A Comparison of German Talking Heads in a Smart Home Environment

Sascha Fagel [1], *Christine Kuehnel* [2], *Benjamin Weiss* [2], *Ina Wechsung* [2], *Sebastian Moeller* [2]

[1] Berlin Institute of Technology
[2] Deutsche Telekom Laboratories, Berlin Institute of Technology

sascha.fagel@tu-berlin.de,
[christine.kuehnel,benjamin.weiss,ina.wechsung,sebastian.moeller]@telekom.de

Abstract

The authors describe a newly developed German Text-To-audiovisual-Speech (TTavS) synthesis system based on the English speaking HeadZero. Targets of the control parameters of the talking head are generated by mapping of German phonemes to the originally English visemic blend shapes controls. The resulting German version of HeadZero and the German talking head MASSY were extended to generate audiovisual speech utterances from text with voices of both the MARY and the MBROLA audio speech synthesizers. A test was designed to evaluate the quality of the talking heads combined with the two synthetic voices in the context of a smart home environment. The results show a significant user preference for the new German HeadZero. Both heads are rated better when combined with the MARY voice.

Index Terms: audiovisual speech synthesis, talking heads, smart home environment

1. Introduction

The quality of synthetic speech is affected by functional aspects such as intelligibility and aesthetic aspects such as naturalness or pleasantness [1]. Functional aspects and aesthetic aspects can be regarded as nearly orthogonal, i.e. a system can be very natural and completely unintelligible or highly intelligible while obviously synthetic. Both aspects can only be determined by investigating the user's perception.

An embodied agent that displays appropriate non-verbal behavior can improve user satisfaction and engagement and enhance the interaction with a computer system [2]. In case of the transmission of verbal information from a computer to a human, i.e. speech output, a synthetic face can increase the intelligibility and enhance the robustness of the information transmission [3] as known from natural speech [4]. But incoherent audiovisual speech, i.e. when the voice does not mach the face with respect to phonemes/visemes [5], timing [6], expression [7][8] and age [9], the face may affect the perception of the voice, e.g. decrease the intelligibility or change the perceived emotion to one that was not intended in either the voice or the face. The application in which a talking head (an audiovisual speech synthesizer) is used may severely change its requirements and hence the user's judgment of the talking head.

2. The German HeadZero

In the original English version of HeadZero each control parameter is implemented in the form of a blend shape to activate a certain viseme (i.e. a set of phonemes that are visually nearly undistinguishable). As there is no one-to-one mapping from English to German visemes an activation setting as target position was manually created for each German phoneme.

The trajectories between each pair of neighboring phones are generated at a frame rate of 30fps by a transition of fixed length (here: one frame of 33ms) and a stationary phase that fills the phone duration. As the mixture of more than one viseme always led to disturbing artifacts – this kind of control was obviously not intended by the blend shape animation of the original version of HeadZero – only each two neighboring visemes were blended. This is considered to results in insufficient modeling of co-articulation. Only 8 out of 16 possible visemes were used as the others either did not represent a German viseme or produced unintentional artifacts. All visemes except one (that looked over articulated at maximum value) were used with a magnitude of 100% at the stationary phase. The intensity of a viseme in the transition from the preceding viseme was set to 50% to account for anticipatory co-articulation. The intensity of a viseme was set to 66% in the transition to the next viseme as carry-over co-articulation. Figure 1 shows trajectories for HeadZero generated for the first sentence, figure 2 (see page 3) shows HeadZero.

3. MASSY

MASSY is a parametric 3-D talking head. See [10] for a detailed description. Its controls are articulatory parameters – lip spreading/rounding, jaw opening, vertical lip opening, tongue tip height, tongue body height, tongue dorsum advance, and velum height – which are a complete set of basic movements (except voiced/unvoiced excitation) to display all German phonemes. Trajectories of these parameters over time are generated for the utterance (i.e. the phone chain) as a whole by a modified version of the dominance model of Cohen and Massaro [11]. Co-articulated targets are calculated from ideal targets (e.g. in isolation) by mutual influence of potentially all phonemes of an utterance. A target of a phoneme is represented by a setting of all articulatory targets. Hence, in contrast to the visemic control of HeadZero, all virtual articulators are permanently incorporated instead of "activating" one or two at a time. Figure 3 shows trajectories generated for the first sentence.

As HeadZero is male, the texture and the geometry of the default (female) head of MASSY were adjusted to look more male. Freckles were removed from the texture and the eyebrows were strengthened. The chin and other face features were made more chiseled and the hair was shortened and darkened. Figure 4 (see page 3) shows the new synthetic male head of MASSY.

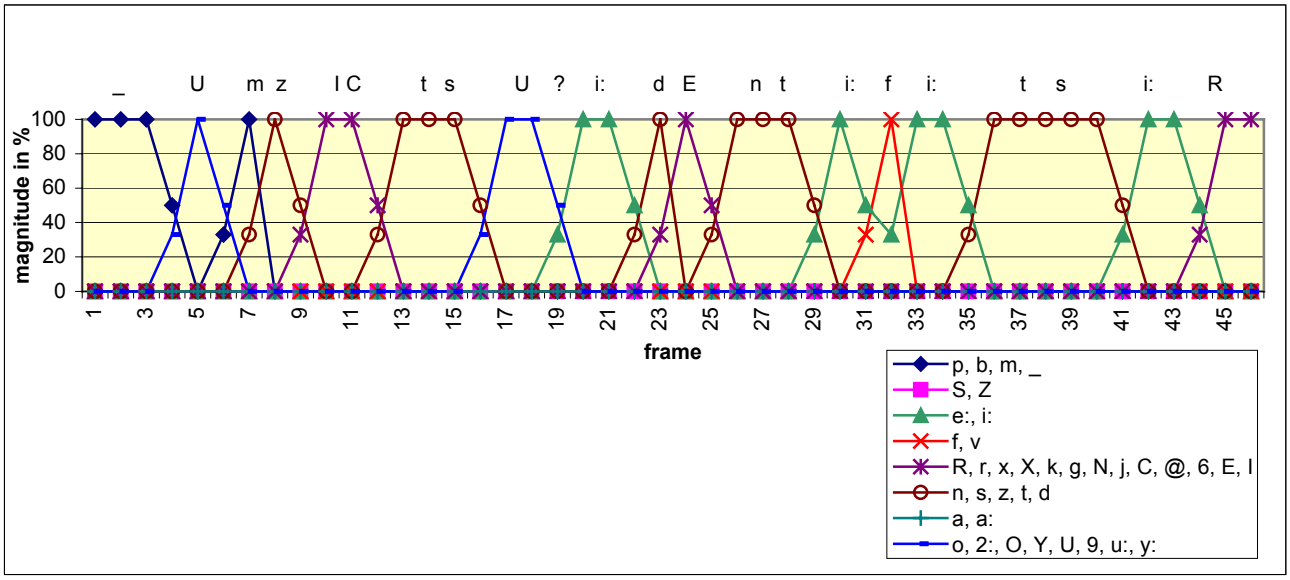

Figure 1: *Visemic control parameters for the first sentence for HeadZero for the initial part of "Um sich zu identifizieren ...".*

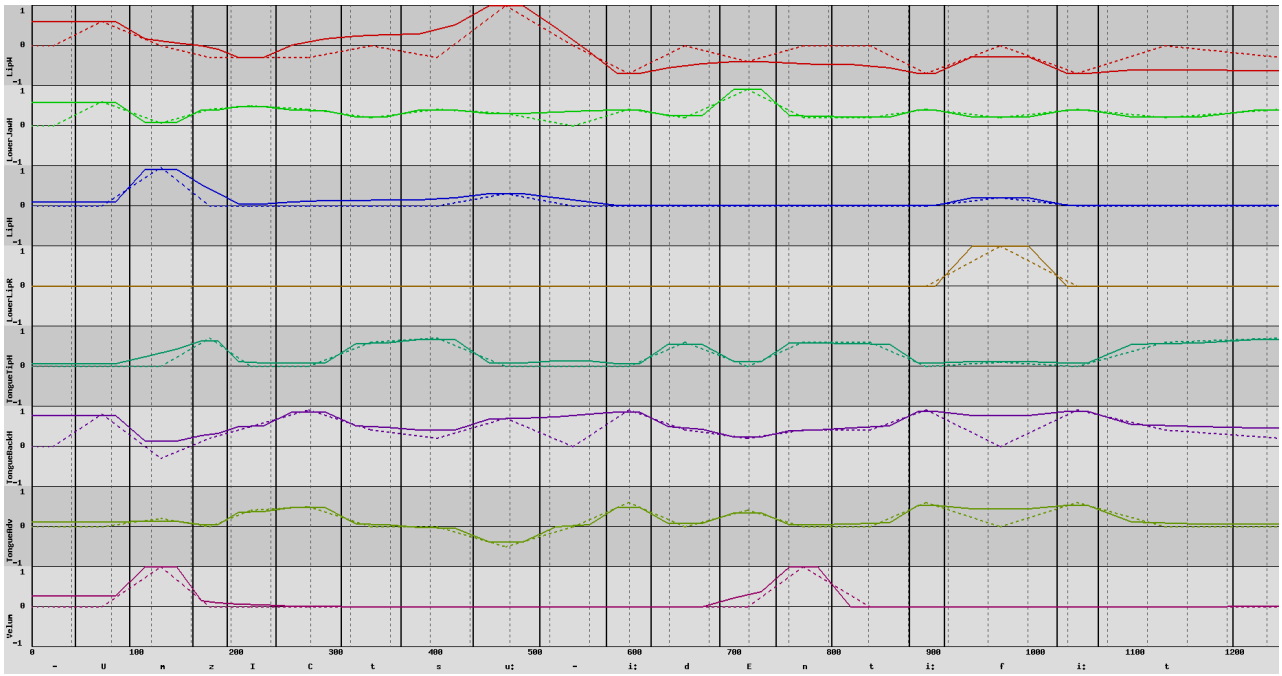

Figure 2: *Articulatory control parameters for MASSY's virtual articulators for the initial part of "Um sich zu identifizieren ...". Articulatory parameters from top to bottom: lip spreading/rounding, jaw opening, vertical lip opening, tongue tip height, tongue body height, tongue dorsum advance, velum height. Dashed lines show linear interpolations between ideal (not co-articulated) target positions, solid lines show co-articulated trajectories which are used in the synthesis.*

4. Evaluation

4.1. Subjects

Overall 14 subjects participated in the test (seven female and seven male, aged between 20 and 32 years, mean = 27 years, stdev = 4.21). One of them produced an immoderately high number of outliers and another one stated after the test that he did not follow the instructions. Therefore, data from only 12 subjects were analyzed. The subjects were paid for their participation.

Figure 3: *HeadZero.*

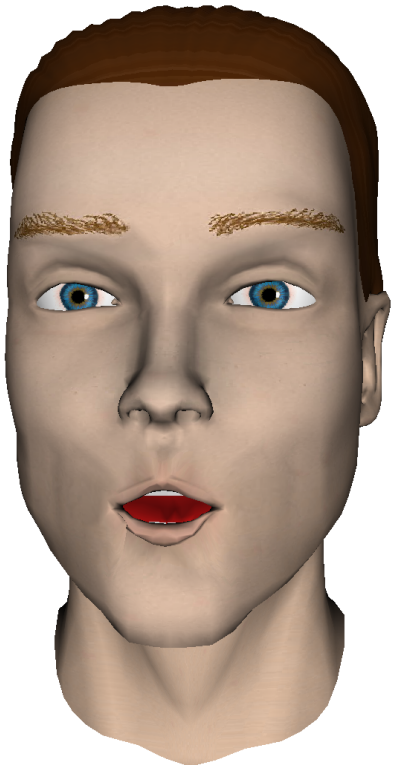

Figure 4: *The male head of MASSY.*

4.2. Stimuli

Stimuli were synthesized for the INSPIRE domain, i.e. possible speech output of a system to control domestic appliances, by producing ten different sentences like "Um sich zu identifizieren, nennen Sie bitte Ihren vollständigen Vor- und Zunamen!" ("To identify yourself, please give your first name and your last name!"). These sentences vary in length to receive one phrasal and two phrasal stimuli and include questions and statements. The sentences were synthesized by the German version of HeadZero, the talking head MASSY and a third TTavS system (based on 3-D speaker cloning; [12]) which was not included in the analysis. All heads were combined with both the MBROLA [13] de2 synthetic voice and the male HMM-bits3 voice of MARY [14]. The prosodic parameters phone durations and F0 contour were generated by HADIFIX [15] for the MBROLA voice and by MARY itself for the MARY voice. Phone durations were passed to the synthetic faces in order to generate lip synchronous facial animations.

4.3. Presentation

Three presentation schemes were used in the evaluation. In the first part of the experiment all stimuli were presented to the subjects in random order. Each stimulus was rated concerning overall, visual and (audio) speech quality on a five point scale. In addition, a context question had to be answered to keep the subjects concentrated on listening to what was spoken and not only on the impression of the talking head. In the second part, six stimuli in the same condition were presented blocked. Analyses of this part are not presented here. All conditions were rated this way in pseudo-random order. In the third part one sentence was presented in every condition and the subjects had to order the conditions concerning overall quality. The subjects were allowed to play the sentences as often as they wanted.

5. Results

Only results for HeadZero and MASSY are presented here. Concerning the overall quality assessed in the first part of the experiment (per-stimulus ratings padded by context questions), MARY is rated significantly better than MBROLA and HeadZero is rated better than MASSY (Tukey's post hoc test: $\alpha = .05$). The perceived overall quality was significantly affected by the sentence that was played to the subject and, as expected, by both the voice and the face (ANOVA: $p < .001$).

The speech quality assessed in the first part of the experiment was significantly affected by the sentence and the voice ($p < .001$) and, interestingly, by the head that was presented with the voice ($p < .05$).

When rating the faces (visual quality), one subject did not rate one of the heads better than the other one, seven subjects rated HeadZero consequently better than MASSY whereas four subjects rated MASSY consequently better than HeadZero. Overall, HeadZero was rated significantly better (Tukey's post hoc test: $\alpha = .05$) than MASSY. The rating of the face was not influenced by the voice or the sentence.

The outcomes of the forced ordering differed slightly from the results of the per-stimulus ratings. This part of the evaluation reveals a statistically significant (Wilcoxon signed-rank test: $Z = -2.45$, $p = .007$) user preference – with only few exceptions – for the new German HeadZero. In this part of the experiment there was no significant preference for one of the two voices (MARY: mean rank = 3.72, stdev = .66; MBROLA: mean rank = 3.28, stdev = .66)

6. Conclusions

The ratings of the speech quality was influenced by the head that was displayed along. This cross-modal interaction shows that the subjects could not ignore the quality of the face when asked to rate the voice. This is coherent with earlier findings on the perceived age of faces and voices [8]. The opposite interaction did not occur revealing that – in this study – vision was the predominant modality for the judgment of a talking head in a smart home environment.

The influence of the sentence on the ratings of the speech quality and the absence of that influence on the visual quality show that the voice depends on the phonetics of the utterance where the face does not.

The German articulatory movements of HeadZero are assumed to be less adequate than those of MASSY. However, the texture of HeadZero taken from the Australian actor Stelarc (Stelios Arcadiou) is more sophisticated. The presentation scheme – that can be regarded as different parts of an application – seems to clearly affect the perceived quality of the talking head. In the forced ranking, the user preference for HeadZero shows that the (static) appearance of a talking head has an important impact on the users' impression even in smart home scenarios. The fact that MASSY was rated better in the part with context questions that is more sensitive to the linguistic content compared to the sentence playback with forced ranking suggests that the more the linguistic content is important the better a clear articulation is rated by users. As a consequence, a talking head that is embedded in a smart home environment should meet both requirements, appropriate static and dynamic appearance.

7. Acknowledgements

We thank Rob Looijmans and David van der Pol for their help in setting up, conducting and analyzing the experiment and we thank the Thinking Head team for providing their HeadZero.

8. References

[1] Sonntag, G. P., Evaluation von Prosodie, Aachen: Shaker Verlag, 1999.

[2] Foster, M. E., "Enhancing Human-Computer Interaction with Embodied Conversational Agents", Proceedings of the International Conference on Human-Computer Interaction, Beijing, 2007.

[3] Ouni, S., Cohen, M. M., Ishak, H. and Massaro, D. W., "Visual Contribution to Speech Perception: Measuring the Intelligibility of Animated Talking Heads", EURASIP Journal on Audio, Speech, and Music Processing, 2007.

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

[5] MacDonald, I. and McGurk, H., 1978, "Visual Influences on Speech Perception Process", Perception & Psychophysics, 24, 253-257, 1978.

[6] Grant, K. W., van Wassenhove, V. and Poeppel, D., "Discrimination of Auditory-Visual Synchrony", Proceedings of the International Conference on Audio-Visual Speech Processing, St. Jorioz, 31-35, 2003.

[7] de Gelder, B. and Vroomen, J., "The Perception of Emotions by Ear and by Eye", Cognition & Emotion, 14(3), 289-311, 2000.

[8] Fagel, S., "Emotional McGurk Effect", Proceedings of the International Conference on Speech Prosody, Dresden, 2006.

[9] Fagel, S., "Auditory-Visual Integration in the Perception of Age in Speech", Proceedings of the International Conference of Phonetic Sciences, Saarbrucken, 2007.

[10] Fagel, S. and Clemens, C., "An Articulation Model for Audiovisual Speech Synthesis - Determination, Adjustment, Evaluation", Speech Communication, 44, 141-154, 2004.

[11] Cohen, M. M. and Massaro, D. W., "Modeling Coarticulation in Synthetic Visual Speech", in: N. M. Thalmann, D. Thalmann (Eds.) Models and Techniques in Computer Animation, 139-156, 1993.

[12] Fagel, S., Elisei, F. and Bailly, G., "From 3-D Speaker Cloning to Text-to-Audiovisual-Speech", Proceedings of INTER-SPEECH, Brisbane, 2008.

[13] The MBROLA Project, URL http://tcts.fpms.ac.be/synthesis/mbrola.html, 2005.

[14] Schröder, M. and Trouvain, J., "The German Text-to-Speech Synthesis System MARY: A Tool for Research, Development and Teaching", International Journal of Speech Technology, 6, 365-377, 2003.

[15] Portele, T., "Das Sprachsynthesesystem Hadifx", Sprache und Datenverarbeitung, 21, 5-23, 1997.