

An Expandable Web-based Audiovisual Text-to-Speech Synthesis System

Sascha Fagel and Walter F. Sendlmeier

Department of Communication Science

Technical University Berlin

sascha.fagel@tu-berlin.de

sendl@kgw.tu-berlin.de

Abstract

The authors propose a framework for audiovisual speech synthesis systems [1] and present a first implementation of the framework [2], which is called MASSY - Modular Audiovisual Speech SYnthesizer. This paper describes how the audiovisual speech synthesis system, the 'talking head', works, how it can be integrated into web-applications, and why it is worthwhile using it.

The presented applications use the wrapped audio synthesis, the phonetic and visual articulation modules, and a face module. One of the two already implemented visual articulation models, based on a dominance model for co-articulation, is used. The face is a 3D model described in VRML 97. The facial animation is described in a motion parameter model which is capable of realizing the most important visible articulation gestures [3][4]. MASSY is developed in the client-server paradigm. The server is easy to set up and does not need special or high performance hardware. The required bandwidth is low, and the client is an ordinary web browser with a freely available standard plug-in.

The system is used for the evaluation of measured and predicted articulation models and is also suitable for the enhancement of human-computer-interfaces in applications like e.g. virtual tutors in e-learning environments, speech training, video conferencing, computer games, audiovisual information systems, virtual agents, and many more.

1. Introduction

Producing and perceiving speech is the usual way of communication. Under specific circumstances, human-computer-interaction might be improved by speech, where one aspect is speech output [5][6][7]. But research has shown that the possible advantage decreases by growing level of abstraction from natural speech [8]. As human speech communication consists of several information streams, a coherent presentation of audible and visible speech like that provided by MASSY should enhance several quality parameters compared to audio only presentation. In addition, speech synthesis is an appropriate instrument for perception experiments, because – in comparison to natural speech – every variable is strictly under control.

An analysis of visible articulation movements – initially for German – was carried out. This resulted in a set of motion parameters to control the speech movements of the talking head. An application was created to estimate still positions of phonemes, respectively groups of visually identical phonemes (visemes). See figure 1 for some examples. The contextual

dependencies, i.e. the influence of neighboring phonemes, have to be taken into account in order to generate realistic visual phoneme chains. Then audible speech has to be generated to be played synchronously with the visible synthetic speech.

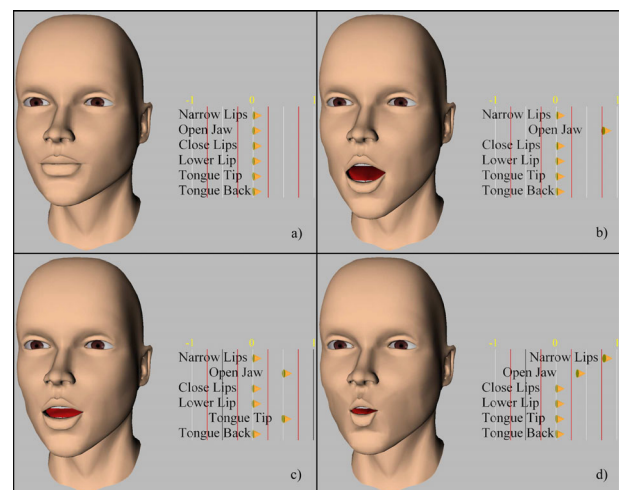


Figure 1: Estimated parameter settings for the phonemes /m/ (a), /a/ (b), /l/ (c), and /o/ (d).

2. The system

Figure 2 shows a system overview. The internal interfaces are specified and can be used to control the behavior of MASSY. Special functionality thus may be realized.

2.1. Phonetic articulation module

The English text to phoneme transcription is using the high level synthesis of the festival speech synthesizer of the University of Edinburgh [9]. Supported are female and male intonation contours. The German female and male text to phoneme conversion is realized by the integration of the high level speech synthesis part of HADIFIX, a speech synthesizer of the University of Bonn [10].

2.2. Audio synthesis

The MBROLA speech synthesis algorithm of the Polytechnic Faculty of Mons is wrapped in MASSY's audio synthesis module [11].

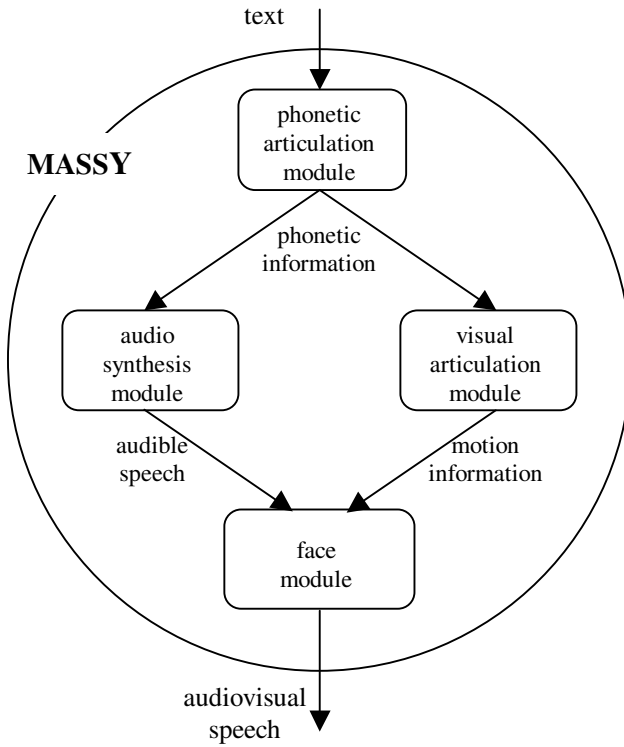


Figure 2: Schematic system overview of MASSY.

2.3. Visual articulation module

The visual articulation module implements a dominance model to generate the motion information for visualizing phoneme chains. The model considers the differently realized phonemes depending on their phonetic context. The real target position of an articulator is calculated for each phoneme based on fictitious ideal positions and the strength (the dominance) to control this articulator. In the current implementation only one parameter describes the influence on neighboring phonemes and the susceptibility to neighbors. Right sided and left sided dominance are assumed to be equal. Equation (1) defines the context-dependently adjusted target position.

$$T_n = \frac{D_n I_n + \frac{1}{2} \sum_{i=n-1}^0 \left(D_i I_i \prod_{k=n}^{i+1} (1 - D_k) \right) + \frac{1}{2} \sum_{j=n+1}^N \left(D_j I_j \prod_{l=n}^{j-1} (1 - D_l) \right)}{D_n + \frac{1}{2} \sum_{i=n-1}^0 \left(D_i \prod_{k=n}^{i+1} (1 - D_k) \right) + \frac{1}{2} \sum_{j=n+1}^N \left(D_j \prod_{l=n}^{j-1} (1 - D_l) \right)} \quad (1)$$

where T is the target position,
 D is the dominance,
 I is the ideal value,
the index n means the current phoneme,
indices greater n indicate phonemes on the right side,
indices smaller n indicate phonemes on the left side.

The calculated target position of a phoneme is always held for a fixed fraction (currently 60%) of the phoneme duration centered in the phoneme. The relative duration and position of the quasi-stationary phase may be varied per articulator and phoneme in order to achieve a more exact articulation model.

The transitions between the calculated target positions of the phoneme chain are either retrieved via linear interpolation or with an especially developed spline interpolation. These splines are defined with the boundary condition not to overshoot the target positions. Additionally a criterion of minimal corrugation is used, i.e. the sum of the absolute values of the acceleration is minimized following the paradigm of minimal articulation effort.

2.4. Face module

The face module controls a 3D face model, that is described in VRML 97 [12] with additional elements (nodes) according to the H-Anim 2001 standard [13]. The difference of the neutral 3D model to the model deformed to the maximum position of one articulator constitutes an articulator displacement. The motion vectors of all affected vertices besides the concerning vertex index are stored as a so called displacer (H-Anim 2001). The motion information is translated into an animation, i.e. each set of articulator positions is converted to a complete set of vertices of the face model. For this purpose the face module linearly combines the amounts of displacements of all articulators. The face module then dynamically generates a 3D scene with the animated face in the VRML file format.

3. Embedding the talking head

The implementation of the framework is realized in the server-sided scripting language PHP (a project of the Apache Software Foundation, [14]) which is especially suited for web development. The system gets a plain text and returns the file with the animated talking head. The animation can either be displayed in a separate frame or can be embedded as an object into any HTML document. In both cases the text is transferred to the text-to-audiovisual-speech system as query parameter (spaces replaced by control characters). The system then returns the dynamically generated animation, which includes a reference to the also generated audio file. The HTML code might look as follows:

```
<OBJECT
  CLASSID="CLSID:86A88967-7A20-11d2-
    8EDA-00600818EDB1"
  ID=CORTONA width="128" height="196"
  CODEBASE="http://www.parallelgraphi
    cs.com/bin/cortvrml.cab#Version=4,0
    ,0,76">
  <PARAM
    NAME="Scene"
    VALUE="talkinghead.php?text=Hello%2
    0Peter.%20You%20have%20some%20email
    s%20and%20a%20voice%20message.%20It
    %20will%20be%20around%2025%20degree
    s%20today.%20Have%20a%20nice%20day.
    ">
</OBJECT>
```

4. Applications

4.1. Evaluating articulation models

MASSY is used for the evaluation of different articulation models and their options. Intelligibility tests are carried out to investigate the magnitude of articulation movements with other parameters constant, the duration of the quasi-stationary phase, and the time center of the quasi-stationary phase. The same server on which MASSY is running serves as experiment controller which plays the stimuli and records the answers of the test persons. Figure 3 shows a screenshot of this test.

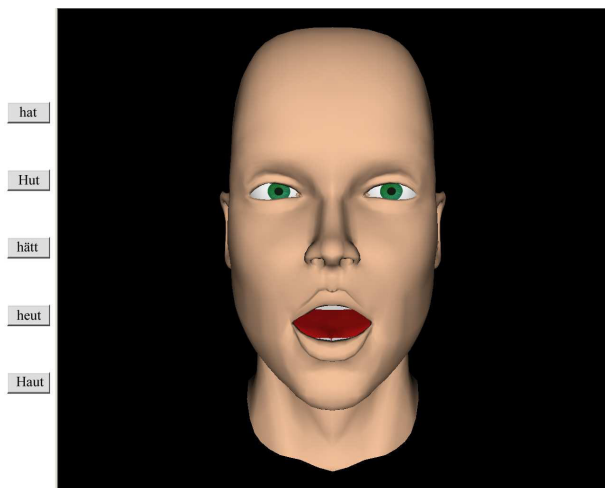


Figure 3: Stimulus and answer alternatives of an audiovisual intelligibility test to evaluate articulation models.

4.2. Virtual agent

MASSY may serve as virtual agent, whereas the currently retrieved data can dynamically be displayed by audiovisual speech. In the example application (figure 4), the web browser is started on logon and loads the personal virtual

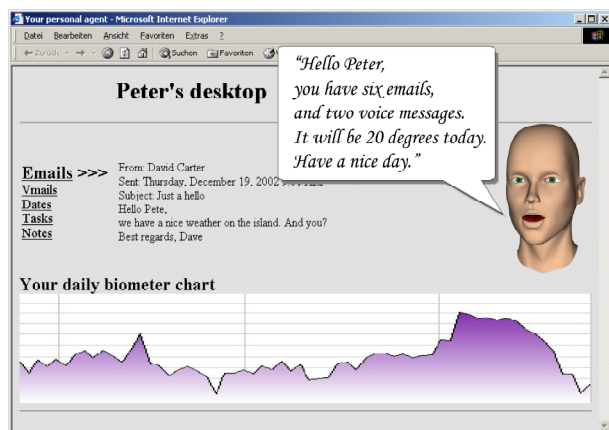


Figure 4: MASSY embedded as virtual agent into a dynamically generated desktop view.

desktop site of the user. The addressed server collects the data that is specified in the user profile (this procedure is held simple, because it is not in the research focus). Then the server accordingly generates an utterance and the browser displays it audiovisually. This usage is also transferable to virtual lecturers in e-learning environments.

5. Conclusions

The presented expandable web-based text-to-audiovisual-speech synthesis system is capable of serving as an extension to current human-computer-interaction. The specified internal system interfaces enable the developer to control or exchange modules, to evaluate their specific advantages, and to detect their weak points. The effective gain of intelligibility and comprehension achieved by the system will be measured in perception experiments. Furthermore it will be used to approve effects like hyperarticulation or speaker-dependent phoneme realization measured in real utterances.

References

- [1] Fagel, S., Sendlmeier, W. F., "Entwurf eines Frameworks für audiovisuelle Sprachsynthesysteme", *Tagungsband der 13. Konferenz Elektronische Sprachsignalverarbeitung ESSV*, pp. 372-378, Dresden, 2002.
- [2] Fagel, S., "MASSY - a Prototypic Implementation of the Modular Audiovisual Speech SYNthesizer", *Proc of the 15th International Congress of Phonetic Sciences ICPHS* (to appear), Barcelona, 2003.
- [3] Cohen, M. M., Massaro, D. W., "Modeling Co-articulation in Synthetic Visual Speech", in Magnenat Thalmann, N. & Thalmann, D. (eds.), *Models and Techniques in Computer Animation*, pp. 139-156, Tokyo, 1993.
- [4] Löfqvist, A., "Speech as Audible Gestures", in Hardcastle, W. J., Marchal, A. (eds.), *Speech Production and Speech Modeling*, Dordrecht, 1990.
- [5] Sumby, W. H., Pollack, I., "Visual Contribution to Speech Intelligibility in Noise", *Journal of the Acoustical Society of America* (26), pp. 212-215, 1954.
- [6] Erber, N. P., "Interaction of Audition and Vision in the Recognition of Oral Speech Stimuli", *Journal of Speech and Hearing Research* (12), pp. 423-425, 1969.
- [7] Guiard-Marigny, T., Benoît, C., Ostry, D. J., "Speech Intelligibility of Synthetic Lips and Jaw", *Proceedings of the 3rd International Congress on Phonetic Sciences*, Sweden, 1995.
- [8] C. Benoît, "On the Production and the Perception of Audio-Visual Speech by Man and Machine", in Y. Wang et al. (eds.), *Multimedia & Video Coding*, New York, 1996.
- [9] University of Edinburgh, Centre for Speech Technology Research, Festival Speech Synthesis System, <http://www.cstr.ed.ac.uk/projects/festival/>.
- [10] Universität Bonn, Institut für Kommunikationsforschung und Phonetik, Speech Synthesis System HADIFIX, <http://www.ikp.uni-bonn.de/~tpo/Hadifix.en.html>.



- [11] Faculté Polytechnique de Mons, Circuit Theory and Signal Processing Lab, MBROLA Project, <http://tcts.fpms.ac.be/synthesis/mbrola.html>.
- [12] The Web3D Consortium, VRML - Virtual Reality Modeling Language, [http://www.web3d.org/technical/info/specifications/ISO IEC 14772-All](http://www.web3d.org/technical/info/specifications/ISO_IEC_14772-All).
- [13] The Web3D Consortium, H-Anim 2001 - Humanoid Animation Specification, <http://www.h-anim.org/Specifications/H-Anim2001>.
- [14] The Apache Software Foundation, PHP - Hypertext Preprocessor, <http://www.php.net>.