Visual Information and Redundancy Conveyed by Internal Articulator Dynamics in Synthetic Audiovisual Speech

Katja Grauwinkel, Britta Dewitt, Sascha Fagel

Institute for Speech and Communication, Berlin University of Technology, Germany
Katja.Grauwinkel@tu-berlin.de, brittadewitt@freenet.de, Sascha.Fagel@tu-berlin.de

Abstract
This paper reports results of a study investigating the visual information conveyed by the dynamics of internal articulators. Intelligibility of synthetic audiovisual speech with and without visualization of the internal articulator movements was compared. Additionally speech recognition scores were contrasted before and after a short learning lesson in which articulator trajectories were explained, once with and once without motion of internal articulators. Results show that the motion information of internal articulator dynamics did not lead to significant different recognition scores at first, and that only in case of this additional visual information the training lesson was able to significantly increase visual and audiovisual speech intelligibility. After the learning lesson with all internal articulatory movements the visual recognition could be enhanced to a higher degree than the audiovisual recognition. The absolute increase of visual recognition could not be integrated completely into audiovisual recognition. It could be shown that this was due to redundant information conveyed by auditory and visual sources of information.

Index Terms: talking head, speech visualization, internal articulators, speech intelligibility, audiovisual speech

1. Introduction
Since five decades many investigations have documented that pertinent visual information enhances speech intelligibility when added to audible speech (e.g. [1], [2]). Previous work of the authors could show that this can also be achieved by synthetic audiovisual speech [3]. The present study investigates the gain of visual information which can normally not be observed: the internal articulator dynamics. Computer-based audiovisual speech synthesizers and speech visualization systems provide an interesting tool for investigating bimodal sensory integration, because each parameter of each source of information can easily be manipulated and therefore is experimentally controllable. Furthermore, a talking head that might be used as speech trainer (e.g. in foreign language acquisition or speech therapy) should be able to show the internal articulators in order to explain the production of different speech sounds. Static pictures which are used by speech therapists to explain place and manner of articulation do not consider the coarticulatory interactions of real speech trajectories. Especially the dynamic information of the articulatory displacements of internal articulators are of importance to explain coarticulatory effects. Therefore a talking head was supplemented by the internal passive and active articulators: alveolar ridge, palatum, velum and pharynx wall. To what extent the dynamic information of these internal articulators is capable to enhance speech intelligibility was evaluated in a second step.

2. System Description
The Modular Audiovisual Speech Synthesizer (MASSY) [4] was used as experimental tool in the study. MASSY is a web-based text-to-speech synthesizer with a 3-dimensional animated head. The embedded modules are a phonetic articulation module, an audio synthesis module, a visual articulation module, and a face module. The phonetic articulation module creates phonetic and prosodic information: an appropriate phone chain, phone and pause durations and a fundamental frequency course. From this data, the audio synthesis module, in which the MBROLA speech synthesizer [5] is embedded, generates the audio signal and the visual articulation module generates motion information. The audio signal and the motion information are merged by the face module to create the complete animation. The visual articulation module generates a series of parameter settings for the (virtual) articulators. The parameters which were embedded before the study were:
- tongue tip height,
- tongue dorsum height,
- lip width,
- lip height,
- lower lip retraction and
- lower jaw height.

The face module animates the 3-dimensional head by linearly displacing its vertices. For each articulator a set of displacement vectors for the head’s vertices is defined. Each vertex displacement is a linear combination of the articulators’ displacement vectors. The VRML file format (Virtual Reality Modeling Language) [6] is used as output. The facial skin can be displayed either opaquely or transparently in order to see the internal articulator movements.

In this study the talking head’s articulators were supplemented by the alveolar ridge, hard and soft palate, uvula, and pharynx wall. For recreating the 3-dimensional geometric models of the internal articulators, the results of medical imaging techniques, in particular mid-sagittal MRI slices, were used as templates (for forms, sizes and positions). The system configuration before (opaque and transparent) and after changes (transparent) can be seen in figure 1.

An additional articulatory parameter was assigned to the tongue, i.e. forward/backward movement, and another one was defined for the opening/closing of the velum. For these new motion parameters new values for control parameters had to be defined. Because of the refinement of tongue movement new values for tongue dorsum height had to be assigned, too. The parameter values for the visual articulation module were derived from previous measurement data of electromagnetic articulography [4]. Thus it is based on human articulation movements. The electromagnetic articulography allows to survey articulatory movements at discrete flesh points of articulators very precisely in space and time, even if they take
place inside the mouth. The articulation model implements the dominance principle as suggested by Löfqvist [7] in order to deal with coarticulation.

Figure 1: Talking head in non-transparent and transparent view before (top) and transparent view after (bottom) supplement of internal articulators.

3. Experimental Setup

3.1. Items and conditions

The corpus contained ten German consonants with different places of articulation all being voiced except the alveolar voiceless fricative [S]; as for its voiced counterpart [Z] no orthographic symbol exists in German and hence could not have been presented to the subjects. The consonants [b, d, g, z, s, v, l, m, n, N] were chosen and combined with the vowels [a, i, u] in a vowel-consonant-vowel (VCV) structure, respectively. These vowels were chosen because they span the articulatory/acoustic vowel space. The items were synthesized in the conditions audio (A), visual (V) and audiovisual (AV). The German female MBROLA voice de⁷ was selected for creating the stimuli. The audio signal both in A and AV conditions was embedded in white noise with a signal-to-noise-ratio of 0 dB. On the one hand the visual part of the stimuli were synthesized with all movements of the internal articulators, on the other hand without the movements: forward/backward movement of the tongue, and motion of tongue dorsum and velum. Without these movements only the articulatory movements remain, which can also be observed in a face-to-face conversation.

3.2. Method

20 subjects (23 to 59 years old, mean: 31 years) with normal hearing and normal or corrected-to-normal vision participated in the experiment voluntarily. The subjects had no explicit phonetic knowledge. Stimuli were presented by use of three different quasi-random orders. All the stimuli were presented to the subjects three times before they had to give the answer, where no feedback was given whether or not the answer was correct. The test was designed as a forced choice test. The recognition of the vowels of the VCV-stimuli were not tested.

All subjects were divided into two groups of 10 subjects (groups A and B). The stimuli with movements of the internal articulators were presented to group A, the stimuli without these movements were presented to group B. After this pre-test the groups were again divided into two subgroups (groups A1, A2, B1 and B2). Groups A1 and B1 received a training lesson of about 30 minutes length with the same motion information as in pre-test, respectively. Afterwards they had a break of 20 minutes. Groups A2 and B2 did not perform the training lesson, they only had a break of the same length. Afterwards a post-test was performed. This test was the same as the one performed as pre-test.

3.2.1. Training Lesson

The training lesson was a video presentation in which the articulatory movements for each consonant in each vowel context were explained from the side view, while the internal articulators were shown. In order to explain differences in place and manner of articulation pairs of consonants were set in contrast to one another.

Articulatory movements of all displayed internal articulators were explained to group A1, whereas the motion information of velum, tongue dorsum and forward/backward movements of the tongue were not shown to group B1; only lip, jaw and tongue tip movements were explained to them. The articulation process of each consonant-vowel combination was shown at reduced speed while giving explanations. Then the stimuli were shown like those in the test situation but without noise. The training lesson was not interactive, the subjects were told to listen and watch carefully.

3.3. Analyses

Besides scores for correct stimuli recognition two other measurements were used for analyses, i.e. Visual Integration Coefficient (VIC) and cross-modal redundancy [8].

3.3.1. Visual Integration Coefficient (VIC)

The Visual Integration Coefficient measures the amount of visual information which is integrated by the subjects for recognition in AV condition by equation (1).

$$\text{VIC} = \frac{E_{av} - E_a}{E_v}$$  \hspace{1cm} (1)

where:

- VIC: Visual Integration Coefficient
- $E_{av}$: recognition for AV condition in %
- $E_a$: recognition for A condition in %
- $E_v$: recognition for V condition in %

The absolute enhancement by visual information when added to auditory information ($E_{av}$-$E_a$) is related to the actual measured visual recognition ($E_v$). Redundancies are not taken into account by the VIC as it does not consider whether the non-used fraction of visual information yields information that is redundant to the audio channel or is omitted due to incomplete sensory integration.

3.3.2. Redundancy

Redundancy is calculated by equation (2). It measures the overlap of response distributions to each stimulus in A and V conditions as the responses are assumed to reflect the information that is accessible by subjects.
As group B1 had only limited motion information (movements of internal articulators were not shown), the visualizations of the consonants [b,m], [d,n,l], [g,N], [z,S] and [v], respectively, did not differ from one another, which corresponds to the visual distinguishability of consonants from the outside view. Hence, the recognition scores were also analyzed in terms of viseme classes. Nevertheless, neither the recognition scores for consonants nor the recognition scores for viseme classes changed significantly from pre- to post-test in group B1. In contrast to group A1, group B1 was not able to benefit from training lesson. Without the dynamic information of tongue dorsum height, velum opening/closing and tongue forward/backward movements the recognition scores could not be enhanced significantly.

The increase of recognition scores for group A1 between pre- and post-test was higher for V (30%) than for AV (16%). Hence, the information which was retrieved in visual condition by group A1 could not be integrated completely in audiovisual recognition. For further analyses the Visual

**Table 1:** Recognition scores for all subgroups in pre- and post-test for each condition.

<table>
<thead>
<tr>
<th>sub-groups</th>
<th>tests</th>
<th>condition</th>
<th>A</th>
<th>V</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>AV</td>
</tr>
<tr>
<td>A1</td>
<td>pre</td>
<td>40.7</td>
<td>24.7</td>
<td>52.7</td>
</tr>
<tr>
<td></td>
<td>post</td>
<td>44.0</td>
<td>54.7</td>
<td>68.7</td>
</tr>
<tr>
<td>A2</td>
<td>pre</td>
<td>40.7</td>
<td>22.7</td>
<td>59.3</td>
</tr>
<tr>
<td></td>
<td>post</td>
<td>46.0</td>
<td>30.7</td>
<td>52.0</td>
</tr>
<tr>
<td>B1</td>
<td>pre</td>
<td>44.0</td>
<td>30.7</td>
<td>64.0</td>
</tr>
<tr>
<td></td>
<td>post</td>
<td>41.3</td>
<td>32.7</td>
<td>66.7</td>
</tr>
<tr>
<td>B2</td>
<td>pre</td>
<td>36.0</td>
<td>26.0</td>
<td>54.0</td>
</tr>
<tr>
<td></td>
<td>post</td>
<td>38.0</td>
<td>32.0</td>
<td>57.3</td>
</tr>
</tbody>
</table>

As group B1 had only limited motion information (movements of internal articulators were not shown), the visualizations of the consonants [b,m], [d,n,l], [g,N], [z,S] and [v], respectively, did not differ from one another, which corresponds to the visual distinguishability of consonants from the outside view. Hence, the recognition scores were also analyzed in terms of viseme classes. Nevertheless, neither the recognition scores for consonants nor the recognition scores for viseme classes changed significantly from pre- to post-test in group B1. In contrast to group A1, group B1 was not able to benefit from training lesson. Without the dynamic information of tongue dorsum height, velum opening/closing and tongue forward/backward movements the recognition scores could not be enhanced significantly.

The increase of recognition scores for group A1 between pre- and post-test was higher for V (30%) than for AV (16%). Hence, the information which was retrieved in visual condition by group A1 could not be integrated completely in audiovisual recognition. For further analyses the Visual

**Table 2:** Recognition scores for pre- and post-test conditions AV and V. Non-significantly different scores are displayed in same columns.

<table>
<thead>
<tr>
<th></th>
<th>AV</th>
<th>V</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1 pre</td>
<td>52.7</td>
<td>24.7</td>
</tr>
<tr>
<td>A2 pre</td>
<td>59.3</td>
<td>22.7</td>
</tr>
<tr>
<td>A2 post</td>
<td>52.0</td>
<td>30.7</td>
</tr>
<tr>
<td>A1 post</td>
<td>68.7</td>
<td>54.7</td>
</tr>
</tbody>
</table>

Even the nasals, for which the manner of articulation was visualized by lowering the velum for group A, were identified only marginally better by group A in V and AV condition. In pre-test subjects of group A identified stimuli neither significantly better nor worse than group B. Hence subjects were not able to use this additional motion information on the underlying sound production without further explanation, but at least they also did not get confused by it.
Integration Coefficient (VIC) and cross-modal redundancy were calculated. VIC and redundancy for both groups A1 and B1 with learning lesson in pre- and post-test are displayed in Table 3.

Table 3: VIC and redundancy for groups A1 and B1 in pre- and post-test.

<table>
<thead>
<tr>
<th></th>
<th>VIC pre-test</th>
<th>VIC post-test</th>
<th>Redundancy pre-test</th>
<th>Redundancy post-test</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>0.49</td>
<td>0.45</td>
<td>0.37</td>
<td>0.51</td>
</tr>
<tr>
<td>B1</td>
<td>0.65</td>
<td>0.78</td>
<td>0.35</td>
<td>0.37</td>
</tr>
</tbody>
</table>

The VIC which shows the actual used proportion of \( V \) in \( AV \) only changed marginally from pre- to post-test for group A1. That means that before and after the learning lesson the proportion of the actual measured recognition in \( V \) condition which was integrated in \( AV \) recognition almost remained the same. Consequently, part of the visual information that became accessible by training was integrated in audiovisual recognition. As another part of this additional information was not used it is assumed that also redundancy between auditory and visual information was introduced. Redundancy for group A1 increased from 0.37 in pre-test to 0.51 in post-test. Hence, group A1 was not able to better integrate the visual information, but to better visually recognize the speech stimuli by using the additional information. Because of redundancies in auditory and visual sources of information the audiovisual recognition could not be enhanced as much as the visual recognition and the VIC did not increase.

To compare pre- and post-test for group B1 it is assumed that the little loss in \( A \) recognition was compensated by a better integration of available visual information so that the same \( AV \) recognition score could be achieved. This is expressed in an increased VIC while redundancy is almost the same in pre- and post-test. Although the VIC increased, no additional information was made available for group B1 during the learning lesson. Thus \( AV \) recognition did not increase significantly.

5. Conclusions and Future Work

On the one hand it could be shown that additional information, which is not available in a natural face-to-face communication, cannot be used by subjects to enhance speech intelligibility without further explanation. However, through a short learning lesson a significant enhancement of recognition scores in visual and audiovisual conditions can be achieved. On the other hand it was shown that a limited amount of visual information of internal articular movements does not result in such an enhancement. It is assumed that the perception and interpretation of articular movements which can to some extent also be observed in a natural face-to-face communication cannot be enhanced by a learning lesson of 30 minutes. Furthermore, the additional motion information could be used after the learning lesson to increase audiovisual recognition only to a certain degree, because visual information of internal articulators seems to convey a high amount of information that is redundant to auditory information.

An important finding is that the given visual information of internal articular dynamics could be used in order to interpret articularatory speech production. As visual feedback which normally is not available can be learned from subjects with normal hearing and speaking abilities, the next step will be to investigate whether people with speech or hearing impairment are also able to interpret, and learn from the articularatory dynamics of a talking head.

A growing number of computer based methods for speech visualization are introduced in therapy for voice, speech and language disorders. In a study by Massaro & Light [9] the talking head BALDI [10] was used as speech trainer for children with hearing disorders from 8 to 13 years of age. The subjects’ ability to perceive and produce the test items before and after the training program could be enhanced. Another study by Albert [11] could show that children from 4 to 8 years of age were able to recognize articulatory visual patterns of speech sounds from the visual model of articulation called SpeechTrainer [12]. After a short introduction into the system, which displayed two-dimensional animations of mid-sagittal MRI slices, the children were able to interpret the information in terms of their own sound production. The authors of the present study currently investigate the applicability of talking heads with three-dimensional animations of internal articulator dynamics as a tool of speech visualization in speech therapy for children with sизмatismus interdentalis.

6. References