14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, September 4-8, 2006, copyright by EURASIP

SPEECH TO FACIAL ANIMATION CONVERSION FOR DEAF CUSTOMERS

György Takács, Attila Tihanyi, Tamás Bárdi, Gergely Feldhoffer, Bálint Srancsik

Faculty of Information Technology, Péter Pázmány Catholic University
H 1083 Práter u. 50/a., Budapest, Hungary, phone: + (36) 1886 4763 , fax: + (36) 1 886 4724,
email: {takacs.gyorgy, tihanyia, bardi, flugi, sraba}@itk.ppke.hu

ABSTRACT

A speech to facial animation direct conversion system was developed as a communication aid for deaf people. Utilizing the preliminary test results a specific database was constructed from audio and visual records of professional lip-speakers. The standardized MPEG-4 system was used to animate the speaking face model. The trained neural net is able to calculate the principal component weights of feature points from the speech frames. The control coordinates have been calculated from PC weights. The whole system can be implemented in standard mobile phones. Deaf persons were able correctly recognize about 50% of words from limited sets in the final test based on our facial animation model.

1. INTRODUCTION

The primary aim of our project was to develop a communication aid for deaf persons which can be implemented in a mobile telephone. In our system a partially animated face is displayed in interaction with deaf users. The control parameters of the animation are calculated directly from the input speech signal. It is known that showing the face itself is a limited representation of the human speech process and contains inherent errors, though deaf people have fantastic abilities in understanding speech based on lip reading only. In spite of the limitations deaf persons aided with special software on the platform of high-end class second or third generation mobile phones could naturally communicate with hearing people.

KTH's Synface system [5] can successfully aid the communication of hard of hearing people by speech to animation conversion. Synface shows the lip movements of the speaker at the other telephone synchronized with the speech sound. Our intention is to aid completely deaf people which can be based on visual modality only.

Speech to animation conversion in Synface is divided to a phoneme recognition module [7, 11] and a visual speech syntheser [2], which is driven by the phoneme string. In our system only continuous types of transformations are used in the complete audio to visual conversion, no discrete classification method is applied. One of the benefits of our direct solution is that the original temporal and energy structure of the speech are retained, so the naturalness of rhythm is guarantied. Further benefit is the relatively easy implementation in mobile phone environments with limited memory and computation power. A rather promising feature of our system is the potentially language independent operation.

A very important element of our concept is to train the system on a unique audio-visual database collected from professional interpreters/lip-speakers. Their articulation style and level are adapted to deaf communication partners.

One of the known solution groups use animated talking faces to increase speech intelligibility gaining additional information related to the only auditory situation in cases of noisy environment or hearing-impaired listeners. This so-called superadditive combination of auditory and visual speech can produce a bimodal accuracy, which is greater than the simple sum of their separate unimodal performances [8, 9, 10]. In our application only a unimodal communication is supposed owing to the deaf users. We focus on the visual input only and calculate with the enhanced abilities of deaf persons.

The dynamics of mouth movements and the naturalness of face animation models seem to be critical parts of audio-to-visual conversion. Usually researchers elaborate very sophisticated procedures to produce dynamic and natural talking heads [3, 6]. We have selected the speakers for the data base recording with special attention to the high dynamic requirements.

2. DATABASE DESIGN AND COLLECTION

2.1 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. In this paper only a brief summary of the results and conclusions is presented.

One of our important conclusions was that visual lip-readability of the speech greatly depends 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.

We also found that hearing impaired persons usually have difficulties with grammatical rules. The context of traditional speech reference databases is too complex for them.

They often do not employ suffixes properly and are unable to follow sudden changes in the topic. They are usually unable to repeat complete sentences word by word. They keep only the essential parts of messages in their mind instead. We have planned our audiovisual speech database for both algorithm training and intelligibility 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 speaking faces 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 compared to the 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.

2.2 Recording

Our database contains synchronised audio and video records from speaking persons. Audio and video data are processed in synchronized way; each video frame has one audio frame. (see Fig. 1.) 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.

The MPEG-4 standard describes the face with 86 feature points (FP). We selected 15 FP around the mouth according to our preliminary results. During recording the feature points were marked by yellow dots on the face of speakers. The advantages of MPEG-4 FP-s are the compressed description of face parameters and the compatibility to standard facial animation models. Our speakers were professional lip speakers

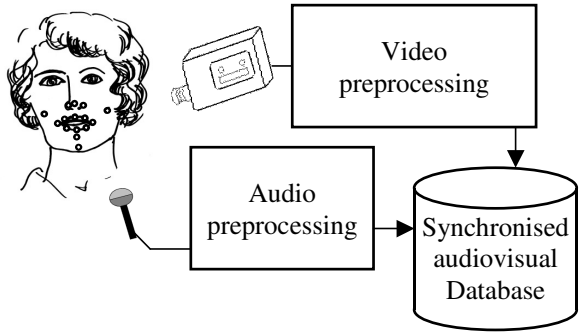

Figure 1 - Database collection

We used commercially available video cameras with 720x576 resolution, 25 fps PAL format video – which means 40ms for audio and video frames. The video recordings have concentrated only on the area of the mouth and vicinity to let maximum resolution for FP extraction. The video records were then processed automatically. After de-interlacing, we balanced the contrast, brightness and saturation of the video records to enhance the marker points. The selection of the marker points was based on RGB components. On the binarized, dilated and eroded picture every spot covers only one pixel in the centre of the marker. 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 (Fig. 2.) moves the least. The input speech sound is sampled at 16 bit/48 kHz.

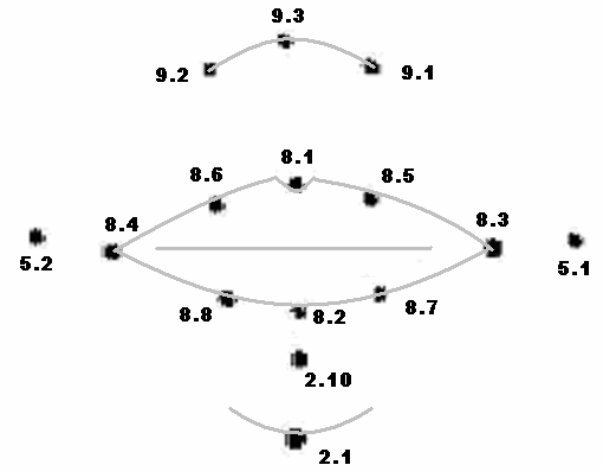

Figure 2 - Selected subset of MPEG-4 feature points

3. CONVERTING SPEECH SIGNAL TO FACIAL ANIMATION

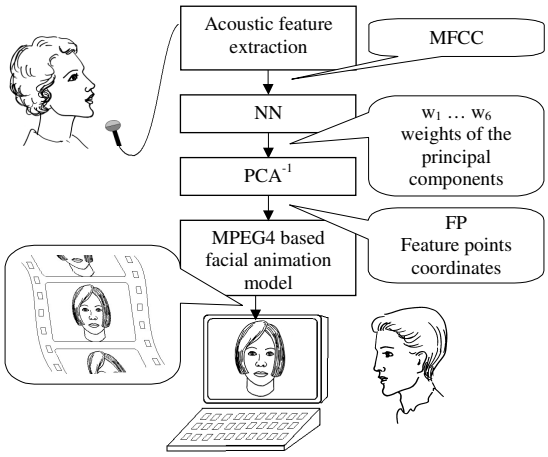

Figure 3 - Structure of the implemented speech to facial animation system

Our implemented conversion system is PC-based software. Here we survey the complete system at a glance, as it is shown in Figure 3, and the details of the building blocks are detailed in points 3.1.-3.4.

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 special weighting vector $[w_1,...w_6]$ that is a 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 our component vectors with the weights coming from the neural net. This coordinate-recovery operation is denoted by the term "PCA^{-1}" in the block diagram, because the predefined component vectors come from Principal Component Analysis (PCA). The FP positions are computed in this way for 25 frames per second. (Fig. 5)

The final component in our system is a modified LUCIA talking head model [4]. We control it with the computed FP coordinates and then the facial animation model appears on the screen. See chapter 3.4.

3.1 Acoustic feature extraction

The input speech is pre-emphasis filtered with $H(z)=1-0.983z^{-1}$. Then 21.33ms long Hamming windows are applied to the signal, and 16 Mel-Frequency Cepstrum Coefficients (MFCC) are extracted from each analysis window. 5 analysis windows are processed this way for every frame of the animation, the middle one is centred to the time position of the actual frame. The windows are placed with 40 ms distance between them (Fig. 4).

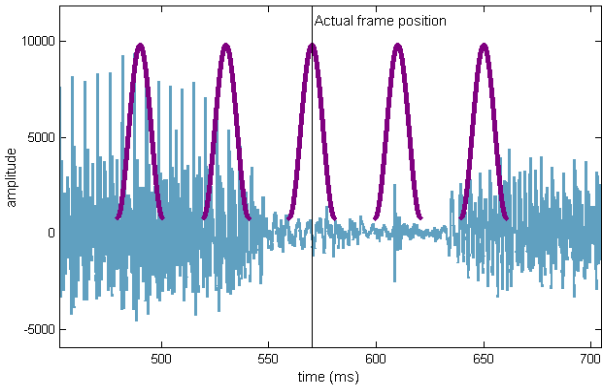

Figure 4 - Acoustic information from 5 windows for each frame of animation

Co-articulation phenomenon has great importance in both visual speech and acoustics, but in different ways. Phonemes can be visually dominant or flexible. The dominant ones have typical visual shape and highly affect the figure of the neighbouring flexible phonemes. The flexible ones tend to suffer the effects of the dominant neighbours. Analysing our database the closest relation between acoustic and visual parameters was found at the steady state part of the visually dominant phonemes, these parts behave as anchors between the two modalities. During flexible phonemes the relation is much less determinate. In our experience it is advantageous if the conversion algorithm accesses some acoustic information from the steady state part of at least one of the neighbouring phonemes.

When professional lip-speakers talk to deaf people the speech rate falls down to 5-10 phoneme/sec. The steady state phases of dominant phonemes are emphasized. Our 5 windows partially cover about 180 ms, and likely at least one of them reaches the quasi-stationary phase of a dominant phoneme, which provides reliable information about the visual shape.

MFCCs from the 5 consecutive windows are sent to the input layer of the artificial neural network (ANN).

3.2 Neural Network

Matrix Back-propagation algorithm (elaborated by David Anguita) was used to train the network [1]. The NN has 3 layers: 80 elements in the input layer to receive 16 MFCC from 5 frames. The hidden layer consists of 40 nodes. The output layer has 6 nodes, providing 6 centered PCA weights to reproduce the FPs.

The training file contained 5450 frames. The network was trained by 100000 epochs. This NN model uses interval [-1,1] both for inputs and outputs. MFCC and PCA vectors are scaled linearly by the same value to fit to the appropriate range except the energy coefficient of MFCC.

The program of the trained neural network runs very fast in the working system, since all the database is represented in the weights of the network, consisting from 3440 coefficients only. So the calculation has constant and tolerable time consumption and realisable by mobile phones.

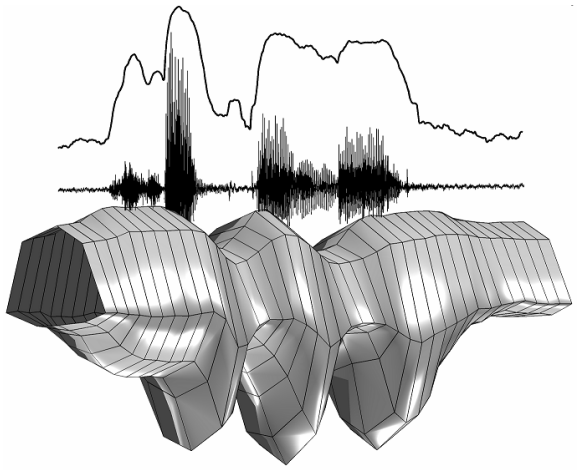

Figure 5 - The x-y components of 8.1-8.8 FP-s as a function of time pronouncing word "September". The upper solid line shows the frame energy in dB, the middle graph represents the waveform, the lower surface represents the lip contours.

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:

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

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, which is 1-2 pixel error in *xy*-coordinates which is acceptable in lip-reading. This operation needs only 180 multiplications. PCA is widely used in speech animation systems due to its orthogonality feature which is utilized also in MPEG-4 Facial Animation standard.

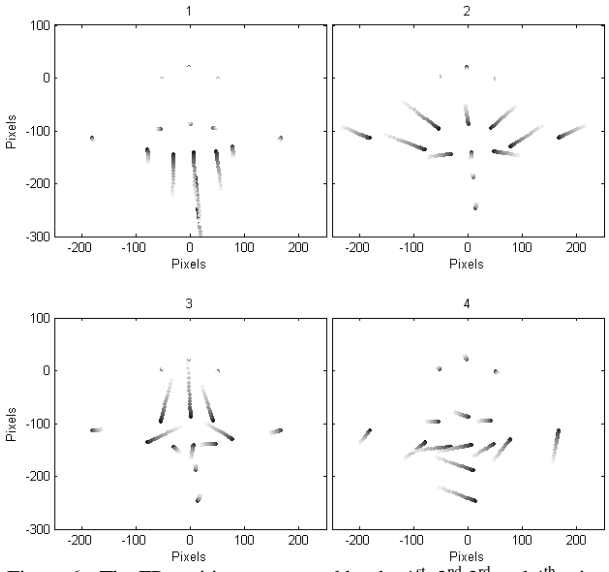

Figure 6 - The FP positions expressed by the 1st, 2nd 3rd and 4th principal components using professional lip-speaker

PCA is capable for more than dimension reduction. The PCA vectors contain important information about the lip speaker and the recording. There are principal components, which have recognizable role in viseme distinction. (Fig. 6) For example the vertical movement of the jaw-bone is the strongest principal component. Also recognizable principal component are horizontal tension ("cheese" mouth at photos) and lip-rounding. These components have viseme distinctive functions.

In this point of view, only the order of the components is important. If the speaker is professional lip speaker, the strongest components are viseme distinctive. However, if the speaker is an unqualified speaker, the correctional components (eg. emotional) can outrun viseme distinctive components. As it can be seen in Figure 7, the second strongest component shows that the speaker's articulation is not acceptable for training.

3.4 Talking head

We applied Lucia talking head model [10] 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.

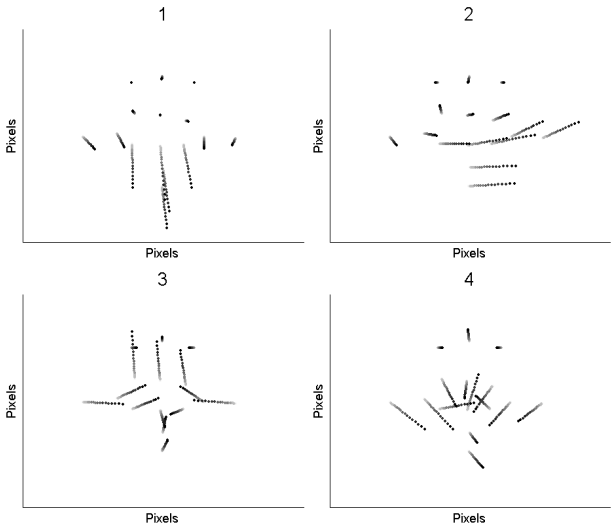

Figure 7 - The FP positions expressed by the 1st, 2nd 3rd and 4th principal components using unqualified lip-speaker

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 a discussions were organised and remarks, comments, questions of deaf test persons were collected and carefully considered in refinement steps of our system.

4.2 Final tests

Lip reading the speech only, phonemes could not be distinguished perfectly, because 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 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).

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 video clip was repeated. 18 deaf persons were involved in the tests. See Fig 8.

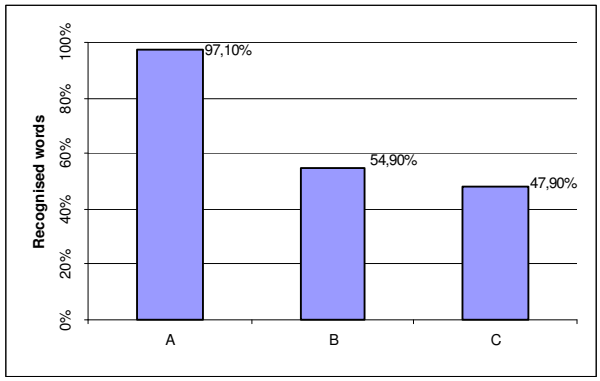

Figure 8 - The ratio of correctly recognised words.
A – video records of interpreter/lip-speaker,
B – face animation model controlled by FPs os speaker,
C – face model controlled by FP coordinates calculated from speech signal

4.3 Discussion

The visual word recognition even in the case of natural 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 parts of the face others then the mouth 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.

5. 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. Remember, that deaf people have no other natural and efficient telephone based communication with the hearing society.

6. 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 and remarks.

REFERENCES

[1] D. Anguita, "Matrix Back Propagation - An efcient implementation of the BP algorithm" *Technical Report*, DIBE - University of Genova, Nov. 1993.

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

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

[4] P. Cosi, A. Fusaro, G. Tisato, "LUCIA a New Italian Talking-Head Based on a Modified Cohen-Massaro's Labial Coarticulation Model*", Proceedings of Eurospeech 2003*, Geneva, Switzerland, September 1, 2003, Vol. III, pp. 2269-2272

[5] B. Granström, I. Karlsson, K-E Spens: „SYNFACE – a project presentation" *Proc of Fonetik* 2002, TMH-QPSR, 44: 93-96

[6] 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. pp. 33-42, February 2005.

[7] M. Johansson, M. Blomberg, K. Elenius, L.E.Hoffsten, A. Torberger, "Phoneme recognition for the hearing impaired," *TMH-QPSR*. vol 44 –Fonetik pp. 109-112, 2002.

[8] D. W. Massaro, *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle* Cambridge, Mass: MIT Press, 1998.

[9] D. W. Massaro, D. G. Stork, "Speech Recognition and Sensory Integration," *American Scientist,* 86. 1998

[10] J. Ostermann, "Animation of Synthetic Faces in MPEG-4", *Computer Animation*, pp. 49-51, Philadelphia, Pennsylvania, June 8-10, 1998.

[11] G. Salvi: „Truncation error and dynamics in very low latency phonetic recognition" *Proc of ISCA workshop on Non-linear Speech Processing (2003)*