IST-2001-33327   SYNFACE
Synthesised Talking Face Derived from Speech for Hearing Disabled Users of Voice Channels

Summary

The SYNFACE project has developed 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 providing additional support to the user.

Hard of hearing people normally compensate for their hearing problems by lip-reading as well as listening to the person they are talking to. This means that they need to be able to clearly see the face of the person they are talking to which is impossible over the telephone.

SYNFACE consists of a phoneme recogniser and a visual speech synthesiser, that is a talking face. The phoneme recogniser identifies the speech sounds and the face synthesiser then recreates their articulation. The movements of the talking face are synchronised with the audio speech signal and shown on a screen attached to the SYNFACE user's telephone. To use SYNFACE, only the hard of hearing user will need to have the device installed.

Setting the Scene

The number of people with hearing problems is large. For example in the UK it is estimated that one in seven has some form of hearing loss and this number is growing. This is partly due to the ageing population, as hearing problems increase with age. The noise pollution of everyday life and exposure to very loud music also causes a growing number of afflicted.

Hard-of-hearing people often rely on lip-reading to follow conversations This works well in face-to-face situations but over the telephone this visual information is missing, and people are left to rely only on what they can hear. This means that telephone conversations can be difficult for hard of hearing people and frequently these problems are greater when the person at the other end is a stranger.

In an earlier project (Teleface) it was demonstrated that an artificial face where the lip movements were derived from the speech signal using automatic speech recognition gave good support to hard of hearing listeners. The time delay that the technology introduced was too large though. In the present project the automatic speech recognition is developed to work with a very small delay allowing it to be used in ordinary conversation. In addition to the time constraints the recogniser needs to be able to work for a wide variety of voices. A further complication for the recogniser is that the conversations may be about any topic, unlike many typical speech recognition applications where the domain is known.

Today two main alternatives to telephone communication are available when simple amplification of the speech signal is insufficient: textphone or videophone. Textphones consist of a keyboard and small display, allowing users to hold a typed conversation over a telephone line. Textphones can also be connected to a relay service where an operator acts as intermediary for the call, typing what the hearing person says to the deaf person, and speaking what the deaf person types to the hearing person. When using a videophone a camera, screen and a wideband video connection is necessary at both ends. There are drawbacks with both techniques. When using SYNFACE only the hard of hearing user will need any extra equipment, no intermediaries are needed and the speech is transferred as in ordinary telephony.

Approach

In SYNFACE the lip movements related to the speech signal are recreated on an artificial talking face. The incoming speech signal is fed into a computer where the speech sounds are automatically identified with a phoneme recogniser. The recognition result serves as the input to a visual speech synthesiser which

produces an animation on a synthetic face with the proper lip-, jaw- and tongue movements for the recognised phone sequence. The speech signal will be slightly delayed, by 200 ms, so that the acoustic speech synthesis is synchronised with the visual speech signal. This delay will allow time for the processing of the speech signal.

The nature of the task poses demanding constraints on the automatic speech recognition (ASR) system. Whilst conventional speech recognition systems may take advantage of long-term linguistic information through time lags in speech signal processing, the specific time-constrained needs of the SYNFACE application preclude the use of such information. Therefore, a unique phoneme recogniser was developed for SYNFACE that was optimised to deliver phonetic output with very low latency (on the order of a phoneme.). Because the automatic phonetic recognition and the visual animation software need to run simultaneously on a compact machine, the computational resources are quite limited. Given these constraints, the most effective yet computationally efficient method for phoneme recognition chosen employs hybrid recurrent neural networks (RNN) and hidden Markov models (HMM) for the acoustic decoding.

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. A control model needs to include coarticulation to account for the way that the realisation of a phonetic segment is influenced by neighbouring segments. Coarticulation reflects articulatory planning, inertia in the biomechanical structures of the vocal tract, and economy of production. It makes the speech signal more robust to noise by introducing redundancies, since the phonetic information is spread out over time. For the prototype SYNFACE system, the articulatory targets have been adapted to Dutch and English, in addition to Swedish. 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 that required access to the full utterance prior to synthesis. The current prototype uses a look-ahead window of 200 ms.

Articulation data that informs the synthesis system for the three project languages has been collected using a Qualisys system, http://www.qualisys.se. The data has been used to formulate articulation rules for the visual speech synthesis. Perception tests have been performed to evaluate the rules.

Results and achievements
Prototype SYNFACE systems have been produced for Dutch, English and Swedish. In the prototypes the incoming speech signal is picked up before it is entering the earphone and entered into the computer. The signal is processed continuously and the results are delivered to the visual speech synthesiser. The speech signal is delayed and fed back to the earphone. The artificial face movements are synthesised from the recognition results and synchronised with to speech signal. The artificial face is shown on a computer screen. The speech signal coming from the SYNFACE user is blocked out to avoid confusions for the user.

Conclusions
The SYNFACE prototypes have been evaluated by hard-of-hearing users in the UK, the Netherlands and Sweden. These results are very encouraging; a majority, 70%, of the evaluators of SYNFACE, have found it helpful and effective. An even higher percentage, 90%. felt that they gained at least some support in conversations from SYNFACE. A majority, 80%, thought SYNFACE was a useful product. In particular many felt that SYNFACE helped in understanding some special types of information for example telephones numbers and names and unknown words. There were suggestions on ways to improve SYNFACE, the users would prefer a more natural face and they were sometimes annoyed by errors made by SYNFACE.

Further results will be published on the project homepage www.speech.kth.se/synface.

Coordinator
Dr Inger Karlsson
KTH (Kungl Tekniska Högskolan)
Speech, Music and Hearing, Lindstedtsv 24
SE-100 44 Stockholm, Sweden
Tel.: +46 8 790 7563
Fax.: +46 8 790 7854
Email: inger@speech.kth.se