19th INTERNATIONAL CONGRESS ON ACOUSTICS
MADRID, 2-7 SEPTEMBER 2007

SYNCHRONIZATION OF ACOUSTIC SPEECH DATA FOR MACHINE LEARNING BASED AUDIO TO VISUAL CONVERSION

PACS: 43.71.Ky

Takács, György; Tihanyi, Attila; Feldhoffer, Gergely; Bárdi, Tamás; Oroszi Balázs
Pázmány Péter Catholic University, Faculty of Information Technology, Práter u. 50/a Budapest 1083 Hungary {takacsgy, tihanyia, flugi, bardi, oroba}@itk.ppke.hu

ABSTRACT

The present paper studies the role of synchronisation in lip-readable speech to animation conversion for hearing impaired users. Formerly, we presented a neural network based speaker dependent audio to visual conversion system, trained on the data of a professional lip-speaker. Our current effort aims to decrease the speaker dependency of the system. Our proposed solution is to train the system on the acoustic data of numerous everyday speakers coupled with the visual data of professional lip-speakers. Their articulation is much more lip-readable. This method needs time synchronization between data files of lip-speakers and other speakers. This paper details the acoustic data synchronization procedure applied on our database. Dynamic Time Warping (DTW) was used on the acoustic feature vector sets to match them with the reference acoustic vectors and the same time warping function was applied to the visual feature vectors. The quality of the matching of audio-visual feature sets was evaluated by subjective tests. The simple DTW gave acceptable solution. Our iterative DTW gave even better matching. The listeners were not able to differentiate whether the original speech and video pairs or the warped video and speech from different speakers were presented.

INTRODUCTION

Our goal is to create an application which attends to convert voice to facial animation [1-3]. This would help deaf people to understand voice-only communication channels as telephone systems [4-6]. Recent research projects on conversion of speech audio signal to facial animation have concentrated on development of feature extraction methods, database construction and sys-tem training [7,8].

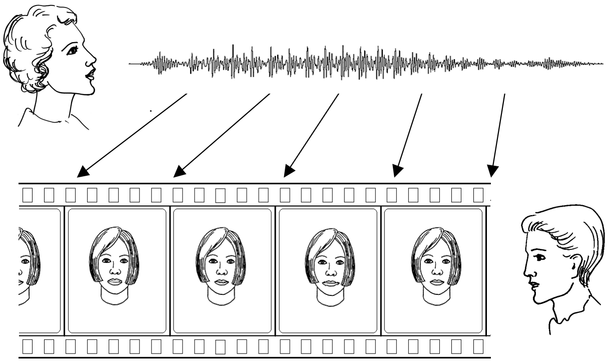

Fig. 1. The main goal of our system: allowing deaf users to understand the far end speaker in real time.

The task is similar to speech inversion, which tries to predict the state flow of the speech organs from the voice signal. Speech inversion research uses some measurement of speech organs as laringograph or MRI motion capture. Our task is easier, because we want to predict only the motion of the visible speech organs which are the source of the lip-reading. Data capturing of visible speech organs is easy, it can be done with commercially available video cameras.

The main system components are the acoustic feature extraction, the feature point coordinate vector calculation and running a standard MPEG4 face animation model [9] as it is shown in Figure 2.

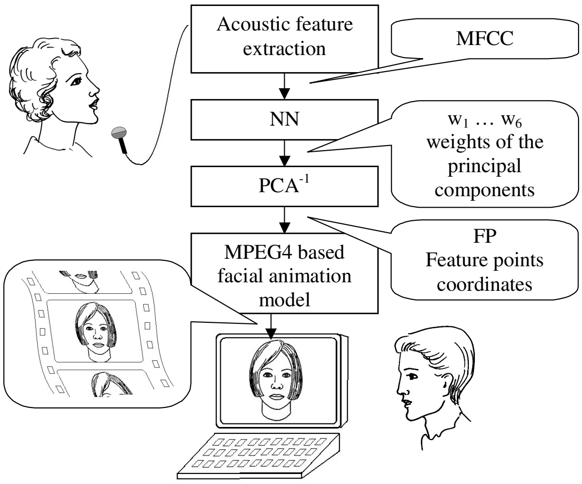
Acoustic feature extraction
MFCC
NN
$w_1 \ldots w_6$ weights of the principal components
PCA^{-1}
FP Feature points coordinates
MPEG4 based facial animation model

Figure 2.-System structure

The most important feature of the system is that all the modules are continuous, so no classification or database lookup happens during the conversion. This fundamental principle stands for database construction also.

DATABASE CONSTRUCTION

The traditional audio-visual databases are elaborated for testing by hearing people. Our database was constructed according to our special specifications. As a continuous database, the basic scheme of the data is a set of pairs of preprocessed audio and video data. Continuity means that the database content is calculated by continuous functions, and all of the segments are processed uniformly, so no labeling or any discrete information were used.

Content selection

The speakers and the text were specially selected for deaf customer needs. The speakers are professional lip-speakers, and the text was collected to represent each lip-reading situation evenly.

Video records

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.

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.

Audio records

Our database was recorded in small sized (2.5m x 4m) acoustically damped room to get low background noise and reverberation. The signal to noise ratio is better then -30 dB throughout the records. Speech audio was picked up with one Sennheiser dynamic microphone. The microphone signal was pre-amplified and sampled at 16 bit/48 kHz.

Audio and video signals were manually synchronized to eliminate the effects of sound velocity and different time latencies and clock-source inaccuracies of the digital camera and the soundcard. Bilabial occlusive consonants have easily identifiable effects on both video and audio signal. Therefore, the speakers were told to say *“pa-pa-pa-pa”* at the beginning and the end of the records. The initial frames of mouth openings just after the bilabial occlusions were sought in the videos and synchronized to the corresponding burst phase initials in the audio signals. Then we obtained the sample position of each frame by linear interpolation.

Processing of 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 [9] 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.

Processing of audio files

The digitized speech signal was pre-emphasis filtered. Then 1024 samples long (21.33 ms) Hamming windows were centered to the sample position of each video frame. 16 dimensional feature vectors were extracted from the analysis windows. The feature vectors are Mel-Frequency Coefficients (MFC) representing the log-energy of the window on 16 Mel-scaled triangular bands between 80 and 8000 Hz. The spectrum is computed using Fast Fourier Transform(FFT). The Euclidean distance of these feature vectors were used as metric for time warping. Generally, in speech signal processing, this computation is complemented with Discrete Cosine Transform (DCT). We avoided DCT because it would change the distance metric as it is not a unitary transform.

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.

The voice records were used for warping. Voice frames are characterized by MFC feature vectors. The distance metric in the DTW algorithm was the sum of MFC coefficient differences. The total frame number in records A and B might be different. MFC is better than MFCC with this metric since frequency bands are more evenly important than cepstrum quefrencies.

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.

Iterative DTW procedure

DTW can combine only three possible steps (based on omitting, doubling, and common step), and any alignment which can not be expressed with these steps are unable to come out. There are two possible solutions for this problem.

The first solution is to use the basic recursive steps as omitting any of the elements, or doubling any of the elements. This will not give satisfactory result, because the repeated omitting and doubling steps will avoid important information in the cumulating process. It may be possible to limit the number of repeating the same step.

The other solution is to use steps only which are including the common step. This will cumulate all of the important information during the calculation, but the possible results are limited to the valid combinations of these steps. If the optimal alignment is not one of them, the result is only a close estimation of it. To help this situation we are using an iterative approach.

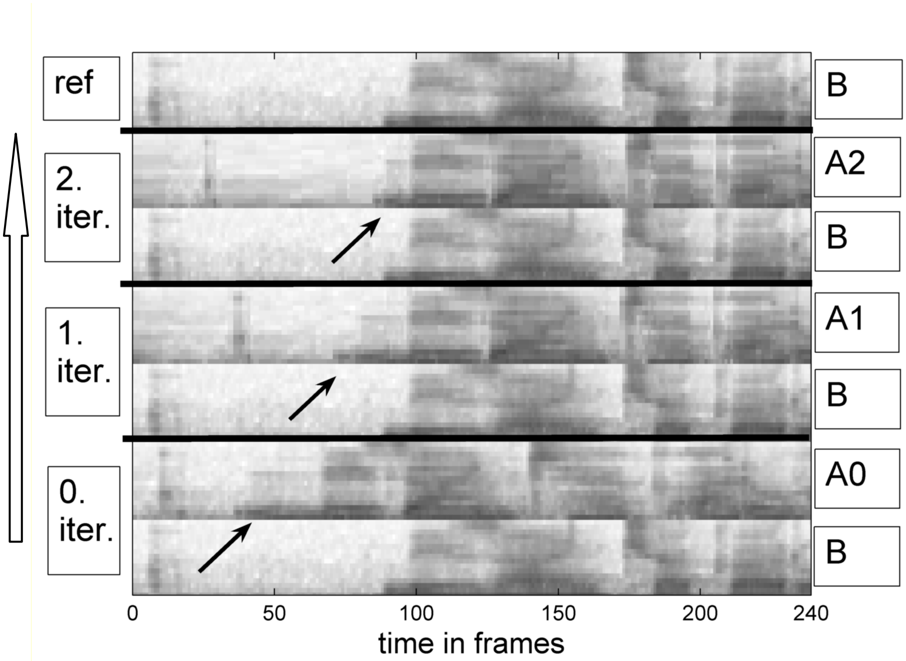

Figure 3.-Iterative DTW effect. On the iteration 0 the signals are not aligned. On the iteration 1 the alignment is acceptable on the most important sections (frame 100-200 and above), but still have problems signed by the arrow. On the second iteration this is eliminated.

The DTW is applied on the source information, which is the MFC vector set. The result of the DTW is an indexing vector. If this indexing vector is applied to the MFC vectors, it gives another feature vector set which is aligned to the original MFC matrix, but this alignment is not perfect, as it was described above, but after a second DTW procedure, a new indexing will make the first estimation more correct. These steps can be repeated, the process will be convergent to an indexing vector which equal to the identical index.

The main advantage of iterative DTW is the combination of the cumulating completeness and the free form of indexing. The disadvantage is the over-represented repeating and omitting since the optimal warping curve is frequently extreme on short sections, for example if a speaker have a pause where the other has not. Many repeats make the indexed video information look clogging, so post-filtering of the output is necessary.

On the Figure 3 an example is shown on different pause lengths. One step DTW can not handle this situation correctly.

EXPERIMENT AND RESULTS

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. Randomly 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.

21 test persons watched the 15 audio-video pairs in random order. The results are summarized in Figure 4 and 5.

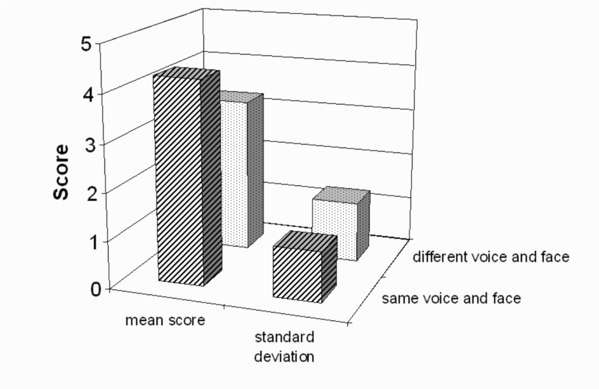

Figure 4.- Results of the subjective DTW tests. The opinion score values: 5- surely identical, 4-probably identical, 3-uncertain, 2-probably different, 1-surely different the origin of the voice and video

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. 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 5. The standard deviation values are overlapped within the two groups.

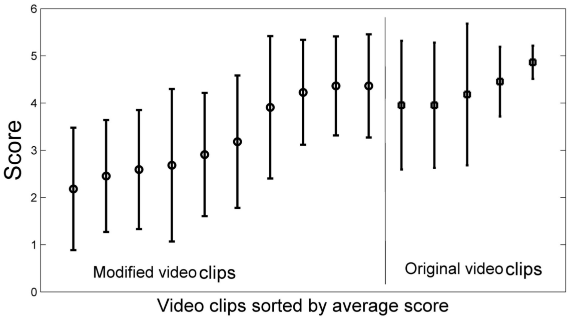

Figure 5: Ratings (average score and deviation) of video clips grouped by modified and original video data, sorted by average score.

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. 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.

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

[7] 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.

[8] P. Kakumanu,A. Esposito, O. N. Garcia, R. Gutierrez-Osuna, "A comparison of acoustic coding models for speech-driven facial animation", Speech Communication 48 pp 598-615, 2006

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