Talking Heads and Hearing Impaired Persons

Kenneth Wilhelmsson, kw@ling.gu.se
Dept of Linguistics, Göteborg University

Abstract

This paper aims at giving a brief introduction to the research area of computer animated images of faces used for helping the hearing-impaired. The focus is especially on the progress being made in the Synface project and its predecessor. Some of the basic methods are mentioned and some description of the current state of research is given.

1. Introduction

An example of what applied speech technology can result in is a helpful type of device for the hearing impaired. By analysing the sound of speech technically and produce the correspondent facial movements on a computer image of a face, hearing impaired can be presented with a situation much like that of true face to face conversation. Encouraging results from this area of research come from the work in the Teleface project which is at the time being continued in a new project, Synface [1].

The construction of this particular application relies much on the development in the area of automatic speech recognition (ASR), and a successful, helpful behaviour of course confirms some sort of validity in the analysis of acoustic speech.

In this article the question of the relevance of this type of application will first be posed. After that, some description and results from the two projects Teleface [2] and Synface [1] will be presented. The term visemes is then examined and some general questions finish this paper.

2. Why Talking Heads?

Language technology has in many ways been able to help different groups of disabled people. A large group of people likely to benefit by development in speech technology is that of hearing impaired persons. It was earlier perhaps not the most obvious purpose for development in speech technology to serve this group of people, for whom the reading of lips and facial movements play a particularly important role. The bimodal nature of speech is for this group a great advantage; many deaf and hard of hearing people have a trained ability of lip-reading (only in Western Europe the size of this group is three million [3]). Also, people without any hearing difficulties benefit by the various facial expressions conveyed together with the actual speech signal. The important and sometimes crucial effect of the visual information is particularly known through the McGurk effect [12] – which is an interesting phenomenon. In brief, it concludes that an unclear or even a clear sound can be interpreted differently only depending on the visual information conveyed at the same time. For example *"spim"* might be interpreted as *"spym"* if there is a clear lip-rounding as visual information. (The McGurk effect stated, after some initially misleading conclusions, that *the least ambiguous* source of information takes higher precedence than the other in terms of influence for interpretation.)

The McGurk effect also seems to imply that the received impression sometimes is a mix of the two modes, e.g., the visual message for "ga" together with an audio signal conveying "ba" has been showed to give the impression of "da" to the receiver.

Talking heads can for example be defined as "a parametrically controlled three-dimensional polygonal model that can be animated in synchrony with natural speech" (Beskow, 1997), [4]. A pioneer in this area was Frederic I Parke in the 1970s. Facial animation using direct parametric models seems often to mean developing a number of *key poses*, from which the rest of the facial moves are interpolated using a certain algorithm. There are also other approaches to facial animation, e.g., those based on modelling an assembly of tissues such as skin, muscles, fatty tissues etc.

The definition above will cover many of the cases found in animated environments, like for example in the film Toy Story. But all types of talking heads are not such that they provide help for the hearing impaired and one of the key questions in this area of research is what the relevant aspects of facial articulation are for helping this group of people? A test concerning the intelligibility of sound accompanied with different visual information was carried out in [5], see figure 1.

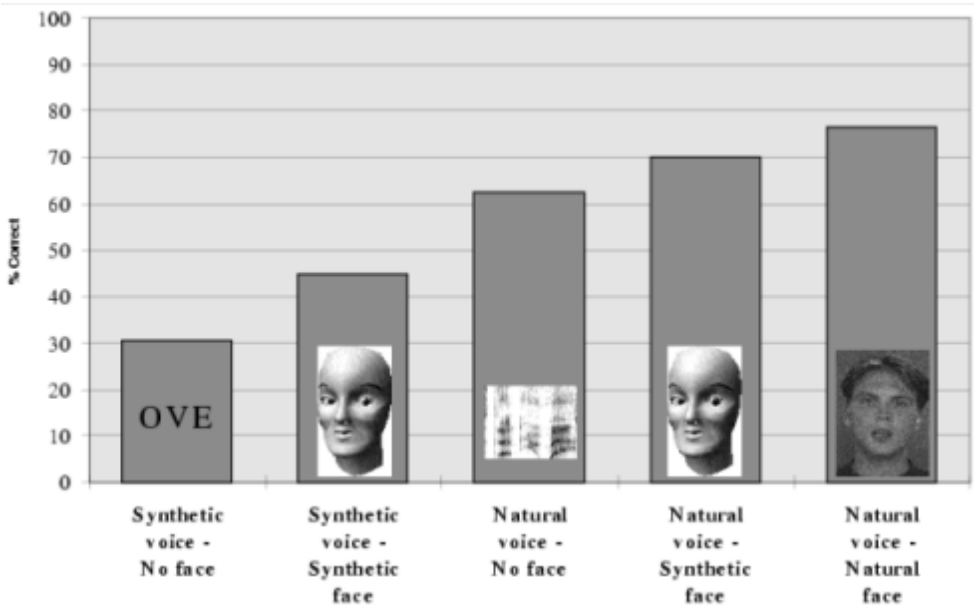

Figure 1: In an intelligibility test with normal hearing persons (students at KTH) the above result was given [5]. It seems to state that naturalness correlates with comprehensiveness. These results came from a situation where a natural voice was heard together with either of three the visual alternatives (meaning without any visual information in the first case). White noise was mixed into the acoustic signal, making the importance of visual information more evident. This test was part of the Teleface project, (see below).

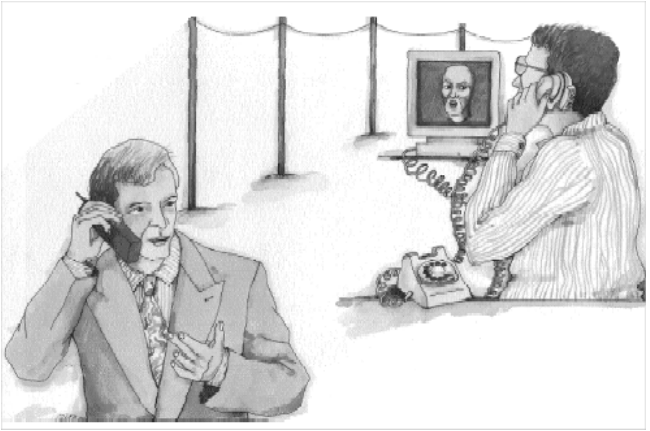

Figure 2: The idea in the Synface project is to examine how an animated face can convey the speech information to a hearing impaired listener [13].

From the results of figure 1 (from the Teleface project), it seems the grade of naturalness roughly resulted in the grade of intelligibility, but also that a synthetic face helps for making a correct interpretation. The use of a synthetic face is better than no representation at all for hearing impaired people [3]. So, the initial standpoint ought to be that video telephones should deliver the very best results for the telephone situation, these may support sign language and text too. But for this purpose, cameras and microphones are needed on both sides when a conversation takes place. This demand, together with that of a sufficiently high bandwidth to transmit a reasonable video quality, makes that solution less flexible than the systems presented here [3].

The aim of the Synface project is to develop a means for personal computers that improves a telephone conversation with a visual modality. A talking head appearing on the screen provides the lip-reading possibility for the hearing-impaired. The helping aid is thus only located on the side of the conversation where the help is needed, and will also hopefully be voice-independent.

Furthermore, it should be noted that people differ much in how much they articulate during speech. It seems to be the case too that different recipients of speech receive a spoken message better through different persons' articulations. An interesting task is to come up with a facial model and strength of articulation that seem suitable in a general perspective.

3. Teleface, Synface and their Predecessors

At KTH the first step of research was taken by work with an audio-visual text-to-speech-synthesis framework in 1995. Talking animated agents were used together with synthetic speech in some earlier projects, like in the Waxholm [6], and the Olga projects. The role of those was not as in the later Teleface and Synface to be used for hearing impaired, but as stated above, also persons with normal hearing can benefit from this. A face, like that found in the August project can help keeping track of the other participants' moods in a conversation. In the August project other facial expressions than those connected with the actual speech have an important role in displaying a communicative personality on the screen.

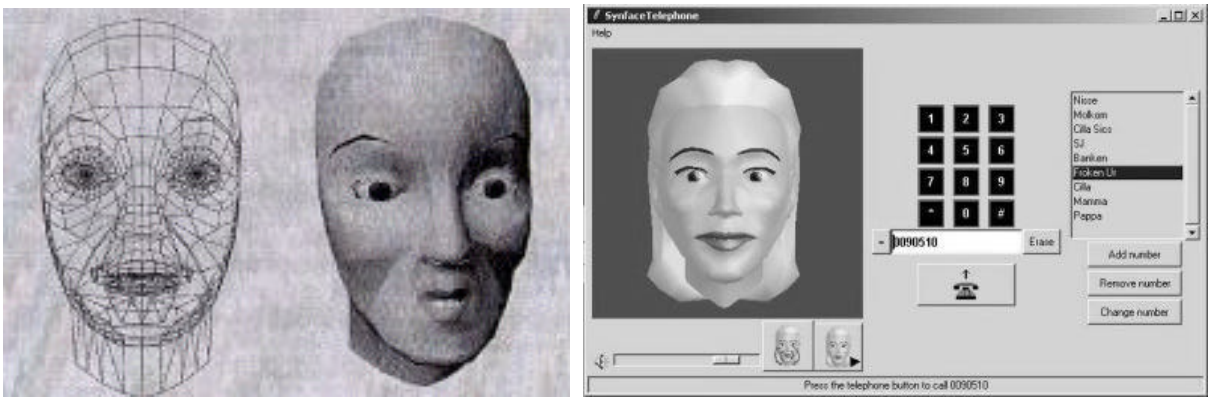

Figure 3: The visual interfaces of the Teleface project (left) and the Synface project (right) bear some resemblance to each other. To the left modelling by the Parke model is shown clearer. In the Synface project it is possible to choose from some different faces and the rest of the phone interface is built-in.

The systems developed in the projects Teleface and Synface aimed at providing any speech received by telephone with the right behaviour of the facial articulators, namely the lips, the tongue and the jaw, of a synthetic face on a computer screen.

The purpose of the Teleface project, which is now finished, was to estimate the value of a synthesised animated face helping a hearing impaired by the phone. The result was positive in the sense that it was concluded that a working device would be appreciated. In the project, many different investigations

concerning speech perception using the computer image (see for example figure 2) were carried out. (From the research of the project it was concluded, among other things, that front consonants logically enough are made easier to distinguish when using visual information, [5]) In the project a framework was constructed for the study of how the various parts of the face were moving during natural speech.

The project examined the potential usefulness with respect to the grade of impaired hearing. The highest absolute benefit from speech-reading a synthetic face was that of those perceiving between 40 and 80% of keywords in an audio only condition [7]. People who had a higher score than approximately 80% of the keywords seemed to rely on their hearing, that group seemed in the test not to gain anything from either a natural or a synthetic face.

In the Teleface project one parametric speech model was used. The parameterisation approach comes from Parke and works by describing *the surface* of the synthetic face. It consists of about 800 polygons and movements are described using 50 parameters [8]. Furthermore, this model was modified; tongue movements were introduced and a new set of parameters was used for the lips. The tongue was produced from 64 polygons and parametrically controlled by length, width, thickness and apex. As for the modification of the lip parameters, the original way of parametric control in Parke's modelling is said to be difficult. Important additions were lip rounding, bilabial closure and labiodental closure. In systems with a modelled artificial agent "with a mind of its own" other features become interesting than those directly involved in speech production. This is what can be referred to as *visual prosody* and means movement of eyebrows and eyes, nodding etc. This sort of behaviour seems not applicable in the situation dealt with here but it is present in for example the August spoken dialogue system [9], see figure 4.

Figure 5: The animated face in the August system did not aim specifically at hearing impaired users [9].

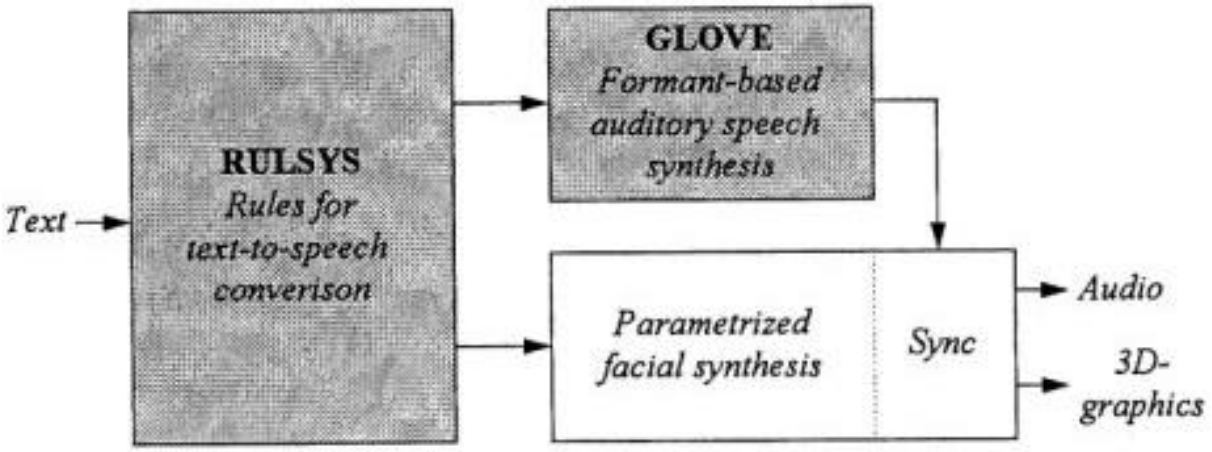

Figure 6: An audiovisual speech model was created [8] that worked schematically like this: RULSYS (a synthesis framework) operated on incoming text according to morphologic, syntactic and phonetic rules. The output from here was a multi-channel data file containing parameter values for both the auditory and the visual synthesiser. The sample rate was 10 samples per second. Glove is a

synthesiser module that created an audio sample file based on this information. The face synthesis module then played back the audio synchronised with the face model.

In the Synface project, which will be evaluated using Swedish, Dutch and English, an investigation about the minimum required precision of visual speech information is carried out. The evaluation is conducted by comparing the information gain with two reference points: correctly received information through speech only and correctly received information in a situation where a voice is heard and a natural face also conveys the message [1]. It is stated that systems developed this far do not achieve the task of displaying the right animation state corresponding to a true facial movement at any given moment, although systems that more accurately display natural movements are being developed at KTH.

4. Baldi

Like the mentioned projects at KTH, the American Baldi project [11] is aiming at producing a talking head that moves to a speech signal. In this project a system for collecting a gallery of different faces is being built. The specific Baldi face can be used as a prototype for facial animation and new customised models can be made to adapt the movements of Baldi after a series of mapping points to the new face. In this project 11 facial control parameters are used: jaw rotation, lower lip f-tuck, upper lip raising, lower lip roll, jaw thrust, cheek hollow, philtrum indent, lip zipping, lower lip raising, rounding and retraction. The results from here are also said to be promising.

Figure 7: In the leftmost figure a training situation is shown; LED markers are used for measuring movements. The figures to the right show the alignment points placed on a laser scanned image and on the Baldi face.

5. Visemes

Visemes are generic facial images that can be used to describe a particular sound. One might come a definition like this. *"A viseme is the visual equivalent of a phoneme or unit of sound in spoken language."* [a] I suppose the analogy of this way of definition cannot be taken to apply precisely the same way in all practical applications. A phoneme is the smallest contrastive unit in the sound system of a language, whereas visemes probably sometimes seem to coincide. The corresponding visemes for /g/ and /k/, will probably not help differentiating between "game" and "came".

However, at the KTH projects the above definition of visemes is *not* used. Visemes are not regarded as the smallest contrastive units of the visual mode of speech, instead they are regarded as poses in the animation modelling, sometimes for coinciding phonemes.

6. Articulation Strength

It is interesting to note that the strength of articulation (i.e., how much movement of the facial articulators) that is useful for persons using a synthetic face, is an open and probably an individual

question. From the approach with a digitally synthesised composition of moving polygons follows the possibility to give the facial articulation an extra emphasis in order to possibly become even clearer to the user. In [10] the question of optimal articulation strength was investigated. The parameters which were used in articulatory control are shown in Table 1.

Parameter	**Examined by altering strength**
Jaw rotation	X
Labiodental occlusion	
Bilabial occlusion	
Lip rounding	
Lip protrusion	X
Mouth spread	X
Tounge tip elevation	X

Table 1: The parameters involved in the testing of effects from changing the amount of articulation were four of the total number of seven parameters. (The not examined properties are probably not worth controlling by altering strength)

The conclusion drawn from this experiment was that a slightly over-articulated variant (by 25% in average) of the originally used measures resulted in the most comprehensive expression. Furthermore, as opposed to the comprehensiveness aspect, similar tests suggested that a decreased level of articulation (by 25%) was making a more natural impression than an over-articulation of 25%. This difference is interesting in the sense that it might imply that what is here *most natural* may not be the ideal behaviour for this application.

A question might be whether people also have this type of rating for natural faces (which would of course be hard to test, and ought certainly to result in normal articulation strength being most natural), or if the synthesised faces differ from natural faces in that they from the start have a lack of naturalness, and become even more unnatural when aiming at comprehensiveness.

7. Some Questions

It seems relevant to compare the development in this area with the research in automatic speech-to-text, where a process is tied to the accurate analysis of an incoming speech signal. In fact, since the same source of transformation allows going to *visemes* as going to text, it could be argued that an aid for the hearing impaired such as that developed within the Synface project probably could be augmented with a text representation of the signal too.

The goal of going to correctly spelled text from a representation of phonemes is perhaps not reached yet (especially not in the case where an unidentified voice is transmitted via the phone net[1]). But it seems some sort of textual representation in addition to merely the facial expressions of the talking head could be helpful. A reason for my thoughts here is that the articulation of a natural, as well a synthetic face, sometimes hinders a correctly analysed sound from being clearly evident from the facial articulation. An example of this situation is how consonant sounds between two /u/ (Swedish U) phonemes are harder to distinguish compared to when surrounded by /a/ [9]. It seems facial expressions do not always convey the expression as distinct as would normal text in this case. Sometimes the importance of a text source (with segmented words) is stressed. The more multimodality is used for conveying information the better, is a personal hypothesis. But, of course all

[1] The difficulties involved in correctly analysing an input signal can be experienced through the automatic travel information at SJ (0771-75 75 75).

this relies on getting the speech signal right in the first place, and it is not an easy task to divide the speech signal into words.

In [3] some general difficulties concerning the methods are also presented. The telephone situation is dependent on quick responses and not too long times of delay – this means the analysis of speech followed by the synchronised visual and sonic output needs to be a fast process (the delay should be less than 250 ms) or the communication will suffer from lack of cohesion.

8. Web Links

The Teleface project (KTH): http://www.speech.kth.se/teleface/index.html
The Synface project (KTH): http://www.speech.kth.se/synface/
Talking Heads (Yale): http://www.haskins.yale.edu/haskins/heads.html
Baldi (University of California): http://mambo.ucsc.edu/

9. References

1. Granström, B. et al. 2002. Synface – a project presentation. From *TMH-QPSR Vol. 44 – Fonetik 2002*
2. Beskow, J. et al. 1997. The Teleface Project – Multi-modal Speech-communication for the Hearing Impaired
3. Ward, K. 2002. User interface for the Synface project. M.Sc. Thesis, Department of Speech, Music and Hearing, Kungliga Tekniska Högskolan
4. Beskow, j. et al. 1997. Animation of Talking Agents. From *Proceedings of AVSP'97. ESCA Workshop on Audio-Visual Speech Processing,* Rhodes Greece
5. Beskow, J. et al. 1997. The Teleface project - disability, feasibility and intelligibility in *Proceedings of Fonetik-97, Swedish Phonetics Conference*, Umeå, Sweden
6. Carlson R. & Granström B. 1996. The Waxholm spoken dialogue system. From *Palková Z, ed. Phonetica Pragensia IX. Charisteria viro doctissimo Premysl Janota oblata. Acta Universitatis Carolinae Philologica 1;* 39-52.
7. Agelfors E. et al. 1998. The synthetic face from a hearing impaired view The synthetic face from a hearing impaired view. *Fonetik98*
8. Granström, B. 1999. Multi-modal speech synthesis with applications From *G. Chollet, M. G. Di Benedetto, A. Esposito, M. Marinaro, (Eds) Speech Processing, Recognition and Artificial Neural Network, Proceedings of the 3rd International School on Neural Nets "Eduardo R. Caianiello" Springer London 1999,* 327-346
9. Gustavson, J., Lindberg, N., Lundeberg, M. 1999. The August spoken dialogue system. From *Proc. Eurospeech'99, Budapest, Hungary, 1999.*
10. Beskow J., Granström B. & Spens K-E. 2002. Articulation strength - Readability experiments with a synthetic talking face. *Proc of Fonetik 2002, TMH-QPSR,* 44: 97-100
11. Cohen, M.M. and Massaro, D.W., and Clark R. 2002. Training a talking head From *ICMI'02, IEEE Fourth International Conference on Multimodal Interfaces.*
12. McGurk H. and MacDonald J. 1976. Hearing lips and seeing voices, From *Nature* 264, 746-748
13. Faulkner A. SYNFACE: A Speech-driven Syntetic Face as a Communication Aid for Hearing-Impaired People From *Elsnews*, Autumn 2001

From the Internet:

a. What is a Viseme? *Whatis.com* http://iroi.seu.edu.cn/books/ee_dic/whatis/viseme.htm