interspeech 2008 incorporating SST 08

Two- and Three-Dimensional Visual Articulatory Models for Pronunciation Training and for Treatment of Speech Disorders

Bernd J. Kröger[1], *Verena Graf-Borttscheller*[1] & *Anja Lowit*[2]

[1] Department of Phoniatrics, Pedaudiology and Communication Disorders, University Hospital Aachen and Aachen University, Germany

[2] Speech and Language Therapy Division, Department of Educational and Professional Studies, University of Strathclyde, Glasgow, UK

bkroeger@ukaachen.de, verenagraf76@aol.com, a.lowit@strath.ac.uk

Abstract

Background: Visual articulatory models can be used for visualizing vocal tract articulatory speech movements. This information may be helpful in pronunciation training or in therapy of speech disorders. **Method:** For testing this hypothesis, speech recognition rates were quantified for mute animations of vocalic and consonantal speech movements generated by a 2D and a 3D visual articulatory model. The visually based speech sound recognition test (mimicry test) was performed by two groups of eight children (five to eight years old) matched in age and sex. The children were asked to mimic the visually produced mute speech movement animations for different speech sounds. **Results:** Recognition rates stay significantly above chance but indicate no significant difference for each of the two models. **Conclusions:** Children older than 5 years are capable of interpreting vocal tract articulatory speech sound movements without any preparatory training in a speech adequate way. The complex 3D-display of vocal tract articulatory movements provides no significant advantage in comparison to the visually simpler 2D-midsagittal displays of vocal tract articulatory movements.

Index Terms: visual articulatory model, audio-visual speech synthesis, visual perception, development of visual perception

1. Introduction

This paper focuses on the *problem of visual complexity* or on the *need for visual simplicity* especially for the case of 2D- versus 3D-visual data perceived by children. It is known that the development of visual perception in children – i.e. the development of abilities for perceiving and recognizing colors, shapes, spatial relations, visual concepts, and for performing the figure-ground distinction, etc. – is a part of their overall cognitive development (Snowdon et al. 2006). Thus processing of visual information is different and not as elaborate for children as it is for adults.

The visual performance of children while recognizing speech sounds on the basis of visual data generated by two different visual articulatory models were tested in this study. Both visual models produce animations of vocal tract speech movements for single sounds, syllables, words, or for short utterances. The models can be used in pronunciation training or in treatment of speech disorders. The *two-dimensional visual articulatory model* generates two-dimensional midsagittal views of the vocal tract. The *three-dimensional visual articulatory model* generates three-dimensional views of the vocal tract organs and their movements by displaying the vocal tract walls in a transparent mode.

It is hypothesized that visual speech sound information generated by a three-dimensional visual articulatory model does not necessarily result in better visual speech sound recognition rates than speech sound information generated by a simpler two-dimensional visual articulatory model. This is because two-dimensional midsagittal views comprise all essential visual information of speech movements and at the same time are less complex and less difficult to process for children than the three-dimensional views.

2. The 2D-model

The two-dimensional visual articulatory model (Kröger 2003, see also Fig. 1and Fig. 2) is a geometrical model based on static and dynamic MRI-data of a speaker of Standard German with no known speech anomalies (Kröger et al. 2000 and 2004).

Figure 1: *Midsagittal view of the vocal tract organs generated by the two-dimensional visual articulatory model for central tongue position, lowered velum and abducted vocal folds (i.e. starting and final position for each animation).*

Nine functional parameters were extracted for controlling the movements of the model articulators (jaw, tongue, lips, velum, and glottis). Three parameters are used for controlling the whole *global* vocal tract shape with respect to vocalic articulation, i.e. high-low, front-back, unrounded-rounded. The remaining six parameters are used for controlling *local parts* of the vocal tract shape with respect to consonantal articulation, i.e. degree of lip closure, degree of tongue tip closure, degree of tongue body closure, velic aperture, glottal aperture, tongue tip position. In addition three different types of consonantal constrictions can be differentiated: (i) critical

Accepted after peer review of full paper

constriction type for the production of fricatives, (ii) central closure together with lateral opening for the production of laterals (lateral opening is visualized by a lateral tongue contour line superimposed on the midsagittal view) and (iii) vibrant constriction type for the production of vibrants.

Local consonantal closing and opening movements are superimposed on basic global vocalic articulatory movements. This concept of separating *global vocalic* and *local consonantal* articulation and superimposing consonantal articulatory closing-opening movements on basic vocalic articulatory movements is capable of describing a huge amount of consonant-vowel-coarticulation (Kröger 1998). For example the tongue body movement from [ʃ] to [ʊ] in [ʃpʊtnik] (Fig. 2, "Sputnik" in German pronunciation) occurs "behind" the consonantal closure during the consonantal closure time interval of [p]. In general, articulatory positions of subsequent sounds are already approached during the production time intervals of preceding sounds for all articulators (anticipatory coarticulation).

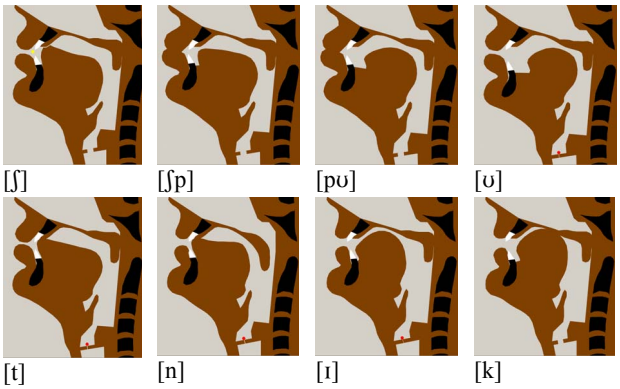

Figure 2: *Midsagittal views of the segments of the word [ʃpʊtnɪk] ("sputnik" in German pronunciation) generated by the two-dimensional visual articulatory model. The [p]-closure is displayed at its beginning and end. Red points indicate phonation, yellow points indicate frictional noise generation.*

3. The 3D-model

The three-dimensional visual articulatory model (Birkholz et al. 2006, Fig. 3 and Fig. 4) is a geometrical model based on static and dynamic MRI-data of the same speaker of Standard German (Birkholz and Kröger 2006). 23 geometrical parameters are used for controlling the positioning of all model articulators. A *dominance model* is implemented in order to define the degree of participation of each articulator in the realization of a vocalic or in the realization of a consonantal vocal tract movement (see the gestural control concept developed by Birkholz et al. 2006 and cf. Cohen and Massaro 1993). The resulting vowel-consonant-coarticulation of this model has been tested by fitting natural MRI-data (Birkholz and Kröger 2006). Glottal aperture, phonation, and frictional noise source location – which is visualized in the 2D-model – are not visualized in this 3D-model.

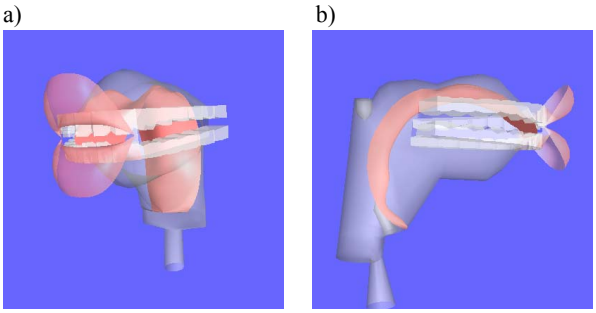

Figure 3: a) *Right-frontal and b) left-lateral 3D-views of the vocal tract organs generated by the three-dimensional visual articulatory model for central tongue position, lowered velum and abducted vocal folds (i.e. starting and final position for each animation).*

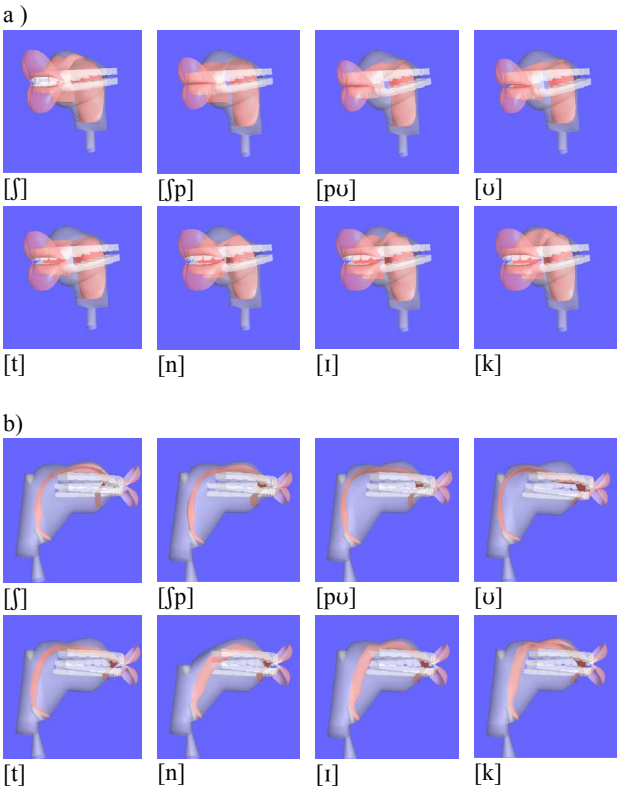

Figure 4: *a) Right-frontal and b) left-lateral 3D-views for each segment of the word [ʃpʊtnɪk] "sputnik" generated by the three-dimensional visual articulatory model. The [p]-closure is displayed at its beginning and end.*

4. The mimicry test

A mimicry test was designed for estimating the visual sound recognition rates for both models. 15 sounds, i.e. 4 vowels [i, a, u, y] and 11 consonants (3 plosives [p, t, k], 5 fricatives [f, s, ʃ, ç, x], and 3 nasals [m, n, ŋ]) were presented visually by

both models. The vowel gesture started from a neutral articulator position (Fig 1 and Fig. 3) and ended in the vowel target position. The consonants were embedded in [aCa]-context. 11 children (4m, 7f) from 4;6 to 8;3 years (mean 5;10 years) were tested using the 2D-model (Albert 2005) and 14 children (3m, 11f) from 5;3 to 10;7 years (mean 7;9 years) were tested using the 3D-model (Graf-Borttscheller 2007). None of these children had experience with the visual articulatiory models, i.e. none of the children used these models before in speech therapy or in articulation training. The children were native speakers of German but showed articulation disorders on single sounds but no severe deficits in language development.

45 mute video-stimuli were presented in random order comprising each of the 15 speech sounds three times. The video-stimuli are continuous movie-like sequences (not just a sequence of steady state still images) slowed down to the half of normal speech rate.

In the case of the 3D-model each animation stimulus consisted of 2 animation sequences, displaying the right-front and the left-lateral view of the vocal tract (cf. Fig. 3 and Fig. 4) consecutively. Both groups of children were asked to mimic the speech sounds visualized by the mute animations of the two models. The children's productions were phonetically transcribed by the examiners of the studies. In order to compare the results of the mimicry test across models a subgroup of children matched in age and sex was selected from the 2D-model the 3D-model group. Both subgroups included 8 children (3m, 8f) from 4;11 to 8;3 years (mean 6;5 years).

5. Results

Children's recognition rates for these mute purely visual stimuli were evaluated using two different procedures, i.e. phoneme and feature evaluation. For the *phoneme evaluation* the ratings of the correct sound recognition were calculated for both groups. 19% of all 3D-model stimuli were produced correctly while 22% of all 2D-model stimuli matched the stimulus. The difference in recognition rate between both models (2D and 3D) is not significant (p=0.96, two-tailed Wilcoxon sign rank test for matched samples). The children's recognition rate increased significantly with age for both models (2D and 3D).

Feature evaluation accounts for the fact that even in the case of a mistake in a *sound* recognition sample, a lot of sound *features* may be detected correctly. For example, if a [n] is produced by the 2D- or 3D-model and a [t] is recognized by the child, this child's answer is not completely wrong. The child has already recognized (i) the correct consonantal constriction forming articulator (i.e. tongue tip), (ii) the correct place of articulation (i.e. alveolar), and (iii) has recognized correctly that a consonantal vocal tract closure is produced for example in contrast to a vocalic opening of the vocal tract or in contrast to a consonantal critical constriction as it occurs in the case of fricatives.

Thus a set of *visual articulatory features* was defined as a basis for calculating the correct *feature* recognition rate. Six visual articulatory features were identifiable for the selected sounds (Tab. 1). However, as the *voice* feature was only visible in the 2D-model, it was excluded and only five visual articulatory features were used for the comparison of the feature recognition rates.

The results for *feature evaluation* show that 63% of all visual articulatory features were correctly recognized with the 3D-model, and 61% with the 2D-model. The difference in recognition rate was not significant (p=0.994, two-sided Wilcoxon sign rank test for matched samples). In contrast to the phoneme evaluation, the children's feature recognition rate did not increase significantly with age for either model. Furthermore the recognition scores of any individual visual articulatory feature did not differ significantly across the two models (Fig. 5). However, the various articulatory features showed significantly different recognition scores for both models (Table 2). In the case of the both models the visual articulatory features *nasality, articulator* and *place* do not exhibit significant different recognition scores (two-tailed Wilcoxon sign rank test) while the feature rounding exhibits the highest and the feature narrowness the lowest recognition scores. An ordering of articulatory visual features with respect to the recognition rates is given in Tab. 2 for both models.

Table 1. *System of visual articulatory features, the area of appearance of the feature within the animation, and the possible specifications for each feature.*

Area	**feature**	**specification**		
		(1)	**(2)**	**(3)**
lips	rounding	rounded	neural	spreaded
oral region	articulator	lips	tongue body	tongue tip
oral region	narrowness of obstruction	open	narrow	closed
tongue	place of obstruction	hard palate or alveolar ridge	soft palate	pharyngeal wall
velum	nasality	nasal	non-nasal	-
larynx	voice	voiced	voiceless	-

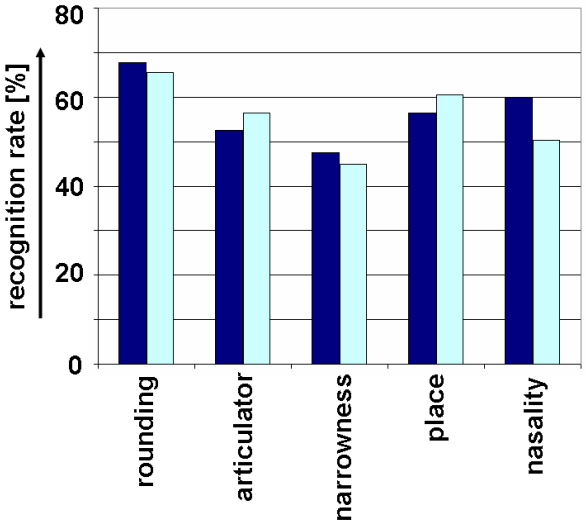

Figure 5: *Percentage of recognition scores for five visual articulatory features (cf. Tab. 1). Dark blue: recognition scores for 3D-model; light blue: recognition scores for 2D-model.*

Table 2. *Ordering of visual articulatory features with respect to the feature recognition rates for the 2D- and for the 3D-model (see figure 5). Significant different features are ordered in different boxes.*

3D-model	**2D-model**
rounding	rounding
nasality place articulator	place articulator nasality
narrowness	narrowness

6. Discussion

Despite the fact that children doing the mimicry test were not familiar with the 2D or 3D visual articulatory models, recognition rates occur significantly above the chance probability level. Ad hoc rates for whole visual sound recognition are around 20% while the probability for a right decision by chance is lower than 7%. Ad hoc rates for articulatory visual feature recognition are around 60% for both models while the probability for a right decision by chance is around 37%. This indicates that children of *5 and older* are capable of processing the visual information of speech movements produced by these models to some degree. These children are capable of recognizing speech sounds from mute visual speech sound movements significantly above chance level.

In addition, the current data have shown age effects in the phoneme recognition rate, although this was not replicated when feature recognition was evaluated. This might point to the fact that children are able to process the visual information from an early age on, but might need e.g. more cognitive maturity to put all the information together and arrive at the correct phoneme.

A detailed inspection of the visual articulatory feature recognition rates indicates that the lip feature *rounding* is recognized with highest accuracy for both models. The features *nasality*, *place of obstruction*, and *obstruction producing articulator* are recognized with medium accuracy and the visual articulatory feature *narrowness of obstruction* is recognized with lowest accuracy.

The high recognition scores for *lip rounding* are not surprising as this is the most visible feature during speech and thus the most familiar concept to untrained observers. However, the fact that the other features were recognized with rates above chance level provides clear evidence for the children's ability to reconstruct articulatory movements from visual models.

7. Conclusions

In conclusion this work supports the hypothesis that the information of vocal tract articulator movements displayed by visual articulatory 2D- or 3D-models is *intuitively perceived by persons older than 5 years in a speech adequate way* even if the visual information is as simple as in 2D midsagittal visual articulatory models. More complex 3D visual information does not lead to higher visual speech sound perception rates despite the fact, that stimuli were presented here successively in two views (left-lateral and right-frontal). Thus we suggest the use of very easily understandable 3D visualizations of articulation.

Our results are encouraging the clinical use of 2D and 3D models in therapy of speech disorders while the school use for foreign language training was not tested here. We have shown that the output of visual articulatory 2D- and 3D-models can be interpreted even by young children. Further evidence for the usefulness of such models is provided by our previous research indicating that the interpretation of the visual information improves with training (Kröger et al. 2005).

8. Acknowledgements

This work was supported in part by the German Research Council Grant Nr KR 1439/13-1 and by the EU-Project-Nr. 133920-LLP-ECSMSD.

9. References

[1] Snowden R, Thompson P, Troscianko T (2006) Basic Vision: An Introduction to Visual Perception. Oxford University Press, Oxford.

[2] Kröger BJ (2003) Ein visuelles Modell der Artikulation. Laryngo-Rhino-Otologie 82: 402-407

[3] Kröger BJ, Hoole P, Sader R, Geng C, Pompino-Marschall B, Neuschaefer-Rube C (2004) MRT-Sequenzen als Datenbasis eines visuellen Artikulationsmodells. HNO 52: 837-843

[4] Kröger BJ, Winkler R, Mooshammer C, Pompino-Marschall B. (2000) Estimation of vocal tract area function from magnetic resonance imaging: Preliminary results. Proceedings of 5th Seminar on Speech Production: Models and Data (Kloster Seeon, Bavaria) pp. 333-336 (see also http://www.speechtrainer.eu)

[5] Birkholz P, Jackel D, Kröger BJ (2006) Construction and control of a three-dimensional vocal tract model. Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2006) (Toulouse, France) pp. 873-876

[6] Birkholz P, Kröger BJ (2006) Vocal tract model adaptation using magnetic resonance imaging. Proceedings of the 7th International Seminar on Speech Production (Belo Horizonte, Brazil) pp. 493-500

[7] Cohen MM, Massaro DW (1993) Modeling coarticulation in synthetic visual speech. Models and techniques. In: Thalmann NM, Thalmann D. (eds.) Computer Animation. Springer, Tokyo, pp. 139–156

[8] Albert S (2005) Einsatz des SpeechTrainers in der Artikulationstherapie bei Kindern. Unpublished MA thesis, Aachen University, Aachen, Germany

[9] Graf-Borttscheller V (2007) Visuelles Erkennen von Lautbewegungen am dreidimensionalen Artikulationsmodell. Unpublished MA thesis, Aachen University, Aachen, Germany

[10] Kröger BJ (1998) Ein phonetisches Modell der Sprachproduktion. Niemeyer-Verlag, Tübingen

[11] Kröger BJ, Gotto J, Albert S, Neuschaefer-Rube C (2005) A visual articulatory model and its application to therapy of speech disorders: a pilot study. In: S Fuchs, P Perrier, B Pompino-Marschall (eds.) Speech production and perception: Experimental analyses and models. ZASPiL 40: 79-94