Are real tongue movements easier to speech read than synthesized?

Olov Engwall, Preben Wik

Centre for Speech Technology, CSC, KTH, Stockholm, Sweden

engwall@kth.se, preben@kth.se

Abstract

Speech perception studies with augmented reality displays in talking heads have shown that tongue reading abilities are weak initially, but that subjects become able to extract some information from intra-oral visualizations after a short training session. In this study, we investigate how the nature of the tongue movements influences the results, by comparing synthetic rule-based and actual, measured movements. The subjects were significantly better at perceiving sentences accompanied by real movements, indicating that the current coarticulation model developed for facial movements is not optimal for the tongue.

Index Terms: multimodal speech perception, augmented reality, visual speech synthesis

1. Introduction

Several recent studies [1, 2, 3] have investigated if an augmented reality (AR) view of the face, where tongue movements are visible, can support speech perception. It has been proposed that such a display, exemplified in Fig. 1, could be useful for hearing-impaired listeners and second language learners. It is well established that visual information in the speaker's face, such as the lip shape, is beneficial for speech perception, especially in cases when the auditory information is not sufficient, due to noise or a hearing-impairment. This is true even if the face is synthetic [4, 5].

However, a normal view of the face can not (efficiently) transfer information about the place of articulation of the tongue for many phonemes, since it is too far back in the oral cavity to be seen. Cued speech [6] has therefore emerged as a complement to the acoustic signal and speech reading of the face for hearing-impaired listeners with residual hearing. The method relies on the speaker providing additional phonetic information with iconic hand gestures, but since these gestures are arbitrary, it has been proposed that an augmented reality animated face, in which both external and intraoral information is given, could be a viable alternative. Second language learners may also have difficulties perceiving unfamiliar phonetic features and may hence be helped by additional visual information when trying to establish the contrasts.

1.1. Previous studies

Since the augmented reality view of the tongue movements is unfamiliar to most listeners it is far from certain that such a display is beneficial. In fact, perception studies [1, 2, 3] have found that the inherent tongue reading skills are weak, i.e., initial recognition scores are no better than for a normal face view, where no tongue movements are visible. However, all three studies found that the ability to interpret animations of tongue movements can be acquired fast and effectively by some listeners, through a short, explicit or implicit, training session.

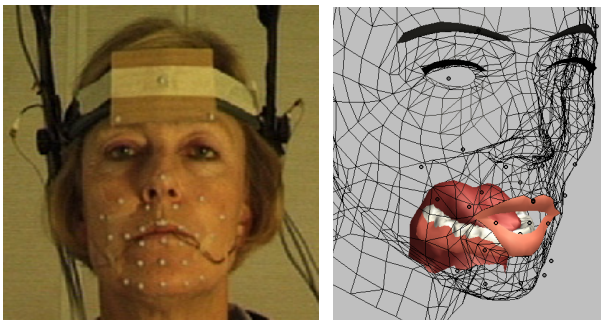
Grauwinkel at al. [2] showed an instruction video that explained the movement of the articulators for all consonants in all vowel contexts in an AR side view of the face to one of their subject groups. This group performed significantly better in the consonant recognition task in noise than both the subjects who had not received any training before seeing the AR display and those who saw a normal side view of the face, without tongue movements visible. Badin et al. [1] conducted a VCV perception study in noise, in which one group was presented the stimuli in order of increasing acoustic signal-to-noise ratio and the other in decreasing. The extremes were clear speech and visual only presentation. The subjects who started with clear speech, and hence received implicit training, were as a group significantly better. Wik & Engwall [3] started the sentence test with a familiarization phase, in which the subjects could listen to and watch the tongue movements in a set of five sentences with normal audio and five sentences with vocoded speech as many times as they wanted. This training did not lead to any general improvement compared to the condition when a normal, front face view was displayed, but some sentences were better perceived with the AR view. These sentences contained more phonetic features that are difficult to see in a normal face view, such as the tongue dorsum lift for [k, g] or tip lift for [l, r], especially when they appeared in clusters (e.g., [kl, kr]).

1.2. Animating tongue movements

Two of the studies above [2, 3] used rule-based synthesis, whereas one [1] instead controlled the animations through inversion of actual tongue movements for the stimuli, measured with Electromagnetic articulography (EMA).

In this study, we investigate if the type of animation influences the speech perception results. It could be the case that real tongue movements are easier to speech read, because the listener unconsciously maps displayed movements to his or her own. This would happen, e.g., if mirror neurons [7] are acti-

Figure 1: *An augmented reality view of the talking head displaying tongue movements.*



(a) Speaker (b) Talking Head

Figure 2: *(a) Placement of the reflective markers. (b) Corresponding virtual markers in the talking head model.*

vated when seeing tongue movements or if the speech motor theory is applicable [8]. These theories stipulate that neurons used when performing an action are activated by seeing or hearing the same action and if this is the case when seeing tongue movements, perception results may improve with realism. It could however also be the case that rule-based movements give more support, since the targets are often more hyperarticulated (as they are determined from static articulations) and the movements have less variability.

In this paper we investigate whether any of these two cases can be observed, or if real and rule-based tongue movements give equal support for speech perception; either because the coarticulation model is adequate enough to activate the same cognitive processes as real movements, or because the two types of movements are equally unfamiliar.

2. Experimental condition

The experiments conducted to test the above hypotheses were performed using the AR side-view of a talking head shown in Fig. 1, created by making the skin at the cheek transparent.

2.1. The talking head

The model consists of 3D-wireframe meshes of the face, jaw and tongue that are shaped by articulatory parameters, such as jaw opening, lip rounding and tongue body and tip raise. The tongue model is based on a statistical analysis of Magnetic Resonance Imaging (MRI) data of a Swedish subject producing static vowels and consonants [9].

Two different types of audiovisual animations were shown to the subjects of the study; one with real articulatory movements (AVR) and one with synthetic movements (AVS). For the AVR condition, the parameter values were determined directly from simultaneous and spatially aligned measurements using the Qualisys motion capture system (for the face) and the Movetrack EMA (for the tongue movements) of one female speaker of Swedish [10]. The Qualisys data consists of the 3D-coordinates in each frame of 28 small reflectors glued to the speaker's face, as shown in Fig. 2(a). The EMA coils were placed on the tongue (approximately 8, 20 and 52 mm from the tip), on the jaw and on the upper incisor.

The estimation of the parameter trajectories is described in detail in [10], but in brief it consisted of minimizing the error function $\varepsilon(\boldsymbol{y}) = \varepsilon_{fit}(\boldsymbol{y}) + w_{vol} \cdot \varepsilon_{vol}(\boldsymbol{y}) + w_{range} \cdot \varepsilon_{range}(\boldsymbol{y})$ by finding the optimal combination of parameter values $\boldsymbol{y}$. $\varepsilon_{fit}(\boldsymbol{y})$ is the absolute difference between the positions of the real Qualisys/Movetrack markers and the corresponding virtual markers in the model, c.f. Fig. 2(b). $\varepsilon_{vol}(\boldsymbol{y})$ penalizes differences in tongue volume and $\varepsilon_{range}(\boldsymbol{y})$ is a penalty function for breaking the parameter ranges defined in the statistical model. w_{vol} and w_{range} are empirically derived weights.

The AVS animations were generated with a rule-based visual speech synthesizer [11] with time-alignment to the original acoustic signal. The rule-based synthesizer uses a coarticulation model developed for facial animation, in which the control parameters for each viseme have either specified targets or are left undetermined. If undetermined, their values are inferred from the context through linear interpolation and smoothing of the resulting curve. The targets and the timing for the tongue movements are based on data from static MRI and dynamic EMA [9]. However, the interpolation algorithm is the same as for the face, which may or may not be adequate to create realistic intraoral movements.

In addition, an acoustic only (AO) condition was included in order to provide a baseline for recognition scores without the support of tongue movements.

2.2. Stimuli and subjects

The stimuli consisted of 27 VCV words and 50 short, simple Swedish sentences spoken by a female speaker who has been rated as highly intelligible by hearing-impaired listeners. The stimuli was acoustically degraded using a noise-excited three-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 [12].

The VCV stimuli were non-sense words with the consonants C=[v, d, g, l, r, n, s, ʃ, ç] in the different vowel contexts V=[a, i, u]. The sentences were 3, 4 or 5 words long with an "everyday content", e.g., "Gardinen var för kort" (The curtain was too short) and are part of a set of 270 sentences designed for audiovisual speech perception tests, based on [13]. VCV words and sentences were divided into three sets S1, S2 and S3. For the VCV words the sets were equally large, with one instance of each of the 9 consonants per set, but with varied vowel context, so that no VCV word was presented twice. For the sentences, S1 contained 10 stimuli and S2 and S3 20 stimuli each.

The main goal of the current study is to investigate potential differences between AVR and AVS and this was made with two subject groups I-II that were presented the animations of S2 and S3 in opposite conditions (Group I: S2 in AVS, S3 in AVR; Group II: S2 in AVR, S3 in AVS). The subjects were also presented set S1 in AO to provide an acoustic only baseline, but since this set is different, any comparison with AVS or AVR results is based on the assumption that S1 is equivalent to S2 and S3. For the VCV words, the assumption may hold, since vowel contexts were randomly distributed between sets, but for the sentences, semantic or phonetic complexity could differ between the sets. A matched control group III that heard all stimuli in AO was therefore used for comparisons between the audiovisual and acoustic only conditions. The results on set S1 were used to adjust the scores of the control group so that the AO baseline performance corresponded to that of groups I-II, since inter-group difference could otherwise make inter-condition comparisons invalid.

20 normal-hearing native subjects participated in the tongue reading test, 13 male and 7 female (aged 22-45 years), and 10 subjects in the matched audio only control group.

2.3. Experimental set-up

The acoustic signal was presented over headphones and the graphical interface was displayed on a 17" flat screen. The stimuli order was semi-random, i.e., the distribution of different display conditions was balanced to avoid artifacts caused by learning effects. The sentence order was the same for all subjects, which means that the relative AVR-AVS condition order was reversed for groups I and II.

Each stimulus was presented three times, as in [2], before the subjects should type in their answer into the allocated frames in the graphical interface. One entry frame was presented for the VCV words and five for the sentences, regardless of how many words the sentence contained.

The entire experiment, including familiarization and test of the 27 VCV words and 50 sentences, lasted 30-40 minutes.

3. Results

The consonant and word accuracy rates were counted manually, disregarding spelling and alignment errors for the sentences. We focus on the difference between the real and synthetic tongue movements, but the results compared to the acoustic only condition should also be considered. Fig. 3 shows two different AO scores, one that is for the same subjects but different stimuli (AOb) and one for the control group, i.e. different subjects, but same stimuli, with inter-group differences removed (AO). Compared to the AO score, the visualizations of tongue movements resulted in significantly better speech perception, with levels of significance indicated in Fig. 3. Using the assumption that sentences and subjects are similar, the results can also be compared on a general level with those of [3], where the AO score was 59% and the AV score with a front face view (with or without the AR side view with synthetic tongue movements) was 69%. It hence appears that the AR side view with synthetic tongue movements gives less support than a more familiar front view, but gives a similar improvement with real movements (but this can not be concluded with certainty, since different subjects and partly different sentences were used).

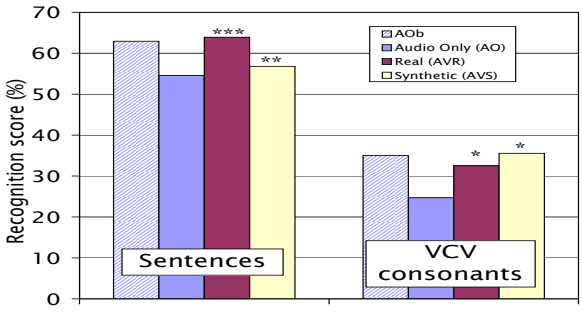


Figure 3: *Average number of correctly identified words in the sentences and consonants in the VCV words for different presentation conditions. AOb is the Acoustic Only baseline score for Groups I-II. *, ** and *** indicate that the difference compared to AO is significant at $p<0.05$, $p<0.005$ and $p<0.00005$, respectively, using a paired two-tailed t-test.*

3.1. VCV words

Since an open response format was used and no information on voicing is provided by the vocoded acoustics, corresponding

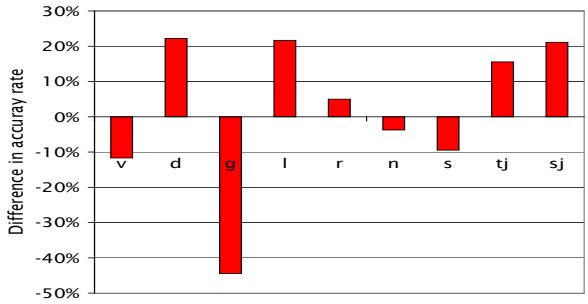


Figure 4: *AVR–AVS difference in average recognition score for each of the nine consonants. Values above zero indicate that the recognition score was better in AVR condition.*

unvoiced phonemes were counted as correct for [v, d, g]. In addition, correct spelling alternatives (tj, sj, sh, sch, ch) were accepted for the two fricatives [ɕ, ʃ].

The average results show that the rule-based movements result in a slightly better (3%) perception for VCV words, but the difference is non-significant. The VCV words could hence be interpreted as giving some weak support to the hypothesis that the more hyperarticulated synthetic tongue shapes give more information. Fig. 4 indicates that the superiority of the rule-based movements is almost entirely due to the clearer animation of [g], for which the recognition is significantly better (at $p<0.05$ with a two-tailed t-test). A similar case, but without a significant difference, appears for [v], where the lip retraction in the labiodental closure is hyperarticulated for the rule-based synthesis. It is hence not the case that the rule-generated animations are better in general, but rather that the score is similar to or inferior to that of the real movements, except for some phonetic features that may be better perceived with hyperarticulation.

3.2. Sentences

For the sentences on the other hand, the real movements resulted in a 7% higher word accuracy rate and the difference is highly significant ($p<0.005$). As illustrated in Fig. 5, the recognition score is higher in AVR for a large majority of the sentences (28 out of 40), and for a handful the difference is striking (for one it is significant at the sentence level at $p<0.0005$). Analyses of the sentences to identify the causes of the differences are difficult, since the context have a high influence. The six sentences that were more than 15% better in AVR nevertheless suggest some issues, even though the following examples are far from exhaustive. a) Real tongue movements may clarify the number and nature of phonemes better, hence avoiding insertions, e.g., [poː isɛn] ("on the ice") perceived as [polisɛn] ("the police") by six of the subjects in the AVS condition and by five in AO, compared to one for the AVR. b) Consonant clusters combining alveolars and palatals, such as [lk, rj], seem to have been easier to perceive in the AVR display. c) Words beginning with [h] were more problematic in AVS condition, which might be explained by the fact that [h] was not included in the MRI database [9] and its tongue shape target for the rule-based synthesis is hence created through interpolation. d) The rule-based lip articulation sometimes caused confusions, such as between [b] and [v].

The results suggest that even though the subjects are unaccustomed to seeing tongue movements they have an unconscious notion of how the tongue should move, and this is dif-

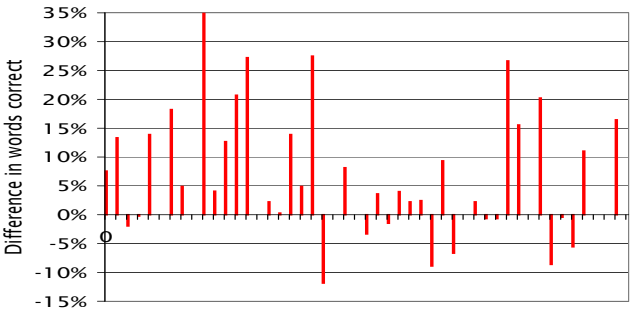


Figure 5: *AVR–AVS difference in word accuray rate for each sentence. Values above zero indicate that the recognition score was better in AVR condition.*

ferent from what the rule-based synthesis creates. The consequences of this finding are further discussed in Section 4.

Finally, the intersubject variations should be considered. As shown in Fig. 6, ten of the subjects were more than 10% better with AVR animations (one being 31% better) and only one was more than 5% better with AVS. As each individual subject did not see the same sentences in the two conditions, one should be careful in attributing importance to intrasubject differences, since it may instead be due to the sentence content. A weighted intrasubject difference was therefore also calculated to remove the influence of sentence content, i.e., the results for the two sets S1 and S2 were weighted to have the same average over all three subject groups and conditions. This gave scale factors of 0.97 and 1.03 for S1 and S2, respectively. With this weighted measure, four subjects were more than 10% better with AVR, seven were 5-10% better, six were 5-10% worse and one was more than 10% worse. The importance of the type of tongue movements hence differed substantially between subjects, but Fig. 6 indicates that it is nevertheless reasonable to argue that the AVR animations are better for a majority of the subjects.

4. Discussion and conclusions

The results of this study indicate that prototypic, hyperarticulated tongue shapes may result in a better audiovisual perception, but only for some specific phonemic features (e.g., the velar closure of [g]) and in single phoneme classification.

For sentences, real tongue movements gave significantly better results, indicating that the current rule-based coarticulation model is not optimal for tongue movements. The coarticulation model was specified for the face, for which it has proved to be effective in speech perception tests, but since tongue movements are generally much faster and with larger articulation ranges, this model seems to be insufficient.

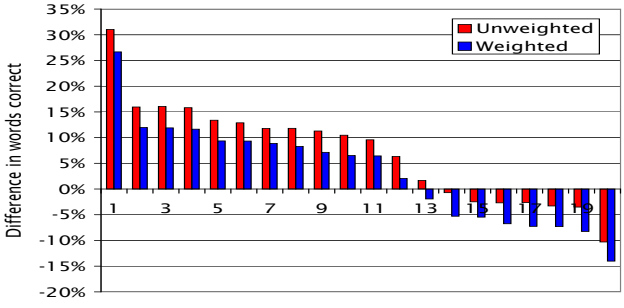


Figure 6: *AVR–AVS difference in word accuracy rate for each subject. Subjects have been sorted in order of decreasing (AVR-AVS) value, and values above zero indicate that the recognition score was better in AVR condition.*

We are therefore currently working on a new coarticulation model specific for intraoral movements, based on articulatory data from EMA to create more realistic tongue movements.

As a follow up study, we are further planning to investigate if the subjects are aware of the nature of the tongue movements (real or synthetic), or if the interpretation is subconscious. We foresee a combination of a classification task, in which the subjects should indicate if one animation was real or synthetic, and a discrimination task, in which the subjects should indicate if two animations are different, both real or both synthetic. The awareness test will be performed with the new, data-based coarticulation model as well as with the rule-based. These tests could shed additional light on the issues of motor theory in speech perception [8]. If subsequent tests confirm that real tongue movements are indeed perceived better, then this may be an indication that speech perception involves articulatory interpretation by the listener.

5. Acknowledgments

This work is supported by the Swedish Research Council project 80449001 Computer-Animated LAnguage TEAchers (CALATEA). The parameter fitting to articulatory data described in Section 2 was performed by Jonas Beskow.

6. References

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

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

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

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

[5] Agelfors, E., Beskow, J., Dahlquist, M., Granström, B., Lundeberg, M., Spens, K.-E., and Öhman, T., "Synthetic faces as a lipreading support," in ICSLP, 3047–3050, 1998.

[6] Cornett, O. and Daisey, M., *The Cued Speech Resource Book for Parents of Deaf Children*, National Cued Speech Ass., 1992.

[7] Rizzolatti, G. and Arbib, M., "Language within our grasp," Trends Neuroscience, 21:188–194, 1998.

[8] Liberman, A., *Speech: A special code*, MIT Press, 1996.

[9] Engwall, O., "Combining MRI, EMA & EPG in a three-dimensional tongue model," Speech Commun., 41/2-3:303–329, 2003.

[10] Beskow, J., Engwall, O., and Granström, B., "Resynthesis of facial and intraoral motion from simultaneous measurements," in ICPhS, 431–434, 2003.

[11] Beskow, J., "Rule-based visual speech synthesis," in Eurospeech, 299–302, 1995.

[12] Siciliano, C., Williams, G., Beskow, J., and Faulkner, A., "Evaluation of a multilingual synthetic talking face as a communication aid for the hearing impaired," in ICPhSc, 131–134, 2003.

[13] MacLeod, A. and Summerfield, Q., "A procedure for measuring auditory and audiovisual speech-reception thresholds for sentences in noise. Rationale, evaluation and recommendations for use," Brit. J. Audiol., 24:29–43, 1990.