SYNTHETIC VISUAL SPEECH DRIVEN FROM AUDITORY SPEECH

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

Department of Speech, Music and Hearing, KTH, Sweden
{eva, beskow, bjorn, magnusl, giampi, kalle, tobias}@speech.kth.se
www.speech.kth.se/teleface/

ABSTRACT

We have developed two different methods for using auditory, telephone speech to drive the movements of a synthetic face. In the first method, Hidden Markov Models (HMMs) were trained on a phonetically transcribed telephone speech database. The output of the HMMs was then fed into a rule-based visual speech synthesizer as a string of phonemes together with time labels. In the second method, Artificial Neural Networks (ANNs) were trained on the same database to map acoustic parameters directly to facial control parameters. These target parameter trajectories were generated by using phoneme strings from a database as input to the visual speech synthesis

The two methods were evaluated through audio-visual intelligibility tests with ten hearing impaired persons, and compared to "ideal" articulations (where no recognition was involved), a natural face, and to the intelligibility of the audio alone. It was found that the HMM method performs considerably better than the audio alone condition (54% and 34% keywords correct respectively), but not as well as the "ideal" articulating artificial face (64%). The intelligibility for the ANN method was 34% keywords correct.

1. INTRODUCTION

The idea to use auditory speech to generate visual information about the linguistic message, and thereby making it accessible to persons who rely to a larger extent on their visual sensory mode, is not new. Previous work where an established visual code, such as facial, or hand gestures, has been utilized were performed by, for instance, Erber [1] who synthesized lip shapes for vowels from the auditory speech code. More recently, Duchnowski et al. [2] used automatic speech recognition to provide speech readers with hand gestures in cued speech. The great advantage of using face gestures as the visual code is that it is often already trained by most people, either directly or indirectly, as they lipread as soon as the auditive signal is weak or disturbed. Earlier related studies where lip movements or face gestures have been generated from the auditory speech signal include [3, 4, 5].

In the Teleface project at KTH, we have since 1996 investigated the possibility of using a synthetic face as a speech reading aid for hard of hearing persons during telephone conversation. The idea is to use the telephone speech to drive a synthetic face at the listener's end, so that it articulates in synchrony with the speech [6].

The project is divided into two stages. In the first stage, we evaluated the possible use of such a technique. We found that the intelligibility of VCV-syllables was considerably increased when the synthetic visual speech was added to the auditory speech. This is also true for everyday sentences [7].

However, the visual stimuli presented in these tests were prepared by applying manually created and fine-tuned transcriptions to synthesis rules, to achieve stimuli that is perfectly synchronized and consistent with the audio. Artificial face articulations of this sort will in the remainder of this paper be referred to as rule-based ideal or just "ideal." In a real life application, such ideal artificial face movements are very hard to accomplish. This paper describes the second phase of the project, where we have implemented algorithms to automatically derive artificial facial articulations from telephone quality speech, and evaluated whether these articulations are sufficient to improve the intelligibility over the audio alone condition, and approach the ideal synthetic face.

2. METHOD

Two methods, described in this paper, make use of statistical models trained on the same speech material. Utterances containing single words and sentences were selected from the SpeechDat database [8]. This subset of the database contains about 13,000 recordings of telephone speech from 1000 speakers. For training the HMMs and the ANN we selected 750 speakers (433 females and 317 males, 14 hours of speech) using a controlled random selection algorithm [9] in order to maintain the same gender and dialect proportions as in the full database. Two hundred of the remaining subjects were used for evaluation. Aligned phonetic transcriptions of the material have been obtained by forced alignment using the SpeechDat orthographic transcriptions. The 8 kHz sampled speech signal, was parameterized into 10ms frames of thirteen

parameters (twelve mel-cepstral and energy), which were used as input to the systems. Dynamic parameters (delta and acceleration) were added for the HMM method.

2.1. The HMM Method

In a first step the acoustic signal is analyzed and each frame is classified by HMMs into linguistic units. The resulting time aligned transcription is in a second step converted into face parameter trajectories by a rule-based visual speech synthesis system, see [10] and [11].

Two main methods have been tested. In the first, HMMs are trained to recognize phonemes. In the second method, classes of visually similar phonemes (termed visemes) are recognized. All phonemes in a viseme class are identically modeled in the visual synthesis rules.

The phoneme models are three or four states HMMs, trained on the SpeechDat material [12], by using the HTK toolkit [13]. These mono-phone models have been improved by adding up to eight Gaussian distributions for each state, and by including context information, resulting in a set of tri-phones. Viseme models are obtained by merging mono-phone models belonging to the same visual class, hence preserving the same HMM structure. These models have been retrained and improved adding up to sixteen Gaussian terms for each state mixture. In the HMM-networks, which specify the allowed sequences of HMMs in recognition, any model may follow any other. In the case of tri-phone models, context information is taken into account when generating recognition hypothesis.

2.2. The ANN Method

An alternative to using the HMM method is to train ANNs to map the audio directly to parameter values of the synthetic face. In this way, no intermediate classification errors come into play. Another possible advantage is that coarticulation is handled directly, without applying any rules.

To generate the target values for training the ANN, we ran the phoneme strings and time labels of the training speech through the visual speech synthesis system. The resulting eight trajectories (one for each visual speech synthesis parameter) were then used for training, see Figure 1.

A three layered net, with 13 units in the input layer, 50 units in the hidden layer and eight units in the output layer, was created using the NICO toolkit [14]. The input speech parameters were the same as the static ones in the HMM method, and each output node corresponds to one visual synthesis parameter. Each layer was connected to the other two, and the hidden layer was also connected to itself (i.e., a recurrent network). A time-delay window of six frames (10 ms per frame) was used. This gives the net three frames of context, both forward and backward in time for the parameter values at any frame. The total number of connections in the network was 15.636.

Three different speaker independent ANNs with the typology described above were trained on the same speech material as the HMMs: one for males, one for females, and one for mixed speakers.

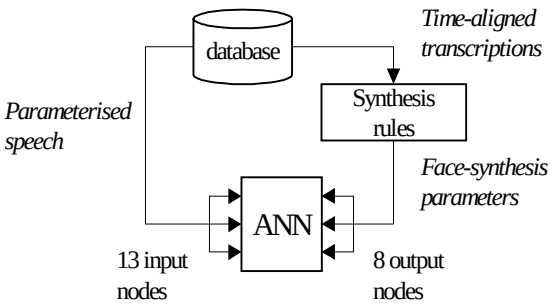

Figure 1: Procedure to train the ANN. The target values are generated by applying the visual synthesis rules to phoneme strings, which are obtained by forced alignment

3. PRELIMINARY EVALUATION

During the development of the two methods, we needed to test intermediate versions against each other. For the HMMs, this was done by computing the recognition accuracy of viseme classification for different types of models (mono-phones, tri-phones and visemes). In the case of mono, and tri-phones, the clustering into visemes was done after recognition.

Results have shown that the pre-clustering method, in which models for visemes are employed, never performs as good as the post-clustering solution. The accuracy for mono-phones with eight mixtures is 47.0%, while visemes with sixteen mixtures obtain only 42.4% accuracy. The best results were obtained with tri-phones, eight mixtures (56.2%).

For the ANNs, on the other hand, accuracy scoring is not possible because there is a direct mapping of the audio to the face synthesis parameters, and no classification is done.

When evaluating the results, we are not primarily interested in the classification of linguistic units.

Rather we want to know how well the methods produce the articulations of the synthetic face, i.e., how well they produce the parameter trajectories. The evaluation is therefore done by referring to the target trajectories used as target values for the ANNs.

It is important to keep in mind that, in our study, those target trajectories were generated by applying the same synthesis rules as in the HMM method. These rules are not necessarily optimal for the application. Therefore, trajectories obtained from a perfectly recognized utterance, even if optimal (according to our evaluation method), may not be so regarding the end result, since they keep the limitations which are intrinsic in the rule-based synthesizer. This is not the case for the ANN method, since the target trajectories may be obtained in any way, e.g., by manual adjustments or by re-synthesis from measurements of human speakers.

The differences between the methods can be visualized in Figure 2. For the HMM method, the result is perfect if the phonemes are recognized into the correct viseme classes (before 900 ms and after 1400 ms in Figure 2). However, when the recognition fails, the generated movements may be completely misleading (between 900 and 1400 ms in Figure 2). The results from the ANN, on the other hand, are usually neither perfect nor completely wrong. The articulation is carried out in the right direction, most of the time but seldom all the way. For example, bilabial occlusion (one of the synthesis parameters) is often near, but never exactly, 100% for bilabials (which it should be). Since speech readers are sensitive even for small deviations in, e.g., bilabial occlusion, this is a serious drawback.

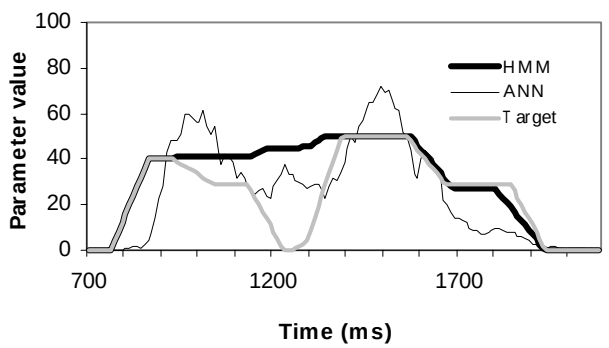

Figure 2: Trajectories of the visual speech synthesis parameter for lip rounding for the utterance /kΠ♣ρεκ]/. The thick gray curve is the target trajectory obtained from the forced alignment

Another characteristic feature for the ANN method is that, since it works on a frame-by-frame basis, the trajectories tend to become jerky, as can be seen in Figure 2. This was a disturbing feature, which contributed to the generally high subjective mental effort the subjects experienced when speechreading the synthetic face movements generated by the ANN method [15].

4. INTELLIGIBILITY TESTS

In the previous section, we discussed the performance of the HMM method in terms of recognition accuracy. We also compared the results obtained from both methods to the target trajectories. This analysis is however not necessarily relevant, since we do not know how deviations in a given parameter will affect the audio-visual perception. For that reason, we have performed intelligibility tests to see how the end result is perceived and subjectively evaluated by humans.

The tests were performed using a computer-based test environment [16]. The subjects were ten hearing-impaired persons with a high motivation for using the synthetic face. All but one of the subjects have been active in previous tests.

The visual stimuli were presented on the computer screen and the auditive stimuli were presented using a separate loudspeaker with the volume set to a comfortable level. During a few training sentences the subjects were allowed to adjust their hearing aid to obtain as good hearing as possible. The test material consisted of short everyday sentences, which were presented without any information about the context. The subjects' task was to verbally repeat the perceived sentence. The number of correctly repeated keywords (three per sentence) was counted and expressed as the percent keywords correct.

Stimuli were presented in three basic modes: natural voice and no face (the audio-only test condition, labeled A), natural voice and synthetic face (AS), and natural voice and video recordings of a natural face (AN). The natural voice used in all modes and, for the recognition, was filtered to a bandwidth of 3.7 kHz to simulate the audio quality of an ordinary telephone conversation.

For the test condition with natural voice and synthetic face (AS), articulation for the synthetic face was prepared in three different ways. In the first method, the ideal rule-based trajectories were used. This is the way the articulation of the synthetic face has been created for our previous intelligibility tests. In the other two cases, the trajectories were obtained by the HMM and ANN methods respectively. For the HMM method, we chose to use the tri-phone models, trained on both males and females. The

ANN was a speaker-independent net, trained on the male speakers. Figure 3 shows the result of the intelligibility test.

Mean values from ten hearing impaired subjects are presented, and the standard deviation (one above and one below the mean) is shown as vertical bars. The HMM method improved the intelligibility over the audio alone condition (54.0% keywords correct compared to 33.7%), and approaches the ideal rule-based condition (64.0%). The ANN method did not improve intelligibility significantly (34.3%). The audio-visual intelligibility with the natural face was nearly the same for all the subjects (85.9%), whereas it varied considerably for the other test conditions.

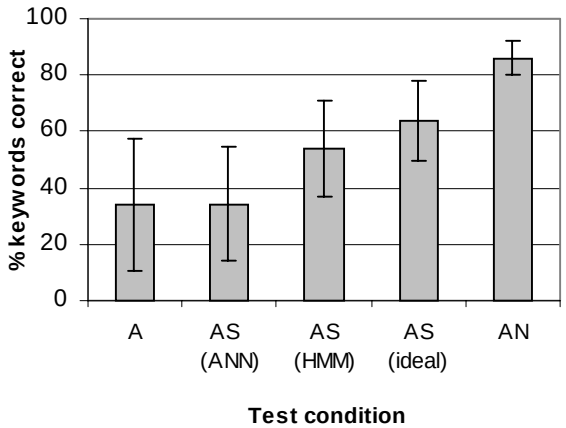

Figure 3: Results of the intelligibility tests for different test conditions described in the text. The bars show one standard deviation above and below the mean values

5. DISCUSSION

In this paper we have presented two methods for generating the movements of an artificial face with telephone quality speech as input. In intelligibility tests, with sentence speech material, we have seen that the HMM method increases the percentage of correctly perceived key words considerably compared to the audio alone condition. For the ANN method, the improvement compared to the audio only was not significant.

In this study, we did not process the output of the ANNs in any way. In future work we will experiment with different filtering of the trajectories to smooth the movements of the synthetic face. Another possible way to improve the ANN method is to divide the parameters into smaller groups and train separate nets for each group. Current nets are trained for all eight parameters, including the parameters controlling the length and elevation of the tongue. These parameters are probably not as important as, e.g., the parameters for rounding and bilabial occlusion. Since the ANN generated parameters are often not reaching its extreme values, a further possible improvement might be to expand the amplitude of the parameter trajectories.

Apart from improving details, the next important issue for our project is to implement the algorithms in real time.

6. ACKNOWLEDGEMENT

This work was funded by KFB, the Swedish Transport and Communications Research Board.

7. REFERENCES

1. Erber, N.P. (1979). Real-time synthesis of optical lip shapes from vowel sounds. *JASA 66(5)*, pp. 1542-1544.

2. Duchnowski, P., Braida, L., Lum, D., Sexton, M., Krause, J., & Banthia, S. (1998). Automatic generation of cued speech for the deaf: status and outlook. In *Proceedings of the International Conference on Auditory-Visual Speech Processing*, Terrigal, Australia.

3. Yamamoto E., Nakamura, S., and Shikano, K. (1998). Lip movement synthesis from speech based on Hidden Markov Models. *Speech Communication, 26*, pp. 105-115.

4. Masuko, T. Kobayashi, T., Tamura, M., Masubuchi, J., & Tokuda, K. (1998) Text-to-visual speech synthesis based on parameter generation from HMM. In *Proceedings of the IEEE International Conference of Acoustics, Speech and Signal Processing*, pp. 3745-3748, Seattle.

5. Morishiba, S. (1998) Real-time Talking Head Driven by Voice and its Application to Communication and Entertainment. In *Proceedings of the International Conference on Auditory-Visual Speech Processing*, Terrigal, Australia.

6. Beskow, J., Dahlquist, M., Granström, B., Lundeberg, M., Spens, K-E & Öhman, T. (1997). The Teleface project - Multimodal Speech Communication for the Hearing Impaired. In *Proceedings of Eurospeech'97*, Rhodes, Greece.

7. Agelfors, E., Beskow, J., Dahlquist, M., Granström, B., Lundeberg, M., Spens, K.-E., and Öhman, T. (1998) Synthetic faces as a lipreading support. In *Proceedings of the International Conference on Spoken Language Processing*, Sydney, Australia.

8. Höge, H., Tropf, H.S., Winski, R., van den Heuvel, H., Haeb-Umbach, R., & Choukri, K. (1997). European Speech Databases for Telephone Applications. In *Proceedings of the IEEE International Conference of*

Acoustics, Speech and Signal Processing, Munich, Germany.

9. Chollet, G., Johansen, F.T., Lindberg, B., and Senia, F. (1998). Test set and specification. *Technical Report LE2-4001-SD1.3.4, Consortium and CEC.*

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

11. Beskow, J. (1995): Rule-based Visual Speech Synthesis. In *Proceedings of Eurospeech'95*, Madrid, Spain.

12. Salvi, G. (1998). Developing acoustic models for automatic speech recognition. *Master of Science Thesis, TMH, KTH*, Stockholm, Sweden.

13. Young, S., Woodland, P., and Byrne, W. (1997). HTK: Hidden Markov Model Toolkit V2.1. Entropic Research Laboratory.

14. Ström, N. (1997). Phoneme Probability Estimation with Dynamic Sparsely Connected Artificial Neural Networks. *The Free Speech Journal (http://www.cse.ogi.edu/CSLU/fsj/), Issue #5.*

15. Öhman, T., and Salvi, G. (1999). Using HMMs and ANNs for mapping acoustic to visual speech. *STL QPSR 1-2/1999.*

16. Lundeberg, M. (1997). Multimodal talkommunikation - Utveckling av testmiljö (in Swedish). *Master of science thesis, Department of Speech Communication and Music Acoustics, KTH*, Stockholm, Sweden.