Acta Cybernetica **18** (2007) 355–362.

Conversion of continuous speech sound to articulation animation as an application of visual coarticulation modeling

Gergely Feldhoffer* and Tamás Bárdi*

Abstract

A voice to facial animation conversion system is presented in this paper. In particular the temporal structure of the multimodal speech is discussed. Mutual information and neural network training is used to estimate the optimal temporal scope for audio to video conversion.

Keywords: Voice to animation, visual coarticulation, mutual information

1 Introduction

The most common form of the language is the personal talk which is an audiovisual speech process. Our research is focused on the relation of the audio and the visual part of talking to build a system converting voice signal into face animation. A voice to animation conversion system (VACS) targets hearing impaired persons to help them understand voice only communication channels. A VACS get a speech signal as input, and produce a face animation which is understandable for persons who can lip-read. This task is similar to speech inversion which tends to extract information from speech signal about the state sequence of the speech organs. However, speech inversion aims to reproduce every speech organ to exactly the same state as the speaker used his organs, with every speaker dependent property. VACS is different, the target is to produce a lip-readable animation which depends only on the visible speech organs and not depends on the speaker dependent features of the speech signal.

Recent research activities are on speech signal processing methods specially for lip-readable face animation [7], face representation and controller methods[8], and more natural face animation systems [3]. In this paper a working system is presented focusing on the temporal structure of the audiovisual speech process.

VACS are not speech recognition systems, the target is to produce an animation without recognizing any of the language layers as phonemes or words, as this part

*Pázmány Péter Catholic University, Faculty of Information Technology, Budapest, Hungary, E-mail: {flugi, bardi}@itk.ppke.hu

of the process is left to the lip-reader. Because this, our VACS uses no phoneme recognition, furthermore there is no classification part in the process. This is the continuous VACS, avoiding any discrete type of data in the process. Discrete VACS are using visemes as the visual match of phonemes to describe a given state of the animation of a phoneme, and using interpolation between them to produce coarticulation.

Training a continuous VACS needs audio-video data pairs. Since plenty of speech audio databases exist but only a few audiovisual ones, building a continuous VACS means building a multimodal database first. A discrete VACS is a modular system, it is possible to use existing speech databases to train the voice recognition, and separately train the animation part on phoneme pairs or trigraphs[1]. So continuous VACS needs a special database, but the system will handle energy and rhythm naturally, meanwhile a discrete VACS has to reassemble the phonemes into a fluid coarticulation chain of viseme interpolations. Let we call the overall time of a coarticulation phenomena as temporal scope which means that the state of the mouth is depending on this time interval of the speech signal. In continuous VACS the calculation of a frame is based on this audio signal interval. In discrete VACS the visemes and the phonemes are synchronized and interpolation is applied between them, as it is popular in text to visual speech systems[5]. Figure 1 shows this difference.

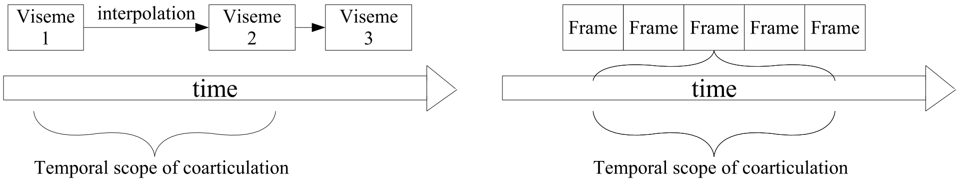

Figure 1: Temporal scope of discrete and continuous VACS

In this article we will describe a continuous VACS, and show how the system handles the visual coarticulation. We will show that using the average phoneme length as the length of temporal scope is a confirmable choice.

2 Database

The database for continuous VACS contains audio and video feature vector pairs. Basically it is a preprocessed multimedia material specially to use it as training set for neural networks. For this purpose the data should not contain strong redundancy for optimal learning speed.

2.1 Audio

The feature extraction starts with windowing. The length of a window depends on the frequency of the video camera which is 25 fps in this case, this means 40 ms long windows. For temporal scope estimations we used 1 ms window step, and 40 ms window step for system training. Preemphasis is used, and FFT after Hamming window. Radix-2 FFT was applied for CPU efficiency, so the first 2^n element of the window is processed. The spectrum is mel-scaled to 16 bands, and logarithm and DCT is applied. The result is the mel-scaled cepstrum, the MFCC.

2.2 Video

For video processing we used two methods. Both methods are based on video recording of a speaker and feature tracker applications. The first method is based on markers only which are placed around the mouth. The markers were selected as a subset of the MPEG-4 face description standard. Tracking the markers is a computer aided process; a 98% precise marker tracker algorithm was developed for this phase. The mistakes were corrected manually. The marker positions as a function of time were the raw data which was normalized by control points as the nose to eliminate the motion of the whole head. This gives a 30-36 dimensional space depending on marker count. This data is very redundant and high dimensional, it is not suitable for neural network training, so PCA was applied to reduce the dimensionality and eliminate the redundancy. PCA can be treated as lossy compression as only the first 6 parameters were used for training. Using only 6 coefficients can cause about 1 pixel error on PAL screen which is the precision of the marker tracking.

The second method uses only 2 markers but uses color information of the mouth to extract markers which can not be painted as the inner contour of the mouth. This technique is still under development. In this paper the results were measured on video data which was extracted by the first method.

2.3 Training

The synchrony of the audio and video data is checked by word ”papapa” in the beginning and the end of the recording. The first opening of the mouth by this bilabial can be synchronized with the burst in the audio data. This synchronization guaranties that the pairs of audio and video data were recorded in the same time. For the best result the neural network has to be trained on multiple windows of audio feature vectors where the window count have to be chosen based on the optimal temporal scope.

3 Temporal structure

To achieve the best results a good estimation of temporal scope is needed. Using a too short temporal scope can cause losing information about coarticulation. Using a too long temporal scope results longer training time without any quality improvement since the training will calculate with data which is independent from the actual state. In this section the method of mutual information estimation will be shown as a possible solution for the question.

3.1 Mutual information

The mutual information is (1):

$$MI_{X,Y} = \sum_{x \in X} \sum_{y \in Y} P(x,y) \log \frac{P(x,y)}{P(x)P(y)} \tag{1}$$

Mutual information is high if knowing X helps to find out what is Y, and it is low if X and Y are independent. To use this measurement for temporal scope the audio signal will be shifted in time compared to the video. If the time shifted signal has still high mutual information, it means that this time value should be in the temporal scope. If the time shift is too high, mutual information between the video and the time shifted audio will be low due to the relative independency of different phonemes.

Using a and v as audio and video frames:

$$\forall \Delta t \in [-1s, 1s] : MI(\Delta t) = \sum_{t=1}^{n} P(a_{t+\Delta t}, v_t) \log \frac{P(a_{t+\Delta t}, v_t)}{P(a_{t+\Delta t})P(v_t)} \tag{2}$$

where $P(x, y)$ is estimated by a 2 dimensional histogram convolved with Gauss window. Gauss window is needed to simulate the continuous space in the histogram in cases where only a few observations are there. Since audio and video data are multidimensional and MI works with one dimensional data, all the coefficient vectors were processed, and the results are summarized.

The channels were calculated by Independent Component Analysis (ICA) to keep down the interchannel dependency. The 16 MFCC channel was compressed into 6 independent component channels. The 6 PCA channels of video information was transformed into a ICA based basis. Interchannel independency is important because the measurement is the sum of all possible audio channel – video channel pairs, and we have to prove that each member of mutual information sum is not from the correlation of different video channels or different audio channels which would cause multiple count of the same information.

Since mutual information is a commutative, 6 x 6 estimations gives 15 different pairs.

Figure 2 shows the result of the estimation. Certain asymmetry can be observed in the sum of mutual information curves of all channel pairs of audio and video data. This measurement was done on a recording which aimed deaf people for lip-reading. This is a special situation; the speech speed is decreased to 5-6 phonemes per second. This gives an average phoneme length of 200 ms. As it can be seen on the figure, there is a high mutual information at 200 ms in the future of the voice, but a relatively low value in the past. This result shows that the visible speech organs are preparing for the next phoneme during the visual coarticulation

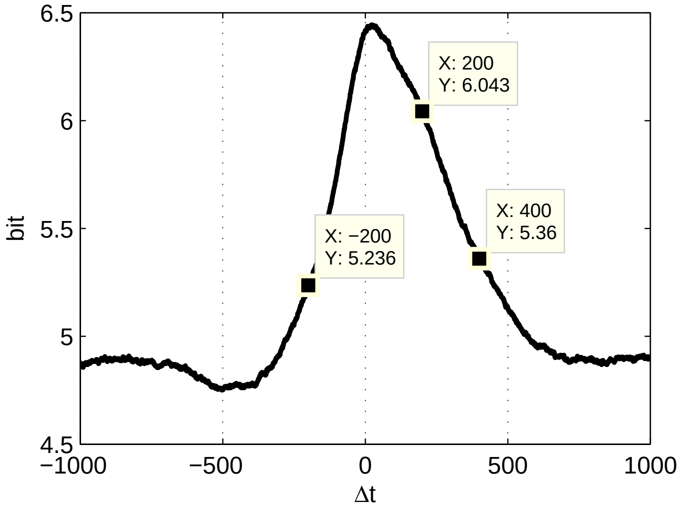

Figure 2: Sum of $MI(\Delta t)$ results of all channel audio-video pairs (6 x 6 : 15 pairs). Positive Δt means voice in the future was measured to the video frame in $\Delta t = 0$. The unit of time is millisecond.

while the speech audio signal is not changing. If both modalities would be changing together, there would be no asymmetry in mutual information.

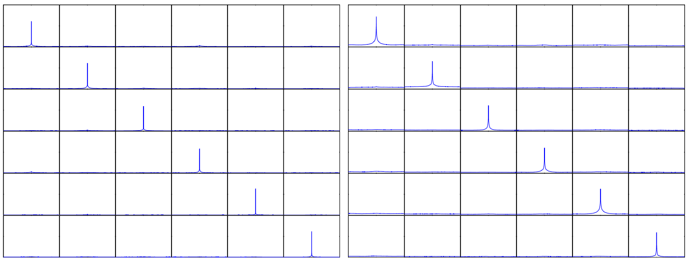

Figure 3: Interchannel $MI(\Delta t)$ results show independency in audio (left) and video (right) channels. The scaling of the curves are Δt = -1..1 in seconds on x axis, and 0..10 range in bits on y axis

On Figure 3 can be seen the independency of the channels. A channel with itself produces high mutual information in $\Delta t = 0$ because of equality. Short rising and decreasing phases can be observed in both modalities, much shorter than on Figure 2, however video data shows longer window of autocorrelation. This difference between audio and video data is partly because video information is from a 25fps recording which is 40ms of window length but the audio information was measured on every milliseconds, so video data was interpolated to fit to the audio data, and possibly partly because of the difference between invisible and visible speech organs, this question is a part of our future work.

3.2 Network training

In practical way the measurement of the temporal scope is to estimate it with training efficiency. Efficiency is measured in this case by training error after a given epoch number. The same data were trained with different window counts, and after 10.000 epochs the training error was compared. Training error means the average difference of the network's output and target values in the training set. Using the training error of single frame training as 100%, we found that training errors are nearly linearly decreasing to 50% at 200ms and stay around 50% (even higher due to the increased difficulty and fixed epoch count) if the scope is increased further. See Figure 4. This confirms in practice the mutual information measurement.

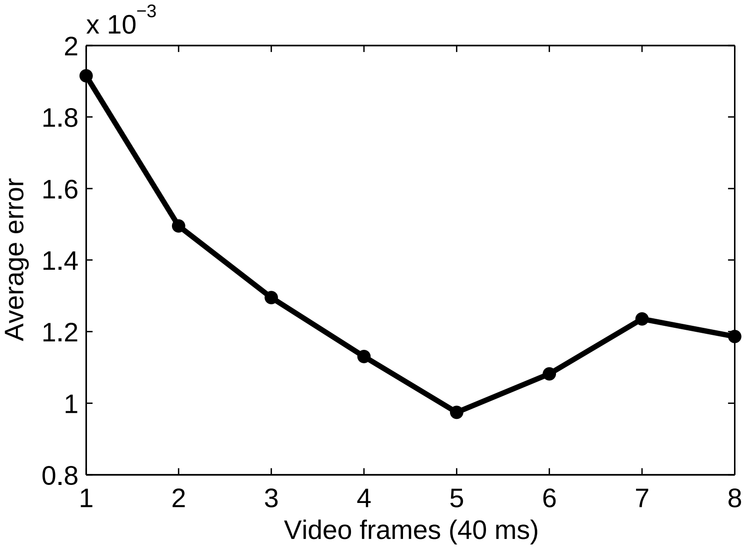

Figure 4: Training errors of different temporal scopes. The error is given in neural networks training data which is normalized to [-1..1] interval.

3.3 Visual coarticulation model

As mutual information estimation resulted that any given state of the video data stream can be calculated fairly on a defineable relative time window of the speech signal. This model predict that the transient phase of the visible speech can be calculated in the same way as in the steady phase as Figure 1 shows. This model gives a prediction about asymmetries in the multimodal speech process. This asymmetry can be explained with mental predictivity in the motion of the facial muscles to fluently form the next phoneme. This explanation needs more proof, and it is an important part of our future work.

4 Results

The described modules were implemented and trained. The system was measured with a recognition test with deaf people. To simulate a measurable communication situation, the test covered numbers, names of days of the week and months. As the measurement aimed to tell the difference between the VACS and a real person's video, the situation had to be in consideration of average lip-reading cases. As we found [4] deaf persons recline upon context more than hearing people. In the cases of numbers or names of months the context defines clearly the class of the word but leave the actual value uncertain. During the test the test persons had to recognize 70 words from video clips. One third of the clips were original video clip from the recording of the database, other one third were output of the VACS from audio signals and the remaining one third were synthesized video clips from the extracted video data. The difference between the recognition of real recording and the face animation from the extracted video data gives the recognition error from the face model and the database, as the difference between animations from video data and audio data gives the quality of the audio to video conversion. Table 1 shows the results.

Table 1: Recognition rates of different video clips.

Material	Recognition rate
original video	97%
face model on video data	55%
face model on audio data	48%

The results show that our VACS have satisfactory precision in audio to video conversion, but the face model has to be more fine. As it was mentioned, a new video feature extraction method is in progress.

The system uses 200 ms temporal scope. It was showed that this time interval is confirmable with both mutual information estimation and neural network training experiments.

A visual coarticulation model was introduced based on the results of mutual information estimation.

5 Acknowledgements

We are thankful for György Takács and Attila Tihanyi for the guidance and useful remarks. Thanks for the SINOSZ (National Association of the Deaf and Hard of

Hearing in Hungary) and all the hearing impaired people who lent a helping hand in testing and development advices.

References

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

[2] 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.

[3] E. Sifakis, A. Selle, A. Robinson-Mosher and R. Fedkiw *imulating Speech with a Physics-Based Facial Muscle Model*, ACM SIGGRAPH/Eurographics Symposium on Computer Animation (SCA), pp. 261-270, 2006.

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

[5] M. Cohen and D. Massaro *Modeling coarticulation in synthetic visual speech*, Computer Animation 93. Springer-Verlag, 1993.

[6] P. Kakumanu and R. Gutierrez-Osuna and A. Esposito and R. Bryll and A. Goshtasby and O. Garcia *Speech Driven Facial Animation*, Proc. of the Workshops on Perceptual/Perceptive User Interfaces (PUI), Orlando, FL, PUI 2001.

[7] 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.

[8] P. Scanlon, G. Potamianos, V. Libal, and S. M. Chu *Mutual Information Based Visual Feature Selection for Lipreading*, in Proc. of ICSLP 2004, South Korea, 2004.