A Framework for Data-Driven Video-Realistic Audio-Visual Speech-Synthesis

Christian Weiss

Institut für Kommunikationsforschung und Phonetik, Universität Bonn
Poppelsdorfer Allee 47, 53115 Bonn / Germany
cwe@ikp.uni-bonn.de

Abstract

In this work, we present a framework for generating a video-realistic audio-visual "Talking Head", which can be integrated in applications as a natural Human-Computer interface where audio only is not an appropriate output channel especially in noisy environments. Our work is based on a 2D-video-frame concatenative visual synthesis and a unit-selection based Text-to-Speech system. In order to produce a synchronized audio-video-stream with novel utterances the speaker never made in the previously recorded corpus, we deploy data-driven selection and concatenation techniques, which we borrowed from known TTS-algorithms. Our framework is organized in an offline-processing and an online-processing stage. The offline module handles data preparation which is used during runtime. The online module manages and generates the audio-visual-synthesis. We use multimedia algorithms to produce a lip-movement, speech segment synchronized audio-video stream. The generated output stream has the form of a camera recorded video. We offer the possibility to choose different sources for the generated "Talking Head". Within the framework it is possible to use the visual output with different speakers or one speaker with different visual mappings. Our framework is built on German speech and video data but could easily be adjusted to other languages.

Introduction

In a face to face communication, the facial gesture plays a major role in disambiguating spoken words and in improving the intelligibility of a spoken utterance (Massaro, 1998), (Beskow, 2003) especially in noisy environments (also known as the cocktail party effect). Therefore a lot of research has been done in the domain of audio-visual synthesis. Some examples of computer animated "Talking Heads" which produce audio and lip synchronized speech can be found in various applications (Chen, Rao, 1998). In this work, we describe a framework of how to produce a video-realistic audio-visual "Talking Head", which can be used as a multimodal Human-Computer-Interface. The framework includes a database that comprises recorded video and audio data of two speakers. The video corpus consists of a video camera recorded speaker in a frontal position who reads our preselected sentences in a newsreader style. The audio corpus is composed of recorded speech that simulates a spontaneous business conversation. With our framework we are able to produce a video-realistic audio-visual synthesis output with precise lip and audio synchrony. Thus it can be seen as a variation of the Video Rewrite algorithm (Bregler, 1997) and the photo-realistic unit selection approach (Cosatto, Graf, 2000). The Video Rewrite algorithm uses viseme-triphone mapping sequences and stores them in a database to retrieve the viseme–triphones when a corresponding sequence is required for the generation of the visual utterance of underlying speech. The photo-realistic unit selection algorithm extracts mouth and eye regions of the speaker and fits them into a base image of the speakers head. To create an appropriate mouth movement of the spoken utterance, the mouth and eye region units which fit best in the base image according to the phonetic transcription, are selected from a database. In distinction to these two approaches, we extract sequences not only of facial regions like mouth or eyes but use the recorded speaker as a video-frame sequence with no preprocessed manipulation of the video source. This means we use the video corpus and cut out complete video-frame sequences which we concatenate to a new sequence. Thus we can maintain the natural movement of all facial gestures. This follows the unit-selection approach, an algorithm which is used in corpus-based concatenative speech synthesis introduced by (Campbell, Black, 1995). The length of the speech and video segments are of variable size. This means that for selection and concatenation of speech segments and visual segments, we select the longest matching segments we stored in our database in order to minimize the distortions while concatenating the segments to create a new one. This is known as "non-uniform unit-selection". The phonetically transcribed input string is the initial step for the segment selection. To synchronize the lip movement with the audio stream, we compute a transition factor between the single 2D-video-frames and the speech segments to control the audio-lip synchrony. The resulting video-realistic audio-visual output stream provides a natural interface to the user. This kind of multimodal output can be integrated in various application scenarios such as supportive functions in educational environments as well as in the area of entertainment or even in the e-commerce domain. This paper is organized as follows: In section 2 we will specify how we built our framework and describe the modules used for the audio-visual output generation process. In section 3 we will describe the visual and the audio segment selection algorithms which are deployed. Furthermore, we will show how we receive synchrony of the visual output with the underlying audio. And in section 4, we will introduce the resulting software. Our conclusion and future work will be discussed in section 5.

2. System Overview

Our framework architecture for generating video-realistic audio-visual-synthesis output is divided into two stages which we call offline- and online-processing. To get a fast and robust way to access our recorded data for selection and concatenation during runtime, we have to prepare it

first. We organized the data preparation in the offline-processing module. The module includes segmentation and annotation of the audio and the video corpus where the segmentation and annotation is semi-automatic and has to be corrected manually. We use an XML-style annotation for our data. The main task, which is the production of the audio-visual output stream, is organized in the online module. The online-processing module includes the symbolic preprocessing of any given input text like text normalization and phonetic transcription. Further on the online-processing module controls the audio and video segment selection and additionally controls the synchrony of lip-movement according to the speech segments. It also includes the final concatenation of the audio segments and video sequences and merges them to produce the audio-visual output-stream.

2.1 Corpora

Our applied corpora consist of recorded speech and recorded video. The speech data comprises about 3500 sentences of spontaneous speech. The recorded video corpus contains one hundred sentences, one hundred word bigrams and one hundred triphone- and diphone-sequences extracted of the speech corpus text. The recorded video has the format PAL 25 fps, 720 x 576 pixels and has been stored as an uncompressed AVI file. The speaker was told to read the sentences in a news announcer style moving the head as little as possible. We recorded the speaker in front of a neutral blue background to simplify fast and robust head pose detection.

2.2 Offline-Processing

The offline-processing module constitutes the initial step to prepare the recorded audio-visual data corpus. We are using a semi automatic approach where the preprocessed data is being manually corrected afterwards.

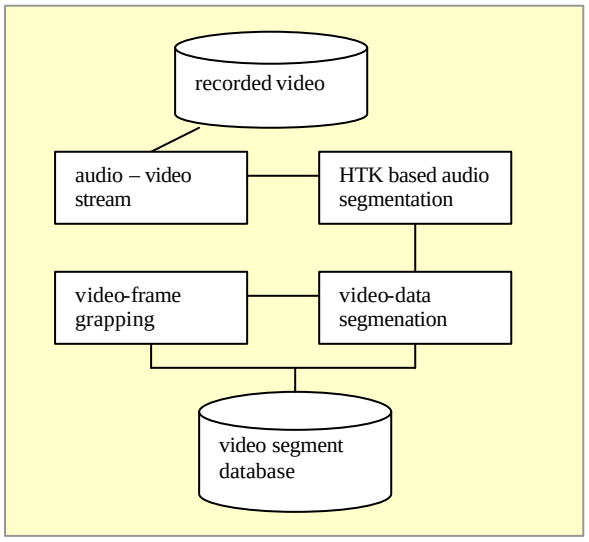

Fig. 1: Offline-Processing-Module Overview

As shown in Fig.1, the audio and video streams are identified in the video and are automatically extracted from the recorded video corpus. Through a Hidden-Markov-Toolkit (http://www.htk.org) based audio segmentation, we get the time stamps of word and phone

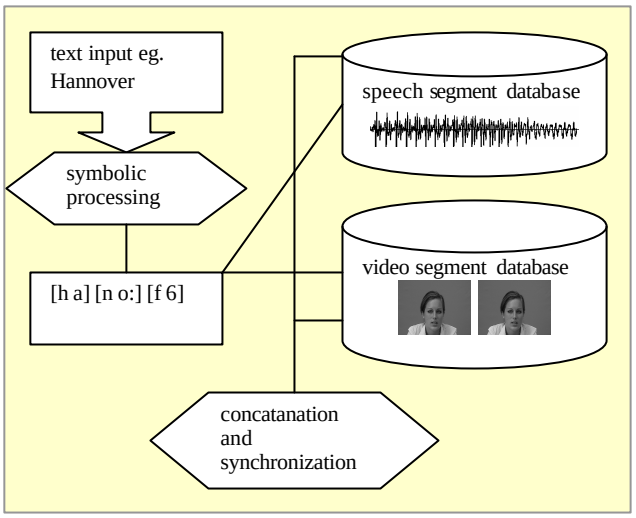

Fig. 2: Online-Processing-Module Overview

boundaries to segment the video stream into video chunks according to the spoken utterance. A post processing manual correction is essential to get proper video and audio segments. In addition to the clipped video sequences we extract the all 2D-video-frames out of the video-segment-stream as denoted in a), b) and c).

a) $T_g = \sum_{j=a}^{e} \Delta t(S_{ij})$, b) $F_g = T_g \cdot f_{ps}$

c) $F_{eff} = K_a + F_g + K_e$

where T_g is the length of the required segment. F_g is the number off all frames, and F_{eff} is the number of the required frames to be extracted. K_a and K_e are constants for frame extraction normalization which is necessary for the alignment of the audio segment duration and the according frame boundaries. This is due to the allegation of the video-frame time which has in our video corpus the default setting of 40 milliseconds. From the extracted 2D-video-frames, we get all the features required in the online processing module (Fig 2) to select the appropriate video segments for concatenation during runtime. The features we need to select are described in section 3.2.

2.3 Online-Processing

Responsible for the audio-visual output generation is the online-processing, which is separated into three modules: a symbolic preprocessing module, a selection and concatenation module and a synchrony and generation module. The symbolic preprocessing module includes the resolution of abbreviations, date, time and numbers of any given input text. In our system only German input text is accepted. This module also provides the phonetic transcription. The transcribed input text is essential for the selection of the speech segments and the visual segments. As shown in Fig. 2 in our speech and visual databases, we are searching for the adequate units to choose for concatenation. The selection of the speech units follows the unit-selection algorithm and is described in section 3.1. After having retrieved the speech and visual segments, we need to synchronize them to get a lip-movement synchrony according to the underlying

concatenated speech. The synchronization is described in section 3.3.

3 Audio-Visual-Synthesis

In general, people are very sensitive in judging the synchrony of mouth gestures according to the speech perceived. As in dubbed movies, we are able to identify minimal distortions in audio-lip synchrony and distortions in the head movement. Leaps are occurring while concatenating non matching head pose video-frames which make the resulting audio-video stream useless. This is a non trivial challenge to the synchrony and concatenation module of our audio-visual speech-synthesis system. We provide our solution to this question in section 3.2 and 3.3. The synthesis of the acoustic representation and of the visual representation from the input text is following the unit selection algorithm introduced by Hunt and Black (1996). In our system, we use a spontaneous speech corpus and a test video corpus to generate new audio-visual utterances resulting in a video-stream. The audio source is independent from the applied video source. This makes the system very flexible to adapt the video to new voices and vice versa.

3.1 Audio Segment Selection

To synthesize any grapheme input text, we use a unit selection based Text-to-Speech synthesis approach. Each segment that is selected out of our speech segment database is concatenated and results in a complete utterance. For concatenation we use speech segments of variable size. This method preserves the naturalness of a speaker's voice and does not need any manipulation of the speech signal. Fig. 3 denotes the selection and concatenation process of the appropriate speech units.

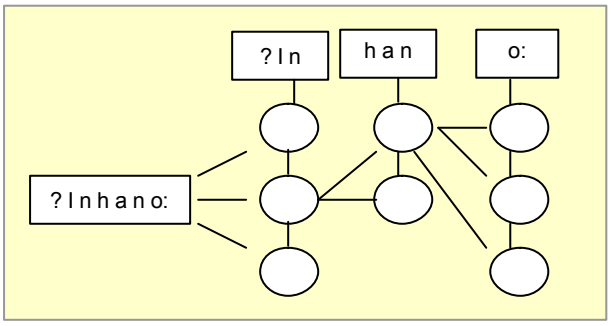

Fig. 3: Unit Selection Overview

In order to select the adequate speech unit candidate, a two-step mechanism is deployed. The first step is to pre-select the possible units while evaluating quantitative and qualitative features. Then we include contextual information, while the best fitting candidate is chosen according to the given constraints. The next step is to analyze the acoustic context with regard to its features pitch, duration and the spectral context represented by the mel-frequency cepstral coefficients. To avoid distortions of the intonation contour and to avoid spectral distortions during concatenation, a two dimensional cost function is applied, where the segments with the least costs are selected for concatenation. The audio synthesis is integrated in our video-realistic audio-visual synthesis system. Besides the internal speech synthesis module, the system supports the BOSS TTS system which was developed at IKP (Stöber, 2002) and is a fully functional non restricted unit selection based TTS system.

3.2 Video Segment Selection

In order to produce a mouth movement sequence synchronized with the synthesized audio, we have to select the 2D-video-frame sequences which map the appropriate visemic representation. We define here the visemic representation as a counterpart to the phonetic transcription of text (Fig. 4). The video segment selection and generation process works in the same way as the speech segment selection and concatenation process:

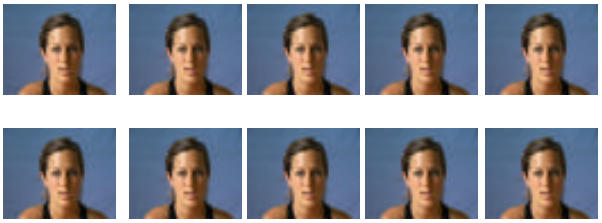

Fig. 4: Visemic representation of German syllable

We select the video-frame sequences according to the phonetically transcribed input text. The first frame of the selected 2D-video-frame sequence is our base-frame, which provides the features used to select the following 2D-frame sequences. This results in three head pose coordinates of the basis frame. Here we choose a pixel-color based determination of the appropriate head pose coordinates. Due to our common blue background in the 2D- video-frames, we are able to detect the head in a iterative way. Then we compute the Euclidian distance to the following frames which are suitable according to the phonetic transcription. The computation follows: a), b) and c).

$$\text{a) } \min \Delta P(n), where \Delta P(n) = | P_b - P_n |$$

$$\text{b) } \min \Delta Q(n), where \Delta Q(n) = | Q_b - Q_n |$$

$$\text{c) } \min \Delta R(n), where \Delta R(n) = | R_b - R_n |$$

The reference frame coordinates are denoted as P_b, Q_b and R_b, the coordinates of the following frames are denoted with P_n, Q_n and R_n. The distance threshold of the next frame sequences can be set in a range of 1-5 pixels where a difference of five pixels reflects a strong distortion in the natural movement of the speakers head. A distance measure on an acceptable deviation of the basic frame and the suitable following frames has to be further examined. Edge detection based techniques will also be evaluated in future work for adopting robust frame selection. After all 2D-video-frames have been selected, we have to align the video-frames with the audio-stream before we concatenate them to an entire video-stream. This audio-video synchronization step is described in the following section.

3.3 Audio-Video Synchronization

A minimum of distortion in naturalness is crucial for the user acceptance of a video-realistic audio-visual synthesis.

Thus, synchrony of lip-movement and acoustic output is as important for a seamless and smooth concatenation of the video frame sequences particularly in regard to the head pose. For lip-movement according to audio synchrony, we defined a frame transition factor TF. This frame transition factor controls the transition speed of the selected 2D-video-frames depending on the length of the selected speech segment. We compute this transition factor as shown below:

$$TF = T_{Sij} / N_{FpVS},$$

where T_{Sij} is the time of the speech segment and N_{FpVS} is the number of frames per video frame sequence. For each speech segment, we want to concatenate, we re-compute the transition factor and thus get a non-linear synchrony for the whole utterance. This allows us to produce a natural audio-visual synthesis.

4. The AVISS Software

We implemented our audio-visual synthesis framework in the resulting AVISS software (Fig. 5). Our software can be applied for educational and research purposes. It is capable to produce video-realistic audio-visual output from underlying audio and video databases while the user can choose whether he wants to generate the output video-stream manually or automatically. Within the manually option, all steps are comprehensive and the user is asked to select the adequate speech segment himself.

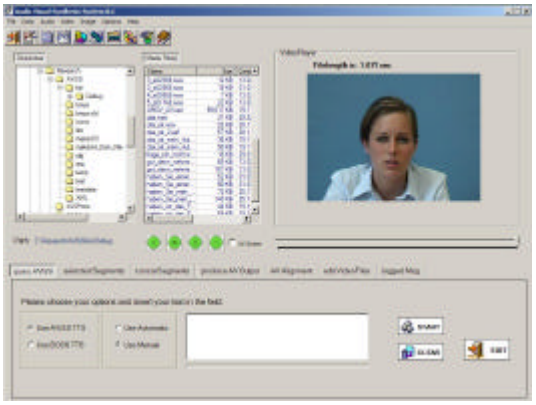

Fig. 5: Audio-Visual Synthesis System Software

Examples of the generated audio-visual output-stream are provided on the author's webpage (http://www.ikp.uni-bonn.de/~cwe).

5. Conclusion and Future Work

Our research concentrates on data-driven, video-realistic audio-visual speech-synthesis resulting in a framework-software to generate audio-visual-synthesis output. As an outcome, the user sees a video of a speaker who makes utterances which seem entirely natural. However these utterances do not occur as a whole in the previous recoded data. We dispose a data-driven video-realistic audio-visual synthesis framework with resulting software to generate such an audio-visual synthesis output. We use a well known algorithm for speech synthesis and a derived algorithm to generate the visual output. The best results while synthesizing audio-visual sequences are achieved with longer 2D-video-frame sequences such as a video-frame sequence in word length. Nevertheless, segments of syllable size also produce an appropriate visual output. In our future work, we will concentrate on removing distortions at segment concatenation points for audio and visual concatenation. For an unrestricted audio-visual synthesis, we need to record a phonetically balanced visual corpus. Further on we will have to make a broad evaluation of the audio-visual synthesis output according to the visual quality, including synchrony and intelligibility.

Acknowledgements

We have to thank Bianca Aschenberner who reads our sentences for the video recordings. Additionally, we want to thank our colleagues for their helpful comments on our work.

References

Black, A. and Campbell, N., (1995). Optimizing selection of units from speech databases for concatenative synthesis. Eurospeech, Madrid, Spain (1): 581-584.

Beskow, J., (2003). Talking Heads. Models and Applications for Multimodal Speech Synthesis. Dissertation, Stockholm.

Bregler, C., Covell, M., Slaney, M., (1997). Video Rewrite: Driving Visual Speech with Audio. Proc. SIGGRAPH, ACM SIGGRAPH: (pp 353 – 360).

BOSS, Bonn Open Synthesis System. http://www.ikp.uni-bonn.de/dt/forsch/phonetik/boss/index.html

Cosatto, E., Graf, H.-P.,(2000). Photo-Realistic Talking-Heads from Image Samples. In: IEEE Transactions on Multimedia, 2(3): (pp 152 – 163).

Ezzat, T., Poggio, T., (1998). MikeTalk: Facial Display on Morphing Visemes. Proceedings of the Computer Animation Conference Philadelphia, PA, June 1998.

Hunt, A. and Black, A., (1996). Unit selection in a concatenative speech synthesis system using a large speech database. Proceedings of ICASSP, Atlanta, Georgia: (pp 373-376).

Massaro, D.,W., (1998). Perciving talking faces: From speech perception to a behavioral priciple. Cambridge, MA: The MIT Press.

Chen, T., Rao, R.E., (1998). Audio-Visual Integration in Multimodal Communication. Special Issue on Multimedia Signal Processing, IEEE 5.

Stöber, K., et al., (2000). Speech Synthesis by Multilevel Selection and Concatenation of Units from Large Speech Corpora. Verbmobil: Foundations of Speech-to-Speech Translation, Symbolic Computation, Springer, Berlin: (pp 519-537)

Stöber, K. (2002): Bestimmung und Auswahl von Zeitbereichseinheiten für die konkatenative Sprach-synthese. Dissertation, Bonn.

Yang, J., Xiao, J., Ritter, M, (2000). Automatic Selection of Visemes for Image-based Visual Speech Synthesis. Proceedings of First IEEE International Conference on Multimedia IEEE ME.