SYNFACE – a talking face telephone

Inger Karlsson[1], *Andrew Faulkner*[2], *Giampiero Salvi*[1].

[1]Department of Speech, Music and Hearing; KTH, Stockholm
[2]Department of Phonetics, UCL, London
inger@speech.kth.se, andyf@phonetics.ucl.ac.uk, giampi@speech.kth.se

Abstract

The SYNFACE project has as its primary goal to facilitate for hearing-impaired people to use an ordinary telephone. This will be achieved by using a talking face connected to the telephone. The incoming speech signal will govern the speech movements of the talking face, hence the talking face will provide lip-reading support for the user.

The project will define the visual speech information that supports lip-reading, and develop techniques to derive this information from the acoustic speech signal in near real time for three different languages: Dutch, English and Swedish. This requires the development of automatic speech recognition methods that detect information in the acoustic signal that correlates with the speech movements. This information will govern the speech movements in a synthetic face and synchronise them with the acoustic speech signal.

A prototype system is being constructed. The prototype contains results achieved so far in SYNFACE. This system will be tested and evaluated for the three languages by hearing-impaired users.

SYNFACE is an IST project (IST-2001-33327) with partners from the Netherlands, UK and Sweden. SYNFACE builds on experiences gained in the Swedish Teleface project.

1. Introduction

SYNFACE has as its main purpose to increase the possibilities for hearing impaired persons to communicate by telephone. To achieve this goal a talking face controlled by the incoming telephone speech signal is developed, see Figure 1. The talking face will facilitate speech understanding by providing lip-reading support. This method is intended to work with any telephone and is cost-effective compared to video telephony that need compatible equipment at both ends and text telephony that need a relay service where somebody types the spoken message. Furthermore, the participants in the spoken interaction can maintain privacy. Systems for three European languages: Dutch, English and Swedish that work in real time are under development.

SYNFACE builds on experiences gained in the Swedish Teleface project carried out at the Department of Speech, Music and Hearing, KTH [1]. Teleface evaluated the possibilities of using synthetic visual speech in tools for hearing-impaired people. This included a simulation of a telephone communication aid for the hearing-impaired. This non-real-time device generated a synthetic face that articulated in synchrony with the telephone speech using only the information contained in the telephone speech signal. In tests Teleface was proved able to deliver useful visual speech information for profoundly hearing-impaired users [2].

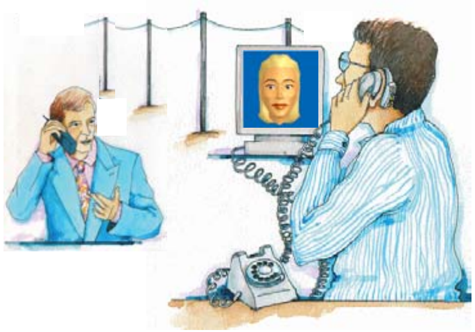


Figure 1. SYNFACE telephone in use, a vision of the future.

The key technological task in SYNFACE is to control a 3-D model of a talking face so as to generate, in real-time, prominent information-bearing oral movements derived from arbitrary acoustic speech signals. While current technology is able to synchronise lip movements with arbitrary speech, the result is neither very natural nor does the visual information contain enough information for lip-readers. Thus two main research areas are addressed in the SYNFACE project. There is a need to

- clearly define the visual speech information requirements of auditory-visual communication,
- develop techniques to derive this information from the acoustic speech signal in very close to real time. The definition of targets for such a speech recognition system requires the development of automatic speech recognition methods that are attuned to the detection of information in the acoustic signal that correlates with prominent and informative facial movements [3]

The project's results will be combined in a multilingual prototype for the generation of articulatory movements of a talking head from live telephone speech. Hearing-impaired users in three countries and languages will evaluate this, in everyday home or workplace situations. The project progress can be followed on the project homepage http://www.speech.kth.se/synface

1.1. Partners and their roles

The department of Speech, Music and Hearing, KTH have provided the project with background knowledge and a prototype system (not real-time) for Swedish gained in the Teleface project [1]. KTH is mainly responsible for the automatic recognition of speech and extraction of articulatory movements and of the synthetic talking face development. Department of Phonetics, UCL, London, is mainly responsible for the definition and evaluation of the visual speech information requirements of auditory-visual communication. The industrial partner Babel-Infovox AB, Stockholm, is responsible for a market survey and for building a SYNFACE prototype. Royal National Institute for Deaf People (RNID), London, UK and Viataal, Sint-Michielsgestel, the Netherlands are both user-based organisations whose main responsibility in the SYNFACE project is the user evaluation of the prototype.

2. Perceptual studies of synthetic face

The face synthesis rules were originally developed in the prior Swedish Teleface project [1]. One activity of the SYNFACE project has been to extend these rules to two other languages, English and Dutch, and to examine the usefulness of synthetic face in these languages when synthesis is driven by ideal data. Sentence intelligibility measures have been obtained in these two languages and in Swedish from normal listeners, and from a group of English and Swedish hearing-impaired listeners [2][4]. Additionally, visual consonant confusion data have been collected from English normal listeners to identify weaknesses in the consonant synthesis rules.

2.1. Sentence Intelligibility

Video recordings of simple everyday sentences from a native adult talker were used: BKB sentences for English [5], the Plomp and Mimpen sentences for Dutch [6] and a translation of the IHR adaptive sentence lists for Swedish [7]. The audio track was processed using a noise-excited vocoder with spectral resolution limited to two or three spectral bands. This processing leads to low auditory intelligibility. Each sentence was phonetically transcribed and automatically aligned with hand-correction. This data was used to drive the face synthesis. The processed audio was recombined with either the original video of the natural talker or with the synthetic face video.

Twelve normally hearing listeners from each language group took part, along with 10 hearing-impaired English listeners (average hearing loss of 86 dB). 24 Swedish hearing-impaired listeners (average hearing loss 86 dB) had performed a similar test earlier [2]. Hearing-impaired listeners heard speech that was unprocessed but telephone-band limited. Normal listeners heard noise-vocoded speech. Intelligibility scores are shown in figure 2.

Both for normal and hearing-impaired listeners, intelligibility with the added synthetic face was always significantly higher than for audio alone. The average improvement in intelligibility was 22% for both hearing-impaired and normal listeners. Except for the Swedish normal listeners, the addition of the natural face gave significantly higher intelligibility than did the synthetic face.

There was a considerable spread of scores in the hearing-impaired group, especially in the sound alone and synthetic face conditions, reflecting a wide range of auditory abilities and of usefulness of the synthetic face. The subjects not being used to synthetic faces could to a part explain this. There was no clear relation between hearing loss and either auditory alone performance or the advantage gained from the addition of the synthetic face. However, as figure 3 shows for the English subjects, the synthetic face was generally more effective for those hearing-impaired listeners with poorer auditory-alone scores, although two listeners with virtually zero auditory-alone scores showed 10% or less intelligibility gain when the synthetic face was added. The results for 24 Swedish hearing-impaired subjects are in agreement with the English data.

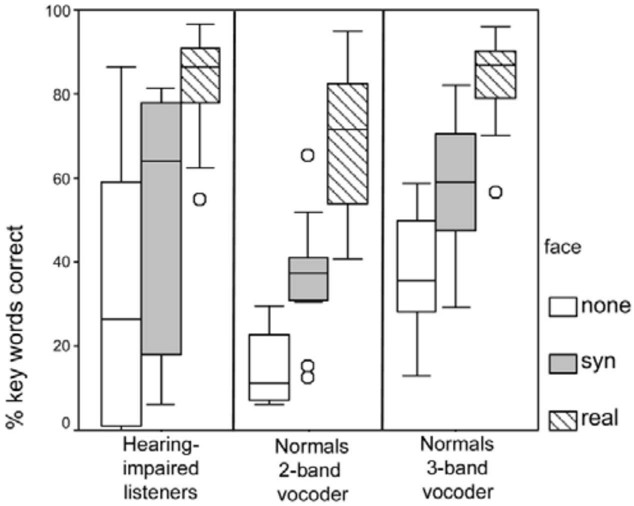


Figure 2: Sentence intelligibility for English hearing-impaired listeners (left), and for Dutch, English, and Swedish normal listeners with degraded audio (center and right). Boxes show interquartile ranges, with median as the vertical bar. Whiskers show the range of scores excluding outliers (circles).

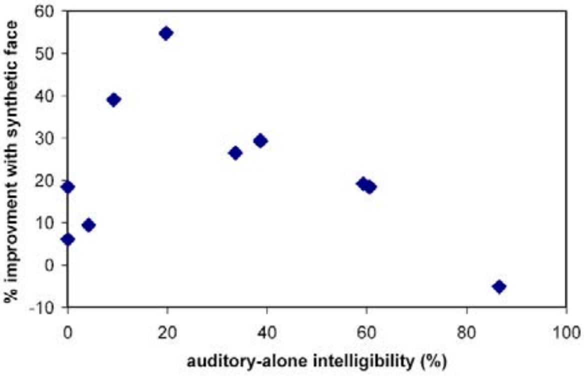


Figure 3: Intelligibility gain with synthetic face compared to auditory-alone scores for English hearing-impaired listeners.

2.2. Consonant information from English synthetic face.

Purely visual identification of consonants in English VCV tokens has also been investigated for all 24 English consonants in three vowel contexts. The synthetic face allowed identification accuracy of 14% compared to 23% for the natural faces (not significant). Place of articulation is the most important aspect of the consonantal information provided by the face, and the current synthesis rules proved significantly poorer than the natural faces in signalling the place of dental, palatal, velar and glottal consonants [8].

2.3. Summary and conclusions from perceptual studies

Both auditory-visual sentence intelligibility and visual consonant data show that the current visual speech synthesis methods provide useful information. Importantly, this information appears to be usable by those hearing-impaired listeners whose purely auditory speech perceptual abilities are rather poor, and who are likely to want to make use of a synthetic face in telephone communication.

3. Visual speech synthesis

. One task for SYNFACE is to develop a synthetic talking face for Dutch and English as well as refine the Swedish synthesis. The perception tests reported in this paper are one part of that. Another is the collection of audio-visual speech databases containing annotated face articulation. Such databases with articulation movement trackings have been recorded for the three project languages. The optical motion tracking is done using a Qualisys system (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 speech material consisted of simple everyday sentences [5],[6],[7] and VCV tokens. The procedure is described more fully in [9].

3.1. New methods for visual speech synthesis

Data recorded with the Qualisys system has been used to investigate alternative face synthesis methods [10]. 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.

4. Speech recognition

A crucial part of the SYNFACE system is the speech recognition. The recognition must occur with as little delay as possible. As has been found in delay tests, the visual signal should lag less than about 175 ms compared to the audio signal if the face articulation shall be of help [11]. 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. Transmission delays in the range of 500 ms give considerable subscriber difficulties in telecommunication [12].

Speech recognition based on word recognition relies on the use of too long speech segments. Instead phoneme recognition is employed. Two alternative methods for phoneme recognition have been investigated:

1. SYNFACE I uses Gaussian mixture HMMs and a two pass decoder with limited look-ahead.
2. SYNFACE II uses recurrent neural networks (RNNs). Here, a two-pass decoder may also be used in a HMM framework, even if with much more limited effects.

Two similar two-pass decoders were independently implemented [13] and will be referred to respectively as DecI and DecII. Both allow the choice of the look-ahead length, corresponding to one of the delays introduced by the recogniser. The influence of this parameter on the performance was tested using DecI and the SYNFACE I method and using the DecII and both SYNFACE I and SYNFACE II [13].

For SYNFACE I results obtained with DecI and DecII are in good agreement. The correct frame rate (CFR) remains stable (41.6% - 41.7%) when the look-ahead length is greater then 100ms. It drops to 34.7% for 10ms look-ahead. For SYNFACE II a similar behaviour is encountered with respect to the look-ahead parameter. In this case, though, the increase in correct frame rate when using the two-pass decoder is negligible if compared to the frame-by-frame MAP decoder, in which the phoneme with highest *a posteriori* probability is selected for each frame. The two-pass decoder, with look-ahead length of 300ms, gives 55% CFR, while the MAP method gives 54.2% CFR. The reason is that the recurrent neural networks in SYNFACE II retain their own representation of the dynamic characteristics of the speech signal, making the decoding step redundant. The great advantage of the RNN/MAP method in this application is, beside the higher accuracy, the fact that results are available with no look-ahead (delay).

The scoring method adopted in the studies computes the number of correctly classified frames of the total number of frames (CFR). This method was chosen instead of the minimum edit distance on the sequences of phonetic symbols (commonly called accuracy), because the alignment of the recognised segments is as important in this application as their correctness. A drawback is that the stability of the result is not taken into account: a recogniser could change hypothesis at every time step, but still provide the highest CFR. In our application though, the visual synthesis interpolation algorithm will filter out the frequent switches between hypotheses sometimes seen with the RNN/MAP method.

5. Prototype

A complete prototype for the target languages, Dutch, English and Swedish is under development. The design of the prototype will reflect the needs of users and market factors that have been established in the project. A graphical telephone interface for SYNFACE has been created as a thesis project for a master’s degree at KTH [14]. The graphical design of the interface can be seen in Figure 4. The interface picks up the incoming speech signal from the telephone and feeds it to the recognition unit. The SYNFACE II method employing neural networks has been chosen as the main recogniser for the SYNFACE prototype, given its properties of speed and low latency processing. With this recogniser the total delay on the data flow (from sound card to visual synthesis, excluding computation time) is estimated to be in the order of 60ms. The speech gestures that the recogniser extracts are combined with the speech signal in the talking face unit. The interface then shows the talking face synchronously with the speech signal. The speech signal can be delayed by the interface to allow the system some processing time. The interface also handles the connection between the user and the telephone net and allows the user to build up a personal phonebook, as can be seen in Figure 4. The user can also choose between different talking faces, both male and female.

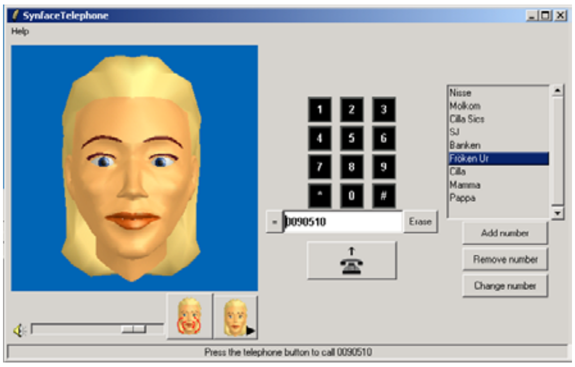
The graphical interface design has been evaluated by RNID and some changes will be implemented according to this. The main changes are a larger face and a different layout for the telephone controls. This prototype will be tested with hard-of-hearing users in three countries, the Netherlands, UK and Sweden.

6. Project evaluation

A substantial part of the project effort is directed at the user-based evaluation of the prototype. It will be evaluated in trials involving hearing-impaired users, with a focus on those persons with severe and profound hearing impairments. Tests will be performed in the three project languages both in home and work environments. The tests will start during the later half of this year.

7. Conclusions

The SYNFACE project aims to develop a new system for hard of hearing telephone users. It uses an artificial face to recreate the lip movements of the person at the other end of the telephone line. The primary application is for hearing-impaired users, where this information will substantially enhance access to telephone and other voice channels. The system can also be used more widely for public voice information channels in noisy environments. There is a wide range of other applications, for example that of avatars, and in audio-visual tools for language training, to which the results of the project will also contribute.



Press the telephone button to call 0090510

Figure 4. The first SYNFACE graphical interface prototype. The synthetic face can be seen to the left and the controls to the right.

8. Acknowledgements

The SYNFACE project is financed by the European Union (EU) under the FP5 IST Key Action I Programme: Systems and Services for the Citizen. The project also benefits from an equipment donation within the HP VoiceWeb Initiative.

The Teleface project was funded by VINNOVA (previously KFB and NUTEK) and the Swedish Handicap Institute.

We would like to thank B Lyberg at Linköping University who allowed us use their equipment for recording the audiovisual speech databases.

9. References

[1] Beskow, J., Dahlquist, M., Granström, B., Lundeberg, M., Spens, K-E., and Öhman, T., "The Teleface project: Multimodal speech communication for the hearing impaired", *Proc of Eurospeech*, 1997.

[2] Agelfors, E., Beskow, J., Dahlquist, M., Granström, B., Lundeberg. M., Spens, K-E. & Öhman, T. Synthetic faces as a lipreading support, *Proc of ICSLP'98*, 1998.

[3] Brooke N. and Summerfield Q "Analysis, synthesis and perception of visible articulatory movements". *J. of Phonetics*, 11: 63-76,1983

[4] Siciliano, C., Faulkner, A., Williams, C., "Lipreadability of a Synthetic Talking Face in Normal Hearing and Hearing-Impaired Listeners" *Proc. of AVSP*, 2003.

[5] Bench J and Bamford J (Eds.) *Speech-hearing Tests and the Spoken Language of Hearing-Impaired Children.* London: Academic, 1979.

[6] Plomp, R. and. Mimpen, A., "Improving the reliability of testing the speech-reception threshold for sentences", *Audiology*, 18, pp. 43-52, 1979.

[7] Öhman T., An audio-visual speech database and automatic measurements of visual speech. *KTH, Speech, Music and Hearing, Quat rept*, 34/1-2, pp. 61-76, 1998

[8] Siciliano, C., Williams, G., Beskow, J. and Faulkner, A., "Evaluation of a multilingual synthetic talking face as a communication aid for the hearing impaired", *15th Int. Cong. of the Phonetic Sciences*, 2003.

[9] Beskow, J., Engwall, O. and Granström, B., "Resynthesis of Facial and Intraoral Articulation from Simultaneous Measurements", *Proc. of 15th Int. Cong. of Phonetic Sciences,* 2003

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

[11] Molander, M. *Experiment with asynchrony in multimodal speech communication*. MSc Thesis. KTH-TMH 2003

[12] Kitawaki, N. "Pure delay Effects on Speech Quality in Telecommunications", *IEEE Journal on selected areas in communications*, 9, 586-593. 1991

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

[14] Ward, K. *User interface for the SYNFACE-project*. MSc Thesis report, TMH-KTH, 2002.