Rhodes, 2003—2006.

Breeuwer, M., Plomp, R. 1985, Speechreading Supplemented With Formant Frequency Information from Voiced Speech — Journal of the Acoustical Society of America 77, 314—317.

Bregler, C., Covell, M., Slaney, M. 1997, Video Rewrite: Visual Speech Synthesis from Video. — Proceedings of the Workshop on Audio-Visual Speech Processing, Rhodes, 153—156.

Cohen, M., Massaro, D. 1993, Modeling Coarticulation in Synthetic Visual Speech. — Models and Techniques in Computer Animation, Tokyo, 141—155.

Ekman, P., Friesen, W. 1978, Manual for the Facial Action Coding System, Palo Alto.

Ezzat, T., Poggio, T. 1997, Videorealistic Talking Faces. A Morphing Approach. — Proceedings of the Workshop on Audio-Visual Speech Processing, Rhodes, 141—144.

Le Goff, B., Benoît, C. 1997, A French-Speaking Synthetic Head. — Proceedings of the Workshop on Audio-Visual Speech Processing , Rhodes, 145—148.

McGurk, H., MacDonald, J. 1976, Hearing Lips and Seeing Voices. — Nature 264, 746—748.

Parke, F., 1974, A Parametric Model for Human Faces. PhD Thesis, University of Utah, Salt Lake City.

—— 1982, Parameterized Models for Facial Animation. — IEEE Computer Graphics 2(9), 61—68.

Parke, F., Waters, K. 1996, Computer Facial Animation, Wellesley.

Pearce, A., Wyvill, B., Wyvill, G., Hill, D. 1986, Speech and Expression. A Computer Solution to Face Animation. — Proceedings of Graphics Interface, Calgary, 136—140.

Platt, S., Badler, N. 1981, Animating Facial Expressions — Computer Graphics 15 3, 245—252.

Risberg, A., Lubker, J., 1978, Prosody and Speechreading. — Quarterly Progress & Status Report 4, KTH, Speech Transmission Laboratory, Stockholm, 1—16.

Sumby, W., Pollack, I. 1954, Visual Contribution to Speech Intelligibility in Noise. — Journal of the Acoustical Society of America 26, 212—215.

Summerfield, Q. 1992, Lipreading and Audio-Visual Speech Perception. — Transactions of the Royal Society of London 335, 71—78.

Waters, K. 1987, A Muscle Model for Animating Three-Dimensional Facial Expression. — Computer Graphics 21 4, 17—14.