Umeå University, Department of Philosophy and Linguistics
PHONUM 9 (2003), *61-64*
Available online at http://www.ling.umu.se/fonetik2003/

The SYNFACE project – a status report

Inger Karlsson

Department of Speech, Music and Hearing, KTH

The status and progress of the SYNFACE project is described. SYNFACE is a European project with partners in three countries. The project aim is to produce a talking face telephone. The present report is concentrated on the work performed by the group at KTH during the first half of the project.

1. Introduction

SYNFACE is a European project that aims at developing a talking face telephone that can assist hard of hearing people in their use of an ordinary telephone. The project partners come from the Netherlands, Great Britain and Sweden. SYNFACE prototypes will be developed for the three languages Dutch, English and Swedish. See also information on the project homepage http://www.speech.kth.se/synface

This report describes work performed at KTH during the first half of the three-year project. The main tasks have been to develop automatic phoneme recognition methods that will provide results with only a very short time delay and to improve the articulation of a talking face to facilitate lip-reading. The output from the phoneme recogniser will decide the movements of the talking face. The face articulation should be synchronised with the delayed speech signal to facilitate lip-reading.

Two different perception tests have been performed. A multilingual test was run to prove the gain in understanding that SYNFACE can give. A test of different delays between audio and visual signals was performed to learn about sensitivity to visual delay. The visual synthesis has been improved using more recorded data. Two recognition methods have been evaluated keeping in mind the special demands of SYNFACE.

2. Multilingual perception test

Multilingual perceptual studies have been performed by three of the SYNFACE partners. The aim of the tests was to characterise the potential gain in intelligibility derived from a synthetic talking head controlled by phonetically transcribed speech. Speech materials were simple Swedish, English and Dutch sentences. Speech was degraded to simulate severe-to-profound hearing impairment. Degradation was produced by vocoder-like processing using either two or three frequency bands, each excited by noise. 12 native speakers of each of the three languages took part in intelligibility tests in which each of the two degraded auditory signals were presented alone, with the synthetic face, and with a natural video of the face of the original talker. Intelligibility in the purely auditory conditions was low (7% for the 2-band vocoder and 30% for the 3-band vocoder). The results are shown in Figure 1. The average intelligibility increase for the synthetic face compared to no face was 20%, and was

statistically highly reliable ($p<0.001$). The synthetic face fell short of the advantage of a natural face by an average of 18% (Siciliano, Williams, Beskow, Faulkner, 2003).

2.1. Swedish test

During the Swedish test the subject was placed in front of a screen and a loudspeaker in a sound proofed room. The subject was told that he/she would hear ordinary sentences. After each sentence the subject was asked to repeat what he/she heard. The test leader sitting outside the sound proofed room wrote down the answer and then the next test sentence was played. The test was divided into two sessions of equal duration. Each session contained all six stimulus conditions in randomised order. All stimulus conditions contained 12 sentences. The time interval between the test sessions was normally at least a day, for one subject it was only half a day. The first test session began with a training session where six sentences for each stimulus condition were presented starting with the natural face and 3-band vocoder speech and ending with audio only and 2-band vocoder. The test subject was not informed about the results of the training session. 5 grown-up women and 7 grown-up men participated; all were normal-hearing, had normal or corrected-to-normal eyesight and had Swedish as their mother tongue. None had previously performed any speech reading tests including a synthetic talking face, two had earlier (more than 20 years ago) experiences of speech reading tests.

The sentences were unrelated everyday Swedish sentences developed at KTH by G Öhngren (Öhman 1998) and were three to eight words long. Each sentence contains three keywords. Each sentence occurred only once. The test material was read by a normal male speaker.

The number of correctly repeated keywords was expressed as percent correct of keywords presented. The results are given in Figure 1. The difference between natural and synthetic face is smaller for Swedish than for the other languages. This may depend on that the face synthesis was developed for Swedish and only recently adapted to English and Dutch.

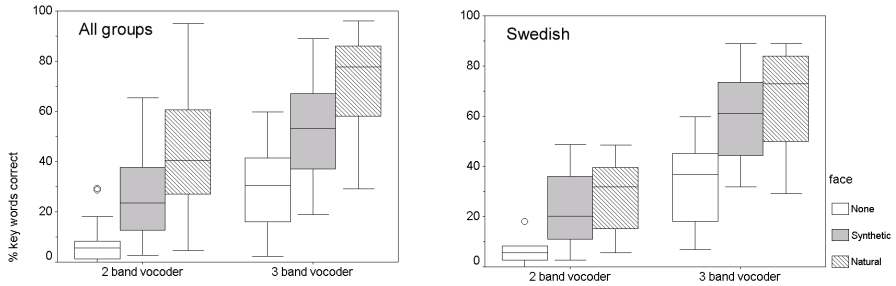

Figure 1. Sentence intelligibility for degraded speech with synthetic and natural speech reading cues. The box and whisker plots show the median (bar), interquartile range (box), full range excluding outliers (whiskers) and outlying values (o – outliers are values outside the interquartile range by more than 1.5 times the interquartile range). Left is the data for all three groups together, right is the results for Swedish (from Siciliano et al., 2003).

3. Delay test

Two different types of delays may occur in the SYNFACE system: 1) the audio and visual signals can be non-synchronous, and 2) both audio and visual signals can be delayed. The

amount of delay of the audio signal relative to the visual signal that is acceptable to a listener has been tested in a thesis project performed by Molander (2003). She used the 3-band vocoder material described above with both synthetic and natural face and delayed the audio and visual signal in relation to each other. The audio only condition was used as reference. For the synthetic face the visual signal was lagging 0, 50, 175 and 300 ms, for the natural face the visual signal was both leading and lagging these amounts. The tests were performed using the same procedure as in the perception test described above. 12 subjects were tested. The results from the tests are summarised in Figure 2. The results that are most relevant for SYNFACE are that the results for 0 and 50 ms delays are very similar and that even with a visual delay of 175 ms the subjects performed better than with only the audio signal.

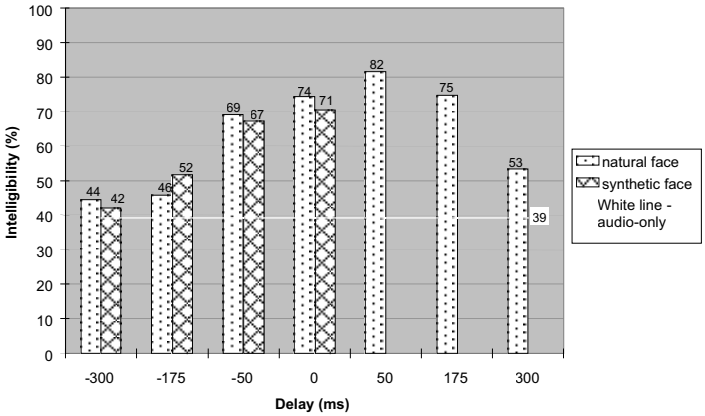

Figure 2. Perception results from the delay test (Molander 2003)

4. Visual speech synthesis

Audio-visual speech databases have been recorded for the three project languages using a optical motion tracking system Qualisys (http://www.qualisys.se) with four IR cameras. The system tracks about 30 small reflectors (4 mm diameter) glued to the subject's jaw, cheeks, lips, nose and eyebrows and a pair of glasses (to serve as reference for head movements) and calculates their 3D-coordinates at a rate of 60 frames per second. The procedure is described more fully in (Beskow, Engwall, Granström, 2003).

In a recent study (Beskow, 2003), data recorded in this way has been used to investigate alternative synthesis methods. Automatic extraction of face articulation parameters for visual speech synthesis from Qualisys recordings has been obtained using frame-by-frame minimisation of error between measured points and face model, yielding a set of "optimal" control parameter trajectories. Using this data, four coarticulation models have been implemented and trained. Two of them are based on previously described coarticulation models from speech production theory and two of them are based on artificial neural networks (ANNs). The models have been evaluated objectively (by comparing RMS error between target and prediction) as well as perceptually through audiovisual sentence intelligibility testing.

5. Speech recognition

A crucial part of the SYNFACE system is the speech recognition. The recognition must occur with as little delay as possible. As found in the delay test, the visual signal should lag less than about 175 ms compared to the audio signal if the face articulation shall be of help (Molander 2003). Delaying the total signal should not to exceed 200/300 ms in order to avoid communication problems on the phone causing e.g. mutual silence, doubletalk. Kitawaki (1991) found that transmission delays in the range of 500 ms give considerable subscriber difficulties in telecommunication.

Speech recognition based on word recognition is accordingly too slow. Instead phoneme recognition is employed. Two alternative methods for phoneme recognition have been investigated: 1. SYNFACE I using HMM and 2. SYNFACE II using ANN (Salvi, 2003). In SYNFACE I the delay can be varied. The influence of the delay time was tested by (Pihl 2003). He found that the delay needs to be at least 100 ms to give recognition results comparable to a standard recogniser. SYNFACE II has in tests proved to give equivalent results with a delay of only 10 ms. As there will be other delays as well in the complete system, the SYNFACE II recogniser using neural nets has been chosen as the main recogniser for the SYNFACE prototype. This prototype will be tested with hard-of-hearing users in three countries, the Netherlands, UK and Sweden during the next year.

6. Acknowledgements

The IST-2001-33327 SYNFACE project is financed by the European Union (EU) under the FP5 IST Programme: Systems and Services for the Citizen. The work described in this paper has been performed by E Agelfors, J Beskow, B Granström, M Molander, E Pihl, G Salvi, A Seward, K-E Spens and the author at KTH-TMH.

We thank B Lyberg at Linköping University for making their Qualisys system available.

7. Reference.

Beskow, J. (2003) *Talking Heads - Models and Applications for Multimodal Speech Synthesis*. Doctoral Thesis, KTH-TMH

Beskow, J., Engwall, O. & Granström, B. (2003) Simultaneous Measurements of Facial and Intraoral Articulation. *Proceedings of Fonetik 2003 (this volume)*

Kitawaki, N. (1991) Pure delay Effects on Speech Quality in Telecommunications. *IEEE Journal on selected areas in communications*, 9, 586-593.

Öhman T., (1998) An audio-visual speech database and automatic measurements of visual speech. *TMH-QPSR, vol. 34/1-2, pp. 61-76*

Molander, M. (2003) *Experiment with asynchrony in multimodal speech communication*. MSc Thesis report. KTH-TMH

Pihl, E. (2003) *Bottlenecks in the SYNFACE telephone*. BSc thesis report. KTH-TMH

Salvi, G. (2003) Truncation error and dynamics in very low latency phonetic recognition. *ISCA workshop on Non-linear Speech Processing*, 2003.

Siciliano, C., Williams, G., Beskow, J. & Faulkner, A. (2003) Evaluation of a Multilingual Synthetic Talking Face. as a Communication Aid for the Hearing Impaired. *Proceedings of 15th International Congress of Phonetic Sciences*