# SYNFACE - A Talking Head Telephone for the Hearing-impaired

Jonas Beskow*[1], Inger Karlsson[1], Jo Kewley[2], and Giampiero Salvi[1]

[1] Kungl Tekniska Högskolan (KTH), Dept of Speech, Music and Hearing, SE-100 44 Stockholm, Sweden
[2] Royal National Institute for Deaf People (RNID), London EC1Y 8SL, UK

**Abstract.** SYNFACE is a telephone aid for hearing-impaired people that shows the lip movements of the speaker at the other telephone synchronised with the speech. The SYNFACE system consists of a speech recogniser that recognises the incoming speech and a synthetic talking head. The output from the recogniser is used to control the articulatory movements of the synthetic head. SYNFACE prototype systems exist for three languages: Dutch, English and Swedish and the first user trials have just started.

## 1 Introduction

For a hearing-impaired person it is often necessary to be able to lip-read as well as hear the person they are talking with to communicate successfully. This is particularly relevant for telephone communication, where the hearing-impaired user is at a distinct disadvantage. It has been shown that the videophone can be a valuable form of communication for hearing-impaired people, providing essential visual speech information. However, videophones require expensive equipment at both ends and high bandwidth. In this paper a different solution is presented where the hearing-impaired telephone user is supplied with a system that is connected to the ordinary telephone. This system, SYNFACE [1], shows the lip movements of the speaker at the other end synchronised with the speech. SYNFACE has the distinct advantage that only the user on the receiving end needs special equipment. The SYNFACE technology has its background in the Teleface project [2, 3], which demonstrated that synthesised facial movements driven by an automatic speech recogniser can provide phonetic information that is not available in the auditory signal to a hearing-impaired user. In the SYNFACE project this has been further developed into a multilingual synthetic talking face, Fig. 1. A phoneme recogniser that works with very low latency has been developed as it is crucial in conversations that the visual and auditory speech signals are delayed relative to each other.

The key technological task for the SYNFACE system 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

* authors in alphabetical order

current technology is able to synchronise lip movements with arbitrary speech, the result is neither very natural nor does the visual information carry substantial phonetic information. Thus there are two main research areas that have been addressed in the SYNFACE project. The visual speech information requirements of auditory-visual communication have been defined, and techniques to derive this information from the acoustic speech signal in near real time have been developed. A multilingual prototype of the SYNFACE system has developed for Dutch, English and Swedish. This prototype will be evaluated by hearing-impaired users during this year. The user trials are about to start at three sites: RNID in the UK, Viataal in The Netherlands and KTH in Sweden. During the coming year the use of SYNFACE for two further languages, Finnish and Italian, will be demonstrated. The different parts of the SYNFACE system and the planned user trials are described in detail in this paper.

**Fig. 1.** The SYNFACE system

## 2 The SYNFACE Recogniser

The conditions faced by a speech recogniser for SYNFACE are adverse in many respects. The system is required to be speaker independent as the identity of the caller is not known in advance, task independent as the conversation is not restricted to any particular domain, narrow band as the conversation is supposed to happen across the telephone line, low latency as very short delay is allowed between the incoming speech and the lip movements, if the turn taking mechanism in the telephonic conversation is to be preserved. Furthermore the system is to be developed for three different languages.

Task independence can be achieved either by very large dictionary word recognition, or by avoiding decoding the speech signal at the lexical level, e.g. with phone recognition. The last was chosen here as the task of mapping acoustic to visual articulatory movements is independent of the higher levels of speech understanding. The speech signal is thus first classified into phonetic classes. The rule-based system described later in this paper is then used to generate the visual parameters from the phonetic sequence.

The task could in principle be considered as one of regression rather than classification, and the visual parameter trajectories could be directly estimated from the acoustic signal avoiding any phonetical representation. This solution, even if appealing from the research point of view, was abandoned, due to earlier experience [4], and the constraint of building a fully working system, imposed by the project plan.

The system is based on a hybrid of recurrent neural networks (RNNs) and hidden Markov models (HMMs) [5, 6]. The RNNs are used as frame-by-frame posterior probability estimators given the acoustic evidence. Probabilities are then fed into the HMMs that bear a model of time evolution. A Viterbi-like decoding scheme is employed in order to obtain the best phonetic sequence for a given speech segment. The latency constraints are met by approximating the Viterbi solution with a limited look-ahead length. This reduces the effectiveness of the time dependent information contained in the HMM. As shown in [5], this degradation is limited by the ability of the RNNs to learn their internal representation of time evolving problems. The interaction of these factors is thoroughly examined in [6].

The system was trained on the SpeechDat databases for Swedish, English and Flemish (Dutch). These databases are suitable for this application as they contain recordings of a large number of speakers (speaker independence) over telephone lines. Moreover they are available in over 20 different European languages, allowing easy development for other languages.

## 3 Synthetic Talking Head

The visual signal generation is based on 3D polygon models, which are parametrically articulated and deformed. The facial models are parameterised using weighted geometric transformations (translation, scaling and rotation) of the vertices. This is a generalised version of the scheme proposed by Parke [7] see [8] for details. The control parameter set used for the synthetic face articulation is chosen to reflect the articulatory phonetic features often used to describe speech production, and most parameters are defined in terms of articulatory targets rather than general geometric measures. The most important articulatory control parameters are jaw rotation, lip rounding, lip protrusion, mouth width, bilabial closure, labiodental closure and apex (tongue tip elevation).

### 3.1 Articulatory Control Models

To animate the movements of the talking head, an articulatory control model is used, that takes time-stamped phonetic symbols as input and produces articulatory control parameter trajectories to drive the face model. One of the problems that a control model has to deal with is coarticulation. Coarticulation refers to the way in which the realisation of a phonetic segment is influenced by neighbouring segments. It is the result of articulatory planning, inertia in the biomechanical structures of the vocal tract, and economy of production [9]. But

coarticulation also serves a communicative purpose in making the speech signal more robust to noise by introducing redundancies, since the phonetic information is spread out over time. Backward, or carry-over coarticulation, refers to the way in which articulation at some point in time is affected by the articulation at some previous point in time. Forward, or anticipatory coarticulation, on the other hand, is a term used to describe how articulation at some point in time is affected by articulation of segments not yet realised. This means that a model that accounts for anticipatory coarticulation will need to look at future segments to calculate movements at the current point in time. In the SYNFACE scenario, this poses a problem, since the control model has to react with very short delay to the phonemes received from the speech recogniser.

A talking head that does not account for coarticulation will seem jerky and unnatural in its articulation. In the SYNFACE project, we have investigated several alternative articulatory control models, both rule-based and data-driven. In the rule-based model [10], each phoneme is assigned a target vector of articulatory control parameters. To allow the targets to be influenced by coarticulation, the target vector may be under-specified, i.e. some parameter values can be left undefined. If a target is left undefined, the value is inferred from context using interpolation, followed by smoothing of the resulting trajectory. As an example, consider the lip rounding parameter in the word "askew": an unrounded vowel, followed by a consonant cluster and a final rounded vowel. In this case, lip rounding would be unspecified for the consonants /s/ and /k/, leaving these targets to be determined from the leading and trailing vowel context by linear interpolation from the unrounded vowel /a/, across the consonant cluster, to the rounded /ew/. Definition of the targets in the rule-based control model is a manual labour, based on comparisons between animation and video recordings. For the prototype SYNFACE system, the articulatory targets for the synthetic talking face has been adapted to Dutch and English, in addition to Swedish [11]. In addition, a special real-time version of the rule-based control model has been developed, that uses a finite time-window of articulatory anticipation, as opposed to the original model of [10] that required access to the full utterance prior to synthesis. The current prototype uses a look-ahead window of 200 ms.

As an alternative to the rule-based control model, we have investigated several data-driven (trainable) methods of generating articulatory parameter trajectories to control the face model [12]. The data-driven models are trained on a corpus of articulatory movements recorded from a human speaker, and learn to reproduce the articulatory patterns. We have recorded audio-visual speech databases with articulation movement tracking for the three project languages. The optical motion tracking is done using a Qualisys system with four IR cameras (`http://www.qualisys.se`). 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 spectacles (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 [13]. In the comparison study, four different data-driven models were trained on the same data (Swedish only). Two of the models are

based on coarticulation models from speech production theory, that have previously been employed for visual speech synthesis [14, 15] and two of them are based on artificial neural networks (ANNs). The models are trained by estimating the free parameters, to minimise the error between predicted and measured parameter trajectories over a training set of 200 sentences. The models have been evaluated objectively (by comparing RMS error and correlation between target and prediction) over a set of sentences not part of the training material. The evaluation showed a small advantage for one of the speech production theory-based models [14]. In addition, an audiovisual sentence intelligibility test with 25 normal-hearing subjects was conducted, where the four data-driven models were compared against the rule-based model as well as a no-face condition. This evaluation showed that all models provide significantly increased intelligibility over the audio alone case, and the rule-based model produced the most intelligible articulation. One possible explanation why the rule-based model outperformed the data-driven ones is that the rule-based model was developed with clear articulation and high intelligibility as the primary goal, and as such it almost tends to hyper-articulate. The data-driven models on the other hand, are trained to mimic the speaking style of the target speaker, who could be characterised as having a rather relaxed pronunciation. On the basis of these experiments, the rule-based control model has been chosen for the current version of the prototype. However, work to improve the intelligibility of the data-driven models by re-training on data from a more intelligible speaker is underway.

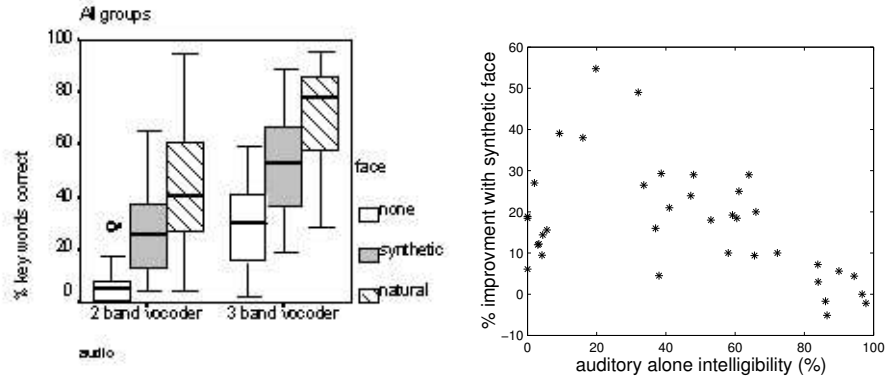
### 3.2 Intelligibility Experiments

Intelligibility testing has been performed during the development of the prototype system based on the rule-based control model. Sentence intelligibility measures have been obtained for Dutch, English and Swedish from normal listeners using filtered speech, and from a group of English and Swedish hearing-impaired listeners [16]. Additionally, visual consonant confusion data have been collected from English normal listeners to identify weaknesses in the consonant synthesis rules. Some intelligibility test will also be included in the final user tests of the complete SYNFACE system with hearing-impaired users.

In the sentence intelligibility tests the filtered auditory signal was presented either alone, with the synthetic face or with a video of the original talker. Normal-hearing listeners heard noise-vocoded speech. Hearing-impaired listeners heard speech that was unprocessed but telephone-band limited. 12 native speakers for each language 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 [3]. The speech material consisted of lists of everyday sentences in Swedish, Dutch and English designed to test speech perception ability. Subjects were seated in front of a computer screen and a loudspeaker or headphones, were presented with sentences in their native language, and were then asked to repeat what they perceived. The test leader noted the subjects' response. The number of correctly perceived keywords was counted. Scores are expressed as percent of keywords correct. Intelligibility

scores for the normal-hearing subjects are shown in Fig. 2 a). 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 Fig. 2 b) shows, the synthetic face was generally more effective for those hearing-impaired listeners with poorer auditory-alone scores, although listeners with virtually zero auditory-alone scores showed around 10% intelligibility gain when the synthetic face was added.

For the VCV intelligibility tests video recordings from two female native British English talkers were used in addition to the synthetic face. 24 different British English consonants were produced in left and right /i/, /a/ and /u/ contexts. Five male and five female native speakers were shown a silent movie of the VCV productions and were asked to indicate the correct consonant in a forced choice task. Accuracy was low overall with 13.6% for the synthetic face and 23.4% for the natural faces. The difference between intelligibility was small for bilabial and labiodental articulations. For back articulations the synthetic face tended to give a large proportion of /l/ responses compared to the natural faces. This indicates that the production rules for tongue movements in the synthetic face need to be improved.




a) b)

**Fig. 2.** a) Sentence intelligibility for degraded speech with synthetic and natural visual speech. Results from all languages and normal-hearing subjects. b) Intelligibility gain with synthetic face compared to auditory-alone scores for hearing-impaired listeners

## 4 SYNFACE Prototype

The present prototype is implemented on a portable Windows PC connected to a standard telephone via off-the-shelf hardware. The remote and local talkers voices are fed into the line-in of a USB sound card. In order to facilitate programmatically controlled delay of the incoming speech, the ear-piece/loudspeaker of the telephone is connected to the line-out of the USB sound card. The prototype system software consists of three distinct modules: the recogniser (that is also responsible for input and delayed output of audio), the articulatory control model, and the facial rendering engine. To facilitate system integration and promote rapid prototype development, we use a scripting language (Tcl/Tk [17]). The recogniser, the articulatory control model and the facial rendering engine are implemented in C/C++ as compiled extensions to the Tcl interpreter.

## 5 User Trials

User trials will be carried out in the UK, Sweden and the Netherlands to evaluate the usability, effectiveness and perceived value of the SYNFACE prototype. In the UK about 40 users will participate in the trials while in Sweden and the Netherlands about 10 subjects will test the system.

In the UK, trials will initially be carried out in the lab with hearing impaired users discussing set topics/scenarios with an experimenter, using SYNFACE. To gain an greater understanding of the potential uses of SYNFACE and to illustrating the speech recognition aspects of the system, users will also be shown a video of SYNFACE being used for a conversation, and will then watch several short stories and sentences that have been pre-recorded. Information on the usability, effectiveness and perceived value of SYNFACE will be gathered using a questionnaire, as well as a de-briefing interview. From the users that complete these UK lab-based trials, a number of users will be asked to use the SYNFACE in their home or workplace for 2-3 weeks to investigate how users attitudes and experiences change with prolonged use of SYNFACE. In the Netherlands, trials will be comprised of two parts. The first part will present sentences to users in three different ways - SYNFACE, Video telephony and sound only. This will allow comparisons to be made between the comprehension of the sentences for each method of display, and reveal the extent to which SYNFACE facilitates telephone conversations. The second part of the Dutch trials will get the participants to use SYNFACE for spontaneous conversation. In Sweden, it is planned that trials will include a perception test on sentences and then the system will be left with participants to use for one day in their homes, to call friends and relatives.

To allow comparisons to be made between the findings of trials in each country, the same questionnaires will be used in all trials. The trials will start during May this year, some preliminary results will be presented at the conference.

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

We would also like to thank the other project members at KTH and Babel-Infovox, Sweden, UCL and RNID, the UK, Viataal the Netherlands.

## References

1. (http://www.speech.kth.se/synface)
2. Beskow, J., Dahlquist, M., Granström, B., Lundeberg, M., K-E.Spens, Öhman, T.: The teleface project: Multimodal speech communication for the hearing impaired. In: Proc. of Eurospeech, Rhodes (1997) 1651–1654
3. Agelfors, E., Beskow, J., Dahlquist, M., Granström, B., Lundeberg, M., Spens, K.E., Öhman, T.: Synthetic faces as a lip-reading support. In: Proc. of ICSLP, Sydney, Australia (1998)
4. Öhman, T., Salvi, G.: using HMMs and ANNs for mapping acoustic to visual speech. TMH-QPSR **1-2** (1999)
5. Salvi, G.: Truncation error and dynamics in very low latency phonetic recognition. In: ISCA workshop on Non-linear Speech Processing. (2003)
6. Salvi, G.: Dynamic behaviour of connectionist speech recognition with strong latency constraints. Speech Communication (submitted)
7. Parke, F.: Parametrized models for facial animation. ieee computer graphics. IEEE Computer Graphics **2** (1982) 61–68
8. Beskow, J.: Animation of talking agents. In: Proc. of AVSP'97, Rhodes (1997)
9. Lindblom, B.: Economy of speech gestures. In: The production of speech. P. MacNeilage, New York: Springer-Verlag (1983) 217–245
10. Beskow, J.: Rule-based visual speech synthesis. In: Proc. of Eurospeech, Madrid, Spain (1995) 299–302
11. Karlsson, I., Faulkner, A., Salvi, G.: SYNFACE a talking face telephone. In: Proc. of EuroSpeech. (2003) 1297–1300
12. Beskow, J.: Trainable articulatory control models for visual speech synthesis. Journal of Speech Technology (in press)
13. Beskow, J., Engwall, O., Granström, B.: Resynthesis of facial and intraoral articulation from simultaneous measurements. In: Proc. of ICPhS. (2003) 431–434
14. Cohen, M., Massaro, D.: Modelling Coarticulation in Synthetic Visual Speech. In: Models and Techniques in Computer Animation. Springer Verlag, Tokyo (1993) 139–156
15. Reveret, L., Bailly, G., Badin, P.: Mother: a new generation of talking heads providing a flexible articulatory control for video-realistic speech animation. In: Proc. of ICSLP, Bejing, China (2000) 755–758
16. Siciliano, C., Williams, G., Beskow, J., Faulkner, A.: Evaluation of a multilingual synthetic talking face as a communication aid for the hearing impaired. In: Proc. of ICPhS. (2003) 131–134
17. Ousterhout, J.: Tcl and the Tk Toolkit. Addison Wesley (1994)