

Visualization of Internal Articulator Dynamics for Use in Speech Therapy for Children with Sigmatismus Interdentalis

Katja Grauwinkel & Sascha Fagel

Institute for Speech and Communication, Berlin University of Technology, Germany

katja.grauwinkel@tu-berlin.de, sascha.fagel@tu-berlin.de

Abstract

This paper presents an ongoing study in which the applicability of a talking head with its three-dimensional animation of internal articulator dynamics is investigated as a method of speech visualization for use in speech therapy for children with sigmatismus interdentalis. Previous work of the authors showed that people with normal hearing, vision and speaking abilities were able to benefit from visual information provided by the internal articulatory movements of the talking head during speech production after a short learning lesson. In the present study a perception and a production test for children with sigmatismus interdentalis were designed. In a learning lesson the correct production of the alveolar sibilants /s/ and /z/ was auditorily presented and visualized by use of the talking head; the correct tongue position was explained and set in contrast to an interdental production of the sounds. Afterwards a visual identification test was performed and the perception and production tests were repeated. In order to determine the effect of the learning lesson, the production data of three of the children were evaluated by 15 raters. The results of the perception test showed that children were able to auditorily differentiate between correct and wrong production of the alveolar fricatives /s/ and /z/. Furthermore, most of the children are able to visually identify correct and wrong productions of the talking head. The evaluations showed that the learning lesson was capable to improve the sibilant production of two of the three children.

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

1. Introduction

Speech therapists use visual methods to explain articulatory processes so that a specific imagination of the underlying place and manner of articulation of sounds can be learned. The view of the therapist's mouth or the own view in a mirror is regarded as an effective method [1]. Static pictures of isolated articulatory positions from mid-sagittal or frontal view are used by speech therapists in order to explain the positions of the articulators in different articulatory contexts. Nevertheless, such static pictures do not consider the coarticulatory interactions of real speech trajectories. Especially the dynamic information of the articulatory displacements of internal articulators are of importance in order to explain coarticulatory effects. Computer-based audiovisual speech synthesizers and speech visualization systems provide an interesting tool for investigating the applicability as speech trainer (e.g. in foreign language acquisition or speech therapy), if the dynamic information of internal articulator displacements can be visualized.

A growing number of computer-based methods of speech visualization are introduced in therapy for voice, speech and language disorders. In a study by Massaro and Light [2] the

talking head BALDI [3] 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 could be enhanced after a learning program which contained six hours of lessons over the course of 21 weeks. Other studies could show that on the one hand an adult with apraxia [4] and on the other hand children from 4 to 8 years of age with heterogeneous speech and language disorders [5] were able to recognize articulatory visual patterns of speech sounds from the visual model of articulation called SpeechTrainer [6]. SpeechTrainer displays two-dimensional animations of sketches of mid-sagittal MRI slices.

The authors of the present study investigate the applicability of a talking head with three-dimensional animations of internal articulator dynamics for use as a method of speech visualization in speech therapy for children with sigmatismus interdentalis. In previous work [7] the talking head called MASSY (*Modular Audiovisual Speech SYnthesizer*) [8] was supplemented by the internal passive and active articulators: alveolar ridge, palate, velum and pharynx wall. In this previous study it could be shown that after a short learning lesson the dynamic information of the internal articulators was capable to enhance speech intelligibility for adults with normal hearing and speaking abilities. The given visual information of internal articulator dynamics could be used in order to interpret articulatory speech production. Since visual feedback of a kind that normally is not available can be learned by adults with normal hearing and speaking abilities, the authors of the present study investigated whether children are also able to interpret and learn from the articulatory dynamics of the talking head. Furthermore, the present work investigates whether children with sigmatismus interdentalis are able to benefit from the talking head as method for speech visualization. The working hypothesis of the present study is that synthetic visible speech can be beneficially included in the training of speech production.

2. System Description

The talking head MASSY that was used as experimental tool in the study is a web-based text-to-speech synthesizer with a three-dimensional animated head. The modules used in this study were the audio synthesis module, the visual articulation module, and the face module. In the present work the system was driven by phonetic and prosodic information of phone labels, phone durations and fundamental frequency courses instead of plain text. The audio synthesis module, in which the MBROLA speech synthesizer 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 facial skin can be displayed either opaquely or transparently in order to see the internal articulator movements. The parameters which define the

(virtual) articulator positions are: lower jaw height, lip width, lip height, lower lip retraction, tongue tip height, tongue dorsum height, velum height, and tongue forward displacement. Passive Articulators are alveolar ridge, palate, and pharynx wall. The parameter values for the visual articulation module were derived from previous measurement data of electromagnetic articulography [8]. Thus it is based on human articulation movements. The electromagnetic articulography allows to survey articulatory movements at discrete flesh points of the 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 [9] in order to take coarticulation into account.

3. Experimental Setup

3.1. Method

Seven children from 4;11 to 7;11 years of age participated in the study. Two of them (subjects *S* and *C*) have a sigmatismus interdentalis. They were selected by speech-language pathologists, who diagnosed the sigmatism. None of the two children had been to speech therapy before and/or while participating in the study. Testing and training was carried out individually. MASSY did not serve as instructional agent for testing and training, because it turned out that children – especially beneath the age of six – had difficulties to pay attention only to the monitor during the entire test. The children were better able to concentrate on the task in an interactive session with the experimenter.

The whole test was carried out on a notebook with *Creative CSW 5300 Travelsound* external loudspeakers. The speech productions were recorded with an *AKG C4201* microphone at 44.1 kHz/16 bit/mono. The whole test started with a perception test to determine whether correct and wrong productions of the /s,z/-sound (/s/ or /z/, respectively) can be differentiated auditorily by the children. Then a production test determined the actual state of the children's sound productions. In a learning lesson the correct production of the /z/ sound was auditorily presented and visualized by use of the talking head. The children's ability to recognize the visual information conveyed by the talking head was determined in a visual identification test. Afterwards the perception and production tests were repeated (post-test) in order to determine the effect of the learning lesson. After one week the learning lesson and the post-test were repeated. In total, each child performed the perception and production tests three times and the learning lesson two times across two test sessions. For the youngest child (subject *C*) the test was shortened and divided into two parts, because *C* was not able to keep attention during the entire test. Therefore, subject *C* performed the perception and production test in one session and the learning lesson, the visual identification test and the post-test in a second session, which took place after one week. The stimuli for speech perception and production tests consisted of German words containing the /s/ word-final, the /z/ word-initial and both /s/ and /z/ word-medial.

3.1.1. Perception Test





For creating the test items for the perception test the first author produced six words with correctly produced /s,z/-sound and with the /s,z/-sound produced with the tongue protruded between the front teeth (interdental). In order to confirm that the interdental production could be identified as lisped speech, the test stimuli were classified by five subjects,

all members of the Institute for Speech and Communication. All test stimuli were identified correctly by all subjects (100% correct classification), so they were considered to be adequate for the perception test. In the perception test the stimuli were presented over loudspeakers in a quasi random order. The children had to decide whether the /s,z/-sound was correctly spoken or not.

3.1.2. Production Test

In order to record the children's /s,z/-sound productions, twelve pictures were presented on the notebook display in a quasi random order. The children were asked to attend to the display and say aloud what they were seeing. Examples of the words which the children were asked to say aloud are displayed in figure 1. Like in the perception test, these words contained the /s/-sound word-final (e.g. /haUs/ engl. "house"), the /z/-sound word-initial (e.g. /zOn@/ engl. "sun") and both /s/- and /z/-sound word-medial (e.g. /tas@/ engl. "cup" or /kE:z@/ engl. "cheese").

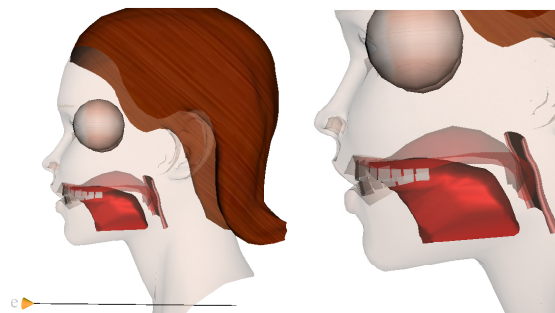
Figure 1: Examples of pictures which were presented to the children in order to record their /s,z/-sound productions.

			
/haUs/	/zOn@/	/tas@/	/kE:z@/
word-final	word-initial	word-medial	

3.1.3. Learning Lesson

Although the /s,z/-sound can be produced in different ways, a prototypical production was always used to explain the articulators' positions: the tongue lying behind the upper incisors not touching them. Exemplarily for both voiced and voiceless counterparts, the place of articulation was explained along with the voiced /z/-sound. First the talking head was presented from side view with transparent skin and the articulators were shown and explained. Then the articulators' positions for the /z/-sound were explained once in the context of the word "Nase" (/na:z@/, engl. "nose") and once as isolated segment. A static picture of the /z/-sound was presented to give a closer explanation on the place of articulation (see figure 2). Afterwards the correct tongue position was set in contrast to an interdental place of articulation of the /z/-sound. The interdental production was presented visually and auditorily. Finally, the trajectory from non-correct to correct place of articulation was shown. The children were asked to reproduce this motion and to say /z/ and /zi:b/ ("Sieb" engl. "sieve") when the tongue was at the targeted position.

Figure 2: Talking head in transparent view showing the place of articulation for /s,z/.



3.1.4. Visual Identification Test

After the learning lesson a visual identification test was performed. The stimuli were presented visually-alone. The correct (prototypical) and the incorrect (interdental) places of articulation were shown once in static pictures and once with dynamic information. On the one hand the dynamic information was the movement from neutral articulatory position to the target position and back. On the other hand the trajectories in the context /a:/ - target position - /@/ were shown. The children were asked to say whether the articulatory positions/trajectories were correct or not.

3.1.5. Evaluation of Speech Productions

An evaluation test was performed in order to examine the degree of lisping and potential qualitative variations in the /s,z/-sound productions of the children with sigmatismus interdentalis (subject *C* and *S*) before and after the learning lessons. Since it is expected that children without sigmatismus will not differ significantly between the different test trials, because their production is already accurate, the evaluation test was performed for the two children with sigmatismus interdentalis only. Nevertheless, two experts of speech science evaluated whether the children without sigmatismus in fact were able to produce the /s,z/-sound accurately. In one case the experts decided to let the speech productions also be evaluated in the evaluation test together with the speech production recordings of the children with sigmatismus. For the child without sigmatismus only pre- and post-test productions were taken into account, in order to reduce the evaluation test duration. Word productions of the three children, and pre-/mid-/post-test utterances were randomized and presented auditorily one by one to 15 subjects. The subjects had normal hearing abilities, they aged from 21 to 63 years (mean 30). The degree of lisping was evaluated on a five-point-scale (from 1="not at all" to 5="very strong"). The raters had no knowledge of details of the word production experiment.

4. Results

4.1. Perception Test

The perception test revealed that 93.5% of all stimuli were identified correctly. The misidentifications mostly originated from one child (subject *C*). From 12 stimuli *C* misidentified six in the first and four in the second turn of the perception test. With 4;11 years of age *C* was the youngest of the children. It is not obvious whether *C* was not able to hear the difference, whether the task was too difficult, or whether this was due to a lack of attention. The investigator who was present during the test had the impression that *C* was able to hear the difference but that he alternated between the two answer categories more or less for fun. The second child with sigmatismus interdentalis (subject *S*) was able to differentiate between both kinds of stimuli to 100%. *S* was aware of the fact that the lisped stimuli sounded the way like her own /s,z/-sounds. Table 1 shows the identification results for the three perception tests trials (pre-, mid- and post-test) for subject *C*, for subjects except *C*, and for all subjects (rightmost column). The differences between the three test trials are not significant. Except for subject *C*, misidentifications occurred only sporadically and all children were well able to auditorily differentiate between the lisped and non-lisped test stimuli.

Table 1: Identification scores for subject *C*, subjects except *C* and in total in the three trials of the perception test.

	correct identification for perception		
	subject <i>C</i>	others	total
pre	50.0%	95.8%	89.3%
mid	66.7%	98.6%	94.0%
post	-	97.2%	97.2%
total	58.4%	97.2%	93.5%

4.2. Visual Identification Test

Results for the visual identification test are displayed in table 2. Besides subject *C*, who performed the visual identification test once, only one child (subject *T*) had problems solving this task in the first trial. One week later in the second trial, *T* was able to identify all presented stimuli. The other children identified all stimuli in both trials correctly. Although this task can be considered as a simple pattern selection task, it is not as trivial because of the different amount of dynamic visual information which is conveyed. Nevertheless, the 3D presentation and animation of the internal articulators which were completely unknown to the children before, could be interpreted at least after the second learning lesson. The children were able to identify the correct place of articulation of the /s,z/-sound production. Whether children are also able to convert the visual information into their own sound production was investigated by analyses of the evaluation test.

Table 2: Results for the visual identification test for subject *C*, subject *T*, all other subjects and in total

	Correct visual identification (<i>number of stimuli</i>)			
	subject <i>C</i>	subject <i>T</i>	others	total
trial 1	66.7% (<i>N</i> =6)	33.3% (<i>N</i> =6)	100.0% (<i>N</i> =30)	85.7% (<i>N</i> =42)
trial 2	-	100.0% (<i>N</i> =6)	100.0% (<i>N</i> =30)	100.0% (<i>N</i> =36)

4.3. Evaluation Test

Evaluations of the different test trials from the speech production recordings of subjects *A*, *C* and *S* were compared in order to investigate whether the children were able to benefit from the visual information conveyed by the talking head. Results are displayed in table 3. For statistical significance the Wilcoxon signed ranks test was performed. Non-significantly different mean scores of each subject in table 3 are displayed in same sub-columns of mean.

As can be seen, the mean values of the /s,z/-productions of subject *A* were rated significantly higher ($p < 0.01$) in pre-test (mean 1.54) than in post-test (mean 1.28). Even with a mean evaluation of 1.54 in pre-test, the /s,z/-productions of subject *A* can be regarded as being appropriate at this stage of speech acquisition (subject *A* is 6;11 years old). Most of subject *A*'s sound productions were rated with 1 (median and mode are 1). At that age sporadically occurring mispronunciations are not regarded as pathological, but are considered as part of normal speech development. Nevertheless, the mean values of the ratings decrease from pre- to post-test. Hence, subject *A* could improve the /s,z/-production after the learning lesson.

The /s,z/-productions of subject *C* were evaluated with a mean value of 3.48 in pre-test and a non-significantly higher value of 3.76 after the learning lesson. The median value was

4 in both test trials. Thus, subject *C* did not benefit from the learning lesson. During the learning lesson subject *C* had difficulties in placing the tongue at the target position, although *C* understood where it has to be placed. It was assumed that this was due to an insufficient motor control of the tongue. Therefore it was recommended to the parents of *C* to consult a speech therapist in order to train the tongue's motor control.

The evaluation scores of subject *S* decreased significantly from pre-test (mean 2.79, median 3) to mid-test (mean 1.95, median 2) and further to post-test (mean 1.53, median 1). With a mean of 1.53 and a median of 1 in post-test, *S* reaches evaluation scores that can be compared to those of a child with normal speech development. *S* was able to interpret and learn from the talking head's articulatory visualizations of the /s,z/-production.

Table 3: Ratings of the sound productions of subjects *A*, *C* and *S*: mean, median (Mdn), mode, and standard deviation (s.d.) for different test trials (pre-, mid-, post-test). Non-significantly different mean scores of each subject, are displayed in same sub-columns of mean. The scale for evaluating the degree of lisping ranged from 1="not at all" to 5="very strong".

		Mean	Mdn	Mode	s.d.
<i>A</i>	pre	1.54	1	1	0.79
	post	1.28	1	1	0.51
<i>C</i>	pre	3.48	4	5	1.24
	mid	3.76	4	5	1.19
<i>S</i>	pre	2.79	3	2	1.21
	mid	1.95	2	1	0.95
	post	1.53	1	1	0.81

5. Discussion and Conclusions

We are aware that the small amount of data examined in this study only permits tentative conclusions. Sigmatism has various physiological and psychological causes [10]. Due to individual characteristics and various causes of sigmatism interdentalis, the analyses in this study must be regarded as case studies. One child was not able to improve the production of /s,z/. It was assumed that this was due to an insufficient motor control of the tongue and hence could not be learned by one training lesson. A consultation with a speech therapist was recommended for this child. Especially for sigmatism that is caused by pervasive physiological developmental disorders, a more intense (long-term) speech therapy is indicated. Subject *S* was aware of the fact that its own productions of /s,z/ sounded different than those from other children. *S* said that it is very difficult and tiring to perform the correct production. At the first meeting *S* did not consider it necessary to change her /s,z/-production. After the second meeting *S* regarded it as necessary and wanted to train the /s,z/-production with a speech therapist further. An informal interview with the children turned out that they all were fascinated by the talking head. Some of the children asked for more training lessons, next time addressing the sound /r/ for example. A positive attitude is a precondition for a successful training, especially for training with children.

Correct perception of all phonemes in a language is essential for spoken language learning. All participating children either with or without sigmatism were able to differentiate between lisped and non-lisped speech stimuli. For the participants of this study it can be assumed that lisping is not caused by a disability to auditorily differentiate

correct and wrong productions, what often comes along with a mild hearing disorder involving high frequencies.

The main goal of the study was to investigate the applicability of a talking head as a method of speech visualization for use in speech therapy. Ratings of the post-test productions of two of the three children were significantly lower than pre-test ratings, indicating a significant learning effect. Therefore the talking head is assumed to be a useful tool for speech visualization for children in speech therapy. For further use as a method for speech visualization in speech therapy, the talking head can be modified to suit individual needs. In future work the 3D presentation and animation of the talking head's internal articulators will be compared to a two-dimensional presentation and animation. Furthermore, future work will be dedicated to a more detailed analysis on a larger amount of data in order to gain more general results. Given that the analyses presented here must be regarded as case studies and that we did not have a control group, it is always possible that some of the learning effects occurred independently of our speech visualization method and was simply based on routine practice. However, the results provided some evidence that the improvement must at least partly be due to the visualization of place of articulation features during /s,z/-production.

6. Acknowledgements

This research is carried out within the project "Forschungsstand und zukünftige Entwicklung von computeranimierten Sprechbewegungen in realen Anwendungen" funded by the German research council (BMBF).

7. References

- [1] Rosenbek, J.C., Wertz, R.T., "Treatment of apraxia of speech in adults", Clinical Aphasiology Conference Proc., Madison, 191-198, 1972.
- [2] Massaro, D.W., Light, J., "Using Visible Speech to Train Perception and Production of Speech for Individuals With Hearing Loss", J. Speech, Language, and Hearing Research (47): 304-320, 2004.
- [3] Cohen, M.M., Beskow, J., and Massaro, D.W., "Recent developments in facial animation: An inside view", AVSP, Sydney, 1998.
- [4] Gotto, J., "PC-gestützte Therapie bei Sprechapraxie – eine Einzelfallstudie". Diploma thesis at the RWTH Aachen, 2004.
- [5] Albert, S., "Einsatz eines visuellen Artikulationsmodells in der Artikulationstherapie bei Kindern", Diploma thesis at the RWTH Aachen, 2005.
- [6] Kröger, B.J., "Ein phonetisches Modell der Sprachproduktion", Niemeyer Verlag, Tübingen, 1998.
- [7] Grauwinkel, K., Dewitt, B., Fagel, S., "Visualization of Internal Articulator Dynamics and its Intelligibility in Synthetic Audiovisual Speech", Proc. ICPhS, Saarbrücken, 2007 (accepted).
- [8] Fagel, S., Clemens, C., "An Articulation Model for Audiovisual Speech Synthesis - Determination, Adjustment, Evaluation", Speech Comm. (44): 141-154, 2004.
- [9] Löfqvist, A., "Speech as Audible Gestures", In W. J. Hardcastle, A. Marchal (eds.), Speech Production and Speech Modeling, Dordrecht: Kluwer, 1990.
- [10] Schindler, A., "Störungen des Spracherwerbs", Deutsche Gesellschaft für Sprachheilpädagogik. Lechte Druck, Emsdetten, 1998.