

# Real vs. rule-generated tongue movements as an audio-visual speech perception support

Olov Engwall and Preben Wik

Centre for Speech Technology, CSC, KTH

engwall@kth.se, preben@kth.se

## Abstract

*We have conducted two studies in which animations created from real tongue movements and rule-based synthesis are compared. We first studied if the two types of animations were different in terms of how much support they give in a perception task. Subjects achieved a significantly higher word recognition rate in sentences when animations were shown compared to the audio only condition, and a significantly higher score with real movements than with synthesized. We then performed a classification test, in which subjects should indicate if the animations were created from measurements or from rules. The results show that the subjects as a group are unable to tell if the tongue movements are real or not. The stronger support from real movements hence appears to be due to subconscious factors.*

## Introduction

Speech reading, i.e. the use of visual cues in the speaker's face, in particular regarding the shape of the lips (and hence the often used alternative term lip reading), can be a very important source of information if the acoustic signal is insufficient, due to noise (Sumby & Pollack, 1954; Benoît & LeGoff, 1998) or a hearing-impairment (e.g., Agelfors et al., 1998; Siciliano, 2003). This is true even if the face is computer animated. Speech reading is much more than lip reading, since information is also given by e.g., the position of the jaw, the cheeks and the eye-brows. For some phonemes, the tip of the tongue is visible through the mouth opening and this may also give some support. However, for most phonemes, the relevant parts of the tongue are hidden, and "tongue reading" is therefore impossible in human-human communication. On the other hand, with a computer-animated talking face it is possible to make tongue movements visible, by removing parts in the model that hide the tongue in a normal view, thus creating an augmented reality (AR) display, as exemplified in Fig. 1.

Since the AR view of the tongue is unfamiliar, it is far from certain that listeners are able to make use of the additional information in a similar manner as for animations of the lips. Badin et al. (2008) indeed concluded that the tongue reading abilities are weak and that subjects get more support from a normal view of the face, where the skin of the cheek is shown instead of the tongue, even though less information is given. Wik & Engwall (2008) similarly found that subjects in general found little additional support when an AR side-view as the one in Fig. 1 was added to a normal front view.

There is nevertheless evidence that tongue reading is possible and can be learned explicitly or implicitly. When the signal-to-noise ratio was very low or the audio muted in the study by Badin et al. (2008), subjects did start to make use of information given by the tongue movements – if they had previously learned how to do it. The subjects were presented VCV words in noise, with either decreasing or increasing signal-to-noise ratio (SNR). The group with decreasing SNR was better in low SNR conditions when tongue movements were displayed, since they had been implicitly trained on the audiovisual relationship for stimuli with higher SNR. The subjects in Wik & Engwall (2008) started the word recognition test in sentences with acoustically degraded audio by a familiarization phase, where they could listen to, and look at, training stimuli with both nor-



Figure 1. Augmented reality view of the face.

mal and degraded audio. Even though the total results were no better with the AR view than with a normal face, the score for some sentences was higher when the tongue was visible. Grauwinkel et al. (2007) also showed that subjects who had received explicit training, in the form of a video that explained the intra-oral articulator movement for different consonants, performed better in the VCV recognition task in noise than the group who had not received the training and the one who saw a normal face.

An additional factor that may add to the unfamiliarity of the tongue movements is that they were generated with a rule-based visual speech synthesizer in Wik & Engwall (2008) and Graunwinkel et al. (2007). Badin et al. (2008) on the other hand created the animations based on real movements, measured with Electromagnetic Articulography (EMA). In this study, we investigate if the use of real movements instead of rule-generated ones has any effect on speech perception results.

It could be the case that rule-generated movements give a better support for speech perception, since they are more exaggerated and display less variability. It could however also be the case that real movements give a better support, because they may be closer to the listeners' conscious or subconscious notion of what the tongue looks like for different phonemes. Such an effect could e.g., be explained by the direct realist theory of speech perception (Fowler, 2008) that states that articulatory gestures are the units of speech perception, which means that perception may benefit from seeing the gestures. The theory is different from, but closely related to, and often confused with, the speech motor theory (Liberman et al, 1967; Liberman & Mattingly, 1985), which stipulates that speech is perceived in terms of gestures that translate to phonemes by a decoder linked to the listener's own speech production. It has often been criticized (e.g., Traunmüller, 2007) because of its inability to fully explain acoustic speech perception. For *visual* speech perception, there is on the other hand evidence (Skipper et al., 2007) that motor planning is indeed activated when seeing visual speech gestures. Speech motor areas in the listener's brain are activated when seeing visemes, and the activity corresponds to the areas activated in the speaker when producing the same phonemes. We here investigate audiovisual processing of the more unfamiliar visual gestures of the tongue, using a speech perception and a classi-

fication test. The perception test analyzes the support given by audiovisual displays of the tongue, when they are generated based on real measurements (AVR) or synthesized by rules (AVS). The classification test investigates if subjects are aware of the differences between the two types of animations and if there is any relation between scores in the perception test and the classification test.

## Experiments

Both the perception test (PT) and the classification test (CT) were carried out on a computer with a graphical user interface consisting of one frame showing the animations of the speech gestures and one response frame in which the subjects gave their answers. The acoustic signal was presented over headphones.

### The Augmented Reality display

Both tests used the augmented reality side-view of a talking head shown in Fig. 1. Movements of the three-dimensional tongue and jaw have been made visible by making the skin at the cheek transparent and representing the palate by the midsagittal outline and the upper incisor. Speech movements are created in the talking head model using articulatory parameters, such as jaw opening, shift and thrust; lip rounding; upper lip raise and retraction; lower lip depression and retraction; tongue dorsum raise, body raise, tip raise, tip advance and width. The tongue model is based on a component analysis of data from Magnetic Resonance Imaging (MRI) of a Swedish subject producing static vowels and consonants (Engwall, 2003).

### Creating tongue movements

The animations based on real tongue movements (AVR) were created directly from simultaneous and spatially aligned measurements of the face and the tongue for a female speaker of Swedish (Beskow et al., 2003). The Movetrack EMA system (Branderud, 1985) was employed to measure the intraoral movements, using three coils placed on the tongue, one on the jaw and one on the upper incisor. The movements of the face were measured with the Qualisys motion capture system, using 28 reflectors attached to the lower part of the speaker's face. The animations were created by adjusting the parameter values of the talking head to optimally fit the Qualisys-Movetrack data (Beskow et al., 2003).

The animations with synthetic tongue movements (AVS) were created using a rule-based visual speech synthesizer developed for the face (Beskow, 1995). For each viseme, target values may be given for each parameter (i.e., articulatory feature). If a certain feature is unimportant for a certain phoneme, the target is left undecided, to allow for coarticulation. Movements are then created based on the specified targets, using linear interpolation and smoothing. This signifies that a parameter that has not been given a target for a phoneme will move from and towards the targets in the adjacent phonemes. This simple coarticulation model has been shown to be adequate for facial movements, since the synthesized face gestures support speech perception (e.g., Agelfors et al. 1998; Siciliano et al., 2003). However, it is not certain that the coarticulation model is sufficient to create realistic movements for the tongue, since they are more rapid and more directly affected by coarticulation processes.

### **Stimuli**

The stimuli consisted of short (3-6 words long) simple Swedish sentences, with an “everyday content”, e.g., “Flickan hjälpte till i köket” (The girl helped in the kitchen). The acoustic signal had been recorded together with the Qualisys-Movetrack measurements, and was presented time-synchronized with the animations.

In the perception test, 50 sentences were presented to the subjects: 10 in acoustic only (AO) condition (Set S1), and 20 each in AVR and AVS condition (Sets S2 and S3). All sentences were acoustically degraded using a noise-excited three-channel vocoder (Siciliano, 2003) that reduces the spectral details and creates a speech signal that is amplitude modulated and bandpass filtered. The signal consists of multiple contiguous channels of white noise over a specified frequency range.

In the classification test, 72 sentences were used, distributed evenly over the four conditions AVR or AVS with normal audio (AVRn, AVSn) and AVR or AVS with vocoded audio (AVRv, AVSv), i.e. 18 sentences per condition.

### **Subjects**

The perception test was run with 30 subjects, divided into three groups I, II and III. The only difference between groups I and II was that they saw the audiovisual stimuli in opposite conditions (i.e., group I saw S2 in AVS and S3 in AVR; group II S2 in AVR and S3 in AVS).

Group III was a control group that was presented all sets in AO.

The classification test was run with 22 subjects, 11 of whom had previously participated in the perception test. The subjects were divided into two groups I and II, again with the only difference being that they saw each sentence in opposite condition (AVR or AVS).

All subjects were normal-hearing, native Swedes, aged 17 to 67 years old (PT) and 12 to 81 years (CT). 18 male and 12 female subjects participated in the perception test and 11 of each sex in the classification test.

### **Experimental set-up**

Before the perception test, the subjects were presented a short familiarization session, consisting of five VCV words and five sentences presented four times, once each with AVRv, AVRn, AVSv and AVSn. The subjects in the perception test were unaware of the fact that there were two different types of animations.

The stimuli order was semi-random (PT) or random (CT), but the same for all groups, which means that the relative AVR-AVS condition order was reversed between groups I and II. The order was semi-random (i.e., the three different conditions were evenly distributed) in the perception test to avoid that learning effects affected the results.

Each stimulus was presented three times in the perception test and once in the classification test. For the latter, the subjects could repeat the animation once. The subjects then gave their answer by typing in the perceived words in the perception test, and by pressing either of two buttons (“Real” or “Synthetic”) in the classification test. After the classification test, but before they were given their classification score, the subjects typed in a short explanation on how they had decided if an animation was from real movements or not.

The perception test lasted 30 minutes and the classification test 10 minutes.

### **Data analysis**

The word accuracy rate was counted manually in the perception test, disregarding spelling and word alignment errors. To assure that the different groups were matched and remove differences that were due to subjects rather than conditions, the results of the control group on sets S2 and S3 were weighted, using a scale factor determined on set S1 by adjusting the average

of group III so that the recognition score on this set was the same as for the group I+II.

For the classification test, two measures  $\mu$  and  $\Delta$  were calculated for all subjects. The classification score  $\mu$  is the average proportion of correctly classified animations  $c$  out of  $N$  presentations. The discrimination score  $\Delta$  instead measures the proportion of correctly separated animations by disregarding if the label was correct or not. The measures are in the range  $0 \leq \mu \leq 1$  and  $0.5 \leq \Delta \leq 1$ , with  $\{\mu, \Delta\} = 0.5$  signifying answers at chance level. The discrimination score was calculated since we want to investigate not only if the subjects can tell which movements are real but also if they can see differences between the two animation types. For example, if subjects A and B had 60 and 12 correct answers,  $\mu = (60+12)/72 = 50\%$  but  $\Delta = (36+24)/72 = 67\%$ , indicating that considered as a group, subject A and B could see the difference between the two types of animations, but not tell which were which.

The  $\mu$ ,  $\Delta$  scores were also analyzed to find potential differences due to the accompanying acoustic signal, and correlations between classification and word recognition score for the subjects who participated in both tests.

## Results

The main results of the *perception test* are summarized in Fig. 2. Both types of tongue animations resulted in significantly better word recognition rates compared to the audio only condition (at  $p < 0.005$  using a two-tailed paired t-test). Moreover, recognition was significantly better ( $p < 0.005$ ) with real movements than with synthetic. For 28 of the 40 sentences, the rec-

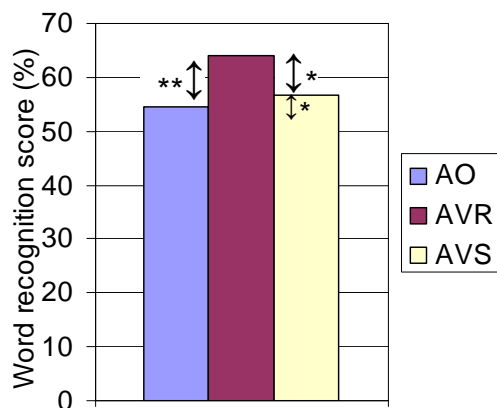


Figure 2. Percentage of words correctly recognized when presented in the different conditions Audio Only (AO), Audiovisual with Real (AVR) or Synthetic movements (AVS). The level of significance for differences is indicated by \* ( $p < 0.005$ ), \*\* ( $p < 0.00005$ ).

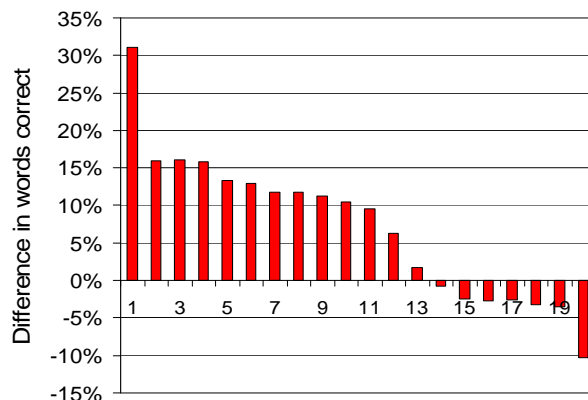


Figure 3. Difference of words correctly recognized when presented with real movements compared to synthetic movements for each of the 20 subjects.

ognition score was higher in AVR and for six sentences the difference was over 20%.

Fig. 3 however shows that there were large differences between subjects in how important the AVR-AVS difference was. Since individual subjects did not see the same sentences in the two conditions, the per-subject results may be unbalanced by sentence content. A weighted difference was therefore calculated that removes the influence of sentence content by scaling the results so that the average for the two sets of sentences was equal when calculated over all three conditions (AO, AVR, and AVS) and all three subject groups. The calculated weighted difference, displayed in Fig. 5 for the 11 subjects participating in both tests, indicates that while some subjects were relatively better with AVR animations, others were in fact better with AVS.

The main results of the *classification test* are shown in Fig. 4, where it can be seen that the subjects as a group were unable to classify

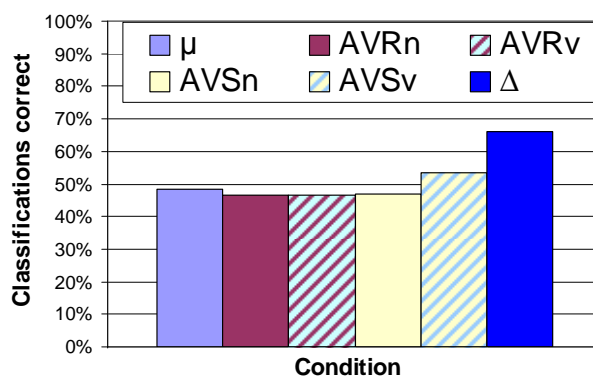


Figure 4. Mean classification score ( $\mu$ ) for all subjects, for all stimuli, and animations with real (AVRn, AVRv) or synthetic movements (AVSn, AVSv), accompanied by normal (n) or vocoded (v) audio.  $\Delta$  is the discrimination, i.e. the mean absolute deviation from chance.

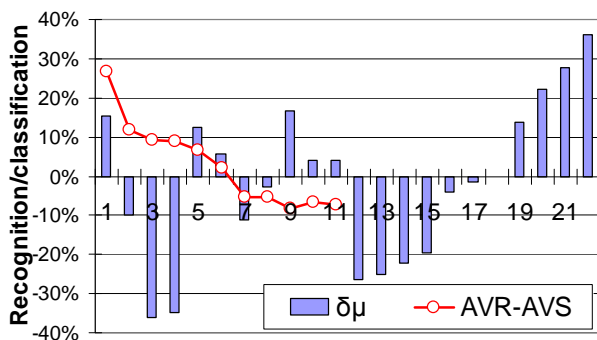


Figure 5. Classification score  $\delta\mu$  relative chance-level ( $\delta\mu=\mu-0.5$ ). The x-axis crosses at chance level and the bars indicate scores above or below chance. For subjects 1–11, who participated in the perception test, the weighted difference in word recognition rate between the AVR and AVS conditions is also given.

the two types of animations correctly, with  $\mu=48\%$  at chance level. The picture is to some extent altered when considering the discrimination score,  $\Delta=0.66$  (standard deviation 0.12) for the group. Fig. 5 shows that the variation between subjects is large. Whereas about half of them were close to chance level, the other half were more proficient in the discrimination task.

The classification score was slightly influenced by the audio signal, since the subjects classified synthetic movements accompanied by vocoded audio 6.5% more correctly than if they were accompanied by normal audio. There was no difference for the real movements. It should be noted that it was not the case that the subjects consciously linked normal audio to the movements they believed were real, i.e., subjects with low  $\mu$  (and high  $\Delta$ ) did not differ from subjects with high or chance-level  $\mu$ .

Fig. 5 also illustrates that the relation between the classification score for individual subjects and their difference in AVR-AVS word recognition in the perception test is weak. Subject 1 was indeed more aware than the average subject of what real tongue movements look like and subjects 8, 10 and 11 (subjects 15, 16 and 18 in Fig. 3), who had a negative weighted AVR-AVS difference in word recognition, were the least aware of the differences between the two conditions. On the other hand, for the remaining subjects there is very little correlation between recognition and classification scores. For example, subjects 4, 5 and 9 were much more proficient than subject 1 at discriminating between the two animation types, and subject 2 was no better than subject 7, even though there were large differences in perception results between them.

## Discussion

The *perception test* results showed that animations of the intraoral articulation may be valid as a speech perception support, since the word recognition score was significantly higher with animations than without. We have in this test not investigated if it is specifically the display of tongue movements that is beneficial. The results from Wik & Engwall (2008) and Badin et al. (2008) suggest that a normal view of the face without any tongue movements visible would be as good or better as a speech perception support. The results of the current study however indicate that animations based on real movements were significantly higher, and we are therefore currently working on a new coarticulation model for the tongue, based on EMA data, in order to be able to create sufficiently realistic synthetic movements, with the aim of providing the same level of support as animations from real measurements.

The *classification test* results suggest that subjects are mostly unaware of what real tongue movements look like, with a classification score at chance level. They could to a larger extent discriminate between the two types of animations, but still at a modest level (2/3 of the animations correctly separated).

In the explanations of what they had looked at to judge the realism of the tongue movements, two of the most successful subjects stated that they had used the tongue tip contact with the palate to determine if the animation was real or not. However, subjects who had low  $\mu$ , but high  $\Delta$  or were close to chance level also stated that they had used this criterion, and it was hence not a truly successful method.

An observation that did seem to be useful to discern the two types of movements (correctly or incorrectly labeled) was the range of articulation, since the synthetic movements were larger, and, as one subject stated, “reached the places of articulation better”. The subject with the highest classification rate and the two with the lowest all used this criterion.

A criterion that was not useful, ventured by several subjects who were close to chance, was the smoothness of the movement and the assumption that rapid jerks occurred only in the synthetic animations. This misconception is rather common, due to the rapidity and unfamiliarity of tongue movements: viewers are very often surprised by how fast and rapidly changing tongue movements are.

## Conclusions

The word recognition test of sentences with degraded audio showed that animations based on real movements resulted in significantly better speech perception than rule-based. The classification test then showed that subjects were unable to tell if the displayed animated movements were real or synthetic, and could to a modest extent discriminate between the two.

This study is small and has several factors of uncertainty (e.g., variation between subjects in both tests, the influence of the face movements, differences in articulatory range of the real and rule-based movements) and it is hence not possible to draw any general conclusions on audiovisual speech perception with augmented reality. It nevertheless points out a very interesting path of future research: the fact that subjects were unable to tell if animations were created from real speech movements or not, but received more support from this type of animations than from realistic synthetic movements, gives an indication of a subconscious influence of visual gestures on speech perception. This study cannot prove that there is a direct mapping between audiovisual speech perception and speech motor planning, but it does hint at the possibility that audiovisual speech is perceived in the listener's brain terms of vocal tract configurations (Fowler, 2008). Additional investigations with this type of studies could help determine the plausibility of different speech perception theories linked to the listener's articulations.

## 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

- Agelfors, E., Beskow, J., Dahlquist, M., Granström, B., Lundeberg, M., Spens, K.-E. and Öhman, T. (1998). Synthetic faces as a lipreading support. *Proceedings of ICSLP*, 3047–3050.
- Badin, P., Tarabalka, Y., Elisei, F. and Bailly, G. (2008). Can you "read tongue movements"? *Proceedings of Interspeech*, 2635–2638.
- Benoît, C. and 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.
- Beskow, J. (1995). Rule-based visual speech synthesis. *Proceedings of Eurospeech*, 299–302.
- Beskow, J., Engwall, O. and Granström, B. (2003). Resynthesis of facial and intraoral motion from simultaneous measurements. *Proceedings of ICPhS*, 431–434.
- Branderud, P. (1985). Movetrack – a movement tracking system, *Proceedings of the French-Swedish Symposium on Speech*, 113–122.
- Engwall, O. (2003). Combining MRI, EMA & EPG in a three-dimensional tongue model. *Speech Communication* 41/2-3, 303–329.
- Fowler, C. (2008). The FLMP STMPed, *Psychonomic Bulletin & Review* 15, 458–462.
- Grauwinkel, K., Dewitt, B. and Fagel, S. (2007). Visual information and redundancy conveyed by internal articulator dynamics in synthetic audiovisual speech. *Proceedings of Interspeech*, 706–709.
- Lieberman A, Cooper F, Shankweiler D and Studdert-Kennedy M (1967). Perception of the speech code. *Psychological Review*, 74, 431–461.
- Lieberman A & Mattingly I (1985). The motor theory of speech perception revised. *Cognition*, 21, 1–36.
- Siciliano, C., Williams, G., Beskow, J. and Faulkner, A. (2003). Evaluation of a multilingual synthetic talking face as a communication aid for the hearing impaired, *Proceedings of ICPhS*, 131–134.
- Skipper J., Wassenhove V. van, Nusbaum H. and Small, S. (2007). Hearing lips and seeing voices: how cortical areas supporting speech production mediate audiovisual speech perception. *Cerebral Cortex* 17, 387 – 2399.
- Sumby, W. and Pollack, I. (1954). Visual contribution to speech intelligibility in noise, *Journal of the Acoustical Society of America* 26, 212–215.
- Traunmüller, H. (2007). Demodulation, mirror neurons and audiovisual perception nullify the motor theory. *Proceedings of Fonetik 2007, KTH-TMH-QPSR 50*: 17–20.
- Wik, P. and Engwall, O. (2008). Can visualization of internal articulators support speech perception?, *Proceedings of Interspeech 2008*, 2627–2630.