

# DISTRIBUTED AUDIO-VISUAL SPEECH SYNCHRONIZATION

*Peter Poller, Jochen Müller*

DFKI

German Research Center for Artificial Intelligence

Stuhlsatzenhausweg 3

66123 Saarbrücken, Germany

*{poller|jmueller}@dfki.de*

## ABSTRACT

The main scientific goal of the SmartKom project is to develop a new human-machine interaction metaphor for multimodal dialog systems. It combines speech, gesture, and facial expression input with speech, gesture and graphics output. The system is realized as a distributed collection of communicating and cooperating autonomous modules based on a multi-blackboard architecture. Multimodal output generation is consequently separated in two steps. First, the modality-specific output data are generated. Second, an inter-media synchronization of these data is realized on independent media devices to perform the multimodal presentation to the user. This paper describes the generation of appropriate lip animations that are based on a phonetic representation of the speech output signal and as a second computational step the timestamp based realization of audio-visual speech output on distributed media devices.

## 1. INTRODUCTION

SmartKom ([www.smartkom.org](http://www.smartkom.org)) is a multimodal dialog system that supports the situated understanding of possibly imprecise, ambiguous, or partial multimodal input and the generation of coordinated, cohesive, and coherent multimodal presentations based on the situated delegation-oriented dialog paradigm (SDDP, [9]) in which the user delegates a task to a virtual communication assistant, visualized as a life-like character on a graphical display.

According to SDDP, SmartKom breaks with the traditional desktop interaction metaphor that is based on WIMP (windows, icons, mouse pointer) interaction. We radically reduce the content of the graphical user interface to only those elements (e.g., graphics) that are relevant to the user. These are presented on a black background. Thereby, our communication assistant also gets multimodal capabilities that humans don't have. We investigate a new multimodal interaction metaphor within a multimodal dialog system aiming to reach new fields of human-machine interaction by taking advantage of the unlimited virtuality of the communication assistant.

There are three different instances of SmartKom. SmartKom PUBLIC is a system which can be used as a multimodal information kiosk. The gestures of the user are tracked by the Siemens Virtual Touch screen (SIVIT®) and the display is projected on a screen. In the SmartKom HOME scenario, a Tablet PC is used to show the visual output of SmartKom. Here, the system can be used as an information system at home, e.g., as an EPG. SmartKom MOBILE is an instance of the SmartKom system with its

display on a PDA. Here, the presentation agent Smartakus guides an user through a city and shows him interesting sights. HOME and MOBILE use a pen as the gesture input device.

In order to reduce the complexity of modules and representations, SmartKom can be divided into a set of version specific output modules that have to be adapted to the concrete applications and scenario specific hardware devices while a so called multimodal dialog backbone, i.e., a set of modules that are responsible for recognition, analysis, and interpretation of user input data, works independently of the current applications and hardware devices.

The system characteristics described above impose several specific requirements on multimodal presentations and also on inter-media lip synchronized speech output as a part of them. Thus, a major goal of our lip synchronization approach are scenario-independent usability and online speed behavior in the distributed system environment as part of the ambitious multimodal SDDP human-machine interaction research.

Under these circumstances and due to the fact that phonetic representations of the speech output signal are available, we chose a less complex sample-based approach to generate synchronized lip animations. This approach is described in the rest of the paper.

## 2. THE MODULAR OUTPUT ARCHITECTURE

In this section, we focus on the modular architecture of the multimodal generation and presentation tasks of SmartKom. Two central modules control the realization of multimodal presentations: the Presentation Planner and the Display Manager. The task of the Presentation Planner is the computation of media and modality specific representations that are passed to the Display Manager whose task is to perform the multimodal presentation on the modality specific output devices. Figure 1 illustrates the main computational steps of the two modules in the right column.

The data flow in SmartKom to achieve visual speech output is indicated in figure 1. After the identification of modality specific presentation tasks, the Presentation Planner publishes the speech generation task to a Text Generator, which in turn sends a conceptual representation of the text (including phrase structure) to Speech Synthesis. Beside the speech output data, the synthesis module also publishes a phonetic representation of the speech data including the phonemes and timestamps for them. This representation is used as input by the Presentation Planner again in order to generate a corresponding lip animation script that is finally passed to the Display Manager to perform the multimodal presentation

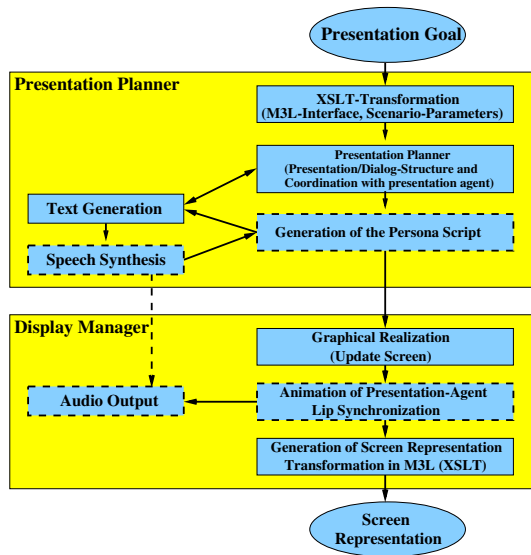


Fig. 1. Presentation-Pipeline in SmartKom

including audio-visual speech output.

### 3. PLANNING OF THE MODALITY-SPECIFIC OUTPUT

The task of the Presentation Planner is the planning of multimodal presentations including graphics that are combined with gestures and audio-visual speech (see [6] for details). Figure 2 shows an example presentation resulting from a TV schedule request.

A special sub-task of this planning process is the generation of a lip synchronized animation script for the audio-visual speech output of our presentation agent Smartakus.

#### 3.1. Knowledge Sources

Smartakus is modeled in 3D as an artificial life-like character with 3D-Studio-Max. The lip synchronization is based on a predefined indexed set of different mouth position pictures (visemes) developed with 3D-Studio-Max and stored as static unmoved GIF pictures. In our system, a viseme is defined as a specific mouth position picture (www.whatis.com: generic facial image).

However, closely cooperating with the speech synthesis group in SmartKom (IMS, University of Stuttgart [7], <http://www.ims.uni-stuttgart.de/phonetik/synthesis>), we found that due to the cartoon-like character of Smartakus (neither tongue nor teeth visible) only a limited variety of mouth/jaw positions or movements are possible at all. Consequently, the only parameters that we found relevant to describe mouth positions are the lip rounding and the jaw opening. Lip rounding is a binary parameter because lips are either rounded or not. On the other hand, for jaw opening we identified 4 different values being reasonable and sufficient especially with respect to the precision that is needed in SmartKom in all three scenarios. Figure 3 shows the 8 different visemes currently used in the PUBLIC scenario. They resulted from several optimization procedures involving our designer and the speech synthesis group of SmartKom. The first row shows the visemes with unrounded lips and 4 different opening degrees, the second row the corresponding rounded lips. The visemes are named by concatenating a letter for

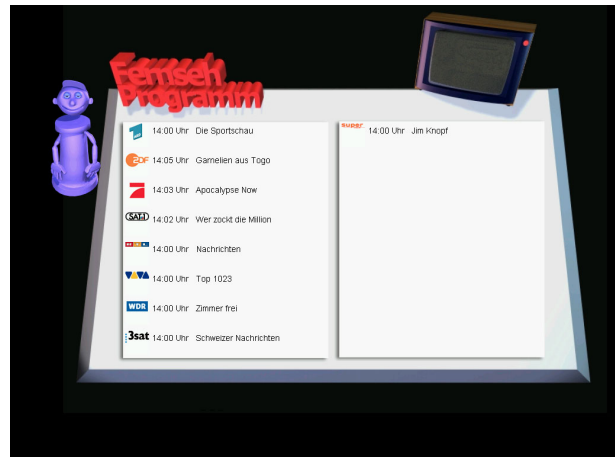


Fig. 2. Smartakus showing the current TV schedule (German: “Fernsehprogramm”)

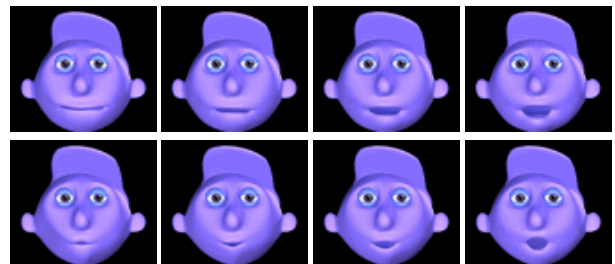


Fig. 3. Smartakus' Visemes  $u_0, \dots, u_3, r_0, \dots, r_3$

the lip rounding value ('r' = rounded, 'u' = unrounded) with a number for the jaw opening degree ('0' = closed, '1' = slightly opened, '2' = opened, '3' = widely opened).

Smartakus has been designed as an artificial human-like character intended to play a central role in investigating a new multimodal human-machine interaction metaphor. Consequently, we found the simple face model for lip animation described above reasonable and sufficient with respect to the research focus of SmartKom. Other approaches (e.g., [2] or [4]) deeply model a human face as a whole. To some extent, SmartKom tries to go beyond the various aspects of approaching human-human conversation in virtual models of real worlds described in the papers in [3] by extending them to new multimodal areas that are unreachable in reality.

However, in cooperation with the synthesis group at IMS, we developed an underspecified mapping of phonemes to visemes. We found that almost every phoneme has only one corresponding viseme, while just a few of them (e.g., plosives and diphthongs) have to be mapped to at least two visemes to visualize their articulation appropriately [7]. Thus, the mapping becomes a 1:n mapping. Furthermore, the mapping has to be partly underspecified in lip rounding and jaw opening as well to be able to take coarticulation effects into account.

The underspecified viseme(s) corresponding to a phoneme that result from this mapping are structurally integrated into the phonetic representation that is published by the synthesis module. Figure 4 show a simplified excerpt of this XML representation con-

```

<synthesizedTextTime>
...
<phoneme> n </phoneme>
<viseme>
  <StartTime> 0.242 </StartTime>
  <EndTime> 0.282 </EndTime>
  <opening>
    <minimum> 1 </minimum>
    <maximum> 2 </maximum>
  </opening>
  <rounding> unspec </rounding>
</viseme>
...
</synthesizedTextTime>

```

**Fig. 4.** Phonetic representation as M3L document

forming to the XML Schema based “Multi-Modal Markup Language – M3L” that was developed in SmartKom.

Beside general information about the utterance (e.g., the sentence type and the individual words), the representation consists of representations as shown in figure 4 for each phoneme in the speech signal. The information given about a phoneme includes the exact times in milliseconds and underspecified information that restricts the selection of appropriate visemes. The jaw <opening> degree is restricted by a <maximum> value and a <minimum> value. The lip <rounding> value is either unspecified ('unspec') as in figure 4 or specified ('rounded', 'unrounded').

### 3.1.1. The Algorithm

Based on the phonetic representation shown in figure 4, the Presentation Planner generates a lip animation script to be executed appropriately by the Display Manager during speech output. This script is generated by a stepwise procedure iterating over the phonemes that consecutively specifies the corresponding viseme(s) and their exact time points at which they have to be displayed. The determination of a viseme considers the neighboring (possibly still underspecified) visemes and follows the following criteria to fix an underspecified viseme:

- avoid “extreme” mouth opening degrees if possible
- prefer less differing consecutive visemes

In terms of figure 3 the procedure always tries to select a viseme that has a common borderline with the previous viseme whenever possible (also by inserting intermediate visemes if necessary).

The concrete time points of the visemes do not coincide with the beginning time of phonemes. Currently, the viseme timestamps are decremented by a constant time (20 ms) because lip and jaw movements take place before the articulation starts (e.g., when articulating an 'm', the lips have to be closed first before the articulation is possible at all). We are currently investigating the question whether the time shift is the same for all phonemes/visemes or is phoneme/viseme dependent. But at this point, we benefit from the extremely small amounts of time (milliseconds) exceeding the recognition capacity of the human eye.

An example lip animation script for the German word “Kino” is shown in figure 5.

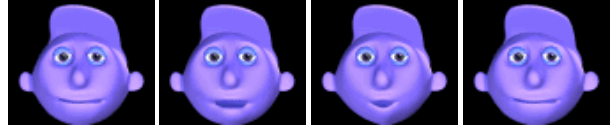
The number 4 after the command name 'skspeak' indicates that this lip animation consists of 5 pairs of viseme and duration

```

skspeak 4 ap_1 u0 78 u2 127 r2 123 u0 166
in_talk out_talk

```

**Fig. 5.** Lip animation script



**Fig. 6.** Smartakus' Visemes for “Kino”

(in milliseconds). The identifier ap\_1 is a unique name to identify the corresponding audio stream. This is followed by the data defining the lip animation itself, consisting of 4 visemes that have to be displayed for the indicated number of milliseconds. All animations are visually surrounded by 'u0' such that the mouth is always closed before and after lip animations. Figure 6 shows the resulting viseme sequence. The phonemes 'l' and 'n' are both mapped to 'u2' in this example.

The last two elements “in\_talk” and “out\_talk” name the preparation and the retraction gestures of the speech animation, which visually enclose the mouth animation itself. Currently, preparation and retraction consist of lifting the right arm before speaking.

Finally, the lip animation script is sent to the Display Manager for execution during the presentation.

## 4. REALIZATION OF LIP-SYNCHRONIZED SPEECH OUTPUT

The animations of the presentation agent Smartakus have to be related to the graphical output for pointing gestures and to the speech output for synchronous mouth movements. In SmartKom, gesture animations have to be aware of two kinds of synchronization in order to realize a natural behavior of the communication assistant on the output screen (figure 2):

- All deictic gestures have to be “graphically” synchronized with the display of the corresponding graphical output they are related to.
- The acoustic speech output has to be synchronized by appropriate lip movements of the presentation agent Smartakus.

In the SmartKom system, the script for our presentation agent Smartakus is sent to the Display Manager, which is responsible for showing the generated graphics, screens, and animations. The script is generated by the Presentation Manager which uses a presentation planning system called PrePlan ([1],[8]).

The Display Manager is responsible for the realization of visible output while the Audio Output module realizes speech output. Consequently, audio output and visual output are performed independently from each other and can even be processed and realized on different machines. When presenting both modalities, we have to merge them to get lip synchronized output.

First, the Display Manager updates the screen according to the screen definition delivered by the presentation planner.

After the static elements of the screen are updated, the PersonaPlayer [5] which is integrated in the Display Manager runs the generated animations.

For each primitive gesture specified in the presentation script, it selects a sequence of bitmaps from an indexed data base. At run time complex gestures are constructed out of primitive gestures.

In the script, there are not only commands for displaying gestures, but also commands for the interaction with the other elements of the user interface and other modules of the SmartKom-System. For example, the presentation agent Smartakus is able to dial a number on the phone which is shown on the SmartKom display. Also the synchronization of the speech gesture with the audio output is done with a special command of the engine of the Persona-Player.

In SmartKom, the module which is responsible for the output of the audio-data (called the Audio-Device of SmartKom) and the Display Device run on different computers. In order to synchronize the audio and the video output, we use synchronized clocks. The Persona-Player defines a point in time when the audio output should be started. Then the the player waits until this point in time is reached. Assuming that the clocks of the computers are synchronized, the animation and the audio output are synchronized according to the script and techniques shown with figure 5.

In order to let the system run on slow machines, we must ensure that the display of the frames does not consume too much time and the synchronization between audio and video gets lost. Therefore, we continuously check if the time for displaying the frames exceeds the defined values. For each frame we measure the duration needed to display it. If it is greater than the precomputed duration for the current frame given within the `skspeak`-command, we reduce the duration of the next frame appropriately. On slow hardware, the remaining duration of a frame can become negative. Then, the frame is dropped. This procedure is repeated until positive values are achieved again.

## 5. IMPLEMENTATION AND EVALUATION

The Presentation-Manager and the Display-Manager of SmartKom are implemented in Java. The SmartKom-System uses a cluster of computers running Linux and Windows NT. Since we use Java to implement the two modules, we are able to start these modules under both operating systems.

To check the “correctness” of the audio-visual output, we evaluated the distributed synchronization on different machines working with different sound cards and different graphic boards and found that there is an inherent, unpredictable time shift between the two media streams ranging up to around 50 ms. Therefore, we implemented a menu permitting the users to manipulate the time shift based on their subjective impression.

So far, we only evaluated the resulting viseme frequency for a set of 50 example sentences. The result was that the number of visemes per second ranges from 7 up to 15. Although this frequency seems to be relatively low, it seems to suffice to exceed the recognition capacity of the human eye because the 15 people we demonstrated the system did not recognize that the lip animation output stream is realized as a synchronous presentation of unmoved mouth position pictures. Currently, the SmartKom system is being extensively evaluated with respect to various usability and quality criteria.

## 6. CONCLUSION

In this paper, we described the lip synchronization of the presentation agent Smartakus of the multimodal dialog system SmartKom.

Since we use precalculated bitmaps for the graphical visualization of the life-like character, we are able to use various display devices like ordinary PC, Webpads or PDAs for output. For the synchronization of the lip movements with the audio output, we use time and phoneme information delivered by the speech synthesizer. The script calculated from this data is evaluated by the Persona-Engine to display the animation.

## 7. ACKNOWLEDGMENTS

This work was funded by the German Federal Ministry of Education, Science, Research and Technology (BMBF) in the framework of the SmartKom project under Grant 01 IL 905 K7. The responsibility for the contents lies with the authors.

We'd like to thank Norbert Reithinger and Tilman Becker for their invaluable and helpful comments on earlier versions of this paper.

Finally, we'd like to thank Antje Schweitzer of the IMS, University of Stuttgart, for the collaboration and fruitful discussions.

## 8. REFERENCES

- [1] E. André. *Ein planbasierter Ansatz zur Generierung multimedialer Präsentationen*. PhD thesis, Universität des Saarlandes, 1995.
- [2] J. Beskow, M. Dahlquist, B. Granström, M. Lundeberg, K.-E. Spens, and T. Ohman. The teleface project multi-modal speech-communication for the hearing impaired. In *Proc. Eurospeech '97*, pages 2003–2006, Rhodes, Greece, 1997.
- [3] J. Cassell, J. Sullivan, S. Prevost, and E. Churchill. *Embodied Conversational Agents*. The MIT Press, Cambridge, MA, USA, 2000.
- [4] D. Massaro, M. Cohen, J. Beskow, and R. Cole. Developing and evaluating conversational agents. In J. Cassell, J. Sullivan, S. Prevost, and E. Churchill, editors, *Embodied Conversational Agents*, pages 287–318. MIT Press, Cambridge, MA, 2000., 2000.
- [5] J. Müller. *Persona: Ein anthropomorpher Präsentationsagent für Internet-Anwendungen*. PhD thesis, Universität des Saarlandes, Saarbrücken, 2000.
- [6] J. Müller, P. Poller, and V. Tschernomas. Situated delegation-oriented multimodal presentation in smartkom. Proceedings of AAAI-02 Workshop on Intelligent Situation-Aware Media and Presentations, Edmonton, Alberta, Canada, 2002.
- [7] A. Schweitzer, G. Dogil, and P. Poller. Gesture-speech interaction in the smartkom project. Poster presented at the 142nd meeting of the Acoustical Society of America (ASA), 2001. Ft. Lauderdale, FA, USA, <http://www.ims.uni-stuttgart.de/schweitz/documents.shtml>.
- [8] V. Tschernomas. *PrePlan Dokumentation (Java-Version)*. Deutsches Forschungszentrum für Künstliche Intelligenz, Saarbrücken, 1999.
- [9] W. Wahlster, N. Reithinger, and A. Blocher. Smartkom: Multimodal communication with a life-like character. In *Proceedings of Eurospeech 2001, 7th European Conference on Speech Communication and Technology*, volume 3, pages 1547 – 1550, Aalborg, Denmark, 2001.