Audio-visual quality as combination of unimodal qualities: environmental effects on talking heads

Benjamin Weiss, Christine Kühnel, Sebastian Möller

Quality & Usability Lab, Telekom Laboratories, TU Berlin, 10587 Berlin, E-Mail: BWeiss@telekom.de

Introduction

Talking heads provide a multimodal output component for human-computer-interfaces. They consist of facial visual models that are synchronized with speech synthesis modules concerning speech articulation. Due to their reduction to a human head or upper body, articulation is often more clearly visible compared to a full human body due to the possibly bigger display of the head. Therefore, talking heads are especially suited for applications like robust speech understanding and language acquisition. Evaluation is typically concerned with function test to assess the synthesis quality with e.g. metrics like word error rate of human listeners or perceived naturalness (cf. [8]). But as more and more talking heads are used as interfaces for speech-based dialogue systems and are enhanced with facial expressions, the overall quality experienced by the user is in scope.

The topic addressed in this paper is the relationship between the modalities audio and vision in terms of perceived quality and their impact on perceived overall talking head quality. The focus lies on the goodness of fit of models, describing the multimodal quality of the talking head as a linear combination of ratings from single modalities, i.e. visual and speech. Another aim is to assess the relevance of two important factors emerging from using talking heads in real applications: The degree of interactivity and distraction from the talking heads originating e.g. from other system output. Such models are already used in the domain of audio-visual quality for IP-based transmission systems like IP-TV and videotelephony (cf. [1,6]) and multimodal interactive systems without embodiment [9].

Method

Overall quality of six different talking heads has been evaluated in four different settings: a passive rating scenario in our laboratory (14 subjects), a passive web-experiment (42 subjects), an interaction scenario with a talking head interface only (24 subjects), an interaction scenario with a second screen showing information in addition to the talking head (different 24 subjects), and an interaction scenario in a real living room instead of a test laboratory, also with an additional output screen (22 subjects).

Quality was rated on a 5-point scale to assess user's subjective perception. Except for the last test in a real living room, pre-recorded videos of the talking heads were presented to the subjects. Ratings of visual quality, speech quality and overall quality were assessed after each sentence (in the passive scenario), each of two task blocks for every condition (in the interaction scenarios) or after each condition (in the real living room).

Material

The first head (TH) originates from the Thinking Head Project [2]. This head is based on a 3D model of a human being, in this case the Australian artist STELARC. In addition to having a humanlike texture build from pictures of STELARC, it exhibits random head movements and extra-linguistic facial expressions like smiling and winking. As the control of the visual articulation was built for English and does not define separate phonetic articulators (like lip spreading or jaw opening), but target visemes, a German synthesis was made by hand using the most appropriate English visemes and applying basic co-articulation rules from Massy (see below). The original English visemes were created from motion-capture data. The two following head components do not exhibit facial expressions or movements apart from visual articulation. The second head was developed at TU Berlin: Massy (MS), the **M**odular **A**udiovisual **S**peech **Sy**nthesizer is a parametric 3D head model and provides accurate audio-visual speech synchronization and includes articulators like the velum and tongue body which are not always visible. MS also accounts for co-articulation with rules based on empiric data of German [7]. The third head is a 2D German Text-To-Audiovisual-Speech synthesis system based on speaker cloning (CL) using motion capture. The coarticulation behaviour was extracted from the videos. CL was developed in cooperation between TU Berlin and GIPSA-lab Grenoble [4]. Pictures of the three head components are displayed in Fig. 1. Because of the low quality scores obtained in the first experiment, Clone was only used in the first passive scenario.

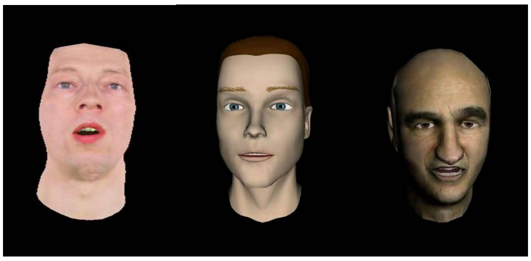

Figure 1: The facial models used (left to right): CL, MS, TH

The speech synthesis systems producing the respective voices include the **M**odular **A**rchitecture for **R**esearch on speech s**Y**nthesis (Mary) [5] and the Mbrola system (Mbrola) [3]. A male German voice was selected for both systems, namely 'hmm-bits3' for Mary and 'de2' for Mbrola. Both were considered best from a selection of the two TTS-systems in an earlier informal listening test.

Results

As in the passive scenario every single sentence was randomly presented and rated, there are a lot of data points compared to the non-passive ones. Therefore, for each scenario mean ratings were computed over all conditions tested for all subjects to be comparable; such that the final values represent all subjects and talking heads tested:

- *passive (laboratory)*: averaged over 10 different sentences, resulting in 84 values: 14 subjects, 3 visual models, 2 TTS modules
- *passive (web)*: averaged over 6 different sentences, resulting in 168 values: 42 subjects, 2 visual models, 2 TTS modules
- *interactive (1 screen)*: averaged over 2 different task blocks, resulting in 96 values: 24 subjects, 2 visual models, 2 TTS modules
- *interactive (2 screen)*: averaged over 2 different task blocks, resulting in 96 values: 24 subjects, 2 visual models, 2 TTS modules
- *interactive (living room)*: not averaged, 88 values: 22 subjects, 2 visual models, 2 TTS modules

Although there are differences in the data used due to the varying scenarios, the results of the regression analysis can be compared, see Table 1 for the linear models and the goodness of fit (Pearson's r and root-mean-squared-error for the error of prediction (Ep)). Please note that excluding the CL visual model from the stimuli does not result in major differences of the models fit (passive lab vs. passive web). It just increases the impact of the speech modality, as the better fit of the model might also been explained by the number of ratings.

There are two major findings: Firstly, with interaction and increasing distraction due to the additional screen and the environment of the living room, the models fit decreases: The error of prediction increases, R decreases, and the constant (the models offset) becomes higher.

Secondly, the relevance of the visual quality is stronger than the speech quality for both scenarios with two screens, which is opposed to the results from the other scenarios.

Table 1: Results of linear regression analysis: Model description (factors for single modalities, the offset) Pearson's r and Ep of the estimated multimodal quality, number of data points (N).

Scenario	Visual	Speech	Offset	R	Ep	N
passive lab	.37	.44	0.64	.85	.37	84
passive web	.30	.57	0.39	.90	.30	168
interaction 1 screen	.29	.46	0.91	.78	.41	96
interaction 2 screens	.34	.29	1.48	.53	.48	96
living room	.40	.26	1.31	.57	.62	88

Conclusion

Results show that in the passive scenario talking head quality can be described quite well as a linear combination of visual and auditory aspects. However, the more distraction the environment offers, the less variation in the ratings is explained by visual and speech quality. Of course, other modelling approaches might estimate the assessed talking head ratings better. It is, however, more relevant and interesting to further study environmental effects on user quality ratings; in particular to get insight into cognitive processes of assessment. For example, it is unclear, if the results presented here are caused by distraction during the perception process or rating process and thus represent a reduced ability in assessment or in separating talking heads from the overall system.

Literature

[1] Belmudez, B., Möller, S., Lewcio, B., Raake, A., and Mehmood, A.: Audio and video channel impact on perceived audio-visual quality in different interactive contexts. In *Proc. IEEE Int. Workshop on Multimedia Signal Processing* (MMSP 2009)

[2] Burnham, D., Abrahamyan, A., Cavedon, L., Davis, C., Hodgins, A., Kim, J., Kroos, C., Kuratate, T., Lewis, T., Luerssen, M., Paine, G., Powers, D., Riley, M., Stelarc, Stevens, K.: From talking to thinking heads: 2008. In: *Proc. International Conference on Auditory-Visual Speech Processing* (AVSP 2008)

[3] Dutoit, T., Pagel, V., Pierret, N., Bataille, F., Van der Vreken, O.: The MBROLA project: Towards s set of high-quality speech synthesizers free of use for non-commercial purposes. In: *Proc. International Conference on Spoken Language Processing* (ICSLP 1996), 1393–1396

[4] Fagel, S., Bailly, G., Elisei, F.: Intelligibility of natural and 3d-cloned German speech. In: *Proc. International Conference on Auditory-Visual Speech Processing* (AVSP 2007), Paper L2-1.

[5] Fagel, S., Clemens, C.: An articulation model for audiovisual speech synthesis – determination, adjustment, evaluation. *Speech Communication 44* (2004), 141–154

[6] Garcia, M. and Raake, A.: Impairment-factor-based audio-visual quality model for IPTV. In *Proc. 1st Int. Workshop on Quality of Multimedia Experience* (QoMEX 2009).

[7] Schroeder, M., Trouvain, J.: The German text-to-speech synthesis system MARY: A tool for research, development and teaching. *International Journal of Speech Technology 6* (2003), 365–377

[8] Theobald, B.-J., Fagel, S., Bailly, G., Elisei, F.: Lips2008: Visual speech synthesis challenge. In: *Proc. INTERSPEECH* (2008), Brisbane, 2310–2313

[9] Wechsung, I., Engelbrecht, K.-P., Nauman, A., Schaffer, S., Seebode, J., Metze, F., and Möller, S.: Predicting the quality of multimodal systems based on judgements of single modalities. In *Proc. INTERSPEECH* (2009), 1827–1830