A 3-D Virtual Head as a Tool for Speech Therapy for Children

Sascha Fagel & Katja Madany

Department of Language and Communication, Berlin Institute of Technology, Germany

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

Abstract

A virtual talking head was adapted to show articulatory movements of the lips, jaw, tongue, and velum. A JavaScript program with a graphical user interface in HTML was developed for speech therapists to use the talking head as a tool in speech therapy. In an initial study children's productions of words containing the sounds /s/ and /z/ were recorded and evaluated before and after two short learning lessons with an experimenter using the virtual head to explain the correct (prototypic) pronunciation of these sounds. Results show that several children could significantly enhance their speech production of the /s,z/ sound.

Index Terms: computer animated speech, speech therapy, talking head

1. Introduction

Speaking is a physiological procedure that manifests in the acoustic as well as in the optic domain. Hence, human speech is both audible and visible. Several properties of audible speech – especially intelligibility – can be enhanced by the visibility of natural speech movements [1] or synthetic speech movements [2].

A two-dimensional display of the tongue position was used for speech therapy since the 1980ies [3]. With current standard computer graphics technology, surfaces of 3-D objects can be displayed transparently. In this way a look through the facial skin of a virtual head can show articulators, which are usually hidden inside the oral cavity. This functionality enables a talking head to explain the articulatory movements that are necessary to produce a given phone chain in 3-D without slicing the head.

2. The Speech Training Software "Vivian"

Vivian is a software tool that was developed for speech therapists to show and explain articulatory processes to patients with phonetic disorders. The software is written in HTML/JavaScript and hence is platform independent and can be run from the program's CD-ROM without installation. The software requires a standard web browser equipped with a 3-D plug-in for the display of VRML animations. No persistent internet connection is needed.

2.1. The Graphical User Interface

The graphical user interface was designed to the needs of the therapist. Figure 1 shows a screenshot. The left column of the screen is dedicated for controlling the software. A table of speech sounds, currently all German consonants and four affricates (in SAMPA: C = {p, b, m, t, d, n, k, g, N, f, v, s, z, S, C, j, x, R, l, h, pf, ts, tS, ks}) is located on top of the control section. The therapist selects the sound of interest and gets a list of items that contain this sound. The list is structured in up to five sections (depending whether or not the sound appears in the respective position due to phonotactic constraints):

- "isoliert" lists the phone alone (followed by a schwa only in case of sounds that need a release) and in clusters with other consonants without context.
- "initial", "medial" and "final" list words where the sound occurs in word-initial, word-medial or word-final position, respectively.
- The section "cluster" contains words where the sound occurs as first sound in word-initial consonantal clusters.

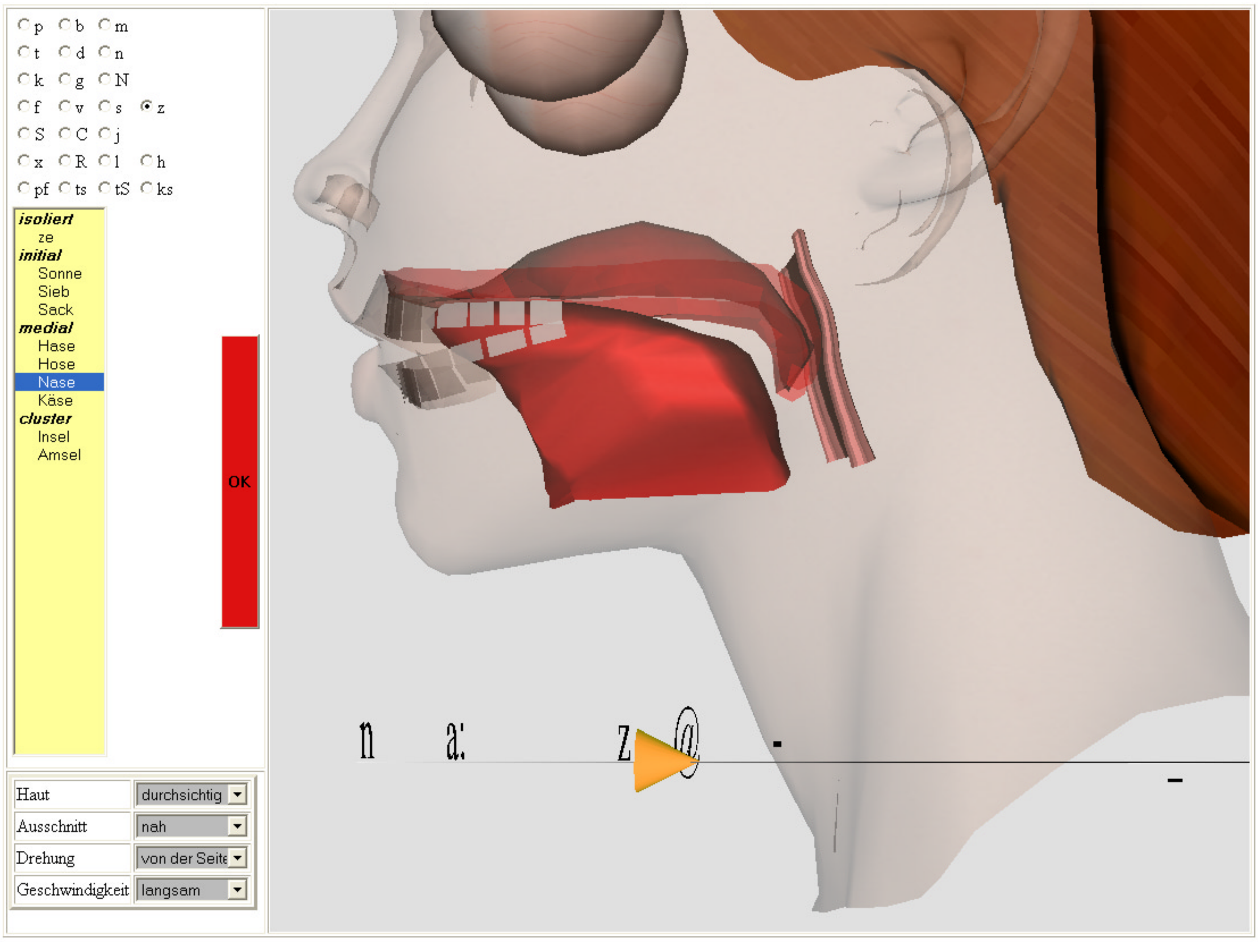
The "OK" button to start the display of the selected word or sound is placed close to the word list. Underneath the word list and the "OK" button the therapist can change the display settings:

- Skin ("Haut"): normal or transparent
- Detail ("Ausschnitt"): full face or zoom ("weit" or "nah")
- Orientation ("Drehung"): frontal or side ("von vorne" or "von der Seite")
- Speed ("Geschwindigkeit"): normal or slow ("normal or "langsam")

The main area of the software on the right of the screen shows the animated head uttering the selected item in the required setting. On the bottom of this area a time slider enables the therapist to scroll through the utterance in a desired speed with moving articulators but without slow motion sound. As a guidance the phonemic symbols (also in SAMPA) are displayed at those points on the time slider where the respective phones occur.

2.2. Training Material

The training material was selected according to standard diagnostic literature (PLAKSS: Psycholinguistic Analysis of Children's Speech Disorders [4], AVAK: Analysis Procedure for Speech Disorders of Children [5] and Patholinguistic Diagnostics for Disorders of Speech Development [6]. Simple nouns were chosen for the word-initial, word-medial, word-final and consonant clusters. Those words assumed to be known by children and those that can easily be displayed by pictures (for a future extension of Vivian) were preferred. All German consonants and their phonotactically possible combinations were included in an isolated form.



Haut: durchsichtig
Ausschnitt: nah
Drehung: von der Seite
Geschwindigkeit: langsam

Figure 1: *Screenshot of Vivian's graphical user interface: Settings are made in the left column of the screen, the articulation is displayed on the right side.*

2.3. Generation of Audiovisual Speech Examples

The system that is used to generate the audiovisual speech utterances is MASSY (Modular Audiovisual Speech SYnthesizer) [7]. The system consists of four modules for: 1) phonetic transcription (and prosody), 2) audio synthesis, 2) visual articulation, and 4) the virtual face. A plain text serves as system input. The phonetic module generates an appropriate phone chain, phone and pause durations and a fundamental frequency course from the text. This functionality is realized by embedding the external program txt2pho from the HADIFIX speech synthesizer [8]. Alternatively, a phonetic transcription with phone durations (and f0 values) can be entered in the system directly. This is especially useful if non-words or specific pronunciations shall be generated. The limiting factor for the generation of an utterance with a specific pronunciation is the inventory of the used voice database. The audio synthesis module generates the audio signal from the phonetic and prosodic data. The audio is rendered by the MBROLA speech synthesis engine [9]. The visual articulation module generates motion information in terms of control parameters for virtual articulators. These articulatory parameters are generated with a simplified dominance model adapted from Cohen and Massaro [10]. The face module moves the virtual articulators of the virtual head according to the control parameters and adds the audio signal to create the complete audiovisual speech output. The virtual head was created by Head Designer [11], a plugin for 3ds Max [12]. The tongue, the velum and a part of the pharyngeal wall was manually designed on the basis of midsagittal images from MRI (magnetic resonance imaging).

The articulation parameters of the virtual head used in the present paper are:

- lip width
- jaw height
- lip closure (independent from the jaw height)
- lower lip retraction
- tongue tip height
- tongue back height
- tongue advancement
- velum height

3. Initial Evaluation

This section describes part of a study that was concerned with testing and evaluation of real world applications of talking heads [13]. An initial evaluation investigated the applicability of the virtual head for speech therapy. Only one specific speech disorder was chosen at this stage: sigmatismus interdentalis is a pathological production of the /s/ and /z/ sounds (no interdental phoneme exists in German) – one of the most frequent speech disorders in children in Germany.

3.1. Method

Eight children from 4;11 to 7;11 years of age participated in the study. Testing and training was carried out individually by one experimenter in an interactive session with each child. The test was carried out on a notebook with external microphone and loudspeakers. It was verified that the child could differentiate auditorily between correct alveolar and wrong interdental productions of the /s,z/ sound before the test started. The test itself consisted of three phases as follows.

A initial production test determined the actual state of the child's sound production. In a learning lesson the correct production of the /z/ sound was auditorily presented and visualized by use of the talking head. Afterwards the production tests was repeated in order to determine the effect of the learning lesson. After one week the learning lesson and the production test were repeated. In total, each child performed the production test 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 production test in one session and the learning lesson and the repeated production test in a second session, which took place after one week.

3.1.1. Production Test

The speech productions were recorded with a condenser microphone at 44.1 kHz/16 bit/mono. In order to record the child'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. 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").

3.1.2. Learning Lesson

The experimenter used the full face as well as zoomed display of the virtual head in side view with transparent skin. The program was controlled by a predecessor version of the above described graphical user interface. Firstly, the virtual head introduced itself by saying an initial hello statement with non-transparent skin and showing the sequence "halalala" once with non-transparent and once with transparent skin. In addition to the word "Nase" (nose) and the sound in isolation the wrong interdental realization and a transition from an interdental to the alveolar fricative was used. 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.

3.1.3. 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 before and after the learning lessons. Word productions of the children and from the three repetitions of the production test were randomized and presented auditorily one by one to 15 listeners. 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 production experiment.

3.2. Evaluation Test Results

Evaluations of the three production test passes from the speech production recordings of subjects reveal whether the children were able to benefit from the visual information conveyed by the talking head. The Wilcoxon signed ranks test for non-parametric data was used to determine statistical significance. For better comparison table 1 displays mean degrees of lisping (i.e. listeners' ratings). Means of non-significantly different values of each child are displayed in same sub-columns.

Table 1. *Results of the evaluation of the children's sound productions (mean degree of lisping per child and per test pass) and remarks whether the child was under therapy during the test and whether he or she has myofunctional disorders. In cases where mean values for different test passes of a child are significantly different, these values appear in different sub-columns. If a child performed non-significantly different in two test passes, the according mean values are displayed in the same sub-column.*

child	test pass	mean degree of lisping			under therapy
c	1	3.48			no; myofunctional disorders
	2	3.76			
	3				
an	1	1.54			no
	2		1.32		
	3		1.28		
ant	1	1.76			no
	2		1.46		
	3	1.60			
le	1	1.57			no
	2	1.58			
	3	1.52			
lin	1	3.66			no
	2		3.24		
	3		3.18		
lu	1	3.58			yes: myofunctional disorders
	2		2.26		
	3			2.54	
s	1	2.79			no
	2		1.95		
	3			1.53	
ta	1	1.44			no
	2		1.16		
	3		1.19		

When entering the test, the mean degree of lisping was between 1 and 2 for four children, for one child it was between 2 and 3, and for three children between 3 and 4. There is no noticeable dependency of these values before the first learning lesson and the possible changes over the test.

The evaluations of the /s,z/ productions of one child (*c*) were not significantly different before and after the learning lesson. Thus, subject *c* did not benefit from the one learning lesson he received (*c* passed a shortened procedure as mentioned above). 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 myofunctional disorders. Therefore it was recommended to the parents of *c* to consult a speech therapist in order to train the tongue's motor control. One more child (*le*) did not show a significant difference of the mean degree of lisping between the passes of the production test.

The first learning lesson had a significantly positive effect on the productions of six of the eight children. Their mean degrees of lisping decreased significantly from the first test to the second test (over the first learning lesson). For one of these children (*ant*) the mean degree of lisping after the second learning lesson was non-significantly lower than the initial score. For another child (*lu*) the mean degree of lisping increased from the second to the third pass of the test, but still showing a significant decrease compared to the initial value. For three children (*an*, *lin* and *ta*) the positive effect persisted after the second learning lesson, for one child (*s*) the mean degree of lisping decreased significantly further from the second to the third test i.e. over the second learning lesson. Over all children (excluding the one that did not complete all test passes) the mean degree of lisping was 2.33 before the first learning lesson, 1.85 after the first, and 1.83 after the second learning lesson.

The small amount of data examined in this study only permits tentative conclusions. The results can be seen as a collection of case studies as the children that participated in the study entered with different prerequisites. However, six of the eight children could interpret and learn from the talking head’s articulatory visualizations of the /s,z/ production.

Sigmatism has various physiological and psychological causes [14]. 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. Especially for sigmatism that is caused by pervasive physiological developmental disorders, a more intense (long-term) speech therapy is indicated. Another child was aware of the fact that its own productions of /s,z/ sounded different from those of other children. She did not consider it necessary to change her /s,z/ production in the first meeting but after the second meeting. Informal interviews 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 other sounds. A positive attitude – a precondition for a successful training, especially for training with children – could be observed.

4. Discussion and Outlook

The present paper describes a new tool for speech therapy that was designed in co-operation with speech therapists to meet their needs. It gives him or her a huge and still extensible variety of synthesized utterances at hand that can be used to visualize articulatory movements inside the oral cavity. The outcomes of the initial evaluation of the system indicate that at least for several children and for training of the correct pronunciation of the /s,z/ sound a three-dimensional talking head is an applicable tool for speech therapy. It is possible that some of the learning effects occurred independently of the speech visualization method – a simple practice effect is unlikely due to the short duration of the learning lessons of only five minutes each.

Although not explicitly tested, children seem to be able to learn the display of internal articulators quickly, which is coherent with results obtained by Tarabalka et al. [15] (and by Grauwinkel et al. [16] with adult subjects). However, comparisons of this kind of display to results obtained with a normal view of a synthetic head – also Tarabalka et al. [15] and Massaro and Light [17] – could not give clear evidence for the benefit of the display of internal articulators.

The next stage of evaluation will be carried out together with the academic speech therapy practice of the LMU Munich. It will incorporate a higher number of children, various speech disorders, and a control group treated without the presented software tool.

5. References

[1] Sumby, W. and Pollack, I. (1954). “Visual Contribution to Speech Intelligibility in Noise”, Journal of the Acoustical Society of America, 26:212-215.

[2] Benoît, C., “On the Production and the Perception of Audio-Visual Speech by Man and Machine”, in: H. Bertoni, Y. Wang and S. Panwar (Eds.): Multimedia and Video Coding. Plenum Press, New York, 1996.

[3] Heike, G., “Computergestützte Therapie”, in: Grohnfeldt, M. (Ed.): Störungen der Aussprache, Handbuch der Sprachtherapie 2, Wissenschaftsverlag Volker Spiess, Berlin, 1990.

[4] Fox, A.V., Kindliche Aussprachestörungen, Verlag Schulz-Kirchner, Idstein, 2007.

[5] Hacker, D. and Wilgermein, H., Aussprachestörungen bei Kindern, Ernst Reinhardt Verlag, München, 2001.

[6] Siegmüller, J. and Kauschke, C., Patholinguistische Therapie bei Sprachentwicklungsstörungen, Elsevier, Munich, 2006.

[7] Fagel, S., Audiovisuelle Sprachsynthese - Systementwicklung und –bewertung, Logos Verlag, Berlin, 2004.

[8] Portele, T., “Das Sprachsynthesesystem Hadfix”, Sprache und Datenverarbeitung, 21:5-23, 1997.

[9] The MBROLA Project, URL http://tcts.fpms.ac.be/synthesis/mbrola.html, 2005.

[10] Cohen, M. M. and Massaro, D. W., “Modeling Coarticulation in Synthetic Visual Speech”, in N. M. Thalmann, D. Thalmann (Eds.): Models and Techniques in Computer Animation, 139-156, Springer-Verlag, Berlin, 1993.

[11] Digimation Head Designer, URL http://www.digimation.com

[12] Autodesk 3ds Max, URL http://www.autodesk.de/3dsmax

[13] Fagel, S. and Madany, K., Computeranimierte Sprechbewegungen in realen Anwendungen, Universitätsverlag der Technischen Universität Berlin, Berlin, 2008.

[14] Schindler, A., “Störungen des Spracherwerbs”, Deutsche Gesellschaft für Sprachheilpädagogik / Lechte Druck, Emsdetten, 1998.

[15] Tarabalka, Y., Badin, P., Elisei, F. and Bailly, G., “Can you read tongue movements? Evaluation of the contribution of tongue display to speech understanding”, Proceedings of Conférence internationale sur l'accessibilité et les systèmes de suppléance aux personnes en situation de handicaps (ASSISTH'2007), Toulouse, pp. 187-193, 2007.

[16] Grauwinkel, K., Dewitt, B., and Fagel, S., “Visual Information and Redundancy Conveyed by Internal Articulator Dynamics in Synthetic Audiovisual Speech”, Proceedings of Interspeech, pp. 706-709, 2007.

[17] Massaro, D. W. and Light, J., “Read My Tongue Movements: Bimodal Learning To PerceiveAnd Produce Non-Native Speech /r/ and /l/”, Proceedings of INTERSPEECH, Geneva, pp. 2249-2252, 2003.