THE TELEFACE PROJECT
MULTI-MODAL SPEECH-COMMUNICATION FOR THE HEARING IMPAIRED

Jonas Beskow, Martin Dahlquist, Björn Granström, Magnus Lundeberg, Karl-Erik Spens & Tobias Öhman
(in alphabetical order)
Department of Speech, Music and Hearing, KTH
S-100 44 Stockholm, Sweden.
Tel. +46 8 790 7879, FAX: +46 8 790 7854, E-mail: teleface@speech.kth.se

Abstract

The Teleface Project, a project that aims at evaluating the possibilities for a telephone communication aid for hard of hearing persons, is presented as well as the different parts of the project: audio-visual speech synthesis, visual speech measurement and multimodal speech intelligibility studies. The experiments showed a noticeable intelligibility advantage for the addition of the face information, both for natural and synthetic faces.

INTRODUCTION

It is well known that visual information obtained by speechreading and interpretation of body gestures improves perception of speech, especially in a noisy environment. The amount of visual information from lip reading has been described as a function of the signal-to-noise level [1,2]. The visual information is even more important to persons with a hearing loss.

The Teleface project at KTH focuses on the usage of multimodal speech technology for hearing impaired people. The first phase of the project aims at evaluating the increased intelligibility hearing impaired people might experience from an auditory signal if it is supplemented with a synthesized face. The main focus of this paper will be an intelligibility study.

In the second phase, which has not yet been initiated, we will try to implement a demonstrator of a telephone communication aid for hard of hearing persons. This device will generate a synthetic face that articulates in synchrony with the telephone speech. Control parameters for the face will be extracted from the telephone speech signal. Such a device would support the user with speechreading during any telephone conversation. It should be emphasized that this is in contrast to video telephony, which requires both parties to be equipped with compatible video telephone hardware. This visual hearing aid could be implemented as software running on a PC, or as a dedicated stand-alone unit (the "Teleface" unit).

Our research on multimodal speech synthesis is also targeted at speech-based user-interfaces and spoken dialogue systems. Talking animated agents, employing visual speech synthesis, have been used in recent spoken dialogue system research projects such as Waxholm [3] and Olga [4]. When implemented as real life application, such systems will also prove useful to hearing impaired persons.

AUDIOVISUAL SPEECH SYNTHESIS AND ANALYSIS

The project's different stages involve different kinds of processing of acoustic and visual speech data. Parametrically controlled synthetic visual speech is used in the intelligibility study and will also form the basis for the intended telephone conversation aid of Phase Two of the project. Automatic extraction of facial parameters from the acoustic signal require extensive analysis of the relationship between the facial parameters and the acoustics. To this end, we have built a framework for automatic measurements of visual speech movements [5]. In the intelligibility study, we utilize a rule-based audiovisual text-to-speech-synthesis framework [6] to generate synthetic acoustic as well as visual speech stimuli.

A set of phonetic rules calculate parameter trajectories from a phoneme string. A formant synthesizer [7] is used to generate synthetic voices. Facial images are generated using a three-dimensional facial model, which is a descendant of Parke's model [8], that includes teeth and a tongue, (Figure 1 left). The model is implemented as a polygon surface that can be articulated and deformed through a set of parameters, rendered with lighting and smooth shading and animated at 25 frames per second on a graphics workstation. Parameters for speech movements include jaw rotation, lip rounding, bilabial occlusion, labiodental occlusion and tongue tip raise.

In the intelligibility study, we use two different face models: in addition to the extended version of the Parke-face [6], there is a cartoon-like female character, developed for an agent-based spoken dialogue system Olga [4]. The Olga character employs a parametrisation technique similar to that of the Parke model and can be controlled using the same set of phonetic rules.

For the optical measurements, a data base of video sequences of a speaker uttering 270 Swedish sentences and 51 VCV words has been recorded. Parts of the speakers face have been marked with a blue colour to facilitate image analysis of lips and other parts of the face that are important for lip-reading. For each frame in the video (25 frames per second), the following data were automatically extracted from the front view of the speaker: upper lip area, lower lip area, mouth opening area, lip width, mouth opening width, lip height, mouth opening height, mouth circumference,

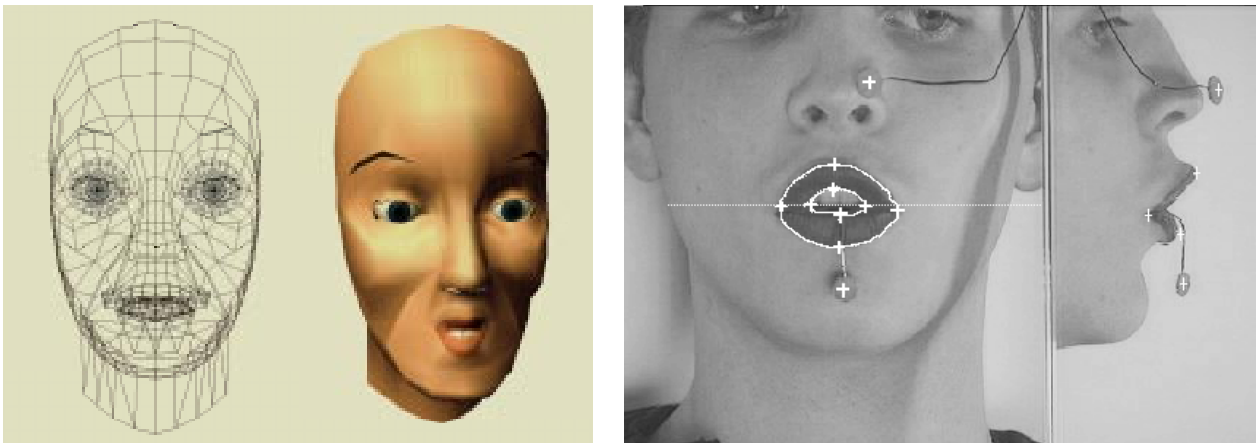

Figure 1: Facial synthesis model (left). Points and borders extracted for the face measurements (right).

outer lip circumference, and jaw opening (using fix points for the mandible and the skull), and from the profile of the speaker: lip area, protrusion of upper and lower lip, and displacement of the mandible in the anterior-posterior direction (Figure 1 right).

The resulting 14 trajectories will be mapped onto the parameter set used in the generation model and then statistically analysed together with the acoustic signal, providing knowledge about the relationship between the visual and acoustic modes of speech. The optical measurements will also be used to improve naturalness of the visual speech synthesis.

INTELLIGIBILITY STUDY

We have created a database of video recordings of a male speaker's face pronouncing Swedish VCV-words and everyday sentences. The audio track has been separated from the video recordings and phonetically labelled. By processing the label file through a rule system for audio-visual text-to-speech synthesis [6], parameter trajectories for face model animation as well as formant synthesizer control have been calculated. In the synthesis procedure, all the phoneme durations were copied from the original utterances in the video sequences. The parameter files have been used to generate two different synthetic voices as well as animations of two different synthetic faces. With the natural face and the natural voice, this adds up to three faces and three voices, which are all synchronised with each other when played together. Tests are currently performed in two rounds, the first one with normal hearing subjects and VCV-words, and the second also incorporating hard-of-hearing subjects and everyday sentences. Results below are from Round One and one of the synthetic versions only.

Method

Subjects

In the first test series, the subjects were 18 fourth-year MSc-students at KTH. The test was made as part of a mandatory laboration in the Speech Communication course given by the departement. A screening test was performed to check that all the subjects had normal hearing.

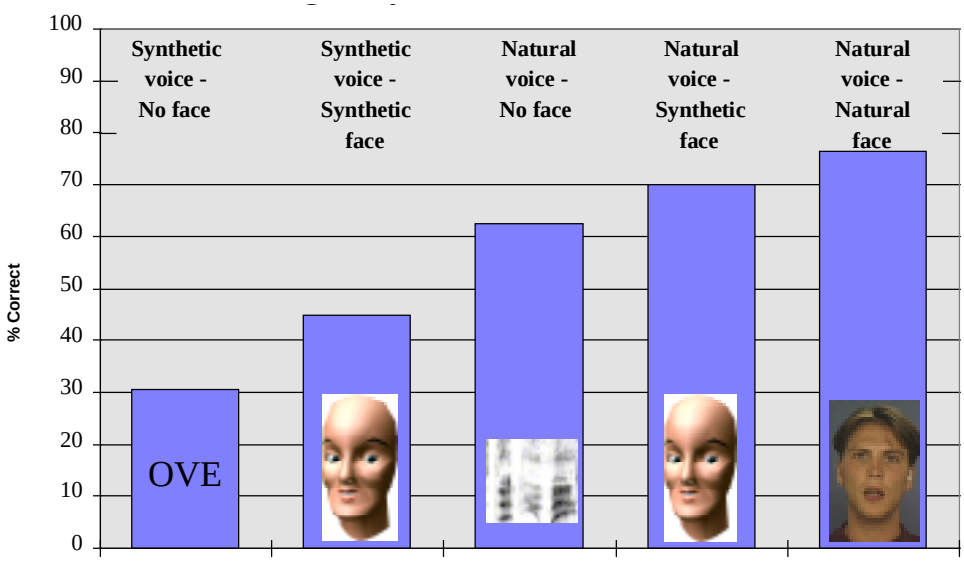

Figure 2. Results from intelligibility tests. Number of correct responses (in %). Average for 18 normal hearing subjects.

	b	d	g	p	t	k	s	ʃ	ç	f	v	m	n	ŋ	j	l	r
b	**24**		1			1	1				9						
d		**24**												1	6	3	2
g			**19**			4									13		
p				**15**	4	8			1	8							
t				3	**22**	4	1	2	2	2							
k				1	2	**32**								1			
s				13	11	2	**5**		1	4							
ʃ							1	**25**	10								
ç							1	19	**16**								
f			1	9	2	1	1	2	2	**18**							
v	4		1			1					**30**						
m											3	**18**	10	2		3	
n												2	**22**	4	1	7	
ŋ											4	5		**26**	1		
j		2	1								1			1	**28**		3
l															8	**28**	
r					1						1				3		**31**
	28	*26*	*23*	*41*	*42*	*53*	*10*	*48*	*32*	*32*	*48*	*25*	*32*	*35*	*60*	*41*	*36*

	b	d	g	p	t	k	s	ʃ	ç	f	v	m	n	ŋ	j	l	r
b	**36**																
d		**23**	2				1								5	5	
g			**25**											1	10		
p				**38**		1										1	
t			2	1	**21**	6	2	3		1							
k				1		**35**											
s					20	2	**6**	4	1	4						1	
ʃ					3		3	**23**	7								
ç								19	**13**								
f				2						**34**							
v	1	1									**34**						
m											1	**29**	1	3			
n													**14**	6	2	13	1
ŋ			1				1					1	2	**26**	3		2
j		1	8				1	1						1	**22**		2
l															11	**24**	1
r		1	1				1							2	3	3	**25**
	37	*26*	*39*	*42*	*44*	*44*	*15*	*50*	*21*	*39*	*35*	*30*	*17*	*39*	*56*	*47*	*31*

	b	d	g	p	t	k	s	ʃ	ç	f	v	m	n	ŋ	j	l	r
b	**33**										1						
d		**23**	1						1		2				4	1	1
g			**20**			4								1	8		
p				**34**													
t					**27**	5			1	1							
k					2	**30**		2									
s					19		**4**	6	2	1	1						
ʃ					1		2	**25**	6								
ç								18	**16**								
f										**34**							
v											**33**						
m											2	**30**		1			
n													**28**		5	1	
ŋ													4	**24**	4	1	1
j		1	3					1	1					1	**22**	1	4
l													1		7	**24**	1
r		1												1	2		**29**
	33	*25*	*24*	*34*	*49*	*39*	*6*	*52*	*27*	*36*	*39*	*30*	*33*	*28*	*52*	*28*	*36*

Figure 3. Confusion matrices. Natural voice accompanied with - from top to bottom - no face, synthetic face and natural face. Stimuli on the vertical axis and response on the horizontal axis. Bottom line shows total number of responses.

Stimuli

In the first test series we used lists consisting of VCV-words with 17 Swedish consonants, /b, d, g, p, t, k, s, ʃ, ç, f, v, m, n, ŋ, j, l, r/, in symmetric context with the vowels /a/ and /ʊ/. When performing tests with this material on normal hearing persons, the audio signal was degraded by adding white noise. The signal-to-noise ratio in these tests was 3 dB. Each subject performed with 8 different combinations of voices and faces.

Procedure

The tests were performed in a computer-based test environment [9]. This gave us the opportunity to play video sequences of the faces with sound files of the voices. In this way, it was possible to evaluate the intelligibility of different audio-visual combinations. A monitor was used for presenting the visual stimuli and a loud-speaker for the audio. A forced choice response for the VCV-words was made using the mouse on the computer screen presenting all consonants in the stimuli set. There was no time limit for the response.

Results

Data from the tests were analysed using confusion matrices and feature analysis. Results from a subset of the eight combinations are presented here. Overall results are shown in Figure 2. Adding a synthetic face to a natural male voice improves the correct response rates from 63% to 70%. The corresponding result when adding a natural face is 76%. Synthetic male voice gave 31% correct responses compared to 45% with a synthetic face added.

Confusion matrices for normal hearing subjects (Figure 3) show that a number of confusions are reduced by adding a synthetic face to the natural voice. Results for bilabials and labiodentals are equally well improved for synthetic and natural faces.

Negative effects from the synthetic face such as the increased tendency for /n/ being identified as /l/ can also be found. This type of information may be useful when improving the quality of the synthetic face. Generally the results for /s/, /ʃ/, and /ç/ were poor, which was expected because of the characteristics of the masking noise.

Results for the consonants in the context of /a/ was generally better than in the context of /ʊ/, both acoustically and visually.

The proportion of correct responses with respect to place of articulation (‘total’ in Figure 4), increased from 72% for natural voice only, to 83% with a synthetic face and to 86% with a natural face. Figure 4 shows that the highest improvement for bilabials and labiodentals is achieved when adding a synthetic or natural face. This is hardly surprising, since they are two of the most salient visemes [10]. The effect is enhanced by their poor voice-only scores. Dentals showed almost no difference between the audio and audiovisual conditions. Palatal and velar consonants did not benefit at all from adding a face to the natural voice condition. This is not surprising considering the high voice-only score in combination with the back articulatory movements.

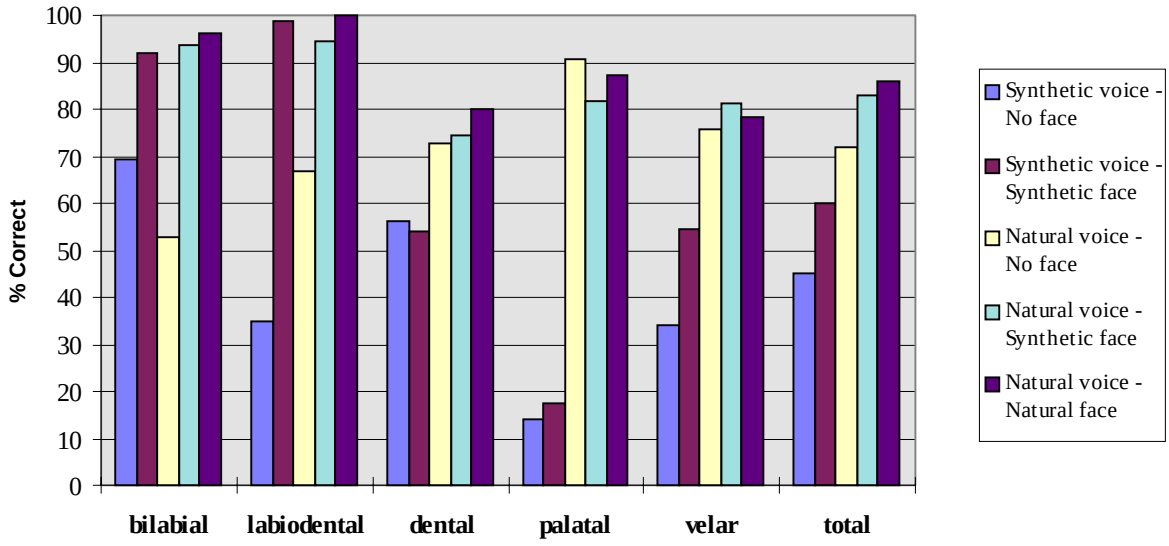

Figure 4. Results from feature analysis of intelligibility tests. Correct responses (in %) with respect to five places of articulation: bilabials /p, b, m/, *labiodentals* /f, v/, *dentals* /t, d, s, l, r, n/, *palatals* /j, ç, ʃ/, *and velars* /k, g, ŋ/. *The overall result with respect to place of articulation is shown as 'total'. Average for 18 normal hearing subjects.*

CONCLUSIONS AND FUTURE WORK

The results give a good illustration of the complementary nature of multimodal speech. The visual mode is most important for the front consonants (bilabials and labio-dentals received low auditory but high audio-visual score), whereas acoustic-only information is most important for back consonants. Results show that a synthetic face substantially improves the intelligibility of synthetic and natural speech. Further tests will be made with hard-of-hearing persons as subjects. While the use of VCV-words serves a phonetic analytical purpose, it may not tell the whole truth about the visual support in a communicative situation like the intended Teleface application. Therefore tests with a material of short everyday sentences will also be performed. The work with visual articulatory measurements is in progress, and the results from these, along with the results from the intelligibility studies, will serve as a basis for improving the audio-visual speech synthesis.

ACKNOWLEDGEMENTS

The Teleface project is supported by a grant from KFB, KommunikationsForskningsBeredningen, the Swedish Transport and Communications Research Board.

REFERENCES

[1] Sumby, W. H. and I. Pollack (1954). "Visual contributions to speech intelligibility in noise", *Journal of the Acoustical Society of America* 26: 212-215.

[2] Benoît, C., Guiard-Marigny,T., Le Goff, B. And Adjoudani, A. (1996). "Which Components of the Face Do Humans and Machines Best Speechread?", In *Speechreading by Humans and Machines*, edited by D.G. Stork and M.E. Hennecke. Springer-Verlag.

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

[4] Beskow, Elenius & McGlashan (1997). "Olga - A dialogue system with an animated talking agent", In *Proceedings of Eurospeech '97*, Rhodes, Greece.

[5] Öhman, T. (1997). "Measuring visual speech", In *Proceedings of Fonetik 97*, Umeå, Sweden.

[6] Beskow, J. (1995). "Rule-based Visual Speech Synthesis", *Proceedings of Eurospeech '95*, Madrid, Spain.

[7] Carlson, R., Granström, B., Karlsson, I. (1991), "Experimentswith voice modelling in speech synthesis", Speech Communication 10, pp 481-489.

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

[9] [Lundeberg, M. (1997). "Multimodal talkommunikation - Utveckling av testmiljö", *Master of science thesis (in swedish)*. TMH-KTH, Stockholm, Sweden.

[10] Jeffers, J., Barley, M. (1971). "Speechreading (Lipreading)", C.C. Thomas.