SYNTHETIC FACES AS A LIPREADING SUPPORT

Eva Agelfors, Jonas Beskow, Martin Dahlquist, Björn Granström, Magnus Lundeberg, Karl-Erik Spens and Tobias Öhman (in alphabetical order)

Centre for Speech Technology (CTT),
Department of Speech, Music and Hearing, KTH

ICSLP98

SE-100 44 Stockholm, Sweden
email: teleface@speech.kth.se

ABSTRACT

In the Teleface project the possibility to use a synthetic face as a visual telephone communication aid for hearing impaired persons is evaluated. In an earlier study, NH, a group of normal hearing persons participated. This paper describes the results of two multimodal intelligibility tests with hearing impaired persons, where the additional information provided by a synthetic as well as a natural face is evaluated.

In a first round with hearing impaired persons, HI:1, twelve subjects were presented with VCV-syllables and ”everyday sentences” together with a questionnaire. The intelligibility score for the VCV-syllables presented as audio alone, was 30%. When adding a synthetic face the score improved to 55% and when instead adding the natural face it was 58%. In a second round, HI:2, fifteen hearing impaired persons were presented with the sentence material and a questionnaire. The audio track was filtered to simulate telephone bandwidth. The intelligibility score for the audio only condition was 57% correctly identified keywords. Together with a synthetic face it was 66% and with a natural face 83%. Answers in the questionnaires were collected and analysed. The general subjective rating of the synthetic face was positive and the subjects would like to use such a type of aid if available.

1. INTRODUCTION

At KTH our work with multimodal speech synthesis started with a rule-based audio-visual text-to-speech-synthesis framework [1], developed in 1995. In projects such as Waxholm [2] and Olga [3] talking animated agents were using visual speech synthesis. Another project that uses this technology is the ongoing August project; a dialogue system with a user interface that is multimodal both in its input and output (see www.speech.kth.se/august). Employing multimodal speech input and output in dialogue systems increases both the user's intelligibility and the recognition rate of the speech recognition system.

The Teleface project focuses on the usage of multimodal speech technology for hearing impaired persons. The aim of the first phase of the project is to evaluate the increased intelligibility hearing impaired persons experience from an auditory signal when it is complemented by a synthesised face. We are also interested in the difference between a synthetic and a natural face from a lipreading point of view. A demonstrator of a system for telephony with a synthetic face that articulates in synchrony with a natural voice will be implemented in phase two of the project.

2. SYNTHESIS AND ANALYSIS OF VISUAL SPEECH

The project's different stages involve different kinds of processing of acoustic and visual speech data. In the intelligibility studies, we utilise a rule-based audio-visual text-to-speech-synthesis framework to generate synthetic acoustic, as well as visual, speech stimuli. A set of phonetic rules generates parameter trajectories from a phoneme string. A formant synthesizer is used to generate synthetic voices [4]. Facial images are generated using a three-dimensional facial model, which is a descendant of Park's model [5], augmented with teeth and a tongue. The model is implemented as a polygon surface that can be manipulated and deformed through a set of parameters, rendered with lighting and smooth shading and animated at 25 frames per second on a UNIX workstation. Parameters for speech movements include jaw rotation, lip rounding, bilabial occlusion, labiodental occlusion and tongue tip raise. This parametrically controlled visual speech synthesis will also form the basis for the intended telephone conversation aid in phase two of the project.

Automatic extraction of facial parameters from the acoustic signal requires extensive analysis of the relationship between the facial parameters and the acoustics. To this end, we have built a framework for automatic measurements of visible speech movements [6]. A database of video sequences of a male speaker has been recorded. The speaker is a Swedish male from Stockholm. 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 speech reading. The parameters from the optical measurements are being statistically analysed together with the acoustic signal, providing knowledge about the relationship between the visual and acoustic modes of speech.

The difference in intelligibility for a natural face and a synthetic face is being analysed and measurements of the speaker's face will be used to find parameters that need to be refined or features that should be modelled to improve the synthesis.

3. INTELLIGIBILITY TESTS

A further audio visual speech database of video sequences, used to set up the intelligibility tests, has been recorded. This database is identical to the database that we use for the visual measurements, but without colours and markers in the face. The database consists of two parts. The speech material of the first part consists of 153 hyper articulated VCV-syllables, with the vowels V={a, ʊ, ɪ} symmetrically surrounding the consonants C={b, d, g, p, t, k, s, ʃ, ç, f, v, m, n, ŋ, j, l, r}. The second part of the database consists of 270 normally articulated Swedish "every day sentences". The test lists were developed at TMH by Öhngren based on MacLeod and Summerfield [7] (1990). During recording of the database, the speech rate was kept constant by prompting the speaker with text-to-speech synthesis set to normal speed.

3.1. VCV-Syllables

A previously reported [8] preliminary test, NH, was performed with normal-hearing subjects. The subjects were 18 fourth-year-students in engineering at KTH. The audio signal was degraded by adding white noise. Three test lists of (3 x 17 stim./list) were presented in different conditions for the subjects (2 audio-visual and 1 audio-alone). Subjects were asked to respond with the consonant. Responses for the VCV corpus were forced-choice and made with a graphical interface on a computer screen. The results for the test, using only the /a/ and /ɪ/ surroundings, show that adding a synthetic face to a natural male voice increases correct responses from 63% to 70%. Corresponding result for adding a natural face is 76% (Figure 1).

In the first test round with hearing impaired persons, HI:1, the twelve subjects had a mean hearing loss of 88.4 dB hearing level (HL), (range 62-103 dB). They were between 23 and 76 years old with a median age of 57 years. The subjects are experienced hearing-aid users and were allowed to adjust their hearing aid to their most comfortable listening level during a training session. The score for the natural voice only was 30% correctly identified syllables. When adding a synthetic face the score significantly improved to 55%, and with a natural face it was 58%. (Figure 1).

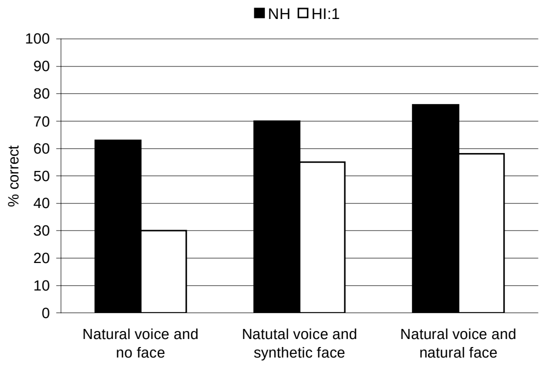


Figure 1: Mean VCV-syllable scores for the normal hearing subjects, NH, and for the first round with hearing impaired persons, HI:1.

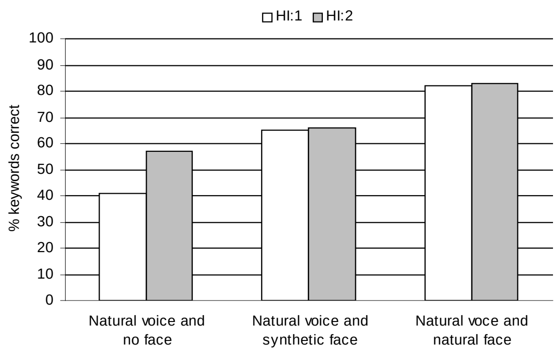
3.2. Sentences

Apart from VCV-syllables, sentences were used as stimuli in the first test with hearing impaired persons, HI:1. In this part of the test, the performance was measured as the percentage of correctly repeated keywords. The sentences were organised in lists with 15 sentences in each list and three keywords per sentence. Six test lists were presented; two test lists in each condition (2 audiovisual and 1 audio alone). The responses were given verbally. The score for the subjects in the auditory alone condition (natural voice) was 41% (standard deviation, SD=26). The intelligibility score increased to 65% (SD = 24) when adding a synthetic face, and to 82% (SD = 11) for a natural face added. In a second round with hearing impaired subjects, HI:2, 12 new subjects participated and three subjects, who also were tested in HI:1 four months earlier, were retested. The median age of this group was 54 years (range 37-82), and the subjects mean hearing loss is 83.2 dB HL (range 32-113 dB). Apart from the subjects, HI:2 differed from HI:1 in the following ways:

- Only the sentence corpus was used.
- The order of the sentence lists was changed compared to HI:1.
- The audio signal was filtered to telephone bandwidth in order to get the test conditions closer to those of the intended Teleface application.
- Visual distraction was introduced as one condition of the test without the subject's knowledge. This was done by displaying the synthetic face with articulatory movements controlled by the output of a simple phoneme recogniser not trained for our purpose.
- A visual-only condition was introduced. This had not previously been tried with the synthetic face.

The mean scores for HI:2 was 57% (SD = 38) correctly identified keywords for the audio alone condition, 66% (SD=33) with the synthetic and 83% (SD = 23) with the natural face. (Figure 2). When the synthetic face was controlled by a phoneme recogniser, intelligibility score dropped to 51%, i.e. visual distraction caused intelligibility to fall 6% below the audio alone condition. When showing only the synthetic face, the average score was only 3 %. For the natural face without audio, the result was 16%.

Figure 2: Mean sentence scores for the two groups of hearing



impaired persons, HI:1 and HI:2.

There was no significant difference between the results in HI:1 and HI:2 for the retested subjects. The characteristics of their hearing loss, implies that a telephone bandwidth filtering of the speech signal does not make much difference in intelligibility.

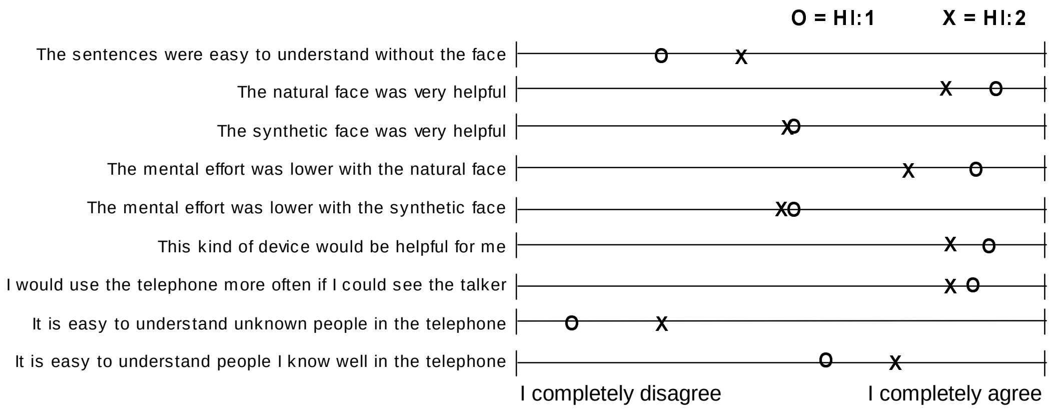


Figure 3: Questions from the questionnaire and corresponding rating by the subjects in HI:1 (O) and HI:2 (X) on an open scale.

Therefore it is possible to compare the results from the first and the second round. However, comparing individual results between the two groups should be done carefully since the subjects' individual differences in hearing loss and lipreading skills are large. Some subjects had a result with the synthetic face very close to the result with the natural face. On the other hand one subject with a profound hearing loss seemed to get very little information from the synthetic face although the she obtained 82 % correctly identified keywords with the natural face. The subjects were not familiar with neither the visual speech synthesis nor the human speaker and the motivation to lipread the synthetic face was different for the different subjects. Between the two test lists, a learning effect was observed for the synthetic, but not for the natural face.

4. QUESTIONNAIRE

In HI:1 and HI:2, the subjects were asked to complete a questionnaire concerning their subjective responses to the synthetic and natural face. They were asked to rate a number of questions, phrased as statements, on an open scale with the end markings: *I completely disagree* and *I completely agree*, where the subjects could indicate to what extent they agreed or disagreed with a given statement (Figure 3). The questions are about telephone use and the benefit of the natural and the synthetic face in a multimodal test condition. The objective of the questionnaire was to compare the intelligibility rates to the subjective ratings and to find out whether hearing impaired persons would welcome and use a device such as the intended demonstrator.

4.1. Subjective Benefit and Need

The mean results rated by the subjects are shown in Figure 3 together with the corresponding questions. In general, the subjects thought that it was hard to perceive auditory stimuli without lipreading support, and both the natural and the synthetic face was helpful for them. About the questions concerning telephone use, they responded that familiar people are easier to understand than unfamiliar and that they would use the telephone more often if they could see the talker. In general, they thought that a device such as the intended artificial lipreading support would be helpful.

4.2. Comparing Subjective Rating and Objective Results

The mean subjective results (Figure 3) showed a positive benefit for the two faces, but the natural face was rated to a higher degree of benefit. The differences in the results between the two groups of subjects regarding the benefit of the natural face (see Figure 3) could be explained by the fact that several subjects in HI:2 performed very well in the audio only condition. They probably relied on their hearing and therefore the natural face was not helpful for them in this test situation (Figure 4b & 4d).

The answers from the questionnaire showed that the subjective opinions about the synthetic face (*The synthetic face was very helpful*, Figure 3) matches the objective results for the subjects with a profound to severe hearing loss (Figure 4). The lesser the degree of hearing loss, the larger the variation between the subjective and objective result. In HI:2, some of the stimuli were not synchronised, because of the recognition errors. It is likely that this lowers their confidence and also their subjective rating of the benefit.

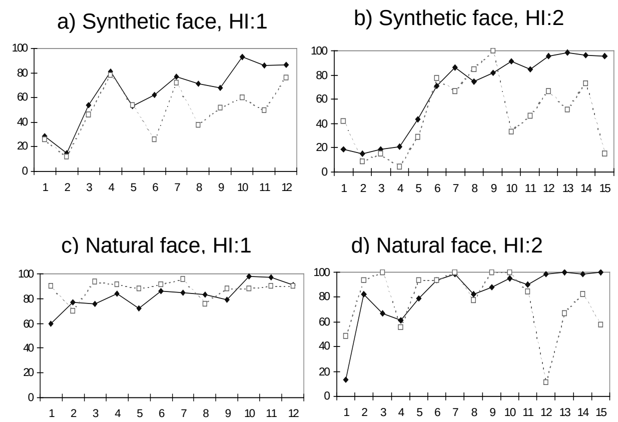


Figure 4a-d: The relation between individual subjective benefit rating of the faces (dashed lines) and audiovisual sentence scores (solid lines). The y-axis shows rating (100 % maximum) and % correctly identified keywords, respectively. The numbers on the x-axis corresponds to the subjects, ordered after their performance in the audio only test condition.

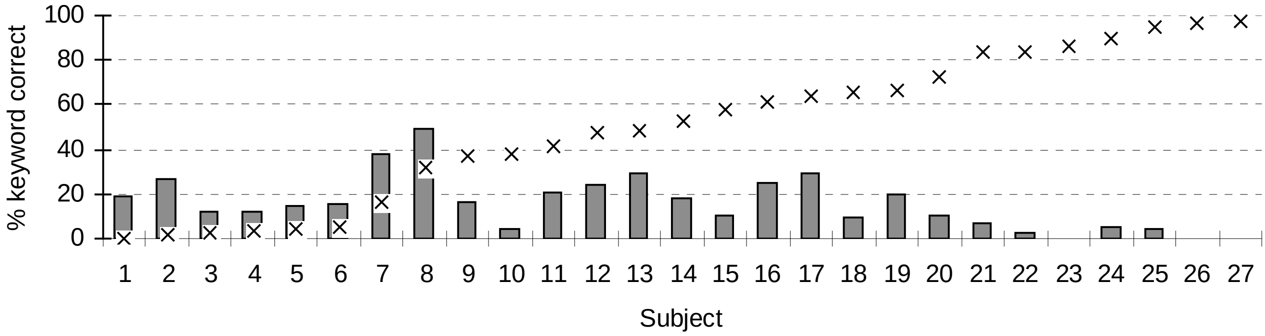


Figure 5: The bars show the contribution of the synthetic face when added to the natural voice in relation to the performance on the audio-only condition (x) for the subjects HI:1 and HI:2. The subjects are presented in order of increasing performance on the audio-only condition of the test.

5. CONCLUSIONS AND FUTURE WORK

The highest absolute benefit from lipreading the synthetic face was obtained by the subjects with such a hearing loss that they reach scores between 40% and 80% correctly identified keywords in the audio only condition in our experiment. These are the ones that seem to have the best use of a visual hearing aid like the intended Teleface application. Subjects with a lower audio-only score often get a good gain from adding a synthetic face to the natural voice, but although the gain is high, they will probably not benefit enough in a telephone situation, since their starting point is too low. Subjects with a score higher than approximately 80 % correctly identified keywords seem to rely very much on their hearing and therefore they do not gain very much from adding a face, synthetic or natural, in our experiment (Figure 5). Depending on a number of individual factors, people with different degrees of hearing loss, ranging from normal hearing to severe hearing impaired persons, will make up this target group.

The subjective ratings of both the need and the benefit of the synthetic face were high. Especially persons in the target group rated the artificial face high. Spontaneously, they were also positive to the synthetic aid. Many of them expressed a desire for such a telephone aid if available. Individually, adding the faces often decreased the mental effort, regardless if there was any increase of intelligibility.

The learning effect that we found for the synthetic face between the first and the second list is promising. Since the articulatory movements are much more consistent than those found in a natural face, a subject could probably learn to get more information out of the synthetic face after a longer period of training. This learning effect was not found for the natural face, where the performance was at the same higher level.

The result for the first phase of the project presupposes a perfect mapping from acoustics to facial gestures. Preliminary studies in the second round of phase one, with a simple phoneme recogniser not trained on our database show that displaying misleading articulatory movements decreases the intelligibility. Our research in this area is now focused on how to train a speech recogniser for the special needs of the intended application.

6. ACKNOWLEDGEMENTS

The work is funded by KFB, the Swedish Transport and Communications Research Board, and CTT, the Centre for Speech Technology, KTH.

7. REFERENCES

1. Beskow, J. “Rule-based Visual Speech Syntheses”, *Proceedings of Eurospeech ’95*, Madrid, Spain, 1995.

2. 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. “The Waxholm system – a progress report”, *Proceedings of Workshop on Spoken Dialogue Systems*, Vigsø, Denmark, May 1995.

3. Beskow, J., Elenius, K. and MacGlashan, S. “Olga - A dialogue system with an animated talking agent”, *Proceedings of Eurospeech ’97*, Rhodes, Greece, 1997.

4. Carlson, R., Granström, B., and Hunnicutt, S. “Multilingual text-to-speech development and applications”, Ainsworth, A. W. (Ed.), *Advances in speech, hearing and language processing*, JAI Press, London, UK, 1991.

5. Parke, F. I. “Parametrized models for facial animation”, *IEEE Computer Graphics*, 2(9), pp 61-68, 1982.

6. Öhman, T. “An audio-visual speech database and automatic measurements of visual speech”, *TMH-QPSR, KTH*, 1/1998.

7. MacLeod, A., and Summerfield, Q. “A procedure for measuring auditory and audiovisual speech-reception thresholds for sentences in noise. Rationale, evaluation and recommendations for use", *British Journal of Audiology*, 24:29-43, 1990.

8. Beskow, J., Dahlquist, M., Granström, B., Lundeberg, M., Spens, K-E., and Öhman, T. “The Teleface project - Multimodal Speech Communication for the Hearing Impaired”, *Proceedings of Eurospeech ’97*, Rhodos, Greece, 1997.