5

Augmented Reality Talking Heads as a Support for Speech Perception and Production

Olov Engwall
Centre for Speech Technology, School of Computer Science and Communication, KTH (Royal Institute of Technology), Stockholm
Sweden

1. Introduction

Visual face gestures, such as lip, head and eyebrow movements, are important in all human speech communication as a support to the acoustic signal. This is true even if the speaker's face is computer-animated. The visual information about the phonemes, i.e. speech sounds, results in better speech perception (Benoît et al., 1994; Massaro, 1998) and the benefit is all the greater if the acoustic signal is degraded by noise (Benoît & LeGoff, 1998; Sumby & Pollack, 1954) or a hearing-impairment (Agelfors et al., 1998; Summerfield, 1979).

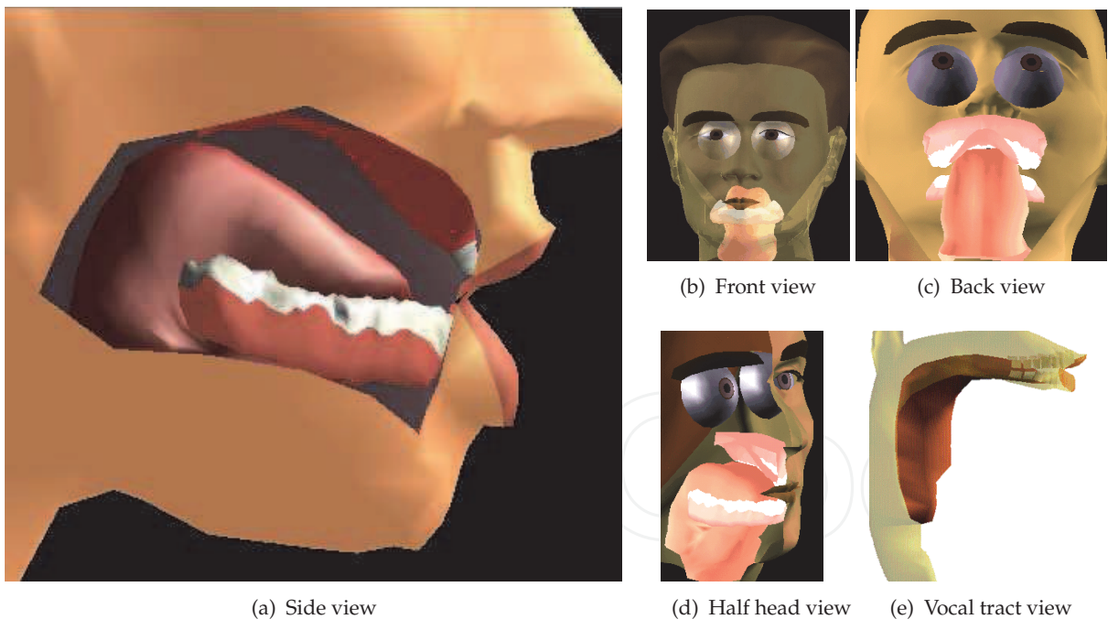
Many phonemes are however impossible to identify by only seeing the speaker's face, because they are visually identical to other phonemes. Examples are sounds that only differ in voicing, such as [b] *vs.* [p], or sounds for which the difference in the articulation is too far back in the mouth to be seen from the outside, such as [k] *vs.* [ŋ] or [h]. A good speech reader can determine to which viseme, i.e. which group of visually identical phonemes, a speech sound belongs to, but must guess within this group. A growing community of hearing-impaired persons with residual hearing therefore relies on cued speech (Cornett & Daisey, 1992) to identify the phoneme within each viseme group. With cued speech, the speaker conveys additional phonetic information with hand sign gestures. The hand sign gestures are however arbitrary and must be learned by both the speaker and the listener. Cued speech can furthermore only be used when the speaker and listener see each other.

An alternative to cued speech would therefore be that the differences between the phonemes are directly visible in an augmented reality display of the speaker's face. The basic idea is the following: Speech recognition is performed on the speaker's utterances, resulting in a continuous transcription of phonemes. These phonemes are used in real time as input to a computer-animated talking head, to generate an animation in which the talking head produces the same articulatory movements as the speaker just did. By delaying the acoustic signal from the speaker slightly (about 200 ms), the original speech can be presented together with the computer animation, thus giving the listener the possibility to use audiovisual information for the speech perception. An automatic lip reading support of this type already exists, in the SYNFACE extension (Beskow et al., 2004) to the internet telephony application Skype. Using the same technology, but adding augmented reality, the speech perception support can be extended to display not only facial movements, but face and tongue movements together, in displays similar to the ones shown in Fig. 1. This type of speech

perception support is less vulnerable to automatic speech recognition errors and is therefore preferred over displaying the recognized text string.

Similarly, second language learners and children with speech disorders may have difficulties understanding how a particular sound is articulated or what the difference compared to another phoneme is. Both these groups may be helped by an augmented reality display showing and describing tongue positions and movements. The AR talking head display allows a human or virtual teacher to instruct the learner on how to change the articulation in order to reach the correct pronunciation.

For both types of applications, augmented reality is created by removing parts of the facial skin or making it transparent, in order to provide additional information on how the speech sounds are produced. In this chapter, we are solely dealing with computer-animated talking heads, rather than the face of a real speaker, but we nevertheless consider this as a good example of augmented reality, rather than virtual reality, for two reasons: Firstly, the displayed articulatory movements are, to the largest extent possible, real speech movements, and hence relate to the actual reality, rather than to a virtual, and possibly different, one. Secondly, the listener's perception of reality (the sounds produced) is enhanced using an augmented display showing another layer of speech production. In addition, many of the findings and discussions presented in this chapter would also be also relevant if the augmented reality information about tongue movements was displayed on a real speaker's cheek.



(a) Side view (b) Front view (c) Back view (d) Half head view (e) Vocal tract view

Fig. 1. Illustration of different alternatives to create the augmented reality display. (a) Skin made transparent in order to show the movements of the articulators. Display used for the experiments described in Section 3.2. (b) Front view with transparent skin, similar to one option in Massaro & Light (2003). (c) Viewer position inside the talking head, similar to one display in Massaro & Light (2003). (d) Front half of the head removed, similar to the display in Badin et al. (2008). (e) Displaying the vocal tract only, similar to the display in Kröger et al. (2008).

Since we are normally unaccustomed to seeing the movements of the tongue, the use of such a display leads to several research questions. AR talking head displays have therefore been created by several research teams, in order to investigate their usefulness as a support for speech perception or for speech production practice. This chapter will first introduce and discuss the different types of augmented reality displays used (Section 2) and then present a set of studies on speech perception supported by AR talking heads (Section 3). The use for speech production practice is more briefly discussed in Section 4, before ending with a general outlook on further questions related to the use of AR talking heads in Section 5.

2. Augmented reality talking heads

Augmented reality displays of the face have been tested for both speech perception (Badin et al., 2008; Engwall, 2010; Engwall & Wik, 2009a; Grauwinkel et al., 2007; Kröger et al., 2008; Wik & Engwall, 2008) and speech production (Engwall, 2008; Engwall & Bälter, 2007; Engwall et al., 2006; Fagel & Madany, 2008; Massaro et al., 2008; Massaro & Light, 2003; 2004). These studies have used different displays to visualize the intraoral articulation, as examplified in Fig. 1 and summarized in Table 1. The list excludes the epiglottis and the larynx, which are only shown in the studies by Badin et al. (2008) and Kröger et al. (2008).

As is evident from Table 1 and Fig. 1, there are several different choices for the presentation of the AR display. It is beyond the scope of this chapter to try to determine if any set-up is superior to others, but it may nevertheless be interesting to compare the different alternatives, as it is not evident what articulators to display and how. In addition to the tongue, all studies show the jaw in some form, since it is needed as a reference frame to interpret tongue movements and since it in itself gives important information for speech reading (Guiard-Marigny et al., 1995). One could argue that all other articulators that are relevant to speech production should be displayed as well, in order to give the viewer all the available information. However, most viewers have a diffuse and superficial understanding of the intraoral anatomy and articulatory movements and may hence be confused or frightened off by too much detail in the display. Some articulators may also hide others if a full three-dimensional representation is used.

Displays that show the entire 3D palate, either fully visible or semi-transparent, may encounter problems in conveying sufficiently clear information about if and where the tongue touches the palate, which is vital in speech production. To overcome such problems, Cohen et al. (1998) proposed to supplement the face view with a separate display that shows the regions were the tongue is in contact with the palate. This additional display would however split the learner's visual attention and it has not been used in the subsequent studies by Massaro et al. The alternative opted for by Badin et al. (2008), Engwall (2010); Engwall & Wik (2009a); Wik & Engwall (2008) and Massaro et al. (2008) is to concentrate on the midsagittal outline of the palate to facilitate the perception of the distance between the tongue and the palate. Engwall (2010); Engwall & Wik (2009a); Wik & Engwall (2008) simplified the display further by showing the tongue and jaw moving inside a dark oral cavity, with the limit of the transparent skin region corresponding to the palate outline (Fig. 1(a)). This choice was made since the children who were shown a line tracing of the palate in Engwall et al. (2006) found it difficult to interpret (they e.g., speculated that it was a small tube where the air passes in the nose). Kröger et al. (2008) on the other hand presented the vocal tract movements without the surrounding face, as in (Fig. 1(e)), and avoided occluding articulators this way.

The velum has been included in some of the studies in Table 1, but the usefulness of displaying it can be discussed. Seeing tongue movements is strange for many viewers, but they are at least conscious of the appearance and proprioceptic responses of the tongue surface, whereas it is much more difficult to internally visualize the placement and movement of the velum.

	B=Badin et al. (2008) E={ E1=Wik & Engwall (2008), E2=Engwall & Wik (2009a), E3=Engwall (2010) } F=Fagel & Madany (2008), K=3D model in Kröger et al. (2008) M=Massaro et al. (2008), ML=Massaro & Light (2003), Massaro & Light (2004)
View	Side view (B, E, F, K) (Fig. 1(a)), with a small angle (M) (Fig. 1(d)) Front view (K, ML) (Fig. 1(b)) Back view (ML) (Fig. 1(c))
Face	Video-realistic. Closer half removed, remoter half a black silhouette (B) Synthetic-looking. Closer half removed (M), semi-transparent skin (F, ML), transparent skin at the oral cavity (E) No face (K) (Fig. 1(e))
Lips	3D and video-realistic (B) 3D and synthetic-looking (F, K, L) 3D for the remoter part of the face (E, M)
Tongue	Midsagittal shape (in red, B; or turquoise, M) and the remoter half (B, M) Upper tongue surface (K) 3D body (E, F, ML)
Jaw & teeth	Midsagittal shape of the incisor (in blue, B; or green, M) and the remoter half of the lower and upper jaw (B, M) Semi-transparent schematic teeth blocks or quadrangles (F, K) Semi-transparent and realistic in 3D (ML) Visible and realistic 3D jaw, lower teeth and upper incisor (E)
Palate	Midsagittal shape (in yellow, B; or green, M) and the remoter half (B, M) Uncoloured semi-transparent tube walls (K) Semi-transparent schematic (F) or realistic (ML) Upper limit of transparent part of the skin corresponds to the midsagittal contour of the palate (E, Fig. 1(a))
Velum	Midsagittal shape (B) and the remoter part (M) Part of the semi-transparent tube walls (K) As part of the palate surface (F)
Pharynx walls	Realistic remoter half (B) Non-realistic surface at the upper part of the pharynx (F, M) Semi-transparent tube wall (K) Limit of transparent part corresponds to upper pharynx walls (E)
Movements	Resynthesis of one speaker's actual movements measured with EMA (B,E2) Rule-based, but coarticulation adapted to measurements (E1, E3, K) Rule-based with coarticulation models from facial animation (F, M, ML)

Table 1. Alternative representations of the articulators in the augmented reality display.

Another simplification, used in several studies, is to present the intra-oral articulations from a side view that makes the display similar to traditional two-dimensional tracings in phonetics, even if the model is in 3D. The side-view is the one that makes different articulations most distinct (which is why this display is used in phonetics), but one may well argue that different viewers may prefer different set-ups. Massaro & Light (2003) in addition used a front (as in Fig. 1(b)) and a back (as in Fig. 1(c)) view of the head, but without attempting to investigate if any view was better than the other. As an alternative, one could choose an interactive display, in which the user can rotate the 3D structure to different view points, but there is a risk that the structure complexity in other views may hide important articulatory features. To the best of our knowledge, the side view is hence the best alternative for displaying intra-oral movements.

The studies also differ in the attempted realism of the articulator appearance, anatomy and movements. For the appearance, several researchers, e.g., Badin et al. (2008); Massaro et al. (2008) intentionally depart from realism by choosing contrasting colours for the different articulators. No user study has yet been performed to investigate whether viewers prefer easier discrimination between articulators or caricaturized realism. The meaning of the latter would be that the appearance does not have to be photo-realistic, but that the articulator colours have the expected hue. Concerning anatomy, the models were created from Magnetic Resonance Imaging (MRI) (Badin et al., 2008; Engwall, 2010; Engwall & Wik, 2009a; Wik & Engwall, 2008) or adapted through fitting of an existing geometric model to data from MRI (Fagel & Madany, 2008; Kröger et al., 2008) or three-dimensional ultrasound (Cohen et al., 1998). For the articulatory movements, Badin et al. (2008); Engwall & Wik (2009a) used actual Electromagnetic articulography (EMA) measurements of the uttered sentences, while the other studies used rule-based text-to-speech synthesis. Experiments reported in Section 3.2 indicate that this choice may have an influence on speech perception. On the one hand, prototypic or exaggerated movements created by rules may be easier to understand than real tongue movements, but on the other, real movements may be closer to the viewer's own production and therefore more easily processed subconsiously.

A final issue regarding realism concerns the appearance of the face and its correspondence with the intra-oral parts. A video-realistic face may have benefits both for pleasantness of appearance and possibly also for speech perception, since finer details may be conveyed by the skin texture. There is however a risk of the so called uncanny valley effect when the intra-oral articulation is shown within a video-realistic face. In the current scope, the uncanny valley effect signifies that users may perceive the talking head as unpleasant if the face has a close-to-human appearance, but includes non-human augmented reality, with parts of the skin removed or transparent. This question is further discussed in Section 5.

3. AR talking heads as a speech perception support

AR talking heads as a speech perception support have been investigated in several studies in the last years (Badin et al., 2008; Engwall, 2010; Engwall & Wik, 2009a; Grauwinkel et al., 2007; Kröger et al., 2008; Wik & Engwall, 2008). The studies have shown that even if the intraoral articulators give much less information than the face, at least some listeners benefit from seeing tongue movements; but only if they have received explicit or implicit training on how to interpret them.

Badin et al. (2008) tested audiovisual identification of all non-nasal French voiced consonants in symmetrical vowel-consonant-vowel (VCV) contexts with [a, i, u, y] and different levels of signal-to-noise ratio (SNR). To one group the stimuli was presented in four decreasing steps of SNR, from clean conditions to muted audio, whereas the steps were reversed with increasing SNR for the other group. The first group hence received implicit training of the relationship between the acoustic signal and the tongue movements. Four different conditions were presented to the subjects, acoustic only and three audiovisual conditions. They were a cutaway display showing the outline of the face, the jaw and palate and pharynx walls, but not the tongue (AVJ); a cutaway display that in addition also showed the tongue (AVT); and a display showing the face with skin texture instead (i.e., a realistic display, rather than AR). The main results of the study were that the identification score was better for all audiovisual displays than for the acoustic only, but that the realistic display was better than the two augmented reality displays (of which AVT was the better). The subjects hence found it easier to employ the less detailed, but familiar, information of the face. The group that had received implicit training was however significantly better in the AR conditions than the one that had not. For the first group, the AVT display was moreover better than the realistic display in mute condition.

Similarly, Grauwinkel et al. (2007) concluded that the additional information provided by animations of the tongue, jaw and velum was not, in itself, sufficient to improve the consonant identification scores for VCV words in noise. Ten German consonants in symmetric [a, i, u] context were presented in white noise at SNR=0 to two groups of subjects who saw either the external face or a semi-transparent face with movements of the tongue and velum. The audiovisual recognition scores were significantly higher than the acoustic ones, but the subject group that saw an AR face was not significantly better than the one that saw a non-transparent face, *unless* subjects had received training prior to the test. The training was in the form of a video presentation that explained the place and manner of articulation of the consonants and the movement of the articulators for all consonants in all vowel contexts in a side view display.

Kröger et al. (2008) performed a visual only test of 4 vowels and 11 consonants with German articulation disordered children. Mute video animations of the articulatory movements at half speed were displayed in a 2D- or 3D-model and the children were asked to acoustically mimic the sound they saw. One repetition was used for the 2D-model and two, with different views, for the 3D-model. The phoneme recognition rates and correct identification of articulatory features (i.e., the case when the child produced a different phoneme, but it had the same type of lip rounding, place of articulation, manner of articulation or used the same articulator, as in the stimuli) were significantly above chance level and similar for the two models.

The implications of these three studies for general speech perception are nevertheless limited, since only forced-choice identification of consonants and four isolated vowels were tested. If the articulatory display is to be used as an alternative to cued speech, a more varied and less restricted corpus needs to be tested as well. It is also of interest to explore the importance of realism of the displayed articulatory movements. Finally, the role of the training merits further investigation to determine if the subjects are learning the audiovisual stimuli as audiovisual templates or if they start to make use of already established articulatory knowledge. In order to do so, we have conducted a series of tests, focused on the use of AR talking heads as a general speech perception support (Section 3.1), on comparing speech perception with authentic and rule-generated articulatory movements (Section 3.2) and on the subjects internalized articulatory knowledge (Section 3.3).

Fig. 2. The dual face display showing a normal front view and an AR side view simultaneously. The picture in addition shows the experimental display set-up with an entry frame, in which the subjects typed in the sentence that they perceived.

3.1 AR talking heads as an alternative to cued speech

In the first study, we tested a setting simulating what a hearing impaired person could use as a speech reading support. A group of listeners were presented vocoded speech accompanied by a dual display, showing a normal front view of the face and an augmented reality side view (c.f. Fig. 2). Vocoded speech is a good simulation of a hearing impairment and a dual display would be used in a speech reading support, since the front view is the best for lip reading, while the side view is better to show the articulation of the tongue.

3.1.1 Stimuli and subjects

The stimuli consisted of acoustically degraded short Swedish sentences spoken by a male Swedish speaker. The audio degradation was achieved by using a noise-excited channel vocoder that reduces the spectral details and creates an amplitude modulated and bandpass filtered speech signal consisting of multiple contiguous channels of white noise over a specified frequency range (Siciliano et al., 2003). In this chapter, the focus is placed on 30 sentences presented with a three-channel vocoder, but Wik & Engwall (2008) in addition give results for sentences presented with two channels.

The sentences have a simple structure (subject, predicate, object) and "everyday content", such as *"Skjortan fastnade på en spik"* (*The shirt got caught on a nail*). These sentences are part of a set of 270 sentences designed for audiovisual speech perception tests, based on MacLeod & Summerfield (1990). The sentences were normally articulated and the speech rate was kept constant during the recording of the database by prompting the speaker with text-to-speech synthesis set to normal speed.

The sentences were presented in three different conditions: Acoustic Only (AO), Audiovisual with Face (AF) and Audiovisual with Face and Tongue (AFT). For the AF presentation a frontal view of the synthetic face was displayed (left part of Fig. 2) and the AFT presentation in addition showed a side view, where intra-oral articulators had been made visible by making parts of the skin transparent (Fig. 2).

18 normal-hearing native subjects (15 male and 3 female) participated in the experiment. All were current or former university students and staff. They were divided into three groups, with the only difference between groups being that the sentences were presented in different conditions to different groups, so that every sentence was presented in all three conditions, but to different groups. The sentence order was random, but the same for all subjects.

3.1.2 Experimental set-up

The acoustic signal was presented over headphones and the graphical interface was displayed on a 15″ laptop computer screen. The perception experiment started with a familiarization set of sentences in AFT condition, in which the subjects could listen to and watch a set of five vocoded and five clear sentences as many times as they wanted. The correct text was then displayed upon request in the familiarization phase. When the subjects felt prepared for the actual test, they started it themselves. For each stimulus, the subjects could repeat it any number of times and they then typed in the words that they had heard (contrary to the familiarization phase, no feedback was given on the answers during the test). No limit was set on the number of repetitions, since the material was much more complex than the VCV words of the studies cited above and since subjects in Badin et al. (2008) reported that it was difficult to simultaneously watch the movements of the lips and the tongue in one side view. Allowing repetitions made it possible for the subjects to focus on the front face view in some repetitions and the augmented side view in others. This choice is hence similar to that in Grauwinkel et al. (2007), where each stimulus was repeated three times. The subjects' written responses were analyzed manually, with the word accuracy counted disregarding morphologic errors.

3.1.3 Results

The results for the two audiovisual conditions were significantly better than the acoustic only, as shown in Fig. 3(a). A two-tailed t-test showed that the differences were significant at a level of $p<0.05$. The word recognition for the two audiovisual conditions was very similar, with word accuracy 70% *vs.* 69% and standard deviation 0.19 *vs.* 0.15 for AF *vs.* AFT. Overall, the augmented reality display of the tongue movements did hence not improve the performance further compared to the normal face view, similar to the findings by Badin et al. (2008) and Grauwinkel et al. (2007). Fig. 3(a) however also shows that the performance differed substantially between the groups, with higher accuracy in AFT condition than in AF for groups 1 and 2, but lower for group 3.

The reason for this may be any of, or a combination of, differences in the semantic complexity between the sentence sets, in the phonetic content of the sentences between the sentence sets or in the distribution of individual subjects' ability between the subject groups. Sentences and subjects were distributed randomly between their three respective groups, but it could be the case that the sentences in one set were easier to understand regardless of condition, or that one group of subjects performed better regardless of condition. Since the sentence sets were presented in different conditions to the subject groups, both differences between sentence sets and subject groups can make comparisons between conditions unfair. The differences between sentence sets and subject groups were therefore first analyzed. For the sets, the average word accuracy was 71% for set 1, 59% for set 2 and 64% for set 3, where the difference between sets 1 and 2 is statistically significant at $p<0.005$, using a paired t-test, whereas the difference between sets 1 and 3 and between sets 2 and 3 is non-significant. For the groups,

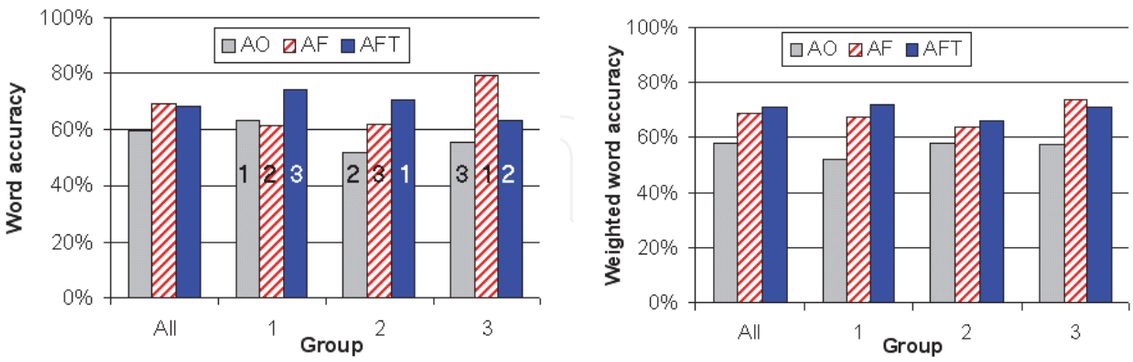
the average word accuracy was 66% for group 1, 62% for group 2 and 66% for group 3, and none of the intra-group differences are significant.

There is hence an artifact of set difficulty that needs to be taken into account in the following analysis. In order to be able to compare display conditions without the influence of the intra-set differences, a weighted word accuracy was calculated, in which the average score of each set was normalized to the average of the three sets (66%). The word accuracy for sentences belonging to set 1 was decreased by multiplying it by a factor 0.66/0.71=0.92, while that of sets 2 and 3 was increased by a factor 1.12 and 1.03, respectively. The weighted word accuracy for the different display conditions is displayed in Fig. 3(b). The difference between the weighted AF and AO conditions is significant at a level of $p<0.05$, while that between AFT and AO is significant at $p<0.001$. The difference between the two audiovisual conditions is still not significant.

The intra-subject differences are a natural consequence of different subjects having different multimodal speech perception abilities to make use of augmented reality displays of intraoral articulations, and this was also observed in the study by Badin et al. (2008) (personal communication). Fig. 4 shows that six subjects (1:3, 1:5, 1:6, 2:3, 2:6, 3:3) clearly benefited from the augmented reality view, with up to 20% higher weighted word accuracy scores in AFT than in AF, while three others (2:4, 3:2, 3:5) were as clearly better in the AF condition.

In future studies we plan to use an eye-tracking system to investigate if the differences between subjects may be due to where they focus their visual attention, so that subjects who have higher recognition scores in the augmented reality condition give more attention to the tongue movements. Such an evaluation has also been proposed by Badin et al. (2008).

In order to analyze how different phonetic content influenced the speech perception in different display conditions, the average word accuracy per sentence was first considered. Fig. 5 shows the weighted word accuracy, where the effect of differences in subject performance between the groups has been factored out through a normalization procedure equivalent to that described for the sentence set influence (however, contrary to the set



(a) Average word accuracy

(b) Average weighted word accuracy

Fig. 3. Word accuracy for all subjects and the three different groups. a) The numbers in the bars indicate which set of sentences that was presented in the different conditions. b) The weighting is a normalization, applied to factor out the influence of intra-set differences.

influence, the effect of the different subject groups was marginal, with scale factors 0.98, 0.98 and 1.05).

From Fig. 5 one can identify the sentences for which AFT was much better than AF (sentences 5, 9, 10, 17, 21, 22, 28) and vice versa (1-3, 6-8, 12, 27). A first observation concerning this comparison of the two audiovisual conditions is that of the first eight sentences, seven were more intelligible in the AF display. This suggests that the subjects were still unable to use the additional information from the AFT display, despite the familiarization set, and were initially only confused by the tongue animations, whereas the more familiar AF view could be used as a support immediately. The very low AFT score for sentence 12 is probably due to a previously unnoticed artifact in the visual synthesis, which caused a chaotic behavior of the tongue for a few frames in the animation.

The analysis of the sentences that were better perceived in AFT than in AF condition is tentative and needs to be supported by more, and more controlled, experimental data, where sentences can be clearly separated with respect to the type of articulatory features they contain. As a first hypothesis, based on the words that had a higher recognition rate in AFT condition, it appears that subjects found additional information in the AFT display mainly for the tongue dorsum raising in the palatal plosives [k, g] and the tongue tip raising in the alveolar lateral approximant [l] and the alveolar trill [r]. In addition, the fricatives [ʃj, ɕ] also seem to have been better perceived, but they appeared in too few examples to attempt hypothesizing. The animations of the tongue in particular appear to have been beneficial for the perception of consonant clusters, such as [kl, ml, pl, sk, st, kt, rd, rt, rn, dr, tr], for which the transitions are difficult to perceive from a front face view.

Note that there is a weak negative correlation (σ=-0.09) between the number of repetitions for a sentence and the accuracy rates, and the accuracy rate is hence not increased if the subjects listened to the stimuli additional times. The word accuracy decreased almost monotonously with the number of repetitions after an initial peak (at 1-2 repetitions for AO and AF and at 3 for AFT), as shown in Fig. 6. A two factor ANOVA with number of repetitions and display condition as factors indicates that there is no interaction between number of listenings and display condition for the word recognition accuracy. Fig. 6 also shows that, on average, the

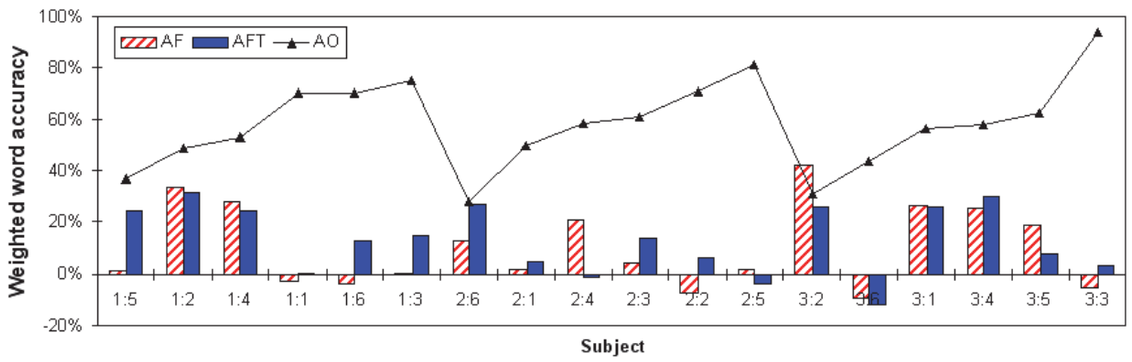


Fig. 4. Average weighted mean word accuracy per subject in the acoustic only (AO) condition and the change compared to the AO condition when the AF or the AFT display is added. Numbers on the x-axis indicate group and subject number within the group. Subjects have been sorted on increasing AO performance within each group.

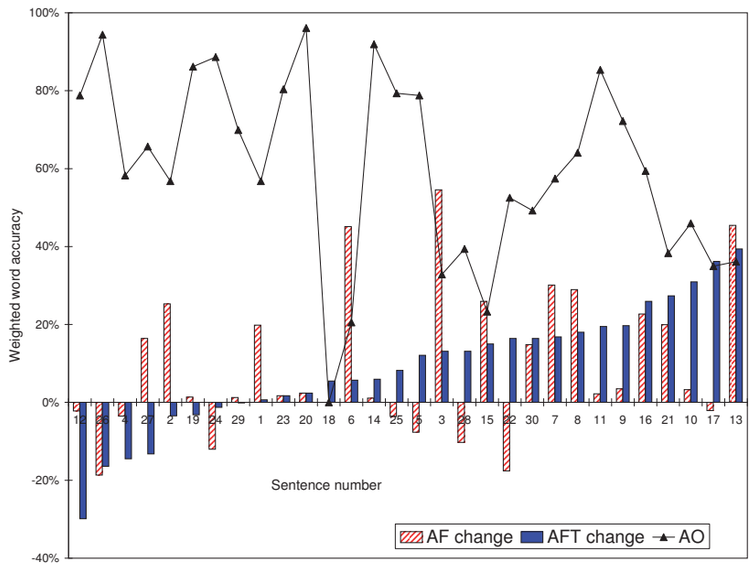


Fig. 5. The weighted mean word accuracy for each stimulus in the acoustic only (AO) condition and the change compared to the AO condition when the AF or the AFT display is added. The sentences have been sorted in order of increasing AFT change.

stimuli were mostly played two, three or more than six times. From the number of repetitions used and the corresponding word accuracy, it appears that the subjects were either certain about the perceived words after 1-3 repetitions, or they used many repetitions to try to decode difficult sentences, but gained little by doing so. Fig. 6 suggests that the additional repetition

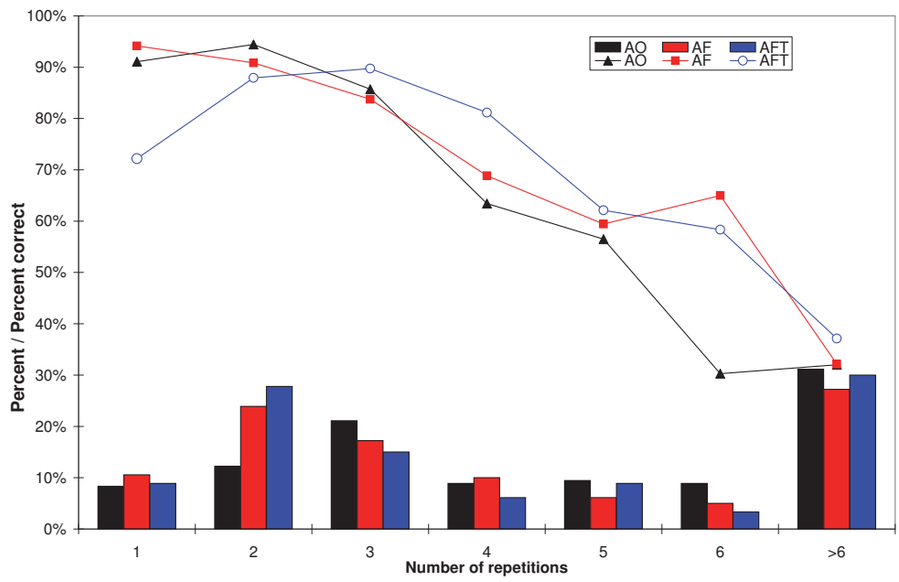


Fig. 6. Lines show the average weighted mean word accuracy in the three display conditions as a function of number of times the stimulus was repeated before the subject gave an answer. Bars show the distribution of the average number of repetitions for the different conditions.

with the AFT display allowed users to take more information from both face views into account.

Due to the rapidity of the tongue movements and the comparably low word recognition for one repetition, it seems unrealistic that the AR talking head could be used as an alternative to cued speech for real-time speech perception for an average person, at least not without large amounts of training. However, the study shows that some subjects are indeed very strong "tongue readers", and such persons could well be helped by an AR talking head display. The following two sections continue to explore how the tongue movements in the augmented reality animations are processed by the listeners.

3.2 On the importance of realism of articulator movements

As described in Section 2, both rule-based and recorded articulator movements have been used in the AR animations. Movements created from rules are more prototypic, may be hyperarticulated (i.e. more exaggerated) and have no variation between repetitions of the same utterance. Recorded movements display speaker specific traits, with variability and more or less clearly articulated words, but they are, on the other hand, *natural* movements. We have performed a study on VCV words and sentences to investigate if the difference between these two types of movements influences the perception results of the viewers. Realistic tongue movements could be more informative, because the listener can unconsciously map the displayed movements to his or her own, either through activation of mirror neurons (Rizzolatti & Arbib, 1998) when seeing tongue movements, or if the theory of speech motor control is applicable (Perkell et al., 2000). It may, on the other hand, be so that the rule-based movements give more information, because the hyperarticulation means that the articulations are more distinct. This was indeed found to be the case for the velar plosive [g], for the part of this test on VCV words (Engwall & Wik, 2009a). The consonant identification rate was 0.44 higher with animations displaying rule-based [g] movements than for those with real movements. For other consonants ([v, d, l, r, n, s ɕ, ɧ]), the difference was either small or with the recorded movements resulting in higher identification rates. For a description of the test with VCV words, please refer to Engwall & Wik (2009a), as the remainder of this section will deal with sentences of the same type as in Section 3.1.

3.2.1 Stimuli and subjects

For the animations based on recorded data (AVR), the movements were determined from measurements with the MacReflex motion capture system from Qualisys (for the face) and the Movetrack EMA (for the tongue movements) of one female speaker of Swedish (Beskow et al., 2003). For the face, 28 small reflective markers were attached to the speaker's jaw, cheeks, lips and nose, as shown in Fig. 7(a). To record the tongue movements, three EMA coils were placed on the tongue, using a placement shown in Fig. 7(b). In addition, EMA coils were also placed on the jaw, the upper lip and the upper incisor. Beskow et al. (2003) describe how the recorded data was transformed to animations in the talking head model through a fitting procedure to minimize the difference between the data and the resynthesis in the model.

For the rule-based synthesis animations (AVS), the movements were created by a text-to-visual speech synthesizer with forced-alignment (Sjölander, 2003) to the recorded acoustic signal. The text-to-visual speech synthesis used is an extension to the one created for the face by Beskow (1995) and determines the articulator movements through targets for each phoneme

(a) Motion capture markers (b) EMA coil placement

Fig. 7. Set-up used to collect data for the AVR animiations. (a) Placement of the Qualisys motion capture markers. (b) The corresponding virtual motion capture markers (+) and articulography coils (circled) in the talking head model.

and interpolation between targets. The targets and the timing for the tongue movements are based on data from static MRI and dynamic EMA (Engwall, 2003), but the interpolation is the same as for the face, which might not be suitable for the tongue movements, since they are much faster and of a slightly different nature. It has been shown that the synthetically generated face animations are effective as a speech perception support (Agelfors et al., 1998; Siciliano et al., 2003), but we here concentrate on if the synthesis is adequate for intraoral animations.

The stimuli were 50 Swedish sentences of the same type (but not necessarily the same content) and with the same acoustic degradation as described in Section 3.1.1. The sentences were divided into three sets S1, S2 and S3, where S1 contained 10 stimuli and S2 and S3 20 stimuli each.

The subjects were 20 normal-hearing native speakers of Swedish (13 male and 7 female). They were divided into two groups, I and II. The sentences in S2 were presented in AVS condition to Group I and in AVR to Group II, while those in S3 were presented in AVR to Group I and in AVS to Group II. Both groups were presented S1 in acoustic only (AO) condition. To determine the increase in word recognition when adding the AR animations to the acoustic signal, a matched control group (Group III) was presented all stimuli in AO. For the comparisons below, the stimuli were hence the same as for Groups I and II, but the subjects were different in Group III. The results on set S1 were therefore used to adjust the scores of the control group so that the AO baseline performance corresponded to that of Groups I-II on S1, since inter-group differences could otherwise make inter-condition comparisons invalid.

3.2.2 Experimental set-up

The AR talking head shown in Fig. 1(a) was used to display the animations and the acoustic signal was presented over high-quality headphones. The sentence order was the same for all subjects and the display condition (AVR, AVS or AO) was random, but balanced, so that all conditions were equally frequent at the beginning, middle and end of the test.

Each sentence was presented three times before the subjects typed in their answer in five entry frames. The five frames were always active, even if the sentence contained fewer words.

Before the test, the subjects were given the familiarization task to try to identify the connection between the sound signal and tongue movements in five sentences presented twice with normal acoustic signal and twice with degraded.

3.2.3 Results

Both types of animations resulted in significantly higher word recognition rates than the acoustic only condition, when comparing the perception results for Groups I and II with those of Group III for sets S2 and S3, as shown in Table 2. When considering the two audiovisual conditions, the word recognition rate was 7% higher when the animations were based on recorded data than when they were synthesized, and the difference is highly significant, using a single factor ANOVA ($p<0.005$).

	AO	AVS	AVR
acc.	54.6%	56.8%	63.9%
std.	0.12	0.09	0.09

Table 2. Word accuracy rates (acc.) and standard deviation (std) when the stimuli were presented as acoustic only (AO), with animations created from synthesis (AVS) and from measurements (AVR). The differences AVR-AO and AVS-AO are significant at $p<0.005$, using a paired two-tailed t-test.

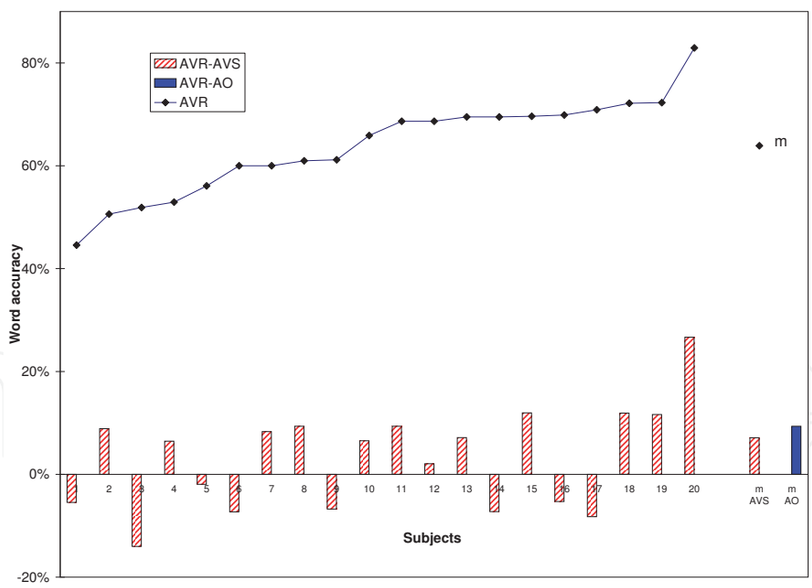


Fig. 8. Rate of correctly recognized words for animations with recorded data (AVR, black line) and difference in recognition rate between AVR and synthetic movements (AVS, striped bars) for each subject. The AVR average for the group (m) and the average improvement for the group, compared to AVS (m AVS, striped bar) and acoustic only (m AO, blue bar) is also given. Subjects are shown in order of increasing AVR score.

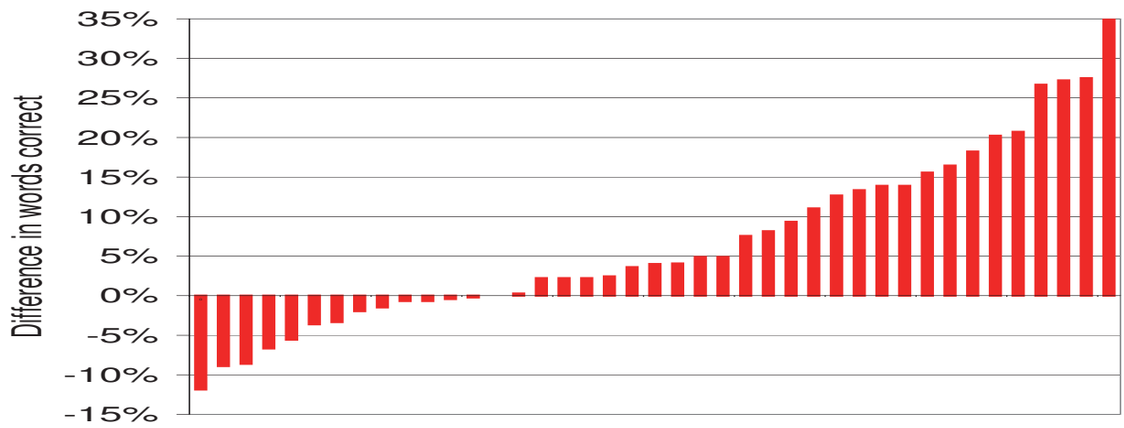


Fig. 9. Difference in recognition rate between AVR and AVS for each sentence. The sentences have been sorted in order of increasing AVR-AVS difference.

Since the same subject was not presented the same sentences in both AVR and AVS, the recognition scores were weighted, so that the average score for the two sets S2 and S3 over all subjects and both conditions AVR and AVS is the same. The scale factors were w_{S1}=0.97 and w_{S2}=1.03. As shown in Fig. 8, the accuracy rate in the AVR condition varied between 45% and 85% for different subjects, and 40% of the subjects actually performed better in AVS condition. The majority of the subjects nevertheless performed better in the AVR condition, and whereas four subjects were more than 10% better with AVR, only one was 10% better with AVS.

The word accuracy rate per sentence, shown in Fig. 9, was higher in AVR for 70% of the sentences, and for about half of these, the difference is large. For one of the sentences (*"Snön låg metertjock på marken"*, i.e. *"The snow lay a meter deep on the ground"*), the word accuracy is 35% higher in AVR, and the difference is significant at the sentence level at $p<0.0005$.

In a follow-up study, published in Engwall & Wik (2009b), it was shown that subjects (of which 11 out of the 22 were the same as in the experiment presented here) could not judge if an animation was created from real recordings or from text-to-speech synthesis. It is hence the case that even though subjects are unaccustomed to seeing tongue movements and can not consciously judge if the animations are truthful representations of the tongue movements, they are, as a group, nevertheless better if the actual articulations that produced the acoustics are displayed.

A possible explanation for this would be that there is a more direct connection between speech perception and articulatory movements, rather than a conscious interpretation of acoustic and visual information by the subjects. There are indeed several theories and evidence that could point in that direction. Skipper et al. (2007) showed that perception of audiovisual speech leads to substantial activities in the speech motor areas of the listener's brain and that the activated areas when seeing a viseme are the same as when producing the corresponding phoneme. However, the connection between visemes and speech perception could be established through experience, when seeing the speaker's face producing the viseme simultaneously with hearing the phoneme, whereas we here deal with a connection between acoustics and visual information that is not normally seen. A potential explanation could be provided by the direct realist theory of speech perception (Fowler, 2008), which

states that speech is perceived through a direct mapping of the speech sounds to the listener's articulatory gestures. Hence, seeing the gestures may influence perception unconsciously. Similarly, the speech motor theory (Liberman & Mattingly, 1985) stipulates that both acoustic and visual gestures are processed in accordance with how the speaker produced them. This would explain why the AVS animations, which are realistic, but are not necessarily in accordance with the speaker's gestures, gave lower recognition rates than AVR, where acoustic and visual gestures correspond.

The above explanations are however problematic, since the speaker's and the listener's oral anatomy differ, and they would use slightly different gestures to produce the same sequence of sounds. It is hence unclear if the listener could really map the speaker's articulatory gesture's to his or her own. An alternative explanation is provided by the fuzzy logical theory of speech perception (Massaro, 1998), which argues that perception is a probabilistic decision based on previously learned templates. Acoustic and visual information is processed independently and then combined in a weighted fusion to determine the most probable match with both sources of information. While this appears to be a plausible explanation for visemes (see further the explanation of the McGurk effect in Section 3.3), it is unclear how the visual templates for the tongue movements could have been learned. In the next section, this issue of learning is investigated further.

3.3 How do people learn to "read" tongue movements?

All perception studies cited above indicated that a training phase in some form was required if the subjects should be able to use the information provided by the AR talking head. A fundamental question is then what the subjects learn during this training phase: Is it a conscious mapping of articulatory movements to corresponding phonemes in a template learning scheme? Or are tongue reading abilities pre-existing, and the role of the training phase is to make subjects sub-consciously aware of how to extract information from animations of articulatory movements?

In order to investigate this issue, the so called McGurk effect (McGurk & MacDonald, 1976) can be used. The McGurk effect describes the phenomenon that if the acoustic signal of one phoneme is presented together with the visual lip movements of another, it is often the case that a third phoneme is perceived, because of audiovisual integration. For example, if auditory [ba] is presented with visual [ga], then for the very large majority of subjects [da] is perceived. The reason is that the visual signal is incompatible with [ba] (since the lip closure is missing) and the acoustic with [ga] (the acoustic frequency pattern in the transition from the consonant to the following vowel is wrong) and the brain therefore integrates the two streams of information to perceive [da], which is more in agreement with both streams. It should be noted that this effect is sub-conscious, that the subject actually perceives [da], and that the effect appears even for subjects who know about the conflicting stimuli.

For the AR talking heads, the McGurk effect was adapted to create mismatches between the acoustic signal and the tongue movements in the AR display, rather than with face movements in a normal display. Subjects were then randomly presented either matching stimuli (the acoustics and the animations were of the same phoneme) or conflicting (McGurk stimuli). The underlying idea was that if the subjects had an existing subconscious notion of general articulatory movements, then the perception score for the matching stimuli should be higher and that some type of McGurk effect should be observed for the conflicting stimuli.

3.3.1 Stimuli and subjects

24 different symmetric VCV words, with C=[p, b, t, d, k, g, l, v] and V=[a, ɪ, ʊ], uttered by a female Swedish speaker, were presented at four different levels of white noise (signal-to-noise ratio SNR=+3dB, -6dB, -9dB and Clean speech) and three different audiovisual conditions. The stimuli were presented in blocks of 48 stimuli at each noise level, in random order between noise levels and audiovisual conditions, but in the same order for all subjects. The 48 stimuli consisted of the 24 VCV words played in acoustic only condition (AO), plus 12 of these VCV words played with the animations of the tongue matching the acoustic signal (AVM) and 12 played with animations of the tongue movements that were in conflict with the acoustics (AVC).

The conflicting animations were created by in turn combining the acoustic signal of each of the bilabials [p, b], alveolars [t, d] and velars [k, g] with tongue movements related to one of the other two places of articulation. The conflicting condition for [l] was visual [v] and vice versa. The display excluded the lip area (in order to avoid that lip movements, rather than those of the tongue, influenced the results), and the labial consonants [p, b, v] therefore constitute a special case for both AVM and AVC. Since the subjects did not see the articulation of the lips, AVM in this case signifies that there were *no conflicting* tongue movements in the animation, and AVC for acoustic [k, g, t, d, l] with the animation showing the articulation of [p, b, v] in this case signifies that there were *no supporting* tongue movements in the animation.

Subjects were divided into two groups, with the only difference between groups being that the were presented the AVM and AVC stimuli in opposite conditions. That is, Group I was presented Set 1=[apːa, ɪdːɪ, ukːu, ɪbːɪ, utːu, agːa, upːu, adːa, ɪkːɪ, alːa, ɪvːɪ, ulːu] with matched animations and Set 2=[abːa, ɪtːɪ, ugːu, ɪpːɪ, udːu, akːa, ubːu, atːa, ɪgːɪ, avːa, ɪlːɪ, uvːu] with conflicting. Group II was, on the other hand, presented Set 1 with conflicting and Set 2 with matching animations. Note that Sets 1 and 2 are balanced in terms of vowel context and consonant place of articulation and voicing, i.e., if Set 1 contains a VCV word with an unvoiced consonant, then Set 2 contains the voiced consonant having the same place of articulation in the same vowel context, and this is reversed for another vowel context.

The 18 subjects (13 male and 5 female, aged 21-31 years, no known hearing impairment) had different language backgrounds. Four were native speakers of Swedish; two each of Greek, Persian and Urdu; and one each of German, English, Serbian, Bangla, Chinese, Korean, Thai and Tamil. The heterogeneous subject group was chosen to investigate if familiarity with the target articulations influenced perception results. The question is relevant in the light of the use of AR talking heads for pronunciation training of a foreign language (c.f. Section 4). The influence of the subjects' first language is further discussed in Engwall (2010), while we here deal with the general results.

The description in this chapter concentrates on the stimuli presented at SNR=-6dB, where the combination of audio and animations was the most important. An analysis of the results at the other noise levels is given in Engwall (2010).

3.3.2 Experimental set-up

Each stimulus was presented once, with the acoustic signal played over high quality headphones and the animations of the tongue movements shown on a 21" flat computer screen. AVM and AVC animations displayed the movements in an AR side view, such as

	AO	AVM	AVC
acc.	36.2%	43.1%	33.8%
std.	0.13	0.15	0.14

Table 3. Word accuracy rates (acc.) and standard deviation (std) when the stimuli were presented as acoustic only (AO), with matching animations (AVM) and with conflicting (AVC). The differences AVM-AO and AVM-AVC are significant at $p<0.05$, using a single factor ANOVA.

the one in Fig. 1(a), but translated to hide the lip area. For AO, an outside view, without any movements, was instead shown.

For the auditory stimuli, the SNR for the added white noise spectrum was relative to the average energy of the vowel parts of each individual VCV word and each VCV word was then normalized with respect to the energy level.

Before the test, a set of 9 VCV words with C=[m, n, ŋ] and V=[a, ɪ, ʊ] was presented in AVM at SNR=Clean, -6dB and -9dB, as a familiarization to the task. No feedback was given and these stimuli were not included in the test. The familiarization did hence not constitute a training phase.

A forced choice setting was used, i.e., subjects gave their answer by selecting the on-screen button for the consonant that they perceived. In the results below, accuracy is always counted with respect to the acoustic signal.

3.3.3 Results

The mean accuracy levels at SNR=-6dB are shown in Table 3. The differences between AVM and AO and between AVM and AVC are significant at $p<0.05$ using a single factor ANOVA. Note that voicing errors were disregarded and responses were grouped as [p/b], [t/d] and [k/g], giving a chance level of 20%. The reasons for this was that several subjects were from language backgrounds lacking the voiced-unvoiced distinction (such as between [t] and [d]) and that the aim was to investigate the influence of the visual information given about the tongue articulation. In the following, /p/ refers to [p, b], /t/ to [t, d] and /k/ to [k, g].

As a general result, the animations with matching articulatory movements hence gave an important support to the perception of the consonants in noise. This is all the more true if only the consonants that are produced with the tongue [t, d, k, g, l] are considered. Fig. 10 summarizes the individual and average perception scores (m_{AVM}=59%) for these consonants. The graph shows that 14 of the 18 subjects performed better with matched animations than with only audio and that 9 performed worse with conflicting animations than with audio only. Curiously, 9 subjects however performed better with conflicting animations than with audio only, indicating that one effect of presenting the animations may have been that the subjects listened more carefully to the acoustic signal than if the same acoustic signal was presented without animations. The graph also shows, just as the results for the studies presented above, that the differences between subjects were very large, with e.g., subject 18 being a particularly gifted tongue reader (100% recognition in AVM compared to 32.5% in AO)

When analyzing the responses with respect to accompanying animation shown in Fig. 11, several patterns appear, both in terms of the strength of the information given by a particular acoustic signal or articulatory movement and integration effects for conflicting acoustic and

visual signals. For the acoustic signal, [l] is the most salient with over 90% correct responses already with AO and consequently only marginal improvement with AVM or decline with AVC. On the other hand, the fricative [v] is particularly vulnerable to the background noise, with the AO accuracy level being half that of the next lowest, /p/. For the visual signal, the articulatory movement of /k/ has the strongest influence: For acoustic /k/, when the movement is shown in AVM, the accuracy in the responses increases with 50%, and when it is lacking in AVC, the accuracy decreases by 25%, regardless of if the animation shows no tongue articulation (for /p/) or a conflicting movement (for /t/). Further, for /t/, a conflicting /k/ animation decreases the recognition score in AVC by 10% compared to AO.

Concerning audiovisual integration, shown in Fig. 12, the changes listed in Table 4 are the most important that can be observed. Several of these changes are similar to the McGurk effect, even if the change is much smaller (and only took place with a noisy acoustic signal).

In conclusion for this study we can argue that the subjects must have a prior knowledge of articulatory movements of the tongue, since the animations were randomly matched and conflicting and the subjects performed significantly better with the matching movements. The conflicting animations further showed that subjects integrated both signals in their perception.

We are currently planning a follow-up study with a training phase prior to the test, in order to investigate if consistency between training and test or between acoustics and articulation is the most important. In this, subjects will be divided into four groups. Group I will be shown matching audiovisual stimuli in both training and test. Group II will be shown conflicting audiovisual stimuli in both training and test, but the audiovisual combinations would be consistent between training and test. Group III will be shown conflicting audiovisual stimuli

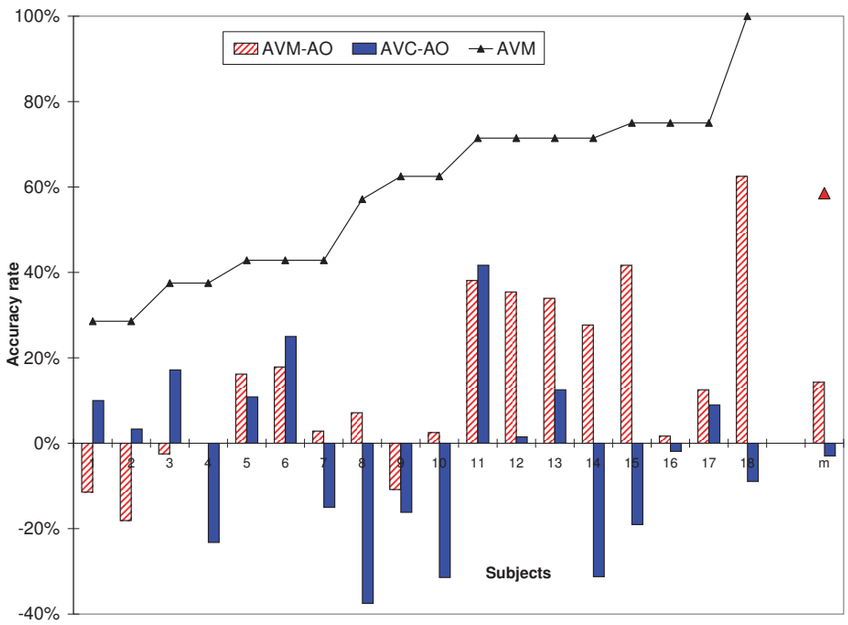


Fig. 10. The accuracy rate for [t,d,k,g,d,l] in the matched (AVM) condition (black line), and the difference between the matched AVM (red and white striped bars) or conflicting AVC (blue bars) conditions and the acoustic only (AO), for each individual subject, and for the group (m). Subjects are presented in order of increasing AVM score.

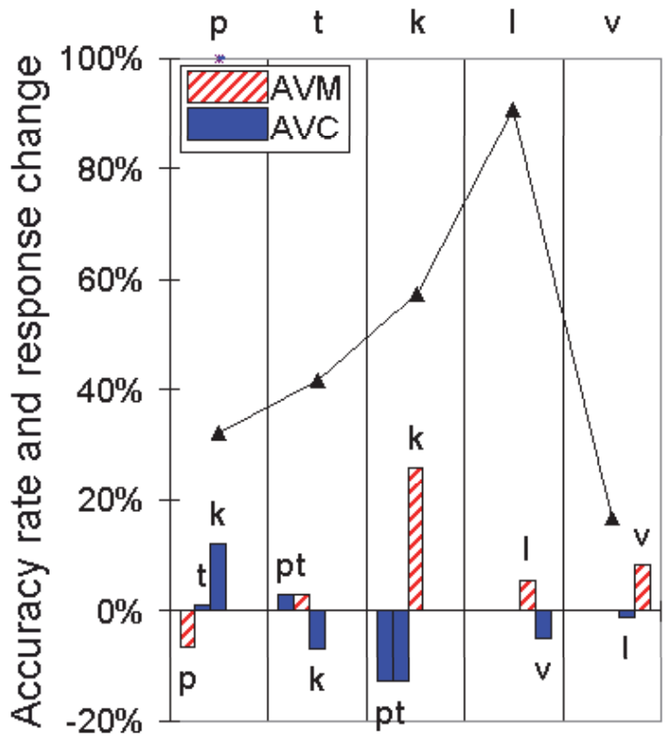


Fig. 11. The accuracy rate in acoustic only AO condition (black line), and the change (bars) when animations were added. The stimuli having the same acoustic signal ([p/b, t/d, k/g, l, v]) are grouped on the x-axis and within each group the bars indicate the difference in perception score, for that stimuli, compared to AO. Red and white striped bars signal matching condition AVM, blue bars conflicting AVC.

in the training, but matching in the test. Group IV will be shown matching audiovisual stimuli in the training, but conflicting in the test.

If match between the acoustic and visual signals is the most important, then Group I and Group III will have higher recognition scores than Groups II and IV. If, on the other hand, consistency between training and test is more important, Groups I and II will perform similarly, and better than Groups III and IV.

4. AR talking heads in speech production training

Several studies on the use of AR talking heads in pronunciation training have been performed (Engwall & Bälter, 2007; Engwall et al., 2006; Fagel & Madany, 2008; Massaro et al., 2008; Massaro & Light, 2003; 2004).

In Massaro & Light (2003), Japanese students of English were instructed how to produce /r/ and /l/ with either a normal front view of the talking face or with four different AR displays that illustrated the intraoral articulation from different views (c.f. Section 2). In Massaro & Light (2004), American hearing-impaired children were instructed how to produce consonant clusters, the fricative-affricate distinction and voicing differences in their native language, using the same four AR displays. In Massaro et al. (2008), English speakers were instructed

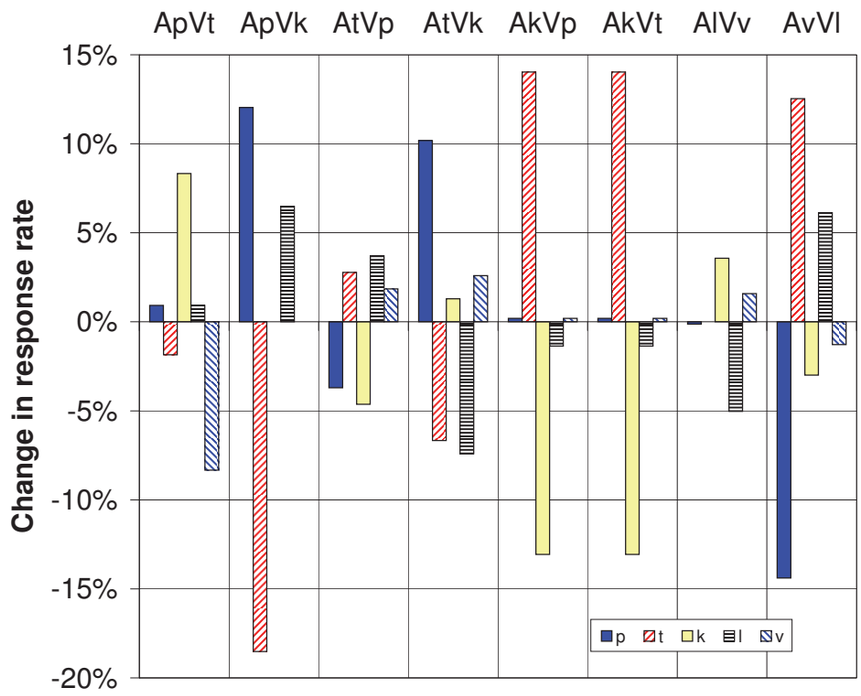


Fig. 12. The change in response rate for different consonant labels when comparing AVC with AO. The conflicting stimuli are given on the x-axis, with AxVy indicating that the acoustic signal was of consonant x and the visual signal of consonant y.

A	V	+	-	Explanation
p	t	k	v	The acoustic signal is incompatible with /t/ & the visual with /p, v/, /k/ is the most compatible with both.
p	k	p	t	The visual signal is incompatible with /t/, the jaw movement in /k/ may be interpreted as signaling a bilabial closure.
t	p		k	The visual signal is incompatible with /k/.
t	k	p	l	The acoustic signal is incompatible with /k/, the visual with /t,l/ & the jaw movement in /k/ may be interpreted as signaling a bilabial closure
k	p	t	k	The acoustic signal is incompatible with /p/ & the visual with /k/, /t/ is the most compatible with both.
k	t	t	k	The visual signal is compatible with /t/ and incompatible with /k/.
l	v		l	The visual signal is incompatible with /l/.
v	l	t	p	The acoustic signal is incompatible with /l/ & the visual with /p/, /t/ is the most compatible with both.

Table 4. Observed changes in response when a conflicting animation is added to an acoustic stimuli. The table lists the acoustic (A) and the visual (V) signals, the main increase (+) and decrease (-) in the subjects' responses compared to the AO condition and a tentative explanation on why the change takes place.

how to produce either one pair of similar phonemes in Arabic or one pair of similar phonemes in Mandarin. For the Arabic phoneme pair, the main articulatory difference was the place of contact between the tongue and the palate, while for the Mandarin pair, the difference was in the lip rounding. For the Arabic pair, a cut-away AR side-view was used to illustrate the position of the tongue, while a normal front view was used for the Mandarin pair.

These three studies did not provide any strong evidence that the AR view was beneficial, as judged by listeners rating the students production before and after the training. In Massaro & Light (2003), the students improved in both training conditions, but those who had been presented the AR displays did not improve more than those who had seen the normal face view. In Massaro & Light (2004), the children did improve, but since they were not compared with other subjects who had not been shown the AR talking heads, it can not be concluded that this was thanks to the augmented reality-based training. In Massaro et al. (2008), the group that practised the Mandarin pair with the normal face view had improved significantly more than a control group that had only been presented the auditory targets, while the group that practised the Arabic phoneme pair was not significantly better than the acoustic only control group. However, many of the subjects in the three studies reported that they really enjoyed the practise with the AR talking head and believed that it was useful.

The outcome in Fagel & Madany (2008) was somewhat better in terms of subject improvement. Children with pathological lisping were instructed how to produce [s, z] during two interactive lessons with a human teacher and the AR talking head was used as a tool to illustrate prototypic correct articulations. Listeners, who rated the degree of lisping before and after the lessons, judged that there was a significant reduction in lisping for the children as a group, but that there were large individual differences.

In Engwall et al. (2006), Engwall & Bälter (2007) and Engwall (2011), the subjects were given *feedback* on their articulation, using the AR talking head display shown in Fig. 1(a). The feedback was in the form of instructions on how to change the articulation, e.g., *"Lower the tongue tip and move the back of the tongue as far back as you can in the mouth and then slightly forward, to create a wheezing sound."* in order to change the articulation from [ʃ] to [ɧ], accompanied by animations illustrating the instructions in the AR talking heads. In the first study, the subjects were Swedish children with pronunciation disorders and in the other two, they were non-native speakers without prior knowledge of Swedish. The first two studies focused on user evaluations of the interface and the subjects' improvement in the production of the practice phoneme, the velar fricative [ɧ], was not investigated quantitatively. In the last study, the articulation change that the subjects did when they received feedback on the production of the Swedish [r] and [ɧ] was measured with ultrasound. Some subjects readily followed the audiovisual instructions, as exemplified in Fig. 13, showing two French speakers who changed their articulation from the French rhotic [ʁ] with a low tongue tip to the Swedish [r] with a raised tip, after a number of attempts and feedback. However, other subjects had great difficulties changing the articulation in the short practice session.

The studies described above indicate that it is not an easy task initially to transfer articulatory instructions to the own production. However, throughout the different studies, the subjects were positive about the usefulness of the AR talking heads. They stated that they thought that they had improved through the practise and that the feedback instructions had been helpful to change the articulation. As an example, in Engwall & Bälter (2007), subjects were asked to rate the pronunciation training system's usability on a number of aspects, using a 1-9 Likert

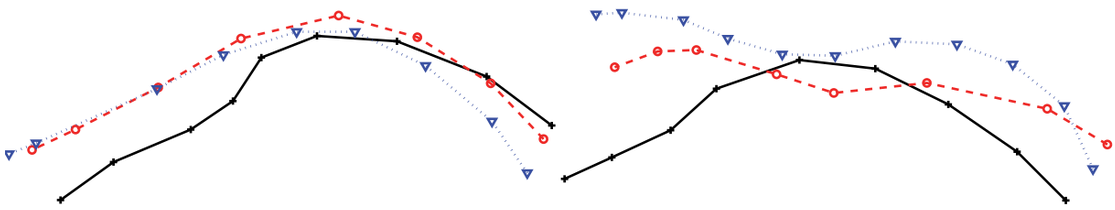


Fig. 13. Change in the tongue articulation of 'r', from a French rothic in the first attempt (black line, +) to a Swedish alveolar trill (red line, o; blue line, ▽) in the sequence "rik".

scale. For the question regarding if the articulatory animations were confusing (1) or clear (9), the mean opinion score was m=7.75 with a standard deviation of ρ=1.18, and the subjects further stated that the interaction with the virtual teacher was clear (m=7.68, ρ=1.04) and interesting (m=8.06, ρ=1.31). They also thought that the practice had helped them improve their pronunciation (m=7.32, ρ=1.37). It may hence well be the case that AR talking heads could provide the learners with useful information after additional familiarization. For a more thorough discussion of the use of AR talking heads in computer-assisted pronunciation training, please refer to Engwall & Bälter (2007) or Engwall (2011).

5. Paths for future research

Many questions concerning AR talking head support remain, regarding both the large differences between subjects and for which phoneme sequences the AR animations are helpful.

It would be interesting to monitor the subjects' visual attention using an eye-tracking system to investigate if there is evidence that the subjects' viewing patterns influence their performance. That is, if subjects with higher perception rates or adequate production changes focus more on the areas on the screen that provide important information about place and manner of articulation. By analysing the looking pattern one could also look into if the important factor is where the subject looks (on what part of the display) or how (focused on the task or watching more casually).

All the studies described above have further used naive subjects, in some cases with a short explicit or implicit training prior to the perception test. An intriguing question is if long-term use would make tongue reading a viable alternative for some hearing-impaired subjects, just as normal speech reading abilities improve with practice and experience. The experiments have shown that some subjects are in fact very apt at extracting information from the AR talking head displays and it would be of interest to investigate recognition accuracy and user opinion for such subjects after several longer practice sessions.

Further, systematic larger studies are required to determine how AR talking heads may be used as a support for speech perception and in production. With larger subject groups and a larger speech material, the potentials for long term use could be more clearly evaluated.

Another question, already introduced in Section 2, concerns the realism of the AR talking head: How are the recognition rates and the subjects' impression of the interface influenced by the visual representation? That is, are recognition rates higher with videorealistic animations of the face and/or the intraoral parts or with schematic, more simplified illustrations of the

most important articulatory features? We are planning for a perception study comparing the 3D AR talking head display with simplified 2D animations of the midsagittal tongue, lips and palate contours. Kröger et al. (2008) found that 5–8 year old children were as successful in mimicking vowels and consonants if they were presented a 2D- as a 3D-representation of the articulation. The authors therefore concluded that the more complex 3D view did not provide any additional information. It could however be argued that the 3D model used (only the surface of the tongue moving in a semi-transparent vocal tract tube, with no face, similar to the display in Fig. 1(e)) may not represent the full potential of AR talking heads.

Other issues that merit further consideration include user preferences for colour coding of different articulators (contrast *vs.* realism), correspondence in realism between the face and intraoral parts (potential problems of combining a videorealistic face with computer graphics for the tongue and mouth cavity) and strategy to create the Augmented Reality display (see-through or cut-away of the facial skin, or hiding the face).

For the last issue, we are interested in investigating the uncanny valley effect for Augmented Reality talking heads. The uncanny valley effect normally refers to the situation in robotics and 3D computer animation when human observers become negative towards the robot or avatar because it appears to be or act *almost* like a human, but either lack or have some aspects that one would or would not expect from a human. In the case of AR talking heads the effect could appear if the face of the avatar is so truthfully videorealistic that viewers feel uneasy about representations suggesting that the head of the avatar has been cut in half or that parts of the skin have been surgically removed. To avoid such reactions it may be suitable either to choose a less realistic face or to project intraoral information on the skin of the avatar's cheek instead. This could for example be in the form of stylized X-ray or MR Images, which many viewers are familiar with and they would hence immediately understand the analogy that these imaging techniques allow to see what is underneath the skin. With further advances in Augmented Reality display technology and response times for Automatic Speech Recognition one can also envisage that such a display could be used to provide information directly on a *real* speaker's cheek.

Even if much research is required to further investigate user reactions, preferences and performance with AR talking heads, we are convinced that they could potentially have an important role to play as a support for speech perception and production. In addition, the methods described above to illustrate tongue movements in an AR setting are also of interest for numerous other applications, such as planning for and rehabilitation after glossectomy surgery, education in phonetics, and animation in computer games and movies.

6. Acknowledgments

This work is supported by the Swedish Research Council project 80449001 Computer-Animated LAnguage TEAchers (CALATEA).

7. References

Agelfors, E., Beskow, J., Dahlquist, M., Granström, B., Lundeberg, M., Spens, K.-E. & Öhman, T. (1998). Synthetic faces as a lipreading support, *Proceedings of International Conference on Spoken Language Processing*, pp. 3047–3050.

Badin, P., Elisei, F., Bailly, G. & Tarabalka, Y. (2008). An audiovisual talking head for augmented speech generation: Models and animations based on a real speaker's

articulatory data, *in* F. Perales & R. Fisher (eds), *Articulated Motion and Deformable Objects*, Springer, pp. 132–143.

Benoît, C. & LeGoff, B. (1998). Audio-visual speech synthesis from French text: Eight years of models, design and evaluation at the ICP, *Speech Communication* 26: 117–129.

Benoît, C., Mohamadi, T. & Kandel, S. (1994). Effects of phonetic context on audio-visual intelligibility of French, *Journal of Speech and Hearing Research* 37: 1195–1203.

Beskow, J. (1995). Rule-based visual speech synthesis, *Proceedings of Eurospeech*, pp. 299–302.

Beskow, J., Engwall, O. & Granström, B. (2003). Resynthesis of facial and intraoral motion from simultaneous measurements, *Proceedings of International Congress of Phonetical Sciences*, pp. 431–434.

Beskow, J., Karlsson, I., Kewley, J. & Salvi, G. (2004). Synface - a talking head telephone for the hearing-impaired, *in* K. Miesenberger, J. Klaus, W. Zagler & D. Burger (eds), *Computers Helping People with Special Needs*, Springer-Verlag, pp. 1178–1186.

Cohen, M., Beskow, J. & Massaro, D. (1998). Recent development in facial animation: an inside view, *Proceedings of International Conference on Audiovisual Signal Processing*, pp. 201–206.

Cornett, O. & Daisey, M. E. (1992). *The Cued Speech Resource Book for Parents of Deaf Children*, National Cued Speech Association.

Engwall, O. (2003). Combining MRI, EMA & EPG in a three-dimensional tongue model, *Speech Communication* 41/2-3: 303–329.

Engwall, O. (2008). Can audio-visual instructions help learners improve their articulation? - an ultrasound study of short term changes, *Proceedings of Interspeech 2008*, pp. 2631–2634.

Engwall, O. (2010). Is there a McGurk effect for tongue reading?, *Proceedings of International Conference on Auditory-Visual Speech Processing*.

Engwall, O. (2011). Analysis of and feedback on phonetic features in pronunciation training with a virtual language teacher, *Computer Assisted Language Learning* .

Engwall, O. & Bälter, O. (2007). Pronunciation feedback from real and virtual language teachers, *Computer Assisted Language Learning* 20(3): 235–262.

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

Engwall, O. & Wik, P. (2009a). Are real tongue movements easier to speech read than synthesized?, *Proceedings of Interspeech*, pp. 824–827.

Engwall, O. & Wik, P. (2009b). Can you tell if tongue movements are real or synthetic?, *Proceedings of International Conference on Audiovisual Signal Processing*.

Fagel, S. & Madany, K. (2008). A 3-D virtual head as a tool for speech therapy for children, *Proceedings of Interspeech*, pp. 2643–2646.

Fowler, C. (2008). The FLMP STMPed, *Psychonomic Bulletin & Review* 15: 458–462.

Grauwinkel, K., Dewitt, B. & Fagel, S. (2007). Visual information and redundancy conveyed by internal articulator dynamics in synthetic audiovisual speech, *Proceedings of Interspeech*, pp. 706–709.

Guiard-Marigny, T., Ostry, O. & Benoît, C. (1995). Speech intelligibility of synthetic lips and jaw, *Proceedings of the International Congress of Phonetical Sciences*, pp. 222–225.

Kröger, B., Graf-Borttscheller, V. & Lowit, A. (2008). Two- and three-dimensional visual articulatory models for pronunciation training and for treatment of speech disorders, *Proceedings of Interspeech*, pp. 2639–2642.

Liberman, A. & Mattingly, I. (1985). The motor theory of speech perception revised, *Cognition* 21: 1–36.

MacLeod, A. & Summerfield, Q. (1990). A procedure for measuring auditory and audiovisual speech-reception thresholds for sentences in noise. Rationale, evaluation and recommendations for use, *British Journal of Audiology* 24: 29–43.

Massaro, D. (1998). *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle*, MIT Press.

Massaro, D., Bigler, S., Chen, T., Perlman, M. & Ouni, S. (2008). Pronunciation training: The role of eye and ear, *Proceedings of Interspeech*, pp. 2623–2626.

Massaro, D. & Light, J. (2003). Read my tongue movements: Bimodal learning to perceive and produce non-native speech /r/ and /1/, *Proceedings of Eurospeech*, pp. 2249–2252.

Massaro, D. & Light, J. (2004). Using visible speech for training perception and production of speech for hard of hearing individuals, *Journal of Speech, Language, and Hearing Research* 47: 304–320.

McGurk, H. & MacDonald, J. (1976). Hearing lips and seeing voices, *Nature* 264(5588): 746–748.

Perkell, J., Guenther, F., Lane, H., Matthies, M., Perrier, P., Vick, J., Wilhelms-Tricarico, R. & Zandipour, M. (2000). A theory of speech motor control and supporting data from speakers with normal hearing and with profound hearing loss, *Journal of Phonetics* 28: 233–272.

Rizzolatti, G. & Arbib, M. (1998). Language within our grasp, *Trends Neuroscience* 21: 188–194.

Siciliano, C., Williams, G., Beskow, J. & Faulkner, A. (2003). Evaluation of a multilingual synthetic talking face as a communication aid for the hearing impaired, *Proceedings of International Conference of Phonetic Sciences*, pp. 131–134.

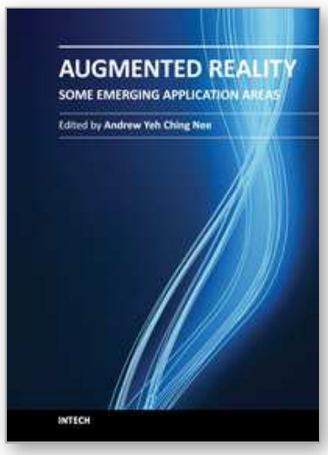
Sjölander, K. (2003). An HMM-based system for automatic segmentation and alignment of speech, *Proceedings of Fonetik*, pp. 93–96.

Skipper, J., Wassenhove, V. v., Nusbaum, H. & Small, S. (2007). Hearing lips and seeing voices: How cortical areas supporting speech production mediate audiovisual speech perception, *Cerebral Cortex* 17: 2387 – 2399.

Sumby, W. & Pollack, I. (1954). Visual contribution to speech intelligibility in noise, *Journal of the Acoustical Society of America* 26: 212–215.

Summerfield, Q. (1979). Use of visual information for phonetic perception, *Phonetica* 36: 314–331.

Wik, P. & Engwall, O. (2008). Can visualization of internal articulators support speech perception?, *Proceedings of Interspeech*, pp. 2627–2630.



Augmented Reality - Some Emerging Application Areas

Edited by Dr. Andrew Yeh Ching Nee

ISBN 978-953-307-422-1
Hard cover, 266 pages
Publisher InTech
Published online 09, December, 2011
Published in print edition December, 2011

Augmented Reality (AR) is a natural development from virtual reality (VR), which was developed several decades earlier. AR complements VR in many ways. Due to the advantages of the user being able to see both the real and virtual objects simultaneously, AR is far more intuitive, but it's not completely detached from human factors and other restrictions. AR doesn't consume as much time and effort in the applications because it's not required to construct the entire virtual scene and the environment. In this book, several new and emerging application areas of AR are presented and divided into three sections. The first section contains applications in outdoor and mobile AR, such as construction, restoration, security and surveillance. The second section deals with AR in medical, biological, and human bodies. The third and final section contains a number of new and useful applications in daily living and learning.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Olov Engwall (2011). Augmented Reality Talking Heads as a Support for Speech Perception and Production, Augmented Reality - Some Emerging Application Areas, Dr. Andrew Yeh Ching Nee (Ed.), ISBN: 978-953-307-422-1, InTech, Available from: http://www.intechopen.com/books/augmented-reality-some-emerging-application-areas/augmented-reality-talking-heads-as-a-support-for-speech-perception-and-production

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen