From 3-D Speaker Cloning to Text-to-Audiovisual-Speech

Sascha Fagel [1], *Frédéric Elisei*[2], *Gérard Bailly*[2]

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

sascha.fagel@tu-berlin.de, [gerard.bailly,frederic.elisei]@gipsa-lab.inpg.fr

Abstract

Visible speech movements were motion captured and parameterized. Coarticulated targets were extracted from VCVs and modeled to generate arbitrary German utterances by target interpolation. The system was extended to synthesize English utterances by a mapping to German phonemes. An evaluation by means of a modified rhyme test reveals that the synthetic videos of isolated words increase the recognition scores from 27 % to 47.5 % when added to audio only presentation.

Index Terms: talking head, intelligibility, evaluation

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 phoneme level often leads to reasonable results [4].

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 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 in isolation. The six main articulatory parameters were extracted by an iterative PCA from 42 poses of vowels and coarticulated consonants [3]. The synthetically reproduced VCV sequences have shown to yield ~70 % of the intelligibility provided by the natural face.

3. Coarticulation modeling and TTavS

The generation of parameter sequences for an utterance to be synthesized is done in two steps: target estimation and transition interpolation. The targets (a set of values of the six articulation parameters) for consonantal segments in asymmetric contexts are linear combinations of the parameters from the two measured context VCVs (or an estimated mixture of the existing data if V≠{a,i,u,E,O}). If consonants occur in clusters the influence of a context vowel decreases linearly with the distance in phones. A symmetric context is assumed if only one context vowel exists due to an utterance boundary. Targets for a vowel segment are taken 1) from the VCVs if the vowel is between two identical consonants, 2) from the vowel measured in isolation in case of pure vocalic context, or 3) an average of both if left context ≠ right context.

The transitions are linear to quadratic interpolations of the targets of each parameter where the exponent (1 to 2) is determined by the degree of coarticulation that occurred in the target estimation of consonants: linear interpolation if the target completely adopts to the neighbors (the parameter value of the consonant in the VCV equals that one of the vowel), quadratic interpolation if the target is not coarticulated at all (the parameter value of a consonant is the same in both contexts).

4. Evaluation

A modified rhyme test without carrier sentence was carried out as a preliminary evaluation of the system in terms of intelligibility. The English phonemes were mapped to German based on articulatory considerations and the American English voice *us1* of the mbrola speech synthesizer were used for the audiovisual synthesis with white noise at 0dB SNR. The test words were taken from [5]: six lists of 50 monosyllabic words each. 12 subjects with normal hearing and normal or corrected to normal vision participated in the test. All word lists were used in the test but distributed over the subjects. One of the six word lists was presented to a subject audiovisually, another one was presented audio alone (blocked conditions). Hence, a higher intelligibility of audiovisual compared to audio alone presentation – if present – cannot be produced by a possible learning effect. After each presented spoken word the subject was requested to select the most probable one from six alternatives.

Results: 11 of 12 subjects benefit from the additional synthetic visual speech. The overall increase of recognition rate is from 27 % (std. 6.1 %) correct answers in audio alone condition to 47.5 % (std. 12.9 %) in audiovisual condition (at a chance level of 16.7 %). This gain is highly significant (ANOVA, $p < .001$). The error reduction due to the synthetic face added to the audio presentation – as called "audiovisual benefit" by [1] – is 28.1 %.

5. Conclusions

The speech visualization of the presented TTavS system shows a significantly enhanced intelligibility compared to audio alone presentation in an evaluation experiment of isolated word recognition. The gain in intelligibility (20.5 %) is somewhat below that measured with directly reproduced motion capture data in [3] (24.5 % to 26.2 % depending on the SNR). However, due to the differing evaluation schemes and the limited transferability to intelligibility of natural speech, the system's performance has to be evaluated further.

6. Acknowledgements

We thank 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).

7. References

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

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

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

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

[5] House, A.S., Williams, C.E., Hecker, M.H.L., Kryter, K. D., "Articulation-Testing Methods: Consonantal Differentiation with a Closed-Response Set", JASA 37:1, pp. 158-166, 1965.