German Text-to-Audiovisual-Speech by 3-D Speaker Cloning

Sascha Fagel [1], *Gérard Bailly* [2]

[1] Berlin Institute of Technology,
[2] GIPSA-lab, Grenoble

sascha.fagel@tu-berlin.de, gerard.bailly@gipsa-lab.inpg.fr

Abstract

Visible speech movements were optically motion captured and parameterized by means of a guided PCA. Co-articulated consonantal targets were extracted from VCVs, vocalic targets were extracted from these VCVs and from sustained vowels. Targets were selected or combined to derive target sequences for phone chains of arbitrary German utterances. Parameter trajectories for these utterances are generated by interpolating targets through linear to quadratic functions that reflect the degree of co-articulatory influence. Videos of test words embedded in a carrier sentence were rendered from parameter trajectories for an evaluation in the form of a rhyme test in noise. Results show that the synthetic videos – although intelligible only somewhat above chance level when played alone – significantly increase the recognition scores from 45.6% in audio alone presentation to 60.4% in audiovisual presentation.

Index Terms: talking head, intelligibility, evaluation, trajectory generation

1. Introduction

Visible speech movements that match audible speech increase the intelligibility [1]. This advantage can be reproduced by synthetic video [2, 3] although the perceptual relevance of all properties of the visualization are not yet completely understood. A playback of separately generated audio and video synthesis synchronous at phone level often leads to reasonable results [4].

There are several approaches to the synthetic visualization of speech movements. Besides data-based systems that use pre-recorded video images [5] or sequences [6] one major method is the manipulation of a 3-D object that represents the geometric properties of a face or a head. The positions of all vertices have to be defined over the time of the synthesized utterance. Commonly not every vertex is controlled separately but a limited number of parameters is used to control the vertex positions. Of whatever kind these controls are – muscle activations [7], purely statistic components [8] or articulatory motivated parameters [9] – an adequate series of control values has to be generated for the synthesis of a previously unknown utterance in case of a TTavS (text-to-audiovisual-speech) system. These control values can be generated by rule [10][11], by statistical methods such as HMM [12] or data-based. The method proposed here is mainly a member of the latter group.

2. 3-D Speaker Cloning

A native speaker of German was filmed from three views with 398 colored beads glued on the face. The speaker uttered 100 symmetric VCVs composed of V = {a,i,u,E,O} and C = {p,b,t,d,k,g,f,v,s,z,S,Z,C,j,x,R,m,n,N,l}. Additionally the vowels {a,e,i,o,u,2,y,E,I,O,U,6,Y,@} and 9 were uttered short-time sustained in isolation. Six main articulatory parameters were extracted by an iterative modeling procedure – a so-called guided PCA [9]. Marker positions were taken from the center frames of 42 captured sequences: the 15 sustained vowels and 13 consonants each co-articulated in VCVs with V = {a,i,u} context [3]. These 42 visemes are selected to cover the articulatory space of all captured sequences. By iteratively defining regions of articulators to guide the PCA, the following articulatory parameters result:

- Jaw opening/closing
- Lip rounding/spreading
- Lip closing/opening (without jaw)
- Upper lip lift/drop
- Jaw advance/retraction
- Throat (tongue root) lowering

The synthetically reproduced VCV sequences have shown to yield approx. 70% of the intelligibility provided by the natural face. The error reduction (as known as audiovisual benefit [13]) was between 32.6% and 41.6%, depending on the signal-to-noise ratio in the audio channel (SNR = –6dB or SNR = 0dB, respectively). Figure 1 shows a screenshot of the synthetic display used in the evaluation (section 4). "Note that the upper jaw was added but in contrast to Fagel et al. [11] the inner contour of the lips and the inner vocal tract has not been modeled yet and is left untextured with background color and no tongue animation is included. The reduction of intelligibility compared to the natural face is in part due to this reduction in display.

3. Coarticulation modeling and TTavS

The parameter values to control the face are taken in principle from measurements (i.e. from data). As only VCV sequences and sustained vowels in isolation are used, a set of additional procedures are necessary to fill the missing data when generating parameter trajectories for an arbitrary utterance. The generation of parameter trajectories is done in two steps: target estimation and transition interpolation.

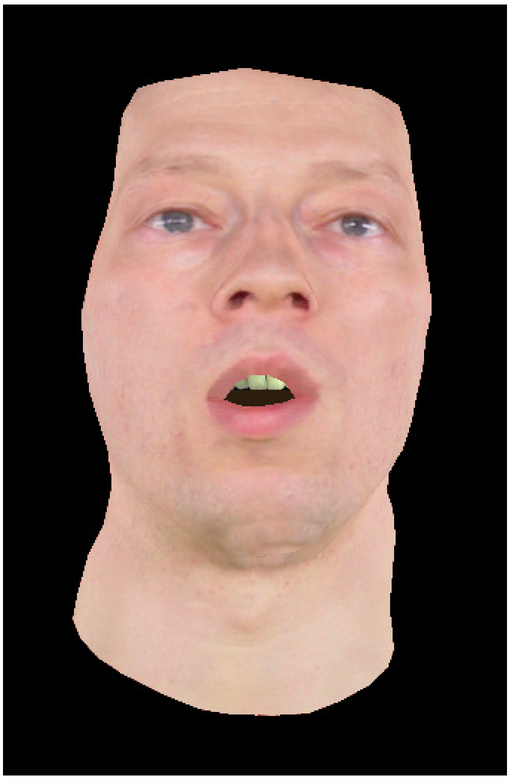

Figure 1: *Synthetic display generated from a mesh of the 398 marker positions and 30 points of an additional lip model, rendered with a static cylindric texture.*

For the determination of targets (sets of values of the six articulatory parameters) the phonetic contexts were differentiated.

1) If consonantal segments occur in asymmetric vocalic contexts: the parameters of the consonant in the two measured symmetric VCVs are linearly combined. In case of V≠{a,i,u,E,O}, i.e. the VCV of the consonantal segment and one of the context vowels does not exist in the motion capture database then a manually estimated mixture of the existing data is used (e.g. /I/ is estimated by taking a part of 0.8 from the according consonant appearing in /i/ context and 0.2 of the consonant in /E/ context).

2) If consonants occur in clusters: the influence of a context vowel decreases linearly with the distance in phones, i.e. given the distance of the consonant to the preceding vowel V_n is n and to the subsequent vowel V_m is m (where n = m = 1 means direct neighbors), then parameter values of the consonant in V_nCV_n are taken by m/(n+m) and from V_nCV_n by n/(n+m).

3) If only one context vowel exists due to an utterance boundary then a symmetric context is assumed and case 1) applies.

4) Targets for a vowel segment are taken

   a. from the VCVs if the vowel is between two identical consonantal visemes,

   b. from the vowel measured in isolation in case of pure vocalic context, or

   c. a balanced average of both if left context ≠ right context.

The transitions are linear to quadratic interpolations between the targets of each parameter. Like in the target estimation, consonantal targets are interpolated by regarding the nearest preceding and subsequent vowel. The exponent of the interpolation (1.0 to 2.0) is determined by the degree of coarticulation that occurred in the target estimation of consonants. Linear interpolation is used if the target completely adopts to the neighbors, i.e. if the parameter values of the consonant in the V_nCV_n and in the V_mCV_m equals that one of the respective context vowel. The linear interpolation leads to transient targets (a "pass through") and hence no quasi-stationary phase of the articulator in the consonantal segment. Quadratic interpolation is used if the target is not co-articulated at all, i.e. the parameter value of the consonant in one contexts equals that one of the parameter value of the consonant in the other context. Quadratic interpolation leads to a quasi-stationary phase of the articulator in the consonantal segment. The exponent of the interpolation is calculated to range from 1.0 to 2.0 between these extreme cases. Figure 2 shows the original and a synthesized parameter trajectories for the sequence /aba/.

4. Evaluation

A rhyme test with carrier sentence was carried out as evaluation of the system in terms of intelligibility. The carrier sentence "Ich sehe … heute" (English: "I see … today") was chosen for its short duration and for the preferably little co-articulatory effect of the phones before and after the test word (/@/ and /h/) on the test word. The test words were taken from Sendlmeier [16]: The test contains three lists of 40 monosyllabic words each, and one set of five answer alternatives (e.g. one test word of a list is "Nest" and the according answer alternatives are "Fest", "best", "Test", "Nest" and "Rest"). The words test word initial and word final consonants and medial vowels. The test is phonetically balanced.

Txt2Pho from HADIFIX [14] was used to generate phone sequences and durations (and f0 contour) from the sentences. The mbrola speech synthesizer [15] with the German voice *de2* was used for the audio synthesis. The phone durations from the HADIFIX output were used to generate the timing of the target sequence as described in section 3. Hence, Audio and video are synchronous on phone level when played back simultaneously. White noise at 0dB SNR was added to the audio channel. The test word was masked in the audio channel by white noise at 0dB SNR (maximum RMS power) with 200ms fade in and out before and after the word (during the carrier sentence). In the visual alone condition the carrier sentence was audible and only the test word was cut out of the audio channel. It was assured that the masking and the cutting covered as well the transition to the first phone of the test word and from the last phone of the test word. Figure 3 shows the signal masked with noise (used for audio alone and audiovisual stimuli), the signal masked with noise with the test word cut out (used for visual alone stimuli), and the noise signal. A black screen was displayed in the audio alone condition.

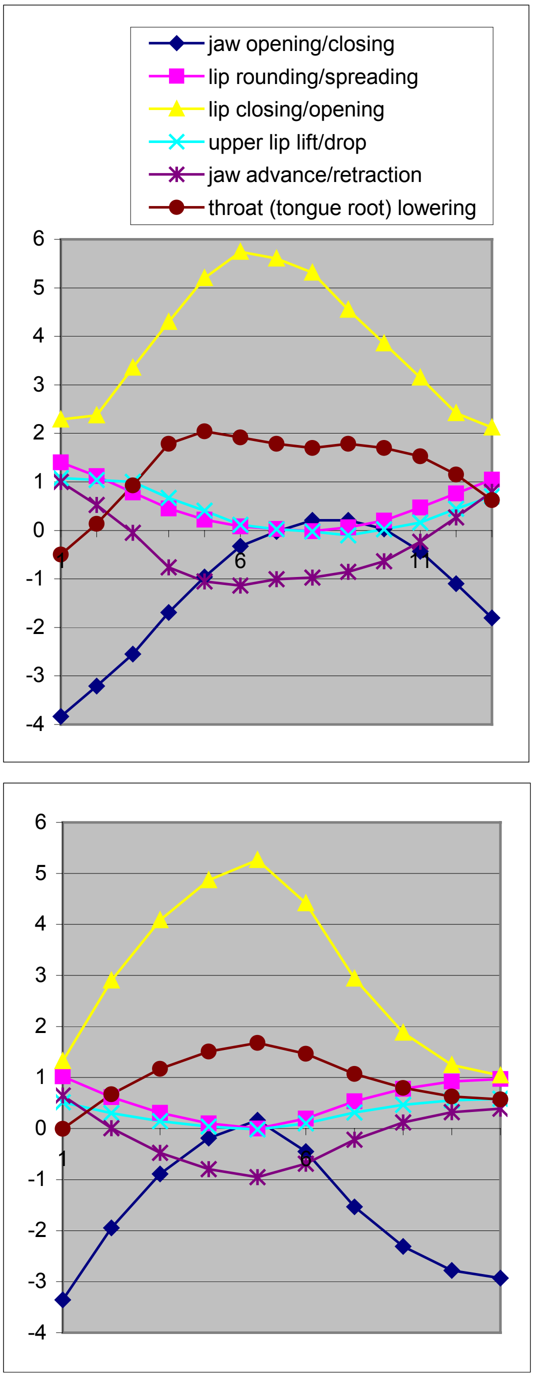

Figure 2: *Original (top) and synthesized (bottom) parameter trajectories for the /b/ in the sequence /aba/. Note that the phone timing and accordingly the duration in frames differ as they are naturally produced in the original and generated by the TTS system (which tends to speak faster) in the synthesized version.*

12 subjects with normal hearing and normal or corrected to normal vision participated voluntarily in the test. All word lists were used in the test but distributed over the subjects. One of the three word lists was presented to a subject audiovisually, another one was presented audio alone and a the third one visual alone. The conditions were blocked, the order of the conditions audiovisual, audio alone and visual alone was varied across subjects (each of the six possible

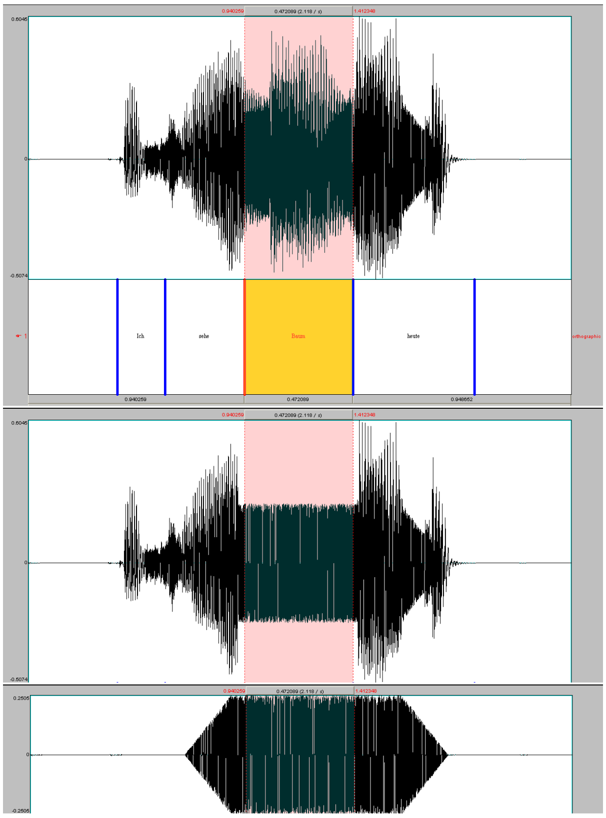

Figure 3: *Test word embedded in the carrier sentence "Ich sehe Baum heute". From top to bottom: the speech signal masked with noise (with orthographic tier, the test word is highlighted), the speech signal masked with noise with the test word cut out, and the noise signal.*

orders was used twice). After each presented word in carrier sentence the subject was requested to select the most probable word from the five alternatives.

5. Results

The recognition rate was 26.9% for the face only, which is only somewhat above the chance level of 20%. 10 of the 12 subjects showed an enhanced recognition with the face displayed along with the audio. The overall recognition rate increased from 45.6% in audio alone condition to 60.4% in audiovisual condition. This difference was highly significant (ANOVA: $p < .01$). Hence, the face causes a reduction of recognition errors (audiovisual benefit, [3]) of 27.2%. When the recognition rates are reduced to chance level, i.e. the percentage above 20% is regarded relative to the absolute possible 80% above chance, visual alone recognition is 8.6%, audio alone recognition is 32.0% and audiovisual recognition is 50.5%.

6. Conclusions and Discussion

The speech visualization of the presented TTavS system shows a significantly enhanced intelligibility in audiovisual condition compared to audio alone presentation in an evaluation experiment of word recognition. The gain in intelligibility measured by error reduction (27.2%) is somewhat below that previously measured with directly reproduced motion capture data [3] (32.6% to 41.6% depending on the SNR). Although the test methods differ and hence the results are not directly comparable, it is assumed that the articulation modeling and visualization as described in [3] and the trajectory generation presented here both decrease the visual intelligibility, but the cloning procedure and the trajectory generation preserve a large part of the benefit of visible speech movements.

The recognition scores reduced to chance level show a super-additive effect. The recognition scores of audio alone plus video alone presentations are below that one of audiovisual presentation. This is assumed to result from the very low information available in one of the channels, here in the visual alone display, where cross-modal redundancy is nearly not existent and cross-modal synergy effects still occur. Other descriptions of this kind of super-additivity also use nearly unintelligible presentations in one channel that nevertheless leads to enhancement in bimodal presentation: Our results are in line with outcomes of the work of Saldaña and Pisoni [18] who use natural video together with hardly intelligible sinewave speech, and Schwartz et al. [19] who use videos of voiced and unvoiced segments that are un-distinguishable from each other when played alone but result in a gain of intelligibility when played with audio.

The behavior of the articulatory parameters at utterance boundaries were not extracted from the motion capture database. When using test words embedded in a carrier sentence this drawback does not appear in the evaluation. An appropriate modeling of parameter trajectories at utterance boundaries is an open issue of the presented work.

The next step towards a data-based trajectory generation system following the one that is presented here will be to take a larger database, to use transitions found the database instead of using only targets where possible, and to adapt selection and concatenation methods from diphone synthesis and unit selection. Objective measures such as RMSE of marker positions will be applied besides perceptual evaluation.

7. Acknowledgements

We thank Frédéric Elisei, Christophe Savariaux and Ralf Baumbach for their help with recording and data preparation. The work was partly supported by DAAD (D/0502124) and DFG (FA 795/4-1).

8. References

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

[2] Le Goff, B., Guiard-Marigny, T., Cohen, M. M., Benoît, C., "Real-time Analysis-Synthesis and Intelligibility of Talking Faces", Proceedings of the Workshop on Speech Synthesis, New York, 53-56, 1994.

[3] Fagel, S., Bailly, G., and Elisei, F., "Intelligibility of Natural and 3D-cloned German Speech", Proceedings of the AVSP, 2007.

[4] Beskow, J., "Talking Heads–Models and Applications for Multi-modal Speech Synthesis", PhD Thesis at KTH Stockholm, 2003.

[5] Ezzat, T., Geiger, G. and Poggio, T., "Trainable Videorealistic Speech Animation", Proceedings of ACM SIGGRAPH, San Antonio, 2002.

[6] Bregler, C., Covell, M. and Slaney, M., "Video Rewrite: Driving Visual Speech with Audio", Proceedings of the International Conference on Computer Graphics and Interactive Techniques, 353-360, 1997.

[7] Kähler, K., Haber, J., Seidel, H.-P., "Geometry-based Muscle Modeling for Facial Animation", Proceedings of Graphics Interface, 37-46, 2001.

[8] Matthews, I., Xiao, J. and Baker, S., "2D vs. 3D Deformable Face Models: Representational Power, Construction, and Real-Time Fitting", International Journal of Computer Vision 75(1), 93-113, 2007.

[9] Maeda, S., "Face Models Based on a Guided PCA of Motion-capture Data: Speaker Dependent Variability in /s/-/S/ Contrast Production", ZAS Papers in Linguistics 40, 95-108, 2005.

[10] 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.

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

[12] Govokhina, O., Bailly, G. and Breton, G., "Learning Optimal Audiovisual Phasing for a HMM-based Control Model for Facial Animation", Proceedings of the ISCA Speech Synthesis Workshop, Bonn, 2007.

[13] Sumby, W., Pollack, I., "Visual Contribution to Speech Intelligibility in Noise", JASA 26, 212-215, 1954.

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

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

[16] Sendlmeier, W. F. and von Wedel, H., "Ein Verfahren zur Messung von Fehlleistungen beim Sprachverstehen - Überlegungen und erste Ergebnisse", Sprache - Stimme - Gehör 10, 164-169, 1986.

[17] Saldaña, H. and Pisoni, D., "Audio-Visual Speech Perception Without Speech Cues", Proceedings of the International Conference on Spoken Language Processing, Philadelphia, 2187-2190, 1996.

[18] Schwartz, J.-L., Berthommier, F. and Savariaux, C., "Audio-Visual Scene Analysis: Evidence for a “Very-Early” Integration Process in Audio-Visual Speech Perception". Proceedings of the International Conference on Spoken Language Processing, Denver, 1937-1940, 2002.