Signal Conversion from Natural Audio Speech to Synthetic Visible Speech

György Takács*, Attila Tihanyi*, Tamás Bárdi*, Gergely Feldhoffer*, Bálint Srancsik*
*Faculty of Information Technology,
Péter Pázmány Catholic University,
50 Práter Street, 1083 Budapest, Hungary,
e-mail: {takacsgy, tihanyia, bardi, flugi, sraba}@itk.ppke.hu

***Abstract*—A speech to facial animation direct conversion system is summarized in this paper. It is planned to provide a communication aid for deaf people. The system is composed from published speech and image processing tools and methods, as neural network, PCA and MPEG-4 compliant talking heads. This system forms a new concept for a direct conversion from audio to visual avoiding any discrete classification inside. The specialty of our system is the network training material obtained from professional lip-speakers, and the whole system can be implemented in mobile phones.**

I. Introduction

Several systems are known to convert audio speech signal into visible form. The purpose and the solution of published systems are different. Hard of hearing people and also normal hearing persons in noisy environment can utilize visual information from the speaker's face in addition to the speech sound. Talking artificial agents can provide very human like interactions between computers and users. In such applications the speech signal and face movements must be well synchronized.

KTH's Synface system [1] can successfully aid the communication of hard of hearing people by speech to animation conversion. Synface shows animated lip movements calculated from the incoming voice call. 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 [2, 3] and a visual speech synthesizer [4], 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 this newly developed 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.

In our system only the mouth and its surrounding part is animated for deaf users. It is known that showing the face itself is a limited representation of the human speech process and contains inherent errors, but deaf people have fantastic abilities in understanding speech based on lip reading only. In spite of the limitations deaf persons could naturally communicate with hearing people using our system.

The dynamics of mouth movements and the naturalness of face animation models seem to be critical in lip-readable audio-to-visual conversion. Usually researchers elaborate very sophisticated procedures to produce dynamic and natural talking heads [5]. We have selected the speakers for the data base recording with special attention to the high dynamic requirements.

II. Database Design and Collection

A. Preliminary lip-reading tests

This research study was started with several lip-reading experiments to measure the communication skills of deaf people [6, 7]. Lip-readability of the speech greatly depends on the quality of articulation. The most lip-readable speakers within the hearing society are interpreters/lip-speakers. We have decided to employ interpreters to record our audiovisual database.

The importance of 3D in lip-reading was also examined with binarized videos, where the depth information is almost completely hidden. There was no significant decreasing in recognition rates, so 2D face models can be adequate for such applications.

B. Database recording

Our audio-visual database contains synchronized audio and video records of speech from professional lip-speakers. The head of speakers was softly fixed to eliminate the motion of the head. 15 selected feature points (FP) out of the 86 in MPEG-4 Facial Animation Standard were marked with yellow dots and tracked in 2D on the pictures.

Commercially available camera was used with 720x576 resolution, 25 fps in PAL format videos. The camera zoomed to the area of mouth for better FP position detection. The input speech sound is sampled at 16 bit/48 kHz.

III. Converting Speech Signal to Facial Animation

Our implemented conversion system is PC-based software. Here we survey the complete system at a glance, as it is shown in Fig. 1, and the details of the building blocks are detailed in subsections.

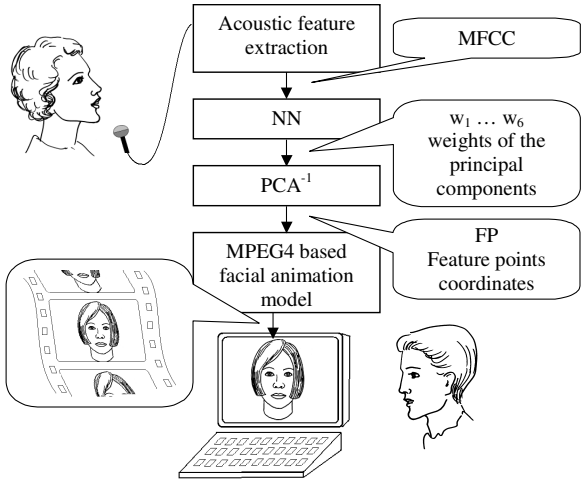

Fig. 1. Structure of the implemented speech to facial animation system

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. 2.)

The final component in our system is a modified LUCIA talking head model [8]. We control it with the computed FP coordinates and then the facial animation model appears on the screen.

A. 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. Five 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. 3).

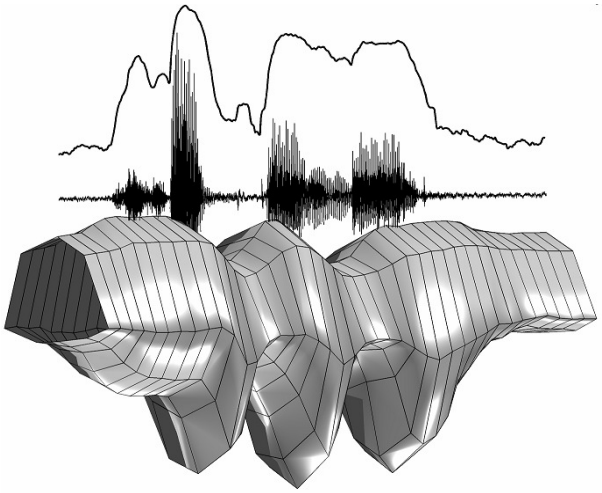

Fig. 2. The x-y components of 8.1-8.8 FP-s as a function of time pronouncing word "September". The upper solid line shows the frame energy in dB, the middle graph represents the waveform, the lower surface represents the lip contours.

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.

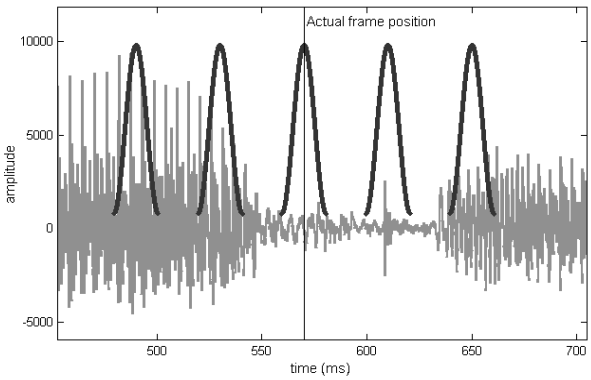

Fig. 3. Acoustic information from 5 windows for each frame of animation

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).

B. Neural Network

Our NN is a 3 layer perceptron network, which maps the acoustic representation of speech to visual. In our implementation Anguita's Matrix Back Propagation open source toolkit [9] is used both for training end evaluation. The input layer has 80 nodes to receive the 5*16 coefficients of the acoustic feature vector. The hidden layer has 40 nodes and the output layer has 6 nodes for the weights of the 6 principal component used. Component weights and also MFCCs are scaled to get them distributed throughout [-0.9, 0.9] interval, which is required for the NN. The scaling is linear.

The network was trained on a set of patterns extracted from our audiovisual database. The training set contains 5450 frames of coupled MFCC feature vector and component weight vector. The training session took 100000 epochs. Reducing the dimension of visual feature vectors highly influences the effectiveness of the training. Without applying PCA the NN was unable to find appropriate mapping even after millions of epochs.

We found that the simple PCA gives better result in our system than Independent Component Analysis (ICA). Using ICA for dimension reduction, learning of NN begins faster with ICA than with PCA, but finally the latter gives better approximation. The learning error of NN was significantly less for ICA during the first 10000 epochs of training. Nevertheless after 20000 epochs that relation always rolled over and PCA gave more accurate approximation than ICA. This interesting turn-on took place presumably because independency provides somehow broader valleys in the error surface, while those valleys are deeper for PCA.

The trained NN runs very fast, since all of its "knowledge" retrieved from the database is represented in the synaptic weights of the network, consisting of 3440 coefficients in aggregate. Each synaptic weight occurs only once per frame in computation of scalar products, so it needs tolerable amount of operations even in the processor of a smart-phone, and can be applied in real-time.

C. 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:

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

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. These weight parameters were used to train NN.

Operating the trained network our conversion algorithm estimates the coordinates from the weights supplied by NN. The recovery operation is:

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

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.

Analysis of principal components in our research study is not for dimension reduction only, but it is also useful in understanding the differences between the articulation style and quality of speakers. We also analyzed what kind of lip or facial motions are represented when the FP coordinates are shifted towards at the direction of a certain principal component. We found that some of the components are related with visual phoneme distinctive features, some of them belong to facial expressions, and the rest of them are for statistically rear lip movements diffused with the noise of pixel errors in marker point detection.

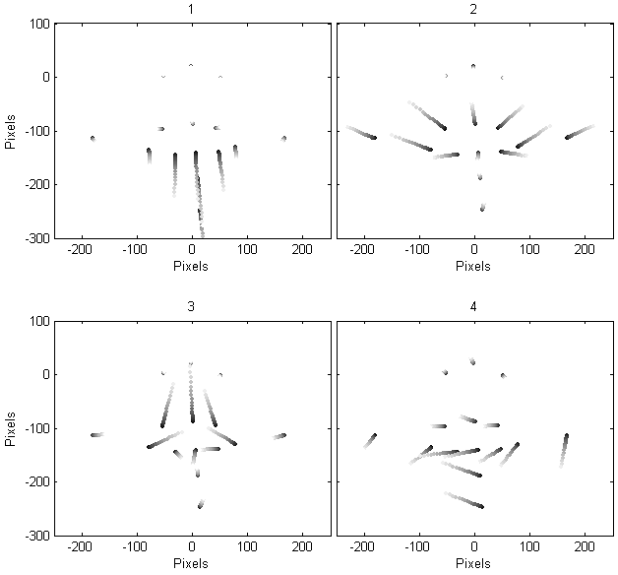

Fig. 4. The FP positions expressed by the 1st, 2nd 3rd and 4th principal components using professional lip-speaker

Fig. 4 shows the motions of the 15 marked FPs in the directions of the first 4 Principal Component vectors. These components are from the analyzed data of a professional lip-speaker (interpreter).

D. Talking-head models

Due to the fact that we used MPEG-4 standard to characterize the visible speech, we were able to drive many talking heads. We tried Greta [10] with 9138 polygon and Lucia [8] with 27327 poly from Cosi & al, Alice [11] with 13412 poly from the X-Face tool, and created our own engine and talking head based on modified free VRML John modell with only 632 polygon, which designed for mobile phones.

All of them use the animation standard of MPEG-4 called face animation parameters (FAP). Since FAP is a viseme

based animation method, we have modified Lucia to accept directly FP-s. 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.

VRML format model was applied in the 3D animation. MPEG-4 feature points fitting to the model were calculated beside the 3D model. So we have an extended model which can be modified flexibly by an independent program procedure engine.

The feature point coordinates are obtained from the lip-speaker persons so this means that the relations of feature point coordinates are automatically fulfilled.

According to the MPEG-4 standard the extended 3D model and the engine are separated.

The engine can drive different 3D extended models.

The whole procedure can be implemented only if the feature extraction, along with the feature point animation analysis of the lip speakers are managed according to the MPEG-4 measure rules.

IV. Results

Lip reading the speech only phonemes could not be distincted perfectly. 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. 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%.

V. 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 of the complete changing of face model control from natural parameters to calculated parameters seems to be the fundamental result of our system.

VI. 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 which work in real time.

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.

Acknowledgment

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

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

[3] G. Salvi: "Truncation error and dynamics in very low latency phonetic recognition" *Proc. of ISCA Workshop on Non-linear Speech Processing*, 2003.

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

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

[6] G. Takács, A. Tihanyi, T. Bárdi, G. Feldhoffer, B. Srancsik: "Speech to facial animation conversion for deaf applications" *14th European Signal Processing Conf.*, Florence, Italy, September 2006.

[7] G. Takács, A. Tihanyi, T. Bárdi, G. Feldhoffer, B. Srancsik: "Database Construction for Speech to Lip-readable Animation Conversion" *48th Int. Symp. ELMAR-2006 on Multimedia Signal Processing and Communications*, Zadar, Croatia, June 2006.

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

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

[10] Pelachaud C., Marno Caldognetto E., Zmarich C., Cosi P., "Modelling an Italian Talkting Head", in Proceedings of AVSP 2001, Aalborg, Denmark, Septembre 7-9 2001, pp 72-77

[11] Balci, K. Xface: MPEG-4 based open source toolkit for 3D facial animation. In Proc. Advance Visual Interface, 2004