TEMPORAL ASYMMETRY IN RELATIONS OF ACOUSTIC AND VISUAL FEATURES OF SPEECH

Gergely Feldhoffer, Tamás Bárdi, György Takács and Attila Tihanyi

Faculty of Information Technology, Péter Pázmány Catholic University
Práter u 50/a, H-1083, Budapest, Hungary
phone: + (36)-1-886-4763, fax: + (36)-1-886-4725, email: {flugi, bardi, takacsgy, tihanyia}@itk.ppke.hu
web: www.itk.ppke.hu

ABSTRACT

The fine temporal structure of relations of acoustic and visual features has been investigated to improve our speech to facial animation conversion system. Mutual information of acoustic and visual features has been calculated with different time shifts. Our result shows that the movement of feature points on the face of professional lip-speakers can precede the changes of acoustic parameters even by 100 milliseconds. Considering the measured time-shifts in synchrony in our system design the quality of our speech driven animations can be improved.

1. INTRODUCTION

Recent research projects on conversion of speech audio signal to facial animation have concentrated on development of feature extraction methods, database construction and system training [1, 2]. Evaluation and comparison of different systems have also had high importance in the literature. Our team demonstrated a complete real time working system at EUSIPCO'06 [3]. The results of subjective evaluation tests indicated lower performance than the target 80% word intelligibility but our system complexity and response time were considerably lower than any competitive system.

In this paper we focus on the temporal integration of acoustic features, optimal for converting the speech to animation in real-time. This study is based on our audiovisual database. The details of our system concept will be described in section 2. The input of the system is the audio speech signal and the output is of the animation control parameter set for MPEG-4 compliant facial animation rendering systems. The training database contains synchronized feature sets extracted from the audio and video signal from one speaker [4]. The critical part of such systems is the building of an optimal statistical model for the calculation of the video features from the audio features. The exact relation of the audio feature set and video feature set is unknown.

The speech signal conveys information elements in a very specific way. Some of speech sounds are related rather to steady states of the articulatory organs, others rather to transition movements [5, 6]. Our application is for providing a communication aid to deaf people. Professional lip-speakers have 5-6 phoneme/s speech rates to adapt the communication to the demand of deaf people so steady state phases and the transition phases of speech sounds are longer than in everyday speech style. The signal features to characterize a sound steady state phase or a transition phase or even to characterize a co-articulation phenomenon when the neighboring sounds are highly interrelated, need a careful selection of the temporal scope to characterize the speech and video signal. In our model we selected 5 analysis windows to describe the actual frame of speech plus two previous and two succeeded windows to cover +/- 80 ms interval. So such 5 element sequence of speech parameters can characterize transient sounds and the co-articulations.

We have recognized that at the beginning of words the lip movements start earlier then the sound production. Sometimes 100ms earlier the lips start to move to the initial position of the sounds. It was the task of the statistical model to handle this phenomenon.

In the refinement phase of our system we have tried to optimize the model selecting the optimal temporal scope of audio and video features. The measure of the fitting has based on the mutual information of audio and video features [7].

2. SYSTEM DESCRIPTION

In our former paper a real-time system was introduced that can convert the audio speech signal into video signal of the animated speaking face. Deaf users can understand the speech message based on the speaking face video. The main elements of the system are described below.

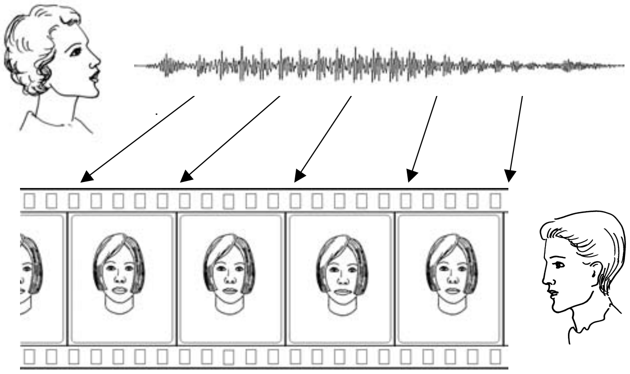

Figure 1 - Direct speech to facial animation conversion system.

The main feature of the system is the direct transformation from voice to animation without any language level operation (see Fig.1).

2.1 Audio processing

The voice signal is processed by 25 frame/s rate to be in synchrony with the processed video signal. One analysis window is 21.33 ms. The input speech is pre-emphasis filtered with $H(z)=1-0.983z^{-1}$. Hamming window and FFT with Radix-2 algorithm are applied. The FFT spectrum is converted to 16 mel-scale bands, and logarithm and DCT is applied. Such MFCC feature vectors are commonly used in general speech recognition tasks. The MFCC feature vectors provide the input to the neural networks.

2.2 Video processing

The video database is a set of video records of professional lip-speakers. Their moving faces are described by the 15 element subset of the standard MPEG-4 feature points (FP) set (84). These feature points were marked by colored dots on the face of the speakers. The coordinates of feature points were calculated by a marker tracking algorithm. The FP coordinates means 30 dimensional vectors which are compressed by PCA. We have realized that the first few PCA basis vectors have close relations to the basic movement components of lips. Such components can differentiate visemes. The marker coordinates are transformed into this basis, and we can use the transformation weights as data (FacePCA). The FacePCA vectors are the target output values of the neural net during the training phase [4].

2.3 Conversion by neural network

The audio processing unit extracts the audio MFCC feature vectors from the input speech signal. Five frames of MFCC vectors are sent to the trained neural net. The NN provide FacePCA weight vectors. These are converted into the control parameters of a MPEG-4 standard face animation model.

This system has an important trait. It does not attempt any classification or discrete database lookup. It has only continuous components, so the voice energy and rhythm flows through the system naturally, there is no need to a posteriori corrections as discrete systems do [8, 9, 10].

3. ESTIMATION OF TEMPORAL SCOPE AND SHIFT

The test of fitting of audio and video features was based on step-by-step temporal shifting of feature vectors. Indicator of the matching was mutual information. Small value of mutual information means that we have low average chance to estimate the facial parameters from the audio feature set, and the maximum location of the mutual information convey the most information about facial parameters.

An alternative method is to calculate cross correlation. We have also tested this method. It needs less computational power but some of relations are not indicated so it is an under estimation of theoretical maximum.

Mutual information between acoustic and visual features was estimated after time shifting the audio signal. The signal was shifted with Δt from -1000 ms to 1000 ms with 1 ms step. Then we get the mutual information as a function of the shifting time:

$$I_{X,Y}(\Delta t) = I(X(t), Y(t+\Delta t)) \qquad (1)$$

where mutual information is calculated according to Shannon's definition:

$$I(X,Y) = \sum_k \sum_n P(X=x_k, Y=y_n) \cdot \log_2 \frac{P(X=x_k, Y=y_n)}{P(X=x_k)P(Y=y_n)} \qquad (2)$$

Pairs of acoustic and visual features are quantized to discrete variables, and the joint probability distribution P(X,Y) is estimated by smoothing and normalizing the histograms of our statistical data. Marginal distributions (P(X),P(Y)) are the columnwise and rowwise sums of the estimated joint distribution. Histograms are 200x200 in size and smoothed with Gaussian window with 2.5 cells deviation.

The feature vectors (MPEG-4 FP coordinates and MFC) are transformed to orthogonal components using PCA (FacePCA and MFCPCA). Dependencies between variables is reduced this way, so that the mutual information of the vector valued acoustic and visual variables can be investigated by analyzing pairs of 1-dimensional variables. That is:

$$I(X,Y) \approx \sum_i \sum_j I(X_i, Y_j) \qquad (3)$$

4. MFCPCA AND FACEPCA MEASUREMENTS

170 seconds of audio and video speech records was processed. The speech frames are described with MFCPCA parameters. The MFCPCA parameters are more readable representation of frames for human experts than PCA of MFCC feature vectors, though they are interchangeable, so no DCT was applied.

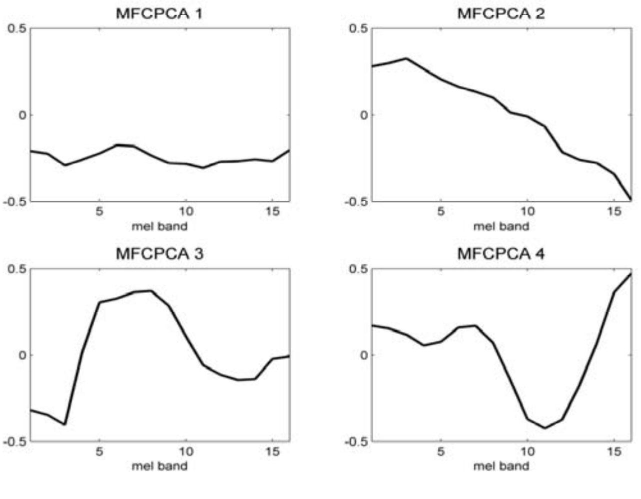

Figure 2 - Principal components of MFC feature vectors

The MFCPCA parameters have direct relations to the log-spectra. The PCA transformation does not consider the sign of the transformed vectors, so the first MFCPCA component shows energy-like representation as can be seen in Fig 2. For another example the second MFCPCA component has positive value in voiced speech frames and negative in frames of fricative speech elements.

The original video records have 40 ms frame rate so to have the possibility of 1 ms step size shifting, the intermediate shifted frame parameters have been calculated by interpolation and low pass filtering.

In the new coordinate system generated by the principal component analysis the coordinates can be characterized by an estimated importance rate. The importance rate can express that in the given direction which portion of the variance has been produced in the original space, so when any errors in the PCA space implies an error in the original space, this caused error can be estimated using this importance concept. The importance rate values in the case of MFCPCA transformation are shown in table 1.

MFCPCA	alone	first n together
1	77%	77%
2	10%	87%
3	5%	93%
4	2%	95%

Table 1 - Estimated importance rate (variance) of the MFCPCA

The importance rate values in the case of FacePCA transformation are shown in table 2.:

FacePCA	Alone	first n together
1	90%	90%
2	6%	96%
3	2%	98%
4	1%	99%

Table 2 – estimated importance rate (variance) of the MFCPCA

A possible estimation of the MFCPCA – FacePCA pairs is the multiplication of the estimated importances. This combines importance shows that which curve is important enough to extend the time scope for it. The important curves are the combinations of the 1-4 principal components. Their general importance is expressed by the darkness of the curves below.

Potential systematic errors have been carefully checked. The real synchrony of the audio-video records has been adjusted based on explosive sounds. The noise burst of explosives and the opening position of lips are real the characteristics. The check has been repeated at the end of the records also. The possible synchrony error is below one video frame (40ms).

5. RESULTS AND CONCLUSIONS

In out former work the earlier movements of the lips and the mouth have been observed in cases of co-articulation and at the beginning of words. The delay value has been estimated only by random sample investigations. Now we measured this co-articulation delay more generally.

Our new experiments produced a general rule with well defined delay values. Some of the strongest relation of audio and video features is not in the synchronous time frames. The mouth starts to form the articulation in some cases 100 ms earlier and the audio parameters follow it with such delay.

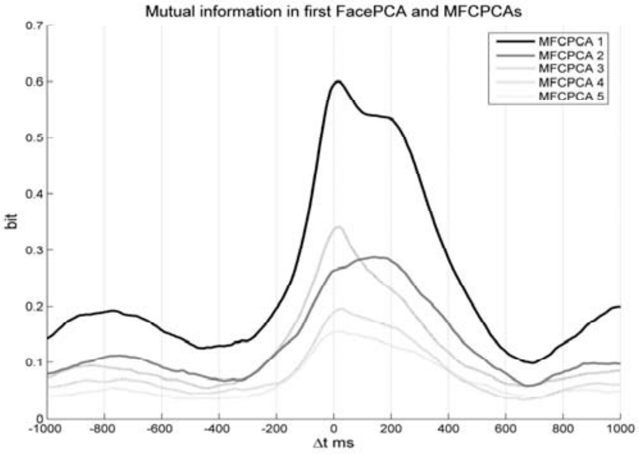

Figure 3 - Shifted FacePCA1 and MFCPCA mutual information. Positive Δt means voice in the future

The curves of mutual information values are asymmetric (Fig 3) and moved towards positive time shift. This means the acoustic speech signal is a better prediction basis to calculate the previous face and lip position than the future position. This fact is in harmony of the mentioned practical observation that articulation movement precedes the speech production at the beginning of words. The excitation signal comes later.

The results underline the general synchrony of audio and video database because the maximum of curves generally fit to Δt=0.

Interesting exception is the mutual information curve of FacePCA1 and MFCPCA2. Its maximum location is above zero Δt.

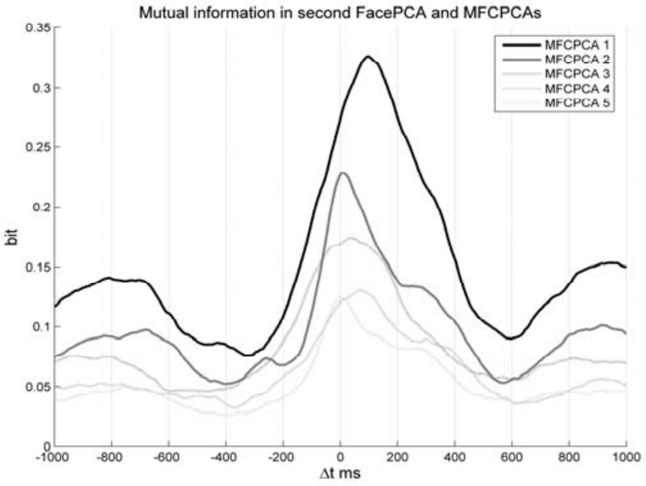

Figure 4 - Shifted FacePCA2 and MFCPCA mutual information. Positive Δt means voice in the future

On the Fig 4 the mutual information of FacePCA 2 and MFCPCA1 has maximum location at Δt=100ms with a very characteristic peek. This means that the best estimation of the FacePCA1 and FacePCA2 have to wait the audio parameters 100 ms later.

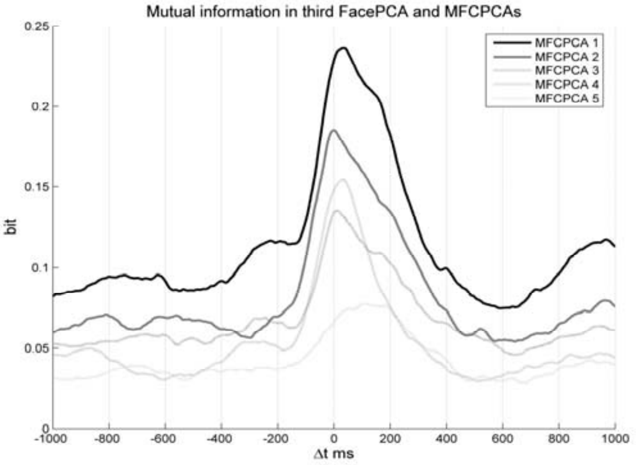

Figure 5 - shifted FacePCA3 and MFCPCA mutual information. Positive Δt means voice in the future

The FacePCA3 (Fig 5) curves have less importance because the weight of this parameter is considerable less in the facial animation process compared to the first two one. The asymmetry of curves is similar, if any peak results so there is no new message on the figure. Fig 6 shows the mutual information between the 6. PCAs for example.

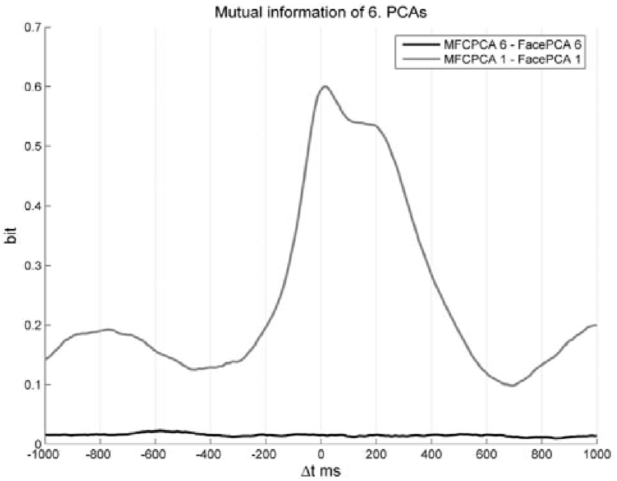

Figure 6 - N-th principal component does not contain substantive information

The curves show a general suitability since the mutual information values and the estimated importance values are in harmony. If there was a pair of FacePCA and MFCPCA which have high importance with a small value of maximum mutual information, it would show that our representation contains considerable elements which are useless in the voice to facial animation conversion.

5.1 A word size example beyond the average curves

The pronounced word was “September”. In figure 7 the bottom plot shows the audio waveform. The MFCPCA1 and MFCPCA2 parameter curves can be seen above. The changes on the MFCPCA curves are in exact synchrony with the waveform chart. The jump-up phases of FacePCA1 parameter curve start a little bit earlier then the transient phases on the waveform. But the waveform and the MFCPCA parameters remain in near steady state phase while the FacePCA parameters fall down towards the next phase. Such phenomenon with 100-200 ms time interval can produce the asymmetry and shoulder like shape on the curves of FacePCA and MFCPCA mutual information.

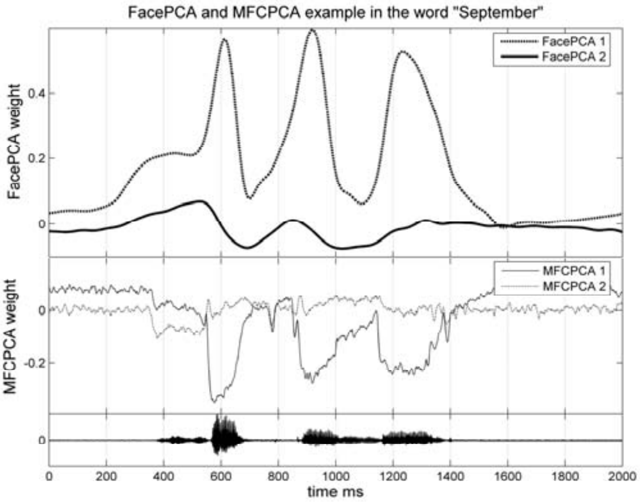

Figure 7 - The “september” word as the first two FacePCA, first two MFCPCA, and the signal.

Fig 7 shows clearly that the FacePCA2 parameter has regular changes during the steady state phases of audio features, usually shifted by 100 ms, as Fig 4 has predicted.

The example shows a possible reason of the shoulder of the MFPCPA1-FacePCA1 mutual information curve. At the “ep”, where the bilabial “p” follows the vowel, the spectral content does not change so fast as the FacePCA. This is because the tongue keeps the spectrum close to the original vowel, but the lips are closing already. This lasts until the mouth closes, where the MFC changes rapidly.

These results are valid in the case of a speech and video signal which is slow enough and lip-readable for deaf persons.

6. ACKNOWLEGDEMENTS

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.

7. REFERENCES

[1] 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.

[2] 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

[3] Gy. Takács, A. Tihanyi, T. Bárdi, G. Feldhoffer, B. Srancsik *"Speech to Facial Animation Conversion for Deaf Customers"* 14th European Signal Processing Conf., Florence, Italy, September 2006.

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

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

[6] Gy. Takács, A. Tihanyi, T. Bárdi, G. Feldhoffer, B. Srancsik *"Signal Conversion from Natural Audio Speech to Synthetic Visible Speech"* ICSES'06 Intl. Conf. on Signals and Electronic Systems, Lodz, Poland, September 2006.

[7] P. Scanlon, G. Potamianos, V. Libal, S. M. Chu *"Mutual Information Based Visual Feature Selection for Lipreading"*, Int. Conf. on Spoken Language Processing, South Korea 2004

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

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

[10] 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