MASSY - a Prototypic Implementation of the Modular Audiovisual Speech SYnthesizer

Sascha Fagel
Technical University Berlin
E-mail: sascha.fagel@tu-berlin.de

ABSTRACT

Audiovisual speech synthesis systems usually are inflexible with respect to the ability to replace the audio and video synthesis and the control algorithms due to the dependencies of the implemented pieces. In order to enable a newly developed system to exchange modules, to evaluate their specific advantages, and to detect their weak points, the author proposes a framework for audiovisual speech synthesis systems which divides the system into several modules and describes their information flow [7]. This paper presents MASSY, the first prototypic implementation of the framework. Besides the embedded audio synthesis, the presented implementation includes a phonetic articulation module, a visual articulation module, and a face module. The visual articulation module implements two alternative models based on a dominance model for co-articulation in terms of Löfqvist's suggestion [10][3] and a pattern selection algorithm, respectively. The realized face is a 3D model described in VRML 97 [16] with additionally implemented functionality according to the H-Anim 2001 standard. The facial animation is described in a motion parameter model which is capable to realize the most important visible articulation gestures [4][1]. MASSY is developed in the client-server paradigm, where 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 standard, non-proprietary plug-ins. The presented system is suitable for the evaluation of measured or predicted articulation models, as well as 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, or virtual agents.

1. MOTIVATION

Human-computer-interfaces might be improved by speech output. But research has shown that this advantage decreases by growing level of abstraction from natural speech [2]. Human speech communication consists of several information streams. Thus a coherent presentation of audible and visible speech like provided by MASSY should enhance several quality parameters compared to audio only presentation. This includes not only intelligibility and comprehension which can be improved by visible articulation [15][6][9], but also naturalness and the transmission of non-verbal information might be advanced by facial expression [5]. In addition, speech synthesis is an appropriate instrument for perception experiments, because - compared to natural speech - every variable is strictly under control. At least for these reasons, audiovisual speech synthesis is worth to be investigated.

2. MODULE INTERFACES

The presented implementation of the framework is realized in the server-sided scripting language php (a project of the Apache Software Foundation, [13]) which is especially suited for web development. Figure 1 shows the system architecture.

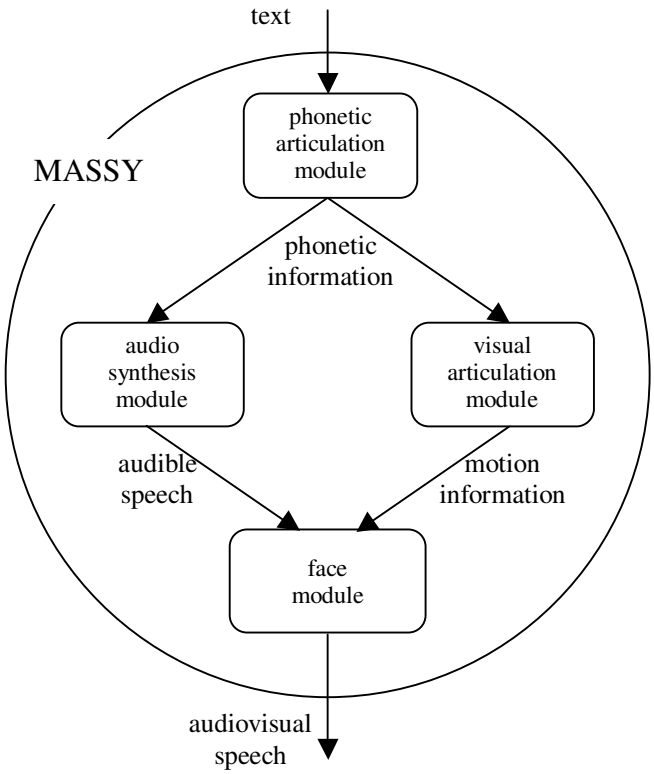

Figure 1: Schematic system overview of MASSY.

A plain text file serves as system input. The phonetic articulation module creates an appropriate phoneme chain and furthermore the prosodic information phoneme and pause durations and fundamental frequency curve. From this data, the audio synthesis module generates the audio

phonetic articulation module:

Name	Text2Pho()
Description	Transcribes a text to an extended phonetic representation.
Input	• the name and path of a plain text file (HADIFIX notations permitted for German) • a gender (male or female), • a language (currently de_DE or en_US)

Output	• an array with the following dimensions:
	  • [0..n] one for each phoneme of the utterance:
	  • ["Phoneme"] contains a phoneme (SAMPA-Notation)
	  • ["Duration"] contains the length of the phoneme in ms
	  • ["Audio"] contains positions in percent of the phoneme duration used as indices:
	  • [position] contains audio parameters as indices:
	  • ["F0"] contains the value of the fundamental frequency

visual articulation module:

Name	AddMotionParameters()
Description	Generates motion values for articulation.
Input	• an array with phonetic information as described in the phonetic articulation module • an articulation model type (pattern or dominance) • optionally a name and path of a model file

Output	• the array with phonetic information with added motion values
	  • for each phoneme:
	  • ["Visual"] contains positions in percent of the phoneme duration used as indices:
	  • [position] contains motion parameters as indices:
	  • ["LipW"] lip spreading / narrowing
	  • ["LowerJawH"] jaw opening
	  • ["LipH"] lip opening /closing height
	  • ["LowerLip"] lower lip retraction
	  • ["TongueTipH"] tongue tip height
	  • ["TongueBackH"] tongue back height
	• an array with the names of the motion parameters
	(these two return arrays are packed into one)

audio synthesis module:

Name	Pho2Wav()
Description	Writes an audio file.
Input	• an array with phonetic information as described in the phonetic articulation module • a compress flag • a gender • a language • optionally the name of a preferred voice
Output	• the name and path of the written audio file

face module:

Name	AddMotionParameters()
Description	Generates an animation and writes it into a multimedia file.
Input	• an array with motion information as described in the visual articulation module • the name and path of an audio file • a compress flag • the name of a face model (currently VRML) • an array with the names of the motion parameters
Output	• the name and path of the written multimedia file

Table 1a-d: Description of the Functions, that implement the modules and their interfaces. The indentation depths of a list (tree view) represent the dimensions of an array.

signal. The visual articulation module adds motion information, which is used by the face module to create a facial animation. The face module also integrates the audio signal into the animation. Table 1a-d describes the module interfaces.

3. PHONETIC ARTICULATION MODULE

The phonetic articulation module embeds a German and an English female and male text to phoneme conversion. The German part is realized by the integration of the high level speech synthesis part of HADIFIX [14], a speech synthesizer of the University of Bonn. The English transcription is using the high level synthesis of the festival speech synthesizer [8] of the University of Edinburgh.

4. AUDIO SYNTHESIS

MASSY's audio synthesis module wraps the MBROLA speech synthesis algorithm [11] of the Polytechnic Faculty of Mons.

5. VISUAL ARTICULATION MODULE

5.1 Dominance model

Phonemes are realized differently depending on the specific context of potentially all preceding and following phonemes. The implemented visual articulation model calculates the real target position of each phoneme based on fictitious ideal positions of each articulator and the strength (the dominance) to control this articulator. As simplification, only one parameter expresses the influence on neighbouring phonemes and the susceptibility to neighbours. Additionally the right sided and left sided dominance are assumed to be equal. Furthermore, as the quasi-stationary phases, the target positions are always held for a fixed fraction (currently 60%) of the phoneme duration centred in the phoneme. All these restrictions are part of the implementation of the visual articulation model, not of the framework. A different or more detailed model may be more exact. The target value of an articulator for a phoneme is calculated from the phoneme's ideal value of this articulator and its dominance on it and these of all neighbours by equation (1),

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

where I is the ideal value, D is the dominance, the index n means the current phoneme, indices greater n mean the phonemes on the right side, and indices smaller n mean the phonemes on the left side. The dominance is a value between 0 and 1. The ideal value of the current phoneme is weighted by the dominance of the current phoneme. The rest ($1-D_n$) is taken half by half from both adjacent phonemes weighted by their dominances and so on. If no abort criterion is used, the denominator equals 1, otherwise it normalizes all weights. The sum in the numerator for the left or right influence is implemented recursively as described in pseudo code:

```
(influence, dominance_residual) :=
if (NOT_END_OF_PHONEMESEQUENCE)
begin
influence +=
    dominance_residual * this_dominance
    * this_ideal_position ;
dominance_residual *= (1 - this_dominance) ;
influence += next_influence() ;
end
```

5.2 Pattern selection

An alternative visual articulation model implements a pattern selection algorithm. Phonemes that are visually identical are grouped to visemes. The real target positions of all consonantic visemes in the context of each vocalic viseme (and vice versa) are stored in a pattern database. For each viseme the most similar patterns regarding the last and next vowel in case of consonant, or the last and next consonant in case of vowel, respectively, are selected. The stored target positions of the selected patterns are averaged weighted with their distance from the current viseme.

6. FACE MODULE

By using the motion information generated by the visual articulation module, the face module dynamically generates an animation of a 3D scene and writes it into a VRML file.

Two modes are supported: In the coordinate interpolation mode, the motion information of each target position is converted to a complete set of points. This procedure shifts the processing effort from the client to the server, but increases the required bandwidth. In the displacer mode, the deformators (called displacers) of the scene, that describe the effect of a single articulator each, can be sent to the client. Then, only the amounts of displacements, that are needed to reach the complete articulation position for each phoneme, have to be transferred. The latter mode minimizes the used bandwidth. Additionally the size of the transferred file increases only marginally with longer utterances synthesized. Playing the animation in displacer mode currently is realized by a slow client-sided VRML-script. But as displacers are part of the H-Anim 2001 standard, the next generation of browser plug-ins will support this feature at much higher performance.

7. FUTURE WORK

A female speaker's articulation (as the currently animated face is female) will be measured by electromagnetic mid-sagittal articulography (EMA) and video analysis. The acquired data will be used to approximate a dominance model and a pattern selection model, respectively. A perception experiment shall be carried out to adjust the magnitude of articulation movements for maximum visual intelligibility. The relative length of the quasi-stationary phase per phoneme and per motion parameter and other effects of hypo- and hyperarticulation shall also be investigated.

A 2D synthetic face implemented in flash and an image based 2D video realistic model are planned. Furthermore, the 3D VRML face shall be extended with exchangeable face topologies and textures based on the description of faces standardized by MPEG-4 FDPs (facial definition parameters, [12]).

The framework for audiovisual speech synthesis systems, on which the presented implementation is based, yield an audio and a visual expression module. These modules are planned to be developed. They shall implement a three-dimensional emotion model (based on the approved dimensions valence, arousal, and potency) and furthermore a basic emotion model (using the well known basic emotions fear, anger, happiness, and sadness).

The phonetic articulation module generates information, that is written into a MBROLA-compliant .pho file. This file format supports only phoneme chains and the prosodic information of rhythm (durations of phonemes and pauses) and intonation (F0 values). Other data needed for emotional speech like voice quality parameters currently is ignored by the audio synthesis module. In future, this could be taken into consideration e.g. by post-processing or by using articulatory or formant synthesis. The interface specification then will have to be adapted.

All modifications and extensions will be evaluated in audiovisual speech perception experiments.

REFERENCES

[1] C. Benoît, C. Abry, M. A. Cathiard, T. Guiard-Marigny, T. Lallouache. "Read my Lips: Where? How? When? And so ... What?", In B. Bardy, R. Bootsma, Y. Guiard (eds.), Poster Book of the 8th International Congress on Event Perception and Action. France: 1995.

[2] 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: Plenum Press, 1996.

[3] M. M. Cohen, D. W. Massaro, "Modeling Co-articulation in Synthetic Visual Speech", In N. Magnenat Thalmann & D. Thalmann (eds.), Models and Techniques in Computer Animation. Tokyo: Springer-Verlag, 1993, pp. 139-156.

[4] M. M. Cohen, D. W. Massaro. "Development and Experimentation with Synthetic Visual Speech", Behavior Research Methods, Instruments and Computers (26). 1994, pp. 260-265.

[5] P. Ekman. "Methods for Measuring Facial Action", In K. R. Scherer, P. Ekman (eds.), Handbook of Methods in Nonverbal Behavior Research. Cambridge: Cambridge University Press, 1982, pp. 45-90.

[6] N. P. Erber. "Interaction of Audition and Vision in the Recognition of Oral Speech Stimuli", Journal of Speech and Hearing Research (12). 1969, pp. 423-425.

[7] S. Fagel, W. F. Sendlmeier. "Entwurf eines Frameworks für audiovisuelle Sprachsynthese-systeme", Tagungsband der 13. Konferenz Elektronische Sprachsignalverarbeitung. Dresden: Universitätsverlag, 2002, pp. 372-378.

[8] The Festival Speech Synthesis System: http://www.cstr.ed.ac.uk/projects/festival/

[9] T. Guiard-Marigny, C. Benoît, D. J. Ostry. "Speech Intelligibility of Synthetic Lips and Jaw", Proceedings of the 3rd International Congress on Phonetic Sciences. Sweden: 1995.

[10] A. Löfqvist. "Speech as Audible Gestures", In W. J. Hardcastle, A. Marchal (eds.), Speech Production and Speech Modeling. Dodrecht: Kluwer Academic Publishers, 1990.

[11] The MBROLA Project: http://tcts.fpms.ac.be/synthesis/mbrola.html

[12] MPEG - Moving Picture Experts Group: http://mpeg.telecomitalialab.com

[13] PHP: Hypertext Preprocessor. http://www.php.net

[14] Speech Synthesis System HADIFIX: http://www.ikp.uni-bonn.de/~tpo/Hadifix.en.html

[15] W. H. Sumby, I. Pollack. "Visual Contribution to Speech Intelligibility in Noise", Journal of the Acoustical Society of America (26). 1954, pp. 212-215.

[16] VRML - Virtual Reality Modeling Language: http://www.vrml.org