Database Construction for Speech to Lip-readable Animation Conversion

Gyorgy Takacs, Atilla Tihanyi, Tamas Bardi, Gergo Feldhoffer, Balint Srancsik
Peter Pazmany Catholic University, Faculty of Information Technology 1083 Budapest, Práter u. 50/a. Hungary
E-mail: *takacsgy@itk.ppke.hu*

Abstract – *The training database was one of the critical element in our speech to facial animation conversion system. This system was developed as a communication aid for deaf people. The specific database was constructed from audio and visual records of professional lip-speakers. The standardized MPEG-4 system was used to animate the talking head model. The trained neural net is able to calculate with acceptable error the principal component weights of feature points from the speech frames. The feature point coordinates are calculated from PC weights. The whole system can be implemented in mobile phones. Deaf persons were able to recognize about 50% of words from the speech driven animation in the final test.*

Keywords – *Audiovisual speech processing, facial animation, multimodal communication, lip reading.*

1. INTRODUCTION

Our aim was to develop a communication aid for deaf persons which can be implemented in a mobile telephone. In our system we provide one part of an animated human face on a display as the output to deaf users. The control parameters of the animation are calculated directly from the input speech signal. We know well that such representation of the human speech process is limited and contains inherent errors. Deaf people have fantastic abilities in understanding speech based on lip reading only. In spite of the limitations deaf persons aided with such an everyday equipment as a high-end class second or third generation mobile phone could naturally communicate with hearing people. The system calculates upon the enhanced capabilities of deaf persons in continuous error correction.

It is important in our system concept, that only continuous type of transformations are used in the complete audio to visual conversion [1]. One of the benefits of our direct solution is that the original temporal and energy structure of the speech are retained. Thus the naturalness of rhythm is guarantied. Further benefit is the relatively easy implementation in mobile phones environments with limited memory and computation power. A rather promising feature of our system is the potentially language independent operation.

A very important element in this new concept is to train the system by an audio-visual database collected from professional interpreters/lip-speakers. Their articulation style and level are adapted to deaf communication partners.

The dynamics of mouth movements and the naturalness of face animation models seem to be critical in lip-readablity of the audio-to-visual conversion. Usually researchers elaborate very sophisticated procedures to produce dynamic and natural talking heads [2,3,4]. We have selected the speakers in the data base composition thoughtfully and with special attention to the high dynamic requirements.

2. SYSTEM DESCRIPTION

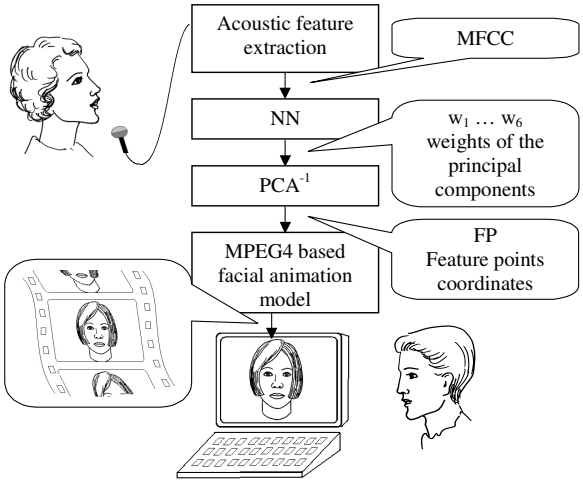

Fig. 1. Structure of the implemented speech to facial animation system

Our implemented conversion system is a PC-based software. Here we survey the complete system at a glance, as it is shown in Fig. 1.

The input speech sound is sampled at 16 bit/48 kHz and then acoustic feature vectors based on Mel-Frequency Cepstrum Coefficients (MFCC) are extracted from the signal. The feature vectors are sent to the neural network (NN), which computes a weighting vector $[w_1,\ldots w_6]$ which is a highly compressed representation of the target frame of the animation. The coordinates of our selected feature point set used to drive the animation are obtained by linear combination of some component vectors with the weights coming out from NN. This coordinate-recovery operation is denoted by the term "PCA^{-1}"

in Fig 1, because the predefined component vectors come from Principal Component Analysis (PCA). The Feature Point (FP) positions are computed in this way for 25 frames per second. An illustration can be seen in Fig 2.

The final component of our system is a talking head model. It is controled with the computed FP coordinates. Then the facial animation model appears on the screen.

In the training process of the neural network matrix back-propagation algorithm was used.

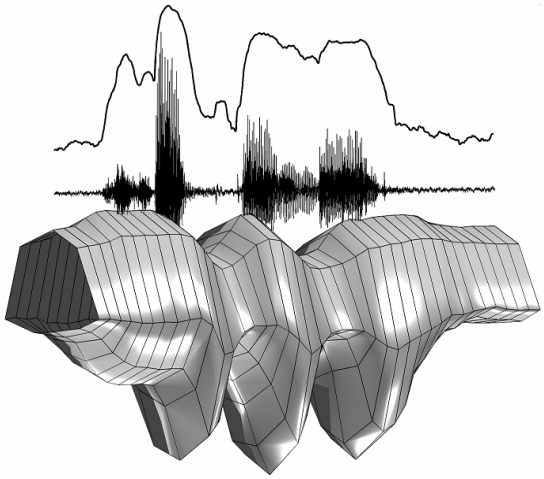

Fig. 2. The x-y components of 8.1-8.8 FP-s as a function of time pronouncing word "September". The upper solid line shows the frame energy in dB, the middle graph represents the waveform, the lower surface represents the lip contours.

We applied Lucia talking head model [5] with some modification. It uses the animation standard of MPEG-4 called facial animation parameters (FAP). Since FAP is a viseme based animation method, we have modified Lucia to work directly with feature point coordinates. Direct control is more generic, so the vertex handling was refined using dynamically weighted moving, which can be used to avoid motion conflicting with anatomic rules. The number of vertexes in Lucia is 60000.

3. DATABASE DESIGN AND COLLECTION

3.1. Preliminary lip-reading tests

This research study was started with several lip-reading experiments to measure the communication skills of deaf people, and to understand their everyday problems better.

One of our important conclusions was that visual lip-readability of the speech is greatly dependent on the quality of articulation. Lip-reading needs higher level attention to understand speech and misunderstandings are more frequent. Therefore clear articulation which emphasizes distinctive features and a slower speech rate can help a lot. The most lip-readable speakers within the hearing society are interpreters/lip-speakers. They have every day contacts with deaf persons so they are able to adapt their articulation for lip-reading. Therefore we have decided to employ interpreters to record our audiovisual database.

Hearing impaired persons usually have difficulties with sophisticated grammatical rules. They keep only the essential parts of messages in their mind. The context of traditional speech reference databases is too complex for them. So we have planned our audiovisual speech database both for the special training and tests. The text material of our database contains only two-digit numbers, names of months, and days of the week.

In the preliminary tests the importance of the third (z) dimension (depth) of space in perception of visual speech were also tested. Two types of distorted videos were presented to deaf subjects. In first the blue component was gained while red and green were attenuated, and in the other type the picture was binarized to black and white on the brightness of the pixels compared with a threshold. In binarized videos the depth information is almost completely hidden related to original ones, and blue videos represent an intermediate level in that sense. Surprisingly there was no significant difference in recognition rates.

Further experiments were organised with small displays using hand held mobile phones and deaf customers recognized well lip represented speech video records. This led us to the conclusion: only the area around the mouth is really important and enough to recognize speech.

3.2. Recording

Our database contains synchronised audio and video records about speaking persons. Audio and video data are processed in a synchronized way, each video frame has one audio frame. (see Fig. 3.) The head of speakers have been softly fixed to eliminate the motion of the head. Our system in the actual status is a speaker dependent solution but we plan a limited speaker independent version.

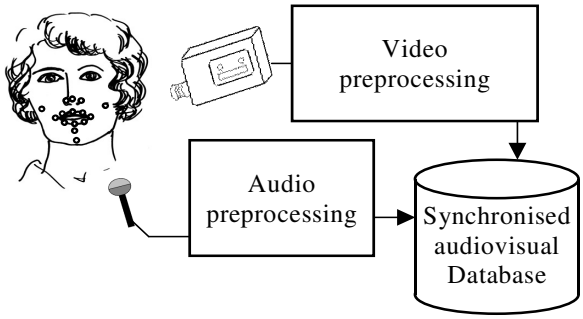

Fig. 3. Database collection

The MPEG-4 standard [6] describes the face with 86 FPs. We selected 15 FP around the mouth according to our preliminary results. (see Fig. 4) During recording the feature points were marked by yellow dots on the face of speakers. Our speakers were professional lip speakers

We used commercially available video cameras with 720x576 resolution, 25 fps PAL format video – which means 40ms for audio and video frames. The camera was focused onto the area of the mouth to get fine resolution for the selected FPs. On the video records deinterlace filter was applied to improve the efficiency of detection of marker points. Fast motions cause horizontal stripes on intelaced pictures, and that could corrupt the exact detection of marker points. The contrast, brightness and saturation were balanced to enhance the saliency of the marker points. The detection of the marker points was based on RGB components. Frames were binarized according to a statistically determinded threshold value. To get solid spots a dilation method was applied. The central pixel of the feature point is obtained by eroding the delateted marker spot. This method has 1-2 pixel maximum error. Since the average horizontal latitude of a FP-s are is 40-60, vertically 80-140 pixel, this error is acceptable. The origo has been chosen near to the nose, because this FP-s (see Fig. 4.) moves the least.

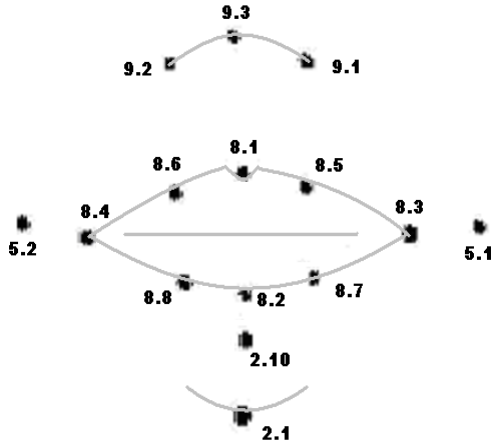

Fig. 4. Selected subset of MPEG-4 feature points

3.3. Principal Component Analysis

15 FP positions were tracked on *xy*-plane, so each frame is represented by 30 coordinates. In order to improve efficiency of training, these highly redundant vectors are compressed into 6 weight parameters ($w_1 \ldots w_6$) using PCA:

$$w_i = \underline{p}_i^T (\underline{x} - \underline{x}_{ref}) \; ; \quad i = 1 \ldots 6 \qquad (1)$$

Where $\underline{x}$ is the coordinate vector of the actual frame, $\underline{x}_{ref}$ is the vector of reference frame when the speaker was silent with closed lips, and $\underline{p}_i$ –s are the principal component vectors. The weight parameters are used to train NN. For convenience in implementation the principal component vector are scaled to get the weights between -1 and +1.

In operation (after training phase) our conversion algorithm estimates the coordinates from the weights supplied by NN. The recovery operation is:

$$\hat{\underline{x}} = \underline{x}_{bias} + \sum_{i=1}^{6} w_i \, \underline{p}_i \qquad (2)$$

The compression in our database causes only 1-3% loss of data, that is 1-2 pixel error in *xy*-coordinates which is acceptable in lip-reading. PCA is widely used in speech animation systems [7] due to its orthogonality feature which is utilized also in MPEG-4 Facial Animation standard.

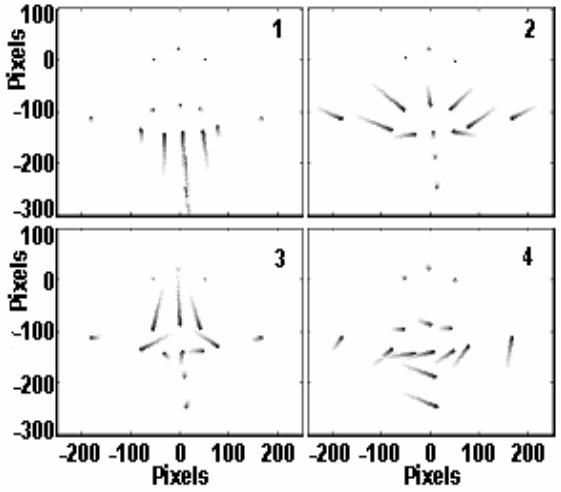

Fig. 5. The FP positions expressed by the 1st, 2nd 3rd and 4th principal components

4. EXPERIMENTS AND RESULTS

4.1. Preliminary tests

The preliminary tests were useful in the tuning of the system and in the modification of the database e.g. using professional interpreters/lip-speakers.

The preliminary tests have highlighted also the importance of inter speech breaks in the system. Low level background noise in audio speech perception do not cause any problem but even very small lip movements calculated from the background noise can disturb very much the lip-reading based speech perception.

After the preliminary test sessions, discussions were organised and remarks, comments, questions of

deaf test persons were collected and carefully considered in the refinement phase of our system.

4.2. Final tests

Lip reading the speech only phonemes could not be distincted perfectly, becouse some of them have identical viseme representation (like b-p). The natural way of recognition in those cases might be the estimation based on context or starting a dialogue to clarify the ambiguity. To avoid this type of interruptions the measuring text has to have some redundancy. Our text words were randomly selected from very limited sets. Two digit numbers, names of months and names of days were used in our test material, similarly to the training set.

During the final tests the complete head of the speaker was visible on large screen. The test subjects were told to answer the questions in a written form. The tests were composed from 70 short video clips. The complete test was taking about 30 minutes. In case of signed requests, the test stimulation were repeated. 18 deaf persons were involved in the tests. The test material has been composed randomly from three lip-reading situations. A- video records of interpreter/lip-speaker (no voice), B- face animation model controlled by 15 FP coordinates of the interpreter/lip-speaker (no voice), C- face animation model controlled by 15 FP coordinates calculated from speech signal (no voice). Final score of correctly recognized words: case A- 97,1%, case B- 54,9%, case C- 47,9%.

5. DISCUSSION

The visual word recognition even in the case of real and professional speaker has about 3% of errors.

The animated face model controlled by 15 FP parameters following accurately the FP parameters of interpreter/lip-speaker's face resulted about 42% of errors. After test discussions it was clarified, that the visible parts of tongue and movement of other parts of the face convey additional information to help the correct recognition. Probably the face model itself needs further improvements.

The decreasing of correct recognition only by about 7% as a result the complete changing of face model control from natural parameters to calculated parameters seems to be the fundamental result of our system.

6. CONCLUSION

The experiments and results have proved that the complete speech to facial animation conversion is possible on the level that provides communication aid for deaf persons.

Several components of the system have been implemented on smart mobile phones working in real time. The rest of the implementation on mobile phone is rather a technical question.

Further improvement of the facial animation model and enhancement of the conversion process could reduce the visual recognition error rate to the absolute tolerable 20% value.

The complete working system will be demonstrated on the conference.

ACKNOWLEDGEMENT

The authors would like to thank the National Office for Research and Technology for supporting the project in the frame of Contract No 472/04. Many thanks to our hearing impaired friends for participating in many-many tests and for their valuable advices, remarks.

Special thanks to Tamás Harczos for the valuable countribution to the project.

REFERENCES

[1] Gy. Takács, A. Tihanyi, T. Bárdi, G. Feldhoffer, B. Srancsik "Speech to facial animation conversion for deaf costumers" , *EUSIPCO 2006*, Florace, Italy, submitted paper.

[2] K. H. Choi, J. N. Hwang, "Constrained optimization for Audio-to Visual Conversion," *IEEE Transactions on Signal Processing*, Vol. 52. No. 6, June 2004, pp. 1783-1790.

[3] R. Gutierrez-Osuna, P.K. Kakumanu, A, Esposito, O. N. Garcia, A. Bojorquez, J.L Castillo and I. Rudomin, "Speech-driven Facial Animation with Realistic Dynamics" *IEEE Transactions on Multimedia*, Vol. 7., February 2005, pp. 33-42.

[4] J. Beskow, *Talking Heads, Models and Applications for Multimodal Speech Synthesis,* Doctoral Dissertation, KTH, Stockholm, 2003.

[5] P. Cosi, A. Fusaro, G. Tisato, "LUCIA a New Italian Talking-Head Based on a Modified Cohen-Massaro's Labial Coarticulation Model", *Eurospeech 2003*, Geneva, Switzerland, September 2003, pp. 2269-2272.

[6] I. S. Pandzic, R. Forchheimer (Eds.), *MPEG-4 Facial Animation: The Standard, Implementation and Applications*, Wiley, Chichester, 2002.

[7] J. Beskow, M. Nordenberg, "Data-driven Synthesis of Expressive Visual Speech using an MPEG-4 Talking Head", *Interspeech 2005*, Lisbon, Portugal, Septemper 2005, pp. 793-796.