Inter-speaker Synchronization in Audiovisual Database for Lip-readable Speech to Animation Conversion

Gergely Feldhoffer, Balázs Oroszi, György Takács, Attila Tihanyi, Tamás Bárdi

Faculty of Information Technology, Péter Pázmány Catholic University, Budapest, Hungary
{flugi, oroba, takacsgy, tihanyia, bardi}@digitus.itk.ppke.hu

Abstract. The present study proposes an inter-speaker audiovisual synchronization method to decrease the speaker dependency of our direct speech to animation conversion system. Our aim is to convert an everyday speaker's voice to lip-readable facial animation for hearing impaired users. This conversion needs mixed training data: acoustic features from normal speakers coupled with visual features from professional lip-speakers. Audio and video data of normal and professional speakers were synchronized with Dynamic Time Warping method. Quality and usefulness of the synchronization were investigated in subjective test with measuring noticeable conflicts between the audio and visual part of speech stimuli. An objective test was done also, training neural network on the synchronized audiovisual data with increasing number of speakers.

Keywords: audiovisual, database, facial animation, dynamic time warping

1 Introduction

Our language independent and direct speech to facial animation conversion system was introduced in [1], [2], and [3]. The main system components are the acoustic feature extraction, the feature point coordinate vector calculation and running a standard MPEG4 face animation model as it is shown in Figure 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 special weighting vector that is a compressed representation of the target frame of the animation.

A conceptual element of our system is to process only continuous acoustical and facial parameters. The traditional solutions transform the continuous process of speech into a discrete set of language elements like phonemes and visemes. The second part of the traditional systems converts discrete text or phoneme strings into animated faces. The modular structure and the separated training of the elements mean the main benefit of such discrete element based systems. [4],[5] Their problems are the accumulated error and the lost original temporal and dynamic structure.

One of the benefits of our direct solution is the reservation of the original temporal and energy structure of the speech process. Thus the naturalness of rhythm is guarantied. Further benefit of our solution is a relatively easy implementation in the environment of limited computational and memory resources. A rather promising feature of our system is the language independent operational capacity.

The single speaker version works quite well. Deaf persons can recognize about 50% of the words correctly. The neural network has been trained by speech parameters at the input and principal components of the facial parameters at the output and parameter pairs have originated from the same speaker [1].

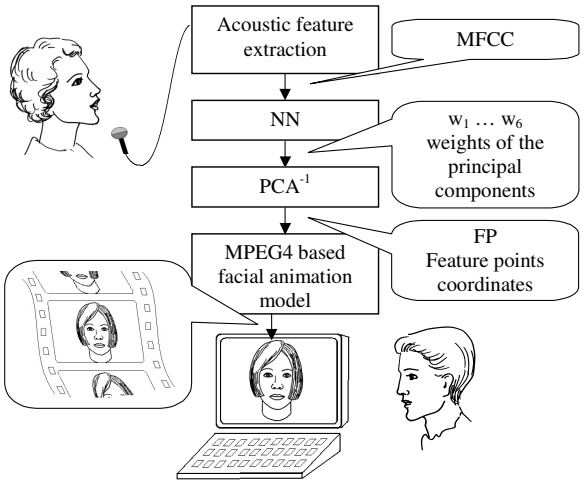

Fig. 1: Structure of the direct speech to facial animation conversion system.

In this paper we report on the results of speaker variation tests in our system. The main technical problem has been caused by timing differences in the utterances of different speakers. In traditional discrete systems the labeling of small units by phoneme codes can eliminate the timing problems. The training and testing procedures are based on phoneme size segments [1], [2].

Dynamic Time Warping algorithm (DTW) is a well refined procedure in time matching of two utterances and widely used in the isolated word recognition systems. This algorithm does not need an exact segmentation within complete utterances. So a special version of DTW has been applied in our program.

Several theoretical questions have been raised in the investigation of multiple speaker situations. Hard of hearing persons stated that the lip-readability of speakers is very different due to the lip level articulation, while the intelligibility of their voice is similar to that of the hearing people. How can we optimally train the neural network? Whether train by a high number of speakers representing the average population or train by a carefully selected group of lip-readable persons? We do not need to transform a speech signal to its original face movement. Rather we need a transformation to one easily lip-readable face. How can we test the efficiency of time matching? How can we measure objectively the performance of the system changing the speakers?

2 Database construction

The traditional audio-visual databases are elaborated for testing by hearing people. Our database was constructed according to our special specifications.

Professional interpreters/lip-speakers were invited to the record sessions. Their speech tempo and articulation style is accommodated to the communication with hard of hearing persons. Their articulation emphasizes visible distinctive features of speech.

Sentences from a popular Hungarian novel were selected. The selection differs from the phonetically balanced criteria. The rare phonetic elements and visually confusable phoneme pairs were represented on higher level, than their average probability. The records consist of 100 sentences.

The head of speakers have been softly fixed to reduce the motion of the head. We used commercially available video cameras with 720x576 resolutions, 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. The text was visible on a big screen behind the camera. The camera produced portrait position picture to maximize the resolution on the desired area of the face.

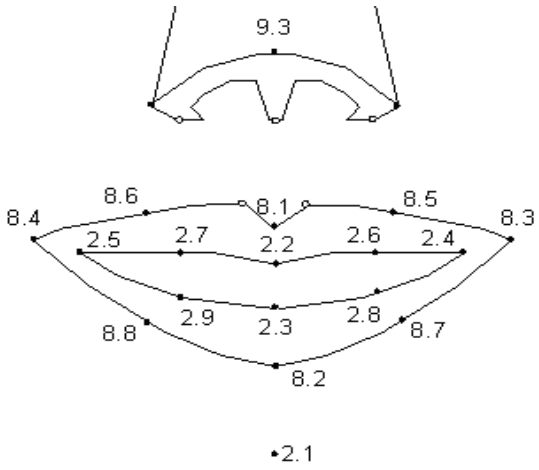

Fig. 2: The applied MPEG-4 feature point (FP) subset.

Yellow markers were used on the nose and chin of speakers as reference points. Red lipstick emphasized the lip color for easier detection of lip contours. The records were stored on digital tape and then copied into a PC.

The records were prepared in an acoustically treaded room by a dynamic microphone. The input speech sound is sampled at 16 bit/48 kHz. The speakers could repeat sentences in case of wrong pronunciations. The audio and video files were synchronized by “papapapa” phrases recorded at the beginning and end of files. The closures end opening of explosives are definite both in audio and video files.

The video signal was processed frame by frame. The first step was the identification of yellow dots on the nose and on the chin based on their color and brightness. The color values of the red lips were manually tuned to get the optimal YUV parameters for identification of internal and external lip contours. On the shape of lips the left most and right most points identified the MPEG-4 standard FP-s 8.4 and 8.3. The further FP coordinates were determined at the cross points of halving vertical lines and lip contours. FPs around the internal contour were located similarly. An extra module calculated the internal FPs in the cases of closed lips. The FP XY coordinates can be described by a 36 element vector frame by frame. The first 6 principal component values (PCA) were used to compress the number of video features.

The input speech was pre-emphasis filtered then in 40 ms frames 1024 element FFT with Hamming windows were applied to gain spectrum vectors. Next step calculated 16 Mel-Frequency Cepstrum Coefficients (MFCC) in each frame.

2.1 The DTW procedure

The temporal matching has the following task: the frame "i" in the audio and video records of Speaker A has a corresponding frame "j" in the records of speaker B. The correspondence means the most similar acoustic features and lip position parameters. Each speaker read the same text so several corresponding frames are evident for example at sentence beginnings. The time warping algorithm in the isolated word recognition can provide a suboptimal matching between frame series. [6]

The voice records were used for warping. Voice frames are characterized by MFCC feature vectors. The distance metric in the DTW algorithm was the sum of MFCC coefficient absolute differences. The total frame number in records A and B might be different.

In the database 40 sentences of 5 speakers were warped to the files of all other speakers. This warp is represented by indexing the video frames for 40 ms audio windows as possible jumps or repeats.

3. Experiments and results

Subjective tests were used to measure the quality of time warping in cases of changing of speakers in audio and visual parts of records.

An objective measure, the average error of calculated FP coordinates related to natural FP parameters of speaker A was investigated as a function of number of speakers involved in the training. The neural net has 16 input nodes, 20 hidden nodes and 6 output nodes. The input parameters are the MFCC parameter values of the voice. The output parameters are the principal component values of the FP coordinates. So the system calculates face animation parameters from the voice features. The training errors express the average value of distances of the calculated FP coordinates from the target FP coordinates. The voice might be from speaker A, B, C and D and the target FP coordinates are the elements of PF coordinates are from the video records of speaker A.

3.1 Subjective Testing of the Time Warping

The quality test of matching of the records needs a subjective assessment. For this reason we prepared a test sequence of voice and video records of feature points. The audio part and video parts were taken randomly from the same records and other cases the voice parts were from different speakers and the video frames were warped to the audio frames spoken by the different speakers. In the case of ideal warping the test persons could not differentiate the audio-video pairs whether they were from the same records or warped different records. In the test they expressed the opinion on a scale: 5- surely identical, 4-probably identical, 3-uncertain, 2-probably different, 1-surely different the origin of the voice and video.

3.2 Results of the DTW tests

21 test persons watched the 15 audio-video pairs in random order. The results are summarized in Figure 3.

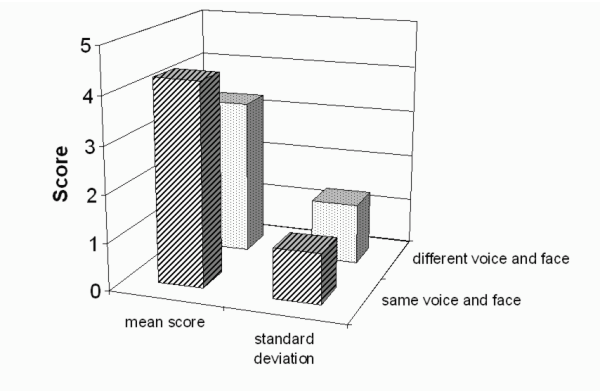

Fig. 3: Results of the subjective DTW tests. The opinion score values: 5- surely identical, 4- probably identical, 3-uncertain, 2-probably different, 1-surely different the origin of the voice and video

The cases when the audio and video parts of the records were from the same persons the average score was 4.2. The average opinion expressed, that the test persons can differentiate original and warped pairs at some level.

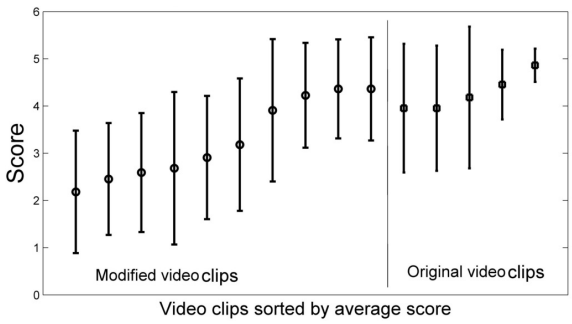

Fig. 4: Ratings (average score and deviation) of video clips grouped by modified and original video data, sorted by average score.

The same time the warped pairs have a score of 3.2 so it is in between probably identical and uncertain value. This score value proves that the warping is good because the result is not on the "different" side.

The video clips were rearranged in a way that we put into one group the modified and in other group the original ones during the evaluation process. The results are in Figure 4. The standard deviation values are overlapped within the two groups.

3.3 Training and testing the system by a single speaker

In this experiment the records of 5 speakers were studied. Matrix Back-propagation algorithm was used to train the neural network. Speaker A, B, C and D were used for the training of the neural net and speaker E was used for testing. In this section the records of speaker A was used for training and Speaker E for testing. The average training error is a good indicator for the average FP coordinate error. Figure 5 shows, that after 10.000 epochs the training error approximately reaches the value below one pixel during independent training processes.
Meantime the test error value with person E reaches the level of 1.5 pixels.

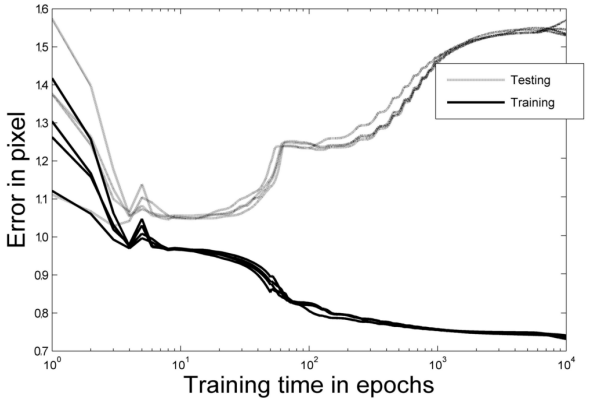

Fig. 5: The training error and test error values as a function of training epoch number. Training with records of person A and test with warped records of person E.

3.4 Training and testing the system by multiple speakers

In this section we repeated the training and testing procedures involving several speakers in the training. The test results below represent the average pixel errors in case of different training situations. The training sequence was the following:

person A voice parameters as input and person A video parameters as target values of the net

the previous + person B voice parameters to A as input and warped A video parameters as target values of the net

the previous + person C voice parameters to A as input and warped A video parameters as target values of the net

the previous + person D voice parameters to A as input and warped A video parameters as target values of the net

The training and testing error values are shown in Figure 6. More and more persons involved in the training caused higher training error and lower testing error values. The ratio of testing error and training error is diminished from more than 100% to 7%.

3.5 Discussion

Voice features of four speakers were involved in the training of the neural net. The average pixel error of the calculated FP coordinates related to the original FP coordinates of speaker A. The values of FP test error (testing with speaker E whose data were not involved in the training) decrease from 1.5 to 1.0 by increasing the number of speakers from 1 to 4. In the case we involved more speakers in the training there was no considerable decrease in FP error. The training error increases because of the higher variations in the training set.

Fig. 6: Variation of training and testing errors involving 1-4 speakers in the training procedure.

The decreased value of the test errors proves that the time warping works well. The test error value in case of 4 speakers in the training is only 7% higher then the training error. This means that four speakers represent well the speaker variations in the system.

It is possible to express the error in ratio of the domains. On a PAL screen which contains only a face from the eyes to the neck, the average domain size of a feature point coordinate is about 40 pixels. 1 pixel error means about 2.5% error this way. The reason of using pixel as a unit of error is the database recording method. The error of FP calculations has 2 pixels in the training material. This calculation error is the ±1 pixel uncertainty of the automatic detection algorithm. The error range of the calculated FP coordinates from the voice parameters and the detection uncertainty error are comparable.

4 Conclusions

The speaker dependent variations in the direct calculation of face animation parameters from the voice parameters can be treated by the applied methodology. DTW is a convenient solution to compensate the lack of phoneme level in multi-personal issues of speech-to-animation conversion.

The subjective tests proved that the time warping based on voice parameters can map well the speech process of speaker A into the speech process of speaker B who has the same text. The level of testing error is lower than the critical error which disturbs the lip reading for hard of hearing persons [1]. So the training of the conversion system performs the speaker independent criteria on the required level.

The increasing of the number of feature points around the inner contour of lips improved the readability of the face animation.

It is easier to implement and train systems by simple DTW matching instead of phoneme level labeling, which is a time consuming manual work.

Holding to continuous speech processing is good to support potential language independency also.

The direct subjective test with hard of hearing persons will be done in the next phase of our research project.

DTW method can be tuned for our purposes with specially designed cumulative distance formulas, for long pauses in composite sentences in particular, which caused the errors of lowest scored videos in the subjective tests.

Acknowledgements This work was supported by the Mobile Innovation Center, Hungary. 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 tests and for their valuable advices, remarks.

References

[1] Takács, G., Tihanyi, A., Bárdi, T., Feldhoffer, G., Srancsik, B., "Speech to Facial Animation Conversion for Deaf Customers", Proceedings of EUSIPCO Florence Italy, 2006.

[2] Takács, G., Tihanyi, A., Bárdi, T., Feldhoffer, G., Srancsik, B., "Signal Conversion from Natural Audio Speech to Synthetic Visible Speech" Proceedings of International Conference on Signals and Electronic Systems, Lodz, Poland, Vol. 2. p.261 2006.

[3] Takács, G., Tihanyi, A., Bárdi, T., Feldhoffer, G., Srancsik, B., "Database Construction for Speech to Lip-readable Animation Conversion", Proceedings of ELMAR Zadar, Croatia p. 151, 2006.

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

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

[6] L. R. Rabiner, B-H. Juang, "Fundamentals of speech recognition", 1993.