Intelligibility of Natural and 3D-Cloned German Speech

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

[1] Institute for Speech and Communication, Berlin University of Technology, Germany
[2] Department of Speech and Cognition, GIPSA-Lab Grenoble, France
sascha.fagel@tu-berlin.de, bailly@icp.inpg.fr, frederic.elisei@icp.inpg.fr

Abstract

We investigate the intelligibility of natural visual and audiovisual speech compared to re-synthesized speech movements rendered by a talking head. This talking head is created using the speaker cloning methodology of the Institut de la Communication Parlée in Grenoble (now department for speech and cognition in GIPSA-Lab). A German speaker with colored markers on the face was recorded audiovisually using multiple cameras. The three-dimensional coordinates of the markers were extracted and parameterized. Spoken VCV sequences were then visually re-synthesized. A perception experiment was carried out to measure the visual and audiovisual intelligibility of natural and synthesized video, using the original audio with and without added noise. Identification scores show that the clone is capable of recovering almost 70% of the intelligibility gain provided by the original face. Part of this loss is due to missing visual cues in the present synthesis, due notably to the lack of a tongue.

Index Terms: visual speech synthesis, speech intelligibility, speech motion capture, talking head

1. Introduction

Speech production generates two coherent information streams. This is due to the fact that the movements of the speech organs that form the utterance become manifest in the acoustical and optical domain. Speech perception is bimodal in nature, too, i.e. humans process both auditory and visual information – if present – when perceiving speech. It is known for decades that audiovisual speech leads to better recognition compared to pure audio speech [1][2]. Audition and vision contain both redundant and complementary phonetic cues that can be jointly used when modalities are combined. This perception process also applies to synthetized speech [3][4]. This perceptive fusion process might be challenging for the current audiovisual speech synthesis systems ("talking heads") were the audio signal and the video signal are in most cases not generated by a unique underlying process but synthesized separately and played back synchronously. Research has also shown that this benefit can decrease when using stimuli too far away from natural speech [5][6]. Strong and well-known evidence for the sensory integration of auditory and visual information was found by McGurk & MacDonald [7] who showed that a visual syllable /ga/ combined with an audible syllable /ba/ mainly leads to the overall auditory perception of /da/, the so-called McGurk effect. Auditory and visual cues are here integrated into one percept despite incoherent stimuli. The integration process takes place whether or not the subject is aware about the effect.

Although they borrow some techniques from one another, the rendering 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 – to which the present method of speech synthesis belongs – models the speech production process by means of physiological, articulatory, or facial parameters.

2. Speaker cloning procedure

A native speaker of German was recorded audiovisually from three synchronized views: left, right and center. 398 colored markers were placed on his skin from the hair line to the neck and from one ear to the other. Figure 1 shows the marked face.

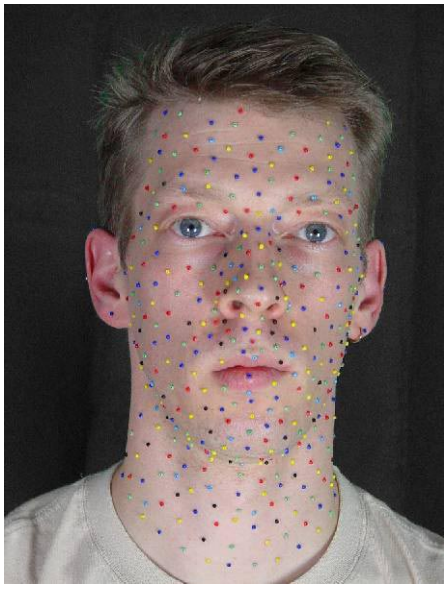

Figure 1: *398 colored beads glued on the speaker's face.*

Figure 2: *Textured synthetic face (from left to right): in neutral position, with open jaw, rounded lips, closed lips, lifted upper lip.*

In a semi-automatic procedure these markers and 10 additional points at the eye corners and incisors were registered for 40 representative images (visemes). 3D coordinates of points that were visible from at least two views were calculated by the use of the calibrated projection matrices of the cameras. Additionally a separate lip model with 30 anchor points was fitted to the measured visemes. With a "guided" principal component analysis (PCA) [8] 6 articulatory parameters and 6 head posture parameters were determined. This procedure consists in injecting *a priori* knowledge in the linear decomposition of facial movements. Initially, the data of the markers on the lower jaw line were isolated and analyzed by a plain PCA. The first component of the PCA on this reduced data set represents jaw opening. The articulatory parameter for jaw opening was determined in the whole data by a linear regression of the first PCA component and the data. Contribution of this articulatory parameter was then removed from the data and another step of the procedure led to a second parameter, and so on. The resulting articulatory parameters are

Jaw opening/closing

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

Figure 2 and 3 illustrate the actions of these articulatory parameters.

Figure 3: *Neutral face shape (black line), advanced jaw (dashed line) and lowered tongue root (dotted line).*

The 6 head position parameters correspond to rotations around the x, y and z axes and the shifts on the axes. It is worth mentioning that these parameters are not simple rotations and translations but they include neck torsion and so model measured head turn, side tilt, nodding, and forward/backward, side and up/down shifts. Figure 4 shows the action of the first three head parameters separately.

Marker positions in all video frames of 36 recorded VCV sequences with V = {a, i, u} and C = {p, b, m, v, t, d, n, z, k, g, N, R} were automatically tracked and inverted by means of an error minimization procedure using the 12 parameters.

Figure 4: *From left to right: neutral head position, side tilt, side turn, head nod. The chest, linked to the base of the neck, is not moving when the position of the head changes.*

3. Evaluation

3.1. Stimulus generation

A triangular mesh of the marker positions was manually defined. Three-dimensional animations were created using an animation software developed at ICP using OpenGL. They are driven by the parameter values of the aforementioned marker tracking procedure. A three-dimensional textured model of the upper teeth was introduced into the face motion data. The position and movements of the teeth were estimated from the measured marker positions. A static texture of the face was added to the face shape. No tongue was included at this stage as no tongue position data was acquired during the recording session. The projection properties of the front view camera were used for the video synthesis of the 36 VCV sequences. Video clips of the VCV sequences were extracted from the front view recordings. Audio clips from the original recordings were extracted accordingly. Two degraded versions of these clips were created: pink noise was added at SNR (sound-to-noise ratio) –6dB and 0dB, respectively. Audiovisual stimuli were merged by combining the video types *natural*, *synthetic* and *no video* with the audio types *clean audio*, *SNR 0dB*, *SNR –6dB* and *no audio* (except *no video* and *no audio*). This leads to 396 stimuli (11 conditions with 36 stimuli each).

3.2. Experimental procedure

The presentation and answering to 396 stimuli would result in a test of about an hour with a lot of noisy stimuli. We thus decided to distribute the stimuli with *SNR 0dB* and *SNR –6dB* stimuli among two subgroups of equal size. All subjects were also presented with *no audio* and *pure audio* stimuli.

Ten undergraduates and graduates of the faculty for humanities at the Berlin University of Technology participated voluntarily. The test was designed as a forced choice test with 36 alternatives and with variable inter-stimulus intervals (subjects were able to start the next stimulus when they had answered to the preceding one). Subjects were instructed to answer intuitively and to react as quickly as possible. Resulting test durations were between 27'30" and 29'45".

3.3. Results: identification

Identification rates in conditions that included visual information were always higher than the corresponding audio-only conditions. Conditions with natural video were always better recognized than those with synthetic video. Clear audio-only condition resulted in very high recognition rates, so enhancements by natural and by synthetic video were non-significant due to a ceiling effect. All other visual enhancements and the advantage of natural versus synthetic display were significant ($p<.05$; see table 1).

audio type	video type	significance groups					
		1	2	3	4	5	6
pure audio	natural	96.9					
pure audio	synthetic	96.4					
pure audio	no video	96.1					
SNR 0dB	natural		76.7				
SNR 0dB	synthetic			65.6			
SNR -6dB	natural			61.7			
SNR -6dB	synthetic				45.6		
SNR 0dB	no video				41.1	41.1	
no audio	natural					32.5	
no audio	synthetic						19.4
SNR -6dB	no video						19.4

Table 1: *Mean recognition rates at the 11 conditions. Non-significant values are grouped.*

Figure 5 shows the recognition scores of stimuli with natural, synthetic and audio-only stimuli at different signal-to-noise ratios. This figure evidences a clear enhancement by visual information and the greater benefit of natural video.

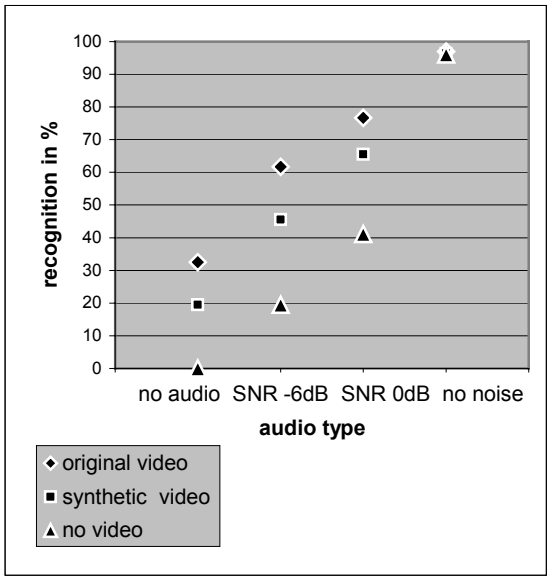

Figure 5: *Mean recognition scores of stimuli with natural, synthetic and without video at different audio types.*

One major drawback of the synthetic display is revealed by a closer analysis of phoneme classes (Figure 6). Alveolars are identified second best in conditions with natural video and degraded audio (added noise or no audio; in combination with pure audio the aforementioned ceiling effect occurs). However, in the corresponding conditions with synthetic video, alveolars are always identified worst. Labials and labiodentals are identified best with both natural and synthetic video. The recognition rates of velars and uvulars in conditions with natural video are almost identical to those in the corresponding conditions with synthetic video.

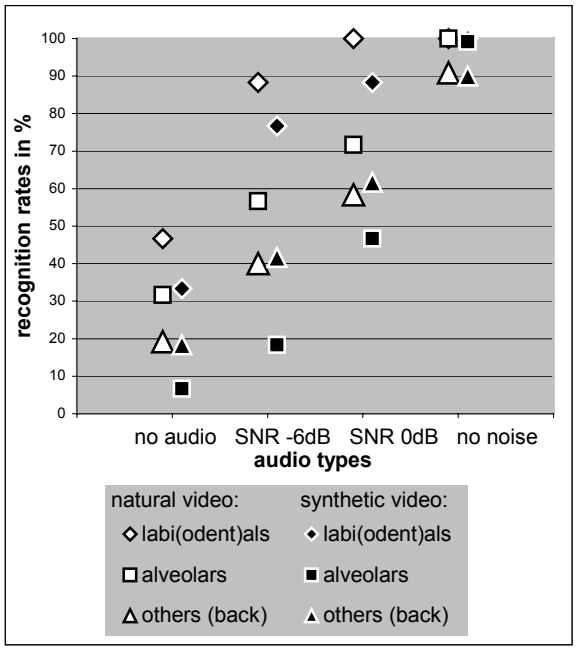

Figure 6: *Mean recognition scores of stimuli in the 8 conditions with natural or synthetic video at different audio types split into phoneme classes according to the place of articulation.*

3.4. Perceptual confusion analysis

Tree diagrams of the confusions in different conditions were built as follows (the method was taken from [9]): First, the similarity of each pair of stimuli was determined from the confusion matrices per condition by equation (1), resulting in one symmetric similarity matrix for each condition. Then the similarity matrices were transformed in dissimilarity matrices by equation (2). Finally, tree diagrams were built from the dissimilarity matrices using the single linkage criterion.

$$S_{ij} = \frac{1}{2}\sum_{k}\left(M_{ki} + M_{kj} - \left|M_{ki} - M_{kj}\right|\right) \qquad (1)$$

$$D_{ij} = 1 - \frac{s_{ij}}{\max(S)} \qquad (2)$$

3.5. Results: confusion trees

Natural video alone presentation (Figure 7) results in four main groups of similarity:

1. all /a/ contexts, all labiodentals and all bilabials in /i/ context
2. all bilabials in /u/ context
3. all alveolars and velars in /i/ context
4. all alveolars and velars in /u/ context

The first main group additionally shows the stimuli clearly in blocks of place of articulation. Context-specific subgroups of alveolars, velars (except /aka/), labiodentals and bilabials occur, and one subgroup of the labiodental in all three vocalic contexts. The last main group also shows subgroups of alveolars (except /uzu/) and velars.

The tree diagram of synthetic video alone presentation (Figure 8) shows three main groups:

1. all stimuli in /u/ context and all non-bilabials in /i/ context
2. all non-bilabials in /a/ context
3. all bilabials in /a/ and in /i/ contexts

The first main group shows context-specific similarities: bilabials in /u/ context and all other stimuli in /u/ context are separated from stimuli in /i/ context, but among the non-bilabials, the stimuli are not grouped according to alveolar or velar place of articulation. The same applies to the second main group, where only the labiodental shows greater dissimilarity. In the third main group the stimuli in /a/ and /i/ contexts are separated from one another.

For main groups of similarity are identified in audio alone (SNR –6dB) condition(Figure 9):

1. /a/ context
2. non-nasals in /i/ and /u/ contexts except /izi/
3. /izi/
4. nasals in /i/ and /u/ contexts

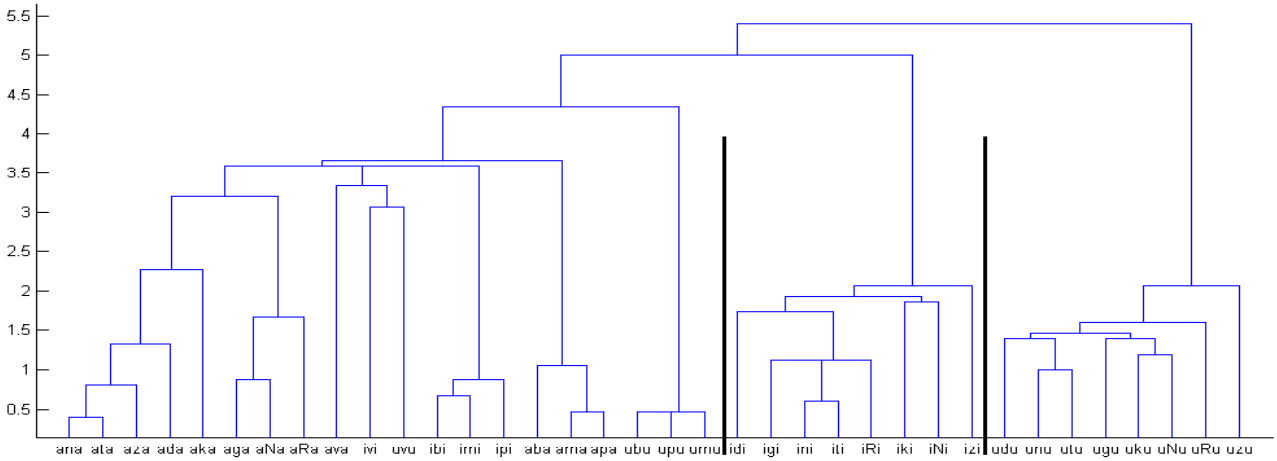

Figure 7: *Confusion tree for natural video alone condition. Main groups are separated by black vertical lines.*

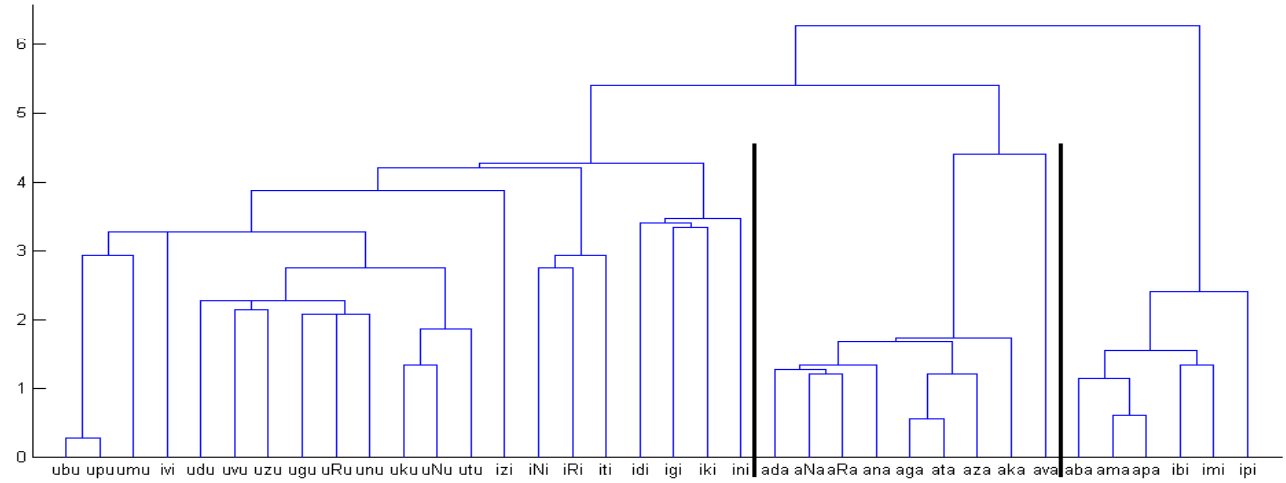

Figure 8: *Confusion tree for synthetic video alone condition. Main groups are separated by black vertical lines.*

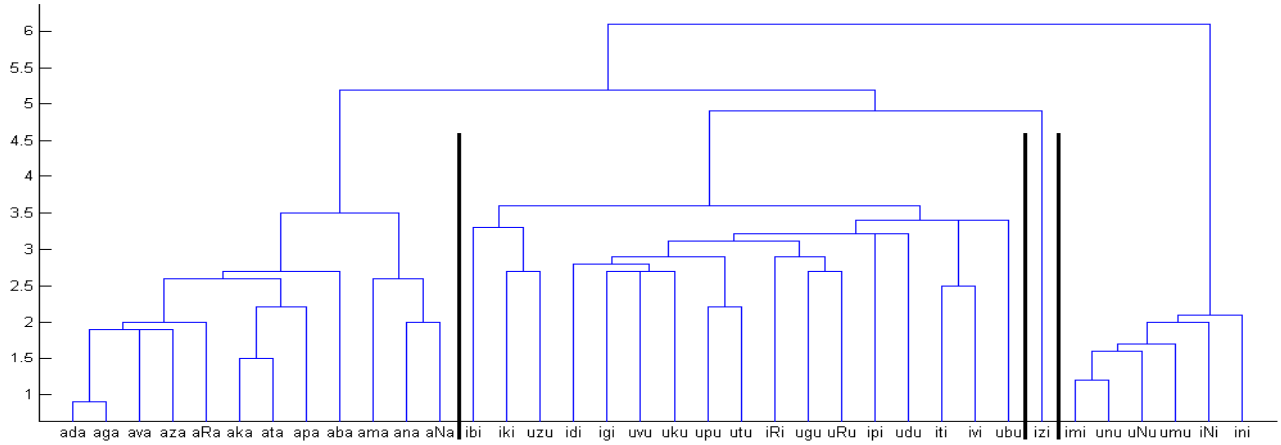

Figure 9: *Confusion tree for audio alone (-6 dB) condition. Main groups are separated by black vertical lines.*

Stimuli in /a/ context can be distinguished auditorily from the other stimuli. In the first group voiced medial consonants – here plosives are separated from fricatives – , voiceless plosives and finally nasals constitute subgroups.

3.6. Face shape analysis

The stimuli were analyzed regarding the geometrical similarity of the displayed consonants. The center of the realization of the consonant in each synthetic video was identified by visual inspection. The first derivates of the articulatory parameters were used to guide the decisions, but as different articulators and articulator combinations were assumed to be crucial for the various consonants, the expert's decision was preferred to a pure automatic approach. A distance matrix was derived as follows: for all pairs of consonants the Euclidian distances between each vertex of the one shape and the same vertex of the other shape were calculated and summed. A tree diagram was built from the distance matrix using the ward method.

3.7. Results: shape analysis

The diagram (Figure 10) shows four main groups:

1. all alveolars and the labiodental in /a/ context and all velars, alveolars and the labiodental in /i/ context
2. all velars in /a/ context
3. all bilabials in /a/ and /i/ context and one in /u/ context
4. all stimuli in /u/ context except /upu/

The first group shows the similarity between stimuli in /a/ and in /i/ contexts and a rudimentary but incomplete subgouping of velars, alveolars, labiodentals and bilabials. The second group shows increasing similarity from plosives to phonemes with oral occlusion to all velars. The third group shows, with the exception of /ipi/, a separation of plosives and nasals. In the fourth group an incomplete separation of bilabials, alveolars and velars can be seen.

4. Conclusions

This paper presents results of intelligibility tests and shape analysis on visually cloned German speech. The clone clearly enhances recognition scores when added to degraded natural audio. However the identification scores lie still below those provided by natural video. Detailed analyses show that velar and uvular consonants are yet identified as often as in natural

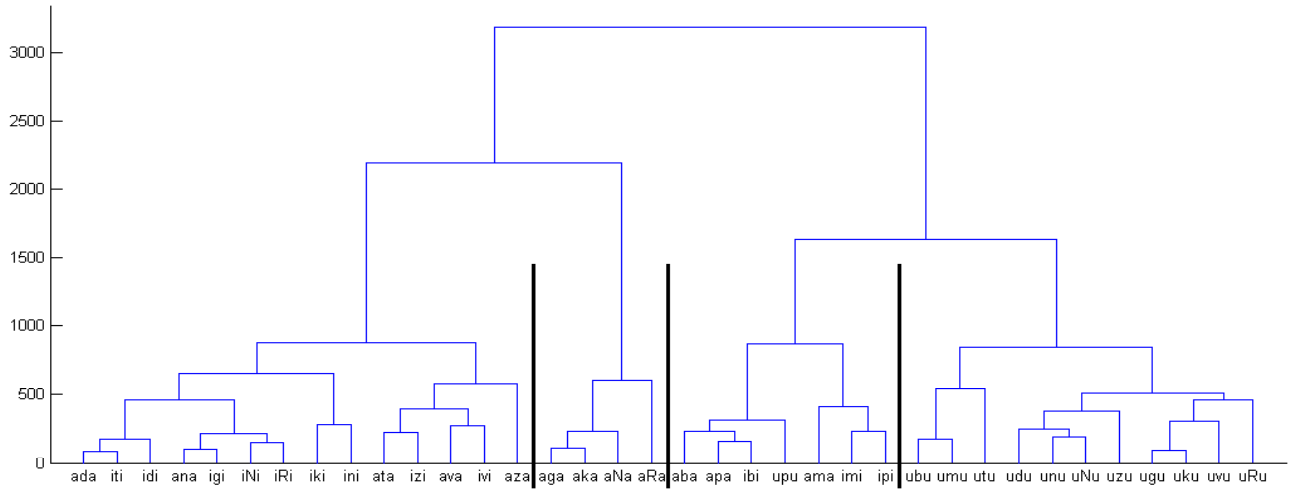

Figure 10: *Confusion tree derived from spatial distances between face meshes. Main groups are separated by black vertical lines.*

speech. It is assumed that this is due to well modeled throat movements which will be investigated explicitly in upcoming experiments. Labial and labiodental consonants are identified best of the synthesized stimuli but nevertheless the lip model still needs improvement. More important, a drop in identification rates for alveolars evidences that the clone needs to be supplemented by a tongue in order to reach the intelligibility of a natural video.

The tree diagrams show typical structures of auditory and visual information, respectively. The audio only presentation shows groupings of vocalic context, voicing, and manner of articulation. The video only presentations show groupings of vocalic context and place of articulation, but the synthetic video results in higher similarity between velars and alveolars and a slightly less clear separation of the vocalic context than the natural video.

The objective evaluation reveals four main groups of similarity where three of them represent geometrically well distinguished properties: rounded lips (although not vowels but consonants in vocalic context are analyzed), bilabial place of articulation in non-rounded context, and velar place of articulation in /a/ context. Where the optical identity of bilabials and of rounded lips was expected, the clear state of velars in /a/ context is somewhat surprising. These results support the assumption that the movements of the tongue root are well visible from the outside and well captured and reconstructed by the described cloning method.

The place of articulation is known to be present in the transition from and to neighboring vowels. The present study shows that the lip rounding is contained in the center of realization of the embedded consonant. This supports the thesis that where in audible speech dynamic information is important, in visible speech static shapes are relatively informative.

5. Acknowledgments

Parts of this work were supported by a grant of the DAAD (Germany) and EGIDE (France) within the PROCOPE program. We thank Christophe Savariaux and Alain Arnal for their technical help in the capture process and Ralf Baumbach who helped with annotating the data.

6. References

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

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

[3] 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.

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

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

[6] Benoît, C. 1996. On the Production and the Perception of Audio-Visual Speech by Man and Machine. In Y. Wang et al. (eds.): *Multimedia & Video Coding*, Plenum Press, New York.

[7] McGurk, H., MacDonald, I. 1976. Hearing Lips and Seeing Voices. *Nature* 264, 746-748.

[8] Bailly, G., Elisei, F., Badin, P., Savariaux, C. 2006. Degrees of Freedom of Facial Movements in Face-to-Face Conversational Speech. *Proceedings of the International Workshop on Multimodal Corpora*, Genoa, 33-36.

[9] Odisio, M. 2005. *Estimation des mouvements du visage d'un locuteur dans une séquence audiovisuelle*. PhD Thesis at the Institut de la Communication Parlée, Institut National Polytechnique, Grenoble.