Audio to visual speech conversion

Gergely Feldhoffer

A thesis submitted for the degree of
Doctor of Philosophy

Scientific adviser:
György Takács

Faculty of Information Technology
Interdisciplinary Technical Sciences Doctoral School
Pázmány Péter Catholic University

Budapest, 2010

Contents

Acknowledgements

I would like to thank the help of my supervisor György Takács, and my actual and former colleagues Attila Tihanyi, Tamás Bárdi, Tamás Harczos, Bálint Srancsik, and Balázs Oroszi. I am thankful for my doctoral school for providing tools and caring environment to my work, especially personally for Judit Nyéky-Gaizler and Tamás Roska.

I am also thankful for Iván Hegedűs, Gergely Jung, János Víg, Máté Tóth, Gábor Dániel "Szasza" Szabó, Balázs Bányai, László Mészáros, Szilvia Kovács, Solt Bucsi Szabó, Attila Krebsz and Márton Selmeci students, who participated in our research group.

My work would be less without the discussions with visual speech synthesis experts as László Czap, Takaashi Kuratate, Péter Mihajlik and Sasha Fagel.

I would like to thank also my fellow PhD students and friends, especially to Béla Weiss, Gergely Soós, Ádám Rák, Zoltán Fodrózci, Gaurav Gandhi, György Cserey, Róbert Wágner, Csaba Benedek, Barnabás Hegyi, Éva Bankó, Kristóf Iván, Gábor Pohl, Bálint Sass, Márton Miháltz, Ferenc Lombai, Norbert Bérci, Ákos Tar, József Veres, András Kiss, Dávid Tisza, Péter Vizi, Balázs Varga, László Füredi, Bence Bálint, László Laki, László Lővei and József Mihalicza for their valuable comments and discussions.

I thank the endless patience and helpfulness to Mrs Vida, Lívia Adorján, Mrs Haraszti, Mrs Körmendy, Gabriella Rumi, Mrs Tihanyi and Mrs Mikesy. I also thank the support of the technical staff of the university, especially Péter Tholt, Tamás Csillag and Tamás Rec.

And finally but not least I would like to thank the patient and loving support of my wife Bernadett and my family.

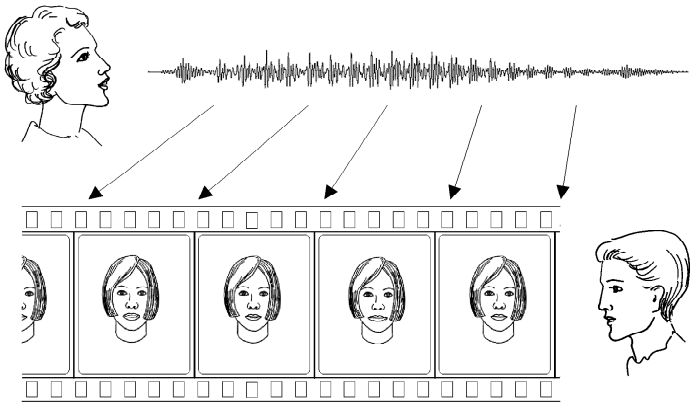

Figure 1: Task of audio to visual speech (ATVS) conversion.

Abstract

In this thesis, I propose new results in audio speech based visual speech synthesis, which can be used as help for hard of hearing people or in computer aided animation. I will describe a synthesis tool which is based on direct conversion between audio and video modalities. I will discuss the properties of this system, measuring the speech quality and give solutions for occurrent drawbacks. I will show that using adequate training strategy is critical for direct conversion. At the end I conclude that direct converison can be used as well as other popular audio to visual speech conversions, and it is currently ignored undeservedly because of the lack of efficient training.

1 Introduction

Audio to visual speech conversion is an increasingly popular applicable research field today. Main conferences such as Interspeech or Eurasip started new sections concerning multimodal speech processing, Interspeech 2008 held a special session only for audio to visual speech conversion.

1.1 Components

Each ATVS consist of the audio preprocessor, the AV (audio to video) mapping, and the face synthesizer.

Audio preprocessing use feature extraction methods to get useful and compact information from the speech signal. The most important aspects of quality here are the extracted representation dimensionality and covering error. For example the spectrum can be approximated by a few channels of mel bands replacing the speech spectrum with a certain error. In this case the dimensionality is reduced greatly by allowing certain noise in the represented data. Databases for neural networks have to consider

dimensionality as a primary aspect.

1.1.1 AV mapping

There are different strategies for performing this audio to visual conversion. One approach is to exploit automatic speech recognition (ASR) to extract phonetic information from the acoustic signal. This is then used in conjunction with a set of coarticulation rules to interpolate a visemic representation of the phonemes [1, 2]. Alternatively, a second approach is to extract features from the acoustic signal and convert directly from these features to visual speech [3, 4].

Recent research activities are on speech signal processing methods specially for lip-readable face animation [5], face representation and controller methods[6], and convincingly natural facial animation systems [7].

1.1.2 Quality issues

An AV mapping method can be evaluated on different aspects.

- naturalness: how much is the similarity of the result of the ATVS and a real persons visual speech
- intelligibility: how the result helps the lip-reader to understand the content of the speech
- complexity: the systems overall time and space complexity
- trainability: how easy is to enhance the system's other qualities by examples, is this process fast or slow, is it adaptable or fixed
- speaker dependency: how the system performance varies between different speakers
- language dependency: how complex is to port the system to a different language. The replacement of the database can be enough, or may rules have to be changed, even the possibility can be questionable.
- acoustical robustness: how the system performance varies in different acoustical environments, like higher noise.

2 Methods of investigation

I have built direct AV mapping systems which I evaluated with subjective and objective measurements such as subjective opinion scores (for naturalness), recognition tests (for intelligibility), neural network training and precision measurements.

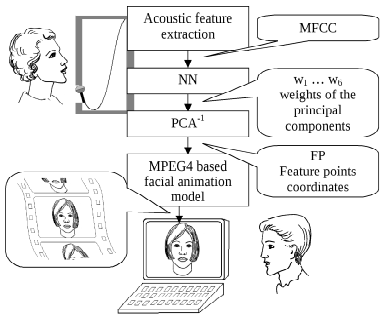

Figure 2: Workflow used in the base system.

2.1 Database building from audiovisual data

The direct conversion needs pairs of audio and video data, so the database should be a (maybe labeled) audiovisual speech recording where the visual information is enough to synthesize a head model. Therefore we recorded a face with markers on the subset of MPEG-4 FP positions, mostly around the mouth and jaw and also some reference points. Basically this 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, so the pre-processing includes the choice of an appropriate representation also.

2.1.1 Base system

The 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 ATVS and a real person's video, the situation had to be in consideration of average lip-reading cases. As we found [3] 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. In the tests we used a real lip-speaker's video as reference, and also measured the recorded visual speech. 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%

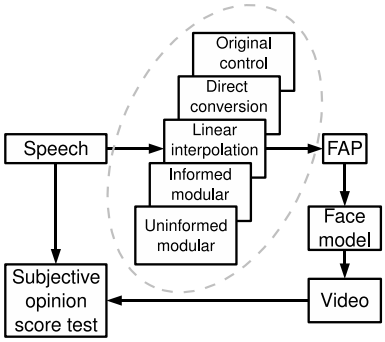

Figure 3: Multiple conversion methods were tested in the same environment. Informed and uninformed ASR based modular system was tested separately

3 New scientific results

3.1 Naturalness of direct conversion

A comparative study of audio-to-visual speech conversion is done. Our direct feature-based conversion system is compared to various indirect ASR-based solutions. The methods are tested in the same environment in terms of audio pre-processing and facial motion visualization. Subjective opinion scores show that with respect to naturalness, direct conversion performs well. Conversely, with respect to intelligibility, ASR-based systems perform better.

> ***I. I showed that our direct AV mapping method, which is more efficient computationally than modular approaches, overperforms the modular AV mapping in aspect of naturalness with a specific training set of professional lip-speaker.*** [8]

3.1.1 Approaches

One of the tested systems is our direct conversion. The rest of the approaches are using an ASR, a Weighted Finite State Transducer — Hidden Markov-Model (WFST-HMM) decoder. Specifically, a system known as VOXerver [9] is used, which can run in one of two modes: *informed*, which exploits knowledge of the vocabulary of the test data, and *uninformed*, which does not.

To account for coarticulation effects, a more sophisticated interpolation scheme is required. In particular the relative dominance of neighboring speech segments on the articulators is required. Speech segments can be classified as dominant, uncertain or mixed according to the level of influence exerted on the local neighborhood. We use the best Hungarian system by László Czap [10].

The videos were produced by the following: direct conversion was applied to a speaker's voice who is not included in the training database. ASR was used on the

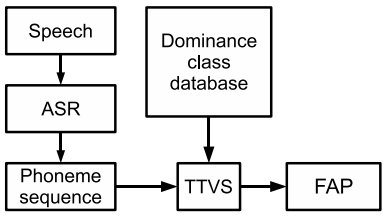

Figure 4: Modular ATVS consists of an ASR subsystem and a text to visual speech subsystem.

Table 2: Results of opinion scores, average and standard deviation.

Method	Average score	STD
Original facial motion	3.73	1.01
Direct conversion	3.58	0.97
UASR	3.43	1.08
Linear interpolation	2.73	1.12
IASR	2.67	1.29

same speech data. The resulting phoneme sequence was used for modular and linear interpolation methods. Modular method was applied ASR outputs where the ASR had a vocabulary of the test material (IASR) and another where no vocabulary was used (UASR). A recorded original visual speech parameter sequence was used as reference.

3.1.2 Results

58 test subjects were instructed to give opinion scores on naturalness of the test videos.

The results of the opinion score test is on Table 2. The difference between original speech and the direct conversion is not significant with $p = 0.06$ but UASR is significantly worse than original speech with $p = 0.00029$. The advantage of correct timing over correct phoneme string is also significant: UASR turned out more precise on timing that IASR which has errors on timing but has a phoneme precision of 100%.

Note that the linear interpolation system is exploiting better quality ASR results, but still performs significantly worse than the average of other ASR based approaches. This shows the importance of correctly handling viseme dominance and viseme neighborhood sensitivity in ASR based ATVS systems.

3.1.3 Conclusion

This is the first direct AV mapping system trained with data of professional lip-speaker. Comparison to modular methods is interesting because direct AV mappings trained on low quality articulation can be easily overperformed by modular systems in aspect of naturalness and intelligibility.

Opinion score averages and deviations shown no significant difference between hu-

man articulation and direct conversion, but significant difference between human and modular mapping based systems.

3.2 Temporal asymmetry

The fine temporal structure of relations of acoustic and visual features has been investigated to improve our speech to facial conversion system. Mutual information of acoustic and visual features has been calculated with different time shifts. The results has shown that the movement of feature points on the face of professional lip-speakers can precede even by 100ms the changes of acoustic parameters of speech signal. Considering this time variation the quality of speech to face animation conversion can be improved.

> ***II. I showed that the features of visible speech organs within an average duration of a phoneme are related closer to the following audio features than previous ones. The intensity of the relation is estimated with mutual information. Visual speech carries preceding information on audio modality.*** **[11]**

The earlier movement of the lips and the mouth have been observed in cases of coarticulation and at the beginning of words.

There are already published results about the temporal asymmetry of the *perception* of the modalities. Czap et al experienced difference in the tolerance of audio-video synchrony between the directions of the time shift: if audio precedes video, the listeners are more disturbed than in the reverse situation. My results show temporal asymmetry in the *production* side of the process, not the perception. This can be one of the reasons why perception is asymmetric in time (along some other things, like the difference between the speeds of sound and the light, which makes perceivers to get used to audio latency while listening to a person in distance)

3.2.1 Mutual information

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

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 independence 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 summarizing is validated by ICA. The mutual information values have been estimated from 200x200 size joint distribution histograms. The histograms have been smoothed by Gaussian window. The window has 10 cell radius with 2.5 cell deviation. The marginal density distribution functions have been calculated from the sum of joint distribution functions.

Audio and video signal are described by 1 ms fine step size synchronous frames. The signals can be shifted related to each other by fine steps. The audio and video representation of the speech signal can be interrelated from Δt= -1000ms to +1000ms. Such interrelation can be investigated only level that a single voice element how can estimate based on a shifted video element and vice versa as an average.

3.2.2 Multichannel Mutual Information estimation

In order to have a representation which is free of interchannel mutual information the data should be transformed by Independent Component Analysis (ICA) which looks for those multidimensional basis vectors which can make the distribution of the data to a uniformly filled hyper quadric shape. This way the joint distribution function of any two dimension will be uniformized.

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 independence 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.

3.2.3 Results

The mutual information curves were calculated and plotted for every important (in the space of first 6 principal component) ICA parameter pair in the range of -1000 to 1000 ms time shift. Also, some of the audio (MFCPCA) and video (FacePCA) mutual information estimation is shown.

The curves of mutual information values are asymmetric and moved towards positive time shift (delay in sound). 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 proceeds the speech production at the beginning of words. The results underline the general

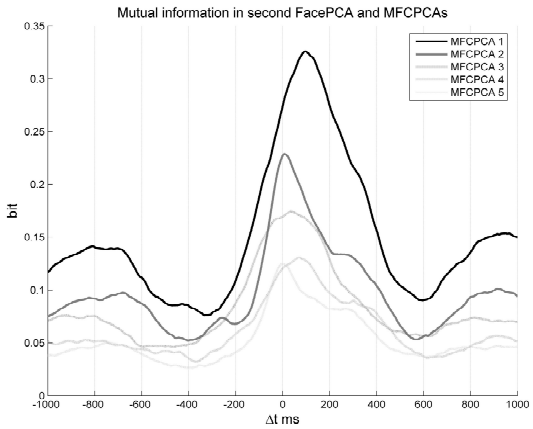

Figure 5: An example of shifted 2. FacePCA and MFCPCA mutual information. Positive Δt means future voice

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 0.

On the Fig 5 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.

Fig 6 shows clearly that the FacePCA2 parameter has regular changes during the the steady state phases of audio features so this parameter is related rather to the transients. 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 wovel, the spectral content does not change so fast as the FacePCA. This is because the tongue keeps the spectrum close to the original wovel, 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.

3.2.4 Conclusion

A multichannel mutual information estimation was introduced. I decreased the inter-channel mutual information of the same modality using ICA. To use only relevant, content distinctive data, the ICA was used on the first few PCA results. This way the traditional mutual information estimation method can be used on each pairs of the channels. The phenomena can not be reproduced in fast speech. There must be enough transient phase between phonemes. The effect is stronger in isolated word database, and weaker but still present in read database.

The main consequence of the phenomena is that the best possible ATVS system

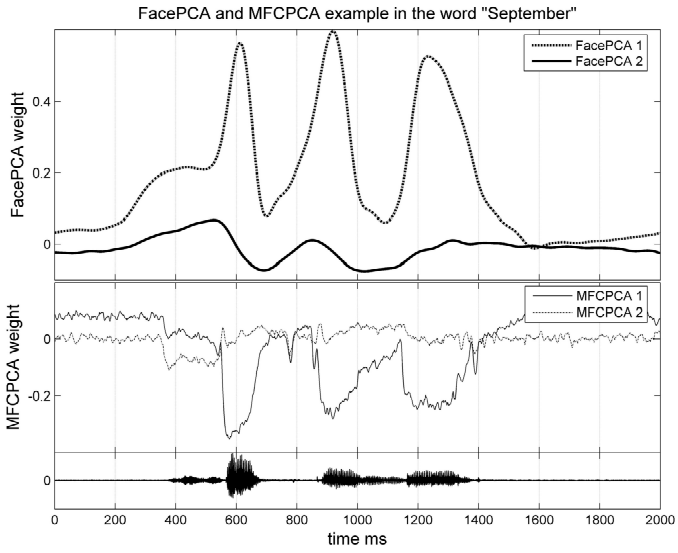

Figure 6: The word "September" as an example of time shifted visual components compared to audio components.

should have 200ms theoretical latency to wait up the future of the audio speech to synthesize the video data with the most extractable information. This phenomena can be useful also in multimodal speech recognition, using the video data to pre-filter the possibilities in the audio representation.

3.3 Speaker independence in direct conversion

The direct ATVS need an audiovisual database which contains audio and video data of speaking face.[12] The system will be trained on this data, so if there is only one person's voice and face in the database, the system will be speaker dependent. For speaker independence the database should contain more persons' voice, covering as many voice characteristics as possible. But our task is to calculate only one but lip-readable face. Training on multiple speaker's voices and faces results a changing face on different voices, and poor lip readability because of the lack of the talent of many people. We made a test with deaf persons, and the lip-readability of video clips is affected mostly by the training person's talent, and any of the video quality measures as picture size, resolution or frame/sec frequency affected less. Therefore we asked professional lip-speakers to appear in our database. For speaker independence the system needs more voice recording from different people. To synthesize one lip-readable face needs only one person's video data. So to create direct ATVS the main problem is to match the audio data of many persons with video data of one person.

III. I developed a time warping based AV synchronizing method to create training samples for direct AV mapping. I showed that

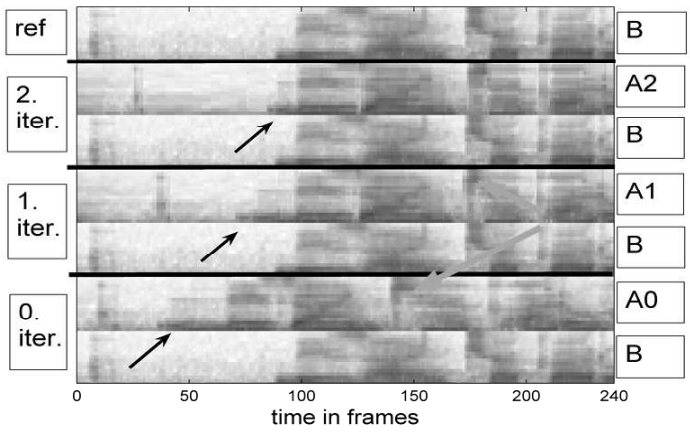

Figure 7: Iterations of alignment. Note that there are features which need more than one iteration of alignment.

> ***the precision of the trained direct AV mapping system increases with each added training sample set on test material which is not included in the training database.***[13]

Because of the use of multiple visual speech data from multiple sources would rise the problem of inconsistent articulation, we decided to enhance the database by adding audio content without video content, and trying to match recorded data if the desired visual speech state is the same for more audio samples. In other words, we create training samples as "How a professional lip-speaker would visually articulate this" for each audio time window.

I use a method based on Dynamic Time Warping (DTW) to align the audio modalities of different occurrences of the same sentence. DTW is originally used for ASR purposes on small vocabulary systems. This is an example of dynamic programming for speech audio.

Applying DTW for two audio signals will result in a suboptimal alignment sequence, how the signals should be warped in time to have the maximum coherence with each other. DTW has some parameters which restricts the possible steps in the time warping, for example it is forbidden in some systems to omit more than one sample in a row. These restrictions guarantee the avoidance of ill solutions, like "omit everything and then insert everything". In the other hand, the alignment will be suboptimal.

I have used iterative restrictive DTW application on the samples. In each turn the alignment was valid, and the process converged to an acceptable alignment. See Fig 7.

This above described matching is represented by index arrays which tell that speaker A in the i moment says the same as speaker B in the j moment. As long as the audio and video data of the speakers are synchronized, this gives the information of how speaker B holds his mouth when he says the same as speaker A speaks in the moment i. With this training data we can have only one person's video information which is

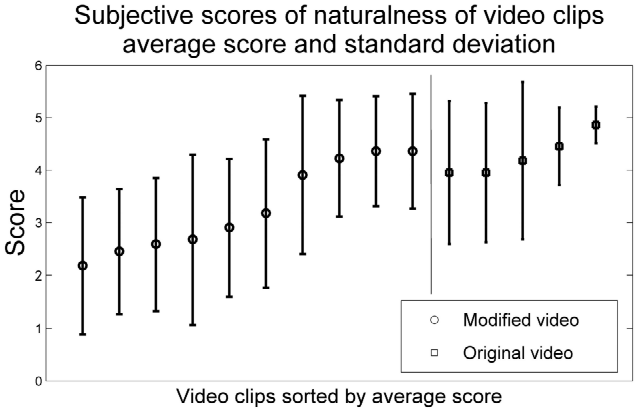

Figure 8: Mean value and standard deviation of scores of test videos.

from a professional lip-speaker and in the same time the voice characteristics can be covered with multiple speakers' voices.

3.3.1 Subjective validation

The DTW given indices were used to create test videos. For audio signals of speaker A, B and C we created video clips from the FP coordinates of speaker A. The videos of A-A cases were the original frames of the recording, and in the case of B and C the MPEG-4 FP coordinates of speaker A were mapped by DTW on the voice. Since the DTW mapped video clips contains frame doubling which feels erratic, all of the clips was smoothed with a window of the neighboring 1-1 frames. We asked 21 people to tell whether the clips are original recordings or dubbed. They had to give scores, 5 for the original, 1 for the dubbed, 3 in the case of uncertainty.

As it can be seen on Fig. 8. the deviations are overlapping each other, there are even better scored modified clips than some of the originals. The average score of original videos is 4.2, the modified is 3.2. We treat this as a good result since the average score of the modified videos are above the "'uncertain"' score.

3.3.2 Objective validation

A measurement of speaker independence is testing the system with data which is not in the training set of the neural network. The unit of the measurement error is in pixel. The reason of this is the video analysis, where the error of the contour detection is about 1 pixel. This is the upper limit of the practical precision. 40 sentences of 5 speakers were used for this experiment. We used the video information of speaker A as output for each speaker, so in the case of speaker B, C, D and E the video information is warped onto the voice. We used speaker E as test reference.

First, we tested the original voice and video combination, where the difference of

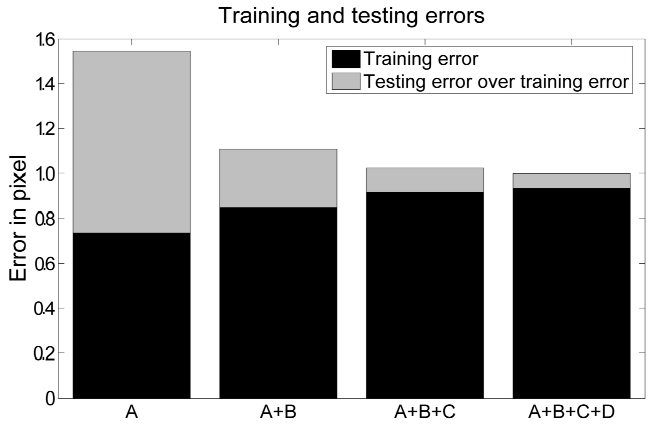

Figure 9: Training with speaker A, A and B, and so on, and always test by speaker E which is not involved in the training set.

the training was moderate, the average error was 1.5 pixels. When we involved more speakers's data in the training set, the testing error decreased to about 1 pixel, which is our precision limit in the database. See Fig. 9

3.3.3 Conclusion

A speaker independent ATVS is presented. Subjective and objective tests confirm the sufficient suitability of the DTW on training data preparing. It is possible to train the system with only voice to broaden the cover of voice characteristics. The speaker independence induces no plus expense on the client side.

Speaker independence in ATVS is usually handled as an ASR issue, since most of the ATVS systems are modular ATVS, and ASR systems are well prepared for speaker independence challenges. In this work a speaker independence enhancement was described which can be used in direct conversion.

Subjective and objective measurements were done. The system was driven by an unknown speaker, and the response was tested. In the objective test a neural network was trained on more and more data which were produced by the described method, and test error was measured with the unknown speaker. In the subjective test the training data itself was tested. Listeners were instructed to tell if the video is dubbed or original.

The method is vulnerable to pronouncation mistakes, the audio only speakers have to say everything just like the original lip-speaker, because if the dynamic programming algorithm lose the synchrony between the samples, serious errors will be included in the resulting training database.

This is a method which greatly enhance a quality without any run-time penalties. Direct ATVS systems should use the method always.

3.4 Visual speech in audio transmitting telepresence applications

Supporting an ATVS system with head models requires information on the representation of the ATVS system. In the base system we used PCA parameters. Face rendering parameters are based on measurements. If the system has to use head models by graphical designers, the main component states of the head should be clearly formulated. Graphical designers can not draw principal components since these abstractions can not be examined in themselves in nature. Designers can draw viseme states, so there is motivation in investigation of ATVS systems mapping to this representation.

> ***IV. I developed and measured a method to enhance audio transmitting telepresence applications to support visual speech with low time complexity and with the ability to handle viseme based head models. The resulting system overperforms the baseline of the widely used energy based interpolation of two visemes.*** **[14]**

For audio preprocessing we used MFCC. In some applications there are other audio preprocessing included, in the case of audio transmission mostly Speex.

3.4.1 Viseme based decomposition

For a designer artist it is easier to build multiple model shapes in different phases than building one model with the capability of parameter dependent motion by implementing rules in the 3D framework's script language. The multiple shapes should be the clear states of typical mouth phases, usually the visemes, since these phases are easy to capture by example. A designer would hardly create a model which is in a theoretical state given by factorization methods.

Every facial state is expressed as a weighted sum of the selected viseme state sets. The decomposition algorithm is a simple optimization of the weight vectors of viseme elements resulting minimal errors. The visemes are given in pixel space. Every frame of the video is processed independently in the optimization. We used partial gradient method with a constraint of convexness to optimize the weights where the gradient was based on the distance of the original and the weighted viseme sum. The constraint is a sufficient but not necessary condition to avoid unnatural results as too big mouth or head, therefore no negative weights allowed, and the sum of the weights is one. In this case a step in the partial gradient direction means a larger change in the direction and a small change in the remaining directions to balance the sum. The approximation is accelerated and smoothed by choosing the starting weight vector from the last result.

$$\vec{G} = \sum_{i=1}^{N} w_i \vec{V_i} \tag{3}$$

where

$$\sum_{i=1}^{N} w_i = 1 \tag{4}$$

Figure 10: Visemes are the basic unit of visual speech. These are those visemes we used for subjective opinion score tests in this (row-major) order.

The state G can be expressed as convex sum of viseme states V, which can be any linear representation, as pixel coordinates or 3D vertex coordinates.

The head model used in subjective tests is three dimensional and this calculation is based on two dimensional similarities, so the phase decomposition is based on the assumption that two dimensional (frontal view) similarity induce three dimensional similarity. This assumption is numerically reasonable with projection.

The base system was modified to use Speex as audio representation, and the results of decomposition as video representation. The neural network was trained and Speex interface was connected to the system.

3.4.2 Results

The details of the trained system response can be seen on Fig 11. The main motion flow is reproduced, and there are small glitches at bilabial nasals (lips not close fully) and plosives (visible burst frame). Most of these glitches could be avoided using longer buffer, but it cause delay in the response.

Subjective opinion score test was done to evaluate the voice based facial animation with short videos. Half of the test material was face picture controlled by decomposed data and the other half by facial animation control parameters given by the neural network based control data from original speech sounds.

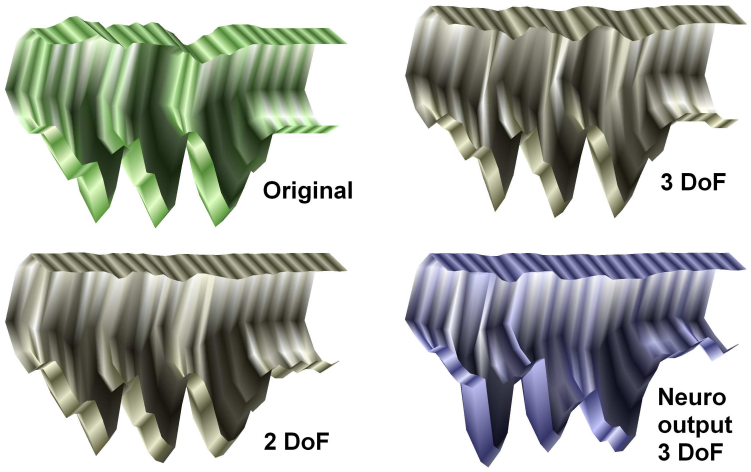

Figure 11: Examples with the hungarian word "Szeptember", it's very close to English "September" except the last e is also open. Each figure is the mouth contour in the time. The original data is from a video frame sequence. The 2 and 3 DoF are the result of decomposition. The last picture is the voice driven synthesis.

The results of the opinion score test show that the best score/DoF rate is at the 2 DoF (Fig 12), in fact the highest numerical error. These results show that the neural network may train to details which are not very important to the test subjects.

3.4.3 Conclusion

The main challenge was the strange representation of the visual speech. We can say our system was successfully used this representation.

The presented method is efficient as the CPU cost is low, there is no network traffic overhead, the feature extraction of the voice is already performed by voice compression, and the space complexity is scalable for the application. The feature is independent from the other clients, can be turned on without explicit support from the server or other clients.

The quality of the mouth motion was measured by subjective evaluation, the proposed voice driven facial motion shows sufficient quality for on-line games, significantly better than the one dimensional jaw motion.

Let us note that the system does not contain any language dependent component, the only step in the workflow which is connected to the language is the content of the database.

Facial parameters usually represented with PCA. This new representation is aware of the demands of the graphical designers. There were no publications before on the usability if this representation concerning ATVS database building or real-time synthesis. Subjective opinion scores were used to measure the resulting quality.

Using the Speex and the viseme combination representation the resulting system is

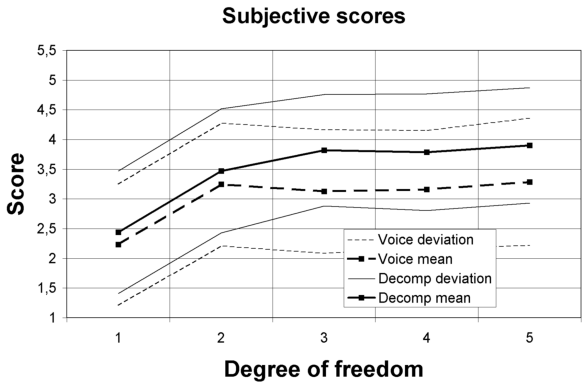

Figure 12: Subjective scores of the decomposition motion control and the output of the neural network. There is a significant improvement by introducing a second degree of freedom. The method's judgment follows the database's according to the complexity of the given degree of freedom.

embeddable very easily.

List of Publications

International transactions

- Gergely *Feldhoffer*, Tamás Bárdi : Conversion of continuous speech sound to articulation animation as an application of visual coarticulation modeling, Acta Cybernetica, 2007

- Gergely *Feldhoffer*, Attila Tihanyi, Balázs Oroszi : A comparative study of direct and ASR based modular audio to visual speech systems, Phonetician 2010 (accepted)

International conferences

- Gyorgy Takacs, Attila Tihanyi, Tamas Bardi, Gergely *Feldhoffer*, Balint Srancsik: Database Construction for Speech to Lip-readable Animation Conversion, Proceedings 48th International Symposium ELMAR, Zadar, 2006

- G. Takács, A. Tihanyi, T. Bárdi, G. *Feldhoffer*, B. Srancsik: Signal Conversion from Natural Audio Speech to Synthetic Visible Speech, Int. Conf. on Signals and Electronic Systems, Lodz, Poland, September 2006

- 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.

- Takács György, Tihanyi Attila, Bárdi Tamás, *Feldhoffer* Gergely,: Feasibility of Face Animation on Mobile Phones for Deaf Users, Proceedings of the 16st IST Mobile and Wireless Communication Summit, Budapest 2007

- Gergely *Feldhoffer*, Balázs Oroszi, György Takács, Attila Tihanyi, Tamás Bárdi: Inter-speaker Synchronization in Audiovisual Database for Lip-readable Speech to Animation Conversion, 10th International Conference on Text, Speech and Dialogue, Plzen 2007

- Gergely *Feldhoffer*, Tamás Bárdi, György Takács and Attila Tihanyi: Temporal Asymmetry in Relations of Acoustic and Visual Features of Speech, 15th European Signal Processing Conf., Poznan, Poland, September 2007

- Takács, György; Tihanyi, Attila; *Feldhoffer*, Gergely; Bárdi, Tamás; Oroszi Balázs: Synchronization of acoustic speech data for machine learning based audio to visual conversion , 19th International Congress on Acoustics, Madrid, 2-7 september 2007

- Gergely *Feldhoffer*: Speaker Independent Continuous Voice to Facial Animation on Mobile Platforms, PROCEEDINGS 49th International Symposium ELMAR, Zadar, 2007.

Hungarian publications

- Bárdi T., *Feldhoffer* G., Harczos T., Srancsik B., Szabó G. D: Audiovizuális beszéd-adatbázis és alkalmazásai, Híradástechnika 2005/10

- *Feldhoffer* G., Bárdi T., Jung G., Hegedûs I. M.: Mobiltelefon alkalmazások siket felhasználóknak, Híradástechnika 2005/10.

- Takács György, Tihanyi Attila, Bárdi Tamás, *Feldhoffer* Gergely, Srancsik Bálint: Beszédjel átalakítása mozgó száj képévé siketek kommunikációjának segítésére, Híradástechnika 3. 2006

- Takács György, Tihanyi Attila, Bárdi Tamás, *Feldhoffer* Gergely, Srancsik Bálint: MPEG-4 modell alkalmazása szájmozgás megjelenítésére, Híradástechnika 8. 2006

- *Feldhoffer* Gergely, Bárdi Tamás: Látható beszéd: beszédhang alapú fejmodell animáció siketeknek, IV. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, 2006.

Bibliography

[1] J. Kewley J. Beskow, I. Karlsson and G. Salvi. Synface - a talking head telephone for the hearing-impaired. *Computers Helping People with Special Needs*, pages 1178–1186, 2004. 1.1.1

[2] M. De Smet S. Al Moubayed and H. Van Hamme. Lip synchronization: from phone lattice to PCA eigen-projections using neural networks. In *Proceedings of Interspeech 2008*, Brisbane, Australia, Sep 2008. 1.1.1

[3] T. Bárdi G. Feldhoffer Gy. Takács, A. Tihanyi and B. Srancsik. Speech to facial animation conversion for deaf customers. In *4th European Signal Processing Conf.*, Florence, Italy, 2006. 1.1.1, 2.1.1

[4] J. Yamagishi G. Hofer and H. Shimodaira. Speech-driven lip motion generation with a trajectory HMM. In *Proc. Interspeech 2008*, pages 2314–2317, Brisbane, Australia, 2008. 1.1.1

[5] O. N. Garcia R. Gutierrez-Osuna P. Kakumanu, A. Esposito. A comparison of acoustic coding models for speech-driven facial animation. *Speech Communication*, 48:598–615, 2006. 1.1.1

[6] V. Libal P. Scanlon, G. Potamianos and S. M. Chu. Mutual information based visual feature selection for lipreading. In *in Proc. of ICSLP*, 2004. 1.1.1

[7] A. Robinson-Mosher E. Sifakis, A. Selle and R. Fedkiw. Simulating speech with a physics-based facial muscle model. *ACM SIGGRAPH/Eurographics Symposium on Computer Animation (SCA)*, pages 261–270, 2006. 1.1.1

[8] A. Tihanyi G. Feldhoffer and B. Oroszi. A comparative study of direct and asr based modular audio to visual speech systems (accepted). *Phonetician*, 2010. 3.1

[9] B. Németh P. Mihajlik, T. Fegyó and V. Trón. Towards automatic transcription of large spoken archives in agglutinating languages: Hungarian ASR for the MALACH project. In *Speech and Dialogue: 10th International Conference*, Pilsen, Czech Republic, 2007. 3.1.1

[10] L. Czap and J. Mátyás. Virtual speaker. *Híradástechnika Selected Papers*, Vol LX/6:2–5, 2005. 3.1.1

[11] Gy. Takács G. Feldhoffer, T. Bárdi and T. Tihanyi. Temporal asymmetry in relations of acoustic and visual features of speech. In *15th European Signal Processing Conf.*, Poznan, Poland, 2007. 3.2

[12] T. Bárdi-G. Feldhoffer B. Srancsik G. Takács, A. Tihanyi. Database construction for speech to lipreadable animation conversion. In *ELMAR Zadar*, pages 151–154, 2006. 3.3

[13] G. Feldhoffer. Speaker independent continuous voice to facial animation on mobile platforms. In *49th International Symposium ELMAR*, Zadar, Croatia, 2007. 3.3

[14] G. Feldhoffer and B. Oroszi. An efficient voice driven face animation method for cyber telepresence applications. In *2nd International Symposium on Applied Sciences in Biomedical and Communication Technologies*, Bratislava, Slovak Republic, 2009. 3.4