DOI:10.15774/PPKE.ITK.2010.007

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

DOI:10.15774/PPKE.ITK.2010.007

DOI:10.15774/PPKE.ITK.2010.007

Contents

DOI:10.15774/PPKE.ITK.2010.007

DOI:10.15774/PPKE.ITK.2010.007

DOI:10.15774/PPKE.ITK.2010.007

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 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.

DOI:10.15774/PPKE.ITK.2010.007

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.

DOI:10.15774/PPKE.ITK.2010.007

DOI:10.15774/PPKE.ITK.2010.007

Chapter 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.

Possible applications of the field are communication aiding tools for deaf and hard of hearing people[1] by taking advantage of the sophisticated lip-reading capabilities of these people, or lip-sync applications in the animation industry, in computer aided animations as well as in real-time telepresence based video games. In this thesis I will describe solutions for both of these applications.

In this chapter I will show the actual status of the topic, motivations and state of the art techniques. To understand this chapter basic speech and signal processing knowledge is needed.

1.1 Definitions

Speech is a multimodal process. The modalities can be classified as audio speech and visual speech. I will use the following terms:

Visual speech is a representation of the view of a face talking.

Visual speech data is the motion information of visible speech organs in any representation.

Phoneme is the basic meaning distinctive segmental unit of the audio speech. It is language dependent.

Viseme is the basic meaning distinctive segmental unit of the visual speech. Also language dependent. There are visemes belonging to phonemes, and there are phonemes which do not have viseme in particular, because the phoneme can be pronounced with more than one ways of articulation.

Automatic speech recognition (ASR) is a system or a method which can extract phonetic information from audio speech signal. Usually a phoneme string is produced.

Audio to visual speech (ATVS) conversion systems are to create an animation of a face according to a given audio speech.

DOI:10.15774/PPKE.ITK.2010.007

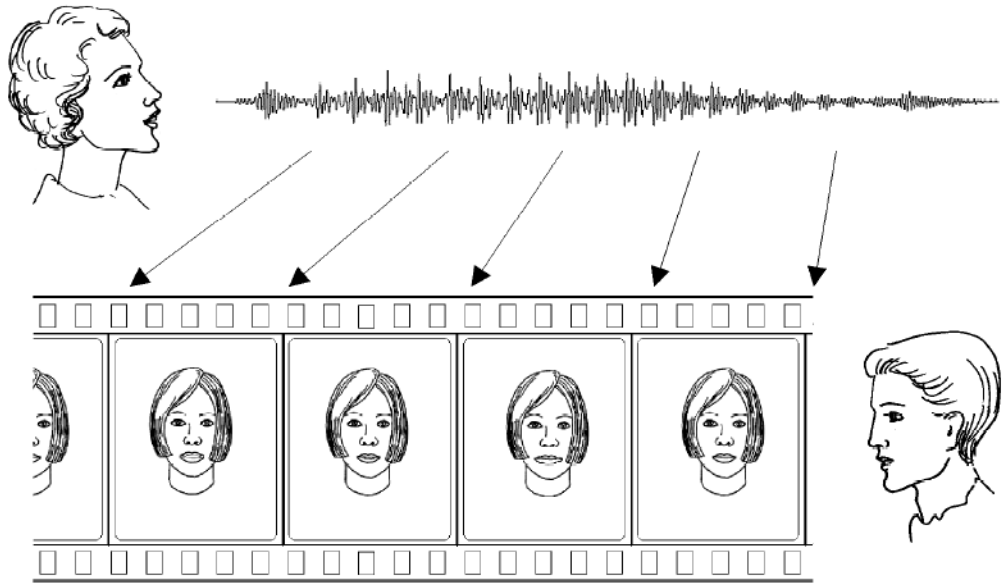

Figure 1.1: Task of audio to visual speech (ATVS) conversion.

Direct ATVS is an ATVS which maps audio representation to video representation by approximation.

Discrete ATVS is an ATVS which uses classification into discrete categories in order to connect the modalities. Usually phonemes and visemes are used.

Modular ATVS is an ATVS which contains ASR subsystem, and phoneme-viseme mapping subsystem. Modular ATVS systems are usually discrete.

AV mapping is an input-output method where the input is audio data in any representation and the output is visual data in any representation. In a discrete ATVS, AV mapping is a phoneme-viseme mapping, in a direct ATVS this is an approximator.

1.1.1 Components

Each ATVS consist of the audio preprocessor, the AV (audio to video) mapping, and the face synthesizer. The most straightforward method is the jaw-opening driven by speech energy, this system is widely used in on-line games, so the audio preprocessor is a frame-by-frame energy calculation expressed in dB, the AV mapping is a linear function, which maps the one dimensional audio data to the one dimensional video parameter, the jaw opening. The face model is usually a vertexarray of the face, and by modifying the vertices of the jaw the face synthesis is done. In below more sophisticated cases will be detailed where naturalness and intelligibility are issues.

Recent research activities are on speech signal processing methods specially for lip-readable face animation [2], face representation and controller method[3], and convincingly natural facial animation systems [4].

DOI:10.15774/PPKE.ITK.2010.007

Audio preprocessing

These systems 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.

Audio preprocessing methods can be clustered in many aspects as time domain or frequency domain feature extractors, approximation or classification, etc. A deeper analysis of audio preprocessing methods concerning audiovisual speech is published by [5] resulting the main approaches are approximately equally well. These traditional approaches are the Mel Frequency Cepstral Coefficients (MFCC) and Linear Prediction Coding (LPC) based methods. A quite convenient property of LPC based vocal tract estimation is the direct connection to the speech organs via the pipe excitation model. It seems to be a good idea to use vocal tract for ATVS as well but according to [5] it has not significantly more usable data.

AV mapping

In this step the modalities are connected, visual speech data is produced from audio data.

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 [6, 7]. Alternatively, a second approach is to extract features from the acoustic signal and convert directly from these features to visual speech [8, 9].

Face synthesis

In this step the visual speech representation is applied to a face model. Usually separated to two independent parts as facial animation representation and model rendering. The face representation maps the quantitative data on face descriptors. An example of this is the MPEG-4 standard. Face rendering is a method to produce picture or animation from face descriptions. These are usually computer graphics related techniques.

1.1.2 Quality issues

An ATVS can be evaluated on different aspects.

- naturalness: how much is the similarity of the result of the ATVS and a real persons visual speech

DOI:10.15774/PPKE.ITK.2010.007

- 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

- context dependency: how the system performance varies between speech contents (eg. a system, which trained on medical content, may perform poorer on financial content)

- 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 acoustic environments, like higher noise.

1.2 Applications

In this section I describe some of the recent systems, and give a short description of them in the quality aspects detailed above.

1.2.1 Synface

An example of ATVS systems is the Synface[1] of KTH, Sweden. This system is designed for hearing impaired but not deaf people to handle voice calls on phone. The system connects the phone line to a computer, where a speech recognition software translates the incoming speech signal to a time aligned phoneme sequence. This phoneme sequence is the basis of the animation control. Each phoneme is assigned to a viseme, and the recognized sequence makes a string of visemes to animate. The speech recognition subsystem not just recognizes the phoneme but makes the segmentation also. The viseme sequence timed by this segmentation information gives the final result of the AV mapping, using a rule-based strategy. The rule set is created by examples of Swedish multimodal speech.

This system is definitely language dependent, it uses the Swedish phoneme set, a Swedish ASR, and a rule set built on Swedish examples. On the other hand the system performs very well in aspects of intelligibility, acoustical robustness, speaker and context dependency.

DOI:10.15774/PPKE.ITK.2010.007

1.2.2 Synthesis of nonverbal components of visual speech

An example of audio to visual non-verbal speech estimation is the system of Gregor Hofer and Hiroshi Shimodaira[9]. Their system targets to extract the correct time of blink in speech. The audio preprocessing in this system concentrates on non-verbal components, such as rhythm, and intonation. Compared to actual videos, the original audio signal was used to test the precision of the estimation, which was above 80% with a decent time toleration of 100 ms. It is important that there are two kinds of blink, one of the regular eye care, fast blink, and the other is the non-verbal visual speech component emphasized blink. Of course this work was focused the second variant.

1.2.3 Expressive visual speech

This field changed the name from "Emotional speech" to "Expressive speech" because of psychological reasons. Expressive speech targets to synthesize or recognize emotional expressions in speech. Expressing emotions is very relevant in visual speech.

I show two approaches to the field. Pietro Cosi el al work on the virtual head "Lucia" [10] to connect an expressive audio speech synthesizer with a visual speech synthesizer. This text based system can be used as an audiovisual agent on any interactive media where text can be used. For expressive visual speech it uses visemes for textual content and four basic emotional states of the face as expressive speech basis. They work on a natural blending function of these states.

Sasha Fagel works on expressive speech in a broad sense[11]. He created a method to help creating expressive audiovisual databases by leading the subject through emotional stages to reach the desired level of expression gradually. This way it is possible to record emotionally neutral content (eg. "It was on Friday") articulated with joy or anger. The trick is to record the sentence multiple times and inserting emotionally relevant content between the occurrences. One example could be the sequence "Trouble happen always with me! It was on Friday. What do you think you are?! It was on Friday. I hate you! It was on Friday." This method gives the speaker the guide to express anger which gradually increases in expressiveness. The database will contain only the occurrences of emotionally neutral content.

1.2.4 Speech recognition today

As of 2010, after a decade, the hegemony of Hidden Markov Model (HMM) based ASR systems[12] is still standing. This approach uses a language model formulated with consecutiveness-functions, and a pronouncation model with confusion functions.

The main reason of the popularity of HMM based ASR systems is the efficiency of handling more thousands of words in a formal grammar. This grammar can be used to focus the vocabulary around a specific topic to increase the correctness and reducing the time complexity. HMM can be trained to a specific speaker but also can be trained on large databases to work speaker independent.

DOI:10.15774/PPKE.ITK.2010.007

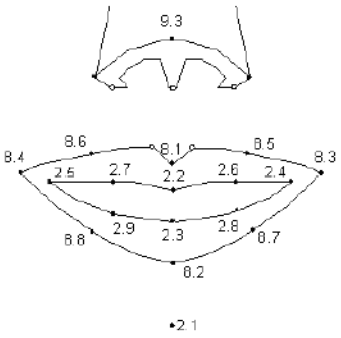

Figure 1.2: Mouth focused subset of feature points of MPEG-4.

1.2.5 MPEG-4

MPEG-4 is a standard for face description for communication. It uses Feature Points (FP) and Facial Animation Parameters (FAP) to describe the state of the face. The constant properties of a face also can be expressed in MPEG-4, for example the sizes in FAPU.

The usage of MPEG-4 is typical in multimedia applications where an interactive or highly compressed pre-recorded behavior of a face is needed, such as video games or news agents. One of the most popular MPEG-4 systems is Facegen[13].

For visual speech synthesis MPEG-4 is a fair choice since there are plenty of implementations and resources. The degree of freedom around the mouth is close to the actual needs, but there are features which can not be modeled with MPEG-4, such as inflation. Gerard Bailly et al showed that using more feature points around the mouth can increase naturalness significantly[14].

1.2.6 Face rendering

The task of the synthesis of the picture from face descriptors is face rendering. Usually 3D engines are used, but 2D systems are also can be found. The spectrum of the approaches and image qualities is very wide from efficient simple implementations to muscle based simulations[4].

Most of the face renderers use 3D acceleration and vertex arrays to interpolate, which is a fairly accelerated operation in today's video cards. In this case a few given vertex array represent given phases of the face, and using interpolation techniques, the status of the face can be expressed as a weighted sum of the proper vertex arrays. The resulting state can be textured and lighted just as one of the original designed facial phases.

DOI:10.15774/PPKE.ITK.2010.007

1.3 Open questions

It is clear that face modeling and facial animation — subtasks of audio to visual speech conversion — are still evolving but mainly a development fields, but there are examples of research areas, such as approximation of the human skin's physical properties, connection of the modalities, evaluation of AV mapping, what is speaker dependent in the articulation, what is the minimal necessary degrees of freedom for perfect facial modeling.

My motivations cover the applicable research on the connection between the modalities. This is also an open question. There are convenient arguments for the physical relation between the modalities: the speech organs are used both for audio and visual speech, although some of them are not visible. There must are physical effects of the visible speech organs to the audio speech.

On the other hand, there are phenomena where the connection between the modalities are minimal. Speech disorders can effect the audio speech without a visible trace. Ventriloquism (the art of speaking without lip movement, usually performed with puppets creating the illusion of a speaking puppet) is also an interesting exception.

To avoid inconsistency I turned to the clarified topic of audio to visual speech conversion.

1.4 The proposed approach of the thesis

The physical connection between the modalities can give guideline to reach basic conversion from audio to video, but this goal is not clear without specified aspects of qualities. The next chapter will detail how our research group met the field through the aid of deaf and hard of hearing people. Their main quality aspects of the resulting visual speech are the lip-readability, and the naturalness. This way the problem can be redefined to search the most appropriate visual speech for the given audio speech signal, not to restore the original visual articulation.

The physical connection can be utilized easily through direct conversion. Direct ATVS systems 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 of the process is left to the lip-reader. Because of this, our ATVS uses no phoneme recognition, furthermore there is no classification part in the process. This is the direct ATVS, avoiding any discrete type of data in the process. Discrete ATVS systems 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.

One of the most important benefits of the direct conversion is the chance to conserve nonverbal content of the speech such as prosody, dynamics and rhythm. Modular ATVS systems have to synthesize these features to maintain the naturalness of the result.

DOI:10.15774/PPKE.ITK.2010.007

1.5 Related disciplines

1.5.1 Speech inversion

Our 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[15, 16]. ATVS is different, the target is to produce a lip-readable animation which depends only on the visible speech organs and does not depend on the speaker dependent features of the speech signal.

Speech inversion aims to recover the state sequence of the speech organs from speech. A very simple model and solution of this problem is the vocal tract. Recent research on this field concerns tongue motion and models in particular.

1.5.2 Computer graphics

Synthesis of human face is a challenging field of computer graphics. The main reason of the high difficulty is the very sensitive human observer. The humankind developed a highly sophisticated communication system with facial expressions, it is a basic human skill to identify emotional and contextual content from a face. An example of cutting edge face synthesis systems is the rendering system of the movie Avatar (2009) where the system parameters were extracted from actors[17]. There are recent scientific results of efficient volume conserving deformations of facial skin based on muscular modeling[4]. These modern rendering methods can reproduce creasing of the face, which is perceptionally important.

1.5.3 Phonetics

The science of phonetics is related to ATVS systems by the ASR based approaches. Phonetically interesting areas are the ASR component, the phoneme string processing, the rules applied on phoneme strings to synthesize visual speech, such as interpolation rules, or dominance rules.

The details of articulation, and the relation of the phonetic content and the facial muscle controls is the topic of articulatory phonetics[18, 19]. This field classifies phonemes by their places of articulation: labial-dental, coronal, dorsal, glottal. ATVS systems are aware of visible speech organs, so labial-dental consonants are important, along vowels and articulations with open mouth. For example the phoneme “l” is identifiable of it’s alveolar articulation since it is done with opened mouth.

Articulatory phonetics have important results for ATVS systems, as we will see the details of visual speech synthesis from phoneme strings.

DOI:10.15774/PPKE.ITK.2010.007

Chapter 2

Motivation and the base system

In this chapter I will describe the main tasks I had to deal with, showing the motivation of my thesis. I will describe a base system as well. The base system itself is not part of the contribution of my thesis, although understanding the base system is important to understanding my motivations.

2.1 SINOSZ project

The original project with SINOSZ (National Association of Deaf and Hearing Impaired) aimed a mobile system to help dealing with audio only information sources for hard of hearing people. The first idea was to visualize the audio data in some learnable representation, but the association rejected any visualization technique which must be learned by the deaf community, so the visualization method had to be some already known representation of the speech. We had basically two choices, to implement an ASR to get text, or translate to facial motion. We expected more efficient and robust quality of facial motion conversion with the capabilities of a mobile device in 2004.

The development of the mobile application was initiated with the project. The mobile branch of the project is out of the scope of my thesis, although the requirement of efficiency is important.

2.1.1 A practical view

When I started to work on audio to visual speech conversion, after examining some of the systems in aspects of requirements and qualities detailed in the previous chapter, I decided to use direct conversion. The main reason in this time was to get a functional and efficient test system as soon as possible to have results and first hand experience with the hope of sufficially efficient implementation later.

Direct conversion can be deployed on mobile platforms easier than database dependent classifier systems. Not only the computational time is moderated, but the memory requirements are also lower. Choosing direct conversion was the option of the guaranteed possibility of the test implementation on the target platform.

DOI:10.15774/PPKE.ITK.2010.007

2.2 Lucia

In the beginning of the project I convinced the team to use direct mapping between modalities. My two important reasons were the efficiency and the lack of the requirement of a labeled database unlike an ASR. Since we did not have any audiovisual databases (and even in 2009 there are quite few publicly available) we had to think on not only the system but the database also. Direct conversion does not need labeled data so manual work can be minimized, which shortens the production time.

So the planned system contained a simple audio preprocessing (LPC or MFCC), a direct mapping to video from audio feature vectors via code-book or neural network, and visualization of the result on an artificial head. We did not have any face models, neither wanted to create one, so we were looking for an available head model.

The first test system used the talking head of Cosi et al[10] called Lucia. The head model was originally used for expressive speech synthesis. The system used MPEG-4 FAP as input, and generated a run-time video in an OpenGL window, and exporting in video files was also available.

2.3 The base system

The base system is an implementation of direct conversion2.1.

2.3.1 Database building from video 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 a training set for neural networks. For this purpose the data should not contain strong redundancy for proactically acceptable learning speed, so the pre-processing includes the choice of an appropriate representation also. With inadequate representation the learning may take months, or may not even converges.

2.3.2 Audio

The voice signal is processed by 25 frame/s rate to be in synchrony with the processed video signal. One analysis window is 20-40 ms, the maximum number of samples in the 40 ms window to be 2^n samples. The input speech can be pre-emphasis filtered with H(z)=1-0.983z-1. Hamming window and FFT with Radix-2 algorithm are applied. The FFT spectra is converted to 16 mel-scale bands, and logarithm and DCT is applied. Such Mel Frequency Cepstrum Coefficients (MFCC) feature vectors are commonly used in general speech recognition tasks. The MFCC feature vectors provide the input to the neural networks after scaling to [-0.9 .. 0.9].

DOI:10.15774/PPKE.ITK.2010.007

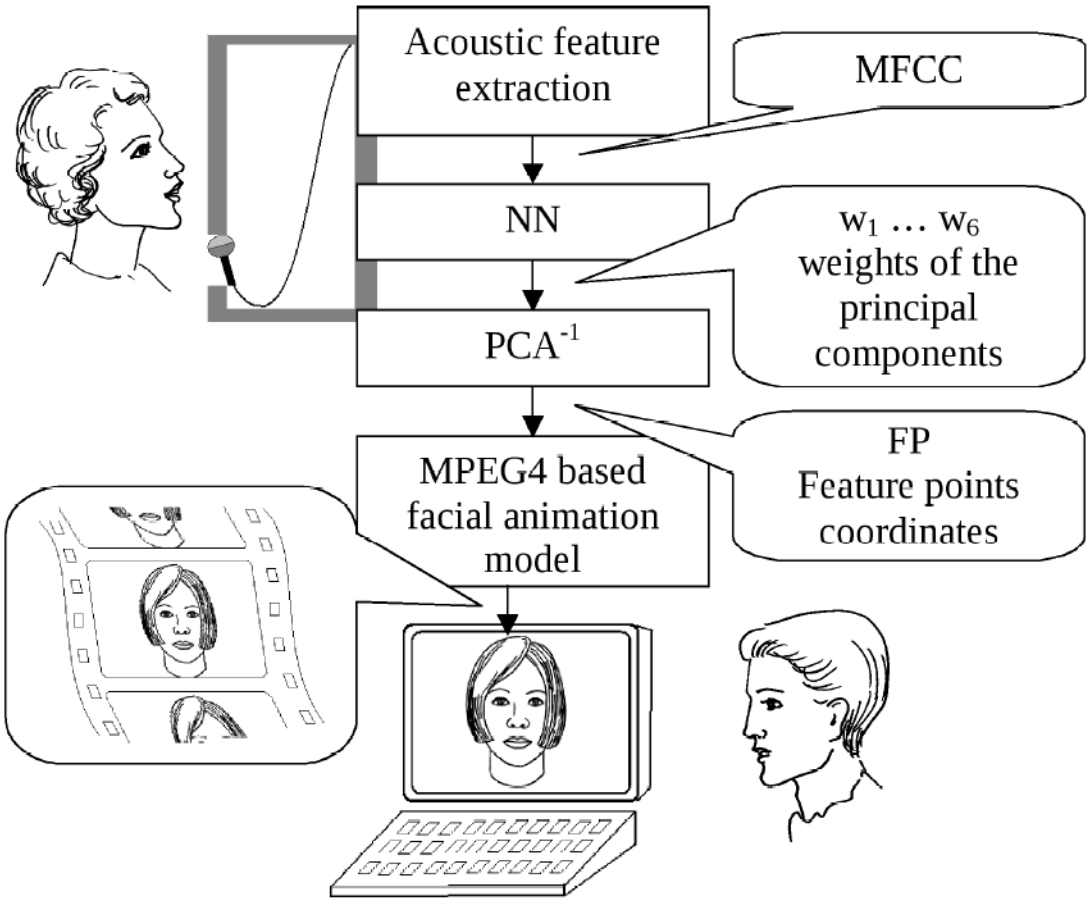

Figure 2.1: Workflow used in Lucia.

DOI:10.15774/PPKE.ITK.2010.007

2.3.3 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 such 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 a lossy compression because 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 first 4 coefficient can be seen on Fig 2.2.

The base systems's 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 marker tracking algorithm used the number of markers (nm) as input, and on each frame it looked for the nm most marker-like areas of the picture. The marker-likeliness was given as high energy fixed sized blob after yellow filtering. The tracking contained a self-check by looking for additional markers, and by comparing the marker-likelinesses of the marker $[.., nm-1, nm, nm+1, ..]$ the good tracking show strong decrease after the nm marker. If the decrease is before nm there are missing markers, if the decrease if after nm there are misleading blobs in the frame. Using the self-check of the tracking, manual corrections was made.

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 [8].

2.3.4 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 guarantees 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 has to be chosen based on the optimal temporal scope.

DOI:10.15774/PPKE.ITK.2010.007

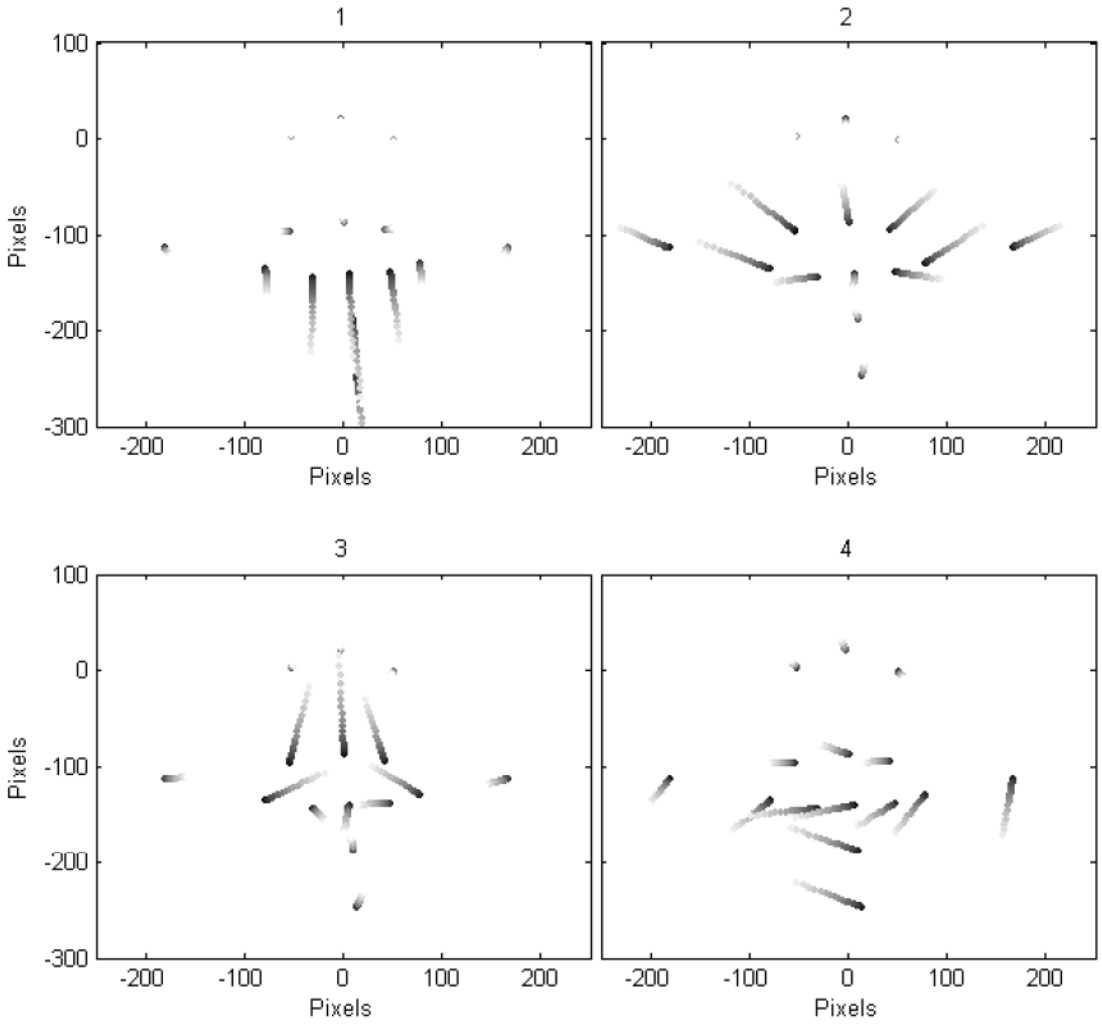

Figure 2.2: Principal components of MPEG-4 feature points in the database of a professional lip-speaker.

DOI:10.15774/PPKE.ITK.2010.007

The neural network is a back-propagation implementation by Davide Anguita called Matrix BackPropagation[20]. This is a very efficient software, we use a slightly modified version of the system to able to continue a training session.

2.3.5 First 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 ATVS and a real person's video, the situation had to be in consideration of average lip-reading cases. As we found [8] 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 subjects 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 ATVS 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 2.1 shows the results.

Table 2.1: Recognition rates of different video clips.

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

2.3.6 Discussion

In this case the 45% should be compared to the 55%. The face models, which is driven by recorded visual speech data is the best possible behavior of the direct ATVS system. The ratio of the results is 87%. It means that the base system can make deaf people to understand 87% of the best possible system. This was very encouraging exprerience.

The 55% is a weak result compared to the 97% of original video. This falloff is because of the artifical head. This ratio could be enhanced by using more sophisticated head models and facial parameters, but this direction of research and development is out of scope of this thesis.

DOI:10.15774/PPKE.ITK.2010.007

2.4 Johnnie Talker

Johnnie Talker is a real-time system with very low time complexity, pure AV mapping application with a simple face model and facial animation. Johnnie was implemented to demonstrate the low time complexity of direct ATVS approach. The application is a development of our research group, it uses my implementation of AV mapping, Tamás Bárdi's audio preprocessing and Bálint Srancsik's OpenGL based head model.

Johnnie is freely downloadable from the webpage of the author of this dissertation[21]. It is a Windows application using OpenGL.

Because of the demand of low latency no theoretical delay was used in this system. In the next few chapters I will describe how the naturalness and the intelligibility can be enhanced by using a time window in the future of audio modality. This can be implemented by delaying the audio dub to maintain audio-video synchrony and using the future audio in the same time. For example a phone line can be delayed to the theoretically optimal time window. But since Johnnie can be used via microphone, which can not be delayed, any additional buffering would cause noticeable latency which ruins the subjective sense of quality. As one of the next chapters will describe, subjective quality evaluation depends heavily on audio-video synchrony, and this phenomena appears strongly in the perception of a synthesized visual speech of one's own speech in real-time.

Johnnie Talker was shown on various international conferences with success. It was a good opportunity to test language independence in practice.

We were looking for techniques to improve the qualities of the real-time system without additional run-time overhead. There will be a chapter about a method which can enhance speaker independence of the system using only database modifications, so no run-time penalty needed.

2.5 Extending direct conversion

2.5.1 Direct ATVS and co-articulation

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.

Co-articulation is the phenomena of transient phases in speech process. In audio modality, co-articulation is the effect of the neighboring phonemes to the actual state of speech in a short window of the time, shorter than a phoneme duration. In speech synthesis, there is a strong demand to create natural transients between the clean states of speech. In visual speech synthesis this issue is also important. In the visual speech process there are visemes even if the synthesizer does not explicitly use this concept. Visual co-articulation can be defined as a system of influences between visemes in time. Because of biological limitations, visual co-articulation is slower than audio co-articulation, but similar in other ways: neighboring visemes can have effect on each

DOI:10.15774/PPKE.ITK.2010.007

other, there are stronger visemes than others, and most of the cases can be described or approximated as an interpolation of neighboring visemes.

Let me call a system *visual speech transient model*, if it generates mediate states of visual speech units, such as visemes. An example of visual speech transient model is the strictly adopted co-articulation concept on visemes, the visual co-articulation, since the viseme string processing has to decide how interpolation should take place between the visemes. Another example of visual speech transient models is the direct conversion's adaptation to longer time windows in order to include more than one phoneme on the audio modality. In this case the transients depend on acoustical properties. In modular ATVS systems, the transients are coded in rules depending on viseme string neighborhoods.

Utilization

Training a direct ATVS needs audio-video data pairs. Since plenty of speech audio databases exist but only a few audiovisual ones, building a direct ATVS means building a multimodal database first. A discrete ATVS is a modular system, it is possible to use existing speech databases to train voice recognition, and separately train the animation part on phoneme pairs or trigraphs[1]. Therefore direct ATVS needs a special database, but the system will handle energy and rhythm naturally, meanwhile a discrete ATVS has to reassemble the phonemes into a fluid coarticulation chain of viseme interpolations. Let use the term "temportal scope" for the overall time of a coarticulation phenomena, which means that the state of the mouth is depending on this time interval of the speech signal. In direct ATVS the calculation of a frame is based on this audio signal interval. In discrete ATVS the visemes and the phonemes are synchronized and interpolation is applied between them, as it is popular in text to visual speech systems [22]. Figure 2.3 shows this difference.

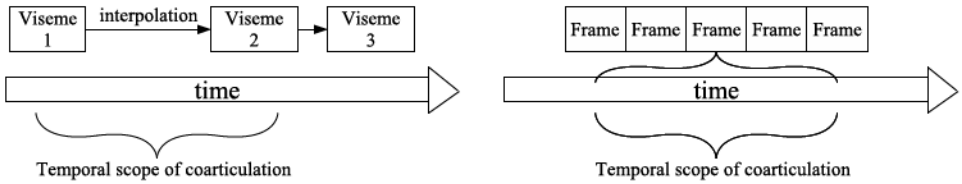

Figure 2.3: Temporal scope of discrete (interpolating) and direct ATVS

Asymmetry

As mutual information estimation resulted any given state of the video data stream can be calculated fairly on a definable relative time window of the speech signal. This model predicts that the transient phase of the visible speech can be calculated in the same way as in the steady phase as Figure 2.3 shows.

This model gives a prediction about temporal asymmetries in the multimodal speech process. This asymmetry can be explained with mental predictivity in the motion of

DOI:10.15774/PPKE.ITK.2010.007

the facial muscles to fluently form the next phoneme.

Details will follow in chapter "Temporal asymmetry".

Speaker independence

Since the direct conversion is usually an approximation trained on a given set of audio and video states, it suffers heavy dependence on the database. As I detailed before, for a good direct ATVS a good lip-speaker needed to share visual data with the system. Talented lip-speakers are rare, and most of the experienced lip-speakers are women. This means that a single recording of one lip-speaker gives not only a speaker dependent system, but collecting more professional lip-speakers would result a gender dependent system, since the statistics of the data would heavily biased, or it is very difficult to collect enough male lip-speaker to the system.

Even if we would have plenty of professional lip-speakers, there is a question about the mixing the video data. People articulate differently. It is not guaranteed that a mixture of good articulations result even an acceptable articulation. The most safe solution is to choose one of the lip-speakers as a guaranteed high quality articulation, and trying to use his/her performance with multiple voices.

I will give a solution for this problem in chapter "Speaker independence in direct conversion" .

2.5.2 Evaluation

The base system was published as a standalone system, and was measured with subjective opinion scores and intelligibility tests with deaf persons. In my chapter about the comparison of AV mappings, I will position the direct ATVS among the others used in the world.

Oddly there are quite few publications on direct ATVS. This is strange, because the system is one of the most simple designs. Let me tell a personal experience from a conference of EUSIPCO, Florence. A young researcher was interested in our Johnnie demo. He was from ATR, Japan, and he praised our system. As I explained the workflow of the system, on each stage he said "We did the same". Even the number of PCA coefficients was the same. At the end, he said that their system produce significantly worse results than ours, it was not even published because it was flawed. We agreed then that the most important difference is the lip-speaker's professionality.

His work can be read in japanese [23] in the annual report of the institute, by the way in the same year we published our results in Hungarian [24, 25, 26]

Another example of direct conversion publicity is a comparison study of an unpublished direct system[27] used as internal baseline.

Because of this underpublicity, it is important to position direct ATVS among the more popular modular ATVS systems, since most of the visual speech synthesis research groups also try to implement a direct ATVS, but their efforts fail because of the quality of the database. At the first glance this may seem as bad news for our research, but the novelty of our system is still unharmed since the work in ATR was identical only in

DOI:10.15774/PPKE.ITK.2010.007

the technical details, and a training system's technology itself, without the database is not a whole system. Our base system is new because of the new training data, and the finding of the need of the professional lip-speaker. Again, I would like to emphasize that difference between our base system and the one developed in ATR is not "only" the database but the training strategy, which is one of the most important and fundamental part of any learning system.

This new and successfull learning strategy makes our base system novel, but in this thesis I focus the results of my own, not the research group. Johnnie Talker is a contribution of the group, and the following extensions and measurements are contribution of the author of this thesis.

In the next chapter I will show how the base system with the essential database of the professional lip-speaker can be ranked among the widely used ATVS systems.

DOI:10.15774/PPKE.ITK.2010.007

Chapter 3

Naturalness of direct conversion

In this chapter I discuss the measurement of the naturalness of synthetic visual speech, and comparison of different AV mapping approaches.

3.1 Method

A comparative study of audio-to-visual speech conversion is described in this chapter. The direct feature-based conversion approach is compared to various indirect ASR-based solutions. The already detailed base system was used as direct conversion. The ASR based solutions are the most sophisticated systems actually available in Hungarian. 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.

The thesis about the results of the comparison is important because no AV mapping comparisons were done before with the novel traning database of professional lip-speaker.

3.1.1 Introduction

A difficulty that arises in comparing the different approaches is that they usually are developed and tested independently by the respective research groups. Different metrics are used, e.g. intelligibility tests and/or opinion scores, and different data and viewers are applied [28]. In this chapter I describe a comparative evaluation of different AV mapping approaches within the same workflow, see Figure 3.1. The performance of each is measured in terms of intelligibility, where lip-readability is measured, and naturalness, where a comparison with real visual speech is made.

3.1.2 Audio-to-visual Conversion

The performance of five different approaches will be evaluated. These are summarized as follows:

DOI:10.15774/PPKE.ITK.2010.007

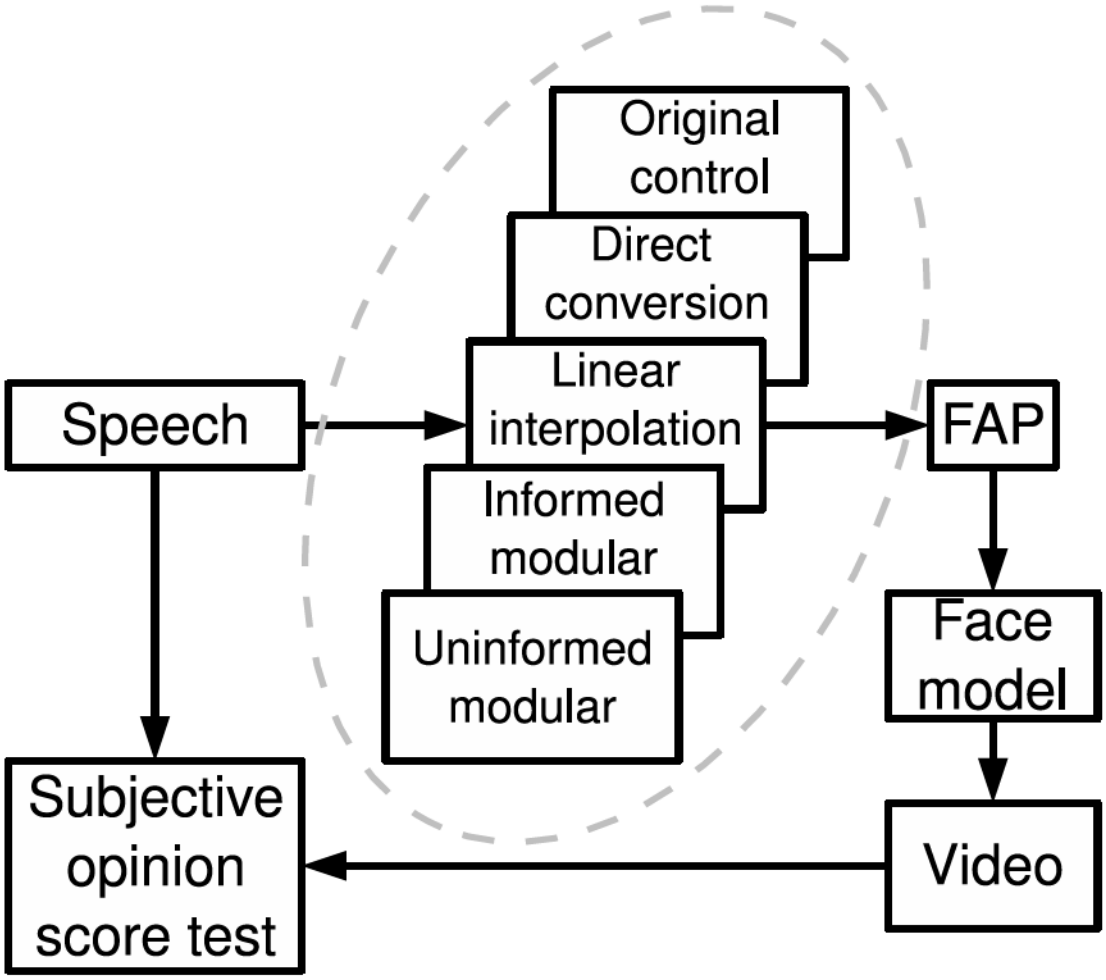

Figure 3.1: Multiple conversion methods were tested in the same environment

DOI:10.15774/PPKE.ITK.2010.007

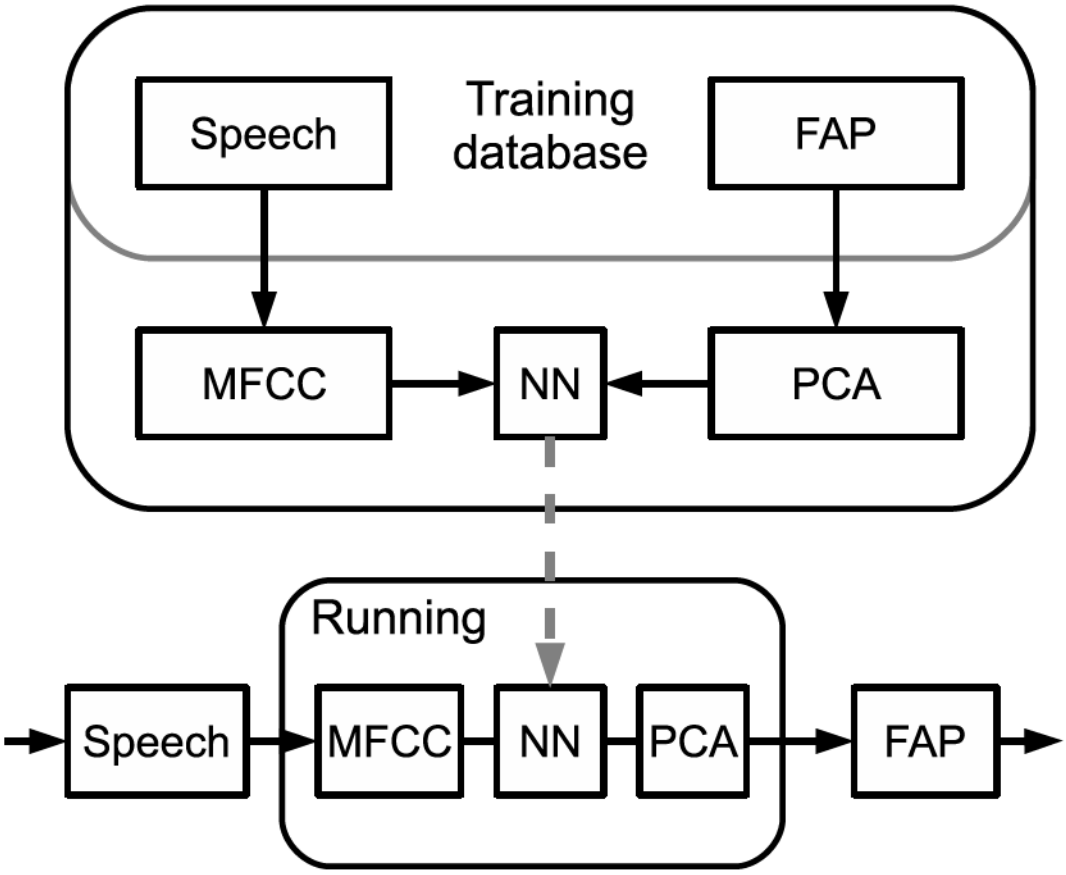

Figure 3.2: Structure of direct conversion.

- A reference based on natural facial motion.
- A direct conversion system.
- An ASR based system that linear interpolates phonemic/visemic targets.
- An *informed* ASR-based approach that has access to the vocabulary of the test material (IASR).
- An *uninformed* ASR (UASR) that does not have access to the text vocabulary.

These are described in more detail in the following sections.

Direct conversion

We used our base system, with a database of a professional lip-speaker. The length of the recorded speech was 4250 frames.

DOI:10.15774/PPKE.ITK.2010.007

ASR-based conversion

For the ASR based approaches a Weighted Finite State Transducer — Hidden Markov-Model (WFST-HMM) decoder is used. Specifically, a system known as VOXerver [29] is used, which can run in one of two modes: *informed*, this exploits knowledge of the vocabulary of the test data, and *uninformed*, which does not. Incoming speech is converted to MFCCs, after which blind channel equalization is used to reduce linear distortion in the cepstral domain [30]. Speaker independent cross-word decision-tree based triphone acoustic models are applied, which previously are trained using the MRBA Hungarian speech database [31], which is a standardized, phonetically balanced Hungarian speech database developen on the Budapest University of Technology and Economics.

The uninformed ASR system uses a phoneme-bigram phonotactic model to constrain the decoding process. The phoneme-bigram probabilities were estimated from the MRBA database. In the informed ASR system a zerogram word language model is used with a vocabulary size of 120 words. Word pronunciations were determined automatically as described in [32].

In both types of speech recognition approaches the WFST-HMM recognition network was constructed offline using the AT&T FSM toolkit [33]. In the case of the informed system, phoneme labels were projected to the output of the transducer instead of word labels. The precision of the segmentation is 10 ms.

Viseme interpolation

To compare the direct and indirect audio-to-visual conversion systems, a standard approach for generating visual parameters is to first convert a phoneme to its equivalent viseme via a look up table, then linear interpolate the viseme targets. This approach to synthesizing facial motion is oversimplified because coarticulation effects are ignored, but it does provide a baseline on expected performance (worst-case scenario).

Modular ATVS

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. To learn the dominance functions an ellipsoid is fitted to the lips of speakers in a video sequence articulating Hungarian triphones. To aid the fitting, the speakers wear a distinctly colored lipstick. Dominance functions are estimated by the variance of visual data in a given phonetic neighborhood set. The learned dominance functions are used to interpolate between the visual targets derived from the ASR output [34]. We use the implementation of László Czap and János Mátyás here which produces Poser script. FAPs are extracted from this format by the same workflow as from an original recording.

DOI:10.15774/PPKE.ITK.2010.007

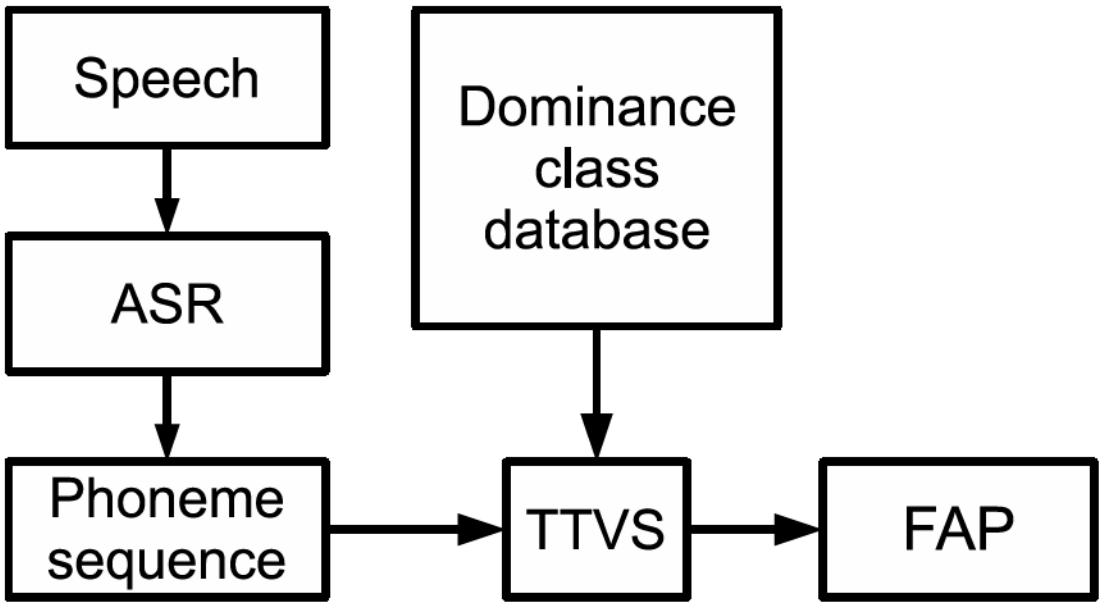

Figure 3.3: Modular ATVS consists of an ASR subsystem and a text to visual speech subsystem.

Rendering Module

The visualization of the output of the ATVS methods is common to all approaches. The output from the ATVS modules are facial animation parameters (FAPs), which are applied to a common head model for all approaches. Note, although better facial descriptors than MPEG-4 are available, MPEG-4 is used here because our motion capture system does not provide more detail than this. The rendered video sequences are created from these FAP sequences using the Avisynth [35] 3D face renderer. As the main components for the framework are common between the different approaches, any differences are due to the differences in the AV mapping methods. Actual frames are shown on Fig 3.5.

3.1.3 Evaluation

Implementation specific noncritical behavior (eg. articulation amplitude) should be normalized to ensure that the comparison is between the essential qualities of the methods. To discover these differences, a preliminary test is done.

Preliminary test

To tune the parameters of the systems, 7 videos were generated by each of the five mapping methods, and some sequences were re-synthesized from the original facial motion data. All sequences started and ended with a closed mouth, and each contained between 2-4 words. The speaker participated in all of the tests was not one of those who

DOI:10.15774/PPKE.ITK.2010.007

Table 3.1: Results of preliminary tests used to tune the system parameters. Shown are the average and standard deviation of scores.

Method	Average score	STD
UASR	3.82	0.33
Original	3.79	0.24
Linear	3.17	0.4
Direct	3.02	0.41
IASR	2.85	0.72

were involved in training of the audio-to-visual-mapping. The videos were presented in a randomized order to 34 viewers whom were asked to rate the quality of the systems using an opinion score (1–5). The results are shown in Table 3.1.

The results were unexpected, the IASR, which uses a more sophisticated coarticulation model, was expected to be one of the best performing systems. Closer investigation of the lower scores showed the reason was rather due to poorer audiovisual synchrony of IASR than for UASR. The reason of this phenomena is the difference of the mechanism of the informed and the uninformed speech recognition process. During informed recognition the timing information is produced as a consequence of the alignment of the correct phonemes to the signal, which presses the segment boundaries by using the certain phonetic information. The uninformed recognition may miscategories the phoneme but the acoustical changes are the driver of the segment boundaries, so the resulting segmentation is closer to the acoustically reasonable than the phonetically driven segmentation.

A qualitative difference between the direct and indirect approaches is the degree of mouth opening — the direct approach tended to open the mouth on average 30% more than the indirect approaches. Consequently, to bring the systems into the same dynamic range, the mouth opening for the direct mapping was damped by 30%. The synchrony of the ASR-based approaches was checked for systemic errors (constant or linearly increasing delays) using cross correlation of locally time shifted windows, but no systematic patterns of errors were detected.

3.1.4 Results

ASR subsystem

The quality of the ASR-based approach is affected by the recognized phoneme string. This typically is 100% for the informed system as the test set consists only of a small number of words (months of the year, days of the week, and numbers under 100), whilst the uninformed system has a typical error rate of 25.21%. Despite this the ATVS using this input performs surprisingly well. The likely reason might be the pattern of confusions — often phonemes that are confused acoustically appear visually similar on the lips. A second factor that affects the performance of the ASR-based approaches is precision of the segmentation. Generally the uninformed systems are more precise on

DOI:10.15774/PPKE.ITK.2010.007

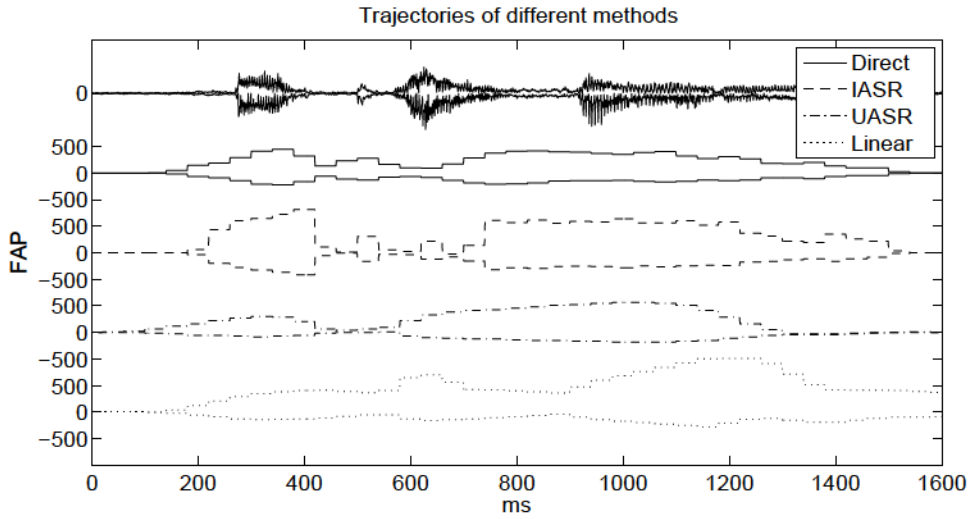

Figure 3.4: Trajectory plot of different methods for the word "Hatvanhárom" (hɔtvɔnha:rom). Jaw opening and lip opening width is shown. Note that the speaker did not pronounce the utterance perfectly, and the informed system attempts to force a match with the correctly recognized word. This leads to time alignment problems.

the average than the informed systems. The precision of the segmentation can severely impact on the subjective opinion scores. We therefore first attempt to quantify these likely sources of error.

The informed recognition system is similar in nature to forced alignment in standard ASR tasks. For each utterance the recognizer is run in forced alignment mode for all of the vocabulary entries. The main difference between the informed and the uninformed recognition process is the different Markov state graphs for recognition. The informed system is a zerogram without loopback, while the uninformed graph is a bigram model graph where the probabilities of the connections depend on language statistics.

While matching the extracted features with the Markovian states, the differences are cumulated in both scenarios. However, the uninformed system allows for different phonemes outside of the vocabulary to minimize the cumulated error. For the informed system only the most likely sequence is allowed, which can distort the segmentation — see Figure 3.4 for an example where the speaker mispronounces the word "Hatvanhárom" (hɔtvɔnha:rom, "63" in Hungarian). The (mis)segmentation of ɔtvɔ means IASR AVTS system opens the mouth after the onset of the vowel. Human perception is sensitive to this error and so this severely impacts the perceived quality. Without forcing the vocabulary, a system may ignore one of the consonants but open the mouth at the correct time.

Note that the generalization of this phenomena is out of the scope of this work. We have demonstrated that this is a problem with certain implementations of HMM-based ASR. Alternative, more robust implementations might alleviate these problems.

DOI:10.15774/PPKE.ITK.2010.007

Table 3.2: 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

Subjective opinion scores

The test setup is similar to the preliminary test described previously to tune the system. However, 58 viewers are used, and only quantative opinion survey was made on the scale of 1 (bad, very artifical) to 5 (real speech).

The result of the opinion score test is on Table 3.2. The advantage of direct conversion against UASR is on the edge of significance with $p = 0.0512$ as well as the difference between the original speech and the direct conversion with $p = 0.06$ but UASR is significantly worse than original speech with $p = 0.00029$. The results compared to the preliminary test also show that with respect to naturalness, the excessive articulation is not significant. The advantage of correct timing over correct phoneme string is also significant.

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.

Intelligibility

Intelligibility was measured with a test of recognition of video sequences without sound. This is not the popular Modified Rhyme Test[36] but for our purposes with hearing impaired viewers it is more relevant, since the keyword spotting is the most common lip-reading task. The 58 test subjects had to guess which word was said from a given set of 5 other words of the same category. The categories were numbers, names of months and the days of the week. All the words were said twice. The sets were intervals to eliminate the memory test from the task (for example “2” , “3”, “4”, “5”, “6” can be a set). This task models the situation of hearing impaired or very noisy environment where an ATVS system can be used. It is assumed that the context is known, so the keyword spotting is the closest task to the problem.

The performance of the audio-to-visual speech conversion methods reverse in this task compared to naturalness. The main result here is the dominance of ASR based approaches (Table 3.3), and the insignificance of the difference between informed and uninformed ATVS results ($p = 0.43$) in this test which may deserve further investigation. Note that as the synchrony is not an issue without voice, the IASR is the best.

DOI:10.15774/PPKE.ITK.2010.007

Table 3.3: Results of recognition tests, average and standard deviation of success rate in percent. Random pick would give 20%.

Method	Precision	STD
IASR	61%	20%
UASR	57%	22%
Original motion	53%	18%
Cartoon	44%	11%
Direct conversion	36%	27%

Table 3.4: Comparison to the results of Öhman and Salvi[27], a HMM and rule based systems intelligibility test. Intelligibility of corresponding methods are similar.

Methods	Prec.	Prec.
IASR / Ideal	61%	64%
UASR / HMM	57%	54%
Direct / ANN	36%	34%

As a comparison with [27] where intelligibility is tested similarly, manually tuned optimal rule based facial parameters are close to our IASR since there was no recognition error, and without voice the time alignment quality is not important, and our TTVS is rule based. Their HMM test is similar to our UASR, because both are without vocabulary, both are targeting time aligned phoneme string to be converted to facial parameters, and our ASR is HMM based. Their ANN system is very close to our direct conversion except the training set, it is a standard speech database audio, and a rule based calculated trajectory video data, while our system is trained on actual recording of a professional lip-speaker. However the results concerning intelligibility are close to each other, see Table 3.4. This is a validation of the results, since the corresponding measurement are close to each other. It is important that [27] tests only intelligibility, and only between three methods of ours, so our measurement is broader.

3.1.5 Conclusion

I presented a comparative study of audio-to-visual speech conversion methods. We have presented a comparison of our direct conversion system with conceptually different conversion solutions. A subset of the results correlate with already published results, validating the approach of the comparison.

We observe higher importance of the synchrony over phoneme precision in an ASR based ATVS system. There are publications on the high impact of correct timing in different aspects [34, 37, 38], but our result show explicitly that more accurate timing achieves much better subjective evaluation than more accurate phoneme sequence. Also, we have shown that in the aspect of subjective naturalness evaluation, direct conversion (trained on professional lip-speaker articulation) is a method which produces the highest opinion score of 95.9% of an original facial motion recording with lower

DOI:10.15774/PPKE.ITK.2010.007

computational complexity than ASR based solutions.

For tasks where intelligibility is important (support for hearing impaired, visual information in noisy environment) modular ATVS is the best approach among those presented. Our mission of aiding hearing impaired people call upon us to consider using ASR based components. For naturalness (animation, entertaining applications) direct conversion is a good choice. For both aspects UASR gives relatively good but not outstanding results.

3.1.6 Technical details

Marker tracking was done for MPEG-4 FP 8.8 8.4 8.6 8.1 8.5 8.3 8.7 8.2 5.2 9.2 9.3 9.1 5.1 2.10 2.1. During synthesis, all FAPs (MPEG-4 Facial Animation Parameter) connected these FPs were used except depth information:

- open_jaw
- lower_t_midlip
- raise_b_midlip
- stretch_l_cornerlip
- stretch_r_cornerlip
- lower_t_lip_lm
- lower_t_lip_rm
- raise_b_lip_lm
- raise_b_lip_rm
- raise_l_cornerlip
- raise_r_cornerlip lower_t_midlip_o
- raise_b_midlip_o
- stretch_l_cornerlip_o
- stretch_r_cornerlip_o
- lower_t_lip_lm_o
- lower_t_lip_rm_o
- raise_b_lip_lm_o
- raise_b_lip_rm_o
- raise_l_cornerlip_o
- raise_r_cornerlip_o

Inner lip contour is estimated from outer markers.

Yellow paint was used to mark the FP locations on the face of the recorded lip-speaker. The video recording is 576i PAL (576x720 pixels, 25 frame/sec, 24 bit/pixel). The audio recording is mono 48kHz 16 bit in a silent room. Further conversions were depended on the actual method.

Marker tracking was based on color matching and intensity localization frame to frame and the location was identified by the region. In overlapping regions the closest location on the previous frame was used to identify the marker. A frame with neutral face was selected to use as the reference to FAPU measurement. The marker on the nose is used as reference to eliminate head motion.

The direct conversion uses a modification of Davide Anguita's Matrix Backpropagation which enables real-time work also. The neural network used 11 frame long window on the input side (5 frames to the past and 5 frames to the future), and 4 principal

DOI:10.15774/PPKE.ITK.2010.007

component weights of FAP on the output. Each frame on the input is represented by 16 band MFCC feature vector. The training set of the system contains standalone words and phonetically balanced sentences.

In the ASR the speech signal was converted to a frequency of 16kHz. MFCC (Mel Frequency Cepstral Coefficients)-based feature vectors were computed with delta and delta-delta components (39 dimensions in total). The recognition was performed on a batch of separated samples. Output annotations and the samples were joined, and the synchrony between labels and the signal was checked manually.

The visemes to the linear interpolation method were selected manually for each viseme in Hungarian from the training set of the direct conversion. Visemes and phonemes were assigned by a table. Each segment is a linear interpolation from the actual viseme to the next one. Linear interpolation was calculated in the FAP representation.

TTVS is a Visual Basic implemented system with a spreadsheet of the timed phonetic data. This spreadsheet was changed to the ASR output. Neighborhood dependent dominance properties were calculated and viseme ratios were extracted. Linear interpolation, restrictions concerning biological boundaries and median filtering were applied in this order. The output is a Poser data file which is applied to a model. The texture of the model is modified to black skin and differently colored MPEG-4 FP location markers. The animation was rendered in draft mode, with the field of view and resolution of the original recording. Marker tracking was performed as described above with the exception of the differently colored markers. FAPU values were measured in the rendered pixel space, and FAP values were calculated from FAPU and tracked marker positions.

This was done for both ASR runs, uninformed and informed.

The test material was manually segmented to 2-4 word units. The lengths of the units were around 3 seconds. The segmentation boundaries were listed and the video cut was automatically done with an Avisynth script. We used an MPEG-4 compatible head model renderer plugin for Avisynth, with the model “Alice” of XFace project. The viewpoint and the field of view was adjusted to have only the mouth on the screen in frontal view.

During the test the subjects watched the videos fullscreen and used headphones.

DOI:10.15774/PPKE.ITK.2010.007

3.2 Thesis

> ***I. I showed that 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.*** [39]

3.2.1 Novelty

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.

3.2.2 Measurements

Naturalness was measured as subjective similarity to human articulation. The measurement was blind and randomized, the number of test subjects was 58, and our direct AV mapping was not significantly worse than original visual speech, but the difference between the modular and the original was significant.

Opinion score averages and deviations shown no significant difference between human articulation and direct conversion, but significant difference between human and modular mapping based systems.

The measurement was done on Hungarian database, fluently read speech. The database contains mixed isolated words and sentences.

3.2.3 Limits of validity

Tests were done on normal speech database, with fully focused perception of the test subjects on good audio and video quality videos.

3.2.4 Consequences

Using direct conversion for areas where naturalness is most important is encouraged. Using professional lip-speaker to record audiovisual database increases the quality to be comparable with the level of human articulation. Other laboratories trained their systems with non-professionals, and those systems were not publicated due to their poor performance.

DOI:10.15774/PPKE.ITK.2010.007

Figure 3.5: An example of the importance of correct timing. Frames of the word "Oktober" show timing differences between methods. Note that direct conversion received best score even though it does not close the lips on bilabial but closes on velar, and it has problems with lip rounding.

DOI:10.15774/PPKE.ITK.2010.007

DOI:10.15774/PPKE.ITK.2010.007

Chapter 4

Temporal asymmetry

In this chapter I discuss the measurement of relevant time window for direct AV mapping, which is important to build a audio to visual speech conversion system since the temporal window of interest can be determined.

4.1 Method

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 result 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 by using the future speech sound to the conversion.

4.1.1 Introduction

Other research projects on conversion of speech audio signal to facial animation have concentrated on development of feature extraction methods, database construction and system training [40, 41]. Evaluation and comparison of different systems have also had high importance in the literature. In this chapter I discuss the temporal integration of acoustic features optimal for real-time conversion to facial animation. The critical part of such systems is the building of an optimal statistical model for the calculation the video features from the audio features. There in no known exact relation of the audio feature set and video feature set currently, this is an open question yet.

The speech signal conveys information elements in a very specific way. Some of speech sounds are related rather to a steady state of the articulatory organs, others rather to the transition movements [42]. Our target application is for providing a communication aid to deaf people. Professional lip-speakers have 5-6 phoneme/s speech rate to adapt the communication to the demand of deaf people so steady state phases and the transition phases of speech sounds are longer then in everyday speech style.

DOI:10.15774/PPKE.ITK.2010.007

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 and fitting of audio and video features. The measure of the fitting has based on the mutual information of audio and video features [43].

The base system uses an adjustable temporal window of audio speech signal. The neural network can be trained to respond to an array of MFCC windows, using the future and/or past audio data. The conversion can be as good as the amount of mutual information between the audio and video representations.

Using the trained neural net for calculation of control parameters of facial animation model

The audio processing unit extracts the audio MFCC feature vectors from the input speech signal. Five frames of MFCC vectors are used as input 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. 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. Low level mutual information means that we have low average chance to estimate the facial parameters from the audio feature set. The time shift value to produce the highest mutual information means the maximal average chance to calculate the one kind of known features from the other one.

Estimation of mutual information needs computation intensive algorithm. The calculation is unrealistic using large database with multidimensional feature vectors. So single MFCPCA and FacePCA parameters were interrelated. Since the single parameters are orthogonal but not independent, they are not additive. For example the FacePCA1 values are not independent from FacePCA2. The mutual information curves even in such complex cases can indicate the interrelations of 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 a lower estimation of theoretical maximum.

DOI:10.15774/PPKE.ITK.2010.007

Mutual information

$$MI_{X,Y} = \sum_{x \in X} \sum_{y \in Y} P(x,y) \log \frac{P(x,y)}{P(x)P(y)} \tag{4.1}$$

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:

$$\forall \Delta t \in [-1s, 1s] : MI(\Delta t) = \sum_{t=1}^{n} P(a_{t+\Delta t}, v_t) \log \frac{P(a_{t+\Delta t}, v_t)}{P(a_{t+\Delta t})P(v_t)} \tag{4.2}$$

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 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.

MFCPCA and FacePCA measurements

170 seconds of audio and video speech records was processed. The time shift has been varied 1 ms steps. Mel frequency coefficients are calculated for each element. Principal component analysis (PCA) has been applied for even more compact representation of audio features since PCA components can represent the original speech frames by minimal average error at given subspace dimensionality. In the following the speech frames are described by such MFCPCA parameters.

The MFCPCA parameters are more readable representation of frames for human experts than PCA of MFCC feature vectors.

The MFCPCA parameters have direct relations to the spectrum. 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 4.1. 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.

DOI:10.15774/PPKE.ITK.2010.007

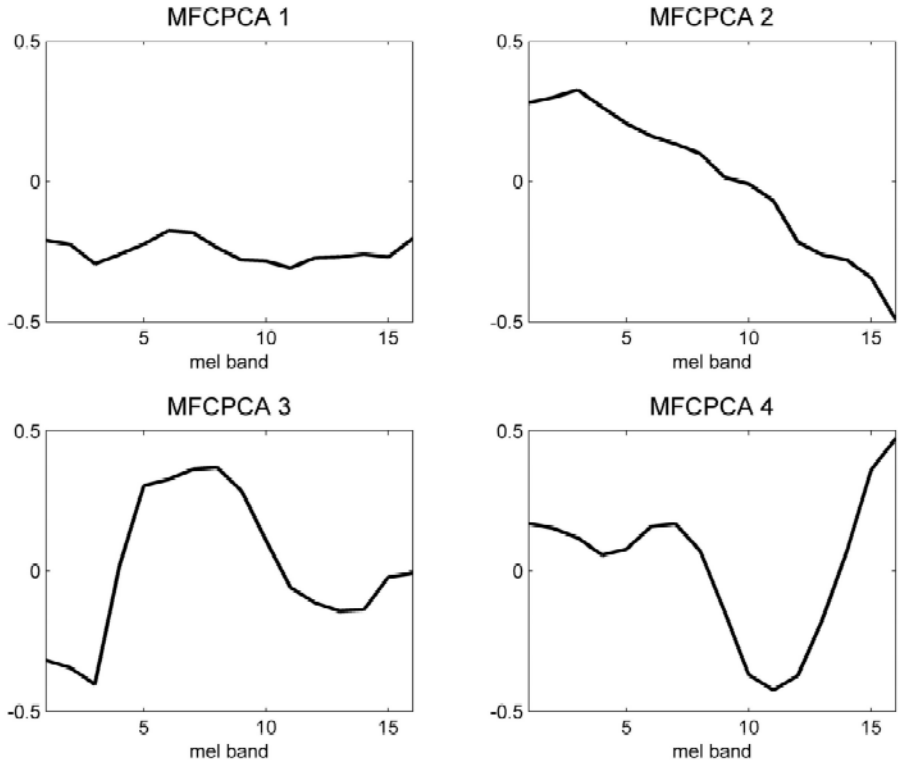

Figure 4.1: Principal components of MFC feature vectors.

DOI:10.15774/PPKE.ITK.2010.007

Table 4.1: Importance rate (variance) of the MFCPCA.

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

Table 4.2: Importance rate (variance) of the FacePCA.

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

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.

Our calculation is not able to explain the value of additional information of shifted signal compared to the 0 shifting value. If it has any additional information it is not subtracted. So the curves do not indicate the need of the extension of the time scope for every non-zero value. Rather the shape of the curve and the shift value of the maximum have a specific meaning.

In the new coordinate system generated by the principal component analysis the coordinates can be characterized by the importance rate. The importance rate can express that in the given direction which portion of the variance has been produced in the original space. The importance rate values in the case of MFCPCA transformation are shown in Table 4.1.

The importance rate values in the case of FacePCA transformation are shown in Table 4.2.

Combining the two tables by multiplication of the two vectors, a common importance estimation can be calculated. The values express the contribution of parameter pairs to the whole multidimensional data.

The really important curves are the combinations of the 1-4 principal components. Their general importance is expressed by the darkness of the curves. 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 have been repeated at the end of the records also. The possible synchrony error is below one video frame (40ms).

DOI:10.15774/PPKE.ITK.2010.007

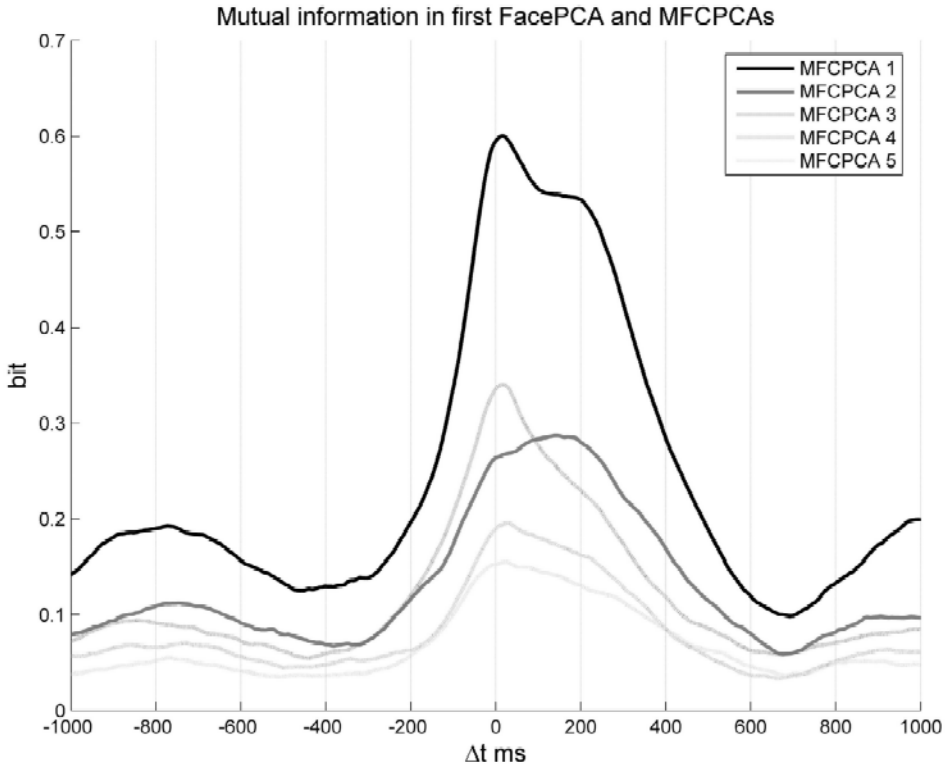

Figure 4.2: Shifted 1. FacePCA and MFCPCA mutual information. Positive Δt means future voice. Darkness show importance.

4.1.2 Results and conclusions

The mutual information curves were calculated and plotted for every possible PCA parameter pair in the range of -1000 to 1000 ms time shift. Only the most important curves are presented below to show the relation of the components having highest eigenvalues. The earlier movement of the lips and the mouth have been observed in cases of coarticulation and at the beginning of words. This delay has been considered as a specific and negligible effect. The delay value has been estimated only. 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.

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 excitation signal comes

DOI:10.15774/PPKE.ITK.2010.007

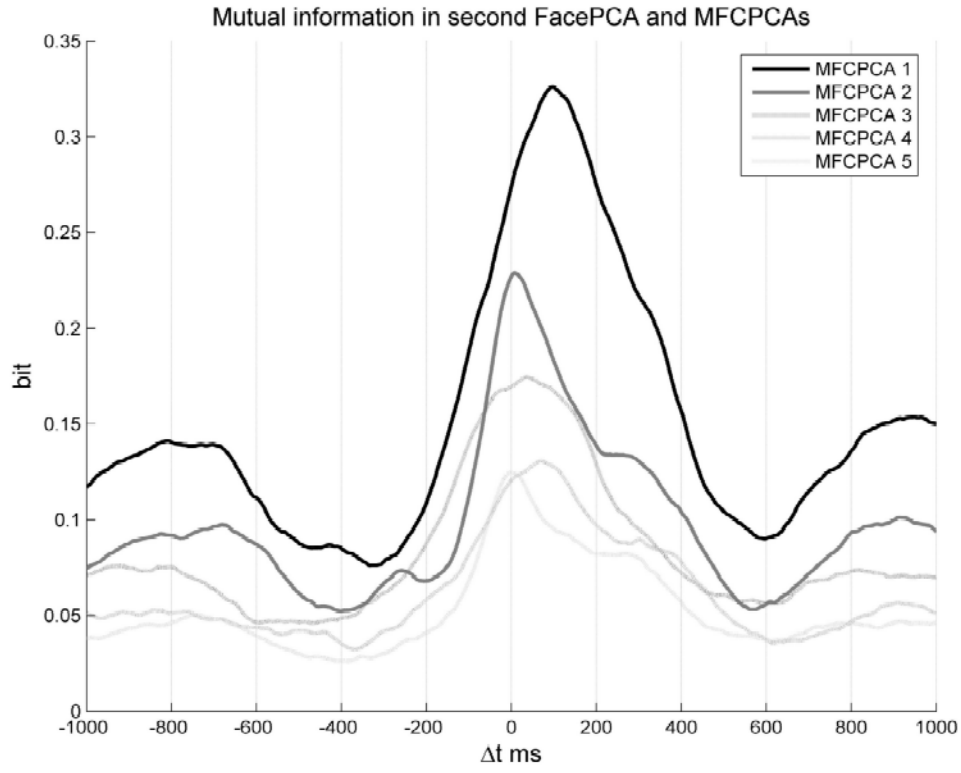

Figure 4.3: Shifted 2. FacePCA and MFCPCA mutual information. Positive Δt means future voice. Darkness show importance.

DOI:10.15774/PPKE.ITK.2010.007

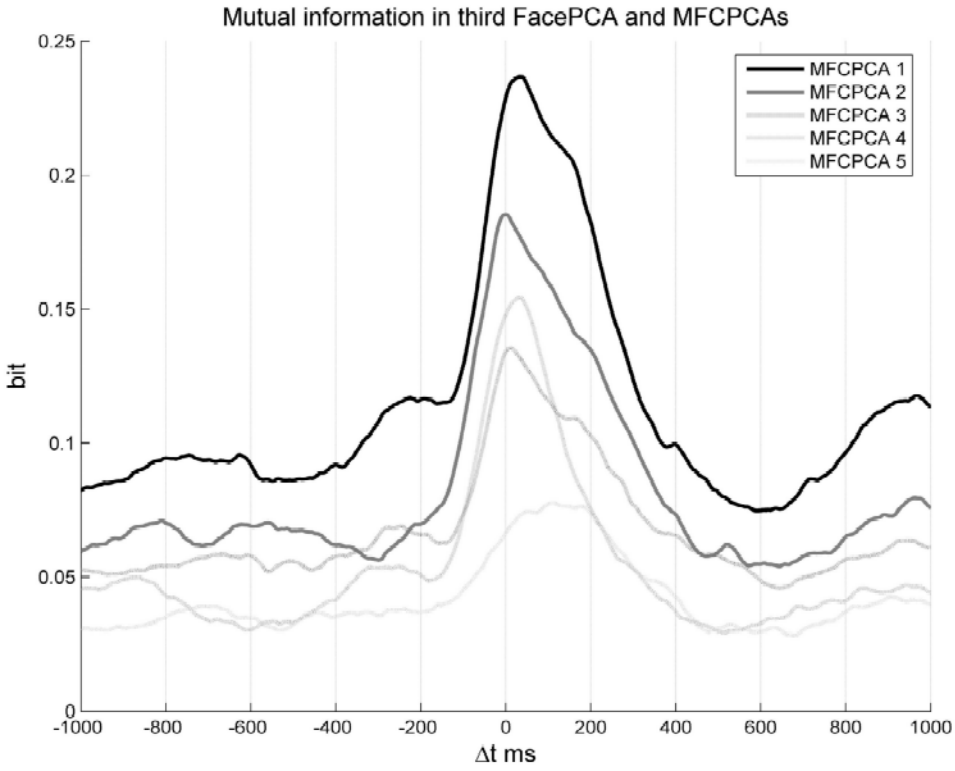

Figure 4.4: Shifted 3. FacePCA and MFCPCA mutual information. Positive Δt means future voice. Darkness show importance.

DOI:10.15774/PPKE.ITK.2010.007

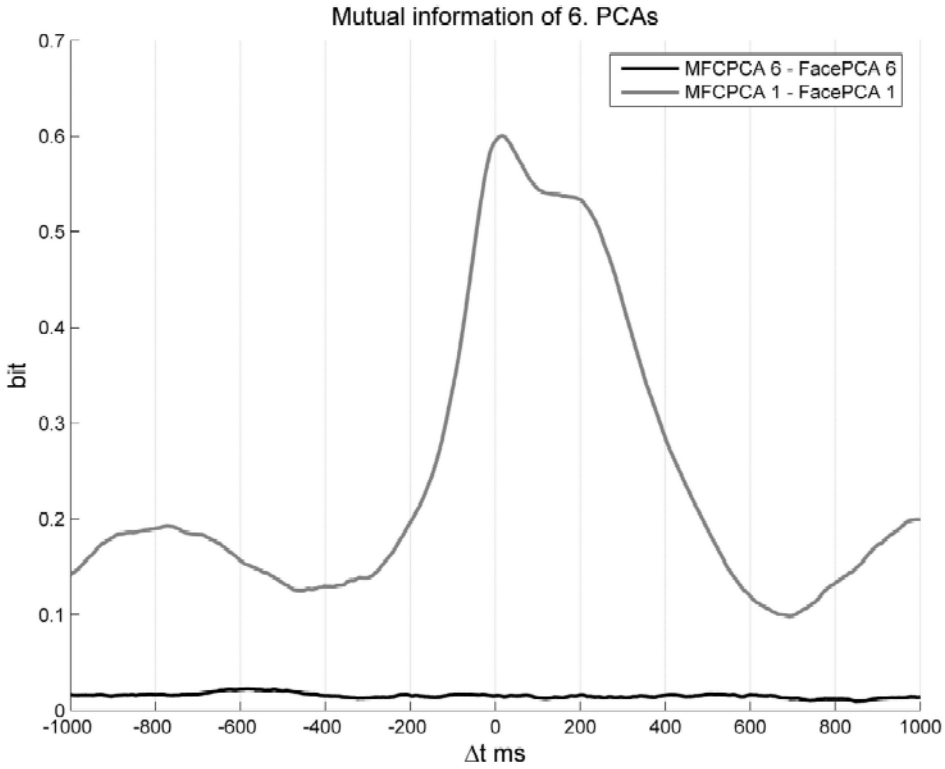

Figure 4.5: n-th principal component (black) does not contain substantive information compared to the main components (gray)

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 0.

On the Fig 4.3 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.

The FacePCA3 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 so there is no new message on the figure.

The pronounced word was "'September"'. In figure 4.6 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

DOI:10.15774/PPKE.ITK.2010.007

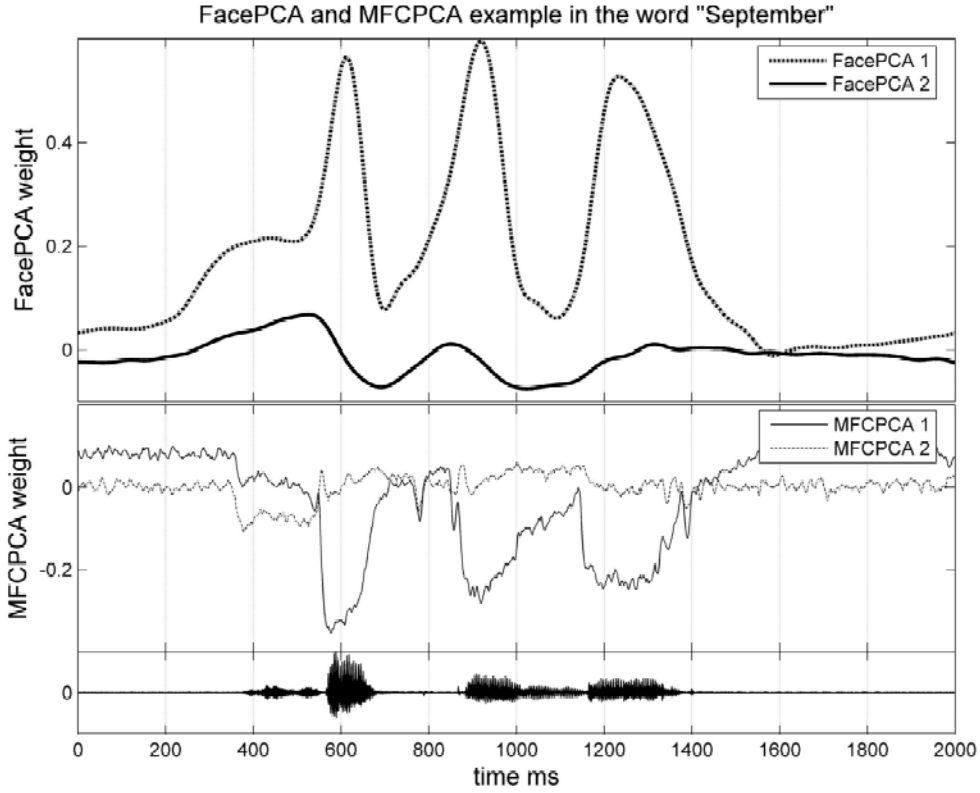

Figure 4.6: The word "September" as an example of time shifted visual components compared to audio components.

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.

Fig 4.6 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 as 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.

4.1.3 Multichannel Mutual Information estimation

The details above investigated the connection between specific facial motions and specific spectral content of audio speech. The use of PCA is convenient if we want to

DOI:10.15774/PPKE.ITK.2010.007

identify the actual motions and actual phonetic content, but it is not suitable for summing up the results because of the interchannel mutual information within the same modality.

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 minimized.

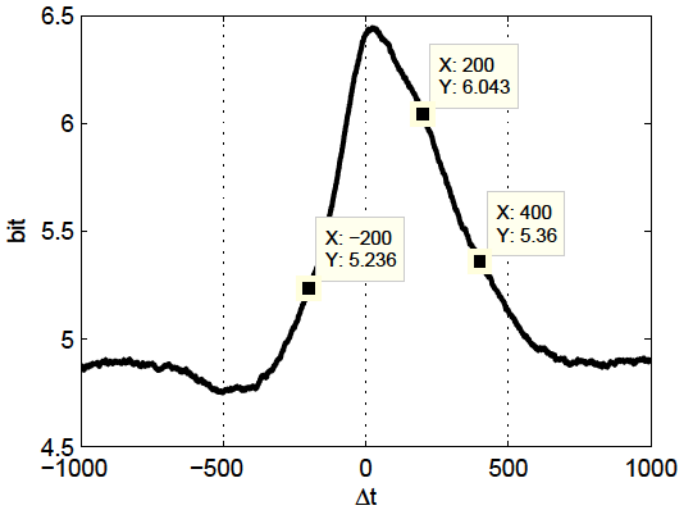

Figure 4.7: Sum of $MI(\Delta t)$ results of all channel audio-video pairs (6 x 6 : 15 pairs). Positive Δt means voice in the future was measured to the video frame in $\Delta t = 0$. The unit of time is millisecond.

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.

Figure 4.7 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 result shows that the visible speech organs are preparing for the next phoneme during the visual coarticulation while the speech audio signal is not changing. If both modalities would be changing together, there would be no asymmetry in mutual infor-

DOI:10.15774/PPKE.ITK.2010.007

mation.

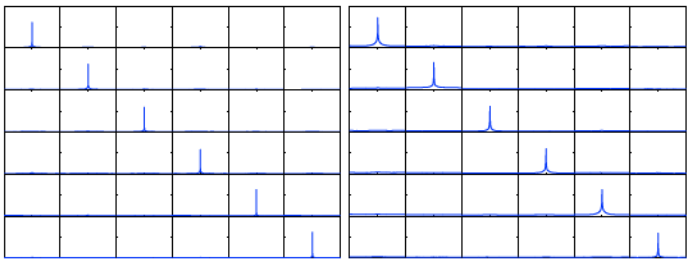

Figure 4.8: Interchannel $MI(\Delta t)$ results show independence in audio (left) and video (right) channels. The scaling of the curves are Δt = -1..1 in seconds on x axis, and 0..10 range in bits on y axis

In Figure 4.8 the independence of the channels can be seen. 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 4.7, 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. The measurement was repeated later with faster camera.

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 4.9. This confirms in practice the mutual information measurement.

4.1.4 Duration of asymmetry

The phenomena was tested with multiple speech tempos. Normal and slow speech tempo show the effect, but in the case of fast speech, there is no temporal asymmetry since there are no shoulders on the slopes of the mutual information curve, because there is no elevation in the value of mutual information. It seems that the existence of the mutual information is related to the clear phases of the speech.

DOI:10.15774/PPKE.ITK.2010.007

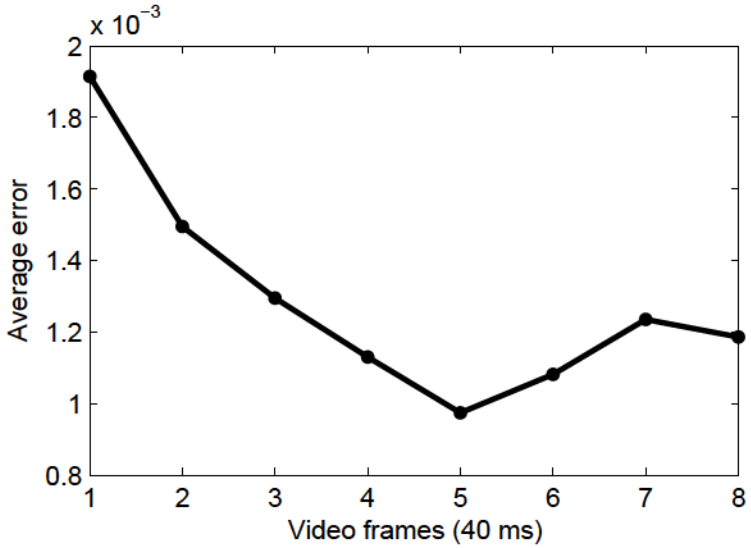

Figure 4.9: Training errors of different temporal scopes. The error is given in neural networks training data which is normalized to [-1..1] interval.

4.2 Thesis

> ***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.*** **[38]**

4.2.1 Novelty

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)

4.2.2 Measurements

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

DOI:10.15774/PPKE.ITK.2010.007

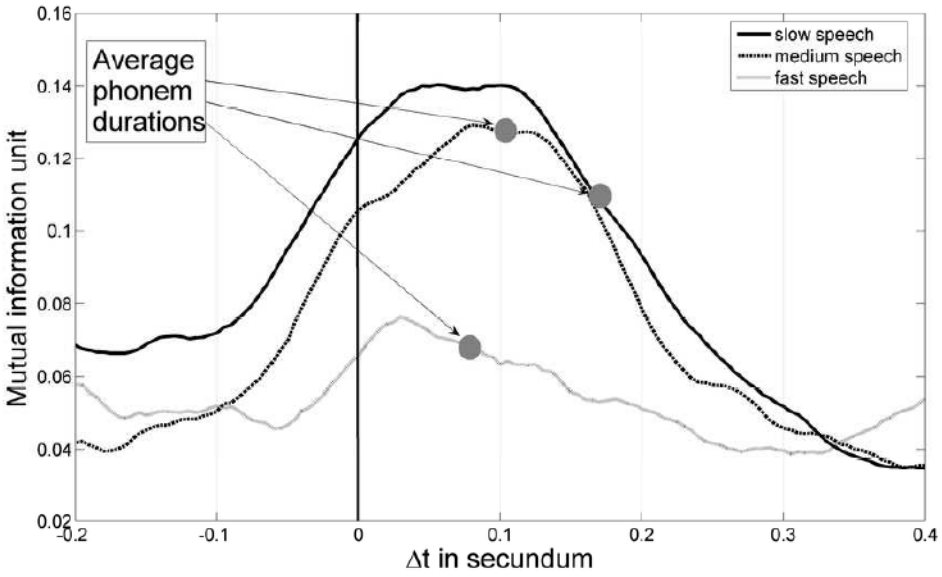

Figure 4.10: Mutual information in multiple speech tempos. The dots show the average phoneme durations. It seems that the mutual information is high in this duration between future audio speech and actual visual speech.

channels. If one would just add the results of the estimations of the channels without ICA or equivalent filter, the sum would contain the same mutual information multiple times because of the interchannel mutual information in the data. By using ICA this redundancy is lowered, depending on the quality of the ICA algorithm used.

Mutual information was calculated by two dimensional distribution histogram. The resolution of the histogram was 200x200. Low resolution give no usable result, high resolution needs more data than we had, so we convolved the histogram with a two dimensional gauss window.

Time shifts were tested in +-1000ms interval.

4.2.3 Limits of validity

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.

4.2.4 Consequences

The main consequence of the phenomena is that the best possible ATVS system should have 200ms theoretical latency to wait up the future of the audio speech to synthesize the video data with the most extractable information.

DOI:10.15774/PPKE.ITK.2010.007

This phenomena can be useful also in multimodal speech recognition, using the video data to pre-filter the possibilities in the audio representation.

DOI:10.15774/PPKE.ITK.2010.007

DOI:10.15774/PPKE.ITK.2010.007

Chapter 5

Speaker independence in direct conversion

This chapter is about handling speaker dependence of direct learning approach.

5.1 Method

In this chapter a speaker independent training method is presented for direct ATVS systems. An audiovisual database with multiple voices and only one speaker's video information was created using dynamic time warping. The video information is aligned to more speakers' voice. The fit is measured with subjective and objective tests. Suitability of implementations on mobile devices is discussed.

5.1.1 Introduction

The direct ATVS need an audiovisual database which contains audio and video data of a speaking face[44]. 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 (see Fig 5.1). But our task is to calculate only one, but lip-readable face. Training on multiple speaker's voices and faces results 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.

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

DOI:10.15774/PPKE.ITK.2010.007

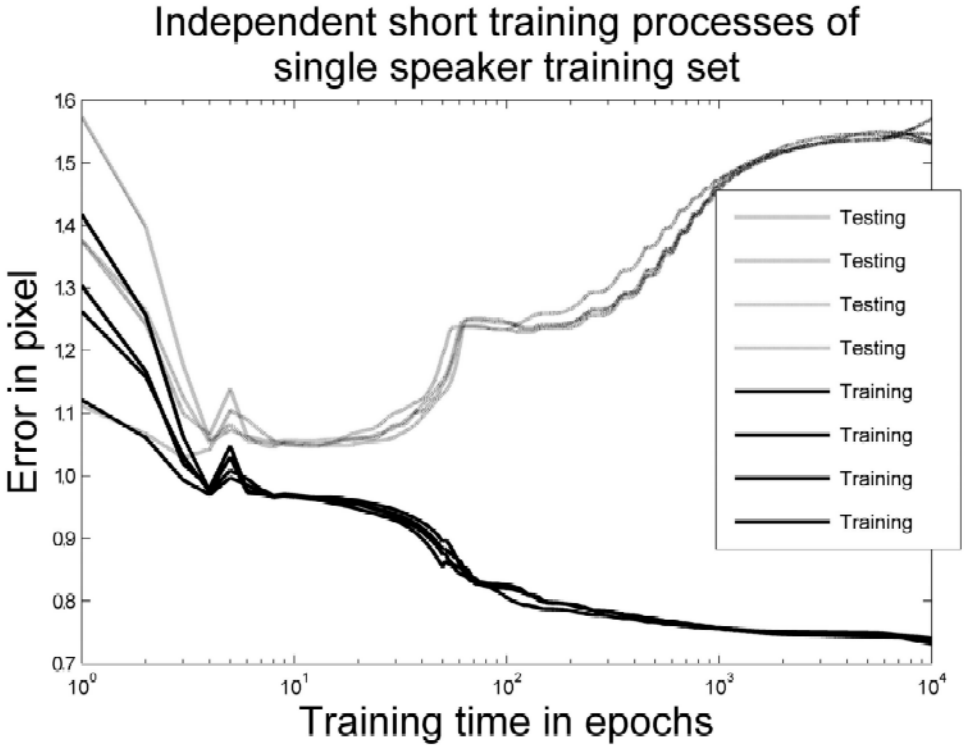

Figure 5.1: Overtraining: the network learns training set dependent details. The train and test runs were independent.

DOI:10.15774/PPKE.ITK.2010.007

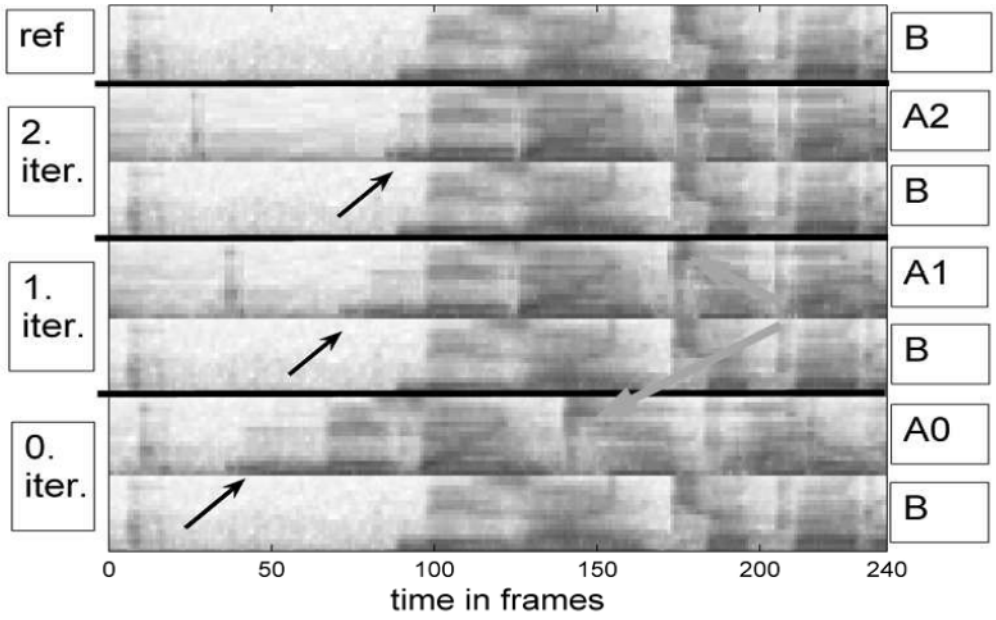

Figure 5.2: Iterations of alignment. Note that there are features which need more than one iteration of alignment.

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 will use a method based on Dynamic Time Warping (DTW)[45] 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". On 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 5.2.

5.1.2 Speaker independence

The described base system works on well defined pairs of audio and video data. This is evident if the database is a single person database. If the video data belongs to a

DOI:10.15774/PPKE.ITK.2010.007

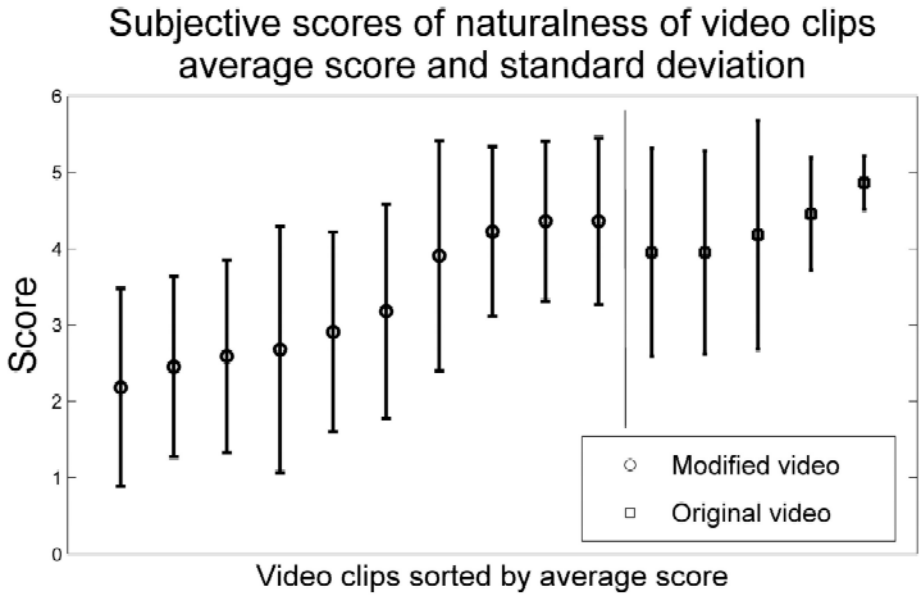

Figure 5.3: Mean value and standard deviation of scores of test videos.

different person, the task is to fit the audio and the video data together. The text of the database was the same for each person. This allows the aligning of audio data between speakers.

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

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. 5.3. the deviations are overlapping each other, there

DOI:10.15774/PPKE.ITK.2010.007

Figure 5.4: Training with speaker A, A and B, and so on, and always test by speaker E which is not involved in the training set.

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.

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 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. 5.4

5.1.3 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

DOI:10.15774/PPKE.ITK.2010.007

the system with only voice to broaden the cover of voice characteristics. Most of the calculations of direct ATVS are cheap enough to implement the system on mobile devices. The speaker independence induces no plus expense on the client side.

5.2 Thesis

> ***III. I developed a time warping based AV synchronizing method to create training samples for direct AV mapping. I showed that 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.***[46]

5.2.1 Novelty

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.

5.2.2 Measurements

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.

5.2.3 Limits of validity

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.

5.2.4 Consequences

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

DOI:10.15774/PPKE.ITK.2010.007

Chapter 6

Visual speech in audio transmitting telepresence applications

Supporting an ATVS system with head models requires information on the representation of the ATVS system. In earlier results 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.

For audio preprocessing we used MFCC. In some applications there are other audio preprocessing included, in the case of audio transmission mostly Speex[47].

I will describe a method of easy enhancement of audio transmitting telepresence applications using it's internal Speex preprocessor and producing results which is capable to render visual speech from viseme states.

6.1 Method

In on-line cyber spaces there are artificial bodies which imitate realistic behavior controlled by remote users. An important aspect is the realistic facial motion of human like characters according to the actual speech sounds. This chapter describes a memory and CPU efficient method for visual speech synthesis for on-line applications using voice connection over network. The method is real-time, can be activated on the receiver client without server support. It is needed only to send coded speech signal and the visual speech synthesis is the task of the receiving client. The animation rendering is supported by graphical accelerator devices, so CPU load of the conversion is insignificant.

DOI:10.15774/PPKE.ITK.2010.007

6.1.1 Introduction

Voice driven visual speech synthesis has a growing popularity in cyber telepresence applications. As of 2009 there are more video games on the market with the benefits of this technology.

The most popular use of visual speech synthesis is the real-time rendered pre-calculated facial animation. This meets all the requirements in an artificial world where the content of the voices is given by the designers, it is recorded with voice actors, and there is time to do all the calculations during production time. An example of this technology is in the title Oblivion or Fallout 3 from Bethesda Softworks[48] which uses the MPEG-4 based FaceGen[13]. However this approach is extendable to real-time applications as well by concatenative synthesis, we will see that it is not a really suitable solution.

In a real-time telepresence application the player activates the transmission, the client side records the voice in small chunks, and send it to the server which forwards it to the given subset of the players, teammates or any characters nearby. During active voice transmission the visual feedback on the receiver client side is some visual effect of the character, like an icon, a light effect, a basic or random facial motion. An example of basic facial motion is in the Counter-Strike Source[49], where the momentary voice energy is visualized by the movement of the jaw. We will use this approach as baseline. Our solution is a replacement of this with improved quality, allowing even lipreading.

6.1.2 Overview

Real-time or pre-calculated motion control

In case of production time methods all of the audio content is available in advance. A typical example of this starts from screenplay, and the voice records are based on the given text. There are solutions to extract phoneme string from text, and to synchronize this phoneme string to the records like Magpie[50] for example. Voice synchronized phoneme strings can be used to create viseme string with visual co-articulation. The viseme is the basic unit of visual speech (Fig 6.2), practically the visual consequence of pronouncation of a phoneme. The viseme string with timing includes the visual information, and co-articulation methods has to form it into a natural visual flow. Viseme combinations were mapped for interactions as domination or modifying, and with this knowledge, viseme pairs or longer subsequences are used for the synthesis.

Also, during production time the speech signal is available as a whole sentence. This makes those methods usable which uses data for a given frame from the voice of next frames. This information definitely important for precise facial motion[38, 6]. Real-time methods are not allowed to use long buffers because of the disturbing delay.

One of the real-time approaches uses automatic speech recognition (ASR) system to extract phoneme string from the voice[6]. The benefit of this approach is the compatibility with viseme string concatenator methods by simply use ASR instead of manually extracted annotated phoneme string information. The ASR system can be trained on

DOI:10.15774/PPKE.ITK.2010.007

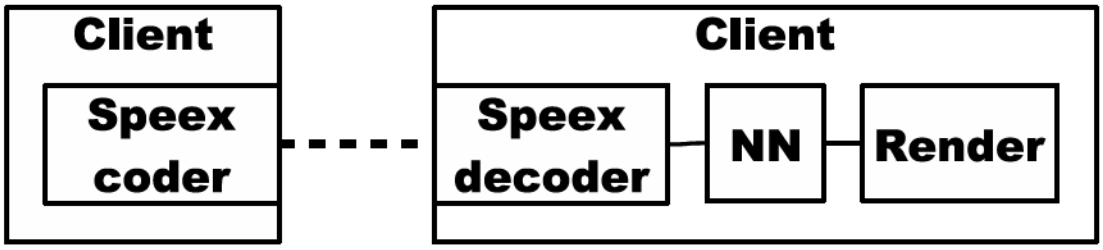

Figure 6.1: Our real-time method: the visual speech animation parameters are calculated from the Speex coding parameters with neural network.

usual speech databases without visual data. The drawback is the time and space complexity of the recognition, and the propagation of the recognition errors, because of the falsely categorized phonemes or words.

Other way is the direct conversion which is simpler and faster but usually less accurate because of the lack of language dependent information.

6.1.3 Face model

For a speaking head model there are two requirements: the artists should design the model easily, and it should have enough degree of freedom. For example in the game Counter-Strike Source the mouth motion has one degree of freedom, the position of the jaw, and it is directly linked to the energy of the signal. Although this behaves obviously artificial, this is numerically a fair approximation since visual speech PCA (Principal Component Analysis) factorization shows that the 90% of the deviation is in the first principal component which mainly shows the motion of the jaws[8]. In order to have a more sophisticated head model there should be more degree of freedom, which includes the horizontal motion of the mouth boundary or more.

In our works we used PCA based facial coordinates to represent a facial state. This representation have some nice properties as mathematically proven maximum compression rate along linear basises in dimension count. Each state is expressed in an optimal basis calculated from visual speech database. In this way a 30 dimensional facial data can be compressed into 6 dimension with only 2% error. As we visualize the calculated basis, the coordinates show motion components as jaw motion, liprounding, and so on. These are not visemes as visemes are not guaranteed to be orthogonals of each other. We used guided PCA in this case, including the most important visemes as long as it is possible.

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.

DOI:10.15774/PPKE.ITK.2010.007

Therefore we need to give a facial animation control based on face states, and the designer can work with examples. The control of the facial animation can be the weights of the drawn shapes. Generally it is not true that every facial state can be expressed from any set of visemes, but there is an approximation of the states for a given viseme set, and depending on the size of this set, and so the degree of freedom, any level of accuracy can be reached (Table 6.1). This approach may use more degrees of freedom than PCA based approach for the same quality, since the PCA is optimal in this point of view.

The rendering of the face is efficient. The graphical interfaces usually provide hardware accelerated vertex blending. There are more sophisticated approaches using volume conserver transformations [51] with slight increase of time complexity. These methods can be used to render our approach as well. Support for features like crinkling skin is out of our interest.

6.1.4 Viseme based decomposition

The video data is from a video recording of a talking person with fixed field of view. The head of the person was fixed to the chair to eliminate the motion of the whole head. The face of the person was prepared with marker points which were tracked automatically and corrected manually[8]. The position of the nose was used as origin, so every frame was translated to common frame. The automatic tracking was based on color sensitive highlight tracking with automatic quality feedback to help the manual corrections. There were 15 markers, placed on a subset of MPEG-4 feature points. The marker tracker results a vector stream in 2D pixel space. This representation is good for measurement, because no special equipment is needed for the recording, and also relatively good for estimation of quality since the generated animation will give the same data in best scenario. To achieve interchangeable metrics, pixel unit must be eliminated. This elimination is done by using the distribution of the given marker position as a reference to the position error. In this case the pixel units are eliminated from the result by transforming to a relative scale.

$$E(G) = \frac{\sum_{i=1}^{N} \sum_{j=1}^{f} \left| \vec{G}_j^i - \vec{S}_j^i \right|}{N f \sigma(\vec{S}^i)} \tag{6.1}$$

where N is the dimensionality of the visual representation, f is the total number of frames, G is the generated signal versus S signal and E is the estimated error value. S can be any linear representation of the facial state, given in pixels or vertices, or MPEG-4 FAP values.

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 (Equation 6.1). The

DOI:10.15774/PPKE.ITK.2010.007

Table 6.1: Errors as a function of viseme number used in compositions on important feature points. Visemes are written in corresponding phoneme codes, except "closed" and ʃ which is not a standalone dominant viseme, it is used at the beginning of the word.

No.	Visemes	Error
2	*closed* ʌ	15.25%
3	*closed* ʌ ɪ	7.48%
4	*closed* ʌ ɪ ɔ	4.17%
5	*closed* ʌ ɪ ɔ ɛ	3.88%
6	*closed* ʌ ɪ ɔ ɛ ʃ*	3.48%

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.

$$\vec{G} = \sum_{i=1}^{N} w_i \vec{V_i} \tag{6.2}$$

where

$$\sum_{i=1}^{N} w_i = 1 \tag{6.3}$$

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 convexness guarantees that the blending is independent of the coordinate system. If the designer use unnormalized vertex coordinates, a weighted sum with more or less of weight sum of one can result translation and magnification of the head.

The results of this simple approximation are acceptable. The quality is estimated by pixel errors of the important facial feature points. The selection of important points is based on deviation, those feature points which are above the average deviation are chosen.

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.

Note that the representation quality is scalable by setting the viseme count. This will make the resulting method scalable on client side.

DOI:10.15774/PPKE.ITK.2010.007

Figure 6.2: Visemes are the basic unit of visual speech. These are those visemes we used for subjective opinion score tests in this (row-major) order.

DOI:10.15774/PPKE.ITK.2010.007

6.1.5 Voice representation

Every voice processing method need to extract useful information from the signal. Those algorithms which use directly the sound pressure signal are called time domain, which uses Fourier transform or other frequency related filter banks are called frequency domain, and those which uses some (lossy) compressed input are called compressed domain methods.

Those applications, where voice driven facial animation can be a matter, use voice transmission. Voice transmission systems use lossy compression methods to minimize the network load. Therefore an efficient voice driven visual speech synthesizer should be a compressed domain method.

A speech coder attends to achieve best voice quality with reasonable sized data packets. This can be treated as a feature extracting method. The question is, what distance function can be used on the given representation? Is there an appropriate metrics what a learning system can approximate?

6.1.6 Speex coding

One of the most popular speech coder for this purpose is the Speex[47]. Speex uses LSP, a member of the linear prediction coding family. Linear prediction use a vector of scalars which can predict the next sample from the previous samples by linear combination.

$$x'_n = \sum_{i=1}^{N} a_i x_{n-i} \tag{6.4}$$

Where N is the size of the prediction vector. The optimal predictor coefficient vector for a given x can be calculated by Levinson-Durbin algorithm. This is short representation, but it is not suitable for quantization or linear operations as linear interpolation, consequently it is not directly used for voice transmission or facial animation conversion. Hence Speex uses LSP which is a special representation of the same information but capable to linear operations, for example the LSP values are linearly interpolated between the compressed frames of Speex.

For LSP coding, instead of storing the predictor vector a we treat it as a polynomial, and store the roots. The roots are guaranteed to be inside the unit circle of the complex plane. To find roots, two dimensional search would be needed, so to avoid this we use a pair of polynomials which are guaranteed to have all the roots on the unit circle, and the mean of the pair is the original root, so one dimensional search is enough.

$$PQ_z = \left\{ a_z \pm z^{-(N+1)} a_{z^{-1}} \right\} \tag{6.5}$$

$$LSP = \bigcup_{z \in \mathbb{C}} \{PQ_z = 0\} \tag{6.6}$$

DOI:10.15774/PPKE.ITK.2010.007

This makes LSP more robust to quantization and interpolation than the predictor vector. Interestingly, PQ values are called vocal tract, with glottis open and closed, which are connected with the topic of audiovisual speech synthesis.

Lossy compression methods use quantization of values of a carefully chosen representation. LSP is a compact and robust representation, and Speex use Vector Quantization to compress these values. We modified the Speex decoding process to export uncompressed LSP values and the energy. This makes only 11 assignments and multiplications for scaling as an extra computational cost.

6.1.7 Neural network training

The data is from an audiovisual recording of a professional lip-speaker. The recording contains 4250 frames. The content is intended for direct voice to visual speech conversion testing for deaf people, it contains numbers, months, etc. The language is Hungarian. The network is a simple straightforward error-backpropagation network with one hidden layer.

Audio

The audio recording is originally 48kHz, and it is downsampled to 8kHz for Speex. We used the modified Speex decoder to extract LSP and gain values to train neural networks as input. There are values for each 20 ms window. LSP has values in $[0, \pi]$, and the neural network use the $[-1, 1]$ interval, so scaling was applied.

Video

The target of the neural network is the viseme weight vector representing facial state. As the original recording is 25 frame per second, and the audio data from Speex uses 20 ms windows, the video data was interpolated from 40 ms to 20 ms frame interval. We used linear interpolation as it not violates convexness. The decomposition weight values are in the range of $[0, 1]$ which is in the neural networks $[-1, 1]$ interval, so no scaling was applied.

Neural network usage

The resulting network is intended to be used directly in the host application. The trained network weights can be exported as a static function of a programming language, for example C++. This source code can be compiled into the client. This function is called with the values exported from the modified Speex codec. The returning values is applied directly for the renderer. With this approach runtime overhead is minimal, no file readings or data structures are needed. The generated source code can be created at speech interested laboratories, the application developers just use the code.

DOI:10.15774/PPKE.ITK.2010.007

6.1.8 Implementation issues

The method can be implemented as a feature on the client, on receiver side. The user may turn on and off the method since the calculation are performed on the receiver clients CPU. There is no extra payload on the network traffic.

The CPU cost of the calculation is 200-400 multiplications depending on the hidden layer size and the degrees of freedom. The space cost of the feature is the multiple shapes of the head models, which depends on the given application, how sophisticated head models are used in it. The space cost is scalable by setting the viseme set, the more head models the better approximation of the real mouth motion.

The head models can be stored on video accelerator device memory and can be manipulated through graphical interfaces as OpenGL or Direct3D. The vertex blending (weighted vector sum) can be calculated on the accelerator device, it is highly parallel since the vertices are independent.

6.1.9 Results

Training and testing set was separated, and during the first 1‘000‘000 epochs (training cycles) of training the error of the testing set still decreased (Fig 6.3). Depending on the degrees of freedom the results are 1-1.5% of average error. Our former measurements gave sufficient intelligibility results at this level of numeric error. This shows that usable training error level can be reached before overtraining even with relatively small databases.

The details of the trained system response can be seen on Fig 6.4. The main motion flow is reproduced, and there are small glitches 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. The opinion score test included from 1 to 5 degrees of freedom of control parameters. Each control source and degrees of freedom combination was represented in 8 short video, 2 of them pronounced numbers 0-9, 2 of them numbers 10-99, 2 with names of the months and 2 with the days of the week. This makes 80 videos.

Test subjects were instructed to evaluate harmony and naturalness of the connection of visual and audio channels. Score 5 for perfect articulation, score 3 for mistakes and score 1 for hardly recognizable connection. The results are interesting since after the second degree of freedom the evaluation is near to constant while numerical error halves between 2. and 3. degree. The possible explanation of this phenomena can be the simpleness of our head model used for scoring. The tongue and the teeth was not independently moved in the videos, the more degree of freedom was used only to approximate the mouth contour more precisely which may was precise enough already at lower degrees of freedom.

DOI:10.15774/PPKE.ITK.2010.007

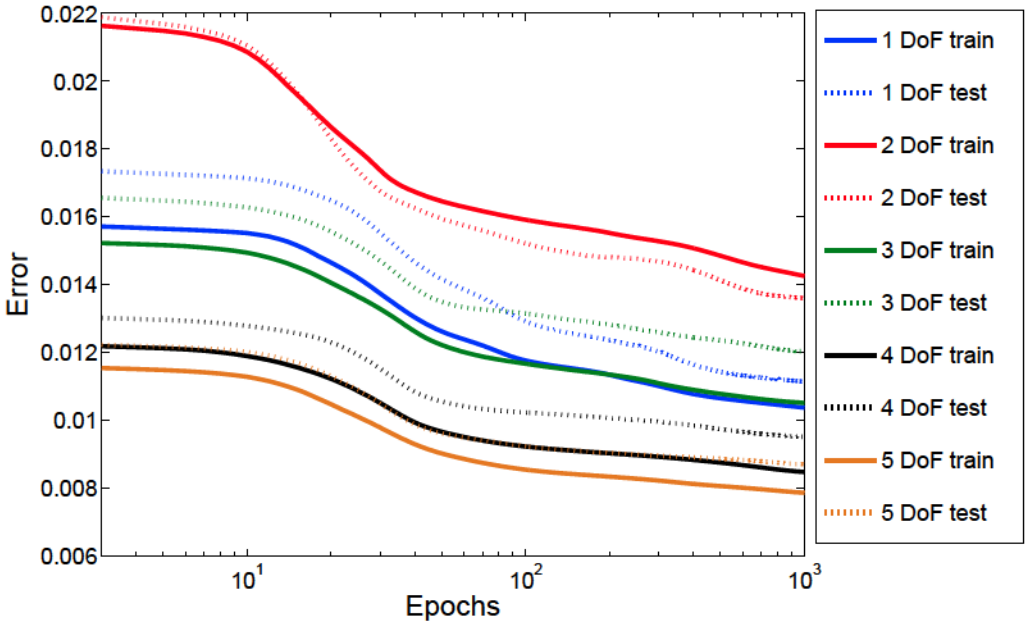

Figure 6.3: Training and testing error of the network during training. The error is the average distance between the weight from the video data and the calculated from Speex input.

DOI:10.15774/PPKE.ITK.2010.007

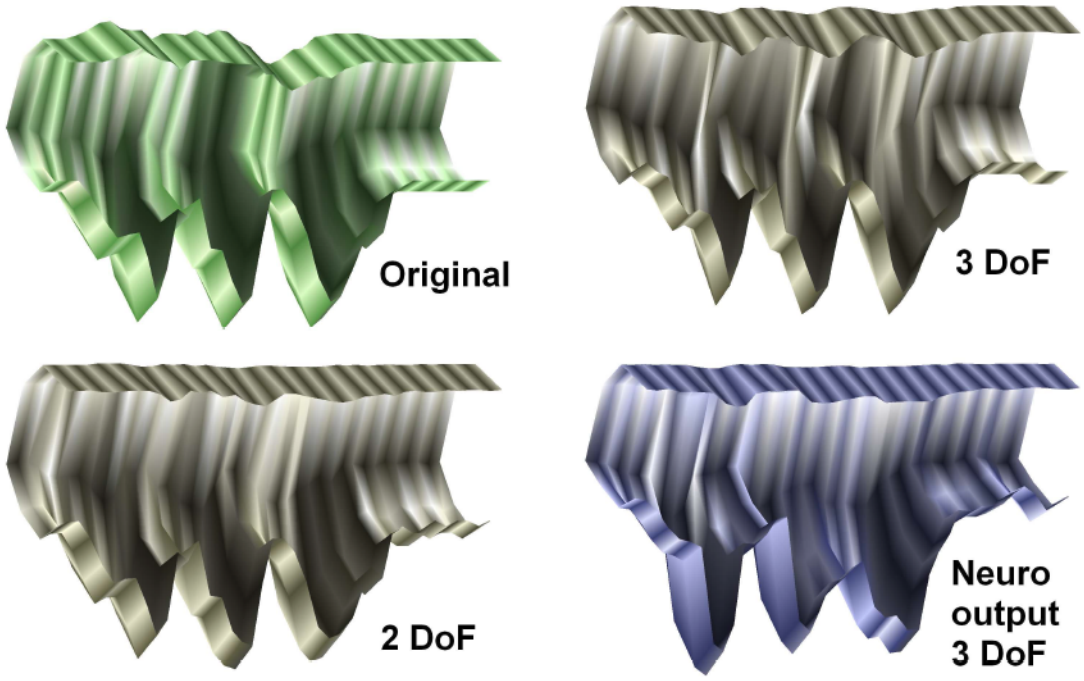

Figure 6.4: Examples with the hungarian word "Szeptember", it's very close to English "September" except the last e is also open. Each figure is the mouth contour in the time. The original data is from a video frame sequence. The 2 and 3 DoF are the result of decomposition. The last picture is the voice driven synthesis.

DOI:10.15774/PPKE.ITK.2010.007

Higher target complexity induce the neural network converge slower, or not at all. But as we increase the degrees of freedom, the neural network's error decreases from the 2. degree.

The results of the opinion score test show that the best score/DoF rate is at the 2 DoF (Fig 6.5), 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. As the decomposition is based entirely and only on mouth contour, it may be not that important. Using correct teeth visibility or tongue movement may improve the results, but in this test we were unable to try this because of the lack of markers on these facial organs. This problem is in the decomposition phase since in the synthesized face we have these facial organs and control them actively, but the control is inaccurate. If the decomposition would be affected by more information, this could be corrected. Active shape modeling or other advanced techniques may improve the decomposition material.

The main consequence of the subjective test that two degrees of freedom can give sufficient quality for audiovisual speech, and the proposed method can give the control parameters in this quality from the voice signal.

6.1.10 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.

6.2 Thesis

> ***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.*** [52]

DOI:10.15774/PPKE.ITK.2010.007

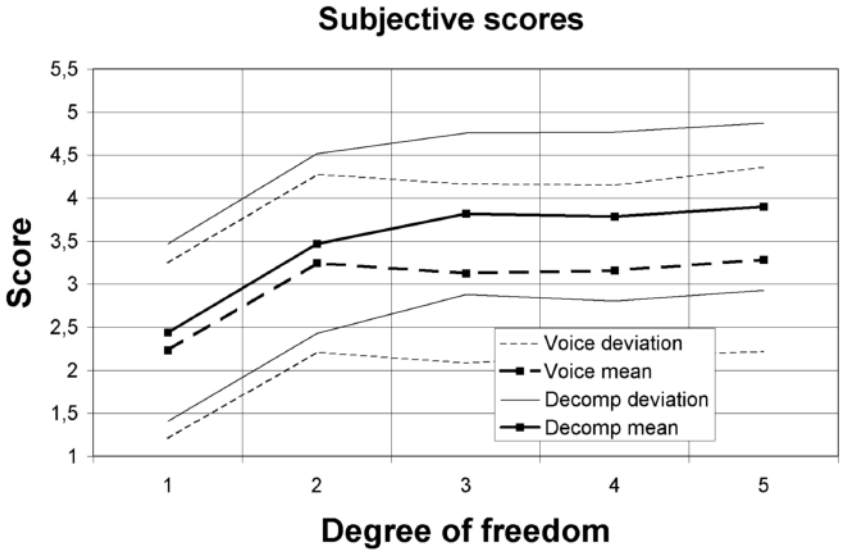

Figure 6.5: Subjective scores of the decomposition motion control and the output of the neural network. There is a significant improvement by introducing a second degree of freedom. The method's judgment follows the database's according to the complexity of the given degree of freedom.

DOI:10.15774/PPKE.ITK.2010.007

6.2.1 Novelty

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.

6.2.2 Measurements

Subjective opinion scores were used to measure the resulting quality.

6.2.3 Consequences

Using the Speex and the viseme combination representation the resulting system is embeddable very easily.

DOI:10.15774/PPKE.ITK.2010.007

Summary

In this proposed thesis I collected my contributions to the field of audio speech conversion to visual speech, especially direct conversion between the modalities.

New scientific results

I positioned direct conversion method among widely used ASR based solutions.

> ***I. I showed that 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.*** [39]

I discovered and measured the phenomena of temporal asymmetry on productional side of the speech process.

> ***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.*** [38]

I solved the problem of speaker dependency of direct conversion of speech.

> ***III. I developed a time warping based AV synchronizing method to create training samples for direct AV mapping. I showed that 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.***[46]

I showed that direct conversion can be used with natural representations instead of mathematically convenient principal components.

> ***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.*** [52]

DOI:10.15774/PPKE.ITK.2010.007

The conclusion of my work is that the direct conversion is an undeservedly ignored method in the world because of the initial failures with inadequate training data. My results clearly show that direct conversion is not only computationally efficient but contributes speaker independent natural visual speech solution for broad range of applications, and the key of the good quality visual speech synthesis is the appropriate database, .

DOI:10.15774/PPKE.ITK.2010.007

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 (submitted)

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.

DOI:10.15774/PPKE.ITK.2010.007

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.

DOI:10.15774/PPKE.ITK.2010.007

Bibliography

[1] K-E Spens B. Granström, I. Karlsson. Synface - a project presentation. *Proc of Fonetik - TMH-QPSR*, 44:93–96, 2002. 1, 1.2.1, 2.5.1

[2] 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

[3] 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

[4] 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, 1.2.6, 1.5.2

[5] O. Garcia P. Kakumanu, A. Esposito and R. Guiterrez-Osuna. A comparison of acoustic coding models for speech-driven facial animation. *Speech Communication*, 48:598–615, 2005. 1.1.1

[6] 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, 6.1.2

[7] 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

[8] 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, 2.3.3, 2.3.5, 6.1.3, 6.1.4

[9] 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, 1.2.2

[10] A. Fusaro P. Cosi and G. Tisato. Lucia a new italian talking-head based on a modified cohen-massaros labial coarticulation model. In *Proceedings of Eurospeech 2003, Geneva, Switzerland, September 1-4, 2003*, volume Vol. III, pages 2269–2272, 2003. 1.2.3, 2.2

DOI:10.15774/PPKE.ITK.2010.007

[11] K. Madany and S. Fagel. Objective and perceptual evaluation of parameterizations of 3D motion captured speech data. In *Proceedings of AVSP*, 2008. 1.2.3

[12] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, volume 77. 1.2.4

[13] http://www.facegen.com. 1.2.5, 6.1.1

[14] M. Odisio and G. Bailly. Shape and appearance models of talking faces for model-based tracking. In *AVSP*, pages 105–110, St Jorioz, France, 2003. 1.2.5

[15] Hedvig Kjellström and Olov Engwall. Audiovisual-to-articulatory inversion. *Speech Communication*, 51(3):195–209, 2009. 1.5.1

[16] S. A. King and R. E. Parent. A 3d parametric tongue model for animated speech. *Journal Of Visualization And Computer Animation*, 12:107–116, 2001. 1.5.1

[17] http://www.genarts.com. 1.5.2

[18] P. Badin, G. Bailly, L. Revéret, M. Baciu, C. Segebarth, and C. Savariaux. Three-dimensional linear articulatory modeling of tongue, lips and face, based on mri and video images. *Journal of Phonetics*, 30(3):533 – 553, 2002. 1.5.3

[19] A. Riecker, H. Ackermann, D. Wildgruber, J. Meyer, G. Dogil, H. Haider, and W. Grodd. Articulatory/phonetic sequencing at the level of the anterior perisylvian cortex: A functional magnetic resonance imaging (fmri) study. *Brain and Language*, 75(2):259 – 276, 2000. 1.5.3

[20] R. Zunino D. Anguita, G. Parodi. An efficient implementation of BP on RISC-based workstations. *Neurocomputing*, 6(1):57–65, 1994. 2.3.4

[21] http://digitus.itk.ppke.hu/flugi/johnnie. 2.4

[22] M. M. Cohen and D. W. Massaro. Modeling coarticulation in synthetic visual speech. In *Models and Techniques in Computer Animation*, pages 139–156. Springer-Verlag, 1993. 2.5.1

[23] T. Kuratate and K. Kinoshita. Real-time talking head system based on principal component analysis. *Journal of the Institute of Image Electronics Engineers of Japan*, 34:336–343, 2005. 2.5.2

[24] T. Harczos-B. Srancsik G. Szabó T. Bárdi, G. Feldhoffer. Audiovizuális beszéd-adatbázis és alkalmazásai. *Híradástechnika*, (10), 2005. 2.5.2

[25] G. Jung-M. Hegedűs G. Feldhoffer, T. Bárdi. Mobiltelefon alkalmazások siket felhasználóknak. *Híradástechnika*, (10), 2005. 2.5.2

[26] T. Bárdi-G. Feldhoffer Gy. Takács, A. Tihanyi and B. Srancsik. Beszédjel átalakítása mozgó száj képévé siketek kommunikációjának segítésére. *Híradástechnika*, (3), 2006. 2.5.2

DOI:10.15774/PPKE.ITK.2010.007

[27] T. Öhman and G. Salvi. Using HMMs and ANNs for mapping acoustic to visual speech. *TMH-QPSR*, pages 45–50, 1999. 2.5.2, 3.4, 3.1.4

[28] B. Theobald, S. Fagel, F. Elsei, and G. Bailly. LIPS2008: Visual speech synthesis challenge. In *Proceedings of Interspeech*, pages 1875–1878, 2008. 3.1.1

[29] 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. 3.1.2

[30] Z. Tüske P. Mihajlik, Z. Tobler and G. Gordos. Evaluation and optimization of noise robust front-end technologies for the automatic recognition of hungarian telephone speech. In *Interspeech 2005 - Eurospeech: 9th European Conference on Speech Communication and Technology*, Lisboa, Portugal, 2005. 3.1.2

[31] http://alpha.tmit.bme.hu/ speech/hdbMRBA.php. 3.1.2

[32] T. Révész P. Mihajlik and T. Tatai. Phonetic transcription in automatic speech recognition. *Acta Linguistica Hungarica*, pages 407–425, 2003. 3.1.2

[33] F. Pereira M. Mohri and M. Riley. Weighted finite-state transducers in speech recognition. *Computer Speech and Language*, pages 69–88, 2002. 3.1.2

[34] L. Czap and J. Mátyás. Virtual speaker. *Híradástechnika Selected Papers*, Vol LX/6:2–5, 2005. 3.1.2, 3.1.5

[35] http://avisynth.org. 3.1.2

[36] A. S. House, C. Williams, M. H. L. Hecker, and K. D. Kryter. Psychoacoustic speech tests: A modified rhyme test. *The Journal of the Acoustical Society of America*, 35(11):1899–1899, 1963. 3.1.4

[37] G. Breton G. Bailly, O. Govokhina and F. Elisei. A trainable trajectory formation model TD-HMM parameterized for the lips 2008 challenge. In *Proceedings of Interspeech 2008*, Brisbane, Australia, Sep 2008. 3.1.5

[38] 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. 3.1.5, 4.2, 6.1.2, 6.2.3

[39] A. Tihanyi G. Feldhoffer and B. Oroszi. A comparative study of direct and asr based modular audio to visual speech systems (accepted). *Phonetician*, 2010. 3.2, 6.2.3

[40] A-Esposito O. N. Garcia A. Bojorquez J.L Castillo R. Gutierrez-Osuna, P.K. Kakumanu and I. Rudomin. Speech-driven facial animation with realistic dynamics. *IEEE Transactions on Multimedia*, 7. 4.1.1

DOI:10.15774/PPKE.ITK.2010.007

[41] O. N. Garcia R. Gutierrez-Osuna P. Kakumanu, A. Esposito. A comparison of acoustic coding models for speech-driven facial animation. *Speech Communication*, 48. 4.1.1

[42] G. Salvi. Truncation error and dynamics in very low latency phonetic recognition. In *Proc of ISCA workshop on Non-linear Speech Processing*, 2003. 4.1.1

[43] V. Libal S. M. Chu P. Scanlon, G. Potamianos. Mutual information based visual feature selection for lipreading. In *Int. Conf. on Spoken Language Processing*, 2004. 4.1.1

[44] 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. 5.1.1

[45] L. Rabiner and B.-H. Juang. *Fundamentals of speech recognition.* Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1993. 5.1.1

[46] G. Feldhoffer. Speaker independent continuous voice to facial animation on mobile platforms. In *49th International Symposium ELMAR*, Zadar, Croatia, 2007. 5.2, 6.2.3

[47] J. Valin and C. Montgomery. Improved noise weighting in celp coding of speech - applying the vorbis psychoacoustic model to speex. In *120th Convention AES*, Paris, France, 2006. 6, 6.1.6

[48] http://www.bethsoft.com. 6.1.1

[49] www.counter-strike.net. 6.1.1

[50] http://www.thirdwishsoftware.com/ magpiepro.html. 6.1.2

[51] J. P. Lewis, Matt Cordner, and Nickson Fong. Pose space deformation: a unified approach to shape interpolation and skeleton-driven deformation. In *SIGGRAPH '00: Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 165–172, New York, NY, USA, 2000. ACM Press/Addison-Wesley Publishing Co. 6.1.3

[52] 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. 6.2, 6.2.3