Can you tell if tongue movements are real or synthesized?

Olov Engwall, Preben Wik

Centre for Speech Technology, CSC, KTH, Stockholm, Sweden
engwall@kth.se, preben@kth.se

Abstract

We have investigated if subjects are aware of what natural tongue movements look like, by showing them animations based on either measurements or rule-based synthesis. The issue is of interest since a previous audiovisual speech perception study recently showed that the word recognition rate in sentences with degraded audio was significantly better with real tongue movements than with synthesized. The subjects in the current study could as a group not tell which movements were real, with a classification score at chance level. About half of the subjects were significantly better at discriminating between the two types of animations, but their classification score was as often well below chance as above. The correlation between classification score and word recognition rate for subjects who also participated in the perception study was very weak, suggesting that the higher recognition score for real tongue movements may be due to subconscious, rather than conscious, processes. This finding could potentially be interpreted as an indication that audiovisual speech perception is based on articulatory gestures.

Index Terms: augmented reality, tongue reading, visual speech synthesis, data-driven animation

1 Introduction

It is well-known that speech reading of the face supports speech perception, even if the face is computer-animated [1, 2, 3, 4]. Information provided by the speaker's lip shape, jaw position and eye-brow movements is important if the acoustic signal is degraded by noise [1, 4, 5] or a hearing-impairment [3, 6]. Speech reading of the tongue is on the other hand an ability that is seldom practiced, since most of the time, most of the tongue is hidden in a normal view of the face. Using augmented reality displays in talking heads, such as the one shown in Fig. 1, several recent speech perception studies [7, 8, 9, 10] have shown that even if animations of the tongue give less support than a normal view of the face, some subjects became able to extract information from the intra-oral animations, after some practice or instructions. Subjects in [7], who saw mute animations of vowels and VCV words in a 3D tube model display, were well above chance in mimicking the articulatory features (lip rounding, articulator used, narrowness, place of articulation and nasality) of the stimuli. Subjects in [8] who had been presented an instruction video that explained how the intra-oral articulators moved for different phonemes performed better in the consonant recognition task in noise than both the group that had not been shown the instruction and the subjects who saw a normal view of the face. Subjects in [9] similarly performed better for low signal-to-noise ratios (SNR) if the consonant recognition task started with clean audio and continued with decreasing SNR down to muted condition than if the SNR order was reversed (i.e., if they received implicit training or not). Subjects in [10], who were presented acoustically degraded sentences accompanied by either a normal front face view or the same view supplemented with an augmented reality side-view, perceived some sentences better if tongue movements were visible. Despite the differences in experimental condition between the four studies, regarding display (semi-transparent vocal tract model without a face, semi-transparent skin of the entire face, midsagittal cut-away view or transparent skin at the oral cavity), type of stimuli (vowels, VCV words or sentences), audio degradation (muted, speech in noise or vocoded speech), language (German, French or Swedish), number of repetitions (2, 3, 1 or unlimited), response format (mimicking, forced-choice or open answers), basis for the animations (synthesis by rules or measurements) and evaluation (accuracy of articulatory features, phonemes or words), the general conclusions were similar: tongue reading is unfamiliar and difficult, but can to some extent be learned.

1.1 AV perception with real *vs.* synthetic tongue movements

In a recent study [11], we investigated whether the type of animation would influence speech perception results. The animations were created either based on real movements, measured with motion capture and Electromagnetic Articulography (EMA), or using a rule-based synthesis. The experimental conditions in [11] were to large extents common with the study presented in this paper and are therefore described further in Section 2.

Compared to the condition when only degraded audio (AO) of short sentences was presented, the visualizations of tongue movements resulted in significantly better word recognition rates (at $p<0.005$ using a two-tailed paired t-test). Moreover, as shown in Fig. 2, real movements (AVR) resulted in a 7% higher word recog-

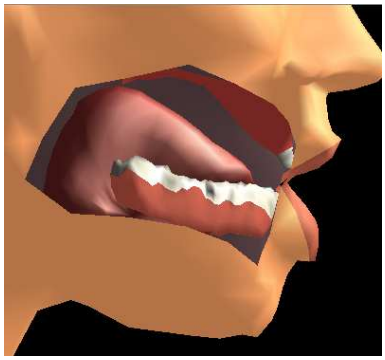


Figure 1: *Augmented reality side-view of the face, showing intra-oral articulatory movements.*

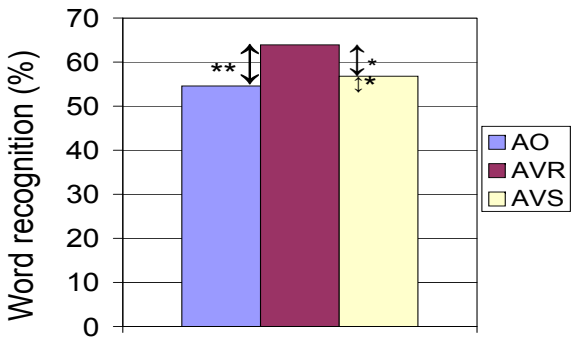


Figure 2: Percentage of words correctly recognized when presented in the different conditions Audio Only (AO), Audiovisual with Real (AVR) or Synthetic movements (AVS) in [11]. Stars indicate significant differences, at $p<0.005$ (*) or $p<0.00005$ (**).

nition rate (WRR) than if rule-based synthetic movements (AVS) were displayed, and the difference is significant ($p<0.005$). The recognition score was also higher in AVR than in AVS for a large majority of the sentences (28 out of 40). Note that, in order to avoid artifacts due to sentence content, the subjects in [11] were divided into three groups that were presented the sentences in different conditions, and the differences between the conditions are calculated on the same set of sentences.

However, Fig. 3 illustrates the fact that the WRR difference between the two types of animations differed between subjects. As each individual subject did not see the same sentences in the two conditions, the sentence content may have an unbalanced influence when the intrasubject difference is considered. A weighted difference, also shown in Fig. 3, was therefore calculated. In the weighted difference, the influence of sentence content has been factored out by scaling the results so that the average for the two sets of sentences was equal over all three conditions (AO, AVR and AVS) and all subjects. This weighted difference gives an indication of how the individual subject performed in the two audiovisual conditions relative a common baseline. Since different subjects performed differently for the two types of tongue animations, this follow-up study focuses on determining if subjects in general, and the subjects of the previous study in particular, can tell whether the displayed animations of the tongue are based on real measurements or not. By partly using the same subjects as in the previous study, we are able to investigate if subjects with higher AVR scores are more aware of what real tongue movements look like. Otherwise, the better recognition would be due to subconscious perception processes.

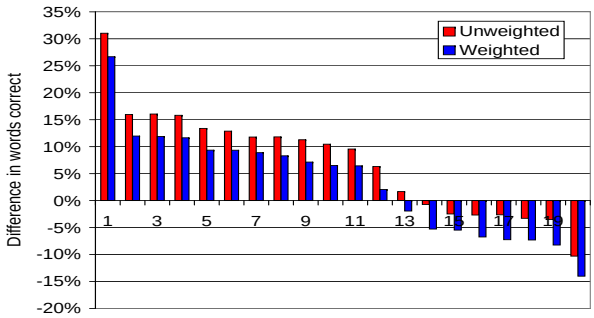


Figure 3: Unweighted and weighted difference of words correctly recognized when presented with real movements (AVR) compared to synthetic (AVS) for the 20 subjects in [11] (sorted in order of decreasing AVR-AVS difference). The weighting is applied to remove influence of sentence content.

If subjects are aware of the differences, it could be an indication of quality problems in the visual speech synthesizer, since it then creates movements that subjects consciously perceive as different from real movements.

On the other hand, if subjects are unaware of which tongue movements that are real, but still perform better for real movements, it might be an indication that visual gestures influence speech perception by a subconscious mapping to the listener's own articulatory gestures. There is evidence [12] 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 corresponds to the activated areas in the speaker when producing the same viseme. Several different theories have been proposed to account for the link between speech perception and production and the fact that it is influenced by visual information. The direct realist theory of speech perception [13] states that speech is perceived through a direct mapping of the speech sounds to the listener's articulatory gestures. This signifies that seeing the gestures may influence the perception, even if the listener is unaware of what the gestures should look like. The speech motor theory [14] is similarly based on gestures, but instead stipulates that the neural representation of the distal object is used to perceive (decode) abstract phoneme units. The listener would hence process both the acoustic and visual gestures in accordance with how the speaker produced them. The modulation theory [15] criticizes the speech motor theory e.g., on accounts of problems with variability of the distal object, because the encoder and decoder would be different due to differences in speaker and listener anatomy. The modulation theory instead proposes a special demodulation processing of acoustic speech to separate different types of information and that the visual gestures are perceived separately. The fuzzy logical theory of speech perception [2] argues that perception is a probabilistic decision that depends on the match compared to previously learned prototypes. Features from different sources of information, including visual, are combined to categorize speech stimuli into different categories. Audiovisual speech perception would hence be the result of a weighted fusion of the probabilities that the stimulus belongs to a certain category when the acoustic and visual information are considered independently.

The combination of this study and the one in [11] contributes to the investigation of the influence on visual gestures on speech perception by considering the unfamiliar visual information given by tongue movements, while the investigated visual gestures in [12] were facial, and therefore familiar to the subjects. Since real tongue movements result in significantly higher word recognition rate than synthesized, there seems to be a pre-established relation between visual tongue gestures and speech perception. In this study we attempt to investigate if this relation is conscious or subconscious and discuss what this could signify for the above audiovisual speech perception theories.

2 Experiments

The experiments were performed using an augmented reality side-view of a talking head, as shown in Fig. 1, in which the tongue movements have been made visible by making the skin at the cheek transparent. The tongue and jaw are shown as three-dimensional structures, whereas the palate is represented by the mid-sagittal outline and the upper incisor, with the other upper teeth removed, so as not to hide details in the tongue movements.

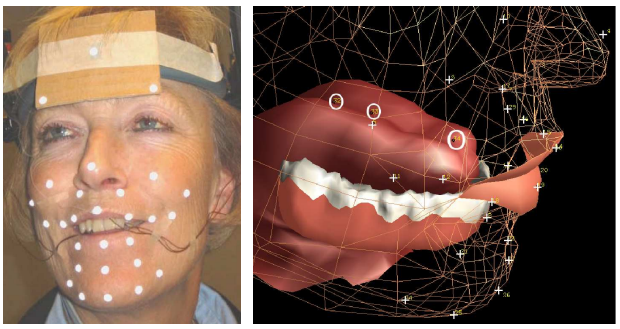
2.1 The talking head display

The talking head model consists of 3D-wireframe meshes of the face, jaw and tongue that are shaped by articulatory parameters. The parameters that are relevant for this study are jaw opening, shift and thrust; lip rounding; upper lip raise and retraction; lower lip depression and retraction; and tongue dorsum raise, body raise, tip raise, tip advance and width. The shape and parameters of the tongue model were created through a statistical analysis of Magnetic Resonance Imaging (MRI) data of a Swedish subject producing static vowels and consonants [16].

2.2 Animating the tongue movements

For the animations based on real tongue movements (AVR), the parameter values were determined directly from simultaneous and spatially aligned measurements of the face and the tongue for one female speaker of Swedish [17]. The Qualisys motion capture system with 28 reflectors, shown in Fig. 4(a), was used to measure the face movements in 3D and the Movetrack EMA [18] was used for the midsagittal movements of the tongue. Three EMA coils were placed on the tongue, one on the jaw and one on the upper incisor, as shown in Fig. 4(b). The animated movements were created by adjusting the parameter values of the face and tongue models to optimally fit the Qualisys-Movetrack (QSMT) data (c.f. [17] for the fitting procedure).

The synthetic (AVS) animations were generated with a rule-based visual speech synthesizer, by forced-alignment [19] of the phoneme input to the synthesizer with the acoustic signal recorded simultaneously with the QSMT measurements. The visual synthesizer was initially developed for the face [20] and it has been shown that the generated face animations are effective as a speech perception support [3, 4]. The movements in the model are created based on articulatory targets for each phoneme. For features that are not important for a certain phoneme, coarticulation occurs, and this is handled in the synthesizer by leaving the parameter value for this feature undetermined and instead letting the adjacent phonemes decide how the parameter varies, through linear interpolation with smoothing. Even if this model is adequate for the face, it is not certain that it is sufficient for the tongue movements, since these are both more rapid and more directly influenced by coarticulation. In fact, the results from [11] presented above suggest that the rule-based animations are not realistic enough, since they provided a weaker speech perception support than the animations created from real movements.



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

Figure 4: *(a) Placement of the motion capture markers. (b) The corresponding virtual motion capture markers (+) and articulography coils (circled) in the talking head model.*

2.3 Stimuli and subjects

The audiovisual stimuli consisted of 72 short, simple Swedish sentences, which were 3-6 words long with an "everyday content", e.g., "Den gamla räven var slug" (The old fox was cunning). For both AVR and AVS animations, half of the stimuli were presented with normal audio (the acoustics recorded together with the QSMT data) and half with degraded audio. The degraded signal was created from the normal audio using a three-channel vocoder that applies bandpass filtering and replaces the spectral details in the specified frequency ranges with white noise [4]. The advantage of using vocoded stimuli over speech in noise is that it is independent of the initial signal amplitude.

22 subjects (11 of each sex) participated in the test. All subjects were normal-hearing, native Swedes, with the age distribution being: 1 subject $<$20 years old, 6 subjects 20–30 yrs, 7 subjects 30–40 yrs, 6 subjects 40–50 yrs and 2 subjects $>$50 yrs. 11 (6 male and 5 female) of the subjects were recruited from the group of subjects in [11]. The subjects were divided into two groups I and II, with the only difference between the groups being that they saw each sentence in opposite condition (AVR or AVS). The two groups were balanced with respect to subjects having participated in the perception study or not, subject sex (both globally and for subjects from [11]) and age.

2.4 Experimental set-up

The animations were presented on a 15" flat screen and the acoustic signal was presented over headphones. The stimuli order was random, but the same for all subjects, which means that the relative AVR-AVS condition order was reversed for groups I and II. The animations were presented once each, but the subjects could repeat any animation once, if they were undecided about the type. The answer ("Real" or "Synthetic") was then given by pressing either of two buttons. After the classification test, but before they knew their score, the subjects were asked to give a short statement about how they had tried to decide whether the tongue movements were real or not.

2.5 Data analysis

The classification score for each subject and each of the four conditions – animations with real or synthetic movements accompanied by normal audio (AVRn, AVSn) or vocoded audio (AVRv, AVSv) – were summarized automatically by the test software. The scores were analyzed with respect to inter-condition differences and, for the subjects who had also participated in the perception test, for potential correlations between classification score and WRR differences.

Two average classification scores were calculated for both in-

dividual subjects i and the group of n=22 subjects. The first

$$\mu = \frac{\sum_{i=1}^{n} \mu(i)}{n}, \qquad \mu(i) = \frac{\sum_{s=1}^{N} c(s)}{N}$$

is the average proportion of correctly classified animations $c(s)$ out of $N = 72$ presentations. The second,

$$\Delta = \frac{\sum_{i=1}^{n} \Delta(i)}{n}, \qquad \Delta(i) = \frac{|\sum_{s=1}^{N} c(s) - C| + C}{N}$$

is the proportion of correctly discriminated animations, calculated using the absolute deviation from chance level C (i.e., C=36 correct answers and $0.5 \leq \Delta \leq 1$, where Δ=1 corresponds to μ=1 or μ=0 and Δ=0.5 is the chance level). This second score is calculated to handle the fact that a low, as well as a high classification score, signifies that the subject did see a difference between the two types of animations, even if they were mislabeled. For the μ score, high and low scores for different subjects will be factored out in the average. We however want to investigate both if the subjects are able to tell which animations that are real and if they can tell the two types apart (e.g., if subject A has 56 correct answers and subject B has 16, μ=(56+16)/(2×72)=50% but Δ=$\Delta(i)$=(20+36)/72=78%, indicating that considered as a group, subject A and B could see the difference between data-driven and rule-based animations, but not tell which were which).

The subjects' statement about the strategy employed to classify the tongue movements was used for analysis of the subjects' visual attention related to the classification results, but also to control for unwanted conscious discrimination strategies that were unrelated to the tongue movements. An additional subject (subject 11 from the perception study) did the classification test, but was removed from the analysis, because he stated that he made the decision purely on the appearance of the lips at the beginning of each animation and had seen no differences in the tongue movements. He had observed that the lips initially displayed small random vibrations for some animations and concluded that it must be for the data-driven movements. The results of this subject (93% correct answers) were hence due to the animations of the lips rather than the tongue, and would have to be considered as an artifact for this study.

3 Results

Fig. 5 summarizes the mean results, indicating that the subjects as a group were unable to tell which movements were real, with μ=48% very close to chance. The audio signal influenced the classification slightly, as synthetic movements were classified 6.5% more correctly if they were accompanied by vocoded audio than by normal. This difference is however not significant (p=0.10) and the only inter-condition difference that was significant (at p<0.05) was the one between AVRv and AVSv (all significance tests in this section use two-tailed paired t-tests). For AVR, there was no difference between the two acoustic conditions. It should further be noted that it does not seem as if subjects consciously grouped vocoded speech with the movements that they thought were synthetic, since 3/4th of the subjects who classified AVS incorrectly as real (i.e., who had μ <0.4) were also better at classifying AVSv than AVSn, just as subjects with higher μ.

The ability to see differences between the two types of animations was higher, but still modest, Δ=67% (standard deviation 0.12), corresponding to 12 answers above or below chance for the group.

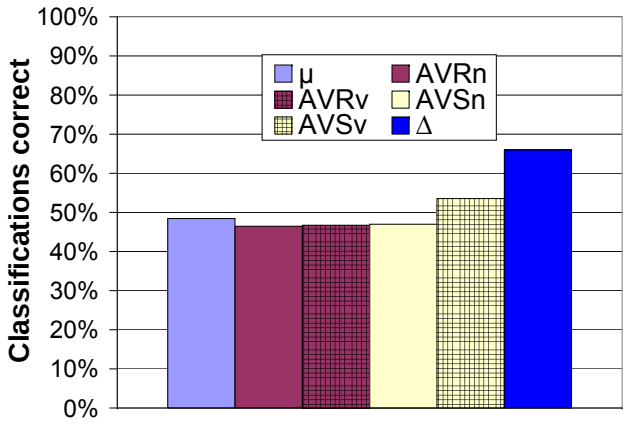


Figure 5: Mean classification score (μ) for all subjects and all stimuli, and for animations with real (AVRn, AVRv) or synthetic movements (AVSn, AVSv), accompanied by normal (r) or vocoded (v) audio. Δ is the mean discrimination score.

There was no significant difference between the two groups in classification or discrimination, but Fig. 6 shows that the variation between subjects was large. About half of the subjects were close to chance level and the other half were well above chance in discriminating between the two animation types, but the success in the choice of which each type was appears to be random (refer to Section 4 for a summary of the criteria that the subjects stated that they used to decide). Fig. 6 further displays the weighted difference in WRR between the AVR and AVS conditions for the 11 subjects who had also participated in the perception test. From the graph, one might be inclined to finding a relation between word recognition and classification score, as subject 1 (who had a very large AVR-AVS difference) had a higher than average classification score and subjects 8, 10 and 11 (subjects 15, 16 and 18 in Fig. 3) were the closest to chance level. However, all tests for possible correlations yield very weak coefficients, as summarized

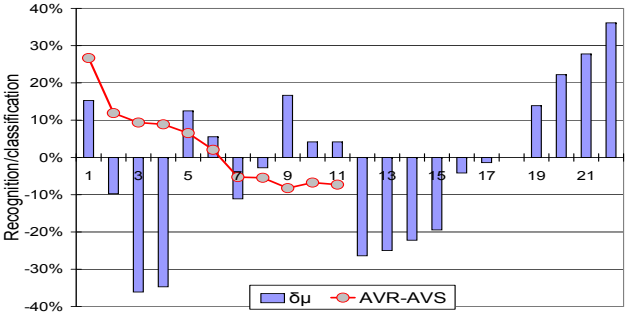


Figure 6: Classification score $\delta\mu$ relative chance-level ($\delta\mu$=μ-0.5). The x-axis crosses at chance level (36 correct answers) and the bars indicate scores above or below chance. For subjects 1–11, who participated in the perception test in [11], the weighted difference in WRR between the AVR and AVS conditions is also given. Subjects 1–11 are sorted according to the AVR-AVS difference, and subjects 12–22 with increasing $\delta\mu$.

Table 1: Correlation between the two classification scores and different WRRs in the perception test. ΔAV and ΔAVO are the differences in WRR for AVR relative AVS and AO, respectively. $\overline{AV}$ is the mean WRR for the two AV conditions.

	ΔAV	AVR	$\overline{AV}$	ΔAVO
μ	-0.11	0.21	0.31	-0.37
Δ	0.42	0.05	-0.14	0.20

in Table 3, with no correlation above 0.5. The tests were chosen to investigate the hypothesis that subjects with higher WRR or larger differences between conditions would reach higher classification or discrimination scores. Their results in the perception test could suggest that they were consciously more aware of the differences or looked more attentively at the tongue animations. The correlation between the classification score and the AVR-AVS WRR difference in [11] is however negative, and the highest correlation, for the combination audiovisual WRR difference and discrimination, is still low. The current data does hence not support the hypothesis that subjects in the perception test reached higher WRR with the real tongue movements because they consciously preferred them over the rule-based animations.

Other factors that could have influenced the classification and discrimination scores are the number of repetitions used (since the additional repetition could allow the subject to see more differences) and the stimuli number (since subjects may have become more apt over time to see differences). The average number of repetitions was 15 (21% of the stimuli repeated) and the mean classification score was indeed 6.8% higher when the animation had been repeated, but the difference is not significant (p=0.34). Since subjects were better at discriminating between the animations, and there were large differences in classification score between subjects, a modified discrimination score $\Delta\rho$ was also calculated to investigate the influence of repetitions and stimuli number. The modification of $\Delta\rho$ compared to Δ is that it calculated for each stimulus s as $\Delta\rho(s) = \sum_{s=1}^{n} \frac{\Delta\rho(i,s)}{n}$, where $\Delta\rho(i,s)$ is calculated for every subject i as

$$\Delta\rho(i,s) = \begin{cases} 1 & \text{if } c(i,s) = 1 \text{ and } \mu(i) \geq 0.5 \\ & \text{or } c(i,s) = 0 \text{ and } \mu(i) < 0.5 \\ 0 & \text{if } c(i,s) = 1 \text{ and } \mu(i) < 0.5 \\ & \text{or } c(i,s) = 0 \text{ and } \mu(i) \geq 0.5 \end{cases}$$

$\Delta\rho$ is a measure of the subjects' classification consistency relative to the animation type that they thought was real. For the repetitions, the discrimination was significantly higher (at p<0.005) for one repetition ($\Delta\rho$=0.71) than for two ($\Delta\rho$=0.48), illustrating that if the subject had not been able to discriminate after the first viewing of the animation, the repetition did only result in a random guess. There was no per-subject correlation between either $\mu(i)$ or $\Delta\rho(i)$ and the average number of repetitions.

Fig. 7 shows that the average classification score decreased slightly with stimuli number. The subjects did hence not learn to see which tongue movements that were real over the course of the test. However, as indicated by the modified discrimination score $\Delta\rho$, there is actually a slight learning effect in discriminating between the two types of animations. The subjects hence became slightly more consistent over time, but also more certain that it was the AVS animations that were the ones based on measurements. The correlation when fitting a line to the $\mu(s)$ or $\Delta\rho(s)$ data points with the least square criterion is moreover weak (Pearson's r=0.41–0.42).

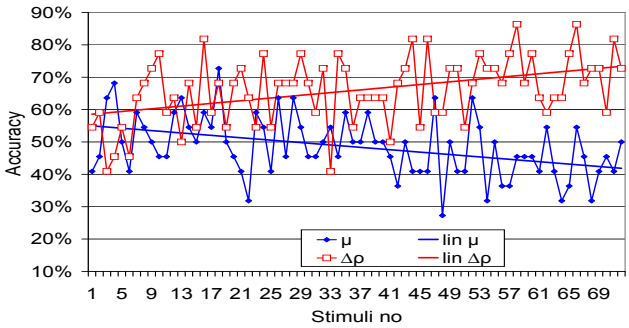


Figure 7: Mean classification and discrimination scores μ and $\Delta\rho$ as a function of stimuli number. $\Delta\rho$ is the correctness of the classification relative to the subjects' own AVR label, i.e., for subjects with μ <0.5, the binary correctness check was reversed.

4 Discussion & Conclusions

The above analysis gives no support to the hypothesis that subjects with larger WRR differences between the AVR and AVS conditions would be better at classifying the animations with respect to how they had been generated. In fact, our subjects were unable to judge which tongue movements were real. They could to a higher, but still modest, extent discriminate between the two types of animations (2/3rds correctly separated). In the explanations of what they had looked at to judge the realism of the tongue movements, two criteria appear to have been useful for discrimination (but not for classification, since they were used by both subjects with high and low μ): 1) The tongue tip contact with the teeth and 2) the range of articulation, since the synthetic movements were larger, and, as different subjects stated, "were more exaggerated" (resulting in high μ), alternatively "reached the places of articulation better" (resulting in low μ).

Several of the chance-level subjects stated that they had looked at the smoothness of the movement, assuming that rapid jerks occurred only in the synthetic animations. It is a rather common misconception that the tongue moves smoothly and graciously, and first-time viewers are very often surprised by how fast and rapidly changing real tongue movements actually are.

In conclusion, the subjects were hence unable to judge if an animation was created from real measurements or was synthesized by rules, and half of them were also unable to tell the different types of animations apart. The significantly higher word recognition rate in the perception test [11] could hence be an indication that subconscious processing of the augmented reality animations of the tongue movements occurs in audiovisual speech perception.

It is not possible to draw definite general conclusions regarding audiovisual speech perception from these two studies, since they are small and have several factors of uncertainty. The most important are that the variability between subjects in both tests was large; that the animation of the face may influence the results (as indicated by the subject who used lip vibrations to classify the animations, c.f. Section 2.5); and that the AVR and AVS animations did not only differ in the movements, but also in the range of articulation, since the targets for the rule-based synthesis were

determined for another speaker (and this could lead to higher discrimination scores).

Despite these caveats, the combination of the two studies opens up an intriguing perspective for future audiovisual perception research: Even though subjects cannot tell if a set of animations displays real articulatory movements, they are able to interpret acoustically degraded sentences much better when data-driven animations are displayed, compared not only to the acoustic only condition, but also to when the animations are generated by a visual speech synthesizer. A hypothesis for future work is hence that there is a coupling between speech motor planning and audiovisual speech perception, not only for visemes, for which prototypes could have been established from earlier face-to-face communication, but also for intra-oral articulatory gestures that are visually unfamiliar to the viewer. Interpretation of intra-oral articulations must instead be based either on conscious or subconscious mapping of the visual gestures to the own articulations, based on motor planning, neural response or active deduction of phonemic features from visual representation. We will not go as far as stating that this signifies that audiovisual speech is directly interpreted by the listener in terms of vocal tract configurations, as argued by [13], but it at least signifies that the processing of visual information in audiovisual speech perception is more than matching to previously established visual prototypes: articulatory gestures do seem to play a role in the perception. Should this be confirmed by future studies with other subjects, stimuli and visual representations, we would be able to judge the plausibility of different theories on audiovisual speech perception. More specific aspects that would be of interest include: does the similarity of the animated movements to the own articulatoty gestures play a role? Does perception benefit from display realism or do subjects perform active visual information retrival that would actually be improved with simplified iconic information on articulatory features (e.g., tongue-palate contact and distance)? Is the visual focus different for subjects with high and low audiovisual speech perception results (which can be evaluated using an eye-tracking system)? Is it the articulatory gesture or the articulatory targets that are important for the perception of realism (i.e., the criteria that subjects who differed from chance-level stated that they used were related to specific articulatory configurations rather than the movement, hence the question how would they do in a classification task with a series of still images, rather than an animation)?

5 Acknowledgements

This work is supported by the Swedish Research Council project 80449001 Computer-Animated LAnguage TEAchers (CALATEA). The estimation of parameter values from motion capture and articulography data was performed by Jonas Beskow.

References

[1] C. Benoît and B. LeGoff, "Audio-visual speech synthesis from French text: Eight years of models, design and evaluation at the ICP," *Speech Communication*, vol. 26, pp. 117–129, 1998.

[2] D. Massaro, *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle*. MIT Press, 1998.

[3] E. Agelfors, J. Beskow, M. Dahlquist, B. Granström, M. Lundeberg, K.-E. Spens, and T. Öhman, "Synthetic faces as a lipreading support," in *Proceedings of the International Conference on Spoken Language Processing*, 1998, pp. 3047–3050.

[4] C. Siciliano, G. Williams, J. Beskow, and A. Faulkner, "Evaluation of a multilingual synthetic talking face as a communication aid for the hearing impaired," in *Proceedings of the International Congress of Phonetic Sciences*, 2003, pp. 131–134.

[5] W. Sumby and I. Pollack, "Visual contribution to speech intelligibility in noise," *Journal of the Acoustical Society of America*, vol. 26, pp. 212–215, 1954.

[6] Q. Summerfield, "Use of visual information for phonetic perception," *Phonetica*, vol. 36, pp. 314–331, 1979.

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

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

[9] P. Badin, Y. Tarabalka, F. Elisei, and G. Bailly, "Can you "read tongue movements"?" in *Proceedings of Interspeech*, 2008, pp. 2635–2638.

[10] P. Wik and O. Engwall, "Can visualization of internal articulators support speech perception?" in *Proceedings of Interspeech*, 2008, pp. 2627–2630.

[11] O. Engwall and P. Wik, "Are real tongue movements easier to speech read than synthetic?" in *Proceedings of Interspeech*, 2009.

[12] J. Skipper, V. v. Wassenhove, H. Nusbaum, and S. Small, "Hearing lips and seeing voices: How cortical areas supporting speech production mediate audiovisual speech perception," *Cerebral Cortex*, vol. 17, pp. 2387 – 2399, 2007.

[13] C. Fowler, "The FLMP STMPed," *Psychonomic Bulletin & Review*, vol. 15, p. 458462, 2008.

[14] A. Liberman and I. Mattingly, "The motor theory of speech perception revised," *Cognition*, vol. 21, pp. 1–36, 1985.

[15] H. Traunmüller, "Evidence for demodulation in speech perception," in *Proceedings of International Conference on Spoken Language Processing*, vol. III, 2000, pp. 790–793.

[16] O. Engwall, "Combining MRI, EMA & EPG in a three-dimensional tongue model," *Speech Communication*, vol. 41/2-3, pp. 303–329, 2003.

[17] J. Beskow, O. Engwall, and B. Granström, "Resynthesis of facial and intraoral motion from simultaneous measurements," in *Proceedings of International Congress of Phonetic Sciences*, 2003, pp. 431–434.

[18] P. Branderud, "Movetrack – a movement tracking system," in *Proceedings of the French-Swedish Symposium on Speech, Grenoble*, 1985, pp. 113–122.

[19] K. Sjölander, "An HMM-based system for automatic segmentation and alignment of speech," in *Proceedings of Fonetik 2003*, 2003, pp. 93–96.

[20] J. Beskow, "Rule-based visual speech synthesis," in *Proceedings of Eurospeech*, 1995, pp. 299–302.