Audiovisual Tools for Phonetic and Articulatory Visualization in Computer-Aided Pronunciation Training

Bernd J. Kröger[1], Peter Birkholz[1], Rüdiger Hoffmann[2], and Helen Meng[3]

[1] Department of Phoniatrics, Pedaudiology, and Communication Disorders,
University Hospital Aachen and RWTH Aachen University, Aachen, Germany
bkroeger@ukaachen.de, pbirkholz@ukaachen.de
[2] Department of Acoustics and Speech Communication,
Dresden University of Technology, Dresden, Germany
ruediger.hoffmann@ias.et.tu-dresden.de
[3] Department of Systems Engineering and Engineering Management,
The Chinese University of Hong Kong (CUHK), Shatin, NT, Hong Kong SAR of China
hmmeng@se.cuhk.edu.hk

Abstract. This paper reviews interactive methods for improving the phonetic competence of subjects in the case of second language learning as well as in the case of speech therapy for subjects suffering from hearing-impairments or articulation disorders. As an example our audiovisual feedback software "Speech-Trainer" for improving the pronunciation quality of Standard German by visually highlighting acoustics-related and articulation-related sound features will be introduced here. Results from literature on training methods as well as the results concerning our own software indicate that audiovisual tools for phonetic and articulatory visualization are beneficial for computer-aided pronunciation training environments.

Keywords: Audiovisual speech tools, pronunciation training, second language learning, speech therapy.

1 Introduction

Second language learners often show severe problems in learning phonetic segmental features of speech sounds as well as prosodic features of target languages since learning is influenced by the phonetic segmental and prosodic knowledge concerning their native language [1, 2]. Since the vocabulary and the grammar of a foreign language can be acquired by using mainly cognitive learning strategies, i.e. processing abstract and discrete linguistic items by grammatical rules, the learner is easily aware of problems at the *linguistic level*. But this is not necessarily the case at the *phonetic level*. Moreover second language learners try to adapt the phonetic segmental and prosodic systems of their first language or mother tongue (L1) to the trained second language (L2). The resulting L2 foreign accent is easily detectable for speakers which are familiar with the target language (e.g. native speakers of the target language), while second language learners often are not even aware of their phonetic faults with respect to L2 [1, 2].

A. Esposito et al. (Eds.): COST 2102 Int. Training School 2009, LNCS 5967, pp. 337–345, 2010.

Phonetic learning is also a problem for hearing-impaired children in the case of first language acquisition, if their hearing-impairment is innate or if it occurred in the first months after birth [3]. Since the auditory perception as well as the auditory feedback loop of these children is impaired, they are not able to learn the phonetic features of sounds or of the prosodic patterns of L1 as easy as normal hearing children. Speech produced by these children may be intelligible but often sounds strange. However, the linguistic capabilities of these children are often completely unaffected, since these children are capable of learning lexical items and grammatical rules, for example, of sign languages [4]. Phonetic problems also occur for acquired hearing impairments, e.g. in adults or the elderly due to a defective auditory feedback control loop. These people may lose control concerning the exact pronunciation of some phonetically difficult sounds of their mother tongue. Problems on the phonetic level of language acquisition also occur for children suffering from articulation or speech sound disorders. Here a common phenomenon is that these children shift posterior plosives (e.g. /g/ and /k/ are realized as /d/ and /t/), or they show a misarticulation of fricatives [5, 6, 7]. Other complex articulatory and coarticulatory problems occur for people suffering from motor speech disorders like apraxia of speech or dysarthria [8].

Thus in both cases (second language learning and phonetic oriented speech therapy) learners need intensive *phonetic treatment* with regard to the target language. Since phonetic treatment requires a lot of interaction between teacher and learner in order (i) to detect phonetic errors of learners (ii) to make the learner aware of his phonetic problems with respect to the target language and (iii) to advice the learner to do corrections of his articulations in direction towards the sound system and towards the prosodic system of the target language, it is not trivial to design phonetic treatment in terms of human-computer interaction.

This paper focuses on the problem of *computer-aided pronunciation training (CAPT*) and addresses three important research issues: (i) Can phonetic errors be detected by machine, for example by using speech recognition algorithms? (ii) Can learners become aware of their phonetic errors by using a human-machine interface exclusively? (iii) Is it possible to develop computer-aided self-learning environments for advising the learner an efficient way to overcome phonetic problems with respect to the target language? It will be argued in this paper that *audiovisual tools* as part of computer-aided self-learning environments are effective in enhancing the phonetic competence of learners with respect to a target language.

2 Auditory Tools

Acoustic speech signals of communication partners as well as the auditory feedback signals of learners provide the most important cues for improving the phonetic competence concerning a target language. In L1 acquisition infants process acoustic speech signals comprising words, phrases, or complete sentences produced by communication partners (e.g. caregivers). They also process the acoustic self-productions of their words and sentences and compare the auditory percept of their self-productions with the temporarily or permanently stored auditory percepts already learned [9]. In addition, visual information of lip and lower jaw movements, which occur during speech, may be helpful, but even if this visual information is not available, as is the case for blind children, speech acquisition proceeds as fast as for normal children [10]. But if children are suffering from severe hearing impairments, speech acquisition can strongly be delayed [11]. Thus many computer-aided language learning environments use acoustic tools that allow learners to listen to words and/or sentences, produced by native speakers of the target language [12]. The next step is to allow speech recordings within the computer-aided language learning environment, in order to enable learners to become aware of their own word and sentence productions [13].

If the improvement of the phonetic competence and thus the elimination of a foreign accent is the primary goal, more specific auditory and acoustic tools should be integrated into the computer-aided language or pronunciation learning environment. In this case tools for pronunciation training should be capable of detecting phonetic errors of learners automatically. Normal speech recognition software does not cater especially for detecting phonetic pronunciation errors, for example in the case of L2 learning or in the case of specific speech errors occurring in different types of speech disorders. For detecting these kinds of pronunciation errors it is necessary to access comprehensive speech data bases comprising knowledge concerning typical errors resulting from a specific L2-L1 interference. Consequently it is necessary to implement knowledge, how L2 learners with a specific L1 background produce sounds, words, and prosodic features in the L2 target language [14, 15, 16, 17].

3 Visual and Audiovisual Tools

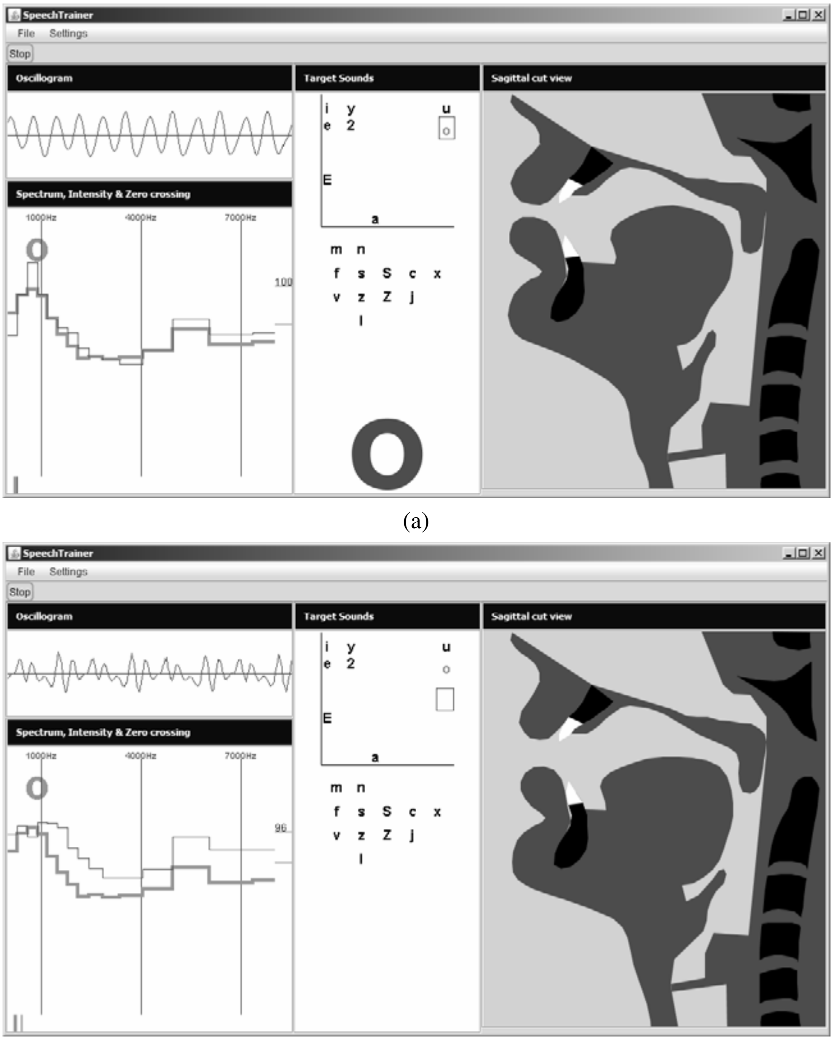
Acoustics-related visual aids display relevant auditory features of the acoustic speech signal. This can be the oscillogram, the spectrogram, the F0- and intensity-contour of a speech item (i.e. syllable, word, or sentence), or the spectrum, F0-, and intensity value of a specific point in time within a speech item. Oscillogram, spectrogram, F0-, and intensity-contour of a whole speech item can be displayed synchronously for self-productions of learners and for pre-recorded correct sample productions done by native speakers of the target language. The visual comparison of both productions allows learners to detect mispronunciations. In the case of prosodic features, i.e. intonation and stress pattern, as well as sound duration (e.g. vowel or consonant productions that are too long or too short), oscillogram, F0-, and intensity contour are helpful not only to detect pronunciation errors but in addition to indicate the direction how the learner can produce the speech item in a more correct way. The comparison of oscillograms indicates how sound durations should be changed. The comparison of intensity-contours indicates how stress patterns should be changed and the comparison of F0-contours indicates how intonation patterns should be changed [18]. The □ompareison of formant trajectories (given in spectrograms) indicates how vocalic and consonantal articulation targets should be changed. In the case of fricatives, the comparison of the spectral energy distribution indicates how fricative production can be changed [19]. The display of these acoustics-related parameters can be *stylized* (e.g. in CoKo for spectra and spectrograms, see [20], in SpeechViewer for intonation contours, see [21]) in order to enable learners to process this visual information easily and in order not to discourage learners without any technical or scientific background to use this training software.

Articulation-related visual aids can be used for training a correct articulation of static targets for vowels and consonants as well as a correct coarticulation within syllables and words. These aids display vocal tract organs or articulators (e.g. lips, tongue, velum, larynx with glottis) and their functioning in speech production, i.e. the positioning and movement of articulators in sound, word, or sentence production. Articulation-related visual aids are already used in some pronunciation training environments. Badin et al. [22, 23] and Bailly et al. [24] developed a realistic 3D virtual talking head comprising a 3D-model of speech articulators (i.e. lips, lower jaw, tongue, velum) on the basis of a comprehensive MRI, CT, EMA, and video database of several speakers. This model is capable of displaying the complete face or a cut-away view of the head including lips, tongue, and velum (Grenoble talking head). Badin et al. [23] were able to show that humans have the capability of "tongue reading" (i.e. interpreting the normally non-visible parts of the tongue) especially in the case of a strongly degraded or in the case of an absent acoustic signal. Engwall et al. [25, 26] developed a data based 3D-virtual tutor (Stockholm talking head, called "Artur"). The Stockholm talking head is embedded in a complex pronunciation training environment capable of detecting mispronunciations in the visual and acoustic domain and capable of giving a set of instructions how to improve the articulation with respect to the target language (Swedish). Kröger et al. [27] developed a 2-D virtual model for the sagittal view of a talker of Standard German (Aachen model, called "Bernie") and developed a 3D-virtual model of the speech articulators [28]. Both models were tested in therapy of developmental speech disorders [29. 30] indicating that even preschoolers can acquire the ability to understand (i.e. to "read") lip, tongue and velum speech movements. In accordance with Badin et al. [23] the results of Kröger et al. [30] indicate that lip reading is predominant and easier than tongue and velum reading. The virtual talking head of Massaro et al. [31, 32, 33, 34, 35] called "Baldi", has been applied already for many languages (i.e. multilingual talking head [33]) and "Baldi" has been used as a tutor in computer-aided environments for spoken and written language training as well as for articulation training.

4 A Preliminary Articulation-Related Audiovisual Tool for Improving Sound Quality in Standard German

As a prototype for an articulation-related audiovisual tool, the 2D-articulatory model "Bernie" (Fig. 1a and Fig. 1b, right side) has been integrated in our pronunciation training environment "SpeechTrainer" for improving the quality of speech sounds in the case of therapy of articulation disorders ([s]-misarticulation) as well as in the case of second language learning with Standard German as L2.

With respect to learners pronunciation problems a target sound can be chosen from the list of German long vowels, nasals, fricatives or laterals (see Fig. 1a and Fig. 1b, middle). The appropriate reference spectral envelope and reference zero crossing rate is displayed (Fig. 1a and Fig. 1b, left side, bold light curve and bold light vertical line below; here for /a:/). Then the learner tries to adapt the spectral envelope and the zero crossing rate of his sound production (Fig. 1a and Fig. 1b, left side, dark curve and dark vertical line below) towards the visually displayed target spectral envelope and



(a)

(b)

Fig. 1. Two screenshots of the sound training tool "SpeechTrainer" for non-correct (Fig. 1a) and for correct production (Fig. 1b) of Standard German long vowel [o:]. Left side in (a) and (b): Oscillogram of current sound produced by the learner (on top); 14 step bark scaled spectral envelope from 0 to 8 kHz (curves below oscillogram); amplitude including db-value (right side adjacent to spectral envelope); zero crossing rate (vertical lines below spectral envelope on the bottom) for current sound produced by the learner (black thin curve or line) and for the selected target sound (bold grey curve or line). Middle in (a) and (b): Vowel space (on top) including Standard German long vowels; consonant list (below). The selected target sound is marked by gray color. The gray rectangle indicates the sound quality currently produced by the learner. The "reward sound symbol" for correct sound production is displayed at the bottom (only in b). Right side in (a) and (b): Midsagittal view corresponding to the current sound quality produced by the learner.

zero crossing rate. If the sound quality is sufficiently reached, the learner will be rewarded by the display of the target sound symbol as huge symbol (letter) below the target sound list (Fig. 1b middle, "reward symbol").

Five realizations of each potential target sound (vowels [iː, eː, ɛː, aː, oː, uː, yː, øː] and consonants [f, s, ʃ, ç, x, v, z, ʒ, j, l]) were recorded by a native Speaker of Standard German (phonetic expert). In addition IPA sounds filling the gaps in the vowel space (i.e. [ɔː, œː, əː]) and in addition typical mispronunciations of consonants (in this preliminary version only for [s], i.e. interdental and addental [s]-realizations) were recorded by the same speaker and stored as "gap sounds". The Mahalanobis distance of the 14 bark-scaled spectral envelope values plus zero crossing rate between current acoustic input on the one hand and all target and gap sounds on the other hand is calculated permanently in 50 ms intervals. The articulatory configuration of the acoustically nearest target or gap sound is displayed (Fig. 1a and Fig. 1b, right side) and the appropriate target or gap sound is marked by a rectangle in the sound table (Fig. 1a and Fig 1b, middle). If in addition this Mahalanobis distance undergoes a specific threshold value for the currently selected target sound (i.e. 30% of the distance between the currently selected target sound and all other target and gap sounds), the currently selected target sound symbol is displayed in an extra region below the sound list (Fig. 1b, middle. "reward symbol"). In this case the learner is rewarded for his correct target sound production.

The tool was evaluated in therapy of speech disorders for children suffering from a specific articulation disorder, i.e. [s]-misarticulation ("sigmatism", a sort of lisping). All reference target and gap sounds were recorded by a female speaker of Standard German, 22 years old (master student of speech therapy). Two tests were performed. (i) The goal of the *sensitivity test* was to evaluate whether the tool is capable of discriminating different [s]-misarticulations on the acoustic level. 11 children participated in this test (6 male, 5 female; age: 5,2 to 6,9). Logopaedic diagnostics resulted in identifying 4 children with addental [s]-realisation, 4 children with interdental [s]-realization and 3 children with normal [s]-production for this group. All children were asked to perform 40 [s]-productions each (40 trials) without looking at the computer monitor. The test conductor (i.e. speech therapist) noted down whether the target sound was reached by the reward display or not for each trial. The resulting rate of initially reaching the target [s] in the "SpeechTrainier"-tool was 0% in the case of children suffering from an addental [s]-production, 2% in the case of children suffering from an interdental [s]-production, and 58% in the case of children with no [s]-production dysfunctions ($p<0.001$). Thus it can be concluded that "SpeechTrainer" is capable of differentiating normal and disordered [s]-production. (ii) The goal of the *learning test* was to evaluate whether the patient is capable of correcting his [s]-articulation by using the "SpeechTrainer"-tool. One subject (male, 5,9 years old) suffering from a [s]-misarticulation problem ([s]-interdentalis) used the "SpeechTrainer"-tool for 5 minutes feedback computer training (FCT) within each of 10 therapy sessions over a time period of 10 weeks. 12 [s]-realizations of this subject produced before and after each FCT were checked as described in the sensitivity test. A significant increase of correct [s]-realizations (before to after FCT) occurred for each FCT (towards 42% at the beginning and towards 67% at the end of the 10 week therapy period, $p<0.01$). Moreover a baseline increase of correct [s]-realizations before FCT over the 10 weeks of speech therapy from 0% to 33% occurred as well ($p<0.01$).

Furthermore the tool was evaluated for second language learning (two Czech L2 learners of Standard German, male, 24 and 26 years old). Here, all reference target and gap sounds were recorded by a male speaker of Standard German, 50 years old (phonetician). The *learning test* was performed by both subjects in order to evaluate whether subjects are capable of correcting their L2-[oː]-articulation towards a more closed vowel quality by using the “SpeechTrainer”-tool. Both subjects used the “SpeechTrainer”-tool for 5 minutes FCT in 5 learning sessions over a time period of 2 weeks. The [oː]-realizations were evaluated before and after each FCT within each learning sessions by performing 12 [oː]-realizations. A significant increase of correct L2-[oː]-realizations (before to after FCT) occurred within each learning session from 27% to 48% (speaker 1, $p<0.01$) and from 35% to 56% (speaker 2, $p<0.05$) while a significant increase of correct produced realizations before FCT was found over the two weeks of training (i.e. over these 5 sessions) only for speaker 1 (from 22% to 32%, $p<0.05$).

5 Discussion

This paper stresses the need of audiovisual tools in computer-aided pronunciation training environments. Beside typical acoustics-related audiovisual tools such as those for comparing oscillograms, sonograms, F0-contours, and intensity-contours between teacher and learner pronunciations of words or sentences, articulation-related audiovisual tools are introduced. An articulation-related audiovisual tool for isolated sound pronunciation developed in the Aachen Lab has been evaluated. Results indicate that this tool can be used successful in therapy of [s]-articulation disorders as well as in second language learning for strengthening vowel quality. Instantaneous and continuous learning benefits (i.e. learning benefits within one training session and over the whole time period of training) were verified for the subject suffering from [s]-articulation disorder and for one of the two second language learners. It was not possible to evaluate to what degree the visualization of the articulatory information (i.e. mid-sagittal view of speech sounds) contributes to these learning outcomes. But all learners reported that they used the visual articulatory information as an additional cue beside the spectral envelope matching in order to correct their articulation.

Acknowledgments. This work was supported in part by the German Research Council DFG grant Kr 1439/13-1 and grant Kr 1439/15-1.

References

[1] Flege, J.E.: Phonetic approximation in second language acquisition. Language Learning 30, 117–134 (1980)

[2] Munro, M.J., Derwing, T.M.: Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. Language Learning 49, 285–310 (1999)

[3] Geers, A.E., Moog, J.S.: Predicting spoken language acquisition of profundly hearing impaired children. Journal of Speech and Hearing Disorders 52, 84–94 (1987)

[4] Strong, M. (ed.): Language Learning and Deafness. Cambridge University Press, Cambridge (1988)

[5] Gibbon, F.E.: Undifferentiated lingual gestures in children with articulation/phonological disorders. Journal of Speech, Language, and Hearing Research 42, 382–397 (1999)

[6] Rvachew, S., Jamieson, D.G.: Perception of voiceless fricatives by children with a functional articulation disorder. Journal of Speech and Hearing Disorders 54, 193–208 (1989)

[7] Rvachew, S., Grawburg, M.: Correlates of phonological awareness in preschoolers with speech sound disorders. Journal of Speech, Language, and Hearing Research 49, 74–87 (2006)

[8] Kent, R.D.: Research on speech motor control and its disorders: A review and prospective. Journal of Communication Disorders 33, 391–428 (2000)

[9] Kröger, B.J., Kannampuzha, J., Neuschaefer-Rube, C.: Towards a neurocomputational model of speech production and perception. Speech Communication 51, 793–809 (2009)

[10] Perez-Pereira, M., Castro, J.: Language acquisition and the compensation of visual deficit: New comparative data on a controversial topic. British Journal of Developmental Psychology 15, 439–459 (1997)

[11] Yoshinaga-Itano, C., Sedey, A.: Early Speech Development in Children Who Are Deaf or Hard of Hearing: Interrelationships with Language and Hearing. Volta Review 100, 181–211 (1999)

[12] Accent School (2008), `http://www.accentschool.com/`

[13] Pronunciation Power (2006), `http://www.englishlearning.com/`

[14] Jokisch, O., Koloska, U., Hirschfeld, D., Hoffmann, R.: Pronunciation learning and foreign accent reduction by an audiovisual feedback system. In: Tao, J., Tan, T., Picard, R.W. (eds.) ACII 2005. LNCS, vol. 3784, pp. 419–425. Springer, Heidelberg (2005)

[15] Harrison, A.M., Lau, W.Y., Meng, H., Wang, L.: Improving mispronunciation detection and diagnosis of learners' speech with context-sensitive phonological rules based on language transfer. In: Proceedings of Interspeech, Brisbane, Australia, pp. 2787–2790 (2008)

[16] Meng, H., Lo, Y., Wang, L., Lau, W.: Deriving salient learners' mispronunciations form cross-language phonological comparisons. In: Proceedings of the IEEE Workshop in Automatic Speech Recognition and Understanding, ASRU, Kyoto, Japan, pp. 437–442 (2007)

[17] Wang, L.X., Feng, X., Meng, H.: Mispronunciation detection based on cross-language phonological comparisons. In: Proceedings of the IEEE IET International Conference on Audio, Language and Image Processing, Shanghai, China, pp. 307–311 (2008)

[18] Better Accent Tutor (2009), `http://www.betteraccent.com/`

[19] Vicsi, K., Csatari, F., Bakcsi, Z.s., Tantos, A.: Distance score evaluation of the visualised speech spectra at audio-visual articulation training. In: Proceedings of EUROSPEECH 1999, Budapest, Hungary, pp. 1911–1914 (1999)

[20] Vicsi, K., Hacki, T.: CoKo - Computergestützter Sprechkorrektor mit audiovisueller Selbstkontrolle für artikulationsgestörte und hörbehinderte Kinder. Sprache-Stimme-Gehör 20, 141–149 (1996)

[21] Öster, A.M.: Teaching speech skills to deaf children by computer-based speech training. STL-Quarterly Progress and Status Report 36(4), 67–75 (1995)

[22] Badin, P., Bailly, G., Boë, L.J.: Towards the Use of a Virtual Talking Head and of Speech Mapping tools for pronunciation training. In: Proceedings of the ESCA Tutorial and Research Workshop on Speech Technology in Language Learning (STiLL 1998), pp. 167–170 (1998)

[23] Badin, P., Tarabalka, Y., Elisei, F., Bailly, G.: Can you "read tongue movements"? In: Proceedings of Interspeech 2008, Brisbane, Queensland, Australia, pp. 2635–2638 (2008)

[24] Bailly, G., Bérar, M., Elisei, F., Odisio, M.: Audiovisual speech synthesis. International Journal of Speech Technology 6, 331–346 (2003)

[25] Engwall, O., Bälter, O., Öster, A.M., Kjellström, H.: Designing the user interface of the computer-based speech training system ARTUR based on early user tests. Journal of Behaviour and Information Technology 25, 353–365 (2006)

[26] Engwall, O., Bälter, O.: Pronunciation feedback from real and virtual language teachers. Journal of Computer Assisted Language Learning 20, 235–262 (2007)

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

[28] Kröger, B.J., Birkholz, P.: A gesture-based concept for speech movement control in articulatory speech synthesis. In: Esposito, A., Faundez-Zanuy, M., Keller, E., Marinaro, M. (eds.) COST Action 2102. LNCS (LNAI), vol. 4775, pp. 174–189. Springer, Heidelberg (2007)

[29] Kröger, B.J., Gotto, J., Albert, S., Neuschaefer-Rube, C.: A visual articulatory model and its application to therapy of speech disorders: a pilot study. In: Fuchs, S., Perrier, P., Pompino-Marschall, B. (Hrsg.) Speech production and perception: Experimental analyses and models. ZAS Papers in Linguistics, vol. 40, pp. 79–94 (2005)

[30] Kröger, B.J., Graf-Bortscheller, V., Lowit, A.: Two- and three-dimensional visual articulatory models for pronunciation training and for treatment of speech disorders. In: Proceedings of Interspeech 2008, Brisbane, Queensland, Australia, pp. 2639–2642 (2008)

[31] Massaro, D.W.: Perceiving Talking Faces: From Speech Perception to a Behavioral Principle. MIT Press, Cambridge (1998)

[32] Massaro, D.W.: A computer-animated tutor for spoken and written language learning. In: Proceedings of the 5th International Conference on Multimodal Interfaces, Vancouver, British Columbia, Canada, pp. 172–175 (2003)

[33] Massaro, D.W.: The psychology and technology of talking heads: Applications in language learning. In: van Kuppevelt, J.C.J., Dybkjær, L., Bernsen, N.O. (eds.) Advances in Natural Multimodal Dialogue Systems, vol. 30, pp. 183–214. Springer, Heidelberg (2005)

[34] Massaro, D.W., Liu, Y., Chen, T.H., Perfetti, C.: A multilingual embodied conversational agent for tutoring speech and language learning. In: Proceedings of Interspeech 2006, Pittsburgh, PA, USA, pp. 825–828 (2006)

[35] Massaro, D.W., Bigler, S., Chen, T., Perlman, M., Ouni, S.: Pronunciation training: the role of eye and ear. In: Proceedings of Interspeech 2008, Brisbane, Queensland, Australia, pp. 2623–2626 (2008)