Merging methods of speech visualization

Sascha Fagel
Technical University Berlin

The author presents MASSY, the MODULAR AUDIOVISUAL SPEECH SYNTHESIZER. The system combines two approaches of visual speech synthesis. Two control models are implemented: a (data based) di-viseme model and a (rule based) dominance model where both produce control commands in a parameterized articulation space. Analogously two visualization methods are implemented: an image based (video-realistic) face model and a 3D synthetic head. Both face models can be driven by both the data based and the rule based articulation model.

The high-level visual speech synthesis generates a sequence of control commands for the visible articulation. For every virtual articulator (articulation parameter) the 3D synthetic face model defines a set of displacement vectors for the vertices of the 3D objects of the head. The vertices of the 3D synthetic head then are moved by linear combinations of these displacement vectors to visualize articulation movements. For the image based video synthesis a single reference image is deformed to fit the facial properties derived from the control commands. Facial feature points and facial displacements have to be defined for the reference image. The algorithm can also use an image database with appropriately annotated facial properties. An example database was built automatically from video recordings. Both the 3D synthetic face and the image based face generate visual speech that is capable to increase the intelligibility of audible speech.

Other well known image based audiovisual speech synthesis systems like MIKETALK and VIDEO REWRITE concatenate pre-recorded single images or video sequences, respectively. Parametric talking heads like BALDI control a parametric face with a parametric articulation model. The presented system demonstrates the compatibility of parametric and data based visual speech synthesis approaches.

1. Introduction

Speech communication usually consists of two coherent information streams, i.e. audition and vision. This is possible due to the fact that the movements of the speech organs that form the utterance become manifest in the acoustical and optical domain and hence are audible and visible. At least under acoustically bad conditions both information streams are used jointly to increase the robustness against transmission errors (Sumby & Pollack, 1954; Erber, 1969). This property of natural speech can also be helpful for speech synthesis (Benoît et al., 1995; Beskow, 2003). This is the case although there is not necessarily a single underlying process like in natural speech but – at least in current unlimited audiovisual speech synthesis systems ("talking heads") – the audio signal and the video signal are synthesized in most cases separately and are played back synchronously.

Although they borrow some techniques from one another, the visualization method of most audiovisual speech synthesis systems can be classified as either image based or parametric. The first class of systems concatenates parts of pre-recorded video speech material (comparable to concatenative audio synthesis systems). The second class models the speech production process by means of physiological, articulatory, or facial parameters. However, the present paper shows that both approaches are not mutually exclusive and that they can be combined in a single system.

2. Parametric and image based talking heads

Parametric visual speech synthesizers generate a sequence of values for a number of fixed parameters. These parameters can be e.g. virtual articulators like tongue tip, tongue back, lip opening, and so on. A synthetic face is then manipulated according to the parameter values. Simple spatial co-articulation, i.e. movements of an articulator caused by the movement of another one near to it, can be modeled in the facial animation. But the temporal co-articulation, i.e. articulators start to move towards their target position for one speech segment in preceding segments and partly carry over the target positions to subsequent segments, is modeled by generating appropriate parameter values.

In contrast, image based visual speech synthesizers use pre-recorded single images (e.g. MIKETALK: Ezzat & Poggio, 2000) or video sequences (VIDEO REWRITE: Bregler et al., 1997). These image databases are indexed by phonemes (or visemes) or phoneme (or viseme) sequences, respectively. Co-articulation can be taken into account by recording a database containing phonemes in possibly all needed contexts. Co-articulation differences as they occur between different languages (e.g. between lip rounding in Turkish and American English:

Boyce, 1990) cannot be realized with the same database. Some main properties of speech visualization systems are summarized in Table 1.

Table 1: General properties of talking heads.

Either	**Or**
synthetic face	natural images
one instance of the face	many instances of one face
(articulation) parameters	database indexed by (classes of) phonemes
co-articulation (mostly) outside the face model	co-articulation (hidden) inside the face model

Some recent developments do not completely fit in the parametric vs. image based distinction: MARY101 (Ezzat, 2002) defines a set of prototypic images and the optical flow (Horn & Schnuck, 1981) between them. The system generates appropriate video frames which do not necessarily have to lead from one prototype to another (which was the case for MARY101's predecessor MIKETALK). VOICE PUPPETRY (Brand, 1999) is trained by audiovisual recordings and then provides a sequence of facial motion vectors related to the audio track. These facial motion vectors can be applied to other prepared faces. The visual extension (Minnis & Breen, 2000) of the concatenative audio synthesis system LAUREATE (British Telecom) associates N-visemes with face deformations which can be applied to a 3D face model.

3. The MODULAR AUDIOVISUAL SPEECH SYNTHESIZER

The system that demonstrates the compatibility of parametric and image based approach is called MASSY (MODULAR AUDIOVISUAL SPEECH SYNTHESIZER). A plain text serves as system input. The phonetic articulation module creates the phonetic information, which consists of an appropriate phone chain on the one hand and - as prosodic information - phone and pause durations and a fundamental frequency curve on the other hand. From this data, the audio synthesis module generates the audio signal and the visual articulation module generates motion information. This motion information consists of control commands for virtual articulators given by an articulation model. Hence, the control part of the visualization follows in general the parametric approach. The face module interprets the motion information and adds the audio signal to create the complete audiovisual speech output. Figure 1 shows a system overview.

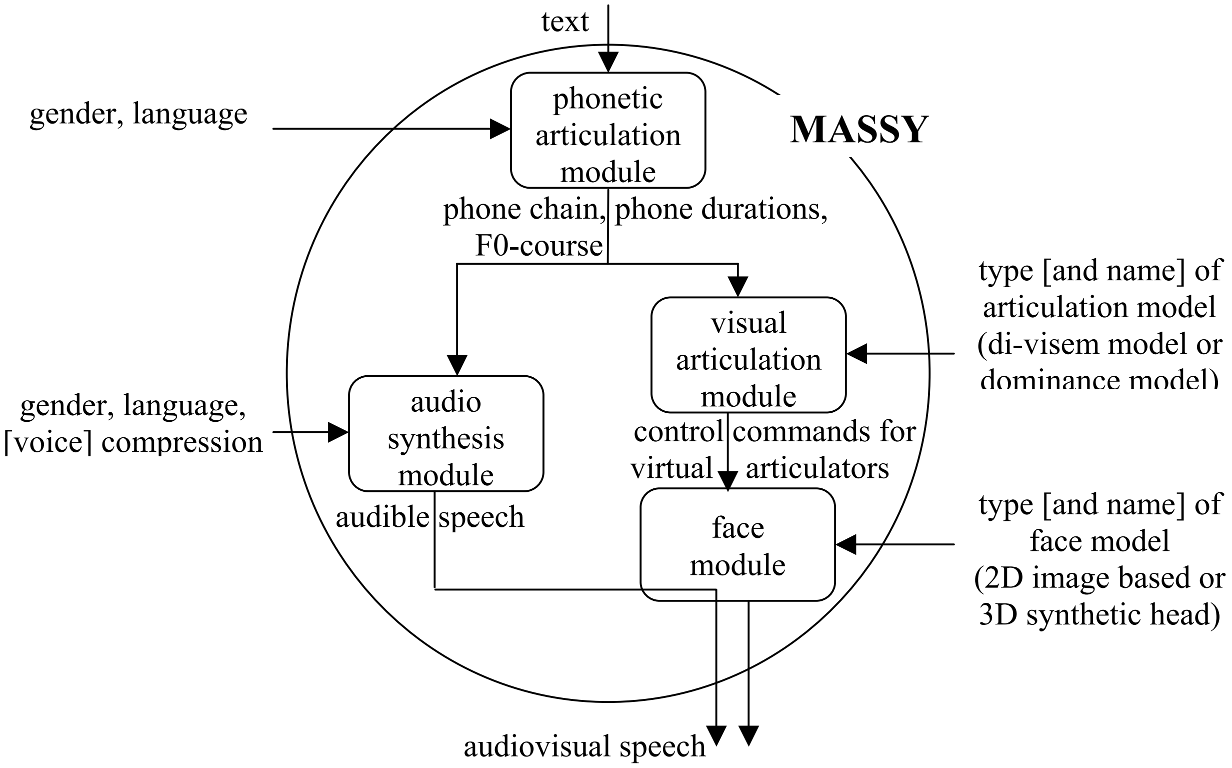
text
gender, language
phonetic articulation module
MASSY
phone chain, phone durations, F0-course
type [and name] of articulation model (di-visem model or dominance model)
visual articulation module
gender, language, [voice] compression
audio synthesis module
control commands for virtual articulators
audible speech
face module
type [and name] of face model (2D image based or 3D synthetic head)
audiovisual speech

Figure 1: Schematic system overview of MASSY.

The system in its present state can be tested at http://avspeech.info. This website provides a user interface to a fully functional text-to-audiovisual-speech synthesis system built of the modules of MASSY. The phonetic transcription and the voice can be German or English, male or female. Both face models described below are available including a simple tool to mark facial feature points in any uploaded image file to make a new face talking. Among others, experimental settings of speaking rate, hyper/hypo-articulation, and phoneme/ viseme replacements to generate McGurk stimuli (McGurk & MacDonald, 1976) are possible.

3.1. Visual articulation module

The visual articulation module generates a sequence of values for articulation parameters to synthesize a phone chain given by the phonetic articulation module. These parameters control the face model implemented in the subsequent module. The articulation parameters of the articulation model currently are

- lip width
- jaw height
- lip height

- tongue tip height
- tongue back height
- lower lip retraction

The lip width is 0 at neutral position (relaxed state), 1 at maximum narrowing and -1 at maximum spreading. For the production of some vowels the real vocal tract is lengthened by lip protrusion, for some other vowels shortened by lip spreading. A negative correlation between lip protrusion and spreading is assumed. At least in German and English which are the languages MASSY currently can "speak" there is no acoustic-articulatory need to spread the lips while protruding them or to narrow them without protrusion. This is an appropriate simplification if the goal is to realize one plausible articulation and not to clone a specific speaker. Hence, lip rounding and narrowing are combined to one articulation parameter. The lower jaw height is 0 at closed jaw and 1 at maximum opening. The lip height is 0 at neutral position relative to the upper and lower teeth. It is 1 for the lips moved maximum towards each other on the upper and lower jaw and -1 if the lips are moved maximum apart. So the vertical lip opening depends on both the jaw height and the lip height and the lips can be closed only if the jaw is not wide open (see Figure 2).

The tongue tip height and tongue back height are 0 at relaxed tongue and 1 at tongue contact at the alveoli or the palate, respectively. For this, the absolute values of the displacement vectors for tongue tip height and tongue back height are scaled by the lower jaw height in order not to break through the palate. The retraction of the lower lip is 0 at neutral position and 1 at retracted position for labiodental constrictions. The set of motion parameters was chosen with respect to the visibility of German phones displayed by MASSY. Motion parameters for tongue advance and velum closure are currently under construction. The visual articulation module implements alternatively two different articulation models to generate the values for the articulation parameters: a di-viseme model and a dominance model (Löfqvist, 1990). Details of the algorithms can be found in Fagel & Sendlmeier (2003).

3.2. Face module

The face module visualizes the articulator movements described by the articulation parameters. The module creates an animation of a face and dubs the synthesized speech audio. One face model is a 3D synthetic head with a set of displacement vectors for each articulation parameter. A second alternative face model is image based. Figure 3 shows image sequences generated by the two face models.

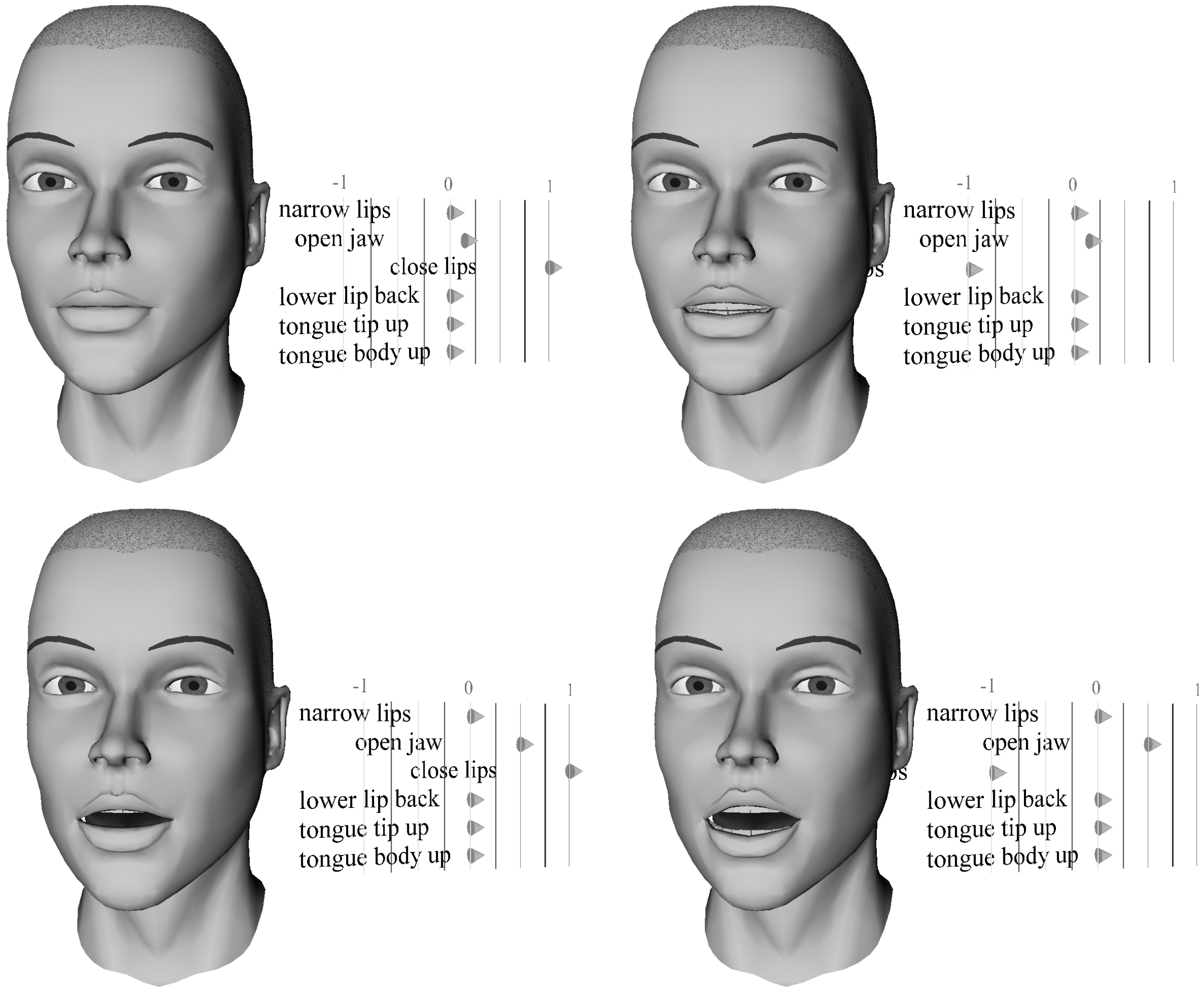

Figure 2: Minimum (both left images) and maximum (both right images) lip height at lower jaw nearly closed (both top images) and half opened (both bottom images) displayed with the 3D synthetic face.

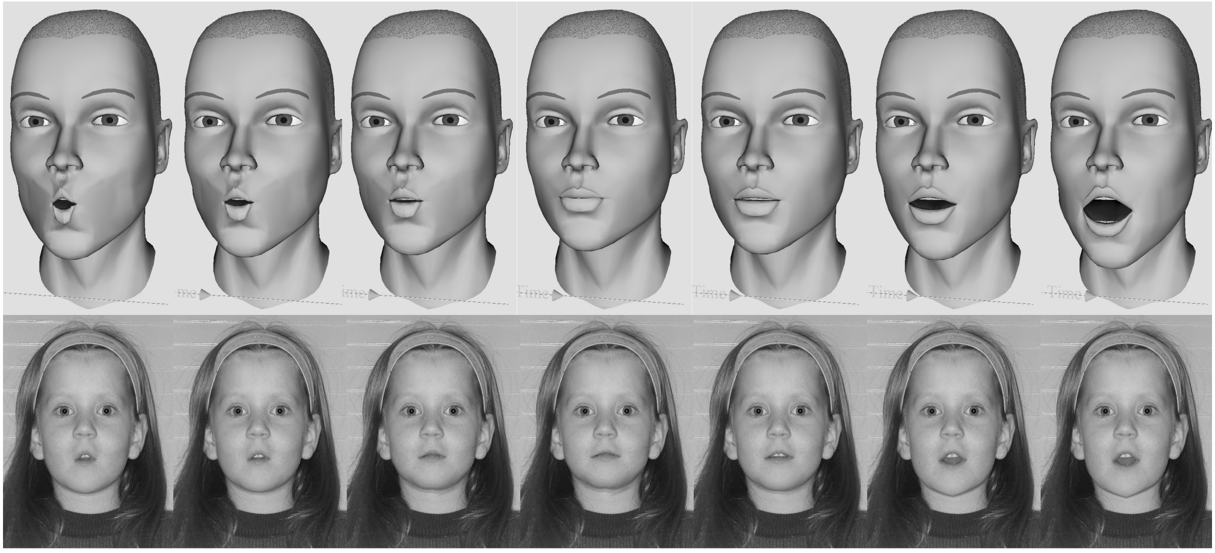

Figure 3: Image sequences of the utterance /oma/ generated by the 3D synthetic head (top) and the image based face model (bottom).

3.2.1. 3D synthetic head

MASSY's 3D face model is realized in VRML and uses the six motion parameters provided by the visual articulation module. The difference of the 3D model in neutral state to the model deformed to the maximum position of one articulator constitutes an articulator excursion. The difference vectors of all affected vertices besides the concerning vertex index are stored as a so called displacer. All possible articulator positions result from linear combinations of the vertex difference vectors contained in the displacers. A facial animation is generated from a sequence of motion parameter values.

3.2.2. Image based face model

The image based face model consists of an image database indexed by facial properties (instead of phonemes/visemes). These facial properties are a subset of the articulation parameters with respect to the visibility of articulators in the images. Currently these facial properties are implemented:

- the lip width and
- joint lip and jaw height and the lower lip retraction.

The image database can be built by deforming a reference image to fit the facial properties. For this procedure 37 feature points are defined in a reference image of a face. 27 of them correspond to feature points standardized in MPEG-4. Five additional feature points define a surrounding of the lower jaw area to prevent sharp edges when the lower jaw is displaced. Another five feature points mark the upper teeth to save them from being deformed or displaced.

Two displacement vectors (one per facial property) are assigned to each of the 37 feature points. These two displacement vectors are linearly combined – weighted with the magnitude of the facial property – before being applied to the feature point for deformation. The pixels of the face image are displaced using a bilinear interpolation between the combined displacement vectors of three feature points surrounding the pixel. Details of the algorithm can be found in Fagel (2004). Figure 4 shows the lower part of a reference image including the feature points, the triangle mesh built of them, and schematically the two displacement vectors for each feature point.

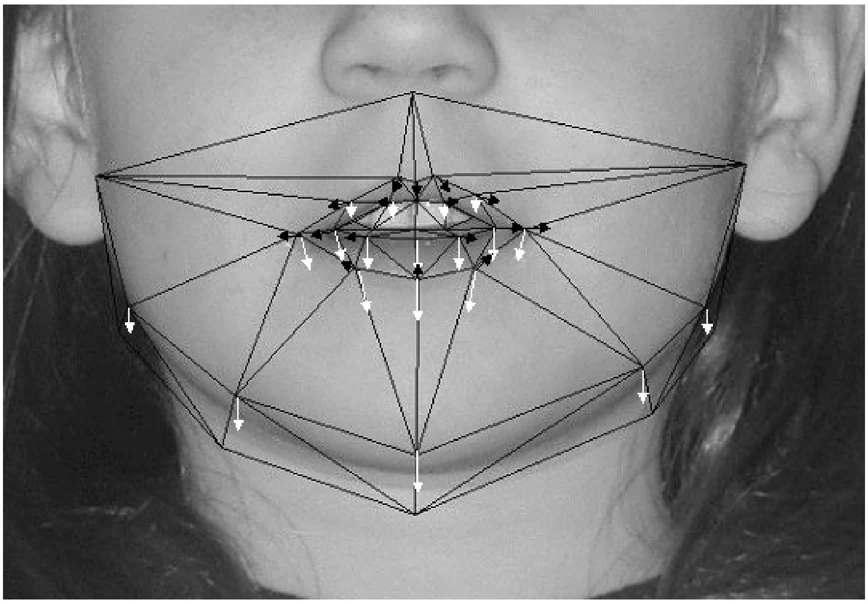

Figure 4: Lower part of a neutral face with the triangle mesh built of the feature points and – schematically – the two displacement vectors for each feature point. Black arrows show the displacements for lip spreading (which also leads to a lip slimming), white arrows show the displacements for vertical mouth opening.

A simple software tool for marking the feature points in an image by hand was developed. This software tool can also be used to mark displaced feature points in an image with one facial property different from neutral. The differences of these feature points and those in the reference image form the displacement vectors. Instead of defining displacement vectors for a new face, a predefined set of displacement vectors for each facial property can be used as a preset. These displacement vectors are scaled by the width of the lips in the reference image of the new face to fit the proportions.

Alternatively to the image deformation approach an example image database was created. A male speaker was videotaped and the frames were extracted. Outer lip height and width were annotated automatically by a lip feature extraction program. In case of duplicates images with upper lip position and vertical lip center near the average upper lip position and average vertical lip center were chosen. A more sophisticated criterion for similarity of images which constitute a database will follow. Figure 5 shows the frames selected from the database for the utterance /oma/ as well as the frames generated by deforming one image of the database with the method described above for the same utterance.

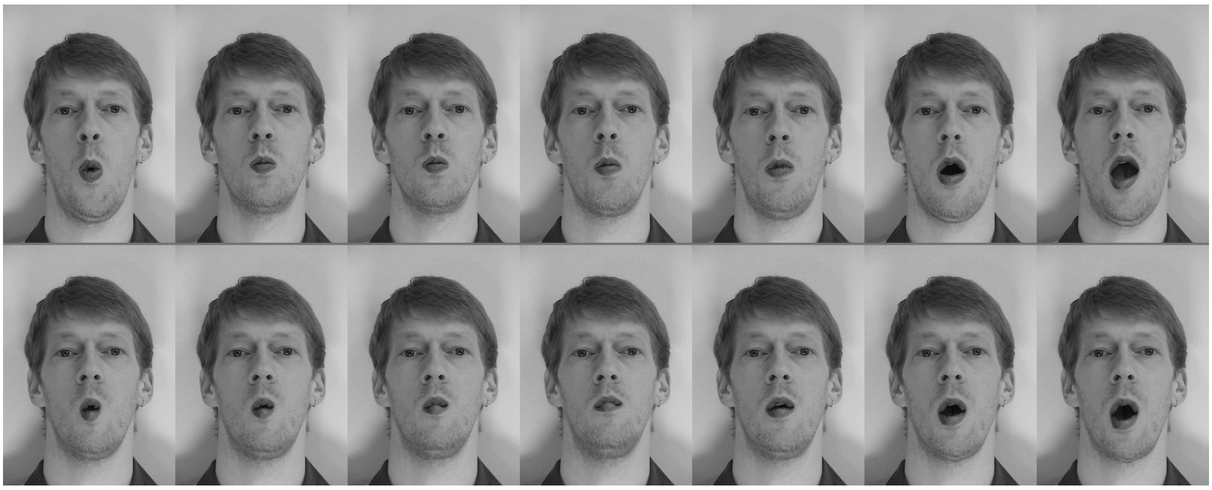

Figure 5: Frames of the utterance /oma/ selected from the database (top) and generated by deforming one image of the database (bottom).

4. Evaluation

4.1. 3D synthetic head

4.1.1. Method

A phonetically balanced rhyme test (Sendlmeier & v. Wedel, 1986) covering initial and final consonants and medial vowels was used in the first experiment for the evaluation of the 3D synthetic head. A trained female speaker uttered the items of the corpus and was videotaped. The recorded items were split and saved as single files. The audio channel was separated. All phones of the recorded items were labeled by hand and the fundamental frequency curves were extracted using the speech analysis software Praat. This information was handed to the face module as input. For each item the synthesizer generated an animated face and an audio file synchronous to the recorded natural utterances.

Both the synthetic head and the videotaped face were paired with both the synthesized and recorded voice. Synthetic and natural audio alone conditions were used as references. Several measures of audiovisual integration (Massaro, 1987, Braida, 1991, Grant & Seitz, 1998) also use visual alone conditions. These were included as well, resulting in a total of eight conditions: synthetic stimuli in audio alone, visual alone and audiovisual conditions (a_s, v_s, a_sv_s), natural stimuli in audio alone, visual alone and audiovisual conditions (a_n, v_n, a_nv_n) , and mixed audiovisual stimuli (a_nv_s, a_sv_n). All audio material was mixed with white noise at -6dB signal-to-noise ratio. The visual alone stimuli were presented without noise. 36 undergraduate students of communication science participated in the test voluntarily. Every subject was presented with the stimuli in a different pseudo-random order.

4.1.2. Results

An analysis of variance was carried out for all pairs of conditions. Table 2 shows the mean recognition scores where conditions producing non-significantly differing results are grouped ($p<0.05$). All audiovisual conditions resulted in higher recognition scores than all unimodal conditions (audio alone or visual alone respectively). Both natural unimodal conditions led to higher scores than the corresponding synthetic condition. In case of synthetic audio in bimodal condition the pairing with natural video showed better recognition scores than the pairing with synthetic video, but the scores reached with natural and synthetic video paired with natural audio did not differ significantly from each other.

Table 2: Mean recognition scores of the first experiment in %.

condition	subgroup				
	1	2	3	4	5
a_s	43.6				
v_s		48.3			
a_n			58.5		
v_n			60.8		
$a_s v_s$				67.2	
$a_n v_s$					75.3
$a_s v_n$					77.2
$a_n v_n$					79.6

4.2. Image based face model

4.2.1. Method

For the image based approach a simpler evaluation experiment was carried out. 12 of the most frequent German phones (Kohler, 1995) were chosen: /p, b, m, t, d, n, l, s, z, k, g, ŋ/. Additionally 3 of the most frequent vowels /a, ɪ, ʊ/ that nearly span the German vowel space were used to build all 36 items of the form VCV. These items were synthesized with identical phone durations and constant fundamental frequency. White noise at -4.5dB signal-to-noise ratio was added to the audio signals. Stimuli in the conditions audio alone (a), visual alone (v) and audiovisual (av) were generated resulting in 108 stimuli in total. All stimuli were presented in the same pseudo-random order to five students of communication science (two male, three female, 23-28 years, mean age 25.4, all normal hearing and normal or corrected to normal vision). The subjects were asked to mark the answer for each stimulus among all possible 36 items.

4.2.2. Results

Vowel and consonant identification was analyzed separately, regarding segments as correctly identified if the chosen answer contained the right of the three vowels or the right of the 12 consonants, respectively. Table 3 shows the mean recognition scores for vowels and consonants in the conditions audio alone, visual alone, and audiovisual. An analysis of variance revealed that vowels and consonants in the audiovisual condition were significantly better recognized ($p<0.05$) than in the audio alone condition. Except for consonant identification in one subject (where the recognition was identical) all subjects reached higher recognition scores in audiovisual than in audio alone condition. The relative error reduction (audiovisual benefit, Sumby & Pollack, 1954) was 55% for vowels and 15% for consonants.

Table 3: Mean recognition scores for vowels and consonants in the second experiment in %.

condition	vowels	consonants
a	61.7	22.8
v	63.9	12.2
av	82.8	34.4

4.2.3. Discussion

The evaluation of the image based face model shows further interesting data which have to be confirmed in forthcoming experiments. The visual information is obviously integrated into the audiovisual perception although the visual alone identification of consonants (12%) is only slightly above chance level (8.3%). If the chance level is taken into account (Equation 1: chance level correction) a super-additive information usage can be seen (Table 4). The correct consonant identification above chance is 4.2% for video alone, 15.8% for audio alone, and 28.5% in the audiovisual condition which is more than the sum of audio+video. This super-additivity of speech perception was already observed by Saldaña & Pisoni (1996) in an audiovisual speech intelligibility test with sine-wave speech as audio signal. Furthermore it was implicitly described by Schwartz (2003) where a non-informative – if presented alone – video increased the distinction of voiced and unvoiced plosives when added to the audio signal.

$$R' = (R-C)/(1-C) \qquad (1)$$

where R': chance level corrected recognition score,
R: recognition score, $0 \leq R \leq 1$
C: chance level, $0 \leq C \leq 1$,
here: C=1/N with N: quantity of response alternatives.

Table 4: Chance level corrected recognition scores for vowels and consonants in the second experiment in %.

condition	vowels	consonants
a	42.6	15.8
v	45.9	4.2
av	74.3	28.5

5. Summary and future work

Both the 3D synthetic head and the image based speech synthesis enhance the intelligibility of audible speech. The visualization methods realize different levels of abstraction from natural static appearance and natural dynamics. With increasing abstraction the benefit of visual speech decreases (Benoit, 1996). There are some studies investigating the influence of spatial and temporal resolution (de Paula et al., 2000, Massaro, 1998) on the speech perception process. Knowledge in this area will help to design maximum intelligible talking heads at minimum system performance requirements and programming effort. Natural speech including shape and appearance (face topology and texture) and dynamics will be simulated with MASSY (methods for “speaker cloning” have been reported e.g. by Odisio & Bailly, 2003). Then the precision of the simulation will successively be reduced in order to determine the crucial synthesis properties.

The MODULAR AUDIOVISUAL SPEECH SYNTHESIZER combines parametric and image based approaches of speech visualization. In this way the visual speech output can be video-realistic but controlled by a specific articulation model not included in the image database. Co-articulation taking into account potentially all preceding and subsequent speech segments (instead of being limited to e.g. neighbors) becomes possible. The separation of articulation and visualization enables the synthesis of different speaking styles within one visual database. But an appropriate database is required for successful synthesis. Ideally all facial properties that are visible should be annotated to the images and not only speech specific properties. So images with similar annotated facial properties look only marginally different from each other. This guarantees smooth transitions when images are concatenated. A database built of deformed versions of a reference image fulfils this requirement. When using databases consisting of naturally recorded material the recording conditions have to be constant (or an accordingly huge corpus has to be recorded) and more facial properties than the two described above have to be annotated to the material. An example database was automatically created by means of a lip feature extraction software. But similarities regarding facial features that are not yet annotated (e.g. eye closure) currently must be detected manually.

Acknowledgements

I thank my students – especially Åsa Wrange – for their active participation in my course on audiovisual speech perception. Parts of the present paper are based on their work.

References

Benoît, C., Abry, C., Cathiard, M., Guiard-Marigny, T. & Lallouache, T. (1995). 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*, Marseille.

Benoît, C. (1996). On the Production and the Perception of Audio-Visual Speech by Man and Machine. In H. Bertoni, Y. Wang & S. Panwar (eds.), *Multimedia and Video Coding*. Plenum Press, New York.

Beskow, J. (2003). Talking Heads – Models and Applications for Multimodal Speech Synthesis. PhD Thesis, Stockholm.

Boyce, S. E. (1990). Coarticulatory Organisation for Lip Rounding in Turkish and English. *Journal of the Acoustical Society of America*, 88: 2584-2595.

Braida, L. D. (1991). Crossmodal Integration in the Identification of Consonant Segments. *Quarterly Journal of Experimental Psychology*, 43, 647-677.

Brand, M. (1999). Voice Puppetry. *Proceedings of the Conference on Computer Graphics and Interactive Techniques*, Los Angeles: 21-28.

Bregler, C., Covell, M. & Slaney, M. (1997). Video Rewrite: Driving Visual Speech with Audio. *Proceedings of the Conference on Computer Graphics and Interactive Techniques*, Los Angeles: 353-360.

de Paula, H. B.,Yehia, H. C., Shiller, D., Jozan, G., Munhall, K. & Vatikiotis-Bateson, E. (2003). Linking Production and Perception Through Spatial and Temporal Filtering of Visible Speech Information. *Proceedings of the 6th International Seminar on Speech Production*, Sydney: 37-42.

Erber, N. P. (1969). Interaction of Audition and Vision in the Recognition of Oral Speech Stimuli. *Journal of Speech and Hearing Research*, 12: 423-425.

Ezzat, T. & Poggio, T. (2000). Visual Speech Synthesis by Morphing Visemes. *International Journal of Computer Vision*, 38: 45-57.

Ezzat, T., Geiger, G. & Poggio, T. (2002). Trainable Videorealistic Speech Animation. *Proceedings of ACM SIGGRAPH*, San Antonio: 388-398.

Fagel, S. & Sendlmeier, W. F. (2003). An Expandable Web-based Audiovisual Text-to-Speech Synthesis System. *Proceedings of the 8th EUROSPEECH European Conference on Speech Communication and Technology*, Geneva: 2449-2452.

Fagel, S. (2004). Video-realistic synthetic speech with a parametric visual speech synthesizer. *Proceedings of the INTERSPEECH*, Korea.

Fagel, S. (2004a). Audiovisuelle Sprachsynthese – Systementwicklung und -bewertung. Logos Verlag, Berlin.

Fagel, S. & Clemens, C. (2004). An Articulation Model for Audiovisual Speech Synthesis – Determination, Adjustment, Evaluation. *Speech Communication*, 44: 141-154.

Grant, K. W. & Seitz, P. F. (1998). Measures of Auditory-Visual Integration in Nonsense Syllables and Sentences. *Journal of the Acoustical Society of America*, 104: 2438-2450.Löfqvist, A. (1990). Speech as Audible Gestures. In W. J. Hardcastle & A. Marchal (eds.), *Speech Production and Speech Modeling*. Kluwer Academic Publishers, Dodrecht.

Kohler, K. J. (1995). Einführung in die Phonetik des Deutschen. Schmidt, Berlin, 1995.

Massaro, D. W. (1987). Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry. Erlbaum, London.

Massaro, D. W. (1998). Illusions and Issues in Bimodal Speech Perception. *Proceedings of Audiovisual Speech Processing*, Sydney: 21-26.

McGurk, H. & MacDonald, I. (1976). Hearing Lips and Seeing Voices. *Nature*, 264: 746-748.

Minnis, S. & Breen, A. (2000). Modeling Visual Coarticulation in Synthetic Talking Heads Using a Lip Motion Unit Inventory with Concatenative Synthesis. *International Conference on Spoken Language Processing*, Beijing: 759-762.

Odisio, M. & Bailly, G. (2003). Shape and appearance models of talking faces for model-based tracking. *Proceedings of Audiovisual Speech Processing*, St. Jorioz: 105-110.

Sumby, W. H. & Pollack, I. (1954). Visual Contribution to Speech Intelligibility in Noise. *Journal of the Acoustical Society of America*, 26: 212-215.

Saldaña, H. & Pisoni, D. (1996). Audio-Visual speech perception without speech cues. *Proceedings of the 4th International Conference on Spoken Language Processing*, Philadelphia: 2187-2190.

Schwartz, J.-L., Berthommier, F. & Savariaux, C. (2002). Audio-Visual Scene Analysis: Evidence for a “Very-Early” Integration Process in Audio-Visual Speech Perception. *Proceedings of the 7th International Conference on Spoken Language Processing*, Denver: 1937-1940.

Sendlmeier, W. F. & v. Wedel, H. (1986). Ein Verfahren zur Messung von Fehlleistungen beim Sprachverstehen – Überlegungen und erste Ergebnisse. *Sprach, Stimme, Gehör*, 10: 164-169.